Deliverable D7.3: “Multilevel modelling and time series analysis in traffic safety research – Methodology”

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
Holywell Way
Loughborough
LE11 3UZ

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

Due Date of Deliverable: 31/10/2005

Submission Date: 20/05/2005

Report Author(s): C. Antoniou (NTUA), R. Bergel (INRETS), C. Brandstaetter (KUSS), J.J.F. Commandeur (SWOV), M. Gatscha (KUSS), E. Papadimitriou (NTUA), W. Vanlaar (IBSR)

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	

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

Table of Contents

THEORY: LITERATURE REVIEW, METHODOLOGY AND DATA 11

2. Multilevel models 11
2.1. Introduction to multilevel models *(W. Vanlaar, IBSR)* 11
2.1.1. An intuitive approach to multilevel models 11
2.2. Basic two level random intercept and random slope models *(W. Vanlaar, IBSR)* 14
2.2.1. Research problem 14
2.2.2. Dataset 14
2.2.3. Model definition of the random intercept model 14
2.2.4. Model definition of the random intercept/random slope model 16
2.2.5. Objectives of the technique 17
2.2.6. Model assumptions 18
2.2.7. Model fit and diagnostics 18
2.2.7.1. The variance partition coefficient (VPC) 18
2.2.7.2. Deviance test 19
2.2.7.3. Residuals 20
2.2.8. Model interpretation 21
2.2.8.1. Variance components model 21
2.2.8.2. Random intercept/random slope model 22
2.2.9. Extending the model 23
2.2.9.1. Categorical predictors 23
2.2.9.2. Contextual effects 25
2.3. Three level models and more 27
2.4. Discrete response models 27
2. 4.1. Generalized linear models 27
2.4.2. Binary and general binomial responses *(W. Vanlaar, IBSR)* 27
2.4.2.1. Research problem 27
2. 4.2.2. Objectives of the technique 27
2.4.2.3. Dataset 28
2. 4.2.4. Model definition 28
2.4.2.5. Model assumptions 30
2. 4.2.6. Model fit and diagnostics 30
2. 4.2.7. Model interpretation 30
2. 4.2.8. Conclusion 32
2. 4.3. Multiple responses 33
2. 4.4. Counts *(E. Papadimitriou & C. Antoniou, NTUA)* 33
2. 4.4.1. Research problem 33
2.4.4.2. Objectives of the technique 34
2.4.4.3. Dataset 34
2.4.4.4. Model definition 35
2. 4.4.5. Model assumptions 36
2. 4.4.6. Model fit and diagnostics 37
2.4.4.7. Model interpretation 43
2.4.4.8. Conclusions 44

2.5. Repeated measures data 44
2.6. Multivariate models *(E. Papadimitriou & C. Antoniou, NTUA)* 45
2.6.1. Research problem 45
2. 6.2. Objectives of the technique 45
2. 6.3. Dataset 46
2. 6.4. Model definition 46
2. 6.5. Model fit and diagnostics 49
2. 6.6. Model interpretation 51
2. 6.7. Conclusions 52
2.7. Factor analysis and structural equations models *(C. Brandstaetter & M. Gatscha, KUSS)* 52
2.7.1. Research problem 52
2. 7.2. Model objectives 53
2. 7.3. Model definition 53
2. 7.4. Model assumptions 53
2.7.4.1. Multivariate normal distribution of the indicators: 53
2.7.4.2. Multivariate normal distribution of the latent dependent variables: 54
2.7.4.3. Linearity: 54
2.7.4.4. Indirect measurement: 55
2.7.4.5. Low measurement error: 55
2.7.4.6. Complete data or appropriate data imputation: 55
2.7.4.7. Not theoretically under-identified or just-identified: 55
2.7.4.8. Recursivity: 56
2.7.4.9. Not empirically identified due to high multicollinearity: 56
2.7.4.10. Interval data are assumed: 57
2.7.4.11. High precision: 57
2.7.4.12. Small, random residuals: 57
2.7.4.13. Uncorrelated residual error: 57
2.7.4.14. Multicollinearity: 57
2.7.4.15. Non-zero covariances: 58
2.7.4.16. Sample size: 58
2.7.5. Dataset 58
2.7.6. Model fit and diagnostics 59
2.7.7. Model interpretation 62
2.7.7.1. Measuring driver characteristics 62
2.7.7.2. Measuring technology characteristics 62
2.7.8. Conclusion 66
2.8 Miscellaneous 67

3. Time series models 68
3.1 Introduction to time series models *(R. Bergel, INRETS)* 68
3.1.1. Descriptive models 68
3.1.1.1. Decomposition models 68
3.1.1.2. Autoregressive, ARMA and ARIMA models 69
3.1.2. Explanatory models 69
3.1.2.1. Decomposition models with explanatory variables 70
3.1.2.2. ARIMA models with explanatory variables 70
3.1.3. Model classes 71

3.1.4. Variables and data 71
3.2 Time series analysis in road safety research *(R. Bergel, INRETS)* 72
3.2.1. The methodological framework 72
3.2.1.1. The diagram of production of the risk 72
3.2.1.2. Risk indicators 72
3.2.1.3. Risk factors 73
3.2.2. Towards an explanatory approach 73
3.2.3. Applications 73
3.2.3.1. Deterministic versus stochastic 73
3.2.3.2. Regression versus ARIMA 75
3.2.3.3. State space models 75
3.2.3.4. ARIMA models 76
3.2.3.5. Non linear models 77
3.2.4. Conclusion 78
3.3 Classical linear and non-linear regression models 78
3.3.1. Classical linear regression models *(C. Brandstaetter & M. Gatscha, KUSS)* 78
3.3.1.1. Research problem 78
3.3.1.2. Model objectives 78
3.3.1.3. Dataset 79
3.3.1.4. Model definition 79
3.3.1.5. Model assumptions 82
3.3.1.6. Model fit and diagnostics 85
3. 3.1.7. Model interpretation 94
3.3.1.8. Conclusion 94
3.3.2. Generalized linear models (GLM) *(E. Papadimitriou & C. Antoniou, NTUA)* 95
3.3.2.1. Research problem 95
3.3.2.2. Dataset 97
3.3.2.3. Model definition 97
3.3.2.4. Objectives of the technique 98
3.3.2.5. Model assumptions 98
3.3.2.6. Model fit and diagnostics 100
3.3.2.7. Model interpretation 106
3.3.2.8. Conclusion 106
3.3.3. Non-linear models 107
3.4 AR(I)MA(X) models *(R. Bergel, INRETS)* 107
3.4.1. Research problem 107
3.4.2. Dataset 108
3.4.3. Model definition 108
3.4.3.1. ARMA and ARIMA models 108
3.4.3.2. AR(I)MAX models 109
3.4.3.3. The application case 110
3.4.4. Objective of the technique 111
3.4.5. Model assumptions 111
3.4.6. Model fit and diagnostics 111
3.4.6.1. Validation and performance 112
3.4.6.2. The application case 113
3.4.7. Model interpretation 113

3.4.8. Conclusion 115
3.5 DRAG models *(R. Bergel, INRETS)* 119
3.5.1. Research problem 119
3.5.2. Dataset 119
3.5.3. Model definition 120
3.5.4. Objective of the technique 120
3.5.5. Model assumptions 120
3.5.6. Model fit and diagnostics 120
3.5.7. Model interpretation 121
3.5.8. Conclusion 121
3.6 State space models *(J. Commandeur, SWOV)* 121
3.6.1. The local level model 122
3.6.1.1. Research problem 122
3.6.1.2. Dataset 122
3.6.1.3. Model definition 123
3.6.1.4. Objectives of the technique 123
3.6.1.5. Model assumptions 123
3.6.1.6. Model fit and diagnostics 128
3.6.1.7. Model interpretation 128
3.6.2. The local linear trend model 128
3.6.3. The local linear trend plus seasonal model 133
3.6.4. Intervention variables 133
3.6.5. Explanatory variables 148
3.6.6. Forecasting 153

2.3 Multilevel models 157

2.4 Time series models 157

ACKNOWLEDGEMENT 158

REFERENCES 159

APPENDIX A- ..... ERROR! BOOKMARK NOT DEFINED.

List of figures

Figure 2.1, 2.2 and 2.3: Higher level distributions for road sites' intercepts and slopes – regression of speed against car length depending on road sites (right), "dotplot" for the distribution of the slopes and intercepts separately (center), "scatterplot" of the joint intercepts and slopes distributions (right), adapted from Jones, 1993, p.251. 12

Figure 2.4: Normal probability plot of residuals for the variance components model with speed and length, centered about its mean, at level 1 (left side) and 2 (right side) 20

Figure 2.5: Normal probability plot of residuals for the variance components model with the natural logarithm of speed and length, centered about its mean, at level 1 (left side) and 2 (right side) 20

Figure 2.6: Regression lines of speed against car length for the various road sites 22

Figure 2.7: Small and long cars' speed as a function of road sites 23

Figure 2.8: Average intercept and random intercepts for the "null" two-level model 39

Figure 2.9: Average and random intercepts and slopes of the single-effects two-level model (effect of alcohol controls) 39

Figure 2.10: Level 1 and 2 residuals of the single-effect model (effect of alcohol) 40

Figure 2.11: Random intercepts and slopes of the mixed-effects two-level model (effect of alcohol controls and effect of speed infringements) 41

Figure 2.12: Level 1 and 2 residuals of the single-effect model with extra-Poisson variation (effect of alcohol) 43

Figure 2.13: 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. 60

Figure 3.1: Scatterplot of accident statistics (number of fatalities) from 1987 to 2004 79

Figure 3.2: Scatterplot with regression line 80

Figure 3.3: Distribution of the residuals and violations of the homoscedasticity assumption 82

Figure 3.4: Histogram and P-P Plot of standardized residuals (in other chapters also the Q-Q Plot is used) 87

Figure 3.5: Table of selected residual plots for identifying heteroscedasticity 89

Figure 3.6: Table of autocorrelations and seasonal adjusted autocorrelations 90

Figure 3.7: Plot of Yearly Fatality Data in Austria from 1987 to 2004 90

Figure 3.8: Histogram and P-P Plot of standardized residuals 92

Figure 3.9: Table of selected residual plots for identifying heteroskedascity 93
Figure 3.10: Table of autocorrelations 94
Figure 3.11: Dataset overview 98
Figure 3.12: Model fit diagnostic plots (Gaussian distribution) 103
Figure 3.13: Model fit diagnostic plots (Poisson distribution) 104
Figure 3.14: Model fit diagnostic plots (Quasi-Poisson distribution) 104
Figure 3.15: Model fit diagnostic plots (Negative binomial distribution) 105
Figure 3.16: Quasi-Poisson model predictions 105
Figure 3.17, 3.18, and 3.19: Monthly impacts of the "Cellier effect" (April - November 1987), of the perspectives of presidential amnesties of 1988 (November 1987 - July 1998) and 1995 (December 1994 - June 1995). 116
Figure 3.20: Deterministic level and irregular component for the log of Norwegian fatalities. 125
Figure 3.21: Stochastic level and irregular component for the log of Norwegian fatalities. 127
Figure 3.22: Deterministic trend (top), deterministic slope (middle), and irregular component for the log of the number of Finnish fatalities. 130
Figure 3.23: Trend of deterministic level and stochastic slope model for the log of Finnish fatalities (top), stochastic slope component (middle), and irregular component (bottom). 132
Figure 3.24: Log of monthly number of UK drivers KSI with time lines for years. 134
Figure 3.25: 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. 137
Figure 3.26: 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. 139
Figure 3.27: Auxiliary residuals for the stochastic level and deterministic seasonal model applied to the log of the UK drivers KSI series. 141
Figure 3.28: Deterministic level plus intervention variable (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series . 143
Figure 3.29: Stochastic level plus intervention variable (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series. 146
Figure 3.30: 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. 147

Figure 3.31: Deterministic level plus intervention and explanatory variable (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series . 149

Figure 3.32: Stochastic level plus intervention and explanatory variables (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series . 152

Figure 3.33: Filtered trend, and seven year forecasts for Finnish fatalities, including their 90% confidence limits. 155

Figure 3.34: Filtered signal, and six months forecasts for the log of UK drivers KSI, including their 90% confidence limits. 156

List of tables

Table 2.1: Figure 2.1 represented as parameters for two higher-level distributions (where + is positive, different from zero and where – is negative, different from zero) 17

Table 2.2: Estimates for the null, variance components, and full random models, with car length as a continuous predictor 21

Table 2.3: Estimates for the null, variance components, and full random models, with car length as a categorical predictor 24

Table 2.4: Estimates for the null model and the models including contextual effects 26

Table 2.5: 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) 31

Table 2.6: Road safety trends in Greece – 1998 to 2002 34

Table 2.7: Variables in the dataset 35

Table 2.8: Estimates for the single-level model 37

Table 2.9: Estimates for the null model, the single-effects models and the mixed-effects model (Poisson assumptions) 38

Table 2.10: Estimates for the null model, the single-effects models and the mixed-effects model (extra-Poisson assumptions) 41

Table 2.11: Variables in the model 46

Table 2.12: Data matrix structure for the simple bivariate model 47

Table 2.13: Effects of the basic two-level bivariate model (intercept only) 49

Table 2.14: Effects of the two-level bivariate model (intercept and slope) 50

Table 2.15: Effects of the three-level bivariate model (intercept only) 50

Table 2.16: Effects of the three-level bivariate model (random intercept and slope) 51

Table 2.17: 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. 64

Table 2.18: 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. 65

Table 3.1: Types of models. 71

Table 3.2: Assumption violations and their consequences 85
Table 3.3: ANOVA table of a sample dataset with time as a predictor and the number of fatal accidents in Austria as a dependent variable 86
Table 3.4: Coefficients table 86
Table 3.5: Model summary table 86
Table 3.7: Table of selected residual statistics 90
Table 3.8: ANOVA table of sample dataset 90
Table 3.9: Coefficient table of sample dataset 91
Tables 3.10 to 3.12: Yearly fatality data in Austria from 1987 to 2004 - regression results 91-92
Table 3.13: Estimation results 102
Table 3.14: Model for the aggregate number of fatalities In France, for 1975-2001. (All exogenous variables kept) 117
Table 3.15: Model for the aggregate number of fatalities In France, for 1975-2001(Exogenous variables kept if T-ratio superior to 1) 118
Table 3.16: Diagnostic tests for deterministic level model and log of Norwegian fatalities. 126
Table 3.17: Diagnostic tests for local level model and Norwegian fatalities 128
Table 3.18: Diagnostic tests of residuals deterministic level and slope model for log Finnish fatalities. 131
Table 3.19: Diagnostic tests for deterministic level and stochastic slope model, and log Finnish fatalities. 133
Table 3.20: Diagnostic tests for deterministic trend and seasonal model for log UK drivers KSI. 138
Table 3.21: Diagnostic tests for stochastic level and deterministic dummy seasonal analysis of log of UK drivers KSI. 140
Table 3.22: Diagnostic tests for deterministic level and seasonal analysis of log of UK drivers KSI, including intervention variable. 144
Table 3.23: Diagnostic tests for stochastic level and dummy seasonal analysis of log of UK drivers KSI, including intervention variable. 146
Table 3.24: Diagnostic tests for deterministic level and dummy seasonal analysis of log of UK drivers KSI, including variables seat belt law and log petrol price. 151
Table 3.25: Diagnostic tests for stochastic level and dummy seasonal analysis of log of UK drivers KSI, including variables seat belt law and log petrol price. 153

Theory: Literature review, methodology and data

As indicated in paragraph 1.1 this chapter is model driven, both for the subchapter about multilevel models as for the subchapter on time series analysis. Furthermore, a standardized discussion format was adhered to, to discuss each model (research problem, dataset, model definition, objectives of the technique, model assumptions, model fit and diagnostics, model interpretation).

2. Multilevel models

2.1. Introduction to multilevel models (W. Vanlaar, IBSR)

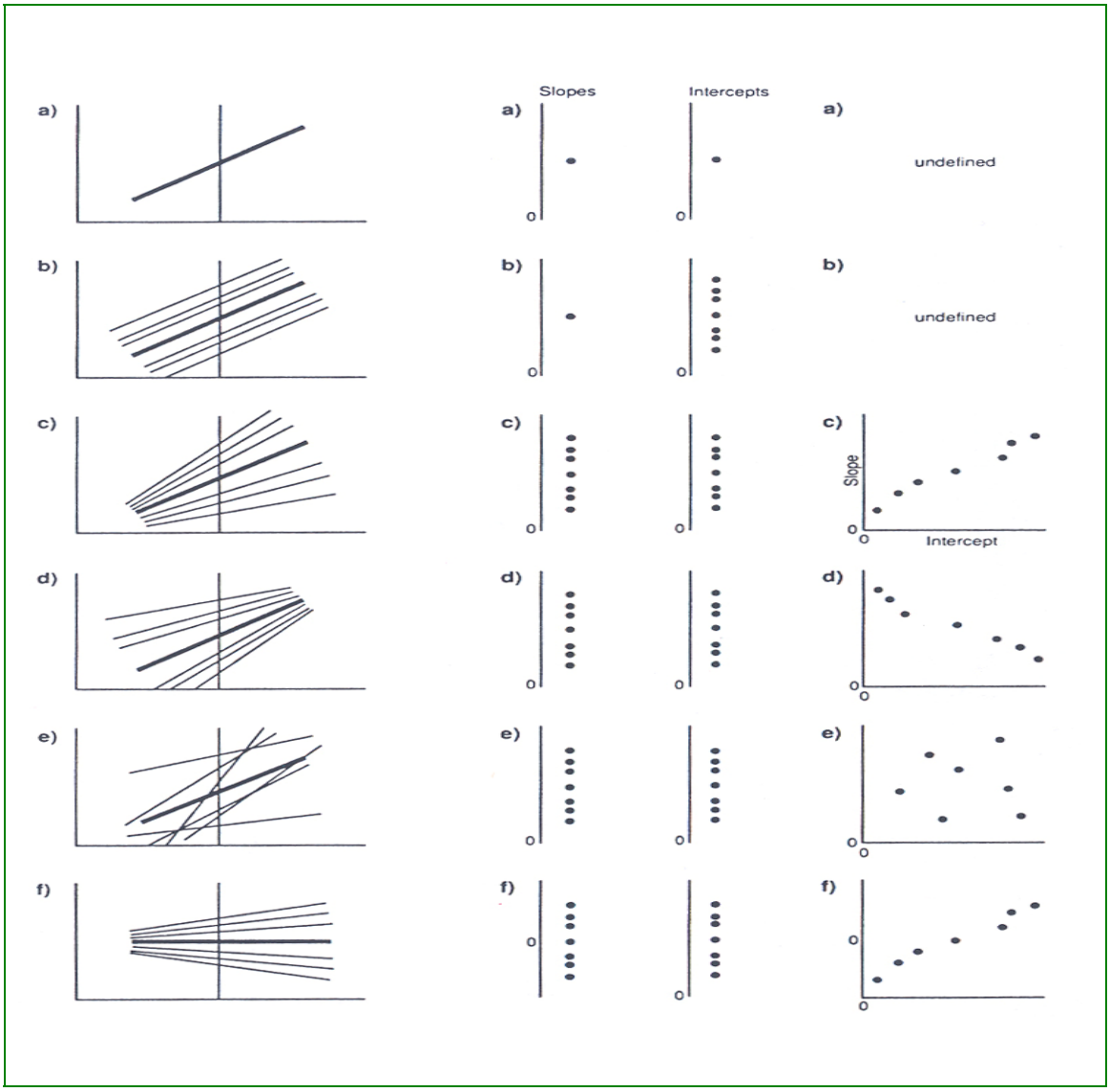
2.1.1. An intuitive approach to multilevel models[1]

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 as 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.

Figures 2.1, 2.2, and 2.3 (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

[1] *This chapter is mainly based on Jones (1993). Dr. Jones kindly granted us permission to use his manuscript as a basis for this section and to insert several of his graphs as an illustration.*

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.



Figures 2.1 to 2.3: 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).

The situation becomes more complicated in 2.1c to 2.1f 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 2.1b only the intercepts of the lines differed. In 2.1c 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 2.1d 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, 2.1e, 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 2.1f 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 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 different forms of Figures 1c to 1f are a result of how the intercepts and slopes are associated" (Jones, 1993: p. 252). In Figure 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 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 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 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 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.

The different forms of 2.1c to 2.1f are a result of how the intercepts and slopes are associated. In 2.1c 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. Put another way, there is positive association between the intercepts and slopes, as shown in Figure 2.3c. In contrast, in Figure 2.1d road sites where the average speed is high have a weak speed/length relationship: a high intercept is associated with a shallow slope. Consequently, Figure 2.3d shows negative association between the slopes and the intercepts. The complex criss-crossing of 2.1e is the result of the lack of pattern between the intercepts and slopes shown in Figure 2.3e. The average speed at a particular road site tells us nothing about the marginal increase in speed with length of cars at that road site. The distinctive feature of the final plot 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. Basic two level random intercept and random slope models *(W. Vanlaar, IBSR)*

In this paragraph we will turn the graphs into equations shifting from an intuitive approach to a more formal, mathematical approach.

2.2.1. 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.

2.2.2. 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 constructed in Belgium for epidemiological reasons. 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.

2.2.3. Model definition of 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:

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 2.1a), the general verbal equation becomes:

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i \quad (2.1)$$

where

- subscript i signifies an individual respondent;
- y and x measure the response and predictor 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 we have to re-specify the *micro-model* by distinguishing road sites with the subscript j. For the random intercept model (cf. figure 2.1b) this yields:

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + e_{oij} \quad (2.2a)$$

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

$$\beta_{0j} = \beta_0 + u_{0j} \quad (2.2b)$$

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

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:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \left(u_{0j} + e_{0ij}\right) \quad (2.2c)$$

All the elaborations have come in the random part, because in addition to allowing individual cars to vary, we have also allowed road sites 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 et al., 2004).

2.2.4. Model definition of the random intercept/random slope model

The formulas look as follows if we also allow the slope to vary from road site to road site besides a random intercept (cf. figures 2.1c-2.1f). The micro model:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{0ij} \quad (2.3a)$$

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

$$\beta_{0j} = \beta_0 + u_{0j} \quad (2.3b)$$

$$\beta_{1j} = \beta_1 + u_{1j} \quad (2.3c)$$

These two macro-models allow respectively for the differential road site intercept (u_{0j}) to vary from road site to road site around the overall intercept (β_0) and for the differential slope (u_{1j}) to vary around the overall slope (β_1) (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_{0ij} 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:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \left(u_{1j} x_{1ij} + u_{0j} + e_{0ij}\right) \quad (2.3d)$$

All the elaborations have come in the random part, for in addition to allowing individual cars to vary, we have also allowed road sites 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 2.1 (after Jones, 1993) summarizes Figure 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 2.1a).

	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 2.1: Figure 2.1 represented as parameters for two higher-level distributions (where + is positive, different from zero and where – is negative, different from zero)

2.2.5. 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 classical 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 classical 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.

2.2.6. 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., 2001):

$e_{0ij} \sim N(0, \sigma^2_{e_0})$, 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 $\sigma^2_{u_1}$;

Random coefficients at level 1 (e_{oij}) 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.

The homoscedasticity assumption, i.e. the assumption that the variances and covariances estimated at the different levels of the data are constant thus holds for multilevel models, just as for many other statistical analysis techniques. However, in multilevel modeling, this assumption can be relaxed. We will see how and why in a next section.

2.2.7. Model fit and diagnostics

2.2..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. However, the former interpretation is the more usual one (Rasbash, 2004). The VPC is denoted by:

$$\frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_{e_0}} \qquad (2.4)$$

In our example the VPC for the variance components model with length as predictor is 0.749, meaning that almost 75% of the variation is due to differences between road sites. This is a strong indication that clustering effects are not to be disregarded in this dataset and that a multilevel approach is preferable.

2.2.7.2. 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 a predictor. The difference between minus twice the natural logarithm of the likelihood (-2xloglikelihood) 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 5). 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 model 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 variance components 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, $\sigma^2_{u_{1j}}$ and σ_{01}) . This yields a p-value of 0.000 and is thus highly significant.

2.2.7.2.1. 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 on model assumptions). 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.

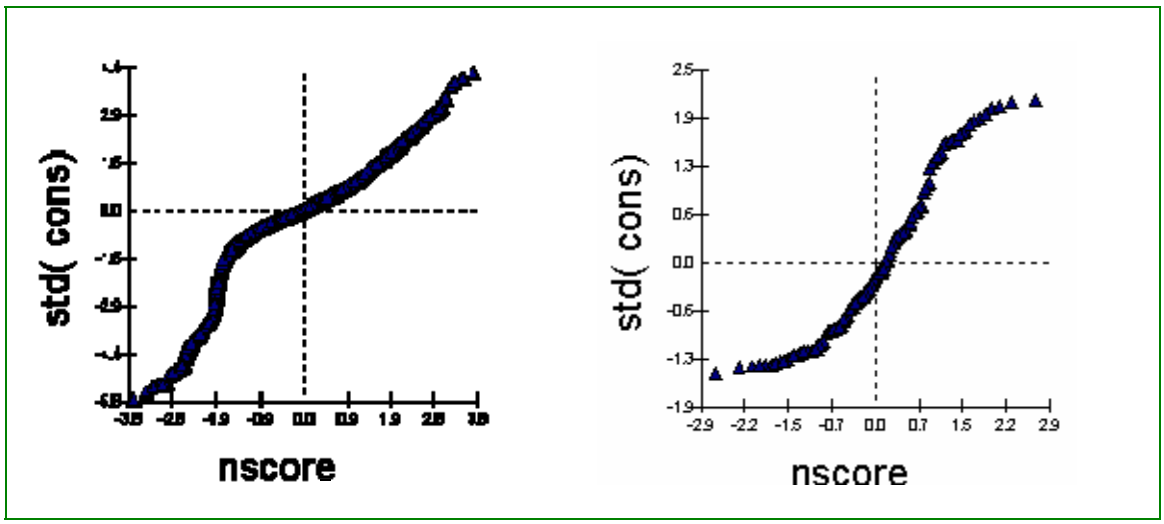


Figure 2.4: Normal probability plot of residuals for the variance components model with speed and length, centered about its mean, at level 1 (left side) and 2 (right side)

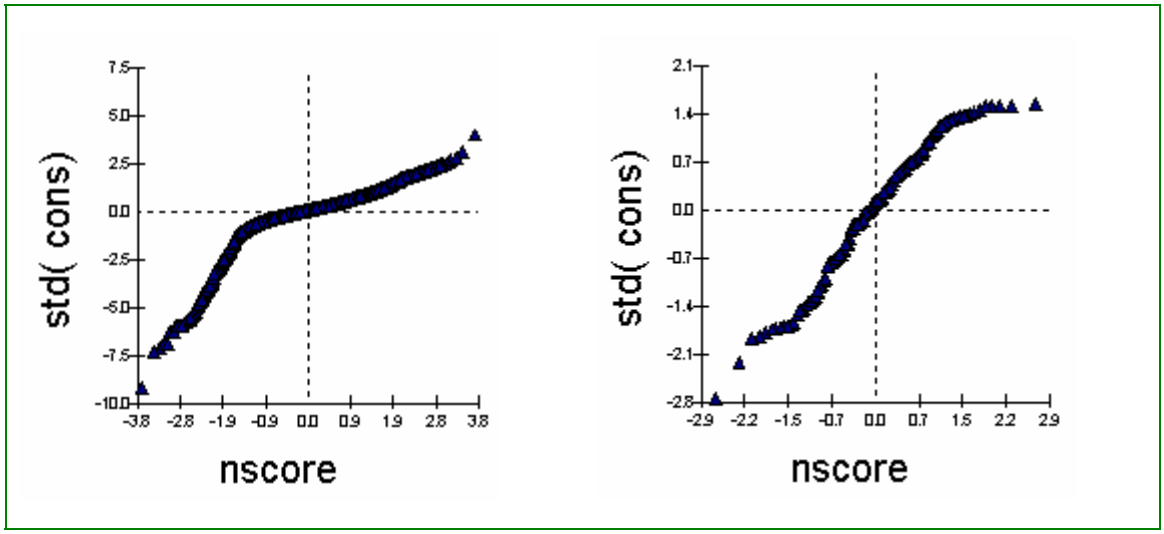


Figure 2.5: Normal probability plot of residuals for the variance components model with the natural logarithm of speed and length, centered about its mean, at level 1 (left side) and 2 (right side)

Clearly, the residuals in figure 2 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 3 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.

2.2.8. Model interpretation

2.2.8.1. Variance components model

The coefficients of the variance components model with a random intercept only are interpreted as follows (see Table 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.

Parameter	Null model	Variance components 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: Estimates for the null, variance components, and full random models, with car length as a continuous predictor

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

2.2.8.2. Random intercept/random slope model

The main difference between the variance components 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.

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.

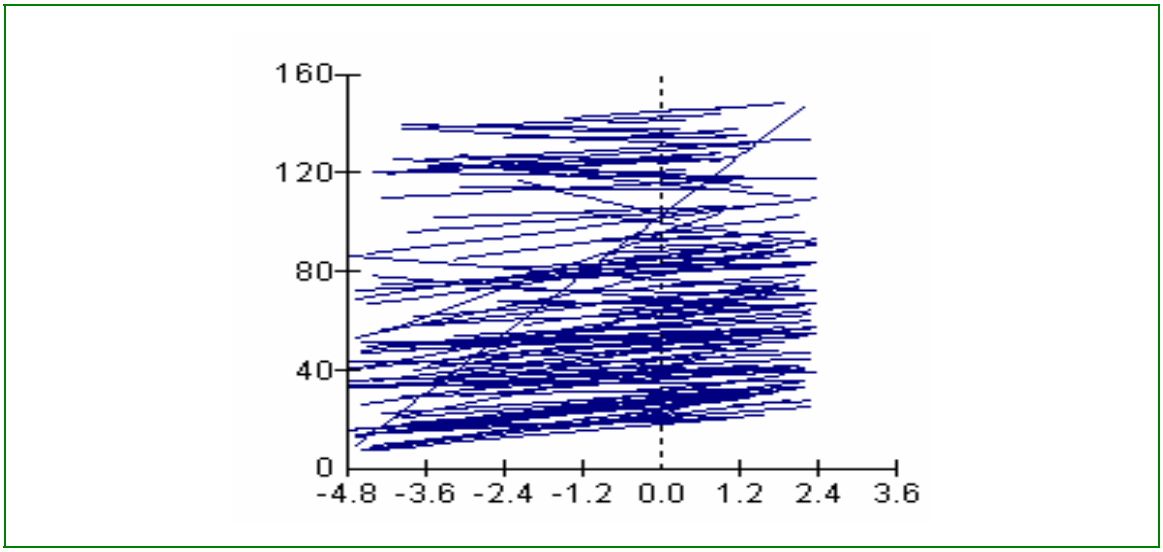


Figure 2.6: Regression lines of speed against car length for the various road sites

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 2.1d and figure 2.6). In other words, a greater intercept corresponds to a smaller slope. The pattern is more easily discerned on figure 2.1d than on the graph based on our dataset. A possible explanation was given previously. Perhaps the attitude of drivers of powerful cars differs from the attitude of drivers of small 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.

