



Deliverable D7.5

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

Dupont, E. and Martensen, H. (Eds.) (2007) Multilevel modelling and time series analysis in traffic research – Manual. Deliverable D7.5 of the EU FP6 project SafetyNet.

Contract No: TREN-04-FP6TR-S12.395465/506723

Acronym: SafetyNet

Title: Building the European Road Safety Observatory

Integrated Project, Thematic Priority 6.2 “Sustainable Surface Transport”

Project Co-ordinator:

Professor Pete Thomas

Vehicle Safety Research Centre

Ergonomics and Safety Research Institute

Loughborough University

Holywell Building

Loughborough

LE11 3UZ

Organisation name of lead contractor for this deliverable:

Belgian Road Safety Institute (IBSR)

Due Date of Deliverable: 28/02/2007

Submission Date:

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

Contributing Authors: A. Angermann (KfV), C. Antoniou (NTUA) R. Bergel (INRETS), E. Berends (SWOV), F. Bijleveld (SWOV), C. Brandstätter (KfV), M. Cherfi (INRETS), C. de Blois (SWOV), E. Dupont (IBSR), M. Gatscha (KfV), H. Martensen (IBSR), E. Papadimitriou (NTUA), G. Yannis (NTUA)

Project Start Date: 1st May 2004

Duration: 4 years

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



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

Table of Contents

EXECUTIVE SUMMARY.....	4
CHAPTER 1 - INTRODUCTION	5
CHAPTER 2 - MULTILEVEL MODELLING	8
2.1 Introduction.....	8
2.2 Multilevel linear regression models.....	11
2.2.1 Basic two level random intercept and random slope models.....	11
2.2.2 Three level models and more	26
2.3 Discrete response models	32
2.3.1 Introduction.....	32
2.3.2 Binary and general binomial responses	32
2.3.3 Multinomial responses.....	40
2.3.4 Counts.....	52
2.4 Longitudinal data.....	72
2.5 Multivariate models	85
2.6 Structural equations models	101
2.7 More complex data structures	101
2.8 Bayesian estimation in multivlevel modelling	101
CHAPTER 3 - TIME SERIES ANALYSIS	102
3.1 Introduction to time series models	102
3.2 Classical linear and non-linear regression models.....	105
3.2.1 Classical linear regression models	105
3.2.2 Generalized linear models (GLM)	130
3.2.3 Non-linear models	130
3.3 Dedicated time series analysis in road safety research	130
3.4 ARMA-type models.....	131
3.4.1 Introduction.....	131
3.4.2 ARIMA models for stationary series (simulated data).....	134
3.4.3 ARIMA models for non seasonal series (Norwegian Fatalities)	134
3.4.4 ARIMA models for seasonal series (UK-KSI Drivers).....	154
3.4.5 Conclusion ARMA-type models.....	183
3.5 DRAG models.....	184

3.6 State space models.....	185
3.6.1 Introduction.....	185
3.6.2 Local level model.....	187
3.6.3 Local linear trend model.....	216
3.6.4 Local linear trend plus seasonal model	232
3.6.5 Intervention variables	243
3.6.6 Explanatory variables	252
3.6.7 Conclusion state space models.....	265
 CHAPTER 4 - CONCLUSION.....	 266

Executive Summary

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

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

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

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

Deliverable 7.5 contains the manual to support the methodology D7.4, where the theoretical background for these two families of analyses is given. For each technique described in the methodology, this manual presents the instructions to fit the models on the basis of user friendly software, as well as guidelines for interpreting the results. The aim of the manual is to enable the reader to conduct all analyses described in the methodology and this way to get hands on experience in the analysis of road safety data. To enable the reader to track every step presented, the data sets discussed in the various sections are available.

Chapter 1 - Introduction

Heike Martensen and Emmanuelle Dupont (IBSR)

This deliverable has been produced in Workpackage 7 (WP7) of the SafetyNet project. WP7 is set up to deal with statistical and conceptual issues that come into play when analysing complex data structures as they arise in road safety research when combining data from different sources or when considering data that have been collected over a long timespan. One of its main objectives is the development of a best practice for the analysis of data structures that require more than the standard statistical tools.

This best practice consists of D7.4 “Multilevel modelling and time series analysis in traffic research – A methodology” (subsequently, simply “the methodology-report”) and the present deliverable. This document contains the practical instructions to support the methodology, where it has been described how to deal with data that are dependent in space (nested data) or in time (time series data). It has been demonstrated how such dependencies can compromise the applicability of standard methods of statistical inferences, because they lead to an underestimation of the standard error and consequently of the probability to classify a result as significant that is in fact due to chance.

Two families of statistical techniques have been presented to deal with these dependencies. Multilevel Modelling is dedicated to the analysis of data that are structured hierarchically and Time Series analyses are employed to overcome dependency issues in time-related data. The methodology is organized in two main chapters, focussing on multilevel modelling (Chapter 2) and time series analysis (Chapter 3) respectively.

For those sections in the methodology where models dedicated to multilevel analysis or to time series analysis are presented, this manual presents the instructions to fit each model on the basis of user friendly software, as well as guidelines for interpreting the results. The aim of this document is to enable the reader to conduct all analyses described in the methodology and this way to get hands-on experience in the analysis of road safety data. To enable the reader to track every step presented, the data sets discussed in the various sections are available. The data are included as a CD and will be available at the SafetyNet website (www.erso.en/safetynet.htm).

This manual is not a stand-alone document. It is intimately related to the methodology and its sections were written under the assumption that the respective part of the methodology report is known. To allow an easy matching of methodology report and manual sections, the numbering in the manual is the same as that in the methodology report. Some sections in the methodology report, however, do not contain data examples or the models presented employ traditional techniques rather than multilevel or dedicated time series models. For these latter sections there is no counterpart in this manual. As a consequence, some sections in the manual are rather short, their main purpose

being, to allow the numbering to continue in the same way as in the methodology report. In Figures 1.1 and 1.2, the structure of Chapters 2 and 3 is presented. Sections that are represented by a colourless box are present in the methodology report, but there is no corresponding example in this manual.

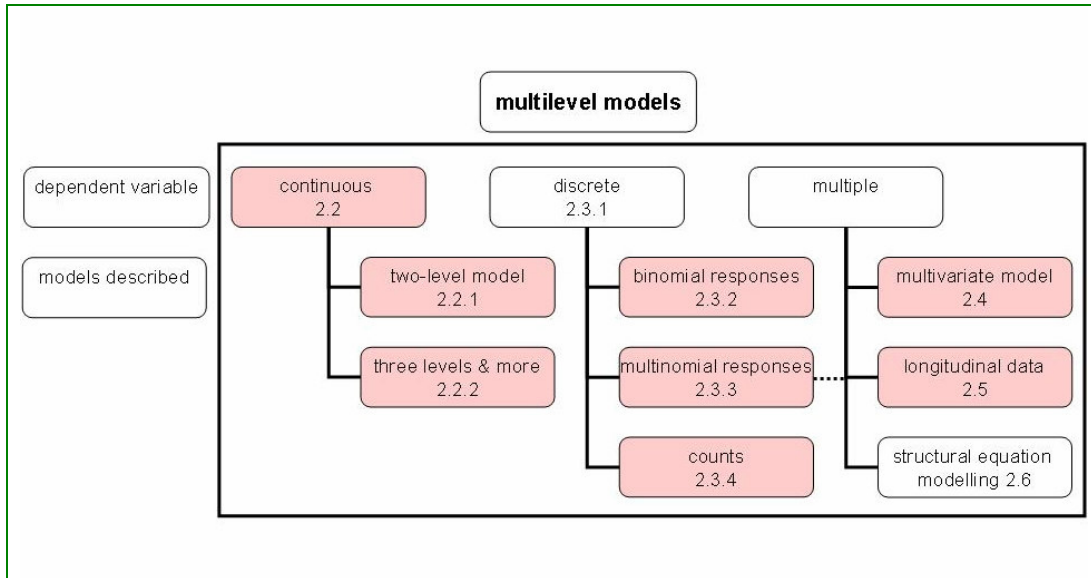


Figure 1.1: Structure of multilevel models presented in Chapter 2. Note: Sections represented in white are present in the methodology report but have no corresponding example in this manual.

Chapter 2 starts with a short description of the principles of multilevel modelling and a software overview in 2.1. Section 2.2 is dedicated to modelling of continuous responses and section 2.3 to the modelling of discrete responses. Section 2.4 presents an example for a multivariate model and section 2.5 for a model for longitudinal data.

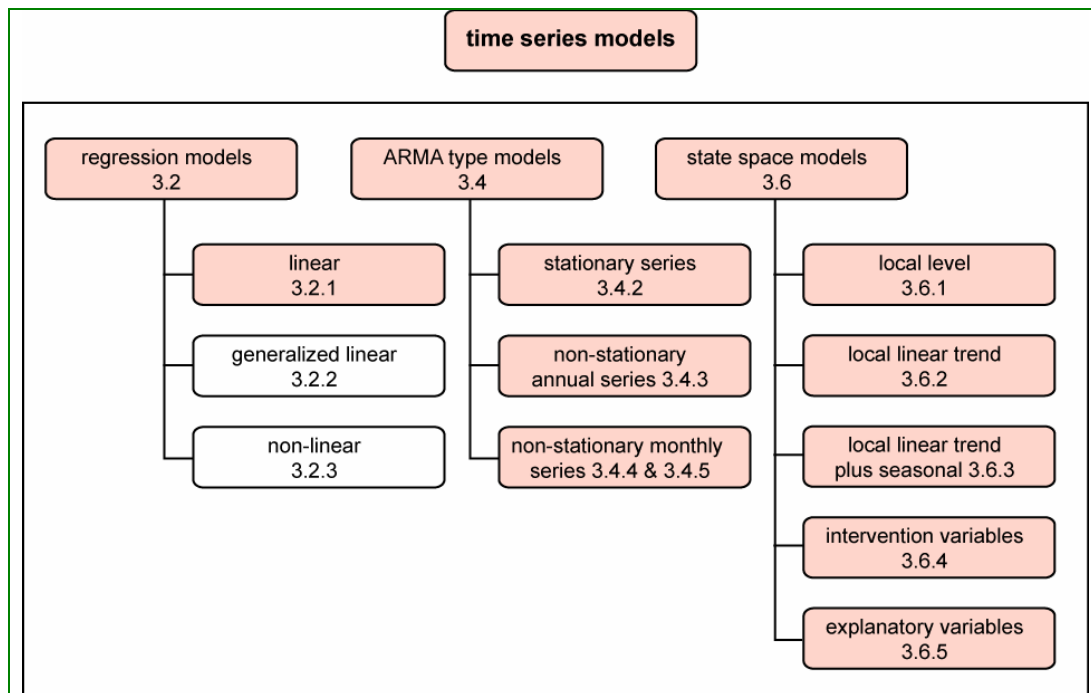


Figure 1.2: Structure of multilevel models presented in Chapter 3. Note: Sections represented in white are present in the methodology report but have no corresponding example in this manual.

Chapter 3 starts with a short introduction to time series analysis and a software overview (3.1). Section 3.2 describes traditional regression analyses models. Traditional regression analyses models were chosen because they are probably the best known type of model, and are often used in the time series context. Special attention is paid to diagnostic tools that serve to detect possible violations of the assumption when dealing with time series data and the possibilities to solve these problems within the traditional framework. In sections 3.4 to 3.6 of the methodology report models dedicated to time series analyses are presented. In the end, these models can be categorised into two classes, one group, including DRAG type models, that can be seen as variants of so-called ARMA-type models and another group of decomposition models that can be regarded as members of state space models. In this manual there are two extensive sections on ARMA-type models (3.5) and on state space models (3.6) respectively. Both contain many empirical examples and detailed instructions for their implementation.

Chapter 4 presents an overview of the methods presented and the examples used.

Chapter 2 - Multilevel Modelling

2.1 Introduction

Heike Martensen and Emmanuelle Dupont (IBSR)

As described in more detail in the methodology report, in traditional regression analyses a dependent variable (y) is predicted by a combination of one or more independent variables (x_1, x_2, \dots), such that y can be modelled by equation 2.1.1.

$$y_i = b_0 + b_1x_1 + b_2x_2 + \dots + e_i \quad (2.1.1)$$

with i being the index of the subjects of study (e.g. accidents, persons, etc.).

As examples, injury severity in an accident can be predicted by the speed of collision or accident frequency can be predicted by the number of alcohol controls and the number of speed infringements. Of course, these predictions are never perfect. Everything that is not predicted is assumed to be due to the randomly distributed error e_i .

One of the most important assumptions upon which the traditional analyses are based is the independence assumption, stating that the residuals, the e_i 's, are independently distributed across all units. Hierarchical structures or nested data often cause the independence assumption to be violated. In hierarchies, the cases within one group are often more similar to each other than the cases in another group. These hierarchical structures have to be represented in the model of analysis, because otherwise the residuals (the variation that cannot be explained by the model) will show the same structure and will therefore not be independently distributed. Examples for such hierarchies are presented in the remainder of the document. To name just a few: In section 2.2 speed data are presented that are collected at a number of randomly selected road sites. The speed of cars at the same road site is jointly influenced by a large number of factors and therefore cars at the same road site are more similar in speed than between different road sites. In sections 2.3.2 and 2.3.3 data on driving under alcohol influence are presented. Again the probabilities of having drunk are more similar for drivers at the same road site as compared to drivers at different road sites. In sections 2.3.4 and 2.4 the fatalities for counties in Greece are presented and it is demonstrated that the numbers of fatalities as well as the effect of certain measures (alcohol and speed controls) vary across regions.

Multilevel modelling offers the possibility to include hierarchical structures into the model of analysis by allowing random variation at each level of the model. Multilevel models also allow the effect of predictor variables to vary across higher level units. In the present chapter, multilevel models are presented for continuous data (section 2.2), dichotomous data (section 2.3.2), and count data (section 2.3.4). Moreover it is demonstrated how multilevel models can be used

to establish a multivariate data structure (section 2.4) that can also be used for multinomial responses (section 2.3.3) and repeated measurement data (section 2.5).

The multilevel models are all implemented with the MLwiN software (Rasbash, Steele, Brown, & Prosser, 2004, www.mlwin.com), dedicated to multilevel modelling. The reason that this software was chosen is its high educational value. In MLwiN (Rasbash et al., 2004), the model formulation is menu-based and can therefore be mastered easily without studying a programming language. The analyses are presented in the form of model equations, allowing a good understanding of the model built. Another advantage of this software is the presence of diagnostic methods tailored to multilevel modelling. Most notably, residuals can be studied at each of the levels included in the model. The program also has excellent plotting functions with an interface that is easy to use, encouraging a thorough inspection of raw data, model predictions, and residuals. The output of the analysis is also presented in the framework of model formulation: the parameters in the model equations are simply replaced by their estimates. This presentation allows maximal understanding of the role of each parameter, and of its possible interpretation.

The downside of this very educational interface is its impracticality: no tables are provided as output, the text in the Equations window cannot be copied; there is no way to export the resulting estimations but to simply type them over. The program is in fact so educational that it forces the user to conduct him/herself many of the calculations necessary for interpretation (variance partition coefficients, test statistics). This policy of not allowing the user to simply take some output without understanding how it came about, can become very tedious once one has passed the initial phase of trying to understand the models and that one simply wants to carry out some routine analyses.

HLM (Bryk, Raudenbush, & Congdon, 1996) is also a special purpose statistical package that will fit many kinds of multilevel models. It has been under active development since the mid 1980s and is now distributed by Scientific Software International (SSI, www.ssicentral.com).

The *MIX* project (Hedeker & Gibbons, 1996 a, b) is a collection of programs for multilevel techniques, including mixed-effects linear regression, mixed-effects logistic regression for nominal or ordinal outcomes, mixed-effects probit regression for ordinal outcomes, mixed-effects Poisson regression, and mixed-effects grouped-time survival analysis. The programs can be downloaded from <http://tigger.uic.edu/~hedeker/mix.html>.

WINBUGS is a software that uses Bayesian estimation algorithms (MCMC, see section 2.7.2 in the Methodology report). Models are represented by a flexible language. Additionally it allows the user to specify their model on a graphical interface.

Multilevel modelling can also be carried out in R (<http://cran.r-project.org>) or its commercial version S-Plus (www.insightful.com/). These programs allow most

of the functions present in MLwiN but without its easily accessible user interface.

The standard statistical software packages allow multilevel modelling to some extent. Most notably SAS, allows the estimation of all models presented in this document (Littell, Milliken, Stroup, and Wolfinger, 1996). However, it does not enable the detailed diagnostics tailored to these models. SPSS only allows the estimation of linear multilevel models. An excellent collection of reviews how to implement multilevel models in a wide variety of statistical software can be found on the MLwiN website (www.mlwin.com/softrev/index.html).

Within the present chapter on multilevel modelling, there is a build-up of information about the use of MLwiN. The chapters on linear models form an introduction to MLwiN and some of its possibilities as well. They contain very detailed information how to address the functions and how to interpret the output. In later chapters this information is more compressed. This all said, the focus of this document is not to learn to deal with MLwiN (for all details the reader is referred to the MLwiN manual by Rasbash, Steel, Brown, & Prosser, 2004) but to get a practical introduction to the multilevel analysis of road safety data.

2.2 Multilevel linear regression models

Heike Martensen and Emmanuelle Dupont (IBSR)

2.2.1 Basic two level random intercept and random slope models

The example data used in this and the following section are based on a national speed survey conducted in Belgium. The speed of 4994 cars was “measured” at 131 randomly selected road sites. Additionally, the length of each car was recorded. The question pursued here, is whether there is a relation between the speed and the length of the car (considered here as rough indicator of its engine power). The data contain the following variables:

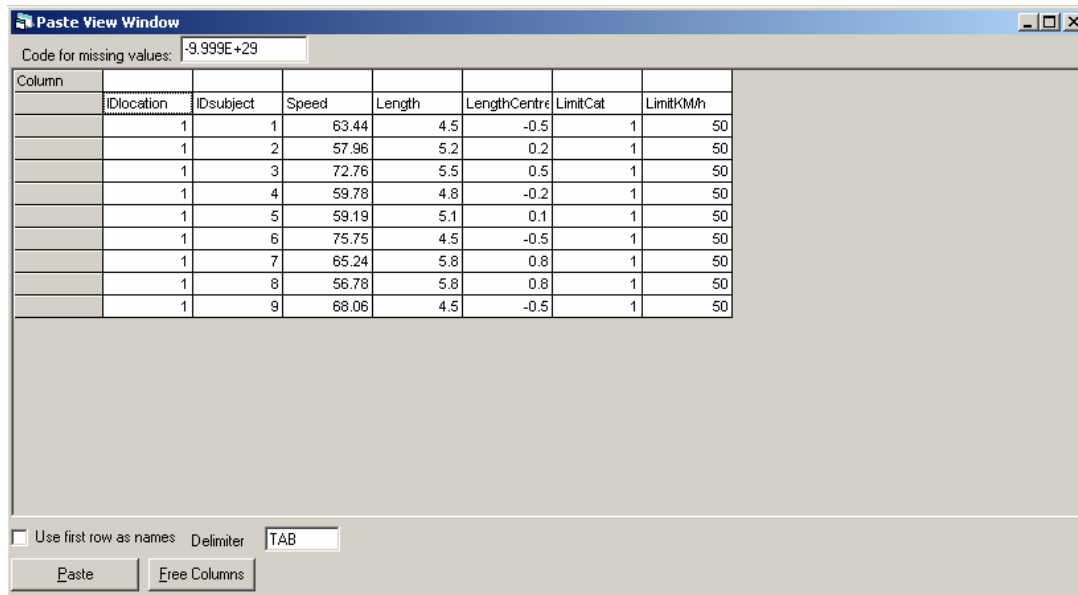
IDlocation	Identifies the road site (i.e. the location)
IDsubject	Identifies the subjects, i.e. the individual cars within each location
Speed	Indicates the speed of that car
Length	Indicates the length of that car
LengthCentred	Indicates the length of that car minus the average length
LengthCat	Indicates whether a car is shorter (0) or longer (1) than 4.3m
TrafficCount	Indicates the number of cars passing a road site during measurement
TrafficCountCat	Indicates whether fewer (0) or more (1) than 100 cars passed
Province	The Belgian Province in which measurement has taken place

Data load

To begin with, we will import the data from Excel.

- Open MLwiN
- Open the data file (SPEED.xls) in Excel
- Select all columns (ctrl A) and copy them (ctrl C) in Excel
- Go to MLwiN, press Ctrl V

The following window appears:



- Check “Use First row as names” in the lower left corner of the Paste View Window
- Click on “Free Columns” (this assigns the first free columns in MLwiN to the imported data)
- Click on “Paste”
- Close the paste window
- Select “File” in the top menu bar and save the worksheet you just created

The centre of MLwiN is the *Equations window*, where the models that you want to fit to the data are built.

- Click on “Model” in the menu bar and select “Equation”
- Click on “Notation” at the bottom of the Equations window
- Uncheck “General” (this changes the notation from the General-linear-model notation to the linear-model notation)
- Click “Done”

2.2.1.1. An “empty” single-level model

The first model built is one that ignores the hierarchical structure in the data. It includes only one level, that of the individual cars (IDsubject). This model, containing only an intercept and no predictors will be used as a point of reference.

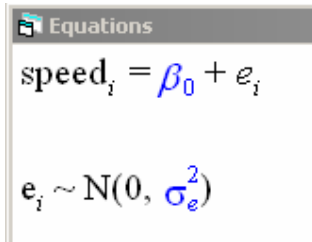
Model formulation

Define the dependent variable:

- Click on the “y” and
 - Select “Speed” from the drop-down menu as dependent variable
 - Select “1-i” from the N-levels drop-down menu
 - Select “ID-subject” from the level 1(j) drop-down menu

- Click “done”

The specified model includes only one random-term (e_i) and so far only an intercept. If your Equations window does not look like this,



Equations

$$\text{speed}_i = \beta_0 + e_i$$

$$e_i \sim N(0, \sigma_e^2)$$

click “Estimates” at the bottom of the Equations window. This changes back and forth between three views:

- the parameter names (e.g. β_0)
- the parameter names in colour coding
- the estimates for each parameter with their Standard Errors in parenthesis.

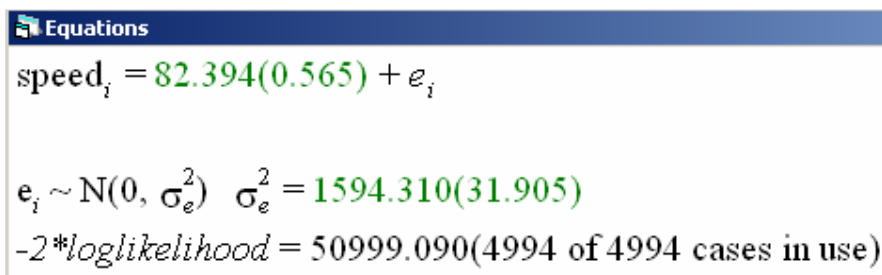
Colour coding of the parameters indicates their status:

- red: not yet specified
- blue: specified but not yet estimated
- green: estimation is complete

- Click “Estimates” until you see blue numbers in the equation.
- Press “Start” in the upper left corner of the MLwiN Window to start the estimation procedure

The Estimation is concluded when the numbers turn from blue to green.

Results and Interpretation



Equations

$$\text{speed}_i = 82.394(0.565) + e_i$$

$$e_i \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 1594.310(31.905)$$

-2*loglikelihood = 50999.090(4994 of 4994 cases in use)

*In this simple model, only two parameters are estimated (you can tell, because there are only two green numbers): the **intercept**, here simply the overall mean of Speed, and σ_e^2 the variance of the individual error-term e_i . The error e_i denotes the derivation of each individual (i.e. the cars) from the model, here simply the variance of the complete sample. Behind each parameter estimate, its **standard error** is indicated in parenthesis. To be significant, a parameter estimate has to be at least twice as large as its standard error. (To be more*

exact, the parameter estimate divided by its standard error is z-distributed. A Z-value of 1.96 indicates a two-tailed probability of 0.05). The third line of the output shows the **deviance** of the model (the $-2\log\text{likelihood}$). This value indicates how well the model fits the data. The smaller it is, the better the model fits. It is used to compare models. We will come back to that later.

A practical tip: MLwiN makes building different models very easy. It does, however, not allow to eyeball the results for two different models in parallel. As a solution we suggest to use a viewing software like **Irfan** (freeware) to produce and save screenshots of the models you build: Start IrfanViewer, click on Options in the menu-bar and select Capture/Screenshots. Press “Start” and close the window (but not the program!). Now go to the equations-window of MLwiN and press Ctrl-F11. Irfan now keeps a picture of your model and overwrites it (unless you save it) when you press Ctrl-F11 again.

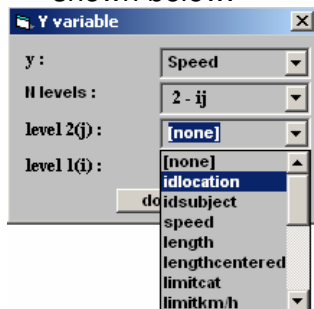
2.2.1.2. An “empty” two-level model

Values measured at the same location can be expected to be more similar to each other than to values measured at different locations. To include this hierarchical structure in the model, we will now define a two-level model with the cars (ID-subject) constituting the first level and the road sites (ID-location) constituting the second.

Model formulation

First define the dependent variable (Speed) as varying over two random factors, namely the individual cars (IDsubject) and the location of measurement (IDlocation)

- Click on the dependent variable and define IDlocation as the second level as shown below.

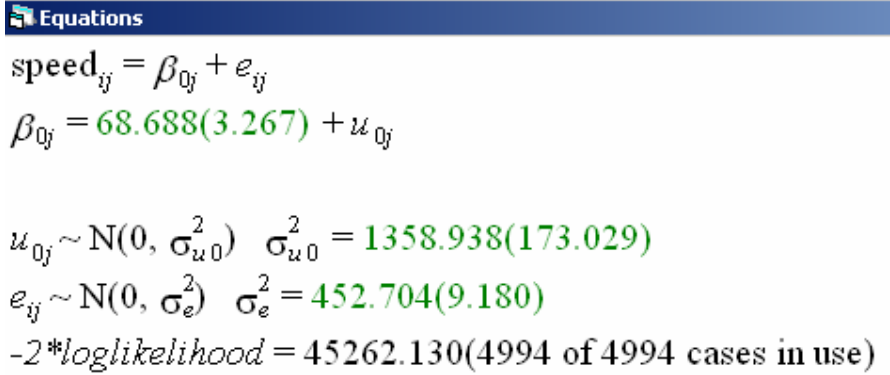


Then define a variance component model by allowing the intercept to vary randomly across locations.

- Click on the intercept
- Check the box $j(\text{IDlocation})$
- Press “Done”
- Press “Start” to estimate the parameters

Results and Interpretation

Your Equations window should now look like this.



$$\text{speed}_{ij} = \beta_{0j} + e_{ij}$$

$$\beta_{0j} = 68.688(3.267) + u_{0j}$$

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad \sigma_{u0}^2 = 1358.938(173.029)$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 452.704(9.180)$$

$$-2 * \text{loglikelihood} = 45262.130(4994 \text{ of } 4994 \text{ cases in use})$$

The model has two parts; the first equation corresponds to the individual car level (Level 1), the second equation to the level of the locations (Level 2). Instead of an overall intercept β_0 , the intercept β_{0j} varies across locations. The mean value (68.688), indicated in the second equation, gives the mean speed across all road sites.

Two error-variances are estimated: σ_{u0}^2 is the variance of u_{0j} , the location error term in the second equation. This location error-term u_{0j} indicates the derivation of each location-intercept from the mean intercept estimated in the second equation. σ_e^2 is the variance of e_{ij} , the individual error term in the first equation.

σ_{u0}^2 , the variation **between** locations, is highly significant and much larger than σ_e^2 , the variation **within** locations. The variance partition coefficient ($\sigma_{u0}^2 / (\sigma_{u0}^2 + \sigma_e^2)$) is .75 indicating that 75% of the total variance is due to variations between the level-two units (here the locations).

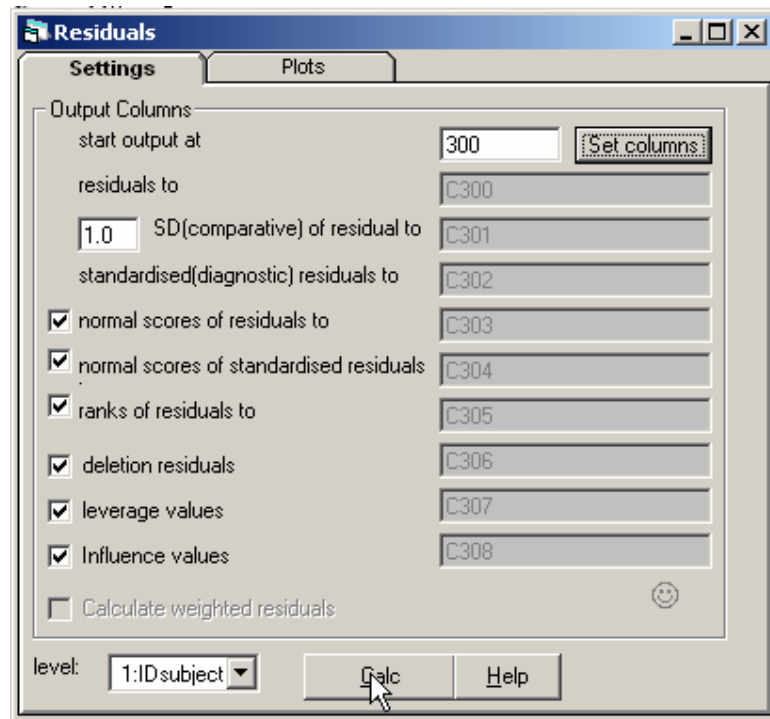
The deviance is now by a factor 10 smaller than that of the single-level model. The difference between the two deviances (in this case 464737) is Chi-square distributed with the difference in numbers of parameters as degrees of freedom (in this case one, because the two-level model estimates one parameter more than the one-level model). As a guideline, the expected value for a Chi-squared distribution is equal to the degrees of freedom, so there is little doubt that 464737 significantly exceeds this value. For a formal check, click on "Data Manipulation" in the top menu bar and select "Tail Areas", select "Chi-squared", fill the X^2 -value (464737) and the degrees of freedom (1) in and press "Calculate".

To conclude, both the variance partition coefficient and the deviance test both strongly suggest that a two-level structure describes the data more adequately than a single-level structure.

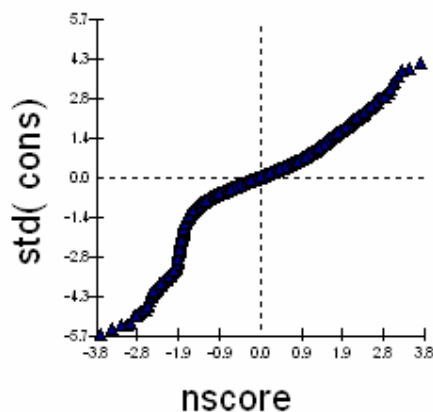
Graphic inspection of the residuals: Level 1

To inspect the residuals of the model,

- Click on Model in the top menu bar
- Select Residuals

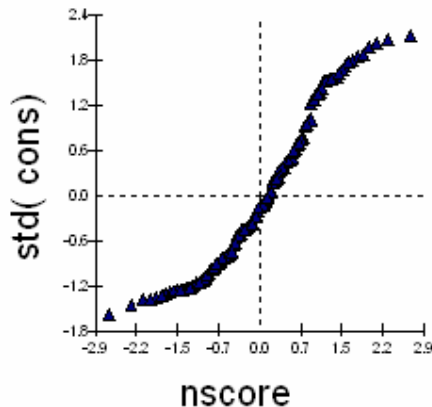


- Click “Calc” to calculate the Level 1 residuals
- Select “Plot” at the top of the Residuals window
- Check radio-button in front of “standardised residual and normal score”
- Click “Apply”

**Graphic inspection of the residuals: Level 2**

- Go back to the “Settings” part of the residuals window
- Select “2:IDlocation” from the “level” dropdown list

- Click “Calc” to calculate the Level 2 residuals
- Select “Plot” at the top of the Residuals window
- Click “Apply”



Normally distributed residuals would result in straight lines. Obviously this is not the case. A log-transformation ($\ln\text{Speed}$) could help normalize the speed distribution. The transformed value would, however, make the interpretation more difficult. Therefore, for the sake of clarity of interpretation the non-transformed speed variable is kept here. As an exercise, the reader is advised, however, to repeat the analyses presented here with the log-transformed speed variable ($\ln\text{Speed}$) as an exercise.

2.2.1.3. A Two-level variance component model with predictor length

Next, include the length of a car as a predictor for its speed. Rather than including the absolute length, include `LengthCentred`, the length of the car centred to its mean.

Model formulation

- Click “Add Term” at the bottom of the Equations window
- Select “LengthCentred” from the “Variable” drop-down window
- Click “Done”
- To estimate this model press “Start”

Results and Interpretation**Equations**

$$\text{speed}_{ij} = \beta_{0j} + 2.303(0.275)\text{lengthCentred}_{ij} + e_{ij}$$

$$\beta_{0j} = 68.878(3.236) + u_{0j}$$

$$u_{0j} \sim N(0, \sigma_u^2) \quad \sigma_u^2 = 1333.164(169.936)$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 446.478(9.053)$$

$$-2*\loglikelihood = 45192.310(4994 \text{ of } 4994 \text{ cases in use})$$

The intercept β_{0j} now presents the average speed at $\text{LengthCentred} = 0$ (i.e. for a car of average length) and the coefficient in front of LengthCentred indicates its **slope**: the change in speed per unit of length (here meters). The deviance ($-2*\loglikelihood$) decreased by 70, i.e. the introduction of car length as a predictor significantly improved the model.

2.2.1.4. Two-level random intercept, random slope model

To investigate whether the relation between speed and the length of a car was the same at all measurement locations, the slope of Length will now be allowed to vary randomly across locations too.

Model formulation

- Click on the Length
- Check the box j (IDlocation)
- Estimate the parameters by clicking on “Start”

Results and Interpretation

$$\text{speed}_{ij} = \beta_{0j} + \beta_{1j}\text{lengthCentred}_{ij} + e_{ij}$$

$$\beta_{0j} = 68.954(3.239) + u_{0j}$$

$$\beta_{1j} = 1.692(0.471) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 1334.843(169.700) & \\ -15.507(17.409) & 12.830(3.157) \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 412.745(8.456)$$

$$-2*\loglikelihood = 44901.500(4994 \text{ of } 4994 \text{ cases in use})$$

The model now has three parts; the first equation specifies the level of the individual cars and the other two the level of the locations. Both, the intercept β_{0j} and the coefficient of Length β_{1j} are now varying across locations with the means indicated in the second and third equation.

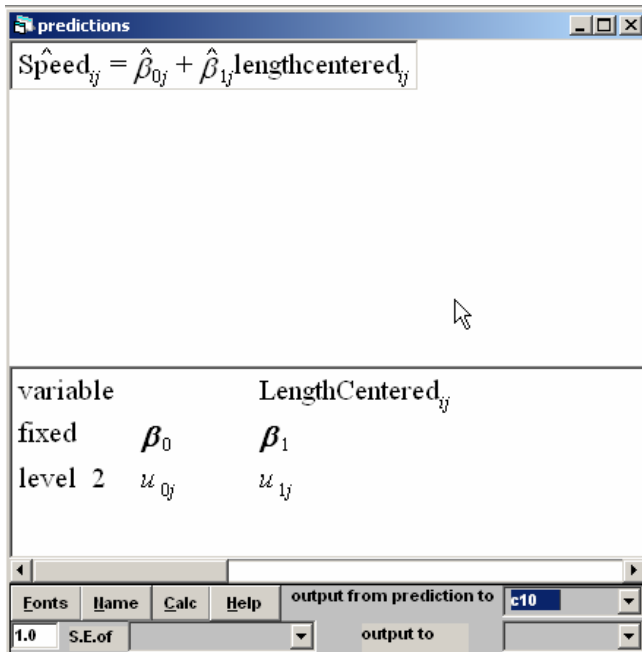
The location variance has now become a variance-covariance matrix Ω_u : The upper left number is σ^2_{u0} , the variance of the intercepts across locations. It indicates how much the general level of “Speed” varies between groups. The lower right number is σ^2_{u1} , the variance of the coefficient for Length across locations. It shows to what extent the relation between “Speed” and “LengthCentered” varies between groups. The lower left number is the covariance between the two, indicating to what extent there is a relation between the intercept (i.e. the general level of “Speed”) and the slope (i.e. the strength of the relation to “LengthCentred”) across locations. While the two variances can only be positive, the covariance can be positive or negative. A positive covariance indicates that larger intercepts are associated with larger slopes. The opposite is true for a negative covariance.

The deviance decreased by 181 as opposed to the variance component model, indicating that there is indeed substantial variation across locations in the effect that length has on speed. This also becomes apparent in the fact that the variance of the slope is significant. However the covariance between slope and intercept is not, indicating that there is no relation between the average level of speed (i.e. the intercept) and the length effect (i.e. the slope).

Graphic inspection of the model predictions

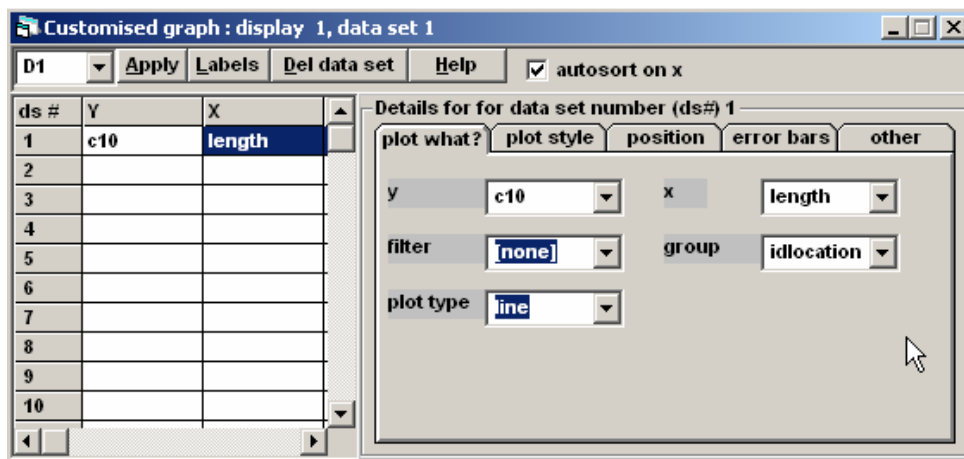
To interpret the results it can often help to make use of MLwiNs great graphical functions. In this case we will plot the predicted speed values for each location. To do so, we have to save the model-predictions as a new variable first.

- Select “Model” in the top menu bar and click on “Predictions”
- Click on all parameters in the lower half of the appearing dialogue window (so that they turn from grey to black)
- Select an empty column (e.g. c10) for “output from prediction to”
- Click on “Calc”
- Now you can close the “Predictions” window



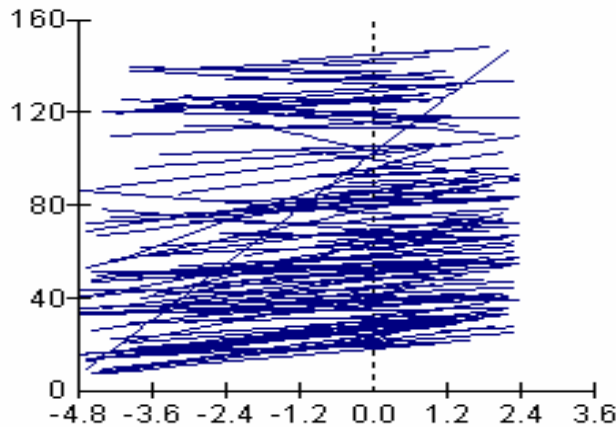
Now we have created a variable that contains the predictions of the model, which we are going to plot against the length.

Open the graph dialogue by clicking on “Graph” in the top menu bar and by selecting “Customized Graphs”. Fill in as shown below (all changes have to be made in the drop-down lists on the right-hand side and appear automatically in the left-hand side table)



- Press “Apply” to create the graph (You can close the Dialogue Window then.)

The resulting graph shows separate regression lines for each location.



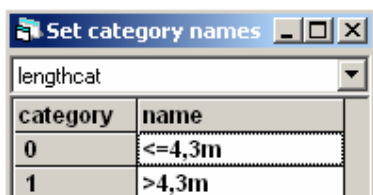
The size of σ^2_{u0} (the intercept-variance across locations) is reflected by variations in the height of regression lines and the size of σ^2_{u1} (the slope-variance across locations) in variation in their steepness. The size of σ_{u01} (the covariance between intercept and slope) would be reflected in the fact that lines that are on a higher level all-together (larger intercept) tend to be more or less steep than those at the bottom of the graph. As the covariance is not significant here, it is however not possible to see such a tendency.

2.2.1.5. Adding a categorical predictor

To demonstrate how a categorical predictor can be included into the model, the continuous variable LengthCentred will be replaced by a categorical one (LengthCat) that simply indicates whether the length of a car is below (0) or above (1) average. The first step is to define this variable as categorical rather than continuous.

Model formulation

- Click on “Data Manipulation” in the top menu bar and select “Names”
- Select LengthCat
- Click “Categories”
- Define the categories as shown below (simply start typing after you clicked on each field)
- Press “Apply”
- Close the “Names” window



- Remove LengthCentred by clicking on the term and then on “Delete Term”
- Include LengthCat with “Add Term”
 - Choose “<=4,3m” as the reference category
- Estimate the model

Results and Interpretation

Equations

$$\text{speed}_{ij} = \beta_{0j} + 4.974(0.755) \text{>4,3m}_{ij} + e_{ij}$$

$$\beta_{0j} = 65.034(3.284) + u_{0j}$$

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad \sigma_{u0}^2 = 1333.842(169.939)$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 448.924(9.103)$$

$$-2 \cdot \log \text{likelihood} = 45218.960(4994 \text{ of } 4994 \text{ cases in use})$$

The coefficient of speedCat indicates that long cars go 4.97 km/h faster than short ones. The other parameters are very similar to the model in 2.2.1.3 with length as a continuous predictor.

As the next step the categorical speed effect is allowed to vary randomly across locations.

Model formulation

- Click on the coefficient for SpeedCat
- Check “j(IDlocation)”
- Click “Done”
- Start the estimation by clicking “Start”

Results and Interpretation

Equations

$$\text{speed}_{ij} = \beta_{0j} + \beta_{1j} \text{Length}_{ij} + e_{ij}$$

$$\beta_{0j} = 65.007(3.482) + u_{0j}$$

$$\beta_{1j} = 5.105(1.333) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 1472.439(195.631) & -132.266(55.116) \\ -132.266(55.116) & 99.254(24.495) \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 418.309(8.567)$$

$$-2 * \log\text{likelihood} = 44963.590(4994 \text{ of } 4994 \text{ cases in use})$$

Contrary to the model in 2.2.1.4, with Length as a continuous predictor, the covariance between intercept and slope is now significant. Its negative value indicates that the difference between long and short cars is smaller at locations with a high overall speed level (i.e. a large intercept) as compared to locations with a low overall speed level. The decreased deviance value (45218-44963 = 255) is highly significant indicating that the effect of speed is indeed not constant across locations.

2.2.1.6. Adding a contextual variable

One of the advantages of multilevel models is the possibility to include predictors situated at different levels simultaneously in the model. As an example of a higher level variable, the traffic count for each road site will be taken up into the model as a contextual predictor (that means it does not vary across Level 1, but only across higher level units, here the locations). The variable TrafCountCat takes the value 0 for each road site with fewer than 100 cars passing during observation and 1 for road sites on which more than 100 cars passed.

Model formulation

First go back to model 2.2.1.4, the random slope model with the continuous length variable:

- Delete LengthCat from the equation
- take up LengthCentered again
- make the effect of LengthCentred vary randomly across locations

Now add the context variable

- Include TrafCountCat into the equation with “Add Term”.

- Select ≤ 100 as the reference category

Results and Interpretation

Equations

$$\text{speed}_{ij} = \beta_{0j} + \beta_{1j} \text{lengthCentred}_{ij} + 33.166(6.505) > 100_j + e_{ij}$$

$$\beta_{0j} = 59.494(3.502) + u_{0j}$$

$$\beta_{1j} = 1.654(0.471) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 1107.560(141.560) & \\ -15.656(15.841) & 12.891(3.136) \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 412.737(8.457)$$

$$-2 * \log \text{likelihood} = 44877.820(4994 \text{ of } 4994 \text{ cases in use})$$

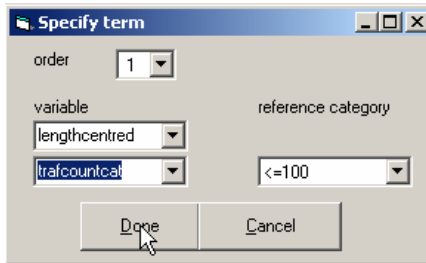
The coefficient for trafficCount, category ">100" indicates that at road sites with more than 100 cars passing, cars went on average 33.2 km/h faster than on road sites with fewer than 100 cars passing. The coefficient is highly significant. Moreover the deviance is reduced by 385 as opposed to the model in 2.2.1.4. Both indicates that the number of cars passing at a road sites is a good predictor of speed at that road site.

2.2.1.7. Testing for a cross-level interaction

In order to test whether the context variable trafficCount modifies the length-effect at the level of the individual cars, the interaction between TrafficCountCat and LengthCentred is added to the model.

Model formulation

- Click on "Add term" and
- Include the interaction between LimitCat and LengthCentred as shown below
 - Select order 1 (this means it is a first-order interaction)
 - Choose again ≤ 100 as reference category for TrafficCountCat
- Estimate the model



Results and Interpretation

$$\text{speed}_{ij} = \beta_{0j} + \beta_{1j} \text{lengthCentred}_{ij} + 33.215(6.534) \text{>100}_j + -0.077(0.973) \text{lengthCentred.>100}_{ij} + e_{ij}$$

$$\beta_{0j} = 59.480(3.506) + u_{0j}$$

$$\beta_{1j} = 1.683(0.596) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 1107.345(141.530) & -15.586(15.870) \\ -15.586(15.870) & 12.850(3.151) \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 412.754(8.456)$$

$$-2 * \log \text{likelihood} = 44877.820 (4994 \text{ of } 4994 \text{ cases in use})$$

The coefficient for the interaction (*lengthCentred.>100*) is clearly not significant, as it is much smaller than its standard error in parenthesis. The negligible weight for this interaction indicates that the length effect at road sites with more than 100 cars passing were not different from those at road sites with fewer cars passing. The same message is conveyed by the likelihood that did not decrease from the model in 2.2.1.6 to the present model, suggesting that adding the interaction introduces complexity that does not explain anything.

To conclude, the speed of cars varies more between road sites than within them. Accordingly, speed is affected by a level-1 variable (*length*) to some extent, but much more so by a level-2 variable. The effect of *length* is not modified by the traffic count.

2.2.1.8. Conclusion

In this chapter it was demonstrated how to extend a linear regression model to a multilevel structure. It was demonstrated how the variance partition coefficient and the deviance test can be used to establish the appropriateness of the multilevel structure and how predictions and residuals at these different levels can be presented graphically. Moreover, it was explained how effects of level-1 predictors (here the length of a car) can be considered together with predictors at higher levels (here the traffic count at the measurement location) and how an interaction between variables at different levels can be investigated.

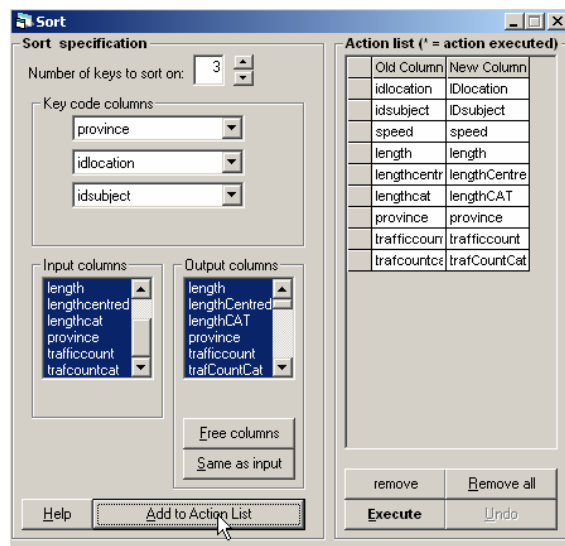
2.2.2 Three level models and more

The research example concerning speed measurement at Belgian road sites is continued here. Each location of measurement is not only characterised by the characteristics of the road site in question but also by the Province (Belgian regions with limited governmental responsibilities) it was situated in. To investigate whether this hierarchical structure is also represented in the data, a three level model will be fit.

Data-load

Open the worksheet you created in the previous section or paste the data from Excel, as described there. Before including a higher level, the data have to be resorted. In general, the data have to be sorted by all levels included (save the lowest one, here IDsubject, which is simply indicated by the rows in the data file). The data imported from Excel are sorted according to the two-level structure, i.e. by IDlocation and then by IDsubject. In order to include Province as a third level, they have to be sorted by Province, then IDlocation, and then IDsubject (note that the order in sorting is opposite to the numbering of the levels).

- Click on “Data Manipulation” in the top menu bar and select “Sort”
- Select 3 for “Number of keys to sort on”
- Select Province, IDlocation, and IDsubject as key codes
- Mark all variables in the “Input columns” list
- Click on “Same as input” so that the same columns appear in the “Output column” list
- Click on “Add to Action List”
- Click on “Execute”
- Close the Sorting Dialogue



2.2.2.1. A three-level variance component model

The first three-level model to test would always be the variance component model in which the intercepts but not the slopes vary across the levels.

Model formulation

Define the dependent variable (Speed) as varying over three random factors, namely the individual cars (IDsubject), the location of measurement (IDlocation), and the province.

- Click on the dependent variable and define as shown below

Then define a variance component model by allowing the intercept to vary randomly across locations and take up LengthCentred as a predictor.

- Click on the intercept
- Check the box j(IDlocation)
- Check the box k(Province)
- Press "Done"
- Include LengthCentred as a predictor with "Add Term"
- Press "Start" to estimate the parameters

Results and Interpretation

$$\text{speed}_{ijk} = \beta_{0jk} + 2.286(0.275)\text{lengthCentred}_{ijk} + e_{ijk}$$

$$\beta_{0jk} = 74.489(5.329) + v_{0k} + u_{0jk}$$

$$v_{0k} \sim N(0, \sigma_{v0}^2) \quad \sigma_{v0}^2 = 218.018(132.889)$$

$$u_{0jk} \sim N(0, \sigma_{u0}^2) \quad \sigma_{u0}^2 = 990.124(132.623)$$

$$e_{ijk} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 446.467(9.054)$$

$$-2*\loglikelihood = 45167.920(4994 \text{ of } 4994 \text{ cases in use})$$

The intercept β_{0jk} now varies across locations and provinces, resulting in the estimation of three variances: σ_{v0}^2 is the variance of u_{0k} , the derivation of each province's intercept from the mean intercept. σ_{u0}^2 estimates the variation of the

intercept between locations but within the provinces and σ_e^2 is the variance of e_{ij} , the individual error term in the first equation.

The deviance of this three-level model as compared to the two-level variance component model presented in 2.2.1.3 decreased by 24, which is significant ($p < .000$). This suggests that the introduction of the three-level structure is justified. Note, however, that σ_{v0}^2 , the variance at level 3 (i.e. the variance of the intercept across provinces), is only marginally significant. (To test this, click on "Basic Statistics", select "Tail Areas", check "Standard Normal Distribution", divide the parameter estimate 218 by its standard error 132.9 and fill this value in the slot next to "Value". Press calculate. The resulting p value is .051). In contrast, the level-2 variance, σ_{u0}^2 , and the level-one variance, σ_e^2 , are both clearly significant.

Another way to estimate the importance of each level is to calculate the variance component coefficients for Level 2 ($\sigma_{u0}^2 / \sigma_e^2 + \sigma_{u0}^2 + \sigma_{v0}^2 = .60$) and for Level 3 ($\sigma_{v0}^2 / \sigma_e^2 + \sigma_{u0}^2 + \sigma_{v0}^2 = .13$).

These results place the third level somewhere in a grey zone: The model including the third level fits the data better than the two-level model (suggesting that there is variation between the provinces that make up the third level), but at the same time, the variance of that third level is not significant. The variance component coefficients indicate that the largest part of the variation is situated at the level of the road sites (60%) and only a small part is situated at the level of the provinces (13%). One can conclude that there is some variation between provinces but that the variation between the locations is much more important.

2.2.2.2. A three-level model with a random slope at Level 2

From the two-level model in 2.2.1.4 we already know that the effect of car length varies across locations. Accordingly, the next model to estimate is the three-level model including a random slope for length at Level 2.

Model formulation

- Click on the Length
- Check the box j (IDlocation)
- Estimate the parameters by clicking on "Start"

Results and Interpretation

Equations

$$\text{speed}_{ijk} = \beta_{0jk} + \beta_{1j} \text{lengthCentred}_{ijk} + e_{ijk}$$

$$\beta_{0jk} = 74.509(5.291) + v_{0k} + u_{0jk}$$

$$\beta_{1j} = 1.675(0.470) + u_{1jk}$$

$$v_{0k} \sim N(0, \sigma_{v0}^2) \quad \sigma_{v0}^2 = 213.109(131.577)$$

$$\begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 995.645(133.383) & -11.867(15.426) \\ -11.867(15.426) & 12.761(3.118) \end{bmatrix}$$

$$e_{ijk} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 412.738(8.457)$$

$$-2 * \log\text{likelihood} = 44877.420(4994 \text{ of } 4994 \text{ cases in use})$$

The variance of the intercept across provinces σ_{v0}^2 is still only marginally significant ($p=.053$). For the rest the estimates look very similar to the output of the two-level model (see section 2.2.1.4). There is a lot of variation in the intercept across locations, a small but significant amount of variation in the slope of length across locations and no significant covariation between slope and intercept. The introduction of a random slope for speed decreased the deviance by 290, which is highly significant.

2.2.2.3. A fully random three-level model

As a last step, the effect of speed will be allowed to vary randomly not only across locations but also across provinces.

Model formulation

- Click on the Length
- Check the box k (province)
- Estimate the parameters by clicking on “Start”

Results and interpretation**Equations**

$$\text{speed}_{ijk} = \beta_{0jk} + \beta_{1jk} \text{lengthCentred}_{ijk} + e_{ijk}$$

$$\beta_{0jk} = 74.528(5.341) + v_{0k} + u_{0jk}$$

$$\beta_{1jk} = 1.642(0.591) + v_{1k} + u_{1jk}$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 219.055(133.882) & -4.181(10.570) \\ -4.181(10.570) & 1.307(1.598) \end{bmatrix}$$

$$\begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 994.165(133.233) & -11.434(15.298) \\ -11.434(15.298) & 11.642(3.121) \end{bmatrix}$$

$$e_{ijk} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 412.743(8.456)$$

$$-2 * \loglikelihood = 44876.820(4994 \text{ of } 4994 \text{ cases in use})$$

The results now contain two variance-covariance matrices: Ω_v and Ω_u . Ω_v indicates the variance and covariance of intercept and slope across provinces and Ω_u across locations. The variance-covariance matrix for the locations is still relatively unchanged as compared to the two-level model (2.2.1.4). For provinces, the intercept variance is still only marginally significant. The slope variance and the slope/intercept covariance that have been added to the model are clearly not significant (to be significant at the .05-level they would have to exceed twice the size of their standard error, which is clearly not the case). The conclusion that this last extension of the model was not necessary is confirmed by the deviance test. Although the present model estimates two extra parameters, the deviance is exactly the same as that of the simpler model with the length effect only varying at the level of the locations.

To conclude, the speed of cars has shown to vary substantially across measurement locations and only to a limited extent across provinces. There is a relation between the length of a car and its speed and this relation varies across measurement locations but not across provinces. The introduction of a third level has proven of limited use. Not only is the variation attributed to this level very limited, but moreover the results for the other levels were almost exactly the same as in the two-level models.

2.2.2.4. Conclusion

It has been shown how to extend the two-level models presented in section 2.2.1 to three level models. It has also been demonstrated how to investigate

whether the additional level improves the model and in the present case it has been concluded that a two-level model would be sufficient.

2.3 Discrete response models

2.3.1 Introduction

In the methodology report the section for discrete responses is introduced by outlining the generalised linear models (GLM) and their hierarchical version, the multilevel GLM. As there is no empirical example in the introduction of the GLM, there is no corresponding manual section. In the following sections, the analysis of general binomial responses (2.3.2), multinomial responses (2.3.3) and counts (2.3.4) will be presented. All these analyses are instances of the multilevel GLM.

2.3.2 Binary and general binomial responses

Heike Martensen and Emmanuelle Dupont (IBSR)

The example data used in this section were gathered in a Belgian drink driving roadside survey. At 413 randomly selected road sites 11,186 drivers were stopped, asked to perform an alcohol breath test and to answer a number of questions. The data contain the following variables:

DrinkDriving	Was the alcohol concentration of the driver above the legal limit of .05 g/l? Yes=1, No=0.
ID_ind	Identification number of each driver tested.
ID_loc	Identification number of each test location.
Gender	Gender of the driver: Male =1, Female =2.
Age	A categorical variable: 16-25=1, 26-39=2, 40-54=3, 55+=4.
Previously	Has the driver been tested for alcohol at a roadside control previously? Yes=1, No=0.
Probability	How high does the driver estimate the probability of being stopped for an alcohol control? Very Low=1, Low=2, Medium=3, High=4, Very High=5.
TrafficCount	The average number of cars passing the test site within 15 minutes.
Intensity	The number of control officers present divided by TrafficCount.

Data load

- Click on “File” in the top menu bar and select “Open worksheet”.
- Open “ALCOHOL.ws”

Two constant variables (“denom” and “cons”) are necessary for building a model for binary data. To generate these variables,

- Click on “Data Manipulation” and select “Generate Vector”
- Select “Constant Vector” as type of vector
- Select an empty output column (e.g. c30)
- Fill in the number of cases (11,186) at “Number of Copies”
- Fill in “1” at “Value”

- Click “Generate”
- Select another empty output column (e.g. c31)
- Click “Generate” again
- Close the “Generate Vector” dialogue window
- Click on “Data Manipulation” and select “Names”
- Select the first variable just generated (i.e. c30)
- Type “cons” in the field at the top of the window and press return
- Name the second variable just generated (i.e. c31) “denom”

Model formulation

Define the dependent variable

- Click on the “y” and
 - Select “DrinkDriving” from the drop-down window as dependent variable
 - Select “2-ij” from the N-levels drop-down window
 - Select “ID_loc” from the level 2(j) drop-down window
 - Select “ID_ind” from the level 1(i) drop-down window
 - Click “done”
- Click on the N in the Distribution statement for DrinkDriving and
 - Check “Binomial”
 - In the appearing link functions, leave “logit” checked
 - Click “Done”
- Click on the red n_{ij} in the distribution statement for DrinkDriving
 - Select “denom” in the variable drop down list

A binomial distribution is characterised by the proportion π_{ij} and the denominator n_{ij} stating the number of instances on which the proportion is based. In the present study the denominator is a constant 1, meaning that the data are binary. This and the choice of the “logit” function as a link function make the model a logistic regression model.

Build a two-level random intercept model

- Click “Add Term” and select “Cons” from the variable drop down list
- Click on “cons” and check “j(id_loc)”

In the General Linear Model notation, the intercept is not automatically included. To do so, a constant variable must be included as a predictor.

Add predictors

- Click on “Add Term”, select “TrafficCount” as a predictor from the variable drop down list
- Click on “Add Term”, select “Intensity” as a predictor
- Click on “Add Term”, select “Gender” as a predictor, chose “Male” as reference category

- Click on “Add Term”, select “Previously” as a predictor, select “not tested previously” as reference category
- Click on “Add Term”, select “Probability” as a predictor; select “very low” as reference category
- Click on “Add Term”, select “Previously” as a predictor, select “Age16-25” as reference category

Estimation

In the binomial distribution it is assumed that the variance is equal to the odds-ratio, $\pi_{ji} (1 - \pi_{ji})$. To test this assumption, first estimate a model assuming an extra-Binomial distribution, where the variance is left free to vary.

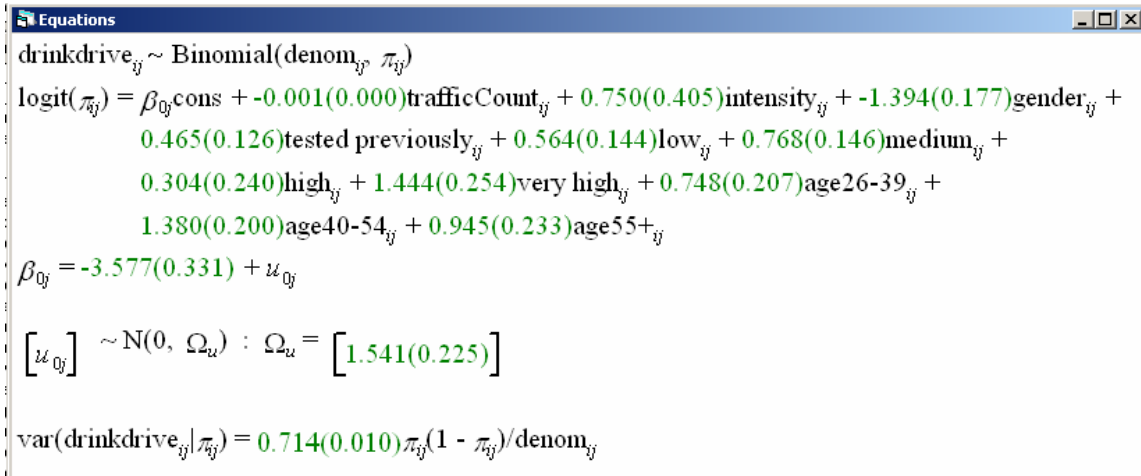
- Click on “Notation” at the bottom of the “Equations” window again
- Select “extra Binomial” in the top row
- Leave the other two options at their default value (1st order linearization and MQL as estimation type)
- Click “Done”
- Press “Start” to start the estimation procedure.

Once the estimation procedure has converged (i.e. all blue numbers turned green)

- Click on “Notation” at the bottom of the “Equations” window again
- Select “2nd order” under “Linearization”
- Select “PQL” under “Estimation type”
- Press “Done”
- Press “More” to continue the estimation

In the estimation procedure, the nonlinear link function is linearized by approximating it with a Taylor Series expansion. A Taylor series consists of an infinite number of terms and the more of them are used, the closer the approximation. The first choice is whether only the first (1st order linearization) or the first two are used (2nd order linearization). The other choice concerns the values that the Taylor series expansion is based on: During each iteration the Taylor series is calculated on the basis of the currently estimated parameter values. In the Marginal Quasi Likelihood method (MQL) only the fixed parameters are included, in the Penalized Quasi Likelihood method (PQL) the residuals are included as well. Generally speaking, 2nd order linearization and PQL are more accurate but computationally intensive and more prone to convergence problems. The 1st order MQL estimates on the other hand are known to be biased downwards. It is therefore suggested to use 1st order linearization and MQL to get rough starting values on which the final estimation using 2nd order linearization and PQL is based. For more information see Goldstein (2003) or Hox (2002). With these methods, slight variations in the estimated values are to be expected.

Results



Equations

$$\text{drinkdrive}_{ij} \sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j}\text{cons} + -0.001(0.000)\text{trafficCount}_{ij} + 0.750(0.405)\text{intensity}_{ij} + -1.394(0.177)\text{gender}_{ij} +$$

$$0.465(0.126)\text{tested previously}_{ij} + 0.564(0.144)\text{low}_{ij} + 0.768(0.146)\text{medium}_{ij} +$$

$$0.304(0.240)\text{high}_{ij} + 1.444(0.254)\text{very high}_{ij} + 0.748(0.207)\text{age26-39}_{ij} +$$

$$1.380(0.200)\text{age40-54}_{ij} + 0.945(0.233)\text{age55+}_{ij}$$

$$\beta_{0j} = -3.577(0.331) + u_{0j}$$

$$[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = [1.541(0.225)]$$

$$\text{var}(\text{drinkdrive}_{ij}|\pi_{ij}) = 0.714(0.010)\pi_{ij}(1 - \pi_{ij})/\text{denom}_{ij}$$

The first interest is to evaluate whether the assumption of a binomial distribution holds. In that case the theoretically expected value for the variance would be 1. At the bottom of the Equations window, the estimated variance is indicated with 0.711. As this is very close to the theoretically expected value, it is probably safe to estimate the model under the more restrictive assumption of a Binomial distribution (rather than an extra-Binomial one).

Estimate the model again assuming a Binomial distribution.

- Click on “Notation” at the bottom of the “Equations” window again
- Click on “Use Defaults” (i.e., Binomial distribution, 1st order linearization and MQL as estimation type)
- Click “Done”
- Press “Start” to start the estimation procedure.

Once the estimation procedure has converged (i.e. all blue numbers turned green)

- Click on “Notation” at the bottom of the “Equations” window again
- Select “2nd order” under “Linearization”
- Select “PQL” under estimation type
- Press “Done”
- Press “More” to continue the estimation

Results and interpretation

Equations

$$\text{drinkdrive}_{ij} \sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j} + \text{cons} + -0.002(0.000)\text{trafficCount}_{ij} + 0.898(0.379)\text{intensity}_{ij} + -1.374(0.206)\text{female}_{ij} + 0.407(0.140)\text{tested previously}_{ij} + 0.536(0.166)\text{low}_{ij} + 0.743(0.168)\text{medium}_{ij} + 0.313(0.277)\text{high}_{ij} + 1.431(0.289)\text{very high}_{ij} + 0.709(0.241)\text{age 26-39}_{ij} + 1.312(0.233)\text{age 40-54}_{ij} + 0.859(0.271)\text{age 55+}_{ij}$$

$$\beta_{0j} = -4.746(0.283) + u_{0j}$$

$$[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = [0.958(0.194)]$$

$$\text{var}(\text{drinkdrive}_{ij} | \pi_{ij}) = \pi_{ij}(1 - \pi_{ij})/\text{denom}_{ij}$$

The parameter estimates of the Binomial model are very similar to that of the extra-Binomial one, confirming that the Binomial distribution assumptions hold for the present analysis. The interpretation will therefore be based on this last model.

Before interpreting the coefficients, their significance has to be tested. For categorical variables with several levels (e.g. probability) there is more than one predictor (here 4: low, medium, high, and very high) which have to be tested jointly. This is done with the Multivariate Wald test. Single coefficients for continuous predictor variables (e.g. TrafficCount) or those with only two levels (e.g. Gender) can be tested with the Z-test.

To conduct the Z-test:

- Divide the coefficients by their standard errors
- Click on “Basic Statistics” in the top menu bar and select “Tail Areas”
- Check “Standard Normal distribution”
- Fill in the result of the division
- Click “Calc”

To conduct a Multivariate Wald test (joint Chi-square test)

- Click on “Model” in the top menu bar and select “Intervals and tests”
- Check “fixed” at the bottom of the appearing dialogue window
- Type a 1 in front of every coefficient that you want to test jointly (e.g. those for “low probability”, “medium probability”, “high probability”, and “very high probability”)
- Click on “Calc”
- Click on “Basic Statistics” in the top menu bar and select “Tail Areas”
- Check “Chi Squared”
- Fill in the resulting Chi-square value from the “Intervals and tests” dialogue
- Click “Calc”

As can be seen in Table 2.3.1, all predictor variables are significant.

Predictor	Coefficient	SE	Z	p(Z)	Chi2	d.f.	p(chi2)	e ^{coefficient}
TrafficCount	-0.002	0.0001	-20.00	0.000				0.998
Intensity	0.898	0.379	2.37	0.009				2.455
Female	-1.374	0.206	-6.67	0.000				0.253
Previously	0.407	0.14	2.91	0.002				1.502
Prob. Low	0.536	0.166			25.46	4	0.000	1.709
Prob. Medium	0.743	0.168						2.102
Prob. High	0.313	0.277						1.368
Prob. Very high	1.431	0.289						4.183
Age26-39	0.709	0.241			18.17	3	0.000	2.032
Age40-54	1.312	0.233						3.714
Age55+	0.859	0.271						2.361

Table 2.3.1: Results of single and joint tests for predictors

One way to interpret the coefficients is to take their exponentials, which is presented in the right-most column of Table 2.3.1. For a one unit increase in the predictor, the odds of the dependent variable have to be multiplied by the exponential of the coefficient.

The odds of an event are calculated as the number of events divided by the number of non-events. For example, on average 3 drivers in every 100 are drunk, so the odds for any randomly chosen car of having a drunk driver are: $3/97 = 0.031$. The odds for an event that is as likely to happen as not ($p=0.5$) are 1. While odds have useful mathematical properties, they can produce counterintuitive results because they are similar to probabilities in the lower ranges (the odds of $p=.01$ are .0101) but not at all in the higher ranges (the odds of $p=.75$ are 3 and those of $p=.99$ are 99). As an example: an 80% probability is four times the chance of a 20% probability but the odds are 16 times higher.

Another way to interpret the coefficients, is to calculate the probability for different values of the predictor. The probability for any chosen value of a predictor x is given by:

$$\pi_i = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_i))} \quad (2.3.1)$$

In Table 2.3.2, the probability for DrinkDriving is given for each predictor taking the value of 1, while all other predictors are 0. In the third column, this probability is divided by the probability at the intercept (i.e. for all predictors being zero), indicating the multiplicative factor on the probability for a one-unit increase. This factor is compared to the exponential of the coefficient, the multiplicative factor on the odds. As can be seen, the two right-most columns are exactly the same, indicating that with such a small proportion of drink driving, the difference between probabilities and odds are negligible.

The interpretation of each coefficient is described elaborately in section 2.3.2 of the methodology report (D7.4) and will not be fully repeated here. As an example we

Predictor	Coefficient	$\pi_{ji} x=1$	$(\pi_{ji} x=0) / (\pi_{ji} x=1)$	$e^{\text{coefficient}}$
Intercept	-4.746	0.009		
TrafficCount	-0.002	0.009	0.998	0.998
Intensity	0.9	0.021	2.460	2.460
Female	-1.374	0.002	0.253	0.253
Previously	0.407	0.013	1.502	1.502
Prob. Low	0.536	0.015	1.709	1.709
Prob. Medium	0.743	0.018	2.102	2.102
Prob. High	0.313	0.012	1.368	1.368
Prob. Very high	1.431	0.036	4.183	4.183
Age26-39	0.709	0.018	2.032	2.032
Age40-54	1.312	0.032	3.714	3.714
Age55+	0.859	0.021	2.361	2.361

Table 2.3.2: Probability of DrinkDriving for each predictor at $x=0$ and $x=1$.

will describe the interpretation of one continuous variable (TrafficCount) and of a categorical one (Age).

TrafficCount has a negative weight, indicating that for each car passing, the proportion of drink driving decreases. The exponential of the coefficient (.998) indicates that for the decrease in one car the odds have to be multiplied by .998, i.e. decrease by 0.2%. Note that the relation between predictor and dependent variable is not linear. To establish the decrease in odds for 100 cars passing, one has to multiply the coefficient by 100 before taking the exponential which results in 0.819 or an 18% decrease.

The coefficients for the age categories 26-39, 40-54, and 55+ are all positive and thus result in exponential coefficients larger than 1. This means all age groups show a higher incidence of drink driving than the youngest drivers (16-25) who constitute the reference category. Most notably, in the age-group of 40-54 year olds the exponential coefficient amounts to 3.714, which indicates that drink driving in this age group occurs almost four times as often as among the young drivers.

The joint chi-square test reported above indicates that there is a difference between the age-groups somewhere. One might also want to test, whether two particular age groups differ from each other significantly. To test whether the 40-54 year olds differ from the 55+ year olds, follow the same procedure as described above, but put 1 in front of "age 40-54" and -1 in front of "age 55+". The difference between those two age groups is significant ($X^2(1)=5.58$, $p=.018$). We can therefore conclude that the 40-54 year olds drink and drive significantly more often than even the group of people older than 55+ who feature the second highest rate of drink driving.

2.3.2.1. Conclusion

A multilevel version of a logistic regression analysis was presented. Special characteristics of the estimation procedure for binary variables were discussed. The binomial model was compared to the extra-binomial model and it was concluded that the binomial distribution holds. It was demonstrated how to use

the joint Wald test to test the significance of categorical variables and shown how the coefficients can be interpreted either by transforming them into odds ratios or calculating the probability for specific values of the predictors.

2.3.3 Multinomial responses

Heike Martensen and Emmanuelle Dupont (IBSR)

The example data used in this section are the Belgian drink driving roadside survey data also used in section 2.3.2. The data contain the same variables as in 2.3.2, with the exception of the dependent variable. Rather than DrinkDriving (a dichotomous variable) in this chapter a variable with three possible outcomes, "Breathtest" will be modelled:

Breathtest	1 = <i>Safe</i> ; blood-alcohol concentration (BAC) is below 0.05 mg/l. 2 = <i>Alarm</i> ; driver's BAC is between 0.05 and 0.08 mg/l. 3 = <i>Positive</i> ; driver's BAC is above 0.08 mg/l.
ID_ind	Identification number of each driver tested.
ID_loc	Identification number of each test location.
Gender	Gender of the driver: Male =1, Female =2.
Age	A categorical variable: 16-25=1, 26-39=2, 40-54=3, 55+=4.

Categorical responses can be perceived in two ways: They can either form an ordered series that is based on some underlying continuous variable or they consist of different categories that are not systematically related. In the present case, the three categories ("safe", "alarm", "positive") are clearly related to the underlying variable blood alcohol concentration (BAC) and will therefore be modelled in an *ordered proportional odds* model. At the end of this section, the *unordered category* model will be presented, so that the reader can see how such a model is fitted and how the results differ from an ordered model.

Data load

Click on "File" in the top menu bar and select "Open worksheet".
Open "ALCOHOL.ws"

A constant ("cons") is necessary for building a model for categorical data. To generate this variable (unless you have done it and saved it in section 2.3.2),

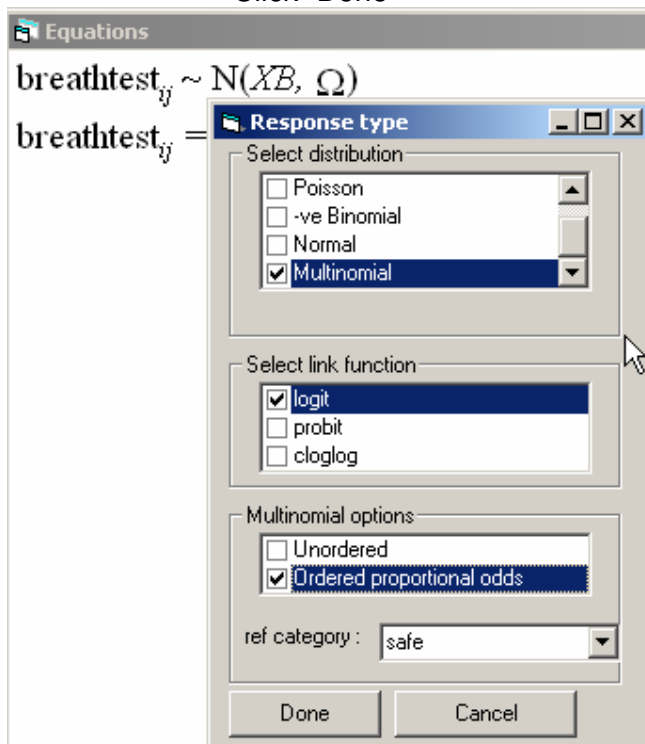
- Click on "Data Manipulation" and select "Generate Vector"
- Select "Constant Vector" as type of vector
- Select an empty output column (e.g. c30)
- Fill in the number of cases (11,186) at "Number of Copies"
- Fill in "1" at "Value"
- Click "Generate"
- Close the "Generate Vector" dialogue window
- Click on "Data Manipulation" and select "Names"
- Select the first variable just generated (i.e. c30)
- Type "cons" in the field at the top of the window and press return

2.3.3.1. Ordered proportional odds: empty single level model

Model formulation

Define the dependent variable

- Click on the “y” and
 - Select “Breathtest” from the drop-down window as dependent variable
 - Select “2-ij” from the N-levels drop-down menu
 - Select “ID_loc” from the level 2(j) drop-down menu
 - Select “ID_ind” from the level 1(i) drop-down menu
 - Click “Done”
- Click on the N in the Distribution statement for Breathtest and
 - Check “Multinomial”
 - In the appearing window, leave “logit” checked
 - Select “Ordered proportional odds”
 - Leave “safe” as reference category
 - Click “Done”



- Click on the red n_{ij} in the distribution statement for Breathtest
 - Select “cons” in the variable drop down list

When a response variable is defined as multinomial, MLwiN automatically generates a number of new variables. To view these variables select “Data Manipulation” in the top menu bar, click “view or edit data”, click on “view” and select “resp”, “resp_indicator”, and “id_ind_long”. Click “Ok”.

Data					
goto line	1	view	Help	Font	
	resp(22372)	resp_indicator(2	id_ind_long(223i	cons.(>=alarm)(2	cons.(>=positive)
487	0	(>=alarm)	276	1	0
488	0	(>=positive)	276	0	1
489	0	(>=alarm)	277	1	0
490	0	(>=positive)	277	0	1
491	1	(>=alarm)	278	1	0
492	1	(>=positive)	278	0	1
493	0	(>=alarm)	279	1	0
494	0	(>=positive)	279	0	1
495	0	(>=alarm)	280	1	0
496	0	(>=positive)	280	0	1
497	0	(>=alarm)	281	1	0
498	0	(>=positive)	281	0	1
499	0	(>=alarm)	282	1	0
500	0	(>=positive)	282	0	1
501	1	(>=alarm)	283	1	0
502	0	(>=positive)	283	0	1

A new response variable has been made (*resp*), which indicates for each individual for each response category, whether this category was the given response (1) or not (0). In the present case we have three categories. Note however, that the data can be fully described with two independent categories. If someone was neither in the category “alarm” nor in “positive”, we know for sure that he is in “safe”. Therefore the variable “*resp*” contains for each individual two values indicating whether or not the response has been “alarm” or “positive”, respectively.

The variable “*resp_indicator*” indicates which response the value in the variable “*resp*” applies to. There are two possible values: (>=alarm) or (>=positive). These labels including the “greater-than or equal” relation are automatically generated by MLwiN. They are due to the fact that the estimated model is an **ordered category** model in which it is assumed that “positive” is greater than “alarm” and “alarm” is greater than “safe”. Therefore, an individual categorized as (>=positive) is automatically also categorized as (>=alarm). The reader can verify this by checking the values in “*resp*”. The majority of the individuals have two zeros, indicating that they were in the safe-category. There are some individuals having a 1 at (>=alarm) and a zero at (>=positive), who were in the alarm category. Note however, that all individuals with a 1 at (>=positive) also have a 1 at (>=alarm). It is easy to understand this when re-translating the category titles to the underlying BAC values: Everybody who is in the positive category (i.e. his BAC was ≥ 0.08) also has a BAC ≥ 0.05 (which defines the alarm category).

A multinomial model has an intercept for each independent category (i.e. one for (>=alarm) and one for (>=positive). To include these intercepts into the model

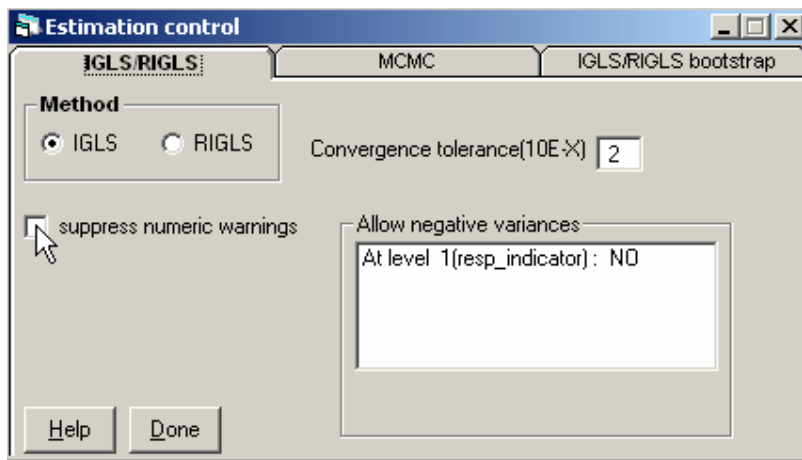
- Click “Add Term” and select “Cons” from the variable drop down list
- Click “add Separate coefficients”

- Click “Done”

Estimation

- Click on “Nonlinear” at the bottom of the Equations window
- Select “Use defaults” or check “Multinomial”, “1st order”-Linearisation, and “MQL”
- Click “Done”

A couple of remarks about the estimation procedure are necessary. At the time of writing this manual, the estimation of multinomial models proved to be problematic. As described in the previous section on binomial data (2.3.2), it is principally advised to use 2nd order PQL estimates for good unbiased results. However, for the multinomial model these estimates did not converge, leaving us with 1st order MQL estimates, which are more stable but downwardly biased. It was decided to include this chapter nevertheless, because due to the rapid development in estimation techniques (see also section 2.7 in the Methodology report) these problems might be solved soon. Because the estimation is problematic at the moment, MLwiN gives a lot of numerical warnings while running. As advised in the MLwiN manual (Rasbash et al., 2004), we suggest to suppress them by clicking “Estimation control” and checking “suppress numeric warnings”.



- Then press “Start” the estimate the model

Results and interpretation

Equations

$$\text{resp}_{ij} \sim \text{Ordered Multinomial}(\text{cons}_j, \pi_{ij})$$

$$\gamma_{3j} = \pi_{3j}, \gamma_{2j} = \pi_{3j} + \pi_{2j}, \gamma_{1j} = 1$$

$$\text{logit}(\gamma_{2j}) = -3.488(0.056)\text{cons.}(=\text{alarm})_{ij}$$

$$\text{logit}(\gamma_{3j}) = -3.909(0.068)\text{cons.}(=\text{positive})_{ij}$$

$$\text{cov}(y_{sj}, y_{tj}) = \gamma_{sj}(1 - \gamma_{tj}) / \text{cons}_j \quad s \leq t$$

The only thing this empty model does is to estimate an intercept for each of the categories. As indicated in the methodology report, these intercepts can be translated into probabilities by filling them into equation 2.3.4 of the methodology report.

$$\gamma_{ij} = \frac{1}{1 + \exp(-(\text{parameter}))} \quad (2.3.4)$$

This way we receive 0.02 as probability to be in category “positive” (i.e. BAC ≥ 0.08) and 0.03 to be in category “alarm” or “positive” (i.e. BAC ≥ 0.05). Note that no variance is estimated, as in the multinomial distribution the variance is determined exclusively by the probability of belonging to a particular category (see section 2.3.3 in the methodology report).

Each component of the model equations is double indexed by ij , indicating that we have a two-level structure. This might be confusing, because conceptually, this model has only one level (the individuals). Structurally, the individuals are the second level and therefore indexed j here. The first level, indexed by i , is defined by the variable “resp_indicator” that indicates which of two categories ($=\text{alarm}$) or ($=\text{positive}$)) the response variable refers to.

2.3.3.2. Ordered proportional odds: empty two-level model

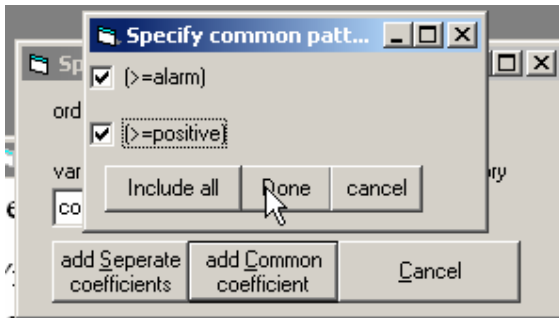
In the previous section on binary responses (2.3.2), it had been demonstrated that the probability of drink driving varied across measurement locations. A model that includes a location-level is conceptually a two-level model. To implement a two-level multinomial model, however, we will build a structural three-level model (the first level being reserved for the response indicator).

Model formulation

Because the three response categories are assumed to have one underlying variable (in this case the BAC) it is assumed that random variation across the locations is the same for each response category. To create this structure, it is necessary to include another constant into the model with a common coefficient

(as opposed to the constant that is already included that has separate coefficients for (\geq alarm) and (\geq positive)).

- Click on “Add Term” at the bottom of the Equations window
- Select “cons” from the variable drop down list
- Click on “Add common coefficient”
- Select “Include all”
- Click “Done”



This common coefficient must be allowed to vary randomly across locations. However, because we already have intercepts we do not want the newly included constant to function as a fixed factor. Therefore

- Click on the constant just added (cons.23)
- Check “k(id_loc_long)”
- Uncheck “Fixed Parameter”
- Click “Done”



Press “Start to estimate the model.

