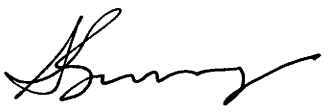
THE REGRESSION ANALYSIS OF GROUP TRUNCATED DATA

A THESIS
SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
OF THE
AUSTRALIAN NATIONAL UNIVERSITY

By
Simon Christopher Barry
September 1995

Statement of Originality

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of a university or other institute of higher learning, except where due acknowledgement is made in the text.



Simon C Barry

Acknowledgements

I would first like to thank my supervisor, Terry O'Neill for his guidance, advice and encouragement during the course of this study. On the occasions that I felt like quitting Terry has always put things into perspective and given me sound advice.

Special thanks to my advisor Alan Welsh for stimulating discussions and occasional psychiatric help. Alan has provided me with help and encouragement over many years. In addition I would like to thank my other advisor Chris Skeels for useful discussions.

Thanks also to Steven Stern for many useful discussions on a range of issues. Thanks to Jenny Hunt for administrative help.

Thanks also to the regulars at the Statistics Department tea room who provided me with useful sanity checks. In this regard I would also like to thank Geoffrey Browne and my other friends at the Canberra North Bowling Club.

Special thanks to David Rowell for providing me with encouragement over the years.

Thanks also to the staff of the Department of Mathematics and Statistics at the Australian Defence Force Academy where this thesis was completed.

Last but not least I would like thank my wife Angela, who has made many sacrifices to make this thesis possible.

This thesis was completed while on an Australian Postgraduate Research Award. This financial support is appreciated.

Abstract

This thesis considers the regression modelling of grouped binary data that is subject to truncation, and explores some general issues relating to truncation.

The likelihood for simple binary and ordinal models is developed and the statistical behaviour of these models is explored. The models are found to be well behaved. The efficiency of the truncated model is compared with that of conditional logistic regression, a competing technique. It is found that the truncated model is always more efficient but requires additional assumptions about the data generation process to be applicable.

The estimation of the full sample size, N, before truncation occurs is considered, in quite general regression models. The case where the covariate distribution is discrete is first considered. This is extended to allow continuous covariates, and the additional difficulties involved are explored. The issue of setting confidence intervals for N is discussed. A simulation study is used to explore the methods behaviour.

Next, the Bayesian analysis of truncated regression models is considered. The use of the empirical distribution of the observed covariates to facilitate the analysis is explored. The posterior distribution of the models parameters under this approach is derived and a Gibbs sampling algorithm implemented to explore the posterior. The convergence properties of the algorithm is considered, and the techniques behaviour assessed in a small simulation study.

The effect of over-dispersion on the analysis of group truncated binary data is considered. The available methods of introducing over-dispersion in clustered binary data are discussed and it is argued that only random effects models provide a viable approach. Parameter estimation in these models is derived via a marginal likelihood. In addition a score test is constructed to test for the presence of random effects in group truncated binary data. The methods performance is demonstrated using a simulation study.

Finally, the use of the bootstrap to estimate the sampling distribution of parameter estimates from truncated data is considered in an appendix. The inherent limitations of using resampling methodologies to investigate truncated data is demonstrated. It is shown that the nonparametric advantages of the bootstrap are not realised with truncated data due to the lack of observations on the truncated class.

Contents

List of Figures

List of Tables

Papers from this thesis

From the material in this thesis there are, at the time of submission, three papers which have been submitted to refereed journals.

Parts of Chapter 2 will appear in O'Neill & Barry (1993a)

Parts of Chapter 3 have appear in O'Neill & Barry (1993b)

Parts of Chapter 5 are included in the manuscript Barry & O'Neill *The Bayesian analysis of truncated regression models* which is currently under review.

It is anticipated that the software from this project will be contributed to the Splus directory at *StatLib*.

Chapter 1

Introduction

1.1 Introduction

This thesis was motivated by a need to analyse the 1988 and 1990 Fatal Files. These files are compiled on a biennial basis by the Federal Office of Road Safety (FORS) in the Australian Department of Transport and Communications. The files consist of records of all road traffic accidents involving fatalities in Australia in the calendar year. The aim of the proposed analysis was to examine the effects of covariates, such as seating position, seat belt usage and vehicle speed on an individual's probability of death in a road traffic accident. A common problem in road safety research is that the data sources are often unreliable. For this reason the Fatal Files are attractive since hopefully all accidents involving fatalities are investigated by the police and included in the file, leading to greater fidelity in the data set.

Similar data is collected on an annual basis in the U.S.A.. The Fatal Accident Reporting System (FARS) is a data file compiled by the National Highway Traffic Safety Administration, an agency of the U.S.A. Department of Transportation. The database began on 1 January 1975 and includes data on all fatal crashes. Fatal crashes are defined as those in which anyone dies within 30 days of the crash as a result of the crash.

In both databases, information is collected on all individuals involved in the crash, and not only those who died. The data is hierarchical since there is information about the overall crash, about the individual vehicles involved in the crash and about the individuals in each of the vehicles.

There is considerable interest among Australian road traffic researchers in whether various factors such as seat belt usage, gender and age for example, have similar effects in Australia compared to the U.S.A.. In continuing to base road safety discussions on estimated effects from the U.S.A., it may be that some inherent differences between the populations are being ignored.

Evans (1985, 1991) has investigated the effect of many factors such as age, sex, and seat belt usage on the risk of dying in a motor vehicle accident using his single and double pair comparison methods. These techniques require very large sample sizes but they are effective on the U.S.A. data because of the vastness of the FARS database. By contrast the Australian database was derived from two years and one tenth of the population base and so is two orders of magnitude smaller. Consequently there is a much greater requirement when analysing the Australian data to use efficient estimation techniques which do not require huge sample sizes.

The basic feature of the FORS and FARS databases that prevents the use of standard techniques such as logistic regression is that the data is group truncated. The individuals fall naturally into groups and the data from a group is collected only if one of the individuals dies. With road safety data the group is determined by the accident. An alternative example of group truncated data is the following. Consider collecting unemployment information on households only if at least one household member was registered unemployed. In this case the truncation group would be the household and the binary response would be employed or unemployed.

The aim of this thesis is to examine ways of analysing group truncated data. We will be paying particular regard to the analysis of data where the response is ordinal. Although other response types can be incorporated into the group truncated model, they do not arise as naturally as the ordinal response does.

We first review methods for modelling truncated data. Truncated data arises when the range of possible responses is restricted in some way. From this definition there are several uses of the term truncation in the statistical literature. The most common, and the meaning used in this thesis, is for data that is assumed to be generated as follows:

1. A sample $Y_1, \ldots, Y_N$ of size N is generated from a distribution F.

2. Only responses within a certain set are observed producing a sample of n observations.

We will refer to a sample generated in this way as observationally truncated data, as the truncation effect is explicitly defined in terms of some observational process. An alternative to observationally truncated data is distributionally truncated data. In this case there need be no observational interpretation. Distributionally truncated data is generated by

1. A sample $Y_1, \ldots, Y_n$ of size n is generated from a distribution F^*.

where F^* is a truncated form of some density. An example of this is continuous non negative data, such as peoples incomes, where the sample is from a population of individuals with positive incomes. In this case the observational arguments do not obviously hold, and a truncated standard distribution may be used as a device to better approximate the underlying populations distribution. With distributionally truncated data all inference is performed with respect to the truncated distribution. For example, with the income data discussed above inferences regarding the mean and variance are all performed with respect to the truncated distribution. Attempting to make inference regarding the observational process, such as the number and distribution of individuals with negative income, is not well defined and sensible interpretation of the results is not possible.

With observationally truncated data inference regarding the underlying observational process may be of primary importance. In fact, the moment parameters of the truncated distribution may be extremely difficult to interpret.

To demonstrate these differences we will consider a simple example. Assume we have a sample $y_1, \ldots, y_n$ of independent observations. Assume that the natural model for this data is some density $f(y; \mu)$ where μ is some parameter, and the density is defined over some set $\mathcal{A}$. Now assume that the observations are restricted in some manner to the set $\mathcal{B} \subset \mathcal{A}$. If the data is assumed distributionally truncated

the likelihood is simply

$$L(\mu; y) = \prod_i^n \frac{f(y_i; \mu)}{P(\mu)} \tag{1.1}$$

where

$$P(\mu) = \int_B f(x)dx.$$

Alternatively, if the data is observationally truncated the complete data likelihood (Little & Rubin, 1987) is

$$\binom{N}{n} P(\mu)^n [1 - P(\mu)]^{N-n} \prod_i^n \frac{f(y_i; \mu)}{P(\mu)}.$$

The value of N is of course unknown (otherwise the data would be type I censored, and different methods would apply) so we analyse the data conditional on n giving the conditional likelihood

$$\prod_i^n \frac{f(y_i; \mu)}{P(\mu)}. \tag{1.2}$$

Note that the resulting likelihoods 1.1 and 1.2 are identical, so the same maximum likelihood estimation techniques can be used. But it would be a mistake to conclude that the two models are interchangeable, as they both have different interpretations. For example, the parameter μ of $f()$ has an interpretation with respect to the underlying distribution if the data is observationally truncated, but its interpretation in the distributionally truncated case is only through its impact on the mean of the density $f(y; \mu)/P(\mu)$.

The occurrence of truncated data was recognised early in the history of statistics, and its analysis discussed since the 1950's. In a long series of papers summarised in Cohen (1991), Cohen and others developed likelihood and moment based estimators to analyse truncated data from the core univariate and multivariate distributions. Numerical techniques for the likelihood analysis of these models has also been of interest. For example Hartely (1958) demonstrated a special case of the EM algorithm (Dempster, Laird, & Rubin, 1978) to estimate parameters from truncated data.

The regression modelling of truncated data has been considered for the case where the response distribution is continuous and assumed to be Gaussian, at least approximately. For this case a large literature has developed as the model has a long history in applied econometrics. A review can be found in Amemiya (1984). In addition to the fully parametric likelihood approaches there have been several nonparametric approaches to the truncated regression model. These models have specified the mean function parametrically, but have relaxed the Gaussian assumption. Examples of these approaches can be found in Powell (1986), who assumed only that the error distribution was symmetric, Bhattacharya, Chernoff, & Yang (1983), who considered the simple linear regression model, and Tsui, Jewell, & Wu (1988) who used a nonparametric approach to estimate the error distribution.

With count data there has been recent interest in the estimation of truncated versions of the standard regression models. Shaw (1988) developed maximum likelihood estimates for the truncated Poisson model and also considered estimation in the case where the sample is endogenously stratified, due to the probability of selection increasing with the magnitude of the response. Grogger & Carson (1991) considered the effect of over-dispersion on the truncated model and derived

likelihood estimates assuming that conditional on the covariates the data follow a negative binomial distribution. Gurmu & Trivedi (1992) considered the construction of tests for over-dispersion in the presence of covariates, which is based on a test of the negative binomial distribution against the limiting case which is the Poisson distribution.

Further afield, there has been research on the estimation of densities from truncated samples (Turnbull, 1976).

We now consider the analysis of grouped data. The analysis in this thesis relies on the grouped nature of the data. This is due to truncated binary data being degenerate. Thus we can only proceed by considering groups of binary responses. Implicit in this is the assumption that we can model the probability that the group is observed, which explicitly defines the probability of the truncated class. An obvious approach is to assume that the untruncated responses are independent and proceed from there. This is the approach taken in the bulk of this thesis. Alternatively, we can note that a distinctive feature of much grouped data encountered in practice is the presence of some pattern of dependence in the response. We will thus investigate ways of introducing dependence into the group truncated model. This topic is delayed until Chapter 6.

There are a number of methods applicable to group truncated binary data that have been reported in the literature. The methods of Evans (1985) and Greenland (1994) are based on a multiplicative relative risk model, either implicitly or explicitly. Because of the range restrictions that this assumption imposes on the parameters we prefer to concentrate on methods that model the odds ratio as a linear function. Conditional Logistic Regression (CLR) meets this criteria and is applicable to group truncated binary data. CLR is one of the most under utilised statistical techniques in the road traffic literature. Early readable references on the application of the technique to matched pair designs are Breslow, Day, Halvorsen, Prentice, & Sabai (1978), Breslow & Day (1980) and Holford, White, & Kelsey (1978). The application of CLR to group truncated data can be found in Lui, McGee, Rhodes, & Pollack (1988) who looked at the effects of variables such as seat belt usage and principal point of impact on the probability of driver death. The data from FARS used were two car collisions which involved a driver death and the two drivers were compared.

This thesis will not dwell heavily on the comparison of these techniques. Of the available approaches only Evans (1985) allows the estimation of group level effects. Unfortunately his matched pair method is extremely inefficient, needing vast data sets, and is not model based. It will therefore not be pursued here. The other methods finesse the group level effects by either conditioning (CLR) or by fortuitous construction (Greenland, 1994). This is a virtue if the group level effects are nuisance parameters, as the specification of plausible models for the group level effects adds another layer of complexity that could be done without in applied work. Unfortunately, there are times when these group level effects may be of interest, both in their own right, as well as in comparison to the other effects in the model. It is for this reason that we consider group truncated regression, in addition to the philosophical attractions in modelling such data directly.

The capture-recapture literature has produced a number of significant papers. Capture-recapture data from closed populations with multiple captures can be considered a form of group truncated binary data, where the group sizes are fixed and equal to the number of capture occasions. In this case the groups are the

members of the population, and the 0/1 variables indicating capture at the ith occasion are the units within the group. In the analysis attention focusses on the estimation of population size, with the covariate effects regarded as nuisance parameters. Alho (1991) considered the case with a single recapture phase, and thus a group size of two. This was demonstrated in Alho, Mulry, Wurdeman, & Kim (1993). Huggins (1989) and Huggins (1991) considered the general case of k capture occasions. These papers did not view the problem from the perspective of truncation.

The presence of truncation has lead to the majority of the analyses in this thesis being based on maximum likelihood methods. The standard regularity conditions for the consistency and asymptotic normality of the Maximum Likelihood estimates are assumed throughout, such as those found in Serfling (1980) or Amemiya (1979). In the specific cases where the theory is applied, the regularity conditions are satisfied, and space will not be wasted demonstrating this. The main assumption that is needed in the regression case is for the usual regularity conditions to hold for each X and that the scaled observed information converges. This of course excludes certain sequences of covariates. The thesis does not concern itself with the determination of esoteric conditions on the covariate distribution for the standard results to hold. In all practical cases considered this was not a problem. Any departures from the usual regularity conditions are stated in the text.

Most Chapters conclude with an analysis of a set of data extracted from the 1988/90 Federal Office of Road Safety Fatal Files. This data is described in Chapter 2. These analyses are for illustration purposes only. With observationally truncated data endless arguments can be made regarding the correct model to use and the correct interpretation of any estimates that are obtained. The truncation literature itself is fairly glib on many of these issues, preferring to assume that the model is taken at face value, without getting involved in the deeper robustness problems. This is understandable. Truncated models are missing data models, and as such exhibit the unfortunate property of relying heavily on the assumptions used regarding the missing data mechanism to produce their estimates. These assumptions are often either untestable, or hard to test with any power in moderately sized data sets.

To examine the model's properties with respect to misspecification is extremely difficult, as it relies on comparing the approach with an infinite number of competing alternatives. Even if specific alternatives are considered the methodology is very ad hoc, and gives no direct insight into any applied problem. Thus any analysis must be interpreted in an exploratory, even slightly subjective Bayesian, way. The Bayesian flavour comes from viewing the untruncated distribution as prior information. The analysis can be interpreted along the following lines: 'Given that I assume the data was generated from this distribution what is it sensible to infer?'. This interpretation does not stand critical scrutiny, but the frequentist interpretations contribution is little better: 'If your model is correct the accuracy is as stated, otherwise I'm not sure'. The problem revolves around the impossibility of constructing consistent nonparametric estimates of quantities such as the mean of the untruncated distribution in many practical problems. Viewed another way, the problem arises due to the truncated analysis being fundamentally based on extrapolation to the unobserved(truncated) class. If there is no information on that class then no consistent non-parametric estimate can be found.

One useful approach is to attempt several analyses under a plausible range of

assumptions and check the variation in the inferences that would be made. Determining a plausible range in a truncated regression problem may present difficulties, but the approach seems the best objective method available. Other approaches such as splitting up the data into sub groups and comparing analyses could also be attempted.

This is not performed in any of the analyses presented here. This is due to a desire to not focus unduly on the particular. This thesis develops methodology to allow group truncated data to be explored, and discusses the behaviour of the methods. It is beyond the scope of the thesis to give a definitive analysis of Road Safety Data. Therefore the analyses presented are merely meant to demonstrate the viability and interpretation of the truncated models.

The thesis is arranged as follows. In Chapter 2 the group truncated binary regression model is introduced, and its estimation is discussed. The method is compared to CLR and efficiency issues are considered. There is a brief discussion of the issue of testing goodness of fit either numerically or graphically. Some examples are then given and the method discussed.

Chapter 3 considers the extension of the binary model to ordinal data. Again, the model is developed and estimation is performed. A number of examples are given.

Chapter 4 considers the estimation of the size of the unobserved sample. This chapter explores quite general truncated regression models and extends and clarifies the current estimates available in the literature. The case of a distinct number of covariate configurations is considered, and this is extended to allow arbitrary covariate configurations. A number of examples are given.

Chapter 5 develops a Bayesian approach to the truncated regression model. In this approach the covariate distribution is estimated for the unobserved population from the covariates observed, and this is used to generate posteriors for the regression parameters and other derived items in the sample via a Gibbs Sampling algorithm. A simulation study and example are presented.

Chapter 6 considers the relaxation of the assumption of independence between responses within each observed group in the group truncated model. Chapter 7 presents discussion and conclusions about the results of the thesis and considers avenues for further research.

Appendix A considers the use of bootstrapping/resampling techniques for the estimation of variability of estimates in the group truncated model and presents simulation results. Appendix B derives a rejection sampling algorithm for the exponentially tilted Dirichlet random variables used in the Gibbs Sampling algorithm in Chapter 5.

Chapter 2

Group Truncated Binary Regression

2.1 Introduction

In this Chapter we will introduce the group truncated binary regression model and compare it to the conditional logistic model. The technique described in this chapter, Truncated Logistic Regression (TLR), is a method for the regression analysis of binary data which is aggregated into groups and observed only if there is at least one positive response in the group. The covariates that are used to model the mean responses can be either individual specific or determined by the group.

In Section 2.2, maximum likelihood estimation of the truncated regression model is considered. In Section 2.3, the model is compared to the related technique of Conditional Logistic Regression (CLR). Section 2.4 presents some generalisations of the truncated model. Section 2.5 investigates the effects of model misspecification and Section 2.6 contains a detailed analysis of the FORS dataset and compares the results with the existing estimates of the effects of factors in the U.S.A.. Section 2.7 discusses the material presented in the chapter.

2.2 Truncated Logistic Regression

2.2.1 The Model

Consider a group of individuals who each exhibit a binary response. Assume that the group is only observed if at least one response is positive. We will assume that we observe N truncated groups and that the jth group involves n_j individuals. Further let

$$y_{ij} = \begin{cases} 1 & \text{if the } i\text{th individual in the } j\text{th group dies,} \\ 0 & \text{otherwise,} \end{cases}$$

x_{ij} = the $p \times 1$ vector of covariates associated with the ith individual in group j,

X_j = the $n_j \times p$ matrix with ith row x_{ij},

$\mathcal{R}_j$ = the set of individuals in group j,

β = the $p \times 1$ vector of regression parameters.

Note that x_{ij} can potentially include group level, as well as individual specific covariates.

If we assume that a logistic model holds for the individual probabilities of positive response, then

$$Pr(y_{ij} = 1) = \frac{exp(\beta' x_{ij})}{1 + exp(\beta' x_{ij})} = p(\beta, x_{ij}) = 1 - q(\beta, x_{ij}).$$

We will consider extensions to other link functions in the next chapter.

The Truncated Logistic Regression (TLR) approach conditions on the probability that a group is observed which is the probability that it results in at least one positive response. This has the effect of introducing a divisor to a conventional binary regression likelihood. The TLR estimator $\hat{\beta}$ is the maximiser of

$$\prod_{j=1}^{N} \frac{\prod_{i \in \mathcal{R}_j} p(\beta, x_{ij})^{y_{ij}} q(\beta, x_{ij})^{1-y_{ij}}}{P_j(\beta)} \tag{2.1}$$

where

$$\begin{aligned} P_j(\beta) = P(\beta, X_j) = 1 - Q_j(\beta) &= 1 - Q(\beta, X_j) \\ &= 1 - \prod_{i \in \mathcal{R}_j} q(\beta, x_{ij}). \end{aligned}$$

The score function for β is then

$$U(\beta) = \sum_{j=1}^{N} \left(\sum_{i \in \mathcal{R}_j} x_{ij} y_{ij} \right) - \mu_j(\beta)$$

where

$$\mu_j(\beta) = \mu(\beta, X_j) = P_j(\beta)^{-1} \sum_{i \in \mathcal{R}_j} p(\beta, x_{ij}) x_{ij}$$

and $\hat{\beta}$ is the solution of $U(\beta) = 0$. The score equations are thus found by setting the sufficient statistics for β equal to their expectation. The sample information matrix is

$$I(\beta) = \sum_{j=1}^{N} \left\{ P_j(\beta)^{-1} \sum_{i \in \mathcal{R}_j} p(\beta, x_{ij}) q(\beta, x_{ij}) x_{ij} x_{ij}' \right\} - Q_j(\beta) \mu_j(\beta) \mu_j(\beta)',$$

and the estimated variance matrix of $\hat{\beta}$ is $V(\hat{\beta}) = I^{-1}(\hat{\beta})$.

2.2.2 Model Fitting

As in Hodeshima (1988) there is no problem with the identifiability of parameters in these truncated models provided that the standard linear constraints are used when parameterising categorical covariates. Inference can be performed using the asymptotic normality of the maximum likelihood estimators as the number of accidents N goes to infinity. Model testing of nested models can be done in the usual way using differences of deviances. The fact that $I(\beta)$ is the sum of the variance's of $\sum_{i \in \mathcal{R}_j} x_j y_j$ evaluated using the truncated distribution ensures that if

a maximum exists it will be unique. This is in agreement with the results of Orme (1989) who examined the issue of uniqueness of the maximum likelihood estimates in the truncated Gaussian regression problem.

The parameter estimates are easily calculated via a Fisher scoring algorithm as

$$\hat{\beta}_{i+1} = \hat{\beta}_i - I(\hat{\beta}_i)^{-1}U(\hat{\beta}_i),$$

where $\hat{\beta}_i$ is the estimate at the ith iteration. An equivalent approach is to define the row vectors, $p_j(\beta) = [p(\beta, x_{1j}), \ldots, p(\beta, x_{n_j j})]$, $y_j = [y_{1j}, \ldots, y_{n_j j}]$, $b_j(\beta) = p_j(\beta)/P_j(\beta)$ and

$$\begin{aligned}
W_j(\beta) &= \frac{1}{P_j(\beta)}\text{diag}[p_j(\beta) * (1 - p_j(\beta))] - \frac{1}{P_j(\beta)^2}p_j' p_j \\
W &= \text{blockdiag}[W_j(\beta)] \\
b &= [b_1(\beta), \ldots, b_N(\beta)]' \\
y &= [y_1, \ldots, y_N]' \\
X &= [X_1', \ldots, X_N']' \\
t &= X\beta - W^{-1}(y - b)
\end{aligned}$$

where diag[a] denotes a square matrix of zeros with the vector a on the diagonal, blockdiag(b_i) denotes the matrix with blocks b_i on its main diagonal, and zeros elsewhere, and $*$ denotes the hadamard product of two matrices.

With these definitions, $\hat{\beta}$ can be found by an iteratively re-weighted least squares regression of t on X with weights $W(\beta)$.

The use of these iterative procedures requires initial estimates of the parameters. Though it would usually be desirable to develop computationally simple consistent estimates for use as the initial values there are no obvious candidates. Instead, the estimates from a fit of the unconditional model are used as the initial values. No problems have arisen from adopting this approach. As mentioned previously the likelihood is well behaved and convergence is rapid.

Two problems can arise in the fitting of these models to sparse data sets. The first problem that arises is when all responses at one level of a categorical covariate are the same. In this case divergent estimates occur. This behaviour also occurs in standard logistic regression (see Albert & Anderson (1984), Clogg, Rubin, Schenker, & Weidman (1991)). Asymptotically, the probability of this occuring tends to zero as the sample size increases, but in finite samples the only remedy is to either collapse categories or collect more data. The other problem that can occur is that there is only a single positive response in each group. This leads to the intercept parameter diverging to minus infinity. This occurs due to the data implying infinite truncation. This can be seen as follows.

We will assume that the intercept is parametrised as a vector of 1's. Let β_1 denotes the intercept parameter, and let β_{-1} denotes the parameter vector excluding β_1. The contribution to the score for the intercept parameter by the jth group is thus

$$U_j(\beta_1) = \sum_{i \in \mathcal{R}_j} y_{ij} - P_j(\beta)^{-1} \sum_{i \in \mathcal{R}_j} p(\beta, x_{ij})$$

By assumption, $\sum_{i\in\mathcal{R}_j} y_{ij} = 1$. For any fixed β_{-1},

$$P_j(\beta)^{-1} \sum_{i\in\mathcal{R}_j} p(\beta, x_{ij})$$

is a strictly decreasing function of β_1. The limit for fixed β_{-1}, as $\beta_1 \to -\infty$ is found using l'Hopitals rule to be

$$\lim_{\beta_1\to-\infty} P_j(\beta)^{-1} \sum_{i\in\mathcal{R}_j} p(\beta, x_{ij}) = 1$$

and so $U_j(\beta_1) < 0$ for all j and the parameter estimate is divergent.

