Statistical Modelling of Road Accident Data via Graphical Models and Hierarchical Bayesian Models

A thesis submitted to Middlesex University

in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

RADU TUNARU

Business School

Middlesex University

October 1999

To my wife

"Some mathematician, I believe, has said that true pleasure lies not in the discovery of truth, but in the research for it."

Tolstoy, *Anna Karenina*

ACKNOWLEDGEMENTS

My first thanks go to the Director of Studies and my first supervisor, Mr. David Jarrett, for his careful supervision and many helpful discussions throughout the research, for his thoughtful reading of the thesis, and for drawing my attention to a new area of research based on Markov Chain Monte Carlo methods.

Thanks are gratefully extended to my second supervisor, Professor C.C. Wright, for his clear and concise comments about the research and the thesis, and for his continuous encouragement.

I would also like to acknowledge Middlesex University for providing a three-year studentship to support the research in this thesis.

I also thank Ken Lupton, the research manager of Transport Management Research Centre at Middlesex University, for helping me to prepare the data.

The Library of Middlesex University at Hendon supported me in getting many articles and books needed for the research. The calculations in the research are made on a Pentium computer provided by Middlesex University. The thesis is typed using LaTex2ε.

I take full responsibility for any errors or omissions in this thesis.

ABSTRACT

The objective of this thesis is to develop statistical models for multivariate road accident data. Two directions of research are followed: graphical modelling for contingency tables cross-classified by accident characteristics, and hierarchical Bayesian models for multiple accident frequencies of different types modelled jointly.

Multi-dimensional tables are analysed and it is shown how to use collapsibility to reduce the dimensionality of the analysis without the problems of Simpson's paradox. It is revealed that accident severity and the number of casualties are associated, and that these variables are mainly influenced by the number of vehicles and speed limit. Graphical chain models allow causal hypotheses to be formulated and it is shown how they are valuable tools for empirical research about road accident characteristics.

The hierarchical Bayesian models developed combine generalized linear models with random effects. The novelty of these models consists in the joint modelling of multiple response variables. The models account for overdispersion and they are used for accident prediction and for ranking hazardous sites. All models are fully Bayesian and are fitted using Markov Chain Monte Carlo methods. It is shown that multiple response variables models are superior to separate univariate response models.

Some theoretical problems are examined regarding the maximum likelihood estimation process for the two parameters negative binomial distribution. A condition is given that is equivalent with unique maximum likelihood estima-

tors.

The two directions of research are connected by using graphs to describe the models. In addition, a new Bayesian model selection procedure for contingency tables is proposed. This is based on Gibbs sampling and avoids problems associated with asymptotic tests.

The conclusions revealed here can help practitioners to design better safety policies and to spend money more wisely on sites that really are dangerous.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The cost to society of road accidents is very high. According to The Institution of Civil Engineers it was estimated in 1996 as being between £14 billion and £19 billion per annum in the UK, although it is unmeasurable in terms of human lives (Carruthers, Bulpitt, Gray, Holmes, MacKinven, Moore, Quinn, Zealley and Huxford, 1996). Since road accidents are random events, their occurrence cannot be predicted. Various factors are thought to contribute to the realisation of road accidents. Valuable information can be extracted from large and complex data sets with the help of statistical methods. Although the exact number of future accidents cannot be calculated, it is possible to predict or estimate this number and to identify some important contributing factors that can be measured and influenced if necessary. What makes all these possible is statistical modelling.

After the second world war the number of accidents increased dramatically but so did the number of vehicles. Governments all over the world were facing a serious problem that needed major attention. Statistical methods were soon starting to be applied in this area of research too. However, the major turning point in the advance of scientific methodologies for analysing road accidents has been the development of the theory of *generalized linear models* (McCullagh and Nelder, 1989). This new class of models is flexible enough to allow modelling of the accident frequencies with a Poisson error. There are statistical methods for measuring the safety effect of engineering treatment and for taking into account the regression-to-mean effect (Hauer, 1980; Hauer, Ng and Lovell, 1989; Hauer, 1997; Wright, Abbess and Jarrett, 1988), and for relating the number of accidents at a site to road network characteristics (Maycock and Hall, 1984; Maher and Summersgill, 1996; Mountain, Fawaz and Jarrett, 1996; Amis, 1996). Comparatively little statistical work has been done on the relationships between accident characteristics such as severity, number of vehicles, pedestrian involvement, time of day and so on. The aim of this research is to contribute to the statistical modelling of large and complex road accident data using and developing appropriate multivariate techniques.

1.1.1 Possible forms of analysis

The statistical investigation of road accident data is a non-randomized study, a kind of observational study in which there is no direct control by the investigator. The analyst just observes what is happening, making it very difficult

to establish causal relationships. The nature of this type of data makes impossible any controlled randomization that would help in designing the study. This is true for data collected for accident characteristics and summarised in contingency tables and it is also true for data collected for regression-like analyses. For the former case, the analyst takes into account the fact that the accidents already occurred so a retrospective view is appropriate. In the latter case, the situation is somehow reversed, the task of the analysis being to predict future numbers of accidents using a statistical model that fits the current set of data, again an observational study. A practitioner aims to understand why accidents occur on a road network and what can be done to reduce the number of accidents to a minimum. There are two ways of extracting valuable statistical information from road accident data and these perspectives divide the thesis into two parts.

First, various characteristics are recorded for all accidents which occur in a given period of time. At a national level this is done in UK each year in a database like STATS 19. Then the practitioners might attempt to understand the associations between these characteristics that will help them to design better safety policies. Primarily, they are interested in identifying the causes of accidents. However, they cannot analyse each accident individually so they rely on a statistical analysis to identify factors contributing to a large number of accidents. Then the local authorities design and implement the safety policies thought to manipulate the identified factors in such a way to reduce the future number of accidents. It has to be remarked that in statistics the

word "causal" is very often avoided in favour of a less powerful term, that is "association". Nevertheless, studies from other areas of research and some external information may help to identify causes and effects. Maycock (1985) studied 20 variables as road accident factors. Writing about future possible research he said :

> "Everyone knows that correlation is not the same thing as causation but the existence of correlations demand explanations and attempting to obtain explanations would lead into different sorts of behavioural studies, but studies which were targeted towards explanations of established accident facts.
>
> Moreover, establishing and following up statistical associations in this way could provide fairly direct clues to the design of remedial measures for those involved in safety legislation, education and training and the design and administration of driving test standards."

For the analysis of accident characteristics the observational units are the accidents themselves. The variables are the characteristics of the accidents together with other more general variables like road network characteristics, time specifications and so on. They are analysed in this thesis as categorical, any continuous variables being categorised, and data is summarised in contingency tables. This type of data is most of the time recorded by police and it is possible to have miscategorization of some observations due to human error. As highlighted above, for this type of data, one purpose is to find a

model which explains how the categorical variables are interrelated. For three variables A, B and C, if the model suggests that only the pairs A, B and B, C are related, this is formulated statistically as a conditional independence between A and C given the values of B. In common language, knowing the values of variable B may provide some information about possible values of C, and moreover, finding out *any* information about A would be irrelevant for discovering *more* information about C other than it is already known from B.

For the first kind of data, the approach proposed in this thesis is based on *graphical* modelling and its derivative, *graphical chain* modelling. With 6 or more road accident characteristics under study, the contingency table can be expected to be sparse. Due to the nature of the data it is a finite population in a fixed period of time. This particularity creates specific problems that are discussed in this thesis. On a real-world example, it is shown that relying on asymptotic inference gives different results than exact conditional inference and the latter should always be used in such instances.

The second type of data is analysed by dividing the road network into small units, called *sites*, and then trying to relate the observed number of accidents to site characteristics, either environmental or socio-economical or geometric. Depending on the results of the statistical analysis, treatment policies are implemented to reduce the number of accidents. The units of the analysis are the sites and the variables are both discrete (e.g. accident frequencies) and continuous (e.g. traffic flow).

This second direction of research aims at modelling the accident counts as

numerical random variables. The units of the statistical investigation are the sites of the road network. The models proposed in this thesis can be used for prediction of future numbers of accidents, for describing possible correlation structures between accident frequencies of different type and for ranking the sites according to different criteria. Practical applications described here show the usefulness of the joint modelling of multiple accident counts.

Analysing multivariate counts by statistical methods has been very difficult because of the lack of well-defined parametric distributions that can explain complex correlation structures. This problem is solved in this thesis using *hierarchical* Poisson multivariate models. The whole methodology used for generalized linear modelling (McCullagh and Nelder, 1989) is incorporated and models with random effects and regression structures are easily and naturally included. However, the complexity of such models makes analytical methods unfeasible. In the modelling process integrals of dimension of hundreds have to be calculated and even numerical methods are not helpful because they are not feasible for dimensions greater than 20. This major difficulty is overcome in this thesis using Markov Chain Monte Carlo (MCMC) methods, in particular Gibbs sampling.

The class of hierarchical Bayesian models proposed here is new to applied statistical modelling of road accident data because multiple responses are jointly modelled, the models are fully Bayesian in specification and they can be used to answer different questions based on the same statistical MCMC output. Although hierarchical Bayesian models have been developed for re-

peated measurements data in other areas of research, the hierarchical models developed in this thesis are tailored for road accident data. The multiple responses studied in this thesis represent counts of different type of accidents, so the possible correlation structure of the responses is not caused by studying the same model over time, like in longitudinal studies. The novel multiplicative equations describing the models can be used by practitioners to predict changes in accident type as well as frequency if treatment policies are implemented.

It is somehow regretable that the term "hierarchical" has different meanings in the two parts of the thesis. In connection with a log-linear model for contingency tables, hierarchical means an imposed rule of model specification, very important for the interpretability of the models. Regarding a predictive accident model, hierarchical is again about model specification but in a totally different manner. The observed data is combined with a prior distribution for the model parameters; the prior also depends on some unknown parameters which follow a hyper-prior and the specification may continue like that on several stages. The hierarchy is ended at some stage where all the parameters are known.

1.1.2 Graphical representation

The two directions of research are related by the basic method of representing hierarchies, which is a **graph**. In the discussion of the articles given by Wermuth and Lauritzen (1990) and Edwards (1990), A.P. Dawid strongly sup-

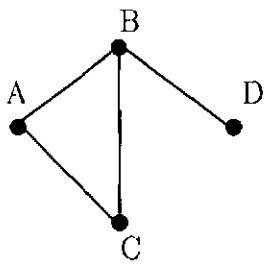


Figure 1.1: Graphical association model

ported the use of graphs for communicating statistical modelling ideas. In this thesis, two types of graphical models, therefore of graphs, are used. The first type, like the one illustrated in Figure 1.1, has vertices associated with observed categorical variables representing accident characteristics. The graph synthesizes the conditional independencies revealed by the graphical model fitting the data. Similar graphs with a mixture of undirected and directed edges will be encountered in the first part of this thesis. Regardless of the nature of the edges, these graphs are built using observed variables.

The second type of graphs are used in this thesis again for model specification, more exactly for expressing conditional independencies. There are only directed edges due to the hierarchical structure of the models. The difference relative to the first type consists in having vertices for observed and unobserved quantities. A simple example is given in Figure 1.2. The program WinBUGS uses such a graphical model for simulation.

In addition, there are some other links between the two main parts of the thesis. The analysis of the characteristics of accidents in Bedfordshire

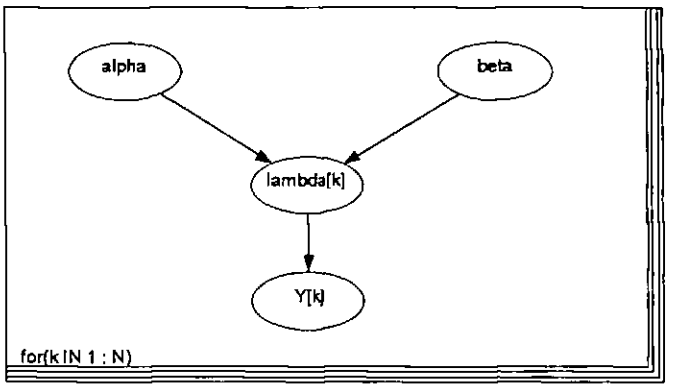


Figure 1.2: Directed graphical model for Bayesian model specification in WinBUGS

and Hampshire data sets reveals that the accident severity and the number of vehicles involved in the accident are directly related. This suggests that developing separate regression models for these two variables may give unreliable results. The research carried out in the second part of the thesis confirms this hypothesis and provides a feasible methodological solution. Regarding the model selection procedures for (hierarchical) graphical models, a new method is proposed in a Bayesian framework, employing similar Markov Chain Monte Carlo ideas as those used for the multiple response variables models. This method provides another link between the two parts of the thesis.

1.1.3 Data sets used

Two separate sources of data were used in this thesis. The first was the STATS 19 database for 1995, obtained from UK ESRC Data Archive by the Trans-

port Management Research Centre at Middlesex University. The two subsets of data extracted from STATS 19 and the subset of variables analysed were the author's choice. Some of the variables, like accident severity, were used as recorded in the database but others were recategorised to have a small number of levels. For example, the number of vehicles involved and the number of casualties were considered with only three levels (one, two, three or more), road surface conditions with only three (dry, wet-damp, snow-ice-frost-flood). Other temporal variables were also categorised as it will be seen in later chapters.

The set of data analysed in the second part of the thesis contains the accident frequencies on 156 single-carriageway link sites between 1984 and 1991 in Kent. The data had been provided by Kent County Council to Middlesex University's Transport Management Research Centre for a previous research project (Mountain, Jarrett and Fawaz, 1995; Mountain, Jarrett and Wright, 1994). The accident counts are known at a disaggregated level; four separate categorises were investigated. The disaggregation was made by the author linking the original set of data with the STATS 19 database. Covariate information, such as estimated traffic flow, speed limit and link length, was also available and used in the modelling process. Speed limit was considered as a binary variable having only two levels: urban meaning 40 mph or less and rural meaning 50 mph or 60 mph.

It is well known that not all road accidents are recorded in STATS 19 database (Department of Transport, 1996). The number of unreported accidents

is not known and the analysts try to make the best of what is available. In this thesis the sets of data are used without trying to account for missing records.

1.2 Aims of the thesis

The overall aim of this thesis is to contribute to the development of sound statistical techniques that can be applied to road accident data. The intention is to develop statistical methods which improve the extraction of relevant information contained in the data, information that can be used subsequently by various organisations and traffic engineers to design safety measures. If the wrong sites are selected for treatment due to bad ranking methods, or policy measures are designed to improve irrelevant (from the safety point of view) characteristics of road accidents, the loss is very high in terms of money and human life.

Graphical models and graphical chain models are described as an exploratory multivariate technique that can be applied to large sets of road accident data. It is intented to find out which variables, "environmental", "road user", and so on, are associated with variables representing very important accident characteristics, such as accident severity, the number of vehicles involved and the number of casualties.

More specifically, the first part of the thesis has the following objectives

1. To investigate the associations and conditional independencies between several road accident characteristics for two fairly large datasets, corre-

sponding to the counties of Bedfordshire and Hampshire, separately and pooled together.

2. To investigate methods of reducing the analysis of large contingency tables to the analysis of a smaller dimensional subtables defined by subsets of variables of particular interest.

3. To investigate various model selection procedures that can be used in practice for selecting a graphical model; to discuss their advantages and limitations.

4. To investigate the application of graphical chain models when substantive research hypotheses are formulated prior to the statistical modelling process and to identify posible causal implications of such hypotheses.

The research carried out in the first part of the thesis will use only categorical variables, but continuous variables such as traffic flow are also important in the study of road accidents. The problem is that the theory of graphical models is less well developed for a mixture of discrete and continuous variables. Partly for this reason, the research continues in the second part of the thesis by separating out the individual accidents according to location, in order to relate the accidents to the road network.

In the second part of the thesis the author's aim is to propose a new class of models for different type of accidents jointly modelled. Models including covariate information as well as models based only on parametric specification are developed. It is shown how computational problems in developing such

complex models can be solved using MCMC. It is important to relate the observed number of accidents to environmental characteristics, such as speed limit, link length and estimated traffic flow and this aim will play a major role in this thesis in developing the hierarchical models for multiple accident frequencies. The objectives in the second part of the thesis are therefore

1. To develop hierarchical Bayesian models for multiple accident counts.

2. To discuss the problem of ranking the sites according to different criteria and considering multiple response variables.

3. To discuss estimation problems for compound Poisson distributions.

This research will benefit authorities in designing new measures for traffic safety control and new methods for collecting data. At the same time it will provide some clues and starting points for future studies. The hierarchical Bayesian models will provide a new and deeper statistical modelling methodology for road accident data.

1.3 Overview of the thesis

This introduction is followed by a statistical literature review, Chapter 2, where some of the statistical problems related to the ideas developed in the thesis are defined and the solutions known so far are illustrated. Although the applications, for which the statistical techniques are developed, concern road accidents, the same models can be adapted for other count data. The

originality of this thesis consists in taking a multivariate approach for statistical modelling, where "multivariate" means several responses modelled jointly. Nevertheless the univariate case is also important and is better known in the literature. The role of the Chapter 2 is to review the most up to date statistical modelling for the univariate case and to identify potential problems worth discussing in the multivariate setting.

Chapter 3 is concerned with graphical modelling. It provides a motivation for applying graphical modelling to road accident data, describes the graph theory concepts used in the thesis, together with a short account of conditional independence, and gives a detailed description of various Markov properties necessary to develop graphical models and graphical chain models. The theory is almost everywhere accompanied by examples using road accident data.

The inference process is described in Chapter 4. The starting point of discussion is the class of log-linear models, a particular case of generalized linear models. When the researcher is interested in identifying conditional independence relationships between the variables (or between groups of variables) under study, graphical models are proposed as one of the best solutions. The theoretical framework and the most important results are described. Moreover, since it is known that any log-linear model can be nested into a graphical model, it seems to be always useful to find out a graphical model fitting the data well and simply enough to assist interpretation. Various model selection procedures for log-linear models and graphical models are reviewed and exemplified. The theoretical aspects of graphical chain models are also developed.

The data subsequently analysed in Chapter 5 are subsets of data extracted from the national road accident database for Great Britain, STATS 19. It is expected that the contingency table summarising such data will be sparse. This particular aspect makes the contingency tables more difficult to analyse. The classical tests based on asymptotic methods are not reliable so exact conditional tests, using Monte Carlo methods to overcome the computational difficulties, are described in the context of graphical models. Graphical models and graphical chain models for very large sets of data are proposed and important conditional independencies between road accident characteristics are identified. A comparison of asymptotic and exact conditional methods is investigated in relation to graphical chain modelling, for a large subset of data regarding accidents with pedestrian casualties in Bedfordshire in 1995.

Methods of reducing the dimensionality of the analysis are extremely useful. Collapsibility is a concept developed in the context of log-linear modelling that proves extremely helpful in reducing the amount of work necessary to extract reliable information from data. This is done in an applied manner in Chapter 6.

Probably the most theoretical chapter of this thesis is Chapter 7 where estimation problems for compound Poisson distributions are studied. Two major cases, the Poisson-gamma and Poisson-log normal distributions, are discussed in greater detail. This chapter has a special importance since many practitioners seem not to be aware of the difficulties presented by these two compound distributions and compound Poisson distributions in general. Chapter 7 con-

tinues the discussion started in Chapter 2 about empirical Bayesian modelling but goes beyond that and opens the door to more complex and realistic models.

Chapter 8 is dedicated to hierarchical Bayesian models for counts. Bayesian methods combining hierarchical models and regression techniques are developed to extract information from a set of road accident data. In the first section the general methodology is explained in the context of univariate models, thus making a straighforward connection with the second chapter of the thesis. MCMC methods are used to solve computational problems related to hierarchical models and are illustrated using two standard models. In the second part of Chapter 8 several complex hierarchical models are developed. At the same time, an attempt is made to model multiple response count models, based solely on the observed frequencies, using distributions such as the multivariate log-normal distribution, hierarchically specified.

A new Bayesian model selection procedure is proposed for log-linear models for contingency tables. The computational side of the new method is solved again by applying MCMC techniques and this is the main reason why this section is included in this chapter.

Given the applied character of this thesis, there is a companion Chapter 9 to Chapter 8 in which a complex set of accident data is investigated at a multiple response level. The set of data concerns accidents on 156 links in Kent between 1984 and 1991. The models analysed are fully Bayesian and range from simple log-linear regression models to mixed Poisson regression models with random effects. First, it is shown how to select a small subset of represen-

tative models (3 models are identified), and then, these models are examined in greater detail. The sites can be ranked according to different criteria using a single MCMC output, and the results are described and discussed towards the end of the chapter.

The last chapter summarises the conclusions of this thesis, from both theoretical and applied points of view. It also contains a section proposing further research that would follow quite naturally from the results of this thesis.

Chapter 2

Statistical modelling of road accident data

2.1 Introduction

The purpose of this chapter is to present the framework of the thesis in terms of the assumptions made and the problems that will be tackled, and also to review critically the contingent literature to these problems.

Road accidents are among the more visible consequences of an enormous number of failures in the daily volume of interaction between the people who use the road networks and the environment in which they travel. An accident that is predictable is a contradiction in terms. In other words, when we are talking about an individual accident, no matter how much knowledge we have about the possible generating mechanisms, we are unable to predict exactly where, when and to whom the next individual accident will occur. The best

that can be done is to predict their approximate number. This is simply because, although an individual accident is impossible to predict, the total number of accidents of some kind may behave with an almost constant overall frequency in the long run.

As defined in Hauer et al. (1989) and Hauer (1997) *safety* is the property of some specific entity, most commonly a site of the road network. The property of safety (or more exactly the non-safety) for a site is quantified as the number of accidents expected to occur per unit of time and their adverse consequences. The important term is "expected" which makes a straightforward connection with the statistical approach. If all conditions that affect safety (traffic, weather, and so on) are frozen, expected means the "average" in the long run.

One aim of collecting and investigating road accident data is to identify significant clusters of accidents having common causal factors and to asses the expected numbers of road accidents. The list of problems includes the evaluation of safety treatments, the ranking and identification of hazardous locations, predicting the numbers of futures accidents and investigating the associations between characteristics of road accidents. The statistical models proposed for solving these problems can be divided into three categories: models for accident frequencies, models for type of accidents and models for both accident frequencies and type of accidents. The first category has been well investigated at univariate level and it is reviewed next.

2.2 Models for accident frequencies

The following methodological framework is followed for studying accidents counts on a road network over a fixed period of time. The network is first divided into units, usually called *sites*, like junctions or stretches of the road. The statistical unit is the road network element and the response variables are accident counts.

2.2.1 The pure Poisson Model

The main probability distribution used in modelling accident data is the Poisson distribution. Accidents occur in time. Consider a fixed site for which accidents are recorded in a fixed period of time T. Partitioning the time period into n intervals of duration T/n, let $Y_{n,i}$ be the number of accidents recorded in the i-th time interval, let $P_{n,i} = \Pr(Y_{n,i} = 1)$ and let $\varepsilon_{n,i} = \Pr(Y_{n,i} \geq 2)$. The following assumptions are made

1. The random variables $Y_{n,i}, (i = 1, 2, \ldots, n)$ are independent over i
2. $\sum_{i=1}^{i=n} P_{n,i} \to \lambda \in (0, \infty)$ as $n \to \infty$,
3. $\max_{1 \leq i \leq n} P_{n,i} \to 0$ as $n \to \infty$,
4. $\sum_{i=1}^{i=n} \varepsilon_{n,i} \to 0$ as $n \to \infty$.

Then it is shown in Durrett (1991, Theorem 6.1) that

$$Y_{n,1} + Y_{n,2} + \ldots + Y_{n,n} \xrightarrow{d} \text{Pois}(\lambda)$$

where d means that the convergence is in distribution. This justifies using the Poisson distribution for modelling road accidents. This derivation is conceptually different from the one based on a homogeneous Poisson process and the Poisson distribution that characterizes it. The assumption of a homogeneous Poisson process is not valid for road accidents since it is natural to expect great variation of accidents by time patterns.

The Poisson distribution is defined mathematically and whether a series of events is in agreement with it is an empirical fact. Denote by Y_k the number of accidents at site k during an observed time period T_k. The first assumption made in modelling accident frequencies (Nicholson, 1985) is that

$$Y_k \mid m_k \overset{ind}{\sim} \text{Pois}(m_k \equiv \lambda_k T_k)$$

where $k = 1, 2, \ldots, N$ and λ_k is the mean accident frequency per unit time at site k. The expected number of accidents, m_k, can then be linked with a covariate vector $X_k = (X_{k1}, X_{k2}, \ldots, X_{kq})'$, representing for instance traffic flows and the geometric characteristics of the site. The connection is made via a multiplicative equation which can be transformed into a linear equation on the logarithmic scale. The unknown coefficients are estimated by fitting the model to data and these will be used for statistical inference. The fitting process, under this generalized linear statistical modelling framework, can be done in GLIM or GENSTAT, where maximum likelihood estimates are obtained using an iterative weighted least squared (WLS) procedure. The

most common goodness-of-fit measures used are

$$G^2 = \sum_{k=1}^{k=N} 2\left[y_k \log\left(\frac{y_k}{\widehat{m}_k}\right) - (y_k - \widehat{m}_k)\right] \tag{2.1}$$

$$X^2 = \sum_{k=1}^{k=N} \frac{(y_k - \widehat{m}_k)^2}{\widehat{m}_k}. \tag{2.2}$$

where $y_k (k = 1, 2, \ldots, N)$ are the observed number of accidents and $\widehat{m}_k$ are the estimated means under the fitted model. The above notation for the Poisson model will be used without any index accounting for different sites when the theoretical model in itself is the same for each site and the model is self-explanatory.

Regarding the accident frequencies observed on a fixed number of sites, there are two broad types of statistical investigations:

1. before-after studies; and

2. regression models regarding the prediction of future number of accidents.

2.2.2 Before-after studies

A safety treatment of a site of a road network aims to reduce the number of accidents at that site. The usual way of assessing the effectiveness of a safety treatment is to compare the accident frequency before the treatment has been implemented with that after treatment.

A reduction in accidents at the treated sites does not necessarily imply that the treatment has been successful. Three reasons may be responsible for this.

- The number of accidents at a site may change in a random manner, increasing or decreasing, whether or not there has been any change at the site. Statistical methods are necessary to consider this random variation.

- The mean number of accidents may decrease without any connection with the treatment. In order to study these systematic factors it is important to compare treated sites with a control group of untreated sites. The confounding effects, such as time, can be overcome by selecting a control group of sites and observe the number of accidents at these sites over the same period as the treated sites. This design is called the before-after study and it uses a 2×2 contingency table

	Control	Treatment
Before	n_{11}	n_{12}
After	n_{21}	n_{22}

  defined by the time dichotomy, before-after, and the control-treatment dichotomy.

- The third problem, is the *regression-to-mean* effect, which means that for the many sites with a "low" accident frequency before treatment there will be a slight rise after treatment, for the few sites with a "high" frequency a greater fall; while for all sites together, no change, (Hauer, 1980).

The first two problems can be solved by standard methods (Hauer, 1986; Hauer, 1980; Hauer, 1997). In terms of improvement due to the statistical

analysis, the third problem is viewed as one of the most important. The regression-to-mean bias inadvertently results from the fact that only locations with a large number of accidents are generally selected for treatment, which may lead to biased conclusions. The standard solution to this problem is to use *empirical Bayes* (EB) models as developed in Abbess, Jarrett and Wright (1981), Jarrett, Abbess and Wright (1982), Brude and Larsson (1988), Morris, Christiansen and Pendleton (1991). Hauer (1997) is a general reference explaining empirical Bayes methods for practitioners.

The empirical Bayes (EB) method for estimation provides a general framework where different distributions can be studied in order to improve the quality of the estimators. The compound model

$$\begin{aligned} Y_k \mid m_k &\overset{ind}{\sim} \text{Pois}(m_k) \\ m_k &\overset{iid}{\sim} G(\cdot) \quad k \in \{1, 2, \ldots, N\} \end{aligned} \tag{2.3}$$

lead to estimates of the individual parameters m_k using information from all sites under study. In studies using EB methods the variation of m_k from site to site is regarded as purely random. Then the Y_k are marginally independent. If the unknown distribution $G(\cdot)$ has probability density g then the marginal density is

$$p_G(y_k) = \int \text{Pois}(y_k|m_k) g(m_k) dm_k$$

and the posterior density of m_k is

$$p(m_k|y_1, y_2, \ldots, y_N) = \frac{\text{Pois}(y_k|m_k)g(m_k)}{p_G(y_k)} \tag{2.4}$$

If g is known, meaning that its parameters are given and do not have to be estimated, then the model is called fully Bayesian; if the parameters of g have to be estimated from data then this approach is called an empirical Bayes (EB) method.

One of the first important empirical Bayes ideas for modelling counts was advocated by Robbins (1955) in a nonparametric form. For the compound Poisson-G model described in (2.3), suppose that G is totally unknown. Under squared error loss (SEL), the Bayes estimator is the posterior mean

$$m^B = \text{E}(m|y) \tag{2.5}$$

$$= \frac{(y+1)p_G(y+1)}{p_G(y)}. \tag{2.6}$$

The MLE of m is Y so m^B is biased. However, m^B is preferred because of lower MSE. When G is known, the estimation is straightforward. For the case when G is unknown Robbins (1955) suggested to estimate $p_G(y)$ by the number of values Y in the sample $Y_1, Y_2, \ldots, Y_N$ that are equal with Y, so

$$m^B = (y+1)\frac{\sum_{j=1}^{j=N} I_{\{y_j=y+1\}}}{\sum_{j=1}^{j=N} I_{\{y_j=y\}}}$$

where $I_{\{\cdot\}}$ is the indicator function. Therefore, the Bayes estimate m^B takes

information from other sites as well. Although this procedure has some good asymptotic properties, it was shown that, even when the sample size is large, this method does not perform very well and a parametric approach is more suitable (Carlin and Louis, 1996).

The prior distribution $g(m)$ is usually assumed to be of gamma form, because the gamma distribution is the conjugate distribution for the Poisson distribution (George, Makov and Smith, 1993). Thus

$$m \sim \text{gamma}(a, b) \equiv \text{gamma}\left[\frac{a}{b}; \frac{a}{b^2}\right]. \tag{2.7}$$

where $\text{gamma}(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}$ and the second parameterisation is in terms of the mean $\frac{a}{b}$ and variance $\frac{a}{b^2}$. Then it follows from the Bayes formula in equation (2.4) that the posterior distribution of m is

$$p(m \mid y) = \frac{(b+1)^{a+y}}{\Gamma(a+y)} m^{a+y-1} e^{-(b+1)m}. \tag{2.8}$$

The marginal distribution of Y is then

$$p(y) = \frac{\Gamma(a+y)}{y!\Gamma(a)} \left(\frac{1}{1+b}\right)^y \left(\frac{b}{1+b}\right)^a, \tag{2.9}$$

which is a negative binomial distribution $\text{NB}(\frac{b}{1+b}, a)$. As described by Morris in discussion of Hauer et al. (1989), the whole parametric modelling methodology for accident counts can be expressed in terms of a *descriptive model* and an *inferential model*. Both describe the distribution for the *observed data* and the

distribution of the *unobserved parameters*. The descriptive model is given by

- observed data

$$Y \sim \text{Pois}(m) \tag{2.10}$$

- unobserved parameters

$$m \mid a, b \sim \text{gamma}(a, b) \equiv \text{gamma}\left[\frac{a}{b}; \frac{a}{b^2}\right] \tag{2.11}$$

The inferential model is then

- observed data

$$Y \sim \text{NB}\left(p \equiv \frac{b}{1+b}, a\right) \tag{2.12}$$

- unobserved parameters

$$m \mid y \sim \text{gamma}(a+y, b+1) \equiv \text{gamma}\left[\frac{a+y}{b+1}; \frac{a+y}{(b+1)^2}\right]. \tag{2.13}$$

The Bayes estimate of m for the subpopulation of those sites at which y accidents occurred is

$$\text{E}(m \mid y) = \frac{a+y}{b+1}. \tag{2.14}$$

The regression effect can then be defined by $\text{E}(m \mid y) - y$. An alternative

definition is the expected percentage change in the number of accidents

$$\begin{aligned} R &= \left[\frac{\mathrm{E}(m \mid y) - y}{y}\right] \times 100 \\ &= \left[\frac{a+y}{b+1}\frac{1}{y} - 1\right] \times 100. \end{aligned}$$

In order to calculate the regression effect R the values of parameters a and b need to be estimated. The values of these parameters can be estimated by fitting the negative binomial distribution, equation (2.9), to the observed data. This can be done in GLIM using macros or more directly in GENSTAT. Some examples of such analyses are in Persaud (1991), Jarrett et al. (1982), Hauer (1997).

The Bayes estimate m^B is a convex combination of the overall expected accident frequency μ and the observed frequency y

$$m^B = \mathrm{E}(m \mid y) = \frac{a+y}{b+1} \tag{2.15}$$

$$m^B = \alpha\mu + (1-\alpha)y \tag{2.16}$$

where $\alpha = \frac{b}{b+1}$, $\mu = \mathrm{E}(m \mid a, b) = \frac{a}{b}$. It is worth pointing out that α depends on var(m) in the population of sites.

Another way of modelling the effect of a safety measure implemented at a site is to define a coefficient θ such that

$$m_{aft} = \theta\, m_{bef}$$

where the two m values represent the expected number of accidents, before and after the implementation. If the remedial treatment has no effect then $\theta = 1$. The difference from this value can be interpreted as an increase or decrease by the same percentage in the expected number of accidents. The value of θ is estimated as shown in Kulmala (1994).

There are other methods for dealing with the regression-to-mean effect, though they are more difficult to apply in practice (Wright et al., 1988). However, only the EB methods are important for the development of the models considered in the second part of the thesis. Wright et al. (1988) describe four main problems about the assumptions made for all methods that need to be carefully considered.

1. The first problem is about the definition of the term "site". For treated sites this is done by local authorities and this may influence the estimate of the true accident rate for that site in future years. However, for the regression models considered in the next subsection and later chapters, the road network is usually divided into nodes (junctions) and links.

2. The second problem is about defining the population. For a given site, do "all" the sites in the study area define the population or only "those" with similar physical characteristics as the treated site? The regression models allow the parameters of the gamma distribution to depend on site characteristics, so the 'population' consists of all sites with the same characteristics.

3. The third problem concerns the "gamma assumption". Following Abbess et al. (1981), this means that the distribution of the true mean accident rates is gamma. This is very convenient from the mathematical point of view but it is a strong assumption. It would be very interesting to know how sensitive the results are to this assumption and whether other distributions such as log normal give satisfactory solutions. Some new approaches are described in this thesis in Chapters 7 and 9.

4. The remedial sites are chosen for treatment because they have a large number of accidents which appear to have causal factors in common. The fourth problem is whether the regression-to-mean effect can be studied in terms of the overall accident frequency at each site. A simultaneous analysis of accident frequencies of various type would certainly be more beneficial. Statistical models for doing this kind of analysis after disaggregation are developed in Chapters 8 and 9.

2.2.3 Regression models for accident frequencies

Very often, a better prediction of future number of accidents is posssible when the covariate information available is linked to the observed number of accidents. This will help in establishing a straightforward method for prediction. Linear regression models using a normal distribution for the error term are not appropriate. Generalized linear modelling gives better modelling flexibility and the predictive accident models developed in the last two decades are included in this general framework. This allows retention of the Pois-

son assumption. Therefore, Poisson log-linear modelling is often used for the regression models for road accident data.

A *generalized linear model*, McCullagh and Nelder (1989), is specified by

$$Y \sim f(\theta, \phi) \tag{2.17}$$

$$\mathrm{E}(Y) \equiv m \tag{2.18}$$

$$h(m) = X'\beta. \tag{2.19}$$

In this, X is a vector of explanatory variables. The relationship between the mean m and the linear predictor $X'\beta$ is modelled by the so called *link function* h. This is possible as long as there is a function h_* such that $\theta = h_*(X'\beta)$. When the error distribution $f(\theta, \phi)$ is Poisson with mean m the canonical link $\theta = \log(m) = X'\beta$ leads to the standard log-linear Poisson-regression model.

Regression models

In the literature there are studied several classes of regression models. A Poisson class of models (Miaou and Lum, 1993) assumes that

$$Y \sim \mathrm{Pois}(m) \tag{2.20}$$

$$m = \mathrm{E}(Y) = \nu[\exp(X'\beta)] \tag{2.21}$$

where ν is an exposure factor, like time for instance. The rate function is $\lambda = \exp(X'\beta)$ which is very convenient being nonnegative. A modified Poisson

regression model, Maycock and Hall (1984), is described by

$$\begin{aligned} Y &\sim \text{Pois}(m) \\ m = \text{E}(Y) &= \nu^{\beta_0}[\exp(X'\beta)] \end{aligned}$$

where the unknown constant β_0 needs to be estimated. If ν is a good exposure measure then the estimated $\widehat{\beta_0}$ should be close to 1.

As pointed out in Miaou and Lum (1993), the Poisson distribution is very useful not only because tests and confidence sets for the estimated regression coefficients can be calculated, but probabilistic statements can be made about Y. This is an important point in favour of using the Poisson distribution, which is discrete. There is no need to look for some other continuous distributions, like the normal that is still used, quite inappropriately, in some investigations, for example Amis (1996).

For predictive accident models traffic flow plays a major role, and should also be considered in before-after studies. Changes in traffic flows influence changes in accident counts between the "before" and "after" periods, and this should be accounted for before making any claims about the effectiveness of any treatment. Traffic flow is also important for estimating the expected accident numbers, and is usually included amongst the explanatory variables X. Quite often accident rates like accidents/vehicle kilometer are used to account for changes in traffic flow as a measure of exposure. This would be correct if the expected accident frequencies like accidents/year were directly

proportional to traffic flow. This common belief is seldom true; the coefficient of flow is significantly different from I.

A further problem is that the exact values for traffic flows are not known and they are replaced by estimates. This may cause further problems if there are random errors in these estimates. If Q is the traffic flow count and z is the true annual average daily traffic (AADT) flow, they can be modelled at the same time using the following model

$$Y \sim \text{Pois}(m \equiv \lambda T) \tag{2.22}$$

$$Q \sim \text{Pois}(zt) \tag{2.23}$$

$$m = T \exp[X'\beta + \log(z)\gamma]. \tag{2.24}$$

An iterative procedure described in Maher and Summersgill (1996), can be used to calculate the estimates of the unknown parameters (β, γ).

The Overdispersion problem

One limitation of the Poisson-regression modelling, well documented in the literature, is that the error variance has to be equal to the mean $E(Y)$ in equation (2.18), see Cox (1983) and Dean and Lawless (1989). However, in practice count data very often shows overdispersion: the error variance is greater than the mean. Ignoring this phenomenon can be very troublesome. Although the maximum likelihood estimators of the regression coefficients are still consistent, the variances of the estimated coefficients tend to be underes-

timated, which means that the significance levels of the estimated coefficients can be misleading. The phenomenon of overdispersion is well-known in many areas of statistics. There are several methods to overcome this difficulty but there is much research under progress searching for better solutions. Some possible reasons for overdispersion in predictive accident models are commented in Maher and Summersgill (1996).

Overdispersion occurs quite often in modelling count data under a Poisson assumption, so the first attempts to solve this problem were based on making more complex distributional assumptions. One solution proposed by Wedderburn (1974) to correct for overdispersion is a quasi-Poisson model (QP). The difference from the classical Poisson model is that $\text{var}(Y) = \tau m$, with the parameter τ accounting for overdispersion. This parameter can then be estimated by any of $G^2/(N-p), X^2/(N-p)$, or $G^2/\text{E}(G^2)$, where N is the number of observations and p is the number of parameters estimated. Simulation studies (Maher and Summersgill, 1996) have shown that the second performs better. For the estimates of the regression parameters β there is no difference compared to the pure Poisson model, but their standard errors are inflated by a factor of $\sqrt{\tau}$. The asymptotic t-statistic for the coefficient of regression can be improved (Agresti, 1990) by multiplying the value for the initial t-statistic, obtained from the Poisson regression model, by $\tau^{-\frac{1}{2}}$. One may obtain the correct adjusted asymptotic standard errors by multiplying the values given by traditional generalized linear modelling software by the scaling factor $\sqrt{\tau} = \sqrt{X^2/(N-p)}$. The inference is then performed in the

classical manner using these adjusted asymptotic standard errors. It can be immediately seen that, when $\tau > 1$, i.e. there is overdispersion, the confidence intervals obtained after adjusting are larger than the unadjusted confidence intervals. Thus, the inferential process is improved by using the correct asymptotic standard errors.

An alternative is to use another discrete distribution instead of the Poisson distribution. Following a Bayesian approach as described above, it seems that the negative binomial distribution (NB) is more suitable, as it allows the variance to be greater than the mean. A third more general solution is to use a more general family of negative binomial distributions for which (QP) and (NB) models are just two special cases (Cameron and Trivedi, 1986). This general model is given by the following assumptions

$$Y_k \sim \text{Pois}(\lambda_k T_k), \quad \text{for all } k \tag{2.25}$$

$$\lambda_k \sim \text{gamma}(\eta, b) \equiv \text{gamma}\left[\mu, \frac{\mu}{b}\right] \tag{2.26}$$

$$\eta = \alpha\mu^j \tag{2.27}$$

where α is a constant factor and the overall mean μ is estimated from the data. From the model specification it follows that

$$p(Y_k \mid \mu, b) = \text{NB}\left(\frac{b}{b+T_k}, \eta\right) \tag{2.28}$$

$$\text{E}(Y_k \mid \mu, b) = \frac{\eta T_k}{b} = \mu T_k \tag{2.29}$$

$$\text{var}(Y_k \mid \mu, b) = \text{E}(Y_k \mid \mu, b)\frac{b+T_k}{b} = \mu T_k\left(1 + \frac{T_k}{b}\right) \tag{2.30}$$

Using equation (2.27) it follows that $b = \alpha\mu^{j-1}$, and this means that

$$\text{var}(Y_k \mid \mu, b) = \mu T_k + \frac{\mu^{2-j} T_k^2}{\alpha}$$

as mentioned in Maher and Summersgill (1996). Thus, $j = 0$ implies that $\eta = \alpha$ and this is the classical NB model used. If $j = 1$ it follows that $\eta = \alpha\mu$ so the shape of the gamma distribution is not constant and it depends on its mean. In this case

$$\text{var}(Y_k \mid \mu, b) = \mu T_k \left(1 + \frac{T_k}{\alpha}\right)$$

and if $T_k = T$ then this model becomes a (QP) model with $\tau = 1 + \frac{T}{\alpha}$.

This methodology can be extended to incorporate covariate information; the parameter μ is then a function of the covariate vector X. In this family of models, for the TRL studies, like the TRL 4-arm roundabout study (Maycock and Hall, 1984), it seems that the (NB) model is more adequate than the (QP) model.

2.3 Selecting sites for treatment

2.3.1 Introduction

The main job of traffic safety engineers is to correct hazardous sites. First, they have to identify the risky locations, then to determine remedial schemes and in the end to implement the best feasible treatment. Choosing the wrong sites is damaging in two ways: firstly, some hazardous sites may be left un-

treated and secondly, large amounts of public money are wasted. Ideally, sites should be ranked by the values of their true means m. These are unknown, but because of random variation, observed numbers of accidents are not entirely reliable. Statistical modelling is often used to improve the methodology. Similar problems are addressed in medicine (Morris and Christiansen, 1996), where profiling hospitals has become very important in recent years, and in education (Laird and Louis, 1989), where ranking schools based on pupil performance data is required for public information and for implementation of better education policies.

Ranking and selection are related to either a "relative" given set of statistical units, in our case sites, and then the units are just compared to each other, or to an "absolute" standard like a given threshold and the purpose is then to identify those units that exceed the threshold. Ranking can be successfully used to indicate good or bad performance. Ranks should contain statistical information that avoid misrepresentation of the precision of estimation. If regression methods can be used to explain the whole between-sites variation there is no basis for ranking.

Generally, sites are ranked according to some safety measure such as accident count or rate. Higle and Witkowski (1988) were the first to propose (EB) methods for ranking locations. The (EB) methods were used to give greater weight to those sites having greater exposure. They were not used because of selection bias, which is not of concern here. The site estimates are different in their *reliability*. For example, if a large number of accidents y_1 is observed

at a site with a high exposure, then there is more confidence that y_1 is close to its true mean value than for a large number y_2 accidents observed at a site with low exposure.

Ranking the sites by their empirical accident frequency, without considering the uncertainty of each estimate, may not correctly identify the worst locations. Nothing can be said about the probability that the worst sites have been selected or about the extent to which the selected sites are really hazardous compared with the non-selected ones. Bayesian and empirical Bayes methods have been used to overcome some of these difficulties, see Hauer (1980), Higle and Witkowski (1988), Davies (1990), Christiansen, Morris and Pendleton (1992). A recent study, proposing hierarchical Bayesian models as a general solution to all the problems highlighted above, is given in Schluter, Deely and Nicholson (1997).

Ranking and selection are based on solving one or more of the following problems (Morris and Christiansen, 1996), here translated for road accident sites.

1. Estimate the maximum or minimum of all means or even find the distribution of this quantity.

2. Determine the site or family of sites that are likely to be the best (or worst).

3. Find the sites that are likely to exceed a given threshold.

4. Obtain the predictive distribution for each of the N sites and calculate

the probability that, for a fixed future period, each site will have the maximum (or minimum) number of accidents

Methods for solving these problems can be based on a Bayesian framework.

Given a tolerance level δ, Higle and Witkowski (1988) called a site k *hazardous* when the probability that λ_k, the expected accident frequency per unit time, is greater than a specified upper limit $\overline{\lambda}$ (a possible acceptable underlying accident mean) exceeds δ. In another study (Davies, 1990) sites were classified by the ratio ρ between the accident mean at each site and the pooled accident means at the remaining sites. For each site under scrutiny, the posterior distribution of ρ is used to obtain the similarity measure

$$\alpha = \Pr(\rho < 1 \mid y_1, \ldots, y_N).$$

When α is small the corresponding site has a higher underlying accident mean than the other sites pooled together and it is therefore selected.

Christiansen et al. (1992) developed a hierarchical Bayesian model for estimation and for ranking the accident sites. The posterior accident mean estimates, adjusted for costs and future traffic volume, are ranked in a decreasing order and sites are selected until a fixed budget constraint is met.

2.3.2 Statistical modelling methodology

Suppose there are N sites labelled $k = 1, 2, \ldots, N$, and at site k there is a total of Y_k accidents over a period of time T_k. The counts Y_k are assumed

independent with means λ_k, where $\lambda_k > 0$.

The hierarchical models are developed in several stages. First of all, the mean per unit time λ_k is considered a random variable with prior distribution $f(\cdot \mid \beta, \nu)$. Then the *hierarchical* Bayesian method considers a hyper-prior distribution h on the parameters β and ν, in a second stage. Under the assumption of exchangeability the prior distribution of $\lambda = (\lambda_1, \ldots, \lambda_N)$ is

$$f(\lambda) = \int_\nu \int_\beta \prod_{k=1}^{N} f(\lambda_k \mid \beta, \nu) h(\beta, \nu) d\beta d\nu. \tag{2.31}$$

The hyper-prior $h(\beta, \nu)$ can be factorised as

$$h(\beta, \nu) = h_1(\beta) h_2(\nu \mid \beta) \tag{2.32}$$

using prior information about the nature of parameters β and ν. The posterior distribution of the parameter of direct interest λ, given the observed data $y = (y_1, \ldots, y_N)$, can be written as

$$f(\lambda \mid y) = \int_0^\infty \int_0^\infty \frac{f(\lambda, y, \beta, \nu)}{p(y)} d\beta d\nu \tag{2.33}$$

$$f(\lambda \mid y) = \int_0^\infty \int_0^\infty f(\lambda \mid y, \beta, \nu) \frac{p(y \mid \beta, \nu)}{p(y)} h_2(\nu \mid \beta) h_1(\beta) d\beta d\nu \tag{2.34}$$

where

$$p(y) = \int_0^\infty \int_0^\infty p(y \mid \beta, \nu) h_2(\nu \mid \beta) h_1(\beta) d\beta d\nu \tag{2.35}$$

is the marginal distribution of the observed data y. Because the Bayesian calculus involves only the expectation of the posterior distribution or other measures such as mean or mode, the exact form of the posterior distribution is not a matter of specific concern. However, it has to be remarked that under the gamma assumption, $p(y \mid \beta, \nu)$ is a product of negative binomial distributions. The specification of the hyper-prior distribution $h(\beta, \nu)$ is not easy. Schluter et al. (1997) provide an interesting discussion in connection with the ranking problem.

Based on the previous methodology, Schluter et al. (1997) proposed three criteria for ranking. These will be explained in turn.

Ranking using the posterior probability that a site is the worst site

For a given type of accident or the total number of accidents, if λ_k is the accident mean at the site k, then the posterior probability that the site k is the worst one can be calculated as

$$p_k(v) = \Pr(\lambda_k > v\,\lambda_j, \quad \text{for all } j \neq k \mid y)$$

where $v \in [0, \infty)$. If $v = 1$ then $p_k(v)$ is the probability that the site k is the worst site. Only for this value of v the sum of $p_k(v)$ equals 1, so they are true probabilities. The practitioners specify v a priori. Then either the first r largest values $p_k(v)$ or the smallest group of sites with summed values $p_k(v)$ greater than some threshold value $P*$, are selected. If the results are not satisfactory, for instance only two or three sites are selected, then the value of

v can be lowered and the ranking process repeated. This criterion is designed for long term projects and calculates a measure of uncertainty, based on a pre-specified distance quantity v.

Ranking using the predictive probability of future accidents

For a given threshold number n_0, if $\tilde{Y}_k$ is the future number of accidents in the next period at site k, then

$$\mathrm{pd}_k(n_0) = \Pr(\tilde{Y}_k \geq n_0|y) = \int_{\{y_k \geq n_0\}} p(y_k \mid \lambda, y) f(\lambda \mid y) d\lambda$$

is the Bayesian predictive probability that the future number of accidents will exceed an important future target accident number. Again, the selection is made by taking either the first r largest $\mathrm{pd}_k(n_0)$ values or all the sites having $\mathrm{pd}_k(n_0) \geq P_0$, where P_0 is fixed. This criterion is designed for short term objectives because it uses the probability of future numbers of accidents in the next period.

Ranking using the posterior mean

The posterior mean

$$\mathrm{E}(\lambda_k \mid y) = \int_0^\infty \lambda_k f(\lambda_k \mid y) d\lambda_k$$

is the most commonly used measure. Selection is made either by taking the r largest $\mathrm{E}(\lambda_k \mid y)$ or by retaining all sites for which $\mathrm{E}(\lambda_k \mid y) \geq e_0$, where e_0 is

a given threshold value. This measure is probably the most easily calculated of all three. It is an estimate of the underlying mean and it can be used for long term forecasts.

However, as pointed out by Laird and Louis (1989) and Morris and Christiansen (1996), this approach can be misleading. A more reliable method is to estimate the actual ranks of the parameters of interest corresponding to the observational units, which in this thesis will be the means λ_k of the Poisson distributions. The beauty of the Bayesian methodology coupled with MCMC methods is that the entire posterior distribution of ranks can be estimated.

It would be very useful if the above methodology could be further developed and hierarchical models for *multiple* counts considered to rank the sites according to different criteria. Nothing has been done apparently about ranking hazardous locations when multiple accident counts are jointly investigated. Practitioners prefer to use data at an aggregated level, mainly because of lack of statistical models that can be used for multiple counts. For the same period of observation, if one site has a total of 30 accidents, out of which 6 are KSI, and another site has a total of 15, out of which 10 are KSI, then, looking only at the totals, the first site seems more hazardous than the second one. But if only the number of KSI accidents is considered then the second site is more hazardous than the first one. Therefore developing models for ranking multiple accident counts would provide a much better analysis. Three hierarchical Bayesian models are investigated for ranking 156 link sites in Chapter 9.

Another hierarchical model used for ranks was proposed by Maher and

Mountain (1988). The model is specified in three stages

$$Y_k \sim \text{Pois}(\lambda_k) \tag{2.36}$$

$$\lambda_k \sim \text{gamma}\left(\alpha, \frac{\alpha}{\delta_k}\right) \tag{2.37}$$

$$\delta_k \sim \text{gamma}\left(\beta, \frac{\beta}{a}\right) \tag{2.38}$$

and the difference $\lambda_k - \delta_k$ represents the quantity by which the mean accident frequency at the site k exceeds the average mean for a site with fixed characteristics of that type. Maher and Mountain (1988) ranked the sites by the potential accident reduction criterion (PAR), that is by $y_k - \widehat{\delta}_k$, where $\widehat{\delta}$ is an estimate. It was shown that this criterion is better than ranking based on annual accident totals, provided that the estimation of δ is accurate enough. This model is an improvement because it is not based only on the observed total accident counts at each site and because covariates can be easily included. Although (PAR) shows great promise there are several drawbacks for using this model in this form. One major criticism is that the estimated average means $\widehat{\delta}$ and the observed counts y are assumed to be sufficient for calculating the ranks. The environment may experience dynamic changes in many unobserved ways with results in increasing or decreasing the number of accidents. The plain observed counts are unreliable for ranking purposes, but fully Bayesian or EB methods combine the data from other sites and therefore are more reliable, especially if random effects are employed, for estimating $\text{E}(\lambda_k \mid y)$ or for ranking the sites. In addition, nothing has been said, re-

garding the (PAR) criterion, about the uncertainty associated with the ranks. Even when two sites have different ranks, if their uncertainty intervals are quite overlapped then the difference may be due to the particular estimation procedure chosen. A solution to all these problems is sketched in Chapters 8 and 9, using hierarchical models combining regression with random effects in a Bayesian framework.

2.4 Models for type of accidents

The first category of studies described in Section 2.2 focused on statistical modelling of accident frequencies as random variables. A second category of applications is looking at the characteristics of the accidents which have occurred, such as the severity of injury, the date (day, month, year), location, speed limit, road classification and so on. The unit of the statistical analysis is different from that in the previous category of studies. Each accident is a unit of the sample and the random variables are the characteristics of the accident, given that the accident has occurred.

There will typically be a large number of variables. There is an obvious interest in identifying the *association* or *independence relationships* among the variables. An example is in Salminen and Heiskanen (1997), where the correlations between accidents in traffic, at work, at home and during sports and leisure time were investigated. The product moment correlation was used as the main tool. Even after logarithmic transformations, the correlations

were unchanged and still low. The study used data for 3 years 1980, 1988 and 1993, the matrix of correlation changing over time.

Another study of which declared purpose was to investigate the characteristics of pedal cycle accidents at T-junctions is Henson (1992). A number of ten variables representing various accident factors were analysed using log-linear models for data summarised in contingency tables. The analysis was conducted on several marginal two dimensional and three dimensional tables and it was inconclusive. Henson (1992) required a larger database to get better results. The data was indeed sparse, comprising only 272 reported injury accidents, but the statistical methodology used, analysing several marginal tables, is potentially misleading. However, there are better techniques available for studying associations between variables that will be described and applied in this thesis in Chapters 3, 4 and 5. It will be shown in this thesis how to conduct an exploratory analysis on a single large table cross-classified by all variables under study. It will also be shown how to avoid model selection problems for sparse tables by using exact conditional tests.

Studies of accident characteristics are observational in the same sense as studies regarding accident frequencies at individual sites. A retrospective view is taken, conditioning on the fact that accidents have occurred, so only characteristics of observed accidents are recorded. In this thesis we will call by "road accident characteristics" features of accidents such as accident severity, the number of casualties, the number of vehices involved in the accident; characteristics of the road network such as road class, speed limit; environmental

conditions such as road surface conditions, hazardous objects on the road; temporal characteristics such as day of the week, hour of the day and so on. The report book, that was used by Thames Valley Police to collect data about contributory factors of accidents, contains a total of 33 variables of this kind.

There are some studies about road accident characteristics (Taylor and Barker, 1994-1995; Maycock, 1985), but the approach is more descriptive rather than trying a statistical inferential approach. Generally there is a lack of exploratory studies of large data sets in this area. Several applications will be given in this thesis in Chapters 5 and 6 continuing the work described in Tunaru and Jarrett (1998*b*) and Tunaru and Jarrett (1998*a*).

For tables of small dimension cross-classified by accident characteristics the class of log-linear models has been used (Fienberg, 1980) successfully for statistical modelling. A subset of data of this type extracted from Kihlberg, Narragon and Campbell (1964) has been analysed in textbooks, see Fienberg (1980) and Christensen (1990). This small table is used in Chapters 3, 4, 6 and 8 as a general example to illustrate the theoretical concepts involved. Another example of a log-linear analysis is described in Agresti (1996), examining the characteristics of passengers in cars and light trucks involved in accidents. The 4-dimensional contingency table contains data on 68,694 passengers in the state of Maine in 1991 and the analysis revealed that, even for a large sample size, asymptotic significance tests can be unreliable. This conclusion will be reconfirmed by the results obtained in Chapter 5.

2.5 Models for accident frequencies and type of accidents

The previous two categories of variables are sometimes studied jointly, developing models for relating road environment factors to both accident frequency and the type of accident. The models falling into this category try to relate the total number of accidents at a junction or along a length of road to a number of explanatory road environment variables, and also to investigate which variables are associated with the type of accident (Amis, 1996; Mountain et al., 1996). In Amis (1996), an exploratory stepwise multiple regression approach was proposed in the first stage in order to determine which covariates should be retained for further regression modelling. If the square root is taken to normalise the Poisson variable, the model proposed first is

$$\sqrt{Y_k} = \alpha + X_k'\beta + \varepsilon$$

where Y_k is the number of accidents at the site k, β is a vector of parameters, X_k is a vector of covariates and ε is an error term having the standard normal distribution. In the second stage generalized linear models are fitted either for accident frequencies or for accident type. For example, if the site is defined as a junction, then the generalized linear model for accident frequencies proposed in Amis (1996) is

$$Y_k \sim \text{Pois}\left(M_k \times (\exp(\alpha + X_k'\beta))\right)$$

where M_k is either the time period or the link length. A logistic model is discussed in the same paper for accident type.

Stepwise multiple regression applied in an automated way can easily lead to misleading results. A particular covariate can be evaluated as significant as well as non-significant, depending on what explanatory terms are included in regression. This model selection procedure should be used with great caution. In addition, instead of attempting to normalise the Poisson variable, it would be better to use a Poisson or NB regression model. However, the idea as a whole is very interesting and further research could usefully be done in this area. This may require a multivariate approach and a general framework is proposed now.

Suppose that there are Y_{ki} accidents of type i, at site k, that are Poisson distributed with mean λ_{ki}, where $i = 1, 2, \ldots, M$, $k = 1, 2, \ldots, N$. Given the means λ_{ki}, the accident frequencies Y_{ki} are assumed independent from site to site, but accident frequencies of different types are not assumed independent. From the properties of Poisson distribution, the total number of accidents of type i is $Y_{+i} \sim \text{Pois}(\lambda_{+i})$. Conditioning on the total number of accidents over all sites Y_{++}, it follows (Santner and Duffy, 1989) that

$$(Y_{+1}, Y_{+2}, \ldots, Y_{+M}) \mid \{Y_{++} = n\} \sim \text{Multi}(n, p_i)$$

where the probabilities $p_i = \frac{\lambda_{+i}}{\lambda_{++}}$. Making a strong assumption that the mean number of accidents can be calculated multiplicatively as a product of a site

effect and an accident type effect, $\lambda_{ki} = \mu_k \theta_i$, it follows that $\lambda_{+i} = \mu_+ \theta_i$ and therefore $p_i = \frac{\theta_i}{\theta_+}$. The next step is to consider a log-linear model for the vector of probabilities $(p_i)_{i=1,2,\ldots,M}$. Therefore, this is a log-linear analysis of accident characteristics. Thus, conditioning on the fact that the accidents have occurred and knowing various information about the characteristics of accidents and road network, the relationships between these categorical variables can be investigated and the conclusion can be drawn about accidents as a whole on that road network. Since there are many variables of interest regarding accident characteristics, the log-linear modelling, in this context, should be able to deal with large probability vectors in an efficient manner. A statistical technique that does just that is graphical modelling which will be the subject of the following four chapters.

2.6 Summary

In this chapter a number of different statistical models for road accidents have been reviewed. Statistical modelling for road accident data was greatly improved by applying generalized linear models. Accident data can be viewed as an example of count data in general and therefore models for counts developed in other areas of research can be also applied here. Nevertheless, accident data has some specific characteristics that makes it more difficult to analyse. Data come from observational studies and it is almost always sparse, that is many counts are zero or very small. This means that classical techniques applied

in other areas cannot be always applied here and various changes need to be made.

More powerful statistical methods are required to handle large and complex road accident datasets. Graphical modelling offers a solution to study the relationships between the variables under study, usually a large number, and multiple response variables models would give accident prediction modelling a new dimension.

Predictive accident models were developed mainly at an univariate level. The lack of models for joint types of accident gives the statistician an opportunity to research a vast area. The benefits would be a better and more structured information for local authorities that could in return spend the money more wisely and help reducing the number of accidents further.

Chapter 3

Graphical log-linear models

3.1 Introduction

In the last decade graphical modelling has become an important tool in applied statistical modelling. A graphical model is usually identified with a pictorial representation of a statistical model, thus making a straightforward connection with graph theory. Graphical models are mainly used to represent conditional independencies and they cover exploratory studies, where all variables are treated as response variables, and more causal approaches where the variables are divided into response and explanatory blocks. The potential of applications includes biostatistics, genetics, sociology, education studies- see the examples in Edwards (1995) and Mohamed, Diamond and Smith (1998), and credit scoring in finance (Hand, McConway and Stanghellini, 1997; Stanghellini, McConway and Hand, 1999) among others. These models can be also applied to econometrics (Lyngaard and Walther, 1993) and theoretical statistics

(Stanghellini, 1997). A directed graph representing an econometric model for traffic fatalities has been illustrated in Roh, Bessler and Gilbert (1999).

In addition, the concept of graphical model is fundamental in the development of Markov Chain Monte Carlo strategies for applied Bayesian statistics. It is also used as a tool for communicating complex statistical models analysed in the computer program WinBUGS: see Spiegelhalter, Thomas and Best (1998).

In this chapter, the theoretical elements on which graphical modelling is based are reviewed, and some terminology from graph theory, used in the subsequent chapters, is introduced. A 4-dimensional contingency table is used throughout to illustrate various concepts related to graphical modelling.

In Section 3.2 the motivation for applying graphical modelling for analysing large contingency tables is given. Section 3.3 contains a short revision of conditional independence and a list of various concepts of graph theory used later on. Then, in Section 3.4, the Markov properties defining the graphical models and the methodological skeleton for practical applications are outlined. Chain graphical modelling is a generalisation of graphical modelling for situations when variables are ordered by some causal a priori assumption. The corresponding Markov properties and other results are summarised in Section 3.4.2. Various model selection procedures are discussed in Chapter 4 and a new battery of Bayesian model selection procedures that can be applied for contingency tables is proposed in Chapter 8, Section 8.4.

Chapters 5 and 6 are complementary to this one, discussing some practical

applications to large and sparse contingency table summarising road accident data and methods of reducing the dimensionality of the statistical analysis with the help of a collapsibility concept and corresponding theoretical results as given in Asmussen and Edwards (1983).

3.2 The need for graphical modelling

A national road accident database will contain a large number of variables representing characteristics of the recorded road accidents. An important problem is then to identify the associations, or in a complementary way, the conditional independence relationships between the variables under study. For statistical analysis, the data can be summarised in a multi-dimensional contingency table cross-classified by the variables under study. Because of the Yule-Simpson paradox (Simpson, 1951), the analysis of marginal tables, involving only two or three variables at a time, can be very misleading.

Consider, for instance, a subset of data reported in Kihlberg et al. (1964). The variables are

- A = Driver ejected (No / Yes)
- B = Car type (Small / Standard)
- C = Injury type (Not severe / Severe).

and the data is shown in Table 3.1.