Results

Equations

$$\text{resp}_{ijk} \sim \text{Ordered Multinomial}(\text{cons}_{jk}, \pi_{ijk})$$

$$\gamma_{3jk} = \pi_{3jk}, \gamma_{2jk} = \pi_{3jk} + \pi_{2jk}, \gamma_{1jk} = 1$$

$$\text{logit}(\gamma_{2jk}) = -3.367(0.075)\text{cons.}(=\text{alarm})_{ijk} + h_{jk}$$

$$\text{logit}(\gamma_{3jk}) = -3.789(0.083)\text{cons.}(=\text{positive})_{ijk} + h_{jk}$$

$$h_{jk} = v_{3k}\text{cons.}23$$

$$\begin{bmatrix} v_{3k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.875(0.180) \end{bmatrix}$$

$$\text{cov}(y_{sjk}, y_{tjk}) = \gamma_{sjk}(1 - \gamma_{tjk})/\text{cons}_{jk} \quad s \leq t$$

As can be seen in the triple index ijk , structurally this is a three level model. Conceptually however, this is a two-level model. The second conceptual level is that of the locations. It is implemented by the joint intercept h_{jk} . This joint intercept has been defined as random factor only, but not as fixed factor. Consequently there is no mean estimated for it, only its variance Ω_v , which indicates how the probability to be either in category “alarm” or in category “positive” varies across locations. Although the present estimations should be interpreted with caution, we can note that the variation Ω_v is quite large compared to its standard error, making it very likely that there is substantial variation across locations in the probability to have a BAC above .05 or above .08. This is also in line with the results from the binary model in section 2.3.2.

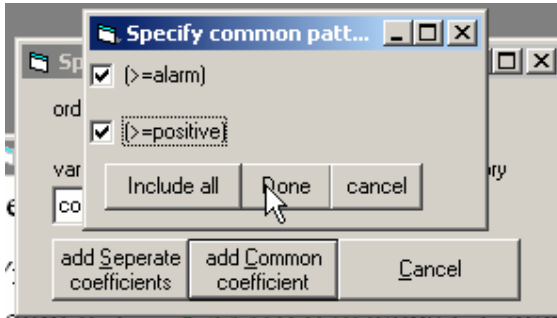
2.3.3.3. Ordered proportional odds: the two-level model with predictors

The next step is to include predictors into the multinomial model. In the ordered model, predictors are usually assumed to apply to all categories in the same way (i.e., if a particular variables is thought to affect the probability to have a BAC above .05 it is also thought to affect the probability to have a BAC above .08). Therefore only one slope is estimated for each predictor. We will take up the variables “gender” and “age”.

Model formulation

- Click on “Add Term” at the bottom of the Equations window
- Select “gender” from the variable drop down list
- Select “male” as reference category

- Click on “Add common coefficient”
- Select “Include all”
- Click “Done”



Results and interpretation

Equations

$$\text{resp}_{ijk} \sim \text{Ordered Multinomial}(\text{cons}_{jk}, \pi_{ijk})$$

$$\gamma_{3jk} = \pi_{3jk}, \gamma_{2jk} = \pi_{3jk} + \pi_{2jk}, \gamma_{1jk} = 1$$

$$\text{logit}(\gamma_{2jk}) = -3.068(0.078)\text{cons.}(\geq \text{alarm})_{ijk} + h_{jk}$$

$$\text{logit}(\gamma_{3jk}) = -3.492(0.086)\text{cons.}(\geq \text{positive})_{ijk} + h_{jk}$$

$$h_{jk} = -1.607(0.227)\text{female.23}_{jk} + v_{3k}\text{cons.23}$$

$$\begin{bmatrix} v_{3k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.890(0.182) \end{bmatrix}$$

$$\text{cov}(y_{sjk}, y_{tjk}) = \gamma_{sjk}(1 - \gamma_{tjk})/\text{cons}_{jk} \quad s \leq t$$

The common part h_{jk} is now defined by the random variation over the locations ($v_{3k}\text{cons.23}$) but also by the effect of gender. The coefficient for female.23_{jk} is negative, indicating that women, as compared to men who form the base line category, have a lower probability to be in categories “alarm” or “positive”. Accordingly the intercepts $\text{cons.}(\geq \text{alarm})_{ijk}$ and $\text{cons.}(\geq \text{positive})_{ijk}$ have increased (i.e., they have lower negative values) as compared to the model without predictors. These intercepts now represent the probabilities for men instead of representing the probabilities for the whole group.

Now include the predictor “age”. This is a categorical variable representing four age-categories (16-25, 26-39, 40-54, 55+).

- Click on “Add Term” at the bottom of the Equations window

- Select “age” from the variable drop down list
- Select “age 16-25” as reference category
- Click on “Add common coefficient”
- Select “Include all”
- Click “Done”
- Click “Start” to estimate the model

Results and interpretation

Equations

$$\text{resp}_{ijk} \sim \text{Ordered Multinomial}(\text{cons}_{jk}, \pi_{ijk})$$

$$\gamma_{3jk} = \pi_{3jk}, \gamma_{2jk} = \pi_{3jk} + \pi_{2jk}, \gamma_{1jk} = 1$$

$$\text{logit}(\gamma_{2jk}) = -3.762(0.226)\text{cons.}(=\text{alarm})_{ijk} + h_{jk}$$

$$\text{logit}(\gamma_{3jk}) = -4.187(0.229)\text{cons.}(=\text{positive})_{ijk} + h_{jk}$$

$$h_{jk} = -1.610(0.226)\text{female.}23_{jk} + 0.513(0.254)\text{age } 26-39.23_{jk} + 1.155(0.241)\text{age } 40-54.23_{jk} + 0.527(0.278)\text{age } 55+.23_{jk} + v_{3jk} \text{cons.}23$$

$$\begin{bmatrix} v_{3jk} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.933(0.185) \end{bmatrix}$$

$$\text{cov}(y_{sjk}, y_{tjk}) = \gamma_{sjk}(1 - \gamma_{tjk}) / \text{cons}_{jk} \quad s \leq t$$

The coefficients for the various age-categories are positive, indicating that all age-categories listed have a higher probability of being in “alarm” or “positive” than the youngest (16-25) that forms the reference category. The highest coefficient is estimated for drivers aged 40-54, indicating that this group is especially at risk of drink driving.

2.3.3.4. Unordered categories: the two-level model with predictors

In the present example we clearly have a variable (the BAC) underlying the response categories, making the ordered proportional odds model likely to be the appropriate. In this way, it was assumed that both probabilities (of having a BAC >.05 and of having a BAC >.08) vary in the same way across locations and that the effects of gender and age on both probabilities are the same. However, even if both probabilities are related to one underlying variable, they might nevertheless show different random or fixed effects. This can be investigated in an unordered category model, which assumes no relation between the different categories to start with. To switch from an ordered to an unordered multinomial model in MLwiN, one has to build a new model from the start.

- Click on “Clear” at the bottom of the Equations window
- Click on the “y” and
 - Select “Breathtest” from the drop-down menu as dependent variable
 - Select “2-ij” from the N-levels drop-down menu

- Select “ID_loc” from the level 2(j) drop-down menu
 - Select “ID_ind” from the level 1(i) drop-down menu
 - Click “Done”
- Click on the N in the Distribution statement for Breathtest and
 - Check “Multinomial”
 - In the appearing window, leave “logit” checked
 - Select “Unorderd”
 - Leave “safe” as reference category
 - Click “Done”
- Click “Add Term” and select “Cons” from the variable drop down list
- Click “add Separate coefficients”
- Click “Done”

To include the location-level into the model, one has to let the intercept vary randomly across locations. In the ordered model, this was done for a joint intercept. In the unordered model, we simply let the intercepts that already define the categories vary across locations.

- Click on “cons.alarm_{ij}”
- Check “k(id_loc_long)”
- Click “Done”
- Do the same with “cons.positive_{ij}”

In the unordered model, predictors are not assumed to have the same effect on all categories. Therefore *separate coefficients* have to be estimated for each category.

- Click on “Add Term” at the bottom of the Equations window
- Select “gender” from the variable drop down list
- Select “male” as reference category
- Click on “Add separate coefficients”
- Click “Done”
- Repeat the procedure to include predictor “age”
 - Select “age16-25” as reference category
- Click “Start” to estimate the model

Results and interpretation

Equations

$$\text{resp}_{ijk} \sim \text{Multinomial}(\text{cons}_{ijk}, \pi_{ijk})$$

$$\log(\pi_{2jk} / \pi_{1jk}) = \beta_{0k} \text{cons.alarm}_{ijk} + -1.053(0.249) \text{female.alarm}_{ijk} + 0.774(0.344) \text{age 26-39.alarm}_{ijk} + 1.019(0.339) \text{age 40-54.alarm}_{ijk} + 0.712(0.375) \text{age 55+.alarm}_{ijk}$$

$$\beta_{0k} = -4.927(0.314) + v_{0k}$$

$$\log(\pi_{3jk} / \pi_{1jk}) = \beta_{1k} \text{cons.positive}_{ijk} + -1.608(0.220) \text{female.positive}_{ijk} + 0.522(0.251) \text{age 26-39.positive}_{ijk} + 1.185(0.238) \text{age 40-54.positive}_{ijk} + 0.590(0.271) \text{age 55+.positive}_{ijk}$$

$$\beta_{1k} = -4.158(0.226) + v_{1k}$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.651(0.236) & \\ 1.056(0.160) & 0.962(0.180) \end{bmatrix}$$

$$\text{cov}(y_{sjk}, y_{ijk}) = -\pi_{sjk} \pi_{ijk} / \text{cons}_{ijk} : s \neq i; \quad \pi_{sjk} (1 - \pi_{ijk}) / \text{cons}_{ijk} : s = i;$$

In the unordered category model for three categories, two contrasts function as dependent variables: The contrast between “safe” (the reference category) and “alarm”, and the contrast between “safe” and “positive”. The values that are predicted by the model are the log-odds of these contrasts (“alarm-safe”: $\log(\pi_{2jk} / \pi_{1jk})$ and “positive-safe”: $\log(\pi_{3jk} / \pi_{1jk})$). For each of these log odds, there is a full prediction model including fixed factors (age and gender) and a random factor (the location effects v_{0k} and v_{1k}). The variation across locations is given by Ω_v which is now a matrix containing the variation of the contrast “alarm-safe” across locations (upper left), the variation of the contrast “positive-safe” across locations (lower right), and the covariance (lower left corner). We can see that there is substantial covariance, indicating that for locations with a large contrast “alarm-safe”, the contrast “positive-safe” is also large. The assumption that the probabilities to be in “alarm” and to be in “positive” vary across locations in the same way, lay at the basis of assuming a common random factor in the ordered model. The large covariance in the present model supports this assumption.

The coefficients for gender and age show the same pattern for both contrasts: The negative coefficients for “female” indicate that men have a higher probability to be in “alarm” or “positive” respectively as compared to “safe”. The positive coefficients for the age categories listed indicate that drivers in those age categories have higher probabilities to be in “alarm” or “positive” than the youngest drivers. And for both contrasts it is the category of 40-54 year olds that received the highest coefficient.

Although both sets of coefficients show exactly the same pattern, one might note that the coefficients for the contrast “positive-safe” all have somewhat higher values than those for the contrast “alarm-safe”. This would indicate that being a man between 40 and 54 years is even a better predictor for being in category “positive” (i.e. having a BAC of .08 or higher) than for being in category “alarm” (i.e., $.05 < \text{BAC} < .08$). It is interesting to test whether this difference is significant. To do so for the two coefficients for “female”:

- Click on “Model” in the top-menu bar
- Select “Intervals and tests”
- Check the radio-button in front of “fixed”
- Type “1” behind “female.alarm”
- Type “-1” behind “female.positive”
- Click on “Calc”

The resulting Chi-square value of 2.792 with 2 degrees of freedom corresponds to a probability of .095. We can conclude that the effect of being a woman does not differ significantly between the contrasts “alarm-safe” and “positive-safe”. As for the other coefficients the differences between both contrasts are smaller than the one for the gender-coefficients, we can conclude that there is no systematic difference between the effects of age and gender on the categories “alarm” and “positive” respectively. Again, the results of the unordered model support the assumptions at the basis of the ordered proportional odds model.

2.3.3.5. Conclusion

It has been shown how categorical responses can be analysed in a multilevel multinomial model. Two different versions were presented. (1) The ordered proportional odds model is based on the assumptions that the response categories result from an underlying continuous variable and that fixed and random effects therefore have the same shape for outcome. (2) The unordered categories model does not assume any systematic relation between the different outcomes. Independent models are estimated for each outcome (in contrast to the reference category). It was shown that even for categorical data that are expected to have a common underlying variable it can be interesting to analyse them in an unordered categorical model, as comparing the independent prediction models for each category can indicate whether an ordered proportional odds model is appropriate.

2.3.4 Counts

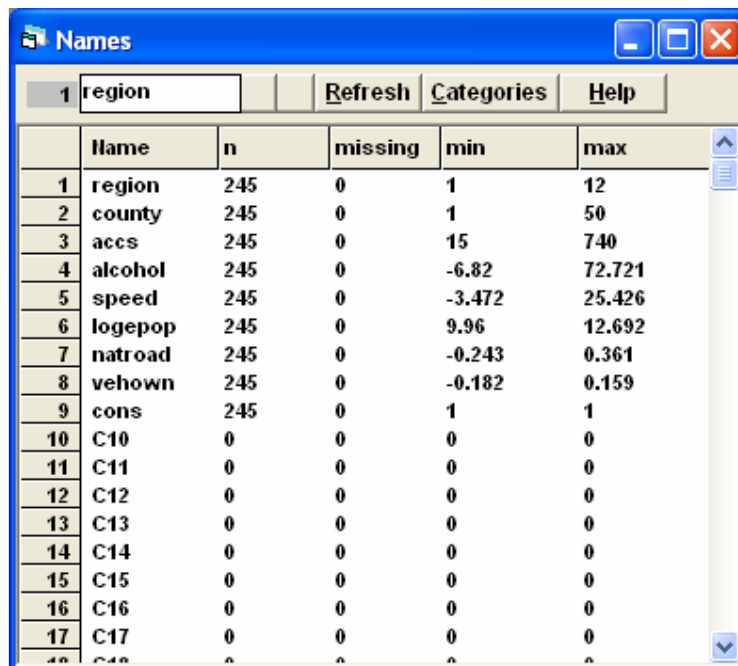
George Yannis, Eleonora Papadimitriou and Constantinos Antoniou (NTUA)

In this section, an example for fitting Poisson multilevel models is presented using the MLwinN 2.01 software. The example concerns an investigation of the regional effect of speeding and drinking-and-driving enforcement on the number of road accidents in Greece. The theoretical background, models fit and results were discussed in section 2.3.4 of the Methodology Report.

The dataset includes accidents data, police enforcement data as well as other demographic data for the 49 counties and 12 regions of Greece for the period 1998-2002. More specifically, the variables and values used are summarized in the following Table:

Region	1-12 regions of Greece
County	1-49 counties of Greece
Accs	The number of accidents of each county
alcohol	The number of alcohol controls of each county (1000 alcohol controls)
Speed	The number of speed infringements of each county (1000 speed infringements)
logepop	The natural logarithm of the population of each county
Natroad	The proportion of National Roads of the road network of each county
Vehown	The vehicle ownership of each county (vehicles per 1000 inhabitants)
Cons	The constant term (1)

Open the dataset PoissonManualData3.ws using the Open Worksheet option from the Files menu. Opening the Names window from the Data Manipulation menu gives the following:



The screenshot shows the 'Names' window in MLwinN software. It displays a list of variables with their names, counts (n), missing values, minimum values, and maximum values. The variables are listed in a table with columns: Name, n, missing, min, and max. The variables are: region, county, accs, alcohol, speed, logepop, natroad, vehown, cons, C10, C11, C12, C13, C14, C15, C16, C17, and C18. The 'region' variable has 12 unique values, while the others have 49 unique values. The 'cons' variable has 1 unique value. The 'C' variables have 0 unique values.

	Name	n	missing	min	max
1	region	245	0	1	12
2	county	245	0	1	50
3	accs	245	0	15	740
4	alcohol	245	0	-6.82	72.721
5	speed	245	0	-3.472	25.426
6	logepop	245	0	9.96	12.692
7	natroad	245	0	-0.243	0.361
8	vehown	245	0	-0.182	0.159
9	cons	245	0	1	1
10	C10	0	0	0	0
11	C11	0	0	0	0
12	C12	0	0	0	0
13	C13	0	0	0	0
14	C14	0	0	0	0
15	C15	0	0	0	0
16	C16	0	0	0	0
17	C17	0	0	0	0
18	C18	0	0	0	0

The response variable (accs) in this dataset is the number (counts) of road accidents in various counties of Greece during the period 1998 to 2002. The data were collected in 49 counties¹; these counties are included into 12 regions, giving two levels of data. As explained in the Methodology Report (section 2.3.4), count data are constrained to be non-negative, therefore we would prefer to model the logarithms of the counts. We will therefore fit a Poisson model to the count data using a log link function, which can be specified through the software.

In order to work with the accident rates rather than the accident counts, we use an additional parameter known as an offset. The variable logpop reflects the expected number of accidents in each county, which is considered to be proportional to the population of each county, and will be used to create an offset variable. It should be noted that, as a log link function is used for the response variable, the offset term should also be transformed accordingly. In this dataset no transformation is required, as the variable logpop already corresponds to the natural logarithm of the population of each county. However, the transformation could be carried out using the Command Interface option from the Data Manipulation menu.

The variables alcohol, speed, natroad and vehown also concern each county and shall be used as explanatory variables. These variables have been centered around their mean, as recommended by the MLwiN users manual, in order to avoid computational instabilities.

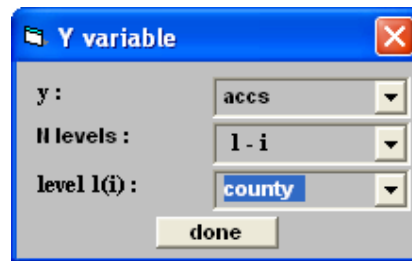
We will start by fitting a simple (single level) model and then proceed to multilevel structure.

2.3.4.1. A single-level Poisson model

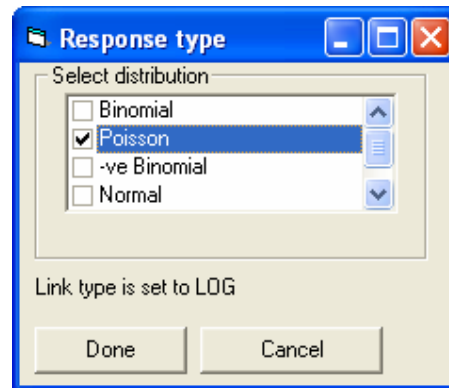
In order to specify a simple (single level) model in MLwiN:

- Open the Equations window from the Model menu and click on y
- In the Y variable window, select accs from the y: drop down list, select i-1 from the N levels: drop down list and county from the level 1(i) drop down list, and click Done.

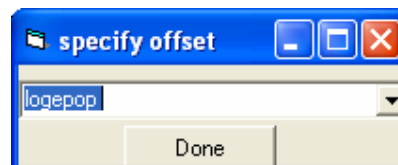
¹ The Athens and Thessalonica metropolitan areas, where a disproportionally high number of accidents and police controls are observed, were not included in the dataset.



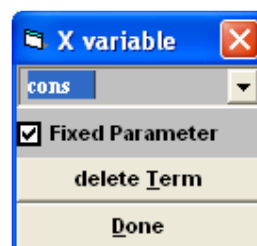
- Click on the $N(\Omega X, B)$ that appears on the first line of the Equations window, select Poisson from the available distributions and click Done.



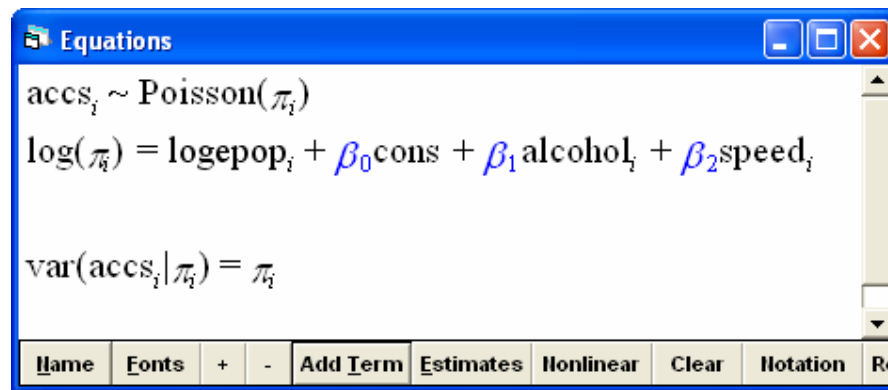
- Click on the (π_i) that appears on the second line of the Equations window, select logepop as offset term and click Done.



- Click on the Add Term button of the Equations window and add cons, alcohol and speed to the model. By clicking on each of the terms in the Equations window, we can see that these are entered by default as fixed parameters.

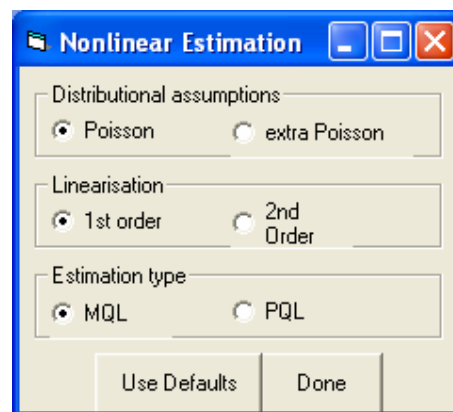


- Click on the Estimates button of the Equations window and the parameters to be estimated will be highlighted in blue. The Equations window will now look as follows:

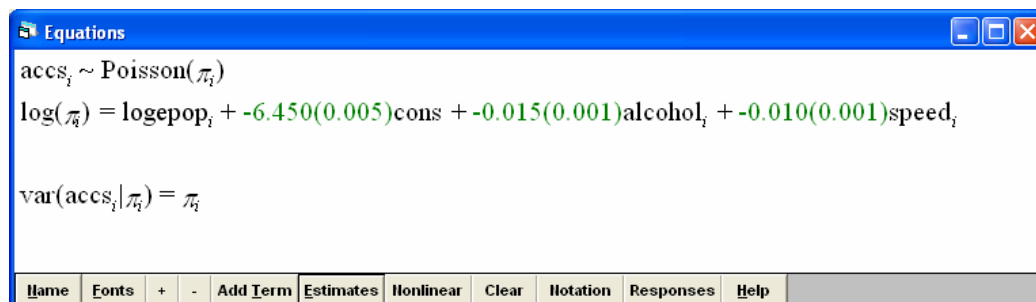


It should be noted that, the last line in the Equations window reflects the Poisson assumption that the variance of the response variable is equal to the mean π_i .

- In order to set the estimation procedure, click on the Nonlinear button of the Equations window select distributional assumptions Poisson, linearization 1st order and estimation type MQL (Marginal Quasi Likelihood) and click Done.



- In order to run the model, click Start on the toolbar of the main window. We then obtain the following results:

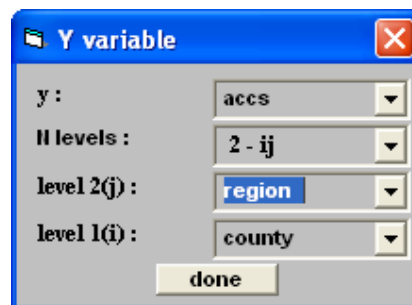


These results are intuitive (i.e. an increase in speeding and drinking-and-driving controls results in a reduction of road accidents). In the next section, we will see how this effect may vary when adding more structure to the data.

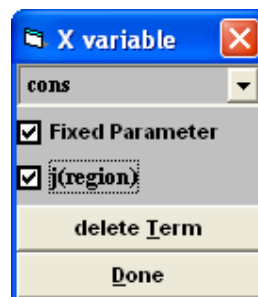
2.3.4.2. A two-level Poisson model

We will now fit a two-level model, in order to investigate the regional variation of the effect of enforcement on the number of road accidents. We shall start with the random intercept model:

- Remove the terms alcohol and speed from the model.
- Click on accs in the Equations window, select j-2 from the N levels: drop down list and region from the level 2(j) drop down list, and click Done.

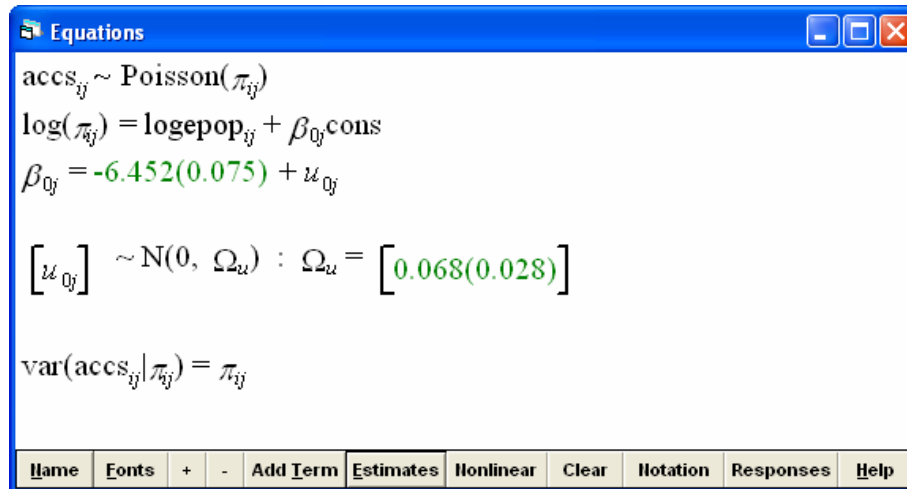


- Click on the variable cons in the Equations window and set cons to be random at the j(region) level.



As there are only 12 regions at the higher level, it is recommended to use the RIGLS estimation method, which provides less biased estimates of the variance than the IGLS when there is limited number of higher level units.

- Select RIGLS from the Estimation menu of the main window
- Click the Start button on the toolbar of the main window to run the model. The results are as follows:



Equations

$$\text{accs}_{ij} \sim \text{Poisson}(\pi_{ij})$$

$$\log(\pi_{ij}) = \text{logpop}_{ij} + \beta_{0j} \text{cons}$$

$$\beta_{0j} = -6.452(0.075) + u_{0j}$$

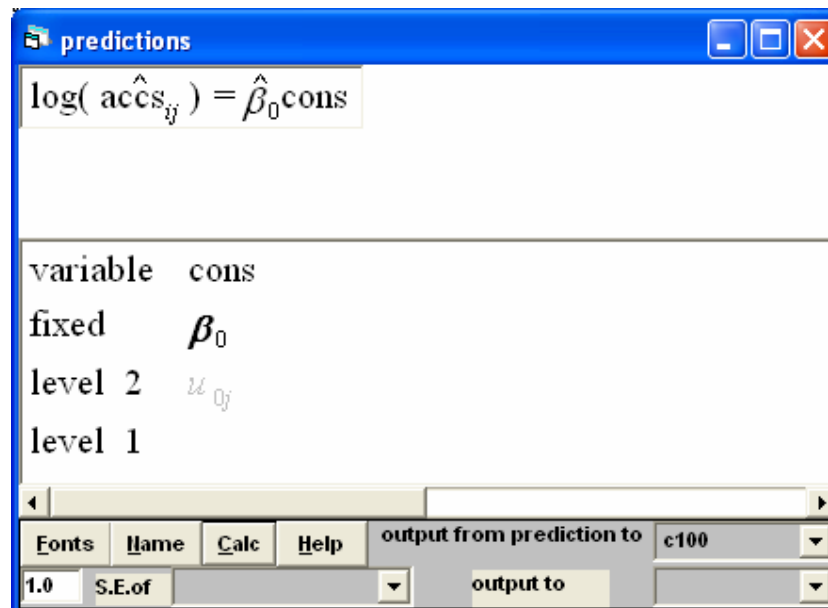
$$[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = [0.068(0.028)]$$

$$\text{var}(\text{accs}_{ij} | \pi_{ij}) = \pi_{ij}$$

Home Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

In order to graphically represent the average intercept and the random intercepts:

- Open the Predictions window from the Model menu. The elements of the model are arranged in two columns. These columns are initially grayed out. We will build a prediction equation in the top of the window, by selecting the elements we want from the bottom section.
- Click on β_0 in order to select only the fixed part of the model.
- Select c100 from the Output from prediction drop down list
- Click the Calc button



predictions

$$\log(\hat{\text{accs}}_{ij}) = \hat{\beta}_0 \text{cons}$$

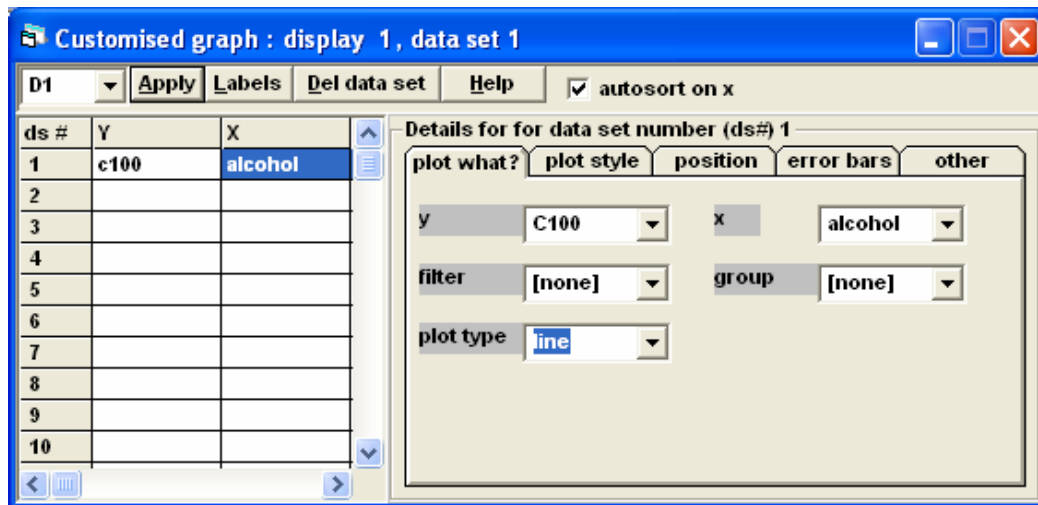
variable	cons
fixed	β_0
level 2	u_{0j}
level 1	

Fonts Home Calc Help output from prediction to c100

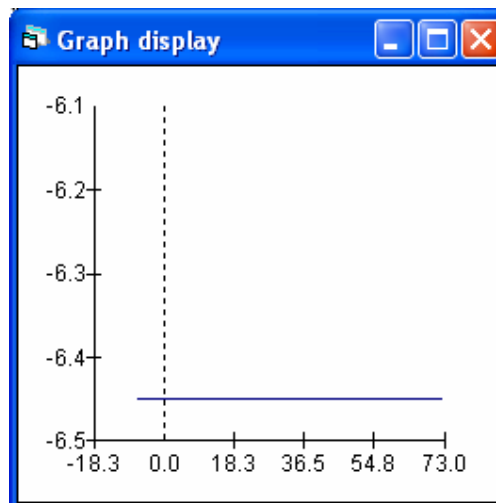
1.0 S.E.of output to

- Open the Customized Graphs window from the Graphs menu

- In the plot what? tab, select c100 from the y drop down list, alcohol in the x drop down list and line in the plot type drop down list.

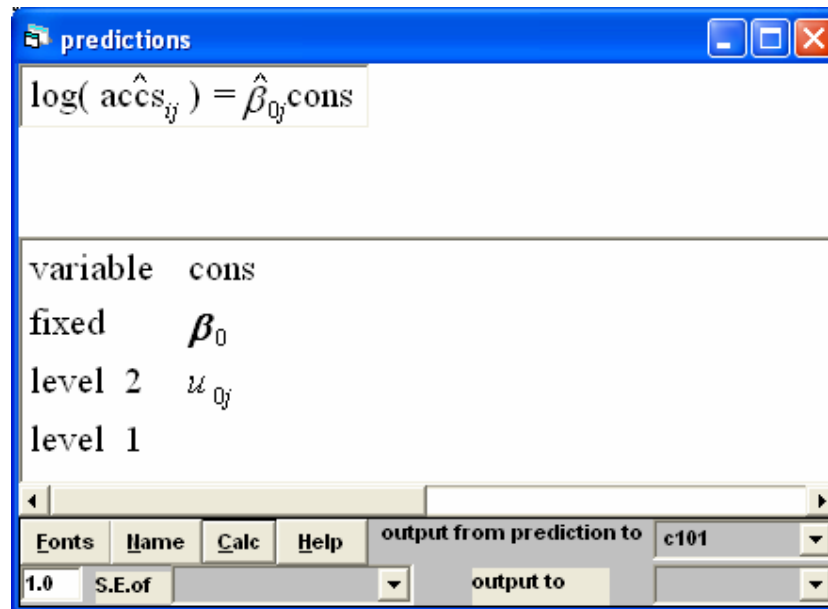


- Click the Apply button to obtain a graph of the average (fixed) intercept

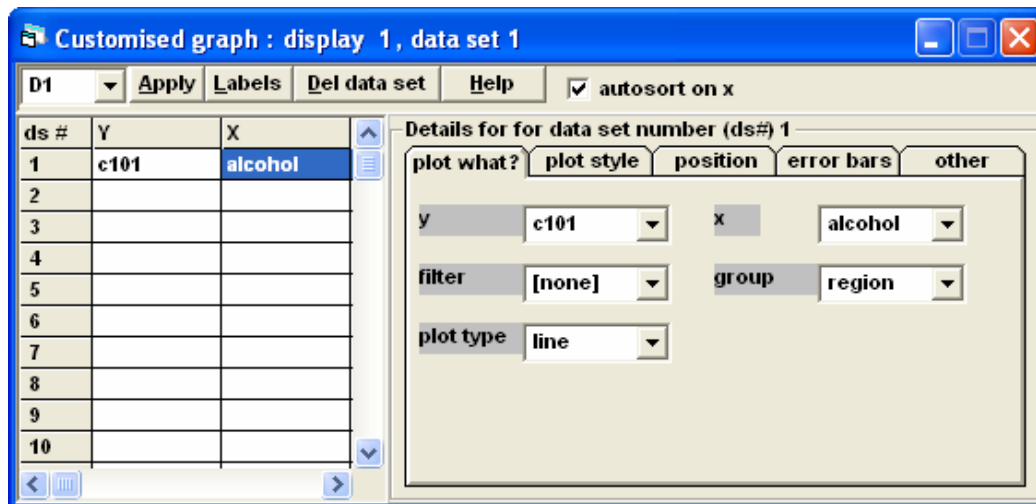


Accordingly, in order to graphically represent the average intercept and the random intercepts:

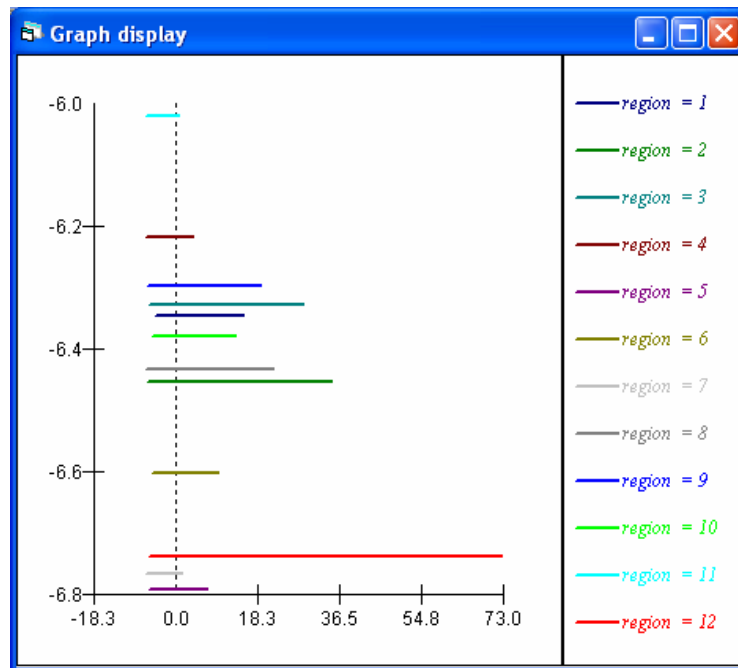
- Open the Predictions window from the Model menu.
- Click on β_0 and u_{0j} in order to select both the fixed and the random part of the model.
- Select c101 from the Output from prediction drop down list
- Click the Calc button



- Open the Customized Graphs window from the Graphs menu
- In the plot what? tab, select c101 from the y drop down list, alcohol in the x drop down list, line in the plot type drop down list and region in the group drop down list.
- In the plot style tab, select 16 rotate from the color drop down list.
- In the other tab, select group code.



- Click the Apply button to obtain a graph of the random intercepts



We will now add a random slope to the model:

- Click on the Add Term button of the Equations window and add alcohol to the model.
- Click on the variable alcohol in the Equations window and set alcohol to be random at the $j(\text{region})$ level.
- Click on the Start button to run the model

The Equations window displays the following model specification:

$$\text{accs}_{ij} \sim \text{Poisson}(\pi_{ij})$$

$$\log(\pi_{ij}) = \text{logpop}_{ij} + \beta_{0j} \text{cons} + \beta_{1j} \text{alcohol}_{ij}$$

$$\beta_{0j} = -6.553(0.085) + u_{0j}$$

$$\beta_{1j} = -0.045(0.015) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.087(0.036) & \\ 0.009(0.005) & 0.003(0.001) \end{bmatrix}$$

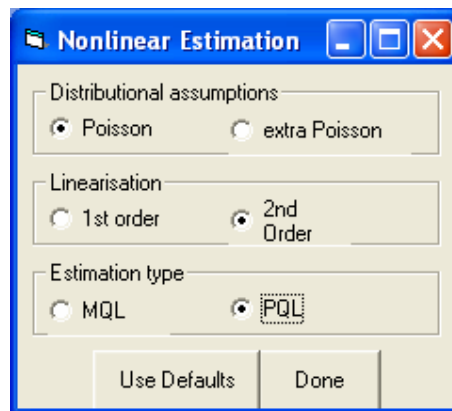
$$\text{var}(\text{accs}_{ij} | \pi_{ij}) = \pi_{ij}$$

The bottom of the window contains a toolbar with buttons: Name, Fonts, +, -, Add Term, Estimates, Nonlinear, Clear, Notation, Responses, and Help.

The results reveal a significant variance in the effect of alcohol controls on the number of accidents.

However, it has been proved that the 1st order MQL estimation method tends to overestimate some of the variance in Poisson multilevel models. We will therefore switch to the 2nd order PQL (Penalized "Predictive" Quasi Likelihood), which is more accurate.

- Click on the Nonlinear button of the Equations and select linearization 2nd order and estimation type PQL and click Done.



- Click on the More on the toolbar of the main window button to run the model.

Equations

$$\text{accs}_{ij} \sim \text{Poisson}(\pi_{ij})$$

$$\log(\pi_{ij}) = \text{logepop}_{ij} + \beta_{0j} \text{cons} + \beta_{1j} \text{alcohol}_{ij}$$

$$\beta_{0j} = -6.642(0.108) + u_{0j}$$

$$\beta_{1j} = -0.059(0.014) + u_{1j}$$

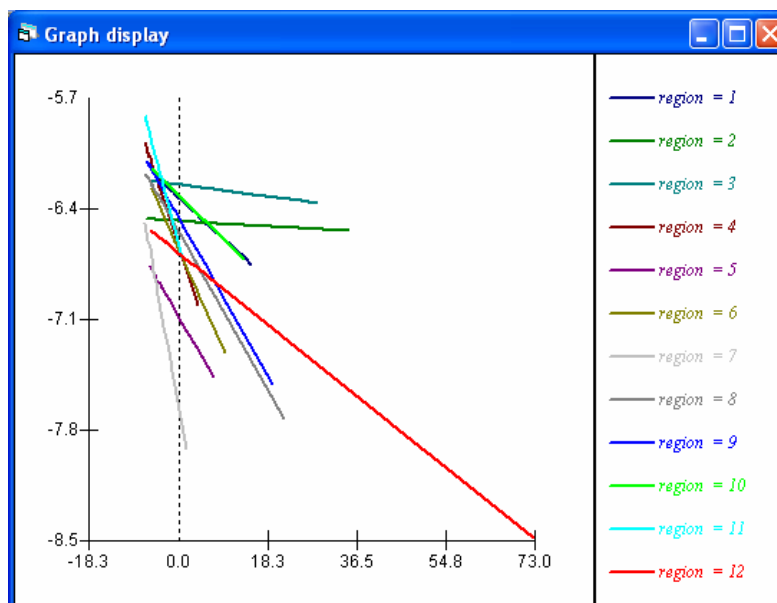
$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.139(0.057) & \\ 0.014(0.007) & 0.002(0.001) \end{bmatrix}$$

$$\text{var}(\text{accs}_{ij} | \pi_{ij}) = \pi_{ij}$$

Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

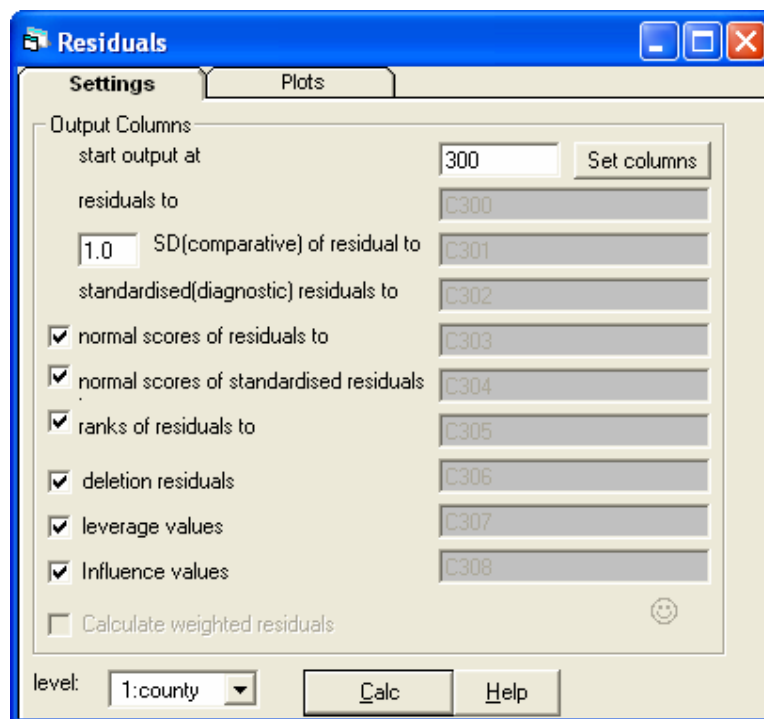
All fixed and random effects are statistically significant.

In order to graphically represent the random slopes, we follow the process described above for the random intercepts. In this case, we should select β_0 , β_1 , u_{0j} and u_{1j} in the Predictions window of the Model menu and output from prediction to another column. We then obtain the following:

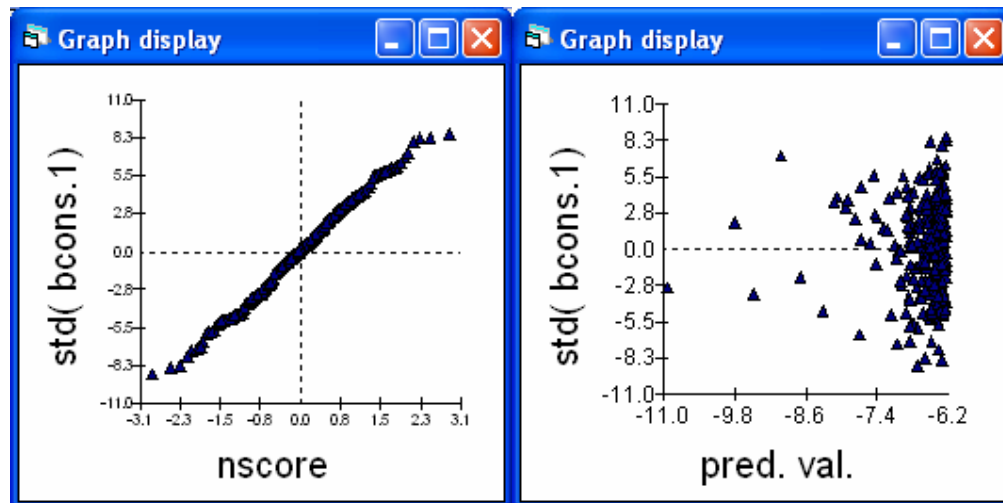


In order to explore the residuals of the model, starting by the level-1 residuals:

- Open the Residuals window from the Model menu
- In the Settings tab, write 300 (or any other appropriate column number) in the start output box and click on Set columns. The boxes beneath this button are then filled in gray with the column numbers that will be used for residuals calculations. Additionally, select 1-county from the level: drop down list.
- Click on the Calc button in the Residuals window.



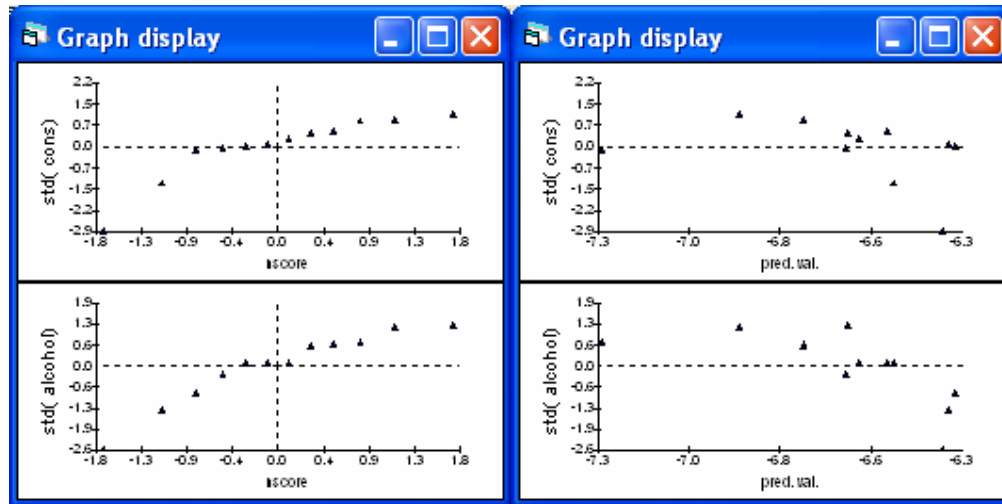
- In the Plots tab, select the first option standardized residual * normal scores
- Click on the Apply button.
- Then in the Plots tab, select the fourth option standardized residual * fixed part prediction
- Click on the Apply button.



We can see that the level-1 residuals present no significant deviation from the Normal distribution. Moreover, the residuals are independent from the predicted values.

In order to explore the level-2 residuals, we will repeat the process described above, except that we will set start output at a different column number and select 2-region in the level: drop down list, in the Settings tab of the Residuals window.

We will then obtain the following results:



These results are less satisfactory compared to the level-1 results, as a consequence of the limited number of higher level units.

Accordingly, the effect of speed enforcement on the number of accidents can be separately examined, by removing the variable alcohol from the model and adding the variable speed, also allowing it to randomly vary between regions. The multilevel model fitted should be as follows:

The 'Equations' window displays the following model:

$$\text{accs}_{ij} \sim \text{Poisson}(\pi_{ij})$$

$$\log(\pi_{ij}) = \text{logpop}_{ij} + \beta_{0j}\text{cons} + \beta_{1j}\text{speed}_{ij}$$

$$\beta_{0j} = -6.689(0.110) + u_{0j}$$

$$\beta_{1j} = -0.131(0.041) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.142(0.058) \\ 0.046(0.020) \quad 0.020(0.008) \end{bmatrix}$$

$$\text{var}(\text{accs}_{ij} | \pi_{ij}) = \pi_{ij}$$

The window also includes a toolbar with buttons: Name, Fonts, +, -, Add Term, Estimates, Nonlinear, Clear, Notation, Responses, and Help.

All fixed and random effects are statistically significant.

In order to examine the combined effect of speeding and drinking-and-driving enforcement, we will add alcohol to the model, allowing it to vary among regions. We will obtain the following results:

Equations

$$\text{accs}_{ij} \sim \text{Poisson}(\pi_{ij})$$

$$\log(\pi_{ij}) = \text{logpop}_{ij} + \beta_{0j} \text{cons} + \beta_{1j} \text{speed}_{ij} + \beta_{2j} \text{alcohol}_{ij}$$

$$\beta_{0j} = -6.654(0.101) + u_{0j}$$

$$\beta_{1j} = -0.058(0.023) + u_{1j}$$

$$\beta_{2j} = -0.037(0.010) + u_{2j}$$

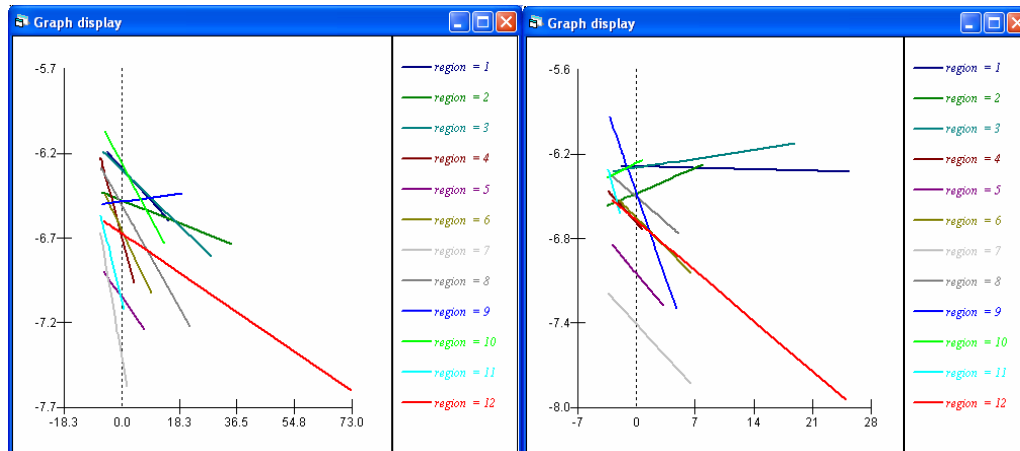
$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.119(0.050) & & \\ 0.013(0.009) & 0.006(0.003) & \\ 0.008(0.004) & 0.000(0.001) & 0.001(0.000) \end{bmatrix}$$

$$\text{var}(\text{accs}_{ij} | \pi_{ij}) = \pi_{ij}$$

File Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

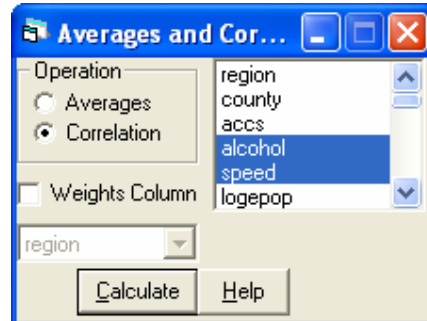
In this case, all fixed effects are significant; however, the covariances related to the number of speed infringements are non significant. This is quite surprising, when considering that both effects were significant when examined separately. It is therefore indicated that there is some bias in the model. This is also identified when plotting the predicted values with alcohol and speed.

In order to plot the effects of alcohol, follow the process described above, but select in the Predictions window only the β_{0j} , β_{2j} , u_{0j} and u_{2j} effects (i.e. cons and alcohol) and another column (e.g. c106) in the output from prediction drop down list. Accordingly, in order to plot the effects of speed, follow the same process, but select in the Predictions window only the β_{0j} , β_{1j} , u_{0j} and u_{1j} effects (i.e. cons and speed) and another column (e.g. c107) in the output from prediction drop down list. The two plots should be as follows:

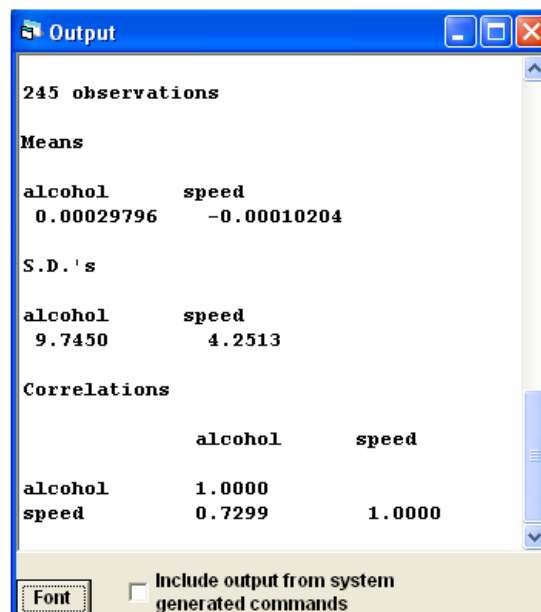


We can see that there are several regions for which the slopes are counter-intuitive. We should examine whether the two variables are correlated.

- Open the Averages and correlations window from the Basic Statistics menu
- Select Correlation in the Operation tab
- Click on alcohol and speed in the variables list
- Click Calculate



The results will appear in the Output window as follows:

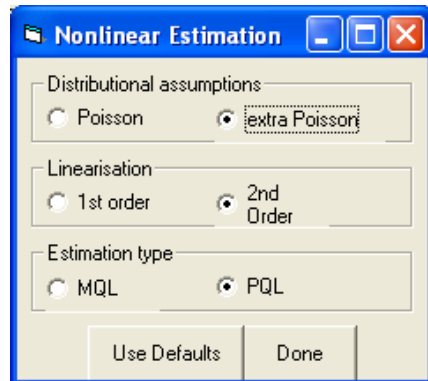


We can see that there is a positive correlation of 0,729 between the variables speed and alcohol, indicating multicollinearity. This explains to some degree the confusing modelling results (see section 2.3.4 of the Methodology Report for more information on multicollinearity effects).

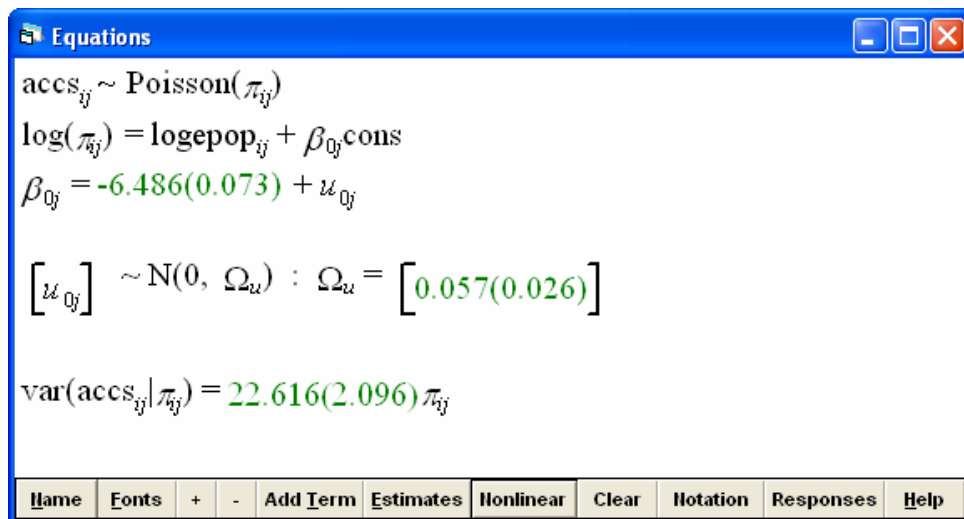
A two-level extra-Poisson model (with overdispersion)

Another issue that needs to be examined in Poisson models is overdispersion (see section 2.3.4 of the Methodology Report). In order to fit a multilevel model with overdispersion, an extra-Poisson distribution is assumed:

- Create a two-level model including only a constant term in the Equations window, as described previously
- Click on the Nonlinear button of the Equations window and select distributional assumptions extra Poisson, linearization 2nd order and estimation type PQL and click Done.



An additional term to be estimated (i.e. the dispersion parameter) will appear in the bottom line of the Equations window, allowing for the mean / variance relationship to be different than 1. Running the model should give the following results:



The results indicate that there is overdispersion in the data, as the dispersion parameter estimated is highly significant.

Adding alcohol to the model gives the following results.

$$\text{accs}_{ij} \sim \text{Poisson}(\pi_{ij})$$

$$\log(\pi_{ij}) = \text{logpop}_{ij} + \beta_{0j}\text{cons} + \beta_{1j}\text{alcohol}_{ij}$$

$$\beta_{0j} = -6.572(0.087) + u_{0j}$$

$$\beta_{1j} = -0.046(0.010) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.082(0.037) & 0.005(0.003) \\ 0.005(0.003) & 0.001(0.000) \end{bmatrix}$$

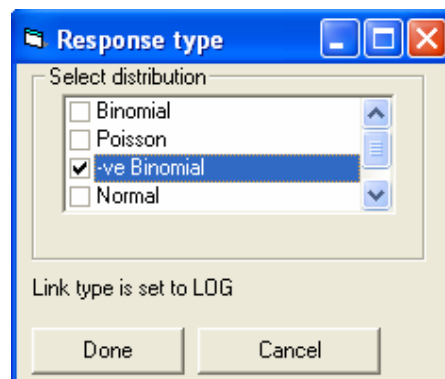
$$\text{var}(\text{accs}_{ij} | \pi_{ij}) = 12.923(1.228) \pi_{ij}$$

In this case, the dispersion parameter is lower (but also significant), indicating that the explanatory variable has accounted for a part of the overdispersion.

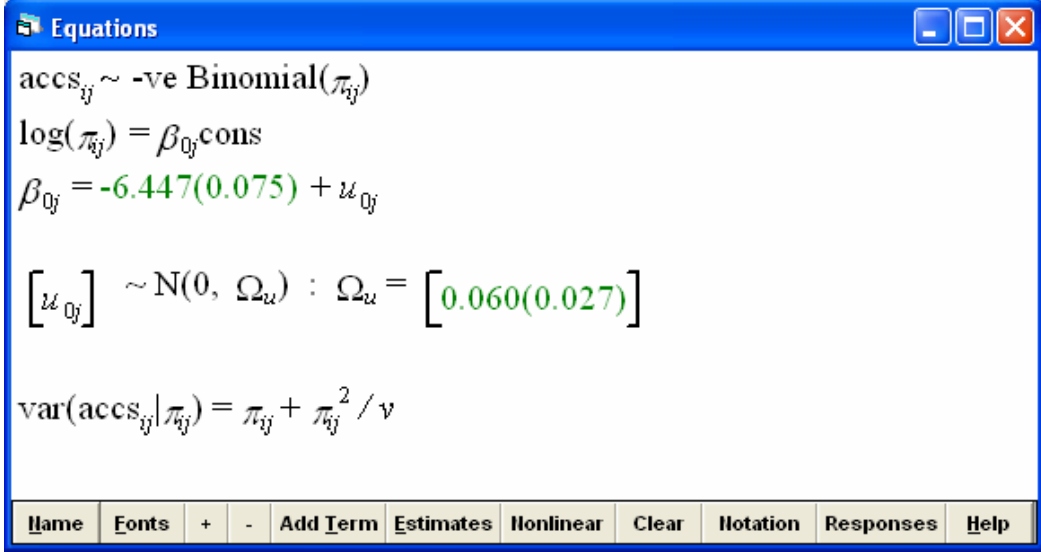
2.3.4.3. A two-level negative binomial model

As explained in the Methodology Report (section 2.3.4), another option for dealing with overdispersion in count data is to assume a Negative Binomial distribution, which includes a more complex variance structure, allowing thus more flexibility. In order to fit a Negative Binomial model in the data:

- Create a two-level model including only a constant term in the Equations window, as described previously
- Click on the N (QX, B) that appears on the first line of the Equations window, select Negative Binomial from the available distributions and click Done.



In the bottom line of the Equations window, the mean / variance relationship is displayed, in which the variance is a quadratic function of the mean. Running the model should give the following results:



The screenshot shows a window titled "Equations" with a blue title bar and standard window controls. The main text area contains the following mathematical expressions:

$$\text{accs}_{ij} \sim \text{-ve Binomial}(\pi_{ij})$$

$$\log(\pi_{ij}) = \beta_{0j} \text{cons}$$

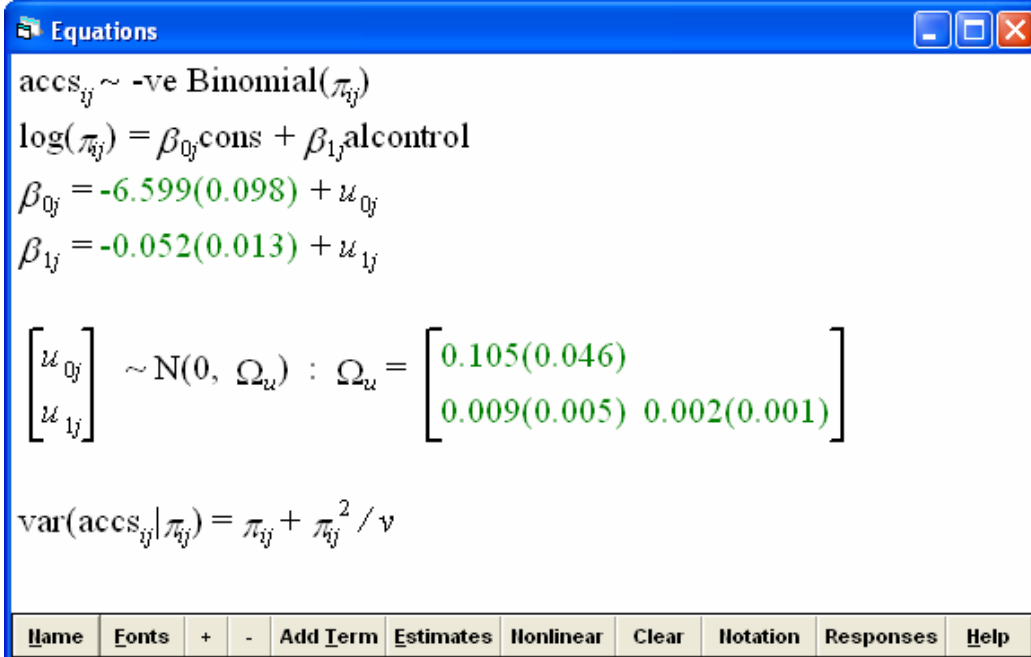
$$\beta_{0j} = -6.447(0.075) + u_{0j}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.060(0.027) \end{bmatrix}$$

$$\text{var}(\text{accs}_{ij} | \pi_{ij}) = \pi_{ij} + \pi_{ij}^2 / v$$

At the bottom, there is a toolbar with buttons: Name, Fonts, +, -, Add Term, Estimates, Nonlinear, Clear, Notation, Responses, and Help.

Adding alcohol to the model gives the following results.



The screenshot shows a window titled "Equations" with a blue title bar and standard window controls. The main text area contains the following mathematical expressions:

$$\text{accs}_{ij} \sim \text{-ve Binomial}(\pi_{ij})$$

$$\log(\pi_{ij}) = \beta_{0j} \text{cons} + \beta_{1j} \text{alcohol}$$

$$\beta_{0j} = -6.599(0.098) + u_{0j}$$

$$\beta_{1j} = -0.052(0.013) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.105(0.046) & \\ 0.009(0.005) & 0.002(0.001) \end{bmatrix}$$

$$\text{var}(\text{accs}_{ij} | \pi_{ij}) = \pi_{ij} + \pi_{ij}^2 / v$$

At the bottom, there is a toolbar with buttons: Name, Fonts, +, -, Add Term, Estimates, Nonlinear, Clear, Notation, Responses, and Help.

These results are very similar to the Extra-Poisson model, in terms of both fixed and random parameter estimates. It is therefore shown that both Extra-Poisson

and Negative Binomial distributional assumptions can efficiently handle overdispersion in count data.

2.4 Longitudinal data

Heike Martensen & Emmanuelle Dupont (IBSR)

The example data used in this section are a set of simulated data for 500 beginning drivers for which a driving-skill score was simulated for 7 consecutive years. Moreover, for each occasion an experience value was generated. This value was always “0” at the first measurement occasion and corresponded to the cumulative number of km driven for all the others. The variable “initial age” indicates for each driver the age at which they acquired their licences.

Data load

Open the file DRIVING SKILL.xls. Like most repeated measures tables, the data are coded in a format that will be called “wide” here. This means that there is one row for each subject with all measurements in it. Below a section of the original wide table is shown. To the right, the table continues with exp3/skill3 to exp6/skill6, below there are more subjects than visible here.

	A	B	C	D	E	F	G	H	I
1	subID	initage	iacen	exp0	skill0	exp1	skill1	exp2	skill2
2	1	24	2	0.00	5.98	0.97	3.35	1.80	5.43
3	2	20	-2	0.00	2.96	0.74	2.52	0.86	1.39
4	3	20	-2	0.00	2.63	0.15	4.81	0.53	5.21
5	4	21	-1	0.00	7.25	0.52	5.10	1.02	4.80
6	5	41	19	0.00	5.67	0.03	4.25	0.97	8.76
7	6	24	2	0.00	4.15	0.02	5.49	0.30	8.65

To analyse these data in MLwiN, they have to be imported (see section 2.2 for instructions to paste data from Excel into MLwiN) and then the wide table must be converted into a long table. This means that all measurements are noted in a long list below each other. Only one variable “experience” and one variable “skill” are present and the measurement occasion is indicated by a third variable. An example for a long table is given here:

	A	B	C	D	E	F
1	subID	initage	iacen	exp	skill	telaps
2	1	24	2	0.00	5.98	0
3	1	24	2	0.97	3.35	1
4	1	24	2	1.80	5.43	2
5	1	24	2	2.69	4.47	3
6	1	24	2	3.17	4.91	4
7	1	24	2	3.33	8.57	5
8	1	24	2	4.29	6.49	6
9	2	20	-2	0.00	2.96	0
10	2	20	-2	0.74	2.52	1
11	2	20	-2	0.86	1.39	2
12	2	20	-2	1.65	4.69	3
13	2	20	-2	1.71	1.95	4
14	2	20	-2	2.23	4.43	5
15	2	20	-2	2.36	2.89	6

To create a long table in MLwiN, click on “Data Manipulation” in the top menu and select “Split record” and fill the dialogue window in as shown below.

Split records

Dimensions

Number of occasions: 7 Number of variables: 2

Stack data

Occasion 2	exp1	skill1
Occasion 3	exp2	skill2
Occasion 4	exp3	skill3
Occasion 5	exp4	skill4
Occasion 6	exp5	skill5
Occasion 7	exp6	skill6
Stacked into	C18	C19

Repeat(carried) data

Input columns: subid, initage, iacen, exp0

Output columns: subid, initage, iacen, exp0

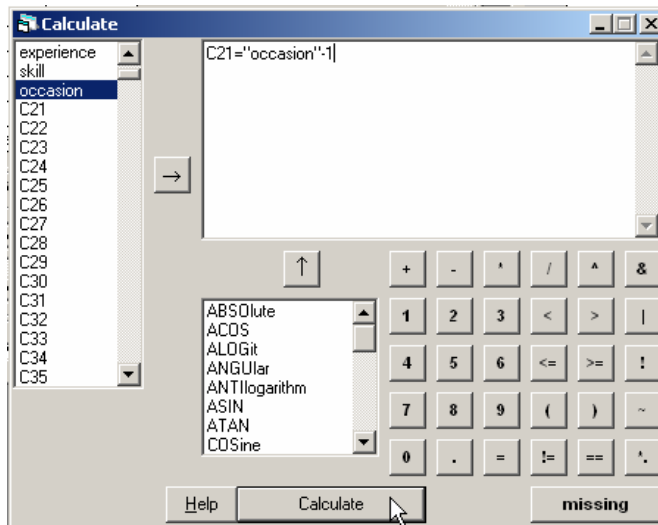
Free columns: Same as input

☐ Generate indicator column: c20

Split **Help**

Click on “Split” to conduct build the long table. Then click on “Data Manipulation” and select “Names”. Change the names of C18, C19, and C20 by selecting each of these numbers from the list, typing their new name into the top frame and pressing return. C18 is the collection of the experience scores. Call it “experience”. C19 is the collection of the skill scores and should be called “skill”. C20 contains an indication from which measurement occasion the data come. This variable, call it “occasion”, has a value from 1 to 7 (1 for skill0, 2 for skill1, ... etc.).

As described in the methodology report, rather than simply indicating the number of the measurements, the *time elapsed* should be coded. This means that values should run from 0 to 6 rather than from 1 to 7. In order to do so, click on “Data Manipulation” in the top menu bar and select “Calculate”. Fill the dialogue window as shown below:



Then open the “Names” window and rename the newly generated variable into “telaps” (short for “time elapsed”).

The long table should contain 3500 cases and consist of the following variables:

subID	Identification number of the driver tested.
initialAge	Age at obtainment of drivers licences.
iacen	Initial age centred to its mean.
experience	Number of km driven (in 1000).
skill	Number between 0 and 15 indicating the driving skill.
telaps	Time elapsed: 0 for first test, 1 – 6 for tests at consecutive years.

2.4.1.1. The empty two-level model

In the case of repeated measurements, the simplest model is the empty two-level model. The first level is that of the single driving skill scores. These skill scores are nested within subjects, as there are 7 skill scores from each person. The repeated measures structure is therefore defined at the second level.

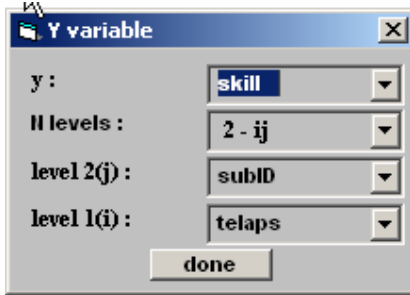
Model formulation

Define the dependent variable:

- Click on the “y” and
 - Select “skill” from the drop-down window as dependent variable
 - Select “1-i” from the N-levels drop-down window
 - Select “subID” from the level 1(j) drop-down window
 - Click “done”

Then define the dependent variable (skill) as varying over two random factors, namely the individual measurements (telaps) and the subject they are taken from (subID)

- Click on the dependent variable and define subID as the second level as shown below.



Then define a variance component model by allowing the intercept to vary randomly across locations.

- Click on the intercept
- Check the box j(subID)
- Press "Done"
- Press "Start" to estimate the parameters

Results and Interpretation

Your Equations window should now look like this.

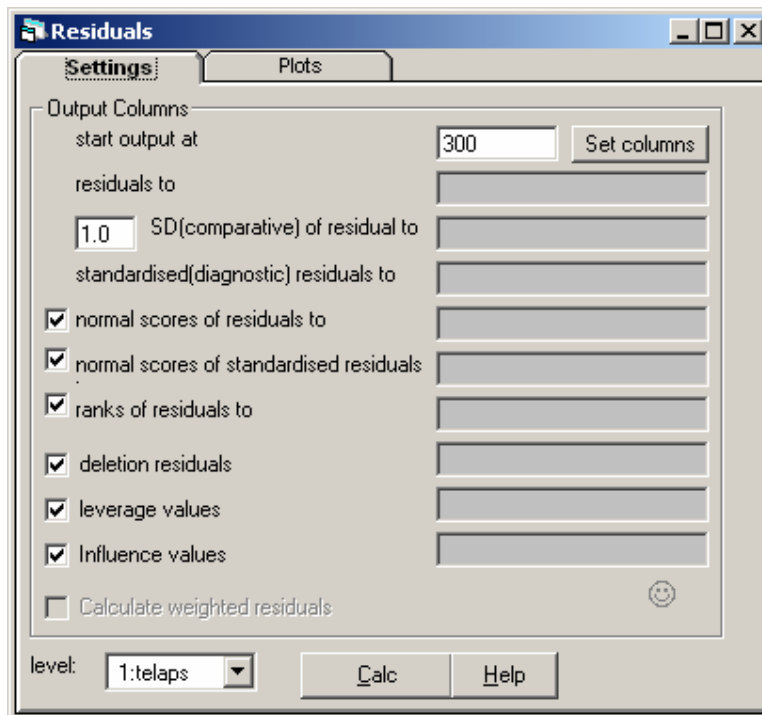
$$\begin{aligned} \text{skill}_{ij} &= \beta_{0j} + e_{ij} \\ \beta_{0j} &= 6.543(0.082) + u_{0j} \\ u_{0j} &\sim N(0, \sigma_{u0}^2) \quad \sigma_{u0}^2 = 2.861(0.212) \\ e_{ij} &\sim N(0, \sigma_e^2) \quad \sigma_e^2 = 3.402(0.088) \\ -2*\loglikelihood &= 15182.690(3500 \text{ of } 3500 \text{ cases in use}) \end{aligned}$$

The within-subject variation is indicated by σ_e^2 and the variation between subjects by σ_{u0}^2 . Note that the results suggest that the differences between repeated driving tests for each participant are larger than those across participants. The mean intercept, β_{0j} , indicates that over all subjects and all times of testing the mean skill score is 6.543.

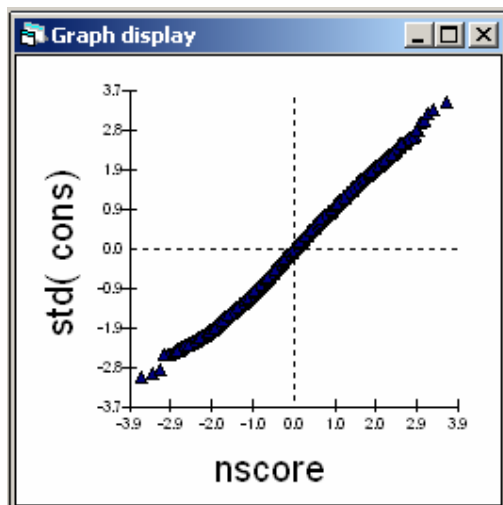
Graphic inspection of the residuals: Level 1

To inspect the residuals of the model,

- Click on Model in the top menu bar
- Select Residuals



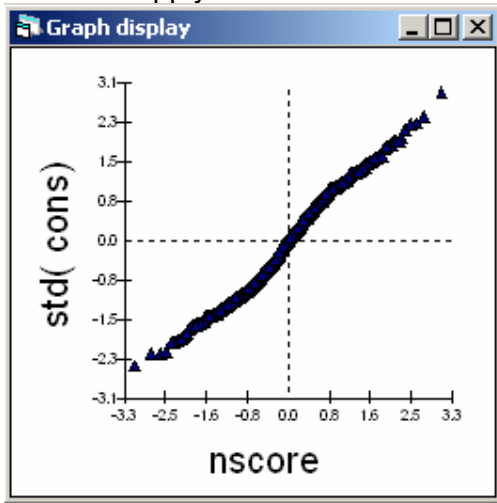
- Click “Calc” to calculate the Level 1 residuals
- Select “Plot” at the top of the “Residuals” window
- Check radio-button in front of “standardised residual and normal score”
- Click “Apply”



Graphic inspection of the residuals: Level 2

- Go back to the “Settings” part of the “Residuals” window
- Select “2:subID” from the “level” dropdown list
- Click “Calc” to calculate the Level 2 residuals
- Select “Plot” at the top of the Residuals window

- Click “Apply”



The residuals are satisfyingly close to a normal distribution (that would have been indicated by a straight line).

2.4.1.2. A Two-level variance component model with one predictor

In the empty model there was more variation within subjects (as indicated by σ_e^2) than between subjects. To estimate the proportion of the within subjects variation that can be attributed to the time elapsed after acquisition of their drivers licence, “telaps” is used as a predictor.

Model formulation

- Click “Add Term” at the bottom of the Equations window
- Select “telaps” from the “Variable” drop-down window
- Click “Done”
- To estimate this model press “Start”.

Results and Interpretation**Equations**

$$\text{skill}_{ij} = \beta_{0j} + 0.496(0.013)\text{telaps}_{ij} + e_{ij}$$

$$\beta_{0j} = 5.055(0.090) + u_{0j}$$

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad \sigma_{u0}^2 = 3.025(0.212)$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 2.254(0.058)$$

$$-2 * \loglikelihood = 13947.380 (3500 \text{ of } 3500 \text{ cases in use})$$

Because “telaps” is coded to be zero at the time of acquirement of the driving licence, the mean intercept β_{0j} indicates the average skill score at that moment. The coefficient for “telaps” indicates that on average the skill score increases by half a point each year that a driver has his/her licence. Note that the within subject variance σ_e^2 is reduced as compared to the null model, suggesting that the measurements for each participant changed over time. However, the between subject variance σ_{u0}^2 as well, suggesting that participants varied in the effect that telaps had for them. Finally, the decrease of the deviance (loglikelihood) confirms that the model with “telaps” fits better than the one without.

2.4.1.3. Two-level random intercept model with two predictors

The question “treated” in this simulated study is whether it is the number of years passing or rather the increase of experience that make older drivers less accident prone than younger ones. To investigate this, the driving experience (measured in 1000 km driven) is taken up into the model in parallel with the time elapsed (telaps).

Model formulation

- Click “Add Term” at the bottom of the Equations window
- Select “experience” from the “Variable” drop-down window
- Click “Done”
- To estimate this model press “Start”

Results and Interpretation

Equations

$$\text{skill}_{ij} = \beta_{0j} + 0.028(0.038)\text{telaps}_{ij} + 0.947(0.073)\text{experience}_{ij} + e_{ij}$$

$$\beta_{0j} = 5.049(0.090) + u_{0j}$$

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad \sigma_{u0}^2 = 3.015(0.209)$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 2.138(0.055)$$

$$-2 * \log\text{likelihood} = 13784.950 (3500 \text{ of } 3500 \text{ cases in use})$$

The coefficient for the variable “experience” is highly significant as it exceeds its own standard error several times. The coefficient for “telaps” however, is not significant anymore. This might be confusing, especially when noting that overall the model has clearly increased in fit (to test the significance of the difference between the two deviances, 162.43, use “basic statistics”, “tail areas”). The change in the predictor for “telaps” is due to the fact that the time elapsed since one has acquired his driver’s licence and the number of kilometres one has driven are two related measures. A significant coefficient indicates that a proportion of the variance can be attributed to the predictor **exclusively**. The fact that “telaps” is not significant anymore when taken up together with “experience” suggests, that all the variance that “telaps” explained can be explained by “experience” as well. Note however, that the reverse is not the case: “experience” is significant even if it is taken up jointly with “telaps”, indicating that there is a proportion of variance in the driving skills that can be uniquely attributed to “experience”.

2.4.1.4. A Two-level random intercept model with predictor experience only

Because the predictor “telaps” (i.e. the time elapsed since acquirement of the driving licence) is not significant anymore one “experience” is taken up into the equation, this term can be dropped and “experience” remains in the model equation.

Model formulation

- Click the term “telaps” in the model equation
- Click on “delete Term”
- To estimate this model press “Start”

Results and Interpretation**Equations**

$$\text{skill}_{ij} = \beta_{0j} + 0.998(0.024)\text{experience}_{ij} + e_{ij}$$

$$\beta_{0j} = 5.057(0.089) + u_{0j}$$

$$u_{0j} \sim N(0, \sigma_{u0}^2) \quad \sigma_{u0}^2 = 3.022(0.210)$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 2.137(0.055)$$

$$-2 * \log\text{likelihood} = 13785.500 (3500 \text{ of } 3500 \text{ cases in use})$$

Removing “telaps” from the model has increased the deviance by only 0.55. With 1 degree of freedom (one parameter that needs to be estimated less), this CHI-2 value has a probability of 0.76 which is not significant at all. Indeed, all the variance in the “skill” scores that can be explained by time elapsed (telaps), can be explained by “experience” as well.

2.4.1.5. A Two-level random intercept random slope model

The coefficient for “experience” indicates that generally driving skills improve with an increase of km driven. To test whether this increase is the same for all participants we will allow the slope of experience to vary randomly across the level-2 units, i.e. the participants.

Model formulation

- Click on the Length
- Check the box j(IDlocation)
- Estimate the parameters by clicking on “Start”

Results and Interpretation

Equations

$$\text{skill}_{ij} = \beta_{0j} + \beta_{1j} \text{experience}_{ij} + e_{ij}$$

$$\beta_{0j} = 5.058(0.077) + u_{0j}$$

$$\beta_{1j} = 1.002(0.027) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 2.094(0.190) & 0.244(0.049) \\ 0.244(0.049) & 0.090(0.023) \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 2.027(0.057)$$

$$-2 * \log \text{likelihood} = 13700.750 (3500 \text{ of } 3500 \text{ cases in use})$$

Both, the intercept β_{0j} and the coefficient of “Experience”, β_{1j} , are now varying across subjects with the means indicated in the second and third equation.

The between subject variance has now become a variance-covariance matrix Ω_u : The upper left number is σ_{u0}^2 , the variance of the intercepts across subjects. It indicates how much the average driving score varies between them. The lower right number is σ_{u1}^2 , the variance of the “experience”- coefficient across locations. It shows that the relation between “experience” and “skill” varies between participants. The lower left number is σ_{u01} , the covariance between the two, indicating that subjects with a higher intercept (i.e. the average “skill”) have steeper slopes (i.e. a stronger increase of skills with experience).

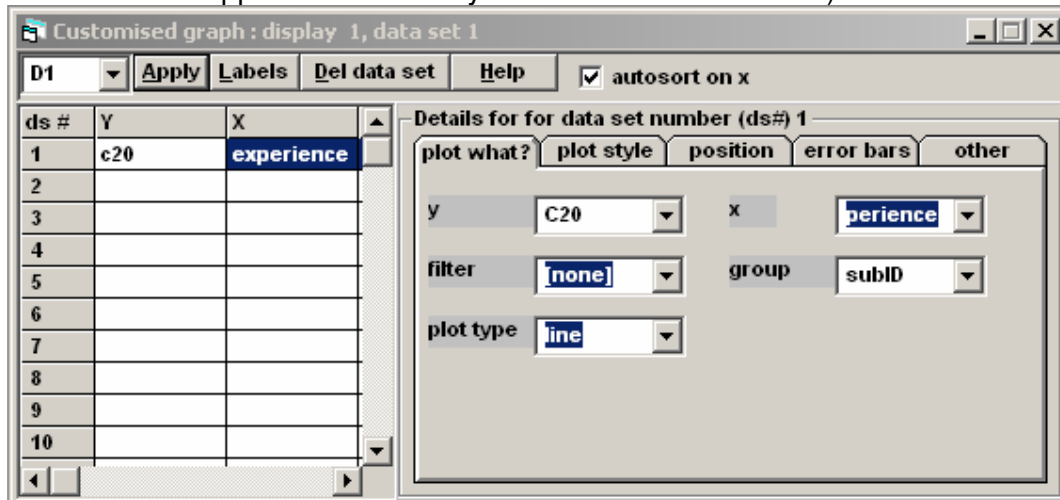
The deviance decreased by 85 as opposed to the variance component model, indicating that there is indeed substantial variation in the size of the experience-effect, which also is apparent in the fact that the variance of the slope, σ_{u1}^2 , is significant.

Graphic inspection of the model predictions

To plot the model predictions, they have to be saved as a new variable first.

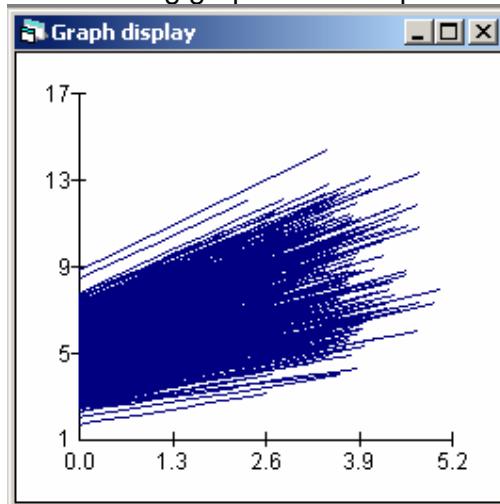
- Select “Model” in the top menu bar and click on “Predictions”
- Click on all parameters in the lower half of the appearing dialogue window (so that they turn from grey to black)
- Select an empty column (e.g. c20) for “output from prediction to”
- Click on “Calc”
- Now you can close the “Predictions” window

Now a variable that contains the predictions of the model is created, which can be plotted against the “experience”. Open the graph dialogue by clicking on “Graph” in the top menu bar and by selecting “Customized Graphs”. Fill in as shown below (all changes have to be made in the drop-down lists on the right-hand side and appear automatically in the left-hand side table)



- Press “Apply” to create the graph. (You can close the Dialogue Window then.)

The resulting graph shows separate regression lines for each location.



The graph shows a “the rich get richer” effect: Those drivers that start at a high level also improve more with additional driving experience. In the model output, this effect becomes apparent in the covariance between intercept and slope, σ_{u01} .

