Deliverable D7.4

Multilevel modelling and time series analysis in traffic safety research – Methodology

Dupont, E. and Martensen, H. (Eds.) (2007) Multilevel modelling and time series analysis in traffic research – Methodology. Deliverable D7.4 of the EU FP6 project SafetyNet.

Contract No: TREN-04-FP6TR-S12.395465/506723
Acronym: SafetyNet
Title: Building the European Road Safety Observatory

Integrated Project, Thematic Priority 6.2 "Sustainable Surface Transport"

Project Co-ordinator:
Professor Pete Thomas
Vehicle Safety Research Centre
Ergonomics and Safety Research Institute
Loughborough University
Holywell Building
Loughborough, LE11 3UZ

Organisation name of lead contractor for this deliverable:
Belgian Road Safety Institute (IBSR)

Due Date of Deliverable: 28/02/2007

Submission Date:

Editors: E. Dupont & H. Martensen (IBSR)

Contributing Authors: C. Antoniou (NTUA), C. Brandstaetter (KfV), R. Bergel (INRETS), M. Cherfi (INRETS) F. Bijleveld (SWOV), J.J.F. Commandeur (SWOV), C. DuBlois (SWOV), E. Dupont (IBSR), M. Gatscha (KfV), H. Martensen (IBSR), E. Papadimitriou (NTUA), W. Vanlaar (IBSR), G. Yannis (NTUA)

Project Start Date: 1st May 2004 **Duration:** 4 years

Project co-funded by the European Commission within the Sixth Framework Programme (2002 -2006)		
Dissemination Level		
CO	Public	

Transport

Project co-financed by the European Commission, Directorate-General Transport and Energy

Table of Contents

Executive Summary

The SafetyNet project is set up to build a European Road Safety Observatory. The data assembled or gathered for the observatory consist of the Community database on Accidents on the Roads in Europe (CARE); data on road safety risk indicators; data on road safety performance indicators and in-depth accident data. Potential users will link data from different data-sets, consider different levels of aggregation jointly, and analyse the development over time. Work package 7 (WP7) is set up to deal with statistical and conceptual issues that come into play when analysing such complex data structures.

One of WP7's main objectives is to develop a best practice advice for the analysis of data structures that require more than the standard statistical tools. This best practice consists of D7.4 "Multilevel modelling and time series analysis in traffic research – A methodology" and D7.5 "Multilevel modelling and time series analysis in traffic research – The manual".

The main goal is to enable the reader to deal with complex data-structures that show dependencies in space (nested data) or in time (time series data). At first it is demonstrated how such dependencies can compromise the applicability of standard methods of statistical inferences, because they can lead to an underestimation of the standard error and consequently of the error in statistical tests.

As a solution to this problem, two families of statistical techniques are presented to deal with these dependencies. *Multilevel Modelling* is dedicated to the analysis of data that are structured hierarchically. It offers the possibility to include hierarchical structures into the model of analysis. In road-safety research, multilevel analyses allow for the introduction of exposure data and of safety performance indicators, even if those are not specified at the same level of disaggregation as the accident data themselves. In this way, multilevel analyses allow a global and detailed approach simultaneously. *Time series analyses* are employed to overcome dependency issues in time-related data. They allow describing the development over time, relating the accident-occurrences to explanatory factors such as exposure measures or safety-performance indicators (e.g., speeding, seatbelt-use, alcohol, etc), and forecasting the development into the near future.

Deliverable D7.4 gives the theoretical background for these two families of analyses. For each technique the objectives, detailed model formulation, and assumptions are described and subsequently the technique is illustrated with an empirical example relevant to traffic safety research.

Chapter 1 - Introduction

(Heike Martensen and Emmanuelle Dupont, IBSR) [1]

This deliverable has been produced in Workpackage 7 (WP7) of the SafetyNet project, an Integrated Project that brings together the most experienced road safety organisations within the EU to assemble a co-ordinated set of data resources that together will meet the EC needs for policy support. The goal of the project is to set up a *Road Safety Observatory* that will enable the European Commission to monitor progress towards targets, identify best practises, and ensure that new regulatory and other safety actions will result in the maximum casualty reduction.

The data assembled or gathered within the SafetyNet project consist of the Community database on Accidents on the Roads in Europe (CARE); road safety risk exposure data; data on road safety performance indicators and in-depth accident data. The data will be available to the entire road safety community and will serve to answer a broad variety of questions.

Road traffic data is structured in space and in time. For example, accident numbers can be disaggregated to countries, regions, and counties, as well as to years, months, weeks and days. In many cases, data at different levels of aggregation will be considered jointly, and the development over time will be analysed. WP7 is set up to deal with statistical and conceptual issues that come into play when analysing such complex data structures. One of its main objectives is the development of a best practice for the analysis of data structures that require more than the standard statistical tools.

In Section 1.1 of this introduction the linear regression model that forms the basis for the majority of all analyses is introduced shortly. It will then be explained why the basic model is not sufficient for many road-safety analyses and demonstrated that additional requirements for the analysis of complex data structures are mainly related to recognizing and dealing with dependencies in space and time. In Subsection 1.2 two families of sophisticated analysis techniques are introduced that allow road-safety researchers to deal with these dependencies: multilevel modelling and time series analysis. Based on several empirical traffic-safety examples it is illustrated that both are very valuable to traffic safety research. The use of those techniques in the field of traffic safety is advocated. At the end of this introductory Section (1.3) an overview of two WP7 deliverables (7.4 & 7.5) will be given that form together the best-practice advice for the analysis of complex data structures.

[1] An earlier version of the introduction was written by Ward Vanlaar. Where relevant, quotes or references to Vanlaar's work have been inserted in the present version."

.

1.1 Best practice for the analysis of complex data-structures

1.1.1 What is a statistical model

Many, if not all, road-safety questions require that different quantities or categories are linked to one another and seek at establishing whether there is a relation between them.

Examples are questions like the following: Is there a relation between the number of speed controls and the number of people killed in a traffic accident? Does the number of errors a driver makes depend on the number of years he/she has been driving? Did the number of people killed in accidents decrease after the introduction of the seat-belt law?

Statistically these questions are expressed as relations between variables. An *observed* or *dependent* or *endogenous* variable[2] y_i (e.g., the number of driving errors person *i* makes) is predicted by one or more *explanatory* or *independent* or *exogenous* variables x_1, x_2.. (e.g., the number of years of driving experience person *i* has, his or her age, gender, etc.). Such a relation is modelled by Equation 1.1.1, where *e* is the *error* term, also called the *disturbance* term and i=1…n, with n the number of persons.

.

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + e_i \qquad (1.1.1)$$

In principle, the number of explanatory variables is not limited, but for simplicity reasons the model here considered as an example includes only one independent variable. The number of errors of driver *i* is predicted by his/her number of years driving experience[3]. This relation is modelled in 1.1.2

$$driving\_errors_i = b_0 + b_1 years\_experience_i + e_i \qquad (1.1.2)$$

The parameters or coefficients (here b_0, and b_1) quantify the relation between the independent and the dependent variable. The intercept b_0 indicates the average value of the dependent variable when all independent variables are zero. Here the intercept is the number of errors at 0 years experience, i.e. during the year following receiving one's driving licence. The coefficient b_1

[2] Different research traditions (e.g., multilevel and time series modelling) have generated different terms for the same concepts. They are listed next to each other here to enable the reader to make the link.

[3] The relation between experience and the number of errors is not linear. In practice this could be solved by transforming one of both variables or applying nonlinear models (see, e.g. section 3.2.3). For simplicity sake the nonlinearity will be ignored at this point.

indicates how much the average number of errors decreases with each year of driving experience.

A statistical model determines an expected value for each observation on the basis of the independent variables. However, in practice the independent variables can never perfectly predict the value of the dependent one; the observations always depart from the values predicted. Every useful statistical model therefore has a *fixed* or *deterministic* or *structural* part -- the variation in the dependent variable that can be predicted by the independent variables -- and a *random* or *stochastic* part -- the variance that cannot be predicted, the error or disturbance.

1.1.2 Assumptions of statistical models and their violations

Defining the relation between variables in equations is called "modelling" because the equations do not describe the true relation between these variables; they rather give a simplified model of it. The common linear regression model described so far contains a number of restrictions, most notably the following:

1. The dependent variable (y) has to follow the normal distribution.
2. The dependent variable (y) can be expressed as a linear combination of the independent ones ($b_0+b_1x_1+b_2x_2\ldots$)
3. The errors e_i (the part in the dependent variable that *cannot* be explained by this linear function) are independently distributed across all observations.

In reality, these assumptions seldom hold. Violations of the first two assumptions can often be dealt with in the Generalized Linear Model (GLM) described in sections 2.3.1 and 3.2.2 of this document. The GLM allows modelling observations that do not follow the normal distribution (e.g. discrete responses). In nonlinear models, described in section 3.2.3, relations between dependent and independent variables are analyzed that do not need to have the linear form (they can follow the exponential function, for example).

The focus of the present document, however, is on the third assumption, the assumption of independence. A statistical model determines an expected value for each observation. In this way the dependency of the observed data is modelled in the structural part with exogenously measured factors (the explanatory variables). Nevertheless, the observations spread around their expected value. The "independence assumption" refers to this random part of the model. By saying that the observations must be independent, we mean that the deviation of any one observation from its expected value must not be linked to the deviation of another. [4]

We will present two examples showing that this assumption of independence can be unrealistic in road safety research. Generally, two types of commonly

[4] In practice this means that the prediction error e is uncorrelated with x and the error associated with one value of y has no effect on the errors associated with other values, i.e. all observed autocorrelations of the errors are 0.

occurring dependencies in data can be distinguished: hierarchically dependent data and time dependent data. We will describe why these dependent data are problematic for the traditional statistical methods and present a variety of techniques allowing to deal with these problems.

1.1.2.1. Hierarchically dependent data (nested data)

In a Belgian study on speeding, cameras were set up at a large number of randomly selected road sites and the speed of all cars passing through was registered. Speeds at the same road-site are usually more similar to each other than data coming from different clusters. For example, if the first car recorded drove 30 km/h, the probability of the next car passing through with 110 km/h is much smaller than if the first car recorded had driven 120 km/h. As mentioned above, this dependency can be modelled in the structural or fixed part of the model, by including explanatory factors that predict the differences between cars at different road-sites. In our example, the speed-limit would come to mind. However, other characteristics of the location (e.g. quality of road, traffic count, … and probably some characteristics the researcher is not aware of) affect the driving speed as well. So it will never be possible to perfectly model the differences between the road-sites in the structural part of the model, which means that the errors will not be independent from each other, as required by the assumptions for linear regression. The next section describes how this problem can be more efficiently and elegantly dealt with by including the road-sites into the random part of the model[5].

1.1.2.2. Time dependent data (time-series)

For many questions in road-safety research, the annual or monthly numbers of road traffic accidents are considered. Again the assumption for traditional statistical methods would be that the numbers at each point in time show an independent deviation from some expected value (e.g. the overall average). Like in the example of road sites above (where local variations in conditions should be considered), the possibility should be considered that temporal variation in conditions stretching over multiple observations could increase or decrease the expected number of accidents in addition to what would otherwise be expected.

Identifying the fact that these variations exist in a particular series of data and quantifying these variations over time allows the researcher to enhance inference and prognosis.

[5]Another source of dependency are cohort effects. Cars that follow each other closely will be more similar in speed than to other, more remote cars, because their speed is dictated by the slowest car driving in front. Again this could partly be captured by taking up traffic concentration in the fixed part of the model but it could be modelled more elegantly by including the cohort structure into the random part of the model.

1.1.3 What to do with dependent observations?

In regression models the variance of a dependent variable $y_1, \ldots, y_n$ is split up in two parts: That part of the variance that can be predicted by a combination of independent variables (the fixed or structural part) and the error, the part of the variance that cannot be predicted (the random part). In traditional regression models the random part consists of only one variable (e_i in equations 1.1 and 1.2), reflecting the idea that there is only one source of random variation, the individual unit of measurements.

With highly structured data, as we often deal with in road-safety research, this assumption is not realistic. Each data point must in fact be considered to be sampled from different populations at the same time. For nested data structures these populations correspond to the levels of the data hierarchy. For example the registered speed depends on which car had been randomly selected, but also on the cohort the car arrives in and also on which road-site had been randomly selected from the population of all road-sites in Belgium. This means that the prediction for this particular car contains a random effect, which is shared by all cars that arrived in the same cohort and another that is shared by all cars at a particular site. These random effects allow those cars that share it to deviate in similar way from the average car in the study.

For time series the resulting situation is similar in that a specific structure is imposed on the random term, for instance by also introducing additional random terms.

While the traditional regression models described above assume that there is only one source of random variation, it is important to structure the random part of a statistical model according to the nature of the statistical units. Multilevel models therefore introduce random variation at each level of the data hierarchy and time series models introduce random variation that is specific to the transition from one point in time to another.

Ignoring the structure of the random variation and thus the dependence of residuals generally causes standard errors to be either over- or underestimated (see for example Rasbash et al., 2004, for a discussion focused on multilevel models where usually underestimation is observed), which will in turn distort the estimated probability of having observed a particular effect on a purely coincidental basis. Both consequences, (1) accepting as significant a result that is actually not so, and (2) rejecting a result as due to chance that is in fact not due to chance, can occur in sometimes unpredictable ways.

1.2 The added value of Multilevel and Time Series Analysis

For the development of a best practice for the analysis of complex data, it is necessary to give an overview of methods to deal in one way or another with dependencies in data. In the following the added value for road safety research for two families of analysis will be described separately: Multilevel modelling that is dedicated to data with hierarchical dependencies and time series analyses that are dedicated to time-dependent data.

1.2.1 Multilevel models

Heike Martensen and Emmanuelle Dupont (IBSR)

1.2.1.1. Definition and conceptual issues

There are several introductory books on multilevel analysis are available (e.g., Goldstein, 2003; Heck and Thomas, 2000; Hox, 2002; Kreft and de Leeuw, 2002; Leyland and Goldstein, 2001; Snijders and Bosker, 1999) and each of those defines them in a specific way. However, these definitions share one concept, namely the concept of hierarchies or nested data structures. There are individuals and variables describing these individuals, but there are also larger units the individuals are grouped into and variables that describe these larger units (Raudenbush and Bryk, 2002).

Multilevel models as they are presented here have mostly been developed in educational and social research (e.g., Aitkin & Longford, 1986, Kreft, 1994, Kreft et al.,1995), where many objects of investigation are hierarchically structured. (e.g.: pupils in classes; classes in schools; employees in departments, departments in firms; suspects in courts; offspring within families). However, structurally identical methods are commonly used in other disciplines. In bio-medical sciences these models are often referred to as mixed-effects or random-effects models (Bates & Pinheiro, 1995) and are used for growth curve analyses, (Lindsey, 1993), survival analyses (Sargent, 1998), and epidemiological analyses (Diez-Roux, 2002, Carrière & Bouyer, 2002) among others. In econometrics the same models are known as random-coefficient regression models (e.g. Longford, 1993) and are, for instance, used for analysing risk-return tradeoffs (Lee, et al., 2006) and panel data (Swamy, 1971; Hsiao & Pesaran, 2004). Although the first multilevel models concerned linear models, they have been extended for the use of binary and count data (Lee & Nelder, 2001) and nonlinear analyses (Pinheiro & Bates, 1995).

Although multilevel models are widespread in many scientific disciplines, they are relatively new to the field of road-safety research and applied only in a small number of studies. This is all the more concerning, as nested data structures are the rule rather than the exception in this field. In *roadside surveys*, like speed measurements (section 2.2), seat-belt counts (1.2.1), or alcohol controls (2.3.2 and 2.3.3), individual cars are nested within measurement locations.

Accident and victim numbers are hierarchically structured according to spatial or administrative units like counties and regions. The same is true for statistics describing enforcement activities, like the number of speeding infringements or alcohol controls. In section 2.3.4 and 2.4 it is described how multilevel modelling can be applied in such a structure and it is demonstrated that not only the number of accidents varies across regions in Greece, but also the relation between accident number and number of enforcement actions.

Accidents show a hierarchical structure because drivers and passengers are nested in vehicles, vehicles in accidents, accidents in regions (Jones and Jørgensen, 2003). Moreover, multilevel models can be applied to *repeated measurements* of, for instance driving performances, where the performance scores are nested within the individuals that produced them (Burns et al. 1999, see also section 2.4 in this document). Meta-analyses (e.g., Delhomme et al., 1999; van Driel et al., 2004) also show a nested data structure, where data points are nested within studies.

As an example we will show a how a traditional linear regression model on a large sample of accidents can be extended to represent multiple levels of the accident. For each accident the severity of injury for each passenger is established. (For simplicity, we will assume that there is a quantitative measure of injury severity that is approximately normal distributed). Simultaneously possible explanatory variables, for example, age of the victim are also measured. The severity of injuries will to some extent be explained by the age of the victim. In model terms:

$$severity_i = \beta_0 + \beta_1 age_i + e_i \qquad (1.2.1)$$

In the present example, we are dealing with a hierarchical data structure, because each injury is not only determined by characteristics of the victim, but also by the accident the victim was part of and the vehicle the victim was in. Factors such as speed, type of collision, and type of vehicle are characteristics of accidents that affect all victims inside a particular vehicle in the same way. As a consequence the injury severities of victims that have been in the same vehicle will be more similar to each other than to those of other victims.

The solution in multilevel modelling is to assume that random variation not only occurs at the level of the basic measurement unit (i.e. occupant), but also at higher-level measurement units (e.g. the vehicle).

A very simple multilevel extension of equation (1.2.1) would be to let the intercept β_0, which indicates the general level of severity, vary across the secondary measurement unit *j* (here the vehicles).

$$severity_{ij} = \beta_{0j} + \beta_1 x_i + e_{ij} \qquad (1.2.2)$$

$$\beta_{0j} = \mu_0 + u_j \qquad (1.2.3)$$

In this model there is random variation associated with each measurement at the lowest level of the hierarchy (e_{ij}, e.g. random variation between victims) but also at a second level (u_j, e.g., random variation between vehicles). Due to this second level variation, there is a different intercept (β_{0j}) for each level-two unit (i.e. a different mean injury severity for each vehicle). In this way, multilevel models explicitly include a hierarchical structure resembling the one present in the data.

The consequences of ignoring a hierarchical structure form two broad categories: statistical problems and conceptual problems. The first type of problems has been mentioned before. Due to the dependence of the observations in hierarchical data structures, there is a risk to underestimate the standard errors and therefore to consider as significant a result significant that is in fact due to chance (Rasbash et al., 2004). The conceptual problems result from the existence of variables affecting different levels in the data hierarchy and from their possible interactions. Variables related to higher-order levels are also referred to as contextual information.

In the following paragraphs both types of problems (statistical and conceptual) will be briefly discussed and illustrated with examples from road-safety studies. First, consequences of ignoring dependence of nested observations are investigated and data from an observational study on seatbelt use are used as an illustration. Then, consequences of impoverished conceptualisation of contextual information are discussed. Finally conclusions regarding multilevel modelling in traffic safety are drawn.

1.2.1.2. Consequences of ignoring dependence of nested observations

In a Belgian study on seatbelt use (Verbeke, Vanlaar, & Silverans, 2005) observers were situated at 150 different road-sites. For each car passing, they determined the gender of the driver and the front passenger and whether they were wearing a seat-belt or not. In total, this information could be determined for 21.785 cars.

Because of the sampling plan, the individual cars were not selected independently from each other but in clusters. Due to randomly selecting a number of road-sites (the clusters) first, not all Belgian cars (and their inmates) had the same probability to be observed. The sampling strategy resulted in a hierarchical data-structure. Many factors that possibly affect seat-belt use (e.g. design speed of the road, weather, time of the day) are the same for all participants observed at the same road site and as a consequence the probability of car-inhabitants wearing seat belts will be more similar for cars measured at the same road site than for cars recorded at different ones. [6]

[6] Another possible clustering effect in seatbelt observation studies could be that of occupants of the same car. Indeed, it is reasonable to assume that the seatbelt wearing behaviour of one person is more similar to occupants of the same car as it is to that of other cars' occupants. The dependence introduced by such a "car effect" does however not apply to the Belgian study discussed here, because the road-side observers in this study registered either whether the

The problem of dependent observations in complex sampling designs is not a new one and for the analysis of such designs elaborate correction procedures are available to correct the standard errors (Cochran, 1963; Kish, 1965; Levy and Lemeshow, 1999). Addressing this problem with multilevel modelling as demonstrated below, has however the advantage that the population structure, insofar as it is mirrored in the sampling design, is not only seen as a 'nuisance factor' but can be used to collect and analyse data about the higher level units in the population Goldstein (2003: p. 5).

Parameter	Single-level logistic model			Two-level logistic model		
	Logit coefficients	s.e.	P	Logit coefficients	s.e.	p
Fixed parameters						
Intercept	0.883	0.169	0.000	0.776	0.184	0.000
Passenger	-0.260	0.130	0.046	-0.205	0.132	0.120
Male	-0.663	0.121	0.000	-0.670	0.114	0.000
Wallonia	-0.454	0.158	0.004	-0.510	0.182	0.005
Brussels	-0.583	0.137	0.000	-0.365	0.140	0.009
50km/h	0.648	0.137	0.000	0.649	0.171	0.000
70km/h	0.921	0.171	0.000	0.665	0.155	0.000
90km/h	0.461	0.159	0.004	0.433	0.191	0.023
120km/h	0.795	0.173	0.000	0.811	0.188	0.000
Weekday night	-0.092	0.214	0.667	0.037	0.156	0.813
Weekend day	-0.091	0.142	0.522	0.151	0.139	0.277
Weekend night	0.312	0.156	0.046	0.197	0.166	0.235
Random parameters						
Level 2 variance: Ω_u	n.a.	n.a.		0.197	0.039	
Level 1 variance: Ω_e	1.000	0.000		1.000	0.000	

Table 1.2.1: Results from Vanlaar 2005a: Comparison of logit coefficients and s.e. of a single-level and a two-level model regarding seatbelt use

To demonstrate how the dependence of observations causes standard errors to be underestimated, Vanlaar (2005a) compared results from a single-level model that does not take the similarity into account to those from a two level model, which explicitly includes road-sites as a source of variation. The results from both models are presented in Table 1.2.1.

The coefficients estimated in both models by Vanlaar (2005a) concerned the type of occupant (*Passenger* as opposed to driver), gender (*Male* as opposed to female), the region (*Wallonia* and *Brussels* as opposed to Flanders), the speed-

driver wore a seatbelt or whether the front passenger wore a seatbelt but never both for the same car.

regime (*50km/h*, *70km/h*, and *120km/h* as opposed to 30 km/h) and the time of testing (*Weekday night*, *Weekend day*, and *Weekend night* as opposed to weekday at daytime). Even though the significance levels of most variables were the same in both the single-level and the two-level model, Vanlaar pointed out there were two variables for which this was not the case. Those two variables were *Passenger* and *Weekend night*. Both were significant at the 5%-level in the single-level model. However, these effects were no longer significant according to the two-level model.

The proportion of level-two variance estimated in the two-level model was significant, which indicated that the data did indeed have a hierarchical structure. This example by Vanlaar (2005a) demonstrated that ignoring this structure can lead to erroneous conclusions. As he warned, based on the significant negative coefficient of front-seat passengers compared to drivers in the single-level model (meaning that the odds of front-seat passengers to wear a seatbelt are lower than those of drivers) it could for example be decided to make front-seat passengers a special target group in a mass media campaign. However, the two-level model suggested that the difference in seatbelt use between those two groups does not exceed the chance-level. More generally, single-level models applied to multilevel structures can lead to overconfident or even plain incorrect conclusions.

1.2.1.3. Consequences of impoverished conceptualisation of contextual information

Many problems in traffic research cannot be understood correctly if only one level is regarded. As an example, consider the following question: Are pedal cyclists safer on roads with cycle paths as compared to roads without? The Netherlands have by far the highest percentage of roads with cycle paths in Europe. They also have the highest rate of accidents involving cyclists. Should we conclude that the presence of cycle paths puts cyclists in particular danger? Probably not. This wrong conclusion would rise from trying to answer a question concerning the individual level (here the roads that do or do not have cycle paths) with data concerning the group level (here the countries), a tendency that is known as the *ecological fallacy* (Robinson, 1950).

In order to avoid the ecological fallacy, one might focus exclusively on the individual level. This strategy, however, might also lead to incorrect or at least incomplete conclusions. For example, it is possible that *in the Netherlands* there is no difference with respect to the number of cyclist accidents between roads with and without cycle paths. The reason would be that the only roads without cycle paths are those that are relatively safe for cyclists and that other road users are used to watch out for them. This would however, not be true in countries with fewer pedal cyclists and a smaller percentage of roads with cycle paths. The difference between cycling on roads with and without cycle path is, therefore, affected by country-level variables (overall number of pedal cyclists,

extension of cycle path network) as well. To ignore these higher-level effects is called the *individualistic* (or psychologistic) *fallacy (*Diez-Roux, 2002)

The interactions between variables measured at different levels in hierarchically structured data are referred to as *cross-level interactions* (Kreft and de Leeuw, 2002). The examination of cross-level interactions is also called *contextual analysis* which has been developed in the social sciences. There, the focus is on the effects of the social context on individual behaviour, which gave rise to the need to consider variables at different levels simultaneously. This has been the motivation for the development multilevel models in the first place (Hox, 2002; Snijders and Bosker, 1999).

Likewise, many road-safety problems involve relationships between micro-level (e.g. presence of cycle-paths) and macro-level variables (e.g. overall number of pedal cyclists). These complex problems could not be solved with analyses on either aggregated or data disaggregated. Multilevel modelling overcomes these obstacles in an elegant and productive way.

1.2.1.4. Conclusion

Although multilevel models are common in many scientific areas, they are relatively new to the field of traffic safety. The advantages of multilevel modelling compared to statistical techniques that ignore hierarchies were discussed and illustrated based on two traffic safety examples.

Two types of problems were demonstrated when ignoring a hierarchical structure in the data: statistical and conceptual. Statistical problems result from the underestimation of standard errors due to the dependence of nested observations. Data from a road-side survey on seatbelt behaviour were analysed according to a single-level model and a two-level model to illustrate this. Two effects that were significant in the single-level model were found not to be significant any longer when including the level of road-sites into the model. The model that ignores the hierarchical data structure would therefore lead to erroneous conclusions regarding variables that could have an impact on seatbelt use and ultimately, on increasing the level of traffic safety (Vanlaar, 2005a).

The second consequence is a conceptually impoverished representation. For traditional types of analyses a choice has to be made considering the level of aggregation. Based on the example of bike-safety, it has been demonstrated that analyses at the country level can lead to wrong conclusions but that analyses that include the level of individual bikers only also leave out important information. As a consequence, it was argued for the need of statistical methods that allow the analysis of variables for different levels in the data structure simultaneously.

Of course, multilevel modelling is no wonder-weapon. The assumptions that have to hold in order to apply them are plenty and will be discussed in the remainder of the document. However, when applied with caution, they can prevent overoptimistic inferences and “allow researchers to translate a research

problem into a design reproducing a lot of the nuances at stake and without giving in too drastically towards simplifying the nature of the issue under evaluation."(Vanlaar, 2005a, p. 315)

1.2.2 Time series models

Jacques Commandeur (SWOV)

Many road traffic data consist of *observations made sequentially through time*. Examples are the annual or monthly number of road traffic accidents in a country, its annual or monthly number of road traffic fatalities, its annual or monthly number of vehicle kilometres driven, its annual or monthly values on safety performance indicators, etc.. Each example is a collection of observations made sequentially through time.

Whenever one is interested in studying and analysing such developments of one and the same phenomenon over time, special issues arise not encountered in cross-sectional data analysis. In this section we will illustrate with a simple example what these special issues are, and how they can be dealt with by using a special family of analysis techniques collectively known as *time series models*.

The example consists of the log of the annual number of road traffic fatalities as observed in Norway for the period 1970-2003. It may be noted that the annual number of road traffic fatalities are count data, which are non-negative. If count data were analysed as they are, one could obtain predicted counts that are negative. By analysing them in their logarithm, however, and then taking the exponent of the predicted values, it is guaranteed that non-negative predicted counts are obtained.

Since the period 1970-2003 spans 34 years, there are $n = 34$ observations. Because the observations (i.e., the annual number of fatalities) are made sequentially through time, they are collectively called a *time series* (see Chatfield, 2004). We will first analyse this time series with traditional linear regression.

Typically, in traditional linear regression a linear relationship is assumed between a criterion or dependent or endogenous variable y, and an explanatory or independent or exogenous variable x such that

$$y_i = a + bx_i + \varepsilon_i, \qquad \varepsilon_i \sim NID(0,\sigma_\varepsilon^2) \qquad (1.2.4)$$

where $i = 1,..., n$, and n is the number of observations. The expression

$$\varepsilon_i \sim NID(0,\sigma_\varepsilon^2)$$

in (1.2.4) is a shorthand notation for: the residuals ε_i are assumed to be Normally and Independently Distributed (NID) with mean equal to zero and variance equal to σ_ε^2.

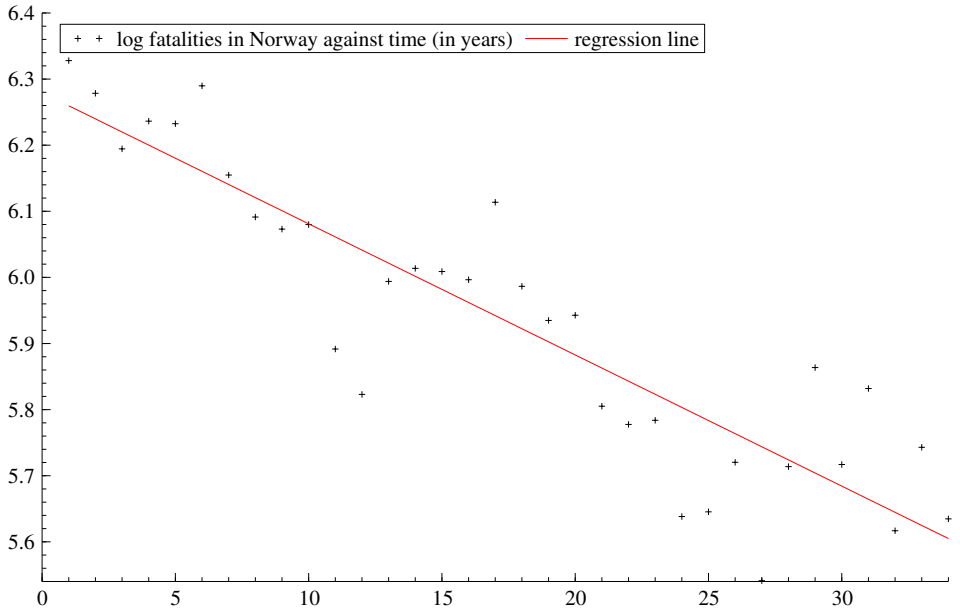


Figure 1.2.1: Scatter plot of log of fatalities in Norway against time (in years), including regression line.

Now suppose that the dependent variable *y* in (1.2.4) is the just mentioned series of the log of Norwegian road traffic fatalities. Also, suppose that the independent variable *x* in (1.2.4) consists of the numbered consecutive time points in the series (thus, $x = i = 1, 2, ..., 34$). The usual scatter plot of these two variables -including the best fitting line according to traditional linear regression- is shown in Figure 1.2.1.

The equation of the regression line in Figure 1.2.1 is

$$\hat{y}_i = 6.2794 - 0.019837 x_i,$$

with residual variance $\sigma_\varepsilon^2 = 0.00985827$. Graphically, the intercept $a = 6.2794$ in model (1.2.4) is the point where the regression line intersects with the *y*-axis. Therefore, the intercept determines the 'height' or *level* of the regression line on the *y*-axis. The value of the regression coefficient or weight $b = -0.019837$ determines the *slope* of the regression line (i.e., the tangent of its angle with the *x*-axis).

The standard *t*-test for establishing whether the regression coefficient $b = -0.019837$ deviates from zero yields

$$t = \frac{b}{\sqrt{\frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}} = \frac{-0.019837}{\sqrt{\frac{0.00985827}{3272.5}}} = -11.43.$$

Since the value of this *t*-test is associated with a *p*-value of 1E-12, the linear relationship between the criterion variable *y* and the predictor variable *x* is extremely significant.

When the assumptions for traditional linear regression are valid, time is a highly significant predictor of the log of the number of Norwegian road traffic fatalities, and there is a negative relation between these two variables: as time proceeds the log of the number of fatalities decreases.

However, one crucial issue has completely been overlooked in this analysis. The just mentioned *t*-test was based on the fundamental assumption that the 34 observations in the time series are *independent* of one another. That the observations are not independent becomes more obvious by connecting the consecutive observations in Figure 1.2.1 with lines, as has been done in the top graph of Figure 1.2.2. Inspection of the latter graph shows that the observations in a certain year tend to be more similar to the observation of the previous year than to any other earlier observation.

The dependencies between the observations are also reflected in the fact that the residuals of traditional linear regression model (equation 1.2.4) shown at the bottom of Figure 1.2.2 are *not independent* of one another. Positive values of the residuals in Figure 1.2.2 tend to be followed by further positive values, while negative values tend to be followed by further negative values.

A useful diagnostic tool for investigating whether the residuals are independent is called the *correlogram*. As will be explained in more detail in Section 3.2.1.2, the correlogram is a graph depicting the correlations between the residuals and the same residuals shifted *k* time points into the future. These correlations are therefore called autocorrelations.

The correlogram containing the first eight autocorrelations of the traditional linear regression residuals in Figure 1.2.2 takes on the form shown in Figure 1.2.3. The two horizontal lines in the correlogram are the 95% confidence limits $\pm 2/\sqrt{n} = \pm 2/\sqrt{34} = \pm 0.343$. If residuals are independently distributed then all autocorrelations in the correlogram are close to zero, and do not exceed the confidence limits. The dependence in the traditional linear regression residuals is therefore confirmed by the fact that three of the eight autocorrelations in the correlogram in Figure 1.2.3 significantly deviate from zero.

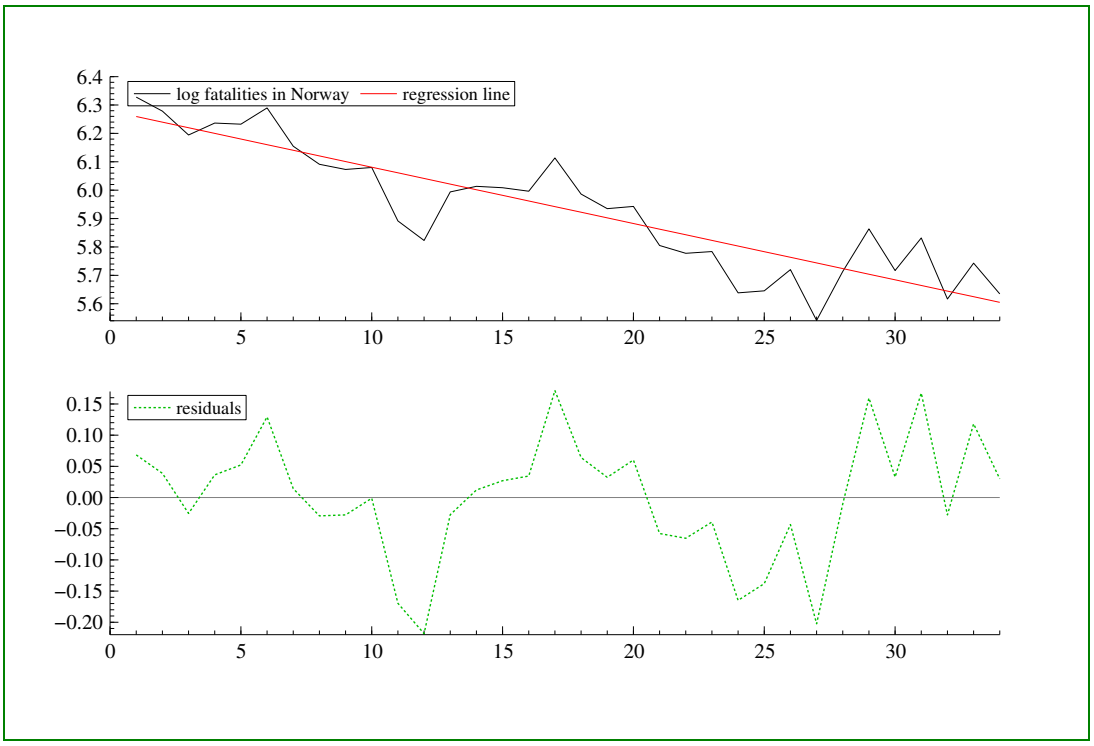


Figure 1.2.2: Log of fatalities in Norway plotted as a time series including regression line (top), and residuals of traditional linear regression analysis (bottom).

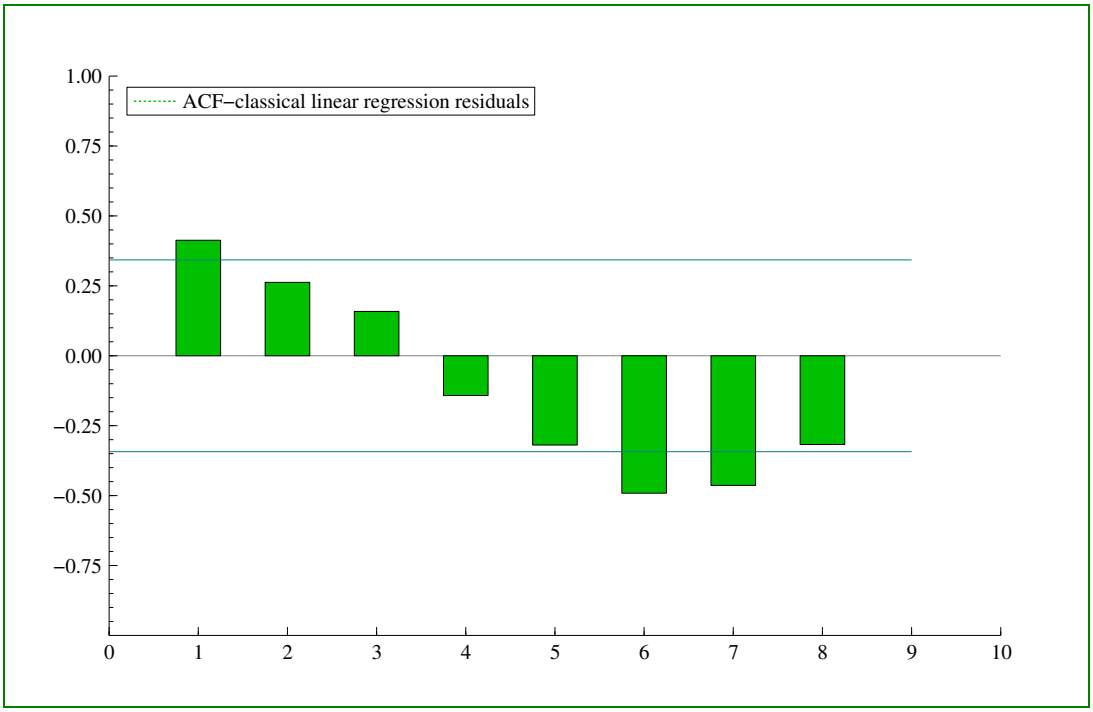


Figure 1.2.3: Correlogram of residuals of traditional linear regression of the log of the Norwegian fatalities on time.

Generally, when the first order residual autocorrelation is positive and significantly deviates from zero, a positive residual tends to be followed by one or more further positive residuals, and a negative residual tends to be followed by one or more further negative residuals. As pointed out in the literature (e.g., Ostrom, 1990; van Belle, 2002), the error variance for standard statistical tests is seriously underestimated in this case. This in turn leads to a large overestimation of the *F*- or *t*-ratio, and therefore to overly optimistic conclusions about the linear relation between the dependent variable and time.

Note that this is exactly what is found to be the case in the traditional linear regression analysis of the log of the Norwegian fatalities series discussed above: the first autocorrelation in the correlogram of the residuals is positive and significantly deviates from zero (see Figure 1.2.3), and positive residuals tend to be followed by one or more further positive residuals, while negative residuals tend to be followed by one or more further negative residuals (see Figure 1.2.2). All this implies that the value of -11.43 for the *t*-test is seriously flawed, and probably much too large.

The problem of dependencies between the residuals in the traditional linear regression analysis of time series data can be solved as follows:

1. additional predictor variables can be added to the regression of the dependent variable on time such that the dependencies are removed from the residuals;
2. the dependent variable can be analysed with (dedicated) time series analysis techniques like ARMA-type, DRAG and state space models.

To give an example in this introductory chapter, we illustrate how the time dependencies between the observations are dealt with in state space methods (Harvey, 1989; Durbin and Koopman, 2001). In state space methods it is assumed that the development over time of the system under study is determined by an unobserved number of components which are collectively called the state, and with which are associated a series of observations y_1, …, y_n. The relation between the state and the observations is specified by the state space model. The purpose of time series analysis by state space methods is to infer the relevant properties of the state given a series of observations y_1, …, y_n. State space methods handle the dependencies between the observations constituting a time series by absorbing them directly into the model. This again is achieved by allowing the intercept and/or the regression coefficient -that are constants in traditional linear regression- to *vary over time*.

The dependencies in the log of the Norwegian fatalities series, for example, can be handled by allowing the intercept in model (1.2.4) to vary over time, as follows:

$$y_t = a_t + bx_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0,\sigma_\varepsilon^2) \qquad (1.2.5a)$$

$$a_{t+1} = a_t + \xi_t, \qquad \xi_t \sim NID(0,\sigma_\xi^2) \qquad (1.2.5b)$$

where t = 1, …, n, and n is the number of observations. The second equation in (1.2.5b) allows the intercept (i.e., the level) to change from time point to time point. Moreover, in this equation dependencies in the observed time series are dealt with by letting the intercept at time t+1 be a direct function of the intercept at time t. Therefore, it takes into account that the observed value of the series at time point t+1 is usually more similar to the observed value of the time series at time point t than to other previous values in the series.

Applying model (1.2.5) to the log of the Norwegian fatalities series, we find

$$\hat{y}_t = a_t - 0.019860\, x_t,$$

for t = 1, …, n, with variances $\sigma_\varepsilon^2 = 0.00367357$ and $\sigma_\xi^2 = 0.0035908$. The values of $\hat{y}_t$ are plotted at the top of Figure 1.2.4, while the values of the residuals ε_t obtained with model (1.2.5) are graphed at the bottom of Figure 1.2.4.

The first eight autocorrelations of the residuals in Figure 1.2.4 are shown in the correlogram in Figure 1.2.5 (see again Section 3.2.1.2 for the exact definition of the correlogram). None of these autocorrelations exceed the 95% limits of ±0.343. In contrast with traditional linear regression, this indicates that the residuals of the state space analysis are independent of one another, and that the value of the t-test can now therefore be trusted.

In this case, the standard t-test for establishing whether the regression coefficient b = -0.019860 deviates from zero yields

$$t = \frac{-0.019860}{0.0106358} = -1.87.$$

Since the value of the latter t-test is associated with a p-value of 0.071, the relation between the Norwegian fatalities and time is no longer significant at the conventional 5% level. Moreover, since the values of the regression coefficient obtained with traditional linear regression and with state space analysis are virtually identical, the large difference between the values of the two t-tests can be almost completely attributed to the large differences in their standard errors: 0.0017356 for traditional regression versus 0.0106358 for time series analysis.

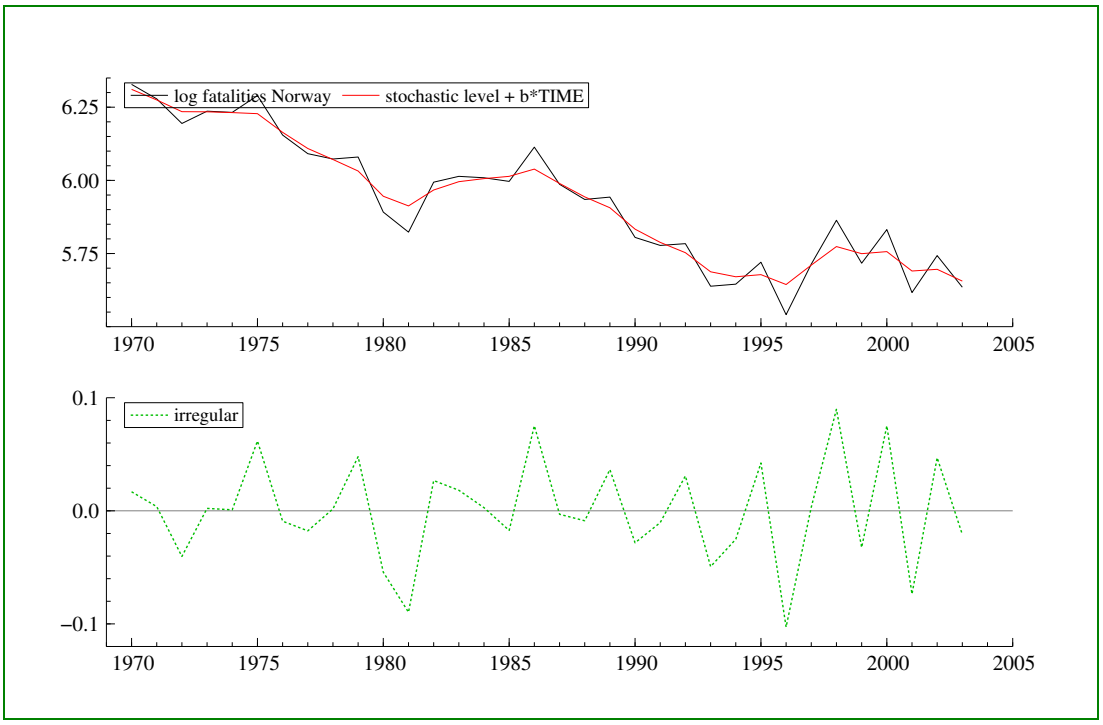


Figure 1.2.4: Correlogram of residuals of traditional linear regression of the log of the Norwegian fatalities on time.

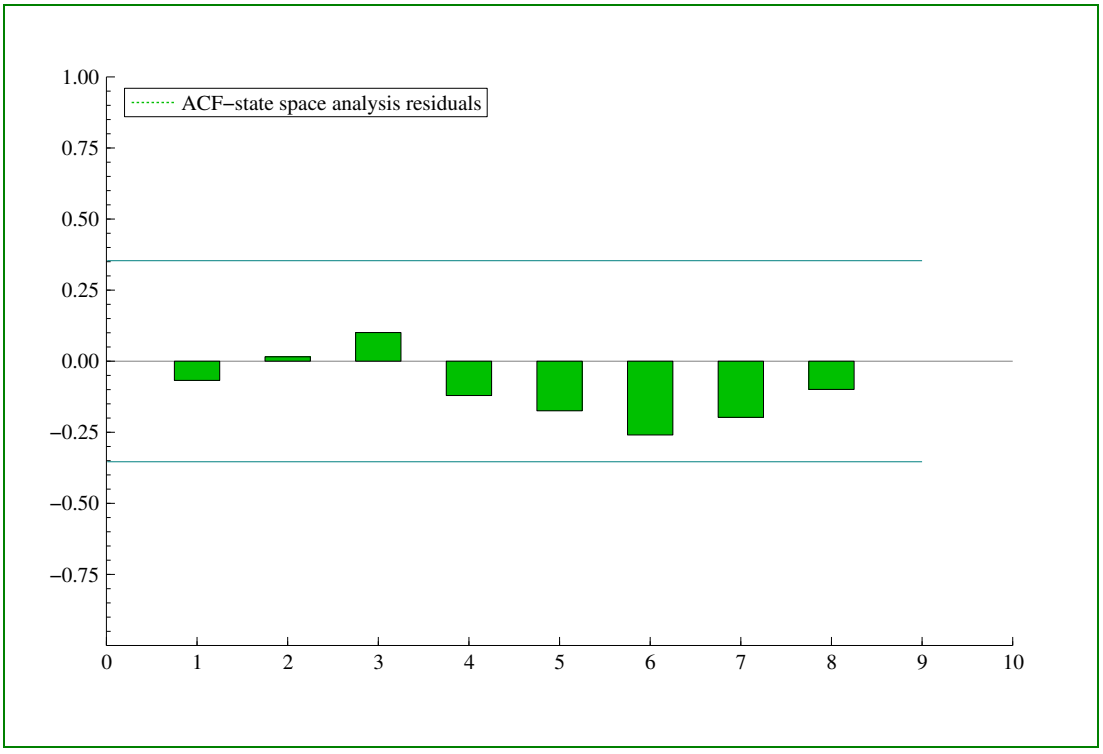


Figure 1.2.5: Correlogram of the residuals of state space analysis of the log of the Norwegian fatalities.

See Durbin and Koopman (2001, Section 6.2.4) for details on how to calculate the denominator of the *t*-statistic.

Generally, time series analysis can serve three purposes. First, time series analysis can be used to obtain an adequate *description* of the time series at hand, as we have illustrated for the log of the Norwegian fatalities series. Second, explanatory variables other than time can be added to the model in order to obtain *explanations* for the development in the time series at hand. In SafetyNet, these explanatory variables are national exposure data (as collected in WP2), national safety performance indicators (as collected in WP3), and national road traffic safety measures. A third important application of time series analysis is the ability to *predict* or *forecast* further developments of a series into the (unknown) future. In traffic safety research, such forecasts can be used to assess whether future national safety targets are likely to be met, for example.

Summarising, when dealing with observations made sequentially through time, statistical tests based on standard techniques like traditional linear regression easily result in overoptimistic or even plain incorrect conclusions, due to the fact that the residuals obtained with these techniques usually do not satisfy the model assumptions. This is true irrespective of whether the interest lies in descriptive analysis, in explanatory analysis, or in forecasting.

Dedicated time series analysis techniques, on the other hand, explicitly take the time dependencies between the observations into account, thus greatly improving the chances of obtaining residuals that do satisfy the model assumptions, and allowing to reliably test whether the estimated relationships between dependent and independent variables in the analysis are statistically meaningful or not. This is not only true for the state space methods illustrated in the present section, but also applies to other dedicated time series techniques like ARIMA (see Section 3.4) and DRAG models (Section 3.5).

Since many data collected in the SafetyNet project consist of observations made sequentially through time, it is essential that the relations between developments in accident data, exposure data, and safety performance indicators in the EU are investigated with dedicated time series analysis techniques.

In the report the following data sets are used to illustrate the results of their analysis with time series models:

- the monthly number of Austrian fatal accidents from January 1987 through December 2004 (Section 3.2.1);
- the monthly number of people killed and seriously injured in road traffic in Greece from January 1998 through December 2003, excluding the cities of Athens and Thessalonica, together with the monthly number of breath alcohol controls and the monthly number of vehicles in circulation for the same period of time (Section 3.2.2);
- the annual number of fatalities, vehicles and population from 1970 through 2002 for seventeen member states of the European Union (Section 3.2.3);
- the annual number of Norwegian road traffic fatalities for the years 1970 through 2003 (Sections 1.2.2, 3.4.3, 3.6.1, 3.6.6, and 3.7);

- the monthly number of French road traffic fatalities from January 1975 through December 2001, together with gasoline and diesel sales, car fuel price, a few weather variables, and three intervention variables: two for presidential amnesties concerning fines, and one for the so-called Cellier incident (a young woman killed by a drunk driver resulting in a lot of media attention); the monthly number of injury accidents and fatalities on French A-level roads and motorways in the same time period are also considered (Section 3.4.5);
- the annual number of Finnish road traffic fatalities for the years 1970 through 2003 (Sections 3.6.2 and 3.6.6);

the monthly number of drivers killed and seriously injured in the UK from January 1969 through December 1984, together with the monthly price of petrol in the UK and the monthly number of vehicle kilometers driven by cars for the same period of time, and an intervention variable for the introduction of the seatbelt law in February 1983 in the UK (Sections 3.4.4, 3.6.3, 3.6.4, 3.6.5, 3.6.6 and 3.7).

1.3 Overview

In this introductory chapter, it has been demonstrated that in traffic-safety research data often form hierarchies (nested data) or time series. It has been demonstrated how the analysis of such complex designs with traditional techniques can lead to erroneous conclusions and two families of analysis techniques were presented that are able to properly represent the dependencies in these complex data structures.

As mentioned above, the independence of the errors is not the only assumption that traditional regression analyses are based on. It is often stated however, that it is the most important one in terms of potential consequences of its violation. Note that this can only be regarded as a very general rule of thumb. Violations of the other assumptions (see above for a general introduction and 3.2.1.2 for details) may also lead to serious, sometimes even more serious consequences. Nevertheless, examples of the dire consequences of ignoring dependence are sufficiently frequent to make it a reasonable rule of thumb.

It is also important to note that the potential (combinations of) violations of the assumptions are abundant, any combination of dependency and distribution may occur. General classes of violations can be treated by available statistical techniques (in the software packages). With respect to the Gaussian assumption, the generalised linear models approach (McCullagh & Nelder, 1989), which is commonly available in statistical software, allows to treat a class of Non-Gaussian distributions, that includes the Poisson and negative binomial distribution among others, but it does not cover all potential distributions (see 2.3.1 and 3.2). Extensions to hierarchical models of the generalised linear models are available and are discussed in section 2.3. Extensions to time series models are currently under development. Although in practice, one might sometimes have to develop a completely new approach, the most important now implemented approaches are discussed in this best practice advice.

Next to variables that do not follow the Gaussian distribution one also often encounters problems that involve multiple dependent variables. There are a multitude of techniques dedicated to this type of data, which to describe is clearly beyond the scope of the present document that is focused on the treatment of dependency. Multivariate methods are addressed only to the extent that they are straight forward extensions of the multilevel models presented (2.4, 2.5, and 2.6).

To conclude, this best practice advice is focussed on the treatment of dependency in complex data structures and therefore introduces multilevel models for the analysis of nested data and time series analysis. These guidelines encompass two deliverables. The present document, Deliverable 7.4, gives theoretical back-ground and details for the two families of analyses sketched in this introduction. Deliverable 7.5, the manual, is developed in parallel with the present document. It contains practical guidelines for the conduction of the analyses that are introduced in this deliverable. It gives an

overview of the software available at the time of writing as well as examples of their actual implementation in exemplary chosen software.

The present document is organized in two main chapters, focussing on multilevel modelling and time series analysis respectively. This separation is based on the general difference between the data structure for multilevel modelling and time series analyses. Typically, multilevel analyses are applied to data from many units of measurements (e.g. drivers, cars, counties, etc.) with one or relatively few observations per unit. In contrast, time series data are usually applied to data from one or relatively few units of measurement (regions, countries, etc) with many observations per unit repeated through time. Because of this general difference, the deliverable is structured in one chapter treating multilevel modelling (Chapter 2) and one treating time series analysis (Chapter 3). For researchers who want to analyse hierarchically structured data it should be enough to read the parts concerning multilevel analyses, while researchers interested in the analysis of time series data can restrict themselves to reading the parts dedicated to time series analyses. Within the chapters, however, information does build up across sections.

In Chapter 2, the general principles of multilevel modelling are at first described in an intuitive way along the lines of a simplified example (Section 2.1). Subsequently, detailed descriptions are given for multilevel versions of analyses that are commonly used in traffic research. As presented in Figure 1.3.1, the sections are structured according to the type of dependent variable. In Section 2.2 the multilevel version of linear regression models for normally distributed data are presented. In Section 2.3, this special case is placed in the broader framework of the generalised linear model approach, which allows to model data resulting from different types of distributions (Section 2.3.1). Under this framework models for discrete data will be presented. More specifically, in Section 2.3.2 it will be described how binary responses can be modelled in multilevel logistic regression analyses. In Section 2.3.3 the analysis two types of models are presented for the analysis of multinomial responses: the ordered proportional odds analysis and an unordered multinomial model. In Section 2.3.4 it is demonstrated how count data can be modelled in multilevel Poisson regression analyses. Further it will be shown how multilevel modelling can be applied to analyse datasets containing repeated measurements in Section 2.4 and multivariate responses in Section 2.5. Finally, the application of the multilevel approach to structural equation models will be discussed in Section 2.6. In Section 2.7 modelling data structures that are not strictly hierarchical will be addressed. In particular modelling cross-classifications and/or multiple memberships will be addressed. In Section 2.8 recently developed estimation based on Bayesian modelling will be addressed. The chapter on multilevel modelling is closed with conclusions (Section 2.9) containing a summary of the methods presented and some general recommendations for the analysis of hierarchical data structures. The structure of Chapter 2 is presented in Figure 1.3.1

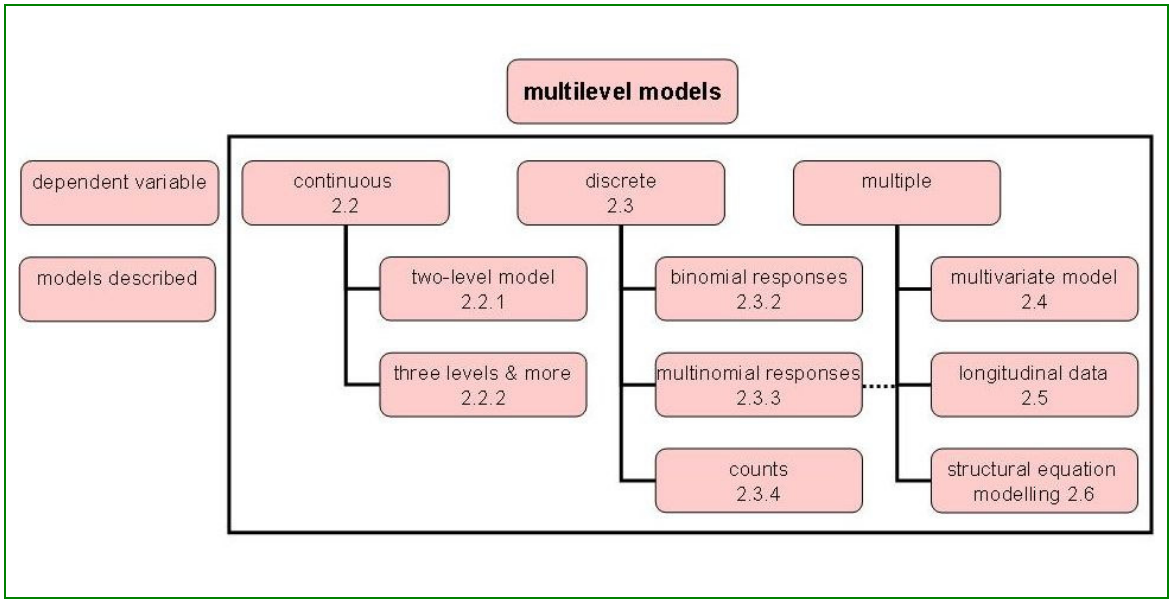


Figure 1.3.1: Structure of multilevel models presented in Chapter 2.

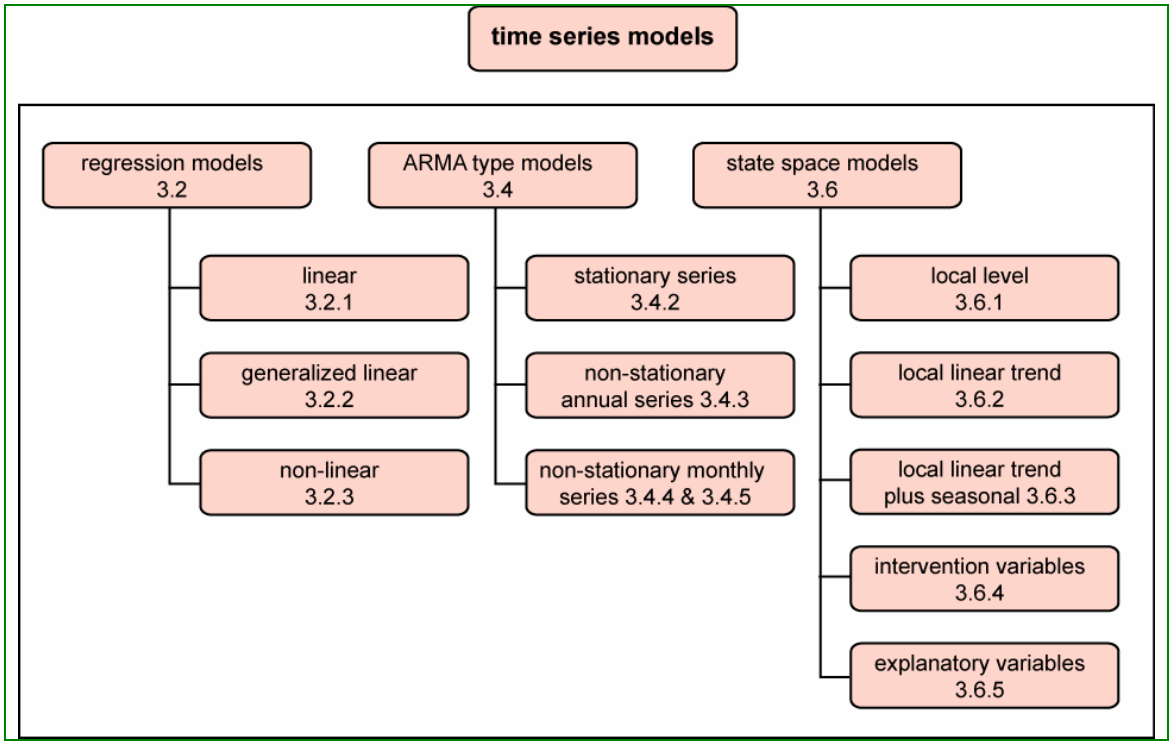


Figure 1.3.2: Structure of multilevel models presented in Chapter 3.

Chapter 3 begins with a short introduction to a few core issues in time series analysis (3.1). Section 3.2 describes traditional regression analyses models

(linear: 3.2.1, generalised linear model: 3.2.2 and non-linear models in 3.2.3) discussing possible violations of the assumption when dealing with time series data and the possibility to solve these problems by adding predictors variables such that the dependencies are removed. An introduction to dedicated time series analysis techniques and their application in road-safety research is presented in Section 3.3. In Sections 3.4 to 3.6 models dedicated to time series analyses are presented. The ARMA-type and DRAG approaches are discussed in Sections 3.4 and 3.5, while the state space methods are presented in Section 3.6. In Section 3.7, the equivalence of ARMA-type and state space models is demonstrated on the basis of a few examples. The chapter on time series is closed with conclusions (Section 3.8) containing a summary of the methods presented and some general recommendations for the analysis time series. The structure of Chapter 3 is presented in Figure 1.3.2.

Chapters 2 and 3 each present a number of analysis models for either nested data or time series that are relevant to traffic safety research. A standardized discussion format was adhered to when scrutinizing each model to maintain a certain consistency throughout this deliverable. Furthermore, theoretical considerations with respect to model building, testing and interpreting are explained by applying them to a real dataset. Therefore, special attention is given to each of the following aspects of a particular model:

- Objectives of the technique
- Model definition
- Model assumptions
- Research example + dataset
- Model fit and diagnostics
- Model interpretation

This standardized format should give the reader a good insight how a particular model is applied, for what sort of data it is suitable and how the results can be interpreted. For each chapter there will also be references for a more in-depth treatment of the method presented.

Throughout the remainder of the document readers are expected to master ordinary regression analysis. Given the different levels of complexity of the models described in the various chapters, the readers' need to depend on earlier acquired information or on extra background material will vary. For the later multilevel chapters it is good to be familiar with the corresponding single level models (more specifically, binomial model, Poisson model, structural equation modelling, etc.). Similarly, references for readers who are interested in the background and different versions of the ARMA-type models, state space models, or non-linear time series analysis will be supplied.

Chapter 2 - Multilevel Modelling

In the introductory chapter it has been shown that many research problems in the area of road-safety contain hierarchical data structures and how this challenges the use and interpretation of traditional analysis. In the following sections it will be demonstrated in detail how the problems sketched in the introduction can be solved by the application of multilevel models. To understand the structure of Chapter 2 the reader has to keep in mind that multilevel modelling is not one type of analysis. It does not even denote one class of analyses; rather it is a technique that has to be applied to traditional statistical analysis of different types. In the last decennium, the problem of hierarchical data structures for traditional analyses[7] has been widely recognised and as a consequence the multi-level approach has now been implemented in a wide range of techniques of analyses (Kreft and de Leeuw 2002). Structurally identical models are also know as mixed effects or random effects models (e.g., Bates & Pinheiro, 1995) or as random coefficient regression models (e.g., Longford, 1993)

In Section 2.1, the general principles of multilevel modelling are at first described in an intuitive way along the lines of a simplified example. Subsequently, detailed descriptions will be given for multilevel versions of analyses that are commonly used in traffic research. Sections 2.2 and 2.3 are dedicated to describing multilevel regression models in more detail. In Section 2.2 the multilevel version of linear regression models for normally distributed data are presented, while in Section 2.3, the analysis of discrete response data will be described. The introduction to this section (2.3.1) places the special case of linear models into the broader framework of the General Linear Model approach, which allows to model data resulting from different types of distributions. Under this framework models for discrete data will be presented. Specifically, in Section 2.3.2 it will be described how binary and binomial data can be modelled in multilevel logistic regression analyses, in Section 2.3.3 how multilevel models can be used to model multinomial responses in either ordered or unordered category models and in Section 2.3.4 it is demonstrated how count data can be modelled in multilevel Poisson regression analyses.

Hierarchical data structures can also arise due to the structure of the variables that are analysed. A dataset with multiple dependent variables has several measurements that are nested under one person. These data structures can therefore be modelled with multilevel models. We will show how multilevel modelling can be applied to the analysis of datasets containing longitudinal data and other types of repeated measurements in Section 2.4 and to the analysis of multivariate responses in Section 2.5. In both cases, just like in the case of the multinomial responses, multilevel modelling is used to represent the structure of the responses themselves and not (at least not in the first place) that of a hierarchical structure from which the data are collected. Finally, in Section 2.6 a multilevel version of structural equation models will be presented. In Section 2.7 modelling data structures that are not strictly hierarchical because they contain

[7] Here and in the following, the term “traditional analyses” denotes analyses in which the random part is not structured – neither hierarchically nor in time.

cross-classifications and/or multiple memberships will be addressed and in Section 2.8 recently developed estimation methods, in particular Bayesian estimation methods are presented. This document on multilevel modelling will be closed with conclusions (Section 2.9) containing a summary of the methods presented and some general recommendations for the analysis of hierarchical data structures.

In each section, a standardised discussion format was adhered to, to discuss each model (objectives of the technique, model definition, model assumptions, introduction of a research example and dataset, model fit and diagnostics, model interpretation).

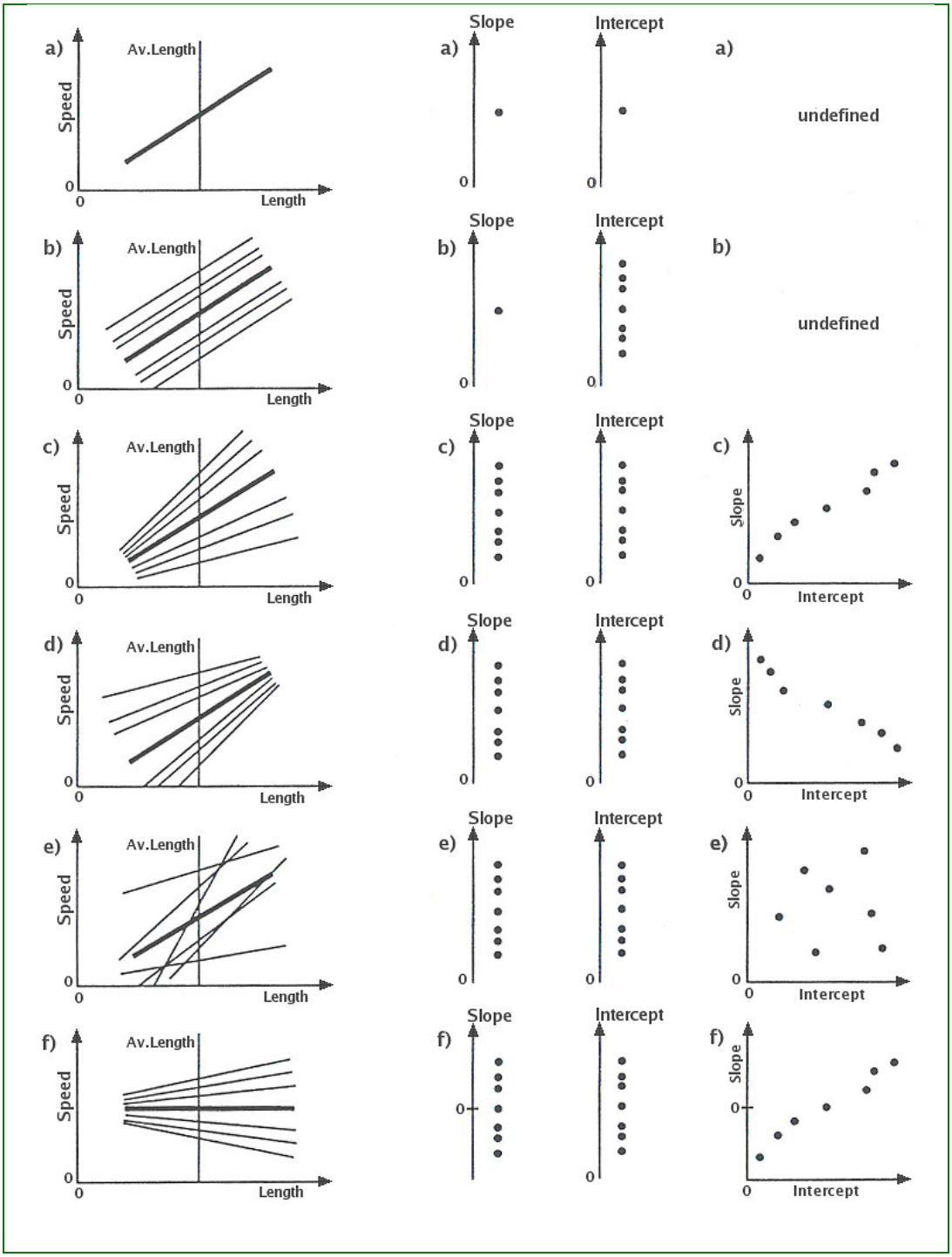
2.1 An intuitive introduction to multilevel modelling[8]

(Ward Vanlaar, IBSR)

To appreciate the basic concepts of the multilevel approach, we first work with a two-level model with drivers at level 1 nested in road sites at level 2 and two variables measured on a continuous scale. The example in this section is an artificial example as an illustration for teaching purposes. Each driver's speed is measured along with some other variables when passing by the road site. The dependent variable in this artificial example is speed, measured in km/h and the independent variable is length of the car, measured in metres and centred about its mean. The underlying hypothesis is that longer vehicles will correlate with higher speeds because a longer vehicle has a more powerful engine. Note that this hypothesis is rather naively formulated for the sake of clarity in this artificial example and that it does not necessarily bear real social relevance.

In a multilevel model distributional assumptions are made at each level of the model, in this case at level 1 – drivers – and at level 2 – road sites. The distributional assumptions at the lower level are assumptions about the variation between drivers; this is comparable to the distributional assumptions in the traditional regression model. The distributional assumptions at the higher level are assumptions about the variation between road sites. Road sites too are now allowed to vary and this variation is summarized in a distribution. For example, road sites can have different intercepts and slopes and they are assumed to be normally distributed around the overall intercept and slope. These distributions at higher levels are called higher-level distributions. Figures 2.1.1a - f (after Jones, 1993) give a range of possible models and the higher-level distributions for the corresponding slope and intercept. These higher-level distributions are the result of the existence of several intercepts and slopes at level 2, corresponding to road sites. Put another way, instead of one regression line with one intercept and slope, there are several regression lines, one per road site, each with their corresponding intercept and slope. The slopes measure the increase in speed associated with a unit increase in length for each road site. Since the vertical axis in these graphs is centred at the mean of length, the intercepts correspond to the speed of a car of average length per road site. In figure 1a the speed/length relation is shown as a straight line with a positive slope; longer cars drive faster. In this graph no account is taken of context; place – i.e. road site – does not matter for the speed of drivers and the relationship is conceived only in terms of individual characteristics. This is remedied in 1b with each of the different road sites (seven in this figure) having its own speed/length relation represented by a separate line. These parallel lines imply that, while the speed/length relation on each road site is the same, some road sites have uniformly higher speeds than others, which is easily explained by the existence of different speed limits. The lowest line could for example represent a road site with a speed limit of 30km/h, while the upper line could represent a road site with a speed limit of 120km/h.

[8] In this section the same format appears as in Jones (1993). Dr. Kelvin Jones kindly gave his permission to use this format.



<u>Figures 2.1.1a to f:</u> Higher level distributions for road sites' intercepts and slopes – regression of speed against car length depending on road sites (graphs on left hand side); dot plot for the distribution of the slopes and intercepts separately, with the variable length centred about its mean (centre); scatter plot of the joint intercepts and slopes distributions, with the variable length centred about its mean (right hand side). Adapted from Jones, 1993, p. 251

The situation becomes more complicated in 2.1.1c to 2.1.1f as the steepness of the lines varies from road site to road site, i.e. each line, representing a road site, has a different slope, while in 2.1.1b only the intercepts of the lines differed. In 2.1.1c the pattern is such that road site makes very little difference for small cars, but road sites have very different speeds for longer cars. An explanation could be that the maximum speed of small cars is so low that they can only reach the lowest speed limit of 30km/h, e.g., if the car fleet of a town would be composed exclusively of small electronic cars, while long powerful cars can easily reach higher speeds leading to a more diverse speed pattern depending on the different existing speed limits at road sites. In contrast, figure 2.1.1d shows relatively large road site-specific differentials for small cars. A possible explanation could perhaps be found in the attitude of drivers of powerful cars: those drivers tend to speed regardless of the speed limit and therefore their speed distribution over different locations has a very small range, while drivers of smaller cars are more conscientious and tend to respect the speed limits resulting in a broad range of speeds. Note again that these possible explanations are only given for didactical reasons; they don't necessarily reflect a relevant or true idea.

The next graph, 2.1.1e, with its criss-crossing, represents a complex interaction between length and road site. Steep lines, indicating strong relationships between the dependent and independent variable, can both be seen at road sites with a high speed limit and with a low speed limit. At some road sites small cars have relatively high speeds, in others long cars have. An explanation could probably be found in other road site-specific characteristics besides the speed limit. Finally, plot 2.1.1f shows that small cars drive with the same speed, regardless of the road site, while the speed of powerful long cars differs according to the road site. This pattern is similar to 2.1.1c, but this time this difference is achieved by some road sites having a high speed for long cars, while at other road sites long cars drive at a lower speed than small cars. An explanation could be the architecture of the roads in combination with the attitude of car owners. Car owners of long powerful – and thus exclusive and expensive cars – will treat their car with a lot of care. Such drivers will take speed bumps in a low speed regime very prudently and therefore perhaps even drive slower than the maximum limit. Car owners of small cars could be less considerate about their car and thus take speed bumps at a more appropriate speed.

"The differing patterns of Figure [2.1.1] are achieved by varying the slopes and intercepts of the lines. […] The key feature of multilevel models is that they specify the potentially different intercepts and slopes for each road site as coming from a distribution at a high level" (Jones, 1993, p. 250). Figure 2.1.1 also shows the higher-level distributions for the slope and intercept that correspond to the different graphs. A separate dot plot for the distributions of the slopes and intercepts and a scatter plot of the joint distribution can be found in the centre part and the part at the right hand side of Figure 2.1.1. These distributions concern road sites, not individuals, and result from treating road sites as a sample drawn from a population of road sites. "It can be seen that:

Figure 1a is the result of a single non-zero intercept and slope; Figure 2.1.1b has a set of intercepts, but a single slope; Figures 2.1.1c-2.1.1f have sets of intercepts and slopes" (Jones, 1993, p. 251).

"The different forms of Figures [2.1.1]c to f are a result of how the intercepts and slopes are associated" (Jones, 1993, p. 252). In Figure 2.1.1c the speed/length relation is strongest at road sites where the average speed is high (as indicated by a greater intercept); a steep slope is therefore associated with a high intercept, meaning there is positive association between the intercepts and slopes, as shown on the right hand side of the figure. In contrast, in Figure 2.1.1d road sites where the average speed is high have a weak speed/length relationship: a high intercept is associated with a shallow slope. Consequently, there is a negative association between the slopes and the intercepts. "The complex criss-crossing of Figure 2.1.1e is the result of the lack of pattern between the intercepts and slopes" (Jones, 1993, p. 252) shown in the graph at the right hand side of Figure 2.1.1e. The average speed at a particular road site contains no information about the marginal increase in speed with length of cars at that road site. The distinctive feature of the final plot in Figure 2.1.1f, results from the slopes varying about zero so that at the "typical" road site there is no relation between speed and length; at some road sites the slope is positive and at others it is negative.

2.2 Multilevel linear regression models

In this section, graphs will be turned into equations shifting from an intuitive approach to a more formal, mathematical approach. For the ease of understanding, multilevel models will be presented for linear models.

2.2.1 Basic two level random intercept and random slope models[9]

Ward Vanlaar (IBSR)

The basic principles of 2-level models will be illustrated on the basis of the assessment of the relationship between the length of cars and their speed described in the introduction.

2.2.1.1. Objectives of the technique

The objectives of this technique correspond to the objectives of ordinary regression analysis, but in addition to that, there is also the objective of taking contextual information into account by letting the intercept and slope vary across road sites. According to Tacq (1997), the four objectives of traditional linear regression analysis are:

- To look for a function, which represents the linear association between the independent variables and the dependent variable better than any other function. This comes down to calculating a regression coefficient for each independent variable.
- To examine the strength of the relationship and to know which share of the variance of the dependent variable is explained by the variances of the independent variables together. This comes down to the calculation of the multiple correlation coefficient R and its square. While the concept of explained variance is well-known in traditional regression analysis, it is problematic in multilevel models according to Snijders and Bosker (1999).
- To investigate whether the associations found in the sample can be generalized to the population. This corresponds to performing significance tests.
- To examine which independent variable is most important in the explanation of the dependent variable, corresponding to calculation of the beta weights.

2.2.1.2. Model definition

2.2.1.2.1. The random intercept model

According to Jones (1993, P. 252) all statistical equations have in essence the same underlying structure, which can be expressed verbally as:

[9] In this section the same format appears as in Jones (1993). Dr. Kelvin Jones kindly gave his permission to use this format.

RESPONSE = SYSTEMATIC COMPONENT + FLUCTUATIONS

Or

RESPONSE = FIXED PARAMETERS + RANDOM PARAMETERS

In the case of a single-level bivariate model, i.e. the usual simple regression model (cf. figure 1a), the general verbal equation becomes:

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i \qquad (2.2.1)$$

where

- subscript i signifies an individual respondent;
- y and x measure the dependent and independent variables, namely the speed and length of a car;
- β_0 and β_1 are fixed and unchanging parameters, namely the intercept and the slope; the former, when x is centred about its mean, represents the speed of a car of average length; the latter is the change in speed for an increase in length with one metre;
- e signifies the random part which allows for fluctuations around the fixed part, where the term random simply means “allowed to vary”.

This equation is specified only at the micro-level of the individual. To build a multilevel model the *micro-model* has to be re-specified by distinguishing road sites with the subscript j. For the random intercept model (cf. figure 2.1.1b) this yields:

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + e_{ij} \qquad (2.2.2a)$$

There is one *macro-model* at the road site level:

$$\beta_{0j} = \beta_0 + u_{0j} \qquad (2.2.2b)$$

This macro-model allows for the differential road site intercept (β_{0j}) to vary from road site to road site around the overall intercept (β_0) by adding the random term u_{0j}.

The micro model is seen as a within-road site equation, while the macro model is a between-road site equation in which the parameter of the within model is the response (Jones, 1993). Both equations are combined to form the random two-level model:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \left(u_{0j} + e_{ij}\right) \qquad (2.2.2c)$$

All the elaborations have come in the random part, because in addition to allowing individual cars to vary, road sites have been allowed to vary in having a differential speed for a car of average length. Such models in which the intercept is the only term allowed to vary at level two are commonly referred to as "variance components models" (Rasbash, Steele, Browne, & Prosser., 2004).

2.2.1.2.2. The random intercept/random slope model

The formulas look as follows if the slope is also allowed to vary from road site to road site in addition to a random intercept (cf. figures 2.1.1c-f). The micro model:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij} \qquad (2.2.3a)$$

and the two macro-models at the road site level:

$$\beta_{0j} = \beta_0 + u_{0j} \qquad (2.2.3b)$$

$$\beta_{1j} = \beta_1 + u_{1j} \qquad (2.2.3c)$$

These two macro-models allow respectively for the differential road site intercept (β_{0j}) to vary from road site to road site around the overall intercept (β_0) by adding the random component u_{0j} and for the differential slope (β_{1j}) to vary around the overall slope (β_1) by adding the random component u_{1j} (Jones, 1993).

Again, the micro model is seen as a within-road site equation, while the macro models are two between-road site equations in which the parameters of the within model are the responses. Note that this is easy to see when using the notation with e_{ij} as part of the micro model as opposed to the macro model because then only the micro-model contains both subscripts i and j, referring to a within situation, while the macro-models then only contain subscript j, referring to a between situation. All three equations are combined to form the fully random two-level model:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \left(u_{1j} x_{1ij} + u_{0j} + e_{ij}\right) \qquad (2.2.3d)$$

All the elaborations have come in the random part, because in addition to allowing individual cars to vary, road sites are also allowed to vary in having a differential speed for a car of average length, and a differential speed/length relationship (Jones, 1993).

As with any other statistical distribution, and making the usual assumptions of normality, homogeneity and independence, these higher-level distributions can

be summarized by measures of the centre, the mean, and spread around the centre, the variance. Relations between the slope and intercept distributions can be summarized by a measure of covariance. "Thus, the higher-level distributions can be summarized in terms of the fixed part (the means β_0 and β_1) and the random part (the variances $\sigma^2_{u_0}$ and $\sigma^2_{u_1}$, and the covariance $\sigma_{u_0u_1}$)" (Jones, 1993, p. 253).

Table 2.2.1 (after Jones, 1993) summarizes Figure 2.1.1 in terms of these parameters. Estimates of these terms effectively summarize the extent to which places differ. The various combinations of substantial and close-to-zero estimates for the variance/covariance tell us in a quantitative manner the way in which context matters. The case of Figure 1f is interesting in this regard, because it suggests that the usual single-level model would find that across the sample there is no relation between speed and length, but the multilevel model would reveal differing relationships at different road sites. If all the variance terms of the higher-level distributions are effectively zero, there is no contextuality and thus there is no need for macro models. These variations in speed are adequately described in terms of a micro model based solely on individual attributes (cf. Figure 2.1.1a).

	Intercepts		*Slope*		*Intercept/slope*
	Mean	*Variance*	*Mean*	*Variance*	*Covariance*
Graph	β_0	$\sigma^2_{u_0}$	β_1	$\sigma^2_{u_1}$	$\sigma_{U_0U_1}$
A	*+*	*0*	*+*	*0*	*/*
B	*+*	*+*	*+*	*0*	*/*
C	*+*	*+*	*+*	*+*	*+*
D	*+*	*+*	*+*	*+*	*-*
E	*+*	*+*	*+*	*+*	*0*
F	*+*	*+*	*0*	*+*	*+*

Table 2.2.1: Figure 2.1.1 represented as parameters for two higher-level distributions (where + is positive, different from zero and where – is negative, different from zero)

2.2.1.3. Heteroscedasticity

Multilevel models share with many traditional models the assumption that the residuals at each level are homoscedastic, i.e., have constant variance and covariances, and do not depend on the particular values of the explanatory variable(s) included in the model. This assumption is partially relaxed, however, once random slopes are specified in the model: Variances at one or both levels

are assumed to depend linearly or quadratically on one or more of the explanatory variable(s)[10].

The following reasoning, borrowed from Snijders and Bosker (1999), and applied to our speed example, illustrates this feature of multilevel models. In case of a fanning-in pattern (see figure 2.1.1d) a random slope for the effect of car length on speed would indicate that road sites affect the speed of small cars to a larger extent than the speed of large cars. This can be seen in figure 2.1.1d as the lines representing the different road sites are farther away from one another at the lower values on the X-axis and closer to one another at the higher values on the X-axis. So at lower values on the X-axis (i.e. for small cars) there is much more variation in speed between road sites, compared to higher values on the X-axis (i.e. for large cars). In other words, if you drive a small car, which road site you are at will matter a lot and may influence your speed considerably, while road site does not matter if you drive a large car. This means that road sites add a large component of variance to the speed of small cars, but little or nothing to the speed of large cars. Therefore, the intra-class correlation for small cars (also known as the Variance Partition Coefficient (VPC)), defined as the proportion of the total residual variation that is due to differences between groups (Goldstein, 2003), will be higher than the intra-class correlation for large cars. This implies that, once random slopes are specified in a model, the intra-class correlation or VPC cannot be uniquely defined any longer because this residual variation (due to differences between groups; road sites in our case) will vary as a function of the explanatory variable's values (small or large cars in this example).

2.2.1.4. Model assumptions

"As all statistical models, the hierarchical linear model is based on a number of assumptions. If these assumptions are not satisfied, the procedures for estimating and testing coefficients can be invalid. […] It is advisable, when analysing multilevel data, to devote some energy to checks of the assumptions. (Snijders & Bosker, 1999, p. 120)" Before investigating checks of the assumptions in the next section, the assumptions themselves are listed below (Snijders & Bosker, 1999; Rasbash et al., 2004):

$e_{0ij} \sim N(0, \sigma^2_{e_0})$, the level-one residuals are assumed to be Normally distributed, with mean zero and constant variance $\sigma^2_{e_0}$;

$u_{0j} \sim N(0, \sigma^2_{u_0})$ and $u_{1j} \sim N(0, \sigma^2_{u_1})$, the level-two random coefficients are assumed to follow a multivariate Normal distribution with mean zero and constant variance respectively $\sigma^2_{u_0}$ and $\sigma^2_{u_1}$;

[10] The reader is referred to Section 2.5 for a similar discussion of heteroscedasticity linked to the introduction of random slopes in the model, and for a mathematical description of the implication of random slopes in the definition of the observations' variance and covariances.

Random coefficients at level 1 (e_{ij}) and at level 2 ($\sigma^2_{u_0}$, $\sigma^2_{u_1}$) are assumed to be uncorrelated;

$y_{ij} = N(XB, \Omega)$, the response variable is assumed to be Normally distributed, where XB is the fixed part of the model and Ω represents the variances and covariances of the random terms over all the levels of the data.

2.2.1.5. Research problem

As explained in the previous section the basic two level model will be explained using an artificial example about the influence of length of a car on the speed of that car. The underlying hypothesis, formulated for teaching purposes only, is that longer vehicles will correlate with higher speed as a longer vehicle has a more powerful engine.

2.2.1.6. Dataset

The dataset used consists of a sample of n=4994 drivers (of cars and motorbikes) passing by m=131 road sites out of a real dataset, which was collected in Belgium for epidemiological purposes. Each driver's speed is measured as a continuous variable in km/h along with some other variables when passing by the road site, the most important being the independent continuous variable length of the car, measured in metres and centred about its mean.

2.2.1.7. Model fit and diagnostics

2.2.1.7.1. The variance partition coefficient (VPC)

The VPC is the proportion of the total residual variation that is due to differences between groups (Goldstein, 2003), more precisely between road sites in our example. It is also referred to as the intra-class correlation (Snijders & Bosker, 1999), which measures the extent to which the y-values of individuals in the same group resemble each other as compared to those from individuals in different groups[11]. However, the former interpretation is the more usual one (Rasbash, 2004). The VPC is denoted by:

$$\frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_{e_0}} \qquad (2.2.4)$$

In our example the VPC for the random intercept model with length as explanatory variable is 0.749, meaning that almost 75% of the variation is due to differences between road sites. This is a strong indication that clustering

[11] As it has been noted earlier, the VPC cannot be uniquely defined once random slopes are included in the model. Snijders & Bosker (1999) propose alternative solutions to partition the observations' variance between the different level of analysis for models including random slopes.

effects are not to be disregarded in this dataset and that a multilevel approach is preferable.

2.2.1.7.2. Deviance test

"The deviance test, or likelihood ratio test, is a quite general principle for statistical testing. […] The general principle is as follows. When parameters of a statistical model are estimated by the maximum likelihood (ML) method the estimation also provides the likelihood, which can be transformed into the deviance defined as minus twice the natural logarithm of the likelihood. This deviance can be regarded as a measure of lack of fit between model and data, but (in most statistical models) one cannot interpret the values of deviance directly, but only differences in deviance values for several models fitted to the same data." (Snijders & Bosker, p. 88).

The deviance can thus be used to make an overall comparison of a more complex model with a less complex one, e.g., for the comparison of the model containing only the constant term with the model with length as an explanatory variable. The difference between minus twice the natural logarithm of the likelihood (-2xloglikelihood, see Tables 2.2.2. to 2.2.4) of both models follows a chi-square distribution with the number of degrees of freedom equal to the difference in the number of parameters being estimated in both models. This chi-square value can be tested against the null hypothesis that the extra parameters have population values of zero (Rasbash et al., 2001).

First, the simplest model of all is fitted, i.e. the model in which the intercept is specified as random at level 2, and in which no explanatory variables are included. For obvious reasons, such a model is referred to as the "null" or "empty" model. The value of the deviance for this null model is 45262.130 (cf. Table 2.2.2). Then, this empty model is extended by adding a fixed slope, representing the effect of car length on speed. The deviance obtained in this case corresponds to 45192.320. Both models can now be compared by performing the deviance test. Subtracting the deviance value of the variance component model with a fixed slope for car length (the "more complex model") from the deviance value of the empty model (the "less complex model) yields a value of 69.81. One extra parameter is estimated in the more complex model. Therefore the associated degree of freedom is 1. Testing this value as a chi-square value of 69.81 with 1 degree of freedom against the null hypothesis shows that this decrease is highly significant (p=0.000), indicating that the more complex model is the better model. Put another way, the deviance decreased after having elaborated the model, meaning the model fit improved.

The same conclusion can be drawn when shifting from the random intercept model to the full random model. The decrease corresponds now to 290.82 (45192.32 minus 44901.50) with 2 degrees of freedom (two additional parameters have been estimated, namely, $\sigma^2_{u_1}$ and $\sigma^2_{u_0}$) . This yields a p-value of 0.000 and is thus highly significant.

2.2.1.7.3. Residuals

Estimated residuals at any level can be used to check model assumptions (Rasbash et al., 2004). The residuals at each level are assumed to follow Normal distributions (see Section 2.2.1.4). At level 2, these residuals are interpreted as group effects, i.e. road site effects, while at level 1, residuals are in general interpreted as the individual error terms.

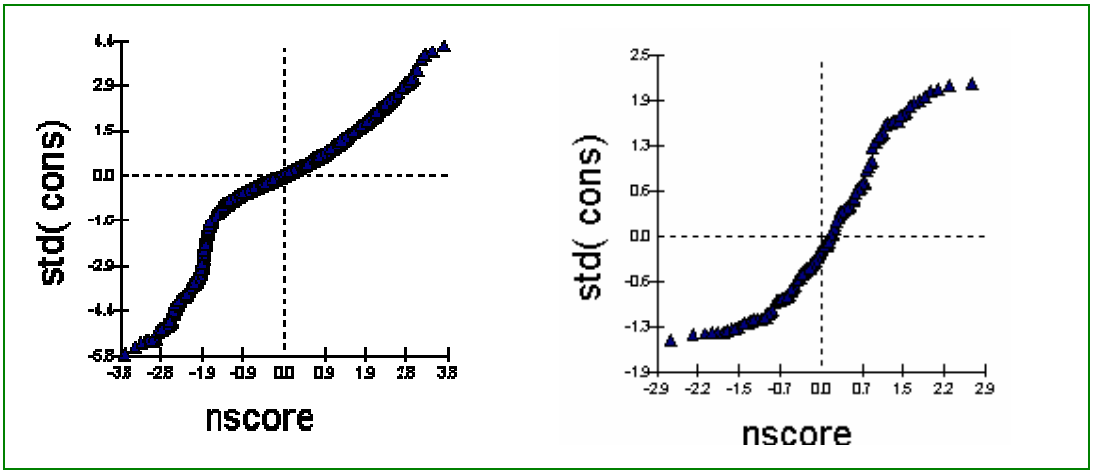


Figure 2.2.1: Normal probability plot of residuals for the random intercept model with speed and length, centered about its mean, at level 1 (left side) and 2 (right side)

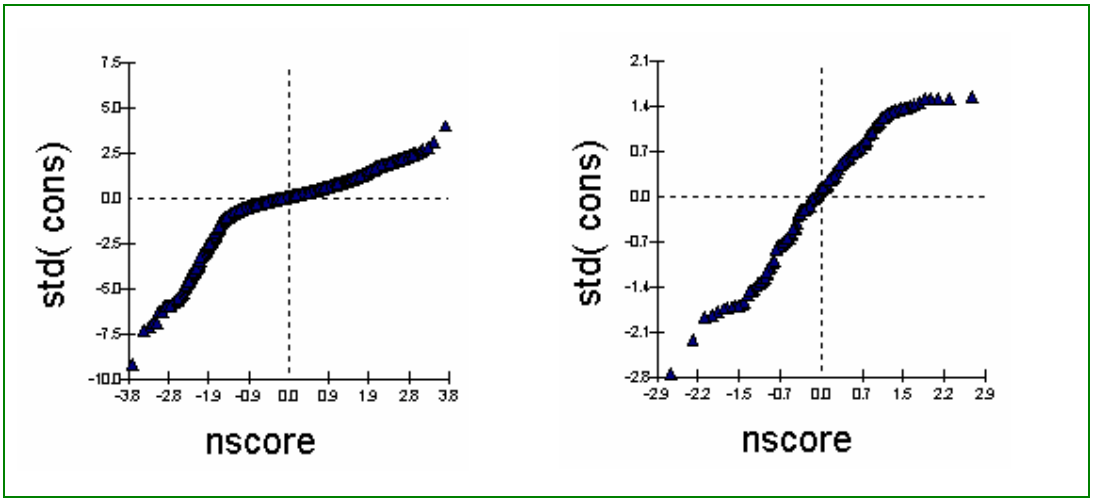


Figure 2.2.2.: Normal probability plot of residuals for the random intercept model with the natural logarithm of speed and length, centered about its mean, at level 1 (left side) and 2 (right side)

Parameter	Null model	Random intercept model	Full random model
	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Fixed			
Intercept	68.69 (3.27)	68.88 (3.24)	68.95 (3.24)

Length	/	2.30 (0.28)	1.69 (0.47)
Random			
Level 2			
$\sigma^2_{u_0}$ (intercept)	1358.94 (173.03)	1333.18 (169.37)	1334.85 (169.70)
$\sigma_{u_0u_1}$ (covariance)	/	/	-15.51 (17.42)
$\sigma^2_{u_1}$ (length)	/	/	12.82 (3.16)
Level 1			
$\sigma^2_{e_0}$	452.70 (9.18)	446.48 (9.05)	412.75 (8.46)
-2xloglikelihood	45262.13	45192.32	44901.50

Table 2.2.2 Estimates for the null, variance components, and full random models, with car length as a continuous explanatory variable

Clearly, the residuals in Figure 2.2.1 do not follow a normal distribution as their normal probability plot does not correspond to a straight diagonal, meaning those assumptions are violated. Therefore, care is warranted when estimating and testing the regression coefficients of the model. A solution could be to transform the dependent or independent variables, for example by calculating their natural logarithm. Figure 2.2.2. contains normal probability plots for the log transformed data. The situation at level 2 has improved as the level 2 residuals seem to follow the Normal distribution more closely after having transformed the data. However, the residuals at level 1 are still problematic. Model fit issues will be studied more extensively in the following chapters when elaborating on the different models.

2.2.1.8. Model interpretation

2.2.1.8.1. Random intercept model[12]

The coefficients of the random intercept model are interpreted as follows: (see Table 2.2.2) On average, over all road sites, the speed of a car with an average length is 68.88km/h. Obviously, there is a lot of variation over road sites, due to the different speed limits at road sites. This was revealed by the VPC.

For each increase of one length unit of a car, the speed of that car increases with 2.30km/h. Put another way, there is a positive relationship between length of a car and speed of that car.

The question now is whether this positive coefficient is significantly different from zero. The answer can be found by comparing the value of the coefficient with its standard error. In our case the standard error is 0.28. Clearly the

[12] Because they allow the calculation of the Variance Partition Coefficient, and thus the partitioning of the variance of the observations between the two levels, random intercepts model are also sometimes referred to as to “Variance Components Models”.

coefficient is significant as it is much greater than twice the value of its standard error.

2.2.1.8.2. Random intercept/random slope model

The main difference between the random intercept model and the full random model (i.e. the random intercept/random slope model) is the random slope, indicated by 2 extra parameters ($\sigma_{u_0u_1}$, $\sigma^2_{u_1}$) in the random part at level 2. A deviance test comparing the -2 loglikelihood value of the random intercept model to the one of the full random model clearly indicates that the inclusion of these two parameters significantly improved the model's fit ($\chi^2_2 = 255.37, p < .001$).

Different road sites can now have different slopes besides different intercepts. The variation between the different slopes is summarized by $\sigma^2_{u_1}$. There is a significant difference between the slopes of the different road sites since the value of the parameter (12.82) is greater than twice the value of its s.e. (3.16).

The average slope over all road sites is 1.69 (s.e.=0.47), meaning that a one unit increase of length of a car results in an average increase of speed with 1.69km/h.

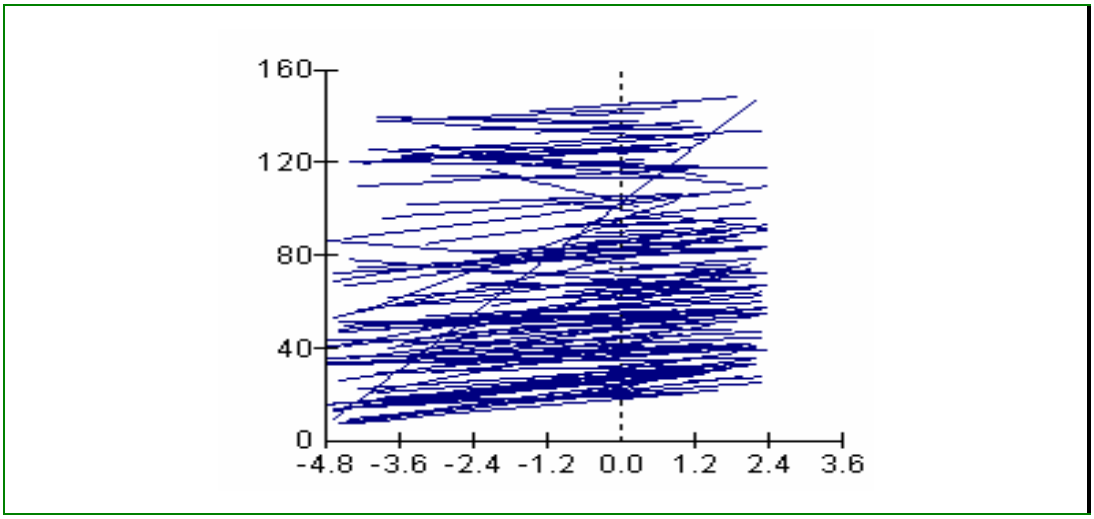


Figure 2.2.3: Regression lines of speed against car length (centred) for the various road sites

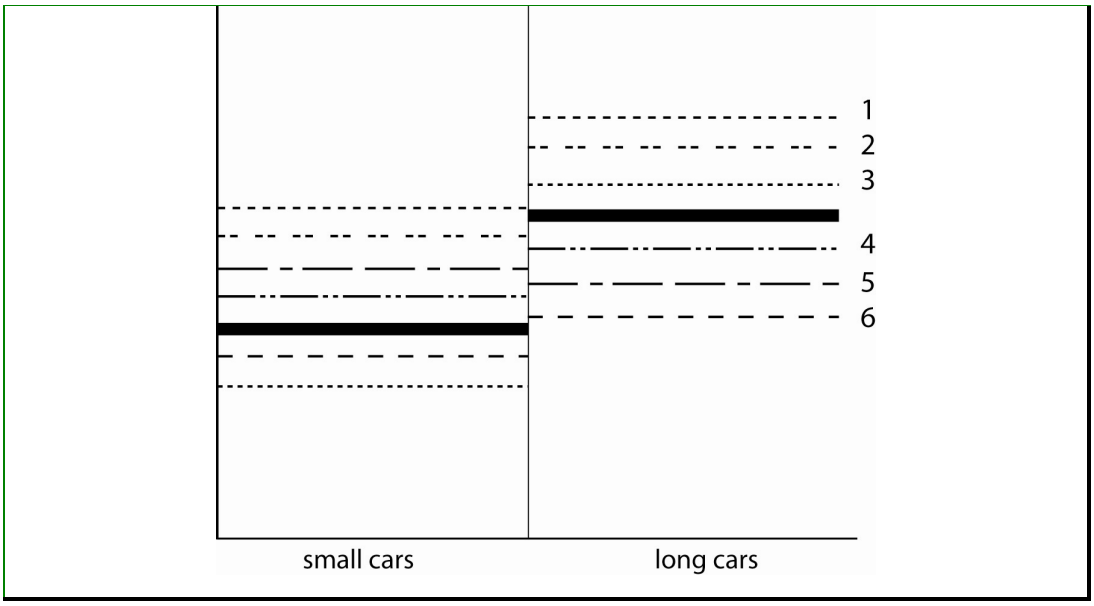


Figure 2.2.4: Small (<4.3m) and long cars' (>=4.3m) speed as a function of road sites

Note that the model also contains a value of the covariance between the random level 2 parameter for the intercept ($\sigma^2_{u_0}$) and length ($\sigma^2_{u_1}$). Its value equals -15.51 with a standard error of 17.42. Although this value clearly is not significant, its negative sign indicates a fanning in pattern (see figure 2.1.1d and Figure 2.2.3). In other words, a greater intercept corresponds to a smaller slope. The pattern is more easily discerned on figure 1d than on the graph based on our dataset. A possible explanation may be the attitude of drivers of powerful cars differs: those drivers tend to speed regardless of the speed limit and therefore their speed distribution over different locations has a very small range, while drivers of smaller cars are more conscientious and tend to respect the speed limits resulting in a broad range of speeds.

2.2.1.9. Extending the model

So far a bivariate two-level model with continuous variables on level 1 has been considered. Two important extensions of this model will now be discussed. First a model with a categorical explanatory variable will be studied. Second, higher level explanatory variables and contextual effects will be considered

2.2.1.9.1. Categorical explanatory variables

According to Jones (1993), level 1 categorical explanatory variables present no special problems and multilevel models can be specified in which some or all of the explanatory variables consist of categories. A random intercept/random slope model with an independent variable with two categories is achieved by specifying a micro-model with two dummy variables (having a value 0 or 1). In our example the continuous independent variable length could for example be divided in two categories: small cars and long cars. The micro-model looks as follows:

Parameter	Null model	Random intercept model	Full random model
	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Fixed			
Intercept	68.69 (3.27)	65.03 (3.28)	65.01 (3.48)
>4.3 meter	/	4.97 (0.76)	5.11 (1.33)
Random			
Level 2			
$\sigma^2_{u_0}$ (intercept)	1358.94 (173.03)	1333.86 (169.48)	1472.44 (195.63)
$\sigma_{u_0 u_1}$ (covariance)	/	/	-132.28 (55.11)
$\sigma^2_{u_1}$ (length)	/	/	99.286 (24.49)
Level 1			
$\sigma^2_{e_0}$	452.70 (9.18)	448.92 (9.104)	418.31 (8.57)
-2xloglikelihood	45262.13	45218.96	44963.59

Table 2.2.3: Estimates for the null, variance components, and full random models, with car length as a categorical explanatory variable

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij} \quad (2.2.5a)$$

and additionally two macro-models:

$$\beta_{0j} = \beta_0 + u_{0j} \quad (2.2.5b)$$

$$\beta_{1j} = \beta_1 + u_{1j} \quad (2.2.5c)$$

If the reference category is small cars (<4.3 meters) and the dummy variable x represents long cars (>4.3 meters), this model allows cars of different length at different road sites to have different speeds (cf. Figure 2.2.4). The solid lines in the figure represent the overall general relationship indicating that smaller cars, on average, drive slower than longer cars. However, at road site 5 a pattern is discerned that differs from the overall general relationship, more precisely, at that site, on average, long cars drive slower than small cars.