The statistical analysis of contingency tables like this, where all variables are viewed as response variables, is based on the class of log-linear models,

Table 3.1: A 3-way contingency table of road accidents

Driver Ejected	*Car type*	*Injury type*	
		Not Severe	Severe
No	Small	410	262
	Standard	2026	1426
yes	Small	45	103
	Standard	133	426

Source: Kihlberg et al. (1964).

which is perhaps the most useful class for contingency tables (Haberman, 1974; Bishop, Fienberg and Holland, 1975; Christensen, 1990). Log-linear models express the logarithms of the cell probabilities as sums of main effects and interaction terms, by analogy with analysis of variance (ANOVA) models for continuous data. The parameter p_{ijk} represents the probability of the cell at the intersection of level i of A, level j of B and level k of C. The saturated log-linear model for a three dimensional contingency table can be parameterised as

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)},$$

where, for instance, the term $u_{12(ij)}$ represents the interaction between variables A and B. The terms $u_{12(ij)}, u_{13(ik)}, u_{23(jk)}$ are called *two-way interaction terms* and $u_{123(ijk)}$ is called a *three-way interaction term.* The mutual independence model, that is the model which specifies that all variables cross-

classifying the table are independent, is

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)},$$

and the all two-way interaction model is

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}.$$

Since there are more parameters on the right side of the equation than on the left side, the representations are overparameterised. Thus, in order to have a unique representation, some ANOVA-like constraints are imposed. The main effects are always kept in the model because it is very hard to interpret a model having a higher-order relative term of a main effect not present in the model.

Since there are only three variables involved in this example, it is possible to test the fit of all possible models. The results are shown in Table 3.2. In this table, the model formula is expressed using only the terms of highest interaction. Thus, for example, $[A][B][C]$ denotes a log-linear model containing the main effects of the factors A, B and C, while $[AB][BC]$ is an abbreviation for $A + B + C + A.B + B.C$, a model containing all the main effects and the interactions between factors A and B and respectively, between B and C. The *scaled deviance* (McCullagh and Nelder, 1989) is the statistic employed to test the model. The saturated model $[ABC]$ fits the data perfectly and has deviance 0. In general, the scaled deviance is the generalised log-likelihood

ratio statistic for comparing each model with the saturated model. Under the null hypothesis that a particular model is correct, the scaled deviance is asymptotically distributed as chi-squared with the indicated number of degrees of freedom (denoted by *df*). This distribution is used to calculate the P-value, the probability of obtaining the observed or a larger deviance. There are two

Table 3.2: Models fitted to the collision-rollover data

Model	*Formula*	*Scaled deviance*	*df*	*P-value*
1	$[A][B][C]$	298.69	4	0.000
2	$[AB][C]$	289.89	3	0.000
3	$[AC][B]$	12.69	3	0.005
4	$[BC][A]$	297.93	3	0.000
5	$[AB][BC]$	289.13	2	0.000
6	$[AB][AC]$	3.89	2	0.143
7	$[BC][AC]$	11.93	2	0.003
8	$[AB][BC][AC]$	1.15	1	0.284
9	$[ABC]$	0.00	0	1.000

models that fit the data well, models $[AB][BC][AC]$ and $[AB][AC]$. The second model is nested within the first one so it is preferred because it has fewer parameters and is easier to interpret in terms of conditional independencies. The model informs us that car type, B, and injury type, C, are independent given driver ejected, A. This means that, knowing whether the driver has been ejected or not in an accident, finding out the type of the car will not help in any way to predict the type of injury in that accident. Therefore, B is irrelevant to C when A is known, or in other words it is only A which is associated with C. This conditional independence relationship is denoted, following

Dawid (1980), by $B \perp\!\!\!\perp C \mid A$. A similar notation will be used throughtout the thesis for the conditional independence of sets of random variables.

The problem here is that there is more information available and there is a fourth variable D = Accident Type (Collision / Rollover) so the Table 3.1 can be further cross-classified. Even if the interest is focused on the relationship between type of injury, type of car and driver being ejected it is not wise to take out of the analysis the variable D, the accident type. The full data is shown in the 4-dimensional Table 3.3.

Table 3.3: 4-way contingency table of road accidents

A	*B*		*C*	
Driver	*Car*	*Accident*	*Injury type*	
Ejected	*type*	*type D*	Not Severe	Severe
No	Small	Collision	350	150
		Rollover	60	112
	Standard	Collision	1878	1022
		Rollover	148	404
yes	Small	Collision	26	23
		Rollover	19	80
	Standard	Collision	111	161
		Rollover	22	265

Source: Kihlberg et al. (1964).

One of the models that fits Table 3.3 well is $[ACD][BCD]$. The importance of this model will be better described in the context of various model selection procedures compared in Chapter 4. This model can be again interpreted in terms of conditional independence such as $A \perp\!\!\!\perp B \mid \{C, D\}$, which means that driver ejected is independent of car type given accident type and

injury type. In addition, B and C are not conditionally independent which seems to contradict the analysis of the 3-dimensional table. Moreover, the conditional independence between A and B is in contradiction with the previous conclusion. This phenomenon, where a relationship between two variables is changing to the opposite when more (or less variables) are considered, is called Yule-Simpson paradox or just Simpson's paradox and it highlights the importance of taking a multivariate approach, by involving all relevant variables under study. Therefore a powerful technique is needed to analyse large tables in an efficient manner without losing important information or arriving at misleading conclusions.

This looks like a problem without any solution. On the one hand all the variables under study should be considered in order to avoid Simpson's paradox, and on the other hand there is a natural tendency to simplify the picture to have more reasonable interpretations. This is where *graphical modelling* comes in as a very useful exploratory technique for describing the conditional independencies between the variables.

3.3 Preliminaries and terminology

3.3.1 Background

The seminal ideas of graphical modelling can be found in several areas of science where statistics plays an important role:

1. in statistical physics, Gibbs (1902) studied a large system of particles of a gas or a solid where, for a subgroup of particles, only the interactions between the particles in the subgroup and the neighbour particles are considered significant.

2. in genetics, path analysis (Wright, 1934) was proposed for studying heritable properties of natural species using graphs with arrows from parents to children. These ideas were later taken up in economics and social sciences for developing causal models (Wold, 1954; Wold, 1960; Blalock, 1971).

3. in theoretical statistics, Bartlett (1935) used interactions for contingency tables in a similar way to their use in statistical physics. The counts in a group of cells of the table were independent of the counts in the rest of the table , given the counts in the boundary of the group.

Graphical models have been developed for categorical variables as a subclass of hierarchical log-linear models (Darroch, Lauritzen and Speed, 1980; Lauritzen, 1996; Whittaker, 1990; Wermuth and Lauritzen, 1990; Edwards, 1995), for continuous Gaussian variables, better known as covariance selection models (Wermuth, 1976; Whittaker, 1990; Lauritzen, 1996), and for a mixture of continuous Gaussian and categorical variables (Lauritzen, 1989; Edwards, 1990; Edwards, 1995). In the first part of the thesis only graphical models for categorical variables and graphical chain models are considered. Other forms of graphical models, like directed graphical models or graphical chain models

with categorical, discrete and continuous (not necessarily normal) variables, are used implicitly in conjunction with Markov Chain Monte Carlo methods applied in the second part of the thesis.

3.3.2 Graph theory concepts

The account of the elements of graph theory in this subsection follows Lauritzen (1996). Formally, a graph is a pair $\mathcal{G} = (V, E)$ where V is a finite set of vertices (which in this thesis correspond to the variables under examination) and E is the set of edges, which is a subset of the set of ordered pairs of distinct elements of V. The number of vertices in V is denoted by $|V|$. All the graphs in this thesis are assumed to be *simple*, that is no multiple edges or loops are allowed. If $(a, b) \in E$ but $(b, a) \notin E$, then the edge is called directed and is represented by an arrow from a pointing towards b; it is said that a is a *parent* of b and b is a *child* of a, denoted by $a \rightarrow b$. The set of parents of b is denoted by $\mathrm{pa}(b)$ and the set of children of a as $\mathrm{ch}(a)$. If both $(a, b) \in E$ and $(b, a) \in E$ then the edge is undirected and represented by a line joining a to b; the vertices are then called *adjacent* or *neighbours*, denoted by $a \sim b$. The set of neighbours of a vertex a is denoted by $\mathrm{ne}(a)$. For a subset of vertices $\mathcal{A}$, the notations $\mathrm{pa}(\mathcal{A}), \mathrm{ch}(\mathcal{A})$ and $\mathrm{ne}(\mathcal{A})$ denote the collection of parents, children and neighbours respectively of vertices in $\mathcal{A}$ that are not themselves elements of $\mathcal{A}$. For example, in the graph in Figure 3.1, S and N are both parents of A whereas $\{S, T, L\}$ is the parental set of the set $\{A, N\}$. In the same time, R has only one neighbour L and N has none. In spite of the fact that N has no

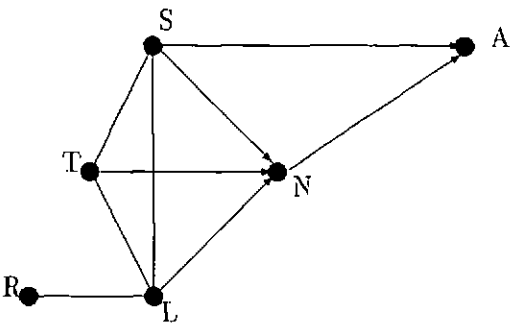


Figure 3.1: A simple graph, neither directed nor undirected

neighbours, this node has parents and children so the concept of *boundary* of a set of vertices $\mathcal{A}$, defined as $\mathrm{bd}(\mathcal{A}) = \mathrm{pa}(\mathcal{A}) \cup \mathrm{ne}(\mathcal{A})$, is just a natural extension of the set of neighbours for situations where there is a mixture of directed and undirected edges. Another useful concept from graph theory that will be used later is the *closure* of a set of vertices $\mathcal{A}$, defined as $\mathrm{cl}(\mathcal{A}) = \mathcal{A} \cup \mathrm{bd}(\mathcal{A})$. A graph with only undirected edges is called an *undirected graph*, and if all edges are directed the graph is called *directed*. The undirected graph obtained from $\mathcal{G}$ by replacing arrows with lines is called the undirected version $\mathcal{G}^{\sim}$ of $\mathcal{G}$. A set of vertices that has all possible pairs adjacent is called *complete*. A subset $\mathcal{A} \subseteq V$ is a *clique* if it is complete and there is not other subset $\mathcal{B}$, $\mathcal{A} \subseteq \mathcal{B} \subseteq V$ that is also complete.

A *path* of length n from a to b is a sequence $a = a_0, \ldots, a_n = b$ of distinct vertices such that $(a_{i-1}, a_i) \in E$ for all $i = 1, 2, \ldots, n$. An *n-cycle* is a path of length n with $a = b$. The cycle is called *directed* if one or more of its

edges are arrows. If there is a path from a to b, denoted by $a \mapsto b$, and if in addition $b \mapsto a$, it is said that a and b are connected. This is an equivalence relationship and the equivalence classes are called connectivity components. A subset $\mathcal{S} \subseteq V$ is said to be an (a, b) *separator* if all paths from a to b intersect $\mathcal{S}$. The subset $\mathcal{S}$ is said to *separate* $\mathcal{A}$ from $\mathcal{B}$ if it is an (a, b) separator for every $a \in \mathcal{A}, b \in \mathcal{B}$. For the graph in Figure 3.1, $\{S, N\}$ separates $\{A\}$ and $\{R, T, L\}$. The vertex a such that $a \mapsto b$ and $b \not\mapsto a$ is called an *ancestor* of b, and the vertex b is called a *descendant* of a. The set of ancestors of all vertices from the subset b is denoted by $\mathrm{an}(b)$ and the set of descendants of all vertices from a subset a is denoted by $\mathrm{de}(a)$. The set of non-descendants of a is denoted by $\mathrm{nd}(a)$. For the graph in Figure 3.1, the ancestors of A are R, L, T, S, N and A has no descendants.

If $\mathrm{bd}(a) \subseteq \mathcal{A}$ for all $a \in \mathcal{A}$ then $\mathcal{A}$ is said to be an *ancestral* set. fn an undirected graph, the ancestral sets are unions of connectivity components. The intersection of a collection of ancestral sets is again ancestral, so there is a smallest ancestral set containing $\mathcal{A}$ which is denoted by $\mathrm{An}(\mathcal{A})$. For example, in the graph of Figure 3.1, $\mathrm{An}(\{R, L, N, S, T\}) = \{R, L, N, S, T\}$ whereas $\mathrm{An}(\{A, L, N\}) = \{A, N, S, L, T, R\}$.

Chain graphs are graphs where the vertex set V can be partitioned into numbered blocks, forming a so-called dependence chain $V = V(1) \cup \cdots \cup V(T)$, such that all edges between vertices in the same block are undirected and all edges between different blocks are directed, pointing from the blocks with lower numbers to the blocks with higher numbers. These graphs are characterised

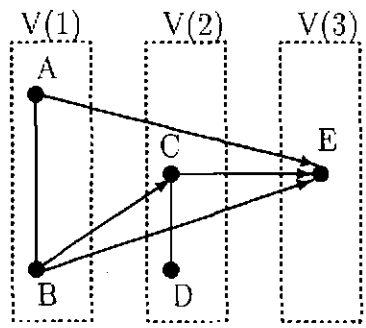


Figure 3.2: Chain graph with dependence chain $\{A, B\} \cup \{C, D\} \cup \{E\}$

by having no directed cycles. The connectivity components indicate the block partitioning of the chain graph. A graph $\mathcal{G}$ is a chain graph if and only if its connectivity components induce undirected subgraphs. It is easy to identify the chain components simply by removing all directed edges before taking connectivity components. The graph in Figure 3.2 illustrates the definitions given above. The boxes are not part of the graph, but are used to indicate the partition into blocks of the chain graph. The connectivity components are easy to determine: they are $\{A, B\}, \{C, D\}$ and $\{E\}$ and these are the blocks. If there had been an arrow from C to A, the graph could not have been a chain graph, even after determining the new connectivity components, which would have been $\{A, B, C, D\}$ and $\{E\}$. The reason is that there would be a directed cycle $A \rightarrow B \rightarrow C \rightarrow A$ and this is not allowed by definition because it will create problems regarding interpretability of the model and model specification.

The *moral graph* $\mathcal{G}^m$ of a chain graph $\mathcal{G}$ is the undirected graph with the same vertex set V but with $a \sim b$ in $\mathcal{G}^m$ if and only if either $a \rightarrow b$ or $b \rightarrow a$

or if there are δ_1, δ_2, connected in the same block, such that $a \rightarrow \delta_1$ and $b \rightarrow \delta_2$. For a directed acyclic graph, that is a directed graph with no cycles, the moral graph is obtained from the original graph by "marrying parents" with a common child and subsequently deleting directions on all arrows.

A triangulated graph is an undirected graph with the property that every cycle of length $n \geq 4$ has a chord, that is two non-consecutive vertices that are neighbours. This type of graph is sometimes also called *chordal*.

3.3.3 Conditional independence

Suppose that $X_V = (X_1, \ldots, X_d)$ is the entire set of random variables of interest (often denoted by the index V), where each variable X_v takes values in a set Ω_v. Then X_V takes values in $\Omega = \Omega_V = \prod_{v \in V} \Omega_v$. If $A \subseteq V$ let $\Omega_A = \prod_{v \in A} \Omega_v$ and the elements of Ω_A will be denoted by $x_A = (x_v)_{v \in A}$ and the corresponding vectors will be denoted as $X_A = (X_v)_{v \in A}$.

In this thesis, f (or sometimes p) is used as a generic symbol for the probability density of the random variables involved. The random vectors X and Y are called conditionally independent given the random vector Z if and only if

$$f(x, y \mid z) = f(x \mid z) f(y \mid z)$$

for all triples (x, y, z) for which $f(z) > 0$. Given the applied character of the thesis, only variables having a positive density probability function, for continuous variables, or a positive mass probability function, for discrete vari-

ables, are taken into account. This will ensure that all conditional densities are defined.

Several equivalent definitions of conditional independence are

$$\begin{aligned} f(x,y,z) &= f(x,z)f(y,z)/f(z), \\ f(x \mid y,z) &= f(x \mid z), \text{ and} \\ f(x,y,z) &= h(x,z)k(y,z), \text{ for some } h,k. \end{aligned}$$

It is very easy to prove the following properties (Lauritzen, 1996), where h denotes an arbitrary measurable function on the sample space Ω_X:

(C1) : if $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$ (Symmetry)

(C2) : if $X \perp\!\!\!\perp Y \mid Z$ and $U = h(X)$ then $U \perp\!\!\!\perp Y \mid Z$ (Reduction)

(C3) : if $X \perp\!\!\!\perp Y \mid Z$ and $U = h(X)$ then $X \perp\!\!\!\perp Y \mid (Z,U)$ (Redundance)

(C4) : if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y,Z)$ then $X \perp\!\!\!\perp (Y,W) \mid Z$

(Contraction)

(C5) : if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$ then $X \perp\!\!\!\perp (Y,Z)$. (Strong contraction)

It should be noted that the assumption of the positive densities or mass functions is needed to prove (C5).

Axioms of irrelevance for road accident characteristics

As pointed out in Lauritzen (1996) the first four properties (C1)–(C4) can be interpreted in a non-probabilistic language as general axioms of *irrelevance*. A model of irrelevance is given by the graph separation property for undirected

graphs. If A, B and C are subsets of the vertex set V of an undirected graph $\mathcal{G} = (V, E)$ the separation of A and B by C is denoted by $A \overset{\mathcal{G}}{\perp} B \mid C$. Then it can be easily checked that

(C1) Symmetry: if $A \overset{\mathcal{G}}{\perp} B \mid C$ then $B \overset{\mathcal{G}}{\perp} A \mid C$

(C2) Reduction: if $A \overset{\mathcal{G}}{\perp} B \mid C$ and $U \subseteq A$ then $U \overset{\mathcal{G}}{\perp} B \mid C$

(C3) Redundance: if $A \overset{\mathcal{G}}{\perp} B \mid C$ and $U \subseteq A$ then $A \overset{\mathcal{G}}{\perp} B \mid (C \cup U)$

(C4) Contraction: if $A \overset{\mathcal{G}}{\perp} B \mid C$ and $A \overset{\mathcal{G}}{\perp} D \mid (B \cup C)$ then

$A \overset{\mathcal{G}}{\perp} (B \cup D) \mid C$

If the subsets are disjoint then

(C5) Strong contraction: if $A \overset{\mathcal{G}}{\perp} B \mid C$ and $A \overset{\mathcal{G}}{\perp} C \mid B$ then $A \overset{\mathcal{G}}{\perp} (B, C)$ is also true. This correspondence shows that graphs can be used to conceptualise and communicate complex scientific ideas. The use of graphs in this way will become particulary important in Chapters 8 and 9 where, in relation with Bayesian graphical modelling, it gives a basis for computation as implemented in WinBUGS (Spiegelhalter, Thomas and Best, 1996).

3.4 Graphical models for contingency tables

3.4.1 Graphical Models

Let $X_V = (X_v)_{v \in V}$ be a vector of $d = |V|$ *categorical* random variables. The categories are labelled by positive integers so that each variable X_v takes values in $\Omega_v = I_v = \{1, 2, \ldots, r_v\}$. Let $\mathcal{I} = \prod_{v \in V} I_v$ denote the set of all possible configurations of X_V. For any subset $A \subseteq V$, let $\mathcal{I}_A = \prod_{v \in A} I_v$. The *cells* of

the contingency table resulting from cross-classification of X_V are indicated by $i \in \mathcal{I}$, and $i_A \in \mathcal{I}_A$ denotes a cell from the marginal table of X_A. Suppose that observational units are classified according to factors in V and the data is summarised in contingency tables, by counts $\mathbf{n} = \{n(i) : i \in \mathcal{I}\}$ where $n(i)$ is the number of units that fall in the ith cell. The table has dimension equal to the number of variables d. For $A \subset V$ the counts on the A-marginal table $\mathbf{n}_A = \{n(i_A) : i_A \in \mathcal{I}_A\}$ are given by summation over all cells in $\mathcal{I}_{V \setminus A}$ so for $A = \emptyset$ it follows that

$$n(i_\emptyset) = \sum_{i \in \mathcal{I}} n(i) = |\mathbf{n}| = N,$$

the total number of observations. Considering $B \subset V$ and a cell $i_B \in \mathcal{I}_B$, the *i_B-slice* of the table is obtained by classifying only those observations for a fixed level of each variable in B. This means that the i_B-slice has cells in $\mathcal{I}_A$ where $A = V \setminus B$, and counts $n^{i_B}(i_A) = n(i_A, i_B)$ where the i_A is variable and i_B is fixed.

For the purposes of this thesis only three different sampling schemes are considered.

1. All cell counts and the total number of observations are random. This situation appears when counting the number of events in fixed time periods (such as traffic accidents) and classifying them accordingly to type of road, accident severity, day of the week etc. The sampling scheme assumes that the cell counts $\{n(i)\}_{i \in \mathcal{I}}$ are independent and Poisson dis-

tributed. The joint distribution of counts is

$$\Pr[n(i), i \in \mathcal{I}] = \prod_{i \in \mathcal{I}} \frac{m(i)^{n(i)}}{n(i)!} \exp\left(-m(i)\right)$$

where $\mathrm{E}(n(i)) = m(i)$.

2. The total number of observations is fixed but cell counts are otherwise random. This sampling scheme assumes that the observations are independent and the probability that a given observation belongs to the cell i is $p(i) \geq 0$, so the joint distribution is a multinomial distribution

$$\Pr[n(i), i \in \mathcal{I}] = \frac{N!}{\prod_{i \in \mathcal{I}} n(i)!} \prod_{i \in \mathcal{I}} p(i)^{n(i)}.$$

By conditioning upon the total number of observations N the Poisson distribution becomes multinomial.

3. The number of observations $n(i_B) = \sum_{i_A \in \mathcal{I}_A} n^{i_B}(i_A)$ in each i_B-slice is fixed for some $B \subset V$. The sampling scheme is based on the assumption that the counts in the slices are independent and multinomially distributed as in case 2, with cell probabilities in slice i_B equal to $p(i_A|i_B)$. The joint distribution is product-multinomial (also called restricted multinomial)

$$\begin{aligned} \Pr[n(i), i \in \mathcal{I}] &= \prod_{i_B \in \mathcal{I}_B} \left[\frac{n(i_B)!}{\prod_{i_A \in \mathcal{I}_A} n^{i_B}(i_A)!} \prod_{i_A \in \mathcal{I}_A} p(i_A \mid i_B)^{n^{i_B}(i_A)} \right] \\ &= \prod_{i_B \in \mathcal{I}_B} \left[\frac{n(i_B)!}{\prod_{i_A \in \mathcal{I}_A} n(i)!} \prod_{i_A \in \mathcal{I}_A} p(i_A \mid i_B)^{n(i)} \right]. \end{aligned}$$

This distribution can be also obtained from the Poisson distribution by conditioning, in addition to the condition shown for 2, on $n(i_B), i_B \in \mathcal{I}_B$.

The log-linear models are based on expansions for either $\log(Np(i))$ or $\log p(i)$ (which is not much different since $\log N$ is a constant for the multinomial sampling) as a sum of main effects and interaction terms

$$\log p(i) = \sum_{a \subseteq V} u_a^V(i),$$

subject to ANOVA-like constraints to make the expansion unique. The terms u_a^V are called $|a|$-order interaction terms. The first order interaction terms are also called *main effects* and should usually be included in the log-linear models. If the u-terms are written in the form u_a^V, then the subscript (in this case a) shows the subset of variables and the superscript (V in this case) shows the set of variables for which the log-linear model is proposed. For small dimensional tables a more straightforward notation, depending on the context, is used.

For a given log-linear model, denoted for convenience by L, a graph can be associated, called the *interaction graph*, which is an undirected graph with vertices corresponding to the variables in V, and an edge between two vertices (variables) v and w if and only if there is an interaction term $u_{\{v,w\}}^V$ in L. The properties of the interaction graph are studied in Darroch et al. (1980).

A hierarchical log-linear model L is specified by its associated *generating class*. This is defined as the class of subsets a of V, maximal with respect to

inclusion, such that $u_a \neq 0$. The subsets a are called the *generators* of the log-linear model and the model may be specified by enumerating the generators in square brackets. For instance, the log-linear model $A+B+C+A.B+A.C$ will be specified as $[AB][AC]$. Different hierarchical models may have the same interaction graph. The simplest example is given by the models $[ABC]$ and $[AB][AC][BC]$.

The restriction L_a, of a log-linear model L to a set $a \subseteq V$, is a log-linear model for the set of probabilities p_a, whose generating class can be determined from the generating class of L by removing all factors in $a^c = V \setminus a$, the variables in V which are not in a, and then removing the redundant subsets.

Graphical models can be described as a sub-class of hierarchical log-linear models with the maximal permissible higher-order interactions corresponding to a given graph. More formally, a graphical model is a family of probability distributions P_G which satisfies some Markov property over a graph $\mathcal{G}$ (Whittaker, 1990; Lauritzen, 1996). More details about the Markov properties of a family of probability distributions, over a graph, are given below. The decomposable models are graphical models whose interaction graphs contain no cycle of length greater than 3 without a chord (Lauritzen, Speed and Vijayan, 1984). This class of log-linear models is better known in the literature (Haberman, 1974; Bishop et al., 1975; Christensen, 1990; Santner and Duffy, 1989), one reason being that, for decomposable models, maximum likelihood estimators have closed forms.

From the inference point of view, the three sampling distributions are

related and without loss of generality, the main results can be illustrated using only one distribution, like the multinomial distribution (Agresti, 1990). The unknown quantities that are the subject of statistical modelling are the probabilities $p(i)$ of the cells i of the contingency table. The table of counts is a sufficient statistic for the parameters $\mathbf{p} = (p(i))_{i \in \mathcal{I}}$ (Whittaker, 1990).

In this thesis all log-linear models are assumed to be hierarchical so in the model formula only the maximal terms need to be specified. A hierarchical model is based on the assumption that if a lower-order interaction term is missing then all its higher level relatives interaction terms are out of the model. So if any of u_{12}, u_{13}, u_{23} is set to zero then u_{123} should be also set to zero. Graphical models require an extra condition in a somewhat opposite direction. For a graphical model, if all interaction terms of some lower level are included in the model then the higher relative interaction term should be also included. For example, if u_{12}, u_{14}, u_{24} are in the log-linear expansion then u_{124} should also be included. Graphical models are fully interpretable in terms of conditional independencies. In addition, it is worth pointing out that, because the saturated model is graphical, any log-linear model can be nested within a graphical model. This suggests that for any log-linear modelling relative to a contingency table, it may be useful to find first the simplest graphical model fitting the table and then try to refine the analysis.

Some graphical models have already been encountered in Section 3.2. Some other simple models are described now to explain the difference between a graphical and a hierarchical model. With only three variables A, B and C for

simplicity, the saturated log-linear model (which is hierarchical and graphical) is

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}.$$

The log-linear model

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}$$

is hierarchical, but

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

is not hierarchical. The hierarchical model of no three-way interaction

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$$

is not graphical because the inclusion of u_{12}, u_{13} and u_{23} would require the inclusion of u_{123} too.

The interaction graph of a graphical model for categorical variables, is equivalent to the *conditional independence graph*, which is the main tool in graphical modelling (Whittaker, 1990). The conditional independence graph (for short the independence graph) is an undirected graph $\mathcal{G} = (V, E)$ where the set of vertices $V = \{1, 2, \ldots, d\}$ is corresponding to the set of variables

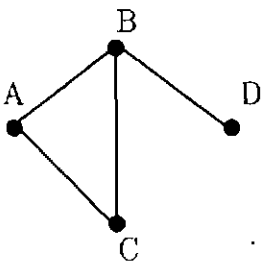


Figure 3.3: Undirected graph $\mathcal{G} = (V, E)$ *where* $V = \{A, B, C, D\}$ *and* $E = \{AB, AC, BC, BD\}$

under study $X_V = \{X_1, \ldots, X_d\}$, and where (i, j) is not in the edge set E if the variables X_i and X_j are independent given the remaining variables $X_{V \setminus \{i,j\}}$. Very often the random quantities are denoted with the labels of their nodes in the graph.

Markov properties on undirected graphs

A probability measure P on Ω has:

(P) the *pairwise Markov* property, relative to $\mathcal{G}$, if for any pair of non-adjacent vertices $a \not\sim b$,

$$a \perp\!\!\!\perp b \mid V \setminus \{a, b\};$$

thus, in Figure 3.3 A is independent of D conditional on B, C;

(L) the *local Markov* property, relative to $\mathcal{G}$, if for any vertex $a \in V$,

$$a \perp\!\!\!\perp (V \setminus \mathrm{cl}(a)) \mid \mathrm{bd}(a);$$

again on the graph in Figure 3.3, D is independent of A, C given $\mathrm{bd}(D) = B$;

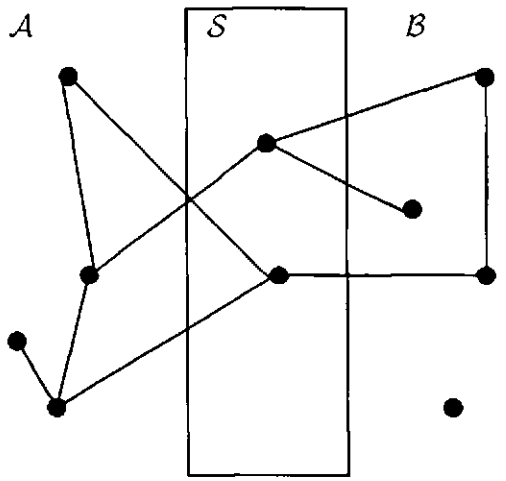


Figure 3.4: The global Markov property

(G) the *global Markov* property, relative to $\mathcal{G}$, if for any triple $(\mathcal{A}, \mathcal{B}, \mathcal{S})$ of disjoint subsets of V such that $\mathcal{S}$ separates $\mathcal{A}$ from $\mathcal{B}$ in $\mathcal{G}$,

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{S}.$$

The global Markov property can be understood as a separation property, see Figure 3.4. If all paths connecting nodes from $\mathcal{A}$ to nodes from $\mathcal{B}$ intersect at least one node from $\mathcal{S}$ then $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{S}$. Because of the general regularity assumptions made, it is true that if $A \perp\!\!\!\perp B \mid C \cup D$ and $A \perp\!\!\!\perp C \mid B \cup D$ then $A \perp\!\!\!\perp (B \cup C) \mid D$ for any disjoint subsets of variables A, B, C, D. The main result regarding these Markov properties is described next.

Theorem 3.1 *If $\mathcal{G}$ is an undirected graph then the global Markov property (G), the local Markov property (L) and the pairwise Markov property (P) are*

all equivalent.

For a proof, see Lauritzen (1996).

The most important property is the global Markov property because it gives a general criterion for deciding when two groups of variables $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given a third group of variables $\mathcal{S}$.

Under the assumption that the joint density $f(V)$ is everywhere positive the local Markov property is also equivalent to the following factorisation of $f(V)$

$$f(V) = \prod_{C \in \mathcal{C}} \psi_C(v_C)$$

where $\mathcal{C}$ is the set of cliques of the graph $\mathcal{G}$ (Lauritzen, 1996). Hence, for the graph in Figure 3.3 the joint density can be factorised as

$$f(V) = \psi_1(A, B, C)\psi_2(D)$$

because the cliques of the graph are $\{A, B, C\}$ and $\{D\}$.

The process of building a graphical model, or equivalently its corresponding conditional independence graph, can be illustrated using the collision-rollover data in Table 3.3. The set of vertices of the graph corresponds to the variables under study. In this case four vertices are needed, that can be denoted again by A, B, C and D. Then an edge is present for each two-way interaction term in the model. This is the same thing as having no edge between two vertices when they are conditionally independent given the remaining set of variables. This pairwise Markov property is used for building the graph which

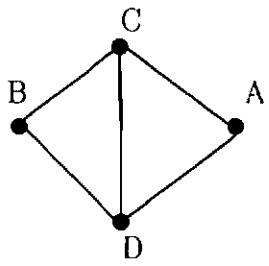


Figure 3.5: Conditional independence graph for collision-rollover data; A is Driver ejected, B is Car type, C is Injury and D is Accident type

is obviously an undirected graph. The independence graph corresponding to the model $[ACD][BCD]$ has the following set of edges: AC, AD, CD, BC, BD and is represented in Figure 3.5.

Because all Markov properties are equivalent, after constructing the conditional independence graph, the independence relationships between the variables can be read directly from the graph, using the global Markov property. The variables B and A are not directly connected on the graph but they are linked via either the variable C or the variable D. This is telling us that B and A are independent given $\{C, D\}$. Although in this case it does not look that global Markov property is more helpful than the pairwise Markov property used to build the graph, when a large number of variables is used, and the final graphical model is proposed as a result of a selection algorithm, the global Markov property (G) is a valuable tool to read the conditional independencies correctly.

3.4.2 Graphical Chain Models

Chain graphs are a combination of directed and undirected edges such that there are no directed cycles in the graph (Section 3.3). They originated in statistical modelling of substantive research hypotheses in the social sciences (Wermuth and Lauritzen, 1990; Cox and Wermuth, 1993). Quite often, the set of the variables under study can be divided into blocks by some prior ordering criterion. The partitioning imposed by the research hypotheses requires naturally that variables in the same block are to be treated on an equal footing, and variables from lower-numbered blocks influence the variables in the blocks with higher order numbers. In a chain graph $\mathcal{G}$, the vertex set V is partitioned into disjoint blocks $V = V(1) \cup \cdots \cup V(T)$ such that the vertices within each $V(t)$ has undirected edges between vertices, and the arrows point from vertices in blocks with lower number to those with higher number. Thus a directed acyclic graph is a chain graph where each block contains only one vertex and an undirected graph is a chain graph with only one block. For $t \leq T$, define $C(t) = V(1) \cup \cdots \cup V(t)$.

Given a particular chain graph $\mathcal{G}$ it is said that a probability P satisfies:

(PB) the *pairwise block-recursive Markov* property if for any pair $a \not\sim b$ it is true that

$$a \perp\!\!\!\perp b \mid C(t^*) \setminus \{a, b\}$$

where t^* is the smallest t that has $a, b \in C(t)$;

(PC) the *pairwise chain Markov* property, if for any pair $a \not\sim b$ with b a

non-descendant of a,

$$a \perp\!\!\!\perp b \mid \mathrm{nd}(a) \setminus \{b\};$$

(LC) the *local chain Markov* property, if for any vertex $a \in V$,

$$a \perp\!\!\!\perp \mathrm{nd}(a) \mid \mathrm{bd}(a);$$

(GC) the *global chain Markov* property if for any triple $(\mathcal{A}, \mathcal{B}, \mathcal{S})$ of disjoint subsets of V such that $\mathcal{S}$ separates $\mathcal{A}$ from $\mathcal{B}$ in $\left(\mathcal{G}_{\mathrm{An}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{S})}\right)^m$, the moral graph of the smallest ancestral set containing $\mathcal{A} \cup \mathcal{B} \cup \mathcal{S}$, it is true that

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{S}.$$

The same chain graph can have attached different dependence chains. The property (PB) is relative to a particular dependence chain. It can be shown (Lauritzen, 1996) that

Theorem 3.2 *For a chain graph $\mathcal{G}$, the global chain Markov property (GC), the local chain Markov property (LC), the pairwise chain Markov property (PC) and the pairwise block-recursive Markov property (PB) are all equivalent.*

A useful practical result is that a chain graph $\mathcal{G}$ possesses the Markov properties of its associated moral graph $\mathcal{G}^m$ (Whittaker, 1990). Frydenberg (1990) shown that (LC) is equivalent to a factorisation of the joint distribution as

$$f(V) = \prod_t f(V(t) \mid \mathrm{pa}[V(t)]), \tag{3.1}$$

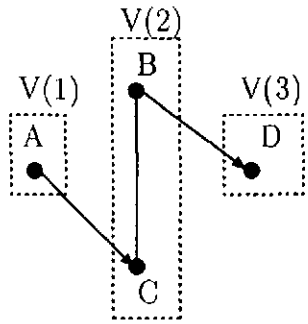


Figure 3.6: Chain graph with the dependence chain $\{A\} \cup \{B, C\} \cup \{D\}$

where $a \in \text{pa}[V(t)]$ if there is a directed link from a to a vertex of $V(t)$. This factorisation is similar to the case of a directed acyclic graph where each block has been considered a single vertex in the directed graph. Moreover, each term in the factorisation (3.1) can be further factorised into

$$f(V(t) \mid \text{pa}[V(t)]) = \prod_{C \in \mathcal{C}_t} \psi_C(v_C), \tag{3.2}$$

where $\mathcal{C}_t$ is the set of cliques of the undirected graph with the set of vertices $(V(t) \cup \text{pa}[V(t)])$, edges consisting of the undirected links between the vertices of $V(t)$, the arrows between $\text{pa}[V(t)]$ and $V(t)$ transformed into undirected lines, and a complete set of lines between the vertices of $\text{pa}[V(t)]$. Thus for the chain graph in Figure 3.6 the following factorisation takes place

$$f(A, B, C, D) = f(D \mid B) f(B, C \mid A) f(A)$$

where

$$f(D \mid B) = \psi_1(B, D)$$

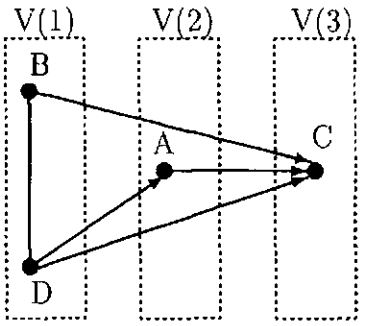


Figure 3.7: Chain graph corresponding to graphical chain model for collision-rollover data, with dependence chain $\{B, D\} \cup \{A\} \cup \{C\}$

and

$$f(B, C \mid A) = \psi_2(B, C)\psi_3(A, C).$$

The variables cross-classifying the collision-rollover data in Table 3.3 can be divided into three blocks: first the car type B and the accident type D, then A driver-ejected and the third block is the injury type C. The chain graph for this graphical chain model is described in Figure 3.7.

Although the research hypotheses are obvious from the chain partitioning, the conditional independencies should be read on the associated moral graph in Figure 3.8. The moral graph is complete so there seem to be no conditional independencies. Apparently this contradicts the conditional independence between car type B and driver ejected A revealed by the conditional independence graph in Figure 3.5. However, the sampling schemes are different. For simple graphical models all variables are treated as response variables in a joint framework, so multinomial sampling is used, whereas for graphical chain models some prior assumptions require the factorisation of the joint dis-

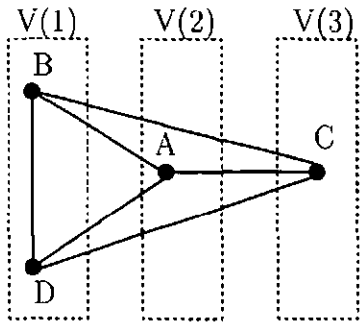


Figure 3.8: Moral graph for the chain graph corresponding to graphical chain model for collision-rollover data, with dependence chain $\{B, D\} \cup \{A\} \cup \{C\}$. *The dependence chain is superimposed for comparison and clarification.*

tribution into conditional distributions according to the block division. The same data was used to exemplify all situations but the modelling problems to be solved are different. Moreover, during the modelling process which is carried out sequentially, it can be noticed that, when just the first two blocks are considered, the arrow from B to A is missing which means that $B \perp\!\!\!\perp A \mid D$. There is nothing wrong with this. If the question is whether $B \perp\!\!\!\perp A \mid D$ in the final chain graph with all three blocks, then the moral graph of the smallest ancestral subset covering $\{B, D, A\}$ needs to be considered. The moral graph $\mathcal{G}^m_{\mathrm{An}(B \cup D \cup A)} = \mathcal{G}^m_{B \cup D \cup A}$ and it is described in Figure 3.9. The conditional independence between B and A given only D is obvious now. The lesson to learn is that the full moral graph can hide some independence relationships.

The graphical model illustrated in Figure 3.5 is different from the graphical chain model with the chain graph in Figure 3.7 in terms of assumptions, fitting process and conclusions implied. The graphical model is based on the assumption that all four variables A, B, C, D are response; the fitting process

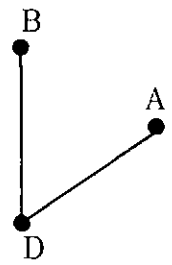


Figure 3.9: Moral subgraph of $\{A, B, D\}$

is based on multinomial sampling and the model fitted uses the factorisation

$$f(A, B, C, D) = f(A, C, D)f(B, C, D). \tag{3.3}$$

The graphical model is the family of probability multinomial distributions satisfying equation (3.3).

On the contrary, the graphical chain model starts by assuming that B, D are pure explanatory variables, A is an intermediate response and C is a pure response. Because of this assumption the joint distribution modelled is not $f(A, B, C, D)$ but $f(C \mid A, B, D)f(A \mid B, D)f(B, D)$, so the product-multinomial distribution is employed. In this case each conditional distribution is fitted to the data separately. The graphical chain model selected is given by

$$f(C \mid A, B, D)f(A \mid D)f(B, D), \tag{3.4}$$

so the data contains statistical evidence of a simplification of only the second factor $f(A \mid B, D)$. The graphical chain model is the family of product-

multinomial distributions satisfying equation (3.4). If another partition of the set variables is chosen, a different graphical chain model may be selected. This highlights the importance of choosing appropriate ordering of the variables in practice. A graphical chain model where car type follows after injury type does not make much sense although the inference process would fit the model to the data and would give some (meaningless) estimates.

More complicated graphical chain models will be investigated in Chapter 5. The process of building the chain graph corresponding to a graphical chain model will be described in detail on an example in Section 5.4. In addition, graphical chain models are mentioned in the context of response variable models in Section 6.2.1.

3.5 Summary

This chapter contained a brief revision through examples of the main concepts from graph theory and probability theory that are needed to understand graphical modelling. The emphasis was on graphical models and graphical chain models because these two classes of models will be applied in the following chapters of this thesis.

Graphical modelling is useful because of the need to analyse large contingency tables. Graphical chain modelling is designed to be applied to situations were some external knowledge is available and the models then become more sophisticated in interpretation.

Although both classes of models are represented by graphs there is a major distinction between them. Graphical chain models are built relative to dependence chains. Great care is needed when interpreting such models because the dependence chain describing the partition plays a major role in extracting the conclusion.

Chapter 4

Inference and model selection

4.1 Introduction

In this chapter the estimation and model selection processes are reviewed. The first section highlights, from an applied perspective, the results on which the whole inference process for graphical log-linear models is based. More details can be found in the standard accounts of Whittaker (1990) and Lauritzen (1996). The second section describes several model selection procedures that can be applied for selecting graphical log-linear models. The problems are explained with the help of the collision-rollover 4-dimensional table. For this particular example, it will be shown that all log-linear models selected by various methods can be nested into the same graphical model. This highlights the idea that graphical modelling can be used to select a small number of models that can be interpreted in terms of conditional independencies and that are good initial models for further analysis.

Estimation and testing procedures are briefly described for graphical models and graphical chain models. Although these two classes are conceptually different, the inferential process for the latter mimics sequentially the fitting and testing process for the former class. The collision-rollover 4-dimensional table used in this thesis as an omnibus example has been analysed in classical textbooks (Fienberg, 1980; Christensen, 1990) in the context of log-linear models but the analysis output and the graphical chain approach presented here are the author's contribution. More complex tables are analysed in Chapter 5.

4.2 Inference

4.2.1 Graphical modelling

Statistical inference can be based on the (scaled) *deviance* (McCullagh and Nelder, 1989) which is a generalised log-likelihood ratio. Denoting the current model by M and the saturated model by M_s, the deviance $\mathrm{dev}(M)$ is twice the difference between the maximised log-likelihood function under the saturated model M_s and the maximised log-likelihood function under the model M:

$$\mathrm{dev}(M) = 2\sum_i n_i \log \frac{n_i}{N\hat{p}^M(i)}.$$

This is the same quantity as G^2 given in equation (2.1) because it is easy to show that $\sum_i (n_i - N\hat{p}^M(i)) = 0$ knowing that $\sum_i n_i = N$. This statistic is asymptotically distributed chi-squared with degrees of freedom equal to the

number of free parameters. Thus the overall deviance can be used as a measure of goodness-of-fit. For testing nested models $M_0 \subset M_1$ the deviance difference $d = \text{dev}(M_0 \mid M_1) = \text{dev}(M_0) - \text{dev}(M_1)$ is appropriate: under the hypothesis that M_0 is true, d has an asymptotic chi-squared distribution with degrees of freedom equal to the difference in the number of free parameters between M_0 and M_1. The asymptotic test based on the deviance difference is more reliable than that based on the overall deviance so it is always better to use the former for model selection. Another generally used measure of goodness-of-fit is the Pearson chi-squared statistic

$$X^2 = \sum_i \frac{[n_i - N\hat{p}_i^M]^2}{N\hat{p}_i^M},$$

having the same asymptotic distribution as the deviance; again this may not provide a reliable test. Both this and the deviance are special cases of the power family of test statistics introduced by Read and Cressie (1988)

$$I^\lambda = \frac{2}{\lambda(\lambda+1)} \sum_i n_i \left[\left(\frac{n_i}{N\hat{p}_i^M} \right)^\lambda - 1 \right]$$

which also covers other well known statistics such as Freeman-Tukey and Neyman; see Bishop et al. (1975). The Pearson X^2 is obtained for $\lambda = 1$ and the deviance is obtained for $\lambda = 0$ by taking the limit of I^λ when $\lambda \to 0$. It has been suggested (Read and Cressie, 1988) that $I^{\frac{2}{3}}$ is more reliable than the more common dev and X^2, especially for sparse tables.

The derivation of the likelihood equations for the maximum likelihood

estimators of the table of probabilities p, satisfying a graphical model given by formula $M = [X_{d_1}][\cdots][X_{d_c}]$, is based on the fact that a set of minimal sufficient statistics is given by the set of marginal tables $\mathbf{n}_a$ corresponding to the generators in the model, that is for all cliques $a = d_1, \ldots, d_c$. Then the maximum likelihood equations are formed by equating the minimal sufficient statistics to their expected values under the model M

$$\mathbf{n}_a = N\widehat{\mathbf{p}}_a^M \tag{4.1}$$

for all cliques $a = d_1, \ldots, d_c$. A proof of this result is given in Whittaker (1990).

If the graphical model M is based on one single conditional independence relationship

$$X_a \perp\!\!\!\perp X_b \mid X_c$$

then there are exactly two cliques in the independence graph, $a \cup c$ and $b \cup c$, and the likelihood equations are

$$N\widehat{p}^M(i_{ac}) = n(i_{ac}) \quad \text{and} \quad N\widehat{p}^M(i_{bc}) = n(i_{bc}). \tag{4.2}$$

The probabilities can then be calculated as

$$\widehat{p}^M(i_{abc}) = \frac{\widehat{p}^M(i_{ac})\widehat{p}^M(i_{bc})}{\widehat{p}^M(i_c)} \tag{4.3}$$

$$= \frac{n(i_{ac})n(i_{bc})}{Nn(i_c)}. \tag{4.4}$$

Denoting by r_a, r_b and r_c the number of cells of the marginal tables given by X_a, X_b and X_c, the deviance of the graphical model $M : X_a \perp\!\!\!\perp X_b \mid X_c$ is

$$\mathrm{dev}(M) = 2\sum_{abc} n(i_{abc}) \log \frac{n(i_{abc})n(i_c)}{n(i_{ac})n(i_{bc})}$$

and it has an asymptotic χ^2 distribution with $r_c(r_a - 1)(r_b - 1)$ degrees of freedom. As an immediate consequence, the deviance for testing the exclusion of only one edge (v, w) in a general independence graph $\mathcal{G}$ is

$$\mathrm{dev}(X_v \perp\!\!\!\perp X_w \mid X_{V\setminus\{v,w\}}) = 2\sum n(i_V) \log \frac{n(i_V)n(i_{V\setminus\{v,w\}})}{n(i_{V\setminus v})n(i_{V\setminus w})}.$$

Decomposable models

The first graphical models investigated were a subclass of log-linear models for contingency tables that have closed-form maximum likelihood estimates (Darroch et al., 1980). Those models were called *decomposable* models because the joint density function can be factorised into the product of marginal density functions on cliques. Recalling the model with the independence graph in Figure 3.5, specified by $[ADC][BCD]$, this is a decomposable model and its joint density function can be calculated from

$$p(A, B, C, D) = p(A, C, D)p(B, C, D).$$

Decomposable models are characterised in terms of graph theory as those that have triangulated (or chordal) independence graphs (Lauritzen et al., 1984).

Being able to calculate maximum likelihood estimators in closed-form is very attractive but numerical methods can overcome this difficulty easily for non-decomposable models. Other reasons why statisticians might restrict their attention to this subclass of graphical models is that exact conditional tests are available only for decomposable models (Lauritzen, 1996) and this is very important for model selection in sparse tables (Kreiner, 1987).

4.2.2 Hypothesis testing

There is specialised software called MIM which was designed for graphical modelling (Edwards, 1995). It includes several methods of model selection and testing. The model selection procedures are discussed in greater detail in Section 4.3 but for grasping a complete view of the graphical modellling from the beginning to the end, MIM's backward elimination procedure is briefly described.

The procedure of backward elimination starts from the saturated model and at each step it removes the edge for which the deviance difference test for edge removal has the largest P-value greater than or equal to a specified significance level α. The edges that are significant (with P-values smaller than α) at one stage of the analysis are not tested again at further stages but always retained in the graph. In the end, when no further edge can be deleted, the corresponding model should fit the data well. Furthermore, the conditional independencies can be read directly from the graph. The backward elimination procedure is usually preferred to a forward inclusion procedure since it is passes

through a sequence of models, all of which fit the data, and the models become simpler at each step.

Exact testing

One aim of this thesis is to apply graphical modelling to large tables. In Chapter 5 tables with 6, 9 and 10 variables are investigated. Even if the sample size is very large, contingency tables summarising road accident data can be expected to be sparse, with many very small cell frequencies. This is due to the nature of road accident data in combination with a large number of cells. For instance, when accident severity is one of the variables, the total number of fatal accidents will be relatively small; when they are distributed across the cells resulting from the cross-classification of the levels of the other (more than 5) variables, many cells are likely to have zero frequencies. This is one important problem that should not be overlooked in modelling accident tables.

The usual methodology employing asymptotic tests for the deviance are then not very reliable. The asymptotic P-values in the case of large sparse tables tend to underestimate the real P-values. Exact tests are required (Kreiner, 1987) to overcome this difficulty and MIM provides options for them. Consider, for instance, a 3-dimensional table of counts. For testing the hypothesis $H_0 : X_1 \perp\!\!\!\perp X_2 \mid X_3$ exact tests are constructed by conditioning on the marginal totals. Denote by Ψ the sample space of all possible 3-dimensional tables $\mathrm{n} = [n_{ijk}]$ with the same fixed margins as the table of observed counts.

Then the P-value for the test criterion $\mathcal{T}$ is

$$P_{obs} = \Pr(\mathcal{T} \geq \mathcal{T}_{obs} \mid H_0) \tag{4.5}$$

$$= \sum_{\mathbf{n} \in \Psi : \mathcal{T}(\mathbf{n}) \geq \mathcal{T}_{obs}} \Pr(\mathbf{n} \mid H_0) \tag{4.6}$$

where

$$\Pr(\mathbf{n} \mid H_0) = \prod_k \frac{\prod_i n_{i+k}! \prod_j n_{+jk}!}{n_{++k}! \prod_i \prod_j n_{ijk}!}. \tag{4.7}$$

This approach is implemented only for decomposable models, so that closed form estimates exist, and it is easily generalised to higher dimensional tables (Whittaker, 1990; Lauritzen, 1996). The *exhaustive enumeration* method, calculating $\mathcal{T}(\mathbf{n})$ and $\Pr(\mathbf{n} \mid H_0)$ for each table $\mathbf{n}$ in Ψ, is not always feasible. The alternative is to use *Monte Carlo sampling.* Following the algorithm in Patefield (1981), K random tables are sampled from Ψ such that the probability of sampling a table $\mathbf{n}$ is from the right distribution. For the table $\mathbf{n}_r$, define z_r to be 1 if $\mathcal{T}(\mathbf{n}_r) \geq \mathcal{T}_{obs}$ and to be 0 otherwise; then estimate P_{obs} by $\widehat{P}_{obs} = \sum_{r=1}^{K} \frac{z_r}{K}$.

4.2.3 Graphical chain modelling

In Chapter 3 it was noted that chain graphs extended graphical modelling to studies where substantive information is available, the variables V being divided into blocks $V(1) \cup \ldots \cup V(T)$, ordered by a prior causal assumption. Great attention should be given to the meaning of the adjective *causal*

(Cox, 1993). The meaning of the concept causal requires both the sense of Suppes underlying Granger-Wiener causality in econometrics and the sense of Rubin (Holland, 1986), and also that there is a substantive process underlying the dependence structure proposed. Subject matter knowledge and theories indicate the type, direction and even the strength of the associations. These hypotheses, describing actual properties of observational units, are called *substantive research hypotheses* or just research hypotheses (Wermuth and Lauritzen, 1990). They are different from statistical null hypotheses which play only the role that they should be rejected by the observed data.

Once again graphs are used to formulate research hypotheses. The specification considers two types of direct association: *directional associations* for pairs of variables where one is a response variable and the other is explanatory; and *symmetric associations* where variables are treated on an equal footing. Changing the direction of some associations would result in changing the research hypotheses.

A graph can serve three purposes: to formulate research hypotheses, to describe conditional independencies, and to characterise a statistical model. In the first case, from subject knowledge, prior to the statistical analysis, the variables under study are divided into several blocks indicated by boxes superimposed on the graph. The boxes are in a one-to-one correspondence with the blocks, of number T say, and they define a *dependence chain* with T *concurrent* sets of variables. A dependence chain can be also defined as an ordered partitioning of the set of vertices V into chain elements such that

edges within chain elements are undirected, and edges between chain elements are directed in the same direction.

In the second case the graph is a mathematical object as used in graph theory. Such a graph is called a chain graph if a dependence chain can be attached to it. Different dependence chains can have the same chain graph which means that they will describe the same conditional independence structure. A cautious approach should be taken when interpreting chain graphs.

The third case relates a chain graph to a statistical model called a graphical chain model by specifying the joint distribution as a product of distributions over the blocks of the dependence chain. Graphical chain models are multivariate response models for $V(t)$ given $V(1) \cup \ldots V(t-1)$. The joint density $f(x_1, x_2, \ldots, x_d)$ can be factorised as

$$f(V(1))f(V(2) \mid V(1)) \ldots f(V(T) \mid V(T-1) \cup V(T-2) \ldots \cup V(1)).$$

The chain graph pictures the conditional independence restrictions on the joint distribution. The case of just two blocks is generic because the inference process is based on fitting two blocks at a time. The first block $V(1)$ is considered a set of covariates $X_1, \ldots, X_p$ and the second block $V(2)$ a set of response variables $Y_1, \ldots, Y_r$. If all $p + r$ variables were responses it can be shown (Whittaker, 1990) that the number of possible models decreases from $2^{\binom{p+r}{2}}$ to $2^{\binom{r}{2}+pr} + 2^{\binom{p}{2}}$, which is an improvement.

To see that note that there are two types of conditional independence rela-

tionships used to build the chain graph; $\binom{r}{2}$ possible conditional independencies between pairs of responses given the remaining responses and all covariates; and pr possible response-covariate pairwise conditional independencies given the remaining responses and remaining covariates. The response-response and covariate-covariate edges are represented by lines, the covariate-response edges by arrows and response-covariate edges are forbidden. For categorical variables, in the class of log-linear models, it was stated in Section 3.4.2 that all chain Markov properties are equivalent and also equivalent with the factorisation of the joint density as given in Equation (3.1). This factorisation implies that the fitting process can be done by focusing on only two blocks at a time The conditional independence structure is then conveyed by combining all $T-1$ conditional independence graphs into a chain graph. The independence relationships can be read using the global Markov property on the associated moral graph, obtained by replacing arrows with lines and by connecting vertices that have connected children in the same block, see Sections 3.3.2 and 3.4.2. Considering just the case of two blocks, the conditional independence graph for a model with the conditional distribution of $V(2) \mid V(1)$ is the same as the conditional independence graph for the model with the joint distribution of $V(1)$ and $V(2)$, having the subgraph corresponding to $V(1)$ complete. Moreover, the graph has the global Markov property with respect to the conditional distribution $f(V(2) \mid V(1))$.

The modelling process can be carried out sequentially. At each step, the current block of variables is considered as response variables and all the pre-

vious blocks are considered explanatory. The conditional model can be fitted in the joint framework making sure that the subgraph of the explanatory variables is complete. The fitting process for a single graphical chain model requires that only $\mathrm{pa}(V(t))$ to be complete, so not necessarily the entire set of explanatory variables. However, for model selection purposes, when all possible models are tested, it is indeed a necessary condition.

Fitting a conditional distribution in a joint distribution is not possible in general but, for contingency tables under multinomial sampling, it is because the multinomial distribution is closed under marginalisation and conditioning. Another example when this is possible is for continuous Gaussian variables, the normal distribution being again closed under marginalisation and conditioning. Thus all the methods of estimating and inference available for graphical models can be used.

In general, two graphical chain models are equivalent if they have identical joint distribution and identical conditional independence structures. A chain graph determines (Frydenberg, 1990) the conditional independence structure and the joint distribution of a graphical chain model. In the same time, substantive research hypotheses based on different chain graphs may have equivalent statistical models. In this case, specific research hypotheses cannot be distinguished just by a statistical analysis of data.

4.3 Model selection

In this section, some model selection procedures for log-linear models, that are used for graphical models as well, are reviewed and compared. The overall deviance can be used as a measure of goodness-of-fit for a given model M. For nested models $M_0 \subset M_1$ it is preferable to have tests on the deviance difference $d = \text{dev}(M_0) - \text{dev}(M_1)$ because it has a better χ^2 approximation and the deviance differences are asymptotically independently distributed when they are components of a single sequence of nested models passing from the maximal to model minimal (Whittaker, 1990). Apart from knowing its asymptotic distribution there are some other advantages for using the deviance as the main tool for statistic inference. Edwards' specialised software MIM for graphical modelling includes several methods of model selection and methods for estimation and testing.

For any log-linear model there is a graphical model such that the log-linear model is nested within the graphical model. Therefore, different model selection procedures for log-linear models can be applied and several models identified. Then, from this set of final models the graphical models can be selected and interpreted in terms of conditional independencies.

The methodology of model selection used below generally follows the stages:

1. Identify some initial models; for example the saturated model is a convenient starting model since it fits the data perfectly. Other initial models can be the main effects model (all variables mutually independent), the

models proposed by Brown's method (Brown, 1976; Christensen, 1990; Whittaker, 1990), models proposed in connection with Aitkin's method (Aitkin, 1979; Christensen, 1990; Santner and Duffy, 1989).

2. From the starting model proposed above use a stepwise model selection (backward, forward or combined) or other method (for example Aitkin's method, Whittaker's method, Edwards and Havranek method) to determine simplified models that fit the data well according to some criteria. The stepwise methods do not necessarily give the best model based on any overall criterion of model fit and they can be very sensitive to the cutoff values used and to the initial model. Consequently, it is better to use several variations and to propose several candidate models.

3. Compare the list of these final models using the Akaike information criterion. This criterion is used for selecting models that maximizes a type of information proposed by Akaike (1973), information that is contained in the statistical model. For log-linear models, in practice this means that the model with the minimum diference between the deviance and twice the number of degrees of freedom is selected.

4. For final models study the residuals, the influential cells and the interpretability of the models.

5. Can a proposed model give some simple answers to some important questions?

There is no doubt that it would be useful to identify a small set of graphical

models, fitting the data well and that can be used for further research.

4.3.1 Aitkin's method

This method is a backward selection procedure. It selects an all j interaction terms model and then it searches all models between all j interaction terms model and all $j-1$ interaction model. This method was designed to control the overall rate for all tests performed using *simultaneous testing*.

Let $M^{(j)}$ denote the model with all possible maximal u-terms of jth order interaction and let d_j be its associated degrees of freedom. Examples for a d-dimensional contingency table are

$$M^{(1)} : \log p_i = u_\emptyset + \sum_{t \in V} u_t,$$

the mutual independence model, and

$$M^{(2)} : \log p_i = u_\emptyset + \sum_{t \in V} u_t + \sum_{j,k \in V} u_j u_k,$$

the all 2-factor effects model.

The initial model for this procedure is the model $M^{(j)}$ that fits the data well while the model $M^{(j-1)}$ does not fit the data. The cutoff points γ_j for $\chi^2(1-\gamma_j, d_{j-1} - d_j)$ should be chosen such that there is a probability no greater than $\gamma \in (0.25, 0.5)$ of rejecting the main effects model when this model is adequate. When complete independence is true the various tests for

j-order interactions are asymptotically independent, so

$$1 - \gamma = \prod_{j=2}^{j=d} (1 - \gamma_j).$$

Aitkin (1979) suggests using $1 - \gamma_j = (1 - \alpha)^{\binom{d}{j}}$ and choosing an α level that yields a $\gamma \in (0.25, 0.5)$. For a 4-dimensional contingency table, choosing

Table 4.1: All j-factors models

all j- factors	Model formula	df	deviance
4	$[ABCD]$	0	0
3	$[ABC][ABD][ACD][BCD]$	1	0.67
2	$[AB][AC][AD][BC][BD][CD]$	5	7.33
1	$[A][B][C][D]$	11	1193.10

$\gamma_4 = 0.05, \gamma_3 = 0.185$ and $\gamma_2 = 0.265$, it is calculated that $\gamma = 0.431$. For the collision-rollover data in Table 3.3, Chapter 3, there are 4 models to be compared, which are described in Table 4.1. Based on calculations in Table 4.2 the model selected is model $M^{(3)}$, given by the largest value j such that

$$\text{dev}(M^{(j-1)}) - \text{dev}(M^{(j)}) > \chi^2(1 - \gamma_j, d_{j-1} - d_j).$$

Table 4.2: Tests for Aitkin's model selection procedure

j-1 vs j	$dev(M^{(j-1)}) - dev(M^{(j)})$	$\chi^2(1 - \gamma_j, d_{j-1} - d_j)$
3 vs 4	0.67 - 0 = 0.67	$\chi^2(.95, 1) = 3.841$
2 vs 3	7.33 - 0.67 = 6.66	$\chi^2(.815, 4) = 6.178$
1 vs 2	1193.10 - 7.33 = 1185.77	$\chi^2(.735, 6) = 7.638$

This overall criterion was criticised by D.R. Cox in the discussion of Aitkin's paper (Aitkin, 1979). Christensen (1990) tried to improve the method by choosing

$$\alpha = \gamma_2 = \ldots = \gamma_k.$$

Following Christensen's idea for $\alpha = 0.1$ it results that $\gamma = 0.271$ and the model $M^{(2)}$ is selected.

Aitkin's model selection procedure continues by examining the models between $M^{(3)}$ and $M^{(2)}$. A model M will be rejected if

$$\mathrm{dev}(M) - \mathrm{dev}(M^{(3)}) > \chi^2(1-\gamma_j, d_{j-1} - d_j) = \chi^2(.815, 4) = 6.178.$$

Using the concept of *coherence* as introduced by Gabriel (1969), the submodels of a rejected model will be definitely rejected too and models which contain an accepted model will be accepted too. This is of great help especially for tables with a large number of variables. The models selected by this procedure are enumerated in Table 4.3. All these are non-graphical log-linear models. If only one model should be proposed then Akaike's information criterion can be

Table 4.3: Models selected by Aitkin's procedure

Model formula	$\mathrm{dev}(M_i \mid M^{(3)})$	$\mathrm{df}(M_i) - \mathrm{df}(M^{(3)})$
$M_1 : [ABC][ABD][CD]$	4.04	2
$M_2 : [ABC][ACD][BD]$	4.86	2
$M_3 : [ABC][BCD][AD]$	4.86	2
$M_4 : [ABD][AC][BC][CD]$	4.51	3
$M_5 : [ACD][AB][BC][BD]$	4.90	3
$M_6 : [BCD][AB][AC][AD]$	4.99	3

used, finding the model M_i for which $\text{dev}(M_i) - \text{df}(M_i)$ is minimum. Using the decomposition of deviance (Whittaker, 1990)

$$\begin{aligned} \text{dev}(M_i) &= \text{dev}(M_i \mid M^{(3)}) + \text{dev}(M^{(3)}) \\ \text{df}(M_i) &= \text{df}(M_i \mid M^{(3)}) + \text{df}(M^{(3)}), \end{aligned}$$

it is easy to calculate the values of Akaike's criterion in Table 4.4. The model

Table 4.4: Akaike's criterion values

Model formula	$\text{dev}(M_i)$	$\text{dev}(M_i) - \text{df}(M_i)$
$M_1 : [ABC][ABD][CD]$	4.04+6.66=10.70	10.70-2=8.70
$M_2 : [ABC][ACD][BD]$	4.86+6.66=11.52	11.52-2=9.52
$M_3 : [ABC][BCD][AD]$	4.86+6.66=11.52	11.52-2=9.52
$M_4 : [ABD][AC][BC][CD]$	4.51+6.66=11.17	11.17-3=8.17
$M_5 : [ACD][AB][BC][BD]$	4.90+6.66=11.56	11.56-3=8.56
$M_6 : [BCD][AB][AC][AD]$	4.99+6.66=11.65	11.65-3=8.65

selected is $M_4 : [ABD][AC][BC][CD]$. Because all the models should have all two-way factors the simplest graphical model that contains this model as a nested submodel is the saturated model, which is not very informative. Therefore nothing can be said about the conditional independencies that might be true.

