UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

School of Mathematics

Examining the effect of daylight on road accidents and investigating a state space time series approach to modelling zero inflated count data

by

James Edward Dartnall

Thesis for the degree of Doctor of Philosophy

May 2007

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

SCHOOL OF MATHEMATICS

Doctor of Philosophy

EXAMINING THE EFFECT OF DAYLIGHT ON ROAD ACCIDENTS AND INVESTIGATING A STATE SPACE TIME SERIES APPROACH TO MODELLING ZERO INFLATED COUNT DATA

By James Dartnall

In this thesis two aspects of the modelling of road accident count data are investigated in detail. Under the first investigation the effect of daylight on road accidents is considered. Here, daylight is established as a significant cause of car occupant casualties in both Scotland and Southwest England using linear and log-linear regression models. It is also shown that there is a noticeable difference in the level of daylight during morning rush hour in December and January between Scotland and Southwest England due to the difference in latitude between two regions. Ad hoc methodology is then introduced to investigate the possibility that the difference in the level of daylight during morning rush hour will result in a significant difference in the numbers of car occupant casualties between the two regions during December and January.

The second investigation considers the use of a conditional Bernoulli truncated Poisson state space time series model for modelling zero inflated count data. Although it is technically complex, its appeal is likely to be broader than the daylight investigation as the methods presented here offer useful insight into the modelling of any zero inflated time series count data. The conditional Bernoulli truncated Poisson model has been used before to model zero inflated data, but it has not been used on time series data and has not been put into state space form. Difficult issues are raised by applying the conditional model to time series data and various methods are introduced and compared to overcome these problems.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank first and foremost my supervisor, Alan Welsh, who has steered the research in this thesis into interesting avenues and has been my statistical guide, always providing help and advice when needed. I would also like to thank my advisor, Phil Prescott, who read through this thesis giving confirmation of its readiness for submission and then made all necessary arrangements regarding its assessment.

For the data used in this thesis I am very grateful to Jeremy Broughton of the Transport Research Laboratory who provided most of the data upon which the analysis in this thesis is based. I am also grateful to the Department of Trade and Industry and the MET Office for the fuel deliveries data and weather data respectively.

For help with technical questions I am indebted to the members of the S-news newsgroup who helped me out on several occasions with questions of S-Plus syntax, particularly Eric Zivot of the University of Washington who provided all answers to my queries relating to the use of state space models in S+FinMetrics. I would also like to thank my good friend and former flat-mate, Adam Brentnall, for his help getting me started with VBA.

On a personal note, I would like to thank my mother and father for their love and support; their belief in my ability to complete this thesis has never wavered, unlike my own. Last but not least, I would like to thank my beloved cat, Minnie; without her constant companionship I would never have been able to make it through the dark days when progress was slow and despair was close at hand.

Chapter 1

Introduction

Modelling road accident data is of great interest to many parties such as road safety groups, government, motor car manufacturers and transport research organisations like the Transport Research Laboratory (TRL). In this thesis there are two investigations into the modelling of road accidents for which new methodology has been introduced: one into the effect of daylight on road accidents and the other into appropriate state space time series models for zero inflated count data. However, the new methodology has not been conceived in isolation and is very much related to existing modelling techniques. As such, together with the new methodology a great deal of existing methodology is also presented in this thesis.

This introduction is divided into three parts: in the first part the data sets used in the analysis are briefly described, in the second section an outline is given of each chapter and the particular technique investigated therein, and in the third section the difficulties encountered during the compilation of the data and completion of the thesis are detailed.

1.1 The data sets used

Most of the data used in this thesis has been provided by the Transport Research Laboratory. The data provided by the TRL is road accident data of various types including car occupant accidents, pedestrian accidents, accidents in snowy weather conditions and many more. Each accident type is given on three levels of severity: fatal, serious and slight. Also, each accident type at each level of severity is given for Scotland, Wales and all of the nine government office regions of England, these being: Northern England, Northwest England, Yorkshire & Humberside, West Midlands, East Midlands, Eastern England, Southwest England, Southeast England and Greater London. All the road accident series are monthly aggregated time series from January 1979 to December 2000.

The rest of the data used in this thesis has been obtained from various sources and has been primarily used as explanatory variables in chapter 2 and elsewhere in the thesis. In detail the variables are:

- National car traffic (Billion vehicle km's travelled); this is the estimated volume of car traffic on roads in Great Britain obtained from TRL records and originally extracted from DTLR (2000) and previous years reports. The data is quarterly adjusted to monthly, ranging from January 1987 to December 2001 and adjusted using the monthly averages over the five year period from 1996 to 2000.
- Total rainfall (mm); monthly data for each of the regions of Great Britain ranging from January 1987 to December 2000, provided by the UK MET Office.
- Monthly average maximum daily temperature (Celsius); monthly data for each of the regions of Great Britain from January 1987 to December 2000, provided by the UK MET Office.
- Monthly average minimum daily temperature (Celsius); monthly data for each of the regions of Great Britain from January 1987 to December 2000, provided by the UK MET Office.
- Cloud cover (oktas); monthly data for each of the regions of Great Britain from January 1987 to December 2000, provided by the UK MET Office. Here 1 okta = 1/8 cloud cover and 8 oktas is total cloud cover.

- National inland deliveries of petroleum (tonnes); monthly data for the UK from January 1980 to December 2001, provided by the Department of Trade and Industry.
- National inland deliveries of diesel (tonnes); monthly data for the UK from January 1980 to December 2001, provided by the Department of Trade and Industry.
- Monthly average amount of daylight per day (hours); calculated using a Visual Basic for Applications (VBA) program translated from a Fortran program provided by the TRL.

Since the shortest range of the variables listed above is January 1987 to December 2000, this time span has been used for most of the analysis throughout the course of the thesis. It provides 14 years, totaling 168 observations, of data which provides an adequate time span to perform regression or time series analysis.

The explanatory variables above have been included as it is thought that they will have an impact on road accidents; indeed, some of them need no explanation. Car traffic makes up the vast majority of traffic on British roads and so the car traffic variable is a direct measure of exposure to road accidents, i.e., the more kilometers travelled, the more accidents there are likely to be. The national petrol and diesel delivery variables are to some respect surrogate measures of exposure. It is more common to use fuel sales as a surrogate for exposure, but since these data were not available, fuel deliveries were used instead. Measures of exposure are generally very difficult to obtain and most exposure data is gathered through surveys. For this reason it has not been possible to obtain regional exposure data, although it is hoped that the national level of traffic will provide a reasonable approximation at the regional level.

Common sense might suggest that increased rainfall would have a similar impact on road accidents as increased traffic, i.e., it would lead to an increase in road accidents. This would be because vehicles have less control when road conditions are wet, and wet weather can affect visibility. However, it is not necessarily the case that the effect of higher rainfall will result in more road accidents. Studies have shown that adverse weather conditions can result in more cautious driving and therefore a reduction in accidents; they may also result in people not making journeys unless they are competent drivers, thereby reducing exposure and increasing the proportion of good drivers on the road at once. Fridstrom et al. (1995) found a reduction in overall road accidents in snowy weather due to this sort of effect.

The temperature variables are highly correlated with one another and are also, to a certain extent, correlated with exposure since it is known that there is generally more traffic on the roads when weather is good. It is likely that should one variable be found significant the other is unlikely to also be significant, and this may be the case for the exposure variables too. It is possible that the minimum daily temperature variable will act as a surrogate for adverse weather and road conditions such as frost or snow. Again, this could go two ways: either frost and snow will cause drivers to more easily lose control of their vehicles, or drivers will act over cautiously to compensate and may not even drive at all.

Daylight and cloud cover have both been included as measures of the amount of daylight received. Again, these could go two ways: either increased light levels may mean better visibility and thus fewer accidents or an increase in light levels will mean more drivers travel and thus more accidents due to increased exposure. It is hoped that the presence of the exposure variables may counter this effect so that the true effect of daylight is revealed.

1.2 Thesis format

A variety of methodologies are considered in this thesis. We begin each chapter by giving an outline of a technique which can be used to model count data, then give applications of that method. Throughout the course of the thesis, the computer program S-Plus together with the S+FinMetrics add-on have been used to derive the solutions given to the practical examples.

The thesis begins with a summary of linear and log-linear regression modelling in Chapter 2. Both types of regression model are applied to Scottish and Southwest English car occupant accident data and conclusions are drawn. Along with the familiar residual plots used for examining the residuals from regression models, the correlogram is introduced as a means for checking the residuals for signs of autocorrelation; that is, serial correlation of the residuals. However, autocorrelation and dependence are not discussed in detail until the following chapter.

In chapter 3 we consider ARIMA time series models which can be used in the analysis of residuals from regression models or alternatively can be used in their own right to analyse time series and make predictions. In this chapter an ARIMA model is fitted to the Scottish car occupant deaths and injuries data used in chapter 2, with the aid of the correlogram.

The focus of chapter 4 is on the development of methodology to determine the effect of daylight on road accidents. The central idea of the chapter is to try to use latitude to examine the effect of daylight on road accidents; it was the difficulty in obtaining suitable explanatory variables for the regression analysis in chapter 2 which initially inspired this idea. Aspects of regression and ARIMA modeling are included in the analysis, and the same Scottish and Southwest English car occupant accident series used in chapter 2 are used to illustrate the methods.

In chapter 5 and 6 we move on to considering state space time series models. In chapter 5 Gaussian state space models are examined and an example is given of their successful application in the seat belt study of Harvey and Durbin (1986). A Gaussian state space model is then applied to the Scottish deaths and injuries data and comparisons are drawn between this model and the linear regression model for the same data in chapter 2.

In chapter 6 we consider non-Gaussian state space models, specifically the Poisson model which is the state space time series analogy of the log-linear regression model. This model is applied to the Scottish fatalities data and comparisons are drawn with the log-linear model for the same data.

New methodology concerning the modelling of zero inflated counts, that is data with an unexpectedly large number of zero counts, is introduced in chapter 7. Here, we consider putting a zero inflated count model - the conditional Bernoulli truncated Poisson distribution - into state space form to analyse zero inflated count data. While this chapter is entirely theoretical, applications of the results are presented in chapter 8.

1.3 Difficulties encountered in the completion of this thesis

The first difficulty encountered in this thesis was the form in which the data from the TRL arrived. Data for each accident type was grouped together in individual year groups where months were in columns and similar accident types were in rows. To perform statistical analysis on the data in this form was not possible. After starting manually to cut and paste the data into single columns for each accident type, it became evident that this would take too long as there were, in all, 1056 individual data series to sort out from one another. Learning Visual Basic for Applications (VBA) to write routines to collate the data for each series into single columns rectified this problem. Once in column form, each series could easily be read into S-Plus where subsequent analysis could be carried out.

The most common difficulty faced in the early stages of the thesis was the lack of explanatory variables for use in the analysis. When performing linear or log-linear regression, such as that in chapter 2, it is necessary to use several explanatory variables to obtain a reasonable fit as there is only so much that fitted ploynomial terms can do. Many of the examples involving road accidents used in this thesis include explanatory variables, the majority of which have not been easy to obtain.

To start with, the factor with possibly the most notable impact on the incidence of road accidents is the volume of traffic on the roads. This, as we noted in chapter 2, has been found to be significant from previous road accident studies such as Fridstrom et al. (1995) and Fridstrom and Ingerbrigtsen (1991); however, it is surprisingly difficult to come by. After a good deal of trying, it was established that we would be able to obtain no more than national quarterly data, which we would have to use as a substitute for regional monthly data. The Department for Transport and other organisations were unable or unwilling to supply any data which was not already available on their website. The national quarterly data on the DfT website, however, only covered recent years, and was neither regional nor monthly nor from 1987. The TRL were eventually able to supply the rest of the data back to 1987.

Another major factor in determining the occurrence of road accidents is weather. Weather data, again, is unavailable to the general public in all but the most basic highly aggregated form which is of no use to a sensible statistical analysis. Eventually, due to lack of success obtaining this data from other sources, we were led to purchase the information from the MET Office for a considerable fee. The form that the weather data came in also needed to be altered to column format in VBA for use in S-Plus.

Finally, the daylight data, which is vital to much of the analysis carried out in chapter 2 and 4, was also only obtained after some time and effort. The TRL provided a Fortran program which would calculate the times of sunrise and sunset for given a time, latitude and longitude. Without the software to run Fortran programs, I converted it into VBA and adjusted it to produce monthly output for the regions of the UK we were studying.

Our experience in trying to obtain data led us to realise that data is an expensive commodity and not free of charge even to non-profit making organisations and individuals like universities or students. It was partly this that led us to more actively pursue alternative ways of modelling road accident data which would give sensible solutions, even without the use of explanatory variables. These alternative ways of modelling count data led to ARIMA time series models, state space time series models and the ad-hoc methodology of chapter 4.

However, even the use of these models was not without difficulty; S-Plus has no feature for the analysis of state space models. After unsuccessful attempts to obtain software such as Eviews, TSP, PcGive and Microfit for the analysis of state space models from the social sciences faculty, we were able to persuade computing services to make available FinMetrics (the financial time series analysis add on to S-Plus) for staff and research students. FinMetrics contains basic routines for state space analysis such as the Kalman filter and smoother and has the added advantage of being totally compatible with S-Plus so that no new language or syntax needs to be learnt. However, fitting a state space time series model is, as we learnt, certainly not a straightforward exercise like fitting a generalised linear model. The various functions that are used come piecemeal and need to be put together coherently; they do not come so that a state space model may be fit using one or two lines of code. Fitting non-Gaussian state space models is even more involved. However, due to having made contacts on the S-Plus news list, we were able to proceed with the fitting of the Poisson and conditional truncated Bernoulli Poisson models by adapting some software we were given for the fitting of the stochastic volatility model (Zivot et al., 2003).

It is somewhat telling of the state of advancement in software for the analysis of state space time series models, that we found after having fitted state space models that there was no way to obtain estimates of standard errors on the state error variance parameters (hyperparameters) in these models. Software for the calculation of these standard errors became available in the next upgrade to FinMetrics which came out in April 2005, but the university was unable to obtain this until shortly before Christmas that year.

In the spring of 2006 a long standing problem that had occurred when trying to analyse non-Gaussian state space models was finally solved. The problem had been that it was not possible to obtain sensible state estimates for non-Gaussian state space models. However, it had been possible to obtain the overall fitted mean for the model as well as the state error variance estimates for all states along with their associated standard errors. Therefore, although a hindrance, for much of the compilation of this thesis this problem had not seriously impaired progress on the analysis of non-Gaussian state space models since state error variances are most important for deciding upon the inclusion or exclusion of states in a state space model. But difficulties had arisen when analysing non-Gaussian state space models with explanatory variables since these models include constant states which are the coefficients of the explanatory variables. These states cannot be judged using their state error variance since they are constant and thus have no state error variance. Since these states are the coefficients of the explanatory variables, it is crucial that sensible estimates are obtained for them, as well as their standard errors, so that conclusions can be drawn about the effect of the explanatory variables they apply to. As it had not been possible to obtain sensible state estimates, it was therefore not possible to obtain sensible estimates for the coefficients. In the spring of 2006 I eventually isolated the cause of the problem to the SimSmoDraw function, which is a function in the FinMetrics software package written to implement the simulation procedure in §6.3. It took a long time to find out that it was the SimSmoDraw function which was at fault since, generally, all possible shortcomings in one's own functions should be examined and exhausted before considering that one of the functions provided in the software may be at fault. Once the error in the SimSmoDraw function had been identified, it was relatively straightforward to create a new amended function, SimSmoDrawJames, which could be used in place of SimSmoDraw to calculate simulated states and errors for the non-Gaussian log-likelihood and fitted model. Reference to the role the SimSmoDrawJames function plays in the analysis of non-Gaussian state space time series models can be found in appendix A, along with the other functions I wrote to calculate the non-Gaussian log-likelihood and fitted model.

Chapter 2

Linear and log-linear regression modelling

In this chapter we investigate the use of linear and log-linear regression models with explanatory variables to model the variation in road accident counts. The methodology for fitting these models is generally well known, but for continuity with the rest of the thesis, where less familiar methodologies are presented, a short summary of the linear and log-linear modelling techniques is presented in the first two sections. The methodology is presented in terms of univariate observations since in the examples presented at the end of the chapter, four univariate time series are examined, although comparisons are drawn between them when appropriate. The examples presented concern linear and log-linear models applied to road accident data from Scotland and Southwest England. These series shall be re-examined in chapter 4 and provide examples to illustrate the modelling techniques elsewhere in this thesis. The primary references for this chapter are Neter et al. (1996) for linear models, and McCullagh and Nelder (1989) for log-linear models.

2.1 Linear regression

2.1.1 Linear model framework

Suppose we have univariate data with n observations, $y_1, ..., y_n$. Under the linear model these are modelled as the sum of m explanatory variables, $x_{i1}, ..., x_{im}$, multiplied by suitable constants, β_j, plus a random error, ϵ_i, which is independently, identically and normally distributed, $N(0, \sigma^2)$:

$$y_i = \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i, \qquad i = 1, ..., n. \tag{2.1}$$

Most regression models will contain a constant; in the representation above the constant can be thought of as the first explanatory variable coefficient, β_1, where the first explanatory variable $x_{i1} = 1$, for $i = 1, .., n$. Putting (2.1) into matrix form, where $\boldsymbol{y} = (y_1 \ ... \ y_n)'$, $\boldsymbol{\beta} = (\beta_1 \ ... \ \beta_m)'$, $\boldsymbol{\epsilon} = (\epsilon_1 \ ... \ \epsilon_n)'$ and

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix},$$

model (2.1) can be summarised by

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}.$$

To estimate the coefficients, $\boldsymbol{\beta}$, we choose estimates, $\hat{\boldsymbol{\beta}}$, so that the model fit using these estimates, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, is as close as possible to the observations, $\boldsymbol{y}$. This is accomplished by minimising the sum of squares of the errors, ϵ_i:

$$\sum_{i=1}^{n} \{y_i - E(Y_i)\}^2 = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{m} \beta_j x_{ij} \right\}^2 = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}). \tag{2.2}$$

The sum of squares is minimised by differentiating with respect to β_j and setting to zero, which gives the estimate for $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}). \tag{2.3}$$

This estimate can also be derived by maximum likelihood estimation, where the log-likelihood of (2.1),

$$\log\{L(\boldsymbol{\beta}, \sigma^2)\} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2,$$

is maximised by differentiation with respect to β_j and set to zero in the same way as the sum of squares. In this way maximum likelihood estimation can also be used to derive an estimate of the variance, σ^2, which is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m}\hat{\beta}_j x_{ij}\right)^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

However, this is a biased estimate and so the unbiased estimate,

$$\hat{\sigma}^2 = \frac{1}{n-m}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m}\hat{\beta}_j x_{ij}\right)^2 = \frac{1}{n-m}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

is generally used instead.

Note that in the above equations the main part of the estimating equation for $\hat{\sigma}^2$ is called the deviance, D, and is given by

$$D = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m}\hat{\beta}_j x_{ij}\right)^2.$$

The deviance will be of relevance in the following subsection on model building and goodness-of-fit.

Finally, the degree of uncertainty surrounding the coefficient estimates, $\hat{\beta}_j$, is given by standard errors which are calculated from the coefficient variance-covariance matrix: $s.e.(\hat{\beta}_j) = \{\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}\}^{1/2}$, where $(\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}$

denotes the diagonal elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. The t-ratios (parameter estimates divided by corresponding standard errors) are judged for significance against a t distribution with $n - m$ degrees of freedom. Where a t-ratio lies in the tails of the t_{n-m} distribution, outside the middle 95% of the distribution say, then the corresponding coefficient estimate is judged as insignificant and the explanatory variable to which it applies is removed from the model. For the examples given in this chapter and subsequent chapters, n is so much greater than m that the t_{n-m} distribution can be approximated by the standard Gaussian distribution.

2.1.2 Model building and goodness-of-fit

For exploratory observational studies the task of choosing the best subset of explanatory variables to include in the final regression model is not straightforward. Generally, if there are m possible explanatory variables available in the study then 2^m possible combinations of those variables can be used as the explanatory variables in the model. Some of the explanatory variables, however, will not have a significant effect on the overall fit of the model to the response variable, so can be eliminated from the model.

F-tests are used to compare one model with a competing model and only nested models are compared to one another, i.e., the explanatory variables of one of the models must be a subset of the explanatory variables of the other. If the explanatory variables of model H_0 are a subset of the explanatory variables of model H_1, then the F-statistic is given by

$$F = \frac{(D_0 - D_1)/(m_1 - m_0)}{D_1/(n - m_1)},$$

where D_0 and D_1 are the deviances for models H_0 and H_1 respectively, and m_0 and m_1 are the numbers of explanatory variables in model H_0 and H_1 respectively. The F-statistic follows an $F_{m_1 - m_0, n-m_1}$ distribution, so the F-test works on the basis that model H_0 will be rejected in favour of H_1 when F is greater than a certain percentage point of the $F_{m_1 - m_0, n-m_1}$

distribution, typically the 95% point. For a large number of explanatory variables, model selection using the F-test can be time consuming and there may not be any one model which offers the best fit to the data. So, an analyst should generally pick the most plausible model and use other methods to assess the overall fit of the model such as the R^2 statistic and residual plots. Neter et al. (1996), chapters 8, 9 and 10, gives an exhaustive treatment of regression model building methodology which fits the F-test into the wider topic of the analysis of variance.

As far as residual plots are concerned, there are two which can be used to assess the fit of practically all linear models, they being the residuals vs fitted values plot, which should show a random scatter of residuals, and the normal probability plot, which should show that the residuals when plotted against their expected values under normality form a virtually straight diagonal line. For data which has been collected over a period of time, such as the data investigated in this thesis, another residual plot that should be examined is a simple plot of the residuals plotted against time. Time series data can be highly correlated and if explanatory variables do not account for all the variation in the data then some of this correlation may remain in the residuals and the simplest way to detect it is by viewing the residual series over time. To find out exactly what the pattern of correlation among the residuals is, if any, another plot called the *correlogram* may be used.

The definition of the correlogram is as follows: For any time series of data or residuals, $y_1, ..., y_n$, the k^{th} sample autocovariance coefficient is defined as

$$\hat{\gamma}_k = \frac{1}{n} \sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y}),$$

where k is an integer; this leads to the k^{th} sample autocorrelation coefficient

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}. \tag{2.4}$$

A plot of the $\hat{\rho}_k$ values against k gives the correlogram. The correlogram shows how strongly correlated residuals are with one another; a description of how to use the correlogram to tell exactly what the pattern of correlation is in a series is given in chapter 3. Also given in chapter 3 is a description of autoregressive and moving average terms which can be used to model dependencies in the data. Most importantly, for a correlogram which shows no sign of dependence structure, all autocorrelations, accept the first, do not differ significantly from zero and the first at $k = 0$ is exactly 1 since any observation is totally correlated with itself.

2.2 Generalised linear models

In the modelling of data it is common to use non-linear regression methods as well as linear methods, depending on the form of the data. Many non-linear methods used belong to a class of models known as generalised linear models (GLM's). In fact, linear models discussed in the previous section also belong to this class of models. The following subsections are a summary of the GLM framework, a more detailed description can be found in McCullagh and Nelder (1989), chapter 2.

2.2.1 GLM framework

In a GLM the observations, y_i, are independent and follow distributions which belong to the exponential family, where the exponential family density takes the form

$$p(y_i|\theta_i, \phi_i) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right\}.$$

Here, θ_i is the canonical parameter and ϕ_i is the nuisance, or dispersion, parameter. A variety of common distributions can be written in exponential form such as the Gaussian, Poisson and Binomial distributions. For linear models where the observations follow the Gaussian distribution we have $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, $a(\phi_i) = \phi_i = \sigma^2$ and $c(y_i, \phi_i) = -1/2\{y_i^2/\sigma^2 + \log(2\pi\sigma^2)\}$.

The systematic part of a generalised linear model, i.e., the explanatory variables, are expressed through a linear predictor, η_i, given by

$$\eta_i = \sum_{j=1}^{m} \beta_j x_{ij}.$$

The linear predictor is then linked to the mean, μ_i, through a link function, $g(\cdot)$, where

$$\eta_i = g(\mu_i).$$

The link function may be any monotonic differentiable function. For the linear model the link function is the identity, i.e., $\eta_i = \mu_i$. For models which belong to the exponential family, $b'(\theta_i) = \mu_i = E(Y_i)$, therefore $\theta_i = b'(\mu_i)$. When $g(\mu_i)$ is chosen so that $g(\mu_i) \equiv b'(\mu_i)$ then $g(\mu_i)$ is known as the canonical link function. The canonical link is desirable as it leads to the simple relationship $\theta_i = \eta_i$.

The coefficient estimates, $\hat{\beta}_j$, are derived using the maximum likelihood approach where the log-likelihood is given by

$$\log\{L(\boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{y})\} = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right\}. \tag{2.5}$$

Solutions may not be obtainable from maximising the log-likelihood analytically, so numerical maximisation procedures such as the Fisher scoring algorithm (McCullagh and Nelder (1989), p.42) are usually used to find the maximum likelihood estimates, $\hat{\beta}_j$, as well as their associated standard errors.

2.2.2 Model building and goodness-of-fit

Essentially, as with linear models, for GLM's we need to find the best set of the explanatory variables to reasonably accurately explain the variation within the observations and then assess the fit of the model once the right set of explanatory variables has been found. However, F-tests are not used to compare the suitability of different sets of explanatory variables; instead, log-likelihood ratios are used.

Log-likelihood ratio tests compare one model with a competing model and, as with linear models, only nested models are compared to one another. Note that a nested model in this context not only means that the explanatory variables of one of the models must be a subset of the explanatory variables of the other, but it also means that the models compared are from the same exponential family distribution and have the same link function. If the explanatory variables of model H_0 are a subset of the explanatory variables of model H_1 then the log-likelihood ratio test statistic is given by

$$L_{01} = 2\log\{L(\hat{\boldsymbol{\theta}}^{(1)}|\boldsymbol{y})\} - 2\log\{L(\hat{\boldsymbol{\theta}}^{(0)}|\boldsymbol{y})\},$$

where $\boldsymbol{\theta} = (\beta_1 \,...\, \beta_m \; \sigma^2)'$. L_{01} has an asymptotic chi-squared distribution with $m_1 - m_0$ degrees of freedom, where m_1 and m_0 are the numbers of explanatory variable coefficients in model H_1 and H_0 respectively. Typically, H_0 will be rejected in favour of H_1 when L_{01} is greater than the 95% point of the $\chi^2_{m_1-m_0}$ distribution. Again, as for the F-test, for a large number of explanatory variables model selection using the log-likelihood ratio test procedure can be time consuming and there may not be any one model which offers the best fit to the data. So an analyst should generally pick the most plausible model and use residual plots to assess the fit of that model.

For GLM's, residuals are generally not calculated simply by subtracting the fit from the observations; instead, usually residual deviances are used, which are calculated from the scaled deviance. The scaled deviance is calculated in the same way as the log-likelihood ratio statistic, where model H_0 is the chosen model and model H_S is the alternative model. So the scaled deviance is given by

$$L_{0S} = 2\log\{L(\hat{\boldsymbol{\theta}}^{(S)}|\boldsymbol{y})\} - 2\log\{L(\hat{\boldsymbol{\theta}}^{(0)}|\boldsymbol{y})\}.$$

Here, the S in model H_S stands for saturated, so called because this model has n parameters, as many as there are observations. Because there are as many parameters as observations, the saturated model provides a perfect fit to the data where the fitted mean for each observation is equal to that observation: $\hat{\mu}_i = y_i$. This means that the scaled deviance expressed in terms of the mean value parameter is

$$D(\boldsymbol{y}|\hat{\boldsymbol{\mu}}) = 2\log\{L(\boldsymbol{y}|\boldsymbol{y})\} - 2\log\{L(\hat{\boldsymbol{\mu}}|\boldsymbol{y})\}. \tag{2.6}$$

The residual deviances are given by

$$r_i^D = sign(y_i - \hat{\mu}_i)\sqrt{d_i}, \qquad i = 1, ..., n, \tag{2.7}$$

where $\sum d_i = D(\boldsymbol{y}|\hat{\boldsymbol{\mu}})$. Other types of residual can be used aside from the deviance residuals; the most widely used of these are Pearson residuals:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}}, \qquad i = 1, ..., n.$$

When residuals have been calculated, in most cases they can be displayed using the same array of plots as given in §2.1.2. Even the normal probability plot is often suitable for residual deviances as long as the counts are not too small, in this case the half-normal probability plot would be more suitable.

2.2.3 The Poisson model

A sensible choice for the modelling of count data, like road accident counts, is to assume the counts are generated by a Poisson process since Poisson random variables can only take non-negative integer values and have no upper limit. The Poisson density is given by

$$f(y_i|\lambda_i) = \frac{\lambda_i^{y_i} e^{\lambda_i}}{y_i!}. \tag{2.8}$$

Written in exponential family form this is

$$f(y_i|\theta_i) = \exp\{y_i\theta_i - \exp(\theta_i) - \log(y_i!)\},$$

where the canonical parameter $\theta_i = \log(\lambda_i)$ and the dispersion parameter $\phi_i = 1$; here, $b(\theta_i) = \exp(\theta_i)$ and $a(\phi_i) = \phi_i$. In a Poisson model the parameter λ_i is the mean, so we have $\mu_i = \lambda_i$; also, for the Poisson model with the canonical link we have $\eta_i = \theta_i$. Therefore in a Poisson model with the canonical link, the linear predictor relates to the mean by a log-link, $\eta_i = g(\mu_i) = \log(\mu_i)$. The Poisson model with the log-link is often known as the log-linear model.

The log-linear model log-likelihood is given by

$$\log\{L(\boldsymbol{\lambda}|\boldsymbol{y})\} = \sum_{i=1}^{n}\{y_i \log(\lambda_i) - \lambda_i\}. \tag{2.9}$$

Substituting (2.9) into (2.6), we obtain the scaled deviance for the log-linear model:

$$D(\boldsymbol{y}|\hat{\boldsymbol{\lambda}}) = 2\sum_{i=1}^{n}\{y_i \log(y_i/\hat{\lambda}_i) - y_i + \hat{\lambda}_i\}.$$

Therefore the residual deviance for the log-linear model is

$$r_i^D = sign(y_i - \hat{\lambda}_i)\surd 2(y_i \log(y_i/\hat{\lambda}_i) - y_i + \hat{\lambda}_i), \qquad i = 1, ..., n. \tag{2.10}$$

Although Poisson variation is a logical assumption to apply to count data, it is seldom the case that after calculating the mean we find that it is exactly equal to the variance. Instead, it is often the case that the variance of the data is greater than the mean: $Var(\boldsymbol{Y}) > E(\boldsymbol{Y})$, which is a departure from the Poisson assumption that $Var(\boldsymbol{Y}) = E(\boldsymbol{Y})$. This could occur for a variety of reasons, for road accident counts the main reason is likely to be that the explanatory variables do not manage to explain all the systematic variation in the data and so some of this systematic variation is being interpreted as random and is thus included in the variance. When a log-linear model is such that the variance is significantly greater than the mean then the model is said to be over-dispersed.

The simplest way of formulating the over-dispersion is to assume that $Var(\boldsymbol{Y}) = \sigma^2 E(\boldsymbol{Y})$ for some constant σ^2. Then σ^2 can be estimated using the Pearson statistic, X^2, divided by the degrees of freedom, $n - m$, for the model with m explanatory variables:

$$\hat{\sigma}^2 = X^2/(n-m) = \frac{1}{n-m}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \tag{2.11}$$

Note that the Pearson statistic is the sum of the Pearson residuals: $X^2 = \sum r_i^P$. In a log-linear model the dispersion parameter is assumed to be 1, therefore an over-dispersed log-linear model is one for which the estimate $\hat{\sigma}^2$ is found to be significantly larger than one. If a log-linear model is found to be over-dispersed then the coefficient variance-covariance matrix is multiplied by $\hat{\sigma}^2$ in order to obtain an approximate measure of precision for $\hat{\boldsymbol{\beta}}$. This translates to multiplying the coefficient estimate standard errors, $s.e(\hat{\boldsymbol{\beta}})$, by $\hat{\sigma}$ to take into account the increased uncertainty over the coefficient estimates.

2.3 Regression methods applied to Scottish and Southwest English car occupant accident data

In this section we examine Scottish and Southwest English car occupant accident data using linear and log-linear regression models. The data examined are the number of car occupant fatalities and the total number of car occupants killed and injured (fatalities, serious injuries and slight injuries) per month in Scotland and Southwest England from 1987 to 2000. The four series are given in figure 2.1. Scottish and Southwest English data are used here rather than data from any other regions of the UK because

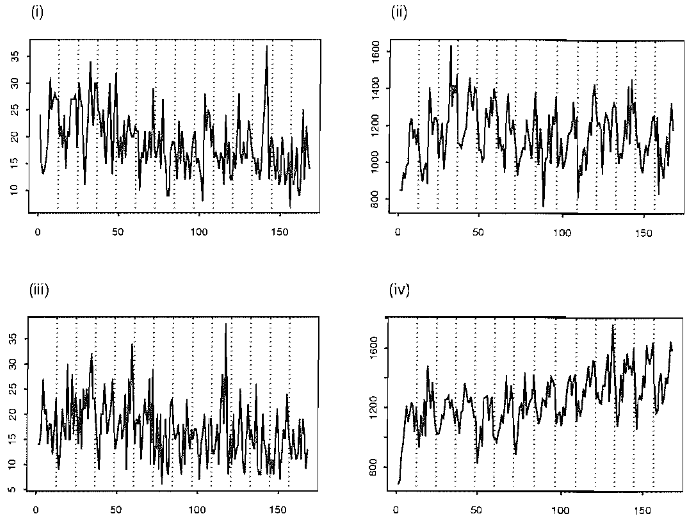

Figure 2.1: Car occupant casualty time series from January 1987 to December 2000: (i) Scottish fatalities, (ii) Scottish deaths and injuries, (iii) Southwest English fatalities, (iv) Southwest English deaths and injuries.

the data and results obtained here will be made use of in the techniques of chapter 4 where a large difference in latitude between two regions is required.

2.3.1 Model fitting

Linear and log-linear models were fitted to all four series and explanatory variables were used to model the variation in the data. The linear models were fitted to the raw data counts since transformations such as logging or taking square roots added unnecessary complexity to the models and yielded no improvement in the model fits. The explanatory variables examined in the model building were those detailed in §1.1; thus the rainfall variable, temperature variables, cloud cover and daylight were specific to each region, but the traffic variable and the two fuel variables were national. Note, the daylight variables were taken at the rough geographical centres of Scotland and Southwest England, they being respectively: the town of Stirling located at latitude 56.12 and longitude -3.94, and the town of Langport located at latitude 51.03 and longitude -2.83. A fitted linear trend designed to represent gradual change in the response variable not captured by the explanatory variables was also used in the analysis; this device was used in a comparable study using log-linear models on road accidents in the Scandinavian countries and Finland (Fridstrom et al., 1995). The linear trend is designed to represent such factors as improvements in the design of safety features in cars, like side impact bars and improved tire traction, as well as general improvements in legislation and traffic management by government and local authorities.

Under the linear regression models, tables 2.1 and 2.2, those explanatory variables found to have a significant effect on car occupant casualties along with their coefficients and coefficient standard errors for the four accident series are shown. The residual variance estimates for the models fitted to the Scottish and Southwest fatalities data are $\hat{\sigma}^2 = 23.00$ and $\hat{\sigma}^2 = 29.17$ respectively; residual variance estimates for the Scottish and Southwest deaths and injuries models are $\hat{\sigma}^2 = 10421$ and $\hat{\sigma}^2 = 11286$ respectively. R-squared values are very low for the Scottish and Southwest fatalities models at 0.30 and 0.16 respectively; they are better for the Scottish and Southwest deaths and injuries models at 0.56 and 0.67. Linear model residual plots for the Scottish fatalities and Southwest fatalities data (figures 2.2 and 2.4) indicate good fits. However, the time plot of residuals for the Scottish and Southwest deaths and injuries linear models show a slight variation of residuals about zero through time suggesting that not all systematic variation has been eliminated from these series. Also, the Southwest correlogram has a significant spike at lag 12, this suggests that there may be some seasonal variation in the data which has not been accounted for in the model. Despite these problems, overall the Scottish and Southwest deaths and injuries models provide reasonable fits as can be seen in figures 2.3 and 2.5.

Tables 2.3 and 2.4 show those explanatory variables found to have a significant effect on the observations under the log-linear models fitted to the four series. It can be seen that exactly the same variables found significant in the linear models are also significant under the log-linear models. Deviance and Pearson residual plots are virtually identical for the log-linear models as for their respective linear counterparts and therefore have not been shown. The log-linear Scottish and Southwest deaths and injuries models again show variation of the residuals around zero through time, and the Southwest deaths and injuries model also has a spike at lag 12 on the correlogram like its linear counterpart. Over-dispersion is present in the log-linear Scottish and Southwest fatalities models, the estimated dispersion parameters are $\hat{\sigma}^2 = 1.207$ and $\hat{\sigma}^2 = 1.717$ respectively. The dispersion estimates for the Scottish and Southwest deaths and injuries models are much larger at $\hat{\sigma}^2 = 9.054$ and $\hat{\sigma}^2 = 9.275$ respectively. The standard errors on the log-linear model coefficient estimates in tables 2.3 and 2.4 have been multiplied by the square roots of the dispersion estimates to get a better measure of precision for the coefficient estimates, as mentioned earlier. Due to the increase in the sizes of standard errors after multiplying by the square root of the dispersion estimates, some of the explanatory variables originally included were removed from the log-linear deaths and injuries models. This made no change to the overall fit of the two models and actually had the effect of marginally reducing the over-dispersion in these models.