2.2.9. 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 predictor variable will be studied. Second, higher level predictor variables and contextual effects will be considered

2.2.9.1. Categorical predictors

According to Jones (1993) level 1 categorical predictors present no special problems and multilevel models can be specified in which some or all of the predictors are 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:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{0ij} \tag{2.5a}$$

and additionally two macro-models:

$$\beta_{0j} = \beta_0 + u_{0j} \tag{2.5b}$$

$$\beta_{1j} = \beta_1 + u_{1j} \tag{2.5c}$$

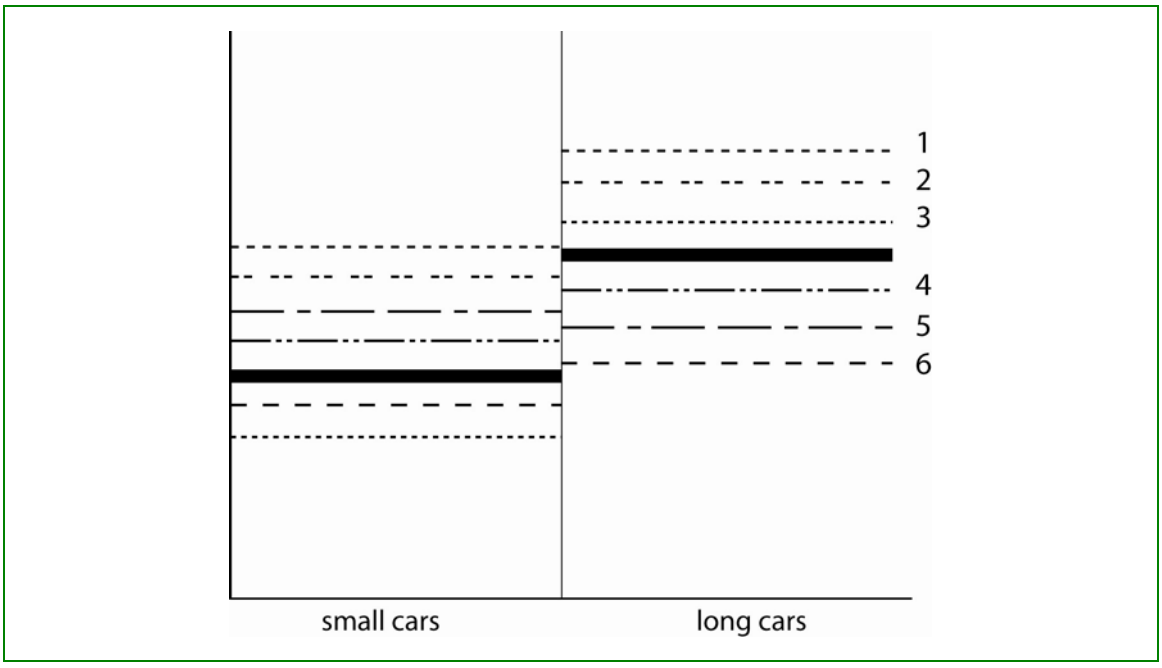


Figure 2.7: Small (<4.3m) and long cars' (>=4.3m) speed as a function of road sites

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 2.7). The solid lines in the figure represent the overall general relationship indicating that smaller cars on average have lower speeds than longer cars. However, there at road site 5 a pattern is found that differs from the overall general relationship.

Table 2.2 contains the estimates of the null model, the variance components model and the full random model. According to the variance components 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 variance components model (deviance: 45262.13-45218.96=43.17; degrees of freedom=1; p=0.000).

Parameter	Null model	Variance components 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)
$\sigma_{u_0u_1}$ (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 2.3: Estimates for the null, variance components, and full random models, with car length as a categorical predictor

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.

2.2.9.2. 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 (2.2b) can be re-specified to include an extra term, as in:

$$\beta_{0j} = \beta_0 + \alpha_1 C_j + u_{0j} \tag{2.6a}$$

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.

$$\beta_{1j} = \beta_1 + \alpha_2 C_j + u_{1j} \tag{2.6b}$$

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):

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \alpha_1 C_j + \alpha_2 C_j x_{1ij} + \left(u_{1j} x_{1ij} + u_{0j} + e_{0ij}\right) \tag{2.6c}$$

Table 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 latter 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. 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.

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 2.4: Estimates for the null model and the models including contextual effects

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. Actually 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 could be hidden because of a high traffic count. Indeed free way is a necessary condition for speed behaviour to become visible. Of course one should not forget that this coefficient is not significant. Not so surprisingly this third model is not a better model compared to the main effect model according to the deviance test.

2.3. Three level models and more

- To be completed-

2.4. Discrete response models

2.4.1. Generalized linear models

- To be completed-

2.4.2. Binary and general binomial responses[2] *(W. Vanlaar, IBSR)*

2.4.2.1. *Research problem*

In 2003 the Belgian Road Safety Institute organised the third national roadside survey to estimate the proportion of drink drivers and their profile. The objective of this initiative is 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 will be 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 generalize 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.

2.4.2.2. *Objectives of the technique*

- 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 (this variable is binary so the responses can only take the values of 0 or 1) (see model definition and interpretation).

- To investigate whether the model that was found fits the data well (see model diagnostics).

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

- To interpret the relationships found and to check whether these relationships can be generalized to the population (see model definition and interpretation).

2.4.2.3. *Dataset*

Data were gathered during a drink driving roadside survey in 2003 according to a stratified two stage cluster sample. 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. 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).

2.4.2.4. *Model definition*

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). Our binary response (0,1) is y_{ij}, which equals 1 if driver i in district j was drink driving, and 0 if he/she was not. We denote the probability that y_{ij}=1 by π_{ij}. Note that other link functions could be used as an alternative to the logit function. This choice should be guided by "the empirical fit, ease of interpretation, and convenience – e.g., availability of computer software" (Snijders and Boskers, 1999: p. 213).

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

$$\log it(\pi_{ij}) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_{0j} + \beta_1 x_{1ij} \qquad (2.7a)$$

$$\beta_{0j} = \beta_0 + u_{0j} \qquad (2.7b)$$

Or as an equation for the outcome y_{ij} (Rice, 2001: p. 31):

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + e_{0ij} \qquad (2.8a)$$

$$\beta_{0j} = \beta_0 + u_{0j} \qquad (2.8b)$$

As in the random intercept model for a continuous response, the intercept in the probability equation consists of two terms: a fixed component β_0 and a road-site-specific component, the random effect u_{0j}.

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 2.7a, we obtain:

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_{0j}) \times \exp(\beta_1 x_{ij}) \qquad (2.9a)$$

If we increase x by 1 unit, we obtain:

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_{0j}) \times \exp(\beta_1 (x_{ij}+1) = \exp(\beta_{oj}) \times \exp(\beta_1 x_{ij}) \times \exp(\beta_1) \qquad (2.9b)$$

This is the expression in 2.9a, multiplied by e^{β_1}. Therefore e^{β_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 e^{β_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 e^{β_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.

2.4.2.5. *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}.

$e_{0ij} \sim \text{logistic}(0, \pi^2/3)$, the driver-specific error term is assumed to have the logistic distribution with mean zero and variance $\pi^2/3 = 3.29$.

$y_{ij} \sim \text{Bin}(1, \pi_{ij})$, the observed binary responses are assumed to be binomially distributed with mean 1 and variance π_{ij}.

2.4.2.6. *Model fit and diagnostics*

The final model fits the data well, which can be derived from the level 1 variance Ω_e=0.712 in the extra binomial model; i.e. a model that does not constrain the level 1 variance to be equal to unity as opposed to the binomial model where the level 1 variance is equal to 1 by definition. Since this model diagnostic is rather close to 1 – which actually means there is little evidence that our model exhibits extra binomial variance, more precisely under dispersion[3] – the binomial distribution holds. Table 2.5 contains the results for the different parameters of both the extra binomial and the binomial model. The strength and the direction of all relationships remain unchanged between both models.

The intraclass correlation coefficient for the multilevel logistic model – a coefficient that indicates "whether a given nesting structure in a data set calls for multilevel analysis" (Snijders and Bosker, 1999: p. 22) – is $\rho = \Omega_u / (\Omega_u + \pi^2/3)$. In our case the intraclass correlation coefficient, 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.

2.4.2.7. *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 2.5, 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

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

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%.

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 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).

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: Ω_u	1.569 (0.229)		0.991 (0.197)	
Level 1 variance: Ω_e	0.712 (0.010)		1.000 (0.000)	

Table 2.5: 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)

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 in comparison with the reference category are multiplied by a factor of 4.187; in other words, those odds increase by 318.7%.

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).

2.4.2.8. *Conclusion*

Regarding the appropriateness of this technique, note that it was shown in the model fit section that the model fits the data well and that the data called for a multilevel approach. Furthermore, taking account of the arguments in favour of multilevel modeling, elaborated on in the introduction, it is concluded that modeling this dataset according to a 2 level binomial model is highly recommended and that a 1 level model would perform less well.

The following conclusions related to drink driving are drawn. The results of the multilevel models 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).

An interesting relationship was identified 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. 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.

We found evidence that drivers who have been tested and provided a breath sample in the past at least once are more likely to drink drive than drivers who have never been tested before. 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.

Another strange result was identified regarding the perception of drivers of being stopped and tested on an average trip – the subjective risk of getting caught. The data clearly support a positive relationship, meaning that drivers who estimate the likelihood of getting tested as very high, are at the highest risk for drink driving. Based on the same model as before, 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).

2.4.3. Multiple responses

- To be completed -

2.4.4. Counts *(E. Papadimitriou & C. Antoniou, NTUA)*

2.4.4.1. *Research problem*

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 are systematically monitored and the related enforcement and casualty results at local and

national level are regularly published, as shown at the following table with basic road safety related trends in Greece.

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 the number of accident represents a random count of events occurring within a population, a Poisson distribution is assumed.

	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 2.6: Road safety trends in Greece – 1998 to 2002

2.4.4.2. *Objectives of the technique*

In this section multilevel models that fit data where the response variable is discrete are further analyzed. Following the analysis concerning binary data shown in the previous section, count data where the response 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 the previous section, and the Poisson distribution 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

2.4.4.3. *Dataset*

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 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:

It should be noted that the Athens and Thessaloniki metropolitan areas, where a disproportionally 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 2.7: Variables in the dataset

2.4.4.4. *Model definition*

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). In case of Poisson multilevel regression, there is a higher level classification of the data across which the probability response is considered to vary. The multi-level 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 of observed cases is assumed, and hence it is necessary to include an offset of expected numbers of cases in the model so that:

$$O \sim \text{Poisson}(\mu) \qquad (2.10a)$$

$$\log(P) = \log(E) + X\beta + Z\theta \qquad (2.10b)$$

where *E* represents the expected numbers of cases for each level 1 unit. When using such fixed offsets, it is recommended to centre them about their mean in order to avoid numerical instabilities (Rasbach et al., 2000).

The Poisson distribution is used to model the level 1 variance, with a logarithmic link function, and assume random parameters at higher levels (e.g. region and nation) as being multivariate normal. An efficient estimation procedure for this non-linear 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.

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.

There are some applications where the response is a count and we do not require an offset, or where the offset is effectively constant. For example, if we were interested in the number of times individuals visited their general practitioner or physician in a year, we could collect data over a one year period for all individuals and study the variation in counts across practitioners (level 2) according to individual and practitioner characteristics.

There are also variations on the Poisson distribution assumption which we may wish to use (for example, the negative binomial distribution). One could add further terms or consider even a nonlinear function.

2.4.4.5. *Model 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 that events occur in the underlying population, according to the following characteristic function:

$$P(x ; \lambda)=\frac{\lambda^{x} e^{-\lambda}}{x!} \qquad (2.11)$$

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, sometimes referred to as overdispersion.

2.4.4.6. *Model fit and diagnostics*

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 (mixed-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:

Parameters	Single-level model
constant	-6.450 (0.005)
alcontrols	-0.015 (0.001)
speedinf	-0.010 (0.001)

Table 2.8: Estimates for the single-level model

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. This model (Model 1) shall be also used as the "null" model for the assessment of models fit, through the calculation of deviance. The results presented in Table 2.9 below indicate a significant random variance among regions:
The significant regional variation of the intercept is presented in Figure 2.8. The first graph concerns the average (fixed) intercept for all regions, whereas the second graph concerns the intercepts corresponding to each one of the 12 regions of Greece. A significant regional effect on the number of accidents is illustrated. Additionally, Model 2 presents a significantly improved fit, as the related deviance statistic is equal to (7038.97-4624.30)=2414.67, which is highly significant compared to a Chi-square distribution with one degree of freedom.

It should be noted that likelihood statistics for discrete response models are very approximate, as quasilikelihood estimation is use. Therefore, likelihood statistics are only examined as a rough assessment of models fit (Rashbach et al., 2000).

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 2.6.

Parameter	Model 1 (Null model)	Model 2 (Effect of alcohol controls)	Model 3 (Effect of speed controls)	Model 4 (Effect of speed and alcohol controls)
	Estimate (s.e)	Estimate (s.e)	Estimate (s.e)	Estimate (s.e)
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
-2*loglikelihood	7038.97	4624.30	4666.03	4360.68

Table 2.9: Estimates for the null model, the single-effects models and the mixed-effects model (Poisson assumptions)

It is noticed that all fixed and random effects are significant. However, the variance 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 higher the effect of alcohol enforcement (reduction of accidents).

The significant regional variation of the slope is presented in Figure 2.9. The first graph concerns the average (fixed) slope for all regions, whereas the second 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.

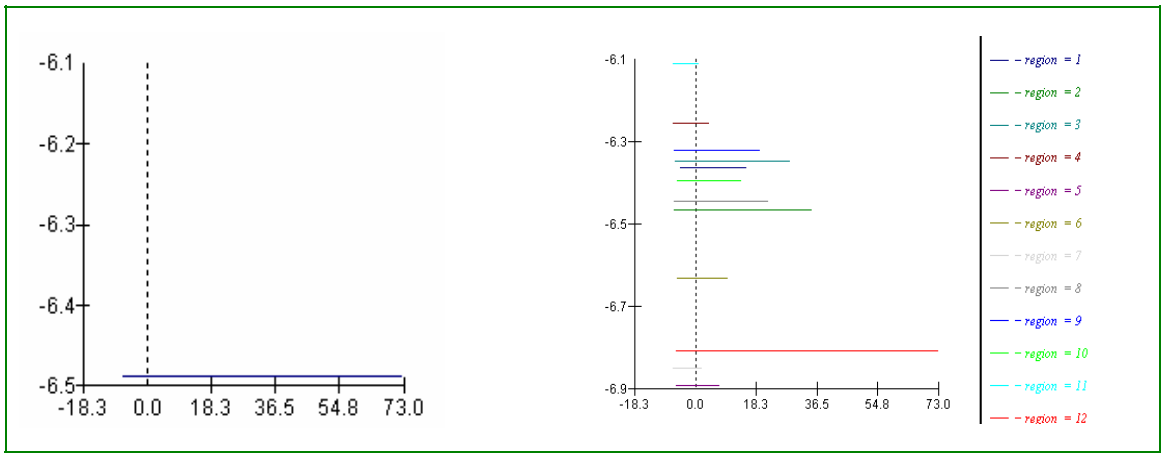


Figure 2.8: Average intercept (left) and random intercepts (right) for the "null" two-level model

In Figure 2.10, the Level 1 and 2 residuals are examined for Model 2. In particular, the top graphs concern Level 1 residuals and the four bottom graphs 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.

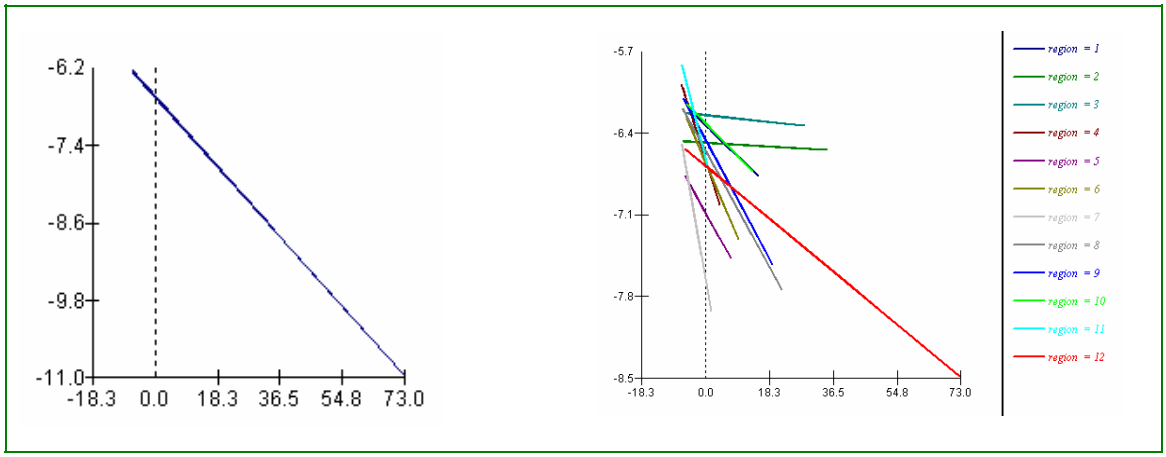


Figure 2.9: Average (left) and random (right) intercepts and slopes of the single-effect two-level model (effect of alcohol controls)

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

Accordingly, the effect of speed enforcement on the number of accidents is separately examined, by removing the number of alcohol controls from the model and adding the number of speed infringements, also allowing it to

randomly vary between regions. Another multilevel model with a random intercept and a random slope is therefore fitted (Model 3).

Also in this case, all fixed and random effects are significant. It should be noted, however, that the variance of the effect of speed infringements is highly significant in this case. 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. The resulting deviance from incorporating the number of speed infringements in the model is equal to 2372.94, which is also significant for one degree of freedom, however somewhat less improving the model fit compared to the number of alcohol controls.

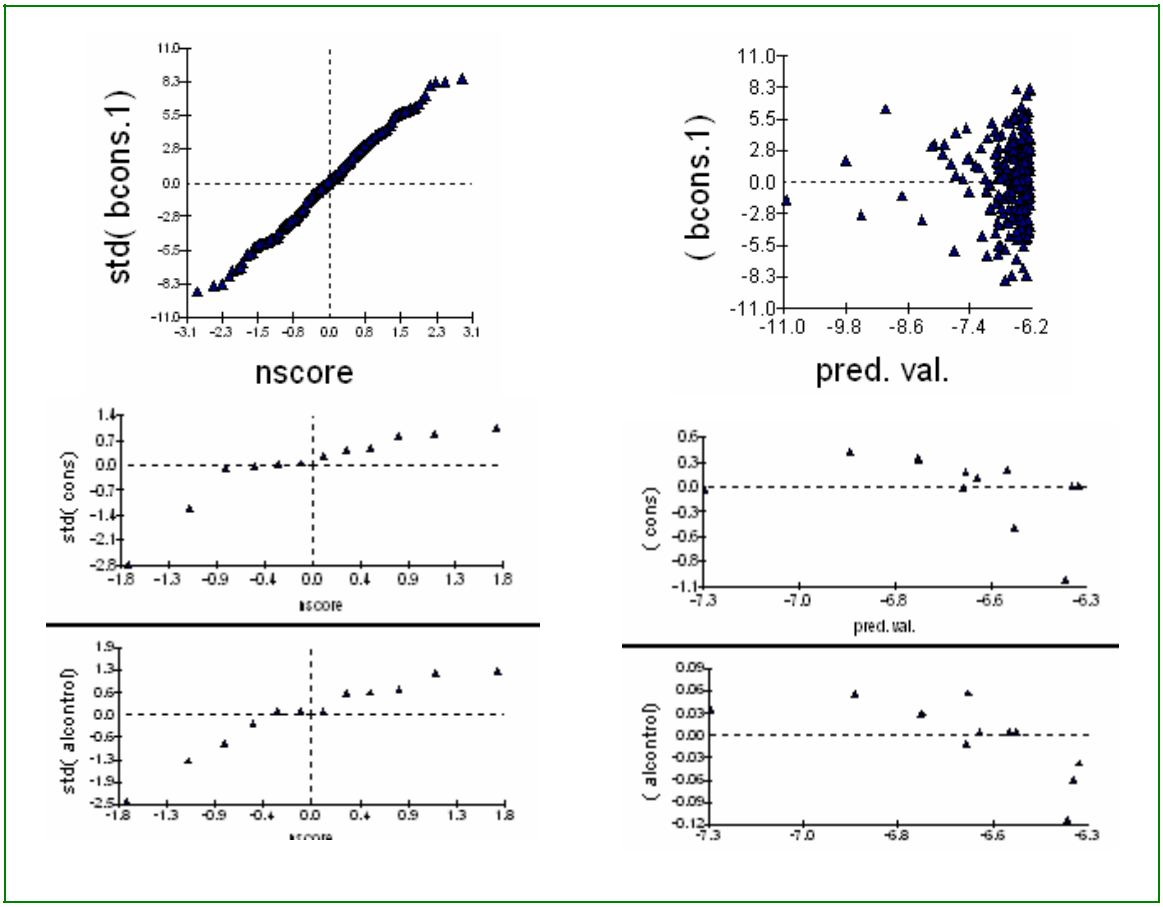


Figure 2.10: Level 1 and 2 residuals of the single-effect model (effect of alcohol)

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), allowing both explanatory variables to vary among regions. In this case, all fixed effects are highly significant. However, the variances and 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. Additionally, the fact that the overall fit of the model was at the same time improved indicates some bias in the estimates.

In Figure 2.11, 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. These results are discussed and interpreted in the following section.

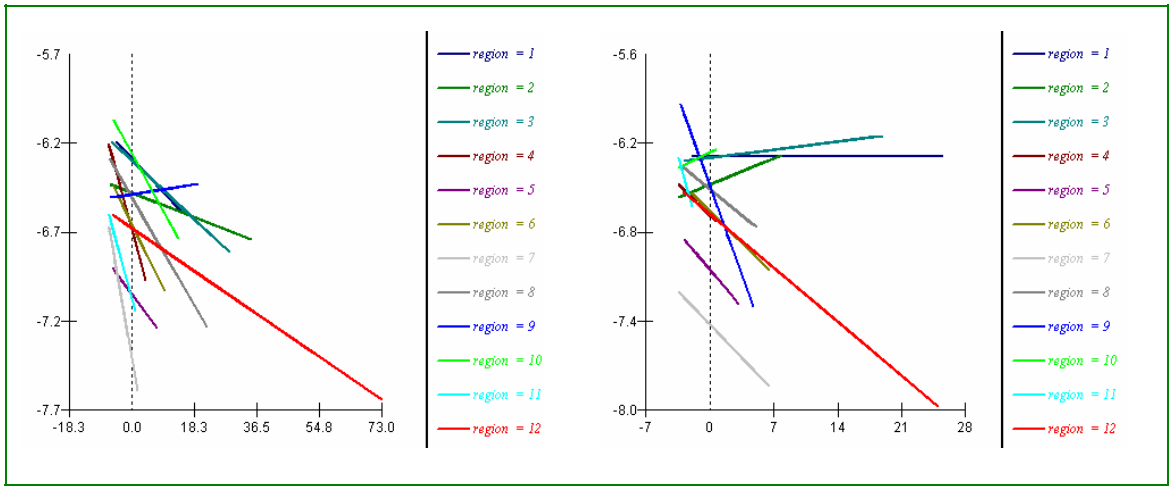


Figure 2.11: Random intercepts (left) and slopes (right) of the mixed-effects two-level model (effect of alcohol controls and effect of speed infringements)

Parameter	Model 5 (Null model)	Model 6 (effect of alcohol)
	Estimate (s.e)	Estimate (s.e)
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)
-2*loglikelihood	2729.07	2621.82

Table 2.10: Estimates for the null model, the single-effects models and the mixed-effects model (extra-Poisson assumptions)

Another issue that should be examined in case of Poisson multilevel models is overdispersion (Dean, Lawless, 1989). In particular, when examining Model 2, it is noticed that the dispersion parameter, calculated as the ratio of the residual deviance to the degrees of freedom (minus the number of estimated parameters) is equal to 2414.67/(245-1)=9,89. The model is proved to be highly overdispersed, and the initial assumption of variance-mean equality is violated.

Overdispersion generally reflects missing parameters, not included in the model, which would account for the extra-variation.

A procedure to account for this overdispersion can be used, by not restricting the variance-mean relationship to be equal to one. It should be noted that this assumption would not significantly affect parameter estimates; however the related significancies 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.

In particular, in Table 2.10 above, parameter estimates are presented for a "null" model (Model 5 - intercept only) and a model examining the effect of alcohol (Model 6). It is noticed that parameter estimates are not significantly different from the ones obtained with Poisson assumptions. Additionally, a significant estimate of the variance/mean ratio is obtained, indicating that the variance-mean equality assumed in the previous examples was not adequate.

The dispersion parameter for Model 6 is equal to 107,25/244=0.44. This parameter is now lower than one (underdispersion), however the overall result is improved from the 9,89/1 overdispersion obtained before. Moreover, taking into account that these deviance estimates are very approximate estimates, it can be deduced that Model 6 is improved in relation to Model 2.

In Figure 2.12, level 1 and 2 residuals are examined for Model 6. Examining the level 1 residuals of the model, it is observed that these are normally distributed and independent. When examining level 2 residuals, 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.

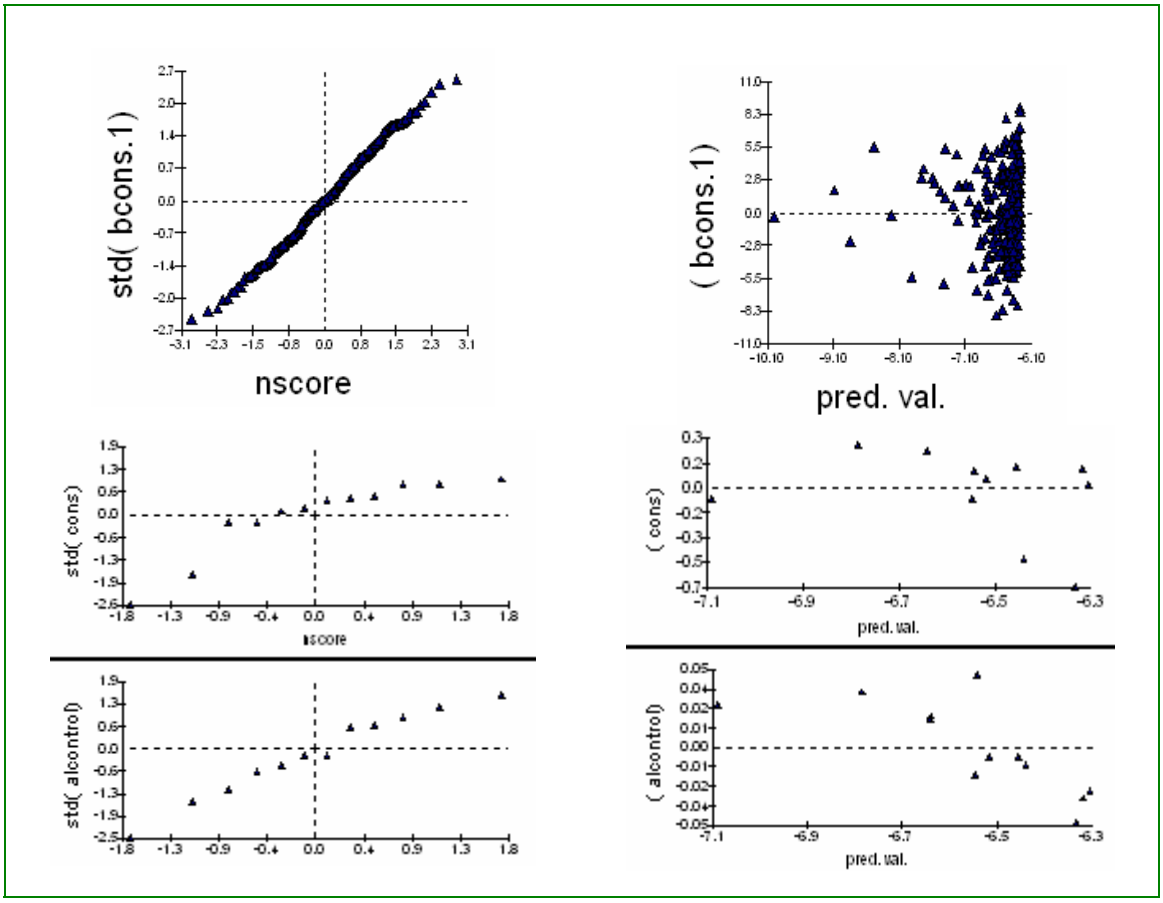


Figure 2.12: Level 1 and 2 residuals of the single-effect model with extra-Poisson variation (effect of alcohol)

2.4.4.7. *Model interpretation*

In order to interpret the above results, especially as far as Model 4 is concerned, the correlation between speed infringements and alcohol controls was examined, resulting to a positive correlation of 0,729. This explains to some degree the counter-intuitive slopes a couple of regions in Model 4. In particular, when variables are highly correlated (multicollinearity), a redundancy of variables is exposed, causing both logical and statistical problems (Washinghton et al. 2003). Therefore, it is recommended that variables with a bivariate correlation of greater than 0.70 are avoided in the same analysis. Redundant variables weaken the analysis, through reduction of degrees of freedom error.

It is interesting to note that model's fit is not significantly affected. If the analysis simply aims to predict the response variable from a set of explanatory variables, then multicollinearity is not a problem. However, if the analysis aims to investigate how the various explanatory variables impact the dependent variable, then multicollinearity is an important problem. As far as multilevel models are concerned, the results of a recent study show that, with multicollinearity presented at Level 1 of a two-level mixed-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). According to the above, Model 4 is rejected against Models 2 and 3.

Another interesting issue rising from the above results is overdispersion in Poisson models. It was demonstrated that, a restricted variance-mean equality assumption may lead to overestimation of the parameter significances, as the underlying degree of overdispersion affects the estimated standard errors. This may not always compromise models fit, it is therefore necessary to examine the dispersion parameter though the models deviance. However, in discrete response models, the deviance estimates are quite rough and can only provide a very general assessment of fit.