However, considering the slight alternative proposed by Christensen, the initial model $M^{(2)}$ is selected. There is only one simpler model that fits the data well and this is

$$M_7 : [AC][AD][BC][BD][CD].$$

This model also is not graphical because AC, AD and CD are included in the model but ACD is not. However, it can be nested within a graphical model, the simplest being $M_8 : [ACD][BCD]$. The conditional independence graph of this graphical model is illustrated in Figure 3.5.

4.3.2 Brown's method

This method can be used to determine an initial model. For each term in the saturated model, *marginal association* and *partial association* (Brown, 1976) are tested. For a 3-dimensional table and interaction between variables A and B, a marginal association test compares $[A][B]$ with $[AB]$ and a partial association test compares $[AC][BC]$ with $[AC][BC][AB]$. The extension to larger tables is obvious. The models considered are built considering

1. either all terms for which either the marginal or the partial test is significant

2. or all terms for which both the marginal and partial tests are significant.

The first method gives the largest model and is suitable for backward elimination and the second method gives the smallest model and can be used for forward selection. For collision-rollover data Brown's tests are described in Table 4.5. The stepwise backward selection can start either from

$$[AB][AC][AD][BC][BD][CD]$$

Table 4.5: Marginal and partial association tests

Interaction	Partial Association dev	P-value	Marginal Association dev	P-value
AB	1.69	0.19	8.79	0.00
AC	220.24	0.00	401.69	0.00
AD	114.84	0.00	285.99	0.00
BC	57.48	0.00	52.96	0.00
BD	15.58	0.00	0.38	0.00
CD	441.89	0.00	601.42	0.00
ABC	1.10	0.29	0.07	0.79
ABD	2.92	0.08	1.15	0.28
ACD	2.94	0.08	1.71	0.19
BCD	1.22	0.27	1.44	0.23

(at 0.05 significance level here), or from

$$[ABD][ACD][BC]$$

with calculations made at 0.1 significance level. This procedure will select the final models $[ACD][BC][BD]$ and $[AD][AC][BC][BD][CD]$. None is a graphical model but both are submodels of the graphical model $[ACD][BDC]$. Applying a forward selection ($\alpha = 0.05$ or $\alpha = 0.01$) started from the initial model $[AC][AD][BD][CD]$ leads to the final model $[BCD][AD][AC]$. This is again a non-graphical log-linear model which can be nested into the graphical model $[ACD][BDC]$.

4.3.3 Edwards-Havranek model selection procedure

This is a fast model selection method with the potential to identify a set of simple models all fitting the data well. In this respect it is different from a stepwise model selection which identifies *one* final model. This single model is most of the time used for any inferences, neglecting uncertainty about the model itself, leading to underestimation of measures of uncertainty such as standard errors. It is always good practice (Christensen, 1990) to look at several well-fitting models and the method proposed by Edwards and Havranek (1985) is perfect for this task. It can search through the class of graphical models between a maximal model and a minimal model that can be specified before starting the search. The models are then classified as 'accepted', which means that they fit the data well, or 'rejected'. The coherence principle from Gabriel (1969) is applied, submodels of rejected models being considered rejected and models containing "accepted" models being accepted without further testing. This principle improves the speed of the model selection procedure. At any step, based on the asymptotic χ^2 distribution of the deviance, a model M is accepted if its corresponding P-value is higher than the significance level α. More details are given in Edwards and Havranek (1985) and Edwards (1995).

For the collision-rollover data summarised in Table 3.3, using a 0.05 significance level and searching between the saturated model $[ABCD]$ and the complete independence model $[A][B][C][D]$, the Edwards-Havranek procedure identifies a unique minimal accepted model $[ACD][BCD]$. This is the same model as selected previously and being the only one gives greater confidence

about the conditional independence implied by the model. Moreover, the model has a deviance equal to 5.87 with 4 degrees of freedom providing a very good fit, $P = 0.21$. This set of data is not very complex but the method can be very useful for a higher dimensional contingency table, as illustrated in the next chapter.

It seems that all methods lead to one graphical model $[ACD][BCD]$. Therefore, various relationships can be studied using this model. In Table 4.6 the deviance residuals $\pm\,(2n_i|\log(n_i/m_i)|)^{1/2}$ are given, where $m_i = \hat{\mathrm{E}}(n_i)$ and "-" sign used when $n_i < m_i$. Overall, the fit seems to be good, although

Table 4.6: Deviance residuals for the model $[ACD][BCD]$

Driver Ejected	*Car type*	*Accident type*	*Injury type*	
			Not Severe	Severe
No	Small	Collision	-.22	.04
		Rollover	-.75	-.29
	Standard	Collision	.10	-.01
		Rollover	.50	.15
yes	Small	Collision	.88	-.10
		Rollover	1.55	.35
	Standard	Collision	-.40	.04
		Rollover	-1.18	-.19

simpler models might be more informative.

4.4 Summary

The estimation and model selection framework was highlighted in this chapter. A graphical log-linear model is built in parallel with its corresponding condi-

tional independence graph. Undirected independence graphs have a missing edge for any pairwise independence of two variables conditioned on the rest. This means setting to zero all two-way and higher-order interaction terms containing that pair. Empirically, the interaction terms, u, are estimated by maximizing the appropriate likelihood function, this being a well-developed process for log-linear models that can be done in general in widely known software like SPSS, SAS, S-Plus and GLIM.

Graphical chain models require the same inferential procedures as graphical models. However, their interpretation is made in a different framework, where the variables under study are partitioned by some partial order relationship with possible causal reasoning. The conditional independencies in this case should be read on the moral graph.

Road accident data is usually sparse and therefore asymptotic tests are unreliable. For model selection exact conditional tests should be used and when an exhaustive enumeration is impossible, Monte Carlo sampling provides a feasible solution.

There are many model selection algorithms that have been proposed for the log-linear models and that can be used for graphical models as well. Applying various model selection procedures can be beneficial in providing a set of good models. Edwards-Havranek procedure is very fast and can be used to select more than one model. The collision-rollover data was used here for exemplification but a better example with six variables is given in Chapter 5.

Chapter 5

Applications to road accident data

5.1 Introduction

This chapter contains several applications for road accident characteristics, following the methodology described in earlier chapters. Large tables are investigated for two UK counties, Bedfordshire and Hampshire. These two counties were chosen because they have a relatively small number of records with missing information; both have a large sample size Hampshire having almost four times more records than Bedfordshire, so some comparisons can be made. There are two aims in this chapter. To investigate the relationships between a relatively large number of characteristics and to show, on a particular case, that asymptotic inference may lead to very different results compared with exact conditional inference.

For these two sets of data graphical models will be selected in an exploratory manner and graphical chain models will be proposed in relation with a prior ordering of the variables given by the temporal order of variables related to the accident.

5.2 Bedfordshire data

5.2.1 Graphical model with 6 variables

The data under study consists of all accidents in the STATS 19 database for the county of Bedfordshire in 1995. The data can be summarised in a contingency table cross-classified by the following variables:

- A = Accident severity (fatal, serious, slight),
- L = Light conditions (daylight, darkness),
- N = Number of vehicles involved in the accident (one, two, three or more),
- R = Road surface conditions (dry, wet-damp, snow-ice-frost-flood),
- T = Road Type (major roads, minor roads, where major roads are motorways and A roads, and minor roads are B, C and unclassified roads), and
- S = Speed Limit ($\leq$ 40 mph, $>$ 40 mph).

The author believes these variables to be the six most important road accident characteristics, but the choice of variables does not affect the principle of graphical modelling. There are 1,951 accident records summarised in a 6-dimensional contingency table of order $3 \times 2 \times 3 \times 3 \times 2 \times 2$. The variables are all considered response variables. The conditional independence relationships between the variables can be studied in an exploratory manner, with the aim of finding an initial model that can be investigated further using more sophisticated techniques. The analysis below follows the lines of Tunaru and Jarrett (1998*a*).

The contingency table summarising the data is sparse. For instance, there are 42 fatal accidents spread over $2 \times 3 \times 3 \times 2 \times 2 = 72$ cells, giving an average cell frequency of 0.58 in this part of the table. Model selection procedures based on the asymptotic χ^2 tests are therefore unreliable (Kreiner, 1987). Exact conditional tests using Monte Carlo simulation are implemented to overcome this difficulty. This can be done in MIM (Edwards, 1995) which is an easy and elegant computer platform for graphical modelling. In this case backward elimination, under exact and asymptotic inferential procedures, leads to the model represented in Figure 5.1. That both procedures lead to the same model might be just a coincidence; for other sets of data, as will be shown later, the differences are striking.

The graph of Figure 5.1 can be interpreted as follows. Grouping the variables as $a = \{A\}, b = \{N, S\}$ and $c = \{L, R, T\}$ it is easy to verify the conditions relative to the global Markov property. Directly on the independence

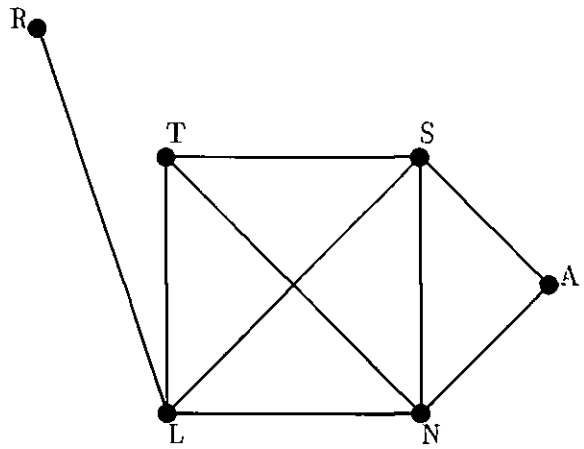


Figure 5.1: The final graphical model for Bedfordshire data with 6 variables; A is accident severity (fatal, serious, slight), N is the number of vehicles involved (1, 2, 3 or more), S is speed limit ($\leq$ 40 mph, $>$ 40 mph), L is lighting conditions (day, night), T is road type (major, minor), and R is road surface (dry, wet-damp, snow-ice)

graph it can be read that, given the number of vehicles N, and the speed limit S, accident severity A is independent of light conditions L, road surface R, and road type T. This is not saying that those three variables are not important regarding accident severity, but conditioning on the fact that an accident has happened, the information provided by those three variables is important for accident severity only as a way of influencing speed limit (which is regarded here as a proxy for the actual speed of the vehicle) and the number of vehicles involved. Thus the important variables for explaining accident severity seem to be speed limit and the number of vehicles.

There are many variables involved in a study of road accident data. Collapsibility (summing over a subset of variables to obtain the marginal table

of the others) breaks large problems into small problems. Looking at the graph in Figure 5.1, let $b = \{L, T, R\}$ and $a = \{A, S, N\}$. The boundary of b is $\{S, N\}$, which is complete, so the model can be collapsed over b, as is proved later in Chapter 6, Section 6.2. This means that the conditional independencies between A, S, N are preserved in the independence graph of any graphical model fitting the marginal table defined by A, S, N. In addition, since the multinomial distribution is closed under marginalisation, the probabilities of this marginal table p_{ASN} can be estimated from the marginal model of $\{A, S, N\}$. In other words, attention may be restricted to the marginal table defined by the variables A, N and S instead of looking at the 6-dimensional table, without introducing problems with Simpson's paradox.

The graphical model presented above suggests that there is a three-way interaction between accident severity, speed limit and the number of vehicles involved, and that studying the marginal three-way table defined by these variables will lead to the same result. This lower dimensional table (the reduction in dimension is from 216 to 18 cells) is more robust to asymptotic tests and it is not sparse as it can be seen from Table 5.1. For this table, the likelihood ratio tests for the three possible conditional independencies are reliable. Therefore the analysis can be further continued on this particular subtable.

The likelihood test for $A \perp\!\!\!\perp S \mid N$ is calculated as

$$\mathrm{dev}(A \perp\!\!\!\perp S \mid N) = 2 \sum_{v=1}^{3} \sum_{s=1}^{2} \sum_{a=1}^{3} n_{asv} \log \frac{n_{asv} n_{++v}}{n_{a+v} n_{+sv}}$$

Table 5.1: 3-way marginal contingency table of road accidents

Accident Severity	*Speed Limit*	*Number of Vehicles* 1	2	3 >
1	1	5	2	1
	2	13	12	9
2	1	77	72	15
	2	39	58	31
3	1	307	640	113
	2	162	271	124

where, for clarity, v is used to index the levels of N; this is equal to

$$\mathrm{dev}(A \perp\!\!\!\perp S \mid N) = \sum_{v=1}^{3} \left\{ 2 \sum_{s=1}^{2} \sum_{a=1}^{3} n_{asv} \log \frac{n_{asv} n_{++v}}{n_{a+v} n_{+sv}} \right\}.$$

If the value of variable N is known, the quantity inside the brackets is the

Table 5.2: Partitioned deviance tests; the P-values are with 3 decimals

Variable	Deviance	df	P-value
$N = 1$	10.41	2	0.005
$N = 2$	28.51	2	0.000
$N = 3$	9.36	2	0.009
Sum	48.28	6	0.000
$S = 1$	24.33	4	0.000
$S = 2$	2.69	4	0.611
Sum	27.07	8	0.000
$A = 1$	1.65	2	0.439
$A = 2$	15.43	2	0.000
$A = 3$	40.83	2	0.000
Sum	57.90	6	0.000

deviance for testing independence between the variables A and S in the sub-table given by $N = v$. In this way the likelihood test can be calculated at each level of the conditioning variable N. Similar calculations and partitions can be made for the other two possible conditional independence hypotheses $A \perp\!\!\!\perp N \mid S$ and $S \perp\!\!\!\perp N \mid A$, as summarised in Table 5.2. The unpartitioned tests are named by the general word "Sum" and it can be remarked that the sums of the partial deviances equal the total deviances and the same for the degrees of freedom. Nevertheless, the situation is not quite the same for P-values, the quantities that are driving the inference process. The tests for the number of vehicles N does not reveal anything new but for speed limit S and accident severity A there are some noticeable exceptions. Although $A \perp\!\!\!\perp N \mid \{S = 1\}$ is strongly rejected by a P-value of 0.0001, the other specified conditional independence hypothesis $A \perp\!\!\!\perp N \mid \{S = 2\}$ cannot be rejected at all and this is in spite of the rejection of the general hypothesis $A \perp\!\!\!\perp N \mid S$. In a similar manner $S \perp\!\!\!\perp N \mid \{A = 1\}$ cannot be rejected because the corresponding P-value is 0.439, although overall $S \perp\!\!\!\perp N \mid A$ has a P-value much smaller than the critical value 0.05.

Consequently, the conditional independence structure revealed by this set of data is more appropriately described by the conditional independence graphs in Figure 5.2. From these graphs it can be easily concluded that, for urban areas, accident severity and the number of vehicles are associated and for rural areas they are not. In addition, for fatal accidents speed limit and the number of vehicles are conditionally independent, the opposite being true for serious

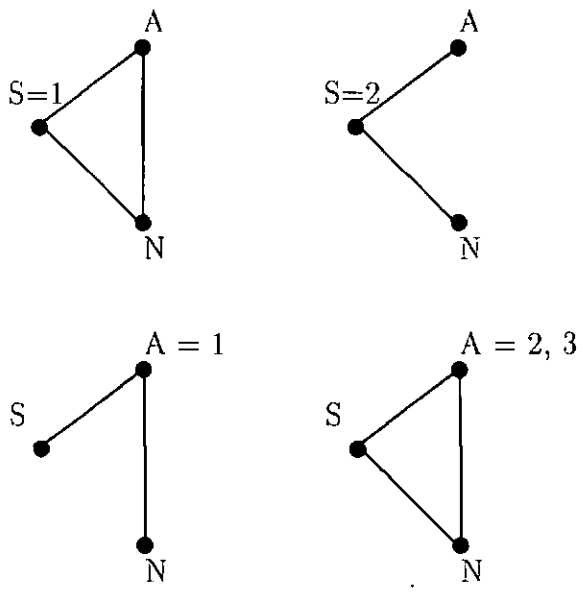


Figure 5.2: Conditional independence graphs revealing a more detailed association structure

or slight accidents.

Edwards-Havranek model selection

As stated in the previous chapter it is better to look at several models instead of basing inference on a single model. The reason for this is that uncertainty in the model may be overlooked and as a consequence parameters of interest be underestimated. The approach proposed by Edwards and Havranek (1985) seeks the simplest models fitting the data well. This searching procedure can screen models between a maximal model known to fit the data well and a minimal model known not to fit the data well; both models are specified in the initialising stage. In MIM, by default the method searches between the saturated model and the complete independence model so all possible graphical

models are eligible for selection. The procedure is very fast. For the same data studied above, at the 0.05 significance level, between the saturated model $[ALNRST]$ and the complete independence model $[A][L][N][R][S][T]$ only 288 models are tested out of $2^{15} = 32768$ possible graphical models which is a great improvement. This happens because once a model M has been accepted all models containing M as a submodel will automatically be accepted without further testing and also, once a model M has been rejected all its submodels are considered rejected too without further testing. This procedure splits the set of possible models in three sets: accepted models, rejected models and non-tested models. The algorithm is testing marginal non-tested models until this set is empty and the minimal accepted models are retained.

For Bedfordshire set of data, the minimal accepted models with the corresponding deviance tests and P-values are given in Table 5.3. The model selected by a stepwise backward elimination procedure using exact conditional tests or approximate asymptotic χ^2 tests, namely $[RL][LTSN][ASN]$, is not included in Table 5.3 because some of its submodels, like the last model $[R][ASN][LST][LSN]$, are listed. If the analyst would like anyway to select a unique model to work with, the Akaike information criterion (Akaike, 1973) is helpful. The idea is to penalise complex models with a large number of parameters and to look for parsimony as recommended by Occam's razor principle. The Akaike information criterion favours the model M with minimum difference between the deviance $\text{dev}(M)$ and the degrees of freedom $\text{df}(M)$. The calculations are made in the last column of Table 5.3.

Table 5.3: Minimal accepted models by Edwards-Havranek procedure

Model M	$\text{dev}(M)$	$\text{df}(M)$	P-value	$\text{dev}(M) - \text{df}(M)$
$[ALR][LRS][LST][LSN]$	211.22	180	0.055	31.22
$[ALR][ALT][LST][LSN]$	210.34	180	0.060	30.34
$[AL][AN][RSN][LST][STN]$	205.21	180	0.095	15.21
$[ALT][LRS][LST][LSN]$	215.48	184	0.056	31.48
$[ALN][LR][LST][LSN]$	209.06	184	0.099	25.06
$[AN][LSTN][RS]$	213.97	182	0.052	31.97
$[R][ALN][LSTN]$	204.75	178	0.083	26.75
$[AL][RS][LSTN]$	211.10	184	0.083	27.10
$[AL][AN][RST][LST][STN]$	216.64	184	0.050	32.64
$[AL][LR][LSTN]$	206.01	184	0.127	22.01
$[AL][AN][LRS][LST][STN]$	211.21	184	0.082	27.21
$[AS][LR][LST][LSN]$	209.81	192	0.180	17.81
$[AS][LR][LST][STN]$	219.68	192	0.083	27.68
$[R][ALS][LST][LSN]$	214.95	190	0.103	24.95
$[AS][LST][LSN][RS]$	214.90	192	0.123	22.90
$[R][AST][LST][LSN]$	222.81	190	0.052	32.81
$[R][AS][LSTN]$	205.50	186	0.156	19.50
$[AL][RSN][LST][LSN]$	216.29	184	0.052	32.29
$[R][ASN][LST][STN]$	210.55	186	0.105	24.55
$[AL][LRN][LST][LSN]$	215.90	184	0.054	31.90
$[ALN][RS][LST][LSN]$	214.15	184	0.063	30.15
$[AS][LST][RS][STN]$	224.77	192	0.053	32.77
$[AS][LST][LSN][RN]$	218.76	190	0.075	28.76
$[R][ASN][LST][LSN]$	200.68	186	0.219	14.68

The last model $[R][ASN][LST][LSN]$ is chosen. This differs from the final model $[RL][LTSN][ASN]$, chosen by other model selection procedures, by having two missing edges RL and TN as can be seen on its independence graph in Figure 5.3. However the main conditional independence relationship $A \perp\!\!\!\perp \{R, T, L\} \mid \{S, N\}$ is still valid and again the model can be collapsed onto A, S, N. Regarding the variables A, S, N there are no differences compared with the model given by the graph in Figure 5.1. The total independence

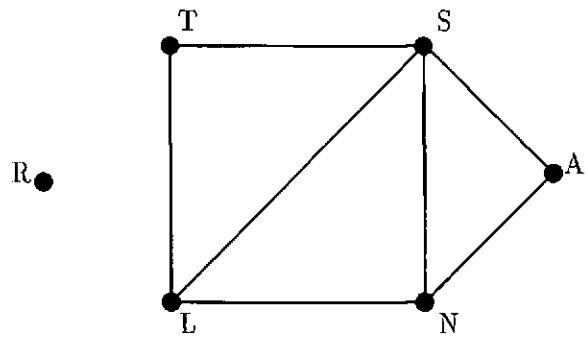


Figure 5.3: Graphical model for Bedfordshire data, chosen by Akaike criterion from the minimal accepted models by Edwards-Havranek model selection procedure

of road surface conditions R as implied by the model in Figure 5.3 seems a bit strong. A possible explanation is that the stepwise backward elimination procedure used in MIM does not test again for removal of the edge RL if it is found significant at one step of the procedure.

This model selection procedure can be used when the aim is to select a subset of models in order to investigate the strength of some relationships between the variables. One major concern is that the testing is done asymptotically. However, decomposable models can be retested using exact conditional tests.

5.2.2 Graphical chain model with 6 variables

In this section, graphical chain modelling is applied to the same six variables investigated in the previous section. This type of analysis has more causal

implications and it will be expanded in the next section to ten variables.

The Bedfordshire data is considered again. The six variables are partitioned into three ordered blocks: $V(1) = \{L, R, S, T\}, V(2) = \{N\}$ and $V(3) = \{A\}$. This partitioning was the author's choice motivated by a temporal argument. Imagine a journey during which an accident happens. Accident severity is decided after the accident takes place, sometimes few days past before an accident can be categorised as fatal or serious. The number of vehicles is established right away at the place of accident and the variables in the first block are known previous to the accident.

The first block contains the variables light conditions, road type, road surface and speed limit and they are considered purely explanatory variables. The independence graph for this block may or may not be of interest. However, it was decided to investigate the conditional independence relationships among the variables in this block. There are two edges missing, between R and T and between R and S. This means that, given daylight conditions L, road surface R is independent of road type T and speed limit S. The first step to build the graphical chain model is to fit the conditional model for the first two blocks. The subgraph defined by L, R, S, T is assumed complete and there is only one missing arrow, between R and N. The next step is to consider accident severity, A, the single variable of the third block, as a response and to keep fixed the complete subgraph defined by all variables in the first two blocks. There are three arrows missing, between R and A, between L and A and between T and A. The chain graph is described in Figure 5.4. The

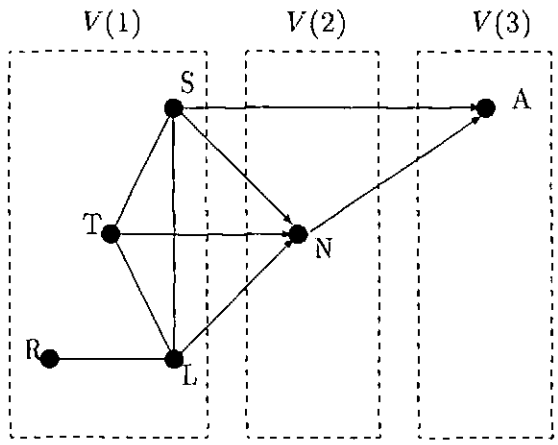


Figure 5.4: Graphical chain model for Bedfordshire data with the dependence chain $\{R, L, T, S\} \cup \{N\} \cup \{A\}$

sequential process of building a graphical chain model is described in greater detail in Section 5.4.

This model has an obvious causal interpretation. The speed limit, road type and daylight conditions all influence directly the number of vehicles involved in the accident. Road surface has no direct influence to the number of vehicles but acts only through its association with daylight conditions. Finally, accident severity is influenced only by the speed limit and the number of vehicles. The conditional independence relationships can be read on the moral graph of the chain graph in Figure 5.4. The moral graph in this case is obtained by replacing the directed edges with undirected edges. So $A \perp\!\!\!\perp \{L, R, T\} \mid \{N, S\}$ which means that accident severity is independent of daylight conditions, road type and road surface given the number of vehicles and the speed limit. This is the same conclusion as before. Using these conditional independencies, the model is given by the following factorisation of the

joint density function

$$\begin{aligned} f(a,l,n,r,s,t) &= f(l,r,s,t)f(n \mid l,r,t,s)f(a \mid l,r,t,s,n) \\ f(a,l,n,r,s,t) &= f(l,r,s,t)f(n \mid l,t,s)f(a \mid s,n) \\ f(a,l,n,r,s,t) &= \frac{f(l,r)f(l,s,t)}{f(l)}f(n \mid l,t,s)f(a \mid s,n) \end{aligned}$$

where the last factorisation is of less interest than the first two.

5.2.3 Graphical chain model with 10 variables

It is possible to consider a larger number of variables. The table will then be more sparse and using exact conditional methods becomes essential. For the same county Bedfordshire, another four variables, regarding time characteristics, location characteristics and accident characteristics, are considered:

- C = Number of casualties in the accident (1, 2, 3 or more),
- D = Day of the week (Sunday, Monday-to-Thursday, Friday, Saturday),
- H = Hour of the accident (0-6, 7-9, 10-14, 15-18, 19-23),
- P = Pedestrian crossing within 50m of the place of the accident (no, yes).

It seems more appropriate not to consider all 10 variables in a symmetric way. The possible history of the accident provides a clue about how the variables can be partitioned into recursive blocks. Consider the first block of variables $\{D, H, L, P, R, S, T\}$; the reason for choosing these variables is

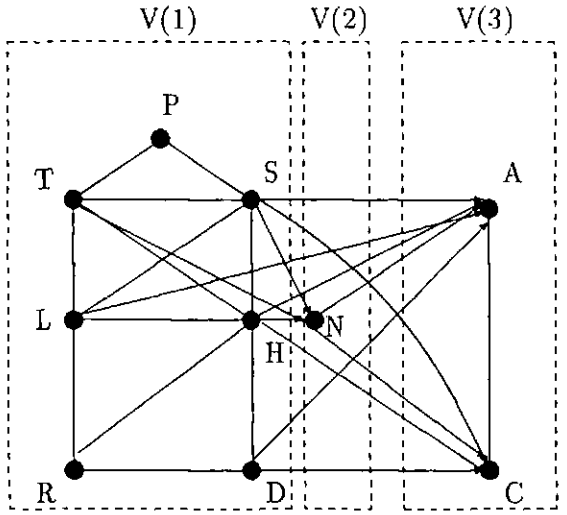


Figure 5.5: Graphical chain model for Bedfordshire data with 10 variables, A is accident severity, C is the number of casualties, N is the number of vehicles, S is speed limit, H is hour of the day, D is day of the week, P is presence of a pedestrian crossing, T is road type, L is daylight conditions and R is road surface conditions

that their values related to a site of a road network are established well in advance of the occurrence of the accident. The number of vehicles is the only variable in the second block and the last block contains accident severity and the number of casualties $\{A, C\}$. The values of these last two variables can be known only after the accident happens. Backward elimination, using exact conditional tests leads to he chain graph of Figure 5.5. The graphical model for the first block of variables may be of interest or not, but directly from the graph it can be seen that

$$P \perp\!\!\!\perp \{D, H, L, R\} \mid \{S, T\}$$

$$\{S, T\} \perp\!\!\!\perp \{R, D\} \mid \{H, L\}.$$

Modelling the number of vehicles, N, as a response variable, it can be seen directly on the chain graph that

$$N \perp\!\!\!\perp \{D, L, P, R\} \mid \{H, S, T\}$$

and for the accident severity, A and the number of casualties, C,

$$\{A, C\} \perp\!\!\!\perp \{P, R, T\} \mid \{D, H, L, N, S\}$$

$$C \perp\!\!\!\perp \{L, P, R, T\} \mid \{A, D, H, N, S\}.$$

These relationships can help us understand what variables influence either the accident severity or other related variables of interest such as the number of vehicles and the number of casualties in the accident. The number of vehicles is independent of daylight conditions, day of the week, road surface and presence of pedestrian crossing given hour of the day, speed limit and road type. Accident severity and the number of casualties are influenced directly only by day of the week, hour of the day, daylight conditions, the number of vehicles and speed limit. The direct association between accident severity and the number of casualties suggests that, when data is disaggregated by these two variables, the analysis should consider modelling multiple accident frequencies jointly. This idea is followed in the second part of the thesis in Chapters 8 and 9.

5.3 Bedfordshire and Hampshire data

5.3.1 Graphical models for the Hampshire data and comparisons

In this section several graphical models are investigated for the Hampshire data and a comparison is made with the models obtained for the Bedfordshire data.

After deleting 68 observations having missing variables, the data corresponding to Hampshire for 1995 contains 7242 accident records, a much greater number than that for Bedfordshire. Starting with the same 6 variables symmetrically treated, the Edwards-Havranek model selection procedure searching between the saturated model $[ALNRST]$ and the complete independence model $[A][L][N][R][S][T]$, tested just 24 models out of 2^{15} possible models. This procedure was used because the 6-dimensional table is not so sparse, having a cell frequency average of 2.76 for fatal accidents. It was aimed to select some models for comparison purposes. Only two minimal models, consistent with the data, are proposed. The first one is $[ASN][LST][RSTN]$ having a deviance equal to 182.06 with 164 degrees of freedom, which has the independence graph in Figure 5.6. The second one is $[ASN][LST][LSN][RST][RSN]$ having a deviance of 173.91 with 172 degrees of freedom. The independence graph for the second model is showed in Figure 5.7. As opposed to the first model, the second model is not decomposable because of the chordless 4-cycle $R - T - L - N$. This means that the estimates have to be calculated by

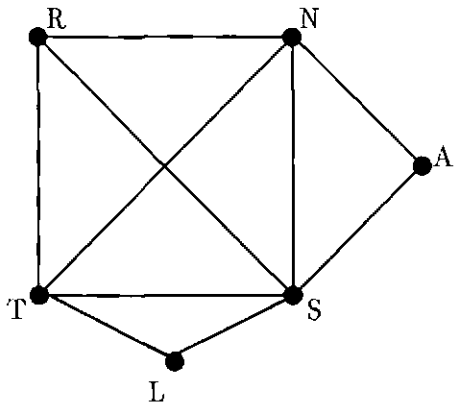


Figure 5.6: Graphical model for Hampshire data with 6 variables, where A is accident severity (fatal, serious, slight), N is the number of vehicles involved (1, 2, 3 or more), S is speed limit ($\leq$ 40 mph, $>$ 40 mph), L is lighting conditions (day, night), T is road type (major, minor), and R is road surface (dry, wet-damp, snow-ice)

iterative methods. Anyway, both models still support the main conditional independence relationship identified in the case of Bedfordshire county, which is

$$A \perp\!\!\!\perp \{R, T, L\} \mid \{S, N\}. \tag{5.1}$$

For both counties, the independence relationship (5.1) is true. It is worth pointing out that this does not necessarily imply that this will be also true for the pooled set of data, combining the accidents from Bedfordshire with the accidents from Hampshire. It could be just another instance of Simpson's paradox. The most general model under which the conditional independence (5.1) can be tested is $[ANS][RTLNS]$. For Bedfordshire and Hampshire combined, the deviance of this model is 178.43 with 132 degrees of freedom, giving a P-

value of 0.004. Therefore, after pooling data for Bedfordshire and Hampshire into one table, it is not true anymore that accident severity is independent of road surface conditions, road type and daylight conditions given the values of speed limit and the number of vehicles involved.

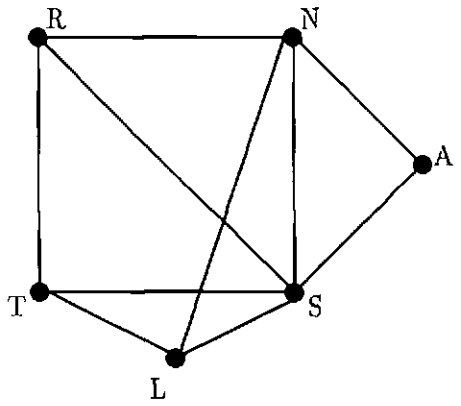


Figure 5.7: A graphical non-decomposable model for Hampshire data with 6 variables, where A is accident severity (fatal, serious, slight), N is the number of vehicles involved (1, 2, 3 or more), S is speed limit ($\leq$ 40 mph, $>$ 40 mph), L is lighting conditions (day, night), T is road type (major, minor), and R is road surface (dry, wet-damp, snow-ice)

This may happen because the two counties have different geographical conditions, different socio-economic characteristics, different percentages of roads of some type and so on. The two sets of data are observational studies from different populations.

5.3.2 Graphical chain model with 10 variables

As revealed in the previous section, an interesting question is what happens when more data is collected. It may be thought that there is no need for exact conditional tests and Monte Carlo methods as there are data available for other counties as well and by pooling the data, the contingency table will cross-classify a larger and larger number of cases keeping fixed the number of cells. However this is not the case. Considering the data from STATS 19 for 1995, for Bedfordshire and Hampshire, cross-classified by the same 10 variables as before, the resulting table is still sparse in spite of the large sample size of 9193 accidents. This is due to the nature of the data and it has nothing to do with the sampling method. The table is expected to have small frequencies in the cells corresponding to fatal accidents and large numbers in the cells corresponding to slight accidents, for example. Applying the same methodology as before the chain graphical model in Figure 5.8 is obtained.

There are some interesting causal relationships revealed by the chain graph. The presence of a pedestrian crossing, P, does not affect the number of vehicles, N, the accident severity, A, or the number of casualties, C. The day of the week, D, influences directly the number of vehicles, the accident severity and the number of casualties. The accident severity and the number of casualties are directly connected, suggesting that a multivariate regression model may be more appropriate than ordinary regression models.

Following the modelling process step by step, it can be informative to describe the conditional independence relationships. From the chain graph it

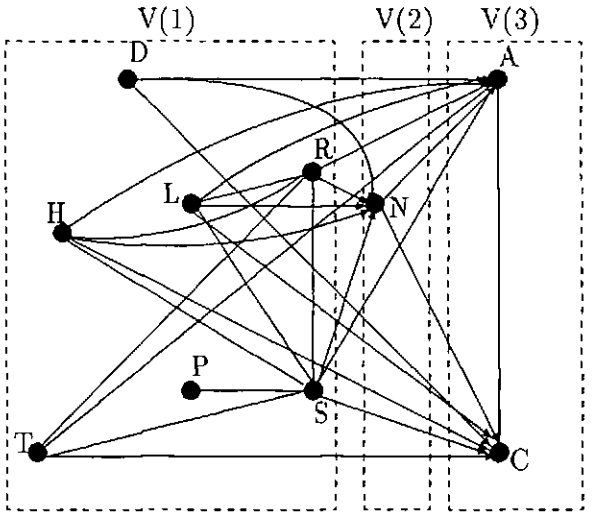


Figure 5.8: Graphical chain model for Bedfordshire + Hampshire data

is easy to see that

$$N \perp\!\!\!\perp \{P, T\} \mid \{D, H, L, R, S\}$$

and secondly

$$C \perp\!\!\!\perp \{P, R\} \mid \{D, H, L, N, S, T\}$$

$$\{A, C\} \perp\!\!\!\perp P \mid \{D, H, L, N, R, S, T\}.$$

These conditional independence relationships suggest that safety measures, aiming at a reduction in accident severity and the number of casualties, should not consider primarily the presence of pedestrian crossings. The variables in the conditioning set are those that should be targeted because they influence

directly the variables of interest, accident severity and the number of casualties. The above conditional independence relationships reveals that, knowing that the accidents have occurred, we expect that the presence of pedestrian crossing to be irrelevant regarding accident severity and the number of casualties, from the *statistical* information point of view. This does not mean that pedestrian crossings are useless. They are designed for reducing pedestrian casualties. A more detailed analysis in the next section, only for accidents with pedestrian casualties, reveals that the presence of pedestrian crossing is directly influencing the number of casualties in such accidents but not the accident severity. Other road characteristics and accident characteristics contribute to accident severity.

5.4 Graphical chain modelling at a disaggregated level

5.4.1 Accidents with pedestrian casualties

The accidents where there is a pedestrian casualty might have different contributory factors from those with no pedestrian casualties. For this reason it seems advisable to analyse separately the two classes of accidents. Table 5.4 and Tables B.1, B.2, B.3 in the Appendix B contain the results needed to build the graphical chain models for Bedfordshire only and for Bedfordshire and Hampshire pooled together, at critical levels $\alpha = 0.05$ and $\alpha = 0.01$. For

comparison the results obtained using decomposable model selection, unrestricted model selection and exact Monte Carlo sampling model selection will be given. Unrestricted models selection means that all graphical decomposable and non-decomposable models are searched. Before doing so, it is helpful

Table 5.4: Bedfordshire 1995 ; $\alpha = 0.05$

Variables	Model formula	Method
D, H, T	$[DH][HT]$	Dec.
	$[DH][HT]$	Unres.
	$[T][DH]$	Exact.
$L, R, S \mid D, H, T$	$[RS][DLST][DHLT]$	Dec.
	$[RS][HR][LS][HL][DHT][DS]$	Unres.
	$[RS][HLST][DHST]$	Exact.
$P, N \mid L, R, S, D, H, T$	$[NRS][DHPRT][DHLRST]$	Dec.
	$[NS][PST][DHLRST]$	Unres.
	$[NS][PRT][DHLRST]$	Exact
$A, C \mid P, N, L, R, S, D, H, T$	$[ADHNPRT][ACHNPR][DHLNPRST]$	Dec.
	$[AST][CS][DHLNPRST]$	Unres.
	$[ART][CHNP][DHLNPRST]$	Exact

to explain the building process of a chain graph using exact testing. The results in Table 5.4 contain all the necessary information. The choice of blocks of variables was based on the same principles as before. However, the previous set of explanatory variables was further divided into a block of temporal variables, day of the week, hour of the day together with road type, which are some sort of fixed variables, and a block of environmental variables: daylight conditions, road surface conditions and speed limit. Speed limit is included in the second block because it may change from time to time. The third block

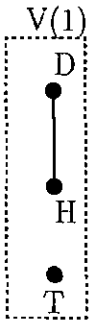


Figure 5.9: Initial step of building the chain graph for accident data with pedestrian casualties in Bedfordshire, 1995; D is day of the week, H is hour of the week and T is road type

includes presence of pedestrian crossing and the number of vehicles as factors that influence directly the number of casualties in the accident. Accident severity and the number of casualites are known only after the accident takes place.

First the initial set of variables $\{D, H, T\}$ is investigated and conditional independencies between these three variables, in the marginal table defined by them, are revealed in the graph of Figure 5.9. This step is not really necessary and can be skipped. The sequential process is modelling just two sets of variables at a time, one explanatory and one response.

The next set of variables to be considered is $\{L, R, S\}$. The edges between $\{D, H, T\}$ are not relevant and they can be left out of the graph. The two blocks are delimited in Figure 5.10 by dash boxes. As described in Chapter 3, there are arrows pointing towards the variables in the new block and undirected lines between the variables inside this block. From Table 5.4 it can be seen that there is only one line between R and S and 5 arrows, 3 pointing

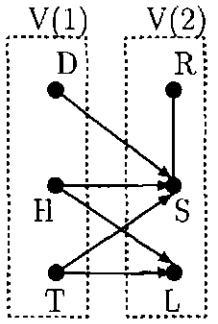


Figure 5.10: First step of building the chain graph for accident data with pedestrian casualties in Bedfordshire, 1995; R is road surface, S is the speed limit and L is lighting conditions

towards S (out of 3 possible) and 2 pointing towards L (out of 3 possible).

The second step consists in considering all the variables in the first two blocks as one single explanatory set, so therefore a single block, and the third block, in order, of variables, that is $\{P, N\}$, takes the place of the response variables block. Again there are two types of edges; arrows pointing towards P or N and a possible line between P and N. The graph at this intermediary stage is presented in Figure 5.11 and is based on the inferential results from Table 5.4.

The last step, the third, brings the last set of variables $\{A, C\}$ as the response block and all the previously investigated variables are playing the role of explanatory variables as in Figure 5.12. From Table 5.4, there is no line between A and C and there are 5 arrows between R, T and A, and between H, N, P and C.

The intermediary graphs look quite simple, revealing simple association structures. Now the chain graph, with the associated dependence chain, can be

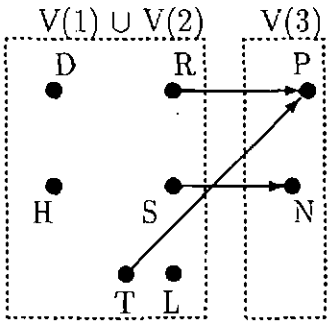


Figure 5.11: Second step of building the chain graph for accident data with pedestrian casualties in Bedfordshire, 1995; P is the presence of pedestrian crossing within 50 m and N is the number of vehicles involved

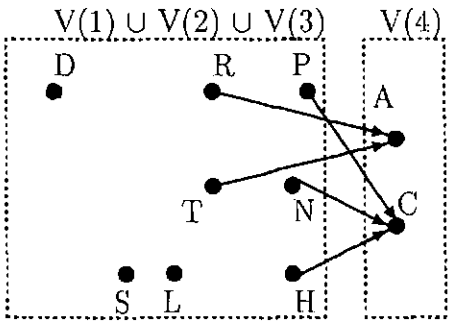


Figure 5.12: Third step of building chain graph for accident data with pedestrian casualties in Bedfordshire, 1995; A is accident severity and C is the number of casualties

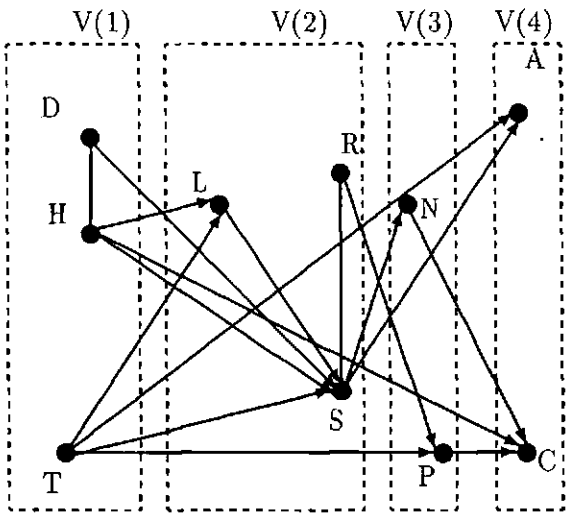


Figure 5.13: Graphical chain model for Bedfordshire data; accidents with pedestrian casualties only

drawn putting all the previous steps together. The graphical chain model has the chain graph in Figure 5.13. Although this graph looks a bit complicated, the actual sequential building process shows the opposite. However, great care should be taken when reading the conditional independencies. The moral graph has to be used, replacing arrows by lines and connecting vertices that have common children. For example N and P should be connected by a line in the moral graph because both have C as their child. It can be seen from the chain graph in Figure 5.13 that accident severity and the number of casualties are not associated, that speed limit is a very important variable absorbing the information from a group of other variables like day of the week, hour of the day, road type, daylight conditions and road surface; that accident severity is directly influenced only by speed limit and road type; that the number of casualties is directly influenced only by the hour of the day, the number of

vehicles involved and the presence of pedestrian crossing.

However, the conclusions are slightly different from a similar previous analysis considering all accidents, with or without pedestrian casualties, pooled together. From the author's point of view, the results found here do make sense. The accident severity is affected by the speed limit and the type of the road where accident occurred. Speed limit is also influenced by the type of the road, as characteristics of accidents with pedestrian casualties, which again is sensible, but speed limit is not enough to explain accident severity, otherwise there would be no arrow from road type T to accident severity A.

It is evident from Table 5.4 that the results are quite different for the other methods, decomposable or unrestricted. This means that some false inference can be made when asymptotic rather than exact conditional methods are used. Since large accident tables are very often sparse it is better to base the inference on exact conditional testing. A drawback of this method is that the selected models are always decomposable so simpler non-decomposable graphical models are not even tested with this approach. In the author's opinion it is better to have a reliable model rather than a simple unreliable one.

5.4.2 Accidents without pedestrian casualties

This section contains the complementary analysis for accidents without pedestrian casualties. For this type of accidents, the presence of pedestrian crossing was considered to have no importance and it was removed. Although the de-

pendence chain is very similar, it will not be surprising if the selected graphical chain models will be different from those discussed earlier for accidents with pedestrian casualties. One major change revealed here is that accident severity and the number of casualties are directly associated.

Table 5.5: Bedfordshire 1995; $\alpha = 0.01$

Variables	Model formula	Method
D, H, T	$[HT][DH]$	Dec.
	$[HT][DH]$	Unres.
	$[HT][DH]$	Exact.
$L, R, S \mid D, H, T$	$[LST][HLT][HLR][DHT]$	Dec.
	$[LST][HLT][HLR][DHT]$	Unres.
	$[LST][HLT][HR][DHT]$	Exact.
$N \mid L, R, S, D, H, T$	$[DHNRST][DHLRST]$	Dec.
	$[HNST][DHLRST]$	Unres.
	$[HNST][DHLRST]$	Exact
$A, C \mid N, L, R, S, D, H, T$	$[ACDHLNRST]$	Dec.
	$[AS][CNS][DHLNRST]$	Unres.
	$[ACNS][ADHNS][DHLNRST]$	Exact

The Tables 5.5, and B.4, B.5 in the appendix B are for accidents without pedestrian casualties. It can be easily seen that exact inferential methods provide different results than asymptotic inferential methods. In addition, there are differences between the graphical chain models for accidents with pedestrian casualties and the graphical chain models for accidents without pedestrian casualties. However this is not a surprise. The analysis at the more disaggregated level is more fragile because of the sparse character of the contingency tables. When there is a particular interest in one type of

the accidents, like accidents with pedestrian casualties, this difficulty can be overcome by collecting more data over a larger period of time or over a larger spatial area.

5.5 Summary

Graphical chain models provide a useful exploratory technique for disentangling the potential factors which influence variables such as accident severity or the number of casualties. However, some care needs to be taken in the choice of statistical test used to select a well fitting model. Using the same 10 variables, the graphical chain models for Bedfordshire, and for Bedfordshire and Hampshire together, are different. This is not surprising since the second model was based on more data. It was pointed out that for Bedfordshire data alone, when just six variables are used, the graphical chain models obtained using different methods of testing and model selection are the same. For the 10-variables table, different final models are obtained if asymptotic (chi-squared) methods of testing are used instead of the exact Monte-Carlo method used here. As the contingency tables becomes larger and more sparse, the classical tests are not reliable and the use of exact tests and Monte Carlo simulation procedures become essential.

Graphical modelling and graphical chain modelling provide a sound alternative for investigating a large number of road accident characteristics at an aggregated level and at a more specific level of aggregation. In addition, there

is strong empirical evidence that, for large sparse tables, asymptotic methods and exact conditional methods give very different results, the second type of inference being more reliable.

Chapter 6

Collapsibility in contingency tables

6.1 Introduction

This chapter aims to show how data analysis can be reduced in dimensionality, in a reliable manner, and questions of particular interest can be answered using other statistical tools following the results of graphical modelling. Collapsibility was briefly used in Chapter 5 for continuing the analysis in a marginal table of interest. There are different concepts of *collapsibility* defined in the literature (Bishop et al., 1975; Whittemore, 1978; Asmussen and Edwards, 1983; Davis, 1986), and although there are some equivalence results (Davis, 1986), the collapsibility concept used here concerns the presence or not of interactions terms in the log-linear expansion. This can be called *model collapsibility* but being the only collapsibility type investigated in this thesis

it will be simply called collapsibility.

For statistical modelling, the more parameters a log-linear model has the better is the fit to the data. The saturated model has one parameter for each data value, so it will fit the data perfectly. However, the saturated model cannot be used for prediction because for another sample from the same population the results will be different. The statistician is confronted with a dillema. One tendency is to put more parameters into the model to explain the complexity of the data. The other is to have less parameters because they are more efficiently estimated, Altham (1984), and the model is more easily interpreted. The solution is collapsibility, which breaks large problems down into small problems. It is very useful to know when lower dimensional marginal tables can be analysed instead of very large high-dimensional tables.

6.1.1 Simpson's Paradox

This phenomenon has been described in many classical textbooks like Bishop et al. (1975), Edwards (1995), Whittaker (1990), which show that collapsing tables can lead to misleading conclusions. This phenomenon is not just of academic interest. A set of examples from the real world is presented by Wagner (1982). An example of Simpson's paradox in the context of road accident data was discussed in Section 3.2.

Simpson's paradox is the result of collapsing a contingency table that should not be collapsed. Possibly the confusion starts with the analogy between log-linear models and ANOVA models. For a three-factor ANOVA

model, when there is no three-way factor interaction, all of the two-factor interactions can be examined from the corresponding two-factor marginal table. On the contrary, for tables of counts, for a log-linear model that has no three-way interaction but all two-factor interactions, it is not correct to draw conclusions about two-factor interactions from the two-factor marginal tables.

Simpson's paradox appears when the complex analysis of large tables is unwisely replaced by a series of investigations of marginal small dimensional tables. Some studies that can be criticised on this ground are Henson (1992) and Taylor and Barker (1994-1995). In analysing large tables there is one last obstacle that needs to be overcome. The tables may be sparse and the asymptotic tests are unreliable. As it was shown in Chapters 4 and 5 exact conditional tests with Monte Carlo sampling can be extremely helpful in such situations.

6.2 Collapsibility

Asmussen and Edwards (1983) introduced a definition of collapsibility based on the relationship between maximum likelihood estimators computed on the joint and marginal tables of counts $\mathbf{n}$.

Definition 6.1 *The hierarchical log-linear model L is collapsible onto the subset of variables a if one of the following equivalent conditions hold:*

1. *for all $p(i) \in L$, it is true that $p(i_a) \in L_a$*

2. *for all i_a, $\widehat{p}(i_a) = \widehat{p}_a(i_a)$.*

The hat denotes the maximum likelihood estimator and p_a is the vector of probabilities under model L_a. The next theorem and its corollary, proved by Asmussen and Edwards (1983), are possibly the most important properties of collapsibility for contingency tables:

Theorem 6.1 (Asmussen and Edwards) *A hierarchical log-linear model L is collapsible onto the subset of variables a if and only if the boundary of every connected component of a^c is contained in a generator of L.*

Corollary 6.1 *If L is a graphical model, the condition in Theorem 6.1 means that the boundary of every connected component of a^c is complete and L is said graphically collapsible onto a.*

The collapsibility as presented above is based on the idea that, for log-linear models, the presence or not of the interaction terms is important, and not the exact values of the log-linear parameters.

The graphical model, proposed for Bedfordshire data following Edwards-Havranek model selection procedure and having the independence graph in Figure 6.1, is not collapsible onto $a = \{T, N, A\}$ because the connected components of $a^c = \{R, L, S\}$ are $\{R\}$ and $\{S, L\}$ and their boundaries are

$$\mathrm{bd}\{R\} = \{\emptyset\} \quad \mathrm{bd}\{L, S\} = \{T, N, A\}$$

and although the empty boundary means that it is possible to collapse, the second boundary is properly incomplete and this means that the graphical model is not collapsible onto $\{T, N, A\}$. More generally, if the variables under

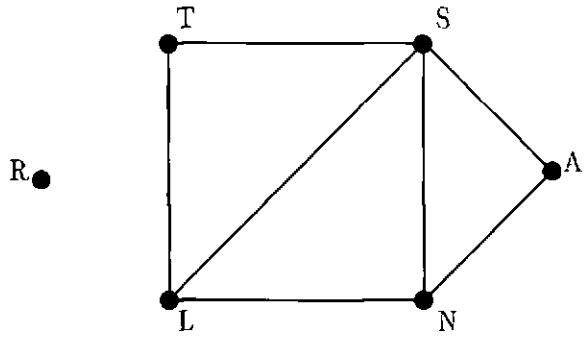


Figure 6.1: Graphical model for Bedfordshire data: A is accident severity, S is speed limit, N is the number of vehicles involved, T is road type, L is lighting conditions, R is road surface

study X_V are partitioned into (X_a, X_b), knowing the independence graph of X_V, what can be said about the independence graph of X_a? This question has an answer in the concept of *graphical collapsibility* as defined in Corollary 6.1. The important result, (Whittaker, 1990), is that, if $X_V = (X_a, X_b)$ is graphically collapsible onto X_a, then the conditional independencies between the variables of X_a, in the independence graph of (X_a, X_b), are preserved in the independence graph of X_a. Again using the graphical model illustrated in Figure 6.1, for the partition $a = \{A, S, N\}$ and $b = \{R, L, T\}$, it can be seen that $\mathrm{bd}(L, T) = \{S, N\}$ which is complete, and so the model is collapsible onto a and the three-way interaction between accident severity, speed limit and the number of vehicles is preserved in the model for the 3-dimensional marginal contingency table defined by these three variables. This means that

the only simpler hierarchical log-linear model that could fit this 3-way table is the model of no three-way interaction $[AS][AN][SN]$, which is not graphical.

Another question of interest concerning collapsibility is whether the predicted distribution, calculated by marginalising the fitted model of the joint distribution, can be recovered by modelling the marginal data. This is a

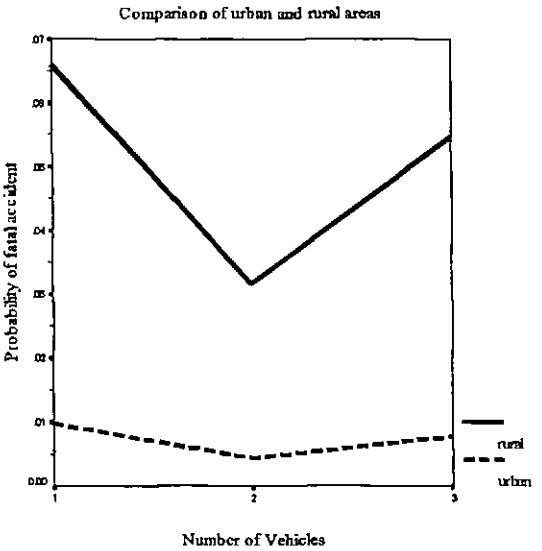


Figure 6.2: Probabilities that an accident on urban and rural roads in Bedfordshire is fatal

question of commutativity of fitting and marginalisation, which means that a model $\widehat{f_{ab}}$ for the joint distribution f_{ab} can be fitted first and then one can marginalise the fitted model to $\widehat{f_a}$ or marginalise first the joint distribution and then fit the marginal distribution f_a and get the same result $\widehat{f_a}$. Collapsibility in this sense means that the fitted cell probabilities are the same irrespective of the order of fitting and collapsing. A necessary and sufficient condition for the commutativity of the maximum likelihood estimates is graphical collapsibility together with the closure under marginalisation of the parametric

distribution. Then the estimated probabilities for X_a are the same calculated using the model for X_V or the model for X_a.

The model of no three way interaction $[AN][AS][NS]$ is the only simpler log-linear model fitting the marginal 3-way table defined by A, N and S. It has a deviance equal to 7.29 with 4 degrees of freedom which gives a P-value of 0.12. For practitioners it might be of interest to compare the probability to have a fatal or serious accident on urban areas and rural areas. From Figures 6.2 and 6.3 it can be easily concluded that a fatal or serious accident is more likely to occur on rural roads than on urban roads.

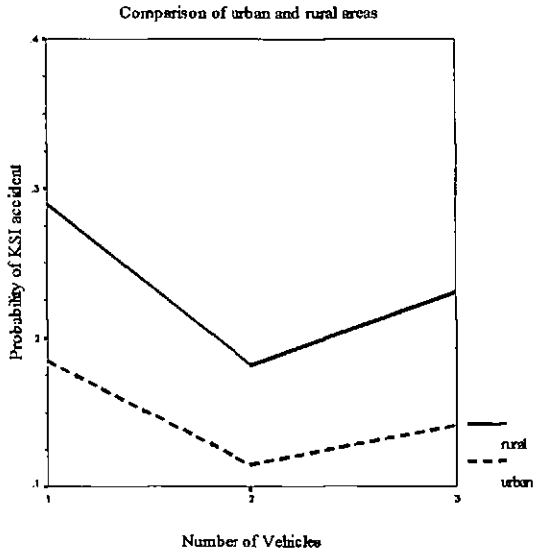


Figure 6.3: Probabilities that an accident on urban and rural roads in Bedfordshire is fatal or serious

Another example where graphical collapsibility can be applied is the model with the independence graph in Figure 3.5, Chapter 3, Section 3.4.1, with $b = \{B\}$, and $a = \{A, C, D\}$. The boundary of $b = \{B\}$ is $\{C, D\}$, which is complete, so the model can be collapsed graphically over B and the conditional

independencies between A, C, D are preserved in the independence graph of X_a. As our model is graphically collapsible over $\{B\}$, and it is well known that the multinomial distribution is closed under marginalisation, the probabilities of interest can be estimated from the marginal model of $\{A, C, D\}$. In other words, the marginal table defined by the variables A, C and D is sufficient for estimation and there is no need to look at the 4-dimensional table. At the same time there are no problems with Simpson's paradox. The same argument is true for the subtable defined by B, C and D. The counts of the two subtables are given in Table 6.1. For each subtable the model of no three-

Table 6.1: Observed counts for subtables BCD and ACD of collision-rollover data

		B				A	
C	D	1	2	C	D	1	2
1	1	376	1989	1	1	2228	137
	2	173	1183		2	1172	184
2	1	79	170	2	1	208	41
	2	192	669		2	516	345

way interaction (Bartlett's model) fits the data well and it is the only one, apart from the saturated model

$$\text{dev}[AC][CD][AD] = 1.71, \quad df = 1, \quad P = 0.19$$

$$\text{dev}[BC][BD][CD] = 1.44, \quad df = 1, \quad P = 0.23.$$

The estimates for the models $[BC][BD][CD]$ and $[AC][AD][CD]$ are given

Table 6.2: Estimates for subtables BCD and ACD of collision-rollover data

		B				A	
C	D	1	2	C	D	1	2
1	1	382.35	1982.65	1	1	2234.10	130.90
	2	166.651	1189.35		2	1165.90	190.10
2	1	72.65	176.35	2	1	201.90	47.10
	2	198.35	662.65		2	522.097	338.90

in Table 6.2. Considering that the variables are standing for rows, columns and layers, the model of no three-way interaction is equivalent to the model of equal odds ratios for rows and columns given the layer. The interpretation can be permuted by fixing either rows or columns. The Bartlett model can be examined by looking at the estimated odds ratios and see if they are approximately equal. This can be done using the unrestricted estimates of the p_{ijk} which are $\hat{p}_{ijk} = \frac{n_{ijk}}{N}$. The index for the ACD table is i for C, j for D and k is for A. Thus, using the counts of Table 6.1, the estimated odds ratios for table ACD are

$$\hat{p}_{111}\hat{p}_{122}/\hat{p}_{112}\hat{p}_{121} = 2.553 \tag{6.1}$$

$$\hat{p}_{211}\hat{p}_{222}/\hat{p}_{212}\hat{p}_{221} = 3.3919 \tag{6.2}$$

Remember that the threshold value for the odds ratio is 1 and its distribution is not symmetric. To overcome this small difficulty, log odds ratios are considered. The hypothesis of interest is whether the two odds ratios are equal. In other words whether the ratio of these odds ratios is 1 or, equiva-

lently whether

$$\log\left(\frac{\widehat{p}_{211}\widehat{p}_{222}/\widehat{p}_{212}\widehat{p}_{221}}{\widehat{p}_{111}\widehat{p}_{122}/\widehat{p}_{112}\widehat{p}_{121}}\right)=0$$

A confidence interval can be easily calculated for this statistic, which has the observed value $3.3919/2.553 = 1.3286$. The standard deviation is

$$SE=\sqrt{\frac{1}{n_{111}}+\ldots+\frac{1}{n_{222}}}=0.7071.$$

The Z variable is $\frac{0.284-0}{0.7071} = 0.4071$. The confidence interval for the ratio of odds ratios is (0.6045 , 2.9182) which includes the value 1. To conclude, for both types of accident, the odds of having a severe injury are almost 3 times larger if the driver is ejected than if the driver is not ejected and the odds of having a not severe injury when the driver has not been ejected are about 3 times larger than the odds of having a not severe injury when the driver has been ejected. This shows that if the driver is ejected in an accident then this substantially increases the probability of being severely injured. Similar conclusions can be deduced by regrouping the variables.

For the table BCD the estimated odds ratios are

$$\widehat{p}_{111}\widehat{p}_{221}/\widehat{p}_{121}\widehat{p}_{211} = 5.2822 \tag{6.3}$$

$$\widehat{p}_{112}\widehat{p}_{222}/\widehat{p}_{122}\widehat{p}_{212} = 6.6164 \tag{6.4}$$

Following the idea described above and fixing variable car type, the hypothesis of interest is

$$\log\left(\frac{\widehat{p}_{112}\widehat{p}_{222}/\widehat{p}_{122}\widehat{p}_{212}}{\widehat{p}_{111}\widehat{p}_{221}/\widehat{p}_{121}\widehat{p}_{211}}\right) = 0$$

The Z variable is $\frac{0.2252-0}{0.7071} = 0.3185$ and a confidence interval for this statistic, which has the observed value $6.6164/5.2822 = 1.2526$, can be constructed. The confidence interval for the ratio of odds ratios is (0.6708 , 2.3382) which includes 1. Thus, for both types of car, the odds of having a severe injury in the case of a rollover accident are 5 or 6 times larger than the odds of having a severe injury in the case of a collision accident.

6.2.1 Response variable models

Very often it is known *a priori* that the variables under study do not play a symmetric role. Some of the variables, say a, are viewed as explanatory (exogenous, treatment-control, independent) for the rest of variables, say b, which are considered response (endogenous, dependent). Ignoring this type of information can be misleading. Graphical chain models described in Chapters 3 and 4 are suitable for this framework. For categorical variables the modelling process was done sequentially as explained in Section 5.4, using the classical log-linear framework. This does not mean that there are no graphical chain models that can be fitted outside the log-linear framework.

This section contains a discussion of collapsibility in relation to a class of models introduced by Goodman (1973) for modelling explanatory and response

variables together. Some connections with the class of graphical chain models will be made and some useful results stated. Goodman's models factorizes the joint density of (a, b) into a product of the marginal density of a and the conditional density of $b \mid a$ such as:

$$p^J(i) = p^M(i_a)p^C(i_b \mid i_a) \tag{6.5}$$

and then a log-linear model M is specified for $p^M(i_a)$ and a log-linear model C for $p^C(i_b \mid i_a)$. The model M is fitted in the marginal table of $\mathbf{n}_a$ and C is fitted as a model for the whole table and since the model is conditioned on a, all the interactions between the variables in a have to be included. The final joint model J has the fitted values $\widehat{m}$ calculated as

$$\widehat{m}^J(i) = \widehat{m}^M(i_a)\{\widehat{m}^C(i)/n(i_a)\}. \tag{6.6}$$

Using the additivity property of the deviance (and the corresponding degrees of freedom of the asymptotic χ^2 distribution) inference for the marginal model and conditional model can be performed separately. However, the class of log-linear models does not coincide with the class of response variable models, see Asmussen and Edwards (1983). In order to determine the intersection of these classes some additional notation is necessary. Let $\mathcal{L}$ be the set of log-linear models for the table of counts $\mathbf{n}$, $\mathcal{M}_a$ be the set of log-linear models for the marginal table of counts $\mathbf{n}_a$, $\mathcal{C}_a$ the set of conditional models (having u^a fixed in the log-linear expansion) and $\mathcal{J}_a$ the set of response variable models

generated from $\mathcal{M}_a$ and $\mathcal{C}_a$. The most important result regarding the response variable models is given in the next theorem (Asmussen and Edwards, 1983). What the author believes to be a more elementary proof is given in detail in Appendix, Section A.