In all, it can be seen that the linear and log-linear models are virtually identical both in terms of the explanatory variables included and the signs of their respective coefficients, so linear models provide a very reasonable approximation to the count data in this example. It is worth noting that the fatality data series in both the linear and log-linear cases are modelled with considerably fewer explanatory variables than the deaths and injuries series. This is likely to be since the sizes of the monthly observations are much smaller for the fatalities data and so the series exhibit more random variation which explanatory variables cannot pick up on.

Variable	Fatal		Total	
	Coef	Std err	Coef	Std err
Constant	29.12	1.607	-377.2	179.7
Linear Trend	-0.05345	0.007690	—	
Nat car traffic	—		53.38	8.675
Nat petrol del	—		3.841 $\times 10^{-4}$	6.872 $\times 10^{-5}$
Nat diesel del	—		-5.328 $\times 10^{-4}$	9.851 $\times 10^{-5}$
Total rainfall	—		1.398	0.2934
Max daily temp	—		11.98	4.025
Min daily temp	0.8012	0.1574	—	
Cloud cover	—		—	
Daylight	-0.8132	0.1561	-36.41	4.138

Table 2.1: Scottish data explanatory variable coefficients and standard errors from fitted linear regression models.

Variable	Fatal		Total	
	Coef	Std err	Coef	Std err
Constant	-4.575	7.039	182.3	142.0
Linear Trend	-0.09542	0.01773	—	
Nat car traffic	1.269	0.3450	45.44	3.505
Nat petrol del	—		—	
Nat diesel del	—		—	
Total rainfall	—		1.524	0.2451
Max daily temp	—		—	
Min daily temp	—		13.41	3.470
Cloud cover	—		-32.18	16.16
Daylight	-0.5877	0.2120	-22.52	4.539

Table 2.2: Southwest English data explanatory variable coefficients and standard errors from fitted linear regression models.

Variable	**Fatal**		**Total**	
	Coef	Std err	Coef	Std err
Constant	3.470	0.07552	5.674	0.1615
Linear Trend	-0.002823	3.692×10^{-4}	—	
Nat car traffic	—		0.04793	0.007685
Nat petrol del	—		3.428×10^{-7}	6.042×10^{-8}
Nat diesel del	—		-4.715×10^{-7}	8.603×10^{-8}
Total rainfall	—		0.001187	2.515×10^{-4}
Max daily temp	—		0.01023	0.003524
Min daily temp	0.04243	0.007565	—	
Cloud cover	—		—	
Daylight	-0.04338	0.007583	-0.03221	0.003644

Table 2.3: Scottish data explanatory variable coefficients and standard errors (over-dispersion included) from fitted log-linear regression models.

Variable	**Fatal**		**Total**	
	Coef	Std err	Coef	Std err
Constant	1.569	0.4146	6.234	0.1165
Linear Trend	-0.005654	0.001040	—	
Nat car traffic	0.07423	0.02016	0.03781	0.002928
Nat petrol del	—		—	
Nat diesel del	—		—	
Total rainfall	—		0.001210	1.979×10^{-4}
Max daily temp	—		—	
Min daily temp	—		0.01063	0.002825
Cloud cover	—		-0.02654	0.01314
Daylight	-0.03462	0.01251	-0.01806	0.003706

Table 2.4: Southwest English data explanatory variable coefficients and standard errors (over-dispersion included) from fitted log-linear regression models.

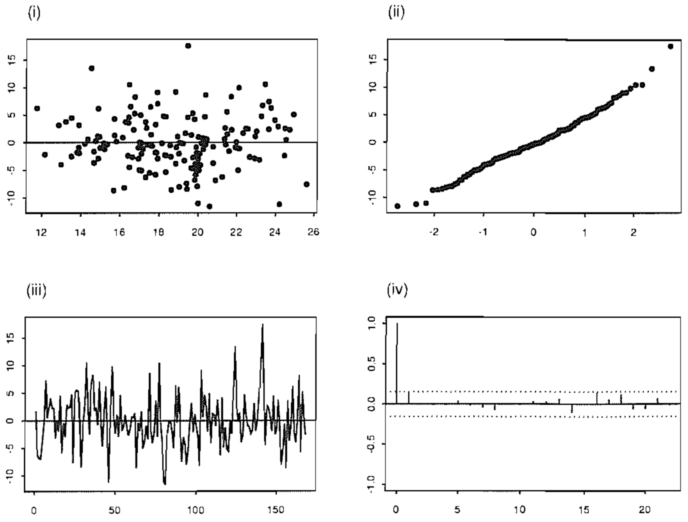

Figure 2.2: Scottish fatalities linear model residual plots: (i) residuals vs fitted values, (ii) normal plot, (iii) time chart of residuals, (iv) correlogram.

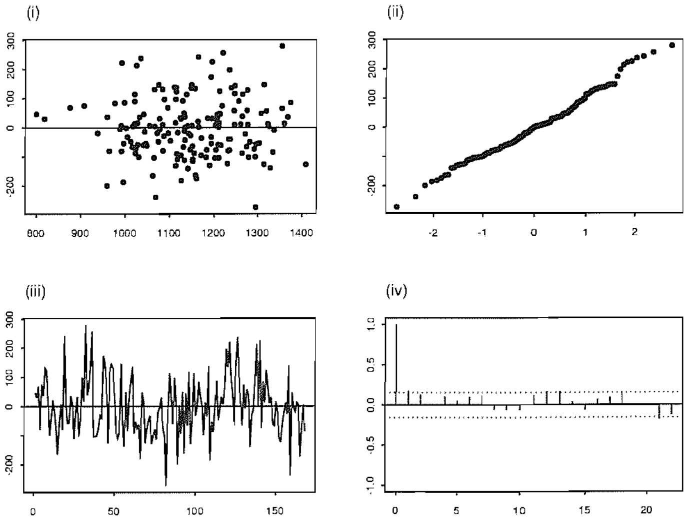

Figure 2.3: Scottish deaths and injuries linear model residual plots: (i) residuals vs fitted values, (ii) normal plot, (iii) time chart of residuals, (iv) correlogram.

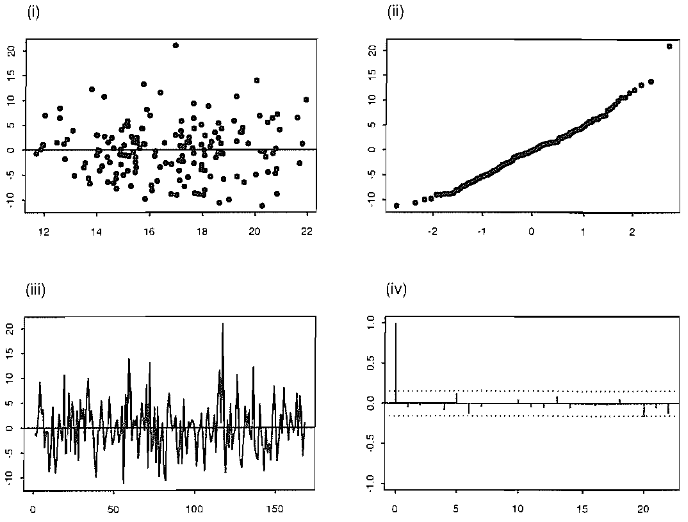

Figure 2.4: Southwest fatalities linear model residual plots: (i) residuals vs fitted values, (ii) normal plot, (iii) time chart of residuals, (iv) correlogram.

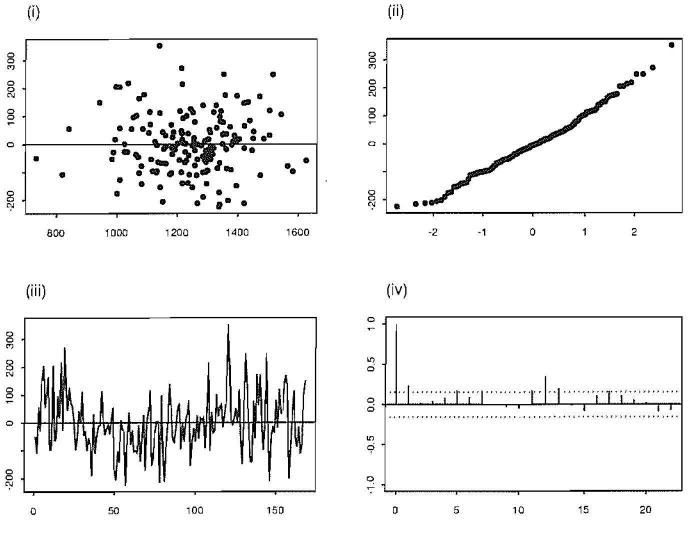

Figure 2.5: Southwest deaths and injuries linear model residual plots: (i) residuals vs fitted values, (ii) normal plot, (iii) time chart of residuals, (iv) correlogram.

2.3.2 Conclusions and interpretations

Scottish data

From the coefficients of table 2.1 we can see that the Scottish fatalities linear model shows that a 1^o Celsius rise in the average minimum daily temperature contributes to an increase of 0.8 fatalities. Also, an increase in the length of day from sunrise to sunset of 1 hour leads to a decrease of 0.8 fatalities. There is a general downward trend in Scottish car occupant fatalities indicated by the negative trend coefficient, albiet rather small. The coefficient indicates that each month there are 0.05 fewer fatalities than in the previous month due to general factors not covered by the explanatory variables such as improvements in road infrastructure and improvements in car safety.

Table 2.1 also gives the coefficients for the total Scottish car occupants killed and injured linear model. These coefficients show that an increase in travel of 1 billion kilometres nationally leads to an increase of 53.38 deaths and injuries in Scotland. An increase in national petrol deliveries of 1000 tonnes contributes to a rise of 0.3841 deaths and injuries in Scotland. And a rise in national diesel deliveries of 1000 tonnes contributes to a fall of 0.5328 deaths and injuries in Scotland. The regional Scottish explanatory variables show that an increase in rainfall of 1mm leads to an increase of 1.398 deaths and injuries, an increase of 1^o Celsius in the average maximum daily temperature contributes to an increase of approximately 12 deaths and injuries, and a 1 hour increase in daylight leads to a fall of 36.4 deaths and injuries.

Table 2.3 indicates a similar story for the Scottish fatalities and Scottish total killed and injured log-linear models as is given in the linear models. The fatalities model shows that a 1^o Celsius rise in the average minimum daily temperature contributes to an increase in fatalities of 4.3% and an increase in the length of day from sunrise to sunset of 1 hour leads to a decrease in fatalities of 4.25%. The linear trend term again indicates a general downward trend in Scottish car occupant fatalities with the coefficient indicating that each month there are 0.28% fewer fatalities than in the previous month caused by factors not covered by the explanatory variables.

The Scottish deaths and injuries model from table 2.3 shows that an increase in travel of 1 billion kilometres nationally leads to an increase in deaths and injuries of 4.9% in Scotland. An increase in national petrol deliveries of 1000 tonnes contributes to a 0.034% increase in deaths and injuries in Scotland. The annual variation in petrol deliveries is in the region of 300,000 tonnes, so such an increase is substantial. A rise in national diesel deliveries of 1000 tonnes contributes to a 0.047% decrease in deaths and injuries in Scotland. For diesel deliveries, annual variation is generally over 10,000 tonnes. The regional Scottish explanatory variables show that an increase in rainfall of 1mm leads to an increase in deaths and injuries of 0.1188%, an increase of 1^o Celsius in the average maximum daily temperature contributes to an increase in deaths and injuries of approximately 1%, and a 1 hour increase in daylight leads to a fall in deaths and injuries of 3.17%.

Southwest English data

From the coefficients of table 2.2, the Southwest English fatalities linear model shows that an increase in travel of 1 billion kilometres nationally leads to an increase of 1.269 fatalities in Southwest England and an increase in the length of day from sunrise to sunset of 1 hour in Southwest England leads to a decrease of approximately 0.59 fatalities. There is a general downward trend in the number of car occupant fatalities in Southwest England indicated by the negative trend coefficient, showing that each month there are 0.095 fewer fatalities than in the previous month due to general factors not covered by the explanatory variables.

Table 2.2 also gives the coefficients for the total Southwest English car occupants killed and injured linear model. These coefficients show that an increase in travel of 1 billion kilometres nationally leads to an increase of 45.44 deaths and injuries in Southwest England. The regional Southwest explanatory variables show that an increase in rainfall of 1mm leads to an increase of 1.524 deaths and injuries, an increase of 1^o Celsius in the average minimum daily temperature contributes to an increase of 13.41 deaths and injuries, and a 1 hour increase in daylight leads to a fall of 22.52 deaths and injuries. Cloud cover is also a significant variable and an increase of 1 okta (1/8 cloud cover) corresponds to a decrease of 32.18 fatalities.

Table 2.4 indicates a similar story for the Southwest fatalities and Southwest deaths and injuries log-linear models as for the linear models. The fatalities model shows that an increase in travel of 1 billion kilometres nationally leads to an increase in fatalities of 7.7% in Southwest England and an increase in the length of day from sunrise to sunset in Southwest England of 1 hour leads to a decrease in fatalities of 3.402%. The linear trend term again indicates a general downward trend in Southwest English car occupant fatalities with the coefficient indicating that each month there are 0.56% fewer fatalities than in the previous month caused by factors not covered by the explanatory variables.

The Southwest deaths and injuries model from table 2.4 shows that an increase in travel of 1 billion kilometres nationally leads to an increase in deaths and injuries of 3.85% in Southwest England. The regional Southwest explanatory variables show that an increase in rainfall of 1mm leads to an increase in deaths and injuries of 0.121%, an increase of 1^o Celsius in the average minimum daily temperature contributes to an increase in deaths and injuries of approximately 1%, a 1 hour increase in daylight leads to a fall in deaths and injuries of 1.79%, and an increase of 1 okta in cloud cover leads to a decrease in deaths and injuries of 2.62%.

Interpretations of the models

Generally, the interpretations of the effects of the explanatory variables on the data in the various models considered are fairly straightforward.

Although there are some exceptions which on the face of it seem to be counter intuitive, there are plausible explanations and these are detailed in the the following paragraphs. Considering the more obvious explanatory variable interpretations firstly though, we have the following: The daylight variable features in all models and has the effect that more hours of daylight leads to fewer accidents, which seems reasonable as with increased daylight hours visibility would be good for a longer period of each day. Also, the absolute value of the daylight coefficients are larger for all the Scottish models than the Southwest models, implying that daylight has a stronger effect on car occupant casualties in Scotland than Southwest England. The rainfall variable features in the Scottish and Southwest deaths and injuries models, the effect in these models is that increased precipitation leads to an increase in accidents; this again seems plausible. The national traffic variable features in all but the Scottish fatalities models and has the effect that an increase in traffic leads to an increase in casualties in whichever model it appears in; the same result has been found in all other studies where traffic has been used as an explanatory variable.

One of the two temperature variables features in all but the Southwest fatalities model; they have the same effect in all the models in which they appear: that is, an increase in temperature leads to an increase in casualties. If the monthly average minimum daily temperature is a proxy for adverse weather conditions such as snow and frost, as was suggested in the previous chapter, then it appears that the argument that these conditions encourage better driving and more skilled drivers is possibly true. However, this variable is highly correlated with the average maximum daily temperature variable and so could simply be showing that good weather encourages more drivers onto the road and so act as a proxy for exposure.

The two fuel variables: petrol deliveries and diesel deliveries, feature in the linear and log-linear Scottish deaths and injuries models. However, it seems on the face of it slightly implausible that while an increase in petrol deliveries leads to an increase in road accidents, an increase in diesel deliveries should lead to a decrease in accidents. But it was decided to keep the model as it was rather than try to change things so there was no such discrepancy because there are differences between the types of vehicles that use petrol and the type that use diesel and these may explain why higher diesel imports seem to decrease road accidents. Diesel, on the whole, tends to be used in commercial and goods vehicles such as lorries, trucks, vans and taxi's, although an increasing number of private cars also run on diesel. For the period under investigation there has been a steady, almost linear, increase in the level of diesel deliveries implying that more and more vehicles on the road are powered by diesel fuel, and over the same period of time there has been an initial rise and then tailing off in the level of petrol deliveries. Diesel powered vehicles are generally slower than petrol powered vehicles, not only because large slow moving vehicles use diesel fuel, but also because diesel powered vehicles of a particular make and model will generally not have quite the speed of petrol powered vehicles of the same make and model. It is possible that the presence of a higher proportion of diesel powered vehicles on the road could generally slow the pace of traffic and thereby reduce accidents.

Another unexpected result is that more cloud cover appears to have the effect of decreasing and not increasing car occupant accidents; this variable appears in the Southwest deaths and injuries models. It was expected that cloud cover would behave in a similar way to daylight since more cloud generally reduces light levels and so should lead to an increase in accidents; also, more cloud cover generally comes with a higher chance of rain and so, again, should lead to more accidents. The fact that cloud cover has been found to have an effect in reducing accidents must be due to the fact that the rainfall and daylight variables are also present in the models in which the cloud cover variable appears, and so must account for all the daylight and rainfall effects. Eliminating the effects of rainfall and daylight, it is not difficult to see that cloud cover could have a positive effect since cloud cover reduces dazzle from the sun, especially at times when the sun is low in the sky and also when the sun is shining during or just after rainfall.

Chapter 3

ARIMA time series models

The topic of Autoregressive Integrated Moving Average (ARIMA) analysis is large and has many different aspects. This chapter summarises the most important elements of ARIMA processes for data taken at discrete time points and then goes on to apply the methods to the Scottish deaths and injuries data analysed in chapter 2. A thorough coverage of this topic can be found in Hamilton (1994) while a more introductory approach is given by Diggle (1990). Other standard texts on this subject are Chatfield (1975) and Harvey (1981). The chapter begins with a summary of the concept of stationarity which is a key component of ARIMA modelling, then §3.2 covers the implementation of ARIMA time series modelling from model building to parameter estimation. Section 3.3 goes on to apply the theory to data. Throughout the chapter, methods are presented in terms of univariate observation data.

3.1 Stationarity

3.1.1 Definitions

An important concept when dealing with ARIMA time series models is that of *stationarity*; we shall consider the definitions of stationarity for discrete time processes as these are the subject of this thesis. A time series is said to be strictly stationary if the joint probability distribution associated with p observations, $y_{t_1}, ..., y_{t_p}$, made at any set of times, $t_1, ..., t_p$, is the same as the joint distribution of $y_{t_1+k}, ..., y_{t_p+k}$, for any value k. In other words, shifting the time origin by k has no effect on the joint distributions which therefore only depend on the intervals between $t_1, ..., t_p$; this definition holds for any value p. Generally speaking, this definition implies that a stationary time series is one in which there is no systematic change in the mean (a trend), no change in the variance (heteroskedasticity) and no periodic variations such as seasonal or cyclic effects. So a stationary time series will appear to look similar whichever point in time it is observed.

In practice strict stationarity is often an uncheckable assumption, so the weaker assumption of second-order stationarity (weak or covariance stationarity), defined by the autocovariance function, is used instead. For a sequence of identically distributed time series observations, y_t, each with mean $E(Y_t) = \mu_t$, the autocovariance function $\gamma_{k,t}$ is defined as

$$\gamma_{k,t} = E\{(Y_t - \mu_t)(Y_{t-k} - \mu_{t-k})\},$$

which is much like the ordinary covariance function except defined for only one sequence of observations at two points in time. For the sequence to be covariance stationary, or weakly stationary, we have

$$\begin{aligned} E(Y_t) &= \mu, \qquad for\ all\ t, \\ E\{(Y_t - \mu)(Y_{t-k} - \mu)\} &= \gamma_k, \qquad for\ all\ t. \end{aligned} \tag{3.1}$$

Therefore, for a process to be weakly stationary, the covariance between y_t and y_{t-k} depends only on k, the length or lag in time separating the observations, and not on t, the point in time. In this definition no assumption is made about higher moments than those of second order. From the definition of the covariance function it can be seen that the marginal variance is given by γ_0.

A good example of a stationary time series is a white noise process, which is simply a sequence of mutually independent random variables, each with mean zero and variance σ^2. Its autocovariance function is

$$\gamma_{k,t} = \begin{cases} \sigma^2, & t = k, \\ 0, & t \neq k. \end{cases} \tag{3.2}$$

Stationarity is an important concept in ARIMA time series analysis since a time series needs to be converted to stationarity before analysis can be carried out on it. The white noise process is used as a benchmark to asses possible serial dependence in an observed time series and is used in the formulation of ARMA models shown later. In this context, the white noise process is usually Gaussian distributed and will be considered so for the rest of the chapter, but it need not be in all cases.

3.1.2 Converting a time series to stationarity

A time series can be converted to stationarity in a variety of ways. One way is to fit a regression model to the series. This should eliminate undesirable characteristics such as trend and seasonal variation giving a sequence of residuals which are stationary; the use of this method is demonstrated in chapter 2. Another way to convert a time series to stationarity is to use differencing. Here, elements of the time series are subtracted from one another in such a way as to yield a stationary series. In most texts a differenced series is usually given by an operator, Δ, such that for time series data, y_t,

$$\Delta_s y_t = y_t - y_{t-s}, \tag{3.3}$$

for all integers s. To remove a linear trend from a series for instance, s is set to one; in this case we usually write $\Delta_1 = \Delta$, i.e., without the subscript. To remove seasonal variation from a series, s is set to the length of the seasonal cycle; so for data collected monthly, for instance, the seasonal is likely to arise in a yearly pattern, i.e., we set $s = 12$. Polynomial trend removal can be carried out by differencing to the degree of the polynomial; e.g., if a quadratic trend is present in the time series then it is eliminated as follows:

$$\Delta^2 y_t = \Delta(\Delta y_t) = \Delta y_t - \Delta y_{t-1} = y_t - 2y_{t-1} + y_{t-2}.$$

This can be extrapolated up to polynomials of any order, d, although it is not usual to go above order 3.

3.2 ARIMA modelling

3.2.1 The ARIMA model

ARIMA time series analysis is designed specifically to handle data that has been collected over time, in comparison with regression which can be used to model the relationships between factors of any sort. Perhaps, more specifically, ARIMA analysis is useful since data collected over time can exhibit signs of *autocorrelation*, that is dependence among observations within the same time series, which the ARIMA framework can handle. For instance, on a small time scale, daily, say, localised temperature readings are likely to exhibit signs of autocorrelation, that is, knowing the weather yesterday is likely to add to the accuracy of a prediction of today's weather compared to not knowing yesterday's weather.

If a time series is stationary, or has been converted to stationarity, then an Autoregressive Moving Average (ARMA) model can be used to model the correlation structure of the stationary series. An Autoregressive *Integarated* Moving Average (ARIMA) model is an ARMA model applied to a time series which has been converted to stationarity using differencing. However stationarity is achieved, the ARMA modelling principle is that the resulting time series can be expressed as a function of past observations and past error terms such that

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j z_{t-j} + z_t, \tag{3.4}$$

where z_t is a white noise term and $-1 < \phi_i, \theta_j < 1$, for a stationary process. ARMA processes with only terms in ϕ and none in θ are called autoreggressive processes, and ARMA processes with only terms in θ and none in ϕ are called moving average processes.

3.2.2 Model building

As with all statistical problems, model building for ARIMA models is a process of formulation, estimation and verification. Formulation for ARIMA models, however, is not a precise science. An initial model may be specified using a mixture of knowledge and examination of the data. Looking at the time plot of the data is very helpful, especially for the type of road accident data presented in chapter 2. Generally, it can be seen whether there is evidence of a trend or seasonal effect in the data so the series can then be differenced according to the methods above to hopefully reduce it to stationarity.

The next stage is the identification of a suitable ARMA model from the stationary series. Again, this is not a precise science but can be helped with the aid of the autocorrelation function, partial autocorrelation function and spectrum. For a stationary series the autocorrelation function, ρ_k, shows the correlation structure among the observations. It is derived from the autocovariance function for a stationary process, (3.1), and is given by

$$\rho_k = \frac{\gamma_k}{\gamma_0}.$$

Here, $\rho_k = \rho_{-k}$, $-1 \leq \rho_k \leq 1$ and if y_t and y_{t-k} are independent then $\rho_k = 0$.

For a real set of observed data the autocorrelation function needs to be estimated and we have already come across the estimated, or sample, autocorrelation function, $\hat{\rho}_k$, in (2.4). Recalling from §2.1.2, the $\hat{\rho}_k$ values can then be plotted against k to give a plot known as the correlogram. The correlogram is interpreted for various features which suggest that autoregressive or moving average terms are needed to model the data, and can also be used for judging whether a series has successfully been transformed to stationarity.

From §2.1.2 we know that a correlogram which exhibits virtually no significant autocorrelations, except at lag zero, shows that the data or residuals from which it has been derived have no dependence structure. If the correlogram shows an exponential decay in the autocorrelations, it suggests that autoregressive terms are needed to model the dependencies in the data. We then use the *partial* autocorrelation function to asses the degree of the autoregression; that is to say, the value of p from (3.4).

If we consider the following succession of autoregressive models:

$$\begin{aligned}
y_t &= \phi_{11}y_{t-1} + z_t, \\
y_t &= \phi_{21}y_{t-1} + \phi_{22}y_{t-2} + z_t, \\
y_t &= \phi_{31}y_{t-1} + \phi_{32}y_{t-2} + \phi_{33}y_{t-3} + z_t, \\
&\vdots \\
y_t &= \phi_{i1}y_{t-1} + \phi_{i2}y_{t-2} + \phi_{i3}y_{t-3} + ... + \phi_{ii}y_{t-i} + z_t, \\
&\vdots
\end{aligned}$$

then the succession $\phi_{11}, \phi_{22}, \phi_{33}, ..., \phi_{ii}, ...,$ is the partial aurocorrelation function. So the partial autocorrelation function is constructed from the series of last coefficients corresponding to autoregressive processes of successively higher orders. It is estimated by fitting successive autoregression models to the series. A plot of the estimated partial autocorrelations against their lag gives the partial correlogram. The partial correlogram of an autoregressive process of order p will show a sharp cut-off at lag p.

The correlogram and partial correlogram can also be used to identify moving average processes. A moving average process of order q shows a sharp cut-off at lag q on the correlogram and an exponential decay on the partial correlogram similar to that shown on the correlogram for an autoregressive process. A mixed process with an order p autoregressive component and an order q moving average component will show a mixture of exponential and damped sine waves after the first $q-p$ lags on the correlogram and a mixture of exponential and damped sine waves after the first $p-q$ lags on the partial correlogram.

A series which has not effectively been reduced to stationarity will also exhibit characteristic behavior in the correlogram and partial correlogram. A linear trend, for example, will appear as a linear decay in the correlogram and partial correlogram, while a seasonal effect which has not been accounted for will appear as damped sine wave effects in both the correlogram and partial correlogram.

Sometimes the correlogram might not show up seasonal effects very well, in cases like this another graph called the periodogram can be used which is more suited to finding unexpected seasonal behavior in a series than the correlogram. Generally, with data such as road accident data taken monthly over a period of years, there is usually evidence of an annual seasonal effect from a time plot of the data and so it would probably not be necessary to confirm this from the periodogram. Model identification can also be aided via the Akaike Information Criterion (AIC) or with likelihood ratio tests similar to those used for explanatory variable selection in log-linear regression models.

Once an ARIMA model has been fitted to the data, the residuals, z_t, from (3.4), can again be assessed using the correlogram and partial correlogram to see whether the fitted model has successfully eliminated the dependence structure in the data. Here, we generally discard z_t for $t \leqslant \max(p,q)$.

3.2.3 Parameter estimation

Once a suitable ARIMA model is proposed, the parameter vectors $\boldsymbol{\phi} = (\phi_1 \, ... \, \phi_p)'$ and $\boldsymbol{\theta} = (\theta_1 \, ... \, \theta_q)'$ can be estimated by maximising the log-likelihood as with regression. When calculating the log-likelihood, the initialisation of the ARMA process is important. If we have a simple AR(1) process for instance (a process with one autoregressive term), $y_t = \phi y_{t-1} + z_t$, then clearly we cannot easily find $E(Y_1)$ since there is no y_0. For this reason, under the conditional approach, the likelihood would only be calculated from observations $y_2, ..., y_n$ conditional on y_1, and more generally for an AR(p) process, from $y_{p+1}, ..., y_n$ conditional on $y_1, ..., y_p$.

There is a similar problem for the moving average terms. For instance, if we have an MA(1) model (a process with one moving average term), $y_t = \theta z_{t-1} + z_t$, then not only can we not calculate $E(Y_1)$ easily but we cannot calculate $E(Y_2)$ easily either, unless we know z_1. For this reason it is generally assumed under the conditional approach that $z_1 = z_2 = ... = z_q = 0$ and the log-likelihood is only calculated for values from $q + 1$ to n.

For an ARMA model with Gaussian errors the conditional log-likelihood is given by

$$\begin{aligned}&\log\{L(\boldsymbol{\phi}, \boldsymbol{\theta}|y_p, ..., y_n, z_p = 0, ..., z_{p-q+1} = 0)\}\\ &\qquad = -\frac{n-p}{2}\log(2\pi) - \frac{n-p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=p+1}^{n} z_t^2.\end{aligned}$$

The conditional log-likelihood provides a way of calculating parameter estimates by hand along with their associated standard errors. To calculate the exact log-likelihood computer software is needed since initialisation methods for the exact log-likelihood depend on simulated draws. Hamilton (1994), chapter 5, gives a full explanation of how the exact log-likelihood is calculated using a variety of numerical procedures.

When missing values are present in the data, S-Plus converts the problem to state space form (§5.1.2) and the Kalman filter (§5.2.1) is used to calculate the log-likelihood and parameter estimates. An example of the ARIMA model in state space form can be found in Durbin and Koopman (2001), §3.3. A detailed discussion of the relationship between ARIMA time series models and structural time series models can be found in Harvey (1989), §2.5.3.

3.3 ARIMA analysis of Scottish car occupant deaths and injuries data

In this section the use of ARIMA analysis is illustrated by applying an ARIMA model to the Scottish deaths and injuries data, i.e., we reduce the series to stationarity using differencing rather than using another method such as the fitting of explanatory variables, then an ARMA model is applied to the resulting series.

The plot of the Scottish total car occupant accidents data, figure 2.1, reveals a possible trend together with a definite 12 monthly seasonal pattern. Thus the observations, y_t, should certainly be seasonally differenced to eliminate the seasonal pattern and should perhaps be differenced to account for trend. Despite the trend having an almost cubic polynomial look to it, it is usually not a good idea to difference a series too many times since it can introduce spurious autocorrelation structure to the resulting series. Usually differencing a series once for trend is enough, especially when higher order polynomial effects are not particularly pronounced as is the case for the Scottish data.

Differencing the Scottish data once, Δ^1, and then seasonally differencing, Δ_{12}, reduces the data to stationarity; the resulting differenced series and corresponding correlogram is given in figure 3.1. The correlogram reveals some signs of autocorrelation in the differenced series. The significant autocorrelations at lags 1 and 12 suggest that any autoregressive or moving average terms fitted to the differenced series should account for the autocorrelations at these lags. Both of these autocorrelations are isolated and do not decay gradually in the lags, so moving average terms are suitable to model them. In fact, the form of the correlogram is quite distinctive and leads us to choose the well known Airline model (Box and Jenkins, 1970) to model the dependencies. The airline model is given by

$$\Delta\Delta_{12}y_t = (1 + \theta_1 L)(1 + \theta_2 L^{12})z_t,$$

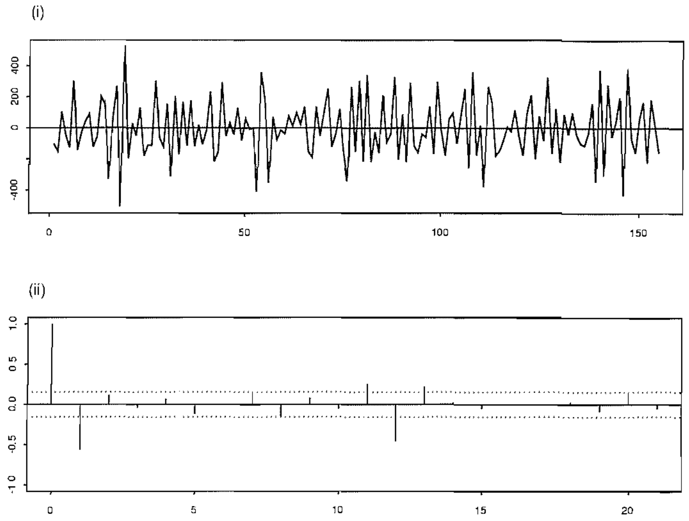

Figure 3.1: (i) Differenced Scottish data, (ii) Correlogram of differenced series.

where L is the backward shift operator which has the effect that $Ly_t = y_{t-1}$; here, we assume $z_t \sim N(0, \sigma^2)$. The form of the airline model is mutiplicative so that when multiplied out, the moving average terms not only model dependencies at lags 1 and 12 but also at lag 13:

$$\Delta\Delta_{12}y_t = z_t + \theta_1 z_{t-1} + \theta_2 z_{t-12} + \theta_1\theta_2 z_{t-13}.$$

The airline model was originally used to forecast passenger numbers on aeroplanes but has been found to be appropriate in many social and economic time series since.

Fitting the airline model to the Scottish deaths and injuries data gives $\hat{\theta}_1 = -0.8019$ and $\hat{\theta}_2 = -0.7988$, with standard errors 0.04818 and 0.04851 respectively; the model variance is $\hat{\sigma}^2 = 12066$. When the model is written out in full, taking into account the initial differencing, we arrive at

$$y_t = y_{t-1} - 0.8019z_{t-1} + y_{t-12} - 0.7988z_{t-12} - y_{t-13} + 0.6406z_{t-13} + z_t.$$

Residual plots for this model look reasonable (fig 3.2); the model seems to have successfully accounted for the correlation structure in the data as there are no significant autocorrelations in the correlogram.

Although ARIMA models like the airline model are easy to fit, their interpretation is not so clear as models which contain explanatory variables, this is especially true of models with moving average terms, i.e., ones that use past error terms to explain the current observation, such as the airline model. However, ARIMA models are mainly used for predictive purposes and in this arena they excel in providing accurate short term predictions. The topic of prediction does not really relate to the topics under investigation in this thesis and as such it is not covered here. However, the referenced texts provide extensive information on predicting with ARIMA models.

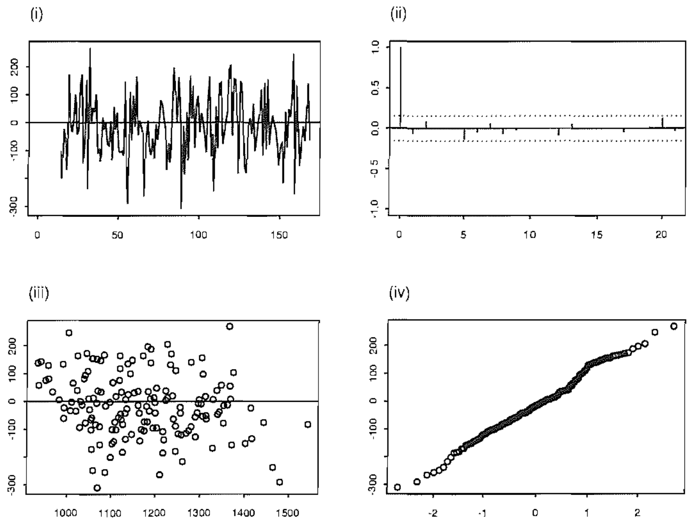

Figure 3.2: Residual plots from airline model fitted to Scottish data: (i) time chart of residuals, (ii) correlogram of residuals, (iii) residuals vs fits plot, (iv) normal plot of residuals.

Chapter 4

Estimating the effect of daylight on road accidents

In chapter 2, one of the explanatory variables used to model the Scottish and Southwest English car occupant accident data was the length of day from sunrise to sunset. It was found that the length of day was a significant predictor for the numbers of car occupant accidents and fatalities in all of the models fitted. In this chapter an alternative method for examining the effect of daylight on road accidents is investigated based on the difference in latitude between Scotland and Southwest England. Aspects of regression and ARIMA modelling are used in the presentation and analysis of the idea.

4.1 Differences in daylight during morning rush hour due to latitude

Latitude is perhaps the most important contributor to the amount of daylight an area or region receives throughout the year; the nearer the equator one is, the less annual variation in daylight there will be, so that the length of a day in one month is much the same as the length of a day in any other month of the year. Nearer the poles, however, there are more marked changes in the length of a day throughout the year, so that beyond the arctic or antarctic circles there is 24 hours of daylight in mid-summer and 24 hours of darkness in mid-winter. The UK is located between 50 and 60 degrees north and so experiences quite marked seasonal variation in the number of daylight hours. Despite its small size, the amount of seasonal variation in daylight in the north is noticeably greater than in the south, this can mean more than an hours difference in the number of daylight hours received in Scotland (the UK's northern most region) when compared to the Southwest of England (the UK's southern most region) during mid-winter and mid-summer.