This is essentially a lack of information problem. With no prior information regarding the parameters there is not any direct solution. If one is willing to entertain Bayesian techniques or consider penalised approaches, progress can be made but that will not be considered here. Again, the probability of this problem occuring tends to zero as the sample size increases.

2.3 The Efficiency Gain from Truncated Logistic Regression

2.3.1 Conditional Logistic Regression

As was mentioned in the introduction, conditional logistic regression has also been used in the analysis of group truncated paired data. As the two models are closely related and are competing for the analysis of group truncated data we will compare and contrast the two approaches.

Instead of conditioning on there being at least one death, the conditional logistic model conditions on the actual number of deaths. Suppose the covariates x_{ij} are arranged such that $x_{ij} = (u_j, v_{ij})$ and $\beta = (\beta_1, \beta_2)$. So u_j is the component of x_{ij} that is constant over the accident and v_{ij} is the component that varies over individuals in the accident. Further, let m_j denote the actual number of deaths in accident j and let

$$\begin{aligned} S_j &= \sum_{i\in\text{observed } m_j \text{ deaths}} v_{ij}, \\ S_{A,j} &= \sum_{i\in A} v_{ij}, \end{aligned}$$

where A is a subset of the individuals in accident j. Also let $\mathcal{R}_{m_j,j}$ denote the set of all subsets of size m_j of individuals in accident j. Then the CLR estimate of β, $\tilde{\beta}$, is the maximiser of

$$\prod_{j=1}^{N} \frac{exp(\beta_2' S_j)}{\sum_{A\in\mathcal{R}_{m_j,j}} exp(\beta_2' S_{A,j})}. \tag{2.2}$$

The terms in the product 2.2 will be one for accidents involving either zero deaths (those which are truncated) or n_j (all) deaths. So the truncation does not affect the conditional logistic likelihood. In the case where only accidents involving

two individuals are considered, only accidents involving exactly one death will contribute to the likelihood 2.2 which becomes

$$\prod_{j \in \text{accidents with one death}} \tilde{p}(\beta, d_j)^{y_{1j}} \tilde{q}(\beta, d_j)^{y_{2j}}$$

where

$$\tilde{p}(\beta, d_j) = 1 - \tilde{q}(\beta, d_j) = \frac{exp(\beta_2' d_j)}{1 + exp(\beta_2' d_j)},$$

and $d_j = v_{1j} - v_{2j}$. In this case $\tilde{\beta}_2$ can be found by solving the score equations $\tilde{U}(\beta_2) = 0$ where

$$\tilde{U}(\beta_2) = \sum_{j=1}^{N} \{y_{1j} - \tilde{p}(\beta, d_j)\} d_j \tag{2.3}$$

and the estimated covariance matrix of $\tilde{\beta}_2$ is

$$\tilde{V}(\tilde{\beta}_2) = \tilde{I}(\tilde{\beta}_2)^{-1}$$

where

$$\tilde{I}(\beta_2) = \sum_{j=1}^{N} \tilde{p}(\beta_2, d_j)\tilde{q}(\beta_2, d_j) d_j d_j'. \tag{2.4}$$

Equations 2.3 and 2.4 have obvious generalisations to accidents involving more than two individuals.

The key feature of the conditional likelihood 2.2 is that covariates that are common to individuals within an accident do not appear in the conditional probabilities. For single vehicle accidents this means that variables like vehicle speed which are constant across individuals in a vehicle cannot appear in a conditional likelihood. By contrast, the truncated logistic regression likelihood 2.1 will include all accident level covariates.

The availability of the group level variables will often determine the choice between TLR and CLR. If these covariates are not available then CLR must be used. If the accident level covariates are available then either method can be used. If there is a finite number of parameters which determine the accident level response and the accident level effects are of interest then TLR should be used. The CLR method can also be computationally more complex for accidents involving larger numbers of individuals and deaths (Thompson, 1991). Another factor in the choice between the methods is the efficiencies of the two methods. There are several levels of efficiency loss in truncated samples:

- The fact that truncated groups are not observed at all causes both the truncated and conditional methods to be less efficient than estimation based on the untruncated data set.

- Let m_j be the number of deaths in a group. Then CLR conditions on m_j while TLR uses the additional information in $Pr(m_j | m_j \geq 1)$. So in general TLR will be more efficient than CLR, although the efficiency gain on the parameters of interest depends on the situation.

- CLR has zero efficiency for the associated parameters of covariates which do not vary within truncated accidents.

- Since $Pr(m_j|m_j \geq 1)$ includes information on the parameters of covariates which vary within accidents, TLR will also be more efficient than CLR for these parameters.

We denote by ELT the efficiency loss caused by the truncation. ELT is the difference between the efficiency of maximum likelihood estimation based on the untruncated sample and TLR. A further efficiency loss, ELC, is incurred by using CLR instead of TLR. This can be split into two components, ELC_1 and ELC_2. ELC_1 is the efficiency loss that CLR incurs by not attempting to estimate accident level effects. ELC_2 is the efficiency loss incurred for those effects which vary within the accident.

Due to the added assumptions required to validly fit truncated regression models we will briefly investigate the magnitude and basis of the efficiency gains.

2.3.2 The efficiencies of TLR and CLR

Consider an accident $\mathcal{R}$, where $\mathcal{R}$ denotes the set of n individuals in the accident, and for clarity of exposition suppress the subscript j. The subscript i will as usual index the individuals in the accident. If the accident is not truncated, then the sample information is equal to the Fisher information and is

$$\mathcal{I}(\beta, X) = \sum_{\mathcal{R}} p(\beta, x_i) q(\beta, x_i) x_i x_i'$$

while the Fisher information from a (possibly unobserved) accident subject to truncation is

$$\begin{aligned} \mathcal{I}_T(\beta, X) &= P(\beta, X)\{\sum_{\mathcal{R}} p(\beta, x_i) q(\beta, x_i)/P(\beta, X) x_i x_i' \\ &\quad -Q(\beta, X)\mu(\beta, X)\mu(\beta, X)'\} \\ &= \mathcal{I}(\beta, X) - P(\beta, X)Q(\beta, X)\mu(\beta, X)\mu(\beta, X)'. \end{aligned}$$

So the information loss by truncation is

$$\mathcal{I}(\beta, X) - \mathcal{I}_T(\beta, X) = P(\beta, X)Q(\beta, X)\mu(\beta, X)\mu(\beta, X)'.$$

Next consider the information loss due to conditioning on the actual number of deaths in the accident. Suppose the accident $\mathcal{R}$ of n individuals is such that $x_i' = (u', v_i')$ where u is constant over the accident and only v_i varies. If m deaths are observed then the sample information from the conditional logistic likelihood 2.2 for estimating β_2 is $V_{m,22}(\beta_2, X)$, the variance of S given the probability distribution 2.2 over $\mathcal{R}$. Hence, since by assumption u does not vary over $\mathcal{R}$, the sample information for estimating β is

$$V_m(\beta, X) = \begin{pmatrix} 0 & 0 \\ 0 & V_{m,22}(\beta_2, X) \end{pmatrix}.$$

So letting $\mathcal{I}_C(X) = \mathcal{I}_C(\beta, X)$ denote the Fisher information matrix for β from the conditional logistic likelihood, we have

$$\mathcal{I}_C(X) = \begin{pmatrix} 0 & 0 \\ 0 & \mathcal{I}_{C,22}(X) \end{pmatrix}$$

where

$$\begin{aligned}\mathcal{I}_{C,22}(X) &= \sum_{m=1}^{n-1} P_m(\beta, X) V_{m,22}(\beta_2, X) \\ &= \sum_{m=1}^{n} P_m(\beta, X) V_{m,22}(\beta_2, X),\end{aligned}$$

where $P_m(\beta, X)$ is the probability that m deaths are observed in the accident. If M is the actual number of deaths in the accident, then $\mathcal{I}_{C,22}(X)$ is the expected information from the accident, with respect to the distribution of M.

Let $\mathcal{I}^*_m(X)$ be the Fisher information in m deaths conditional on there being at least one death. This is the information in the conditional density of $M \mid M \geq 1$,

$$\begin{aligned}P_{M^*}(m) &= Pr(M = m \mid M \geq 1) \\ &= \frac{\sum_{A \in \mathcal{R}_m} \exp(\beta' \sum_{i \in A} x_i)}{\prod_{i \in \mathcal{R}} \{1 + \exp(\beta' x_i\} - 1},\end{aligned}$$

where $\mathcal{R}_m$ denotes the set of all subsets of $\mathcal{R}$ of size m. So the expected information from the $\mathcal{I}^*_m(X)$ component is

$$\mathcal{I}_M(X) = \sum_{m=1}^{n} P_m(\beta, X) \mathcal{I}^*_m(X).$$

Hence if we let $F(X)$ describe the expected distribution of X where X can be chosen either stochastically or deterministically, then for each type of information $\mathcal{I}, \mathcal{I}_T, \mathcal{I}_C$ and $\mathcal{I}_M$ it follows that

$$\mathcal{I}_\bullet = \int \mathcal{I}_\bullet(X) dF(X)$$

and since

$$Pr(y_1, y_2, \ldots, y_n \mid M \geq 1) = Pr(y_1, y_2, \ldots, y_n \mid M) Pr(M \mid M \geq 1)$$

it follows that

$$\mathcal{I}_T(X) = \mathcal{I}_C(X) + \mathcal{I}_M(X),$$

and we have that

$$\mathcal{I}_T = \mathcal{I}_C + \mathcal{I}_M.$$

We will now consider measures of relative efficiency. With a scalar parameter the choice of measure of efficiency is straight forward. With vector parameters the choice becomes less clear. When comparing a method A, to a less efficient method B, we will use as our measure of efficiency

$$\text{Efficiency Loss} = p^{-1} \text{tr} \mathcal{I}_A^{-1} \{\mathcal{I}_A - \mathcal{I}_B\},$$

where $\mathcal{I}_A$ and $\mathcal{I}_B$ are the information in the accident using methods A and B respectively and it is assumed that $\mathcal{I}_A^{-1} \mathcal{I}_B$ is positive semi-definite. This has the obvious property that if $\mathcal{I}_A = \mathcal{I}_B$ then the efficiency loss is zero, and if $\mathcal{I}_B$ is a matrix of zeroes then the efficiency loss is one. It is also invariant under transformations of the parameters. It can be viewed as the average efficiency loss for the orthogonal

parameterisation $\beta^* = \mathcal{I}_A^{1/2}(\beta)\beta$. This parameterisation will be orthogonal for method A, but is not necessarily orthogonal for method B.

With this definition the efficiency loss by truncation is

$$\text{ELT} = p^{-1}\text{tr}\mathcal{I}^{-1}\{\mathcal{I} - \mathcal{I}_T\}.$$

As an overall measure of the efficiency of conditional compared to truncated estimation of β we propose

$$\text{ELC} = p^{-1}\text{tr}\mathcal{I}_T^{-1}\{\mathcal{I}_T - \mathcal{I}_C\}.$$

However we note that in the case where u never varies within $\mathcal{R}$, and thus the information for the corresponding parameters in the conditional likelihood is zero,

$$\text{ELC} \geq 1 - p_2/p$$

where v_j has length p_2. The efficiency of conditional to truncated estimation of β_2 is

$$\text{ELC}_2 = p_2^{-1}\text{tr}\mathcal{I}_{T,22}^{-1}\{\mathcal{I}_{T,22} - \mathcal{I}_{C,22}\},$$

that is the efficiency is only compared over the parameters that can be estimated by conditional logistic regression.

While it is not possible to present detailed conclusions regarding the relative efficiencies of the two procedures, we have evaluated a limited set of examples and some impressions can be given. The efficiency loss caused by truncation, ELT, can be substantial but viable estimation is still possible. TLR is generally more efficient than CLR and in certain situations both ELC and ELC_2 can be substantial. In other words there can be substantial gains in precision even in estimating the individual level parameters. To illustrate these points we consider the following simple example.

Consider the case where the covariates for the group are

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix} = \begin{pmatrix} x'_{+1} \\ x'_{+2} \end{pmatrix}$$

where x_1 and x_2 are chosen independently from a Uniform(0,1) density.

If $p_i = p(\beta, x_{+i})$ and $p_{2i} = p_2(\beta_2, x_i)$ where

$$p(\beta, x) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}$$

and

$$p_2(\beta, x) = \frac{\exp(x'\beta)}{\exp(x'_1\beta) + \exp(x'_2\beta)}$$

then the expected information for a group with these covariates subject to truncation is

$$\mathcal{I}_T = \sum_i p_i q_i x_{+i} x'_{+i} - \frac{q_1 q_2}{1 - q_1 q_2}(p_1 x_{+1} + p_2 x_{+2})(p_1 x_{+1} + p_2 x_{+2})'$$

while the information conditional on exactly one death is

$$\sum_i p_{2i} x_i^2 - \left(\sum_i p_{2i} x_i\right)^2.$$

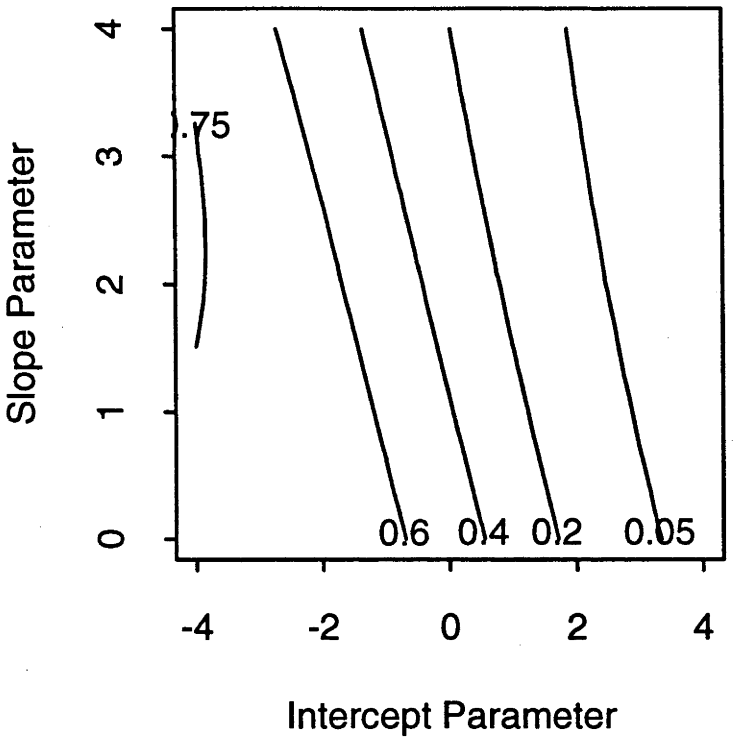


Figure 2.1: Contour of ELT for uniform covariate example.

Also,

$$P(\beta, X) = 1 - q_1 q_2$$

and

$$P_1(\beta, X) = p_1 q_2 + q_1 p_2$$

and the information about the slope coefficient in a conditional logistic likelihood term is thus

$$\mathcal{I}_{C,22}(X) = \frac{p_1 q_1 p_2 q_2 (x_1 - x_2)^2}{p_1 q_2 + q_1 p_2}$$

The truncated information $\mathcal{I}_T$ and the conditional logistic information $\mathcal{I}_{C,22}$ cannot be evaluated explicitly and were calculated via numerical integration over a range of β values. In Figures 2.1, 2.2 and 2.3 we contour ELT and ELC_2 and ELC for a range of values of the intercept and slope parameters.

The efficiency results suggest that if the accident level variables are available and a finite number of parameters determine the accident level response, then TLR should be used. If the accident level variables are unavailable or inaccurate or each accident requires a unique parameter to describe it, then CLR should be chosen. Some efficiency loss will result and any accident level effects will not be estimable.

2.4 Extensions of the truncated model

In this section we will briefly consider possible extensions of the truncated model. No formal details of the estimation methods will be given as they are straightforward modifications of the methods already presented.

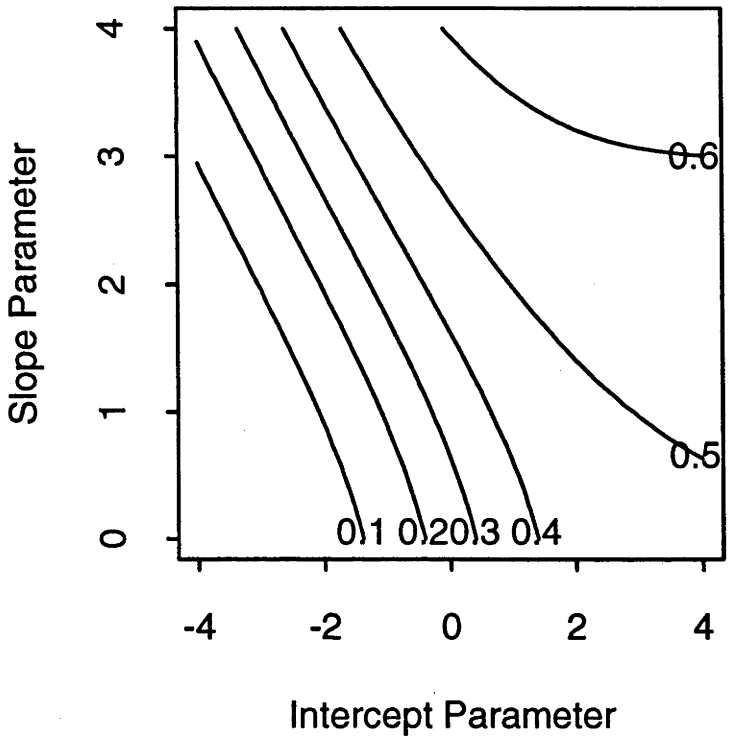


Figure 2.2: Contour of ELC_2 for uniform covariate example.

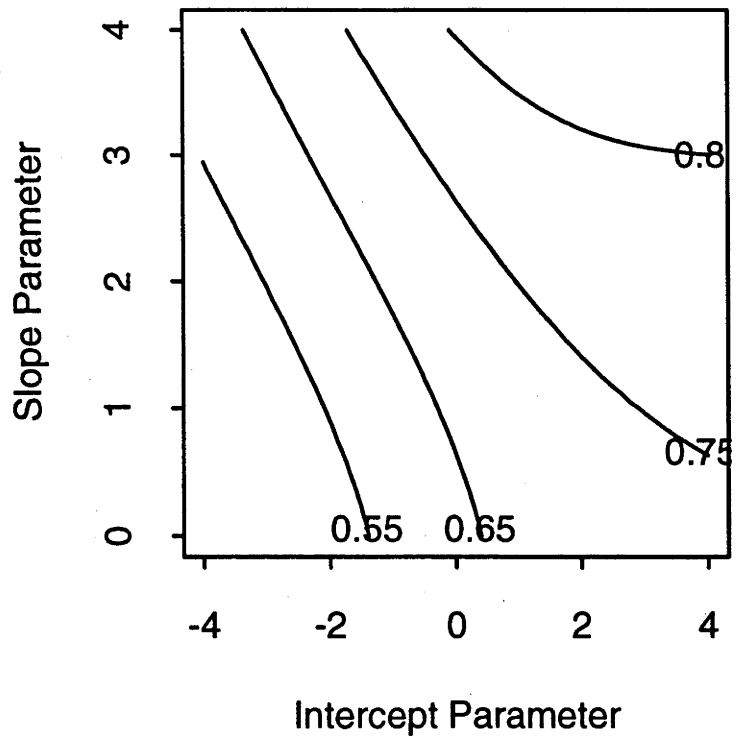


Figure 2.3: Contour of ELC for uniform covariate example

The first extension considered concerns endogenous stratification. Consider a sampling frame consisting of unemployed persons. If we select a sample of these individuals and interviewed the members of the household they lived in we would obtain a sample of group truncated data. If we applied the methods of this chapter to the data, biases could result due to households with multiple unemployed members having a higher probability of being observed. This case has been considered in the univariate case by Shaw (1988) who derived results for truncated Poisson and Gaussian models. We will consider the extension of these ideas to the group truncated setting. Consider a fixed covariate configuration X. We will assume that the frame contains observations from N such groups. We also assume that the probability of sampling two individuals from the same group is negligible, i.e. the frame is large compared to the sample size. For convenience we will use the same notation as Section 2.2, but remembering that the covariates are now fixed across the groups. The number of individuals on the frame from the jth group is

$$Y_j = \sum_i y_{ij}.$$

If we draw a sample of size 1, $y = [y_1, \ldots y_{n_j}]$ from the set of N groups represented on the frame and letting $Y = \sum_i y_i$ it can be shown that, conditional on a fixed number, N, of groups being 'observed', and hence available on the frame, that the distribution of the sampled response, y, is

$$N \frac{\prod_{i \in \mathcal{R}_j} p(\beta, X)^{y_i} q(\beta, X)^{1-y_i}}{P_j(\beta)} E\left[\frac{Y}{Y + \sum_{i=2}^{N} Y_i}\right], \tag{2.5}$$

where the expectation is performed over the truncated distribution, and arises from calculating the marginal distribution over the unselected groups on the frame. We note that as N tends to infinity this converges to

$$\frac{\prod_{j \in \mathcal{R}_i} p(\beta, X)^{y_i} q(\beta, X)^{1-y_i}}{P_i(\beta)} \frac{Y}{\sum_j^{n_i} j Pr(Y = j)} \tag{2.6}$$

where $Pr(Y = j)$ is the probability that there are Y positive responses in a group, calculated with respect to the truncated distribution of Y. The resulting score equations are complicated by the dependence of $Pr(Y = j)$ on the parameters, but estimation is straightforward.

The second modification considered relates to a processing application. Consider a manufacturing process. Each batch consists of N components, each of which exhibits a binary response, being either faulty or not faulty. The components are quickly examined and only if the number of faults is above a certain number k is the batch fully examined and covariates measured. If $k = 0$ the truncated model applies and estimation is as before. If $k > 0$ the likelihood of a response y_j, is

$$\frac{\prod_{i \in \mathcal{R}_j} p(\beta, x_{ij})^{y_i} q(\beta, x_{ij})^{1-y_i}}{Pr(Y > k)}.$$

Again this likelihood is complicated by the term $Pr(Y > k)$, but estimation proceeds normally.

2.5 The effects of model misspecification on the truncated logistic model

2.5.1 Introduction

In this section we consider the problems that occur when the model specified is incorrect. The basic theory regarding the asymptotic distribution of the maximum likelihood estimates when an incorrect probability model is specified was considered by Huber (1967). More recent references considering problems associated with misspecification are White (1982) who presents an alternative development of Huber's results, and applies them constructing tests of misspecification, and Royall (1985) who considers the construction of robust confidence intervals. The literature relating to robustness is large and is beyond the scope of this thesis to review. The result of interest here is that the maximum likelihood estimates produced assuming an incorrect model can still consistently estimate the parameters of the true model. A simple example is the Gaussian model. If we assume that y_i are a sample from a Gaussian distribution with mean μ, it is easy to show that under mild conditions the MLE's obtained provide consistent estimates of the mean and variance, even if the Gaussian model does not hold. All that is required is that the mean and variance of the true model are finite. This can be seen by noting that the score equation for μ is

$$\sum_i (y_i - \mu) = 0$$

Unfortunately this consistency does not extend to truncated data. This lack of consistency is well known in the truncation literature. To see this consider the following. In the exponential family the score equations for the mean with truncated data, y_i take the form

$$\sum (y_i - \mu_T) = 0$$

where μ_T is the expected value of y_i given the model and that the data is truncated. Hence the MLE obtained gives a consistent estimate of the mean of the truncated observations, but if the model is incorrect there is no guarantee that the estimate of the mean of the untruncated observations is consistent. This occurs due to the maximum likelihood equations being solved by parameter estimates that imply the observed truncated mean. This mapping from the untruncated mean to the truncated mean depends critically on the form of the density.

If we are only interested in inference regarding the truncated mean then no problems occur, and the methods of Royall (1985) can be used to set approximate robust large sample confidence intervals. These intervals are robust against model misspecification. Alternatively, if interest lies in the untruncated mean then little can be done. This is due to the probability model being used to essentially extrapolate back to the untruncated case. The correct model is essential to guarantee consistency. Of course, if the assumed model is close to true model, then the bias should be small. The lack of consistency arising here is similar to the example stated by White (1982), where ML estimates of the population variance of a sample from a Gaussian distribution are shown to be inconsistent if the incorrect mean is assumed.

We now consider the effect of under fitting, that is the effect of not fitting relevant covariates in the regression model. This can of course be viewed in the

context of misspecification of the probability model and the above discussion applies. The effect of under fitting is of course dependent on the particular situation being examined. This section seeks only to acknowledge the existence of potential problems and to examine their effects in simple cases. This will at least give an indication of possible problems in higher dimensional cases.