Theorem 6.2 *If $L \in \mathcal{L}$, then $L \in \mathcal{J}_a$ if and only if L is collapsible onto a. In that case $M = L_a$ and $C = [a] \cup L_b$, where $b = \mathrm{cl}(a^c)$.*

The reverse question, when a response variable model is a log-linear model, has an answer in the following theorem, proved in Asmussen and Edwards (1983)

Theorem 6.3 *Let $J = (M, C) \in \mathcal{J}_a$ be a response variable model. Then $J \in \mathcal{L}$ if and only if the boundary of every connected component of a^c is contained in a generator of M. Moreover, $L = M \cup C_b$, where $b = \mathrm{cl}(a^c)$.*

To summarise the results, the log-linear models are appropriate for contingency tables with response and explanatory variables if and only if they are collapsible onto the explanatory variables. For the graphical model in Figure 3.5, Chapter 3, Section 3.4.1, considering car type and accident type, $\{B, C\}$, the explanatory variables and driver ejected and injury type A, D as response variables it is easy to see that $\mathrm{bd}\{A, D\} = \{B, C\}$, which is complete and so the graphical model is collapsible onto the explanatory variables. This means that the graphical model with the independence graph in Figure 3.5 is appropriate. On the contrary, considering just D as a response variable, the same model is not appropriate because it cannot be collapsed onto the explana-

tory variables $\{A, B, C\}$ because $\mathrm{bd}\{D\} = \{A, B, C\}$ which is not complete. In a similar manner, the model in Figure 5.1, Chapter 5, Section 5.2.1, with A accident severity as the only response variable, is appropriate because it can be collapsed onto the explanatory variables R, L, T, S, N since $\mathrm{bd}\{A\} = \{S, N\}$ which is complete.

A generalisation of the class of response variable models is the class of graphical chain models described in Chapter 3. For these models, variables are divided into blocks $V(1) \cup V(2) \ldots \cup V(T)$, by a partial ordering relationship, given by time or any other possible causal prior substantive knowledge. Define the sets $d_0 = V(1), d_i = V(i+1) \cup d_{i-1}$, for all $i \in \{1, \ldots, T-1\}$. Then the class of graphical chain models is defined by the following factorisation of the joint density which describes the log-linear models $C_0, C_1, \ldots, C_{T-1}$ on the corresponding marginal tables

$$p^J = p^{C_0}(d_0) \prod_{i=1}^{T-1} p^{C_i}(d_i \mid d_{i-1}).$$

The collapsibility results for response variable models are generalised, Asmussen and Edwards (1983), in the next theorem.

Theorem 6.4 *A log-linear model $L \in \mathcal{L}$ is a graphical chain model if and only if it is collapsible onto d_i, for all $i \in \{0, 1, \ldots, T-1\}$.*

Conversely, a graphical chain model $J = (C_0, C_1, \ldots, C_{T-1})$ is log-linear if and only if the boundary of each connected component of $V(i+1)$ under C_i is contained in a generator of C_{i-1}, for all $i \in \{1, 2, \ldots, T-1\}$.

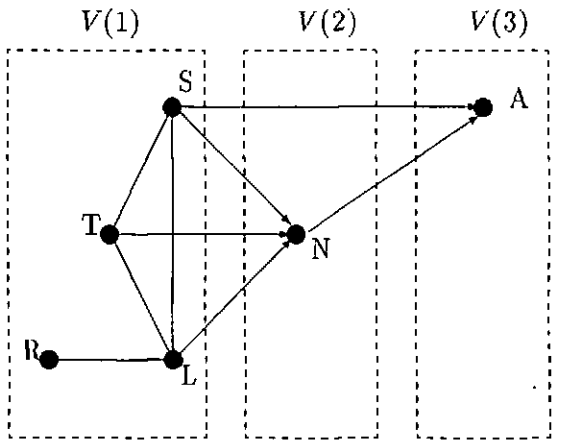


Figure 6.4: Graphical chain model for Bedfordshire data with 6 variables

An example of a graphical chain model that is log-linear is the model in Figure 6.4. It is relatively easy to see that

$$\begin{aligned} d_0 = V(1) &= \{R, L, T, S\} \\ d_1 = V(2) \cup d_0 &= \{N, R, L, T, S\} \\ d_2 = V(3) \cup d_1 &= \{A, N, R, L, T, S\} \end{aligned}$$

and therefore

$$\begin{aligned} \mathrm{bd}(V(2) \mid C_1) = \mathrm{bd}(N \mid C_1) &= \{S, T, L\} \subseteq [STL] \\ \mathrm{bd}(V(3) \mid C_2) = \mathrm{bd}(A \mid C_2) &= \{S, N\} \subseteq [SN] \end{aligned}$$

An example of a graphical chain model which is not a log-linear model is the

model with the chain graph in Figure 3.7. For that model

$$V(1) = \{B, D\}, \quad V(2) = \{A\}, \quad V(3) = \{C\}.$$

But

$$\begin{aligned} \mathrm{bd}(A \mid C_1) &= \{D\} \subseteq [BD] \\ \mathrm{bd}(C \mid C_2) &= \{A, B, D\} \end{aligned}$$

which is not included in any generator of C_1 defined by

$$p(A, B, D) = p(A, D)p(D, B)$$

Conversely, there are log-linear models that are not graphical chain models. For example, the log-linear model

$$L_1 = [RL][TS][LSA][NT][LN]$$

is not a graphical chain model for the dependence chain $\{T, R, L, S\} \cup \{N\} \cup \{A\}$. This is because it should be collapsible onto $d_0 = \{T, R, L, S\}$ and this by definition means that $\mathrm{bd}(N) = \{T, L\}$ and $\mathrm{bd}(A) = \{S, L\}$ are complete, which is not true for the first boundary. This model can be made a graphical chain model if the interaction between L and T is allowed in the model.

Minimal collapsible set

Very often there is some particular interest in a subset b of variables of a larger set of variables V. It is not always possible to collapse onto b, so the problem is then what is the minimal subset $b_1, b \subseteq b_1 \subseteq V$, such that the log-linear model L can be collapsed onto b_1? This problem has an answer when the log-linear model is decomposable.

The results are based on the concept of simplicial vertex and a version of Graham's algorithm known as *Selective Acyclic Hypergraph Reduction*, proposed by Tarjan and Yannakis (1984). A vertex is called *simplicial* if its boundary is complete. The Selective Acyclic Hypergraph Reduction algorithm, (SAHR), follows the steps:

1. draw up a list of cliques of the corresponding interaction graph;

2. remove a simplicial vertex which is not in b;

3. delete from the list of cliques any redundant clique;

4. repeat the last two steps until neither is applicable.

The minimal collapsible set is given by the subset of vertices left. The main result, Madigan and Mosurski (1990), is given by the following theorem.

Theorem 6.5 *Let L be a decomposable log-linear model having the interaction graph $\mathcal{G} = (V, E)$ and let b be a subset of variables of interest $b \subseteq V$. Then the SAHR algorithm provides the minimal set $b_1, b \subseteq b_1 \subseteq V$, such that L can be collapsed onto b_1.*

Let consider again the graphical model for Hampshire data with the corresponding conditional independence graph in Figure 6.5. This model is not collapsible onto $\{A, T\}$ because $\mathrm{bd}\{R, N, S, L\} = \{A, T\}$ is not complete on the graph. However, suppose that there is an interest in collapsing this 6-dimensional table onto a smaller one containing A, T. For the SAHR algorithm let $b = \{A, T\}$ so $b^c = \{R, N, S, C\}$. It is easy to verify that R and L are simplicial, that is that their boundary is complete, and that N and S are not simplicial. The algorithm starts with the cliques

$$[RNST][ASN][SLT]$$

and in the first step it removes the simplicial vertex R. Thus, the next set of cliques is given by

$$[NST][ASN][SLT]$$

and in the second step of the algorithm the vertex L is eliminated. In conclusion the minimal subset, containing the variables $\{A, T\}$, onto which the model in Figure 6.5 can be collapsed is $\{A, T, S, N\}$. This can be checked by seeing that

$$\mathrm{bd}(\{A, T, S, N\}^c) = \mathrm{bd}(\{R, L\}) = \{T, S, N\}$$

which is a complete subset on the graph.

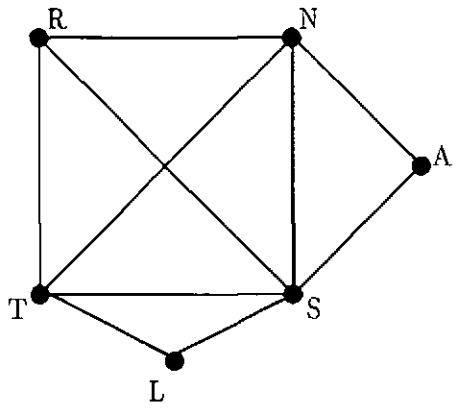


Figure 6.5: Graphical model for Hampshire data

6.3 Summary

The concept of collapsibility relative to log-linear models for contingency tables are extremely important. Not taking into account whether collapsibility equivalent conditions allow a multi-dimensional table to be collapsed and a marginal small-dimensional table to be analysed instead, may lead to Simpson's paradox.

Reducing safely the dimension of the analysis has important benefits, especially when the large table is sparse and asymptotic tests are unreliable. This was shown on a particular example in this chapter. The analysis of a six dimensional table was focused on a three dimensional marginal table defined by accident severity, speed limit and the number of vehicles, doing also estimation of some probabilities of interest. It was also shown how the analysis of a 4-dimensional table can be safely decomposed into two separate analyses

of 3-dimensional tables, the analysis being continued on some odds ratios of interest.

Graphical chain models are helpful for situations where the set of variables under study can be classified as response and explanatory. It was described in this section how to apply some collapsibility results and decide whether a log-linear model is a response variable model.

When the model cannot be collapsed onto a desired subset of variables b it is still possible to find out a minimal subset of variables b_1, containing the subset b, such that the model is collapsible onto b_1. This can help once more to reduce the complexity of the model by analysing a reduced number of variables.

Chapter 7

Problems for compound Poisson distributions

7.1 Introduction

The analysts using likelihood or empirical Bayes methods "estimate" some unknown parameters describing the statistical model and then provide inference as if the data has been generated by the model with those estimated parameters. The estimation process is therefore crucial and bad estimation can lead to false inference.

For count data, it is very common to use a compound Poisson-gamma distribution for modelling since this distribution helps to overcome overdispersion. This implies that the marginal distribution of the observed data follows a negative binomial distribution with two unknown parameters.

In this chapter, an insight into the process of maximum likelihood esti-

mation (MLE) for both parameters is given and a new proof of when there is such an estimator is given. The new approach does not give an answer to the question of whether this bivariate MLE estimator is unique but it does provide a numerical equivalent condition that can be checked on the computer for any set of data.

Because the first part of this chapter suggests that the inference process may be sensitive to the choice of prior a numerical technique is developed in the second part of the chapter for investigating the change in posterior inference due to the change in prior distribution. An example based on road accident data is also described.

7.2 Estimation problems for NB distribution

Let $Y = (Y_1, \ldots, Y_n)$ be a sample of size n from a negative binomial distribution

$$\mathrm{NB}(x \mid p, \kappa) = \binom{\kappa+x-1}{x} p^{\kappa}(1-p)^{x} \tag{7.1}$$

for $x = 0, 1, 2, \ldots$, and where $0 < p < 1$ and $\kappa > 0$. The combinatorial term $\binom{\kappa+x-1}{x}$, which is equal to $\binom{\kappa+x-1}{\kappa-1}$, is generally used for κ positive integer, but when κ is real it is equal to $\frac{\Gamma(\kappa+x)}{x!\Gamma(\kappa)}$. This is equal to 1 when x is zero.

When the parameter κ is known, the negative binomial distribution is of exponential type and the estimation process for p is simple and straightforward. On the contrary, when κ and p are both unknown then the negative binomial distribution is no longer a member of the exponential family and

there are some unforeseen problems regarding the estimation of κ.

This distribution arise often in a Bayesian context. It is not therefore surprising that some parameters are estimated by biased but minimum variance estimators. In the exponential family of distributions there is always a complete sufficient statistic so minimum variance unbiased estimators can be identified. However, this is not the case for the NB distribution with κ unknown. The next theorem, proved in Willson, Folks and Young (1986), is just the tip of the iceberg.

Theorem 7.1 *The order statistic $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ is minimal sufficient but not complete for the negative binomial family of distributions, when $n \geq 3$.*

This means that given an unbiased estimator of (p, κ) the well known Rao-Blackwell theorem for determining an *unique* unbiased estimator, for the same parameters, cannot be applied. Therefore, there may exist several unbiased estimators, all functions of the minimal sufficient statistic, for which we cannot compare their variances. This situation is due to having both parameters of the negative binomial distribution unknown and it gives a hint that there may be some problems regarding the MLE estimators for the NB distribution.

Willson et al. (1986) found that an uniformly mininum variance unbiased estimator of κ cannot be obtained in the usual manner. An explanation was offered by Wang (1996) and it is described in the following theorem:

Theorem 7.2 *There is no unbiased estimator of κ for* $\mathrm{NB}(p, \kappa)$.

Proof : Let $T(Y_1, \ldots, Y_n)$ be an estimator of κ for $\mathrm{NB}(p, \kappa)$. This estimator is

unbiased if and only if

$$\mathrm{E}(T(Y)|\kappa, p) = \kappa, \quad \text{for all } \kappa > 0 \text{ and } p \in (0,1).$$

Hence, using the density function given in (7.1), if T is unbiased then

$$\sum_{y} \left[\prod_{i=1}^{n} \binom{y_i+\kappa-1}{\kappa-1} \right] p^{n\kappa}(1-p)^{\sum_{i=1}^{n} y_i} T(y) = \kappa$$

where $y = (y_1, \ldots, y_n)$ and the summation is taken over all n-uples of positive integers. Rearranging the terms it follows that

$$T(0,\ldots,0)p^{n\kappa} + \sum_{y \neq 0_n} \left[\prod_{i=1}^{n} \binom{y_i+\kappa-1}{\kappa-1} \right] p^{n\kappa}(1-p)^{\sum_{i=1}^{n} y_i} T(y) = \kappa$$

For $\kappa = 1$ and $p \to 1$ we get that $T(0,\ldots,0) = 1$ and taking $\kappa \neq 1$ and $p \to 1$ we get that $T(0,\ldots,0) = \kappa$, which is obviously a contradiction.□

This simple but powerful result is not altogether surprising from a Bayesian point of view. In general, for a univariate parameter θ, the Bayes estimator $T(Y) = \mathrm{E}(\theta|Y)$ is biased no matter what prior distribution $\pi(\theta)$ is used. The following theorem summarises some known results and provides at the same time a motivation for using Bayesian estimators rather than classical frequentist estimators.

Theorem 7.3 *Consider a statistical model with observed data* $y = (y_1, \ldots, y_n)$ *and an univariate parameter* θ. *Then, if* $T(Y) = \mathrm{E}(\theta|Y)$ *is the Bayes estimator,*

1. *for any prior distribution* $\pi(\theta)$, *if* $\mathrm{var}(T(Y)) > 0$ *then* $T(Y)$ *is biased.*

2. $T(Y)$ *is an admissible estimator of* θ *relative to squared error loss*

$$MSE_\theta = \mathrm{E}\left((T(Y) - \theta)^2 | \theta\right)$$

3. *If the risk of* $T(Y)$ *is finite, that is* $\mathrm{E}(MSE_\theta) < \infty$, *then*

$$\mathrm{E}\,(MSE_\theta(T)) \leq \mathrm{E}\,(MSE_\theta(U))$$

for any other estimator $U(Y)$ *and the equality is obtained if and only if* $T(Y) = U(Y)$ *almost everywhere.*

The fact that there is no unbiased estimator for the parameter κ of the negative binomial distribution NB suggests that, in this case, estimators with good properties are very likely to come from a Bayesian approach.

The negative of the corresponding log-likelihood function is, up to a constant factor,

$$f_1(\kappa, p) = \left(-\sum_{i=1}^{n} y_i\right) \log(1-p) - n\kappa \log p - \sum_{i=1}^{n} \sum_{j=0}^{y_i - 1} \log(\kappa + j)$$

where the last sum has a zero term contribution when $y_i = 0$ and this will be true for all the subsequent calculations. The trivial case when the sample contains only zeros, that is $y_i = 0$ for all $i = 1, 2, \ldots, n$, is not of interest in

this thesis and it seems hard to imagine an application where this sample is meaningful. Therefore, the assumption that at least one element of the sample is different than zero, is natural and such a sample will be called non-trivial. The likelihood equations are

$$\frac{\partial f_1(\kappa,p)}{\partial p} = \frac{1}{1-p}\sum_{i=1}^{n} y_i - \frac{n\kappa}{p} = 0 \tag{7.2}$$

$$\frac{\partial f_1(\kappa,p)}{\partial \kappa} = -n\,\log p - \sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{1}{(\kappa+j)} = 0. \tag{7.3}$$

From likelihood equation (7.2) the MLE of p is $\widehat{p} = \frac{\kappa}{\kappa+m}$, where $m = \frac{1}{n}\sum_{i=1}^{n} y_i$. Replacing $\widehat{p}$ in (7.3) the following likelihood equation is obtained

$$\log(1+\frac{m}{\kappa}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{1}{\kappa+j}. \tag{7.4}$$

It can be easily seen that there are no closed form solutions of this equation. If $S^2 = \frac{1}{n}\sum_i (y_i - m)^2$, Anscombe (1950) conjectured that there is only one positive solution $\widehat{\kappa}$ when $S^2 > m$ and none otherwise. Johnson and Kotz (1969) proved that there is at least one positive solution $\widehat{\kappa}$ when $S^2 > m$. Ross and Preece (1985) described how to fit the NB for real data in the computer program MLP. It is not known if the MLE of κ is unique and it seems that it has not been proved that there is no solution when $S^2 < m$. Aragon, Eberly and Eberly (1992) claimed to have proved the existence and uniqueness of the maximum likelihood estimator for the two-parameter negative binomial distribution but Wang (1996) showed that there is a mistake in their proof

and moreover, he could not correct it.

The NB distribution is frequently used for fitting biological data and the related statistical literature has grown considerably over the years. However precise estimation of κ has been elusive and other methods of estimation were proposed and compared (Willson, Folks and Young, 1984; Willson et al., 1986) and simulation and graphic tools like contours and 3-dimensional plots of the log-likelihood function provided to show that the possibilities about MLE of κ are not encouraging. The log-likelihood can be very flat instead of being peaked and this means that the MLE of κ could be sensitive to small changes in sample values. A fully Bayesian approach may be more informative.

It will be shown that there is at least a positive MLE of κ, a different proof being given in Willson et al. (1986), that there is no solution when $S^2 < m$, and a sufficient condition will be identified when there is a unique solution $\widehat{\kappa}$ of the MLE equations. A definite answer is not given, but this criterion can be checked on computer for any set of data.

The profile function $f(\kappa) = f_1(\kappa, \widehat{p}(\kappa))$ is

$$\begin{aligned} f(\kappa) &= n[(\kappa+m)\log(\kappa+m) - \kappa\log\kappa - m\log m] - \sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\log(\kappa+j) \\ &= n[\kappa\log(\kappa+m) - \kappa\log\kappa + m\log(\kappa+m) - m\log m] \\ &\quad -\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\log(\kappa+j). \end{aligned}$$

The next step is to transform the parameter κ by the one-to-one transformation $\beta(\kappa) = \kappa[\log(\kappa+m) - \log\kappa]$, where $\beta : (0,\infty) \longrightarrow (0,m)$. This is a

strictly monotone increasing concave function because

$$\begin{aligned} \beta'(\kappa) &= \log(\kappa+m) - \log\kappa - \frac{m}{\kappa+m} > 0 \\ \text{and } \beta''(\kappa) &= -\frac{m^2}{\kappa(\kappa+m)^2} < 0 \end{aligned}$$

for any $\kappa > 0$. Moreover $\lim_{\kappa \searrow 0} \beta(\kappa) = 0$ and $\lim_{\kappa \to \infty} \beta(\kappa) = m$. The transformation β is one-to-one and instead of studying whether the profile log-likelihood function $f = f(\kappa)$ has a positive root, f can be studied as a function of β. To prepare the grounds, a few preliminary results are proved first.

Lema 7.1 *For a non-trivial sample $y = (y_1, \ldots, y_n)$ from the $\mathrm{NB}(p, \kappa)$ distribution, the application f and parameter β introduced above, it is true that*

$$\lim_{\kappa \searrow 0} \frac{df}{d\beta} = -\infty.$$

Proof: By the chain rule

$$\begin{aligned} \lim_{\kappa \searrow 0} \frac{df}{d\beta} &= \lim_{\kappa \searrow 0} \frac{df}{d\kappa}\frac{d\kappa}{d\beta} \\ &= \lim_{\kappa \searrow 0} \frac{df}{d\kappa} \Big/ \frac{d\beta}{d\kappa} \\ &= \lim_{\kappa \searrow 0} \frac{n[\log(\kappa+m) - \log\kappa] - \sum_{i=1}^{n} \sum_{j=0}^{y_i - 1} \frac{1}{\kappa+j}}{\log(\kappa+m) - \log\kappa - \frac{m}{\kappa+m}}. \end{aligned}$$

Denoting by τ the number of non-zero y_i, $i = 1, \ldots, n$, and separating the

terms in the double sum for which $j = 0$, it follows that

$$\begin{aligned}\lim_{\kappa \searrow 0} \frac{df}{d\beta} &= \lim_{\kappa \searrow 0} \frac{n[\log(\kappa+m) - \log\kappa] - \frac{\tau}{\kappa} - \sum_{i=1}^{n} \sum_{j=1}^{y_i - 1} \frac{1}{\kappa + j}}{\log(\kappa+m) - \log\kappa - \frac{m}{\kappa+m}} \\ &= n + \lim_{\kappa \searrow 0} \frac{\frac{mn}{\kappa+m} - \frac{\tau}{\kappa} - \sum_{i=1}^{n} \sum_{j=1}^{y_i - 1} \frac{1}{\kappa+j}}{\log(\kappa+m) - \log\kappa - \frac{m}{\kappa+m}}\end{aligned}$$

and applying l'Hopital rule for the second term

$$\begin{aligned}\lim_{\kappa \searrow 0} \frac{df}{d\beta} &\overset{\text{l'Hopital}}{=} n + \lim_{\kappa \searrow 0} \frac{-mn\kappa + \frac{\tau(\kappa+m)^2}{\kappa} + \kappa(\kappa+m)^2 \sum_{i=1}^{n} \sum_{j=1}^{y_i - 1} \frac{1}{(\kappa+j)^2}}{-m^2} \\ &= n + \lim_{\kappa \searrow 0} \frac{mn\kappa}{m^2} + \lim_{\kappa \searrow 0} \frac{\tau(\kappa+m)^2}{-m^2\kappa} + \lim_{\kappa \searrow 0} \frac{\kappa(\kappa+m)^2 \sum_{i=1}^{n} \sum_{j=1}^{y_i - 1} \frac{1}{(\kappa+j)^2}}{-m^2} \\ &= n + \lim_{\kappa \searrow 0} \frac{\tau(\kappa+m)^2}{-m^2\kappa} \\ &= -\infty. \square\end{aligned}$$

Lema 7.2 *If m and $S^2 = \frac{1}{n}\sum_i (y_i - m)^2$ are the mean and the sample variance of a non-trivial negative binomial sample $y = (y_1, \ldots, y_n)$, and f and β as above, then it is true that*

$$\lim_{\kappa \to \infty} \frac{df}{d\beta} = \frac{n}{m^2}(S^2 - m).$$

Proof : As before, using the chain rule followed by l'Hopital rule, we can calculate

$$\begin{aligned}\lim_{\kappa \to \infty} \frac{df}{d\beta} &= \lim_{\kappa \to \infty} \frac{df}{d\kappa} \Big/ \frac{d\beta}{d\kappa} \\ &\overset{\text{l'Hospital}}{=} \lim_{\kappa \to \infty} \frac{d^2 f}{d\kappa^2} \Big/ \frac{d^2\beta}{d\kappa^2}\end{aligned}$$

$$
\begin{aligned}
&= \lim_{\kappa\to\infty} \frac{\frac{-mn}{\kappa(\kappa+m)} + \sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{1}{(\kappa+j)^2}}{\frac{-m^2}{\kappa(\kappa+m)^2}} \\
&= \lim_{\kappa\to\infty} \frac{-mn(\kappa+m) + \kappa(\kappa+m)^2\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{1}{(\kappa+j)^2}}{-m^2} \\
&= \frac{1}{m^2}\lim_{\kappa\to\infty}\left[mn(\kappa+m) - (\kappa+m)\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{\kappa(\kappa+m)}{(\kappa+j)^2}\right] \\
&= \frac{1}{m^2}\lim_{\kappa\to\infty}(\kappa+m)\left[mn - \sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{\kappa(\kappa+m)}{(\kappa+j)^2}\right] \\
&= \frac{1}{m^2}\lim_{\kappa\to\infty}(\kappa+m)\left[\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\left(1 - \frac{\kappa(\kappa+m)}{(\kappa+j)^2}\right)\right] \\
&= \frac{1}{m^2}\lim_{\kappa\to\infty}\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{(\kappa+m)[\kappa(2j-m)+j^2]}{(\kappa+j)^2} \\
&= \frac{1}{m^2}\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}(2j-m) \\
&= \frac{1}{m^2}\sum_{i=1}^{n}(y_i^2 - my_i - y_i) \\
&= \frac{1}{m^2}\left(\sum_{i=1}^{n}(y_i-m)^2 + \sum_{i=1}^{n}my_i - \sum_{i=1}^{n}y_i - \sum_{i=1}^{n}m^2\right) \\
&= \frac{1}{m^2}(nS^2 + m^2n - mn - m^2n) \\
&= \frac{n}{m^2}(S^2 - m).\square
\end{aligned}
$$

Therefore, because $\frac{df}{d\beta}$ is a continuous function and using the above lemmas it is obvious that $\frac{df}{d\beta} = 0$ has at least one positive solution when $S^2 > m$.

Theorem 7.4 *For the negative binomial distribution* $\mathrm{NB}(p,\kappa)$*, there is at least one MLE of* κ*. Moreover, the MLE is unique if* $\frac{df}{d\beta}$ *is a strict monotone function. A sufficient condition to have a unique MLE is that* $\frac{d^2f}{d\beta^2} > 0$.

Proof: If there are two roots β_1 and β_2 of the equation $\frac{df}{d\beta} = 0$ then, because $\frac{df}{d\beta}$ is differentiable, there must be at least one solution β^* of the equation

$\frac{d^2f}{d\beta^2} = 0$, where β^* is between β_1 and β_2. It is easy to see that

$$\begin{aligned}\frac{d^2f}{d\beta^2} &= \frac{d}{d\beta}\left(\frac{df}{d\beta}\right) \\ &= \frac{\frac{d^2f}{d\kappa^2}\cdot\frac{d\beta}{d\kappa} - \frac{d^2\beta}{d\kappa^2}\cdot\frac{df}{d\kappa}}{\left[\frac{d\beta}{d\kappa}\right]^3}.\end{aligned}$$

The condition $\frac{d^2f}{d\beta^2} > 0$, which will prove that there is unique MLE of κ, means that

$$\frac{d^2f}{d\kappa^2}\cdot\frac{d\beta}{d\kappa} > \frac{d^2\beta}{d\kappa^2}\cdot\frac{df}{d\kappa}$$

which is equivalent to

$$\left[n\left(\frac{1}{\kappa+m}-\frac{1}{\kappa}\right)+\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{1}{(\kappa+j)^2}\right]\left[\log(\kappa+m)-\log\kappa-\frac{m}{\kappa+m}\right] >$$

$$-\frac{m^2}{\kappa(\kappa+m)^2}\left[n(\log(\kappa+m)-\log\kappa)-\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{1}{\kappa+j}\right]$$

$$\begin{aligned}m^2n+\kappa(\kappa+m)\left[(\kappa+m)\log(1+\frac{m}{\kappa})-m\right]\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{1}{(\kappa+j)^2} - \\ -mn\kappa\log(1+\frac{m}{\kappa})-m^2\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{1}{(\kappa+j)} > 0 \qquad (7.5)\end{aligned}$$

It seems that this complicated formula cannot be further simplified or proved. Therefore, a definitive answer is not known whether the MLE of κ

is unique. However, for a given set of data, it can be checked on a computer whether the left side of equation (7.5) is strictly positive. It cannot be strictly negative because of the results in Lemmas 1 and 2.

The worst situation that may occur is that there are many solutions of the MLE equations and they are widely spread. If the likelihood of κ is not peaked around the mode but it has a very flat top, a small change in the sample may result in large shifts in the MLE solutions and therefore the inference results may change dramatically. Therefore, it would be extremely useful to be able to investigate a large range of priors, or in other words a large range of the mixing distributions G for the compound Poisson-G distributions. Another advantage of being able to do this is that the so called "gamma assumption", discussed in Chapter 2, Section 2.2.2, can be challenged and other types of mixing distributions can be investigated.

A numerical procedure that is doing just that is described in the next section.

7.3 Sensitivity analysis of priors in compound Poisson modelling

In this section, a method is proposed for investigating the sensitivity of prior choice in compound Poisson modelling. After a theoretical derivation, a practical example involving a road accident data set is described.

7.3.1 Theoretical derivation

Most of the time, accident count data shows overdispersion. This is quite natural because of the unobserved changes in environmental conditions, social changes and so on that take place all the time and that are not reflected in the data at the covariate level. The most used model to account for this phenomenon is the compound Poisson-gamma model. This model can be described by

$$\begin{aligned} Y_k \mid \lambda_k &\sim \text{Pois}(\lambda_k), \quad \text{for all } k = 1, 2, \ldots, N \\ \lambda_k \mid a, b &\sim \text{gamma}(a, b). \end{aligned}$$

Assuming that a and b are known quantities, it is relatively straightforward to calculate the posterior means

$$\text{E}(\lambda_k \mid y) = \frac{(y_k + a)}{b + 1}. \tag{7.6}$$

for all sites $k = 1, 2, \ldots, N$ and where $y = (y_1, \ldots, y_N)$ and $\lambda = (\lambda_1, \ldots, \lambda_N)$. The gamma distribution is used as a mixing distribution mainly because of computational simplicity. This prior distribution will be considered in the following as a reference prior and will be denoted by p_{ref}. The distribution of another prior investigated for comparison will be denoted by p_{new}. Following a result due to Kass, Tierney and Kadane (1989), if $\lambda_k \sim p_{new}$ then, the

posterior expectations of λ_k can be approximated by the formula

$$\mathrm{E}_{new}(\lambda_k \mid y) \approx \frac{b(\widetilde{\lambda})}{b(\widehat{\lambda})}\mathrm{E}_{ref}(\lambda_k \mid y) \tag{7.7}$$

where $b(\lambda) = \frac{p_{new}(\lambda)}{p_{ref}(\lambda)}$, $\widetilde{\lambda}$ maximizes $\log[\lambda_k p(y \mid \lambda)p_{ref}(\lambda)]$ and $\widehat{\lambda}$ maximizes the reference log-likelihood $\log[p(y \mid \lambda)p_{ref}(\lambda)]$.

Taking $p_{new} = \mathrm{logN}(\mu, \sigma^2)$, that is the log normal distribution, it can be easily calculated that

$$b(\lambda_1, \ldots, \lambda_N) = \frac{\Gamma(a)^N}{b^{aN}\sigma^N\sqrt{2\pi}^N} \prod_{k=1}^{N} \left(\lambda_k^{-a} \exp\left[\lambda_k b - \frac{1}{2\sigma^2}(\log \lambda_k - \mu)^2\right]\right). \tag{7.8}$$

The only thing left is to calculate $\widetilde{\lambda}$ and $\widehat{\lambda}$. Since

$$\log p(y \mid \lambda)p_{ref}(\lambda) \propto \sum_{k=1}^{N}[(y_k + a - 1)\log \lambda_k - \lambda_k(b+1)]$$

the optimising solutions are

$$\widehat{\lambda_k} = \frac{y_k + a - 1}{b+1}, \quad \text{for all } k = 1, 2, \ldots, N \tag{7.9}$$

under the requirement that $a > 1$. It can be easily remarked that, for $a = 1$ and sites with $y_k = 0$ observed accidents, the above formula is not convenient because it implies that λ_k is zero. Therefore, either a reference gamma prior with the shape parameter a greater than 1 is used or estimation of posterior means is done separately for sites with zero observed accidents.

Similarly, because

$$\log[\lambda_k p(y \mid \lambda) p_{ref}(\lambda)] \propto \sum_{i \neq k}^{N} [(y_i + a - 1)\log\lambda_i - \lambda_i(b+1)] + (y_k + a)\log\lambda_k - \lambda_k(b+1)$$

it can be easily shown that

$$\widetilde{\lambda_i} = \frac{y_i + a - 1}{b+1}, \quad \text{for all } i \neq k \tag{7.10}$$

$$\widetilde{\lambda_k} = \frac{y_k + a}{b+1} \tag{7.11}$$

and again $a > 1$ is required in order to have convenient solutions, otherwise sites with zero accidents must be treated separately.

Plugging the solutions from Equations (7.8), (7.9), (7.10) and (7.11) into formula (7.7), for the compound Poisson-log normal distribution the posterior means are approximately

$$\mathrm{E}_{new}(\lambda_k \mid y) \approx \frac{y_k + a - 1}{b+1} e^{\left\{\frac{b}{b+1} - \frac{1}{2\sigma^2}\log\frac{y_k+a}{y_k+a-1}\left[\log\frac{(y_k+a)(y_k+a-1)}{(b+1)^2} - 2\mu\right]\right\}}. \tag{7.12}$$

7.3.2 Application to road accident data in Kent

In Chapter 8 different compound Poisson models, fully Bayesian specified, are fitted to the total number of accidents between 1984 and 1991, on 156 single-carriageway link sites in Kent. The posterior Bayes estimates for the gamma prior parameters are $a = 0.58$ and $b = 0.02$ and for the log normal prior parameters are $\mu = 2.44$ and $\sigma = 2.45$. There are some weaknesses about these

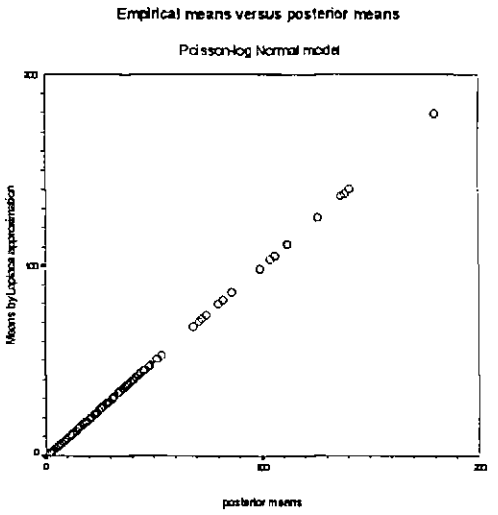


Figure 7.1: Approximate posterior means, calculated from gamma$(0.58, 0.02)$, *against the posterior means of Poisson-log normal model with* $\mu = 2.44$ *and* $\sigma^2 = 2.45$*; sites with zero observed accidents are missing*

two priors that should be acknowledged. From table 7.1 it can be seen that the variances of these two priors are very large. Thus, the value of the means does not play any role. The variance of the log normal prior is 8 times higher than the variance of the gamma distribution but in real terms both can be understood as infinite. Due to this non-informative or largely diffuse character of the priors used it follows that the data will dominate the priors so it is not surprising to see a very close agreement between the posterior estimates and the observations. The elicitation of prior distributions is subject of intensive research and it is known to be difficult. The priors used in this section play a rather illustrative purpose regarding the method proposed for studying the sensitivity of the priors in compound Poisson modelling. The research done by Doss and Narasimhan (1994) can be also useful for investigating, for Poisson-regression modelling, the effects on results of a large range of priors.

Table 7.1: Means and variances of two prior distributions

prior distribution	mean	variance
gamma$(0.58, 0.02)$	29	1450
logN$(2.44, 2.45)$	38.86	11604

In this section the above approximation machinery is used to calculate the posterior means of accidents for all 156 sites, with an unknown log normal distribution as the new prior and gamma$(0.58, 0.02)$ as the reference distribution. Since the shape parameter of the Gamma prior is $a = 0.58 < 1$ sites with zero accidents do not have a solution. For comparison a parallel calculation is made, doubling the value of a to 1.172. In this second situation, with gamma$(1.17, 0.02)$, approximate solutions are possible for all sites.

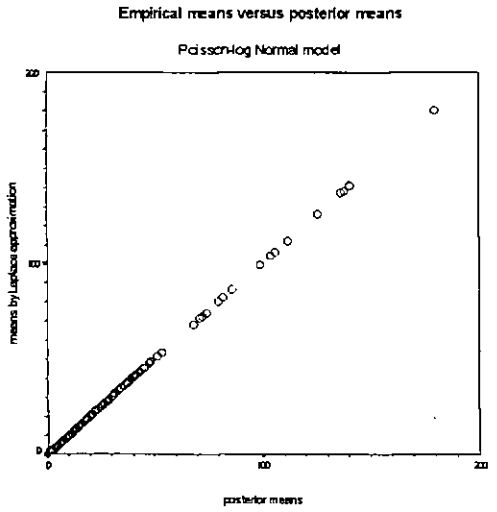


Figure 7.2: Approximate posterior means, calculated from gamma$(1.17, 0.02)$, *against the posterior means of Poisson-log normal model with* $\mu = 2.44$ *and* $\sigma^2 = 2.45$*; all sites represented*

Both sets of posterior means can then be compared with the posterior

means under a fully Bayesian Poisson-log normal model, with the estimated posterior parameters $\mu = 2.44$ and $\sigma = 2.45$. The comparison is made by plotting the approximated means, as calculated from Equation (7.12), against the posterior means given by the fully Bayesian Poisson-log normal model investigated later in Chapter 8. The plots are in Figures 7.1 and 7.2. The fit seems to be very good, so the approximation method advocated in the previous section can provide reliable and easy calculations.

This method can be used as a tool to investigate the use of different priors, like the log-normal already investigated here, or the inverse Gaussian, or other more complicated distributions which are not implemented in standard packages and whose usefulness has not been yet confirmed.

7.4 Summary

Great care should be taken in applying even the most known estimation methods relative to the two-parameter negative binomial distribution. The MLE equations are non-linear and analytical solutions are not tractable. With this excuse, the majority of applied studies using negative binomial fitting for road accident data employes the method of moments for estimation. This circumvent the estimation problems for the parameter κ and the statistical inference is obtained relative to a single estimate.

Another proof of some general inference results for the NB distribution has been given in this chapter. A sufficient condition with the uniqueness

of MLE estimators for NB(p, κ), that can be verified for each set of data, has been proposed. From the applied point of view, after finding by some numerical procedures the MLe of p and κ, the condition given by equation (7.5) gives a straighforward answer to the qnestion whether there could be multiple solutions to the likelihood equations. If there is a unique solntion then the conclusions can be based on this set of estimates; otherwise a more in depth analysis is required.

Compound Poisson models are often proposed for modelling count data in general and accident data in particular. The Poisson-gamma model is one of the well-known instances. The choice of the prior distribution, or the compound distribntion, is a relative matter and although the choice of gamma distribntion is motivated by the conjugacy with Poisson distribution, other distribntions having a positive support may give a better fit to some sets of data. A nnmerical procedure for studying the sensitivity of prior choice has been developed and applied for a set of accident counts. The advantage of this procedure is that avoids complicated calculations and a wide range of distributions can be investigated easily.

Chapter 8

Bayesian models for accident counts

8.1 Introduction

Statistical science was developed in the 19^{th} and 20^{th} centuries by the founders such as Francis Galton, Karl Pearson, Sir Ronald.A. Fisher, Jerzy Neyman and Egon Pearson. Although at the beginning there was no clear distinction between the *frequentist* approach and the *Bayesian* approach, the former was preferred in most of the 20th century because of the mathematical developments supporting the methodologies defining the frequentist school of thought. Bayesian methods experienced a revolution in the last decade due to the development of Markov Chain Monte Carlo methods and are getting more and more enthusiasts attracted by the flexibility of this type of statistical modelling. Paradoxically, the Bayesian approach is older, starting with the

original 1763 paper by the Rev. Thomas Bayes. The controversy surrounding the two approaches is not the subject of this thesis. One of the strongest arguments against the use of Bayesian statistics was the lack of closed-form mathematical results and what frequentist school called the lack of objectivity. It is not the aim of this thesis to discuss the pros and cons of the Bayesian methodology. We are more interested in the benefits of the Bayesian methodology for the applied work. Some of the problems analysed in this thesis, like modelling multiple count response variables, seem to have a solution only in a fully Bayesian framework. There is no free lunch, of course, and the choice of prior distributions can be seen as a lack of objectivity. However, in this thesis the majority of priors were largely spread, a non-informative approach being used for the empirical work. Mathematical solutions could be developed only for a limited range of probability distributions, such as the normal distribution. Multivariate problems in a Bayesian framework lead sooner or later to the calculation of multi-dimensional integrals of very high order. For a while, the inability to calculate such integrals hampered the development of these methods. The computational problems related to hierarchical models concern multi-dimensional integrals of order higher than 20, so a more sophisticated approach is needed.

Helped by the advances in computer science, this major difficulty has been overcome using numerical methods and simulation. For applied statisticians, the real breakthrough was the paper by Geman and Geman (1984). Since then, a new class of methods has emerged, generally called Markov Chain

Monte Carlo methods (MCMC), which are designed to solve specific applied Bayesian problems. For general introductions to Bayesian data analysis and MCMC algorithms see Gelman, Carlin, Stern and Rubin (1995) or Carlin and Louis (1996).

Bayesian methods have been used for statistical analysis of road accident data in the last two decades. The approach was *empirical*, either nonparametric, making use of Robbins' formula as described in Chapter 2 (Robbins, 1955) or parametric, estimating the parameters from the marginal likelihood of those parameters (Morris, 1983; Maritz and Lwin, 1989; Carlin and Louis, 1996). However, in this part of the thesis a fully Bayesian approach is taken and the application of MCMC methods seems to be the only computational solution available. Generalized linear models with random effects are developed for road accident frequencies. The models are hierarchically specified in several stages, assuming that the parameters of probability distributions are random variables with some other probability distributions, up to the last level of hierarchy where all parameters are known. These models can become quite complicated and the level of complexity is substantially increased when multiple response models are considered. The estimation process is in this case very difficult and computational problems are in abundance. MCMC methods, Gibbs sampling in particular, offer a good solution for computational problems and they will be applied in Chapter 9. A good starting point on modelling based on a Gibbs sampling approach can be found in Zeger and Karim (1991). Various other hierarchical Bayesian examples are described in

Gilks, Richardson and Spiegelhalter (1996).

This chapter focuses on models of counts with particular emphasis on practical applications regarding accident frequencies on road networks. There are two problems investigated. Firstly, fully Bayesian models with univariate response are investigated. These are models based on compound Poisson distributions and they are discussed in terms of theoretical improvements and interpretability. The Markov Chain Monte Carlo methodology is explained using a Poisson-gamma model and a Poisson-log normal model. A Poisson-double exponential model is used as an unusual compound Poisson model and all three models are compared on a set of data by the Deviance Information Criterion (Spiegelhalter, Best and Carlin, 1998).

Secondly, the hierarchical Bayesian modelling process is explained in the context of developing two classes of models for multiple response counts: hierarchical Poisson-regression models with random effects and multivariate Poisson-log normal models. Both classes are multiple response models. They are very complex and MCMC methods, employing Gibbs sampling and the Metropolis-Hastings algorithm overcome computational difficulties. The Deviance Information Complexity criterion (DIC) is used in Chapter 9 to compare the fit of 11 models and to choose a small set of good fitting models.

8.2 Univariate Hierarchical Models of Counts

Suppose that for N units (sites) accident counts Y_k, with $k = 1, 2, \ldots, N$ are observed over a fixed time period. The modelling process starts with the assumption that

$$Y_k \overset{ind}{\sim} \text{Pois}(\lambda_k) \text{ for all } k = 1, 2, \ldots, N.$$

This model is not very useful because it is saturated. To improve it, the unobserved parameters λ are modelled as random quantities from the same distribution G,

$$\lambda_k \overset{iid}{\sim} G.$$

The next step is to make some specific distributional assumptions about the prior distribution G.

8.2.1 Choice of the form of prior

Historically, the choice of a suitable parametric class was often governed by mathematical convenience because, until software was widely available, statisticians were restricted to closed analytical calculation. In a Bayesian context, it was helpful to consider the density g of G to be a conjugate distribution of the likelihood distribution.

Therefore, when $Y_k \overset{ind}{\sim} \text{Pois}(\lambda_k)$, the gamma distribution with probability

distribution function

$$g(x \mid \alpha, \beta) = \text{gamma}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

with $\alpha > 0, \beta > 0$, was very convenient. This yields the marginal distribution of the observed counts as the negative binomial distribution

$$p(y \mid \alpha, \beta) = \frac{\Gamma(\alpha + y)}{y!\Gamma(\alpha)} \left(\frac{1}{1+\beta}\right)^y \left(\frac{\beta}{1+\beta}\right)^\alpha \tag{8.1}$$

with $y = 0, 1, 2, \ldots$ as already seen in equation (2.9) in Chapter 2. All that needs to be done is to estimate somehow the hyper-parameters α and β.

This procedure has become standard in modelling count frequencies in the social sciences. Using a negative binomial model seems more appropriate than using a simple Poisson model. The negative binomial distribution is here the result of compounding the Poisson distribution with a gamma distribution. Nevertheless, the parametric distribution G can be any other distribution with non-negative support.

A log normal distribution, for instance, is a possible alternative,

$$g(x) = \text{logN}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2\sigma^2}(\log x - \mu)^2\right),$$

where $\mu \in \mathbf{R}, \sigma^2 > 0, x > 0$.

But now the marginal distribution of Y cannot be calculated in closed form

$$\begin{aligned} p(y|\mu,\sigma^2) &= \int \text{Pois}(y|\lambda)\text{logN}(\lambda|\mu,\sigma^2)d\lambda \\ &= \int \frac{\lambda^y e^{-\lambda}}{y!}\frac{1}{\sqrt{2\pi}\sigma\lambda}\exp\left(-\frac{1}{2\sigma^2}\left[(\log\lambda-\mu)^2\right]\right)d\lambda \\ &\propto \int \lambda^{y-1+\frac{\mu}{\sigma^2}}\exp\left(-\lambda-\frac{1}{2\sigma^2}\left[(\log\lambda)^2\right]\right)d\lambda \end{aligned} \tag{8.2}$$

where "$\propto$" means equality up to a normalizing factor, a convention followed everywhere in this thesis. The last integral cannot be expressed in closed form. It is possible to estimate the parameters of this compound distribution either by moment estimators or maximum likelihood estimators (Shaban, 1988). The MLE estimates require numerical integration techniques. Not very much is known about the properties of MLE estimators for the Poisson-log normal distribution, whether they are unique or not or under what conditions. The computational problems are further complicated when regression terms are involved and where multiple response variables are investigated.

8.2.2 A fully Bayesian approach

However, Markov Chain Monte Carlo methods are designed specifically for situations like this. Under a fully Bayesian framework, some further prior distributions for the hyper-parameters μ and σ^2 have to be set up. An initial approach can be based on setting non-informative priors for the parameters, or in other words, not very much is known a priori about these parameters. Non-informative priors are usually very flat, close in a sense to an uniform

distribution over a large range of values. For computational simplicity it is common to assume that

$$\begin{aligned} p(\mu) &= \mathrm{N}(\mu|0, 0.001) \\ p(\tau) &= \mathrm{gamma}(\tau|0.001, 0.001) \end{aligned} \tag{8.3}$$

where $\tau = 1/\sigma^2$.

The Poisson-log normal model is described by

$$\begin{aligned} Y_k|\lambda_k &\overset{ind}{\sim} \mathrm{Pois}(\lambda_k), \quad \text{for all } k = 1, 2, \ldots, N \\ \lambda_k|\mu, \tau &\overset{iid}{\sim} \mathrm{logN}(\mu, \tau) \\ \mu &\sim \mathrm{N}(0, 0.001) \\ \tau &\sim \mathrm{gamma}(0.001, 0.001) \end{aligned} \tag{8.4}$$

The parameterisation of the normal distribution and of the log normal distribution is not in classical form, the second parameter is the inverse of the variance, also called *precision*. Therefore a very small precision means a very large variance. The actual value of the mean is not important when the variance is so large.

Bayes theorem provides the posterior distribution calculated as

$$p(\lambda, \mu, \tau|y) \propto p(y|\lambda, \mu, \tau)p(\lambda, \mu, \tau). \tag{8.5}$$

where λ and y represent vectors.

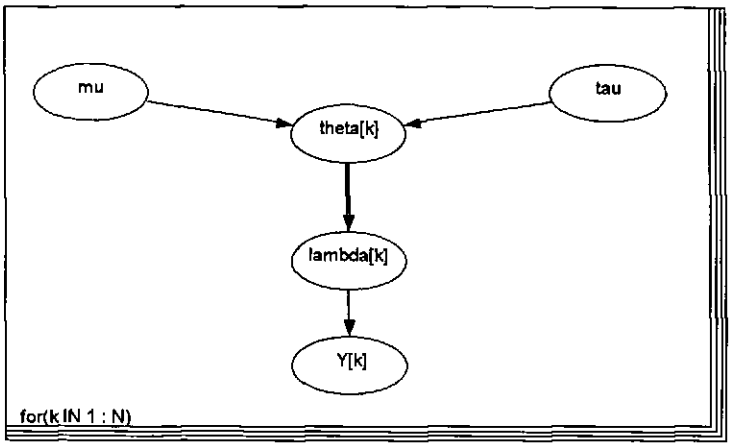


Figure 8.1: Directed graphical model for a mixed Poisson-log normal model

The conditional independencies between the quantities involved, observed data and unobserved parameters are very important. They are used for simplifying the mathematical calculations and to represent mathematically scientific assumptions made before the actual statistical modelling exercise. The best way to communicate these relationships is via a directed graphical model. For example the Poisson-log normal model is based on the graphical model in Figure 8.1. This graph is similar to a directed acyclic graph. In order to be able to define a joint distribution over this type of graph, the graph must be *acyclic*, that is not containing directed cycles. By analogy with chain graphs described in Chapter 3, a *directed local Markov property* can be defined, stating that any vertex v is independent of all vertices that are not descendants of v, given its parents pa[v] (Frydenberg, 1990). No positivity requirement is necessary to prove that this property is equivalent to assuming that the joint distribution

of all quantities V factorizes as

$$p(V) = \prod_{v \in V} p(v \mid \text{pa}[v]).$$

It follows that, in order to specify the joint distribution $p(V)$, only the parent-child distributions need to be provided. In WinBUGS, there are two types of arrows, a normal type corresponding to stochastic relationships and hollow type, corresponding to deterministic functions, that is logical nodes. For reading the conditional independencies only the first type should be considered so the graphical model should be collapsed over all logical variables before attempting to read any conditional independence between the stochastic variables.

The conditional independencies are easy to read directly on the graph

$$Y_k \perp\!\!\!\perp \tau | \lambda_k, \text{ for all } k = 1, 2, \ldots, N$$

$$Y_k \perp\!\!\!\perp \mu | \lambda_k, \text{ for all } k = 1, 2, \ldots, N.$$

The equation (8.5) can then be simplified as

$$\begin{aligned} p(\lambda, \mu, \tau | y) &\propto p(y|\lambda) p(\lambda|\mu, \tau) p(\mu, \tau) \\ &\propto p(y|\lambda) p(\lambda|\mu, \tau) p(\mu) p(\tau) \end{aligned} \tag{8.6}$$

This means that

$$
\begin{aligned}
p(\lambda, \mu, \tau|y) \quad &\propto \quad \prod_{k=1}^{N} \text{Pois}(y_k|\lambda_k) \prod_{k=1}^{N} \text{logN}(\lambda_k|\mu, \tau) \qquad (8.7)\\
&\quad \times N(\mu|0, 0.001)\text{gamma}(\tau|0.001, 0.001)\\
&\propto \quad \prod_{k=1}^{N} \frac{\lambda_k^{y_k} e^{-\lambda_k}}{y_k!} \prod_{k=1}^{N} \frac{\sqrt{\tau}}{\lambda_k} \exp\left(-\frac{\tau}{2}(\log \lambda_k - \mu)^2\right)\\
&\quad \times e^{-\frac{0.001}{2}\mu^2} \tau^{0.001-1} e^{-0.001\tau}\\
&\propto \quad \left[\tau^{\frac{N}{2}} \prod_{k=1}^{k=N} \lambda_k^{y_k-1} e^{-\lambda_k} e^{-\frac{\tau}{2}(\log \lambda_k - \mu)^2}\right] \times\\
&\quad \times e^{-\frac{0.001}{2}\mu^2} \tau^{0.001-1} e^{-0.001\tau}.
\end{aligned}
$$

The joint posterior distribution of all parameters of interest cannot be simplified further. Markov Chain Monte Carlo methods overcomes the lack of closed form analytical methods by a simple and brilliant idea. Denoting by φ all parameters of interest, taking values in a sample space Φ, a Markov chain is simulated with the space state Φ and whose equilibrium distribution is exactly $p(\varphi|y)$, the target distribution. So when a sample from $p(\varphi|y)$ cannot be simulated directly it might be possible to simulate a Markov chain with the properties just described and after a sufficient number of iterations, having some confidence that it has become stationary, any sample from the stationary part of the Markov chain is a (dependent) sample from $p(\varphi|y)$. Methods for simulating a Markov chain with all these properties have been identified and depend on the type of model investigated. The most famous method of sampling is *Gibbs sampling*. This algorithm starts by calculating all conditional distributions of separate parameters, or block of parameters where appropri-

ate, conditioning on everything else. For the above model it follows easily that

$$p(\lambda, \mu, \tau \mid y) \propto \left[\tau^{\frac{N}{2}} \prod_{k=1}^{k=N} \lambda_k^{y_k - 1} e^{-\lambda_k} e^{-\frac{\tau}{2}(\log \lambda_k - \mu)^2}\right] e^{-\frac{0.001}{2}\mu^2} \tau^{0.001-1} e^{-0.001\tau}.$$

The conditional densities of separate parameters (possibly vectors) are calculated by retaining only those terms in the above product that are necessary. For example to calculate the conditional density $p(\lambda_k \mid y, \mu, \tau)$ only the factors containing λ_k are retained, everything else being considered as a part of the normalizing constant, so for every site k

$$p(\lambda_k | y, \mu, \tau) \propto \lambda_k^{y_k - 1 + \mu\tau} e^{-\lambda_k - \frac{\tau}{2}\left[(\log \lambda_k)^2\right]} \tag{8.8}$$

$$p(\mu | y, \lambda, \tau) \propto e^{-\frac{0.001}{2}\mu^2 - \frac{N\tau}{2}\mu^2} \prod_{k=1}^{N} \lambda_k^{\mu\tau} \tag{8.9}$$

$$\begin{aligned} p(\tau | y, \lambda, \mu) &\propto \tau^{\frac{N}{2} + 0.001 - 1} e^{-\tau\left[0.001 + \frac{1}{2}\sum_{k=1}^{N}(\log \lambda_k - \mu)^2\right]} \\ &\propto \text{gamma}\left(\tau \mid \frac{N}{2} + 0.001, 0.001 + \frac{1}{2}\sum_{k=1}^{N}(\log \lambda_k - \mu)^2\right) \end{aligned} \tag{8.10}$$

Starting from some arbitrary points $(\lambda^{(0)}, \mu^{(0)}, \tau^{(0)})$, the Gibbs sampler goes through the following scheme

1. Draw $\lambda_k^{(1)} \sim p(\lambda_k | y, \mu^{(0)}, \tau^{(0)})$, for all $k = 1, 2, \ldots, N$.

2. Draw $\mu^{(1)} \sim p(\mu | y, \lambda^{(1)}, \tau^{(0)})$

3. Draw $\tau^{(1)} \sim p(\tau|y, \lambda^{(1)}, \mu^{(1)})$.

If $(\lambda^{(t)}, \mu^{(t)}, \tau^{(t)}) \mid y$ is the Markov chain resulting from the Gibbs sampler described above then it can be proved under appropriate regularity conditions that $(\lambda^{(t)}, \mu^{(t)}, \tau^{(t)})|y \xrightarrow{d} (\lambda, \mu, \tau)|y \sim p(\lambda, \mu, \tau|y)$ as $t \to \infty$. For a proof and a general description of the conditions under which this theorem is true see Besag (1974), Geman and Geman (1984), Roberts and Smith (1993).

The hierarchical specification of the Poisson-log normal model, equation (8.4), can be followed for the Poisson-gamma model in a similar manner

$$\begin{aligned} Y_k \mid \lambda_k &\overset{ind}{\sim} \text{Pois}(\lambda_k) \\ \lambda_k \mid \alpha, \beta &\overset{iid}{\sim} \text{gamma}(\alpha, \beta) \\ \alpha &\sim \text{logN}(0, 0.0001) \\ \beta &\sim \text{gamma}(0.001, 0.001) \end{aligned} \tag{8.11}$$

The directed graphical model describing the conditional independencies is given in Figure 8.2. This is a full Bayesian model as opposed to an empirical Bayesian model which, instead of setting hyper-priors for the parameters α and β, estimates them from the data. As above, in order to be able to simulate from the joint posterior density $p(\lambda, \alpha, \beta|y)$, the conditional densities are first calculated. From the model assumptions it follows that

$$\begin{aligned} p(\lambda, \alpha, \beta \mid y) &\propto p(y \mid \lambda)p(\lambda \mid \alpha, \beta)p(\alpha)p(\beta) \\ &\propto \left[\prod_{k=1}^{N} \text{Pois}(y_k \mid \lambda_k)\text{gamma}(\lambda_k \mid \alpha, \beta)\right] \text{logN}(\alpha \mid 0, 0.0001) \end{aligned}$$

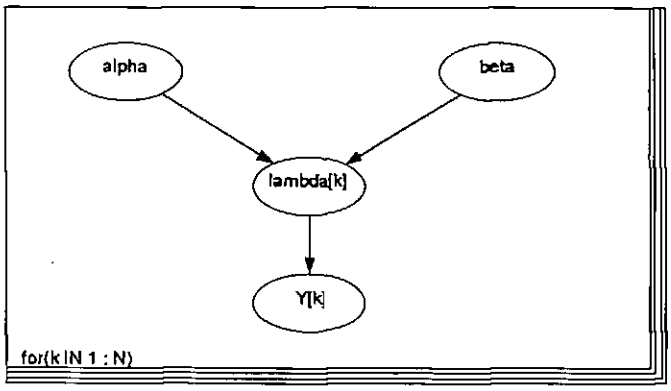


Figure 8.2: Directed graphical model for a mixed Poisson-gamma model

$$\times \text{gamma}(\beta \mid 0.001, 0.001)$$

$$\propto \left[\prod_{k=1}^{N} \lambda_k^{y_k} e^{-\lambda_k} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_k^{\alpha-1} e^{-\beta\lambda_k}\right] \frac{1}{\alpha} e^{-\frac{0.0001}{2}(\log\alpha)^2} \beta^{0.001-1} e^{-0.001\beta}.$$

The conditional densities are calculated now by retaining only the relevant factors from the above product. Therefore

$$p(\lambda_k|y,\alpha,\beta) \propto \lambda_k{}^{y_k+\alpha-1} e^{-\lambda_k(1+\beta)} \tag{8.12}$$

$$\propto \text{gamma}(y_k+\alpha, \beta+1)$$

$$p(\alpha|y,\lambda,\beta) \propto \frac{\beta^{N\alpha}}{[\Gamma(\alpha)]^N} (\Pi_{k=1}^N \lambda_k)^{\alpha-1} \alpha^{-1} \exp\left(-\frac{0.0001}{2}(\log\alpha)^2\right) \tag{8.13}$$

$$p(\beta|y,\lambda,\alpha) \propto \beta^{N\alpha+0.001-1} \exp\left(-\beta(\sum_{k=1}^{N} \lambda_k + 0.001)\right) \tag{8.14}$$

$$\propto \text{gamma}(N\alpha + 0.001, \sum_{k=1}^{N} \lambda_k + 0.001)$$

Compound Poisson models are very useful but cannot provide a good solution for situations when there are several types of counts, possibly correlated.

In Section 8.3 these models are expanded further to allow multiple response counts to be analysed jointly. The computational problems will be more demanding but the same MCMC techniques will be used in a similar manner to solve these problems.

Let see now how the inferential process is executed in practice.

8.2.3 Monitoring the convergence and inference

Markov Chain Monte Carlo methods can be prone to serious errors when the convergence is very slow. If the simulated Markov chain has not converged to the stationary distribution, the inference can be false. Many papers included in Gilks et al. (1996) emphasize how dangerous MCMC methods can be when the convergence is not monitored. The simulated Markov chain should "forget" its starting point after a sufficient number of iterations and the starting point should not influence the inference process.

Based on this simple idea, the following criterion for monitoring convergence has been proposed (Gelman et al., 1995, Section 11.4). Several parallel sequences started from different initial points are simulated. If convergence is attained then the empirical distribution of each sequence is almost identical to the empirical distribution of the sequence obtained by mixing all the sequences together. If convergence is not reached, the variations within each sequence are smaller than the variation within the mixed sequence. By analogy with the analysis of variance, for each parameter of interest, the within-sequence variance W and the between-sequence variance B are calculated, and then

used to estimate the variance of the parameter of interest in the stationary distribution.

Suppose there are m parallel sequences (simulated Markov chains) each with n values, and denote the parameter of interest by ϕ. Denote by $\overline{\phi}_i$ the sample mean and by S_i^2 the sample variance of the ith sequence. If $\overline{\phi} = \frac{1}{m}\sum_{i=1}^{m}\overline{\phi}_i$ then the between-sequence variance is

$$B = \frac{n}{m-1}\sum_{i=1}^{m}(\overline{\phi}_i - \overline{\phi})^2$$

and the within-sequence variance is

$$W = \frac{1}{m}\sum_{i=1}^{m} S_i^2.$$

Under the assumption of stationarity of the simulated Markov chain,

$$\widehat{\text{var}}(\phi) = \frac{n-1}{n}W + \frac{1}{n}B$$

is an unbiased estimate of the variance of ϕ. If the chain has not yet converged then it overestimates the variance; then each sequence has less variability than the mixed sequence, so W underestimates the variance of ϕ. When stationarity is reached both $\widehat{\text{var}}(\phi)$ and W estimate $\text{var}(\phi)$. Gelman et al. (1995) proposed using

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{\text{var}}(\phi)}{W}} \tag{8.15}$$

known as the Gelman-Rubin statistic, as a tool to monitor convergence. When

the simulated Markov chain converges to the stationary distribution then the $\sqrt{\hat{R}}$ decreases to 1. In practice, a value of $\hat{R}$ less than 1.1 indicates convergence.

The program WinBUGS1.2 allows an easy simulation of several parallel chains simultaneously so convergence can be also checked by looking at the dynamic plots of the parameters monitored against iteration number. The Gelman-Rubin convergence statistic as improved by Brooks and Gelman (1998) is calculated in this program. It should be noted that no diagnostic tool can be considered a "proof" of convergence of a MCMC algorithm because it is feasible to use only a finite sample of the chain. However, these monitoring tools help avoiding cases where the mixing is slow and the convergence is unconfirmed. Another recommended practical point (Carlin and Louis, 1996; Gelman et al., 1995) is to simulate several chains starting from dispersed initial points.

At this point, having a sample from the joint posterior distribution $p(\varphi|y)$, any summary inferences (means, medians, quartiles, credible intervals, modes, ranks, density estimation), or predictions of future observations, can be provided.

8.2.4 Residual examination

The particular choice of a model or of a list of models should be checked by comparing the observed statistics with the expectations of these statistics as given by the models. A simple way to check the fit of a model is to consider the residuals $y_k - \mathrm{E}(Y_k)$ or even better, the standardised residuals

$(y_k - \mathrm{E}(Y_k))/\sqrt{\mathrm{var}(Y_k)}$. Large residuals indicate observations that are unlikely to be provided by the probabilistic model proposed. If $Y_k \sim \mathrm{Pois}(\lambda_k)$ it follows that $\mathrm{E}(Y_k) = \mathrm{var}(Y_k) = \lambda_k$. Estimating the unobserved quantity λ_k by the posterior mean $\mathrm{E}(\lambda_k \mid y)$, where y denotes all the data under study, the standardised residual in this case is $(y_k - \mathrm{E}(\lambda_k \mid y))/\sqrt{\mathrm{E}(\lambda_k \mid y)}$.

Another equivalent way to look at the fit of the model is to plot the predicted values $\mathrm{E}(\lambda_k \mid y)$ against the observed values y_k. A good fit would have the points evenly scattered around the line with a 45 degrees slope. This idea will be exploited in Chapter 9 to compare the fit of two hierarchical Bayesian models.