Figure 4.1 shows the Monday to Friday weekday road traffic distribution from the year 2000 which shows a bimodal distribution with peaks between 8 and 9 am, and 5 and 6 pm; this distribution is most likely caused by the influence of rush hour times during the average working weekday. Broughton et al. (1999) showed that the distribution of vehicle casualties throughout the average day of the week also follows a bimodal distribution which peaks in the morning between 8 and 9 am, and in the evening between 5 and 6 pm. It would seem, therefore, that accidents are more likely to occur during times of heavy traffic, which are for the average day of the week during the morning and evening rush hours.

Figure 4.2 shows the monthly average sunrise and sunset times for Scotland and Southwest England with the peak morning and evening rush hour traffic times marked on as straight lines throughout the year. From this graph it can be seen that throughout most of the year there is generally no sizeable difference in the amount of daylight received in Scotland and Southwest England during the rush hour times. The only potentially significant difference occurs during the winter months of December and January at morning rush hour time. The graph shows that in Scotland the sun has not yet risen during the morning rush hour in December and January, and only rises above the horizon at the end of the rush hour period at almost 9am. So it is likely that most of the morning rush hour in Scotland for these months is spent in what is known as 'civil twilight' when

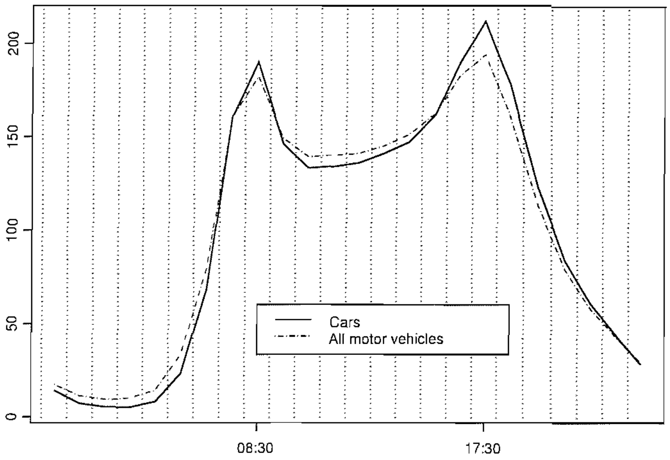

Figure 4.1: Weekday UK traffic distribution by time of day (year 2000), average = 100

the sun's altitude is between -0.83 and -6 degrees; this is the altitude at which large terrestrial objects but no detail can be distinguished according to Broughton et al. (1999). In Southwest England, however, the sun has risen by morning rush hour the whole year round.

For the accident series investigated in chapter 2 we know that daylight is a significant predictor for the total number of car occupant deaths and injuries, and also for just car occupant fatalities, in both Scotland and Southwest England. It is therefore possible that the darkness in morning rush hour in December and January in Scotland has an effect on road accidents which does not exist in Southwest England. The following sections pursue this idea and attempt to use it as an alternative way of measuring the effect of daylight on road accidents.

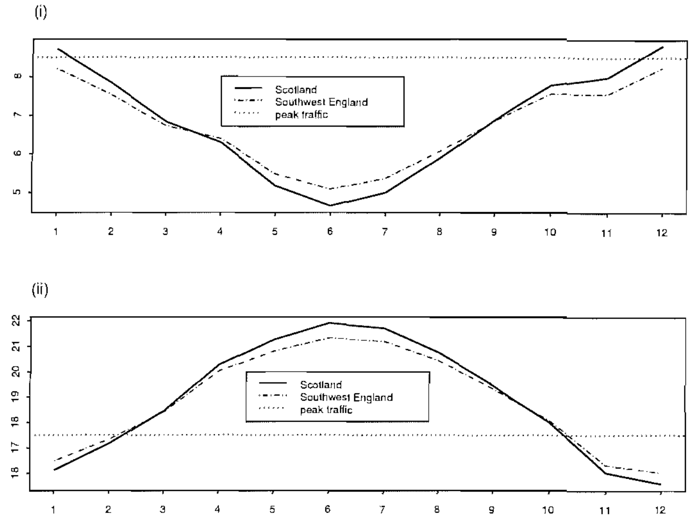

Figure 4.2: Monthly sunrise and sunset times for Scotland and Southwest England throughout the year: (i) sunrise, (ii) sunset.

4.2 Estimating the effect of daylight on road accidents using morning rush hour and latitude

4.2.1 The double-differencing proposal

Here, we use the ideas set out in the previous section to propose an alternative method to regression modelling for measuring the effect of daylight on road accidents. The method proposed is ad hoc and somewhat crude compared to the more sophisticated regression models used in chapter 2; also, it can only really be used to assess whether daylight has an effect on road accidents rather than how strong that effect is. However, this sort of methodology is useful for getting to grips with a problem when no, or very few, explanatory variable are available. Indeed, the idea of using latitude and rush hour to assess the effect of daylight was initially conceived at a time when explanatory variables were proving difficult to obtain for use in this thesis.

From the ideas in the previous section, if we look at the change in the number of road accidents from November to December, in theory there should be a steeper rise in accidents, or at least a smaller fall in accidents, in Scotland than in Southwest England due to the darker morning rush hour time in December in Scotland. Similarly, the change from January to February in Scotland should show a steeper fall, or at least a smaller rise, in accidents than in Southwest England, due to the lighter morning rush hour time in February compared to January in Scotland. The change in other factors which determine the number of accidents should be minimised by comparing December and January with their nearest neighbours, November and February.

To quantify this idea we can do the following: For Scotland and Southwest England, the difference in accident counts from November to December can be measured by subtracting the number of accidents in November from the number in December for each of the 14 years of the study; the Scottish differences should be larger (more positive) than the Southwest differences. Similarly, the difference in accident counts from January to February can be measured by subtracting the number of accidents in February from the number in January for each of the 14 years of the study; here, too, the Scottish differences should be larger than the Southwest differences. To assess whether the Scottish differences are larger than the Southwest differences we may again difference but this time spatially between the two regions; all 28 Southwest differences may be subtracted from the 28 Scottish differences. Since most covariates should not change much from one region to the other, spatially differencing will again reduce the effect that they have, hopefully leaving just the morning rush hour effect. The spatial differencing should result in 28 mainly positive quantities that could be called double-differences as they are a result of differencing twice, first in time and then spatially. A t-test could be used to establish whether the mean of the double-differences is significantly different from zero.

Mathematically, the proposal is as follows: let $y_{s,t}$ be a Scottish observation and $y_{w,t}$ be a corresponding Southwest English observation. Then the quantity δ_i, which is differenced in time and spatially, as above, is given by

$$\delta_i = \begin{cases} \Delta y_{s,t} - \Delta y_{w,t}, & t = 12, 24, ..., 156, 168, \\ -\Delta y_{s,t} + \Delta y_{w,t}, & t = 2, 14, ..., 146, 158. \end{cases} \tag{4.1}$$

Here, we use the Δ notation for differencing from chapter 3. To obtain the desired January/February differences using this notation, we subtract the January observations from the February observations and then multiply by -1. All together there should be 28 values of δ_i. If the mean, $\bar{\delta}$, is significantly greater than zero then it would suggest that the numbers of accidents in Scotland in December and January are significantly higher when compared to the norm (November and February) than in Southwest England. This would mean that the darkness in morning rush hour in December and January in Scotland significantly adds to the numbers of car occupant accidents, which would imply that daylight has a significant effect on car occupant accidents. The converse would mean that the darkness in morning rush hour in December and January in Scotland does not have a strong enough effect to significantly increase accidents; this would not imply that there is no daylight effect, as we already know that there is from chapter 2.

4.2.2 Results and analysis of the double-differencing method

Using the Scottish and Southwest fatalities data to calculate the values of δ_i in (4.1), we obtain a mean of $\bar{\delta} = 0.1071$, which shows that the Scottish differences are larger than the Southwest differences on average. The standard error on the mean is $s.e.(\bar{\delta}) = 1.760$ and therefore a t-test shows insufficient evidence that the mean double difference is significantly greater than zero. Using the Scottish and Southwest deaths and injuries data to calculate the values of δ_i in (4.1), we obtain a mean of $\bar{\delta} = 5.893$ with standard error $s.e.(\bar{\delta}) = 24.79$, which again shows that although the

Scottish differences are larger than the Southwest differences on average; the difference is not significant. This evidence shows that the double-differencing method based on latitude and the time of morning rush hour applied to raw data, as it has been, does not show a significant difference between Scotland and Southwest England in the November/December and January/February changes in car occupant accidents. This would imply that the darkness in morning rush hour in December and January in Scotland is not a strong enough effect to make a significant impact on casualty numbers.

Of course, this analysis is quite crude. Although the double-differencing technique has been designed to minimise the risk of other factors interfering with the estimation of the effect of daylight, it is possible that other factors are having an effect on the analysis leading the mean double difference in both cases to not be significantly greater than zero. To check whether our conclusions are right, we really need to remove the other factors from the data and then apply the double-differencing method to what is left. As it is, we are already in the position of knowing the explanatory variables which determine the means of each car occupant accident series from chapter 2.

From the linear models of chapter 2, removal of the non-daylight explanatory variables from the data is straightforward: we simply subtract them from the data and all that remains is the random variation plus daylight, upon which double-differencing is performed. For each of the models of chapter 2, we have

$$\begin{aligned}
r_{sf,t} &= y_{sf,t} - 29.12 + 0.05345t - 0.8012c_{s,t}, \\
r_{wf,t} &= y_{wf,t} + 4.575 + 0.09542t - 1.269a_t, \\
r_{st,t} &= y_{st,t} + 377.2 - 53.38a_t - 0.0003841p_t + 0.0005328d_t - 1.398f_{s,t} \\
& \quad -11.98C_{s,t}, \\
r_{wt,t} &= y_{st,t} - 182.3 - 45.44a_t - 1.524f_{w,t} - 13.41c_{w,t} + 32.18k_{w,t}.
\end{aligned}$$

Here, $y_{sf,t}$ denotes Scottish fatalities, $y_{wf,t}$ denotes Southwest fatalities, $y_{st,t}$ denotes Scottish deaths and injuries, and $y_{wt,t}$ denotes Southwest deaths and injuries. Also, a_t is the national car traffic variable, p_t is national petrol deliveries and d_t is national diesel deliveries. The regional variable, $C_{s,t}$, is the Scottish maximum temperature variable, while $c_{s,t}$ and $c_{w,t}$ are the Scottish and Southwest minimum temperature variables respectively. The Scottish and Southwest rainfall variables are $f_{s,t}$ and $f_{w,t}$ respectively, and the Southwest cloud cover variable is k_{w_t}.

From the two fatalities series, $r_{sf,t}$ and $r_{wf,t}$, a mean of $\bar{\delta} = -1.481$ with standard error $s.e.(\bar{\delta}) = 1.834$ is obtained by applying the double-differencing method. Similarly, from the two deaths and injuries series, $r_{st,t}$ and $r_{wt,t}$, a mean of $\bar{\delta} = -9.666$ with standard error $s.e.(\bar{\delta}) = 28.95$ is obtained. Again, after this followup analysis on the double-differencing technique, we are led to the same conclusions as before, that there is no significant difference between Scotland and Southwest England in the November/December and January/February changes in car occupant casualties. This implies that the darkness in morning rush hour in December and January in Scotland is not a strong enough effect to make a significant difference to casualty numbers.

4.2.3 Further analysis

A surprising result can be found from testing the double-differencing method on the residuals of the regression models of chapter 2. Looking at the November/December and January/February differences in the residuals of these models shows that, for both Scotland and Southwest England, these differences are significantly greater than zero. That is, if we calculate

$$d_i = \begin{cases} \Delta\epsilon_t, & t = 12, 24, ..., 156, 168, \\ -\Delta\epsilon_t, & t = 2, 14, ..., 146, 158, \end{cases}$$

for either Scotland or Southwest England, then $\bar{d} > 0$. For example, using the linear models of chapter 2 we find that for the Scottish deaths and injuries model, $\bar{d} = 41.39$ with $s.e.(\bar{d}) = 19.14$, for the Southwest deaths and injuries model, $\bar{d} = 81.33$ with $s.e.(\bar{d}) = 24.51$, and for the Southwest fatalities model, $\bar{d} = 2.517$ with $s.e.(\bar{d}) = 1.224$. The only linear model for which $\bar{d}$ is not significantly greater than zero is the Scottish fatalities model where $\bar{d} = 0.4647$ with $s.e.(\bar{d}) = 0.9927$.

This is a surprising result since the regression models of chapter 2 should all contain sufficient explanatory variables, including the daylight variable, to eliminate systematic variation like this. More importantly, it shows that removing the non-daylight related explanatory variables, as we did in the previous subsection, will not leave only the effect of daylight and random variation, it may also leave sources of unaccounted for variation. Unaccounted for sources of variation are an inevitability with linear or log-linear regression modelling, as it is very unlikely that we will have explanatory variables for all the possible influencing factors on a data set. What is unfortunate about the unaccounted for variation mentioned above is that this particular source of variation arises when looking at November/December and January/February differences and it is therefore bound to interfere with the analysis of the double-differencing method.

Chapter 5

Gaussian state space time series models

In this chapter we look at an alternative form of time series modelling, *state space* modelling. The first half of the chapter is a simplified summary of state space modelling and gives the key results needed for the examples in the second half of the chapter. A thorough treatment of state space modelling with the derivation of the Kalman filter and smoother can be found in Durbin and Koopman (2001) and Harvey (1989); Hamilton (1994) also covers state space modelling in chapter 13. Throughout most of the chapter, the presentation is given in terms of multivariate observation data since there is so little difference between the presentation of the univariate and multivariate cases. Also, some of the expressions will be needed in multivariate form for the techniques involving zero inflated data presented in chapter 7.

Two applications of state space time series modelling are presented. Section 5.3 gives a summary of Harvey and Durbin (1986), the case study on the introduction of compulsory seat belt wearing in the front seats of cars in 1983; and in §5.4 a Gaussian state space model is applied to the Scottish deaths and injuries data from chapter 2.

5.1 State space modelling overview

5.1.1 Structural time series models

Structural time series models are formulated directly in terms of the components of interest. If the series being analysed has certain components that we wish to include such as a trend, seasonal or explanatory variables, then these are modelled as separate components much like an ordinary regression model. For this reason, structural time series may have more intuitive appeal than say ARIMA models which seek to eliminate the effect of trends and seasonals through differencing to obtain a stationary series upon which subsequent analysis is performed.

The simplest formulation for a univariate structural time series, with observations $y_1, ..., y_n$, is the local level model, a good description of which is given in Durbin and Koopman (2001), chapter 2:

$$\begin{aligned} y_t &= \mu_t + \epsilon_t, \\ \mu_{t+1} &= \mu_t + \xi_t. \end{aligned} \tag{5.1}$$

Here, ϵ_t is the error or disturbance term and μ_t is the level which varies over time and is itself defined as a random walk with its own error term, ξ_t. The error terms, ϵ_t and ξ_t, are independently and identically Gaussian distributed, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ and $\xi_t \sim N(0, \sigma_\xi^2)$. Since there are no coefficients to estimate in this model, the only quantities we need to estimate here are the variances, σ_ϵ^2 and σ_ξ^2.

A time series with all the most common structural components in it and observed explanatory variables would be formulated as

$$y_t = \mu_t + \gamma_t + c_t + \sum_{j=1}^{b} \beta_j x_{j,t} + \epsilon_t, \qquad t = 1, ..., n, \tag{5.2}$$

where γ_t is a seasonal term, c_t is a cyclic component and there are b explanatory variables, $x_{j,t}$, with coefficients $\beta_{j,t}$ which could themselves vary over time. For this model we could also include a time varying slope component, ν_t, within the trend, μ_t:

$$\begin{aligned}\mu_{t+1} &= \mu_t + \nu_t + \xi_t, \\ \nu_{t+1} &= \nu_t + \zeta_t.\end{aligned} \tag{5.3}$$

This would give another random error, ζ_t, which would again be independently and identically Gaussian distributed, $\zeta_t \sim N(0, \sigma_\zeta^2)$. This formulation for μ_t is known as the local linear trend formulation. We can see that ν_t acts as a slope component since in the deterministic case, where $\zeta_t = 0$, $\nu_{t+1} = \nu_t = \nu$ and $\mu_{t+1} = \mu_t + \nu + \xi_t$, so μ_t increases by ν as t increases by 1 providing a linear trend.

If the coefficients, β_j, are time varying then most commonly they have a form just like μ_t in the local level model, in that they are defined simply as a random walk,

$$\beta_{j,t} = \beta_{j,t-1} + \vartheta_{j,t},$$

with $\vartheta_{j,t} \sim N(0, \sigma_{j,\vartheta}^2)$ for the Gaussian case. Different formulations can be found in Harvey (1989), §7.7. In state space time series literature it is uncommon to find examples of time varying coefficients for explanatory variables. Allowing an explanatory variable coefficient to vary with time is only done if we wish to examine the change in the effect the explanatory variable has on the observations over time. Usually, however, the purpose of modelling is to define a fixed relationship between the observation variable and the explanatory variables as is the case for regression modelling. Putting the explanatory variables into a structural time series model, we have the advantage of being able to use time varying structural terms to account for the random variation in the data unaccounted for by the explanatory variables.

The seasonal term can be modelled in two different ways. One way is the seasonal dummy form. Here, we suppose there are s seasons per year, so for monthly data $s = 12$, for quarterly, $s = 4$ and so on. First, assuming the seasonal is constant over time, the seasonal values for months 1 to s can be modelled by the constants $\gamma_1^*, ..., \gamma_s^*$, where $\sum_{j=1}^{s} \gamma_j^* = 0$. For the jth month in year i we have $\gamma_t = \gamma_j^*$, where $t = s(i-1)+j$ for $i = 1, 2, ...$ and $j = 1, ..., s$. It follows that $\sum_{j=0}^{s-1} \gamma_{t+1-j} = 0$, so $\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j}$ with $t = s-1, s, ...$ Of course, since we wish to allow all terms to vary over time, we can let the seasonal term vary over time by adding an error term, ω_t, to this relation giving the model

$$\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j} + \omega_t, \qquad \omega_t \sim N(0, \sigma_\omega^2), \qquad t = 1, ..., n. \tag{5.4}$$

A suitable trigonometric form to use for the seasonal term is the quasi-random walk model, equation (3.6) in Durbin and Koopman (2001),

$$\gamma_t = \sum_{j=1}^{\lfloor s/2 \rfloor} \gamma_{jt}, \tag{5.5}$$

where

$$\begin{aligned} \gamma_{j,t+1} &= \gamma_{jt}\cos(\lambda_j) + \gamma_{jt}^*\sin(\lambda_j) + \omega_{jt}, \\ \gamma_{j,t+1}^* &= -\gamma_{jt}\sin(\lambda_j) + \gamma_{jt}^*\cos(\lambda_j) + \omega_{jt}^*, \qquad j = 1, ..., \lfloor s/2 \rfloor. \\ \lambda_j &= \tfrac{2\pi j}{s}, \end{aligned} \tag{5.6}$$

Here, the ω_{jt} and ω_{jt}^* terms are independent $N(0, \sigma_\omega^2)$ variables.

Cyclic terms are common in economic time series but are not widely used in applications such as the analysis of road accidents. However, if it is necessary to add a cyclic term, c_t, to the model, the form would be very similar to the trigonometric seasonal, (5.6), with

$$\begin{aligned} c_{t+1} &= c_t\cos(\lambda_c) + c_t^*\sin(\lambda_c) + \hat{\omega}_t, \\ c_{t+1}^* &= -c_t\sin(\lambda_c) + c_t^*\cos(\lambda_c) + \hat{\omega}_t^*, \end{aligned}$$

where $\hat{\omega}_t$ and $\hat{\omega}_t^*$ are independent $N(0, \sigma_{\hat{\omega}}^2)$ variables. Here, the λ_c can be treated as an unknown parameter to be estimated.

5.1.2 The state space form

Once a structural time series model has been specified, the next stage is to convert it into state space form. For now, this is what we shall concentrate on; the subject of model building and goodness-of-fit will be covered in §5.2.3. Converting the structural model to state space form is necessary since the Kalman filter and smoother, which are used to find the fitted states and fitted model, can only be applied to the state space form of the time series.

The general linear Gaussian state space model can be written in several different ways. The form used in this thesis is the form given below and used in Durbin and Koopman (2001); the reason for which is that it more closely relates to the state space form used in S-Plus, which has been used for all calculations in this thesis. In the general multivariate case, the state space equations are given by

$$\begin{array}{rclrcll} \boldsymbol{y}_t & = & \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, & \boldsymbol{\epsilon}_t & \sim & N(\boldsymbol{0}, \boldsymbol{H}_t), & \\ \boldsymbol{\alpha}_{t+1} & = & \boldsymbol{T}_t\boldsymbol{\alpha}_t + \boldsymbol{R}_t\boldsymbol{\eta}_t, & \boldsymbol{\eta}_t & \sim & N(\boldsymbol{0}, \boldsymbol{Q}_t), & t = 1, ..., n, \\ & & & \boldsymbol{\alpha}_1 & \sim & N(\boldsymbol{a}_1, \boldsymbol{P}_1). & \end{array} \quad (5.7)$$

Here, $\boldsymbol{y}_t$ is a $p \times 1$ vector of observations, where p is the number of observation variables; and $\boldsymbol{\alpha}_t$ is an unobserved $m \times 1$ vector called the state vector, where m is the total number of states for all p observation variables. The matrices, $\boldsymbol{Z}_t$, $\boldsymbol{T}_t$, $\boldsymbol{R}_t$, $\boldsymbol{H}_t$ and $\boldsymbol{Q}_t$, are known as the system matrices and have dimensions $p \times m$, $m \times m$, $m \times r$, $p \times p$ and $r \times r$ respectively. The error terms, $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$, are assumed to be independent of one another and have dimensions $p \times 1$ and $r \times 1$ respectively, where r is the number of time varying states. The top line of (5.7) is known as the measurement equation, or observation equation, and the bottom line is known as the transition equation, or state equation. The terms observation and state are somewhat clearer since, after all, $\boldsymbol{y}_t$ is the observation vector and $\boldsymbol{\alpha}_t$ is the state vector, so these terms shall be used in this thesis.

The form of the matrices, $\boldsymbol{Z}_t$, $\boldsymbol{T}_t$, $\boldsymbol{H}_t$, $\boldsymbol{Q}_t$ and $\boldsymbol{R}_t$, is dictated by the components included in the structural time series model. For example, consider the state space form of the local linear trend structural model, i.e., the model with trend expressed as level and slope:

$$\begin{array}{rcl@{\qquad}rcl}
y_t &=& \mu_t + \epsilon_t, & \epsilon_t &\sim& N(0,\sigma^2_\epsilon), \\
\mu_{t+1} &=& \mu_t + \nu_t + \xi_t, & \xi_t &\sim& N(0,\sigma^2_\xi), \\
\nu_{t+1} &=& \nu_t + \zeta_t, & \zeta_t &\sim& N(0,\sigma^2_\zeta).
\end{array} \tag{5.8}$$

This model has the state space form

$$\begin{array}{rcl}
y_t &=& \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \epsilon_t, \\
\begin{pmatrix} \mu_{t+1} \\ \nu_{t+1} \end{pmatrix} &=& \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \xi_t \\ \zeta_t \end{pmatrix},
\end{array} \tag{5.9}$$

where

$$H_t = \sigma^2_\epsilon, \qquad \boldsymbol{Q_t} = \begin{bmatrix} \sigma^2_\xi & 0 \\ 0 & \sigma^2_\zeta \end{bmatrix}.$$

The matrix $\boldsymbol{R_t}$ is known as the selection matrix since it selects which rows of the state equation should have non-zero disturbance terms; it is made up from all or some of the columns of the identity matrix. In some specifications of the state space model $\boldsymbol{R_t}$ is not included and the disturbance vector, $\boldsymbol{\eta_t}$, has dimensions $m \times 1$ and the variance matrix, $\boldsymbol{Q_t}$, has dimensions $m \times m$ where errors or error variances which are non-time-varying are simply given as zero in $\boldsymbol{\eta_t}$ and $\boldsymbol{Q_t}$.

Consider now another example of a state space form where the local linear trend model, (5.8), also has a seasonal component. In state space form the seasonal state space vectors and matrices are combined with the state space vectors and matrices of (5.9) in the following way:

$$\begin{array}{rcl@{\qquad}rcl}
\boldsymbol{Z_t} &=& (\boldsymbol{Z}_t^{[\mu]} \ \boldsymbol{Z}_t^{[\gamma]}), & \boldsymbol{\alpha_t} &=& (\boldsymbol{\alpha}_t^{[\mu]\prime} \ \boldsymbol{\alpha}_t^{[\gamma]\prime})', \\
\boldsymbol{T_t} &=& diag(\boldsymbol{T}_t^{[\mu]}, \boldsymbol{T}_t^{[\gamma]}), & \boldsymbol{R_t} &=& diag(\boldsymbol{R}_t^{[\mu]}, \boldsymbol{R}_t^{[\gamma]}), \\
\boldsymbol{\eta_t} &=& (\boldsymbol{\eta}_t^{[\mu]\prime} \ \boldsymbol{\eta}_t^{[\gamma]\prime})', & \boldsymbol{Q_t} &=& diag(\boldsymbol{Q}_t^{[\mu]}, \boldsymbol{Q}_t^{[\gamma]}).
\end{array} \tag{5.10}$$

Here, the $(\boldsymbol{.})_t^{[\mu]}$ matrices refer to the various matrices of the local linear trend model, (5.9), and the $(\boldsymbol{.})_t^{[\gamma]}$ matrices refer to the seasonal parts that are combined with them.

The forms of the $(\boldsymbol{.})_t^{[\gamma]}$ matrices vary according to whether we are using a trigonometric seasonal component or a seasonal dummy component. For the seasonal dummy specification, (5.4), we have

$$\boldsymbol{Z}_t^{[\gamma]} = (1 \;\; 0 \; ... \; 0), \qquad \boldsymbol{\alpha}_t^{[\gamma]} = (\gamma_t \;\; \gamma_{t-1} \; \cdots \; \gamma_{t-s+2})',$$

$$\boldsymbol{T}_t^{[\gamma]} = \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \qquad \boldsymbol{R}_t^{[\gamma]} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$\eta_t^{[\gamma]} = \omega_t, \qquad \boldsymbol{Q}_t^{[\gamma]} = \sigma_\omega^2,$$

where $\boldsymbol{\alpha}_t^{[\gamma]}$ and $\boldsymbol{R}_t^{[\gamma]}$ have dimensions $(s-1) \times 1$, $\boldsymbol{Z}_t^{[\gamma]}$ has dimensions $1 \times (s-1)$ and $\boldsymbol{T}_t^{[\gamma]}$ has dimensions $(s-1) \times (s-1)$. Recall that s is the number of seasons.

For the seasonal trigonometric formulation, (5.5), we have the $(s-1) \times 1$ vectors $\boldsymbol{\alpha}_t^{[\gamma]} = (\gamma_{1t} \;\; \gamma_{1t}^* \;\; \gamma_{2t} \; ...)'$ and $\boldsymbol{\eta}_t^{[\gamma]} = (\omega_{1t} \;\; \omega_{1t}^* \;\; \omega_{2t} \; ...)'$. We also have the system matrices $\boldsymbol{R}_t^{[\gamma]} = \boldsymbol{I}_{s-1}$ and $\boldsymbol{Q}_t^{[\gamma]} = \sigma_\omega^2 \boldsymbol{I}_{s-1}$ where $\boldsymbol{I}_{s-1}$ is the $(s-1) \times (s-1)$ identity matrix. However, the matrices $\boldsymbol{Z}_t^{[\gamma]}$ and $\boldsymbol{T}_t^{[\gamma]}$ vary according to whether s is odd or even. For even s we have

$$\boldsymbol{Z}_t^{[\gamma]} = (1 \;\; 0 \;\; 1 \;\; 0 \;\; 1 \; ... \; 1 \;\; 0 \;\; 1), \qquad \boldsymbol{T}_t^{[\gamma]} = diag(\boldsymbol{C}_1, ..., \boldsymbol{C}_{s^*}, -1).$$

For odd s we have

$$\boldsymbol{Z}_t^{[\gamma]} = (1 \;\; 0 \;\; 1 \;\; 0 \;\; 1 \; ... \; 1 \;\; 0), \qquad \boldsymbol{T}_t^{[\gamma]} = diag(\boldsymbol{C}_1, ..., \boldsymbol{C}_{s^*}).$$

For both cases the quantities s^* and $\boldsymbol{C_j}$ are

$$s^* = \lfloor \frac{s-1}{2} \rfloor, \quad \boldsymbol{C_j} = \begin{bmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{bmatrix}, \quad \lambda_j = \frac{2\pi j}{s}, \quad j = 1, ..., \lfloor s/2 \rfloor.$$

We shall see a full example of the state space form with a trigonometric seasonal component in §5.3. Note that any other components, such as explanatory variables, are added to the existing state space matrices in the same way as the seasonal components are added in (5.10).

For many applications, such as the cases illustrated above, the system matrices do not vary with time and so the subscript t's need not necessarily be included with them. In fact, for standard Gaussian state space models the matrices $\boldsymbol{H_t}$ and $\boldsymbol{Q_t}$ are always constant and it is partly the purpose of the model estimation process to calculate estimates of the variance parameters contained within those matrices. The variance parameters and any other unknown parameters are put into a parameter vector, $\boldsymbol{\psi}$, which is estimated by the numerical maximisation of the log-likelihood; for example, with the local linear trend model, (5.3), $\boldsymbol{\psi} = (\sigma_\epsilon^2 \;\; \sigma_\xi^2 \;\; \sigma_\zeta^2)'$. However, to obtain the log-likelihood we must first examine the Kalman filter and smoother which are used to find the fitted states and fitted model. Note that the coefficients of explanatory variables are not included among the parameters to be estimated by maximising the log-likelihood since these are expressed as states, even if they are non-time-varying, and so are estimated by the Kalman filter and smoother which handle the fitting of states.

5.2 Model fitting, parameter estimation and goodness-of-fit

5.2.1 The Kalman filter and smoother

The Kalman filter

The Kalman Filter (Kalman, 1960) is an important set of recursions which provides the basis for calculating the quantities of interest from the state space model. The Kalman Filter computes two quantities: $\boldsymbol{a}_t$ and $\boldsymbol{P}_t$, which are the mean and variance respectively of the state vector, $\boldsymbol{\alpha}_t$, given the past observations up to time t. Hence, $\boldsymbol{a}_t$ and $\boldsymbol{P}_t$ are given by

$$\boldsymbol{a}_t = E(\boldsymbol{\alpha}_t|\boldsymbol{y}_T), \qquad \boldsymbol{P}_t = Var(\boldsymbol{\alpha}_t|\boldsymbol{y}_T),$$

where $\boldsymbol{y}_T$ is the stacked vector of past observations up to time t or observation vectors $(\boldsymbol{y}_1' \dots \boldsymbol{y}_{t-1}')'$ up to time t in the general multivariate case. So, in the case of Gaussian distributed state errors, which we shall be dealing with for the rest of this thesis, we have $\boldsymbol{\alpha}_t|\boldsymbol{y}_T \sim N(\boldsymbol{a}_t, \boldsymbol{P}_t)$. Strictly speaking, the expressions above should not only be dependent on $\boldsymbol{y}_T$ but also be dependent on the unknown parameter vector $\boldsymbol{\psi}$, giving $E(\boldsymbol{\alpha}_t|\boldsymbol{y}_T, \boldsymbol{\psi})$ and $Var(\boldsymbol{\alpha}_t|\boldsymbol{y}_T, \boldsymbol{\psi})$. However, since all the expressions in the coming chapters depend on $\boldsymbol{\psi}$, we lose nothing by not explicitly expressing this dependence; also, not including $\boldsymbol{\psi}$ simplifies the notation.

Below, the Kalman Filter recursions for calculating $\boldsymbol{a}_t$ and $\boldsymbol{P}_t$ are given:

$$\begin{array}{rclrcll} \boldsymbol{\nu}_t &=& \boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{a}_t, & \boldsymbol{F}_t &=& \boldsymbol{Z}_t\boldsymbol{P}_t\boldsymbol{Z}_t' + \boldsymbol{H}_t, & \\ \boldsymbol{K}_t &=& \boldsymbol{T}_t\boldsymbol{P}_t\boldsymbol{Z}_t'\boldsymbol{F}_t^{-1}, & \boldsymbol{L}_t &=& \boldsymbol{T}_t - \boldsymbol{K}_t\boldsymbol{Z}_t, & t = 1, ..., n. \\ \boldsymbol{a}_{t+1} &=& \boldsymbol{T}_t\boldsymbol{a}_t + \boldsymbol{K}_t\boldsymbol{\nu}_t, & \boldsymbol{P}_{t+1} &=& \boldsymbol{T}_t\boldsymbol{P}_t\boldsymbol{L}_t' + \boldsymbol{R}_t\boldsymbol{Q}_t\boldsymbol{R}_t', & \end{array} \tag{5.11}$$

The Kalman filter is initialised with the quantities $\boldsymbol{a}_1$ and $\boldsymbol{P}_1$, the mean and variance of the initial state, $\boldsymbol{\alpha}_1$, from (5.7). If $\boldsymbol{a}_1$ and $\boldsymbol{P}_1$ are known, the Kalman filter can simply be run from $t = 1, ..., T$ and this case presents no complications. In practice, however, $\boldsymbol{a}_1$ and $\boldsymbol{P}_1$ are usually unknown so a suitable initialisation technique must be used to start the Kalman filter recursions.

When the initial state vector, $\boldsymbol{\alpha}_1$, is unknown or has unknown mean and variance, it is treated as a diffuse random variable, that is, it is assumed to have infinite variance. To get round this problem it is possible to make an approximation to the variance of $\boldsymbol{\alpha}_1$ by taking $\boldsymbol{P}_1 = \mathbf{10}^7$, i.e., a large number as an approximation to infinity and here we would take $\boldsymbol{a}_1 = \mathbf{0}$. But this method can lead to large rounding errors so special diffuse initialisation methods are used. The theory behind diffuse initialisation methods is to eliminate the infinite variance problem by changing the form of the first few terms of the Kalman filter. The initial state variance matrix, $\boldsymbol{P}_1$, is assumed to take the following form:

$$\boldsymbol{P}_1 = \kappa \boldsymbol{P}_\infty + \boldsymbol{P}_*,$$

where $\kappa \rightarrow \infty$. Initialisation methods alter the first few terms of the Kalman filter so that κ does not come into play.

There are essentially two initialisation methods available, the Exact Initial method and the Augmented approach; the exact initial method is the technique that S-Plus uses. While initialisation methods are important, they do not feature as the subject of investigation in the rest of this thesis and so coverage of these methods is not given here since it is impossible to give a summary of these methods without going into pages and pages of detail; however, chapter 5 of Durbin and Koopman (2001) covers both the exact initial and the augmented approach in depth.

The Kalman smoother

The Kalman filter can be adapted slightly to give the Kalman smoother. The Kalman filter calculates $\boldsymbol{a}_t = E(\boldsymbol{\alpha}_t|\boldsymbol{y}_T)$ and $\boldsymbol{P}_t = Var(\boldsymbol{\alpha}_t|\boldsymbol{y}_T)$, the expectation and variance of $\boldsymbol{\alpha}_t$ given the past observations in the time series $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}$, up to time t. The Kalman smoother, however, calculates the mean and variance of $\boldsymbol{\alpha}_t$ given all the information in the time series, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. If we take $\boldsymbol{y} = (\boldsymbol{y}_1' \ldots \boldsymbol{y}_n')'$, which is the stacked vector of all the observation vectors, then the Kalman smoother calculates $E(\boldsymbol{\alpha}_t|\boldsymbol{y})$ and $Var(\boldsymbol{\alpha}_t|\boldsymbol{y})$.

In the literature $E(\boldsymbol{\alpha}_t|\boldsymbol{y})$ is often called an optimal estimate and is denoted $\hat{\boldsymbol{\alpha}}_t$. This is because $\boldsymbol{\alpha}_t|\boldsymbol{y}$ has a Gaussian distribution and so its mean is equal to its mode and the mode is in some sense the maximum likelihood estimate. However, we try to avoid this notation where possible since it would only seem appropriate once we have calculated the maximum likelihood estimate of $\boldsymbol{\psi}$, i.e., one would expect that $\hat{\boldsymbol{\alpha}}_t = E(\boldsymbol{\alpha}_t|\boldsymbol{y}, \hat{\boldsymbol{\psi}})$ and not $\hat{\boldsymbol{\alpha}}_t = E(\boldsymbol{\alpha}_t|\boldsymbol{y}, \boldsymbol{\psi})$. The fact that the mean of $\boldsymbol{\alpha}_t|\boldsymbol{y}$ is equal to it's mode will be of some importance when dealing with non-Gaussian observations in the next chapter.