Under fitting causes various problems in the simple linear regression model. Consider the model

$$E[Y \mid X = x] = \beta_0 + x\beta_1 \tag{2.7}$$

$$Var[Y \mid X = x] = \sigma^2 \tag{2.8}$$

and assume we observe $(x_1, y_1), \ldots, (x_n, y_n)$. If a simple constant is fitted i.e.

$$E[Y \mid X = x] = \alpha,$$

and the conditional errors are assumed to be Gaussian the maximum likelihood estimate for α is

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Although this is obviously an inferior description with respect to model 2.7 in that it does not capture the 'true' model, it can be argued that if the explanatory variable is unknown this analysis has averaged over the response for the missing variable and incorporated the regression effect into the dispersion parameter. The marginal distribution of the response is captured and the marginal mean is consistently estimated.

In this section we will show that this does not hold for truncated models. That is, if regressors are missing from the analysis, the estimates produced are biased and do not have an interpretation in terms of estimating the marginal expectation over the omitted covariates distribution.

2.5.2 The group truncated model

We consider the group truncated case, although the ideas extend to other truncated models. As before we will consider the fitting of a constant parameter to data involving some treatment effect. We will consider the asymptotic mean estimate for a sequence of groups containing two individuals.

The effect of under fitting is examined as follows. We consider groups of size 2 and a treatment X with distribution $f(x_1, x_2)$, where x_i is the covariate for the ith member of the group. Now, for each group there are four possible outcomes

Individual 1	Individual 2	probability
0	0	$q_1 q_2$
0	1	$p_1 q_2$
1	0	$q_1 p_2$
1	1	$p_1 p_2$

where p_i is the probability individuals i's response is 1, and $q_i = 1 - p_i$. We will model the p_i by the logistic link.

Consider a sequence $(x_{11}, x_{21}), \ldots, (x_{1N}, x_{2N})$ drawn from $f(x_1, x_2)$ and the associated responses $(y_{11}, y_{21}), \ldots, (y_{1N}, y_{2N})$. The notation is logically as before, but now indexes the untruncated sample. For example N is now the size of the untruncated sample, and y_{ij} refers to the response of the ith individual in the jth untruncated group.

If only an intercept term is fitted, the score function for β from the truncated sample of size n is

$$U(\beta) = n_{1,0}(1 - \frac{2p}{2p - p^2}) + n_{1,1}(2 - \frac{2p}{2p - p^2}) = 0$$

where $n_{1,0}$ is the number of groups with response of one and zero, and $n_{1,1}$ is the number of groups with response one and one. Consider the function $I_j(k, l)$ defined by

$$I_j(k,l) = \begin{cases} 1 & \text{if } y_{1j} = k \text{ and } y_{2j} = l \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$n_{1,1} = \sum_{j=1}^{N} I_j(1,1)$$

and

$$n_{1,0} = \sum_{j=1}^{N} I_j(1,0) + I_j(0,1).$$

The marginal expected value of $I_j(k, l)$ is

$$E[I_j(k,l)] = \int\int p_1^k(x_1) q_1^{1-k}(x_1) p_2^l(x_2) q_2^{1-l}(x_2) f(x_1, x_2) dx_1 dx_2. \tag{2.9}$$

The score equation for estimating β is from 2.2.1

$$\frac{\sum_{j=1}^{N} I_j(1,1)}{N}(2 - \frac{2p}{2p - p^2}) + \frac{\sum_{j=1}^{N} I_j(1,0) + I_j(0,1)}{N}(1 - \frac{2p}{2p - p^2}) = 0. \tag{2.10}$$

Note that as p is a one to one function of β we can solve for p to obtain the MLE of β.

We wish to determine the estimate of p that is obtained asymptotically, which is the value of p the MLE converges to for large N. Now

$$\sum_{j=1}^{N} I_j(1,1)$$

is the sum of independent random variables with expectation given by Equation 2.9. Hence provided this integral is convergent which is guaranteed as the integrand is bounded by 1,

$$\frac{1}{N}\sum_{j=1}^{N} I_j(1,1) \xrightarrow{a.s.} E[I_j(1,1)]$$

and

$$\frac{1}{N}\sum_{j=1}^{N} I_j(1,0) + I_j(0,1) \xrightarrow{a.s.} E[I_j(1,0) + I_j(0,1)].$$

So the estimate $\hat{p}$ converges strongly to

$$\frac{2E[I_j(1,1)]}{2E[I_j(1,1)] + E[I_j(1,0) + I_j(0,1)]}.$$

While the integral in Equation 2.9 may not have an explicit solution it can be estimated by numerical methods.

2.5.3 Examples

We will consider four simple cases.

1. 0/1 treatment at the individual level.
2. 0/1 treatment at the group level.
3. Uniformly distributed covariate at the individual level.
4. Uniformly distributed covariate at the group level.

The simplest example is where in each group individual 1 receives treatment 1, and individual 2 receives treatment 2. In this case

$$f(x_1, x_2) = \begin{cases} 1 & x_1 = 0, x_2 = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Figure 2.4 contours the value of $BF = (\hat{p} - \bar{p})/\bar{p}$ for various values of p_1 and p_2. The quantity $\bar{p}$ is the marginal probability of response over the covariates in the untruncated experiment, and thus the displayed quantity is a scaled difference between the two estimates. The use of BF is to some degree arbitrary, and alternative measures may be of interest in particular circumstances. Note that the truncated estimate consistently underestimates the untruncated marginal mean except in the trivial case where there is no treatment effect.

Figure 2.5 contours BF for a 0/1 treatment applied at the group level, with the probability of receiving either treatment equal to .5 . In this case the truncated estimate consistently overestimates the marginal mean except when there is no treatment effect. This can be appreciated when one considers what occurs when the probability of success is 0 for one treatment. In this case only groups from the other treatment are observed, and the truncated estimate is consistent for the mean of this group. Thus the estimated marginal mean is twice the true marginal mean in this case.

With continuous covariates, we consider a uniform distribution

$$f(x_1, x_2) = \begin{cases} 1 & -1/2 \leq x_1 \leq 1/2 \text{ and } -1/2 \leq x_2 \leq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

and the probability of response is modelled by an intercept and linear term in the predictors. In this case the marginal mean is consistently estimated, due to the independence of x_1 and x_2. This can be seen by noting that the integral for $E_{1,1}$ from 2.9 can be written as

$$\left(\int p_1(x_1) f(x_1) dx_1 \right)^2$$

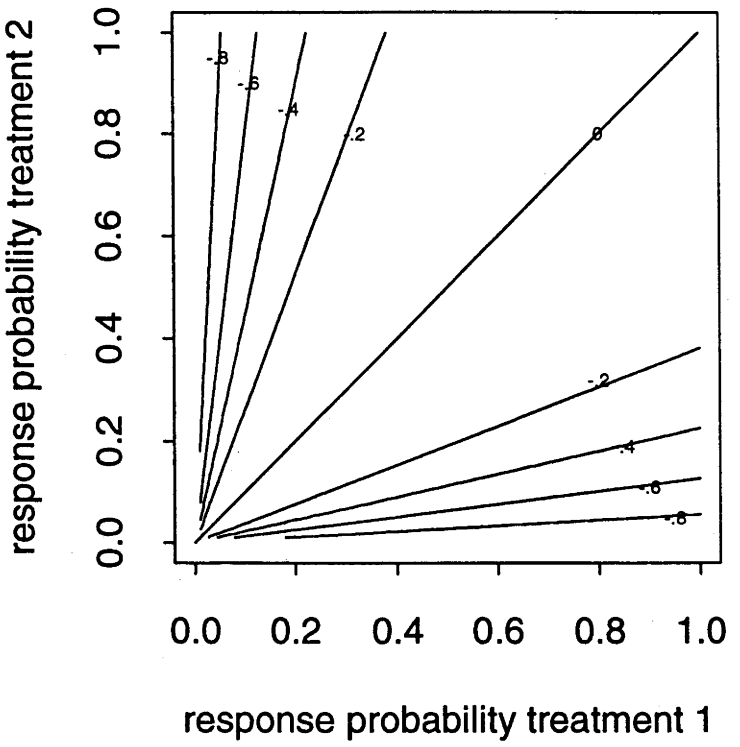


Figure 2.4: Contour plot of BF over the probability of response for treatment 1 versus the probability of response for treatment 2. Treatment disparate within each pair. See text for details.

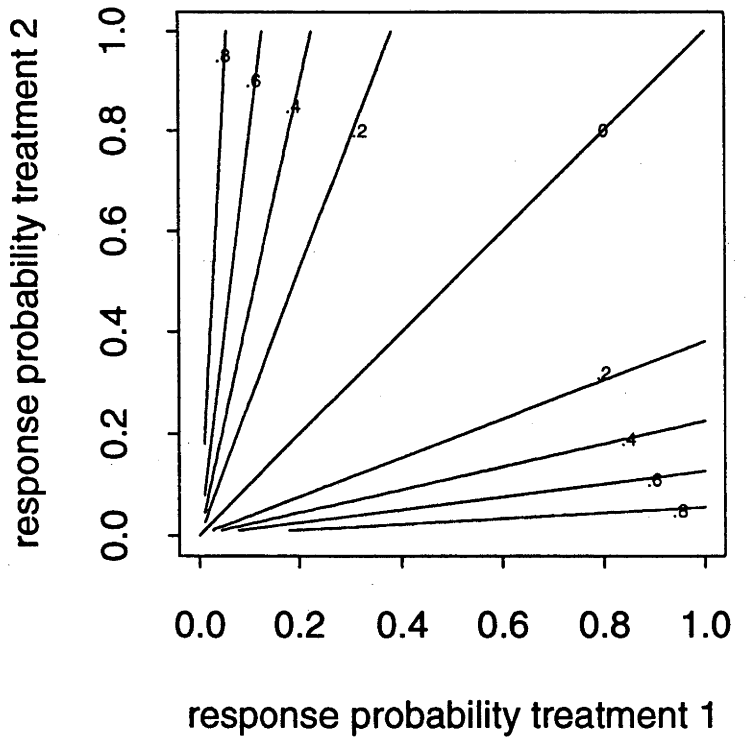


Figure 2.5: Contour plot of BF over the probability of response for treatment 1 versus the probability of response for treatment 2. Treatment applied at the group level. See text for details.

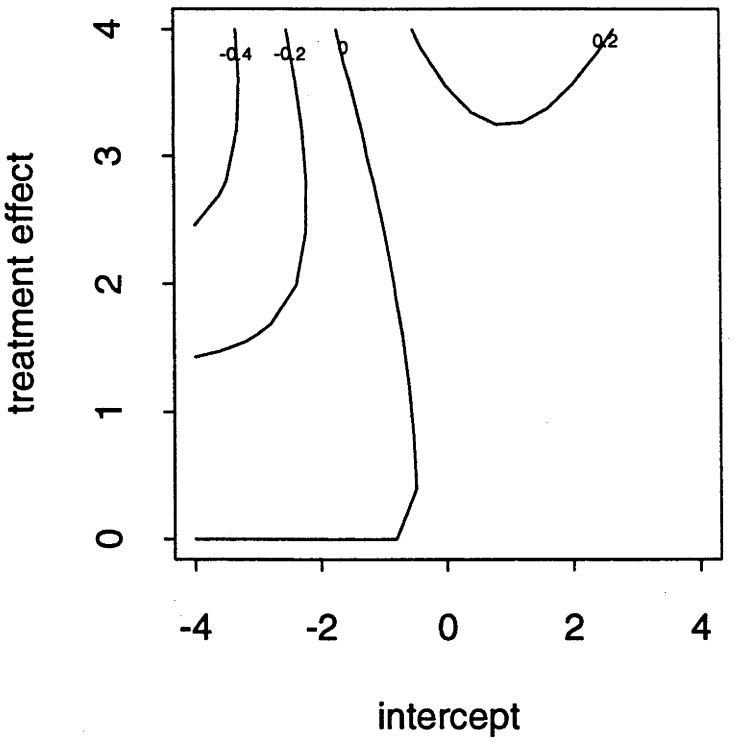


Figure 2.6: Contour plot of BF over the different intercept and treatment effects. Uniform covariate applied at the group level. See text for details.

due to the independence of x_1 and x_2 and hence asymptotically the truncated MLE converges to

$$\int p_1(x_1) f(x_1) dx_1,$$

which is the marginal probability of response for a randomly selected individual in the complete sample. This result only holds if the covariates are independently and identically distributed across the units. Any form of dependence introduces biases.

Figure 2.6 investigates group level uniform covariates distributed as

$$f(x_1, x_2) = \begin{cases} 1 & -1/2 \leq x_1 \leq 1/2 \text{ and } -1/2 \leq x_2 \leq 1/2, x_1 = x_2 \\ 0 & \text{otherwise.} \end{cases}$$

Thus there is dependence in this case due to the covariates being the same for each individual, and the bias is non zero.

Having examined the figures, several points should be made. First, when the treatment effect is zero the estimate is consistent as expected. In addition, the bias increases smoothly as we move away from the consistent case. This is not unexpected, but is worth noting as it ensures that if our model is approximately correct we will get approximately correct answers. Thus there is no catastrophic failure of the model. In addition the bias is smallest when the p's are large as in this case the truncation is mild.

With categorical covariates, note that the bias is reversed when the treatment is applied at the group level, due to the change in the pattern of observation of the two groups. Heterogeneity at the group level leads to an excess of (1,1) responses, while heterogeneity at the individual level leads to an excess of (0,1) responses.

With these problems occuring with the truncated model, the issue of testing lack of fit arises. In this case the data consists of binary observations and thus the usual results relating to the asymptotic distribution of the residual deviance and Pearson chi squared statistics do not hold. This has been discussed in McCullagh (1985), and McCullagh (1986). As in the binary case we can consider grouping the data to construct the tests. For example, consider a group with covariates X_j. If we observe K_j such groups, with response for the kth group in this set being being $[y_{1k}, \ldots, y_{n_j k}]$ we can write the contribution of these to the likelihood as

$$\frac{\Pi_{i \in \mathcal{R}_j} p(\beta, x_{ij})^{\sum_k y_{ik}} q(\beta, x_{ij})^{K_j - \sum_k y_{ik}}}{P_j(\beta)^{K_j}} \tag{2.11}$$

Thus, provided the number of distinct covariate configurations remains fixed as the sample size increases, and thus the number of parameters in the saturated model, we can construct likelihood ratio tests to test our hypothesised model against the saturated model. Then standard results can be applied to show that $-2LR$ has an asymptotic Chi-squared distribution with $T - p$ degrees of freedom, where T is the number of parameters in the saturated model and LR is the log likelihood ratio.

We can evaluate the saturated model for the ith cell of the jth covariate configuration by setting

$$\frac{p(\beta, x_{ij})}{P_j(\beta)} = \frac{\sum_k y_{ik}}{K_j}.$$

These n_j equations are simple to solve and the solution always exists.

Though this procedure is asymptotically valid, problems arise in finite samples. If any continuous covariates are present the number of parameters in the saturated model grows with the sample size and the critical assumption regarding the test will fail. If there are only categorical covariates, then problems still arise due to the grouping which, in light of the possible permutations of covariates within each group, greatly increase the potential number of combinations. This again causes sparse data. Therefore it is only recommended to use this approach when the number of classes of groups is small, or the data is very expansive.

An alternative to the use of residual deviance is to embed the truncated model assuming independence in an over-dispersed alternative and construct tests of the specification. This has the advantage of providing information regarding the nature of the lack of fit. This is considered in Chapter 6.

An alternative to formal testing is the use of graphical techniques. The usefulness of graphical techniques in standard regression problems is well known. The definition and use of residuals from generalized linear models is discussed in Pierce & Schafer (1986). The use of diagnostic techniques in untruncated logistic regression is considered by Pregibon (1981) and Landwehr, Pregibon, & Shoemaker (1984). As Jennings (1986) points out these results rely on asymptotic arguments that are invalid when the data are pure binary as in this case. Fowlkes (1987) considers the binary case and avoids the binary nature of residual distributions by using a smoothing algorithm. The data are smoothed, based on a weighted mean of the responses that are 'near' the specified point in the regression space. The residual is defined as the difference between the smoothed response, and the estimated mean. Thus if the regression space is well behaved, and the weighting is chosen carefully, the residuals, though dependent, have useful diagnostic properties.

Attempts to apply this idea in the truncated case have failed due to the problem of defining an appropriate smoothing algorithm to smooth the truncated response. With group truncated data the concept of closeness in terms of the regression variables is complicated due to the dependence on the marginal response of the other covariates in the group. Thus it is not apparent how to sensibly define distance in this case, and thus provide a useful weighting system. While we can potentially think of groups that are 'near' in terms of covariates, defining a metric for this nearness, which does not introduce systematic patterns in the residuals, is not obvious.

This procedure appears to be only useful if the covariates are fairly densely clustered together in space, and thus effective grouping can be performed to give essentially unbiased estimates of the marginal means. The final point to consider is the potential use of the procedure. If it was successfully implemented in the group truncated case, how would systematic patterns in the residual plot be interpreted. Apart from outliers, the truncation will affect the interpretation of any pattern observed, and in ways that are specific to the case at hand. Thus the construction of such plots may not be as useful as is usually found.

2.6 Federal Office of Road Safety Data

This section examines the effects of various covariates on the probability of death for passengers involved in fatal car accidents. This analysis is based on the so called "Fatal Files" which are collected by the Australian Federal Office of Road Safety on a biennial basis. These files consist of passenger and vehicle information for all fatal accidents that occur in the target year. The analysis is based on the records for the 1988 and 1990 calendar years.

The aim of the analysis was to estimate the effect of various accident level and individual level covariates on the probability of death. We are particularly interested in comparing the effects with those previously reported for the American FARS data (Evans, 1991). Australian road safety decisions have often been made using results originating from the U.S.A.. It is of considerable interest to see if the effects are similar in the two populations.

To simplify the analysis we restrict attention to single vehicle, frontal impact collisions that involved passenger cars. This produced a data set with observations on 306 individuals involved in 111 accidents. Note that cars with only a driver are non informative for all methods.

From this data set, the accidents involving a front seat passenger and a driver where only one dies were extracted for use in the conditional analysis. This parallels the paired analysis of Lui et al. (1988). The resulting data set for the conditional analysis consisted of information from 76 accidents.

We note the relatively modest size of the datasets. This is caused by the small population base in Australia and the short, two year, collection period. By contrast the same criteria in the FARS database would yield many thousands of accidents. Consequently it is important not to be unrealistic in our expectations about the precision with which we can estimate the effects or about the complexity of the model and the model selection procedures that can be used. Nevertheless, we believe that it is an important exercise to take a model using the proven effects for the U.S.A. data and see if the same effects are suggested by the Australian data. In

time when more data becomes available in Australia the effects can be confirmed more conclusively.

From the wide range of variables available for each individual, a subset was chosen because of their previous demonstrated association with fatalities in U.S.A. car accidents (Evans, 1991). The variables used were the sex of the individual, the age in four categories as suggested by previous results (Evans, 1991), restraint usage and seat location.

At the vehicle level we sought to include variables describing the impact speed and the vehicle deformation. The inclusion of speed as a continuous variable is problematic. First the effect of speed on a logistic scale is unknown, and will certainly not be linear exhibiting both threshold and plateau behaviour. The use of B-splines is easily implemented in this regression case, and has been done so successfully. It can be performed as a direct extension of the methods in Sleeper & Harrington (1990) and Stone & Koo (1985). The use of B-splines in this situation is not particularly novel, as the implementation relies only on the regression model, and the fact that if the model matrix is full rank the parameters can be estimated. Therefore the use of these splines will not be considered here. In addition, continuous modelling would require larger datasets than the present one. The second problem with using the speed variable, is that the measurements are imprecise as they are made subjectively after the accidents by police. Variables coding for speed limit and whether the car was speeding were included as surrogate variables. It is hoped that a combination of these variables will adequately approximate the vehicle speed. The degree of damage to the vehicle was also included as an accident level variable. The individual and accident level variables are described more completely in Table 2.1.

The models were fitted using the methods presented in Section 2 and the results are given in Table 2.2. A collection of S-plus functions and C routines were developed to perform TLR. All covariates were treated as factors and the design matrix was constructed using treatment (Chambers & Hastie, 1992) contrasts. The effect of the lowest level of each factor was set to zero.

The coefficients in Table 2.2, particularly for TLR, are in general agreement with the U.S.A. effects. The sex coefficient of .37 can be compared to a relative risk in the U.S.A. of approximately 1.3 for belted females to belted males (Evans, 1991), which converts to a log odds of approximately .26. The age coefficient of 1.21 for 60 plus can be compared to a relative risk in the U.S.A. of approximately 3.5 for belted 70 year old females relative to 20 year old males which converts to a log odds of 1.25. Evans (1991) reports a relative risk in the U.S.A. of rear seat passengers to front seat passengers of approximately .62 which converts to a log odds of -.42. This can be compared to the estimate of -.33 which we have found for the Australian data. The agreement found between the Australian based estimates and the commonly accepted values for the American effects is encouraging. Of course we acknowledge that the standard errors are still fairly large as a result of the small sample size for a truncated analysis and a more conclusive analysis with a more detailed model must await the availability of more data in Australia.

The agreement between CLR and TLR is also encouraging, but must be interpreted cautiously due to the dependence caused by the common dataset used in the estimation. Parameter estimates are in general agreement and TLR generally results in more precise estimates.

There are some additional features of the TLR estimates which are of interest.

Variable	Level	Category
Damage	0	major damage
	1	extensive damage
Resavl	0	restraint worn/restraint use unknown
	1	restraint definitely not worn
Sex	0	male
	1	female
Speedlim	0	speed limit 0-60 kmh
	1	speed limit 61-104 kmh
	2	speed limit 105+ kmh
Speedcat	0	not over speed limit
	1	over speed limit
Age	0	0-15 years old
	1	15-25 years old
	2	25-60 years old
	3	60+ years old
Perloc	0	front seat
	1	not front seat

Table 2.1: Description of variables used from 1988/90 FORS data.

The use of a restraint has a profound positive influence on the odds ratio of surviving a crash. There is an increasing vulnerability with age and females are also more vulnerable. As expected increasing speed and vehicle damage increase the odds ratio of death. It is also interesting to note that the risk of dying in most accidents is low. For example, if an individual has all levels of factors at zero, then the estimated probability of death in an accident is $(1 + \exp 3.5)^{-1} = .029$. Note that only TLR allows us to make conclusions of this type. We again caution that the sample sizes used in this truncated analysis are quite small and the magnitude of the parameter estimates should be interpreted very cautiously at this early stage. The analysis should be regarded as exploratory.

The advantage of TLR in analysing the present dataset is that it is more precise and it enables the estimation of interesting accident level effects. On the other hand since it required the knowledge of those accident level variables, any measurement problems with those variables may flow through to the TLR estimates.

2.7 Discussion

The results and examples given in this chapter are encouraging for the logistic regression analysis of accident level truncated binary data on deaths. We have seen that viable regression estimation can be done using either Truncated Logistic Regression or Conditional Logistic Regression. The methods are applicable to any situations in which binary variates and associated covariates are observed if and only if at least one member of the group has a positive binary response. The efficiency loss due to truncation can be substantial but not catastrophic.

The choice between truncated and conditional logistic regression for the analysis

Variable	Level	Truncated	Conditional
Intercept	1	3.50(.88)	NA
Damage	1	0.53(.61)	NA
Resavl	1	1.10(.35)	1.35(.65)
Sex	1	0.37(.26)	0.61(.37)
Speedlim	1	0.62(.51)	NA
	2	0.83(.83)	NA
Speedcat	1	1.59(.49)	NA
Age	1	0.02(.50)	-0.45(.91)
	2	0.25(.54)	0.24(.53)
	3	1.21(.71)	1.77(1.4)
Perloc	1	-0.33(.26)	NA

Table 2.2: Log odds ratio estimates for single vehicle, frontal impact collisions, using the combined 1988 and 1990 FORS data. (Standard errors are in parenthesis, NA=not applicable.)

of group truncated binary data will be governed by several factors. First consider the accident level effects. For random or unstructured accident level effects, conditional logistic should be used since it eliminates all accident level effects from the likelihood. If the accident level effects that affect the probability structure of the response are known and so the individual probabilities can be adequately modelled, then either truncated logistic regression or conditional logistic regression can be used. Normally the possibility of estimating accident level effects and the greater efficiency of TLR would lead to TLR being the preferred method in such cases. However if the accident level variables are unavailable or very inaccurate, then CLR would be indicated.

Alternatively, the two methods may be regarded as complementary and in situations where it is possible, both can be fitted and the results compared. This may provide insight into the appropriateness of the models and thus the true pattern of the data.

In summary, the two main obstacles to the application of TLR are the availability of accident level covariates and their suitability for modelling systematic effects across accidents. When it can be applied, only TLR offers the possibility of estimating accident level effects and it also has higher efficiency than CLR.

Chapter 3

Group Truncated Ordinal Regression

3.1 Introduction

This chapter considers the modelling of ordinal data that is subject to group truncation. This is a natural extension of the truncated binary model discussed in Chapter 2, and it will be seen that the truncated binary model is a special case of the truncated ordinal model.

As an example consider the following ordinal scale for injuries sustained in a motor vehicle accident:

1. no injury.
2. mild injury.
3. severe injury.
4. death.

Again we consider a grouped response, where the group is only observed if the maximum response in the group is greater than a certain level. For example with the fatality data, if only accidents involving fatalities are observed then the maximum response in the accident must be greater than three.

The results of this chapter centre on the extension of the class of models presented in McCullagh (1980) for modelling ordinal responses. This extension is undertaken both for its own merit, and due to the results of the conditional logistic technique not extending to different link functions and to true ordinal data.