8.2.5 Deviance Information Criterion

Another method to check the fit of a model was proposed by Dempster (1974). It is similar to the use of the deviance measure in generalized linear modelling (McCullagh and Nelder, 1989) but, being in a Bayesian framework, it is the posterior distribution of the log-likelihood of the observed data that is examined.

If the model is given by the data Y and parameters $\varphi = (\theta, \psi)$, the joint distribution can be generally factorised

$$p(y, \varphi) = p(y|\theta)p(\theta|\psi)p(\psi).$$

The fit of the model is directly influenced by the parameters θ, because they

affect directly the observed data y. The models are compared using the posterior distribution of

$$D(\theta) = -2 \log p(y|\theta).$$

The quantity $D(\theta)$ is called *Bayesian deviance* (Spiegelhalter, Best and Carlin, 1998). The posterior distribution of $D(\theta)$ is calculated using $p(\theta|y) \propto p(y|\theta)p(\theta)$ and the fit of a model M is then measured by

$$\overline{D} = \mathrm{E}_{\theta|y}[D] = \int D(\theta)p(\theta|y)d\theta.$$

One aspect that should not be neglected, especially for hierarchical models, is the number of parameters used. Hierarchical models combined with regression models provide a very good solution to fit sparse data. Typically, hierarchical models have more parameters than data observations. However, these models do not provide a perfect fit. This is because the parameters are structured in several layers in a hierarchical structure and they are not independent parameters like in the classical case. These models allow a better description of the stochastic machinery that is assumed to generate the data. The parameters are considered random variables. Thus, the parameters in the second layer are used just to describe the probability distributions of the parameters in the first layer.

Models with large number of parameters should be penalised in the same way the Akaike information criterion (Akaike, 1973) does for regression or log-linear models. The effective number of parameters p_D is a measure of

complexity of the model and is defined by

$$\begin{aligned} p_D &= \overline{D} - D\left(\mathrm{E}_{\theta|y}[\theta]\right) \\ &= \overline{D} - D(\overline{\theta}) \end{aligned}$$

which means the posterior expectation of the Bayesian deviance minus the Bayesian deviance calculated by replacing θ with their posterior expectations $\overline{\theta}$.

The Deviance Information Criterion puts these two measures together

$$\mathrm{DIC} = \overline{D} + p_D$$

and this new measure allows the comparison of arbitrarily complex models. DIC is a measure of fit together with a measure of the effective number of parameters, based on the posterior distribution of the log-likelihood under each model. It was shown, Spiegelhalter, Best and Carlin (1998), that this criterion is a natural generalisation of Akaike's Information Criterion.

Another advantage of using this tool is that $\overline{D}$ and p_D are easy to compute from a MCMC output analysis. Both the Bayesian deviance $D(\theta)$ and parameters θ are monitored during an MCMC and $\overline{D}$ equals the sample mean of the simulated values of $D(\theta)$, while p_D is $\overline{D}$ minus $D(\theta)$ calculated using the sample means $\overline{\theta}$. The models with smaller DIC are preferred.

For a Poisson model, $Y_k \sim \mathrm{Pois}(\lambda_k)$, where the unknown parameters are

the unobserved means λ_k, the scaled deviance is

$$D_S(\theta) = 2\sum_k \left[y_k \log \frac{y_k}{e^{\theta_k}} - (y_k - e^{\theta_k}) \right] \tag{8.16}$$

where $\log \lambda_k = \theta_k$. The Bayesian deviance is obtained by retaining only those terms that depend on θ. The other terms depending only on data do not affect the comparison of different models so they can be left out.

This criterion can be easily calculated in a MCMC analysis, no further calculations being required outside the MCMC output. DIC will be used for model comparison in Section 8.2.7 of this Chapter and in Chapter 9. It should be noted that DIC is not recommended to select a unique model (Spiegelhalter, Best and Carlin, 1998). A unique model should be selected using background knowledge.

8.2.6 Global goodness-of-fit tests based on Bayesian p-values

A compromise between Bayesian and frequentist model checking procedure was introduced by Gelman et al. (1995) and it is described in this section. The discrepancy between the model under scrutiny and the data is measured by a test quantity $T(y, \theta)$, which is a scalar summary of parameters, jointly denoted by θ, and data, jointly denoted by y. In classical statistics, θ is considered known or estimated, and the fit of the data can be measured by the tail-area probability, called the P-value. Then the test statistic depends

only on the observed data y and the P-value is calculated as

$$\Pr(\mathcal{T}(y^{\text{rep}}) \geq \mathcal{T}(y) \mid \theta)$$

where y^{rep} is a replicated set of data, a hypothetical future value of y if the conditions that produced data y are unchanged. Therefore, the probability in calculating the P-value is taken over the distribution of y^{rep} with θ known. An estimate of θ is used in general to calculate this probability.

In a Bayesian framework, point estimates of the parameters θ are not needed. Instead, the fit of the model is measured by comparing the observed data y with the posterior predictive distribution. The test quantity $\mathcal{T}$ depends on the data y and the parameters θ as well, and it is calculated over a sample from the posterior distribution of θ. The P-value is called *Bayesian P-value* and is defined as the probability that the replicated data y^{rep} has a test $\mathcal{T}$ more extreme than the test calculated for the observed data y

$$p_{\mathcal{T}} = \Pr(\mathcal{T}(y^{\text{rep}}, \theta) \geq \mathcal{T}(y, \theta) \mid y). \tag{8.17}$$

A subtle difference is that a Bayesian P-value is conditioned over the data y and not over the parameters θ.

For applications, for each value θ_j, of a sample of size q from $p(\theta \mid y)$, a value for y_j^{rep} is simulated from the posterior predictive distribution. The Bayesian P-value is easily calculated as the proportion of these q draws for which the $\mathcal{T}(y_j^{\text{rep}}, \theta_j) \geq \mathcal{T}(y, \theta_j)$, where $j = 1, 2, \ldots q$. A set of data with a

very small or very large Bayesian P-value provides evidence against the model. However, this does not mean that a single good value qualifies a model as being very good. Other aspects of the models investigated, such as those discussed in the previous two sections, may help in making better decisions regarding model selection and criticism.

A discrepancy measure that will be used for hierarchical multiple response models in Chapter 9 is the χ^2 discrepancy

$$\mathcal{T}(y,\theta) = \sum_k \frac{(y_k - \mathrm{E}(Y_k \mid \theta))^2}{\mathrm{var}(Y_k \mid \theta)} \tag{8.18}$$

where the sum is taken over all observations.

8.2.7 A comparison between different compound Poisson models

For a given set of data, different distributional specifications for G may lead to different results. Here a gamma distribution is used because analytical calculations are possible in this case. This does not mean that other distributions, such as log normal or even the double exponential cannot be used. MCMC methods can easily accommodate complicated calculations required by these two distributions.

Consider road accident data, described in greater detail later at a disaggregated level in Chapter 9, concerning accidents between 1984 and 1991 on 156 single-carriageway link sites in Kent. Without considering any covariate

information, the following three predictive models are compared

M_1: Poisson-gamma model

$$\begin{aligned} Y_k \mid \lambda_k &\sim \text{Pois}(\lambda_k) \\ \lambda_k \mid a, b &\sim \text{gamma}(a, b) \\ a &\sim \text{Exp}(1) \\ b &\sim \text{gamma}(0.1, 1) \end{aligned}$$

where Exp$(\cdot)$ is the exponential distribution. This model is not exactly the same as the Poisson-gamma model given by equation (8.11). It was chosen because it can be shown (George et al., 1993) that this leads to a posterior for b which is a gamma distribution but leads to a non-standard posterior for a which requires the use of Gibbs sampling;

M_2: Poisson-log normal model

$$\begin{aligned} Y_k \mid \lambda_k &\sim \text{Pois}(\lambda_k) \\ \lambda_k \mid \mu, \tau &\sim \text{logN}(\mu, \tau) \\ \mu &\sim \text{N}(0, 0.0001) \\ \tau &\sim \text{gamma}(0.001, 0.001) \end{aligned}$$

M_3: Poisson-log double exponential model

$$Y_k \mid \lambda_k \sim \text{Pois}(\lambda_k)$$

$$\begin{aligned} \log(\lambda_k) \mid \nu, \tau &\sim \text{DE}(\nu, \tau) \\ \nu &\sim \text{N}(0, 0.0001) \\ \tau &\sim \text{gamma}(0.001, 0.001) \end{aligned}$$

where the double exponential probability density function is

$$f(x \mid \nu, \tau) = \frac{\tau}{2} e^{-\tau|x-\nu|}.$$

The posterior summary for the quantities of interest of each model is given in Table 8.1. The inference is based on a sample of 10000 values after a burn-in period of 20000 iterations. The so called *burn-in* period is the part of the Markov chain simulated before the user is confident that the convergence has been reached. This part of the chain is discarded and a sample is selected from the next part of the chain. The actual modelling in WinBUGS took less than 100 seconds for 10000 iterations on a Pentium II personal computer with 100 MHz. The three models can be compared in terms of fit to the data by the Deviance Information Criterion, (DIC).

Before looking at the results one might expect the Poisson-gamma and Poisson-log normal models to be quite close in terms of fit because they have similar shapes and they have been used in the applied statistical literature as compound distributions for the Poisson distribution. Nothing is known from other studies about the Poisson-log double exponential, so we would not be surprised if the third model did not fit the Kent data well. The quantities

Table 8.1: Posterior calculations for all 3 models compared

Model M_1					
node	mean	sd	2.5%	median	97.5%
a	0.58	0.07	0.46	0.58	0.72
b	0.02	0.003	0.02	0.02	0.03
deviance	151.9	17.57	119.7	151.4	188.6
Model M_2					
node	mean	sd	2.5%	median	97.5%
μ	2.44	0.14	2.17	2.44	2.71
τ	0.41	0.06	0.3	0.40	0.54
deviance	173.2	18.5	139.1	172.7	211.7
Model M_3					
node	mean	sd	2.5%	median	97.5%
ν	2.76	0.12	2.51	2.76	3
τ	0.75	0.70	0.62	0.75	0.90
deviance	159.9	18.04	126.7	158.9	197.3

required for calculating DIC are described in Table 8.2 and it can be easily seen that, for this set of data, the gamma distribution is the most appropriate out of the three compared. It is also surprising that the log double exponential distribution gives better results, for this set of data, than the log normal distribution. One explanation for that might be the shape of the distribution. Having some sites with zero counts, the histogram of the data suggests that a gamma distribution with a shape parameter $a \in (0, 1)$ is appropriate. This is the case indeed and the log double exponential distribution is closer in resambling a gamma distribution of this shape than a log normal distribution.

Table 8.2: DIC calculations for all 3 models compared

Model	$\overline{D}$	$D(\overline{\theta})$	p_D	DIC
M_1	152	11.68	140.32	292.32
M_2	173.2	22.96	150.24	323.44
M_3	159.9	16.89	143.01	302.91

Another way to measure the adequacy of the models is to compare the Pearson residuals calculated as $\frac{y_k - \mathrm{E}(\lambda_k|y)}{\sqrt{\mathrm{E}(\lambda_k|y)}}$ for each site $k = 1, 2, \ldots, 156$. The box plots of Pearson residuals for the three models are presented in Figure 8.3. All three models fit the data very well. However, there are a few points worth mentioning. The Poisson-gamma model tends to give higher estimates than the observed numbers of accidents. The Poisson-log normal model would be the best model if the extreme residuals about -1 were ignored. The sites giving these residuals close to -1 are sites with zero accidents observed. The Poisson-log normal model predicts a mean value around 1 for those sites whereas the Poisson-gamma model predicts values around 0.5, closer to the observed data. Therefore, taking out the sites with zero accidents, it is likely that the Poisson-log normal model outperforms the Poisson-gamma model. The Poisson-log double exponential is a good compromise between the previous two, and according to the DIC criterion better than the Poisson-log normal model, because its predictions for sites with zero accidents are better.

In conclusion, for any Bayesian model, the MCMC modelling process will go through the following stages:

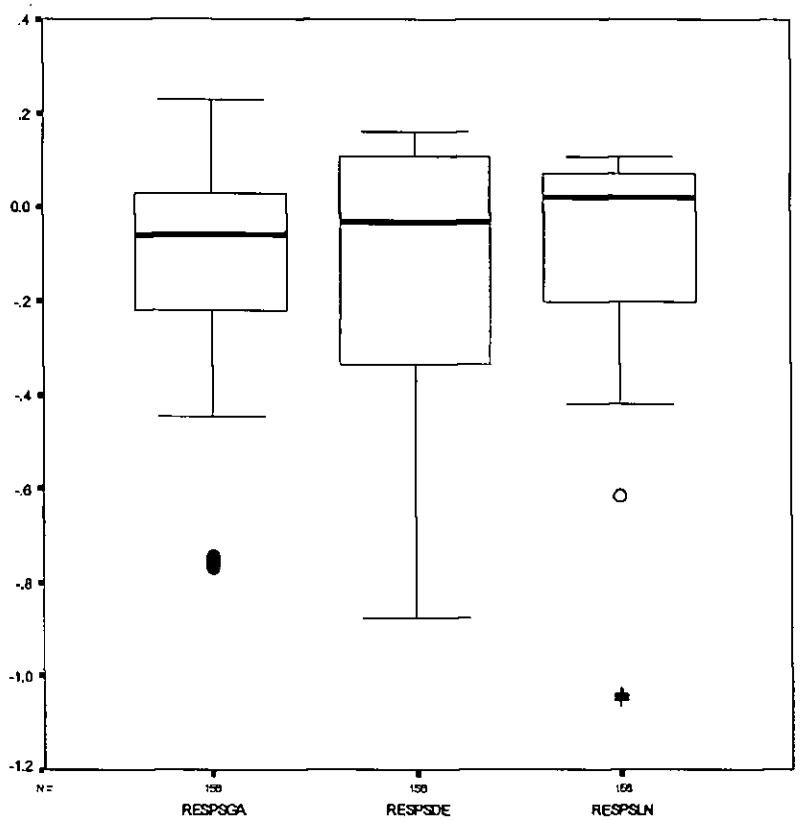


Figure 8.3: Box plots for the models compared; RESPSGA denotes residuals for Poisson-gamma model, RESPSDE denotes residuals for Poisson-log double exponential model and RESPSLN denotes residuals for Poisson-log normal model

1. Start simulating either a single long chain or several parallel chains that will be considered in the end as a mixed chain; it is a good idea to start from some initial values that are not very far from the region in which likelihood is positive.

2. Monitor the convergence of the chain using the Gelman-Rubin convergence tool and the dynamic plots of the values for some of the parameters of interest, and make sure that the chain has become stable.

3. Using the Markov chain output, calculate the Bayesian P-values for one

or more appropriate test criteria and make sure that the model is not rejected by the data, that is the Bayesian P-values are not too small.

4. If everything is fine, any inferential statistics can be calculated now on the same MCMC output.

These four steps are used in MCMC analyses for all models investigated in Chapter 9.

8.3 Multivariate Hierarchical Models of Counts

Techniques for modelling multiple counts jointly have not been extensively developed in the statistical literature, mainly because of the lack of a multivariate discrete distribution that could support complex correlation structures. Bayesian and EB research related to multiple response variables has concentrated on longitudinal studies for clinical trials or biostatistical data (Breslow and Clayton, 1993; Zeger and Karim, 1991; Gilks et al., 1996) and (Carlin and Louis, 1996) or educational studies (Goldstein, 1979). In this section, multiple response models for counts are developed. Several classes of models based on mixing the Poisson distribution with other known distributions are proposed. Some real-world applications involving accident frequencies on the road network are described in detail in Chapter 9. All the models are specified hierarchically and are fully Bayesian.

8.3.1 Hierarchical Poisson-regression models with random effects

Suppose that there are N units of the analysis (for example the sites of a road network). At each unit, M different counts $Y_1, Y_2, \ldots, Y_M$ are recorded (for example the numbers of accidents of different levels of severity in a finite time period). Typically, the counts are modelled with a Poisson likelihood. It is possible that the counts are correlated so multiple response models are desirable. Depending on the information available, the statistical analysis can be based entirely on the observed counts. Alternatively, covariate information (for example environmental characteristics) can be linked to the observed counts through some regression equations.

A framework mean-variance model

The proposed models offer solutions to, at least, two of the well-known problems in modelling counts: overdispersion, and possible correlation between the M counts for each unit. The following mean-variance model can be used as a framework. A similar model has been proposed (Loveday and Jarrett, 1992) at an univariate level for spatially correlated accident frequencies.

For all $k \in \{1, 2, \ldots, N\}$, $i \in \{1, 2, \ldots, M\}$ let Y_{ki} be the count of type i at unit k. Then the assumptions of the model are

$$\begin{aligned} \mathrm{E}(Y_{ki} \mid \lambda_{ki}) &= \mathrm{var}(Y_{ki} \mid \lambda_{ki}) = \lambda_{ki} \qquad (8.19) \\ \text{and} \quad \lambda_{ki} &= \mu_{ki} \exp(X'_{ki}\beta_i). \end{aligned}$$

Here X_{ki} denotes a vector of explanatory variables, each of which can be fixed or random; β_i denotes the vector of the regression coefficients, and μ_{ki} is a random quantity independent of the X_{ki}. In addition, the random variables μ_{ki} (for $k = 1, 2, \ldots, N$) are independently and identically distributed, with

$$\mathrm{E}(\mu_{ki}) = 1, \quad \mathrm{cov}(\mu_{ki}, \mu_{kj}) = \sigma_{ij}$$

for all $k \in \{1, 2, \ldots, N\}$, $i, j \in \{1, 2, \ldots, M\}$. The mean of the random effects μ_{ki} can be always taken equal to 1. If it is not 1 from the beginning then $\mu' = \mu/\mathrm{E}(\mu)$ has mean 1 and the factor $1/\mathrm{E}(\mu)$ can always be included in the regression component. The following proposition illustrates the value of this approach.

Proposition 8.1 *For the mean-variance model described above*

1. $\mathrm{var}(Y_{ki} \mid X) > \mathrm{E}(Y_{ki} \mid X)$

2. $\mathrm{cov}(Y_{ki}, Y_{kj} \mid X) = \exp(X'_{ki}\beta_i + X'_{kj}\beta_j)\sigma_{ij}$

Proof : Because of the independence assumption over units the index k can be dropped to simplify the notation. Moreover, the results can be proved a bit more generally assuming a general positive covariate structure Θ_{ki} instead of $\exp(X'_{ki}\beta_i)$ and this will be used below again for simplicity. Using the properties of conditional expectation it follows immediately that

$$\begin{aligned} \mathrm{E}(Y_i \mid \Theta) &= \mathrm{E}_\mu(\mathrm{E}(Y_i \mid \Theta, \mu)) \\ &= \mathrm{E}_\mu(\mu_i \Theta_i) \end{aligned} \tag{8.20}$$

$$= \Theta_i.$$

Similarly,

$$\begin{aligned} \mathrm{var}(Y_i \mid \Theta) &= \mathrm{E}_\mu(\mathrm{var}(Y_i \mid X, \mu)) + \mathrm{var}_\mu(\mathrm{E}(Y_i \mid \Theta, \mu)) \\ &= \mathrm{E}_\mu(\mu_i \Theta_i) + \mathrm{var}_\mu(\mu_i \Theta_i) \\ &= \mathrm{E}_\mu(\mu_i)\Theta_i + \mathrm{var}_\mu(\mu_i)\Theta_i^2 \\ &= \Theta_i[1 + \sigma_{ii}\Theta_i] \\ &> \Theta_i. \end{aligned} \tag{8.21}$$

Finally,

$$\begin{aligned} \mathrm{cov}(Y_i, Y_j \mid \Theta) &= \mathrm{E}_\mu(\mathrm{cov}(Y_i, Y_j \mid \Theta, \mu)) + \\ &\quad + \mathrm{cov}_\mu(\mathrm{E}(Y_i \mid \Theta, \mu), \mathrm{E}(Y_j \mid \Theta, \mu)) \\ &= \mathrm{cov}_\mu(\mu_i \Theta_i, \mu_j \Theta_j) \\ &= \Theta_i \Theta_j \sigma_{ij}. \end{aligned}$$

Since the Θ_i are positive

$$\begin{aligned} |\mathrm{corr}(Y_{ki}, Y_{kj} \mid \Theta)| &= \frac{\Theta_i \Theta_j |\sigma_{ij}|}{\sqrt{[1 + \sigma_{ii}\Theta_i][1 + \sigma_{jj}\Theta_j]}} \\ &= |\rho_{ij}| \frac{\Theta_i \Theta_j}{\sqrt{\left(\Theta_i + \frac{1}{\sigma_{ii}}\right)\left(\Theta_j + \frac{1}{\sigma_{jj}}\right)}} \end{aligned} \tag{8.22}$$

$$\begin{aligned} &= |\rho_{ij}| \frac{\sqrt{\Theta_i \Theta_j}}{\sqrt{\left(1+\frac{1}{\sigma_{ii}\Theta_i}\right)\left(1+\frac{1}{\sigma_{jj}\Theta_j}\right)}} \\ &< |\rho_{ij}| \sqrt{\Theta_i \Theta_j} \end{aligned}$$

where $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ is the correlation coefficient of the random effects μ_i and μ_j. It can also be seen that $\text{sgn}(\text{cov}(Y_{ki}, Y_{kj} \mid \Theta)) = \text{sgn}(\sigma_{ij})$.

Taking a more specific parametric approach, the following hierarchical models, combining random effects with log-linear regression, are suitable for disentangling the complex structure of multivariate discrete data.

A Poisson-regression model with gamma random effects

This is a hierarchical Bayesian model combined with fixed explanatory variables X, that is specified in three stages. For all sites $k = 1, 2, \ldots, N$ and all types of accident $i = 1, 2, \ldots, M$

$$\begin{aligned} Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}), \qquad (8.23) \\ \log \lambda_{ki} = \theta_{ki} &= \log \mu_{ki} + X'_{ki}\beta_i, \\ \mu_{ki} \mid \alpha_i &\overset{ind}{\sim} \text{gamma}(\alpha_i, \alpha_i) = \text{gamma}\left[1; \frac{1}{\alpha_i}\right], \\ \beta_{ij} &\overset{iid}{\sim} \text{N}(0, 0.001), \text{ and} \\ \alpha_i &\overset{iid}{\sim} \text{gamma}(a, b), \end{aligned}$$

where a, b are known values. The shape and scale parameters of the gamma distribution of μ_{ki} are chosen to be equal in order to ensure that the random effects are distributed with mean equal to 1. The hyperprior for the regression

coefficients (β_{ki}) is non-informative, and typically a and b are chosen to make the hyperprior for the gamma precision parameters (α_i) non-informative also. The random effects (μ_{ki}) account for missing information, uncollected data or unobserved changes in data over the observed period of time. There is no correlation structure for μ_{ki} so this model is a simplification of the general mean-variance model.

The directed graphical model encapsulating the conditional independencies of the above model is illustrated in Figure 8.4. This graphical model is different from the graphical models investigated in the first part of the thesis in connection with road accident characteristics, because some vertices correspond to unobserved quantities. For example the vertex denoted on the graph by lambda[k,i] does not correspond to an observed variable. It is just a variable used for model specification. The regression part of the model is concentrated into the variable denoted on the graph by theta[k,i]. It can be easily seen that, given the values of lambda[k,i], the variable y[k,i] is conditionally independent of all the other variables in the model. This is in agreement with the hierarchical specification of the model given in (8.24). A similar graph was illustrated in the Section 8.2 with a reduced number of vertices also representing observable and unobservable variables. The graph illustrated here is more complex. The joint posterior distribution of all parameters μ, β and α can be calculated as:

$$p(\mu, \beta, \alpha \mid y) \propto p(y \mid \mu, \beta)\, p(\mu \mid \alpha)\, p(\alpha)\, p(\beta)$$

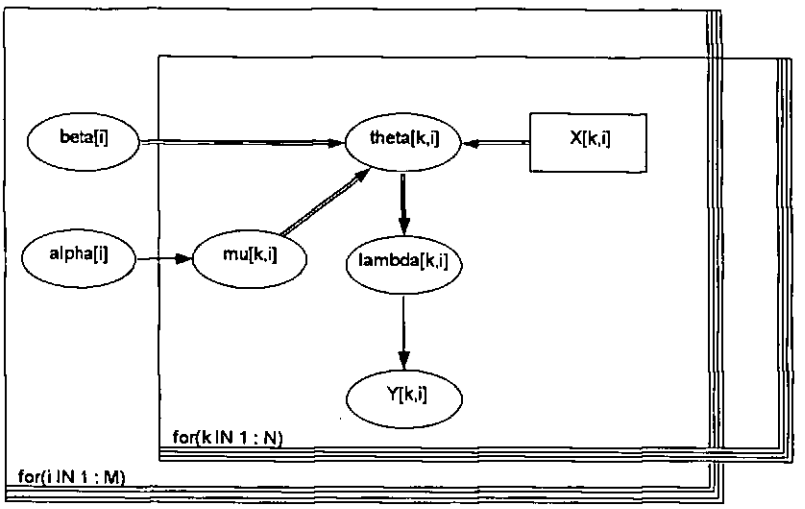


Figure 8.4: Directed graphical model for Poisson-regression model with gamma random effects

$$
\begin{aligned}
&\propto \prod_{k=1}^{N}\prod_{i=1}^{M} \mathrm{Pois}(y_{ki} \mid \mu, \beta) \prod_{k=1}^{N}\prod_{i=1}^{M} \mathrm{gamma}(\mu_{ki} \mid \alpha_i, \alpha_i) \\
&\quad \times \prod_{i=1}^{M} \mathrm{gamma}(\alpha_i \mid a, b) \prod_{i=1}^{M}\prod_{j} \mathrm{N}(\beta_{ij} \mid 0,\ 0.001) \\
&\propto \prod_{k=1}^{N}\prod_{i=1}^{M} \left\{ \mu_{ki}^{y_{ki}} e^{y_{ki} X'_{ki}\beta_i} e^{-\mu_{ki} e^{X'_{ki}\beta_i}} \cdot \frac{\alpha_i^{\alpha_i}}{\Gamma(\alpha_i)} \mu_{ki}^{\alpha_i - 1} e^{-\alpha_i \mu_{ki}} \right\} \times \\
&\quad \times \prod_{i=1}^{M} \alpha_i^{a-1} e^{-b\alpha_i} \prod_{i=1}^{M}\prod_{j} e^{-\frac{0.001}{2}\beta_{ij}^2}
\end{aligned}
\tag{8.24}
$$

If $\dim \beta$ denotes the number of regression coefficients used then there are $MN + M + \dim \beta$ parameters. The model is very complex and it is not possible to simulate directly a sample from $p(\mu, \beta, \alpha \mid y)$. Again the Gibbs sampler is a simple, feasible solution, at the cost of computational effort. The conditional distributions required are

$$
p(\mu_{ki} \mid y, \beta, \alpha) \quad \propto \quad \mu_{ki}^{y_{ki}+\alpha_i-1} e^{-\mu_{ki}(\alpha_i + e^{X'_{ki}\beta_i})}
$$

$$= \text{gamma}(\mu_{ki} \mid y_{ki} + \alpha_i, \alpha_i + e^{X'_{ki}\beta_i}) \tag{8.25}$$

$$p(\alpha_i \mid y, \mu, \beta) \propto \frac{\alpha_i^{N\alpha_i}}{(\Gamma(\alpha_i))^N} \left(\prod_{k=1}^{N} \mu_{ki}\right)^{\alpha_i - 1} e^{-\alpha_i\left(\sum_{k=1}^{N} \mu_{ki} + b\right)} \alpha_i^{a-1} \tag{8.26}$$

$$\propto \frac{\alpha_i^{N\alpha_i + a - 1}}{(\Gamma(\alpha_i))^N} e^{-\alpha_i\left(-\sum_{k=1}^{N} \log \mu_{ki} + \sum_{k=1}^{N} \mu_{ki} + b\right)} \tag{8.27}$$

$$p(\beta_i \mid y, \mu, \alpha) \propto e^{\sum_k y_{ki} X'_{ki}\beta_i} e^{-\sum_k \mu_{ki} \exp X'_{ki}\beta_i} e^{-\frac{0.001}{2}\beta'_i\beta_i} \tag{8.28}$$

where we shall use block conditional distribution for all the regression parameters, that is a multivariate normal distribution instead of a set of separate univariate normal distributions will be used for updating the priors.

A Poisson-regression model with multivariate normal random effects

Starting from the previous model several alternative models are possible. For example, instead of a gamma distribution with mean 1, a multivariate M-dimensional normal distribution for the random effects μ might be considered as more appropiate. In addition, other hyper-priors are required. The model is given by

$$\begin{aligned}
Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}) \qquad (8.29)\\
\log \lambda_{ki} &= \mu_{ki} + X'_{ki}\beta_i \\
(\mu_{ki})_{i=1,2,\ldots,M} \mid T &\overset{iid}{\sim} \text{N}_M(0_M, T) \\
\beta_{ij} &\overset{iid}{\sim} \text{N}(0,\ 0.001) \\
T &\sim \text{Wishart}(R, p)
\end{aligned}$$

where $\mathrm{N}_M(0_M, T)$ is the M-dimensional multivariate normal distribution with mean vector having all elements equal to 0 and with T the inverse of the covariance matrix, also called the precision matrix. The hyper-prior parameters R and $p \geq M$ are known, usually taking $p = M$ for non-informative priors. The Wishart probability density function, as used in this thesis, is

$$f(\mathbf{X} \mid R, p) \propto |R|^{\frac{p}{2}} |\mathbf{X}|^{\frac{p-M-1}{2}} e^{-\frac{1}{2}\mathrm{Tr}(R\mathbf{X})}$$

where $\mathbf{X}$ is an $M \times M$ symmetric and positive-definite matrix, $p \geq M$ is the degrees of freedom and R is a $M \times M$ symmetric and positive-definite (non-singular) matrix. The Wishart prior is used for the inverse of the covariance matrices of multivariate normal distributions and because $\mathrm{E}(\mathbf{X}) = p(R)^{-1}$, R^{-1} is best interpreted as the expected prior precisions of the random effects μ. Small values of p correspond to vaguer prior distributions and it is recommended (Spiegelhalter, Thomas and Best, 1998) to take $p = M$.

This is a complicated version of the mixed Poisson-gamma model and the differences can be seen easily on the graph in Figure 8.5. The Gibbs sampler requires the knowledge of conditional distributions of the unknown quantities of the target distribution, the posterior joint distribution

$$p(\varphi \mid y) = p(\lambda, \mu, \beta, T \mid y)$$

in this case, and this requirement cannot always be satisfied. For these situations a more general MCMC method, called the *Metropolis-Hastings* algo-

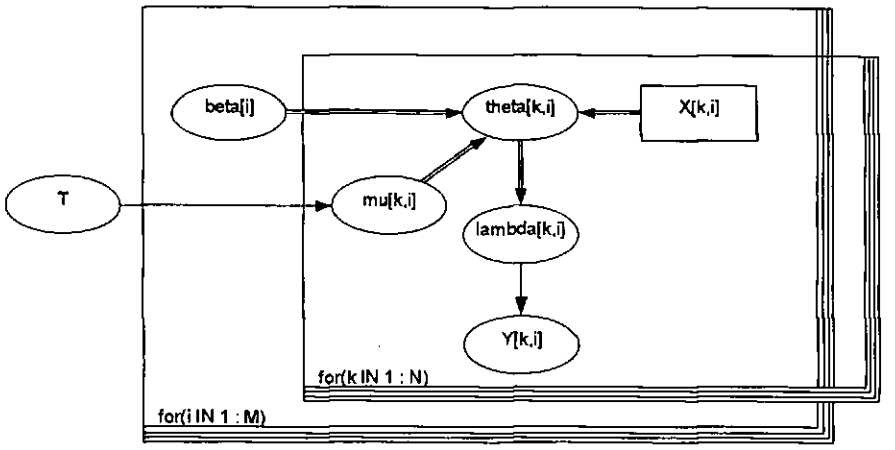


Figure 8.5: Directed graphical model for Poisson-regression model with multivariate normal random effects

rithm, offers a solution: see Carlin and Louis (1996) Section 5.4.3 and Gelman et al. (1995) Section 11.2.. The simulation process moves around in the φ-space according to a candidate probability density $q(\varphi, y)$ from which a draw φ^* is made. Then the jump from the current value to the candidate value φ^* is made with probability r where

$$r = \min\left(1, \frac{p(\varphi^*)q(\varphi \mid \varphi^*)}{p(\varphi)q(\varphi^* \mid \varphi)}\right). \tag{8.30}$$

This is called the acceptance probability and is always equal to 1 for the particular case of Gibbs sampling. The Metropolis-Hastings algorithm is used by default in WinBUGS for situations where Gibbs sampling is not possible. The acceptance rate can be easily monitored and together with other measures on the output it is an indication of the performance of the algorithm. The Poisson-regression model with multivariate normal random effects requires the

use of Metropolis-Hastings algorithm.

A simplified version of the Poisson-regression model with gamma random effects is obtained by approximating the logarithm of the gamma distributed random effects μ_{ki} by a normally distributed quantity b_{ki}. Then

$$\begin{aligned} \log \lambda_{ki} &= b_{ki} + X'_{ki}\beta_i, \\ b_{ki} &\overset{iid}{\sim} \mathrm{N}(0, \tau), \\ \tau &\sim \mathrm{gamma}(0.001, 0.001). \end{aligned} \tag{8.31}$$

Sometimes the random effects b_{ki} can be separated into effects arising from variation among the sites and from variation among accident types

$$\begin{aligned} b_{ki} &= u_k + v_i \\ u_k &\overset{iid}{\sim} \mathrm{N}(0, \tau_u) \\ v_i &\overset{iid}{\sim} \mathrm{N}(0, \tau_v) \\ \tau_u &\sim \mathrm{gamma}(0.001, 0.001) \\ \tau_v &\sim \mathrm{gamma}(0.001, 0.001) \end{aligned} \tag{8.32}$$

All models described in this class are hierarchical and for inference MCMC methods are necessary. These models will be applied and further discussed in Chapter 9.

8.3.2 Bayesian models using the multivariate Poisson-log normal distribution

For multivariate continuous data the multivariate normal distribution provides a sound base for statistical modelling. By contrast, for multivariate counts, there is a lack of discrete multivariate distributions that could play the role of Poisson distribution in the univariate case. A consequence is that sometimes inappropriate methods employing continuous multivariate distributions are proposed in order to support a complex correlation structure. The study of Amis (1996) is an example of a good applied statistical work that can be further improved by applying the hierarchical Bayesian methodology proposed in the previous section. Because the aim of Amis' paper was to investigate accident counts and the associations between accident types and some environmental variables, hierarchical models seem to be perfectly suitable for this. The probability distribution described bellow can also improve another example of applied work involving road accident, done by Salminen and Heiskanen (1997).

In this section, a discrete multivariate distribution is described as a feasible solution for discrete data modelling with multiple responses. The idea is simple, (Aitchison and Ho, 1989), but powerful computational methods are needed to put it into practice. For all $k \in \{1, 2, \ldots, N\}$, $i \in \{1, 2, \ldots, M\}$ we write

$$Y_{ki}|\lambda_{ki} \overset{ind}{\sim} \text{Pois}(\lambda_{ki}), \tag{8.33}$$

$$(\log(\lambda_{k1}), \ldots, \log(\lambda_{kM}))' | \mu, T \overset{iid}{\sim} \mathrm{N}_M(\mu, T) \tag{8.34}$$

where $T = \sum^{-1}$ is the precision matrix. The probability density function of the M-dimensional log normal distribution is

$$p(\lambda|\mu, T) = (2\pi)^{-\frac{M}{2}} (\prod_{i=1}^{M} \lambda_i)^{-1} |T|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\log\lambda - \mu)'T(\log\lambda - \mu)\right)$$

The multivariate Poisson-log normal distribution, that will be denoted by $P\Lambda^M(\mu, T)$, is the mixture of independent Poisson distributions with multivariate log normal distribution for the Poisson means. The probability density function of $P\Lambda^M(\mu, T)$ is exactly the marginal density of Y's conditioned on μ and T only.

$$p(y_1, \ldots, y_M|\mu, T) = \int_{\mathbf{R}_+^M} \prod_{i=1}^{M} \mathrm{Pois}(y_i|\lambda_i) p(\lambda_i|\mu, T) d\lambda_1 \cdots d\lambda_M \tag{8.35}$$

where $y_1, \ldots, y_M = 0, 1, \ldots$. The important moments of this distribution can be easily calculated. If $\Sigma = (\sigma_{ij})$ then

$$\begin{aligned} \mathrm{E}(Y_i) &= \mathrm{E}(\mathrm{E}(Y_i|\lambda_i)) = \mathrm{E}(\lambda_i) \qquad (8.36) \\ &= \exp(\mu_i + \frac{1}{2}\sigma_{ii}) = a_i \\ \mathrm{var}(Y_i) &= \mathrm{E}(\mathrm{var}(Y_i|\lambda_i)) + \mathrm{var}(\mathrm{E}(Y_i|\lambda_i)) \qquad (8.37) \\ &= \mathrm{E}(\lambda_i) + \mathrm{var}(\lambda_i) \\ &= a_i + a_i^2(\exp(\sigma_{ii}) - 1) \end{aligned}$$

$$\begin{aligned} \mathrm{cov}(Y_i, Y_j) &= \mathrm{E}(\mathrm{cov}(Y_i, Y_j|\lambda)) + \mathrm{cov}(\mathrm{E}(Y_i|\lambda_i), \mathrm{E}(Y_j|\lambda_j)) \qquad (8.38) \\ &= \mathrm{cov}(\lambda_i, \lambda_j) = a_i a_j (\exp(\sigma_{ij}) - 1) \end{aligned}$$

Two immediate consequences are that, for each unit $k \in \{1, 2, \ldots, N\}$,

$$\mathrm{var}(Y_{ki}) > \mathrm{E}(Y_{ki})$$

which means that there is overdispersion for the marginal distributions, and

$$|\mathrm{corr}(Y_{ki}, Y_{kj})| < |\mathrm{corr}(\lambda_{ki}, \lambda_{kj})|$$

$$\mathrm{sgn}(\mathrm{corr}(Y_{ki}, Y_{kj})) = \mathrm{sgn}(\mathrm{corr}(\lambda_{ki}, \lambda_{kj}))$$

which are special cases of the results of the mean-variance model. Altogether $\frac{M(M+3)}{2}$ parameters are needed to specify the $P\Lambda^M(\mu, T)$ distribution. Negative and positive correlations are supported by this mixed distribution, which gives it an advantage over other multivariate discrete distributions such as multinomial or negative multinomial. However, the estimation of the parameters is not straightforward. For maximum likelihood estimation, a reparameterization and a mixture of Newton-Raphson and steepest ascent methods are helpful but computationally intensive, (see Aitchison and Ho (1989)).

Here we shall use MCMC methods (Metropolis-Hastings algorithm) to obtain inference summaries about the parameters μ and and T. In a fully Bayesian context, further prior distributions, probably non-informative, are

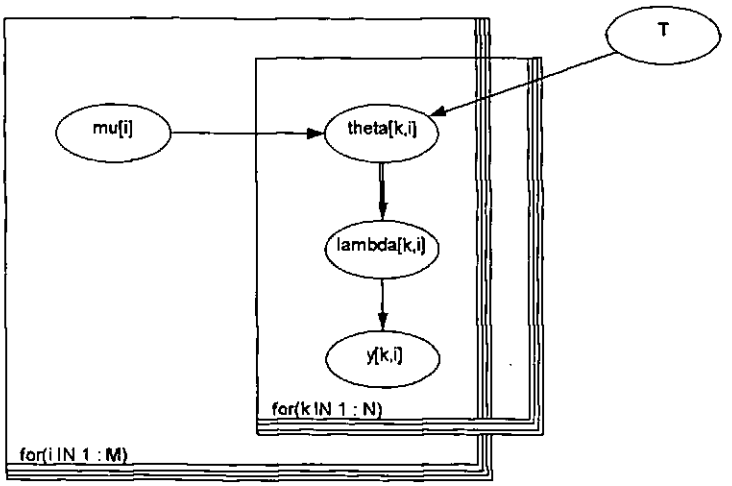


Figure 8.6: Directed graphical model for a multivariate Poisson-log normal model

required for μ and T. The recommended parametric distributions are normal for μ and Wishart for T (Carlin and Louis, 1996; Gelman et al., 1995). Such a model is described by the directed graph in Figure 8.6 and it can be easily seen that it is a straightforward generalisation of the directed graphical model in Figure 8.1 that represented a univariate Poisson-log normal model.

Covariate information can be introduced easily in this model by regression components like $\mu_{ki} = X'_{ki}\beta_i$. The information value of the explanatory variables X_{ki} can be examined by comparing models without regression with models with regression. Many other variations on this model structure are possible by making small changes, like considering that the regression coefficients β do not depend on accident count type or reparametrising $\mu_{ki} = u_k + v_i$, a site random effect and an accident type random effect. Models that have more parameters tend to fit data better. In a Bayesian context, for hierarchi-

cal and regression models, the number of parameters is usually very large but not all are effective. This class of models will be added to the class of models proposed in Section 8.3 and both will be applied and compared on a set of data in Chapter 9.

8.4 Bayesian model selection

In this section a new group of model selection procedure for hierarchical log-linear models is proposed in a Bayesian framework. Other model selection procedures were investigated in the first part of the thesis. The reason why these methods are discussed here is that they employ Gibbs sampling for solving the computational side. The objective of these model selection procedures belong in the first part of the thesis but the solution belongs in this second part.

Since graphical models are log-linear models this method can be used for this subclass as well. The idea on which this model selection algorithm is based is similar to a suggestion of Lindley (1969, Section 5.6) in connection with a classical test of a point null hypothesis. It is a compromise between Bayesian and classical statistics. The significance test at level α is conducted using the credible set, which is roughly equal to the highest posterior region (Carlin and Louis, 1996).

For hierarchical models only the maximal interaction u-terms (the generators of the model as they were described in Chapter 4) need to be specified and,

as will be seen later, they effectively drive the technique proposed here. For log-linear models many selection algorithms have been proposed; see Chapter 4. Some of these methods were applied to the collision-rollover data in Chapter 4 and the Edwards-Havranek method was also applied to Bedfordshire data in Chapter 5. Each of these known methods has a different motivation, but all share the same drawbacks: practical sensitivity to the choice of stopping rule and of initial model; lack of information about the power of different procedures; being able to apply standard distribution theory only for a fixed model; and lack of information about the influence of the model selection procedure on the sampling distribution of the model that is fitted. In addition, some of these algorithms are based on asymptotic distributions of deviance (G^2) or Pearson chi-squared (X^2), which are unreliable when the data is sparse (Kreiner, 1987). Forward selection procedures starting from the mutual independence model are dubious because this model rarely fits the data, and the hypothesis testing of nested models involves models known to fit the data badly.

The idea of this new approach in this section is to overcome these difficulties by avoiding classical hypothesis testing and asymptotic methods. Instead, forward, backward and bidirectional procedures are proposed using fully Bayesian inference for the maximal u-terms eligible for selection (inclusion or elimination). These model selection algorithms can be used for data summarised in contingency tables thus making a straightforward connection with the first part of the thesis. However, the computational side of these algo-

rithms is mainly based on Gibbs sampling, used in this second part of the thesis, motivating the inclusion of these model selection procedures here.

Each stage of the selection methods presented below is in correspondence with an order of interaction and, in any stage, only the maximal u-terms for that order of interaction are tested. The main effects u_k, $k \in V$, are always kept in the models. Other terms might be included in the models because of sampling design specification. For example, if the sampling scheme is a product-multinomial, some interactions terms need to be included in the model without any further testing. The procedures end when there are no maximal eligible u-terms left. The final model is a hierarchical interaction model. If the maximal u-terms correspond to the cliques of the interaction graph then the final hierarchical model is graphical (Whittaker, 1990) and the model is interpretable in terms of conditional independencies.

The idea driving the model selection methods proposed here is to consider the maximal u-terms under scrutiny as random effects and all the other u-terms in the model as fixed effects. For each interaction term u_a (where a is a subset of vertices from V) considered random effect, the following distributional assumptions are made

$$u_a \sim \mathrm{N}(0, \tau_a) \tag{8.39}$$

$$\tau_a = \frac{1}{\sigma_a^2} \sim \mathrm{gamma}(0.001, 0.001) \tag{8.40}$$

and all the other fixed terms have very flat normal priors

$$u_b \sim \mathrm{N}(0, 0.0001)$$

The posterior distribution of the random effects u_a is $p(u_a|n(i) : i \in \mathcal{I})$, and this distribution is used to calculate the equal tail credible set $\mathrm{CS}(u_a)$ for u_a, (see Carlin and Louis (1996, Section 2.3.2)), simply taking the $\alpha/2$ and $(1-\alpha)/2$-quartiles of the posterior distribution $p(u_a|n(i) : i \in \mathcal{I})$. The equal tail credible set is not always equal to the highest posterior density credible set (unless the posterior distribution is symmetric and unimodal) but being just a bit wider it is more convenient for the applied statistician to work with the former. If $0 \notin \mathrm{CS}(u_a)$ then u_a should not be eliminated from the model.

8.4.1 Bayesian forward selection

This procedure starts from the mutual independence model

$$M^{(0)}: \quad \log \lambda_i^{(0)} = u_\emptyset + \sum_{t \in V} u_t. \tag{8.41}$$

In any stage $S^{(j)}$ the inclusion of *each* maximal eligible j-order of interaction terms u_a is tested. If the final model selected at the end of stage $S^{(j-1)}$ is $M^{(j-1)}: \log \lambda_i^{(j-1)}$ then, in the stage $S^{(j)}$, the posterior distributions of each eligible u_a is calculated from the model

$$\log \lambda_i^{(M)} = \log \lambda_i^{(j-1)} + u_a. \tag{8.42}$$

Therefore the credible sets are calculated independently for each u_a. The final model of this stage, $M^{(j)}$, is obtained by adding to $M^{(j-1)}$ all u_a selected for inclusion, that is all u_a with $0 \notin \mathrm{CS}(u_a)$. In the following stages these terms are considered fixed terms.

It is worth mentioning that, in a given stage, maybe not all maximal u-terms are eligible for inclusion. For example, if u_{12} has been eliminated in a previous stage then u_{12a} are automatically excluded because we require the log-linear model to be hierarchical.

8.4.2 Bayesian backward elimination

The procedure starts with the saturated model

$$M^{(S)}: \quad \log \lambda_i = \sum_{a \in 2^V} u_a \tag{8.43}$$

where the sum is over all subsets of V, including the empty set. In any stage $S^{(j)}$, *all* the maximal $(d-j)$-order u-terms, without a higher order relative in the final model of stage $S^{(j-1)}$, are considered *simultaneously* random effects and all the other terms fixed effects. As before, the final model selected at the end of stage $S^{(j)}$ is denoted by $M^{(j-1)}: \log \lambda_i^{(j-1)}$. The posterior distributions of all eligible maximal interaction terms u_a, $a \in A \in 2^V$, are calculated according to the model $M^{(j-1)}$ but the status of the terms $u_a, a \in A$, is changing from fixed effects (stage $S^{(j-1)}$) to random effects (stage $S^{(j)}$).

The credible sets $\mathrm{CS}(u_a)$ are calculated for each of the u_a and only those

terms whose credible sets do not overlap zero are kept in the model. The difference between this procedure and the Bayesian forward selection procedure is that, in a given stage, the credible sets for the eligible u-terms are calculated based on the same log-linear model. Note that, at any given level of interaction excluding trivial cases, there may be lower, higher or the same order of interaction terms that are fixed terms and not eligible.

8.4.3 Bayesian bidirectional selection

A bidirectional procedure combining the above two procedures of Sections 8.4.1 and 8.4.2 can be easily developed. Starting from an initial model, such as the mutual independence model of equation (8.41), a one-stage forward inclusion is performed followed by a one-stage backward elimination. This combined computation is made until all eligible terms are screened.

For any of these three model selection methods, at any stage, the u-terms giving the model under consideration can be partitioned into random effects $u_a, a \in A \in 2^V$, and fixed terms $u_b, b \in B \in 2^V$. If $u = ((u_a)_{a\in A}, (u_b)_{b\in B})$ and $\tau = (\tau_a)_{a\in A}$ are the corresponding precision parameters, the joint posterior distribution of all parameters is

$$\begin{aligned} p(u, \tau|\mathbf{n}) &\propto \prod_{i\in I} p(n(i)|u) \prod_{a\in A} p(u_a|\tau)p(\tau_a) \prod_{b\in B} p(u_b) \qquad (8.44) \\ &\propto \prod_{i\in I} \text{Pois}(n(i)|u) \prod_{a\in A} \text{N}(u_a|0,\tau)\text{gamma}(\tau_a|0.001, 0.001) \\ &\quad \times \prod_{b\in B} \text{N}(u_b|0, 0.0001) \end{aligned}$$

For forward inclusion, $A = \{a\}$ because the inference is done separately for each maximal eligible u_a. Replacing in (8.44) the known densities, the joint posterior density of all quantities, observed and unobserved, is,

$$p(u,\tau|\mathbf{n}) \propto \prod_{i\in\mathcal{I}} \lambda_i(u)^{n(i)} e^{-\lambda_i(u)} \prod_{a\in A} (\tau_a)^{0.001-\frac{1}{2}} e^{-\frac{\tau_a}{2}u_a^2} e^{-0.001\tau_a} \prod_{b\in B} e^{-\frac{0.0001}{2}u_b^2} \quad (8.45)$$

and it is obvious that this expression cannot be manipulated analytically. For example, the marginal posterior density of u_a with $a \in A$ is

$$p(u_a \mid \mathbf{n}) \propto \int p(u,\tau|\mathbf{n}) d\tau du_{A\setminus a} du_B$$

which requires the calculation of a complicated multidimensional integral. Such calculation is impossible to be done in closed form.

However, an advanced Gibbs sampling method can be used to overcome this computational problem (Gilks, 1992) and the analysis can be done in WinBUGS, (see Spiegelhalter, Thomas and Best (1998)).

8.4.4 Applications to road accident tables

Collision-rollover data

It was shown that the 4-dimensional table 3.3 in Chapter 3, summarising the collision-rollover data from Kihlberg et al. (1964), can be safely decomposed into two 3-dimensional subtables ACD and BCD without evoking problems with Simpson's paradox. The model selection procedures used in log-linear

modelling cannot identify a simpler graphical model and the only simpler log-linear model fitting the data well is the no three-way interaction model, for both subtables. In this subsection, forward and backward Bayesian model selection are applied to each subtable. The results are presented in Tables 8.3 and 8.4. It can easily be seen that for the subtable ACD the model of no

Table 8.3: Bayesian model selection for ACD subtable

Forward Model	u-term	mean	CS 2.5%	CS 97.5%
$[AC]$	u_{12}	1.73	1.56	1.90
$[AD]$	u_{13}	1.8	1.64	1.95
$[CD]$	u_{23}	1.45	1.27	1.63
$[ACD]$	u_{123}	0.15	-0.12	0.59
Backward				
$[ACD]$	u_{123}	0.15	-0.12	0.59
$[AC][AD][CD]$	u_{12}	1.38	1.20	1.56
	u_{13}	1.60	1.44	1.76
	u_{23}	1.01	0.82	1.21

three-way interaction is selected both by the forward and backward bayesian model selection criteria. For the second subtable BCD, by forward bayesian selection the model $[BC][CD]$ is selected and by backward elimination the model $[BC][BD][CD]$ is selected. This illustrates the point that forward and backward Bayesian procedures do not necessarily select the same model.

A corner point parameterisation (Bishop et al., 1975) was used for the log-linear expansion, that is all u-terms having at least one index equal to 1 is

Table 8.4: Bayesian model selection for BCD subtable

Forward Model	u-term	mean	CS 2.5%	CS 97.5%
$[BC]$	u_{12}	-0.61	-0.78	-0.43
$[BD]$	u_{13}	0.04	-0.08	0.17
$[CD]$	u_{23}	1.79	1.63	1.95
$[BCD]$	u_{123}	0.15	-0.12	0.59
Backward				
$[BCD]$	u_{123}	0.11	-0.14	0.45
$[BC][BD][CD]$	u_{12}	-0.74	-0.91	-0.56
	u_{13}	0.29	0.11	0.45
	u_{23}	1.83	1.67	2.00

set to 0, and because all the variables are binary the tested u_{ij}-terms are all $u_{ij}[2,2]$, the other values being constrained to zero.

Bedfordshire data 3-dimensional subtable

The collapsibility results discussed in Chapter 6 suggests that it may be worthwhile to analyse the 3-dimensional subtable defined by three variables, accident severity, number of vehicles involved and speed limit. There is no simpler graphical model than the saturated model for this subtable and it was shown in Section 5.2.1, Chapter 5, that the conditional independence structure is worth further exploration. In this subsection, forward and backward bayesian model selection procedures are applied to this small subtable, in an attempt to understand whether a simpler log-linear model can be selected or

Table 8.5: Bayesian forward selection for Bedfordshire data, ANS subtable

Forward Model	u-term	mean	CS 2.5%	CS 97.5%
$[AN]$	$u_{12}[2,2]$	0.12	-0.36	0.78
	$u_{12}[2,3]$	-0.15	-0.69	-0.16
	$u_{12}[3,2]$	0.66	0.11	1.29
	$u_{12}[3,3]$	0.02	-0.40	0.38
$[AS]$	$u_{13}[2,2]$	-1.31	-2.36	-0.45
	$u_{13}[3,2]$	-1.71	-2.76	-0.90
$[NS]$	$u_{23}[2,2]$	-0.09	-0.30	0.06
	$u_{23}[3,2]$	0.85	0.56	1.14
$[ANS]$	$u_{123}[2,2,2]$	0.15	-0.54	0.95
	$u_{123}[2,3,2]$	0.20	-0.53	1.20
	$u_{123}[3,2,2]$	-0.37	-1.24	0.19
	$u_{123}[3,3,2]$	-0.23	-1.30	0.39

not.

The results in Tables 8.5 and 8.6 show that indeed the saturated model cannot easily be simplified. In both approaches, forward or backward, there are two problems. Consider the u_{12} term, where index 1 stands for accident severity A and 2 stands for the number of vehicles N. This u term, which accounts for the pairwise interaction between A and N, would be rejected from the model if the corresponding CS for $u_{12}[3,2]$ overlapped zero. Unfortunately this does not happen and therefore, the interaction between A and N can neither be included or excluded from the model. A similar situation occurs with the u_{23} interaction term, for which 2 values, $u_{23}[2,2]$ and $u_{23}[3,2]$, should be tested. The three-way interaction term is rejected from the model by the backward elimination procedure so the model $[AS][SN][AN]$ remains a

Table 8.6: Bayesian backward elimination for Bedfordshire data, ANS subtable

Backward Model	u-term	mean	CS 2.5%	CS 97.5%
$[ASN]$	$u_{123}[2,2,2]$	0.15	-0.54	0.95
	$u_{123}[2,3,2]$	0.20	-0.53	1.20
	$u_{123}[3,2,2]$	-0.37	-1.24	0.19
	$u_{123}[3,3,2]$	-0.23	-1.30	0.39
$[AN][AS][SN]$	$u_{12}[2,2]$	0.03	-0.46	0.57
	$u_{12}[2,3]$	-0.14	-0.65	0.24
	$u_{12}[3,2]$	0.55	0.04	1.08
	$u_{12}[3,3]$	0.11	-0.26	0.60
	$u_{13}[2,2]$	-1.58	-2.53	-0.78
	$u_{13}[3,2]$	-1.97	-2.9	-1.21
	$u_{23}[2,2]$	-0.06	-0.26	0.09
	$u_{23}[3,2]$	0.86	0.57	1.14

candidate. This is not a graphical model so from the conditional independence point of view it does not reveal any new information. However it can be used for other purposes, like making inference about odds ratios.

The major drawback of this procedure is that each model investigated during the model selection process has to be fitted separately. There is no program available that would make possible an automated implementation.

8.5 Summary

In this chapter two classes of hierarchical Bayesian models have been introduced. The first class was based mainly on mixed Poisson-regression models with random effects. They are all specified hierarchically in three stages and

use the same regression part. The difference between them consists in the distributional assumption for the random effects. A general mean-variance framework model was introduced at the beginning of this chapter, which offers a good solution for accounting for overdispersion and correlation between observed frequencies.

All the models investigated were fully Bayesian and have computational difficulties given by the lack of closed-form analytic inferential methods. The main points of the methodology for applying MCMC techniques, in particular the Gibbs sampler and the Metropolis-Hastings sampler, were pointed out and some relatively simple applications were given. A simple case of three possible compound Poisson distributions for the accident totals on 156 sites in Kent was discussed, the same road accident data that will be investigated at a more disaggregated level in the next chapter.

A new group of model selection procedures for hierarchical log-linear models for contingency tables has also been introduced. The novelty of these procedures consists in being formulated entirely in a Bayesian framework and avoiding classical hypothesis testing.

The emphasis was more theoretical in this chapter, the applications being discussed in greater detail in Chapter 9. The main idea of this chapter is that hierarchical Bayesian models coupled with MCMC techniques offer a statistical modelling solution to a wide range of problems related to analysing complex datasets such as road accident data. Moreover, the multiple response models proposed here open a new area of research and they can be easily adapted to

count data sets from other areas of research.

Chapter 9

Multiple response models for road accident data

9.1 Introduction

In this chapter the techniques introduced in Chapter 8 are applied to a set of road accident data. Several models are fitted and the results are discussed and compared. The ability to model joint responses provides another dimension to statistical modelling of road accidents. It is shown that the ranking of hazardous sites can be improved by looking at several types of accidents simultaneously. The advantage of using MCMC techniques is that the same model output can be used to provide inference on several problems like model selection, goodness-of-fit, ranking the units of the analysis according to different criteria, and so on. This type of analysis is believed to be the first of this kind in the area of statistical modelling of road accident data.

9.1.1 Data analysed

The units of the analysis are 156 single carriageway link sites in Kent and the data includes all accidents between 1984 and 1991. The links are defined as road sections between two major junctions, or between changes in carriageway type (single or dual), or between changes in speed limits. Figure 9.1 shows a map of the relevant part of the Kent road network. The nodes on the road network defining the junctions and the carriageway types were taken from digital maps supplied by Kent County Council and the speed limits were taken from the STATS 19 records. The speed limit plays another important role as a

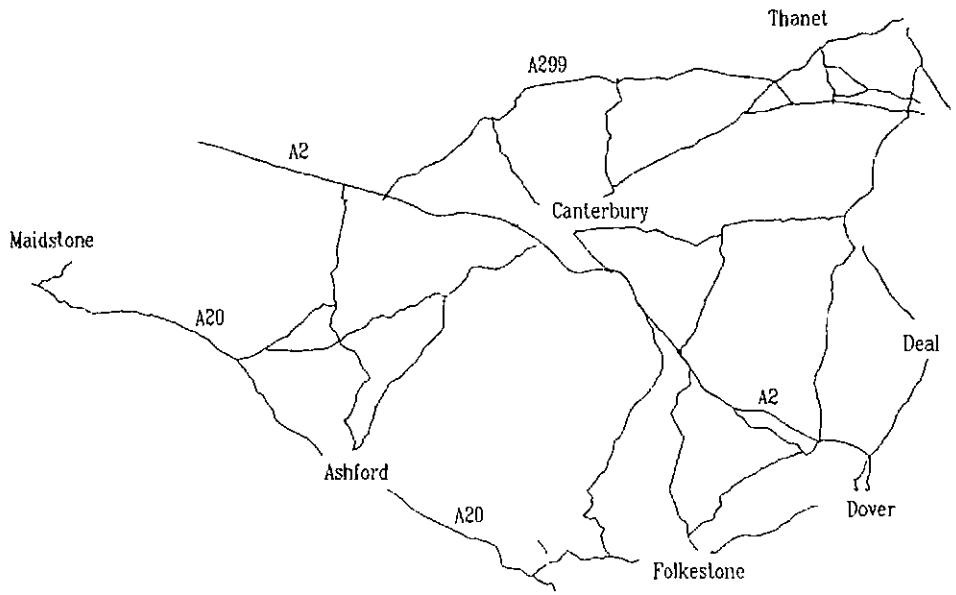


Figure 9.1: Part of Kent road network

proxy for the actual speed. Speed is a variable that is known to have a major impact on the number and severity of accidents (Taylor and Barker, 1994-1995; Tunaru and Jarrett, 1998*a*; Baruya, Finch and Wells, 1999). The other

explanatory variables used in this chapter are link length, in kilometers, and estimated traffic flow (AADT×365), in millions of vehicles per year. These two variables are continuous but speed limit, as it was used here, is dichotomous (40 mph or less, greater than 40 mph) so the interactions between speed limit and the other two were also considered. The original set of data had speed limit with several levels. It was the author's choice to dichotomise this explanatory variable. The traffic flows came from mostly manual counts with some automatic counts. The manual counts can be sparse in both location and time but simple linear regression was used to fill in the missing years and account for some of the variation in individual counts. The estimated traffic flow was averaged over all seven years.

This set of data was provided by the Transport Management Research Centre at Middlesex University which took it from Kent county council for a previous research project. The number of accidents at each site was disaggregated by accident severity, having two levels KSI = fatal or serious and S = slight, and the number of vehicles involved, with two levels, 1 vehicle and 2 or more vehicles. Therefore, there are four accident counts for each site. This further classification of the observed accidents was entirely the author's choice and it was motivated by the direct association between accident severity and the number of vehicles revealed by the graphical models proposed in Chapter 5.

The cross-classification of these two categorical variables gives four possibly correlated groups of observations and the log-linear Poisson regression

equations for each group of accidents might be different. The observed number

Table 9.1: Total number of accidents for each category of accidents

Severity	Number of vehicles involved	Total number of accidents
fatal or serious	1	443
	2 or more	852
slight	1	796
	2 or more	2160

of accidents in each group is given in Table 9.1 and it is also worth pointing out that there are sites with zero accidents for any type of accident and for the total number of accidents as well.

9.2 Hierarchical Poisson-regression models for multiple accidents

This section contains the applied statistical modelling results for the models combining hierarchical Bayesian specification with covariate information. The models reveal qualitative and quantitative relationships between the numbers of road accidents on one side and speed limit, estimated traffic flow and link length on the other. Some parts of this section have been published in Tunaru (1999).

Three road characteristics, speed limit, link length and traffic flow, mea-

sured for each site, are used in the regression equation. The traffic flow was averaged over all years and denoted by Q. Speed limit S was coded -1 for less than 40 mph and $+1$ for less than 60 mph, link length l was transformed on a logarithmic scale to $\log l$, and the same for traffic flows to $\log Q$.

The multiple responses analysed in this paper correspond to the four types of accidents according to severity and the number of vehicles involved. The numbers of fatal or serious accidents with only one vehicle involved are denoted by Y_1, the fatal or serious accidents with two or more vehicles involved are denoted by Y_2, the slight accidents with only one vehicle are denoted by Y_3 and the slight accidents with two or more vehicles by Y_4. A more detailed analysis might consider multiple responses obtained by cross-classifying the accidents according to more than two criteria. For example, pedestrian involvement might be of interest in addition to the criteria used in this paper. Moreover, other explanatory variables can be used in addition to those studied here. The data as provided by the Transport Management Research Centre contained dual carriageway sites as well so an explanatory variable with two levels single-dual would be a natural candidate. However, there were very few dual carriageway sites and a preliminary analysis revealed that it was not worth including those sites.

The explanatory variables were standardised in order to improve the speed with which the simulated Markov chain approach its stationary distribution, as recommended in Spiegelhalter, Thomas and Best (1998). Therefore l^* and Q^*, the standardised values of the logarithms of the link length and estimated link

traffic, were in place of $\log l$ and $\log Q$. This standardisation was calculated by subtracting the sample mean from each value in the sample and then dividing it by the sample standard deviation. The terms accounting for the interactions between speed limit S and link length or link traffic were transformed in a way similar to that used for centring second order terms in polynomial regression. Hence, these terms are given by

$$\begin{aligned} \mathrm{SL}_k &= (S_k - \overline{S})(\log l_k - \overline{\log l}) - \overline{(S_k - \overline{S})(\log l_k - \overline{\log l})} \\ &= S_k \log l_k - \overline{S \log l} - (S_k)\overline{\log l} - (\log l_k)\overline{S} + 2\overline{S}\,\overline{\log l} \end{aligned}$$

and

$$\begin{aligned} \mathrm{ST}_k &= (S_k - \overline{S})(\log Q_k - \overline{\log Q}) - \overline{(S_k - \overline{S})(\log Q_k - \overline{\log Q})} \\ &= S_k \log Q_k - \overline{S \log Q} - (S_k)\overline{\log Q} - (\log Q_k)\overline{S} + 2\overline{S}\,\overline{\log Q} \end{aligned}$$

where the bar indicates the sample mean of the corresponding variable. This transformation helped to reduce the autocorrelation between successive sampled values of the Markov chain. Otherwise the Gibbs sampling algorithm would stay for too many iterations in a small region of the sample space, and it would be necessary to simulate a much larger number of values than usual in order to cover the whole sample space.