The Kalman smoother recursions are given by

$$\begin{aligned} \boldsymbol{r}_{t-1} &= \boldsymbol{Z}_t'\boldsymbol{F}_t^{-1}\boldsymbol{\nu}_t + \boldsymbol{L}_t'\boldsymbol{r}_t, & \boldsymbol{N}_{t-1} &= \boldsymbol{Z}_t'\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t + \boldsymbol{L}_t'\boldsymbol{N}_t\boldsymbol{L}_t, \\ E(\boldsymbol{\alpha}_t|\boldsymbol{y}) &= \boldsymbol{a}_t + \boldsymbol{P}_t\boldsymbol{r}_{t-1}, & Var(\boldsymbol{\alpha}_t|\boldsymbol{y}) &= \boldsymbol{P}_t - \boldsymbol{P}_t\boldsymbol{N}_{t-1}\boldsymbol{P}_t. \end{aligned} \tag{5.12}$$

The recursions are initialised with $\boldsymbol{r}_n = \boldsymbol{0}$ and $\boldsymbol{N}_n = \boldsymbol{0}$ and the rest of the quantities used are obtained from the Kalman filter. Since the Kalman smoother moves backwards, ending with $\boldsymbol{r}_0$ and $\boldsymbol{N}_0$, the Kalman filter must always be run first for the whole series to obtain quantities such as $\boldsymbol{F}_t$, for all n, before the smoother can be started.

In practice it is often the case in analysis, as we shall see in chapter 7, that we do not wish to estimate $\boldsymbol{\alpha}_t$ but rather the response and state disturbances: $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$. These can, of course, be derived from the Kalman smoother, but for completeness we give these quantities together with their variances below:

$$\begin{aligned} E(\epsilon_t|\boldsymbol{y}) &= \boldsymbol{H}_t(\boldsymbol{F}_t^{-1}\boldsymbol{\nu}_t - \boldsymbol{K}_t'\boldsymbol{r}_t), \\ Var(\epsilon_t|\boldsymbol{y}) &= \boldsymbol{H}_t - \boldsymbol{H}_t(\boldsymbol{F}_t^{-1} + \boldsymbol{K}_t'\boldsymbol{N}_t\boldsymbol{K}_t)\boldsymbol{H}_t, \\ E(\boldsymbol{\eta}_t|\boldsymbol{y}) &= \boldsymbol{Q}_t\boldsymbol{R}_t'\boldsymbol{r}_t, \\ Var(\boldsymbol{\eta}_t|\boldsymbol{y}) &= \boldsymbol{Q}_t - \boldsymbol{Q}_t\boldsymbol{R}_t'\boldsymbol{N}_t\boldsymbol{R}_t\boldsymbol{Q}_t. \end{aligned} \tag{5.13}$$

5.2.2 The log-likelihood

Presented below is the Gaussian log-likelihood used for diffuse initialisation via the exact initial approach, and specifically, for the case when $\boldsymbol{F}_{\infty,t}$ is positive definite as it is for all the models considered in this thesis.

$$\log[L(\boldsymbol{y}|\boldsymbol{\psi})] = -\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{d}\log|\boldsymbol{F}_{\infty,t}| - \frac{1}{2}\sum_{t=d+1}^{n}(\log|\boldsymbol{F}_t| + \boldsymbol{\nu}_t'\boldsymbol{F}_t^{-1}\boldsymbol{\nu}_t).$$

Here, d is the number of states whose initial variances are unknown. When initial conditions are known, the part involving $\boldsymbol{F}_{\infty,t}$ can be omitted since in this case $d = 0$. The dependence on $\boldsymbol{\psi}$ is expressed in the log-likelihood since it is this vector which is to be estimated.

The log-likelihood is maximised via a numerical optimisation procedure which alters the values of $\boldsymbol{\psi}$ for each iteration until convergence whereupon the optimal values, $\hat{\boldsymbol{\psi}}$, are found. The starting values of $\boldsymbol{\psi}$ before the log-likelihood is maximised can essentially be educated guesses unless prior knowledge is available. FinMetrics uses the SsfFit function to calculate and maximise the Gaussian log-likelihood, given a state space form and initial parameter estimates. The numerical hessian and its inverse are also calculated by this function in S-Plus version 7 and FinMetrics version 2, although not in previous versions. The standard errors of the parameters can be calculated in the usual way from the square root of the diagonal elements of the inverse hessian. Durbin and Koopman (2001), §7.3.2, gives details of maximization procedures upon which the SsfFit function is based. A description of the functions I wrote which incorporate the SsfFit function for the estimation of Gaussian state space models is given in appendix A.

5.2.3 Model building and goodness-of-fit

Generally, structural models are initially specified after examination of the time plot of the data which should point to obvious features such as trends and seasonal patterns. Whether there are explanatory variables available or not, it is probably best to initially specify a model with a full complement of structural terms and then refine the model by eliminating terms which do not contribute to the model fit. If there is a trend present it is best to specify the trend initially to include both level and slope components. Also, it is sensible to make all terms in the model time varying to start with, i.e., start with the most complex model. Generally, if explanatory variables are not included in the model and we are modelling data that does not include cycles in it, then model building begins with the following model:

$$y_t = \mu_t + \gamma_t + \epsilon_t, \qquad t = 1, ..., n, \tag{5.14}$$

where μ_t is the local linear trend of (5.3) and γ_t is either the dummy variable seasonal component of (5.4) or the trigonometric seasonal component of (5.5). This model is referred to as the *basic structural model* by Harvey (1989) and it shall often be referred to during the course of the remaining chapters of this thesis.

The next stage is the estimation of parameters from the initial model so that the model may be fitted and assessed for its appropriateness and goodness-of-fit. It is possible to use a likelihood ratio test in exactly the same way as for a generalised linear model, using the state space log-likelihood in place of the GLM log-likelihood; however, this is usually not necessary for the structural terms in the model as, generally, it will be fairly clear what changes to the structural components need to be made. For instance, if the variance estimate of a seasonal component is approximately zero or has a large standard error, a model with a fixed seasonal term could simplify the initial specification; or if a series has very little trend, a trend term with a time varying level only and without a slope component may be adequate. The likelihood ratio test is probably best used for explanatory variables although even here it may not be necessary since commonly in time series analysis there are one or two explanatory variables available which are known to affect the observed data and we wish to examine the effect of these variables, so they are unlikely to be removed from the model regardless of fit. However, these explanatory variables are not usually enough to explain all the variation in the data and so the structural terms are used to model the other features of the time series; the structural terms themselves do not have a direct interpretation.

In terms of assessing the goodness-of-fit of a particular model, we may use residual plots; although, for state space time series models there is more than one type of residual we may examine. For a Gaussian state space model there will always be the residuals from the observation equation, $\hat{\epsilon}_t$, but there may also be various residuals from the state equation such as $\hat{\xi}_t$, $\hat{\zeta}_t$ and $\hat{\omega}_t$, the level, slope and seasonal residuals respectively. To save examination of all of these different residuals, the standardised prediction error residuals or standardised *innovations* may be used. The standardised innovations are derived from the Kalman filter and are given by

$$e_t = \nu_t/(F_t)^{1/2}.$$

The assumptions underlying state space models are that all the residuals, whether they be model residuals or level residuals or any other, are iid Gaussian. Under these assumptions the standardised innovations are also iid Gaussian which is why they are frequently used to measure goodness of fit.

5.3 Harvey and Durbin's seat belt study

State space time series methods were successfully used in the study of the compulsory introduction of seat belt wearing in the front seats of vehicles in the UK (Harvey and Durbin, 1986). The aim of the study was to see whether there had been a significant change in the numbers of casualties, for various road user groups, since the introduction of the law. Of primary interest was the change in the number of car drivers killed or seriously injured (KSI) since the seat belt law affected them particularly directly. The study was carried out on monthly data from January 1969 to December 1984 and the compulsory wearing of seat belts in front seats became law on 31st of January 1983. Harvey and Durbin (1986) investigated a variety of formulations for this problem; a fairly straightforward formulation is given in Durbin and Koopman (2001) and this is what we shall show here.

The introduction of the seat belt law was formulated as an explanatory variable which takes the value zero from January 1969 to January 1983 and then the value one from February 1983 to December 1984. Analysis was performed on the log of the data and the log of the price of petrol was included as an explanatory variable along with the seat belt introduction variable. To explain the rest of the random variation, a time varying level, $\mu_t = \mu_{t-1} + \xi_t$, was used (as in (5.1)) along with a trigonometric seasonal term, (5.5). Thus, the finalized structural time series model for the car drivers KSI data was given by

$$y_t = \mu_t + \gamma_t + \beta_1 p_t + \beta_2 \lambda_t + \epsilon_t,$$

where y_t is the log of the observations, p_t is the log of the price of petrol, λ_t is the seat belt variable and all random variables are Gaussian distributed. From the theory of §5.1.2, this model can be written in state space form as follows:

$$\begin{aligned}
\boldsymbol{\alpha}_t &= (\mu_t \;\; \gamma_{1t} \;\; \gamma^*_{1t} \;\; \gamma_{2t} \cdots \gamma^*_{5t} \;\; \gamma_{6t} \;\; \beta_1 \;\; \beta_2)', \\
\boldsymbol{T}_t &= \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\
0 & 0.866 & 0.5 & & 0 & 0 & 0 & 0 & 0 \\
0 & -0.5 & 0.866 & & 0 & 0 & 0 & 0 & 0 \\
\vdots & & & \ddots & & & & & \vdots \\
0 & 0 & 0 & & -0.866 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0 & & -0.5 & -0.866 & 0 & 0 & 0 \\
0 & 0 & 0 & & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 1
\end{bmatrix},
\end{aligned}$$

$$\boldsymbol{\eta}_t = (\xi_t \;\; \omega_{1t} \;\; \omega^*_{1t} \;\; \omega_{2t} \ldots \omega^*_{5t} \;\; \omega_{6t})',$$

$$\boldsymbol{R}_t = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & & 1 \\ 0 & 0 & & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \boldsymbol{Q}_t = \begin{bmatrix} \sigma^2_\xi & 0 & \cdots & 0 \\ 0 & \sigma^2_\omega & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_\omega \end{bmatrix},$$

$$\boldsymbol{Z}_t = (1 \;\; 1 \;\; 0 \;\; 1 \;\; 0 \ldots 1 \;\; 0 \;\; 1 \;\; p_t \;\; \lambda_t), \qquad H_t = \sigma^2_\epsilon.$$

The state space model was estimated via the Kalman Filter and Smoother and the log-likelihood was maximised. To avoid numerical instability in the optimisation of the log-likelihood, the parameter estimates calculated by the optimisation procedure are actually the logs of the variance parameter estimates; it is these log-estimates that the standard errors are calculated on. So the table below shows the parameter estimates as well as log-parameter estimates and log-parameter standard errors. The variance estimates for the three random errors are all small, but since the seasonal variance is so small it shows that the seasonal term could probably be treated as a constant without loss of model fit.

	parameter	log-parameter	log-parm Std err
$\hat{\sigma}^2_\epsilon$	3.788×10^{-3}	-5.576	0.1517
$\hat{\sigma}^2_\xi$	2.676×10^{-4}	-8.226	0.6056
$\hat{\sigma}^2_\omega$	1.157×10^{-6}	-13.67	1.206

Table 5.1: Seat belt study variance parameter estimates, log-parameter estimates and standard errors on log-parameter estimates.

The coefficient estimate for the log of the price of petrol was found to be $\hat{\beta}_1 = -0.2914$ with standard error $s.e(\hat{\beta}_1) = 0.09832$. This indicated a reduction in the number of car drivers killed or seriously injured of 0.29% for an increase of 1% in the price of petrol. The seat belt introduction coefficient was found to be $\hat{\beta}_2 = -0.2377$ with standard error

$s.e(\hat{\beta}_2) = 0.04632$. This indicated a reduction in car drivers killed or seriously injured of 21% due to the introduction of seat belts. Figure 5.1 illustrates the effect of the introduction of seat belts by plotting the level and seat belt variable over the data; the other plots are residual plots of the standardised innovations.

5.4 Application of state space modelling to Scottish deaths and injuries data

The explanatory variables used in the regression models of chapter 2 for the Scottish and Southwest deaths and injuries data do not fully account for all the variation in the data since in the log-linear models there is a large amount of over-dispersion, and in both the linear and log-linear models there is a slight fluctuation of the residuals plotted over time about the zero line which has been left unaccounted for. There are several structural time

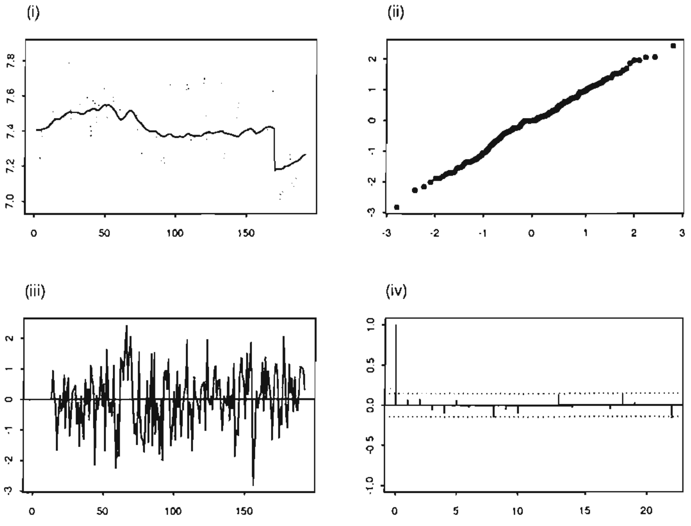

Figure 5.1: Plots for the seat belt study: (i) estimated level with seat belt variable, (ii) normal plot of standardised innovations, (iii) time chart of standardised innovations, (iv) correlogram of standardised innovations.

series choices that could be used to tackle this issue. One way could be to make some of the coefficients of the explanatory variables time varying; this, logically, could be quite sensible since it is quite possible that the effect rainfall has on road accidents, for example, could change over time due to improvements in tyre grip and traction. However, without a good knowledge of tyres we cannot be certain whether this would be so or not. A simpler stochastic way of dealing with the unaccounted variation might be to include a stochastic term for the level of the series, as in (5.1), rather than in one of the explanatory variable coefficients. This would act in a similar way as a linear trend term would act, measuring the changes in the underlying level of risk but hopefully more effectively. Written as a structural time series the model would simply be

$$y_t = \mu_t + \sum_{j=1}^{m-1} \beta_j x_{j,t} + \epsilon_t,$$

where the time varying level would replace the constant, which is usually present in all regression models, leaving only the exogenous explanatory variables.

Putting the above structural model into state space form, and using the Kalman filter and smother to derive the fitted model, then optimising the fit using the SsfFit function in S-Plus, gives the parameter and coefficient estimates in the tables below. Again, optimisation is carried out using the log-parameters in the SsfFit function and so standard errors are calculated on the log-parameters. The explanatory variable coefficients in table 5.3 are very similar to those of the linear regression model.

	parameter	log parameter	log-parm Std err
$\hat{\sigma}^2_\epsilon$	8596	9.059	0.1249
$\hat{\sigma}^2_\xi$	222.9	5.407	0.6928

Table 5.2: Scottish deaths and injuries state space model variance parameter estimates, log-parameter estimates and standard errors on log-parameter estimates.

Variable	Coefficient	Std err
Nat car traffic	51.67	10.52
Nat petrol del	5.722 $\times 10^{-4}$	1.037 $\times 10^{-4}$
Nat diesel del	-1.337 $\times 10^{-3}$	2.198 $\times 10^{-4}$
Total rainfall	1.540	0.2766
Max daily temp	10.61	4.118
Daylight	-35.47	4.087

Table 5.3: Scottish deaths and injuries state space model covariate coefficients and standard errors.

The time series residuals which are most directly comparable with the residuals of the linear model of chapter 2 are the estimated model residuals, $\hat{\epsilon}_t$. Figure 5.2 shows that there seems to be a slight improvement in the variability of the time chart of model residuals around the zero line, compared to figure 2.3 for the linear regression model.

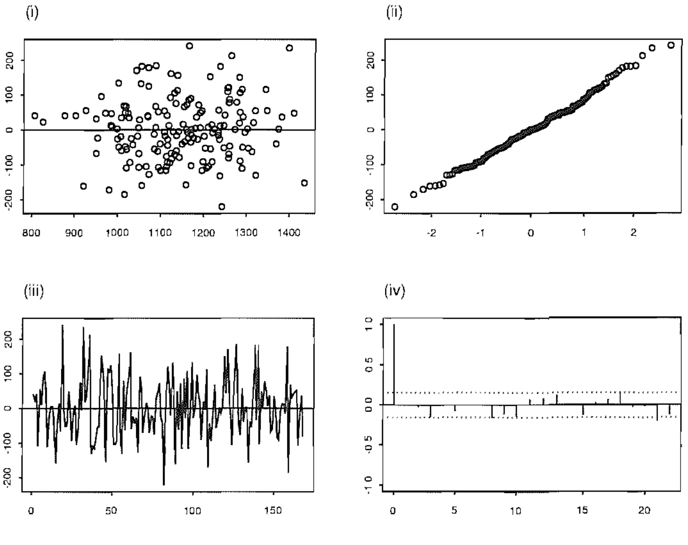

Figure 5.2: Scottish model residual, $\hat{\epsilon}_t$, plots: (i) residuals vs fitted values, (ii) normal plot, (iii) time chart of residuals, (iv) correlogram

Chapter 6

Non-Gaussian state space models

The methods presented in this chapter are concerned with state space models in which the marginal distribution of the observations is non-Gaussian and in particular from the exponential family of distributions. The Exponential family includes the Poisson distribution and so relating the means of counts to structural terms and explanatory variables by a log-link gives the state space analogue of the log-linear model used in regression. In §6.5 the Scottish car occupant fatality data is investigated again and two log-linear state space models are fitted to the data, one with only structural terms and the other with a mix of explanatory variables and structural terms. Durbin and Koopman (2001) is the primary reference for this chapter.

6.1 The non-Gaussian state space modelling approach

The non-Gaussian state space modelling approach is somewhat more involved than that of Gaussian state space modelling. For Gaussian modelling, firstly, a structural time series model is specified, then it is written in state space form, next Kalman filtering and smoothing are applied to the state space form and the log-likelihood is derived. The log-likelihood is maximised by a numerical method which alters the values in the parameter vector, $\boldsymbol{\psi}$. The final parameter estimates are then used in the Kalman filter and smoother to derive the fitted states and fitted model. Finally, the model is checked using diagnostic tests and if necessary it is modified.

For non-Gaussian state space modelling the specification of the structural model and state space form is slightly different and the derivation of the log-likelihood is much more complicated and relies on simulation as well as Kalman filtering and smoothing. However, once the log-likelihood has been obtained, the parameter values in the unknown parameter vector, $\boldsymbol{\psi}$, are updated by a numerical maximisation procedure in the same way as the Gaussian log-likelihood. Perhaps the main difference here is that $\boldsymbol{\psi}$ may not contain a model variance parameter, σ_ϵ^2, such as in the Poisson case.

6.1.1 The formulation of structural and state space models

The state space structure of the general multivariate non-Gaussian model is similar to that of the Gaussian model, (5.7), in that the observations are determined by a relationship of the form

$$f(\boldsymbol{y}_t|\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_t, \boldsymbol{y}_1, ..., \boldsymbol{y}_{t-1}) = f(\boldsymbol{y}_t|\boldsymbol{Z}_t\boldsymbol{\alpha}_t). \tag{6.1}$$

The state vectors are determined independently of previous observations by the relationship

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{T}_t\boldsymbol{\alpha}_t + \boldsymbol{R}_t\boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim f(\boldsymbol{\eta}_t), \tag{6.2}$$

for $t = 1, ..., n$, where the $\boldsymbol{\eta}_t$'s are serially independent and where either $f(\boldsymbol{y}_t|\boldsymbol{Z}_t\boldsymbol{\alpha}_t)$ or $f(\boldsymbol{\eta}_t)$ or both may be non-Gaussian.

In this chapter we shall be dealing with observations distributed according to the exponential family of distributions, which in the general multivariate case is given by

$$f(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = \exp\{\boldsymbol{y}_t'\boldsymbol{\theta}_t - b_t(\boldsymbol{\theta}_t) + c_t(\boldsymbol{y}_t)\}, \qquad -\infty < \boldsymbol{\theta}_t < \infty. \tag{6.3}$$

Rather confusingly, $\boldsymbol{\theta}_t$, which always denotes the canonical parameter in generalised linear models has a different meaning in non-Gaussian state space models; in these models $\boldsymbol{\theta}_t = \boldsymbol{Z}_t\boldsymbol{\alpha}_t$ and is known as the signal. In this respect $\boldsymbol{\theta}_t$ in non-Gaussian state space models plays an identical role to $\boldsymbol{\eta}_i$, the linear predictor, in generalised linear models.

Unfortunately, for many non-Gaussian models, including most from the exponential family, it is not possible to write down an observation equation in quite the straightforward manner that is used for the Gaussian state space form, (5.7). So, when writing down the structural model that we wish to analyse, we must express the structural terms and explanatory variables through the signal. For example, say we are modelling a univariate time series using a non-Gaussian state space model with a time varying level and seasonal dummy term. The trend and seasonal terms would be expressed through the signal, θ_t, and then the formulation for the trend and seasonal terms themselves would be exactly the same as for the Gaussian model except for the possibility of non-Gaussian distributed error terms:

$$\begin{aligned}
\theta_t &= \mu_t + \gamma_t, & \\
\mu_{t+1} &= \mu_t + \xi_t, & \xi_t &\sim f(\xi_t), \\
\gamma_{t+1} &= -\textstyle\sum_{j=1}^{s-1}\gamma_{t+1-j} + \omega_t, & \omega_t &\sim f(\omega_t).
\end{aligned}$$

Once a time series model has been specified, the fitting and estimation procedure is complicated as it involves repeated Kalman filtering and smoothing as well as simulation, the details are given in the following subsections and sections.

6.1.2 Overview of importance sampling

State space modelling cannot be called straightforward even in the case of Gaussian models, but in the non-Gaussian case it is even more involved. As with Gaussian models, the idea is to calculate $E(\boldsymbol{\alpha}_t|\boldsymbol{y})$ and $Var(\boldsymbol{\alpha}_t|\boldsymbol{y})$, the mean and variance respectively of $\boldsymbol{\alpha}_t|\boldsymbol{y}$. This objective can be achieved by finding $E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\}$, the mean of a function, $x(\boldsymbol{\alpha})$, of $\boldsymbol{\alpha}$ given the observation vector $\boldsymbol{y}$, where in the multivariate case we take $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1' \dots \boldsymbol{\alpha}_n')'$ and $\boldsymbol{y} = (\boldsymbol{y}_1' \dots \boldsymbol{y}_n')'$, i.e., as stacked vectors. We start by giving the integral definition for this mean:

$$E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\} = \int x(\boldsymbol{\alpha}) f(\boldsymbol{\alpha}|\boldsymbol{y}) d\boldsymbol{\alpha}. \tag{6.4}$$

The $x(\boldsymbol{\alpha})$ formulation leads to options such as taking $x(\boldsymbol{\alpha}) = \boldsymbol{\alpha}_t$ to obtain the mean $E(\boldsymbol{\alpha}_t|\boldsymbol{y})$, or $x(\boldsymbol{\alpha}) = \{\boldsymbol{\alpha}_t - E(\boldsymbol{\alpha}_t|\boldsymbol{y})\}^2$ to obtain the variance $Var(\boldsymbol{\alpha}_t|\boldsymbol{y})$. Note that as for the Gaussian state space models of the previous chapter, all quantities are conditional on the unobserved parameter vector $\boldsymbol{\psi}$. However, since there is no quantity that is not conditional on $\boldsymbol{\psi}$, nothing is lost when comparing one quantity with another by not including it in the notation.

Ideally at this stage we would draw a random sample from the density $f(\boldsymbol{\alpha}|\boldsymbol{y})$ and estimate $E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\}$ by the sample mean of the values of $x(\boldsymbol{\alpha})$; however, the practice is rather different. The problem is that since $f(\boldsymbol{\alpha}|\boldsymbol{y})$ cannot be written in an explicit form, we cannot sample from it. So, the trick is to choose a density as close to $f(\boldsymbol{\alpha}|\boldsymbol{y})$ as possible for which random draws are available and sample from this instead, making an appropriate adjustment to the integral of (6.4). This density is called the importance density and the technique of sampling from it is called importance sampling.

A Gaussian density, $g(\boldsymbol{\alpha}|\boldsymbol{y})$, is chosen as the importance density since random draws are available from Gaussian densities. With $g(\boldsymbol{\alpha}|\boldsymbol{y})$, the following adjustment to the integral (6.4) is made:

$$E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\} = \int x(\boldsymbol{\alpha}) \frac{f(\boldsymbol{\alpha}|\boldsymbol{y})}{g(\boldsymbol{\alpha}|\boldsymbol{y})} g(\boldsymbol{\alpha}|\boldsymbol{y}) d\boldsymbol{\alpha} = E_g\left\{x(\boldsymbol{\alpha}) \frac{f(\boldsymbol{\alpha}|\boldsymbol{y})}{g(\boldsymbol{\alpha}|\boldsymbol{y})}\right\}, \tag{6.5}$$

where E_g denotes expectation with respect to the importance density $g(\boldsymbol{\alpha}|\boldsymbol{y})$. The densities $f(\boldsymbol{\alpha}|\boldsymbol{y})$ and $g(\boldsymbol{\alpha}|\boldsymbol{y})$ can be algebraically complicated, whereas, the corresponding joint densities, $f(\boldsymbol{\alpha},\boldsymbol{y})$ and $g(\boldsymbol{\alpha},\boldsymbol{y})$, are generally straightforward. We therefore put $g(\boldsymbol{\alpha}|\boldsymbol{y}) = g(\boldsymbol{\alpha},\boldsymbol{y})/g(\boldsymbol{y})$ and $f(\boldsymbol{\alpha}|\boldsymbol{y}) = f(\boldsymbol{\alpha},\boldsymbol{y})/f(\boldsymbol{y})$ into (6.5) giving

$$E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\} = \frac{g(\boldsymbol{y})}{f(\boldsymbol{y})}E_g\left\{x(\boldsymbol{\alpha})\frac{f(\boldsymbol{\alpha},\boldsymbol{y})}{g(\boldsymbol{\alpha},\boldsymbol{y})}\right\}. \tag{6.6}$$

Putting $x(\boldsymbol{\alpha}) = 1$ in (6.6) gives

$$1 = \frac{g(\boldsymbol{y})}{f(\boldsymbol{y})}E_g\left\{\frac{f(\boldsymbol{\alpha},\boldsymbol{y})}{g(\boldsymbol{\alpha},\boldsymbol{y})}\right\}. \tag{6.7}$$

Finally, taking the ratio of (6.6) and (6.7) gives

$$E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\} = \frac{E_g\{x(\boldsymbol{\alpha})w(\boldsymbol{\alpha},\boldsymbol{y})\}}{E_g\{w(\boldsymbol{\alpha},\boldsymbol{y})\}}, \qquad w(\boldsymbol{\alpha},\boldsymbol{y}) = \frac{f(\boldsymbol{\alpha},\boldsymbol{y})}{g(\boldsymbol{\alpha},\boldsymbol{y})}. \tag{6.8}$$

These two equations now provide the basis for a solution. We draw a sample of N independent simulated state vectors, $\breve{\boldsymbol{\alpha}}^{(1)}, ..., \breve{\boldsymbol{\alpha}}^{(N)}$, where $\breve{\boldsymbol{\alpha}}$ denotes a simulated state vector, from the importance density $g(\boldsymbol{\alpha}|\boldsymbol{y})$ and use them in (6.8) to estimate $E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\}$ as follows:

$$\hat{E}\{x(\boldsymbol{\alpha})|\boldsymbol{y}\} = \frac{(1/N)\sum_{i=1}^{N}\boldsymbol{x_i}w_i}{(1/N)\sum_{i=1}^{N}w_i} = \frac{\sum_{i=1}^{N}\boldsymbol{x_i}w_i}{\sum_{i=1}^{N}w_i}, \tag{6.9}$$

where, $\boldsymbol{x_i} = x(\breve{\boldsymbol{\alpha}}^{(i)})$ and $w_i = w(\breve{\boldsymbol{\alpha}}^{(i)},\boldsymbol{y})$. Note that the conditional variance, $Var\{x(\boldsymbol{\alpha})|\boldsymbol{y}\}$, can simply be estimated by

$$\widehat{Var}\{x(\boldsymbol{\alpha})|\boldsymbol{y}\} = \frac{\sum_{i=1}^{N}\boldsymbol{x_i^2}w_i}{\sum_{i=1}^{N}w_i} - [\hat{E}\{x(\boldsymbol{\alpha})|\boldsymbol{y}\}]^2. \tag{6.10}$$

When $f(\boldsymbol{\eta_t}) = g(\boldsymbol{\eta_t})$ in (6.2), i.e., the state errors, $\boldsymbol{\eta_t}$, are Gaussian distributed, then the states themselves are also Gaussian distributed, so

$f(\boldsymbol{\alpha}) = g(\boldsymbol{\alpha})$. This allows $w(\boldsymbol{\alpha}, \boldsymbol{y})$ in (6.8) to be expressed as

$$w(\boldsymbol{\alpha}, \boldsymbol{y}) = \frac{f(\boldsymbol{\alpha}, \boldsymbol{y})}{g(\boldsymbol{\alpha}, \boldsymbol{y})} = \frac{f(\boldsymbol{\alpha})f(\boldsymbol{y}|\boldsymbol{\alpha})}{g(\boldsymbol{\alpha})g(\boldsymbol{y}|\boldsymbol{\alpha})} = \frac{f(\boldsymbol{y}|\boldsymbol{\alpha})}{g(\boldsymbol{y}|\boldsymbol{\alpha})} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})}{g(\boldsymbol{y}|\boldsymbol{\theta})}.$$

So instead of using (6.8), we may write

$$E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\} = \frac{E_g\{x(\boldsymbol{\alpha})w(\boldsymbol{y}|\boldsymbol{\theta})\}}{E_g\{w(\boldsymbol{y}|\boldsymbol{\theta})\}}, \qquad w(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})}{g(\boldsymbol{\epsilon})}, \tag{6.11}$$

where $g(\boldsymbol{\epsilon}) = g(\boldsymbol{y}|\boldsymbol{\theta})$. The advantage of using this expression over (6.8) is that the dimensionality of $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ is usually less than that of $\boldsymbol{\alpha}$; so we may draw a sample of simulated $\boldsymbol{\theta}$'s or $\boldsymbol{\epsilon}$'s for less computational cost than drawing a sample of simulated $\boldsymbol{\alpha}$'s.

6.1.3 Calculation of the log-likelihood

The likelihood function, $L(\boldsymbol{\psi})$, of the unknown parameter vector, $\boldsymbol{\psi}$, is derived in much the same way as $E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\}$ in (6.4):

$$L(\boldsymbol{\psi}) = \int f(\boldsymbol{\alpha}, \boldsymbol{y})d\boldsymbol{\alpha}. \tag{6.12}$$

Using the same manipulations as for the calculation of $E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\}$ gives

$$\begin{aligned} L(\boldsymbol{\psi}) &= \int \frac{f(\boldsymbol{\alpha}, \boldsymbol{y})}{g(\boldsymbol{\alpha}|\boldsymbol{y})} g(\boldsymbol{\alpha}|\boldsymbol{y})d\boldsymbol{\alpha} \\ &= g(\boldsymbol{y}) \int \frac{f(\boldsymbol{\alpha}, \boldsymbol{y})}{g(\boldsymbol{\alpha}, \boldsymbol{y})} g(\boldsymbol{\alpha}|\boldsymbol{y})d\boldsymbol{\alpha} \\ &= L_g(\boldsymbol{\psi})E_g\{w(\boldsymbol{\alpha}, \boldsymbol{y})\}, \end{aligned} \tag{6.13}$$

where $L_g(\boldsymbol{\psi}) = g(\boldsymbol{y})$ is the likelihood of the linear Gaussian approximating model used to obtain the importance density $g(\boldsymbol{\alpha}|\boldsymbol{y})$. When states are Gaussian distributed this likelihood can be expressed corresponding to (6.11) by $L(\boldsymbol{\psi}) = L_g(\boldsymbol{\psi})E_g\{w(\boldsymbol{y}|\boldsymbol{\theta})\}$.

The estimate of the log-likelihood may then be given as

$$\log\{\hat{L}(\boldsymbol{\psi})\} = \log\{L_g(\boldsymbol{\psi})\} + \log(\bar{w}), \tag{6.14}$$

where $\bar{w} = (1/N)\sum_{i=1}^{N} w_i$. To calculate an estimate for the unknown parameter vector, $\hat{\boldsymbol{\psi}}$, $\log \hat{L}(\boldsymbol{\psi})$ is maximised by a convenient numerical optimisation technique. Since the log-likelihood is non-Gaussian it cannot be maximised by the SsfFit function in FinMetrics used for the Gaussian state space models of the previous chapter; instead, the S-Plus function, nlminb, must be employed. The nlminb function is a more general function than SsfFit and must be passed a non-Gaussian log-likelihood function, as well as state space form, observations and initial parameter estimates, before it can calculate estimates for the unknown parameters. The calculation of the log-likelihood is rather complex as we shall see from the following sections; however, a description of the functions I wrote and the way in which they fit together, for the calculation of the log-likelihood and fitted model is given in appendix A.

Both the calculation of $E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\}$ and $\log\{L(\boldsymbol{\psi})\}$ depend on $w(\boldsymbol{\alpha}, \boldsymbol{y})$ which in turn depends on the construction of the importance density, $g(\boldsymbol{\alpha}|\boldsymbol{y})$, so that it is as close as possible to $f(\boldsymbol{\alpha}|\boldsymbol{y})$. In importance sampling as close as possible means choosing the importance density, $g(\boldsymbol{\alpha}|\boldsymbol{y})$, so that its mode is equal to that of $f(\boldsymbol{\alpha}|\boldsymbol{y})$. In the next section we will see how to equate these modes and in the following section we will see how to draw the sample of independent simulated vectors, $\breve{\boldsymbol{\alpha}}^{(1)}, ..., \breve{\boldsymbol{\alpha}}^{(N)}$, from $g(\boldsymbol{\alpha}|\boldsymbol{y})$.

6.2 Constructing $g(\boldsymbol{\alpha}|\boldsymbol{y})$ to be as close as possible to $f(\boldsymbol{\alpha}|\boldsymbol{y})$

6.2.1 The linear Gaussian approximating model

From the previous subsection, $g(\boldsymbol{y})$ is the likelihood of the linear Gaussian approximating model used to obtain the importance density $g(\boldsymbol{\alpha}|\boldsymbol{y})$. The linear Gaussian approximating model referred to here is a linear Gaussian state space given by

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, & \boldsymbol{\epsilon}_t &\sim N(\boldsymbol{\delta}_t, \boldsymbol{H}_t), & & \\ \boldsymbol{\alpha}_{t+1} &= \boldsymbol{T}_t\boldsymbol{\alpha}_t + \boldsymbol{R}_t\boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim N(\boldsymbol{0}, \boldsymbol{Q}_t), & t &= 1, ..., n, \\ & & \boldsymbol{\alpha}_1 &\sim N(\boldsymbol{a}_1, \boldsymbol{P}_1). & & \end{aligned} \tag{6.15}$$

This model is similar to (5.7) but allows greater flexibility in the errors since $E(\boldsymbol{\epsilon}_t) = \boldsymbol{\delta}_t$ rather than in (5.7) where $E(\boldsymbol{\epsilon}_t) = \boldsymbol{0}$ for all t. Note that here $\boldsymbol{\delta}_t$ has no connection to the double-differencing coefficient, δ_i, in chapter 4. The extra flexibility in the errors is needed when $f(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ is an exponential family distribution but is not needed if $f(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ is a symmetric or heavy tailed distribution. There is no similar change to the state error term $\boldsymbol{\eta}_t$ since for most cases in practice $f(\boldsymbol{\eta}_t)$ is a symmetric distribution, and for the cases we consider it will be Gaussian.

Recall that we wish to equate the mode of the importance density, $g(\boldsymbol{\alpha}|\boldsymbol{y})$, with that of $f(\boldsymbol{\alpha}|\boldsymbol{y})$ so that the densities will be as close as possible to one another. To accomplish this we can at least start by finding the mode of $g(\boldsymbol{\alpha}|\boldsymbol{y})$. Applying the Kalman filter and smoother to (6.15) for all values of t gives the mean $E(\boldsymbol{\alpha}|\boldsymbol{y})$ of $g(\boldsymbol{\alpha}|\boldsymbol{y})$. But since $g(\boldsymbol{\alpha}|\boldsymbol{y})$ is Gaussian, the mean is equal to the mode, so $E(\boldsymbol{\alpha}|\boldsymbol{y})$ is also the mode of $g(\boldsymbol{\alpha}|\boldsymbol{y})$. However, the Kalman filter and smoother cannot be appled to (6.15) without modification since they only work for state space models where $E(\boldsymbol{\epsilon}_t) = \boldsymbol{0}$ for all t. To solve this problem the observation equation of (6.15) can be modified to look more like the observation equation of (5.7) by transferring the term $\boldsymbol{\delta}_t$ from $\boldsymbol{\epsilon}_t$ to $\boldsymbol{y}_t$. Taking $\boldsymbol{\epsilon}_t^* = \boldsymbol{\epsilon}_t - \boldsymbol{\delta}_t$ and $\boldsymbol{y}_t^* = \boldsymbol{y}_t - \boldsymbol{\delta}_t$ means that $E(\boldsymbol{\epsilon}_t^*) = \boldsymbol{0}$, as in the standard Gaussian state space model, (5.7). Using $\boldsymbol{\epsilon}_t^*$ and $\boldsymbol{y}_t^*$ in the linear Gaussian approximating model gives

$$\begin{aligned} \boldsymbol{y}_t^* &= \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, & \boldsymbol{\epsilon}_t^* &\sim N(\boldsymbol{0}, \boldsymbol{H}_t), & & \\ \boldsymbol{\alpha}_{t+1} &= \boldsymbol{T}_t\boldsymbol{\alpha}_t + \boldsymbol{R}_t\boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim N(\boldsymbol{0}, \boldsymbol{Q}_t), & t &= 1, ..., n, \\ & & \boldsymbol{\alpha}_1 &\sim N(\boldsymbol{a}_1, \boldsymbol{P}_1). & & \end{aligned} \tag{6.16}$$

The two linear Gaussian approximating models, (6.15) and (6.16), are equivalent and so calculating $E(\boldsymbol{\alpha}|\boldsymbol{y}^*)$ from (6.16) gives the same as we would get if we were able to calculate $E(\boldsymbol{\alpha}|\boldsymbol{y})$ from (6.15). Therefore calculating $E(\boldsymbol{\alpha}|\boldsymbol{y}^*)$ gives the mode of the importance density $g(\boldsymbol{\alpha}|\boldsymbol{y})$.