Table 2.2.3 contains the estimates of the null model, the random intercept model and the full random model. According to the random intercept model drivers of long cars (>4.3 meters) drive on average 4.97 km per hour faster than drivers of small cars (<4.3 meters). This variable is significant, which can be derived from its standard error (the value of the coefficient is greater than twice the value of the standard error). The variation of the intercept is also significant

for the same reason (1333.86>2x169.48). Furthermore, there is a significant decrease in -2loglikelihood when shifting from the null model to the random intercept model (deviance: 45262.13-45218.96=43.17; degrees of freedom=1; p=0.000).

The full random model allows for the difference in speed between small and long cars to vary from road site to road site. On average, there is an increase in speed of 5.11km/h for long cars compared to small cars. This value is significant (s.e.=1.33). The variance of the intercept, of the slope and of the covariance between intercept and slope are all three significant. The negative sign of the covariance indicates again that greater intercepts correspond to smaller slopes. A possible explanation of this pattern was given in a previous section.

2.2.1.9.2. Contextual effects

Another type of extension is to include higher-level variables in the model. Higher-level variables are also referred to as aggregate or ecological variables (Snijders & Bosker, 1999). They describe the higher-level structures in the dataset. This is achieved by including such variables in the relevant macro-models (Jones, 1993). For example, if road site average speed is thought to be affected by traffic count at that road site (C), the random intercept macro model of equation (2.2.2b) can be re-specified to include an extra term, as in:

$$\beta_{0j} = \beta_0 + \alpha_1 C_j + u_{0j} \qquad (2.2.6a)$$

This could for example mean that the average speed at a road site would decrease with increasing traffic count at that road site.

Similarly, the slope terms can also be related to traffic count at a road site.

$$\beta_{1j} = \beta_1 + \alpha_2 C_j + u_{1j} \qquad (2.2.6b)$$

This could for example be explained as follows. At road sites with a low traffic count the real relationship between length and speed is revealed and consists of a strong association between both variables in that a unit increase in length corresponds to a high increase in speed. At road sites with a high traffic count the real relationship is hidden because there is no free flow of traffic; cars are obstructed by one another and therefore a unit increase in length only corresponds to a small increase in speed.

This formulation results in the introduction of an interaction term (the product of x and C) in the combined model. This was defined in the introduction as a cross-level interaction term: interactions between variables measured at different levels in hierarchically structured data (Kreft and de Leeuw, 2002):

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \alpha_1 C_j + \alpha_2 C_j x_{1ij} + \left(u_{1j} x_{1ij} + u_{0j} + e_{ij}\right) \qquad (2.2.6c)$$

Parameter	Null model	Context (level 2 variable) Main effect	Context (level 2 variable) Cross level interaction
	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Fixed			
Intercept	68.69 (3.27)	59.49 (3.50)	59.48 (3.51)
Length	/	1.65 (0.47)	1.68 (0.60)
>100	/	33.17 (6.51)	33.22 (6.53)
>100xlength	/	/	-0.08 (0.97)
Random			
Level 2			
$\sigma^2_{u_0}$ (intercept)	1358.94 (173.03)	1107.59 (141.57)	1107.33 (141.56)
$\sigma_{u_0u_1}$ (covariance)	/	-15.65 (15.87)	-15.59 (15.86)
$\sigma^2_{u_1}$ (length)	/	12.85 (3.15)	12.87 (3.15)
Level 1			
$\sigma^2_{e_0}$	452.70 (9.18)	412.75 (8.46)	412.75 (8.46)
-2xloglikelihood	45262.13	44877.82	44877.82

Table 2.2.4: Estimates for the null model and the models including contextual effects

Table 2.2.4 contains the results of the null model and of two additional models with a level-2 variable. This level-2 variable is a dummy variable with the value 0 representing those road sites where less than 100 cars passed by during observation, while the value 1 was given to those road sites where more than 100 cars passed by during observation. The former is the reference category.

The first model with the main effect of the dummy variable only calculates the influence of traffic count on the average speed at a road site. The second model includes an interaction term between traffic count and length of cars. It shows how the relationship between length and speed changes according to the value of traffic count.

The coefficient of the level-2 variable in the main effect model is 33.17, meaning the average speed of cars at a road site with a traffic count of at least 100 cars increases with 33.17km/h on average compared to road sites where traffic count is below the threshold value of 100. This coefficient is significant (s.e.=6.51). Traffic count somehow reflects the speed regime: higher traffic count corresponds to higher speed regimes, which makes sense because roads that have higher speed regimes are typically busier roads with a higher traffic count. The random parameters show the same pattern as the previous models discussed before: there is a fanning in pattern, although the covariance is not significant. Finally there is significant reduction in the -2xloglikelihood-value: it

drops from 45262.13 to 44877.82 with a difference of 4 degrees of freedom yielding a p-value of 0.000.

Although the coefficient of the interaction term in the third model clearly is not significant, it is interesting from a conceptual point of view to interpret it anyway. It shows that the relationship between length and speed differs according to different values of traffic count. More precisely, for road sites with a traffic count of at least 100 cars, the slope is reduced with 0.08. Put another way, on road sites with a low traffic count the speed increases with 1.68km/h for each unit increase in length of cars, while the speed only increases with 1.60km/h per unit increase in length of cars for road sites with high traffic count. This confirms the previously formulated hypothesis that the real relationship between length and speed may be hidden because of a high traffic count. This coefficient, however, is not significant, hence this third model is not a better one than the main effect model according to the deviance test.

2.2.2 Three level models and more

(Emmanuelle Dupont and Heike Martensen, IBSR)

Section 2.2.1 introduced the "simplest" multilevel model, namely the 2-level model. The present section will show that the same statistical principles apply when the structure of the data at hand contains more than 2 levels. For the ease of comparison with Section 2.2, the same research example and data set will be used here. The relationship between the length of cars and their speed will thus be further assessed, but this time taking account of a presence of a third level in the data hierarchy, namely: the Belgian provinces from which the level-2 units (the road sites) have been selected.

2.2.2.1. Objectives of the technique

The objectives underlying the modelling of data structures with three levels and more are in all points similar to those of 2-level models. The reader is thus referred to Section 2.2 for more information on this topic.

2.2.2.2. Model definition

2.2.2.2.1. The random intercept model:

A first step to take in examining how the 3-level structure affects the relationship between car length and speed would consist of fitting a random intercept model, defining the effect of car length as fixed. The dependent variable "speed" will now be noted "Y_{ijk}" to indicate the speed of car "i" within road site "j" within province "k":

$$Y_{ijk} = \beta_{0jk} + \beta_1 x_{1ijk} + e_{ijk} \qquad (2.2.7a)$$

There are now two equations defining the intercept term "β_{0jk}": one at the second - and the other at the third - level of the data structure. At level 2, the intercept is defined as:

$$\beta_{0jk} = \beta_{0k} + u_{0jk} \qquad (2.2.7b)$$

"β_{0k}" represents the average intercept in level-3 unit "k", and is itself defined by the following equation:

$$\beta_{0k} = \beta_0 + v_{0k} \qquad (2.2.7c)$$

The complete model for Y_{ijk} is thus:

$$Y_{ijk} = \beta_0 + \beta_1 x_{1ijk} + v_{0k} + u_{0jk} + e_{ijk} \qquad (2.2.7d)$$

This model describes speed as being a function of an average speed value (the fixed coefficient for the intercept, or β_0), of the fixed effect of car length ($\beta_1 x_{1ijk}$), and of 3 random deviations from the average intercept value: deviations that occur at the province level (v_{0k}), deviations at the road-site level (u_{0jk}), and deviations occurring within road-sites, between cars (e_{ijk}).

Section 2.2.1 already introduced the VPC, defining it as "the proportion of total residual variation that is due to differences between groups" and related it to the intra-class correlation coefficient – or "the extent to which the y-values of individuals in the same group resemble each other as compared to those from individuals in different groups".

It is nevertheless important to note that this is *only* in the limited context of a *2-level* random intercept model that the two constructs can be so equated (Goldstein, 2003). In the case of 3-level models the two concepts turn out to be close but different, because they actually refer to "two different aspects of the data, which happen to coincide when there are only 2 levels" (Hox, 2002, p. 32).

To illustrate this distinction, the meaning of the VPC and of the intra-class correlation as applied to the 3-level speed dataset must be examined. As it is defined, the VPC corresponds to the ratio of a single level's variation to the total variation. By applying this principle to partition the total variance, the level-2 variance is clearly separated from the level-3 one. The formulas to be used to calculate the variance at level 2 (2.2.8a) and level 3 (2.2.8b) are straightforward extensions of the ones given in Section 2.2.1:

$$\rho_{lev2} = \frac{\sigma^2_{u_0}}{\sigma^2_{v_0} + \sigma^2_{u_0} + \sigma^2_e} \qquad (2.2.8a)$$

$$\rho_{lev3} = \frac{\sigma^2_{v_0}}{\sigma^2_{v_0} + \sigma^2_{u_0} + \sigma^2_e} \qquad (2.2.8b)$$

Matters are different, however, when the intra-class correlation coefficient must be estimated. As a reminder, this coefficient corresponds to the expected correlation between two elements randomly selected within the same higher-level unit (2 or 3). When the model comprises 3 levels, account must be taken of the fact that two level-1 units in the same level-2 one also are *de facto* included in the same level-3 unit (2 speed values recorded at the same road sites were also inevitably recorded in the same province!). Calculating the intra-class correlation at level 2 thus requires that variance components at both level 2 *and* 3 variance components be included in the numerator:

Thus, while for level 3 the same formula (formula 2.2.4b) will be used for the calculation of the VPC and of the intra-class correlation, the intra-class correlation at level 2 will be estimated by 2.2.8c:

$$\rho_{lev2} = \frac{\sigma^2_{v_0} + \sigma^2_{u_0}}{\sigma^2_{v_0} + \sigma^2_{u_0} + \sigma^2_e} \qquad (2.2.8c)$$

2.2.2.2.2. The random intercept and slope model:

The full random model - or a model in which the slope for the effect of car length on speed is specified as being random at level 2 and 3 – would then be defined in the following way:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} x_{ijk} + e_{ijk} \qquad (2.2.9a)$$

$$\beta_{1jk} = \beta_{1k} + u_{1jk} \qquad (2.2.9b)$$

$$\beta_{0jk} = \beta_{0k} + u_{0jk} \qquad (2.2.9c)$$

$$\beta_{1k} = \beta_1 + v_{1k} \qquad (2.2.9d)$$

$$\beta_{0k} = \beta_0 + v_{0k} \qquad (2.2.9e)$$

$$Y_{ijk} = \beta_0 + \beta_1 x_{ijk} + v_{0k} + u_{0jk} + v_{1k} x_{ijk} + u_{1jk} x_{ijk} + e_{ijk} \qquad (2.2.9f)$$

The 3-level model now defines the total variation in the speed of cars as the result of 2 fixed factors (the average intercept and slope), and of 5 sources of random variations. Both level-3 (provinces) and level-2 units (road sites) are said to entail random departure in the cars' speed from the "average speed value" (the fixed intercept) and from the "average length-speed relationship" (the fixed slope). The covariances between the random intercepts and slopes at each level are also part of the model, which raises up to 7 the number of random parameters to be estimated.

Of course, it is by no means compulsory that the effect of any level-1 explanatory variable added to the model were defined as random at both level 2 and 3. A given effect can be declared random at level 3 without being so at level 2, and the other way around. Explanatory variables at either level 2 or 3 can be included in the model, and level-2 explanatory variables can themselves be defined as random at level 3.

2.2.2.3. Model assumptions

The random coefficients at level 2, 3, or higher are all considered representative of distributions of individual effects in the population. These parameter distributions themselves are assumed to be normal, with means 0 and variances $\sigma^2_{u_0}, \sigma^2_{v_0}, \sigma^2_{u_1}, \sigma^2_{v_1}$ for the intercepts and slopes, respectively. At level 2, this implies:

$$\begin{pmatrix} u_{oj} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2{}_{u_0} \sigma_{u_0 u_1} \\ \sigma_{u_0 u_1} \sigma^2{}_{u_1} \end{pmatrix}\right) \qquad (2.2.10a)$$

And, at level 3:

$$\begin{pmatrix} v_{0k} \\ v_{1k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{v_0} \sigma_{v_0 v_1} \\ \sigma_{v_0 v_1} \sigma^2{}_{v1} \end{pmatrix}\right) \qquad (2.2.10b)$$

The level-1 residuals (e_{ij}), are in turn assumed to be normally distributed with mean 0 and variance σ^2 ($\varepsilon_{ij} \sim N(0, \sigma^2)$), and to be independent from one another.

Finally, the level-2 residuals are assumed to be independent over j (i.e.: across the level-2 units, or road sites), the level-3 coefficients are assumed to be independent over k (the level-3 units, or provinces). The residuals at all levels (1, 2, and 3) are assumed to be independent from each other.

2.2.2.4. Research problem

The research problem and the dataset used in this section are identical to those used in Section 2.2.1. Taking account of the full structure of this "length-speed dataset", attempt will be made at determining whether the different Belgian provinces from which the road sites were sampled can be considered to: (1) contribute to the variation of the speed of cars, (2) affect the relationship between car length and speed. Again, it is important to stress that this empirical question was chosen more on the basis of didactical than of theoretical objectives.

2.2.2.5. Dataset

As a reminder, the speed data were collected in Belgium, and consist of a sample of n = 4994 drivers passing by m = 131 road sites. These road sites were themselves selected among 11 provinces.

2.2.2.6. Model fit and diagnostic

With the exception of the Variance Partition Coefficient (VPC), all tools available for diagnostics and for assessing the fit of the model are identical to those outlined with respect to the basic 2-level model. The reader is thus referred to Section 2.2 for a description of deviance tests and tests of single parameters.

The two-level variance-components model that was fitted in Section 2.2.1.8.1 provided indications that a substantial part of the total variation of speed was attributable to the second level of the data hierarchy (road sites). Having now

included the third level – "province" – in the model, this level's contribution to the variation of cars speed can also be assessed. One could, for example, imagine that the average speed is generally lower in some provinces than in others.

Parameter	Random intercept model	Length random at level 2 only	Length random at levels 2 and 3
	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Fixed			
Intercept	74.47 (5.33)	74.49 (5.30)	74.51 (5.34)
Length	2.28 (0.28)	1.68 (0.47)	1.64 (0.59)
Random			
Level 3			
$\sigma^2_{v_0}$ (intercept)	218.01 (132.89)	213.11 (131.59)	219.14 (133.93)
$\sigma_{v_0 v_1}$ (covariance)	/	/	-10.43 (14.37)
σ^2_{v1} (length)	/	/	1.31 (1.60)
Level 2			
$\sigma^2_{u_0}$ (intercept)	990.13 (132.62)	995.89 (133.41)	994.40 (133.26)
$\sigma_{u_0 u_1}$ (covariance)	/	- 12 (15.43)	- 11.55 (15.30)
$\sigma^2_{u_1}$ (length)	/	12.761(3.12)	11.64 (3.12)
Level 1			
$\sigma^2_{e_0}$	446.47 (9.051)	412.74 (8.46)	412.74 (8.46)
-2xloglikelihood	45167.92	44877.42	44876.820
Deviance test	$\chi^2_1 = 24.4$; $p <$ .000[13]	$\chi^2_2 =$ 33.73; $p <$ .000	$\chi^2_2 =$ 0.6; $p =$ .74, *n.s.*

Table 2.2.5: Estimates for the 3-level models: variance components, random slope for car length at level 2, and full random model

Table 2.2.5 summarises the results of the different steps of the 3-level model specification. The deviance test comparing the log-likelihood values associated with the 2-level variance component model (-2 loglikelihood = 4519232, see Section 2.2.1.7.2) and the one associated with the 3-level model reveals a significant improvement in the model's fit ($\chi^2_1 = 24.4$; $p < .000$). Nevertheless, the variance of the random intercept at level 3 is not in itself significant (see Table 2.2.5). The results are thus rather ambiguous with respect to the necessity of including the province level in the model.

Deciding whether the inclusion of an additional level in a model is empirically justified is eased by the estimation of the amount of variation in the observations at this level. Substituting the level 2 and level 3 variance estimates

[13] This deviance test compares the log-likelihood value of the current 3-level random intercept model with the one of the 2-level random intercept model fitted in Section 2.2.

presented in Table 2.2.5 into formulas 2.2.8a and 2.2.8b reveals that 60% of the variation in the speed of vehicles is attributable to the second level (road sites), while 13% only are accounted for by the third level (provinces). Clearly, the amount of variation in cars' speed that is "located" at the third level is far less important than the observations' variance at level 2. The calculation of the intra-class correlation coefficient at each level (formulas 2.2.8c and 2.2.8b for level 2 and 3, respectively) indicates that, should 2 observations be randomly selected from the same level-2 unit, a correlation of about 0.73 between them would be expected. The expected correlation between 2 observations randomly selected from the same level-3 unit is 0.13, a considerably lower value. Both the VPC and the intra-class correlation thus converge to suggest that the dependency among data is much stronger at level 2 than at level 3 (and that more variation in Y is accounted for by level 2 than by level 3). However, the intra-class coefficient value observed at level 3 is not negligible.[14]

The above results did not offer stronger indication of the necessity to take account of the province level when modelling the effect of car length on speed. Defining this effect as random *at level 2* yielded highly similar conclusions in the context of the present three level model than when the model comprised two levels only. However, the third model, in which the effect of car length was defined as random at both level 2 and 3 was not associated with any significant fit improvement. The estimates for the random effects at level 3 are not significant.

2.2.2.7. Model interpretation

Adding the province level to the model resulted in a significant improvement of the model's fit, although not dramatic. Including this level also resulted in a decrease of the random variation of the intercept at level 2 ($\sigma^2_{u_0}$). This is in line with previous observations indicating that the random variation associated with levels that are present in a given data hierarchy, yet omitted from a model is "added" to the residual variation associated with the levels that are specified in the model (Moerbeek, 2004). Imagining, for example, that the random variation in speed records associated with the third level of the data hierarchy would have been very important but that this level had not been explicitly included in the model, then the failure to specify this third level would have resulted in the associated variance "summing" up to the level 2 and level 1 residual variation. These two would have looked more important than what they actually are, simply as the result of a model misspecification.

Compared to level 2, however, our third level cannot be said to contribute much to the variation of the criterion variable. This level does introduce some dependency in the observations, although to a far lower extent than level 2

[14] As a reference : An intra-class correlation of about 0.01 is considered small, while 0.20 is considered a large value (see Kreft & De Leeuw, 1999 for a more detailed discussion of this topic and of the relation between the size of the intra-class coefficient and the standard errors of the estimated parameters).

does. Including the third level in our model of car speed entailed almost no change in the estimation of the parameters as compared to when the 2 level model was fitted. This suggests that "omitting" this level from the model does not result in serious model misspecification.

The 3 level model fitted here can of course be extended, for example through the inclusion of explanatory variables at level-3. In Section 2.2.1, the analysis of the cross level interaction between "traffic count" (an explanatory variable at level-2) and "car length" (at level 1) was described. Although this interactive effect was not significant, the associated coefficient was interpreted there for the purpose of illustration: This coefficient was negative, suggesting that the speed-length relationship was lower at road-sites with less important traffic flows (i.e.: with low traffic count). In the framework of the present 3-level model, the length-traffic count interaction term could itself be defined as random at level 3. The finding of a significant random slope would indicate that the variation of the effect of length on speed depends on traffic count and besides varies (randomly) between provinces. How should the level-3 covariance between this random slope for the interaction effect and the random intercept be interpreted? - A negative covariance would indicate that the traffic count-by-car-length interactive effect is weaker for provinces that are characterised by higher average speed values. In other words, the higher the province's average speed, the more homogeneous the effect of length – or the less affected it would be by the different traffic count values associated with each road sites.

These hypothetical considerations make it clear that models with three levels and more offer the same possibilities as their 2-level counterparts, but that these possibilities are multiplied by the number of levels under analysis. One should bear in mind, however, that this comes at the cost of parsimony, on the one hand, and of ease of interpretation, on the other. The example of the cross level (1 and 2) interaction made random at level 3 illustrates the fact that multiple level models can quickly become "difficult to follow from a conceptual point of view" (Hox, 2002, p. 30).

The number of parameters to be estimated increases in a multiplicative way along with the number of levels included in the model: The simple definition of an effect as being random at both level 2 and 3 implies the estimation of 7 parameters. Defining the effect of another explanatory variable at level-1 as being random at level 2 will involve the estimation of 3 additional parameters (the fixed effect, random slope, and random intercept and slope covariance); raising the total number of parameters to ten. Further specifying this additional parameters' effect as being random at level 3 as well amounts to estimating 2 parameters more (the random slope and intercept-slope covariance at level 3), and so on…

Independently of increasing the difficulty of interpretations, estimating important number of parameters also augment the risk of encountering estimation problems (algorithms failing to converge…). Caution is thus required when fitting models with 3-levels and more. It is usually recommended that the definition of effects at the various levels be grounded on sound theoretical

reasons, or empirical evidence, rather than on mere exploratory attempts (Hox, 2002; Snijders & Boskers, 1999; Kreft & De Leeuw, 2002).

2.3 Discrete response models

2.3.1 Introduction

(Emmanuelle Dupont and Heike Martensen, IBSR)

Sections 2.3.2, 2.3.3, and 2.3.4 of this deliverable respectively focus on the multilevel analysis of three different types of discrete data, namely: dichotomous responses, counts, and multinomial responses. The aim of this introductory section is to provide the reader with useful preliminaries that will allow him/her to apprehend the general framework of the multilevel analysis of discrete data analysis, i.e., the Multilevel Generalised Linear Model (MGLM). This introduction starts with a reminder of the structure underlying the familiar linear model and describes the main properties of discrete response variables. On this basis, the risks associated with the straightforward application of the linear model to discrete response variables are illustrated, and the solution provided by the Generalised Linear Model is outlined. The general principles underlying the *multilevel* generalised linear model are then defined.

2.3.1.1. Reminder: The linear model

Response =	***Systematic component***	***+***	***Random component***
y_i =	η_i	+	ε_i
	"How does response vary with covariates /predictors/explanatory variables?"		"What kind of distribution do data follow?"
The *linear* model:	$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} = \beta_0 + \sum_{j=1}^{k}$		$y_i \sim N(\eta_i, \sigma^2)$ $\varepsilon_i \sim N(0, \sigma^2)$
	$E(y_i) = \sum_{j=0}^{k} \beta_j x_{ij} = \eta_i$		$\mathrm{Var}(y_i) = \mathrm{Var}\left(\sum_{j=0}^{k} \beta_j x_{ij} + \varepsilon_i \right.$

Table 2.3.1: The linear model and related assumptions

As it is already mentioned in Section 2.2, any statistical model defines a response variable ($y_i, i = 1,\ldots,n$) as the result of a systematic and of a random component. The *systematic component* of the model describes how the response varies with explanatory variables or predictors ($x_k, h = 1,\ldots,r$). This component is the one that defines the *expected value* of the response variable. The generic term used to refer to the systematic component and, by extension, to the expected value is η_i. The *random component* of the model defines the variation of the observations that the model cannot explain. It defines the

distribution that the observations and the residual follow (Mc. Cullogh & Searle, 2001).

When the particular model adopted is the linear one, two main assumptions are associated with the definition of the observations that the systematic and the random components respectively provide: (a) the expected value for each observation corresponds to a linear combination of unknown parameters, considered constants, and (b) the data come from a normal distribution with mean η_i and variance σ^2 (McCulloch & Searle, 2001).

2.3.1.1.1. Properties of discrete variables

Discrete response variables often happen to be the focus of road safety research. Attempts at modelling the probability of occurrence of given events - such as the survival of vehicle occupants after a crash, or the infringements that drivers commit - are common. Count data, such as the number of accidents occurring within a given time frame, are also regularly encountered as response variables. As can be seen by comparing the features of the normal distribution to those of the discrete distributions listed in Table 2.3.2, there are two general properties of discrete data that prevent a straightforward application of the linear model. The first is these data's restricted ranges, the second is that they have related mean and variance.

Binary outcomes correspond to data that can take two values only: "1" (usually defined as "success", such as the survival of car drivers following a crash) or "0" (usually defined as "failure", such as the dead of car drivers following a crash). The number of successes in *m* samples can be described by a stochastic variable which is binomially distributed, with parameters *φ* and *m*. Now assume that in several regions the number of crashes and the number of dead drivers resulting from these crashes within a certain period are known. Then the number of drivers that survived a crash in region i, denoted by y_i, is binomially distributed with parameters φ_i and m_i, where φ_i is the probability that a driver survives a crash in region i and m_i is the number of crashes in region i. Then $\mathrm{E}(y_i) = m_i\varphi_i$ and $\mathrm{Var}(y_i) = m_i\varphi_i(1-\varphi_i)$.

Count data are to be conceived of as the number of events occurring during an interval of time having length m_i, or within an area having size m_i. They also have restricted range in the sense that they can only take positive values. When the counted occurrences are rare,[15] such data can be considered to follow the Poisson distribution. In several cases, the events being counted are actually the outcomes of discrete trials, and would more precisely be modeled using the binomial distribution. However, the binomial distribution with parameters n and λ/n, i.e., the probability distribution of the number of successes in n trials, with

[15] i.e., when there are less than 10 cases of the counted event within the time period or the area considered (m_i), according to Langford & Day, 2001.

probability λ/n of success on each trial, approaches the Poisson distribution with expected value λ as n approaches infinity. When the occurrence assessed is frequent, the binomial distribution is more appropriate. The Poisson distribution is characterized by the "exposure" term, m_i and by the event rate λ_i. This distribution has a variance equal to the expected value, namely the mean, so that the two parameters are related, as was also the case for binary data.

Distribution of the response variable:	***Sampling model (Raudenbusch & Bryk, 2001)***
Normal:	$y_i \sim \mathrm{NID}(\mu_i, \sigma^2)$ $\mathrm{E}(y_i) = \mu_i$ $\mathrm{Var}(y_i) = \sigma^2$
Bernouilli and Binomial	$y_i \sim \mathrm{B}(m_i, \varphi_i)$ $\mathrm{E}(y_i) = m_i\varphi_i$ $\mathrm{Var}(y_i) = m_i\varphi_i(1-\varphi_i)$
Poisson	$y_i \sim \mathrm{P}(m_i, \lambda_i)$ $\mathrm{E}(y_i) = m_i\lambda_i$ $\mathrm{Var}(y_i) = m_i\lambda_i$
Multinomial responses	$\mathrm{E}(y_m) = n\varphi_m$ $\mathrm{Var}(y_m) = n\varphi_m(1-\varphi_m)$ $\mathrm{Cov}(y_m, y_{\hat{m}}) = -n\varphi_m\varphi_{\hat{m}}$

Table 2.3.2: Sampling models for normal, binary/binomial, counts and multinomial responses .

The term "multinomial responses" refers to categorical data, or to responses that can take one of a few number of values. Assume that *Y* is a random variable which can take its value in M categories and let φ_m be the probability that Y is in category m. If there are n observations of the random variable Y and y_m is the number of observations in category m, then Y is multinomially distributed with parameters $\mathrm{M}, \varphi_1, \ldots, \varphi_M$.

2.3.1.1.2. Applying linear models to discrete data

Given the particular properties of discrete observations, and the assumptions made by the linear model, two main problems would result from a straightforward application of a linear model to discrete data.

First, the response variable being defined as a *linear* function of some explanatory variables or predictors, the fitted values generated on the basis of the model's systematic component are likely to lie outside the actual range of the observations. What would be modelled in this case would be something that conceptually differs from the observations (values outside the 0-1 range cannot be considered as probabilities). The distribution specified by the model would neither correspond to the actual distribution of the observations, nor to the residual distribution.

Second, the relation existing between the expected value and the variance of discrete observations implies that, once predictors are included in the model, the variance of the error term is not homoscedastic any more (i.e., is not constant and depends on the particular values taken by the predictor(s))[16].

Similar problems are encountered if the response variable consists of non-normal data (and not of well-defined discrete distributions such as the Poisson or Binomial ones). One solution could be to apply an appropriate transformation to these observations, then submit them to a "standard" linear analysis. Such an approach has been – and still is – common practice in data analysis. Although it remains useful and appropriate in particular circumstances (such as in the exploration phase of data analysis), it offers no certainty that the application of linear methods to the transformed data will allow safe inference. First, transformation of data may not be an option at all: One can wonder, for example, which transformation could ever make dichotomous data resemble a normal distribution (Hox, 2002). Second, transforming the observed response offers no guarantee that the error distribution will be normally distributed, an essential condition to be met when applying linear models (Hox, ibid). As it will be explained in the next section, the generalised linear model is a far more advanced technique than transformation, in the sense that it includes "the necessary transformation *and* the choice of the appropriate error distribution (…) explicitly in the statistical model" (Hox, 2002, p. 104).

2.3.1.2. The generalised linear model (GLM)

The Generalised Linear Model is more than a particular statistical technique that conveniently allows overcoming the problems posed by discrete and/or non-normal data. It must be conceived as a broad class of statistical models, in which the linear model itself is encompassed.

[16] In a model fitting binary responses, for example, the residual can take only two values : "1- ($\beta_0 + \beta_1 x_{i1}$)" and "$\beta_0 + \beta_1 x_{i1}$"

The appropriateness of the GLM to analyse discrete data relates to the fact that it "generalises" both the distributional assumptions made about the data, and the systematic component defining the expectations.

With respect to the observed distribution, the GLM makes the general assumption according to which the response variable has a probability distribution that pertains to the "exponential family" (see Dobson, 1990 for a formal definition). This family of distributions encompasses a large number of probability distributions, both continuous and discrete. As a consequence, all specific distributions pertaining to this broad class can be used in the GLM to specify the distribution of the observations.

Distribution of the response variable:	***Link function (Raudenbusch & Bryk, 2001)***
Normal:	- The identity link - $\eta_i = \mu_i$
Bernouilli and Binomial	- The logit link[9] - $\eta_i = \log\left(\frac{\varphi_i}{1-\varphi_i}\right)$
Poisson	- The log link - $\eta_i = \log(\lambda_i)$
Multinomial responses	- The logit link – $\eta_j = \log\left(\frac{\varphi_j}{\varphi_l}\right)$ (with l being the reference category)

Table 2.3.3: Examples of link functions used for normal, binary/binomial, counts and multinomial responses.

The GLM also renders possible fitting "correct" predicted values. Indeed, the usual linear component (the sum of predictors expected to affect the response) is not directly equated to the expected values any more, but to some function of them, called a *link function*. This transformed version of the original response variable (probabilities, counts…) is not restricted in range (it can take values outside 0 and 1, and positive and negative values). There exists some "inverse function" on the basis of which these predicted values can be transformed back into the "metric" of the units initially measured (i.e., probabilities, counts…).

The use of link functions to relate expected values to the predictors included in the model thus prevents fitting "out-of-range" expected values. Various link functions are available; the choice of the appropriate one depending on the nature of the data that are to be modelled. Table 2.3.3 describes a number of

link functions, which are used in Sections 2.3.2 to 2.3.4. As a point of reference, Table 2.3.3 provides the link function corresponding to the normal distribution (the "identity link"). Because the functions listed in this table equate the linear component of the model to the natural parameter of the distribution at hand (φ_m ,etc.), they are also termed *canonical* link functions. Other link functions are available, however (see Dobson, 1990 for examples).

As many data in traffic safety are binary, the logit link for binomially distributed data is the one that will mostly be used in next sections over discrete data analysis. The logit link is defined as the logarithm of odds ratio. The odds ratio themselves correspond to ratio of probabilities. As an example, the log of the odds of survival following an accident amounts to the log of the ratio of the probability to survive (φ_i) to the probability of dying ($1-\varphi_i$) as a consequence of the i-th accident[17].

The link function included, the systematic component of the GLM for binomial or binary data with two explanatory variables writes out:

$$\eta_i = \log\left(\frac{\varphi_i}{1-\varphi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}. \qquad (2.3.1)$$

This indicates that the predicted values fitted by means of the link function are predicted log-odds. How should the coefficients for the different predictors making up the linear component be interpreted? Predicted log-odds can be converted to odds, by taking their exponential[18].

$$\exp\left(\log\left(\frac{\varphi_i}{1-\varphi_i}\right)\right) = \frac{\varphi_i}{1-\varphi_i} \text{ (the predicted odds-ratio)} \qquad (2.3.2)$$

Since the exponential function is applied to the predicted values, it has to be applied to the predictors making up the linear component of the model, in order to obtain:

$$\frac{\varphi_i}{1-\varphi_i} = e^{\beta_0} \times e^{\beta_1 x_{i1}} \times e^{\beta_2 x_{i2}} \qquad (2.3.3)$$

The relation between the different predictors that was additive when the log-odd function was applied is now multiplicative. Consequently, the coefficients for the predictors must be interpreted as the multiplicative effect, associated with a one-unit increase on x_{i1} (for coefficient β_1) , x_{i2} (for coefficient β_2) , etc, on the odds-ratio.

The estimated values of the predictors can in turn be converted into predicted probabilities using the formula:

[17] The binomial model is actually a special case of the multinomial model, for which the numerator $1-\varphi_i$ has to be replaced by the probability of a reference category φ_l .

[18] The exponential function is the inverse of the log function.

$$\varphi_i = \frac{1}{1+\exp(-\eta_i)} = \frac{1}{1+\exp(-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))} \qquad (2.3.4)$$

2.3.1.3. The Multilevel Generalised Linear Model (MGLM)

The essential feature of a *multilevel* generalised linear model is the fact that the individuals from which the data are received belong to groups and the groups themselves are a random sample from a population of groups. In the linear case, higher levels were accounted for by assigning the "j" subscript to the systematic component - μ_i - allowing this parameter to vary randomly across the higher-level units. The corresponding variance was then estimated. The same principle applies in the framework of the multilevel generalised linear model, to the difference that what will be declared to vary across higher-level units are the *transformed* values of the parameters of the distribution at hand (i.e., $m_i\varphi_i$; $m_i\lambda_i$, and so on…) . For the sake of simplicity, the distribution parameter will now be referred to as to "π".

In a two-level generalized linear model, the expected value of the response y_{ij} – provided by individual *i* in group *j* – is defined as being a probability, or a count, or whatever the particular form taken by the observations. The model must account for the particular type of distribution that these observations follow. In the case of binomial responses, for example, the sampling model corresponding to the observations would be;

$$y_{ij} \approx Bin(n_{ij}, \pi_{ij}) \qquad (2.3.5)$$

The expected value for the response y_{ij} is consequently defined as

$$E(y_{ij}/\pi_{ij}) = n_{ij}\pi_{ij} \qquad (2.3.6)$$

… and the variance as:

$$Var(y_{ij}/\pi_{ij}) = \frac{\pi_{ij}(1-\pi_{ij})}{n_{ij}} \qquad (2.3.7)$$

On this basis, the first level of the multilevel model is written as:

$$y_{ij} = \pi_{ij} + e_{ij}Z_{ij} \qquad (2.3.8a)$$

And,

$$Z_{ij} = \sqrt{\frac{\pi_{ij}(1-\pi_{ij})}{n_{ij}}}, \ \sigma^2 e = 1 \qquad (2.3.8b)$$

The two parameters in this model adequately reflect the distribution specified by the sampling model (2.3.5). In this case, the data are assumed to follow the

binomial distribution, hence the formula for Z_{ij}. For another distribution, Z_{ij} would be defined in another way (variance functions for different types of discrete responses are listed in Table 2.3.2). Note that $\sigma^2 e = 1$ can be estimated instead of being constrained to 1. This is a general device in GLMs that allows testing whether the variance of the observations indeed follows the distribution specified by the sampling model. This is achieved by estimating this parameter, and examining whether the obtained value significantly differs from 1. In such a case, it is recommended to keep on working with the estimated parameter to perform the remainder of the analyses.

Depending on the type of observations, the mean response value - π_{ij} - is either a probability, a count, etc… With the adequate link function, it can be expressed as a linear function of parameters:

$$\pi_{ij} = f(\beta_0 + \beta_1 x_{1ij}) \qquad (2.3.9)$$

Higher-level effects can be incorporated in this linear combination of predictors just as they were in the case of the linear models. Thus, the effect(s) of the higher-level units is defined on the values of the level-1 units that are transformed according to the link function used. So, in the case of the logit link function:

$$Logit\,(\pi_{ij}) = \gamma_0 + u_{0j} \qquad (2.3.10)$$

It is important to note that although the variation of the residual variation at level 1 (*2.3.8a)* is defined as following a discrete distribution, the *level 2* random variation of transformed π_{ij} values ($\sigma^2 u_0$, for example) are expected to be normally distributed..

2.3.1.3.1. The empty model

The empty model for the *linear* hierarchical model, defines a response variable as a function of an average value, the intercept, which is specified to vary randomly across the level-2 units. For the logit-link this gives:

$$y_{ij} = \pi_{ij} + e_{ij} Z_{ij} \qquad (2.3.10a)$$

$$Logit\,(\pi_{ij}) = \gamma_0 + u_{0j} \qquad (2.3.10b)$$

where γ_0 represents the average of logit (π_{ij}) across groups and u_{0j} the deviation of the logit in group j from the population average logit (γ_0). These deviations are assumed to be normally distributed, with mean 0 and variance $\sigma^2_{u_0}$, just as this was the case with the linear multilevel model. The model does not contain a parameter for the level-1 variance. This is because for discrete responses the variance follows directly from the expected value as indicated in Table 2.3.2.

Using the inverse of the logit function, the logits can be reconverted into probabilities.

$$Logit\ (\pi_{ij}) = \gamma_0 + u_{0j}$$
$$\pi_{ij} = \frac{1}{1+\exp(-(\gamma_0 + u_{0j})} \qquad (2.3.10c)$$

The reader should however bear in mind that there is no direct relation between population average of the logits (γ_0) and the population value of the discrete variable itself (π_0). The same is true for the level-2 variance ($\sigma^2_{u_o}$) - which concerns the variation of the logits and cannot directly be equated to the variance of the discrete values π_{ij} themselves. Although in each case the reconverted former value can be considered a proxy for the latter, they cannot be considered equivalent, . This is so because the link between them (the logit link) is a nonlinear one.

Another important difference between the linear and the GLM hierarchical model concerns the level-1 residual. For discrete responses, the individual residual variance $e_{ij}Z_{ij}$ is a function of the mean π_{ij} ($Z_{ij} = \sqrt{\frac{\pi_{ij}(1-\pi_{ij})}{n_{ij}}}$). Consequently, the residual variance in a MGLM model cannot be constant as it is the case in linear models. In the MGLM, the groups will have different within-group variances, because π_{ij} depends on u_{oj}. Given that π_{ij} constrains the value of Z_{ij}, this will lead to a different $e_{ij}Z_{ij}$ value for each group.

2.3.1.3.2. The random intercept model

Generally speaking, the random intercept model differs from the empty model in the sense that – besides specifying the intercept as being random – fixed explanatory variables may also be included in the model.
Once explanatory variables enter the model, the expected value of the discrete variable (π_{ij}) cannot any more be considered as a sole function of the level-2 units. Indeed, if some of these predictors are characteristics of the lowest level units (i.e. if they are level-1 predictors), the values fitted are likely to differ for all individuals within the groups. Consequently, the expected discrete value must now be denoted by the "ij" subscript, so that the model becomes:

$$y_{ij} = \pi_{ij} + R_{ij} \qquad (2.3.11a)$$
$$Logit\ (\pi_{ij}) = \gamma_0 + \sum_{h=1}^{r} \gamma_h x_{hij} + u_{oj} \qquad (2.3.11b)$$

The log of the odds of π_j are now defined as being a function of a linear combination of an average population value (γ_0), of the effect of level-1 (and/or) level-2 predictors ($\sum_{h=1}^{r} \gamma_h x_{hij}$) and of a group-related random deviation u_{oj}.

2.3.1.3.3. The random intercept-and-slope model:

The specification of a random intercept and slope MGLM poses no particular additional difficulty, once the general multilevel structure in the GLM is made clear. The model would then describe the expected value as being a nonlinear function of predictors and random effects at higher level(s), and replace this expected value in the framework of the sampling model that is appropriate for the outcome variable. It can be written as:

$$Logit\,(\pi_{ij}) = \gamma_0 + \sum_{h=1}^{r} \gamma_h x_{hij} + u_{oj} + u_{1j} x_{1ij} \qquad (2.3.12)$$

2.3.1.3.4. Over- or underdispersion

When fitting a (hierarchical) generalised linear model, the choice of the distribution at level 1 is often dictated by the nature of the empirical data. For example, Poisson regression analysis is commonly used to model count data, while binary data are modelled under the binomial distribution. It is however possible that the data do not exactly follow the assumed distribution. If the observed level 1 variance is larger than the variance of the distribution assumed, *overdispersion* has occurred. Conversely, *underdispersion* means that there was less variation in the data than predicted.

Overdispersion often indicates heterogeneity in the sample. This can be due to underspecification of the model in terms of predictor variables or in terms of hierarchical levels (i.e., there is variation introduced in the observations by them being clustered into higher levels, without this being specified in the model). Although the parameter estimates are usually still correct, in the case of overdispersion the variance is underestimated suggesting a higher confidence in the estimates than is actually appropriate. The opposite is the case with underdispersion. In both cases it is possible to generalise the model by estimating a scalar variance component α. The variance originally specified by the distribution has to be multiplied by this estimated factor in order to match the observed variance (Raudenbush & Bryk, 2002). Estimating this scalar component is actually a way to test for over- or underdispersion (see Sections 2.3.2 and 2.3.4).

2.3.1.3.5. Estimation methods and tests of the parameters

Although the underlying general principle appears simple, fitting multilevel GLMs can not yet be considered pure routine. This is related to the fact that

“level-1 sampling for discrete models is not normal, while the higher-level model involves multilevel normal assumptions poses a problem for conventional estimation theory” (Raudenbush & Bryk, 2002, p. 352). Without entering details of estimation methods, it is important for the reader to be aware that most current software – at the time of writing - can be considered to use approximate methods. Three main approaches can be distinguished on this basis: The first involve the computation of Maximum Likelihood values. This is the most computationally intensive method, performed by some software (such as SAS). The second involves the approximation of Maximum Likelihood values. This is the approach followed in the analyses described in the present document. The main consequence of using approximate likelihood value is that the estimated likelihood values are not reliable any more, and cannot be used to perform the usual Likelihood-Ratio Test (Leyland & Goldstein, 2001; Rasbash et al., 2004). The third approach relies on the use of Bayesian estimation, such as the use of Markov Chain Monte Carlo methods (see Section 2.1.8.4).

2.3.1.4. Conclusion

For the statistical analyses of discrete responses, the generalised linear model (GLM) and its multilevel extension, the hierarchical GLM was introduced. This introduction provides the theoretical background required for proceeding to fitting the multilevel GLMs that are presented in the following sections. The clear advantage of GLMs is their flexibility to model response variables of very different types. The cost this comes with is an increased complexity, less straightforward interpretation of the parameters and less reliable estimation procedures. As noted above, matters are still evolving with respect to the implementation of these methods on software. The reader interested in the use of the MGLM is strongly recommended to read the Section 2.8 àver Bayesian estimation methods.

In road-safety research many of the important response variables are non-linear and therefore require the GLM approach. This will be demonstrated in the following sections. In Section 2.3.2 and 2.3.3 data from a road-site survey with respect to drink driving will be presented. In Section 2.3.2, this data will be analysed as binary responses (driver has drunk or not) and in 2.3.3 as multinomial responses (not drunk, moderately drunk, drunk). In both cases the effect of the particular road-site at which measurement has taken place is included as a second level in a hierarchical GLM. In Section 2.3.4, counts of fatal accidents are modelled with a hierarchical GLM in order to detect regional variation in the number of accidents and in the effect of law-enforcement measures. It can be concluded that hierarchical GLM forms a tool that cannot be missed in the analysis of road-safety data.

2.3.2 Binary and general binomial responses

(Ward Vanlaar, IBSR[19])

Many variables observed in traffic research are binary variables with only two possible values, rather than continuous variables. As an example we will consider the results of a Belgian roadside survey in which drivers were stopped at randomly selected road-sites. In addition to a number of explanatory control variables, the blood alcohol concentration (BAC) was measured as well. Results of this continuous variable were stored and analyzed according to a binary format; zero indicates a BAC below the legal limit while one corresponds to a BAC at or above the legal limit. Such a binary dependent variable can be modelled using logistic regression analysis.

2.3.2.1. Objectives of the technique

As for other regression techniques, the objective is to look for an appropriate function to model the relationship between a set of explanatory variables (this set can consist of continuous variables, categorical variables or a mixture of both types of variables) and the dependent variable. Specific to the logistic regression analyses presented here is that the dependent variable is binary so the responses can only take the values of 0 or 1.

The multilevel version of logistic regression presented here allows assigning the observed variance to different hierarchical levels and investigating whether the model that was found fits the data well. A proper multilevel representation allows for reliably testing whether the relationships found in the data can be generalized to the population.

2.3.2.2. Model definition

Models for binary data concern the probability π_{ij} that the observed variable y_{ij} from person *i* in cluster *j* takes the value 1 (as opposed to 0). In our example with BAC as an underlying continuous variable, the logistic model can be construed as a threshold model (Snijders and Bosker, 1999). The threshold is the legal limit; if BAC is equal to or greater than this threshold then the dependent variable is one, if BAC is smaller than the threshold, then it is zero. The model can then be written in terms of the underlying continuous variable y^*_{ij} – note that the asterix is used as a symbol to denote the underlying continuous or latent variable, rather than the observed variable.

$$y_{ij}^{*} = \beta_{0j} + \beta_1 x_{1ij} + e_{ij} \qquad (2.3.13)$$

Where

[19] *This section is mainly based on Vanlaar, 2005b.*

$e_{ij} \sim \text{logistic}(0, \pi^2/3)$, with mean zero and variance $\pi^2/3 = 3.29$ (in this case π does not denote a parameter but the number 3.141).

The advantage of constructing the model on the basis of an underlying continuous variable is that the level 1 errors can be assumed to follow the logistic distribution and therefore the error variance is known. More generally, binary data are assumed to follow the binomial distribution, whether they are derived from an underlying continuous variable (e.g., above/below average, severely injured/slightly injured, passed /failed, etc.) or not (e.g. male/female, yes/no, dead/alive, etc.). The model for logistic regression is based on this distribution.

In order to analyse the probability π_{ij} that the observed variable y_{ij} takes the value 1 (as opposed to 0) in the generalised linear model, a link function has to be chosen. For a discussion of possible link functions e.g., logit, probit, or log-log functions) see Snijders and Bosker (1999). In this document the most popular link-function, the logit function, will be used, meaning the analyses that are conducted, are multilevel logistic regression analyses.

A 2 level logistic variance components model for binary responses as an equation for the probability π_{ij} is (Rasbash et al., 2004, p. 111):

$$\text{logit}(\pi_{ij}) = \text{logit}\frac{\pi_{ij}}{1-\pi_{ij}} = \beta_{0j} + \beta_1 x_{1ij} \qquad \textit{(2.3.14a)}$$

$$\beta_{0j} = \beta_0 + u_{0j} \qquad \textit{(2.3.14b)}$$

To interpret the relationship between the binary response and an explanatory variable, logit coefficients were transformed into odds ratios using the exponential transformation (see Rasbash et al. 2000 and Rasbash et al. 2004 for a detailed explanation). These odds ratios compare the odds for drink driving of a certain category of a variable (for example the odds for drink driving for the category “female” of the variable “gender”) to the reference category of that variable (in this example the reference category is “male”).

Taking the exponentials of each side of (2.3.14a), we obtain:

$$\frac{\pi_{ij}}{1-\pi_{ij}} = \exp(\beta_{0j}) \times \exp(\beta_1 x_{ij}) \qquad \textit{(2.3.15a)}$$

If we increase x by one unit, we obtain:

$$\frac{\pi_{ij}}{1-\pi_{ij}} = \exp(\beta_{0j}) \times \exp(\beta_1 (x_{ij}+1)) = \exp(\beta_{oj}) \times \exp(\beta_1 x_{ij}) \times \exp(\beta_1) \qquad \textit{(2.3.15b)}$$

This is the expression in (2.3.15a), multiplied by $\exp(\beta_1)$ (i.e., e^{β_1}). Therefore $\exp(\beta_1)$ can be interpreted as the multiplicative effect on the odds for a 1-unit

increase in x. If x is binary (like gender), then $\exp(\beta_1)$ is interpreted as the odds ratio, comparing the odds for units with $x=1$ relative to the odds for units with $x=0$, i.e., the reference category. More generally, if x is categorical, then $\exp(\beta_1)$ is interpreted as the odds ratio, comparing the odds for units with a value for x, different from 0 (1, 2, 3, etc. depending on how many categories the categorical variable consists of) with $x=0$, i.e., the reference category.

2.3.2.3. Model assumptions

The model assumptions for the binomial model are listed below.

$u_{0j} \sim N(0, \sigma^2_{u0})$, the road-site-specific component of the intercept is assumed to be normally distributed with mean zero and variance σ^2_{u0}.

$y_{ij} \sim Bin(1, \pi_{ij})$, the observed binary responses are assumed to follow the binomial distribution with denominator 1, expected value π_{ij} and variance $\pi_{ij}(1-\pi_{ij})$.

2.3.2.4. Research problem and Data set

In 2003 the Belgian Road Safety Institute organised the third national roadside survey to estimate the proportion of drink drivers and their profile (Vanlaar, 2005 b). The objective of this initiative was to gather epidemiological data as a basis to formulate theory- and research-based recommendations to policymakers with the intention of decreasing the number of alcohol related accidents and victims on Belgian roads. This roadside survey is repeated every two years to study trends in drink driving.

According to the official statistics on police enforcement 6% of all tested drivers were at or above the legal limit (BIVV, 2002). This result corresponds to the results from the SARTRE survey (2004): 6% of fully licensed, active Belgian car drivers report they may have been driving during 1 or more days in the past week while over the legal limit for drinking and driving. The first percentage, however, is based on a non-representative sample as a result of a selective way of sampling drivers. Therefore, it is impossible to generalise this result to the Belgian population of car drivers as a whole. The second percentage most probably suffers from a bias due to social desirability.

The data presented here were gathered during a drink driving roadside survey in 2003 according to a stratified two stage cluster sample. The first stage of the roadside survey consisted of randomly selecting road sites (m=413) in each region using a Geographical Information System (Arcview). The road sites are also called primary sampling units (PSU's). Once the sampling of road sites was completed, each site was randomly linked to one out of four possible time spans (weekday; weekday nights; weekend days; weekend nights). Therefore, the

sampling design is not only stratified in space (per region) but also in time. The second stage of the roadside survey consisted of randomly stopping drivers (n=11,186). Once stopped, they were asked by the police to perform an alcohol breath test.

The outcome variable is a binary variable based on the blood alcohol concentration (BAC) of each driver. For the purpose of the multilevel analysis it has been recoded with 0 representing those drivers with a BAC below the legal limit and 1 representing those drivers with a BAC at or above the legal limit. Drivers at or above the legal limit are referred to as drink drivers.

The individual explanatory variables (level 1 explanatory variables) are Gender, Age (a categorical variable consisting of the following age groups: 16-25, 26-39, 40-54, 55+), Previously (a binary variable distinguishing between drivers who previously have been stopped and tested at a road site at least once and drivers who have never been stopped and tested at a road site before) and Probability (a categorical variable representing the driver's perception of the probability of being tested for drink driving; drivers could answer: very low, low, medium, high, very high).

The aggregated explanatory variables (level 2 explanatory variables) are Traffic count (a continuous variable indicating the total number of vehicles driving by the road site during the police check) and Intensity (a continuous variable calculated by dividing the number of policemen per road site by traffic count for that road site).

2.3.2.5. Model fit and diagnostics

A two-level binomial model was fit with drivers at level 1 and road sites (the PSU's) at level 2. To model the relationship between the binary response and the set of explanatory variables, the logit function was used as a link function, meaning a multilevel logistic regression was performed (Rice, 2001).

The results for the final model, containing all explanatory variables described in the previous section, are presented in Table 2.3.4. Two versions were estimated, a binomial model, in which the variance is constrained to be 1 and an extra binomial model, which does not impose such a constraint. The final model fits the data well, which can be derived from the level 1 variance σ_e=0.712 in the extra binomial model, which is close to the theoretical value of 1 (restriction imposed by the binomial distribution). This means there is little evidence that our model exhibits extra binomial variance, more precisely underdispersion[20] – the binomial distribution holds. As can be seen in Table 2.3.4, the strength and the direction of all relationships remain unchanged between both models.

With threshold models the Variance Partition Coefficient (VPC), as defined in Section 2.2.1.7.1, can be applied to the latent variable. "Since the logistic distribution for the level one residual implies a variance of $\pi^2/3$=3.29", the VPC

[20] *Underdispersion refers to the situation in which the total variance is less than 1; conversely, overdispersion corresponds to a total variance, greater than 1.*

formula simplifies to $\rho = \sigma_u / (\sigma_u + \pi^2/3)$ – with σ_u being the level 2 variance and π being the number π (Snijders and Bosker, 1999, p. 224). In our case the VPC, while controlling for the explanatory variables, is 0.231. This means 23.1% of the total variance is level 2 variance, which justifies modelling the data according to a multilevel structure.

2.3.2.6. Model interpretation

The influence of the independent variables on the outcome variable is interpreted based on the exponential coefficients (i.e., odds ratios) of the binomial model in Table 2.3.4, using the definition explained in the section on model definition.

There is a significant (joint chi square test=10.464, df=1, p=0.001) negative relationship between *Traffic count* and the odds of drink driving when controlling for intensity of stopping drivers and for the other independent variables. For each additional car at a road site the odds of drink driving are multiplied by a factor of 0.998. This means that the odds of drink driving decrease by 0.2%, or, per 100 extra cars on a site, the odds are multiplied by a factor of 0.819 (exp(-0.002x100)), meaning that the odds of drink driving decrease by 18.1%.

One could argue that this relationship is of a spurious nature caused by the fact that drink driving takes place primarily on weekend nights with low traffic while there are less drink drivers during the day when there is much more traffic. Therefore another series of analyses per time span was performed to rule out this explanation. The result confirmed our findings regarding the negative relationship between traffic count and odds for drink driving. Note that a more sophisticated way to investigate this relationship is by extending the two-level model to a-three level model by including the variable time as an extra level. Locations would then be at level 3, time at level 2 and drivers at level 1.

The odds of drink driving for women in comparison with men (*Female*) are multiplied by a factor of 0.253, meaning that women's odds for drink driving decrease significantly (joint chi square test=44.123, df=1, p=0.000) by 74.7% compared to men.

The reference category for the variable Age is the category of drivers in the age group 16-25. The odds of drink driving for drivers with an age in the range 26-39 in comparison with the reference category are multiplied by 2.034. This means that drivers with an age in the range 26-39 have 103.4% more chance to be a drink driver than drivers with an age in the range of 16-25. The odds of drink driving for drivers with an age in the interval 40-54 in comparison with the reference category are multiplied by 3.721 and thus those odds increase by 272.1%. Finally, the odds of drivers aged 55 or older in comparison with the reference category are multiplied by a factor of 2.370; those odds increase by 137.0%. This relationship between age and the dependent variable is also significant (joint chi square test=38.666, df=3, p=0.000).

The odds of drink driving for drivers who previously have been stopped and tested at a road site at least once in comparison with drivers who have never been stopped and tested (*Previously*) are multiplied by a factor of 1.505. This means that the former drivers have a 50.5% higher risk for drink driving than the latter drivers. This relationship was also found to be significant (joint chi square test=8.476, df=1, p=0.004).

This result seems to be in contradiction with the SORC-model, explained in the GADGET-project, stating that past experiences with law enforcement – as one aspect of the objective risk of getting caught – lead to obedience (Christ et al., 1999). It can, however, be explained by the selective way in which police checks in general are carried out in Belgium. Normally police officers focus on drivers who are more likely to be drink driving based on observable criteria like gender. This eventually results in a population of drivers consisting of drink drivers who, relatively speaking, have been tested for drink driving more often than the non-drinking drivers. The evidence we found in this roadside survey is based on a random sampling mechanism that allocates equal probabilities for selection to drink drivers and non-drinking drivers, reflecting the result of the selective way in which police checks are carried out in general. This rationale is of course conditional on the assumption that drink drivers in general are recidivists who will continue to drink drive even if they have been caught and sentenced before. In other words, the explanation for the evidence we found could simply be the nature of the group of drink drivers which might be composed for the largest part by hard core drink drivers (Simpson et al., 2004) for whom this SORC-model does not hold.

The reference category for the following variable (*Probability*) is the category of drivers who answered that they perceive the probability of being tested to be very low. The relationship as a whole is significant (joint chi square test=36.378, df=4, p=0.000). The odds of drink driving for drivers who answered they perceive the probability of being tested as low in comparison with the reference category are multiplied by a factor of 1.711, meaning the odds of drink driving increase by 71.1% compared to the reference category. The odds of those who answered they perceive the probability of being tested medium in comparison with the reference category are multiplied by a factor of 2.104, so the odds increase by 110.4% compared to the reference category. The odds of those drivers who answered they perceive the probability of being tested high in comparison with the reference category are multiplied by a factor of 1.366 and thus are 36.6% higher than the reference category's odds (but this dummy variable is not significant). Finally, the odds of drink driving of those drivers who answered they perceive the probability of being tested as very high compared to the reference category are multiplied by a factor of 4.187; in other words, those odds increase by 318.7%.

Parameter	Extra binomial model		Binomial model	
	Logit coefficients (s.e.)	Exponential coefficients	Logit coefficients (s.e.)	Exponential coefficients
Fixed				

Intercept	-4.981 (0.265)		-4.757 (0.285)	
Traffic count	-0.001 (0.000)	0.999	-0.002 (0.000)	0.998
Intensity	0.746 (0.407)	2.109	0.896 (0.383)	2.450
Female	-1.395 (0.177)	0.248	-1.375 (0.207)	0.253
Previously	0.467 (0.126)	1.595	0.409 (0.141)	1.505
Probability low	0.565 (0.144)	1.759	0.537 (0.167)	1.711
Probability medium	0.769 (0.146)	2.158	0.744 (0.169)	2.104
Probability high	0.304 (0.239)	1.355	0.312 (0.278)	1.366
Probability very high	1.445 (0.254)	4.242	1.432 (0.290)	4.187
Age26-39	0.749 (0.206)	2.115	0.710 (0.242)	2.034
Age40-54	1.382 (0.200)	3.983	1.314 (0.234)	3.721
Age55+	0.948 (0.233)	2.581	0.863 (0.272)	2.370
Random				
Level 2 variance: σ_u	1.569 (0.229)		0.991 (0.197)	
Level 1 variance: σ_e	0.712 (0.010)		1.000 (0.000)	

Table 2.3.4: Logit and Exponential coefficients for the fixed and random effects of the extra binomial and the binomial 2 level multilevel logistic model (significant coefficients are printed in italic)

Based on the SORC model (Christ et al., 1999), mentioned above, one would expect the opposite. A possible explanation is that the perception of drivers who are caught on the spot is influenced by this event. An alternative explanation could be related to a selective memory bias for alcohol cues (Franken et al., 2003).

To summarise, it was shown in the model fit section that the model fits the data well and that the data called for a multilevel approach. The results of the multilevel models revealed an interesting relationship between traffic count and odds for drink driving indicating that drink drivers tend to avoid places with higher traffic counts. In practice this means that police officers should not restrict their enforcement activities to sites where the frequency of vehicle traffic is high. The results for gender and age are in line with previous findings: women are less at risk for drink driving, as are the youngest drivers aged 16-25 (Vanlaar, 2002). Finally it was demonstrated that, in contradiction with the SORC model, drivers who have been controlled previously and/or perceive the probability of being controlled for alcohol are particularly prone to drink driving.

2.3.2.7. Conclusion

A multilevel version of logistic regression analysis was presented. Transforming coefficients of the fixed effects of such a model into easy-to-interpret odds ratios was demonstrated. Differences between the binomial and the extra-binomial model were discussed and it was illustrated how to interpret these differences appropriately.

2.3.3 Multinomial responses

Emmanuelle Dupont and Heike Martensen (IBSR)

The response variable to be modelled can be made of several categories (i.e., two or more). In this case, it is assumed that the response follows the multinomial distribution. The analyses for these data can be considered as an extension of binomial data analysis: What is being modelled is the probability of the observations falling into each of the response category[21]. Contrary to the binomial analysis, however, more than 2 possible responses must be considered altogether. It is important – in order to properly perform the analysis – to distinguish between cases where these categories are related by some meaningful order, and cases where they can not be ordered so. The first case requires the application of "ordered" category analysis (also called ordered proportional odds analysis), the other an unordered model, sometimes simply termed a "multinomial analysis". In order to highlight the statistical implications of conceiving response categories as ordered or not, the models' objectives, definitions and assumptions will be developed in parallel for ordered and unordered responses models.

The Belgian drink-driving study presented in Section 2.3.2 will also be used as a research example in this section. In the present case, however, the drink-driving response variable will be handled as it had been recorded, namely, as made up of 3 categories ("safe", "alarm", and "positive"). Models will be fitted first assuming that there is a meaningful order underlying category numbers 1 to 3, then without making this assumption.

2.3.3.1. Objectives of the technique

The primary aim of the analysis of multinomial responses data is to model the probability of y_{ijk} - the observation for individual *j* belonging to group *k* - to fall into one of the various categories (the i's) making up the response variable. This probability itself is represented as a function of one or more explanatory variables. In its multilevel version, such an analysis additionally allows examining whether these probabilities - and the way they are influenced by the predictors – vary as a function of higher-level units.

2.3.3.2. Model definition

Applying multilevel techniques to multinomial responses implies that the model's lowest level will serve essentially pragmatic purposes, namely the specification of the *structure* of the response variable. Therefore, even a model accounting for single-level data will take a 2-level form. Level 1 will be made of several dummy variables (as many as the total number of categories minus one, designated as the reference). Each of these variables will take on the value "1" when a given observation corresponds to the category it figures in, "0" otherwise. Level 2 will represent the individuals sampled in the study, or

[21] Another option consists of modelling the frequencies or counts of the responses in each of the response category as the response variable, therefore using the Poisson distribution as sampling model at level 1 (see Dobson, 2001; or Goldstein, 2003 for details on this option).

whatever units the observations are made on. To each level 2 unit will thus corresponds a set of dummy values, among which only one "1". In this framework, level 1 does not in itself consist of observations, but rather defines their structure. Therefore, this is at the second level that the lowest-level units are to be found. This device is thus similar to the one applied in the case of repeated measurements (see Section 2.4), or of multivariate multilevel analysis (see Section 2.5): Level 1 in all these models establishes the "measurement model" (Raudenbush & Bryk, 2002). Examining 2-level versions of these data, therefore, requires 3 levels to be included in the model.

The general principles outlined when describing the Multilevel Generalised Linear Model (see section 2.3.1) are applicable to the modelling of multinomial responses data: The distributional features of the data to be modelled are incorporated into the level 1 sampling model, allowing to define the response variable at this level as the result of some particular expected values (probabilities in this case, as in the case of binomial data). This expected value is in turned modelled as being some function of a linear arrangement of predictors, among which random variation at higher levels of the data hierarchy (bearing in mind that level-2 in the present case actually corresponds to the lowest level, the one of the observations).

As indicated earlier, although the models fitted for ordered or non-ordered category data lie upon the same multilevel structure (level 1 defining the structure of the response, level 2 corresponding to the observations), they also substantially differ from each other. In both cases the logit link is used, the odds of two probabilities. The two probabilities that are contrasted, however, differ. The log odds in the unordered response categories contrast the simple probabilities with those of a reference category. In contrast, in the case of ordered categories, it is a cumulative probability that is contrasted with a reference category. The link function for the ordered proportional odds model is therefore called 'cumulative logit'. The use of the cumulative logit link, as explained below is what allows the model to preserve the ordered nature of the categories.

2.3.3.2.1. Ordered categories

The response variable y_{ijk} represents the test result of the j^{th} driver at the k^{th} road site as belonging to one of these three categories: 1 = "Safe", 2 = "Alarm", 3 = "Positive". Considering that a meaningful order underlies these test results, one could conceive of them as reflecting some unobservable dimension – say "z". In the present example, this dimension would be the blood alcohol concentration (BAC) [22]: The higher an individual's stand on this underlying dimension, the higher the probability that this individual's test result will fall into the upper categories of the response variable.

The *cumulative logit* link that is used in ordered models is based on *cumulative probabilities*. Because of the ordered nature of the categories, it makes sense to calculate for each category the probability of an observation falling *into that*

[22] "Safe" corresponds to a blood alcohol concentration below 0.22 mg/l, "alarm", to a BAC between 0.22 mg/l and 0.35 mg/l, and "positive" to a BAC exceeding 0.35 mg/l.

category or above. Throughout this section, the notation "γ_{ijk}" will be used to refer to *cumulative* probabilities, while "π_{ijk}" will be employed to designate "ordinary" probabilities. Formally, cumulative probabilities are defined as:

$$\gamma_{ijk} = \text{Prob}(y_{ijk} \geq i) = \sum_{i}^{I} \pi_{ijk} \,, \qquad \textit{(2.3.16a)}$$

where i is the rank of the response category in question. Thus, for a response made up of three categories, we have:

$$\gamma_{1jk} = \pi_{1jk} + \pi_{2jk} + \pi_{3jk} = 1 \qquad \textit{(2.3.16b)}$$

$$\gamma_{2jk} = \pi_{2jk} + \pi_{3jk} \qquad \textit{(2.3.16c)}$$

$$\gamma_{3jk} = \pi_{3jk} \qquad \textit{(2.3.16d)}$$

Given that $\gamma_1 = 1$, only $I - 1$ (with I being the total number of categories) cumulative probabilities will have to be estimated.

In the example of the alcohol-breathtest the reference category is the lowest one (BAC<.05: safe). The cumulative probability γ_i denotes the probability to have a BAC that defines category in question *or more*. Thus, the cumulative probability for the category "alarm" is the sum of the proportion in "alarm" *and* in "positive" because drivers in both categories have a BAC of .05 or more. The cumulative probability for "positive" is simply the proportion of this category, because there is no category defined by an even higher BAC. Finally, the cumulative probability for the category "safe" is 1, because everybody has a BAC <.05 or more. Figure 2.3.1 provides an illustration of the cumulative probabilities in the case of a 3-categories response.

Cumulative logits are the ratio of the probability of one observation falling into the reference category *or* above ($\Pr(y_{ijk} \geq i)$) to the probability of the observation falling in a lower category ($\Pr(y_{ijk} \prec i)$). This odds ratio is formally defined as $\log\left(\frac{\gamma_{ijk}}{1-\gamma_{ijk}}\right)$.

The model's systematic component can now completely be defined as:

$$\eta_{ijk} = \log\left(\frac{\gamma_{ijk}}{1-\gamma_{ijk}}\right) = \log\left(\frac{\Pr(y_{ijk} \geq i)}{\Pr(y_{ijk} > i)}\right) = \beta_{oj(i)} + \sum_{h=1}^{p} \beta_{hj} x_{hjk} + u_{ok} + \sum_{l=1}^{q} u_{lk} x_{ljk} \qquad \textit{(2.3.17)}$$

The first part of this component, namely, the link function, has already been explained. The second part of the equation indicates that the predicted

cumulative log-odds are expected to be a function of some fixed (population) intercept value ($\beta_{oj(i)}$), of a random effect of the level-3 units on this intercept value (u_{ok}), of fixed effects of explanatory variables ($\sum_{h=1}^{p} \beta_{hj} x_{hjk}$), and of random variation of these effects across level-3 units ($\sum_{l=1}^{q} u_{lk} x_{ljk}$).

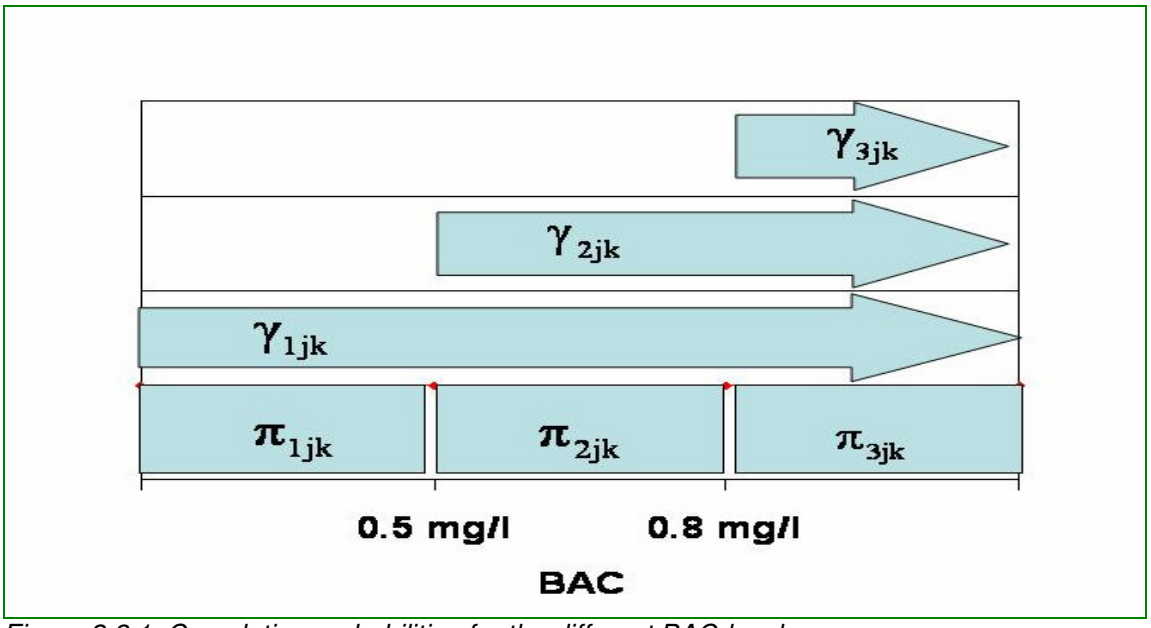


Figure 2.3.1: Cumulative probabilities for the different BAC-levels

The notation used in Equation 2.3.17 indicates that each response category has a different intercept value (β_0 is the only term of the model to which the *i* subscript for categories is assigned). These intercepts, or "thresholds" for the response categories must be understood as the average cumulative log-odds for each category. They can thus be interpreted as the ratio of the probability of an observation falling into category *i* or above to the probability of the observation falling into a lower category[23], when all predictors are set to 0. $\beta_{oj(1)}$ corresponds to the log-odds of being in category one rather than in category 2, or 3, $\beta_{oj(2)}$ corresponds to the log-odds of being in category 1 or 2 rather than in category 3. This series of intercepts accounts for the order of proportional odds, and is what confers the model its cumulative nature (Leyland & Goldstein, 2001). They correspond to predicted log-odds which, once transformed into predicted probabilities, can be interpreted as the probability for a given observation to appear in category *m* before appearing in category *i* (Ibid).

The model thus specifies different intercepts for the response categories, but provides one and only one estimate for the random variation of these intercepts

[23] This interpretation holds because the category chosen as the reference is the first one. This interpretation is to be "reversed", however, when the last category is the reference.

at level 3 (or higher). Indeed, allowing them to vary randomly at level 3, but all in a different way would render the interpretation of the results quite difficult, and the model rather costly to estimate (many parameters would have to be estimated). *All* the components of the proportional model – except the intercepts - are defined as being common to the different response categories. This is the case not only for the level-3 random effects, but for any fixed effect specified in the model. This feature - separate intercepts but common slopes to all categories – reflects a fundamental assumption of the proportional model: All the effects (both fixed and random) are assumed to be independent from the particular category considered. In the same way that the fixed effects of individual-level predictors are defined as being homogeneous across response categories, it can be assumed that the random effects related to the observations' clustering into higher-level units is homogeneous across response categories[24].

Consequently, an empty, single-level proportional model will appear as a series of $I-1$ equations, but these would differ from each other only on the ground of the fixed intercept values. The possible predictors do not vary across categories, and do not take the category index *i,* indicating that they are estimated for all categories jointly.

2.3.3.2.2. Unordered categories

Supposing that the categories making up "breathtest", the response variable of the drink-driving study are not ordered, one would model it as a function of the probability of each category of result rather than on cumulative probabilities. One of the appropriate link functions would in this case be the "usual" logit function rather than the cumulative logit. The main difference between this model - the "unordered" model – and the ordered one is that the former does not assume homogeneous predictor effects across categories. The explanatory variables entered in the model are seen as likely to have a different effect on the different response categories.

As a reminder, the logit link function corresponds to the log of the odds of being in one given category (*i*) rather than in another, designated as the reference category (*m*). The odds themselves are defined as the ratio of the probability of being into category *i* to the probability of being in category *I*. These probabilities are defined as:

$$\pi_{ijk} = \text{Prob}(y_{ij_k} = i) = \pi_{ijk} \text{ , for } i = 1,..., I \qquad (2.3.18a)$$

[24] Fitting an unordered model to verify that the predictors' effects indeed are homogeneous across categories is useful. However, failure to meet the "proportional odds assumption" and concluding that predictors do have different effects across the response categories does not necessarily imply that the ordered nature of the categories has to be questioned, but simply that the effects of predictors are not homogeneous across the different response categories (Raudenbush & Bryk, 2002). In such a case, it is nevertheless necessary to treat the latter as unordered, and to use the multinomial model.

And,

$$\pi_{1jk} = \pi_{1jk} \qquad (2.3.18b)$$

$$\pi_{2jk} = \pi_{2jk} \qquad (2.3.18c)$$

$$\pi_{3jk} = 1 - \pi_{1jk} - \pi_{2jk} \qquad (2.3.18d)$$

Given that $\pi_{3jk} + \pi_{2jk} + \pi_{1jk} = 1$, only $I - 1$ probabilities will have to be estimated.

The systematic component of the unordered model is:

$$\eta_{ijk} = \log\left(\frac{\pi_{ijk}}{\pi_{Ijk}}\right) = \log\left(\frac{\Pr(y_{ijk} = i)}{\Pr(y_{ijk} = m)}\right) = \eta_{ijk} = \beta_{o(i)} + \sum_{h=1}^{pi} \beta_{hjk(i)} x_{hjk} + u_{ok(i)} + \sum_{l=1}^{qi} u_{ljk(i)} x_{ijk} \qquad (2.3.19)$$

The *i* category index is assigned to all components of this equation, implying that there is one separate model for each response category. For this reason, the unordered model can be described as "contrast-specific": It is made of several "sub-models" that compare each response category to the reference one (Rasbah, Steele, Browne, & Posser, 2004). This allows for much flexibility in the way the predictor-probability relationship is specified across the categories: The effect of one given predictor may differ depending on the particular categories contrasted, and so may the random effects at higher levels of the data hierarchy. The unordered model also allows specifying different predictors for different response categories (e.g.: the sub-model contrasting category 1 with the reference would contain predictor x and the sub-model contrasting category 2 with the reference would contain predictor z). To summarize, in contrast to the ordered model, the unordered model conceives of the effects of the predictors (both fixed and random) and of the category probabilities as *interactive* effects. This is the major difference between this model and the proportional odds one, which assumes them to be independent and additive. For this reason, the proportional model is also far more parsimonious than the unordered model (i.e.: the number of parameters to be estimated is greatly reduced in the case of the ordered model).

2.3.3.3. Model assumptions

As already indicated, there is no variance associated with level 1 in a model for multinomial responses. It is at the second level of the data hierarchy (the "individual" level) that the variance specified by the sampling model describing y_{ijk} is to be found.

Level 2 thus describes the inter-individual variation in the data, and the error structure of the response. In the case of the ordered model – working with cumulative probabilities – the variance and covariances of the observations are defined by the ordered multinomial sampling model:

$$Var(y_{ijk}\left|\gamma_{ijk}\right.) = \gamma_{ijk}(1-\gamma_{ijk}) \quad (2.3.20)$$

$$Cov(y_{ijk}, y_{i'jk}\left|\gamma_{ijk},\gamma_{i'jk}\right.) = \gamma_{ijk}(1-\gamma_{i'jk}) \quad (2.3.21)$$

In the case of the unordered multinomial model, the variance and covariances of the observations should be:

$$Var(y_{ijk}\left|\pi_{ijk}\right.) = \pi_{ijk}(1-\pi_{ijk}) \quad (2.3.22)$$

$$Cov(y_{ijk}, y_{i'jk}\left|\pi_{ijk},\pi_{i'jk}\right.) = -\pi_{ijk}(1-\pi_{i'ijk}) \quad (2.3.23)$$

The fact that the observed variance and covariances at level 2 follow the specification of the multinomial sampling model of course depends on whether or not the data indeed exactly follow the multinomial distribution, and do not show over - or under – dispersion (see section 3.1.2).[25]

In the case of a 3-level random intercept model (conceptually a 2-level model, thus), these are the variance and covariances of logit values that compose the variance-covariance structure at level 3. Both the ordered and unordered models assume the intercepts and slopes of the level-3 units to be normally distributed, with mean 0 and variance σ^2. In other words, the (cumulative) logits are assumed to be normally distributed around the level-3 units.

The variance-covariance structure at level 3 also depends on the (un)ordered nature of the model. Indeed, the ordered model assumes random variation at this level to be the same for all response categories, while the random effects are expected to differ across response categories in the case of the unordered model. In this latter case, given that the random effects are allowed to differ for the various categories, the covariance between the categories' random effects also has to be estimated as part of the variance-covariances structure, in addition to the usual covariance between intercepts and slopes.

2.3.3.4. Research problem

The response variable that is assessed here ("breathtest") is conceptually the same as the one examined in the section over binomial data (Section 2.3.2.). In the binomial case, however, observations were indiscriminately treated as the same indication that a driver had been drinking when his/her breathtest result exceeded the "alarm" or "positive" thresholds. These two categories were indeed merged in order to constitute a single "success result", so that the response could be considered dichotomous. In the present analysis the two categories will be treated distinctly, and the response variable will thus be

[25] This assumption can be checked through estimating an additional parameter, known as the "scale factor" that is associated to the "canonical" variance defined by the sampling model of the discrete distribution concerned. This parameter should appear close to one if indeed the data closely follow the discrete distribution described. Values lower than 1 reflect a situation termed "underdispersion", values greater than 1 indicate "overdispersion".

analysed as a "3-categories" response. The effect of gender and age on the drivers' test results will be examined, first by means of the proportional logit model, then by means of the unordered model. This last analysis will provide useful indications about whether the predictors included in the model indeed can be considered to have homogeneous effects across the different response categories.

2.3.3.5. Dataset

The dataset used for the present illustration is the same as the one analysed in Section 2.3.2. The reader is therefore referred to this section for a complete description.

2.3.3.6. Model fit and diagnostic

2.3.3.6.1. A word of caution on estimation methods

As already mentioned in the introductory section over Multilevel Generalised Linear Models, when the estimation method employed in this framework consists of quasi-likelihood estimation (i.e., the approximation of maximum likelihood estimation via linearization) rather than of maximum likelihood itself, the deviance test cannot be trusted any more. Consequently, there is no criterion available to gauge the improvement of the models successively specified. Tests of single parameters remain one option, but should also be used with caution, at least for random parameters.

2.3.3.6.2. Ordered models

The first type of model fitted - the proportional logit model - specifies the following sampling model for the response variable at level 1 ("the test result i of individual j") :

$$y_{ij} \sim \textit{Ordered multinomial}\,(n_{ij},\ \gamma_{ij}) \qquad \textit{(2.3.24a)}$$

The expected value for the response variable is the cumulative probability, γ_{ij}, and is therefore the value that will have to be modelled on the joint basis of the linear predictor and the cumulative logit link function. The predicted cumulative probability γ_{ij} is defined the following way for each of the i response category:

$$\gamma_{3j} = \pi_{3j}, \gamma_{2j} = \pi_{3j} + \pi_{2j}, \gamma_{1j} = 1 \qquad \textit{(2.3.24b)}$$

The first category ("Safe") is designated as the reference, and thus has cumulative probability $\gamma_{1j} = 1$. The lowest category being the reference, the other 2 cumulative probabilities must be understood as "the probability that an observation falls into the next higher category or above" (namely, in the "alarm or positive" categories for γ_{2j} and in the "positive" category for γ_{3j}) [26].

[26] Had the last category – « positive » - been designated as the reference, the cumulative probabilities would have been defined the other way around, with $\gamma_{3j} = 1$, and γ_{2j} and γ_{1j}

As an initial step, the model is left empty. It does not include any predictor. Two levels are specified ("i", and "j"), but one should bear in mind that such a model conceptually is a single-level model. The systematic component is at this stage composed of an intercept only, and fits the "baseline" predicted cumulative logits of the response categories[27]. Given that the response variable is made up of three categories, only two cumulative logits will be estimated:

$$\text{logit}(\gamma_{2j}) = -3.49(0.06) \qquad \textit{(2.3.24c)}$$

$$\text{logit}(\gamma_{3j}) = -3.91(0.07) \qquad \textit{(2.3.24d)}$$

Once exponentiated, these values can be interpreted as the "baseline" probability for an observation to correspond to an "alarm" or "positive" test result (γ_{2j}), or to a "positive" result (γ_{3j}):

$$\gamma_{2j} = \frac{1}{1+\exp-(-3.49)} = 0.03 \qquad \textit{(2.3.24e)}$$

$$\gamma_{3j} = \frac{1}{1+\exp-(-3.91)} = 0.02 \qquad \textit{(2.3.24f)}$$

The results from the empty model indicate that the predicted probability for a driver to have a test result of "alarm" or "positive" is 0.03, and does not differ much from the probability for a driver to obtain a "positive" result to the test rather than a "safe" or an "alarm" one. Both coefficients are negative and significant, indicating that drivers are on average more likely to be tested as "safe" rather than as "positive" or "alarm". The ordered model assumes all effects, save the fixed intercepts, to be homogeneous across the response categories. In order to reflect this assumption, a single term "h_{jk}" is added to each of the cumulative logit's equation, which will contain any effect that will be further specified in the model and remain identical for all the response categories.

The next model fitted is the random intercept model. This model will allow assessing whether "road sites", the 2nd conceptual level in the data hierarchy (but the 3rd level in terms of the present model), introduces random variation in

being the cumulative probabilities of the observations falling in the next lower category or below , and $\gamma_{1j} = \pi_{1j}$, given that there exists no lower category in the response.

[27] As a reminder, the logit of a probability (whatever cumulative or not) has to be understood as the log of the ratio of 2 probabilities (log-odds). In the present model, logit(γ_{2j}) corresponds to the log of the odds of the probability of being into category 2 or above as compared to category 1, while logit γ_{3j} corresponds to the odds of the probability of being in category 3 as compared to category 2 or below.

the cumulative logit of the intercept probabilities for the response's categories. The random intercept model is the first one for which we use the h_{jk} which in this case contains only the level-3 variance of each of the categories' intercepts ($\text{logit}(\gamma_{2jk})$ and $\text{logit}(\gamma_{3jk})$): These are expected to vary across higher-level units in the same way for the two response categories. :

$$\text{logit}(\gamma_{2jk}) = -3.37(0.08) + h_{jk} \qquad \textit{(2.3.25a)}$$

$$\text{logit}(\gamma_{3jk}) = -3.79(0.08) + h_{jk} \qquad \textit{(2.3.25b)}$$

And, $$h_{jk} = v_{3k}\text{cons}.23\,, \qquad \textit{(2.3.25c)}$$

Multiplying the random variation of the intercepts by a constant is what allows it being common to both cumulative logits. This random variation is also defined so as not to have any fixed part. The fixed – or population – values that are estimated for the intercepts of each category are to be considered as the fixed counterpart of this random variation. The random variation of the intercept is defined as:

$$[v_{3k}] \sim N(0, \Omega_v) : \Omega_v = [0.88(0.18)] \qquad \textit{(2.3.25d)}$$

The results obtained at this step seem to indicate that there is significant random variation in the probabilities to drink and drive across road sites. The reader is reminded, however, that the estimates obtained for the random parameters can in this case be severely biased.

The effect of gender is then added to the model. The gender predictor is defined as a dummy variable, with the "0" (and thus, the reference) value corresponding to "men" and "1", to "women". Adding this effect to the model results in somewhat lower values for the intercepts of the two cumulative probabilities:

$$\text{logit}(\gamma_{2jk}) = -3.07(0.08) + h_{jk} \qquad \textit{(2.3.26a)}$$

$$\text{logit}(\gamma_{3jk}) = -3.49(0.08) + h_{jk} \qquad \textit{(2.3.26b)}$$

The two intercept values now correspond to the predicted logits of being tested as alarm or as positive *among men*: Overall, these are less likely to be tested as "positive" or "alarm" than as "safe". The predicted probabilities for men drivers to be tested as either "alarm" or "positive", and to be tested as "positive" are 0.04 and 0.03, respectively. The coefficient for women is now added to the h_{jk} component, and expresses the change entailed in these predicted logits by being a woman rather than a man: Clearly, the odds of being tested as "alarm" or "positive" rather than as "safe" are even lower among women than among men. And,

$$h_{jk} = -1.61(0.23)\text{women}.23_{jk} + v_{3k}\text{cons}.23\,, \qquad \textit{(2.3.26c)}$$

The predicted probabilities for women can also be obtained using the exponential function:

$$\gamma_{2jk} = \frac{1}{1+\exp-(-3.07-1.61)} = 0.009 \qquad (2.3.26d)$$

$$\gamma_{3jk} = \frac{1}{1+\exp-(-3.49-1.61)} = 0.006 \qquad (2.3.26e)$$

These are the predicted probabilities of a female driver to be tested as « alarm » or « positive » or as "positive" rather than "safe", respectively. Obviously, they are both much lower than those obtained for males.

The estimate for the random variation of the intercepts at level 3 remains the same as the one calculated on the basis of the empty model:

$$[v_{3k}] \sim N(0, \Omega_v) : \Omega_v = [0.89(0.18)] \qquad (2.3.26f)$$

An additional model is then fitted to add another categorical predictor – age – to the model. The age variable is made up of 4 categories (16-25; 26-39; 40-54;

	Male		*Female*	
	Alarm or positive	*Positive*	*Alarm or positive*	*Positive*
Age				
- 16-25	0.02	0.01	0.004	0.003
- 26-39	0.04	0.03	0.008	0.005
- 40-54	0.07	0.05	0.01	0.01
- 55+	0.04	0.03	0.008	0.005

Table 2.3.5: Cumulative predicted probabilities for the different age and gender categories.

and 55+), with the first being designated as the reference. This predictor is introduced in the model by means of three dummies, each comparing one of the remaining age categories with the "youngest" one. The intercept values for the cumulative logits are:

$$\text{logit}(\gamma_{2jk}) = -3.76(0.23) + h_{jk} \qquad (2.3.27a)$$

$$\text{logit}(\gamma_{3jk}) = -4.19(0.23) + h_{jk} \qquad (2.3.27b)$$

These logits values are now the ones estimated for male drivers aged 16 to 25. The corresponding predicted probabilities are, 0.02 and 0.01 γ_{2j} and γ_{3j}, respectively.

The h_{jk} component now includes the coefficients associated with the three dummies representing the age variable:

$$h_{jk} = -1.61(0.23)\text{women}.23_{jk} + 0.51(0.2)26-39.23_{jk} + 1.16(0.24)40-54.23_{jk} + 0.53(0.28)55^{+}.23_{jk} + v_{3k}\text{cons}.23 \quad \textit{(2.3.27c)}$$

Overall, the three older age categories seem to be more likely than the 16-25 ones to exceed the legal BAC limit. This is especially true, however, for the 40-54 age range. Appropriately summing the coefficients and exponentiating them provides the corresponding predicted probabilities for all categories of the predictors in the model. For example, in the case of female drivers:

$$\gamma_{2jk} = \frac{1}{1+\exp-(-3.76-1.61)} = 0.005 \quad \textit{(2.3.27d)}$$

…corresponds to the probability of a woman aged 16 to 25 to be tested as "alarm" or "positive".. Computing:

$$\gamma_{2j} = \frac{1}{1+\exp-(-3.76-1.61+1.16)} = 0.01 \quad \textit{(2.3.27e)}$$

… provides the predicted probability of a female driver aged 40 to 54 to be tested as "alarm" or "positive". All the predicted probabilities for the different predictors categories are summarised in Table 2.3.6.

2.3.3.6.3. Unordered models:

Finally, an unordered version of the model including gender and age as predictors should allow examining whether their effects can indeed be considered homogeneous across the response categories. The sampling model at level 1 in this case is:

$$y_{ij} \sim \textit{Multinomial}\,(n_{jk},\ \pi_{ijk}) \quad \textit{(2.3.28a)}$$

The logit model is one for "ordinary" probabilities, and will readily be expressed as the log of the odds of each category to the reference category, which will in this case remain the "safe" one. Given that there are three categories, 2 logit models are estimated:

$$\log\left(\frac{\pi_{2jk}}{\pi_{1jk}}\right) \quad \textit{(2.3.28b)}$$

comparing the "alarm" and "safe" categories, and

$$\log\left(\frac{\pi_{3jk}}{\pi_{1jk}}\right) \quad \textit{(2.3.28c)}$$

comparing the “positive” and “safe” categories.

Contrary to the ordered model, no common term underlies the logit models for the different categories:

$$\log(\frac{\pi_{2jk}}{\pi_{1jk}}) = \beta_{0k} - 1.05(0.25)\text{Women.Alarm}_{ijk} + 0.77(0.34)26-39.\text{Alarm}_{ijk} + 1.02(0.34)40-54.\text{Alarm}_{ijk} + 0.71(0.38).55^{+}.\text{Alarm}_{ijk} \quad \textit{(2.3.28d)}$$

With

$$\beta_{Ok} = -4.93(0.31) + v_{ok} \quad \textit{(2.3.28e)}$$

$$\log(\frac{\pi_{3jk}}{\pi_{1jk}}) = \beta_{1k} - 1.61(0.22)\text{Women.Positive}_{ijk} + 0.52(0.25)26-39.\text{Positive}_{ijk} + 1.19(0.24)40-54.\text{Positive}_{ijk} + 0.59(0.27).55^{+}.\text{Positive}_{ijk} \quad \textit{(2.3.28f)}$$

With

$$\beta_{1k} = -4.16(0.23) + v_{1k} \quad \textit{(2.3.28g)}$$

The intercepts for the log-odds of “alarm” to “safe”, and of “positive” to “safe” are each defined as being made of different fixed values (β_0, β_1) and of different random components (v_{ok}, v_{1k}). As a consequence, the covariance between these two random intercepts is also part of the parameters estimated by the model:

$$\begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} \sigma^2_{v0} \\ \sigma_{v01}\sigma^2_{v1} \end{bmatrix} = \begin{bmatrix} 0.66(0.24) \\ 1.05(0.16)0.96(0.18) \end{bmatrix} \quad \textit{(2.3.28h)}$$

The fixed intercept values are both significant, and highly similar for the “alarm” and “positive” categories. Overall, the estimated effects are similar for the two categories. The effect of gender is in each case significant and negative: Women are less likely to be tested as “alarm” or “positive” than men are. The 40 to 54 age category is the one that differ most from the 16 to 25 one, the positive coefficient associated with this age category revealing that people of this age are more likely to exceed both legal limits (i.e.: “alarm” and “safe”). The estimates for the random variation of the intercepts at level 3 for the two categories ($\sigma^2_{v0}, \sigma^2_{v1}$) also do not differ much from each other. All in all, the results suggests that the two probabilities tend to be homogeneously affected by the fixed and random effects specified in the model, and consequently that

the unordered model is not worth the cost it entails in terms of the number of coefficients to be estimated. The covariance observed between the random variations of the two intercepts at level 3 is quite elevated, further sustaining this conclusion. This positive covariance suggests that, at those road sites at which the probability of being tested as “alarm” rather than as “safe” is higher, the probability of being tested as “positive” also tends to be higher.

2.3.3.7. Model interpretation

The results of the models fitted in the previous section converge to suggest that gender and age are important predictors of the likelihood of the drink-driving behaviour. The pattern of effects observed for these fixed predictors on the basis of the ordered and unordered models proved similar for the probability of exceeding the “alarm” legal limit and for the probability of exceeding the “positive” one. In such a case, the ordered model is certainly the one to be preferred, first because the assumption of ordered response categories appears highly sensible, but also because of this model’s parsimonious value. It can also be noted that the results agree with those found in the previous section (2.3.2), based upon a logistic regression analysis. The two categories “alarm” and “positive” analysed in the present section had been joined into one in the previous section. The unordered model analyses showed that the predictor effects for those two categories do not differ. Consequently, it is to be expected that dichotomising the response variable and analysing it in a logistic regression leads to comparable results, which is indeed what was found.

Several indications were obtained that the baseline or intercept probabilities for the “alarm” and “positive” responses vary randomly as a function of the road sites at which the tests were made. As it has already been stressed however, the present analysis allows few conclusions with respect to random effects.

2.3.3.8. Conclusion

Clearly, the present data call for a more complete multilevel analysis, one that would for example integrate level-3 effects such as the intensity of the traffic characterising the different road sites, or the time span during which tests have been performed at the different road sites. Cross-level interactions between these level-3 predictors and predictors at the individual level (e.g.: gender, age) are also potentially important aspects to address. Techniques permitting the exploitation of the multilevel structure of complex data are still under development, and the multilevel analysis of discrete data is certainly no exception. The available software keeps on evolving, and does so quickly. Before plainly satisfactory solutions can be offered, some alternatives can be used: Complementary information can be gained by relying on software directly using maximum likelihood estimation, although not specifically designed for multilevel analysis (such as “SAS”). Advanced estimation methods, such as MCMC (see Section 2.8) are also likely to provide valuable complementary information.

2.3.4 Counts

George Yannis, Eleonora Papadimitriou and Constantinos Antoniou (NTUA)

2.3.4.1. Objective of the technique

In this section, multilevel models that fit data with discrete response variables are further analysed. Following the analysis concerning binary or multinomial data shown in the previous sections, count data - or data that can take any positive integer value - are discussed. This count may be the number of times an event occurs out of a fixed number of "trials", in which case the resulting proportion is usually dealt with as response: an example is the proportion of fatalities in a population. It is common practise to use the Binomial distribution to fit models to proportional data, as shown in Section 2.3.2, and the Poisson - family distributions to fit models to count data.

The present analysis has the following objectives:

- Present the Poisson distributional assumptions and discuss the related properties and particularities
- Describe the related multilevel structure
- Use the above techniques to explore the regional effect of police enforcement on the number of road accidents in Greece.

2.3.4.2. Model definitions and assumptions

Count data have restrictions on the values they take; they must take positive integer values (or zero) and so if count responses were to be fitted as normal responses, one could obtain predicted counts that were negative. Consequently, the Poisson distribution is used instead (Langford et al., 1999). In this section, the basic Poisson assumptions for count data are presented.

The Poisson distribution has a parameter λ that represents the rate at which events occur in the underlying population, according to the following characteristic function:

$$P(x\,;\lambda)=\frac{\lambda^{x}e^{-\lambda}}{x!} \qquad (2.3.29)$$

The Poisson distribution is based on four assumptions. The term "interval" refers to either a time interval or an area, depending on the context of the problem.

- The probability of observing a single event over a small interval Δτ is approximately proportional to the size of that interval.
  $P(1\,;\Delta\tau) = \lambda\,\Delta\tau$ for small $\Delta\tau$
- The probability of two events occurring in the same narrow interval is negligible.
  $P(0\,;\Delta\tau) + P(1\,;\Delta\tau) = 1$ for small $\Delta\tau$

- The probability of an event within a certain interval does not change over different intervals.
- The probability of an event in one interval is independent of the probability of an event in any other non-overlapping interval.

These assumptions should be examined carefully, especially the last two. If either of these last two assumptions is violated, they can lead to extra variation, generally referred to as overdispersion, as discussed below (see also section 2.3.1).

Generally, modelling count data is known as Poisson regression and is not in itself a multilevel technique. To translate Poisson regression to multilevel Poisson regression is analogous to moving from linear modelling to normal response multilevel modelling (Langford et al, 1998, see also sections 2.1 and 2.2). In case of Poisson multilevel regression, there is a higher level classification of the data across which the response is considered to vary. The multilevel model fitted to the data is based on iterative generalized least squares estimation. Assuming multivariate normality, calculations alternate between estimation of fixed and random parameter vectors until convergence is reached. However, in this case, a Poisson distributed response vector (O) of observed cases is assumed, and hence it is necessary to include an offset of expected numbers of cases in the model, so that:

$$O_{ij} \sim Poisson\ (\pi_{ij}\, E_{ij})$$

$$log\ (\pi_{ij}) = \beta_{0j} + \beta_{1j}\, x_j \qquad (2.3.30)$$

$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

where E_{ij} represents the expected numbers of cases for each level 1 unit. When using such fixed offsets, it is recommended to centre them around their mean in order to avoid numerical instabilities (Rasbash et al., 2000).

The Poisson distribution is used to model the level 1 variance, by using a logarithmic link function, and normal distribution is assumed for the random variances at higher levels. An efficient estimation procedure for this nonlinear model is predictive quasi-likelihood, where estimation of random parameters and associated residuals, is made using a Taylor series expansion around the current values of the fixed and random parts of the model.

It should be underlined though that no random structure can be specified at the lowest level of a Poisson multilevel model. In particular, there is nothing random to estimate as in the Poisson model the relationship between mean and variance is known, so that there is no need to separately estimate the latter. However, the opposite is true in the classical linear regression model, where the mean of the error term is assumed equal to zero but the variance is unknown and must therefore be estimated. Consequently, one would be interested in making the intercept term vary randomly at the 1st level of a normal model but not at the 1st level of a Poisson model.

A basic additive model will have explanatory variables consisting of an intercept, and one or more dummy variables. One would normally also wish to include interactions between variables.

To determine whether the Poisson-assumption of equal means and variance holds, a dispersion parameter at level 1 is estimated, so that

$$var(O_{ij} / \pi_{ij}) = \sigma_1^2 \pi_{ij} E_{ij} \qquad (2.3.31)$$

If $\sigma_1^2=1$, then variation is assumed to be Poisson, if $\sigma_1^2>1$ then there is extra-Poisson variation present (overdispersion), and if $\sigma_1^2 <1$ the model is underdispersed as can happen when many of the counts are zero. However, quite often there are theoretical reasons to assume that extra-Poisson variation may be present in the data (Dean, 1992, Hauer, 2001). For instance, if the counts examined come from significantly heterogeneous populations, the expected values may vary significantly more than the mean of the distribution would allow.

In order to handle the overdispersion, one option is to include an additional parameter α, resulting in an extra - Poisson or quasi - Poisson distribution, so that:

$$var(O_{ij} /\pi_{ij}) = \alpha \sigma_1^2 \pi_{ij} E_{ij} \qquad (2.3.32)$$

This situation may be further described by stating that the counts in each level 1 unit are being modelled as Poisson conditional on the distribution of rates between units. These rates may be assumed to follow a gamma distribution, and hence the mixture of these two distributions can be expressed as a negative binomial distribution of counts, so that:

$$O_{ij} \sim Negative\ Binomial\ (\pi_{ij} E_{ij},\ v)$$

$$log\ (\pi_{ij}) = \beta_{0j} + \beta_{1j} x_j \qquad (2.3.33)$$

$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

where the variance is a quadratic function of π_{ij}:

$$var(O_{ij} /\pi_{ij}) = \pi_{ij} E_{ij} + (\pi_{ij} E_{ij})^2 / v = \sigma_1^2 \pi_{ij} E_{ij} + \sigma_2^2 (\pi_{ij} E_{ij})^2 \qquad (2.3.34)$$

It should be noted that, ignoring extra-Poisson variation would not significantly affect parameter estimates; however the related significances may be slightly affected (Dean, 1992).

2.3.4.3. Research problem and dataset

In 1998, the Greek Traffic Police started the intensification of road safety enforcement, having set as target the gradual increase of road controls for the two most important infringements: speeding and drinking-and-driving. Since then, all controls and related infringements recorded have systematically been monitored and the related enforcement and casualty results at local and national level are regularly published, as shown in Table 2.3.6 with basic road safety related trends in Greece.

	1998	1999	2000	2001	2002	5-year change
injury road accidents	24.819	24.231	23.127	19.710	16.852	-32%
persons killed	2.182	2.116	2.088	1.895	1.654	-24%
vehicles (x1000)	4.323	4.690	5.061	5.390	5.741	33%
speed infringements	92.122	97.947	175.075	316.451	418.421	354%
drink & drive infringements	13.996	17.665	30.507	49.464	48.947	250%
drink & drive controls	202.161	246.611	365.388	710.998	1.034.502	412%

Table 2.3.6: Basic road safety trends in Greece 1998-2002

It is important, however, to further quantify the effect of this intensification of enforcement on road accidents. Additionally, the examination of regional effects might be particularly interesting. For that purpose, a multilevel model is developed, as a different amount and type of police activity in regions with different characteristics is likely to result in different effects of enforcement. It should be noted that the administrative structure of the Greek police also follows the geographical (e.g. geopolitical) structure of Greece. As the number of accident represents a random count of events occurring within a population, a Poisson distribution is assumed.

The dataset that is used in the framework of this analysis concerns regional data from 50 counties of Greece (245 observations in total), nested within 12 regions in the period 1998-2002. The response variable is the number of road accidents with casualties and the explanatory variables are the number of alcohol controls, the number of speed infringements, as well as socioeconomic parameters such as vehicle ownership and road network type. The population of each county is used as offset term, to express the expected number of accidents. It should be noted that explanatory variables are centred around their mean, to avoid numerical problems in the estimation. The dataset variables are summarized in the following Table 2.3.7.

It should be noted that the Athens and Thessalonica metropolitan areas, where a disproportionably high number of accidents and police controls are observed, were not included in the dataset.

Region	1-12 regions of Greece
County	1-50 counties of Greece
Accs	The number of accidents of each county
alcontrol (1000)	The number of alcohol controls of each county
speedinf (1000)	The number of speed infringements of each county
logepop (offset)	The natural logarithm of the population of each county
Cons	The constant term

Table 2.3.7: Variables and values considered in the analysis

2.3.4.4. Model fit, diagnostics and interpretation of results

In the following sections, an application of multilevel Poisson models is presented. The analysis aims at examining the regional effect of speed and alcohol enforcement on the number of road accidents. It should be noted that the demonstration follows a stepwise procedure, both in terms of multilevel model building and variables selection. As far as model building is concerned, the analysis starts from the simplest (single level) model to the most complex (multilevel models). Accordingly, variables are initially examined separately (single-effects models), and then jointly (multiple-effects models).

The initial stage of the analysis concerns a single level model (level 1: i-county), ignoring the geographical hierarchy in the data. This approach gives the following results (Table 2.3.8):

Parameters	Single-level model
Constant	-6.450 (0.005)
Alcontrols	-0.015 (0.001)
Speedinf	-0.010 (0.001)

Table 2.3.8: Poisson single-level model for the effect of enforcement on road accidents

The coefficients of this initial model, all highly significant, as indicated by the respective standard errors in parentheses, indicate a reduction of road accidents when speeding and drinking-and-driving controls increase. This result is reasonable. However, in the following sections it will be demonstrated how this effect may vary significantly among regions.

The next stage is adding the hierarchical structure to the data, by including a second level (level 2: j-region). We first consider a two-level model with a random intercept term only, in order to examine the variation due to the regional effects. The results presented in Table 2.3.9 below indicate a significant random variance among regions (Model 1):

	Model 1 (constant term)	Model 2 (Effect of alcohol controls)	Model 3 (Effect of speed controls)	Model 4 (Effect of speed and alcohol controls)
Fixed effects				
Constant	-6.488 (0.076)	-6.672 (0.108)	-6.691 (0.115)	-6.654 (0.101)
Alcontrols		-0.059 (0.014)		-0.036 (0.010)
Speedinf			-0.131 (0.043)	-0.058 (0.023)
Random effects				
Level 2				
σ_{u0}^2 (constant)	0.070 (0.029)	0.140 (0.057)	0.157 (0.065)	0.119 (0.050)
σ_{u1}^2 (alcontrols)		0.002 (0.001)		0.001 (0.000)
σ_{u2}^2 (speedinf)			0.022 (0.009)	0.006 (0.002)
σ_{u01}^2 (covariance)		0.013 (0.006)		0.008 (0.004)
σ_{u02}^2 (covariance)			0.051 (0.023)	0.013 (0.009)
σ_{u12}^2 (covariance)				0.000 (0.000)
Variance/mean	1.000	1.000	1.000	1.000

Table 2.3.9: Poisson multilevel models for the regional effect of enforcement on road accidents

The significant regional variation of the intercept is presented in Figure 2.3.2 The top graph in Figure 2.3.2 concerns the average (fixed) intercept for all regions, whereas the bottom graph concerns the intercepts corresponding to each one of the 12 regions of Greece. It is noted that the x-axis concerns the number of alcohol controls (in thousands), centred around the mean. A significant regional variation of the number of accidents is illustrated.