9.2.1 A Poisson-regression model with gamma random effects

The model given below in (9.1) will be called (P-ga) and it is a particular case of the Poisson-regression model with gamma random effects defined in Section 8.3. The explanatory variables are specific to the set of data analysed in this chapter. For all sites $k = 1, 2, \ldots, 156$ and accident groups $i = 1, 2, 3, 4$

$$Y_{ki} \mid \lambda_{ki} \overset{ind}{\sim} \text{Pois}(\lambda_{ki}), \tag{9.1}$$

where

$$\log(\lambda_{ki}) = \log(\mu_{ki}) + \beta_{i1} + \beta_{i2} l_k^* + \beta_{i3} Q_k^* + \beta_{i4} S_k + \beta_{i5} \text{SL}_k + \beta_{i6} \text{ST}_k,$$

and

$$\mu_{ki} \mid \alpha_i \overset{ind}{\sim} \text{gamma}(\alpha_i, \alpha_i),$$

$$\beta_{ij} \overset{iid}{\sim} \text{N}(0.0, 0.0001),$$

$$\alpha_i \overset{iid}{\sim} \text{gamma}(3, 1).$$

The precision parameter for the regression coefficients is very small so the normal prior distribution is vague (quite flat). In practical terms this means that we do not have any information about what the actual values of regression coefficients might be. In other words, the regression coefficients may take almost any real value. The gamma(3, 1) prior for the α parameters

is motivated by approximating the logarithm of a gamma random variable with a normal variable. This particular choice of the parameters is explained later in conjunction with a mixed Poisson-log normal model.

The directed graphical model associated with this particular model is presented in Figure 9.2. The two plates correspond to the two different indices, k for sites and i for accident type. For this model, the results were calculated

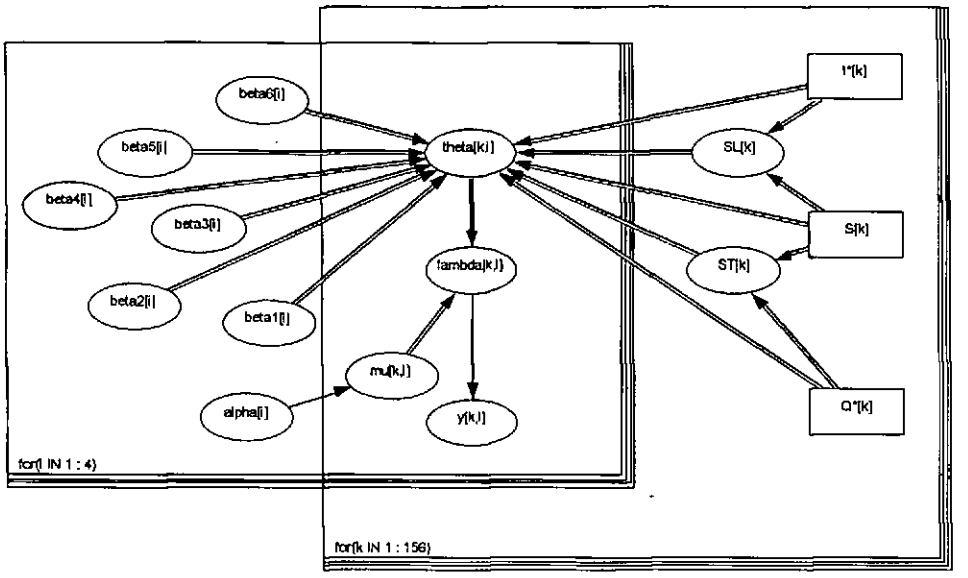


Figure 9.2: Directed graphical model for the hierarchical Bayesian model with gamma random effects

from a sample of 10000 values of a single long chain, with a burn-in period of 45000 iterations, and from a mixed sample of 10000 of two parallel chains, after a burn-in period of 10000 iterations. The Gelman-Rubin monitoring statistic was very good, less than 1.05 for all parameters of interest, and also the

dynamic plots showed that the chain had attained convergence.

The Bayesian P values for the χ^2 discrepancy, equation (8.18) in Section 8.2.6, for each type of accident, are 0.72 for KSI accidents with one vehicle, 0.62 for KSI with two or more vehicles, 0.51 for S with one vehicle, 0.53 for S with two or more vehicles. These values shows that the data does not contradict the model so the inferences are reliable. The hyper-parameters (α_i)

Table 9.2: Posterior means of regression coefficients for mixed Poisson-gamma model

Response	1	l^*	Q^*	S	SL	ST
Y_1	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
	0.55	1.20	0.60	-0.29	0.04	-0.38
Y_2	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
	0.85	1.49	0.72	-0.29	0.09	-0.01
Y_3	β_{31}	β_{32}	β_{33}	β_{34}	β_{35}	β_{36}
	1.09	1.28	0.57	-0.16	-0.10	-0.26
Y_4	β_{41}	β_{42}	β_{43}	β_{44}	β_{45}	β_{46}
	2.00	1.29	0.69	-0.35	-0.00	-0.08

can be estimated by the following posterior means $\overline{\alpha_1} = 4.5, \overline{\alpha_2} = 6.0, \overline{\alpha_3} = 3.5, \overline{\alpha_4} = 3.0$. This shows that the random effects for different types of accident have different gamma distributions. The parameters α are the precision of the random effects μ. The difference in values of α by different accident types may be due to different missing information for each type of accident.

The largest precision $\overline{\alpha_2} = 6.0$ is for killed or seriously injured accidents with two or more vehicles involved and the lowest precision $\overline{\alpha_4}$ is for slightly injured accidents with two or more vehicles. This means that for the latter type the corresponding random effects μ_4 are more volatile so there is more missing information.

The posterior means of the regression coefficients are shown in Table 9.2, other quantiles are described in Appendix G extracted from Tunaru (1999). The four regression equations are

$$\begin{aligned}
\log(\lambda_{k1}) &= \log(\mu_{k1}) + 0.55 + 1.20 l_k^* + 0.60 Q_k^* - 0.29 S_k + 0.04 \mathrm{SL}_k \\
&\quad -0.38 \mathrm{ST}_k \\
\log(\lambda_{k2}) &= \log(\mu_{k2}) + 0.85 + 1.49 l_k^* + 0.72 Q_k^* - 0.29 S_k + 0.09 \mathrm{SL}_k \\
&\quad -0.01 \mathrm{ST}_k \\
\log(\lambda_{k3}) &= \log(\mu_{k3}) + 1.09 + 1.28 l_k^* + 0.57 Q_k^* - 0.15 S_k - 0.10 \mathrm{SL}_k \\
&\quad -0.26 \mathrm{ST}_k \\
\log(\lambda_{k4}) &= \log(\mu_{k4}) + 2.00 + 1.29 l_k^* + 0.69 Q_k^* - 0.35 S_k - 0.005 \mathrm{SL}_k \\
&\quad -0.08 \mathrm{ST}_k
\end{aligned}$$

These estimated equations point to some interesting conclusions.

- The interaction SL between speed limit and link length is almost null or very weak for all 4 regression equations. This can be easily seen

looking at the posterior median, 2.5% and 97.5% percentiles given in Appendix G. All credible intervals overlaps zero. This means that the difference in number of accidents between a rural road of 10 km and an urban road of 10 km is the same between a rural road of 5 km and an urban road of 5 km.

- There is a non-negligible interaction between speed limit and link traffic only for the regression equations corresponding to single vehicle accidents. This means that, for accidents with only one vehicle, reducing the traffic flow by a factor equal to δ, that is from Q to δQ, and keeping all the other covariates the same, will result in a reduction of the number of accidents depending on δ and the speed limit S. It is shown bellow that the percentage in accident reduction is $(1 - \delta^{0.88-0.38S})$ for fatal or serious accidents with one vehicle and $(1-\delta^{0.96-0.26S})$ for slight accidents with one vehicle. There would be no speed limit S in these formulae if there were no interactions between speed limit and link traffic.

- For slight accidents the speed limit effect for accidents with two or more cars is more than double in absolute value the speed limit effect for accidents with only one vehicle involved. A possible explanation might be that, for this category of accidents, the interaction between speed limit and the other two variables, link length and traffic flow, is very weak, whereas for slight accidents with only one vehicle, there is considerable interaction between speed limit and the other variables.

- Speed limit has a negative effect in the linear regression equation given above for all four types of accidents and the effects seem to be the same for fatal or serious accidents with one vehicle and with two or more vehicles, respectively. However, because of the definition of the interaction terms, the effect of speed limit on the number of accidents can be better understood from the multiplicative equations (9.3)–(9.6).

The log-linear regresion equations can be re-expressed in multiplicative form as

$$\lambda_{ki} = \mu_{ki} \exp(\beta^*_{i1} + \beta^*_{i4} S_k) l_k^{(\beta^*_{i2} + \beta_{i5} S_k)} Q_k^{(\beta^*_{i3} + \beta_{i6} S_k)}. \tag{9.2}$$

The new coefficients marked with a star can be recalculated from the initial β_{ki}. The coefficient β^*_{i1} can be included in the constant factor but the above form was preferred for symmetry. For (P-ga) model, the regression equations can be rewritten as

$$\lambda_{k1} = \mu_{k1} \exp(-0.50 + 0.09 S_k) l_k^{(0.89 + 0.04 S_k)} Q_k^{(0.88 - 0.38 S_k)} \tag{9.3}$$

$$\lambda_{k2} = \mu_{k2} \exp(-0.26 - 0.31 S_k) l_k^{(1.09 + 0.09 S_k)} Q_k^{(0.96 - 0.01 S_k)} \tag{9.4}$$

$$\lambda_{k3} = \mu_{k3} \exp(0.15 + 0.11 S_k) l_k^{(0.98 - 0.10 S_k)} Q_k^{(0.81 - 0.26 S_k)} \tag{9.5}$$

$$\lambda_{k4} = \mu_{k4} \exp(0.94 - 0.27 S_k) l_k^{(0.96 - 0.005 S_k)} Q_k^{(0.94 - 0.08 S_k)} \tag{9.6}$$

where the posterior distribution of random effects can be inferred as

$$\mu_{k1} \sim \text{gamma}(4.5, 4.5), \quad \mu_{k2} \sim \text{gamma}(6, 6)$$

$$\mu_{k3} \sim \text{gamma}(3.5, 3.5), \quad \mu_{k4} \sim \text{gamma}(3, 3).$$

The multiplicative equations have different forms, implying that a single response model rather than a multiple response model would lead to unreliable conclusions. For example, a single response model using a single regression equation would have only one value for the regression coefficient corresponding to the speed limit or to the interaction between speed limit and link length. It can be easily seen from equations (9.3)–(9.6) that there is a lot of variation across the four types of accidents for this coefficients. A single value cannot synthesize the whole picture.

The regression equations developed as a major part of the hierarchical Bayesian models proposed can be used by practitioners to understand the behaviour of the mean number of accidents given the explanatory variables. They can also be used to predict how the mean number of accidents at a given site would change if some or all the explanatory variables were changed in some way, and to predict future accident rates given that the conditions are unchanged. If the local authority were to build a bypass around one of the villages on the road network (and they have since 1991) they would want to predict the effect on accidents. Most of the traffic that used to travel through the village would use the bypass and the only traffic using the old road would be traffic travelling to the village. The cost savings in accidents can then be calculated. The novelty of this approach is that predictions can be made simultaneously about the changes in accident type as well as the frequency. It is worth pointing out that many TRL studies investigated, at the univariate level, the relationships between traffic flows and various types of accident such

as lorries, trucks, bikes, accidents classified by manoeuvre and so on. The work presented here is multivariate, looking at several correlated types of accident at the same time.

We can see what happens when the traffic flow Q is changed to δQ, where $\delta > 0$. Momentarily we will drop the site index k. For all accident types, from equations (9.3)–(9.6), it can be easily shown that there is a reduction in the mean number of accidents λ if and only if $\delta < 1$, that is if the traffic is reduced. The reduction in the number of accidents can be calculated as

$$\frac{\lambda_i - \lambda_i^{\delta}}{\lambda_i} \times 100\% = \left(1 - \frac{\lambda_i^{\delta}}{\lambda_i}\right) \times 100\%.$$

This formula is applied for each type of accident based on the multiplicative equations (9.3)–(9.6) and the calculations can be finalised by specifying the speed limit variable, urban $S = -1$ and rural $S = 1$.

For fatal or serious accidents with 1 vehicle

$$\left(1 - \frac{\lambda_1^{\delta}}{\lambda_1}\right) \times 100\% = (1 - \delta^{0.88-0.38S}) \times 100\%.$$

For fatal or serious accidents with 2+ vehicles

$$\left(1 - \frac{\lambda_2^{\delta}}{\lambda_2}\right) \times 100\% = (1 - \delta^{0.96-0.01S}) \times 100\%.$$

For slight accidents with 1 vehicle

$$\left(1-\frac{\lambda_3^{\delta}}{\lambda_3}\right)\times 100\% = (1-\delta^{0.81-0.26S})\times 100\%.$$

For slight accidents with 2+ vehicles

$$\left(1-\frac{\lambda_4^{\delta}}{\lambda_4}\right)\times 100\% = (1-\delta^{0.94-0.08S})\times 100\%.$$

The final results are presented in Table 9.3.

Table 9.3: Proportional reductions ia accidents when traffic flow is reduced, as resulted from the Poisson-regression model with gamma random effects

Severity	No of vehicles	Speed limit	Reduction
fatal or serious	1	urban	$(1-\delta^{1.26})$
		rural	$(1-\delta^{0.5})$
	2+	urban	$(1-\delta^{0.97})$
		rural	$(1-\delta^{0.95})$
slight	1	urban	$(1-\delta^{1.07})$
		rural	$(1-\delta^{0.55})$
	2+	urban	$(1-\delta^{1.02})$
		rural	$(1-\delta^{0.86})$

The conclusion of this analysis is that reducing the traffic flow by a factor of δ will reduce different type of accidents in different ways. In a similar way

the percentage increase in accidents can be calculated if the traffic is increased by a multiplicative factor δ.

9.2.2 Comparison with a simpler scenario

One may wonder why a multiple response approach would give better results than fitting separate Poisson-regression models for accident counts of each type. Therefore in this section the following model will be investigated

$$\begin{aligned} Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}) \\ \log(\lambda_{ki}) &= \beta_{i1} + \beta_{i2} l_k^* + \beta_{i3} Q_k^* + \beta_{i4} S_k + \beta_{i5} \text{SL}_k + \beta_{i6} \text{ST}_k \end{aligned} \tag{9.7}$$

$$\beta_{ij} \overset{iid}{\sim} \text{N}(0.0, 0.0001)$$

and the results will be compared with those given by the (P-ga) model. Following the usual MCMC modelling steps, two chains were simulated in parallel and after a burn-in period of 15000 iterations a sample of 10000 values was retained for inference. The regression coefficients were estimated by their posterior means given in Table 9.4. Comparing Table 9.2 with Table 9.4, the only major differences are between the coefficients of the interaction terms between traffic flow and speed limit. If this model was proposed for inference the following multiplicative predictive equations would be used

$$\lambda_{k1} = \exp(-0.50 + 0.25 S_k) l_k^{(0.87+0.01 S_k)} Q_k^{(0.88-0.48 S_k)} \tag{9.8}$$

Table 9.4: Posterior means of regression coefficients for Poisson-regression model

Response	1	l^*	Q^*	S	SL	ST
Y_1	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
	0.56	1.17	0.58	-0.24	0.01	-0.48
Y_2	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
	0.91	1.37	0.71	-0.21	0.09	-0.13
Y_3	β_{31}	β_{32}	β_{33}	β_{34}	β_{35}	β_{36}
	1.12	1.20	0.59	-0.07	-0.12	-0.46
Y_4	β_{41}	β_{42}	β_{43}	β_{44}	β_{45}	β_{46}
	2.00	1.28	0.74	-0.30	0.01	-0.24

$$\lambda_{k2} = \exp(-0.19 - 0.09S_k) l_k^{(1.00+0.09S_k)} Q_k^{(0.97-0.13S_k)} \quad (9.9)$$

$$\lambda_{k3} = \exp(0.04 + 0.41S_k) l_k^{(0.92-0.12S_k)} Q_k^{(0.88-0.46S_k)} \quad (9.10)$$

$$\lambda_{k4} = \exp(0.79 - 0.05S_k) l_k^{(0.96+0.01S_k)} Q_k^{(1.04-0.24S_k)} \quad (9.11)$$

Comparing these equation with equations (9.3)–(9.6) it is easy to see that the values for traffic flow are different, and this will change predictions in model.

The percentages of reduction in accidents when the traffic flow Q is reduced by a factor of δ are different from before as can be seen from Table 9.5. The simple Poisson-regression model overestimates the reductions in accidents resulting from reducing the traffic for all urban areas, that is it gives higher reduction percentages for $S = -1$, and underestimates the reductions in acci-

Table 9.5: Reductions in accident percentages when traffic flow is reduced, as resulted from the Poisson-regression model without random effects

Severity	No of vehicles	Speed limit	Reduction
fatal or serious	1	urban	$(1-\delta^{1.36})$
		rural	$(1-\delta^{0.4})$
	2+	urban	$(1-\delta^{1.10})$
		rural	$(1-\delta^{0.84})$
slight	1	urban	$(1-\delta^{1.34})$
		rural	$(1-\delta^{0.38})$
	2+	urban	$(1-\delta^{1.28})$
		rural	$(1-\delta^{0.80})$

dents in rural areas, that is it gives smaller percentages for $S = 1$.

In conclusion fitting accident counts of different type at an univariate level would result in different inferential results. In a Bayesian framework, it can be easily seen that this simpler model is rejected by the data since the Bayesian P-values for χ^2 for each type of accident are respectively 0.013, 0.005, 0.000 and 0.000. These values are calculated as described in Section 8.2.6 by formula (8.17) for the test statistic given by formula (8.18). The model is rejected by the data if the Bayesian P-values are too small. It can be easily seen that, for the first type of accident, it is just accepted at 0.01 level but for all the other three the rejection is clear. Thus, the model without random effects cannot model the data well and therefore the inclusion of random effects seem

to be necessary. This is not surprising since only three explanatory variables are used as covariate information. The role of the random effects is to account for missing explanatory variables.

This emphasizes that even after ensuring that the Markov chain has converged and parameters are reliably estimated it is necessary to check the goodness-of-fit of the model before applying the results. Hence the joint-response model is superior to four separate univariate response models.

9.2.3 A Poisson-regression model with log normal random effects

The logarithm of a gamma distributed random variable is approximately normal so it is worth considering a model where the random effect is normally distributed. This assumption can be exploited to simplify computation since all the parameters describing the regression equations are normally distributed. This model is also specified hierarchically in 3 stages and it will be called (P-logN). For all sites $k = 1, 2, \ldots, 156$ and $i = 1, 2, 3, 4$

$$Y_{ki} \mid \lambda_{ki} \overset{ind}{\sim} \text{Pois}(\lambda_{ki}), \tag{9.12}$$

where

$$\log(\lambda_{ki}) = b_k + \beta_{i1} + \beta_{i2} l_k^* + \beta_{i3} Q_k^* + \beta_{i4} S_k + \beta_{i5} \text{SL}_k + \beta_{i6} \text{ST}_k$$

and

$$b_k \mid \tau \overset{iid}{\sim} \mathrm{N}(0.0, \tau),$$

$$\beta_{ij} \overset{iid}{\sim} \mathrm{N}(0.0, 0.0001),$$

$$\tau \sim \mathrm{gamma}(3, 1).$$

For the precision distribution τ a prior gamma$(3, 1)$ was used. The reason

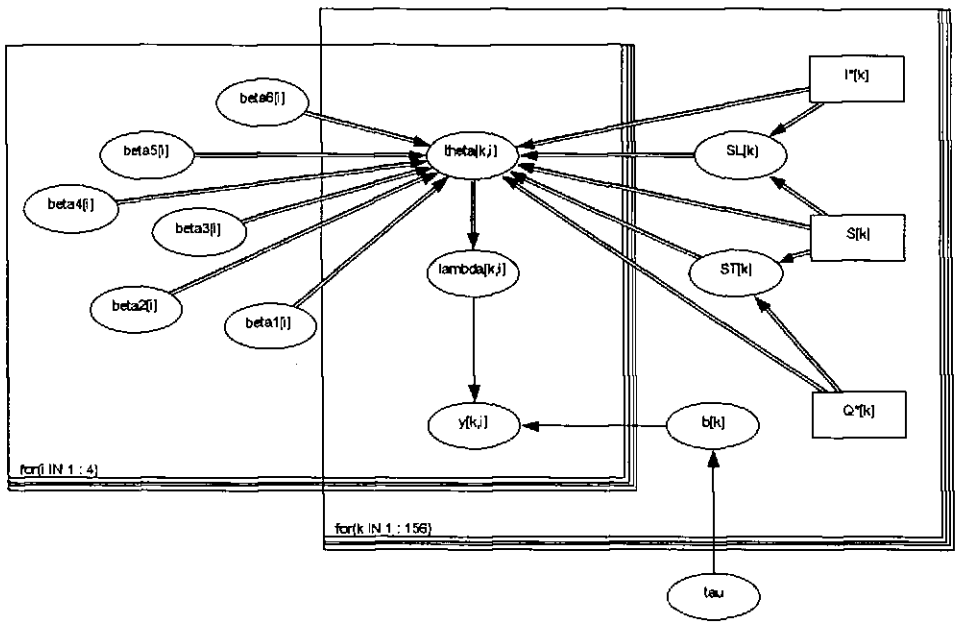


Figure 9.3: Directed graphical model for the hierarchical Bayesian model with log normal random effects

for choosing a gamma$(3, 1)$ prior is based on an idea described in Smith, Spiegelhalter and Thomas (1995). The model just described might support

the assumption that 95% of the sites having identical covariates

$$X'_{ki} = (1, l^*_k, Q^*_k, S_k, \mathrm{SL}_k, \mathrm{ST}_k),$$

will have a $\log(\lambda)$ between $-1.96/\sqrt{\tau}$ and $1.96/\sqrt{\tau}$. Assuming (from a subjective point of view) that sites with the same covariates have expectations varying within one order of magnitude $\log 10 = 2.3$ but not over two orders of magnitude $\log 100 = 4.6$, the equation

$$2 \times 1.96/\sqrt{\tau} \approx 2.3$$

implies that $\tau \approx 2.9$ is a good approximation for $\mathrm{E}(\tau)$. In addition, an approximation of a low value for τ is obtained from

$$2 \times 1.96/\sqrt{\tau} \approx 4.6$$

and this lower limit equals 0.73. With the prior distribution gamma$(3, 1)$, τ has the mean 3 and $\Pr(\tau > 0.73) = 0.96$ which shows that this gamma distribution is appropriate for our subjective assumption.

The joint distribution of the observed and unobserved quantities, data and parameters, factorises in a similar way to the previous model. Although the graphical model in Figure 9.3, representing the conditional distributions assumed by the model, is very similar to that describing the model (P-ga) in Figure 9.2, it should be noted that different parametric distributions are used

Table 9.6: Posterior means for mixed Poisson-log normal regression coefficients

Response	1	l^*	Q^*	S	SL	ST
Y_1	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
	0.41	1.25	0.60	-0.29	0.05	-0.34
Y_2	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
	0.73	1.49	0.75	-0.29	0.12	0.04
Y_3	β_{31}	β_{32}	β_{33}	β_{34}	β_{35}	β_{36}
	0.97	1.28	0.59	-0.13	-0.10	-0.29
Y_4	β_{41}	β_{42}	β_{43}	β_{44}	β_{45}	β_{46}
	1.82	1.39	0.78	-0.38	0.04	-0.06

for corresponding nodes of the graph, and different logical expressions for the means lambda[k,i]. Furthermore, the precision parameter tau for the normal random effects b[k], does not depend on the accident type or the site.

Two parallel chains were simulated and after a burn-in period of 15000 iterations a sample of 10000 iterations was selected for inference. The Gelman-Rubin statistics were very good for all parameters of interest and the Bayesian P-values for the four types of accidents were 0.58 for KSI with 1 vehicle, 0.53 for KSI with 2+ vehicles, 0.07 for S with 1 vehicle and 0.04 for S with 2+ vehicles. The fit of this model seems to be good for the first two accident types and not very good for the last two types. The variance of the random effects can be estimated by its posterior mean 0.50. The results obtained for this model are

shown in Table 9.6. They are very similar to those obtained using a gamma random effect. However, the observed ranges of the standardised logarithmic link length l_k^* and the standardised logarithmic estimated traffic flow Q_k^* are between -2.7 and 1.6, which is quite narrow. Therefore, small differences in the estimated values of regression coefficients may result in substantial differences in the fit of the two models. This problem will be investigated further in the next subsection.

Model comparison

A simple way to check the fit of a model is to compare the posterior predicted mean given by the model with the data values. A close linear relationship would suggest a close fit. This can be done in parallel for the two models (P-ga) and (P-logN). First, the observed pooled number of accidents at each site is plotted against the sum of the predicted means of accidents at the same site. It seems that model (P-logN) performs slightly better.

A better insight is plotting each type of accident separately and this is done in Appendix C. There, the model (P-ga) seems to fit the data well for all four types, and for each type better than the (P-logN) model. This can be expected since it was shown in Chapter 8 that the log normal distribution will not perform very well with the extreme observed values. In addition, the random effects depend only on site so they can account only for missing information about the site and not about the accident class. In summary, both models fit the data well at an aggregated level, the second model appreciably

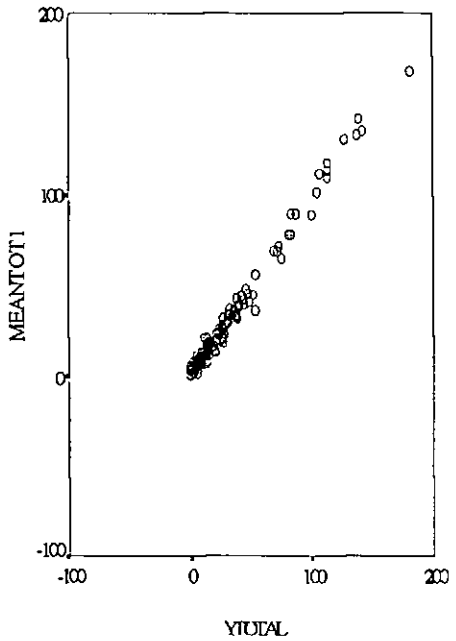


Figure 9.4: Scatter plots of totals for model (P-ga)

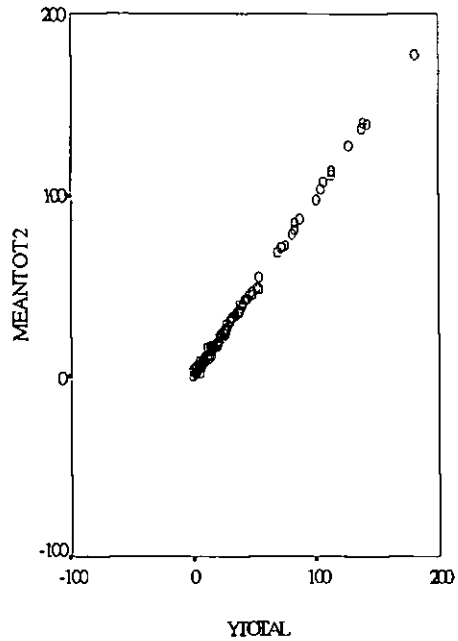


Figure 9.5: Scatter plots of totals for model (P-logN)

better. But at a disaggregated level, clearly the first model fits the data better than the second one. The aim of this analysis was to extract statistical information at a disaggregated level so it is vital to have a good fit for each type of accident. The difference in the form of random effects and the distribution used is important.

9.2.4 Poisson-regression model with multivariate normal random effects

The (P-ga) and (P-logN) models studied above provide a good start for the statistical modelling process but there is no correlation structure assumed for the random effects μ and as a consequence these two models may overlook an important aspect of the real data. The next model, that will be called

(P-MNre), tries to overcome this difficult problem as well. It can be viewed as an extension of the two previously studied models, the only difference being in the distributional assumption for the random effects μ.

For all sites $k = 1, 2, \ldots, 156$ and accident groups $i = 1, 2, 3, 4$

$$Y_{ki} \mid \lambda_{ki} \overset{ind}{\sim} \text{Pois}(\lambda_{ki}), \tag{9.13}$$

where

$$\log(\lambda_{ki}) = \mu_{ki} + \beta_{i1} + \beta_{i2} l_k^* + \beta_{i3} Q_k^* + \beta_{i4} S_k + \beta_{i5} \text{SL}_k + \beta_{i6} \text{ST}_k,$$

and

$$(\mu_{ki})_{i=1,\ldots,4} \mid T \overset{ind}{\sim} \text{N}_4(0, T),$$

$$\beta_{ij} \overset{iid}{\sim} \text{N}(0.0, 0.0001),$$

$$T \sim \text{Wishart}(R, 4).$$

The covariance structure of the random effects is given by the covariance matrix $\Sigma = T^{-1}$, so the parameterisation is again based on the inverse of the variance-covariance matrix. For computational simplicity, a Wishart hyper-prior distribution is required for the matrix T and the matrix $(R)^{-1}$ accounts for our prior beliefs about the precisions between random effects μ of different types of accidents; the second parameter of the Wishart distribution is chosen to be as small as possible (in this case 4) to reflect our ignorance about T.

For this model the inference process is based again on MCMC methods but the Markov chain has to be generated using the Metropolis-Hastings algorithm (see Gelman et al. (1995)), because the Gibbs sampler does not work in this case. WinBUGS has both methods implemented so the models can be fitted using the same software platform. To improve speed, an initial run can be made using some arbitrary values for R, and then the posterior means of the elements of $\Sigma^{-1} = T$ are used for the R values.

The posterior distribution for the model (P-MNre), with multivariate normal random effects, gave the posterior means for the regression coefficients in Table 9.7. A burn-in period of 30000 iterations was used before a sample of 10000 was taken as representative for the posterior distribution of all parameters of the model. Very similar results were obtained when two parallel chains were simulated. After a burn-in period of 15000 iterations a sample of 10000 iterations was taken. The Gelman-Rubin convergence statistics were all less than 1.05 for all parameters of interest β, T, Σ and also, the Bayesian P-values indicated a good fit of this model. The values were 0.87, 0.80, 0.65 and 0.65 for the four types of accident in order, showing a good fit to the data. Apart from the intercept terms β_{i1}, there are no major differences in the signs and absolute values of the regression coefficients as compared with the Poisson model with gamma random effects, (P-ga). The matrix given in (9.14) contains the posterior means of the elements of T, the inverse of the

Table 9.7: Posterior means of regression coefficients for the Poisson-regression model with multivariate normal random effects

Response	1	l^*	Q^*	S	SL	ST
Y_1	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
	-0.65	1.27	0.63	-0.30	0.04	-0.34
Y_2	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
	-0.31	1.55	0.78	-0.31	0.11	0.00
Y_3	β_{31}	β_{32}	β_{33}	β_{34}	β_{35}	β_{36}
	-0.14	1.35	0.60	-0.16	-0.08	-0.25
Y_4	β_{41}	β_{42}	β_{43}	β_{44}	β_{45}	β_{46}
	0.73	1.36	0.77	-0.34	0.00	-0.06

covariance matrix.

$$T = \begin{pmatrix} 4.65 & -1.75 & -1.5 & -0.1 \\ -1.75 & 6.12 & -0.86 & -1.35 \\ -1.5 & -0.86 & 3.25 & -0.5 \\ -0.1 & -1.35 & -0.5 & 2.3 \end{pmatrix} \tag{9.14}$$

Elements close to zero in the inverse covariance matrix T indicate that the corresponding random effects μ are conditionally independent given the values of the random effects not in the pair. For instance, $T_{14} = -0.1$ indicates conditional independence between the random effects for fatal or serious accidents with only one vehicle and slight accidents with two or more vehicles,

which is not surprising. What is surprising is the weak association between slight accidents with one vehicle and slight accidents with two or more vehicles, $T_{34} = -0.5$.

9.3 Multivariate Poisson-log normal model

In the previous chapter, the mixture of a Poisson distribution with a multivariate log normal distribution was described, equation (8.35), as a discrete multivariate distribution for modelling multiple counts. Starting from this multivariate Poisson-log normal distribution a hierarchical, fully Bayesian model, that will be called (P-MN1), is proposed.

$$\begin{aligned} Y_{ki}|\lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}) \\ (\log(\lambda_{ki}))_{i=1,\ldots,4}|\mu, T &\overset{iid}{\sim} \text{N}_4(\mu, T) \\ \mu_i &\overset{iid}{\sim} \text{N}(0, 0.0001) \\ T &\sim \text{Wishart}(R, 4) \end{aligned} \tag{9.15}$$

where the parameterisations are the same as used for the previous models. A variant of this model, (P-MN2), would be to add another level to the hierarchy:

$$\begin{aligned} \mu_i &\sim \text{N}(\nu, \tau_\alpha) \\ \nu &\sim \text{N}(0, 0.001) \\ \tau_\alpha &\sim \text{gamma}(0.001, 0.001) \end{aligned} \tag{9.16}$$

and keeping everything else the same. This allows a comparison between two nested models. The matrix parameter R for the Wishart distribution is proposed by analogy with model (P-MNre).

The graphical model associated with model (P-MN1) is illustrated in Figure 8.6. The conditional independence structure is remarkably simple. The graphical model for model (P-MN2) would have an extra two vertices for ν and τ_a as parents of the vertex μ_i.

The same strategy for simulation was used as for the model (P-ga). Thus the inference results were based on either a sample of 10000 values, taken from a single chain after a burn-in period of 45000 iterations, or on a mixed sample of 10000 taken from two parallel chains, after a burn-in of 20000 iterations. The Gelman-Rubin convergence statistics, equation (8.15) in Section 8.2.3, were very good, with values less than 1.1 for all parameters of interest.

The Bayesian P values for the χ^2 discrepancy, equation (8.18), for each type of accident, are 0.87 for KSI accidents with one vehicle, 0.77 for KSI accidents with two or more, 0.71 for S accidents with one vehicle and 0.62 for S accidents with two or more vehicles. These values are a bit larger than the corresponding Bayesian P-values for model (P-ga), but they are still good and shows that the data does not contradict the model so the inferences are reliable.

The posterior estimates of the parameters of interest for the multivariate Poisson-log normal model (P-MN1) is given in Table 9.8. The covariance matrix $\Sigma = T^{-1}$ is provided because it makes a straightforward link with

possible covariance structure of the observed data.

Table 9.8: Posterior estimation of parameters of multivariate Poisson-log normal model

parameter	mean	sd	2.5%	97.5%
σ_{11}	2.15	0.41	1.46	3.09
σ_{12}	2.34	0.41	1.65	3.29
σ_{13}	2.10	0.39	1.48	3.00
σ_{14}	2.25	0.38	1.62	3.11
σ_{21}	2.34	0.41	1.66	3.29
σ_{22}	3.04	0.54	2.16	4.27
σ_{23}	2.48	0.45	1.78	3.54
σ_{24}	2.78	0.45	2.04	3.79
σ_{31}	2.10	0.39	1.48	3.00
σ_{32}	2.48	0.45	1.78	3.54
σ_{33}	2.49	0.47	1.75	3.61
σ_{34}	2.39	0.41	1.73	3.32
σ_{41}	2.25	0.38	1.62	3.11
σ_{42}	2.78	0.45	2.04	3.79
σ_{43}	2.39	0.40	1.73	3.32
σ_{44}	3.04	0.47	2.25	4.09
μ_1	0.28	0.15	-0.034	0.56
μ_2	0.67	0.16	0.34	0.97
μ_3	0.79	0.16	0.46	1.08
μ_4	1.65	0.15	1.35	1.94

The matrix given in (9.17) contains the posterior means of the elements of T, the inverse covariance matrix.

$$T = \begin{pmatrix} 4.42 & -1.55 & -1.65 & -0.55 \\ -1.55 & 3.76 & -0.99 & -1.52 \\ -1.65 & -0.99 & 3.18 & -0.36 \\ -0.55 & -1.52 & -0.36 & 2.44 \end{pmatrix} \tag{9.17}$$

There are weak partial correlations between KSI accidents with 1 vehicle and

slight accidents with 2+ vehicles, between KSI accidents with 2+ vehicles and slight accidents with 1 vehicle and between slight accidents with 1 vehicle and slight accidents with 2+ vehicles.

9.4 Model selection using DIC

Having so many models under study, some of them nested, some of them not, it would be helpful to check the fit of the models and identify the best ones for further analysis. In a Bayesian context, this can be done using the posterior distribution of the log-likelihood, as suggested by Dempster, but because extremely complex models should be penalised for using a large number of parameters, the deviance information criterion, DIC, offers a better solution.

For comparison, several simpler nested and non-nested models are investigated. The previously discussed models were denoted by (P-ga), (P-logN), (P-MNre), (P-MN1) and (P-MN2). The following models are considered as well

(P-difreg): A Poisson-regression model without random effects but different regression coefficients for different accident types

$$\begin{aligned} Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}) \\ \log(\lambda_{ki}) &= \beta_{i1} + \beta_{i2} l_k^* + \beta_{i3} Q_k^* + \beta_{i4} S_k + \beta_{i5} \text{SL}_k + \beta_{i6} \text{ST}_k \\ \beta_{ij} &\overset{iid}{\sim} \text{N}(0.0, 0.0001) \end{aligned} \tag{9.18}$$

(P-ureg): A Poisson-regression model with random effects and with the same

regression coefficients for all types of accidents

$$\begin{aligned} Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}) \\ \log(\lambda_{ki}) &= \log(\mu_{ki}) + \beta_1 + \beta_2 l_k^* + \beta_3 Q_k^* + \beta_4 S_k + \beta_5 \text{SL}_k + \beta_6 \text{ST}_k \\ \mu_{ki} \mid \alpha_i &\overset{ind}{\sim} \text{gamma}(\alpha_i, \alpha_i) \\ \beta_j &\overset{iid}{\sim} \text{N}(0.0, 0.0001) \\ \alpha_i &\overset{iid}{\sim} \text{gamma}(3, 1) \end{aligned} \tag{9.19}$$

(P-classic): A Poisson-regression model with identical regression coefficients for all types of accidents and without random effects

$$\begin{aligned} Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}) \\ \log(\lambda_{ki}) &= \beta_1 + \beta_2 l_k^* + \beta_3 Q_k^* + \beta_4 S_k + \beta_5 \text{SL}_k + \beta_6 \text{ST}_k \\ \beta_j &\overset{iid}{\sim} \text{N}(0.0, 0.0001) \end{aligned} \tag{9.20}$$

(P-logN2): The same Poisson-log normal regression model as before but with different hyper-prior parameters, gamma(0.001, 0.001)

$$\begin{aligned} Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}) \\ \log(\lambda_{ki}) &= b_k + \beta_{i1} + \beta_{i2} l_k^* + \beta_{i3} Q_k^* + \beta_{i4} S_k + \beta_{i5} \text{SL}_k + \beta_{i6} \text{ST}_k \\ b_k \mid \tau &\overset{iid}{\sim} \text{N}(0.0, \tau) \\ \beta_{ij} &\overset{iid}{\sim} \text{N}(0.0, 0.0001) \\ \tau &\sim \text{gamma}(0.001, 0.001) \end{aligned} \tag{9.21}$$

(P-add): An additive random effects model; there is no covariate information and the effects are split into terms accounting for site variation and terms trying to explain site by accident type variation

$$\begin{aligned} Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}) \\ \log(\lambda_{ki}) &= b_k + \log \mu_{ki} \\ b_k \mid \tau &\overset{iid}{\sim} \text{N}(0.0, \tau) \\ \mu_{ki} | \alpha_i &\overset{ind}{\sim} \text{gamma}(\alpha_i, \alpha_i) \\ \alpha_i &\overset{iid}{\sim} \text{gamma}(3, 1) \\ \tau &\sim \text{gamma}(3, 1) \end{aligned} \tag{9.22}$$

(P-add2): The same additive model as before with different hyper-prior parameters, gamma$(0.001, 0.001)$.

$$\begin{aligned} Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}) \\ \log(\lambda_{ki}) &= b_k + \log \mu_{ki} \\ b_k \mid \tau &\overset{iid}{\sim} \text{N}(0.0, \tau) \\ \mu_{ki} | \alpha_i &\overset{ind}{\sim} \text{gamma}(\alpha_i, \alpha_i) \\ \alpha_i &\overset{iid}{\sim} \text{gamma}(0.001, 0.001) \\ \tau &\sim \text{gamma}(0.001, 0.001) \end{aligned} \tag{9.23}$$

In the table 9.9, DIC is calculated for all models using the method described in Section 8.2.5. If the analysis had been based only on the posterior mean of

Table 9.9: Deviance Information Criterion calculations

MODEL	random effects	with regression	different β	$\overline{D}$	$D(\overline{\theta})$	p_D	DIC
(P-ga)	√	√	√	427.50	255.30	172.20	599.70
(P-difreg)		√	√	1180.40	579.35	601.05	1781.45
(P-ureg)	√	√		427.10	68.52	358.58	785.68
(P-classic)		√		2621.80	1308.27	1313.53	3935.33
(P-MNre)	√	√	√	383.32	127.42	255.90	639.22
(P-logN)	√	√	√	530.40	212.44	317.96	848.36
(P-logN2)	√	√	√	530.50	213.35	317.15	847.65
(P-add)	√			409.79	60.66	349.13	758.92
(P-add2)	√			412.62	60.78	351.84	764.46
(P-MN1)				389.38	119.40	269.98	659.36
(P-MN2)				389.65	60.00	329.65	719.30

the Bayesian deviance, $\overline{D}$, then the models (P-MNre), (P-MN1) and (P-MN2) would have been preferred. Taking into account the complexity of the models using DIC as a yardstick, the model (P-ga) is preferred followed closely by models (P-MNre) and (P-MN1).

In making this comparison, several points are worth noting. First the simplest model (P-classic) has a very large DIC = 3935.33. Therefore the improvement due to including random effects or allowing the regression coefficients to depend on accident type can be gauged relative to this basic model. Just adding the random effects μ, as in the model (P-ureg), results in a reduction of 3149.65 in DIC. Secondly, nested models like some pairs in the first five or the last two models in the table can be compared in terms of DIC. For example, the model (P-MN2) and its submodel (P-MN1) perform equally well in terms of the posterior mean of the deviance, but when we take into account

the number of parameters the DIC clearly indicates that the model (P-MN1) is better. Thirdly, the models retained, (P-ga), (P-MNre) and (P-MN1), are very flexible and reveal various aspects of the data analysed. Notice that model (P-MN1) does not use any regression structure, although to do that is quite easy linking the mean of the multivariate log normal distribution with explanatory variables.

9.5 Ranking the sites

Identifying hazardous sites is the first important step for developing road engineering measures. This problem is vital since designing and implementing remedial measures is based on the characteristics and factors related to those sites. Moreover, engineering treatment is applied only to sites selected. Large amounts of money can be wasted just because the right sites have not been identified as dangerous. Several approaches proposing some solutions were discussed in Chapter 2. All previous work was developed for univariate response models, nothing apparently having been done for multiple responses. This section investigates ranking the 156 sites from Kent, with four types of accidents.

The three hierarchical multiple response models that have been selected by the DIC criterion, that is (P-ga), (P-MNre) and (P-MN1), will be used. Therefore, for each measure, the ranking calculations are made for three models by four types of accident. Under a restricted budget, the analysis proposed

here (using several criteria of ranking) would help the practitioners to select the hazardous locations, where "hazardous" has many facets. Modelling multiple counts jointly makes ranking just a bit more difficult but more rewarding in the same time. A practitioner may compare the sites according to different point of views and hidden aspects might come to the surface in this way.

9.5.1 Ranking by the probability that a site is the worst

The posterior probability that the site k is worse than all the others by a factor of v, for the accident type i, is

$$p_{ki}(v) = \Pr(\lambda_{ki} > v\,\lambda_{ji} \quad \text{for all } j \neq k|y)$$

where $v > 0$. For example, when $v = 1$ this is the probability that the site is the worst one. The factor v should be established prior to the analysis by the practitioner. The posterior probability that is used as a criterion for ranking represents a measure of how much worse one accident site is compared with all the others. In practice arbitrarily selected v- values like $v = 1, 1.1, 1.25$ are used. The practitioner then can see different lists and make an ad-hoc decision accordingly. The point to bear in mind is that the list of selected sites should not contain just a few sites or too many sites. The value $v = 1$ is always a good start and depending on the results obtained, the practitioner can modify

v accordingly. When $v = 1$ it is true that

$$\sum_k p_{ki}(1) = 1$$

and this is convenient for checking that the calculations are correct.

Table 9.10: Ranking probabilities for KSI accidents with 1 vehicle

model (P-ga)		model (P-MNre)		model (P-MN1)	
Site No	Pr	Site No	Pr	Site No	Pr
11	0.0023			11	0.0020
12	0.0070				
14	0.0257	14	0.0048	14	0.0076
23				23	0.0002
38		38	0.0012	38	0.0002
41	0.1427	41	0.1244	41	0.1332
42	0.0003	42		42	0.0004
46	0.0330	46	0.0204	46	0.0280
50	0.0007				
68		68	0.0004		
76	0.0023	76		76	0.0004
77	0.0046	77	0.0028	77	0.0058
90	0.7573	90	0.8132	90	0.7934
91		91	0.0016	91	0.0004
95		95	0.0048	95	0.0058
118				118	0.0008
143		143	0.0264	143	0.0218

Only the sites with corresponding probabilities larger than 10^{-4} are presented in the tables summarising the results. The tables contain the probabilities for the same type of accident, given by all three models for comparison. The sites with the largest probabilities need to be treated. If fatal or seriously injured accidents with only one vehicle involved are of particular interest, it is obvious from Table 9.10 that the worst site is number 90, urgent measures

Table 9.11: Ranking probabilities for KSI accidents with 2+ vehicles

model (P-ga)		model (P-MNre)		model (P-MN1)	
Site No	Pr	Site No	Pr	Site No	Pr
4		4	0.0004	4	0.0018
11	0.1200	11	0.1444	11	0.1252
12	0.2173	12	0.1900	12	0.1194
14	0.3630	14	0.3144	14	0.2732
24	0.0023	24	0.0004	24	0.0018
41				41	0.0014
46	0.2061	46	0.2520	46	0.3228
76	0.0076	76	0.0060	76	0.0032
77	0.0007	77		77	0.0002
90		90	0.0004	90	0.0064
98	0.0596	98	0.0408	98	0.0348
102	0.0169	102	0.0040	102	0.0028
118	0.0062	118	0.0472	118	0.1070

being required; also sites 41, 46, 14 and possibly 143 should be investigated. Site 90 is the worst site for accidents with slight injuries as well, see Tables 9.12 and 9.13, but, as can be seen from Table 9.11, it is not as bad regarding fatal or seriously injured accidents with two or more vehicles. Therefore, the statistical analysis at the disaggregated level provides practitioners with more valuable information as what might be the problems at a specific site.

Table 9.12: Ranking probabilities for slight accidents with 1 vehicle

model (P-ga)		model (P-MNre)		model (P-MN1)	
Site No	Pr	Site No	Pr	Site No	Pr
14		14	0.0008	14	0.0008
41	0.0019	41	0.0004	41	0.0010
46	0.0062	46	0.0008	46	0.0036
90	0.9919	90	0.9980	90	0.9946

The results are quite similar for all three models. By the measure studied in this section, it seems that there are not many dangerous sites for slight accidents with only one vehicle. One reason might be that site 90 is so bad that almost the whole probability is concentrated on this site, and there is not very much left to distinguish between the others. This site is particulary interesting. It is the urban link that runs along the sea front at the resort of Margate. Thus, there would be a high volume of holiday makers both pedestrian and drivers. The high pedestrian flow distinguishes it from the other links and special safety measures need to be implemented.

Table 9.13: Ranking probabilities for slight accidents with 2+ vehicles

model (P-ga)		model (P-MNre)		model (P-MN1)	
Site No	Pr	Site No	Pr	Site No	Pr
11				11	0.0004
12	0.1200	12	0.0820	12	0.1054
14	0.0923	14	0.0512	14	0.0626
24	0.0185	24	0.0304	24	0.0220
41	0.2338	41	0.2368	41	0.2144
46	0.0035	46	0.0048	46	0.0036
76	0.0031	76	0.0028	76	0.0022
77	0.0007	77	0.0004		
90	0.4869	90	0.5640	90	0.5688
98	0.0412	98	0.0276	98	0.0206

9.5.2 Ranking by posterior distributions of ranks

The second criterion for ranking sites investigated here is based on the ranks r_{ki} of the mean parameters λ_{ki} which are the site specific parameters. The ranking process is made again for each type of accident i. The posterior means

$\mathrm{E}(\lambda_{ki} \mid y)$ are optimal estimates when the aim is to produce inference about λ_{ki}. However, if the ranks of λ_{ki} are of interest, the conditional expected ranks (or a discretized version of them when they are not integers) are optimal. It is known that ranking the observed data or even the posterior means can perform poorly (Laird and Louis, 1989; Morris and Christiansen, 1996). Consequently, this ranking method is developed using the posterior distribution of the ranks, that is $p(r \mid y)$, and not the posterior distribution $p(\lambda \mid y)$. This differs than the approach proposed by Schluter et al. (1997).

Ranks are notoriously uncertain and it is useful to know the uncertainty associated with them. The approach followed here easily calculates the corresponding credible intervals of the estimated ranks. The ranks will be estimated by the posterior medians, mainly because they are easier to calculate. For each model and each accident type, the posterior median ranks and the associated $2.5\% - 97.5\%$ credible intervals are plotted together for comparison. Sites with ranks to the far right are more dangerous and sites with ranks to the far left are more safe.

The ranking process should be adjusted for including covariate information. There are two ways for doing that. A weak adjustment is already implicit in a Bayesian framework on the estimation process. For Poisson-regression models like (P-ga) and (P-MNre), a stronger approach considers the ranks not of λ_{ki} but of some quantities like random effects or regression-line intercepts or their sum, after removing the covariate information. Note that if the covariates included in the model are sufficient to explain all the variation between the

sites then there is no reason for ranking. When no covariate information is used, as in Schluter et al. (1997), obviously no adjustment of this type needs to be done. Model (P-MN1) is specified without any covariates so the ranking is based on the ranks of λ_{ki}.

The plot in Figure 9.6 illustrates the estimated statistics of the ranks of λ_{ki} relative to the model (P-ga), the Poisson-regression model with gamma random effects. It can be remarked on the plot in Figure 9.6 that sites with the lowest and, respectively highest, rank values, have quite small credible intervals. The local authorities may decide to treat all the sites that are

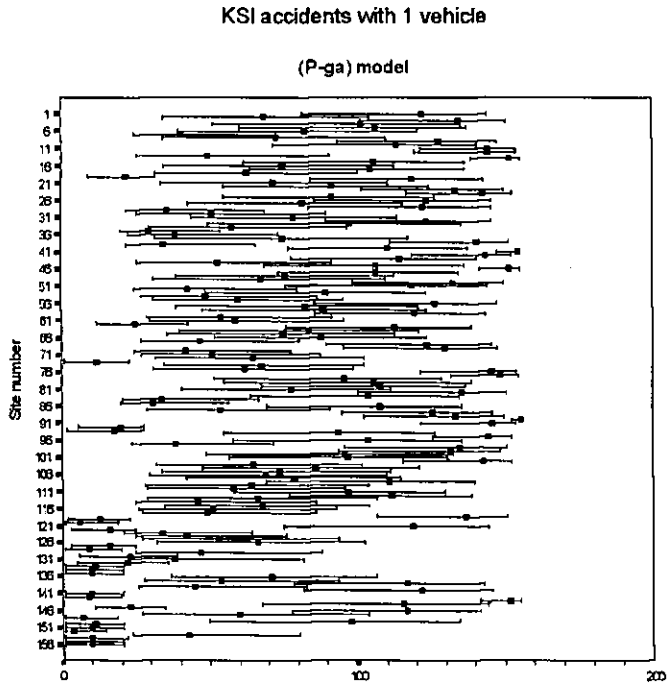


Figure 9.6: Ranks of means; Poisson-regression with gamma random effects model

ranked after 120, for example, where 156 is the worst. The plots like the one in Figure 9.6 can be used to draw a vertical line at the point rank 120 and

select all the sites whose credible intervals are intersected by this threshold line or to the right of this line. In this way it is accounted for the uncertainty in the calculation of ranks. Although two sites may have ranks with a difference of 20 between them, if their credible intervals overlap this means that it may be possible that the situation to be not so different, so both should be selected. This discussion applies to the other two models (P-MNre) and (P-MN1) and all other types of accident, as well. All four plots of this type, corresponding to the four type of accidents and also for the other two models (P-MNre) and (P-MN1) are given in the Appendix D.

The ranks and their credible intervals can be plotted ordering the sites firstly by the rank, secondly by the 2.5% percentile and thirdly by the 97.5% percentile. The pattern of the change in rank and associated credible interval can then be seen. For ranking based on ranks of λ_{ki}, these ordered plots are given in Appendix E, for all three models by accident type. All plots have a leaf shape pattern suggesting that the models give more credible ranks in the extremes, that is for sites with very low and very high ranks.

The advantage of using ranks of the residual terms after removing the covariates, is that exposure variables like traffic flow and link length is taken into account. It can be argued that a site A having double the length of a site B is "expected" to have a greater number of accidents if all the other conditions are the same. A similar argument can be followed for traffic flow. The idea is therefore to rank the "residual" information left after accounting for the covariates.

There are only two Poisson-regression models investigated. The residual terms are $\log(\mu_{ki}) + \beta_{i1}^{*}$ as calculated from the multiplicative equations of the type given in equation (9.2). The sites are presented as ordered by the ranks and corresponding percentiles. For the Poisson-regression model with gamma random effects, the four plots in Figures (9.7–9.10) show how uncertain the ranks may be. The similar plots provided by the Poisson-regression model with multivariate normal random effects, provided at the end of this chapter in Figures 9.11–9.14, tell a similar story. It might be useful to compare to ranks given by different models. This is done in the next section.

9.5.3 Comparison of ranks by three models

It is of course of interest to know how close the rankings are, as given by the three models investigated. An easy way to do that is to plot the estimated ranks given by one model against the estimated ranks given by another model. The comparison should be made for the same type of ranking. This means that either the models are compared for ranks of mean parameters λ_{ki}, as shown in Appendix F, or for the Poisson-regression models, the models are compared for ranks of $\log(\mu_{ki}) + \beta_{i1}^{*}$, as shown below. Overall it can be noticed immediately, from Figures 9.15–9.18, that the two Poisson-regression models provide similar rankings for all types of accident.

9.6 Summary

This chapter is a continuation of Chapter 8, applying the theoretical ideas emphasized there to some real data. A set of road accident data concerning road accidents in Kent on 156 single-carriageway link sites has been analysed for predictive purposes, for ranking the sites according to two criteria and for understanding the relationship between four types of accident and covariate information like link length, speed limit and estimated traffic flow. The main models investigated were (P-ga), (P-MNre), (P-MN1) and (P-logN). The inference process was possible due to MCMC methods and the results were compared from several points of view.

The first three models have been selected by DIC from a set of 11 models. Each models has its advantages and disadvantages and none should necessarily be rejected in favour of the others. There is some evidence that the accident numbers of different types are correlated and this could bias the analysis if multiple response accident frequencies were modelled separately at the univariate level.

This chapter provides an important tool for identifying hazardous locations and for forecasting the reduction in accidents that would result if the traffic could be reduced by a known factor. It was shown that the reduction is not similar for all four types of accidents investigated and generally depends on rural-urban areas.

The selection of hazardous sites followed some ideas reviewed in Chapter 2 for univariate models. The sites were categorised as dangerous according to

either the probability of a site to be the worst or the posterior distribution of the rank of a parameter of interest of a site. The results were then compared with the Poisson-regression with gamma random effects model as a base model.