2.4.1.6. Adding a level-two predictor

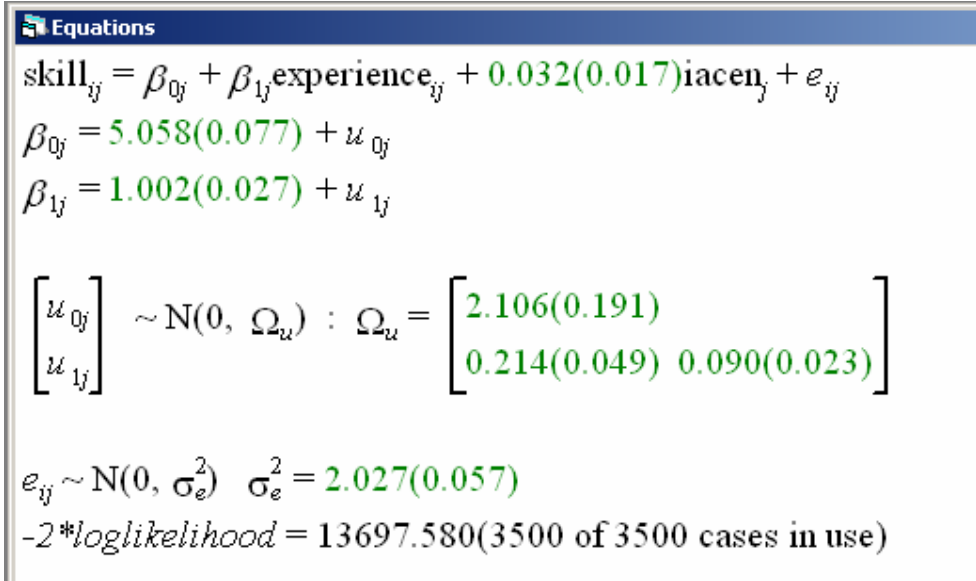
The beginning level of driving skills and the effect of “experience” vary systematically across participants. It is therefore sensible to search for predictors at the second level (here the subject level) to explain these

variations. As an example we will include the variable “iacen”, the initial age (ia) centered to its mean (cen).

Model formulation

- Click “Add Term” at the bottom of the Equations window
- Select “iacen” from the “Variable” drop-down window
- Click “Done”
- To estimate this model press “Start”

Results and Interpretation



The screenshot shows the 'Equations' window with the following results:

$$\text{skill}_{ij} = \beta_{0j} + \beta_{1j}\text{experience}_{ij} + 0.032(0.017)\text{iacen}_j + e_{ij}$$

$$\beta_{0j} = 5.058(0.077) + u_{0j}$$

$$\beta_{1j} = 1.002(0.027) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 2.106(0.191) & \\ 0.214(0.049) & 0.090(0.023) \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 2.027(0.057)$$

$$-2 * \log\text{likelihood} = 13697.580 (3500 \text{ of } 3500 \text{ cases in use})$$

The coefficient of “iacen” is marginally significant ($Z = 1.88$; $p = .060$). Its positive value would indicate that drivers who acquired their driving licences at a higher age tend to have higher skill scores.

2.4.1.7. Adding a cross-level interaction

As the next step it will be tested whether the initial age (iacen) modifies the effect of “experience”. To do this, the interaction between these two variables is included into the model.

Model formulation

- Click on “Add term” and
- Include the interaction between “iacen” and “experience”
 - Select order 1 (this means it’s a first-order interaction)
 - Select “iacen” as first variable
 - Select “experience” as second

- Click “Done”
- Estimate the model by pressing “Start”

Results and Interpretation

Equations

$$\text{skill}_{ij} = \beta_{0j} + \beta_{1j}\text{experience}_{ij} + 0.010(0.017)\text{iacen}_j + 0.046(0.006)\text{iacen}.\text{experience}_{ij} + e_{ij}$$

$$\beta_{0j} = 5.062(0.077) + u_{0j}$$

$$\beta_{1j} = 0.998(0.025) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 2.093(0.190) & \\ 0.237(0.045) & 0.043(0.020) \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma_e^2) \quad \sigma_e^2 = 2.030(0.057)$$

$-2 * \loglikelihood = 13634.490$ (3500 of 3500 cases in use)

Taking up the interaction term leaves the already marginal coefficient of “iacen” non-significant. The interaction between initial age (iacen) and “experience” itself, however, is significant. Its positive coefficient indicates that drivers who acquired their licence at a later age improved more per 1000 km driven than those who acquired their licence at an earlier age. (Please remember that all these “conclusions” are based on simulated data and generated for this manual).

2.4.1.8. Conclusion

In this chapter it was demonstrated how to use a two-level model to analyse repeated measurements taken from a group of participants. It was shown that the first level indicates the variation between measurements taken from the same subject, while the second level contains variation between subjects. Accordingly, variables that vary within subject across measurements (e.g. time or growth variables) should be included at level one, while variables that characterise individuals should be taken up as level-two variables. In the case of a repeated measurements analysis, a cross-level interaction then indicates how person-characteristics can modify the effect of time- or growth variables.

2.5 Multivariate models

George Yannis, Eleonora Papadimitriou and Costas Antoniou (NTUA)

In this section, an example for fitting multivariate multilevel models is presented using the MLwinN 2.01 software. The example concerns an investigation of the regional effect of drinking-and-driving enforcement on the number of road accidents and related persons killed in Greece. The theoretical background, models fit and results were discussed in section 2.5 of the Methodology report.

It is noted that in section 2.5 of the Methodology Report two model formulations were defined and presented: a normal bivariate multilevel model and a hybrid normal-poisson bivariate multilevel model. However, only the latter is demonstrated in this section, as this formulation was proved to be more efficient in the estimation of the models. However, apart from the different level-1 distributional assumptions (see section 2.5. of the Methodology Report) the same process would be followed for fitting the first formulation as well."

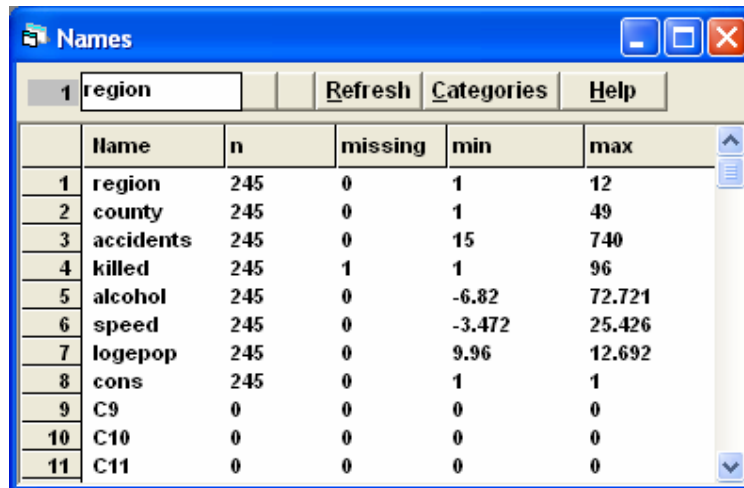
The dataset includes data on accidents and persons killed, as well as alcohol and speed police controls data for the 49 counties and 12 regions of Greece for the period 1998-2002. Part of this dataset was also used in section 2.3.4 of the Methodology report (multilevel models for count data) and in the related demonstration for the Manual. In this section, a variable corresponding to the number of persons killed was also included.

More specifically, the variables and values used are summarized in the following Table:

Region	1-12 regions of Greece
County	1-49 counties of Greece
Accidents	The number of accidents of each county
Killed	The number of persons killed of each county
alcohol	The number of alcohol controls of each county (1000 alcohol controls)
Speed	The number of speed infringements of each county (1000 speed infringements)
logepop	The natural logarithm of the population of each county
Cons	The constant term (1)

It is reminded that the counties of Athens and Thessalonica (large metropolitan areas with disproportionally high numbers of road accidents, persons killed and police controls) are not included in the dataset. It is also reminded that the explanatory variables (alcohol and speed controls) are centered around their mean to avoid numerical problems in the estimations.

- Open the dataset MultivariateManualData2.ws using the Open Worksheet option from the Files menu. Opening the Names window from the Data Manipulation menu gives the following:

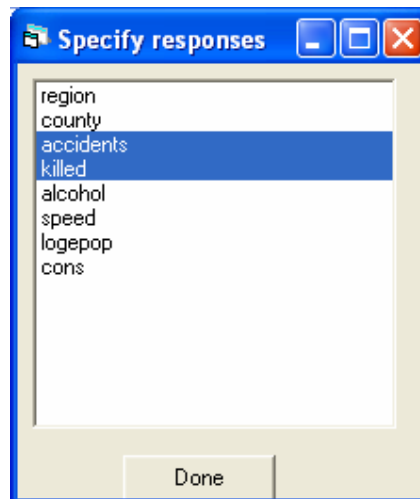


The 'Names' window displays a table of variables and their statistics. The table has columns for 'Name', 'n', 'missing', 'min', and 'max'. The variables listed are region, county, accidents, killed, alcohol, speed, logpop, cons, C9, C10, and C11.

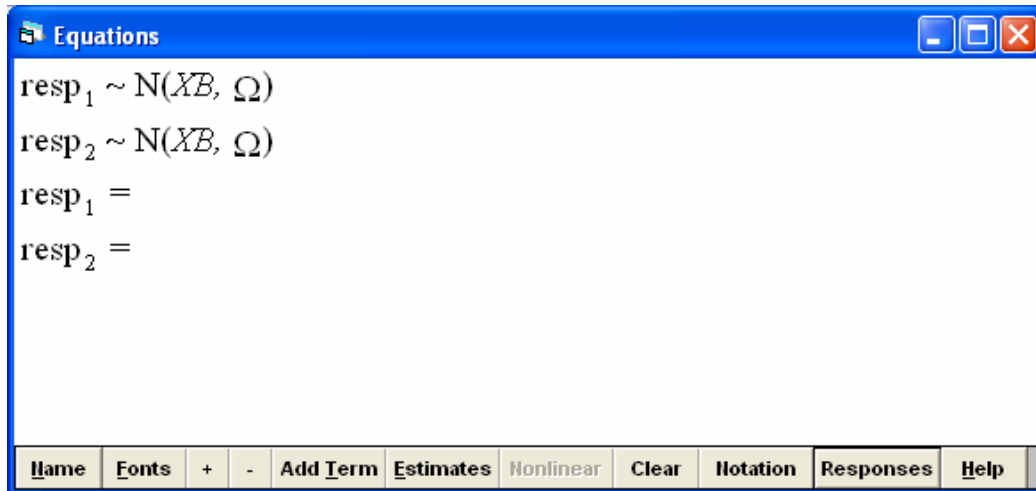
	Name	n	missing	min	max
1	region	245	0	1	12
2	county	245	0	1	49
3	accidents	245	0	15	740
4	killed	245	1	1	96
5	alcohol	245	0	-6.82	72.721
6	speed	245	0	-3.472	25.426
7	logpop	245	0	9.96	12.692
8	cons	245	0	1	1
9	C9	0	0	0	0
10	C10	0	0	0	0
11	C11	0	0	0	0

As described in the multivariate multilevel models Methodology report, the two responses will be treated as 2nd level grouping and the actual values of both responses will be treated as 1st level units. In order to define the bivariate structure:

- Click on the Responses button in the Equations window
- In the Specify responses window, click on accidents and killed and then click Done:



The Equations window should now look like this:



Moreover, in the Names window, we can see that two new variables were created; the variable the variable `resp_indicator`, which is a binary variable separating the two responses (i.e. indicating to which response the current data row applies to), and `resp`, which contains the respective actual values of the two responses.

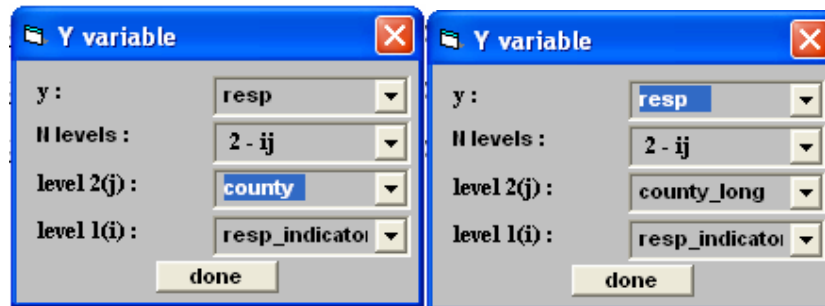
Note also that the variable `resp` includes exactly twice the number of entries of each response, i.e. $2 \times 245 = 490$. Moreover, the minimum and maximum values of this variable are the minimum and maximum values of the grouped values of both responses.

	Name	n	missing	min	max
1	region	245	0	1	12
2	county	245	0	1	49
3	accidents	245	0	15	740
4	killed	245	1	1	96
5	alcohol	245	0	-6.82	72.721
6	speed	245	0	-3.472	25.426
7	logpop	245	0	9.96	12.692
8	cons	245	0	1	1
9	resp_indicator	490	0	1	2
10	resp	490	1	1	740
11	C11	0	0	0	0
12	C12	0	0	0	0
13	C13	0	0	0	0
14	C14	0	0	0	0

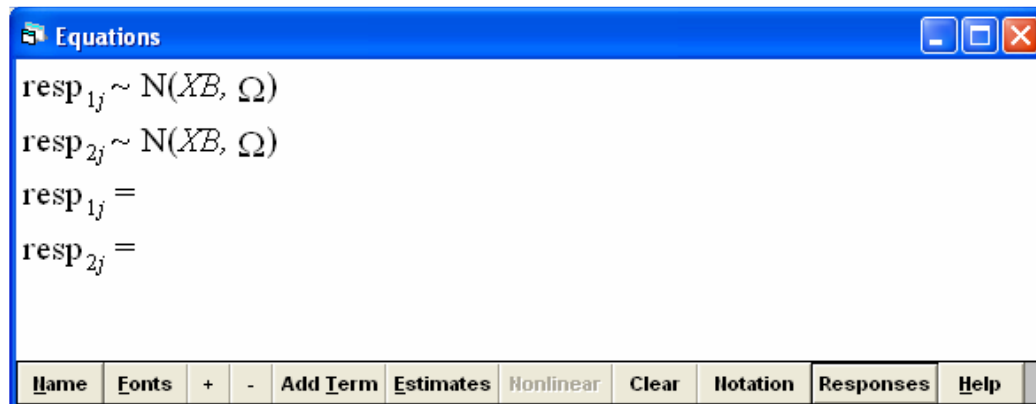
However, so far we have specified a single level model. In order to define the multivariate two-level structure, we should specify that counties are nested within responses.

- Click on resp1 or resp2 in the Equations window
- In the Y variable window, select 2-ij from the N levels: drop down list and county from the level 2(j) drop down list, and click Done.

If we click on resp1 or resp2 again, we can see that the county has been replaced by a new variable county_long (see below picture on the right).

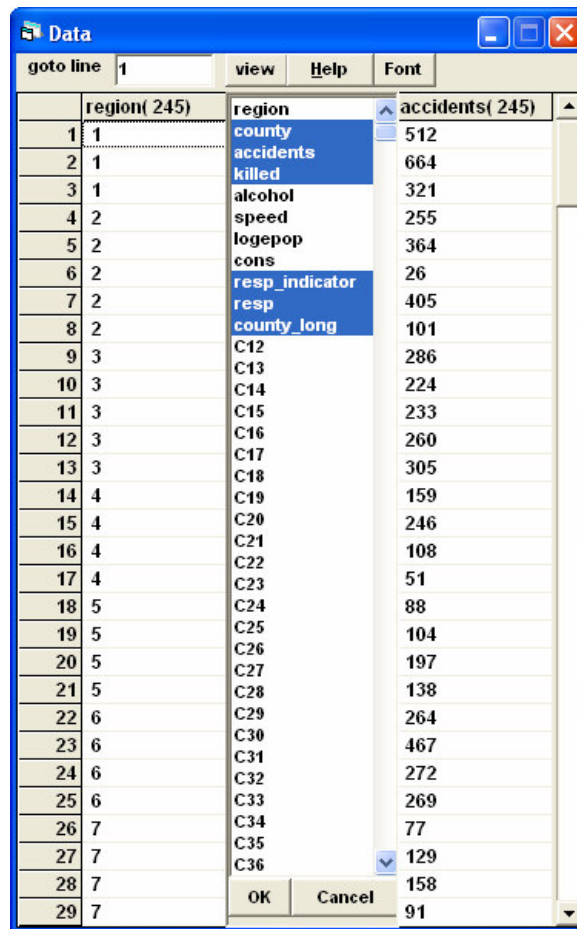


The Equations window should now look like this:



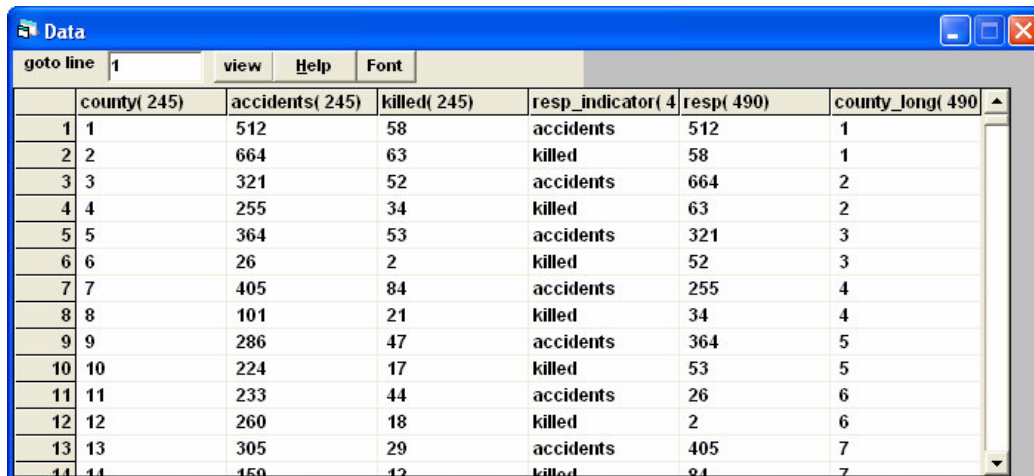
It is interesting to see how the multilevel modelling properties are exploited to build the multivariate structure out of the initial dataset.

- From the Data Manipulation menu on the main toolbar, select View or Edit Data.
- Click on the view button in the Data window and select: county, accidents, killed, resp_indicator, resp and county_long.



region(245)	region	accidents(245)
1	1	512
2	1	664
3	1	321
4	2	255
5	2	364
6	2	26
7	2	405
8	2	101
9	3	286
10	3	224
11	3	233
12	3	260
13	3	305
14	4	159
15	4	246
16	4	108
17	4	51
18	5	88
19	5	104
20	5	197
21	5	138
22	6	264
23	6	467
24	6	272
25	6	269
26	7	77
27	7	129
28	7	158
29	7	91

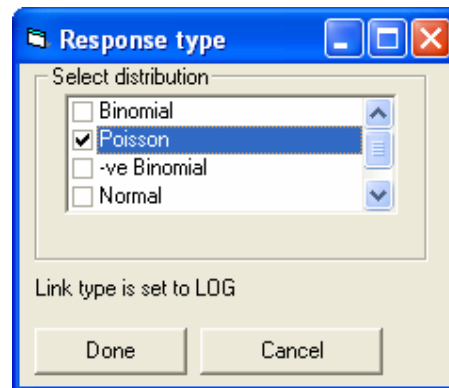
We can see how the `resp_indicator` variable separates the two responses, while their respective values are stored in a single column (`resp`). Moreover, the new variable `county_long` is the 2nd level grouping variable. This demonstration fully corresponds to the general theoretical multivariate structure presented in Table 2.5.1 of section 2.5 of the Methodology report.



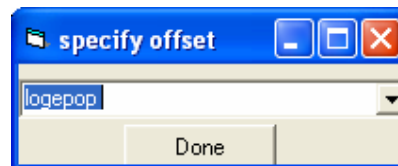
county(245)	accidents(245)	killed(245)	resp_indicator(4)	resp(490)	county_long(490)
1	512	58	accidents	512	1
2	664	63	killed	58	1
3	321	52	accidents	664	2
4	255	34	killed	63	2
5	364	53	accidents	321	3
6	26	2	killed	52	3
7	405	84	accidents	255	4
8	101	21	killed	34	4
9	286	47	accidents	364	5
10	224	17	killed	53	5
11	233	44	accidents	26	6
12	260	18	killed	2	6
13	305	29	accidents	405	7
14	159	12	killed	91	7

Before we proceed in fitting the multivariate model, the distributional assumptions of the two responses should be specified. As discussed in section 2.3.4, the counts of road accidents and persons killed are random counts of events occurring within a population and consequently they can only take positive integer values. Therefore, a Poisson distribution is assumed and a log link function should be used together with an appropriate offset term.

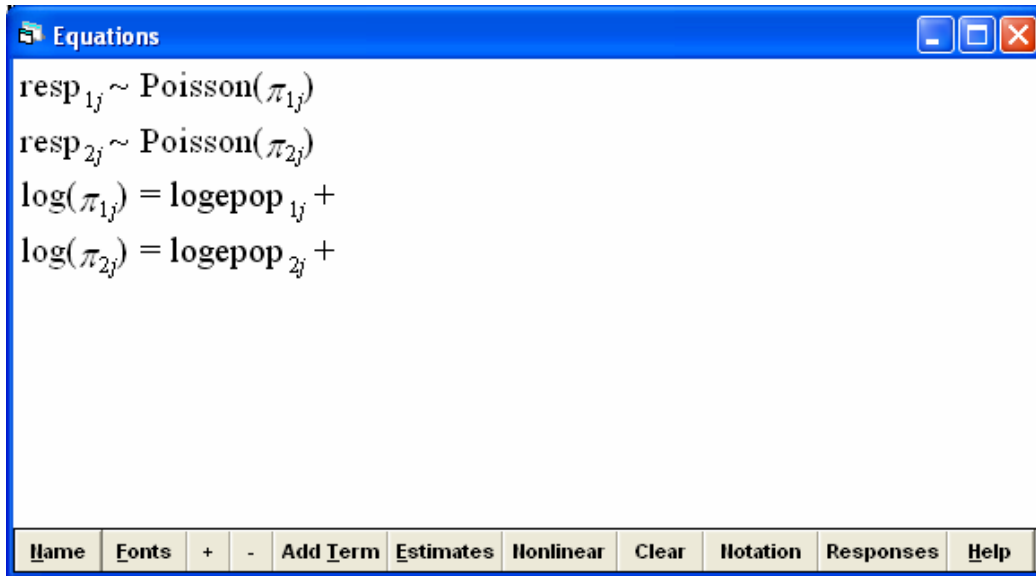
- Click on the $N(\Omega X, B)$ that appears for each response of the Equations window, select Poisson from the available distributions and click Done.



- Click on the (π_i) that appears for each response in the Equations window, select logepop as offset term and click Done.

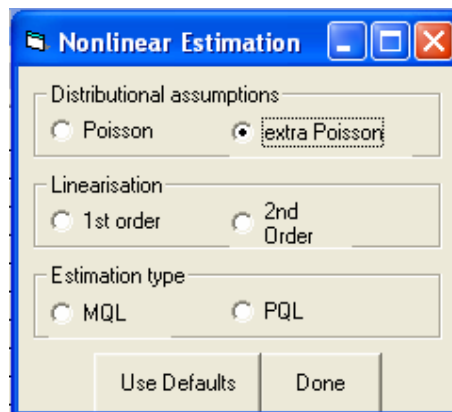


The Equations window should now look like this:



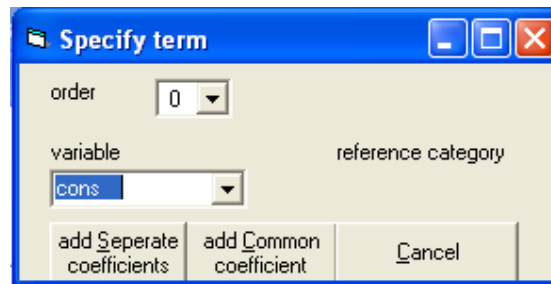
In section 2.3.4 of the Methodology report, it was shown that overdispersion was present in the accidents data and that extra-Poisson or Negative Binomial distributional assumptions would be required in order to handle this unexplained variation. As Negative Binomial responses are not available in this latest version of the software, we will model the two responses by assuming extra-Poisson distributions (for details see section 2.3.4 of the Methodology report).

- Click on the Nonlinear button of the Equations window and select Distributional assumptions extra Poisson.

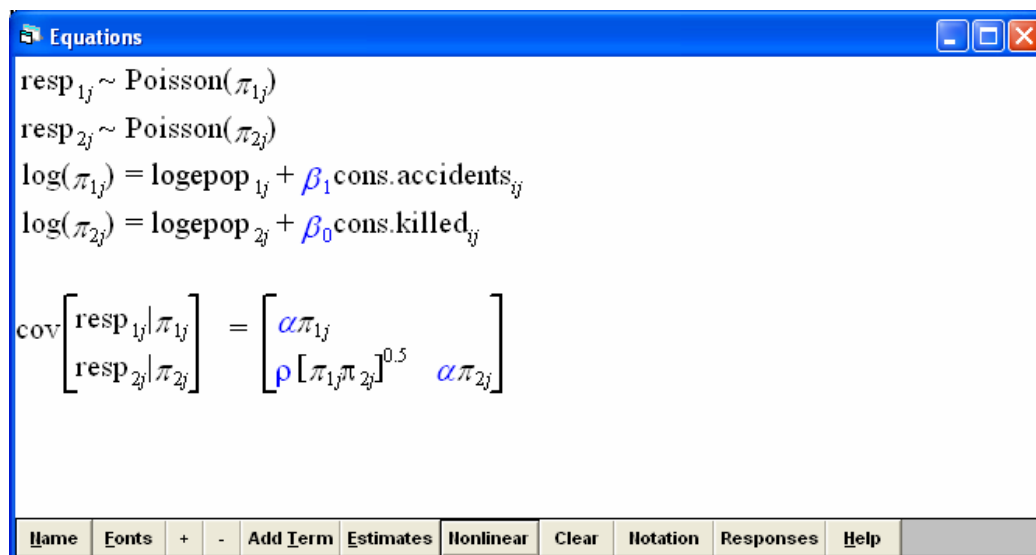


We will now enter variables in the model, starting by an intercept term.

- Click on the Add Term button of the Equations window
- In the Specify term window, select cons from the variable drop-down list and click on Add Separate coefficients.



- Click on the Estimates button of the Equations window and the parameters to be estimated will be highlighted in blue. The Equations window will now look as follows:



First of all, we can see that the coefficients β_0 and β_1 are fixed by default (i.e. the option of random variation is not available when clicking on the related term of the model). This is due to the fact that, as explained in section 2.3.4 of the Methodology report, no random structure can be defined at the lowest level of a Poisson model, as the level-1 variance is assumed to be equal to the mean, and therefore known. In this case, though, the lowest level of the Poisson variables is level-2 of the multivariate model.

Moreover, a covariance matrix for the two responses is created. In this matrix, two dispersion parameters α are to be estimated, one for each response, in order to fit extra-Poisson models, in which the variance-mean equality assumption is relaxed. Finally, a covariance ρ between the two responses will be estimated.

- Click on the Estimation Control button of the main Toolbar and select RIGLS
- Click Start to run the model. The results are as follows:

Equations

$$\text{resp}_{1j} \sim \text{Poisson}(\pi_{1j})$$

$$\text{resp}_{2j} \sim \text{Poisson}(\pi_{2j})$$

$$\log(\pi_{1j}) = \text{logpop}_{1j} + -6.471(0.025)\text{cons.accidents}_{ij}$$

$$\log(\pi_{2j}) = \text{logpop}_{2j} + -8.380(0.023)\text{cons.killed}_{ij}$$

$$\text{cov} \begin{bmatrix} \text{resp}_{1j} | \pi_{1j} \\ \text{resp}_{2j} | \pi_{2j} \end{bmatrix} = \begin{bmatrix} 30.511(2.755)\pi_{1j} & 4.691(0.742) [\pi_{1j}\pi_{2j}]^{0.5} \\ 3.700(0.334)\pi_{2j} & \end{bmatrix}$$

Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

The intercept terms of the two responses are both highly significant. Moreover, a significant between-response covariance (ρ) indicates that more road accidents per county correspond to more persons killed per county. The significant dispersion parameters (α) of the two responses indicate that the extra-Poisson distributional assumptions adopted were reasonable.

We will now introduce a (fixed) slope term in the model term.

- Click on the Add Term button of the Equations window
- In the Specify term window, select alcohol from the variable drop-down list and click on Add Separate coefficients.
- Click More to run the model. The results are as follows:

Equations

$$\text{resp}_{1j} \sim \text{Poisson}(\pi_{1j})$$

$$\text{resp}_{2j} \sim \text{Poisson}(\pi_{2j})$$

$$\log(\pi_{1j}) = \text{logpop}_{1j} + -6.455(0.023)\text{cons.accidents}_{ij} + -0.019(0.003)\text{alcohol.accidents}_{ij}$$

$$\log(\pi_{2j}) = \text{logpop}_{2j} + -8.372(0.023)\text{cons.killed}_{ij} + -0.006(0.002)\text{alcohol.killed}_{ij}$$

$$\text{cov} \begin{bmatrix} \text{resp}_{1j} | \pi_{1j} \\ \text{resp}_{2j} | \pi_{2j} \end{bmatrix} = \begin{bmatrix} 24.555(2.216)\pi_{1j} & 4.139(0.657) [\pi_{1j}\pi_{2j}]^{0.5} \\ 3.614(0.326)\pi_{2j} & \end{bmatrix}$$

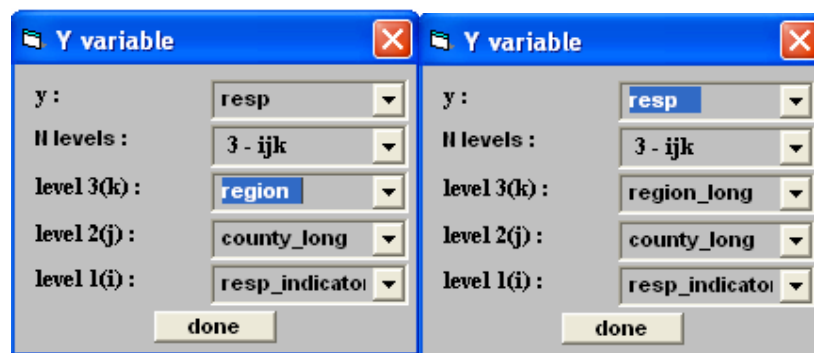
Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

These results indicate that the effect of alcohol enforcement (*alcohol.accidents* and *alcohol.killed*) is significant both for the number of accidents and for the number of persons killed.

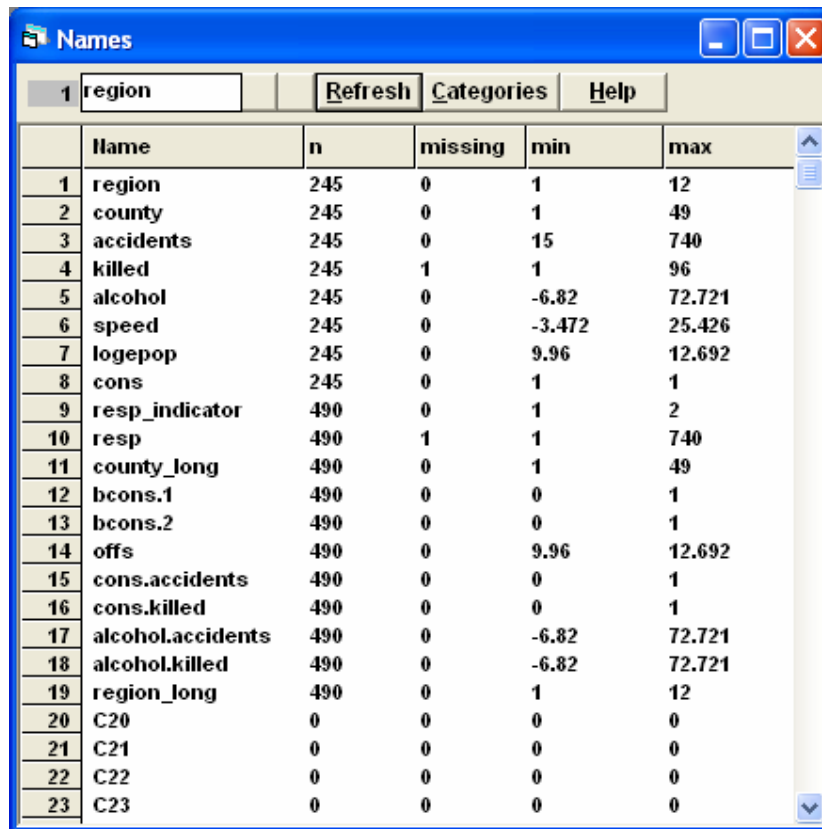
We will now proceed in building a three-level model, in order to investigate regional effects. This third level is created above the existing 2nd level, which corresponds to the counts of the two responses at the county-level.

- Click on resp1 or resp2 in the Equations window
- In the Y variable window, select 3-ijk from the N levels: drop down list and region from the level 3(k) drop down list, and click Done.

If we click on resp1 or resp2 again, we can see that the region has been replaced by a new variable region_long (see below picture on the right), in order to comply to the multivariate structure specified previously.



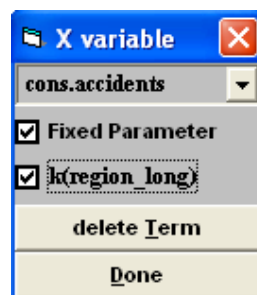
It is interesting to note that, in the Names window, several new variables have been created (cons.accidents, cons.killed, alcohol.accidents, alcohol.killed), as a result of the previous two-level modeling, in order to define the intercept and slope terms for each one of the responses. The additions of the 3rd level resulted in the creation of the variable region_long.



	Name	n	missing	min	max
1	region	245	0	1	12
2	county	245	0	1	49
3	accidents	245	0	15	740
4	killed	245	1	1	96
5	alcohol	245	0	-6.82	72.721
6	speed	245	0	-3.472	25.426
7	logpop	245	0	9.96	12.692
8	cons	245	0	1	1
9	resp_indicator	490	0	1	2
10	resp	490	1	1	740
11	county_long	490	0	1	49
12	bcons.1	490	0	0	1
13	bcons.2	490	0	0	1
14	offs	490	0	9.96	12.692
15	cons.accidents	490	0	0	1
16	cons.killed	490	0	0	1
17	alcohol.accidents	490	0	-6.82	72.721
18	alcohol.killed	490	0	-6.82	72.721
19	region_long	490	0	1	12
20	C20	0	0	0	0
21	C21	0	0	0	0
22	C22	0	0	0	0
23	C23	0	0	0	0

Delete the terms alcohol.accidents and alcohol.killed from the models in the Equations window, in order to fit a random intercept model.

- Click on cons.accidents in the Equations window
- In the X variable window, click in the k(region_long) box to specify the random variation among regions
- Repeat for cons.killed



Note that now a 3rd level covariance matrix is also to be estimated, including the level-3 variances of the two intercepts and their covariance. When running the model, the following output is produced:

Equations

$$\text{resp}_{1jk} \sim \text{Poisson}(\pi_{1jk})$$

$$\text{resp}_{2jk} \sim \text{Poisson}(\pi_{2jk})$$

$$\log(\pi_{1jk}) = \text{logpop}_{1jk} + \beta_{1k} \text{cons.accidents}_{ijk}$$

$$\beta_{1k} = -6.453(0.044) + v_{1k}$$

$$\log(\pi_{2jk}) = \text{logpop}_{2jk} + \beta_{0k} \text{cons.killed}_{ijk}$$

$$\beta_{0k} = -8.382(0.028) + v_{0k}$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.016(0.008) & 0.025(0.010) \\ 0.025(0.010) & 0.092(0.021) \end{bmatrix}$$

$$\text{cov} \begin{bmatrix} \text{resp}_{1jk} | \pi_{1jk} \\ \text{resp}_{2jk} | \pi_{2jk} \end{bmatrix} = \begin{bmatrix} 15.163(1.573) \pi_{1jk} & 2.898(0.556) [\pi_{1jk} \pi_{2jk}]^{0.5} & 3.248(0.333) \pi_{2jk} \end{bmatrix}$$

Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

A significant regional variation of both road accidents and road accident casualties and a significant covariance between the two intercepts are obtained. However, the regional variation of the intercept is higher for the number of persons killed. Moreover, the covariance between responses (ρ) and its significance is reduced. We may conclude that the variations of accidents and persons killed follow the same trend both at national level and within different regions i.e. some of the covariance between accidents and persons killed is situated at the regional level.

The final stage of the modeling concerns the introduction of a random slope.

- Click on the Add Term button of the Equations window
- In the Specify term window, select alcohol from the variable drop-down list and click on Add Separate coefficients, as shown above.
- Right-click on the Ω_v term in the Equations window and select Set diagonal matrix from the menu displayed.

Equations

$$\begin{aligned} \text{resp}_{1jk} &\sim \text{Poisson}(\pi_{1jk}) \\ \text{resp}_{2jk} &\sim \text{Poisson}(\pi_{2jk}) \\ \log(\pi_{1jk}) &= \text{logpop}_{1jk} + \beta_{0k} \text{cons.accidents}_{ijk} + \beta_{2k} \text{alcohol.accidents}_{ijk} \\ \beta_{0k} &= \beta_0 + v_{0k} \\ \beta_{2k} &= \beta_2 + v_{2k} \\ \log(\pi_{2jk}) &= \text{logpop}_{2jk} + \beta_{1k} \text{cons.killed}_{ijk} + \beta_{3k} \text{alcohol.killed}_{ijk} \\ \beta_{1k} &= \beta_1 + v_{1k} \\ \beta_{3k} &= \beta_3 + v_{3k} \end{aligned}$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \\ v_{2k} \\ v_{3k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01} & \sigma_{v02} & \sigma_{v03} \\ \sigma_{v01} & \sigma_{v1}^2 & \sigma_{v12} & \sigma_{v13} \\ \sigma_{v02} & \sigma_{v12} & \sigma_{v2}^2 & \sigma_{v23} \\ \sigma_{v03} & \sigma_{v13} & \sigma_{v23} & \sigma_{v3}^2 \end{bmatrix}$$

set diagonal matrix
set full matrix
cancel

$$\text{cov} \begin{bmatrix} \text{resp}_{1jk} | \pi_{1jk} \\ \text{resp}_{2jk} | \pi_{2jk} \end{bmatrix} = \begin{bmatrix} \alpha \pi_{1jk} & \rho [\pi_{1jk} \pi_{2jk}]^{0.5} \\ \rho [\pi_{1jk} \pi_{2jk}]^{0.5} & \alpha \pi_{2jk} \end{bmatrix}$$

Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

By setting a diagonal matrix, the covariances among intercepts and/or slopes are all assumed to be equal to zero. Although a number of limitations might be considered to arise from this assumption, in the framework of the present demonstration it is adopted mainly for practical reasons (i.e. numerical instabilities and convergence problems were encountered in the full matrix consideration).

Consequently, the Equations window will now look like this:

Equations

$$\text{resp}_{1jk} \sim \text{Poisson}(\pi_{1jk})$$

$$\text{resp}_{2jk} \sim \text{Poisson}(\pi_{2jk})$$

$$\log(\pi_{1jk}) = \text{logepop}_{1jk} + \beta_{1k} \text{cons.accidents}_{ijk} + \beta_{2k} \text{alcohol.accidents}_{ijk}$$

$$\beta_{1k} = \beta_1 + v_{1k}$$

$$\beta_{2k} = \beta_2 + v_{2k}$$

$$\log(\pi_{2jk}) = \text{logepop}_{2jk} + \beta_{0k} \text{cons.killed}_{ijk} + \beta_{3k} \text{alcohol.killed}_{ijk}$$

$$\beta_{0k} = \beta_0 + v_{0k}$$

$$\beta_{3k} = \beta_3 + v_{3k}$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \\ v_{2k} \\ v_{3k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & & & \\ 0 & \sigma_{v1}^2 & & \\ 0 & 0 & \sigma_{v2}^2 & \\ 0 & 0 & 0 & \sigma_{v3}^2 \end{bmatrix}$$

$$\text{cov} \begin{bmatrix} \text{resp}_{1jk} | \pi_{1jk} \\ \text{resp}_{2jk} | \pi_{2jk} \end{bmatrix} = \begin{bmatrix} \alpha \pi_{1jk} & \\ \rho [\pi_{1jk} \pi_{2jk}]^{0.5} & \alpha \pi_{2jk} \end{bmatrix}$$

Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

- Click More to run the model. The following results are displayed:

Equations

$$\text{resp}_{1jk} \sim \text{Poisson}(\pi_{1jk})$$

$$\text{resp}_{2jk} \sim \text{Poisson}(\pi_{2jk})$$

$$\log(\pi_{1jk}) = \text{logepop}_{1jk} + \beta_{0k} \text{cons.accidents}_{ijk} + \beta_{2k} \text{alcohol.accidents}_{ijk}$$

$$\beta_{0k} = -6.472(0.037) + v_{0k}$$

$$\beta_{2k} = -0.024(0.005) + v_{2k}$$

$$\log(\pi_{2jk}) = \text{logepop}_{2jk} + \beta_{1k} \text{cons.killed}_{ijk} + \beta_{3k} \text{alcohol.killed}_{ijk}$$

$$\beta_{1k} = -8.380(0.026) + v_{1k}$$

$$\beta_{3k} = -0.005(0.002) + v_{3k}$$

$$v \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.052(0.014) & 0 & 0 & 0 \\ 0 & 0.009(0.007) & 0 & 0 \\ 0 & 0 & 0.000(0.000) & 0 \\ 0 & 0 & 0 & 0.000(0.000) \end{bmatrix}$$

$$\text{cov} \begin{bmatrix} \text{resp}_{1jk} | \pi_{1jk} \\ \text{resp}_{2jk} | \pi_{2jk} \end{bmatrix} = \begin{bmatrix} 14.575(1.542) \pi_{1jk} & 3.667(0.564) [\pi_{1jk} \pi_{2jk}]^{0.5} \\ 3.667(0.564) [\pi_{1jk} \pi_{2jk}]^{0.5} & 3.344(0.339) \pi_{2jk} \end{bmatrix}$$

File Edit View Help

Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

In order to display more decimals in the values of the variance matrix:

- In the Options main menu, select Numbers(display precision and missing value code)
- In the Settings window that appears, set digits after decimal point equal to 5 and click Apply. Then click Done.

Settings

Worksheet Numbers Directories

Specify numerical precision to be used when displaying numbers:

☒ bdp/adp format ☐ signif digit format

digits before decimal point 4

digits after decimal point 5

exponent ☐

Set all values of to be missing

Help Apply Done Cancel

The Equations window should now look like this:

Equations

$$\begin{aligned} \text{resp}_{1jk} &\sim \text{Poisson}(\pi_{1jk}) \\ \text{resp}_{2jk} &\sim \text{Poisson}(\pi_{2jk}) \\ \log(\pi_{1jk}) &= \text{logepop}_{1jk} + \beta_{1k} \text{cons.accidents}_{ijk} + \beta_{2k} \text{alcohol.accidents}_{ijk} \\ \beta_{1k} &= -6.47545(0.03752) + v_{1k} \\ \beta_{2k} &= -0.02514(0.00524) + v_{2k} \\ \log(\pi_{2jk}) &= \text{logepop}_{2jk} + \beta_{0k} \text{cons.killed}_{ijk} + \beta_{3k} \text{alcohol.killed}_{ijk} \\ \beta_{0k} &= -8.38056(0.02621) + v_{0k} \\ \beta_{3k} &= -0.00415(0.00246) + v_{3k} \end{aligned}$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \\ v_{2k} \\ v_{3k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.01039(0.00727) & 0 & 0 & 0 \\ 0 & 0.05311(0.01430) & 0 & 0 \\ 0 & 0 & 0.00038(0.00021) & 0 \\ 0 & 0 & 0 & 0.00001(0.00002) \end{bmatrix}$$

$$\text{cov} \begin{bmatrix} \text{resp}_{1jk} | \pi_{1jk} \\ \text{resp}_{2jk} | \pi_{2jk} \end{bmatrix} = \begin{bmatrix} 14.57073(1.54799) \pi_{1jk} & 3.66116(0.56525) [\pi_{1jk} \pi_{2jk}]^{0.5} \\ 3.66116(0.56525) [\pi_{1jk} \pi_{2jk}]^{0.5} & 3.32403(0.33863) \pi_{2jk} \end{bmatrix}$$

Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help

The fixed effect of enforcement on the number of accidents is higher compared to the related effect on persons killed. The regional variation of the effect of alcohol enforcement effects is only significant as far as the number of accidents is concerned. In particular, the effect of alcohol controls on persons killed does not vary significantly among regions. These results are further interpreted in section 2.5 of the Methodology report.

2.6 Structural equations models

As explained in the methodology report, the application of a structural equation model requires large amounts of data and also of a certain quality. In practice, these models are only applied in studies that have been planned to produce data suitable for this type of analysis. Moreover, the conduction of structural equation modelling requires a high level of expertise, which to deliver would exceed the scope of the present document. A number of good tutorials, exclusively dedicated to structural equation modelling are on the market (c.f. www.ssicentral.com). Consequently it was refrained from presenting the practical instruction for recapitulating the example in the methodology report.

2.7 More complex data structures

In the respective section of the Methodology report (D7.4), several particular cases of multilevel models, mainly referred to as "non-hierarchical" models were presented (i.e. cross-classified structures, multiple membership structures). These issues were mainly presented for completeness' sake and no practical examples were provided; non-hierarchical structures are seldom encountered in road safety research

2.8 Bayesian estimation in multivlevel modelling

Estimation methods based on simulation techniques (i.e. Monte Carlo Markov Chain methods, bootstrap methods) for fitting these models (and multilevel models in general) were presented. The models for dealing with these structures are still under further development. Moreover, a detailed presentation advanced estimation methods (e.g. simulation techniques) is beyond the scope of this document. Consequently, no manuals are provided for this section.

Chapter 3 - Time series analysis

3.1 Introduction to time series models

Ellen Berends and Frits Bijleveld (SWOV)

In the SafetyNet project, many road traffic data are collected that consist of *repeated measurements over time*. Examples are the annual or monthly number of road traffic accidents in a country, its annual or monthly number of road traffic fatalities, its annual or monthly number of vehicle kilometres driven, its annual or monthly values on safety performance indicators, etc., all repeatedly measured over a certain period of time. Whenever one is interested in studying and analysing such developments of one and the same phenomenon over time, special issues arise not encountered in cross-sectional data analysis. An important issue is that the residuals, although assumed to be independent in the (cross-sectional) model specification, as demonstrated in the methodology report may in fact not be independent of one another. This violation may result in unreliable test statistics, and thus unreliable inferences from the models.

The problem of dependencies between the residuals in the traditional linear regression analysis of time series data may sometimes be solved in a number of different ways:

- additional predictor variables can be added to the regression of the dependent variable on time such that the dependencies are removed from the residuals, and/or
- the relation between the dependent variable and time can be analysed with generalised linear models and/or non-linear models, and/or
- the dependent variable can be analysed with a special family of analysis techniques collectively known as *time series models*. The most common dedicated time series analysis techniques used in road safety analysis are ARIMA, its special case DRAG and state space models.

In the manual, we only deal with the dedicated time series methods. As in principle the DRAG method can be regarded as a special case of ARMA-type modelling, this approach is also not covered. However, linear regression model is included because it is used in the methodology report to demonstrate the identification and consequences of dependency of residuals and it is well known. Because linear regression is so well known, it is assumed to be a better starting point than dedicated time series analysis methods, namely ARMA-type models and state space models, which are discussed as well.

The first part of the chapter on Time Series Analysis shows when linear regression can be used for repeated measurements over time. It is recommended to read this chapter before starting with one of the dedicated time series methods. Austrian fatalities from the period from 1987 to 2004 was used as example to show which tests have to be carried out to test the Gauss-Markov assumptions, which are the conditions for linear regression. Linear

regression is not the preferred model for the Austrian fatalities due to heavy violations of two basic assumptions.

In the parts about ARMA-type models and state space analysis in the methodology report, several data series were used as examples of using the dedicated time series analysis techniques. We name a few interesting examples. An ARMA-type analysis was conducted on the monthly total number of *French fatalities* collected between 1975 and 2001. It was shown that next to various seasonal and economic variables, the number of fatalities is also affected by certain media events. Furthermore, the presidential amnesty that is usually given to traffic offenders during the French elections appeared to be associated to an increase in fatalities. A state space type of analysis was carried out on the monthly number of *drivers killed or seriously injured (KSI)* in the United Kingdom. It was shown that the number of KSI depends on the introduction of the seat belt law, the petrol price (as a indicator of mobility) and seasonal influences. The introduction of the seat belt law resulted in a 21.1% reduction of the number KSI in the UK.

Numerous software packages can be used to carry out the above mentioned analysis techniques. The choice of one software package for this manual does not mean that the user, after having worked with this manual, could not apply this type of analysis in other software environments or new versions of the same software. The software is just used as a means to demonstrate the opportunities that dedicated time series modelling offers to road safety analysis and to instruct in the design and interpretation of classical linear regression, ARMA-type models or state space models.

For linear regression and ARMA-type models, SPSS was retained because it is a mainstream statistics program and it is very suitable for these analysis techniques, in addition to being a user-friendly software (<http://www.spss.com>). Other existing dedicated softwares, such as E-views, R and SAS Proc ARIMA are also appropriate for performing ARMA-type analysis).

The three probably best-known software packages which can be applied for state space analysis, i.e. Ox/SsfPack, STAMP, and SAS Proc UCM, while noting the availability in other packages, such as Splus, R, matlab and Mathematica, will be shortly compared below. SsfPack contains a lot of routines for state space analysis, and can be used in the Ox programming environment. By programming the user has a lot of freedom in modelling. A disadvantage of Ox/SsfPack is that it requires in addition to experience in statistics and modelling, some experience in programming. STAMP is a user friendly, menu-driven package specifically designed for state space analysis and is therefore more appropriate for instructing the possibly inexperienced road safety analyst in state space modelling. SAS Proc UCM can handle univariate models and in the future multivariate models as well (Yaffee, 2003), whereas STAMP 6.0 handles both univariate and multivariate models. Yaffee (2003) states that SAS Proc UCM is "powerful and easy to use" and that "STAMP handles a wide variety of models". Furthermore, STAMP has good graphical options, can display forecasts with error margins, and its algorithms are fast. Judge and Ninomiya (2000) make the following remark, which is very relevant in the light of the

manual's objective: "even those who are inexperienced with structural time series modelling can use STAMP to familiarise themselves with this approach". It was this, its relatively small size (and price), and its dedicatedness to state space analysis which made us choose STAMP for structural time series modelling.

3.2 Classical linear and non-linear regression models

3.2.1 Classical linear regression models

Christian Brandstätter and Andrea Angermann (KfV)

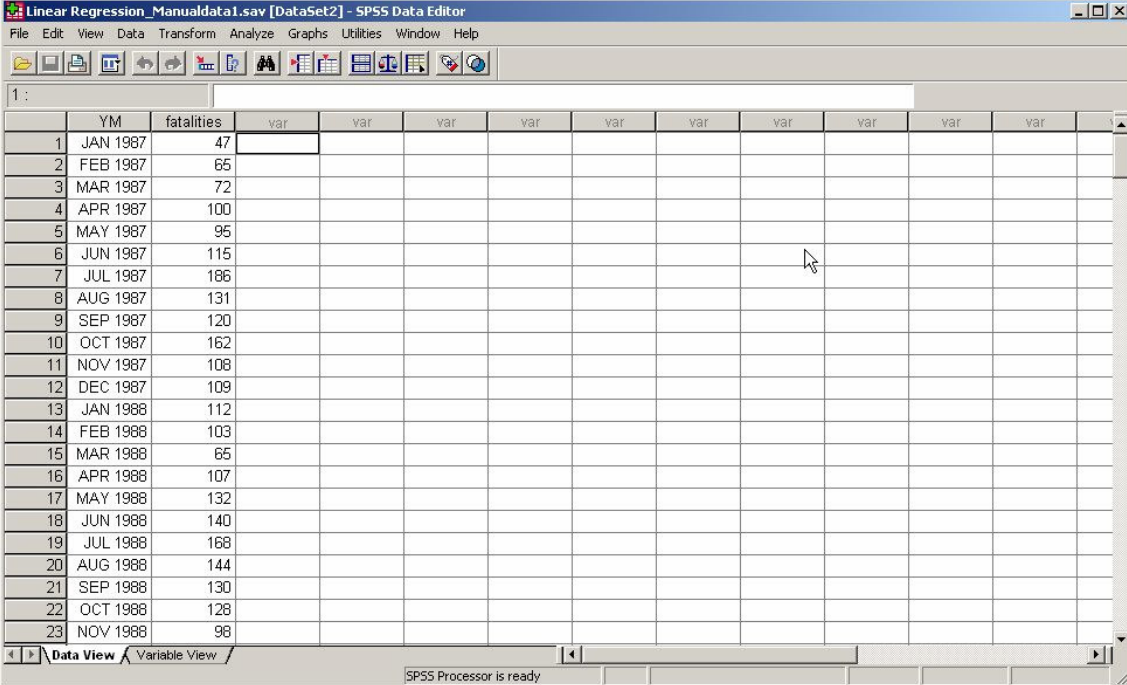
3.2.1.1. Introduction

The goal of this chapter is to demonstrate how to apply linear regression models to Austrian road accident fatalities data and how to determine if a trend in the counts of road accident fatalities can be derived. Some practical computations using SPSS software (version 14.0) on these data are presented. However, the different views on the assumptions underlying data testing are considered more important than the resulting regression line with its parameters.

Screenshots and descriptions are used to explain the steps and results in the process of data analysis; for more detailed information on the theoretical background, please see the corresponding theorie chapter. For further explanations regarding SPSS, we refer the reader to the SPSS user manuals.

3.2.1.2. Dataset description

Austrian data from the period from 1987 to 2004 is analyzed in this tutorial. The raw dataset imported to SPSS is shown in the following table. The time variable “Y/M” and number of “Fatalities” can be seen.



Linear Regression_Manualdata1.sav [DataSet2] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

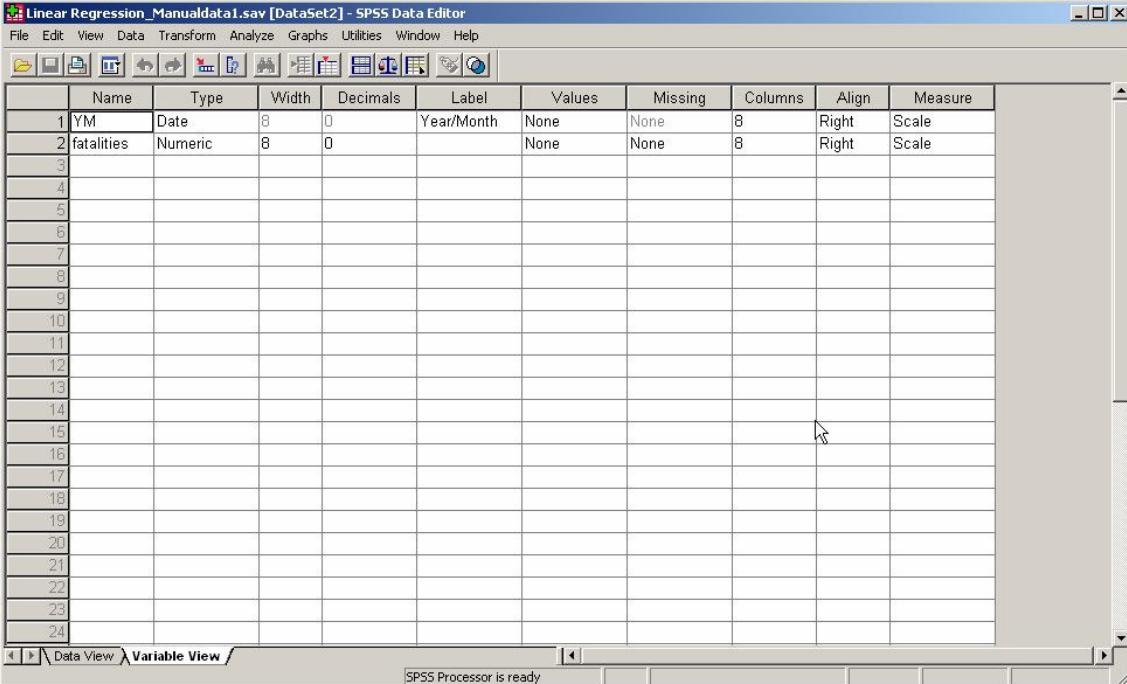
1 :

	YM	fatalities	var	var	var	var	var	var	var	var	var	var
1	JAN 1987	47										
2	FEB 1987	65										
3	MAR 1987	72										
4	APR 1987	100										
5	MAY 1987	95										
6	JUN 1987	115										
7	JUL 1987	186										
8	AUG 1987	131										
9	SEP 1987	120										
10	OCT 1987	162										
11	NOV 1987	108										
12	DEC 1987	109										
13	JAN 1988	112										
14	FEB 1988	103										
15	MAR 1988	65										
16	APR 1988	107										
17	MAY 1988	132										
18	JUN 1988	140										
19	JUL 1988	168										
20	AUG 1988	144										
21	SEP 1988	130										
22	OCT 1988	128										
23	NOV 1988	98										

Data View Variable View

SPSS Processor is ready

The option “Variable View” displays all definitions and attributes of the used variables in the data set:



Linear Regression_Manualdata1.sav [DataSet2] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	YM	Date	8	0	Year/Month	None	None	8	Right	Scale
2	fatalities	Numeric	8	0		None	None	8	Right	Scale
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										

Data View Variable View

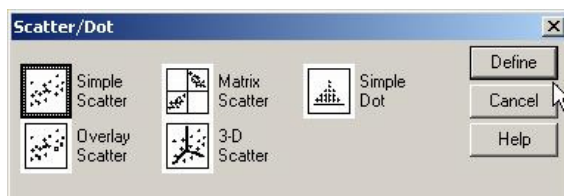
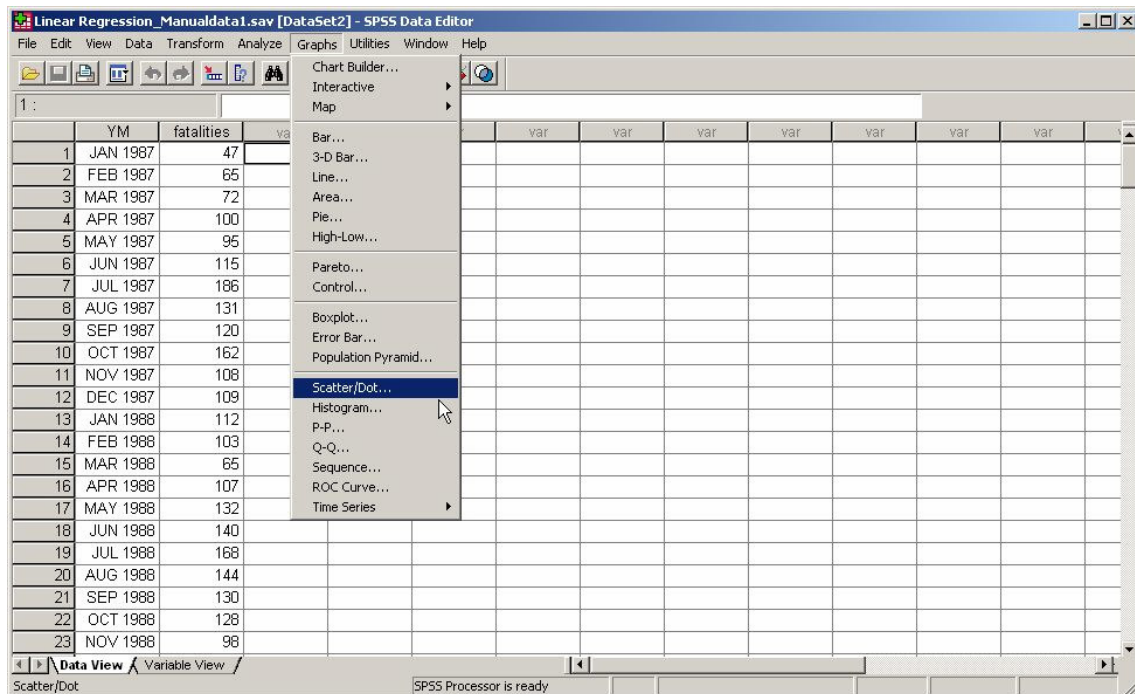
SPSS Processor is ready

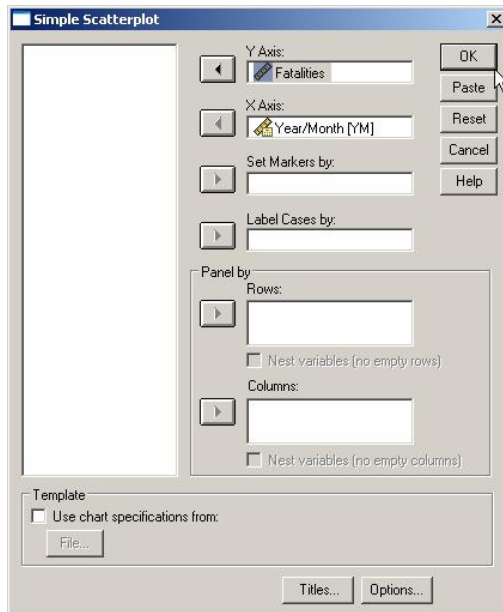
3.2.1.3. First heuristic view of data

In order to get a first impression of the time series data, it is recommended to generate a simple scatter plot of fatalities over time. In the following screenshots the necessary steps for this procedure in SPSS are being

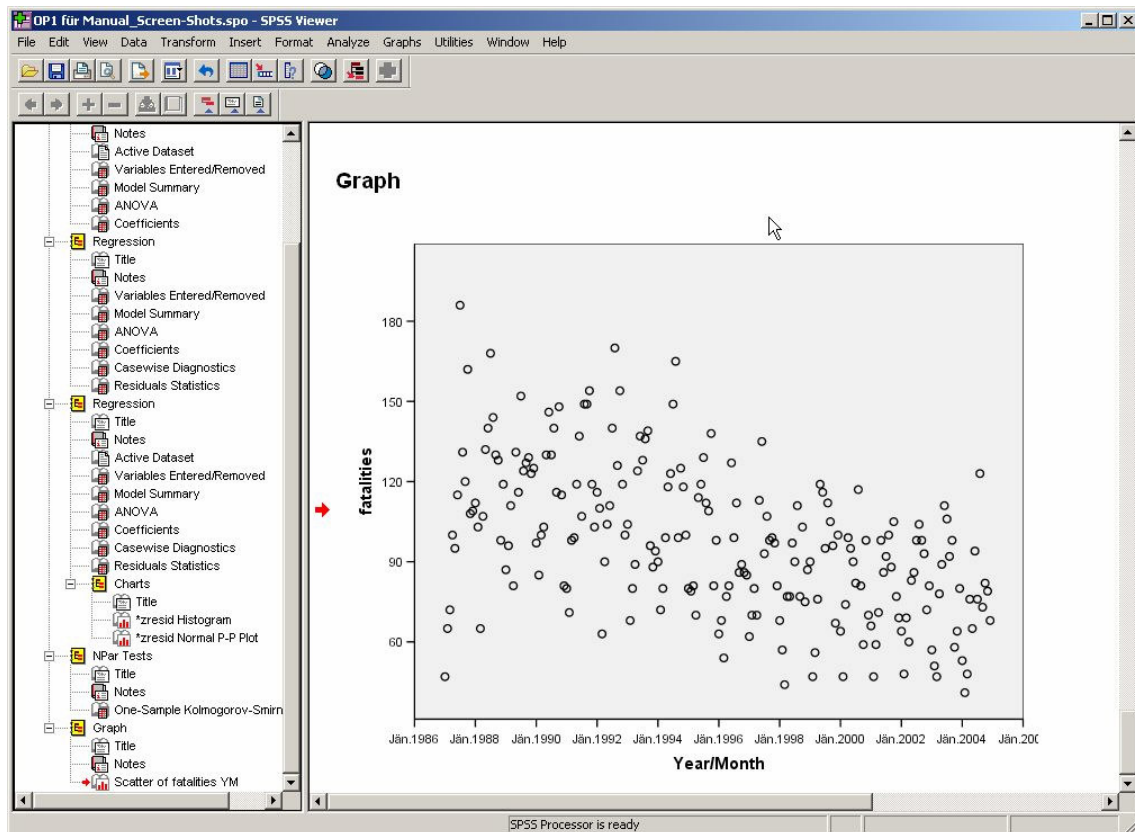
explained. Simple scatter plot has been chosen because the data only contains one data series.

Choose from the menu “Graphs” the option “Scatter/Dot...”. After choosing a “Simple Scatter”, click “Define”. Mark and click the variable “Year/Month” in the X-axis, the variable “Fatalities” in the Y-axis. After clicking “Ok”, the SPSS-processor starts with producing the scatter graph which is shown in Output 1.



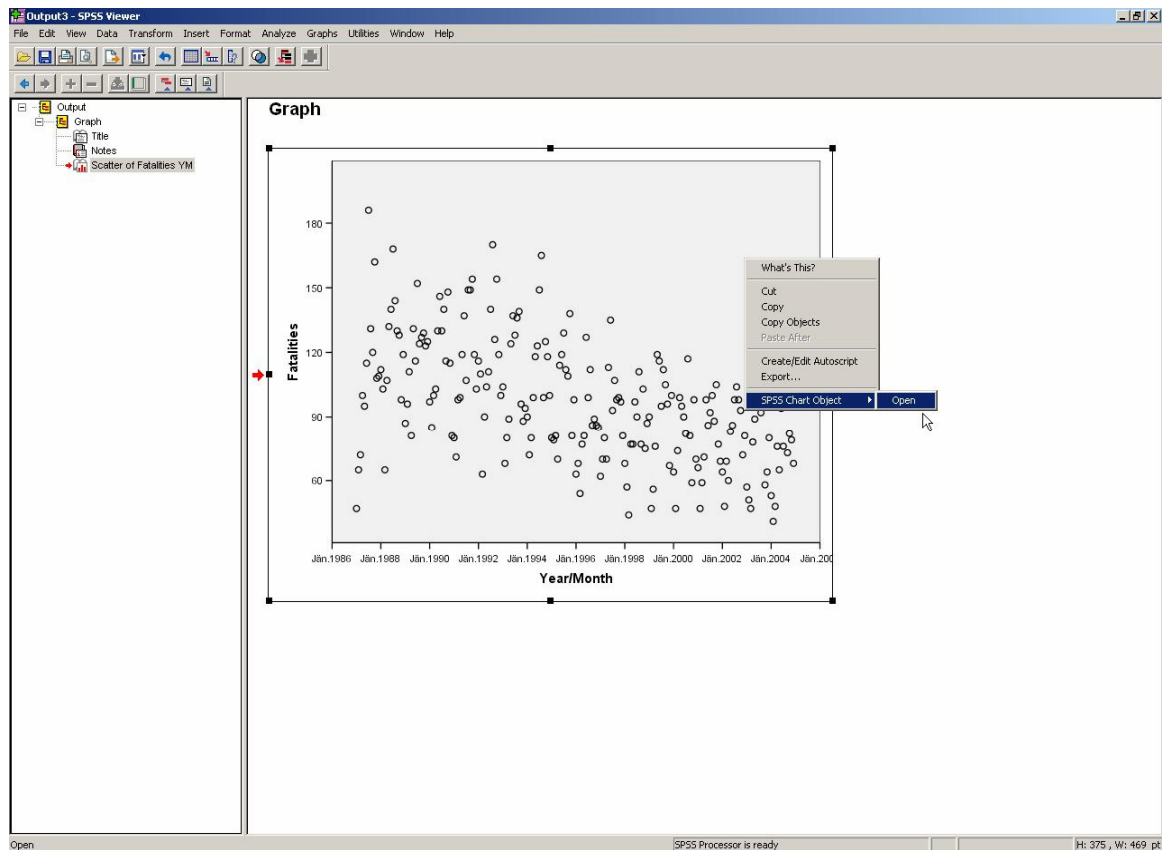


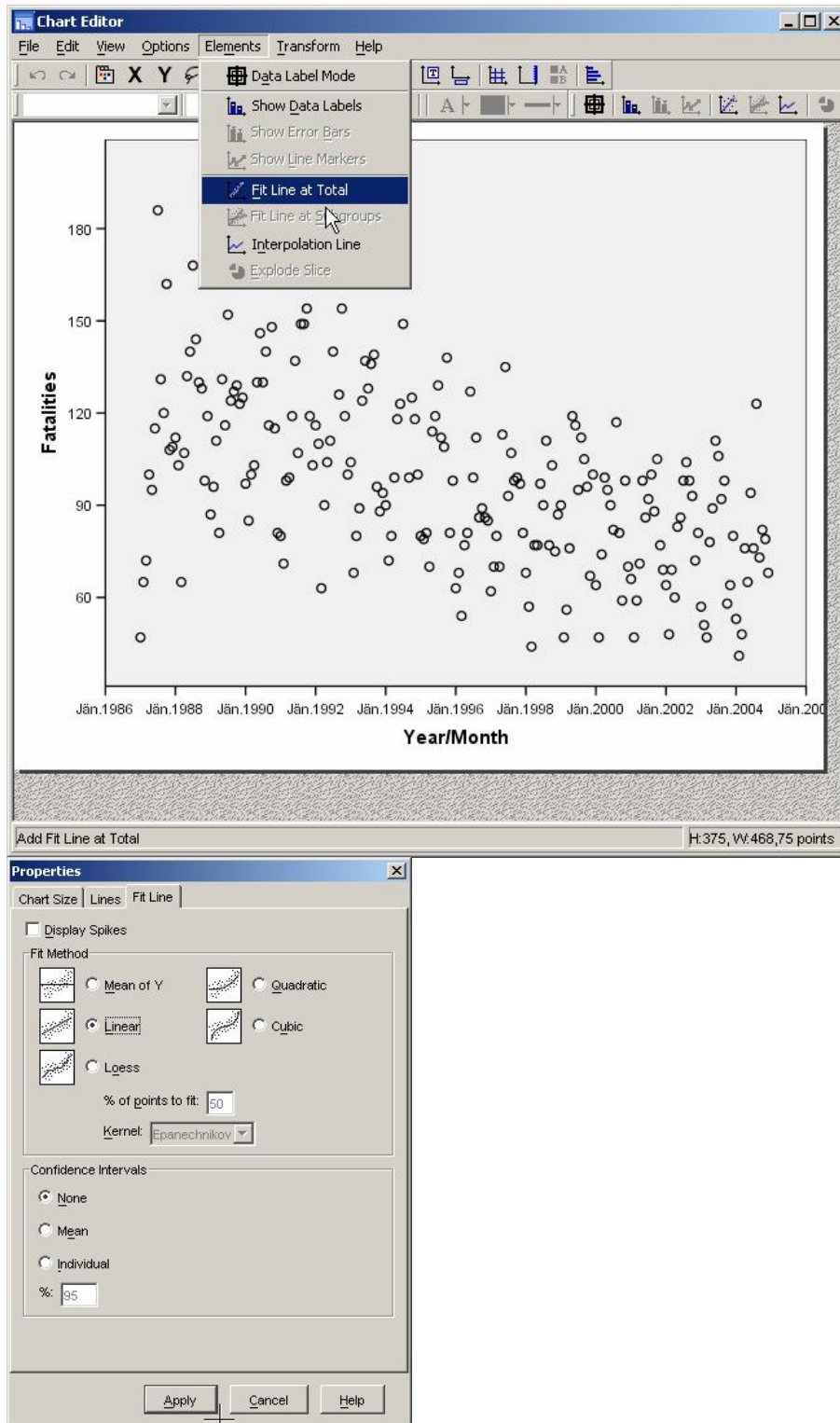
The figure displayed below shows the structure of data at a glance: a decreasing development of counts of fatalities over time.

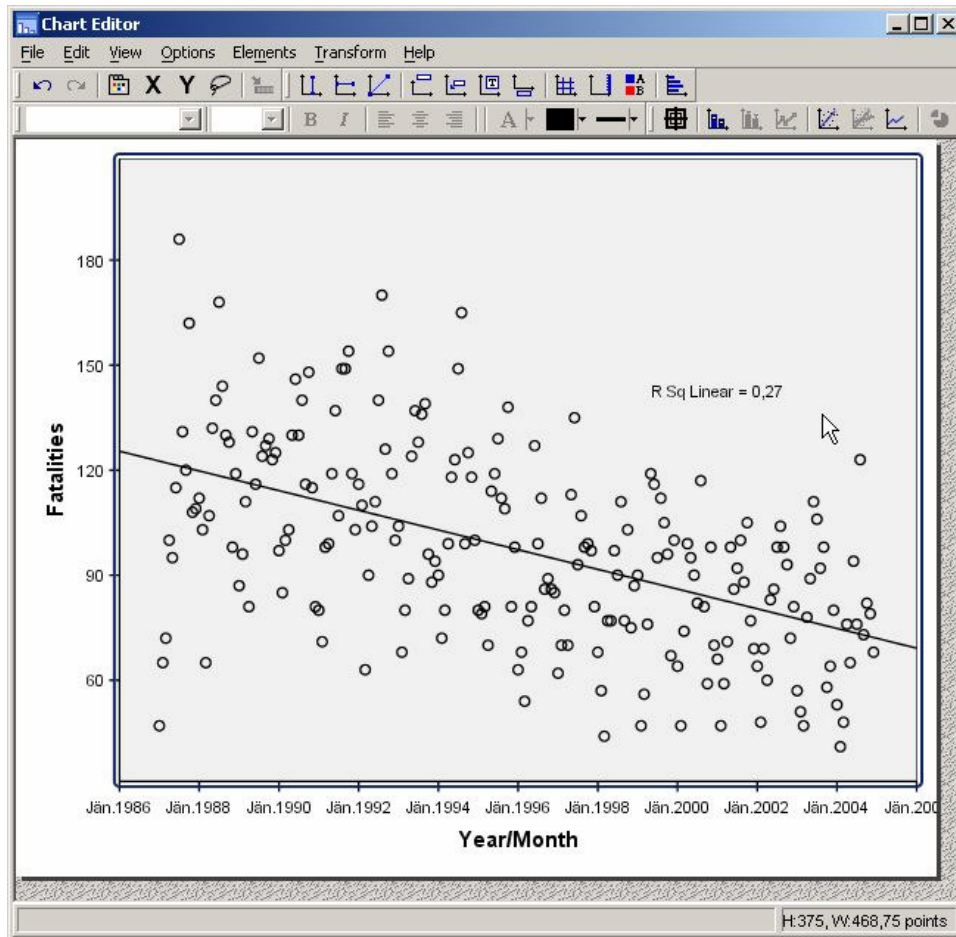


A first attempt to describe this development mathematically could be carried out by fitting a simple linear regression. The next step to get a better impression of the data is to generate a linear regression line in the scatter plot.

For this step click on the graph, click the right mouse button and open the “SPSS Chart Object”. In the now new opened window “Chart Editor”, you choose “Elements” and click “Fit Line at Total”. In the new opened window “Properties” you choose “Linear” and click “Apply”. As a result you can see the linear regression line in the chart and you can close the “Chart Editor” in order to return to the “Output”-window.







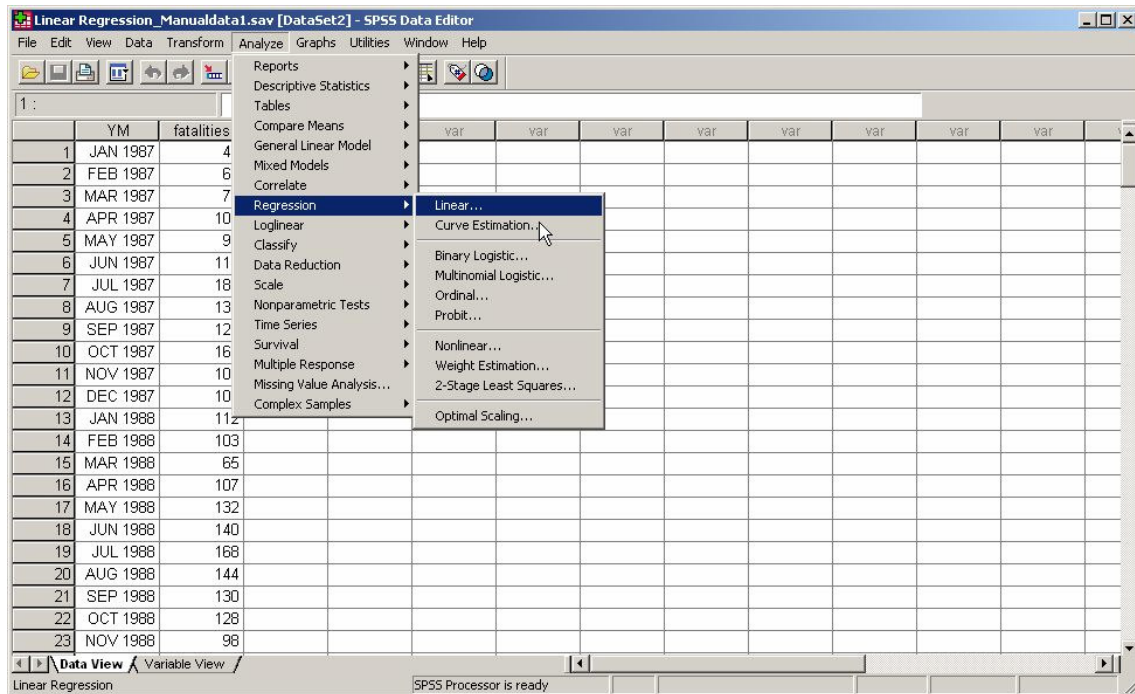
The scatter plot above may be roughly interpreted with the help of a linear regression model. At this stage of data processing there is no reason for not applying this model; therefore next steps and analyses should be started.

3.2.1.4. Linear regression:

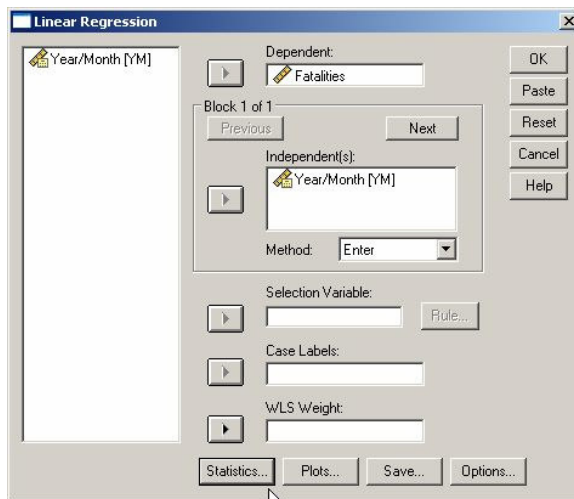
The linear regression is generated with road accident “Fatalities” in Austria as dependent variable and “Year/Month” as independent variable. The time series starts in January 1987 and ends in December 2004.

The next figures show the necessary operations for calculating a linear regression:

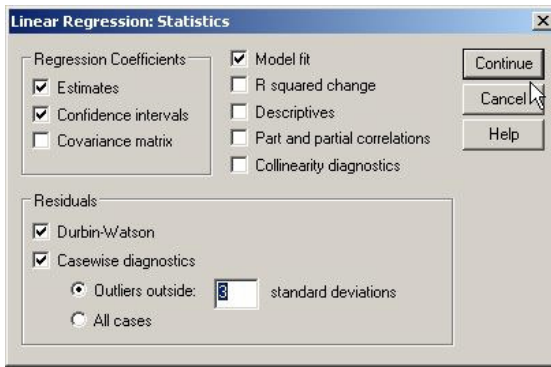
Choose from the menu “Analyze” “Regression” and click “Linear”.



Then mark and click the variable “Fatalities” in the “Dependent” field, the variable “Year/Month” you click in the “Independent” field. Then click the button “Statistics”:



In the “Linear Regression: Statistics”-window choose the regression coefficients “Estimates” and “Confidence intervals”, as well as the option “Model fit” and the residuals “Dubin-Watson” and “Casewise diagnostics” [1]. Define the “Outliers outside:” with “3” standard deviations and click “Continue”.



The result (see the Output-window stated below) shows a highly significant decrease in the count of fatalities; this is explained in the coefficients-table:

- The variable “Year/Month” is negative (-0.520 for the standardized coefficient Beta, the unstandardized coefficients can be used on the original data with formula 3.2.1 in the corresponding chapter in the methodology report) and shows a very “high” significance of < 0.0001 : a significance value of less than 0.05 means that the variation explained by the model is not due to change.

Furthermore, the regression model has a reasonable fit:

- The ANOVA table reports a significant F statistic because its significance value is below 0.05.
- R Square in the Model Summary table shows that the regression explains 27.0% of the variance of the data.

SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Output

- Graph
 - Title
 - Notes
 - Scatter of Fatalities YM
- Regression
 - Title
 - Notes
 - Variables Entered/Removed
 - Model Summary
 - ANOVA
 - Coefficients
 - Casewise Diagnostics
 - Residuals Statistics
- Charts
 - Title
 - Notes
 - *zresid Histogram
 - *zresid Normal P-P Plot
- NPar Tests
 - Title
 - Notes
 - One-Sample Kolmogorov-Smirnov
- Graph
 - Title
 - Notes
 - Scatter of fatalities YM

Regression

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	Year/Month ^a	.	Enter

a. All requested variables entered.
b. Dependent Variable: Fatalities

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,520 ^a	,270	,267	24,130	,953

a. Predictors: (Constant), Year/Month
b. Dependent Variable: Fatalities

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	46129,857	1	46129,857	79,228	,000 ^a
	Residual	124600,5	214	582,245		
	Total	170730,3	215			

a. Predictors: (Constant), Year/Month
b. Dependent Variable: Fatalities

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta				Lower Bound	Upper Bound
1	(Constant)	1259,417	130,558			9,646	,000	1002,072	1516,761
	Year/Month	-8,9E-008	,000	-,520		-8,901	,000	,000	,000

a. Dependent Variable: Fatalities

SPSS Processor is ready

The residual analysis in the table Casewise Diagnostics shows one outlier with case "1" in January 1987. In this stage of the calculations/analysis no assumptions can be made why case "1" is an exceptional case. Presently, no further analysis on the cause of this outlier is being made.

SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Output

- Graph
 - Title
 - Notes
 - Scatter of Fatalities YM
- Regression
 - Title
 - Notes
 - Variables Entered/Removed
 - Model Summary
 - ANOVA
 - Coefficients
 - Casewise Diagnostics
 - Residuals Statistics
- Charts
 - Title
 - Notes
 - *zresid Histogram
 - *zresid Normal P-P Plot
- NPar Tests
 - Title
 - Notes
 - One-Sample Kolmogorov-Smirnov
- Graph
 - Title
 - Notes
 - Scatter of fatalities YM

Casewise Diagnostics^a

Case Number	Std. Residual	Fatalities	Predicted Value	Residual
1	-,3133	47	122,60	-,75,602

a. Dependent Variable: Fatalities

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	72,21	122,60	97,41	14,648	216
Residual	-,75,602	64,791	,000	24,074	216
Std. Predicted Value	-,1720	1,720	,000	1,000	216
Std. Residual	-,3133	2,685	,000	,998	216

a. Dependent Variable: Fatalities

SPSS Processor is ready

3.2.1.5. Condition testing (Gauss-Markov Assumptions)

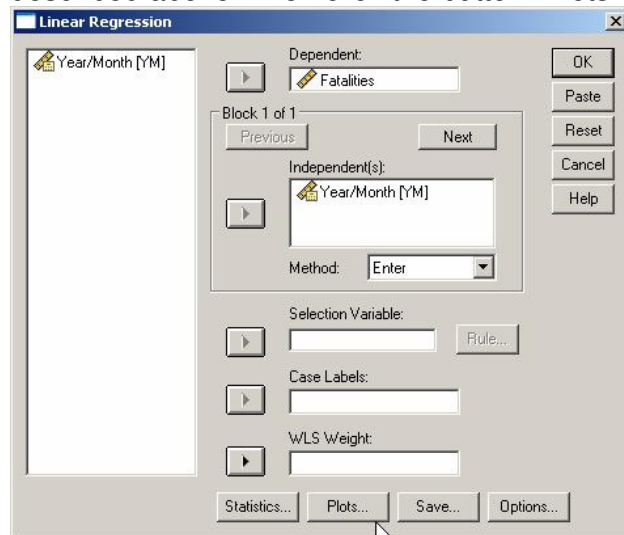
The testing of the Gauss-Markov assumptions can be done by clicking “Plots” in the linear regression. The underlying assumptions are [1]:

- The random errors are distributed normally.
- 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.
- 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 error. An error term with non-constant variance is said to be heteroscedastic.
- The prediction error ε is uncorrelated with x , the independence assumption. This assumption is fulfilled when dealing with road accident time series. As we are dealing with univariate data in this example the problem of co-linearity is not relevant.