However, it is still not possible to derive the mode of the importance density since $E(\boldsymbol{\alpha}|\boldsymbol{y}^*)$ cannot be calculated from model (6.16) until $\boldsymbol{\delta}_t$, and hence $\boldsymbol{y}_t^*$, is known. Also, due to the extra flexibility needed to equate the modes, $\boldsymbol{H}_t$ is no longer constant and $\boldsymbol{Q}_t$ may no longer be constant if the state errors from the state equation are non-Gaussian. In the following two subsections $\boldsymbol{H}_t$ and $\boldsymbol{y}_t^*$ are derived so that not only can $E(\boldsymbol{\alpha}|\boldsymbol{y}^*)$, the mode of the importance density, be calculated, but this mode will also be the mode of $f(\boldsymbol{\alpha}|\boldsymbol{y})$.

6.2.2 Deriving the mode equations for $g(\boldsymbol{\alpha}|\boldsymbol{y})$ and $f(\boldsymbol{\alpha}|\boldsymbol{y})$

The solution of the vector equation $\partial \log g(\boldsymbol{\alpha}|\boldsymbol{y})/\partial\boldsymbol{\alpha} = \mathbf{0}$ gives the mode of $g(\boldsymbol{\alpha}|\boldsymbol{y})$. As mentioned earlier, however, the function $g(\boldsymbol{\alpha}|\boldsymbol{y})$ is often complex algebraically, but since $\log\{g(\boldsymbol{\alpha}|\boldsymbol{y})\} = \log\{g(\boldsymbol{\alpha},\boldsymbol{y})\} - \log\{g(\boldsymbol{y})\}$, the mode of $g(\boldsymbol{\alpha}|\boldsymbol{y})$ is also the solution to $\partial \log g(\boldsymbol{\alpha},\boldsymbol{y})/\partial\boldsymbol{\alpha} = \mathbf{0}$, and it is this equation that shall be considered.

From the linear Gaussian approximating model, (6.16), we see that $g(\boldsymbol{\alpha},\boldsymbol{y}) = g(\boldsymbol{\alpha},\boldsymbol{y}^*)$, and therefore $\log\{g(\boldsymbol{\alpha},\boldsymbol{y})\}$ may be obtained from the unconditional densities of the linear Gaussian approximating model, $g(\boldsymbol{\alpha}_1) = N(\boldsymbol{a}_1, \boldsymbol{P}_1)$, $g(\boldsymbol{\alpha}_{t+1}) = N(\boldsymbol{T}_t\boldsymbol{\alpha}_t, \boldsymbol{Q}_t)$ and $g(\boldsymbol{y}_t^*) = N(\boldsymbol{Z}_t\boldsymbol{\alpha}_t, \boldsymbol{H}_t)$, as follows:

$$\begin{aligned}
\log\{g(\boldsymbol{\alpha},\boldsymbol{y})\} &= \log\{g(\boldsymbol{\alpha},\boldsymbol{y}^*)\} \\
&= \mathit{constant} - \frac{1}{2}(\boldsymbol{\alpha}_1 - \boldsymbol{a}_1)'\boldsymbol{P}_1^{-1}(\boldsymbol{\alpha}_1 - \boldsymbol{a}_1) \\
&\quad -\frac{1}{2}\sum_{t=1}^{n}(\boldsymbol{\alpha}_{t+1} - \boldsymbol{T}_t\boldsymbol{\alpha}_t)'\boldsymbol{R}_t\boldsymbol{Q}_t^{-1}\boldsymbol{R}_t'(\boldsymbol{\alpha}_{t+1} - \boldsymbol{T}_t\boldsymbol{\alpha}_t) \\
&\quad -\frac{1}{2}\sum_{t=1}^{n}(\boldsymbol{y}_t^* - \boldsymbol{Z}_t\boldsymbol{\alpha}_t)'\boldsymbol{H}_t^{-1}(\boldsymbol{y}_t^* - \boldsymbol{Z}_t\boldsymbol{\alpha}_t). \qquad (6.17)
\end{aligned}$$

Differentiating with respect to $\boldsymbol{\alpha}_t$ and equating to zero gives the equations

$$\begin{aligned}
\frac{\partial \log\{g(\boldsymbol{\alpha}, \boldsymbol{y})\}}{\partial \boldsymbol{\alpha}_t} &= (d_t - 1)\boldsymbol{P}_1^{-1}(\boldsymbol{\alpha}_1 - \boldsymbol{a}_1) \\
&\quad -d_t\boldsymbol{R}_{t-1}\boldsymbol{Q}_{t-1}^{-1}\boldsymbol{R}'_{t-1}(\boldsymbol{\alpha}_t - \boldsymbol{T}_{t-1}\boldsymbol{\alpha}_{t-1}) \\
&\quad +\boldsymbol{T}'_t\boldsymbol{R}_t\boldsymbol{Q}_t^{-1}\boldsymbol{R}'_r(\boldsymbol{\alpha}_{t+1} - \boldsymbol{T}_t\boldsymbol{\alpha}_t) + \boldsymbol{Z}'_t\boldsymbol{H}_t^{-1}(\boldsymbol{y}_t^* - \boldsymbol{Z}_t\boldsymbol{\alpha}_t) \\
&= \mathbf{0}, \\
\frac{\partial \log\{g(\boldsymbol{\alpha}, \boldsymbol{y})\}}{\partial \boldsymbol{\alpha}_{n+1}} &= \boldsymbol{R}_n\boldsymbol{Q}_n\boldsymbol{R}'_n(\boldsymbol{\alpha}_{n+1} - \boldsymbol{T}_n\boldsymbol{\alpha}_n) \\
&= \mathbf{0}, \qquad (6.18)
\end{aligned}$$

where $d_1 = 0$ and $d_t = 1$ for $t = 2, ..., n$.

Equivalently, for the non-Gaussian case we have

$$\log\{f(\boldsymbol{\alpha}, \boldsymbol{y})\} = \textit{constant} + \log\{f(\boldsymbol{\alpha}_1)\} - \sum_{t=1}^{n}\{q_t(\boldsymbol{\eta}_t) + h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)\}. \qquad (6.19)$$

For the above equation $h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = -\log\{f(\boldsymbol{y}_t|\boldsymbol{\theta}_t)\}$, $q_t(\boldsymbol{\eta}_t) = -\log\{f(\boldsymbol{\eta}_t)\}$ and $\boldsymbol{\eta}_t = \boldsymbol{R}'_t(\boldsymbol{\alpha}_{t+1} - \boldsymbol{T}_t\boldsymbol{\alpha}_t)$. Differentiating with respect to $\boldsymbol{\alpha}_t$ and equating to zero gives the equations

$$\begin{aligned}
\frac{\partial \log\{f(\boldsymbol{\alpha}, \boldsymbol{y})\}}{\partial \boldsymbol{\alpha}_t} &= (1 - d_t)\frac{\partial \log\{f(\boldsymbol{\alpha}_1)\}}{\partial \boldsymbol{\alpha}_1} - d_t\boldsymbol{R}_{t-1}\frac{\partial q_{t-1}(\boldsymbol{\eta}_{t-1})}{\partial \boldsymbol{\eta}_{t-1}} \\
&\quad +\boldsymbol{T}'_t\boldsymbol{R}_t\frac{\partial q_t(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t} - \boldsymbol{Z}'_t\frac{\partial h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t} \\
&= \mathbf{0}, \\
\frac{\partial \log\{f(\boldsymbol{\alpha}, \boldsymbol{y})\}}{\partial \boldsymbol{\alpha}_{n+1}} &= \boldsymbol{R}_n\frac{\partial q_n(\boldsymbol{\eta}_n)}{\partial \boldsymbol{\eta}_n} \\
&= \mathbf{0}, \qquad (6.20)
\end{aligned}$$

where, as before, $d_1 = 0$ and $d_t = 1$ for $t = 2, ..., n$.

Now that we have the mode equations for $g(\boldsymbol{\alpha}, \boldsymbol{y})$ and $f(\boldsymbol{\alpha}, \boldsymbol{y})$, we need to equate the mode of $g(\boldsymbol{\alpha}, \boldsymbol{y})$ with that of $f(\boldsymbol{\alpha}, \boldsymbol{y})$. When the state disturbance term, $\boldsymbol{\eta}_t$, is Gaussian distributed, as it is in the applications we consider, then all but the component involving $h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ in (6.20) are identical to the components of (6.18). Therefore to equate the modes we just need to choose $\boldsymbol{H}_t$ and $\boldsymbol{y}_t^*$ in (6.18) so that the term $\boldsymbol{Z}_t'\boldsymbol{H}_t^{-1}(\boldsymbol{y}_t^* - \boldsymbol{Z}_t\boldsymbol{\alpha}_t)$ provides the best approximation to $\boldsymbol{Z}_t'\partial h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)/\partial\boldsymbol{\theta}_t$ in (6.20). The next subsection will give details of how to linearise $h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ so that $\boldsymbol{H}_t$ and $\boldsymbol{y}_t^*$ may be derived.

Although most of the non-Gaussian modelling approach has been kept as general as possible so far, we shall not present the details of how to linearise $h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ when $f(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ is not from the exponential family since the linearisation process is quite different in that case. Also, details shall not be presented on the linearisation of $q_t(\boldsymbol{\eta}_t)$ since in the coming examples only Gaussian distributed state errors are considered. These linearisation techniques can be found in §11.5 and §11.6 of Durbin and Koopman (2001) respectively.

6.2.3 Linearisation of $h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ for exponential family distributions

There are several methods of linearising $h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ in equation (6.20); the method considered in this section is applicable to exponential family observations as well as observations from the stochastic volatility model, (Sandmann and Koopman (1998) and Zivot et al. (2003)).

Suppose that $\tilde{\boldsymbol{\alpha}} = (\tilde{\boldsymbol{\alpha}}_1 \ldots \tilde{\boldsymbol{\alpha}}_{n+1})'$ is a trial value of $\boldsymbol{\alpha}$; consequently $\tilde{\boldsymbol{\theta}}_t = \boldsymbol{Z}_t\tilde{\boldsymbol{\alpha}}_t$. Also, the following definitions are given

$$\dot{\boldsymbol{h}}_t = \left.\frac{\partial h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_t}\right|_{\boldsymbol{\theta}_t=\tilde{\boldsymbol{\theta}}_t}, \qquad \ddot{\boldsymbol{h}}_t = \left.\frac{\partial^2 h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_t\partial\boldsymbol{\theta}_t'}\right|_{\boldsymbol{\theta}_t=\tilde{\boldsymbol{\theta}}_t}. \tag{6.21}$$

A Taylor expansion about $\tilde{\boldsymbol{\theta}}_t$ gives, approximately,

$$\frac{\partial h_t(\boldsymbol{y}_t|\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_t} = \dot{\boldsymbol{h}}_t + \ddot{\boldsymbol{h}}_t(\boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}_t).$$

Substituting the above into the final term of (6.20) gives the linearised form

$$-\boldsymbol{Z}_t'(\dot{\boldsymbol{h}}_t + \ddot{\boldsymbol{h}}_t\boldsymbol{\theta}_t - \ddot{\boldsymbol{h}}_t\tilde{\boldsymbol{\theta}}_t).$$

To put this into the same format as the final term of (6.18) we take

$$\tilde{\boldsymbol{H}}_t = \ddot{\boldsymbol{h}}_t^{-1}, \qquad \tilde{\boldsymbol{y}}_t^* = \tilde{\boldsymbol{\theta}}_t - \ddot{\boldsymbol{h}}_t^{-1}\dot{\boldsymbol{h}}_t; \tag{6.22}$$

thus the final term then becomes $\boldsymbol{Z}_t'\tilde{\boldsymbol{H}}_t^{-1}(\tilde{\boldsymbol{y}}_t^* - \boldsymbol{\theta}_t)$ as required.

Now at last we have some trial values, $\tilde{\boldsymbol{H}}_t$ and $\tilde{\boldsymbol{y}}_t^*$, to use in the linear Gaussian approximating model, (6.16). The Kalman filter and smoother are applied to this model to calculate a new trial $\tilde{\boldsymbol{\theta}}_t$ obtained from $\tilde{\boldsymbol{\theta}}_t = E\{\boldsymbol{\theta}_t|\tilde{\boldsymbol{y}}^*\}$. This new $\tilde{\boldsymbol{\theta}}_t$ is then substituted into (6.22) to obtain a new $\tilde{\boldsymbol{H}}_t$ and $\tilde{\boldsymbol{y}}_t^*$ and again these updated values are substituted into the linear Gaussian approximating model, whereupon Kalman filtering and smoothing is carried out to obtain a new $\tilde{\boldsymbol{\theta}}_t$. This process is repeated over and over until $\tilde{\boldsymbol{H}}_t$ and $\tilde{\boldsymbol{y}}_t^*$ do not change with subsequent iterations. We now use $\tilde{\boldsymbol{H}}_t$ and $\tilde{\boldsymbol{y}}_t^*$ for $\boldsymbol{H}_t$ and $\boldsymbol{y}_t^*$ in the linear Gaussian approximating model (6.16) and apply the Kalman filter and smoother to derive $E(\boldsymbol{\alpha}|\boldsymbol{y}^*)$, the mode of both the Gaussian importance density, $g(\boldsymbol{\alpha}|\boldsymbol{y})$, and the non-Gaussian density, $f(\boldsymbol{\alpha}|\boldsymbol{y})$.

6.3 Simulation

The following is a summary of the results from Durbin and Koopman (2002) for constructing a sample of simulated state space disturbance vectors. This is the simulation smoother used in S-Plus. Recall that we wish to draw a sample of simulated state vectors, $\breve{\boldsymbol{\alpha}}^{(1)}, ..., \breve{\boldsymbol{\alpha}}^{(N)}$, from the importance density $g(\boldsymbol{\alpha}|\boldsymbol{y})$. Drawing simulated state vectors from the importance density, $g(\boldsymbol{\alpha}|\boldsymbol{y})$, is equivalent to drawing simulated disturbance vectors, $\breve{\boldsymbol{\varepsilon}}$, from the importance density $g(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$, where $\boldsymbol{\varepsilon}^* = (\boldsymbol{\epsilon}^{*\prime} \quad \boldsymbol{\eta}')'$ from the linear Gaussian approximating model (6.16) and $\breve{\boldsymbol{\varepsilon}}$ denotes a simulated $\boldsymbol{\varepsilon}^*$ vector. Once a sample of simulated disturbance vectors has been drawn they can be substituted back into the linear Gaussian approximating model to derive the desired sample of simulated state vectors from $g(\boldsymbol{\alpha}|\boldsymbol{y})$.

The density $g(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$ can be written $g(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*) = N\{E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*), Var(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)\}$. Here, $Var(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$ does not depend on $\boldsymbol{y}^*$ because the conditional variance matrix of a vector given that a second vector is fixed does not depend on the second vector, which can be seen for this particular case from $Var(\boldsymbol{\epsilon}_t|\boldsymbol{y})$ and $Var(\boldsymbol{\eta}_t|\boldsymbol{y})$ in (5.13). The simulated $\breve{\boldsymbol{\varepsilon}}$ vectors are generated from $g(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$ by drawing vectors from $N\{\mathbf{0}, Var(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)\}$ independently of $\boldsymbol{y}^*$ and adding these to the known vector $E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$.

Looking at the linear Gaussian approximating model we can see that the density of $\boldsymbol{\varepsilon}^*$ is

$$g(\boldsymbol{\varepsilon}^*) = N(\mathbf{0}, \boldsymbol{\Omega}), \qquad \boldsymbol{\Omega} = diag(\boldsymbol{H}_1, ..., \boldsymbol{H}_n, \boldsymbol{Q}_1, ..., \boldsymbol{Q}_n).$$

Let $\boldsymbol{\varepsilon}^+$ be a random vector drawn from $g(\boldsymbol{\varepsilon}^*)$. Then the vector $\boldsymbol{\varepsilon}^+$ is used to generate a $\boldsymbol{y}^+$ and an $\boldsymbol{\alpha}^+$ vector simply by utilising the state space form (5.7). The only issue here is choosing the initial value $\boldsymbol{\alpha}_1^+$.

Next, $E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+)$ is computed using the Kalman filter (5.11) and disturbance smoother (5.13). Using the Kalman filter with diffuse initialisation means that the issue of initialisation is not a problem; the values of $\boldsymbol{\alpha}_1^+$ can be chosen arbitrarily. Since $Var(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$ is independent of $\boldsymbol{y}^*$, then

$$Var(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+) = Var(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*).$$

This leads to

$$g(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+) = N\{E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+), Var(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+)\} = N\{E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+), Var(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)\},$$

and therefore

$$g(\boldsymbol{\varepsilon}^+ - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+)|\boldsymbol{y}^+) = N\{\mathbf{0}, Var(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)\},$$

meaning that $\boldsymbol{\varepsilon}^+ - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+)$ is the desired draw from $N\{\mathbf{0}, Var(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)\}$.

Now, if we let $\breve{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^+ - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+) + E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$ then $\breve{\boldsymbol{\varepsilon}}$ is a draw from density $g(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$ by the following argument:

$$\begin{aligned}
E(\breve{\boldsymbol{\varepsilon}}|\boldsymbol{y}^*) &= E\{\boldsymbol{\varepsilon}^+ - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+) + E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)|\boldsymbol{y}^*\} \\
&= E\{\boldsymbol{\varepsilon}^+ - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+)|\boldsymbol{y}^*\} + E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*) \\
&= E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*), \qquad (6.23)
\end{aligned}$$

and

$$\begin{aligned}
Var(\breve{\boldsymbol{\varepsilon}}|\boldsymbol{y}^*) &= Var\{\boldsymbol{\varepsilon}^+ - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+) + E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)|\boldsymbol{y}^*\} \\
&= E[\{\boldsymbol{\varepsilon}^+ - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+)\}\{\boldsymbol{\varepsilon}^+ - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+)\}'|\boldsymbol{y}^*] \\
&= Var(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+) \\
&= Var(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*). \qquad (6.24)
\end{aligned}$$

As a summary, the algorithm for drawing the simulated $\breve{\boldsymbol{\varepsilon}}$ vectors from $g(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$ is presented below.

1. Draw a random vector $\boldsymbol{\varepsilon}^+$ from density $g(\boldsymbol{\varepsilon}^*) \sim N(\mathbf{0}, \boldsymbol{\Omega})$.
2. Use $\boldsymbol{\varepsilon}^+$ to generate $\boldsymbol{y}^+$ from the state space form (5.7), where $\boldsymbol{\alpha}_1^+$ can be arbitrarily chosen.
3. Compute $E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$ and $E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+)$ using the Kalman filter and disturbance smoother with diffuse initialisation.
4. Take $\breve{\boldsymbol{\varepsilon}} = E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*) - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+) + \boldsymbol{\varepsilon}^+$.

From $\breve{\boldsymbol{\varepsilon}}$ we now have random draws, $\breve{\boldsymbol{\eta}}$ and $\breve{\boldsymbol{\epsilon}}$, of the state and model errors from the Gaussian importance density $g(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$. The algorithm is repeated over and over to get as many vectors $\breve{\boldsymbol{\varepsilon}}$ as are needed to obtain a reasonably sized sample; usually $N = 10$ is sufficient. When states are Gaussian distributed, the simulated observation error component of $\breve{\boldsymbol{\varepsilon}}$, $\breve{\boldsymbol{\epsilon}}$, can be used to derive simulated $\breve{\boldsymbol{\theta}}$ vectors by taking $\breve{\boldsymbol{\theta}}_t = \boldsymbol{y}_t^* - \breve{\boldsymbol{\epsilon}}_t$ from the linear Gaussian approximating model. A sample of $\breve{\boldsymbol{\theta}}$ vectors can then be used in (6.11) to derive $E\{x(\boldsymbol{\alpha})|\boldsymbol{y}\}$.

6.4 Model building and goodness-of-fit for exponential family models

Most of what is discussed in §5.2.3 about goodness-of-fit in Gaussian state space models also applies to non-Gaussian state space models. A visual inspection of the data is vital and model building will frequently begin with the basic structural model, or at least the non-Gaussian equivalent of it:

$$\theta_t = \mu_t + \gamma_t. \tag{6.25}$$

Here, in accordance with the basic structural model, the trend component, μ_t, is expressed as a random walk plus trend and we assume that all errors are Gaussian distributed. So μ_t is given as follows:

$$\begin{aligned} \mu_{t+1} &= \mu_t + \nu_t + \xi_t, \qquad & \xi_t &\sim N(0, \sigma_\xi^2), \\ \nu_{t+1} &= \nu_t + \zeta_t, \qquad & \zeta_t &\sim N(0, \sigma_\zeta^2). \end{aligned}$$

The seasonal component, γ_t, may be of the trigonometric seasonal variety or the seasonal dummy variety.

As with Gaussian models, likelihood ratio tests may be used to assess the effect a particular structural term or explanatory variable has on the model fit. However, when it comes to goodness-of-fit there is no equivalent to the GLM saturated model that can be used in a likelihood ratio to derive a deviance statistic and hence derive deviance residuals. In fact the literature is remarkably scant on the subject of residuals for non-Gaussian models; for example, of the four examples on non-Gaussian state space models given in Durbin and Koopman (2001) ch 14, not one refers to residuals. However, there is no reason in principle why Pearson residuals should not be used to assess model fit, and therefore these have been used and are referred to in the assessment of fit in the non-Gaussian models of this chapter and chapter 8.

Over-dispersion does not feature in the literature on non-Gaussian state space models; this is perhaps because the stochastic terms in a state space model will often absorb much of the variation which would usually be left unaccounted for in a log-linear regression model, say. However, it is none the less interesting to compare the dispersion parameter from a stochastic model with that of an ordinary regression model. It can be calculated in exactly the same way as the dispersion in a regression model by using the Pearson statistic divided by $n - m$ where m is the number of explanatory variables. Note that for state space models each state is an explanatory variable and the seasonal dummy term has $s - 1$ states.

6.5 A Poisson state space model applied to the Scottish Fatalities data

In this section we demonstrate the use of the Poisson state space model in the analysis of the Scottish fatalities data of chapter 2. We start modelling the Scottish fatality data by specifying the marginal distribution of the observations, which in our case is the Poisson distribution; given in exponential family form, this is

$$f(y_t|\theta_t) = \exp\{y_t\theta_t - \exp(\theta_t) - \log(y_t!)\}.$$

The next step is to decide upon the structural terms and explanatory variables to be used to model the data. To demonstrate the model fitting procedure and the ability of structural terms to fit a model, we shall first consider a model with only structural terms in it.

The plot of the Scottish fatalities data in figure 2.1 (i), shows there is little evidence of a trend or indeed a seasonal pattern, but since both the linear and log-linear regression models indicated that a linear trend and two highly seasonal terms, the minimum temperature and daylight variables, were significant, it would be prudent to consider a model with a trend and seasonal component to start with. So to begin with we take the conservative course of action and express θ_t by the non-Gaussian basic structural model (5.14). Since the seasonal dummy term is slightly simpler conceptually than the seasonal trigonometric term and there is generally no difference between the two in terms of model fit, we shall choose the seasonal dummy formulation for our basic structural model:

$$\gamma_{t+1} = -\sum_{j=0}^{s-1} \gamma_{t+1-j} + \omega_t, \qquad \omega_t \sim N(0, \sigma_\omega^2).$$

Next, sensible initial estimates of σ_ξ^2, σ_ζ^2 and σ_ω^2 are chosen and put into the parameter vector $\boldsymbol{\psi}$. The initial estimates can essentially be arbitrary although calculation of the maximum likelihood will take longer the further the initial estimates are from the final estimates.

We now move onto the specification of the linear Gaussian approximating model. Recall that the term $h_t(y_t|\theta_t)$ must be linearised so that the final term of (6.20) is in the same form as the final term of (6.18). An arbitrary trial value, $\tilde{\boldsymbol{\alpha}}$, of $\boldsymbol{\alpha}$ is chosen, which leads to $\tilde{\theta}_t = Z_t\tilde{\alpha}_t$. For models from the exponential family such as the Poisson distribution, $\tilde{\theta}_t$ is used in the linearisation to obtain

$$\tilde{H}_t = \ddot{h}_t^{-1} \quad and \quad \tilde{y}_t^* = \tilde{\theta}_t - \ddot{h}_t^{-1}\dot{h}_t,$$

where the definitions of $\dot{h}_t$ and $\ddot{h}_t$ are given in (6.21). In the Poisson case, the term $h_t(y_t|\theta_t)$ is given by

$$h_t(y_t|\theta_t) = -\log\{f(y_t|\theta_t)\} = -y_t\theta_t + \exp(\theta_t) + \log(y_t!),$$

which leads to $\dot{h}_t = -y_t + \exp(\theta_t)$ and $\ddot{h}_t = \exp(\theta_t)$. These then yield

$$\tilde{H}_t = \exp(-\tilde{\theta}_t) \quad and \quad \tilde{y}_t^* = \tilde{\theta}_t - 1 + \exp(-\tilde{\theta}_t)y_t. \tag{6.26}$$

The quantities $\tilde{H}_t$ and $\tilde{y}_t^*$ are then substituted into the linear Gaussian approximating model (6.16). The Kalman filter and smoother are applied to this model and from this a new trial value of θ_t is obtained by calculating $E(\theta_t|\tilde{\boldsymbol{y}}^*)$. Substituting this into the linearisation equations above gives new trial values $\tilde{H}_t$ and $\tilde{y}_t^*$. These new trial values are then substituted back into the linear Gaussian approximating model which is solved with the Kalman filter and smoother to generate a new trial value $\tilde{\theta}_t$. This process is repeated over and over until subsequent iterations do not alter the values of $\tilde{H}_t$ and $\tilde{y}_t^*$. We may now find the log-likelihood of the linear Gaussian approximating model, $L_g(\boldsymbol{\psi})$.

In the log-likelihood of the the full model, (6.14), an extra term, $\bar{w}$, must be calculated. The calculation of this term requires simulation. Since we are dealing with Gaussian distributed state vectors we may use $w(\boldsymbol{y}|\boldsymbol{\theta})$ in (6.11) as the basis for calculating $\bar{w}$. This means that only a sample of simulated $\breve{\boldsymbol{\epsilon}}$ vectors is required, where we take $\breve{\boldsymbol{\theta}} = \boldsymbol{y}_t^* - \breve{\boldsymbol{\epsilon}}$ to generate the simulated $\breve{\boldsymbol{\theta}}$ vectors. The simulation smoother of §6.3 is used to draw the simulated $\breve{\boldsymbol{\epsilon}}$ vectors.

The log-likelihood (6.14) is maximised with an optimisation procedure such as nlminb in S-Plus. Eventually the parameters within the parameter vector $\boldsymbol{\psi}$ are estimated, we now have estimates for the state error variance parameters, σ_ξ^2, σ_ζ^2 and σ_ω^2. Under the basic structural model we obtain: $\hat{\sigma}_\xi^2 = 6.778 \times 10^{-4}$, $\hat{\sigma}_\zeta^2 = 3.142 \times 10^{-12}$ and $\hat{\sigma}_\omega^2 = 6.371 \times 10^{-5}$. As with the maximisation of the log-likelihood for Gaussian state space models, the parameter estimates calculated by the optimisation procedure are actually the logs of the parameter estimates given above to avoid numerical instability in the optimisation procedure; it is these log estimates that the standard errors are calculated on. So the log estimates are $\log(\hat{\sigma}_\xi^2) = -7.297$, $\log(\hat{\sigma}_\zeta^2) = -26.49$ and $\log(\hat{\sigma}_\omega^2) = -9.661$, and their standard errors are 0.9709, 312.0 and 9.087 respectively.

Clearly some improvements can be made to the model since the slope variance estimate is so small and the standard error for the log of the slope estimate is so large. The standard error is also large for the log of the seasonal dummy variance estimate. Firstly, a model with fixed slope and seasonal components should be tried and then subsequent alterations may be made according to the parameter estimates. It happens that, in fact, the most parsimonious model with the best fit is a simple local level model, i.e., there is no need for a trend or seasonal component at all.

$$\begin{aligned} \theta_t &= \mu_t, \\ \mu_{t+1} &= \mu_t + \xi_t. \end{aligned} \tag{6.27}$$

This model gives a parameter estimate of $\log(\hat{\sigma}_\xi^2) = -6.304$ with standard error 1.312, so $\hat{\sigma}_\xi^2 = 0.001828$.

With the parameters estimated, we may now calculate the signal estimate, $\hat{\theta}_t = E(\theta_t | \tilde{y}_t^*, \hat{\psi})$, i.e., the mean. Since we have assumed the canonical link throughout, the signal estimate is linked to the series mean through the log-link, $\hat{\theta}_t = \log\{E(Y)\}$. From this we may calculate the Pearson statistic, Pearson residuals and estimate the dispersion. From the Pearson residuals for the above model we see from figure 6.1 that the fit of the basic structural model is perfectly reasonable. The estimated dispersion is $\hat{\sigma}^2 = 1.130$, which is less than the dispersion of 1.207 on the log-linear regression model fitted to the same data in chapter 2.

The reason for the improved fit could be that we have only fitted structural terms to the data and have excluded explanatory variables; since structural terms are usually time varying they tend to fit the data better than just explanatory variables which normally have fixed coefficients. However, combining structural terms and explanatory variables can result in a better fit still. Using a time varying level term in place of the constant and linear trend terms of the Scottish fatalities model of chapter 2 but keeping the same explanatory variables gives the model

$$\begin{aligned} \theta_t &= \mu_t + \beta_1 c_t + \beta_2 \ell_t, \\ \mu_{t+1} &= \mu_t + \xi_t, \end{aligned}$$

where β_1 and β_2 are non time varying coefficients, c_t is the minimum temperature variable (in Celsius) and ℓ_t is the daylight variable. Fitting and estimating this model gives $\log(\hat{\sigma}_\xi^2) = -7.473$ with standard error 0.5858, so $\hat{\sigma}_\xi^2 = 0.0005685$. The coefficient estimates are $\hat{\beta}_1 = 0.03848$ and $\hat{\beta}_2 = -0.04056$ with standard errors 0.007611 and 0.007178 respectively. The estimated dispersion is $\hat{\sigma}^2 = 1.091$, which is better than that of the purely structural model. The Pearson residual plots (figure 6.2) are perfectly reasonable too.

This model, unlike the previous one, has a direct interpretation in terms of the explanatory variables. The estimates show that a 1 degree Celsius rise in the average minimum daily temperature contributes to an increase in fatalities of 3.92% and every extra hour of daylight leads to a decrease in fatalities of 3.97%. For comparison with the regression model of chapter 2, the effects of temperature and daylight there were an increase of 4.3% and a decrease of 4.25% respectively in the numbers of Scottish car occupant fatalities.

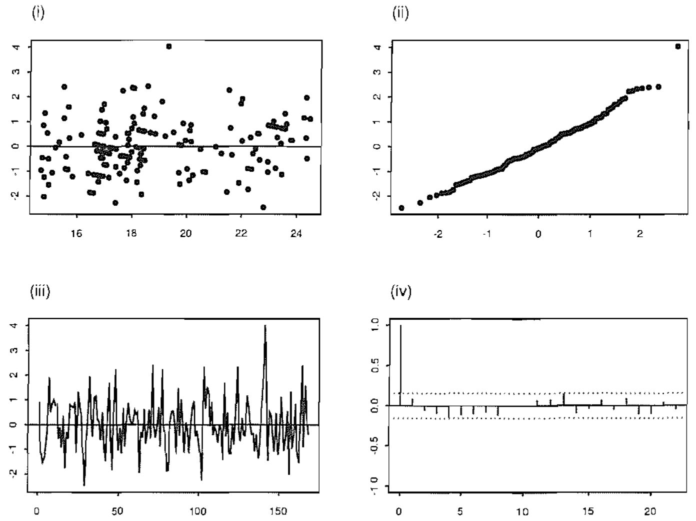

Figure 6.1: Local level Poisson model residual plots: (i) residuals vs fitted values, (ii) normal plot, (iii) time chart of residuals, (iv) correlogram.

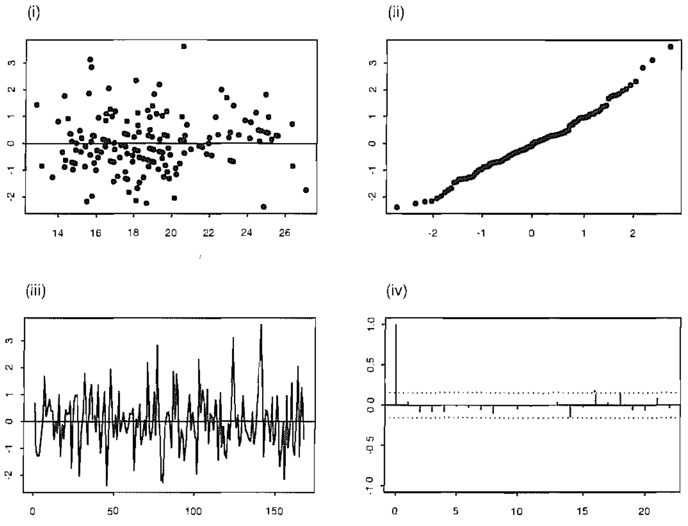

Figure 6.2: Local level and explanatory variable Poisson model residual plots: (i) residuals vs fitted values, (ii) normal plot, (iii) time chart of residuals, (iv) correlogram.

Chapter 7

Zero inflated counts

In the previous two chapters we have seen that road accident counts can be modelled using Gaussian state space models, or log-linear state space models when the Gaussian modelling assumptions are less valid. However, some count data can exhibit a high frequency of zero counts, more than would be expected say under ordinary Poisson assumptions, so that when we try to model these series with Gaussian or Poisson assumptions the model does not provide an adequate fit to the data. High numbers of zero counts in data is known as zero-inflation and as well as occurring in some road accident count series, can occur in other situations such as measuring the abundance or rare animals in a specified region (Dobbie, 2001). In this chapter we examine the conditional Bernoulli truncated Poisson distribution, or conditional distribution for short, which is particularly well suited to the modelling of high numbers of zero counts in data sets. The model has been used before with success in a generalised linear model context (Welsh et al., 1996) and so it is hoped that for time series data it may also prove fruitful. The theory will be presented in terms of univariate observations since later a clear distinction will need to be made between the dimension of the observations and that of the estimating equations. The chapter begins with an overview of the conditional distribution.

7.1 The conditional Bernoulli truncated Poisson distribution

7.1.1 The general form

The conditional model handles zero-inflated count data by separating the zero data from the non-zero data. The data is viewed in two ways: firstly as a Bernoulli model for the presence or absence of data, i.e., for zero counts and greater than zero counts; and secondly as a Poisson distribution truncated at zero, hereafter referred to as the truncated Poisson distribution, to model the positive integer counts.