The next step in model fitting with this dataset is to add explanatory (predictor) variables into the multilevel model. Firstly, the effect of alcohol controls on the number of accidents is examined, allowing it to randomly vary between regions. A multilevel model with a random intercept and a random slope is therefore fitted (Model 2) and the results are presented in Table 2.3.9.

It is noticed that all fixed and random effects are significant. However, the variance of the effect of alcohol controls is less significant than the variance of the intercept, suggesting that the regional effect itself (in geographical terms) is a stronger determinant of the number of accidents than the effect of enforcement. It is also interesting to note that there is a significant covariance among intercept and slope, indicating that, the higher the number of accidents of a region, the stronger the effect of alcohol enforcement (reduction of accidents).

It should be noted that, as mentioned previously, quasi-likelihood estimation is used for discrete response models. Consequently, likelihood statistics for these models are very approximate and are not examined for the assessment of models fit (Rasbash et al., 2000).

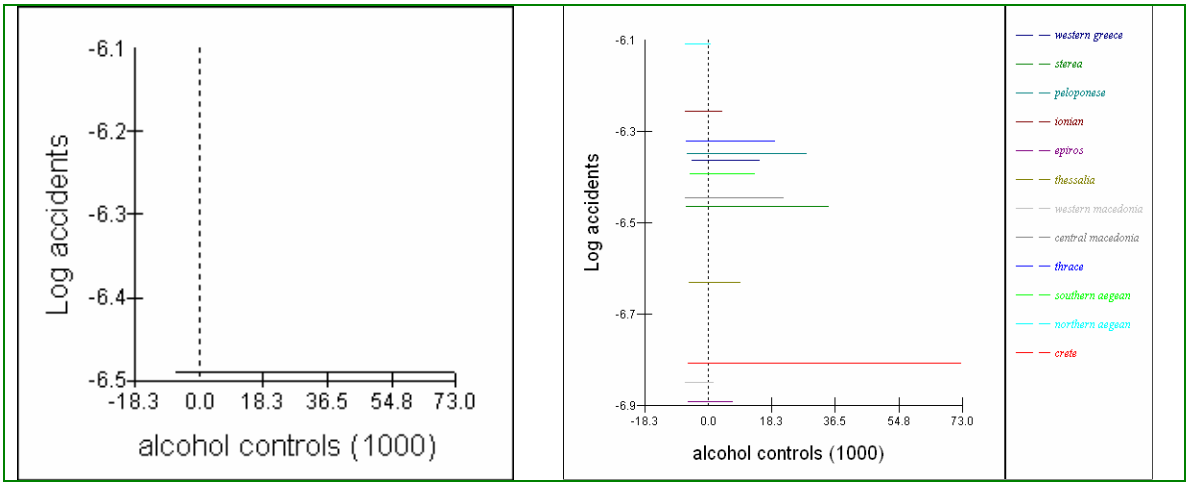


Figure 2.3.2: Average intercept (top graph) and random intercepts (bottom graph) for Model 1

The significant regional variation of the slope of alcohol controls is presented in Figure 2.3.3. The top graph in Figure 2.3.3 concerns the average (fixed) slope for all regions, whereas the bottom graph concerns the slopes corresponding to each one of the 12 regions of Greece. A significant effect of alcohol controls on the number of accidents at regional level is illustrated.

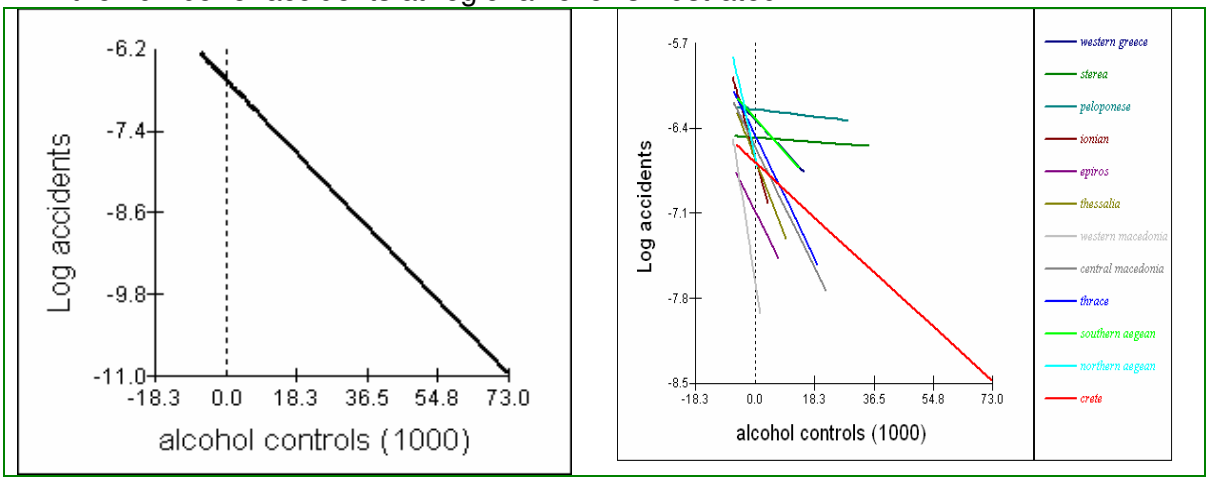


Figure 2.3.3: Average (top graph) and random (bottom graph) intercepts and slopes for Model 2 (effect of alcohol controls)

In Figure 2.3.4, the Level 1 and 2 residuals are examined for Model 2. In particular, the top graphs in Figure 2.3.4a concern Level 1 residuals and the four bottom graphs in Figure 2.3.4b concern Level 2 residuals. Moreover, the left-side graphs concern standardized residuals against normal scores and the

right-side graphs concern standardized residuals against fixed part predicted values.

It is observed that Level 1 residuals are normally distributed and independent. However, Level 2 residuals are less in keeping with the Normal distribution and present more dependence to the predicted values.

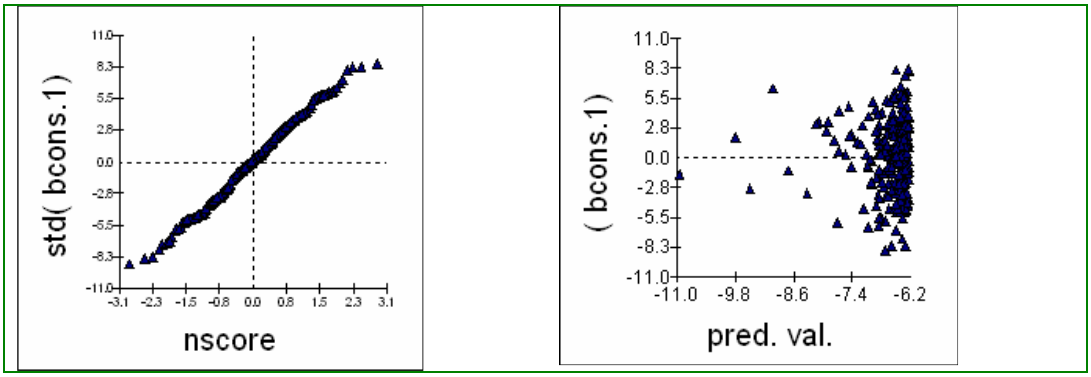


Figure 2.3.4a. Level 1 residuals and normal scores (left graph), Level 1 residuals and predicted values (right graph) for Model 2

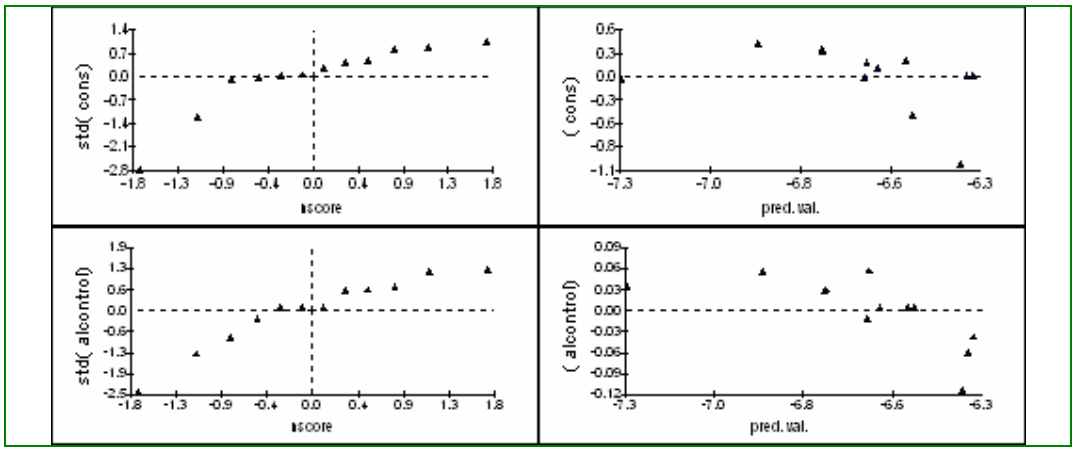


Figure 2.3.4b Level 2 residuals and normal scores (left graphs), Level 1 residuals and predicted values (right graphs) for Model 2

As a next step, the effect of *speed* enforcement on the number of accidents is examined separately. In parallel to the model including alcohol controls, the effect of the number of speed infringements is also allowed to randomly vary between regions. Another multilevel model with a random intercept and a random slope is therefore fitted (Model 3 in Table 2.3.9).

All fixed and random effects are again significant. Contrary to the effect of alcohol controls, the variance of the effect of speed infringements is, however, highly significant. There is also a significant covariance among intercept and slope, indicating that, the higher the number of accidents of a region, the higher the effect of speed enforcement. Although the variables 'alcontrols' and 'speedinf' are measured on the same scale, their parameter estimates are not directly comparable because the first one concerns number of controls and the second one concerns number of violations. In that sense, the fact that the

parameter for speed is higher can be explained by the fact that a given increase of violations results from a more important increase of related controls. Therefore, an equal increase of alcohol controls and speed violations corresponds to a higher increase of speed controls, making the effect of speed enforcement to appear more important, when expressed in number of violations.

The last stage of the analysis concerns the incorporation of both speed and alcohol enforcement effects in the model, in order to examine the related combined effect. A two-level model is therefore fitted (Model 4 in Table 2.3.9), allowing both explanatory variables to vary among regions. In this case, all fixed effects are highly significant, as well as the random variances. However, the covariances related to the number of speed infringements are non significant. This is quite surprising, when considering that both effects were significant when examined separately.

In Figure 2.3.5 the predicted intercepts and slopes of alcohol controls and speed infringements are plotted. It is noticed that the various regional effects differ significantly from the ones obtained previously, when effects were examined separately. Additionally, several slopes present an inversed effect, not directly attributable to regional characteristics.

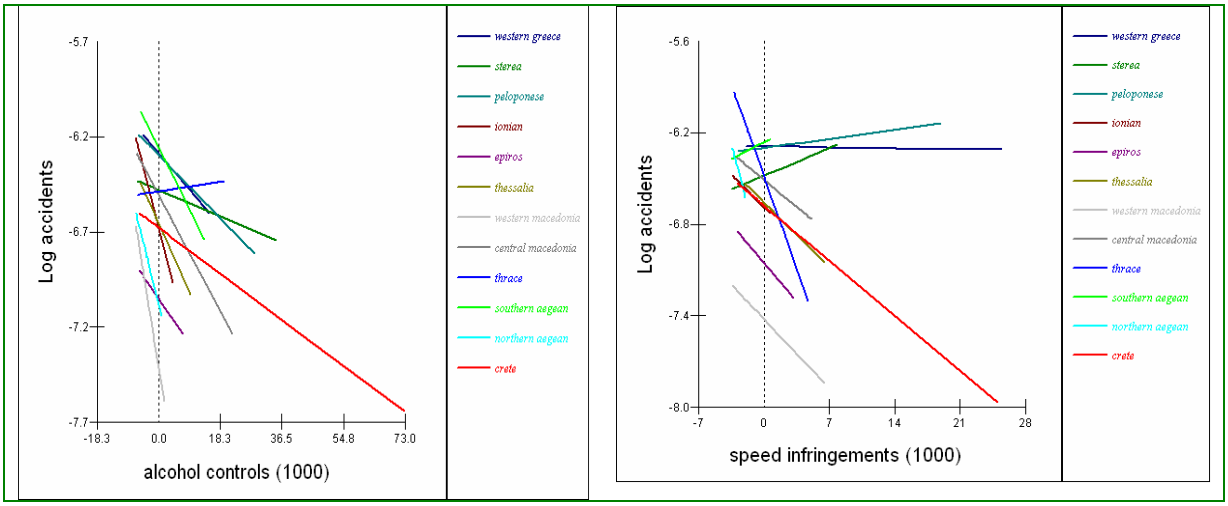


Figure 2.3.5. Random intercepts and slopes the effect of alcohol controls (top graph) and the effect of speed infringements (bottom graph) of Model 4.

This is probably due to the fact that both variables may be seen practically as measurements of one parameter (i.e. police enforcement). The correlation between speed infringements and alcohol controls was examined, resulting to a positive correlation of 0.729. In this case (multicollinearity), a redundancy of variables is exposed, causing both logical and statistical problems and weakening the analysis through reduction of degrees of freedom error (Washington et al. 2003). As far as multilevel models are concerned, the results of a recent study show that, with multicollinearity present at Level 1 of a two-

level multiple-effects linear model, the fixed-effect parameter estimates produce relatively unbiased values; however, the variance and covariance estimates produce downwardly biased values (Shieh, Fouladi, 2003).

Another issue that should be examined in case of Poisson multilevel models is overdispersion (Dean, Lawless, 1989). Overdispersion generally reflects missing parameters, not included in the model, which would account for the extra-variation.

A procedure to investigate and account for overdispersion can be used, by not restricting the variance-mean relationship to be equal to one as in equation 2.3.31. It should be noted that this assumption would not significantly affect parameter estimates; however the related significances may be slightly affected (Dean, 1992). In the framework of the present demonstration, the regional effect of alcohol controls on the number of accidents was examined assuming extra-Poisson variation, as in equation 2.3.32.

In particular, in Table 2.3.10, parameter estimates are presented for an intercept only model (Model 5) and a model examining the effect of alcohol (Model 6). It is noticed that parameter estimates and their standard errors are not significantly different from the ones obtained with Poisson assumptions. However, a significant estimate of the variance/mean ratio is obtained, indicating that the variance-mean equality assumed in the previous examples was not adequate and that overdispersion was present and is sufficiently handled in this model.

	Model 5 (Constant term)	Model 6 (effect of alcohol)
Fixed effects		
Constant	-6.486 (0.073)	-6.587 (0.092)
Alcontrols		-0.047 (0.010)
Random effects		
Level 2		
σ_{u0}^2 (constant)	0.064 (0.029)	0.094 (0.042)
σ_{u1}^2 (alcontrols)		0.001 (0.000)
σ_{u01}^2 (covariance)		0.006 (0.004)
Variance/mean	22.622 (2.096)	12.892 (1.226)

Table 2.3.10: Extra - Poisson multilevel models for the regional effect of enforcement on road accidents

In Figure 2.3.6, level 1 and 2 residuals are examined for Model 6. Examining the level 1 residuals of the model (Figure 2.3.6a), it is observed that these are normally distributed and independent. When examining level 2 residuals (Figure 2.3.6b), it can be noticed that their distribution is improved in relation to Model 2 above, both in terms of normality and independence from predicted values.

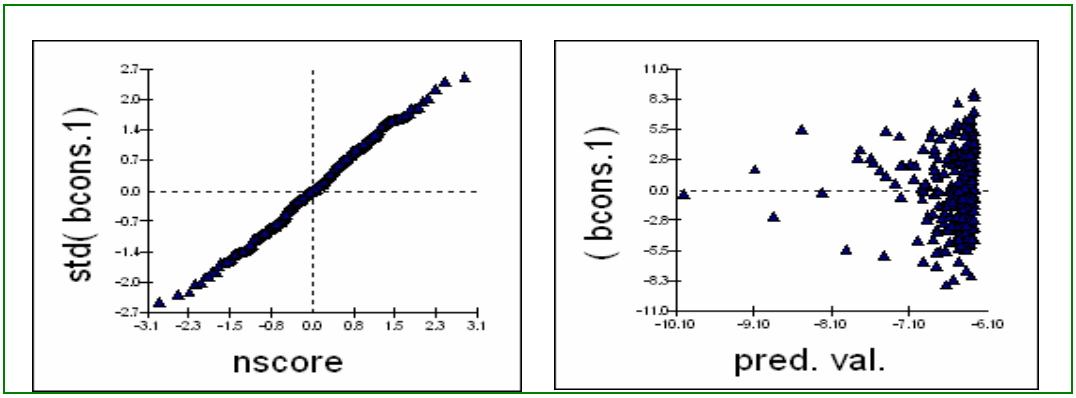


Figure 2.3.6a. Level 1 residuals and normal scores (left graph), Level 1 residuals and predicted values (right graph) for Model 6

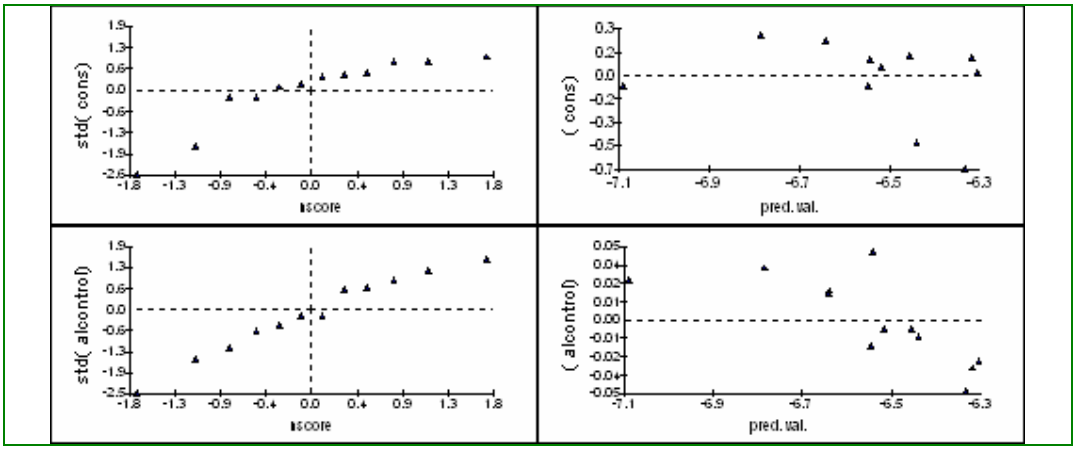


Figure 2.3.6b. Level 2 residuals and normal scores (left graphs), Level 2 residuals and predicted values (right graphs) for Model 6

As explained previously, another option for overdispersed counts data is to assume a Negative Binomial distribution, allowing for a more flexible variance structure, as in equation 2.3.34. The results for the examined dataset are presented in Table 2.3.11. It is interesting to note that the Negative Binomial models are very similar to the Extra-Poisson models, in terms of both fixed and random parameter estimates. It is therefore shown that both Extra-Poisson and Negative Binomial distributional assumptions can efficiently overcome overdispersion in count data. The results of the above analysis models indicate that Models 6 and 8 are the best Models for the purposes of the present analysis.

Summarizing, a Poisson multilevel modelling process was demonstrated by means of an example concerning road accidents and speed-and-alcohol enforcement in Greece. The dataset used includes the number of road accidents and the related speeding and drinking-and-driving violations for 50 counties nested within 12 regions of Greece. The analysis aimed at examining

the effect of police enforcement intensification on the road safety level. Moreover, the regional variation of this effect was quantified.

	Model 7 (Constant term)	Model 8 (effect of alcohol)
Fixed effects		
Constant	-6.477 (0.075)	-6.599 (0.098)
Alcontrols		-0.052 (0.013)
Random effects		
Level 2		
σ_{u0}^2 (constant)	0.064 (0.029)	0.105 (0.046)
σ_{u1}^2 (alcontrols)		0.002 (0.001)
σ_{u01}^2 (covariance)		0.009 (0.005)

Table 2.3.11: Negative Binomial multilevel models for the regional effect of enforcement on road accidents

The multilevel modelling revealed a marginally significant different decrease of road accidents in different regions within the examined period. Moreover, a significant regional variation of the effect of enforcement was obtained. It is interesting to note that no other variables were found to add explanatory effect in the reduction of road accidents in Greece. This was not surprising, as no other parameter (e.g. vehicle ownership, road network length etc.) presented a significant overall variation, comparable to the increase of enforcement, in the examined period. Consequently, the intensification of enforcement is considered to be the main cause of the improvement of road safety in Greece. However, the models developed above are not considered to fully describe this trend. Additional explanatory variables might be required, but not among those for which data were available. However, the models are considered to adequately describe the regional variation of this trend and the relative regional effect of the main causal factor and they are efficient as such.

As far as the regional effect is concerned, the results confirmed the initial suspicion of a significant regional variation of the effect of enforcement. It would be reasonable to assume that the regional variation of the effect is mainly the result of different practices in the implementation of enforcement, as the Greek police is organized according to an administrative structure in full accordance with the examined geographical hierarchy.

2.3.4.5. Conclusions over techniques

In this chapter, several aspects of multilevel models, in which the response variable is a count, were presented and discussed. It was shown that these models are an extension of the classical multilevel models for Normal responses, with a log link function used, in order to satisfy the restriction of positive integer values of the response variable. Within this framework, the Poisson-family distributions (i.e. Poisson, extra-Poisson and Negative Binomial) and their properties were presented.

Multilevel analysis was used to test different Poisson model structures, starting from the basic single-level model and adding fixed and random intercepts and slopes. It was underlined, though, that in Poisson models, random effects are only considered at higher levels, as the level 1 variance is assumed to be known.

During the modelling process, several issues concerning particularities and limitations of data and techniques were discussed. In particular, the effects of multicollinearity (i.e. inclusion of two or more highly correlated covariates/predictors in a model) in multilevel models were discussed, although this problem does not exclusively concern Poisson models

Moreover, the issue of overdispersion in count data was presented. It was shown that extra-Poisson and Negative Binomial distributional assumptions can efficiently handle overdispersion detected in the count data. Modelling results were presented to demonstrate these procedures.

2.3.4.6. Ecological and spatial analysis in road safety research

Spatial analysis refers to a vast group of formal techniques used in various fields of research which study entities using their topological, geometric, or geographic properties. Spatial analytic techniques have been developed in geography, biology, epidemiology, statistics, mathematics, and scientific modelling. A fundamental concept in spatial analysis is that nearby entities often share more similarities than entities which are far apart (Tobler, 1970). Different types of spatial analysis exist, including spatial autocorrelation statistics (which measure the degree of dependency among observations in space), spatial interpolation techniques (which estimate the variables at unobserved locations in geo-space based on the values at observed locations), spatial interaction or "gravity" models (which estimate the flow of people, material or information between locations in geo-space and spatial regression models (which aim at describing spatial relationships among the variables examined) (Miller, 2004). Performing a spatial analysis implies determining an appropriate spatial unit, which may range from a point in space to a large area or zone.

The example presented in this chapter is an example "aggregate spatial modelling", in which the information on spatial variability is available in aggregate form, such as spatial zones. It can also be referred to as "ecological analysis", which uses aggregate group level data to estimate individual level relationships. A concern that often arises in such aggregate analyses is whether the results derived depend more on the type of zones being studied, than on the variables examined (Anselin, 1994).

In this section, a review of spatial and ecological analyses applications in road safety research is presented, also in the light of the fundamental issues mentioned above. A lot of research during the last few years is devoted on spatial analysis of road safety phenomena, mainly focusing on the issue of

spatial dependence of road safety outcomes (road accidents, casualties etc). These studies are particularly relevant in the context of this chapter, not only because they often use hierarchical models, but also because they always assume Poisson-family distributions.

LaScala et al. (2000) explore geographic correlates of pedestrian injury collisions through a spatial autocorrelation corrected regression model. Another study examines ecological and contextual determinants of motor vehicle accident injury in relation to socio-economic indicators, residential environment indicators, medical services availability and utilization, population health, proportion of recent immigrants, crime rates, rates of speeding charge and rates of seatbelt violation (MacNab,.2004). Meliker et al. (2004) evaluated geographic patterns of alcohol-related motor vehicle crashes in a cross-sectional analysis of individual-level blood alcohol content, traffic report information, census block group data, and alcohol distribution outlets, and found that areas of low population density had more alcohol-related motor vehicle crashes than expected. Aguero-Valverde and Jovanis (2006) developed Bayesian[28] negative binomial hierarchical models (with spatial and temporal effects and space–time interactions) to investigate the annual county-level crash frequency in Pennsylvania for 1996–2000, in relation to socio-demographics, weather conditions, transport infrastructure and amount of travel. McMillan et al. (2007) investigate county-level variability in changes in alcohol-related crash rates while adjusting for county socio-demographic characteristics, spatial patterns in crash rates and temporal trends in alcohol-related crash rates through a Bayesian hierarchical binomial regression model.

In these studies, it is often outlined that the level of spatial aggregation may play an important role in the selection of analysis method and the analysis outcome. It is suggested that generalizations made at one level of spatial aggregation may not necessarily hold at another level. Conclusions derived at one scale may be invalid at another. Preliminary examination of the data is important, as one best or unique level of aggregation is not available: it depends upon the objective of the study (Thomas, 1996).

For instance, in a study on child pedestrian casualty data from Devon County UK, the data have been aggregated by two methods: a simple ecological model relating casualty with a child's home location and a more complex spatial model with data aggregated in terms of the collision location. In the first case, it was proved that spatial independence could not be assumed for the data; on the contrary the more complex spatial model resulted in spatially independent counts of accidents (Hewson, 2005).

A relevant issue is also known as the "Modifiable Areal Unit Problem (MAUP)", which may occur when aggregation zones are arbitrary in nature and different spatial units (e.g. counties or census zones) could be just as meaningful in displaying the same base level data (e.g. road accidents counts) (Openshaw, 1984). Although most spatial studies tend towards aggregating units which have

[28] For more information on Bayesian modeling and its applications please see Chapter 2.8

adjacent geographical boundaries, it is possible (and may also be more meaningful) to aggregate spatial units which are spatially distinct.

In Yannis et al., (2007), the example on the effect of alcohol enforcement on road accidents, presented in this section, is also modelled under a different regional classification; counties are grouped on the basis of qualitative similarities, rather than geographical adjacency. The results show that this type of aggregation may be more meaningful for the interpretation of the results, especially as regards the regional variation of the effect.

2.4 Longitudinal measures data

Emmanuelle Dupont and Heike Martensen (IBSR)

The relevance of multilevel models to data that are characterised by complex hierarchical structure (e.g.: speed observations nested within road sites themselves nested within regions…) is easy to conceive of. The fact that multilevel models are very useful when one is to deal with longitudinal data or repeated measurements[29] is in comparison far less obvious: What's hierarchical about repeated measurements? The answer is: The various measurements are to be considered as the lowest level units that are nested within higher level units – the individuals on which these measurements were made. Because they allow such a conception of data, multilevel models offer a handful way to deal with repeated observations. In this section, we will focus on longitudinal data only, but the reader has to bear in mind that ML models similarly allow handling other kinds of repeated measurements (see Section 2.5 about the multilevel analysis of multivariate data). The research example that will be used in this section is the one of a **fictitious** study in which the driving skills of newly licensed drivers are measured at several successive time points. In this case, the repeated observations of the participants' driving skills constitute the lowest level units (level 1, or "the i's"), and are nested within higher-level units, the individuals who each provided a set of observations. The impact of various predictors (the participants' age, their experience with driving before each measurement…) on the evolution of driving skills is assessed. The data examined here are thus characterised by two main dimensions: Time, on the one hand, and the various individuals on which the measurements were made, on the other. This two-dimensional structure is typical of panel data research (Little, Schnabel, and Baumert, 2000), and is what differentiates them from both time-series (see Chapter 3) and cross-sectional data[30], who are characterised by only one dimension (time and "individuals", respectively).

2.4.1 Objectives of the technique

The main objectives of the multilevel analysis of longitudinal observations are identical to those of most techniques allowing their analysis. However, conceptualising longitudinal observations as being hierarchically structured allows for several "extra" objectives to be attained.

Firstly, although one of the basic aims of longitudinal data analysis is the obtainment of an adequate model of the evolution of the criterion variable over time (e.g.: an individual's driving skills), these analyses also render possible the examination of whether the "time – criterion variable" relationship varies *among* individuals. In most "traditional" tools for the analysis of repeated measurements (e.g.: standard linear regression, analysis of variance or multivariate analysis of

[29] Both terms are here understood as the recurrent observation of a dependent variable(s) over time.

[30] The term « cross-sectional data » refers to data that are collected at one point of time as opposed to data collected at several points in time.

variance[31]), however, the question of knowing how the "time – criterion variable" relationship varies among individuals is disregarded: It is assumed to be the same for all individuals, because the slope of the time effect is bound to be fixed.

Second, just as other, more traditional techniques do, the multilevel analysis of longitudinal measurements allows assessing the influence of explanatory variables on the dependent one. Yet, when longitudinal data are at hand, predictors which change with time may also appear to be of particular interest. In the example used in this section, the number of km driven by each participant during the year preceding each measurement of their driving skills was included as a predictor. This latter value, however, is likely to be different for each individual at each occasion measurement, i.e.: It is changing over time. The ability to handle such varying predictors, also called time-varying covariates (Hedeker, 2000; Snijders & Bosker,1999) is another specific feature of the application of multilevel modelling to longitudinal data. What is actually estimated in this case is a relationship that occurs "within individuals" (one participant's skills are likely to be affected by the number of km driven by this very individual, not by another); but this is also likely to differ from one individual to the other (the relation between driving skills and the number of km driven may be stronger for some individuals than for others). Generally speaking, multilevel analyses allow modelling such complex phenomena, which are typical of longitudinal designs.

Longitudinal observations have proved difficult to handle by techniques initially developed for the analysis of cross-sectional data because they are *nested within* individuals (they consist of "sets" of observations, each one being generated by one and the same individual). Indeed, longitudinal data are usually dependent, and their variances and covariances are also unlikely to remain constant over time. These two features violate common assumptions upon which depends the validity of many "traditional techniques". Moreover, because longitudinal designs are more demanding in terms of observations, they most frequently result in unbalanced data sets. Up to two decades ago, no statistical technique appeared to adequately handle these peculiarities *altogether*[32]. Multilevel analyses offer a way to overcome these problems, precisely because their general aim is to take account of hierarchical structures, and hence of dependence among data. So, not only is ML modelling useful to take account of the two-dimensional nature of the data, it is also the adequate means to deal with the occurrence of missing values in panel data sets (see

[31] Applied to longitudinal measurements, analysis of variance (ANOVA) would test the null hypothesis that the means of the observations are equivalent for all occasions, and would take no account of the possible random effects introduced by the individuals from which the observations originate. The same would be true of the multivariate analysis of variance (MANOVA), to the difference that the repeated measurements would in this case be specified on a multivariate response vector and would be transformed in order to test contrasts among the repeated measures (see Hedeker, 2000 for more detailed information on that topic).

[32] MANOVA, for example, can adequately handle heterogeneous variances, but imposes the deletion of all incomplete data sets (Hedeker, 2000).

Section 2.4.3.3 for an extended discussion of this topic). In this respect, the ML analysis may prove particularly useful for road safety research. Indeed, it is recurrently necessary in this research field to have to deal altogether with the need to observe the same phenomenon repeatedly over time while taking account of it being nested into larger units, and while additionally having handle the occurrence of missing values.

2.4.2 Model definition

The multilevel analysis of longitudinal data is a straightforward extension of the multilevel modelling of cross-sectional data. The main difference between both methods is more of a conceptual than of a statistical nature. When longitudinal data are considered, the repeated observations make up the lowest level (level-1) of the data hierarchy. The individuals providing the data at the different occasions thus constitute the level-2 units. By analogy with earlier developments on multilevel cross-sectional analysis (sections 2.2.1 and 2.2.2), one could say that individuals here constitute the "context" in which the repeated data arise. Just as with cross-sectional data, level-1 and level-2 predictors can be included in the model. A predictor is qualified as "level 1" when its value is likely to vary as a function of the measurement occasions. They are, indeed, explanatory variables "at the within-individual" level. For this reason, they are often termed "time-varying covariates" (Hedeker, 2000; Snijders & Bosker,1999). In the study over the evolution of driving skills, for example, the cumulative number of kilometres driven by participants during the year preceding each measurement was included as an explanatory variable. Given that this predictor's value is likely to change over the different occasion measurements, it must be conceived as a level-1 explanatory variable, or as a time-varying covariate. Participants' gender, on the contrary, is a level-2 explanatory variable: It does not change over time.

In order to put forth the similarities between the multilevel analysis of cross-sectional and longitudinal data, the notation used for level-2 units will be "j", and the one for level-1 units will be "i", by analogy with section 2.2.1 ("Basic two-level model"). This will help making clear that the models described in both sections are identical. The reader must nevertheless remain aware that the "j's" here refer to the individual participants while the "i's" designate the measurement occasions.

Before defining the multilevel model as applied to longitudinal data, a comment is necessary about the particular option chosen to code the time effect: In all subsequent developments, the latter will be noted "$\beta_1(t - t_0)$". The 7 measurement occasions were indeed coded as "$t = 0,1,2,...,6$". The value 0 has been assigned to the first occasion in order to make it the reference point in the analyses. Subtracting the value t_0 from t allowed the intercept referring to this first measurement occasion rather than to a possibly meaningless 0 value. Various sensible coding options exist, and could prove more suited to other research problems. It is also important to note that specifying the effect of time as "$\beta_1(t - t_0)$", implies that this effect is a linear one: The evolution of driving skills from one occasion to the other is here assumed to be the same, whatever

the pair of occasions considered. Again, more elaborated functions are available, and would probably allow a more realistic representation of the time effect[33]. Actually, the quality of a model fitted to longitudinal data greatly depends on the particular function chosen to depict the effect of time on the criterion variable. A necessary step in the analyses consists of probing the model's fit with different such functions, using as a guide empirical and theoretical knowledge of the problem investigated. However, this topic will not be further discussed here, because this is a general question with respect to the analysis of longitudinal data, and not a specific issue of the application of multilevel analyses to these data. Furthermore, this is vast an issue, and could constitute the object of a whole chapter in the present document. For more detailed discussions of the modelling of time effects in the context of multilevel models, the reader is referred to Hedeker (2005) and Snijders and Boskers (1999).

2.4.2.1. Definition of the random intercept model

As its name indicates, the random intercept model specifies that the value of the intercept is allowed to vary randomly at level 2 (between individuals). Given the particular coding option chosen here for the time variable, the intercept refers here to the individuals' level of driving skills at the 1st occasion measurement.

The random intercept model is:

$$Y_{ij} = \beta_{0j} + \beta_1 (t - t_0)_{ij} + e_{ij} \qquad (2.4.1a)$$

The coefficient for the time effect (β_1) is bound to be fixed, while the one for the intercept, which is assigned the subscript "j" is defined as random at level-2. Unfolding the model's hierarchical nature, the following equation defines the intercept:

$$\beta_{0j} = \beta_0 + u_{0j} \qquad (2.4.1b)$$

This equation describes the level of driving skills at the first measurement occasion as being a function of a fixed population value (β_O) and of individuals' random departure from this value (u_{0j}). The term u_{0j} thus describes the individual-specific influence on the intercept's value.

The "complete" model is thus written as:

$$Y_{ij} = \beta_0 + \beta_1 (t - t_0)_{ij} + u_{0j} + e_{ij} \qquad (2.4.1c)$$

The random part of the model ($u_{0j} + e_{ij}$) specifies that two random sources determine the value of the observations (Y_{ij}): the individual and occasion-level.

[33] Examples are : polynomial, piecewise, or spline functions. Each is discussed in details in Snijders & Bosker (1999).

Each of these parameters' variances ($\sigma^2_{u_{oj}}$ and σ^2 for the individual and occasion-level residuals, respectively) indicates the magnitude of the variation in the observations that is attributable to each level. The variance of the observations themselves (Y_{ij}) is defined as being composed of individual and occasion-level random departures:

$$\mathrm{Var}(Y_{ij}) = \sigma^2_{u_0} + \sigma^2 \qquad (2.4.2)$$

Note that this model estimates 2 and only 2 parameters to define the random variance of the observations. These parameters are bound to remain the same at all occasion measurements. This is an important property of the random intercept model: It assumes compound symmetry. This issue will again be addressed in the section devoted to the model assumptions.

It is also important to clarify the nature of the parameter "$\sigma^2_{u_0}$". This parameter actually corresponds to the covariance between two observations randomly selected among the whole set of observations provided by one (randomly selected) individual:

$$\mathrm{Cov}(y_{ij}, y_{i'j}) = \sigma^2_{u_0} \qquad (2.4.3)$$

The intra-class correlation coefficient is obtained by calculating the ratio of the level-2 variance to the total variance, and thus "quantifies" the degree of resemblance of two observations taken among those generated by one individual (as compared to observations selected among the observations of *different* individuals):

$$\rho_I = \rho\{Y_{ij}, Y_{i'j}\} = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2} = \frac{\mathrm{Cov}(y_{ij}, y_{i'j})}{\mathrm{Var}(y_{ij})} \qquad (2.4.4)$$

Applied to the research example of young driver's skills, the intra-class correlation coefficient would thus correspond to the average correlation between the driving skills of the same driver measured at any two different time points.

2.4.2.2. Definition of the random intercept and slope model

One can assume that substantial between-individual variation affects the relationship between time and the criterion variable (e.g.: driving skills). In other words, the effect of time on the dependent variable could be larger for some individuals than for others. Such a supposition leads to the following model specification:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(t - t_0)_{ij} + e_{ij} \qquad (2.4.5a)$$

The slope parameter β_1 is now assigned the subscript j and is thus allowed to vary randomly among individuals. The two macro-models defining the level-2 intercept and slope are:

$$\beta_{0j} = \beta_0 + u_{0j} \qquad (2.4.5b)$$

$$\beta_{1j} = \beta_1 + u_{1j} \qquad (2.4.5c)$$

β_0 and β_1 represent the population intercept and slope, u_{0j} and u_{1j} respectively represent the individuals' departure from these population intercept and slope.

Combining the equations for the two macro-models gives:

$$Y_{ij} = \beta_0 + \beta_1 (t - t_0)_{ij} + u_{0j} + u_{1j}(t - t_0)_{ij} + e_{ij} \text{ “} \qquad (2.4.5d)$$

The random part of the model ($u_{0j} + u_{1j}(t - t_0)_{ij} + e_{ij}$) now specifies three random sources to determine the value of the observations: Within-individuals random deviations (e_{ij}) and between-individuals variations, which are now further subdivided into random departures from the intercept and random departures from the slope.

Together, the variances of the random intercept and slope ($\sigma^2_{u_o}$ and $\sigma^2_{u_{1j}}$) provide a rough indication of the importance of inter-individual variation from the population intercept and slope. The covariance between the random intercept and the random slope ($\sigma_{\mu_{01}}$) is also part of the parameters estimated in the model. This parameter represents the co-variation between individual-related variation of initial driving skills (the random intercept) and individual-related variation of the effect of time on these skills (the random slope). A negative covariance (the higher the intercept, the weaker the slope), for example, indicates that time has a more important impact on the driving skills of those individuals who were initially poor at driving.

The introduction of a random slope for a given effect in a model thus substantially complicates the latter. It amounts to estimating 3 parameters for one predictor (the fixed coefficient, the random slope's variance and its covariance with the random intercept). Random slopes also introduce complexity with respect to the definition of the observations' *variance*. Indeed:

$$\mathrm{Var}(Y_{ij}) = \sigma^2_{u_0} + 2\sigma_{u_{01}}(t - t_0) + \sigma^2_{u_1}(t - t_0) + \sigma^2 \qquad (2.4.6)$$

This model expresses the total variance of the drivers' skills to be a function of (1) the various "between-individual variations" (variation in the average skill value at the first measurement occasion ($\sigma^2_{u_0}$), in the average time effect ($\sigma^2_{u_1}(t - t_0)$) and the co-variation between both) and (2) the within-individual variation (σ^2). In contrast to the random intercept model, both the slope variance and the slope-intercept covariance depend on time ($2\sigma_{u_{01}}\underline{(t - t_0)}, \sigma^2_{u_{1j}}\underline{(t - t_0)}$). Consequently, the observations' variance itself is

allowed to vary over time. The same is true for the observations' co-variances (i.e.: the co-variance between two driving skills measurements made at different time points on the same individual):

$$\mathrm{Cov}(Y_{ij}, Y_{i'j}) = \sigma^2_{u_0} + \sigma_{u_{01}}\{ (t-t_0)_i + (t-t_0)_{i'} \} + \sigma^2_{u1}(t-t_0)_i(t-t_0)_{i'} + \sigma^2 \qquad (2.4.7)$$

The random intercept-and-slope model does not assume compound symmetry for the matrix of the observations' variance-covariances. For this reason, it allows a more realistic representation of longitudinal data.

When defining the random intercept model, it was established that the intra-class correlation coefficient expresses the ratio of the level-2 variance (i.e.: between individual) to the total variance (between- *plus* within-individuals) in the observations. In the case of the random intercept and slope model, however, the level-2 variance is made up of the random intercept and the random *slope* variance, as well as of their covariance. These two parameters depend on time. Being variable over time, they render impossible the definition of a unique intra-class correlation coefficient.

2.4.3 Model assumptions

2.4.3.1. Random parameters

The level-2 random coefficients (or level-2 residuals) are considered representative of distributions of individual effects in the population. These distributions parameters themselves are assumed to be normally distributed with means 0 and variances $\sigma^2_{u_0}, \sigma^2_{u_1}$ for the intercept and slope, respectively:

$$\begin{pmatrix} u_{oj} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u_0} \sigma_{u_0u_1} \\ \sigma_{u_0u_1} \sigma^2_{u_1} \end{pmatrix} \right) \qquad (2.4.8)$$

The level-2 coefficients are also assumed to be independent over j (i.e.: across the level-2 units, or individuals), and independent from the level-1 residuals, e_{ij}.

The level-1 residuals (e_{ij}) are assumed to be normally distributed with mean 0 and variance σ^2 ($\varepsilon_{ij} \sim N(0, \sigma^2)$), and to be independent from one another.

These assumptions must be understood in the framework of the conditional nature of the model. From this perspective, saying that the level-1 residuals are independent amounts to stating that they are independent, *conditional* on other effects in the model. To put it in other words: Once the individual-level effects (u_{oj}, u_{1j}) are specified in the model, the level-1 error term is "cleaned", and the residuals at level 1 can be considered independent from one another. By contrast, when they are not specified in the model, the individual-level random effects are confounded with the level-1 error term and introduce dependence among the level-1 residuals.

2.4.3.2. Structure of the observations' variances and covariances

For the random intercept model, the variance and covariances of the observations were defined as:

$$\mathrm{Var}(Y_{ij}) = \sigma^2{}_{u_0} + \sigma^2 \qquad (2.4.9)$$

$$\mathrm{Cov}(Y_{ij,i'j}) = \sigma^2{}_{u_o} \qquad (2.4.10)$$

The random intercept model assumes the observations' variances and covariances to remain the same whatever the moment at which the observations are made. This is the compound symmetry assumption. The introduction of a random slope to qualify the effect of time implies a more complex variance-covariance structure for the observations: The variance of the individuals' observations is allowed to vary as a function of time, and so is their covariance. Defining the time coefficient as random is thus one way to relax the compound symmetry assumption.

2.4.3.3. Assumptions about missing values

As noted by Hox: "In longitudinal research, a major problem is the occurrence of panel attrition: Individuals who, after one or more measurement occasion, drop out of the study altogether" (2002, p. 95). Depending on the causes underlying the occurrence of panel attrition (or missing data), three broad situations can be distinguished (Goldstein & Woodhouse, 2001).

First, data can be said to be *Missing Completely At Random* (MCAR): Their missingness is independent of all other variables included in the model. In the driving skills example, it is likely that the requirement to come back seven times in order to have one's driving skills assessed would appear too much of a burden to many participants, and that consequently far fewer of them would have completed the 7th measurement occasion as compared to the first one. In such a case, panel attrition is linked to a broad feature of the study (its longitudinal nature), but is neither related to the true value of the response (participants' actual driving skills), nor to any of the predictors measured and included in the model. As Wothke puts it: "The fact that a variable's data is missing is not thought to affect its distribution, that is: *P (Y|y missing) = P (Y|y unobserved)*" (2000, p. 224).

The second situation termed: "*Missing At Random*" (MAR) is one in which the probability of being missing depends on predictor variables in the model, or on previous *observed* values of the dependent variable; but is otherwise unrelated to the model's parameters, in particular, to the level-1 and level-2 random effects (Goldstein & Woodhouse, p. 25). In the case of the example again, one could imagine that males are, generally speaking, less compliant or conscientious than females. This could lead them to drop out from the study more easily than female participants do. Yet, it remains reasonable to assume that missing data occur at random, *conditional on the other variables included in*

the model (in this case, gender), and, still conditional on these variables, that they are independent of the values of the response variable. Provided that gender is measured and included in the model, missingness does not constitute a problem. Again quoting Wothke, one could say: "missing and observed distributions of y are identical, *conditional on a set of predictors* or stratifying variables x, that is, *P (Y|y missing, X) = P (Y|y unobserved,X)*" (2000, p. 224).

When data are MCAR or MAR the property or mechanism that caused some data points to be missing does not endanger the interpretation of the results, because it is either unrelated to the observed values (MCAR) or taken up into the model by the inclusion of a predictor (MAR). Finally, a third and more problematic scenario can be faced: The one in which "the probability of a non-response depends on the *unobserved value* of the observation itself". One can imagine, for example, that participants with the poorest driving skills produce more missing values because the risk of them being involved in a crash is larger, making them more likely to be unavailable for further tests in the course of the study, because of hospitalisation - or worse - death. In such a case, missing data cannot be said to occur at random any more and, in contrast to MAR data, the "mechanism" underlying missingness is ignored and cannot be controlled for[34].

Multilevel models assume data to be Missing At Random (MAR). On this point they differ from other statistical models, such as Multivariate Analysis of Variance, which assume data to be missing completely at random. MANOVA is used not only to assess effects of predictors on panels of dependent variables, but also on repeated measurements. Yet multilevel models, when applied to either type of data offer the additional advantage of being able to handle missing values, because they assume MAR rather than MCAR data. "Individuals with incomplete response vectors may be included in the analysis on the basis of the assumption that the association between their responses will, on average, mimic that which is observed for individuals with complete response vectors". (Goldstein & Woodhouse, p. 25). A note of caution is nevertheless necessary: Incomplete responses on the *explanatory* variables *cannot* be included in the analyses. This problem is, obviously, more likely to be encountered when time-varying predictors or covariates are included in the model. The only solution, in this case, is to proceed to a "completer analysis", just as in MANOVA, and to exclude all cases with missing values on explanatory variable(s) from the analyses.

2.4.4 Research problem

Given the lack of appropriate data, the present analyses are based upon a fictitious example study for which a dataset had to be simulated. Although the tests and predictions presented here are coherent with existing literature and empirical evidence on the topic assessed - namely, the evolution of driving skills

[34] Strategies for dealing with non-random missing data are discussed in Goldstein & Woodhouse, 2001.

with age and the experience acquired - **the results of the analyses reported here thus remain to be interpreted as hypothetical**.

Accident records unequivocally indicate that young drivers have considerably higher risks of being involved in a crash than older drivers do. This risk declines most dramatically during the first years following the acquisition of the driving license, the trend then becomes smoother. Although the phenomenon itself is well documented, the factors that determine the overrepresentation of younger drivers in accident records are still poorly understood. The relative importance of age on the one hand, and of driving experience, on the other is the object of many discussions (see Catchpole, Macdonald, and Bowland, 1994, and Vlakveld, 2005 for reviews). Empirically distinguishing these two factors is, of course, uneasy a task: The older one gets, the more one has driven, the more experience acquired and the better developed one's driving skills! The empirical problem which is focused on in this section is the one of the relative impact of age and experience (here understood as the number of kilometres a driver drives in his/her daily life) on the evolution of driving *skills* among newly licensed drivers. Is age important in itself, or is it the intensity of the drivers' training that counts?

The design of this fictitious study would be the following: Upon reception of their brand new driving license, a large panel of 500 young drivers was invited to take part in a 7-year long study. The first test of their driving skills took place 1 to 4 weeks after obtainment of the driving license. During six years afterwards, the participants took every year the same practical evaluation of their driving skills. A single "driving skills" score was calculated on the basis of the test. The occasion variable was coded as "0" for the first measurement occasion; the others were assigned numbers 1 to 6. A similar coding scheme was adopted for the "experience" predictor, which was defined as the number of km driven before each measurement occasion. It was coded "0" at the first measurement occasion and corresponded to the cumulative number of km driven for all the others. In order to avoid as much as possible confounding effects between the age, occasion and experience predictors, age was defined as the "initial age", i.e., the individuals' age when they started to drive. Defined in this way, age is made an individual-level predictor and tracks are kept of the only meaningful "age aspect" from the point of view of the study described[35]. For this reason, and for the sake of clarity, the term "initial age" will be used from now on to refer to this predictor.

[35] Defined as a time-varying covariate of the kind of the « occasion » or « experience » one, the age coefficient would have indicated how driving skills vary with 1-unit increase in age (i.e.: a one year increase). This would have been an information identical to the one conveyed by the "occasion" predictor. The way the age predictor is currently defined, on the opposite, indicates the change in driving skills associated with a one-unit increase of the age the individual has *when starting to drive*…

2.4.5 Dataset

The total number of participants in the simulated data-set is n = 500. The total number of observations amounts to N = 3500. The driving skills score ranges from 0 to 15 (mean 6.54, SD 2.5). The effects of the following predictors were assessed: "occasion measurement", "experience", and "initial age"[36].

Parameter	Model 0 "Empty model"	Model 1 Fixed "occasion" effect	Model 2 "Occasion" and "Experience"
	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Fixed			
Intercept	6.54 (0.08)	5.10 (0.09)	5.05 (0.10)
Occasion	/	0.50 (0.01)	0.03 (0.04)
Experience	/	/	0.95 (0.07)
Random			
Level 2			
$\sigma^2_{u_0}$ (intercept)	2.86 (0.21)	3.03 (0.21)	3.02 (0.30)
$\sigma_{u_0u_1}$ (covariance)	/	/	/
$\sigma^2_{u_1}$ (occasion)	/	/	/
Level 1			
$\sigma^2_{e_0}$	3.40 (0.09)	2.25 (0.06)	2.14 (0.06)
-2xloglikelihood	15182.69	13947.38	13784.95
Deviance test	/	χ^2_1 =; $p < .000$	$\chi^2_2 = 0.6$; $p = .74$, *n.s.*

Table 2.4.1: Models fitted and associated estimates

[36] The data were simulated as follows: For each individual for each year an experience value was created by adding a random number between 0 and 1 to the experience value from the preceding year (starting with 0 at telaps=0). In this way experience was highly correlated to telaps (0.89). For the simulation of the skill score, there was a random number for each individual that constituted this persons intercept (driving skill at telaps=0). To calculate the increase of the driving skill, the experience value was multiplied by a slope-value. The slope value consisted of the following summation: 1) a constant, 2) the same random number as that for the intercept, so that intercept and slope would be moderately correlated, 3) the initial age value for that individual (a random number between 18 and 54, with 75% between 18 and 23), so that the increase in driving skill per experience unit would be higher for persons with a higher initial age and 3) another random number unique to the slope of that particular individual. By construction, driving skills are therefore directly related to experience (r=.42), while the relation between driving skills and telaps (r=.40) runs exclusively via experience.

2.4.6 Model fit and diagnostic

The simplest model that can be fitted to account for the evolution of drivers' skills is the empty model (model 0). No explanatory variable is included in this model, and the average skill value is merely defined as being determined by two sources of random variation: one taking place between individuals and the other taking place within individuals, between each occasion measurement. Table 2.4.1 provides the estimated parameters associated with each model fitted. As it indicates, the average skill value at the first measurement occasion is 6.54. The random variation around this value is lower at level 2 – or at the individual level – than at level 1.

Taken at face value, this result would suggest that there is more variation between the driving skills of the same individual measured at two different time-points than between the average driving skills of two different individuals. This is simply due to the fact that the time effect (i.e., the "occasion" variable that also indicates how many years have passed since the acquisition of the driver's licence) has not been included in the model yet. Once the "time effect" will be included in the analyses and that the occasion-level variance will be properly modelled, the estimate for the individual-level variance will become much more realistic. This is what is done in model 1, and in this case the level 1 variance estimate is indeed lower than the level 2 variance estimate.

The results associated with model 1 reveal a significant effect of the "occasion" predictor (Z = 50, $p < .000$). However, once the fixed effect of experience is also included in the model, the occasion coefficient decreases substantially and turns out not to be significant any more (Z = 0.75, $p = .22$, *n.s.*). This suggests that occasion affected the participants' driving skills essentially because it shared an important part of variance with experience. This is not surprising; given the way the two predictors were respectively coded (see section 2.4.6). The correlation between these predictors is indeed extremely high ($r_{occ,\exp} = .89$, $p < .000$). For this reason, the occasion predictor was dropped from subsequent analyses, experience then constituting the only time-varying predictor remaining in the model (model 3).

Table 2.4.2 presents the estimated coefficients for the models fitted once the occasion predictor is excluded from the analyses. The fixed effect of experience

on skills is highly significant (Z = 33, $p < .000$). The slope coefficient reveals that a 1-unit increase in the number of km driven is associated with a 1-unit increase in driving skills.

Model 5 specifies this effect as being random at the individual level. The estimate for the random variation of the experience slope at level 2 ($\sigma^2_{u_1}$) is small, but significant (Z = 45, $p < .000$). The intercept-slope covariance ($\sigma_{u_0 u_1}$) is positive and also significant (Z = 48, $p < .000$). This indicates that the effect of experience was larger among the drivers who had good driving skills from the

start. In other words: The higher the initial driving skills, the faster progresses are made as a function of the number of kilometers one drives.

How does age affect the evolution of driving skills over the years? Model 6 was specified to assess the effect of initial age on driving skills. It also includes the "initial age x experience" cross-level interaction. The deviance test comparing this model to model 6 indicates a significant fit improvement, although the cross-level interaction is the only significant effect ($Z = 5$, $p < .000$). The coefficient for this interaction is positive; suggesting that the experience effect increases with the individual's age at the time he/she has begun to drive. The slope variance of the experience effect ($\sigma^2_{u_1}$) also decreased substantially in model 6 as compared to model 5, suggesting that the "experience x age" interaction explains part of the random variation of the experience effect among individuals.

Parameter	Model 4 "Experience only"	Model 5 Random slope for "experience"	Model 6 "Experience, age, and their interaction"
	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Fixed			
Intercept	5.06 (0.09)	5.06 (0.08)	5.06 (0.08)
Experience	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)
Age			0.01 (0.02)
Age x Experience			0.05 (0.01)
Random			
Level 2			
$\sigma^2_{u_0}$ (intercept)	3.02 (0.21)	2.09 (0.19)	2.09 (0.19)
$\sigma_{u_0u_1}$ (covariance)	/	0.24 (0.05)	0.24 (0.05)
$\sigma^2_{u_1}$ (occasion)	/	0.09 (0.02)	0.04 (0.02)
Level 1			
$\sigma^2_{e_0}$	2.14 (0.06)	2.03 (0.06)	2.03 (0.06)
-2xloglikelihood	13785.50	13700.75	13634.49
Deviance test	/	$\chi^2_2 = 84.75$; $p < .000$	$\chi^2_2 = 66.26$; $p < .000$

Table 2.4.2: Models fitted and associated estimates

2.4.7 Model interpretation

The application of multilevel techniques is also truly beneficial for the analysis of longitudinal data. The theoretical developments made earlier in this section made several points pleading for the statistical advantages of conceptualising

longitudinal measurements into a multilevel structure (relaxed variances-covariances assumptions…).

The fictitious example dataset and the results presented above in addition illustrated the conceptual interest of doing so. These results clearly indicate that driving skills do increase over time, but that this was mostly a function of the additional practice that the driver acquires over the years. Should single-level modeling have been used, however, the heterogeneity of this effect among individuals would have go unnoticed, and no attempt could have been made at determining what individual-related factors affect the impact of the experience acquired over time on the driving skills of individuals. The present analysis in contrast allows concluding that older individuals, as well as those who already are “gifted” for driving (those with an initially high level of skills), will benefit from intensive driving practice to a larger extent.

2.5 Multivariate models

George Yannis, Eleonora Papadimitriou and Constantinos Antoniou (NTUA)

2.5.1. Objectives of the technique

All the models described in the previous sections considered only a single response variable. In this section, models that allow the inclusion of several responses simultaneously as functions of explanatory variables are examined. Interest in these data lies in the relationship between the responses at various hierarchical levels, in whether there are significant differences in this relationship explained by other variables, and in whether the variability differs among responses.

The analysis has the following objectives:

- Present the assumptions and properties of multivariate Normal multilevel models in relation to univariate models
- Describe the assumptions and particularities of multivariate models for Poisson responses
- Use the above techniques to explore the regional effect of police enforcement on the number of road accidents and road accident fatalities in Greece, testing both Normal and Poisson assumptions for the two responses.

2.5.2. Model definition and assumptions

In order to define a multivariate model, the individual component should be treated as a level 2 unit and the "within-component" measurements (e.g. the different responses) as level 1 units. Each level 1 entry has a response, which is one of the multiple responses. The basic explanatory variables are a set of dummy variables that indicate which response variable is present. Further explanatory variables are defined by multiplying these dummy variables by unit level explanatory variables (Rasbash et al, 2000).

In particular, in the simplest case of a Normal bivariate model, each level 1 entry would consist of either of the two responses, with dummy variables indicating which of the two variables is present (for each response there is a dummy that is one whenever the level-1 value belongs to that particular response variable and 0 otherwise). Further explanatory variables would be created by multiplying their values with the dummy variables indicating which response variable is present (Table 2.5.1).

Individual	Given Response	Intercepts		Explanatory Variable (X)	
		Response 1	Response 2	X.R1	X.R2
1	Response 1	0	1	0*x	1*x
1	Response 2	1	0	1*x	0*x
2	Response 1	0	1	0*x	1*x
2	Response 2	1	0	1*x	0*x
3	Response 1	0	1	0*x	1*x
3	Response 2	1	0	1*x	0*x

Table 2.5.1: Data matrix structure for the simple bivariate model

The statistical formula for the two level basic Normal bivariate model, including one explanatory variable, is written as follows:

$$y_{ij} = b_0 z_{1ij} + b_1 z_{2ij} + b_2 z_{1ij} x_j + b_3 z_{2ij} x_j + u_{1j} z_{1ij} + u_{2j} z_{2ij} \qquad (2.5.1)$$

Where $z_{1ij} = \begin{cases} 1 & \text{if response 1} \\ 0 & \text{if response 2} \end{cases}$, $z_{2ij} = 1 - z_{1ij}$,

$$Var(u_{1j}) = \sigma^2_{u1},\ Var(u_{2j}) = \sigma^2_{u2},\ covar(u_{1j}, u_{2j}) = \sigma_{u12}$$

It is interesting to note that there is no level 1 variation specified, as level 1 exists solely to define the multivariate structure. The level 2 variances and covariance are the (residual) between-units variances. In the case where only the intercept dummy variables are fitted and in the case where every unit has both responses, the model estimates of these parameters become the usual estimates of the between-units variances and covariance. The multilevel estimates are statistically efficient even where some responses are missing (Rasbash et al. 2000).

It should be noted that the estimates obtained are not necessarily the same as those that would be obtained by fitting two separate univariate models. If there is a tendency, for instance, to report or measure only one of the responses, or if the occurrence rate of one response is different from the occurrence rate of the other response, the omitted values of the other response are not missing completely at random. In the univariate analysis there is no way to correct for this bias, as it is considered that any absent values are missing completely at random (MCAR). The multivariate model contains the covariance between the responses, assuming that the absent values are missing at random given the value of the other response (MAR), which is a weaker assumption (Hox, 2002).

Thus, the formulation as a 2-level model allows for the efficient estimation of a covariance matrix with missing responses, where the missingness is at random. This means, in particular, that multilevel analyses are particularly suited to analyse research designs in which not every unit (e.g., not every individual) has a value on every measurement but rather measurements are randomly

allocated to units. Such "rotation" or "matrix" designs are common in many areas and may be efficiently modelled in this way.

A third level can be incorporated and this is specified by inserting a third subscript, k, and two associated random intercept terms:

$$y_{ijk} = b_0 z_{ijk} + b_1 z_{21jk} + b_2 z_{1ijk} x_{jk} + b_3 z_{2ijk} x_{jk} + v_{0k} z_{1ikj} + v_{1k} z_{2ijk} + u_{0jk} z_{1ikj} + u_{1jk} z_{2ijk} \quad (2.5.2)$$

Where $z_{1ijk} = \begin{cases} 1 & \text{if response 1} \\ 0 & \text{if response 2} \end{cases}$, $z_{2ijk} = 1 - z_{1ijk}$, $x_{jk} = \begin{cases} 1 \\ 0 \end{cases}$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N(0, \Omega_v) \qquad \Omega_v = \begin{bmatrix} \sigma^2_{v0} & \\ \sigma_{v01} & \sigma^2_{v1} \end{bmatrix}$$

$$\begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim N(0, \Omega_u) \qquad \Omega_v = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

The 2 by 2 covariance matrix between response 1 and response 2 is partitioned into a level-2 between-units component Ω_v and a level-3 between-units component Ω_u.

This model could be extended further, by allowing the effect of the explanatory variable for each response to vary on level 3. Further explanatory variables can be added and their coefficients can vary randomly at either level. It should be noted that, multiplying each explanatory variable with all the dummy variables, each regression coefficient in the model is different for each response. In a considerably simplified model, one could impose an equality constraint across all response variables, which is equal to adding the explanatory variables directly, without multiplying with the available dummies of level 1. This produces common coefficients for the two responses, resulting in a model that can be considered as "nested" within the above detailed model.

It should be noted that formulae 2.5.1 and 2.5.2 concern the bivariate Normal multilevel case. However, in most cases in road safety the level 1 response is discrete (Binomial, Poisson etc.). In this case, the two-level bivariate model can also be specified in the usual way, by assuming e.g. a Poisson distribution at the lowest level of the multilevel structure and a multivariate Normal distribution at the higher levels of the multilevel structure, as follows:

$$Log(y_{ij}) = b_0 z_{1ij} + b_1 z_{2ij} + b_2 z_{1ij} x_j + b_3 z_{2ij} x_j + u_{1j} z_{1ij} + u_{2j} z_{2ij} \quad (2.5.3)$$

Where $z_{1j} = \begin{cases} 1 & \text{if response 1} \\ 0 & \text{if response 2} \end{cases}$, $z_{2j} = 1 - z_{1j}$,

$$covar(y_{1j}, y_{2j} / u_{1j}, u_{2j}) = 0$$
$$covar(y_{1j}, y_{2j}) = \sigma_{12}$$
$$covar(y_j / z_{1j}, z_{2j}) = \rho$$

It should be underlined though that this formulation is not the formulation of a (full) bivariate Poisson model, in which the variation in all levels is assumed to be Poisson, and whose formulation is much more complex. This case of multilevel models could be considered as a hybrid Normal - Poisson bivariate model, where the bivariation comes from the normal side of the random factors, i.e. is estimated at the 3rd level of the multilevel structure.

A typical example to illustrate the multilevel normal multivariate response model is given by Rasbash et al. (2000) and concerns the scores on two components of a science examination taken in 1989 by 1905 students in 73 schools in England. The first component is a traditional written question paper, and the second consists of coursework. Interest in these data centres on the relationship between the component marks at both the school and student level, whether there are gender differences in this relationship and whether the variability differs for the two components.

An example of fitting multivariate models with Poisson responses can also be found in Langford et al. (1999), where deaths from cancer and cardiovascular diseases in Glasgow are examined simultaneously in a spatial model.

Another, interesting example of multilevel multivariate modelling is given in Duncan et al. (1999); the first response is a binary response indicating whether or not an individual smokes, and the second response is only present for those individuals who smoke and is the number of cigarettes smoked. This model has two interesting features. Firstly, if the number of cigarettes smoked was modelled as a continuous univariate response, there would be a large spike at zero, which would violate any simple Normal theory. However, in the multivariate framework, these individuals are properly included by the first binary response. Secondly, the covariance between the two responses at higher levels can be very informative. In Duncan et al. the individuals were nested within neighbourhoods. A positive covariance at the neighbourhood level means that smokers who are in an area where the probability of smoking is high will tend to smoke more cigarettes than smokers in an area where the probability of smoking is low. In other words: if you are a smoker and a lot people around you are smoking you will smoke greater numbers of cigarettes than if you are not surrounded by smokers.

2.5.3. Research problem and dataset

In Section 2.3.4 a Poisson multilevel model was fitted to the counts of road accidents to identify within-county and within-region variability of the effect of speeding and drinking-and-driving police controls on road accidents. An offset term was included (see section 2.3.4), in order to model the accident rates per population. Results had indicated a significant regional variation in road accident occurrence, as well as a significant effect of both types of police

enforcement explaining the accident reduction within the examined period. Additionally, models with extra-Poisson variation assumptions (overdispersion) and Negative Binomial assumptions were proved to be more flexible in relation to standard Poisson variation assumptions, correcting for the overestimation of the significances of parameter estimates due to overdispersion.

In this section, the effect of alcohol enforcement on both road accidents and road fatalities is examined. The interest of this analysis lies in the fact that road accident severity (number of fatalities) may or may not be related to accident frequency (number of accidents). In particular, an improved road environment or an increase in traffic may be the causes of fewer fatalities within the same number of accidents. Accordingly, the intensification of police enforcement may or may not have the same effect on the number of accidents as on the number of related fatalities.

Therefore, the dataset presented in Section 2.3.4. is used to demonstrate bivariate multilevel modelling. This dataset includes the number of road traffic accidents and related fatalities in 49 counties nested within 12 regions of Greece for the period 1998-2002. As mentioned in Section 2.3.4, this period corresponds to a considerable intensification of police enforcement.

Bivariate models are therefore developed, with the following variables (Table 2.5.2):

region	1-12 regions of Greece
county	1-49 counties of Greece
accidents	The number of accidents of each county
killed	The number of fatalities in the road accidents of each county
alcontrol (1000)	The number of alcohol controls of each county
Pop (10000)	The population of each county
Cons	The constant term

Table 2.5.2. Variables and values considered in the analysis

It should be noted that, as in the example of univariate Poisson models, the Athens and Thessalonica metropolitan areas, where a disproportional high number of accidents and police controls are observed, were not included in the dataset. Additionally, only the number of alcohol controls is examined as explanatory variable, since in the previous example (section 2.3.4) it was proved that alcohol and speed enforcement are significantly correlated and therefore they should not be examined jointly.

In order to demonstrate the particularities of multivariate multilevel models in case of non - normal responses, two examples are shown:

- An example concerning the Normal bivariate multilevel model; on that purpose, the rates of accidents and fatalities per population were log-transformed and assumed to be normally distributed
- An example concerning the hybrid Poisson - Normal model, by assuming extra-Poisson distributions for the counts of accidents and fatalities and Normal distribution for the higher-level variation. It should be noted that all the assumptions of Poisson multilevel models described in Section 2.3.4 also apply in this case.

2.5.4. Model fit, diagnostics, and interpretation of results

2.5.4.1. A Normal multivariate multilevel model

In this example, the rates of accidents and fatalities per county population were log-transformed, allowing assuming a Normal distribution for the two responses. The initial stage of the analysis concerns a two-level model, which is specified in order to define the bivariate response variable. In particular, level 1 is defined as a dummy variable indicating the presence of each response and level 2 is defined as the respective value of each response. Therefore, a response variable of 98 units (counties) is created; 49 units corresponding to the 1st response (accidents per population) and 49 units corresponding to the 2nd response (fatalities per population). Results are presented in Table 2.5.3.

	Model 1	
	Log (accs/pop)	Log (killed/pop)
Fixed effects		
constant	2.691 (0.029)	0.739 (0.026)
Level 2		
Random effects		
σ_{u0j}^2 (constant)	0.213 (0.019)	
σ_{u1j}^2 (constant)		0.168 (0.015)
σ_{u01j}^2 (constant/constant)	0.077 (0.013)	
-2*Log - likelihood	528.92	

Table 2.5.3: Effects of the basic two-level Normal bivariate model (intercept only)

The intercept terms of the two responses are both highly significant. Additionally, a significant between-response covariance indicates that the two responses follow similar trends. When proceeding in adding a fixed slope for alcohol controls, the results presented in Table 2.5.4 indicate that, although the effect of alcohol enforcement is intuitive (i.e. a negative parameter is obtained) for both responses, it is significant only for accidents. Moreover, the variance of the effect across counties is marginally significant for both responses and no covariance of the effect of enforcement between responses is obtained.

Given that the higher-level variation for the two-level model is not significant, it is unlikely that a three-model would be more efficient (i.e. further partitioning of the random variation would not be meaningful). Moreover, convergence problems were encountered, not allowing for the estimation of the three-level Normal bivariate model, whose results could confirm this assumption. In the following section, the same research problem is estimated under extra-Poisson assumptions for the two responses.

	Model 2	
	Log (accs/pop)	Log (killed/pop)
Fixed effects		
constant	2.776 (0.036)	0.757 (0.033)
Alcontrols	-0.014 (0.003)	-0.003 (0.003)
Level 2		
Random effects		
$\sigma^2_{u0j_2}$ (constant)	0.237 (0.029)	
$\sigma^2_{u1j_2}$ (constant)		0.190 (0.024)
$\sigma^2_{u2j_2}$ (alcontrols)	0.000237 (0.000117)	
σ^2_{u3j} (alcontrols)		0.000125 (0.000087)
σ_{u01} (covariance constant/constant)	0.080(0.020)	
σ_{u02} (covariance constant/alcontrols)	-0.006 (0.002)	
σ_{u03} (covariance constant/alcontrols)		0.000 (0.000)
σ_{u12} (covariance alcontrols/constant)	-0.003 (0.003)	
σ_{u13} (covariance alcontrols/constant)		-0.003 (0.002)
σ_{u23} (covariance alcontrols/alcontrols)	0.00013 (0.00009)	
-2*Log - likelihood	479.16	

Table 2.5.4: Effects of the two-level Normal bivariate model (intercept and slope)

2.5.4.2. A hybrid Poisson - Normal multivariate multilevel model

In this case, the untransformed accidents and fatalities counts are used , and assumed to be extra - Poisson[37] distributed. However, the higher level variation is assumed to be Normally distributed. As previously, a two-level model is initially considered, in order to define the bivariate response variable. The natural logarithm of the population is used as an offset in both responses, and so the accident rates per county population are modelled. It should be noted that extra-Poisson distributional assumptions are made so as to allow for more flexibility in the estimations. In particular, the basic assumption of Poisson multilevel models, being that the "real" level-1 variance is assumed to be known (i.e. the variance at the county level), reduces the number of fixed and random parameters that need to be estimated.

The modelling results for the simple examination of variability between responses (two-level model with fixed intercept) are presented in Table 2.5.5.It is noted that conceptually this is the equivalent of a single-level bivariate model.

It is interesting to notice that the intercept terms of the two responses are both highly significant. Additionally, a significant between-response covariance indicates that the two responses follow similar trends. When proceeding in adding a fixed slope for alcohol controls, the results presented in Table 2.5.6 indicate that the effect of alcohol enforcement is significant both for the number of accidents and for the number of fatalities.

[37] In section 2.3.4, extra-Poisson distributional assumptions were found to be suitable for modeling this data

At the next stage, it is examined whether the regional effect on the responses is significant, by adding a 3rd level to the model (which would correspond to the 2nd level of the respective univariate model) and introducing a random intercept.

	Model 3	
	Accidents	Killed
Fixed effects		
constant	-6.471 (0.025)	-8.380 (0.023)
Cov (accs/killed)	4.691 (0.042)	

Table 2.5.5:. Effects of the basic two-level Poisson - Normal bivariate model (intercept only)

	Model 4	
	Accidents	Killed
Fixed effects		
constant	-6.455 (0.023)	-8.372 (0.023)
alcontrols	-0.019 (0.003)	-0.006 (0.002)
Cov (accs/killed)	4.139 (0.657)	

Table 2.5.6. Effects of the two-level Poisson - Normal bivariate model (intercept and slope)

The results presented in Table 2.5.7 show a significant regional variation of both accidents and fatalities, as well as a significant covariance between the two intercepts. Additionally, the regional variability of the intercept is higher for the number of accidents, as indicated by the values of the related mean variances. Moreover, it is interesting to notice that the covariance between responses and its significance is reduced in comparison to those of Model 4. It can be deduced that the variation of accidents and persons killed also follows the same trend within different regions, i.e., some of the covariance between accidents and persons killed is situated at the regional level.

	Model 5	
	Accidents	Killed
Fixed effects		
constant	-6.453 (0.044)	-8.382 (0.028)
alcontrols		
Random effects		
Level 3		
σ_{u0}^2 (constant)	0.092 (0.021)	0.016 (0.008)
σ_{u1}^2 (alcontrols)		
σ_{u01}^2 (covariance)	0.025(0.010)	
Cov (accs/killed)	2.898 (0.556)	

Table 2.5.7. Effects of the three-level Poisson - Normal bivariate model (random intercept only)

By adding a random slope to the model, the results shown in Table 2.5.8 are obtained (Model 6). It is noted that, for practical reasons, only variances (diagonal matrix) are presented. It appears that the mean effect of enforcement on the number of accidents is higher compared to the related effect on persons killed. However, the regional variation of alcohol enforcement effects is very low as far as both number of accidents and persons killed are concerned and only significant as far as the number of accidents is concerned.

	Model 6	
	Accidents	Killed
Fixed effects		
Constant	-6.475 (0.038)	-8.381 (0.026)
alcontrols	-0.025 (0.004)	-0.004 (0.002)
Random effects		
Level 3		
σ_{u0}^2 (constant)	0.053 (0.014)	0.010 (0.007)
σ_{u1}^2 (alcontrols)	0.0004 (0.0002)	0.0001 (0.002)
Cov (accs/killed)	3.313 (0.556)	

Table 2.5.8. Effects of the three-level Poisson - Normal bivariate model (random intercept & slope)

At this stage, there is enough evidence that road accidents and road fatalities present a significantly different regional variation. Additionally, the increase of alcohol controls is associated to a significantly different reduction on accidents and persons killed at national level. However, while the effect of alcohol controls on accidents varies significantly between regions, the respective effect on persons killed does not. .

The above example concerns a multivariate modelling process under Poisson - Normal assumptions. A significant regional variation was observed in both responses. However, a significant variation related to the effect of alcohol controls was observed for accidents only. A less complex univariate model was successfully fitted on the accidents data in Section 2.3.4, and the results had indicated a somewhat higher regional effect of enforcement than the one obtained in the present bivariate analysis. It should be underlined that, for validation purposes, a univariate Poisson model for the number of persons killed was also fitted to the data and the non-significant regional variation of the effect of alcohol enforcement was confirmed. Additionally, the magnitude of fixed effects was also slightly different.

Summarizing, the multivariate structure provides slightly different results as far as the magnitude of the examined effects is concerned, which is due to the fact that dependencies among the two responses are taken into account. In the present example, the number of persons killed in accidents is strongly related to the number of accidents. However, the alcohol enforcement mainly affects the number of accidents. It can therefore be deduced that an increase of alcohol controls is related to a significant decrease of accidents. The number of persons

killed probably decreases because the number of accidents decreases and not because of a direct effect of alcohol controls.

The results seem to indicate that the nationwide intensification of enforcement had an important effect mainly on severe accidents (which may resulting from more risk-taking behaviour, such as speeding). In particular, drivers may have perceived an overall increase of the presence of the Police and adopted their behaviour accordingly, resulting in a significant decrease of severe accidents at national level, and a related decrease on fatalities. However, the effect of enforcement on less severe accidents (resulting from less risk-taking behaviour) varies significantly among regions, and appears to be more dependent to the regional / local presence of the Police on the road network.

2.5.5. Conclusions over techniques

In this section, a multivariate multilevel modelling process was demonstrated. The main interest of the examples presented lies in the illustration of the lower-level structuring to build a multiple response model. In particular, the basic multilevel model structure is exploited to create a multivariate analysis, by shifting the hierarchical structure one level higher and substituting the bottom-level with dummy variables to account for the multiple responses. This process provides several interesting features, mainly concerning the treatment of missing values and the consideration of dependencies among responses.

The examples presented above concerned the effect of alcohol enforcement on the number of road accidents and related casualties. Two approaches were explored as far as the distributional assumptions of the responses are concerned: a bivariate Normal model (resulting from a log-transformation of the responses) and a bivariate hybrid Poisson - Normal model (in which extra-Poisson assumptions were considered for the two responses, with the variation at higher levels to be assumed as Normal). It is underlined that the latter approach is different from the full Poisson bivariate model, in which the variation at all levels is Poisson.

The bivariate Normal modelling approach was proved to be less efficient for the investigation of the research question examined, as convergence problems were encountered in the more complicated (and more interesting) models. On the other hand, a three-level hybrid Poisson - Normal model was successfully fitted to the data, providing some insightful results. It can be said that this model, having fewer parameters to be estimated, is more parsimonious and thus more flexible.

The multivariate modelling processes described above can be applied accordingly to normal, binary, count or mixed responses. Some of the particularities of multivariate multilevel modelling of discrete responses in relation to Normal responses were briefly discussed in the framework of the above examples. However, it is always recommended to begin by fitting simple

univariate models for each response, in order to explore the variability of regional or other effects and the explanatory power of variables, before proceeding to a more complex structure.

2.6 Structural equations models

Christian Brandstaetter and Michael Smuc (KfV)

2.6.1 Objective of the technique

In this chapter, we will introduce concepts for latent dimensions. Often the most important variables are not directly observable. This is true especially for most concepts in psychology, e.g. attitudes, motives or personality traits. In these cases the underlying construct cannot be measured directly, but nevertheless can be assessed indirectly by measuring a number of relevant indicators. Furthermore, the interdependency between these latent dimensions should be analysed. Structural equation modelling, and the special case of factor analysis, was developed for this purpose.

It is important to carry out such analyses where individuals are grouped within hierarchies in a multilevel framework. For example, one may be interested in attitudes with regard to new technologies relevant for traffic safety correlated with driver characteristics. Data on such indicators may be available in different countries and one can postulate a model whereby the underlying attitudes and characteristics vary from country to country (level 2) and also vary randomly over individuals within countries (level 1).

2.6.2 Model definition and assumptions

The theory and application of single level structural equation models, including the special cases of observed variable path models and factor analysis models, is well known (Joreskog and Sorbom, 1979, McDonald, 1985). In this chapter, we look at multilevel generalisations of these models. We will not give details of estimation procedures that are set out in Goldstein and McDonald (1987), McDonald and Goldstein (1988) with elaborations by Muthen (1989) and Longford and Muthen (1992). McDonald (1994) presents an informal overview.

One first considers a basic 2-level factor model where a set of measurements for each person within a sample of countries is available. For the *i* level 1 responses, we first write a multivariate model with *i* responses, where in general some may be randomly missing.

$$y_{ij} = (X\beta)_{ij} + \sum_i e_i z_{ij} \qquad (2.6.1)$$

One may wish to identify some of these factors as the ‘same’ factors at each level, for example by constraining certain loadings to be zero. This means for example that he observe variables for each level have the same correlation with the underlying factor, the latent variable.

A straightforward and consistent procedure for estimating the parameters of this factor model is to perform it in two stages. The first stage involves the estimation of the separate level 1 and level 2 residual covariance matrices. The

second stage involves the factor analysis of these separate matrices using any standard procedure.
All structural equation models, in short SEM, have important assumptions, which have to be known when applying such a concept.

2.6.2.1. Multivariate normal distribution of the indicators

Each indicator whih means he observed variables should be normally distributed for each value of each other indicator. Even small departures from multivariate normality can lead to large differences in the chi-square test, undermining its utility. In general, violation of this assumption inflates chi-square, but under certain circumstances may deflate it. Use of ordinal or dichotomous measurement is a cause of violation of multivariate normality. Please note that multivariate normality is required by maximum likelihood estimation (MLE), which is the dominant method in SEM for estimating structure coefficients. Specifically, MLE requires normally distributed endogenous (i.e. latent or factor) variables.

The Bollen-Stine bootstrap and Satorra-Bentler adjusted chi-square are used for inference of exact structural fit when there is reason to think there is lack of multivariate normality or another distributional misspecification. Other non-MLE methods of estimation exist; some do not require the assumption of multivariate normality.

Under conditions of severe non-normality of data, SEM parameter estimates (ex., path estimates) are still fairly accurate, but corresponding significance coefficients are too high. Chi-square values, for instance, are inflated. Recall for the chi-square test of goodness of fit for the model as a whole, the chi-square value should not be significant if there is a good model fit; the higher the chi-square, the more the difference of the model-estimated and actual covariance matrices, hence the worse the model fit. Inflated chi-square could lead researchers to think that their models were more in need of modification than they actually were. Lack of multivariate normality usually inflates the chi-square statistic such that the overall chi-square fit statistic for the model as a whole is biased toward Type I error (rejecting a model which should not be rejected). The same bias also occurs for other indexes of fit besides the chi-square model. Violation of multivariate normality also tends to deflate (underestimate) standard errors moderately to severely. These smaller-than-they-should-be standard errors mean that regression paths and factor/error covariances are found to be statistically significant more often than they should be.

2.6.2.2. Multivariate normal distribution of the latent dependent variables

Each dependent latent variable in the model should be normally distributed for each value of the other latent variables. Dichotomous latent variables violate this assumption. In this case, other classes of models should be used.

2.6.2.3. Linearity

SEM assumes linear relationships between indicator and latent variables, and between latent variables themselves. However, as with regression, it is possible

to add exponential, logarithmic, or other non-linear transformations of the original variable to the model.

One might think SEM's use of MLE estimation means that linearity is not assumed, as in logistic regression. However, in SEM, MLE estimates the parameters that best reproduce the sample covariance matrix, and the covariance matrix assumes linearity. That is, while the parameters are estimated in a non-linear way, they are in turn reflecting a matrix requiring linear assumptions.

2.6.2.4. Indirect measurement

Typically, all variables in the model are latent variables. Multiple indicators (three or more) should be used to measure each latent variable in the model. Regression can be seen as a special case of SEM in which there is only one indicator per latent variable. Modelling error in SEM requires there should be more than one measure of each latent variable. If there are only two indicators, they should be correlated so that the specified correlation can be used, in effect, as a third indicator and thus prevent under-identification of the model.

2.6.2.5. Low measurement error

Multiple indicators are part of a strategy to lower measurement error and increase data reliability. Measurement error attenuates the correlation and covariance on which SEM is based. Measurement error in the exogenous variables biases the estimated structure (path) coefficients, but in unpredictable ways (up or down) dependent on specific models. Measurement error in the endogenous variables is biased towards underestimation of structure coefficients if exogenous variables are highly reliable, but otherwise bias is unpredictable in direction.

2.6.2.6. Complete data or appropriate data imputation

As a corollary of low measurement error, the researcher must have a complete or near-complete dataset, or must use appropriate data imputation methods for missing cases.

2.6.2.7. Not theoretically under-identified or just-identified

A model is just identified or saturated if there are as many parameters to be estimated as there are elements in the covariance matrix. For instance, consider the model in which V1 causes V2 and also causes V3, and V2 also causes V3. There are three parameters in the model, and there are three covariance elements (1,2; 1,3; 2,3). In this just-identified case, one can compute the path parameters, but in doing so, uses up all the available degrees of freedom. Therefore, one cannot compute goodness of fit tests on the model. AMOS and other SEM software will report degrees of freedom as 0, chi-square as 0, and then p cannot be computed.

A model is under-identified if there are more parameters to be estimated than there are elements in the covariance matrix. The mathematical properties of

under-identified models prevent a unique solution to the parameter estimates and prevent goodness of fit tests on the model.

In most cases, researchers want an over-identified model, which means one where the number of known (observed variable variances and covariances) is greater than the number of unknowns (parameters to be estimated). When one has over-identification, the number of degrees of freedom will be positive (recall AMOS has a DF tool icon to check this easily). Thus, in SEM software output, the listing for degrees of freedom for the chi-square model is a measure of the degree of over-identification of the model.

The researcher is well advised to run SEM on pre-test or fictional data prior to data collection, since this will usually reveal under-identification or just-identification. One good reason to do this is because one solution to under-identification is adding more exogenous variables, which must be done prior to collecting data.

2.6.2.8. Recursivity

Recursive models are never under-identified (that is, they are never models which are not solvable because they have more parameters than observations). A model is recursive if all arrows flow one way, with no feedback looping, and disturbance (residual error) terms for the endogenous variables are uncorrelated. That is, recursive models are ones where all arrows are unidirectional without feedback loops and the researcher can assume covariances of disturbance terms are all zero, meaning that unmeasured variables that are determinants of the endogenous variables are uncorrelated with each other and therefore do not form feedback loops. Models with correlated disturbance terms may be treated as recursive only as long as there are no direct effects among the endogenous variables. Note hat recursivity is just a guarantee for identification and that non-recursive models may also be solvable (not under-identified) under certain circumstances.

2.6.2.9. Not empirically identified due to high multicollinearity

A model can be theoretically identified but still not solvable due to such empirical problems as high multicollinearity in any model, or path estimates close to zero in non-recursive models. There are some signs of high multicollinearity:

- Since all the latent variables in a SEM model have been assigned a metric of 1, all the standardized regression weights should be within the range of plus or minus 1. When there is a multicollinearity problem, a weight close to 1 indicates the two variables are close to being identical. When these two nearly identical latent variables are then used as causes of a third latent variable, the SEM method will have difficulty computing separate regression weights for the two paths from the nearly-equal variables and the third variable. As a result it may well come up with one standardized regression weight greater than +1 and one weight less than -1 for these two paths.
- Likewise, when there are two nearly identical latent variables, and these two are used as causes of a third latent variable, the difficulty in computing separate regression weights may well be reflected in much

larger standard errors for these paths than for other paths in the model, reflecting high multicollinearity of the two nearly identical variables.
- Likewise, the same difficulty in computing separate regression weights may well be reflected in high covariances of the parameter estimates for these paths - estimates much higher than the covariances of parameter estimates for other paths in the model.
- Another effect of the same multicollinearity syndrome may be negative error variance estimates. In the example above of two nearly identical latent variables causing a third latent variable, the variance estimate of this third variable may be negative.

2.6.2.10. Interval data are assumed

Unlike traditional path analysis, SEM explicitly models error, including error arising from use of ordinal data. Exogenous variables may be dichotomies or dummy variables, but unless special approaches are categorical, dummy variables may not be used as endogenous variables. Use of ordinal or dichotomous measurement to represent an underlying continuous variable is, of course, truncation of range and leads to attenuation of the coefficients in the correlation matrix used by SEM.

2.6.2.11. High precision

Whether data are interval or ordinal, they should have a large number of values. If variables have a very small number of values, methodological problems arise in comparing variances and covariances, which is central to SEM.

2.6.2.12. Small, random residuals

The mean of the residuals (observed minus estimated covariances) should be zero, as in regression. A well-fitting model will have small residuals. Large residuals suggest model misspecification (i.e. paths may need to be added to the model, AMOS or LISREL provide tools to help the researcher in model building based on tests of size of the residuals).

Uncorrelated error terms are assumed, as in regression, but if present and specified explicitly in the model by the researcher, correlated error may be estimated and modelled in SEM.

2.6.2.13. Uncorrelated residual error

The covariance of the predicted dependent scores and the residuals should be zero.

2.6.2.14. Multicollinearity

Complete multicollinearity is assumed to be absent, but correlation among the independents may be modelled explicitly in SEM. Complete multicollinearity will result in singular covariance matrices, on which one cannot perform certain calculations (e.g. matrix inversion) because division by zero will occur. Hence complete multicollinearity prevents a SEM solution. Also, when the correlation

between indicator variables r>=0.85, multicollinearity is considered high, and empirical under-identification may be a problem. Even when a solution is possible, high multicollinearity decreases the reliability of SEM estimates. Strategies for dealing with covariance matrices that are not positive definitely add a ridge constant, which is a weight added to the covariance matrix diagonal (the ridge) to make all numbers in the diagonal positive. However, this strategy can result in markedly different chi-square fit statistics. Other strategies include removing one or more highly correlated items to reduce multicollinearity: using different starting values, using different reference items for the metrics, using ULS rather than MLE estimation (ULS does not require a positive definite covariance matrix), replacing tetrachoric correlations with Pearsonian correlations in the input correlation matrix, and making sure to handle missing data list-wise rather than pair-wise because otherwise the result is often a non positive definite correlation matrix.

2.6.2.15. Non-zero covariances

Measures of fit compare model-implied covariances with observed covariances, measuring the improvement in fit compared to the difference between a null model with covariances as zero, on the one hand, and the observed covariances on the other. As the observed covariances approach zero, there is no "lack of fit" to explain it (the null model approaches the observed covariance matrix). More generally, "good fit" will be harder to demonstrate as the variables in the SEM model have low correlations with each other. That is, low observed correlations often will bias model chi-square and other fit measures towards indicating good fit.

2.6.2.16. Sample size

Sample size should not be small as SEM relies on tests that are sensitive to sample size, as well as to the magnitude of differences in covariance matrices. In the literature, sample sizes commonly run 200-400 for models with 10-15 indicators. With over ten variables, sample size under 200 generally means parameter estimates are unstable and significance tests lack power.

One rule of thumb found in the literature is that sample size should be at least 50 more than 8 times the number of variables in the model. Another rule of thumb is to have at least 15 cases per measured variable or indicator. The researcher should go beyond these minimum sample size recommendations, particularly when data are non-normal (skewed, kurtotic) or incomplete. Note also that to compute the asymptotic covariance matrix, one needs k(k+1)/2 observations, where k is the number of variables.

2.6.3 Dataset and research problem

Many expectations are connected with new technical developments, both from the safety side and from the consumer side. SARTRE 3 will yield data that tells us about the acceptance of various systems and also how realistic the drivers will perceive the effects of such systems. This is of great importance as new features in road traffic may change the perception of risk and safety; this know-how is important for designing measures to counteract wrong safety beliefs. We will use data from the SARTRE 3 survey to investigate if there are any factors

that support the acceptance and use of safety relevant systems, which might even restrict some freedom of the drivers. Acceptance of new technologies, driving experience, nationality, profession and economic status will be relevant factors of special interest. A multivariate SEM analysis was applied to take the complex relationship of these factors into account.

The aim of this is to describe how characteristics of the drivers and characteristics of specific technologies are related. When considering the introduction of new measures in traffic it is important to know if different types of drivers will react in a different way to these changes, or if there will be a common effect. This issue also applies to the introduction of new technologies. Still, the qualities of new technologies are also quite different from a psychological perspective.

Therefore the analysis undertaken distinguishes three different aspects of drivers and three different aspects of new technologies:

Driver (User) characteristics
- Emotional driving
- Professional car use
- Socio-economic characteristics

These three aspects have been extracted by principal component analysis from the SARTRE 3 questionnaire data and can shortly be described as follows:

Emotional driving covers a mix of driving habits and feelings when driving. Professional car use is a description of exposure characteristics. Emotional driving and professional car use are dimensions that are related to some extent. Socio-economic characteristics bring in another dimension, which is more or less independent from the other dimensions.

Technology characteristics (benefits)
- Assistance and guidance systems
- Warning and intervention systems
- Enforcement systems

LISREL was used (software AMOS, v5.0) for data analysis. LISREL stands for linear structural relation. By analysing the covariance matrix, the tool allows for the estimation of the weights of paths for defined models. Goodness of fit characteristics show how well the model represents the data.

The goal of this type of analysis was to aggregate data with factor analysis from many questions of the survey to a few distinct latent dimensions on the driver and on the technology side. This leads to a reduction of effect parameters to a manageable size. The relations between the factors – called the structural equation model in LISREL terms – can then be interpreted as an underlying, inner structure between driver and technology characteristics.

2.6.4 Model fit diagnostics and interpretation

In practice, the multilevel software available at this point in time offers only limited possibilities to estimate structural equation models. On the contrary, LISREL, which is the most appropriate software for structural equation modelling, does not allow the inclusion of multiple levels. Therefore, the model presented in the following research example, is not really a dedicated multilevel analysis. To illustrate the consequences of a multilevel structure, a two-step analysis was conducted:

First, data from all available 23 countries was put together to find a general model that fits to all countries. In the next step, a confirmatory analysis was conducted for every single country. Various goodness of fit statistics were calculated to indicate whether the factor-structure given by the general model could be applied to the country in question. This was the case for 19 countries. For the UK and the Czech Republic, an alternative model with extrapolated missing cases produced an acceptable fit. For four countries, the given factor structure did not lead to an acceptable fit. Their results are not considered in the following analysis. These countries were Belgium, Ireland, Portugal and Croatia.

In the future, especially with the newer versions of LISREL it will be possible to do real multilevel analysis of SEM models.

It is proposed that there are clearly defined relations between the six characteristics (arrows, whose weights point out the influence between factors) – the three driver characteristics and the three technology characteristics – in the following graph (Figure 2.6.1.), displayed as ellipsis.

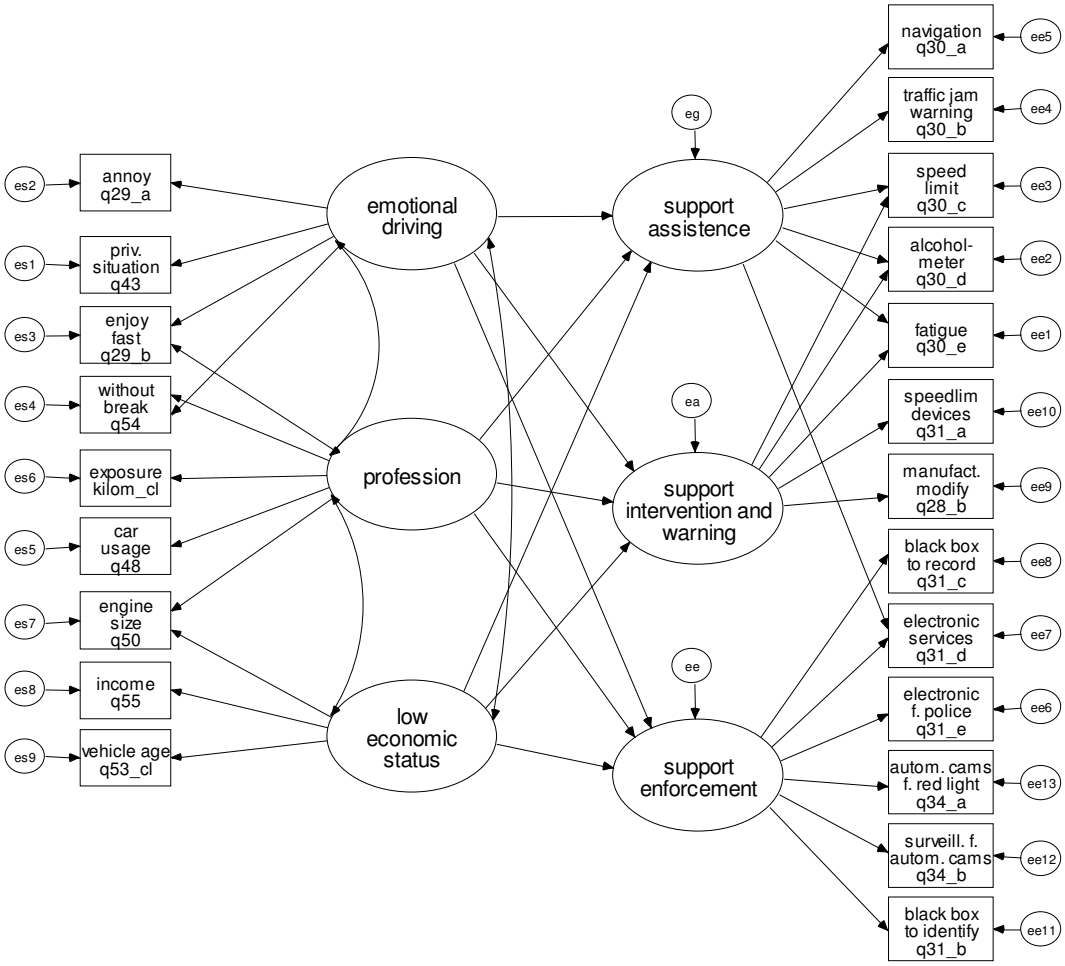


Figure 2.6.1: Proposed relations between driver and technology characteristics and questions used for operationalisation of those characteristics (short description of abbreviations/questions in the next section). Small circles represent the error terms.

These "true" dimensions are operationalised - measured by items of the SARTRE 3 questionnaire. In the graph, a set of questions is displayed on the left side; each question is presented by a box. These questions were used for measuring driver characteristics. The boxes on the right side are those that are used for distinguishing technology characteristics.

2.6.4.1. Measuring driver characteristics

There were only a few items in the questionnaire that really helped to distinguish different characteristics of drivers. We have chosen the following 10 items to identify the three proposed driver characteristics:

- ✓ *car usage (Q48: What applies most to you? I drive for my profession; I need to drive during my work; I drive to and from work)*

- ✓ *private situation (Q43: Which of the following applies best to you at the moment? Single; Living under common law marriage; Married; Separated or divorced; Widowed)*
- ✓ *How much do you agree with the following statements:*
- ✓ *annoyed by other drivers (Q29a: I sometimes get very annoyed with other drivers)*
- ✓ *enjoy driving fast (Q29b: I enjoy driving fast)*
- ✓ *driving without a break (Q54: What is the longest period of time in hours you would spend driving without taking a break?)*
- ✓ *exposure (In total about how many kilometres/miles have you driven in the last 12 months? in classes of 5,000 km)*
- ✓ *engine size (Q50: About the car you usually drive, is it a car with engine size of...? in classes of 1,000 CC)*
- ✓ *income (Q55: total annual income level per family unit)*
- ✓ *vehicle* age (Q53: How old is the vehicle you normally drive?)

2.6.4.2. Measuring technology characteristics

For distinguishing technology characteristics, we used the following items from the SARTRE 3 questionnaire:

- ✓ *manufacturers should modify their vehicles to restrict their maximum speed (Q28b)*
- ✓ *Do you find it useful to have a device like:*
  - o *navigation system (Q30a)*
  - o *congestion warning system (Q30b)*
  - o *system which prevented from exceeding the speed limit (Q30c)*
  - o *alcometer (Q30d)*
  - o *system which detected 'fatigue' (Q30e)*
- ✓ *Are you in favour of:*
  - o *speed limiting device (Q31a: Speed limiting devices fitted to cars that prevented drivers exceeding the speed limit)*
  - o *black box to record...speeding (Q31c)*
  - o *black box to identify...accident causes (Q31b)*
  - o *electronic identification to give access to services (Q31d)*
  - o *electronic identification for police enforcement (Q31e)*
  - o *cameras for red light enforcement (Q34a)*
  - o *speed cameras (Q34b)*

The results for the measurement model of the driver characteristics (left side of Figure 2.6.1.) and technology characteristics (right side of Figure 2.6.1.) are collected in Table 2.6.1.:

driver characteristics	Mean	StdDev
annoyed (q29_a) <--- emotional driving	-0,2	0,2
enjoy fast (q29_b) <--- emotional driving	-0,5	0,2
priv. situation (q43) <--- emotional driving	-0,3	0,1
without break (q54) <--- emotional driving	0,2	0,1
without break (q54) <--- profession	0,3	0,1
exposure (kilom_cl) <--- profession	0,7	0,1
enjoy fast (q29_b) <--- profession	-0,2	0,2
car usage (q48) <--- profession	-0,6	0,2
engine size (q50) <--- profession	0,3	0,3
engine size (q50) <--- low economic status	-0,2	0,3
income (q55) <--- low economic status	-0,4	0,1
vehicle age (q53_cl) <--- low economic status	0,2	0,1

technology characteristics	Mean	StdDev
navigation (q30_a) <--- assistance & guidance	-0,7	0,1
traffic jam warning (q30_b) <--- assistance & guidance	-0,8	0,0
speed delimiter (q30_c) <--- assistance & guidance	-0,2	0,1
alcohol meter (q30_d) <--- assistance & guidance	-0,3	0,1
fatigue (q30_e) <--- assistance & guidance	-0,3	0,1
electronic services (q31_d) <--- assistance & guidance	-0,2	0,1
speed delimeiter (q30_c) <--- warning & intervention	-0,7	0,1
alcohol meter (q30_d) <--- warning & intervention	-0,3	0,1
fatigue (q30_e) <--- warning & intervention	-0,4	0,1
speedlim. device (q31_a) <--- warning & intervention	-0,9	0,0
manufact. modify (q28_b) <--- warning & intervention	-0,5	0,2
black box to record (q31_c) <--- enforcement	-0,7	0,0
electronic services (q31_d) <--- enforcement	-0,4	0,1
electronic serv. for police (q31_e) <--- enforcement	-0,7	0,1
autom. cams f. red light (q34_a) <--- enforcement	-0,4	0,1
surveill. f. autom. cams (q34_b) <--- enforcement	-0,6	0,1
black box to identify (q31_b) <--- enforcement	-0,6	0,1

Table 2.6.1.: Mean factor loadings and standard deviations for the general model. For technology characteristics, high negative values indicate higher support. For driver characteristics, high negative values, i.e. q29a,b, indicate more emotional driving, higher positive values in exposure more profession.

The dimension *assistance and guidance systems* represents, with high weights, the support for navigation (0.7) and congestion warning with 0.8. But this dimension also represents systems that were previously classified in the technologies “that impose behaviours” - alcohol meter and fatigue warning (0.3) and speed limiting device and electronic services (0.2).

The dimension *Support for warning and intervention* largely represents the previous classification of systems that impose behaviour. It represents the questions about the usefulness of speed limiting devices (0.7), alcohol meter (0.3), and fatigue warning (0.4). These variables are also considered in the dimension assistance and guidance systems. Furthermore, the answers are represented in the dimension if speed-limiting devices (0.9) are favoured, and if

car manufacturers should modify their vehicles to restrict their maximum speed (0.5).

Support for enforcement systems, the third dimension, corresponds with the previously used classification of enforcement systems. It represents the questions about black box to record drivers' behaviour (0.7) or to identify what caused an accident (0.6), electronic identification to give access to services (0.4; also in dimension assistance and guidance) and electronic identification for enforcement by the police (0.7). Also, the questions about automated cameras for red light surveillance (0.4) and speed excess (0.6) have been taken into account.

In the central, structural part of the model, all dimensions between the driver and the technology part are connected to each other. Due to technical, LISREL-specific reasons, the driver characteristics relate to each other by covariance. While the covariance values between emotional driving and profession (0.1) and economic status (0.0) are low, the interrelation between profession and low economic status are weighted higher by -0.6.

Compared to the outer parts of the model, which consist of factor weights from specific questions, dimensions behave almost stable over different countries. There is little variation in driver characteristics and even less variation in technology characteristics (see Table 2.6.1.); much more variation could be found in the central part of the model. These findings were taken into consideration in the following part of this report, which takes the structure between drivers and technology as a starting point.

Overall, the main results in the structural pattern for all technological systems are:

- Low economic status drivers are most supportive,
- Professional drivers are also supportive, though less so than the above group, and
- Emotional drivers do not support new technologies (except assistance and guidance systems).

Driver characteristics derived from various variables by principal component analysis are interrelated in the following way: The covariance between low economic status and professional driving (mean -0.6 for general) is very high in Cyprus (0.8). Emotional driving and profession (mean 0.1) are highly interrelated in France, Spain and the UK. Low relations can be found in Germany and Slovakia. Low economic status and emotional driving do not show any coherence in the general model (0.0). Above-mean values can be found in Greece, the Netherlands and Finland. Poland and the UK have below mean values.