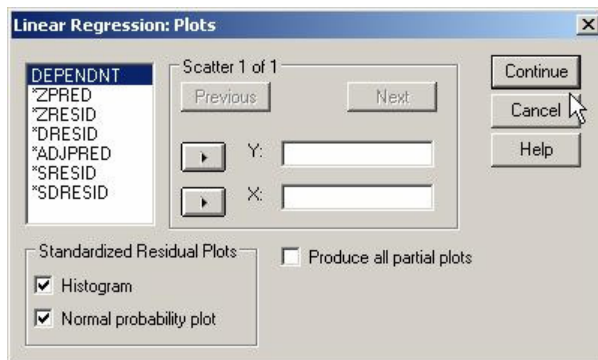
For all these assumptions visual and numeric representations are being generated (based on statistic inference). The statistical tests give detailed information whether the statement is valid or not. The advantage of the graphical analysis is that deviations and type of deviations from the tested conditions/assumptions can be detected more easily.

Normal distribution of random error

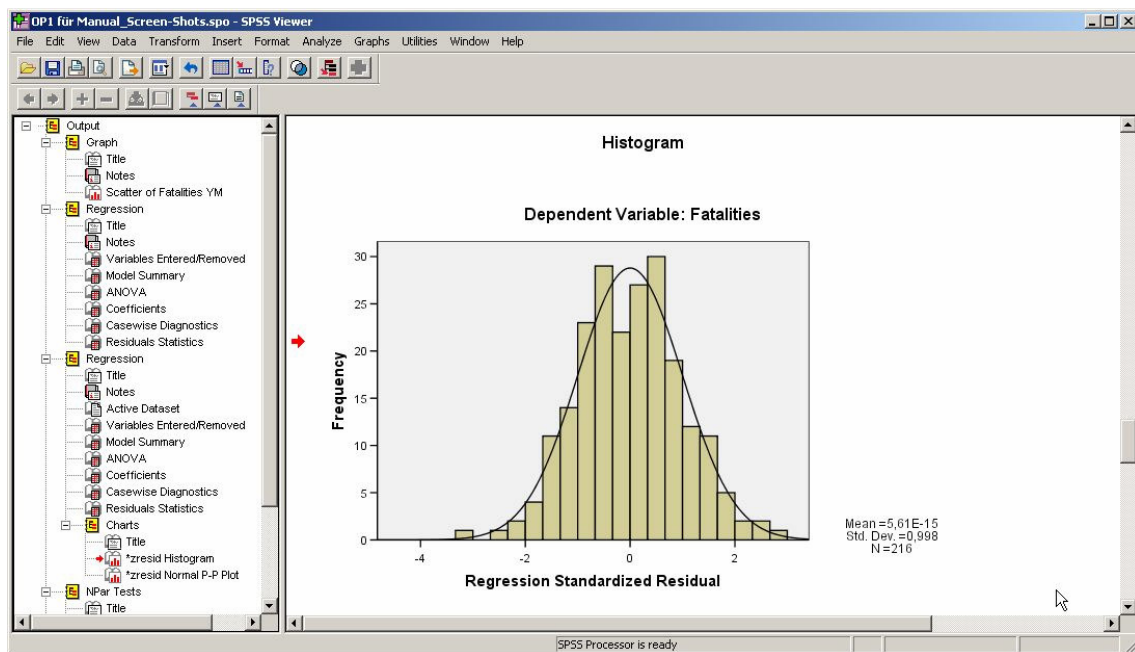
For a simple overview of the distribution of the variables, the graphical representation can be used. Choose the linear regression from the menu “Analyze” and choose the dependent and independent variable respectively, as described above. Then click the button “Plots”.

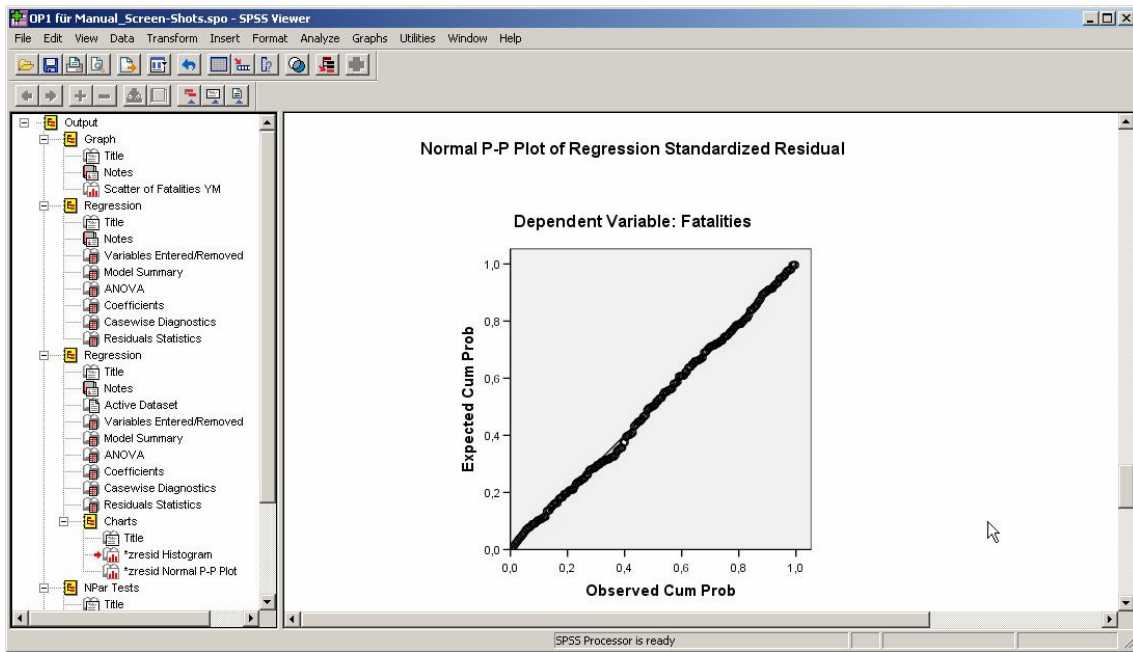


A new window “Linear Regression: Plots” will be opened. Choose for the “Standardized Residual Plots” the options “Histogram” and “Normal Probability Plot”, and then click the button “Continue”:



The results you can see in the “Output”-window:





The assumption of a normal distribution of random error can be confirmed with the two graphs stated above.

For the numeric testing the 1-Sample Kolmogorov-Smirnov test [1] has been used, this is an inference statistic using a non-parametric test. If the Kolmogorov-Smirnov test is not significant, the assumption of a normal distribution of random error can be confirmed.

To start this test, choose from the menu “Analyze” the “Nonparametric Test” “1 Sample K-S”. You can see the result in the Output-Window.

Linear Regression_Manualdata_forPrints1.sav [DataSet1] - SPSS Data Editor

	YM	Fatalities	ZPR_1	ZRE_1	ZRE2	ZRE_ABS	var	var	var
1	JAN 1987	4	1,71974	-3,13316	9,82	3,13			
2	FEB 1987	6	1,70345	-2,37730	5,65	2,38			
3	MAR 1987	7	1,68873	-2,07827	4,32	2,08			
4	APR 1987	10	1,67243	-,90798	,82	,91			
5	MAY 1987	9	1,65666	-1,10562	1,22	1,11			
6	JUN 1987	11	1,64036	-,26688	,07	,27			
7	JUL 1987	18	1,62459	2,68512	7,21	2,69			
8	AUG 1987	13							
9	SEP 1987	12							
10	OCT 1987	16							
11	NOV 1987	10							
12	DEC 1987	10							
13	JAN 1988	112							
14	FEB 1988	103							
15	MAR 1988	65							
16	APR 1988	107							
17	MAY 1988	132							
18	JUN 1988	140							
19	JUL 1988	168							
20	AUG 1988	144							
21	SEP 1988	130							
22	OCT 1988	128							
23	NOV 1988	98							
24	DEC 1988	119							
25	JAN 1989	87							
26	FEB 1989	96							
27	MAR 1989	111							

OP1 for Manual_Screen-Shots.spo - SPSS Viewer

NPar Tests

One-Sample Kolmogorov-Smirnov Test

	Year/Month	Fatalities
N	216	216
Normal Parameters a,b		
Mean	DEC 1995	97,41
Std. Deviation	1902 07:02	28,180
Most Extreme Differences	Absolute	,059
	Positive	,059
	Negative	-,059
Kolmogorov-Smirnov Z		,869
Asymp. Sig. (2-tailed)		,437

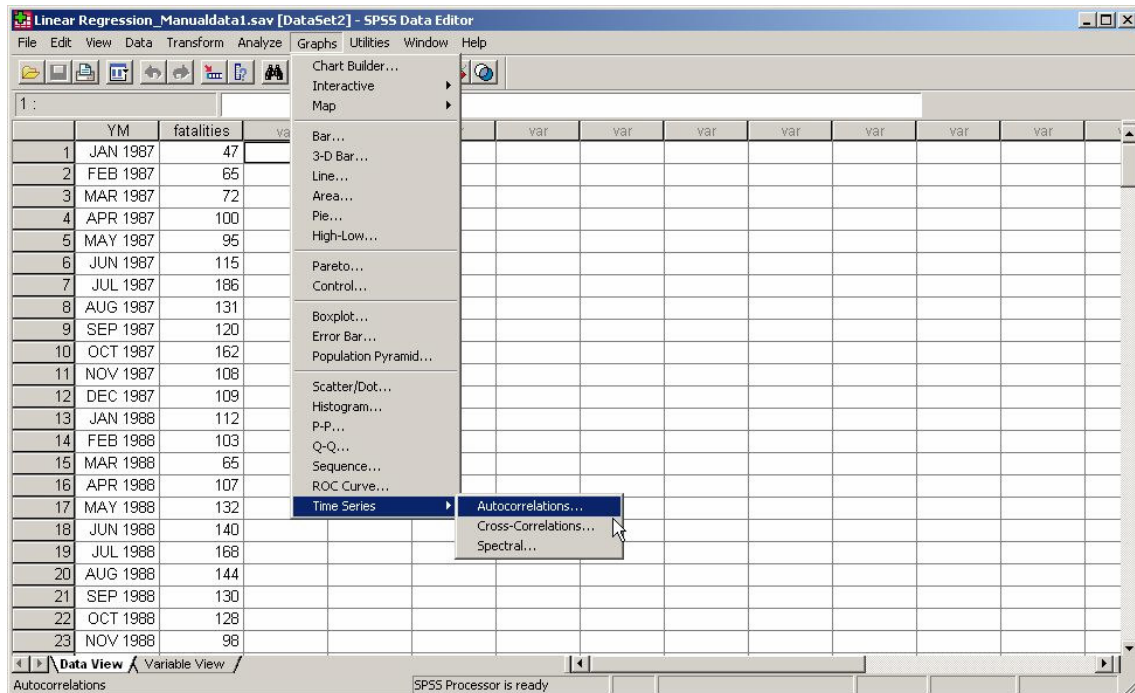
a. Test distribution is Normal.
b. Calculated from data.

The result stated above shows that also the not significant Kolmogorov-Smirnov test (Asymp. Sig. (2-tailed) = 0.316) confirms the normal distribution of the random error.

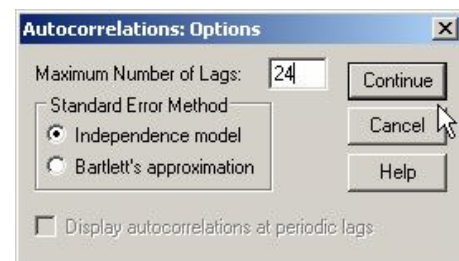
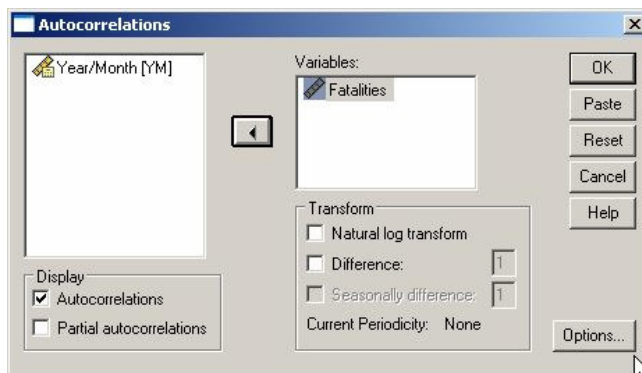
Independency of Variables

The autocorrelation function (ACF) is used to verify the assumption that the error term associated with any different observations is independent of any other. For these computations both a graphical and a numeric method exists.

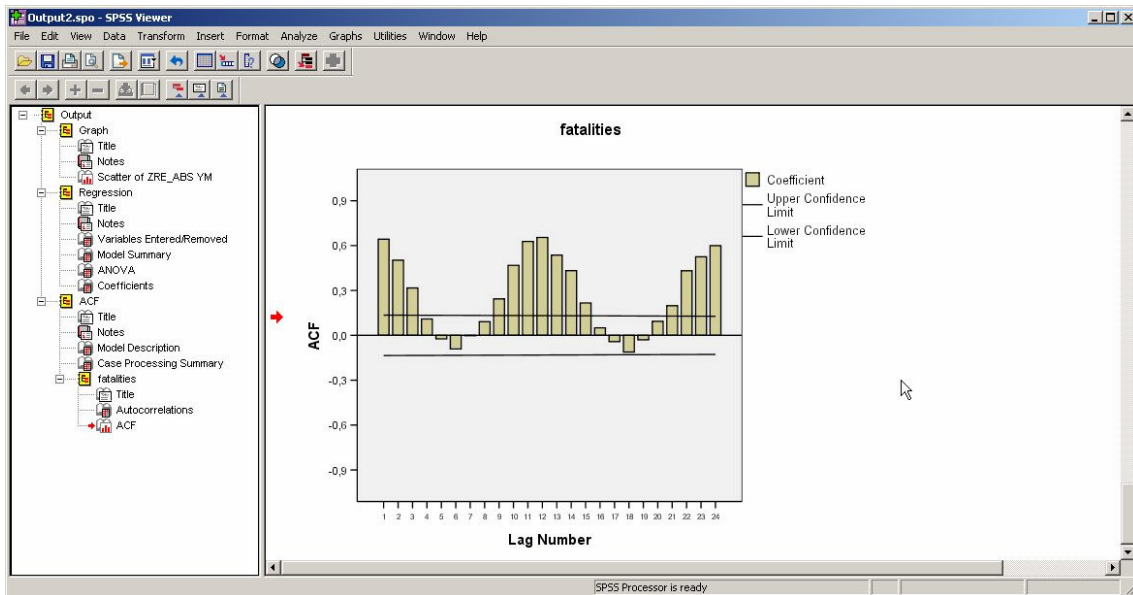
To start the computation of the autocorrelation function, choose the option “Time Series” from the “Graphs”-menu and click “Autocorrelation”.



The variable you choose is the dependent one (“Fatalities”): click in the “Autocorrelation”-window the button “Options” and define there the maximum number of lags to “24”. Close this window with “Continue” and start the analysis in the “Autocorrelation” window with the button “OK”.



A clear seasonal pattern of the autocorrelations can be identified in the diagram stated below with large peaks at 12 and 24 month. this pattern also exceeds the confidence band. Through this test the assumption that the error term associated with any different observations is independent of any other, can not be confirmed.

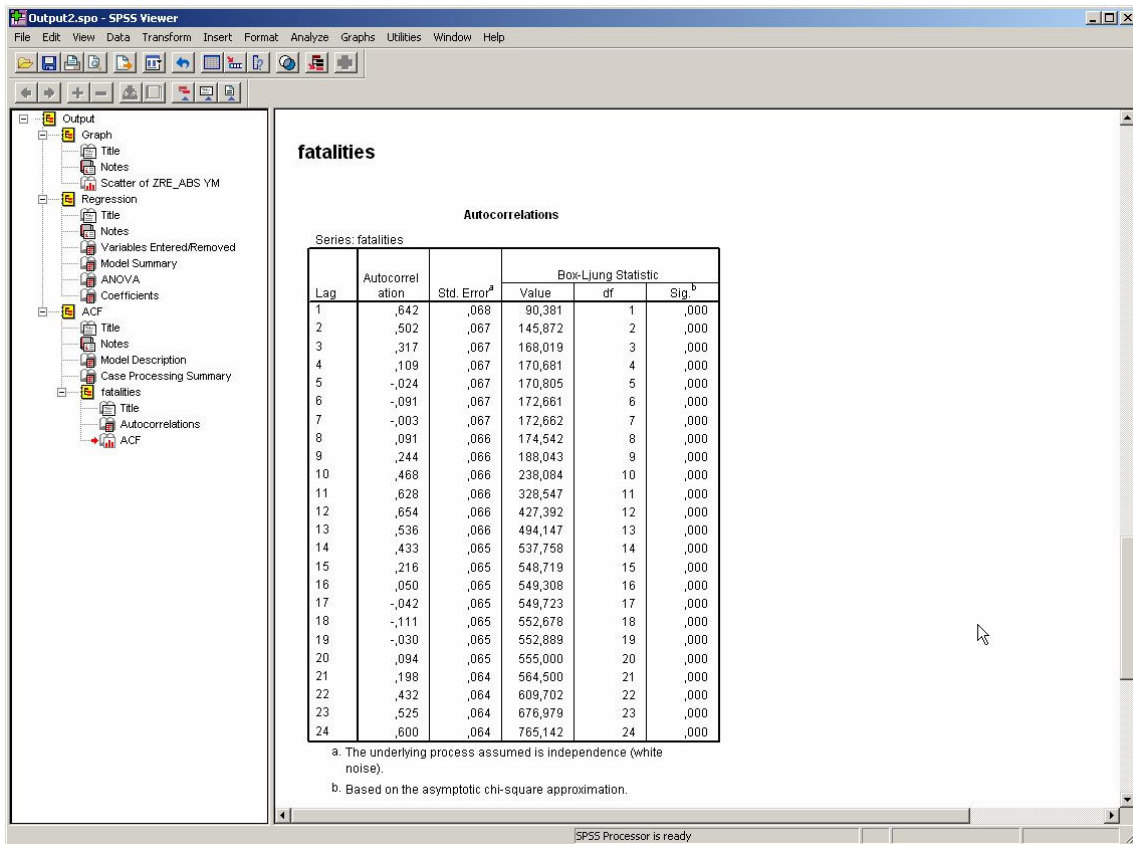


The screenshot shows the SPSS Viewer window with the 'ACF' table selected in the left-hand tree. The table provides a detailed description of the model used for the ACF analysis.

Model Description	
Model Name	MOD_1
Series Name	1 fatalities
Transformation	None
Non-Seasonal Differencing	0
Seasonal Differencing	0
Length of Seasonal Period	No periodicity
Maximum Number of Lags	24
Process Assumed for Calculating the Standard Errors of the Autocorrelations	Independence(white noise)
Display and Plot	All lags

Applying the model specifications from MOD_1

The same result is obtained with the following Box-Ljung-statistics, which is part of the executed ACF and is also presented in the Output window below. The Box-Ljung test tests the significance of autocorrelation at each lag. All 24 lags show a highly significant autocorrelation.

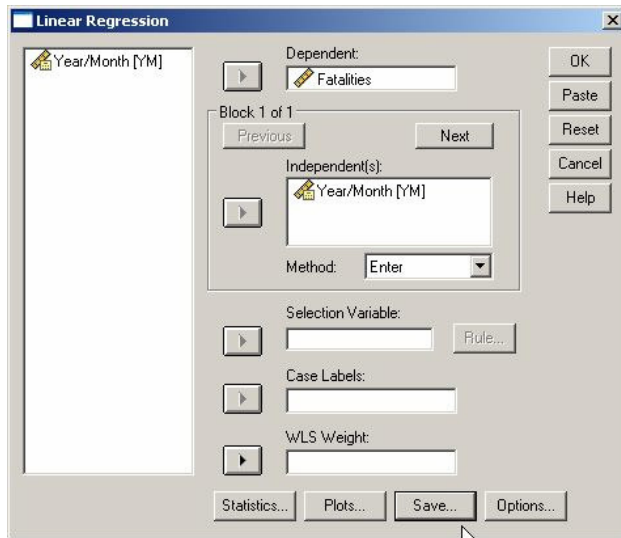


Homoscedasticity Assumption

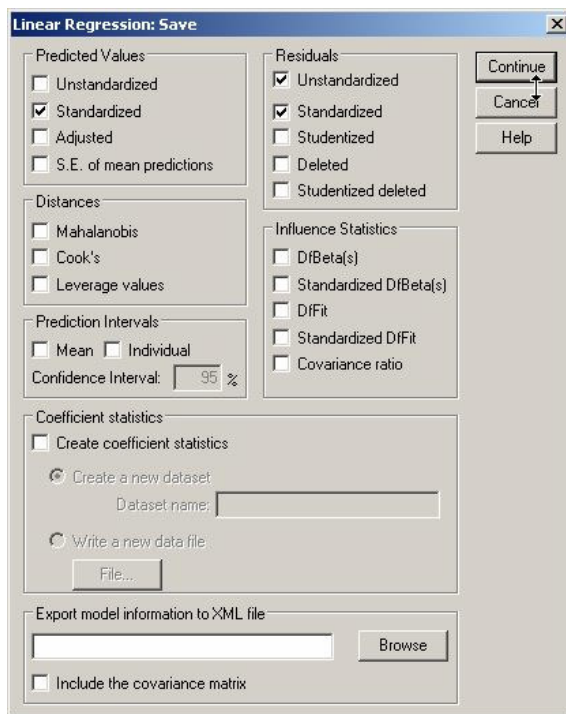
The homoscedasticity assumption specifies that the variance of the error term is constant across cases and independent of the variables in the model. This is the last tested and described assumption in this chapter; again a graphical and inference-statistical method is being used.

To compute the necessary plots, a number of new variables have to be created with the regression procedure.

Choose the linear regression model as shown in the example above, before clicking “OK”, click the button “Save”.



In this new opened window you choose the predicted values “Standardized” and the “Unstandardized” and “Standardized” residuals. Go on with the button “Continue” and then click “OK” in the “Linear Regression” window.



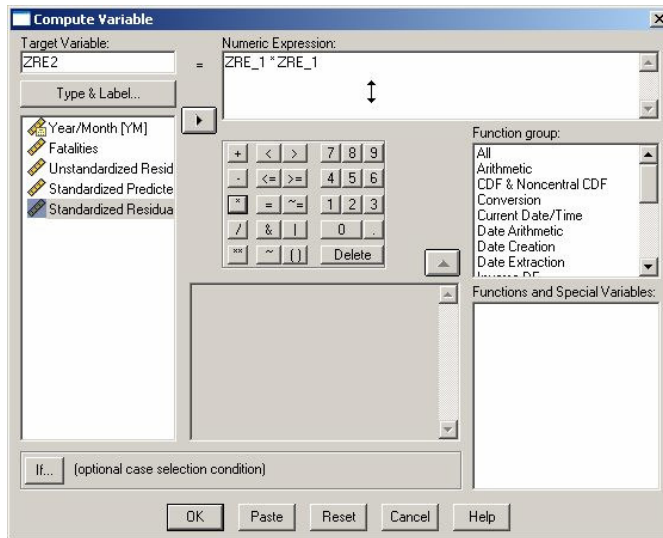
For the three new variables RES_1, ZPR_1, ZRE_1 see the data view in the SPSS-data file:

	YM	Fatalities	RES_1	ZPR_1	ZRE_1	var	var	var	var
1	JAN 1987	47	-75,60244	1,71974	-3,13316				
2	FEB 1987	65	-57,36374	1,70345	-2,37730				
3	MAR 1987	72	-50,14813	1,68873	-2,07827				
4	APR 1987	100	-21,90943	1,67243	-,90798				
5	MAY 1987	95	-26,67843	1,65666	-1,10562				
6	JUN 1987	115	-6,43973	1,64036	-,26688				
7	JUL 1987	186	64,79127	1,62459	2,68512				
8	AUG 1987	131	10,02998	1,60830	,41567				
9	SEP 1987	120	-,73132	1,59200	-,03031				
10	OCT 1987	162	41,49968	1,57623	1,71985				
11	NOV 1987	108	-12,26162	1,55993	-,50815				
12	DEC 1987	109	-11,03062	1,54416	-,45714				
13	JAN 1988	112	-7,79191	1,52787	-,32292				

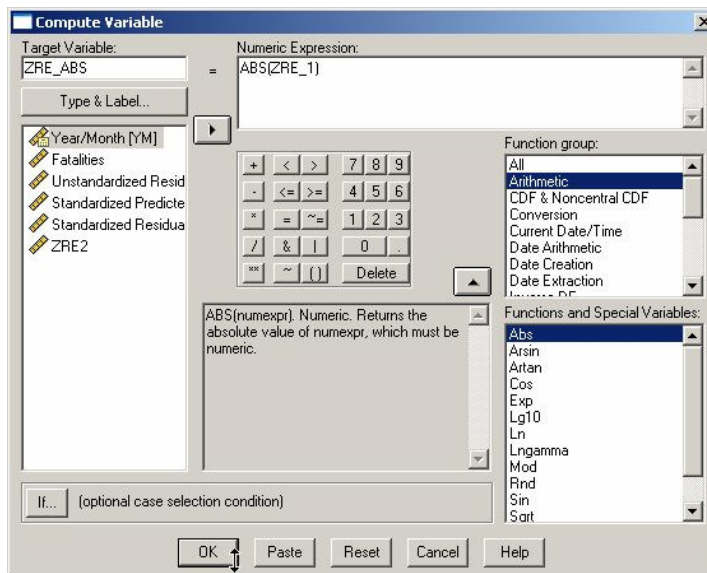
The three new variables are used to generate the variables “square of standardized residuals” and “absolute value of standardized residuals”. To generate them, choose from the menu “Transform” “Compute”.

	YM	Fatalities	RES_1	ZPR_1	ZRE_1	var	var	var	var
1	JAN 1987	47	-75,60244	1,71974	-3,13316				
2	FEB 1987	65	-57,36374	1,70345	-2,37730				
3	MAR 1987	72	-50,14813	1,68873	-2,07827				
4	APR 1987	100	-21,90943	1,67243	-,90798				
5	MAY 1987	95	-26,67843	1,65666	-1,10562				
6	JUN 1987	115	-6,43973	1,64036	-,26688				
7	JUL 1987	186	64,79127	1,62459	2,68512				
8	AUG 1987	131	10,02998	1,60830	,41567				
9	SEP 1987	120	-,73132	1,59200	-,03031				
10	OCT 1987	162	41,49968	1,57623	1,71985				
11	NOV 1987	108	-12,26162	1,55993	-,50815				
12	DEC 1987	109	-11,03062	1,54416	-,45714				
13	JAN 1988	112	-7,79191	1,52787	-,32292				

In the newly opened window “Compute Variable” you name the target variable “ZRE2”. Then you have to enter the numeric expression: Mark and click from the list of variables the “Standardized Residual” (ZRE_1) in the numeric expressions field, insert from the key pad * (for multiplication) and put once more the ZRE_1 variable in the numeric expressions field. Click “OK” for starting the computation process.



To compute the last variable (ZRE_ABS), choose from the menu “Function group” the type “All” and double-click “ABS” to put it in the field “Numeric Expression”. In the given bracket mark and click the variable “ZRE_1” (standardized residual). Click “OK” to compute the variable ZRE_ABS.



As a result five new variables have been retrieved.

Untitled [DataSet3] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : ZRE2 9,81670359463612

	YM	Fatalities	RES_1	ZPR_1	ZRE_1	ZRE2	ZRE_ABS	var	var
1	JAN 1987	47	-75,60244	1,71974	-3,13316	9,82	3,13		
2	FEB 1987	65	-57,36374	1,70345	-2,37730	5,65	2,38		
3	MAR 1987	72	-50,14813	1,68873	-2,07827	4,32	2,08		
4	APR 1987	100	-21,90943	1,67243	-,90798	,82	,91		
5	MAY 1987	95	-26,67843	1,65666	-1,10562	1,22	1,11		
6	JUN 1987	115	-6,43973	1,64036	-,26688	,07	,27		
7	JUL 1987	186	64,79127	1,62459	2,68512	7,21	2,69		
8	AUG 1987	131	10,02998	1,60830	,41567	,17	,42		
9	SEP 1987	120	-,73132	1,59200	-,03031	,00	,03		
10	OCT 1987	162	41,49968	1,57623	1,71985	2,96	1,72		
11	NOV 1987	108	-12,26162	1,55993	-,50815	,26	,51		
12	DEC 1987	109	-11,03062	1,54416	-,45714	,21	,46		

Data View Variable View

SPSS Processor is ready

Descriptions and definitions can be seen in the “Variable View”-window:

Untitled [DataSet3] - SPSS Data Editor

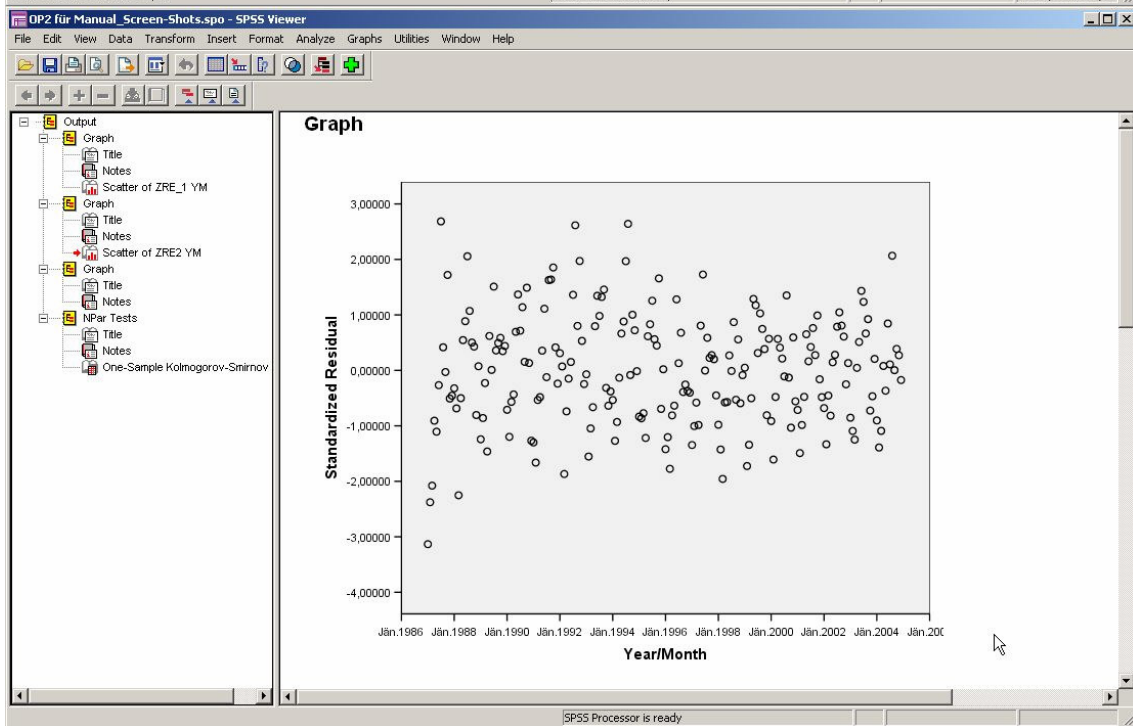
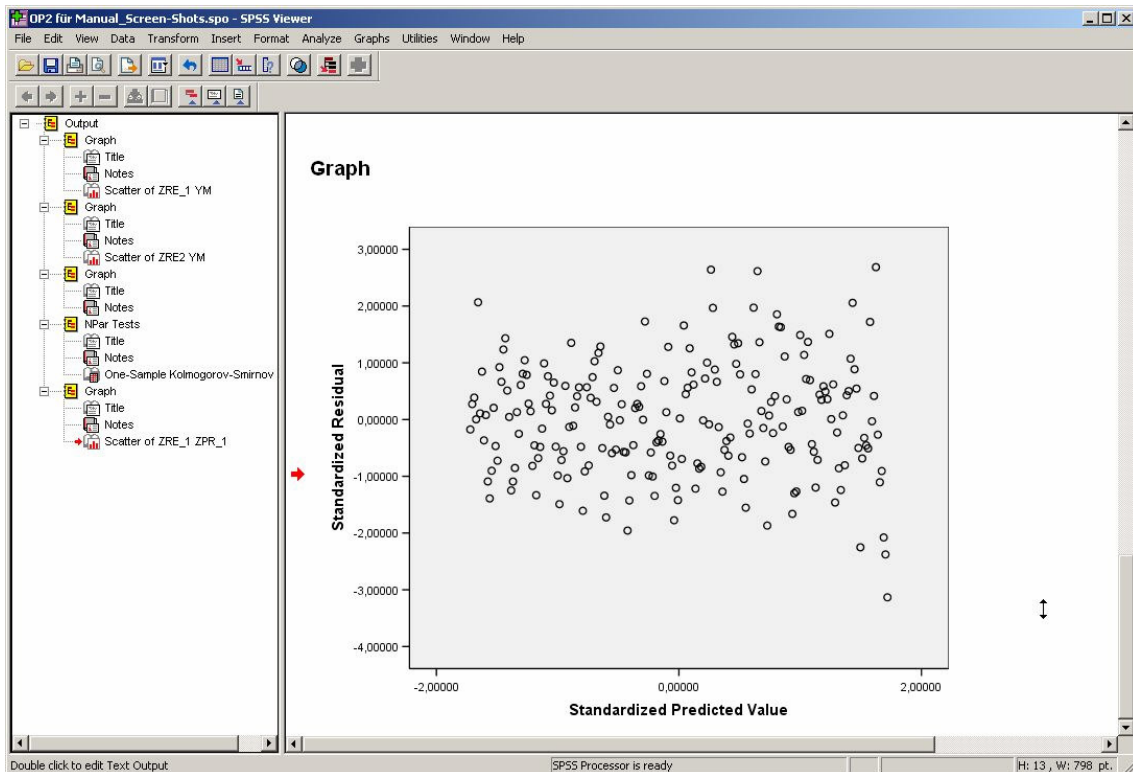
File Edit View Data Transform Analyze Graphs Utilities Window Help

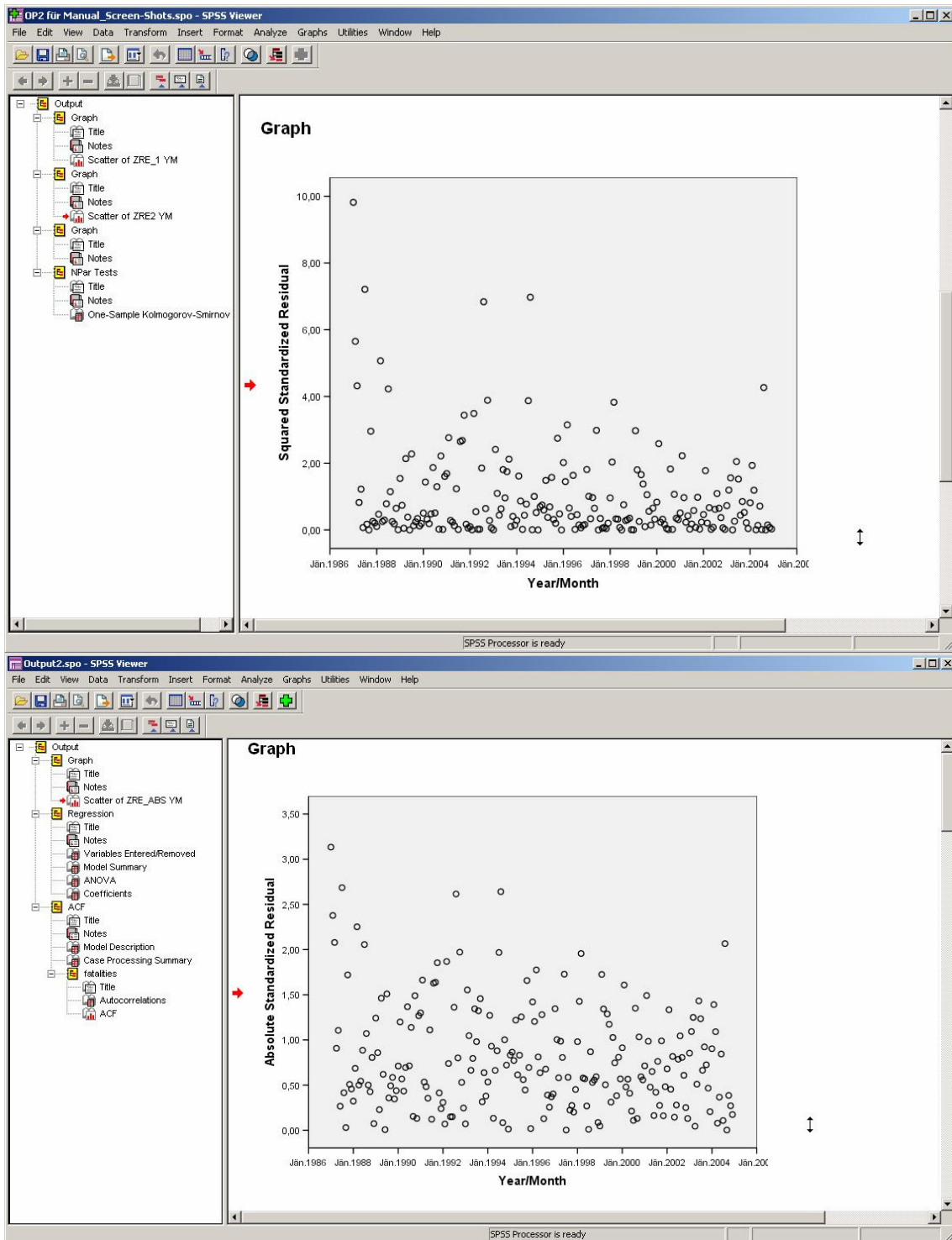
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	YM	Date	8	0	Year/Month	None	None	8	Right	Scale
2	Fatalities	Numeric	8	0	Fatalities	None	None	8	Right	Scale
3	RES_1	Numeric	11	5	Unstandardized Residual	None	None	13	Right	Scale
4	ZPR_1	Numeric	11	5	Standardized Predicted Value	None	None	13	Right	Scale
5	ZRE_1	Numeric	11	5	Standardized Residual	None	None	13	Right	Scale
6	ZRE2	Numeric	8	2	Squared Standardized Residual	None	None	10	Right	Scale
7	ZRE_ABS	Numeric	8	2	Absolute Standardized Residual	None	None	10	Right	Scale
8										
9										
10										
11										
12										

Data View Variable View

SPSS Processor is ready

With these new variables the assumption testing can be started. The scatter plots are derived with the same steps that are already explained above.

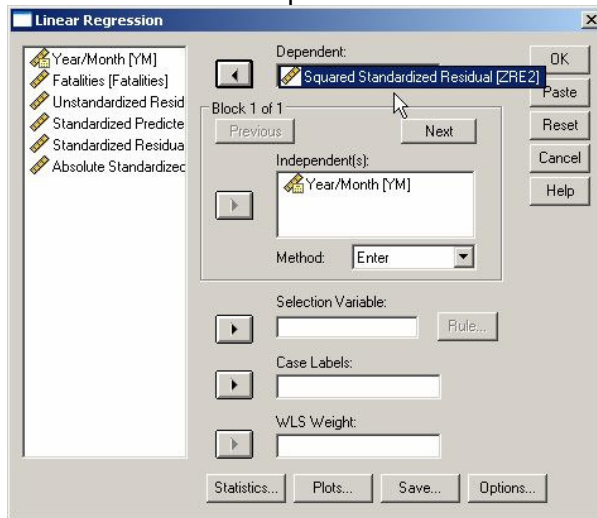




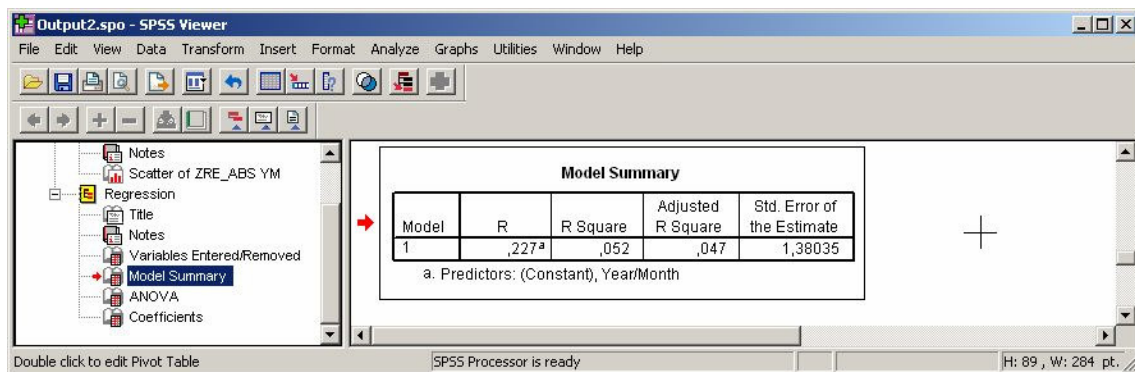
All four plots, especially the last two, show a distinctive pattern that indicates heteroscedasticity.

To complete the analysis with an inference-statistical model, the White-Chi Square test [1] is used. To compute this test, "R square" of the regression of the

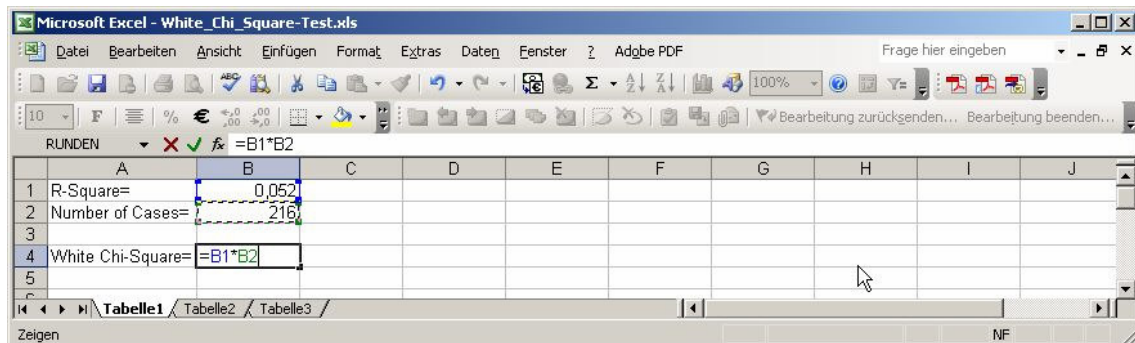
time series on “Squared Standardized Residuals” (ZRE2) is needed. Start the linear regression as explained above with “ZRE2” as dependent and “Year/Month” as independent variable.



The output of “Model Summary” contains the R square with 0.052.



The actual test can be easily done in Excel. Enter in a new excel spread sheet the computed R-square and the number of cases (n=216 in our example). The White Chi square is the multiplication of both.



To calculate the “Chivert” “1-alpha”, insert the Excel-function “Chivert” in the cell B6 and put in brackets the White Chi square (cell B4) and, after the semicolon, “1” for the number of the degrees of freedom.

	A	B	C	D	E	F	G	H	I	J	K
1	R-Square=	0,052									
2	Number of Cases=	216									
3											
4	White Chi-Square=	11,232									
5											
6	1-alpha=	=CHIVERT(B4;1)									
7		CHIVERT(x; Freiheitsgrade)									
8											
9											
10											

The result of “1-alpha” presents a highly significant deviation from the homoscedasticity-assumption:

	A	B	C	D	E	F	G	H	I
1	R-Square=	0,052							
2	Number of Cases=	216							
3									
4	White Chi-Square=	11,232							
5									
6	1-alpha=	0,00080399							
7									
8									

3.2.1.6. Conclusion

The conclusion of the results from the presented tests is that linear regression is not the preferred model in this case. Due to heavy violations of two basic assumptions of the model to fit time series data, a different and more advanced model specialised on time series data should be applied and will be demonstrated in the following chapters.

3.2.2 Generalized linear models (GLM)

As explained in the Introduction, this manual focuses on state-of-the-art dedicated time series techniques. Generalized Linear Models (GLM) are a useful and flexible technique that can be applied for time series analysis, however they do not by definition take into account the time dependence between observations; these dependences can only be considered by extending the GLM approaches with time series properties. In this sense, GLM are not dedicated time series techniques, like the ARMA-type and state space techniques. Consequently, no manuals are provided for GLM in time series analysis.

3.2.3 Non-linear models

Similarly to the Generalized Linear Models (GLM), Non Linear models are also a possible technique that can be used for time series analysis. However, they also can take into account the time dependence between observations only by extending the models accordingly; therefore they can not be considered as a dedicated time series techniques, like the ARMA-type and state space techniques. Consequently, no manuals are provided for Non Linear models in time series analysis.

3.3 Dedicated time series analysis in road safety research

The respective section in the methodology report is an overview section that does not contain empirical examples. Consequently there is no manual part for this section.

3.4 ARMA-type models

Ruth Bergel and Mohamed Cherfi, INRETS

3.4.1 Introduction

The objective of this part of the manual is to introduce technical aspects of time series modelling using ARMA-type models, applied to road safety analysis.

The section is structured in three parts. In Section 3.4.2, it is briefly recalled to the reader that he should refer to the respective section 3.4.2 of the Methodology Report, in which several ARMA models were fitted on simulated stationary datasets. Section 3.4.3 presents an example of *ARIMA* model estimated on non seasonal (yearly) real data, without any use of exogenous variables. And in Section 3.4.4, *ARIMA* models estimated on seasonal (monthly) real data have been chosen and exogenous variables, whether intervention variables or explanatory variables, have been succeedingly introduced in the model.

Each case of real data retained in this manual is a non-stationary case: of first order (the mean of the process varies over time) and of second order (the variance of the process varies over time), this being the general case for risk indicators in the road safety field.

Because of that non stationarity in variance, the dependent data were systematically Log-transformed. As for the exogenous (independent) variables, which are always considered as known (non stochastic, or in other words, not under measurement errors), the ones retained for measuring the traffic volume or the petrol/gasoline price were Log-transformed; whereas the intervention variable was used in a linear form (see the Methodology Report).

The detailed specification of the ARMA-type estimated models cannot be found in the manual but is described in the Methodology Report. Note that explanations for the ARMA structure, applied to the stationary datasets derived from the initial ones, are to be found in section 3.4.2 of the Methodology Report.

For each of the data cases, we first followed the succeeding steps for fitting pure ARIMA models on the datasets:

- Data description
- Model identification
- Model estimation and validation
- Graphical results and additional (normality) test

We then used external information by means of adding intervention and explanatory variables into the pure ARIMA models in view of taking account for certain risk factors, road safety measures or special events.

For each example the parameters of the exogenous variables were interpreted, and the gain in fit statistics was measured.

SPSS (Version 14.0) was used for this work.

Regarding the output obtained after each model estimation, the following results are systematically given:

- three tables (model fit, model statistics, and ARIMA model parameters),
- the ACF plot (and PACF plot) of the residuals
- two graphs (the observed and the fitted series on the one hand, and the residuals on the other hand),
- and the results of additional tests of normality of the residuals.

The first of the three tables provides Goodness-of-Fit Measures* (which enable to evaluate the model's empirical performance):

- Stationary R-squared
- R-squared
- RMSE
- MAPE
- MAE
- MaxAPE
- MaxAE
- Normalized BIC

The second table provides mainly the Ljung-Box statistic** (which enables to evaluate the model specification),

The third one provides the Model parameters (the estimated model parameters and their significance).

As for the normality of the residuals test, two plots were systematically given (the histogram and the QQ-plot), and the non-parametric Kolmogorov-Smirnov statistic ***

(*) The first of the three output tables provides **Goodness-of-Fit Measures:**

Goodness-of-fit statistics are based on the original series $Y(t)$. Let k = number of parameters in the model, n = number of non-missing residuals.

- **Stationary R-squared.** It compares the stationary part of the model to a simple mean model. This measure is preferable to ordinary R-squared when there is a trend or seasonal component in the series.

$$R_s^2 = 1 - \frac{\sum_t (Z(t) - \hat{Z}(t))^2}{\sum_t (\Delta Z(t) - \Delta \bar{Z})^2}$$

Where:

The sum is over the terms in which, both $(Z(t) - \hat{Z}(t))$ and $\Delta Z(t) - \Delta \bar{Z}$ are not missing.

$\Delta\bar{Z}$ is the simple mean model for the differenced transformed series, which is equivalent to the univariate baseline model $ARIMA(0,d,0)(0,D,0)$.

- **R-squared.** An estimate of the proportion of the total variation in the series that is explained by the model.

$$R^2 = 1 - \frac{\sum (Y(t) - \hat{Y}(t))^2}{\sum (Y(t) - \bar{Y})^2}$$

- **RMSE.** Root Mean Square Error. The square root of mean square error.

$$RMSE = \sqrt{\frac{\sum (Y(t) - \hat{Y}(t))^2}{n - k}}$$

- **MAPE.** Mean Absolute Percentage Error.

$$MAPE = \frac{100}{n} \sum |(Y(t) - \hat{Y}(t)) / Y(t)|$$

- **MAE.** Mean absolute error.

$$MAE = \frac{1}{n} \sum |Y(t) - \hat{Y}(t)|$$

- **MaxAPE.** Maximum Absolute Percentage Error. The largest forecasted error, expressed as a percentage

$$MaxAPE = 100 \max |(Y(t) - \hat{Y}(t)) / Y(t)|$$

- **MaxAE.** Maximum Absolute Error. The largest forecasted error, expressed in the same units as the dependent series

$$MaxAE = \max |Y(t) - \hat{Y}(t)|$$

- **Normalized BIC.** Normalized Bayesian Information Criterion.

$$NormalizedBIC = \ln(MSE) + k \frac{\ln(n)}{n}$$

(**) The second table provides mainly the **Ljung-Box statistic** (which enables to evaluate the model specification),

$$Q(K) = n(n+2) \sum_{k=1}^K r_k^2 / (n-k),$$

where r_k is the k th lag ACF of residual.

$Q(K)$ is approximately distributed as $\chi^2(K-m)$, where m is the number of parameters other than the constant term and predictor related-parameters.

(***) The two one-sided Kolmogorov-Smirnov test statistics are given by:

$$D_n^+ = \max(F_n(x) - F(x))$$

$$D_n^- = \max(F(x) - F_n(x))$$

where $F(x)$ is the hypothesized distribution.

3.4.2 ARIMA models for stationary series (simulated data)

Stationary series are usually not found in the road safety field. Therefore, simulated stationary data samples were used in a first approach, on which ARMA models were fitted. The structure of these simple models is similar to the structure of the more elaborated models which will be fitted on real road safety data, as far as handling their stationary part is required.

The reader will therefore refer to the respective section of the Methodology Report, in which the modelling stages are described and the modelling results given.

3.4.3 ARIMA models for non seasonal series (Norwegian Fatalities)

3.4.3.1. Data description

1. Start of analysis and data load

- First, we start SPSS.

Use the menu <File, Open, Data ...> to open the file 'Norw_Fatalities.sav'.

	NorwFatalities	LNorw.Fatalities	YEAR	DATE	var	var	var	var	var	var
1	560,00	6,33	1970	1970						
2	533,00	6,28	1971	1971						
3	490,00	6,19	1972	1972						
4	511,00	6,24	1973	1973						
5	509,00	6,23	1974	1974						
6	539,00	6,29	1975	1975						
7	471,00	6,15	1976	1976						
8	442,00	6,09	1977	1977						
9	434,00	6,07	1978	1978						
10	437,00	6,08	1979	1979						
11	362,00	5,89	1980	1980						
12	338,00	5,82	1981	1981						

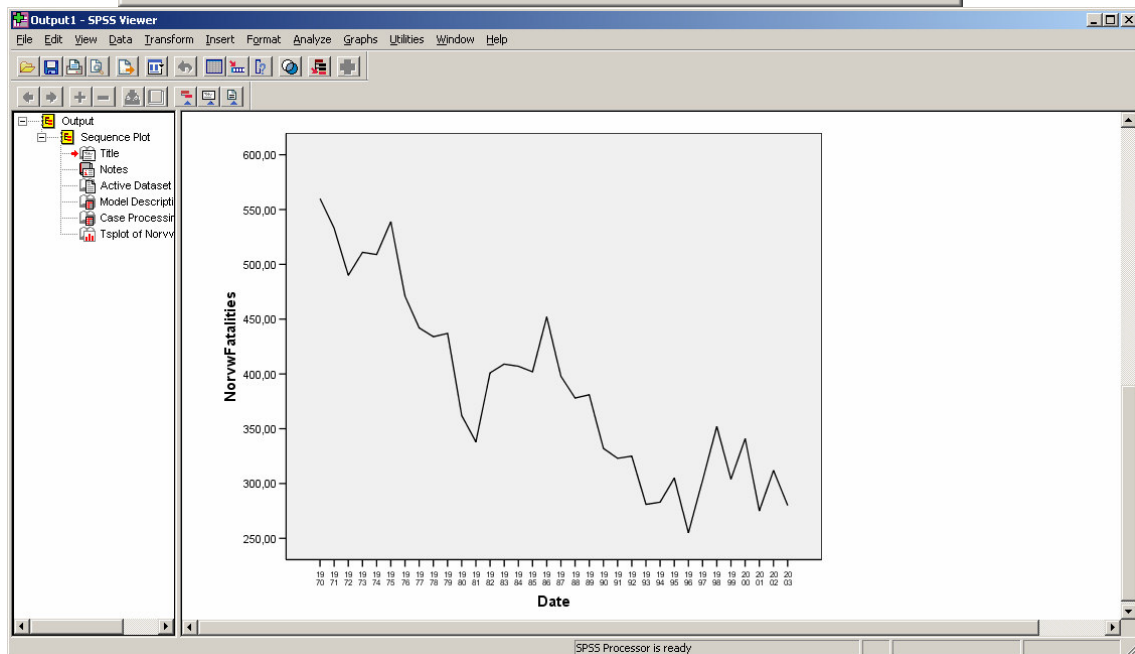
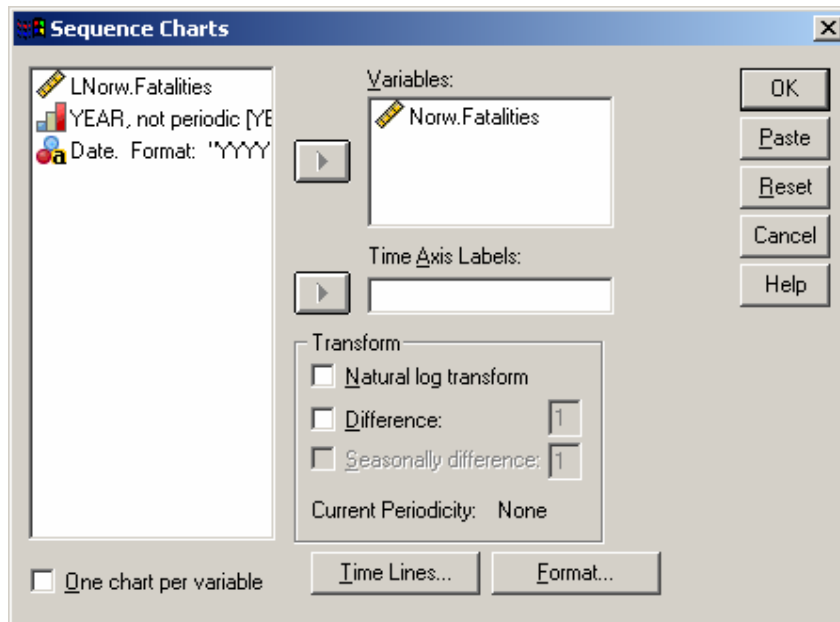
This data file consists of the annual number of people killed in road traffic in Norway for the years 1970 to 2003 ('Norw.Fatalities') and of the logarithm of the latter time series ('LNorw.Fatalities'), and of two additional variables, labelled YEAR and DATE (note that all variables are described in the Variable View).

2. Graphical diagnostics

The data are represented graphically in a time series plot. This will help show up important features such as trend and, eventually, seasonality. The time

series plot can also help deciding whether a preliminary transformation (logarithmic transformation, filtering,) is required on the data.

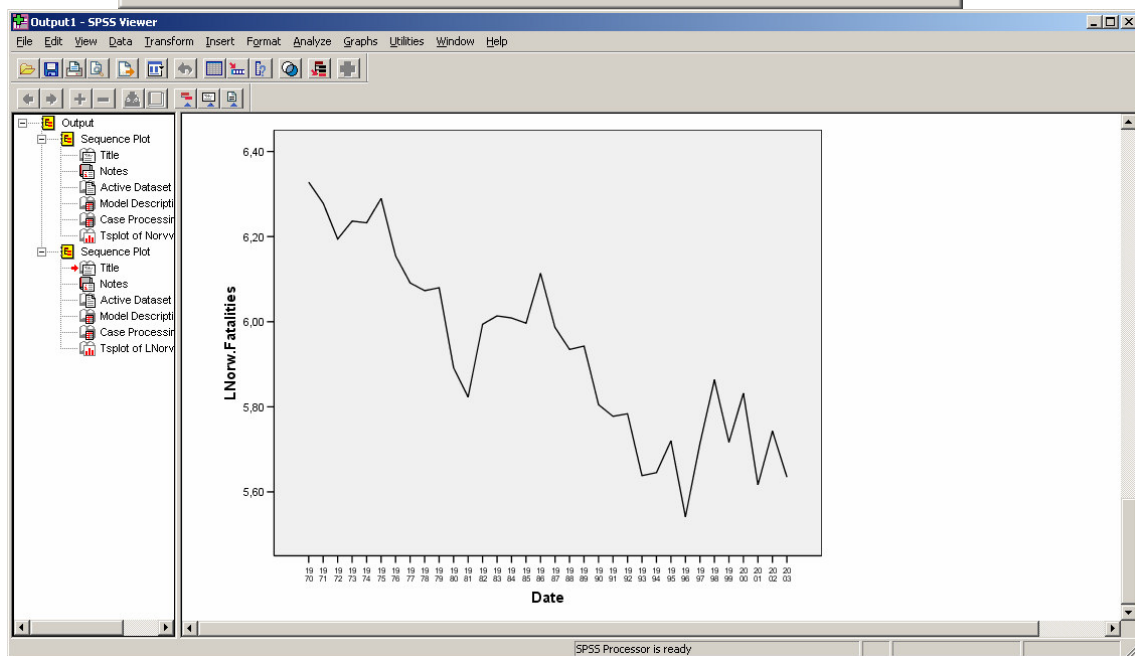
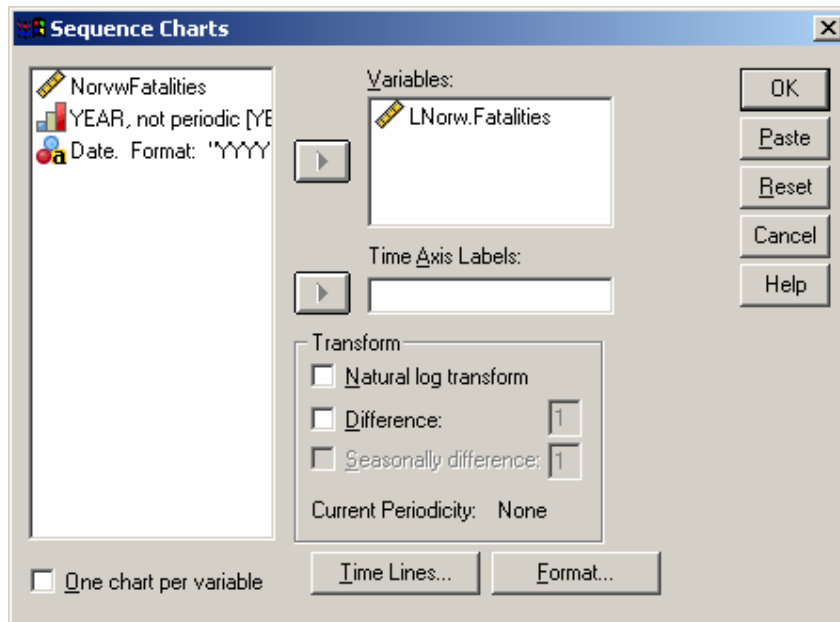
- Click on Graphs..Sequence
- Move Norw.Fatalities into the Variables list box



The plot shows the existence of a decreasing trend in the time series (first order non stationarity).

By taking the logarithm of these data, we can stabilize the data variance (second order non stationarity).

- Click on Graphs..Sequence
- Remove Norvw.Fatalities from the Variables list box
- Move LNorw.Fatalities into the Variables list box

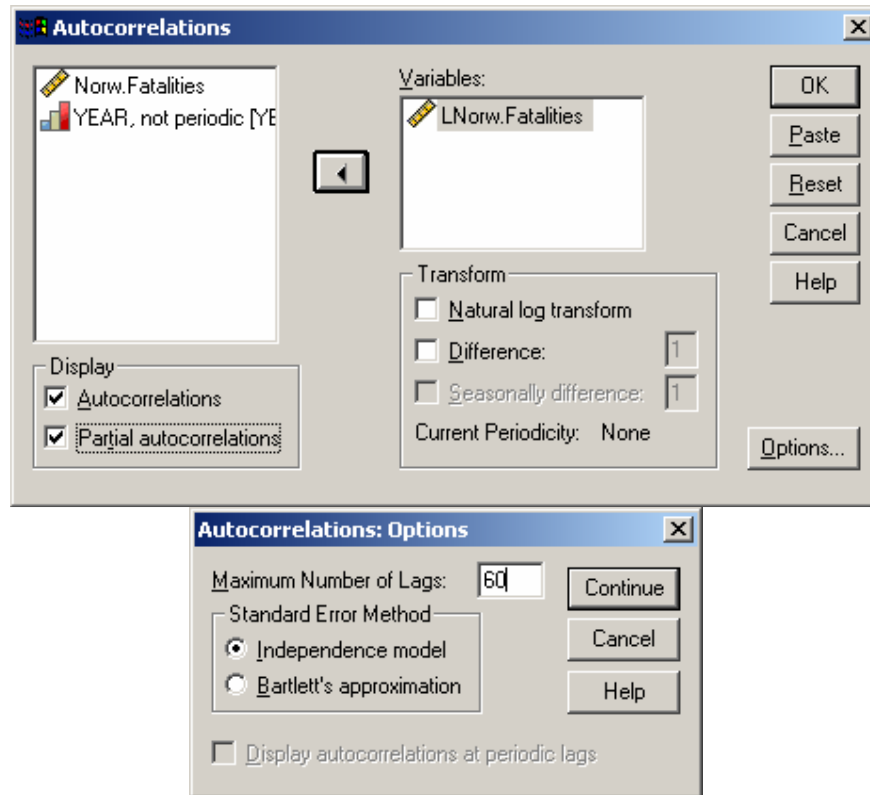


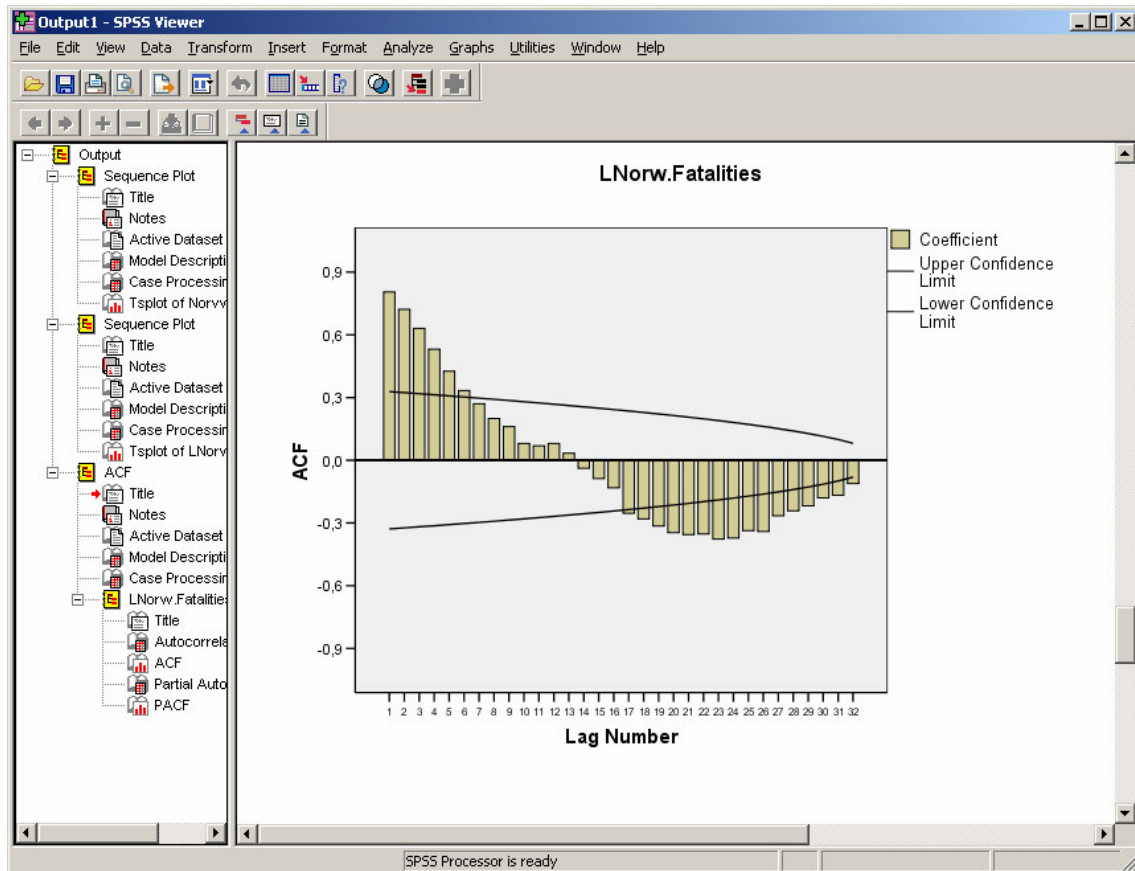
3.4.3.2. Model identification

The model identification consists in determining the three integers p , d , and q in the $ARIMA(p,d,q)$ process generating the series.

The ACF plot will be used to detect the presence of non stationarity in the data.

- Click on Graphs..Time Series..Autocorrelations
- Move the variable LNorw.Fatalities into the Variables list box
- Click on the Options pushbutton
- Replace 16 with 60 in the Maximum number of lags text box




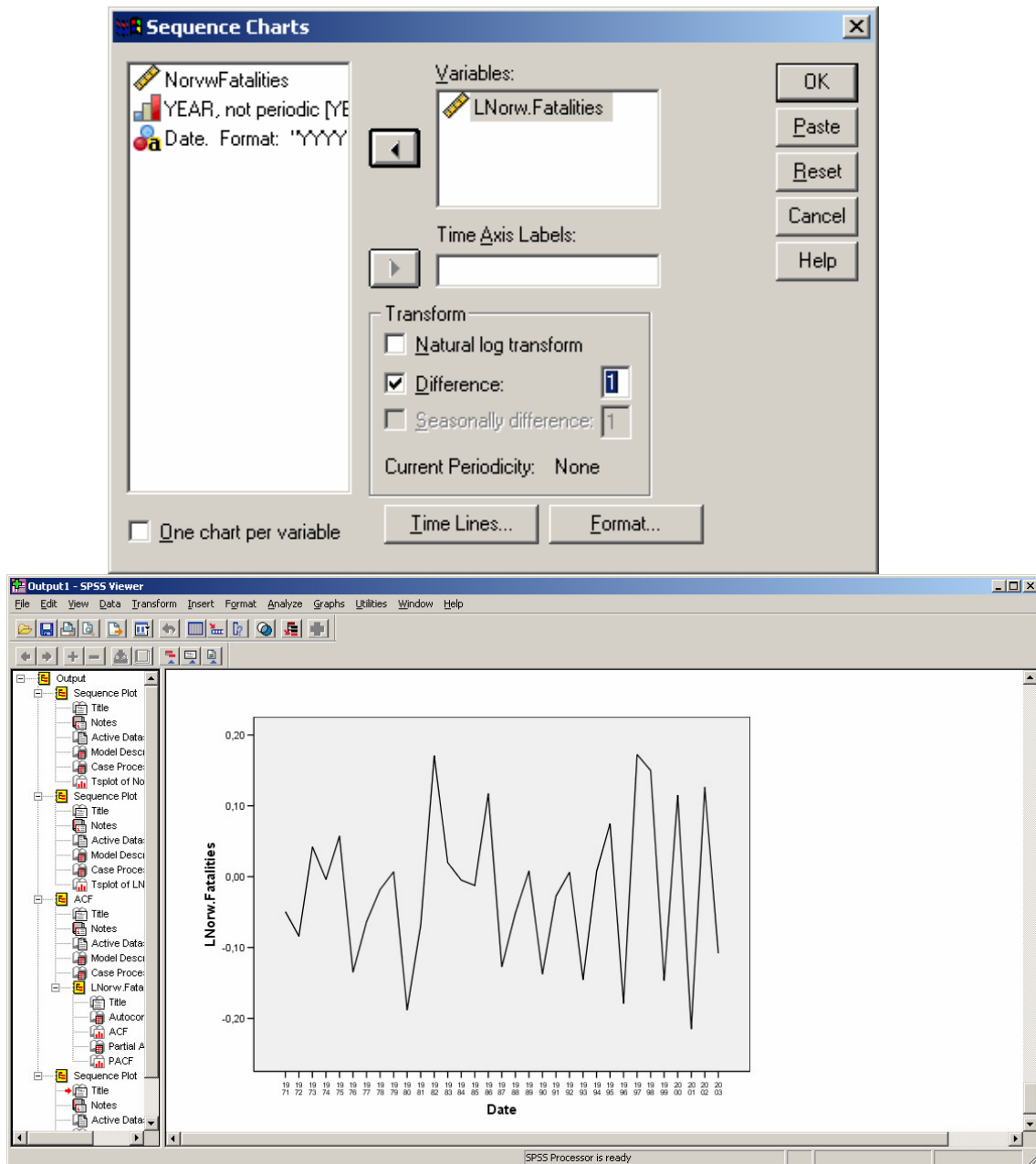


The ACF plot indicates the presence of non stationarity: this is due to the fact that the autocorrelations do not decrease at an exponential rate, after a certain order.

We shall therefore now differentiate the series, by applying the difference filter $F(B) = 1 - B$ to the data B being the backshift operator.

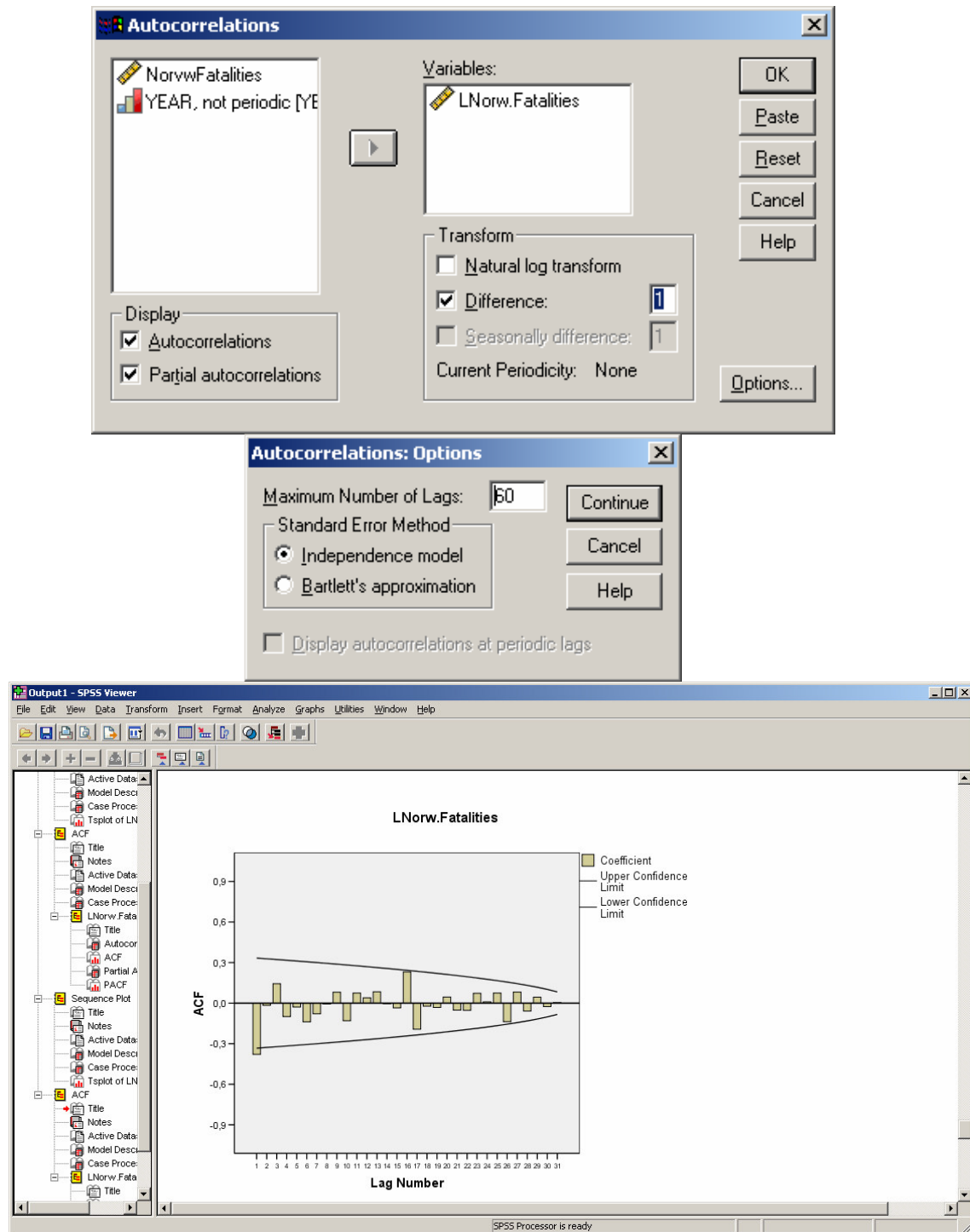


- Click on the Dialog Recall button  , and then click on Sequence Charts
- Click on the Difference check box, and verify that 1 is in the Difference text box.



The ACF plot of the filtered series will be used once again, to detect the presence of nonstationarity in this filtered dataset.

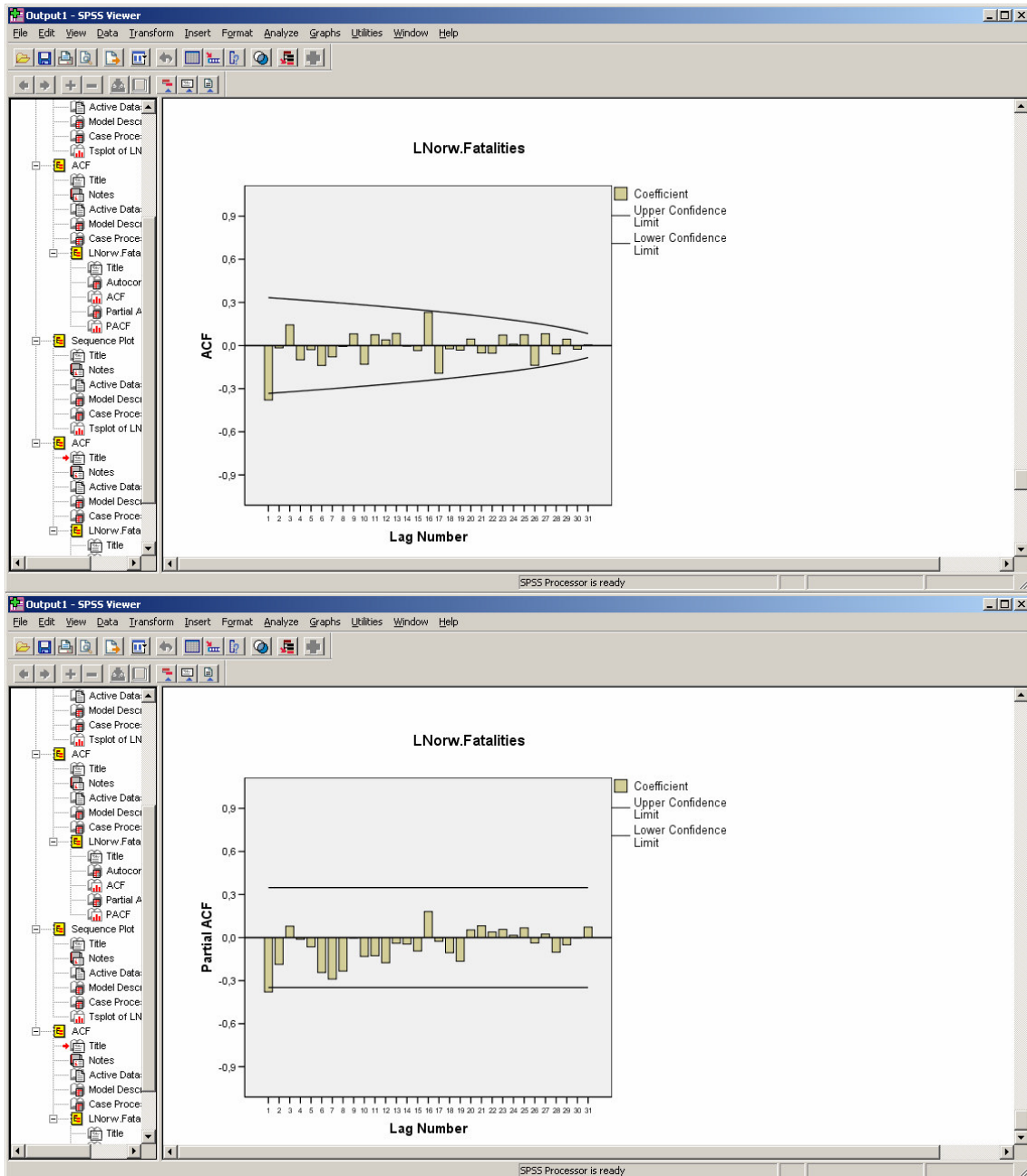
- Click on Graphs..Time Series..Autocorrelations
- Move the variable LNorw.Fatalities into the Variables list box
- Click on the Difference check box, and verify that 1 is in the Difference text box.
- Click on the Options pushbutton
- Replace 16 with 60 in the Maximum number of lags text box



The ACF plot of the filtered series does not indicate the presence of remaining non stationarity.

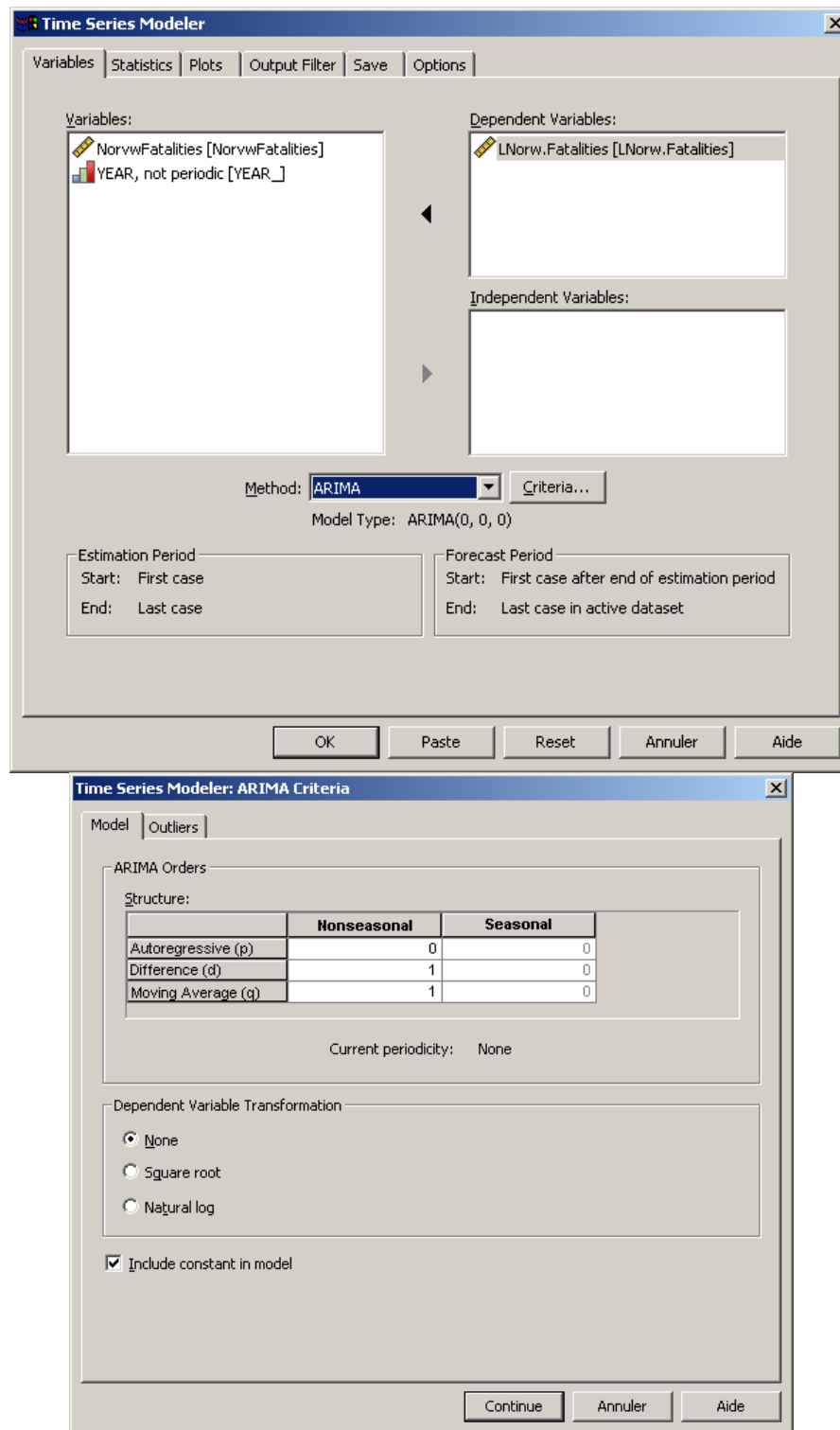
We shall therefore accept the hypothesis that this filtered series is stationary, which enables to retain $d=1$ for the value of the integer d .

Second, the choice of $p=0$ and $q=1$ is made by examining the ACF and the PACF plots taken together: we choose to fit the data with an $ARIMA(0,1,1)$ model.

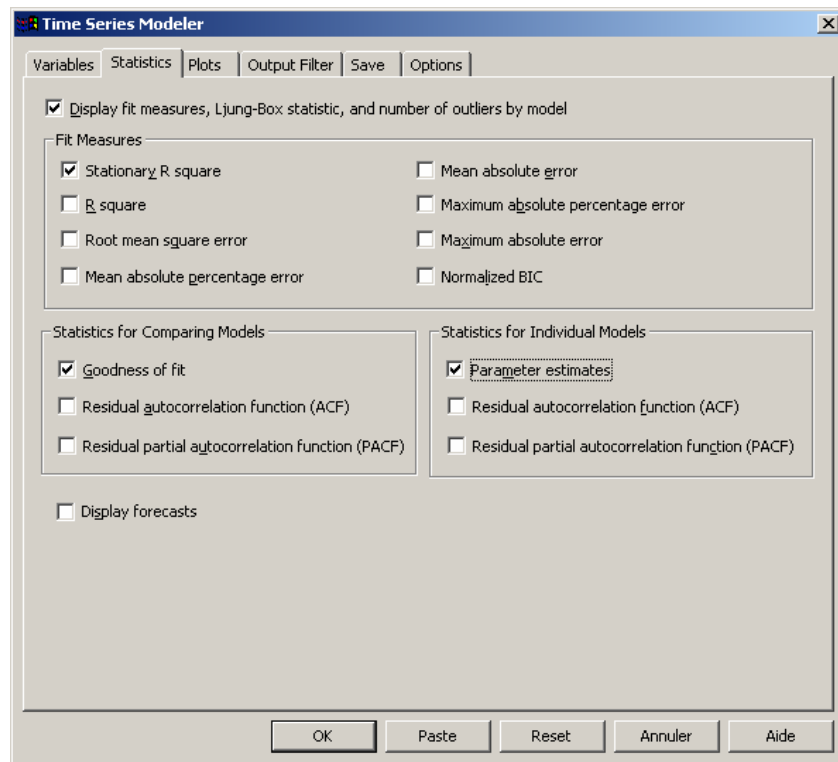


3.4.3.3. Model estimation and validation

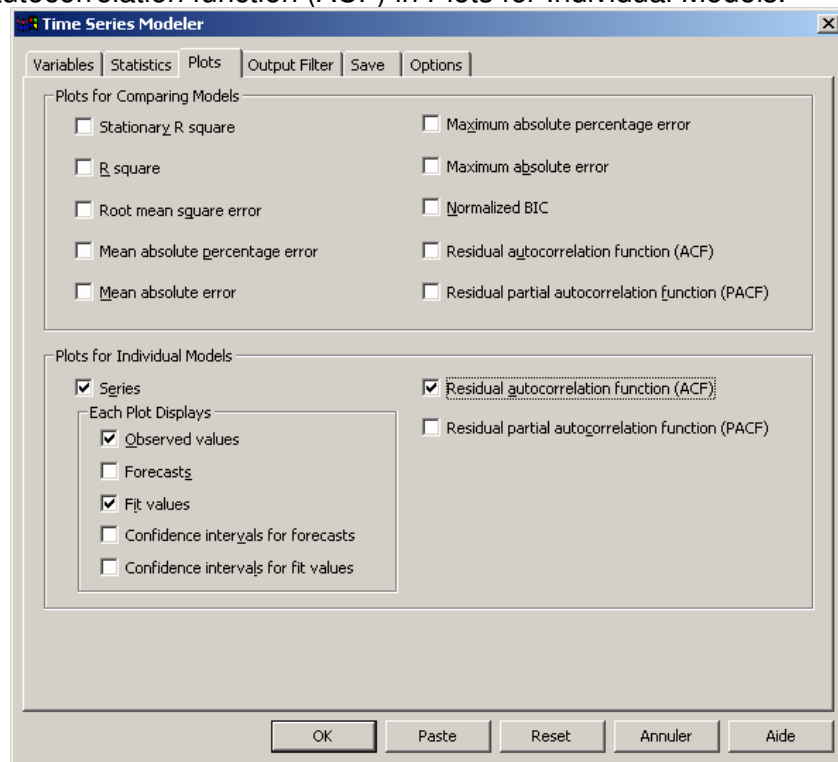
- Click on Analyze..Time Series..Create Models
- Move the variable LNorw.Fatalities into the Dependent Variable(s) list box.
- From the Method box, select ARIMA modelling method
- Click on Criteria then enter values for the three integers p,d,q.