In the above examples, the extra-Poisson variation assumption has led from a highly overdispersed model (Model 2) to a quite underdispersed model (Model 6). However, the overall fit and diagnostics of the two models indicate that Model 6 is the best Model for the purposes of the present analysis.

2.4.4.7.1. Conclusions

In this section, a Poisson multilevel modelling process was demonstrated. The dataset used includes the number of road accidents and the related speeding and drinking-and-driving violations, as well as some socioeconomic parameters, for 50 counties and 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.

The MLwin software for 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. The development of separate models for the effect of speeding enforcement and the effect of alcohol enforcement produced interesting results and revealed statistically significant regional variations in the examined effects. The fitting of a model including both effects produced a couple of counter-intuitive results on specific regions, which are attributed to the fact that the two effects were found to be correlated.

The above analysis reflects the potential and usefulness of using multilevel analysis to identify complex relationships within hierarchical data. It also demonstrates some limits of the analysis, mainly rising from the limited number of higher level units and the existence of correlation between some explanatory variables in the particular dataset.

2.5. Repeated measures data

- To be completed -

2.6. Multivariate models *(E. Papadimitriou & C. Antoniou, NTUA)*

2.6.1. Research problem

In Chapter 2.1.4 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 frequency. 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. It was also found that separate models for each explanatory variable provided more significant and stable results than one model including both variables, as a significant correlation of speed and alcohol enforcement was found.

Additionally, models with extra-Poisson variation assumptions (overdispersion) were proved to be more flexible in relation to standard Poisson variation assumptions, correcting for the overestimation of the significances of parameter estimates.

In this Chapter, the effect of alcohol enforcement on both road accidents and road accident casualties is examined. The interest of this analysis lies on the fact that road accident severity (number of casualties) 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 casualties 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 casualties.

2.6.2. Objectives of the technique

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

The analysis has the following objectives:

- Present the assumptions and properties of multivariate multilevel models in relation to univariate models
- Describe the respective assumptions and particularities for multivariate Poisson models
- Use the above techniques to explore the regional effect of police enforcement on the number of road accidents and road accident casualties in Greece

2.6.3. Dataset

In the following sections, an application of a Poisson multivariate multilevel model is demonstrated. On that purpose, the dataset presented in Chapter 2.1.4 is used. This dataset includes the number of road traffic accidents and related casualties in 50 counties nested within 12 regions of Greece for the period 1998-2002. As mentioned in the previous section, this period corresponds to a considerable intensification of police enforcement for two of the most important traffic violations i.e. exceeding speed limits and driving under the influence of alcohol.

A bivariate model is therefore developed, with the following variables:

region	1-12 regions of Greece
county	1-50 counties of Greece
accidents	The number of accidents of each county
killed	The number of persons killed in the road accidents of each county
alcontrol (1000)	The number of alcohol controls of each county
logepop (offset)	The natural logarithm of the population of each county
Cons	The constant term

Table 2.11: Variables in the model

It should be noted that, as in the example of univariate Poisson models, the Athens and Thessaloniki metropolitan areas, where a disproportionally 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 2.1.4) it was proved that alcohol and speed enforcement are significantly correlated and therefore they should not be examined jointly.

It should be noted that all the assumptions of Poisson multilevel models described in Chapter 2.1.4 also apply in the case of multivariate models.

2.6.4. Model definition

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.

In particular, in the simplest case of a bivariate model, each level 1 entry would be a response indicating one of the two response variables for each unit, the basic explanatory variables would be a set of binary variables indicating which

of the two responses is present and further explanatory variables would correspond to unit level variables. This structure is illustrated in Table 2.12:

Individual	Response	Intercepts		Slopes	
1	Response 1	0	1	0	1
1	Response 2	1	0	1	0
2	Response 1	0	1	0	1
2	Response 2	1	0	1	0
3	Response 1	0	1	0	1
3	Response 2	1	0	1	0

Table 2.12: Data matrix structure for the simple bivariate model

The statistical formula for the two level basic bivariate model, is written as follows:

$$y_{ij} = b_0 z_{ij} + b_1 z_{21j} + b_2 z_{1ij} x_j + b_3 z_{2ij} x_j + u_1 j z_{1ij} + u_{2j} z_{2ij} \quad (2.12a)$$

$$\text{Where } z_{1i} = \begin{cases} 1 & \text{if response 1} \\ 2 & \text{if response 2} \end{cases}, \; z_{2i} = 1 - z_{1ij} \quad (2.12b)$$

There are several interesting features in this model. 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 between-units estimates of the variances and covariance. The multilevel estimates are statistically efficient even where some responses are missing.

It should be noted that the estimates obtained are not necessarily the same as the estimates that would be obtained by fitting two separate univariate models. If there is a tendency, for instance, to report/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 (MAR), which is a weaker assumption.

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 studies can be designed in such a way that not every unit (individual) has every measurement, with measurements randomly allocated to units. Such "rotation" or "matrix" designs are common in many areas and may be efficiently modelled in this way.

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

$$y_{ij} = b_0 z_{ijk} + b_1 z_{21jk} + b_2 z_{1ijk} x_{jk} + b_3 z_{2ijk} x_{jk} + v_{0jk} z_{1ikj} + v_{1jk} z_{2ijk}\ u_{0jk} z_{1ikj} + u_{1jk} z_{2ijk} \quad (2.13a)$$

$$\text{Where } z_{1i} = \begin{Bmatrix} 1 & \text{if response 1} \\ 2 & \text{if response 2} \end{Bmatrix}, \ z_{2i} = 1 - z_{1ij} \quad (2.13b)$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N(0, \Omega_v) \qquad \Omega_v = \begin{bmatrix} \sigma^2_{v0} & \\ \sigma^2_{v01} & \sigma^2_{v1} \end{bmatrix} \quad (2.13c)$$

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

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

This model could be extended further, by allowing 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.

A typical example to illustrate the multilevel multivariate response model is given by Rasbach 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.

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.

An example of fitting multivariate Poisson models can also be found in Langford et al. (1999).

2.6.5. Model fit and diagnostics

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 100 units (counties) is created; 50 units corresponding to the 1st response (number of accidents) and 50 units corresponding to the 2nd response (number of persons killed).

The natural logarithm of the population is used as an offset in both responses. It should be also noted that extra-Poisson distributional assumptions are considered, in order to allow for more flexibility in the estimations. The modeling results for the simple examination of variability between responses (two-level model with fixed intercept) are presented in Table 2.13.

	Model 1	
	Accidents	Killed
Fixed effects		
constant	-6.471 (0.025)	-8.380 (0.023)
Cov (accs/killed)	4.691 (0.042)	

Table 2.13: Effects of the basic two-level bivariate model (intercept only)

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 below indicate that the effect of alcohol enforcement is significant both for the number of accidents and for the number of persons killed:

It should be underlined that no random structure can be specified at the lowest "real" level (i.e. the county level and not the response level) of a Poisson model whether it is univariate or multivariate (bivariate in this case). 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 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 in a 1-level normal model but not in a 1-level Poisson model.

	Model 2	
	Accidents	Killed
Fixed effects	-6.455 (0.023)	-8.372 (0.023)
constant		
alcontrols	-0.019 (0.003)	-0.006 (0.002)
Cov (accs/killed)	4.139 (0.657)	

Table 2.14: Effects of the two-level bivariate model (intercept and slope)

At the next stage, it is interesting to examine 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	-6.453 (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 2.15: Effects of the three-level bivariate model (intercept only)

The results presented in Table 2.15 above show a significant regional variation of both road accidents and road accident casualties, 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. However, it is interesting to notice that the

covariance between responses and its significance is reduced. It can be deduced that the variation of accidents and persons killed does not follow the same trend within different regions i.e., an increase in the number of accidents does not result in the same increase in the number of persons killed among regions.

By adding a random slope to the model, the results shown in Table 2.16 below are obtained (Model 4). 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.

At this stage, there is enough evidence that road accidents and road accident casualties present a significant and different regional variation. Additionally, the increase of alcohol controls causes a different reduction on accidents and persons killed at national level. However, alcohol controls do not appear to significantly affect persons killed at regional level.

	Model 4	
	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)	0.000 (0.000)
Cov (accs/killed)	3.313 (0.556)	

Table 2.16: Effects of the three-level bivariate model (random intercept and slope)

2.6.6. Model interpretation

The above example concerns a typical multivariate modeling process under Poisson assumptions. A significant regional variation was observed in both responses. However, a significant variation related to the number of alcohol controls was observed for accidents only. A less complex univariate model was successfully fitted in the accidents data in Chapter 2.1.4, 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 model for the number of persons killed was also fitted to the data and the non-significant

regional 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, 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 effect of alcohol enforcement mainly affects the number of accidents. It can therefore be deduced that an increase of alcohol controls results to a significant decrease of accidents. The number of persons killed decreases because the number of accidents decreases and not because of a direct effect of alcohol controls.

2.6.7. Conclusions

In this section, a Poisson multilevel modelling process was demonstrated. The main interest of the example presented lies on 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 example presented above concerned the effect of alcohol enforcement on the number of road accidents and related casualties. The results showed that accidents and casualties present (significant) regional variation; however the effect of enforcement on the number of casualties does not vary significantly at regional level.

The modelling process described above can be applied accordingly to normal, binary or count responses, or mixed responses. Some of the particularities of modelling Poisson responses in relation to Normal responses were briefly discussed in the framework of the above example. Additionally, multiple responses can also be modelled in the same way. 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.

2.7. Factor analysis and structural equations models *(C. Brandstaetter & M. Gatscha, KUSS)*

2.7.1. Research problem

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.

2.7.2. Model objectives

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).

2.7.3. Model definition

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 *p* level 1 responses, we first write a multivariate model with *p* responses, where in general some may be randomly missing.

$$y_{ij} = (X\beta)_{ij} + \sum_i e_i z_{ij} \qquad (2.14)$$

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.

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.

2.7.4. Model assumptions

All structural equation models, in short SEM, have important assumptions, which have to be known when applying such a concept. Although it utilizes path analysis, SEM relaxes many (but not all) of its assumptions pertaining to data level, interactions, and uncorrelated error.

2.7.4.1. Multivariate normal distribution of the indicators:

Each indicator 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 (path) coefficients. Specifically, MLE requires normally distributed endogenous 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 (like ADF) 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.

2.7.4.2. 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.

2.7.4.3. 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. These transformations are added alone to model power effects, or along with the original variable to model a quadratic effect with an unanalysed correlation (curved double-headed arrow), connecting them in the diagrammatic model. It is also possible to model quadratic and non-linear effects of latent variables.

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.

2.7.4.4. 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.

2.7.4.5. 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.

2.7.4.6. 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.

2.7.4.7. 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 (arrows) 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 knowns (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.

2.7.4.8. 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 that non-recursive models may also be solvable (not under-identified) under certain circumstances.

2.7.4.9. 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.

2.7.4.10. 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.

2.7.4.11. 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.

2.7.4.12. 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).

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.

2.7.4.13. Uncorrelated residual error:

The covariance of the predicted dependent scores and the residuals should be zero.

2.7.4.14. 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.

2.7.4.15. Non-zero covariances:

CFI and other 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, CFI, NFI, RMSEA, RMR, and other fit measures towards indicating good fit.

2.7.4.16. 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.

2.7.5. Dataset

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 LISREL 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 confirmatory 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.

2.7.6. Model fit and diagnostics

First, data from all available 23 countries was put together to find a general model that fits to all countries. In the next step, various goodness of fit statistics for every single country were computed. Due to the large number of missing cases in a few countries, an alternative model with extrapolated cases – computed by the standard AMOS 5 algorithm for missing cases - was used mainly for comparison purposes.

In the end, the general model worked well for 19 countries with acceptable fit statistics. No models could be calculated for four countries, and their results are not considered in the following analysis. These countries were Belgium, Ireland, Portugal and Croatia. For the UK and the Czech Republic, we have chosen the alternate model with extrapolated missing cases due to their better goodness of fit statistics.

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 2.13), displayed as ellipsis.

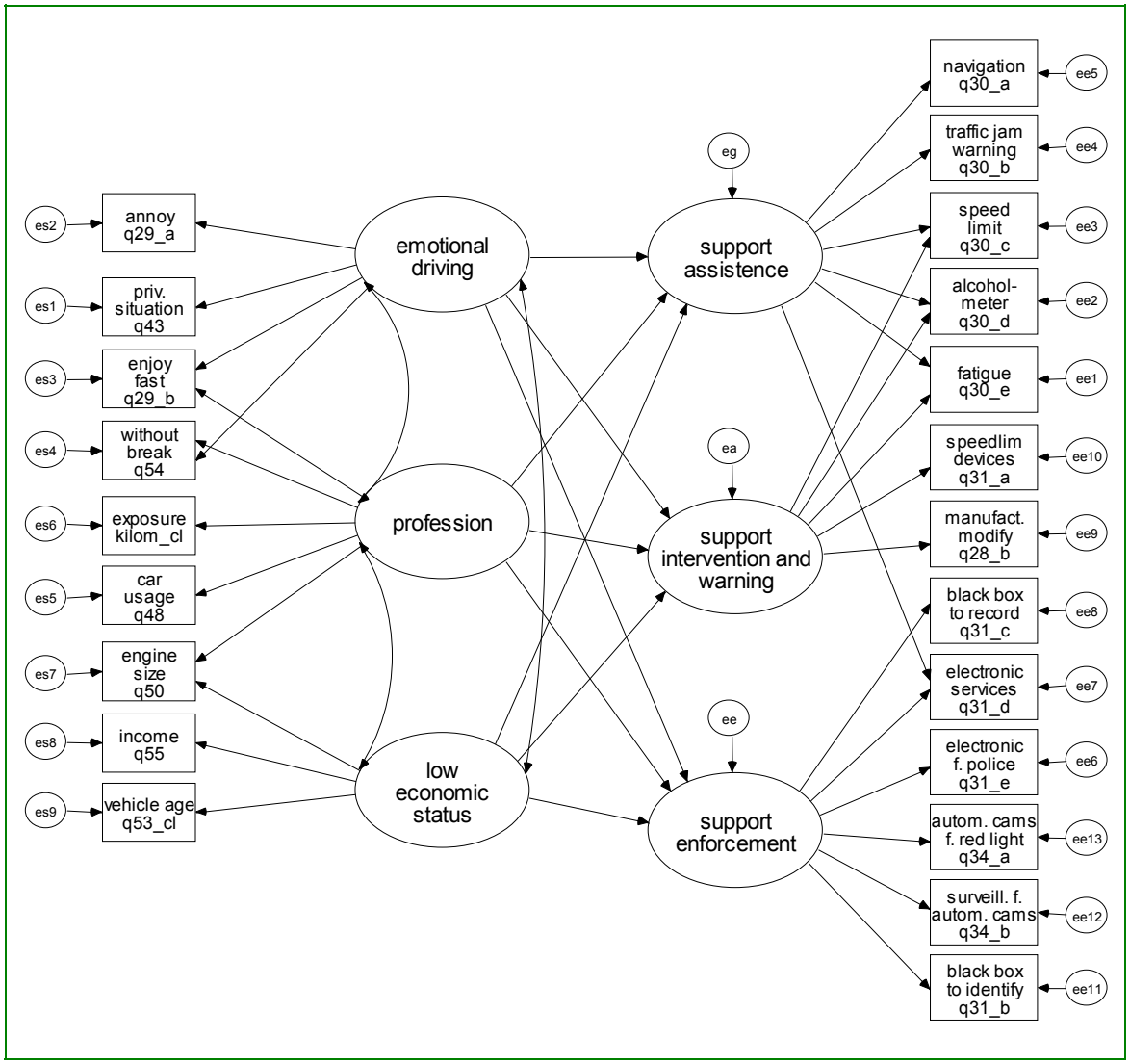


Figure 2.13: Proposed relations between driver and technology characteristics and questions used for operationalisation of those characteristics (short description of abbreviations/questions in the next section). Small circles represent the error terms.

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.

2.7.8. Model interpretation

2.7.8.1. 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?)

2.7.8.2. 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 2.13) and technology characteristics (right side of Figure 2.13) are collected in Table 2.17:

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 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	Mean	Standard deviation
Annoyed (q. 29 a) ← Emotional Driving	-0,2	0,2
Enjoy fast (q. 29b) ←Emotional Driving	-0,5	0,2
Priv. Situation (q. 43) ←Emotional Driving	-0,3	0,1
Without break (q. 54) ← Emotional Driving	0,2	0,1
Without break (q. 54) ← Profession	0,3	0,1
Exposure (kilom_cl) ← Profession	0,7	0,1
Enjoy fast (q. 29b) ←Profession	-0,2	0,2
Car usage (q. 48) ← Profession	-0,6	0,2
Engine Size (q. 50) ← Profession	0,3	0,3
Engine Size (q. 50) ← Low economic status	-0,2	0,3
Income (q. 55) ← Low economic status	-0,4	0,1
Vehicle age (q. 53_cl) ← Low economic status	0,2	0,1
Technology Characteristics	**Mean**	**Standard deviation**
Navigation (q. 30 a) ← Assistance & guidance	-0,7	0,1
Traffic Jam Warning (q. 30b) ← Assistance & guidance	-0,8	0,0
Speed delimiter (q. 30c) ← Assistance & guidance	-0,2	0,1
Alcohol meter (q. 30d) ← Assistance & guidance	-0,3	0,1
Fatigue (q. 30e) ← Assistance & guidance	-0,3	0,1
Electronic services (q. 31d) ← Assistance & guidance	-0,2	0,1
Speed delimiter (q. 30c) ← Warning & Intervention	-0,7	0,1
Alcohol meter (q. 30d) ← Warning & Intervention	-0,3	0,1
Fatigue (q. 30e) ← Warning & Intervention	-0,4	0,1
Speed lim. device (q. 31a) ← Warning & Intervention	-0,9	0,0
Manufact. Modify (q. 28b) ← Warning & Intervention	-0,5	0,2
Black box to record (q. 31c) ← Enforcement	-0,7	0,0
Electronic services (q. 31d) ← Enforcement	-0,4	0,1
Electronic services for police (q. 31e) ← Enforcement	-0,7	0,1
Autom. cams. F. red lights (q. 34a) ← Enforcement	-0,4	0,1
Surveill. f. autom. Cams (q 34b) ← Enforcement	-0,6	0,1
Black box to identify (q. 31b) ← Enforcement	-0,6	0,1

Table 2.17: 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.

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 2.18). 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	3,98	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 2.18: 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.

2.7.9. 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 model. 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.

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).

2.8. Miscellaneous

- ***To be completed –***

3. Time series models

3.1. Introduction to time series models *(R. Bergel, INRETS)*

In this section, types of models are addressed : the different types of models which are usually distinguished, when one aims to formulate the evolution over time of a theoretical stochastic[4] process (Y_t), for t being 1,2,3…., given a sample of observations $Y = (y_1, y_{2,}.., y_n)$.
The models addressed are the models proposed for the theoretical stochastic process (Y_t).

Two main kinds of models are usually distinguished : the ***descriptive models*** on the one hand - models for which the only exogenous variable used is time, which is then not considered as an explanatory variable - , and the ***explanatory models*** on the other hand - models which do use exogenous, or explanatory, variables (see Table 3.1 - Types of models).

3.1.1. Descriptive models

Descriptive models take account for the trend/seasonal/irregular decomposition of the variable Y_t. Here again, two main kinds of models are considered : ***decomposition models*** on the one hand**,** which adjust for each of the components, and AR, ARMA and more generally ***ARIMA models*** on the other hand, which adjust for the irregular component, after it has been filtered for the trend and the seasonal.

3.1.1.1. Decomposition models

Descriptive decomposition models can generally be written as

$$Y_t = f(t, u_t) = f(T_t, S_t, u_t) \qquad (3.1)$$

The process is a function of time t, and of a random disturbance u_t. The non-observed components of the process appear, rather naturally : the long term tendency T_t , the seasonal component S_t, and a random residual component u_t.

In the case of an additive decomposition, we shall write :

$$Y_t = T_t + S_t + u_t, \qquad (3.2)$$

[4] *The process (Y_t) is stochastic, or random, in the sense that the values taken by Y_t are under measurement errors.*

with : T_t the trend of the process Y_t,
S_t the seasonal, periodic, component,
and u_t the random, centred, component, assumed to be stationary[5] and qualified as irregular.

The trend is often thought as a function of certain variables, which determine it, although these variables can not always be quantified easily. But the trend can also be considered as a random walk (Harvey, 1989). The same remarks apply to the seasonal component. As can be seen, the structural modelling proposed by Harvey is another form of the preceding decomposition, in which the trend and the seasonal component are both random.

3.1.1.2. Autoregressive, ARMA and ARIMA models

The descriptive autoregressive, ARMA and ARIMA models can generally be written as

$$Y_t = f(Y_{t\text{-}1}, Y_{t\text{-}2,}.., u_t, u_{t\text{-}1,}..) \tag{3.3}$$

In the particular case where Y_t is stationary, an autoregressive or AR model is used, to express that Y_t is a function of its past values, and of a disrturbance u_t. The fact of knowing the dynamics of the process enables to extrapolate it, assuming that the dynamic's structure will stay unchanged in the future, at least at the forecast's horizon. The reference to the near past makes the model adaptive.

Different specifications, which are equivalent, can be chosen to model the same process Y_t . An autoregressive and moving average, or ARMA, specification for Y_t is often preferred, because of the advantage of it's smaller number of parameters. In that case, Y_t is a function of its past values (this number of past values being now smaller), of a disturbance u_t and of the past values of the disturbance.

In the general case where Y_t is not stationary, it is convenient to assume that another stationary process exists, which is derived from Y_t by removing its trend and its seasonal component, using a filter of differences.

3.1.2. Explanatory models

The explanatory time series models can be written as

$$Y_t = f(Z_t, u_t) \tag{3.4}$$

[5] *its mean, variance and covariance structure are constant over time (see a precise definition in section ..).*

where Y_t, the endogenous variable, is a function of the exogenous or explanatory[6] variable Z_t, and of the disturbance u_t.

Explanatory models can be seen as descriptive models to which exogenous variables have been added, and thus can also been classified as either decomposition models with explanatory variables, or ARIMA models with explanatory variables.

We shall now address these two kinds of models.

3.1.2.1. Decomposition models with explanatory variables

The decomposition models with explanatory variables can generally be written as

$$Y_t = g(Z_t) + f(T_t, S_t, u_t) \tag{3.5}$$

The common example is the regression model, of the dependent variable - or endogenous variable - on explanatory variables - or exogenous variables. The exogenous variables can be explanatory of the trend, of the seasonal component, or of 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 quartermonth, ..)

Harvey's structural model with explanatory - and intervention - variables is a kind of stochastic decomposition model more general than the basic structural model, mentioned before.

3.1.2.2. ARIMA models with explanatory variables

The ARIMA models with explanatory variables can generally be written as

$$YC_t = Y_t - g(Z_t) = f(YC_{t-1}, YC_{t-2}, .., u_t, u_{t-1}, ..) \tag{3.6}$$

Let's recall that ARIMA modelling consists of the estimation of the irregular component of a process, after the trend and the seasonal component have been filtered - this preliminary having thus stationarised the initial process.

ARIMA models with explanatory variables can also be seen as regression models with ARIMA residuals, the two formulations being equivalent. But it is important to determine whether the exogenous variables do have an effect on Y or on the variations of Y, after the trend and the seasonal component have been filtered.

---

[6] *Exogenous or explanatory because used in a model explanatory of the endogenous process* Y_t

3.1.3. Model classes

Different classes of models are usually considered. Let's note that the fact that a model belongs to one of these classes, or to one of the categories listed before, is not exclusive.

In the Safetynet project, the following classes will be considered :

- The classical regression models - linear and non-linear,
- The ARMA or ARMA - type models,
- The state-space models, which are decomposition models with stochastic components.

Descriptive models	*Explanatory models*
Decomposition models	*Decomposition models with explanatory variables*
$Y_t = f(T_t, S_t, u_t)$	$YC_t = Y_t - g(Z_t) = f(T_t, S_t, u_t)$
Autoregressive models	*Autoregressive models with explanatory variables*
$Y_t = f(Y_{t-1}, Y_{t-2,}.., u_t)$	$YC_t = Y_t - g(Z_t) = f(YC_{t-1}, YC_{t-2}, .., u_t)$
Autoregressive and moving average models	*Autoregressive and moving average models with explanatory variables*
$Y_t = f(Y_{t-1}, Y_{t-2,}.., u_t, u_{t-1,}..)$	$YC_t = Y_t - g(Z_t) = f(YC_{t-1}, YC_{t-2}, .., u_t, u_{t-1,}..)$
and, as extensions :	
AR(I)MA models	*AR(I)MAX models*

Table 3.1: Types of models.

3.1.4. Variables and data

The data which are strictly necessary for estimating a time series model have to be periodic, and numerous enough.

We have to distinguish:

- the observations of the endogenous stochastic process, i.e. the sample of data $Y = (y_1, y_{2,}.., y_n)$
- the values taken by k exogenous variables Z_{it}, i=1 à 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, their effect can be local – over time - , or permanent. It seems quite natural, again, to distinguish *the dummy variables*, which are built (outside the model) as witnesses of a local, isolated or repeated, effect, *and the variables of measure* of a phenomena, assumed to be linked with the process Y_t, and which 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.

3.2. Time series analysis in road safety research *(R. Bergel, INRETS)*

3.2.1. 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 *aggregate 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.

3.2.1.1. 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)in an accident,
- And three levels of risk: risk exposure, accident's risk, and accident's gravity.

Within that diagram, risk indicators and risk factors are defined at each of these three levels.

3.2.1.2. Risk indicators

The usual 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 considered as depending on a non-linear function of mileage[7].

[7] *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).*

Finally, the indicators that measure accident's gravity are ***the fatality rate***, i.e. the number of victims (fatalities, seriously injured, lightly 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.

3.2.1.3. Risk factors

Risk factors are classified in ***internal (to the transport system) factors*** on the one hand, related to the vehicle, to the driver and to infrastructure; and in ***external factors*** on the other hand, representing the environment, and related to the climatic, economic, demographic and state systems (Gaudry, Lassarre, 2000).

3.2.2. 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 of 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 the approaches differ on the necessity of taking account for an important number of explanatory factors, and on the nature of the models that should preferably be used. The examples given now illustrate these different approaches.

3.2.3. Applications

3.2.3.1. 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) :

$$R_t = \exp(\alpha t + \beta)$$

with : $R_t = \frac{F_t}{V_t}$,

F_t the number of fatalities, (3.7)

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:

$$\frac{F_t}{V_t^{\eta}} = \exp(m_t) \tag{3.8}$$

A stochastic form has then been proposed by Lassarre (1997) for the temporal function m_t , which becomes locally linear, that is to say by supplementing the basic structural model formulation :

$$\begin{aligned} LogF_t &= \eta LogV_t + m_t + \varepsilon_t \\ m_t &= m_{t\ 1} + b_{t\text{-}1} + \eta_t \\ b_t &= b_{t\ 1} + \xi_t \end{aligned} \tag{3.9}$$

with b the slope of the trend m,

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 - we shall discuss this now.

As has just been seen, a model descriptive of the fatality rate may be considered as a model explanatory 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 for a larger number of risk factors. We shall now give examples of such models.

It may 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.

3.2.3.2. 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 taking account for the trend and the seasonal component, in the form of a time variable and of seasonal dummies, in the regression equation.:

$$\log ACC_t = a + bt + S_t + \sum \beta_i LogX_{it} + \sum_j \beta_j X_{jt} + \lambda\omega_{1t} + \lambda_2\omega_{2t} + u_t \qquad (3.10)$$

with: $a + bt$ 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 speed limitation in rural areas.

3.2.3.3. 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, mentioned before. Used on KSI data in Great Britain, 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.

$$\begin{cases} LogACC_t = \mu_t + \gamma_t + \sum_{i=1}^{I} \beta_i Logx_{it} + \lambda\omega_t + \varepsilon_t \\ \mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t = \beta_{t-1} + \zeta_t \\ \gamma_t = \sum_{j=1}^{s/2} \gamma_{jt} \\ \gamma_{jt} = (\cos\frac{2\pi j}{s})\gamma_{j,t-1} + \omega_{jt} \end{cases} \qquad (3.11)$$

with : 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 Lassare (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 :

$$\begin{aligned} LogF_t &= \eta LogV_t + m_t + \sum_i \lambda_i \omega_{it} + \varepsilon_t \\ m_t &= m_{t-1} + b_{t-1} + \sum_j \lambda_j \omega_{jt} + \eta_t \\ b_t &= b_{t-1} + \sum_k \lambda_k \omega_{kt} + \xi_t \end{aligned} \qquad (3.12)$$

Applied to aggregate data of several European countries, this formulation allowed to asses 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.

3.2.3.4. ARIMA models

ARIMA models with explanatory variables are very often used on monthly data in the road safety field, in order to assess the effect of road safety measures. They generally take account for 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. Transfer functions are used in the case the explanatory variables are stochastic, and intervention variables for assessing road safety measures.

As examples we shall mention the models proposed for aggregate data in Spain and France.
Two variables of oil sales (gasoline and diesel) in the place of 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 june1992 off $LS^{6/92}$, are used for modelling the number of injury accidents in Spain from January 1982 to December 1996 (Rebollo, Rivelott, Inglada Lopez de Sabando, 2004):

$$\begin{aligned} \log ACC_t &= \sum_i \beta_i LogX_{it} + \eta WEND + \gamma LS_t^{6/92}{}_t + N_t \\ \nabla\nabla_{12} N_t &= (1-\theta_1 B)(1-\theta_{12} B^{12})\varepsilon_t \end{aligned} \qquad (3.13)$$

The same econometric specification was used for modelling the aggregate numbers of injury accidents and fatalities in France. The models take 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 fist 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 dummy variables for taking account of the calendar configuration as well (Bergel, Vizatelle, 1990).

3.2.3.5. 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 independant ; this renders the model estimation easier.

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 agregate model at a country's level, proposed by Smeed(1949), relate 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 :

$$D = c(MP^2)^{\frac{1}{3}} \tag{3.14}$$

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 autocorrelated 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 the following way, in which the $\lambda=(\lambda_Y, \lambda_{X1}, \ldots, \lambda_{Xk})$ Box-Cox parameter is estimated simultaneously with the other parameters :

$$\begin{cases} Y_t^{(\lambda_Y)} = \sum_{k=1}^{K} \beta_k X_{kt}^{(\lambda_{X_k})} + u_t \\ u_t = v_t \sqrt{\exp\left(\sum \delta_m Z_{mt}^{(\lambda_{Z_m})}\right)} \\ v_t = \sum_{l=1}^{p} \rho_l v_{t-l} + w_t \end{cases} \tag{3.15}$$

with the Box-Cox transformation defined as a power transformation, of parameter λ, on any positive real variable V_t by :

$$V_t \rightarrow V_t^{(\lambda)} = \frac{V_t^{\lambda} - 1}{\lambda} \; si \; \lambda \neq 0$$
$$V_t^{(0)} = Log \, V_t \qquad (3.16)$$

3.2.4. Conclusion

As we have 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.