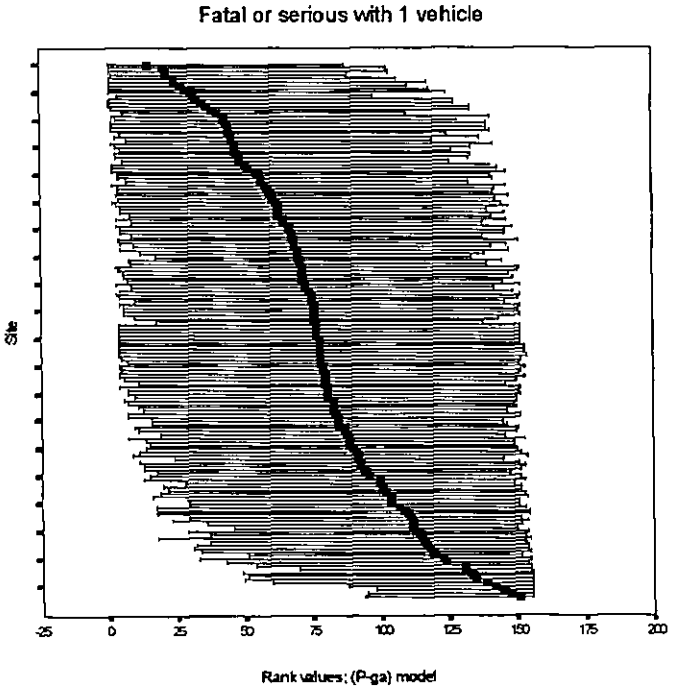


Figure 9.7: Ordered posterior medians and credible intervals of ranks; model (P-ga) for first type of accidents

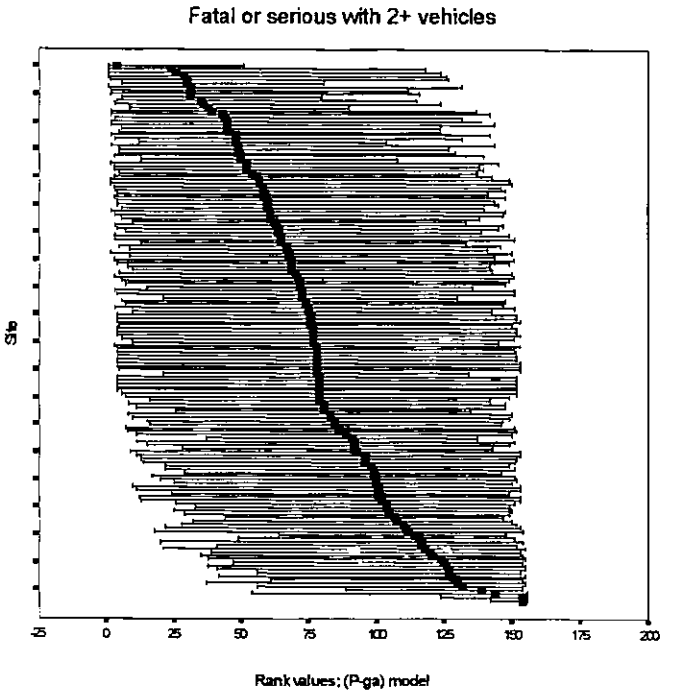


Figure 9.8: Ordered posterior medians and credible intervals of ranks; model (P-ga) for second type of accidents

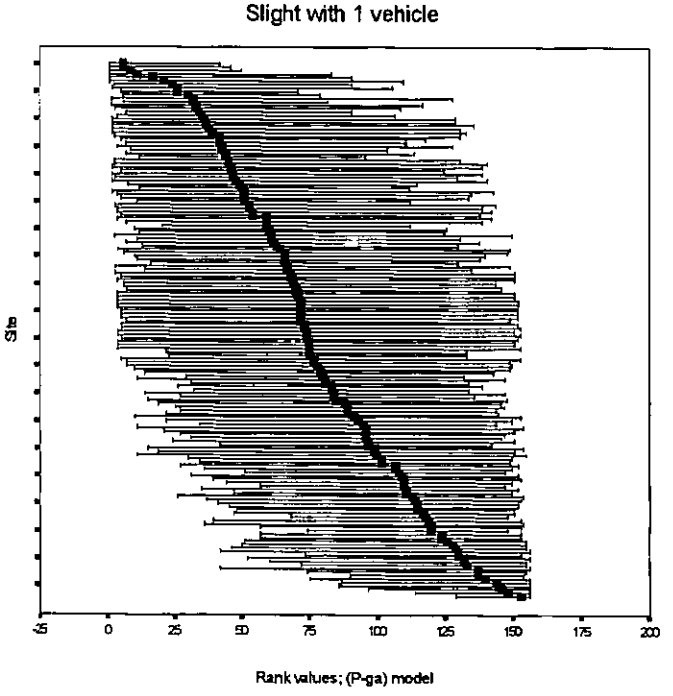


Figure 9.9: Ordered posterior medians and credible intervals of ranks; model (P-ga) for third type of accidents

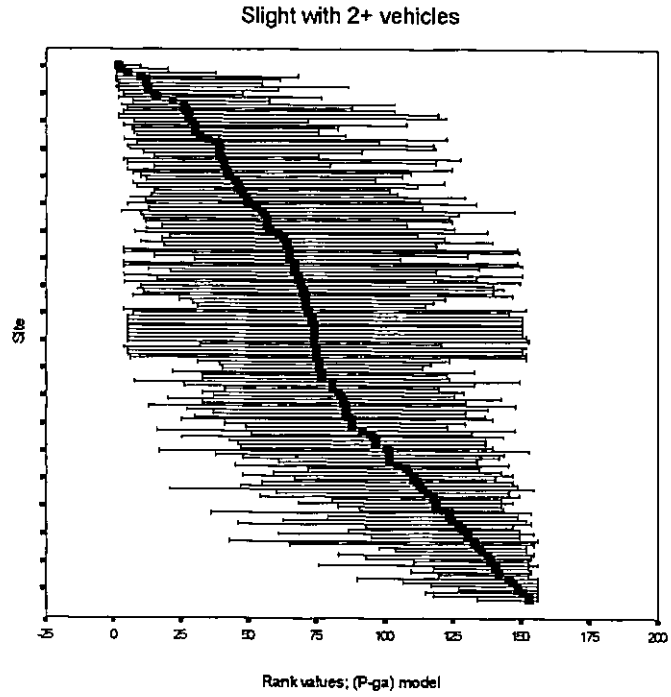


Figure 9.10: Ordered posterior medians and credible intervals of ranks; model (P-ga) for fourth type of accidents

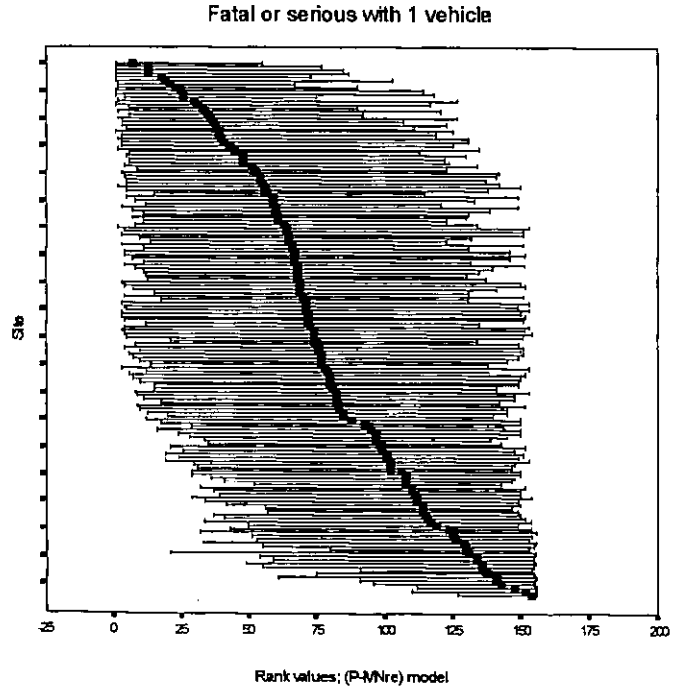


Figure 9.11: Ordered posterior medians and credible intervals of ranks; model (P-MNre) for the first type of accidents

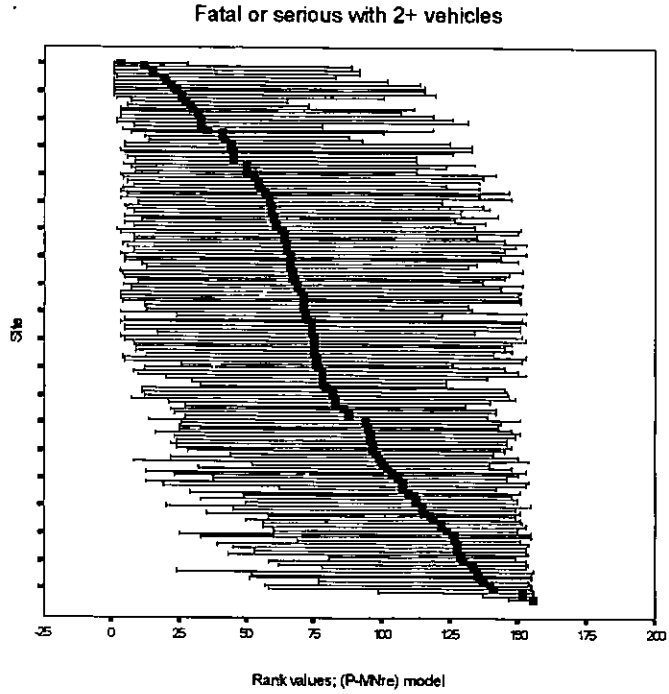


Figure 9.12: Ordered posterior medians and credible intervals of ranks; model (P-MNre) for the second type of accidents

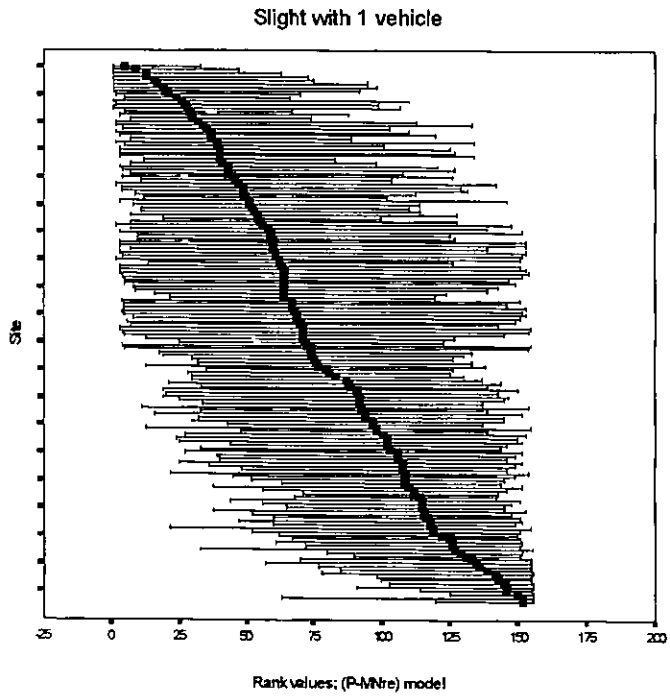


Figure 9.13: Ordered posterior medians and credible intervals of ranks; model (P-MNre) for the third type of accidents

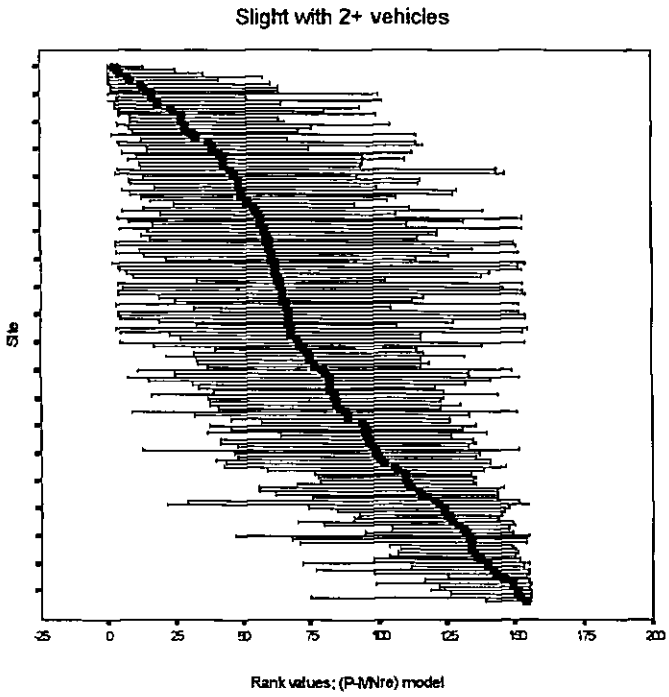


Figure 9.14: Ordered posterior medians and credible intervals of ranks; model (P-MNre) for the fourth type of accidents

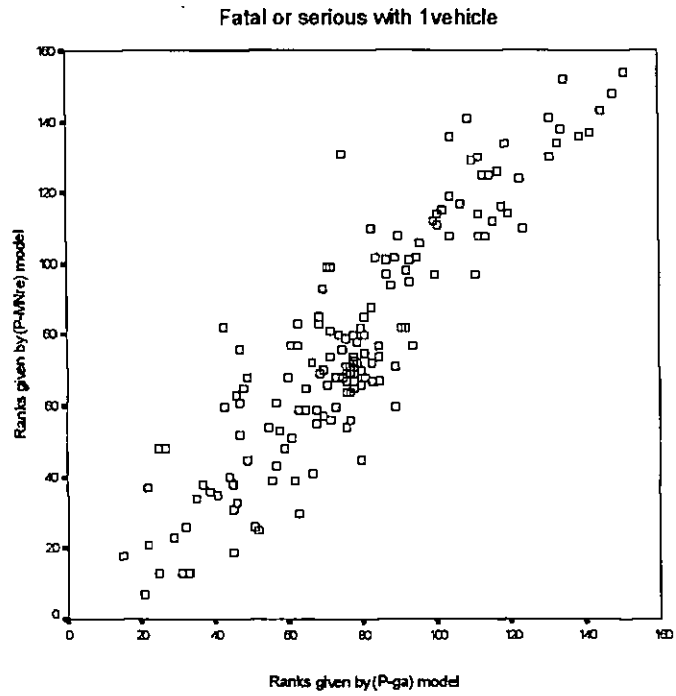


Figure 9.15: Comparison of posterior medians of ranks of residual information; fatal or serious accidents with 1 vehicle, (P-MNre) against (P-ga)

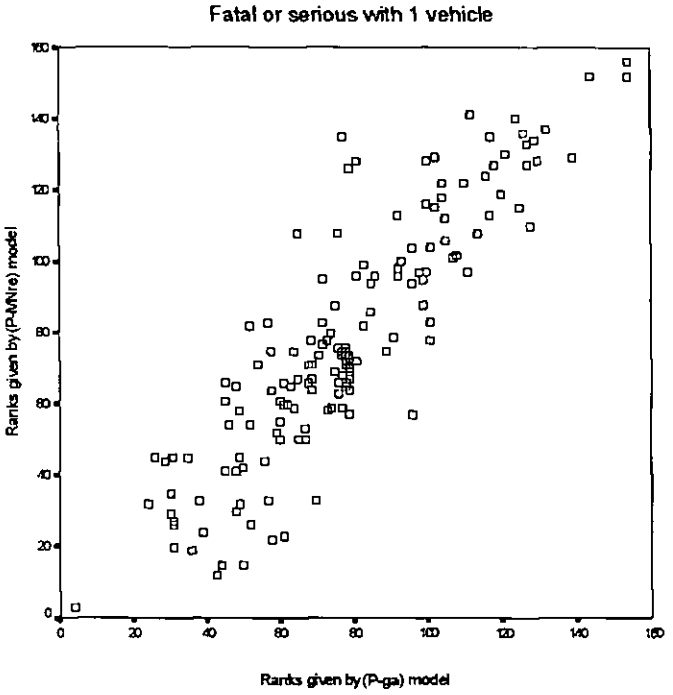


Figure 9.16: Comparison of posterior medians of ranks of means; fatal or serious accidents with 2+ vehicles, (P-MNre) against (P-ga)

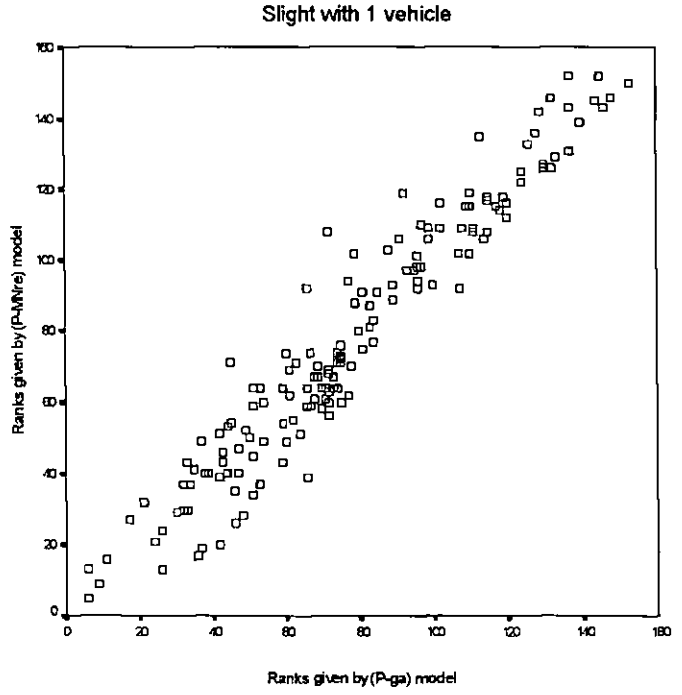


Figure 9.17: Comparison of posterior medians of ranks of means; slight accidents with 1 vehicle, (P-MNre) against (P-ga)

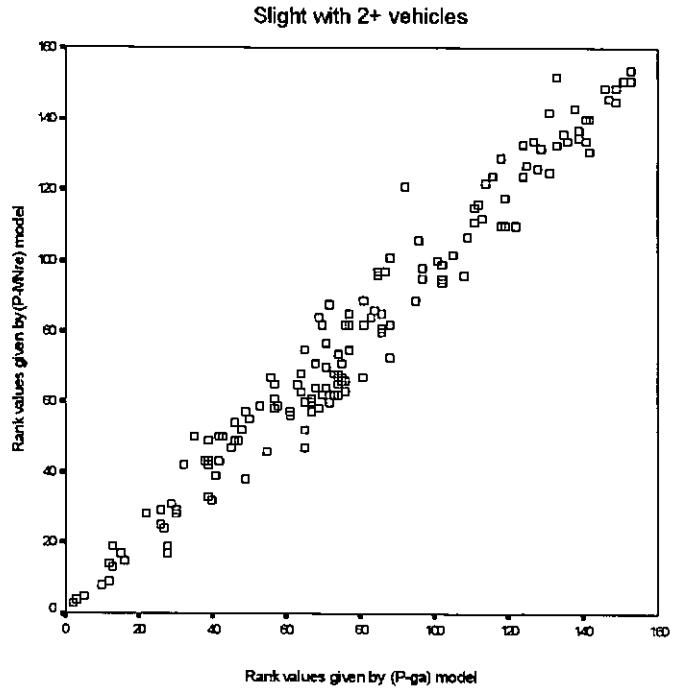


Figure 9.18: Comparison of posterior medians of ranks of means; slight accidents with 2+ vehicles, (P-MNre) against (P-ga)

Chapter 10

Conclusion

10.1 Summary of the thesis

10.1.1 Multivariate modelling of road accident data

The development of computer technology and computational techniques allows scientists to analyse more and more complex sets of data. Applied statistical modelling offers solutions for extracting valuable information from data. Simpson's paradox indicates that the modelling must be done at a multivariate level. One area of research which has not yet extensively exploited multivariate statistical modelling is road accident analysis. This thesis aimed to make a step forward and to develop statistical procedures that can be used in this area of research and possibly in other similar areas of research.

The thesis had two main directions of research given by the type of variables modelled, categorical variables representing characteristics of accidents in the first part, and multiple response variables representing accident numbers of

different type in the second part. The two parts were joined by using a similar tool of modelling, a graphical model represented by a graph. In the first part of the thesis this tool was mainly used for interpretation purposes at the end of the analysis, whereas in the second part of the thesis it was used to set up a hierarchical model before the actual fitting process.

The difficulty of analysing road accident data has several facets. Firstly, data is collected as an observational study, no randomisation being possible. Secondly, the data is bound to be sparse, either when it is summarised in a contingency table or when it is modelled by regression-like techniques bringing covariate information. Therefore, inference based on classical asymptotic tests is most of the time unreliable and other methods are needed. This has been clearly demonstrated for graphical models with about 10 variables, during the course of a comparative model selection in Chapter 5. This is also true for modelling multiple accident frequencies simultaneously, the task dealt with in the second part of this thesis. The disaggregation by accident type was not possible for a larger number of types because the data would have been so sparse that a statistical analysis could have not revealed reliable conclusions.

10.1.2 Graphical models

The complexity of road accident contingency tables requires multivariate statistical models and exact conditional testing. Graphical modelling is a useful multivariate statistical technique for disentangling the potential factors which influence important accident characteristics such as accident severity or the

number of casualties.

It was shown in Chapters 3, 4 and 5 that graphical modelling offers a very good solution for investigating road accident characteristics in an exploratory manner, being useful for small and large contingency tables. It was shown that speed limit, the number of vehicles involved and the number of casualties are directly associated with accident severity, one characteristic of major interest to road safety. Various other conditional independence relationships were established.

With a help of a small table it was shown that most of the log-linear models fitting the data could be nested into a graphical model. Therefore, even when the objective is to find some specific type of log-linear model it would be useful to identify first a graphical model fitting the data well and to refine the analysis starting from this model. The advantage of using a graphical model is that it is interpretable in terms of conditional independencies which can be visualised on a graph.

The analysis of large contingency tables summarising road accident data was further improved when substantive external knowledge was made available. This type of analysis had a causal flavour and the models, called graphical chain models, are a direct generalization of graphical models. The inference process for this class of models is a sequential one, but at each step, it is the same process as developed for graphical models. It was shown on an example in Chapter 5 how this process should be developed. Graphical chain models were developed for a set of data concerning the county of Bedfordshire, for a

set of data with accidents from Bedfordshire and Hampshire pooled together, and for two disaggregated sets of data for Bedfordshire. Reading conditional independencies on chain graphs can be sometimes difficult. It was shown using examples how to avoid traps by considering the moral graph of the smallest ancestral set of the subset of variables investigated.

A class of precursor models to graphical chain models consists of the response variable models introduced by Goodman (1973). Generally speaking, neither this class nor the class of graphical chain models coincides with the class of log-linear models. A result indicating when this equivalence is true, proved in Asmussen and Edwards (1983) using collapsibility, was restated in Chapter 6. Some examples and counterexamples using models encountered in the thesis were also exemplified in Chapter 6.

Collapsibility also helped in showing how the analysis of a 6-dimensional table could be refined using a 3-dimensional marginal table without having problems with Simpson's paradox. Furthermore, it was concluded that what seemed a natural graphical chain model for the collision-rollover table was not a log-linear model.

It was also noticed that Simpson's paradox can appear in a negligent analysis of contingency tables summarising road accidents. In conjunction with the need for analysing large tables this was one of the main motivations for applying graphical modelling to road accident tables. It was also shown how the concept of collapsibility of maximum likelihood estimators could be used to reduce safely the dimension of the analysis.

10.1.3 Hierarchical joint-response models

The hierarchical modelling approach with multiple responses and random effects, developed in Chapters 8 and 9, give a solution to the problem of modelling multiple response variables in a joint manner, that is a multivariate approach on the left hand side of the equations of the models as well as on the right hand side. The inference process can be done by employing MCMC techniques. The model output contains all the ingredients to answer various questions of interest, like predicting future values or ranking the observational units according to different measures.

A framework model was proposed and it was proved that, under its assumptions, this model offers a solution for modelling overdispersion and correlation of the observed counts. This general model can be followed by other researchers in developing other hierarchical models for other sets of road accident data and for other areas where modelling counts is of interest.

Using the models developed in this thesis, for the first time, practitioners can predict changes in accident type as well as the frequency. The predicted percentage reduction in accidents, if the traffic flow is reduced by a known factor δ, was calculated. The results were different for different types of accident and this could provide valuable information to local authorities.

In Chapter 7 some theoretical aspects regarding the compound Poisson distributions were re-examined, a new proof of when there is a maximum likelihood estimator for the two parameters of a negative binomial distribution was given, and a condition for this estimator to be unique was also identified.

An approximation result was given that helps in studying the sensitivity of changing priors in compound Poisson modelling.

MCMC techniques were successfully used for inferential purposes. One advantage of MCMC methods is that the same output can be used for answering many questions. The Gibbs sampler was also helpful in developing a new group of selection procedures of log-linear models for contingency tables, thus making a direct connection between the two parts of the thesis. It is likely that many other complex models proposed for road accident data will have computational problems that could be easily solved by MCMC methods.

Another problem investigated in the context of multiple accident frequencies was ranking the sites. The ability to rank the sites using multiple response models gives another dimension to practical efforts in this area, selecting the hazardous sites according to different criteria. The ranking process was done for the four types of accidents investigated by three models selected by the DIC criterion, that is (P-ga), (P-MNre) and (P-MN1). One ranking measure used was the probability that a site is the worst one. Other ranking measures used were the posterior distribution of the rank of the mean parameter λ_{ki}, for all three models, and the posterior distribution of the rank of the residual terms after removing the covariates, that is $\log(\mu_{ki}) + \beta_i^*$, for (P-ga) and (P-MNre). The posterior distribution was described by its median and 2.5% and 97.5% percentiles. The rankings given by different models were compared using some scatterplots and found to be similar.

10.2 Conclusion

The main conclusion emerging from this research is that it is better to start the analysis of accident data by a multivariate approach. For both types of accident data, either contingency tables or accident frequencies at sites, it is advisable and feasible to do this as shown in this thesis.

Sometimes, the graphical models proposed can be collapsed onto a smaller subset of variables. Then the analysis can be continued with other statistical techniques. An incorrect simplification of the analysis could lead to Simpson's paradox.

Graphical representations are an useful instrument for communicating results and models to a large audience. Graphs can help to extract conclusions from the statistical analyses by, for example, reading conditional independencies between subsets of variables on the conditional independence graph, or to specify models, like the fully Bayesian models analysed in WinBUGS.

It was revealed that speed limit and the number of vehicles involved influence directly accident characteristics responsible for road safety, like accident severity and the number of casualties. Other variables like road class, road surface conditions and the presence of a pedestrian crossing within 50m are not directly associated. The conditional independencies emphasized for the subset of STATS 19 data for Bedfordshire county, for Hampshire county and for those two sets of data pooled together, show that it is wrong to extend conclusions found at county level to a more aggregated level. The type of conclusions revealed by graphical models and graphical chain models developed

in the first part of the thesis can help local authorities in designing better policies and in planning research. The results may also be useful in fundamental research involving the conceptualisation of data structures of road accidents as described in Lupton, Wing and Wright (1998).

Since MCMC methods help to overcome many computational problems, almost any fully Bayesian model can be fitted and any arbitrary function of the parameters of the model can be posteriori estimated. This suggests that fully Bayesian models deserve more attention and more complex questions can be answered in this context.

A statistical approach that can be used for inference on *any* aspect of the data, modelling multiple accident frequencies of different type, was shown in the second part of the thesis. This is the first analysis of this type in this area of research.

The predictive accident models developed here can be used for a wide range of applications. The novelty of these models is that, for the first time in this area, qualitative as well as quantitative conclusions can be drawn at the same time. It was proved that a parallel approach, fitting several univariate regression models, leads to unreliable inference and should be avoided.

Practitioners use either the observed accident frequencies or the posterior mean of the expected number of accidents at a site, in an empirical Bayes approach, to rank hazardous sites. *Both* are wrong and a better approach is described and applied in this research. As emphasized in Chapter 9, ranking the sites ought to be done by the posterior distributions of ranks of the

expected accident rates, that is of the ranks r_{ki} of λ_{ki}. The posterior distribution is then used for a point estimation of the ranks and for calculating the associated credible intervals. This seems to be an almost impossible task for traditional methods because it is not easy at all to provide estimates of ranks of parameters. However, as it was shown in this thesis, under a fully Bayesian framework, it is possible to find a whole sample from the posterior distribution of any arbitrary functions of parameters, so for ranks as well. The ranks of observational units, such as sites in this thesis, are notoriously uncertain and a measure of uncertainty associated with rank estimates should be considered in the final analysis. Credible intervals are a perfect solution to this problem and there is no additional modelling effort for calculating them. Once we have the MCMC output for the model investigated, any empirical summaries can be calculated easily.

Another way of identifying the hazardous sites, presented in this thesis, is to calculate the posterior probability that a site is worst. This second method can be used for long term projects. Applying bad statistical techniques may have extremely bad consequences for the public. If some really hazardous sites are left out of the list of sites to be treated, then, not only will large amounts of public money be wasted, but human lives could be lost as well.

It was also shown that all three explanatory variables used in the second part of the thesis, that is speed limit, estimated traffic flow and link length, have a significant contribution in explaining accident frequencies. However, the interactions between speed limit and the other two explanatory variables

are not significant for all types of accident investigated. The predictive accident models developed in the second part imply that reducing the traffic flow will reduce the number of accidents and it was calculated by how much.

The hierarchical Bayesian models developed here for multiple response variables have been motivated by road accident data. However, they can be adapted to other areas of research where the modelling of counts is of interest.

10.3 Limitations of the research

The research carried out in the first part of this thesis focused on only two counties, Bedfordshire and Hampshire, due to time limitations and to the research for the second part of the thesis. However, a more general investigation would be very much appreciated from the practical point of view by local autorities.

The data used for developing graphical models in the first part of the thesis contained only accidents recorded in 1995. A larger set of data, containing road accidents from several years, may lead to other useful results. Unfortunately, this extension of the analysis to several sets of STATS 19 data was not possible given the period of time of this research.

Another idea not exploited here is to consider all counties in Great Britain with all accidents in the same period of time. Then an additional variable can be defined for county and it would be interesting to see how this spatial variable affects the conclusions revealed by graphical models. This would

be a vast project in itself, involving a lot of data preparation and a lot of computation.

Some categorical variables such as accident severity, the number of vehicles involved and the number of casualties involved are ordinal. It would have been ideal if it had been possible to take this information into account. There is little or no theory of graphical models for variables of this type, only marginal tests developed for log-linear models being implemented in MIM.

Moreover, the Bayesian model selection procedures proposed in this thesis may be improved and a software program able to handle large tables would be a big step forward.

There is no single package that can be used, in a user friendly manner, to develop the type of modelling proposed in this thesis. However, graphical models can be quite easily investigated with the package MIM (Edwards, 1995), and WinBUGS 1.2 (Spiegelhalter, Thomas and Best, 1998) is one of the most advanced packages that can handle hierarchical Bayesian models. A list of other packages having implemented various MCMC techniques for various statistical modelling methodologies is given in Carlin and Louis (1996).

An improved model selection procedure using Akaike information criterion is available on a new version of MIM. However, this version was not available when the research for the relevant part of this thesis was carried out.

The problem whether the maximum likelihood estimators of the two parameters of a negative binomial distribution are unique is very important. If the estimators are not unique then the results of the analysis should be carefully

interpreted. Although the condition given by equation (7.5) can be checked for any set of data, it would be useful to know a definitive answer. A simulation study can provide some hints.

There are some limitations concerning the elicitation of prior distributions. For the hierarchical Bayesian models the priors used in this thesis followed the general trend in the literature for modelling generalized linear models with random effects (Zeger and Karim, 1991; Spiegelhalter, Thomas and Best, 1998; Gilks et al., 1996). Some researchers may prefer more informative priors. The Bayesian methodology can be improved from this point of view and this is an area of intensive research. For Poisson-regression models, Doss and Narasimhan (1994) provided a computing environment within which one can immediately see the changes in the posterior distribution, corresponding to the changes in the prior distribution. Unfortunately, this program seems to be available only for Unix workstations. Subject matter information may help in developing better informative priors.

The specification of a covariance structure for the random effects μ_i in Chapter 8 is not straightforward. A possible model is described in the next section. The difficulty is due to the fact that the random effects account for information not included in the explanatory variables. Thus, it is difficult to interpret the covariance between two random effects.

10.4 Suggestions for further research

Graphical models for all counties

It was remarked in Section 10.3 that one limitation of this research was the focus on only two counties, Bedfordshire and Hampshire. As mentioned in Chapter 5, for the same set of variables, the data sets of different counties may be fitted by different graphical models. Without relying on unique models for each county, a question of interest would be what conditional independencies are supported by the data across the counties. More specifically, is accident severity independent of road type, daylight conditions and road surface conditions given the speed limit and the number of vehicles involved?

Another interesting question is what happens when there are several sets of data corresponding to several years for the same county, with the same variables investigated. For example, if there are data for Bedfordshire for all years between 1995 and 1998, relative to the six variables studied in Chapter 5, can a graphical model fit all these sets of data separately? Some theoretical developments on this direction are described in Lyngaard and Walther (1993).

Error in flow estimates

The traffic flow count at a site is usually a rough estimate because measurements are taken not over the entire period under study but over a limited interval (or intervals) of time. The flows should be calculated as AADTs over the entire time period for which the accident counts are taken. If Z_k is the real unknown AADT for site k in a multiplicative model, $\log Z_k$ would be one of

the explanatory variables. As Z_k is usually not known, an estimate Q_k is used, calculated over a period of time $t_k < T_k$. A functional model in which just one of the explanatory variables is flow, has been briefly described in Maher and Summersgill (1996). The flow with the true AADT Z_k is separated from the other variables

$$\mathrm{E}(Y_k) = \lambda_k = T_k \exp[\beta' X + \gamma \log Z_k]$$

and assuming that the estimated traffic flow Q_k is Poisson distributed with mean $Z_k t_k$, the log-likelihood is partitioned into two parts, one modelling the accidents and the other the flows. A fully Bayesian specification of this type of modelling is given by

$$\begin{aligned} Y_k \mid \lambda_k &\sim \mathrm{Pois}(\lambda_k) \\ \lambda_k &= \exp[\beta' X + \gamma \log Z_k] \\ Q_k \mid Z_k, t_k &\sim \mathrm{Pois}(Z_k t_k) \end{aligned}$$

$$\begin{aligned} \beta_j &\sim N(0, 0.0001) \\ \gamma &\sim N(0, 0.0001) \end{aligned}$$

and it can easily be extended to multiple response models along the lines described in Chapter 8.

A more complex hierarchical Bayesian model

The Poisson-regression model with gamma random effects specified in (8.24) does not impose a covariance structure on the random effects μ. The following model suggests a possible structure. For all $k = 1, 2, \ldots, N$ and $i = 1, 2, \ldots, M$

$$\begin{aligned} Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \text{Pois}(\lambda_{ki}) \qquad (10.1) \\ (\log \lambda_{ki}) = \theta_{ki} &= \log \mu_{ki} + X'_{ki}\beta_i \\ \mu_{ki} \mid b_{ki}, a_i &= b_{ki} + a_i \\ \beta_{ij} &\overset{iid}{\sim} \text{N}(0, 0.001) \\ b_{ki} &\overset{ind}{\sim} \text{gamma}(w_i, \delta) \\ a_1 = F_1 + V_1 \qquad & a_2 = F_1 + V_2 \\ a_3 = F_2 + V_1 \qquad & a_4 = F_2 + V_2 \end{aligned}$$

where F_1, F_2, V_1, V_2 are mutually independent and all independent of b_{ki}. The variables F_1, F_2 model missing information concerning accident severity and the variables V_1, V_2 concerning number of vehicles. It is also assumed that

$$\begin{aligned} F_1 &\sim \text{gamma}(f_1, \delta), \; F_2 \sim \text{gamma}(f_2, \delta) \\ V_1 &\sim \text{gamma}(v_1, \delta), \; V_2 \sim \text{gamma}(v_2, \delta) \end{aligned}$$

which implies that

$$a_1 \sim \text{gamma}(f_1 + v_1, \delta)$$

$$
\begin{aligned}
a_2 &\sim \text{gamma}(f_1 + v_2, \delta) \\
a_3 &\sim \text{gamma}(f_2 + v_1, \delta) \\
a_4 &\sim \text{gamma}(f_2 + v_2, \delta).
\end{aligned}
$$

The condition that $\mathrm{E}(\mu_{ki}) = 1$ for $i = 1, 2, \ldots, 4$ is equivalent to the following system of linear equations, subject to the strict positivity restrictions for all unknowns.

$$
\begin{aligned}
w_1 + f_1 + v_1 &= \delta \\
w_2 + f_1 + v_2 &= \delta \\
w_3 + f_2 + v_1 &= \delta \\
w_4 + f_2 + v_2 &= \delta.
\end{aligned}
\tag{10.2}
$$

It must be checked first that this system has proper solutions. This system of linear equations can be solved on computer, using for example MAPLE V. The idea behind this model was described in Maher (1991) and it was later followed in Loveday and Jarrett (1992).

The covariance structure of μ can be easily calculated as

$$
\mathrm{cov}(\mu_{ki}, \mu_{kj})_{i,j} = \begin{pmatrix}
\frac{w_1+f_1+v_1}{\delta^2} & \frac{f_1}{\delta^2} & \frac{v_1}{\delta^2} & 0 \\
 & \frac{w_2+f_1+v_2}{\delta^2} & 0 & \frac{v_2}{\delta^2} \\
 & & \frac{w_3+f_2+v_1}{\delta^2} & \frac{f_2}{\delta^2} \\
 & & & \frac{w_4+f_2+v_2}{\delta^2}
\end{pmatrix}
\tag{10.3}
$$

and using the system of equations (10.2) this can be further simplified to

$$\operatorname{cov}(\mu_{ki}, \mu_{kj})_{i,j} = \begin{pmatrix} \frac{1}{\delta} & \frac{f_1}{\delta^2} & \frac{v_1}{\delta^2} & 0 \\ & \frac{1}{\delta} & 0 & \frac{v_2}{\delta^2} \\ & & \frac{1}{\delta} & \frac{f_2}{\delta^2} \\ & & & \frac{1}{\delta} \end{pmatrix}. \tag{10.4}$$

It can be remarked that this model may be further refined by choosing the scale parameter of the gamma distribution of b_{ki} to be different from δ, the calculations being adjusted accordingly. An immediate consequence will be that the system of equations (10.2) is nonlinear and the covariance structure becomes more complicated.

Multiple response empirical Bayes models

Many researchers are more interested in empirical Bayes models rather than in a fully Bayesian approach. For univariate responses, these methods are thoroughly investigated in textbooks (Carlin and Louis, 1996; Maritz and Lwin, 1989) and applied on a large scale in modelling road accidents (Hauer, 1997; Mountain et al., 1996; Wright et al., 1988; Jarrett et al., 1982). However, for multiple responses, empirical Bayes methods are less developed. Taking either a nonparametric approach in Robbins' style (Robbins, 1955) or a parametric approach, the results of empirical Bayes models could usefully be compared to the fully Bayesian results developed in this thesis. One of the advantages of empirical Bayes methods is that they are not sensitive to prior

elicitation. The models can be still specified hierarchically in several stages but the parameters of the distribution at the penultimate level of the hierarchy are estimated from data. The estimation process can be very difficult.

10.5 A final comment

Statistical modelling is recognized as an art. All models are false, otherwise they will explain the data entirely, but some are useful. Road accident data is an example of large and complex data requiring advanced statistical techniques for a good analysis.

The graphical modelling methodology emphasized in this thesis can be applied in the future to a large range of studies in this area of research. Similarly, multiple response models as those proposed here can be adapted by other researchers to investigate other questions of interest related to traffic and safety transport. All these contributions can make a difference to a better world.

Appendix A

Proof of a collapsibility result

The following corollary of Theorem 6.1 shows how collapsibility helps in calculating the maximum likelihood estimates for large tables using known maximum likelihood estimates for marginal tables, and it will be used to prove a collapsibility result for response variable models in this appendix.

Corollary A.1 *Let the log-linear model L be collapsible onto a. Then*

$$\widehat{p}(i) = \widehat{p}_a(i_a) \prod_b \left[\widehat{p}_{\mathrm{cl}(b)}(i_{\mathrm{cl}(b)}) / \{ n(i_{\mathrm{bd}(b)})/N \} \right] \tag{A.1}$$

where the product is taken over all connected components b of a^c.

The next result, given in Asmussen and Edwards (1983), can be proved in a different, more explicit way as it is shown below.

Theorem A.1 *If $L \in \mathcal{L}$, then $L \in \mathcal{J}_a$ if and only if L is collapsible onto a. In that case $M = L_a$ and $C = [a] \cup L_b$, where $b = \mathrm{cl}(a^c)$.*

Proof. If $L \in \mathcal{J}_a$ then

$$p^L(i) = p^M(i_a)p^C(i_{a^c} \mid i_a) \tag{A.2}$$

so it follows easily that

$$\begin{aligned} p^L(i_a) &= \sum_{a^c} p^L(i) = \sum_{a^c} p^M(i_a)p^C(i_{a^c} \mid i_a) \\ &= p^M(i_a) \end{aligned}$$

If it can be shown that $M \subseteq L_a$ then it will follow that $p^L(i_a) \in p^{L_a}(i_a)$ and this is exactly the definition of collapsibility of L onto a. Note that p_a is denoted here by p^{L_a}. The inclusion can be shown using the log-linear expansions, and this is the main difference compared to the constructive proof given in Asmussen and Edwards (1983). It is obvious that

$$\begin{aligned} \log p^M(i_a) &= \sum_{f \subseteq a} u_f^a \\ \log p^C(i_{a^c} \mid i_a) &= \sum_{g_1 \subseteq a^c} u_{g_1}^{a \cup a^c} + \sum_{g_2 \subseteq a^c, g_3 \subseteq a} u_{g_2 \cup g_3}^{a \cup a^c} \end{aligned}$$

Therefore from equation (A.2) it follows that

$$\begin{aligned} \log p^L(i) &= \log p^M(i_a) + \log p^C(i_{a^c} \mid i_a) \\ &= \sum_{f \subseteq a} u_f^a + \sum_{g_1 \subseteq a^c} u_{g_1}^{a \cup a^c} + \sum_{g_2 \subseteq a^c, g_3 \subseteq a} u_{g_2 \cup g_3}^{a \cup a^c} \end{aligned}$$

and so

$$\log p^{L_a}(i_a) \;=\; \sum_{f \subseteq a} u_f^a.$$

Now it is obvious that $M = L_a$. To show that $C = [a] \cup L_b$ it follows from equation A.2 that

$$\begin{aligned} \log p^C(i_{a^c} \mid i_a) &= \log p^L(i) - \log p^M(i_a) \\ &= \sum_{g_1 \subseteq a \cup a^c} u_{g_1}^{a \cup a^c} - \sum_{g_2 \subseteq a} v_{g_2}^a \end{aligned}$$

and it is known that there are interaction terms corresponding to all subsets of a, that is a is a generator. Thus,

$$\log p^C(i_{a^c} \mid i_a) = \sum_{s_1 \subseteq a} u_{s_1}^a + \sum_{s_2 \subseteq a^c} u_{s_2}^{a \cup a^c} + \sum_{s_3 \subseteq a} u_{a^c \cup s_3}^{a \cup a^c}.$$

The first sum gives [a]; the second sum contains all u-terms from L that are given by variables in a^c and the third sum contains all u-terms from L that are given by variables in a connected with variables from a^c, that is those variables in $\mathrm{bd}(a^c)$. Therefore, the last two sums give a log-linear expansion of $L_{a^c \cup \mathrm{bd}(a^c)}$ which is L_b, with $b = \mathrm{cl}(a^c)$.

Conversely, if L is collapsible onto a then let $M = L_a, C = [a] \cup L_b$, where $b = \mathrm{cl}(a^c)$. Then $\widehat{p}^M(i_a) = \widehat{p}^{L_a}(i_a) = \widehat{p}^L(i_a)$ by the definition of collapsibility in Section 6.2. and it has to be proved that $\widehat{p}^L(i) = \widehat{p}^J(i)$, where $J = (M, C)$. Let $a^c = b_1 \cup \ldots \cup b_q$ be the connected components of a^c. Using the global

Markov property

$$\widehat{p}^{C}(i_{a^c} \mid i_a) = \widehat{p}^{C}(i_{a^c} \mid i_{\mathrm{bd}(a^c)})$$

and using the independence of the subsets of variables corresponding to the connected components

$$\widehat{p}^{C}(i_{a^c} \mid i_{\mathrm{bd}(a^c)}) = \prod_{k=1}^{q} \widehat{p}^{C}(i_{b_k} \mid i_{\mathrm{bd}(a^c)}).$$

It is obvious that

$$\widehat{p}^{C}(i_{b_k} \mid i_{\mathrm{bd}(a^c)}) = \frac{\widehat{p}(i_{b_k}, i_{\mathrm{bd}(a^c)})}{\widehat{p}(i_{\mathrm{bd}(a^c)})} = \frac{\widehat{p}(i_{\mathrm{cl}(b_k)}, i_{\mathrm{bd}(a^c)})}{\widehat{p}(i_{\mathrm{bd}(a^c)})} \tag{A.3}$$

the last equality following because $b_k \cup \mathrm{bd}(a^c) = \mathrm{cl}(b_k) \cup \mathrm{bd}(a^c)$. Since $\mathrm{bd}(a^c) \cap \mathrm{cl}(b_k) = \mathrm{bd}(b_k) \subseteq a$ and a is a generator for C the following Lemma, proved by Haberman (1974) the first part, and Lauritzen (1982) the second part, can be applied

Lema A.1 (Haberman-Lauritzen) *If a_1 and b_1 are two subsets of variables of the set of variables of interest X such that*

1. $a_1 \cup b_1 = X$
2. *a_1 and b_1 are separated by $a_1 \cap b_1$*
3. *$a_1 \cap b_1 \subseteq c_1$, where c_1 is a generator of the log-linear model L_1*

then

$$\widehat{p}(i) = \widehat{p}_{a_1}(i_{a_1})\widehat{p}_{b_1}(i_{b_1}) / \{\frac{n(i_{a_1 \cap b_1})}{N}\}, \text{ and } \widehat{p}(i_{a_1}) = \widehat{p}_{a_1}(i_{a_1}).$$

Taking $a_1 = \mathrm{bd}(a^c)$ and $b_1 = \mathrm{cl}(b_k)$, it follows from equation (A.3) and Haberman-Lauritzen lema that

$$\begin{aligned}\widehat{p}^{\mathcal{C}}(i_{b_k} \mid i_{\mathrm{bd}(a^c)}) &= \frac{\widehat{p}(i_{\mathrm{cl}(b_k)}, i_{\mathrm{bd}(a^c)})}{\widehat{p}(i_{\mathrm{bd}(a^c)})} \\ &= \frac{\widehat{p}_{\mathrm{bd}(a^c)}(i_{\mathrm{bd}(a^c)})\widehat{p}_{\mathrm{cl}(b_k)}(i_{\mathrm{cl}(b_k)})}{p(i_{\mathrm{bd}(a^c)})\{n(i_{\mathrm{bd}(b_k)})/N\}}\end{aligned}$$

which shows, putting all together that

$$\widehat{p}^{J}(i) = \widehat{p}_a(i_a)\prod_{k=1}^{q}\left[\widehat{p}_{\mathrm{cl}(b_k)}(i_{\mathrm{cl}(b_k)})/\{n(i_{\mathrm{bd}(b_k)})/N\}\right]$$

and using Corollary A.1 that $\widehat{p}^{J} = \widehat{p}^{L}$. Hence $L = J \in \mathcal{J}_a$ as required.□

Appendix B

Tables for graphical chain modelling

Table B.1: Accidents with pedestrian casualties in Bedfordshire, 1995; $\alpha = 0.01$

Variables	Model formula	Method
D, H, T	$[T][DH]$	Dec.
	$[T][DH]$	Unres.
	$[T][DH]$	Exact.
$L, R, S \mid D, H, T$	$[R][LS][HL][DHT]$	Dec.
	$[RS][HR][LS][HL][DHT][DS]$	Unres.
	$[R][LS][HL][DHT]$	Exact.
$P, N \mid L, R, S, D, H, T$	$[PRT][NS][DHLRST]$	Dec.
	$[NS][PT][DHLRST]$	Unres.
	$[NS][PT][DHLRST]$	Exact
$A, C \mid P, N, L, R, S, D, H, T$	$[AHRST][CS][DHLNPRST]$	Dec.
	$[AS][CS][DHLNPRST]$	Unres.
	$[A][CN][DHLNPRST]$	Exact

Table B.2: Accidents with pedestrian casualties in Bedfordshire and Hampshire, 1995; $\alpha = 0.05$

Variables	Model formula	Method
D, H, T	$[HT][DH]$	Dec.
	$[HT][DH]$	Unres.
	$[HT][DH]$	Exact.
$L, R, S \mid D, H, T$	$[HRT][LST][HLT][DHT]$	Dec.
	$[HST][HR][HLS][DHT]$	Unres.
	$[HST][HR][HLS][DHT]$	Exact.
$P, N \mid L, R, S, D, H, T$	$[DHLPRT][NS][DHLRST]$	Dec.
	$[NS][LPT][DHLRST]$	Unres.
	$[NS][DLPRT][DHLRST]$	Exact
$A, C \mid P, N, L, R, S, D, H, T$	$[ACDHPRST][CDHNPRST][DHLNPRST]$	Dec.
	$[AST][CNPS][DHLNPRST]$	Unres.
	$[ACHST][CHNST][DHLNPRST]$	Exact

Table B.3: Accidents with pedestrian casualties in Bedfordshire and Hampshire, 1995; $\alpha = 0.01$

Variables	Model formula	Method
D, H, T	$[T][DH]$	Dec.
	$[T][DH]$	Unres.
	$[T][DH]$	Exact.
$L, R, S \mid D, H, T$	$[HR][LST][HLT][DHT]$	Dec.
	$[ST][HR][HL][LS][DHT]$	Unres.
	$[HST][HR][HLS][DHT]$	Exact.
$P, N \mid L, R, S, D, H, T$	$[DHLPRT][NS][DHLRST]$	Dec.
	$[NPS][LPST][DHLRST]$	Unres.
	$[NS][DLPT][DHLRST]$	Exact
$A, C \mid P, N, L, R, S, D, H, T$	$[ACHP][ADHPRST][DHLNPRST]$	Dec.
	$[AS][CNP][DHLNPRST]$	Unres.
	$[ACHS][CHNS][DHLNPRST]$	Exact

Table B.4: Accidents without pedestrian casualties in Bedfordshire and Hampshire, 1995; $\alpha = 0.05$

Variables	Model formula	Method
D, H, T	$[HT][DH]$	Dec.
	$[HT][DH]$	Unres.
	$[HT][DH]$	Exact.
$L, R, S \mid D, H, T$	$[HLRST][DHRST]$	Dec.
	$[HLRST][DHRT]$	Unres.
	$[HLRST][DHRT]$	Exact.
$N \mid L, R, S, D, H, T$	$[DHLNRST]$	Dec.
	$[DHNRST][DHLRST]$	Unres.
	$[DHLNRST]$	Exact
$A, C \mid N, L, R, S, D, H, T$	$[ACDHLNRST]$	Dec.
	$[ALNS][CHNRS][DHLNRST]$	Unres.
	$[ACDHNS][CDHLNST][DHLNRST]$	Exact

Table B.5: Accidents without pedestrian casualties in Bedfordshire and Hampshire, 1995; $\alpha = 0.01$

Variables	Model formula	Method
D, H, T	$[HT][DH]$	Dec.
	$[HT][DH]$	Unres.
	$[HT][DH]$	Exact.
$L, R, S \mid D, H, T$	$[HLRST][DHRST]$	Dec.
	$[HLRST][DHRT]$	Unres.
	$[HLRST][DHRT]$	Exact.
$N \mid L, R, S, D, H, T$	$[DHLNRST]$	Dec.
	$[DHNRST][DHLRST]$	Unres.
	$[DHNRST][DHLRST]$	Exact
$A, C \mid N, L, R, S, D, H, T$	$[ACDHLNRST]$	Dec.
	$[ALNS][CHNRS][DHLNRST]$	Unres.
	$[ACHNS][CDHNST][DHLNRST]$	Exact

Appendix C

Comparison of the (P-ga) and (P-logN) models

The plots on the left correspond to model (P-ga) (model 1 here) given in Chapter 9 by equations (9.1) and those plots on the right correpond to model (P-logN) (model 2 here) given in Chapter 9 by equations (9.12). The fit is better for model (P-ga) for each type of accident.

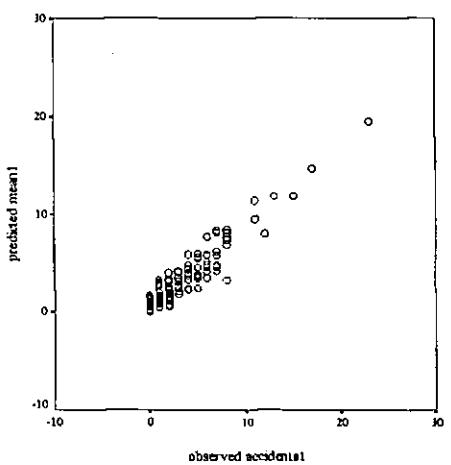


Figure C.1: KSI with 1 vehicle for Model 1

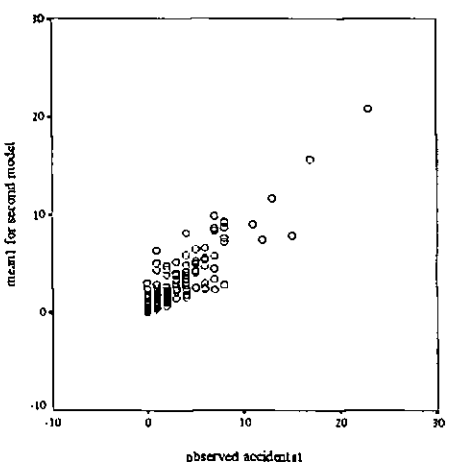


Figure C.2: KSI with 1 vehicle for Model 2

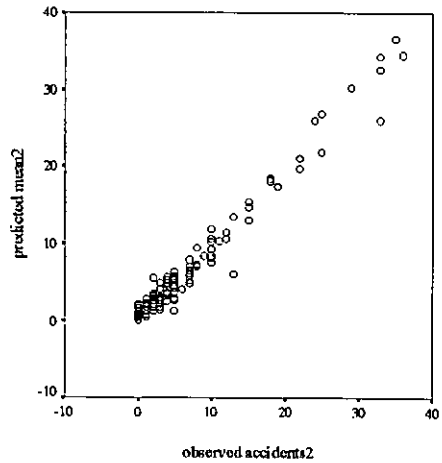


Figure C.3: KSI with 2+ vehicles for Model 1

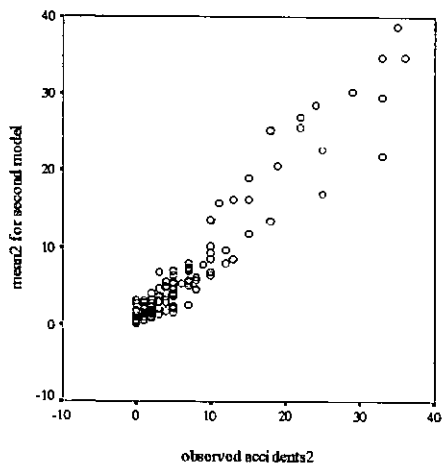


Figure C.4: KSI with 2+ vehicles for Model 2

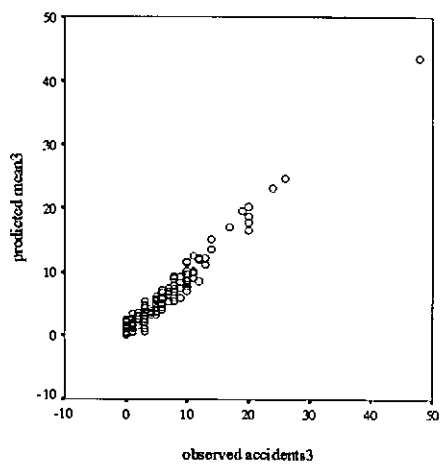


Figure C.5: S with 1 vehicle only for Model 1

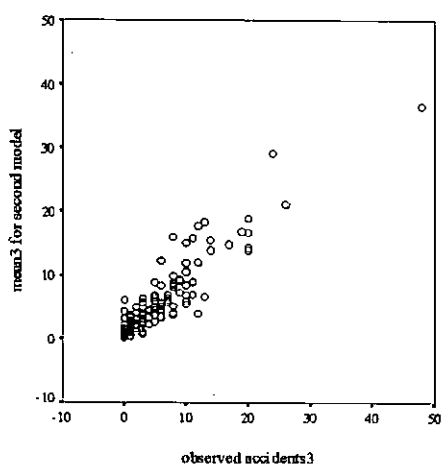


Figure C.6: S with 1 vehicle only for Model 2

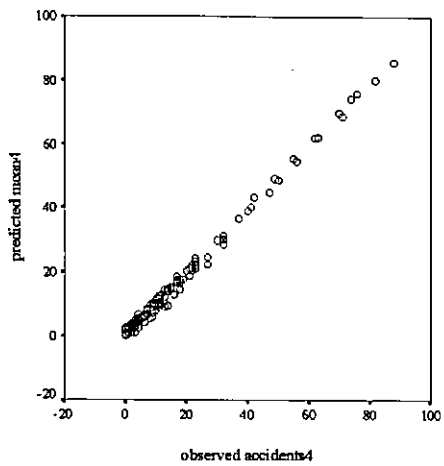


Figure C.7: S with 2+ vehicles for Model 1

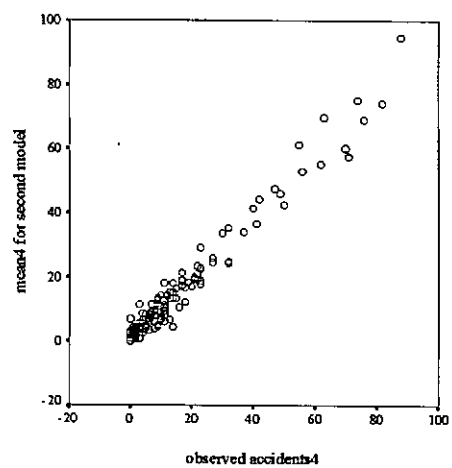


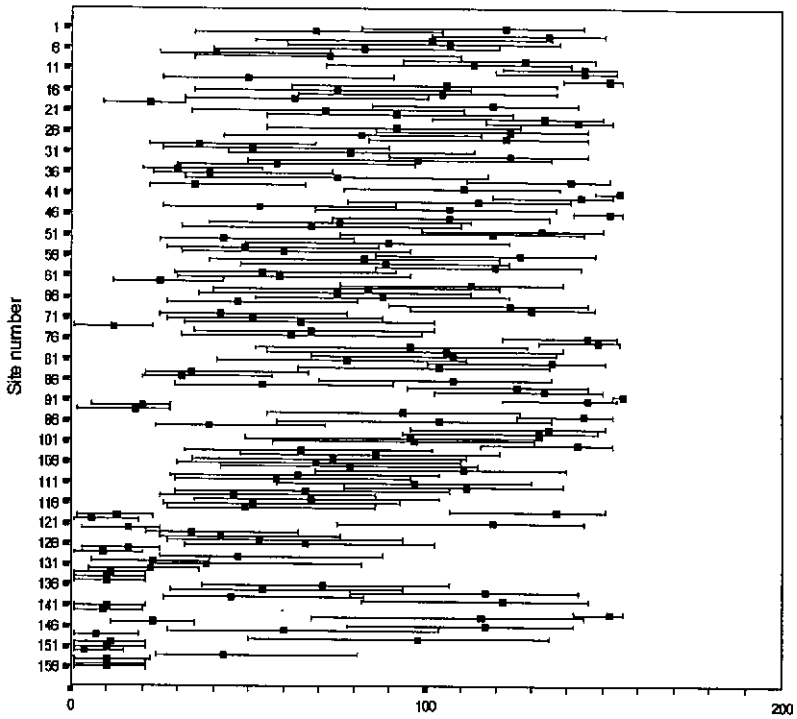
Figure C.8: S with 2+ vehicles for Model 2