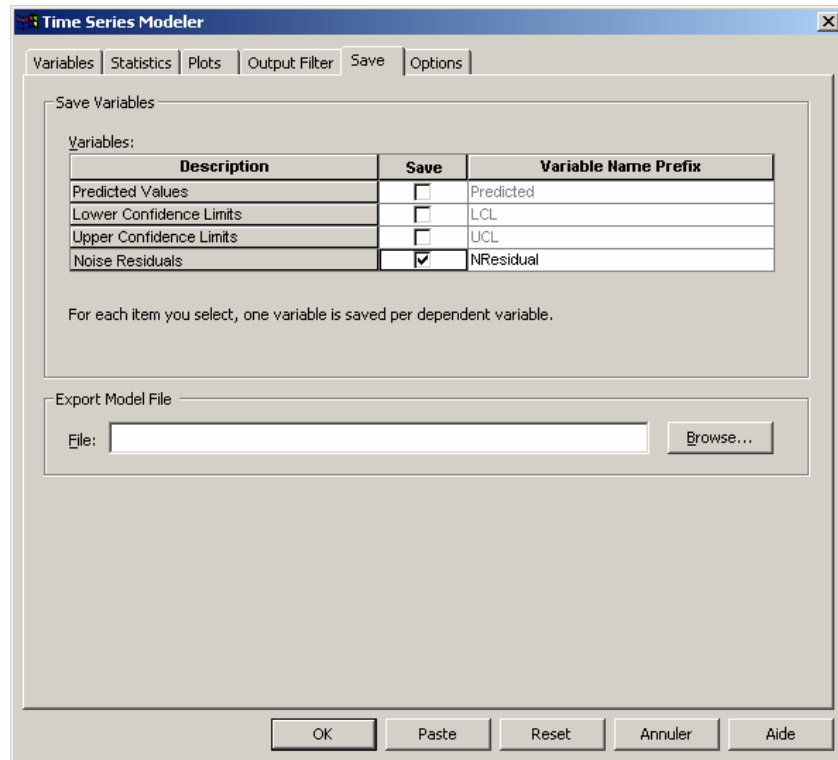
- Click on Statistics then check parameter estimates in the Statistics for Individual Models list.



- Click on Plots and check Observed Values, Fit values, Residual autocorrelation function (ACF) in Plots for Individual Models.



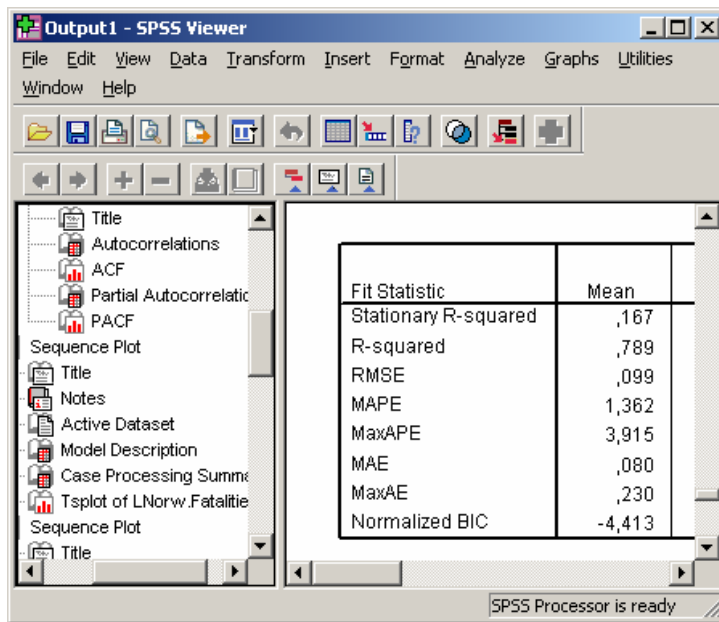
- Click on Save and check Noise Residuals then Click on OK.



Model Description ARIMA(0,1,1)

The results obtained after the estimation procedure (maximum likelihood) are presented and commented below.

Model Fit



Due to the presence of the trend, the stationary R -squared is only 16,7% (the model explains 16,7% of the variance of the filtered data, compared to a regression model), and much smaller than the R -square which is 78,9% (the model explains 78,9% of the variance of the initial data).

As for the different measures of the error made:

the mean absolute percentage error (MAPE) is 1,36% %, its highest value observed being 3,915%,

the mean absolute error (MAPE) is 0,08, its highest value observed being 0,23, the root mean square error (RMSE) is 0,099, and is a little higher than if it were computed as an arithmetic mean (0,08),

At last, the normalized BIC, which is -4,413, is a goodness of fit measure that takes account of the parsimony of the model. Note that, as it is the case for the R -squared, its interest lies in comparisons between several nested models, and not in its absolute value.

Output1 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Model Statistics

Model	Number of Predictors	Model Fit statistics		Ljung-Box Q(18)		Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
LNorw.Fatallie-Model_1	0	,167	16,199	17	,510	0

ARIMA Model Parameters

				Estimate	SE	t	Sig.
LNorw.Fatallie-Model_1	LNorw.Fatallie	No Transformation	Constant	-,020	,010	-1,969	,058
			Difference	1			
			MA	,432	,164	2,636	,013
			Lag 1				

SPSS Processor is ready

Model Statistics

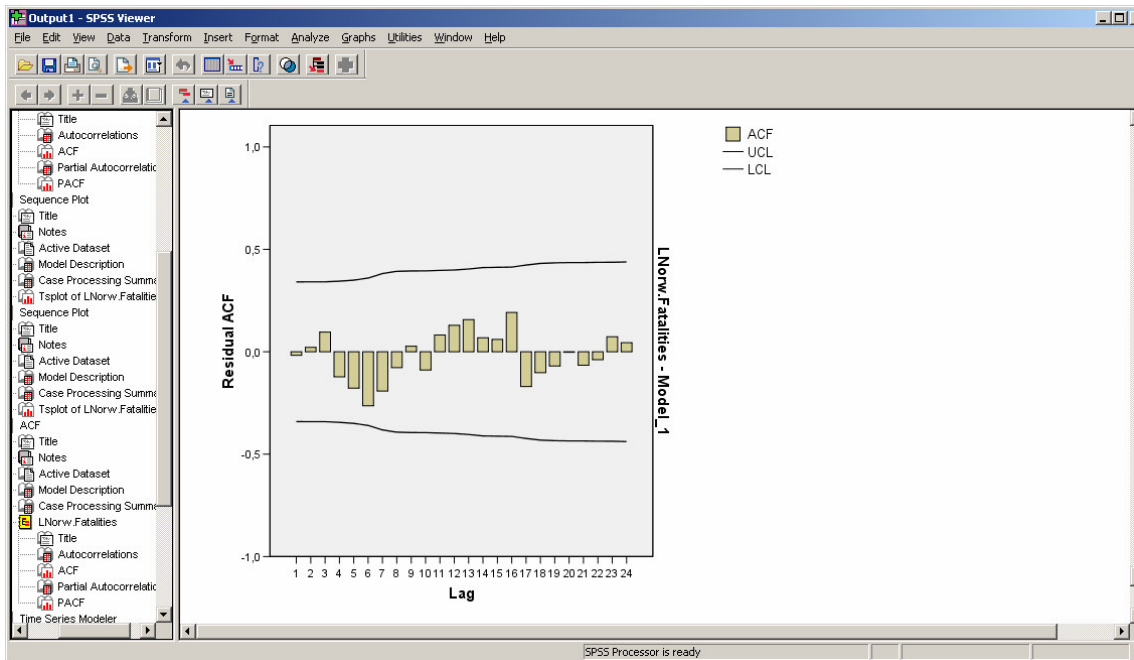
The Ljung-Box statistic provides an indication of whether the model is correctly specified, in the sense it allows testing the global nullity of the autocorrelation of the residuals (of each autocorrelation, of order 1 up to order 18).

In our case, this hypothesis is accepted, because the 0,510 value of the Ljung-Box statistic is more than 0.05.

ARIMA Model Parameters

The ARIMA model parameters table provides estimates of the model parameters and associated significance values (at the usual 95% confidence level). A t-value higher than 1,96 indicates that the hypothesis of nullity of the parameter has to be rejected, and that the parameter can thus be considered as significantly different from 0.

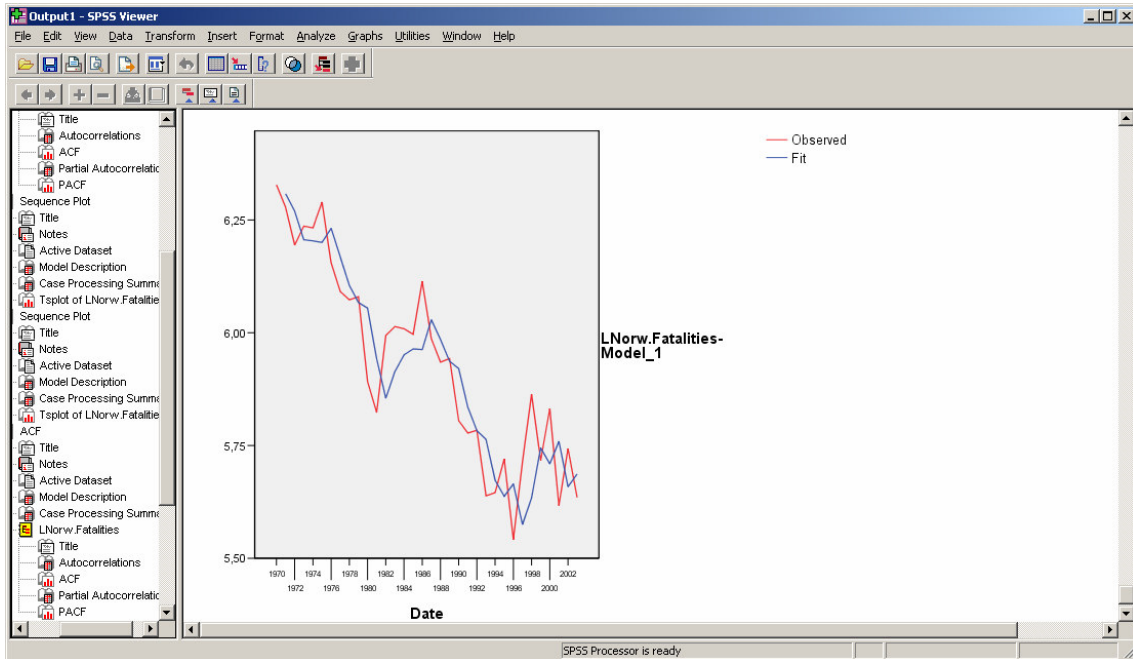
In this case, the hypothesis of nullity is rejected in both cases, and the two parameters of the ARIMA models are to be considered as different from zero.



In addition to the preceding Ljung-Box test, of global nonautocorrelation of the residuals (correct specification of the model), the hypothesis that each autocorrelation of the residuals is zero can be tested using the above ACF plot. The computation of confidence regions enable to determine visually whether it is the case (at the usual 95% confidence level). *It is the case indeed for this example.*

3.4.3.4. Graphical results and additional test

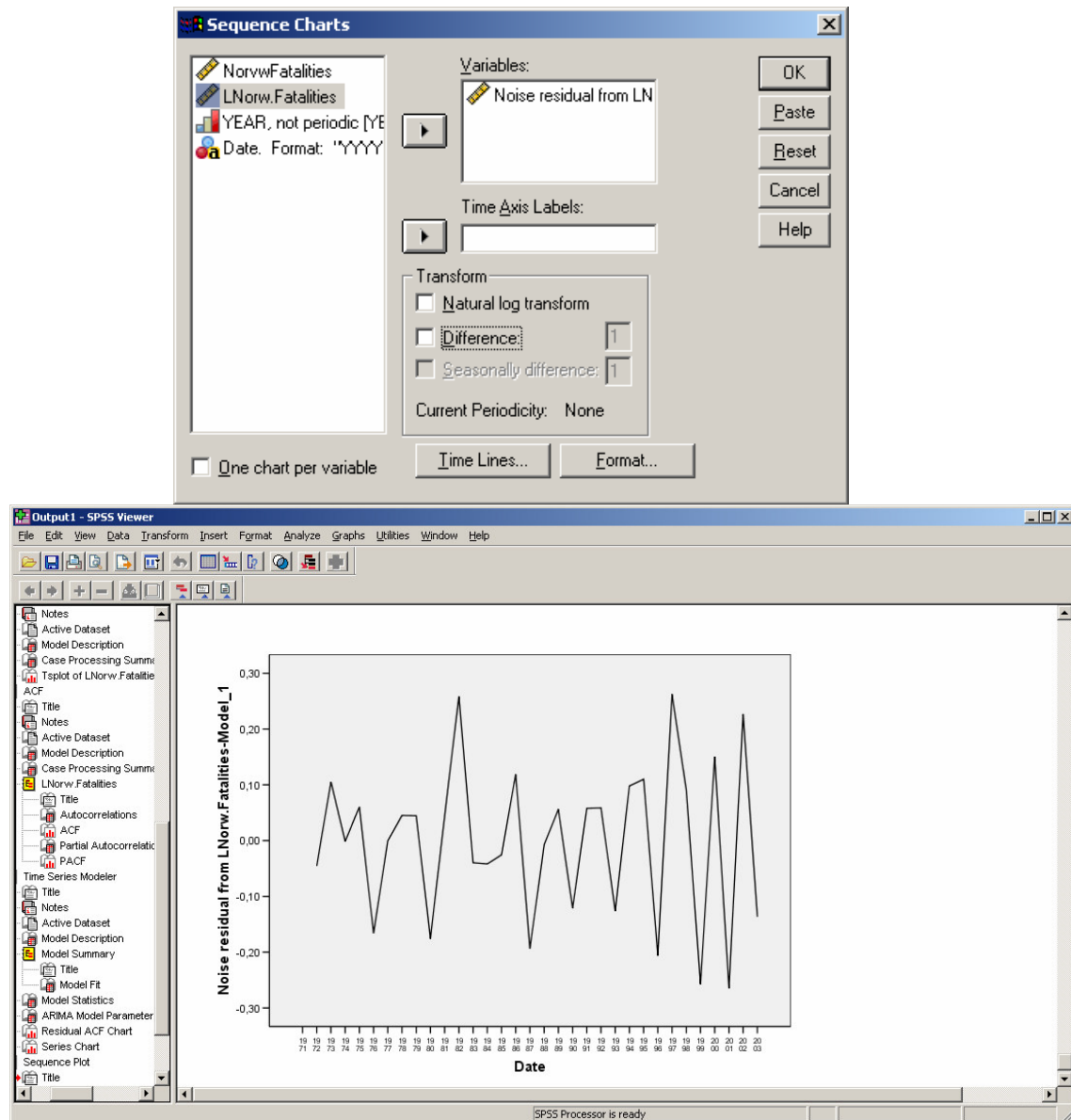
Graphical outputs



The preceding plot describes the development of the observed and fitted series. Note that, due to the use of the filtered state used for computing the fitted values, the fitted data appear to stay one step behind the observed data.

Second, the plot of the estimated residuals - the difference between the observed and the fitted data plotted first - is to be considered:

- Click on Graphs..Sequence
- Remove LNorw.Fatalities from the variables list box
- Move Noise residual into the variables list box
-



Note that it is very difficult, on this example, to determine visually whether the residuals are a white noise or not.

In addition to the nonautocorrelation hypothesis, the Gaussian hypothesis has to be validated too, which then enables to consider the residuals as an independent series - or white noise.

Nevertheless, note that the Gaussian hypothesis is not necessary, and that the residuals can be a white noise even if this condition is not fulfilled. The hypothesis of independence is then to be tested directly, which will not be the case in this manual.

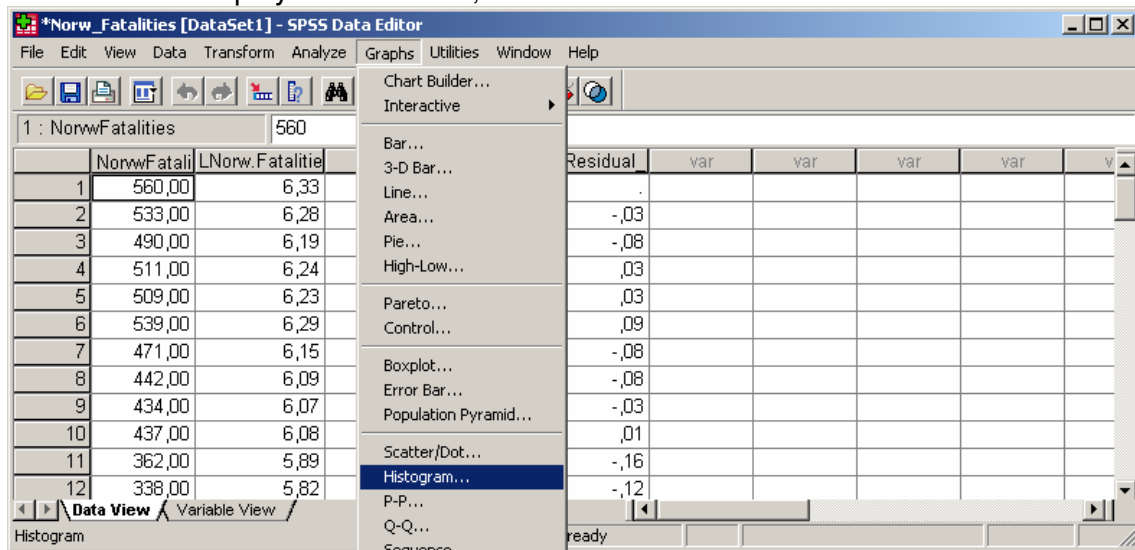
Normality test

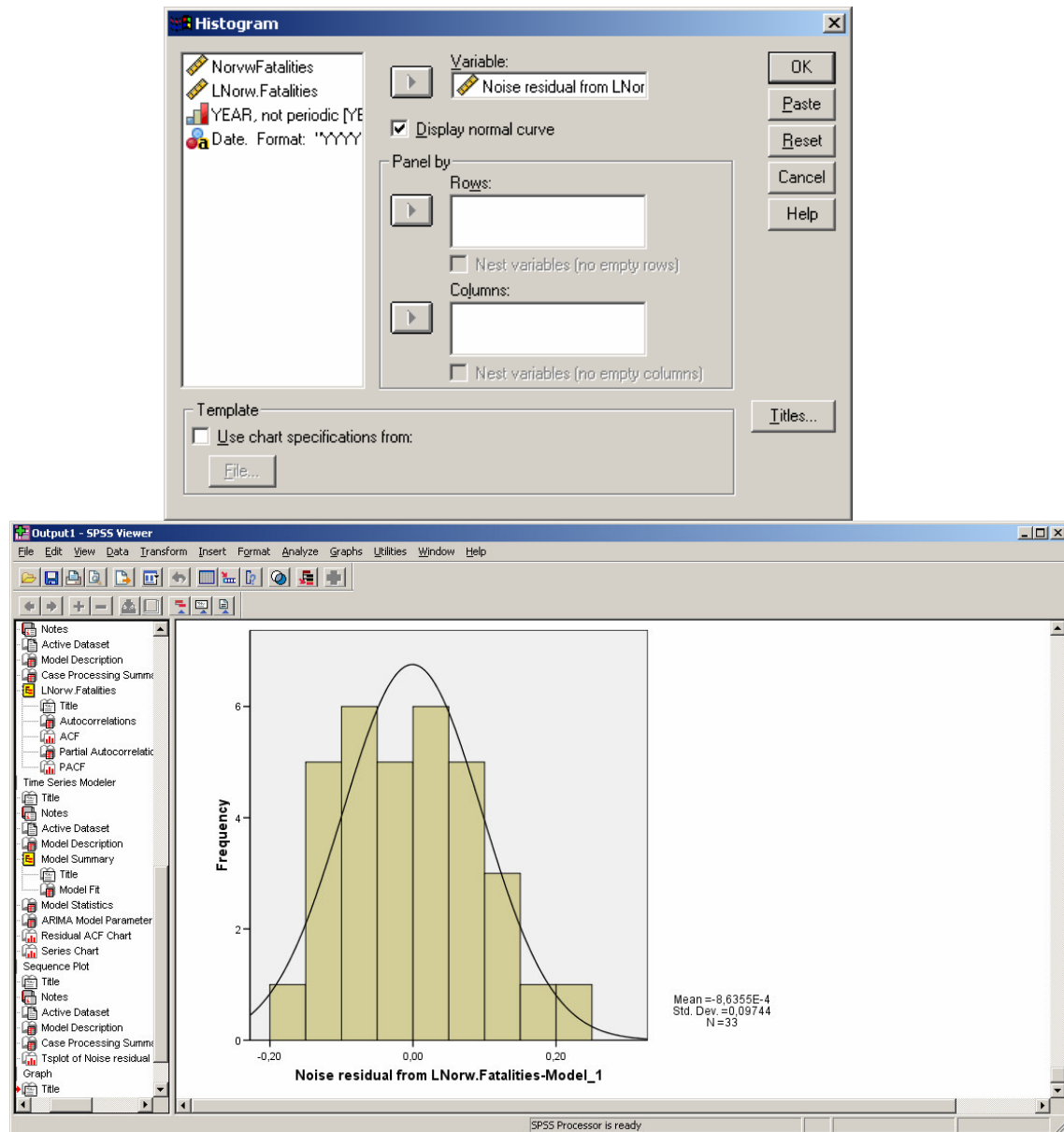
We shall now give the histogram of the residuals (which gives a general idea of whether the residuals are Gaussian), the QQ-plot (which is a graphical test of

this hypothesis), and at last the result of the Kolmogorov-Smirnov test (which is a non-parametric test of this hypothesis).

Histogram

- Click on Graphs..Histogram...
- Move the variable NResidual_LNorv_Model_1 into the Variable list box.
- Check Display normal curve, then click on OK

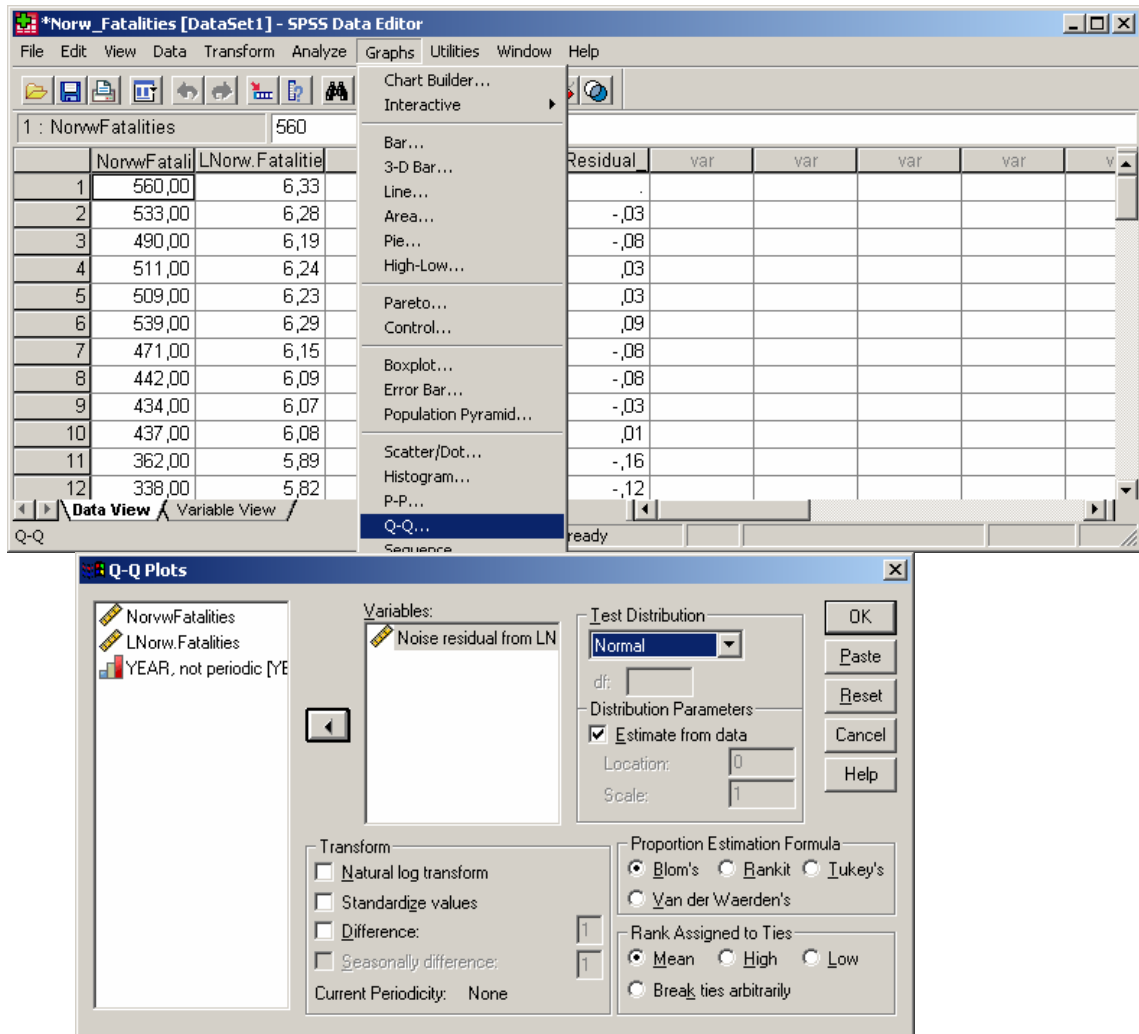


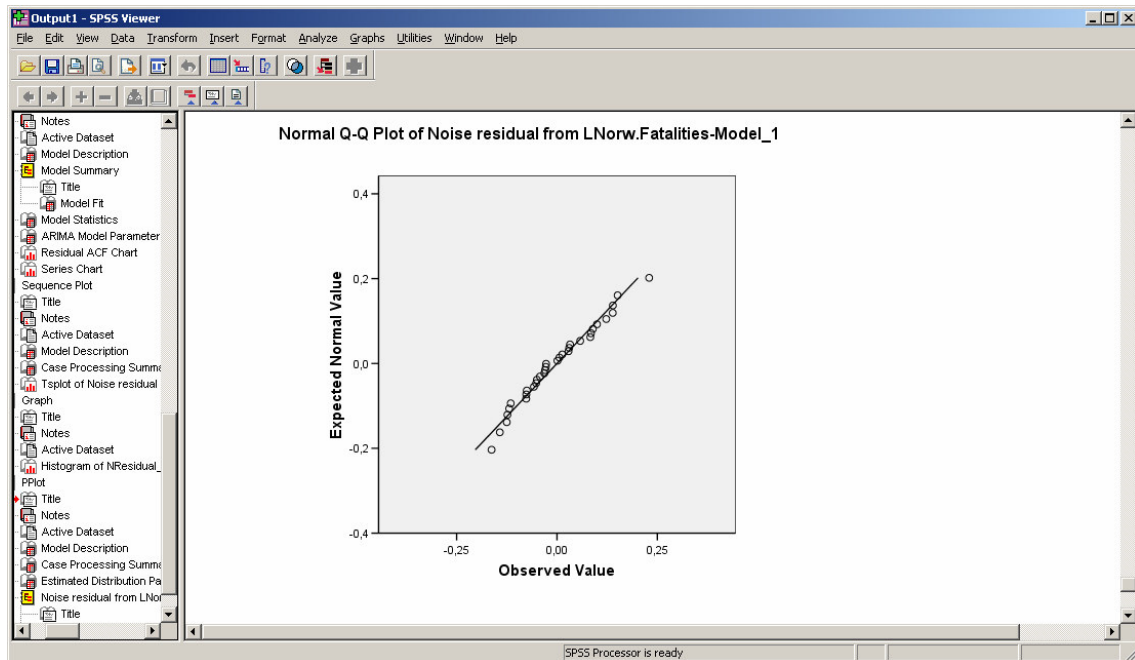


In the case the Residuals have a normal distribution, the histogram looks approximately like the normal curve, and the difference area between the two graphs is minimal.

QQ-plot

- Click on Graphs..QQ..
- Move the variable NResidual_LNorv_Model_1 into the Variables list box.
- Choose Normal from Test Distribution, then click on OK.

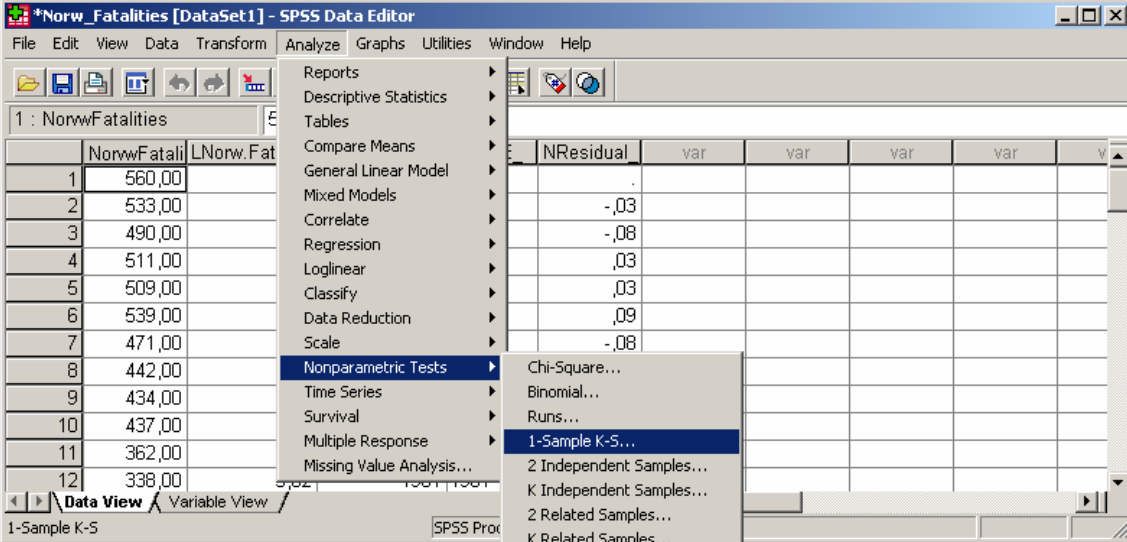




The normal Q-Q plot compares the distribution of a given variable to the normal distribution (represented by a straight line). The straight line represents what the residuals would look like if they were perfectly normally distributed. The residuals are represented by the circles plotted along this line. The closer the circles are to the line, the better the normality hypothesis is fulfilled.

Kolmogorov-Smirnov Test

- Click on Analyze..Non parametric Tests..1-Sample K-S...
- Move the variable NResidual_LNorv_1 into the Test Variable List box
- Check Normal in Test Distribution list, then click on OK.



The screenshot shows the SPSS Data Editor window with the 'Norw_Fatalities' dataset. The 'Analyze' menu is open, and the 'Nonparametric Tests' sub-menu is selected, leading to the '1-Sample K-S...' option. The 'One-Sample Kolmogorov-Smirnov Test' dialog box is open, showing the 'Test Variable List' with 'Noise residual from LN' and the 'Test Distribution' set to 'Normal'.

One-Sample Kolmogorov-Smirnov Test

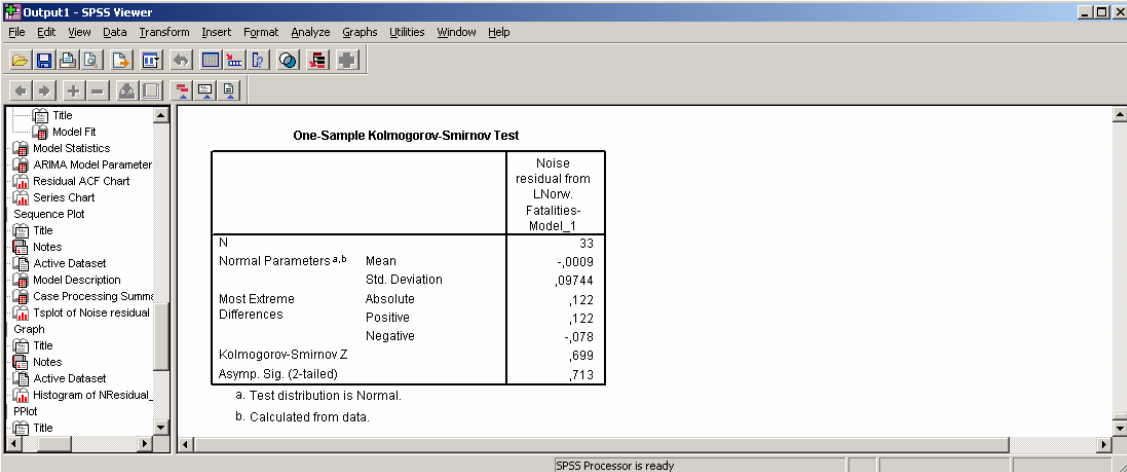
Test Variable List:

- Noise residual from LN

Test Distribution:

- ☒ Normal
- ☐ Uniform
- ☐ Poisson
- ☐ Exponential

Buttons: OK, Paste, Reset, Cancel, Help, Exact..., Options...



The screenshot shows the SPSS Output Viewer window displaying the results of the One-Sample Kolmogorov-Smirnov Test. The results are summarized in a table.

One-Sample Kolmogorov-Smirnov Test

		Noise residual from LN
N		33
Normal Parameters ^{a,b}	Mean	-,0009
	Std. Deviation	,09744
Most Extreme Differences	Absolute	,122
	Positive	,122
	Negative	-,078
Kolmogorov-Smirnov Z		,699
Asymp. Sig. (2-tailed)		,713

a. Test distribution is Normal.
b. Calculated from data.

In the case the Kolmogorov-Smirnov test is significant, the normal distribution of the residuals hypothesis is to be rejected.

In our case, this hypothesis is accepted, because the 0,713 value of the Asymp. Sig. (2-tailed) is more than 0.05 (at the usual 95% confidence level).

3.4.4 ARIMA models for seasonal series (UK-KSI Drivers)

The dataset presented in this section is the monthly number of drivers killed and seriously injured in the United Kingdom (UK-KSI), for the period January 1969 - December 1984 (as described in Harvey and Durbin, 1986). A pure ARIMA model will first be fitted on these data, and an intervention variable and two explanatory variables will be introduced in the model in the two following steps (see the Methodology Report).

3.4.4.1. Data description

1. Start of analysis and data load

- Use the menu <File, Open, Data ...> to open the file 'UK_KSI.sav'.

	DRIVERS	interv	TRKM	PPRICE	LDRIVERS	LTRKM	LPPRICE	YEAR	MONTH	DATE
1	1687,00	,00	9058,96	,10	7,43	9,11	-2,27	1969	1	JAN 1969
2	1508,00	,00	7685,03	,10	7,32	8,95	-2,28	1969	2	FEB 1969
3	1507,01	,00	9962,97	,10	7,32	9,21	-2,28	1969	3	MAR 1969
4	1385,01	,00	10954,99	,10	7,23	9,30	-2,29	1969	4	APR 1969
5	1632,00	,00	11822,98	,10	7,40	9,38	-2,29	1969	5	MAY 1969
6	1511,00	,00	12391,05	,10	7,32	9,42	-2,30	1969	6	JUN 1969
7	1559,00	,00	13460,03	,10	7,35	9,51	-2,27	1969	7	JUL 1969
8	1630,01	,00	14054,95	,10	7,40	9,55	-2,26	1969	8	AUG 1969
9	1579,00	,00	12106,04	,10	7,36	9,40	-2,27	1969	9	SEP 1969
10	1653,00	,00	11372,01	,10	7,41	9,34	-2,27	1969	10	OCT 1969
11	2151,99	,00	9833,99	,10	7,67	9,19	-2,28	1969	11	NOV 1969
12	2147,99	,00	9267,05	,10	7,67	9,13	-2,28	1969	12	DEC 1969

The data file consists of the following variables:

DRIVERS: The number of drivers killed or seriously injured.

Interv: The intervention variable

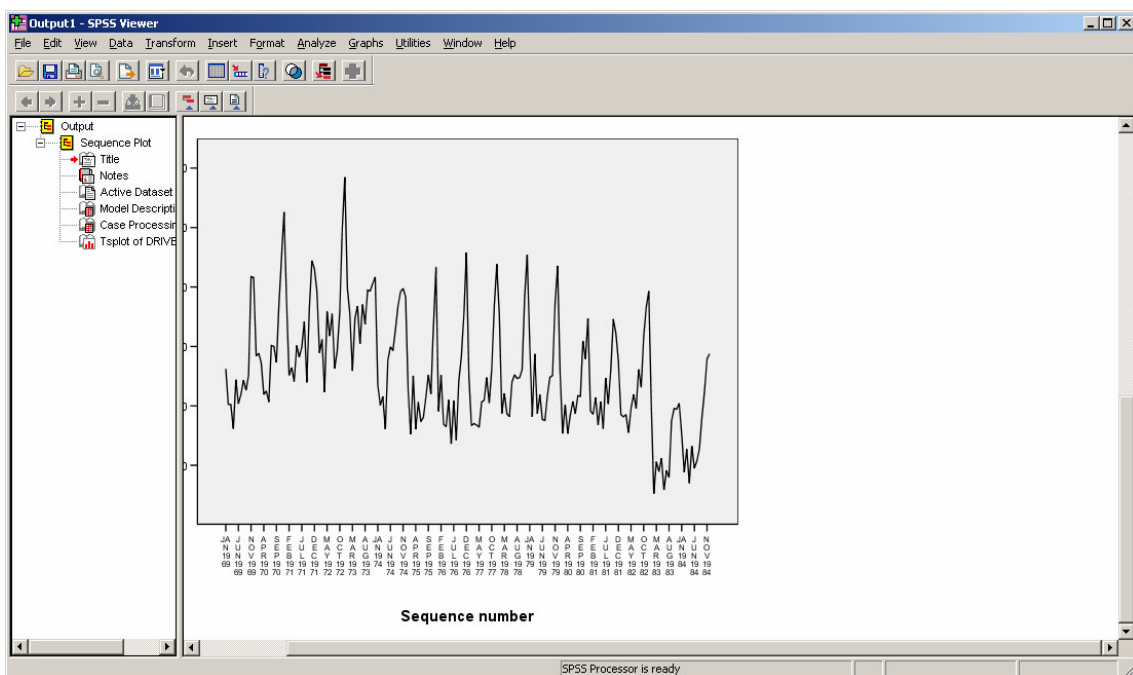
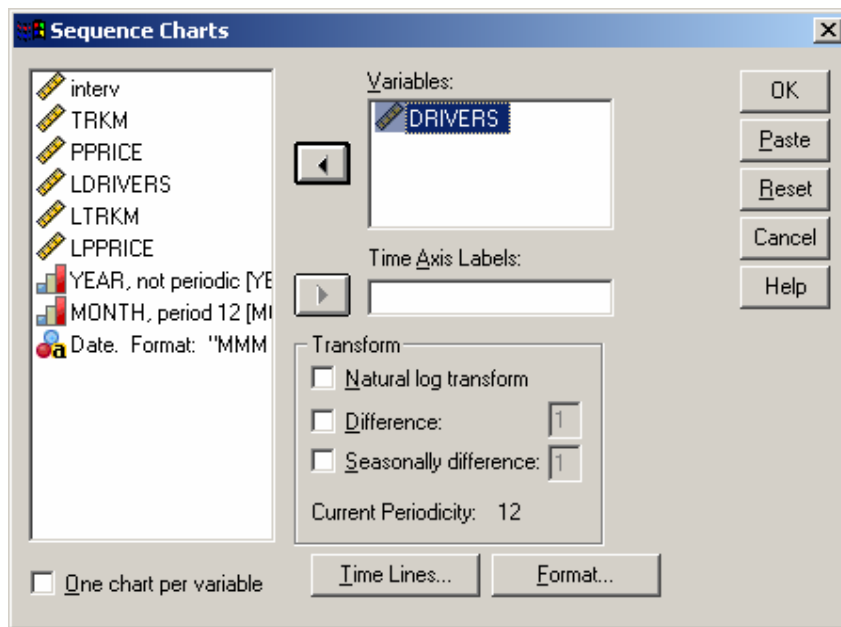
TRKM: The car traffic index

PPRICE: The petrol price.

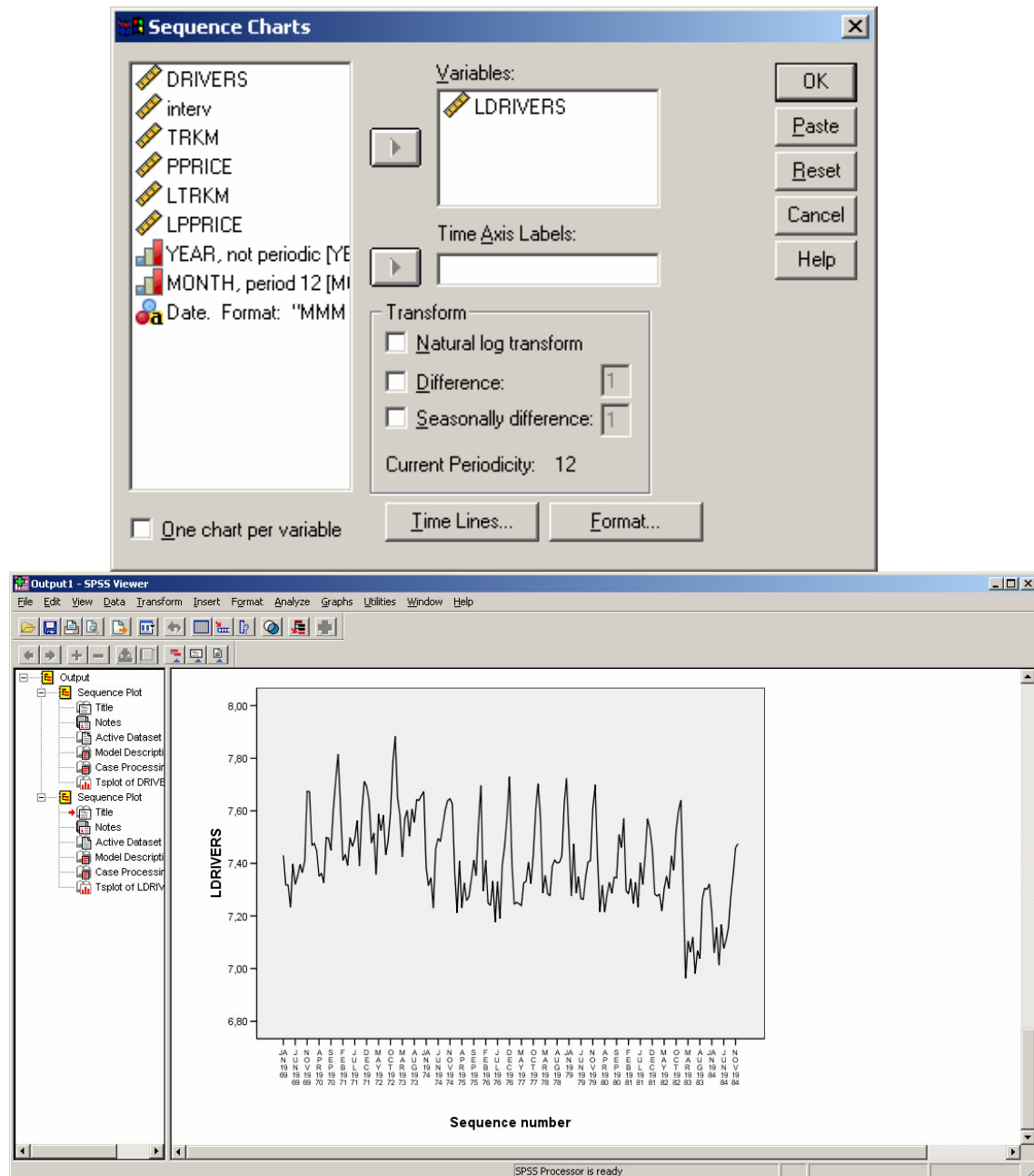
The log transformed variables of the preceding ones, to the exception of the intervention variable, are included in the data file, and three additional variables YEAR MONTH and DATE (see the Variable View, in which all variables are described).

2. Graphical diagnostics

- Click on Graphs..Sequence
- Move DRIVERS into the Variables list box



- Click on Graphs..Sequence
- Remove DRIVERS from the Variables list box
- Move DRIVERS into the Variables list box



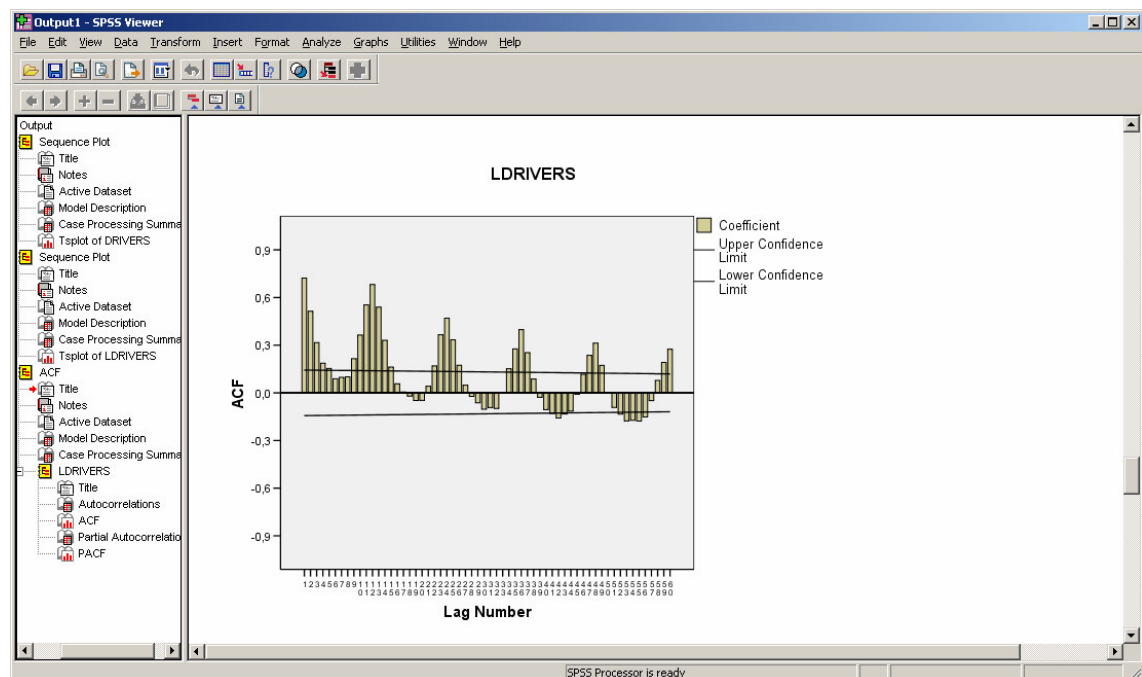
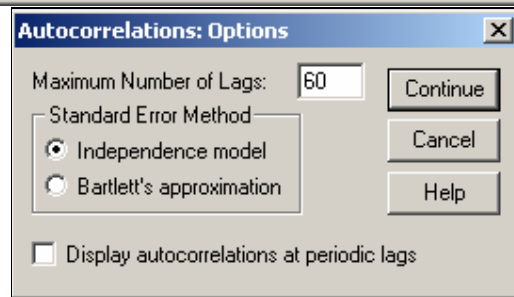
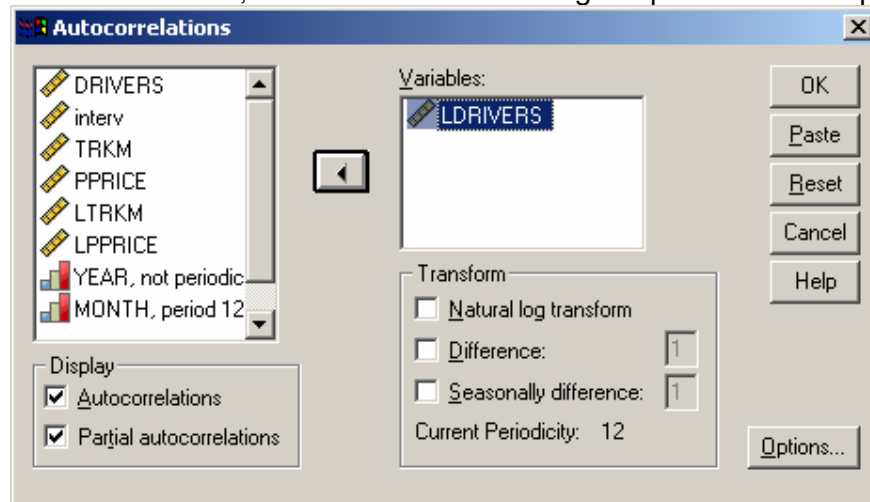
3.4.4.2. Model identification

The model identification consists in determining the six integers p , d , q (related to the non seasonal part of the model), and P, D, Q (related to the seasonal part of the model) in the multiplicative $ARIMA(p,d,q)(P,D,Q)_S$ formulation.

As already mentioned before, the ACF plot will be used to detect non stationarity in the data.

- Click on Graphs..Time Series..Autocorrelations
- Move the variable LDRIVERS into the Variables list box


- Click on the Options pushbutton
- Replace 16 with 60 in the Maximum number of lags text box
- Click on Continue, and then click on OK to get a plot of the ACF plot

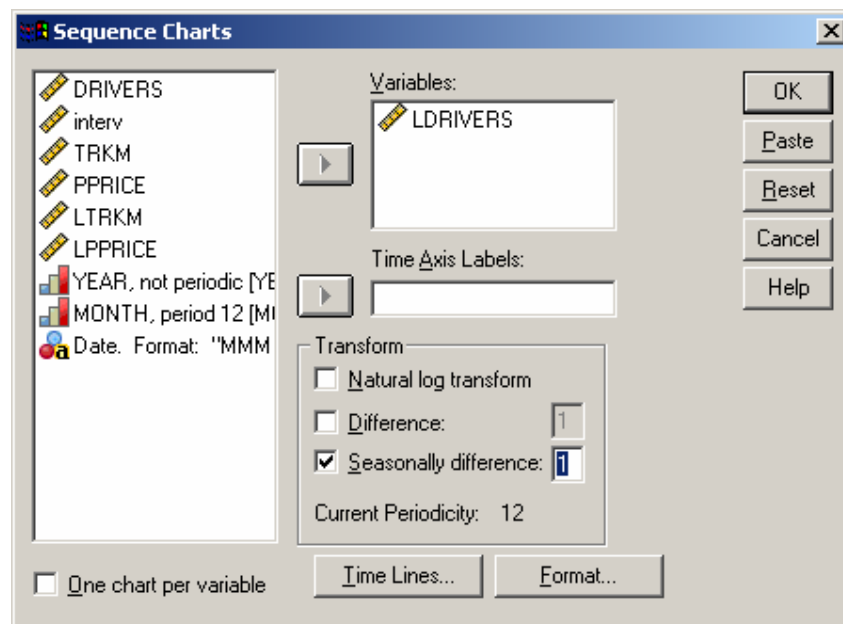


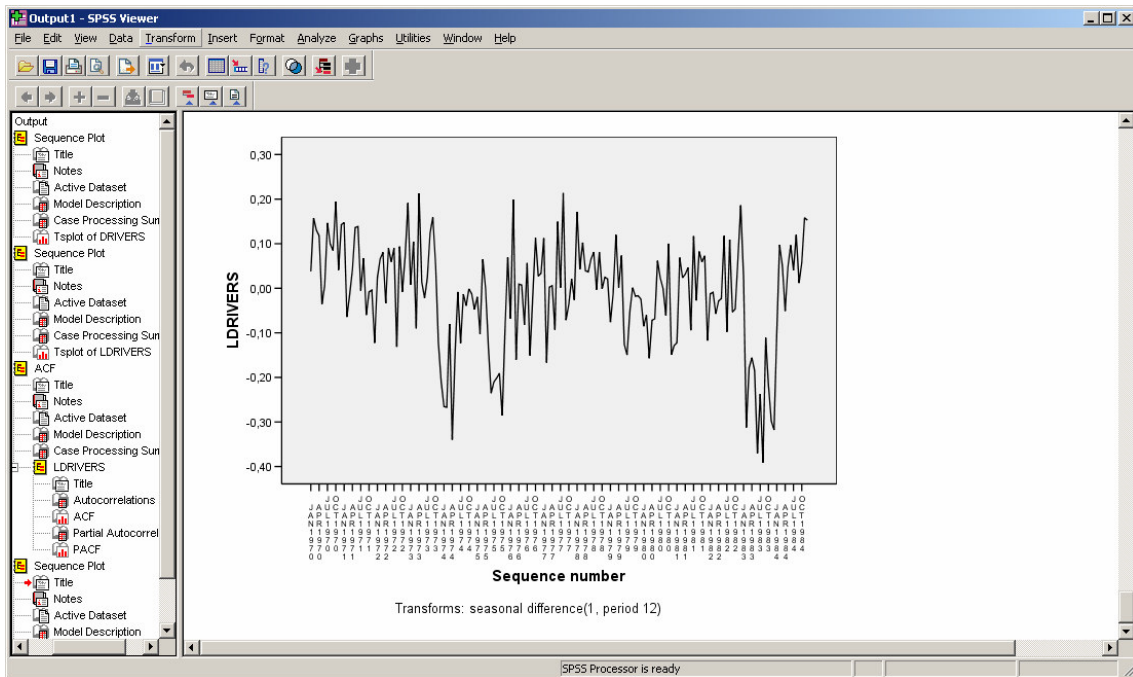
The ACF plot indicates obvious non stationarity, here again this is due to the fact that the autocorrelations do not decrease at an exponential rate, after a certain order.

We shall differentiate the series, by applying the “seasonal” difference filter $F(B) = 1 - B^{12}$ to the data, B being the backshift operator.



- Click on the Dialog Recall button  , and then click on Sequence Charts
- Click on the Seasonally difference check box, and verify that 1 is in the Seasonally difference text box.

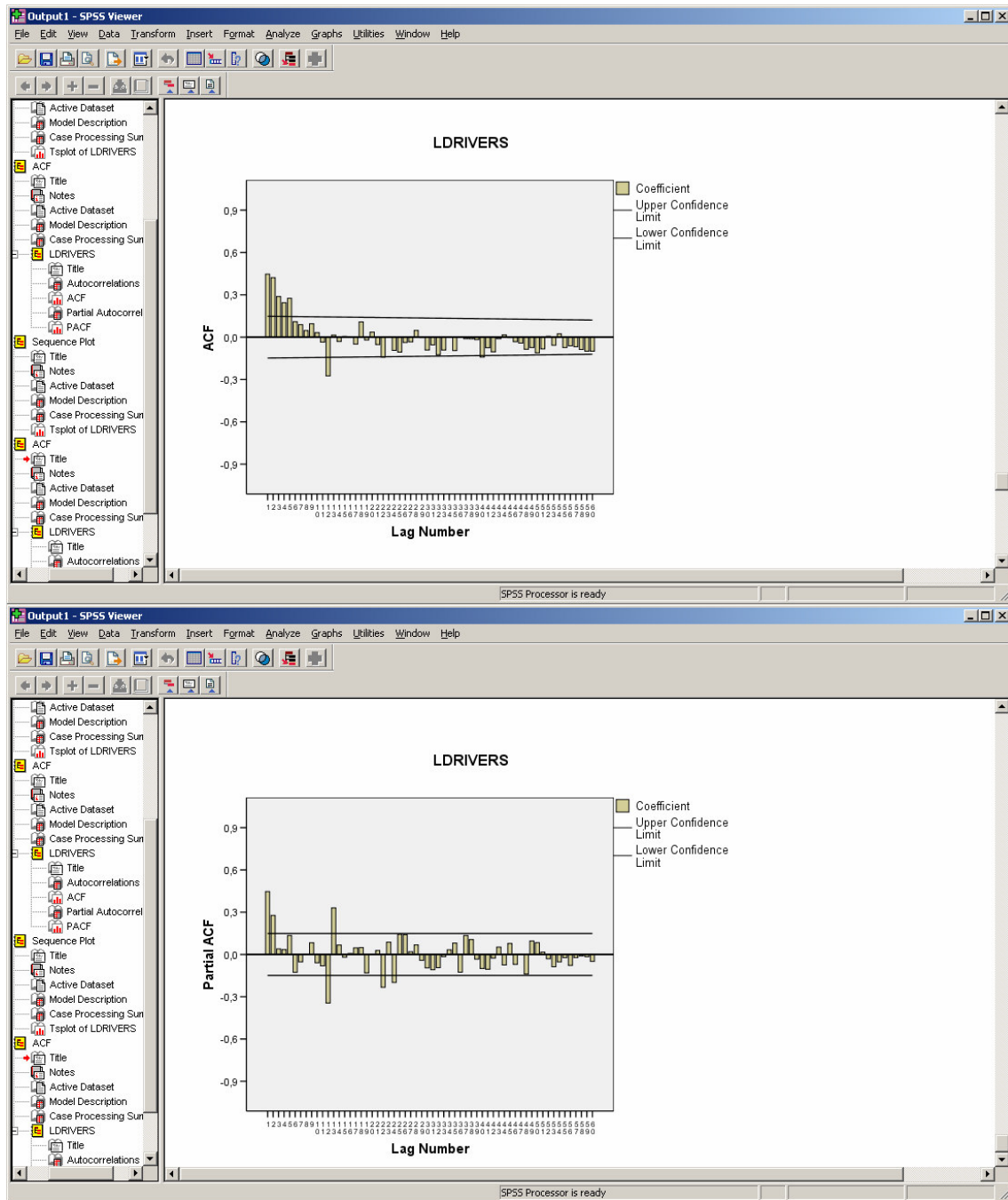




No indication of remaining non stationnarity in the seasonally filtered data can be found in the next-coming ACF plot.

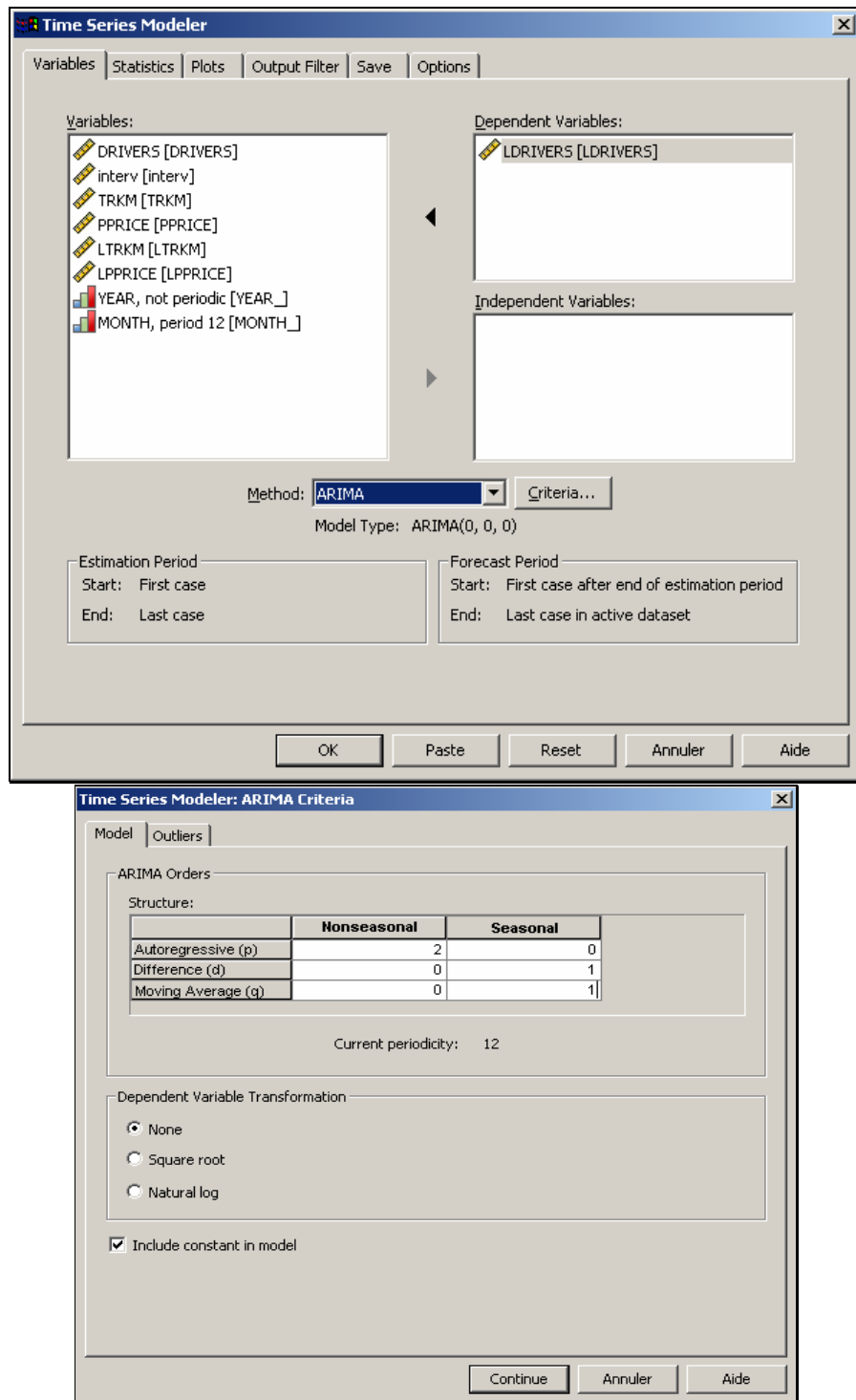
We shall therefore accept the hypothesis that this filtered series is stationary, which enables to retain $d=0$ for the non seasonal part of the general filter and $D=1$ for its seasonal part (see the Methodology Report)..

Second, the ACF and the PACF plots, taken together, lead to choose $p=2$, $q=0$ for the non seasonal part of the model, and $P=0$, $Q=1$ for the seasonal part of the model, indicating that the model is an $ARIMA(2,0,0)(0,1,1)_{12}$.

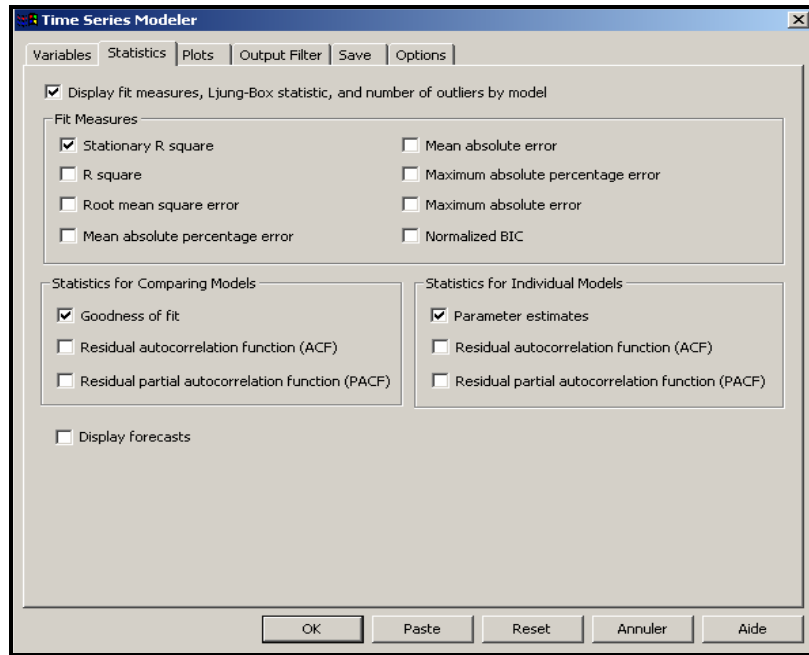


3.4.4.3. Model estimation and validation

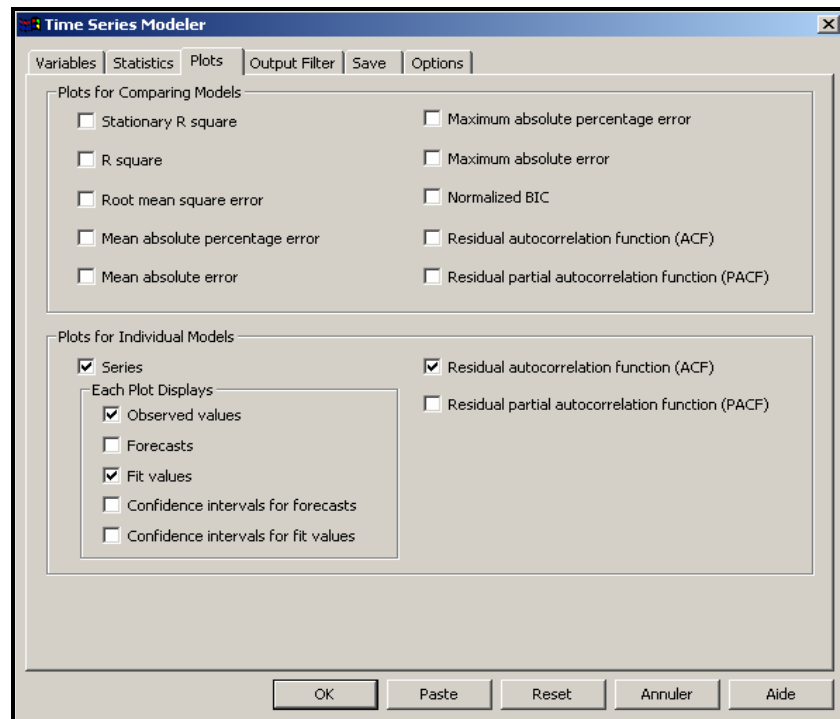
- Click on Analyze..Time Series..Create Models
- Move the variable LDRIVERS into the Dependent Variable(s) list box.
- Choose ARIMA in the Method list.
- Click on Criteria then specify the ARIMA model



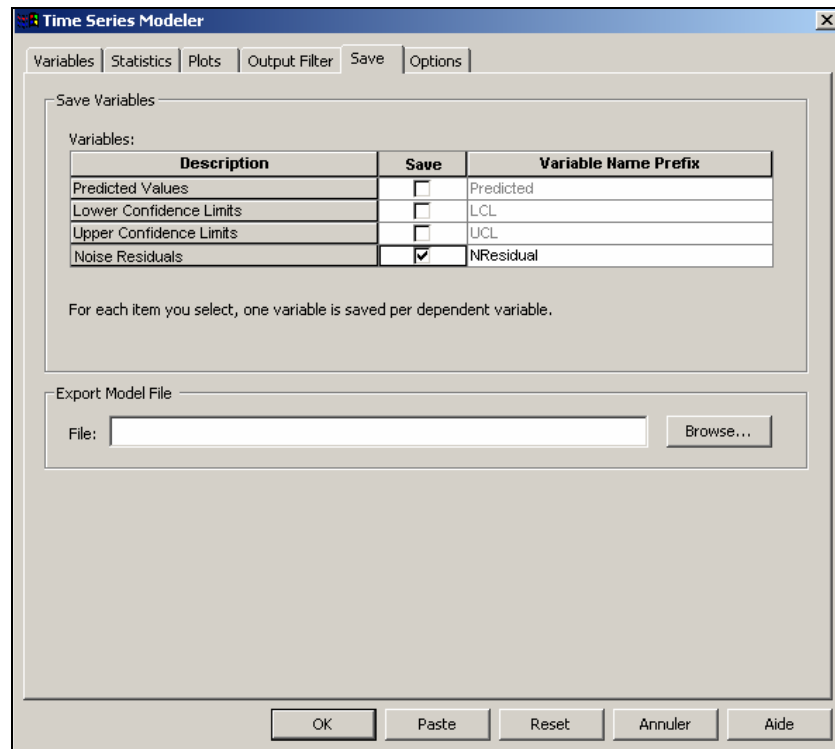
- Click on Statistics tab then check parameter estimates in the Statistics for Individual Models list.



- Click on Plots tab and check Observed Values, Fit values, Residual autocorrelation function (ACF), in Plots for Individual Models.



- Click on Save and check Noise Residuals then Click on OK.

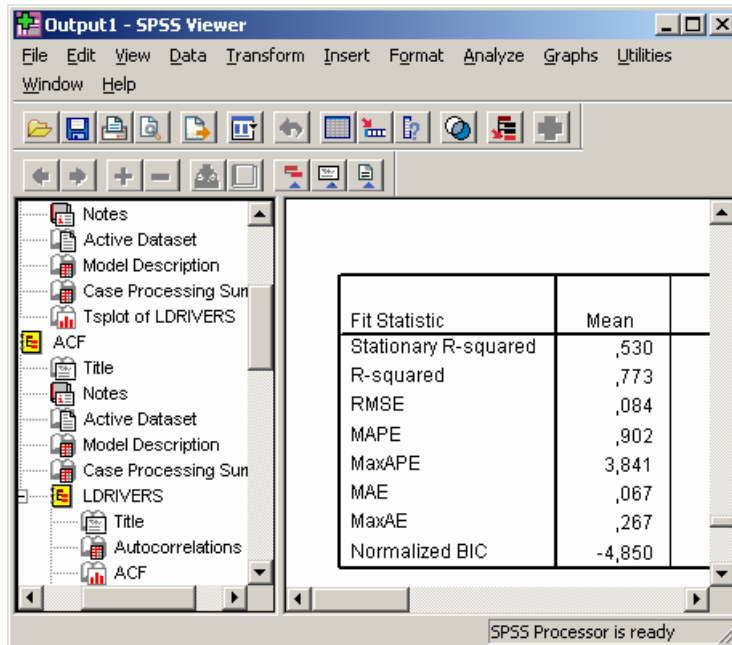


In the output window of SPSS, we find the statistics for the chosen model as follows:

Model Description ARIMA (2,0,0)(0,1,1)

The results obtained after the estimation procedure (maximum likelihood) are presented and commented below.

Model Fit



The screenshot shows the SPSS Output 1 - SPSS Viewer window. The left pane displays a tree view of the output, with 'ACF' selected. The right pane shows a table of fit statistics.

Fit Statistic	Mean
Stationary R-squared	,530
R-squared	,773
RMSE	,084
MAPE	,902
MaxAPE	3,841
MAE	,067
MaxAE	,267
Normalized BIC	-4,850

The stationary R-squared is only 53,0% (the model explains 53,0% of the variance of the filtered data, compared to a regression model), and much smaller than the R-square which is 77,3% (the model explains 77,3% of the variance of the initial data).

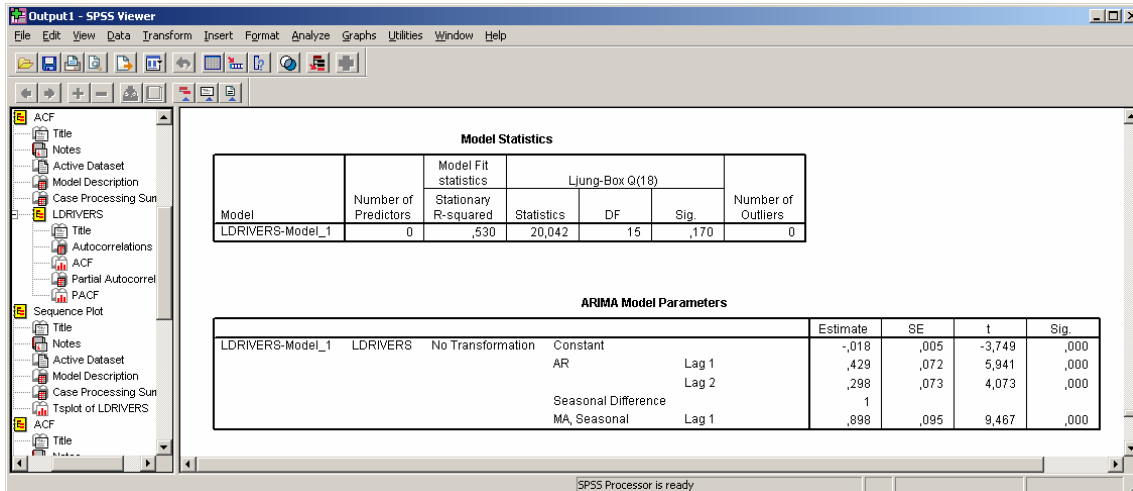
As for the different measures of the error made:

The mean absolute percentage error (MAPE) is 0,902% %, its highest value observed being 3,841%,

the mean absolute error (MAE) is 0,067, its highest value observed being 0,267,

the root mean square error (RMSE) is 0,084, and is a little higher than if it were computed as an arithmetic mean (0,067),

At last, the normalised BIC, which is -4,4850, is a goodness of fit measure that takes account of the parsimony of the model. Note that, as it is the case for the R-squared, its interest lies in comparisons between several models, and not in its absolute value.

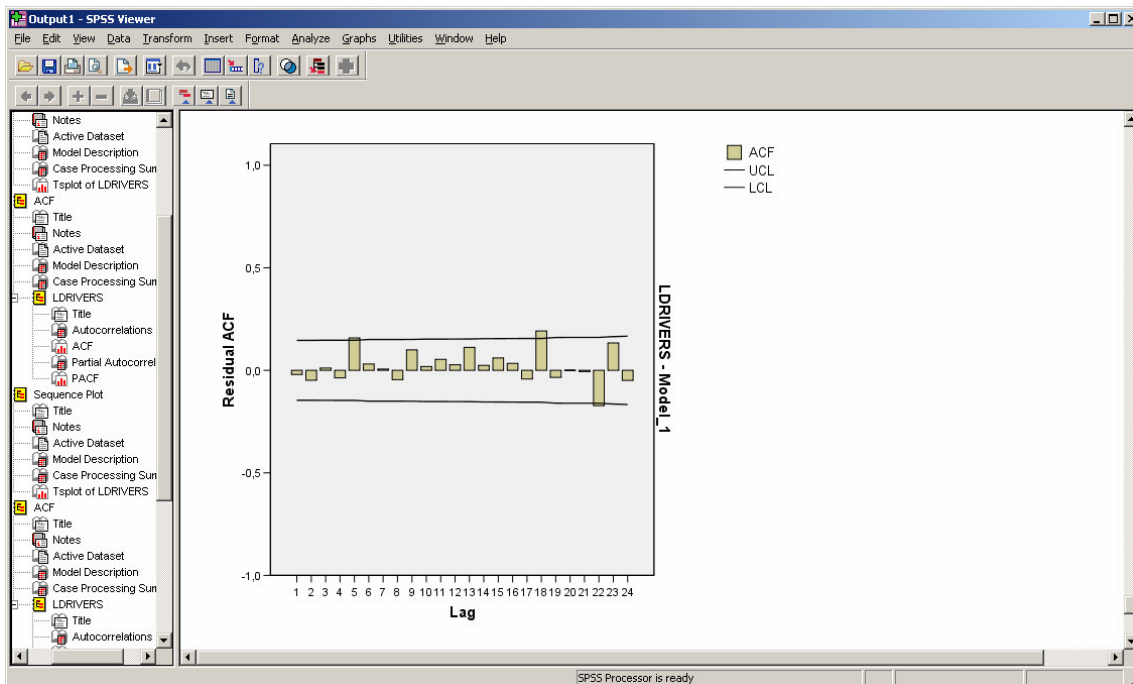


Model Statistics

In this case, this hypothesis of correct specification (global nullity of the autocorrelation of the residuals) is accepted, as the 0,17 value of the Ljung-Box statistic is more than 0.05.

ARIMA Model Parameters

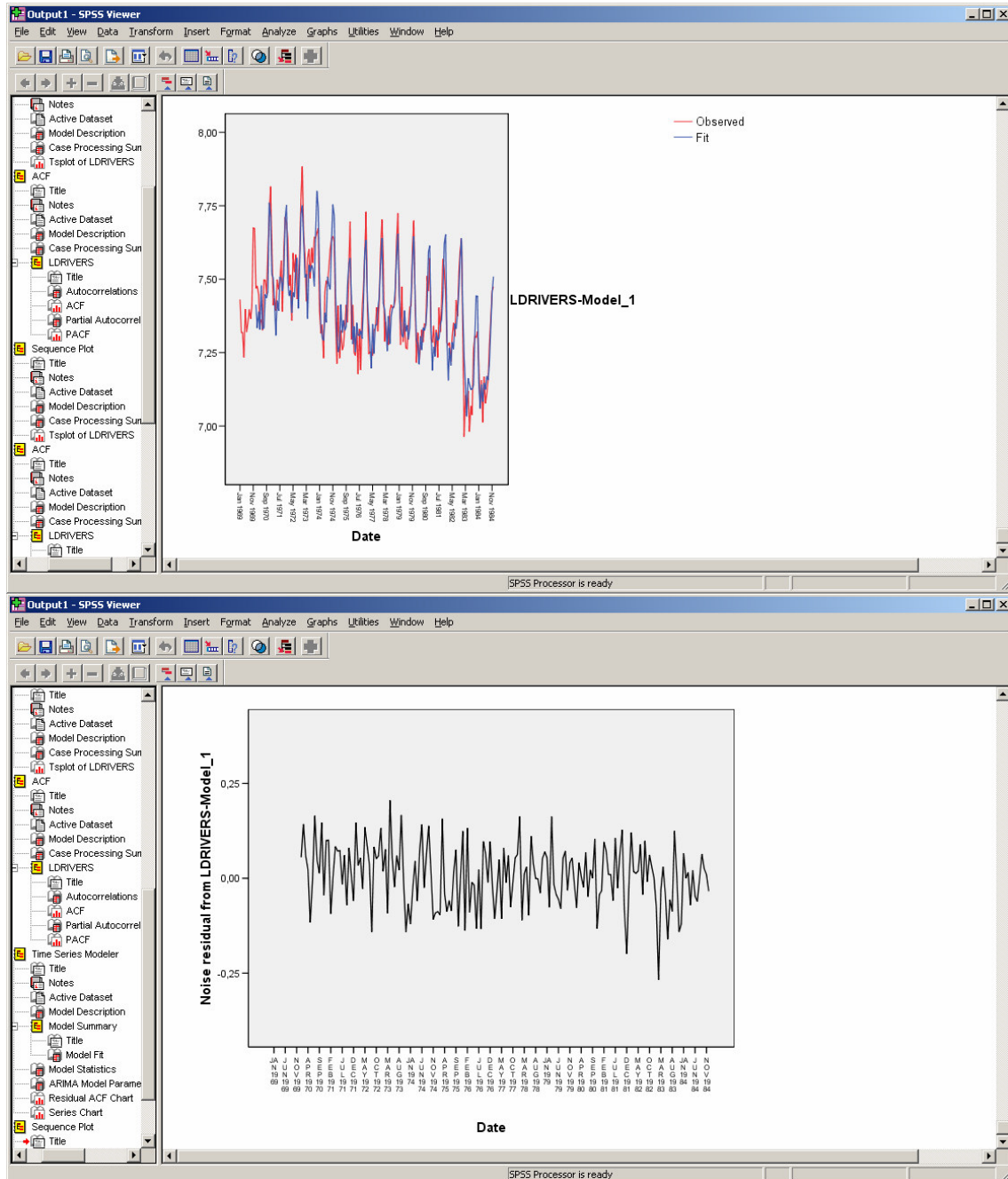
In this case, the hypothesis of nullity of each of the four parameters is rejected, and all parameters of the ARIMA model are to be considered as different from zero.



The residual ACF plot indicates that no very significant autocorrelation remains, up to order 24.

3.4.4.4. Graphical results and additional test

Graphical outputs

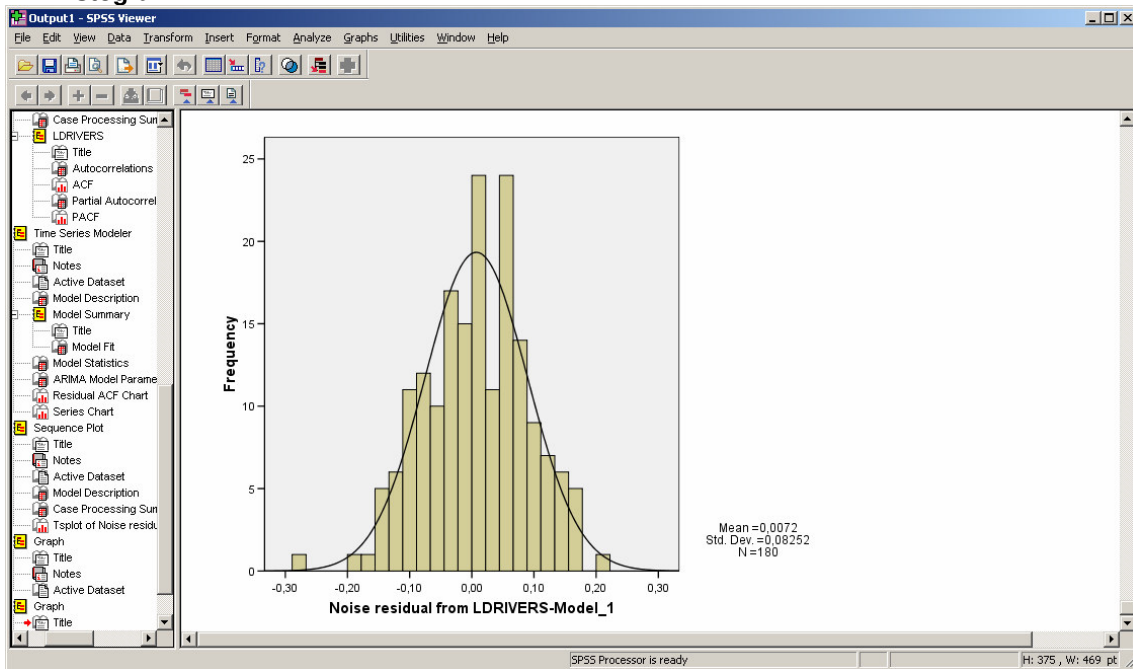


As in the preceding case, the first plot describes the development of the observed and fitted series, and the second plot the development of their difference (the estimated residuals)

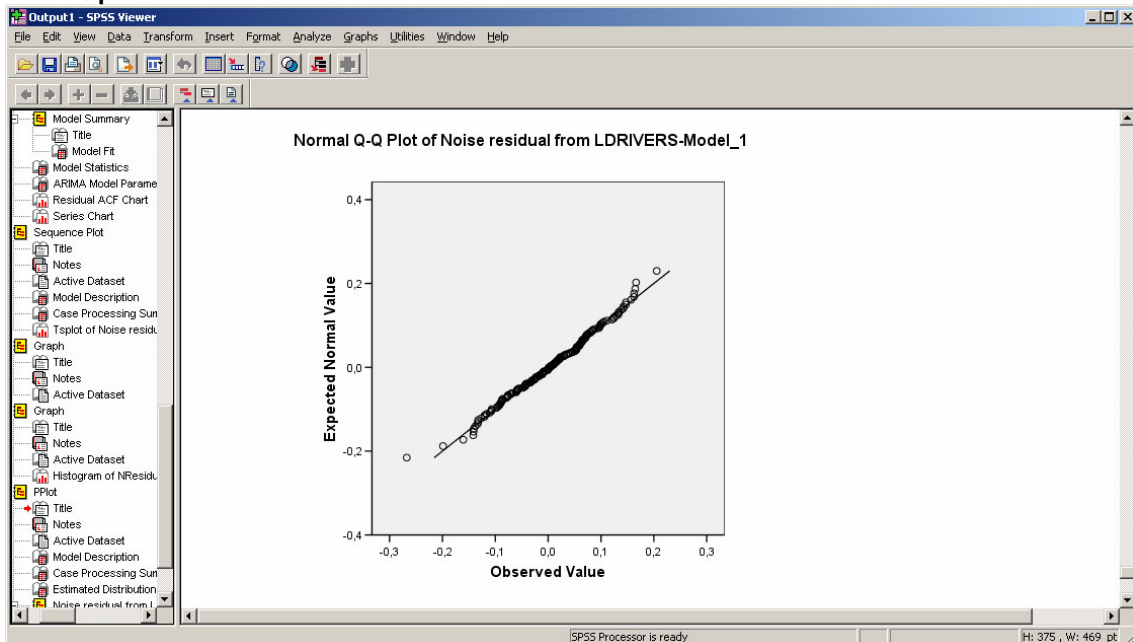
Note that the fitted data do not appear to stay one step behind the observed data, as it was the case for the Norwegian fatalities model. The strong seasonal pattern is reproduced, which takes over the adjustment of the trend.