3.3. Classical linear and non-linear regression models

3.3.1. Classical linear regression models *(C. Brandstaetter & M. Gatscha, KUSS)*

3.3.1.1. Research problem

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.

3.3.1.2. Model objectives

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:

$$y = bx + a \qquad (3.17)$$

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. But 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.

3.3.1.3. Dataset

The dataset used to demonstrate linear regression is derived from accident data from Austria. In this example, the distribution and development of people who were killed in accidents based on monthly observations is shown in Figure 3.1.

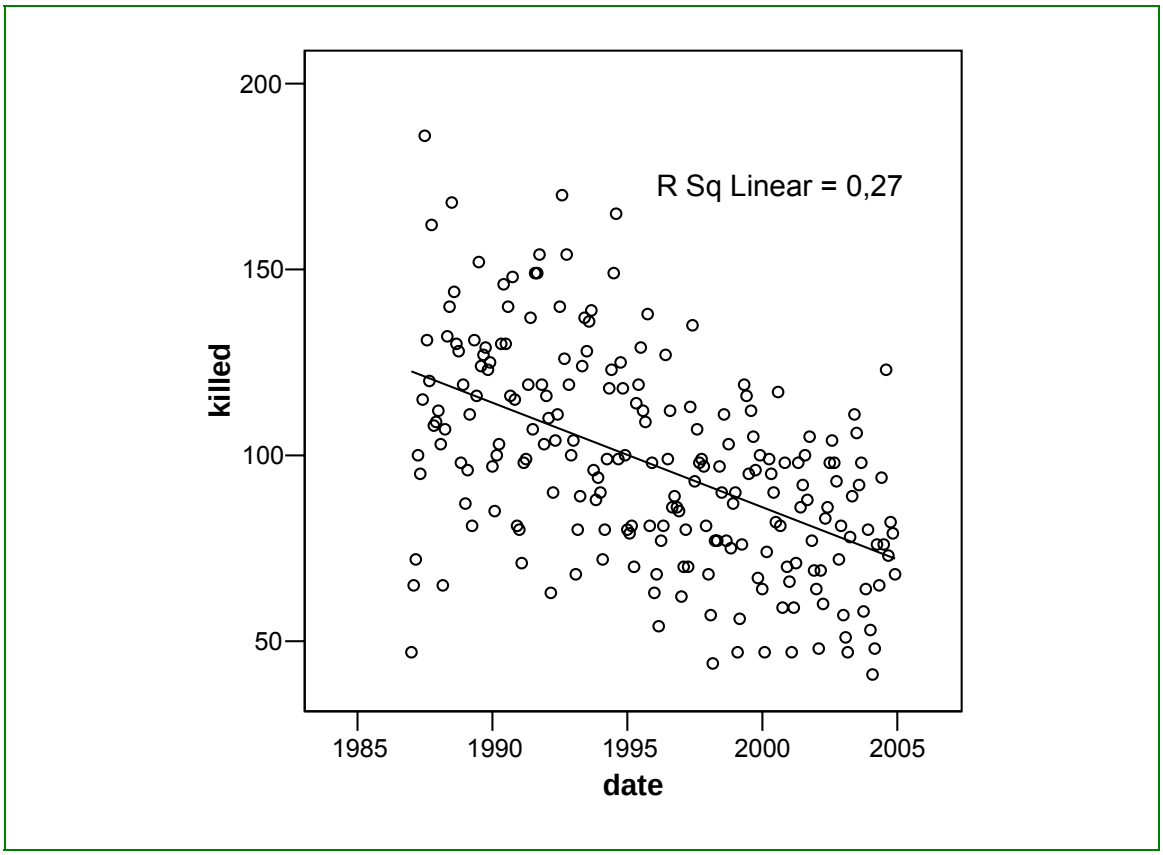


Figure 3.1: Scatterplot of accident statistics (number of fatalities) from 1987 to 2004

3.3.1.4. Model definition

The most basic relationship between two or more interval-scaled variables is explained by the following equation to determine the regression:

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + e_i \qquad (3.18)$$

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 like the first mentioned term:

$$y = bx + a \tag{3.19}$$

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 scatterplot (see Figure 3.2). 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:

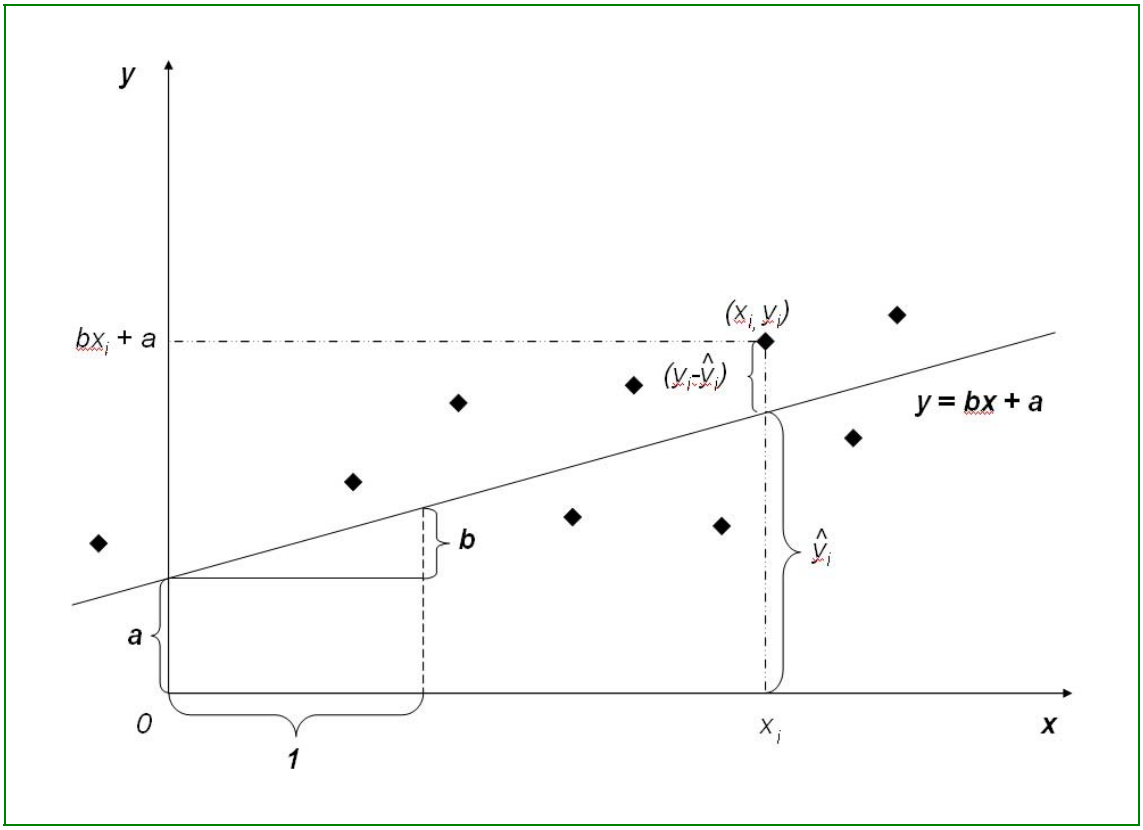


Figure 3.2: Scatterplot with regression line

In Figure 3.2, 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 mistake". 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 $(y_i - \hat{y}_i)$. But it is also possible that this sum is a negative value, therefore it can also be assumed that many regression lines exist where 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:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \tag{3.20a}$$

using ($bx_i + a$) instead of $\hat{y}_i$, the equation looks like:

$$\sum_{i=1}^{n}[y_i - (bx_i + a)]^2 = \min \qquad (3.20b)$$

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:

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x} \qquad (3.20c)$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (3.20d)$$

In the equations mentioned above, *n* is the number of data points in the time series, i.e. 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 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 unbiased and efficient 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, when 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 equal with the definition of a dependent and independent variable, although the latter relationship characterises a narrower, causal relationship.

3.3.1.5. Model assumptions

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 line $y = bx + a$. 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 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. normal distribution of accidents.

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 4 different conditions; 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 SE_b (SE is standard error) will be inefficient - too large or too small. This yields wrong significance tests because $t=(b/SE_b)$.

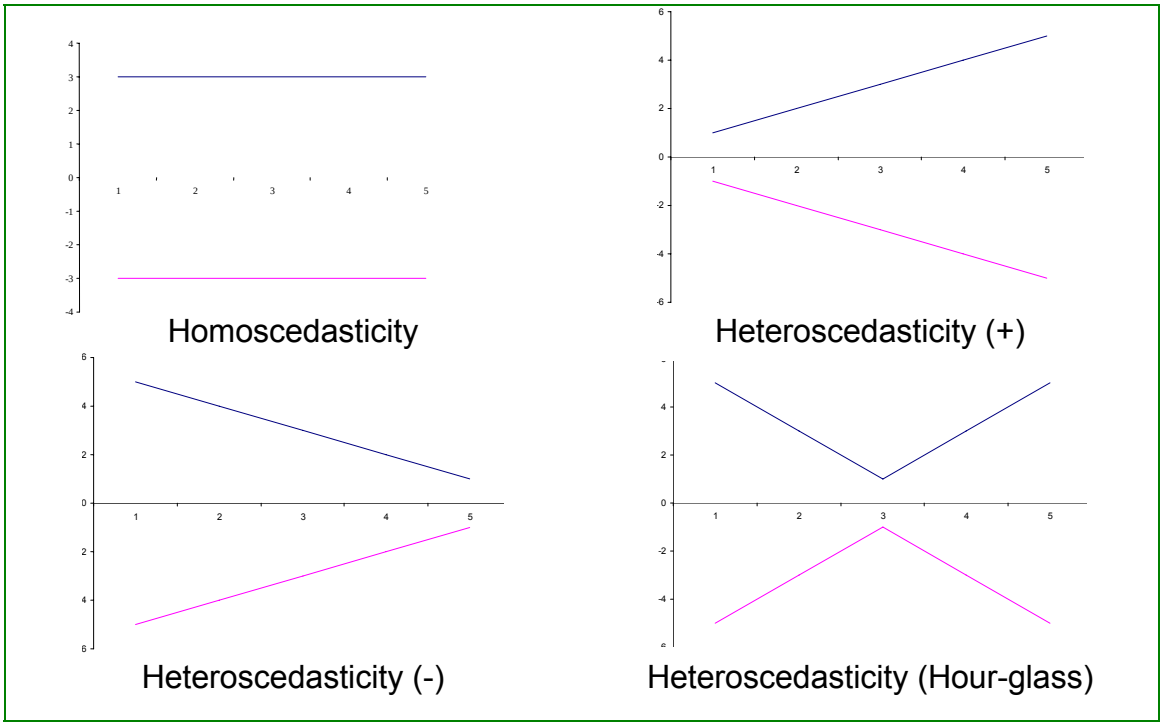


Figure 3.3: Distribution of the residuals and violations of the homoscedasticity assumption

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 (White, 1980) has published a direct test for heteroscedasticity:
$\chi^2 = R^2 n$,
where n is the number of cases, R^2 is the squared multiple correlation coefficient for the regression of the squared residuals on x, and 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:

$$H(h) = \frac{\sum_{t=n-h+1}^{n} e_t^2}{\sum_{t=d}^{h} e_t^2} \qquad (3.21)$$

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-distribution with df=h,h.

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. 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 negative residuals. As pointed out in the literature (see Ostrom, 1990, and 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 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 (G. M. Ljung and G. E. P. Box, 1978) test 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

$$Q_{LB} = n(n+2)\sum_{j=1}^{h}\frac{\rho^2(j)}{n-j} \tag{3.22}$$

with H_0 that the data is random, n is the sample size, $\rho(j)$ is the autocorrelation at lag j, and h is the number of lags being tested. The hypothesis of randomness is rejected if

$$Q_{LB} > \chi^2_{1-\alpha;h} \tag{3.23}$$

where χ^2 is the percent point function of the chi-square distribution.

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 which 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 shows a summary of the different assumption violations and their consequences.

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: 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.

3.3.1.6. Model fit and diagnostics

The used dataset is based on accident data from Austria in order to show the relationship/development of fatal accidents all over the country from 1987 to 2004 on a monthly observation basis. 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).

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. The next Table 3.4 shows the coefficients of the regression:

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.

Model	*Sum of Squares*	*df*	*Mean Squares*	*F*	*Significance test*
Residual	46129,86	1	46129,86	79,23	,000[a]
Regression	124600,50	214	582,25		
Total	170730,30	215			

Table 3.3: ANOVA table of sample dataset with time as predictor and the number of fatal accidents in Austria as dependent variable

a. Predictors: (Constant), Time

Model		*Non-standardized coefficients*		*Standardized Coefficients*	*T*	*Significance test*
		B	*Standard Error*			
1	*Constant*	1259,42	130,56		9,65	,000
	Time	-8,91E-08	,000	-,52	-8,90	,000

Table 3.4 Coefficients table

Finally, the model summary table reports the strength of the relationship between the independent and the dependent variable (Table 3.5):

Model	*R*	*R-square*	*Adjusted R-Square*	*Standard Error of the Estimates*
1	,52	,27	,27	24,13

Table 3.5: Model summary table

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.
These results shown in the point above are only true if the basic conditions stated in the Gauss-Markov assumptions hold true. Based on the fact that linearity is only assured if the residual value (*e=error*) 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):

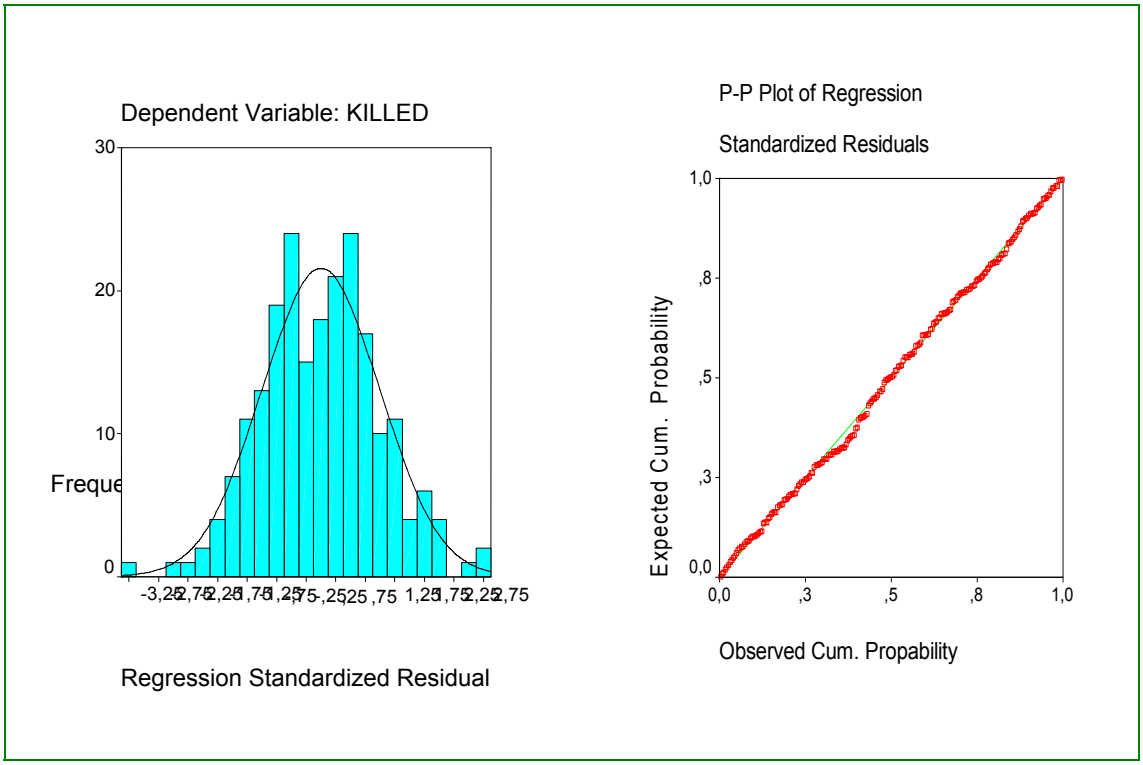


Figure 3.4: Histogram and P-P Plot of standardized residuals (in other chapters also the Q-Q Plot is used)

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 3.4). 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 3.6) shows the following:

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 somewhat around zero, there is also no sign of outliers.

	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 3.6: Table of selected residual statistics

For testing the assumption of homoscedasticity, there are some heuristic ways by looking at different scatterplots (Figure 3.5). All plots in Figure 3.5 below 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 χ^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 are 216, df is 1) which is highly significant and shows the presence of heteroscedasticity in the data. Finally, we will look 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 3.6.

The two plots in Figure 3.6 below 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. These results are not an artifact of the seasonal component in the data series that is shown on the second plot above and will be outlined a bit more by performing the analysis again. Firstly, this is done by expanding the regression equation by adding a dummy variable for the month as a second predictor in the model. The second approach analyses aggregated yearly data as dependent variable.

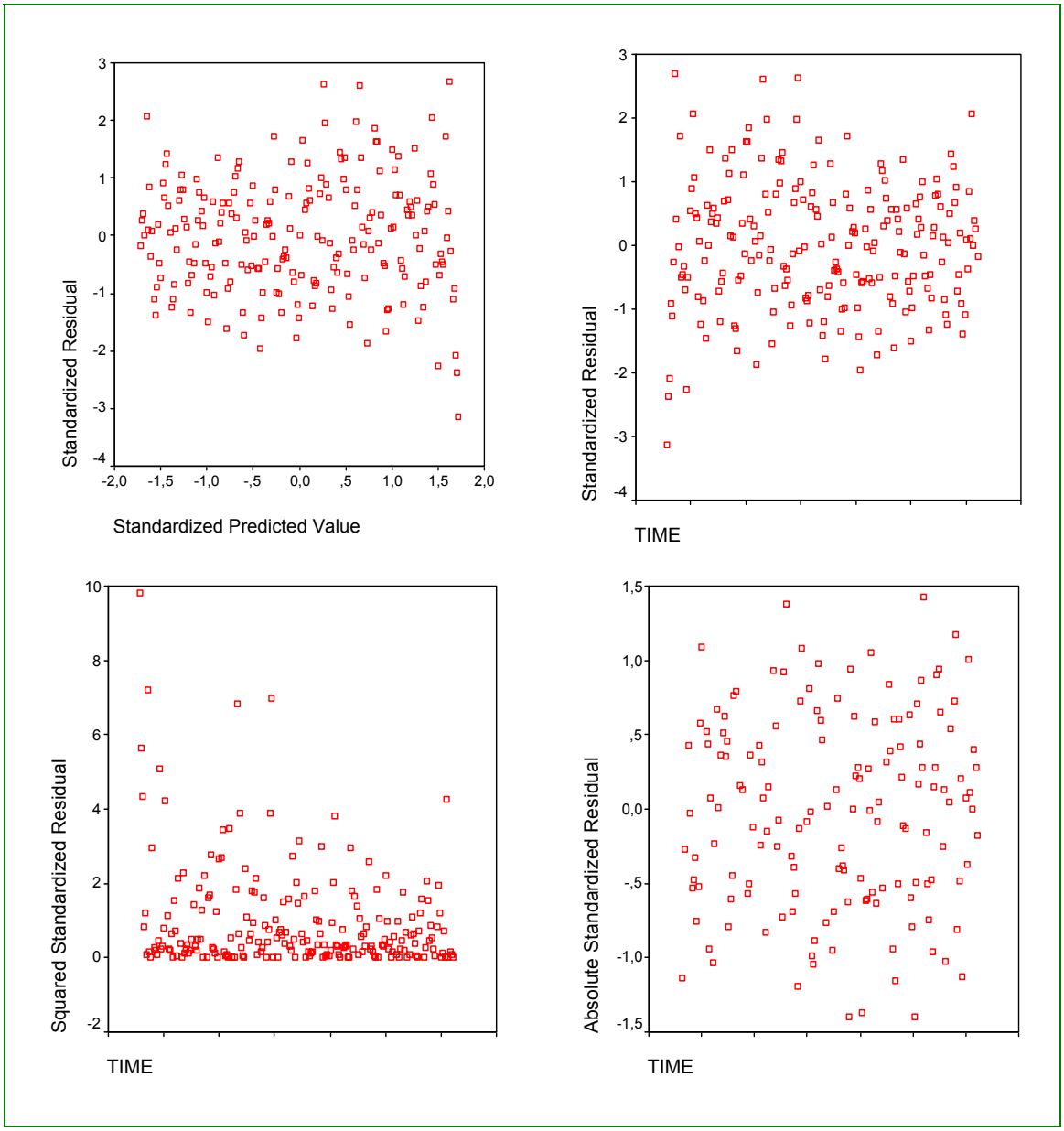


Figure 3.5: Table of selected residual plots for identifying heteroscedasticity

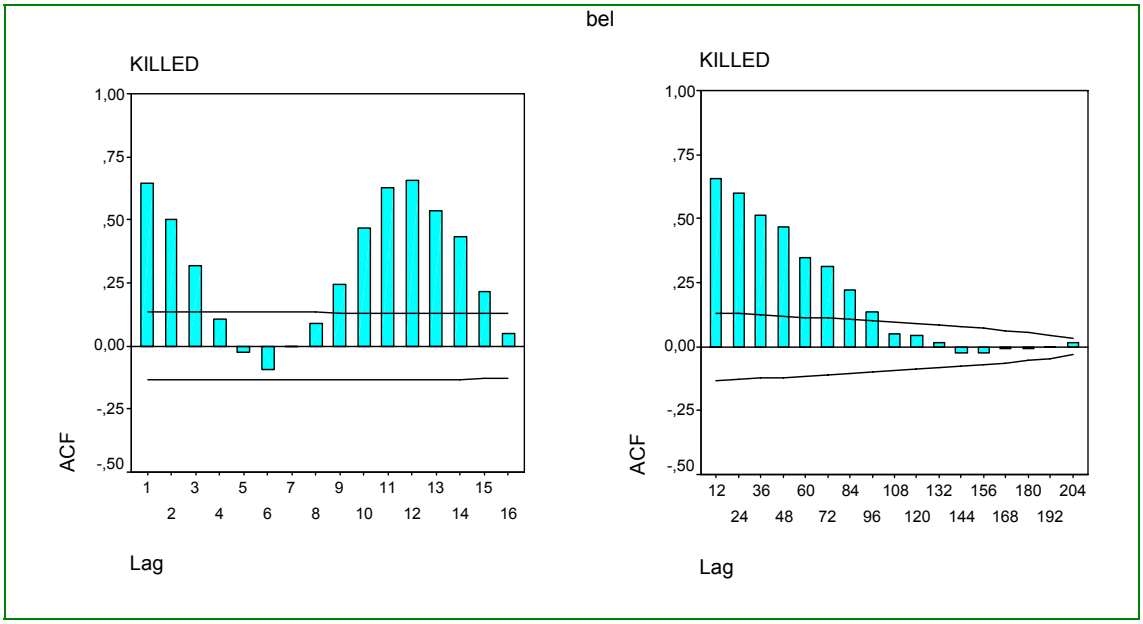


Figure 3.6: Table of autocorrelations and seasonal adjusted autocorrelations. Left side: Autocorrelations up to order 16 for the original series of Austrian fatalities (the lines indicate two standard errors) – Right side: Autocorrelations of the same month over the entire series

The model fit statistic (R Square) in Table 3.7 shows a better fit of the linear regression model including the dummy predictor month compared to the simple model above (0.423 vs. 0.27)

Model	*R*	*R-Square*	*Adjusted R-Square*	*Standard error of the Estimate*
1	,65 (a)	,42	,42	21,51

Table 3.7: Model summary table

a. Predictors: (Constant), month, time

Model		*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
1	*Regression*	72178,51	2	36089,26	78,00	,000(a)
	Residual	98551,82	213	462,69		
	Total	170730,33	215			

Table 3.8: ANOVA table of sample dataset

a Predictors: (Constant), month, time
b Dependent Variable: killed

The results in Table 3.8 are very similar compared to the results in tables 3.3 and 3.4. The inclusion of the second predictor month does not improve whether the F-test nor the parameters significantly.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1287,126	116,443		11,054	,000
	time	,000	,000	-,541	-10,385	,000
	month	3,186	,425	,391	7,503	,000

Table 3.9: Coefficient table of sample dataset

a Dependent Variable: killed

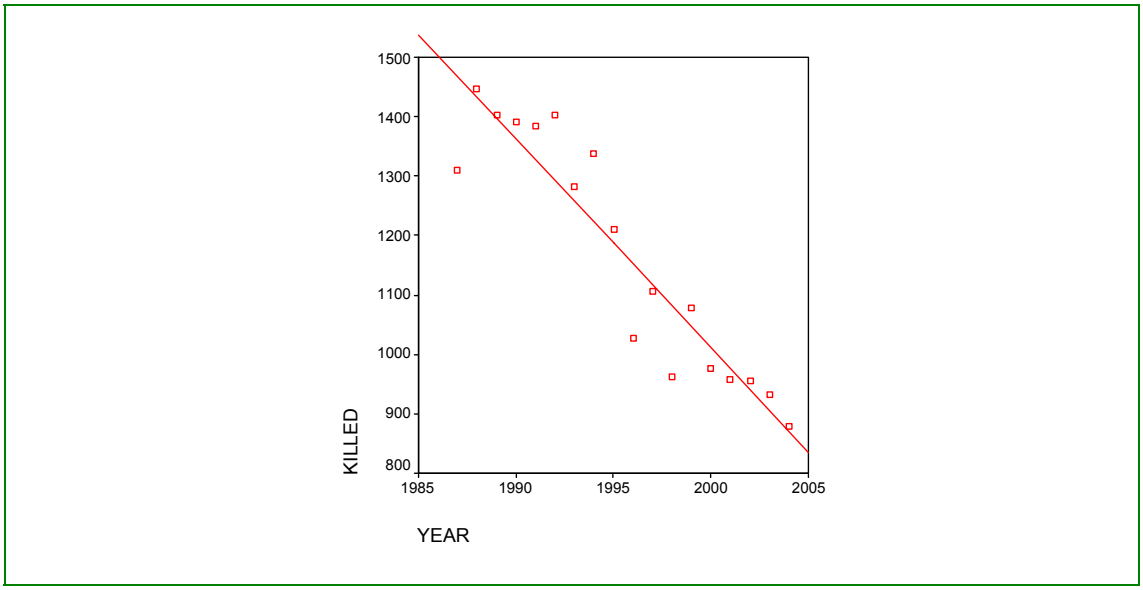


Figure 3.7: Plot of Yearly Fatality Data in Austria from 1987 to 2004

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	598633,50	1	598633,50	107,07	,000(a)
	Residual	89459,45	16	5591,22		
	Total	688092,90	17			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	71312,11	6778,90		10,52	,000
	Year	-35,15	3,40	-,93	-10,35	,000

Model	R	R-Square	Adjusted R-Square	Standard error of the Estimate
1	,93	,87	,86	74,77

Tables 3.10 to 3.12: Yearly Fatality Data in Austria from 1987 to 2004 - regression results

The previous result on monthly data is replicated on yearly data again. The number of fatalities in road accidents has been decreasing since 1987. The result is more significant than before because there are have no seasonal artefacts in the yearly data which introduce high variation not due to the general trend in the model. 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.