Appendix D

Ranks with credible intervals

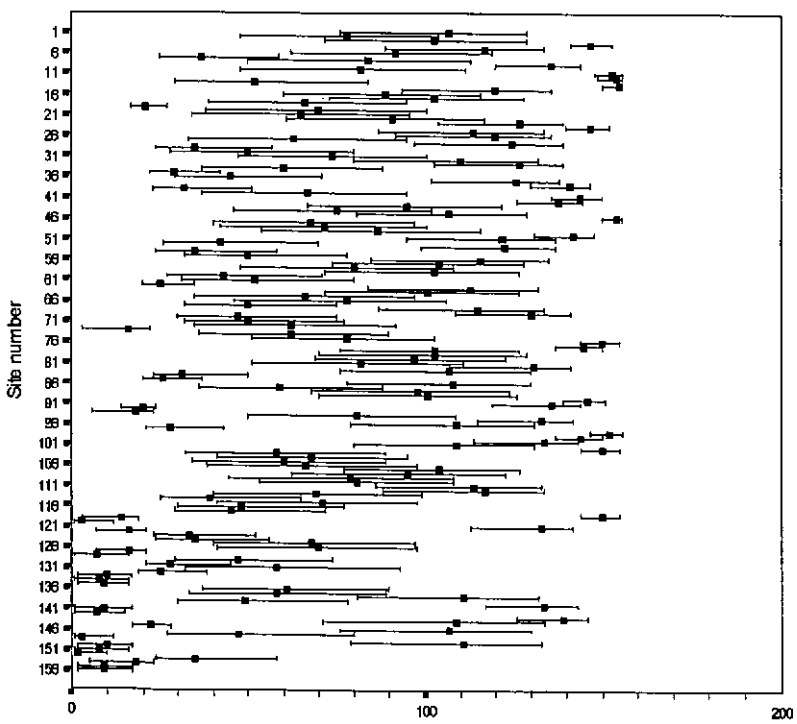
KSI accidents with 1 vehicle

(P-ga) model



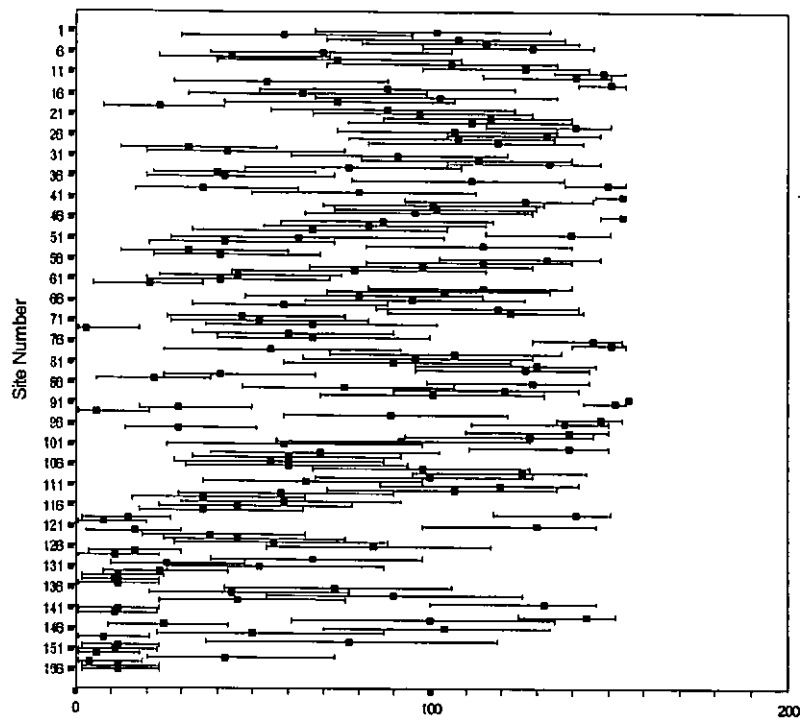
KSI accidents with 2+ vehicles

(P-ga) model



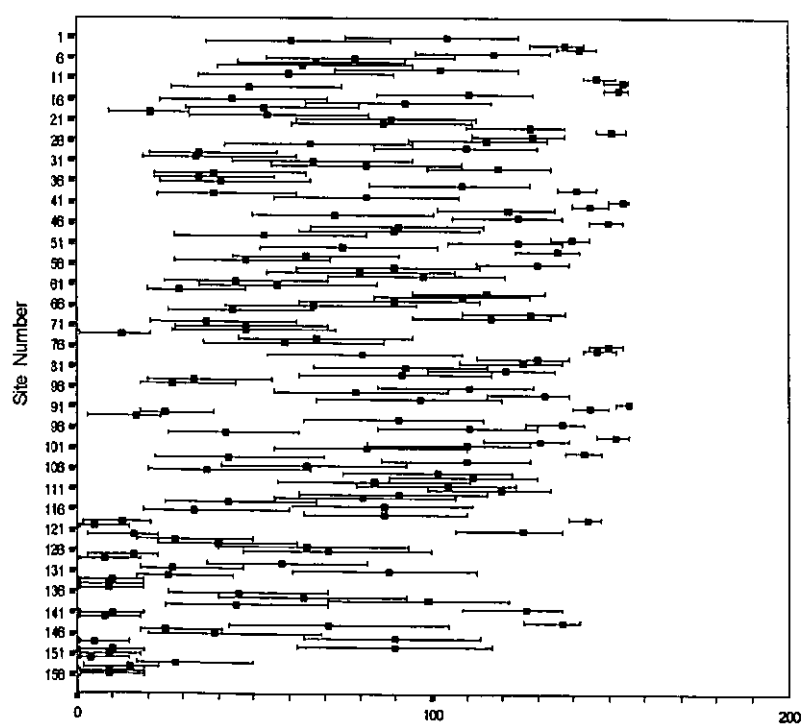
S accidents with 1 vehicle

(P-ga) model



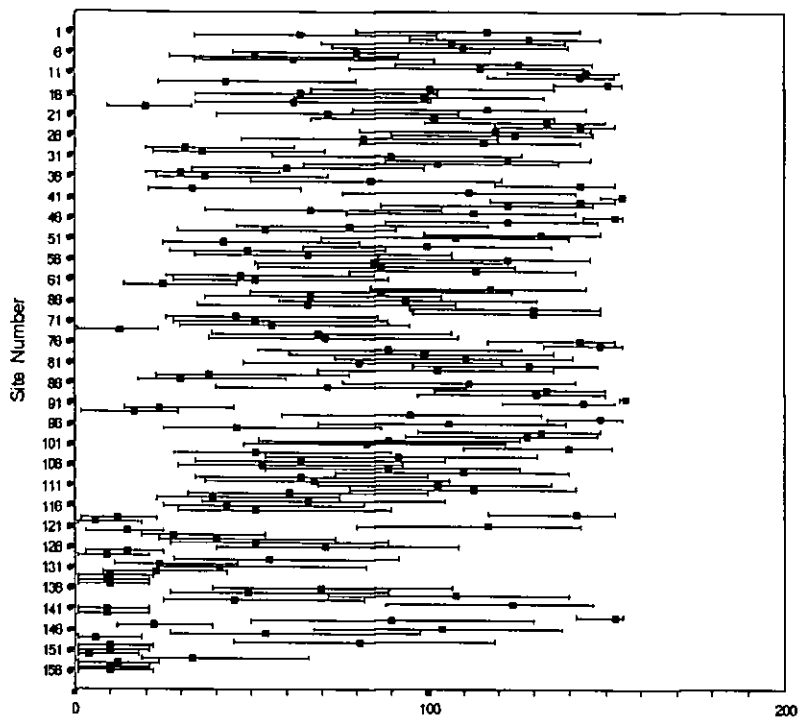
S accidents with 2+ vehicles

(P-ga) model



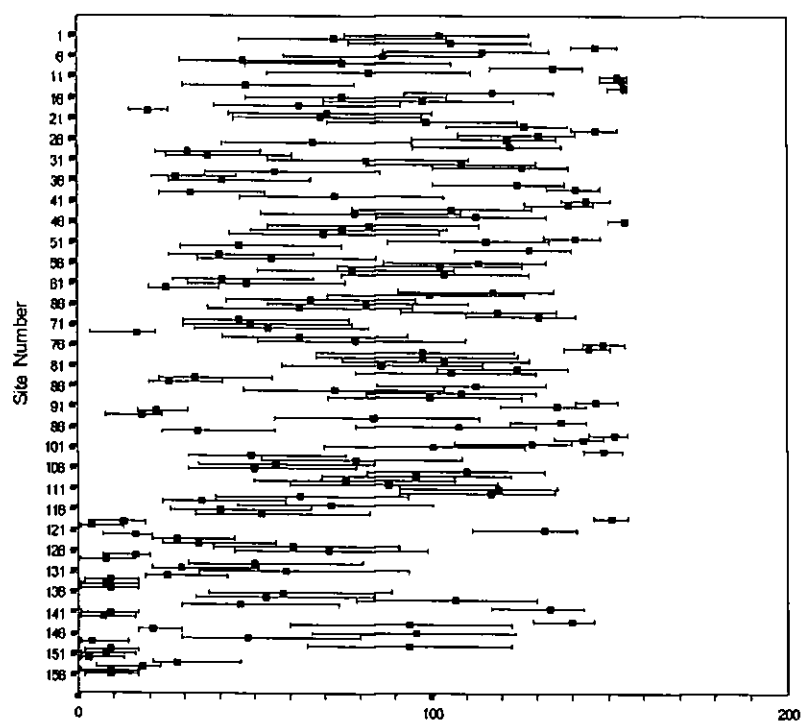
KSI accidents with 1 vehicle

(P-MNre) model



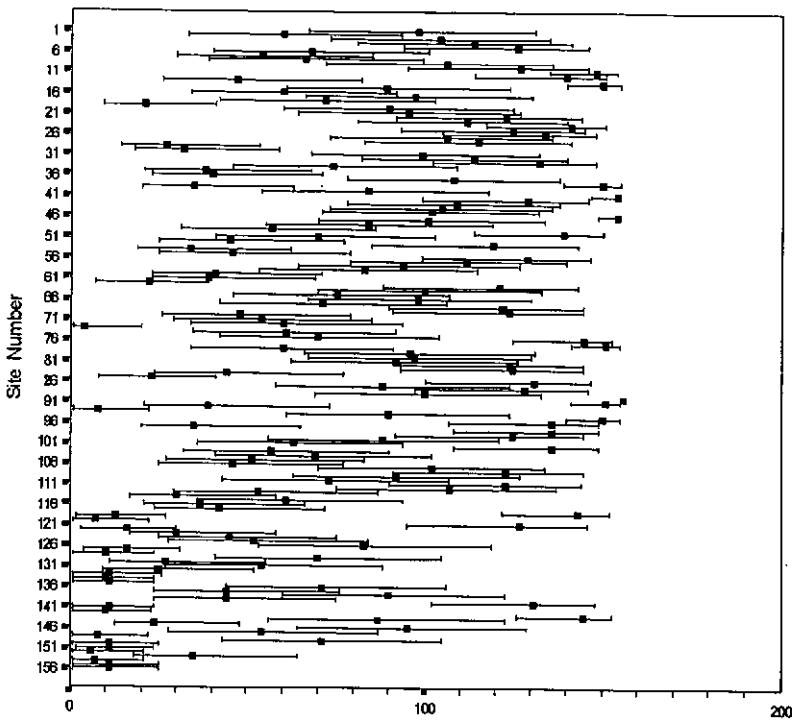
KSI accidents with 2+ vehicles

(P-MNre) model



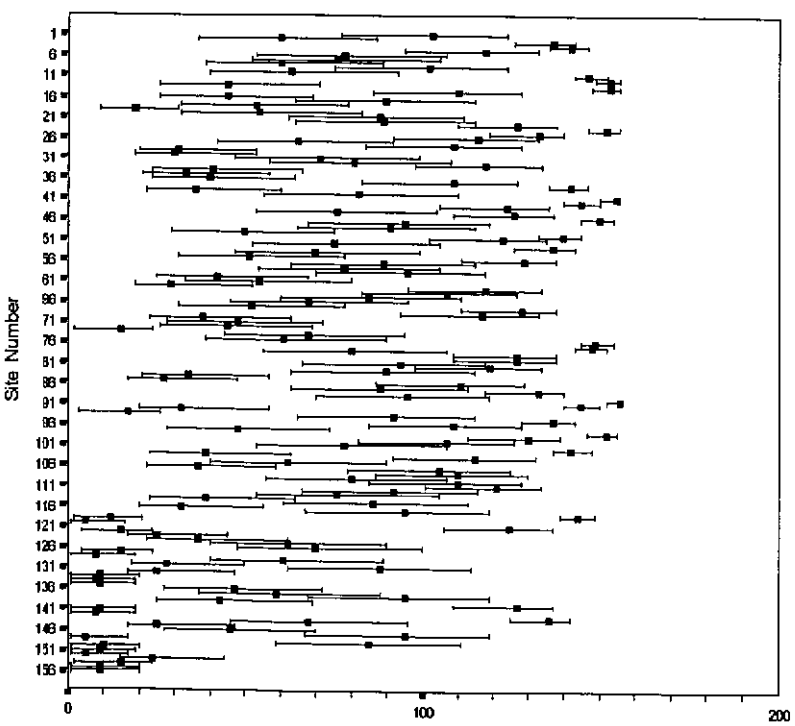
S accidents with 1 vehicle

(P-MNre) model



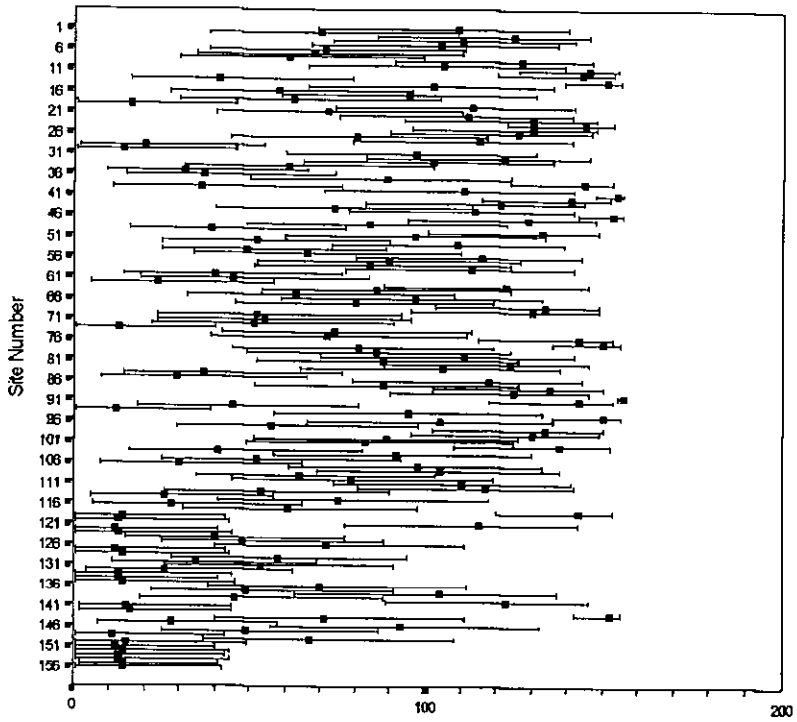
S accidents with 2+ vehicles

(P-MNre) model



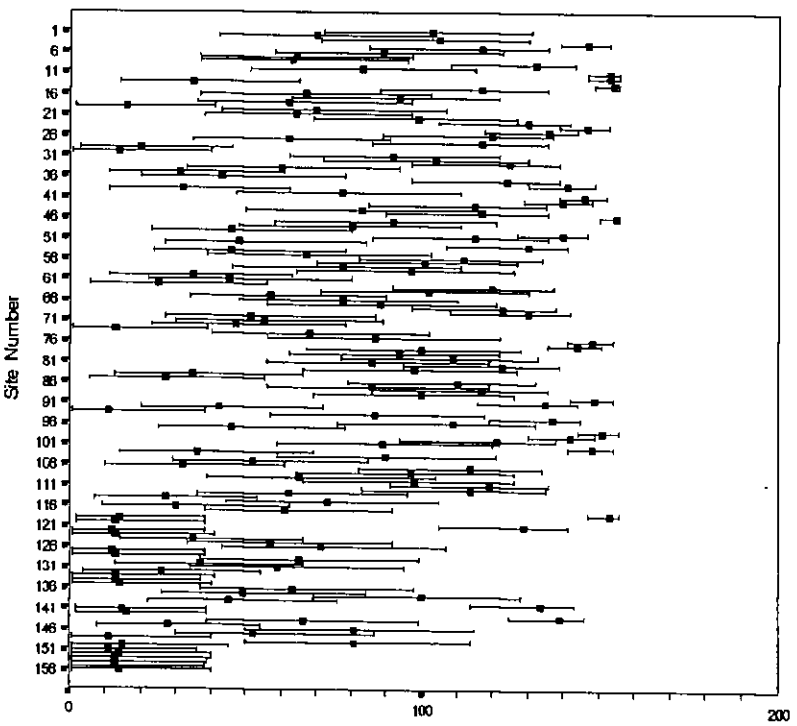
KSI accidents with 1 vehicle

(P-MN1) model



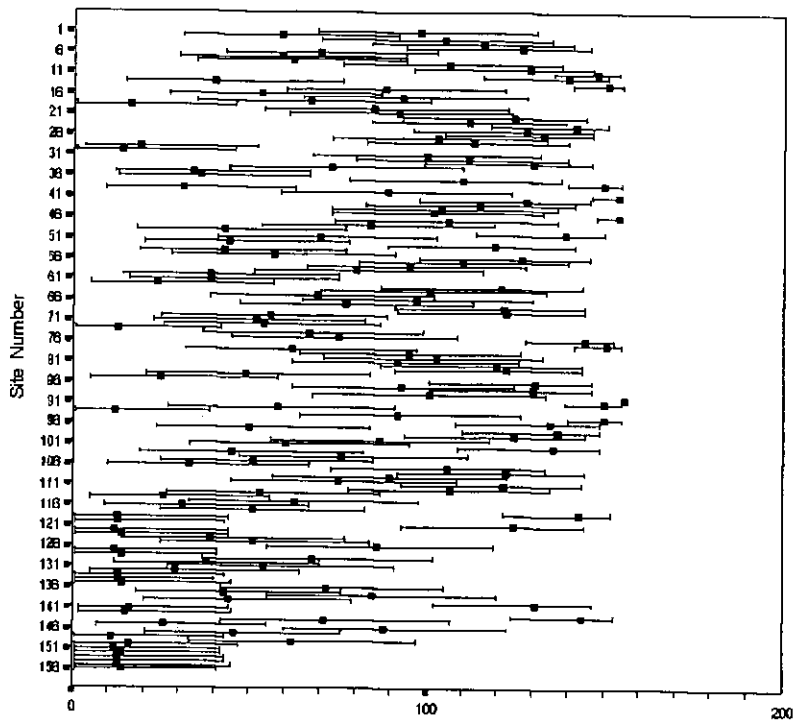
KSI accidents with 2+ vehicles

(P-MN1) model



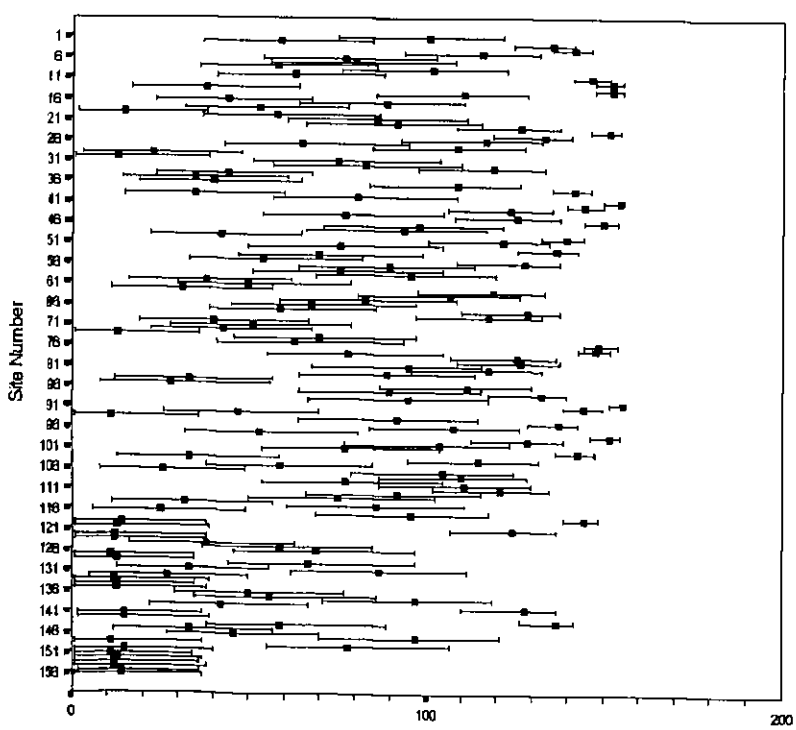
S accidents with 1 vehicle

(P-MN1) model



S accidents with 2+ vehicles

(P-MN1) model

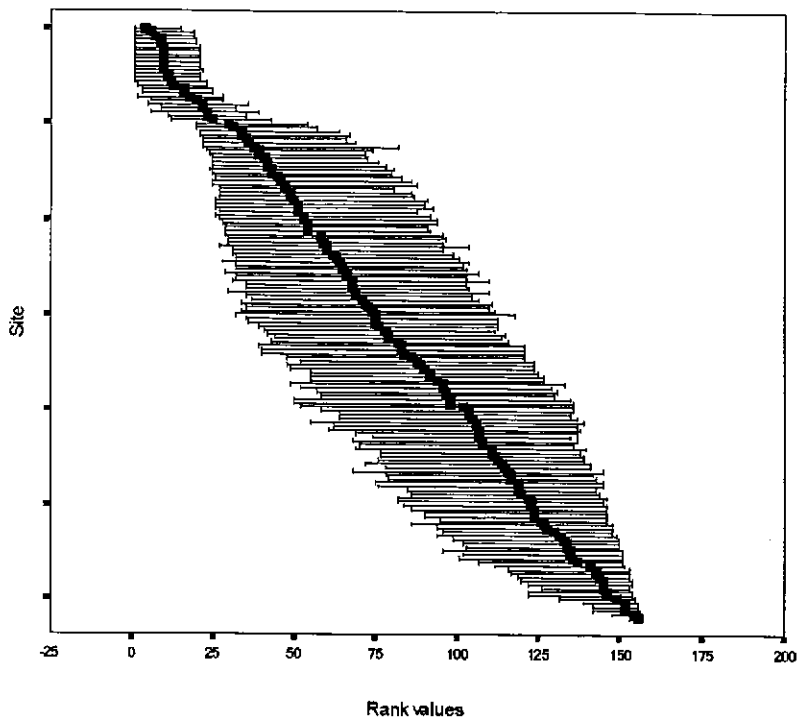


Appendix E

Ordered ranks with credible intervals

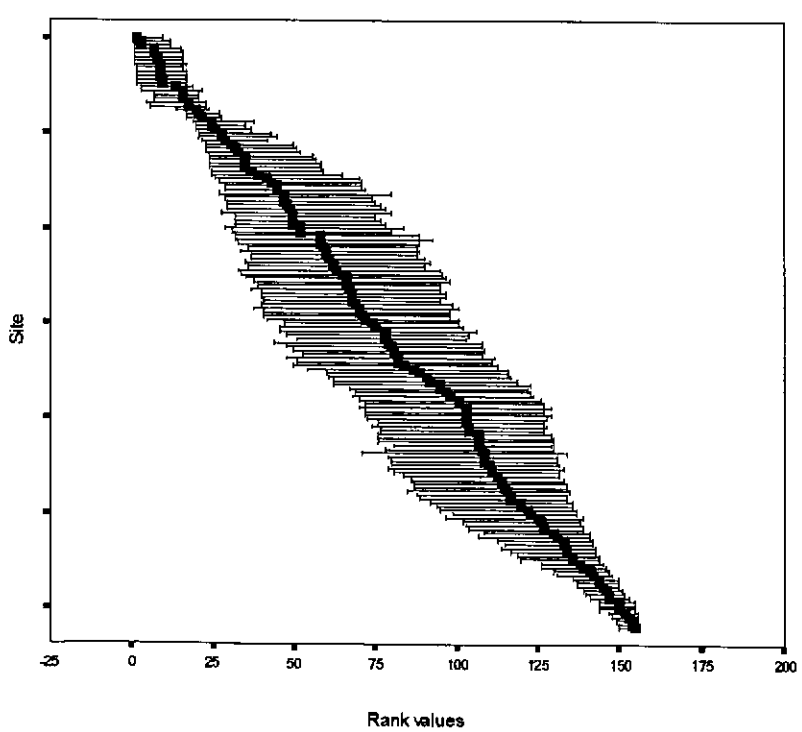
KSI with 1 vehicle

(P-ga) model



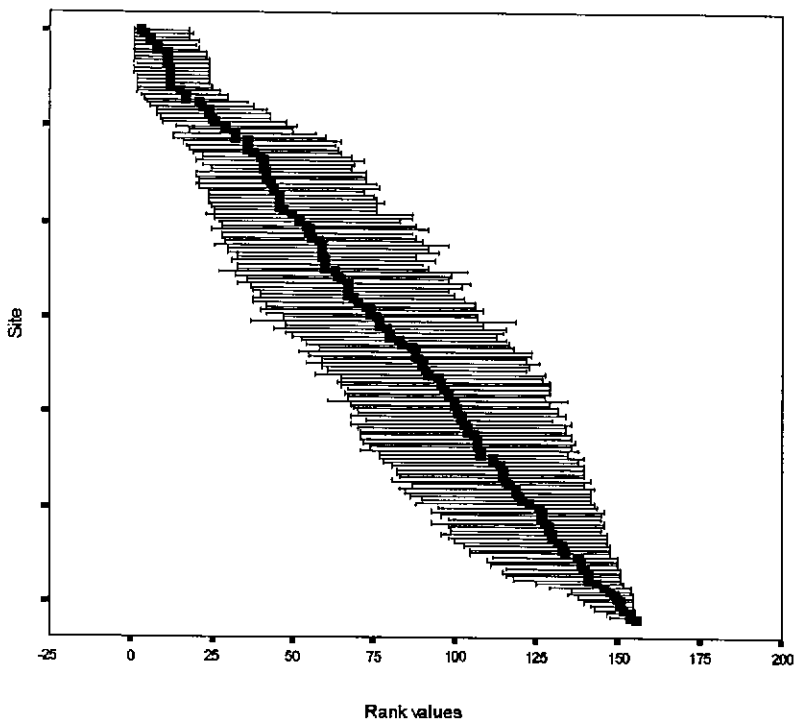
KSI with 2+ vehicles

(P-ga) model



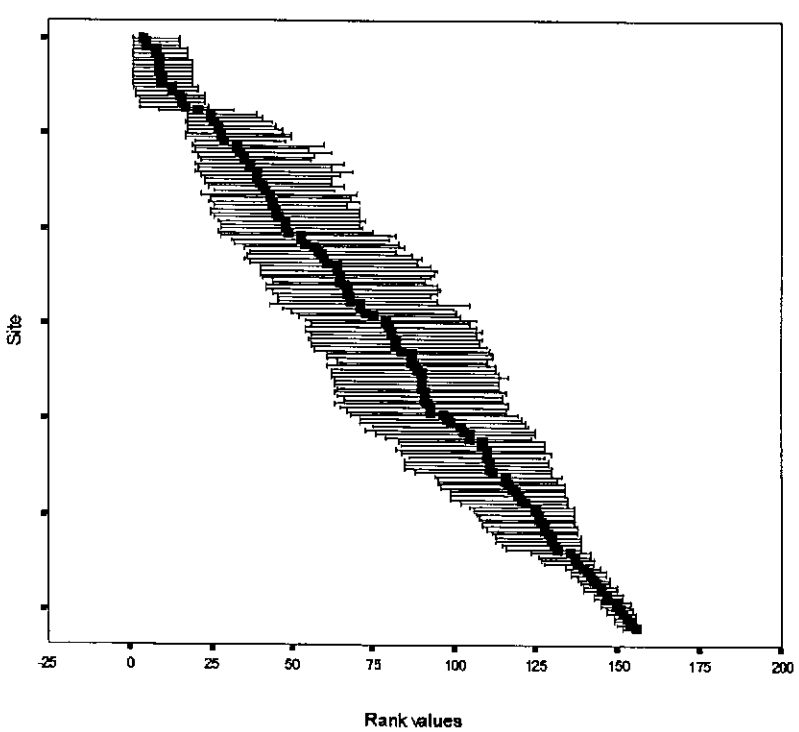
S with 1 vehicle

(P-ga) model



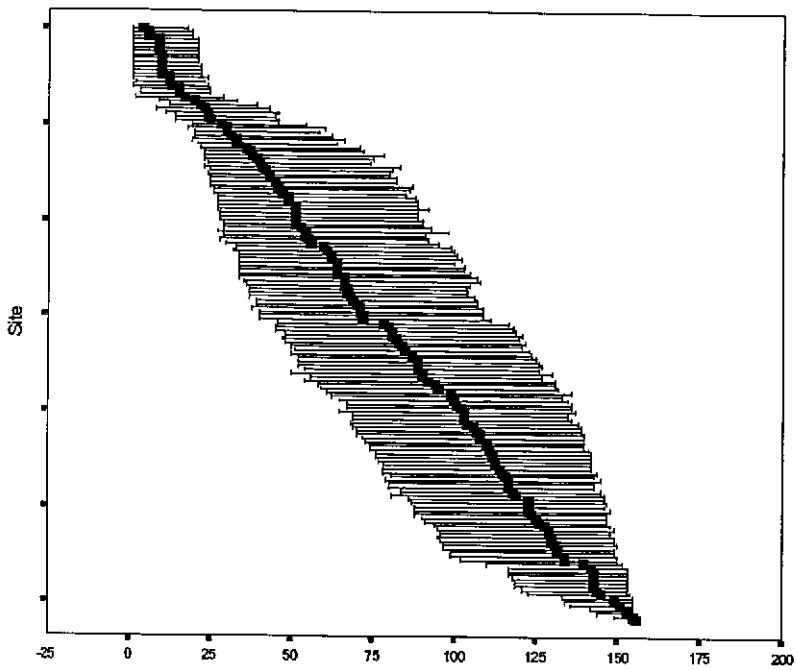
S with 2+ vehicles

(P-ga) model



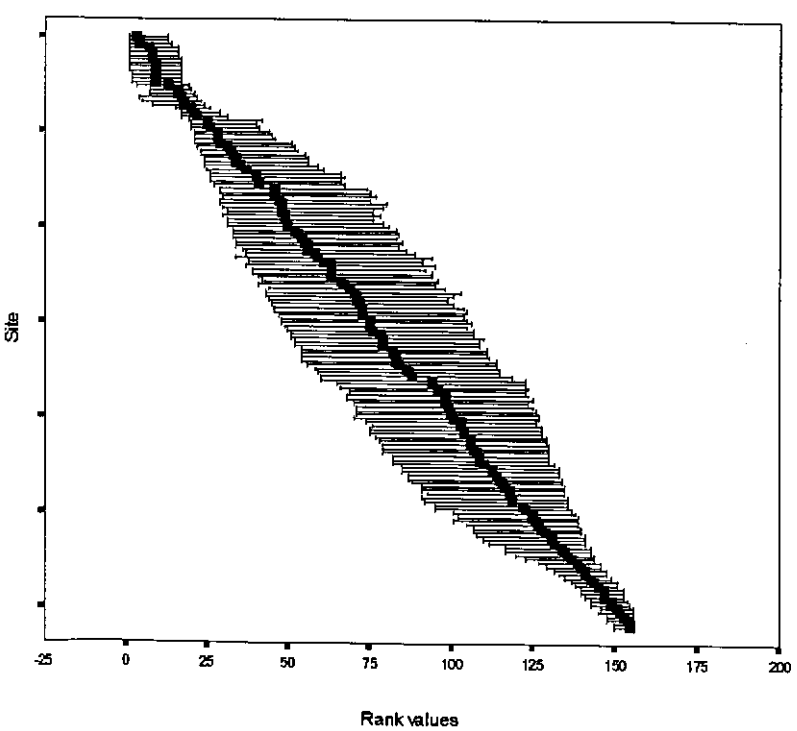
KSI with 1 vehicle

(P-MNre) model



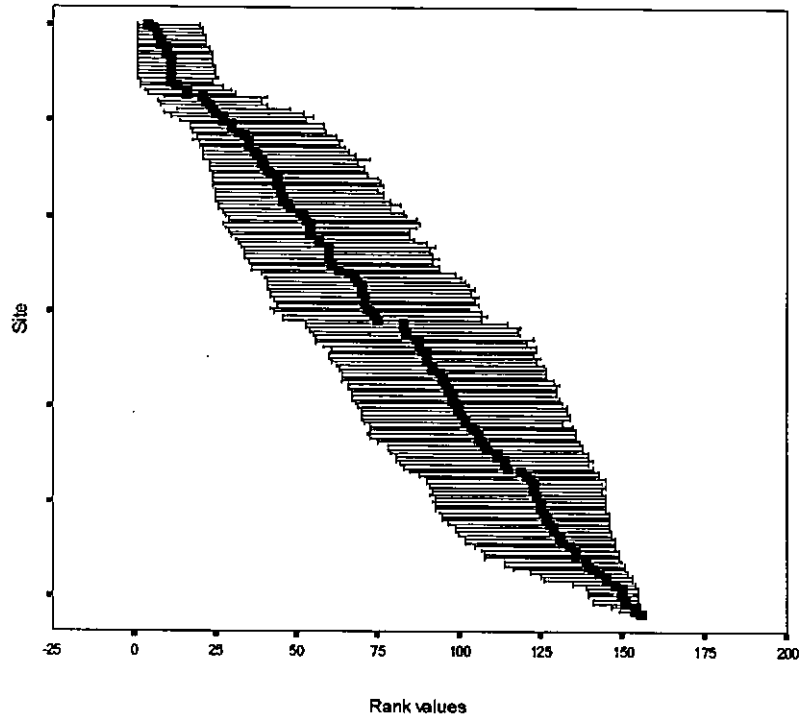
KSI with 2+ vehicles

(P-MNre) model



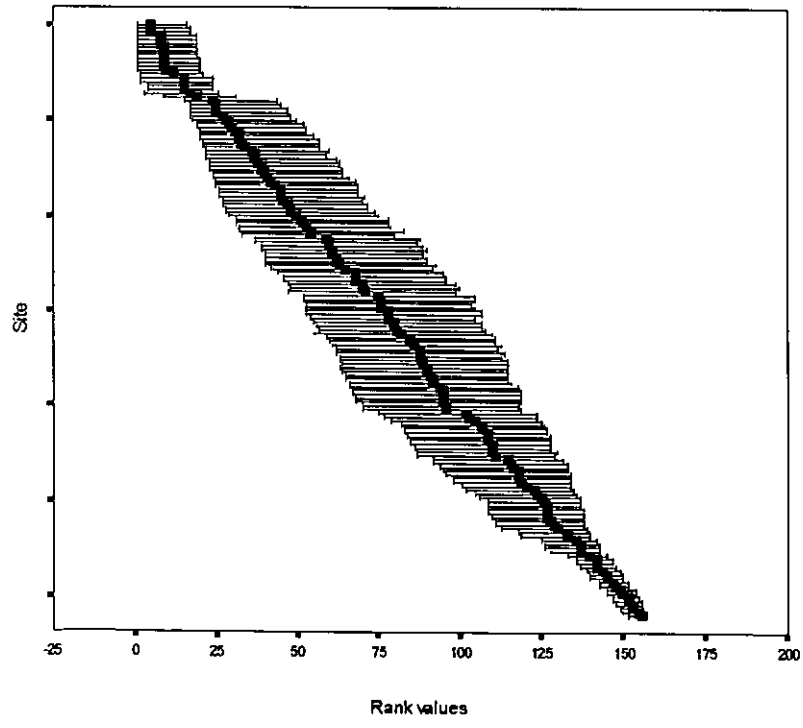
S with 1 vehicle

(P-MNre) model



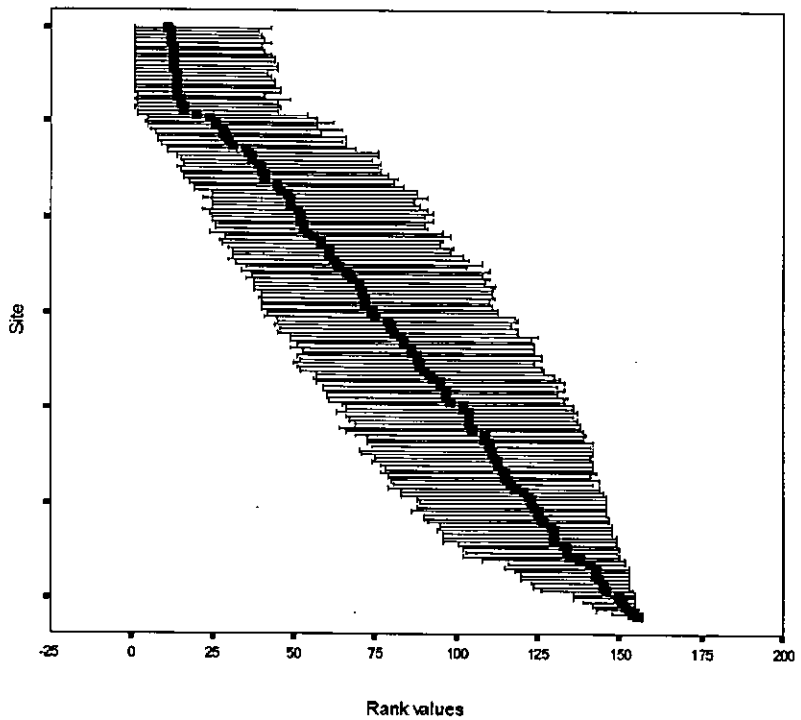
S with 2+ vehicles

(P-MNre) model



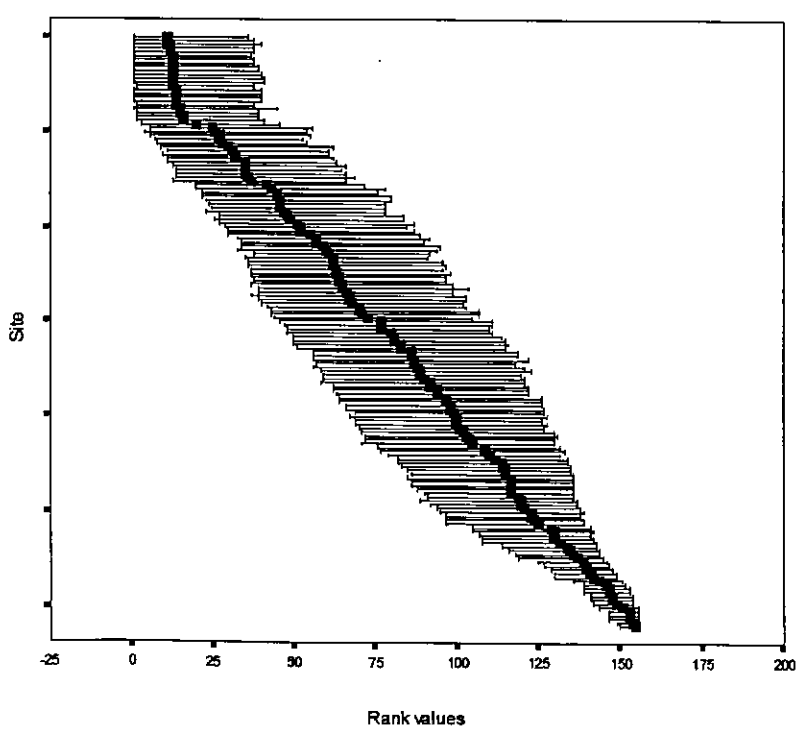
KSI with 1 vehicle

(P-MN1) model



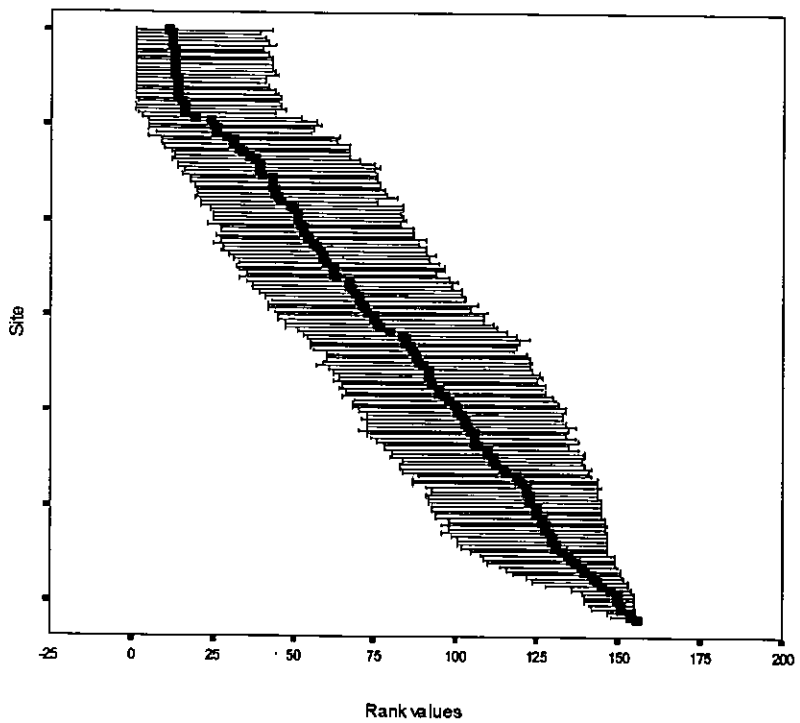
KSI with 2+ vehicles

(P-MN1) model



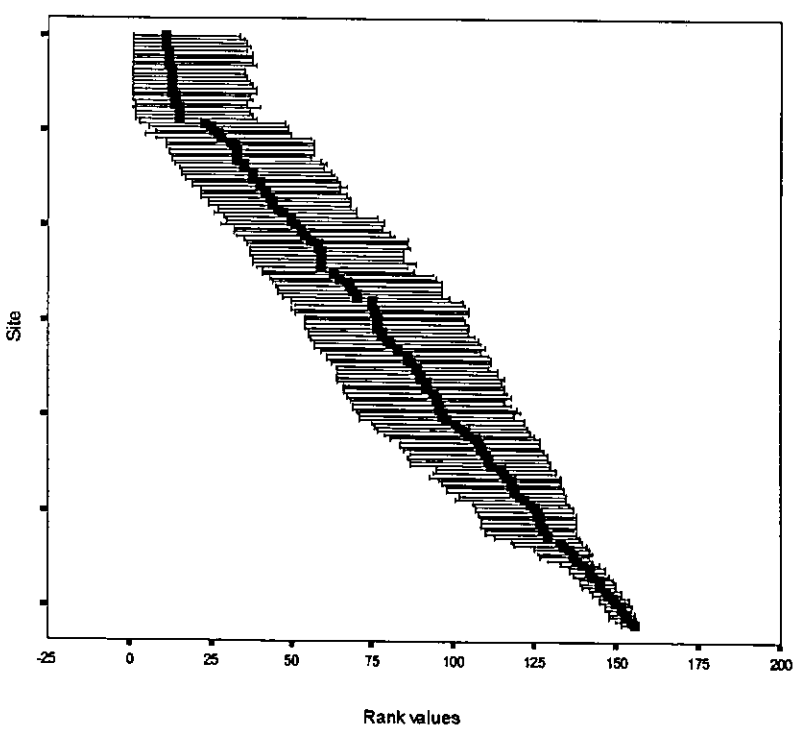
S with 1 vehicle

(P-MN1) model



S with 2+ vehicles

(P-MN1) model

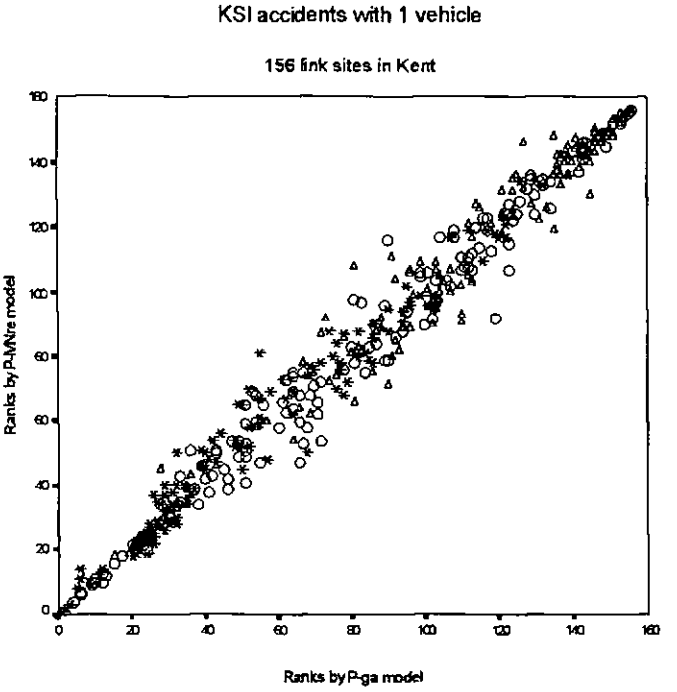


Appendix F

Comparison of ranks

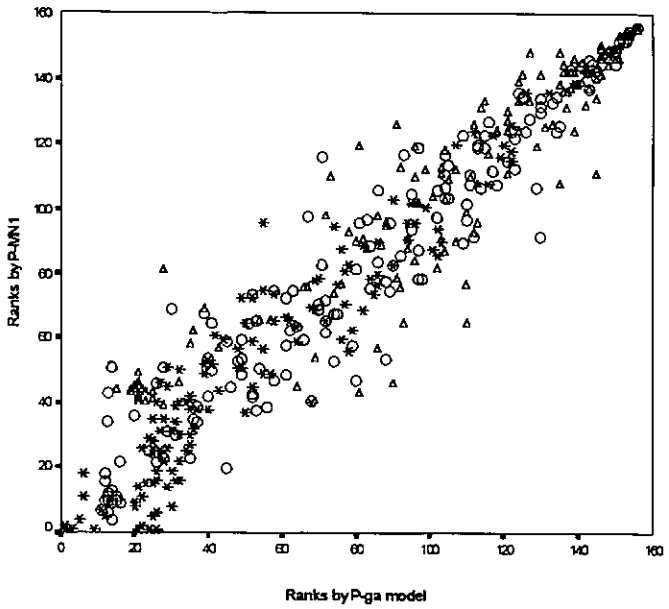
In the following scatterplots the posterior median of ranks, as given by (P-MNre) and (P-MN1), are compared to the posterior median of ranks as given by (P-ga). The model (P-MNre) gives closer matchings of ranks with the base model (P-ga) than the matchings of model (P-MN1). This is not very surprising, (P-ga) and (P-MNre) having a similar model specification and using the same covariate information. On the contrary, the model (P-MN1) is based on different "distributional" assumptions, more exactly on the multivariate Poisson-log normal distribution, and it does not use any covariate information. For fatal or serious accidents, the plots of (P-MN1) against (P-ga) are more volatile but still close in the right extreme of the plot, where is the interest of the practitioner. For slight accidents, the plots of (P-MN1) against (P-ga) are improving; this suggests that the sparsity of the data may be the cause of the difference in ranking. From the plot comparing ranks given by (P-MN1) with those given by (P-ga), for fatal or serious accidents with 2+ vehicles, it can be

seen that there are more triangles above the diagonal line for the sites in the middle of the ranks. This means that the (P-MN1) model gives larger right ends of the credible intervals of the ranks than those given by model (P-ga).



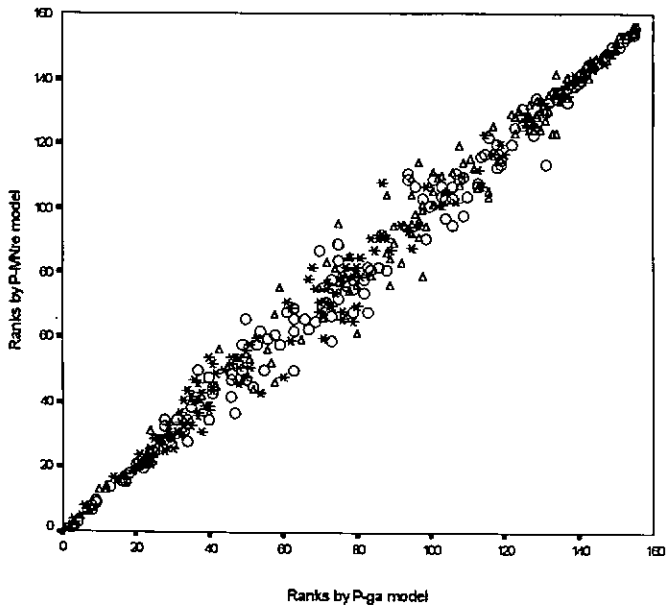
KSI accidents with 1 vehicle

156 link sites in Kent



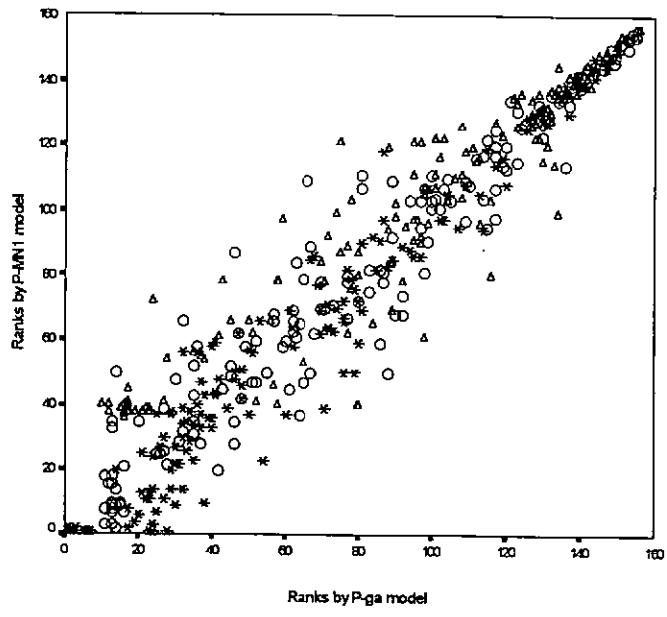
KSI accidents with 2+ vehicles

156 link sites in Kent



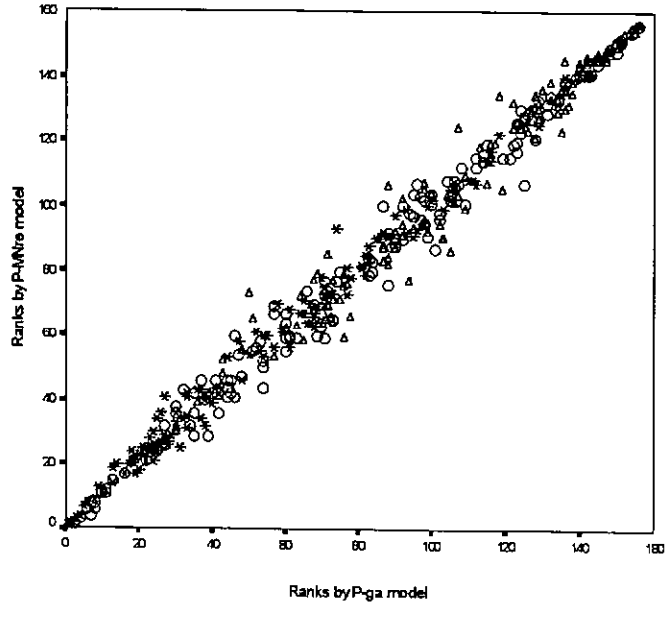
KSI accidents with 2+ vehicles

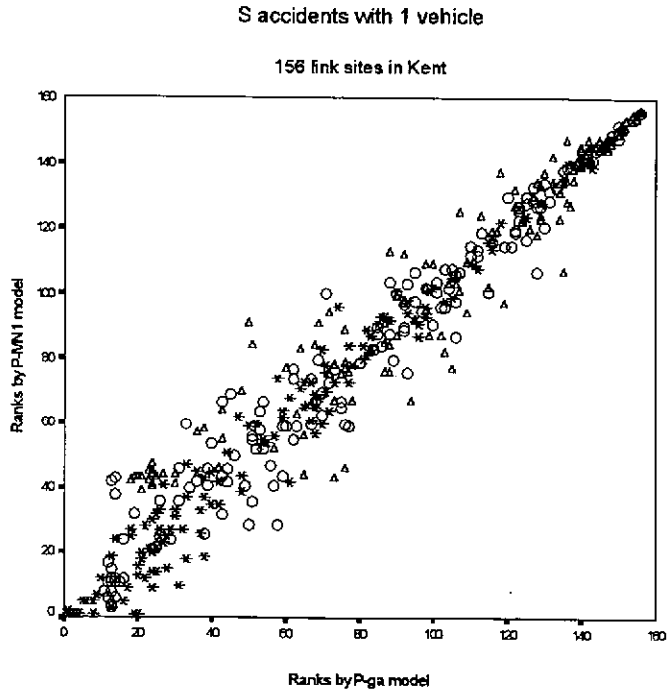
156 link sites in Kent

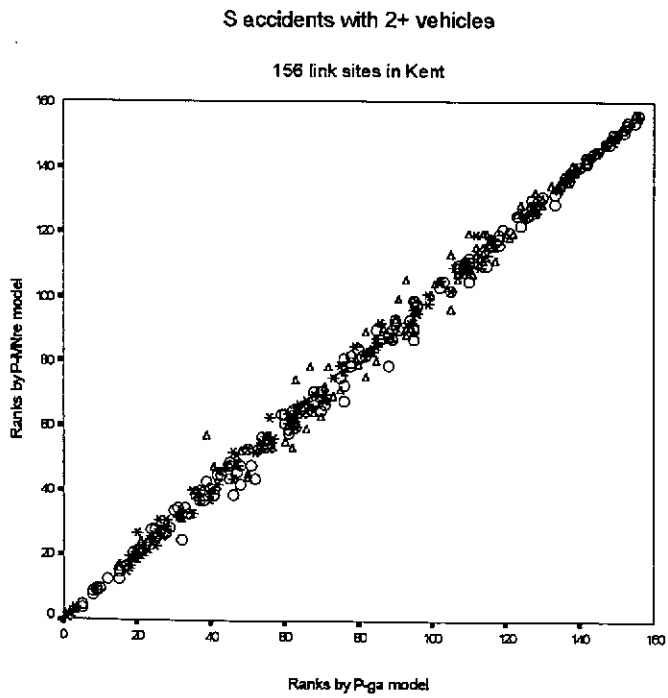


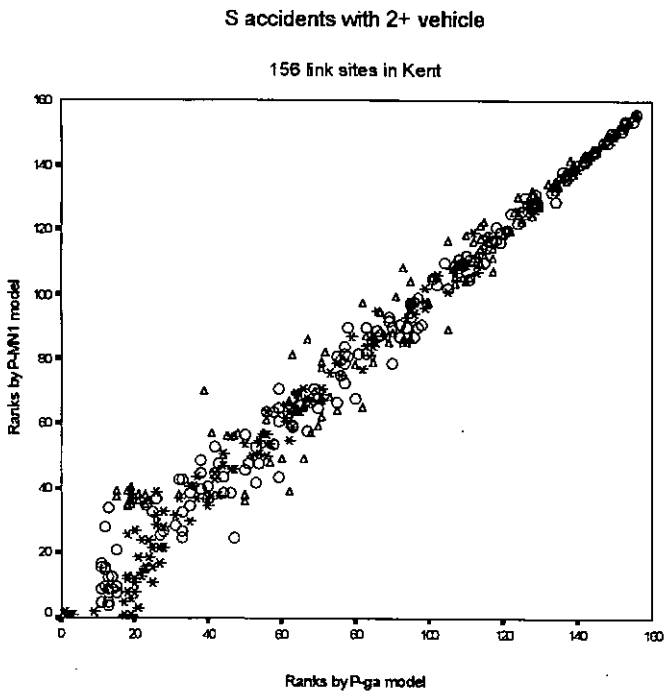
S accidents with 1 vehicle

156 link sites in Kent




S accidents with 1 vehicle
156 link sites in Kent
Ranks by P-MN1 model
Ranks by P-ga model
0
20
40
60
80
100
120
140
160


S accidents with 2+ vehicles
156 link sites in Kent
Ranks by P-MNre model
Ranks by P-ga model
0
20
40
60
80
100
120
140
160


S accidents with 2+ vehicle
156 link sites in Kent
Ranks by P-MN1 model
Ranks by P-ga model
0
20
40
60
80
100
120
140
160

Appendix G

Posterior statistics for regression coefficients

Table G.1: Estimates for mixed Poisson-gamma regression model

	1	$ST(\log l)$	$ST(\log F)$	S	SL	ST
Statistic	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
mean	0.55	1.20	0.60	-0.29	0.04	-0.38
sd	0.09	0.11	0.09	0.10	0.08	0.14
2.5 %	0.36	1.00	0.41	-0.49	-0.13	-0.66
median	0.54	1.20	0.60	-0.29	0.04	-0.37
97.5 %	0.72	1.43	0.79	-0.09	0.20	-0.10
Statistic	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
mean	0.85	1.49	0.72	-0.29	0.09	-0.01
sd	0.08	0.09	0.08	0.08	0.07	0.11
2.5 %	0.70	1.30	0.56	-0.47	-0.06	-0.25
median	0.85	1.49	0.72	-0.29	0.09	-0.01
97.5 %	1.00	1.70	0.87	-0.12	0.24	0.20
Statistic	β_{31}	β_{32}	β_{33}	β_{34}	β_{35}	β_{36}
mean	1.09	1.28	0.57	-0.16	-0.10	-0.26
sd	0.07	0.10	0.09	0.08	0.08	0.13
2.5 %	0.94	1.08	0.47	-0.34	-0.26	-0.52
median	1.09	1.28	0.57	-0.15	-0.10	-0.26
97.5 %	1.23	1.49	0.75	0.00	0.05	-0.02
Statistic	β_{41}	β_{42}	β_{43}	β_{44}	β_{45}	β_{46}
mean	2.00	1.29	0.69	-0.35	-0.00	-0.08
sd	0.06	0.09	0.07	0.07	0.06	0.10
2.5 %	1.87	1.12	0.55	-0.48	-0.12	-0.28
median	2.00	1.30	0.69	-0.35	-0.00	-0.08
97.5 %	2.13	1.46	0.85	-0.21	0.13	0.12

Bibliography

Abbess, C., Jarrett, D. and Wright, C. (1981), 'Accidents at blackspots: estimating the effectiveness of remedial treatment with special reference to the regression-to-mean effect', *Traffic Engineering and Control* **22**(10), 535–542.

Agresti, A. (1990), *Categorical Data Analysis*, Wiley, New York.

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, Wiley, New York.

Aitchison, J. and Ho, C. (1989), 'The multivariate Poisson-log normal distribution', *Biometrika* **76**(4), 643–653.

Aitkin, M. (1979), 'A simultaneous test procedure for contingency table models', *Applied Statistics* **28**, 233–242.

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *in* B. Petrov and F. Csaki, eds, 'Second Int. Symp. Information Theory', Akademiai Kiado: Budapest, pp. 267–281.

Altham, P. (1984), 'Improving the precision of estimation by fitting a model', *J. Roy. Statist. Soc.* **B**(46), 118–119.

Amis, G. (1996), 'An application of generalised linear modelling to the analysis of traffic accidents', *Traffic Eng. and Control* .

Anscombe, F. (1950), 'Sampling theory of the negative binomial and logarithmic series distributions', *Biometrika* **37**, 358–382.

Aragon, J., Eberly, D. and Eberly, S. (1992), 'Existence and uniqueness of the maximum likelihood estimator for the two-parameter negative binomial distribution', *Statistics and Probability Letters* **15**, 375–379.

Asmussen, S. and Edwards, D. (1983), 'Collapsibility and response variables in contingency tables', *Biometrika* **70**(3), 566–578.

Bartlett, M. (1935), 'Contingency table interactions', *Journal of the Royal Statistical Society* **2**, 248–52.

Baruya, A., Finch, D. and Wells, P. (1999), 'A speed-accident relationship for european single-carriageway roads', *Traffic Engineering and Control* **40**(3), 135–139.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems (with discussion)', *J. Roy. Statist. Soc. Ser. B.* **36**, 192–236.

Bishop, Y., Fienberg, S. and Holland, P. (1975), *Discrete Multivariate Analysis*, MIT Press, Cambridge, Massachusetts.

Blalock, H. (1971), *Causal models in the social sciences*, Aldine-Atheston, Chicago.

Breslow, N. and Clayton, D. (1993), 'Approximate inference in generalized linear mixed models', *Journal of American Statistical Association* **88**, 9–25.

Brooks, S. and Gelman, A. (1998), 'Alternative methods for monitoring convergence of iterative simulations', *Journal of Computational and Graphical Statistics* **7**, 434–455.

Brown, M. (1976), 'Screening effects in multi-dimensional contingency tables', *Appl. Statist.* **25**(1), 37–46.

Brude, U. and Larsson, J. (1988), 'The use of prediction models for eliminating effects due to regression-to-the-mean in road accident data', *Accident Analysis and Prevention* **20**(4), 299–310.

Cameron, A. and Trivedi, P. (1986), 'Econometric models based on count data: comparisons and and applications of some estimators and tests', *Journal of Applied Econometrics* (1), 29–53.

Carlin, B. and Louis, T. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall, London.

Carruthers, D., Bulpitt, M., Gray, G., Holmes, A., MacKinven, D., Moore, P., Quinn, D., Zealley, H. and Huxford, R. (1996), A vision for road

safety beyond 2000, Technical report, The Institution of Civil Engineers, London.

Christensen, R. (1990), *Log-Linear Models*, Springer-Verlag, New York.

Christiansen, C., Morris, C. and Pendleton, O. (1992), A hierarchical Poisson model, with beta adjustments for traffic accident analyses, Technical Report 103, Center for Statistical Sciences, University of Texas, Austin.

Cox, D. (1983), 'Some remarks on overdispersion', *Biometrika* **70**(1), 269–274.

Cox, D. (1993), 'Causality and graphical models', *Bulletin of International Statistical Institute* **Proceedings of 49th session**, 365–372.

Cox, D. and Wermuth, N. (1993), 'Linear dependencies represented by chain graphs', *Statist. Sci.* **8**, 204–218.

Darroch, J., Lauritzen, S. and Speed, T. (1980), 'Markov fields and log-linear interaction models for contingency tables', *Annals of Statistics* **8**, 522–539.

Davies, J. (1990), 'A Bayesian analysis of some accident data', *Statistician* **39**, 11–17.

Davis, L. (1986), 'Whittemore's notion of collapsibility in multidimensional contingency tables', *Communications in Statistics Theory* **15**, 2541–2554.

Dawid, A. (1980), 'Conditional independence for statistical operations', *Annals of Statistics* **8**, 598–617.

Dean, C. and Lawless, J. (1989), 'Tests for detecting overdispersion in poisson regression models', *Journal of the American Statistical Association* **84**(406), 467–472.

Dempster, A. (1974), The direct use of likelihood for significance testing, *in* P. B. O. Barndorff-Nielsen and G. Schou, eds, 'Proceedings of Conference on Foundational Questions in Statistical Inference', Department of Theoretical Statistics: University of Aarhus, pp. 335–352.

Department of Transport (1996), Road accidents Great Britain: 1995 the casualty report, Technical report, Department of Transport, London.

Doss, H. and Narasimhan, B. (1994), Bayesian Poisson regression using the Gibbs sampler: Sensitivity analysis through dynamic graphics, Technical report, Department of Statistics, The Ohio State University, Ohio.

Durrett, R. (1991), *Probability: Theory and Examples*, The Wadsworth and Brooks, California.

Edwards, D. (1990), 'Hierarchical interaction models (with discussion)', *Journal of the Royal Statistical Society, B* (52), 3–20.

Edwards, D. (1995), *Introduction to Graphical Modelling*, Springer-Verlag, New York.

Edwards, D. and Havranek, T. (1985), 'A fast procedure for model search in multi-dimensional contingency tables', *Biometrika* **72**(2), 339–351.

Fienberg, S. (1980), *The Analysis of Cross-Classified Categorical Data*, MIT Press, Cambridge, Massachusetts.

Frydenberg, M. (1990), 'The chain graph Markov property', *Scandinavian Journal of Statistics* **17**, 333–353.

Gabriel, K. (1969), 'Simultaneous test procedures: some theory of multiple comparisons', *Ann. Math. Statist.* **40**(1), 224–250.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995), *Bayesian Data Analysis*, Chapman and Hall, London.

Geman, S. and Geman, S. (1984), 'Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

George, E., Makov, U. and Smith, A. (1993), 'Conjugate likelihood distributions', *Scandinavian Journal of Statistics* **20**, 147–156.

Gibbs, W. (1902), *Elementary principles of statistical mechanics*, Yale University Press, NewHaven.

Gilks, W. (1992), Derivative-free adaptive rejection sampling for Gibbs sampling, *in* J. Bernardo, J. Berger, A. Dawid and A. Smith, eds, 'Bayesian Statistics 4', Oxford University Press, pp. 641–649.

Gilks, W., Richardson, S. and Spiegelhalter, D., eds (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Goldstein, H. (1979), *The design and analysis of longitudinal studies*, Academic Press, London.

Goodman, L. (1973), 'The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach', *Biometrika* **60**, 179–192.

Haberman, S. (1974), *The Analysis of Frequency data*, University of Chicago Press, Chicago.

Hand, D., McConway, D. and Stanghellini, E. (1997), 'Graphical models of applicants for credit', *IMA Journal of Mathematics Applied in Business and Industry* (8), 143–155.

Hauer, E. (1980), 'Bias-by-selection: Overestimation of the effectiveness of safety countermeasures caused by the process of selection for treatment', *Accid. Anal. and Prev.* **12**(2), 113–118.

Hauer, E. (1986), 'On the estimation of the expected number of accidents', *Accid. Anal. and Prev.* **18**(1), 1–12.

Hauer, E. (1997), *Observational before-after studies in road safety: estimating the effect of highway and traffic engineering measures on road safety*, Elsevier Science, Oxford.

Hauer, E., Ng, J. and Lovell, J. (1989), 'Estimation of safety at signalized intersections', *Trans. Res. Record* **1185**, 48–61.

Henson, R. (1992), Pedal cycle accidents at T-junctions, *in* J. Griffiths, ed., 'Mathematics in Transport Planning and Control', Clarendon Press, pp. 355–366.

Higle, J. and Witkowski, J. (1988), 'Bayesian identification of hazards locations', *Trans. Res. Rec.* **1185**, 24–36.

Holland, P. (1986), 'Statistics and causal inference', *Journal of the American Statistical Association* **81**, 945–970.

Jarrett, D., Abbess, C. and Wright, C. (1982), Bayesian methods applied to road accident blackspot studies: some recent progress. SWOV Conference, Amsterdam.

Johnson, N. and Kotz, S. (1969), *Discrete distributions*, Houghton Mifflin Company, Boston.

Kass, R., Tierney, L. and Kadane, J. (1989), 'Approximate methods for assessing influence and sensitivity in Bayesian analysis', *Biometrika* **76**, 663–674.

Kihlberg, J., Narragon, E. and Campbell, B. (1964), Automobile crash injury in relaion to car size, Technical Report VJ-1823-R11, Cornell Aero. Lab.

Kreiner, S. (1987), 'Analysis of multi-dimensional contingency tables by exact conditional tests: techniques and strategies', *Scand. J. Statist.* **14**, 97–112.

Kulmala, R. (1994), 'Measuring the safety effect of road measures at junctions', *Accid. Anal. and Prev.* **26**(6), 781–794.

Laird, N. and Louis, T. (1989), 'Empirical Bayes ranking methods', *Journal of Educational Statistics* **14**, 29–46.

Lauritzen, S. (1982), *Lectures on Contingency Tables*, 2 edn, University of Aalborg Press, Denmark.

Lauritzen, S. (1989), 'Mixed graphical association models (with discussion)', *Scand. J. Statist.* **16**, 273–306.

Lauritzen, S. (1996), *Graphical Models*, Oxford University Press, Oxford.

Lauritzen, S., Speed, T. and Vijayan, K. (1984), 'Decomposable graphs and hypergraphs', *J. Austral. Math. Soc. A* **36**, 12–29.

Lindley, D. (1969), *Introduction to Probability and Statistics from a Bayesian viewpoint*, Cambridge University Press, London.

Loveday, J. and Jarrett, D. (1992), Spatial modelling of road accident data, *in* J. Griffiths, ed., 'Mathematics in Transport Planning and Control', Clarendon Press, pp. 433–446.

Lupton, K., Wing, M. and Wright, C. (1998), Conceptual data structures and the statistical modelling of road accidents, *in* J. Griffiths, ed., 'Third IMA International Conference on Mathematics in Transport Planning and Control', Elsevier, Oxford, pp. 267–277.

Lyngaard, H. and Walther, K. (1993), Dynamic modelling with mixed graphical association models, Technical Report VJ-1823-R11, Institute for Electronic Systems, University of Aalborg.

Madigan, D. and Mosurski, K. (1990), 'An extension of the results of Asmussen and Edwards on collapsibility in contingency tables', *Biometrika* **77**(2), 315–319.

Maher, M. J. (1991), 'A new bivariate negative binomial model for accident frequencies', *Traffic Engineering and Control* **32**(9), 422–423.

Maher, M. and Mountain, L. (1988), 'The identification of accident blackspots: A comparison of current methods', *Accident Analysis and Prevention* **20**(2), 143–151.

Maher, M. and Summersgill, I. (1996), 'A comprehensive methodology for the fitting of predictive accident models', *Accid. Anal and Prev.* **28**(3), 281–296.

Maritz, J. and Lwin, T. (1989), *Empirical Bayes Methods*, second edn, Chapman and Hall, London.

Maycock, G. (1985), 'Accident liability and human factors: researching the relationship', *Traffic Engineering and Control* **26**(6), 330–335.

Maycock, G. and Hall, R. (1984), Accidents at 4-arm roundabouts, Technical Report 1120, TRRL Laboratory Report Crowthorne, U.K.: Transport and Road Research Laboratory.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, second edn, Chapman and Hall, London.

Miaou, S. and Lum, H. (1993), 'Modeling vehicle accidents and highway geometric design relationships', *Accid. Anal. and Prev.* **25**(6), 689–709.

Mohamed, W., Diamond, I. and Smith, P. (1998), 'The determinants of infant mortality in malaysia: a graphical chain modelling approach', *J.R. Statist. Soc. A* **161**, 349–366.

Morris, C. (1983), 'Parametric empirical Bayes inference: Theory and applications', *J. Amer. Statist. Assoc.* **25**(78), 47–65.

Morris, C. and Christiansen, C. (1996), Hierarchical models for ranking and for identifying extremes, with applications, *in* J. Bernardo, J. Berger, A. Dawid and A. Smith, eds, 'Bayesian Statistics 5', Oxford University Press, pp. 277–296.

Morris, C., Christiansen, C. and Pendleton, O. (1991), Application of new accident analysis methodologies, Technical report, U.S. Department of Transportation.

Mountain, L., Fawaz, B. and Jarrett, D. (1996), 'Accident prediction models for roads with minor junctions', *Accid. Anal. and Prev.* **28**(6), 695–707.

Mountain, L., Jarrett, D. and Fawaz, B. (1995), The safety effects of highway engineering schemes, *in* 'Proc. Instn Civ. Engrs Transp.', Vol. 111, pp. 298–309.

Mountain, L., Jarrett, D. and Wright, C. (1994), Road accident migration. EPSRC project, GR/G53415.

Nicholson, A. (1985), 'The variability of accident counts', *Accident Analysis and Prevention* **17**, 47–56.

Patefield, W. (1981), 'An efficient method of generating random R x C tables with given row and column totals', *Appl. Statist.* **30**, 91–97.

Persaud, B. (1991), 'Estimating accident potential of ontario road sections', *Transportation Research Record* (1327), 47–53.

Read, T. and Cressie, N. (1988), *Goodness-of-fit Statistics for Discrete Multivariate Data*, Springer-Verlag, New York.

Robbins, H. (1955), An empirical Bayes approach to statistics, *in* 'Proceedings of 3rd Berkeley Symp. on Math. Statist. and Prob.', Vol. 1, Univ. of California Press, pp. 157–164.

Roberts, G. and Smith, A. (1993), 'Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms', *Stochastic Processes and their Applications* **49**, 207–216.

Roh, J., Bessler, D. and Gilbert, R. (1999), 'Traffic fatalities, Peltzman's model, and directed graphs', *Accident Analysis and Prevention* **31**(1/2), 55–62.

Ross, G. and Preece, D. (1985), 'The negative binomial distribution', *The Statistician* **34**, 323–336.

Salminen, S. and Heiskanen, M. (1997), 'Correlations between traffic, occupational, sports and home accidents', *Accid. Anal. and Prev.* **29**(1), 33–36.

Santner, T. J. and Duffy, D. (1989), *The Statistical Analysis of Discrete Data*, Springer-Verlag, New York.

Schluter, P., Deely, J. and Nicholson, A. (1997), 'Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model', *The Statistician* **46**(3), 293–316.

Shaban, S. (1988), Poisson-lognormal distributions, *in* E. Crow and K. Shimizu, eds, 'Lognormal distributions: theory and applications', Marcel Dekker, New York, pp. 195–210.

Simpson, C. (1951), 'The interpretation of interaction in contingency tables', *J. Roy. Statist. Soc. (B)* **13**, 238–241.

Smith, T., Spiegelhalter, D. and Thomas, A. (1995), 'Bayesian graphical modelling applied to random effects meta-analysis', *Statistics in Medicine* **14**, 2685–2699.

Spiegelhalter, D., Best, N. and Carlin, B. (1998), 'Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models'. unpublished paper.

Spiegelhalter, D., Thomas, A. and Best, N. (1996), Computation on bayesian graphical models, *in* J. Bernardo, J. Berger, A. Dawid and A. Smith, eds, 'Bayesian statistics 5', Oxford University Press, Oxford, pp. 407–425.

Spiegelhalter, D., Thomas, A. and Best, N. (1998), *WinBUGS: User Manual, version 1.1.1*, MRC Biostatistics Unit, Cambridge.

Stanghellini, E. (1997), 'Identification of a single-factor model using graphical gaussian rules', *Biometrika* **84**(I), 241–244.

Stanghellini, E., McConway, D. and Hand, D. (1999), 'A discrete variable chain graph for applicants for credit', *Appl. Statist.* **48**, 239–251.

Tarjan, R. and Yannakis, M. (1984), 'Simple linear time algorithms to test chordality of graphs, test acyclicity of hypegraphs, and selectively reduce acyclic hypergraphs', *SIAM Journal on Computing* **13**, 566–579.

Taylor, C. and Barker, K. (1994-1995), Injury accidents on rural single-carriageway roads - an analysis of stats19 data, Technical Report 304, TRL.

Tunaru, R. (1999), 'Hierarchical bayesian models for road accident data', *Traffic Engineering and Control* **40**(6), 318–324.

Tunaru, R. and Jarrett, D. (1998*a*), An analysis of causality for road accident data using graphical models, *in* J. Griffiths, ed., 'Third IMA International Conference on Mathematics in Transport Planning and Control', Elsevier, Oxford, pp. 279–290.

Tunaru, R. and Jarrett, D. (1998*b*), Graphical models for road accident data. Universities Transport Study Group 30th Annual Conference, Dublin.

Wagner, C. (1982), 'Simpson's paradox in real life', *The American Statistician* (36), 46–48.

Wang, Y. (1996), 'Estimation problems for the two-parameter negative binomial distribution', *Statistics and Probability Letters* **26**, 113–114.

Wedderburn, R. (1974), 'Quasi-likelihood functions, generalized linear models and the Gauss-Newton method', *Biometrika* **61**, 439–447.

Wermuth, N. (1976), 'Analogies between multiplicative models in contingency tables and covariance selection', *Biometrics* **32**, 95–108.

Wermuth, N. and Lauritzen, S. (1990), 'On substantive research hypotheses, conditional independence graphs and graphical chain models', *J. Roy. Statist. Soc. B* **52**(1), 21–50.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester.

Whittemore, A. (1978), 'Collapsibility of multidimensional contingency tables', *Journal of Royal Statistical Society* **B 40**(3), 328–340.

Willson, L. J., Folks, J. and Young, J. (1984), 'Multistage estimation compared with fixed-sample-size estimation of the negative binomial parameter k', *Biometrics* **40**, 109–117.

Willson, L. J., Folks, J. and Young, J. (1986), 'Complete sufficiency and maximum likelihood estimation for the two-parameter negative binomial distribution', *Metrika* **33**, 349–362.

Wold, H. (1954), 'Causality and econometrics', *Econometrica* **22**, 162–177.

Wold, H. (1960), 'A generalisation of causal chain models', *Econometrica* **28**, 443–463.

Wright, C., Abbess, C. and Jarrett, D. (1988), 'Estimating the regression-to-mean effect associated with road accident black spot treatment: Towards a more realistic approach', *Accid. Anal. and Prev.* **20**(3), I99–214.

Wright, S. (1934), 'The method of path coefficients', *Annals of Mathematical Statistics* **5**, 161–215.

Zeger, S. and Karim, M. (1991), 'Generalized linear models with random effects; a Gibbs sampling approach', *Journal of the American Statistical Association* **86**(413), 79–86.