If we take a closer look at similarities in driver characteristics between countries, emotional drivers show, in general, similar patterns in France and Spain (Table 2.6.2.). Neither supports any new technology. In contrast, the support of new technologies from Polish and Slovakian emotional drivers lies clearly above the average, whose support is even at the highest level.

	austria	cyprus	czech	denmark	estonia	finland	france	germany	greece	hungary	italy	netherlands	poland	slovakia	slovenia	spain	sweden	uk	switzerland	mean
enforcement <--- low economic status		–				+	+	–			–							+	+	1,0
warning & intervention <--- low economic status		–				+		–				+				–		+		1,0
assistance <--- low economic status		–					–	+					+	+					+	0,8
enforcement <--- profession		– –					+ +	– –	– –							+ +		+ +		0,7
warning & intervention <--- profession		–						–	–				–					+	+	0,6
assistance <--- profession		– –												+ +		+ +			– –	0,6
enforcement <--- emotional driving							– –						+ +	+ +		– –			– –	-0,5
warning & intervention <--- emotional driving			+ +			– –	– –						+ +	+ +		– –				-0,6
assistance <--- emotional driving							– –		– –				+ +	+ +		– –			+ +	0,2
goodness of fit (chi-square/df)	3,19	4,02	3,93	2,63	4,85	3,13	2,60	3,67	3,41	2,15	3,13	3,66	3,06	3,71	3,53	4,14	2,78	3,37	3,96	

Table 2.6.2.: Weight differences in the structural part of the model for 19 countries in comparison to the general model. The '+' symbol stands for higher support, '-' for lower support, where a difference in standard deviation can be found. If standard deviation is higher than 0.5, '++' and '- -' are used instead. The highest values are marked in orange; the lowest values are marked in blue. Means of weights for the general model can be found in the last column on the right hand side, goodness of fit statistics in the bottom row.

Another distinct pattern can be found for drivers characterised by low economic status. In Finland and the UK, there is high support for warning and intervention systems as well as enforcement systems in this driver group.

Cyprus and Germany often show similar patterns: The low economic status group and the professional drivers group do not support new technology systems. A possible explanation could be that Cypriot drivers' scepticism concerning new technologies might be affected by the fact that these technologies are not easily affordable in their country. In contrast, German drivers' expectations might have been scaled down due to experience. There are, however, many differences in driver characteristics in both countries, hence these results do not support the "saturation effect" hypothesis. To conclude, because the differences regarding driver mentalities between these two countries seem to be very decisive, the experience effect cannot easily be separated.

Nevertheless, there are still some arguments for the "saturation by experience effect". Many traffic experts see Germany as a prime example for the spread of traffic-related new technologies. German drivers have similar characteristics to the general model and they show the highest saturation effect. Cypriot driver characteristics show that prestige plays an important role. Furthermore, the strong support from the low economic status group reinforces the saturation hypothesis: The less affordable these systems are, the higher expectations are.

In conclusion, a short summary of the application of structural equation models is introduced using the relationship of driver characteristics and their acceptance of new technologies in traffic.

For this analysis we have used a LISREL model, which led to an acceptable fit for 19 countries. With this method, it was possible to carry out a detailed analysis about support for different characteristics of new technologies in relation to different driver characteristics.

Drivers were characterised by dimensions of "emotional driving", "professional driving" and drivers with "low economic status". For new technologies, the dimensions were distinguished between for "assistance/guidance systems", "warning/intervention systems" and "enforcement systems".

Three main results in driver characteristics can be seen regarding support of new technologies:

- Low economic status drivers are most supportive of all new technologies, with their highest support for warning and interventions systems, as well as for enforcement systems.
- Professional drivers are also supportive, although in general they are less supportive than the low economic status group. This group shows the highest support for enforcement systems and slightly lower support for assistance/guidance and warning/intervention systems.
- Emotional drivers do not support new technologies (except moderate support for assistance/guidance systems).

2.6.5 Conclusion

Structural equation modelling offers one of the most complex data analyses in multivariate research methods. It connects confirmatory factor analysis with linear regression, creating a latent structure of the analysis. Hypothetical constructs are taken as latent variables in this approach.

On one hand, this chapter shows the basic form of such models in the multilevel case, dealing mainly with assumptions on data. On the other hand, this chapter discusses the necessary theoretical concepts of these models.

Analysis with structural equation models places high requirements on data. The requirements depend on the selected method of estimation of the unknown parameters. Assumptions can be divided into general conditions and statistical conditions. General assumptions consist of: the relationships between the variables is linear, the effects of explanations on dependant variables is additive, the relationship between the variables is stochastic. The most important statistical assumptions are: the variables have to be measured continuous and are interval-scaled, and they can be represented by the mean, variance and covariance which is known as a multivariate normal distribution.

At first these models seem ideal to use with a large variety of data but in practice they turn out to be difficult to implement. One is generally successful if data collection is carried out with a theoretically-based structural equation model already in mind. These models are not appropriate for use with exploratory approaches.

2.7 More complex data structures

Eleonora Papadimitriou, Constantinos Antoniou, George Yannis (NTUA)

2.7.1 Introduction

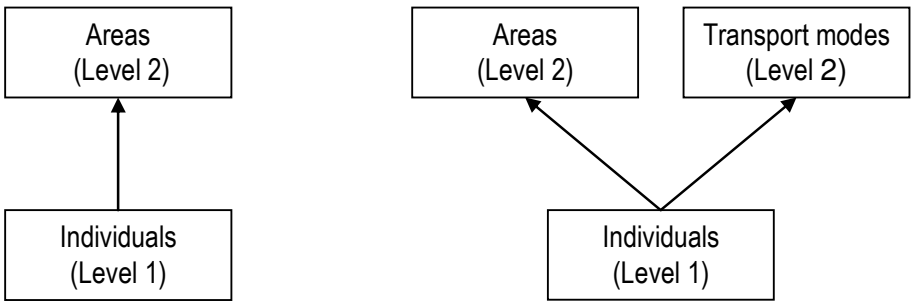
In the previous sections the concepts of multilevel modelling were introduced and it was shown how to develop simple models under Normal distribution assumptions for hierarchical data structures in the context of transport and road safety. It was demonstrated how multilevel models can be applied in the framework of generalized linear modelling, i.e. under non Normal distributional assumptions. Moreover, more advanced multilevel models were presented, including multivariate models, factor analysis and structural equations models. In these sections the emphasis was on the theoretical background, the models assumptions and the interpretation of results, by means of modelling examples.

It was shown that the motivation for multilevel modelling in road safety analysis is that the processes we wish to model often take place in the context of a hierarchical structure (Rasbash et al., 2000), each level of this hierarchy contributing to a random variation of the variable of interest. Accordingly, all the examples presented in the previous sections concerned classical hierarchical data structures e.g. accidents and fatalities nested into regions, speed measurements nested into different road sites etc. However, the assumption that the structures we wish to model are purely hierarchical is often an over-simplification (Rasbash et al., 2000). Individuals or cases may be classified according to more than one group at a given higher level of a hierarchy (cross-classification) and each group can be a source of random variation. For example, in a mobility analysis, individuals may be classified according to the transport mode they use and the area they live, while each area may include all transport modes and each transport mode may serve all areas. Moreover, individuals or cases may belong to more than one sub-groups of the higher level group (multiple memberships). For example, in a longitudinal study, individuals may change area and finally belong to more than one area in the study. These special cases of hierarchical data structure and the resulting multilevel models, often referred to as "non-hierarchical" multilevel models (Browne et al., 2001), are described in the following sections.

2.7.2 Cross - classified data

An example of cross-classified hierarchical structure in the context of road safety may be the following: within a mobility survey, both the area type an individual lives in and the travel mode an individual uses have an important effect on mobility. Therefore, there are two possible higher level classifications of the individuals examined in the survey and these classifications are not mutually exclusive.

While a classical nested multilevel structure can be described as in Figure 2.7.1.A, a crossed multilevel structure can be described as in Figure 2.7.1.B.



A. Nested multilevel structure B. Crossed multilevel structure

Figure 2.7.1. Nested and crossed multilevel structures

In this case, however, each area includes individuals using different transport modes and each transport mode also includes individuals from different areas. Consequently, not only are there two different higher level classifications of the individuals, but also these two classifications are not mutually exclusive. No pure hierarchy can be found and individuals are contained within a cross-classification of transport modes by areas, as shown in Figure 2.7.2.

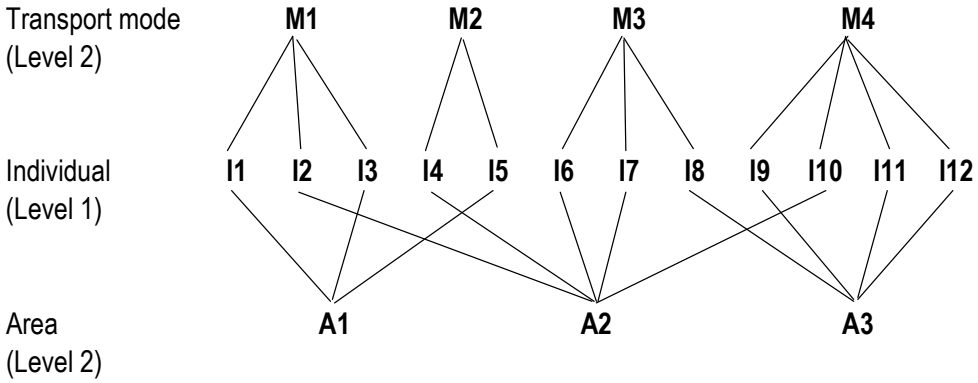


Figure 2.7.2. Cross-classification of individuals within transport modes and areas

It is obvious that individuals can be sorted by transport modes within areas or areas within transport modes, but not both. The consequences of ignoring an important cross-classification are similar to those of ignoring an important hierarchical classification (Rasbash et al. 2000). A simple model describing this situation can be formulated as:

$$y_{i(jk)} = \alpha + u_j + u_k + e_{i(jk)} \qquad (2.7.1)$$

Where

$y_{i(jk)}$ is the mobility of the i^{th} individual from the $(jk)^{th}$ area type / transport mode combination
α is the overall mean
u_j is a random departure due to transport mode j
u_k is a random departure due to area k
$e_{i(jk)}$ is an individual level random departure

Obviously, the model can be further elaborated by adding level-1 explanatory variables, whose coefficients may vary across areas or modes. Also, level-2 variables can be incorporated to explain variation across areas or modes.

Starting from the above baseline model, more complex models can also be formulated, which may include multiple cross-classifications, and/or mixed nested-and-crossed structures. The following example shows the different structures and the related formulations of the multilevel equations, which can be considered according to the specifications of the problem: Within a mobility survey, individuals (i) are interviewed by interviewers (j); the individuals come from (k) transport modes and (l) area types.

If each individual is interviewed by a more than one interviewers (in case, for instance, that a survey has more than one questionnaire and each questionnaire is processed by a different interviewer), there is an individual / interviewer cross-classification at Level 1. Moreover, if a different set of interviewers operates in each transport mode, the Level-1 individual/interviewer cross-classification is nested within transport mode at level 2. A model describing this situation can be formulated as:

$$Y_{(ij)k} = \alpha + u_k + e_{ik} + e_{jk} \qquad (2.7.2)$$

The interviewer and interviewee (individual) effects are modelled by the level-1 random variables e_{ik} and e_{jk}, while the transport mode random effects are modelled by the level-2 random departure u_k (note that the area type effects are not considered in this model). It should be noted that, in such a model, the cross-classification does not need to be balanced i.e. some individuals may not be interviewed by all the interviewers.

If each individual is interviewed by only one interviewer and the same set of interviewers is used for all transport modes, interviewers are cross-classified with transport modes. An equation such as (2.7.1) can be used to model this situation (in this case though, k would refer to interviewers rather than areas). If transport modes are also crossed by area types, then individuals are nested within a three-way interviewer/transport mode/area type classification. In this case, equation (1) can be extended by adding a term u_l for the interviewer classification:

$$yi(jkl) = \alpha + uj + uk + ul + ei(jkl) \qquad (2.7.3)$$

Where now *i* refers to individuals, *j* refers to interviewers, *k* refers to transport mode, and *l* refers to area type.

Furthermore, if interviewers are not crossed with transport modes (i.e. a different set of interviewers is used in each transport mode), but transport modes are crossed with areas (i.e. the same transport mode is used in different areas), the formulation would become:

$$y_{i(jkl)} = \alpha + u_k + u_l + e_{i(kl)} + e_{j(kl)} \qquad (2.7.4)$$

It is obvious that, according to the context of the problem, different structures can be considered; a cross-classification may be present at any level of the hierarchy, from the lowest (equation 2.7.2) to the highest level (equation 2.7.1). Moreover, a cross-classification may be multiple (equation 2.7.3). Finally, a higher level classification may include one cross-classification and one simple higher level classification (equation 2.7.4). In any case, a related multilevel formulation is available.

In order to fit a cross-classification multilevel model, a special procedure is required. For instance, in a level 2 cross-classification with 10 transport modes drawing individuals from 30 areas (as in equation 1), if the data is sorted by transport mode and the cross-classification with areas is ignored, the transport modes impose a block-diagonal structure[38] on the N by N covariance matrix of responses, where N is the number of individuals in the data set. In order to account for the cross-classification of transport modes and areas, a non-block-diagonal covariance structure needs to be estimated (Rasbash et al. 2000).

This can be achieved by setting a third (higher) level in the model. First, a "constant" variable is created, with one unit value which covers the entire data set, and this variable is declared as the third level. Then, thirty dummy variables are created, one for each area, and their coefficients are set to vary randomly at level 3, with a separate variance for each of the 30 area. Finally, all 30 variances are constrained to be equal (Rasbash et al, 2000). This constraint is necessary in order to obtain one common estimate of the level-3 variance. Other coefficients can be set to vary randomly across modes, as level-2 random parameters, in the usual way.

However, if a coefficient of a slope is set to vary randomly across areas, the procedure becomes more complicated, as thirty additional variables need to be created; these would be obtained as the product of the area dummy variables and the examined slope. The new variables are set to vary randomly at level 3, with the respective equality constraint in order to obtain a common estimate. Furthermore, in order to examine the covariance between intercept and slope, 90 random parameters are needed at level 3: an intercept variance, a slope variance and an intercept/slope covariance for each of the 30 areas, each set of

[38] A block matrix is a matrix that is defined using smaller matrices, called blocks. A block diagonal matrix is a block square matrix, having main diagonal blocks square matrices, such that the off-diagonal blocks are zero matrices.

them (30 intercept variances, 30 covariances and 30 slope variances) constrained to produce 3 common (level-3) estimates, and so on.

It should be noted that, although a 3-level structure is defined, conceptually only a 2-level model is considered, in which transport mode and area are crossed at level 2. The third level is only used as a tool to convert the crossed structure into a nested structure and allow for estimation of the crossed structure (Rasbash, Goldstein, 1994).

2.7.3 Multiple membership models

Multiple membership models refer to a situation where, in a 2-level model for instance, level-1 units belong to two or more level-2 units. Thus, for example, in a longitudinal study, some individuals (i) may change region and may finally "belong" to more than one region (j) during the study. This kind of classification is graphically presented with a double arrow, as in Figure 2.7.3:

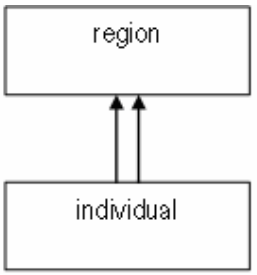


Figure 2.7.3. Multiple membership structure

When modelling such data, level-2 effects are shared between all the units (regions) to which an individual belongs. It is therefore necessary to allocate a set of weights for each individual to attach to these units (Rasbash et al, 2000). First, it is assumed that an individual belongs to more than one region and this set of regions is defined as j_2. If the weight π_{ij2}, associated with the j_2^{th} region for individual i, is known, (e.g. the proportion of time spent in that region) with:

$$\sum_{j2=1}^{J2} \pi_{ij2} = 1$$

a simple variance components model can be formulated as:

$$y_{i(j2)} = (X\beta)_{i(j2)} + \sum_{j2} u_{j2}^{(2)} \pi_{ij2} + e_{i(j2)} \qquad (2.7.5)$$

$$\sum_{j2} u_{j2}^{(2)} \pi_{ij2} = \pi_i u^{(2)}$$

$$u^{(2)T} = \{ u_1^{(2)}, \ldots, u_{j2}^{(2)} \}$$

$$\pi = \{\pi_1, \ldots, \pi_{j2}\}$$

$$\pi_{j2}^T = \{ \pi_{1j2}, \ldots, \pi_{Nj2} \}$$

$$\text{Var}\,(u_{j2}^{(2)}) = \sigma_{u2}^2, \qquad Cov\,(u_{j1}^{(1)}, u_{j2}^{(2)}) = 0, \qquad Var\,(\sum_{j2} u_{j2}^{(2)}\, \pi_{ij2}) = \sigma_{u2}^2 \sum_{j2} \pi_{ij2}$$

In the above formulae, N is the total number of individuals and $u^{(2)}$ is the (J_2 x *1)* vector of the regions j_2 effects. This is therefore a 2-level model, in which the level 2 variation among regions is modelled using j_2 sets of weights for individual *i* ($\pi_{i1}, \ldots, \pi_{iJ2}$) as explanatory variables, with π_{j2} the *(N x 1)* vector of individuals weights for the j_2^{th} region.

For a basic example, we may consider five individuals and three regions according to the weights of Table 2.7.1 (proportion of time spent in each region):

	Region 1 (j=1)	Region 2 (j=2)	Region 3 (j=3)
Individual 1 (i=1)	0.5	0	0.5
Individual 2 (i=2)	1	0	0
Individual 3 (i=3)	1	0	0
Individual 4 (i=4)	1	0	0
Individual 5 (i=5)	0	0.25	0.75

Table 2.7.1. Multiple membership weights

In this case, individual 1 spent half of his time in region 1 and half of his time in region 3, and individual 5 spent 25% of his time in region 2 and 75% of his time in region 3. If we use these weights into formula (2.7.5), we obtain the following set of formulae:

$$Y_1 = X\beta + 0.5\, u_1^{(2)} + 0.5\, u_3^{(2)} + e_i$$
$$Y_2 = X\beta + u_1^{(2)} + e_i$$
$$Y_3 = X\beta + u_1^{(2)} + e_i$$
$$Y_4 = X\beta + u_1^{(2)} + e_i Y_5 = X\beta + 0.25\, u_2^{(2)} + 0.75\, u_3^{(2)} + e_i$$

The above example, where an individual sequentially moves from one region to another is the most frequent case of multiple membership. However, it may also be the case that individuals alternate between regions and may be considered as simultaneously belonging to more than one region. This case can also be dealt with, by using weights which reflect time spent within each region (Goldstein et al. 2000).

In order to fit a multiple membership model, a process similar to the one used for the cross-classified models is adopted. Considering the above example, in which individuals change regions over time, it is necessary to create a set of variables to attribute the weights corresponding to each region for each individual. As in cross-classified models, level-2 is defined by a "constant" unit value variable, which covers the entire dataset. Then, a set of weighted

indicator variables are created, showing the proportion of time spent in each region by each individual. For example, if there are 20 regions, 20 new variables will be created, one for each region, providing the proportion of time spent in each region by each individual. These weighted indicator variables are set to randomly vary at level-2. In order to obtain one single variance estimate for all regions, an equality constraint is imposed for the variances of the 20 regions (Browne et al. 2001).

It is obvious that, technically, this two-level structure is different from the classical one; the higher level is not defined by a "real" variable (but from a "constant" variable) and the higher level variation is not estimated on this higher level itself (but obtained from a set of variances constrained to be equal). However, this structure allows for an efficient estimation of higher-level variance in multiple membership models. For details of these models and examples see also Hill and Goldstein (1997), Browne et al. (2001).

An interesting sub-case of multiple membership models is the case of spatial modelling with neighbourhood matrix. In section 2.3.4.6, the applications of ecological and aggregate spatial analysis in road safety research were briefly presented. In some of these applications (e.g. MacNab, 2004) spatial variability is expressed on the basis of neighbourhood (instead of e.g. distance, as in other studies). The multiple membership structure arises from the fact that each unit of analysis (e.g. county) neighbours with more than one other unit, leading to a neighbourhood matrix in which each diagonal element is equal to the number of neighbours of the corresponding area, and the off-diagonal elements in each row are equal to –1 if the corresponding areas are neighbours and 0 otherwise, allowing to weight the data accordingly. These models are mainly fitted by means of Bayesian approaches (see chapter 2.8).

2.7.4 Summary

The multilevel models described in the previous Chapters of this document were proved to be capable of dealing with a wide variety of hierarchical data structures within the context of road safety analysis (accidents analysis, road users' behaviour, monitoring of road safety measures etc.). These models allow for both continuous and discrete responses to be modelled, as well as for different levels of hierarchies to be considered (spatial, qualitative etc.). They can also handle multiple responses (multinomial responses or multivariate analyses), as well as longitudinal data (e.g. repeated measurements).

In this section, cases of data having a structure which is not purely hierarchical were briefly presented. It was shown that level-1 units may be clustered not only into hierarchically ordered units (e.g., individuals nested within regions, within countries etc.), but may also belong to more than one type of unit at a given level of a hierarchy (cross-classification). Moreover, individuals may belong to more than one sub-groups of a given higher level group (multiple memberships). It was shown that multilevel models can be extended to handle

such complex "non-hierarchical" data structures, and the basic formulation of these models was presented by means of simple indicative examples.

These models allow for such complex structures to be defined and explored, which would be a difficult or impossible task with conventional techniques. It should be noted, however, that in practice difficulties may be encountered when fitting such complex models, both in terms of obtaining satisfactory numerical convergence and interpreting results (Goldstein et al. 2000). In particular, conventional estimation methods like maximum likelihood or quasi-likelihood, which exploit the nested structure of the data in multilevel hierarchical models, are not efficient in this case. As these two types of structures are not purely nested, they need to be converted into nested (purely hierarchical) structures, with a set of constraints reflecting the particularities of the structure (Browne et al. 2001). More advanced (simulation-based) estimation methods, which are presented in the next section, apart from their other advantages compared to the default estimation methods, are also more powerful in dealing with these complex structures.

2.8 Bayesian estimation in multilevel modelling

Eleonora Papadimitriou, Constantinos Antoniou, George Yannis (NTUA)

2.8.1 General

In all the models presented in the previous sections, conventional default estimation methods were used in the modelling process and little or no mention was given to alternative approaches to fitting multilevel models. These default estimation methods are either maximum likelihood or some approximation of maximum likelihood (e.g. quasi-likelihood), which are based on Generalized Least Squares (GLS) estimation. In the present document, maximum likelihood values were used for Normal models and quasi-likelihood methods were used for generalized linear models, according to the common practice (Browne et al. 2001).

However, it was mentioned that an important problem rises from the use of approximation methods; the estimated likelihood ratio is very approximate and can not be used for the assessment of models fit. Moreover, when default methods are applied to more complex data structures, such as the "non-hierarchical" structures mentioned above, numerical and convergence difficulties are often encountered.

In this section, a group of alternative estimation methods for multilevel models are described, namely the Markov Chain Monte Carlo (MCMC) and the bootstrap methods. These advanced estimation methods are both based on simulation techniques and the estimates they produce are dependent on randomly generated numbers (Rasbash et al. 2000). In contrast to the default estimation methods, where a single estimate (described by a mean and a variance) for a parameter is obtained by a single sample, these simulation methods generate a large number of samples from the initial sample, and yield thus a sample of means and a sample of variances, allowing for the calculation of intervals for parameter estimates. For this reason, they are also able to provide accurate likelihood statistics.

More specifically, a single sample gives one estimate for the mean and one estimate for the variance of each parameter. Obviously, the larger the sample size, the more accurate the mean estimate will be. Accordingly, if a sample of means estimates and a sample of variances estimates could be available, interval estimates for the parameters could be calculated. This idea of generating a large number of samples to create interval estimates is the motivation behind most simulation methods (Rasbash et al. 2000). In the following sections two groups of simulation methods that can be used in multilevel modelling, namely MCMC methods and bootstrap methods, are presented.

2.8.2 MCMC methods and Bayesian modelling

In this section, the aim is to provide some background for understanding the general concepts behind Bayesian statistics (Barnett, 1999) and MCMC - Monte Carlo Markov Chain methods (see e.g. Casella and George, 1992, Smith and Gelfland, 1992), as well as their use in the context of multilevel analysis.

The motivation for MCMC comes from the need to obtain accurate statistics (such as point estimates and confidence intervals) with small samples.

The Generalized Least Squares methods (IGLS - Iterative Generalized Least Squares and RIGLS - Restricted Iterative Generalized Least Squares) were considered and used in the previous sections of this document in order to parameter estimates. As the random variables were assumed to have a multivariate Normal distribution at each level, IGLS gave maximum likelihood estimates and RIGLS gave restricted maximum likelihood estimates. These methods are based on iterative procedures and the process involves iterating until two consecutive estimates for each parameter are sufficiently close together and hence convergence has been achieved. These methods give point estimates for all parameters of the model, estimates of the parameter standard errors and large sample hypothesis tests and confidence intervals (Rasbash et al., 2000).

Markov chains (or processes) are a way of representing multi-state stochastic systems, whose states (discrete or continuous) are defined by a transition probability. In a Markov chain of order n, the current state depends on the n previous states. For example, in the most commonly used 1^{st} order Markov chain, the state only depends on the previous state. A Markov chain can be represented by a transition matrix, with the (i,j) cell representing the transition probability that the current state will be (j) given that the previous state was (i). Monte Carlo is used to describe sampling techniques that are based on random variables (equivalent to draws of a fair dice).

MCMC is a general technique for the generation of fair samples from a probability distribution using random numbers from uniform probabilities. The idea behind MCMC is to draw a sample from the full posterior distribution and make inferences using the sample (instead of the posterior distribution). For example, instead of computing the mean and variance of a parameter of a distribution, the sample mean and sample variance of the parameter is calculated from the sample. A posterior distribution of a parameter can be obtained by a histogram/empirical density function of the distribution of the parameter in the sample (Rasbash et al., 2000).

MCMC is particularly interesting in the context of Bayesian statistics (Barnett, 1999). The simple Bayes rule dictates that the posterior is equal to the prior times the likelihood of available data:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \qquad (2.8.1)$$

where P(A) and P(B) are the prior (or marginal) distributions of A and B respectively, and P(B|A) (respectively P(A|B)) is the posterior (or conditional) distribution of B given A (respectively A given B).

A simple variance components multilevel model can be written as follows:

$$y_{ij} = \beta_{0ij}\, x_0 + \beta_1\, x_{1ij} \qquad (2.8.2)$$

$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$$

$$u_{0j} \sim N(0,\sigma_{u0}^2)$$

$$e_{0ij} \sim N(0,\sigma_{e0}^2)$$

In a Bayesian formulation of this model, prior information about the fixed and random parameters, β_0, β_1, σ_{u0}^2, σ_{e0}^2, are combined with the data (Rasbash et al., 2000). These parameters are regarded as random variables described by probability distributions, and the prior information for a parameter is incorporated into the model via a prior distribution. After fitting the model, a posterior distribution is produced for the above parameters, which combines the prior information with the data. MCMC methods make a large number of simulated random draws from the joint posterior distribution of all the parameters, and use these random draws to provide a summary of the underlying distribution. From the random draws of the parameter, it is then possible to calculate the posterior mean and standard deviation, as well as density plots of the complete posterior distribution.

It should be noted that, in Bayesian statistics, every unknown parameter must have a prior distribution, describing all information known about the parameter prior to data collection. Often little is known about the parameters a priori, and so default prior distributions are required to overcome this lack of knowledge. The most natural distribution for this application is the conceptual equivalent of a uniform distribution, i.e. a distribution that assumes that all states have equal probability of occurring or, in other words, that a parameter has the same probability of taking each value. These rather uninformative priors are sometimes called diffuse or vague priors.

Multilevel models contain many unknown parameters and the objective of MCMC estimation of these models is to generate a sample of points in the space defined by the joint posterior distribution of these parameters. In the Normal variance components model, this consists of generating samples from the distribution

$P(\beta_0, \beta_1, u_0, \sigma_{u0}^2, \sigma_{e0}^2 \mid y)$, where u_0 is the vector of u_{0j}'s.

Unfortunately, to calculate this distribution directly would involve integrating many parameters, which can be extremely complicated; however, an alternative

approach is available. This is due to the fact that although the joint posterior distribution is difficult to simulate from, the conditional posterior distributions for each of the unknown parameters often have forms that can be simulated from easily (Rasbash et al., 2000).

MCMC is not a new concept. The ideas have actually been around for several decades. However, as these techniques are computationally very demanding, widespread use followed the emergence of computing. These techniques have been in widespread use for some 15 years and have made Bayesian statistics more practical and accessible to researchers and practitioners. Metropolis-Hastings sampling is based on the seminal paper by Metropolis et al. (1953), which was later expanded by Hastings (1970). Gibbs sampling was first described in Geman and Geman (1984). The name Gibbs is associated with statistician J. Willard Gibbs (1839-1903). Even though Metropolis-Hastings sampling precedes Gibbs sampling, Gibbs sampling is the simpler and more easily implemented sampling method for MCMC.

A) The Gibbs Sampling method

Gibbs sampling works by simulating a new value for each parameter in turn from its conditional distribution, assuming that the current values for the other parameters are the true values. For example, in the Normal variance components model, the parameters and level 2 residuals would be split up into 4 subsets: β, u_0, σ_{u0}^2, and σ_{e0}^2, where $\beta = (\beta_0, \beta_1)$.[39]

Firstly, it is necessary to choose starting values for each set of parameters, $\beta(0)$, $u_0(0)$, $\sigma_{u0}^2(0)$, and $\sigma_{e0}^2(0)$. These can be taken from fitting a multilevel model with the standard estimation methods before MCMC estimation is applied. In fact, it is common practice to use IGLS or RIGLS methods before using MCMC estimation, in order obtain good starting values. The method then works by sampling from the following conditional posterior distributions, firstly

- $P(\beta \mid y, u_0(0), \sigma_{u0}^2(0), \sigma_{e0}^2(0))$ to generate $\beta(1)$, and then from
- $P(u_0 \mid y, \beta(1), \sigma_{u0}^2(0), \sigma_{e0}^2(0))$ to generate $u_0(1)$, and then from
- $P(\sigma_{u0}^2 \mid y, \beta(1), u(1), \sigma_{e0}^2(0))$ to generate $\sigma_{u0}^2(1)$, and then from
- $P(\sigma_{e0}^2 \mid y, \beta(1), u_0(1), \sigma_{u0}^2(1))$ to generate $\sigma_{e0}^2(1)$.

By performing all 4 steps, all of the unknown quantities in the model are updated.

A random walk is generated from this initial point by propagating in a similar way. For k=2…n:

[39] It should be noted that, given the values of the fixed parameters and the level 2 residuals, the level 1 residuals e_{0ij} can be calculated by subtraction. Therefore, they are not included in the algorithms.

- $\beta(k) \sim P(\beta \mid y, u_0(k-1), \sigma_{u0}^2(k-1), \sigma_{e0}^2(k-1))$
- $u_0(k) \sim P(u_0 \mid y, \beta(k), \sigma_{u0}^2(k-1), \sigma_{e0}^2(k-1))$
- $\sigma_{u0}^2(k) \sim P(\sigma_{u0}^2 \mid y, \beta(k), u(k), \sigma_{e0}^2(k-1))$
- $\sigma_{e0}^2(k) \sim P(\sigma_{e0}^2 \mid y, \beta(k), u_0(k), \sigma_{u0}^2(k))$

where ~ means that the left-hand value is drawn from the right-hand distribution. When a new value is drawn, it immediately replaces the previous one and therefore only one set of values is stored at any given time. The resulting sequence is a Markov chain, as the values at the k-th step only depend at the values in the previous step. This chain tends to a stationary distribution that corresponds to the desired distribution $P(\beta_0, \beta_1, u_0, \sigma_{u0}^2, \sigma_{e0}^2 \mid y)$.

This method is very efficient when the conditional posterior distributions are easy to simulate from, as in the case for Normal models. However, when the conditional posterior distributions do not have simple forms, a second MCMC method should be considered, called Metropolis Hastings sampling.

B) The Metropolis Hastings sampling

In general MCMC estimation methods generate new values from a "proposal" distribution that determines how to choose a new parameter value given the current parameter value. As the name suggests, a "proposal" distribution suggests a new value for the parameter of interest. This new value is then either accepted as the next iteration or rejected and the current value is used as the next iteration. The Gibbs sampler, discussed above, has as its "proposal" distribution the conditional posterior distribution, and is a special case of the Metropolis Hastings sampler where every proposed value is accepted.

In the case of the Metropolis Hastings sampler, almost any distribution can be used as a "proposal" distribution. In most cases (e.g. in the MLwiN software), the Metropolis Hastings sampler uses Normal "proposal" distributions centred at the current parameter value. This is known as a random-walk proposal. This "proposal" distribution for parameter θ has the property that it is symmetric in $\theta(t-1)$ and $\theta(t)$, that is:

$$P(\theta(t) = a \mid \theta(t-1) = b) = p(\theta(t) = b \mid \theta(t-1) = a)$$

MCMC sampling with a symmetric proposal distribution is known as pure Metropolis sampling. The proposals are accepted or rejected in such a way that the chain values are indeed sampled from the joint posterior distribution (Rasbash et al., 2000).

As an example of how the method works, the procedure for the parameter β_0 at time step t in the Normal variance components model is as follows:

- Draw β_0^* from the proposal distribution $\beta_0(t) \sim N(\beta_0(t-1), \sigma_p^2)$ where σ_p^2 is the proposal distribution variance.
- Define $r_t = p(\beta_0^*, \beta_1, u_0, \sigma_{u0}^2, \sigma_{e0}^2 \mid y) / p(\beta_0(t-1), \beta_1, u_0, \sigma_{u0}^2, \sigma_{e0}^2 \mid y)$ as the posterior ratio and let $a_t = min(1, r_t)$ be the acceptance probability.

- Accept the proposal $\beta_0(t) = \beta_0^*$ with probability a_t, otherwise let $\beta_0(t) = \beta_0(t-1)$.

In this algorithm, the method either accepts the new value or rejects the new value. The difficulty with Metropolis Hastings sampling is finding a "good" proposal distribution that generates a chain with low autocorrelation. The problem is that, since the output of an MCMC algorithm is a realisation of a Markov chain, autocorrelated (rather than independent) draws from the posterior distribution are made. This autocorrelation tends to be positive, which can mean that the chain must be run for many thousands of iterations to produce accurate posterior summaries. When using the Normal proposals as above, reducing the autocorrelation to decrease the required number of iterations corresponds to finding a "good" value for the "proposal" distribution variance σ_p^2.

As the Gibbs sampling is a special case of the Metropolis Hastings sampling, it is possible to combine the two algorithms so that some parameters are updated by Gibbs sampling and other parameters by Metropolis Hastings sampling.

It should be underlined that there is a restriction on the MCMC techniques that can be used on discrete response models for a different reason. In the previous sections, where discrete response models were discussed, it was noted that we could no longer use simple maximum likelihood based techniques, but instead had to use quasi-likelihood techniques. The Normal models discussed above are a special set of models, as all the parameters in these models have conditional posterior distributions that have standard forms. This means that the standard Gibbs sampling method can be used for all parameters. For discrete response models the conditional posterior distributions for both the fixed effects and the residuals do not have standard forms and consequently Metropolis Hastings sampling must be used for these parameters.

2.8.3 Bootstrapping

Bootstrap can be used to estimate the parameters of a model and their standard errors strictly from the sample, without assuming a theoretical sampling distribution. A number of n samples are drawn with replacement from the available sample. The statistics of interest are then estimated for each of the n samples, and the observed distribution of the n statistics is used as an empirical sampling distribution, from which estimates of the expected value and the variability of the statistics of interest can be obtained. For an introduction to bootstrap, cf. e.g. Efron, 1982, Efron and Tibshirani, 1993, or Davidson and Hinkley, 1997. An overview of bootstrapping in the context of multilevel models can be found in Hox, 2002.

Bootstrapping relies on the available sample for the inference about the population statistics. Therefore, the original sample must have a reasonable sample size. Based on a review of available literature, Yung and Chan (1999)

conclude that a general recommendation for the minimum sample size required for the sample size is not possible. Good (1999) suggests a minimum sample size of 50 in the case of non-symmetric underlying distributions. Nevitt and Hancock (2001) on the other hand suggest that for accurate results despite large violations of normality assumptions, the bootstrap needs an observed sample of more than 150. The number of bootstrap iterations n is typically in the order of thousands (Booth and Sarkar, 1998, Carpenter and Bithell, 2000).

Bootstrap also has some assumptions and restrictions. A key assumption of the bootstrap is that the resampling properties of the statistic resemble the sampling properties (Stine, 1989). It is also not ideal for properties that involve only a narrow subset of observations, such as the maximum value (Stine, 1989). Another assumption that is particularly relevant to the use of bootstrap in multilevel modelling commands that the resampling scheme that is used must reflect the actual sampling mechanism used to collect the data (Carpenter and Bithell, 2000). This last property is important and must be followed so that the hierarchical sampling mechanism of multilevel models bootstrap procedure is simulated correctly.

Bootstrapping can be either based on resampling complete cases or resampling residuals (Stine, 1989, Mooney and Duvall, 1993). Resampling complete cases is perhaps the most intuitive approach, but also more difficult in practice, especially in the context of multilevel models. When sampling residuals, it is assumed that the predictor variables have exactly the same value for each case, and therefore the only difference is in the residuals. To bootstrap residuals one first needs to run a multiple regression to estimate the regression coefficients and a set of residuals. In each bootstrap iteration the fixed values of the regression coefficients are used to predict outcomes, to which bootstrapped sets of residuals are added. The resulting bootstrapped responses are used to estimate the required statistics.

Bootstrapping cases is more complicated in multilevel models because it implies bootstrapping units at all available levels. This does not only change the values of the explanatory and outcome variables, but also the way the variance is partitioned over the different levels (Hox, 2002). This redistribution of the variance affects all other estimates. Two bootstrap approaches can be used: parametric and non-parametric.

Parametric bootstrapping uses assumptions about the distribution of the data to construct the bootstrap datasets, usually the multivariate normality assumption. For instance, for a sample of 100 cases with mean μ and standard deviation of σ, parametric bootstrap would draw a large number n of samples of size 100 from a Normal (μ, σ^2) distribution. Then for each sample the parameter of interest would be calculated and used for the calculation of the statistics of the population.

Non-parametric bootstrapping does not assume a distribution for the data but instead generates a large number of datasets by sampling (with replacement) from the original sample. In the above example, lots of samples of size 100 would be generated, with replacement of the values of the initial sample. This

approach is called non-parametric, because it preserves the possibly non-normal distribution of the original data. Obviously, if the data are normally distributed, then parametric and non-parametric bootstrap would be equivalent.

Spiegelman and Gates (2005) describe a non-parametric double bootstrapping procedure for direct comparison of quantiles of two or more sample populations. The first bootstrap simulation is used to produce estimates of standard errors for the desired quantiles and thereby overcome the inability to make reasonable variance estimations. The second layer of bootstrap simulations is used to determine the threshold cut-off values based on a desired level of confidence for the test of hypothesis. The cut-off values also may be used to form confidence intervals.

In multilevel modelling, bootstrapping can be used for two main purposes (Rasbash et al., 2000). Firstly, it can be used as an alternative procedure to MCMC methods, to make accurate inferences on the basis of simulated parameter estimates. Thus, for example, while in Normal response models we can construct confidence intervals for functions of the fixed parameters assuming Normality, this may not be appropriate for the random parameters, unless the number of units at the level to which the parameter refers is large.

The bootstrapping methods are used to construct the bootstrap datasets and then the classical Generalized Least Squares estimation methods can be used to find estimates for each dataset. The parametric bootstrap works exactly as mentioned above, i.e. the datasets are generated (by simulation) based on the parameter estimates for the original dataset. Due to the multilevel structure, the simple non-parametric approach introduced above can not be used; a new approach is used, based on sampling from the estimated residuals (Rasbash et al., 2000).

The second purpose for which bootstrap estimation can be used is to correct any bias in the parameter estimation (again as an alternative to MCMC methods). This is useful in models with discrete responses, where the standard estimation procedure based upon quasi-likelihood estimation produces estimates, especially of the random parameters, that are downwardly biased when the corresponding number of units is small (Goldstein and Rasbash, 1996). The severity of this bias can be trivial in some data sets and severe in other data sets. A complicating feature in these models is that the bias is a function of the underlying "true" value so that the bias correction needs to be iterative.

The following example, presented in Rasbash et al (2000), can be considered: suppose a data set for a simple variance components model is simulated, where the standard estimation procedure has a downward bias of 20% for the variance of level 2, and the true value for the variance of level 2 equal to 1. Then if this model is estimated for several simulated datasets using the standard procedure, an average estimate of 0.8 would be obtained.

If there is just one simulated data set with a level 2 variance estimate that happens to be 0.8, together with fixed parameter estimates to which the same procedure can be applied, a large number of new response vectors from the model with level 2 variances of 0.8 can be simulated (parametrically bootstrapped), and the average of the variance estimates across these new replicates can be estimated. A value of 0.64 would be expected, since the level 2 variance is estimated with a downward bias of 20% (0.8*0.8 =0.64). If the downward bias of 0.16 is added to our starting value of 0.8, a bias corrected estimate of 0.96 would be obtained. Another set of simulations can be run then, taking the bias corrected estimates (0.96 for the variance) as the starting simulation values. Averaging across replicates, an average of 0.768 for the variance parameter would be expected, resulting in a bias estimate of 0.192. This estimated bias would then be added to 0.8 to give a new bias corrected estimate of 0.992. Another set of replicates from the latest bias corrected estimate could be then simulated; the process could be repeated until the successive corrected estimates converge[40].

2.8.4 Applications of simulation methods and Bayesian multilevel modelling in road safety

The use of Bayesian approaches to highway safety research began with the introduction of empirical Bayes (EB) into the field by Hauer and colleagues (see e.g. Persaud and Hauer, 1984, Hauer, 1986, Hauer and Persaud, 1987, Hauer et al., 1988, Hauer, 1996a, Hauer, 1996b, Hauer, 1996c, Hauer, 1997, Hauer et al, 2002a, Hauer et al., 2002b, Hauer et al., 2004). Since then, much research using EB has emerged. Over the past years, "full" Bayesian modelling in general, and MCMC methods in particular, are becoming increasingly popular, especially as computational power of recent computers makes them practical (Davis and Guan, 1996, Davis, 2000, Davis and Yang, 2001, Miaou and Lord, 2003, MacNab, 2004).

Qin et al. (2005) use crash and physical characteristics data for highway segments from several US states to investigate the relationship between crash count and traffic volume. A hierarchical Bayesian framework has been used to fit zero-inflated-Poisson regression models for predicting counts for each crash type as a function of the daily volume, segment length, speed limit and lane/shoulder width using Markov Chain Monte Carlo methods.

Carriquiry and Pawlovich (2006) discuss the basic differences between various Bayes approaches to traffic safety data analysis and use data from a four-lane to three-lane conversion study to illustrate the implementation of these methods.

Pawlovich et al. (2006) used Bayesian methods and MCMC estimation to assess whether the reduction of number of lanes (“road diets”) resulted in crash reductions on Iowa roads. Crash data at each site was collected before and after the conversions were completed. Given the random and rare nature of

[40] In models where the bias is independent of the underlying true value (additive bias) only a single set of bootstrap replicates is needed for bias correction.

crash events, a hierarchical Poisson model was fit to crashes, where the log mean was expressed as a piece-wise linear function of time period, seasonal effects, and a random effect corresponding to each site. The posterior distributions of the parameters in the model were estimated using Markov chain Monte Carlo (MCMC) methods.

As regards Bayesian multilevel modelling in road safety, several applications have been published in the last decade within the context of spatial analyses. The most important applications are briefly described below; nevertheless, a more detailed presentation of these applications is provided in section 2.3.4.6. MacNab (2004) examines ecological and contextual determinants of area-aggregated motor vehicle accident injury in relation to socio-economic, residential and environmental indicators by means of Bayesian multilevel modelling (MacNab, 2004). Hewson (2005) examined child casualty rates aggregated within different areas and compared a simple generalized linear model, with an extension of it, in which a spatial structure is assumed for the random effects, and eventually with a Bayesian model, in which the "random effect" can be given a spatial prior structure and "shrink" the estimates of casualty rates across adjacent areas.

McMillan et al. (2007) developed Bayesian hierarchical binomial regression models in order to measure county-level variability in changes in alcohol-related crash rates while adjusting for county socio-demographic characteristics, spatial patterns in crash rates and temporal trends in alcohol-related crash rates. Aguero-Valverde and Jovanis (2006) compared full Bayes hierarchical models (including spatial effects, temporal effects and space–time interactions) to traditional negative binomial estimates of annual county-level crash frequency in Pennsylvania, and found that, in general, highly significant variables in the negative binomial models were also significant in the Bayesian models; however, variables marginally significant in the negative binomial models were non-significant in the Bayesian models. Because the FB models address spatial correlation and take into consideration all sources of uncertainty, the authors believe the FB models more accurately associate covariates with crash risk and are better suited for this type of data.

2.8.5 Summary

It is obvious that both simulation techniques presented in these sections include a substantial amount of computation. For this reason bootstrapping, like MCMC estimation should not be used for model exploration, but rather to obtain unbiased estimates and more accurate interval estimates at the final stages of analysis.

Moreover, it should be noted that the estimates these methods produce are dependent on random numbers. Consequently, using a different set of random numbers or a longer simulation run can produce (slightly) different estimates. For this reason, and because these methods are fairly new, compared to GLS

estimation methods, and have only recently started to be widely used in the context of multilevel analysis, it is important that they are implemented with care.

2.9 Conclusion multilevel modelling

Heike Martensen and Emmanuelle Dupont (IBSR)

Throughout this chapter a number of examples have been shown for hierarchically structured road-safety data. Accident data have a hierarchical structure because accidents can involve several vehicles, which may contain several passengers. Road safety data that are sampled from larger populations are often structured hierarchically when a limited number of primary sampling units (e.g. road sites) are selected from which the secondary units (e.g. cars) are randomly sampled. Hierarchies can also arise due to administrative structures like counties that are nested in regions that are nested again in countries.

Many researchers in road-safety are not aware of the consequences a hierarchical structure has for the appropriate analysis. The main goal of Chapter 2 of this deliverable is therefore to give guidelines how to deal with hierarchical data of different types.

2.9.1 Summary of multilevel techniques

It has been shown how the multilevel approach can be applied to a wide range of analysis techniques to solve the problems inherent to hierarchically dependent data in a productive way. Multilevel versions of those techniques that are most commonly used in road safety research have been presented.

2.9.1.1. Regression analyses

The regression techniques described in sections 2.2 (linear regression of normally distributed data); 2.3.2 (logistic regression for binomial response data); and 2.3.4 (Poisson regression for count data) are powerful tools to link different types of variables to each other. They can help to describe how a number of observed predictor variables (e.g., number of police controls) affect a particular outcome (e.g. number of fatalities). The limitations inherent to the original analyses concern the distribution of the outcome data, the form of the function that links dependent and independent variable, the overlap between predictors, and the distribution of the residuals; all of which – if not taken into account correctly – can jeopardize the interpretation of the results.

One of the most important assumptions is the independence of the residuals. Multilevel modelling is necessary in situations where this assumption is violated, as often the case when dealing with a hierarchical data structure. When data are collected from a nested structure (e.g. drivers nested in road site), the data coming from the same unit of the higher-order structure (e.g., all drivers checked at a particular road-site) are often more similar to each other than to those from another higher-order unit (e.g. the drivers checked at a different road site). While in traditional regression techniques such a hierarchical structure can cause violations of the independence assumption, this structure is explicitly

included in multilevel analyses by allowing the specification of sources of random variation at different levels of a hierarchy. Therefore, the most obvious advantage of multilevel analyses is to allow the researcher to respect one of the most important assumptions for regression techniques.

Moreover, the hierarchical structure itself can be a source of information. It can be interesting to know whether there is variation between the units of a higher level. For example, by comparing models that include “region” as a higher order level and those who do not one can investigate whether there is regional variation with respect to a particular phenomenon. By conducting a residual analysis, it is also possible to identify higher order units that behave differently from the others.

The multilevel structure of the analysis also allows investigating the relationship between variables that are situated at different levels of the hierarchy, for example the weather (a road site variable) might influence the effect that speed (a car specific variable) has on the probability of an accident. Another advantage concerns variables that are conceptually situated at a lower level (e.g. accidents, drivers, road sites, etc) but are available only at some higher aggregated level (e.g. county, region, country). As an example, traffic density is known to affect the risk of an accident. This density varies within a particular region but different regions also have different overall densities. Although in theory it would be preferable to include the density at the accident level, this information will often be unavailable. Multilevel modelling offers the solution to include traffic density as a regional characteristic, while still analysing the effect of some other variable at accident level.

2.9.1.2. Multivariate responses and repeated measures

Multilevel modelling was also introduced as a new way of dealing with more than one independent variable. This can be the case when several dependent variables of interest are analysed in parallel (Section 2.5) or with response types that are in fact represented by several variables. The latter case concerns, for example, multinomial responses, i.e. categorical variables that can take more than two different values (section 2.3.3). When such a variable forms the dependent variable, each response option is considered as a variable apart and they are jointly submitted to a multivariate analysis. When analysing multiple dependent variables, the lowest level of analysis consists of a dummy variable specifying to which response variable a particular value belongs, while the individual from which the values are obtained are coded at a higher level.

In a similar way, multilevel modelling can be used to analyse repeated measurements from the same subject (section 2.4). Instead of regarding the different measurements as levels of a factor as in traditional repeated-measurement approaches, one can enter all measurements simultaneously from each subject in the first level and consider the subject as a higher-level variable that groups the different measurements together.

Defining the multivariate or repeated measurement structures as a level in a multilevel analysis allows an easier handling of missing values as compared to

traditional multivariate methods. Values yoked to the ones missing can be kept in the analysis. Moreover, the assumptions with respect to the cause of missing values are less strict. While traditional multivariate models are based on the assumption that all missing values are missing completely at random, multilevel models can cope with values not missing completely at random, as long as the source of non-randomness is specified in the model.

The section on structural equation models (2.6) shows the basic form of such models in the multilevel case, dealing mainly with assumptions on data. On the other hand, this chapter discusses the necessary theoretical concepts of these models. Finally, a short summary of the application of structural equation models is introduced using the relationship of driver characteristics and their acceptance of new technologies in traffic.

2.9.2 When is the use of multilevel modelling necessary?

Generally, when dealing with a hierarchical data-structure, one should consider using multilevel modelling. In some cases the dependency among cases can be compensated by taking up higher order variables that cause this dependence without actually introducing a higher level into the analysis. As an example: The speed of cars is measured throughout the country by cameras at randomly selected road sites. The cars measured at the same road site will resemble each other more with respect to speed than those measured at different road sites. One might try to capture this dependency by taking up variables in the model equation that are responsible for the speed-differences between road sites. An obvious candidate is the speed limit which varies across road-sites and will indeed affect the speed of all cars at a particular road site in the same way. If the speed limit was the sole reason for the speed of cars resembling each other at the same road site, including it as a predictor would solve the dependency problem. The reason for this is that the assumption of independence must be applied to the residuals after all variables in the model have been accounted for. If one can include all sources of dependencies as variables into the models, there will be no dependency among the residuals anymore.

Practically however, it is usually a large numbers of factors that lie at the basis of the dependence. To keep with our example, road sites do not only differ with respect to the speed-limit but also with respect to the number of lanes, road conditions, traffic density, viewing conditions and probably a number of other factors of which the researcher might not even be aware that they affect the driving-speed. Consequently, the attempt to capture the dependencies with higher-level variables taken up in a single-level model will often, if successful at all, result in a large number of predictors many of which in them selves are not of interest to the researcher. As mentioned before, including many predictors can create problems with respect to interpretation. Moreover, they reduce the degrees of freedom which might make it more difficult to get a clear picture about the variables concerning the actual research question.

Apart from these practical problems with the inclusion of many predictors, by applying a single-level model, one misses out on important information about the data structure. In the case of random intercept models, the variance partition coefficient gives information to what extent cases within a second-level unit resemble each other more than cases between units. In the case of models with random slopes and random intercepts, the covariance between these two sources of variance tells the researcher whether there is a relation between the general level of measurement in the second order units (i.e. the intercepts) and the slope of the variable of interest. In sum, capturing hierarchical structures in multilevel models is easier and more informative than capturing the hierarchy by including second-level predictors in a single-level model.

While multilevel models offer an elegant solution to the problem of hierarchical data-structures, they inherit all other advantages and limitations of the regression models from which they are derived. An example is the treatment of correlated predictors. As demonstrated in Sections 2.3.3 and 2.4, the regression weights for correlated predictors are difficult to interpret. Non-significant weights can either mean that the variable has no effect, or that the effect is simultaneously captured by another variable included in the equation. A careful investigation of the correlation among the predictors and/or comparisons of various versions of the model (including each predictor singularly and then together) are necessary for a proper interpretation of the results. As the readers of this deliverable are expected to master the traditional analyses that each particular multilevel model is based on, it exceeds the scope of this document to give a full account of the possible limitations and problems in the interpretation that multilevel models inherited from traditional regression analyses. It must be kept in mind though that all other assumptions of a traditional model, except that of independent distribution of data, still have to hold in order to safely interpret the results of its multilevel version.

It is also important to realise that the possibility to carefully check whether there are hierarchies in the data, is actually a two-way street. Sometimes, one might think of possible higher-level variables (e.g. regions or countries) but it turns out that there is little variation between the units. The great advantage of the multilevel approach is that it is possible to represent a hierarchical structure, but of course that only makes sense if that structure is actually present in the data. It should also be noted that models can grow very complex very quickly. This is already the case with traditional multiple regression models where the inclusion of many predictors can lead to patterns of results that are difficult to interpret. With the introduction of multilevel models each predictor can moreover be defined as having a random slope at each level in the model (or not). This way the set of possible models is growing very quickly. We advise to introduce random slopes for particular predictors very sparsely, preferably on the basis of theoretical reasons.

2.9.3 Recommendations

The most important message we would like the reader to take home is the following: Always check the assumptions your analysis model is based on. If

you have hierarchically dependent data, statistical tests conducted with traditional methods of analysis might be flawed. Use multilevel modelling to deal with these dependency issues and make optimal use of all the information present in the data.

Multilevel modelling enables researcher to specify models that resemble complex hierarchical data-structures, allowing the parallel analysis of data at different levels of aggregation and the investigation of interactions between variables at different levels. With all these great opportunities, keep in mind, however, to make your model as complex as necessary but to keep it as simple as possible.

Chapter 3 - Time series analysis

3.1 Introduction to time series models

F. Bijleveld (SWOV)and Ruth Bergel (INRETS)

This chapter introduces the fundamentals of time series analysis as it is commonly used in road safety analysis and other fields. In road safety research, most time series are constructed by aggregating or averaging some quantity over a specific period of time, and then tabulating its value for subsequent periods. Probably the most common example of a time series used for road safety analysis is the annual or monthly number of fatalities in a country. For obvious reasons, the number of fatalities recorded every month at any space-aggregated level is the risk indicator of interest for road safety analysis, But to give another example, a series consisting of the maximum temperature recorded in the day at some meteorological station, averaged over several such stations and per month is also a time series, and is also of interest as risk factor for road safety analysis. In road safety research as in other fields, for commodity reasons, the time periods are almost always taken equal in length.[41]

A distinguishing feature of models for time series data over models for traditional cross-sectional data is that the order of observations is important: a linear regression on the data presented in reverse (or any other) order than the original one would yield exactly the same results. This will typically not be the case with time series analysis because effectively, an estimate for a particular observation may be dependent (among others) on the previous observation, which may be another observation in another ordering of the data.

Time series analysis can be regarded as an extension of regression analysis. In particular, it extends regression analysis by allowing for a certain type of relations between the residuals of the regression model, while in regression analysis residuals have to be fully 'independent'.

As discussed in the introduction (1.1.2), the error term of a model is assumed to be identically and independently (Gaussian) distributed. In practice, this assumption is tested with the help of the residuals of the model once a dataset of observations of the time series of interest is available and a model has been estimated on this dataset, and, as an extension, it is said that this assumption is related to the residuals. A similar, but more general concept has to be introduced for time series analysis: *stationarity.*

[41] Technically this is often not true. A year has either 365 or 366 days, a difference that is mostly ignored, which is not the case for monthly data (28, 29, 30 or 31 days). However, such effects may be corrected for in the model or directly in the data.

In this document, stationarity[42] is understood as follows: a time series is stationary if its expected value and covariance remain the same across and between time points. Although this seems restrictive, compliance with this (or a similar) assumption is essential for inferences to be made: you have to be able to assume your model is valid up to the year you want to make prognosis for. If you cannot do this, you cannot draw inferences for the future based on past observations.

Many time series analysis techniques require the time series to be stationary, or at least its random structure to be homogenous. See further definitions in Section 3.4.2.2.

A time series y_t is a sequential series of measurements over time. A time series model is a model for such a time series. In the introduction (Section 1.1.1), a simple model for driving errors as a function of driving experience is introduced:

$$driving\_errors_i = b_0 + b_1 years\_experience_i + e_i.$$

Basically, a time series model may not be that different from this model. In fact, if the annual number of driving errors is recorded for one person over his or her driving carreer, we already have a time series model:

$$driving\_errors_t = b_0 + b_1 years\_experience_t + e_t,$$

for t= the first year, the second year, and so on. Labelling these years by 1, 2, …, we obtain the familiar form:

$$driving\_errors_t = b_0 + b_1 t + e_t.$$

At this point the model is in fact no different from an ordinary linear regression model. It is called a *descriptive model, because no other variable than time is used to predict the driving errors.* It would be called an *explanatory model* when additional variables were used (Section 3.3.1.1). The differences between ordinary linear regression models and time series models are determined by how the residuals e_t of the regression model are treated as a consequence of the correlation property described above, and the fact that past observations can be considered. Schematically (and slightly simplified), the treatment of the e_t and the fact that past observations can be considered can be added to the model formula as follows:

$$driving\_errors_t = function1(present_t) + function2(past_t) + e_t,$$

where *function1 and function2* are generic -- but usually linear or at least additive -- functions and $past_t$ is all information that became available in the past

[42] This definition of stationarity is by far the most commonly used: it is called covariance or second order sationarity, or also weak stationarity - in lieu of strict stationarity which assumes identical joint distributions.

(at times t-1, t-2, ..), whether they be driving errors, explanatory variables *or* residuals, and $present_t$ is all new information at time t, Note that in most cases this distinction between $past_t$ and $present_t$ is more conceptual than practical, but in the end, the presence of the component which is a function of $past_t$ in the model distinguishes time series models from cross-sectional models.

Usually, the practical model representation is different from the one above. Instead of referring to the distinction present vs. past, it may be rearranged into components that have an interpretable role, such as for instance the trend, the general tendency of time series, and the periodic component in the case of a periodic pattern. The possibilities for the model specification in relationship with the components of interest, and in relationship with additional variables too, are numerous. A general model specification, referring to the nature of the variables used within the model, is discussed in Section 3.3.1.

However, besides identifying components based on the role they play in a model, two other important categorisations can be considered: based on whether a component is observed or not and based on whether a component is deterministic (non-random) or random. In practice the potential combination, unobserved random components can be important.

It is commonly said that that the dependent variable y_t is the observed one, and that its unobserved components are: the cycle, the trend, the seasonal component and the irregular component. In practice, some components may not be relevant, and for instance the cycle will never be considered in the applications to road safety analysis presented in this document. The seasonal component exists only in the case of a periodic (seasonal) pattern, and will mainly be estimated on the monthly datasets presented. See Section 3.3 for precise definitions.

In its simplest form, it is possible to construct a linear trend component using a linear function of the time index $at+b$, and to construct a seasonal pattern using dummy variables or a trigonometric function. The real interest of considering such 'parts' as components emerges when such components can be regarded as being *random*. A component can be regarded as being random (also called stochastic) when it changes over time. It is not (necessarily) meant that its value changes with time - a trend for instance in general is supposed to – rather it is meant that its *structure* changes. For instance the slope (a in the linear trend $at+b$ may level off or increase a little over time, or the seasonal pattern may change. Such a phenomenon may be modelled using a random component. See in Section 3.3 the discussion on decomposition models for precise definitions.

In many cases, data are transformed before analysis. Although this is not strictly a time series feature, many time series in road safety are log transformed before analysis.

In time series analysis two types of transform are most common:

1) at each individual time point, transformations where just one observation at a time is considered. The most common in traffic safety research being the logarithmic transform. In practice, one or both of two goals is intended to be achieved by applying the logarithmic transform: making a multiplicative model additive, and obtaining (approximately) equal observation error variances, attempting to satisfy second order stationarity requirements. The logarithmic transform, however, belongs to a special class of transforms called the Box-Cox transforms, which play for instance an important role in the DRAG-modelling context, but are also used in other contexts.

2) along the time axis. This type of transform in addition considers observations at other (usually previous) time points. This type of transform is mostly used to remove certain properties from a time series *before* the time series is analysed, in order to have the transformed time series satisfy requirements imposed by the technique that is intended to be used, or because these components are not of immediate interest for the analysis. Usually the requirement to be satisfied is first order stationarity.

Among these transforms, differencing is the most common. Differencing is performed to create a series of differences

$$\nabla y_t = y_t - y_{t-1}$$

If the resulting series ∇y_t is not stationary, the process is repeated by differencing again. See Section 3.4 for more details.

Nevertheless, some time series fail the stationarity assumption because for instance the expected value at a time point is a more complicated function than what can be removed by repeated differencing. Another reason for failing the stationarity assumption is non homogeneity in covariance, as it was said above. In such cases often a non-linear transformation is applied to the data before they are modelled and a time series analysis technique is performed on the residuals of this model. An example of a non-linear time series model is shown in Section 3.2.3.

The remainder of this chapter is organised as follows:

The linear regression model is used as a starting point, and treated in Section 3.2.1. This type of model is deterministic as it only contains deterministic components. The same section also discusses the identification of dependence in more detail (as well as discussing the other assumptions). Although knowledge of linear regression models is assumed in this document, it is strongly advised to read this section.

Due to the potential importance of the distributional assumptions, the generalised linear model (McCullagh & Nelder, 1989), which -- in particular its time series aspects -- is the topic of much ongoing research is treated in Section 3.2.2. Nonlinear least squares models are the subject of Section 3.2.3. In both

sections the treatment of time dependence is informally introduced. In the generalised linear models section this is done using functions. In the non-linear least squares models section this is done using lagged residuals, thereby informally introducing autoregressive models, which are discussed in more detail in Section 3.4.

The above-mentioned sections discuss the time series aspect by extending the GLM and nonlinear models approaches to time series. After a general introduction to dedicated time series models in road safety is given in Section 3.3, two sections devoted to specific dedicated time series analysis approaches based on linear Gaussian models are given. Of the dedicated models, Auto Regressive Moving Average (ARMA) type models are discussed in Section 3.4. This type of model is by far the most often used for fitting stationary and not stationary data, and calling on additional variables as well. The following section, 3.5, discusses the closely related DRAG model (Demand for Road use, Accidents and their Gravity), a three level approach using many explanatory variables, where certain nonlinear transformations on the data, both dependent and independent, are considered. An alternative, based on state space techniques is the topic of Section 3.6. These models, which are unobserved and stochastic components models, and also referred to as structural time series models, are directed at decomposing the time series into interpretable (un)observed components structures. To conclude, in Section 3.7 the state space approach and the ARIMA approach are discussed in terms of similarity, and two examples of equivalences between well-defined specifications of models of these two classes are given, on datasets already modelled in Sections 3.4 and 3.6. Finally in section 3.8 the conclusions and recommendations of this chapter are summarized.

3.2 Classical linear and non-linear regression models

3.2.1 Classical linear regression models

Christian Brandstaetter and Michael Gatscha (KfV)

3.2.1.1 Objective of the technique

In the field of social science, no other statistical procedure has offered so many impulses as the procedures of analysing correlations. The knowledge of a correlation between two variables is an essential pre-condition in order to draw conclusions by predicting one variable through another.

Time series data are often used in conjunction with linear regression techniques in terms of predicting statistical trends. In time series analysis, the independent variable *x* is given as time. The equation of a straight line is used to calculate the trend that the dependent variable *y* adheres to as time passes:

$$y = bx + a \qquad (3.2.1)$$

where *y* represents the dependent variable, *x* is the independent variable, *b* describes the gradient of the straight line and *a* the altitude in geometrical terms. The gradient *b* of a straight line can be positive or negative. If the gradient is positive, the *y*-values increase with increasing *x*-values. In the case that *b* is negative, *y*-values decrease with increasing *x*-values.
When time is used as the independent variable, a number of complications that are introduced to the regression method are expected. The most important complication is caused by the time dependencies between the values of y. However, there is also an influence affected by the units that are used to measure time. For example, if annual data are used, it will be impossible to identify the seasonal factors that may well influence the data. So, when looking at data with regard to accidents, one would probably want to view quarterly figures rather than merely annual data, as one would expect there to be an increase in accidents e.g. in the summer quarter when analysing motorcycle accidents.
However, in order to identify a trend value of the time series data that is analysed, a linear regression line can be drawn by using averages over periods of time to smooth out fluctuations and, as a result, show the general trend.

3.2.1.2 Model definition and assumptions

The most basic relationship between two or more interval-scaled variables is explained by the following equation to determine the regression:

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + e_i \qquad (3.2.2)$$

where
y_i is the i^{th} value of the dependent scale variable
p is the number of predictors

b_j is the number of the j^{th} coefficient, $j=0,\ldots,p$
x_i is the value of the i^{th} case of the j^{th} predictor
e_i is the error in the observed value for the i^{th} case

For visualization reasons in the following text, the equation can be simplified to formula 3.2.1.

If one has obtained n pairs of observations x_i, y_i (i =1, . . ., n), it is possible to illustrate these observations by means of a scattergram (see Figure 2.15). Graphically, the principle of a linear regression is to construct a straight line in a two-dimensional system of coordinates such that all data points within the system of coordinates lie as near as possible to this line, as measured in the direction parallel to the y-axis:

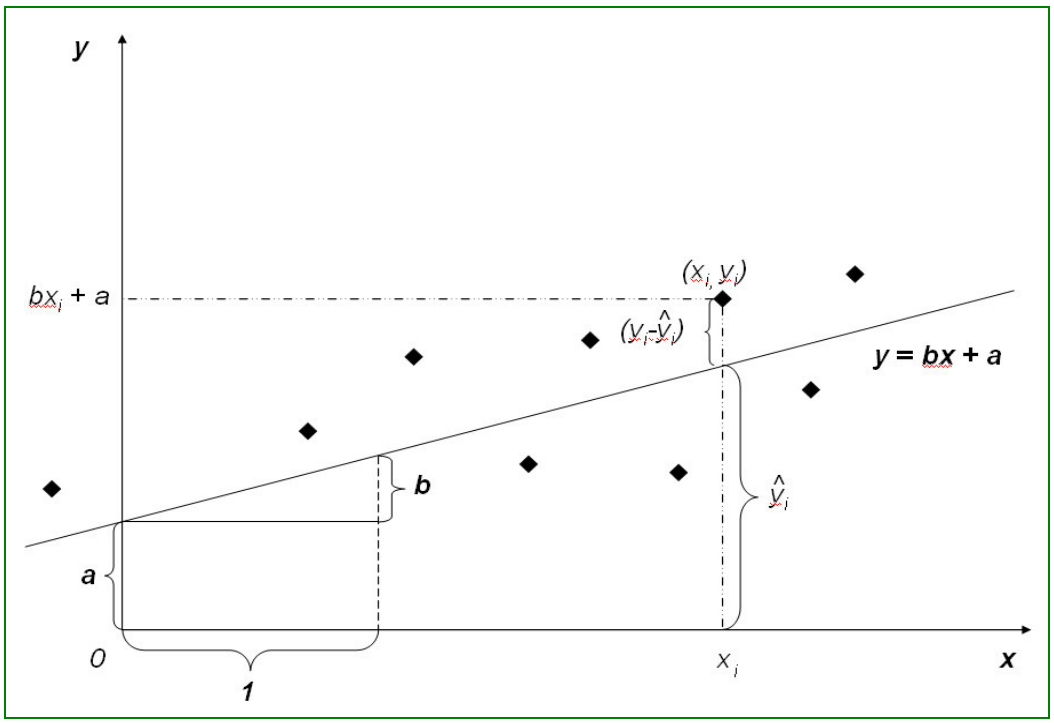


Figure 3.2.1 Scattergram with regression line

In Figure 2.15, y_i is the observed value and $\hat{y}_i$ is the predicted value. As a consequence, the general term $(y_i - \hat{y}_i)$ describes the size of the "prediction error". One could assume now, that the regression line with the best fit to describe the data is characterized through the minimization of the sum of $(y_i - \hat{y}_i)$. However,it is also possible that this sum is a negative value and therefore it can also be assumed that many regression lines exist for which the sum of the differences $(y_i - \hat{y}_i)$ is zero. Hence, the best criterion for the fit of a regression line is not the sum of the differences, but the sum of squared differences, or in other words: the minimized sum of squared distances between the individual observation points and the regression line measured in the direction parallel to the y-axis:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \qquad (3.2.3)$$

using ($bx_i + a$) instead of $\hat{y}_i$, the equation looks like:

$$\sum_{i=1}^{n} [y_i - (bx_i + a)]^2 = \min \qquad (3.2.4)$$

With that criterion in mind, it is possible to generate *n* values to draw the regression line, but one has to hope that the calculated values are as small as possible. It is also possible that another regression line, based on squared differences, describes the observed values even better. For this reason, variables *a* and *b* are defined by a differential equation, *f(a,b),* partially differentiated with respect to *a* and *b.* Solving this equation yields to the following explicit solution for a and b:

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x} \qquad (3.2.5)$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (3.2.6)$$

In the equations mentioned above, *n* is the number of data points in the time series, e.g. the number of months. That is to say, *y*-values exist only for the natural numbers (i = 1, …, n) on the *x*-axis. Thus, the regression line of the time series arises through the connection of all points $\hat{y}_i$ (for i = 1, …, n).

If *a* and *b* are calculated through these equations, the result is a regression line for which the sum of squared differences is really minimized. This estimation procedure is called ordinary least squares, or OLS, and is one of the basic concepts of linear regression. The Gauss-Markov Theorem shows that:

- ✓ *b* is an unbiased estimate of the regression coefficient β, which means that on repeated estimates, the distribution of *b* will be centred around β.
- ✓ The sampling distribution of b will be normal if the samples are large and a sufficient number of samples are taken.
- ✓ OLS provides the best linear unbiased estimate of β (BLUE).
- ✓ "Best" means: OLS provides the most efficient unbiased estimate of β. Efficiency refers to the size of the standard error of *b* (σ_b));

Most commonly, regression is used to predict the value of one variable from the value of another, if the two are related. Therefore, one variable is normally

defined as a predictor, whereas the other is determined by a criterion. This categorization is quite similar to the definition of a dependent and independent variable, although the latter relationship characterizes a narrower, causal relationship.

In order to fit a simple linear regression model to a set of data, one has to find estimators for the unknown parameters *a* and *b*, which are expected to have a linear relationship of the shape $y = bx + a$. Since the sampling distributions of these estimators will depend on the probability distribution of the random error *e*, it is necessary to make several specific assumptions about its properties. The mean of the probability distribution of the random error is 0. That is, the average of the errors over an infinitely long series of experiments is 0 for each setting of the independent variable *x*. This assumption implies that the mean value of *y* for a given value of *x* is $y = bx + a$.

For estimated linear regression following the OLS procedure shown above, we have four basic assumptions about the prediction error ε. Corresponding to the above-mentioned Gauss-Markov Theorem, they are called Gauss-Markov assumptions:

1. The prediction error ε is uncorrelated with x, the independence assumption.

2. The variance of the error term is constant across cases (*x*) and independent of the variables in the model. This is called homoscedasticity, or homogeneity of the variance of ε. An error term with non-constant variance is said to be heteroscedastic.

3. The value for the error term associated with any different observations is independent. The error associated with one value of *y* has no effect on the errors associated with other values. This means that all observed autocorrelations of the errors are near 0.

4. The random errors are distributed normally.

As mentioned earlier, when it comes to analysing time series with regard to accident data, one can suppose that at least one of the listed assumptions is often violated in practice, e.g. the assumption of nonautocorrelated.

The first assumption was the independence of the prediction errors and x. We can find three different possibilities of problems:

- ✓ Spurious relationship: ε and x may be correlated because z is a common cause of x and y. In this case b is a biased estimate of the regression coefficient β.
- ✓ Collinear Relationship: If x_2 is correlated with x_1 and y, but is not the cause of either, b_1 will be a biased estimate of β_1.
- ✓ Intervening Relationship: x_2 intervenes in the relationship between x_1 and y. In this case b_1 will not be a biased estimate of β, but it will reflect both the direct and indirect effects of x_1 on y.

The second assumption is the homoscedasticity of the residuals. Here we can find four different conditions (see Figure 3.2.2.; the lines represent the pattern of the dispersion of the residuals. In all three conditions with heteroscedasticity, b will be an unbiased estimate of β, but σ_b (σ is standard deviation) will be incorrect - too large or too small. This yields wrong significance tests because significance is tested with the Student's t-statistic $t=(b/\sigma_b)$.

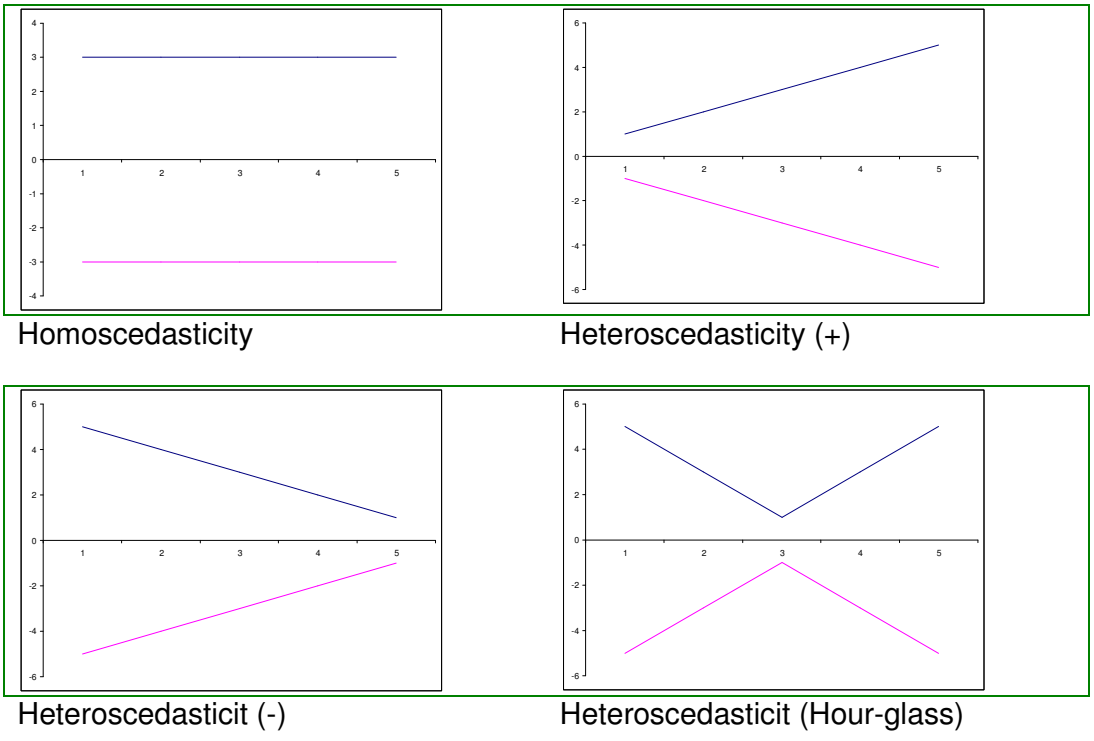


Figure 3.2.2 Overview of patterns of homoscedasticity und heterocsedasticity

In the case of Heteroscedasticity (+), SE_b is underestimated and a type I error may occur. In the case of Heteroscedasticity (-), in contrast, SE_b is overestimated and a type II error may occur.

White (1980) has published a direct test for heteroscedasticity:
χ^2 (df)= R^2n,
where n is the number of cases, R^2 is the squared multiple correlation coefficient for the regression of the squared residuals on predictor? x, and the number of degrees of freedom df is the number of independent variables. The null hypothesis is that the residuals are homoscedastic.

Another widely used test for homoscedasticity is given by the following test statistic:

$$H(h) = \frac{\sum_{t=n-h+1}^{n} e_t^2}{\sum_{t=d}^{h} e_t^2} \qquad (3.2.7)$$

where h is some time point in the series cutting the series in two parts: one before and one after time point h. This statistic can be tested against an F(h,h)-distribution.

The third assumption of non-autocorrelated errors is most often violated in time series regression. Plotting the residuals of the classical regression analysis against time can confirm that the observations are not independent. Since these residuals are assumed to be completely independent, they should be randomly distributed.

A useful diagnostic tool for investigating the randomness of a time series is called the correlogram. The correlogram is a graph containing the correlations between an observed time series and the same time series shifted t time points into the future, for a (limited) number of t. Thus, the correlogram of the residuals e_i consists of the correlation between e_i and e_{i+1}, the correlation between e_i and e_{i+2}, the correlation between e_i and e_{i+3} and so on. Using a more general notation, the correlogram contains the correlations between e_i and e_{i+k}, for k = 1, 2, 3, etc. Since k equals the distance the observations are set apart in time, it is called the lag. Moreover, since the correlations are computed between a variable and itself (albeit shifted in time), they are called autocorrelations.

When the first order residual autocorrelation (i.e., the residual autocorrelation for lag 1) is positive and significantly deviates from zero, a positive residual tends to be followed by one or more further positive residuals. As pointed out in the literature (see Ostrom, 1990, and Belle, 2002), the error variance for standard statistical tests can be seriously underestimated in this case. This in turn leads to a large overestimation of the F- or r-ratio, and therefore overly optimistic conclusions from the analysis.

On the other hand, when the first order residual autocorrelation is negative and significantly deviates from zero, then a positive residual tends to be followed by a negative residual, and vice versa. In this case, the error variance for the standard statistical tests is seriously overestimated, leading to a large underestimation of the F- or r-ratio, and therefore overly pessimistic conclusions.

The Ljung-Box test (Ljung and Box, 1978) is based on the autocorrelation plot. However, instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags. More formally, the Ljung-Box test can be defined as follows. The test statistic is

$$Q_{LB} = n(n+2)\sum_{j=1}^{h}\frac{\rho^2(j)}{n-j} \qquad (3.2.8)$$

with n the sample size, $\rho(j)$ the autocorrelation at lag j, and h the number of lags being tested. The null hypothesis of randomness is rejected if

$$Q_{LB} > \chi^2_{1-\alpha;h} \qquad (3.2.9)$$

where χ^2 is the percent point function of the chi-square distribution.

Excursion: The sample autocorrelation and partial autocorrelation (F. Bijleveld, SWOV).

A plot of the sample (partial) autocorrelation (the word sample is often dropped in applied studies) is often used to identify time dependence in residuals. It is also used to identify the order of dependence, where the partial autocorrelation is used to determine the order of the autoregressive dependence and the autocorrelation is used to determine the order of the moving-average dependence. See the "ARMA type models" section (3.4) for details. All introductory time series books cover this subject extensively, including Brockwell and Davis (1998, page 57 and 136) and Box and Jenkins (1976, page 32 and page 64).

Following the introduction of this document, it is argued that the presence of time dependence in the residuals of road safety models is discovered by the phenomenon that adjacent residuals tend to have the same sign, or tend to have the opposite sign. It may also occur that, for instance the residuals of winter-time observations share the same sign. For that reason, not only the immediate adjacent residuals are compared, but in addition also residuals at reasonable distance in time (lag). For instance, when monthly data are analysed, it is common to compare a january residual with the january residual the year before, february with february the year before, and so on, in order to identify seasonal patterns. The maximum considered lag is usually determined by the problem at hand. For monthly data, the maximum lag considered is longer than 12, but often not longer than 24.

The *sample autocorrelation* coefficients from the residuals e_t are computed as follows:

$$r_k = \frac{\sum_{i=1}^{n-k}(e_i - \bar{e}) \times (e_{i+k} - \bar{e})}{\sum_{i=1}^{n}(e_i - \bar{e})^2} \qquad (3.2.10)$$

where n is the total number of observations and $\bar{e}$ is the average of the e_i.

Please note that r_k differs slightly from what would be obtained when the classical correlation coefficient between $\{e_1,\ldots,e_{n-k}\}$ and $\{e_{k+1},\ldots,e_n\}$ would be calculated. Also note that r_0 is always equal to one. Also note that $r_k = r_{-k}$. Finally note that the autocorrelations for larger lags (larger values of k) are calculated using less terms in the numerator, while the (number of terms in the) denominator remains the same.

Once calculated, the sample correlations are then displayed like in Figure 1.2.3 in the introduction and Figure 3.2.7 later in this section. These figures are called autocorrelation plots and often abbreviated to ACF plots. *Approximate* confidence intervals (usually 95%) for each k are indicated by two lines (± 0.343 in Figure 1.2.3). For the confidence intervals it is assumed that under the null

hypotheses of no autocorrelation at all, each r_k has a standard error of approximately $1/\sqrt{n}$. Please note that these tests for all r_k are not independent.

However, for very large lags compared to the total number of observations these approximate confidence intervals may not be accurate. Therefore, in the plot in Figure 2.18 (produced by SPSS, a different approximation is used (see SPSS algorithms, page 4):

$$\mathrm{var}(r_k) \cong \frac{1}{n}\left(\frac{n-k}{n+2}\right) \qquad (3.2.11)$$

Obviously, in most practical cases this correction can be ignored.

It is important to note that these tests on the (partial) autocorrelations are only valid for stationary residuals, which is usually the case with residuals of a satisfactory fitting model. However, the plot of the autocorrelation function is also used as an indicator for non-stationarity.

The *sample partial autocorrelation* (PAC) indicates what correlation cannot be accounted for by the sample autocorrelation. It is computed by means of linear equations from the sample autocorrelations. The partial autocorrelations for the residuals are usually assumed to have a standard error of approximately $1/\sqrt{n}$ (similar to the autocorrelation). How precise to compute the partial autocorrelations is relatively complicated and can be found in introductory time series books, for instance Brockwell and Davis (1998, page 136) and Box and Jenkins (1976, page 64), Chatfield (2004, page 61).

Although the ACF and PACF are well suited to determine at what lags significant correlations exist (and other conclusions to be discussed in the Section ARMA type models), they may not be very practical to capture the whole picture in one test. To that end, the Box-Ljung statistic is often used.

Sample (partial) autocorrelation can be contrasted with a *theoretical* (partial) autocorrelation. The same is true for, autocovariances. In general, for simplicity focussing on covariance instead of correlation here, two stochastic variables X and Y have a theoretical covariance EXY-EX.EY.

In a sample, this quantity if mostly estimated by

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - (\frac{1}{n}\sum_{i=1}^{n}X_i)) \times (Y_i - (\frac{1}{n}\sum_{i=1}^{n}Y_i)), \qquad (3.2.12)$$

where it is assumed that the X_i and Y_i all have identical distributions (this assumption can be weakened somewhat). Obviously, the figure S_{XY} is not very meaningful when this (in practice a weaker) assumption cannot be upheld. In fact, it is only useful when all X_i and Y_i have the same theoretical covariance and we can thus talk about *the* covariance. In a similar fashion, the *correlation* is defined.

Please note that, like the estimate S_{XY} of the theoretical covariance above, sample estimates (the final outcomes) tend to differ from theoretical "true" values, and only if the estimates are unbiased, will they on average, if the samples get larger, tend to the true value.

The issue of X_i and Y_i having the same theoretical covariance extends to time series analysis in a more complicated way. In time series analysis, the notion of stationarity, already briefly mentioned above, is defined for this purpose. It is discussed in the "ARMA-type models" section below.
End of excursion: The sample autocorrelation and partial autocorrelation

For testing the last assumption about normality, most statistical packages provide both estimates of skewness and kurtosis and standard errors for those estimates. One can divide the estimate by it's standard error to obtain a z test of the null hypothesis that the parameter is zero (as would be expected in a normal distribution). There are other tests that in this situation are more powerful, for example the Kolmogorov-Smirnov statistic (for larger samples) or the Shapiro-Wilks statistic (for smaller samples). These have very high power, especially with large sample sizes, in which case the normality assumption may be less critical for the test statistic whose normality assumption is being questioned.

Table 3.2.1 shows a summary of the different assumption violations and their consequences.

It has to be mentioned that some assumptions are more important than others. In the case of linear regression in time series applications, the most important violation concerns the independence assumption. The second most important assumption is the homogeneity of the residuals. The least important assumption is that the residuals are normally distributed.

Assumption Violation	Consequences
Errors correlated with x	
Spurious relationship	b biased estimate of β
Collinear relationship	b biased estimate of β
Intervening relationship	b unbiased estimate of β, but reflects both direct & indirect effects
Heteroscedasticity ($R_{XS_e}^2 \neq 0.0$)	b unbiased, but not efficient; SE_b too small/large; Type I or II error may result
Autocorrelated errors	b unbiased but not efficient; SE_b too small/large; Type I or II error may result
Errors non-normally distributed	b may be unbiased if homescedasticity & independence assumptions meet & n is large; if n is small, t distribution may be biased

Table 3.2.1 Summary of assumptions and consequences of violations

3.2.1.3 Dataset and research problem

The dataset used is based on accident data from Austria and shows the development of fatal accidents all over the country from 1987 to 2004 on a monthly observation basis.

In this example, the distribution and development of people who were killed in accidents based on monthly observations is shown in Figure 3.2.3.:

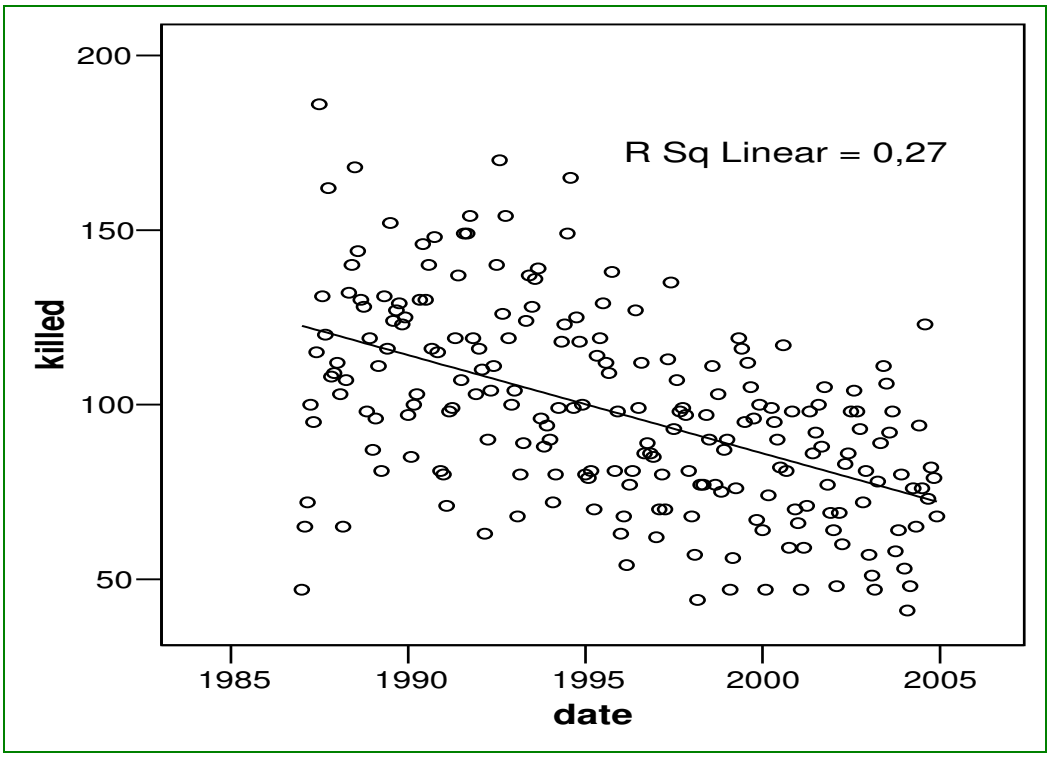


Figure 3.2.3 Scatterdiagram of the monthly number of fatalities in Austria from 1987 to 2004

3.2.1.4 Model fit, diagnostics, and interpretation

The model estimation of the example dataset was calculated with SPSS (www.spss.com). First, the ANOVA table test procedure tests the acceptability of the regression model. It shows that the unexplained variation (sum of squares, residual row) is higher than the explained variation (sum of squares, regression row).

ANOVA[b]

Model		Sum of Squares	df	Mean Squares	F	Sig.
1	Regression	46129,857	1	46129,857	79,228	,000[a]
	Residual	124600,5	214	582,245		
	Total	170730,3	215			

a. Predictors: (Constant), TIME

b. Dependent variable: KILLED

Table 3.2.2 ANOVA table of the linear regression analysis applied to the monthly number of fatal accidents in Austria in the period 1987-2004

The significance value of the F-statistic is less than 0.05, which means that the variation explained by the model is not due to chance. While the ANOVA table is a useful test of the model's ability to explain any variation in the dependent

variable, it does not directly address the strength of this relationship. Table 3.2.3 shows the coefficients of the regression:

Coefficients[a]

Model		Non standardized coefficients		standardized coeffizients		
		B	Standard error	Beta	T	Sig.
1	Constant	1259,417	130,558		9,646	,000
	TIME	-8,91E-08	,000	-,520	-8,901	,000

a. Dependent variable: KILLED

Table 3.2.3 Coefficients table of the linear regression analysis applied to the monthly number of fatal accidents in Austria in the period 1987-2004

The gradient of the regression line is negative, whereas the beta-coefficient (i.e. the coefficient of correlation) between *x* and *y* is –0.520. The gradient of the regression line is checked by a t-test, which is equal to the square root of the F-test in the ANOVA table mentioned before. The result suggests a highly significant decrease in the number of fatalities in Austria since 1987.

Finally, the model summary table reports the strength of the relationship between the independent and the dependent variable (Table 3.2.4)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,520[a]	,270	,267	24,130

a. Predictors: (Constant), Year/Month

Table 3.2.4 Model summary table of the linear regression analysis applied to the monthly number of fatal accidents in Austria in the period 1987-2004

R, the multiple correlation coefficient, is the linear correlation between the observed and model-predicted values of the dependent variable. Its value (0.52) indicates a moderate relationship. The R Square value (the coefficient of determination) is the squared value of the multiple correlation coefficient. It shows that 27 percent of the variation in the number of fatalities is explained by time.

The results shown above are only true if the basic conditions stated in the Gauss-Markov assumptions hold. Based on the fact that linearity is only assured if the residual value varies unsystematically, one can check the validity of the model. All model checks are based on the assumption that the error term is independent of the variables (*x, y*). So, when checking the plot, it must not show any systematic relationships. If this is the case, the use of the linear regression is not justified due to non-linear relationships in the data.

The histogram of the residuals reveals that the assumption of normality of the error term is justified (the standard Kolmogorov-Smirnov test yields a z-value of

0.594, which indicates no significant deviation from the normal distribution): see Figure 3.2.5.

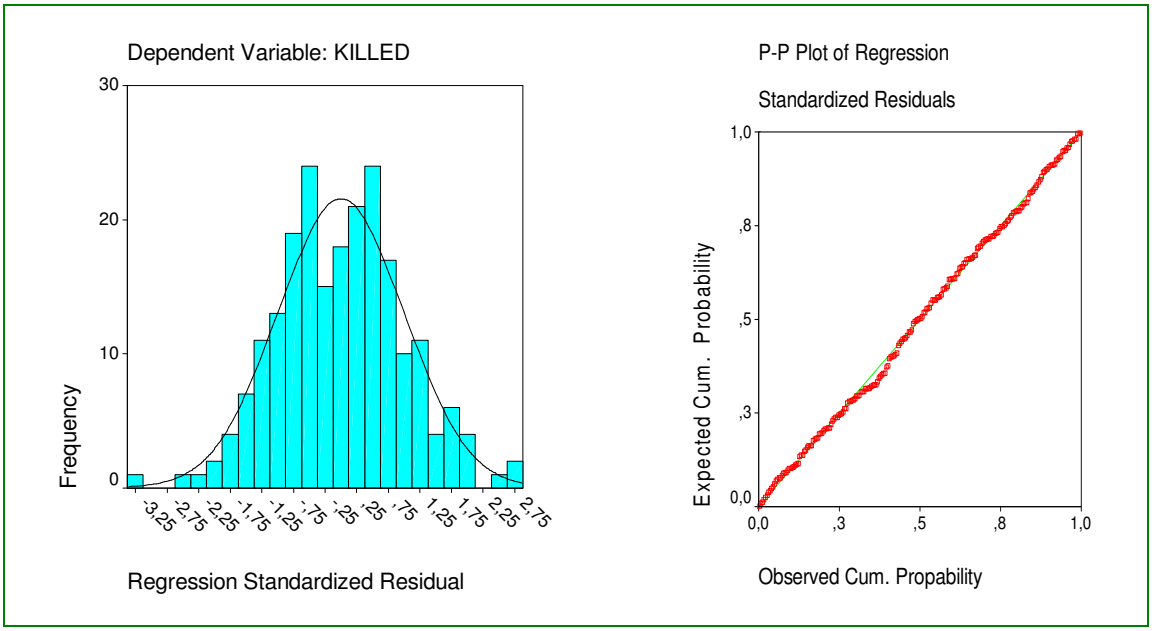


Figure 3.2.5 Histogram and P-P Plot of standardized residuals (in other chapters also the Q-Q Plot is used)

The shape of the histogram approximately follows the shape of the Gaussian curve; the P-P plotted residuals also follow the 45-degree line (Figure 3.2.5). Therefore, it can be concluded that the histogram is acceptably close to the normal curve. Again, the assumption of normal distribution of the example data is reasonable.
Additionally, a (shortened) table of residual statistics (Table 3.2.5) shows the following:

	Minimum	Maximum	Mean	Std. Deviation	N
Stud. Deleted Residual	-3,191	2,527	-,001	1,006	216
Cook's Distance	,000	,097	,004	,008	216
Centred Leverage Value	,000	,046	,005	,006	216

Table 3.2.5: Table of selected residual statistics

One can find the most important indices of the residuals in the row “Studentized Deleted Residuals”. In the example dataset, the maximum for this value is 2.527. As a consequence, there is no evidence for extremely high or low observation values. Furthermore, the values of “Cook's Distance” and “Centred Leverage Value” are also good checks for very influential values (Stevens, 1996). As both are around zero, this also indicates that there is also no sign of outliers. For testing the assumption of homoscedasticity, there are some heuristic ways by looking at different scatterplots (Figure 3.2.6.).

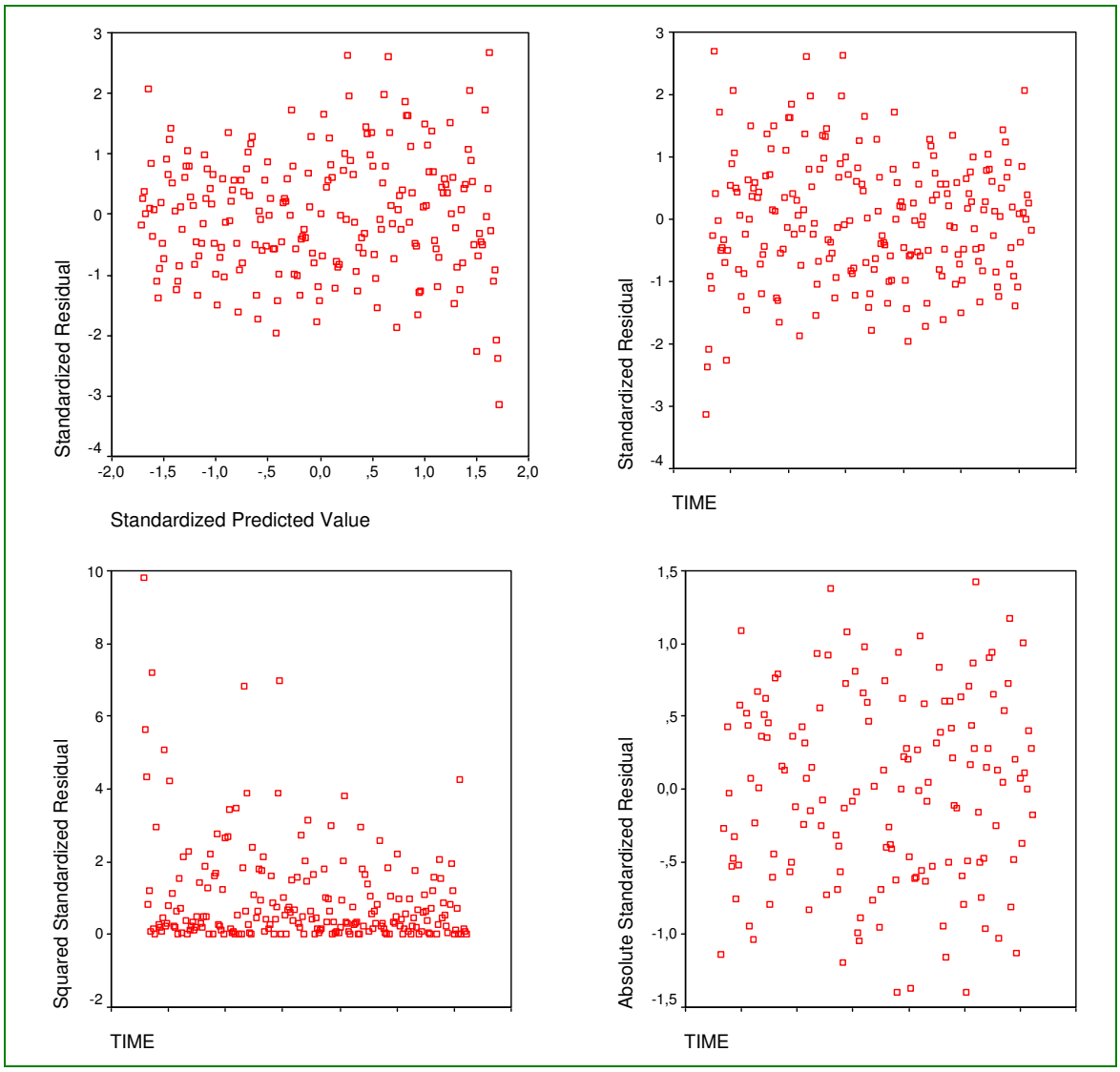
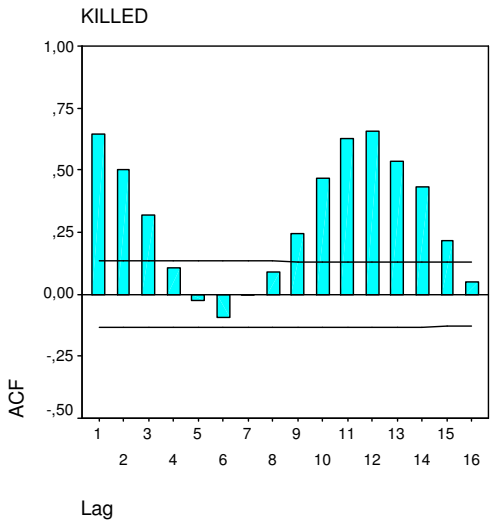


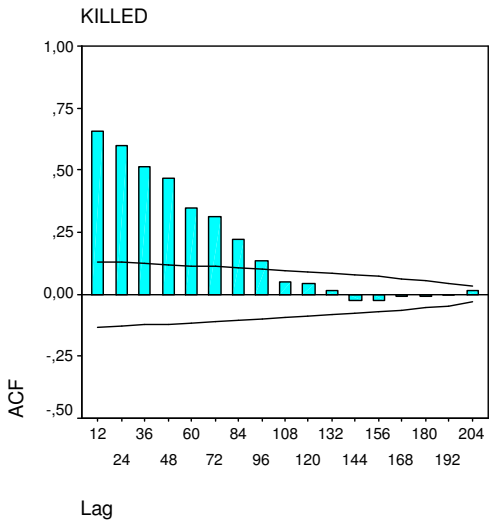
Figure 3.2.6: Table of selected residual plots for identifying heteroscedasticity

All plots in the table above show no indication of the presence of heteroscedasticity except the one in the lower left, in which slightly higher squared residuals in the early years are found. By using the previously introduced White's test, we find a $\chi^2 = 11.232$ (R-Square of the regression of the squared standardized residuals on the date variable is 0.052, the number of time points in the analysis is 216, df is 1) which is highly significant and shows the presence of heteroscedasticity in the data.

Finally, we are looking at the problem of autocorrelated errors, which is the most likely violation in time series regression. This is also true in the data example used in this chapter, as shown in figure 3.2.7.



Autocorrelations up to lag 16 for the original series of Austrian fatalities (The lines indicates two standard errors)



Autocorrelations of the same month over the entire series

Figure 3.2.7: Table of autocorrelations and seasonal adjusted autocorrelations

The two plots in figure 3.2.7 show very high dependencies of consecutive errors. Despite the fact that b will remain an unbiased estimate of β, the significance tests shown above in the outlined example are wrong. When the first order residual autocorrelation (i.e., the residual autocorrelation for lag 1) is positive and significantly deviates from zero, a positive residual tends to be followed by one or more further positive residuals, and a negative residual tends to be followed by one or more further negative residuals. The error variance for standard statistical tests is seriously underestimated in this case. This leads to an overestimation of the *F*- or r-ratio, and therefore overly optimistic conclusions from the analysis.

The above results are not an artefact of the seasonal component in the data series, which is shown in the right-hand plot in Figure 3.2.7 and will be outlined a bit more by performing two more analyses. Firstly, the regression equation will be expanded by adding dummy variables for the month as a second set of predictors in the model (the 11 variables feb thru dec are 0/1 dummys, January is collinear with the 11 others) . Secondly, aggregated yearly data will be used as dependent variable.

The model fit statistic (R Square) in Table 3.2.6 shows a much better fit of the linear regression model including the dummy predictors for the month effect compared to the simple model above (0.856 vs. 0.27).

a Predictors: (Constant), dec, YM, jun, jul, mai, aug, apr, sep, mar, feb, oct, nov

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,846(a)	,716	,699	15,456

a Predictors: (Constant), dec, Year/Month, jun, jul, mai, aug, apr, sep, mar, feb, oct, nov

Table 3.2.6: Model summary table of multiple linear regression analysis applied to the monthly number of fatal accidents in Austria in the period 1987-2004 with dummy variables for the month of year as a second predictor.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	122233,758	12	10186,147	42,638	,000(a)
	Residual	48496,570	203	238,899		
	Total	170730,329	215			

a Predictors: (Constant), dec, Year/Month, jun, jul, mai, aug, apr, sep, mar, feb, oct, nov

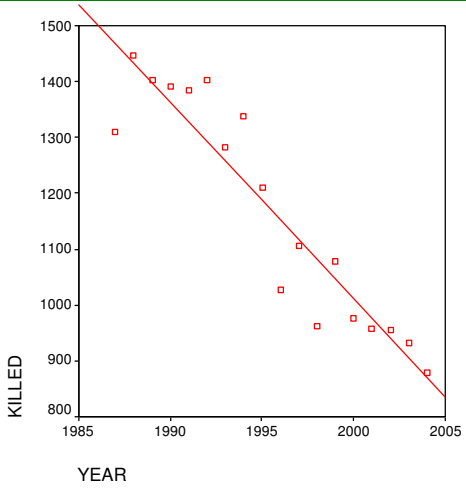
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1286,700	83,738		15,366	,000
	Year/Month	-9,28E-008	,000	-,541	-14,451	,000
	feb	-9,474	5,152	-,093	-1,839	,067
	mar	-6,136	5,152	-,060	-1,191	,235
	apr	7,502	5,152	,074	1,456	,147
	mai	28,020	5,153	,275	5,438	,000
	jun	39,324	5,153	,387	7,632	,000
	jul	41,454	5,153	,408	8,044	,000
	aug	48,869	5,153	,480	9,483	,000
	sep	30,784	5,154	,303	5,973	,000
	oct	36,525	5,154	,359	7,086	,000
	nov	18,218	5,155	,179	3,534	,001
	dec	16,570	5,155	,163	3,214	,002

Table 3.2.7 ANOVA[a] and coefficients[b] table of multiple linear regression analysis applied to the monthly number of fatal accidents in Austria in the period 1987-2004 with dummy variables for the month of year as a second predictor.