The analysis of ordinal accident data has been performed by Weiss (1992) who applied econometric methods based on Gaussian latent variables to data on the severity of injury suffered by motorcycle riders. In its simplest form this model is contained in McCullagh (1980), with the probit link, but the various modifications incorporated to allow the modelling of the variability invalidate this. The data analysed did not exhibit truncation. In a second paper, Weiss (1993) has analysed motorcycle accident data where the injury severity for head and body injuries was obtained, and the accidents were only observed if some injury occured. The analysis was a direct extension of the previous work and explicitly modelled the correlation, over-dispersion and means of the response, while adjusting for the truncation. This

chapter differs from the work of Weiss by allowing more arbitrary groupings, link functions and truncation patterns.

This chapter is arranged as follows. In section 3.2, the conditional approach is briefly reviewed and it is shown that it depends crucially on the assumption of the logistic link. In section 3.3 the truncated ordinal model is presented and the maximum likelihood estimates derived. Section 3.4 presents some examples. The chapter concludes in Section 3.5 with discussion.

3.2 Conditional Logistic Regression

In this section we extend the discussion of Chapter 2 regarding the use of CLR with truncated data and consider its extension to other link functions and the ordinal case. Recall that with binary data the contribution of the truncated groups to the log likelihood was zero and hence these groups could be ignored without loss of efficiency. With a group truncated ordinal response we would like to construct a conditional likelihood with this feature. In general, for ordinal data with ordered responses $1, \ldots, k+1$ the conditioning event would be the number of responses $C = c$ which are greater than some truncation point l where $l \leq k$. Now if $\mathcal{R}$ denotes the set of individuals in the group, then

$$P(C = c) = \sum_{\mathcal{I} \in \mathcal{R}_c} \left\{ \prod_{\mathcal{I}} \frac{(1 - \gamma_{il})}{\gamma_{il}} \right\} \prod_{\mathcal{R}} \gamma_{il}$$

where $\mathcal{R}_c$ is the set of all subsets of $\mathcal{R}$ of size c and $\gamma_{i1}, \ldots, \gamma_{ik+1}$ is the set of cumulative probabilities for individual i. Thus γ_{ij} is the probability that individual i's response is less than or equal to category j. As an example $\gamma_{ik+1} = 1$ for all i. Following McCullagh & Nelder (1989) we consider link functions of the form

$$g(\gamma_{ij}) = \eta_{ij} = \theta_j - \beta' x_i$$

where β, a $p \times 1$ vector, and $\theta_j, j = 1, \ldots, k$ are the regression parameters and x_i is a $p \times 1$ vector of covariates measured on the ith individual. The negative sign is used by convention to ensure that positive parameter values imply increased probability of the response being in a higher category. If we let $h()$ be the inverse of $g()$, implicitly defined if necessary, then

$$P(C = c) = \sum_{\mathcal{I} \in \mathcal{R}_c} \left\{ \prod_{\mathcal{I}} \frac{(1 - h(\eta_{il}))}{h(\eta_{il})} \right\} \prod_{\mathcal{R}} h(\eta_{il})$$

and if y is the set of observed levels

$$P(y|C = c) = \frac{\prod_{\mathcal{R}} h(\eta_{iy_i}) - h(\eta_{iy_i - 1})}{\prod_{\mathcal{R}} h(\eta_{il}) \sum_{\mathcal{I} \in \mathcal{R}_c} \left\{ \prod_{\mathcal{I}} \frac{(1-h(\eta_{il}))}{h(\eta_{il})} \right\}}. \tag{3.1}$$

Note that when $k > 1$ and $C = 0$ the likelihood 3.1 is not degenerate as occurs in the binary case. Thus while the conditional approach is still feasible in the ordinal case, it does not have its full appeal, in that the truncated groups still contain information regarding the parameters. It thus becomes one of a competing set of conditioning events. Its viability for estimation is due to the conditioning in effect reducing the sampling process to observing a sequence of group level experiments,

i.e. the configuration of the response within each accident, given C. Thus the conditional approach can still be used to estimate the parameters, but with reduced efficiency compared to the estimate produced from the full, untruncated data.

The cancellation of the group level effects in the binary case is the primary reason for the popularity of conditional logistic regression. However it is easily seen that the group level effects do not cancel from 3.1, even for the logistic link. Note though that McCullagh (1984) has investigated techniques to eliminate nuisance parameters from the proportional odds model, but his method is not based on 3.1.

To check that the cancellation implies the logistic link for $k = 1$, take $f = h/(1-h)$. Then cancellation must apply for groups of size 2 where the individual with positive response has $x = 0$. So

$$\frac{f(\theta)}{f(\theta) + f(\theta + \eta_1)} = c(\eta_1)$$

where $c()$ is some function and $\eta_1 = \beta_1 x$ for the individual not dead, or $f(\theta + y) = f(\theta)c_1(y)$, where

$$c_1(y) = \frac{1 - c(y)}{c(y)}.$$

Taking $\theta = 0$, we obtain $c_1(y) = f(y)/f(0)$ and $f(\theta + y) = f(\theta)f(y)/f(0)$. These equations have solution $f(\eta) = exp(a + b\eta)$ for some a and b which implies that the link function is logistic.

Note that in O'Neill & Barry (1993b) it was stated that the fact that 3.1 was not independent of the parameters for $C = 0$ meant that it cannot be used for analysing group truncated data. In light of the above discussion this statement is false. What can be concluded is that the estimate may be inefficient due to the conditioning, and that the attractive feature of the CLR likelihood, the cancellation of group level effects, does not occur. Thus the use of Truncated Ordinal Regression, which is fully efficient with truncated data, should be favoured.

In generalized linear modelling we are used to inference being fairly robust to the choice of link function, so it is a matter of some concern that the inclusion of group level effects in the conditional likelihood is determined by whether the logistic link function is assumed. Thus the use of conditional logistic models in situations where the link function is not logistic will cause the group level effect to behave as missing predictors with the resulting attenuation in the estimates produced by the model. Of course, the effect of this misspecification is small if the link function is closely approximated by the logistic link. For instance, numerous authors have shown the closeness between the logistic and probit link (Zeger, Liang, & Albert, 1988), although this similarity does not extend into the tails where the modelling of any rare event will occur.

A more serious problem with the use of conditional models is that even if the logistic link is assumed the property of no group level effect in the conditional likelihood 3.1 is not preserved when we split categories. Thus instead of conditioning on a limited set of response configurations, we will consider the extension of truncated logistic regression to group truncated data by conditioning on the event that the group is observed.

3.3 Group truncated Ordinal Regression

3.3.1 Proportional Odds Model

Since the likelihood equations for group truncated ordinal regression are similar to those for conventional ordinal regression, we begin with a brief summary of the standard results presented in McCullagh (1980) and McCullagh & Nelder (1989). We suppose that there are N distinct experimental situations, and that for each experimental situation there is a response with cumulative frequencies $z_i' = (z_{i1}, \ldots, z_{ik})$, where the response has $k+1$ categories, and there are m_i individuals observed. So z_{ij} is the number of individuals in the ith sample with response at most j and $E(z_{ij}) = m_i\gamma_{ij}$, $\gamma_i' = (\gamma_{i1}, \ldots, \gamma_{ik})$.

Consider link functions of the form

$$g(\gamma_{ij}) = \eta_{ij} = \theta_j - \beta_1' x_i$$

where β_1 is a $p \times 1$ vector of parameters and x_i is the $p \times 1$ vector of covariates associated with the ith experimental situation. Let

$$D_i = \text{diag}\left(\frac{\partial\gamma_{ij}}{\partial\eta_{ij}}\right) = \text{diag}\left(\frac{1}{g'(\gamma_{ij})}\right)$$

which is $\text{diag}(\gamma_{ij}(1-\gamma_{ij}))$ for the logistic link and

$$\Gamma_i = m_i[\gamma_{ij}(1-\gamma_{ij'})], \quad j \le j'$$

where Γ_i is the covariance matrix of the ith response and is thus symmetric. Thus Γ_i has a tri- diagonal inverse with entries

$$\begin{aligned} \Gamma_i^{j,j} &= \frac{\gamma_{i,j+1} - \gamma_{i,j-1}}{(\gamma_{i,j+1} - \gamma_{i,j})(\gamma_{i,j} - \gamma_{i,j-1})} \quad j = 1, \ldots, k \\ \Gamma_i^{j,j+1} &= \frac{-1}{\gamma_{i,j+1} - \gamma_{i,j}}, \quad j = 1, \ldots, k-1. \end{aligned}$$

Then for the sample of size N let

$$\begin{array}{llll} D = \text{blockdiag}(D_i), & \Gamma = \text{blockdiag}(\Gamma_i), \\ M = \text{diag}(m_i) \otimes I, & z' = (z_1', \ldots, z_N'), \\ \gamma' = (\gamma_1', \ldots, \gamma_N'), & X_i = (I_k, -ex_i'), \\ X' = (X_1', \ldots, X_N') \end{array} \tag{3.2}$$

where $\otimes$ is the Kronecker product, I_k is the $k \times k$ identity matrix and e is a $k \times 1$ column vector of ones. If

$$\beta' = (\theta_1, \ldots, \theta_k, \beta_1')$$

then the score functions are

$$D_\gamma l = M\Gamma^{-1}(z - M\gamma)$$

and

$$D_\beta l = X'DM\Gamma^{-1}(z - M\gamma)$$

where D_β is the vector differential operator $(D_\beta)' = (\partial f/\partial\theta_1, \ldots, \partial f/\partial\beta_p)$. The Fisher information is

$$\mathcal{I}(\beta) = X'DM\Gamma^{-1}MDX$$

so the Fisher scoring method for estimation is

$$(X'WX)(\hat{\beta}_{i+1} - \hat{\beta}_i) = X'W(MD)^{-1}(z - M\gamma)$$

where

$$W = DM\Gamma^{-1}MD.$$

3.3.2 Truncated Proportional Odds Model

We now consider the truncated case. For simplicity we will now assume that $M = I$, i.e. that we only observe a single event for each individual. This case is also of the most practical importance as it relates to the traffic data where the event is the injury received. The case where $M \neq I$ is of less practical importance and will be discussed later.

Consider a single group subject to group truncation, and for clarity suppress the i (individual) level subscript. This case is equivalent to the situation outlined in the previous section with the number of individuals in the group equal to N. The likelihood for an observed group is then

$$L_T(\beta, X) = \frac{L(\beta, X)}{1 - \prod_{\mathcal{R}} \gamma_l} = \frac{L(\beta, X)}{P_l(\beta, X)}$$

where as before the group is only observed if the maximum response over the group is greater than l and

$$P_l(\beta, X) = 1 - \prod_{\mathcal{R}} \gamma_l = 1 - Q_l(\beta, X)$$

is the probability of observing the group. Now

$$\begin{aligned} D_\beta log P_l(\beta, X) &= \frac{-D_\beta \prod_{\mathcal{R}} \gamma_l}{P_l(\beta, X)} \\ &= \frac{-Q_l(\beta, X)}{P_l(\beta, X)} \sum_{\mathcal{R}} \frac{\partial \gamma_l}{\partial \eta_l} \frac{D_\beta \eta_l}{\gamma_l} \\ &= \frac{-Q_l(\beta, X)}{P_l(\beta, X)} X'DE_T\gamma^{-1} \end{aligned}$$

where $E_T = I \otimes E_{l,l}$ with $E_{l,l}$ being a $k \times k$ matrix with a one in the (l, l) position and zeros elsewhere and I has dimension the number in the group. Also, γ^{-1} is the vector of inverses of the elements of γ and matrices X and D are defined in 3.2. Thus writing $P = P_l(\beta, X)$ and $Q = Q_l(\beta, X)$ the score function for the group is

$$X'D\left\{\Gamma^{-1}(z - \gamma) + (Q/P)E_T\gamma^{-1}\right\}$$

which can be written as

$$X'D\Gamma^{-1}\left\{(z - \gamma) + (Q/P)\Gamma E_T\gamma^{-1}\right\}. \tag{3.3}$$

Consider the sub vector of the score corresponding to a particular individual. The jth element of the vector 3.3 contained in braces is

$$\{(z-\gamma)+(Q/P)\Gamma E_T\gamma^{-1}\}_j = \begin{cases} z_j - \gamma_j/P + (Q/P)\gamma_j/\gamma_l & j \leq l \\ z_j - \gamma_j/P + (Q/P) & j > l. \end{cases}$$

Now for $j \leq l$

$$\begin{aligned} E(z_j|observed) &= \frac{Pr(y \leq j, \text{observed})}{Pr(\text{observed})} \\ &= \gamma_j/P - (Q/P)\gamma_j/\gamma_l \end{aligned}$$

and for $j > l$

$$\begin{aligned} E(z_j|\text{observed}) &= \frac{Pr(y > j, \text{observed})}{Pr(\text{observed})} \\ &= \gamma_j/P - (Q/P). \end{aligned}$$

So the score statistic for β is

$$U_T(\beta, X) = X'D\Gamma^{-1}(z - \mu_T) \tag{3.4}$$

where μ_T is the mean of an observed z given that it is subject to truncation. Now if we use Fisher scoring to estimate β, we require

$$\mathcal{I} = -E\{\frac{\partial^2 log L_T}{\partial\beta\partial\beta'}\}.$$

But,

$$\begin{aligned} \mathcal{I}(\beta) &= E\{U_T(\beta, X)U_T(\beta, X)'\} \\ &= X'D\Gamma^{-1}V_T\Gamma^{-1}DX \end{aligned} \tag{3.5}$$

(3.6)

where $V_T = V_T(\beta, X) = Var(z|\text{observed})$. The diagonal blocks of this matrix contain the covariance of the response vector within an individual response. This is analogous to the Γ matrix in the non-truncated case. Now for a given individual, since for $j \leq j'$

$$E(z_j z_{j'}|\text{observed}) = E(z_{j'}|\text{observed}),$$

it follows that the (j, j') entry of the ith diagonal block of V_T for $j \leq j'$ is

$$\mu_{T(i,j)}(1 - \mu_{T(i,j)})$$

where $\mu_{T(i,j)}$ is the jth element of the component of μ_T corresponding to individual i.

In a departure from the non-truncated model, the truncation causes the responses between individuals to be correlated. As defined previously, let $z_{i1}, \ldots z_{ik}$ be the response for individual i in the group. Then for two individuals i and j, straightforward calculation leads to:

for $a \leq l$ and $b \leq l$

$$E[z_{ia}z_{jb}] = \frac{\gamma_{ia}\gamma_{jb}}{P} - \frac{Q}{\gamma_{il}\gamma_{jl}P}$$

for $a > l$ and $b > l$

$$E[z_{ia}z_{jb}] = \frac{\gamma_{ia}\gamma_{jb}}{P} - \frac{Q}{P}$$

for $a \leq l$ and $b > l$

$$E[z_{ia}z_{jb}] = \frac{\gamma_{ia}\gamma_{ib}}{P} - \frac{\gamma_{ia}Q}{\gamma_{jl}P}.$$

It is thus simple to complete the off diagonal blocks of V_T, given that the expectations are already known.

We now consider the score function for the sample of groups. Assume we have a sample of G groups. Letting $X_g, D_g, \Gamma_g, \mu_{gT}, V_{gT}$ and z_g denote the above defined quantities for the gth group and defining

$$\begin{array}{rclrcl} D & = & \text{blockdiag}(D_g), & \Gamma & = & \text{blockdiag}(\Gamma_g), \\ z' & = & (z'_1, \ldots, z'_G) & \gamma' & = & (\gamma'_1, \ldots, \gamma'_G), \\ X' & = & (X'_i, \ldots, X'_G) & \mu'_T & = & (\mu'_{1T}, \ldots, \mu'_{GT}), \\ V_T & = & \text{diag}(V_{gT}) & & & \end{array} \tag{3.7}$$

we have, due to the independence of the responses between groups, that the score statistic and Fisher information using this new notation and partitioning are given again by Equations 3.4 and 3.5. Thus the Fisher scoring algorithm for the estimation of β is defined by the equation

$$(X'D\Gamma^{-1}V_T\Gamma^{-1}DX)(\hat{\beta}_{i+1} - \hat{\beta}_i) = X'D\Gamma^{-1}(z - \mu_T),$$

where $\hat{\beta}_i$ is the estimate at the ith iteration. Under mild regularity conditions the MLE's produced from this model have the usual asymptotic properties (see Amemiya (1979)) which can be used for inference.

3.3.3 Model Fitting

The estimation scheme given above was implemented using a suite of C routines and an interface to the Splus language. This implementation is efficient and allows for great flexibility in model specification via the Splus model formula functions. Several approaches were considered for the initial estimate for the Fisher scoring algorithm. Attempts were made to devise a simple consistent estimate of the parameters but there are no obvious candidates. As an alternative the untruncated proportional odds model was fitted to the data. This obviously produces biased estimates, especially of the intercept parameters (θ_j) but no convergence problems were observed, except in cases with *extreme* truncation, where the bias is obviously at its greatest. An alternative estimate when the θ_j are constant across the individuals was to fit simple logistic regressions to each of the z_j and use the intercept terms from these regressions and the average of the treatment effects to provide the initial estimate. This is in effect a poor man's version of the proportional odds model.

When the problem was well defined convergence was rapid, and the sampling distribution of the parameter estimates achieved their asymptotic approximations rapidly. The usual problems of sparse data sets producing divergent parameter

estimates occured. Sparse in the ordinal case refers to cases were no response is observed in a particular category (singular model), only one response is above the truncation point in each group (intercept divergence) or all individuals with a particular treatment exhibit an extreme response (treatment divergence).

Model selection can be done via any of the asymptotic results applying to regular maximum likelihood problems under mild regularity conditions. Thus subsets of the parameters can be tested via likelihood ratio, Wald, or score tests. The asymptotic normality of the MLE's can be used to test for effects and to choose subsets to test with the more attractive likelihood methods, instead of facing the problems inherent in multiple comparisons.

Unfortunately, standard goodness of fit tests do not apply here due to the sparseness of the sample when $m_i = 1$. Thus the usual residual deviance does not have an asymptotic χ^2 distribution, due to the number of parameters in the saturated model growing with N. In addition, the use of visual techniques such as examining the empirical logits is not possible due to the truncation bending the linear predictor space in unexpected ways. This is a major problem in the use of these models as it means that the plausibility of the model cannot be statistically assessed. In the presence of only discrete covariates with a finite number of covariate configurations goodness of fit tests can be derived by careful choice of the null model, in an analogous way to the binary case. In this case the problem is regular and standard methods apply, although it is noted that the number of potential covariate configurations increase rapidly with the number of covariates and categories thus limiting the situations where this grouping is possible. The device of making discrete the continuous covariates to produce categorical data is available, but must be used with caution with truncated data due to the biases produced in the parameter estimates.

3.3.4 Alternative Truncated Models

The model outlined in Section 3.3 can be extended in several ways. In this subsection we briefly consider some of these extensions.

The first extension we will consider is when $M \neq I$, and thus each unit exhibits a clustered response. An artificial example of this is the following. We consider a production run of components. Assume there are N components in the run and that each component has m_i subunits that exhibit an ordinal response. The covariates are measured at the component level. The response from the run is only sampled if at least one response over all the sub-units is above some threshold, l. In this case

$$P(\beta) = 1 - \prod_{\mathcal{R}} \gamma_l^{m_i}$$

but the results of Section 3.3 can be easily modified to produce estimates in this case. This model though may be implausible, as we would reasonably expect some form of dependence structure at the component level.

Another extension to consider is the feasibility of estimating a more complicated set of parameters. In the derivation given it was assumed that the θ_j were constant across individuals. This is only appropriate if we can assume that the probability model is constant across individuals. To allow for more complicated patterns in the intercept parameters is straightforward and only requires suitable modification of the design matrix X_i, with the rest of the derivation remaining the same. Thus

if there were two treatments say, each with a different set of intercept parameters, it would only be necessary to augment the design matrix and parameter vector with the additional intercept parameters required. Note though that the linear predictors for a particular individual must remain parallel, due to the ordinal nature of the model, or strange/uninterpretable results will occur.

3.4 Examples

3.4.1 Example 1

The first example demonstrates the connection between group truncated ordinal regression and group truncated binary regression when k=2 and the logistic link is assumed. In this case, we have a single category response z_{i1} for each individual. Therefore, we have the single equation:

$$g(\gamma_{i1}) = \eta_{i1} = \theta - \beta_1' x_i \tag{3.8}$$

and $P(\beta, X) = 1 - \prod_{\mathcal{R}} \gamma_{i1}$. Note though that care must be taken in interpreting the coefficients. If γ_{i1}, is the probability the response is in category 1, we are modelling the probability of being in the lowest category. In this case model 3.8 implies the model

$$g(1 - \gamma_{i1}) = -\eta_{i1} = -\theta + \beta_1' x_i$$

for the probability of success. Hence the parameter estimates are the same except for changes in sign. Note that this symmetry does not extend to asymmetric link functions, and care must be exercised in these cases.

The question arises as to how much extra information is attained by including the ordinal response in the analysis, instead of collapsing categories to produce a truncated binary model. It is impossible to give a succinct answer to this as the problem is multidimensional, and will depend on the design matrix of the regression, but some observations can be made.

We are analysing the response vector $z_1, \ldots, z_k$. The likelihood of a particular group can be written

$$L = \frac{\prod_{\mathcal{R}} f(z_{ik}) f(z_{ik-1}|z_{ik}) \cdots f(z_{i1}|z_{ik}, \ldots, z_{i2})}{1 - \prod_{\mathcal{R}} \gamma_l}.$$

Now the sequence $z_{i1}, \ldots, z_{ik}$ exhibits a Markov property so that $f(z_{ij}|z_{ik}, \ldots, z_{ij+1}) = f(z_{ij}|z_{ij+1})$. We can thus write the log likelihood as

$$logL = \sum_{\mathcal{R}} f(z_{ik}) - log(1 - \prod_{\mathcal{R}} \gamma_l) + \sum_{\mathcal{R}} \sum_{j=1}^{k-1} logf(z_{ij}|z_{ij+1}). \tag{3.9}$$

If the truncation point is $l = k$, i.e. truncation occurs at the top category, then the log likelihood 3.9 consists of two components. The first is the log likelihood for the truncated binary model, and the second relates to the likelihood of the ordinal response that is collapsed in the binary model. As these two terms are linear the information for the parameters will also be additive,

$$I^{ORD} = I^{TLR} + I^{z_1, \ldots, z_{k-1}}$$

where I^{ORD} is the information in the truncated ordinal model, I^{TLR} is the information in the truncated binary model, and $I^{z_1,\ldots,z_{k-1}}$ is thus the information gained by modelling the ordinal response. Note that this information does not depend on the truncation, excluding the loss caused by groups not being observed.

Attempts to define simple meaningful measures of average efficiency loss comparable to those presented in Chapter 2 have foundered due to I_β^{TLR} having no information about the intercept parameters except for θ_k. Attempts to compare the methods over the common parameters are complicated due to the effect of the intercept parameters on these measures. Thus, even if an attractive measure was formulated, the results are multidimensional, and provide little insight into plausible comparisons. A moment's reflection suggests that if the probability of being in in the kth category is large, relative to the $1,\ldots,k-1$ categories then, very little information is lost, whereas if it is small more information will be lost.

3.4.2 Simulation

This section presents a tiny simulation to demonstrate the behaviour of the MLE's in a finite situation. This simulation is small as it is only intended to reassure the reader that the estimates produced are well behaved. It is entirely feasible to perform expansive simulations with the technique, but the additional insight gained from such an approach is small.

As the asymptotic results are derived conditional on a fixed design space so is the simulation. The design was based on a sample of 10 groups of size 5 and 25 groups of size 2, giving a sample of 35 groups and 100 individuals. For each unit within a group a 0/1 covariate, x_1, was generated from a Bernoulli distribution with probability .5. At the group level a uniform (0,1) covariate, x_2 was generated. The model used was:

$$\begin{aligned} logit(\gamma_1) &= \theta_1 - (x_1\beta_1 + x_2\beta_2) \\ logit(\gamma_2) &= \theta_2 - (x_1\beta_1 + x_2\beta_2) \end{aligned}$$

with $(\theta_1, \theta_1, \beta_1, \beta_2) = (-1, 1, -1, -1)$. This gives, for covariates at the lowest level, category probabilities of $(.26, .46, .28)$, and at there highest level $(.73, .23, .04)$. Data was generated using a fixed design matrix from the truncated distribution. For each simulation an approximate 90% CI was calculated using the inverse of the Fisher information found from the fit. The estimated coverage are given in Table 3.4.2. The estimated sampling distribution can be found in Figure 3.1. As can be seen the asymptotic approximations perform extremely well for this sample size.

3.4.3 Road Safety Example

In this section we examine the effect of various covariates on the severity of injury for passengers involved in fatal car accidents. This analysis is based on the so called "Fatal Files" introduced in Section 2.6 which are collected by the Australian Federal Office of Road Safety on a biennial basis. The aim of the analysis was to estimate the effect of various group level and individual level covariates on the category of injury severity. The data set used is the same as that in Section 2.6.