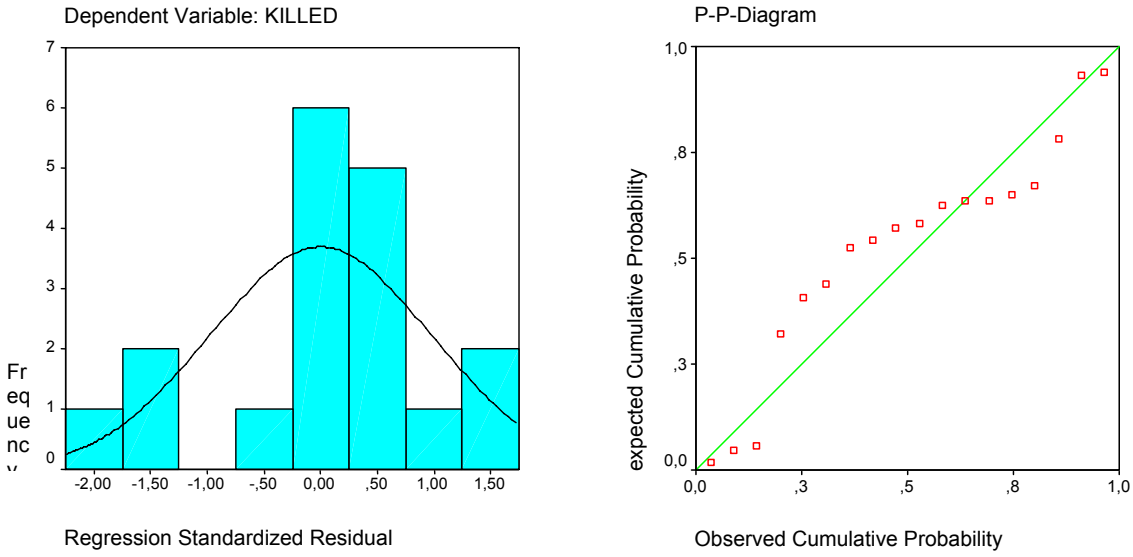


Figure 3.8: Histogram and P-P Plot of standardized residuals

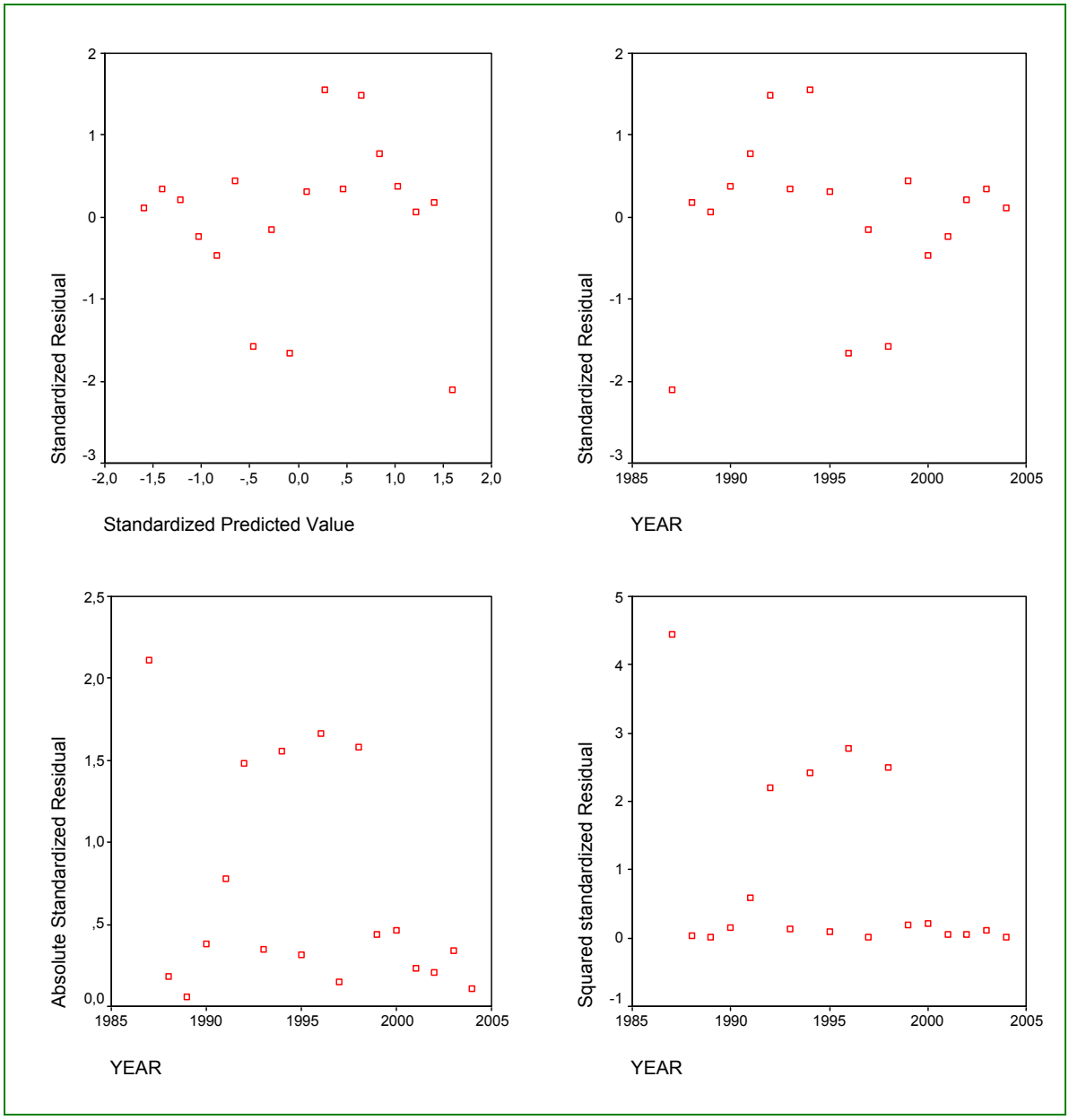


Figure 3.9: Table of selected residual plots for identifying heteroscedascity

All plots show a light trend of smaller residuals in the later years. This cannot be proven by White's test: χ^2 = 2,057, df=1, therefore we can assume ascity 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 last 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. But 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 about 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.

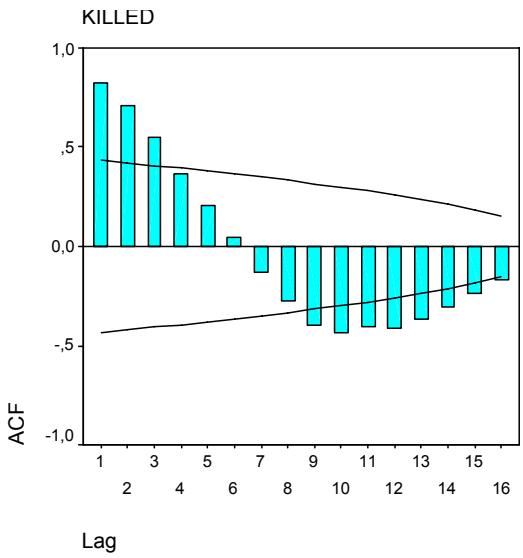


Figure 3.10: Table of autocorrelations

3.3.1.7. Model interpretation

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.

3.3.1.8. 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. Not only from a statistical view 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 costly decisions have to be performed 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 assumptions of these tests. 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 heteroscedasticit data.
Other chapters in this work will lead to an in depth view of the various options to handle the specific properties of accident data in more sophisticated model environments.

3.3.2. Generalized linear models (GLM) *(E. Papadimitriou & C. Antoniou, NTUA)*

3.3.2.1. Research problem

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).

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 parameter θ. 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:

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \tag{3.24}$$

where a, b, s, and t are known functions. The symmetry between y and θ becomes more evident if we rewrite it as:

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \tag{3.25}$$

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 $y_1, \ldots, y_N$, 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):

$$f(y_i;\theta_i) = \exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)] \tag{3.26}$$

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

$$f(y_i;\theta_i) = \exp\left[\sum_{i=1}^{N}(y_i b(\theta_i) + c(\theta_i) + d(y_i))\right]$$

(3.27)

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.

$$g(\mu_i)=\mathbf{x}_i^T\beta \quad (3.28)$$

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
$\beta=[\beta_1, \ldots, \beta_p]^T$ 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 β and explanatory variables $\mathbf{X}=[\mathbf{x}_1^T, \ldots, \mathbf{x}_M^T]^T$
3. A monotonic link function g such that

$$g(\mu_i)=\mathbf{x}_i^T\beta \quad (3.29)$$

where $\mu_i=E(Y_i)$

3.3.2.2. Dataset

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 Thessaloniki). Monthly data from January 1998 to December 2001 have been used for this research (Figure 3.11). The data of the first three years (36 observations) are used for the model estimation, while the data for the last year (12 observations) are used for validation.

3.3.2.3. Model definition

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 36 (for December 2000). Seasonal effects are captured by the addition of sinusoid components (used e.g. by Zeger, 1988, and Campbell, 1994). Several frequencies have been investigated, but the most useful proved to be the annual and its first (six month) harmonic.

Furthermore, besides specifying a trend and a seasonal component, the impact of explanatory variables is also tested, with an emphasis on enforcement data (number of breath alcohol controls per month).

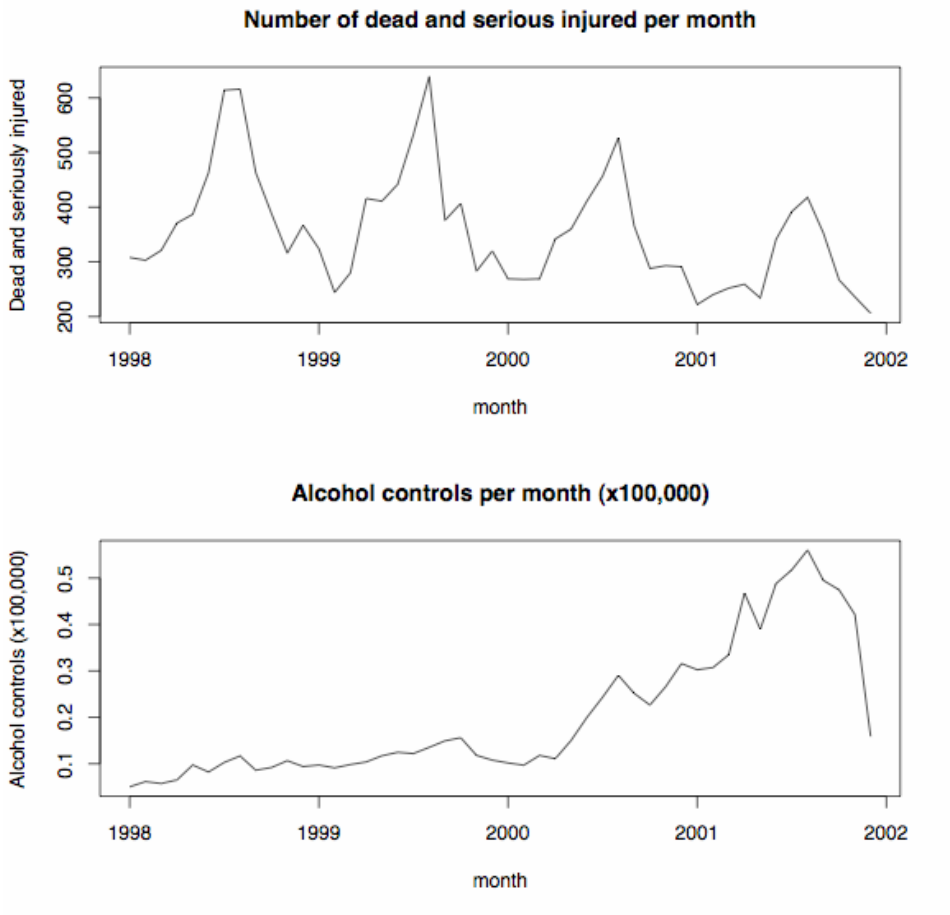


Figure 3.11: Dataset overview

3.3.2.4. Objectives of the technique

The objective of GLM is to allow for more flexible error structures (besides the Gaussian which is assumed by regression – linear or nonlinear). 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.

3.3.2.5. Model assumptions

Generalized linear models require uncorrelated observations. Time-series data require special consideration, since the observations typically fail to meet this assumption, as neighboring 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). Zeger concludes that "generalized linear models with linear and log links can be extended to parameter-driven models". 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 modeled 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).

3.3.2.6. Model fit and diagnostics

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 (R Development Core Team, 2005). First, the Gaussian (Normal) distribution is used. Since the identity link function is used and the model specification is linear additive, this is equivalent to the linear regression model. A Poisson model is also fitted, along with a quasi-Poisson that relaxes the assumption that the dispersion parameter is equal to 1. A negative binomial model is also fitted. A log-link function has been used for the Poisson, quasi-Poisson and negative binomial models.

Estimation results and model fit for the four models are shown in Table 2.23. Variables (such as the linear trend) that were found to be insignificant at the 10% level were removed from the model specification. 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 dead and seriously injured decreases as the intensity of breath alcohol controls increases, which is an intuitive result. Other explanatory variables (such as the number of speeding violations) were also entered into the model. However, explanatory variables relating to enforcement were highly correlated (corr=0.97). Therefore, while using either one resulted in intuitive results, their combination resulted in multicollinearity.

Due to the different link that is used in these models (identity for the normal, and log for the other three), the magnitude of the estimated coefficients is very different for the normal. The coefficient signs, however, are consistent for all models. The intercept and sinusoid terms are very significant, while the alcohol controls are significant at a 5% to 10% level.

An exception to this rule is the Poisson model, which shows very high significance of all coefficients. A closer look at the model statistics, however, suggests that the data maybe 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 scale parameter not equal to one does not necessarily imply overdispersion. This can 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.

The dispersion factor for the data at hand is equal to 270.73/32=8.46, which is significantly different than 1. The assumption of a Poisson model (with a dispersion parameter equal to 1) 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 1) 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 indeed underestimated due to the overdispersion. The scale parameter for the quasi-Poisson model is 31.763/32=0.99, i.e. very close to 1, indicating that overdispersion has been effectively handled by the estimated dispersion parameter.

Finally, a negative binomial model has been fitted. The estimated coefficients are identical (up to two significant digits) with those obtained from the quasi-Poisson. This confirms the findings of Maher and Summersgill (1996) who state that the two approaches may give similar estimation results. Slightly lower standard errors for the binomial, however, lead to more significant statistics.

Further model diagnostics are presented in Figures 3.12 through 3.15. 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.

		Normal	
Coefficient	Estimate	Std. error	t-value
Intercept	303.48	34.44	8.813
sin(pi*Month/6)	-89.65	17.00	-5.266
sin(pi*Month/12)	191.27	33.54	5.702
Alcohol controls (x100,000)	-322.82	180.93	-1.784
Null deviance:		376861	(35 d.o.f.)
Residual deviance:		128560	(32 d.o.f.)

		Poisson	
Coefficient	Estimate	Std. error	z-value
Intercept	5.699	0.029	196.471
sin(pi*Month/6)	-0.231	0.013	-17.118
sin(pi*Month/12)	0.532	0.029	18.471
Alcohol controls (x100,000)	-0.876	0.149	-5.854
Null deviance:		931.49	(35 d.o.f.)
Residual deviance:		270.73	(32 d.o.f.)

		Quasi-Poisson	
Coefficient	Estimate	Std. error	z-value
Intercept	5.718	0.082	70.122
sin(pi*Month/6)	-0.227	0.040	-5.665
sin(pi*Month/12)	0.483	0.079	6.071
Alcohol controls (x100,000)	-0.776	0.428	-1.810
Null deviance:		103.019	(35 d.o.f.)
Residual deviance:		31.763	(32 d.o.f.)

		Negative binomial	
Coefficient	Estimate	Std. error	z-value
Intercept	5.718	0.077	74.630
sin(pi*Month/6)	-0.227	0.037	-6.037
sin(pi*Month/12)	0.484	0.075	6.471
Alcohol controls (x100,000)	-0.778	0.402	-1.932
Null deviance:		120.575	(35 d.o.f.)
Residual deviance:		36.009	(32 d.o.f.)

Table 3.13: Estimation results

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):

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_i)}{(n-1)s_x^2} \qquad (3.30)$$

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.

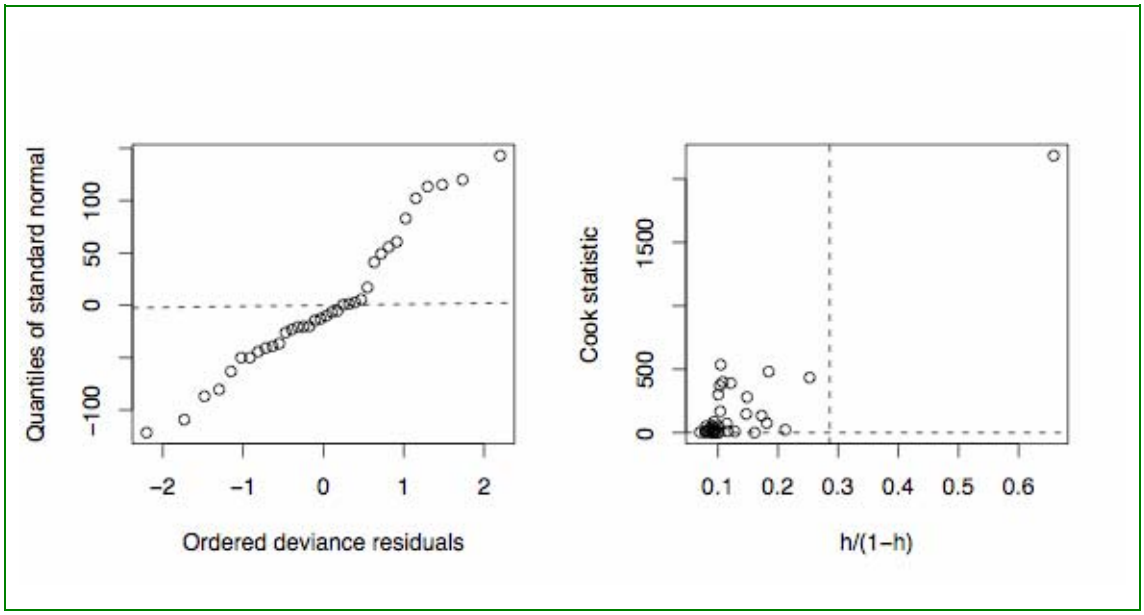


Figure 3.12: Model fit diagnostic plots (Gaussian distribution)

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.)

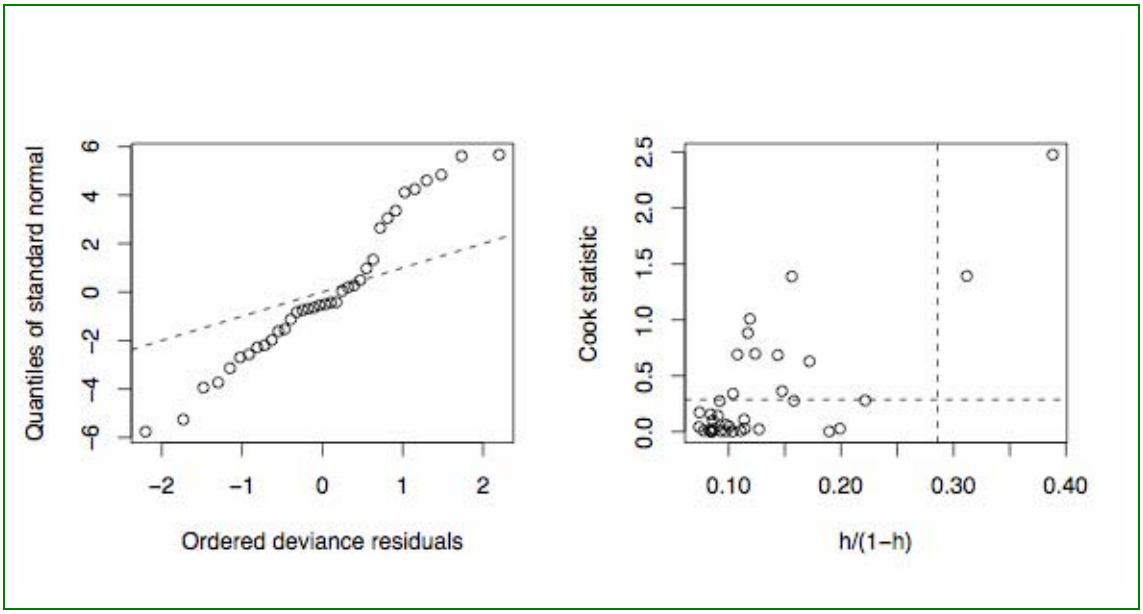


Figure 3.13: Model fit diagnostic plots (Poisson distribution)

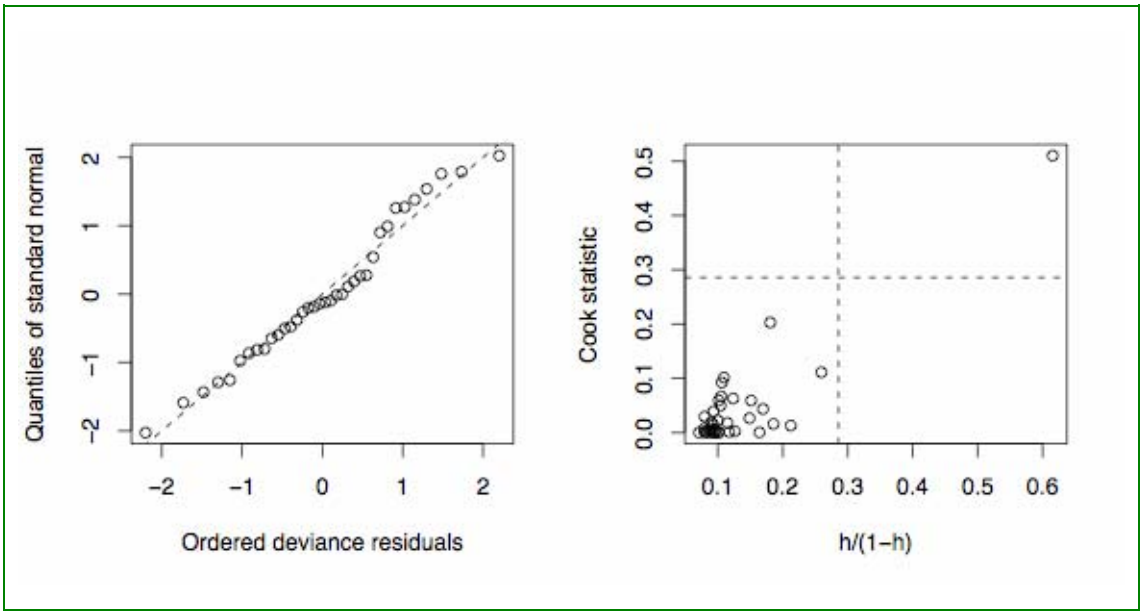


Figure 3.14: Model fit diagnostic plots (Quasi-Poisson distribution)

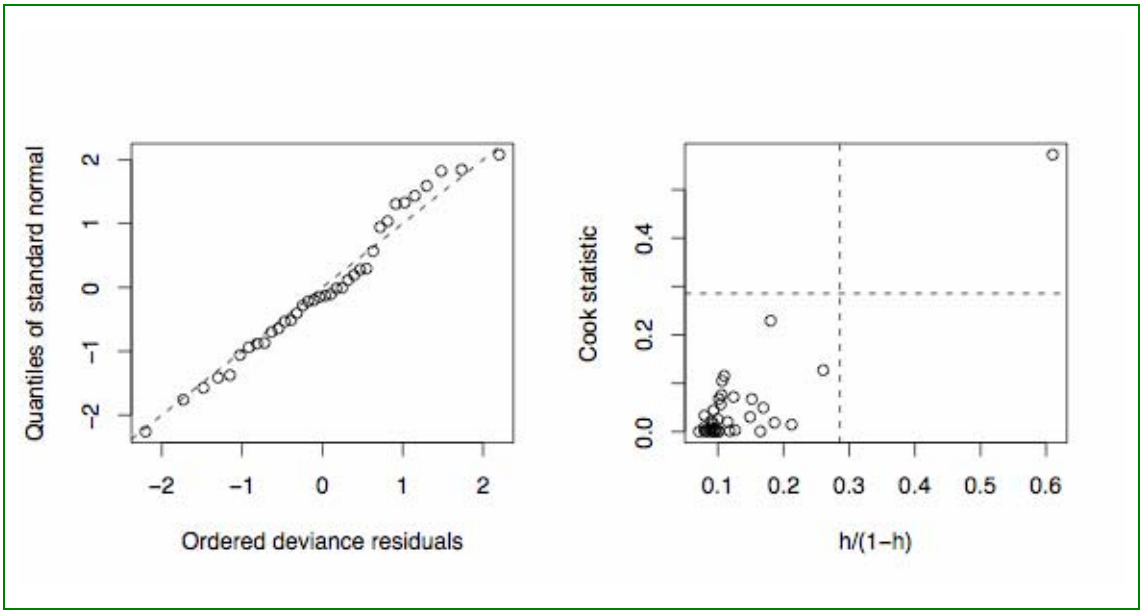


Figure 3.15: Model fit diagnostic plots (Negative binomial distribution)

Figure 3.16 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 Thessaloniki). The thick solid line represents the model predictions and 95% confidence intervals are also shown with thinner solid lines.

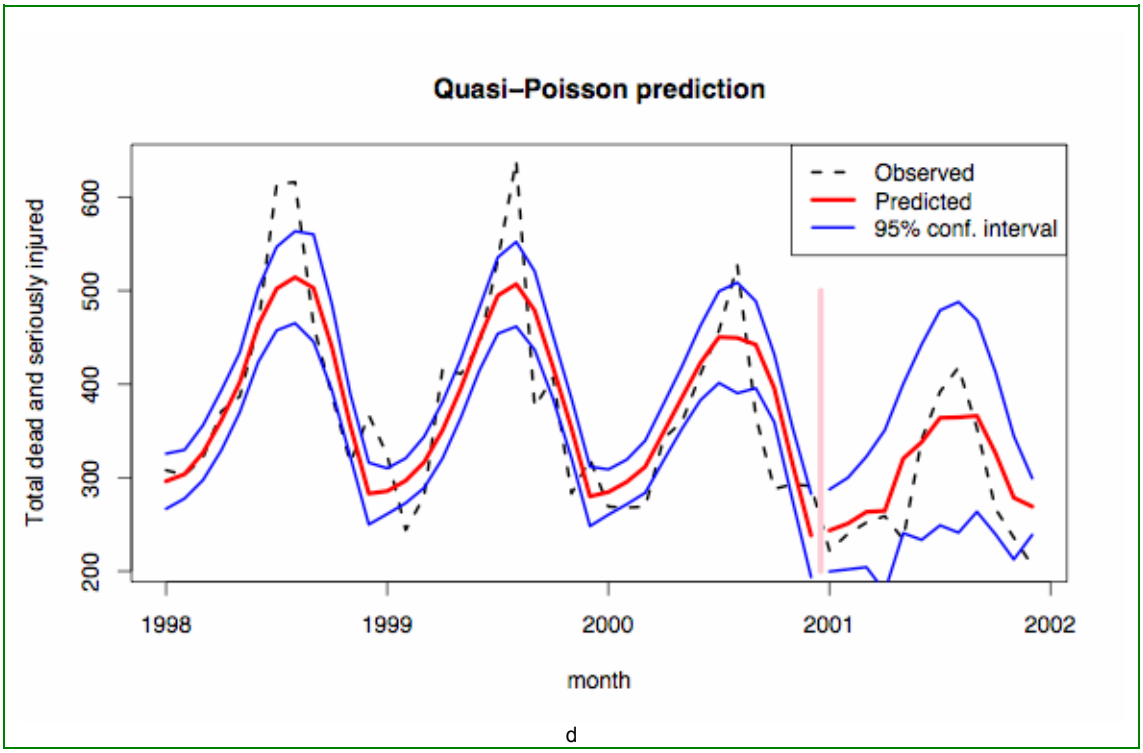


Figure 3.16 Quasi-Poisson model predictions

3.3.2.7. Model interpretation

The above discussion illustrates the impact of the distributional assumptions for the dependent variables on the model estimation results. While it is not easy to compare the estimated parameters between the Gaussian and the other models models (due to the different link function, i.e. identity for the normal and log for the Poisson, quasi-Poisson and negative binomial), the signs are consistent. The usual tests for comparing models (such as the Akaike Information Criterion, AIC, and the Bayes Information Criterion, BIC) are not suitable for comparison across these models. (While a detailed discussion is outside of the scope of this section, and there is a lot of specialized research on the topic, the AIC-type is best suited for the comparison of nested models and models with similarly computed log-likelihood measures.

The estimated coefficients between the two Poisson models are very similar (in terms of magnitude). The restrictive assumption of setting the dispersion parameter equal to one seems to lead to an underestimation of standard errors. This would in turn artificially increase the significance of the estimated coefficients. For example, the explanatory variable (alcohol controls) that appears to be significant at the p=0.01 level when dispersion is set equal to 1, is only significant at the p=0.10 level when the dispersion coefficient is estimated (quasi-Poisson model) or when the negative binomial model is used. The negative sign of the explanatory variable (alcohol controls) confirms the intuitive expectation that as the number of alcohol controls increases, the number of fatalities should decrease.

3.3.2.8. 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 modeling approach to capture serial correlation through the introduction of sinusoid latent processes has also been demonstrated.

While it is not easy to compare the estimated parameters between models with different link function (i.e. identity for the normal and log for the Poisson, quasi-Poisson and negative binomial), 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 quasi-Poisson and the negative binomial, 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. Furthermore, 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.

3.3.3. Non-linear models

- To be completed -

3.4. AR(I)MA(X) models *(R. Bergel, INRETS)*

3.4.1. Research problem

As it has already been emphasised before, 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[8]- , it is very practical to use the class of ARMA models, which enables to describe the dynamics of the process 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 (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 $Y = (y_1, y_{2,}.., y_n)$, at hand, 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.

The interest in this section is not in forecasting.
In the road safety field in France, as already mentioned in 2.2.2, 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, constructed on a monthly basis and for the last 25 years - period, for analysing 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

[8] *its mean, variance and covariance structure are constant over time (see a precise definition in section 2.2.4.3.1.)*

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 of1988 and 1995 – for which the information was carried by the media.

3.4.2. Dataset

The data set is the monthly number of fatalities in France, for the period january 1975-december 2001.

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 and calendar variables that take account for transitory effects, 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 section 2.2.4.3.3.

3.4.3. Model definition

3.4.3.1. ARMA and ARIMA models

We want to model a process $(X_t)t \in Z$, of second order[9], of which we have a set of data $X = (x_1, x_{2,}.., x_n)$.
This process is stationary if its mean, variance and covariance structure do not depend on time :

$$\begin{aligned} &E(X_t) = \mu \\ &\text{var}(X_t) = \sigma^2 \\ &\text{cov}(X_t, X_{t+l}) = \text{cov}(X_k, X_{k+l}) \end{aligned} \qquad (3.31)$$

The fist condition define the fist order stationarity, and the two following conditions define the second order stationarity.

That constant covariance structure allows separating X_t in two parts : the one related to the past at time t , and the part that is new at time t. This latter part of X_t that is not correlated to its past is also called « innovation » white noise.

The canonical ARMA(p,q) representation :

$$(1) \qquad X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + u_t + \sum_{j=1}^{q} \theta_j u_{t-j} \qquad (3.32a)$$

is usually written in the following way :

[9] *Having a finite mean and a finite variance*

$$\Phi(B)Y_t = \Theta(B)u_t, \quad (3.32b)$$

with : $\Phi(B)$ and $\Theta(B)$ the two polynomials of the delay operator *B*, of degrees p and q,
unitary, with no common root,
the roots of Φ are (strictly) outside the unit circle,
the roots of Θ are outside the unit circle
and u_t the « innovation » white noise.

In the general case where X_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 by :

$$Y_t = F(B)X_t, \quad (3.33)$$

and $F(B) = (I - B)^d$, with *B* the delay operator and d a positive value,

becomes stationary, and then model 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 X_t :

$$\Phi(B)F(B)X_t = \Theta(B)u_t \quad (3.34a)$$

which 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 :

$$\Phi(B)\Phi_s(B^s)F(B)X_t = \Theta(B)\Theta_s(B^s)u_t, \quad (3.34b)$$

and $F(B)=(I-B)_d(I-B_s)_D$, with *B* the delay operator, d and D two positive values and s the periodicity of the seasonal.

This last representation will be used in the case of the application.

3.4.3.2. AR(I)MAX models

There are different manners of taking account of exogenous (or explanatory) variables. The following form can be preferred for commodity reasons, in the case the data corrected for the exogenous effects are stationary :

$$\Phi(B)\left[Y_t - \sum_{i=1}^{K} \Phi_i(B)Z_{it}\right] = \Theta(B)W_t \quad (3.35)$$

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.

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.

3.4.3.3. The application case

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 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, specially low values of the number o 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.

Finally, 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:

$$\Phi(B)(I-B12)\left[\log Y_t - \sum_{i=1}^{I}\alpha_i LogX_{i,t} - \sum_{j=1}^{J}\beta_j X_{j,t} - \sum_{k=1}^{3}\sum_{l=0}^{n_k}\delta_{l,k}P^{T_{0,k}}(t-l)\right] = \mu + \Theta(B)a_t \quad (3.36)$$

with :

- Y the number of fatalities,
- $X_{i,i=1à I}$ the I variables measuring risk exposure and the economic factors,

- $X_{j,j=1àJ}$ 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=$T_{0,k}$ et 0 elsewhere, $T_{0,k}$ the first month of the intervention period n° k,
- n_{k+1} le 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.