The truncated Poisson density is a modification of the standard Poisson density (2.8), and for observations y_t (we shall use subscript t rather than i so as to link with the rest of this chapter), for $t = 1, ..., n$, is given by

$$p(y_t|\lambda_t) = \frac{e^{-\lambda_t}\lambda_t^{y_t}}{y_t!(1-e^{-\lambda_t})}, \qquad y_t = 1, 2, ..., \qquad \lambda_t > 0 \tag{7.1}$$

Here, unlike the standard Poisson density, λ_t is neither the mean nor the variance. Deriving the mean and variance from the density (this can be seen for the conditional model in appendix B), we find

$$E(Y_t) = \frac{\lambda_t}{1-e^{-\lambda_t}}, \qquad Var(Y_t) = \left(\frac{\lambda_t}{1-e^{-\lambda_t}}\right)\left(1+\lambda_t-\frac{\lambda_t}{1-e^{-\lambda_t}}\right).$$

The truncated Poisson density is then combined with the standard Bernoulli density,

$$b(y_t|\pi_t) = \pi_t^{y_t}(1-\pi_t)^{1-y_t}, \qquad y_t \in \{0,1\}, \quad 0 < \pi_t < 1,$$

using the function $I(y_t)$. The conditional density is thus given by

$$f(y_t|\pi_t, \lambda_t) = \{1-\pi_t\}^{(1-I(y_t))}\left\{\pi_t\frac{e^{-\lambda_t}\lambda_t^{y_t}}{y_t!(1-e^{-\lambda_t})}\right\}^{I(y_t)}. \tag{7.2}$$

where the parameter ranges are $\lambda_t > 0$ and $0 < \pi_t < 1$, exactly as for the parameters of the truncated Poisson and Bernoulli distributions. The function $I(y_t)$ is an indicator function applied according to whether $y_t = 0$ or $y_t > 0$ and takes the form

$$I(y_t) = \begin{cases} 0, & y_t = 0, \\ 1, & y_t = 1, 2, \ldots. \end{cases}$$

Deriving the mean and variance of the conditional density, as shown in appendix B, we find

$$E(Y_t) = \frac{\pi_t \lambda_t}{1 - e^{-\lambda_t}}, \qquad Var(Y_t) = \left(\frac{\pi_t \lambda_t}{1 - e^{-\lambda_t}}\right)\left(1 + \lambda_t - \frac{\pi_t \lambda_t}{1 - e^{-\lambda_t}}\right). \quad (7.3)$$

7.1.2 The conditional distribution in exponential family form

Like the Bernoulli and Poisson distributions, the conditional distribution can also be put into exponential family form so that the methods of estimation set out in the previous chapter can be applied with little modification. However, it is not possible to express the conditional distribution in the standard exponential family form (6.3). Instead the general exponential family form must be used. For univariate observations this is given by

$$f(y_t|\boldsymbol{\theta_t}) = \exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta_t})T_j(y_t) + d(\boldsymbol{\theta_t}) + S(y_t)\right\}, \qquad -\infty < \boldsymbol{\theta_t} < \infty, \quad (7.4)$$

where the dimension of the vector $\boldsymbol{\theta_t}$ is k, where $k \geq 1$.

It is convenient that in the general exponential distribution the dimension of $\boldsymbol{\theta_t}$ can be greater than one even when the dimension of y_t is equal to one since the conditional distribution has two parameters, λ_t and π_t. Putting the conditional distribution into the exponential family form gives

$$\begin{aligned} f(y_t|\lambda_t,\pi_t) &= \exp\left\{I(y_t)\left[\log\left(\tfrac{\pi_t}{1-\pi_t}\right)-\lambda_t-\log\{1-\exp(-\lambda_t)\}\right]\right. \\ &\quad \left.+I(y_t)y_t\log(\lambda_t)+\log(1-\pi_t)-I(y_t)\log(y_t!)\right\}. \end{aligned}$$

By noting that $I(y_t)y_t\log(\lambda_t)=y_t\log(\lambda_t)$ and $I(y_t)\log(y_t!)=\log(y_t!)$, due to the form of the function $I(y_t)$, the above density can be simplified to

$$\begin{aligned} f(y_t|\lambda_t,\pi_t) &= \exp\left\{I(y_t)\left[\log\left(\tfrac{\pi_t}{1-\pi_t}\right)-\lambda_t-\log\{1-\exp(-\lambda_t)\}\right]\right. \\ &\quad \left.+y_t\log(\lambda_t)+\log(1-\pi_t)-\log(y_t!)\right\}. \end{aligned} \tag{7.5}$$

Taking $\boldsymbol{\theta_t}=(\theta_{1,t}\ \ \theta_{2,t})'$ allows for density (7.5) to be expressed in terms of $\boldsymbol{\theta_t}$. We are now compelled to choose links to relate the parameters λ_t and π_t to $\theta_{1,t}$ and $\theta_{2,t}$. As in Welsh et al. (1996), we choose the log link to relate $\theta_{1,t}$ to λ_t and the logistic link to relate π_t to $\theta_{2,t}$.

$$\begin{aligned} \theta_{1,t} &= \log(\lambda_t) & \Rightarrow\quad \lambda_t &= \exp(\theta_{1,t}), \\ \theta_{2,t} &= \log(\tfrac{\pi_t}{1-\pi_t}) & \Rightarrow\quad \pi_t &= \tfrac{\exp(\theta_{2,t})}{1+\exp(\theta_{2,t})}. \end{aligned} \tag{7.6}$$

Then writing (7.5) in terms of $\theta_{1,t}$ and $\theta_{2,t}$ gives the following

$$\begin{aligned} f(y_t|\theta_{1,t},\theta_{2,t}) &= \exp\left[I(y_t)\{\theta_{2,t}-\exp(\theta_{1,t})-\log[1-\exp\{-\exp(\theta_{1,t})\}]\}\right. \\ &\quad \left.+y_t\theta_{1,t}-\log\{1+\exp(\theta_{2,t})\}-\log(y_t!)\right]. \end{aligned} \tag{7.7}$$

Relating this back to the general exponential family density, (7.4), gives the following functions:

$$\begin{aligned} c_1(\boldsymbol{\theta_t}) &= \theta_{2,t}-\exp(\theta_{1,t})-\log[1-\exp\{-\exp(\theta_{1,t})\}], \\ T_1(y_t) &= I(y_t), \\ c_2(\boldsymbol{\theta_t}) &= \theta_{1,t}, \\ T_2(y_t) &= y_t, \\ d(\boldsymbol{\theta_t}) &= -\log\{1+\exp(\theta_{2,t})\}, \\ S(y_t) &= -\log(y_t!). \end{aligned}$$

Expressing the conditional density, (7.2), in exponential family form, as in

(7.5) or (7.7), highlights the orthogonal parameterisation of this distribution. This is a feature of the conditional model that other distributions used for modelling zero inflated data, such as the zero inflated Poisson distribution, do not have. Although philosophically there is no intuitive advantage to orthogonal parameterisation in the modelling of road accidents, there are practical advantages that will become apparent in the following sections.

Following the procedures of the previous chapter we must generate a sample of simulated state vectors, $\breve{\boldsymbol{\alpha}}^{(1)}, ..., \breve{\boldsymbol{\alpha}}^{(N)}$, from the importance density, $g(\boldsymbol{\alpha}|\boldsymbol{y})$, and the mode of the importance density must be equated with the mode of density $f(\boldsymbol{\alpha}|\boldsymbol{y})$. Thus, the first step is to equate the mode of $g(\boldsymbol{\alpha}|\boldsymbol{y})$ with that of $f(\boldsymbol{\alpha}|\boldsymbol{y})$, and the next step is to sample from $g(\boldsymbol{\alpha}|\boldsymbol{y})$.

7.2 Equating the mode of $g(\boldsymbol{\alpha}|\boldsymbol{y})$ with the mode of $f(\boldsymbol{\alpha}|\boldsymbol{y})$

7.2.1 The linear Gaussian approximating model

Since $g(\boldsymbol{\alpha}|\boldsymbol{y})$ is a symmetric distribution, its mode is equal to its mean which, theoretically, can be obtained by calculating $E(\boldsymbol{\alpha}|\boldsymbol{y})$ from the linear Gaussian approximating model, (6.15). The density $f(y_t|\boldsymbol{\theta_t})$ in (7.7), used to determine the observations for the conditional model, has bivariate $\boldsymbol{\theta_t}$ where $\boldsymbol{\theta_t} = (\theta_{1,t} \;\; \theta_{2,t})'$. Therefore in the linear Gaussian approximating model $\boldsymbol{\theta_t}$ is also bivariate, where $\boldsymbol{\theta_t} = \boldsymbol{Z_t\alpha_t}$. However, in model (6.15) a bivariate $\boldsymbol{\theta_t}$ implies a bivariate y_t and ϵ_t, and yet we know that y_t is univariate.

Let us accept the bivariate nature of the linear Gaussian approximating model by defining a bivariate observation vector, $\boldsymbol{y_t}$, in (6.15), where both elements of $\boldsymbol{y_t}$ are defined simply by the observation y_t, thus $\boldsymbol{y_t} = (y_t \;\; y_t)'$.

In this formulation the two elements of $\boldsymbol{\epsilon_t}$ are potentially different from one another, so we take $\boldsymbol{\epsilon_t} = (\epsilon_{1,t} \;\; \epsilon_{2,t})'$ and $\boldsymbol{H_t} = diag(H_{1,t}, H_{2,t})$.

The bivariate forms of the vectors and matrices relating to $\boldsymbol{\theta_t}$ in the linear Gaussian approximating model are obtained similarly to adding a seasonal term to a state space form as was shown in (5.10); the only exception being the matrix $\boldsymbol{Z_t}$. The bivariate state space quantities presented below can easily be generalised to any dimension.

$$\begin{array}{rclcrcl} \boldsymbol{Z_t} & = & diag(\boldsymbol{Z_{1,t}}, \boldsymbol{Z_{2,t}}), & & \boldsymbol{\alpha_t} & = & (\boldsymbol{\alpha}'_{\boldsymbol{1,t}} \;\; \boldsymbol{\alpha}'_{\boldsymbol{2,t}})', \\ \boldsymbol{T_t} & = & diag(\boldsymbol{T_{1,t}}, \boldsymbol{T_{2,t}}), & & \boldsymbol{R_t} & = & diag(\boldsymbol{R_{1,t}}, \boldsymbol{R_{2,t}}), \\ \boldsymbol{\eta_t} & = & (\boldsymbol{\eta}'_{\boldsymbol{1,t}} \;\; \boldsymbol{\eta}'_{\boldsymbol{2,t}})', & & \boldsymbol{Q_t} & = & diag(\boldsymbol{Q_{1,t}}, \boldsymbol{Q_{2,t}}). \end{array}$$

Recall that we cannot calculate the mean $E(\boldsymbol{\alpha}|\boldsymbol{y})$ of $g(\boldsymbol{\alpha}|\boldsymbol{y})$ from model (6.15) using the Kalman filter and smoother since $E(\boldsymbol{\epsilon_t}) \neq \mathbf{0}$. Therefore we must instead calculate $E(\boldsymbol{\alpha}|\boldsymbol{y}^*)$ from model (6.16), where $\boldsymbol{y_t^*} = (y_{1,t}^* \;\; y_{2,t}^*)'$ and $\boldsymbol{\epsilon_t^*} = (\epsilon_{1,t}^* \;\; \epsilon_{2,t}^*)'$. The vector $E(\boldsymbol{\alpha}|\boldsymbol{y}^*) = E(\boldsymbol{\alpha}|\boldsymbol{y})$ and therefore is the mean and mode of the importance density $g(\boldsymbol{\alpha}|\boldsymbol{y})$.

7.2.2 Linearisation

For non-Gaussian state space models the states are determined by a relationship of the type given in (6.2). If we assume that the distribution of the state errors for the conditional model is Gaussian then the states are determined exactly by the state equation of the linear Gaussian approximating model (6.16). As such we just need to find the quantities $\boldsymbol{y_t^*}$ and $\boldsymbol{H_t}$ in the linear Gaussian approximating model so that $E(\boldsymbol{\alpha}|\boldsymbol{y}^*)$ may be calculated.

We find $\boldsymbol{y_t^*}$ and $\boldsymbol{H_t}$ by linearising the quantity $h_t(y_t|\boldsymbol{\theta_t})$ using the methods of §6.2.3. For the conditional density we take

$$\begin{aligned}
h_t(y_t|\boldsymbol{\theta_t}) &= h(y_t|\theta_{1,t}, \theta_{2,t}) \\
&= -\log\{f(y_t|\theta_{1,t}, \theta_{2,t})\} \\
&= -I(y_t)\,(\theta_{2,t} - \exp(\theta_{1,t}) - \log[1 - \exp\{-\exp(\theta_{1,t})\}]) \\
&\quad -y_t\theta_{1,t} + \log\{1 + \exp(\theta_{2,t})\} + \log(y_t!).
\end{aligned}$$

Evaluating $\dot{\boldsymbol{h}}_{\boldsymbol{t}} = (\dot{h}_{1,t} \;\; \dot{h}_{2,t})'$ at the trial value $\tilde{\boldsymbol{\theta}}_{\boldsymbol{t}} = (\tilde{\theta}_{1,t} \;\; \tilde{\theta}_{2,t})'$ leads to

$$\begin{aligned}
\dot{h}_{1,t} &= \frac{I(y_t)\exp(\tilde{\theta}_{1,t})}{1 - \exp\{-\exp(\tilde{\theta}_{1,t})\}} - y_t, \\
\dot{h}_{2,t} &= -I(y_t) + \frac{\exp(\tilde{\theta}_{2,t})}{1 + \exp(\tilde{\theta}_{2,t})}.
\end{aligned}$$

Then from (6.21), evaluating $\ddot{\boldsymbol{h}}_{\boldsymbol{t}}$ at the trial value $\tilde{\boldsymbol{\theta}}$ leads to a 2×2 matrix of the form

$$\begin{pmatrix} \frac{\partial^2 h_t(\boldsymbol{y_t}|\boldsymbol{\theta_t})}{\partial\theta_{1,t}^2} & \frac{\partial^2 h_t(\boldsymbol{y_t}|\boldsymbol{\theta_t})}{\partial\theta_{1,t}\partial\theta_{2,t}} \\ \frac{\partial^2 h_t(\boldsymbol{y_t}|\boldsymbol{\theta_t})}{\partial\theta_{2,t}\partial\theta_{1,t}} & \frac{\partial^2 h_t(\boldsymbol{y_t}|\boldsymbol{\theta_t})}{\partial\theta_{2,t}^2} \end{pmatrix} = \begin{pmatrix} \ddot{h}_{1,t} & 0 \\ 0 & \ddot{h}_{2,t} \end{pmatrix}.$$

The fact that the off diagonal elements of the above matrix are zero is a direct result of the orthogonal parameterisation of (7.7). Finally, differentiating $\dot{h}_{1,t}$ and $\dot{h}_{2,t}$ gives

$$\begin{aligned}
\ddot{h}_{1,t} &= \frac{I(y_t)\exp(\tilde{\theta}_{1,t})[1 - \exp\{-\exp(\tilde{\theta}_{1,t})\} - \exp(\tilde{\theta}_{1,t})\exp\{-\exp(\tilde{\theta}_{1,t})\}]}{[1 - \exp\{-\exp(\tilde{\theta}_{1,t})\}]^2}, \\
\ddot{h}_{2,t} &= \frac{\exp(\tilde{\theta}_{2,t})}{\{1 + \exp(\tilde{\theta}_{2,t})\}^2}.
\end{aligned}$$

The quantities $\dot{h}_{1,t}$ and $\dot{h}_{2,t}$ lead to the trial variance matrix $\tilde{\boldsymbol{H}}_{\boldsymbol{t}} = \ddot{\boldsymbol{h}}_{\boldsymbol{t}}^{-1}$, which has no covariance terms, vastly reducing the complexity of subsequent analysis:

$$\tilde{\boldsymbol{H}}_{\boldsymbol{t}} = \begin{pmatrix} \tilde{H}_{1,t} & 0 \\ 0 & \tilde{H}_{2,t} \end{pmatrix}.$$

Below, $\tilde{H}_{1,t}$ and $\tilde{H}_{2,t}$ are given in terms as simplified as possible to avoid numerical rounding errors during Kalman filtering.

$$\begin{aligned}
\tilde{H}_{1,t} &= \ddot{h}_{1,t}^{-1} \\
&= \frac{\exp\{\exp(\tilde{\theta}_{1,t})\} - 2 + \exp\{-\exp(\tilde{\theta}_{1,t})\}}{I(y_t)\exp(\tilde{\theta}_{1,t})[\exp\{\exp(\tilde{\theta}_{1,t})\} - 1 - \exp(\tilde{\theta}_{1,t})]}, \qquad (7.8) \\
\tilde{H}_{2,t} &= \ddot{h}_{2,t}^{-1} \\
&= \exp(-\tilde{\theta}_{2,t}) + 2 + \exp(\tilde{\theta}_{2,t}). \qquad (7.9)
\end{aligned}$$

Again, in as simplified a form as possible, the observations $\tilde{\boldsymbol{y}}_t^* = (\tilde{y}_{1,t}^* \;\; \tilde{y}_{2,t}^*)'$ from the linear Gaussian approximating model are given by the expressions

$$\begin{aligned}
\tilde{y}_{1,t}^* &= \tilde{\theta}_{1,t} - \ddot{h}_{1,t}^{-1}\dot{h}_{1,t} \\
&= \tilde{\theta}_{1,t} - \tilde{H}_{1,t}\left[\frac{I(y_t)\exp(\tilde{\theta}_{1,t})}{1 - \exp\{-\exp(\tilde{\theta}_{1,t})\}} - y_t\right], \qquad (7.10) \\
\tilde{y}_{2,t}^* &= \tilde{\theta}_{2,t} - \ddot{h}_{2,t}^{-1}\dot{h}_{2,t} \\
&= \tilde{\theta}_{2,t} + \frac{\{1 + \exp(\tilde{\theta}_{2,t})\}^2}{\exp(\tilde{\theta}_{2,t})}\left\{I(y_t) - \frac{\exp(\tilde{\theta}_{2,t})}{1 + \exp(\tilde{\theta}_{2,t})}\right\} \\
&= \tilde{\theta}_{2,t} - 1 - \exp(\tilde{\theta}_{2,t}) + I(y_t)\{\exp(-\tilde{\theta}_{2,t}) + 2 + \exp(\tilde{\theta}_{2,t})\}. \; (7.11)
\end{aligned}$$

7.3 Overcoming problems with the linear Gaussian approximating model

Closer inspection of (7.8) and (7.10) reveals a difficult issue concerning the nature of these quantities when the observations take the value zero. When $y_t = 0$, $I(y_t) = 0$, and so in the variance $\tilde{H}_{1,t}$ there is problem of the form $1/0$ which implies infinite variance. Similarly, since $\tilde{H}_{1,t}$ features as part of the formula for $\tilde{y}_{1,t}^*$, there is a problem of the sort $0/0$ in $\tilde{y}_{1,t}^*$ when $y_t = 0$. These problems are caused because the truncated Poisson component of the conditional model, representing all observations greater than or equal to one, assigns a probability of zero to observations at zero. The truncated

Poisson model, however, must continue to play a part in the conditional model while observations are zero valued. So the question is what to do with $\tilde{H}_{1,t}$ and $\tilde{y}^*_{1,t}$ while $y_t = 0$. There are conceivably many ways of dealing with this problem; the following subsections describe four methods which could be used.

7.3.1 Treating zero observations as missing in the truncated Poisson model

Model formulation

One way to proceed while $y_t = 0$ is to treat $\tilde{H}_{1,t}$ and $\tilde{y}^*_{1,t}$ as missing. If this is the assumption then the problem can be dealt with entirely via the Kalman filter and smoother, meaning that we do not have to come up with alternatives for $\tilde{H}_{1,t}$ and $\tilde{y}^*_{1,t}$. This method has the advantage that the treatment of missing observations via the Kalman filter and smoother is relatively simple.

In general, for a series $\boldsymbol{y_t}$, for $t = 1, ..., n$, with observations missing at times $t = \tau, ..., \tau^*$, the Kalman filter quantities $\boldsymbol{K_t}$, $\boldsymbol{\nu_t}$ and $\boldsymbol{F_t}^{-1}$ simply take the value zero for these times. Substituting zero for these quantities into the Kalman filter and smoother implies that $\boldsymbol{L_t} = \boldsymbol{T_t}$ and the filtering and smoothing updating equations become

$$\begin{array}{rclrcll} \boldsymbol{a}_{t+1} & = & \boldsymbol{T}_t\boldsymbol{a}_t, & \boldsymbol{P}_{t+1} & = & \boldsymbol{T}_t\boldsymbol{P}_t\boldsymbol{T}_t' + \boldsymbol{R}_t\boldsymbol{Q}_t\boldsymbol{R}_t', & \\ \boldsymbol{r}_{t-1} & = & \boldsymbol{T}_t'\boldsymbol{r}_t, & \boldsymbol{N}_{t-1} & = & \boldsymbol{T}_t'\boldsymbol{N}_t\boldsymbol{T}_t, & t = \tau, ..., \tau^*. \end{array}$$

Note that it is quite possible to have a series of multivariate observations with obsevations missing from only one of the series and not both. For the other series in the multivariate model the Kalman filter quantities would be estimated as normal.

It can be seen from the equations of (5.13) that while the quantities $\boldsymbol{K_t}$, $\boldsymbol{\nu_t}$ and $\boldsymbol{F_t}^{-1}$ are equal to zero, $E(\boldsymbol{\epsilon_t}|\boldsymbol{y}) = \boldsymbol{0}$ and $Var(\boldsymbol{\epsilon_t}|\boldsymbol{y}) = \boldsymbol{H_t}$. However, in the case of the linear Gaussian approximating model for the conditional distribution, $\tilde{H}_{1,t}$ is considered a missing value meaning that $Var(\epsilon^*_{1,t}|\tilde{\boldsymbol{y}}^*)$ is still undefined. This problem is shared by some other models when missing values are present such as the stochastic volatility model (Zivot et al., 2003); here $\tilde{H}_t$ is dependent on y_t, which causes it to be undefined when y_t is missing. However, the problem does not arise in cases such as the Poisson model, where $\tilde{H}_t$ in the linear Gaussian approximating model is only dependent on $\tilde{\theta}_t$.

Simulation

The missing values in $\tilde{H}_{1,t}$ have an effect on how the simulation methods of §6.3 are implemented. Recall that we wish to draw a simulated vector, $\breve{\boldsymbol{\varepsilon}}$, which is a draw from density $g(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$, where $g(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*) = N\{E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*), Var(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)\}$. The vector $\breve{\boldsymbol{\varepsilon}}$ is calculated by $\breve{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^+ - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+) + E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$. Therefore we start the process of drawing $\breve{\boldsymbol{\varepsilon}}$ by generating a random vector, $\boldsymbol{\varepsilon}^+$, from the distribution $p(\boldsymbol{\varepsilon}^*) = N(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = diag(\boldsymbol{H_1, ..., H_n, Q_1, ..., Q_n})$. However, there are missing values in $\boldsymbol{\Omega}$ corresponding to the missing $H_{1,t}$ values. This means that missing $\epsilon^+_{1,t}$ values are generated in the vector $\boldsymbol{\varepsilon}^+$ and consequently missing values are generated in $\breve{\boldsymbol{\varepsilon}}$ since $\breve{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^+ - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+) + E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*)$.

The most sensible option at this stage is to use the Expected values of $\breve{\epsilon}_{1,t}$ in place of the missing values; $E(\breve{\epsilon}_{1,t}) = E(\epsilon^*_{1,t}) = 0$ for $t = \tau, ..., \tau^*$. If we are using a sample of $\breve{\boldsymbol{\theta}}$ vectors for the calculation of $w(\boldsymbol{\theta}|\boldsymbol{y})$ then when we have missing values we must take $\breve{\theta}_{1,t} = E(y^*_{1,t}) - E(\breve{\epsilon}_{1,t}) = \theta_{1,t}$, for $t = \tau, ..., \tau^*$, where $\theta_{1,\tau}, ..., \theta_{1,\tau^*}$ are real valued and are calculated from the linear approximating model.

Practically, the effect of missing observations in a Gaussian state space model is that when we calculate the model fit, it is interpolated for the missing observations. For example, in the case of the univariate local level model, (5.1), while observations are missing, the model fit, $E(\theta_t|\boldsymbol{y})$, takes the form of a simple straight line connecting $E(\theta_{\tau-1}|\boldsymbol{y})$ to $E(\theta_{\tau^*+1}|\boldsymbol{y})$.

7.3.2 Using continuous time state space models

In general, continuous time state space models are used when the observed series is unevenly spaced. This can, of course, be applied to our case where we consider $\tilde{y}^*_{1,t}$ to be missing when $y_t = 0$. Therefore we end up with an unevenly spaced series in $\tilde{y}^*_{1,t}$ and so continuous time methods can be applied.

The way in which continuous time series are expressed is slightly different to discrete time series since the exact time that an observation appears is what is important rather than the order of the observation in the series. We take $y(t)$ as a continuous function of time, for t in an interval $0 \leq t \leq T$. Observations are taken at time points $t_1 \leq ... \leq t_n$ from this interval.

First, we consider the local level model, (5.1). In continuous time all differences become differentials, so writing the state equation of (5.1) in difference form, $\mu_{t+1} - \mu_t = \xi_t$, where $\xi_t \sim N(0, \sigma^2_\xi)$ now leads to its conversion to the continuous time process $\mu(t)$ expressed in differential form:

$$\frac{d\mu(t)}{dt} = \frac{dw(t)}{dt}, \qquad \frac{dw(t)}{dt} \sim N(0, \sigma_\xi),$$

where $w(t)$ is the Brownian motion process or Wiener process. Here, $w(0) = 0$, $w(t) \sim N(0, t\sigma^2_\xi)$ for $0 < t < \infty$ and increments between observation time points $w(t_2) - w(t_1)$, $w(t_4) - w(t_3)$,..., for $0 \leq t_1 \leq t_2 \leq t_3 \leq t_4$,..., are independent. Integrating and combining with the measurement equation gives the continuous time state space form of (5.1):

$$\begin{aligned} y(t) &= \mu(t) + \epsilon(t), \\ \mu(t) &= \mu(0) + w(t), \qquad 0 \leq t \leq T, \quad T > 0. \end{aligned} \tag{7.12}$$

It is easy to think of the state equation here as being continuous since it is the mean of the observations; the observation equation, however, is defined by the observations and therefore cannot in some sense be continuous like the state equation since for most applications, and certainly the ones considered in this thesis, we only take observations at discrete points in time. To define $y(t)$ in (7.12) we must define $\epsilon(t)$; here, $Var\{\epsilon(t)\}$ is allowed to vary with time even when we are not dealing with the non-Gaussian case, this is to allow for different sized gaps between observations. The complete continuous time state space model analogue for (5.1) is, then,

$$\begin{array}{rclllcl} y(t) & = & \mu(t)+\epsilon(t), & t=t_1,...,t_n, & \epsilon(t_i) & \sim & N\{0,\sigma_\epsilon^2(t_i)\}, \\ \mu(t) & = & \mu(0)+w(t), & 0\leq t\leq T, & w(t) & \sim & N\{0,t\sigma_\xi^2\}. \end{array} \tag{7.13}$$

Finally, however, the estimation of unknown parameters via the log-likelihood leads to a further discretisation of the state space form. The log-likelihood depends on $\mu(t)$ only at the values $t_1,...,t_n$, so to actually estimate parameters, the following model is applied:

$$\begin{array}{rcll} y_i & = & \mu_i+\epsilon_i, & \\ \mu_{i+1} & = & \mu_t+\xi_i, & i=1,...,n, \end{array} \tag{7.14}$$

where $y_i=y(t_i)$, $\mu_i=\mu(t_i)$, $\epsilon_i=\epsilon(t_i)$ and $\xi_i=w(t_{i+1})-w(t_i)$. This model is different from (5.1) only in that the variances of the ϵ_i's can vary. Of course, for the models we are using, the variance of the observation equation also varies with time because of the non-Gaussian distributed observations.

The problem with equation (7.14) is that it really is no different to equation (5.1) in terms of its estimation. In a bivariate situation, say, there must be as many $y_{1,i}$'s as $y_{2,i}$'s; that is, the $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ vectors must be the same length. If there are missing observations in one of the observation series but not the other, as is the case for the observations of the linear Gaussian approximating model in the conditional model, then we are back to the situation of applying NA's to the missing values and adapting the Kalman filter to deal with them. Also, the method above is presented just for the local level model; when other states are added, let alone other observation series, the system matrices take on more complicated forms than in the discrete time case. So, all in all, although continuous time state space models are theoretically more intuitively appealing, in practice no advantage is gained by their application to the conditional model and a lot of unnecessary complexity is added to deal with continuous time series which are actually only observed at discrete time points. Continuous time models would, of course, be of great use if the data were unevenly spread or if there were missing observations.

7.3.3 The joint Bernoulli truncated Poisson approach

There is another way of treating $\tilde{H}_{1,t}$ and $\tilde{y}^*_{1,t}$ while $y_t = 0$. For this method we do not model the observations y_t; instead, a bivariate series, $K(y_t) = (I(y_t) \quad J(y_t))'$, is modelled based on a joint Bernoulli truncated Poisson distribution. Here $J(y_t)$ is a function which constrains y_t to be one or more:

$$J(y_t) = \begin{cases} 1, & y_t = 0, \\ y_t, & y_t = 1, 2, ... \end{cases}$$

The joint Bernoulli truncated Poisson distribution is simply a Bernoulli distribution modelling $I(y_t)$ joined with a truncated Poisson distribution modelling $J(y_t)$. Strictly speaking it is not a conditional Bernoulli truncated Poisson distribution at all since it has a different distribution function:

$$k(I(y_t), J(y_t)|\pi_t, \lambda_t) = \{1 - \pi_t\}^{(1-I(y_t))}\pi_t^{I(y_t)} \cdot \frac{e^{-\lambda_t}\lambda_t^{J(y_t)}}{J(y_t)!(1 - e^{-\lambda_t})}.$$

However, $I(y_t)$ and $J(y_t)$ are not independent of one another, and they both depend on y_t, so that the form of the distribution is still conditional on the presence or absence of data; also, the mean and variance are still the same as for the conditional distribution. For simplicity, we shall consider it to be a type of conditional distribution.

Using the same link functions as before, (7.6), we obtain the following distribution in exponential form:

$$\begin{aligned} k(J(y_t), I(y_t)|\theta_{1,t}, \theta_{2,t}) &= \exp\left[I(y_t)\theta_{2,t} - \exp(\theta_{1,t})\right. \\ &\quad - \log[1 - \exp\{-\exp(\theta_{1,t})\}] + J(y_t)\theta_{1,t} \\ &\quad \left. - \log\{1 + \exp(\theta_{2,t})\} - \log\{J(y_t)!\}\right]. \end{aligned}$$

Now, by going through the linearisation process outlined in §6.2.1, but using $k(K(y_t)|\theta_{1,t}, \theta_{2,t})$ rather than $f(y_t|\theta_{1,t}, \theta_{2,t})$, we find that the linear Gaussian approximating model quantities $\tilde{H}_{1,t}$ and $\tilde{y}^*_{1,t}$ are given by

$$\begin{aligned} \tilde{H}_{1,t} &= \frac{\exp\{\exp(\tilde{\theta}_{1,t})\} - 2 + \exp\{-\exp(\tilde{\theta}_{1,t})\}}{\exp(\tilde{\theta}_{1,t})[\exp\{\exp(\tilde{\theta}_{1,t})\} - 1 - \exp(\tilde{\theta}_{1,t})]}, \qquad (7.15) \\ \tilde{y}^*_{1,t} &= \tilde{\theta}_{1,t} - \tilde{H}_{1,t}\left[\frac{\exp(\tilde{\theta}_{1,t})}{1 - \exp\{-\exp(\tilde{\theta}_{1,t})\}} - J(y_t)\right]. \qquad (7.16) \end{aligned}$$

These are similar to those given in (7.8) and (7.10) but now do not have infinite and undefined values caused by the presence of $I(y_t)$. Note that $\tilde{H}_{2,t}$ and $\tilde{y}^*_{2,t}$ remain as they are in equations (7.9) and (7.11).

It must be stated that, in theory at least, there is a disadvantage to using the approach outlined above, which seems to negate the point of using a conditional model to begin with. The problem is that by using the function $J(y_t)$ to convert zero counts to one, in the truncated Poisson part of the model, we end up with a disproportionately large number of 1 counts to be modelled by the truncated Poisson state space model and there is no reason why a truncated Poisson form should model an inflated number of 1 counts any better than a Poisson form can model an inflated number of zero counts. We shall see in the following chapter whether this means that the joint model cannot give a better fit than the Poisson state space model.

7.3.4 The conditional model with univariate signal θ_t

Finally, another solution to the problem of infinite and undefined values in $\tilde{H}_{1,t}$ and $\tilde{y}^*_{1,t}$ is to take $\theta_{1,t} = \theta_{2,t} = \theta_t$ and thus have a univariate signal, θ_t, corresponding to a single set of explanatory variables and structural terms. Below, the conditional exponential form density (7.5) is restated:

$$\begin{aligned} f(y_t|\lambda_t, \pi_t) &= \exp\left\{I(y_t)\left[\log\left(\tfrac{\pi_t}{1-\pi_t}\right) - \lambda_t - \log\{1 - \exp(-\lambda_t)\}\right]\right. \\ &\quad \left. + y_t \log(\lambda_t) + \log(1 - \pi_t) - \log(y_t!)\right\}. \end{aligned}$$

Now, again as before, we are compelled to choose links to relate λ_t and π_t to the signal; this time, however, they are linked to the same signal, θ_t. We take the same links as before, the log-link for λ_t and the logistic link for π_t:

$$\begin{array}{rclcrcl} \theta_t & = & \log(\lambda_t) & \Rightarrow & \lambda_t & = & \exp(\theta_t), \\ \theta_t & = & \log(\frac{\pi_t}{1-\pi_t}) & \Rightarrow & \pi_t & = & \frac{\exp(\theta_t)}{1+\exp(\theta_t)}. \end{array} \tag{7.17}$$

In this formulation λ_t and π_t are not actually separate parameters at all; they are merely different functions of θ_t. Translating into (7.5) we obtain

$$\begin{array}{rcl} f(y_t|\theta_t) & = & \exp\left[I(y_t)\left\{\theta_t - \exp(\theta_t) - \log[1-\exp\{-\exp(\theta_t)\}]\right\}\right. \\ & & \left. +y_t\theta_t - \log\{1+\exp(\theta_t)\} - \log(y_t!)\right]. \end{array} \tag{7.18}$$

From this density we follow the linearisation process from the theory of §6.2.1. We first take $h_t(y_t|\theta_t) = -\log\{f(y_t|\theta_t)\}$ and then evaluate the differentials $\dot{h}_t$ and $\ddot{h}_t$; from these, the linear Gaussian approximating model quantities $\tilde{H}_t$ and $\tilde{y}_t^*$ are obtained. In (7.19) we use the link function $\tilde{\lambda}_t = \exp(\tilde{\theta}_t)$ to fit the equation onto the page.

$$\begin{array}{rcl} \tilde{H}_t & = & \ddot{h}_t^{-1} \\ & = & \dfrac{\{\exp(\tilde{\lambda}_t) - 2 + \exp(-\tilde{\lambda}_t)\}\{\tilde{\lambda}_t^{-1} + 2 + \tilde{\lambda}_t\}}{I(y_t)\{\exp(\tilde{\lambda}_t) - 1 - \tilde{\lambda}_t\}\{1+\tilde{\lambda}_t\}^2 + \exp(\tilde{\lambda}_t) - 2 + \exp(-\tilde{\lambda}_t)}, \end{array} \tag{7.19}$$

$$\begin{array}{rcl} \tilde{y}_t^* & = & \tilde{\theta}_t - \ddot{h}_t^{-1}\dot{h}_t \\ & = & \tilde{\theta}_t - \tilde{H}_t\left[\dfrac{I(y_t)\exp(\tilde{\theta}_t)}{1-\exp\{-\exp(\tilde{\theta}_t)\}} - I(y_t) - y_t + \dfrac{\exp(\tilde{\theta}_t)}{1+\exp(\tilde{\theta}_t)}\right]. \end{array} \tag{7.20}$$

It would seem that the individual expressions for $\tilde{H}_t$ and $\tilde{y}_t^*$ in the above equations are more complicated than those of $\tilde{H}_{1,t}$, $\tilde{H}_{2,t}$, $\tilde{y}_{1,t}^*$ and $\tilde{y}_{2,t}^*$ in (7.8) to (7.11). This complexity, however, is countered by the fact that there are only two equations as opposed to four; and further, when $y_t = 0$, we notice that there are no longer problems with undefined values in either $\tilde{H}_t$ or $\tilde{y}_t^*$ since these expressions simplify exactly to the forms of $\tilde{H}_{2,t}$ and $\tilde{y}_{2,t}^*$ in (7.9) and (7.11) respectively.

Although this method has no conceptual or technical difficulties associated with it, unlike the three outlined in the previous subsections, it does have the disadvantage of being less flexible than the other three. This is because the same set of explanatory variables and structural terms must account for the Bernoulli and Poisson variation in the data, whereas, in the other three methods, different terms may be used to model the Bernoulli and Poisson variation meaning that a wider selection of models may be used and in theory, therefore, the chances of finding a model which fits the data well are higher for those methods.

In the following chapter we shall apply this method and the methods of §7.3.1 and §7.3.3 to various zero inflated data sets where we shall compare and contrast the different methods with each other as well as with the Poisson state space model.

Chapter 8

Applications of the conditional Bernoulli truncated Poisson state space model

In this chapter several data sets which show signs of an inflated number of zero counts are investigated; each example highlights different aspects of the modelling of zero inflated count data. Three of the four different methods presented in the previous chapter for treating the conditional model are compared in the analysis, they being the missing observation method (§7.3.1), the joint method (§7.3.3) and the univariate signal method (§7.3.4). To establish whether it is actually advantageous to use the conditional models at all, a Poisson model is also fitted to each series for comparison. For all applications the state error terms, η_t, are assumed to be Gaussian distributed as has been the case in the theory so far.