The results in Table 3.2.7 are very similar to the results in Tables 3.2.2 and 3.2.3 The inclusion of the predictor 'month' in the model significantly improves neither the F-test nor the parameter tests. But the change in the model fit is highly significant: the R^2 increases from .270 to .716 (F-Change=28.96, df=11).

The previous result on monthly data (see Table 3.2.6 and 3.2.7.) is replicated on yearly data again. The number of fatalities in road accidents has been decreasing since 1987 (see Table 3.2.8). The result of the regression on the

yealy data is more significant than in the case of the monthly fatalities in the above regressions because there are no seasonal artefacts in the yearly data which introduce high variation not due to the general trend in the model.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,933[a]	,870	,862	74,77443

a. Predictors: (Constant), Year

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	598633,5	1	598633,499	107,067	,000[a]
	Residual	89459,445	16	5591,215		
	Total	688092,9	17			

a. Predictors: (Constant), Year

b. Dependent Variable: Fatalities

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	71312,108	6778,903		10,520	,000
	Year	-35,151	3,397	-,933	-10,347	,000

a. Dependent Variable: Fatalities

Table 3.2.8: Plot of Yearly Fatality Data in Austria from 1987 to 2004 and regression results

On the other side, we find that the distribution assumptions are also met in this case, but not as close as in the monthly model (see Figure 3.2.8).

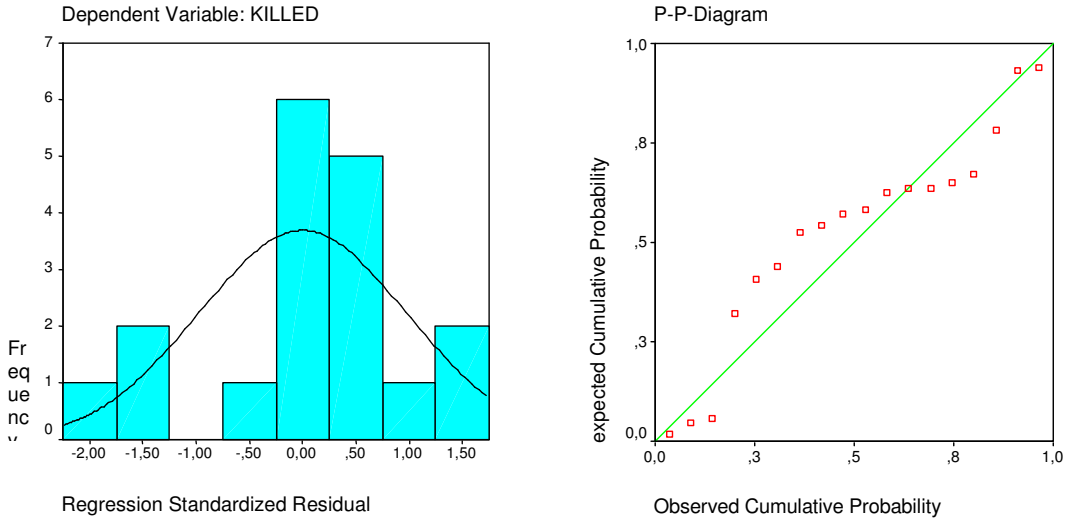


Figure 3.2.8: Histogram and P-P Plot of standardized residuals for the regression model on yearly Austrian fatalities data.

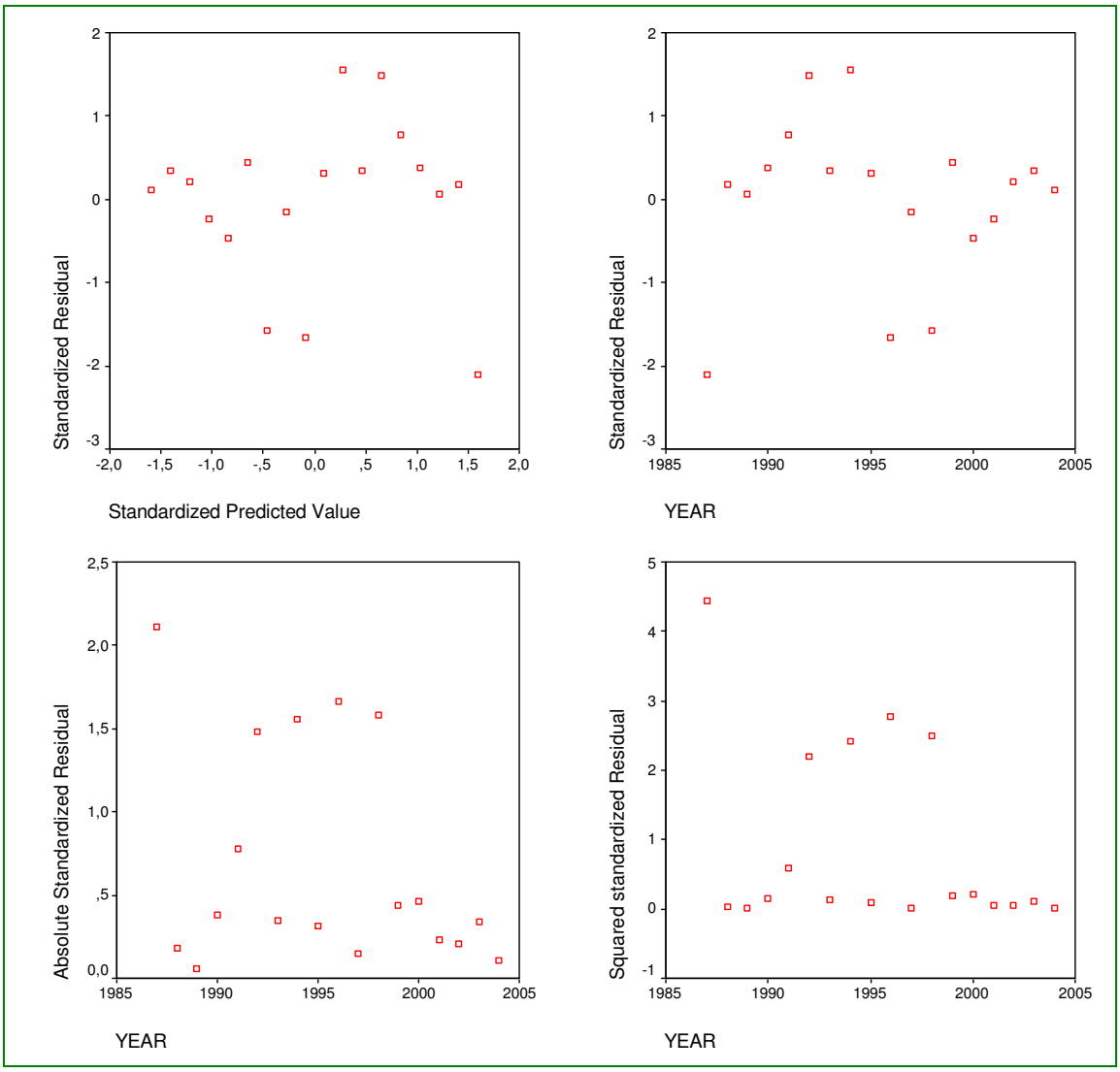


Figure 3.2.9 Table of selected residual plots for identifying heteroscedascity

All plots (see Figure 3.2.9) show a light trend of smaller residuals in the later years. This cannot be proven by White's test: $\chi^2 = 2{,}057$, df=1, therefore we can assume homoscedastiscity in the yearly fatalities data in Austria.
As expected from the previous analysis of the monthly data corrected for the season, we find seriously high autocorrelations in the yearly data as well.

The one of the four Gauss-Markov assumptions about exogenous independent variables has not been covered yet in this paper. This is because this cannot be done with the limited dataset used in this section about introducing linear regression. However, this is also true for most research problems in time series analysis. It is not feasible in practical work to include all possible factors in multivariate models and analyse the problem by co-linearity analysis or factor models. So, the researcher needs a good theoretical understanding of the context of the data on which he wants to fit a model. This is not only true for simple linear regression models, but also for more sophisticated extensions covered in other sections of this book.

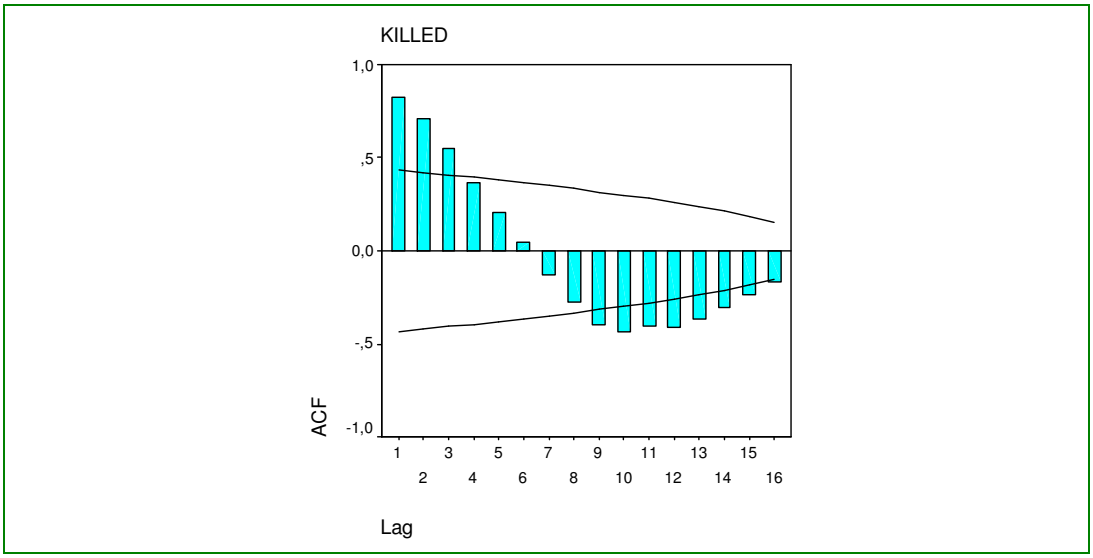


Figure 3.2.10 Table of autocorrelations

As already mentioned in the introduction to the time series analysis section, in principle there is nothing wrong in fitting a classical regression model with Austrian fatality data to obtain a rough idea of the linear trend in the series. The results show a negative relation between the number of Austrian fatalities and time, suggesting that the number of fatalities have decreased over the last 18 years. However, as soon as standard statistical tests are applied to ascertain whether or not the relationship should be attributed to chance, serious problems arise. As noted above, the *F*-test (or, equivalently, the *t*-test for the regression weight) would lead one to conclude that the negative relationship between the number of driver fatalities and time is highly significant. These tests are based on the fundamental Gauss-Markov assumptions. In the examples shown, especially the most important assumption of randomly distributed errors was clearly violated, implying that the results of the statistical tests regarding the regression could not be trusted.

3.2.1.5 Conclusion

For most studies, the fit of a linear regression model is a good start to examine the different properties of the data, and if all conditions hold true, it is the most efficient way to estimate a trend in a time series. This is true not only from a statistical viewpoint, but also for communicating the solution. The parameters in the model are simple and also non-statisticians can have an intuitive understanding of the results. This is an important issue in road safety work, where people have to make decisions which are costly both in terms of money and fatalities.

In a risk management environment, not only the general trend is important, but decisions are most often based on statistical inference. Therefore, it is important to analyse all the model assumptions. This analysis is also a good start to decide the direction of more advanced modelling of the data. In the example

shown above with time dependent errors, further investigation of the data will lead to dedicated time series models, which can handle this problem much better than classical regression. Other violations of the assumptions may lead to alternate estimation procedures. Weighted least squares or maximum likelihood techniques are options in the case of heteroscedastic data.

Other sections in this chapter on time series analysis will lead to an in-depth view of the various options to handle the specific properties of accident data in more sophisticated model environments.

3.2.2 Generalized linear models (GLM)

George Yannis, Constantinos Antoniou and Eleonora Papadimitriou (NTUA)

3.2.2.1. Objective of the technique

While the linear regression model is simple (to run and interpret), elegant and efficient, it is subject to the fairly stringent Gauss-Markov assumptions (Washington et al., 2003). The Gauss-Markov assumptions require:

- Linearity (in the parameters; nonlinearity in the variables is acceptable);
- Homoscedasticity;
- Exogenous independent variables;
- Uncorrelated disturbances; and
- Normally distributed disturbances

If these assumptions hold, it can be shown that the solution obtained by minimizing the sum of squared residuals ('least squares') is BLUE, i.e. best linear unbiased estimator (in other words, it is unbiased and has the lowest total variance among all unbiased linear estimators). These assumptions, however, are often violated in practice. In this research, two of these violations -that are relevant to road safety data- are considered, in particular correlated disturbances; and non-normal error structures.

Generalized linear models (GLM), a generalization of the linear regression, can be used to overcome these restrictions (McCullagh and Nelder, 1989, Dobson, 1990, Gill, 2000). The objective of GLM is to allow for more flexible error structures (besides the Gaussian which is assumed by –linear and nonlinear– regression). The allowable distributions belong in the exponential family. In this section, we investigate the suitability of each distribution for road safety data that are temporally correlated.

3.2.2.2. Model definition and assumptions

Generalized linear models facilitate the analysis of the effects of explanatory variables in a way that closely resembles the analysis of covariates in a standard linear model, but with less confining assumptions. This is achieved by specifying a *link function*, which links the systematic component of the linear model with a wider class of outcome variables and residual forms (McCullagh and Nelder, 1989, Dobson, 1990, Gill, 2000).

A key point in the development of GLM was the generalization of the normal distribution (on which the linear regression model relies) to the exponential family of distributions. This idea was developed by Fisher (1934). Consider a single random variable y whose probability (mass) distribution (if it is discrete) or probability density function (if it is continuous) depends on a single parameterθ. Probability (mass) distribution is the set of values x taken by a discrete random

variable X (the domain of the variable) and their associated probabilities. If X is a continuous random variable, the probability associated with any particular point is zero; therefore, positive probabilities can only be assigned to intervals in the range over which x is defined. In that case, the probability density function is defined by the area under the distribution in the range of the interval of interest.

The distribution belongs to the exponential family if it can be written in the form:

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \qquad (3.2.13)$$

where a, b, s, and t are known functions. The symmetry between y and θ becomes more evident if we rewrite it as:

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \qquad (3.2.14)$$

where s(y)=exp[d(y)] and t(θ)=exp[c(θ)]. If a(y)=y then the distribution is said to be in the canonical form. Furthermore, any additional parameters (besides the parameter of interest θ) are regarded as nuisance parameters forming parts of the functions a, b, c, and d, and they are treated as though they were known. Many well-known distributions belong to the exponential family, including –for example– the Poisson, normal, and binomial distributions. On the other hand, examples of well-known and widely used distributions that cannot be expressed in this form are the student's t-distribution and the uniform distribution.

The generalized linear model can be defined in terms of a set of N independent random variables y_1, … , y_N, each with a distribution from the exponential family with the following properties:

1. The distribution of each y_i is of the canonical form and depends on a single parameter θ_i (not necessarily the same parameter for all variables):

$$f(y_i;\theta_i) = \exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)] \qquad (3.2.15)$$

2. The distributions of all the y_i s are of the same form (e.g. all normal or all binomial) so that the subscripts on b, c, and d are not needed.

The joint probability density function of $y_1, \ldots, y_n$ is then

$$f(y_i;\theta_i) = \exp\left[\sum_{i=1}^{N}(y_i b(\theta_i) + c(\theta_i) + d(y_i))\right] \qquad (3.2.16)$$

When specifying a model, the *N* parameters θ_i are usually not of direct interest (the number of parameters θ is N, since there is one for each y). Instead, for a GLM, a smaller set of *p* parameters $\beta_1, \ldots, \beta_p$ is considered (where $p < N$), such that a linear combination of the βs is equal to some function of the expected value μ_i of y_i, i.e.

$$g(\mu_i) = x_i^T \beta \qquad (3.2.17)$$

where,
g is a monotonic, differentiable function called the link function;
$\mathbf{x}_i$ is a (p x 1) vector of explanatory variables (covariates and dummy variables for levels of factors); and
$\boldsymbol{\beta}=[\beta_1, \ldots, \beta_p]^T$ is the (p x 1) vector of parameters.

To recapitulate, in the univariate case, a generalized linear model has three components:
1. A response variable y assumed to follow a distribution from the exponential family;
2. A set of parameters $\boldsymbol{\beta}$ and explanatory variables $\mathbf{X}=[\mathbf{x_1}^T, \ldots, \mathbf{x_p}^T]^T$
3. A monotonic link function g such that

$$g(\mu_i) = x_i^T \beta \qquad (3.2.18)$$

where $\mu_i=E(Y_i)$

Generalized linear models require uncorrelated observations. Time-series data require special consideration, since the observations typically fail to meet this assumption, as neighbouring observations are likely to be correlated. It is often possible to include a large number of explanatory variables in a linear regression model, resulting in seemingly serially uncorrelated residuals (and, therefore, the linear model theory would apply). There are, however, two problems with such a strategy. First, it may not be easy to identify the appropriate explanatory variables that would reflect the serial correlation. Second, and perhaps more important, the additional variables included in the model to reduce the serial correlation may dilute the effects of the main variables of interest, thus potentially affecting the power and the interpretation of the model.

In a very different (with respect to road safety) context, Zeger (1988) introduced a method for regression when the outcomes are a time series of counts (as is often the case in road safety applications). The critical point about this model is that the serial correlation in the observed data is captured through some unobserved (or latent) process and conditional on this unobserved process, the counts are independent. This is a reasonable assumption for road safety data, since the occurrence of an accident (or a fatality or injury) is *usually* not directly caused by another.

The data, however, are serially correlated because they are ordered in time, and other factors (also ordered in time) are affecting the underlying risk. A discussion on these properties, albeit in a totally different context, can be found in Campbell (1994), who also presents a practical application of the approach, where the only assumption that is made on the distribution of the error structure is that it is mean stationary. Davis et al. (2000) developed a practical approach

to diagnose the existence of a latent stochastic process in the mean of a Poisson regression model.

For the Poisson model, the covariance matrix, and hence the standard errors of the parameter estimates, are estimated under the assumption that the Poisson model is appropriate. Occasionally one may observe more variation in the response than what is expected by the Poisson assumption. This is called overdispersion and implies that the estimates of the standard errors of the parameters will not be correct. Overdispersion typically occurs when the observations are correlated, and therefore it is very relevant in the context of time-series analysis. Underdispersion (less variation than expected) is also possible, although not as common.

The Poisson distribution has been considered suitable to counts of car crashes for a long time (Nicholson and Wong, 1993). However, the Poisson model (while arguably more appropriate than the Gaussian) is not without weaknesses and technical difficulties. For example, the assumption of a pure Poisson error structure may prove inadequate in the presence of "overdispersed" data (Maycock and Hall, 1984). A straightforward approach to overcome this issue is to use a quasi-Poisson model (i.e. estimate a dispersion parameter for the Poisson model, thus allowing it to take values other than 1). Maycock and Hall (1984) showed that the negative binomial model could also be used as an extension to the Poisson. Miaou (1994) and Wood (2002) have also used the negative binomial model for road safety applications. Maher and Summersgill (1996) mention that, quite often, the two approaches (i.e. quasi-Poisson and negative binomial) may give very similar estimation results. One may then be tempted to think that the two models are equivalent and that it does not really matter which model is selected. Maher and Summersgill further warn that this may not be the case, as the two models may have different prediction properties, as measured, e.g. by the prediction error variance.

Furthermore, few processes are adequately modelled by linear models in practice. For example, several researchers have shown that conventional linear regression models lack the distributional property to adequately describe collisions. This inadequacy is due to the random, discrete, non-negative, and typically sporadic nature that characterizes the occurrence of a vehicle collision. Several researchers (including Hauer et al.1988, Hakim et al., 1991; Cameron et al., 1993; Newstead et al., 1995), using road accident statistics, have presumed that the explanatory variables have a multiplicative effect on accidents (as opposed to e.g. additive).

3.2.2.3. Introduction of dataset and research problem

The use of generalized linear models for road safety research is demonstrated using accident casualties and police enforcement data from Greece (excluding the two largest cities, i.e. Athens and Thessalonica). Monthly data from January 1998 to December 2003 have been used for this research (Figure 3.2.12). The

data of the first five years (60 observations) are used for the model estimation, while the data for the last year (12 observations) are used for validation.

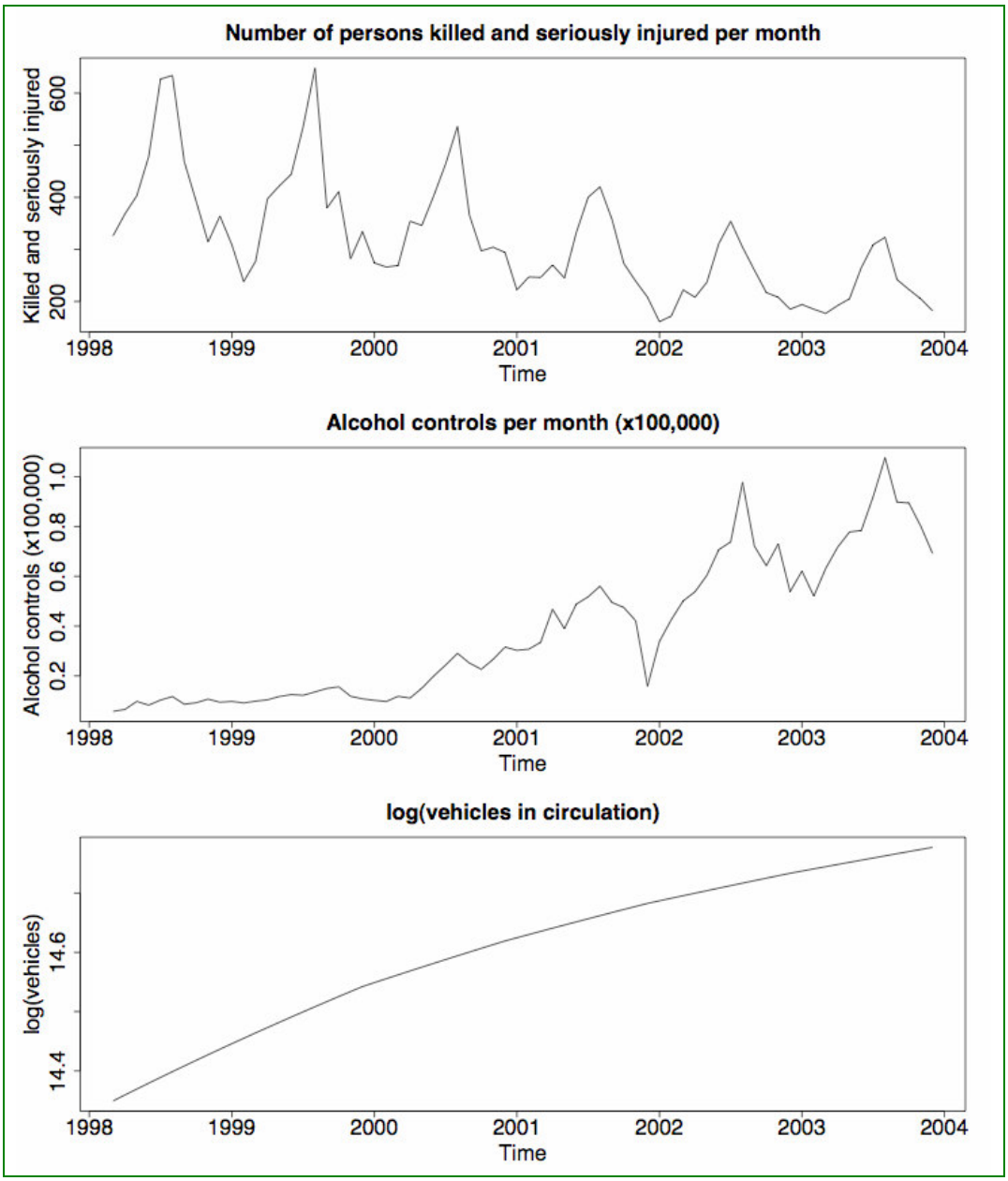


Figure 3.2.12: Dataset overview

The model specification comprises three main effects: trend, seasonal effects, and explanatory variables. The trend captures the evolution of the dependent variable over time. This is captured in the specification by the addition of the "Month" variable, which ranges from 1 (for the first month, i.e. January 1998) to 72 (for December 2003). Seasonal effects are captured by the incorporation of sinusoid components (similar to those used e.g. by Zeger, 1988, and Campbell, 1994). Several frequencies have been investigated (from 1 to 15 months), but the most useful proved to be the annual and its first (six month) harmonic.

Furthermore, besides specifying trend and seasonal components, the impact of explanatory variables is also tested, with an emphasis on enforcement data (number of breath alcohol controls per month) and (the log of) vehicles in circulation. To account for the delayed impact of enforcement in road safety (as the word-of-mouth spreads) the number of breath alcohol controls has been lagged by two intervals, capturing the impact of enforcement intensification two months after it occurs. The log of vehicles in circulation has been entered as an offset. This modeling decision was based on the comparison of this model and a model in which the vehicles in circulation were entered as a regular variable (however, that model led to counterintuitive parameter estimates). Naturally, the two major Greek urban areas excluded from the casualty data have also been excluded from the data of breath alcohol controls and registered vehicles. The number of registered vehicles has been interpolated from annual figures. Finally, a high number of casualties was observed during the month of August. Therefore, a binary dummy variable has been introduced, that takes the value of one for August and zero otherwise. Further exploration of the available monthly data did not reveal any new insight in the seasonality of the road safety phenomenon. The "August phenomenon" remained predominant.

Seasonality (August peak) observed mainly in the persons killed and seriously injured but also on the enforcement can be attributed to increased summer traffic in Greece as a holiday destination. The exceptional enforcement low value on December 2001 cannot be explained by any other reason than the internal enforcement programming of the Police.

3.2.2.4. Model fit, diagnostics and interpretation

In this section, different error structures -that are allowable within the GLM framework and are also theoretically supported- are applied. Model estimation and analysis has been performed using the R Software for Statistical Computing (RDCT, 2006). First, the Gaussian (Normal) distribution is used. A Poisson model is also fitted, along with a quasi-Poisson that relaxes the assumption that the dispersion parameter is equal to one. Finally, a negative binomial model is fitted. The link function used for all four models (Normal, Poisson, quasi-Poisson and negative binomial) is the log function.

Estimation results and model fit for the four model families are shown in Table 3.2.10 A sinusoid term with an annual frequency and its (6 month) harmonic capture periodicity. A negative coefficient value for the number of breath alcohol controls indicates that the number of persons killed and seriously injured decreases as the intensity of breath alcohol controls increases, which is an intuitive result.

A binary dummy variable, taking the value of one for August and zero otherwise, was also found to be significant. Other explanatory variables (such as the number of speeding violations) were also originally entered into the model. However, explanatory variables relating to enforcement were highly correlated

(in particular the number of breath alcohol controls and speeding violations had a correlation of 0.97). Therefore, while using either variable resulted in intuitive results, their combination resulted in multicollinearity problems.

The coefficient signs, however, are consistent for all models and all retained parameters are significant at the 1% level (with the exception of the enforcement data in the quasi-Poisson and negative binomial models, which are still significant at the 10% level). A comparison of the standard errors shows that the values obtained for the Poisson model are significantly lower than those obtained from the other three models. Therefore, the z-values obtained for the Poisson model seem unusually high. A closer look at the model statistics suggests that the data may be overdispersed.

Potential overdispersion can be identified by dividing the residual deviance (defined -up to a constant- as twice the log-likelihood ratio statistic) by the residual degrees of freedom (i.e. the number of observations minus the number of parameters in the model). The resulting measure is an approximately unbiased estimator of the dispersion parameter (Venables and Ripley, 2002). If the deviance is equal to the degrees of freedom then there is no evidence of overdispersion. Note that a dispersion parameter not equal to one does not necessarily imply overdispersion, but could also indicate other problems, such as an incorrectly specified model or outliers in the data. An incorrectly specified model can be due to an incorrectly specified functional form (an additive rather than a multiplicative model may be appropriate) or, more likely, that important explanatory variables (or interactions) are missing from the model. However, overdispersion can also be a property of the data, typically indicating a lack of independence or heterogeneity among observations, sampling issues, etc.

The dispersion factor for the data at hand is equal to 151.11/51=2.96, which is significantly different from one. The assumption of a Poisson model (with a dispersion parameter equal to one) is therefore unlikely to be realistic. A quasi-Poisson model (an extension of the Poisson model, in which the dispersion parameter is allowed to vary from one) has also been estimated. The estimation is based on the iterative algorithm proposed by Breslow (1984) for fitting overdispersed log-linear Poisson models. The magnitude of the estimated coefficient values is similar to that obtained by the Poisson model, and the signs are the same. The significance of the coefficients, however, has significantly decreased, indicating that in the Poisson model the standard errors were underestimated due to the overdispersion. As expected, the dispersion parameter for the quasi-Poisson model is 51.38/51=1.01, i.e. very close to one.

Finally, a negative binomial model has been fitted. The estimated coefficients were similar to those obtained from the quasi-Poisson. This confirms the findings of Maher and Summersgill (1996) who state that the two approaches may provide similar estimation results. Slightly lower standard errors for the binomial, however, lead to more significant statistics.

Further model diagnostics are presented in Figures 3.2.13 through 3.2.16. Normal scores plot (QQ plot) of standardized deviance residuals is presented in

the left subfigure of each figure. The x-axis represents the standardized deviance residuals, while the y-axis represents the quantiles of the standard normal. The dotted line in the QQ plot (left) is the expected line if the standardized residuals are normally distributed, i.e. it is the line with intercept 0 and slope 1. If the deviance residuals are normally distributed, all points on the plot would fall on this dotted line. The deviance residuals of the normal model are far from normally distributed. The Poisson model is a slight improvement, but still far off. The quasi-Poisson and the negative binomial model deviance residuals, on the other hand, are practically normally distributed. While normality of the residuals is not a requirement of the generalized linear model, it is an indication of a well-behaved model specification.

On the right subfigure is a plot of the Cook statistics against the standardized leverages. The standardized leverage of the i-th observation x_i can be computed as (Belsley et al., 1980):

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_i)}{(n-1)s_x^2}$$

where n is the number of observations, the overbar indicates the predicted value, and s_x is the standard error. There are two dotted lines on each plot. The horizontal line is at 8/(n-2p) where n is the number of observations and p is the number of parameters estimated. Points above this line may be points with high influence on the model. The vertical line is at 2p/(n-2p) and points to the right of this line have high leverage compared to the variance of the raw residual at that point. If all points are below the horizontal line or to the left of the vertical line then the line is not shown.

A large number of points appear to be influential (i.e. above and to the right of the two dashed lines) in the Gaussian and the Poisson models, while only one point has a high leverage for the quasi-Poisson and negative binomial models.

The estimation results and the model diagnostics suggest that the quasi-Poisson and the negative binomial assumptions are more valid for the considered problem (while this may not be always the case). The output of the resulting models is very similar and therefore a clear decision regarding the most appropriate model cannot be made. One observation relates to the estimated standard errors, which are higher for the quasi-Poisson. Choosing to err in the side of caution, one could retain this model.

It should be noted that the usual tests for comparing models, such as the Akaike Information Criterion, AIC, (Akaike, 1973) or the Schwarz/Bayesian Information Criterion, BIC, (Schwarz, 1978), are not suitable for comparison across these models. (While a detailed discussion is outside of the scope of this document, and there is a lot of specialized research on the topic, the AIC is best suited for the comparison of nested models and models with similarly computed log-likelihood measures. In this application, for example, the quasi-Poisson model is not estimated using maximum likelihood.)

		Normal	
Coefficient	Estimate	Std. error	t-value
Intercept	-7.9608	0.1175	-67.763
	-0.0154	0.0017	-9.054
August dummy	0.1995	0.0355	5.628
sin(pi*Month/6)	-0.2279	0.0215	-10.580
sin(pi*Month/12)	-0.5326	0.1826	-2.917
cos(pi*Month/6)	-0.4434	0.0781	-5.674
Laggedx2 alcohol controls (x100,000)	-0.2949	0.1481	-1.992
Null deviance:		*725 608*	*(57 d.o.f.)*
Residual deviance:		*79 290*	*(51 d.o.f.)*
		Poisson	
Coefficient	Estimate	Std. error	z-value
Intercept	-7.9881	0.0641	-124.548
Trend	-0.0157	0.0010	-15.921
August dummy	0.1919	0.0241	7.963
sin(pi*Month/6)	-0.2229	0.0123	-18.162
sin(pi*Month/12)	-0.4859	0.0985	-4.932
cos(pi*Month/6)	-0.4214	0.0430	-9.803
Laggedx2 alcohol controls (x100,000)	-0.2629	0.0821	-3.201
Null deviance:		*2 042.30*	*(57 d.o.f.)*
Residual deviance:		*168.42*	*(51 d.o.f.)*
		Quasi-Poisson	
Coefficient	Estimate	Std. error	z-value
Intercept	-8.0038	0.1066	-75.068
Trend	-0.0159	0.0017	-9.470
August dummy	0.1838	0.0466	3.949
sin(pi*Month/6)	-0.2206	0.0212	-10.427
sin(pi*Month/12)	-0.4582	0.1623	-2.824
cos(pi*Month/6)	-0.4087	0.0718	-5.692
Laggedx2 alcohol controls (x100,000)	-0.2410	0.1368	-1.761
Null deviance:		*568.12*	*(57 d.o.f.)*
Residual deviance:		*51.41*	*(51 d.o.f.)*
		Negative binomial	
Coefficient	Estimate	Std. error	z-value
Intercept	-8.0027	0.1007	-79.434
Trend	-0.0159	0.0016	-10.022
August dummy	0.1843	0.0436	4.229
sin(pi*Month/6)	-0.2208	0.0199	-11.071
sin(pi*Month/12)	-0.4602	0.1534	-2.999
cos(pi*Month/6)	-0.4096	0.0678	-6.038
Laggedx2 alcohol controls (x100,000)	-0.2425	0.1293	-1.875
Null deviance:		*682.57*	*(57 d.o.f.)*
Residual deviance:		*58.78*	*(51 d.o.f.)*

Table 3.2.10. Estimation results

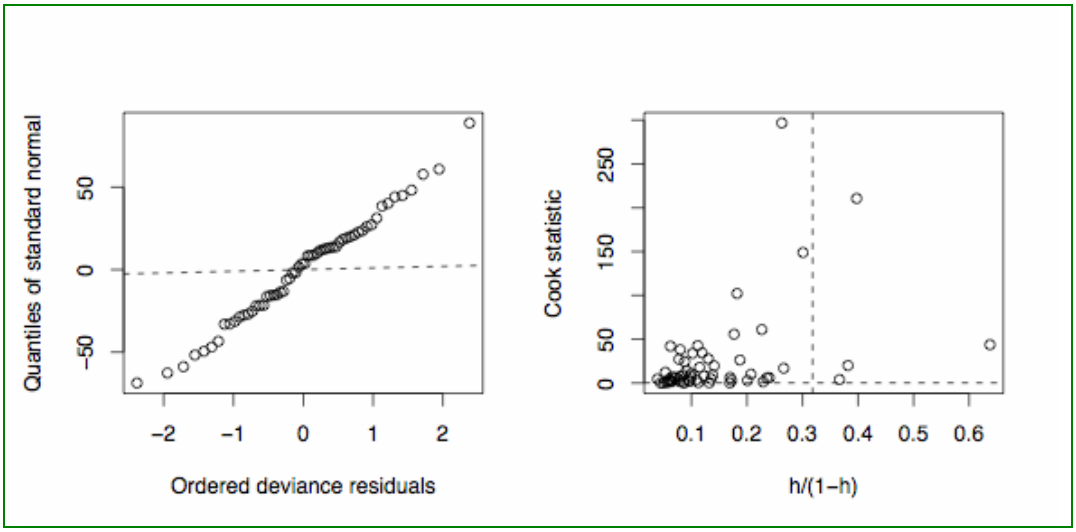


Figure 3.2.13: Model fit diagnostic plots (Gaussian distribution)

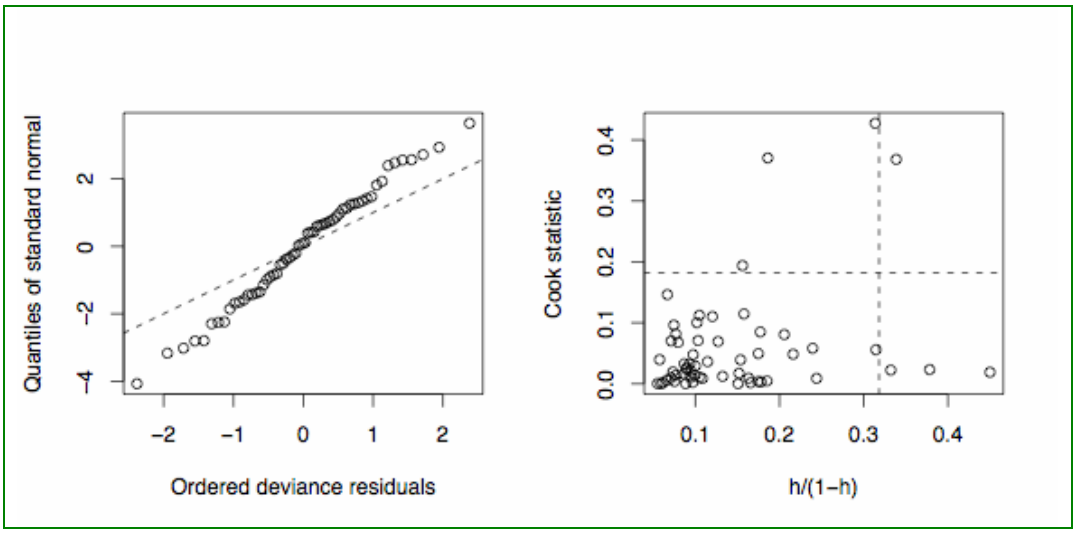


Figure 3.2.14: Model fit diagnostic plots (Poisson distribution)

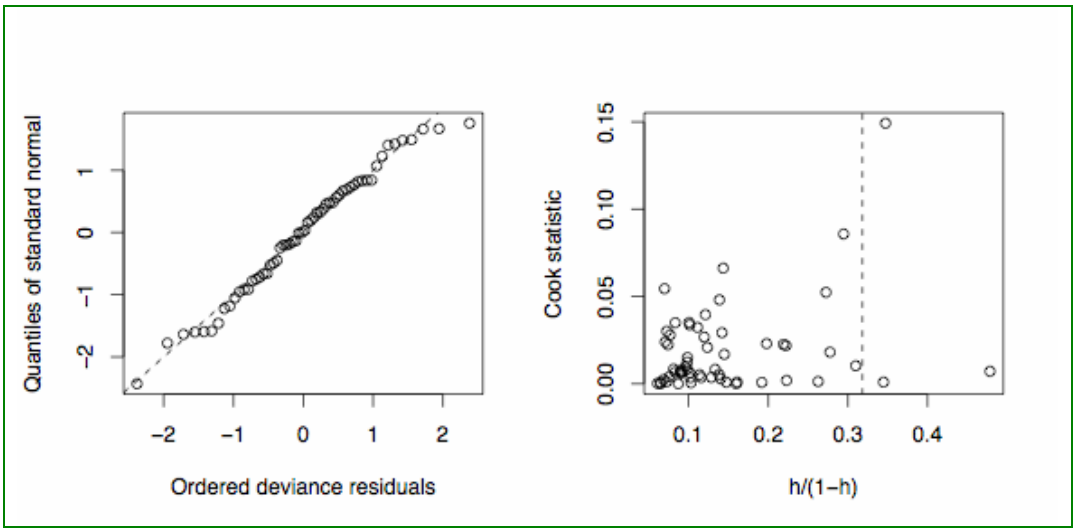


Figure 3.2.15: Model fit diagnostic plots (Quasi-Poisson distribution)

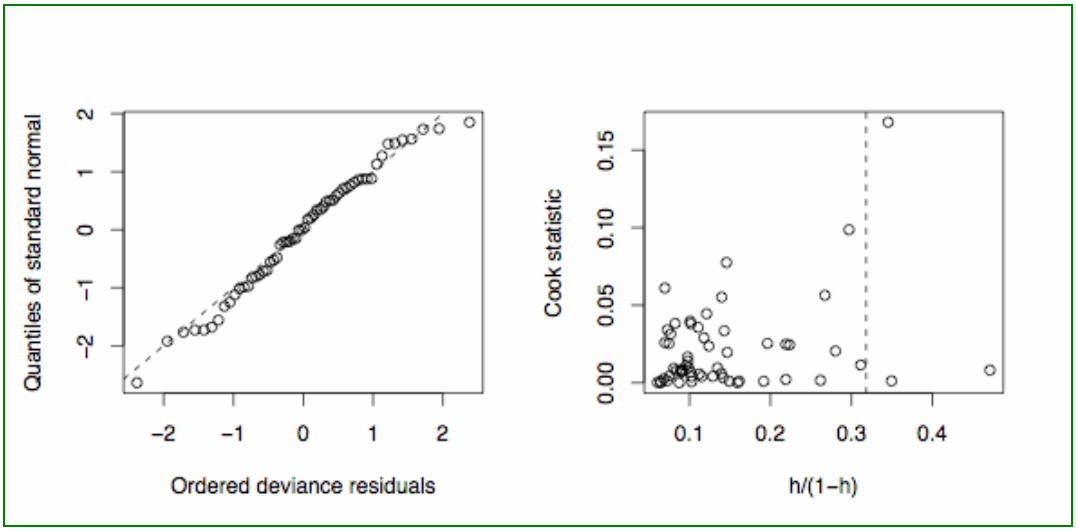


Figure 3.2.16: Model fit diagnostic plots (Negative binomial distribution)

Figure 3.2.17 shows the values predicted by the quasi-Poisson model. The dashed line shows the actual observed number of dead and seriously injured in Greece (excluding the two major metropolitan areas of Athens and Thessalonica). The thick solid line represents the model predictions and 95% confidence intervals are also shown with thinner solid lines.

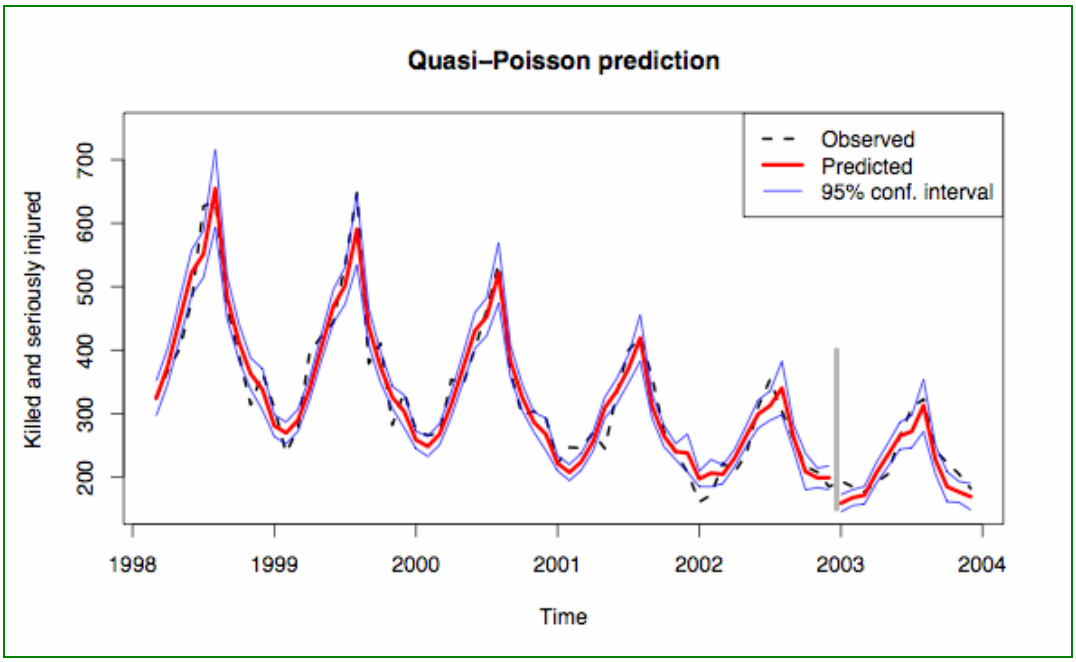


Figure 3.2.17: Quasi-Poisson model predictions

3.2.2.5. Conclusion

The impact of different distributional assumptions for the dependent variables on the model estimation results is demonstrated in this research within the unified framework of generalized linear models. Due to the time-series nature of the data, a modelling approach to capture serial correlation through the introduction of sinusoid latent processes has also been demonstrated.

The signs of the estimated coefficients for all models are consistent and intuitive. The estimated coefficients for the Poisson model are close to those estimated by the other three models, but the standard errors are severely underestimated (due to overdispersion), leading to artificially high t-statistic values. Even though these values were indeed significant in this application, this issue could have led to incorrect retention of insignificant variables in the Poisson model. As a result, the use of the Poisson model in this case is not recommended, and the quasi-Poisson or the negative binomial models should be used instead. However, even though the magnitude of the estimated coefficients for the quasi-Poisson and negative binomial is very similar, the different models may have different predictive properties and therefore may not be used interchangeably without further analysis.

3.2.3 Non-linear models

George Yannis, Constantinos Antoniou and Eleonora Papadimitriou (NTUA)

3.2.3.1. Objective of the technique

While the linear regression model is relatively simple (to run and interpret), elegant and efficient, few processes are adequately modeled by linear regression models in practice. Linear regression models might have been a practical necessity in the past, but theoretical and computational developments have made the use of more elaborate (appropriate, accurate) methods practical. This can also be seen in road safety research, where while early work used multiple linear regression modeling (assuming normally distributed errors and homoscedasticity), over the past two decades there has been a departure from this model. Generalized linear models (GLM) allow for some nonlinear relationships to be modeled and relax some restrictions on the distributional assumptions of linear regression (McCullagh and Nelder, 1989, Dobson, 1990). Although many scientific and engineering processes can be described well using linear models, or other relatively simple types of models, there are many processes that are inherently nonlinear. Non-linear models need then be used. The biggest advantage of nonlinear regression over many other techniques is the broad range of functions that can be fit.

Non-linear regression is widely used in road-safety related research. Several researchers (including Oppe, 1979, 1989, Hauer, 1988, Hakim et al., 1991, Cameron et al, 1993; Newstead et al., 1995), using road accident statistics, have presumed that the explanatory variables have a multiplicative effect on accidents (as opposed to e.g. additive). Henning-Hager (1986) presented a non-linear regression model to express the relationship between road safety, traffic volumes and the quality of transportation supply and demand in urban areas. Qin et al. (2004) showed that the relationship between crashes and the daily volume (AADT) is non-linear and varies by crash type, and is significantly different from the relationship between crashes and segment length for all crash types.

A commonly used macroscopic road-safety model is based on Smeed's original relationship (Smeed, 1968):

$$\frac{F_n}{V_n} = \alpha\left(\frac{V_n}{P_n}\right)^{\beta} + Z_n \qquad (3.2.20)$$

where F is the number of fatalities, V is the number of (motor) vehicles (in thousands), P is the population (in thousands), n indicates the country, α and β are model parameters to be estimated and Z_n are the disturbances. Using data

for road fatalities, vehicles and population from 20 (mostly European) countries, Smeed (1968) estimated the values of a and b as 0.0003 and -0.66 respectively. Jacobs (1986) repeated this analysis for a number of developed and developing countries using data between 1968 and 1975 and obtained values of 0.000204 and -0.84 for a and b respectively. Gharaybeh (1994) applied Smeed's formula to assess the development of road safety in Jordan, relative to that of other middle-eastern and developing countries.

In this section, four different models are presented in the context of road safety. Using the model shown in Equation 3.2.20 as a starting point, a log-transformed version is developed. Autoregressive versions of these two models are also derived, to account for serially correlated disturbances. The four models are developed and estimated in parallel, and their results are compared.

3.2.3.2. Model definition and assumptions

A non-linear regression model is most commonly written as:

$$Y_n = f(x_n, \theta) + Z_n \tag{3.2.21}$$

where f is the expectation function, x_n is a vector of associated regressor variables or independent variables for the nth case, Y_n is the dependent variable, θ is a vector of parameters to be estimated and Z_n are random disturbances. This model is of the same general form as the linear model, with the exception that the expected responses are nonlinear functions of the parameters. More formally, for non-linear models, at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the parameters. The presentation of non-linear models on the following sections relies on Bates and Watts (1988), while the following paragraphs presenting the advantages and disadvantages of non-linear regression are based on NIST (2006).

Non-linear regression is estimated using "least squares" procedures, using the same underlying concepts as linear least squares regression. As a result, nonlinear least squares regression has some of the same advantages (and disadvantages) that linear least squares regression has over other methods. One common advantage is efficient use of data. Nonlinear regression can produce good estimates of the unknown parameters in the model with relatively small data sets. Another advantage that nonlinear regression shares with linear regression is a fairly well-developed theory for computing confidence, prediction and calibration intervals to answer scientific and engineering questions. In most cases the probabilistic interpretation of the intervals produced by nonlinear regression are only approximately correct, but these intervals still work very well in practice.

The major cost of moving to nonlinear least squares regression from simpler modeling techniques like linear least squares is the need to use iterative optimization procedures to compute the parameter estimates. With functions

that are linear in the parameters, the least squares estimates of the parameters can always be obtained analytically, while that is generally not the case with nonlinear models. The use of iterative procedures requires the user to provide starting values for the unknown parameters before the software can begin the optimization. The starting values must be reasonably close to the as yet unknown parameter estimates or the optimization procedure may not converge to the optimal point. Bad starting values can also cause the software to converge to a local minimum rather than the global minimum that defines the least squares estimates.

Disadvantages shared with the linear least squares procedure includes a strong sensitivity to outliers. Just as in a linear least squares analysis, the presence of one or two outliers in the data can seriously affect the results of a nonlinear analysis. In addition there are unfortunately fewer model validation tools for the detection of outliers in nonlinear regression than there are for linear regression.

The flexibility of non-linear regression is also a caveat, since similarly good fits can be obtained with very different functional forms (whereas presumably only one of them captures the modeled process). These different models may be adequate for interpolation purposes, but may produce very different predictions when used to extrapolate, i.e. predict values outside of the support of the estimation dataset. This is a very important point, that should never be treated lightly. It can (and it has) lead to seriously erroneous models, with potentially very misleading predictive properties, when applied to slightly different data. As a result, it is important to use appropriate tests and checks to ensure that the selected functional form is appropriate for the problem at hand. It should be noted, however, that this is not an exclusive property of non-linear regression, and other methods, including linear regression, suffer from this.

The assumptions from ordinary least square (OLS) procedures (normal, i.i.d. disturbances etc., sometimes collectively referred to as Gauss-Markov assumptions) still apply in non-linear regression. Therefore, whenever time or distance is involved as a factor in a regression analysis, it is important to check the assumption of independent residuals. When the residuals are not independent, the model for the observations must be altered to account for dependence (e.g. moving average or autoregressive models of variable order).

Road safety data are often correlated in space and/or time, raising the suspicion of correlated data (and hence residuals), which violates one of the underlying model assumptions (that of independent disturbances). Correlation of the disturbances can, for instance, be detected from an ordered time series plot of the residuals versus time or from a lag plot of the residuals on the nth case versus the residuals on the (n-1)th case. If a violation of independent disturbances is detected, then the model needs to be altered to account for this. Common forms for dependence, or autocorrelation, of disturbances are (combinations of) moving average and autoregressive models of a certain order (see e.g. Box et al., 1994).

A moving average process of order 1 can be written as:

$$Z_n = \varepsilon_n - \omega_1 \varepsilon_{n-1} \tag{3.2.22}$$

while an autoregressive process of order 1, can be expressed as:

$$Z_n = \varepsilon_n + \phi_1 Z_{n-1} \tag{3.2.23}$$

where $\varepsilon_n, n = 1,2,\ldots,N$ are white noise terms (i.e. independent normal error terms with zero mean and constant unit variance).

The problem to be fitted is

$$Y_n = f(x_n, \theta) + Z_n \tag{3.2.24}$$

where $Z_n = \varepsilon_n + \phi_1 Z_{n-1}$. In order to solve this problem by reducing it to a non-linear least squares problem, one can subtract ϕ times the equation for Y_{n-1} from Y_n, thus obtaining:

$$Y_n - \phi \cdot Y_{n-1} = f(x_n, \theta) - \phi \cdot f(x_{n-1}, \theta) + Z_n - \phi \cdot Z_{n-1} \tag{3.2.25}$$

which is equivalent to

$$Y_n = \phi \cdot Y_{n-1} + f(x_n, \theta) - \phi \cdot f(x_{n-1}, \theta) + \varepsilon_n \tag{3.2.26}$$

Substituting Equation 3.2.20 into Equation 3.2.26, the autoregressive non-linear model that corrects for temporal correlation is:

$$\left(\frac{F}{V}\right)_n = \phi \cdot \left(\frac{F}{V}\right)_{n-1} + \alpha\left(\frac{V}{P}\right)_n^{\beta} - \phi \cdot \alpha\left(\frac{V}{P}\right)_{n-1}^{\beta} + \varepsilon_n \tag{3.2.27}$$

The original non-linear model (Equation 3.2.20) can be converted to a similar (but not equivalent) linear model through a simple log transformation. For demonstration purposes, this model is presented next. Furthermore, in the rest of this section, the various models are developed in parallel. This approach demonstrates both the close relation of the models, but also highlights their differences.

Taking the log of both sides of Equation 3.2.20 (temporarily ignoring the additive error term), the following linear model is obtained (where log(α) has been simplified to α):

$$\log\left(\frac{F_n}{V_n}\right) = \alpha + \beta \log\left(\frac{V_n}{P_n}\right) \tag{3.2.28}$$

Adding an additive error term, the equation becomes:

$$\log\left(\frac{F_n}{V_n}\right) = \alpha + \beta \log\left(\frac{V_n}{P_n}\right) + Z_n \tag{3.2.29}$$

This equation is similar, but not equivalent to Equation 3.2.20. The difference is in the error term. If one takes the exponent of equation 3.2.29, the resulting equation is:

$$\frac{F_n}{V_n} = \alpha \left(\frac{V_n}{P_n}\right)^{\beta} \exp(Z_n) = \alpha \left(\frac{V_n}{P_n}\right)^{\beta} Z_n' \tag{3.2.30}$$

i.e. there is a multiplicative error term (as opposed to an additive error term in Equation 3.2.20).

An autoregressive version of Equation 3.2.30 can be constructed in a similar way to Equation 3.2.27:

$$\begin{aligned} \log\left(\frac{F}{V}\right)_n &= \phi \cdot \log\left(\frac{F}{V}\right)_{n-1} + \alpha - \phi \cdot \alpha + \beta \cdot \log\left(\frac{V}{P}\right)_n - \phi \cdot \beta \cdot \log\left(\frac{V}{P}\right)_{n-1} + \varepsilon_n \\ &= \phi \cdot \log\left(\frac{F}{V}\right)_{n-1} + (1-\phi) \cdot \alpha + \beta \cdot \log\left(\frac{V}{P}\right)_n - \phi \cdot \beta \cdot \log\left(\frac{V}{P}\right)_{n-1} + \varepsilon_n \end{aligned} \tag{3.2.31}$$

Note that the above model (Equation 3.2.31) is not linear in the parameters, due to the second and fourth right-hand terms (in particular $(1-\phi) \cdot \alpha$ and $\phi \cdot \beta$).

Rather than choosing a single model and using it to demonstrate nonlinear regression models in this section, four models are developed in parallel. The comparison of the model parameters, goodness-of-fit properties, and predictive ability of these models may help the reader better comprehend the theory, practice, advantages, and caveats of non-linear regression. In particular, Equations 3.2.20 (nonlinear), 3.2.26 (AR nonlinear), 3.2.29 (log-transformed) and 3.2.31 (AR log-transformed) are used. While the log-transformed model ends up being linear, the other three models are nonlinear.

3.2.3.3. Introduction of dataset and research problem

Aggregate fatality, population and vehicle data from European countries between 1970 and 2002 have been used. The first 25 years of the data (i.e. 1970-1994) have been used for fitting the models, while the last seven years (1995-2002) have been used for validating the estimated models. This way, i.e. through separating the available data into two groups, issues such as over-fitting are overcome. The data have been obtained primarily from IRTAD. Official representatives of the countries with missing data were contacted directly, and several responses with additional data were incorporated to the database. In the end, out of the 25 countries of the enlarged EU, sufficiently

complete data have been available for 17 of them, for which this model has been applied.

Figure 3.2.18 presents the following variables for the entire data-set:
- Fatalities per thousand vehicles (solid line, decreasing trend)
- Vehicles per population (dashed line, increasing trend)

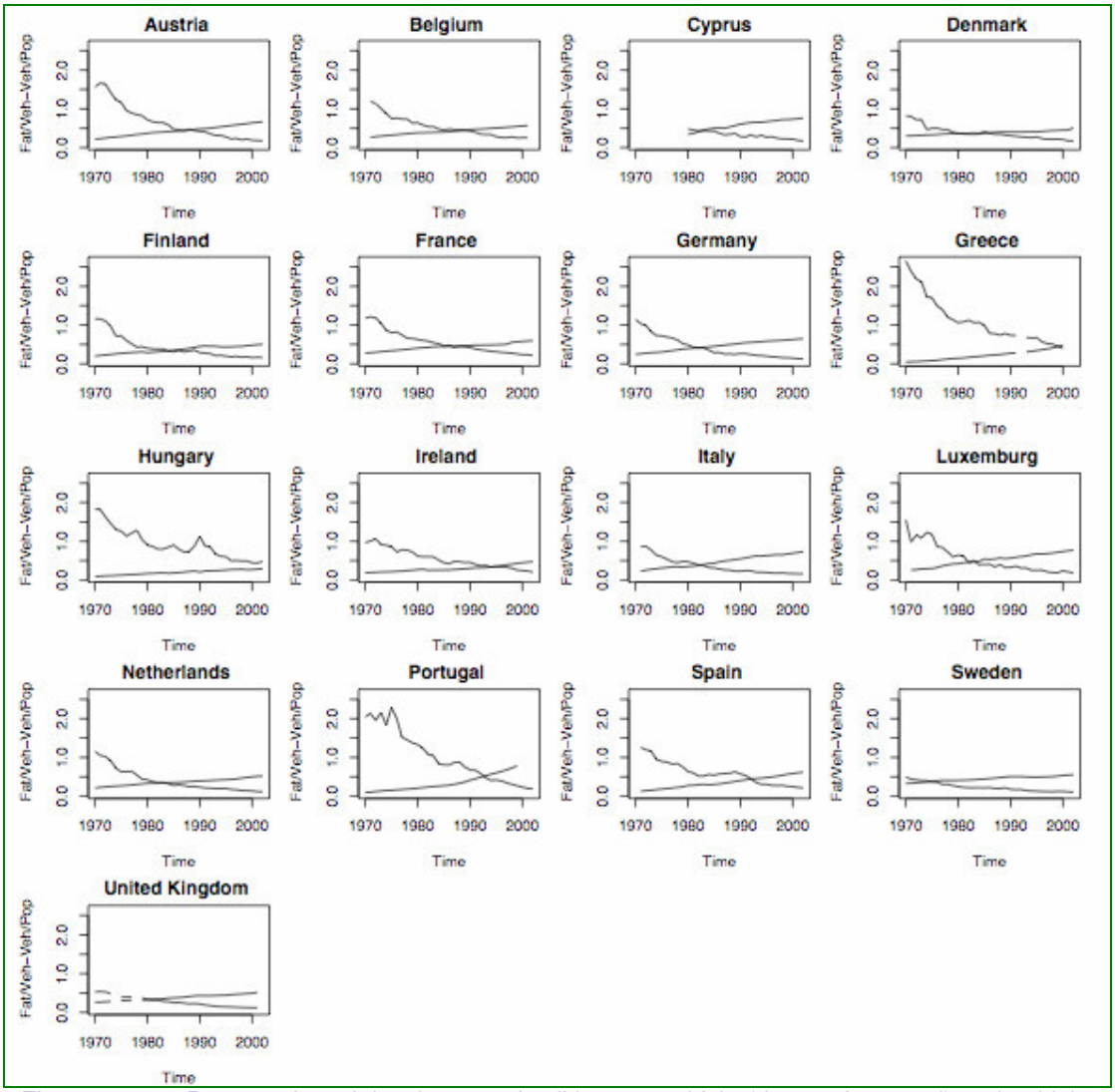


Figure 3.2.18. Presentation of the data set: fatalities per vehicle (decreasing trend) and vehicles per population (increasing trend)

3.2.3.4. Model fit, diagnostics and interpretation

The model shown in Equations 3.2.20 and 3.2.27 were estimated for the countries mentioned above and the estimated coefficients and statistics are shown in Table 3.2.11. All models in this section have been estimated using the R Software for Statistical Computing v. 2.3.0 (R, 2006). With the exception of the results for the autoregressive model for Sweden (SE), all parameters are very significant. The issue with the autoregressive model for Sweden

(highlighted in the Table 3.2.11) is the high estimated value for the coefficient φ, which is approaching 1.

	Coefficient a			Coefficient b		
	Estimate	Standard error	t-test	Estimate	Standard error	t-test
AT	0,099	0,007	14,962	-1,962	0,054	-36,252
BE	0,080	0,006	13,215	-2,068	0,069	-30,091
CY	0,219	0,018	12,262	-0,770	0,108	-7,158
DK	0,012	0,004	3,204	-3,477	0,291	-11,958
FI	0,026	0,006	4,597	-2,475	0,162	-15,263
FR	0,083	0,006	13,151	-2,153	0,073	-29,698
DE	0,070	0,006	12,469	-2,012	0,070	-28,597
EL	0,288	0,016	18,252	-0,711	0,023	-31,058
HU	0,172	0,028	6,260	-0,984	0,082	-11,987
IE	0,035	0,008	4,540	-2,075	0,151	-13,762
IT	0,081	0,006	14,078	-1,677	0,060	-27,834
LU	0,156	0,018	8,626	-1,542	0,104	-14,815
NL	0,017	0,002	8,384	-2,844	0,091	-31,123
PT	0,290	0,039	7,398	-0,956	0,075	-12,753
ES	0,212	0,017	12,716	-0,876	0,049	-17,784
SE	0,028	0,005	6,117	-2,596	0,175	-14,830
UK	0,030	0,003	11,403	-2,210	0,076	-28,933

	Coefficient a			Coefficient b			Coefficient phi		
	Estimate	Standard error	t-test	Estimate	Standard error	t-test	Estimate	Standard error	t-test
AT	0,090	0,010	9,303	-2,051	0,096	-21,484	0,3387	0,1255	2,699
BE	0,077	0,012	6,215	-2,111	0,158	-13,396	0,4487	0,197	2,277
CY	0,214	0,027	7,994	-0,815	0,180	-4,524	0,2047	0,2905	0,705
DK	0,015	0,010	1,550	-3,227	0,617	-5,228	0,5686	0,1695	3,355
FI	0,021	0,009	2,209	-2,687	0,370	-7,273	0,4647	0,1429	3,252
FR	0,068	0,016	4,329	-2,382	0,251	-9,494	0,5339	0,1798	2,970
DE	0,069	0,011	6,215	-2,034	0,153	-13,329	0,5282	0,1752	3,015
EL	0,294	0,025	11,740	-0,701	0,037	-19,013	0,3005	0,2131	1,410
HU	0,155	0,062	2,516	-1,045	0,218	-4,794	0,5825	0,1784	3,265
IE	0,034	0,015	2,295	-2,119	0,309	-6,865	0,6081	0,152	4,001
IT	0,071	0,009	8,243	-1,818	0,116	-15,687	0,3571	0,1819	1,964
LU	0,131	0,015	8,521	-1,757	0,114	-15,438	0,2458	0,1454	1,691
NL	0,015	0,003	4,986	-2,969	0,163	-18,200	0,3247	0,1364	2,380
PT	0,219	0,068	3,230	-1,154	0,196	-5,893	0,5303	0,1314	4,037
ES	0,135	0,054	2,473	-1,306	0,418	-3,122	0,7992	0,0688	11,619
SE	0,096	0,147	0,653	-0,146	2,131	-0,069	0,913	0,035	26,065
UK	0,025	0,006	4,244	-2,374	0,230	-10,333	0,5916	0,2184	2,709

Table 3.2.11. Non-linear model estimation results (top: base, bottom: after correcting for correlation)

Figure 3.2.19 and Figure 3.2.20 show the main diagnostics for the estimated nonlinear models (as per Equation 3.2.20). For each country, the residuals per observation are plotted, followed by the autocorrelation function (ACF) and the partial ACF (PACF). Note that PACF plots start at lag 1 while ACF plots start at 0. It is clear from these figures that the assumption of independent disturbances is violated for most countries. For some countries (such as Hungary, Spain and Sweden), both the residuals' plot and the ACF plot suggest violations. The different combinations of violations in these plots suggest that different approaches may be required to correct the model for each country. In this

paper, however, we will instead follow a unified approach. In most of the ACF plots, the correlation decays quickly and falls below the limits (indicated with the dotted lines) after one or two intervals. Please note that lag-0 autocorrelations have a value of 1 by definition. Therefore the fact that these values exceed the limits should not be interpreted as a violation of assumptions. Both the apparent exponential decay of the autocorrelations and the presence of a significant partial autocorrelation of order 1 suggest that a first order autoregressive process may be able to capture the correlation of the residuals. This is confirmed, as the autocorrelation are mostly dealt in the residuals of the autoregressive models (as per Equation 3.2.27), diagnostics for which are provided in Figure 3.2.21 and Figure 3.2.22.

The estimated coefficients of the log-transformed models are shown in Table 3.2.12. The model shown in equation 3.2.29 is shown on top, followed by the model presented in equation 3.2.31. Similarly to the non-linear model (Table 3.2.11), the estimation results are unreliable for models with estimated values for φ very close to 1 (such as Finland, Germany, Ireland, Sweden and United Kingdom, highlighted in the table). The term "unreliable" here is used to convey inconsistency with expectations about these values, e.g. sign and magnitude.

This issue requires some further discussion. This finding can be an indication that the data need to be differenced (an approach that is discussed later in this document). Furthermore, it appears that values of φ up to at least 0.85 result in “stable” models, while values above 0.94 result in “unreliable” models. This is an indication that the critical value lies somewhere between these two values. Finally, an indication of an “unreliable” model may be the very high t-statistic of the estimated φ coefficients.

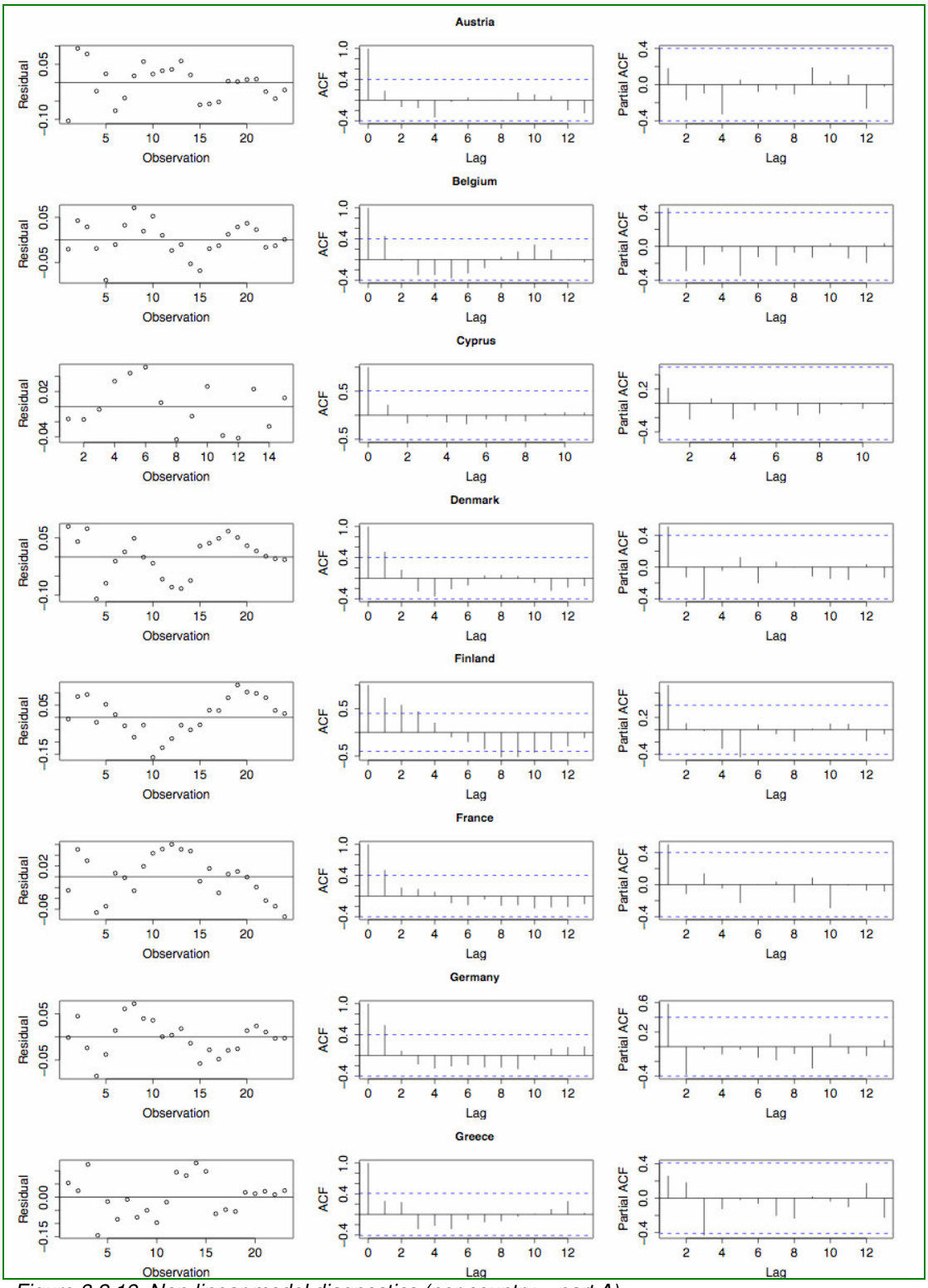


Figure 3.2.19. Non-linear model diagnostics (per country – part A)

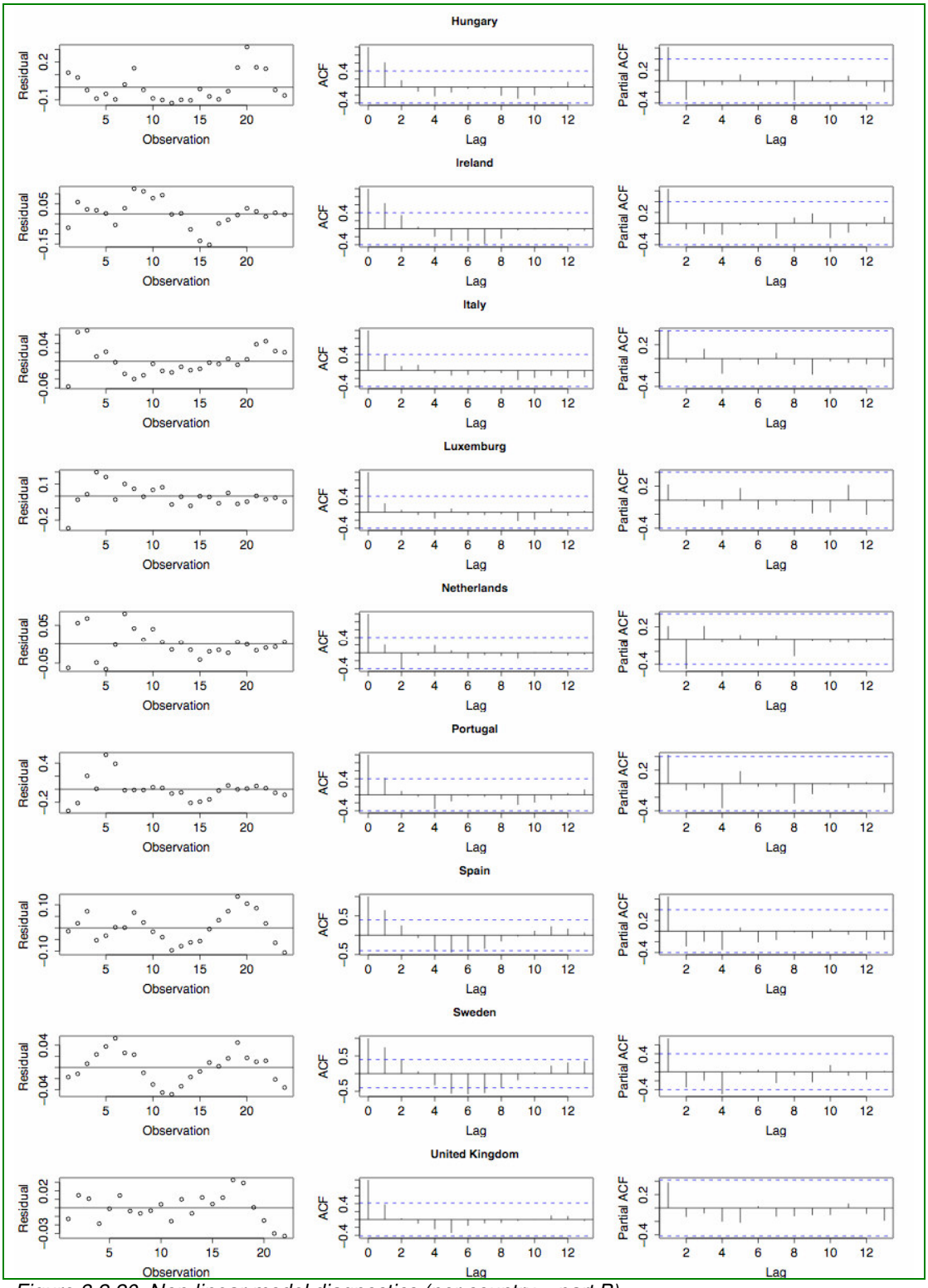


Figure 3.2.20. Non-linear model diagnostics (per country – part B)

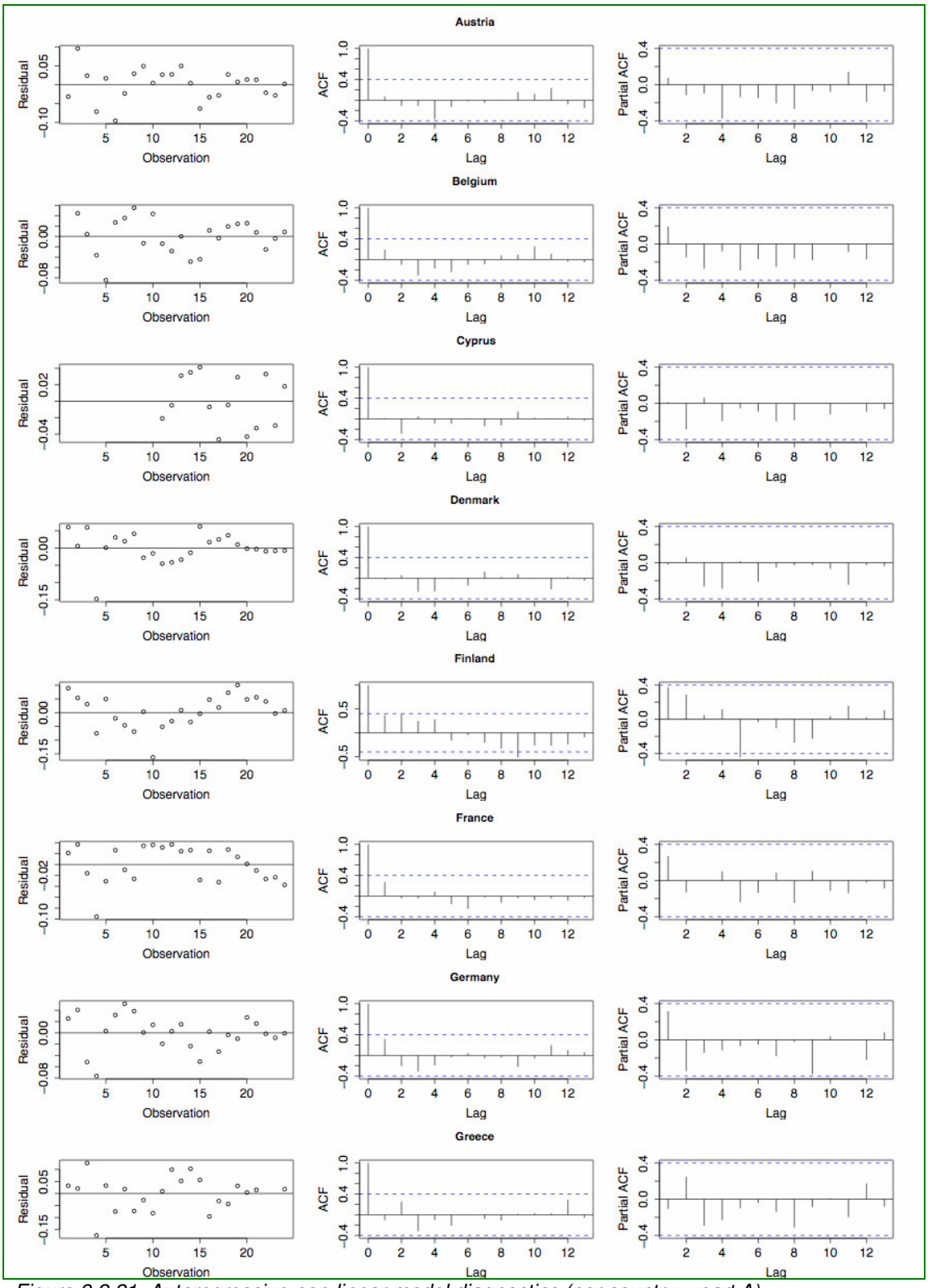


Figure 3.2.21. Autoregressive non-linear model diagnostics (per country – part A)

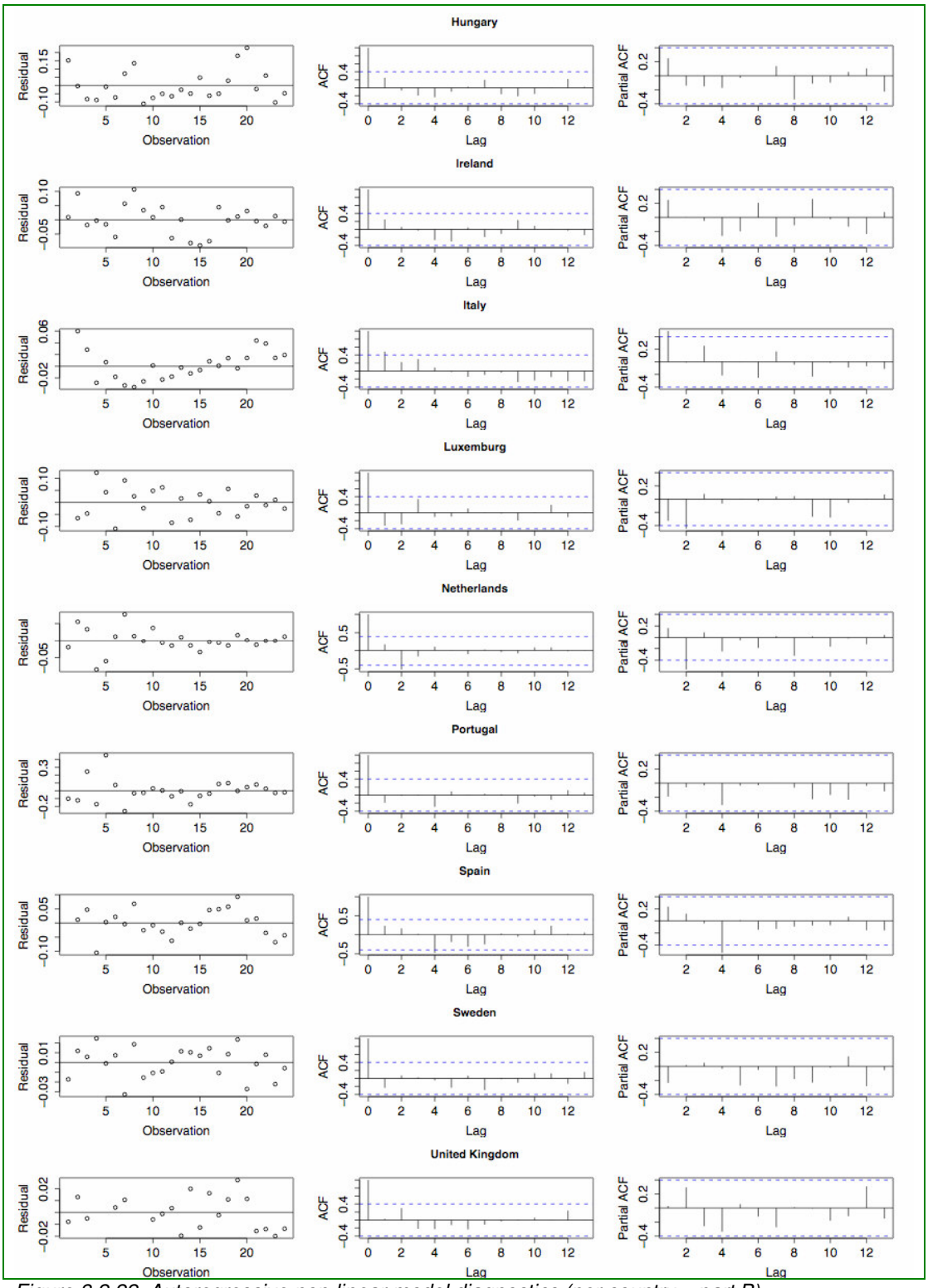


Figure 3.2.22. Autoregressive non-linear model diagnostics (per country – part B)

	Coefficient a			Coefficient b		
	Estimate	Standard error	t-test	Estimate	Standard error	t-test
AT	-2,395	0,057	-42,122	-2,031	0,057	-35,857
BE	-2,521	0,074	-33,904	-2,056	0,077	-26,896
CY	-1,555	0,083	-18,642	-0,818	0,120	-6,808
DK	-4,004	0,278	-14,423	-3,047	0,273	-11,160
FI	-2,985	0,188	-15,884	-1,916	0,166	-11,528
FR	-2,565	0,091	-28,050	-2,236	0,100	-22,443
DE	-2,715	0,068	-40,134	-2,056	0,073	-28,224
EL	-1,210	0,048	-25,150	-0,694	0,024	-28,465
HU	-1,647	0,172	-9,569	-0,919	0,096	-9,567
IE	-3,353	0,224	-14,995	-2,073	0,164	-12,682
IT	-2,395	0,051	-46,635	-1,558	0,054	-29,118
LU	-1,994	0,071	-27,923	-1,671	0,082	-20,433
NL	-4,187	0,088	-47,684	-2,928	0,078	-37,472
PT	-1,340	0,078	-17,150	-1,012	0,053	-19,184
ES	-1,558	0,094	-16,500	-0,878	0,070	-12,540
SE	-3,554	0,171	-20,833	-2,551	0,200	-12,782
UK	-3,566	0,117	-30,529	-2,248	0,112	-20,050

	Coefficient a			Coefficient b			Coefficient phi		
	Estimate	Standard error	t-test	Estimate	Standard error	t-test	Estimate	Standard error	t-test
AT	-2,452	0,097	-25,263	-2,100	0,103	-20,420	0,451	0,155	2,910
BE	-2,509	0,159	-15,772	-2,045	0,173	-11,850	0,539	0,188	2,862
CY	-1,581	0,109	-14,456	-0,865	0,167	-5,191	0,079	0,300	0,263
DK	-3,526	0,662	-5,324	-2,540	0,675	-3,765	0,688	0,150	4,594
FI	-1,871	1,036	-1,805	0,149	0,713	0,209	0,940	0,040	23,539
FR	-2,954	0,571	-5,177	-2,697	0,711	-3,795	0,748	0,185	4,035
DE	-2,567	1,269	-2,022	0,738	1,270	0,581	0,961	0,020	49,141
EL	-1,184	0,087	-13,568	-0,680	0,045	-14,979	0,431	0,205	2,097
HU	-1,977	0,508	-3,896	-1,110	0,304	-3,653	0,709	0,165	4,290
IE	-2,817	5,391	-0,522	-0,523	0,582	-0,898	0,980	0,066	14,977
IT	-2,360	0,126	-18,768	-1,521	0,149	-10,193	0,686	0,165	4,165
LU	-2,053	0,063	-32,483	-1,760	0,075	-23,359	-0,033	0,184	-0,178
NL	-4,219	0,134	-31,500	-2,963	0,122	-24,254	0,358	0,177	2,030
PT	-1,437	0,145	-9,879	-1,094	0,109	-10,010	0,548	0,158	3,472
ES	-1,948	0,687	-2,837	-1,216	0,760	-1,601	0,849	0,191	4,451
SE	-2,830	2,391	-1,184	0,638	1,019	0,626	0,968	0,035	27,998
UK	0,123	2,283	0,054	0,044	0,766	0,057	1,040	0,040	26,168

Table 3.2.12. Log-transformed model estimation results (top: base, bottom: autoregressive model)

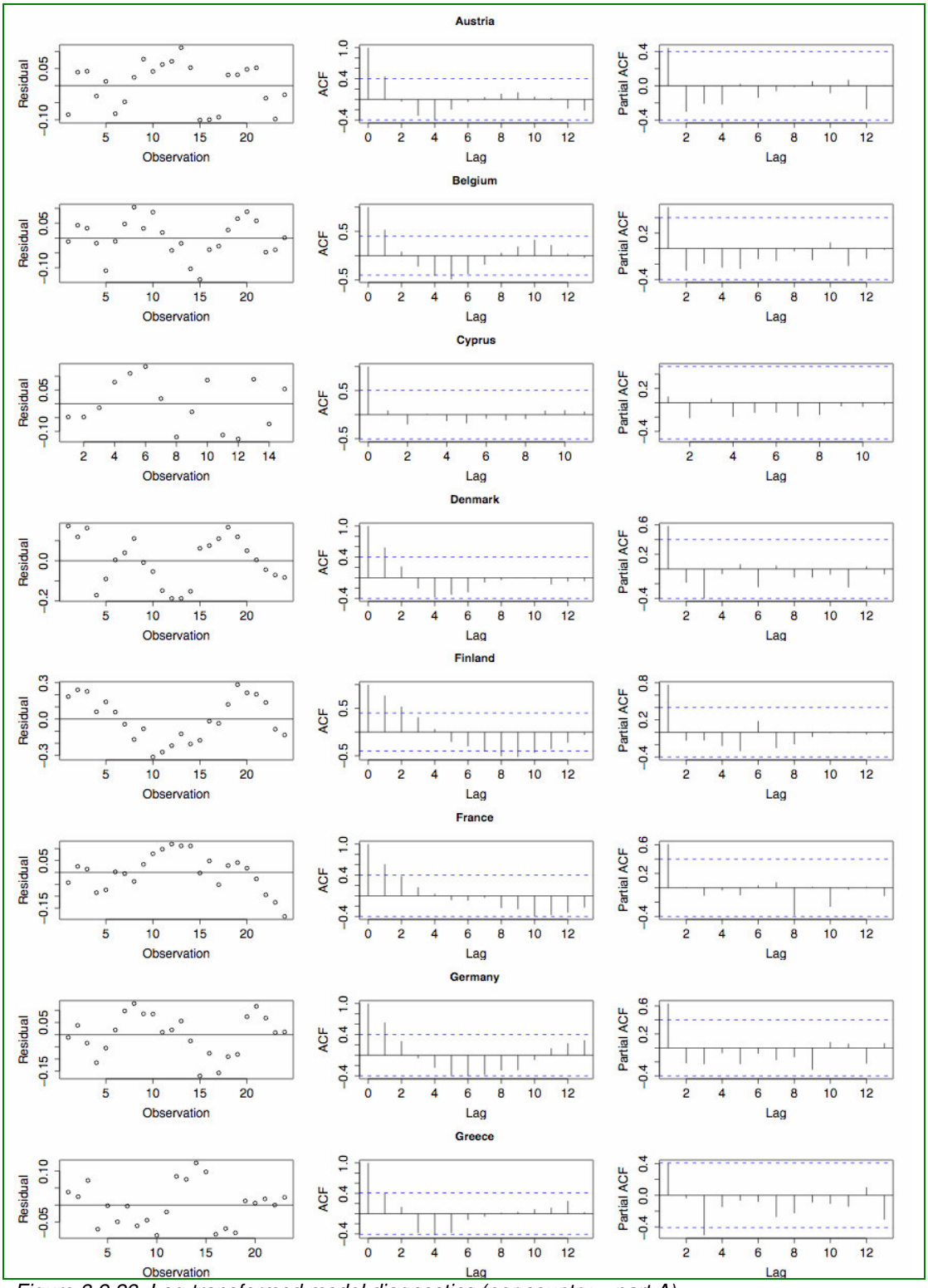


Figure 3.2.23. Log-transformed model diagnostics (per country – part A)

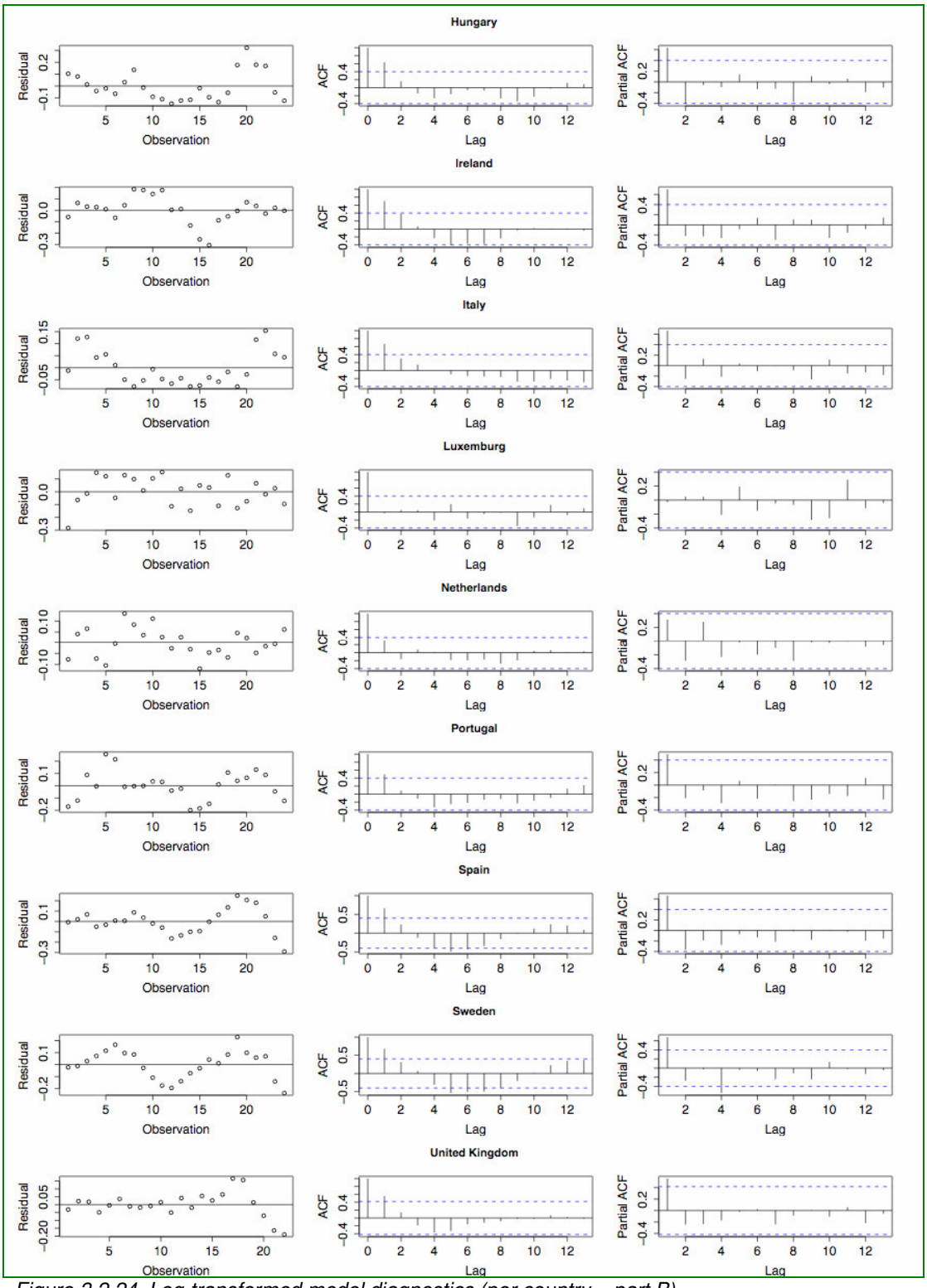


Figure 3.2.24. Log-transformed model diagnostics (per country – part B)

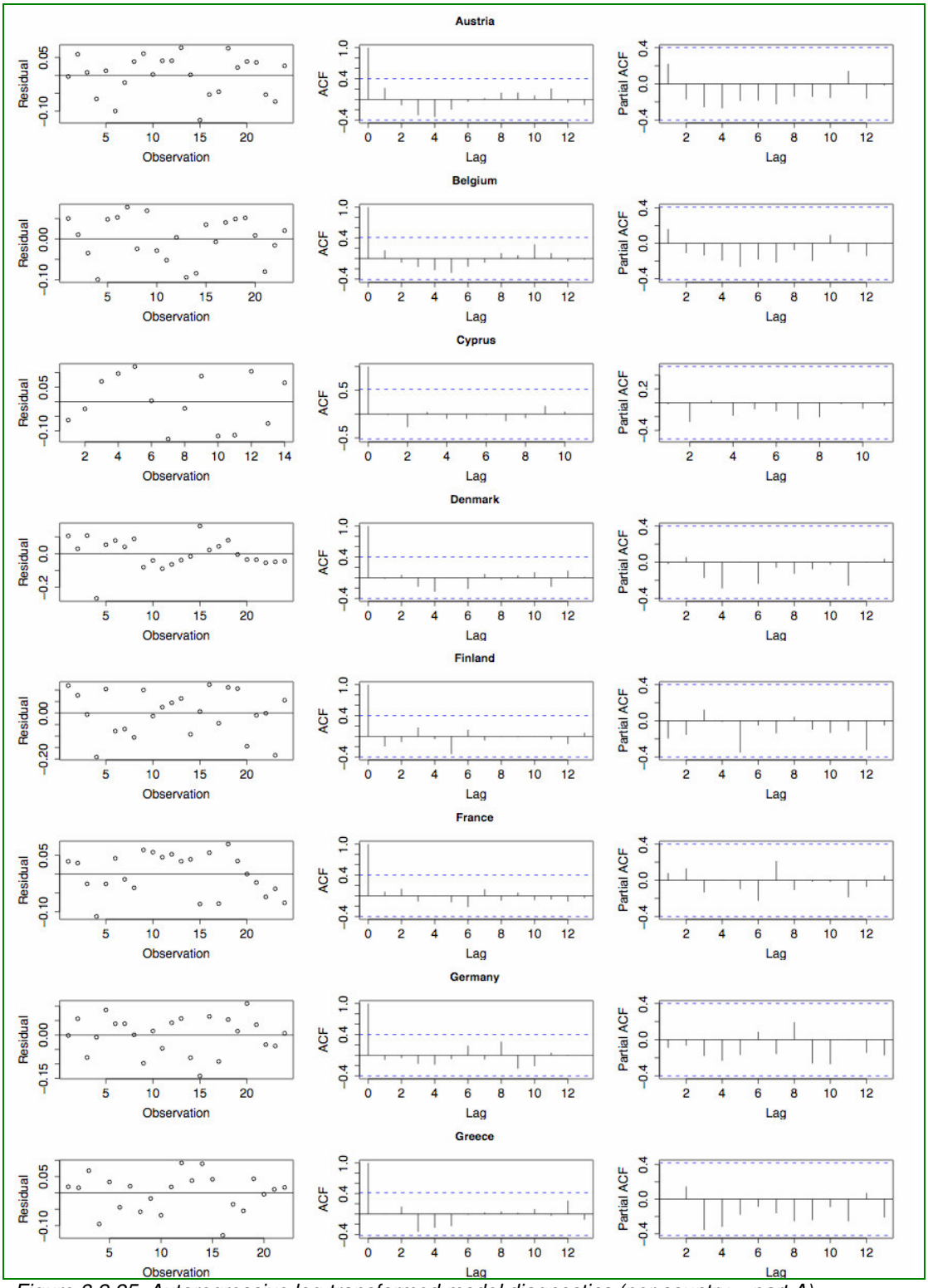


Figure 3.2.25. Autoregressive log-transformed model diagnostics (per country – part A)

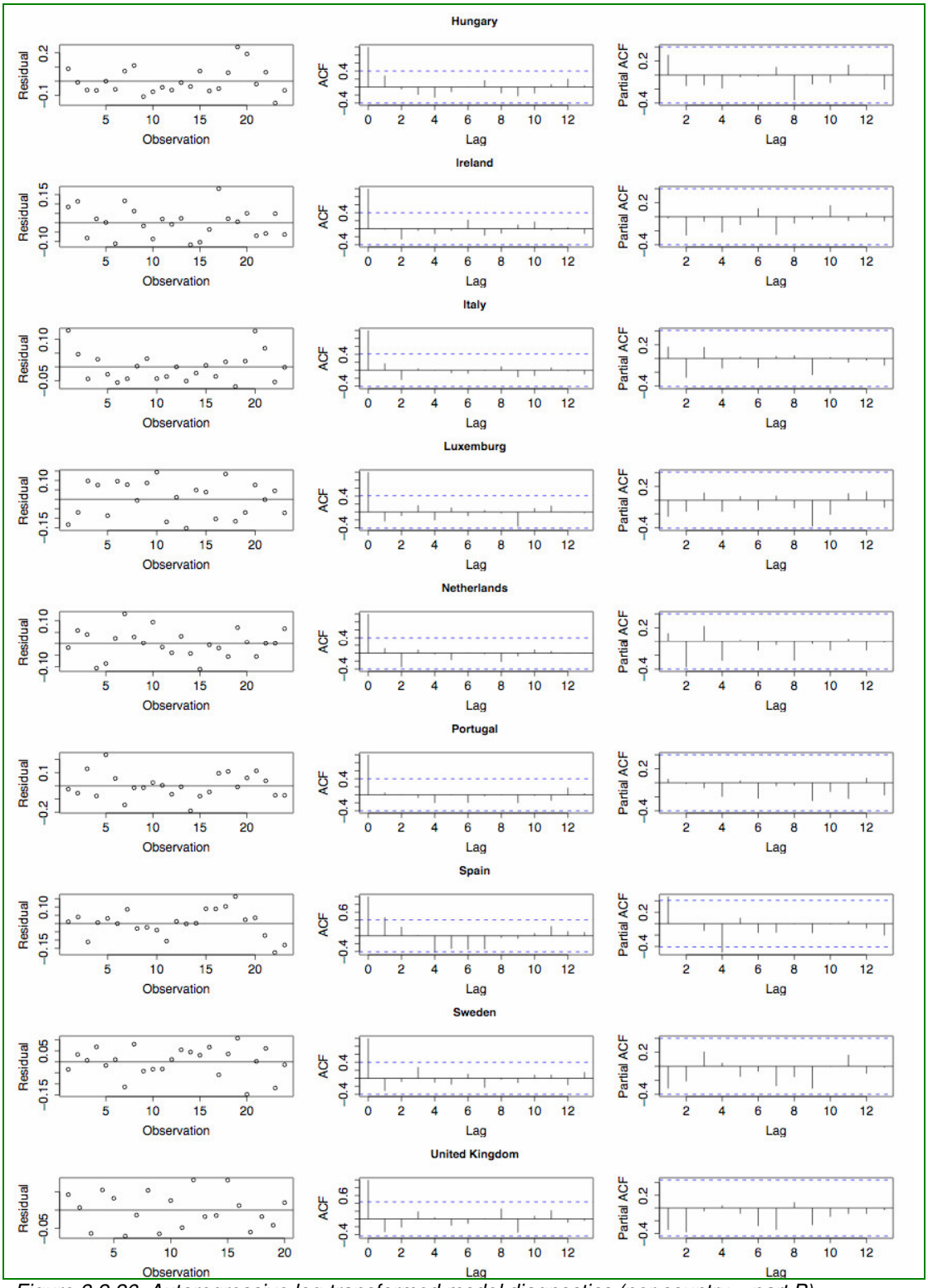


Figure 3.2.26. Autoregressive log-transformed model diagnostics (per country – part B)

Figure 3.2.23 and Figure 3.2.24 show the residual plots, ACF and PACF plots for the log-transformed models, while Figure 3.2.25 and Figure 3.2.26 show the

same statistics for the autoregressive log-transformed models. As with the non-linear model, the autocorrelation of the residuals has been significantly reduced in most cases due to the autoregressive process, and it has been practically eliminated from the ACF and PACF plots. However, while only one country (Sweden) seemed to face the issue with the high estimated parameter for φ, there are now four more countries with the same issue.

The autocorrelations for the various lags have been considered individually. A different way to test this type of lack-of-fit of a model is to consider the first e.g. 12 autocorrelations as a whole. It should be noted that this value depends on the data and is probably a bit high for this application. A lag of 4 or 5 might be sufficient, and using a lower lag might not illustrate the temporal dependency. Larger lags do not add to the inference, but are also rather harmless in this context. Denoting the first K autocorrelations as $r_k(\hat{a})$ (k=1,2,…K) Box and Pierce (1970) showed that if the fitted model is appropriate then

$$Q = n\sum_{k=1}^{K} r_k^2(\hat{a}) \tag{3.2.32}$$

is approximately distributed as $\chi^2(K-p-q)$ where n=N-d is the number of residuals used to fit the model. On the other hand, if the model is inappropriate, the average values of Q will be inflated. Therefore a so-called "**portmanteau**" test of the hypothesis of model adequacy can be obtained by comparing the value of Q against a standard χ^2 table. Ljung and Box (1978) argued that the chi-squared distribution does not provide an adequate approximation of the distribution of the Q-statistic under the null hypothesis, while Davies et al. (1977) provided empirical evidence to support this argument. Ljung and Box (1978) proposed a modified statistic (Ljung-Box-Pierce statistic):

$$\tilde{Q} = n(n+2)\sum_{k=1}^{K}(n-k)^{-1}\, r_k^2(\hat{a}) \tag{3.2.33}$$

A more detailed presentation of these tests is available in several texts, including Box et al. (1994), on which the development of this section is based. In the following application, Equation 3.2.33 is used.

Figure 3.2.27 visually presents the Ljung-Box-Pierce test results for the four groups of models. While the interpretation of the obtained p-values cannot be easily quantified, smaller p-values indicate lack of fit. Both the non-linear and the log-transformed models show mostly low p-values (and consequently a lack of fit). A threshold of 5% (indicated by a horizontal dashed line) exceeds several models' lines for the non-linear model and all-but-three for the log-transformed. The situation is substantially improved for the autoregressive models, with the p-values being considerably increased. Actually, only a couple of models fall below the 5% threshold for the non-linear AR model, and only one for the log-transformed AR model.