The category of injury was measured on a three point scale. Level 1 injury was assigned when the individual received no injury, or the injury received did

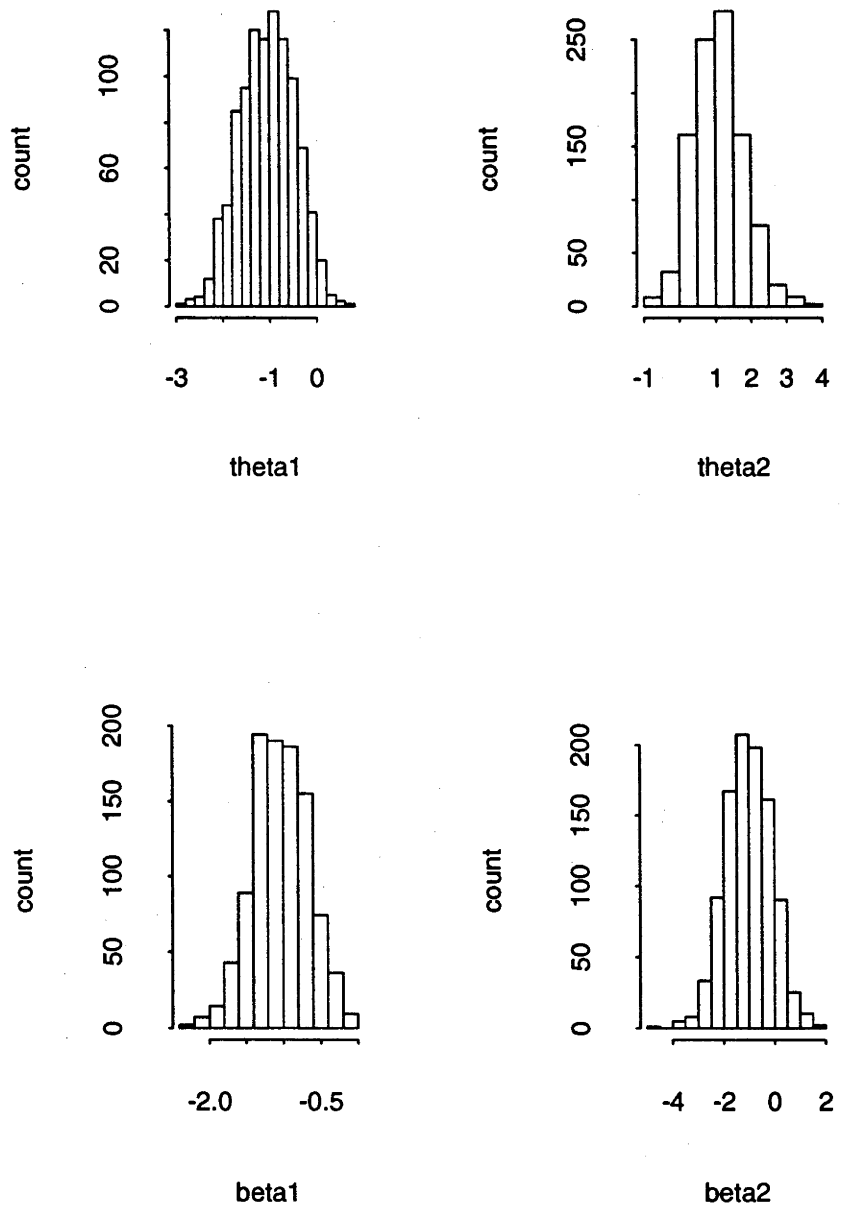


Figure 3.1: Estimated sampling distribution of maximum likelihood estimates of the parameters

Parameter	Coverage
θ_1	.886
θ_2	.906
β_1	.897
β_2	.909

Table 3.1: Estimated coverage for approximate 90% CI with simulated truncated ordinal data.

not require hospitalisation. Level 2 injury corresponded to injury that required hospitalisation. Level 3 injury indicated that the individual died.

The definition of the variables used is given in Table 2.1. The maximum likelihood estimates are given in Table 3.2. The model was parametrised by setting the effect of the lowest level of each factor to zero. Hence the parameter estimates reflect the difference in effect with respect to the lowest parameter level.

The results of the analysis are as follows. For an individual with all covariates at level 0 the fitted γ's are

$$\left(\frac{e^{.12}}{1+e^{.12}}, \frac{e^{2.8}}{1+e^{2.8}}\right) \approx (.52, .94).$$

Thus the probability of death is estimated to be $1-.94 = .06$ while the probability of sustaining a type 2 or 3 injury is $1 - .52 = .48$. The treatment effects have their usual interpretation as log odds ratios. From the analysis it can be seen that there is evidence that seat belt usage significantly lowers the probability of injury and that females are more likely to receive a higher injury score. This is in accordance with the prevailing views in the road safety literature. The effect of the variable *speedlim* is to be expected as cars travel faster when the speedlimit is higher. The effect of *speedcat* is as expected, but its interpretation is complicated by the subjective nature of the assessment that must surely be affected by the number of injuries in the accident. The effect of *age*, though non significant for the individual parameters is significant at the .05 level when tested together using the likelihood ratio ($-2 * LR = 8.662, \chi^2_{3df,.95} = 7.81$). The signs and magnitude of the age effects are in rough agreement with current beliefs.

Comparing this fit to that produced based on the binary response several features are apparent. First, the truncation structure is roughly similar with both models producing low probabilities of death. The effect of sex estimated here is much stronger than that estimated using the binary model. Though this feature could arise due to variability it may also have arisen due to the effect of sex being stronger on the probability of sustaining a Level 2 injury instead of a Level 1 injury. This can be clearly seen by remembering that the proportional odds model is a parallel series of logistic models, and that any deviation from this assumption will bias the results.

3.5 Discussion

This chapter has extended the grouped truncated model to an ordinal response, and has demonstrated its use in a plausible example.

Variable	Level	Estimate
Intercept	1	0.12(.56)
	2	2.80(.62)
Damage	1	0.16(.38)
Resavl	1	.71(.29)
Sex	1	.61(.23)
Speedlim	1	.82(.36)
Speedcat	1	1.14(.34)
Age	1	-0.25(.40)
	2	-0.19(.45)
	3	1.10(.61)
perloc	1	-.26(.26)

Table 3.2: Log odds ratio estimates for single vehicle, frontal impact collisions, using the combined 1988 and 1990 FORS files. Approximate standard errors are given in parentheses.

The problems associated with truncated models have been discussed in the previous chapter and will not be reiterated here. All that will be stated is that the ordinal model still appears to produce reliable inference regarding the individual level covariates.

Weiss (1993) has also proposed a model for group truncated data. Weiss considers grouping based on a bivariate response within an individual. He models this response by assuming that there is underlying, unobserved latent data, assumed to follow a bivariate Gaussian distribution. At this level the model is equivalent to a generalised linear model with probit link and Gaussian random effects defined at the group level and linear on the scale of the linear predictor. Weiss also allows for the variability of the underlying normal distribution to be effected by covariates, and considers truncation at the lowest level of the response.

The analysis outlined in this chapter approaches the problem from a different perspective, which extends the results of Weiss in several ways. First, it allows for arbitrary link function, which could be important in the analysis of road safety data given the interest in analysing rare events, and develops the results for the logistic case. Second, it allows for an arbitrary truncation point which is important for data that is truncated at the highest level of response, a feature of the data compiled in the so called "Fatal Files" collected by many government agencies. A third difference between the approaches is in the nature of the grouping. The model presented here allows for general patterns of groupings of individuals.

The lack of facility for the direct incorporation of correlation between responses before truncation occurs is a concern with this model, and is the main advantage of the model presented by Weiss for paired data. It is well known that parameter estimates for the mean response cease to be consistent in truncated models when the probability density is misspecified. It is hoped that the careful modelling of systematic components of the model will reduce these dependencies, but no guarantee can be given. Miller & Landis (1991) have presented methods for introducing correlation in clustered categorical data but the method is not likelihood based and does not extend naturally to truncated data, where there is correlation due to the

truncation, as well as endogenous correlation. Jansen (1990) has presented a random effects approach which utilises the full likelihood. The binary case of this will be considered in Chapter 6.

Chapter 4

Approaches to the Estimation of N

4.1 Estimation of N

4.1.1 Introduction

In this chapter we consider the estimation of the size of the truncated sample. By this we mean that we view the truncated sample $y_1, \ldots, y_n$ as being formed by the truncation of a larger, unobservable sample $y_1, \ldots, y_N$. The aim of this chapter is to develop an estimate of N.

This problem is of interest in a variety of situations. For example with road traffic data it may be of interest to estimate the number of 'potentially' fatal accidents that occur from the actual fatal accidents that are observed. This information may be useful in determining the prevalence of certain accident types. Another example of the potential use of this methodology is for the analysis of data from complicated capture-recapture studies. In this case the estimation of population size can be the primary focus of the analysis.

The frequentist estimation of N in the presence of truncation has been considered by a number of authors. Sanathanan (1972b) considered the estimation of N from data generated by a truncated multinomial distribution. This is extended to more general truncated distributions in Sanathanan (1977). Blumenthal & Marcus (1975) considered the estimation of N from a truncated sample from an exponential distribution, Blumenthal, Dahiya, & Gross (1978) considered the truncated Poisson distribution, while Blumenthal (1977) considered the estimation of N in a general univariate setting. Higher order expansions for the estimation of N were considered by Watson & Blumenthal (1980a). Other references include Blumenthal (1985) who considers the estimation of N from type II censored data, where the number of truncated observations is fixed. This differs from the approach in this thesis which assumes type I censoring, i.e. groups are truncated at random. This problem will therefore not be considered. Watson & Blumenthal (1980b) consider a two stage procedure to estimate N in the case that the proportion observed is small, and thus the variance becomes unbounded. This technique ensures the variance is bounded, but is only appropriate if the proportion of the population observed can be manipulated by the experimenter. This is not true in all truncated experiments, such as traffic accident data. It may also be uneconomic.

All of these truncated approaches have not considered the general problem of

estimating N in the presence of arbitrary covariates. The work of Blumenthal is based on the assumption of i.i.d observations and has concentrated on the development of asymptotic expansions to assess the behaviour of estimators. In particular he has considered the use of prior distributions as penalty functions to protect against divergence in parameter estimates, which occur due to extreme truncation.

The methods of Sanathanan can be directly applied to the group truncated binary case, provided there is a finite set of covariate configurations, the group size is constant and the number of groups is large. In fact the theory was specifically developed for such a case. To apply these results, the grouped binary data is expressed as multinomial data, with each response configuration corresponding to a cell in the multinomial table. The probability of each cell is easily determined from the model. In an application in visual scanning, where each observation records a set of scanners ability to detect a particle, Sanathanan (1972c) presents a model where the covariates for each group consists of the scanners used (constant across groups) and the difficulty of detecting an event which is included as a random effect.

The estimation of population size also arises in the analysis of capture-recapture experiments. This is reviewed in Seber (1982). Fienberg (1972) examines the estimation of N using contingency table methods to incorporate heterogeneity in the probabilities of capture. The use of covariates has been considered in the case of multiple capture-recapture data from closed populations by Pollock, Hines, & Nichols (1984), who considered categorical covariates, or discretised continuous covariates, and Huggins (1989), who considered a linear logistic model to model covariates in a multiple capture-recapture experiment. The use of a logistic model of the probability of capture was also considered for the single recapture case by Alho (1991). These authors do not treat the problem from the perspective of truncated data, and only consider grouped binary data modelled via the logistic link with fixed group size.

The work in this chapter differs from the available literature by considering the estimation problem in terms of general truncated regression models. General results for the estimation of N are obtained for truncated regression models which are applied to group truncated data as a special case. In addition the estimation of N for subgroups of the population is considered.

This chapter will be arranged as follows. Section 4.2 develops an estimate of N for general truncated data, where only categorical covariates are used to model the mean of the response. Section 4.3 considers the application of this result to the group truncated case, while Section 4.4 develops the results for more general covariate distributions. Section 4.5 considers the construction of approximate confidence intervals. Section 4.6 presents the results of some simulation studies designed to examine some of the finite sample properties of the estimates. Finally, Section 4.7 presents discussion and conclusions.

4.2 The estimation of N

4.2.1 Background

The predominate approach in the literature is as follows (Sanathanan, 1972b). If the total population size is N, and we consider the observed response y and the

truncated sample size n as random variables, then we can write the likelihood of the sample as

$$L(y, n; N, \beta) = L_1(n; N, \beta) L_2(y|n; N, \beta)$$

where β is a vector of parameters. In the estimation considered previously we have considered the likelihood of the sample conditional on n, and have thus maximised the conditional likelihood $L_2()$, which does not depend on N. There are two likelihood based estimators used in the literature for the estimation of N. The first is termed the conditional estimate and involves maximising $L_2(\beta)$ over β to obtain $\hat{\beta}_c$, and then maximising $L_1(N, \hat{\beta}_c)$ over N to obtain $\hat{N}_c$. This method takes advantage of the fact that the term $L_2(\beta)$ does not depend on N due to the conditioning on n. The second estimate, the unconditional estimate, is found by maximising $L()$ jointly over N and β to obtain $\hat{\beta}_u$ and $\hat{N}_u$.

When maximising over N an issue which arises is that the likelihood is defined only for integer values. This issue is discussed in a variety of sources e.g. Blumenthal (1977) and will not be repeated here. In summary, treating N as continuous adequately approximates the discrete distribution provided N is large. This is seen by noting that $\hat{N} - [\hat{N}] < 1$, where $[\hat{N}]$ is the estimate assuming N is discrete.

Sanathanan (1972b) and Blumenthal (1977) have compared the conditional and unconditional methods of estimating the parameter N when analysing truncated data. Under general conditions they compared the methods and showed their asymptotic equivalence as N tends to infinity. Noting this Sanathanan (1972b) proves that $\hat{N}_c \leq \hat{N}_u$ in finite samples. The relative advantages of the two approaches will not be considered here, and for most practical purposes differences are negligible. We will focus attention on the conditional estimate, confident that for large N there will be little difference.

4.2.2 Estimation of N with categorical covariates

For group truncated data with fixed group sizes and no covariates the estimation of N is well understood. For example, this case is equivalent to the multiple capture-recapture census (see Seber (1982)) where the estimation of N has been analysed in some detail by Darroch (1958). In addition the problem is easily dealt with by representing the observations as truncated multinomial data and applying the methods of Sanathanan (1972b), expressing the multinomial probabilities as functions of β as found from the probability of each response configuration. Fienberg (1972) has considered quite general approaches to this estimation problem by considering the estimation of incomplete 2^k contingency tables.

With covariates present, problems arise due to the variation in the probability of being observed. In this case, as estimation of the parameters is performed over the whole data set for efficiency reasons, the variation in the estimated probability of being observed must be taken into account, in addition to the truncation process producing a random number, n, of observations. We will do this by considering a more general formulation of the estimation problem and then use the results obtained to infer the behaviour in the particular situations of interest.

We consider a sequence of covariates $x_i, i = 1, \ldots, N$ and responses y_i, where x_i is defined to be the covariate set for the ith unit subject to truncation. For example if we have univariate observations y_i on individuals, and the individual is only observed if $y_i > t$ for some constant t, then the individual is the unit of truncation and x_i refers to covariates specific to individual i. Alternatively, with

group truncated data, x_i would refer to the configuration of covariates for the entire group. In this case y_i is the vector of responses for the units in the ith group. We assume that the untruncated distribution of y_i conditional on x_i is known, and is $f(y_i|x_i;\beta)$, where β is a $p \times 1$ vector of parameters. Note that the distribution of y is independent of N.

We consider a sequence of sets $\mathcal{O}_i$. If $y_i \in \mathcal{O}_i$ then it is observed, otherwise it is not. We assume that there is a finite set of covariate configurations, $\mathcal{G}$ and that elements in this set can be indexed by $k = 1, \ldots, g$. Let $x_{(k)}$ refer to the kth covariate configuration. Without loss of generality we assume that $\mathcal{O}_i$ is constant within k, meaning that all units with covariates $x_{(k)}$ have common $\mathcal{O}_i$. This set will be referred to as $\mathcal{O}_{(k)}$. This last assumption can be easily relaxed, by simply increasing g so that the set of responses in any group has the same region of truncation. We also assume that in the untruncated sample there are N_k units with covariates $x_{(k)}$ and we observe n_k of these. In addition, to derive asymptotic results, we make the assumption that

$$\frac{N_k}{N} \to c_k$$

as $N \to \infty$ where $N = \sum_{k=1}^{g} N_k$, and $c_k > 0$ for all k. This assumption is not overly restrictive, and only ensures that the average information converges in the limit. In practical terms we require the N_k to be 'large'.

We consider the likelihood for a sample of truncated observations $y_{(k)i}$, where $k = 1, \ldots, g; i = 1, \ldots, n_k$, and the index i is nested within k. Thus $y_{(k)i}$ is the ith observation with covariates $x_{(k)}$. We assume that $y_{(k)i}$ can be either a scalar or vector quantity. Conditional on the covariates in the truncated sample the likelihood is

$$L(\beta, N_1, \ldots, N_g; y) = \prod_{k=1}^{g} \binom{N_k}{n_k} P_k(\beta)^{n_k}(1-P_k(\beta))^{N_k-n_k} \prod_{i=1}^{n_k} \frac{f(y_{(k)i}|x_{(k)};\beta)}{P_k(\beta)} \quad (4.1)$$

where $P_k(\beta)$ is the probability of $y_{(k)i}$ being in $\mathcal{O}_{(k)}$. We will assume that $P_k(\beta)$ is bounded away from zero to ensure that there is a finite probability that a response with the kth covariate configuration is observed.

In this case the components $n_1, \ldots, n_g$ are independent, so we can write the likelihood as

$$L(\beta, N_1, \ldots, N_g; y, x) = \prod_{k=1}^{g} L_{1k}(\beta, N_k; y_k, x_{(k)}) L_2(\beta; y_k, x_{(k)}) \quad (4.2)$$

with

$$L_{1k}(\beta, N_k; y_k, x_{(k)}) = \binom{N_k}{n_k} P_k(\beta)^{n_k}(1 - P_k(\beta))^{N_k-n_k} \quad (4.3)$$

$$L_2(\beta; y_k, x_{(k)}) = \prod_{k=1}^{g} \prod_{i=1}^{n_k} \frac{f(y_{(k)i}|x_{(k)},\beta)}{P_k(\beta)}. \quad (4.4)$$

Note that $L_2(\beta; y_k, x_k)$ is the truncated likelihood seen previously. Thus the methods used in the usual maximum likelihood approach can be applied in this case to estimate β. In addition the independence of the n_k allows us to maximise the $L_{1k}()$ independently over the N_k. This leads to the conditional estimates,

$$\hat{\beta} = \hat{\beta}_{ML}$$

and

$$\hat{N}_k = \frac{n_k}{P_k(\hat{\beta})}$$

where $\hat{\beta}_{ML}$ is the maximum likelihood estimate.

Note that $\hat{N} = \sum_k \hat{N}_k$ and that

$$\hat{N} = \sum_k \sum_{n_k} P_k(\hat{\beta})^{-1}$$

which is attractively simple and interpretable, as $P_k(\hat{\beta})^{-1}$ is the expected number each observation 'represents' in the full sample.

4.2.3 Asymptotic Distribution of N

To investigate the asymptotic distribution of the estimates we will modify the results of Sanathanan (1972b) and Sanathanan (1977). These modifications combine elements of Theorem 3 of Sanathanan (1972b) and Theorem 1 of Sanathanan (1977), and extends and clarifies Sanathanan's results to apply to truncated regression problems with a finite number of distinct covariate configurations, where the response is dependent on the covariates. The theorem is quite general in that it provides conditions for any estimates to have the same asymptotic distribution as the maximum likelihood estimates. This generality was introduced by Sanathanan (1972b) who considered alternative estimators of N, and wished to show their asymptotic equivalence to the MLE's. For our purposes we are only interested in the application of the theorem to obtain the asymptotic distribution of the MLE's. Assumption 3 of the theorem ensures the validity of this.

Consider the Likelihood 4.1. Let $\hat{\beta}$ and $\hat{N}_k, k = 1, \ldots, g$ be estimates of β and N_k respectively. Let the partial derivatives of $logL(\beta, N_1, \ldots, N_g; y, x)$, $P_k(\beta)$, $log(f(y_{(k)i}|x_{(k)}, \beta)$ and $log(f(y_{(k)i}|x_{(k)}, \beta)/P_k(\beta))$ with respect to β_m, the mth element of β, exist and be $l^{(m)}, P_k^{(m)}, f_{ik}^{(m)}, g_{ik}^{(m)}$ if evaluated at β and $\hat{l}^{(m)}, \hat{P}_k^{(m)}, \hat{f}_{ik}^{(m)}, \hat{g}_{ik}^{(m)}$ if evaluated at $\hat{\beta}$. Note that this implies that all functions are continuous in β for all β_m. We make the standard assumptions about the derivatives of $log(f(y_{(k)i}|x_{(k)}, \beta)$ to ensure the expected value of the score is zero with finite variance.

Theorem 1 *Assume:*

1. $\hat{\beta} \to \beta$ *almost surely.*
2. $N_k^{-1/2}[\hat{N}_k - n_k/P_k(\hat{\beta})] \to 0$ *almost surely.*
3. $N^{-1/2} l^m(\hat{\beta}, \hat{N}_1, \ldots, \hat{N}_g) \to 0$ *almost surely,* $m = 1, \ldots, p$.

Then the vector $[N^{1/2}(\hat{\beta} - \beta), N^{-1/2}(\hat{N}_1 - N_1), \ldots, N^{-1/2}(\hat{N}_g - N_g)]$ *is asymptotically normal with mean zero and covariance matrix* Σ^{-1} *where*

$$\begin{aligned}
\Sigma_{lj} &= \sum_k \left[\frac{c_k P_k^{(l)} P_k^{(j)}}{P_k(1 - P_k)} + c_k P_k E(g_{ik}^{(l)} g_{ik}^{(j)}) \right] \quad l = 1, \ldots, p;\ j = 1, \ldots, p \\
\Sigma_{p+k,j} &= \frac{\sqrt{c_k} P_k^{(j)}}{1 - P_k} \quad k = 1, \ldots, g;\ j = 1, \ldots, p \\
\Sigma_{j,p+k} &= \Sigma_{p+k,j} \quad k = 1, \ldots, g;\ j = 1, \ldots, g \\
\Sigma_{p+k,p+j} &= I(k = j) \frac{P_k}{1 - P_k} \quad k = 1, \ldots, g;\ j = 1, \ldots, g
\end{aligned}$$

and $E()$ is the expectation operator, and expectation is performed over the relevant truncated distribution.

We prove this as follows. All summations in k are from 1 to g, those in i are from 1 to n_k. By appropriate addition and subtraction of terms we can write

$$N^{-1/2}\hat{l}^{(m)} = N^{-1/2}\sum_k\sum_i \hat{g}_{ik}^{(m)}$$
$$+N^{-1/2}\sum_k (n_k - N_k P_k)\frac{\hat{P}_k^{(m)}}{\hat{P}_k(1-\hat{P}_k)} + N^{-1/2}\sum_k (N_k P_k - N_k\hat{P}_k)\frac{\hat{P}_k^{(m)}}{\hat{P}_k(1-\hat{P}_k)}$$
$$+N^{-1/2}\sum_k (N_k - \hat{N}_k)\frac{\hat{P}_k^{(m)}}{1-\hat{P}_k}$$

By Taylor's theorem we can write

$$\hat{P}_k - P_k = \sum_o (\hat{\beta}_o - \beta_o)P_k^o(\beta_o^{1*}) \tag{4.5}$$

$$\sum_k\sum_i \hat{g}_{ik}^{(m)} = \sum_k\sum_i g_{ik}^{(m)} + \sum_{o=1}^{p}(\hat{\beta}_o - \beta_o)\sum_k\sum_i g_{ik}^{(m)(o)}(\beta_o^{2*}) \tag{4.6}$$

where β_o^{1*} and β_o^{2*} are on the line segment joining $\hat{\beta}_o$ and β_o , while $g_{ik}^{(m)(o)}$ is the derivative of $g_{ik}^{(m)}$ with respect to β_o.

We can then write

$$-N^{-1/2}\hat{l}^{(m)} = \sum_o N^{1/2}(\hat{\beta}_o - \beta_o)(\hat{a}_{om} - \hat{E}_{om})$$
$$+\sum_k N_k^{-1/2}(\hat{N}_k - N_k)\hat{a}_{m,p+k} - z_m \tag{4.7}$$

where

$$\hat{a}_{om} = \sum_k \frac{N_k}{N}\frac{\tilde{P}_k^{(o)}\hat{P}_k^{(m)}}{\hat{P}_k(1-\hat{P}_k)}$$
$$\hat{a}_{m,p+k} = (\frac{N_k}{N})^{1/2}\frac{\hat{P}_k^{(m)}}{1-\hat{P}_k}$$
$$\hat{E}_{om} = N^{-1}\sum_k\sum_i \tilde{g}_{ik}^{(o)(m)}$$
$$z_m = N^{-1/2}\sum_k (n_k - N_k P_k)\frac{\hat{P}_k^{(m)}}{\hat{P}_k(1-\hat{P}_k)} + N^{-1/2}\sum_k\sum_i g_{ik}^{(m)}$$

and the tilde denotes evaluation at the point β_o^{1*} or β_o^{2*} as required for equality.