3.4.4. Objective of the technique

The objective of the technique is to estimate the parameters of the dynamics of the stationary process corrected for exogenous effects, and to simultaneously estimate the parameters of the exogenous variables.

Tests are then used to validate the model, and criteria to evaluate the model's empirical performance.

3.4.5. Model assumptions

The main assumption is the stationary of the data, corrected for the exogenous effects.
In fact, this hypothesis is tested on the residual of the model, which should be a white noise in the case that this hypothesis is valid.

3.4.6. Model fit and diagnostics

In the case of pure ARIMA models, the well known following stages[10] are succeedingly considered : stabilisation , identification, estimation, and validation.

In the general case where exogenous variables are introduced, it is not feasible to consider these stages, before the functional form between Y and each exogenous variable Z_i has been established, and before a preliminary estimation of the exogenous effects has been obtained, in such a way that a realisation of the stationary process, corrected for the exogenous effects, is at hand.
In practise, the parameters are estimated altogether, whether related to the endogenous or exogenous variables; and the diagnostic tests, carried out after

[10] *see Brockwell and Davis(1998)*

the model has estimated, will help for replacing the two first stages (stabilisation and identification) which could not be considered before.

3.4.6.1. Validation and performance

Tests are used for validating the model.

We have to be distinguish between the tests used for validating each parameter (Student's test), and the tests related to the residual. These latter ones consist in testing the « randomness » property (up and down test) , the white noise property (test based on the Shapiro-Wilk's statistic), the gaussian property (Fisher's test), non-correlation property (Ljung-Box's test) - this last property being fundamental. Thus, in the case the assumption of normality is not validated, the log likelihood computation can be blamed, but the estimators may nevertheless have good asymptotic convergence properties. However, the assumption of non correlation of the residual is fundamental, because in the case it is rejected the model's specification has necessarily to be changed

Criteria 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 part of explained variance (R2[11] or adjusted R2), as well as the different criteria which enable to evaluate the estimation fit : MSE, RMSE, and the widely used MAPRE[12], and in the second group the BIC or the AIC[13], and the SBC[14].

Several models proposed for the same sample of data will be compared after the test and criteria, just mentioned above, have been performed. Two nested models will be compared by using a likelihood ratio test, which can lead to reduce the number of parameters of an over-parameterised model.

A practical question finally is, after the model has been validated, the question of the model's stability over time. The parameters' stability will be discussed by comparing estimations obtained from different samples of data covering

---

[11] *Let's recall the part of explained variance or R2 :* $\frac{\sum_{t=1}^{n}\left(Y_t-\hat{Y}_t\right)^2}{\sum_{t=1}^{n}\left(Y_t-\bar{Y}\right)^2}$

[12] *The MAPRE is defined by :*

$MAPRE\_R=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t-\hat{Y}_t}{Y_t}\right|$ *or by* $MAPRE\_E=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t-\hat{Y}_t}{\hat{Y}_t}\right|$*, depending on whether one refers to the estimation's deviation to the realisation (model's performance) or to the realisation's deviation to the estimation (realisation's estimation)*

[13] *The Akaike Information Criteria (AIC) is given by :* $-2\ln(V_m)+2k$*, with* V_m *the likelihood at the maximum, and k*

[14] *The bayesian criterium of Schwarz (SBC) is given by :* $-2\ln(V_m)+\ln(n)k$

different time intervals. The responses to the validation tests and empirical performance criteria might also differ with each new sample of data.

3.4.6.2. The application case

The forms suggested by the autoregressive polynomial $\sum_{l=0}^{n_k} \delta_{l,k} P^{T_{0,k}}(t-l)$ is a step in all the three cases (see graphs n° 1 to 3) . The initial model (3.36) has been simplified by using three variables representing steps :

$$\Phi(B)(I-B12)\left[\log Y_t - \sum_{i=1}^{I} \alpha_i LogX_{i,t} - \sum_{j=1}^{J} \beta_j X_{j,t} - \sum_{k=1}^{3} \gamma_k Step_{k,t}\right] = \mu + \Theta(B)a_t \qquad (3.37)$$

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.

At last, the model (3.37) 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. Doing this way leaded to restrict the second period to December 1994 - June 1995, without modifying the first one.

The results obtained by estimating model (3.37) are given in Tables 2.24 and 2.25.
All parameters related to the exogenous variables were kept in the model in the first case, whether significant or not, whereas they were only kept in the model if significant at a given confidence level (T-ratio superior to 1).
As for the parameters related to the dynamics, they were only kept if significant at the usual 95% confidence level (T-ratio superior to 2).

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.

3.4.7. 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, the simple filter $(I-B12)$ was used in equation (3.37).

As for the exogenous part, it's natural to try to interpreter the relationship[15] between the exogenous variables Z_{it}, i=1 to k and the endogenous variable Y, regardless of the dynamics.

The parameters related to explanatory variables seem to be acceptable, although 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 explanory variables, correlated oi oilsales.

The parameters related to climate and calendar variables are consistant with other results (Bergel, Depire, 2004).

The following comments focus on the intervention step variables.
Thus, succeeding to a "Cellier effect" of -6,1 % per month (average decrease of 6,1% in the number of fatalities between April and October 1987), the effect of the amnesty's perspectives of 1988 is estimated at +6,4% per month (average increase in the number of fatalities of 6,4% between November1987 and july1988), and the effect of 1995 is estimated at +3,8% per month (average increase of 3,8% in the number of fatalities per month between December 1994 and June 1995), see Table 3.14.

Measured with an absolute number of deaths, the effects of both perspectives of amnesty are estimated at 512 and 183 fatalities respectively. The associated confidence levels are 0,06 et 0,25 respectively.

When the number of exogenous variables are reduced – best model in terms of forecasting – the average impacts are respectively estimated at -6,5%, +6,5% and +3,5% per month during the three intervention periods, see Table 3.15. The effects of both perspectives of amnesty are then estimated at 512 and 183 fatalities respectively, with confidence levels of 0,04 and 0,28.

---

[15] 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:

$$\varepsilon_{Y/Zi} = \frac{\Delta Y / Y}{\Delta Z_k / Z_i}.$$

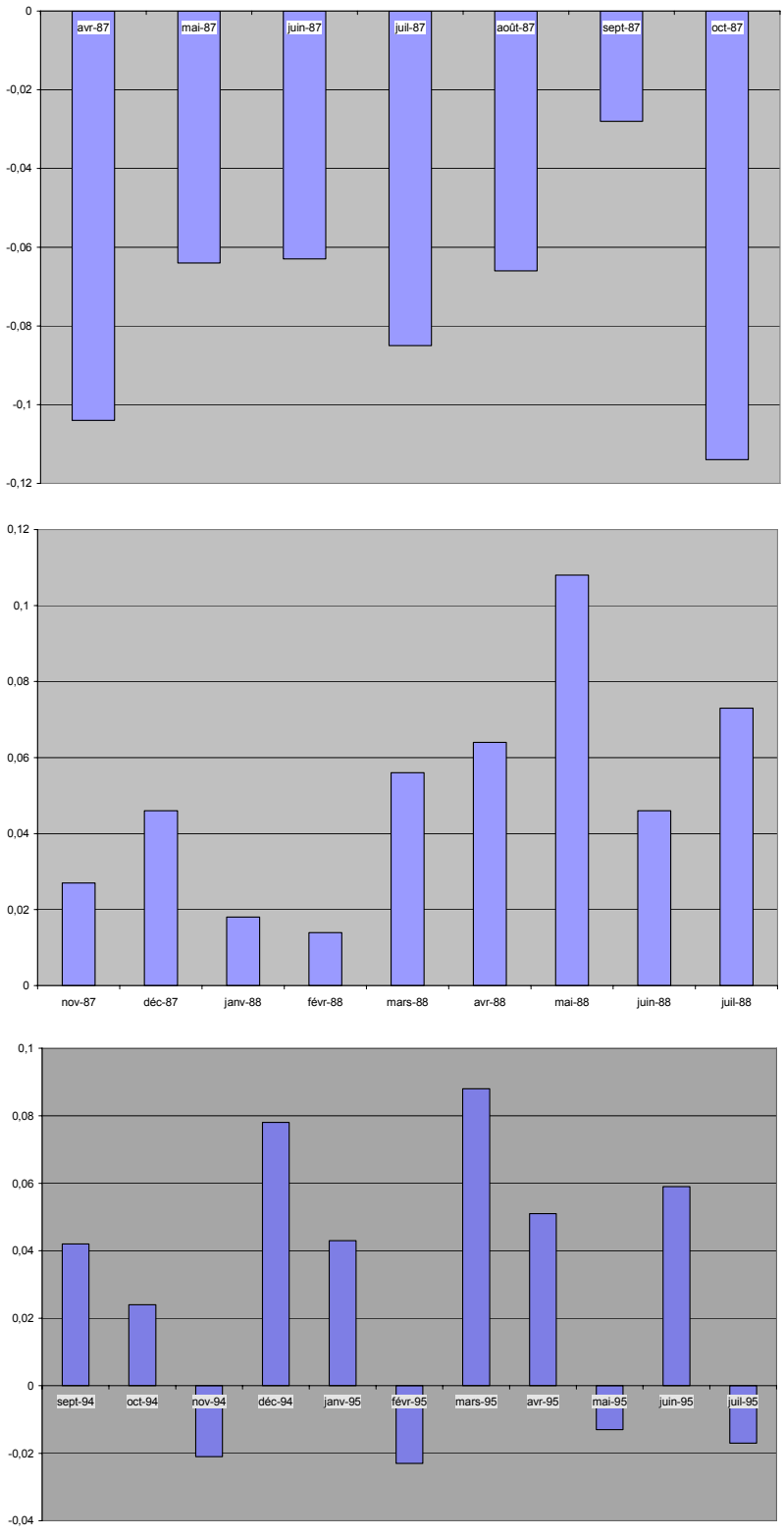
In the very special case where both variables have been log transformed, the parameter β_i indeed represents the elasticity of Y with respect to Zi, which is then constant. But it is important to note that one does generally comment an « apparent elasticity » of Y to Zi, because the condition of mutual orthogonality of the exogenous variables Z_{it}, i=1 to k, is rarely valid.

3.4.8. Conclusion

In this chapter, 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 and calendar nature. The possible effects of two presidential amnesties of driving faults, in 1988 and in 1995, on the number of fatalities in France was questioned by the means of an intervention analysis.

The amplitude of the effects of the perspectives of amnesty of 1988 is larger (400 to 500 additional fatalities, between September 1987 and July 1988) than it is in for the amnesty of 1995 (100 to 180 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. 517[16] additional fatalities, with a confidence level of 0,04.

[16] *The annual number of fatalities in France was around a thousand in the years 1990.*



Figures 3.17, 3.18 and 3.19: Monthly impacts of the "Cellier effect" (April - November 1987), of the perspectives of presidential amnesties of 1988 (November 1987 - July 1998) and 1995 (December 1994 - June 1995).

Parameter	Estimate	Std Error	T Ratio	Lag	Variable	Shift	
MU	-0.02505	0.0027828	-9.00	0	LTUEFE		0
MA1,1	0.78905	0.04110	19.20	12	LTUEFE		0
AR1,1	0.16913	0.06196	2.73	1	LTUEFE		0
AR1,2	0.16157	0.06374	2.53	2	LTUEFE		0
AR1,3	0.22444	0.06290	3.57	3	LTUEFE		0
NUM1	0.10347	0.08102	1.28	0	LCARBUB		0
NUM2	-0.0022755	0.08441	-0.03	0	LICARB		0
NUM3	0.0013002	0.0003383	3.84	0	DTE		0
NUM4	0.0008768	0.0005836	1.50	0	DTH		0
NUM5	0.00002586	0.00001965	1.32	0	DHPLUI		0
NUM6	-0.0035809	0.0026823	-1.34	0	DNGEL		0
NUM7	0.0043532	0.01137	0.38	0	ATYTES		0
NUM8	0.0013106	0.0038428	0.34	0	ATYTHS		0
NUM9	-0.01157	0.0038852	-2.98	0	ATYTHI		0
NUM10	-0.0022139	0.0030156	-0.73	0	ATYHS		0
NUM11	0.0015597	0.0028888	0.54	0	ATYNGLS		0
NUM12	-0.0083433	0.0081969	-1.02	0	S1		0
NUM13	-0.0060422	0.0089395	-0.68	0	S2		0
NUM14	0.0055669	0.0047984	1.16	0	S3		0
NUM15	0.01411	0.0050085	2.82	0	DP		0
NUM16	0.0026091	0.0026722	0.98	0	VESADI		0
NUM17	-0.06081	0.03512	-1.73	0	STEP1		0
NUM18	0.06417	0.03358	1.91	0	STEP2		0
NUM19	0.03813	0.03346	1.14	0	STEP3		0

Constant	Estimate =	-0.0111419
Variance	Estimate =	0.00414369
Std Error	Estimate =	0.06437153
AIC	=	-774.14637*
SBC	=	-685.17573*
Number of	Residuals=	301

* Does not include log determinant

Table 3.14: Model for the aggregate number of fatalities In France, for 1975-2001. (All exogenous variables kept)

with :
LTUEFE the log of the monthly number of fatalities,
LCARBUB the log of the monthly oil sales,
LICARB the log of the monthly car fuel price,
DTE-DTH the highest temperature of the day, in summer and in winter, DHPLUI the rainfall height and DNGEL the occurrence of frost, averaged on a hundred of meteorology stations and on the month,
ATYTES et ATYTHS the number of « superior » atypical days in the month regarding temperature, in the summer and in the winter,
ATYTHI the number of « inferior » atypical days regarding temperature, in the summer,
ATYHS the number of « superior » atypical days regarding rainfall height,
ATYNGLS the number of « superior » atypical days regarding occurrence of frost,
S1,S2,S3 the number of days coding a calendar exceptional effect related to bank holidays, gathered in three classes,
DP the number of days coding days a calendar effect related to holiday movements,

VESADI the number of week-end days (Friday/Saturday/Sunday),
STEP1 the dummy variable for the period April - November 1987,
STEP2 the dummy variable for the period November 1987 - July 1998,
STEP3 the dummy variable for the period December 1994 - June 1995

Parameter	Estimate	Std Error	T Ratio	Lag	Variable	Shift	
MU	-0.02452	0.0026660	-9.20	0	LTUEFE		0
MA1,1	0.77833	0.04043	19.25	12	LTUEFE		0
AR1,1	0.16357	0.05849	2.80	1	LTUEFE		0
AR1,2	0.14335	0.05910	2.43	2	LTUEFE		0
AR1,3	0.23408	0.05985	3.91	3	LTUEFE		0
NUM1	0.08976	0.07641	1.17	0	LCARBUB		0
NUM2	0.0013158	0.0003284	4.01	0	DTE		0
NUM3	0.0010321	0.0004655	2.22	0	DTH		0
NUM4	0.0000163	0.00001198	1.36	0	DHPLUI		0
NUM5	-0.0023010	0.0021175	-1.09	0	DNGEL		0
NUM6	-0.01055	0.0031936	-3.30	0	ATYTHI		0
NUM7	-0.0090827	0.0078523	-1.16	0	S1		0
NUM8	0.0044649	0.0044660	1.00	0	S3		0
NUM9	0.01387	0.0048195	2.88	0	DP		0
NUM10	0.0029882	0.0026133	1.14	0	VESADI		0
NUM11	-0.06497	0.03388	-1.92	0	STEP0		0
NUM12	0.06483	0.03206	2.02	0	STEP1		0
NUM13	0.03504	0.03238	1.08	0	STEP2		0

Constant	Estimate =	-0.0112535
Variance	Estimate =	0.00403826
Std Error	Estimate =	0.06354729
AIC	=	-819.52059*
SBC	=	-752.08894*
Number of	Residuals=	313

* Does not include log determinant

Table 3.15: Model for the aggregate number of fatalities In France, for 1975-2001(Exogenous variables kept if T-ratio superior to 1)

with :
LTUEFE the log of the monthly number of fatalities,
LCARBUB the log of the monthly oil sales,
LICARB the log of the monthly car fuel price,
DTE-DTH the highest temperature of the day, in summer and in winter, DHPLUI the rainfall height and DNGEL the occurrence of frost, averaged on a hundred of meteorology stations and on the month,
ATYTHI the number of « inferior » atypical days regarding temperature, in the summer,
S1, S3 the number of days coding a calendar exceptional effect related to bank holidays,
DP the number of days coding days a calendar effect related to holiday movements,
VESADI the number of week-end days (Friday/Saturday/Sunday),
STEP1 the dummy variable for the period April - November 1987,
STEP2 the dummy variable for the period November 1987 - July 1998,
STEP3 the dummy variable for the period December 1994 - June 1995

3.5. DRAG models *(R. Bergel, INRETS)*

3.5.1. Research problem

We address here 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) which relies on a multiple regression structure with autocorrelated 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.

A DRAG model is defined on the basis of (at least) 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 framework of the DRAG approach is well defined in (Gaudry, Lassarre, 2000).
In this framework, one 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 exposure, 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.

3.5.2. Dataset

No condition is required from the data.
Six DRAG models have already been performed whether at a national or at a regional level. Because of the voluminous database necessary for estimating a DRAG model, the DRAG approach can not be achieved without enough time and financial support An example of a DRAG-type model is given by the RES Model for France[17] .

[17] *With the financial support of the French Observatory for Road Safety, A DRAG-type model was applied to the french main road network in France (A-level roads and motorways), the two rnetworks on which the number of vehicle-kilometers are measured on a monthly basis. Only two risk factors were taken into account : the traffic volume, and the climatic factor.*

3.5.3. Model definition

The model is written as follows, the parameter $\lambda=(\lambda_Y, \lambda_{X1}, \ldots, \lambda_{Xk})$ being estimated :

$$\begin{cases} Y_t^{(\lambda_Y)} = \sum_{k=1}^{K} \beta_k X_{kt}^{(\lambda_{Xk})} + u_t \\ u_t = v_t \sqrt{\exp\left(\sum \delta_m Z_{mt}^{(\lambda_{Zm})}\right)} \\ v_t = \sum_{l=1}^{p} \rho_l v_{t-l} + w_t \end{cases} \quad (3.38)$$

with : Y_t the endogenous variable to be modelled,
X_{it}, i=1 to k, the exogenous (or explanatory) variables,
u_t the first residual, and v_t the final residual.

And with: the Box-Cox transformation defined as a power transformation, of parameter λ, on any positive real variable V_t by :

$$\begin{aligned} V_t \to V_t^{(\lambda)} &= \frac{V_t^{\lambda} - 1}{\lambda} \; si \; \lambda \neq 0 \\ V_t^{(0)} &= Log \, V_t \end{aligned} \quad (3.39)$$

Two well-known particular cases are obtained when the parameter λ is identically equal to 0 (we then have a log-log specification), or to 1 (we then have the linear specification).

3.5.4. Objective of the technique

The main objective of the technique, compared to a multiple linear regression is that the use of the Box-Cox transformation to 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.

3.5.5. Model assumptions

The endogenous variable is supposed to be gaussian (the data are aggregate, and their frequency is supposed to be larger than 30).
The stationarity is not required. The explanatory variables take account for trend and seasonality, 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.

3.5.6. Model fit and diagnostics

The model fit is performed with the TRIO program, all the parameters - linear and non-linear - being estimated simultaneously. Note that no procedure in the

SAS system, for instance, enables to estimate the parameters of the linear and non-linear parts simultaneously.

3.5.7. Model interpretation

Multicolinearity between the numerous explanatory variables is a source of difficulties in interpreting the estimated parameters. In some cases, the Box-Cox parameters may not be stable and interpretable either[18], and the model's specification seems to be over-parametrised.

In the general case, parameters are not interpreted directly, but provide ***elasticity values***, of the endogenous variables with respect to the exogenous variables - that is to say of risk indicators with respect to risk factors. These elastcity values, calculated at a country's level independently of the units of measure of risk indicators and risk factors, are used for international comparisons.
Detailed interpretations can be found in (Gaudry, Lassarre, 2000).

3.5.8. Conclusion

Because of the need of voluminous databases, it would not be feasible to apply the DRAG model to European data, in the SafetyNet project. Nevertheless, the theoretical framework is powerful, and is actually used for time series analysis for road safety research purposes.

3.6. State space models *(J. Commandeur, SWOV)*

In this section we present 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 explanatory and intervention variables. A third important

[18] *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.*

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). In the next section we start the presentation with the most simple structural time series model: the local level model.

3.6.1. The local level model

The first and most simple structural time series model is the local level model.

3.6.1.1. Research problem

The research problem addressed with the local level model is how to obtain an adequate description of an observed time series.

3.6.1.2. Dataset

The dataset in an analysis with the local level model simply consists of only one variable: a time series y_t consisting of repeated measurements of one and the same phenomenon at time points $t = 1, \ldots, n$.

3.6.1.3. Model definition

The local level model is defined as

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \qquad & \varepsilon_t &\sim NID(0,\sigma_\varepsilon^2) \\ \mu_{t+1} &= \mu_t + \xi_t, \qquad & \xi_t &\sim NID(0,\sigma_\xi^2) \end{aligned} \tag{3.40}$$

for $t = 1, \ldots, n$, where μ_t is the unobserved so-called *level* at time t, ε_t is the observation error or disturbance at time t, and ξ_t is the so-called 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 (3.40) 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 (3.40) 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.

3.6.1.4. Objectives of the technique

The objective of the local level model is therefore to establish whether an observed time series can be adequately described with a time-varying level component.

3.6.1.5. Model assumptions

In definition (3.40) 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.

3.6.1.6. Model fit and diagnostics

In the remaining part of this section we will first discuss and illustrate what happens when the level disturbances ξ_t in (3.40) are all fixed on zero, and then show 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 traffic 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).

If the level disturbances ξ_t in (3.41) 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

$$y_t = \mu_1 + \varepsilon_t, \qquad \varepsilon_t \sim NID(0,\sigma_\varepsilon^2) \tag{3.41}$$

for t = 1, …, 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 (3.41) are

$$\mu_1 = \bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t \tag{3.42}$$

and

$$\sigma_\varepsilon^2 = s_y^2 = \frac{\sum_{t=1}^{n}\left(y_t - \bar{y}\right)^2}{n-1} \tag{3.43}$$

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 (3.40) is simply

$$y_t = \bar{y} + \left(y_t - \bar{y}\right). \tag{3.44}$$

Applying deterministic level model (3.40) 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 $\sigma_\varepsilon^2 = 0.0485829$. 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.

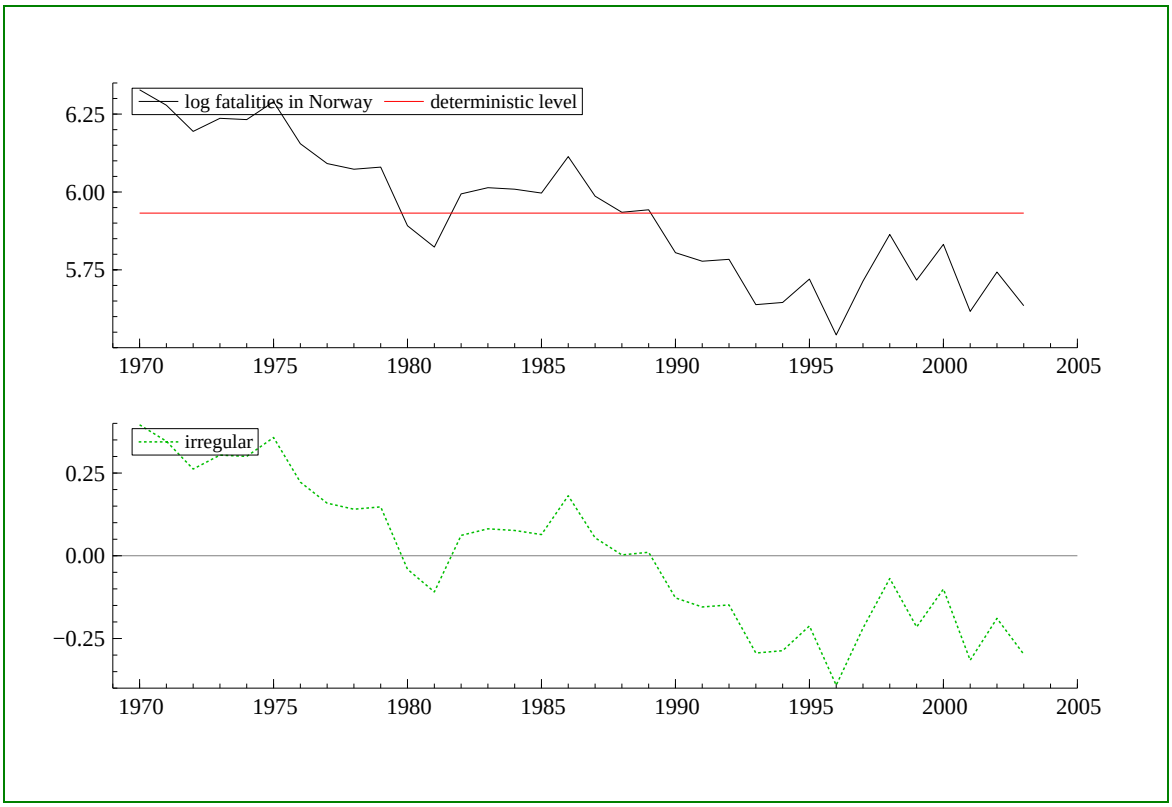


Figure 3.20: Deterministic level and irregular component for the log of Norwegian fatalities.

The level for model (3.41) is displayed at the top of Figure 3.20, 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 3.20 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 3.20 simply consists of the deviations of the observed time series from its mean, as already implied by (3.44).

Diagnostic tests for the assumptions of independence, homoscedasticity, and normality of the residuals of the analysis are presented in Table 3.16. For the exact definition, computation and interpretation of these diagnostic tests we refer 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 2.26. 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 3.16), 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 3.16 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 3.16: 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:

$$\mathrm{AIC} = \frac{1}{n}\left[-2n\log L_d + 2(q+w)\right], \tag{3.45}$$

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 (3.41) 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

$$\text{AIC} = \frac{1}{34}\left[-2(34)(0.038701012)+2(1+1)\right] = 0.040245.$$

Below, this value will be used for purposes of comparison with other state space models.

On the other hand, when the level in (3.40) 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.

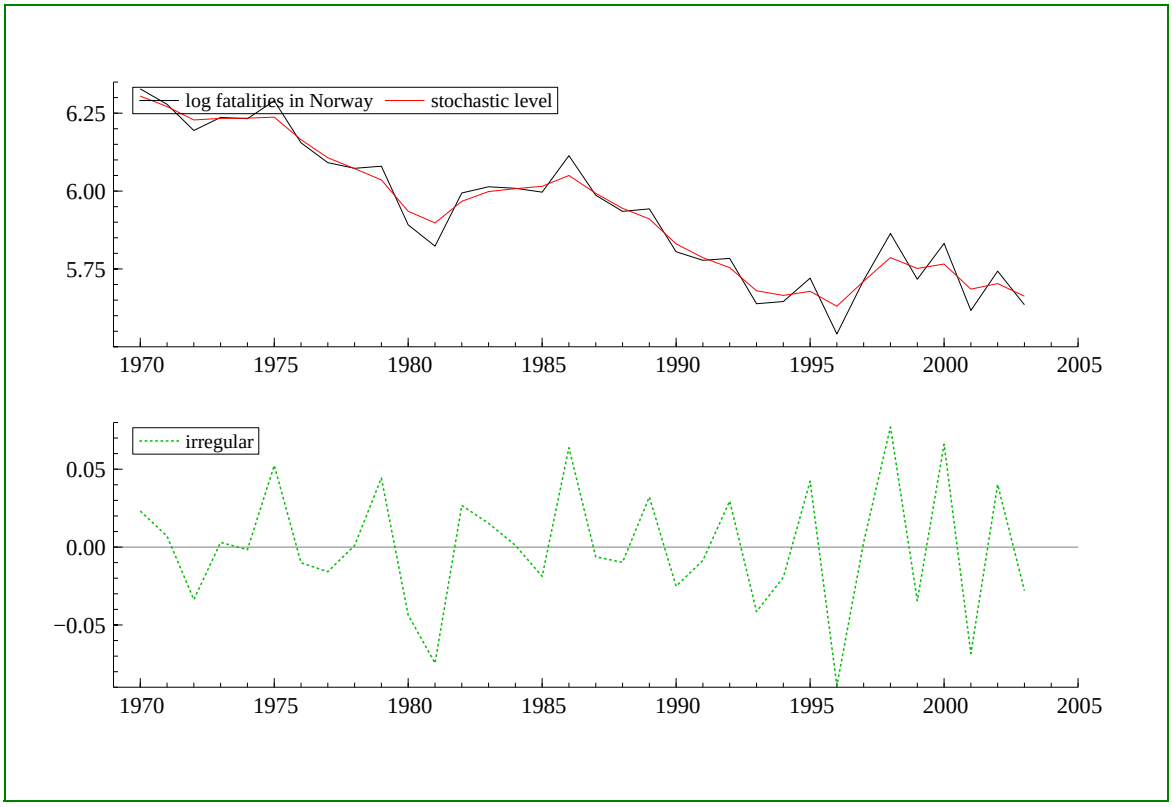


Figure 3.21: Stochastic level and irregular component for the log of Norwegian fatalities.

The local level for model (3.40) is illustrated at the top of Figure 3.21, together with the observed time series. As can be seen in Figure 3.21, 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 3.17: 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 3.21. The diagnostic tests for these observation disturbances are given in Table 3.17. 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

$$\text{AIC} = \frac{1}{34}\left[-2(34)(0.8468622) + 2(1+2)\right] = -1.51725.$$

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 (3.40) does not improve the description of the time series, since this results in an AIC value of only -1.28035.

3.6.1.7. Model interpretation

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. A second more general conclusion is that 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.

3.6.2. The local linear trend model

In this section we discuss the effects of adding a new component to the local level model, called the *slope* component. The *research problem* addressed with this model is again how to obtain an appropriate description of an observed time series. 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 repeated measurements of one and the same phenomenon at time points $t = 1, \ldots, n$.

The local linear trend model is obtained by adding a slope component ν_t to the local level model, and is *defined* as follows:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim NID(0,\sigma_\varepsilon^2) \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim NID(0,\sigma_\xi^2) \\ \nu_{t+1} &= \nu_t + \zeta_t, & \varsigma_t &\sim NID(0,\sigma_\varsigma^2) \end{aligned} \qquad (3.46)$$

for t = 1, …, 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 (3.46) 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 6.3.1). 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 (3.46) is allowed to change over time.