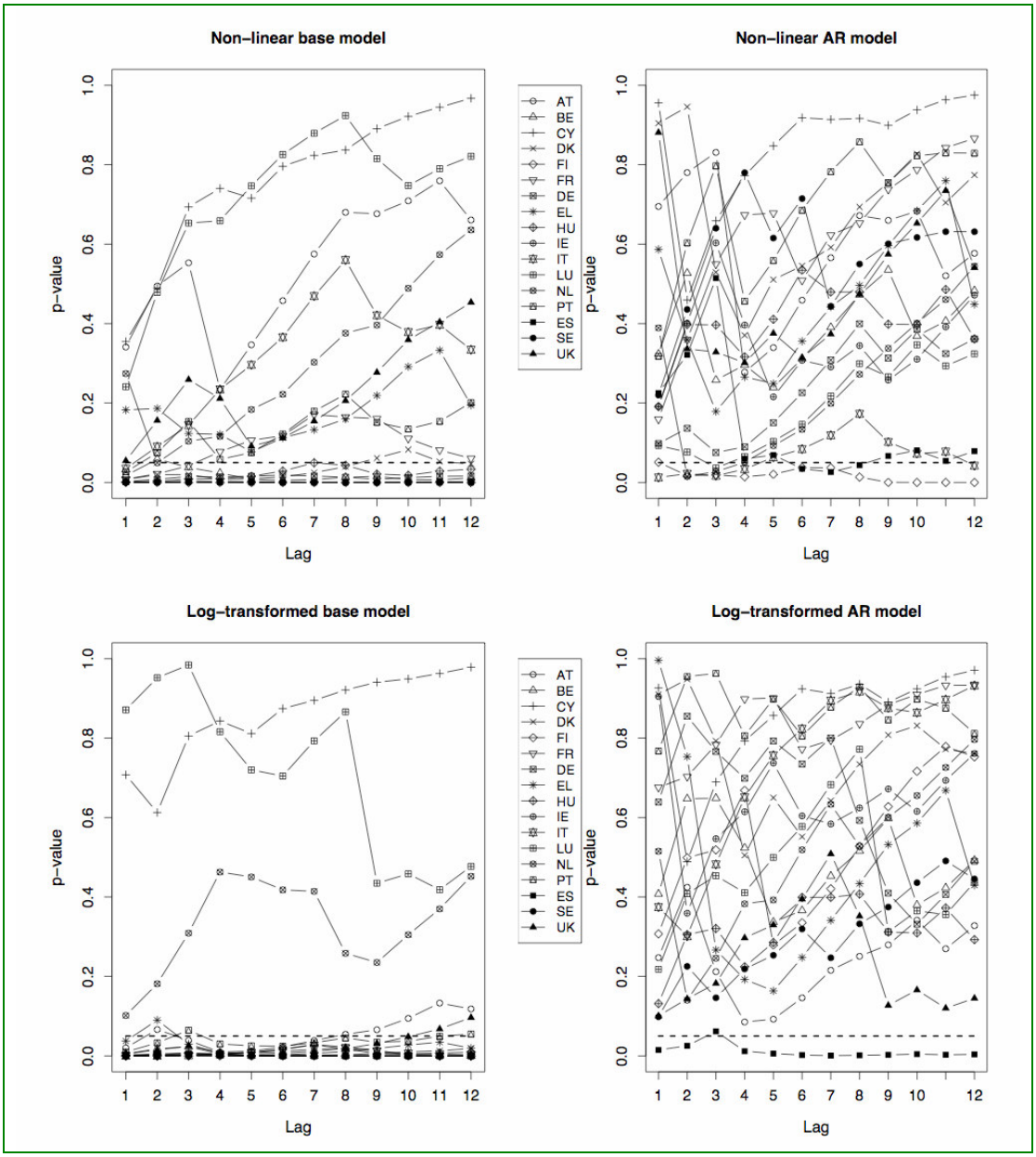


Figure 3.2.27. Ljung-Box-Pierce test p-values

Summary validation statistics using all four models are presented in Figure 3.2.28. As it has been mentioned before, this data is different from the data set that was used for the estimation of the models (i.e. years 1970-1994), in order to avoid issues such as over-fitting. In particular, while estimation used data from years 1970-1994, the validation used data from years 1995-2002.

The root mean square percent error (RMSPE) statistic (Pindyck and Rubinfeld, 1997) is used in Figure 3.2.28:

$$RMSPE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n^0 - x_n^1}{x_n^0}\right)^2} \tag{3.2.24}$$

where x is the variable of interest, N is the number of observations (years) and superscripts 0 and 1 denote observed and fitted measures respectively.

The impact of the autoregressive process in the prediction results is clear, with both autoregressive models consistently outperforming the base models. The non-linear AR model performs on average 39% better than the nonlinear model, while the autoregressive log-transformed model performs on average 49% better than the log-transformed model. This is a substantial improvement at the cost of just one extra parameter (the AR coefficient phi). Also, the AR log-transformed model also performs on average more than 13% better than the AR non-linear model.

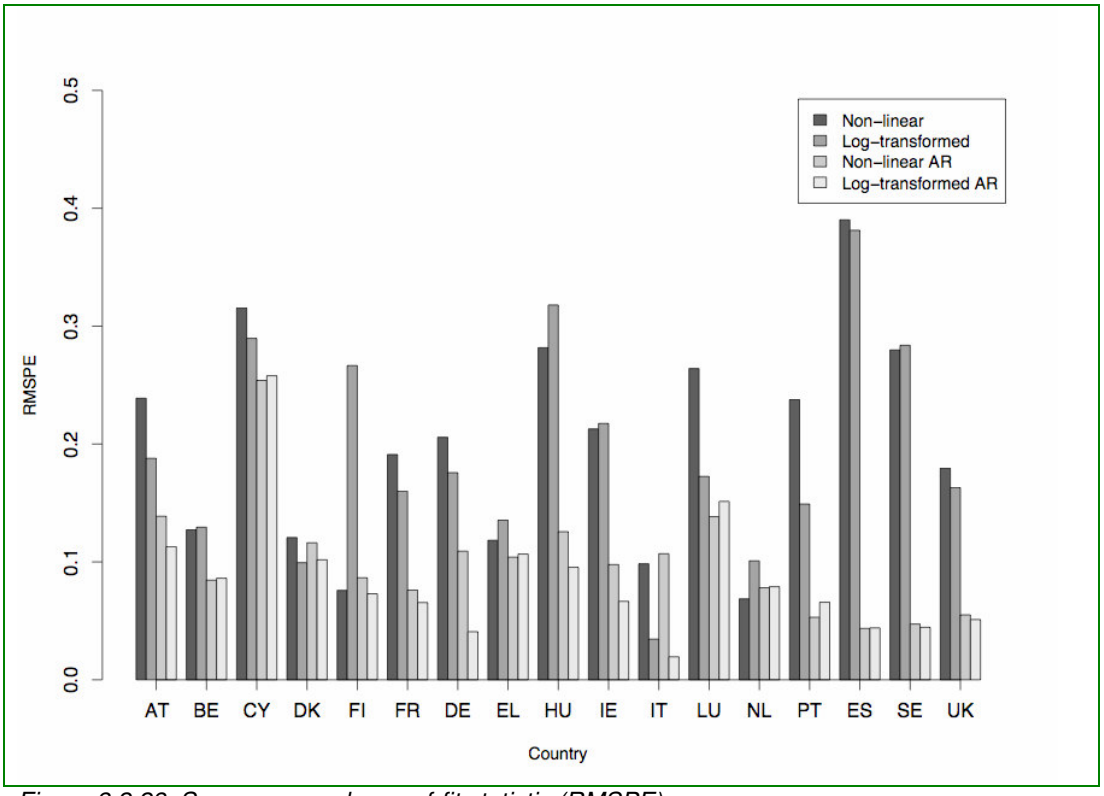


Figure 3.2.28. Summary goodness-of-fit statistic (RMSPE)

Figure 3.2.29 presents a plot of the estimated model parameters per country. While the log-transformed AR model seemed to provide a superior overall performance in terms of RMSPE, the non-linear AR model is used for this analysis, as its parameters are more reliable. In particular, significant overall improvement in the fit of the autoregressive models over the nonlinear models (39%) and over the log-transformed models (49%) has been obtained. On average, the AR log-transformed models outperform the AR non-linear models

by 13%. However, the estimated coefficients of the AR log-transformed model for five of the 17 countries are not reasonable, suggesting that this model should be applied with caution, as its estimation results may not be reliable.

The interpretation of parameter α is fairly straightforward, as it is a positive multiplicative parameter, and as such it can be considered as an indicator of the level of traffic safety in the country. Naturally, these parameters are not always directly comparable, as the value of the second parameter β also affects the total number of fatality rate. As the base of the exponent term is the car ownership rate, which is usually less than one, a larger negative value implies a higher overall term.

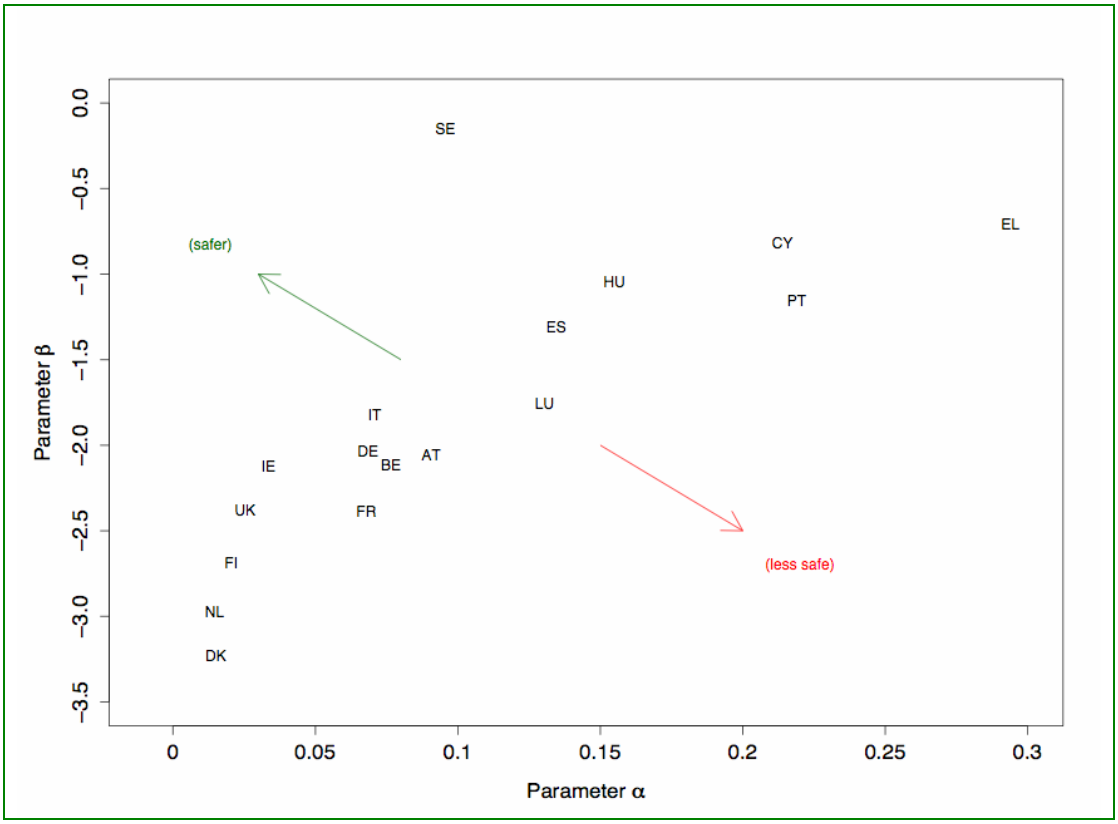


Figure 3.2.29. Interpretation of parameters (non-linear AR model)

Combining these two observations, safer countries should be to the left and top of Figure 3.2.29 and less safe countries should be in the right and bottom. No countries are located in the lower right triangle of the plot, which is a reflection of the fact that the considered countries are developed and have a decent level of road safety. It is expected that developing countries may be located closer to the lower right corner of the plot. The least safe countries in terms of safety in Europe today are Greece and Portugal, and indeed the respective points are located closer to the right and top of the plot. Similarly, the United Kingdom and the Netherlands (two of the safest countries in Europe) are closer to the left and

bottom. Therefore, this analysis (using a simple model and few explanatory variables) reflects the prevailing safety patterns, as evidenced by the literature on the subject.

3.2.3.5. Conclusion

The non-linear regression model has been presented, in relation to the linear regression model and the generalized linear model discussed above. Extensions of the non-linear regression model that correct for correlation of the disturbances have also been presented and applied to the estimation of fatality rate for 17 European countries.

A simple non-linear regression model has been fitted first, and the model diagnostics have been scrutinized to identify correlation of residuals and determine an appropriate line of action to correct for it. An autoregressive model has been selected and the approach to incorporate it into the non-linear regression model has been shown. The results of the autoregressive non-linear regression model have been presented. The model diagnostics demonstrate that the correlation of the disturbances has been effectively dealt with. An interesting finding is that Smeed's widely used relationship may produce serially correlated residuals, which –however- can easily be remedied by the presented auto-regressive models.

While a single global recommendation about a "best" model cannot be made based on the presented analysis, these results indicate that the autoregressive non-linear model generally outperforms the other models, while also overcoming the issue of serially correlated disturbances.

The ability to predict traffic safety trends is useful in setting and evaluating road-safety targets, policies and initiatives. The predictions obtained by the presented models can be used to evaluate the traffic safety performance of various countries, identifying poor performers, as well as traffic safety leaders. Given predictions of a country's expected performance, the actual traffic safety performance of that country over the past few years may be assessed. Furthermore, the study of more advanced (in terms of traffic safety and in general) countries may be applied to predict the evolution of less developed or successful (in terms of traffic safety) countries.

3.3 Dedicated time series analysis in road safety research

Ruth Bergel (INRETS)

We shall now focus on *Gaussian processes*, i.e.: processes having *a normal distribution.* The types of models described in this section are dedicated at handling time dependence, and will be discussed in more detail in the following Sections 3.4 to 3.6.

Time series models are defined in several specific manners, depending on the point of view adopted.

Technically speaking, a random process can be regarded as being made of a certain number of components. Whereas only the process can be observed - through a sample of observations - its components can only be estimated with a model's help. Thus, the *unobserved components models* are meant to provide estimates of each of these components, which it is not generally the case for a model meant to provide an estimate of the observed component only.
In the case the unobserved components are estimated, the main components (all components to the exception of the irregular component, which is stochastic by nature, see *3.3.1*) can be estimated while treated as being deterministic, or as being stochastic: thus, *deterministic vs stochastic components* will be considered.
At last, note that for different reasons the observed process may be pretransformed before being modelled: in the case the transformation is a filter of differences, filtered or *integrated components* will be considered.

Summarising these concepts, the following types of Gaussian times series models can be given:

- models with deterministic main components (decomposition model with deterministic trend/seasonality for instance.)
- models with stochastic main components (decomposition model with stochastic trend/seasonality for instance.)
- models with integrated components (integrated model).

The basic structure of these models can be enriched in different ways, in order to bring additional information, related to the past of the observed process, or related to other processes. The reference to the past of the process is performed by introducing *autoregressive/moving average* parts, whereas the reference to the environment is performed by introducing other (explanatory) variables. Finally, the form itself of the model, which generally is linear with respect to the the parameters and components, might also be enlarged by introducing non-linearity.

However, the systemic-approach of road safety adopted by researchers since the beginning of the 1980's aims at taking account of all explanatory risk factors and at assessing road safety measures (Hakim and al., 1980), and at focusing

on explanatory time series models.. As this research direction still holds and has been enforced, the main distinction retained in section 3.3 is the distinction between mere *descriptive models* and *explanatory models* (the second group being an extension of the first one, as it will be demonstrated below). Within each of these two groups, the basic model structures defined before hold, but the focus in this chapter will be on decomposition models - whether with deterministic or stochastic main components - , on the one hand, and on ARMA and ARIMA models on the other hand.

The plan retained in Section 3.3 is as follows. The main distinction between descriptive and explanatory models is first introduced in a general manner, without reference to the road safety field, in Section 3.3.1; the methodological framework for understanding time series analysis in the road safety context is then recalled in Section 3.3.2, and a brief overview of time series analysis, as performed since the beginning of the 1980's in road safety research, is given in Section 3.3.3. In the concluding Section 3.3.4, the models main features are summarised.

3.3.1 Types of models

Two main kinds of models are usually distinguished, when one aims to formulate the evolution over time of a stochastic[43] process (Y_t), for t being 1,2,3,…, having a number of observed values of the process - a sample of observations *:* $Y = (y_1, y_2, ..., y_n)$ - at hand. The ***descriptive models*** on the one hand - they will be defined here as models for which the only exogenous variable used is time, which is thus not considered as an explanatory variable - , and the ***explanatory models*** on the other hand - models which do on the opposite use exogenous, also called independent or explanatory[44] , variables.
The different types of models, whether descriptive or explanatory, will be recalled and summarized in Table 3.3.1.

3.3.1.1. Descriptive models

Descriptive models take account for the trend/seasonal/irregular decomposition of the process Y_t. Here again, two main kinds of models are considered : *decomposition models* on the one hand*,* which adjust for each of the components by explicitly modelling it, often modelled using *state space* methodology, and *ARMA and ARIMA models* on the other hand, which adjust for the irregular component, after the process has (if necessary) been filtered in such a way as to remove its trend and seasonal components.

3.3.1.1.1. Decomposition models

Descriptive decomposition models can be written, in the simple case of an additive decomposition, as

---

[43] The process (Y_t) is stochastic, or random, in the sense that the values taken by Y_t are under measurement errors.

[44] "Explanatory" as being used in an explanatory model
"Explanatory variable" is used here in the general acception, and no difference is to be made yet between pure explanatory and intervention variables (as it will be made in Sections 3.4 to 3.6)

$$Y_t = T_t + S_t + u_t \qquad (3.\ 3.1)$$

with: T_t the trend of the process Y_t,
S_t the seasonal (periodic) component[45],
and u_t the random component (also called irregular component), assumed to be stationary[46].

These *unobserved components*, emerge very naturally: the long term tendency T_t, the seasonal component S_t and the random residual component u_t.

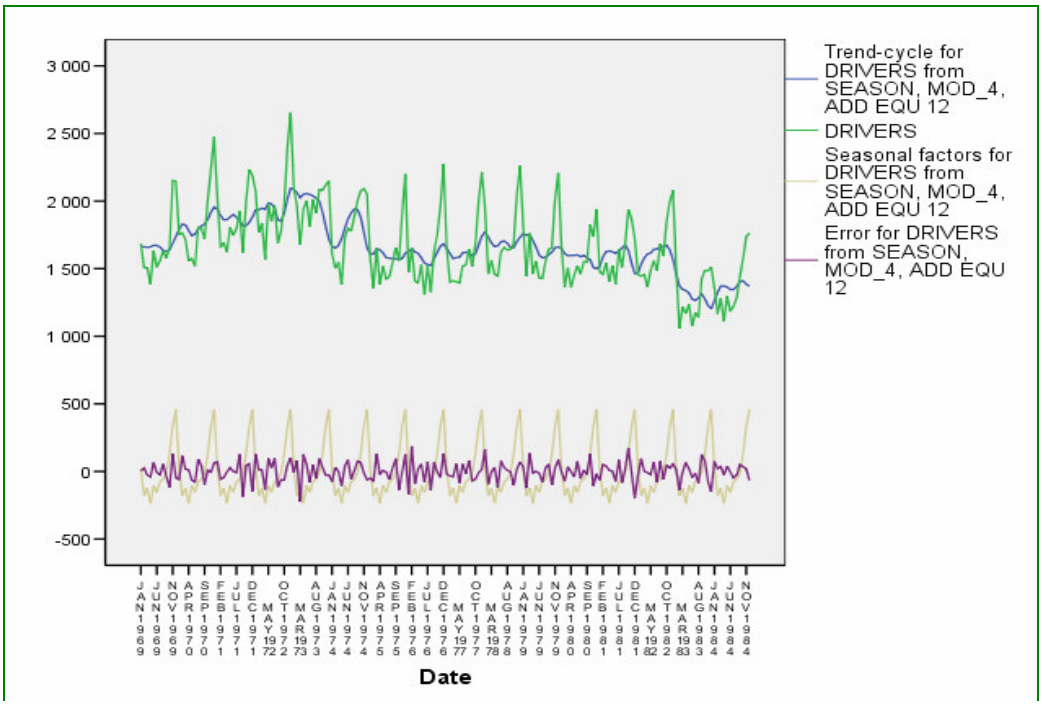


Figure 3.3.1: The monthly number of UK-KSI drivers, for the period January 1969 - December 1984 - original data and unobserved components. The sample process (green line), the trend (blue line), the seasonal component (grey line) and the irregular component (violet line).

Figures 3.3.1 and 3.3.2 describe the development of the unobserved components, as obtained with SPSS, of the two seasonal data sets modelled in the following sections: the monthly number of drivers killed and seriously injured in the UK (UK-KSI drivers), for the period January 1969 - December 1984, and the monthly number of fatalities in France, for 1975-2001.

[45] the sum of the seasonal component terms within a season (period) is also zero
[46] its mean, variance and covariance structure are constant over time (see a precise definition in 3.4.2.2), and moreover its mean is zero.

The data (green line) are the sum of the trend (blue line), of the seasonal component (grey line) and of the irregular (violet line).
The data corrected for the seasonal component are the so-called seasonal corrected, or seasonal adjusted data.
Once they are also corrected for the irregular component, only the trend remains..

Of the three unobserved components, interest goes first to the trend. The trend is often thought as a function of certain variables, which determine it, although these variables can not always be quantified easily. In such cases, the trend is modelled with a deterministic form, and is qualified as deterministic - the same approach being retained for the seasonal component.

But the trend can also be considered as a random walk (Harvey, 1989). The same remarks apply to the seasonal component. The structural modelling proposed by Harvey is another form of the decomposition previously described, in this case, the trend and the seasonal component may also be random. In such cases, the trend and the seasonal are, as the irregular part of the model, subject to random fluctuations. This approach is taken in most of the state space models presented in this document (see Section 3.6).

At last, it should be mentioned that a more general model can be retained for the decomposition of the process as a function of its unobserved components: If it is not additive, as in (3.3.1), the decomposition form can be multiplicative or semi-multiplicative accordingly to the Census decomposition method's available options (Dagum, 1980).

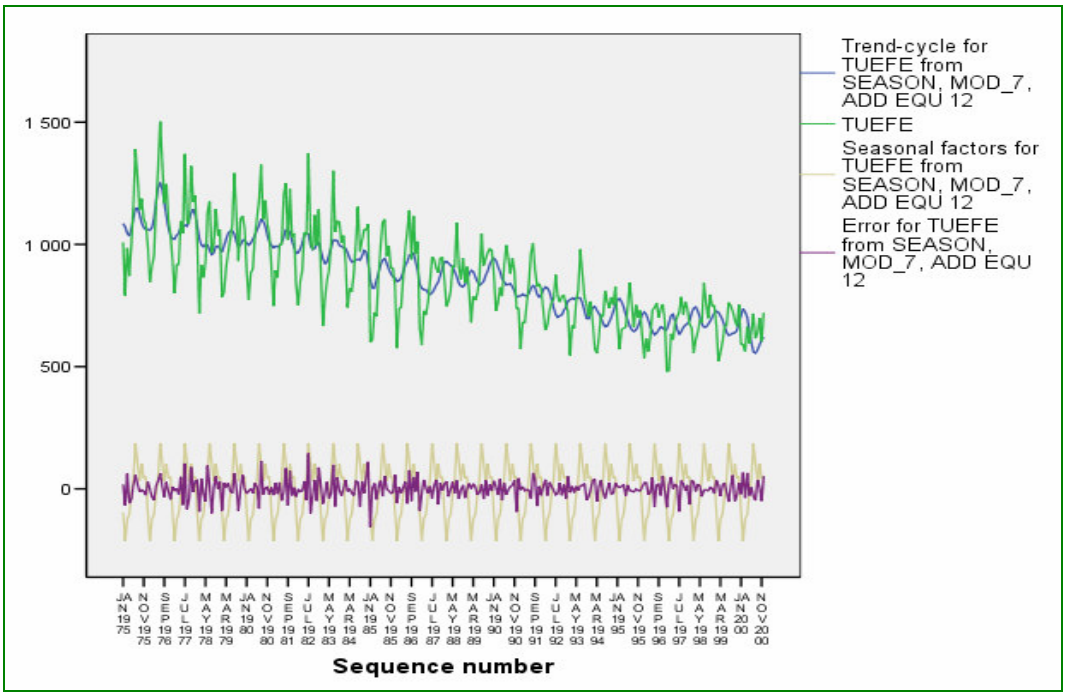


Figure 3.3.2: The monthly number of fatalities in France, for the period 1975-2001. For explanation see Figure 3.3.1.

3.3.1.1.2. ARMA and ARIMA models

The descriptive autoregressive and moving average (ARMA) models focus on describing the *dynamics* (the relationship between its values at different time points) of the *stationary* sample process $Y = (y_1, y_2, ..., y_n)$. This relevant property of stationarity allows separating Y_t in two parts: the one related to the past at time t, and the part that is new at time t - which is therefore called the « innovation » - .in such a way that this later component is a white noise[47], and is therefore called the innovation white noise.

Thus, the value taken by the process at time t: Y_t, can be expressed as a function, and more precisely as a linear combination of its passed values Y_{t-1}, Y_{t-2}, ... , and of the innovation white noise u_t. For parsimony reasons, as different equivalent formulations can all be retained for describing the process dynamics, the formulation currently chosen[48] is that Y_t is expressed as a linear combination of a small number (p) of its past values, and of a small number (q) of the past values of the disturbances.

This can be written the following way:

$$Y_t = \phi_1 y_{t-1} + ... + \phi_p Y_{t-p} + u_t + \theta_1 u_{t-1} + .. + \theta_q u_{t-q}, \qquad (3.\ 3.2)$$

with: ϕ_1, ϕ_2, .., ϕ_p, θ_1, θ_2,.., θ_q p+q real values,
and u_t the innovation disturbance,

The fact of knowing the dynamics of the process enables to extrapolate it without any call to additional variables, assuming that the dynamic's structure will stay unchanged in the future, at least at the forecast's horizon (hence we need to assume the process is stationary). The reference to the near past makes the model adaptive.

In the general case where stationarity cannot be assumed, it is convenient to assume that another stationary process exists, which is derived from Y_t by removing its trend and its seasonal component. An easy manner for doing this, as recommended by Box and Jenkins in 1976, is to apply a so-called filter of differences[49] to the process Y_t, as many times as necessary until the result, the filtered process, can be considered as fulfilling the property of stationarity, and therefore be fitted with an ARMA model itself. This comes to removing the trend

[47] see a precise definition in 3.4.2.2

[48] see a precise definition in 3.4.2.2 or in (Box,Jenkins, 1976)

[49] *see* a definition of the difference filter as a function of the backshift operator B in 3.4.2.2

and seasonality from a non stationary process, in other words to solving the first order non stationarity[50]. The fact that the filtered or *integrated process* obtained by applying an appropriate filter of differences to Y_t is fitted with an ARMA is equivalent to say that Y_t is fitted with an ARIMA (or integrated ARMA).

The second order stationarity can also be obtained by deriving another process from the initial one. The logarithmic transformation is therefore currently[51] applied to the initial data in order to stabilizing their variance.

3.3.1.2. Explanatory models

3.3.1.2.1. Explanatory variables

In this subsection, *exogenous* - also called *independent* or *explanatory* - variables will be considered. Note that other terms, such as “predictor” or “regressor” are also commonly used, when a specific role is expected from them.

Several data sets of different nature will now be considered, and used within one and the same model:

- the observations of the endogenous stochastic process, i.e. the sample of data $Y = (y_1, y_2, ..., y_n)$
- the values taken by the k exogenous variables Z_{it}, i=1 to k, assumed to be known.

It is natural to distinguish several kinds of exogenous variables, depending on whether they affect the trend, the seasonal component, or the irregular component of the process Y_t.Moreover, effects of exogenous variables can be local - over time (the effect may be 'short-lived') - , or permanent. It seems quite natural, again, to distinguish *the dummy variables*, which are created (outside the model) as witnesses of a local, isolated or repeated, effect usually having values zero or one, *and the variables of measure* of a phenomenon (of which the value is actually measured), assumed to be linked with the process Y_t, and which may have a permanent effect. As an example, climate and calendar variables can be used for modelling the seasonal component, or the residual; the variables used to model the trend are of a different nature, insofar as one can expect their effect to extend over time.

[50] There are different ways for removing the trend a non-stationary process: the trend itself being modelled, as a function of time for instance.

[51] Among other transformations (see section 3.5). In case an independent variable is added in the model, the form of the relationship between the dependent and the independent should be fixed accordingly to the knowledge of existing additive or multiplicative effects.

The explanatory time series models take into account the relationship between the endogenous variable Y_t and exogenous or explanatory variables - gathered in a vector of the k exogenous variables $Z_t = (Z_{1t}, Z_{2t}, ..., Z_{kt})'$

For instance, in the case in which there are two explanatory variables, Y_t will be expressed as a linear combination of Z_{1t} and Z_{2t} . The residual $Y_t - \beta_1 Z_{1t} - \beta_2 Z_{2t}$ of the regression of Y_t on Z_{1t} and Z_{2t} will be noted YC_t, and modelled as described in the previous subsection.

For the commodity of the coming formulation, the function g will be used in the general case where there are more than 2 exogenous variables, for representing the relationship between Y_t and $Z_t = (Z_{1t}, Z_{2t}, ..., Z_{kt})'$, and l=1 to k.

Explanatory models can be seen as descriptive models to which exogenous variables have been added, and thus can also be classified as either decomposition models with explanatory variables, or ARIMA models with explanatory variables.

We shall now address these two kinds of models.

3.3.1.2.2. Decomposition models with explanatory variables

The decomposition models with explanatory variables can generally be written, in the case of an additive decomposition, as

$$YC_t = Y_t - g(Z_t) = T_t + S_t + u_t \qquad (3.3.3)$$

with: YC_t the process corrected for the exogenous effects,
T_t , S_t and u_t the trend, the seasonal component and the random component of YC_t.

A basic example is the regression model, of the dependent variable - or endogenous variable - on explanatory variables - or exogenous variables, described in Section 3.2.1. The exogenous variables can account for the trend, for the seasonal component, or for the residual. For instance, in the case of periodic data, the regression model will contain dummy variables in order to model the season (the day, the month, the quarter month),

Harvey's structural model with explanatory - and intervention - variables is another kind of stochastic decomposition model more general than the basic structural model, mentioned before.

3.3.1.2.3. ARMA and ARIMA models with explanatory variables

The ARMA and ARIMA models with explanatory variables can generally be written as

$$YC_t = \phi_1 YC_{t-1} + ... + \phi_p YC_{t-p} + u_t + \theta_1 u_{t-1} + .. + \theta_q u_{t-q}$$
$$YC_t = Y_t - g(Z_t) \qquad (3.3.4)$$

with: YC_t the process corrected for the explanatory variables,
and u_t the innovation disturbance of the process YC_t .

In that general specification, Y_t and $Z_t = (Z_{1t}, Z_{2t}, ..., Z_{kt})'$ $_{l=1 \text{ to } k}$. may have been pretransformed (filter of differences, logarithmic transformation, ..) in such a way that YC_t can be assumed to be stationary.

ARMA or ARIMA models with explanatory variables can also be seen as regression models with ARMA or ARIMA residuals, the two formulations being equivalent. It is relevant to determine whether the exogenous variables do have an effect on Y or on the variations of Y, after the trend and the seasonal components have been filtered out.

Descriptive models	*Explanatory models*
Decomposition models $Y_t = f(T_t, S_t, u_t)$	*Decomposition models with explanatory variables* $YC_t = Y_t - g(Z_t) = f(T_t, S_t, u_t)$
Autoregressive models $Y_t = f(Y_{t-1}, Y_{t-2,}.., u_t)$	*Autoregressive models with explanatory variables* $YC_t = Y_t - g(Z_t) = f(YC_{t-1}, YC_{t-2}, .., u_t)$
Autoregressive and moving average models $Y_t = f(Y_{t-1}, Y_{t-2,}.., u_t, u_{t-1,}..)$	*Autoregressive and moving average models with explanatory variables* $YC_t = Y_t - g(Z_t) = f(YC_{t-1}, YC_{t-2}, .., u_t, u_{t-1,}..)$
and, as extensions: *Integrated autoregressive and moving average models (ARIMA models)*	*Integrated autoregressive and moving average models with explanatory variables (ARIMAX models)*

Table 3.3.1: Types of models.

3.3.2 The methodological framework

In this section, we recall the methodological framework which enables us to quantify the influence of the different factors related to the transport system, to mobility, and to road safety's economy on road risk (Lassarre, 1994).
We address *aggregated time series* - on an annual, monthly or daily basis. The dependent variables are in all cases aggregated at a territory's or at a network's level, or aggregated according to a typology of injury accidents or victims.

3.3.2.1. The diagram of production of the risk

Risk analysis is based on the ***exposure/accident/victim*** triad.

We have to distinguish between:

- Two types of road risk : the accident's risk, and the risk of being a victim (killed, seriously injured, lightly injured) of an accident,
- And three levels of risk: risk exposure, accident's risk, and accident's gravity.

Risk indicators and risk factors are defined at the three levels of this framework.

3.3.2.2. Risk indicators

The usual, but not always available, measure of risk exposure is an indicator which measures the traffic volume: the mileage, measured in ***number of vehicle kilometres*** driven on a road network.

The accident rate (number of injury accidents in a billion of vehicle kilometres) is usually retained to measure the accident's risk on a network; but, in order to overcome the hypothesis that the number of accidents would be proportional to the traffic volume, ***an absolute number of accidents*** is also retained, but is then mostly considered as being a non-linear function of mileage[52].

Finally, the indicators that measure accident's gravity are like ***the fatality rate***, i.e. the number of victims (fatalities, seriously injured, slightly injured) by accident; one may prefer to measure directly ***the absolute number of victims***, but it will then be considered as depending on the number of accidents, or directly on the traffic.

It may be noted, at that stage, that the absolute numbers of accidents and victims are also considered as accident's risk and accident's gravity indicators.

3.3.2.3. Risk factors

Risk factors are classified either as ***internal (to the transport system) factors***, related to the vehicle, to the driver and to infrastructure; or as ***external factors***, representing the environment, and related to the climatic, economic, demographic and legislative systems (Gaudry, Lassarre, 2000).

---

[52] *The same remark applies to the risk of being killed (or fatality rate, i.e. the number of fatalities in a billion of vehicle-kilometres).*

3.3.2.4. Towards an explanatory approach

Since the beginning of the 1980's, time series analysis in the road safety field is directed at taking into account all explanatory factors of accidents frequency and gravity, and at assessing road safety measures (Hakim and al., 1990). *Descriptive* models have been followed by *explanatory* models - models with explanatory variables - , built on the basis of a rich economic formulation, with an elaborate econometric specification.

By examining the numerous models proposed for aggregate accident data of European countries, it appears that they differ on the necessity of taking into account an important number of explanatory factors, and on the nature of the models that should preferably be used. The examples given now illustrate the different approaches.

3.3.3 Applications in road safety research

Details about the examples of road safety analysis given in this section can be found in (Cost 329, 2004), and additional references can be found at the end of the Methodology Report

3.3.3.1. Deterministic versus stochastic

The purely descriptive models (without any explanatory variable, except for time) have mainly been used to model a road safety indicator: ***the fatality rate***. The objective of these decomposition models was to adjust the trend as a function of time. The trend/residual decomposition retained on an annual basis is extended to a trend/seasonal/residual decomposition on a monthly basis. The trend, and the seasonal component as well, is deterministic or stochastic.

Thus, on annual data, an example of a deterministic model is provided by Oppe (1993), who proposes an exponential decreasing trend for the fatality rate R_t (the number of fatalities per billion of vehicle-kilometre):

$$R_t = \exp(at + b)$$

$$\text{with:} \quad R_t = \frac{F_t}{V_t},$$

F_t the number of fatalities, *(3.3.5)*

and V_t the traffic volume.

This form proposed for the trend of the fatality rate R_t has been enlarged afterwards, and a transformation on the traffic variable was retained, to account for the non- proportionality of the number of fatalities to the traffic volume, the additional parameter η representing the elasticity of the number of fatalities with respect to traffic:

$$\frac{F_t}{V_t^{\eta}} = \exp(\mu_t) \qquad (3.3.6)$$

In both previous cases, the trend of the fatality rate was modelled in a deterministic manner, as a function of time. The simplest function being the linear function $\mu_t = at + b$ as in the first case. But the trend can also be random itself, in which case a specific error term is added to the model, for taking account of its randomness. More precisely, there are as many additional error terms as there are random components in the model.

A stochastic form has been proposed by Lassarre (1997) for the temporal function μ_t , which becomes locally linear, that is to say by supplementing the basic structural model formulation:

$$\begin{aligned} LogF_t &= \eta LogV_t + \mu_t + \varepsilon_t \\ \mu_t &= \mu_{t\ 1} + \beta_{t\text{-}1} + \eta_t \\ \beta_t &= \beta_{t\text{-}1} + \xi_t \end{aligned} \qquad (3.3.7)$$

with β the slope of the trend μ,

ε, η, ζ white noises of variances σ_ε^2, σ_η^2 and σ_ζ^2, mutually non-correlated.

In the case of monthly data, a seasonal component is added, which can also be deterministic or stochastic. In fact, due to the larger number of data available on a monthly basis, additional parameters can be estimated - i.e. additional exogenous variables can be used - as this will now be discussed.

As has just been seen, a descriptive model of the fatality rate may be considered as an explanatory model of the absolute number of fatalities, with as single explanatory variable: the traffic volume. This kind of explanatory model with a single exogenous variable has been enriched with additional variables, more or less numerous. In fact, the real explanatory models take account of a larger number of risk factors. Examples of such models will now be described.

It must be noted that the same formulation proposed for modelling the number of fatalities can also be used for modelling the number of accidents, as a function of the traffic volume and of additional variables.

3.3.3.2. Regression versus ARIMA

As an example of a decomposition model with a deterministic trend and with explanatory variables, we shall mention Scott (1986) who uses an ARIMA structure for modelling the monthly number of accidents in the United Kingdom from 1970 to 1978, after having first regressed the data on exogenous variables measuring the traffic volume, the petrol price, temperature, rainfall height and the number of working days (in fact a regression with an ARIMA residual); he then demonstrates that the ARIMA structure on the residuals of the regression can be omitted, subject modelling the trend and the seasonal component with the help of a time variable and of seasonal dummies, in the regression equation.:

$$\log Y_t = a + bt + S_t + \sum \beta_i LogX_{it} + \sum_j \beta_j X_{jt} + \lambda\omega_{1t} + \lambda_2\omega_{2t} + u_t \qquad (3.3.8)$$

with: Y_t the monthly number of accidents in the UK,

$a + bt$ the trend,

S_t the seasonal, modelled with 11 dummy variables,

$X_i, i = 1,2$: the traffic volume for two kinds of vehicles and the petrol price,

$X_j, j = 1,2,3$: the two climate variables and the number of working days,

ω_{1t} and ω_{21t} two dummies indicating the oil crisis of 1974 and the introduction of speed limitation in rural areas.

3.3.3.3. State space models

Among the different types of state space models, Harvey's structural model with explanatory - and intervention - variables (1986) is a type of stochastic decomposition model more general than the basic structural model. Used on the number of drivers killed and seriously injured (KSI) in the UK, it included two explanatory variables x_{it} (the petrol price and the number of travel kilometres) which have an effect on the trend of y_t, as well as the dummy variable $\omega_t=1_{t\geq\tau}$ which is used to assess the effect $\lambda\omega_t$ of the seat belt law.

$$\begin{cases} LogY_t = \mu_t + \gamma_t + \sum_{i=1}^{I} \beta_i Logx_{it} + \lambda\omega_t + \varepsilon_t \\ \mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t = \beta_{t-1} + \zeta_t \\ \gamma_t = \sum_{j=1}^{s/2} \gamma_{jt} \\ \gamma_{jt} = (\cos\frac{2\pi j}{s})\gamma_{j,t-1} + \omega_{jt} \end{cases} \qquad (3.3.9)$$

with : Y_t the monthly number of drivers KSI in the UK,

ε, η, ζ et ω_{jt} white noises of variances σ_ε^2 , σ_η^2, σ_ζ^2 and σ_ω^2, mutually uncorrelated.

In an equivalent way but on annual data, the largest formulation proposed by Lassarre (2001) for the local linear trend model incorporates intervention dummy variables ω_{it}, ω_{jt} and ω_{kt} , which may modify the irregular component, the level or the slope of the trend of the number of fatalities :

$$\begin{aligned} LogF_t &= \eta LogV_t + \mu_t + \sum_i \lambda_i \omega_{it} + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \sum_j \lambda_j \omega_{jt} + \eta_t \\ \beta_t &= \beta_{t-1} + \sum_k \lambda_k \omega_{kt} + \xi_t \end{aligned} \qquad (3.3.10)$$

Applied to aggregate data of several European countries, this formulation allowed to assess the effect of the main road safety measures. For France, the main measures taken in 1973 - the speed limitation and the seat belt wearing obligation - caused a significant drop of 17% from 1973 onwards, in the fatality rate. A drop of 9,3% in 1978 is caused by the introduction of random alcohol tests on the road.

3.3.3.4. ARIMA models

One type of ARIMA model with explanatory variables is very often used on monthly data in the road safety field, in order to assess the effect of road safety measures. These models, fitted on monthly aggregate numbers of injury accidents and victims, generally take into account recognised exogenous effects - such as the effect of risk exposure, the climate influence with the help of one or two meteorological variables, and the calendar configuration influence - and the effect of specific road safety measures.

As examples we shall mention the models proposed for aggregate data in Spain and France.

In Spain, two variables of oil sales (gasoline and diesel) as a proxy for traffic, the number of week-end days in the month WEND and another intervention variable taking account for a great number of road safety measures gradually enforced from June 1992 off $LS^{6/92}$, were used for modelling the number of injury accidents Y_t from January 1982 to December 1996 (Rebollo, Rivelott, Inglada Lopez de Sabando, in COST 329, 2004):

$$\begin{aligned} \log Y_t &= \sum_i \beta_i LogX_{it} + \eta WEND + \gamma LS_t^{6/92} + N_t \\ \nabla\nabla_{12} N_t &= (1-\theta_1 B)(1-\theta_{12} B^{12})\varepsilon_t \end{aligned} \qquad (3.3.11)$$

The same econometric specification was used for modelling the aggregate numbers of injury accidents and fatalities in France. The models account for the mileage and the speed, but they mainly allow for assessing the safety measures enforced during the period. It's the case of the first speed limitation of 1973, of the oil crisis of 1974, of the legislation of 1978 introducing random alcohol tests on the road (Lassarre, Tan, 1981, 1982, 1989).

Other models of the same type were also proposed for modelling the number of injury accidents and fatalities on the main network categories in France: A-level roads and motorways, secondary roads and urban roads, with the help of climate and calendar variables for taking account of transitory factors as well (Bergel, Depire, 2004).

3.3.3.5. Non linear models

As can be seen, non-linear models have often been transformed into linear models, by applying a log-transformation to some of the variables, whether dependent or independent; this renders the model estimation easier. Other examples of dealing with non-linearity have been given in Section 3.2.3.

The multiplicative relationship between exposure and casualties, and between exposure and fatalities, is generally accepted. It is worth recalling here, as an example, that the first aggregate model at a country's level, proposed by Smeed (1949), relates the number of road injuries to the number of motorised vehicles and to the corresponding population (i.e. D, M and P respectively) in a multiplicative manner :

$$D = c(MP^2)^{\frac{1}{3}} \qquad (3.3.12)$$

Other transformations may also be chosen, preferably to the Log-transformation, and applied to the observed data. Let's mention the three-level explanatory model constructed on a monthly basis, the DRAG-model (Demand for Road use, Accidents and their Gravity) proposed by Gaudry(1984), which relies on a multiple regression structure with auto correlated and heteroscedastic errors, and takes account for a type of non-linearity. The fact that numerous explanatory variables are introduced allows the trend and the seasonal component to be modelled, which thus do not need to be filtered. The use of the Box-Cox transformation allows a more flexible form (linear form, logarithmic form or a compromise) of the link between the endogenous variable and each of the exogenous variables.

The generic model is written as follows:

$$\begin{cases} Y_t^{(\lambda_Y)} &= \sum_{k=1}^{K} \beta_k \, X_{kt}{}^{(\lambda_{X_k})} + u_t \\ u_t &= v_t \sqrt{\exp\left(\sum \delta_m Z_{mt}^{(\lambda_{Z_m})}\right)} \\ v_t &= \sum_{l=1}^{p} \rho_l \, v_{t-l} + w_t \end{cases} \qquad (3.3.13a)$$

with: Y_t the endogenous variable to be modelled,
X_{kt}, k=1 to K, the exogenous (or explanatory) variables,
u_t the first residual, and v_t the final residual,
w_t a white noise.

and the Box-Cox transformation defined as a power transformation, of parameter λ, on any positive real variable V_t by:

$$\begin{aligned} V_t \to V_t^{(\lambda)} &= \frac{V_t^{\lambda} - 1}{\lambda} \; if \; \lambda \neq 0 \\ V_t^{(0)} &= Log \, V_t \end{aligned} \qquad (3.3.13b)$$

In that general formulation, the Box-Cox parameters $\lambda_Y, \lambda_{X_1}, .. \; \lambda_{X_K}$ are estimated in addition to the other parameters β_k, δ_m and ρ_l, for k=1 to K, m=1 to M and l=1 to L.

3.3.4 Conclusion

In this introduction to dedicated time series models applied to road safety research, different types of models were defined. The need for a systemic, comprehensive approach - and the related methodological framework - are recalled. The major examples of aggregate time series modelling and analysis are given, and commented.

As it has been seen, different kinds and different classes of time series models have been selected for modelling aggregate risk indicators, at a country's level in Europe. The main difference between the models is the use of many versus few explanatory variables, but an important feature is their nature, whether deterministic or stochastic. The choice for a specific model is often governed by the purpose of the analysis, and unfortunately, often also by the availability of data.

The following sections of the methodology address mainly ARMA-type models and state space models. Nevertheless, it will be demonstrated on real road safety examples that the fact that a model belongs to one of the classes, or to one of the categories referred to in this chapter, is not exclusive.

3.4 ARMA-type models

Ruth Bergel and Mohamed Cherfi(INRETS)

3.4.1 Introduction

As it has already been emphasised, the dependencies over time of a stochastic, theoretical process (Y_t), for t being 1,2,3,…., can be modelled in different manners.

In a very special case of dependency over time - where the process in question (Y_t) is stationary[53]- , it is very practical to use the class of ARMA (autoregressive moving average) models, which enables us to describe the dynamics of the process and to extrapolate it in the future, without any call on additional variables, and with the only assumption that the process dynamics will stay unchanged at the forecast's horizon (see 2.2.1).

Nevertheless, the processes with dependencies over time usually are not stationary, because of the presence of a cycle, of a trend, or of a seasonal component: the sample of observations $Y = (y_1, y_2, ..., y_n)$, can rarely be considered as a sample of realisations of a stationary process. In that general case, it will be assumed that another stationary process exists, derived from Y_t by means of filtering, or by means of modelling before correcting for them, the non-stationary components of Y_t with the help of additional variables. It is this other stationary process, derived from Y_t, that will be modelled with an ARMA representation. In all cases, ARMA-type models will be used, which includes all the following cases: ARIMA models in the non-stationary case, ARMAX models in the case exogenous variables are used, and ARIMAX models in the non-stationary case and exogenous variables being used.

In all these cases, a stationary process, derived from Y_t, will be considered, and its dynamics estimated with the sample of observations at hand; as in the traditional ARIMA case, the model will constitute a tool for monitoring and for forecasting as well, if the exogenous variables used can also be forecasted or if scenarios for their development in the future can be established.

This section dedicated to ARMA-type models is structured as follows.
In Section 3.4.2, several ARMA models, fitted on simulated stationary data samples, are described. The interest of this preliminary section is that the structure of these simple models is very similar to the structure of the more elaborated models that will be fitted in the following sections on real road safety data, as far as handling their stationary part is required.
In Section 3.4.3, an ARIMA model is fitted on the annual number of road traffic fatalities observed in Norway for the period 1970-2003, as already described in Section 1.1.2 of the general introduction.

[53] its mean, variance and covariance structure are constant over time (see a precise definition in Section 2.2.4.3.1.)

In the two following Sections 3.4.4 and 3.4.5, ARIMA models are fitted on seasonal monthly data, with very similar structures in the sense that in both cases independent variables are used. The first dataset consists of the monthly number of drivers Killed or seriously injured on the road in UK, for the period January 1969 - December 1984, and the second one consists of the monthly number of fatalities registered in France, for the period January 1975 - December 2001. In both cases, the effects of several risk factors, of road safety measures and of special events were taken into account, and the related significant parameters were interpreted.
In the last and concluding Section 3.4.6, a summary of the models results as obtained on the real road safety data of the chapter is given: the estimated parameters are interpreted and the goodness of fit commented.

3.4.2 ARMA-models for stationary series (simulated data)

3.4.2.1. Objective of the technique

An ARMA-model is constructed for descriptive and forecasting purposes. It aims at giving account for the *dynamics* of a *stationary* process Y_t, when having a sample of observations $Y = (y_1, y_2, ..., y_n)$ at hand.

3.4.2.2. Model definition and assumption

A process $(Y_t)t \in Z$, of second order[54] is (weakly) stationary if its mean, variance and covariance structure do not depend on time:

$$\begin{aligned} &E(Y_t) = \mu \\ &\text{var}(Y_t) = \sigma^2 \\ &\text{cov}(Y_t, Y_{t+l}) = \text{cov}(Y_k, Y_{k+l}) \end{aligned} \qquad (3.4.1)$$

The fist equation defines the fist order stationarity, and the two following equations define the second order stationarity.

The constant covariance structure allows separating Y_t in two parts: the one related to the past at time t, and the part that is new at time t, which has a white noise property. This latter part of Y_t that is not correlated to its past is called « innovation » - as it is what is new to the process at time t -, and more precisely "innovation white noise" - as it is a white noise, due to the stationarity of Y_t.

There are different ARMA equivalent representations which could be retained for modelling a stationary process. Therefore, the "canonical" form, which is unique, is currently retained as the simpler manner for expressing Y_t as a linear combination

[54] Having a finite mean and a finite variance

of a small number (p) of its past values, and of a small number (q) of the past errors (or disturbances, as it will be explained later on).

The canonical ARMA (p,q) representation:

$$Y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + u_t + \sum_{j=1}^{q} \theta_j u_{t-j} \qquad (3.4.2)$$

is usually written in the following way:

$$\Phi(B) Y_t = \Theta(B) u_t , \qquad (3.4.3)$$

with: $\Phi(B)$ and $\Theta(B)$ two polynomials[55] of the delay operator B, of degrees p and q,
and u_t the « innovation » white noise.

The backshift operator B used in the previous representation is an operator on an element of a time series, that produces the previous element in time of that time series: $BY_t = Y_{t-1}$.

Note that $B^2 Y_t = B(BY_t) = BY_{t-1} = Y_{t-2}$, and so on. In particular, for a monthly time series, $B^{12} Y_t = B \cdot B \cdot ... \cdot BY_t = ... = B \cdot BY_{t-10} = BY_{t-11} = Y_{t-12}$ yields the observation exactly one year before.

Similar to classical polynomials, a polynomial of order p in B can be written as:

$$\Phi(B) = \phi_0 + \phi_1 B + ... + \phi_p B^p ,$$

and, in the case ϕ_0 is 1, we have the unitary polynomial in B:

$$\Phi(B) Y_t = 1 Y_t + \phi_1 BY_t + ... + \phi_p B^p Y_t = Y_t + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} .$$

Please note that the polynomial representation described above is a convenient notation for specifying ARMA models - rather than higher mathematics. In the case of the canonical ARMA (p,q) representation, two unitary polynomials in B, $\Phi(B)$ and $\Theta(B)$ of orders p and q, are used: the first one is applied to the process Y_t , and the second one to the innovation white noise u_t.

The few examples that are given now will help understand this representation.

[55] Conditions are required from the polynomials of the canonical representation: to be unitary, with no common root, the roots of Φ (strictly) outside the unit circle, and the roots of Θ outside the unit circle, see(Box, Jenkins, 1976).

3.4.2.3. Research problem and dataset

Four stationary datasets have been simulated with the help of the following formulas, in which a Gaussian "white noise" (a_t are independently normally distributed with mean 0 and variance 1 (N *(0,1)*) and t = 1,…, 1000) was generated:

$$Y_t = 0.8Y_{t-1} + a_t + 3 \qquad (3.4.5a)$$

$$Y_t = 0.5Y_{t-1} + 0.3Y_{t-2} + a_t + 3 \qquad (3.4.5b)$$

$$Y_t = a_t - 0.6a_{t-1} + 5 \qquad (3.4.5c)$$

$$Y_t - 0.5Y_{t-1} - 0.3Y_{t-2} = a_t - 0.6a_{t-12} + 8 \qquad (3.4.5d)$$

The four fsample processes are, as constructed:

- two *autoregressive* of order 1 and 2 processes,
- a *moving average* of order 1 process,
- an *autoregressive and moving average* of orders 2 and 12 process.

Figures 3.4.1 to 3.4.4 describe the development over time of the 200 hundred first values of the sample processes.

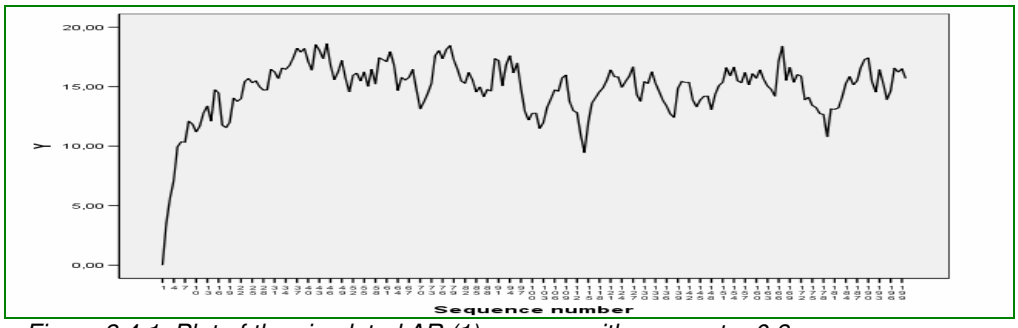


Figure 3.4.1: Plot of the simulated AR (1) process with parameter 0.8

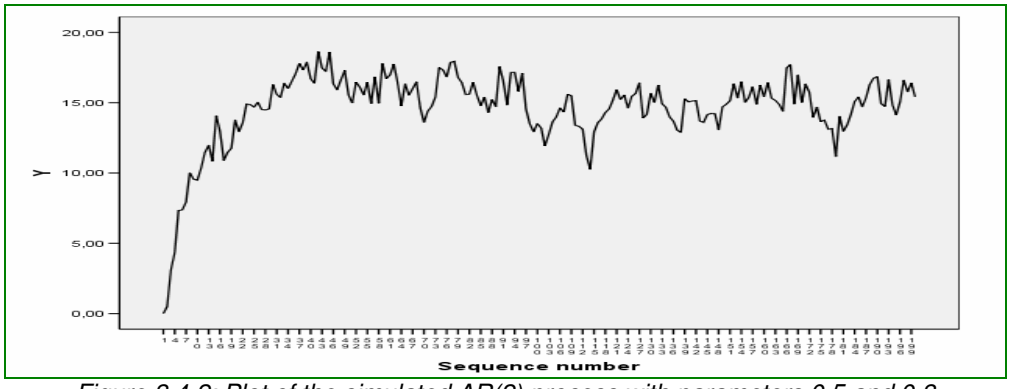


Figure 3.4.2: Plot of the simulated AR(2) process with parameters 0.5 and 0.3

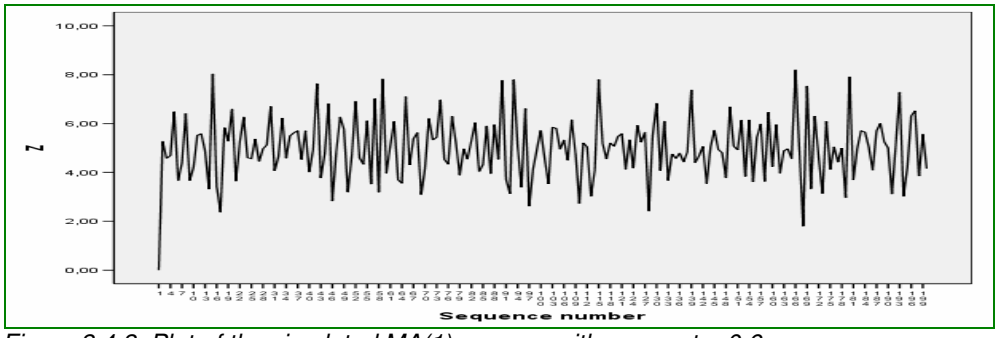


Figure 3.4.3: Plot of the simulated MA(1) process with parameter 0.6

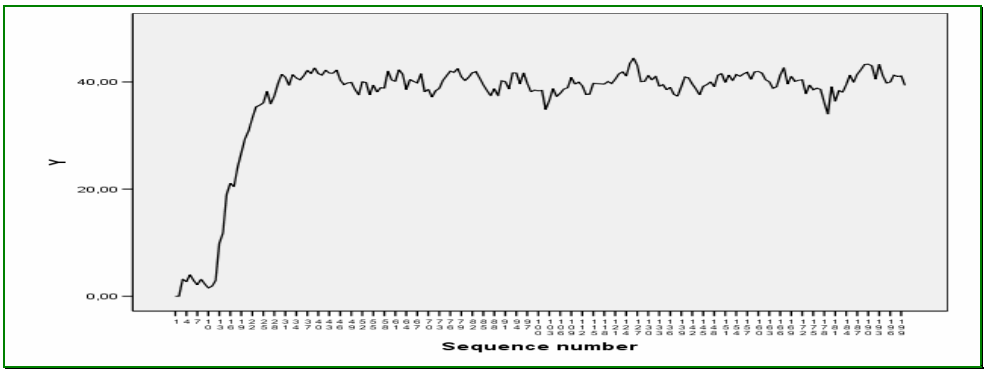


Figure 3.4.4: Plot of the simulated ARMA(2,12) process with parameter 0.5, 0.3 and 0.8

The property of stationarity can be briefly described by the fact that, whatever the initial values of a process are, the values it takes will rapidly reach a certain level (its mean) and stay around it, and vary constantly around that mean. This can be observed in Figures 3.4.1 to 3.4.4.

3.4.2.4. Model fit

3.4.2.4.1. Identification

The model identification (choice of the two integers *p* and *q*) is performed by examining both the autocorrelation function (ACF)) and the partial autocorrelation function (PACF) plots (see Box and Jenkins, see also Section 3.2.1).

The model identification is classically performed *in two stages*:

- First by fitting a pure AR (of order p_0), and a pure MA model (of order q_0).
- Second, by fitting the parsimonious canonical ARMA(p,q) as satisfying the condition: ($p<=p_0$,$q<=q_0$).

The first stage is the difficult part of the identification, as theoretical properties are tested for determining the orders of the pure AR and MA specifications.

Autoregressive processes of order p_0 have exponentially (or sinusoidal) decaying AC values, and their PAC values of order larger than p_0 are zero. Moving average of order q_0 have exponentially (or sinusoidal) decaying PAC values, and their AC values of order larger than $q_0$are zero.

In practice, the hypothesis of nullity of an AC (or PAC) is rejected when the 95% level confidence interval, centered on the estimated AC (or PAC) value, does not include zero. But the risk of rejecting the nullity hypothesis, and thus of considering as significant an AC (or PAC) value which should not be considered as significant, leads to over parameterised and mis-specificated models. Therefore, the decision of rejecting the nullity hypothesis should be taken cautiously, and the test confidence level should preferably be lowered in practise.

The related ACF and PACF plots, used for identifying the four models which will be fitted on the simulated datasets, are summarised in Figure 3.4.5.

In the three first cases of Figure 3.4.5, the classical patterns of two autoregressive processes and of a moving average process are found, indicating that there is no need for further identification. The exponential (or sinusoidal) decay appears to be more or less obvious, but the relevant information has to be taken where it appears to be highly significant, whether from the ACF plot or from the PACF plot.

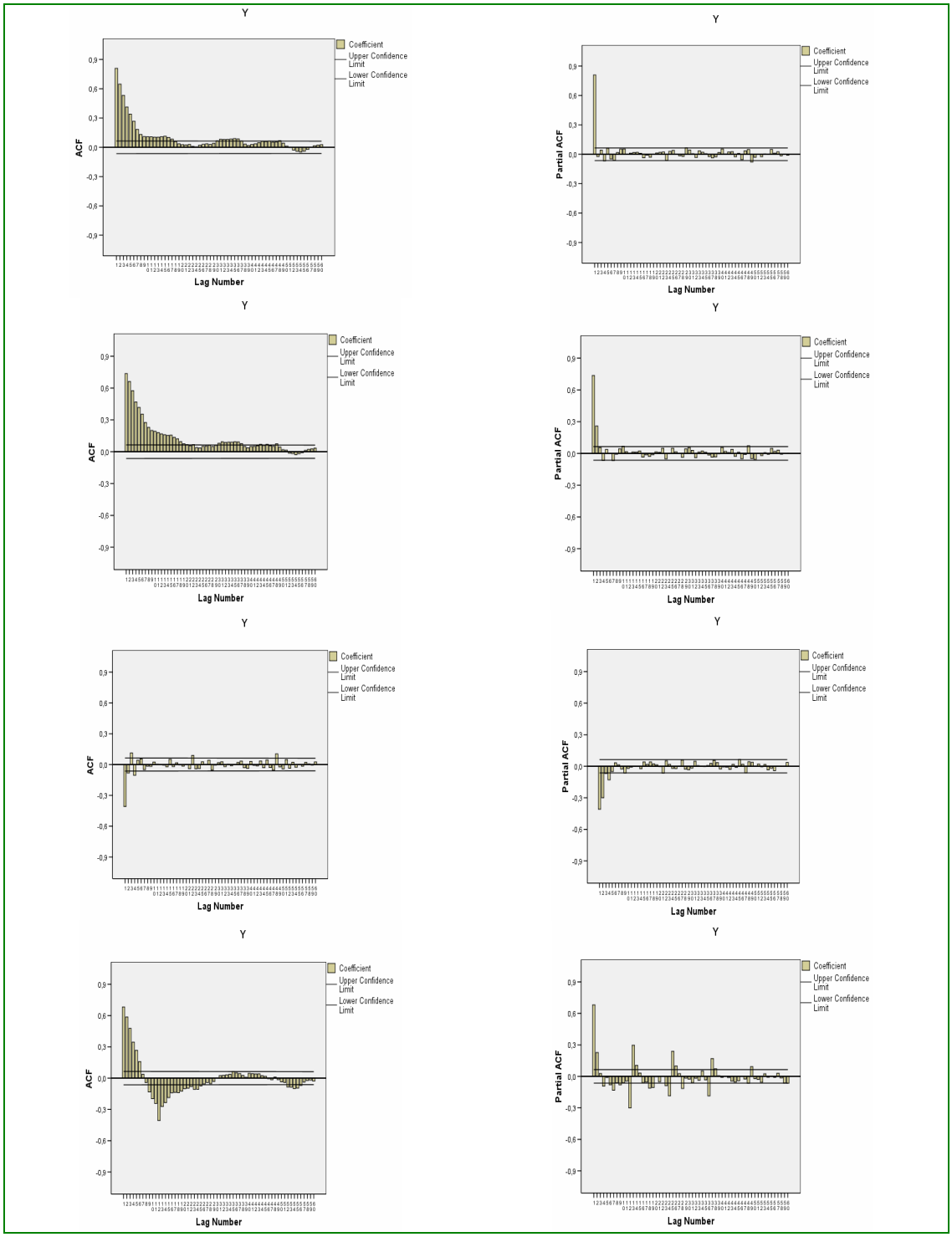


Figure 3.4.5: ACF and PACF plots of the simulated samples (Summary),

Note that in the particular case of a seasonal process of period s, the seasonal part of the model is often separated from the non-seasonal part, in a multiplicative manner. In that case, the four integers of the seasonal part (P and Q) and of the non-seasonal part (p and q) have to be determined. In the last case of Figure

3.4.5, the seasonal pattern and the non-seasonal (short term) pattern are to be visually considered separately.

3.4.2.4.2. Estimation

For each model, the p+q +1 (the p ϕ_i and q θ_j , for i=1 to p and j=1 to q, and in addition the variance of the residuals) parameters are then estimated by means of maximising the log likelihood function, which comes to minimize the sum of the squares of the residuals.

The estimation results are given in the following Table 3.4.1, and will be commented in 3.4.2.5.2.
For efficiency reasons, the initial values of each simulated sample have been excluded from the sample on which the model was fitted..

					Estimate	SE	t	Sig.
Y-Model_1	Y	Simulated AR(1) sample	Constant		15,058	,172	87,751	,000
			AR	Lag 1	,809	,019	42,690	,000

					Estimate	SE	t	Sig.
Y-Model_1	Y	Simulated AR(2) sample	Constant		15,066	,169	89,408	,000
			AR	Lag 1	,543	,031	17,407	,000
				Lag 2	,263	,031	8,426	,000

					Estimate	SE	t	Sig.
Y-Model_1	Y	Simulated MA(1) sample	Constant		5,006	,013	373,354	,000
			MA	Lag 1	,586	,026	22,715	,000

					Estimate	SE	t	Sig.
Y-Model_1	Y	Simulated ARMA (2,12) sample	Constant		39,996	,039	1035,172	,000
			AR	Lag 1	,540	,031	17,271	,000
				Lag 2	,259	,031	8,265	,000
			MA, Seasonal	Lag 1	,772	,021	35,904	,000

Table 3.4.1: Estimation results - Models fitted on the simulated samples.

At that point, two relevant outputs are available: the *" estimated"* series on the one hand - also called *" fitted"* , *"adjusted" or "predicted" series* -, and the *"residual" series* on the other hand, over the estimation period. Note that the first one is made of the one step ahead predictions; whereas the second one, which is the difference between the sample series and the adjusted series, is the best estimation of the innovation white noise, the part of the process Y_t which is not correlated to the past of Y_t at time t.

Figures 3.4.6 and 3.4.7 describe the development of these two series, in the case of the ARMA(2,12) sample.

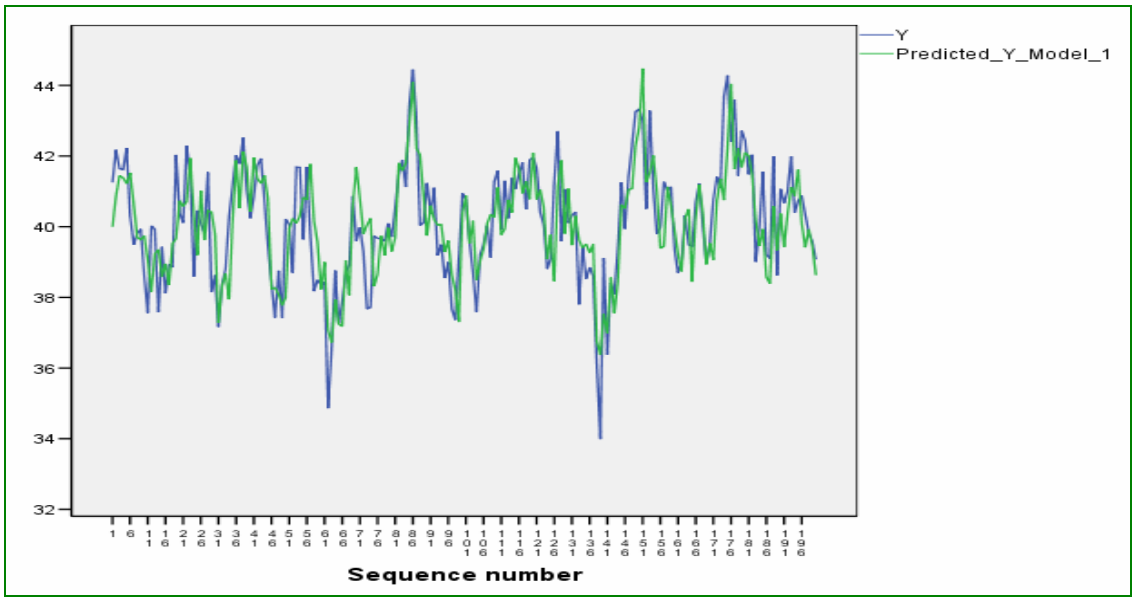


Figure 3.4.6: Plot of the simulated ARMA(2,12) process, and of the adjusted series

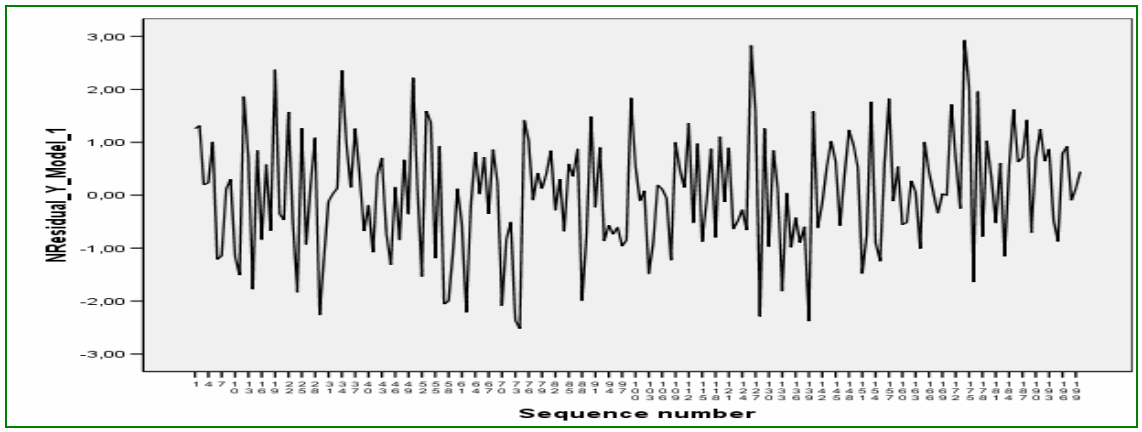


Figure 3.4.7: Plot of the estimated innovation white noise.

3.4.2.5. Model diagnostics

3.4.2.5.1. Validation and empirical performance

Tests are used to validate the model, and criteria to evaluate the model's empirical performance. These tests and criteria will be exposed first, and then the results related to the application case.

Tests are used for validating the model

A difference is to be made between the tests related to the residuals, and the test used for validating each parameter (Student's test).

The test related to the residuals consist in testing the « white noise » property - ie mainly the non-correlation property (Ljung-Box's test, for instance) – and the Gaussian property (Shapiro-Wilk's or Kolmogorov-Smirnov's test) of the error term of the model. Among these properties, non-correlation is fundamental. If the assumption of normality is violated, the log likelihood computation can be compromised, but the estimators may nevertheless have good asymptotic convergence properties. However, it is fundamental that the assumption of non correlation of the residual is tested, because if it is rejected the model's specification has to be changed

In practice, all tests related to the residuals are not performed: the non correlation and the Gaussian property are tested, and in case they are not rejected, the independence of the residuals is assumed.

Criteria[56] are used for evaluating the model's empirical performance.

They relate to the model's adjustment, or forecasting power. Let's mention in the first group the proportion of explained variance (R-squared or stationary R-squared), as well as the different criteria which enable to evaluate the estimation fit: the root mean square error (RMSE), and the widely used mean absolute percentage error (MAPE), and in the second group the Bayesian information criterium (BIC) or the Akaike information criterium (AIC), and the Bayesian criterium of Schwarz (SBC).

Several models proposed for the same sample of data will be compared after the test and criteria, mentioned above, have been performed. Two nested models will be compared by using a likelihood ratio test, which can lead to a reduction in the number of parameters in an over-parameterised model.

A practical question finally is, after the model has been validated, whether the model is stable over time. The parameters' stability will be discussed by comparing estimations obtained from different samples of data covering different time intervals. The responses to the validation tests and empirical performance criteria might also differ with each new sample of data.

3.4.2.5.2. Application cases

In the example cases, the Student's tests lead to conclude that all parameters were significant (the null hypothesis is rejected at the 95% confidence level).

As for the tests performed on the model residuals, the hypothesis of non-correlation, at each order, is accepted, as shown in Figure 3.4.8, and the hypothesis of global uncorrelation (from order 1 to 18) is also accepted, as shown in Table 3.4.2.

[56] Whereas the R-squared, the RMSE and the MAPE are currently computed by all softwares, it is not allways the case for all information criteria (see the manual).

Model	Ljung-Box Q(18)		
	Statistics	DF	Sig.
AR(1)	24,768	17	,100
AR(2)	23,066	16	,112
MA(1)	25,423	17	,086
ARMA(2,12)	24,731	15	,054

Table 3.4.2: Ljung-Box statistics (Summary)

Note that no additional normality test was performed on the residuals, as the white noise used for simulating the datasets was generated as Gaussian.

Finally, some criteria enabling to evaluate the model's empirical performance are given in Table 3.4.3. The model's performance is lower in the case of the MA model, as the R-squared is at the lowest, around 25%, and the absolute error made on the estimation period, measured in mean and in percentage, takes its highest value, around 19%. The best performance is obtained for the ARMA model, with a R-squared value around 69%, and an average error around 2%.

Fit Statistic	AR (1)	AR (2)	MA (1)	ARMA (2,12)
R-squared	,655	,574	,249	,685
MAPE	5,531	5,513	18,592	2,052
Normalized BIC	,053	,061	,049	,071

Table 3.4.3: Goodness of fit criteria (Summary)

3.4.2.6. Model interpretation

The small number of parameters of the autoregressive and moving average polynomials of the ARMA(p,q) canonical representation - three at the most, in the example given - enables to define the past of the process, and to determine its future. In this parsimonious expression, the (one step ahead) forecasted value of the process is determined with the only knowledge of a small number of past values and a small number of past (one step ahead) forecast errors. In the example given, the memory of the process is taken into account with the help of the two parameters 0.5 and 0.3, and the link with the past error with the help of the parameter 0.8.

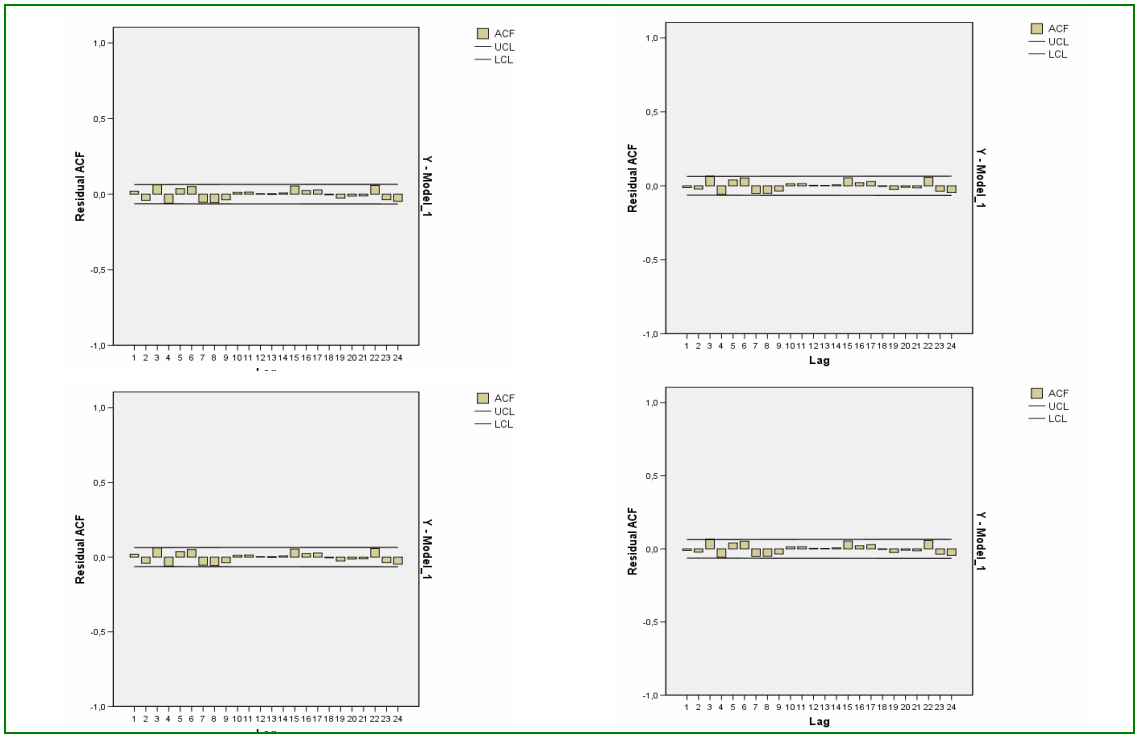


Figure 3.4.8: ACF plots of the models residuals (Summary).

3.4.2.7. Conclusion

In this section, ARMA models were fitted on four simulated stationary datasets of 1000 observations from which, for efficiency reasons in the estimation stage, the first values where excluded. In each case, the model identification was described, the model estimation results were validated with the help of tests, and the model's empirical performance evaluated with the help of criteria. The goodness of fit statistics showed very important differences among the datasets: the worse performance was obtained with the MA(1) model and the best one with the ARMA(2,12) model.

It is worth mentioning that the estimated (dynamics) parameters were in all case significant, and very near to the real - and, in these cases, known - values. However, in practice, the fact that every parameter is subject to estimation errors and that a model is generally estimated with numerous parameters on a dataset of smaller size, may lead to lower the confidence level of significance tests.

.

In the following Sections 3.4.3 to 3.4.5, the ARMA structure of the models that will be fitted on real road safety datasets is similar to the one described in Section 3.4.2.

3.4.3 ARIMA models for non seasonal series (Norway fatalities)

3.4.3.1. Objective of the technique

An ARIMA-model is, as in the preceding subsection, constructed for descriptive and forecasting purposes. It aims at giving account for the *dynamics* of a *non stationary* process Y_t, when having a sample of observations $Y = (y_1, y_2, ..., y_n)$.
In this section, the general case of an ARIMA model will be considered, without any consideration of seasonality.

3.4.3.2. Model definition and assumptions

In the general case where Y_t is not stationary, it is possible to apply a filter of differences to the process, in such a way that the transformed process Y_t defined as:

$$F(B)Y_t,$$

with $F(B) = (I - B)^d$, B the delay operator and d a positive value,

becomes stationary, and then model this transformed process $F(B)Y_t$ with an ARMA(p,q) model.

In such a case, we shall have an ARIMA (p,d,q) representation for the non-stationary process Y_t :

$$\Phi(B)F(B)Y_t = \Theta(B)u_t \qquad (3.4.6a)$$

Note that for d being 1, $(1-B)Y_t = 1Y_t - BY_t = Y_t - Y_{t-1}$, so the approach above would in one turn change a linear trend into a stationary series. The integer d is often taken as 1, and is rarely larger than 2. Differencing twice would for instance turn a quadratic development into a stationary one.

3.4.3.3. Research problem and dataset

The dataset consists of the annual number of road traffic fatalities observed in Norway for the period 1970-2003[57], as already described in Section 1.2.2. of the general introduction The research problem consists in determining a time series model *both for descriptive and forecasting purposes*. In this particular case, the explanatory capacity of the model will not be addressed, as no additional independent variable will be used for modelling the sample observations.

[57] More precisely, the log of the annual number of fatalities will be the modelled data - and not the absolute annual number.

3.4.3.4. Model fit

In the case of ARIMA models without any exogenous variables, the well known following stages are succeedingly considered: stabilisation, identification, estimation and validation, as described for instance in (Box and Jenkins, 1976).
The three last stages dedicated to stationary time series have already been described in 3.4.2. The first stabilisation stage is necessary every time the sample dataset is not a stationary one. An easy manner for doing this consists[58] in applying a difference filter to the initial dataset, in such a way that the filtered dataset can be considered as stationary

In the application case, one difference was applied to the initial log transformed data, and no presence of non stationarity could be detected in the ACF plot[59], which led to accept the hypothesis that the one difference filtered data were a sample of a stationary process.

The observation of both the ACF and PACF plot then led to retain a moving average model of order 1 to fit the filtered data. Finally, the model fitted on the log Norwegian fatalities is an ARIMA (0,1,1) model.

3.4.3.5. Model diagnostics

As indicated in Table 3.4.4, the moving average parameter (θ_1) was estimated at 0,432 and the constant term at -0,020. In both cases, the result of the student test was that the hypothesis of nullity of these two (theoretical and unknown) parameters had to be rejected, at the usual 95% confidence level).

As for the residuals, the hypothesis of nullity of each autocorrelation, from order 1 to order 24, was to be accepted, as shown in Figure 3.4.9.
Moreover, the hypothesis of global non correlation of the residuals was also tested. The Ljung-Box statistic provides an indication of whether the model is correctly specified, in the sense it allows testing the global nullity of the autocorrelation of the residual (from order 1 up to order 18). The hypothesis was accepted, as the 0,510 value of the Ljung-Box statistic is more than 0.05, as indicated in Table 3.4.5.

The normality of the residuals was graphically tested with the help of the histogram and of the QQ-plot, shown in Figures 3.4.10 and 3.4.11.
Moreover, the non-parametric Kolmogorov-test was also performed on the residuals.

[58] There are several manners for deriving a stationary dataset from the initial one: extracting the trend as a function of time, for instance, is currently performed.
[59] The property of stationarity can not be checked visually, because the sample length is generally too short to give the right overview of the dynamics of the process. Nevertheless, the presence of non stationarity can be detected visually: in the case of a stationary dataset, the autocorrelations should decrease exponentially after a certain order.

In the case the Kolmogorov-Smirnov test is significant, the normal distribution of the residual hypothesis is to be rejected. This hypothesis was accepted, as the 0,713 value of the Asymp. Sig. (2-tailed) is more than 0.05 (at the usual 95% confidence level), as indicated in Table 4.3.6. Therefore, the hypothesis of independence of the residuals is also accepted.

The model's empirical performance was evaluated by computation of different kinds of goodness of fit statistics, given in Table 3.4.7; but this evaluation really makes sense in the case several (nested) models have been fitted on the same data, and their empirical performance can thus be compared

Due to the presence of the trend, the stationary R-squared is only 16,7% (the model explains 16,7% of the variance of the filtered data, compared to a regression model), and much smaller than the R-square which is 78,9% (the model explains 78,9% of the variance of the initial data).

As for the usual measure of the error made: the mean absolute percentage error (MAPE) is 1,36%, whereas its highest value observed on the estimation period is 3,915%,

At last, the normalized BIC, which is -4,413, is a goodness of fit measure that takes account of the parsimony of the model. Note that, as it is the case for the R-squared, its interest lies in comparisons between several models, and not in its absolute value.

3.4.3.6. Model interpretation

In the case the initial data are filtered, to interpret the model's parameters is not easy as the formulation is slightly more complicated. However, the same global interpretation can be given as in the preceding section, in the sense that the fitted value is a function of a small number of the past values of the process, and of a small number of the past forecast errors.

,

However, this ARIMA(0,1,1) representation has an equivalent local level representation, which will be described in Section 3.6 dedicated to state space methods. As such, the local level fitted on this dataset will be interpreted in Section 3.6.1.

As demonstrated in (Harvey, 1989), the relationship between the parameter θ_1 of an ARIMA(0,1,1) and the parameter $q = \sigma_\eta^2 / \sigma_\varepsilon^2$ of the local level model, is the following one:

$$\theta_1 = \frac{1}{2}\left(\left(\sqrt{q^2 + 4q}\right) - (q+2)\right) \qquad (3.4.7)$$

The two noise variances of the local level were estimated by using Ox/SsfPack, which led to q=0;0047026/0.00326838 and thus θ_1 was calculated to be -0,32070152; this value is very close to the one estimated by SPSS, which is given

in Table 3.4.8 in the case of the ARIMA(0,1,1) model without constant term, and which is precisely θ_1=0,32069194.

					Estimate	SE	t	Sig.
LNorv-Model_1	LNorw	Norwegian Fatalities	Difference		1			
			MA	Lag 1	,321	,170	1,888	,068

Table 3.4.8: Estimation results for the ARIMA(0,1,1) model without constant term

3.4.3.7. Conclusion

In this section, an ARIMA (0,1,1) model was fitted on the log-transformed annual number of road traffic fatalities observed in Norway for the period 1970-2003.
The models diagnostics were satisfactory, in the sense that all parameters were significant, and that the residuals could be considered as independent.

At last, it was shown in this example that the ARIMA(0,1,1) representation is equivalent to a local level representation, of the class of state space presented in Section 3.6.

					Estimate	SE	t	Sig.
LNorv-Model_1	LNorw	Norwegian Fatalities	Constant		-,020	,010	-1,969	,058
			Difference		1			
			MA	Lag 1	,432	,164	2,636	,013

Table 3.4.4: Estimation results for the ARIMA(0,1,1) model

Model	Ljung-Box Q(18)		
	Statistics	DF	Sig.
LNorv-Model_1	16,199	17	,510

Table 3.4.5: Ljung-Box statistic for the residuals of the ARIMA(0,1,1) model

		Noise residual from LNorw-Model_1
N		33
Normal Parameters(a,b)	Mean	-,0009
	Std. Deviation	,09744
Most Extreme Differences	Absolute	,122
	Positive	,122
	Negative	-,078
Kolmogorov-Smirnov Z		,699
Asymp. Sig. (2-tailed)		,713

Table 3.4.6: Kolmogorov-Smirnov statistic for the residuals of the ARIMA(0,1,1) model

Fit Statistic	
Stationary R-squared	,167
R-squared	,789
RMSE	,099
MAPE	1,362
MaxAPE	3,915
MAE	,080
MaxAE	,230
Normalized BIC	-4,413

Table 3.4.7: Goodness of fit criteria for the ARIMA(0,1,1)

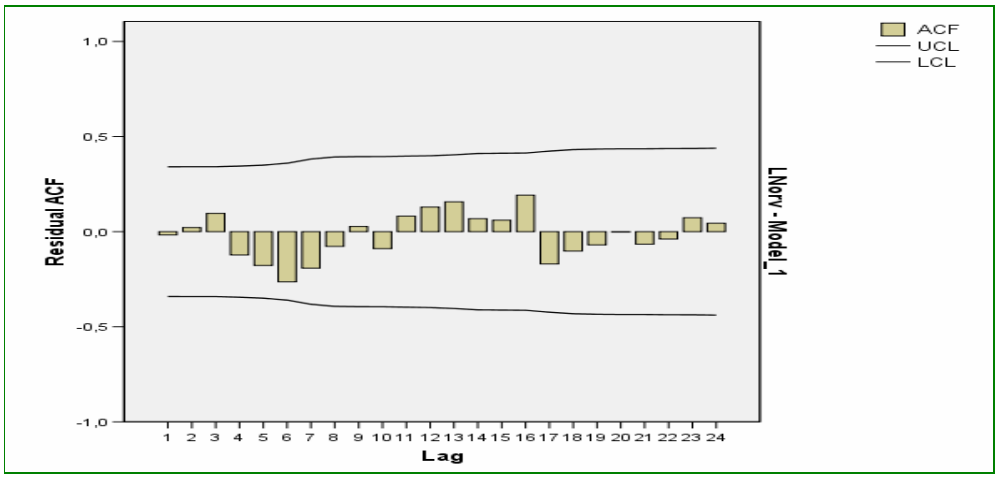


Figure 3.4.9: The ACF plot of the residuals and their confidence interval.

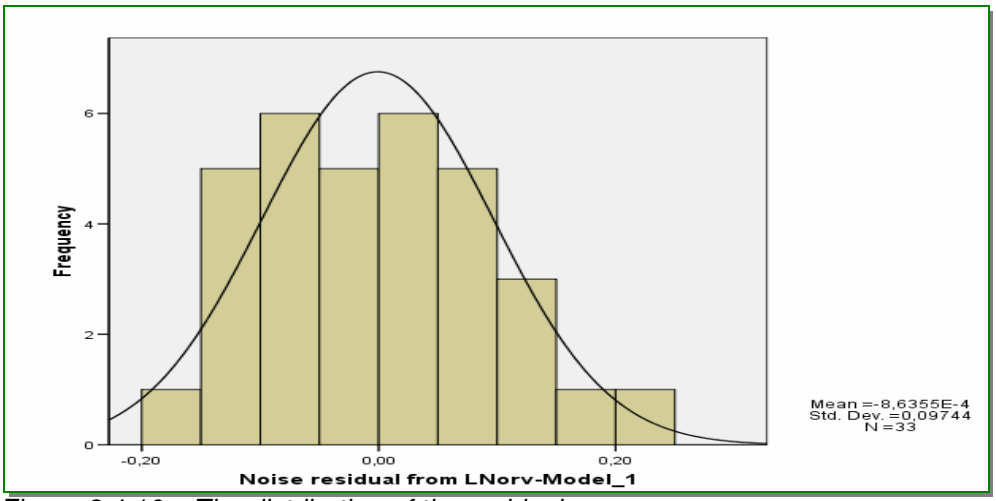


Figure 3.4.10. : The distribution of the residuals

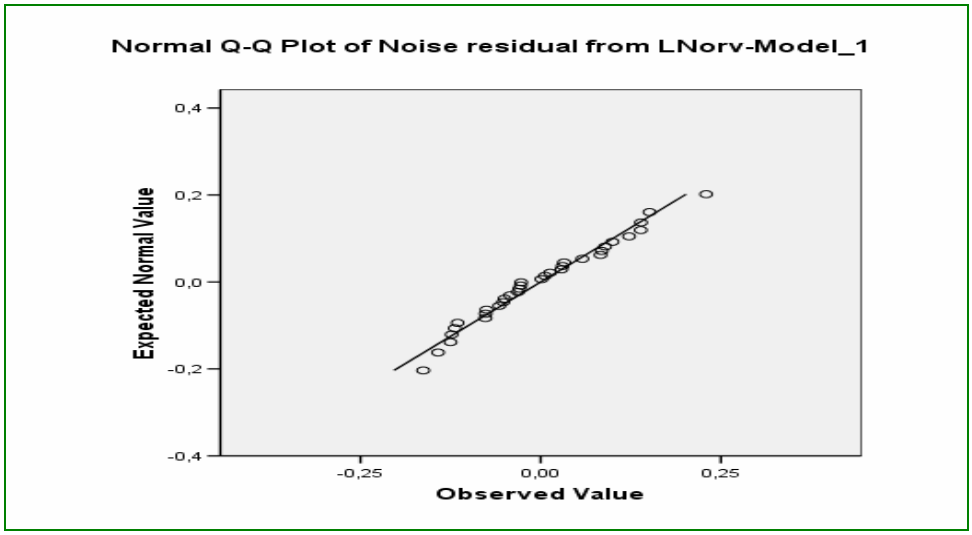


Figure 3.4.11.: The QQ-plot

3.4.4 ARIMA models for seasonal series (UK-KSI drivers)

3.4.4.1. Objective of the technique

In this subsection, an ARIMA-model with exogenous variables is constructed for descriptive, explanatory and forecasting purposes. It aims at giving account for both the *dynamics* of a *non stationary* process Y_t and the *influence of exogenous factors*, when having a sample of observations $Y = (y_1, y_2, ..., y_n)$ at hand.

In this section, seasonality of the process is considered, and treated in a multiplicative manner.

3.4.4.2. Model definition and assumptions

In the general case where Y_t is not stationary and has a seasonal (periodic) component, the ARIMA (p,d,q) representation defined in (3.34a) can be extended to the more general ARIMA $(p,d,q)(P,D,Q)_s$ representation, in which the seasonal and non seasonal parts of the dynamics can be separated in a multiplicative manner :

$$\Phi(B)\Phi_s(B^s)F(B)Y_t - \Theta(B)\Theta_s(B^s)u_t, \qquad (3.4.6b)$$

with $F(B) = (I - B)^d (I - B^s)^D$, B the delay operator, d and D two positive values and s the periodicity of the seasonal process.

In other words, first the seasonal pattern is removed, and then the remaining trend. Note that, In the case of this multiplicative ARIMA representation, 4 polynomials in B will be estimated instead of 2.

Moreover, when independent variables are introduced in the model, there are different manners of taking account of them. The following form is retained for commodity reasons, if the data corrected for the exogenous effects are stationary:

$$\Phi(B)\left[Y_t - \sum_{i=1}^{K} \Phi_i(B) Z_{it} \right] = \Theta(B) W_t \qquad (3.4.7)$$

with: Y the endogenous variable to be modelled (eventually filtered with a difference filter F(B)),

Z_i the K exogenous variables (eventually filtered with difference filters F_i (B)),

W a white noise not correlated with the past Y and of the Z_i ,

and $\Phi . \Phi_i . \Theta$ polynomials in B.

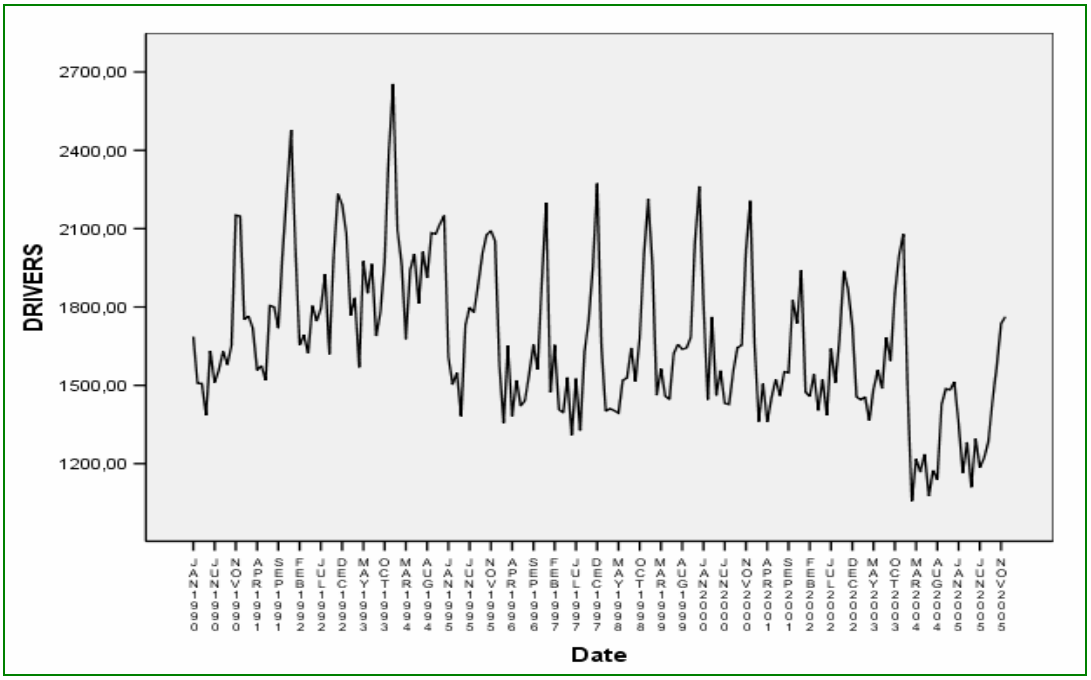


Figure 3.4.11: The monthly number of UK-KSI driver, for the period January 1969 - December 1984.

In this specification, the endogenous variable and the K exogenous variables are if required filtered with difference filters F(B) and F_i (B), but it may as well not be necessary, if the exogenous variables help to correct for the trend and the seasonality of the process (Y_t) .

The main assumption is the stationary of the data, corrected for the exogenous effects, as written in the general specification (3.4.7).
This hypothesis of stationarity is tested on the residual of the model, which is a white noise if this hypothesis is valid.

3.4.4.3. Research problem and dataset

The example retained in this section is the one described in (Harvey, Durbin, 1986).
The dataset consists of the monthly number of drivers, killed or seriously injured in the UK, for the period January 1969 - December 1984 (UK-KSI drivers), described in figure 3.4.11. Three exogenous variables (an intervention variable and two explanatory variables) will be introduced in the model in two successive steps.
The intervention analysis takes account for the obligation, from February 1983 onwards, for motor vehicle drivers and front seat passengers to wear a seat belt: thus, an intervention variable, equal to 1 from February 1983 onwards, and equal to 0 before, was constructed.
The two explanatory variable are: the monthly car traffic index (more precisely the monthly number of vehicle-kilometres driven by cars in the UK), and the monthly prices of petrol in UK; for the period January 1969 - December 1984.
Note that this dataset is the one used in Section 3.6.3. for fitting a local linear trend plus seasonal model of the class of state space models.

3.4.4.4. Model fit

When exogenous variables Z_i are introduced into ARIMA models, it is not feasible to consider the usual stages (stabilisation, identification, estimation, validation) before the functional form between Y and each exogenous variable Z_i has been established, because a preliminary estimation of the exogenous effects has to be obtained, so that the stationarity of the process, corrected for the exogenous effects, can be evaluated.
In practise, an econometric specification is retained, all parameters are estimated together, whether related to the endogenous or exogenous variables; and the diagnostic tests, carried out after the model has been estimated, replace the two first stages (stabilisation and identification) which could not be considered before.

The results obtained after estimating the model (3.4.8) are given in Table 3.4.9.

$$\Phi(B)(I-B12)\left[\log Y_t - \sum_{i=1}^{I} \alpha_i LogZ_{i,t} - Step_t\right] = \mu + \Theta(B)a_t \qquad (3.4.8)$$

with: Y the number of UK-KSI drivers,
$Z_{i,i=1toI}$ the car traffic index and the petrol price,
$Step_t$ a dummy variable equal to 1 starting February 1983 and to 0 before,
$\Phi(B)$ and $\Theta(B)$, two polynomials of the delay operator B,
and a_t a white noise.

3.4.4.5. Model diagnostics

The hypothesis of nullity of the model parameters is rejected (at the 95% confidence level), except for the log of the traffic index variable parameter: Thus, all parameters related to the dynamics are to be considered as different from zero, and the petrol price parameter and the intervention parameter too.
Note that, in case the confidence level is lowered to 70% for instance (t-value between 1 and 2), the parameter related to the traffic index variable would also be considered as different from zero.

The hypothesis of global non-autocorrelation of the residuals is accepted, and the hypothesis of normality of the residuals is accepted too, as can be seen from Tables 3.4.10 and 3.4.11, which therefore enables to accept the independence hypothesis.

Regarding the model fit criteria given in Table 3.4.12, the stationary R-squared is only 59,0% (the model explains 59,0% of the variance of the filtered data, compared to a regression model), whereas the R-square is 80,2% (the model

explains 80,2% of the variance of the initial data); the mean absolute percentage error is 0,902% %, its highest value observed being 3,841%,

3.4.4.6. Model interpretation

The dynamics estimated is related to the corrected for exogenous effects process, the one that is assumed to be stationary. It is worth noting here that, at the difference of the well-known "airline model" structure proposed by Box and Jenkins[60], for which the filter $(I-B)(I-B12)$ was applied to the log transformed data, the simple filter $(I-B12)$ in (3.4.8).

As for the exogenous part, it's natural to try to interpreter the relationship[61] between the exogenous variables Z_{it}, i=1 to k and the endogenous variable Y, regardless of the dynamics.

In the application case, a special effort is to be paid to the three exogenous effects parameters. As this dataset has already been used for fitting state space models, estimations for these three parameters were already given (Harvey, Durbin, 1986).

Thus, since the intervention parameter is estimated at -0,163, and due to the relation exp (-0,163) =0,85, the reduction in the number of drivers killed and seriously injured in the UK February 1983 onwards is estimated at 15%.

The two other parameters are (constant) elasticity values : the elasticity value of the number of drivers killed and seriously injured with respect to the traffic volume index is estimated at 0,134 (at the 70% confidence level), whereas the elasticity value of the same indicator with respect to the petrol price is estimated at -0,297 (at the usual 95% confidence level). In the case the traffic index is not kept in the model, this later elasticity value is estimated at -0,323, which is not very different from the preceding estimation, whereas the other parameters vary very little.

---

[60] *This model was fitted on the monthly number of international airline passengers in thousands, for 1949- 1960, series G in (Box, Jenkins, 1976)*

[61] *Apart from commenting on the value of the parameter β_i of the variable Zi, the interest often goes to the related elasticity function, given by: $\frac{d(LogY)}{d(LogZ_i)}$.*

For small variations of Z_i, at a given time, the following formulation for the elasticity of the endogenous variable Y with respect to an exogenous variable Z_i, is used:

$$\varepsilon_{Y/Zi} = \frac{\Delta Y / Y}{\Delta Z_k / Z_i}.$$

In the very special case where both variables have been log transformed, the parameter β_i indeed represents the elasticity of Y with respect to Zi, which is then constant. But it is important to note that one does generally comment an « apparent elasticity » of Y to Zi, because the condition of mutual orthogonality of the exogenous variables Z_{it}, i=1 to k, is rarely valid.

With a state space model fitted on the same dataset, Harvey and Durbin estimated at 23% the reduction in the number of drivers killed and seriously injured in the UK February 1983 onwards and at -0,31 the elasticity value of the number of drivers KSI in the UK with respect to the petrol price - whereas the traffic index effect appeared to be non significant, far beyond the 70% confidence level.

3.4.4.7. Conclusion

In this section, a multiplicative ARIMA $(2,0,0)(0,1,1)_{12}$ model was fitted on the log transformed monthly number of drivers killed and seriously injured in the UK, for the period January 1969 - December 1984 (UK-KSI drivers).
The effect of the obligation of wearing a seat belt in the UK, from February 1983 onwards, for motor vehicle drivers, was investigated by the call of an intervention variable. The effects of the risk exposure and the petrol price variations, were also investigated by the call to two other additional variables: the monthly car traffic index (more precisely the monthly number of vehicle-kilometres driven by cars in the UK), and the monthly prices of petrol in UK.

The models diagnostics were satisfactory, in the sense that all parameters were significant, and that the residuals could be considered as independent. One exception is to be made for one exogenous effect parameter, related to the traffic index variable, which could only be considered as significant at the 70% lower confidence level. Thus, a 15% reduction in the number of UK-KSI February 1983 ownwards was observed, and an elasticity of -0.32 of the number of UK-KSI with regard to the petrol price was obtained.

The model's empirical performance was evaluated by computation of different kinds of goodness of fit measures, and the model's performance increased about 5% with the introduction of all exogenous variables, as indicated in Table 3.4.18.

					Estimate	SE	t	Sig.
LDRIVERS-Model_1	LDRIVERS	UK-KSI drivers	Constant		-,015	,006	-2,463	,015
			AR	Lag 1	,283	,075	3,800	,000
				Lag 2	,235	,077	3,072	,002
			Seasonal Difference		1			
			MA, Seasonal	Lag 1	,857	,078	10,930	,000
	LPPRICE	Traffic volume	Numerator	Lag 0	-,297	,095	-3,132	,002
			Seasonal Difference		1			
	LTRKM	Petrol price	Numerator	Lag 0	,210	,134	1,561	,120
			Seasonal Difference		1			
	interv	Seat belt law Introduction	Numerator	Lag 0	-,163	,037	-4,464	,000
			Seasonal Difference		1			

Table 3.4.9: Estimation results for the ARIMA(2,0,0)(0,1,1)$_{12}$ model

Model	Ljung-Box Q(18)		
	Statistics	DF	Sig.
LDRIVERS-Model_1	23,289	15	,078

Table 3.4.10: Ljung-Box statistic for the residuals of the ARIMA(2,0,0)(0,1,1)$_{12}$) model

		Noise residual from LDRIVERS-Model_1
N		180
Normal Parameters(a,b)	Mean	,0048
	Std. Deviation	,07727
Most Extreme Differences	Absolute	,050
	Positive	,042
	Negative	-,050
Kolmogorov-Smirnov Z		,670
Asymp. Sig. (2-tailed)		,761

Table 3.4.11: Kolmogorov-Smirnov statistic for the residuals of the ARIMA(2,0,0)(0,1,1)$_{12}$ model

Fit Statistic	
Stationary R-squared	,590
R-squared	,802
RMSE	,079
MAPE	,860
MaxAPE	2,336
MAE	,064
MaxAE	,177
Normalized BIC	-4,881

Table 3.4.12: Goodness of fit criteria for the ARIMA(2,0,0)(0,1,1)$_{12}$ model

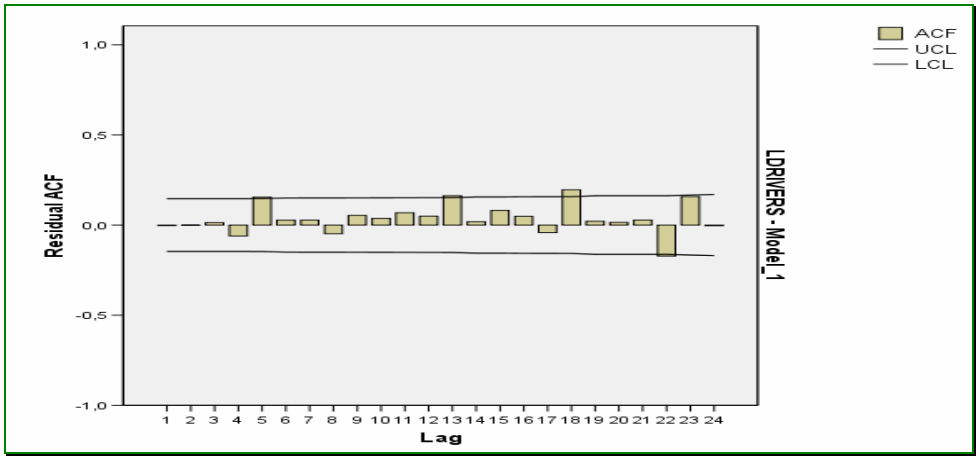


Figure 3.4.12: The ACF plot of the residuals and their confidence interval.