Due to the assumed continuity of P, and consistency of $\hat{\beta}$,

$$\hat{a}_{om} \xrightarrow{a.s.} \sum_k c_k\frac{P_k^{(o)}P_k^{(m)}}{P_k(1-P_k)} = a_{om}$$
$$\hat{a}_{m,p+k} \xrightarrow{a.s.} \sqrt{c_k}\frac{P_k^{(m)}}{1-P_k} = a_{m,p+k}$$
$$\hat{E}_{om} \xrightarrow{a.s.} -\sum_k c_k P_k E[g_{ik}^{(o)}g_{ik}^{(m)}] = E_{om}.$$

Now by similar manipulation

$$N_k^{-1/2}\frac{(n_k-\hat{N}_k\hat{P}_k)}{1-P_k} = N_k^{-1/2}\frac{\hat{P}_k(\hat{N}_k-N_k)}{1-P_k} + N^{1/2}(\frac{N_k}{N})^{1/2}\frac{\sum_o(\hat{\beta}_o-\beta_o)P_k^{(o)}(\beta_o^{1*})}{1-P_k} - N_k^{-1/2}\frac{(n_k-N_kP_k)}{1-P_k}$$

for $k=1,\ldots,g$, *with* β_o^{1*} *as before . We can thus write*

$$N_k^{-1/2}(N_k-\hat{N}_k)\hat{a}_{p+k,p+k}+N^{1/2}\sum_o(\hat{\beta}_o-\beta_o)\hat{a}_{p+k,o}-z_{p+k}=N_k^{-1/2}\frac{(n_k-\hat{N}_k\hat{P}_k)}{1-P_k} \quad (4.8)$$

where

$$\begin{aligned}\hat{a}_{p+k,p+k} &= \frac{\hat{P}_k}{1-P_k}\\ \hat{a}_{p+k,o} &= (\frac{N_k}{N})^{1/2}\frac{\tilde{P}_k^{(o)}}{1-P_k}\\ z_{p+k} &= N_k^{-1/2}\frac{N_kP_k-n_k}{1-P_k}.\end{aligned}$$

Again, it is easily seen that

$$\begin{aligned}\hat{a}_{p+k,p+k} &\xrightarrow{a.s.} \frac{P_k}{1-P_k}=a_{p+k,p+k}\\ \hat{a}_{p+k,o} &\xrightarrow{a.s.} (c_k)^{1/2}\frac{P_k^{(o)}}{1-P_k}=a_{p+k,o}.\end{aligned}$$

Defining the vector $U=[N^{1/2}(\hat{\beta}-\beta),N_1^{-1/2}(\hat{N}_1-N_1),\ldots,N_g^{-1/2}(\hat{N}_g-N_g)]$ *and* $z=[z_1,\ldots,z_p,z_{p+1},\ldots,z_{p+g}]$ *we can combine the two sets of equations, 4.7 and 4.8 and use assumptions 2 and 3 to obtain the equation*

$$\hat{\Sigma}U-z=\mathbf{o}_p(1),$$

where

$$\hat{\Sigma}=\begin{pmatrix}\hat{a}_{11}+\hat{E}_{11} & \cdots & \hat{a}_{1p}+\hat{E}_{1p} & \hat{a}_{1,p+1} & \cdots & \hat{a}_{1,p+g}\\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ \hat{a}_{p1}+\hat{E}_{p1} & \cdots & \hat{a}_{pp}+\hat{E}_{pp} & \hat{a}_{p,p+1} & \cdots & \hat{a}_{p,p+g}\\ \hat{a}_{p+1,1} & \cdots & \hat{a}_{p+1,p} & \hat{a}_{p+1,p+1} & 0 & 0\\ \vdots & \ddots & \vdots & 0 & \ddots & 0\\ \hat{a}_{p+g,1} & \cdots & \hat{a}_{p+g,p} & 0 & 0 & \hat{a}_{p+g,p+g}\end{pmatrix}$$

and $\mathbf{o}_p(1)$ *denotes a vector with each element* $o_p(1)$.

We can then write (implicitly assuming $\hat{\Sigma}$ *is of full rank, which is assured for sufficiently large* N *if* Σ *is full rank)*

$$U-\hat{\Sigma}^{-1}z=\mathbf{o}_p(1).$$

Now z *has a limiting multivariate normal distribution, and with some calculation it is easily shown that* $z\sim AN(0,\Sigma)$. *In addition*

$$\hat{\Sigma}\xrightarrow{p}\Sigma.$$

Hence by the multivariate extension of Slutsky's Theorem, $U\sim AN(0,\Sigma^{-1})$.

4.3 The estimation of N from group truncated ordinal data

4.3.1 Group truncated estimation

We consider applying the theorem to the group truncated case. This reduces to showing that the conditions stated in the theorem are satisfied. If we consider the conditional estimation of N the following results hold. The MLE for β is strongly consistent (Serfling, 1980) and so Condition 1 is satisfied. By the definition of the conditional estimate of N, Condition 2 is satisfied. Condition 3 is seen to be satisfied by considering the following:

$$\begin{aligned} N^{-1/2}l^{(m)}(\hat{\beta},\hat{N}_k) &= N^{-1/2}\sum_k \frac{\partial}{\partial\beta_m}\log L_{1k}(\hat{\beta},\hat{N}_k) \\ N^{-1/2}l^{(m)}(\hat{\beta},\hat{N}_k) &= N^{-1/2}\sum_k P_k^{(m)}(\hat{\beta})(n_k/P_k(\hat{\beta}) - (\hat{N}_k - n_k)/(1-P_k(\hat{\beta})) \\ & \quad m = 1,\ldots,p. \end{aligned}$$

This is trivially satisfied if we consider N continuous, and is satisfied if we use discrete $\hat{N}$ as the term on the right hand side is $O(N^{-1/2})$ so it is also $o(1)$ and the condition is satisfied.

Note also that this result readily extends to group truncated ordinal model and hence viable estimation of N is possible in that case as well.

4.3.2 Finite sample considerations

In finite samples the sampling distribution of the $\hat{N}_k$ has undefined mean and variance. This is due to the existence of sample points which lead to divergent estimates of β and infinite estimates of N, as described in Chapter 2. This occurs when the likelihood is maximised by parameters that indicate extreme truncation. Blumenthal (1977) has used penalised estimates to, in part, ensure a proper distribution for N, but in the absence of any prior knowledge this appears to some extent a mathematical convenience with limited practical usefulness. Note though that the techniques of Blumenthal also serve to reduce bias in finite samples (Blumenthal, 1977), which is an important consideration.

The undefined mean and variance of the finite sampling distribution for $\hat{N}$, though important to recognise, does not cause major problems in applied work. As mentioned previously the lack of finite variance occurs in many situations. For example in logistic regression the variance of the sampling distribution is undefined in finite samples due to sample points (i.e. all responses 0) that produce divergent estimates. The sample points that produce this aberrant behaviour are extreme and are deficient in information about the parameters. In addition, as N increases the probability of observing one of these extreme samples tends to zero. Provided the asymptotic results relate adequately to the non extreme samples no serious practical bias should result, and the probability of it occuring becomes small.

A feature of the result is that the estimate of N is not consistent in the usual sense. Thus

$$\hat{N} \neq N + o_p(1).$$

Instead,

$$\hat{N} = N + O_p(N^{1/2}).$$

This is due to the estimate of N_k being based in effect on a sample of size 1. This is because n_k is a single draw from a binomial distribution. Thus information about N does not increase with N.

This feature of the estimates produces complications when setting confidence intervals. This is discussed in Section 4.5.

4.3.3 Investigation of bias

Although Theorem 1 gives the asymptotic distribution of $N_k^{-1/2}(\hat{N}_k - N_k)$ further questions remain regarding the finite sample properties of $\hat{N}_k$. In finite samples the distribution can be significantly skew, and questions regarding the bias must be answered. Note though that asymptotic biases can be calculated but finite biases are not defined due to the undefined expectation of N in finite samples.

We wish to investigate the bias of $\hat{N}$. These results, though conceptually simple, are considerably more complicated than the first order results and to state general results would entail extreme notational complexity to deal with the various configurations of group sizes and covariates available. The situation considered here is extremely simple, but the techniques used extend to more complicated cases. Unfortunately, the algebra rapidly becomes daunting. When this effort is compared to the insights the higher order expansions provide, it must be concluded that these techniques would not be considered except in exceptional cases, or if appropriate symbolic computational packages were available (see Andrews and Stafford, 1993).

The approach presented provides a more detailed analysis than the results in the previous sections and utilise the techniques of Blumenthal (1977). These methods are based on the usual local approximations to the likelihood produced using Taylor series expansions. An equivalent approach is to use implicit differentiation to derive the required derivatives from the score equations. We will assume that the group size is constant, i.e. that the group size is G and that there are no covariates.

We adopt a new notation to the previous sections.

We wish to write our estimate $\hat{N}$ as

$$\hat{N} = N + a\sqrt{N} + b + O(N^{-1/2}) \tag{4.9}$$

where a and b are to be determined. Let n_g denote the number of groups observed with g positive responses, $g = 1, \ldots, G$, and let p be the probability of a positive response, $q = 1 - p$. The score equation 2.2.1 is

$$\sum_{g=1}^{G} n_g \left(g - \frac{G\hat{p}}{\hat{P}} \right) = 0 \tag{4.10}$$

where $\hat{P} = 1 - \prod_k (1 - \hat{p})$ is the estimated probability of being observed, $Q = 1 - P$, and $\sum_{g=1}^{G} g n_g$ is the number of positive responses in the truncated sample. We wish to form a function that implicitly defines $\hat{N}$ in terms of the observed sample quantities n_g.

Now

$$\hat{N} = \frac{n}{\hat{P}}$$

and therefore

$$\hat{p} = 1 - \sqrt[G]{1 - n/\hat{N}}$$

so we can re-write Equation 4.10 as

$$\sum_{g=1}^{G} g n_g = GV[1 - (1 - n/V)^{1/G}] \tag{4.11}$$

which defines an equation in V with solution $\hat{N}$. If we let $Y_N = \sqrt{GN}(\bar{Y} - p)$ and $T_N = (n - NP)/\sqrt{N}$ where $\bar{Y} = \sum_{g=1}^{G} g n_g / GN$ we have

$$\sqrt{N}\sqrt{G}Y_N + NGp = GV\left[1 - \sqrt[G]{1 - \frac{\sqrt{N}T_N + NP}{V}}\right].$$

Now substituting $V = N + a\sqrt{N} + b + O(N^{-1/2})$, we can solve for a and b as the equality will be maintained for large N, as the error is $O(N^{-1/2})$. We use the Taylor expansion for $(1+x)^{1/g}/(1+y)^{1/g}$ about $x = y = 0$, and some tedious algebra yields

$$\begin{aligned} NGp + \sqrt{N}\sqrt{G}Y_N &= NGp - \sqrt{N}\left[\frac{q(aP - T_N)}{Q} - aGp\right] \\ &\quad -\frac{q}{2GQ^2}\left[2bGPQ + (1-G)(a - T_N)^2 + (1+G)a^2Q^2\right. \\ &\quad \left. -2aQ(a - T_N) + 2aGQ(aP - T_N) - 2bG^2Q^2p/q\right] + O(N^{-1/2}). \end{aligned}$$

Equating coefficients yields

$$\begin{aligned} a &= \frac{\sqrt{G}QY_N - qT_N}{GQp - Pq} \\ b &= \frac{(1-G)q(aP - T_N)^2}{2GQ(GQp - Pq)}. \end{aligned}$$

Now T_N and Y_N are random variables. The vector $[Y_N, T_N]$ for fixed N has expected value zero and covariance matrix

$$\Sigma = \begin{pmatrix} pq & \sqrt{G}p(1-P) \\ \sqrt{G}p(1-P) & P(1-P) \end{pmatrix}.$$

We can write $[Y_N, T_N] = 1/\sqrt{GN}\sum_j U_j$ where $U_j = [\sum_i Y_{ij} - Gp, \sqrt{G}(I_j - P)]$ where i indexes the individuals within a group, j indexes the group, I_j is an indicator for the group being observed and Y_{ij} is the response for the ith individual in the jth group. Thus as $N \to \infty$, $[Y_N, T_N]$ is asymptotically Gaussian with mean zero and variance Σ. Therefore as a and b are functions of $[Y_N, T_N]$ we can determine their asymptotic distribution. Thus for large N

$$\begin{aligned} a &\sim N(0, \frac{GQ^2pq + PQq^2 - 2GQ^2pq}{(GQp - Pq)^2}) \\ b &\sim \frac{G(G-1)QPp\chi_1^2}{2(GQp - Pq)^2} \end{aligned}$$

where χ_1^2 is a Chi squared distribution with 1 degree of freedom.

Thus the approximate asymptotic bias is the expectation of b which is

$$\frac{G(G-1)QPp}{2(GQp-Pq)^2}.$$

The device of using the representation in Equation 4.9 with the additional representation

$$\hat{\beta} = \beta + \frac{a'}{\sqrt{N}} + \frac{b'}{N}$$

is well documented in Blumenthal (1977) and its extension to group truncated data with categorical covariates is conceptually simple. As mentioned previously the algebra is heavy going.

4.4 Estimation of N with continuous covariates

4.4.1 Introduction

The results shown so far have all been derived by assuming that there is a finite set of distinct covariate values. With continuous covariates, writing the likelihood in the form of Equation 4.2 results in g increasing with N, and N_k equal to 1 for all k. Thus the asymptotic results developed previously do not hold. This is restrictive as there are plausible situations where continuous covariates are available. In addition the biases that are introduced by omitting them are extremely serious due to the use of our model to essentially extrapolate to the unobserved class. Of course the device of discretizing over the range of the continuous covariate could be applied but this leads to the introduction of biases into the estimation scheme.

In this section an approach will be demonstrated which provides valid inference in the presence of continuous covariates.

4.4.2 Continuous covariates

Consider continuous covariates, X, with density $f(x)$, defined over some set $\mathcal{A}$. We allow the covariate distribution to contain mixtures of potentially different densities. For example with group truncated data, we assume that the covariate distribution is a mixture of the covariate density for each group size. Assume that we observe a truncated response y, with associated covariates x, and that the conditional distribution of y given x is $g(y|x;\beta)$. If we assume the population is generated by drawing a sequence of covariates from $f(x)$, and observations from $g(y|x;\beta)$, we can write the marginal(over the unobserved covariates) likelihood of the observed data as

$$L(\beta, N; y, x) = L_1(\beta, N; y, x, n) L_2(\beta; y, x, n)$$

with

$$\begin{aligned} L_1(\beta, N; y, x, n) &= \binom{N}{n} P(\beta)^n (1 - P(\beta))^{N-n} \\ L_2(\beta; y, x, n) &= \prod_n \frac{g(y_i|x_i, \beta) f(x_i)}{P(\beta)} \\ P(\beta) &= \int_{\mathcal{A}} P(\beta, x) f(x) dx \end{aligned}$$

where the integration is defined so as to sensibly evaluate any discrete components of the distribution of X, and $P(\beta, x)$ is again the probability that the unit was observed, which we assume is a continuous function of β. We assume that

$$\int_{\mathcal{A}} \frac{1}{P(\beta, x)} f(x) dx$$

is convergent. For practical analysis this is not extremely restrictive. For instance if we assume group truncated binary data, and the probabilities are modelled by a linear logistic model and the covariates have a Gaussian distribution then the assumption is satisfied. This assumption ensures that units with $P(\beta, x) \to 0$ do not exist in the population with high probability. If this is not true the problem is not well defined and an alternative analysis should be considered. We assume that $P(\beta)$ is a continuous function of β, which will be assured by the continuity of $P(\beta, x)$.

The distribution of the observed covariates is

$$f_T(x) = \frac{f(x)P(\beta, x)}{P(\beta)}$$

so if we condition on the observed covariates the likelihood becomes

$$L^*(\beta, N; y, x) = L_1(\beta, N; y, x, n) L_2^*(\beta; y, x, n)$$

with

$$L_2^*(\beta; y, x, n) = \prod_n \frac{g(y_i|x_i, \beta)}{P(\beta, x)}.$$

We note that $L_2^*(\beta; y, x, n)$ is the usual truncated likelihood which does not depend on n and is maximised by the maximum likelihood estimate $\hat{\beta}$. We assume that this estimate is consistent as N tends to infinity.

We use the conditional estimate of N,

$$\hat{N} = \frac{n}{P(\hat{\beta})}.$$

We can state the following theorem.

Let P denote $P(\beta)$ and $P^{(i)}$ denote the derivative of P with respect to the ith parameter evaluated at the true parameter point and $g^{(i)}$ denote the derivative of the truncated density function with respect to the ith parameter.

Theorem 2 *Assume we have estimates $\hat{\beta}$ and $\hat{N}$ and that:*

1. $\hat{\beta} \xrightarrow{a.s} \beta$.
2. $N^{-1/2}(\hat{N} - n/P(\hat{\beta})) \xrightarrow{a.s} 0$.
3. $N^{-1/2} l^{(m)}(\hat{\beta}, \hat{N}) \xrightarrow{a.s} 0$, $m = 1, \ldots, p$.

Then the vector $[N^{1/2}(\hat{\beta} - \beta), N^{-1/2}(\hat{N} - N)]$ is asymptotically normal with mean zero and symmetric covariance matrix Σ^{-1} given by

$$\begin{aligned}
\Sigma_{lj} &= \frac{P^{(l)} P^{(j)}}{P(1-P)} + E_x[P(\beta, x) E_y(g^{(l)}(y_i|x_i, \beta) g^{(j)}(y_i|x_i, \beta))] \\
&\quad l = 1, \ldots, p;\ j = 1, \ldots, p \\
\Sigma_{p+1,j} &= \frac{P^{(j)}}{1-P} \quad j = 1, \ldots, p \\
\Sigma_{p+1,p+1} &= \frac{P}{1-P}
\end{aligned} \tag{4.12}$$

where E_x is performed over the covariate distribution and E_y is performed with respect to the truncated distribution of y.

Proof: The proof is similar to that given for Theorem 1 and is omitted.

Note that $E_x[P(\beta,x)E_y(g^{(i)}(y_i|x_i,\beta)g^{(j)}(y_i|x_i,\beta))]$ can be calculated either explicitly or consistently estimated by $\hat{N}^{-1}\mathcal{I}$ where $\mathcal{I}$ is the observed information.

This result is useful if the distribution of the covariates is considered known. In practice, cases may arise where the distribution is unknown. The form of the distribution of X, $f(x)$ is clearly a nuisance parameter in this case, so we propose using the conditional likelihood found from L_2^* to estimate β as before. To estimate N it is not possible to find $P(\hat{\beta})$ as the distribution of X is unknown and thus no 'plug in' estimate of $P(\hat{\beta})$ exists. Instead we construct an estimate as follows.

$$\hat{P}(\hat{\beta}) = \frac{N^{-1}\sum_i I(i\text{th observed})}{N^{-1}\sum_i I(i\text{th observed})P(\hat{\beta},x_i)^{-1}} \tag{4.13}$$

where summation is over the full sample, and $I(i\text{th observed})$ is the indicator function for the ith unit being observed. This is the expected value of $P(\hat{\beta},x)$ calculated over the empirical distribution of the observed covariates, with weight

$$w_j = \frac{P(\hat{\beta},x_j)^{-1}}{\sum_l P(\hat{\beta},x_l)^{-1}}$$

at each of the observed points. In this case the weights are defined to adjust for the varying probabilities of observation of the observed covariates.

The derivatives of $P(\beta)$ can be approximated by the consistent estimate

$$\hat{P}^{(m)}(\hat{\beta}) = \frac{N^{-1}\sum_i I(i\text{th observed})P^{(m)}(\hat{\beta},x_i)P(\hat{\beta},x_i)^{-1}}{N^{-1}\sum_i I(i\text{th observed})P(\hat{\beta},x_i)^{-1}}. \tag{4.14}$$

We propose using as our estimate of N,

$$\hat{N} = \frac{n}{\hat{P}(\hat{\beta})}.$$

Note that this reduces to

$$\hat{N} = \sum_l \frac{1}{P(\hat{\beta},x_l)}$$

where $\sum_l$ is the sum over the observed covariates. This is the estimate that would have been obtained if we had naively used the estimator for discrete covariates. This result uses the fact that we can consistently estimate the probability of being observed in a particular sample, by extrapolation based on the parametric model. It can also be viewed by considering that an observed value effectively 'represents' observations that are unobserved but close to this observation in terms of the probability of observation.

To see that the marginal probability of being observed is consistently estimated, consider the following.

$$\frac{1}{N}\sum_i I(i\text{th observed}) \xrightarrow{a.s.} \int P(\beta,x)f(x)dx$$

and

$$\frac{1}{N}\sum_i \frac{I(i\text{th observed})}{P(\beta, x_i)} \xrightarrow{a.s.} 1$$

due to the weak law of large numbers. This is applicable because of the assumptions regarding the summands, which ensure finite mean and variance. Therefore

$$\hat{P}(\beta) \xrightarrow{p} \int P(\beta, x) f(x) dx = P(\beta)$$

by the weak law of large numbers.

Now for fixed $\hat{P}()$, the continuity of $\hat{P}()$ and the strong consistency of the MLE of β ensures

$$\hat{P}(\hat{\beta}) = \hat{P}(\beta) + o_p(1).$$

Therefore,

$$\hat{P}(\hat{\beta}) - P(\beta) = o_p(1)$$

and thus the marginal probability of being observed is consistently estimated by $\hat{P}(\hat{\beta})$.

To determine the asymptotic variance it is tempting to apply Theorem 2 to the present case, as we have consistent estimates of all the required constants, but problems arise. Note that conditions 2 and 3 cease to hold. For example

$$N^{-1/2}(\hat{N} - n/P(\hat{\beta})) = N^{-1/2}\frac{n(P(\hat{\beta}) - \hat{P}(\hat{\beta}))}{P(\hat{\beta})\hat{P}(\hat{\beta})}$$

which has a non degenerate asymptotic distribution. The reason for this problem is that the estimate produced if the covariate distribution were known, $\tilde{N}$, say, and $\hat{N}$ are not asymptotically equivalent. We can write

$$N^{-1/2}(\hat{N} - N) \;=\; N^{-1/2}(\tilde{N} - N) + N^{-1/2}(\hat{N} - \tilde{N})$$

which shows that the variability of the estimate $\hat{N}$ would be underestimated by the results of Theorem 2 which does not take account of the variability in the estimate of $P()$, $\hat{P}()$.

Now

$$\begin{aligned} N^{-1/2}(\hat{N} - \tilde{N}) &= N^{-1/2}\sum_i \frac{I(i\text{th observed})}{P(\hat{\beta}, x_i)} - \frac{I(i\text{th observed})}{P(\hat{\beta})} \\ &= N^{-1/2}\sum_i \frac{I(i\text{th observed})}{P(\beta, x_i)} - \frac{I(i\text{th observed})}{P(\beta)} \\ &+ N^{1/2}A(\hat{\beta} - \beta) + o_p(1) \end{aligned} \tag{4.15}$$

where

$$A = N^{-1}D_\beta\left[\sum_i \frac{I(i\text{th observed})}{P(\beta, x_i)} - \frac{I(i\text{th observed})}{P(\beta)}\right]$$

with D_β the vector differential operator, and A is evaluated at β. The pth element of $N^{-1}A$ is

$$-N^{-1}\sum_i \frac{I(i\text{th observed})P^{(p)}(\beta, x_i)}{P(\beta, x_i)^2} - \frac{I(i\text{th observed})P^{(p)}(\beta)}{P(\beta)^2}.$$

This converges almost surely to

$$\frac{\int P^{(p)}(\beta,x)f(x)dx}{\int P(\beta,x)f(x)dx} - \int \frac{P^{(p)}(\beta,x_i)}{P(\beta,x)}f(x)dx$$

which is consistently estimated by

$$\frac{\hat{P}^{(p)}(\hat{\beta})}{\hat{P}(\hat{\beta})} - \sum_j w_j \frac{P^{(p)}(\beta,x_i)}{P(\beta,x)}.$$

The terms in the summation in Equation 4.15 are random variables with expected value

$$E\left[\frac{I(i\text{th observed})}{P(\beta,x_i)} - \frac{I(i\text{th observed})}{P(\beta)}\right] = 0$$

and variance

$$var\left[\frac{I(i\text{th observed})}{P(\beta,x_i)} - \frac{I(i\text{th observed})}{P(\beta)}\right] = \int \frac{1}{P(\beta,x)}f(x)dx - \frac{1}{P(\beta)} = V(\beta,f(x)).$$

Hence provided the variance is finite, which is assured by our earlier assumptions, a central limit theorem applies, and

$$N^{-1/2}(\hat{N}-\tilde{N})$$

is asymptotically normal with mean zero and variance $V(\beta,f(x)) + A\Sigma_\beta^{-1}A'$, where Σ_β^{-1} is the variance matrix of $\sqrt{N}(\hat{\beta}-\beta)$ found from Theorem 2. $V(\beta,f(x))$ can be consistently estimated by

$$\hat{V}(\hat{\beta},f(x)) = \sum_j w_j \frac{1}{P(\hat{\beta},x_j)} - \frac{1}{\hat{P}(\hat{\beta})}.$$

Note that the variance is zero when $f(x)$ is degenerate, as expected, because in this case the estimate collapses to the no covariate case. Also, the estimate $\hat{V}(\beta,f(x))$ will always be positive due to Jensen's inequality.