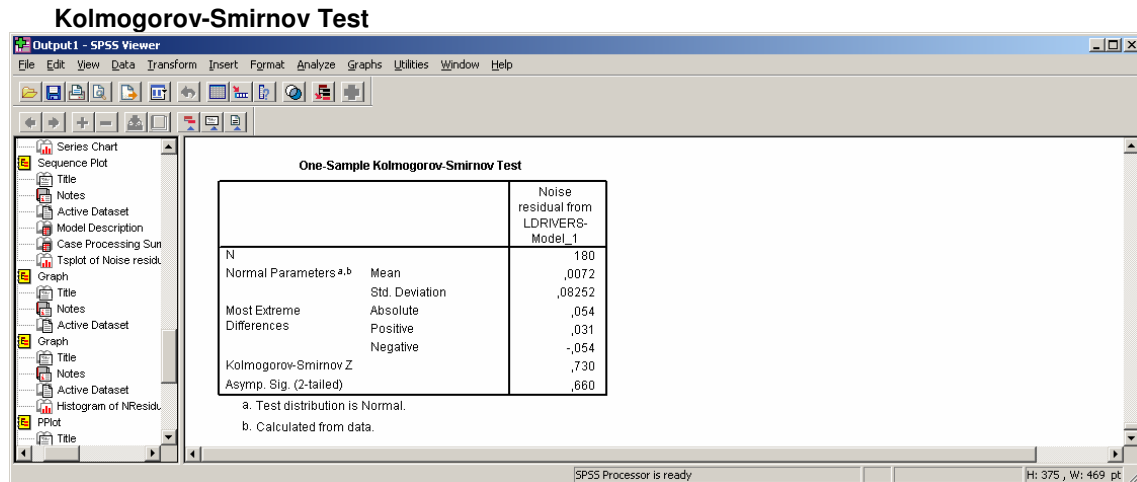
Normality test

Histogram



QQ-plot





In this case, this hypothesis of normality of the residuals is accepted, because the 0,66 value of the Asymp. Sig. (2-tailed) is more than 0.05 (at the usual 95% confidence level).

3.4.4.5. Intervention variable

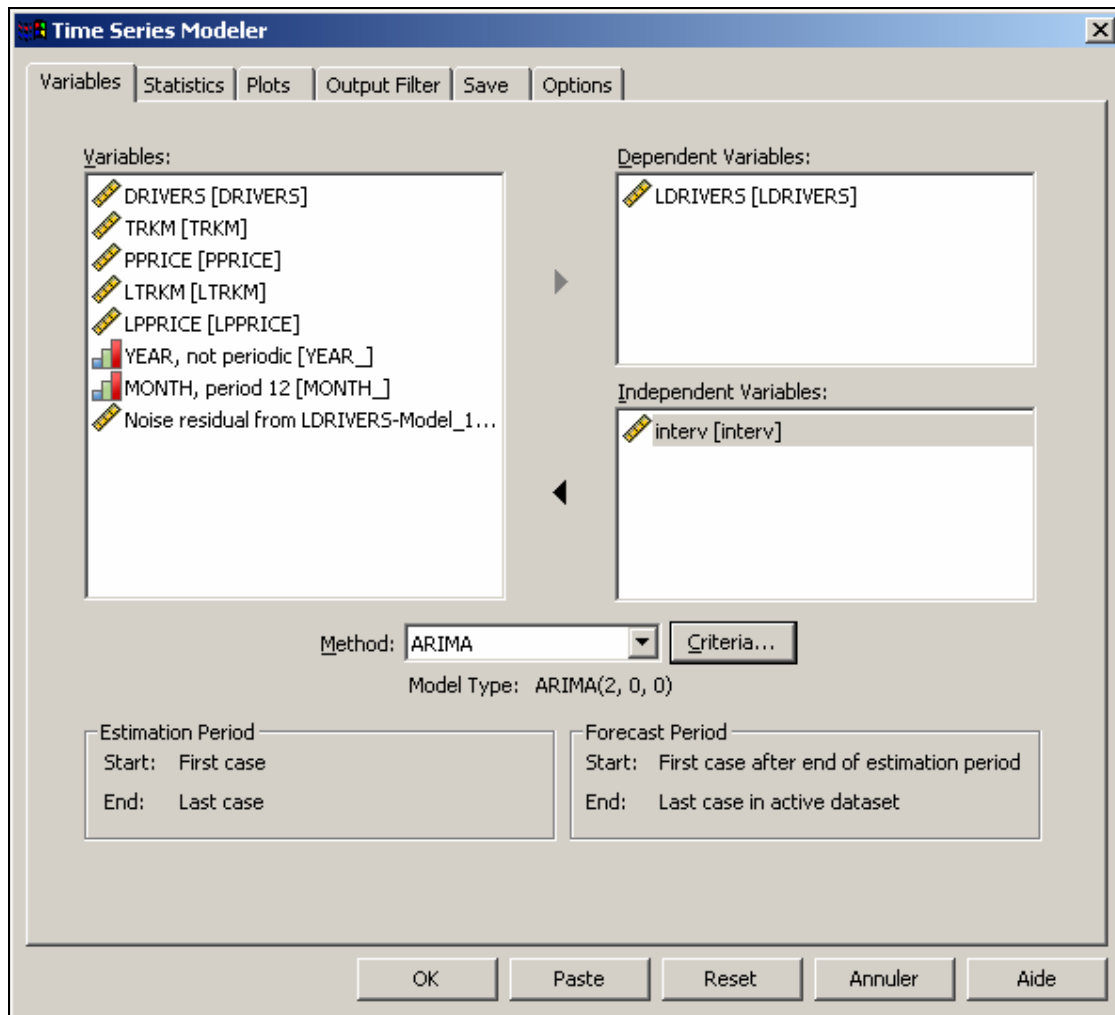
In this section, we will add an intervention variable to the model, in view of performing a so-called intervention analysis.

Data description

The reason for the introduction of the intervention variable is the introduction of the seat belt law in February 1983. The variable will therefore be equal to 1, February 1983 onwards, and equal to 0 before (see the Methodology Report).

Model estimation and validation

- Click on Analyze..Time Series..Create Models
- Move the variable LDRIVERS into the Dependent Variables list box and the variable interv in the Independent variables list box.
- Choose ARIMA in the Method list, click on Criteria and then specify the model you want to estimate.



The Transfer Function tab (only present if independent variables are specified) will now be used to define the call to the intervention variable.

The Transfer Function tab allows defining transfer functions for the independent variables specified on the Variables tab. In this case, the intervention variable is the only independent variable, and the indication to be given is that a seasonal difference filter is used for that variable (see the Methodology Report).

Time Series Modeler: ARIMA Criteria

Model Transfer Function Outliers

Independent Variables:

interv [interv]

Transfer Function Orders

Structure:

	Nonseasonal	Seasonal
Numerator	0	0
Denominator	0	0
Difference	0	1

Current periodicity: 12

Delay: 0

Transformation

☒ None

☐ Square root

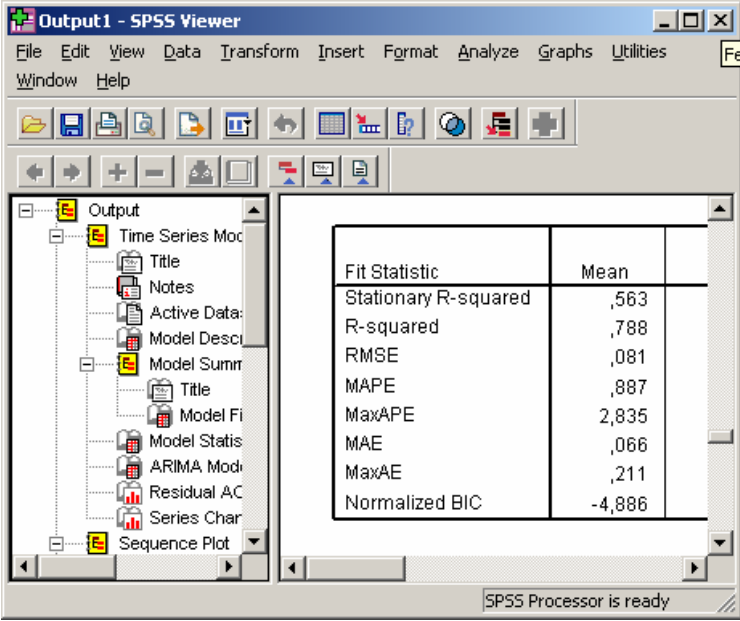
☐ Natural log

Continue Annuler Aide

Here are the SPSS results for the specified model:

Model Description ARIMA(2,0,0)(0,1,1) with intervention variable

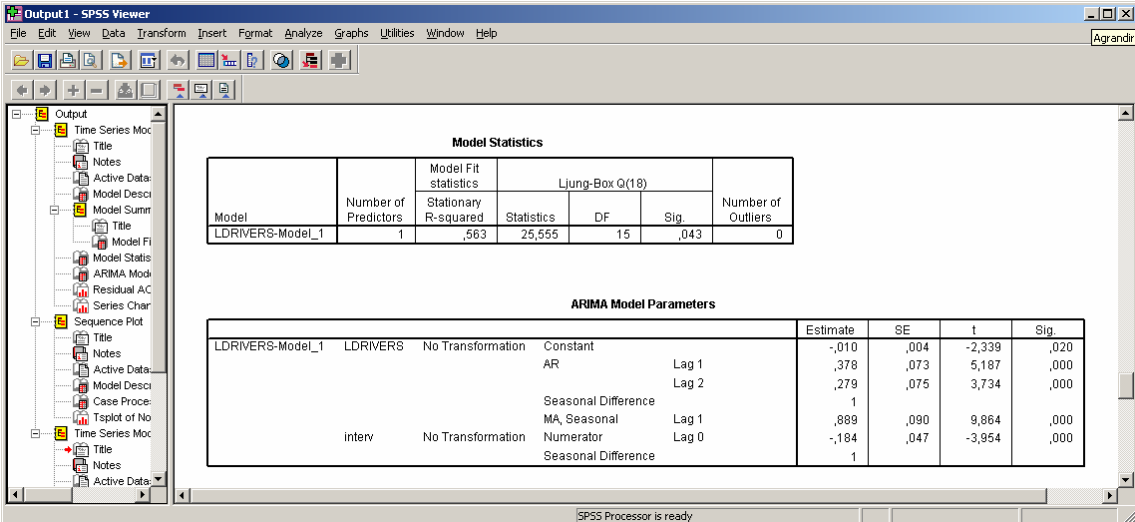
Model Fit



The screenshot shows the SPSS Output1 - SPSS Viewer window. The left pane displays a tree view of the output, with 'Model Fit' selected. The right pane shows the 'Fit Statistic' table with the following data:

Fit Statistic	Mean
Stationary R-squared	,563
R-squared	,788
RMSE	,081
MAPE	,887
MaxAPE	2,835
MAE	,066
MaxAE	,211
Normalized BIC	-4,886

The addition of the intervention to the model has improved the goodness-of-fit: the Stationary R-squared has increased (from 0,53 to 0,563).
 The R-squared has increased (from 0,773 to 0,788)
 the MAPE has decreased (from 3,841 to 2,835).
 the Normalized BIC has decreased (from -4,590 to -4,886).



The screenshot shows the SPSS Output1 - SPSS Viewer window with two tables displayed. The first table is 'Model Statistics' and the second is 'ARIMA Model Parameters'.

Model Statistics

Model	Number of Predictors	Model Fit statistics		Ljung-Box Q(18)		Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
LDRIVERS-Model_1	1	,563	25,555	15	,043	0

ARIMA Model Parameters

				Estimate	SE	t	Sig.
LDRIVERS-Model_1	LDRIVERS	No Transformation	Constant	-,010	,004	-2,339	,020
			AR Lag 1	,378	,073	5,187	,000
			AR Lag 2	,279	,075	3,734	,000
			Seasonal Difference	1			
			MA, Seasonal Lag 1	,889	,090	9,864	,000
	interv	No Transformation	Numerator Lag 0	-,184	,047	-3,954	,000
			Seasonal Difference	1			

Model Statistics

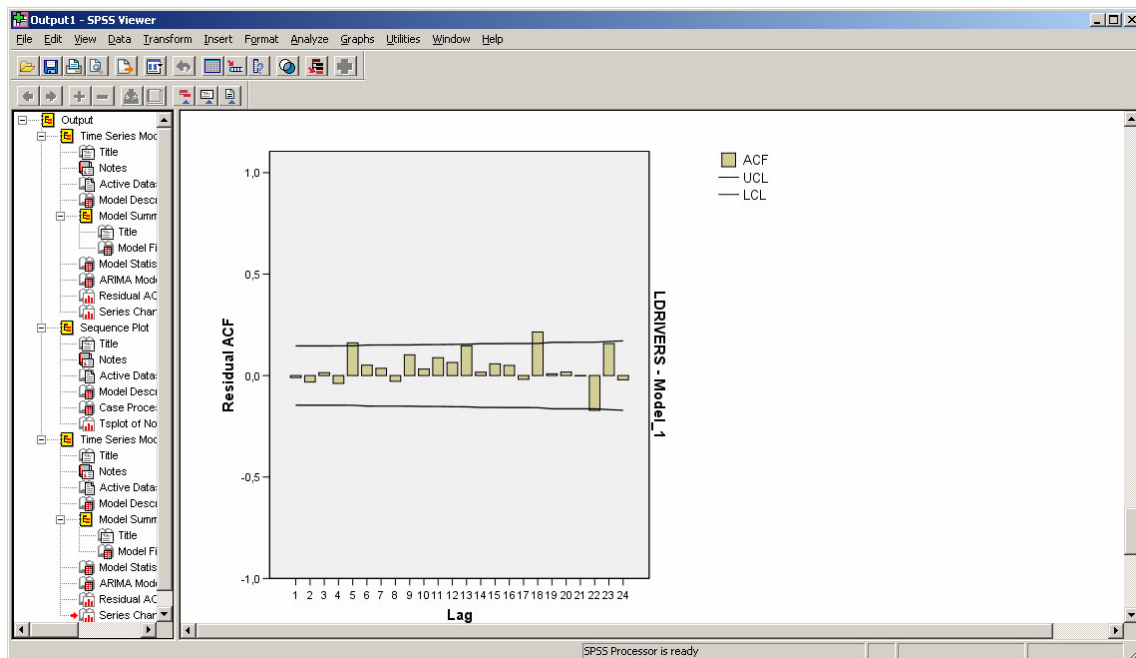
Note that, in this case, the Ljung-Box statistic value of 0,043 is smaller than 0,05, which indicates that the hypothesis of global nullity of the autocorrelation

of the residuals, up to order 18, has to be rejected, at the usual 95% confidence level.

ARIMA Model Parameters

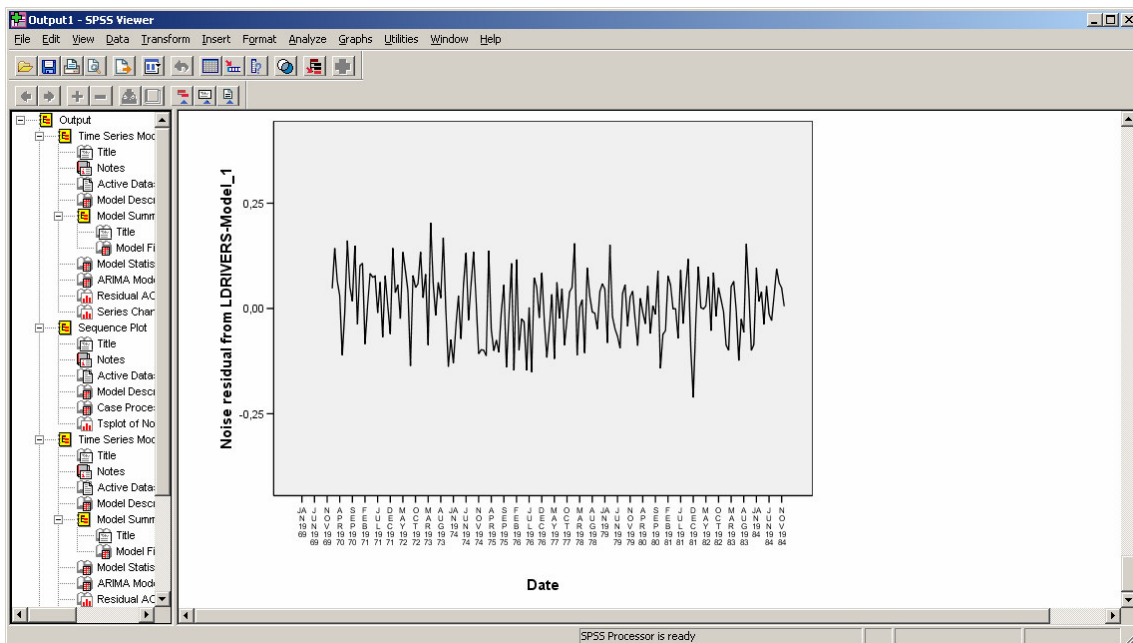
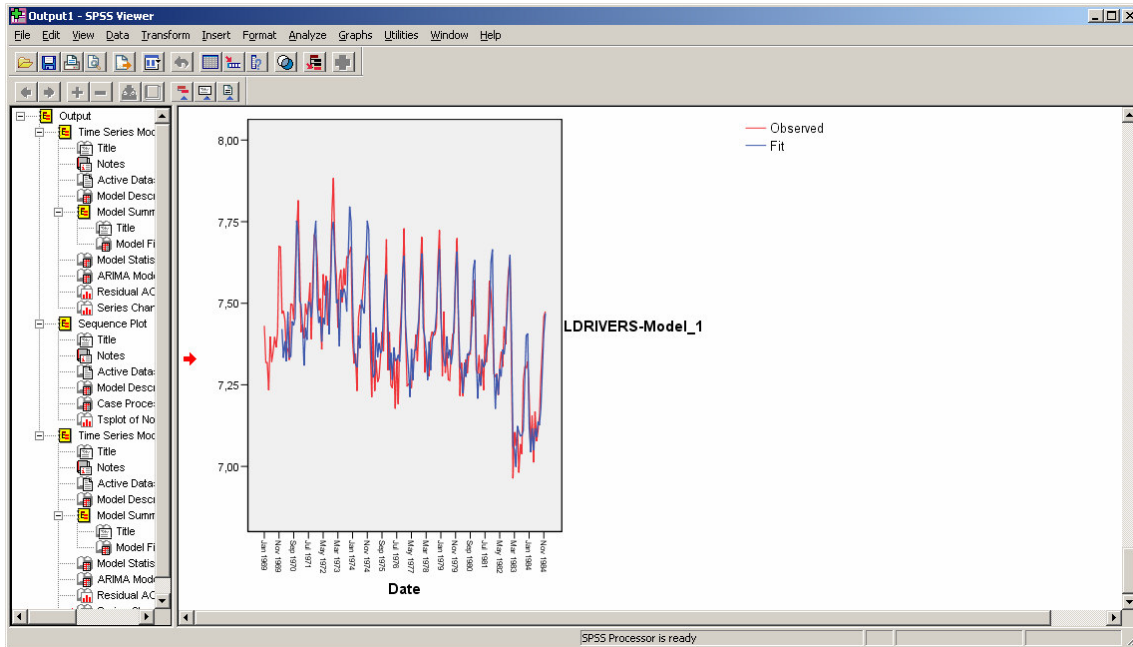
In this case, the hypothesis of nullity of all parameters is rejected: all four parameters related to the dynamics are to be considered as different from zero and the intervention parameter too.

The following ACF plot of the residuals indicate that the autocorrelation of order 5, and of order 18, for instance, differ significantly from zero, at this usual confidence level

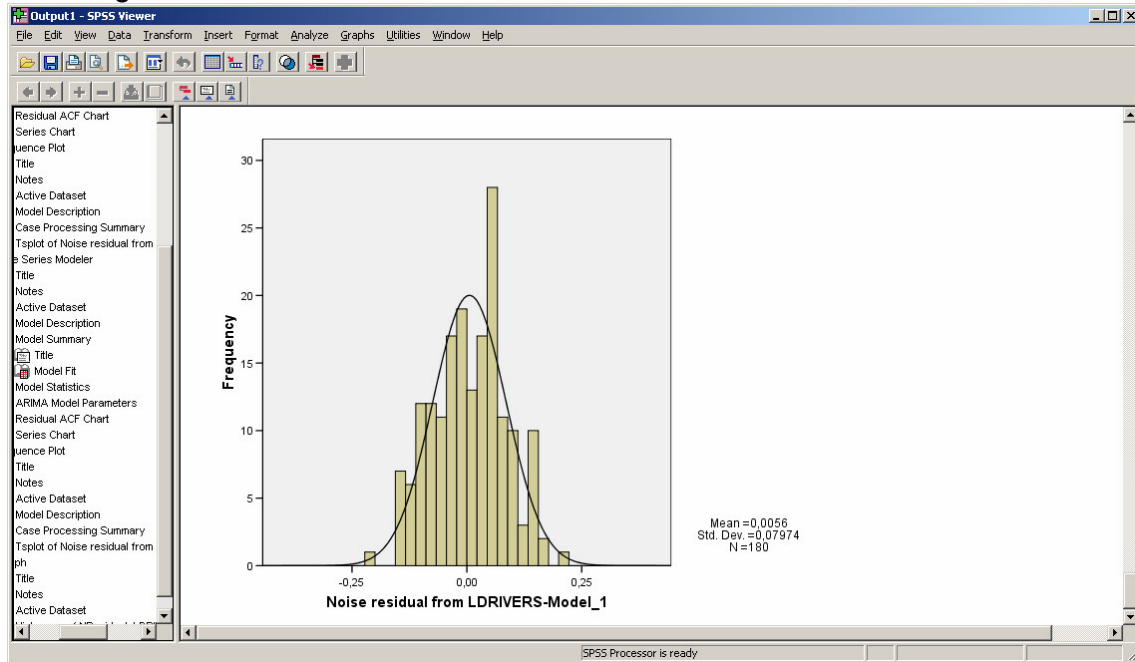


Graphical results and additional test

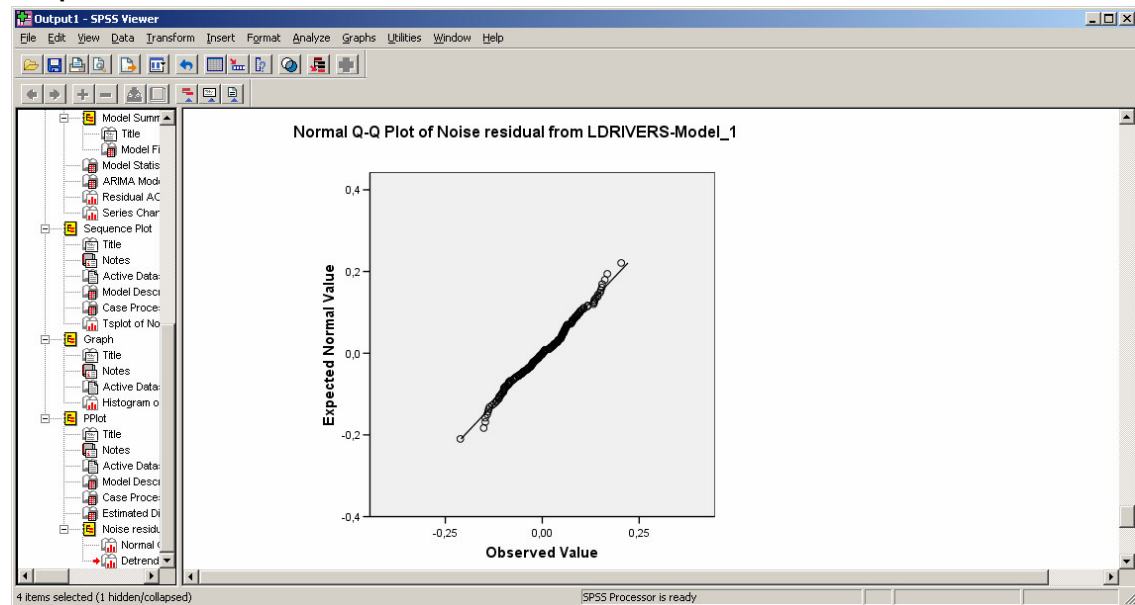
The two usual graphical outputs are given below, followed by the normality test results.

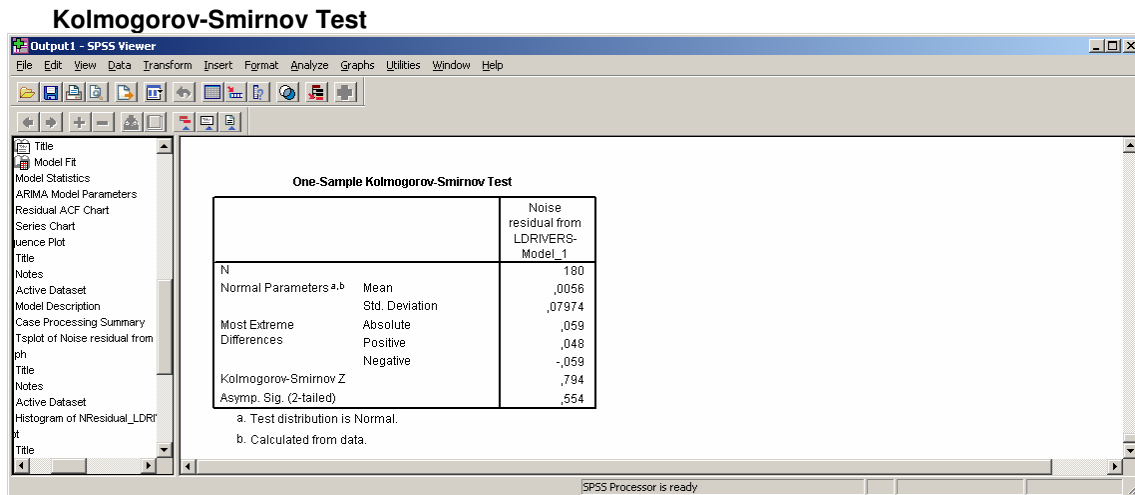


Histogram



QQ-plot





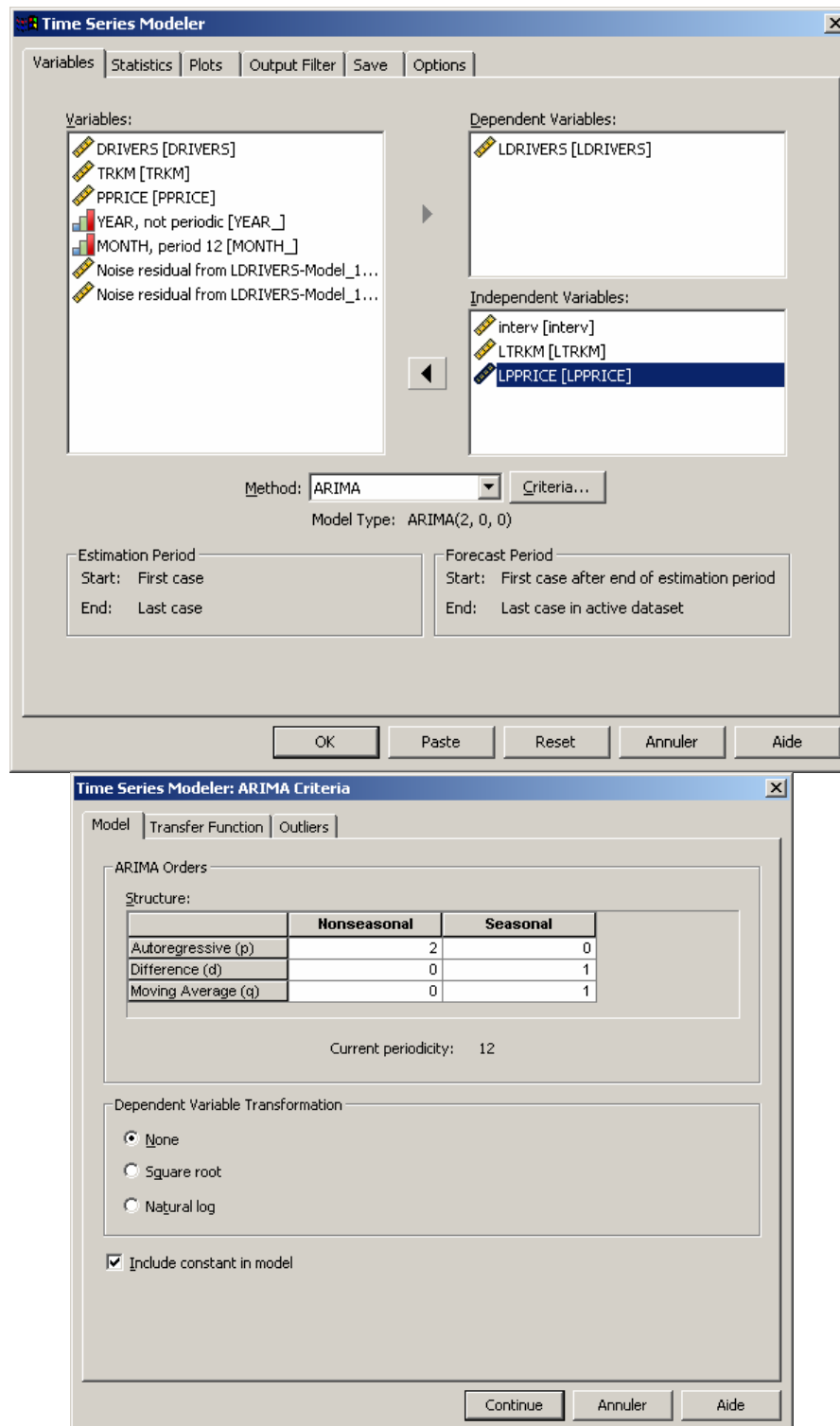
In this case, this hypothesis of normality of the residuals is accepted, because the 0,554 value of the Asymp. Sig. (2-tailed) is more than 0.05 (at the usual 95% confidence level).

3.4.4.6. Intervention and explanatory variables

In this section, we will add two explanatory variables, LTRKM, LPPRICE, as defined above in the database. The seat belt law intervention variable from the previous section will be kept in the model.

Model estimation and validation

- Click on Analyze..Time Series..Create Models
- Move the variable LDRIVERS into the Dependent Variables list box and the variables interv, LTKRM and LPPRICE in the Independent variables list box.



As the intervention variable is the only independent variable, the indication to be given is that a seasonal difference filter is performed on that variable (see the Methodology Report).

Time Series Modeler: ARIMA Criteria

Model Transfer Function Outliers

Independent Variables:

- interv [interv]
- LTRKM [LTRKM]
- LPPRICE [LPPRICE]

Transfer Function Orders

Structure:

	Nonseasonal	Seasonal
Numerator	0	0
Denominator	0	0
Difference	0	1

Current periodicity: 12

Delay: 0

Transformation

☒ None

☐ Square root

☐ Natural log

Continue Annuler Aide

Model Description ARIMA(2,0,0)(0,1,1) with explanatory variable

Model Fit

Output 1 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

4 items selected (1 hidden/collapsed) SPSS Processor is ready

Fit Statistic	Mean
Stationary R-squared	,590
R-squared	,802
RMSE	,079
MAPE	,860
MaxAPE	2,336
MAE	,064
MaxAE	,177
Normalized BIC	-4,881

The addition of the two explanatory variables to the model has still improved the goodness-of-fit:

The Stationary R-squared has increased to 0,59.

The R-squared has increased to 0,802

The MAPE has decreased to 0,86 %.

The only exception is that the Normalized BIC has increased a little, from -4,886 to -4,881, but remains smaller than in the pure ARIMA model (64,841).

Output 1 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

4 items selected (1 hidden/collapsed) SPSS Processor is ready

Model	Number of Predictors	Model Fit statistics		Ljung-Box Q(18)		Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
LDRIVERS-Model_1	3	,590	23,289	15	,078	0

					Estimate	SE	t	Sig.
LDRIVERS-Model_1	LDRIVERS	No Transformation	Constant		-,015	,006	-2,463	,015
			AR	Lag 1	,283	,075	3,800	,000
				Lag 2	,235	,077	3,072	,002
			Seasonal Difference		1			
			MA, Seasonal	Lag 1	,857	,078	10,930	,000
	Interv	No Transformation	Numerator	Lag 0	-,163	,037	-4,464	,000
			Seasonal Difference		1			
	LTRKM	No Transformation	Numerator	Lag 0	,210	,134	1,561	,120
			Seasonal Difference		1			
	LPPRICE	No Transformation	Numerator	Lag 0	-,297	,095	-3,132	,002
			Seasonal Difference		1			

Model Statistics

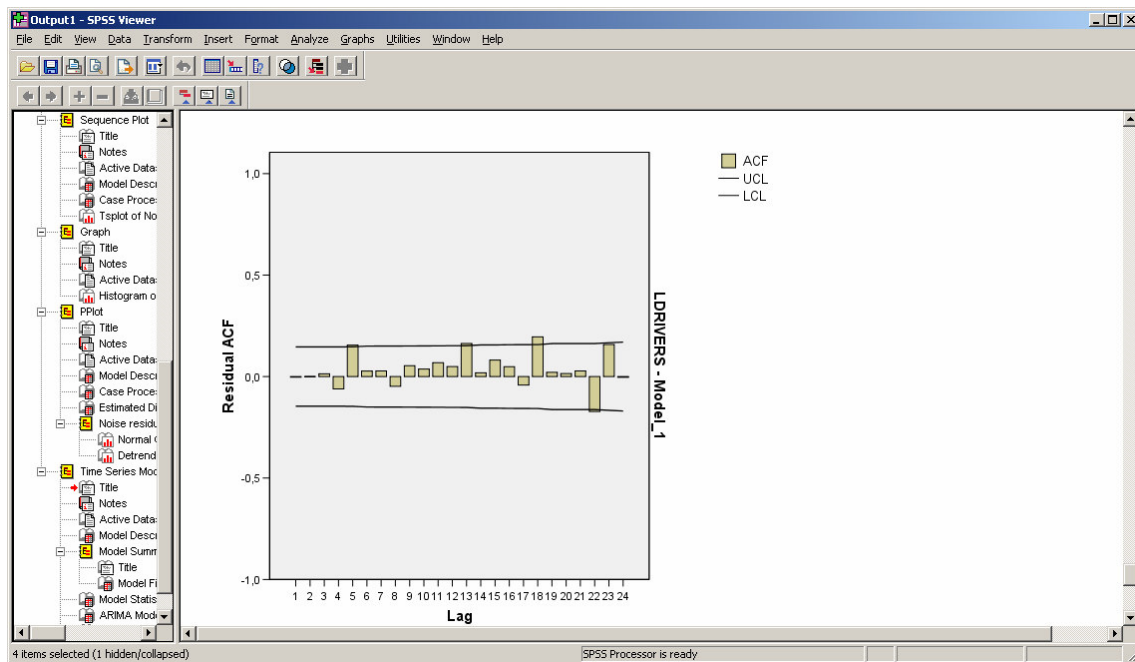
The hypothesis of global nullity of the autocorrelation of the residuals is still accepted, as the statistic is 0,078, and higher than 0,05 (at the 95% confidence

level).

ARIMA Model Parameters

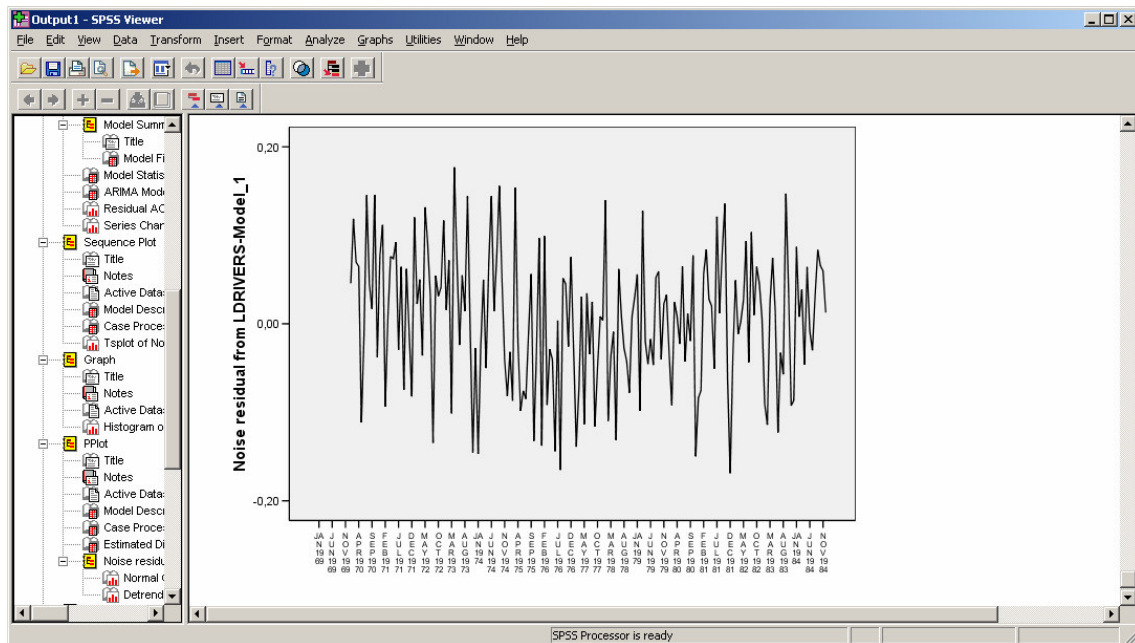
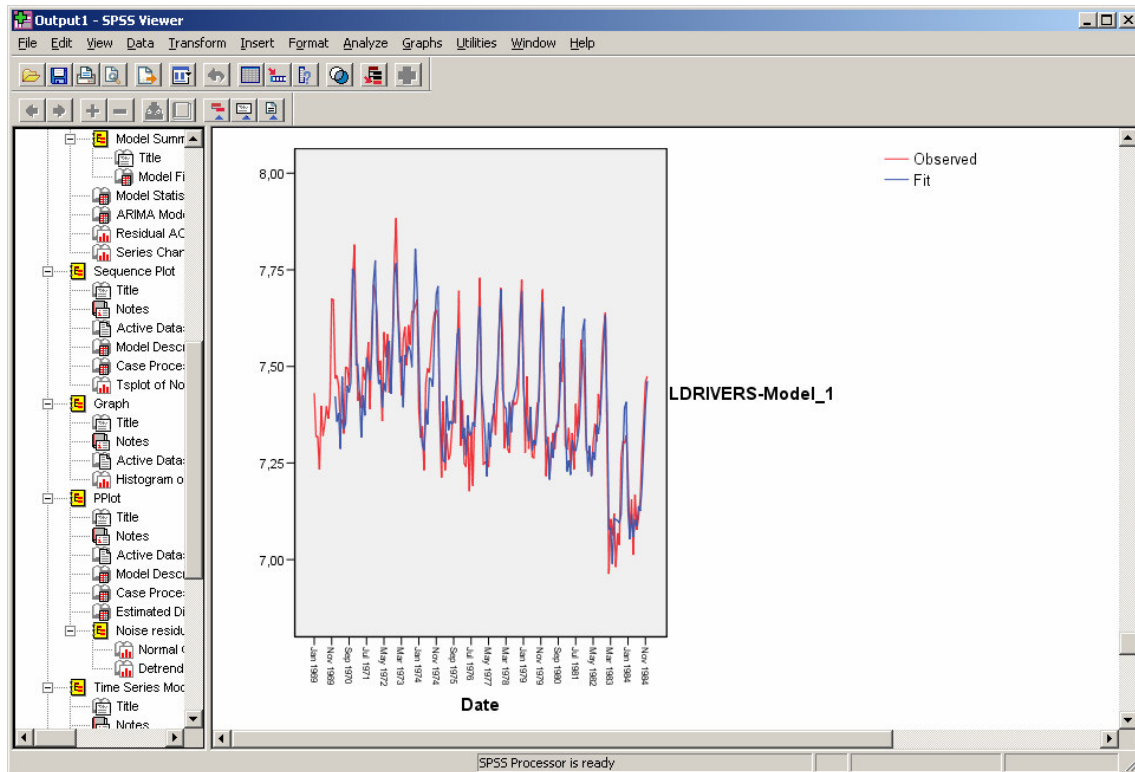
In this case, the hypothesis of nullity of the model parameters is rejected (at the 95% confidence level), except for the log of the traffic index variable parameter: all parameters related to the dynamics are to be considered as different from zero, the petrol price parameter and the intervention parameter too.

Note that, in case the confidence level is lowered to 70% for instance (t-value between 1 and 2), the parameter related to the traffic index variable would also be considered as different from zero too.

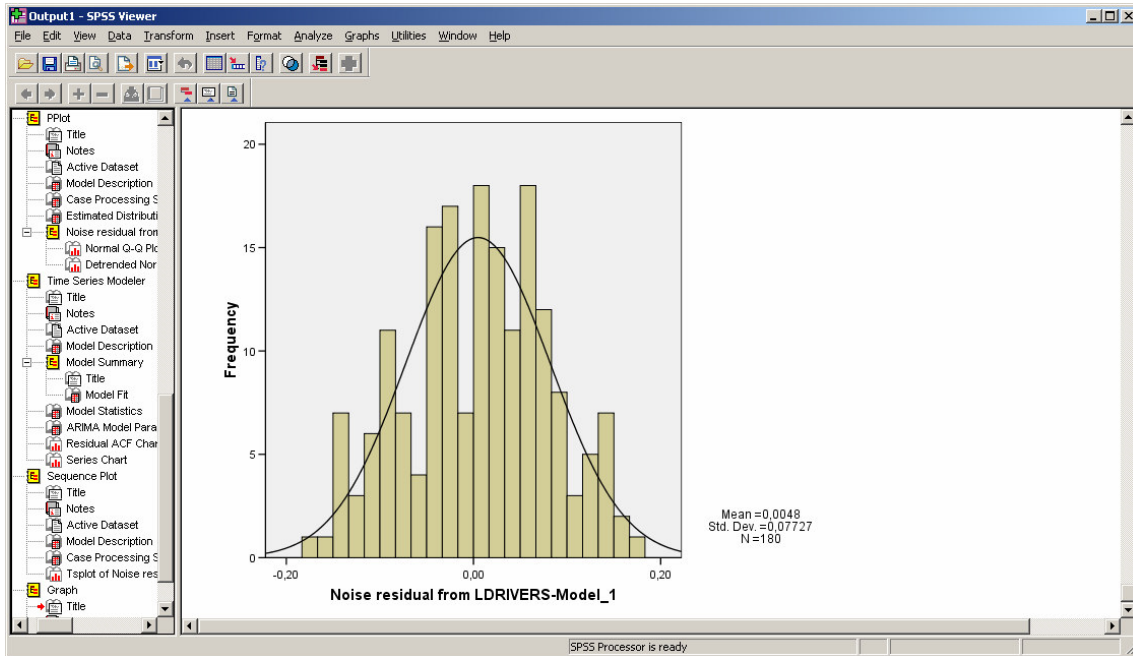


Graphical results and additional test

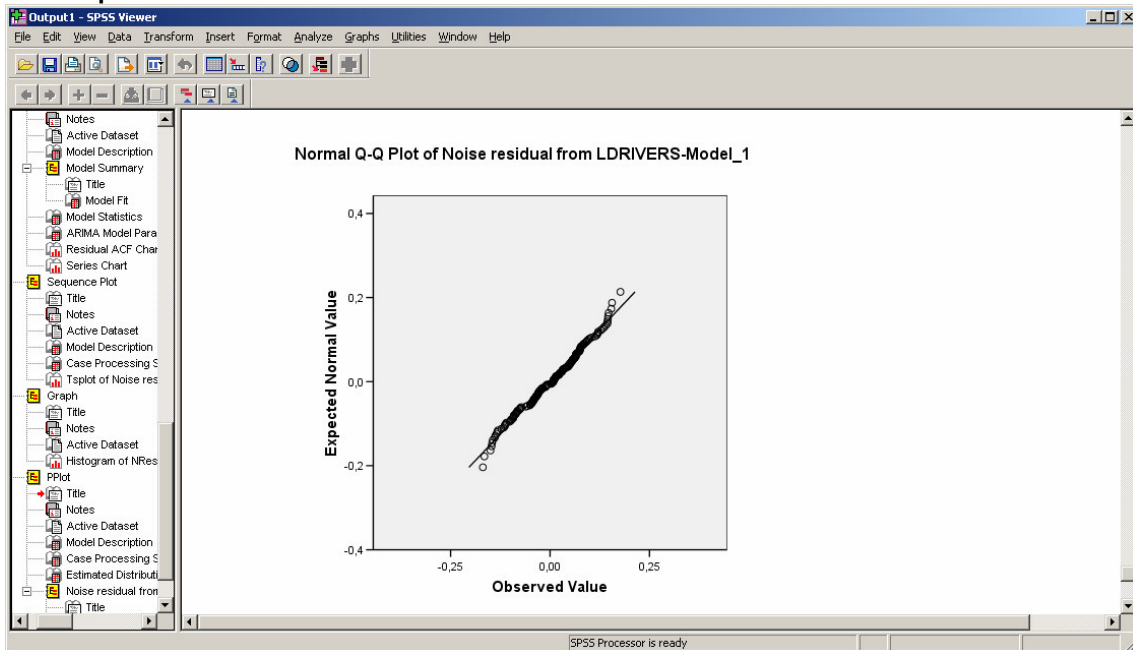
The two usual graphical outputs are still given below, followed by the normality test results.

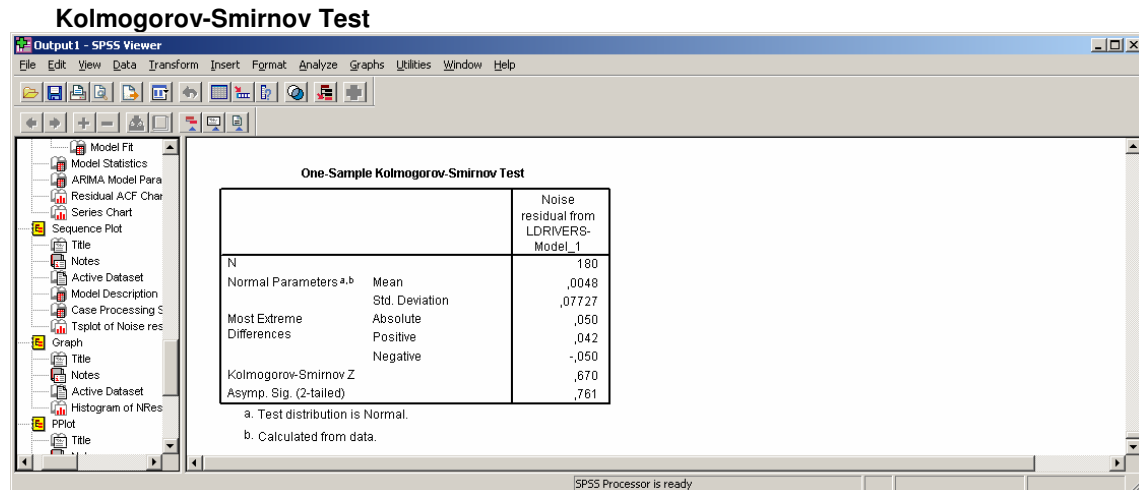


Histogram



QQ-plot





In this case, this hypothesis of normality of the residuals is accepted, because the 0,761 value of the Asymp. Sig. (2-tailed) is more than 0.05 (at the usual 95% confidence level).

Note that the statistic value has kept on increasing, and has the highest value in this very last case.

3.4.5 Conclusion ARMA-type models

In this manual, the general approach for performing an ARMA-type analysis was given and demonstrated on several examples of ARMA-type models.

A pure ARIMA model was estimated on non stationary annual data (section 3.4.3.) and ARIMA models including additional variables (intervention and explanatory) were estimated on non stationary monthly data (sections 3.4.4).

In the case a pure ARIMA model was estimated, for descriptive or forecasting purposes, without any call to additional variables), the first relevant task consists in pretransforming the initial dataset in order to obtain another stationary data set. The second relevant task consists in identifying a parsimonious ARMA model on this second data set, and to test whether the main hypothesis related to the residuals of the model (non correlation, and normality) are valid.

In the case ARIMA models including additional variables were estimated, for descriptive and explanatory purposes, the global model was fitted directly, and all parameters - whether related to the dynamics or to the exogenous effects – estimated altogether.

The main hypotheses related to the residuals were tested in the same manner as in the preceding case.

Apart of the added value due to the exogenous variables - in terms of interpretation of the exogenous estimated parameters - , this modelling in successive stages described in this manual, highlighted a general increase of the model fit, which was observed at each stage (see the Methodology Report for more results about parameters interpretation and gain in the model fit).

3.5 DRAG models

The DRAG model is an important theoretical contribution to road safety analysis and has for the sake of completeness been described in the methodology report. In this manual, however, there is no section dedicated to DRAG-models in this manual, because they require large amounts of data which, in practice, are seldom available in road safety research. This manual is written as part of the SafetyNet project. The databases this project produces are not meant to be exhaustive enough for such a purpose. Moreover, the software retained for performing time series analysis within the project do not allow for estimation of a DRAG-model (note that a software: the TRIO program, is dedicated to estimating DRAG models).

3.6 State space models

Chris de Blois (SWOV)

3.6.1 Introduction

The objective of this manual for state space analysis is 1) to demonstrate the opportunities that state space modelling offers to road safety analysis, 2) to instruct the reader in setting up a state space model, and 3) to instruct the reader how to interpret the model results. The reader does not need to be an expert in statistics, modelling, or programming.

This state space analysis manual is closely related to Section 3.6 of the Methodology report, which deals with the theory behind state space modelling. This manual section demonstrates how the analyses discussed in Section 3.6 of the Methodology report are performed with a software package for state space analysis called STAMP 6.0². Therefore, the datasets used in this manual are the same as the datasets which are considered in Section 3.6. In addition to the theoretical sections, this manual also describes the general approach recommended for state space analysis of time series. This approach is illustrated using the dataset representing the monthly number of drivers killed or seriously injured (KSI) for the years 1969-1984 in the UK, which is one of the datasets employed in the Methodology report.

STAMP 6.0, a software package dedicated to state space modelling, is powerful and easy to use, and is therefore also used for the state space analyses in this manual. STAMP 6.0 has independently been reviewed by several authors: Teyssière (2005), Hallahan (2003), Judge and Ninomiya (2000), and Yaffee (2003).

Section 3.6.2 first describes in detail how to set up a deterministic level model in STAMP 6.0 and how to interpret the results. Then, the stochastic level model is described less extensively and the results of the analysis are compared with the results of the analysis with the deterministic model. Section 3.6.3 deals with the local linear trend model. The deterministic variant of this model, the deterministic level and deterministic slope model, corresponds to a classical linear regression. So, if this model is applied to the same dataset as used in the classical linear regression manual (Section 3.2), then the results should correspond to the results presented in that manual. Section 3.6.4 introduces an additional component: the seasonal. In Section 3.6.5 and Section 3.6.6 another two components are added: intervention variables and explanatory variables, respectively. Sections 3.6.4 through 3.6.6 demonstrate the recommended approach to state space analysis of time series. Finally, Section 3.6.7 contains

²System requirements for STAMP 6.0 are: Windows XP/2000/NT/98/95. STAMP 6.0 is available from Timberlake Consultants Ltd, Ujit 3, Broomsleigh Business Park, Worsley Bridge Road, London SE26 5BN, United Kingdom. Telephone +44 (0)20 86973377, Fax +44 (0)20 86973388. E-mail: info@timberlake.co.uk. Website: www.timberlake.co.uk. The main website for STAMP is: www.STAMP-software.com.

a summary of results and some general recommendations for the analysis of road safety data with state space techniques.

The analysis of each model is subdivided into the following steps:

1. Start of analysis and data load;
2. Model formulation;
3. Model estimation and inspection of results;
4. Graphics of model components;
5. Test of model residuals;
6. Test of auxiliary residuals;
7. Conclusion of analysis;
8. Forecasting (optional);
9. Exercise (optional).

Forecasting is discussed for some of the models. Forecasts can be made only if the model performs well and the residuals satisfy the model assumptions. The additional exercise for the reader is optional as well.

The following conventions concerning notation are used throughout Section 3.6:

The basic explanations are in standard print,

- Instructions are preceded by a bullet point,

The model output is printed in Courier, 10 pnt,

More elaborate explanatory texts, which can be skipped without missing essential information, are printed in italics,

<Menu selections are placed between triangular brackets>.

3.6.2 Local level model

This section first presents an extensive step-by-step description of the analysis of the Norwegian fatalities time series using the deterministic level model in STAMP. The analysis includes trend description, residual testing, and outlier testing. Then, the analysis with the stochastic level model, or *local level model*, is described more succinctly. The latter analysis also includes forecasting over seven years.

3.6.2.1. Deterministic level model

The above mentioned steps will now be taken one-by-one for the deterministic level model.

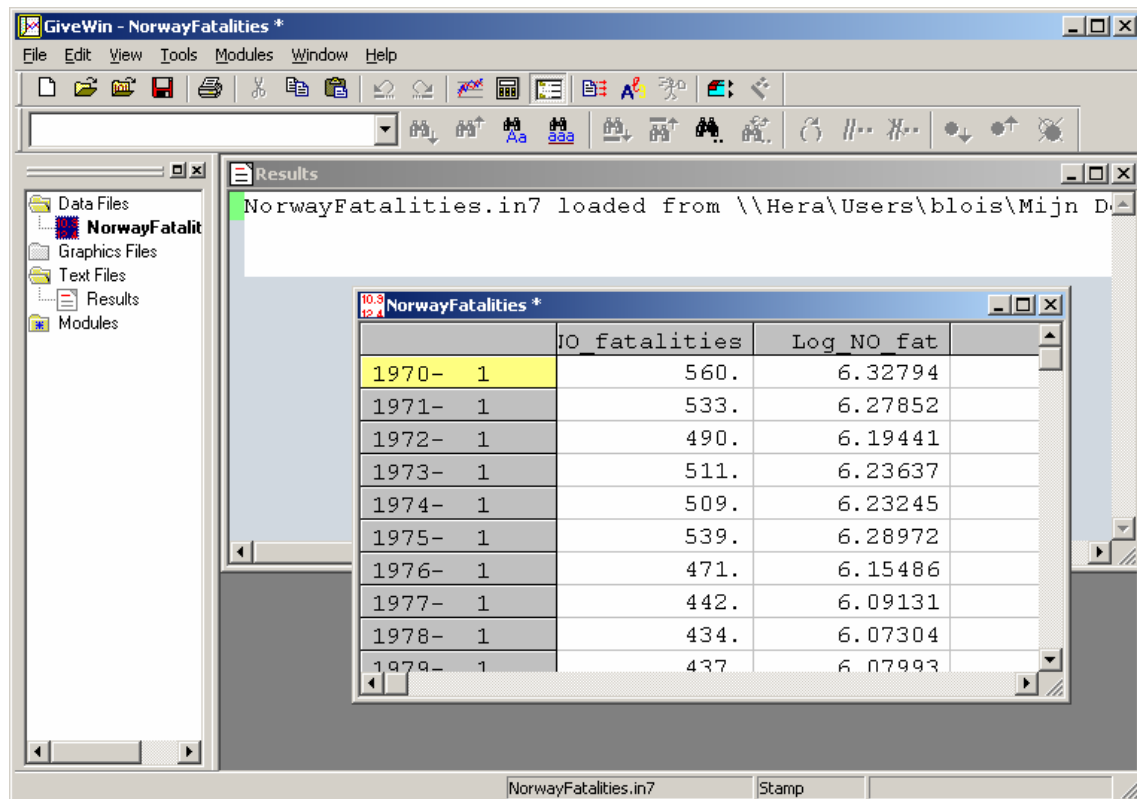
Step 1: Start of analysis and data load

First, we open GiveWin, load the data, and start STAMP.

- Start the GiveWin2 program.
- Use the menu <File, Open Data File...> to open the file "NorwayFatalities.in7".

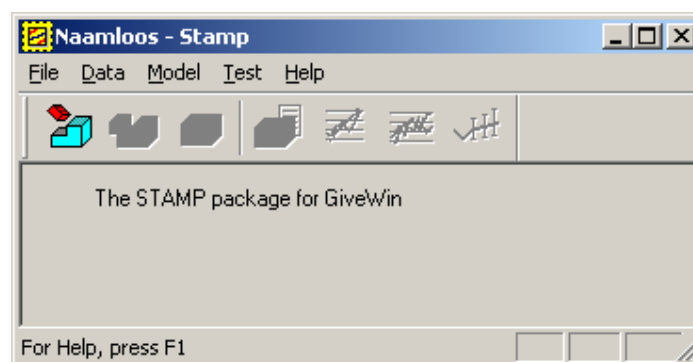
The data file is loaded and displayed in a minimized window at the bottom of the GiveWin main window. To view the data file:

- Click on the icon with the two overlapping boxes:



This data file consists of two variables: the annual number of people killed in road traffic in Norway (see Sections 1.2.2 and 3.6.1 of the Methodology report) for the years 1970 through 2003 (“NO_fatalities”) and the logarithm of the latter time series (“Log_NO_fat”).

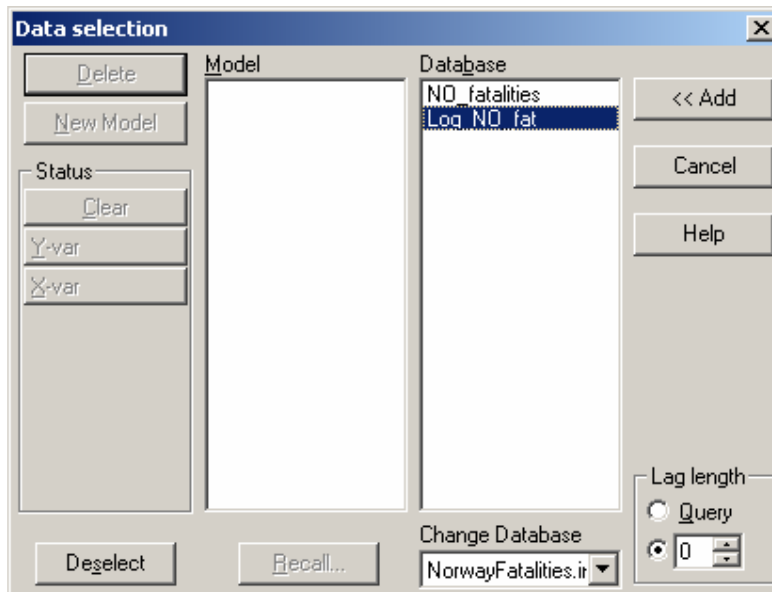
- Minimize the data file window again and use the menu <Modules, Start Stamp> to start the STAMP program. The STAMP window appears:



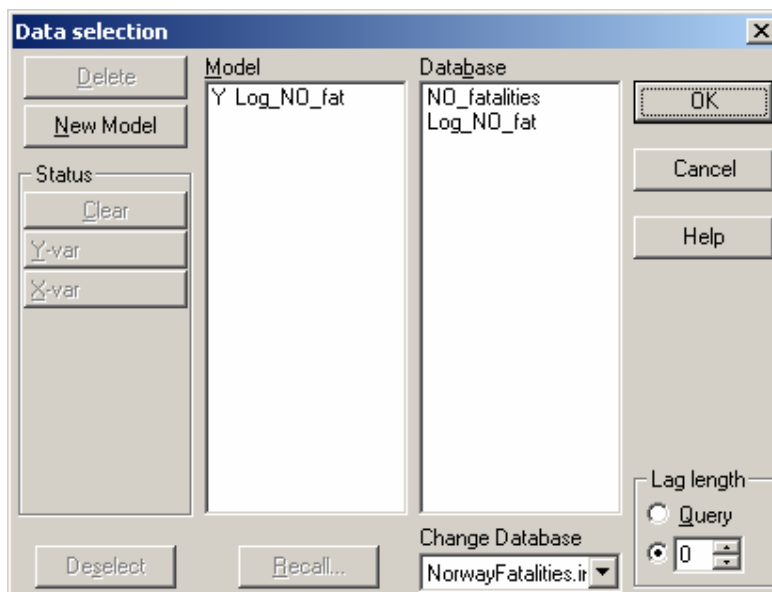
Step 2: Model Formulation

In this step, we define the deterministic level model:

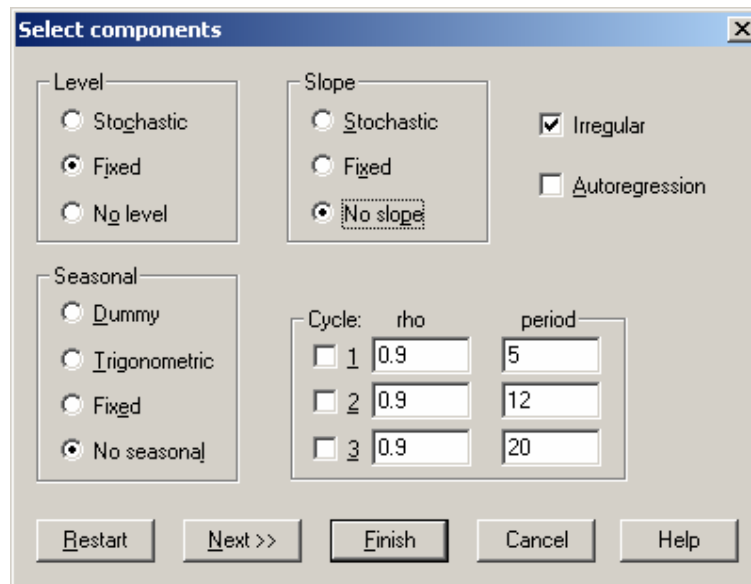
- In the STAMP window choose the menu <Model, Formulate...>.
- In the Data selection window select the variable Log_NO_fat.



- Click the Add button.



- Then click OK.
- In the Select components window, choose a Fixed Level, No slope, Irregular, and No seasonal:

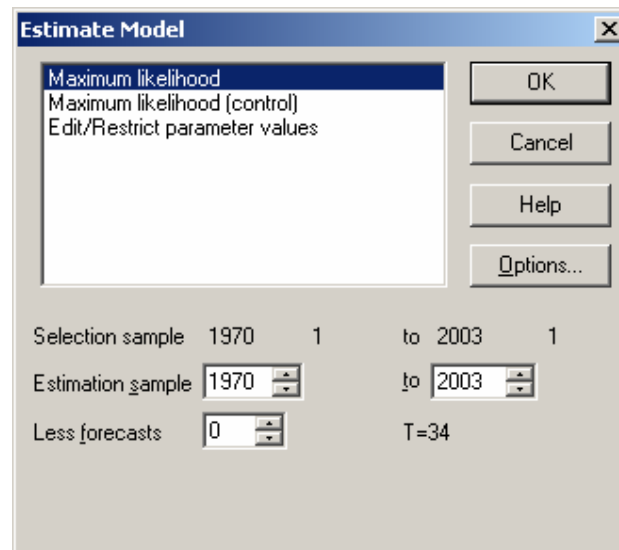


- Then click on the Finish button.

Step 3: Model estimation and inspection of results

The third step is to estimate the model and inspect the results.

- In the Estimate Model window, select Maximum Likelihood:



- Click OK.

The model is estimated, and the following output appears in the GiveWin Results window:

```
---- STAMP 6.30 session started at 13:05:05 on Monday 20 February 2006
----
Please cite STAMP as:
```

Koopman S.J., Harvey, A.C., Doornik, J.A. and Shephard, N. (2000).
 Stamp: Structural Time Series Analyser, Modeller and Predictor,
 London: Timberlake Consultants Press.

Method of estimation is Maximum likelihood
 The present sample is: 1970 to 2003

Equation 1.

Log_NO_fat = Level + Irregular

Estimation report
 Model with 1 parameters (1 restrictions).
 Parameter estimation sample is 1970.1 - 2003.1. (T = 34).
 Log-likelihood kernel is 0.
 No estimation done.

Eq 1 : Diagnostic summary report.

Estimation sample is 1970.1 - 2003.1. (T = 34, n = 33).
 Log-Likelihood is 48.1408 (-2 LogL = -96.2816).
 Prediction error variance is 0.047433

Summary statistics

	Log_NO_fat
Std.Error	0.21779
Normality	1.3457
H(11)	3.6612
r(1)	0.58763
r(6)	-0.073609
DW	0.22639
Q(6, 6)	28.814
R^2	0.00000

Eq 1 : Estimated variances of disturbances.

Component	Log_NO_fat (q-ratio)
Irr	0.048583 (1.0000)

- In the first part of the output (estimation report and above), check the output on the estimation method (maximum likelihood), sample period (1970-2003), model components (level and irregular), the number of parameters estimated (1), and the number of observations (T=34).

The diagnostic summary report gives some additional information: number of degrees of freedom (T-1), log-likelihood, and prediction error variance. The log-likelihood value given is the log-likelihood function at its maximum value after estimation. This value is different from the value in Section 3.6.1.4 of the Methodology report, which is obtained from the above value by extracting a constant and dividing by another constant. Both constants depend on the number of observations T. The prediction error variance (PEV) is a basic measure of goodness-of-fit (the smaller the PEV, the better the fit).

Next, the summary of statistics can be used to evaluate model performance with respect to the diagnostic tests (see Section 3.6.1.4 of the Methodology report). For this evaluation, we make a table like Table 3.6.1. A “+” in the last column of Table 3.6.1 means that the assumption is satisfied, a “-” indicates violation of the assumption.

Statistic	Value	Critical 5% value ^a	Assumption satisfied
-----------	-------	-----------------------------------	-------------------------

Independence	Q(6,6)	28.8	12.59	-
	r(1)	0.588	0.34	-
	r(6)	-0.0736	0.34	+
Homoscedasticity	H(11)	3.66	3.47	-
Normality	N	1.35	5.99	+

Table 3.6.1: Diagnostic test results for the deterministic level model applied to the log of Norwegian fatalities. ^aProbability that statistic exceeds critical value is 0.05.

Comparison of Table 3.6.1 and the corresponding Table 3.6.1 in the Methodology report, shows that STAMP uses other choices with respect to the statistics than in the analysis presented in the Methodology report, i.e. Q(6) instead of Q(10) and r(6) instead of r(4). Below, STAMP's choices are amplified.

Koopman et al. (2000) give more information on the summary statistics. Here, we restrict ourselves to some concise remarks.

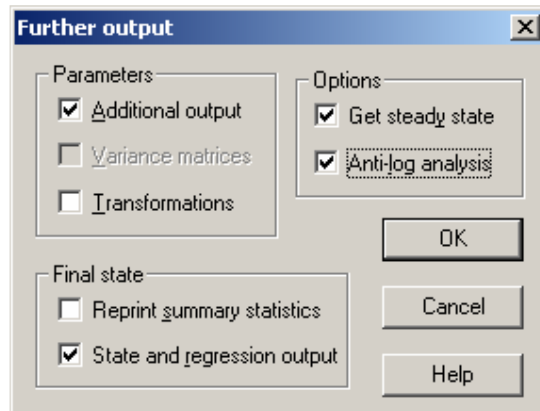
- For testing normality, the Doornik-Hansen statistic is used, which is, under the null hypothesis of normally distributed residuals, approximately $\chi^2(2)$ distributed.
- $H(h)$ is a heteroscedasticity test, which is approximately $F(h,h)$ distributed. In STAMP, "h" is determined as the number of degrees of freedom divided by three and rounded down to the nearest integer.
- $r(\tau)$ is the residual autocorrelation at lag τ , distributed approximately as $N(0, 1/T)$.
- DW is the Durbin-Watson statistic, which tests for residual autocorrelation at lag 1, and is approximately $N(2, 4/T)$ distributed.
- $Q(P,d)$ is the BOX-Ljung Q-statistic based on the first P residual autocorrelations, which is distributed approximately as $\chi^2(d)$, where d is $P-m+1$ with m the number of parameters.
- R^2 is the coefficient of determination, which is a measure of the proportion of observational variance which is explained by the model and as such a measure of goodness-of-fit.

- At the bottom of the GiveWin results window, check whether the estimated variances of the disturbances are sufficiently large.

A near zero variance is an indication of a deterministic component. In this model, the only model component which can vary, the irregular component (i.e., the observation disturbances), has unequal to zero. The level is fixed. In the deterministic level model, the estimated variance of the irregular component is equal to the variance of the series. The variance of the log of the number of Norwegian fatalities is therefore equal to 0.048583. The q-ratio (in the output between brackets) is the ratio of each variance to the largest and is equal to one, because there is only one variance, which therefore is the largest.

Next, we will produce some additional output:

- In the STAMP window choose <Test, Further output...> in the menu.
- Select Additional output, Get steady state, Anti-log analysis, and State and regression output:



- Click OK.

In the GiveWin Results window, the following additional results are displayed:

Eq 1 : Estimated standard deviations of disturbances.

Component	Log_NO_fat (q-ratio)
Irr	0.22042 (1.0000)

Eq 1 : Estimated coefficients of final state vector.

Variable	Coefficient	R.m.s.e.	t-value
Lvl	5.9323	0.037801	156.93 [0.0000]

Anti-log trend analysis
Trend value at end of period is 377.005.

The estimated standard deviation of the irregular is the square root of the estimated variance of the irregular (see above). In the deterministic level model, the estimated standard deviation of the irregular is equal to the standard deviation of the observations in the series.

- Check the values of the estimated coefficients of the final state vector.

The final state vector contains the values of the model components for the last time step of the observed time series. The state only consists of a level component in this case, and the estimate for the value of the level in 2003 equals 5.9323, which is the mean of the log of the Norwegian fatalities series. Moreover, since the level is treated deterministically in this analysis, its estimated value is actually 5.9323 for all $T=34$ time points of the series. The t -statistic is computed as the coefficient (5.9323) divided by its root mean square error (0.0378), and is used to test whether the estimated value of the level significantly deviates from zero.

- Test the significance of the estimated coefficients of the final state vector.

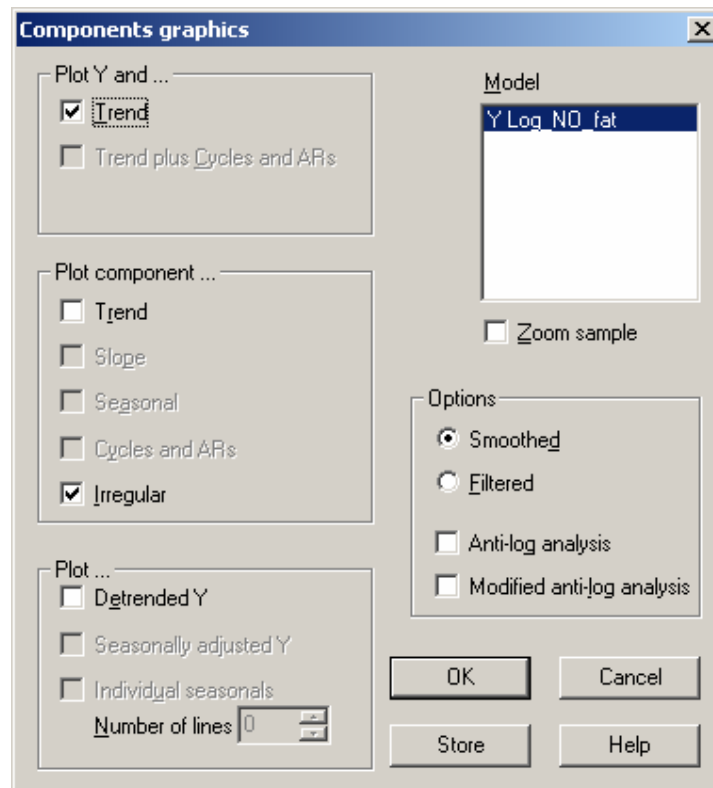
Under the null hypothesis, the t -statistic has a Student's t -distribution with $T-1$ degrees of freedom. Between square brackets behind the t -value, the model output gives the probability that the absolute value of a Student's t -distributed variable X exceeds the actual, absolute value of the t -statistic, i.e. $\text{Prob}(|X| > |t|)$. This probability is very small here, so it may seem that the level significantly deviates from zero. However, since the residuals do not satisfy the assumptions of independence and homoscedasticity (see Table 3.6.1), this t -test is seriously flawed, and one should be careful not to draw any conclusions from this test.

The anti-log trend analysis presents the value of the estimated level for the original series at the end of the series (2003). In the deterministic level model, this value is equal to $\exp(\text{mean of the log-transformed series})$.

Step 4: Graphics of model components

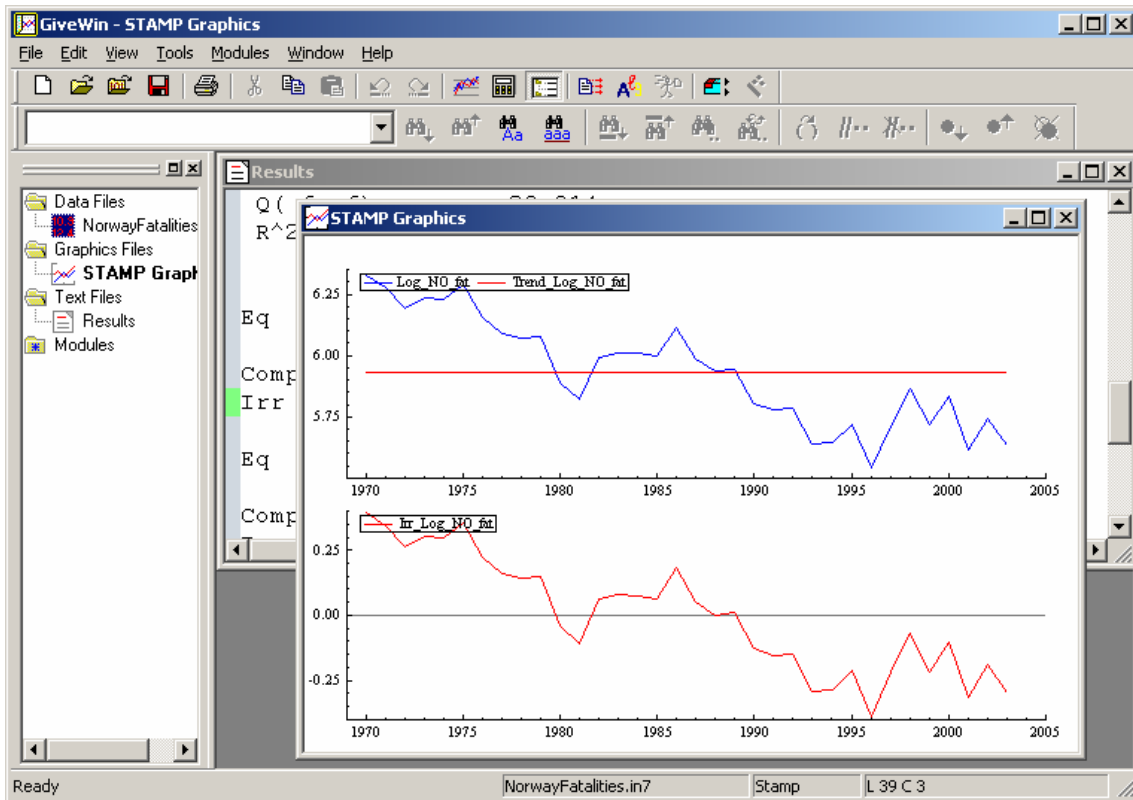
Graphics enlarge the insight in the data and the model results. Therefore, we will generate figures of the observed and log-transformed time series and the estimated trend.

- In the STAMP window choose menu <Test, Components graphics...>. Select Trend, Irregular and Smoothed:



- Click OK.

The STAMP Graphics window appears with graphs of the observed log-transformed time series and the modelled trend and irregular:



- Examine the figures in the STAMP Graphics window and visually inspect the bottom figure of observation disturbances for possible serial correlation.

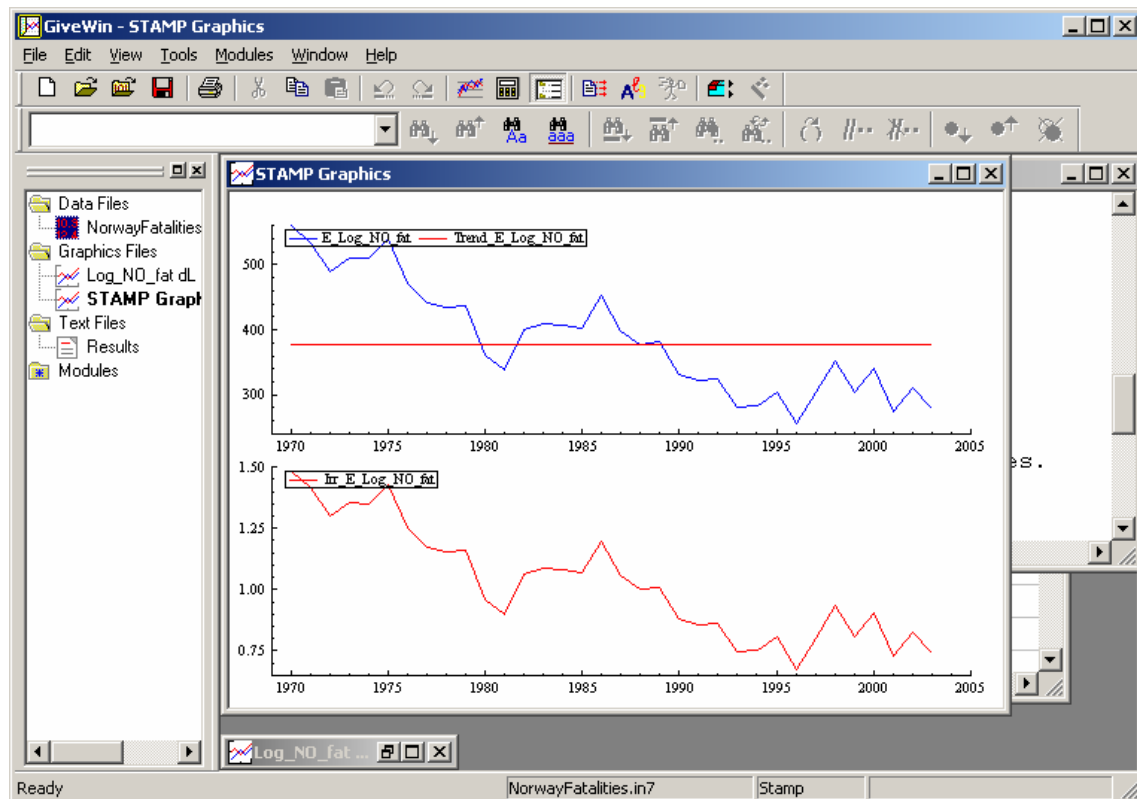
The top figure shows the log-transformed observations and the estimated level, which is just the mean of the series. The bottom figure shows the irregular component or observation disturbances, which, in this case, are equal to the deviations of the observations from their mean value. Visual inspection of the latter figure clearly reveals serial correlation because a positive disturbance tends to be followed by other positive disturbances while a negative disturbance tends to be followed by more negative disturbances. This is confirmed by the more formal tests for independence presented in Table 3.6.1.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps). Minimize the STAMP Graphics window.

Now, we will have a look at the graphs re-expressed in terms of the original data (i.e., in terms of the original number of fatalities instead of their logarithm):

- Go back to the STAMP window and choose <Test, Components graphics...>.
- Select Trend, Irregular, Smoothed, and Anti-log analysis.
- Click OK.

The STAMP Graphics window appears with graphs of the original observed time series and the modelled (anti-logged) level and irregular components:



- Shortly examine the figures in the STAMP Graphics window and check possible serial dependence of the observation disturbances.

The top figure shows the original observations and the estimated level, which is equal to the anti-logged mean of the log-transformed series (which is not exactly the same as the mean of the original series!). The bottom figure shows the irregular component or observation disturbances, which, in this case, are equal to the deviations from the mean value. Notice that the mean of the irregular is around one and not zero, because of the anti-logging.

- Again use the menu <File, Save> or <Ctrl+S> to save these graphs and minimize the STAMP Graphics window.

Step 5: Test of model residuals

STAMP provides the most relevant graphical residual tests as well as more extensive test statistics for normality, goodness-of-fit, and serial correlation.

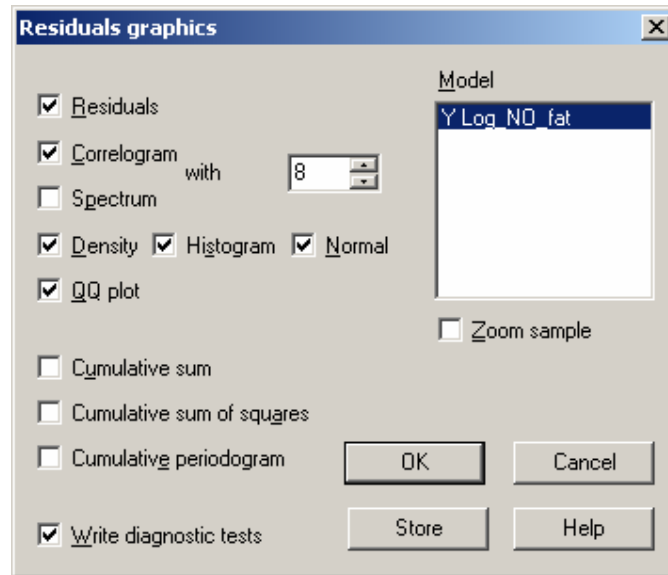
The residuals are not the same as the observation disturbances (i.e., the irregular component). In state space modelling, the estimation process consists of two main steps:

- *filtering, in which only the preceding observations are used and which leads to the “filtered state” and the “one-step ahead predictions”, and*
- *smoothing, in which all observations are used and which leads to the “smoothed state” and the “smoothed predictions”.*

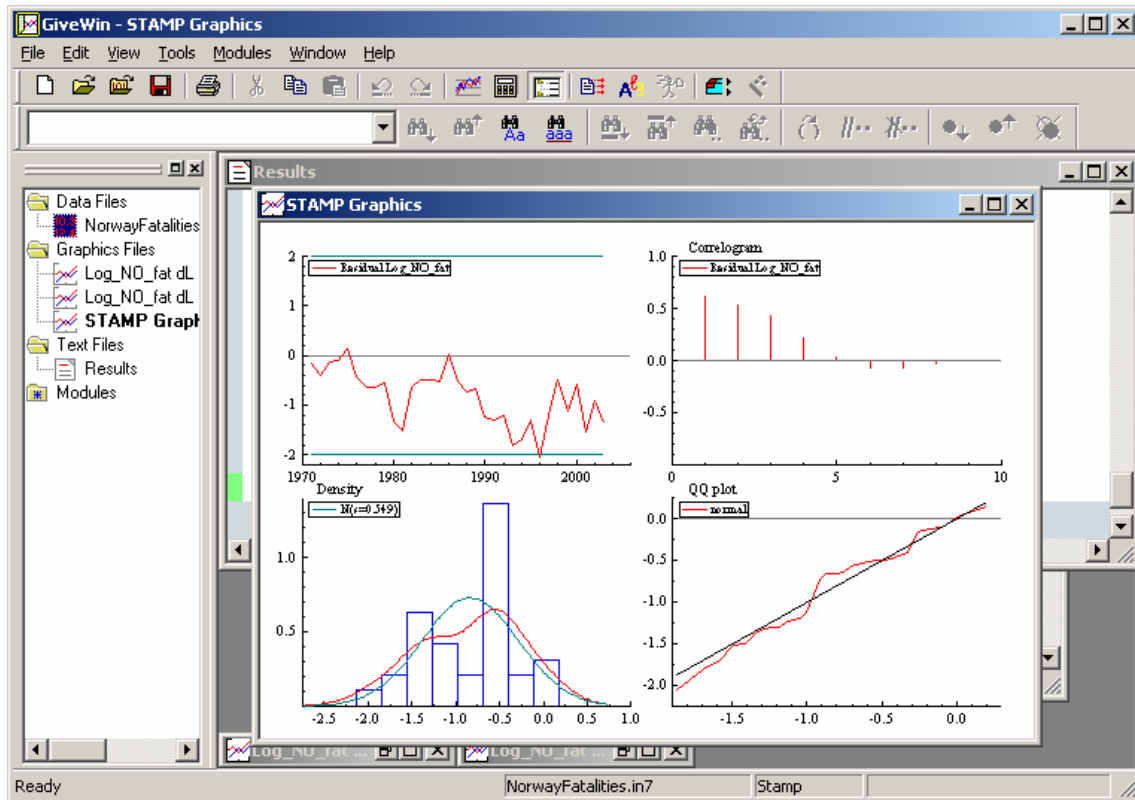
The residuals correspond to the filtered state, the observation disturbances to the smoothed state. In fact, the residuals are the standardized one-step ahead prediction errors, whereas the observation disturbances are the smoothed prediction errors. For more information about the filtered and smoothed state, see Harvey (1989) or Durbin and Koopman (2001).