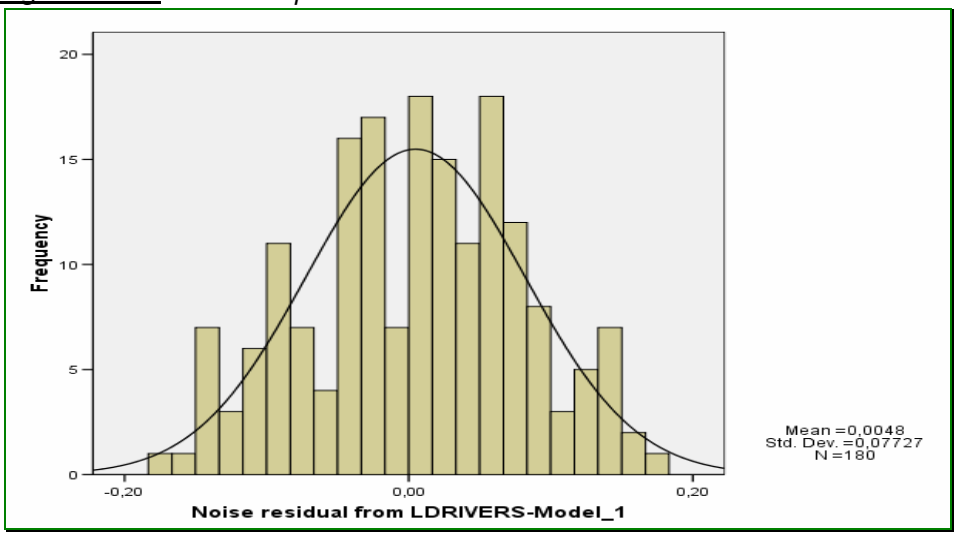


Figure 3.4.13 : The distribution of the residuals

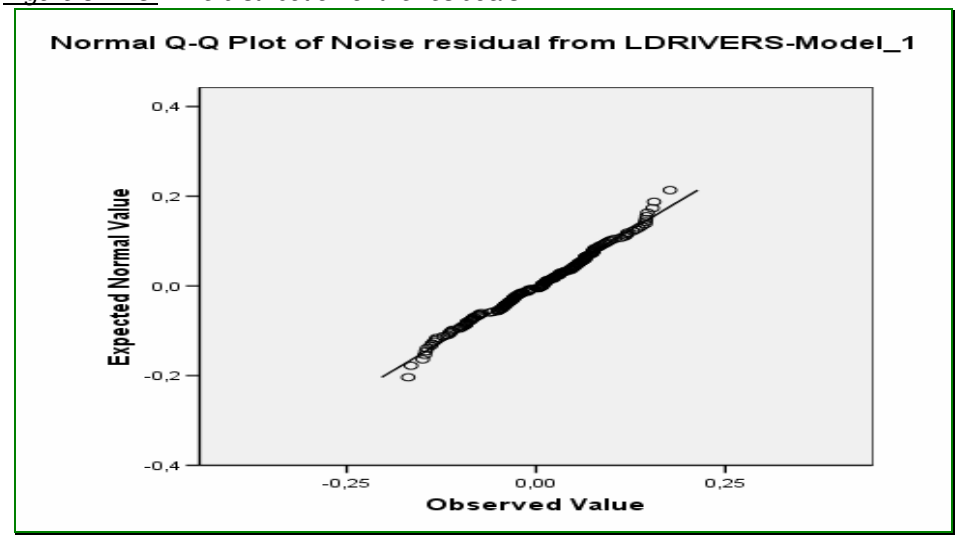


Figure 3.4.14: TheQQ-plot

3.4.5 ARIMA models for seasonal series (French injury accident and fatalities)

3.4.5.1. Objective of the technique

As in Section 3.4.4.1.

3.4.5.2. Model definition and assumptions

As in Section 3.4.4.2.

3.4.5.3. Research problem and dataset

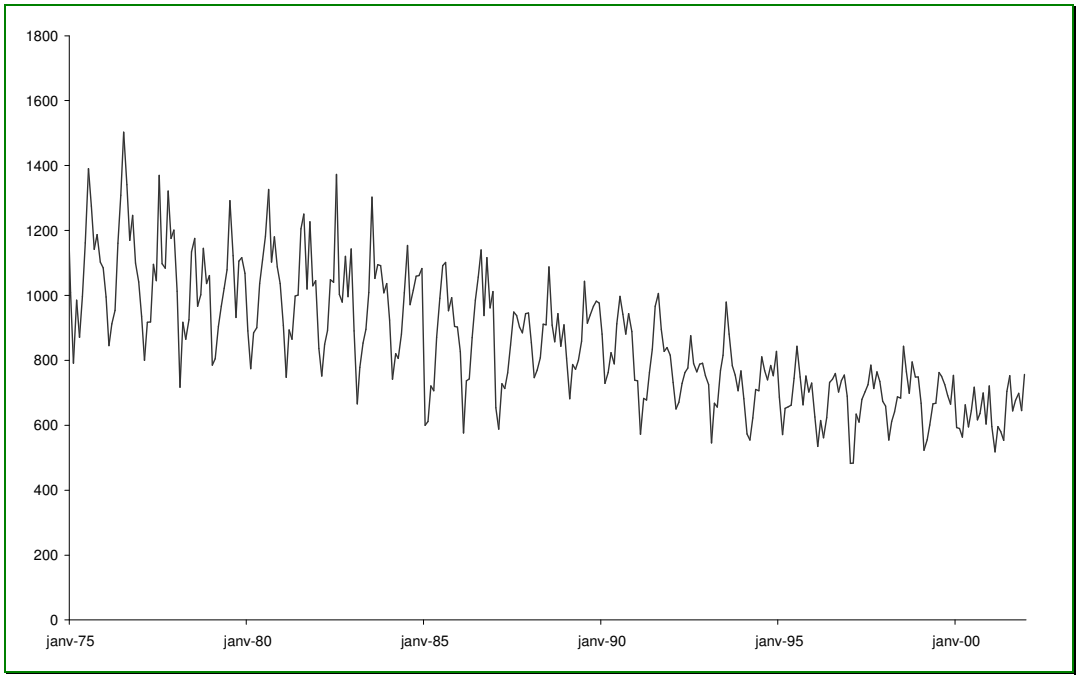


Figure 3.4.15: The aggregate number of fatalities in France, for 1975-2001.

In the road safety field in France, as already mentioned in Section 3.3, ARMA-type models were very often used on monthly aggregate data for assessing road safety measures (Lassare and al., 1993). We shall now describe an application of another ARMA-type model, based on monthly data over a period of 25 years, implemented to analyse the development of the aggregate number of fatalities in France. The purpose is to determine whether a relationship can be established between the amnesty of driving faults that traditionally accompanies the presidential election in France and the road safety level. The analysis presented here is limited to the statistics of fatalities, and to the two elections of 1988 and 1995 - for which the information was carried by the media.

The dataset is the monthly number of fatalities in France, for the period January 1975-December 2001, as presented in Figure 3.4.15.

Oil sales (gasoline and diesel) as a proxy for risk exposure (the total number of vehicle-kilometres is not measured on a monthly basis, in France), the car fuel price, and a small number of weather variables that take account for transitory effects (the highest temperature of the day, the rainfall height and the occurrence of frost, averaged or aggregated on the month) were used as exogenous variables in an ARIMA model.

Because of the purpose described above, three intervention variables were also constructed and the form of their intervention function then determined. This will be described precisely in more detail in the next-coming paragraphs.

3.4.5.4. Model fit

Regarding the application case, an intervention analysis is carried out, in order to determine whether the perspectives of the presidential amnesty of 1998, and of 1995, eventually had an effect on the development of the monthly number of fatalities.

This can be achieved in two stages:

- First by determining a period during which the perspectives of the presidential amnesty eventually had an impact on the drivers and policemen behaviour,
- Second by identifying the form of intensity of that impact with an intervention function.

The even nature of the presidential amnesty leads to delimit its impact in time (transitory effect). The two first intervention periods are, in a first approach, fixed as November 1987 - July 1988 and September 1994 – July 1995 (month of first announcement, last month before the amnesty law is voted). The form of the intervention function is then determined depending on the values of the monthly impacts of the dummy variables defined on the period (Box,Tiao, 1975),(Gourieroux and Monfort, 1990).

In addition, particularly low values of the number of fatalities were detected, between February 1987 and October 1987: the media effect of the Anne Cellier case (a young woman died in an accident, whereas the person responsible for the accident was drunk driving, and was only lightly condemned) followed by the introduction of a new law related to drink driving, certainly contributed to diminish accidents' gravity in France. Because of its proximity to the election of 1988, the "Cellier effect" was also modelled, and the period April - October 1987 also retained as a third intervention period, with here again the hypothesis of a limited effect in time.

In sum, three intervention variables were constructed, and for three predefined periods. In each of the three cases, the form of the intervention function still has to be determined.

The form of the three intervention functions has been established using the following model:

$$\Phi(B)(I-B12)\left[\log Y_t - \sum_{i=1}^{I}\alpha_i LogZ_{i,t} - \sum_{j=1}^{J}\beta_j Z_{j,t} - \sum_{k=1}^{3}\sum_{l=0}^{n_k}\delta_{l,k}P^{T_{0,k}}(t-l)\right] = \mu + \Theta(B)a_t \quad (3.4.9)$$

with:

Y the number of fatalities,

$X_{i,i=1toI}$ the I variables measuring risk exposure and the economic factors,

$Z_{j,j=1toJ}$ the J variables measuring the transitory factors,

$P^{T_{0,k}}$, k=1 to 3, three dummy variables given by $P^{T_{0,k}}(t)=1$ in t= $T_{0,k}$ and 0 elsewhere, $T_{0,k}$ the first month of the intervention period n° k,

n_{k+1} the number of months of the intervention period n° k,

$\Phi(B)$ and $\Theta(B)$, two polynomials of the delay operator B,

and a_t a white noise.

The forms suggested by the autoregressive polynomial $\sum_{l=0}^{n_k}\delta_{l,k}P^{T_{0,k}}(t-l)$ is a step[62] in all the three cases. The initial model (3.4.9) has therefore been simplified by using three variables representing steps:

$$\Phi(B)(I-B12)\left[\log Y_t - \sum_{i=1}^{I}\alpha_i LogZ_{i,t} - \sum_{j=1}^{J}\beta_j Z_{j,t} - \sum_{k=1}^{3}\gamma_k Step_{k,t}\right] = \mu + \Theta(B)a_t \quad (3.4.10)$$

with: $Step_{k,t}$, k=1 to 3, three dummy variables equal to 1 in $[T_{0,k}, T_{0,k}+n_k]$ and 0 elsewhere.

Finally, the model was still adjusted by allowing the beginning and the end of the two intervention periods corresponding to the presidential amnesties to vary, in order to maximise the likelihood of the model. As a consequence the second period was restricted to December 1994 - June 1995, while the first one remained unchanged.

The results obtained by estimating model (3.4.10) are given in Table 3.4.13.

3.4.5.5. Model diagnostics

All parameters related to the exogenous variables were kept in the model if significant at the 70% confidence level (T-ratio larger than 1).

[62] In all three cases, the intervention effect was assumed to be the constant every month inside the intervention period, and zero outside.

As for the dynamics' parameters, they were kept if significant at the usual 95% confidence level (T-ratio larger than 2).

The reason for keeping less significant variables, is that the best model - in terms of adjustment - , is obtained when all exogenous variables are kept, whether significant or not. This is equivalent to considering that each variable's contribution must be taken account for, in order to estimate in the best manner the effects of the perspectives of presidential amnesties, which remains the main objective. The main argument for reducing the number of exogenous variables to the most significant ones is to aim at the best model - in terms of forecasting -, which is not the objective here.

The values of the Ljung-Box and Kolmogorov-Smirnov statistics, given in Tables 3.4.14 and 3.4.15 lead to accept the non-correlation and normality of the residuals, which can therefore be considered as independent.

As for the model fit criteria given in Table 3.4.16, the stationary R-squared is 59,0% whereas the R-square reaches 91,6%; the mean absolute percentage error is only 0,75% , its highest value observed being 3,33 1% aver the 25 years,

3.4.5.6. Model interpretation

The parameters related to explanatory variables given in Table 3.4.13 appear to be acceptable.

Those related to climate are consistent with other results (Bergel, Depire, 2004). Rainfall height is linked, positively, to the total number of fatalities: an increase of 100 mm in the average rainfall height leads to an increase of 0,3% in this indicator. Temperature is also linked, positively, to the total number of fatalities: an increase of one degree in the average temperature in the month leads to an increase of 1% in the summer and 2% in the winter of the number of fatalities. On the contrary, no link was found between the occurrence of frost and the number of fatalities.

Only the elasticity value of the number of fatalities with respect to oil sales is small, around 0.1, and this is probably due to the presence of the other explanatory variables, correlated to oil sales.

The following comments focus on the intervention step variables.
Succeeding to a "Cellier effect" of -5,4 % per month (average decrease of 5,4 % in the number of fatalities between April and October 1987), the effect of the amnesty's perspectives of 1988 is estimated at +7,1% per month (average increase in the number of fatalities of 7,1% between November1987 and July1988), and the effect of 1995 is estimated at +4,2% per month (average increase of 4,2% in the number of fatalities per month between December 1994 and June 1995).

Measured in absolute number of deaths, the effects of both perspectives of amnesty are estimated at 565 and 202 additional fatalities respectively. The associated confidence levels are 0.036 and 0.215 respectively, which confirms that the effects of the first amnesty is the only significant one at the usual confidence level.

						Estimate	SE	t	Sig.
LTUEFE-Model_1	LTUEFE	French Fatalities	Constant			-,026	,002	-10,799	,000
			AR	Lag 1		,149	,059	2,536	,012
				Lag 2		,191	,059	3,248	,001
				Lag 3		,231	,060	3,822	,000
			Seasonal Difference			1			
			MA, Seasonal	Lag 1		,883	,045	19,477	,000
	LCARBUB	Oil Sales	Numerator	Lag 0		,096	,080	1,210	,227
			Seasonal Difference			1			
	LICARB	Petrol price	Numerator	Lag 0		-,012	,084	-,138	,890
			Seasonal Difference			1			
	TE	Summer temperature	Numerator	Lag 0		,001	,000	3,960	,000
			Seasonal Difference			1			
	TH	Winter temperature	Numerator	Lag 0		,002	,000	4,600	,000
			Seasonal Difference			1			
	HPLUI	Rainfall	Numerator	Lag 0		2,81E-005	1,29E-005	2,176	,030
			Seasonal Difference			1			
	NGEL	Frost	Numerator	Lag 0		,000	,002	-,208	,836
			Seasonal Difference			1			
	Step0	Cellier effect	Numerator	Lag 0		-,054	,035	-1,535	,126
			Seasonal Difference			1			
	Step1	1988 Amnisty	Numerator	Lag 0		,071	,033	2,106	,036
			Seasonal Difference			1			
	Step2	1995 Amnisty	Numerator	Lag 0		,042	,033	1,243	,215
			Seasonal Difference			1			

Table 3.4.13: Estimation results for the ARIMA(3,0,0)(0,1,1)$_{12}$ model

Model	Ljung-Box Q(18)		
	Statistics	DF	Sig.
LTUEFE-Model_1	31,570	14	,005

Table 3.4.14: Ljung-Box statistic for the residuals of the ARIMA(3,0,0)(0,1,1)$_{12}$) model

		Noise residual from LTUEFE-Model_1
N		300
Normal Parameters(a,b)	Mean	,0024
	Std. Deviation	,06353
Most Extreme Differences	Absolute	,030
	Positive	,024
	Negative	-,030
Kolmogorov-Smirnov Z		,519
Asymp. Sig. (2-tailed)		,950

Table 3.4.15: Kolmogorov-Smirnov statistic for the residuals of the ARIMA(3,0,0)(0,1,1)$_{12}$ model

Fit Statistic	
Stationary R-squared	,595
R-squared	,916
RMSE	,065
MAPE	,750
MaxAPE	3,331
MAE	,050
MaxAE	,206
Normalized BIC	-5,201

Table 3.4.16: Goodness of fit criteria for the ARIMA(3,0,0)(0,1,1)$_{12}$ model

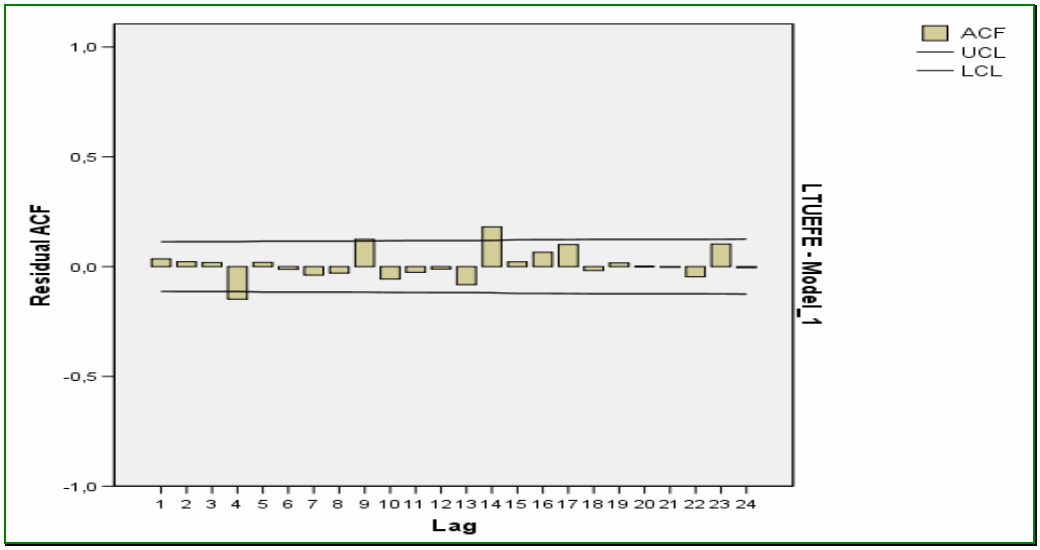


Figure 3.4.16: The ACF plot of the residuals and their confidence interval.

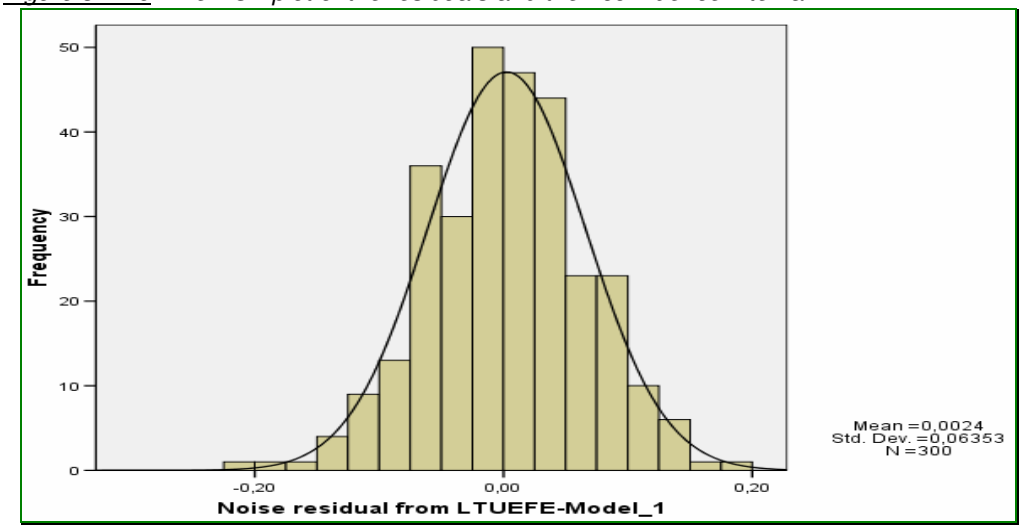


Figure 3.4.17 : The distribution of the residuals

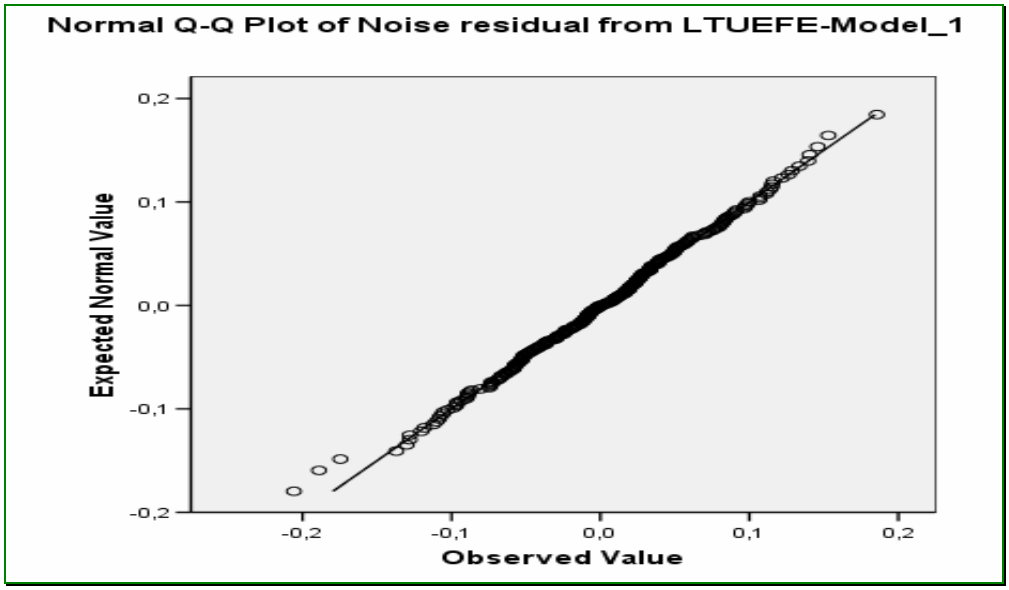


Figure 3.4.18: TheQQ-plot

3.4.5.7. Conclusion and similar results

In this application case, it was demonstrated that an ARIMA model with exogenous (explanatory and intervention) variables is an efficient tool for analysing the development of the aggregate number of injury accidents and fatalities in France, by taking account for risk exposure (measured with oil sales as a proxy of risk exposure) and transitory factors of climatic nature. The possible effects of two presidential amnesties of driving faults, in 1988 and in 1995, on the number of fatalities in France were questioned by the means of an intervention analysis.

The amplitude of the effects of the perspectives of amnesty of 1988 is larger (over 500[63] additional fatalities, between September 1987 and July 1988) than

[63] The annual number of fatalities in France was around a thousand in the years 1990.

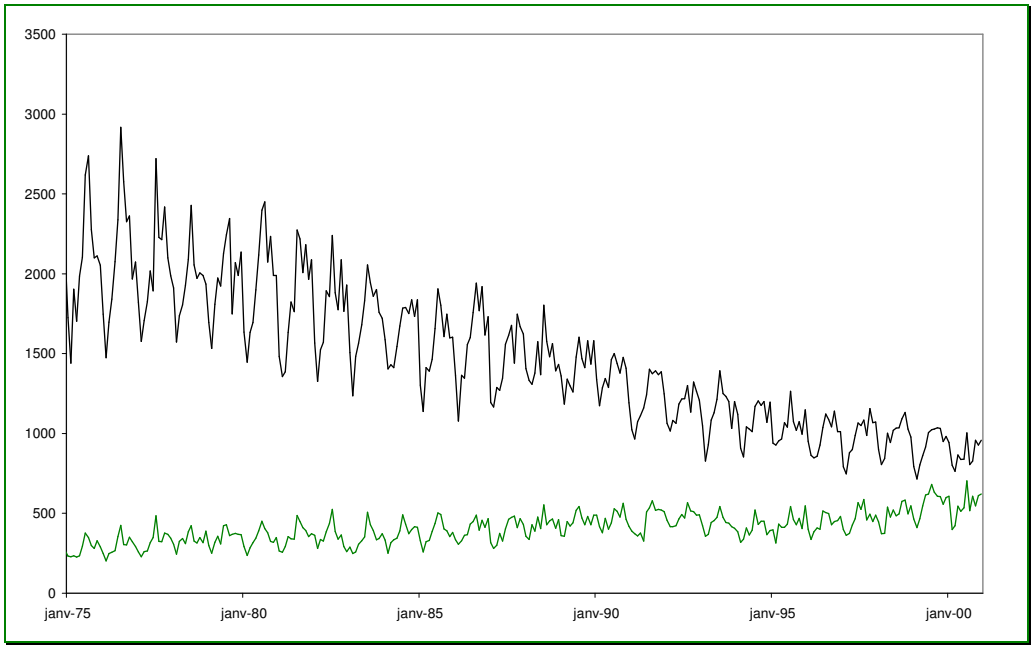


Figure 3.4.19: The number of injury accidents in France, on A-level roads and motorways, for 1975-2001.

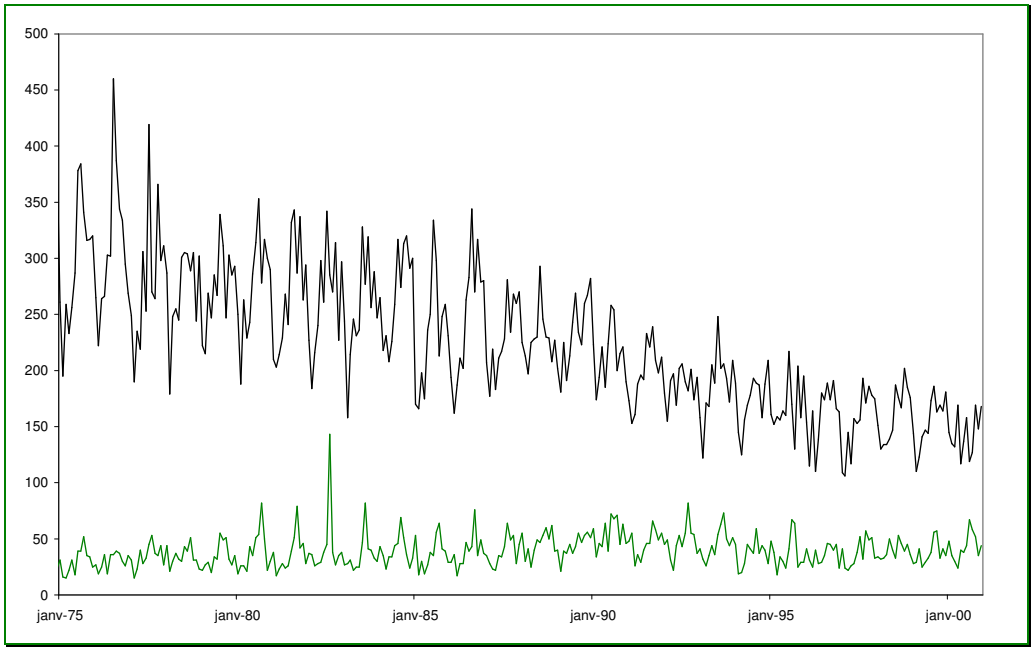


Figure 3.4.20: The number of fatalities in France, on A-level roads and motorways, for 1975-2001.

it is in for the amnesty of 1995 (around 200 additional fatalities, between December 1994 and June 1995).
The increase related to the presidential election of 1988 is the only one that is statistically significant, at the usual level - i.e. 565 additional fatalities, with a confidence level of 0,04.

This approach was extended and applied to other risk indicators, such as the number of injury accidents and fatalities, on A-level roads and on motorways (see

Figures 3.4.19 and 3.4.20). Similar results, given in Table 3.4.17, were obtained and confirmed the previous parameters' interpretations.

Thus, to the exception of one case, the elasticity value of the risk indicators with respect to the traffic volume is smaller than 1 (between 0,5 and 0,8) but much superior that the estimated elasticity value of the number of fatalities with respect to oil sales, given in 3.4.5.6.

The climate parameters are consistent with those estimated previously, and appear to be even larger. Thus, the rainfall height influence is generally larger on the disaggregate risk indicators, whereas the temperature effect is about the same. Note that the occurrence of frost comes out to be very significant in two cases, with a positive link between the number of days of frost in the month and the risk indicators.

As for the intervention step variables' parameters, a general result is that the effect of the perspectives of amnesty of 1988 is significant at the 70% confidence level, whatever the risk indicator, and is estimated at 5,9% and 8,2% per month regarding the number of injury accidents on A-level roads and motorways, and at 9% and 14% per month regarding the number of fatalities on A-level roads and motorways, between November1987 and July1988. These increase levels are higher than the increase level of the number of fatalities estimated on the whole territory, and the highest values are found on motorways.

3.4.6 Conclusion on ARMA-type models

As a general conclusion of the chapter, it will be recalled that ARMA-type models are very widely used for purposes of road safety research. The so-defined ARMA-type models include all the following cases: ARMA models in the stationary case, ARIMA models in the non-stationary case, ARMAX models in the case exogenous variables are used, and ARIMAX models in the non-stationary case and exogenous variables being used.

The use of transformations applied to the initial data, and the call to exogenous variables (whether pure explanatory variables or intervention variables) allows another process, derived from the initial one and corrected from exogenous effects, to be modelled with an ARMA model, as fulfilling the hypothesis of stationarity.

Two relevant features in all these models, related to the additional independent variables, are to be highlighted in this general conclusion: the higher capacity for the the model interpretation, and the gain in the model-fit.

3.4.6.1. Model interpretation

A summary of all parameters estimated with ARMA-type models fitted on real data, and described in this section, given in Tables 3.4.17a & b, enables to

conclude that, in addition to the dynamics-parameters, numerous exogenous effects - parameters appeared to be highly significant .

Whereas all parameters related to the dynamics were only kept if significant at the usual confidence level, the other parameters were kept even if significant at the 70% confidence level (t-value larger than 1).

The main results are the following:

- the risk exposure indicator was the most significant when measured with the number of vehicle-kilometers on disaggregated networks (the French motorways and A-level roads),
- the (petrol) price was the only price indicator which was found to be significant (in the case of the UK-KSI drivers),
- the climatic variables, happened to have distinct effects, at the aggregate level and on disaggregated networks (in the case of the French fatalities)
- the intervention variables, which were significant at an aggregate level in both cases of the UK-KSI drivers and French fatalities) were less significant on disaggregated networks (the French motorways and A-level roads).

3.4.6.2. Model fit

Second, a summary of the goodness of fit criteria, given in Table 3.4.18, leads to conclude that the introduction of exogenous variables in the pure ARIMA models enabled the part of variance explained by the model to increase significantly (between 2,1% and 24% according to the indicator) and the absolute error made, measured in mean over the period and in percentage, to decrease significantly (between 4,4% and 11,9% respectively). Nevertheless, the normalized BIC decreased less significantly, and even happened to increase (varying between -0,5% and +1,3%), and this is due to the fact that this criteria is meant to take account of the parsimony of the model.

	Traffic volume	Price	Summer temp	Winter temp	Rainfall height	Occurrence of Frost	Interv. Var. 1	Interv. Var. 2	Interv. Var. 3
Norwegian fatalities									
UK-KSI drivers									
							-0,184 (***)		
	0,21 (**)	_0,297 (***)					-0,163 (***)		
French fatalities									
	0,096 (**)	-0,012 (*)	0,001 (***)	0,002 (***)	2,81E-05 (***)	0 (*)	-0,054 (**)	0,071 (***)	0,042 (**)
French injury accidents on motorways									
	0,765 (***)		0,002 (***)	0,001 (**)	8,76E-05 (***)	0,007 (***)	-0,025 (*)	0,078 (**)	-0,039 (*)
French injury accidents on A-level roads									
	0,526 (***)		0 (*)	-4,19E-05 (*)	6,07E-05 (***)	-0,001 (*)	-0,036 (**)	0,057 (**)	0,007 (*)
French fatalities on motorways									
	1,788 (***)		0,001 (*)	0,002 (**)	1,73E-05 (*)	0,012 (**)	-0,044 (*)	0,145 (**)	-0,105 (**)
French fatalities on A-level roads									
	0,598 (***)		0,001 (**)	0,001 (**)	8,38E-05 (***)	0,004 (**)	-0,054 (*)	0,09 (**)	0,086 (**)

Tables 3.4.17a & b : The exogenous and dynamics parameters - Summary
(*) T-value smaller than 1, (**) T-value between 1 and 2, (***) T-value larger than 1.

	phi1	phi2	phi3	Theta1	Theta12	Mu
Norwegian fatalities						
				-0,432		-0,02
				(***)		(**)
UK-KSI drivers						
	0,429	0,298			-0,898	-0,018
	(***)	(***)			(***)	(***)
	0,378	0,279			-0,889	-0,01
	(***)	(***)			(***)	(***)
	0,283	0,235			-857	-0,015
	(***)	(***)			(***)	(***)
French fatalities						
	0,264	0,187	0,064		-0,907	-0,022
	(***)	(***)	(**)		(***)	(***)
	0,149	0,191	0,231		-0,883	-0,026
	(***)	(***)	(***)		(***)	(***)
French injury accidents on motorways						
	0,328	0,262			-0,841	0,023
	(***)	(***)			(***)	(***)
	0,339	0,259			-0,845	-0,023
	(***)	(***)			(***)	(***)
French injury accidents on A-level roads						
	0,337	0,192			-0,831	-0,036
	(***)	(***)			(***)	(***)
	0,341	0,225			-0,837	-0,046
	(***)	(***)			(***)	(***)
French fatalities on motorways						
					-0,794	0,01
					(***)	(***)
					-0,932	-0,096
					(***)	(***)
French fatalities on A-level roads						
	0,158	0,226	0,146		-0,917	-0,03
	(***)	(***)	(***)		(***)	(***)
	0,103	0,274	0,212		-0,94	-0,042

	(***)	(***)	(***)	(***)	(***)

	R2	BIC	MAPE
Norwegian fatalities			
ARIMA model	0,789	-4,413	1,362
UK-KSI drivers			
ARIMA model	0,77	-4,85	0,9
with intervention variables	0,788	-4,886	0,887
with intervention and explanatory variables	0,802	-4,881	0,86
Gain in the model fit	***4,2%***	***-0,6%***	***-4,4%***
French fatalities			
ARIMA model	0,891	-5,145	0,795
with intervention and explanatory variables	0,916	-5,201	0,75
Gain in the model fit	***2,8%***	***-1,1%***	***-5,7%***
French injury accidents on motorways			
ARIMA model	0,813	-4,557	1,311
with intervention and explanatory variables	0,849	-4,591	1,155
Gain in the model fit	***4,4%***	***-0,7%***	***-11,9%***
French injury accidents on A-level roads			
ARIMA model	0,95	-5,319	0,745
with intervention and explanatory variables	0,96	-5,344	0,668
Gain in the model fit	***1,1%***	***-0,5%***	***-10,3%***
French fatalities on motorways			
ARIMA model	0,375	-2,595	5,982
with intervention and explanatory variables	0,465	-2,568	5,486
Gain in the model fit	***24,0%***	***1,1%***	***-8,3%***
French fatalities on A-level roads			
ARIMA model	0,846	-4,326	1,63
with intervention and explanatory variables	0,864	-4,269	1,534
Gain in the model fit	***2,1%***	***1,3%***	***-5,9%***

Table 3.4.18: Goodness of fit criteria - Summary

3.5 DRAG models

Ruth Bergel (INRETS)

In this section, we address the three-level explanatory model constructed on a monthly basis, proposed by Gaudry (1984), the DRAG-model (Demand for Road use, Accidents and their Gravity). As it will be seen now, that ARMA-type model constitutes in itself an application to the road safety field. Apart from the strict statistical aspects, the technique cannot be described without referring to the road safety methodological framework, described in 3.2.1.

3.5.1 Objective of the technique

The main objective of the DRAG approach is to model, altogether and at an aggregate level, *several levels of risk*, as described in 3.2.1.
As the model is meant to be a *comprehensive (explanatory)* model, it aims at taking account for numerous risk factors, and at measuring their influence on pre-defined risk indicators.
The advantage of the technique, compared to a multiple linear regression, is that the use of the Box-Cox transformation for all data allows for *a more flexible form* (linear form, logarithmic form or a compromise) of the link between the endogenous variable and each of the exogenous variables.

3.5.2 Model definition and assumptions

3.5.2.1. Economic formulation

Summarising the preceding description of the technique, a DRAG-model can shortly be defined on the basis of the following three criteria:

- to model (at least) the three following levels : road demand, risk's accident and accident's gravity,
- to be explanatory,
- to rely on a flexible functional form.

The general and precise framework of the DRAG approach is well defined in (Gaudry, Lassarre, 2000). In this framework, one demand level (the exposure to risk) and two risk levels (the risk of accident and the risk of being victim in an accident) are defined, as well as indicators and factors at each of these levels.

Numerous explanatory variables are introduced, related to exposure, economic factors, transitory factors, behavioural factors and road safety measures. By modelling road demand (exposure to risk), and the two risk levels with the same explanatory factors, it is possible to quantify the direct and indirect effects - via the traffic volume - on the two types of risk indicators.

It is worth noting here that the human behaviour, measured with the practised speed, is also modelled as an additional level in the TAG-1 model for France, but this four-level approach is not generalized within the DRAG-family models yet.

3.5.2.2. Econometric specification

Let us first recall that the Box-Cox transformation, which is used in the econometric specification of the DRAG-model, is defined as a power transformation, of parameter λ, on any positive real variable V_t by:

$$\begin{aligned} V_t \rightarrow V_t^{(\lambda)} &= \frac{V_t^{\lambda} - 1}{\lambda} \; if \; \lambda \neq 0 \\ V_t^{(0)} &= Log \; V_t \end{aligned} \tag{3.38}$$

The DRAG-model relies on a multiple regression structure with auto correlated and heteroscedastic errors, and takes account for a type of non-linearity. The fact that many explanatory variables are introduced allows the trend and the seasonal component to be modelled, which thus do not need to be filtered. The use of the Box-Cox transformation allows a more flexible form (linear form, logarithmic form or a compromise) of the link between the endogenous variable and each of the exogenous variables.

The model is written as follows:

$$\begin{cases} Y_t^{(\lambda_Y)} &= \sum_{k=1}^{K} \beta_k \; X_{kt}^{(\lambda_{X_k})} + u_t \\ u_t &= v_t \sqrt{\exp\left(\sum \delta_m Z_{mt}^{(\lambda_{Z_m})} \right)} \\ v_t &= \sum_{l=1}^{p} \rho_l \; v_{t-l} + w_t \end{cases} \tag{3.39}$$

with: Y_t the endogenous variable to be modelled,
X_{kt}, k=1 to K, the exogenous (or explanatory) variables,
u_t the first residual, and v_t the final residual,
w_t a white noise.

In that general formulation, the Box-Cox parameters λ_Y, λ_{X_1},.. λ_{X_K} are estimated in addition to the other parameters β_k, δ_m and ρ_l, for k=1 to K, m=1 to M and l=1 to L.

In practice, all parameters are not estimated, and some of them may be fixed to 0 or to 1, for specific reasons. Two well-known particular cases are obtained when the parameter λ is identically equal to 0 (we then have the log-log specification), or to 1 (we then have the linear specification).

3.5.2.3. Assumptions

The main assumption is that the endogenous variable is supposed to be Gaussian (as the observed data are aggregate, their frequency is easily larger than 30).

The assumption of stationarity of the process Y_y is not required. The explanatory variables take account for trend and seasonality of the transformed process $Y_t^{(\lambda_Y)}$, whereas heteroscedasticity on the first residual u_t is also modelled separately, in such a way that the final residual V_t is supposed to be stationary.

3.5.3 Research problem and data set

Six DRAG models have already been constructed on aggregate data, whether at a national (Germany, Norway, France), regional (Quebec, California) or at an urban (Stockholm) level. Their latest versions available are the following ones (Gaudry, Lassarre, 2000):

The DRAG-2 model for Quebec
The SNUS 2-5 model for Germany
The TRULS-1 model for Norway
The STOCKHOLM-2 model for the city of Stockholm
The TAG-1 model for France
The TRACS-CA model for California

No condition is required from the data, but the constitution of a voluminous database covering a long-time period requires time and financial support.

Nevertheless, a major difficulty of the DRAG approach lies in modelling the first level of road demand - *the monthly number of vehicle-kilometres driven on the defined aggregate network.* The monthly data to be modelled may not be available over a long time period or may not be measured at all, and therefore need to be estimated first. This can be achieved by means of modelling, or by other means (Yannis et al., 2005). This preliminary step - estimating unknown numbers of vehicle-kilometres, on a monthly basis and over a long-time period - is a source of additional error in the global model.

On French data for instance, a DRAG-type model was applied to the French main road network (A-level roads and motorways, the two networks on which the number of vehicle-kilometres driven are measured on a monthly basis).

3.5.4 Model fit and diagnostics

The model fit is performed with the TRIO program, all the parameters - linear and non-linear - being estimated simultaneously; the usual statistical tests and criteria being also computed by the program. It is worth mentioning that no other existing softwares, the SAS system for instance, allow estimating the parameters of the linear and non-linear parts of the DRAG-model simultaneously.

3.5.5 Model interpretation

3.5.5.1. Multicolinearity

Multicolinearity between the numerous explanatory variables is a source of difficulties in interpreting the estimated parameters related to the explanatory variables.

3.5.5.2. Box-Cox parameters

In some cases, the Box-Cox parameters may not be stable and interpretable either[64], and the model's specification seems to be over-parameterised. A general important question is to determine whether the estimated values of the Box-Cox parameters significantly differ from 0 and from 1. If it is not the case, the related Box-Cox parameter should be fixed to 0 or to 1 instead of being estimated, which may lead to diminish the total number of parameters of the model in an important manner.

3.5.5.3. Elasticity values

Most of the estimated parameters are not interpreted directly: *elasticity values* are computed, of the endogenous variables with respect to the exogenous variables - that is to say of risk indicators with respect to risk factors. These elasticity values, calculated at a country's level independently of the units of measure of risk indicators and risk factors, are used for international comparisons.

3.5.5.4. International comparisons

Detailed interpretations of elasticity values of risk indicators with respect to risk factors, as well as evaluations of the major road safety measures that appear to be significant at an aggregate level, can be found in (Gaudry, Lassarre, 2000).

3.5.6 Conclusion

Because of the need of a voluminous database for estimating a DRAG model, the DRAG approach can not be achieved without enough time and financial support, and it would not be feasible to apply it to European data within the SafetyNet project.
In some cases where the monthly number of vehicle-kilometres is not available on the defined aggregate network and over a long period, the constitution of the first level model - the road demand model - may be the real difficulty.

[64] *In the case of the RES Model, an analysis of the advantage of the Box-Cox transformation was produced for this application (Bergel, Depire, 2004).The Box-Cox transformation was retained for the main exogenous variable, whereas the logarithmic transformation was retained for the endogenous variable. Tests of comparison of the initial specification with two particular cases were carried out. No significant difference could be found between the model with the Box-Cox transformation on the main exogenous variable and the model with the logarithmic transformation on the main exogenous variable, which indicates that the second specification, widely used, can be preferred for reasons of parsimony. Nevertheless, the use of the optimal functional form permits to relax the hypothesis of a constant elasticity to the traffic, and to take account for certain saturation effects with regard to the traffic.*

Nevertheless, the underlying theoretical framework is powerful, and is used for time series analysis in road safety research purposes far beyond the application of the DRAG-approach itself.

3.6 State space models

Jacques Commandeur and Chris de Blois (SWOV)

This section presents the subclass of state space methods collectively known in the literature as *structural time series models* or *unobserved components models*. Important references in this field are Harvey (1989), and Durbin and Koopman (2001). In structural time series models, an observed time series is typically decomposed into a number of *components*. The state of a structural time series model may consist of several components, which will be introduced one by one in the following sections.

First, in Sections 3.6.1, 3.6.2, and 3.6.3, those components are addressed that are useful for obtaining an adequate *description* of an observed time series. These components are the level, the slope and the seasonal. Then, in Sections 3.6.4 and 3.6.5, components of the state are presented that are helpful in finding *explanations* for the observed development in the series. These components are intervention and explanatory variables. A third important application of structural time series models is the ability to *predict* or *forecast* further developments of a series into the (unknown) future. This aspect of structural time series models is presented in Section 3.6.6. Finally, throughout these models will be compared with their equivalent in terms of classical linear regression models. These comparisons are particularly easy to make because, as will become clear below, classical regression models are easily fitted in the framework of structural time series analysis, and are in fact just a subclass of these models.

All the analyses presented below were performed with SsfPack (Koopman, Shepard and Doornik (1999)), which is a set of C routines collected in a library that can be linked to the Ox matrix programming language of Doornik (2001). The next section starts the presentation of models with the most simple structural time series model: the local level model.

3.6.1 Local level model

3.6.1.1. Objective of the technique

The objective of the local level model is to establish whether an observed time series can be adequately described with a time-varying level component.

3.6.1.2. Model definition and assumptions

The local level model is defined as

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \\ \mu_{t+1} &= \mu_t + \xi_t, \qquad \xi_t \sim NID(0, \sigma_\xi^2) \end{aligned} \tag{3.6.1}$$

for $t = 1, \ldots, n$, where μ_t is the unobserved *level* at time t, ε_t is the observation error or disturbance at time t, and ξ_t is the *level error* or *disturbance* at time t. In the literature on state space models, the observation disturbances ε_t are also referred to as the *irregular component*. The first equation in (3.40) is called the *observation* or *measurement equation*, while the second equation is called the *state equation*.

The level μ_t in model (3.6.1) can be conceived of as the equivalent of the intercept a in classical linear regression (see Section 3.3.1). Just as the intercept of a regression line determines the "height" or level of the regression line, so does the level determine the "height" of the state in state space modelling. The important difference is that the "height" of a regression line is fixed (i.e. constant over time), whereas the "height" of the state in the local level model is allowed to change from time point to time point.

As the measurement equation in (3.6.1) shows, with this model the observed time series is effectively decomposed into *two* components: the level component μ_t, and the irregular component ε_t.

In definition (3.6.1) the assumptions of the local level model are given algebraically by $\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$ and $\xi_t \sim NID(0, \sigma_\xi^2)$, where NID is a short-hand for Normally and Independently Distributed. The observation and level disturbances ε_t and ξ_t are therefore all assumed to be mutually independent, and normally distributed with zero means, and variances equal to σ_ε^2 and σ_ξ^2, respectively.

3.6.1.3. Dataset and research problem

In general, the dataset in an analysis with the local level model simply consists of only one variable: a time series y_t consisting of observations made sequentially through time points $t = 1, \ldots, n$.

The remaining part of this section first discusses and illustrates what happens when the level disturbances ξ_t in (3.6.1) are all fixed on zero, and then shows the effect of letting the level vary over time. In both cases, the same time series will be used as already presented in Section 1.2.2: the log of the annual number of road fatalities as observed in Norway for the period 1970-2003. As already mentioned in Section 1.2.2, the reason that the analysis is applied to the log of the fatalities is that the numbers of fatalities themselves are non-negative count data, meaning that the predicted values obtained with a time series analysis should also be non-negative. This is achieved by analysing count data in their logarithm, and parallels the use of the log link for count data in generalised linear models (see Section 3.3.2).

The research problem addressed with the local level model is how to obtain an adequate description of the log of the observed annual number of road fatalities in Norway in the period 1970-2003.

3.6.1.4. Model fit, diagnostics, and interpretation of results

If the level disturbances ξ_t in (3.6.2) are all fixed on zero (or, equivalently, the level disturbance variance σ_ξ^2 is fixed on zero), then it is not very difficult to show that the local level model simplifies into

$$y_t = \mu_1 + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \tag{3.6.2}$$

for $t = 1, \ldots, n$. Therefore, in this special situation everything hinges on the value of μ_1, which is the value of the level right at the beginning of the time series. Once this value is established, it remains constant throughout the remainder of the series. In this situation the level is said to be treated *deterministically*. When the level is allowed to vary over time, on the other hand, it is said to be treated *stochastically*.

Generally, in state space models the value of the unobserved state at the beginning of the time series (i.e., at $t = 1$) is unknown. There are two ways to deal with this problem. Either the researcher provides the first value, based on theoretical considerations, or some previous research, for example. Or this very first value is *estimated* by the very same procedure that is used to fit the state space model at hand. Since nothing is usually known about the initial value of the state, the second approach is most often followed in practice, and will be used in all further structural time series analyses discussed in the present report. In state space modelling, the second approach is called *diffuse initialisation*.

It can be proved that the best estimates for μ_1 and σ_ε^2 in model (3.6.2) are

$$\hat{\mu}_1 = \bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t \tag{3.6.3}$$

and

$$\hat{\sigma}_\varepsilon^2 = s_y^2 = \frac{\sum_{t=1}^{n}(y_t - \bar{y})^2}{n-1} \tag{3.6.4}$$

respectively. This extremely simple structural time series model thus actually computes the mean and variance of the observed time series, and the best fitting model for (3.6.2) is simply

$$\hat{y}_t = \bar{y} + (y_t - \bar{y}). \tag{3.6.5}$$

Applying deterministic level model (3.6.2) to the log of the annual number of road traffic fatalities in Norway for the period 1970 through 2003, yields

$$y_t = 5.9323 + \varepsilon_t,$$

with $\hat{\sigma}_\varepsilon^2 = 0.0485829$. Thus the mean of this series is 5.9323, and its variance equals 0.0485829. For these parameter estimates, the value of the log-likelihood function that is maximised in state space methods equals 0.038701012.

The level for model (3.6.2) is displayed at the top of Figure 3.6.1, together with the observed time series. As the figure illustrates, the deterministic level is indeed a constant, which does not vary over time.

The bottom graph in Figure 3.6.1 contains a plot of the observation disturbances ε_t corresponding to the deterministic level model. As the latter graph shows, the disturbances ε_t of the deterministic level model are not independently distributed at all, but follow a very systematic pattern. In fact, the irregular component in Figure 3.6.1 simply consists of the deviations of the observed time series from its mean, as already implied by (3.6.5).

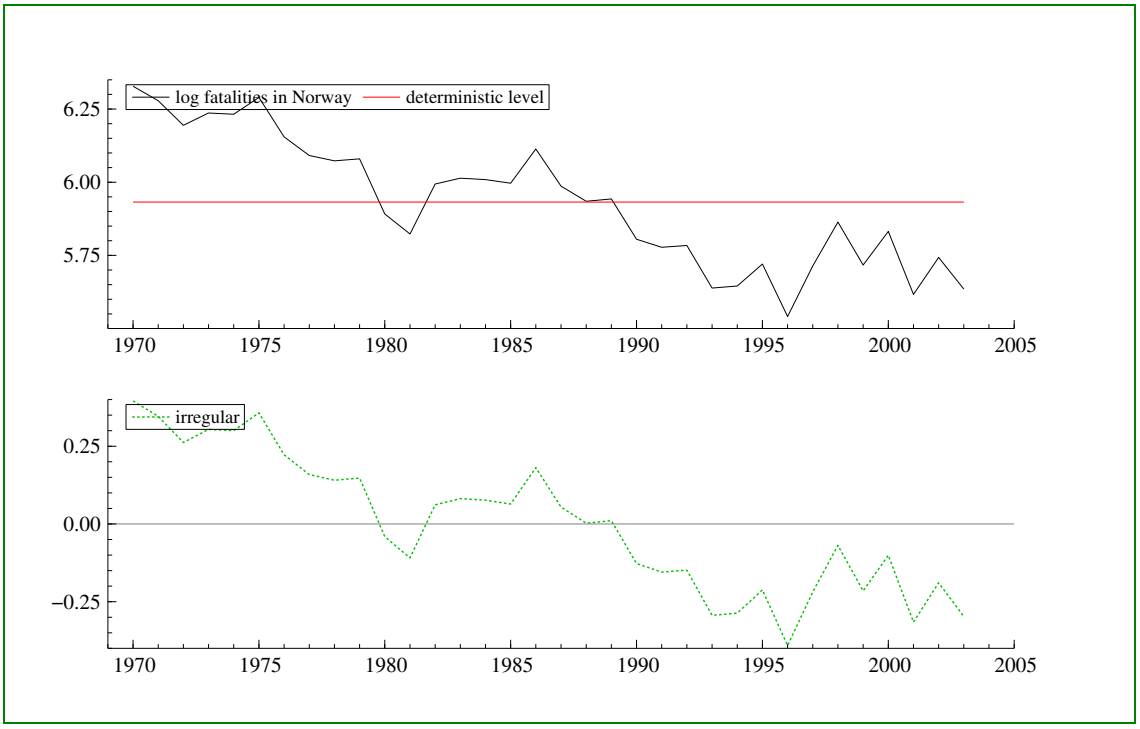


Figure 3.6.1: Deterministic level and irregular component for the log of Norwegian fatalities.

Diagnostic tests for the assumptions of independence, homoscedasticity, and normality of the residuals of the analysis are presented in Table 3.6.1. For the exact definition, computation and interpretation of these diagnostic tests the reader is referred to Section 3.3.1.

The value of the autocorrelation at lag 1, which is $r(1) = 0.588$, exceeds the 95% confidence limits of $\pm 2/\sqrt{n} = \pm 2/\sqrt{34} = \pm 0.343$ for this time series. The high amount of dependency between the residuals is also confirmed by the very large value of the *Q*-test in Table 3.16. Since $Q(10) = 29.259$ and because this value is much larger than the critical value of $\chi^2_{(10;0.05)} = 16.92$ (see Table 3.6.1), evaluated as a whole the first ten autocorrelations significantly deviate from zero, meaning that the null hypothesis of independence of the residuals must be rejected.

The two-tailed *H*-statistic in Table 3.6.1 shows that the variance of the first 11 elements of the residuals is unequal to the variance of the last 11 elements of the residuals, because $H(11) = 3.661$ is larger than the critical value of $F_{(11,11;0.025)} \approx 3.28$. This means that the assumption of homoscedasticity of the residuals is also not satisfied in the present analysis.

	statistic	value	critical value	assumption satisfied
independence	Q(10)	29.259	16.92	-
	r(1)	0.588	0.34	-
	r(4)	0.178	0.34	+
homoscedasticity	H(11)	3.661	3.28	-
normality	N	1.241	5.99	+

Table 3.6.1: Diagnostic tests for deterministic level model and log of Norwegian fatalities.

Finally, since N = 1.241 is smaller than the critical value of $\chi^2_{(2;0.05)} = 5.99$ (see Table 3.16), the null hypothesis of normally distributed residuals is not rejected.

Summarising, for the log of Norwegian fatalities series the residuals of the deterministic level model neither satisfy the assumption of independence nor that of homoscedasticity; only the least important assumption of normality is not violated.

In order to compare the different state space models, throughout Section 3.6 the Akaike Information Criterion (AIC) will be used:

$$\text{AIC} = \frac{1}{n}\left[-2n\log L_d + 2(q+w)\right], \tag{3.6.6}$$

where n is the number of observations in the time series, $\log L_d$ is the value of the diffuse log-likelihood function that is maximised in state space modelling, q is the number of initial values in the state, and w is the total number of disturbance variances estimated in the analysis. When comparing different models with the AIC, the following rule holds: smaller values denote better fitting models than larger ones. Compared with the more simple maximum log-likelihood criterion, a very useful property of the AIC criterion is that it compensates for the number of estimated parameters in a model, thus allowing for a fair comparison between models involving different numbers of parameters.

In the deterministic level model (3.6.2) only one variance is estimated (σ_ε^2) and one initial value (μ_1). Therefore, the Akaike information criterion for the analysis of the log of the number of Norwegian fatalities with the deterministic level model equals

$$\text{AIC} = \frac{1}{34}\left[-2(34)(0.0387010112) + 2(1+1)\right] = 0.040245.$$

Below, this value will be used for purposes of comparison with other state space models.

On the other hand, when the level in (3.6.1) is allowed to vary over time the following results are obtained. For the log of the annual number of Norwegian

fatalities series, the maximum likelihood estimates of the disturbance variances are $\sigma_\varepsilon^2 = 0.00326838$ and $\sigma_\xi^2 = 0.0047026$, respectively. For these parameter estimates, the value of the log-likelihood function equals 0.84686222.

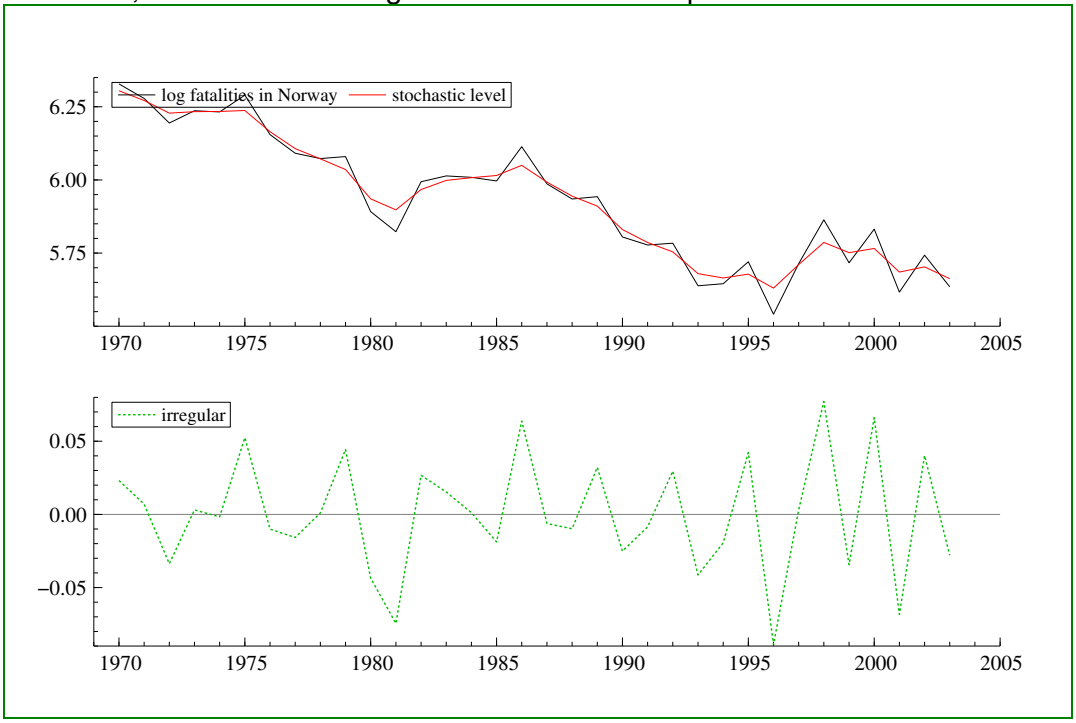


Figure 3.6.2: Stochastic level and irregular component for the log of Norwegian fatalities.

The local level for model (3.6.1) is illustrated at the top of Figure 3.6.2, together with the observed time series. As can be seen in Figure 3.6.2, when the level is allowed to vary over time, the observed time series is recovered quite well.

	statistic	Value	critical value	assumption satisfied
independence	Q(10)	6.228	16.92	+
	r(1)	-0.127	0.34	+
	r(4)	-0.105	0.34	+
homoscedasticity	1/H(11)	1.746	3.28	+
normality	N	1.191	5.99	+

Table 3.6.2: Diagnostic tests for local level model and Norwegian fatalities

The irregular component of the local level model applied to the log of Norwegian fatalities is displayed at the bottom of Figure 3.6.2. The diagnostic tests for these observation disturbances are given in Table 3.6.2. In contrast with the deterministic level model, the observation disturbances of the local level model satisfy all of the distributional assumptions for this model: independence, homoscedasticity, and normality.

The disturbance variances of a state space model are often called *hyper-parameters*. Since the local level model requires the estimation of two hyper-parameters (σ_ε^2 and σ_ξ^2), and of one initial value (μ_1), the Akaike information criterion for this analysis equals

$$\text{AIC} = \frac{1}{34}\left[-2(34)(0.8468622) + 2(1+2)\right] = \text{-}1.51725.$$

which is a clear improvement upon the deterministic level model applied to these data, since the AIC value for the latter model was 0.040245. It may be noted that the addition of a slope component (see Section 3.6.2) to model (3.6.1) does not improve the description of the time series, since this results in an AIC value of only -1.28035.

A time varying level suffices to provide a good description of the development in the log of the annual road traffic fatalities in Norway for the period 1970 through 2003, yielding residuals that satisfy all the model assumptions.

3.6.1.5. Conclusion on the technique

The analysis of a time series with the deterministic level model is identical to a classical regression analysis with only an intercept in the regression equation. In fact, it is simply a horizontal line through the mean value of a series. As the analysis in this section showed, making the level component stochastic can be sufficient to adequately describe a time series.

3.6.2 Local linear trend model

This section discusses the effects of adding a new component to the local level model, called the *slope* component.

3.6.2.1. Objective of the technique

The objective of the local linear trend model is to establish whether an observed time series can be described with a trend consisting of a time-varying level and a time-varying slope component.

3.6.2.2. Model definition and assumptions

The local linear trend model is obtained by adding a slope component ν_t to the local level model, and is defined as follows:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim NID(0,\sigma_\varepsilon^2) \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim NID(0,\sigma_\xi^2) \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\sim NID(0,\sigma_\zeta^2) \end{aligned} \qquad (3.6.7)$$

for $t = 1, \ldots, n$. The local linear trend model therefore contains *two* state equations: one for modelling the level, and one for modelling the slope. The slope ν_t in (3.6.7) can be conceived of as the equivalent of the regression coefficient b in the simple classical regression model of y_t on time (see also Section 2.2.3.1). Just as the value of b determines the angle of the regression line with the x-axis, so does the slope determine the angle of the trend with the x-axis in state space modelling. Again, the important difference is that the regression coefficient or weight b is fixed in classical regression, whereas the slope in (3.6.7) is allowed to change over time.

The *assumptions* of the local linear trend model (3.6.7) are that the observation, level, and slope disturbances ε_t, ξ_t, and ζ_t are all mutually independent, and normally distributed with zero means, and variances equal to σ_ε^2, σ_ξ^2, and σ_ζ^2, respectively.

3.6.2.3. Dataset and research problem

In general, the dataset in an analysis with the local linear trend model again simply consists of only one variable: a time series y_t consisting of observations made sequentially through time points $t = 1, \ldots, n$.

The remaining part of this section will first discuss and illustrate the effect of fixing all state disturbances ξ_t and ζ_t in (3.6.7) on zero, and then present the effect of allowing the level and slope components to vary over time. In both cases, the

model will be applied to the log of the number of fatalities as observed in Finland for the period 1970 through 2003.

The research problem addressed with this model is how to obtain an appropriate description of the log of the observed number of fatalities in Finland during the period 1970-2003.

3.6.2.4. Model fit, diagnostics, and interpretation of results

Fixing all state disturbances ξ_t and ς_t in (3.6.7) on zero, that is, not allowing the level and slope component to vary over time, it is not too difficult to verify that the linear trend model simplifies into

$$y_t = \mu_1 + \nu_1(t-1) + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \tag{3.6.8}$$

for $t = 1, \ldots, n$, where the independent or predictor variable $(t-1) = 0, 1, \ldots, n-1$ is time itself, and μ_1 and ν_1 are the initial values of the level and the slope components, respectively.

Applying the deterministic level and slope model (3.6.8) to the log of the logarithm of the annual number of road traffic fatalities in Finland for the period 1970 through 2003, it is found that $\hat{\mu}_1 = 6.8717$, $\hat{\nu}_1 = -0.028733$, and therefore

$$y_t = 6.8717 - 0.028733(t-1) + \varepsilon_t$$

with $\sigma_\varepsilon^2 = 0.0213603$. For these maximum likelihood estimates, the value of the log-likelihood function is 0.3036367. The latter regression equation can also be written as

$$y_t = 6.8717 - 0.028733t + 0.028733 + \varepsilon_t = 6.9004 - 0.028733t + \varepsilon_t .$$

This is exactly the same result as a classical linear regression of the log of the Finnish fatalities on time $t = 1, \ldots, n$. Thus, treating the level and the slope components of the local linear trend model deterministically is the same as performing a linear regression of the dependent variable on time.

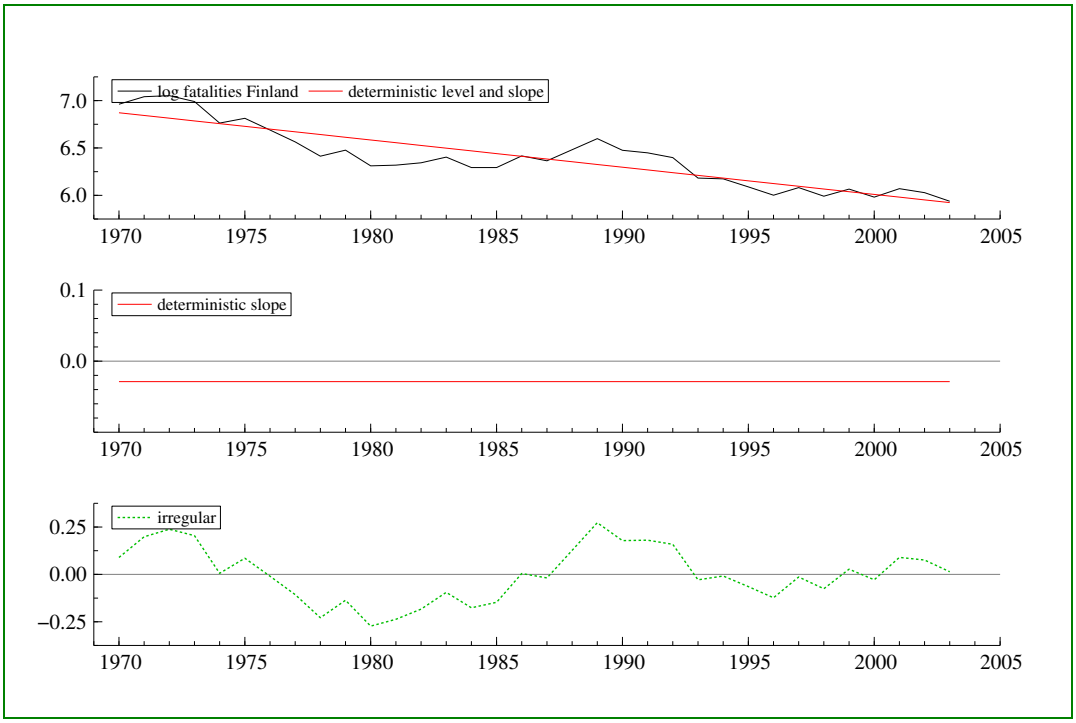


Figure 3.6.3: Deterministic trend (top), deterministic slope (middle), and irregular component for the log of the number of Finnish fatalities.

The best fitting regression line obtained with the deterministic linear trend model is shown at the top of Figure 3.6.3, while the bottom of Figure 3.6.3 contains the graph of the residuals of this classical regression analysis. Just a visual inspection of these residuals already reveals that they are not independent of one another.

	statistic	value	critical value	assumption satisfied
independence	Q(10)	73.199	16.92	-
	r(1)	0.767	0.34	-
	r(4)	0.271	0.34	+
homoscedasticity	1/H(11)	1.783	3.28	+
normality	N	2.226	5.99	+

Table 3.6.3: Diagnostic tests of residuals deterministic level and slope model for log Finnish fatalities.

This is confirmed by the results of the diagnostic tests for the residuals given in Table 3.6.3. The tests for homoscedasticity and normality are satisfactory, but the most important assumption of independence is clearly violated. The value of the AIC for this analysis is

$$\text{AIC} = \frac{1}{34}\left[-2(34)(0.3036367)+2(2+1)\right] = -0.430803.$$

Allowing both the level and the slope to vary over time in model (3.6.7), on the other hand, at convergence the value of the log-likelihood function equals 0.7864746. The value of the AIC for this analysis is therefore

$$\text{AIC} = \frac{1}{34}\left[-2(34)(0.7864746) + 2(2+3)\right] = -1.27883. \qquad (3.6.9)$$

The maximum likelihood estimates of the variances corresponding to the irregular, level, and slope components are $\hat{\sigma}_{\varepsilon}^2 = 0.00320083$, $\hat{\sigma}_{\xi}^2 = 9.69606\text{E}^{-}26$, and $\hat{\sigma}_{\zeta}^2 = 0.00153314$, respectively.

Since the variance of the level disturbances σ_{ξ}^2 is, for all practical purposes, equal to zero, the analysis is repeated with a deterministic level component, yielding the following results.

At convergence the value of the log-likelihood function equals 0.7864746. The maximum likelihood estimates of the variances of the observation and slope disturbances are $\hat{\sigma}_{\varepsilon}^2 = 0.00320083$, and $\hat{\sigma}_{\zeta}^2 = 0.00153314$, respectively. The maximum likelihood estimates of the values of the level and the slope right at the start of the series are $\hat{\mu}_1 = 7.0133$ and $\hat{\nu}_1 = 0.0068482$.

The trend (consisting of a deterministic level and a stochastic slope) of this analysis is displayed at the top of Figure 3.6.4, while the stochastic slope is shown separately in the middle of the figure. Since the time varying slope component in Figure 3.6.4 models the rate of change in the series, it can be interpreted as follows. When the slope component is *positive*, the trend in the series is *increasing*. Thus, log of the number of fatalities in Finland was increasing in the years 1970, 1982, 1984 through 1988, and in 1998 (see Figure 3.6.4). On the other hand, the trend is *decreasing* when the slope component is *negative*. The log of the number of fatalities in Finland was therefore decreasing in the remaining years of the series.

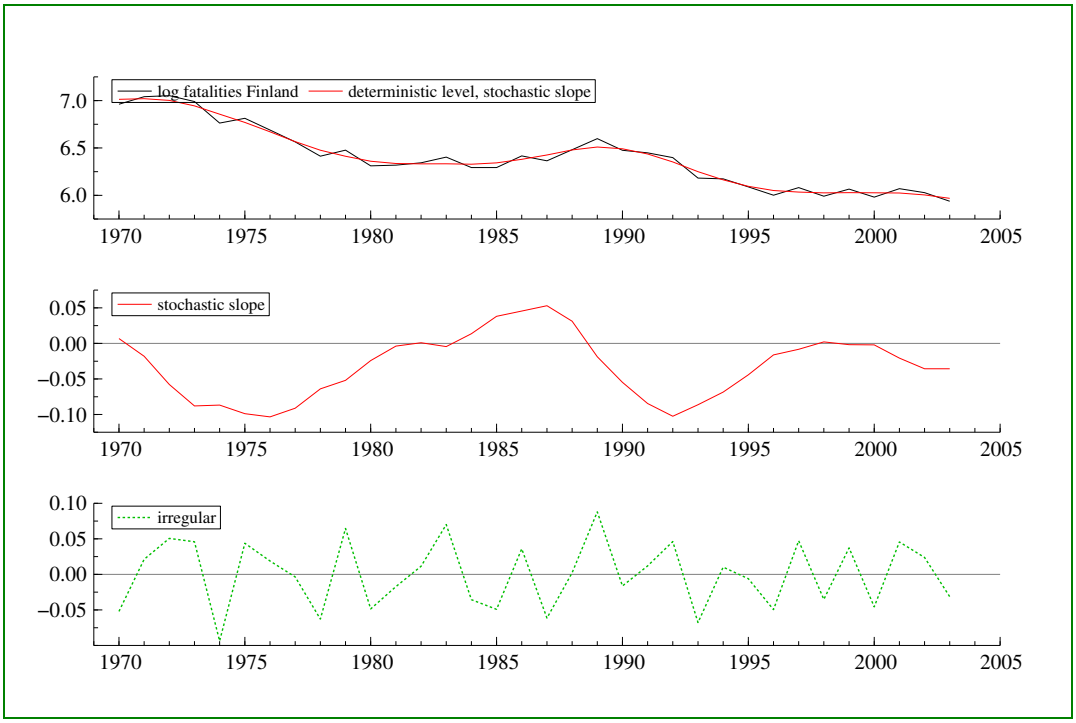


Figure 3.6.4: Trend of deterministic level and stochastic slope model for the log of Finnish fatalities (top), stochastic slope component (middle), and irregular component (bottom).

Moreover, when the slope is positive and increasing then the increase becomes more and more pronounced, while the increase becomes less and less pronounced (i.e., levels off) when the slope is positive but decreasing. Conversely, when the slope is negative and decreasing then the decrease becomes more and more pronounced, while the decrease levels off when the slope is negative but increasing.

The irregular component of this analysis is shown at the bottom of Figure 3.6.4, and the diagnostic tests for the residuals of the analysis are given in Table 3.6.4. As the table shows, the assumptions of independence, homoscedasticity, and normality are all satisfied, indicating that the deterministic level and stochastic slope model yields an appropriate description of the log of the annual traffic fatalities in Finland.

	statistic	value	critical value	assumption satisfied
independence	Q(10)	7.044	16.92	+
	r(1)	-0.028	0.34	+
	r(4)	-0.094	0.34	+
homoscedasticity	1/H(11)	1.348	3.28	+
normality	N	0.644	5.99	+

Table 3.6.4: Diagnostic tests for deterministic level and stochastic slope model, and log Finnish fatalities.

The Akaike information criterion for the deterministic level and stochastic slope model equals

$$\text{AIC} = \frac{1}{34}\left[-2(34)(0.7864746) + 2(2+2)\right] = \text{-1.33766.}$$

Thus, the fit of this model is slightly better than the fit of a model with stochastic level and stochastic slope. Since the log-likelihood values are identical for the two models, the improved fit of the second model can be completely attributed to its greater parsimony. The model with a deterministic level and stochastic slope is also called the *smooth trend* model, reflecting the fact that the trend of such a model is relatively smooth compared to a trend with a level disturbance variance unequal to zero.

Concluding, a smooth trend model with a constant level and a time-varying slope component yields a good description of the log of the annual road traffic fatalities in Finland for the period 1970 through 2003.

3.6.2.5. Conclusion on the technique

As the present section illustrates, the deterministic linear trend model actually performs a classical linear regression analysis of the dependent variable on the predictor variable time. This is an important and very useful result. By way of the Akaike information criterion, and of the residual tests for independence, homoscedasticity, and normality, this allows for a straightforward, fair and quantitative assessment of the relative merits of state space methods and classical regression models when it comes to the analysis of time series data. The reverse is also true: the state space models discussed in Section 3.6 are regression models in which the parameters (intercept and regression coefficient(s)) are allowed to vary over time.

In the following section, the effects of adding yet another component to the state are discussed: the *seasonal*.

3.6.3 Local linear trend plus seasonal model

Whenever a time series consists of hourly, daily, monthly, or quarterly observations with respective periodicity of 24 (hours), 7 (days), 12 (months), or 4 (quarters), one should always be on the alert for a special type of recurring pattern, called a *seasonal*. As an example, consider the plot of the log of the monthly number of drivers killed or seriously injured (KSI) in the United Kingdom (UK) for the period January 1969 through December 1984 in Figure 3.6.5. In the figure, vertical lines have been added through each year in the observed time series.

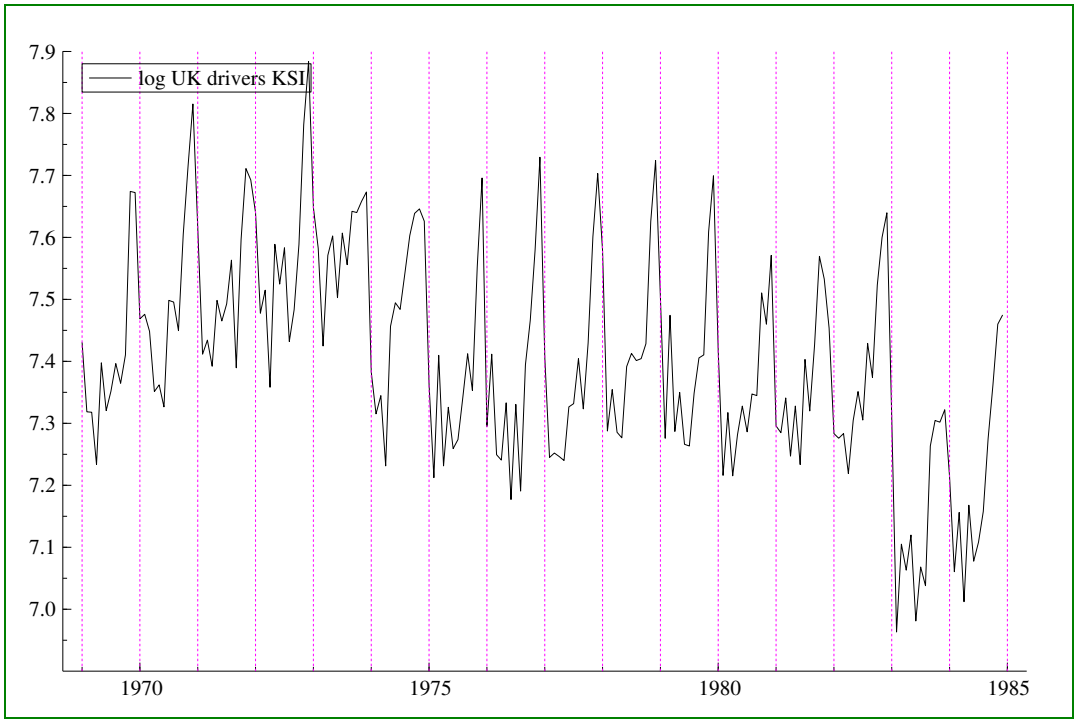


Figure 3.6.5: Log of monthly number of UK drivers KSI with time lines for years.

Inspecting the monthly development for each year in Figure 3.6.5, the following regularity emerges: in every year in this series more drivers are killed or seriously injured at the end of the year than during the rest of the year.

3.6.3.1. Objective of the technique

The objective of the local linear trend and seasonal model is to establish whether an observed time series containing a seasonal pattern can be described with a trend consisting of a time-varying level and a time-varying slope component, and a time-varying seasonal component.

3.6.3.2. Model definition and assumptions

In state space methods, a seasonal can be modelled by adding it either to the local level model or to the local linear trend model. Temporarily assuming quarterly data, adding a seasonal to the local linear trend model takes the following form:

$$\begin{aligned}
y_t &= \mu_t + \gamma_{1,t} + \varepsilon_t, & \varepsilon_t &\sim NID(0, \sigma_\varepsilon^2) \\
\mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim NID(0, \sigma_\xi^2) \\
\nu_{t+1} &= \nu_t + \zeta_t, & \varsigma_t &\sim NID(0, \sigma_\varsigma^2) \\
\gamma_{1,t+1} &= -\gamma_{1,t} - \gamma_{2,t} - \gamma_{3,t} + \omega_t, & \omega_t &\sim NID(0, \sigma_\omega^2) \\
\gamma_{2,t+1} &= \gamma_{1,t}, \\
\gamma_{3,t+1} &= \gamma_{2,t},
\end{aligned} \tag{3.6.10}$$

for $t = 1, ..., n$, where $\gamma_{1,t}$ denotes the seasonal component. The disturbances ω_t in (3.6.10) allow the seasonal to change over time.

In contrast with the level and slope components, which each only require one state equation, the modelling of a seasonal generally requires (s-1) state equations, where s is the periodicity of the seasonal. For quarterly data (where $s = 4$), for example, three state equations are needed, as is shown in (3.6.10). Irrespective of its periodicity, the seasonal always satisfies

$$\sum_{j=1}^{s} \gamma_{1,j} = 0, \tag{3.6.11}$$

thus ensuring that the seasonal is not confounded with the other components of the model. The type of seasonal that is modelled in (3.6.10) is called a *dummy* seasonal. There are other ways in which the seasonal component can be specified, one of them being the *trigonometric seasonal*. For the latter and other specifications of the seasonal the reader is referred to Durbin and Koopman (2001), as these specifications are beyond the scope of the present report.

The *assumptions* of the local linear trend and seasonal model (3.6.10) are that the observation, level, slope, and seasonal disturbances ε_t, ξ_t, ς_t, and ω_t are all mutually independent, and normally distributed with zero means, and variances equal to σ_ε^2, σ_ξ^2, σ_ς^2, and σ_ω^2, respectively.

3.6.3.3. Dataset and research problem

In general, the dataset in an analysis with the local linear trend plus seasonal model consists of only one variable: a time series y_t consisting of observations made sequentially through time points $t = 1, \ldots, n$.

As before, the remaining part of this section will first discuss and illustrate the effect of fixing all state disturbances ξ_t, ς_t, and ω_t in (3.6.10) on zero, and then present the effect of letting the level, slope, and seasonal components vary over time. In both cases, the model will be applied to the log of the monthly number of drivers killed or seriously injured (KSI) in the United Kingdom (UK) for the period January 1969 through December 1984, as presented in Figure 3.6.5.

The research problem addressed in this section is how to obtain an appropriate description of an observed time series with a seasonal pattern, i.e. the log of the monthly number of drivers KSI in the UK, January 1969 – December 1984.

3.6.3.4. Model fit, diagnostics, and interpretation of results

When the state disturbances ξ_t, ς_t, and ω_t in (3.6.10) are all fixed on zero, the model reduces to the following deterministic model:

$$y_t = \mu_1 + \nu_1(t-1) - \sum_{i=1}^{s-1} \gamma_{i,t-1} + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2). \tag{3.6.12}$$

Applying the latter model to the series shown in Figure 3.24 (with eleven instead of four state equations for the seasonal, since the UK series consists of monthly instead of quarterly data) the following results are obtained. The maximum likelihood estimate of σ_ε^2 equals 0.00981585, and the value of the log-likelihood function is 0.69830186. The values of $\hat{\mu}_1$ and $\hat{\nu}_1$ are 7.5540 and -0.00155, respectively. Thus, for these data the following holds:

$$y_t = 7.5540 - 0.00155(t-1) - \sum_{i=1}^{s-1} \gamma_{i,t-1} + \varepsilon_t,$$

which can also be written as

$$y_t = 7.5556 - 0.00155t - \sum_{i=1}^{s-1} \gamma_{i,t-1} + \varepsilon_t.$$

The estimates for the eleven initial values of the dummy seasonal are not mentioned here because these are not very informative in the present context.

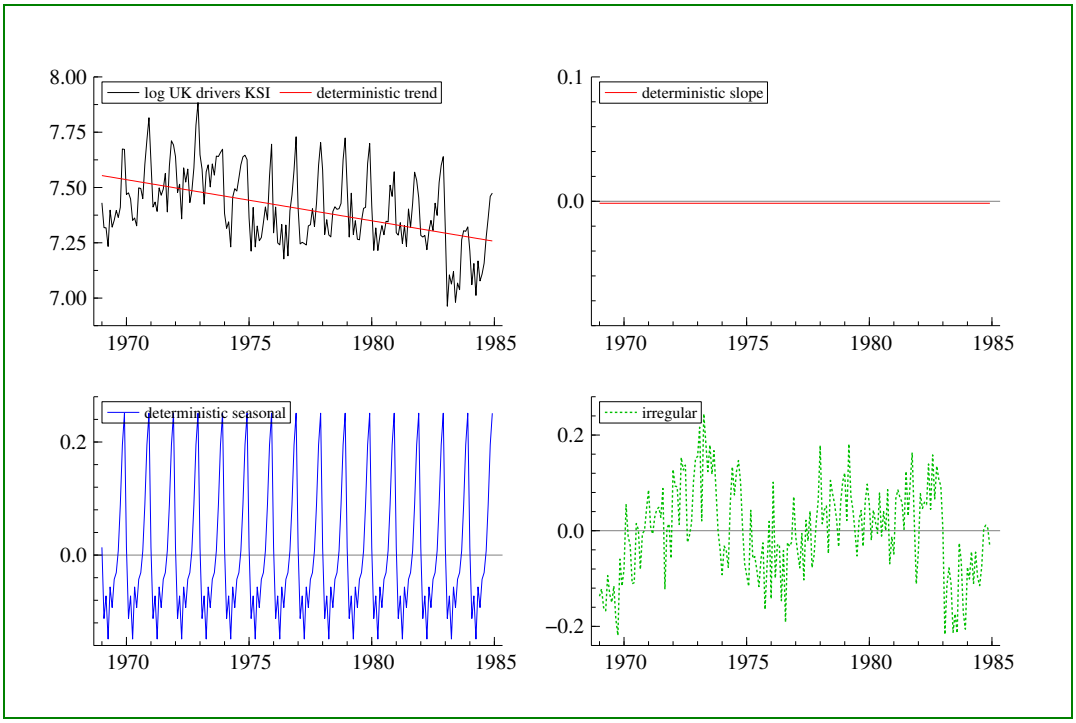


Figure 3.6.6: Deterministic trend (top left), deterministic slope (top right), deterministic seasonal (bottom left), and irregular component (bottom right) of deterministic trend and seasonal model for log UK drivers KSI.

The deterministic trend (which is the part equal to 7.5556-0.00155*t* in the just mentioned equation) of the analysis is shown at the top left of Figure 3.6.6, which also contains plots of the deterministic slope (top right), the deterministic seasonal (bottom left), and the irregular component (bottom right). The diagnostic tests in Table 3.6.5 of the irregular component in Figure 3.6.6 indicate that the residuals of this completely deterministic model neither satisfy the assumption of independence nor that of normality.

	statistic	value	critical value	assumption satisfied
independence	Q(15)	180.100	25.00	-
	r(1)	0.504	0.14	-
	r(12)	0.158	0.14	-
homoscedasticity	1/H(60)	1.008	1.67	+
normality	N	7.655	5.99	-

Table 3.6.5: Diagnostic tests for deterministic trend and seasonal model for log UK drivers KSI.

Since only one hyper-parameter was estimated (σ_ε^2), and a total of thirteen initial values for the state (i.e., one for the level, one for the slope, and eleven for the seasonal component), the Akaike information criterion for the completely deterministic trend and seasonal model equals

$$\text{AIC} = \frac{1}{192}[-2(192)(0.69830186) + 2(13+1)] = -1.25077.$$

In the previous sections, it was found that deterministic state space models are identical to some form of classical regression analysis. This suggests that the deterministic level, slope, and seasonal model must also have its counterpart in classical regression analysis. This is indeed the case. Results identical to those of the deterministic level, slope, and seasonal model presented above are obtained by performing the following classical multiple regression analysis.

Eleven dummy variables are constructed as follows. The first dummy variable is coded eleven (i.e., *s*-1) whenever an observation in the time series falls in the month of January, and minus one for all the other months of the year. The second dummy variable is coded eleven whenever an observation in the time series falls in the month of February and minus one elsewhere. And so on, until the eleventh and last dummy variable, which is coded eleven for the month of November and minus one elsewhere. A classical multiple regression analysis with the log of UK drivers KSI as dependent variable, and time *t* and these eleven dummy variables as independent variables yields the same results as those in Figure 3.6.6: the sum of the eleven dummy variables weighted by their respective regression coefficients is identical to the seasonal shown at the bottom left of Figure 3.6.6. The estimates for the intercept and for the regression coefficient for the independent variable time *t* are 7.5556 and –0.00155, respectively, meaning that the linear trend is identical to the linear trend in the top left of the figure. The residuals, finally, are therefore identical to those shown at the bottom right of Figure 3.6.6.

Allowing the level, slope and seasonal components in (3.6.10) all to vary over time, on the other hand, the following results are obtained. The algorithm converges to a log-likelihood value of 0.95650011, with disturbance variances $\hat{\sigma}_{\varepsilon}^2 = 0.00346783$, $\hat{\sigma}_{\xi}^2 = 0.00100094$, $\hat{\sigma}_{\zeta}^2 = 6.74681\text{E}^{-}52$, and $\hat{\sigma}_{\omega}^2 = 7.28648\text{E}^{-}025$. The values of μ_1 and ν_1 are 7.4133 and -0.00090532, respectively. Since the analysis requires the estimation of four hyper-parameters (i.e., disturbance variances), the Akaike information criterion now equals

$$\text{AIC} = \frac{1}{192}[-2(192)(0.95650011) + 2(13+4)] = -1.73592,$$

which is a big improvement upon the deterministic trend and seasonal model discussed above.

Since the slope and seasonal disturbance variances σ_{ζ}^2 and σ_{ω}^2 are found to be extremely small in the last analysis, these two components probably may as well be treated deterministically. This is confirmed by performing an analysis where the slope and seasonal disturbances ζ_t and ω_t in (3.6.10) are all fixed on zero. At convergence the value of the log-likelihood function is still 0.95650011, as before, while the maximum likelihood estimates of the disturbance variances are now

$\hat{\sigma}_\varepsilon^2 = 0.00346757$ and $\hat{\sigma}_\xi^2 = 0.0010011$. The values of μ_1 and ν_1 are now 7.4133 and -0.00090531, respectively. For this model, the Akaike information criterion equals

$$\text{AIC} = \frac{1}{192}\left[-2(192)(0.95650011) + 2(13+2)\right] = \text{-}1.75675\,,$$

which is a slight improvement upon the previous model. Since the values of the log-likelihood functions are for the two models are identical, this slight improvement can completely be attributed to the greater parsimony of the last model.

Finally, since the slope component is not only found to be best treated deterministically, but also obtains the fixed very small value of -0.00090531, it is allowed to consider completely dropping the slope component from the structural time series analysis of the log of the UK drivers KSI series. This yields the following results. Treating the level component stochastically and the dummy seasonal component deterministically, at convergence the value of the log-likelihood function equals 0.98299654. The value of μ_1 is 7.4118, and the maximum likelihood estimate of the variance of the irregular component is $\hat{\sigma}_\varepsilon^2 = 0.00351385$, and that of the level component equals $\hat{\sigma}_\xi^2 = 0.000945723$. This implies that the Akaike information criterion now equals

$$\text{AIC} = \frac{1}{192}\left[-2(192)(0.98299654) + 2(12+2)\right] = \text{-}1.82016.$$

The latter value of the AIC for the local level and deterministic dummy seasonal model is the smallest of all the seasonal models discussed so far, which is the reason why this model can be considered as the best model for describing the log of the UK drivers KSI series.

The three components of the latter analysis are all displayed in Figure 3.6.7. Moreover, the figure also contains a blown-up version of the dummy seasonal for the first year of the series, clearly indicating that April is the safest month for drivers in the UK, while December is the most dangerous month. Since the seasonal was treated deterministically in this analysis, this pattern is identical for all the other years in the series.

Finally, the diagnostic tests in Table 3.6.6 indicate that the residuals of this best fitting model satisfy all of the assumptions of the model, although the test for normality seems somewhat close to the critical value.

Concluding, a stochastic level and deterministic seasonal model yields the best description of the log of the monthly number of UK drivers killed or seriously injured for the period 1969 through 1984.

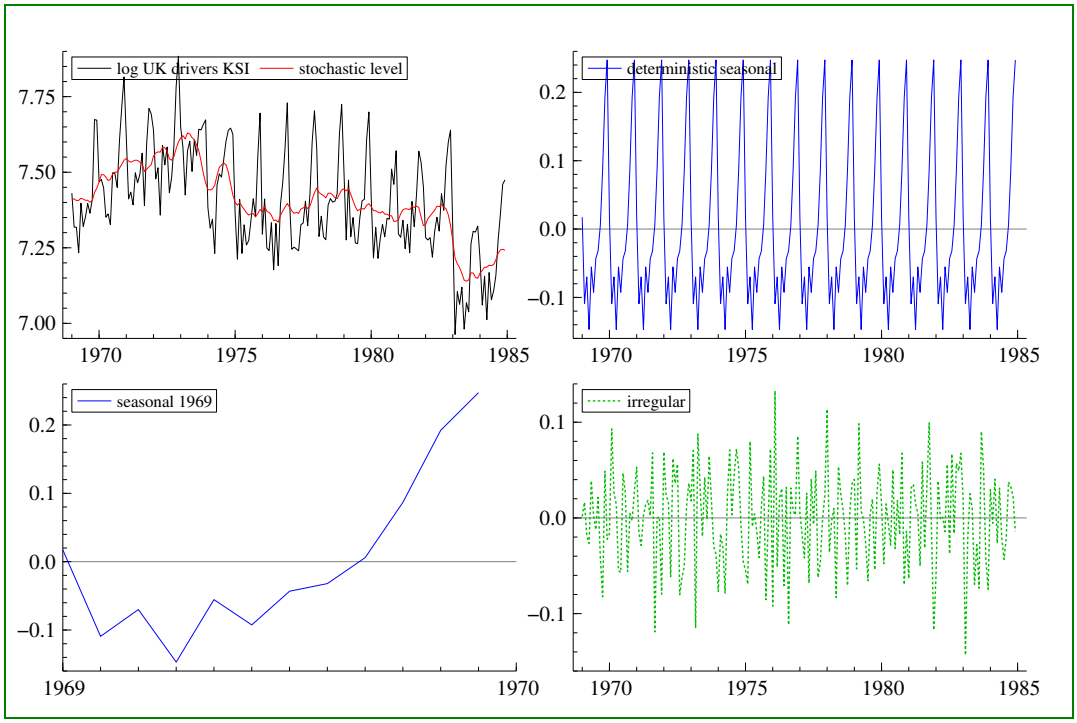


Figure 3.6.7: Stochastic level (top left), deterministic seasonal (top right), the seasonal for 1969 (bottom left), and irregular component (bottom right) for stochastic level and deterministic seasonal analysis of log of UK drivers KSI.

	statistic	value	critical value	assumption satisfied
independence	Q(15)	14.370	23.68	+
	r(1)	0.040	0.14	+
	r(12)	0.033	0.14	+
homoscedasticity	H(60)	1.093	1.67	+
normality	N	5.157	5.99	+

Table 3.6.6: Diagnostic tests for stochastic level and deterministic dummy seasonal analysis of log of UK drivers KSI.

3.6.3.5. Conclusion on the technique

The seasonal component in state space models facilitates the analysis of within-year patterns of quarterly, monthly, weekly, or even daily data.

So far, state components have been discussed that are useful for obtaining an adequate *description* of a time series. In the next two sections those components are presented that can be used to also obtain *explanations* for the observed developments in a time series.

3.6.4 Intervention variables

Apart from the diagnostic tools discussed in the previous sections for testing the assumptions of independence, homoscedasticity, and normality of the residuals in time series analysis, a second important diagnostic tool for determining the appropriateness of a model is provided by the inspection of its so-called *auxiliary residuals*. These auxiliary residuals are standardised versions of the observation disturbances ε_t and of the state disturbances ξ_t, ς_t, ω_t, etc. Inspection of the standardised observation disturbances allows for the detection of possible *outlier* observations, while the inspection of the standardised state disturbances makes it possible to detect *structural breaks* in the underlying development of a time series.

For the stochastic level and deterministic dummy seasonal model applied to the log of the UK drivers KSI series (see Section 3.6.3) for example, the standardised level disturbances of the analysis are presented at the top of Figure 3.6.7, while the standardised observation disturbances are shown at the bottom of the same figure.

Each of the auxiliary residuals at the top of Figure 3.6.7 can be considered as a *t*-test for the null hypothesis that there was no structural break in the level of the observed time series. The usual 95% confidence limits of ±1.96 for a two-tailed *t*-test are shown in the figure as two parallel horizontal lines. The auxiliary residuals exceed these limits at five time points, which is less than the $n/20 = 192/20 \approx 10$ that would be expected purely based on chance for this series. Still, the value of the residual for January 1983 particularly stands out as being very extreme.

Similarly, each of the auxiliary residuals at the bottom of Figure 3.6.8 can be considered as a *t*-test for the null hypothesis that the corresponding observation is not an outlier. Only seven out of the 192 observations exceed the 95% confidence limits of ±1.96, which is less than the ten that would be expected according to chance. Since, moreover, none of these are very extreme the conclusion is that the series does not contain outlier observations.

Summarising, inspection of the auxiliary residuals of the stochastic level and deterministic seasonal model applied to the log of the UK drivers KSI series suggests that there was a shift in the level in January 1983. This coincides with an actual event in the United Kingdom, which was the obligation from February 1983 onwards for motor vehicle drivers and front seat passengers to wear a seat belt.

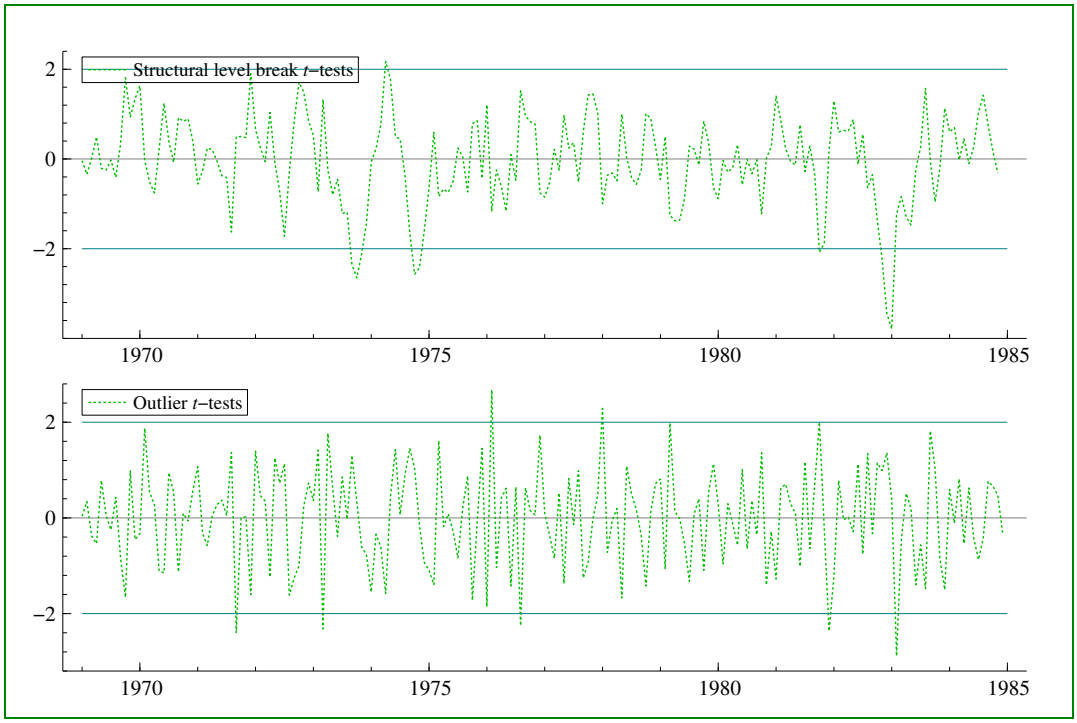


Figure 3.6.8: Auxiliary residuals for the stochastic level and deterministic seasonal model applied to the log of the UK drivers KSI series.

The effect of the introduction of this seat belt law can be investigated by adding an *intervention variable* to the model at hand. There are several ways in which an intervention can affect the development of a time series. One possible effect is that of a *level shift*, where the level of the time series suddenly changes and this level change continues after the intervention. A second possible effect is that of a *shift in the slope component*, where the value of the slope shows a continuous change after the intervention. A third possible effect is that of a *pulse*, where the value of a state component suddenly changes at the moment of the intervention, but then returns back to its previous value, in which case the effect is only temporary.
Since the auxiliary residuals in Figure 3.6.8 suggest a break in the level of the log of the UK drivers KSI, a level shift intervention variable will be added to the level and seasonal model discussed in the previous section.

3.6.4.1. Objective of the technique

The objective of the local level and seasonal model with an intervention variable is to establish the type, size and significance of the effect of the intervention variable on the development of an observed time series containing a seasonal pattern.

3.6.4.2. Model definition and assumptions

The level, the seasonal, and the level shift intervention variable for the introduction of the seat belt law in February 1983 are combined into the following state space model:

$$
\begin{aligned}
&y_t = \mu_t + \gamma_{1,t} + \lambda_t w_t + \varepsilon_t, && \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \\
&\mu_{t+1} = \mu_t + \xi_t, && \xi_t \sim NID(0, \sigma_\xi^2) \\
&\gamma_{1,t+1} = -\gamma_{1,t} - \gamma_{2,t} - \gamma_{3,t} + \omega_t, && \omega_t \sim NID(0, \sigma_\omega^2) \\
&\gamma_{2,t+1} = \gamma_{1,t}, \\
&\gamma_{3,t+1} = \gamma_{2,t}, \\
&\lambda_{t+1} = \lambda_t + \rho_t, && \rho_t \sim NID(0, \sigma_\rho^2)
\end{aligned}
\qquad (3.6.13)
$$

for $t = 1, \ldots, n$, where w_t is a dummy variable consisting of zeroes at all time points before the introduction of the seat belt law in February 1983, and ones at time points at and after the introduction in February 1983. To keep the number of state equations low, model (3.6.13) is presented as if dealing with quarterly data. In reality, however, there are thirteen state equations involved: one for the level, one for the regression coefficient λ_t of the intervention variable, and eleven for the seasonal. It may be noted that, although it would be technically possible to treat the regression component in the last state equation of (3.6.13) stochastically, in practice this is never done when dealing with intervention variables.

The *assumptions* of the local level and seasonal model (3.6.13) are that the observation, level, seasonal, and intervention disturbances ε_t, ξ_t, ω_t, and ρ_t are all mutually independent, and normally distributed with zero means, and variances equal to σ_ε^2, σ_ξ^2, σ_ω^2, and σ_ρ^2, respectively.

3.6.4.3. Dataset and research problem

In general, the *dataset* in a state space analysis with one intervention contains two variables: a dependent variable y_t which is a time series as before, and an independent intervention variable which is denoted by w_t.

The remaining part of this section will first discuss and illustrate the effect of fixing all state disturbances ξ_t, ω_t, and ρ_t in (3.51) on zero and then present the effect of letting the level component vary over time. In both cases, the local linear trend plus seasonal model from Section 3.6.3 extended with one intervention, i.e. the introduction of the seat belt law, will be applied to the log of the monthly number of drivers killed or seriously injured (KSI) in the United Kingdom (UK) for the period January 1969 through December 1984 (see Figure 3.6.5).

The *research problem* addressed in this section is how to assess the effect of the introduction of the seat belt law in February 1983 on the log of the number of drivers KSI in the UK, January 1969 – December 1984.

3.6.4.4. Model fit, diagnostics, and interpretation of results

Treating all the state components in (3.6.13) deterministically, it is not very difficult to prove that the model simplifies into the following classical regression model:

$$y_t = \mu_1 - \sum_{i=1}^{s-1} \gamma_{i,t-1} + \lambda_1 w_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0,\sigma_\varepsilon^2). \tag{3.6.14}$$

Estimating model (3.6.14) by fixing all the state disturbances in (3.6.13) on zero, the value of the log-likelihood function equals 0.71553091. The optimal values of μ_1 and λ_1 are 7.4373 and -0.26075, respectively, and the maximum likelihood estimate of the irregular variance is $\hat{\sigma}_\varepsilon^2 = 0.0100188$. The best fitting classical regression model can therefore be written as

$$y_t = 7.4373 - \sum_{i=1}^{s-1} \gamma_{i,t-1} - 0.26075 w_t + \varepsilon_t.$$

The effect of the intervention variable on the deterministic level of the model is clearly seen in the top graph in Figure 3.6.9. The level which is equal to 7.4373 until January 1983 suddenly shifts down to the value of 7.4373 - 0.26075 = 7.17655 in February 1983. Since the dependent variable is analysed in its logarithm, the following formula must be used to re-express the level change in a percentage change in the absolute numbers of drivers KSI:

$$e^{\hat{\lambda}_1} - 1 = e^{-0.26075} - 1 = -0.2295,$$

meaning that -according to this model- the introduction of the seat belt law resulted in a change of (100)(-0.2295) = -23% in the number of drivers KSI.