It remains to determine the asymptotic covariance of $N^{-1/2}(\hat{N}-\tilde{N})$ and $N^{-1/2}(\tilde{N}-N)$. Noting that

$$E\left[\sum_i\sum_j \frac{I(i\text{th observed})I(j\text{th observed})}{P(\beta)P(\beta,x_i)} - \frac{I(i\text{th observed})I(j\text{th observed})}{P(\beta)^2}\right] = 0,$$

due to the marginal independence of the indicator functions, the covariance of $N^{-1/2}(\hat{N}-\tilde{N})$ and $N^{-1/2}\tilde{N}$ is $A\Sigma_{p+1,j}^{-1}$, where $\Sigma_{p+1,j}^{-1}$ is the column vector of covariances found from Theorem 2. Thus $N^{-1/2}(\hat{N}-N)$ is asymptotically normal with mean zero and variance

$$V(\beta,f(x)) + A\Sigma_\beta^{-1}A' + V_{\tilde{N}} + 2A\Sigma_{p+1,j}^{-1} \tag{4.16}$$

where $V_{\tilde{N}}$ is the asymptotic variance of $N^{-1/2}(\tilde{N}-N)$ found from Theorem 2.

4.4.3 Estimation of the Covariate Distribution

In the univariate case the empirical distribution of the observed covariates converges uniformly in x to the true truncated distribution, $f_T(x)$ (Serfling (1980)). This is of little use if we wish to investigate the untruncated distribution, which as stated before may be the only distribution that is readily interpretable. We propose the estimate

$$\hat{F}(x) = \frac{N^{-1}\sum_i I(x_i \leq x \text{ and observed}) P(\hat{\beta}, x_i)^{-1}}{N^{-1}\sum_i I(i\text{th observed}) P(\hat{\beta}, x_i)^{-1}}$$

where $F(x)$ is the distribution function of X. Due to the consistency of $\hat{\beta}$ and the assumptions on the behaviour of $f(x)$ we can show that for any x, $\hat{F}(x) \to F(x)$ almost surely, provided the correct probability model is specified for the response y.

4.5 Confidence intervals for N

We consider the problem of setting approximate confidence intervals for N using the results of Theorem 1. The results for continuous covariates follow easily.

4.5.1 g=1

4.5.1.1 Method 1

If $g = 1$ and thus there is a single component for N we can set confidence intervals in the following manner. From Σ^{-1} we can extract the $(p+1, p+1)$ element which is the asymptotic variance of

$$\frac{N - \hat{N}}{\sqrt{N}}.$$

Call this element d. We can thus write for large N that

$$Pr(z_{\alpha/2} < \frac{N - \hat{N}}{\sqrt{dN}} < -z_{\alpha/2}) \approx 1 - \alpha$$

where $z_{\alpha/2}$ is the $\alpha/2$th percentile of the standard normal distribution. We can then consider inverting this relationship to find a confidence interval for N. This is complicated due to the appearance of N in the denominator. After some manipulation the approximate $1 - \alpha$ confidence interval is found to be

$$\left[\hat{N} + z^2_{\alpha/2}d/2 - \sqrt{z^2_{\alpha/2}\hat{N}d + z^4_{\alpha/2}d^2/4}, \hat{N} + z^2_{\alpha/2}d/2 + \sqrt{z^2_{\alpha/2}\hat{N}d + z^4_{\alpha/2}d^2/4}\right]. \tag{4.17}$$

Note that this interval is symmetric about the point $\hat{N} + z^2_{\alpha/2}d$ where $z^2_{\alpha/2}d$ is a fixed constant that does not increase with N.

4.5.1.2 Method 2

An alternative CI can be constructed as follows. As $\hat{N}$ is approximately normally distributed for large N, we can consider setting a $1 - \alpha$ confidence interval as

$$\left[\hat{N} - z_{\alpha/2}\sqrt{dN}, \hat{N} + z_{\alpha/2}\sqrt{dN}\right]. \tag{4.18}$$

Unfortunately , N is unknown. We consider substituting $\hat{N}$ in this case. The effect of this is complicated by the weaker consistency of $\hat{N}$. Heuristically we can justify this choice by noting that as $\hat{N} = N + O_p(N^{1/2})$

$$\frac{\sqrt{\hat{N}}}{\sqrt{N}} = \frac{\sqrt{N^{1/2}\left(\sqrt{N^{1/2}+B}\right)}}{\sqrt{N}} \sim 1 \tag{4.19}$$

as B is bounded in probability for all N. We note that this interval is asymptotically equivalent to the interval 4.17, where equivalence in this case is defined as the ratio of the endpoints tending to 1. Thus the asymptotic coverage of the intervals is the same.

4.5.2 g> 1

We now consider setting confidence intervals for some quantity $H = \sum I_k N_k$ where I_k is one if the kth population is included in the total and zero otherwise. We could attempt to extend method 1 to this case, but this is complicated by the number of parameters. To perform this inversion is equivalent to setting a joint CI for the individual N_k's, and mapping this to H. This would be complicated to perform. Instead we consider a simpler approach based on the interval 4.18, and argue that the equivalence of the two procedures for large N extends from the univariate case.

If we define the vector $V = [0_p, I_1\sqrt{N_1}, \ldots, I_g\sqrt{N_g}]$, where 0_p is a p-vector of zeros then, for large N, by Theorem 1 $V'U$ is approximately normally distributed with mean zero and variance $V'\Sigma^{-1}V$.

We can therefore set approximate $1-\alpha$ confidence intervals as $[\hat{H}-z_{\alpha/2}V'\Sigma^{-1}V, \hat{H}+z_{\alpha/2}V'\Sigma^{-1}V]$. To do this we must find an estimate of $V'\Sigma^{-1}V$. We can consistently estimate c_k by

$$c_k = \frac{\hat{N}_k}{\sum_k \hat{N}_k}$$

and E_{jm} by the elements of

$$\hat{N}^{-1}\mathcal{I}(\hat{\beta})$$

where $\mathcal{I}(\hat{\beta})$ is the information matrix found from the conditional estimation of β. In addition $P_k(\hat{\beta})$ and $P_k^{(m)}(\hat{\beta})$ consistently estimate $P_k(\beta)$ and $P_k^{(m)}(\beta)$.

As mentioned before a consistent estimate of V is not available. Instead we will use $\hat{V} = [0_p, I_1\sqrt{\hat{N}_1}, \ldots, I_g\sqrt{\hat{N}_g}]$.

4.5.3 Comments

A problem with these intervals is that they do not enforce the range restriction that $N_k \geq n_k$. Although this presents no theoretical problem, provided the coverage is accurate, its appearance in applied work is unfortunate. It could potentially be avoided by setting confidence intervals for

$$a = \frac{n}{N}$$

and then inverting the resulting intervals. Thus provided that the confidence interval for a was was contained in $(0,1)$ the range restriction would be preserved. Unfortunately this is not easily obtained. It is not sufficient to set a CI for the

marginal proportion truncated, P, which is easily obtained via the delta method. The CI constructed for this quantity would have asymptotically the correct coverage, but the coverage of the inverted interval is not guaranteed as for fixed n, n/P does not necessarily equal N. Hence even if P is in the CI, the true N need not be in the inverted interval.

An obvious candidate for the estimation of the variability of the estimate is the bootstrap. This is attractive because it can hopefully take account of the skewness of the sampling distribution. To implement the bootstrap, we need to define at least approximately independent quantities to resample from. In this case the n_k are in effect a sample of size of 1. Thus resampling from them, though simple, would produce spurious results. If the sampling process was repeated, so that we obtained a series of n_k, for fixed N_k, then we could use a bootstrapping algorithm, but in this case the dynamics of the estimation problem have changed, with N_k being a fixed parameter and consistent estimates now available. Thus the bootstrap does not appear a panacea in this particular problem.

Huggins (1989) has suggested the use of the conditional bootstrap, based on the distribution of the response conditional on the truncated sample size and the observed covariates, to estimate the variability of $\hat{N}$. Its use in this case is incorrect, and the procedure only calibrates variability in the estimate due to the distribution of $\hat{\beta}$. It ignores the variation due to the random nature of n. This can be seen by noting that if $\hat{\beta}$ is consistent for β the parametric bootstrap distribution of $\hat{\beta}$ for large n is degenerate. Thus the conditional coverage tends to 0, or 1 depending on the n observed, and the unconditional coverage tends to 0. Any approximate coverage results for small samples are spurious and have no theoretical foundation. The use of the bootstrap to estimate the variability of the parameter estimates is explored in Appendix 1.

4.6 Examples

4.6.1 Approximate CI's

The first issue to consider is the finite sample behaviour of the confidence intervals set via the methods discussed in Section 4.5. We consider an analogous situation. In this case we draw $\hat{N}$ from a distribution that is Normal with mean N and variance Nd, for some constant d. In this example we consider d known. This is done to focus attention on the use of $\hat{N}$ in the estimation of variance. In applied situations we would require a consistent estimate, $\hat{d}$ of d.

The confidence intervals calculated by method 1 will have exact coverage, by definition. In the case of method 2 we set $1-\alpha/2$ confidence intervals via

$$[\hat{N} - z_{\alpha/2}\sqrt{\hat{N}d}, \hat{N} + z_{\alpha/2}\sqrt{\hat{N}d}].$$

We can calculate the coverage, p, of these intervals as

$$p = \Phi[(N_l - N)/\sqrt{Nd}] + 1 - \Phi[(N_u - N)/\sqrt{Nd}]$$

where Φ is the distribution function for the standard normal distribution and N_l, and N_u are the lower and upper bounds and are found from the solutions of

$$N_l + z_{\alpha/2}\sqrt{dN_l} - N = 0$$

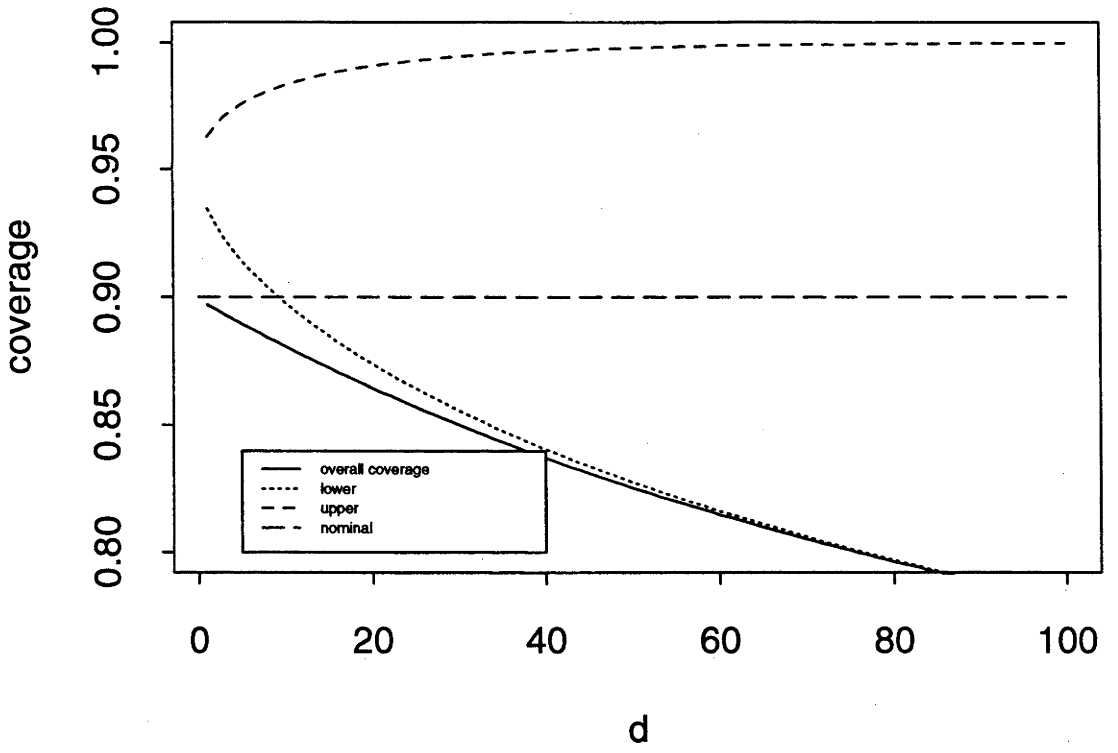


Figure 4.1: Theoretical coverage for $N = 100$. Upper and lower refer to $\Phi[(N_l - N)/\sqrt{Nd}]$ and $1 - \Phi[(N_u - N)/\sqrt{Nd}]$ respectively.

and

$$N_u - z_{\alpha/2}\sqrt{dN_u} - N = 0.$$

Figures 4.1 to 4.3 graph the coverage for various levels of d and increasing N.

The figures display two obvious features. First, the bias in coverage increases with d. This is as expected given that as d increases the variability of N increases, and the errors in coverage are amplified. Second, the bias of the coverage goes to zero as N goes to infinity, due to the standard error depending on $\sqrt{N}$. Thus for large N the coverage should be accurate, but the determination of large N is dependent on d.

Of course the coverage of a CI is only one measure of its performance. The mean width of the CI is another important feature. Given the likelihood theory used in its construction the asymptotic performance of the proposed CI should be good. The behaviour in finite samples will now be investigated.

4.6.2 Simulation Study

4.6.2.1 Categorical covariates

In this section we present a small simulation study to demonstrate the techniques presented in this chapter. We first consider a simple case where there are three distinct groups. The probability of a positive response is assumed to depend on a single covariate, with the probability being modelled by a logistic model as described in Chapter 2, including an intercept term. The covariate is assumed to have two levels, 0 and 1. We have N_1 groups of size 2 with covariate levels 1 and 0, N_2 groups of size 2 with covariate levels 1 and 1, and N_3 groups of size 3 with

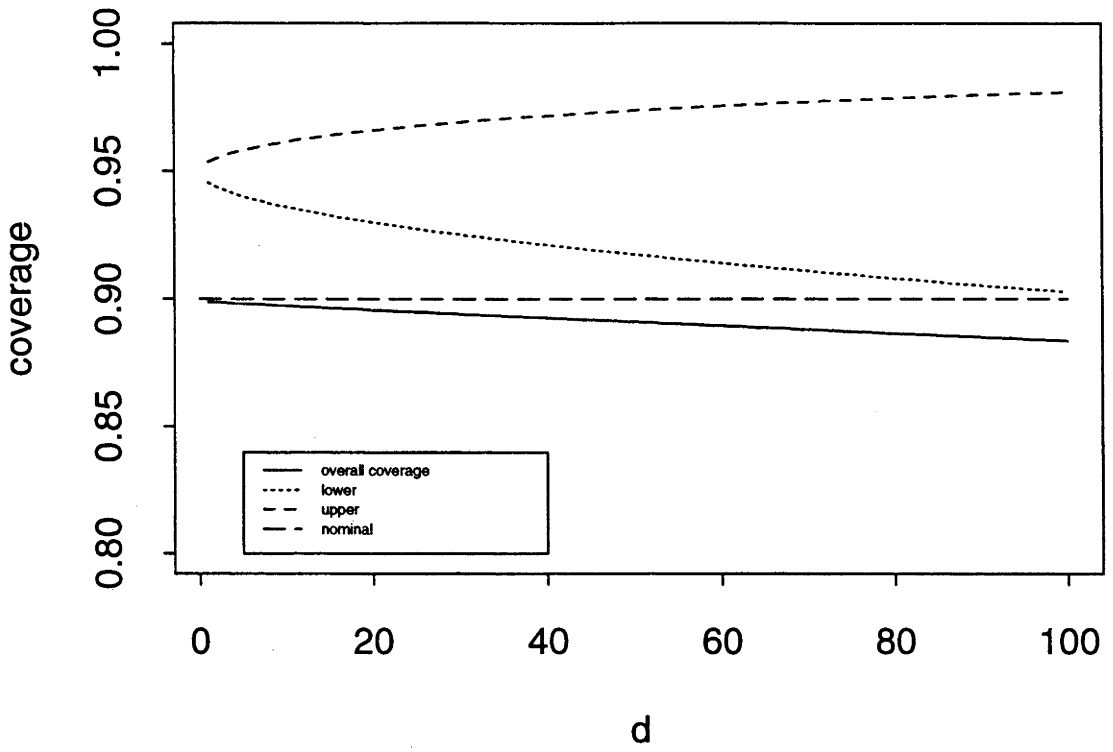


Figure 4.2: Theoretical coverage for $N = 1200$. Upper and lower refer to $\Phi[(N_l - N)/\sqrt{Nd}]$ and $1 - \Phi[(N_u - N)/\sqrt{Nd}]$ respectively.

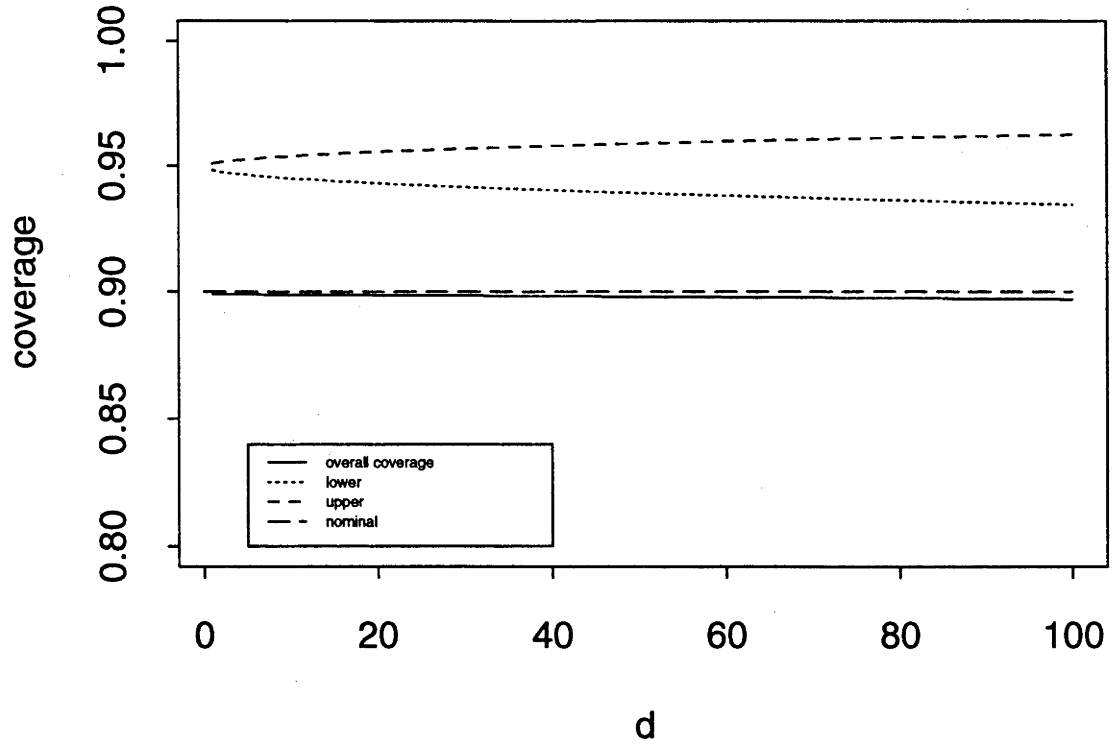


Figure 4.3: Theoretical coverage for $N = 100000$. Upper and lower refer to $\Phi[(N_l - N)/\sqrt{Nd}]$ and $1 - \Phi[(N_u - N)/\sqrt{Nd}]$ respectively.

covariates 1,1 and 0. We consider the behaviour of the estimates of N over a range of parameter values and population sizes.

The proportion $N1 : N2 : N3$ was kept constant over the simulation at 2.5 : 2.5 : 1 and the simulation was performed over the grid $\beta_1 = [-4, -2, 0]$ for the intercept term and $\beta_2 = [.5, 2]$ for the covariate. These values were chosen to give a range of situations, from extreme truncation to mild truncation, and from a small treatment effect to a large treatment effect. The probabilities of being observed are given in table 4.1. The population size was taken to be 600 and 1200, and for each population size the covariates were kept fixed. The lower population size was chosen to ensure that a reasonable number of truncated observations remained, and that divergent estimates were not routinely encountered. The samples chosen give a range of values for d, the asymptotic variance of the scaled estimate, of 500 to .5. The simulation was performed as follows:

1. For given covariate values the logistic model

$$logit(p) = \beta_1 + \beta_2 x$$

   was used to determine p and a sample was generated.

2. The sample was group truncated.

3. The parameters were estimated via the conditional maximum likelihood approach.

The above steps were iterated 500 times for each set of parameters and population size. Non convergence could occur due to sparse data sets. For each set of parameter values the mean and variance of the estimates was calculated, along with the estimated coverage of the approximate 90% confidence intervals for N and N_3. To calculate the estimated variance the methods outlined in Section 4.2.3 and Section 4.4 were used. The results from these simulations are presented in Tables 4.2 to 4.5. In these tables $\bar{\hat{N}}$ and $VAR(\hat{N})$ are the mean and variance respectively of the estimates over the simulation. $\widehat{VAR(\hat{N})}_1$ and $\widehat{VAR(\hat{N})}_2$ are the means of the variance estimates calculated using the results of Theorem 1 and Section 4.4 respectively. Coverage N_{v1}, and coverage N_{v2} refer to the coverage of the confidence intervals calculated using the variance estimate from Theorem 1 and Equation 4.16 respectively. Numbers in parentheses refer to the actual number of estimates used to produced the results, which is only of interest in the cases were divergent estimates were encountered. As can be seen this occured when there was extreme truncation.

4.6.2.2 Continuous covariates

The second simulation presented involves continuous covariates. The design has the same number of groups as before but in this case the population consists of pairs to the desired population size, and the covariates are now drawn from a uniform distribution over $(0, 1)$. In this case estimation via Theorem 1 is not feasible and only the results outlined in Section 4.4 are used. The results from these simulations are presented in Tables 4.6 to 4.9. The behaviour of the procedure is further considered in Figures 4.4 to 4.9, which are the simulations with the treatment effect, β_2 equal to .5. The figures consist of four components. The upper two

β_1	β_2	$P_{(1,0)}$	$P_{(1,1)}$	$P_{(1,1,0)}$	$\bar{P}$
-4	.5	.046	.057	.074	.056
	2	.135	.224	.238	.189
-2	.5	.279	.331	.361	.301
	2	.559	.750	.779	.675
0	.5	.811	.857	.928	.850
	2	.940	.985	.992	.968

Table 4.1: Marginal probabilities of being observed for example 1. P_u is the probability of being observed for a group with covariates u. $\bar{P}$ is the marginal probability over the groups.

β_1	β_2	$\bar{\hat{N}}$	$VAR(\hat{N})$	$\widehat{VAR(\hat{N})}_1$	$\widehat{VAR(\hat{N})}_2$
-4	.5	315(201)	1423(201)	99401(201)	98941(201)
	2	700	149323	223130	221337
-2	.5	616	13286	14736	14575
	2	601	1028	1068	1010
0	.5	599	212	223	219
	2	599	26	27	27

Table 4.2: Results of simulation with $N = 600$ and categorical covariates. See text for definitions.

β_1	β_2	coverage N_{v1}	coverage N_{v2}	coverage N_3
-4	.5	.66(201)	.66(201)	.82(201)
	2	.88	.88	.95
-2	.5	.904	.898	.97
	2	.904	.888	.992
0	.5	.906	.906	.988
	2	.906	.906	.970

Table 4.3: Estimated coverage for $N = 600$ and categorical covariates. See text for definitions.

β_1	β_2	$\bar{\hat{N}}$	$VAR(\hat{N})$	$\widehat{VAR(\hat{N})}_1$	$\widehat{VAR(\hat{N})}_2$
-4	.5	961(318)	178790(318)	947714(318)	945514(318)
	2	1308	189774	199433	196991
-2	.5	1219	24414	26823	26520
	2	1203	2083	2119	2004
0	.5	1200	424	450	443
	2	1200	55	56	55

Table 4.4: Results of simulation with $N = 1200$ and categorical covariates. See text for definitions.

β_1	β_2	coverage N_{v1}	coverage N_{v2}	coverage N_3
-4	.5	.76(318)	.76(318)	.910(318)
	2	.916	.914	.962
-2	.5	.926	.922	.980
	2	.896	.886	.994
0	.5	.918	.910	.988
	2	.902	.902	.980

Table 4.5: Estimated coverage for $N = 1200$ and categorical covariates. See text for definitions.

β_1	β_2	$\bar{\hat{N}}$	$VAR(\hat{N})$	$\widehat{VAR(\hat{N})}_2$
-4	.5	203(131)	8183	4214
	2	685(411)	130725	486736
-2	.5	639	32267	33369
	2	603	3802	3899
0	.5	600	385	372
	2	600	97	91

Table 4.6: Results of simulation with $N = 600$ and continuous covariates. See text for definitions.

panels display the sampling distribution of $\hat{N}$. The lower two panels display the sampling distribution of the confidence intervals, allowing the behaviour of the CI's to be examined. In these panels the CI's over the simulation are ordered by the lower bound found using the approximations in Theorem 1. The lower bound, upper bound and n are then plotted with respect to this ordering. This plot allows the variation in the width of the CI's to be assessed.

4.6.2.3 Comments

The results of the simulations exhibit several features. First, for reasonably large N, the coverage appears accurate for the overall population total. Although this

β_1	β_2	coverage N_{v2}	coverage N_{v2} inv.
-4	.5	.29(131)	.66(131)
	2	.84(411)	.901(411)
-2	.5	.91	.892
	2	.902	.908
0	.5	.872	.868
	2	.882	.878

Table 4.7: Estimated coverage for $N = 600$ and continuous covariates. See text for definitions.