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.

The *assumptions* of the local linear trend model (3.46) are that the observation, level, and slope disturbances ε_t, ξ_t, and ς_t are all mutually independent, and normally distributed with zero means, and variances equal to σ_ε^2, σ_ξ^2, and σ_ς^2, respectively.

In the remaining part of this section we will first discuss and illustrate the effect of fixing all state disturbances ξ_t and ς_t in (3.46) 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.

Fixing all state disturbances ξ_t and ς_t in (3.46) 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

$$y_t = \mu_1 + \nu_1(t-1) + \varepsilon_t, \qquad \varepsilon_t \sim NID(0,\sigma_\varepsilon^2) \qquad (3.47)$$

for t = 1, …, n, where the independent or predictor variable (t –1) = 0, 1, ..., 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 (3.47) to the log of the logarithm of the annual number of road traffic fatalities in Finland for the period 1970 through 2003, we find that $\mu_1 = 6.8717$, $\nu_1 = -0.028733$, 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, …, *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.

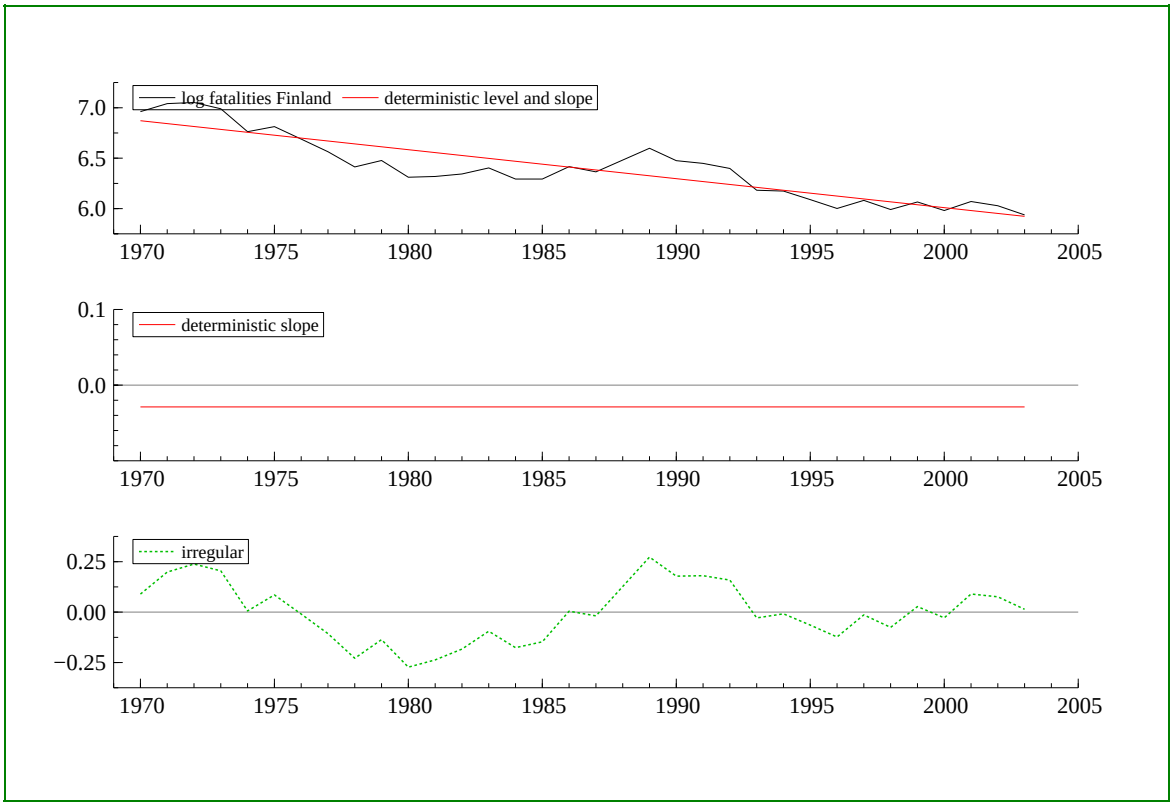


Figure 3.22: Deterministic trend (top), deterministic slope (middle), and irregular component for the log of the number of Finnish fatalities.

The best fitting regression line obtained with the deterministic linear trend model is shown at the top of Figure 3.22, while the bottom of Figure 3.22 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 3.18: 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 3.18. 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

$$\text{AIC} = \frac{1}{34}\left[-2(34)(0.3036367)+2(2+1)\right] = \text{-}0.430803.$$

Allowing both the level and the slope to vary over time in model (3.46), 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

$$\text{AIC} = \frac{1}{34}\left[-2(34)(0.7864746)+2(2+3)\right] = \text{-}1.27883.$$

The maximum likelihood estimates of the variances corresponding to the irregular, level, and slope components are $\sigma_\varepsilon^2 = 0.00320083$, $\sigma_\xi^2 = 9.69606\text{E}^{-}26$, and $\sigma_\zeta^2 = 0.00153314$, 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 $\sigma_\varepsilon^2 = 0.00320083$, and $\sigma_\zeta^2 = 0.00153314$, respectively. The maximum likelihood estimates of the values of the level and the slope right at the start of the series are $\mu_1 = 7.0133$ and $\nu_1 = 0.0068482$.

The trend (consisting of a deterministic level and a stochastic slope) of this analysis is displayed at the top of Figure 3.23, while the stochastic slope is shown separately in the middle of the figure. Since the time varying slope component in Figure 3.23 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 3.23). 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.

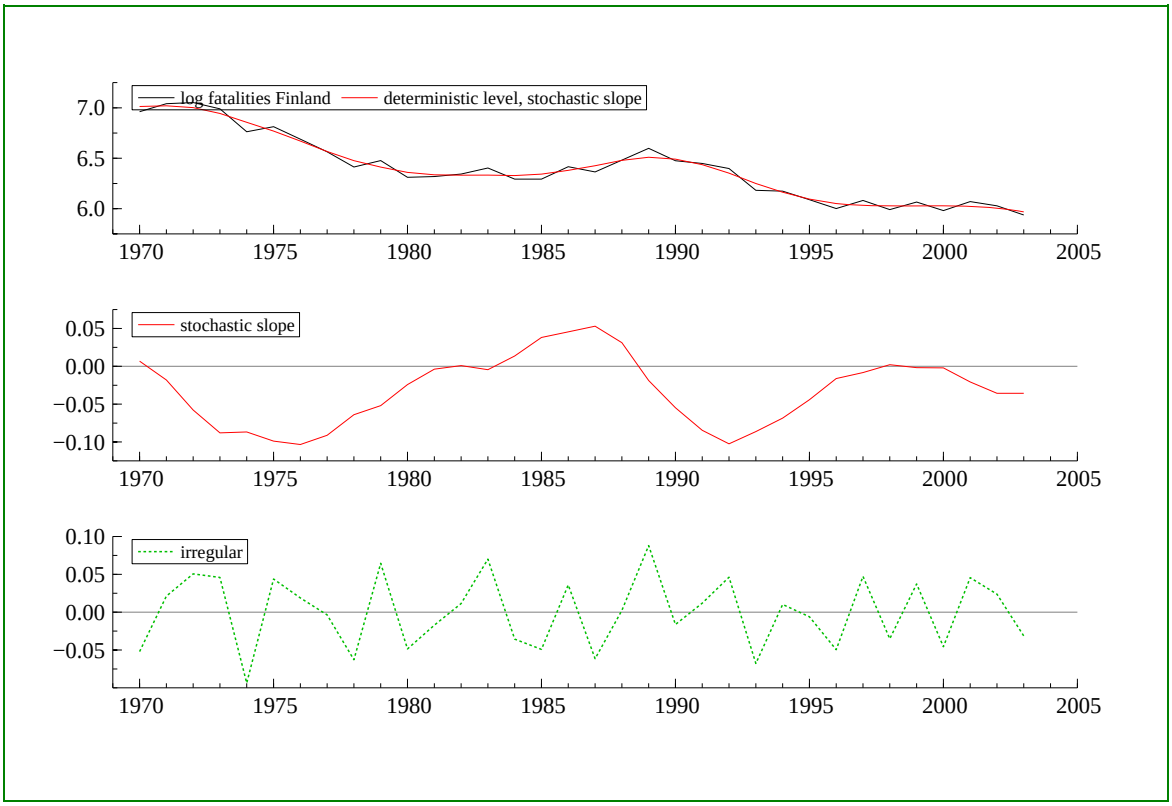


Figure 3.23: Trend of deterministic level and stochastic slope model for the log of Finnish fatalities (top), stochastic slope component (middle), and irregular component (bottom).

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 3.23, and the diagnostic tests for the residuals of the analysis are given in Table 3.19. 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 3.19: 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

$$\text{AIC} = \frac{1}{34}\left[-2(34)(0.7864746)+2(2+2)\right] = -1.33766.$$

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.

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. The 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 3.24. In the figure, vertical lines have been added through each year in the observed time series.

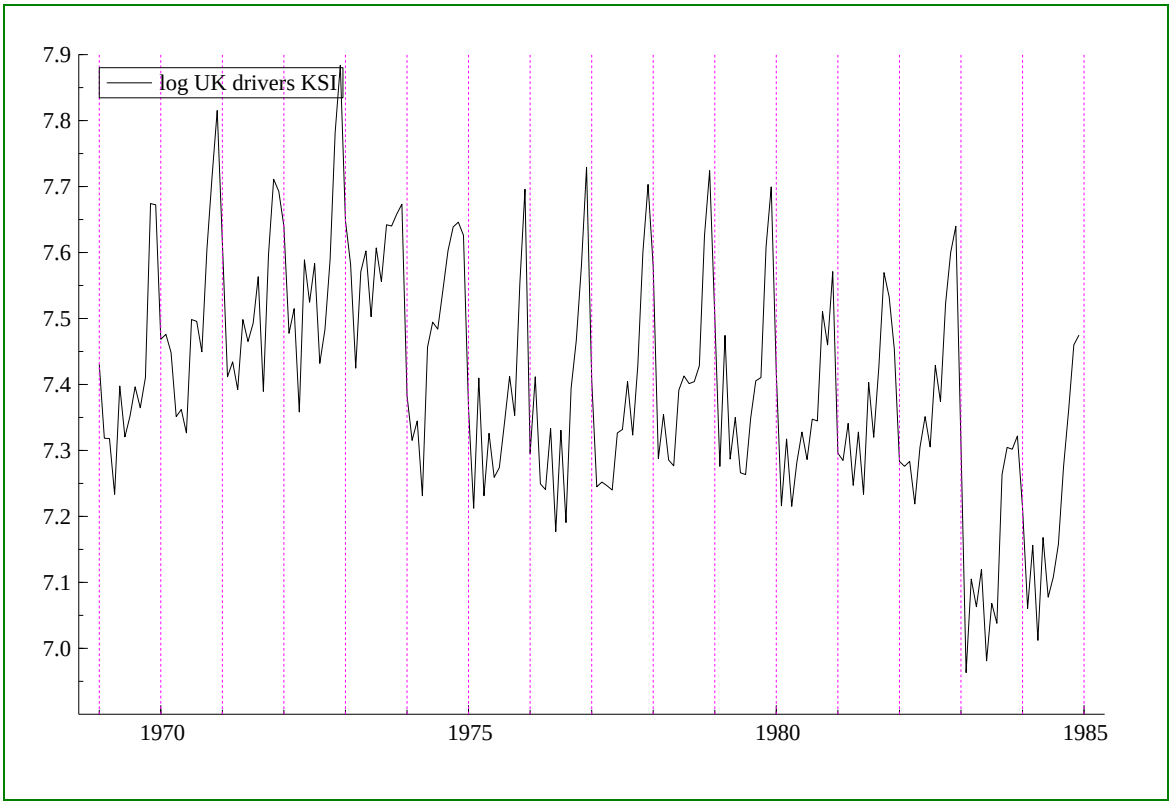


Figure 3.24 Log of monthly number of UK drivers KSI with time lines for years.

Inspecting the monthly development for each year in Figure 3.24, 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.

The *research problem* addressed in this section is how to obtain an appropriate description of an observed time series when it contains a seasonal pattern. The *dataset* in such an analysis still consists of only one variable: a time series y_t consisting of repeated measurements of one and the same phenomenon at time points $t = 1, \ldots, n$.

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:

$$y_t = \mu_t + \gamma_{1,t} + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$$

$$
\begin{aligned}
&\mu_{t+1} = \mu_t + \nu_t + \xi_t, && \xi_t \sim NID(0, \sigma_\xi^2) \\
&\nu_{t+1} = \nu_t + \zeta_t, && \zeta_t \sim NID(0, \sigma_\zeta^2) \\
&\gamma_{1,t+1} = -\gamma_{1,t} - \gamma_{2,t} - \gamma_{3,t} + \omega_t, && \omega_t \sim NID(0, \sigma_\omega^2) \\
&\gamma_{2,t+1} = \gamma_{1,t}, \\
&\gamma_{3,t+1} = \gamma_{2,t},
\end{aligned}
\tag{3.48}
$$

for t = 1, ..., n, where $\gamma_{1,t}$ denotes the seasonal component. The disturbances ω_t in (3.48) 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 (3.48). Irrespective of its periodicity, the seasonal always satisfies

$$
\sum_{j=1}^{s} \gamma_{1,j} = 0, \tag{3.49}
$$

thus ensuring that the seasonal is not confounded with the other components of the model. The type of seasonal that is modelled in (3.48) 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 we refer to Durbin and Koopman (2001), as these specifications are beyond the scope of the present report.

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.

The *assumptions* of the local linear trend and seasonal model (3.48) are that the observation, level, slope, and seasonal disturbances ε_t, ξ_t, ζ_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_ζ^2, and σ_ω^2, respectively.

As before, in the remaining part of this section we will first discuss and illustrate the effect of fixing all state disturbances ξ_t, ζ_t, and ω_t in (3.48) on zero, and then present the effect of letting the level, slope, and seasonal components vary over time.

When the state disturbances ξ_t, ς_t, and ω_t in (3.48) are all fixed on zero, the model reduces to the following deterministic model:

$$y_t = \mu_1 + \nu_1(t-1) - \sum_{i=1}^{s-1} \gamma_{i,t-1} + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2). \tag{3.50}$$

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 μ_1 and ν_1 are 7.5540 and -0.00155, respectively. Thus for these data we obtain

$$y_t = 7.5540 - 0.00155(t-1) - \sum_{i=1}^{s-1} \gamma_{i,t-1} + \varepsilon_t,$$

which can also be written as

$$y_t = 7.5556 - 0.00155t - \sum_{i=1}^{s-1} \gamma_{i,t-1} + \varepsilon_t.$$

We do not mention the estimates for the eleven initial values of the dummy seasonal because these are not very informative in the present context.

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 3.25, 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 3.20 of the irregular component in Figure 3.25 indicate that the residuals of this completely deterministic model neither satisfy the assumption of independence nor that of normality.

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

$$\text{AIC} = \frac{1}{192}\left[-2(192)(0.69830186) + 2(13+1)\right] = -1.25077.$$

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.

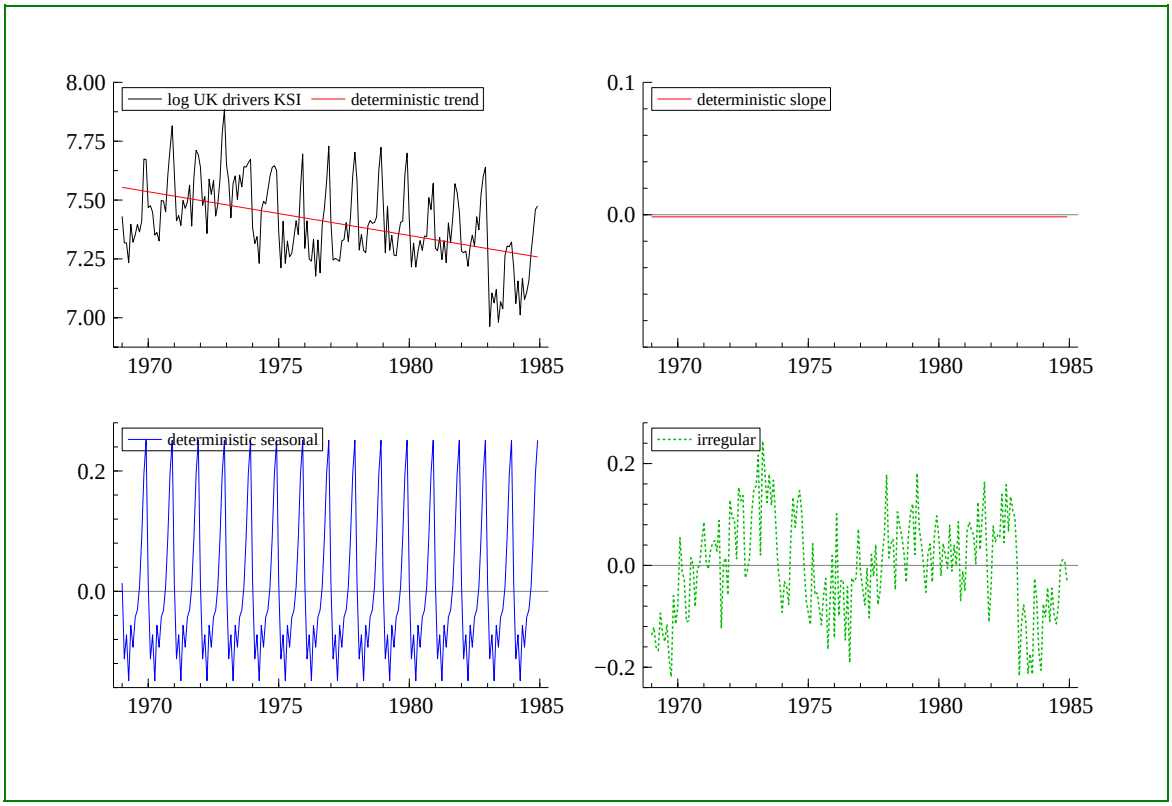


Figure 3.25: Deterministic trend (top left), deterministic slope (top right), deterministic seasonal (bottom left), and irregular component (bottom right) of deterministic trend and seasonal model for log UK drivers KSI.

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 3.25: 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 3.25. 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 3.25.

	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 3.20: Diagnostic tests for deterministic trend and seasonal model for log UK drivers KSI.

Allowing the level, slope and seasonal components in (3.48) 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 $\sigma_\varepsilon^2 = 0.00346783$, $\sigma_\xi^2 = 0.00100094$, $\sigma_\zeta^2 = 6.74681\text{E}^{-}52$, and $\sigma_\omega^2 = 7.28648\text{E}^{-}025$. 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

$$\text{AIC} = \frac{1}{192}\left[-2(192)(0.95650011) + 2(13+4)\right] = -1.73592,$$

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 (3.48) 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 $\sigma_\varepsilon^2 = 0.00346757$ and $\sigma_\xi^2 = 0.0010011$. The values of μ_1 and ν_1 are now 7.4133 and -0.00090531, respectively. For this model, the Akaike information criterion equals

$$\text{AIC} = \frac{1}{192}\left[-2(192)(0.95650011) + 2(13+2)\right] = -1.75675,$$

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, we may 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 $\sigma_\varepsilon^2 = 0.00351385$, and that of the level component equals $\sigma_\xi^2 = 0.000945723$. This implies that the Akaike information criterion now equals

$$\text{AIC} = \frac{1}{192}\left[-2(192)(0.98299654) + 2(12+2)\right] = -1.82016.$$

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 we keep it as the best model for describing the log of the UK drivers KSI series.

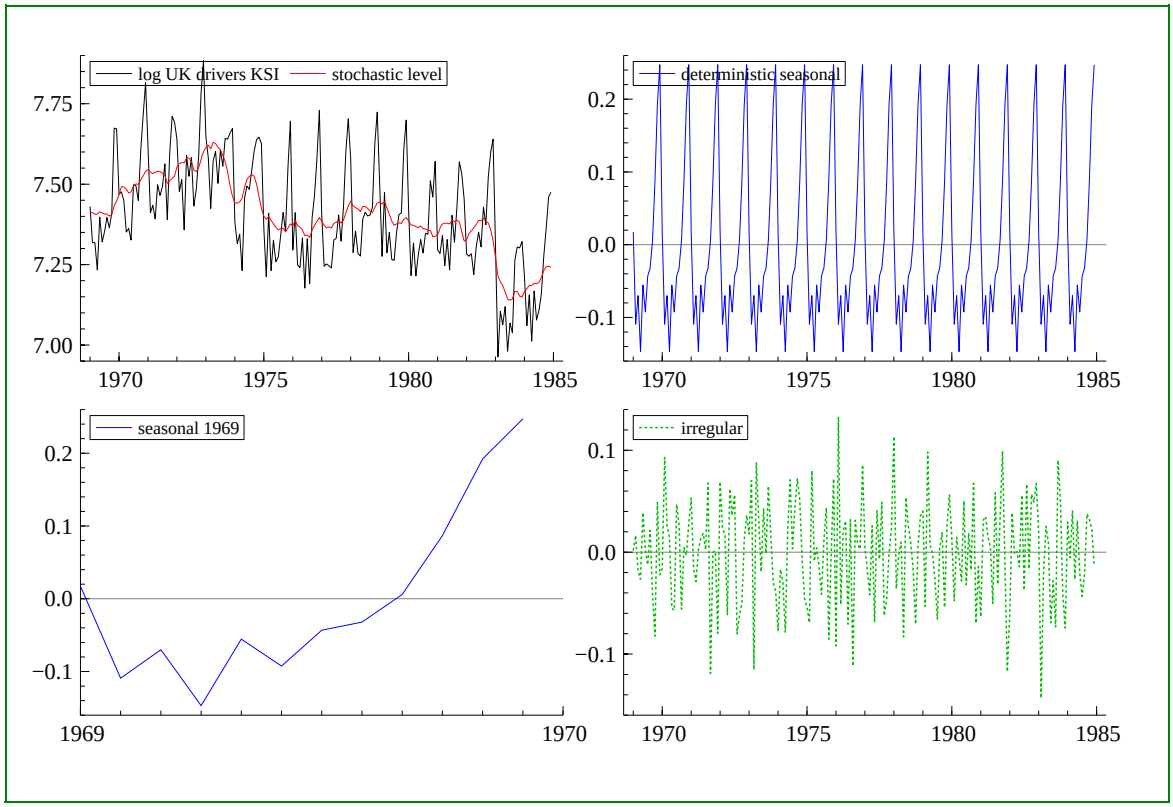


Figure 3.26: Stochastic level (top left), deterministic seasonal (top right), the seasonal for 1969 (bottom left), and irregular component (bottom right) for stochastic level and deterministic seasonal analysis of log of UK drivers KSI.

The three components of the latter analysis are all displayed in Figure 3.26. 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 3.21 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.

	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 3.21: Diagnostic tests for stochastic level and deterministic dummy seasonal analysis of log of UK drivers KSI.

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.

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 3.27, while the standardised observation disturbances are shown at the bottom of the same figure.

Each of the auxiliary residuals at the top of Figure 3.27 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 ≈ 10 that we would expect purely based on chance for this series. Still, the value of the residual for January 1983 particularly stands out as being very extreme.

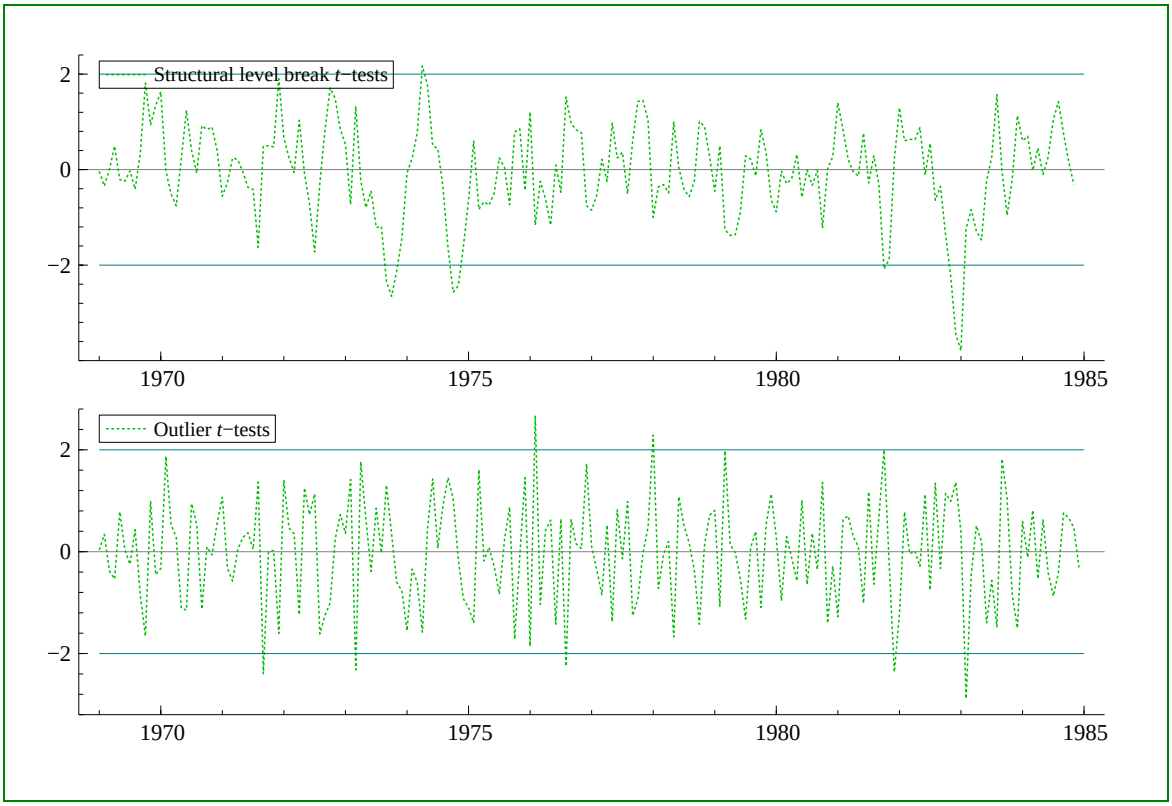


Figure 3.27: Auxiliary residuals for the stochastic level and deterministic seasonal model applied to the log of the UK drivers KSI series.

Similarly, each of the auxiliary residuals at the bottom of Figure 3.27 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 we would expect according to chance. Since, moreover, none of these are very extreme we conclude 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.

The effect of the introduction of this seat belt law can be investigated by adding a so-called *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 3.27 suggest a break in the level of the log of the UK drivers KSI, we will add a level shift intervention variable to the level and seasonal model discussed in the previous section.

The *research problem* addressed in this section is how to assess the effect of an intervention variable on a time series. The *dataset* in such an analysis now contains two variables: a dependent variable y_t which is a time series as before, and an independent intervention variable which we will denote by w_t.

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:

$$
\begin{aligned}
&y_t = \mu_t + \gamma_{1,t} + \lambda_t w_t + \varepsilon_t, && \varepsilon_t \sim NID(0,\sigma_\varepsilon^2)\\
&\mu_{t+1} = \mu_t + \xi_t, && \xi_t \sim NID(0,\sigma_\xi^2)\\
&\gamma_{1,t+1} = -\gamma_{1,t} - \gamma_{2,t} - \gamma_{3,t} + \omega_t, && \omega_t \sim NID(0,\sigma_\omega^2)\\
&\gamma_{2,t+1} = \gamma_{1,t},\\
&\gamma_{3,t+1} = \gamma_{2,t},\\
&\lambda_{t+1} = \lambda_t + \rho_t, && \rho_t \sim NID(0,\sigma_\rho^2)
\end{aligned}
\qquad (3.51)
$$

for t = 1, …, 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, we present model (3.51) as if we are 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 (3.51) stochastically, in practice this is never done when dealing with intervention variables.

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.

The *assumptions* of the local level and seasonal model (3.51) 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.

In the remaining part of this section we 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.

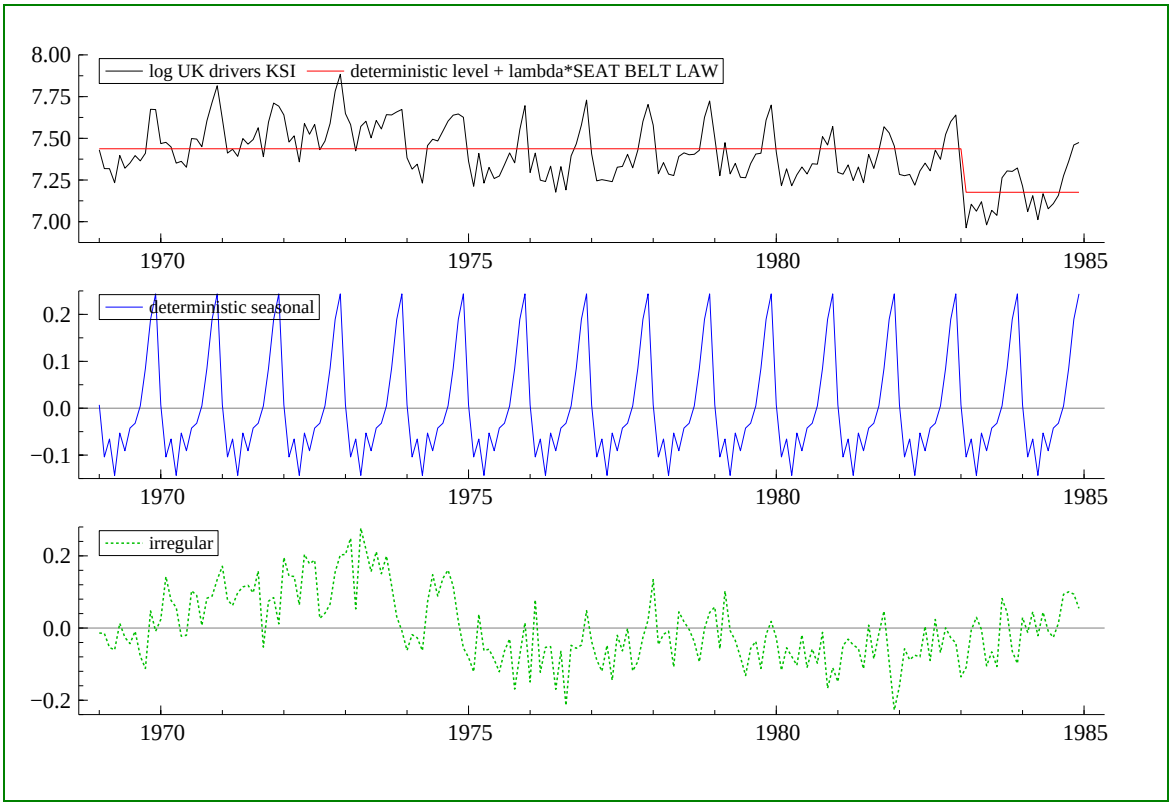


Figure 3.28: Deterministic level plus intervention variable (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series .