8.1 Goodness of fit for zero inflated count data

Assessing the fit of conditional models is not straightforward since, generally, if we are fitting a conditional model it is because there are a large number of zero counts, so the data is sparse. McCullagh and Nelder (1989) define sparseness in §4.4.5 to mean that a sizeable proportion of the observed counts are small. In the context in which they are writing, they apply the term sparseness to binomial models where a large number of counts are 5 or less. But whether a binomial, a Poisson or a conditional model is being used, extremely sparse data means that generally measures of fit, such as the estimated dispersion, do not work well when used to assess the fit of the model. In a binomial model, for instance, over-dispersion can only occur when the number of trials is greater than 1, which means there can be no over-dispersion for Bernoulli models, but this does not mean all Bernoulli models fit data equally as well as each other. Despite these difficulties, in this chapter we shall attempt to make an assessment of the fit of the various models fitted in the following examples by estimating the dispersion using the three measures given below, and where appropriate comment upon residual plots based on the Pearson residuals corresponding to each of the dispersion estimates.

Firstly, estimating the dispersion for the whole of a conditional model using the Pearson statistic gives the following:

$$\hat{\sigma}_{cP}^2 = \frac{1}{n-m}\sum_{t=1}^{n} \frac{\left(y_t - \frac{\hat{\pi}_t\hat{\lambda}_t}{1-e^{-\hat{\lambda}_t}}\right)^2}{\left(\frac{\hat{\pi}_t\hat{\lambda}_t}{1-e^{-\hat{\lambda}_t}}\right)\left(1+\hat{\lambda}_t - \frac{\hat{\pi}_t\hat{\lambda}_t}{1-e^{-\hat{\lambda}_t}}\right)},$$

where n is the total number of observations and m is the total number of states. However, the data used here are going to contain a high proportion of zero's, so we know at the outset that the above estimated dispersion is unlikely to be a particularly good measure of goodness-of-fit. A better statistic to use would be the following:

$$\hat{\sigma}_{tP}^2 = \frac{1}{n-m_1}\sum_{t=1}^{n} \frac{\left(J(y_t) - \frac{\hat{\lambda}_t}{1-e^{-\hat{\lambda}_t}}\right)^2}{\left(\frac{\hat{\lambda}_t}{1-e^{-\hat{\lambda}_t}}\right)\left(1+\hat{\lambda}_t - \frac{\hat{\lambda}_t}{1-e^{-\hat{\lambda}_t}}\right)},$$

where m_1 is the number of states in the truncated Poisson part of the model. It can be seen that this statistic only applies to the truncated

Poisson part of the model and cannot be used to asses fit in the Bernoulli part. This method of estimating the dispersion follows from the joint method of dealing with zero observations, §7.3.3. The only possible problem with it is the same problem that the joint method has; that is, it will have a disproportionately large number of 1-counts due to all the zero counts being changed to 1 in the function $J(y_t)$. Perhaps a better model still would be to exclude the zero data all together by only summing the non-zero data so that the estimated dispersion function is given by

$$\hat{\sigma}^2_{nz} = \frac{1}{\sum I(y_t) - m_1} \sum_{t=1}^{n} \frac{I(y_t)\left(y_t - \frac{\hat{\lambda}_t}{1-e^{-\hat{\lambda}_t}}\right)^2}{\left(\frac{\hat{\lambda}_t}{1-e^{-\hat{\lambda}_t}}\right)\left(1 + \hat{\lambda}_t - \frac{\hat{\lambda}_t}{1-e^{-\hat{\lambda}_t}}\right)}.$$

This function would ensure that only the counts of 1 or more would be used in estimating the dispersion of the truncated Poisson part of the model.

8.2 Child road fatalities in Southwest England

The first series we examine is road fatalities involving children aged 0 to 11 years inclusive, in Southwest England from 1987 to 2000. In this example all the explanatory variables used in the regression models of chapter 2 are considered and the combination of explanatory variables and structural terms which provides the best fit to the data is used in the analysis.

8.2.1 The missing observation method conditional model fit

Under the missing observations method conditional model, the following model provides a reasonable fit to the data:

$$\begin{aligned}
\theta_{1,t} &= \mu_{1,t} + \beta_{1,1}c_t + \beta_{1,2}r_t, \\
\mu_{1,t+1} &= \mu_{1,t} + \xi_{1,t}, \qquad \xi_{1,t} \sim N(0, \sigma^2_{1,\xi}), \\
\\
\theta_{2,t} &= \mu_{2,t} + \beta_{2,1}c_t, \\
\mu_{2,t+1} &= \mu_{2,t} + \xi_{2,t}, \qquad \xi_{2,t} \sim N(0, \sigma^2_{2,\xi}).
\end{aligned} \tag{8.1}$$

Here, $\beta_{1,1}$, $\beta_{1,2}$ and $\beta_{2,1}$ are non time varying coefficients, c_t is the average monthly minimum temperature in degrees Celsius, and r_t is the total number of millimeters of rainfall per month. With this model there is under-dispersion in one of the three dispersion estimates, $\hat{\sigma}^2_{tP} = 0.90$, the other two are slightly over-dispersed, $\hat{\sigma}^2_{cP} = 1.12$ and $\hat{\sigma}^2_{nz} = 1.18$. The parameter and coefficient estimates and associated standard errors are given in the tables below.

	parameter	log parameter	log-parm Std err
$\hat{\sigma}^2_{1,\xi}$	0.003127	-5.768	1.180
$\hat{\sigma}^2_{2,\xi}$	0.008309	-4.790	1.088

Table 8.1: Missing observation method conditional model parameter estimates, log parameter estimates and standard errors on log parameter estimates for model (8.1).

	Coefficient	Std err
$\hat{\beta}_{1,1}$	0.06771	0.02252
$\hat{\beta}_{1,2}$	-0.01007	0.003267
$\hat{\beta}_{2,1}$	0.1245	0.04186

Table 8.2: Conditional model coefficient estimates and standard errors under the missing observations method.

The coefficient estimates show that the odds of one or more child fatalities occurring increases by 13.26% for an increase of 1 degree Celsius in the minimum daily temperature. The interpretation of the coefficients in the truncated Poisson part of the model is less straightforward because the truncation leads to a complex mean function; however, for small coefficient values, a unit increase in one of the explanatory variables will have virtually the same effect on the observed data under the truncated Poisson model as it would under the Poisson model. Therefore the effect of a unit change in an explanatory variable under the truncated Poisson model can be interpreted in approximately the same way as under the Poisson model. For this example we see that given that one or more child fatalities have occurred, the number increases by approximately 7% for an increase of 1 degree Celsius in the minimum daily temperature; also, given that one or more child fatalities have occurred, the number decreases by approximately 9.6% for every extra 10 millimeters of rainfall. Overall, the model suggests that there are fewer child fatalities when the weather is cold and wet. Assuming that the majority of child fatalities occur on the school run and assuming that children in cars are safer than children on foot or bicycles, the result seems reasonable since in cold or wet weather fewer children are likely to travel to school by foot or bike.

8.2.2 The joint method conditional model fit

Under the joint method conditional model the best fit to the data is obtained again by the use of model (8.1). However, unlike the missing observations method, under the joint method the same model shows over-dispersion using all of the three dispersion estimates: $\hat{\sigma}^2_{cP} = 1.55$, $\hat{\sigma}^2_{tP} = 1.21$ and $\hat{\sigma}^2_{nz} = 1.86$. The parameter and coefficient estimates and associated standard errors are given in the tables below.

	parameter	log parameter	log-parm Std err
$\hat{\sigma}^2_{1,\xi}$	0.008155	-4.809	0.8671
$\hat{\sigma}^2_{2,\xi}$	0.008298	-4.792	1.088

Table 8.3: Joint observation method conditional model parameter estimates, log parameter estimates and standard errors on log parameter estimates for model (8.1).

	Coefficient	Std err
$\hat{\beta}_{1,1}$	0.07826	0.02530
$\hat{\beta}_{1,2}$	-0.01268	0.003300
$\hat{\beta}_{2,1}$	0.1232	0.04154

Table 8.4: Conditional model coefficient estimates and standard errors under the joint method.

The coefficient estimates show that the odds of one or more child fatalities occurring increases by 13.11% for an increase of 1 degree Celsius in the minimum daily temperature. Given that one or more child fatalities have occurred, the number increases by approximately 8.14% for an increase of 1 degree Celsius in the minimum daily temperature; also, given that one or more child fatalities have occurred, the number decreases by approximately 11.9% for every extra 10 millimeters of rainfall.

Despite the different dispersion estimates, the coefficient estimates for this model are remarkably similar to those of the missing observations method and the interpretation of the model is therefore the same. Fewer children walk or cycle to school in cold or wet weather and as a result the decreased exposure leads to a decrease in child fatalities in cold or wet weather. The mismatch in dispersion estimates is interesting considering the similarity of the parameter estimates; also, there does not appear to be a great difference in the model fits when the means are plotted over the data as we see in figure 8.1. However, what figure 8.1 does not show is the difference between the separate plots, that is to say the plots of the separate Bernoulli and truncated Poisson means before being multiplied together in the conditional mean. These separate plots can be seen in figures 8.2 and 8.3. They suggest that there is no difference between the Bernoulli means in either plot, but there is a difference in the truncated Poisson means whereby the joint method mean is slightly lower than the missing observations mean. The 1-counts are the lowest value the truncated Poisson part of the models can take, the presence of a lot of 1-counts will have the effect of reducing the overall mean, despite the presence of a stochastic term, which goes to explain why the truncated Poisson mean in the joint method is lower than in the missing values method, which essentially interpolates between counts of zero instead of replacing them with one. The plot in figure 8.4 shows the actual difference in the truncated means. The plot suggests that the means are most different when there are more zeros, whereas, at the beginning of the series, where there are very few zero counts, the difference is more variable and less one-sided; this is precisely what we would expect given the above explanation. Figure 8.10 in the following example on fatalities in Scotland in snowy weather shows much more clearly the difference in the way the missing observations method and the joint method truncated means vary when observarions are zero-valued.

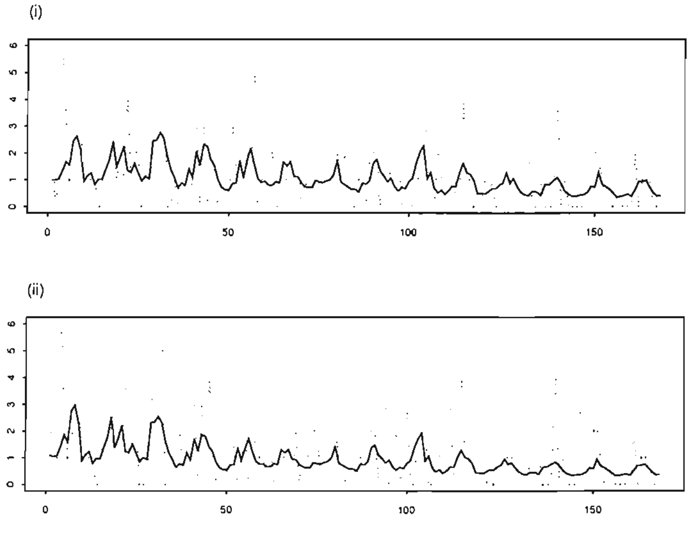

Figure 8.1: Conditional model mean plots using (i) the missing observations method, (ii) the joint method.

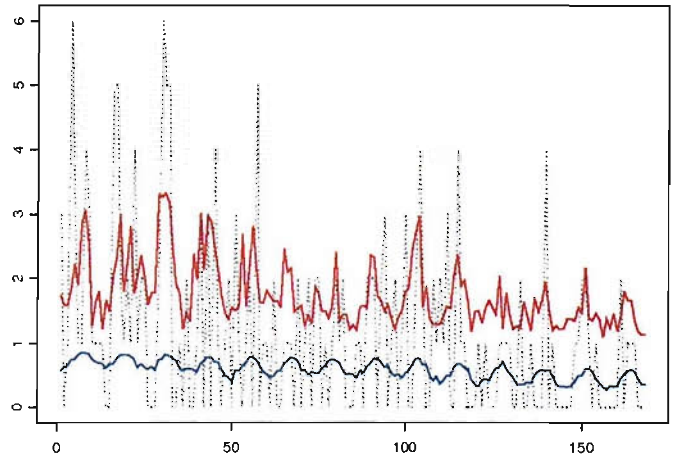

Figure 8.2: Separate Bernoulli and truncated Poisson means for the missing observations method conditional Poisson model.

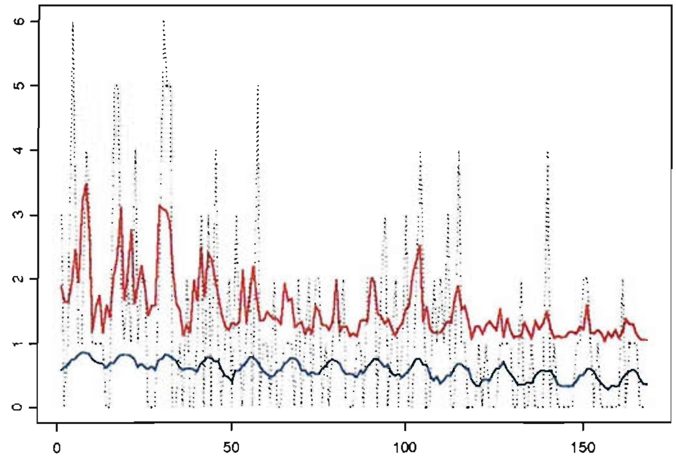

Figure 8.3: Separate Bernoulli and truncated Poisson means for the joint method conditional model.

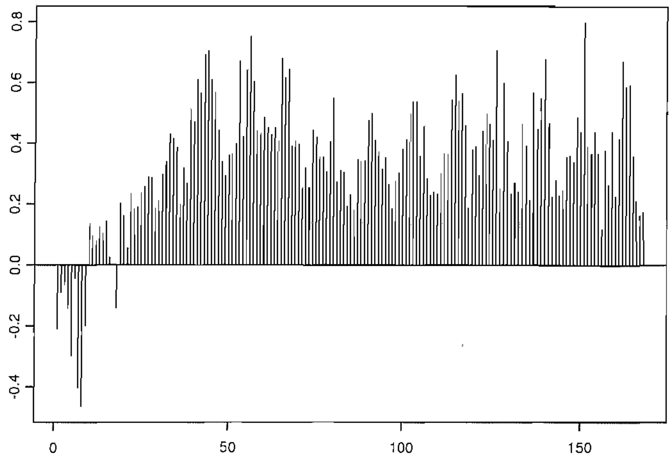

Figure 8.4: Difference plot: truncated Poisson joint method mean subtracted from truncated Poisson missing observations method mean.

8.2.3 The univariate signal conditional model fit

Under the univariate signal conditional model the best fit to the data is obtained by the use of a univariate analogue similar to model (8.1). Note that here there is not the flexibility to include different explanatory variables for the Bernoulli and truncated Poisson parts of the model:

$$\begin{aligned} \theta_t &= \mu_t + \beta_1 c_t + \beta_2 r_t, \\ \mu_{t+1} &= \mu_t + \xi_t, \qquad \xi_t \sim N(0, \sigma_\xi^2), \end{aligned} \tag{8.2}$$

Here, β_1 and β_2 are non time varying coefficients, and again c_t is the minimum temperature variable and r_t is the rainfall variable. The estimated log-level variance for this model is $\log(\hat{\sigma}_\xi^2) = -5.708$ with standard error 0.8666, so $\hat{\sigma}_\xi^2 = 0.003319$. Coefficient estimates and associated standard errors are given in the table below.

It is arguable that we may only use $\hat{\sigma}_{cP}^2$ as a measure of dispersion under the univariate signal method since only one θ_t is used to explain the variation in the data. However, the mean is still constructed from two

	Coefficient	Std err
$\hat{\beta}_1$	0.05806	0.02362
$\hat{\beta}_2$	-0.007418	0.002461

Table 8.5: Conditional model coefficient estimates and standard errors under the univariate signal method.

different functions of θ_t, one which describes the variation in the presence or absence of data, π_t, and the other which describes the variation in the non-zero data, $\lambda_t/(1-e^{\lambda_t})$, so it is still valid and interesting to compare the three different estimates of dispersion in the univariate signal case. The model shows under-dispersion using all of the three dispersion estimates: $\hat{\sigma}^2_{cP} = 0.93$, $\hat{\sigma}^2_{tP} = 0.78$ and $\hat{\sigma}^2_{nz} = 0.94$.

Because the univariate signal conditional model uses the same signal in both the Bernoulli and truncated Poisson part of the model, it is not necessary to interpret the two parts of the series separately. However, the univariate signal conditional mean is a complex function of θ_t and does not easily lend itself to a simple interpretation. But despite its apparent complexity, the univariate signal conditional mean and the Poisson mean are actually very similar to one another:

$$E(Y) = \frac{\pi_t\lambda_t}{1-e^{\lambda_t}} = \frac{\exp(2\theta)}{[1+\exp(\theta)][1-\exp\{-\exp(\theta)\}]} \approx \exp(\theta), \quad \forall\ \theta.$$

This means that we can interpret the coefficient values in the univariate signal conditional model in approximately the same way as for a standard log-linear Poisson model. So we may say that the number of child fatalities increases by approximately 6% for every 1 degree Celsius rise in the minimum daily temperature and the number of child fatalities decreases by approximately 7.15% for every extra 10 millimeters of rainfall. Again, these estimates are in line with those of the previous subsections.

8.2.4 The Poisson model fit

The best fit under the Poisson model is obtained by the use of exactly the same model as for the univariate signal conditional model, (8.2). The estimated log-level variance for this model is $\log(\hat{\sigma}_{\xi}^{2}) = -5.570$ with standard error 0.7781, so the estimated level variance is $\hat{\sigma}_{\xi}^{2} = 0.003809$. For the Poisson model there is, of course, only one dispersion estimate using the Pearson statistic, but it shows very little over-dispersion with an estimated dispersion of $\hat{\sigma}_{cP}^{2} = 1.16$. The coefficient estimates and associated standard errors are given in the table below.

	Coefficient	Std err
$\hat{\beta}_1$	0.06762	0.02124
$\hat{\beta}_2$	-0.007176	0.002248

Table 8.6: Poisson model coefficient estimates and standard errors.

The coefficient estimates show very similar results to the univariate signal conditional model; the number of child fatalities increases by 6.996% for every 1 degree Celsius rise in the minimum daily temperature and the number of child fatalities decreases by 6.925% for every extra 10 millimeters of rainfall.

Figure 8.5 shows the univariate signal conditional mean and the Poisson mean plotted over the data. It is apparent that both these plots appear remarkably similar to one another and also to the fitted mean plots of the two bivariate conditional models in figure 8.1. Both the Poisson and the univariate signal conditional model are based on model (8.2), so we can compare the mean of the Poisson and univariate signal models using a difference plot as we did for the truncated means of the two bivariate models; figure 8.6 shows this difference plot. The pattern of the seasonal variation in this plot suggests that the Poisson mean has a consistently larger amplitude than the univariate signal conditional mean, i.e., peaks are higher and troughs are lower for the Poisson mean. Also apparent in this plot is that for the larger valued counts at the beginning of the series, the Poisson mean is higher than the univariate conditional mean.

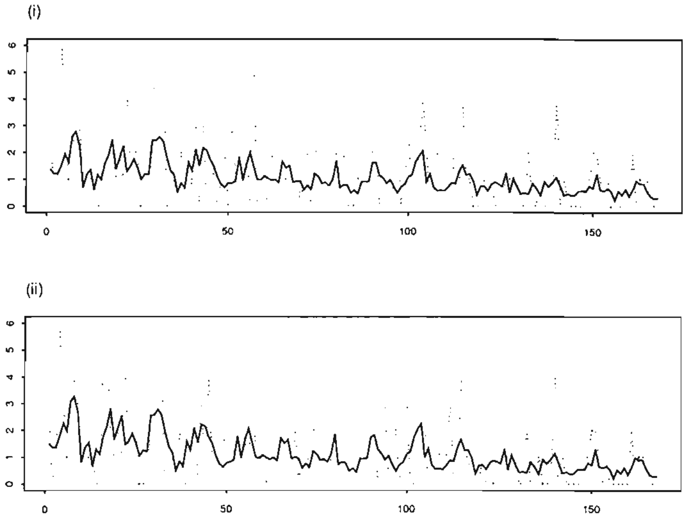

Figure 8.5: Fitted means plots using (i) the univariate signal conditional model, (ii) the Poisson model.

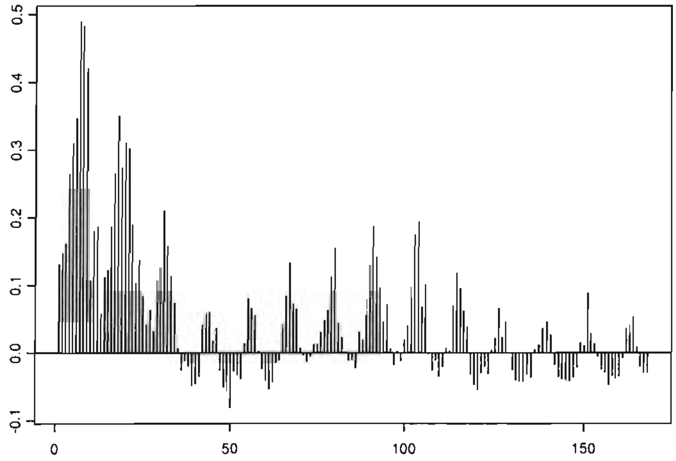

Figure 8.6: Difference plot: univariate signal conditional mean subtracted from Poisson mean.

8.3 Analysis of Scottish road fatalities occurring in snowy weather

8.3.1 Using a structural model

This is a somewhat contrived example since the series in question is the road fatalities which occurred when it was snowing in Scotland from January 1979 to December 2000. Because the data is based upon a highly seasonal event, it is no surprise that the nature of the series is highly seasonal as illustrated in figure 8.7. In this example the highly regular pattern of the data combined with the use of only structural terms are used to illustrate a potential pitfall of zero inflated count modelling.

When fitting a univariate conditional model or a Poisson model to this series, a good choice for a structural model would be something which

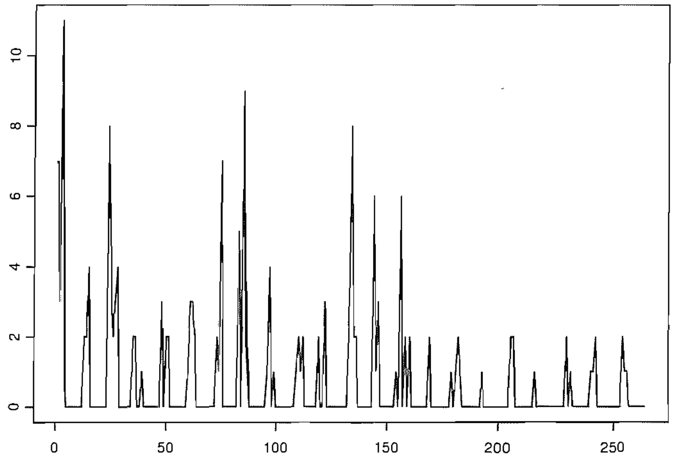

Figure 8.7: Scottish road fatalities occurring in snowy weather

captures the seasonality, such as the following model:

$$\begin{aligned}
\theta_t &= \mu_t + \gamma_t, \\
\mu_{t+1} &= \mu_t + \xi_t, \qquad \xi_t \sim N(0, \sigma_\xi^2), \\
\gamma_{t+1} &= -\sum_{j=0}^{s-1} \gamma_{t+1-j} + \omega_t, \qquad \omega_t \sim N(0, \sigma_\omega^2).
\end{aligned} \tag{8.3}$$

The highly seasonal nature of the series extends to both parts, that is to say, the non-zero data, $y_t \geq 1$, is highly seasonal and the presence-absence data, $I(y_t)$, is also highly seasonal. This means that when considering a conditional model with bivariate signal, the best choice of model is one with a seasonal term in it for both $\theta_{1,t}$ and $\theta_{2,t}$. Thus we might use the bivariate signal equivalent to (8.3):

$$\begin{aligned}
\theta_{1,t} &= \mu_{1,t} + \gamma_{1,t}, \\
\mu_{1,t+1} &= \mu_{1,t} + \xi_{1,t}, \qquad \xi_{1,t} \sim N(0, \sigma_{1,\xi}^2), \\
\gamma_{1,t+1} &= -\sum_{j=0}^{s-1} \gamma_{1,t+1-j} + \omega_{1,t}, \qquad \omega_{1,t} \sim N(0, \sigma_{1,\omega}^2), \\
\\
\theta_{2,t} &= \mu_{2,t} + \gamma_{2,t}, \\
\mu_{2,t+1} &= \mu_{2,t} + \xi_{2,t}, \qquad \xi_{2,t} \sim N(0, \sigma_{2,\xi}^2), \\
\gamma_{2,t+1} &= -\sum_{j=0}^{s-1} \gamma_{2,t+1-j} + \omega_{2,t}, \qquad \omega_{2,t} \sim N(0, \sigma_{2,\omega}^2).
\end{aligned} \tag{8.4}$$

8.3.2 Model estimation failure

Models (8.3) and (8.4) contain only structural terms and are therefore not particularly informative, in fact the series that they are modelling itself is not particularly interesting since it is obvious that fatalities which occur during snowy weather will occur during the winter months. However, the above models do fit the data very well; in fact it is because the models fit the data so well that the fitting process fails before the optimal models are found. The reason that the optimisation process fails is that the estimating equations used in the linear Gaussian approximating model become numerically unstable. The numerical instability is generally caused by the estimates of the signal, $\tilde{\theta}_t$ or $\tilde{\theta}_{1,t}$ and $\tilde{\theta}_{2,t}$, becoming so large that the computer is no longer able to effectively calculate the linear Gaussian approximating model quantities such as $\tilde{H}_{1,t}$. It is easier to understand what is going on if $\tilde{\theta}_{1,t}$ and $\tilde{\theta}_{2,t}$ are converted into a form involving $\tilde{\pi}_t$ and $\tilde{\lambda}_t$ using the link functions (7.6). Here it becomes clear that a huge value of $\tilde{\theta}_{1,t}$ or $\tilde{\theta}_{2,t}$ corresponds to $\tilde{\pi}_t$ or $\tilde{\lambda}_t$ coming too close to its limiting value; that is, $\tilde{\pi}_t$ getting too close to 0 or 1 and $\tilde{\lambda}_t$ getting too close to 0. For example if we take $\tilde{H}_{1,t}$ in (7.8) and convert the equation into a form using $\tilde{\lambda}_t$ and then take $\tilde{\lambda}_t = 0$, a problem of the form 0/0 occurs regardless of whether $I(y_t)$ is zero or one:

$$\begin{aligned}\tilde{H}_{1,t} &= \frac{\exp\{\exp(\tilde{\theta}_{1,t})\} - 2 + \exp\{-\exp(\tilde{\theta}_{1,t})\}}{I(y_t)\exp(\tilde{\theta}_{1,t})[\exp\{\exp(\tilde{\theta}_{1,t})\} - 1 - \exp(\tilde{\theta}_{1,t})]} \\ &= \frac{\exp(\tilde{\lambda}_t) - 2 + \exp(-\tilde{\lambda}_t)}{I(y_t)\tilde{\lambda}_t[\exp(\tilde{\lambda}_t) - 1 - \tilde{\lambda}_t]}. \end{aligned} \tag{8.5}$$

To avoid these numerical instabilities occurring, upper and lower limits are needed on $\tilde{\theta}_{1,t}$ and $\tilde{\theta}_{2,t}$ so that $\tilde{H}_{1,t}$ and other linear Gaussian approximating model quantities do not become incalculable. However, rather than randomly choosing these limits it would be better to make them coincide with upper and lower constraints on the values of $\tilde{\pi}_t$ and $\tilde{\lambda}_t$. For example if $\tilde{\lambda}_t$ is constrained so that $\tilde{\lambda}_t \geq 0.01$, then this corresponds to constraining $\tilde{\theta}_{1,t} \geq \log(0.01)$; by applying this minor restriction $\tilde{H}_{1,t}$ now becomes a calculable quantity. Similarly to the log-link function relating $\tilde{\lambda}_t$ to $\tilde{\theta}_{1,t}$, we find from the logistic link function in (7.6) that constraints on $\tilde{\pi}_t$ of $0.01 \leq \tilde{\pi}_t \leq 0.99$ correspond to constraints of $-\log(99) \leq \tilde{\theta}_{2,t} \leq \log(99)$ on $\tilde{\theta}_{2,t}$.

Applying the above constraints to $\tilde{\theta}_t$ in the univariate signal conditional model is not so straightforward as for the two bivariate signal methods because we may derive the constraints from $\tilde{\pi}_t$ or from $\tilde{\lambda}_t$, since $\tilde{\pi}_t$ and $\tilde{\lambda}_t$ are simply different functions of $\tilde{\theta}_t$. However, the problem is not serious since the constraints derived for $\tilde{\theta}_t$ are very similar whether they have been obtained from $\tilde{\pi}_t$ or from $\tilde{\lambda}_t$. It is probably best, however, to apply the limits corresponding to $\tilde{\pi}_t$ since $\tilde{\pi}_t$ has upper and lower constraints corresponding to upper and lower constraints on $\tilde{\theta}_t$, whereas, $\tilde{\lambda}_t$ is only restricted in one direction allowing $\tilde{\theta}_t$ to become very large in the other. So using the logistic link, (7.17), we use the limits of $-\log(99) \leq \tilde{\theta}_t \leq \log(99)$ on $\tilde{\theta}_t$ under the univariate signal method corresponding to $0.01 \leq \tilde{\pi}_t \leq 0.99$.

Another cause of model estimation failure with the conditional model applies solely to the missing observations method, §7.3.1. The problem here is that the Kalman filter used in the calculation of $\tilde{\theta}_{1,t}$ will not start if any of the first few $\tilde{y}^*_{1,t}$ values are missing. In fact, $m_1 + 1$ non missing values are needed from $\tilde{y}^*_{1,t}$ to $\tilde{y}^*_{m_1+1,t}$ for the Kalman filter to successfully initialise, where m_1 is the total number of states in the truncated Poisson part of the model. If fitting a local-level model to the data, this may not present a problem as only the first two $\tilde{y}^*_{1,t}$ must not be missing. However, for a seasonal model with local level trend such as (8.4), the first 13 $\tilde{y}^*_{1,t}$ must be present for the initialisation of the Kalman filter to work. In the missing observations method $\tilde{y}^*_{1,t}$ is treated as a missing value when $y_t = 0$, and since there are plenty of months each year when there is no snowy weather and thus no road fatalities during snowy weather, we know that some of the first 13 $\tilde{y}^*_{1,t}$ values must be missing. This amounts to having to treat the first few $\tilde{y}^*_{1,t}$ as they would be treated in the joint method (§7.3.3), that is, using $J(y_t)$ in the calculation of the first 13 $\tilde{y}^*_{1,t}$ so that the Kalman filter may be initialised.

8.3.3 The conditional and Poisson model fits

Applying the constraints in the previous section on the values of $\tilde{\theta}_{1,t}$, $\tilde{\theta}_{2,t}$ and $\tilde{\theta}_t$, and treating the first 13 $\tilde{y}^*_{1,t}$ values in the missing observations method as they are treated in the joint method means that models (8.4) and (8.3) are now estimable. It can be seen from figure 8.8 and 8.9 that the fit is very similar and very good under all three conditional model estimation methods, and again we find that the fit of the Poisson model is almost identical to the that of the conditional models. Difference plots have not been shown for this example because they indicate an identical story to the difference plots of the child fatality example. That is, they show again that the truncated mean of the joint method conditional model is lower than that of the missing observations method conditional mean; and the seasonal amplitude of the Poisson mean is slightly greater than that of the univariate conditional mean. Parameter estimates for all four models are shown in the tables below.

	parameter	log parameter	log-parm Std err
$\hat{\sigma}^2_{1,\xi}$	0.01720	-4.063	0.9067
$\hat{\sigma}^2_{1,\omega}$	0.1026	-2.277	0.6049
$\hat{\sigma}^2_{2,\xi}$	0.003752	-5.585	1.152
$\hat{\sigma}^2_{2,\omega}$	≈ 0	-31.54	7098

Table 8.7: Conditional model parameter estimates, log parameter estimates and standard errors on log parameter estimates under the missing observations method for model (8.4).

	parameter	log parameter	log-parm Std err
$\hat{\sigma}^2_{1,\xi}$	0.02896	-3.541	0.6764
$\hat{\sigma}^2_{1,\omega}$	0.1108	-2.199	0.5517
$\hat{\sigma}^2_{2,\xi}$	0.004364	-5.434	1.160
$\hat{\sigma}^2_{2,\omega}$	≈ 0	-30.56	3701

Table 8.8: Conditional model parameter estimates, log parameter estimates and standard errors on log parameter estimates under the joint method for model (8.4).

	parameter	log parameter	log-parm Std err
$\hat{\sigma}^2_{\xi}$	0.01447	-4.236	0.8062
$\hat{\sigma}^2_{\omega}$	0.04919	-3.012	0.8073

Table 8.9: Conditional model parameter estimates, log parameter estimates and standard errors on log parameter estimates under the univariate signal method for model (8.3).

	parameter	log parameter	log-parm Std err
$\hat{\sigma}^2_\xi$	0.02346	-3.753	0.8123
$\hat{\sigma}^2_\omega$	0.1074	-2.231	0.5838

Table 8.10: Poisson model parameter estimates, log parameter estimates and standard errors on log parameter estimates for model (8.4).

The log-parameter estimate standard errors are all reasonable apart from the standard errors on $\log(\hat{\sigma}^2_{2,\omega})$ in the bivariate method tables, which are large. This either suggests that the terms modelling the seasonal aspect of $I(y_t)$ need not be time variable and could be held constant under those two methods, or it could suggest that seasonal terms are not needed because they do not fit the $I(y_t)$ data very well. However, it can be seen by looking at a split plot for those two methods (figure 8.10) that the former reason is more likely as the Bernoulli mean is a definite regular seasonal pattern. Under the missing observations method there is under-dispersion using all three measures of dispersion defined in §8.1: $\hat{\sigma}^2_{cP} = 0.56$, $\hat{\sigma}^2_{tP} = 0.57$ and $\hat{\sigma}^2_{nz} = 0.95$. Under the joint method there is over-dispersion in two of the measures and under-dispersion for one: $\hat{\sigma}^2_{cP} = 1.44$, $\hat{\sigma}^2_{tP} = 0.54$ and $\hat{\sigma}^2_{nz} = 2.05$. The differences in dispersion estimates are again a result of the difference in the truncated Poisson means, as was the case in the previous example. We find that under the univariate signal method there is more under-dispersion in each estimate than either of the other two conditional methods: $\hat{\sigma}^2_{cP} = 0.50$, $\hat{\sigma}^2_{tP} = 0.31$ and $\hat{\sigma}^2_{nz} = 0.80$.

Under the Poisson state space model using (8.3), we must again use limits on $\tilde{\theta}_t$ in the linear Gaussian approximating model since here, too, the fitting procedure fails due to infinite or undefined values. For consistency, and since no advantage is gained by doing otherwise, we choose the same constraint on λ_t as before, namely $\lambda_t \geq 0.01$, which implies $\tilde{\theta}_t \geq \log(0.01)$ under the log-link. The estimated dispersion under the Poisson model is 0.49 showing a lot of under-dispersion. Note that because of the regularity of the seasonal pattern in this series and the models with only structural terms fitted to it, it is no surprise that most of the models in this example show under-dispersion.

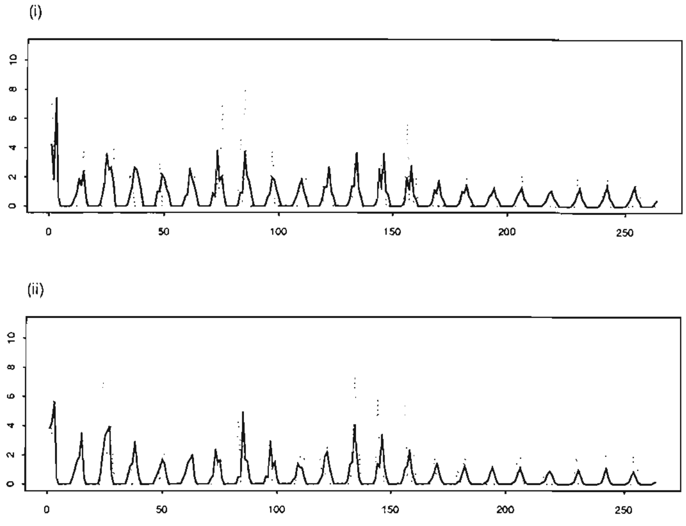

Figure 8.8: Fitted conditional model to Scottish snow data using model (8.4) under (i) the missing observations method, (ii) the joint method.