Tests of the model assumptions are usually applied to the residuals, not to the observation disturbances.

- Go back to the STAMP window and choose <Test, Residuals graphics...>.
- In the Residual graphics window select Residuals, Correlogram, with 8, Density, Histogram, Normal, QQ plot, and Write diagnostic tests:



- Click OK.



The STAMP graphics window in GiveWin shows the standardized residuals and their correlogram, density function, and normal probability plot: see above.

- In the top left figure of the STAMP Graphics window, check how many residuals are outside the 95% confidence interval. The two confidence bounds are indicated by straight lines at the level of about 2 and -2.

Under the assumption of normality, about 95% of the residuals should lie between the two confidence bounds. The figure shows that only one residual is located outside the confidence bounds, indicating that the residuals are acceptable with respect to this test.

- In the top right figure of the STAMP Graphics window, check the correlogram for possible serial dependence of the residuals.

The correlogram presents the residual autocorrelation for lags 1 to 8. Using a 95% confidence level, the autocorrelation should be between $-2/\sqrt{T} = -0.34$ and $2/\sqrt{T} = 0.34$ (see also Table 3.6.1). As we can see, for the first three out of eight lags the autocorrelation is outside this range, indicating serial dependence of the residuals.

- In the bottom left figure of the STAMP Graphics window, compare the estimated density function with the normal density function with the same mean and standard deviation, in order to evaluate the degree of normality of the residuals.

The density diagram shows the distribution of the residuals over discrete intervals in the histogram. The density function is estimated by "a smoothed function of the histogram using a normal or Gaussian kernel" (Koopman et al., 2000). From this density diagram, we can conclude that normality seems to be ok. This conclusion is in agreement with the conclusion on normality based on the diagnostic tests in Table 3.6.1.

- In the bottom right figure of the STAMP Graphics window, check the QQ plot to evaluate the degree of normality of the residuals.

The QQ plot or normal probability plot is created by rank ordering the residuals from small to large and comparing them with the normal probability values corresponding to the cumulative probabilities of $1/(T+1)$, $2/(T+1)$, etc. If the residuals are approximately normally distributed, the plot is an (almost) straight line. From this QQ plot, we can conclude that the residuals are approximately normally distributed, which is in agreement with the conclusions based on the density diagram and the diagnostic tests (see Table 3.6.1).

- Use the menu <File, Save> or <Ctrl+S> to save these graphs and minimize the STAMP Graphics window.

In the Results window of GiveWin the following residual test results (normality, goodness of fit, serial correlation) have been added:

```

Normality test for Residual Log_NO_fat
Sample Size      33
Mean              -0.835567
Std.Devn.         0.549388
Skewness          -0.282549
Excess Kurtosis   -0.763819
Minimum           -2.048279
Maximum           0.148184
Skewness Chi^2(1)  0.43909 [0.5076]
Kurtosis Chi^2(1)  0.8022 [0.3704]
Normal-BS Chi^2(2) 1.2413 [0.5376]
Normal-DH Chi^2(2) 1.3457 [0.5102]
Goodness-of-fit results for Residual Log_NO_fat
Prediction error variance (p.e.v)      0.047433
Prediction error mean deviation (m.d)   0.040164
Ratio p.e.v. / m.d in squares           0.887888
Coefficient of determination              R2      -0.005917
... based on differences                  RD2      -3.292629
Information criterion of Akaike           AIC      -2.989614
... of Schwartz (Bayes)                   BIC      -2.944721

Serial correlation statistics for Residual Log_NO_fat.
Durbin-Watson test is 0.226385.
Asymptotic deviation for correlation is 0.174078.

```

Lag	dF	SerCorr	BoxLjung	ProbChi2(dF)
1	0	0.5876		
2	1	0.5052	21.9744	[0.0000]
3	2	0.3724	27.3134	[0.0000]
4	3	0.1785	28.5818	[0.0000]
5	4	0.0032	28.5822	[0.0000]
6	5	-0.0736	28.8140	[0.0000]
7	6	-0.0734	29.0532	[0.0001]
8	7	-0.0637	29.2405	[0.0001]

- In the Results window of GiveWin, check the results of the normality test. Use the Doornik-Hansen statistic. The probability value [between square brackets] should be larger than 0.05.

The normality test gives the sample size and the mean, standard deviation, skewness, kurtosis, minimum, and maximum of the residuals. The values of the skewness and kurtosis are tested against a $\chi^2(1)$ distribution, whereas the Bowman-Shenton statistic and the Doornik-Hansen statistic are tested against a $\chi^2(2)$ distribution. Koopman et al. (2000) note that the first three tests are only suitable when applied to very large samples. In this case, with a sample size of

34, we better use the Doornik-Hansen statistic only. The corresponding probability value is larger than 0.05 and therefore indicates normality of the residuals. The Doornik-Hansen statistic was also included in the summary of statistics (see step 1) and in the diagnostic test results table (Table 3.6.1).

- Consider the goodness-of-fit results and check
 - whether the ratio between PEV and MD in squares is close to one;
 - whether the coefficient of determination, R^2 , is positive and close to one;
 - the value of the Akaike Information Criterion (AIC).

The goodness-of-fit results include the prediction error variance (PEV), the prediction error mean deviation (MD), and the ratio of their squares computed as $2*PEV^2/(\pi*MD^2)$. The prediction error variance is the variance of the residuals in the steady state, i.e. when the recursive computation procedure, known as the “Kalman Filter” (Harvey, 1989; Durbin and Koopman, 2001), has converged. If the Kalman Filter does not converge, the finite PEV is used. The prediction error mean deviation is the mean deviation of the residuals in the steady state. In a correctly specified model, the ratio between the squared PEV and the squared MD should be close to one (Koopman et al., 2000).

Furthermore, the goodness of fit can be evaluated by means of the coefficient of determination, which is a measure of the extent to which the variance of the observations is explained by the variance of the model predictions. Koopman et al. (2000) give three variants and define their area of application as in Table 3.6.2. They note that the coefficient of determination may become negative, which is an indication of a worse fit than in a simple random level model (or random level and slope model, or random level, slope and seasonal model).

Coefficient of determination	Appropriate to be applied to time series with ...
R^2	no slope, no seasonal (stationary)
R^2_D	slope, no seasonal
R^2_S	slope and seasonal

Table 3.6.2: The coefficients of determination and their application area. Source: (Koopman et al., 2000).

Finally, often used goodness-of-fit variables are the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC). The smaller the AIC (or BIC), the better the model. These variables are computed as $\log(PEV)+c*m/T$, where T is the sample size (34), m the number of parameters estimated (1), and c is 2 in the AIC and $\log(T)$ in the BIC. Note that the AIC is defined differently in Section 3.6.1.4 of the Methodology report, where it is based on the log-likelihood.

- Check on serial correlation by using the Box-Ljung statistic as computed for lags 1 to 8.

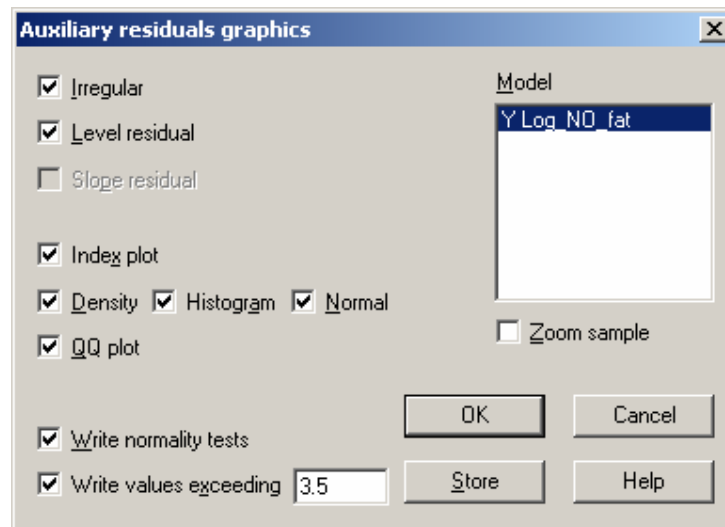
The serial correlation statistics include the Durbin-Watson test, the asymptotic deviation for correlation, and the serial correlation for lags 1 to 8 with the corresponding value of the Box-Ljung statistic and the corresponding probability value. These probability values should be larger than 0.05, which is the case for none of the eight lags considered. The Durbin-Watson and the Box-Ljung statistic were already described above, when dealing with the summary of statistics (see Step 1: Start of analysis and data load).

Not all of these residual test results have to be checked always. We recommend to use the AIC (and/or BIC) to test the goodness-of-fit and the Box-Ljung statistics to test serial independence. The Doornik-Hansen test for normality is already included in the summary of statistics (see Step 1: Start of analysis and data load).

Step 6: Test of auxiliary residuals

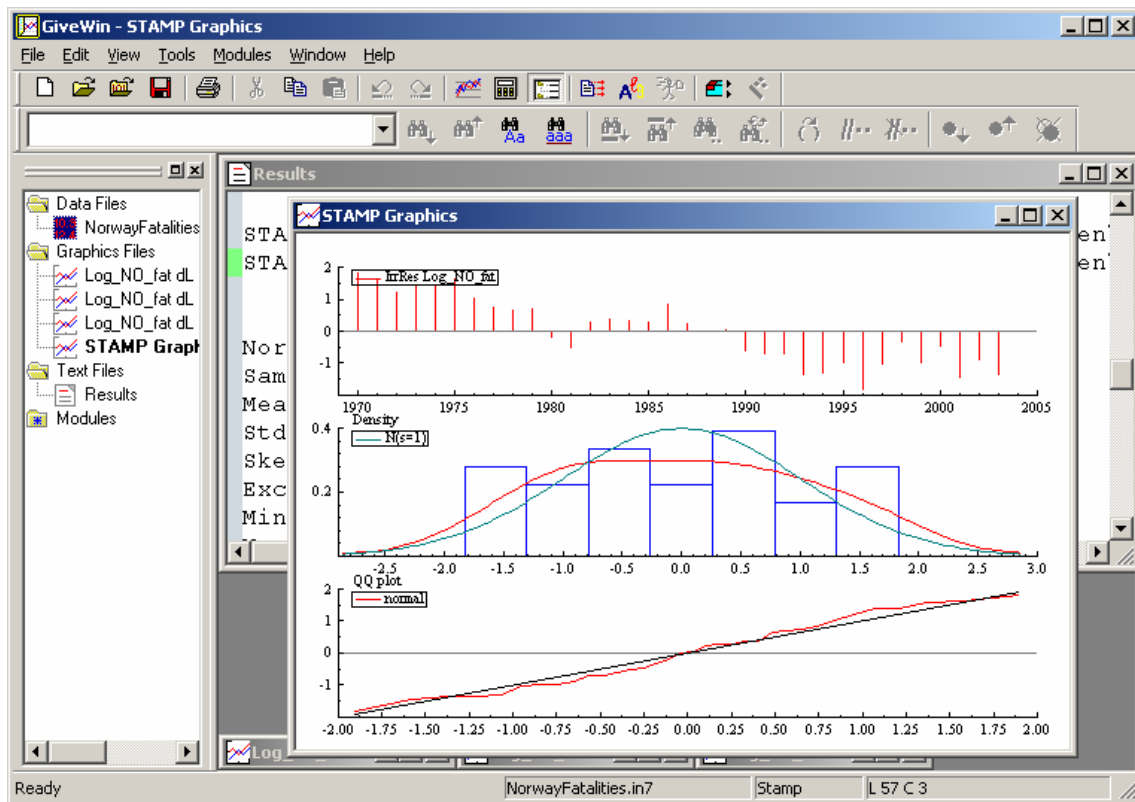
The so-called *auxiliary residuals* are very helpful in finding possible outliers among the observations and structural breaks, e.g. caused by interventions.

- Go to the STAMP window again and choose <Test, Auxiliary residuals graphics...>.
- In the Auxiliary residuals graphics window select Irregular, Level residual, Index plot, Density, Histogram, Normal, QQ plot, Write normality tests, and Write values exceeding (3.5):



- Click OK.

The STAMP graphics window in GiveWin displays the auxiliary residuals of the irregular and their density function and normal probability plot:



The auxiliary residuals are standardised smoothed observation and state disturbances. The auxiliary residuals should be normally distributed with zero mean and unity standard deviation. The auxiliary residuals of the irregular help to detect possible outlier observations, the auxiliary residuals of level, slope, and seasonal help to identify structural breaks in the level, slope, and seasonal component, respectively. For example, if for a certain time point the auxiliary residual of the irregular is larger than 2 or smaller than -2, then this indicates a possible outlier observation. However, one should note that according to a normal distribution 5% of the auxiliary residuals are expected to lie outside the 95% confidence interval of ± 2 .

- Using the figures in the STAMP Graphics window, check whether the auxiliary residuals are approximately normally distributed.

In the top figure, we see that none of the 34 standardised smoothed observation disturbances, i.e. unmistakably less than 5%, is larger than 2 or smaller than -2. The middle and the bottom figure show that the standardised smoothed observation disturbances are approximately normally distributed.

Because the level was assumed deterministic in this model, no standardised smoothed level disturbances were estimated.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs and minimize the STAMP Graphics window.

In the Results window of GiveWin the following auxiliary residual test results (normality, goodness of fit, serial correlation) have been added:

```

Normality test for IrrRes Log_NO_fat
Sample Size      34
Mean              0.000000
Std.Devn.        1.000000
Skewness          0.103696
Excess Kurtosis  -1.063128
Minimum          -1.800584
Maximum          1.822140
Skewness  Chi^2(1)  0.060933  [0.8050]
Kurtosis  Chi^2(1)  1.6012   [0.2057]
Normal-BS  Chi^2(2)  1.6621   [0.4356]
Normal-DH  Chi^2(2)  2.003    [0.3673]

```

- Check the result of the Doornik-Hansen test for normality.

The normality test for residuals were already described above. From the Doornik-Hansen test, we can conclude that the hypothesis of normally distributed auxiliary residuals is accepted; the probability value between square brackets is larger than 0.05.

Note that in the Results window of GiveWin no values exceeding 3.5 have been added.

Under the null hypothesis of normality, the probability of an auxiliary residual whose absolute value exceeds 3.5 is very small, about 0.0005. So, when this happens this is a very strong indication of an outlier.

Step 7: Conclusion of analysis

The residuals obtained with the analysis of the log of the annual Norwegian fatalities from 1970 to 2003 with the deterministic level model do not satisfy the important model assumptions of independence and homoscedasticity (see Table 3.6.1). It is therefore not the appropriate model for describing this series. We still discussed all the output that can be obtained from STAMP in this case, so as to make the reader familiar with the possible options, and for reasons of later reference.

Step 8: Forecasting

Because the deterministic level model is clearly not appropriate, it does not make much sense to compute forecasts with this model.

3.6.2.2 Stochastic level model

The stochastic level model, which is also known as *local level model*, will be described more briefly than the deterministic level model, since much of the output from STAMP has already been discussed and explained in Section 3.6.2.1. The focus will be on the new aspects and the comparison of the results of this model with the results of the deterministic model.

Step 1: Start of analysis and data load

If you just fitted the deterministic level model, GiveWin and STAMP have already been started and data is still loaded in the GiveWin window. If you start here or if you have closed the database, STAMP, or GiveWin after the previous exercise, please follow the instructions under step 1 of Section 3.6.2.1.

Step 2: Model Formulation

The stochastic (or: local) level model can be fitted in STAMP as follows:

- Choose the menu <Model, Formulate...> in the STAMP window.
- If needed, select the variable Log_NO_fat in the Data selection window and click the Add button.
- Then click OK.
- In the Select components window, choose a Stochastic Level, No slope, Irregular, and No seasonal, as follows:

The 'Select components' dialog box is shown with the following settings:

- Level:** ☒ Stochastic, ☐ Fixed, ☐ No level
- Slope:** ☐ Stochastic, ☐ Fixed, ☒ No slope
- Seasonal:** ☐ Dummy, ☐ Irigonometric, ☐ Fixed, ☒ No seasonal
- Irregular:** ☒ Irregular, ☐ Autoregression
- Cycle:**

Cycle	rho	period
<input type="checkbox"/> 1	0.9	5
<input type="checkbox"/> 2	0.9	12
<input type="checkbox"/> 3	0.9	20

Buttons at the bottom: Restart, Next >>, Finish, Cancel, Help.

- Then click on the Finish button.

Step 3: Model estimation and inspection of results

- In the Estimate Model window, select Maximum Likelihood.
- Click OK.

The model is estimated, and the following output appears in the GiveWin Results window:

Method of estimation is Maximum likelihood
The present sample is: 1970 to 2003

```
MaxLik initialising...
it 1 f= 2.09171 e0= 0.52237 step= 1.00000
it 2 f= 2.22405 e0= 0.00514 step= 1.00000

MaxLik iterating...
it 2 f= 2.22407 df= 0.00000 e1= 0.00000 e2= 0.00000 step=
0.00000
```

Equation 2.

Log_NO_fat = Level + Irregular

Estimation report
Model with 2 parameters (1 restrictions).
Parameter estimation sample is 1970. 1 - 2003. 1. (T = 34).
Log-likelihood kernel is 2.224067.
Very strong convergence in 2 iterations.
(likelihood cvg 0
gradient cvg 2.435829e-007
parameter cvg 9.737567e-012)

Eq 2 : Diagnostic summary report.

Estimation sample is 1970. 1 - 2003. 1. (T = 34, n = 33).
Log-Likelihood is 75.6183 (-2 LogL = -151.237).
Prediction error variance is 0.00989161

Summary statistics

	Log_NO_fat
Std.Error	0.099457
Normality	1.2746
H(11)	1.7464
r(1)	-0.12735
r(7)	-0.15301
DW	2.0513
Q(7, 6)	5.4955
R^2	0.79023

Eq 2 : Estimated variances of disturbances.

Component	Log_NO_fat (q-ratio)
Irr	0.0032682 (0.6949)
Lvl	0.0047030 (1.0000)

- Check the results (sample period, log-likelihood, estimated variances of disturbances).

The output first reports about the estimation process, which is subdivided into initialisation and further maximisation of the log-likelihood kernel f (see Koopman et al., 2000). In the deterministic level case, no iterations were needed. Now, however, two parameters have to be estimated: the observation disturbance variance and the level disturbance variance. Convergence is reached in two iterations, and is very strong. This implies that all of the

convergence criteria used in the iterative process for parameter estimation are satisfied. These convergence criteria are likelihood, gradient, and parameter convergence (cvg).

At convergence, the value of the log-likelihood function is 75.6 which is larger than in the deterministic level case (48.1). The prediction error variance (0.00989) is clearly smaller than for the deterministic level model (0.0474). These results indicate that the stochastic model yields a better fit than the deterministic model, albeit at the expense of having estimated one extra parameter (the variance of the level disturbances).

- The STAMP output results concerning the summary statistics can again be condensed into the following table (see also Table 3.6.2 in the Methodology report):

	Statistic	Value	Critical 5% value ^a	Assumption satisfied
Independence	Q(7,6)	5.50	12.59	+
	r(1)	-0.127	0.34	+
	r(7)	-0.153	0.34	+
Homoscedasticity	H(11)	1.75	3.47	+
Normality	N	1.27	5.99	+

Table 3.6.3: Diagnostic test results for the stochastic level model applied to the log of Norwegian fatalities. ^aProbability that statistic exceeds critical value is 0.05.

When we compare the results in Table 3.6.3 with those in Table 3.6.1, we see that also with respect to the diagnostic tests the stochastic level model performs better than the deterministic level model, because the present model satisfies all assumptions.

Finally, the output gives the estimated disturbance variances. The irregular disturbance variance is about 30% smaller than the level disturbance variance.

- In the STAMP window choose <Test, Further output...> in the menu.
- Select Additional output, Get steady state, Anti-log analysis, and State and regression output.
- Click OK.

The GiveWin Results window will display the following additional results:

Eq 3 : Estimated standard deviations of disturbances.

Component	Log_NO_fat (q-ratio)
Irr	0.057168 (0.8336)
Lvl	0.068578 (1.0000)

Eq 3 : Estimated coefficients of final state vector.

Variable	Coefficient	R.m.s.e.	t-value
Lvl	5.6627	0.047118	120.18 [0.0000]

Anti-log trend analysis
Trend value at end of period is 287.92.

The value of the level at the end of the period as presented by the anti-log trend analysis (288) is considerably smaller than the corresponding value in the deterministic model (377).

Step 4: Graphics of model components

- In the STAMP window choose menu <Test, Components graphics...>. Select Trend, Irregular and Smoothed.
- Click OK.

The STAMP Graphics window appears with graphs of the observed log-transformed time series and the modelled trend and irregular.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps).

An eps. file can be loaded into a Word document. In the remainder of this manual we will do so instead of adding a screen print as in Section 3.6.2.1. Figure 3.6.1 shows the observed log-transformed time series and the stochastic level component (top), and the irregular component (bottom).

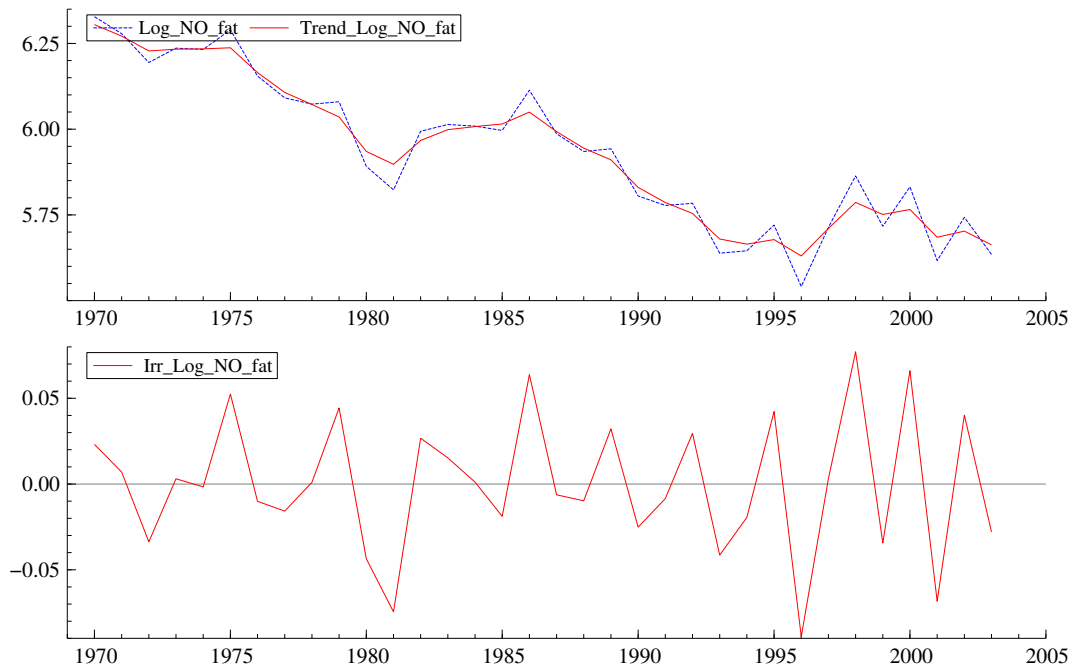


Figure 3.6.1: Observed log-transformed time series and the local level and irregular components for the log of Norwegian fatalities.

In the top part of Figure 3.6.1, we see that the estimated trend is close to the (log-transformed) observations. Furthermore, we can see that the irregular shows a much more random pattern than in the deterministic case.

- Minimize the STAMP Graphics window.

Step 5: Test of model residuals

- Go back to the STAMP window and choose <Test, Residuals graphics...>.
- In the Residual graphics window select Residuals, Correlogram, with 8, Density, Histogram, Normal, QQ plot, and Write diagnostic tests.
- Click OK.
- Use the menu <File, Save> or <Ctrl+S> to save these graphs.

Figure 3.6.2 shows the standardized residuals and their correlogram, density function, and normal probability plot as depicted by the STAMP graphics window in GiveWin.

The top left graph of Figure 3.6.2 illustrates that none of the 34 residuals is outside the 95% confidence interval, which is very good. From the top right graph, we learn that for none of the eight lags considered the autocorrelation is outside the 95% confidence interval, which is defined by the boundaries $-2/\sqrt{T} = -0.34$ and $+2/\sqrt{T} = 0.34$. This is an indication of absence of serial dependence. The bottom graphs show that the assumption of normality of the residuals is better satisfied than in the deterministic level model.

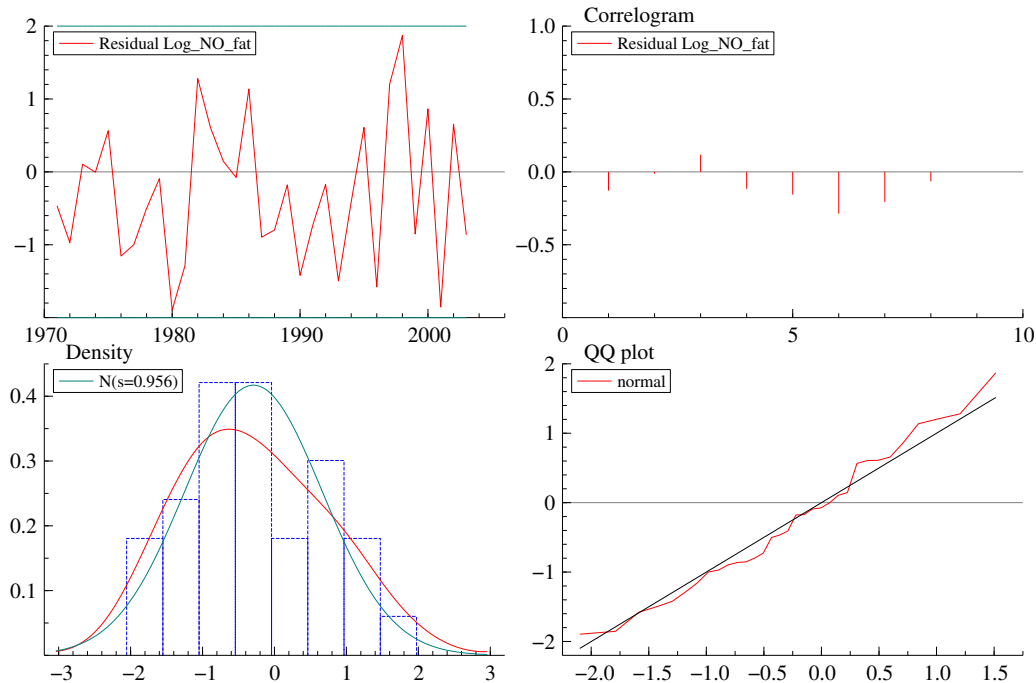


Figure 3.6.2: Residuals and residual tests for the stochastic level model applied to the log of Norwegian fatalities.

In the Results window of GiveWin, the following residual test results have been added (only part of the results is printed):

```

Goodness-of-fit results for Residual Log_NO_fat
Information criterion of Akaike      AIC      -4.498421
... of Schwartz (Bayes)            BIC      -4.408635

Serial correlation statistics for Residual Log_NO_fat.
Lag  dF      SerCorr    BoxLjung    ProbChi2(dF)
1     0      -0.1273
2     0      -0.0124
3     1       0.1095      1.0526      [ 0.3049]
4     2      -0.1054      1.4951      [ 0.4735]
5     3      -0.1382      2.2833      [ 0.5157]
6     4      -0.2253      4.4556      [ 0.3478]
7     5      -0.1530      5.4955      [ 0.3584]
8     6      -0.0478      5.6010      [ 0.4693]

```

The goodness-of-fit results are undoubtedly better than in the deterministic case: in the stochastic model, the AIC is smaller (-4.50 instead of -3.00), just as the BIC (-4.41 instead of -2.94). For all lags considered, the Box-Ljung test indicates that the most important assumption of independence is satisfied.

Step 6: Test of auxiliary residuals

- Go to the STAMP window again and choose <Test, Auxiliary residuals graphics...>.
- In the Auxiliary residuals graphics window select Irregular, Level residual, Index plot, Density, Histogram, Normal, QQ plot, Write normality tests, and Write values exceeding (3.5).
- Click OK.
- Use the menu <File, Save> or <Ctrl+S> to save these graphs.

The STAMP graphics window in GiveWin displays the auxiliary residuals of the irregular and of the level component and their density function and normal probability plot: see Figure 3.6.3.

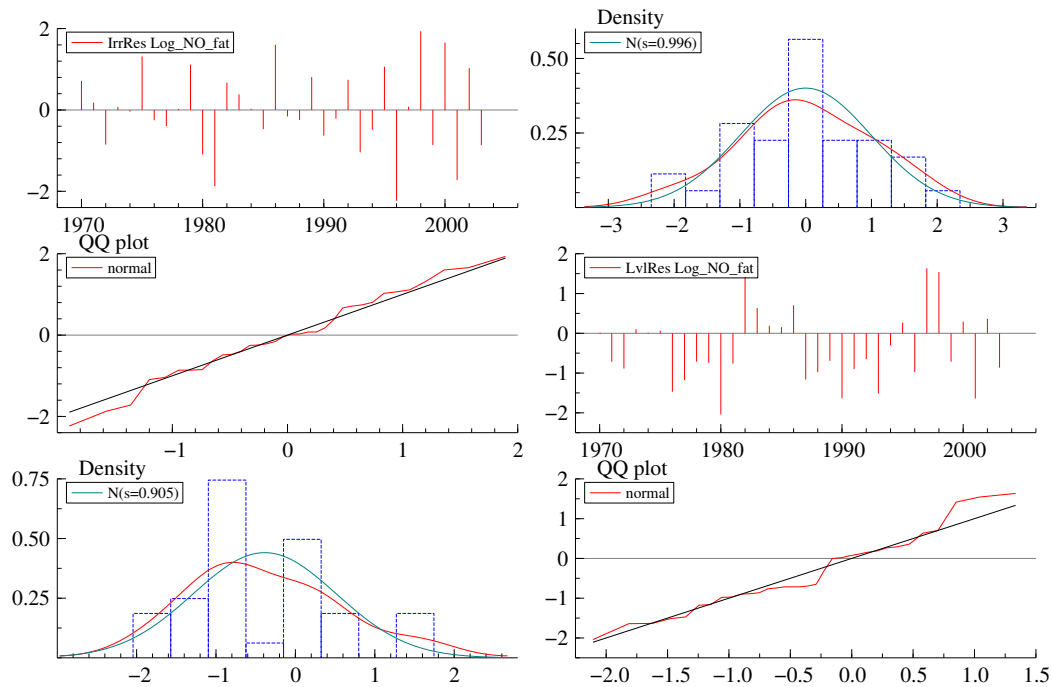


Figure 3.6.3: Auxiliary residuals and corresponding tests for the stochastic level model applied to the log of Norwegian fatalities.

The following output describes the auxiliary residual test results (normality, goodness of fit, serial correlation) as can be found in the Results window of GiveWin.

```

Normality test for IrrRes Log_NO_fat
Sample Size      34
Mean              -0.000312
Std.Devn.         0.995882
Skewness          -0.122988
Excess Kurtosis   -0.405802
Minimum          -2.234331
Maximum           1.935788
Skewness Chi^2(1)  0.085714 [0.7697]
Kurtosis Chi^2(1)  0.23329 [0.6291]
Normal-BS Chi^2(2)  0.319 [0.8526]
Normal-DH Chi^2(2)  0.12802 [0.9380]

Normality test for LvlRes Log_NO_fat
Sample Size      34
Mean              -0.386541
Std.Devn.         0.904762
Skewness          0.485177
Excess Kurtosis   -0.311715
Minimum          -2.040092
Maximum           1.633678
Skewness Chi^2(1)  1.3339 [0.2481]
Kurtosis Chi^2(1)  0.13765 [0.7106]
Normal-BS Chi^2(2)  1.4716 [0.4791]
Normal-DH Chi^2(2)  1.9093 [0.3849]

```

Both Figure 3.6.3 and the auxiliary residual tests demonstrate that the auxiliary residuals of both the irregular and the level component satisfy the assumption of normality.

Note that, just as in the deterministic case, in the Results window of GiveWin no values exceeding 3.5 have been added.

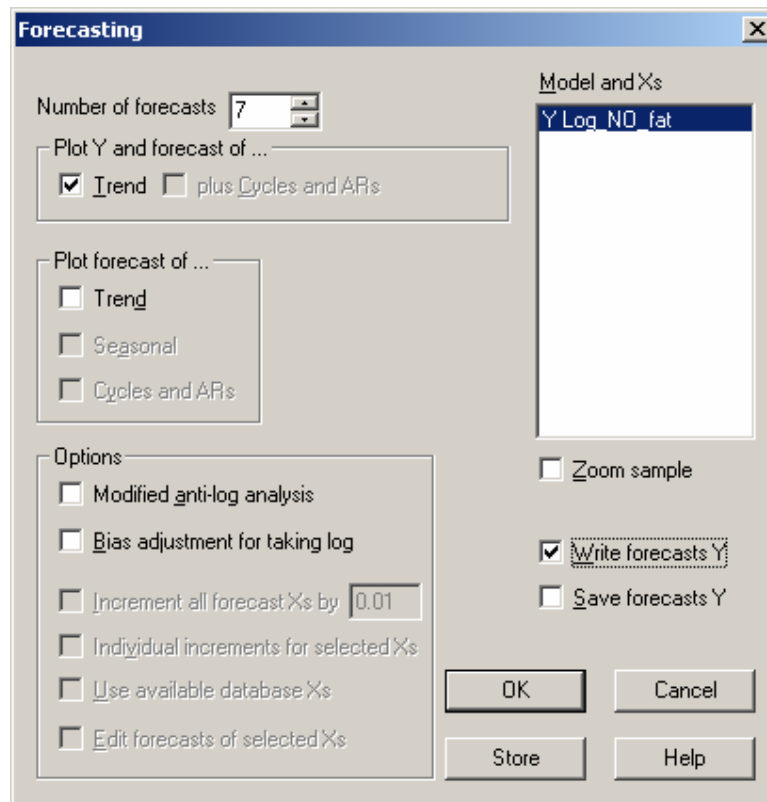
Step 7: Conclusion of analysis

The residuals obtained with the analysis of the log of the annual Norwegian fatalities from 1970 to 2003 with the local level model satisfy all the model assumptions of independence, homoscedasticity, and normality. It seems therefore to be the appropriate model for describing this series.

Step 8: Forecasting

Since the local level model provides an appropriate description of the log of the Norwegian fatalities series, as a final step in the analysis we will compute seven-year forecasts for this series. Furthermore, by performing an anti-log analysis the forecasts will be re-expressed in terms of the original count data.

- Go to the STAMP window again and choose <Test, Forecasting...>.
- In the Forecasting window select 7 as the number of forecasts, Trend, and Write forecasts Y:



- Click OK.

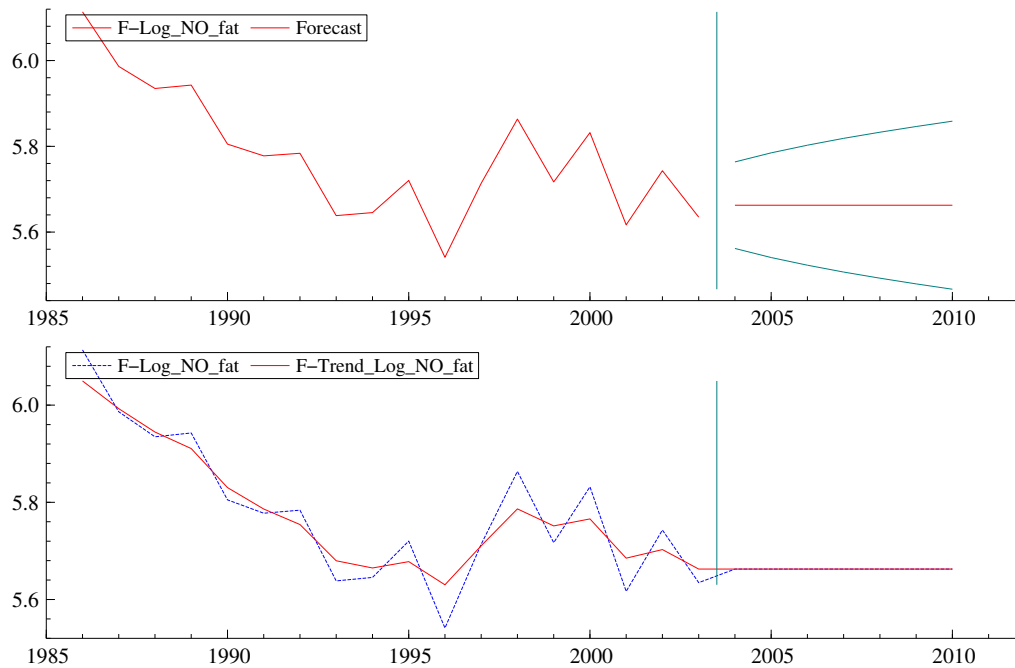


Figure 3.6.4: Seven-years forecasts (2004-2010) of the stochastic level model applied to the log of annual Norwegian fatalities, 1970-2003.

The STAMP graphics window in GiveWin displays the log-transformed observations extended with the seven-years forecasts with 70% confidence interval (plus and minus one estimated standard deviation) in the top figure and the log-transformed observations and the extrapolated trend in the bottom figure: see Figure 3.6.4.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps).

*The bottom figure clearly illustrates that the local level model always yields forecasts that are equal to the last value of the level component in the series. This is in complete agreement with the fact that we are dealing with a local **level** model.*

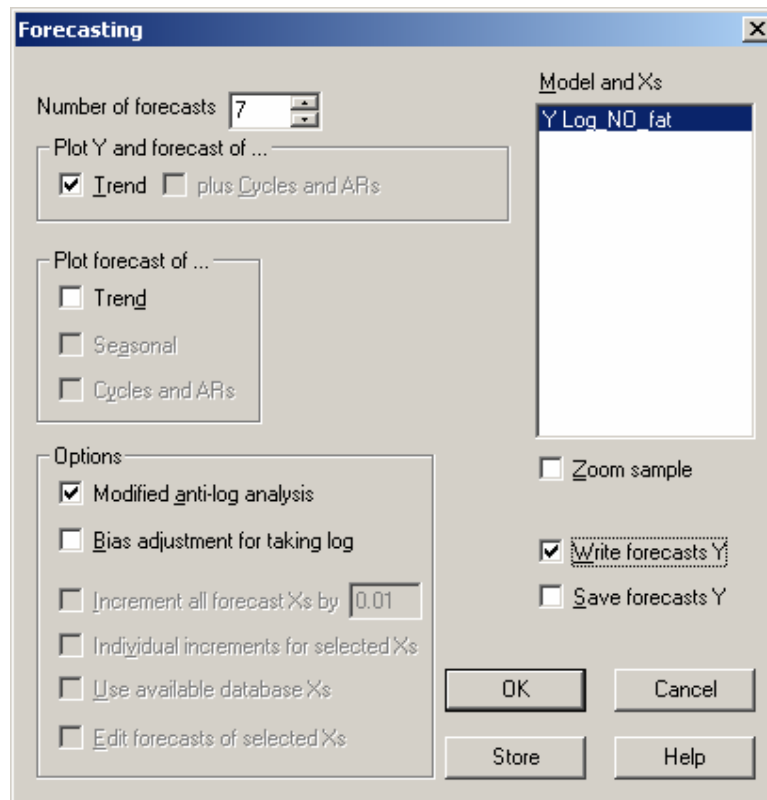
In the Results window of GiveWin the forecasts for the log-transformed time series have been added:

Eq 2 : Forecasts for F-Log_NO_fat.

Period	Forecast	R.m.s.e.	- Rmse	+ Rmse
2004. 1	5.6627	0.10095	5.5617	5.7636
2005. 1	5.6627	0.12204	5.5406	5.7847
2006. 1	5.6627	0.13999	5.5227	5.8027
2007. 1	5.6627	0.15589	5.5068	5.8186
2008. 1	5.6627	0.17030	5.4924	5.8330
2009. 1	5.6627	0.18359	5.4791	5.8463
2010. 1	5.6627	0.19598	5.4667	5.8587

The list of forecast results gives for each time point the forecast, its standard error, and the lower and upper bound of the 70% confidence interval.

- Go to the STAMP window again and choose <Test, Forecasting...>.
- In the Forecasting window select 7 as the number of forecasts, Trend, Modified anti-log analysis, and Write forecasts Y:



- Click OK.

The STAMP graphics window in GiveWin displays the original observations extended with the seven-years forecasts with 70% confidence interval (plus and minus one estimated standard deviation) in the top figure and the original observations and the extrapolated trend in the bottom figure: see Figure 3.6.5.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs.

In the Results window of GiveWin the forecasts for the original observed time series have been added:

Eq 2 : Forecasts for E_F-Log_NO_fat.
Anti-log

Period	Forecast	R.m.s.e.	- Rmse	+ Rmse
2004. 1	287.92	30.584	257.34	318.50
2005. 1	287.92	37.373	250.55	325.29
2006. 1	287.92	43.264	244.66	331.18
2007. 1	287.92	48.570	239.35	336.49
2008. 1	287.92	53.457	234.46	341.38
2009. 1	287.92	58.023	229.90	345.94
2010. 1	287.92	62.336	225.58	350.26

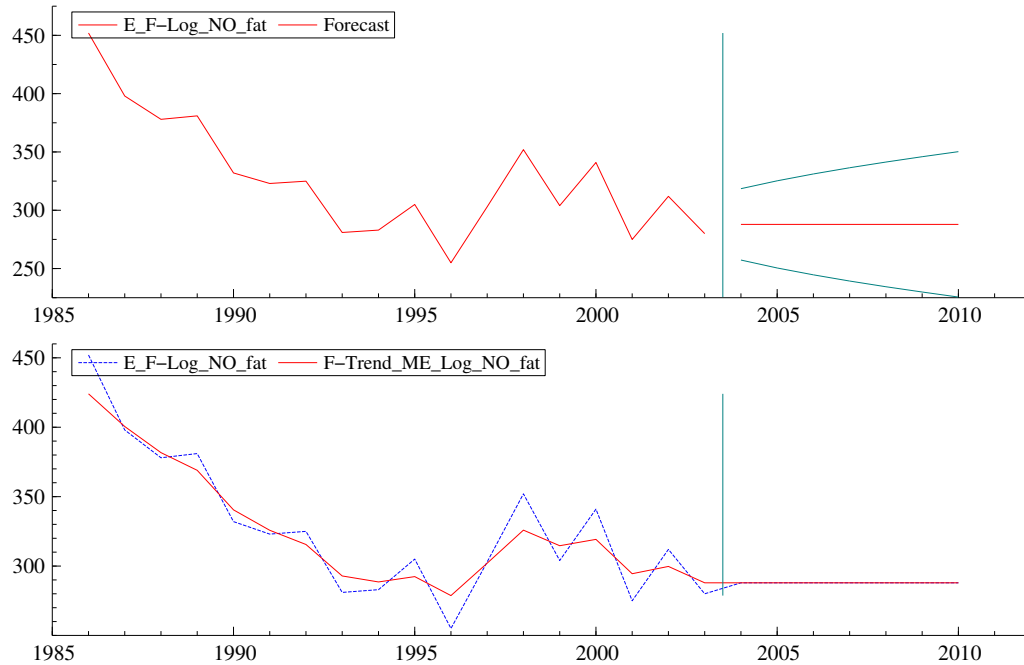


Figure 3.6.5: Anti-logged seven-year forecasts (2004-2010) of the stochastic level model applied to the log of annual Norwegian fatalities, 1970-2003.

3.6.3 Local linear trend model

In this section, the slope component will be added to the local level model, so as to obtain the *local linear trend model*. The model will be applied to the number of fatalities as observed in Finland for the period 1970 through 2003. The theory on this model and the results of its application to the Finnish data are described in Section 3.6.2 of the Methodology report. This section explains how the model is built in STAMP.

First, this section presents a step-by-step description of the analysis of the Finnish fatalities time series using a linear trend model with deterministic level and deterministic slope, also called the *deterministic linear trend model*. As in Section 3.6.2 of the Methodology report, we will show that this model, which is equivalent to a classical linear regression model, does not satisfy important model assumptions. Then, we describe the analysis with the linear trend model with stochastic level and stochastic slope, which is also known as the *stochastic linear trend model* or *local linear trend model*. The latter analysis also includes forecasting over seven years.

3.6.3.1. Deterministic linear trend model

Step 1: Start of analysis and data load

First, we open GiveWin, load the data, and start STAMP.

- If GiveWin is not yet open, then start GiveWin2.
- If GiveWin is still open from the previous analysis, then close all results, data, and graphics windows in GiveWin by clicking on the cross in the top right corner of each window.
- Use the menu <File, Open Data File...> to open the file "FinlandFatalities.in7".

The data file is loaded and displayed in a minimized window at the bottom of the GiveWin main window. To view the data file:

- Click on the icon with the two overlapping boxes.

The data file consists of two variables: the annual number of people killed in road traffic in Finland for the years 1970 through 2003 ("FI_fatalities") and the logarithm of the latter time series ("Log_FI_fat").

- Minimize the data file window again and use the menu <Modules, Start Stamp> to start the STAMP program.

Step 2: Model Formulation

In this step, we define the deterministic linear trend model:

- In STAMP, choose the menu <Model, Formulate>.

- In the Data selection window select the variable Log_FI_fat and click Add.
- Then click OK.
- In the Select components window, choose a Fixed level, Fixed slope, Irregular, and No seasonal:

- Then click on the Finish button.

Step 3: Model estimation and inspection of results

- In the Estimate Model window, select Maximum Likelihood.
- Click OK.

The model is estimated, and the following output appears in the GiveWin Results window:

Eq 1 : Diagnostic summary report.

Estimation sample is 1970. 1 - 2003. 1. (T = 34, n = 32).
Log-Likelihood is 55.7297 (-2 LogL = -111.459).
Prediction error variance is 0.0205839

Summary statistics

	Log_FI_fat
Std.Error	0.14347
Normality	3.5468
H(10)	0.63283
r(1)	0.76735
r(6)	-0.080113
DW	0.38632
Q(6, 6)	49.093
Rd^2	-1.2238

Eq 1 : Estimated variances of disturbances.

Component Log_FI_fat (q-ratio)

Irr 0.021360 (1.0000)

- Check the results (sample period, log-likelihood, estimated variance of disturbances).
- The output results concerning the summary statistics can again be condensed into the following table (see also Table 3.6.3 in the Methodology report):

	Statistic	Value	Critical 5% value ^a	Assumption satisfied
Independence	Q(6,6)	49.1	12.59	-
	r(1)	0.767	0.34	-
	r(6)	-0.0801	0.34	+
Homoscedasticity	H(10)	0.633	3.72	+
Normality	N	3.55	5.99	+

Table 3.6.4: Diagnostic test results for the deterministic linear trend model applied to the log of Finnish fatalities. ^aProbability that statistic exceeds critical value is 0.05.

Table 3.6.3 in the Methodology report gives the reciprocal value of the homoscedasticity test, because both the original value and its reciprocal should be smaller than the critical 5% value and, in this case, the reciprocal is larger than the original value. To stay close to the STAMP results, Table 3.6.4 does not present the reciprocal value of the homoscedasticity test statistic. Since the reciprocal of $H(10)$ equals $1/H(10) = 1/0.633 = 1.580$, and because this value is still smaller than the critical value of 3.72, the assumption of homoscedasticity is satisfied. However, the most important assumption of independence is clearly not satisfied in this analysis.

- In the STAMP window choose <Test, Further output...> in the menu.
- Select Additional output, Get steady state, Anti-log analysis, and State and regression output.
- Click OK.

In the GiveWin Results window, the following additional results are displayed:

Eq 1 : Estimated standard deviations of disturbances.

Component Log_FI_fat (q-ratio)
Irr 0.14615 (1.0000)

Eq 1 : Estimated coefficients of final state vector.

Variable	Coefficient	R.m.s.e.	t-value	
Lvl	5.9235	0.049044	120.78	[0.0000]
Slp	-0.028733	0.0025548	-11.246	[0.0000]

Anti-log trend analysis

Trend value at end of period is 373.731.

Growth rate at end of period is -0.0287328 (-2.87328 % per "year").

From the estimated coefficients of the final state vector and their t-values, we can conclude that, in the final state, both the level and the negative slope component are significantly different from zero. In classical linear regression terms, we would say that both the intercept and the regression coefficient significantly deviate from zero. However, these tests are flawed because the residuals do not satisfy the assumption of independence (see Table 3.6.4).

From the anti-log trend analysis, we can see that the estimated trend value at the end of the period is 374, with a reduction rate of almost 3% per year.

Step 4: Graphics of model components

- In the STAMP window choose menu <Test, Components graphics...>. Select (Plot Y and ...) Trend, Slope, Irregular and Smoothed.
- Click OK.

The STAMP Graphics window appears with graphs of the observed log-transformed time series and the modelled trend, slope, and irregular (see Figure 3.6.6).

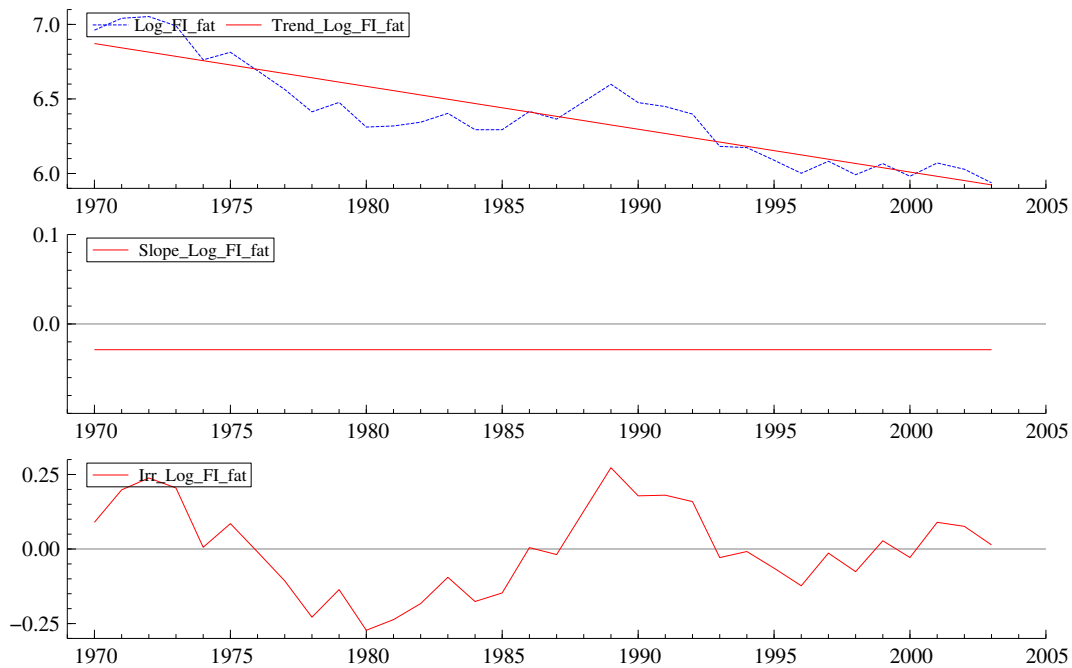


Figure 3.6.6: Observed log-transformed time series and the deterministic linear trend (top graph), slope (middle graph), and irregular component (bottom graph) for the log of Finnish fatalities.

In the top part of Figure 3.6.6, we see that the estimated trend is a straight line; it is in fact a classical linear regression line. The middle graph displays the constant, negative slope. The bottom graph clearly indicates serial dependence in the observation disturbances.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps). Minimize the STAMP Graphics window.

Step 5: Test of model residuals

- Go back to the STAMP window and choose <Test, Residuals graphics...>.

- In the Residual graphics window select Residuals, Correlogram, with 8, Density, Histogram, Normal, QQ plot, and Write diagnostic tests.
- Click OK.
- Use the menu <File, Save> or <Ctrl+S> to save these graphs and minimize the STAMP Graphics window.

Figure 3.6.7 shows the standardized residuals and their correlogram, density function, and normal probability plot as depicted by the STAMP graphics window in GiveWin.

The top left graph of Figure 3.6.7 shows that 1 out of the 34 residuals, i.e. 3%, lies outside the 95% confidence interval; this is acceptable since we would expect about 1 out of 20 to fall outside this range. From the top right graph, we learn that for the first 3 of the 8 lags considered the autocorrelation is outside the 95% confidence interval, which is defined by the boundaries $-2\sqrt{T} = -0.34$ and $+2\sqrt{T} = 0.34$. The bottom graphs show that the assumption of normality of the residuals is quite well satisfied, which confirms the normality test results in Table 3.6.4.

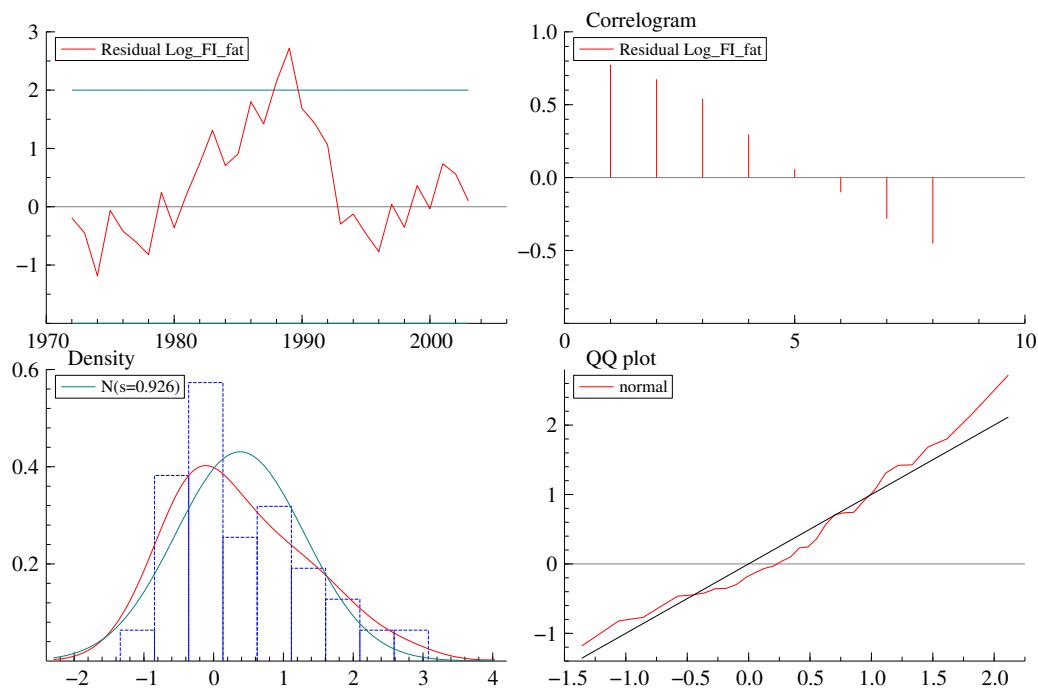


Figure 3.6.7: Residuals and residual tests for the deterministic linear trend model applied to the log of Finnish fatalities.

In the Results window of GiveWin, the following residual test results can be found (only part of the results are printed below):

Goodness-of-fit results for Residual Log_FI_fat
 Information criterion of Akaike AIC -3.765598
 ... of Schwartz (Bayes) BIC -3.675812

Serial correlation statistics for Residual Log_FI_fat.

Lag	dF	SerCorr	BoxLjung	ProbChi2 (dF)
1	0	0.7673		
2	1	0.6601	36.4701	[0.0000]
3	2	0.5001	45.8532	[0.0000]
4	3	0.2712	48.7118	[0.0000]
5	4	0.0528	48.8243	[0.0000]
6	5	-0.0801	49.0928	[0.0000]
7	6	-0.2189	51.1785	[0.0000]
8	7	-0.3337	56.2274	[0.0000]

For all lags considered, the Box-Ljung test indicates that the most important assumption of independence is not satisfied, as we already noted on the basis of Figure 3.6.7 and Table 3.6.4.

Step 6: Test of auxiliary residuals

Because the residuals obtained with the deterministic linear trend model analysis does not satisfy the important model assumption of independence, we skipped this analysis step.

Step 7: Conclusion of analysis

The residuals obtained with the analysis of the log of the annual Finnish fatalities from 1970 to 2003 with the deterministic linear trend model do not satisfy the model assumption of independence. Therefore, it is not the appropriate model for describing this time series. This also means that a classical linear regression model (without explanatory variables) would not be able to appropriately describe this series.

Step 8: Forecasting

Because the deterministic linear trend model is not appropriate, it does not make much sense to compute forecasts with this model.

Step 9: Exercise

Apply a deterministic linear trend model to the dataset which was used in the part of the manual dedicated to classical linear regression (Section 3.2.1). Use the annual averages and compare the results with the corresponding results in Section 3.2.1.

3.6.3.2. Stochastic linear trend model

Step 1: Start of analysis and data load

If you just fitted the deterministic linear trend model, GiveWin and STAMP have already been started and data is still loaded in the GiveWin window. If you start here or if you have closed the database, STAMP, or GiveWin after the previous exercise, please follow the instructions under step 1 of Section 3.6.3.1.

Step 2: Model Formulation

The stochastic (or: local) linear trend model can be fitted in STAMP as follows:

- Choose the menu <Model, Formulate...>.
- If needed, select the variable Log_FI_fat in the Data selection window and click the Add button.
- Then click OK.
- In the Select components window, choose a Stochastic Level, Stochastic slope, Irregular, and No seasonal:

The 'Select components' dialog box is shown with the following settings:

- Level:** ☒ Stochastic, ☐ Fixed, ☐ No level
- Slope:** ☒ Stochastic, ☐ Fixed, ☐ No slope
- Seasonal:** ☐ Dummy, ☐ Trigonometric, ☐ Fixed, ☒ No seasonal
- Irregular:** ☒ Irregular, ☐ Autoregression
- Cycle:**

Cycle	rho	period
<input type="checkbox"/> 1	0.9	5
<input type="checkbox"/> 2	0.9	12
<input type="checkbox"/> 3	0.9	20

Buttons at the bottom: Restart, Next >>, Finish, Cancel, Help.

- Then click on the Finish button.

Step 3: Model estimation and inspection of results

- In the Estimate Model window, select Maximum Likelihood.
- Click OK.

The model is estimated, and the following output appears in the GiveWin Results window:

Equation 1.

Log_FI_fat = Trend + Irregular

Estimation report

Model with 3 parameters (2 restrictions).

Parameter estimation sample is 1970. 1 - 2003. 1. (T = 34).

Log-likelihood kernel is 2.121946.

Very strong convergence in 12 iterations.

(likelihood cvg 3.160187e-014
gradient cvg 2.331024e-007
parameter cvg 1.878562e-008)

Eq 1 : Diagnostic summary report.

Estimation sample is 1970. 1 - 2003. 1. (T = 34, n = 32).

Log-Likelihood is 72.1462 (-2 LogL = -144.292).

Prediction error variance is 0.0100779

Summary statistics

	Log_FI_fat
Std.Error	0.10039
Normality	0.39376
H(10)	0.50989
r(1)	-0.028431
r(8)	-0.15533
DW	2.0137
Q(8, 6)	5.8640
Rd^2	-0.088796

Eq 1 : Estimated variances of disturbances.

Component	Log_FI_fat (q-ratio)
Irr	0.0032008 (1.0000)
Lvl	0.00000 (0.0000)
Slp	0.0015332 (0.4790)

From the estimated variances of disturbances, we see that the variance corresponding to the level component (Lvl) is (almost) equal to zero, meaning that this component does vary over time, and that we may as well treat it deterministically. Therefore, we repeat the analysis (steps 2 and 3) with a deterministic instead of a stochastic level component (select Fixed level in the STAMP Select components window). This yields the following output in the GiveWin results window:

Equation 2.

Log_FI_fat = Trend + Irregular

Estimation report

Model with 2 parameters (1 restrictions).

Parameter estimation sample is 1970. 1 - 2003. 1. (T = 34).

Log-likelihood kernel is 2.121946.

Very strong convergence in 2 iterations.

(likelihood cvg 1.21594e-013
gradient cvg 1.811884e-008
parameter cvg 4.042691e-008)

Eq 2 : Diagnostic summary report.

Estimation sample is 1970. 1 - 2003. 1. (T = 34, n = 32).
 Log-Likelihood is 72.1462 (-2 LogL = -144.292).
 Prediction error variance is 0.0100779

Summary statistics

	Log_FI_fat
Std.Error	0.10039
Normality	0.39376
H(10)	0.50989
r(1)	-0.028427
r(7)	-0.059707
DW	2.0137
Q(7, 6)	4.7703
Rd^2	-0.088796

Eq 2 : Estimated variances of disturbances.

Component	Log_FI_fat (q-ratio)
Irr	0.0032008 (1.0000)
Slp	0.0015331 (0.4790)

- Check the results (sample period, log-likelihood, estimated variance of disturbances).

The estimation report tells that there was very strong convergence in 2 iterations. The estimated variances of both the irregular and the slope component are unequal to zero. The diagnostic summary report shows that the value of the log-likelihood function is 72.1, which is larger than in the deterministic linear trend case (55.7). The prediction error variance (0.0101) is clearly smaller than for the deterministic level model (0.0206). These results indicate that the deterministic level and stochastic slope model, also known as the "smooth trend model", is performing better than the fully deterministic model.

- The STAMP output results concerning the summary statistics can again be condensed into the following table (see also Table 3.6.4 in Section 3.6.2 of the Methodology report):

	Statistic	Value	Critical 5% value ^a	Assumption satisfied
Independence	Q(7,6)	4.77	12.59	+
	r(1)	-0.0284	0.34	+
	r(7)	-0.0597	0.34	+
Homoscedasticity	H(10)	0.510	3.72	+
Normality	N	0.394	5.99	+

Table 3.6.5: Diagnostic test results for the deterministic level and stochastic slope model applied to the log of Finnish fatalities. ^aProbability that statistic exceeds critical value is 0.05.

When we compare the results in Table 3.6.5 with those in Table 3.6.4 of the manual, we see that also with respect to the diagnostic tests the deterministic level and stochastic slope model is better than the deterministic linear trend model. The stochastic model satisfies all model assumptions, whereas the deterministic model did not satisfy the most important assumption, i.e. the assumption of independence.

- In the STAMP window choose <Test, Further output...> in the menu.
- Select Additional output, Get steady state, Anti-log analysis, and State and regression output.
- Click OK.

The GiveWin Results window will display the following additional results:

Eq 2 : Estimated standard deviations of disturbances.

Component	Log_FI_fat (q-ratio)
Irr	0.056576 (1.0000)
Slp	0.039155 (0.6921)

Eq 2 : Estimated coefficients of final state vector.

Variable	Coefficient	R.m.s.e.	t-value	
Lvl	5.9689	0.047371	126	[0.0000]
Slp	-0.035603	0.053297	-0.66802	[0.5089]

Anti-log trend analysis

Trend value at end of period is 391.056.

Growth rate at end of period is -0.0356035 (-3.56035 % per "year").

From the estimated coefficients of the final state vector and their t-values, we can conclude that, in the final state, the level component is significantly different from zero whereas the negative slope component is not.

The trend value at the end of the period as presented by the anti-log trend analysis (391) is larger than the corresponding value in the deterministic model (374).

Step 4: Graphics of model components

- In the STAMP window choose menu <Test, Components graphics...>. Select Trend, Slope, Irregular and Smoothed.
- Click OK.

The STAMP Graphics window appears with graphs of the observed log-transformed time series and the modelled trend, slope, and irregular (see Figure 3.6.8).

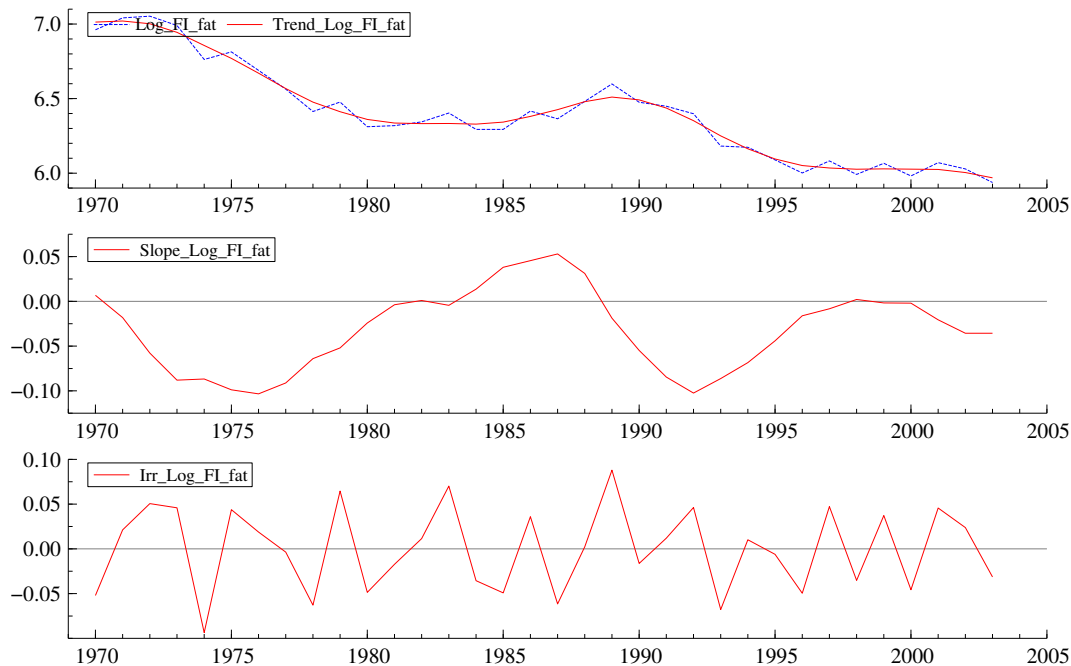


Figure 3.6.8: Observed log-transformed time series and the trend of the deterministic level stochastic slope model (top graph), slope component (middle graph), and irregular component (bottom graph) for the log of Finnish fatalities.

In the middle graph of Figure 3.6.8, we see that a negative slope component corresponds to a decreasing trend, whereas a positive slope component corresponds to an increasing trend. The slope is negative in 1971-1980, (almost) zero in 1981-1983, positive in 1984-1988, negative in 1989-1996, and (almost) negative or slightly negative in 1997-2003. Note that the negative slope component in the final state was found to be insignificantly different from zero (see step 3).

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps). Minimize the STAMP Graphics window.

Step 5: Test of model residuals

- Go back to the STAMP window and choose <Test, Residuals graphics...>.
- In the Residual graphics window, select Residuals, Correlogram, with 8, Density, Histogram, Normal, QQ plot, and Write diagnostic tests.
- Click OK.

Figure 3.6.9 shows the standardized residuals and their correlogram, density function, and normal probability plot as depicted by the STAMP graphics window in GiveWin.

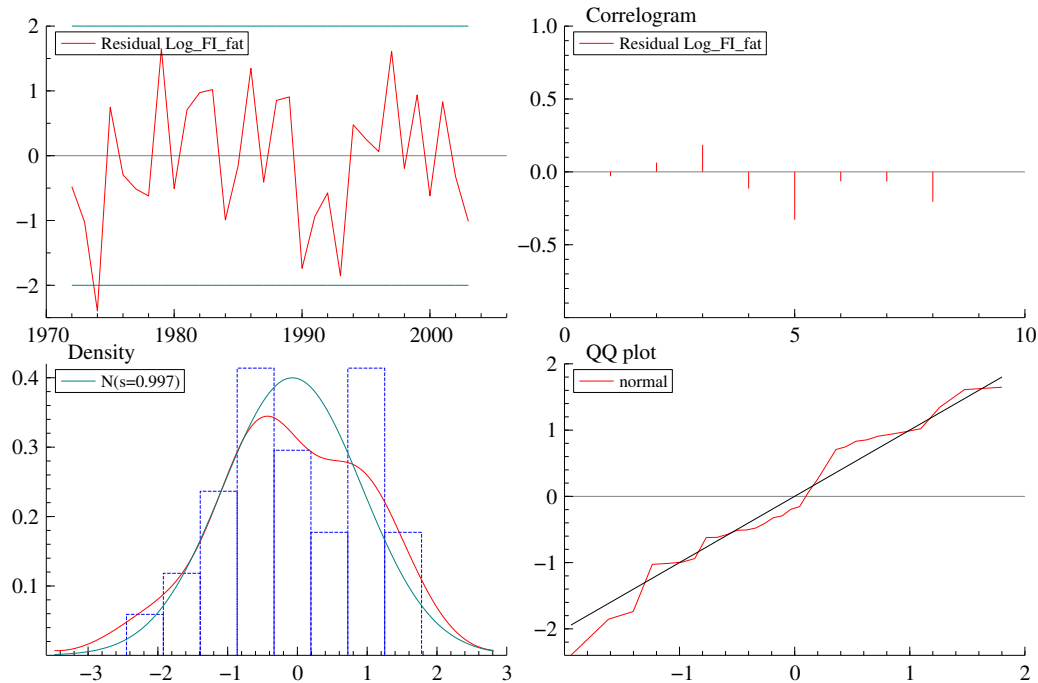


Figure 3.6.9: Residuals and residual tests for the deterministic level stochastic slope model applied to the log of Finnish fatalities.

The top left graph of Figure 3.6.9 shows that 1 out of the 34 residuals, i.e. 3%, lies outside the 95% confidence interval; this is acceptable since we would expect about 1 out of 20 to fall outside this range. From the top right graph, we learn that for none of the 38 lags considered the autocorrelation is outside the 95% confidence interval, which is defined by the boundaries $-2/\sqrt{T} = -0.34$ and $+2/\sqrt{T} = 0.34$. The bottom graphs show that the assumption of normality of the residuals is quite well satisfied, which confirms the normality test results in Table 3.6.5.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs and minimize the STAMP Graphics window.

In the Results window of GiveWin, the following residual test results can be found (only part of the results are printed below):

```
Goodness-of-fit results for Residual Log_FI_fat
Information criterion of Akaike      AIC      -4.420941
... of Schwartz (Bayes)            BIC      -4.286262
```

```
Serial correlation statistics for Residual Log_FI_fat.
```

Lag	dF	SerCorr	BoxLjung	ProbChi2 (dF)
1	0	-0.0284		
2	0	0.0631		
3	1	0.1610	1.1450	[0.2846]
4	2	-0.0937	1.4865	[0.4756]
5	3	-0.2734	4.4980	[0.2125]
6	4	-0.0529	4.6151	[0.3291]
7	5	-0.0597	4.7703	[0.4446]
8	6	-0.1553	5.8640	[0.4386]

The goodness-of-fit is clearly better than in the fully deterministic case: the AIC is smaller (-4.42 instead of -3.77) as well as the BIC (-4.29 instead of -3.68).

For all lags considered, the Box-Ljung test indicates that the most important model assumption, i.e. independence, is satisfied, as we already noted on the basis of Figure 3.6.9 and in Table 3.6.5.

Step 6: Test of auxiliary residuals

- Go to the STAMP window again and choose <Test, Auxiliary residuals graphics...>.
- In the Auxiliary residuals graphics window select Irregular, Level residual, Index plot, Density, Histogram, Normal, QQ plot, Write normality tests, and Write values exceeding (3.5).
- Click OK.
- Use the menu <File, Save> or <Ctrl+S> to save these graphs.

The STAMP graphics window in GiveWin displays the auxiliary residuals of the irregular and of the level component and their density function and normal probability plot: see Figure 3.6.10.

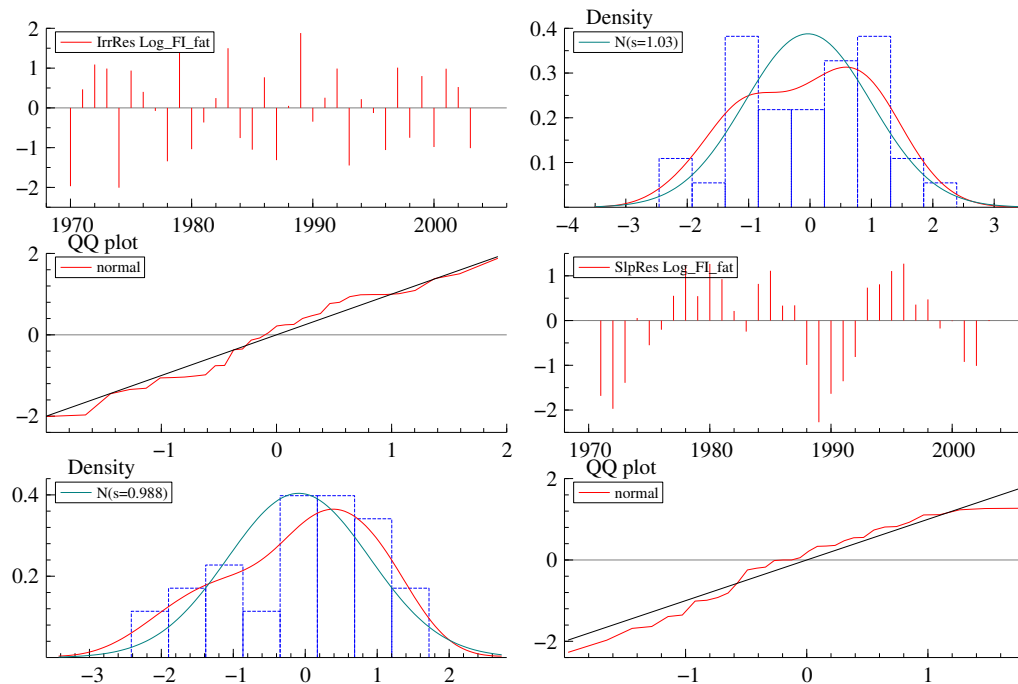


Figure 3.6.10: Auxiliary residuals and corresponding tests for the deterministic level stochastic slope model applied to the log of Finnish fatalities.

The following output describes the auxiliary residual test results for normality as can be found in the Results window of GiveWin.


```

Normality test for IrrRes Log_FI_fat
Sample Size      34
Mean              -0.034110
Std.Devn.         1.028845
Skewness          -0.165816
Excess Kurtosis   -1.021443
Minimum           -2.007568
Maximum           1.881536
Skewness Chi^2(1)  0.15581 [0.6930]
Kurtosis Chi^2(1)  1.4781 [0.2241]
Normal-BS Chi^2(2) 1.6339 [0.4418]
Normal-DH Chi^2(2) 2.0182 [0.3645]

Normality test for IrrRes Log_FI_fat
Sample Size      34
Mean              -0.034110
Std.Devn.         1.028845
Skewness          -0.165816
Excess Kurtosis   -1.021443
Minimum           -2.007568
Maximum           1.881536
Skewness Chi^2(1)  0.15581 [0.6930]
Kurtosis Chi^2(1)  1.4781 [0.2241]
Normal-BS Chi^2(2) 1.6339 [0.4418]
Normal-DH Chi^2(2) 2.0182 [0.3645]

```

Figure 3.6.10 and the auxiliary residual tests demonstrate that the auxiliary residuals of both the irregular and the slope component satisfy the assumption of normality. Note that none of the auxiliary residuals of the irregular (top left graph in Figure 3.6.10) is larger than 2 or smaller than -2, whereas only one of the auxiliary residuals of the slope component (middle left graph) slightly exceeds these bounds. So, there is no indication of outlier observations or structural breaks in the slope component.

Step 7: Conclusion of analysis

The residuals obtained with the analysis of the log of the annual Finnish fatalities from 1970 to 2003 with the deterministic level and stochastic slope model (or: smooth trend model) satisfy all the model assumptions of independence, homoscedasticity, and normality. Therefore, it is an appropriate model for describing this series.

Step 8: Forecasting

Since the deterministic level stochastic slope model provides an appropriate description of the log of the Finnish fatalities series, as a final step in the analysis we will compute seven-year forecasts for this series. By performing an anti-log analysis the forecasts will be re-expressed in terms of the original count data. The forecasts are made according to the instructions in step 8 of the analysis of the local level model (Section 3.6.2.2).

- Go to the STAMP window again and choose <Test, Forecasting...>.
- In the Forecasting window select 7 as the number of forecasts, Trend, Modified anti-log analysis, and Write forecasts Y.
- Click OK.

The STAMP graphics window in GiveWin displays the original observations extended with the seven-years forecasts with 70% confidence interval (i.e., plus and minus one estimated standard deviation) at the top, and the original observations and the extrapolated trend at the bottom of Figure 3.6.11.

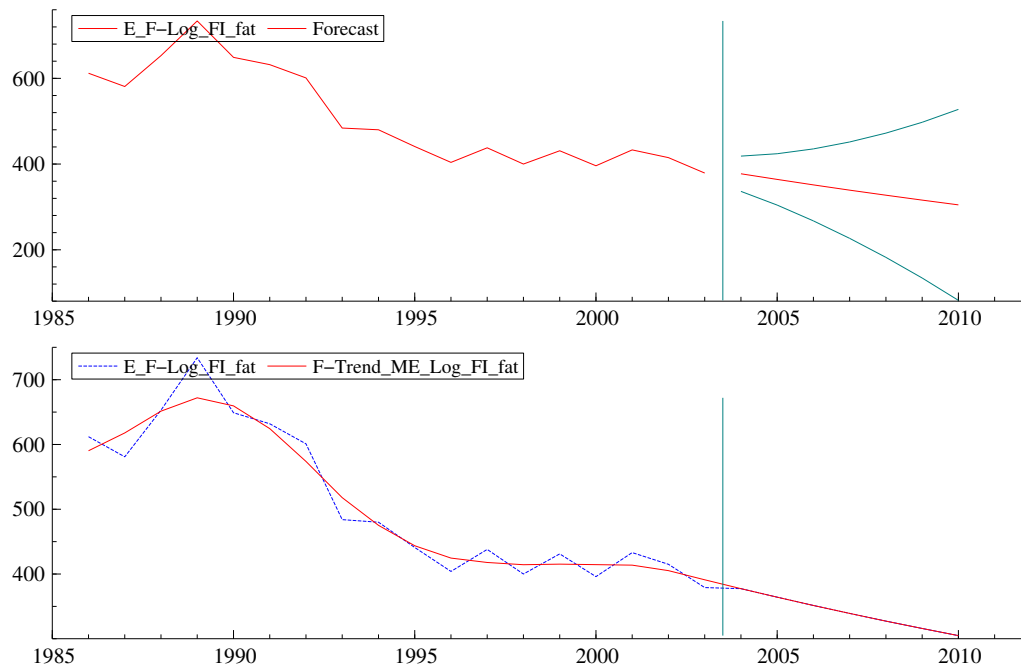


Figure 3.6.11: Anti-logged seven-year forecasts (2004-2010) of the deterministic level stochastic slope model applied to the log of annual Finnish fatalities, 1970-2003.

The bottom figure illustrates that a linear trend model always yields forecasts by extending the last value of the trend (i.e., level plus slope) in the series. This is in complete agreement with the fact that we are dealing with a **linear trend** model.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps).

In the Results window of GiveWin the forecasts for the original observed time series have been added:

Eq 1 : Forecasts for E_F-Log_FI_fat.
Anti-log

Period	Forecast	R.m.s.e.	- Rmse	+ Rmse
2004. 1	377.38	41.142	336.24	418.52
2005. 1	364.18	59.896	304.28	424.08
2006. 1	351.44	84.011	267.43	435.45
2007. 1	339.15	112.49	226.66	451.64
2008. 1	327.29	145.04	182.25	472.32
2009. 1	315.84	181.72	134.11	497.56
2010. 1	304.79	222.82	81.973	527.61

The list of forecast results gives for each time point the forecast, its standard error, and the lower and upper bound of the 70% confidence interval.

3.6.4 Local linear trend plus seasonal model

In this section, the seasonal component will be added to the local linear trend model, so as to obtain the *local linear trend plus seasonal model*. The model will be applied to the monthly number of drivers killed or seriously injured (KSI) as observed in the UK for the period January 1969 through December 1984. Modelling a seasonal component only makes sense when we are dealing with seasonal (monthly, quarterly, weekly, etc.) time series data, and this type of component was therefore not considered in the analysis of the *annual* data discussed in Section 3.6.2 and 3.6.3. In addition to the theory on this model and the results of its application to the UK data, as presented in Section 3.6.3 of the Methodology report, this section explains how the model is built in STAMP and how the results can be interpreted.

This section presents a step-by-step description of the analysis of the UK drivers KSI time series using a local linear trend plus seasonal model. Contrary to the setup of Sections 3.6.2 and 3.6.3, in this section all components will be assumed stochastic from the start. As such, this analysis example follows the recommended way of analysing time series data by state space techniques (see Section 3.6.7).

Step 1: Start of analysis and data load

First, we open GiveWin, load the data, and start STAMP.

- If GiveWin is not yet open, then start GiveWin2.
- If GiveWin is already open from a previous exercise, then close all results, data, and graphics windows in GiveWin by clicking on the icon with the cross in the top right corner of each window.
- Use the menu <File, Open Data File...> to open the file "UKdriversKSI.in7".

The data file is loaded and displayed in a minimized window at the bottom of the GiveWin main window. To view the data file:

- Click on the icon with the two overlapping boxes.

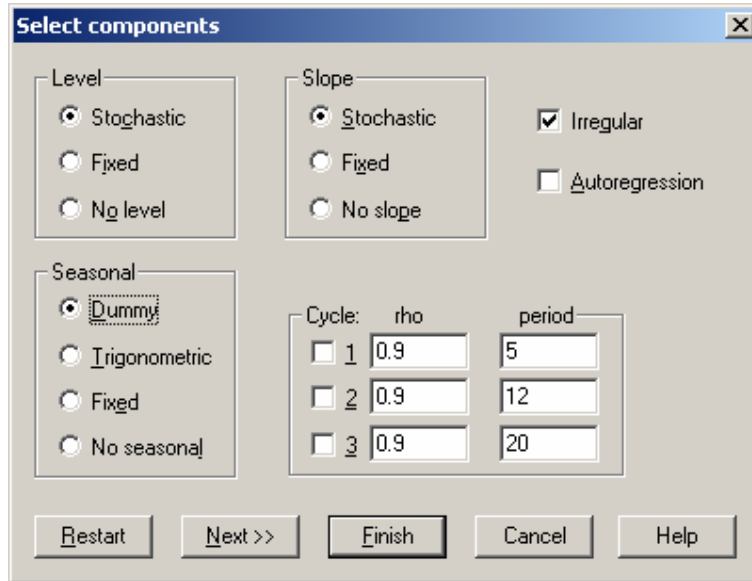
The data file consists of four variables: the monthly number of drivers KSI in the UK for the months January 1969 through December 1984 (UKdriversKSI), the petrol price in the UK in the same months (PetrolPrice), and the logarithm of the same time series (Log_UKdriversKSI and Log_PetrolPrice). The variable PetrolPrice will be used in Section 3.6.6.

- Minimize the data file window again and use the menu <Modules, Start Stamp> to start the STAMP program.

Step 2: Model Formulation

In this step, we define the local linear trend plus seasonal model:

- In STAMP, choose the menu <Model, Formulate...>.
- In the Data selection window select the variable Log_UKdriversKSI.
- Then click OK.
- In the Select components window, choose a Stochastic Level, Stochastic slope, Irregular, and Dummy seasonal:



- Then click on the Finish button.

Step 3: Model estimation and inspection of results

- In the Estimate Model window, select Maximum Likelihood.
- Click OK.

The model is estimated, and the following output appears in the GiveWin Results window:

Equation 4.

Log_UKdriversKSI = Trend + Dummy seasonal + Irregular

Estimation report

Model with 4 parameters (3 restrictions).

Parameter estimation sample is 1969. 1 - 1984.12. (T = 192).

Log-likelihood kernel is 2.279365.

Very strong convergence in 11 iterations.

(likelihood cvg 7.253531e-013
gradient cvg 7.576162e-008
parameter cvg 4.127312e-006)

Eq 4 : Diagnostic summary report.

Estimation sample is 1969. 1 - 1984.12. (T = 192, n = 179).

Log-Likelihood is 437.638 (-2 LogL = -875.276).