Treating all the state components in (3.51) deterministically, it is not very difficult to prove that the model simplifies into the following classical regression model:

$$y_t = \mu_1 - \sum_{i=1}^{s-1} \gamma_{i,t-1} + \lambda_1 w_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2). \tag{3.52}$$

Estimating model (3.52) by fixing all the state disturbances in (3.51) 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 $\sigma_\varepsilon^2 = 0.0100188$. 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 3.28. 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:

$$e^{\lambda_1} - 1 = e^{-0.26075} - 1 = -0.2295,$$

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.

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 3.22: 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 $\lambda_1 = -0.26075$ deviates from zero yields

$$t = \frac{-0.2607515908}{0.02227747268} = -11.70472049, \qquad (3.53)$$

which is very significant. In order to investigate whether this test is reliable, we must also check whether the model assumptions of independence, homoscedasticity and normality of the residuals are satisfied. However, as Table 3.22 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 we allow the level component in model (3.51) 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 $\sigma_\varepsilon^2 = 0.00378397$ and $\sigma_\xi^2 = 0.000473516$, respectively.

The estimated effect of the seat belt law re-expressed in the percentage change in the absolute numbers of drivers KSI is now

$$e^{\lambda 1} - 1 = e^{-0.23981} - 1 = -0.2132,$$

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 $\lambda_1 = -0.23981$, yielding

$$t = \frac{-0.239806756}{0.05307021883} = -4.5187\,. \qquad (3.54)$$

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 (2.65) in the previous completely deterministic model.

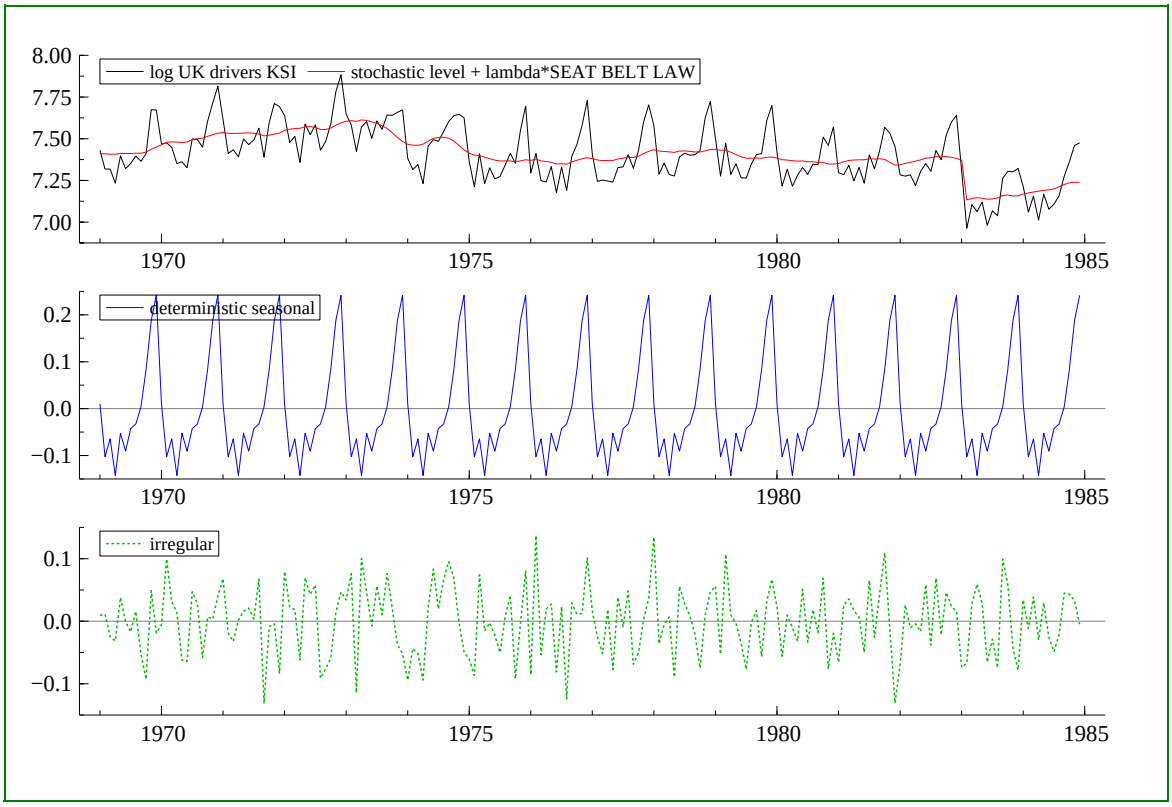


Figure 3.29: Stochastic level plus intervention variable (top)

The stochastic level plus intervention variable is shown in Figure 3.29, together with the deterministic dummy seasonal, and the irregular component. The diagnostic tests for the model assumptions are given in Table 3.23. Since all three assumptions are satisfied in the present analysis, we can now rest 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 3.23: Diagnostic tests for stochastic level and dummy seasonal analysis of log of UK drivers KSI, including intervention variable.

In Figure 3.30, we have again plotted 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 3.31) has now completely disappeared. This is the effect of adding the intervention variable to the model.

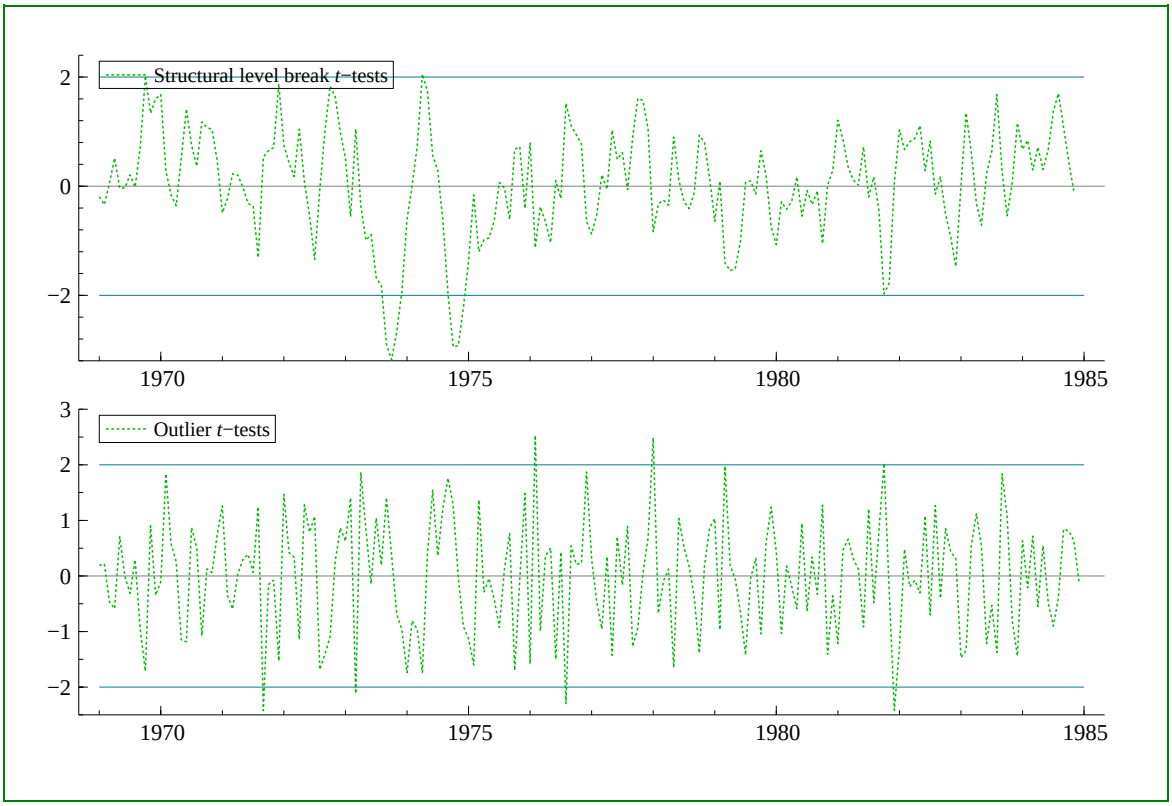


Figure 3.30: Auxiliary residuals for the stochastic level and deterministic seasonal model applied to the log of the UK drivers KSI series, including a level shift intervention variable for the introduction of the seat belt law.

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, we see 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.

3.6.5. Explanatory variables

Apart from binary intervention variables, it is also possible to investigate the effects of continuous explanatory variables on the development of a time series, which is the *research problem* addressed in the present section. 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:

$$y_t = \mu_t + \gamma_{1,t} + \lambda_t w_t + \sum_{j=1}^{k} \beta_{jt} x_{jt} + \varepsilon_t , \tag{3.55}$$

where the x_j are k continuous explanatory variables (j = 1, …, k), and the β_j are unknown regression weights or coefficients. The *dataset* in such an analysis 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.

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. We also keep the same intervention variable in the model that was used in the previous section: 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:

$$\begin{aligned}
&y_t = \mu_t + \gamma_{1,t} + \lambda_t w_t + \beta_t x_t + \varepsilon_t , && \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \\
&\mu_{t+1} = \mu_t + \xi_t , && \xi_t \sim NID(0, \sigma_\xi^2) \\
&\gamma_{1,t+1} = -\gamma_{1,t} - \gamma_{2,t} - \gamma_{3,t} + \omega_t , && \omega_t \sim NID(0, \sigma_\omega^2) \\
&\gamma_{2,t+1} = \gamma_{1,t} , && \\
&\gamma_{3,t+1} = \gamma_{2,t} , && \\
&\lambda_{t+1} = \lambda_t + \rho_t , && \rho_t \sim NID(0, \sigma_\rho^2)
\end{aligned} \tag{3.56}$$

$$\beta_{t+1} = \beta_t + \tau_t, \qquad \tau_t \sim NID(0, \sigma_\tau^2)$$

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 (3.56) is presented as if we are 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 (3.56), thus allowing the regression coefficient to vary over time. Here, however, we will only consider deterministic regression components.

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 on the development of an observed time series containing a seasonal pattern.

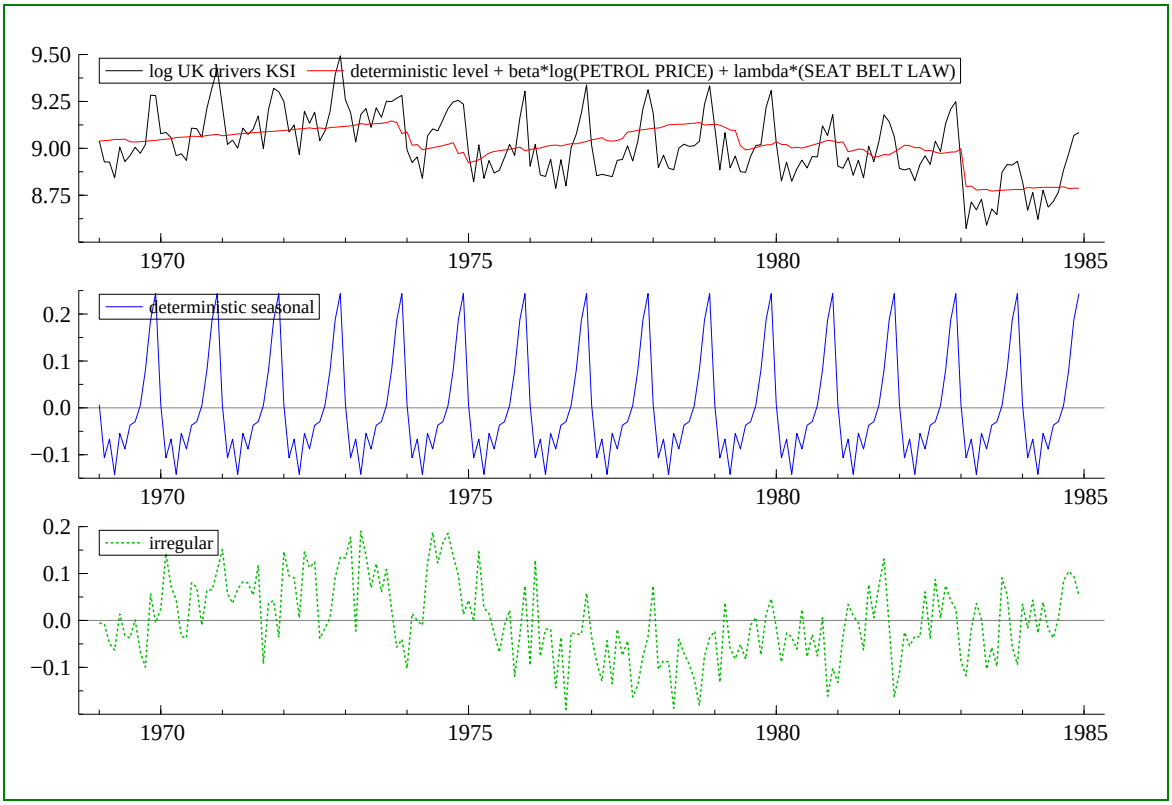


Figure 3.31: Deterministic level plus intervention and explanatory variable (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series .

The *assumptions* of model (3.56) 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.

In the remaining part of this section we will first discuss and illustrate the effect of fixing all state disturbances ξ_t, ω_t, ρ_t, and τ_t in (3.56) on zero, and then present the effect of letting the level component vary over time.

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 $\sigma_\varepsilon^2 = 0.00740223$.

The model therefore reduces to a classical regression model with regression equation

$$y_t = 6.4016 - \sum_{i=1}^{s-1} \gamma_{i,t-1} - 0.19714 w_t - 0.45213 x_t + \varepsilon_t .$$

The plot of the deterministic level plus intervention and explanatory variables is shown in Figure 3.31, 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 β_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

$$\text{AIC} = \frac{1}{192}\left[-2(192)(0.84903819) + 2(14+1)\right] = \text{-}1.54183 ,$$

which is a clear improvement upon the completely deterministic model without 'log petrol price'.

The standard *t*-test for establishing whether the regression coefficient $\lambda_1 = \text{-}0.19714$ for the intervention variable deviates from zero yields

$$t = \frac{-0.197139471 6}{0.02072756003} = -9.510983022,$$

which is very significant. The standard *t*-test for establishing whether the regression coefficient $\beta_1 = -0.45213$ 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 3.24: 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 we must also check whether the residuals satisfy the model assumptions. As Table 3.24 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 3.32.

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.

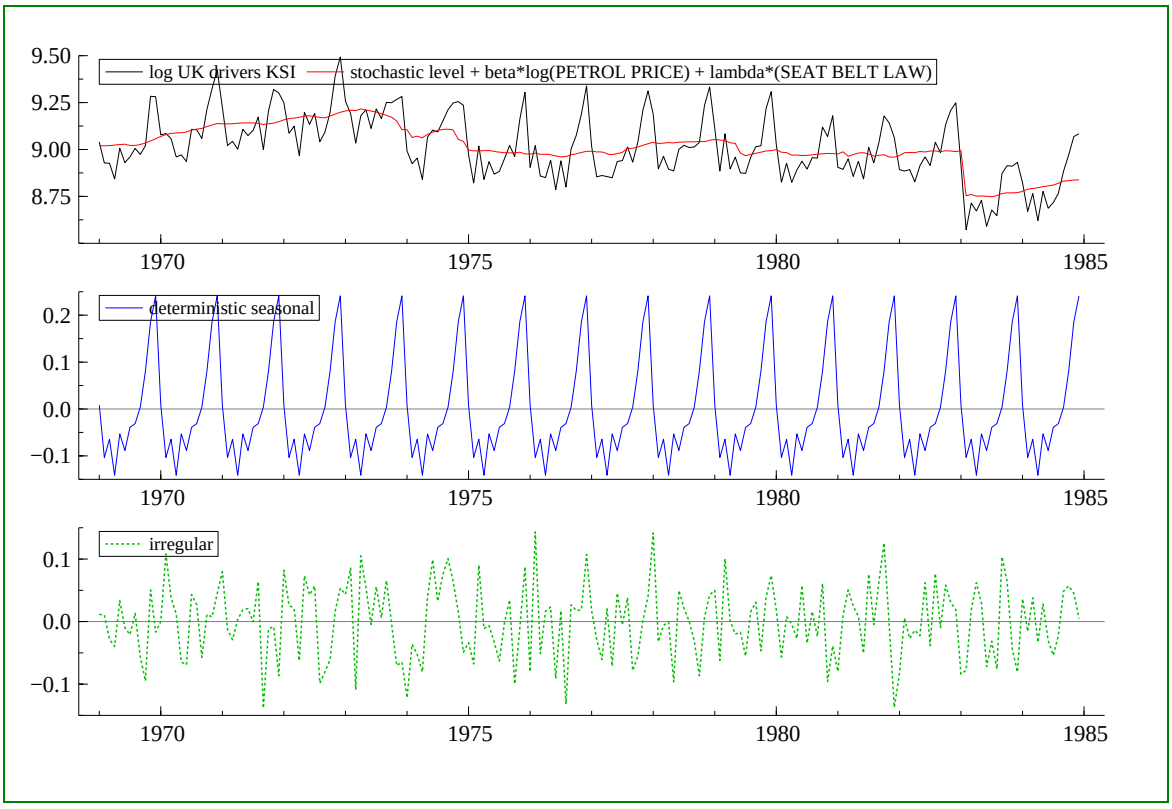


Figure 3.32: Stochastic level plus intervention and explanatory variables (top), deterministic seasonal (middle), and irregular component (bottom) for the log of the UK drivers KSI series .

The value of the Akaike information criterion for this model equals

$$\text{AIC} = \frac{1}{192}\left[-2(192)(1.0265254)+2(14+2)\right] = \text{-}1.88638\,,$$

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 $\lambda_1 = \text{-}0.23759$ 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 $\beta_1 = \text{-}0.45213$ 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 3.25: 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 3.25 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. Comparing the *t*-tests with a fixed intercept with those with a time-varying intercept, we see 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. We finally note 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).

Until now we have focused on the descriptive and explanatory aspects of state space methods. In the next section we 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, we first need to mention that the state components of state space models can be estimated in a number of ways. In all the previous sections on the theory of state space methods we have presented that 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.

In Chapter 1 we already presented one example of forecasting with state space methods. As mentioned in Section 1.2.2, the local level model yields the most appropriate description for the log of the annual number of road traffic fatalities in Norway in the period 1970-2003. The local level model was therefore also used to obtain the forecasts for the log of the annual number of road traffic fatalities in Norway in the period 2004-2010 displayed in Figure 1.7. 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.

In this section we will present two more examples of forecasting: one with the local linear trend model, and one with the local level and seasonal model with an explanatory and intervention variable.

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 3.33, 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.

In state space methods, all estimates of the components of the state also have associated so-called 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 Var(μ_t) denote the estimation error variance of the trend μ_t of the local linear trend model, therefore, the 90% confidence limits are computed with the well-known formula

$$\mu_t \pm 1.64\sqrt{\text{Var}(\mu_t)},$$

where +1.64 and –1.64 are the *z*-scores corresponding to the 90% interval around the mean of a normal distribution.

The thus computed 90% interval for the filtered and predicted trend of the smooth trend model is displayed in Figure 3.33. As the figure shows, 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.

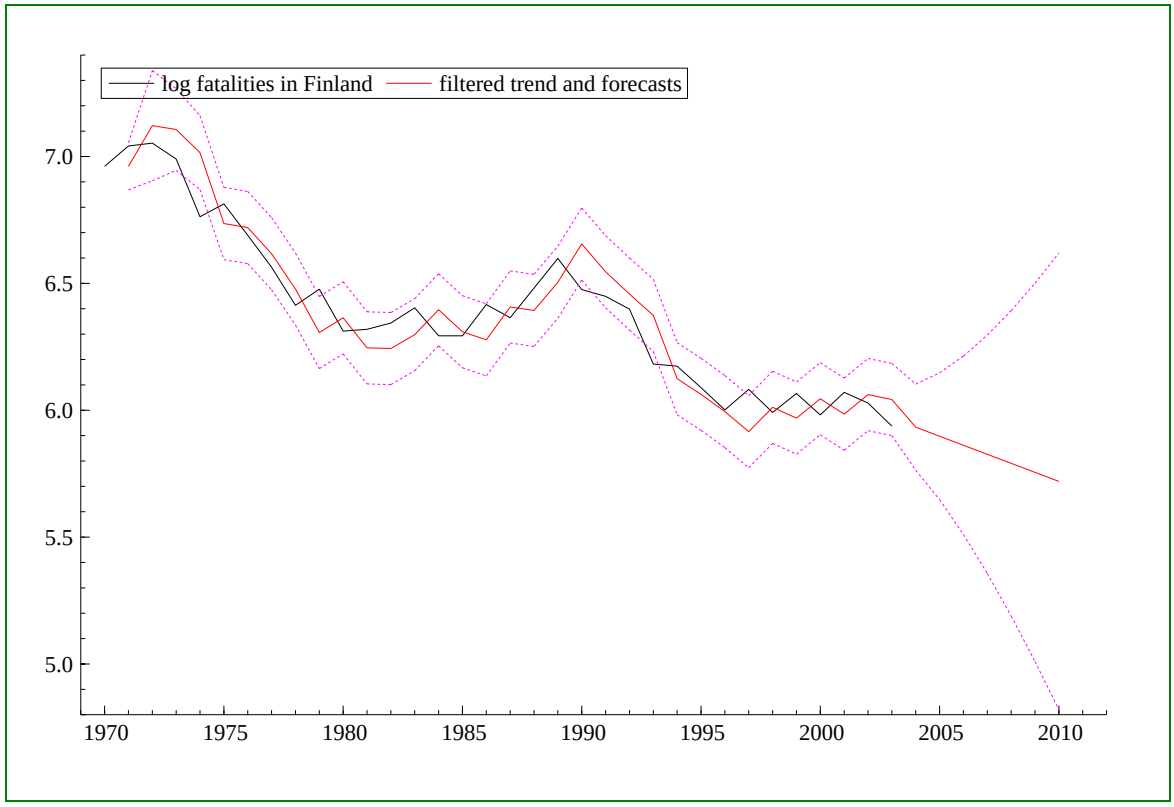


Figure 3.33: Filtered trend, and seven year forecasts for Finnish fatalities, including their 90% confidence limits.

As a last example, we re-analysed the log of the UK drivers KSI series (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, we treated the last six observations in the dependent and independent variables for July through December 1984 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.

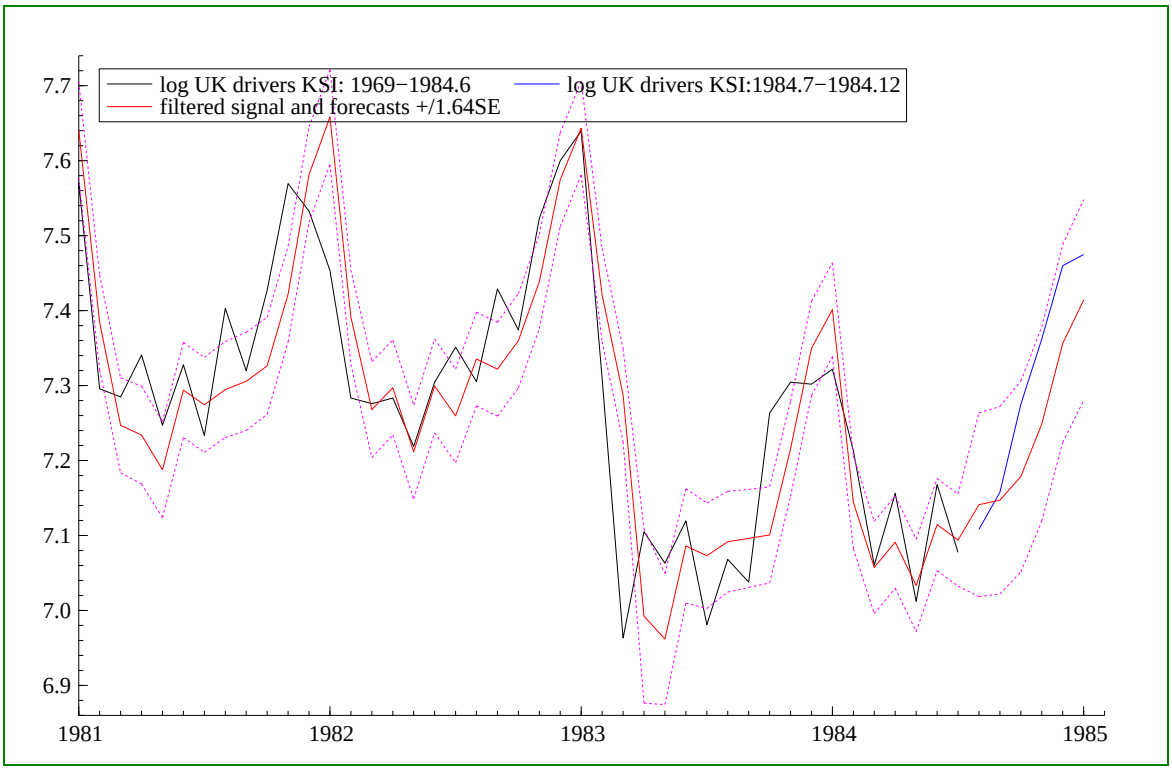


Figure 3.34: Filtered signal

The results are shown in Figure 3.34 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, we see that the 90% confidence limits become larger and larger as the forecasts are located further into the future. We also see that the actual observations fall within the 90% confidence limits of the estimated forecasts, which is a good sign.

We end by noting 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.

Practice: The Manual

3.7. Multilevel models

3.8. Time series models

Acknowledgement

We would like to thank the following people:

References

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.

Belsley, D., Kuh, E., and Welsch, R. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley and Sons, New York.

Box G.E.P., Cox D.R. "An analysis of transformations*". Journal of the Royal Statistical Society*, B(2):211-243, 1964

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.

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

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.

Cochran, W. G. (1963). *Sampling Techniques*, second edition, New York: John Wiley & Sons.

COST 329. *Models for traffic and safety development and interventions*. European Commission, Directorate general for Transport, Brussels , final report of the Action, 2004.

Davis, R., Dunsmuir, W., & Wang, Y. (2000). On autocorrelation in a Poisson regression model. *Biometrika*, Vol. 87, No. 3, pp. 491-505.

Dean, C, "Testing for overdispersion in Poisson and binomial regression models", J. Amer. Statist. Assoc. 87, 451-457 (1992).

Dean, C, Lawless, J.F, "Tests for detecting overdispersion in Poisson regression models", J. Amer. Statist. Asoc. 84, 467-472 (1989).

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.

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 neighbourhood effects?" Social Science and Medicine, 48, p.497-506.

Durbin, J. & Koopman, S.J. (2001). *Time series analysis by state space methods.* Oxford: Oxford University Press.

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.

Gaudry M. *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, 1984.

Gaudry M. *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, 2002 (revision of Gaudry 1984).

Gaudry M., Lassarre S. *Structural Road Accident Models - The International DRAG Family*, Pergamon, *2000*

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.

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.

Gourieroux Christian et Monfort Alain : *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. *Forecasting structural time series and the Kalman filter*. Cambridge University Press, Cambridge, 1989

Harvey A.C., Durbin J., **"**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, 1986

Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter.* Cambridge: Cambridge University Press.

Hauer, E., Ng, J. C. N., & Lovell, J. (1988). *"Estimation of Safety at Signalized Intersections"*, Transportation Research Record, 1185, Transportation Research Board, 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.

Hox J. (2002). *Multilevel Analysis. Techniques and Applications*, Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

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.

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. (2002). *Introducing multilevel modeling*, second edition, London: Sage publications.

LABS, INRETS, INVS. *Presidential amnesty and road safety*. Report of expertise, 2001.

Langford, I, Bentham, G, McDonald, 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 (1998).

Langford, I.H., Leyland, A.H., Rashbach, j., Goldstein, H., (1999), "Multilevel modelling of the geographical distribution of diseases", Applied Statistics 48 part 2, pp 253-268

Lassarre S., **"**Analysis of progress in road safety in ten European countries**"**. *Accident Analysis & Prevention*, Vol. 33 pp.743-751, *2001*

Lassarre S., *Cadrage méthodologique d'une modélisation pour un suivi de l'insécurité routière,* Synthèse INRETS, n°26, Arcueil, 1994.

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.

Longford, N. (1993) *Random coefficient models*, Oxford: Clarendon Press.

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.

McMillan, N. J., Berliner, M. J. (1994). *A spatially correlated hierarchical random effect model for Ohio corn yield*, Technical report 10, National Institute for Statistical Sciences, Research Triangle Park, NC.

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.

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.

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.

Oppe S., **"***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, *1993*

Ostrom, C.W. (1990). *Time series regression techniques.* Second edition. London: Sage Publications.

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).

R. Bergel, A. Depire : *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.

R. Bergel, A. Depire. *Climate, road traffic and road risk - an aggregate approach*. Proceedings of the WCTR'04, Istanbul, 4 - 8 July 2004.

R. Bergel, B. Girard : The RES Model by road type in France . In : Gaudry M., Lassarre S. *Structural Road Accident Models - The International DRAG Family*, Pergamon, *2000*

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.

Schwarz, G. (1978). Estimating the Dimension of a Model. Annals of Statistics. 6 461–464.

Scott P. P., **"**Modelling Time-Series of British Road Accident Data**"**, *Accident Analysis & Prevention*, Vol. 18 n°2 pp.109-117, *1986*

Shieh, Y, Fouladi, R, "The Effect of Multicollinearity on Multilevel Modeling Parameter Estimates and Standard Errors", Educational and Psychological Measurement, Vol. 63, No. 6, 951-985 (2003).

Smeed J.R., "Some statistical aspects of road safety research". Journal of the Royal Statistical Society, A1, 1-34, 1949.

Snijders, T, & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*, London: Sage Publications.

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.

van Belle, G. (2002). *Statistical rules of thumb.* New York: John Wiley and Sons.

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. (2005). 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.

Washington, S. P., Karlaftis, M. G., and F. L. Mannering (2003). Statistical and Econometric Models for Transportation Data Analysis. Chapman & Hall/CRC (2003).

Yang, M., Goldstein, H., Browne, W., Woodhouse, G., (2001) "Multivariate multilevel analyses of examination results", J. Royal Statistical Society, A.

Zeger, S. (1988). A Regression Model for Time Series of Counts. *Biometrika*, Vol. 75, No. 4, pp. 621-629.