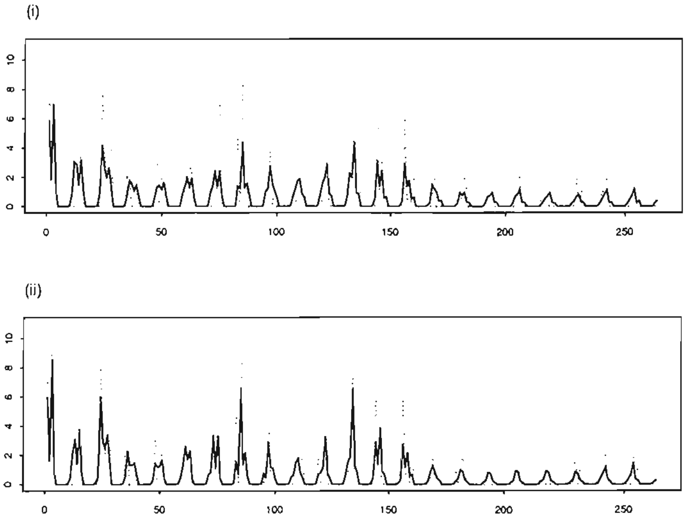

Figure 8.9: Fitted models to Scottish snow data using model (8.3) under (i) the univariate signal conditional model, (ii) the Poisson model.

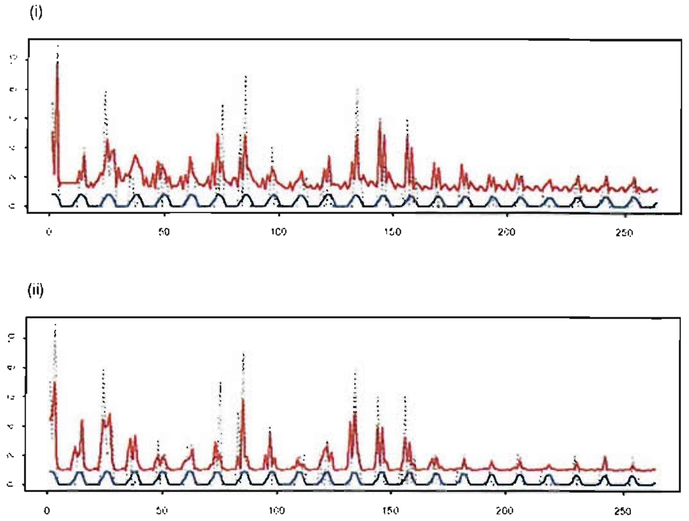

Figure 8.10: Conditional model split into Bernoulli and truncated Poisson parts fitted to the Scottish snow data using model (8.4) under (i) the missing observations method, (ii) the joint method.

8.4 Noticeable differences between conditional and Poisson models

So far in this chapter, although there are differences in dispersion estimates, it would appear that the various conditional models and the Poisson model all produce fairly similar fits, and it is difficult to tell which offers the best model for the data; also, the dispersion estimates are probably unreliable due to the sparse nature of the data. However, the similarity of fits could be due to the dominance of the explanatory variables in the child fatality example and the regularity of the data series in the Scottish snow example. When simple local-level structural models are fitted to zero inflated count series there can be more variation in the fits.

Figures 8.11 and 8.12 show local-level structural models applied to data of the monthly total of fatalities on non built-up roads in Greater London from 1979 to 2000 inclusive. The local-level models are applied under all three conditional modelling methods as well as the Poisson model.

Surprisingly, the two models which appear to have the better fit, the joint method conditional model and the Poisson model, have worse dispersion estimates than the other two models. The joint method dispersion for the whole model is $\hat{\sigma}^2_{cP} = 1.68$ and the dispersion for the Poisson model is $\hat{\sigma}^2 = 1.36$. Comparatively, the dispersion estimates for the missing observations method and the univariate signal method are $\hat{\sigma}^2_{cP} = 1.13$ and $\hat{\sigma}^2_{cP} = 1.16$ respectively. The non-zero only dispersion estimates for the three conditional models also show more over-dispersion for the joint method, with $\hat{\sigma}^2_{nz} = 2.08$, than the missing observations method and the univariate signal method, with dispersion estimates of $\hat{\sigma}^2_{nz} = 1.39$ and $\hat{\sigma}^2_{nz} = 1.46$ respectively.

It has to be said, though, that if one were to choose between the Poisson and conditional models at this point, one would probably choose the Poisson model as it has the advantage of simpler modelling equations and a simpler interpretation. However, there are some circumstances where the conditional model can at least appear to provide a better fit to the data than the Poisson model.

The above example is one in which the Poisson model differs from the conditional models by showing more variation in the mean; although, as the dispersion estimates suggest, this does not necessarily mean the Poisson model provides the better fit. However, for some data sets this situation is reversed and the conditional model mean appears to vary more and provide the best fit to the data. The number of fatalities on motorways in Northern England from 1979 to 2000 is one such series. Figure 8.13 shows that a univariate signal conditional local-level model appears to give a better fit to the data than a Poisson local-level model.

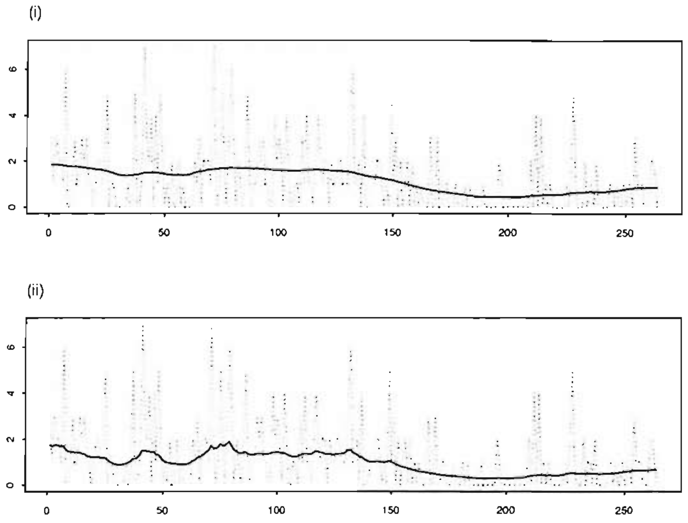

Figure 8.11: Fatalities on non built-up roads in Greater London with (i) fitted local level missing observations method conditional mean, (ii) fitted local level joint method conditional mean.

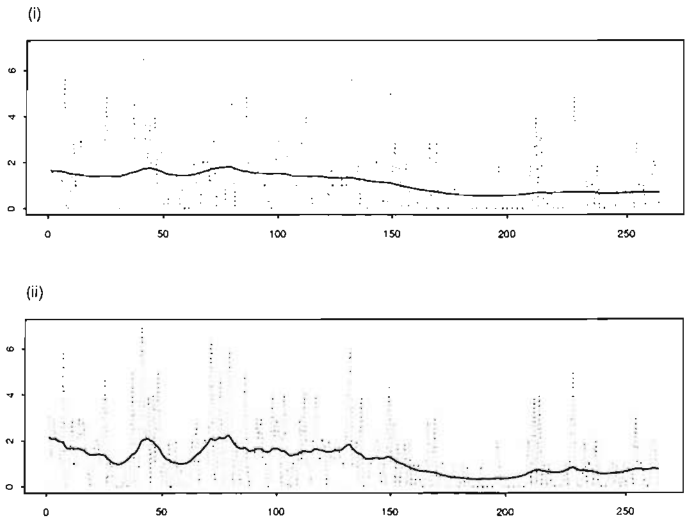

Figure 8.12: Fatalities on non built-up roads in Greater London with (i) fitted local level univariate conditional mean, (ii) fitted local level Poisson mean.

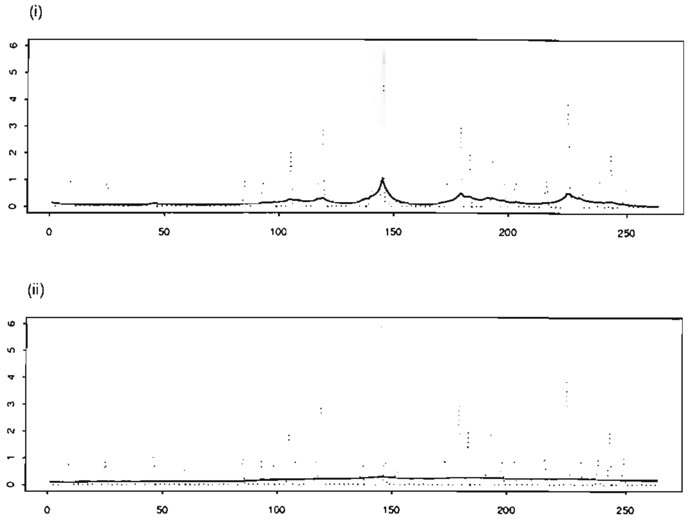

Figure 8.13: Fitted local level models to Northern English motorways data using model under (i) the univariate signal conditional model, (ii) the Poisson model.

A curious feature of this example is the large variation in the three dispersion estimates for the univariate signal conditional model. The dispersion estimates are $\hat{\sigma}^2_{cP} = 1.02$, $\hat{\sigma}^2_{tP} = 0.38$ and $\hat{\sigma}^2_{nz} = 2.47$. For comparison, the Poisson dispersion estimate is $\hat{\sigma}^2 = 1.71$. More generally, while it is true that the local-level model is unlikely to be the best fit to most data series, the problem with this series is that with only 34 non-zero observations out of 264, it is really too zero inflated and there is not enough information in the non-zero part of the series to fit any sort of a meaningful model. Also, it is only in the case of the univariate signal conditional model that a seemingly better fit is achieved for this series; for the other two methods there is even less variation than for the Poisson model.

Chapter 9

Conclusion

It is difficult to draw any one conclusion from this thesis as it has been as much about the application of existing modelling techniques as it has been an investigation into new methodology. However, where the new methodology is concerned, the thesis is split into two investigations: one into the effect of daylight on road accidents and the other into appropriate state space time series models for zero inflated count data. As such, the conclusion has been divided into two sections which conclude the two investigations separately.

9.1 The investigation into the effect of daylight on road accidents

The analysis of car occupant accidents in chapter 2 showed that daylight, as measured from sunrise to sunset, was a significant predictor under linear and log-linear regression models for the total number of car occupant deaths and injuries, and also the number of fatalities, in Scotland and Southwest England. In short, every model for which the daylight variable was tested showed that it had a significant effect and that effect was negative such that more daylight corresponded to fewer accidents. It was also evident from the daylight coefficients that in all models considered, daylight had a stronger effect on accidents in Scotland than in Southwest England, which is probably due to the greater seasonal variation in daylight in Scotland.

The means by which the conclusions in the previous paragraph were drawn were, of course, conventional in statistical terms, they being linear and log-linear regression models. A cruder, less conventional method of assessing the effect of daylight was introduced in chapter 4 based on using the difference in latitude between Scotland and Southwest England, the northern-most and southern-most regions of the UK respectively. The double-differencing method was an ad hoc idea designed to show that daylight had a significant effect on road accidents by using the fact that morning rush hour in Scotland during the months of December and January is spent in darkness while it is light in Southwest England the whole year round. Applying the technique to the raw Scottish and Southwest data revealed that although there was a weak effect, as the means of the double-differences were positive as hoped, they were not significantly different from zero.

Although double-differencing was designed to minimise the effects of other factors influencing the level of road accidents, there was always the possibility that such factors had not been entirely eliminated. To ascertain whether the method would have worked had there not been factors other than daylight influencing the numbers of accidents in the two regions, the double-differencing method was applied to the Scottish and Southwest data which had had all non-daylight variables removed. The idea was that this would provide a fairer comparison between the Scottish and Southwest English data. However, despite the removal of all non-daylight effects from the data, the results showed again that there was no effect on road accidents due to the darkness in morning rush hour in December and January in Scotland using the double-differencing method.

Further analysis identified another unknown source of seasonal variation in three of the four series under investigation which was conflicting with the darkness in morning rush hour in December and January that the double-differencing method was trying to estimate. It is possible that this extra variation was the cause of the double-differencing technique not showing a significant result. One option for dealing with this problem might be to fit a dummy variable to account for the extra unforeseen seasonal variation. A simple example would be a variable which takes the value 1 for December and January and takes the value zero for all other months. However, the drawback with this approach is that the new variable may not only account for the unforeseen variation but may also eliminate the variation that we are trying to estimate.

Apart from the unknown source of variation, there are other possible reasons why the double-differencing method failed to show that darkness in morning rush hour in December and January in Scotland is a strong enough effect to make a significant difference to casualty numbers. Firstly, the morning rush hour is only one hour during the day and although there is a peak in fatalities during this time, it is still only going to account for a small proportion of the overall number of road accidents each day, and consequently each month. So, any analysis based on differences in total monthly accident numbers is not likely to show big differences when the only thing influencing these differences is a change in accidents numbers during a small part of each day. Comparing one months casualty figures with another months is possibly too blunt an instrument to use to analyse potentially quite small changes. Secondly, in previous studies on the effect of daylight on road accidents, a popular idea has been to use the hour change from BST to GMT or vice versa, and measure the numbers of accidents for several weeks before the change and several weeks after. The idea is that the hour change is abrupt and therefore catches some road users by surprise, so it is seen that there is often a sudden jump or sudden fall in road accidents after the hour change. The light level in morning rush hour in Scotland, despite being noticeably darker than in Southwest England for the months of December and January, will nonetheless change gradually over the days and weeks from November to December and from January to February, giving drivers and other road users time to adjust, to a certain extent. Again, comparing one months data with the next is possibly too crude a method to measure this gradual change.

To improve the chances of the double-differencing method showing significant differences between the December and January casualty levels in Scotland and Southwest England, it could be refined by the use of higher frequency data such as weekly data. In this way there would be more data to work with and instead of using the difference between one week and the next, the method could be modified to measure the rate of change over a period of six weeks or so from November to December and again from January to February. This would be a small enough time period for the effects of other factors influencing the numbers of casualties to be minimised and should show a steeper rate of change in Scotland than Southwest England.

9.2 The investigation into zero inflated count data and the conditional model

From the evidence of the previous chapter it can be concluded that a conditional model can provide a satisfactory fit to zero inflated count time series data. However, from the examples considered, what is more evident is that in all but the most extreme case the conditional models do not seem to provide a markedly different or better fit to the data than the Poisson model. It also has to be said that the Poisson model has a much simpler interpretation than the conditional models, which is evident in the interpretation of the coefficient estimates in the child fatality example.

On the face of it, it does not seem to make sense that the conditional model, which is theoretically so well suited to the modelling of zero inflated counts, should be no better in practice than the Poisson model, but there are various reasons why the Poisson model does well modelling zero inflated data. Firstly, none of the data examined in any of the examples actually violate the principal Poisson modelling assumptions, i.e., they are all integer valued, all non-negative and theoretically have no upper limit; so looking at the situation from this point of view, the Poisson model will always be plausible even if the fit is not good. Of course, the other key assumption for using a Poisson model is that there should be Poisson variation in the data. In a study on the abundance of Leadbeaters possum in south east Australia, Welsh et al. (1996) point out that from the Poisson model the predicted number of sites with no animals, for a total of n sites, should be $n \exp\{-\lambda(z)\}$, where $\lambda(z)$ is the Poisson mean for the set of explanatory variables, z. Since there are typically many more sites with no animals than would be expected from this model, the fit is usually poor. Figure 9.1 illustrates this point, with regards to the Scottish snow data, showing the actual distribution of counts with the idealised Poisson distribution of counts for all $n = 264$ months. However, despite this theoretical objection to the Poisson model, the conditional model does not seem able to improve upon the fit in the examples we have investigated.

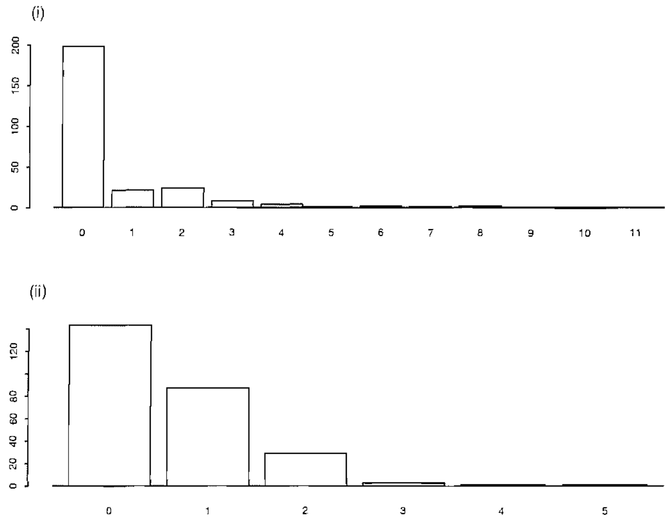

Figure 9.1: (i) Distribution of counts in the Scottish snow fatalities data, (ii) Idealised Poisson distribution with equal mean to the Scottish snow fatalities data.

But the fact of the matter is that each observation is a realisation of a random variable with a different mean dictated by explanatory variables and structural terms; so, if under the Poisson model the mean is near to zero then a zero observation is quite probable. In other words, the explanatory variables and structural terms which make up the mean may soak up the apparent zero inflation in the data, leaving us without any at the end.

If the Poisson model does not provide a good fit to the data, at least in theory, then perhaps the similarity in fits between the Poisson and conditional models may instead be due to some inadequacy in the conditional models, meaning that they also do not provide a good fit. According to the Leadbeaters possum study, the conditional model provided a good fit to the data in that instance; however, there are two differences between the examples in chapter 8 and the Leadbeaters possum study. Firstly, the data used in the possum study was not time series data and secondly, a generalised linear model, rather than a state space time series model, was fitted to it.

The first point is a key difference between the two studies. The fact that the data is time series means that difficulties arise with the conditional model which would not arise from a conditional model applied to non time series data. Specifically, the difficulties relate to the treatment of the truncated Poisson part of the model while the data is zero. Four methods for dealing with this problem were introduced in chapter 7 and three of those methods were tested in chapter 8. Although there were differences in the fits using the three methods, such as the differences in the dispersion estimates due to the different way in which the truncated Poisson part of the model was dealt with, none of the three methods seemed to yield a markedly different or better fit to the data than the Poisson model.

With regards to the second point, it is not so obvious why the type of model fitted to the data should make a difference to the performance of the conditional model in relation to the Poisson model. After all, if a state space time series Poisson model compares similarly to a state space time series conditional model, then one would assume that a generalised linear Poisson regression model would also compare similarly to a generalised linear conditional regression model. But the difference is that the state space time series models can contain structural terms and not just explanatory variables, and those structural terms can behave differently in Poisson and conditional state space models. The Northern motorways example illustrates this difference well, but for some less extremely zero inflated data there can still be a discernible difference between Poisson and conditional models.

It is difficult to establish any one factor as the main reason why the conditional model does not seem to provide a better fit to the data than the Poisson model. Even establishing what a good fit means for zero inflated data modelled by a state space model is not straightforward and it cannot conclusively be said that the conditional model is definitely not as good as the Poisson model. But, it would be useful to more thoroughly investigate the relative merits of the Poisson and conditional state space models. For instance, it would certainly be informative to test Poisson and conditional state space models on different sources of time series data, not road accident data. It would also be useful to examine the performance of state space models in general, which do not necessarily assume Gaussian distributed state errors. Another possibility would be to find alternatives to the rather ad hoc procedures developed in chapter 7 for treating the zero observations in the truncated Poisson part of the conditional model. Also, it would be informative to compare generalised linear Poisson and conditional regression models for the data in this study, to see if the better fit achieved in the Leadbetter's possum study was a result of the type of model used rather than the nature of the data analysed. Finally, another possibility still would be to try a different approach to the modelling of zero inflation all together, such as using the zero inflated Poisson model.

Appendix A

Descriptive summary of the state space modelling code used for the examples in this thesis

This appendix gives a summary of the code I wrote and used for the analysis of state space time series models. Presenting the code in its entirety here is overly cumbersome and takes up some 35 pages when shown using the script size in the extracts of code that are presented, therefore only the key points are drawn to the readers attention. The first section outlines the code used for the Gaussian state space models and the second section outlines that used for the non-Gaussian state space models.

A.1 Gaussian state space modelling code

The key function in FinMetrics for calculating the optimal error variance parameter estimates for Gaussian state space models is the SsfFit function. The SsfFit function must be supplied the initial log-parameter estimates, the data and a state space model; from these it will calculate and maximise the log-likelihood and when it has done so, will output the optimised parameter estimates. In S-Plus version 7 and FinMetrics version 2, it will also calculate the standard errors on the parameter estimates.

From the basis of the SsfFit function I wrote two functions. The SsfGaussAnalysis function took any data, stochastic components and explanatory variables the user wished to use and created from them a structural time series model. The functional form of SsfGaussAnalysis is shown below to illustrate what information is needed to create Gaussian state space analysis function flexible enough to handle the majority of structural time series.

```
SsfGaussAnalysis = function(mY, covar=NULL, covar.split=rep(0,ncol(mY)),
                            irregular=rep(0.01,ncol(mY)), level=rep(0.01,ncol(mY)),
                            slope=rep(NA,ncol(mY)), seasonalDummy=rep(NA,ncol(mY)),
                            seasonalTrig=rep(NA,ncol(mY)), seasons=12)
{
    ## mY = input data (matrix for multivariate)
    ## covar = explanatory variables matrix
    ## covar.split = which covar's are associated with which input data vector
    ## irregular = model error variance (epsilon)
    ## level = level error variance (xi)
    ## slope = slope error variance (zeta)
    ## seasonalDummy/seasonalTrig = seasonal error variance (omega)
    ## seasons = number of seasons, i.e. 12 per year, 7 per week, etc
```

The SsfFit function takes the components of a structural time series model and puts them into a state space form, which must also be provided. The Kalman filter and smoother are run on this state space form and from them a log-likelihood is calculated which is maximised with subsequent iterations. The SsfFit function will, of course, change parameter estimates with each iteration, so the state space form must be able to handle this. FinMetrics provides a routine called GetSsfStsm, which will convert Gaussian structural models into state space form, but this routine cannot handle explanatory variables or time varying error variance parameters such as are used in the linear Gaussian approximating models of chapters 6, 7 and 8. As such, the GetSsfStsm routine is far too limited for the purposes of this thesis. So I wrote my own state space form function, SsfExfSsf, which could convert almost any structural time series model from any exponential family distribution into state space form; this allowed me to use the same function for non-Gaussian state space analysis.

Once the optimal error variance parameter estimates had been calculated, the SsfGaussAnalysis function then put them back into the state space form function, SsfExfSsf, which was then passed to the FinMetrics routines: SsfCondDens and SsfMomentEst. These would calculate the final Kalman smoothed state estimates, $E(\boldsymbol{\alpha}_t|\boldsymbol{y})$, and state variances, $Var(\boldsymbol{\alpha}_t|\boldsymbol{y})$, which would hence give the model fit, coefficients and associated standard errors for the explanatory variables.

A.2 Non-Gaussian state space modelling code

FinMetrics does not have a function for calculating and maximising non-Gaussian state space log-likelihoods; instead, the S-Plus function nlminb (non-linear minimisation subject to box-constraints) must be used. Like the SsfFit function used for Gaussian state space models, the nlminb function is the key function upon which the model fitting process hinges. The difference between the SsfFit function and the nlminb function is that nlminb does not calculate the log-likelihood to be maximised, so the log-likelihood along with the other information such as the data, initial parameter estimates and the state space form must be passed to it for it to be able to proceed with optimisation.

The non-Gaussian log-likelihood, (6.14), is somewhat complicated to calculate; it has two parts: the linear Gaussian approximating model likelihood, $L_g(\boldsymbol{\psi})$, and the weight, $\bar{w}$. I wrote a non-Gaussian log-likelihood function, SsfExfLoglike, which comprised of two routines: SsfExfLgam and SsfExfLoglikeCalc. The SsfExfLgam routine calculated the linear Gaussian approximating model quantities $\tilde{y}_t^*$ and $\tilde{H}_t$, and the SsfExfLoglikeCalc function calculated the linear Gaussian approximating model log-likelihood, $\log\{L_g(\boldsymbol{\psi})\}$, and added this to $\log(\bar{w})$ to get the overall log-likelihood. These routines in turn required other routines which are outlined below.

The SsfExfLgam routine was a procedure which updated the values of $\tilde{y}_t^*$ and $\tilde{H}_t$ until they did not change with subsequent iterations. In the SsfExfLgam routine, initial guesses were made of the values of $\tilde{y}_t^*$ and $\tilde{H}_t$ and these were put into vectors of the form $\tilde{\boldsymbol{y}}^* = (\tilde{y}_1^* \dots \tilde{y}_n^*)'$ and $\tilde{\boldsymbol{H}} = (\tilde{H}_1 \dots \tilde{H}_n)'$; these were matrices in the case of the bivariate models. The vectors were then fed into the S-Plus function SsfCondDens, which performd Kalman filtering and smoothing, and produced a fitted mean, $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1 \dots \tilde{\theta}_n)'$. The signal vector was then fed into another routine I wrote, SsfLgamExpansion, which updated the values of $\tilde{\boldsymbol{y}}^*$ and $\tilde{\boldsymbol{H}}$ according to the linear Gaussian approximating model updating equations (6.22) for various different exponential family distributions. This process was then repeated for as many times as it took until the values of $\tilde{\boldsymbol{y}}^*$ and $\tilde{\boldsymbol{H}}$ did not change with subsequent iterations.

Once $\tilde{\boldsymbol{y}}^*$ and $\tilde{\boldsymbol{H}}$ were fixed, the SsfExfLoglike function passed them to the SsfExfLoglikeCalc routine. Here, $\tilde{\boldsymbol{H}}$ was put into the state space form routine, SsfExfSsf, which was now a complete Gaussian state space form with state error variances from the initial parameter vector, $\tilde{\boldsymbol{\psi}}$, and time varying model error variances from $\tilde{\boldsymbol{H}}$. The linear Gaussian approximating model state space form, along with the linear Gaussian approximating model observations, $\tilde{\boldsymbol{y}}^*$, were then fed into the FinMetrics function, SsfLoglike, which calculates Gaussian state space log-likelihoods given data and a state space form. Hence the linear Gaussian approximating model log-likelihood, $\log\{L_g(\boldsymbol{\psi})\}$, was obtained.

The next stage was to calculate the weight, $\bar{w}$, so that the complete non-Gaussian log-likelihood could be obtained. The first step here was to draw a sample of simulated $\boldsymbol{\theta}$'s and $\boldsymbol{\epsilon}$'s to be used in (6.11) to calculate $\bar{w}$. As noted in the introduction, the simulation smoother provided by FinMetrics, SimSmoDraw, was faulty such that it could not calculate simulated $\boldsymbol{\alpha}$'s or $\boldsymbol{\epsilon}$'s properly. The fault I found in the function was that in step 4 of the algorithm for calculating simulated errors, the SimSmoDraw function was calculating $\breve{\boldsymbol{\varepsilon}} = E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*) - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+)$ rather than $\breve{\boldsymbol{\varepsilon}} = E(\boldsymbol{\varepsilon}^*|\boldsymbol{y}^*) - E(\boldsymbol{\varepsilon}^+|\boldsymbol{y}^+) + \boldsymbol{\varepsilon}^+$ as it should have been. Also, incidentally, it did not recognise that the square root of zero is zero and failed whenever it tried this operation. So I created my own amended version of the function, SimSmoDrawJames, and I used it to calculate the sample of simulated $\boldsymbol{\epsilon}$'s and obtained the simulated $\boldsymbol{\theta}$'s by taking $\breve{\boldsymbol{\theta}} = \boldsymbol{y}^* - \breve{\boldsymbol{\epsilon}}$.

Once the sample of simulated $\boldsymbol{\epsilon}$'s and $\boldsymbol{\theta}$'s had been obtained, they were fed into another of my routines, the SsfLogDensity routine, which used equation (6.11) to calculate w_i values, where $w_i = w(\boldsymbol{y}|\breve{\boldsymbol{\theta}}^{(i)})$. Then calculating $1/N \sum w_i$ from the N simulated $\boldsymbol{\theta}$'s gave $\bar{w}$ as required.

As well as the data, the initial parameter estimates, the state space form and the log-likelihood function, the nlminb function must also be supplied a function to calculate the hessian matrix for the parameter estimates. In the absence of a hessian function being supplied, the nlminb function is supposed to calculate a numerical hessian; unfortunately, as I found, it does not actually do this. However, I eventually found that simply inputting just a single number into the field where a hessian function should be supplied, seemed to induce a numerical hessian function to be displayed in the S-Plus report window. So I copied and adapted this function to work with my routines. If any reader of this thesis has difficulty trying to induce nlminb to produce a numerical hessian then I recommend this course of action.

The graphics routine I wrote, SsfExfGraphics, for calculating the fitted mean and fitted states, used many of the same functions as for the calculation of the log-likelihood. In (6.9) the calculation of $\hat{E}\{x(\boldsymbol{\alpha})|\boldsymbol{y}\}$ relies on simulation in exactly the same way as above to obtain the w_i's. Also, depending on the function $x(\boldsymbol{\alpha})$, accurately simulated $\boldsymbol{\alpha}$ vectors are needed so that the constant states, which are the coefficients of the explanatory variables, may be obtained; this was another good reason for amending the faults of the SimSmoDraw function.

Appendix B

Derivation of the conditional mean and variance

The mean and variance of the conditional Poisson density (7.2) are derived below. Firstly, for convenience of notation, we shall drop the subscript t from y_t, π_t and λ_t. Now the mean $E(Y)$ may be calculated as follows, where the range of y is $y = 0, 1, 2, ...$:

$$\begin{aligned} E(Y) &= \sum_{y=0}^{\infty} y f(y|\pi, \lambda) \\ &= \sum_{y=0}^{\infty} y\{1-\pi\}^{(1-I(y))} \left\{\pi \frac{e^{-\lambda}\lambda^{y}}{y!(1-e^{-\lambda})}\right\}^{I(y)}. \end{aligned}$$

Now, for $y = 0$, we get 0, and for $y = 1, 2, ...$, we have $\{1-\pi\}^{(1-I(y))} = 1$ and $I(y) = 1$, so we may therefore simplify the equation to

$$E(Y) = \sum_{y=1}^{\infty} y\pi \frac{e^{-\lambda}\lambda^{y}}{y!(1-e^{-\lambda})}.$$

We may now treat the equation in much the same way as we would to derive the mean of a Poisson distribution:

$$\begin{aligned}
E(Y) &= \frac{\pi e^{-\lambda}}{1-e^{-\lambda}}\sum_{y=1}^{\infty} y\frac{\lambda^y}{y!} \\
&= \frac{\pi e^{-\lambda}}{1-e^{-\lambda}}\sum_{y=1}^{\infty} \frac{\lambda\lambda^{y-1}}{(y-1)!} \\
&= \frac{\pi\lambda e^{-\lambda}}{1-e^{-\lambda}}\sum_{y=0}^{\infty} \frac{\lambda^y}{y!}.
\end{aligned}$$

Now, since $e^{\lambda} = \sum_{y=0}^{\infty}\frac{\lambda^y}{y!}$, we therefore have

$$E(Y) = \frac{\pi\lambda}{1-e^{-\lambda}}.$$

We start calculation of the variance by finding $E(Y^2)$ so that we may then derive $Var(Y) = E(Y^2) - \{E(Y)\}^2$.

$$\begin{aligned}
E(Y^2) &= \sum_{y=0}^{\infty} y^2 f(y|\pi,\lambda) \\
&= \sum_{y=0}^{\infty} y^2\{1-\pi\}^{(1-I(y))}\left\{\pi\frac{e^{-\lambda}\lambda^y}{y!(1-e^{-\lambda})}\right\}^{I(y)} \\
&= \sum_{y=1}^{\infty} y^2\pi\frac{e^{-\lambda}\lambda^y}{y!(1-e^{-\lambda})} \\
&= \frac{\pi e^{-\lambda}}{1-e^{-\lambda}}\sum_{y=1}^{\infty} y^2\frac{\lambda^y}{y!} \\
&= \frac{\pi e^{-\lambda}}{1-e^{-\lambda}}\sum_{y=1}^{\infty} y\frac{\lambda\lambda^{y-1}}{(y-1)!} \\
&= \frac{\pi\lambda e^{-\lambda}}{1-e^{-\lambda}}\sum_{y=0}^{\infty} (y+1)\frac{\lambda^y}{y!} \\
&= \frac{\pi\lambda e^{-\lambda}}{1-e^{-\lambda}}\left(\sum_{y=0}^{\infty} y\frac{\lambda^y}{y!} + \sum_{y=0}^{\infty}\frac{\lambda^y}{y!}\right).
\end{aligned}$$

Now, when $y = 0$ we have that $y\frac{\lambda^y}{y!} = 0$, so we can start the summation from $y = 1$ without affecting the result.

$$\begin{aligned}
E(Y^2) &= \frac{\pi\lambda e^{-\lambda}}{1-e^{-\lambda}}\left(\sum_{y=1}^{\infty} y\frac{\lambda^y}{y!} + e^{\lambda}\right) \\
&= \frac{\pi\lambda e^{-\lambda}}{1-e^{-\lambda}}\left(\sum_{y=1}^{\infty} \frac{\lambda\lambda^{y-1}}{(y-1)!} + e^{\lambda}\right) \\
&= \frac{\pi\lambda e^{-\lambda}}{1-e^{-\lambda}}\left(\lambda\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} + e^{\lambda}\right) \\
&= \frac{\pi\lambda e^{-\lambda}}{1-e^{-\lambda}}\left(\lambda e^{\lambda} + e^{\lambda}\right) \\
&= \frac{\pi\lambda(\lambda+1)}{1-e^{-\lambda}}.
\end{aligned}$$

The variance is therefore given by

$$\begin{aligned}
Var(Y) &= E(Y^2) - \{E(Y)\}^2 \\
&= \frac{\pi\lambda(\lambda+1)}{1-e^{-\lambda}} - \left(\frac{\pi\lambda}{1-e^{-\lambda}}\right)^2 \\
&= \left(\frac{\pi\lambda}{1-e^{-\lambda}}\right)\left(1+\lambda-\frac{\pi\lambda}{1-e^{-\lambda}}\right).
\end{aligned}$$

The mean and variance of the truncated Poisson model may be found by similar means to the above method.

Bibliography

Box, G. E. P. and Jenkins, G. M. (1970) *Time series analysis, forecasting and control.* Holden-Day.

Broughton, J. (1990) Trends in drink/driving revealed by recent road accident data. *Tech. Rep. 266*, TRL.

Broughton, J. (2000a) The numerical context for setting national casualty reduction targets. *Tech. Rep. 382*, TRL.

Broughton, J. (2000b) Survival times following road accidents. *Tech. Rep. 467*, TRL.

Broughton, J., Hazelton, M. and Stone, M. (1999) Influence of light on the incidence of road casualties and the predicted effect of changing 'summertime'. *Journal of the Royal Statistical Society: Series A*, **162**, 137–175.

Brown, B. M. (1961) *The mathematical theory of linear systems.* Hazell Watson and Viney Ltd.

Chatfield, C. (1975) *The analysis of time series, theory and practice.* Chapman and Hall.

Diggle, P. J. (1990) *Time series - a biostatistical introduction.* Oxford Science Publications.

Dobbie, M. J. (2001) *Modelling correlated zero-inflated count data.* Ph.D. thesis, Australian National University.

DTLR (2000) Road Accidents Great Britain: 2000, The Casualty Report. *Tech. rep.*, DTLR.

Durbin, J. and Koopman, S. J. (1997) Monte carlo maximum likelihood estimation for non-gaussian state space models. *Biometrika*, **84**, 669–684.

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

Durbin, J. and Koopman, S. J. (2002) A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, **89**, 603–615.

EU Directorate general for transport (1999) Cost 329, Models for traffic and safety development and interventions. *Tech. rep.*, DTLR.

Fridstrom, L., Ifver, J., Ingerbrigtsen, S., Kulmala, R. and Thomsen, L. K. (1995) Measuring the contribution of randomness, exposure, weather and daylight to the variation in road accident counts. *Accident Analysis and Prevention*, **27**, 1–20.

Fridstrom, L. and Ingerbrigtsen, S. (1991) An aggregate accident model based on pooled, regional time-series data. *Accident Analysis and Prevention*, **23**, 363–378.

Hamilton, J. D. (1994) *Time series analysis.* Princeton University Press.

Harvey, A. C. (1981) *Time series models.* Philip Allan.

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

Harvey, A. C. and Durbin, J. (1986) The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society: Series A*, **149**, 187–227.

Kalman, R. E. (1960) A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, Transactions ASME, Series D*, **82**, 35–45.

McCullagh, P. and Nelder, J. A. (1989) *Generalized linear models (second edition).* Chapman and Hall.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996) *Applied linear statistical models.* WCB/McGraw-Hill.

Nicholls, D. F. (1979) The exact likelihood function of multivariate autoregressive moving average models. *The Australian Journal of Statistics*, **21**, 93–120.

O'Neill, T. J. and Barry, S. C. (1995) Truncated logistic regression. *Biometrics*, **51**, 533–541.

Rodiguez, V. B. (2003) *State-space models for longitudinal data.* Master's thesis, Australian National University.

Sandmann, G. and Koopman, S. J. (1998) Estimation of stochastic volatility models via monte carlo maximum likelihood. *Journal of Econometrics*, **87**, 271–301.

Sullivan, J. M. and Flannagan, M. J. (2002) The role of ambient light level in fatal crashes: inferences from daylight saving time transitions. *Accident Analysis and Prevention*, **34**, 487–498.

Venables, W. N. and Ripley, B. D. (1994) *Modern applied statistics with S-Plus.* Springer-Verlag.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F. and Lindenmayer, D. B. (1996) Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, **88**, 297–308.

Zivot, E., Wang, J. and Koopman, S. J. (2003) State space modelling in macroeconomics and finance using SsfPack for S+FinMetrics. Preprint working paper.