Prediction error variance is 0.00556772

```
Summary statistics
      Log_UKdriver
Std.Error      0.074617
Normality      3.7108
H( 59)         1.0905
r( 1)          0.026288
r(12)          0.029382
DW             1.9311
Q(12, 9)       12.378
Rs^2           0.23046
```

Eq 4 : Estimated variances of disturbances.

```
Component      Log_UKdriversKSI (q-ratio)
Irr            0.0034678 ( 1.0000)
Lvl            0.0010009 ( 0.2886)
Slp            0.00000 ( 0.0000)
Sea            0.00000 ( 0.0000)
```

Inspecting the estimated variances of the disturbances, we see that the variances corresponding to the slope component (Slp) and the seasonal component (Sea) are (almost) equal to zero. Therefore, we repeat the analysis (steps 2 and 3) with a deterministic slope and seasonal component (by choosing Stochastic level, Fixed slope, and Fixed seasonal in the STAMP Select components window). This yields the following output in the GiveWin results window:

Equation 5.

Log_UKdriversKSI = Trend + Fixed seasonal + Irregular

```
Estimation report
Model with 2 parameters ( 1 restrictions).
Parameter estimation sample is 1969. 1 - 1984.12. (T = 192).
Log-likelihood kernel is 2.279365.
Very strong convergence in 3 iterations.
( likelihood cvg 1.23912e-013
  gradient cvg 3.566036e-008
  parameter cvg 3.127122e-007 )
```

Eq 5 : Diagnostic summary report.
Estimation sample is 1969. 1 - 1984.12. (T = 192, n = 190).
Log-Likelihood is 437.638 (-2 LogL = -875.276).
Prediction error variance is 0.00550388

```
Summary statistics
      Log_UKdriver
Std.Error      0.074188
Normality      2.9938
H( 63)         1.0297
r( 1)          0.029490
r(12)          -0.0070854
DW             1.9343
Q(12,11)       12.196
Rs^2           0.23928
```

Eq 5 : Estimated variances of disturbances.

```
Component      Log_UKdriversKSI (q-ratio)
Irr            0.0034678 ( 1.0000)
Lvl            0.0010009 ( 0.2886)
```

- Check the results (sample period, log-likelihood, estimated variance of disturbances).

The estimation report tells that there was very strong convergence in three iterations. The diagnostic summary report shows that the number of observations is 192, that the value of the log-likelihood function at convergence is 437, and that the prediction error variance is 0.00550. The estimated variance of the level disturbances is unequal to zero.

- In the STAMP window choose <Test, Further output...> in the menu.
- Select Additional output, Get steady state, Anti-log analysis, and State and regression output.
- Click OK.

The GiveWin Results window will display the following additional results:

Eq 5 : Estimated standard deviations of disturbances.

```
Component      Log_UKdriversKSI (q-ratio)
Irr             0.058888 ( 1.0000)
Lvl             0.031637 ( 0.5372)
```

Eq 5 : Estimated coefficients of final state vector.

Variable	Coefficient	R.m.s.e.	t-value	
Lvl	7.2404	0.038792	186.65	[0.0000]
Slp	-0.00090532	0.0023076	-0.39233	[0.6953]
Sea_ 1	0.017176	0.016254	1.0567	[0.2920]
Sea_ 2	-0.10933	0.016219	-6.7409	[0.0000]
Sea_ 3	-0.070091	0.016192	-4.3288	[0.0000]
Sea_ 4	-0.14686	0.016171	-9.0817	[0.0000]
Sea_ 5	-0.055472	0.016157	-3.4333	[0.0007]
Sea_ 6	-0.092507	0.016150	-5.7279	[0.0000]
Sea_ 7	-0.043175	0.016150	-2.6734	[0.0082]
Sea_ 8	-0.032024	0.016157	-1.982	[0.0489]
Sea_ 9	0.0058909	0.016171	0.36429	[0.7160]
Sea_10	0.086848	0.016192	5.3637	[0.0000]
Sea_11	0.19221	0.016219	11.851	[0.0000]

Anti-log trend analysis

Trend value at end of period is 1394.63.

Growth rate at end of period is -0.000905317 (-1.08638 % per "year").

From the estimated coefficients of the final state vector and their t-values, we can conclude that the (constant) value of -0.00090532 for the slope component does not significantly deviate from zero. The same applies to the values for the first and ninth estimates of the seasonal component (corresponding to the months of January and September).

Because the deterministic slope component is not significantly different from zero, we drop the slope component from the model and repeat the analysis (steps 2 and 3) with a stochastic level deterministic seasonal model (by selecting Stochastic level and Fixed seasonal in the STAMP Select components window). This yields the following output in the GiveWin results window:

Equation 6.

Log_UKdriversKSI = Level + Fixed seasonal + Irregular

Estimation report

Model with 2 parameters (1 restrictions).

```
Parameter estimation sample is 1969. 1 - 1984.12. (T = 192).
Log-likelihood kernel is 2.313251.
Very strong convergence in 2 iterations.
( likelihood cvg 1.343833e-015
  gradient cvg 2.082778e-008
  parameter cvg 3.620022e-009 )
```

Eq 6 : Diagnostic summary report.

```
Estimation sample is 1969. 1 - 1984.12. (T = 192, n = 191).
Log-Likelihood is 444.144 (-2 LogL = -888.289).
Prediction error variance is 0.00550316
```

```
Summary statistics
      Log_UKdriver
Std.Error      0.074183
Normality      3.2218
H( 63)         1.0686
r( 1)          0.041087
r(12)         -0.00017919
DW             1.9155
Q(12,11)       12.000
Rs^2           0.23938
```

Eq 6 : Estimated variances of disturbances.

```
Component      Log_UKdriversKSI (q-ratio)
Irr             0.0035140 ( 1.0000)
Lvl             0.00094564 ( 0.2691)
```

- Check the results (sample period, log-likelihood, estimated variance of disturbances).

The estimation report tells that there was very strong convergence in two iterations. The estimated variance of the level disturbances is unequal to zero, meaning that the level component should be treated stochastically. The diagnostic summary report shows that the value of the log-likelihood function at convergence is 444, which is a somewhat larger value than in the stochastic level and deterministic slope and seasonal model (437).

- In the STAMP window choose <Test, Further output...> in the menu.
- Select Additional output, Get steady state, Anti-log analysis, and State and regression output.
- Click OK.

The GiveWin Results window will display the following additional results:

Eq 6 : Estimated standard deviations of disturbances.

```
Component      Log_UKdriversKSI (q-ratio)
Irr             0.059279 ( 1.0000)
Lvl             0.030751 ( 0.5188)
```

Eq 6 : Estimated coefficients of final state vector.

Variable	Coefficient	R.m.s.e.	t-value	
Lvl	7.2414	0.038351	188.82	[0.0000]
Sea_ 1	0.017272	0.016225	1.0646	[0.2884]
Sea_ 2	-0.10925	0.016192	-6.7474	[0.0000]
Sea_ 3	-0.070030	0.016165	-4.3321	[0.0000]
Sea_ 4	-0.14682	0.016146	-9.0933	[0.0000]
Sea_ 5	-0.055446	0.016132	-3.4369	[0.0007]
Sea_ 6	-0.092499	0.016126	-5.7361	[0.0000]
Sea_ 7	-0.043184	0.016126	-2.678	[0.0081]
Sea_ 8	-0.032050	0.016132	-1.9867	[0.0484]
Sea_ 9	0.0058471	0.016146	0.36215	[0.7176]


```

Sea_10      0.086786      0.016165      5.3686 [ 0.0000]
Sea_11      0.19213      0.016192     11.866 [ 0.0000]

```

Anti-log trend analysis
Trend value at end of period is 1396.04.

The values for the first and ninth estimates of the seasonal component (corresponding to the months of January and September) do not significantly deviate from zero. At the end of the period, the trend value as presented by the anti-log trend analysis is 1396.

- The STAMP output results concerning the summary statistics can be condensed into the following table (see also Table 3.6.6 in the Methodology report):

	Statistic	Value	Critical 5% value ^a	Assumption satisfied
Independence	Q(12,11)	12.0	19.68	+
	r(1)	0.0411	0.14	+
	r(12)	-0.00018	0.14	+
Homoscedasticity	H(63)	1.07	1.65	+
Normality	N	3.22	5.99	+

Table 3.6.6: Diagnostic test results for the stochastic level and deterministic seasonal model applied to the log of the number of UK drivers KSI. ^aProbability that statistic exceeds critical value is 0.05.

From Table 3.6.6, we can conclude that the stochastic level and deterministic seasonal model satisfies all model assumptions.

Step 4: Graphics of model components

- In the STAMP window choose menu <Test, Components graphics...>.
- Select Trend, Seasonal, Irregular, and Smoothed.
- Click OK.

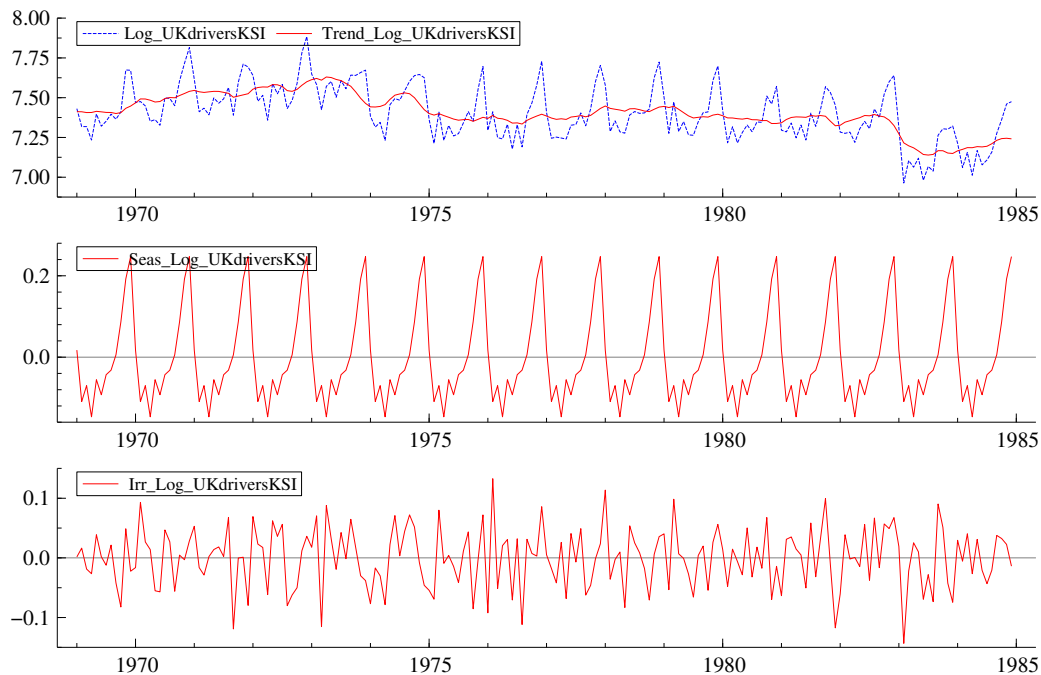


Figure 3.6.12: Observed log-transformed time series and the trend of the stochastic level and deterministic seasonal model (top graph), seasonal component (middle graph), and irregular component (bottom graph) for the log of UK drivers KSI.

The STAMP Graphics window appears with graphs of the observed log-transformed time series and the modelled trend, seasonal, and irregular (see Figure 3.6.12).

The seasonal component is close to zero for the months January and September, is negative for February to August, and is positive for October to December.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps). Minimize the STAMP Graphics window.

Step 5: Test of model residuals

- Go back to the STAMP window and choose <Test, Residuals graphics...>.
- In the Residual graphics window select Residuals, Correlogram, with 14, Density, Histogram, Normal, QQ plot, and Write diagnostic tests.
- Click OK.

Figure 3.6.13 shows the standardized residuals and their correlogram, density function, and normal probability plot as depicted by the STAMP graphics window in GiveWin.

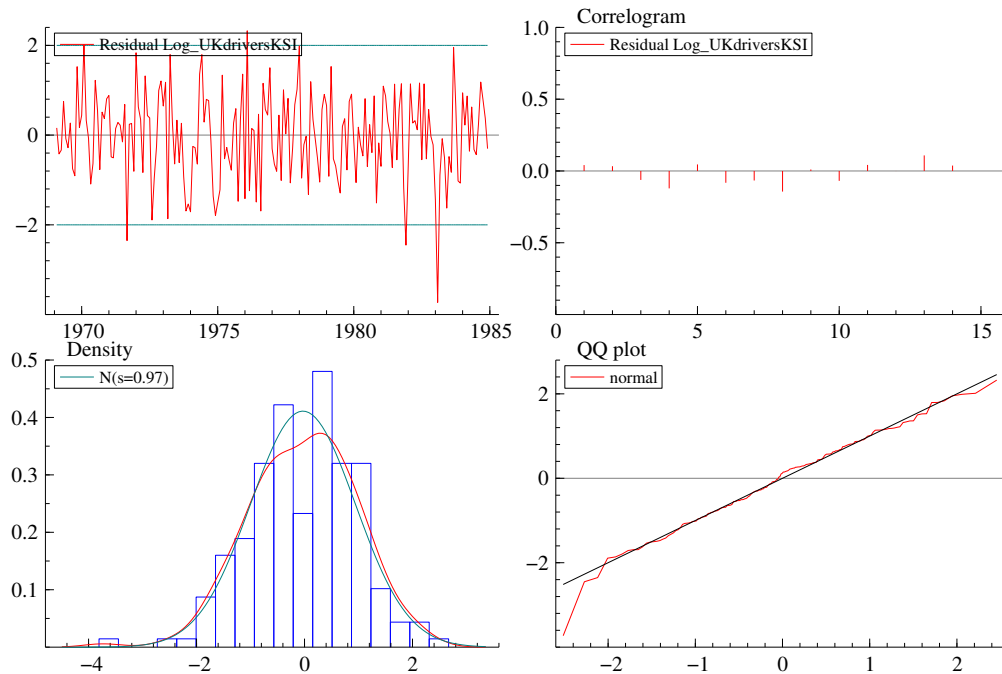


Figure 3.6.13: Residuals and residual tests for the stochastic level deterministic seasonal model applied to the log of UK drivers KSI.

The top left graph of Figure 3.6.13 shows that only five out of the 192 residuals exceed the 95% confidence limits. Still, the residual corresponding to February 1983 is very extreme: -3.73. Under the assumption of a normal distribution with zero mean and unit standard deviation, the probability of a value smaller than -3.73 is 0.01%! From the top right graph, we learn that for 1 out of the 14 lags considered the autocorrelation is (just) outside the 95% confidence interval, which is defined by the boundaries $-2/\sqrt{T} = -0.14$ and $+2/\sqrt{T} = 0.14$. The bottom graphs show that the assumption of normality of the residuals is satisfied, which confirms the normality test result displayed in Table 3.6.6.

- Use the menu <File, Save> or <Ctrl+S> to save the graphs and minimize the STAMP Graphics window.

In the Results window of GiveWin, the following residual test results can be found (only part of the results are printed below):

```
Goodness-of-fit results for Residual Log_UKdriversKSI
Information criterion of Akaike      AIC      -5.067016
... of Schwartz (Bayes)             BIC      -4.846457
```

```
Serial correlation statistics for Residual Log_UKdriversKSI.
Lag  dF      SerCorr      BoxLjung      ProbChi2 (dF)
 1    0        0.0411
 2    0        0.0332
 3    1       -0.0618        1.2906      [ 0.2559]
 4    2       -0.1199        4.1225      [ 0.1273]
 5    3        0.0452        4.5278      [ 0.2098]
 6    4       -0.0812        5.8412      [ 0.2113]
 7    5       -0.0654        6.6993      [ 0.2440]
 8    6       -0.1413       10.7213      [ 0.0974]
 9    7        0.0105       10.7438      [ 0.1502]
10    8       -0.0667       11.6512      [ 0.1675]
11    9        0.0413       12.0002      [ 0.2133]
12   10       -0.0002       12.0002      [ 0.2850]
13   11        0.1043       14.2549      [ 0.2192]
14   12        0.0364       14.5314      [ 0.2681]
```

The value of the AIC is -5.07. Furthermore, the BoxLjung test statistic for the autocorrelations of the first 14 lags shows that the residuals satisfy the assumption of independence.

Step 6: Test of auxiliary residuals

- Go to the STAMP window again and choose <Test, Auxiliary residuals graphics...>.
- In the Auxiliary residuals graphics window select Irregular, Level residual, Index plot, Density, Histogram, Normal, QQ plot, Write normality tests, and Write values exceeding (3.5).
- Click OK.

The STAMP graphics window in GiveWin displays the auxiliary residuals of the irregular and of the level component and their density function and normal probability plot: see Figure 3.6.14. The output below the figure describes the auxiliary residual test results as can be found in the Results window of GiveWin.

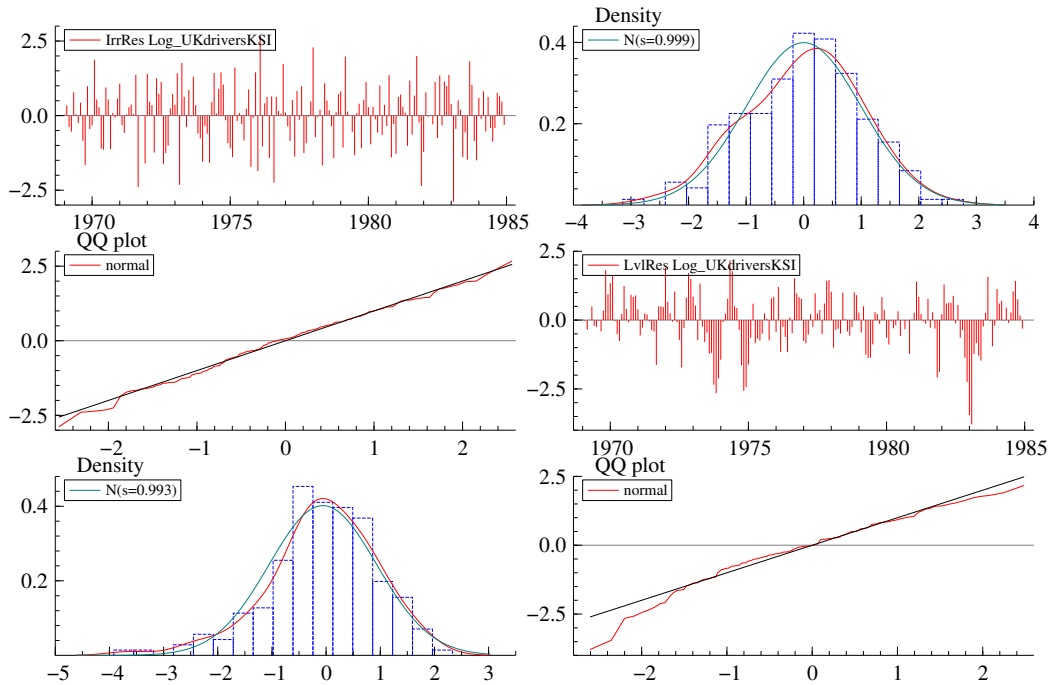


Figure 3.6.14: Auxiliary residuals and corresponding tests for the stochastic level deterministic seasonal model applied to the log of UK drivers KSI.

```

Normality test for IrrRes Log_UKdriversKSI
Sample Size      192
Mean              0.000027
Std.Devn.         0.998613
Skewness          -0.190959
Excess Kurtosis   -0.163838
Minimum          -2.884350
Maximum           2.672501
Skewness  Chi^2(1)  1.1669  [0.2800]
Kurtosis  Chi^2(1)  0.21474 [0.6431]
Normal-BS  Chi^2(2)  1.3816  [0.5012]
Normal-DH  Chi^2(2)  1.4114  [0.4938]

```

```

Normality test for LvlRes Log_UKdriversKSI
Sample Size      192
Mean            -0.059080
Std.Devn.        0.992958
Skewness         -0.669142
Excess Kurtosis   1.084877
Minimum         -3.789216
Maximum           2.172144
Skewness  Chi^2(1)  14.328  [0.0002]
Kurtosis  Chi^2(1)   9.4157 [0.0022]
Normal-BS  Chi^2(2)  23.744  [0.0000]
Normal-DH  Chi^2(2)  13.024  [0.0015]

```

```

Eq 6 : Large values in LvlRes Log_UKdriversKSI.
Period      Value
1983. 2     -3.7892 [ 0.0001]

```

Both Figure 3.6.14 and the auxiliary residual tests demonstrate that the auxiliary residuals of the irregular component satisfy the assumption of normality, whereas those of the level component do not satisfy this assumption. The latter fact means that we must be cautious with the

interpretation of the test statistics of individual auxiliary residuals and it underlines the importance of looking at the graphical output (Koopman et al, 2000).

The auxiliary residual of the level component corresponding to February 1983 is extremely large (-3.79), which is an indication of a structural break in the level component.

- Use the menu <File, Save> or <Ctrl+S> to save the graphs and minimize the STAMP Graphics window.

Step 7: Conclusion of analysis

The residuals obtained with the analysis of the log of the monthly UK drivers KSI from January 1969 to December 1984 with the stochastic level and deterministic seasonal model satisfy all the model assumptions of independence, homoscedasticity, and normality. However, the auxiliary residual of the level component for February 1983 is found to be extremely large (-3.79), which is an indication of a structural break in this component. Furthermore, the auxiliary residuals of the level component do not satisfy the assumption of normality. For these reasons, we expect that the model can be improved by adding an intervention variable to the local level and deterministic seasonal model. In fact, there is a very good reason why February 1983 was a special month for road safety in the UK. It was in that month that the seatbelt law was introduced, requiring front seat passengers in cars to wear a seatbelt. In Section 3.6.5 we will therefore investigate the effect of modelling this event by adding a level shift intervention variable to the the local level and deterministic seasonal model.

Step 8: Forecasting

Because the stochastic level and deterministic seasonal model for describing the log of the monthly number of UK drivers KSI from January 1969 to December 1984 can still be improved, as will be discussed in the following sections, the issue of obtaining forecasts from this series is postponed until Section 3.6.6.

3.6.5 Intervention variables

In the previous section, we discussed that it could be worthwhile to add an intervention variable to the stochastic level deterministic and seasonal model of the (log of the) number of UK drivers KSI for the period January 1969 through December 1984. The reason for this recommendation was the extremely large value of the February 1983 auxiliary residual of the level component. As stated in Section 3.6.4, this point in time coincides with the introduction of the seat belt law in the UK.

Because the extremely large value concerns the auxiliary residual of the level component, in this section we will add a level shift variable to the stochastic level and deterministic seasonal model discussed in the previous section.

Step 1: Start of analysis and data load

In this first step of the analysis, we open GiveWin, load the data, and start STAMP, if needed.

- If GiveWin is not yet open, then start GiveWin2.
- If GiveWin is already open but with another dataset than the UK drivers KSI, then close all results, data, and graphics windows in GiveWin by clicking on the icon with the cross in the top right corner of each window. Use the menu <File, Open Data File...> to open the file “UKdriversKSI.in7”.
- If GiveWin is already open and the UK drivers KSI dataset is already loaded, then proceed to the next instruction.

The data file is loaded and displayed in a minimized window at the bottom of the GiveWin main window. To view the data file:

- Click on the icon with the two overlapping boxes.
- Minimize the data file window again and use the menu <Modules, Start Stamp> to start the STAMP program.

Step 2: Model Formulation

In this step, we will add a level shift in February 1983 to the stochastic level deterministic seasonal model from the previous section. First, formulate the model:

- In STAMP, choose the menu <Model, Formulate>.
- In the Data selection window select the variable Log_UKdriversKSI.
- Then click OK.
- In the Select components window, choose a Stochastic Level, No slope, Irregular, and Fixed seasonal.

Next, we add the level shift:

- Click Next.
- Select the period in sample: year is 1983 and period is 2.
- Click Level.

The text "Lvl 1983.2" appears in the list of interventions.

- Then click on the Finish button.

Step 3: Model estimation and inspection of results

- In the Estimate Model window, select Maximum Likelihood.
- Click OK.

The model is estimated, and the following output appears in the GiveWin Results window:

Equation 1.

Log_UKdriversKSI = Level + Fixed seasonal + Interv + Irregular

Estimation report

Model with 2 parameters (1 restrictions).

Parameter estimation sample is 1969. 1 - 1984.12. (T = 192).

Log-likelihood kernel is 2.339682.

Very strong convergence in 2 iterations.

(likelihood cvg 1.328653e-015
gradient cvg 3.952394e-009
parameter cvg 5.449141e-014)

Eq 1 : Diagnostic summary report.

Estimation sample is 1969. 1 - 1984.12. (T = 192, n = 191).

Log-Likelihood is 449.219 (-2 LogL = -898.438).

Prediction error variance is 0.00501578

Summary statistics

	Log_UKdriver
Std.Error	0.070822
Normality	2.4014
H(63)	0.75510
r(1)	0.079936
r(12)	0.058582
DW	1.8396
Q(12,11)	15.267
Rs^2	0.30674

Eq 1 : Estimated variances of disturbances.

Component	Log_UKdriversKSI (q-ratio)
Irr	0.0037838 (1.0000)
Lvl	0.00047358 (0.1252)

- Check the results (sample period, log-likelihood, estimated variance of disturbances) and compare them with the results from the analysis without intervention in the previous section.

The estimation report tells that there was very strong convergence in two iterations. The diagnostic summary report shows that the addition of the intervention has improved the fit of the stochastic level deterministic seasonal model: the value of the log-likelihood function has increased from 437 to 449 and the prediction error variance has decreased from 0.00550 to 0.00502 (see the previous section).

- In the STAMP window choose <Test, Further output...> in the menu.
- Select Additional output, Get steady state, Anti-log analysis, and State and regression output.
- Click OK.

The GiveWin Results window will display the following additional results:

Eq 1 : Estimated standard deviations of disturbances.

```
Component      Log_UKdriversKSI (q-ratio)
Irr             0.061513 ( 1.0000)
Lvl             0.021762 ( 0.3538)
```

Eq 1 : Estimated coefficients of final state vector.

Variable	Coefficient	R.m.s.e.	t-value	
Lvl	7.2380	0.033960	213.13	[0.0000]
Sea_ 1	0.010335	0.015834	0.65272	[0.5147]
Sea_ 2	-0.10244	0.015814	-6.4775	[0.0000]
Sea_ 3	-0.064452	0.015770	-4.0869	[0.0001]
Sea_ 4	-0.14248	0.015736	-9.0543	[0.0000]
Sea_ 5	-0.052342	0.015711	-3.3315	[0.0010]
Sea_ 6	-0.090631	0.015696	-5.7741	[0.0000]
Sea_ 7	-0.042554	0.015691	-2.7119	[0.0073]
Sea_ 8	-0.032656	0.015696	-2.0805	[0.0388]
Sea_ 9	0.0040042	0.015711	0.25487	[0.7991]
Sea_10	0.083707	0.015736	5.3196	[0.0000]
Sea_11	0.18782	0.015770	11.91	[0.0000]

Anti-log trend analysis

Trend value at end of period is 1391.34.

Eq 1 : Estimated coefficients of explanatory variables.

Variable	Coefficient	R.m.s.e.	t-value	
Lvl 1983. 2	-0.23981	0.053072	-4.5185	[0.0000]

Just as in the model without intervention, the parameter estimates for the first and the ninth month of the seasonal component do not deviate from zero in the final state. At the end of the period, the trend value as presented by the anti-log trend analysis is 1391, which is a somewhat lower value than in the model without intervention (1396).

New in the output are the estimated coefficients of the explanatory variables. In this case there is only one explanatory variable which is the intervention variable for February 1983. The estimated coefficient for the level shift is -0.240, which corresponds to a $100 \cdot (e^{-0.23981} - 1) = -21.3\%$ change in the number of UK drivers KSI as a result of the introduction of the seat belt law. The coefficient is shown to be significant, but to guarantee that the t-value is reliable, we must first check whether the model assumptions are satisfied.

- The summary statistics in the output of STAMP can be used to set up the following table (see also Table 3.6.8 in the Methodology report):

	Statistic	Value	Critical 5% value ^a	Assumption satisfied
Independence	Q(12,11)	15.3	19.68	+
	r(1)	0.0799	0.14	+
	r(12)	0.0586	0.14	+
Homoscedasticity	H(63)	0.755	1.65	+

Normality	N	2.40	5.99	+
-----------	---	------	------	---

Table 3.6.7: Diagnostic test results for the stochastic level deterministic seasonal model with intervention applied to the log UK drivers KSI. ^aProbability that statistic exceeds critical value is 0.05.

Table 3.6.7 shows that the stochastic level and deterministic seasonal model with an intervention variable satisfies all model assumptions. This guarantees that the *t*-test for the regression coefficient of the intervention variable (see output above) is reliable.

Step 4: Graphics of model components

- In the STAMP window choose menu <Test, Components graphics...>. Select Trend, Seasonal, Irregular, and Smoothed.
- Click OK.

The STAMP Graphics window appears with graphs of the observed log-transformed time series and the modelled trend, seasonal, and irregular (see Figure 3.6.15).

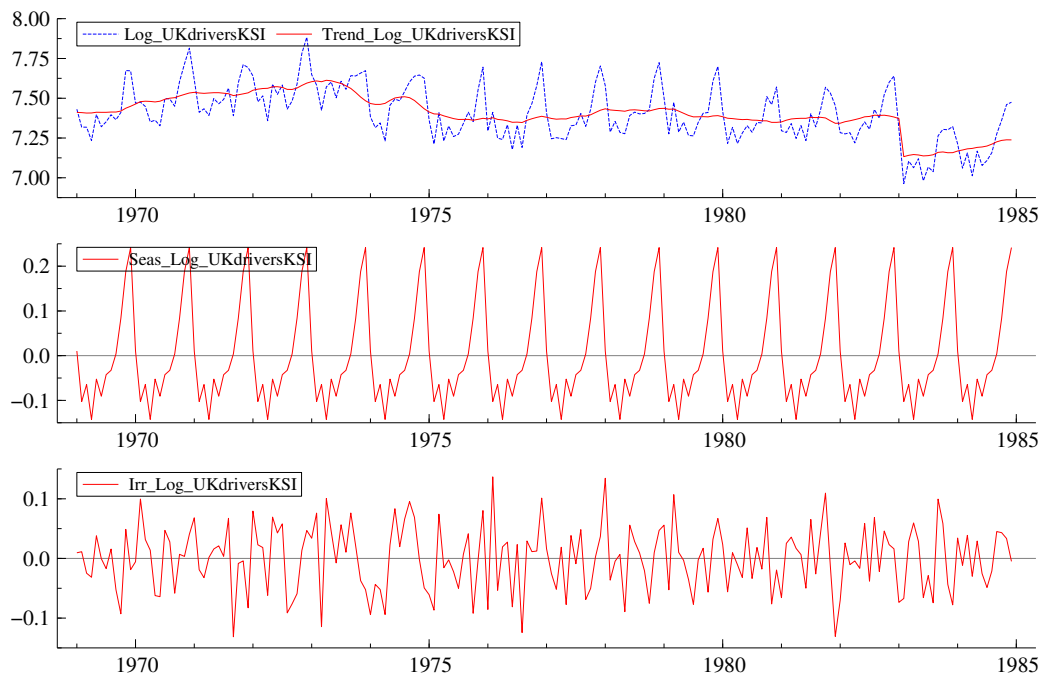


Figure 3.6.15: Observed log-transformed time series and the trend of the stochastic level deterministic seasonal model with intervention (top graph), seasonal component (middle graph), and irregular component (bottom graph) for the log of UK drivers KSI.

Compared to the level of the model without intervention, the level in the present model shows a sudden decrease at the start of 1983 (see Figure 3.6.12 and 3.6.15, top graphs). This results in a value for the irregular component in February 1983 which is considerably smaller, in absolute terms, than in the model without intervention (bottom graphs).

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps). Minimize the STAMP Graphics window.

Step 5: Test of model residuals

- Go back to the STAMP window and choose <Test, Residuals graphics...>.
- In the Residual graphics window select Residuals, Correlogram, with 14, Density, Histogram, Normal, QQ plot, and Write diagnostic tests.
- Click OK.

Figure 3.6.16 shows the standardized residuals and their correlogram, density function, and normal probability plot as depicted in the STAMP graphics window of GiveWin.

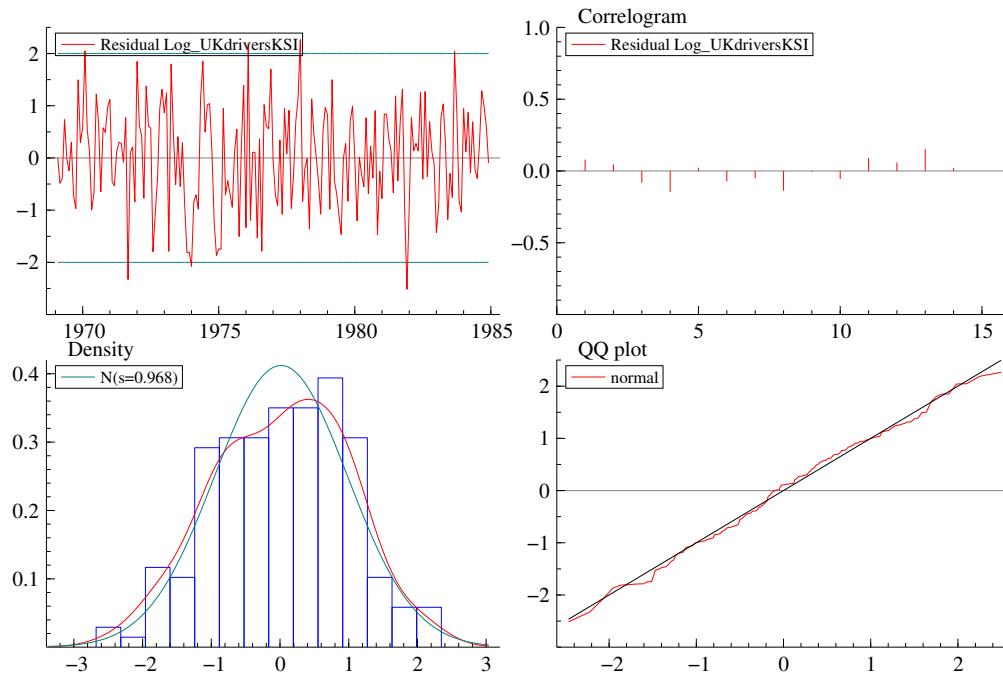


Figure 3.6.16: Residuals and residual tests for the stochastic level deterministic seasonal model with intervention applied to the log of UK drivers KSI.

When we compare the top left graphs of Figure 3.6.13 and 3.6.16, we can see that the extremely large residual in February 1993 has disappeared in the latter figure because of adding the intervention variable. In the model with intervention variable only seven residuals (3.6%) are outside the 95% confidence interval. However, none of them is extremely large.

From the top right graph, we learn that for 2 out of the first 14 lags the autocorrelation is (just) outside the 95% confidence interval, which is defined by the boundaries $-2/\sqrt{T} = -0.14$ and $+2/\sqrt{T} = 0.14$.

The bottom graphs show that the assumption of normality of the residuals is very well satisfied, which confirms the normality test results in Table 3.6.7.

- Use the menu <File, Save> or <Ctrl+S> to save the graphs and minimize the STAMP Graphics window.

In the Results window of GiveWin, the following residual test results can be found (only part of the results are printed below):

Goodness-of-fit results for Residual Log_UKdriversKSI
 Information criterion of Akaike AIC -5.149332
 ... of Schwartz (Bayes) BIC -4.911807

Serial correlation statistics for Residual Log_UKdriversKSI.

Lag	dF	SerCorr	BoxLjung	ProbChi2 (dF)
1	0	0.0799		
2	0	0.0449		
3	1	-0.0809	2.9183	[0.0876]
4	2	-0.1449	7.0597	[0.0293]
5	3	0.0232	7.1660	[0.0668]
6	4	-0.0716	8.1885	[0.0849]
7	5	-0.0489	8.6681	[0.1231]
8	6	-0.1351	12.3438	[0.0547]
9	7	-0.0049	12.3487	[0.0897]
10	8	-0.0541	12.9445	[0.1138]
11	9	0.0888	14.5601	[0.1037]
12	10	0.0586	15.2669	[0.1226]
13	11	0.1458	19.6698	[0.0501]
14	12	0.0201	19.7541	[0.0719]

The addition of the intervention to the model has improved the goodness-of-fit: the AIC has decreased (from -5.07 to -5.15) as well as the BIC (from -4.85 to -4.91).

Step 6: Test of auxiliary residuals

- Go to the STAMP window again and choose <Test, Auxiliary residuals graphics...>.
- In the Auxiliary residuals graphics window select Irregular, Level residual, Index plot, Density, Histogram, Normal, QQ plot, Write normality tests, and Write values exceeding (3.5).
- Click OK.

The STAMP graphics window in GiveWin displays the auxiliary residuals of the irregular and of the level component and their density function and normal probability plot: see Figure 3.6.17.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs and minimize the STAMP Graphics window.

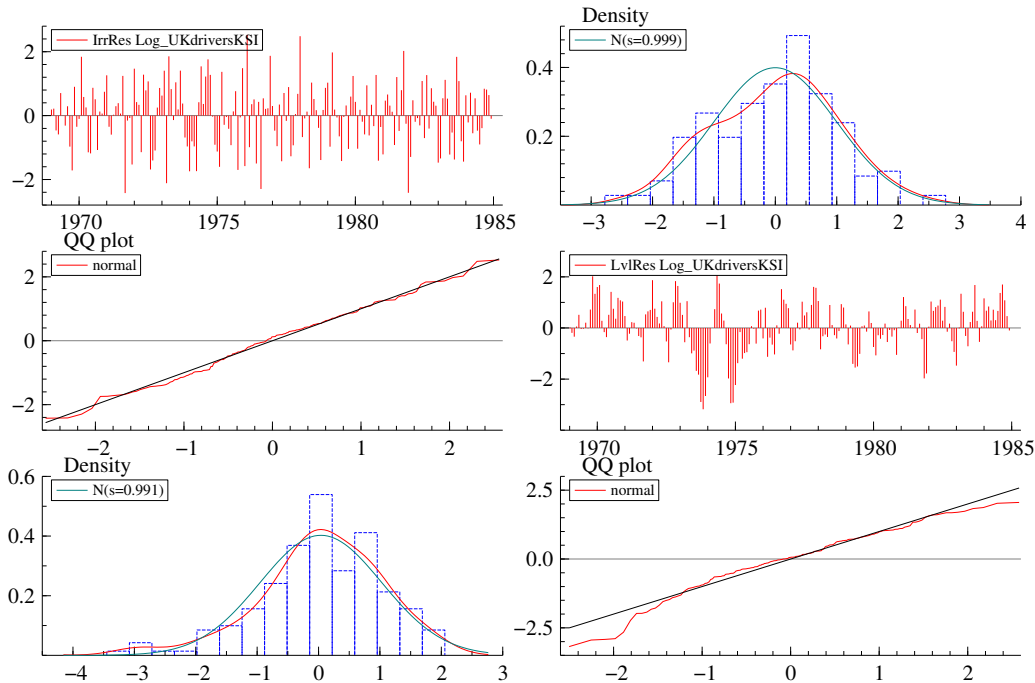


Figure 3.6.17: Auxiliary residuals and corresponding tests for the stochastic level and deterministic seasonal model with intervention applied to the log of UK drivers KSI.

The following output describes the auxiliary residual test results for normality as can be found in the Results window of GiveWin.

```

Normality test for IrrRes Log_UKdriversKSI
Sample Size      192
Mean              -0.000521
Std.Devn.         0.999144
Skewness          -0.081441
Excess Kurtosis   -0.412446
Minimum           -2.424978
Maximum           2.528211
Skewness  Chi^2(1)    0.21225 [0.6450]
Kurtosis  Chi^2(1)    1.3609 [0.2434]
Normal-BS  Chi^2(2)    1.5731 [0.4554]
Normal-DH  Chi^2(2)    1.2611 [0.5323]

```

```

Normality test for LvlResLog_UKdriversKSI
Sample Size      192
Mean              0.039095
Std.Devn.         0.991080
Skewness          -0.656959
Excess Kurtosis    0.780304
Minimum           -3.181119
Maximum           2.053959
Skewness  Chi^2(1)   13.811 [0.0002]
Kurtosis  Chi^2(1)    4.871 [0.0273]
Normal-BS  Chi^2(2)   18.682 [0.0001]
Normal-DH  Chi^2(2)   13.017 [0.0015]

```

Both Figure 3.6.17 and the auxiliary residual tests demonstrate that the auxiliary residuals of the irregular component satisfy the assumption of normality, whereas those of the level component do not satisfy this assumption. This was also the case in the model without intervention. From the middle right graph, we see that in 1973 and 1974 there are too many large negative auxiliary residuals of the level component. As mentioned above, this means that we must be careful with the interpretation of the test statistics for the level component's auxiliary residuals. However, the

extremely large value of the auxiliary residual of the level component for February 1983 that we observed in Figure 3.6.14 has disappeared in Figure 3.6.17. This is the result of including the level shift intervention variable in the model.

Step 7: Conclusion of analysis

The residuals obtained with the analysis of the log of the monthly UK drivers KSI from January 1969 to December 1984 with the stochastic level and deterministic seasonal model with intervention variable satisfy all the model assumptions of independence, homoscedasticity, and normality. The addition of the level shift intervention variable for February 1983 has improved the fit of the model, and the extremely large value of the auxiliary residual of the level component observed in the previous analysis has disappeared in the present one. The regression coefficient for the intervention variable is significant, and indicates that the introduction of the seat belt law resulted in a 21.3% reduction in the number of UK drivers KSI. However, the auxiliary residuals of the level component still do not satisfy the assumption of normality. Therefore, we must be careful with the interpretation of test statistics with respect to structural level breaks.

In Section 3.6.6, we will extend the model with yet another component: a continuous explanatory variable.

Step 8: Forecasting

Because in the next section we will extend the stochastic level and deterministic seasonal model including an intervention variable with yet another explanatory variable, we postpone the forecasting to that section.

3.6.6 Explanatory variables

In the previous section, we extended the stochastic level and deterministic seasonal model for the analysis of the log of the number of UK drivers KSI for the period January 1969 through December 1984 with an intervention variable. In this section, we will add yet another component: a continuous explanatory variable.

This continuous explanatory variable is the log of the monthly prices of petrol in the UK in the period January 1969 to December 1984 of which is assumed that it may have affected car mobility and thus also the number of drivers KSI. The seat belt law intervention from the previous section will be kept in the model.

Step 1: Start of analysis and data load

If needed, we will first open GiveWin, load the data, and start STAMP.

- If GiveWin is not yet open, then start GiveWin2.
- If GiveWin is already open from a previous exercise but with another dataset than the UK drivers KSI, then close all results, data, and graphics windows in GiveWin by clicking on the icon with the cross in the top right corner of each window. Use the menu <File, Open Data File...> to open the file "UKdriversKSI.in7".
- If GiveWin is already open and the UK drivers KSI dataset is already loaded, then proceed with the next instruction.

The data file is loaded and displayed in a minimized window at the bottom of the GiveWin main window. To view the data file:

- Click on the icon with the two overlapping boxes.
- Minimize the data file window again and use the menu <Modules, Start Stamp> to start the STAMP program.

Step 2: Model formulation

In this step, we will add the petrol price variable to the stochastic level and deterministic seasonal model with seat belt law intervention from the previous section:

- In STAMP, choose the menu <Model, Formulate>.
- In the Data selection window select the variable Log_UKdriversKSI and click Add.
- Also select the variable Log_PetrolPrice and click Add.
- Then click OK.
- In the Select components window, choose a Stochastic Level, No slope, Irregular, and Fixed seasonal.

To add the level shift intervention variable for February 1983:

- Click Next.

- Select the the point in the series: year is 1983 and period is 2.
- Click Level.
- Then click on the Finish button.

Step 3: Model estimation and inspection of results

- In the Estimate Model window, select Maximum Likelihood.
- Click OK.

The model is estimated, and the following output appears in the GiveWin Results window:

Equation 1.

Log_UKdriversKSI = Level + Fixed seasonal + Expl vars + Interv + Irregular

Estimation report

Model with 2 parameters (1 restrictions).
 Parameter estimation sample is 1969. 1 - 1984.12. (T = 192).
 Log-likelihood kernel is 2.342043.
 Very strong convergence in 2 iterations.
 (likelihood cvg 3.792323e-016
 gradient cvg 1.065814e-009
 parameter cvg 7.219832e-009)

Eq 1 : Diagnostic summary report.

Estimation sample is 1969. 1 - 1984.12. (T = 192, n = 191).
 Log-Likelihood is 449.672 (-2 LogL = -899.345).
 Prediction error variance is 0.00483573

Summary statistics

	Log_UKdriver
Std.Error	0.069539
Normality	1.9020
H(63)	0.87770
r(1)	0.10275
r(12)	0.052579
DW	1.7930
Q(12,11)	18.706
Rs^2	0.33163

Eq 1 : Estimated variances of disturbances.

Component	Log_UKdriversKSI (q-ratio)
Irr	0.0040344 (1.0000)
Lvl	0.00026772 (0.0664)

- Check the results (sample period, log-likelihood, estimated variance of disturbances) and compare them with the results from the analysis without intervention in the previous section.

We have very strong convergence in two iterations. The diagnostic summary report shows that the addition of the petrol price as explanatory variable has improved the stochastic level and deterministic seasonal model with seat belt law intervention: the value of the log-likelihood

function has increased from 449 to 450 and the prediction error variance has decreased from 0.00502 to 0.00484.

- In the STAMP window choose <Test, Further output...> in the menu.
- Select Additional output, Get steady state, Anti-log analysis, and State and regression output.
- Click OK.

The GiveWin Results window displays the following additional results:

Eq 1 : Estimated standard deviations of disturbances.

```
Component      Log_UKdriversKSI (q-ratio)
Irr             0.063517 ( 1.0000)
Lvl             0.016362 ( 0.2576)
```

Eq 1 : Estimated coefficients of final state vector.

Variable	Coefficient	R.m.s.e.	t-value	
Lvl	6.6317	0.21428	30.948	[0.0000]
Sea_ 1	0.0085360	0.015860	0.5382	[0.5911]
Sea_ 2	-0.10336	0.015836	-6.5267	[0.0000]
Sea_ 3	-0.064435	0.015801	-4.0778	[0.0001]
Sea_ 4	-0.14119	0.015783	-8.9458	[0.0000]
Sea_ 5	-0.052945	0.015757	-3.3601	[0.0009]
Sea_ 6	-0.088490	0.015764	-5.6135	[0.0000]
Sea_ 7	-0.039156	0.015787	-2.4803	[0.0140]
Sea_ 8	-0.031078	0.015754	-1.9727	[0.0500]
Sea_ 9	0.0039760	0.015756	0.25234	[0.8010]
Sea_10	0.080770	0.015815	5.1073	[0.0000]
Sea_11	0.18615	0.015816	11.769	[0.0000]

Anti-log trend analysis

Trend value at end of period is 758.765.

Eq 1 : Estimated coefficients of explanatory variables.

Variable	Coefficient	R.m.s.e.	t-value	
Log_PetrolPrice	-0.27721	0.098431	-2.8163	[0.0054]
Lvl 1983. 2	-0.23757	0.046430	-5.1167	[0.0000]

Just as in the model with seat belt law intervention but without petrol price as explanatory variable, the parameter estimates for the first and the ninth month of the seasonal component do not deviate from zero in the final state. At the end of the period, the trend value as presented by the anti-log trend analysis is 757, which is smaller than in the model with seat belt law intervention but without petrol price variable (1391). This large difference in the trend value is caused by the introduction of the explanatory variable, which explains part of the trend.

The regression coefficients for the seat belt law intervention (-0.23757) and for the log of petrol price (-0.27721) are both significant in this model. However, to guarantee that the t-values are reliable, we must first test the model assumptions.

- The summary statistics in the output of STAMP can be used to set up the following table (see also Table 3.6.10 in the Methodology report):

	Statistic	Value	Critical 5% value ^a	Assumption satisfied
Independence	Q(12,11)	18.7	19.68	+
	r(1)	0.102	0.14	+
	r(12)	0.0526	0.14	+
Homoscedasticity	H(63)	0.878	1.65	+

Normality	N	1.90	5.99	+
-----------	---	------	------	---

Table 3.6.8: Diagnostic test results for the stochastic level and deterministic seasonal model with intervention and explanatory variable applied to the log of UK drivers KSI.
^aProbability that statistic exceeds critical value is 0.05.

Table 3.6.8 shows that the stochastic level and deterministic seasonal model with intervention and explanatory variable satisfies all model assumptions. This guarantees that the *t*-tests for the regression coefficients of the intervention variable and the explanatory variable (see output above) are reliable. According to this analysis, the introduction of the seat belt law resulted in a $100 \times (e^{-0.23757} - 1) = -21.1\%$ change in the number of UK drivers KSI (which is virtually identical to what we found in the previous section), while the regression coefficient for log petrol price indicates that a 1% rise in petrol price was associated with a 0.28% reduction in the number of UK drivers KSI.

Step 4: Graphics of model components

- In the STAMP window choose menu <Test, Components graphics...>. Select Trend plus Xs, Seasonal, Irregular, and Smoothed.
- Click OK.

The STAMP Graphics window appears with graphs of the observed log-transformed time series and the modelled trend, seasonal, and irregular (see Figure 3.6.18).

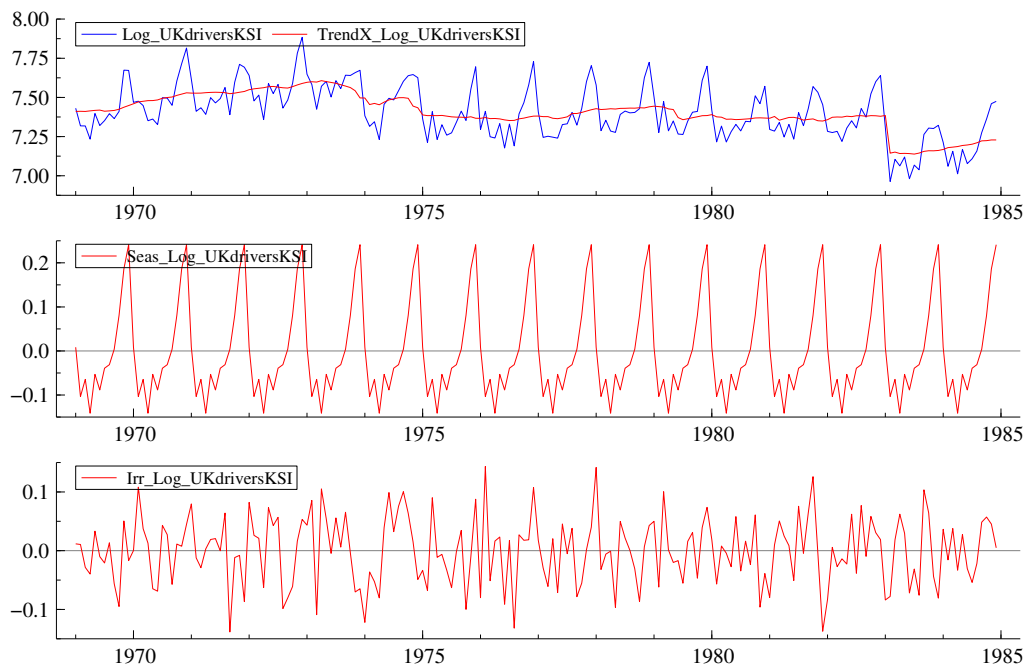


Figure 3.6.18: Observed log-transformed time series and the trend of the stochastic level deterministic seasonal model with intervention and explanatory variable (top graph), seasonal component (middle graph), and irregular component (bottom graph) for the log of UK drivers KSI.

Comparison of Figures 3.6.15 and 3.6.18 leads to the conclusion that there is no difference between the trend, seasonal, and irregular components of the model with and without petrol price as explanatory variable. The difference between the models lies in the composition of the trend component, which will be shown below.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps). Minimize the STAMP Graphics window.

Next, we will make a graph of the part of trend which is explained by the log of petrol price:

- In the STAMP window choose menu <Test, Components graphics...>. Select Trend, Trend plus Xs, and Smoothed.
- Click OK.

The STAMP Graphics window appears with graphs of the observed log-transformed time series and the modelled trend without and with the explained portion (see Figure 3.6.19).

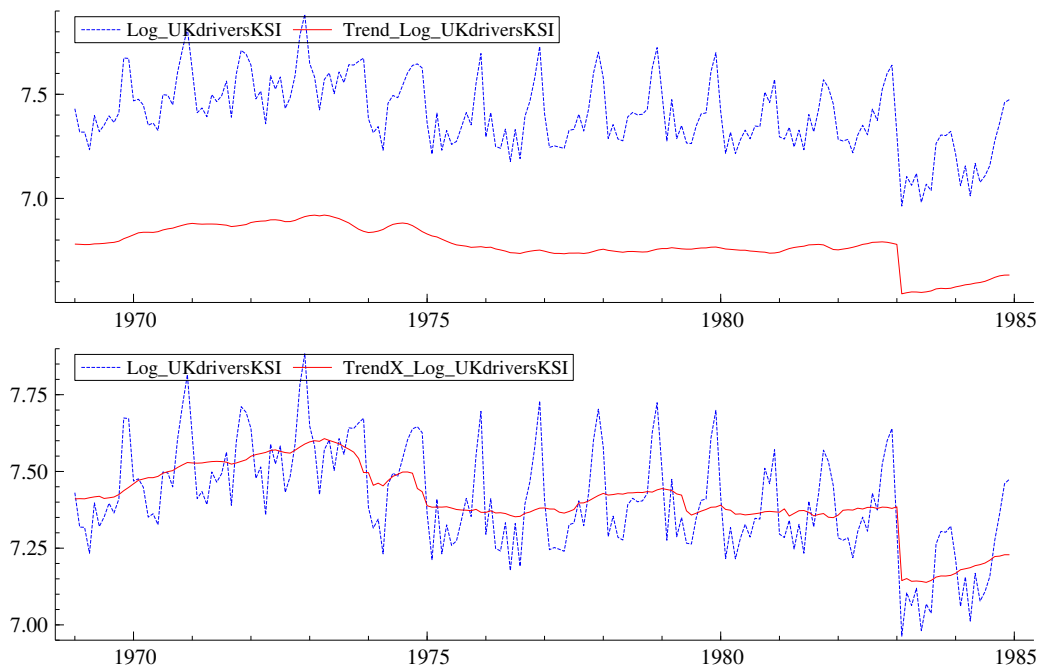


Figure 3.6.19: Observed log-transformed time series and the trend of the stochastic level deterministic seasonal model with intervention and explanatory variable for the log of UK drivers KSI, without the explained portion (top graph) and with the explained portion (bottom graph).

As can be seen in Figure 3.6.19, in the model with explanatory variable a considerable part of the trend is explained by the log of petrol price, whereas the remaining part of the trend is a stochastic level including the effect of the seat belt intervention. In the model without explanatory variable, the trend entirely consists of a stochastic level plus seat belt intervention effect.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps). Minimize the STAMP Graphics window.

Step 5: Test of model residuals

- Go back to the STAMP window and choose <Test, Residuals graphics...>.
- In the Residual graphics window select Residuals, Correlogram, with 14, Density, Histogram, Normal, QQ plot, and Write diagnostic tests.
- Click OK.

Figure 3.6.20 shows the standardized residuals and their correlogram, density function, and normal probability plot as depicted by the STAMP graphics window in GiveWin.

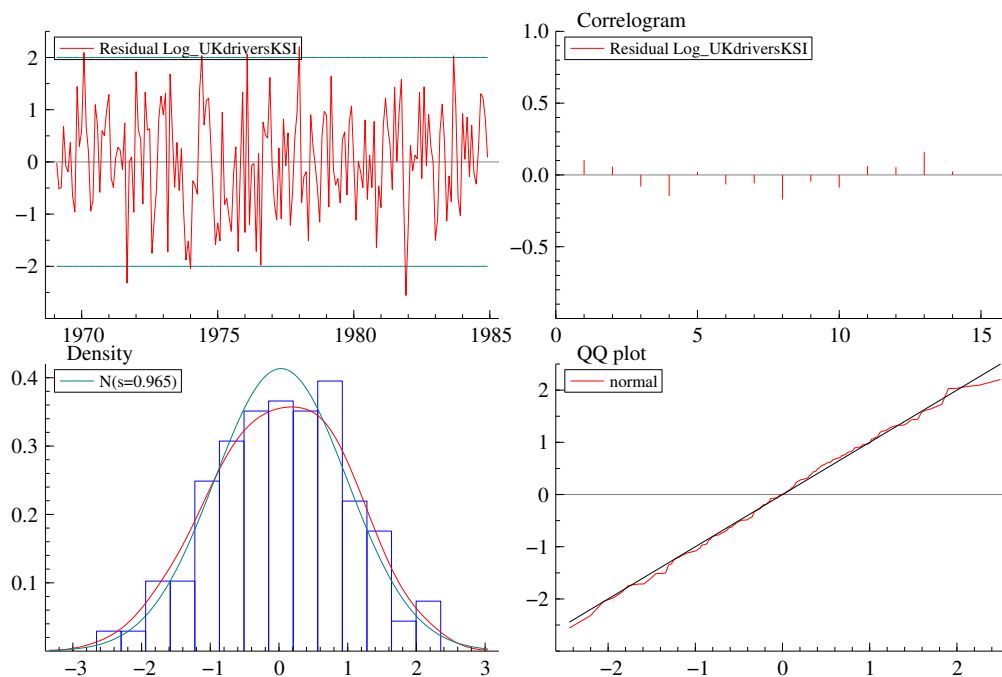


Figure 3.6.20: Residuals and residual tests for the stochastic level and deterministic seasonal model with intervention and explanatory variable applied to the log of UK drivers KSI.

When we compare the residuals and the autocorrelations for lags 1 to 14 of this model including intervention and explanatory variable with the residuals and autocorrelations of the model without explanatory variable (see the top left graphs of Figures 3.6.16 and 3.6.20), we see almost no differences. However, from the bottom graphs of Figures 3.6.16 and 3.6.20 we can conclude that the distribution of the residuals is more close to the normal distribution. This conclusion is confirmed by the Doornik-Hansen statistic, whose value is smaller for this model (1.90, see Table 3.6.8) than for the model without explanatory variable (2.40, see Table 3.6.7).

- Use the menu <File, Save> or <Ctrl+S> to save the graphs and minimize the STAMP Graphics window.

In the Results window of GiveWin, the following residual test results can be found (only part of the results are printed below):

Goodness-of-fit results for Residual Log_UKdriversKSI
 Information criterion of Akaike AIC -5.175474
 ... of Schwartz (Bayes) BIC -4.920982

Serial correlation statistics for Residual Log_UKdriversKSI.

Lag	dF	SerCorr	BoxLjung	ProbChi2 (dF)
1	0	0.1028		
2	0	0.0582		
3	1	-0.0805	3.9804	[0.0460]
4	2	-0.1456	8.1591	[0.0169]
5	3	0.0203	8.2410	[0.0413]
6	4	-0.0653	9.0899	[0.0589]
7	5	-0.0583	9.7707	[0.0820]
8	6	-0.1682	15.4709	[0.0169]
9	7	-0.0467	15.9130	[0.0259]
10	8	-0.0856	17.4054	[0.0262]
11	9	0.0598	18.1372	[0.0336]
12	10	0.0526	18.7065	[0.0442]
13	11	0.1513	23.4503	[0.0153]
14	12	0.0237	23.5674	[0.0233]

The addition of the petrol price variable to the model has improved the goodness-of-fit: the AIC has decreased (from -5.15 to -5.17) as well as the BIC (from -4.91 to -4.92).

Step 6: Test of auxiliary residuals

- Go to the STAMP window again and choose <Test, Auxiliary residuals graphics...>.
- In the Auxiliary residuals graphics window select Irregular, Level residual, Index plot, Density, Histogram, Normal, Write normality tests, and Write values exceeding (3.5).
- Click OK.

The STAMP graphics window in GiveWin displays the auxiliary residuals of the irregular and of the level component and their density function and normal probability plot: see Figure 3.6.21.

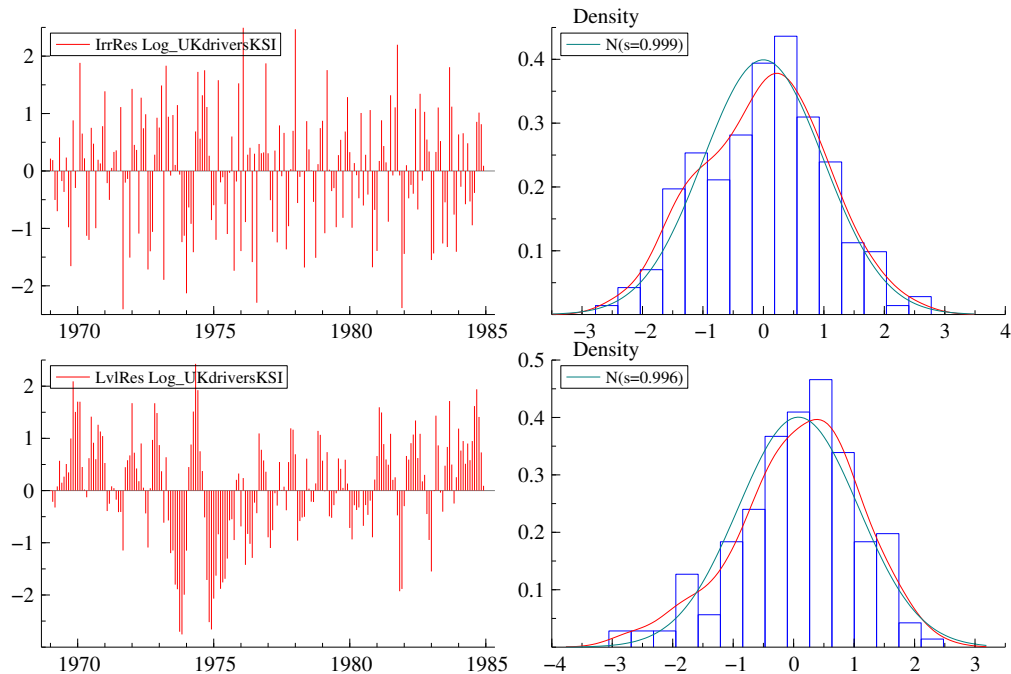


Figure 3.6.21: Auxiliary residuals and corresponding tests for the stochastic level deterministic seasonal model with intervention and explanatory variable applied to the log of UK drivers KSI.

The following output describes the auxiliary residual test results for normality as can be found in the Results window of GiveWin.

```

Normality test for IrrRes Log_UKdriversKSI
Sample Size      192
Mean              -0.000458
Std.Devn.         0.999195
Skewness          -0.063404
Excess Kurtosis   -0.402381
Minimum          -2.408254
Maximum           2.494782
Skewness Chi^2(1)  0.12864 [0.7198]
Kurtosis Chi^2(1)  1.2953 [0.2551]
Normal-BS Chi^2(2) 1.4239 [0.4907]
Normal-DH Chi^2(2) 1.055 [0.5901]

Normality test for LvlRes Log_UKdriversKSI
Sample Size      192
Mean              0.080397
Std.Devn.         0.995941
Skewness          -0.481094
Excess Kurtosis    0.144688
Minimum          -2.756100
Maximum           2.427820
Skewness Chi^2(1)  7.4065 [0.0065]
Kurtosis Chi^2(1)  0.16748 [0.6824]
Normal-BS Chi^2(2)  7.5739 [0.0227]
Normal-DH Chi^2(2)  8.305 [0.0157]

```

Both Figure 3.6.21 and the auxiliary residual tests demonstrate that the auxiliary residuals of the irregular component satisfy the assumption of normality, whereas those of the level component do not satisfy this assumption. This was also the case in the model with intervention but without the petrol price explanatory variable. From the bottom left graph, we see that in 1973 and 1974 there are quite a number of large negative auxiliary residuals of the level component. As

mentioned in the previous section, this means that we must be careful with the interpretation of the test statistics for the auxiliary residuals corresponding to the level component.

Step 7: Conclusion of analysis

In this section, we extended the stochastic level deterministic seasonal model with the seat belt intervention and the log of petrol price as explanatory variable and applied this model to the log of the monthly UK drivers KSI from January 1969 to December 1984. The residuals obtained with the analysis of this model satisfy all the model assumptions of independence, homoscedasticity, and normality. However, the auxiliary residuals of the level component do not satisfy the assumption of normality. Therefore, we must be careful with the interpretation of test statistics with respect to structural level breaks.

According to this analysis, the introduction of the seat belt law resulted in a $100 \cdot (e^{-0.23757} - 1) = -21.1\%$ change in the number of UK drivers KSI, while a 1% rise in petrol price was associated with a 0.28% reduction in the number of UK drivers KSI.

Step 8: Forecasting

Since the stochastic level deterministic seasonal model with the seat belt intervention and the log of petrol price as explanatory variable provides an appropriate description and explanation of the log of the monthly UK drivers KSI series, as a final step in the analysis we will make forecasts for this series. By performing an anti-log analysis, the forecasts will also be re-expressed in terms of the original count data.

As in Section 3.6.6 of the Methodology report, we will make in-sample forecasts so as to validate the model. To this end, the stochastic level and deterministic seasonal model with the seat belt intervention and the log of petrol price as explanatory variable is fitted to the log of the monthly number of UK drivers KSI, but now only for the period January 1969 to June 1984. So, we now do not include the last six observations (i.e., the last half year) of the UK drivers KSI series in the analysis. The results of the analysis without the months July 1984 through December 1984 are very similar to the results presented earlier in this section for the complete series.

- Repeat steps 1 and 2 of this section.
- Then, go to step 3: in the Estimate Model window, select Maximum Likelihood.
- Choose Less forecasts 6.
- Click OK.
- Go through steps 4 to 7 to check the model assumptions.

Next, we use these results to obtain forecasts for July 1984 through December 1984 and compare the forecasts with the observed number of UK drivers KSI

for this period. The actual values of the log of the petrol price are used for making the forecasts.

- Go to the STAMP window and choose <Test, Forecasting...>.
- In the Forecasting window select 6 as the number of forecasts, PlusXs, Seasonal, Use available database Xs, and Write forecasts Y.
- Click OK.

The STAMP graphics window in GiveWin displays the log of the UK drivers KSI series from January 1976 until June 1984 extended with the six-months forecasts including their 70% confidence interval (plus and minus one estimated standard deviation) in the top figure; the log of the UK drivers KSI and the extrapolated trend are shown in the middle figure, and the extrapolated seasonal is displayed in the bottom figure, see Figure 3.6.22.

The GiveWin Results window contains the forecasted values for July 1984 to December 1984, their root mean square errors, and the lower and upper bounds of the 70% confidence intervals.

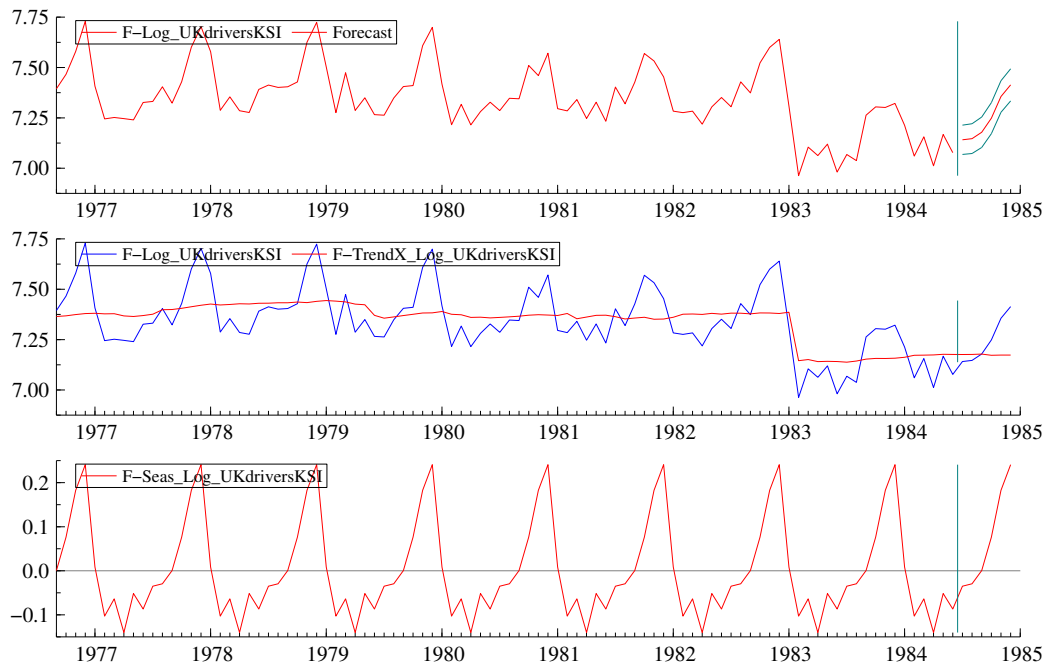


Figure 3.6.22: Six-months forecasts (July-December 1984) of the stochastic level and deterministic seasonal model with intervention and explanatory variable applied to the log of UK drivers KSI, January 1969 – June 1984.

For the forecast, the fixed seasonal is set through to the forecast range (bottom graph). The stochastic level part of the trend is set through as a horizontal line, whereas the explained part of the trend is extrapolated by multiplying the log of the petrol price with the corresponding regression coefficient. The total forecast is obtained by adding up the forecasts for the trend and the seasonal.

- Use the menu <File, Save> or <Ctrl+S> to save these graphs, e.g. as an Encapsulated Postscript file (*.eps). Minimize the STAMP Graphics window.

In Figure 3.6.23, the observed number of UK drivers KSI (not their logs) are compared with the forecasts (also in absolute numbers). The figure also displays the 90% confidence limits, determined as the forecasted values plus and minus 1.64 times the root mean square error (see also Section 3.6.6 of the Methodology report). As can be seen in Figure 3.6.23 the observed numbers of UK drivers KSI for July-December 1984 are all located within the 90% confidence limits of the forecasts. This is a good sign because it means that none of the observed values significantly deviates from the forecasts in this period.

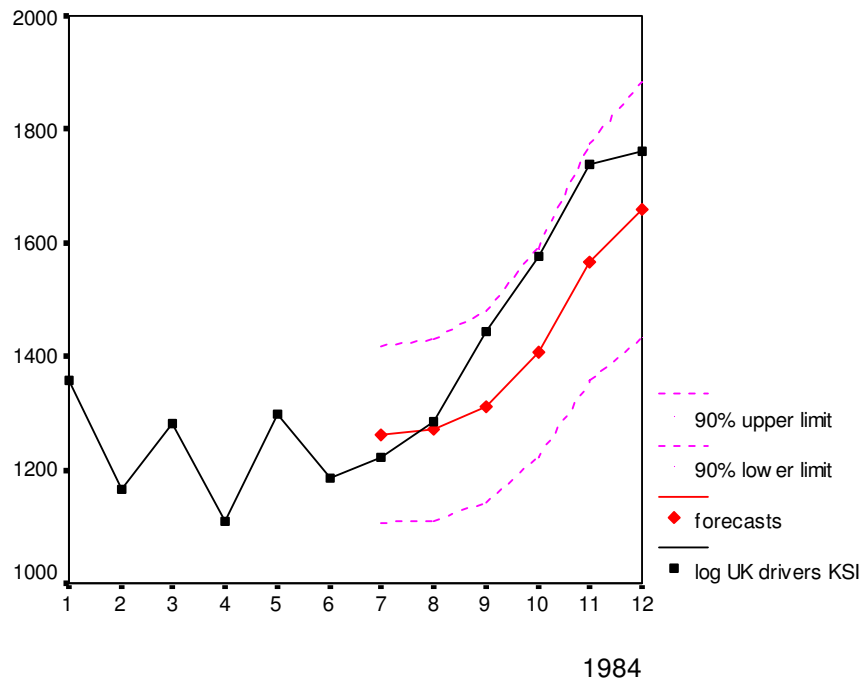


Figure 3.6.23: Observed number of UK drivers KSI in 1984, forecasts of the stochastic level and deterministic seasonal model with intervention and explanatory variable, and 90% confidence interval.

To display the forecasts in terms of absolute numbers in STAMP, we apply an anti-log analysis:

- Again go to the STAMP window and choose <Test, Forecasting...>.
- In the Forecasting window select 6 as the number of forecasts, PlusXs, Modified anti-log analysis, Use available database Xs, and Write forecasts Y.
- Click OK.

The STAMP graphics window in GiveWin displays the original observations from January 1976 until June 1983 extended with the six-months forecasts with 70% confidence interval (plus and minus one estimated standard deviation) in

the top figure and the original observations and the extrapolated trend in the bottom figure: see Figure 3.6.24.

In the Results window of GiveWin the following forecasts for the original observed time series have been added:

Period	Forecast	R.m.s.e.	- Rmse	+ Rmse
1984. 7	1262.9	94.991	1167.9	1357.9
1984. 8	1270.3	97.608	1172.7	1367.9
1984. 9	1311.2	102.84	1208.3	1414.0
1984.10	1406.1	112.49	1293.7	1518.6
1984.11	1565.8	127.67	1438.1	1693.5
1984.12	1659.1	137.79	1521.3	1796.9

The list of forecast results gives for each time point the value of the forecast, its standard error, and the lower and upper limit of the 70% confidence interval.

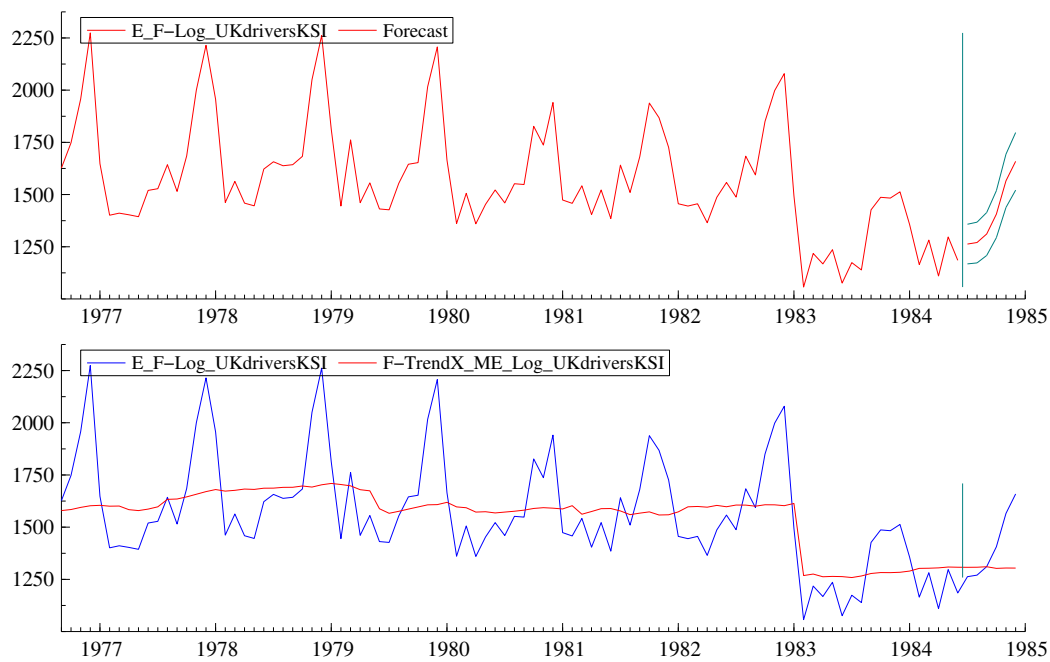


Figure 3.6.24: Anti-logged six-months forecasts (July-December 1984) of the stochastic level and deterministic seasonal model with intervention and explanatory variable applied to the log of UK drivers KSI, January 1969 – June 1984.

3.6.7 Conclusion state space models

This chapter showed examples of the application of state space analysis to road safety data. State space analysis was used to describe the development of road unsafety, to explain part of this development by adding interventions or explanatory variables, and to make forecasts of road unsafety.

For the analysis of a new road safety dataset, we recommend to first analyse the data on fatalities, casualties, drivers KSI, etc. by going through analysis steps 1 to 7 as presented in this chapter. The most efficient way to find the best fitting model is to start from a stochastic level and a stochastic slope component. If there is possible seasonality, e.g. in the case of quarterly, monthly, or weekly data, a dummy or trigonometric seasonal should be included in the model. Then, the estimated variances of disturbances indicate whether the components should be treated stochastically or deterministically. Next, if the final state value of a deterministic component is not significantly different from zero, then the component can be removed from the model.

Always carefully check whether the residuals of the model satisfy the model assumptions of serial independence, homoscedasticity, and normality. The statistics as presented in Table 3.6.1, for example, can be used for that purpose. However, one should not only rely on these statistical tests, but also make use of graphical tests, like the plot of the standardised residuals, the correlogram, the normal density diagram, and the normal probability plot.

The auxiliary residuals of the irregular component are useful to detect outlier observations and the auxiliary residuals of the level, slope, and seasonal components can be employed to find structural breaks in the respective level, slope, and seasonal. The auxiliary residuals should be tested for normality, by using statistical tests (e.g., the Doornik-Hansen statistic) and graphical output as the auxiliary residual plot, the normal density diagram, and the normal probability plot. If the auxiliary residuals of a component are not normally distributed, then this can be interpreted as a warning that the tests of outliers or structural breaks should be interpreted with care.

With state space analysis, forecasts can be made for short and long periods ahead. In doing this, one should be attentive to the fact that the forecasts totally rely on setting through the dynamics of the time series from the past to the future, at least if no additional information regarding the future development of explanatory variables is added. A very useful property of state space analysis is that it produces confidence intervals for the forecasts, which provide insight in the range of possible future values. In general, these ranges become impressively wide when the time span between the forecast and the last observation grows. This prevents the user of state space analysis from drawing too firm conclusions and helps to put the forecast results into perspective.

Chapter 4 - Conclusion

The present document constitutes the practical part of the best practice advice for the analysis of complex data structures by Work Package 7 of the SafetyNet project. This manual is intimately linked to D7.4, “Multilevel modelling and time series analysis in traffic research – A methodology”. While in the methodology report the emphasis is on theoretical background information, the manual gives practical instructions for the conduction of multilevel and time series analyses on the basis of user friendly software. Like the methodology report, this manual is divided into two main chapters each dedicated to one broad family of analyses, multilevel modelling and time series analysis.

In *Chapter 2*, first an overview is given over the various multilevel models presented in this deliverable. Moreover the software used in this deliverable and other available multilevel software is discussed (Section 2.1). Multilevel modelling is then introduced with the simplest case (Section 2.2): A linear variable (speed of a car) was predicted by another linear variable (length). In that example the individual cars constituted the first level, a second level was given by the road sites at which the speed was measured. It was also investigated whether the regions had an effect on the speed measurements (i.e. whether there was a third level), however, this was not the case.

Often in road safety research, variables are not linear. Therefore the linear models are put into the framework of the Generalised Linear Model that allows to model variables from other distributions, for example binary response variables, as well. As an example of a binary response variable, the data from an alcohol study are presented, indicating whether a driver had a BAC above the legal limit or not (Section 2.3.2). The individual drivers were the first level, and again, the road site at which the alcohol level was tested constituted the second level and it was demonstrated that first-level as well as second-level variables could explain some of the variation in drink-driving. This data could also be analysed as multinomial-response data with 3 categories ($BAC < .05$, $.05 < BAC < .08$, and $BAC > .08$) as demonstrated in Section 2.3.3. In this section it is demonstrated how this type of data can be considered as multivariate response structures and can be implemented in a multilevel model.

The number of accidents can be assumed to be Poisson-distributed and in Section 2.3.4 it is demonstrated how the numbers of fatal accidents can be predicted by law-enforcement measures. The first level in this analysis was given by the counties in which the number of fatalities had been established and it was shown that this number varied across regions (the second level) and that the effect that alcohol and speed controls had also varied across regions.

Often in road safety research, the same individual or unit is measured a number of times subsequently. Multilevel modelling can be used for such longitudinal data. This was demonstrated in Section 2.5 with a simulated data set of driving scores taken over 6 consecutive measurements. In this case, the individual measurements constituted the first level and the individuals from whom the

measurements were taken formed the second level. It was demonstrated how this technique allows the inclusion of predictors at the measurement level (i.e. number of km driven at time of measurement) as well as at the individual level (e.g. age at acquirement of driver's licence).

Similarly to a structure of repeated measurements, multilevel models can also serve to analyse multiple responses or measurements from the same individual unit. This was demonstrated in Section 2.4, where accident numbers and fatality numbers were predicted simultaneously by law-enforcement measures in a multivariate model. The multivariate response structure was set up, so that the number of fatalities and the number of accidents jointly formed the data vector. In this case the indicator for the lowest level does not correspond to the unit of measurement (the counties) but to an indicator specifying the type of response given (number of accidents or number of fatalities).

The topic of Chapter 3 is the analysis of road safety time series data. In the introduction (Section 3.1) a short overview is given over the different methods as well as the software used in this manual and other available software.

The first time series approach discussed is the well known linear regression approach. Although technically not a specific time series analysis method, due to the fact that it is well known, it appeared this method is suitable to demonstrate the key issues with time series data (in road safety) in an environment familiar to many readers. It is demonstrated that the ordinary linear regression model is not suitable for the number of fatalities in Austria. In fact, both the heteroscedasticity and the independence assumption have to be questioned.

Next, the ARMA-type section demonstrates how the examples discussed in the methodology report can be fit using SPSS. Details about the Norway fatalities dataset, UK-KSI drivers and the French fatalities examples are given. Finally, in the state space section, the model properties are discussed using examples based on Norwegian and Finnish fatality data. Finally, more extensive models are developed based on the UK-KSI data.

To conclude, a wide range of examples of road safety data analyses was presented in a very detailed way, allowing the reader to understand the necessary decisions about distributional assumptions, variables included, and estimation methods chosen. The data files used are included so that each action and output can be traced by the reader. The interpretations are directly linked to the output, so that they can serve as examples enabling the reader to interpret the output for the same type of analysis with a different set of data.

Together with the methodology report, D7.4, this manual forms the best practice for the analysis of complex data structures. The reader should have understood the necessity to check the assumptions underlying the statistical analyses used, and if necessary to use methods like multilevel modelling and time series analyses that explicitly represent complex data structures and thus allow

researchers conduct valid analyses and gain more information about the structure itself.

References

- Box, G. E. P. & Jenkins, G.M. (1976). *Time series analysis: forecasting and control*. Revised Edition. Oakland, CA: Holden-Day.
- Brockwell P.J., Davis R.A. (1998) *Introduction to time series and forecasting*, Springer Verlag
- Bryk, A. S., S. W. Raudenbush, and R. Congdon (1996). *Hierarchical Linear and nonlinear modeling with the HLM/2L and HLM/3L Programs*. Chicago: Scientific Software International.
- Centre for Multilevel Modelling, www.mlwin.com. Last accessed on 27.02.2007
- Durbin, J. and Koopman, S.J. (2001), *Time series analysis by state space methods*. Oxford University Press.
- Enders W. *Applied econometric time series*, Wiley, 1995
- The European Road Safety Observatory. www.erso.eu/safetynet.htm. Last accessed on 27.02.2007.
- Gourieroux, C. & Monfort, A. (1990). Séries temporelles et modèles dynamiques. *Economica*, 1990.
- Hallahan, C. (2003), "STAMP 6.0", *International Journal of Forecasting*, Vol. 19, Issue 2, pp. 319-325.
- Harvey, A.C. (1989), *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.
- Hedeker, D. & Gibbons, R. D. (1996a). MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, 49:157-176.
- Hedeker, D. & Gibbons, R. D. (1996b). MIXREG: a computer program for mixed-effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine*, 49:229-252.
- HLM - Hierarchical Linear and Nonlinear Modeling. Scientific Software International, www.ssicentral.com/hlm/index.html. Last accessed on 27.02.2007.
- Judge, G. and Ninomiya, Y. (2000), " STAMP 6.0: Structural Time Series Analyser, Modeller and Predictor", *CHEER*, Vol. 14, Issue 1 (virtual edition).
- Littell, R.C., Miliken, G.A., Stroup, W.W., & Wolfinger, R.D. (1996). SAS system for mixed models. Cary: NC: SAS Institute, Inc.

LISREL for Windows. Scientific Software International.
www.ssicentral.com/lisrel/index.html. Last accessed on 27.02.2007

Koopman, S. J., Harvey, A. C., Doornik, J. A. and Shephard, N. (2000),
STAMP: Structural Time Series Analyser Modeller and Predictor, Timberlake
Consultants.

MIX project. <http://tigger.uic.edu/~hedeker/mix.html>. Last accessed on
27.02.2007.

The R project. (<http://cran.r-project.org>). Last accessed on 13 July, 2006

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.

S-Plus 7 – Delivering the power of predicative analytics across the enterprises
(www.insightful.com/products/splus/default.asp). Last accessed on 27.02.2007

Teyssière, G. (2005), "Structural time series modelling with STAMP 6.02",
Journal of Applied Econometrics, Vol. 20, Issue 4, pp. 571-577.

WINBUGS, The bugs project. www.mrc-bsu.cam.ac.uk/bugs/. Last accessed on
27.02.2007.

Yaffee, R. (2003), "Structural Time Series Modeling with SAS Proc UCM and
STAMP", *Social Science, Statistics & Mapping*, Spring 2003 (virtual edition).