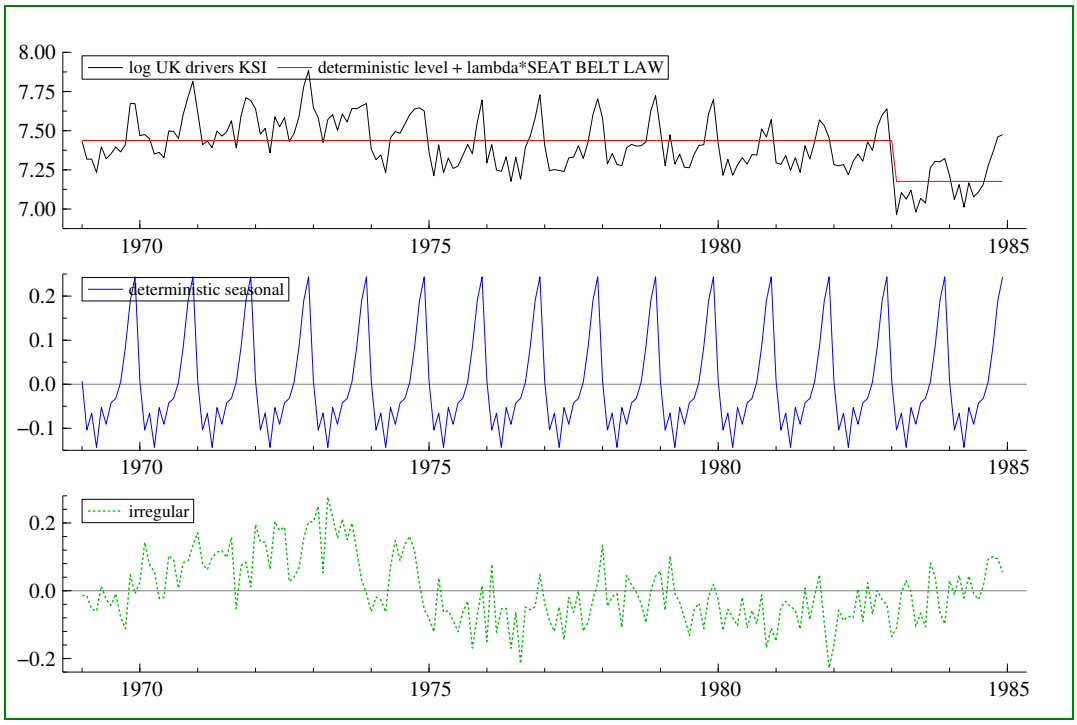


Figure 3.6.9: Deterministic level plus intervention variable (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series .

The value of the Akaike information criterion for this model equals

$$\text{AIC} = \frac{1}{192}\left[-2(192)(0.71553091) + 2(13+1)\right] = -1.28523.$$

The latter value of the AIC indicates that the deterministic level and dummy seasonal model with intervention variable yields a much better fit than the deterministic level and dummy seasonal model without intervention variable, which results in an AIC value of only -0.792879.

	statistic	value	critical value	assumption satisfied
independence	Q(15)	524.110	23.68	-
	r(1)	0.604	0.14	-
	r(12)	0.402	0.14	-
homoscedasticity	1/H(60)	1.475	1.67	+
normality	N	3.604	5.99	+

Table 3.6.7: Diagnostic tests for deterministic level and seasonal analysis of log of UK drivers KSI, including intervention variable.

The standard *t*-test for establishing whether the regression coefficient $\hat{\lambda}_1 = -0.26075$ deviates from zero yields

$$t = \frac{-0.260751590 8}{0.022277472 68} = -11.70472049\,, \qquad (3.6.15)$$

which is very significant. In order to investigate whether this test is reliable, it must be checked whether the model satisfies the assumptions of independence, homoscedasticity and normality of the residuals. However, as Table 3.6.7 indicates, the residuals do not satisfy the most important assumption of independence, meaning that the value of the just mentioned *t*-test (and especially the value of the standard error in the denominator) can not be trusted, and is probably much too large (since the first autocorrelation r(1) is positive).

If the level component in model (3.6.13) is allowed to vary over time, on the other hand, at convergence the value of the log-likelihood function equals 1.0168174. The maximum likelihood estimates of μ_1 and λ_1 are 7.4108 and -0.23981, respectively, and the maximum likelihood estimates of the irregular and level variances are $\hat{\sigma}_\varepsilon^2 = 0.00378397$ and $\hat{\sigma}_\xi^2 = 0.000473516$, respectively.

The estimated effect of the seat belt law re-expressed in the percentage change in the absolute numbers of drivers KSI is now

$$e^{\hat{\lambda}_1} - 1 = e^{-0.23981} - 1 = -0.2132\,,$$

meaning that -according to this model- the introduction of the seat belt law resulted in a change of (100)(-0.2132) = -21.3% in the number of UK drivers KSI.

The Akaike information criterion for this model equals

$$\text{AIC} = \frac{1}{192}\left[-2(192)(1.0168174) + 2(13+2)\right] = -1.87738.$$

The latter value of the AIC for the local level and deterministic dummy seasonal model including a level shift intervention for the introduction of the seat belt law is smaller than that for the same model without intervention variable which is -1.82016 (see the previous section). This means that the intervention variable for the seat belt law improves the fit.

Whether the contribution of the intervention variable is significant can again be tested with the standard *t*-test for the regression coefficient $\hat{\lambda}_1 = -0.23981$, yielding

$$t = \frac{-0.239806756}{0.053070218 83} = -4.5187\,. \qquad (3.6.16)$$

The value of the latter *t*-test is still very significant, but in absolute terms it is much smaller than the value of the *t*-test (3.6.15) in the previous completely deterministic model.

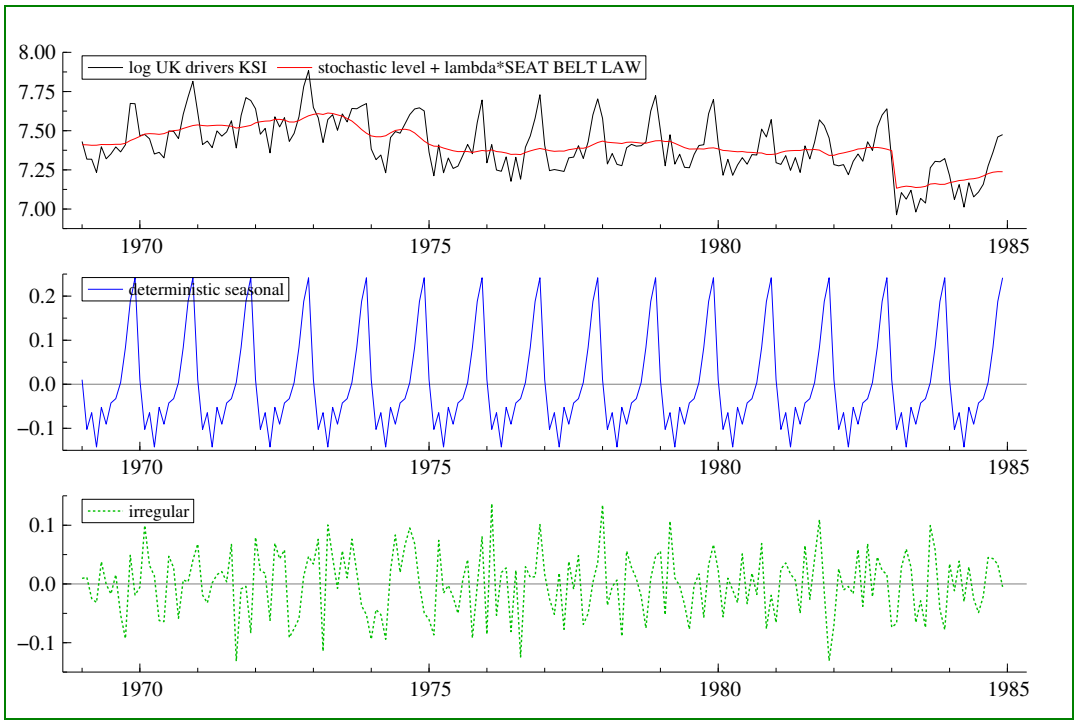


Figure 3.6.10: Stochastic level plus intervention variable (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series.

The stochastic level plus intervention variable is shown in Figure 3.6.10, together with the deterministic dummy seasonal, and the irregular component. The diagnostic tests for the model assumptions are given in Table 3.6.8. Since all three assumptions are satisfied in the present analysis, now it is assured that the *t*-test in (3.54) is a reliable test.

	statistic	value	critical value	assumption satisfied
Independence	Q(15)	17.928	23.68	+
	r(1)	0.080	0.14	+
	r(12)	0.085	0.14	+
homoscedasticity	1/H(60)	1.639	1.67	+
normality	N	2.928	5.99	+

Table 3.6.8: Diagnostic tests for stochastic level and dummy seasonal analysis of log of UK drivers KSI, including intervention variable.

As Figure 3.6.8, Figure 3.6.11 plots the auxiliary residuals of the local level and deterministic seasonal model applied to the log of the UK drivers KSI, but now including the intervention variable for the introduction of the seat belt law. It is interesting to note that the large extreme value that was previously found in January 1983 for the standardised level disturbances (see Figure 3.6.8) has now completely disappeared. This is the effect of adding the intervention variable to the model.

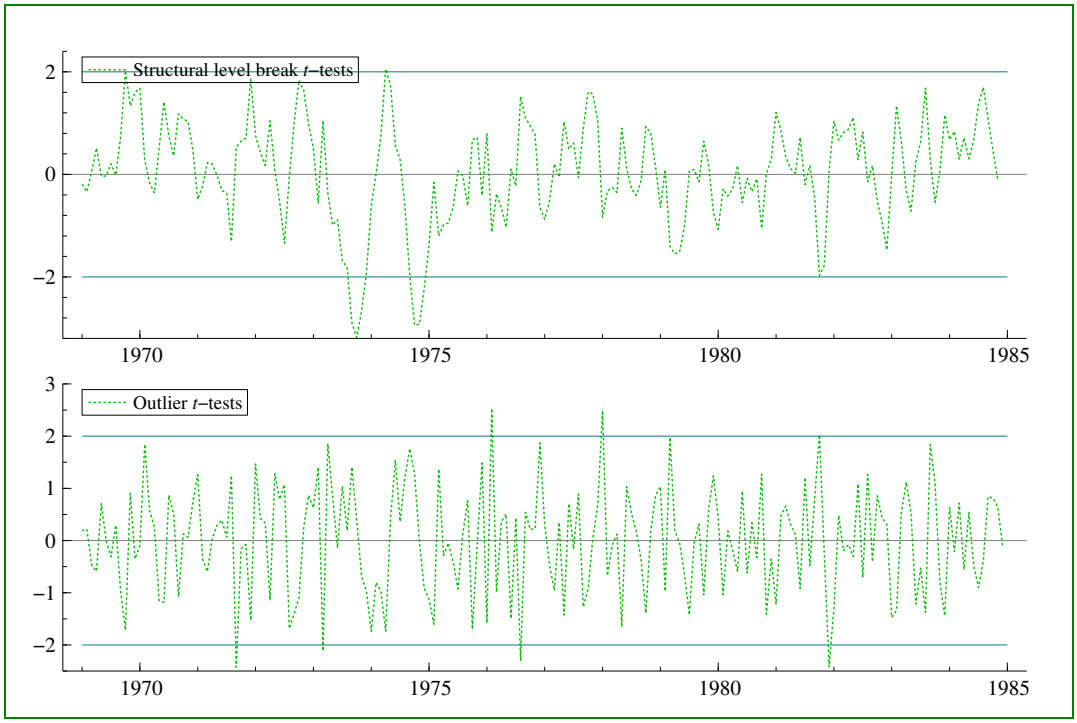


Figure 3.6.11: Auxiliary residuals for the stochastic level and deterministic seasonal model applied to the log of the UK drivers KSI series, including a level shift intervention variable for the introduction of the seat belt law.

Concluding, the fit of the stochastic level and deterministic seasonal model that yields the best description of the log of the monthly number of UK drivers killed or seriously injured for the period 1969 through 1984 can significantly be improved by adding a level shift intervention variable to the model, where the level shift is applied to February 1983 in the series, the month that the seat belt law for drivers and front seat passengers was introduced in the UK. Moreover, the analysis suggests that the introduction of the seat belt law resulted in a 21.3% reduction in the number of UK drivers KSI.

Finally, when comparing the value of the t-test for the regression coefficient of the intervention variable in a completely deterministic (i.e. classical regression) model with that in the stochastic level model, it can be seen that the former test is seriously flawed due to the remaining dependencies in the residuals of the classical regression analysis. In fact, compared to the t-test of the stochastic model the absolute value of the test in the classical regression analysis is 11.7/4.5 = 2.6 times too large.

3.6.4.5. Conclusion on the technique

In state space modelling, the auxiliary residuals are a helpful tool in detecting outlier observations and structural breaks in the level, slope, and seasonal components. As this section demonstrated, a structural break in the level component is an indication of an intervention which suddenly and radically changed the level and, as such, it can be removed by including an intervention

variable. Structural breaks in the slope and seasonal components and outlier observations can be dealt with in a similar way.

Furthermore, the analysis results in this section show that the *t*-test for the regression coefficient in a classical linear regression model can be seriously flawed due to dependencies in the residuals.

3.6.5 Explanatory variables

3.6.5.1. Objective of the technique

The objective of the local level and seasonal model with an intervention variable and a continuous explanatory variable is to establish the type, size and significance of the effects of both the intervention variable and the explanatory variable on the development of an observed time series containing a seasonal pattern.

3.6.5.2. Model definition and assumptions

Just like intervention variables, explanatory variables can simply be added to the measurement equation of any of the state space models discussed so far. If they are added to the local level and seasonal model with an intervention variable, for example, then the measurement equation is:

$$y_t = \mu_t + \gamma_{1,t} + \lambda_t w_t + \sum_{j=1}^{k} \beta_{jt} x_{jt} + \varepsilon_t, \qquad (3.6.17)$$

where the x_j are k continuous explanatory variables (j = 1, …, k), and the β_j are unknown regression weights or coefficients.

We will illustrate the effect of explanatory variables by adding one continuous explanatory variable to the time series analysis of the log of the UK drivers KSI series shown in Figure 3.24. This continuous variable consists of the log of the monthly prices of petrol in the UK in the period 1969 through 1984. The idea is that higher petrol prices may have induced UK car drivers to circulate less in traffic, thus reducing the number of traffic accidents. The model includes the same intervention variable that was used in the previous section, i.e. the introduction of the seat belt law in February 1983 in the United Kingdom.

The level, the dummy seasonal, the introduction of the seat belt law, and the log of petrol price are combined into the following state space model:

$$\begin{aligned}
&y_t = \mu_t + \gamma_{1,t} + \lambda_t w_t + \beta_t x_t + \varepsilon_t, && \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \\
&\mu_{t+1} = \mu_t + \xi_t, && \xi_t \sim NID(0, \sigma_\xi^2) \\
&\gamma_{1,t+1} = -\gamma_{1,t} - \gamma_{2,t} - \gamma_{3,t} + \omega_t, && \omega_t \sim NID(0, \sigma_\omega^2) \\
&\gamma_{2,t+1} = \gamma_{1,t}, && \qquad (3.6.18) \\
&\gamma_{3,t+1} = \gamma_{2,t}, &&
\end{aligned}$$

$$\lambda_{t+1} = \lambda_t + \rho_t, \qquad \rho_t \sim NID(0, \sigma_\rho^2)$$

$$\beta_{t+1} = \beta_t + \tau_t, \qquad \tau_t \sim NID(0, \sigma_\tau^2)$$

for t = 1, …, n, where w_t again contains zeroes at all time points before February 1983, and ones at time points at and after February 1983, and x_t is the continuous predictor variable "log petrol price". Again, the model (3.6.18) is presented as if dealing with quarterly data. In reality, however, there are fourteen state equations involved: one for the level, two for the regression coefficients λ_t and β_t of the intervention and explanatory variables w_t and x_t, respectively, and eleven for the seasonal. It may be noted that state space methods allow for a stochastic treatment of the regression component in the last state equation of (3.6.18), thus allowing the regression coefficient to vary over time. Here, however, only deterministic regression components are considered.

The assumptions of model (3.6.18) are that the observation, level, seasonal, intervention, and explanatory disturbances ε_t, ξ_t, ω_t, ρ_t, and τ_t are all mutually independent, and normally distributed with zero means, and variances equal to σ_ε^2, σ_ξ^2, σ_ω^2, σ_ρ^2, and σ_τ^2, respectively.

3.6.5.3. Dataset and research problem

The dataset in a state space analysis with intervention and explanatory variables consists of the dependent variable y_t which is a time series as before, an independent intervention variable w_t, and the k continuous independent variables x_j which are all time series as well.

The remaining part of this section will first discuss and illustrate the effect of fixing all state disturbances ξ_t, ω_t, ρ_t, and τ_t in (3.6.18) on zero and then present the effect of letting the level component vary over time. In both cases, the local linear trend with seasonal model with the added seat belt law intervention (see Section 3.6.4) and extended with the explanatory variable log petrol price will be applied to the log UK drivers KSI dataset from Figure 3.6.5.

The research problem addressed in the present section is to investigate the effects of a continuous explanatory variable, i.e. log petrol price on the development of a time series, i.e. the log of the number of drivers KSI in the UK, January 1969 – December 1984 .

3.6.5.4. Model fit, diagnostics, and interpretation of results

Treating all the state components deterministically, the value of the log-likelihood function equals 0.84903819. The maximum likelihood estimates of μ_1, λ_1, and β_1 are 6.4016, -0.19714, and -0.45213, respectively, and the maximum likelihood estimate of the irregular variance is $\hat{\sigma}_\varepsilon^2 = 0.00740223$.

The model therefore reduces to a classical regression model with regression equation

$$y_t = 6.4016 - \sum_{i=1}^{s-1} \gamma_{i,t-1} - 0.19714 w_t - 0.45213 x_t + \varepsilon_t .$$

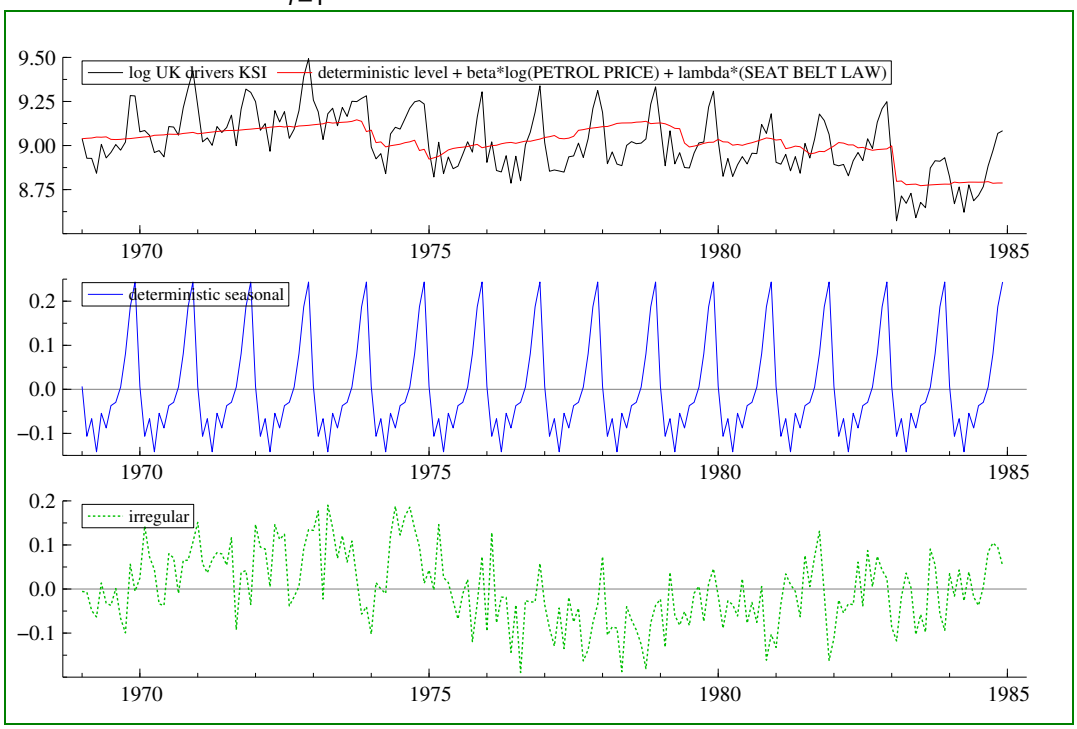


Figure 3.6.12: Deterministic level plus intervention and explanatory variable (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series .

The plot of the deterministic level plus intervention and explanatory variables is shown in Figure 3.6.12, together with the fixed dummy seasonal and the irregular component.

Since exp(-0.19714) – 1 = -0.1789, according to the present analysis the seat belt law resulted in a 17.9% reduction in the number of drivers KSI. Since the variables "number of drivers KSI" and "petrol price" are both analysed in their logarithms, the regression coefficient β_1 may be interpreted as a so-called *elasticity*, meaning that a 1% change in the petrol price is associated with a β_1% change in the number of drivers KSI. If the present analysis were correct, therefore, the conclusion would be that a 1% raise in the price of petrol was associated with a 0.45% *reduction* (since $\hat{\beta}_1$ is negative) in the number of drivers KSI. A nice property of analysing both the number of drivers KSI and the price of petrol in their logarithms is that the value of the elasticity β_1 remains unchanged when the number of drivers KSI is multiplied with a positive number and/or when the price of petrol is multiplied with a positive number.

The value of the Akaike information criterion for this model equals

$$\text{AIC} = \frac{1}{192}\left[-2(192)(0.84903819)+2(14+1)\right] = -1.54183\,,$$

which is a clear improvement upon the completely deterministic model without "log petrol price".

The standard *t*-test for establishing whether the regression coefficient $\hat{\lambda}_1 = -0.19714$ for the intervention variable deviates from zero yields

$$t = \frac{-0.1971394716}{0.02072756003} = -9.510983022\,,$$

which is very significant. The standard *t*-test for establishing whether the regression coefficient $\hat{\beta}_1 = -0.45213$ for the continuous variable "log petrol price" deviates from zero yields

$$t = \frac{-0.452130127}{0.05639609976} = -8.017046017\,,$$

which is also very significant.

	statistic	value	critical value	assumption satisfied
Independence	Q(15)	147.020	23.68	-
	R(1)	0.426	0.14	-
	r(12)	0.198	0.14	-
homoscedasticity	1/H(59)	1.110	1.67	+
Normality	N	0.560	5.99	+

Table 3.6.9: Diagnostic tests for deterministic level and dummy seasonal analysis of log of UK drivers KSI, including variables seat belt law and log petrol price.

However, before drawing any conclusions it must be checked whether the residuals satisfy the model assumptions. As Table 3.6.9 indicates, the most important assumption of independence is clearly violated in this classical regression model, meaning that the values of the just mentioned *t*-tests are seriously inflated since r(1) is positive.

Allowing the level component to vary over time, at convergence the value of the log-likelihood function equals 1.0265254. The estimates for μ_1, λ_1, and β_1 are 6.7814, -0.23759, and -0.27674, respectively. The maximum likelihood estimate of the irregular variance is $\sigma_\varepsilon^2 = 0.00403394$, and that of the level variance is $\sigma_\xi^2 = 0.000268082$. Thus, the measurement equation can be written as

$$y_t = \mu_t - \sum_{i=1}^{s-1} \gamma_{i,t-1} - 0.23759 w_t - 0.27674 x_t + \varepsilon_t .$$

Graphs of the components of the analysis are shown in Figure 3.6.13.

The percent change in the number of drivers KSI due to the seat belt law is now estimated to be equal to (100)(ex(-0.23759) - 1) = -21.1%, while a 1% raise in the petrol price is now associated with a 0.28% reduction in the number of drivers KSI.

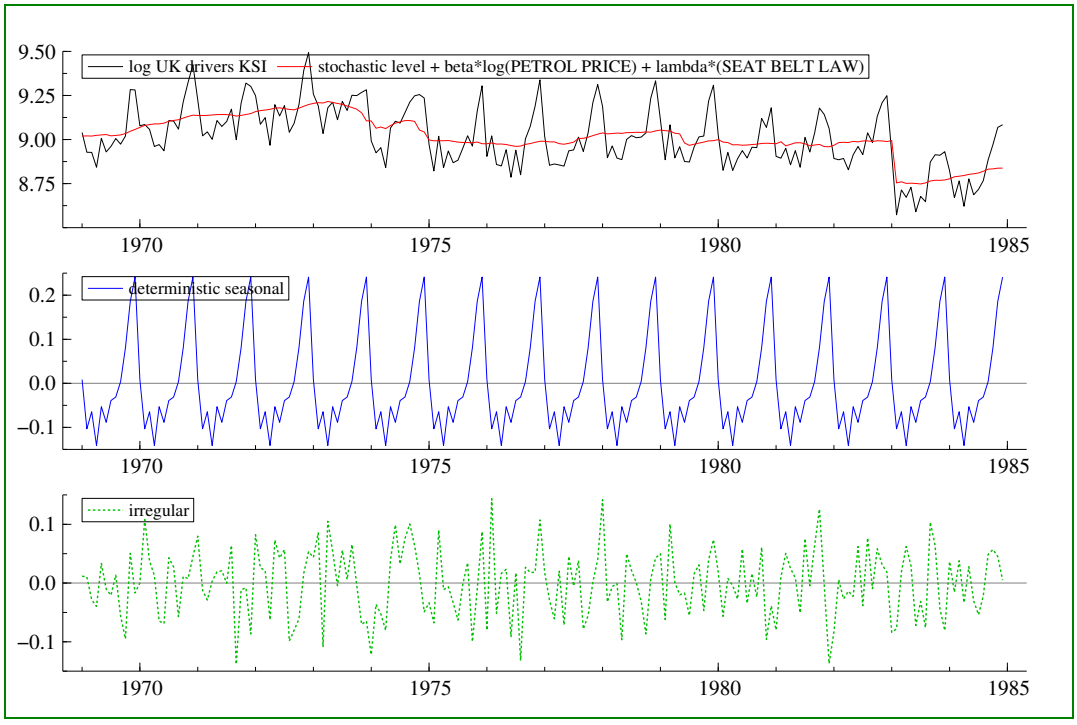


Figure 3.6.13: Stochastic level plus intervention and explanatory variables (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series .

The value of the Akaike information criterion for this model equals

$$\text{AIC} = \frac{1}{192}\left[-2(192)(1.0265254) + 2(14+2)\right] = -1.88638 ,$$

meaning that this is the best fitting of all the models that were used to analyse the log of the UK drivers KSI series.

The standard *t*-test for establishing whether the regression coefficient $\hat{\lambda}_1 = -0.23759$ deviates from zero yields

$$t = \frac{-0.2375871946}{0.04644589627} = -5.115353857 ,$$

which is significant. The standard *t*-test for establishing whether the regression coefficient $\hat{\beta}_1 = -0.45213$ deviates from zero yields

$$t = \frac{-0.276740442}{0.09840666428} = -2.812212405,$$

which is also significant.

	statistic	value	critical value	assumption satisfied
Independence	Q(15)	18.676	23.68	+
	r(1)	0.078	0.14	+
	r(12)	0.068	0.14	+
homoscedasticity	1/H(59)	1.025	1.67	+
Normality	N	1.444	5.99	+

Table 3.6.10: Diagnostic tests for stochastic level and dummy seasonal analysis of log of UK drivers KSI, including variables seat belt law and log petrol price.

As Table 3.6.10 shows, all the model assumptions are satisfied in the present analysis, meaning that the *t*-tests for the regression coefficients are no longer flawed in this case.

Concluding, adding the continuous explanatory variable “log petrol price” to the stochastic level and deterministic seasonal model with a level shift intervention variable also helps in explaining the observed development in the log of the monthly number of UK drivers KSI series.

As before, keeping the intercept (i.e. the level) fixed over time results in residuals that do not satisfy the assumption of independence, and therefore in inflated *t*-tests for the regression coefficients. Allowing the intercept to vary over time, on the other hand, all model assumptions are satisfied, and the *t*-tests are now reliable. The comparison of the *t*-tests in the model with a fixed intercept with those in the model with a time-varying intercept shows that – in absolute value - the test for the regression coefficient of the intervention variable is almost two times too large, while that for regression coefficient of the log of petrol price is almost three times too large.

In the appropriate model, the values of the regression coefficients indicate that the seat belt law resulted in a 21.1% reduction in the number of UK drivers KSI, while a 1% raise in the price of petrol was associated with a 0.28% reduction in the number of drivers KSI. Finally, it is noted that the estimated effect of a 21.1% reduction as a result of the seat belt law in the present analysis is almost identical to the value of 21.3% found with the model without the explanatory variable “log petrol price” (see the previous section).

3.6.5.5. Conclusion on the technique

Explanatory variables can be added to the state space model and their contribution to the dependent variable can be tested reliably. As was shown in this section, in fully deterministic, classical linear regression models the reliability of the *t*-tests for the regression coefficients is not guaranteed, which can lead to incorrect conclusions regarding the significance of those regression coefficients.

Until now, the focus was on the descriptive and explanatory aspects of state space methods. The next section will discuss the issue of *forecasting* with structural time series models.

3.6.6 Forecasting

For a proper understanding of forecasting in state space methods, it is useful to mention that the state components of state space models can be estimated in a number of ways. All the previous sections on the theory of state space methods presented the estimate of the state that is known as the *smoothed* state. The smoothed state at time *t* is typically based on *all* available observations in the time series, therefore including those observations y_{t+1}, …, y_n that occurred after time point *t*.

A second type of estimate is the so-called *filtered* state. The filtered state at time *t* is the estimate of the state only based on all *past* observations y_1, …, y_{t-1}, ànd on the *current* observation y_t.

The third type of estimate is the so-called *predicted* state. The predicted state at time *t* is the estimate of the state purely based on all *past* observations y_1, …, y_{t-1}. This last type of estimate typically yields forecasts as they are obtained with state space methods. It is interesting to note that forecasts in structural time series analysis are actually obtained by treating the future observations in a series as missing.

This section will present three examples of forecasting: one with the local level model, one with the local linear trend model, and one with the local level and seasonal model with an explanatory and intervention variable.

As discussed in Section 3.6.1 the log of the annual number of Norwegian fatalities in the period 1970-2003 can be adequately described with the local level model. The local level model was therefore also used to obtain forecasts for this series in the period 2004-2010. The filtered level and the forecasts obtained with the local level model for the years 2004 through 2010 are shown in Figure 3.6.14, together with their 90% confidence limits.
As the latter figure shows, forecasts of the local level model are always located on a straight horizontal line whose level is equal to the filtered level at time point *n*+1. The values of the forecasts in Figure 3.6.14 are all equal to 5.6627. According this analysis therefore the future number of road traffic fatalities in Norway will remain at a constant level of exp(5.6627) = 288 fatalities per year.

In state space methods, all estimates of the components of the state also have associated *estimation error variances*. This is true irrespective whether the estimate is the smoothed, the filtered or the predicted state. Under the assumption of normality, these estimation error variances allow the construction of confidence intervals for each of the state components, thus making it possible to assess the (un)certainty in the estimates of the state. Letting $\mathrm{Var}(\mu_t)$ denote the estimation error variance of the level μ_t of the local level model, therefore, the 90% confidence limits are computed with the well-known formula

$$\mu_t \pm 1.64\sqrt{\mathrm{Var}(\mu_t)}, \qquad (3.6.19)$$

where +1.64 and –1.64 are the *z*-scores corresponding to the 90% interval around the mean of a normal distribution.

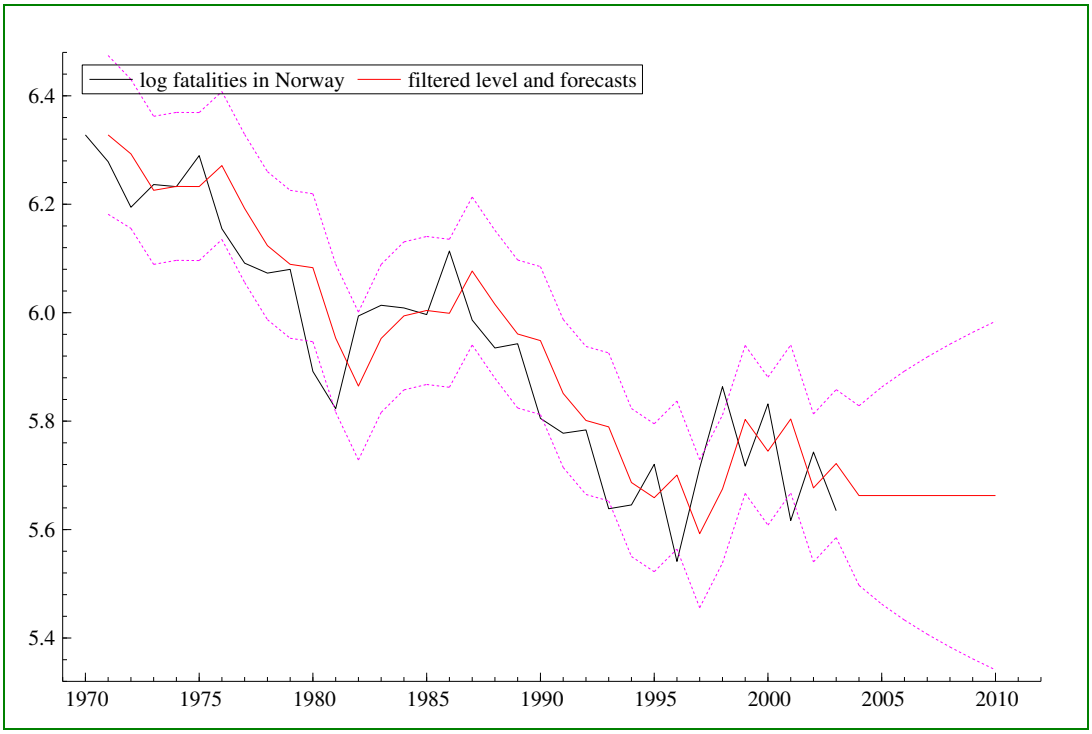


Figure 3.6.14.: Filtered level, and seven years forecasts for log of Norwegian fatalities including their 90% confidence limits.

The thus computed 90% interval for the filtered and predicted level of the local level model is displayed in Figure 3.6.14. As the figure shows, the estimation error variance for the predicted level, and therefore its uncertainty, becomes larger and larger as the forecasts are located further into the future.

The analysis of the log of the annual number of traffic fatalities in Finland with the smooth trend model (see Section 3.6.2) was also used to obtain forecasts using a so-called lead time of seven years. The observations of the series are shown in Figure 3.6.15, together with the filtered state for the years 1970 through 2003, and the predicted state (i.e., the forecasts from the smoothed trend model) for the years 2004 through 2010. As the figure shows, forecasts of the local linear trend model are always located on a straight line with constant level and slope. Again, the estimation error variance for the predicted trend, and therefore its uncertainty, becomes larger and larger as the forecasts are located further into the future.

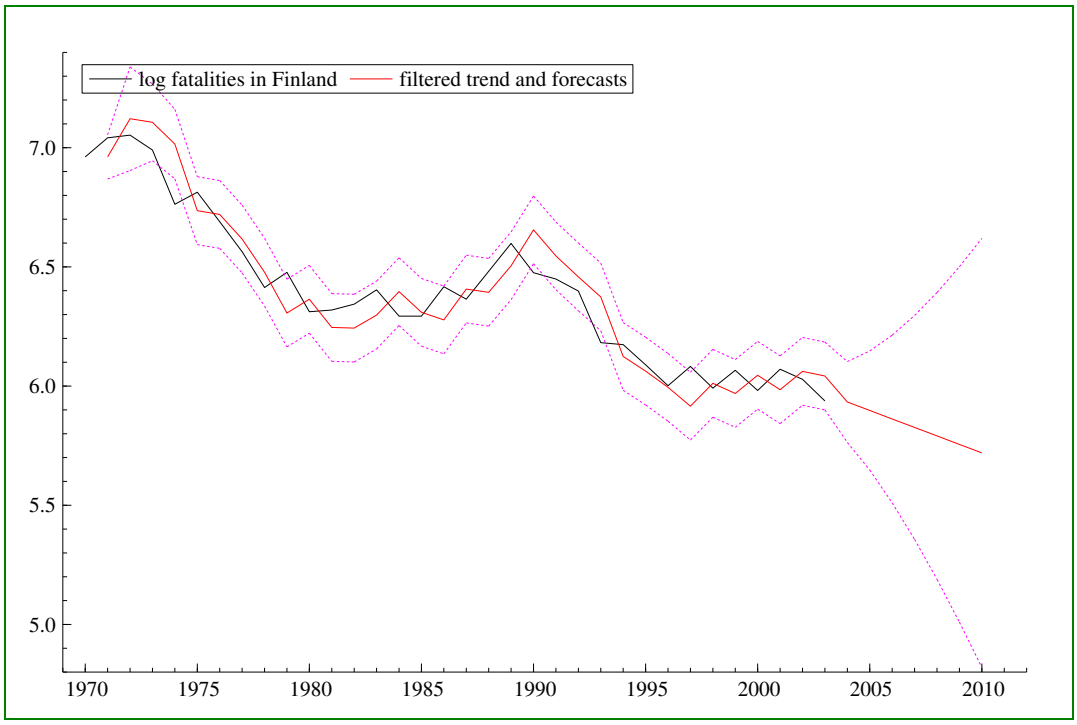


Figure 3.6.15: Filtered trend, and seven year forecasts for Finnish fatalities, including their 90% confidence limits.

As a last example, the log of the UK drivers KSI series was re-analysed (see Sections 3.6.3, 3.6.4, and 3.6.5) with a local level and deterministic dummy seasonal model, including the log of the petrol price and the introduction of the seat belt law as independent variables. In contrast with the analysis discussed in Section 3.6.5, however, the last six observations in the dependent and independent variables for July through December 1984 were treated as missing. The results of this analysis are very similar to those discussed in Section 3.6.5.

Next, based on the results of the latter analysis forecasts were computed for the six missing months July through December 1984. In the calculation of these forecasts the observations for the petrol price and for the seatbelt law intervention were taken into account, but not the numbers of drivers KSI.

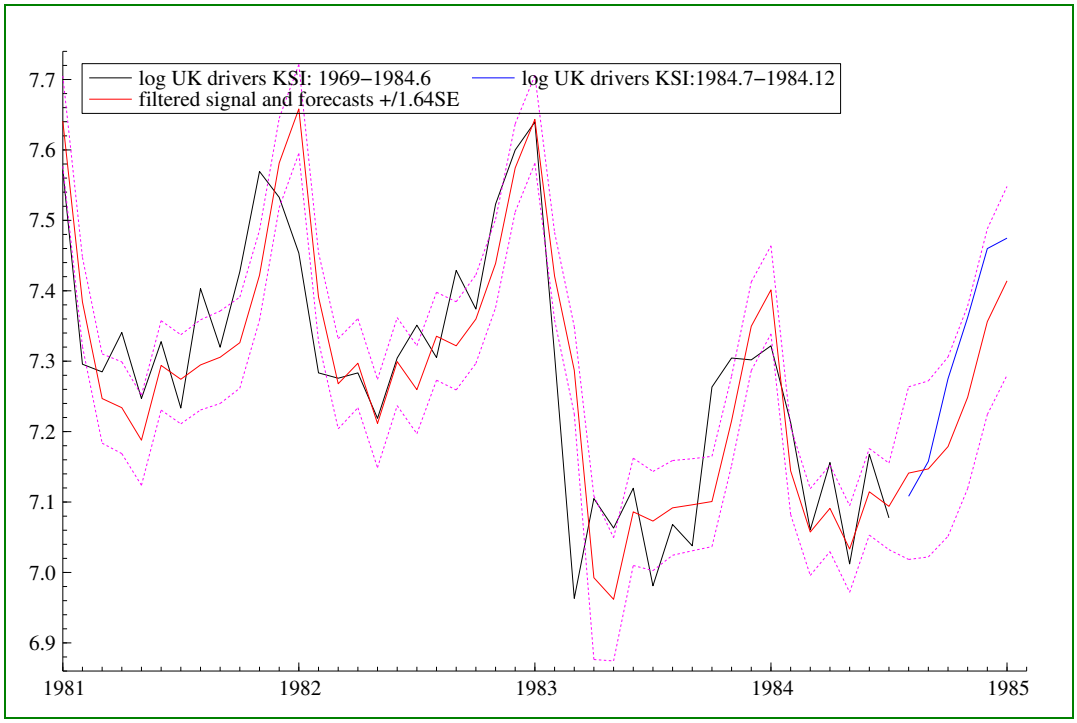


Figure 3.6.16: Filtered signal, and six months forecasts for the log of UK drivers KSI, including their 90% confidence limits.

The results are shown in Figure 3.6.16, which only contains the last four years in the series. Amongst others, the figure displays the filtered signal of the analysis (where the signal is the sum of the filtered state components) as well as the observation forecasts for the months July through December 1984 and the actual observations for the latter six months. Again, the 90% confidence limits become larger and larger as the forecasts are located further into the future. The figure also shows that the actual observations fall within the 90% confidence limits of the estimated forecasts, which is a good sign.

Finally, it is noted that there are a number of diagnostics that can be used to establish the goodness of fit of the predicted values to the observations. The *mean squared error* and the *mean absolute percentage error* of the forecasts obtained with the deterministic level and seasonal model are 0.0080695 and 0.010684, respectively; those obtained with the stochastic level and deterministic seasonal model are 0.0062978 and 0.00946457, respectively.

3.6.7 Conclusion on the state space technique

The examples from this section show that the state space analysis technique is appropriate for the purpose of descriptive analysis as well as explanatory analysis and forecasting in the framework of EU road safety research. As the other techniques described in this chapter, state space analysis assumes independent, homoscedastic, and normally distributed residuals. In state space modelling, stationarity of the data is not required; trend and seasonal are explicitly modelled.

State space analysis can easily handle missing data, which is very practical in road safety research. Furthermore, with a particular parameter setting state space analysis transforms into classical linear regression, which is a useful property with respect to explaining the technique and state space analysis results.

As was shown in this section, in fully deterministic, classical linear regression models the reliability of the *t*-tests for the regression coefficients is not guaranteed due to the fact that they do not take into account the time dependencies of the residuals. This can lead to incorrect conclusions about the significance of those regression coefficients. State space models do take into account the time dependencies, thus improving the reliability of the computed confidence and prediction intervals.

In this section, only univariate state space models, i.e. models with one dependent variable, have been discussed. The state space technique enables the modelling of multivariate time series problems. For example, it can be valuable to analyse the three important road safety components, i.e. exposure, accidents, and accident severity, in one model. In state space analysis this is possible; the components can be modelled simultaneously. Examples of multivariate state space models in the area of road safety can be found in Durbin and Koopman (2001), Commandeur and Koopman (in press), Bijleveld et al. (2005), De Blois et al. (in press), and Goldenbeld et al. (in press).

3.7 Equivalence between ARIMA and state space models

Jacques Commandeur (SWOV) and Ruth Bergel (INRETS)

1.1.1 Introduction

At first sight it may seem that the ARMA-type models and the state space models presented in Sections 3.4 and 3.6 are very different conceptually. When being fitted with ARMA models, time series that do not satisfy stationarity need to be first transformed into a stationary time series. In ARIMA models, a filter of differences is used as preliminary transformation of the original dataset: the trend and seasonal components are first eliminated by differencing before the actual analysis is performed. In state space methods, on the other hand, these two components are explicitly modelled.

However, as pointed out in Harvey (1989) and Durbin and Koopman (2001) ARMA and ARIMA models on the one hand and state space models on the other hand also have much in common. In this section we will focus on the similarities between the two approaches.
It is worth noting that, from a theoretical point of view, any ARMA representation of a stationary process has an equivalent state space representation. Nevertheless, due to the fact that stationary observations are usually not found in the road safety field, we will focus on the equivalencies between ARIMA models and state space models, and discuss them for two particular types of models described in Sections 3.4 and 3.6. For more equivalencies between ARIMA and structural time series models we refer to Appendix 1 in Harvey (1989).

1.1.2 The case of the local level model

As mentioned in Durbin and Koopman (2001), the local level model is equivalent to an ARIMA(0,1,1) model without constant:

$$\Delta y_t = (1 + \theta B)\eta_t ,$$

where B is the backshift operator defined by $B\eta_t = \eta_{t-1}$, Δ is the first-difference operator defined by $\Delta y_t = y_t - y_{t-1}$, θ is the unknown parameter, and η_t is a random process. Further, let σ_ε^2 and σ_ξ^2 denote the disturbance variances of the irregular and level components of a local level model, respectively (see Section 3.6.1), and let $q = \sigma_\xi^2 / \sigma_\varepsilon^2$. Then the equivalence between the parameters of the ARIMA(0,1,1) model and the local level model is given by

$$\theta = \frac{1}{2}\left[\sqrt{(q^2 + 4q)} - (q + 2)\right], \qquad (3.7.1)$$

and

$$\sigma_\eta^2 = -\sigma_\varepsilon^2 / \theta , \qquad (3.7.2)$$

where σ_η^2 is the error variance of the ARIMA(0,1,1) model.

Applying these formulas to the results obtained in the analyses of the log of the annual number of Norwegian road traffic fatalities series discussed in Sections 3.4.3 and 3.6.1, for example, and since $\sigma_\varepsilon^2 = 0.00326838$ and $\sigma_\xi^2 = 0.0047026$ in that case, we find that

$$q = 0.0047026 / 0.00326838 = 1.438816784,$$

and

$$\theta = \frac{1}{2}\left[\sqrt{(1.438816784^2 + (4)(1.438816784))} - (1.438816784 + 2)\right] = -0.3207071305.$$

Within rounding errors this value is equal to the parameter estimate θ = -0.32069194 obtained by applying the ARIMA(0,1,1) model without constant to the same data in SPSS (see Section 3.4.3). Also,

$$\sigma_\eta^2 = -\sigma_\varepsilon^2 / \theta = -0.00326838 / -0.3207071305 = 0.01019,$$

which value is - again within rounding errors - equal to the residual variance $\sigma_\eta^2 = 0.01050984$ obtained by applying the ARIMA(0,1,1) model without constant to the data in SPSS. Moreover, as a consequence, the forecasts obtained with an ARIMA(0,1,1) model are equal to those obtained with the local level model.

1.1.3 The case of the local linear trend with seasonal model

Letting *s* denote the periodicity of the seasonal, the local linear trend with seasonal model is equivalent to an ARIMA(0,1,1)(0,1,1)$_s$ model (also known as the "airline model"[65])

$$\Delta\Delta_s y_t = (1 + \theta B)(1 + \theta_s B_s)\eta_t,$$

when $\theta_s = -1$ and the disturbance variances for the slope and seasonal components of the local linear trend with seasonal model satisfy $\sigma_\zeta^2 = \sigma_\omega^2 = 0$, respectively. In that case, formulas (3.7.1) and (3.7.2) again apply. For example, the variances for the observation and level disturbances of the local linear trend plus seasonal model with deterministic slope and seasonal applied to the log of

[65] The so-called "airline model" was fitted on the monthly number of international airline passengers in thousands, for 1949-1960, series in Box & Jenkins, 1976).

the UK drivers KSI series are found to be equal to $\sigma_\varepsilon^2 = 0.00346757$ and $\sigma_\xi^2 = 0.0010011$, respectively (see Section 3.6.3). The value of *q* therefore equals

$$q = 0.001011 / 0.00346757 = 0.2887036167 ,$$

and substitution of this value in (3.7.1) yields

$$\theta = \frac{1}{2}\left[\sqrt{(0.2887036167^2 + (4)(0.2887036167))} - (0.2887036167 + 2)\right]$$
$$= -0.5879876639.$$

Applying the airline model without constant to the same time series in SPSS yields

$$\Delta\Delta_{12} y_t = (1 + \theta B)(1 + \theta_{12} B_{12})\eta_t = (1 - 0.58796298B)(1 - 0.89666905B_{12})\eta_t ,$$

and the value of the latter parameter θ is indeed remarkably close to the one obtained with (3.7.1), even though parameter θ_{12} is -0.9 instead of -1. Moreover,

$$\sigma_\eta^2 = -\sigma_\varepsilon^2 / \theta = -0.00346757 / -0.5879876639 = 0.005897 ,$$

which value is quite similar to the residual variance $\sigma_\eta^2 = 0.006434$ obtained by applying the ARIMA(0,1,1)(0,1,1)$_{12}$ model without constant to the series in SPSS.

1.1.4 Conclusion and discussion

In this section, two examples of equivalencies between ARIMA models and state space models, already described in Sections 3.4 and 3.6, were discussed. The necessary relationships between the model parameters were checked on the basis of their estimations provided by STAMP (for the state space models) and SPSS (for the ARIMA models).

It should be noted that, as these equivalencies only hold between well-defined specifications, other close specifications may in practice be retained. With the first example, it was demonstrated that the log of the annual number of Norwegian road traffic fatalities could equally be modelled with a local level model or with an ARIMA(0,1,1) model without constant; nevertheless, in practice, an ARIMA(0,1,1) with constant was retained in Section 3.4.3. With the second example, it was demonstrated that the log of the monthly number of UK drivers KSI could equally be modelled with a local linear trend with seasonal model or with an ARIMA(0,1,1)(0,1,1)$_{12}$ model without constant; nevertheless, in practice, an ARIMA(2,0,0)(0,1,1)$_{12}$ model was retained in Section 3.4.4.

It is also worth mentioning that not all ARIMA models have an equivalent in the subclass of state space methods which are collectively known as structural time series models. However, all ARIMA models can be put in state space form (see

Durbin and Koopman, 2001), thus making all the techniques that have been developed for state space models (like diffuse initialisation and the handling of missing values) available for ARIMA models also. Conversely, if required autoregressive components can be added to the structural time series models discussed in Section 3.6.

3.8 Conclusion time series analysis

Chris de Blois and Jaques Commandeur (SWOV)

Main concerns of EU road safety research are to improve insight in the development of road safety in the past and its underlying factors and to make forecasts of road safety in the future. Therefore, through the years much data on exposure, accident frequency and injury severity, as well as several characteristics of road users, vehicles and their use, roads, and accident management have been collected and as such various time series have been created.

Road safety data is voluminous and varied in the sense that several types of data and several dimensions are involved. The frequency of measurement of the road safety data varies: road safety data is mostly measured annually or monthly and sometimes weekly or even daily. Furthermore, the data comprises both national totals and disaggregated data for regions, for sections of the population (e.g. age classes, males, females, etc.), for vehicle types, or for road types among others. Between countries, but also between periods for the same country and between different types of data, there may exist large differences with respect to the availability, the periodicity, and reliability of (disaggregated) data.

The above-mentioned characteristics of the data and the different needs for analysing the several time series and their interrelations - i.e. monitoring, explaining, and forecasting - make road safety analyses complex and not straightforward. Furthermore, it appears that the time dependence structure of road safety developments often does not allow for the application of cross-sectional statistical techniques. As such, the application of dedicated state-of-the-art time series analysis techniques is advocated.

3.8.1 Summary of methods for time series analysis

In this chapter, several techniques used for the analysis of time series are reviewed. Below their main characteristics and their use for time series analysis in general or for road safety analysis in particular will be summarised.

Classical linear regression is a standard technique, which is frequently used for the analysis of time series because of its straightforwardness and efficiency. However, this technique does not properly consider the time dependencies between consecutive observations, nor does it consider alternatives for some other assumptions. Therefore, the residuals obtained with this technique do usually not satisfy the most important model assumptions, f.i. the assumption of independence. The latter problem may lead to statistical test results which are overoptimistic or too pessimistic about the relations between variables and also to poor forecasts, among others.

Generalised linear models can be used to overcome part of the restrictions of classical linear regression. This technique is more flexible than classical linear regression in the sense that it allows for all error distributions within the

exponential family of distributions. Among others, this family includes the normal distribution, which is the one assumed in classical linear regression, the Poisson distribution and the negative binomial distribution. Another extension in comparison with classical linear regression is that what is known as a *link function* can be defined to impose restrictions to the model output, which can be useful, for example, when the log-transformation is used to enforce positive forecasts.

By using *nonlinear models* even more restrictions of classical linear regression can be overcome. The biggest advantage of this technique over the previously mentioned is the broad range of functions that can be fit. Many processes, as in road safety, are inherently nonlinear. This flexibility of nonlinear regression is also a caveat, since similarly good fits can be obtained with very different functional forms, whereas presumably only one of them represents the real underlying process in the best manner. These different models can be adequate for interpolation purposes, but may produce very different predictions when used to extrapolate, i.e. to predict values outside the scope of the estimation dataset (forecasting).
Common advantage of the parametric linear and nonlinear regression models is the efficient use of data. Good estimates of the unknown parameters in the model can be produced with relatively small data sets. Another shared advantage is a fairly well-developed theory for computing confidence, prediction and calibration intervals.

However, for time series analysis the most important drawback of the classical linear, generalised linear and nonlinear regression models is that they do not naturally take into account the time dependencies between the consecutive observations of a time series. To adequately deal with these time dependencies, dedicated time series analysis techniques, such as ARMA (Auto-Regressive Moving Average) - type analysis, its special case DRAG (Demand for Road use, Accidents and their Gravity), and state space analysis, could be employed.

ARMA models (in the case of stationary data) and ARIMA models (in the most general case of non-stationarity data, which is the current case in road safety) enable to describe the dynamics of a process time and to extrapolate it in the future, without any call to additional variables and with the only assumption that the process dynamics will stay unchanged at the forecast's horizon. .Explanatory and intervention variables can also be included in ARMA and ARIMA models, and the additional corresponding regression coefficients can be estimated and ineterpreted.

For the analysis of road safety data, a disadvantage of ARIMA modelling may be its concept: the trend and the seasonal are removed before the modelling itself is performed on the stationary part of the process.. The emphasis is on describing the dynamics of this latter process, by means of estimating a small number of relevant coefficients parameters. This is sufficient for many applications.

The *DRAG model* is an application of a special case of the ARMA models, the AR (AutoRegressive) model with explanatory variables, specially designed for road safety analysis. The DRAG model has (at least) three levels: exposure, accident risk, and accident severity. The trend and the seasonal component are not

removed by filtering but are modelled by the introduction of numerous explanatory variables, whether related to exposure, economic factors, transitory factors, behavioural factors or road safety measures. The use of a particular non-linear transformation allows a flexible form of the link between the dependent variable and the explanatory variables.

The DRAG model has a powerful theoretical framework, but needs voluminous databases and therefore currently cannot appropriately be applied to EU road safety data.

In *state space models*, also known as *structural time series models* or *unobserved components models*, an observed time series is typically decomposed into a number of components. The level, the slope and the seasonal are assumed to be random components – effectively meaning that they may gradually change over time, which may be an important advantage for long time series - , and are estimated for obtaining an adequate *description* of an observed time series. Explanatory and intervention variables also help finding *explanations* for the observed development in the series.
Contrary to what is the case in ARIMA models, in state space modelling the trend and the seasonal are not removed but explicitly modelled. The focus here is on observing the development over time of the - usually unobserved - components, and mainly the development of the trend. Contrary to other decomposition techniques, the randomness of the trend is investigated, and described through its level and slope.

It should however be considered that the core methods used by the state space models and those used for the ARMA-type models have a lot in common if not are identical. As described in Section 3.7, many models have an 'identical twin' in the other approach, but with other parametrisations. This means that in practice, the identification process may end up with formally different but statistically indistinguishable models.

Concluding, EU road safety research requires the monitoring, explaining, and forecasting of road safety on the basis of often restrictedly available repeated measurements in time, with a high level of time dependency and possibly different frequencies of measurement. ARMA-type and state space models can be used for this purpose, for descriptive analysis as well as explanatory analysis and forecasting. It should be noted however that in cases f.i. violations of other assumptions, in particular a hierachical structure, the error distribution or a non-linear model function, may force the researcher to use balance the gravity of the violations, and chose other methods than desribed here (for instance non-linear time series methods).

3.8.2 Recommendations

For the descriptive, explanatory, or forecasting analysis of time series from road safety research, using dedicated time series analysis techniques such as ARMAtype and state space modelling is recommended.

To obtain a 'quick and dirty' insight in the data and the possible interrelations, classical linear regression but also generalised linear models and nonlinear regression can be used. However, the user of all these techniques should continuously be aware of the techniques' limitations and therefore never forget to test the model assumptions. As such, linear and nonlinear regression models can be used as a swift, first step in the analysis of road safety time series data, which should be followed by the application of more dedicated techniques as ARMA-type or state space analysis to obtain more valuable and reliable results.

In the manual (D7.5), these recommendations are followed by supplying manuals for classical linear regression (with an emphasis on the test of the model assumptions), ARMA-type, and state space analysis.

Chapter 4 - Conclusion

Heike Martensen and Emmanuelle Dupont (IBSR)

4.1 Analyzing complex data structures

The present document gave guidelines for analyzing complex data structures, as they commonly occur in traffic safety research. It departed from standard regression methods that were assumed to be known by the reader and focused on the assumptions that have to be met for these traditional methods to lead to valid conclusions. We presented a number of techniques that allow researchers conducting valid analyses even if the assumptions of the standard methods are violated. In particular, the document focused on the independence assumption, which will be recapitulated below.

In a classical linear regression model, an *observed* or *dependent* or *endogenous* variable y_i is predicted by one or more *explanatory* or *independent* or *exogenous* variables x_1, x_2... Such a relation is modelled by Equation 4.1, where *e* is the *error* term, also called the *disturbance* term and i=1…n, with n the number of persons.

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + e \qquad (4.1)$$

It has been shown that this simple regression model contains a number of restrictions, of which the main ones are

4. The dependent variable (*y*) has to follow the normal distribution.
5. The dependent variable (*y*) can be expressed as a linear function of the independent ones ($b_0+b_1x_1+b_2x_2$…)
6. The variance in the dependent variable that *cannot* be explained by this linear function (i.e. the error or disturbance term *e*) is independently distributed across all observations.

In practice, these assumptions are more often violated than not and we have demonstrated ways to deal with such violations: The Generalised Linear Model described in Sections 2.3.1 and 3.2.2 allows modelling observations that do not follow the normal distribution (e.g. discrete responses). In nonlinear models (Section 3.2.3), relations between dependent and independent variables are analysed that do not need to have the linear form (they can follow the exponential function, for example).

The present document is mainly focused on the third assumption, however, the assumption of independence of the error term[66]. It has been demonstrated that many datasets in traffic safety research tend to violate this independence

[66] This assumption is tested with the help of the model's residual, an estimation of the error term of Equation 4.1, computed once a sample of observations of the observed variable is available. It is worth mentioning that, in practice, the hypothesis of independence of the residuals is often referred to in place of the one of independence of the error term.

assumption and that the consequences of such violations can be particularly nasty. More specifically, they can lead to either under- or overestimation of the standard errors of the parameter estimates, which will in turn distort the estimated probability of having observed a particular effect on a purely coincidental basis. Both consequences, (1) accepting as significant a result that is actually not so, and (2) rejecting a result as due to chance that is in fact not due to chance, can occur in sometimes unpredictable ways, as has been demonstrated in the introduction (1.2.1 and 1.2.2).

4.2 Multilevel and time series modelling

Dependencies among the error terms occur often when data have a hierarchical (or nested) structure or are structured in time. *Hierarchically structured data* (or nested data) show random variation at more than one level of observation. Each data point is characterised by the membership to a particular group at each level of the hierarchy (e.g., passengers can be characterised by the car they were travelling in, by the road site at which the car was observed, by the area in which the road site is situated, etc.). Members of the same group tend to be more similar to each other than members of different groups. If this similarity is not represented in the analysis model, the errors (i.e. the part that is not explained by the model) for the members of a particular group also tend to be more similar to each other than to those from other groups. To avoid this, it is necessary to represent the hierarchical data structure in the analysis model. In Chapter 2, *multilevel modelling* is introduced as way to properly represent these structures and to deal with the arising dependencies.

Similarly, for data that develop over time (*times series*), data points can be characterised by time structures on different scales: decades, years, months, days. Between consecutive observations there may be a strong relation, and there can also be repeating patterns, for instance seasonal effects. The form of the relationship depends on many factors. For example monthly data are often most similar to data from the respective month a year earlier or later, while data from adjacent months can be very dissimilar. In almost any case, however, data from consecutive points in time are not independent. If these structures are not properly represented by the model, the errors will show the same dependencies as the data and therefore endanger the conclusion of the analysis. In Chapter 3, a number of *time series analysis* techniques have been introduced that allow for a detailed representation of these structures and therefore overcome the problems mentioned above.

It should be noted that probably the most important part of the model specification and assumptions in traffic safety analysis is the validity of the actual model equation. Misspecifications in the model equation (e.g., a linear model instead of a non-linear one, the choice of independent variables) can have consequences much more important than the violation of any of the further assumptions. Omitting the important explanatory variables may lead to attributing effects to the wrong variables, by absence of the truly influential variable(s). The validity of models however is dependent on the subject at hand and a general discussion on such matters is outside the scope of this document. This document systematically covers what can technically be done to have a valid model, while questions that

concern the content of the models are addressed by giving examples of model building. The application of each technique has been demonstrated on the basis of an empirical example from traffic research. Sometimes, different types of analysis have been illustrated for a single data set.

While the specific multilevel and time-series methods presented in Chapters 2 and 3 are summarized and evaluated in the local conclusions -- Sections 2.9 and 3.8. -- the examples will subsequently be briefly summarized in their order of their first appearance. The analyses performed on each dataset will be shortly described. The reader should remain aware however, that these are just *examples of analyses*. To safely draw conclusions from the results, more theoretical background and more information on the procedures underlying data collection would be necessary. The immediate goal of this description is to demonstrate how one should go about analysing data of a particular type, and how the results would be interpreted. The results presented here cannot in any way, however, serve as definite answer to particular road-safety questions.

4.3 Summary of empirical examples

In the introduction to multilevel modelling (Section 1.2.1), data from a Belgium *roadside survey on seatbelt use* were analysed in a single-level and a multilevel logistic regression analysis. These data were collected at randomly selected road sites in Belgium: For each passing car, it was determined whether the driver and (if present) the front passenger were using a seatbelt. The multilevel model was shown to be more appropriate, because the results showed a significant variation between road-sites in the probability of wearing a seatbelt. The speed limit could explain some of this variation (drivers on roads with higher speed limit have a higher probability of wearing a seatbelt), but not all of it. There was of course also significant variation among the inhabitants themselves, some of which could be explained by gender: Women tend to wear seat-belts more often than men.

In the introduction to time series (Section 1.2.2) the *yearly number of Norwegian fatalities* between 1970 and 2003 (or more specifically, their logarithm) were modelled in different versions of state space models. For the simplest version, which is in fact identical to a classical linear regression, the residuals were not independent. This problem was handled by allowing the intercept (also called level) to vary over time (stochastic intercept). The intercept on a time point t was not constant, but depended on the value of the intercept on the previous time point (t - 1). In that way the residuals are independent and we found a regression coefficient of -0.019860, which means that the number of Norwegian fatalities decreases each year with about 2%. The same dataset was also fitted with a pure ARIMA model (Section 3.4.3), in which case the assumption of independence of the error term, tested on the model's residuals, was accepted. The equivalence of the ARIMA model without constant term and the regression model with stochastic intercept described in the introduction, the so-called local level state space model (Section 3.6.1), was demonstrated on this particular dataset.

The dataset from a *speeding survey* in Belgium was used to demonstrate the use of basic two and three level linear models (Sections 2.2.1 and 2.2.2 respectively). It consists of a sample of drivers passing by a number of randomly selected road sites at which cameras were situated. Each driver's speed is measured as a continuous variable in km/h along with the car's length when passing by the road site. The relation between the length of a car and its speed, and the way it is affected by factors like the traffic count is demonstrated in order to illustrate the multilevel analysis of the relation between two continuous variables. Comparing a two and a three level model revealed that the data clearly had two levels: that of the single car and that of the road-site, while evidence for a third level (regions) was negligible, suggesting that in Belgium there is no substantial variation in speed between the different regions.

In a Belgian roadside *survey on drink driving*, all drivers passing within an hour were stopped at road-sites that were selected randomly with respect to location and point in time. Together with a number of potential explanatory variables, it was established whether the driver's BAC (breath alcohol concentration) was below 0.05 mg per litre (the legal limit), between 0.05 and 0.08 mg, or above. One way to analyse these data is to dichotomize them by joining the two higher categories and simply differentiating between blood-alcohol concentrations under or above the legal limit. These data so aggregated can be analysed by means of a logistic regression analysis, as illustrated in Section 2.3.2. In Section 2.3.3, it is demonstrated that one can also analyse the original three response categories using a multinomial regression model. Both an unordered category-model and an ordered one were estimated. The results however, provided no reason to question the ordered nature of the response categories, a case in which the ordered model is to be preferred given its parsimonious quality. At the test-site level the time of testing was the most important predictor as drink driving on weekend nights by far exceeds that at all other time points by far. At the individual level gender and age were the most notable predictors with men between 40 and 54 having the highest risk of drink driving.

In a Greek study on the *effects of speed infringements and alcohol controls*, the accident and fatality number for each county were collected over a period of 5 years. The yearly data were analysed in *multilevel poisson-family regression models*, including poisson, extra-poisson and negative binomial models (Section 2.3.4, with accidents of counties nested into regions at a higher level). It turned out that both enforcement measures were highly correlated (i.e. counties that executed many alcohol controls also issued many speeding infringements), and that they together lead to a significant decrease in fatalities. Moreover, it was shown that there was significant regional variation in the number of accidents and in the related effect of the enforcement measures. In particular, it was shown that the enforcement measures were the most effective in those regions that had the highest accident rate in the first place.

The same data set was analysed in a multivariate multilevel model (Section 2.5) allowing investigating the effects of enforcement measures on two road safety outcomes (fatality and accident numbers) simultaneously. It was shown that the two outcomes are correlated, and part of their covariance is situated at the regional level. It was also demonstrated that enforcement had a stronger overall

effect on the number of fatalities than on the number of accidents as such, suggesting that the accidents became less severe. Moreover, the significant regional variation of the effect of enforcement on accidents was confirmed, whilst the corresponding variation of the effect on accidents was non significant. It can be said that enforcement has an important overall effect on fatal accidents, which result from more risky behaviours. This effect is uniform in all regions, because drivers perceived an increased nationwide presence of the Police and improved their overall behaviour accordingly; however, the decrease of non-fatal accidents (which result from less risky behaviours) may be more or less important in different regions, according to the local enforcement practices.

A dataset of *monthly fatalities and severe injuries in Greece* were analyzed in relation to enforcement and vehicle ownership, by means of generalized linear models (poisson, extra-poisson, negative binomial, Section 3.2.2.). Vehicle ownership was used as an offset term, in order to model the rates of fatalities and serious injuries per number of vehicles, rather than the fatalities and serious injuries counts themselves. An intuitive negative coefficient between the number of alcohol controls and the number of persons killed and seriously injured in road accidents was identified. This shows that the intensification of enforcement in Greece in the examined period brought an important road safety benefit, in terms of reduction of monthly fatalities and serious injuries, also accounting for the related exposure.

Using the *relationship of driver characteristics and their acceptance of new technologies in traffic* based on data from SARTRE 3, the chapter on structural equation models (2.6) shows the basic form of such models in the multilevel case, dealing mainly with assumptions on data. The chapter also discusses the necessary theoretical concepts of these models.

In the chapter on linear regression models, the *decrease in road accident fatalities in Austria* is modelled leading to significant results. However the focus of this chapter is not on these results but on the underlying assumptions, which are analyzed in detail. In the examples shown, especially the most important assumption of randomly distributed errors was clearly violated, implying that the results of the statistical tests regarding the regression could not be trusted.

Aggregate data on *road accidents, vehicle fleet and population of 17 European countries* for the period 1970-2002 were used in a non-linear time series model (Section 3.2.3). Smeed's original formula on the macroscopic relationship between accidents, vehicles and population was examined and further developed in two additional forms: an auto-regressive and a log-transformed. The analysis of the estimated parameters allowed for a general assessment of the prevailing road safety patterns in the EU. In particular, the least safe countries among the countries examined today appear to be Greece and Portugal, while the United Kingdom and the Netherlands are two of the safest countries in Europe. These results are in accordance to the general trends evidenced by the literature.

The *monthly number of killed and seriously injured drivers registered in the UK* (UK-KSi drivers) for the period January 1969-December 1984 was fitted with an ARIMA model, in which the petrol price and an intervention variable for taking account for the introduction of the seat-belt law (see Harvey, Durbin, 1986)., were introduced and turned out to be significant The assumption of independence of the error term tested on the models residuals was accepted, and the model's performance increased about 5% with the introduction of the additional variables. The main objective of the analysis was the assessment of this road safety measure in the UK, and it was demonstrated that the law caused a reduction of 15% of the number of KSI drivers from February 1983 onwards.
.
The same data was used as an example for a deterministic seasonal model in the State Space Section. It was not possible to find a seasonal model that met the requirements of independence, homoscedasticy and normality of the residuals. Explanatory variables were added, namely seatbelt law (intervention variable) and petrol price. The appropriate model was a deterministic seasonal model with a deterministic level and no slope. The results confirmed those obtained with ARIMA and showed that the seatbelt law resulted in a 21.1% reduction of the number KSI in the UK. Moreover a 1% raise in the petrol price was associated with a 0.28% reduction of the number KSI.

An ARIMA-type analysis, similar to that on the UK drivers was conducted on the monthly total number of *French fatalities* collected between 1975 and 2001: it was shown that next to various seasonal and economic variables, the number of fatalities is also affected by certain media events. In particular, the presidential amnesty that is usually given to traffic offenders during the French elections appeared to be associated to an increase in fatalities. At the same time, the effect of a fatal accident that received much attention in the media (a young woman, Anne Cellier, who was killed by a drunk driver) was taken into account. The results suggested that fatalities increased by 6.4% per month on average during the 10 months preceding the first presidential amnesty in 1988 - and by 3.8% respectively during the 7 months preceding the second one in 1995. In absolute numbers, more than 500 deaths could thus have been attributed to the presidential amnesty in 1988. To the contrary, the attention that the media devoted to the Cellier case seems to have saved lives: The results suggested that fatalities decreased by 6.1% per month on average during the 7 months following this tragic accident.

The same analysis was extended to the monthly number of *injury accidents and fatalities on French A-level roads* and motorways, between 1975 and 2001. Similar results were obtained, and, as risk exposure is precisely measured on these two networks, significant results related to risk factors, namely the traffic volume and specific weather variables measuring temperature, rain and frost, could also be established.

Similar to the Norwegian fatalities modeled in the introduction (1.2.2) and first part of the state space chapter (3.6), the annual *number of Finnish fatalities* was modeled with a state space model. In contrast with the model for the Norwegian fatalities, the Finnish fatalities were best modelled with a constant intercept (or deterministic level) and a stochastic slope. This had been determined by fitting the data with both stochastic slope and level. The small variance of the level indicated

that the level could better be deterministic. Model comparison indicated that indeed the state space model with the deterministic level and stochastic slope fitted the data the best. The stochastic slope varied between 0.05 and -0.10, which means that the number of Finnish fatalities changed between 5% and -10% a year.

All empirical examples given in this document and some of the technical aspects that they were used to demonstrate are summarized in Tables 4.1 and 4.2. Table 4.1 presents the multilevel modelling examples (chapter 2) and Table 4.2, the time-series ones (chapter 3).

Statistical method	Example	Response variable
Basic two-and three level model	Speeding survey in Belgium	Normal
Discrete responses	Drink-driving survey in Belgium	Binary
Discrete responses	Drink-driving survey in Belgium	Categorical
Discrete responses	Effect of enforcement or accidents in Greece	Counts
Multivariate model	Effect of enforcement on accidents and fatalities in Greece	Counts
Repeated Measurements	Driving skills in young drivers (simulated)	Normal
Factor analysis	Attitudes on driving style and on technical devices.	Normal

Table 4.1: Summary of empirical examples for multilevel analyses

Statistical method	Example	Response variable
Linear regression	Austrian fatalities	Normal
GLM	Monthly variation of the effect of enforcement on road accidents in Greece	Counts
Non Linear	Macroscopic relation between accidents, population and vehicle fleet in the EU	Normal
ARMA-type models	Norwagian fatalities	Normal
	Drivers killed and seriously injured in the UK	
	French fatalities	
State space analysis	Norwegian fatalities	Normal
	Drivers killed and seriously injured in the UK	
	Finland fatalities	

Table 4.2: Summary of empirical examples for time series analyses.

As the trend in traffic-safety research is towards large databases containing data from several countries and consecutive years, many datasets have the structure of panel data, in which hierarchical and time-series structures co-occur. For example, accident-counts can be characterised by the regions and the countries they were taken in and by the points in time - such as years or months - at which the accidents happened. As an example, the aggregate yearly Greek fatality data show a multilevel structure and in Section 2.3.4 it has been shown that the effect of enforcement measures varies across regions. However, these data also form a time-series, especially when they are disaggregated over months. Therefore, they are also analysed as a time-series in a Generalized Linear Model in Section 3.2.2. In the area of spatial modelling (see Section 2.3.4) there are now first approaches

to modelling the hierarchical structure and the time-structure simultaneously (Aguero-Valverde & Jovanis, 2006). We are, however, not aware of existing models at the time of writing that allow the inclusion of stochastic components for hierarchical as well as time-structures. Multilevel modelling and time series analysis are both very active areas of research and development. It is probably possible already to combine stochastic components for space and time if one uses flexible software. It will, however,, probably take some time before these combinations become available in standard software packages (see the manual D7.5 for an overview of multilevel (2.1) and time series analysis (3.1) in various types of software).

As these examples illustrate, often there is not one correct method to analyse a particular set of data, with all others being incorrect. Panel data, for instance, are often modelled with time-series analyses – aggregating over possible hierarchical structures or with multilevel models – aggregating over several points in time. Analysing the data in different ways is often a good starting point. In each case one should be aware of the assumptions that do not hold in the particular model of analysis. The final choice for one model cannot be prescribed by a general recipe. One has to carefully weigh the advantages and disadvantages of either procedure and, depending on the research question and the exact characteristics of the data set (e.g. the amount of variation between higher-level units and between time points), a choice has to be made.

4.4 Outlook

The methods presented in this document will be applied in analyses on European road safety data, in particular accidents data, exposure data, and safety performance indicators. The data and the statistical methods will serve answering questions concerning either macroscopic or microscopic data.

Macroscopic data concern the CARE (Community Road Accident) data-base. In this database registering accidents from all EU-member since 1990, there is a clear hierarchical structure (accidents can be characterised by the regions and the countries they took place in) and also a time-series structures as the accidents can be characterised by the point in time (e.g., the year, the month) at which they happened. This data-base offers a wealth of information on each accident and can therefore be aggregated in very different ways, tailored to the particular road safety aspect that needs to be addressed (e.g., county, region, road-type, accident type, vehicle type, participant type, etc.). The research question can be very broad (e.g. did the fatalities in a particular country decrease at the same rate as those in other countries?) or very specific (e.g., did a particular junction become safer after reconstruction?). *Multilevel analyses* allow for the introduction of exposure data and data about safety performance indicators, even if those are not specified at the same level of disaggregation as the accident data themselves. In this way, multilevel analyses allow a global and detailed approach simultaneously. *Time series analyses* allow describing the development over time, relating the accident-occurrences to explanatory factors such as exposure measures or safety-performance indicators (e.g., speeding, seatbelt-use, alcohol, etc), and forecasting the development into the near future.

Microscopic analyses (addressing questions like, did the type of baby-seat affect the risk of young children being killed in an accident?) require in-depth accident data and allow detailed analyses of factors that contribute to the severity of injuries. This type of data involves a high level of detail and is inherently structured in a hierarchical way describing the accident process (persons are nested into vehicles; vehicles are nested into accidents, etc.) Moreover, accidents can be clustered according to geographical or administrative units. In-depth accident data therefore readily call for detailed multilevel analysis.

4.5 In sum

Multilevel modelling and time series analyses form two powerful tools that can help researchers analysing complex data structures that violate the assumptions posed by traditional analyses. A number of empirical examples demonstrated that many (if not most) data sets in traffic safety research are

hierarchically structured and/or form a time-series. Multilevel modelling and time series analysis allow the proper representation of the hierarchical structure of data and their development over time. This representation is crucial to answer questions about these structures themselves, and forms the basis for a proper investigation of possible other factors, allowing experts in road safety to identify different kinds of risk factors and to propose effective and objective policy decisions.

References

Aguero-Valverde J., Jovanis P. P. (2005). Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis and Prevention Vol. 38, pp. 618-625.

Aitkin, M. A. and Longford, N. T. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society A, 149*, 1-43.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory (B. Petrox and F. Caski, eds.), 267–281. Akademia Kiado, Budapest. (Reprinted in Breakthroughs in Statistics, eds Kotz, S. & Johnson, N. L. (1992), volume I, pp. 599–624. Springer, New York.

Amoros, E, Martin, J., & Laumon, B. (2003). Comparison of road crashes incidence and severity between some French counties. *Accident Analysis and Prevention*, Vol 35, pp. 537-547.

Anselin, L. (1995). Local indicators of spatial association - LISA. Geographical Analysis, 27, pp. 93-115.

Barnett, V. (1999). *Comparative Statistical Inference.* Wiley, New York.

Bates, D. M., and Pinheiro, J. C. (1998). *Computational methods for multilevel modelling*. Madison, University of Wisconsin.

Bates, D. M., Watts, D. G. (1988), *Nonlinear regression analysis and its applications*, John Wiley & Sons, Inc.

Belsley, D., Kuh, E., and Welsch, R. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley and Sons, New York.

Bergel R., Depire, A. (2004). *A functional form for an aggregate model of road risk.* Actes Inrets n° 90 du groupe de travail 2001 du Séminaire Modélisation du Trafic, Arcueil , mai 2004.

Bergel R., Depire, A. (2004). *Climate, road traffic and road risk - an aggregate approach*. Proceedings of the WCTR'04, Istanbul, 4 - 8 July 2004.

Bergel, R., Girard, B. (2000). The RES Model by road type in France . In : Gaudry M., Lassarre S. *Structural Road Accident Models - The International DRAG Family*, Pergamon, *2000*

Bijleveld, F.D., Commandeur, J.J.F., Gould, P.G., and Koopman, S.J. (2005). *Model-based measurement of latent risk in time series with applications.* Tinbergen Institute Discussion Paper TI 2005-118/4, Tinbergen Institute, Amsterdam.

Booth, J. G., and Sarkar, S. (1998). Monte Carlo approximation of bootstrap variances. *American Statistician*, 52, p. 354.

Box G.E.P., Cox D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, B(2):211-243.

Box G.E.P., Tiao G.C. "Intervention analysis with applications to economic and environmental problems*". Journal of the American Statistical Association*, 1975,70,349,pp70-79

Box, G. E. P. & Jenkins, G.M. (1976). *Time series analysis: forecasting and control*. Revised Edition. Oakland, CA: Holden-Day.

Box, G. E. P., and Pierce, D. A., (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of American Statistical Association*, 65, pp. 1509-1526.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time series analysis. Forecasting and control*. Prentice Hall International, Inc., New Jersey.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models, *Applied Statistics*, Vol 33, pp. 38-44.

Brockwell P.J., Davis R.A. (1986) *Time series : theory and methods*, second edition, Springer Verlag

Brockwell P.J., Davis R.A. (1998) *Introduction to time series and forecasting*, Springer Verlag

Browne W, Goldstein H, Rabash J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling* 2001;1, pp.103–24.

Burns, N. R., Nettelbeck, T., White, M., & Willson, J. (1999). Effects of car window tinting on visual performance: a comparison of elderly and young drivers, *Ergonomics*, Vol. 42, pp. 428-443.

Cameron, M. H., Haworth, N., Oxley, J., Newstead, S., & Le, T. (1993). *Evaluation of Transport Accident Commission road safety television advertising*. Report No.52, Monash University Accident Research Centre.

Campbell, M. J. (1994). Time Series Regression for Counts: An Investigation into the Relationship between Sudden Infant Death Syndrome and Environmental Temperature. *Journal of the Royal Statistical Society*, Series A (Statistics in Society), Vol. 157, No. 2, pp. 191-208.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicin*, 19, pp. 1141-1164.

Carrière, I. & Bouyer, J. (2002). Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons. *BMC Medical Research Methodology*, 2:15.

Casella, G, and George, E. (1992). Explaining the Gibbs sampler. *American Statistician*, 46, 3, pp. 167-174.

Catchpole, J. E., Macdonald, W. A., and Bowland, L. – Monash University Accident Research Centre (1994). Young driver research program – The influence of age-related and experience-related factors on reported driving behaviour and crashes. Available: www.monash.edu/muarc/reports/atsb143.pdf. (Accessed may, 2006).

Christ, R., Delhomme, P., Kaba, A., Makinen, T., Sagberg, F., Schulze, H., Siegrist, S. (1999). *GADGET Guarding Automobil Drivers through Guidance Education and Technology. Final Report.* Kuratorium fuer Verkehrssicherheit (KfV) Vienna.

Cochran, W. G. (1963). *Sampling Techniques*, second edition, New York: John Wiley & Sons.

Comité d'Accompagnement des Etats Généraux de la Sécurité Routière (2001). Rapport du comité d'accompagnement des EGSR au comité de pilotage : Dossier 4 – Apprentissage de la conduite. Available : http://bivvprint/intranet /FR/Docs/EGSR/dossier%204%20apprentissage%20conduite.doc (accessed may, 2006).

Commandeur, J.J.F., and S.J. Koopman (in press). *An introduction to time series analysis by state space methods.*

COST 329. *Models for traffic and safety development and interventions*. European Commission, Directorate general for Transport, Brussels , final report of the Action, 2004.

Dagum, C. 1980. The generation and distribution of income,the Lorenz curve and the Gini ratio. *Economie Appliquée 33*, pp. 327-367.

Davies, N., Triggs, C. M., and Newbold, P. (1977). Significance levels of the Box-Pierce portmanteau statistic in finite samples. *Biometrika*, 64, pp. 517-522.

Davis, G. A. (2000). Estimating Traffic Accident Rates While Accounting for Traffic-Volume Estimation Error: A Gibbs Sampling Approach. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1717, TRB, National Research Council, Washington, D.C., pp. 94-101.

Davis, G. A. and Guan, Y. (1996). Bayesian Assignment of Coverage Count Locations to Factor Groups and Estimation of Mean Daily Traffic. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1542, TRB, National Research Council, Washington, D.C., pp. 30-37.

Davis, G. A. and Yang S. (2001). Bayesian Identification of High-Risk Intersections for Older Drivers via Gibbs Sampling. *Transportation Research*

Record: Journal of the Transportation Research Board, No. 1746, TRB, National Research Council, Washington, D.C., pp. 84-89.

Davis, R., Dunsmuir, W., & Wang, Y. (2000). On autocorrelation in a Poisson regression model. *Biometrika*, Vol. 87, No. 3, pp. 491-505.

De Blois, C.J., Goldenbeld, Ch. and Bijleveld, F.D. (in press). *Modelling and exploring moped-car KSI crashes.* SWOV, Leidschendam.

Dean, C. (1992). Testing for overdispersion in Poisson and binomial regression models. *J. Amer. Statist. Assoc.* 87, pp. 451-457.

Dean, C., Lawless, J.F. (1989). Tests for detecting overdispersion in Poisson regression models. *J. Amer. Statist. Assoc*. 84, pp. 467-472.

Delhomme, P., Vaa, T., Meyer, T., Harland, G., Goldenbeld, C., Järmark, S., Christie N., & Rehnova, V. (1999). *Evaluated Road Safety Media Campaigns: An Overview of 265 Evaluated Campaigns and Some Meta-Analysis on Accidents*, Paris: INRETS.

Diez-Roux, A. (2002). A glossary for multilevel analysis. *Journal of Epidemiology and Community Health, 56*, 588-594

Dobson, A.J. (1990), *An Introduction to Generalized Linear Models*. Second edition, Chapman and Hall, London.

Doornik, J. A. (2001). *Object-oriented matrix programming using Ox 3.0*. London: Timberlake Consultants Press.

Duncan, C., Jones, K., Moon, G., (1999). Smoking and deprivation: are there

Durbin, J. & Koopman, S.J. (2001). *Time series analysis by state space methods.* Oxford: Oxford University Press.

Efron, B (1982). *The Jackknife, the bootstrap and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics. CBMS-NSF Monographs, 38.

Enders W. *Applied econometric time series*, Wiley, 1995.

Fisher, R. A. (1934), *Two new properties of mathematical likelihood.* Proceedings of the Royal Society A, 144, pp 285-307.

Franken, I.H.A, Rosso, M., Van Honk, J. (2003) Selective memory cues in alcoholics and its relation to craving. *Cognitive Therapy Research, 27(4)*, pp. 481-488.

Gaudry M. (1984). *DRAG, un modèle de la Demande Routière, des Accidents et de leur Gravité, appliqué au Québec de 1956 à 1982*, Publication 359, Centre de Recherche sur les Transports, Université de Montréal.

Gaudry M. *(2002). DRAG, a model of the Demand for Road Use, Accidents and their Severity, applied in Quebec from 1956 to 1982*. Publication 17, Agora Jules Dupuit, Université de Montréal (revision of Gaudry 1984).

Gaudry M., Lassarre S. (2000). *Structural Road Accident Models - The International DRAG Family*, Pergamon.

Geman, S. and D. Geman. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721-741.

Gharaybeh, F. A. (1994). Application of Smeed's formula to assess development of traffic safety in Jordan, *Accident Analysis and Prevention*, Volume 26, Issue 1, pp. 113-120.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London, Chapman and Hall.

Gill, J. (2000), *Generalized Linear Models: A Unified Approach*, Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-134, Thousand Oaks, CA: Sage.

Goldenbeld, Ch., De Blois, C.J., Bijleveld, F.D., and A.L. van Gent (in press). *Modelling and exploring pedestrian-car crashes.* SWOV, Leidschendam.

Goldstein H, & Rasbash J. (1996) Improved approximations for multilevel models with binary responses, *Journal of the Royal Statistical Society A*, Vol. 159, pp. 505-513.

Goldstein H. (2003). *Multilevel statistical models*, London: Arnold.

Goldstein, H., and Woodhouse, G. (2001). Modelling repeated measurements. In A. H. Leyland & H .Goldstein (Eds.), Multilevel modelling of health statistics, Chichester: Wiley.

Goldstein, H., Rasbash, J., Browne, W., Woodhouse, J., Poulain, M. (2000). Multilevel Models in the Study of Dynamic Household Structures. *European Journal of Population* 16: 373–387, 2000.

Good, P. I. (1999). *Resampling methods: a practical guide to data analysis*. Birkhauser, Boston/Berlin.

Gourieroux, C. & Monfort, A. (1990). Séries temporelles et modèles dynamiques. *Economica, 1990*.

Hakim S., Hakkert S., Hochermann I., Shefert D. (1991) **"**A critical Review of macro models for road accidents**"**. *Accident Analysis & Prevention.* Vol. 23, N° 5, pp. 379-400 .

Harvey A.C. (1989). *Forecasting structural time series and the Kalman filter*. Cambridge University Press, Cambridge, 1989

Harvey A.C., Durbin J. (1986). "The Effects of Seat Belt Legislation on British Road Casualties": A Case Study in Structural Time Series Modelling ", *J. R. Statist. Soc.*, Vol. 3 n°149 pp.187-227.

Harvey, A. C., (1993). *Time Series Models,* Second edition, MIT Press.

Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter.* Cambridge: Cambridge University Press.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97-109.

Hauer, E. (1986). On the Estimation of the Expected Number of Accidents. *Accident Analysis and Prevention*, Vol. 18, No. 1, pp. 1-12.

Hauer, E. (1996a). Detection of Safety Deterioration in a Series of Accident Counts. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1542, TRB, National Research Council, Washington, D.C., pp. 38-43.

Hauer, E. (1996b). Identification of Sites with Promise. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1542, TRB, National Research Council, Washington, D.C., pp. 54-60.

Hauer, E. (1996c). Statistical Test of Difference Between Expected Accident Frequencies. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1542, TRB, National Research Council, Washington, D.C., pp. 24-29.

Hauer, E. (1997). *Observational Before-After Studies in Road Safety*. Pergamon, Oxford, U.K..

Hauer, E. and B.N. Persaud (1987). How to Estimate the Safety of Rail-Highway Grade Crossings and the Safety Effects of Warning Devices. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1114, TRB, National Research Council, Washington, D.C., pp. 131-140.

Hauer, E., B. Allery, J. Kononov, and M. Griffith (2004). How Best to Rank Sites with Promise. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1897, TRB, National Research Council, Washington, D.C., pp. 48-54.

Hauer, E., D. W. Harwood, F. M. Council, and M. S. Griffith (2002a). Estimating Safety by the Empirical Bayes Method: A Tutorial. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1784, TRB, National Research Council, Washington, D.C., pp. 126-131.

Hauer, E., J. Kononov, B. Allery, and M. S. Griffith (2002b). Screening the Road Network for Sites with Promise. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1784, TRB, National Research Council, Washington, D.C., pp. 27-32.

Hauer, E., Jerry C.N. Ng, and J. Lovell (1988). Estimation of Safety at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1185, TRB, National Research Council, Washington, D.C., pp. 48-61.

Heck R. H., Thomas S. L. (2000). *An introduction to multilevel modeling techniques*, Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Hedeker, D. (2005). Longitudinal data analysis – Reading materials, overheads, examples, and problem sets [On Line]. Available: http://www.uic.edu/classes/bstt/bstt513/index.html (accessed may, 2006).

Henning-Hager, U. (1986). Urban development and road safety. *Accident Analysis and Prevention*. 18(2):135-45.

Hewson, P.J. (2005). Epidemiology of child pedestrian casualty rates: Can we assume spatial independence? Accident Analysis and Prevention Vol. 37, pp.651-659.

Hill, P. W. and Goldstein, H. (1998). Multilevel modelling of educational data with cross classification and missing identification of units. *Journal of Educational and Behavioural statistics* 23: 117-128.

Hox, J., (2002). *Multilevel Analysis, Techniques and Applications*. Lawrence Erlbaum Associates, Publishers, London.

Hsiao, C. & Pesaran, M. H. (2004). Random Coefficient Panel Data Models. *Institute for the Study of Labor (IZA), Discussion Papers, 1236.*

INRETS, INVS, SETRA, LAB, Champagne-Ardennes University (2002) *Presidential Amnesty and road safety*. Collective expertise report.

Jacobs, G. D. (1986). Road accident fatality rates in developing countries-a reappraisal.. In: *PTRC. Summer Annual Meeting*, University of Sussex, 14-17 July 1986., Proc of Seminar H. London: PTRC Education and Research Services, 107-119.

Jones A. P., Jørgensen S. H. (2003). The use of multilevel models for the prediction of road accident outcomes. *Accident Analysis and Prevention*, Vol. 35, pp. 59-70.

Jones, K. (1993) Using multilevel models for survey analysis. *Journal of the Market Research Society*, Vol. 35, 3, pp. 249-265.

Kish L. (1965). *Survey Sampling*, New York : John Wiley & Sons, Inc.

Koopman, S. J., Shephard, N., & Doornik, J. A. (1999). Statistical algorithms for models in state space using SsfPack 2.2. *Econometrics Journal*, Vol. 2, p.113-166.

Koornstra, M. J. (1992) The evolution of road safety and mobility, *IATSS Research*, 16: 129-148.

Koornstra, M. J. (1997) Trends and forecasts in motor vehicle Kilometrage, road safety, and environmental quality, pp: 21-32 in Roller, D., (ed.) The motor vehicle and the environment – Entering a new century. Proceedings of the 30th International Symposium on Automotive Technology & Automation, Automotive Automation Limited, Croydon.

Kreft, I. G. G. (1994). Multilevel models for hierarchically nested data: Potential applications in substance abuse prevention research, in *Advances in Data Analysis for Prevention Intervention Research*, Ed. LM Collins, LA Seitz, Research Monograph 142, National Institute on Drug Abuse, Washington DC, pp. 140-183.

Kreft, I., De Leeuw, J., (1999). *Introducing multilevel modelling*. SAGE Publications.

LABS, INRETS, INVS. *Presidential amnesty and road safety*. Report of expertise, 2001.

Langford, I, Bentham, G, McDonald (1998), A, "Multi-level modelling of geographically aggregated health data: a case study on malignant melanoma mortality and UV exposure in the European Community", *Statistics in medicine, vol. 17*, 41–57.

Langford, I. H., & Day, R. J. (2001). Poisson Regression. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics*. Wiley series in probability and statistics. Chichester: Wiley.

Langford, I.H., Leyland, A.H., Rasbach, J., Goldstein, H., (1999). Multilevel modelling of the geographical distribution of diseases. *Applied Statistics* 48 part 2, pp. 253-268

LaScala E.A., Gerber D., Gruenewald P.A. (2000). Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. Accident Analysis and Prevention Vol. 32, pp. 651-658.

Lassarre S. (2001). Analysis of progress in road safety in ten European countries. *Accident Analysis & Prevention*, Vol. 33 pp.743-751, *2001*

Lassarre S., (1994). *Cadrage méthodologique d'une modélisation pour un suivi de l'insécurité routière,* Synthèse INRETS, n°26, Arcueil.

Lee & Nelder (2001). Hierarchical generalized linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika 88*, 987-1006.

Lee, H.T., Yoder, J.K., Mittelhammer, R.C., & McCluskey, J. (2006). A random coefficient autoregressive markov regime switching model for dynamic futures hedging. *The Journal of Future Markets, 26*, 2, 103-129.

Levy P. S., Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*, third edition, New York: John Wiley & Sons.

Leyland, A. H, Goldstein, H. (2001). *Multilevel Modeling of Health Statistics*, West Sussex, England: John Wiley & Sons, Ltd.

Lindsey, J. K. (1993). *Models for repeated measurements*, Oxford: Clarendon Press.

Ljung, G. M., and G. E. P. Box (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, pp. 297-303.

Longford, N. (1993) *Random coefficient models*, Oxford: Clarendon Press.

MacNab, Y. (2004). Bayesian Spatial and Ecological Models for Small-area Accident and Injury Analysis. *Accident Analysis and Prevention*, Vol. 36, No. 6, pp. 1019-1028.

Maher, M. J., & Summersgill, I.(1996). A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis and Prevention,* Vol. 28(3), pp. 281-296.

Maycock, G., and Hall, R. D. (1984). "Accidents at 4-Arm Roundabouts." TRRL Laboratory Report 1120, Transport and Road Research Laboratory, Crowthorne, Berkshire, UK.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*. Second edition. Chapman Hall, New York.
McCulloch, C.E. & Searle, S.R., (2001). *Generalized, Linear, and Mixed Models.* Wiley Series in Probability and Statistics.

McMillan G.P., Hanson T.E., Lapham S.C. (2007). Geographic variability in alcohol-related crashes in response to legalized Sunday packaged alcohol sales in New Mexico. Accident Analysis and Prevention Vol. 39, pp. 252-257.

Meliker J.R., Maiob R.F., Zimmerman M.A., Kim H.M., Smith S.C., Wilson M.L., (2004). Spatial analysis of alcohol-related motor vehicle crash injuries in southeastern Michigan. Accident Analysis and Prevention Vol. 36, pp. 1129-1135.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21:1087-1091.

Miaou, S. (1994). *The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions.*

Proceedings of the 73rd Annual Meeting of the Transportation Research Board, Washington, D.C.

Miaou, S. and D. Lord (2003). Modelling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes Methods. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1840, TRB, National Research Council, Washington, D.C., pp. 31-40.

Miller, H. J. (2004) "Tobler's First Law and spatial analysis". Annals of the Association of American Geographers, 94, pp. 284-289.

Mooney, C. Z., and Duval, R. D. (1993). *Bootstrapping. A nonparametric approach to statistical inference*. Sage: Newbury Park, CA.

Mountain, L., Maher, M, & B. Fawaz (1998). The influence of trend on estimates of accidents at junctions. *Accident Analysis and Prevention*, Vol 30, No. 5, pp. 641-649.

neighbourhood effects? *Social Science and Medicine*, 48, p.497-506.

Nevitt, J. and Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modelling. *Structural Equation Modelling*, 8, 3, pp. 353-377.

Newstead, S.; Cameron, M. H; Gantzer, S. & Vulcan, P. (1995). *Modeling of some major factors influencing road trauma trends in Victoria 1989 - 93*. Report No. 74, Monash University Accident Research Centre.

Nicholson, A., & Y-D. Wong. (1993). Are accidents poisson distributed? A statistical test. *Accident Analysis & Prevention*, Volume 25, Issue 1, pp. 91-97.

NIST/SEMATECH (2006). *e-Handbook of Statistical Methods*, National Institute of Standards and Technology, http://www.itl.nist.gov/div898/handbook/ (accessed on 15 May, 2006).

Openshaw, S. (1984). The modifiable areal unit problem. Concepts and Techniques in Modern Geography 38, p. 41.

Oppe S. (*1993). Evolution de la circulation et de sécurité routière dans six pays développés.* Actes du séminaire Tome 2 "Modélisation de l'insécurité routière", Institut de Recherche en Sécurité.

Oppe, S. (1979). The use of multiplicative models for analysis of road safety data, *Accident Analysis and Prevention*, 11(2), pp. 101-115.

Oppe, S. (1989). Macroscopic models for traffic and traffic safety. *Accident Analysis and Prevention* 21 (3), pp. 225–232.

Ostrom, C.W. (1990). *Time series regression techniques.* Second edition. London: Sage Publications.

Pawlovich, M. D., Li, W., Carriquiry, A., and Welch, T. M. (2006). Iowa's Experience with 'Road Diet' Measures: Impacts on Crash Frequencies and Crash Rates Assessed Following Bayesian Approach. Proceedings of the *85th Annual Meeting of the Transportation Research Board*, Washington D.C.

Persaud, Bhagwant N. and Ezra Hauer (1984). Comparison of Two Methods for Debiasing Before-and-After Accident Studies. *Transportation Research Record: Journal of the Transportation Research Board*, No. 975, TRB, National Research Council, Washington, D.C., pp. 43-49.

Pindyck R.S. and Rubinfeld D.L. (1997). *Econometric Models and Economic Forecasts*, 4th Edition. Irwin McGraw-Hill, Boston MA.

Pinheiro, J.C. and Bates D. M. (1995) Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model *Journal of Computational and Graphics Statistics 4*, 12-35

Qin, X., Ivan, J. N., Ravishanker, N., and Liu, J. (2005). Hierarchical Bayesian Estimation of Safety Performance Functions for Two-Lane Highways Using Markov Chain Monte Carlo Modelling. *ASCE Journal of Transportation Engineering*, Volume 131, Issue 5, pp. 345-351.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org (accessed May 26, 2005).

Rasbash J. and Browne W. J. (in press). Non-Hierarchical Multilevel Models. To appear in De Leeuw, J. and Kreft, I.G.G. (Eds.), *Handbook of Quantitative Multilevel Analysis*.

Rasbash, J. and Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross classified random structures using a multilevel model. *Journal of Educational and Behavioural statistics* 19: 337-50.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D, Langford, I, & Lewis, T. (2000). *A user's guide to MLwiN. Version 2.1c*, Centre for Multilevel Modeling, Institute of Education, University of London, UK.

Rasbash, J., Steele, F., Browne, W, Prosser, B. (2004). *A User's Guide to MLwiN. Version 2.1e*, Centre for Multilevel Modeling, Institute of Education, University of London, UK.

Raudenbush, S. W, & Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods* (second edition), Thousand Oaks, California: Sage Publications.

Retting, R. A., & Kyrychenko, S. Y. (2001). *Crash Reductions Associated with Red Light Camera Enforcement in Oxnard,* California. Insurance Institute for Highway Safety, Arlington, VA.

Rice, N. (2001). Binomial Regression, in *Multilevel Modeling of Health Statistics*, Ed. A. H. Leyland & H. Goldstein, pp. 27-43, West Sussex, England: John Wiley & Sons, Ltd.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals, *American Sociological Review*, Vol. 15, pp. 351-357.

Rodriguez, G., Goldman, N. (1995). An assessment of estimation procedure for multilevel models with binary responses, *Journal of the Royal Statistical Society A*, Vol. 158, pp. 73-89.

Sargent D. J. (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics. 54(4)*:1486-97.

Schnabel, K.U., Little, T.D., & Baumert, J. (2000). Modeling longitudinal and multilevel data. London: Lawrence Erlbaum.,

Schwarz, G. (1978). Estimating the Dimension of a Model. Annals of Statistics. 6 461–464.

Scott P. P. (1986), Modelling Time-Series of British Road Accident Data, *Accident Analysis & Prevention*, Vol. 18 n°2 pp.109-117.

Shieh, Y, Fouladi, R. (2003). The Effect of Multicollinearity on Multilevel Modelling Parameter Estimates and Standard Errors. *Educational and Psychological Measurement*, Vol. 63, No. 6, pp. 951-985.

Simpson, H.M., Beirness, D.J., Robertson, R.D., Mayhew, D.R., and Hedlund, J.H. (2004). Hard core drink drivers. Traffic Injury Prevention 5(3), pp. 261 – 269.

Smeed J.R. (1949). Some statistical aspects of road safety research*. Journal of the Royal Statistical Society, A1*, 1-34.

Smeed, R J. (1968) Variations in the pattern of accident rates in different countries and their causes. *Traffic Engineering Control* 10(7), 364-371.

Smith, A. F. M., and Gelfland, F. M. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *American Statistician*, 46, 2, pp. 84-88.

Snijders, T, & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*, London: Sage Publications.

Spiegelman, C., and Gates, T. J. (2005). Post Hoc Quantile Test for One-Way Analysis of Variance Using a Double Bootstrap Method. *Transportation Research Record: Journal of the Transportation Research Board*, 1908, pp. 19-25, Washington, D.C.

Steunpunt Verkeersveiligheid.

Stine, R. (1989). An introduction to bootstrap methods. *Sociological Methods and Research*, 18, 2-3, pp. 243-291.

Swamy, P. A. V. B. (1971). *Statistical Inference in Random Coefficient Regression Models*, New York: Springer.

Tacq, J. (1986). Van *multiniveau probleem naar multiniveau analyse*, Department of Research Methods and Techniques, Erasmus university, Rotterdam.

Tacq, J. (1997). *Multivariate Analysis Techniques in Social Science Research*, London: Sage Publications Ltd.

the DRAG family: literature review (RA-2003-08). Diepenbeek, Belgium:

Thomas I., (1996). Spatial data aggregation: exploratory analysis of road accidents. Accident Analysis and Prevention Vol. 28 No 2, pp. 251 - 264.

Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region. Economic Geography 46, pp. 234–40.

van Belle, G. (2002). *Statistical rules of thumb.* New York: John Wiley and Sons.

Van den Bossche, F., & Wets, G. (2003). *Macro models in traffic safety and*

van Driel, C. J. G., Davidse, R. J., & van Maarseveen M. F. A. M. (2004). The effects of an edgeline on speed and lateral position: a meta-analysis, *Accident Analysis and Prevention*, Vol. 36, pp. 671-682.

Vanlaar, W. (2002). Results of Belgian drink driving roadside survey. In: *Procedings of the 16th International Conference on Alcohol, Drugs, and Traffic Safety in Montreal, Canada.*

Vanlaar, W. (2005a). Multilevel modelling in traffic safety research: Two empirical examples illustrating the consequences of ignoring hierarchies. *Traffic Injury Prevention. 6,* (4), pp. 311-316.

Vanlaar, W. (2005b). Drink driving in Belgium: results from the third and improved roadside survey, *Accident Analysis and Prevention*, Vol. 37, pp. 391-397.

Venables, W. N., and Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth edition, Springer-Verlag, New York.

Verbeke, T., Vanlaar W., Silverans, P. (in press). *Report of seatbelt wearing rates of 2003 and 2004*, IBSR, Brussels.

Vlakveld, W. P. (2005). Jonge beginnende automobilisten, hun ongevalsrisico en maatregelen om dit terug te dringen – Een literatuurstudie. Available: http://www.swov.nl/nl/publicaties/index.htm (accessed may, 2006).

Washington, S. P., Karlaftis, M. G., and F. L. Mannering (2003). *Statistical and Econometric Models for Transportation Data Analysis*. Chapman & Hall/CRC (2003).

Wood, G.R. (2002). Generalized Linear Accident Models and Goodness of Fit Testing. *Accident Analysis & Prevention*, Vol. 34, pp. 417-427.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, D. U. Schnabel, and J. Baumert (Eds.) Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples (pp:13-26). Mahwah: Lawrence Erlbaum.

Yang, M., Goldstein, H., Browne, W., Woodhouse, G., (2001). Multivariate multilevel analyses of examination results. *J. Royal Statistical Society*, A.

Yannis G et al. (2005). *State of the Art Report on Risk and Exposure Data*. Deliverable 2.1 of the EU FP6 project SafetyNet.

Yannis G., Papadimitriou E., Antoniou C., (2007). Multilevel modeling for the regional effect of enforcement on road accidents. In press, Accident Analysis and Prevention.

Yung, Y. F., and Chan, W. (1999). Statistical analyses using bootstrapping: Concepts and implementation. In R. Hoyle (Ed.) Statistical strategies for small sample research, pp. 82-108, Sage, Thousand Oaks, CA.

Zeger, S. (1988). A Regression Model for Time Series of Counts. *Biometrika*, Vol. 75, No. 4, pp. 621-629.