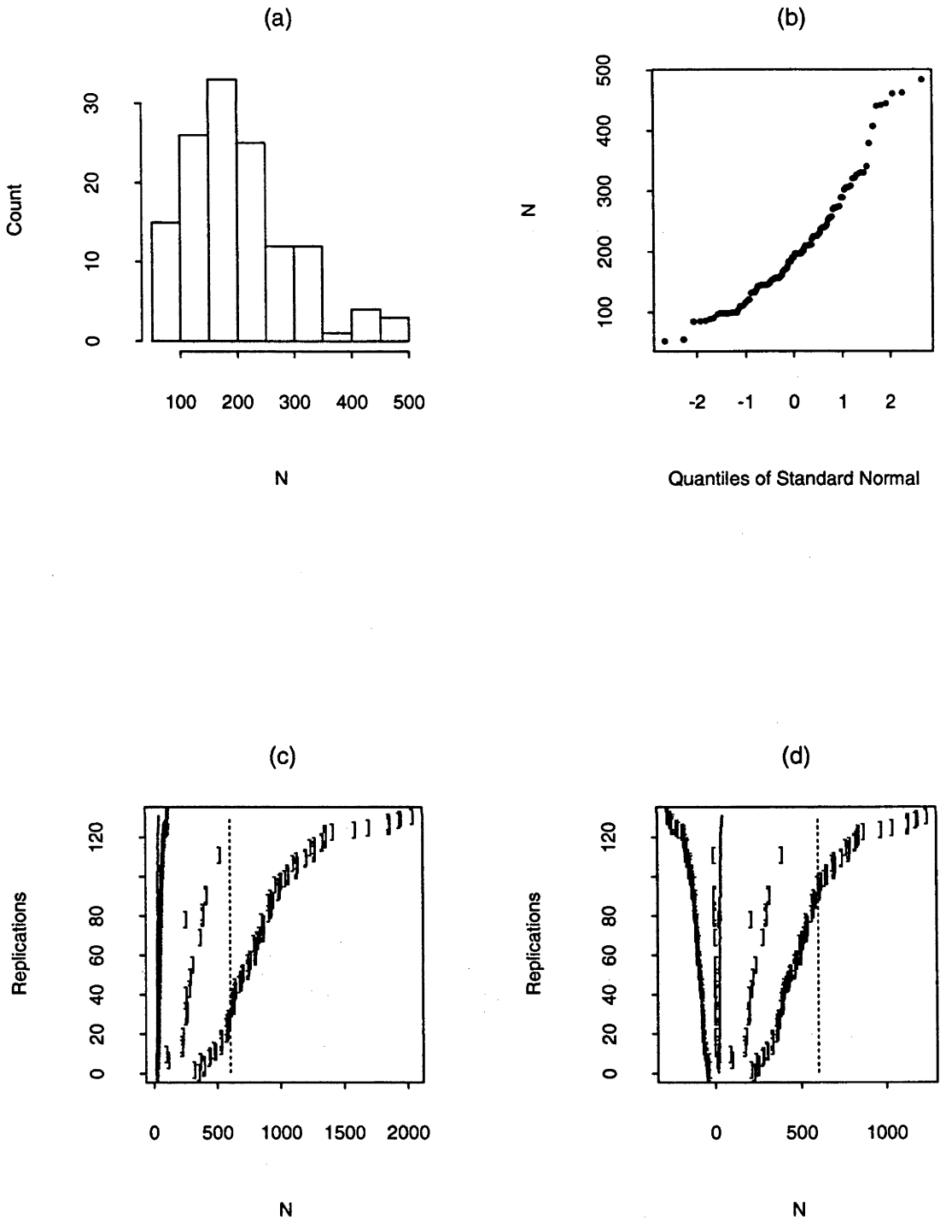


Figure 4.4: Sampling distribution of $\hat{N}$ with $\beta_1 = -4$, $\beta_2 = .5$ and $N = 600$. (a)=sampling distribution, (b)=qq-plot of sampling distribution, (c)=sampling distribution of CI set using method 2, (d)=sampling distribution of CI set using method 1. "["=lower point of CI, "]"=upper point of CI, "."=sample size.

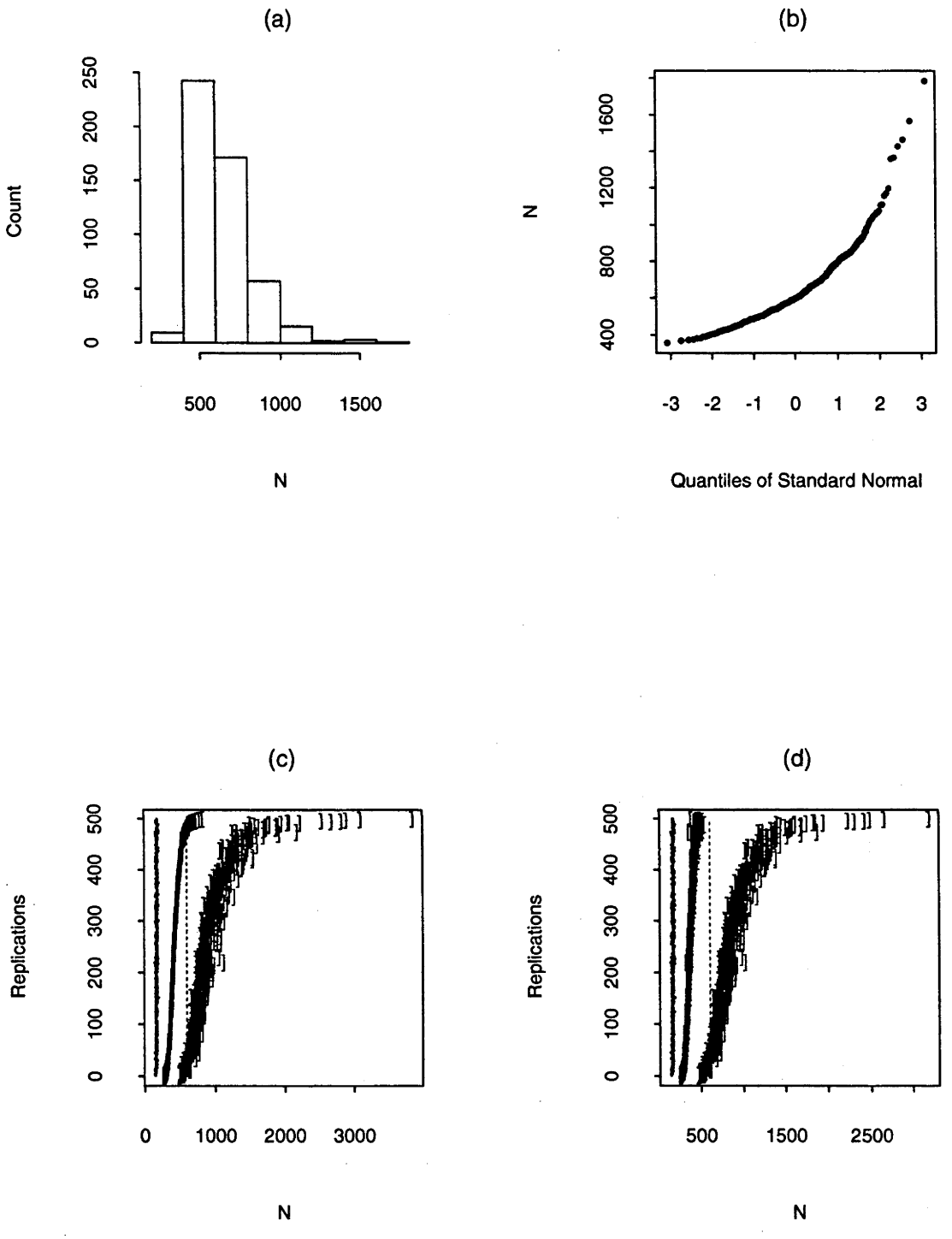


Figure 4.5: Sampling distribution of $\hat{N}$ with $\beta_1 = -2$, $\beta_2 = .5$ and $N = 600$. (a)=sampling distribution, (b)=qq-plot of sampling distribution, (c)=sampling distribution of CI set using method 2, (d)=sampling distribution of CI set using method 1. "["=lower point of CI, "]"=upper point of CI, "."=sample size.

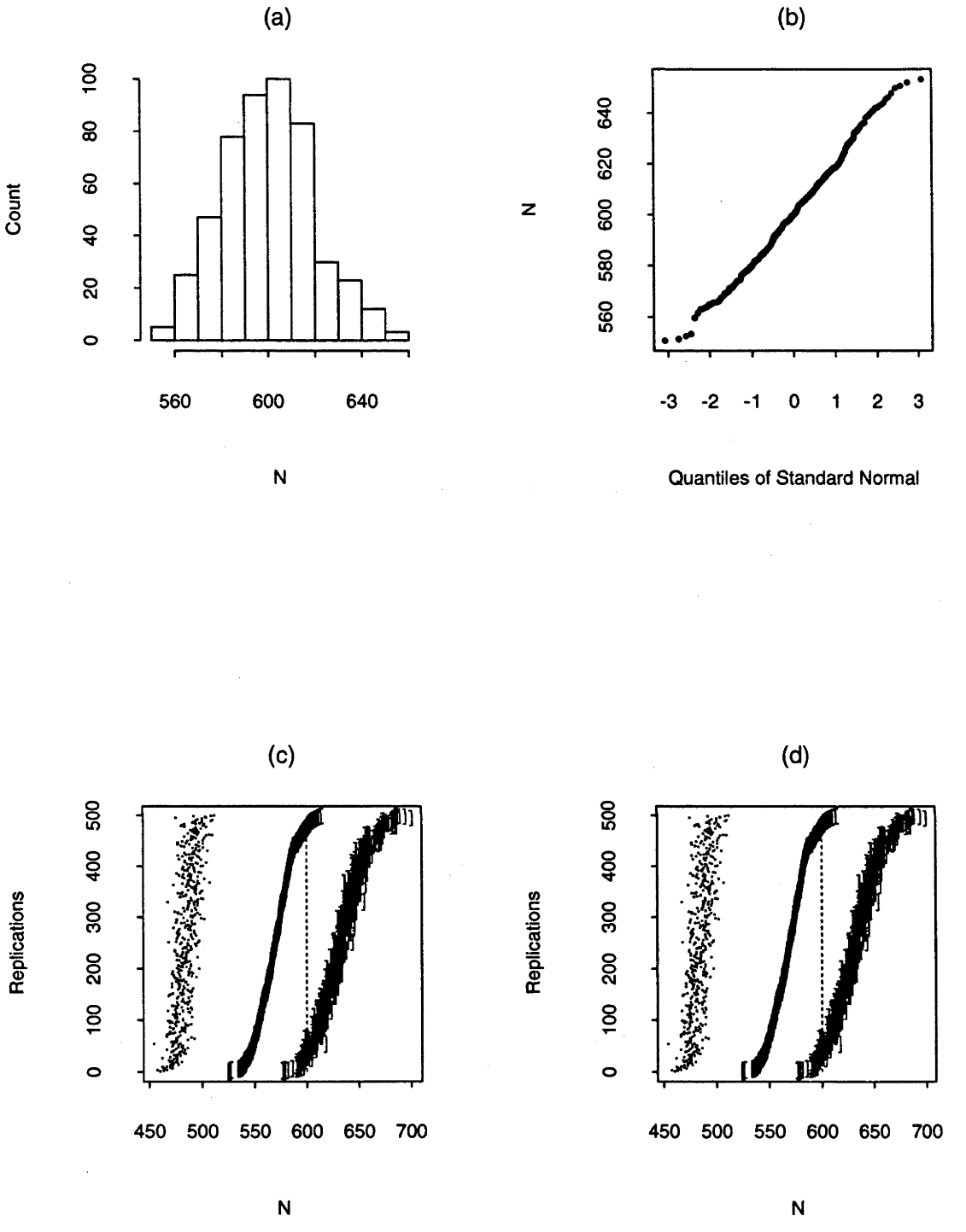


Figure 4.6: Sampling distribution of $\hat{N}$ with $\beta_1 = 0$, $\beta_2 = .5$ and $N = 600$. (a)=sampling distribution, (b)=qq-plot of sampling distribution, (c)=sampling distribution of CI set using method 2, (d)=sampling distribution of CI set using method 1."["=lower point of CI, "]"=upper point of CI, "."=sample size.

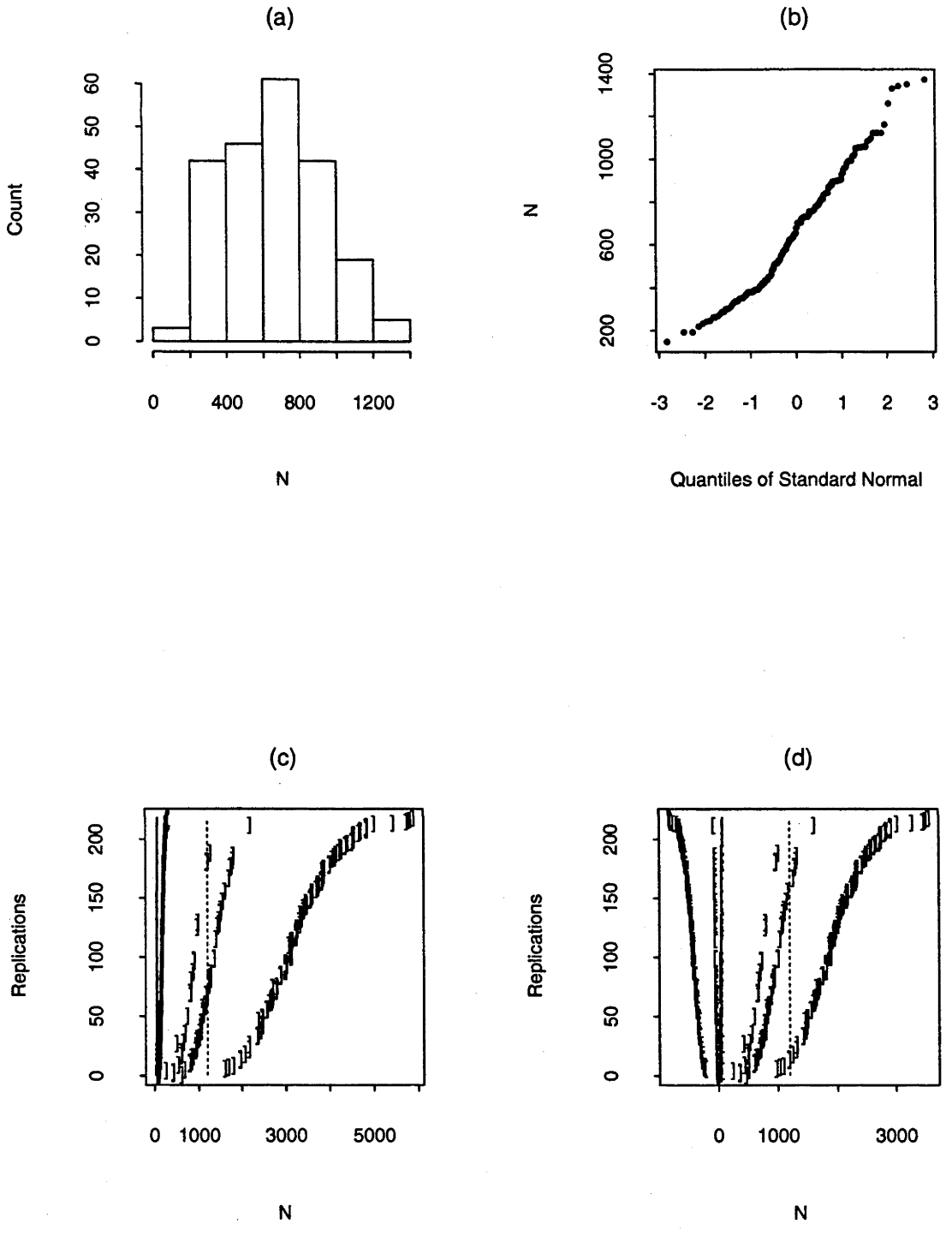


Figure 4.7: Sampling distribution of $\hat{N}$ with $\beta_1 = -4$, $\beta_2 = .5$ and $N = 1200$. (a)=sampling distribution, (b)=qq-plot of sampling distribution, (c)=sampling distribution of CI set using method 2, (d)=sampling distribution of CI set using method 1. "["=lower point of CI, "]"=upper point of CI, "."=sample size.

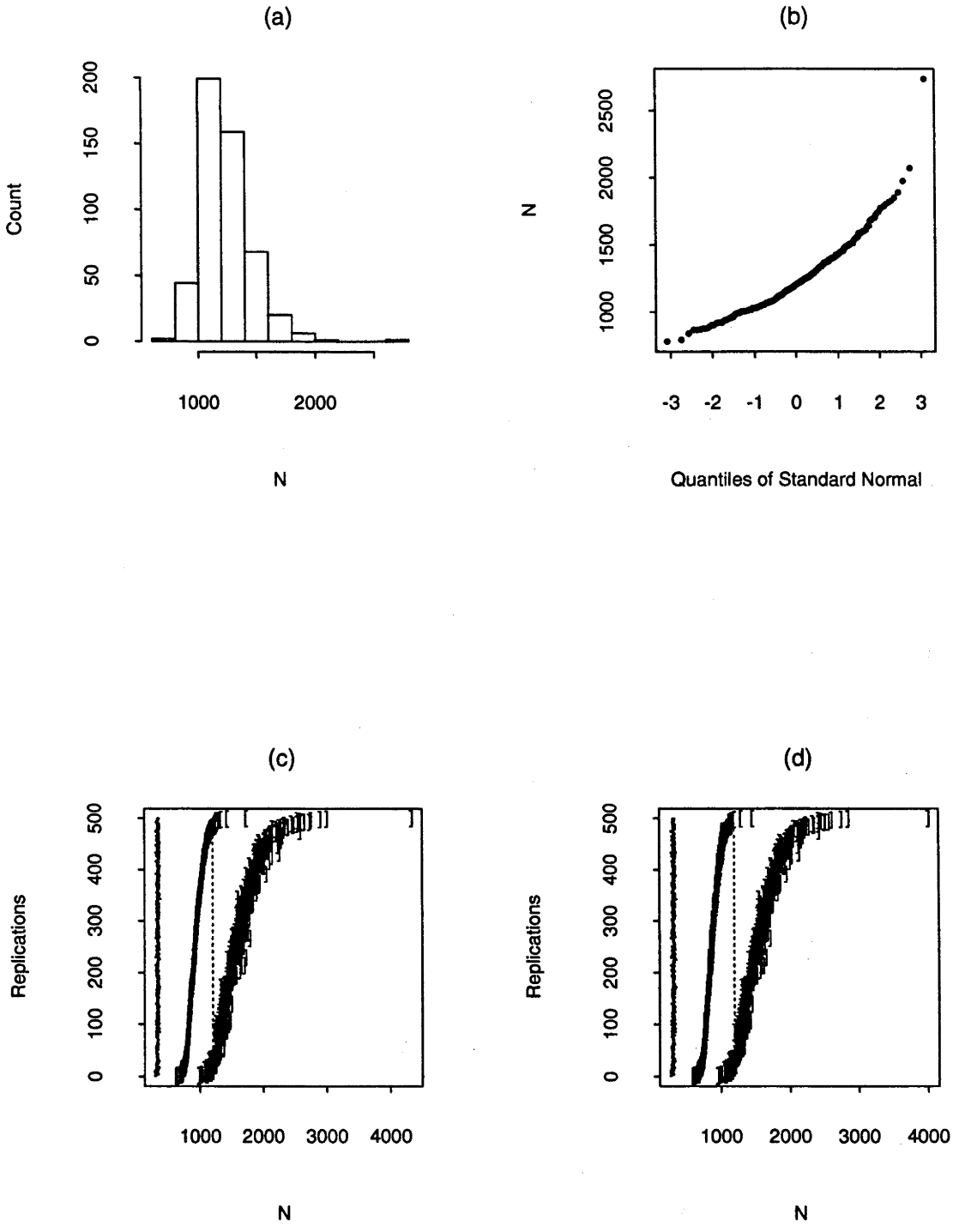


Figure 4.8: Sampling distribution of $\hat{N}$ with $\beta_1 = -2$, $\beta_2 = .5$ and $N = 1200$. (a)=sampling distribution, (b)=qq-plot of sampling distribution, (c)=sampling distribution of CI set using method 2, (d)=sampling distribution of CI set using method 1."["=lower point of CI, "]"=upper point of CI, "."=sample size.

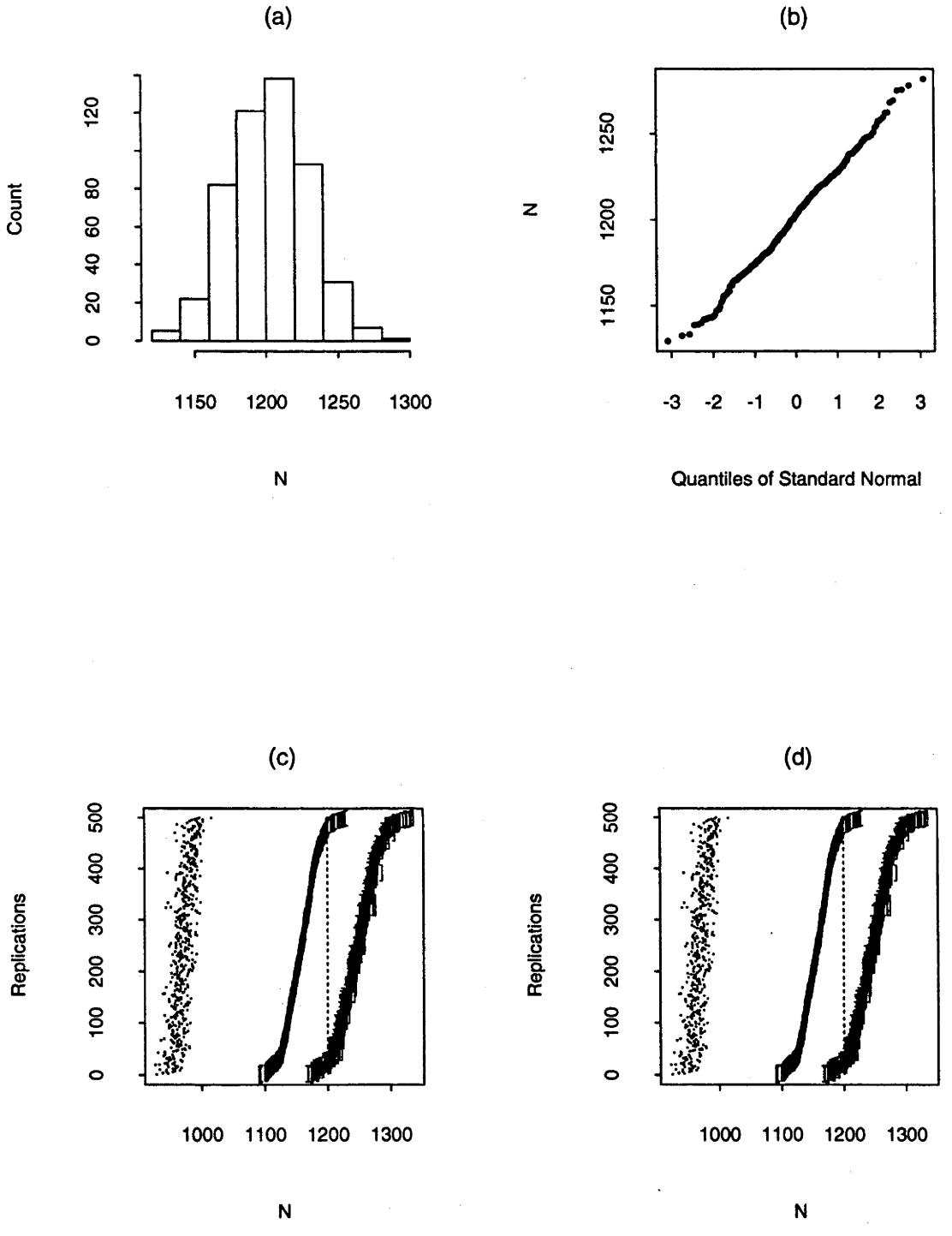


Figure 4.9: Sampling distribution of $\hat{N}$ with $\beta_1 = 0$, $\beta_2 = .5$ and $N = 1200$. (a)=sampling distribution, (b)=qq-plot of sampling distribution, (c)=sampling distribution of CI set using method 2, (d)=sampling distribution of CI set using method 1."["=lower point of CI, "]"=upper point of CI, "."=sample size.

β_1	β_2	$\bar{\hat{N}}$	$VAR(\hat{N})$	$\widehat{VAR(\hat{N})}_2$
−4	.5	665(218)	68890	44230
	2	1501	798085	1.8×10^6
−2	.5	1231	47749	48190
	2	1207	8358	7492
0	.5	1202	738	739
	2	1201	201	183

Table 4.8: Results of simulation with $N = 1200$ and continuous covariates. See text for definitions.

β_1	β_2	coverage N_{v2}	coverage N_{v2} inv
−4	.5	.64(218)	.757(218)
	2	.938(485)	1
−2	.5	.926	.912
	2	.868	.868
0	.5	.918	.914
	2	.882	.882

Table 4.9: Estimated coverage for $N = 1200$ and continuous covariates. See text for definitions.

appears to be at odds with the results presented in Section 4.6.1, which implied that we could reasonably expect significant undercoverage, it can potentially be explained by the positive skewness of the sampling distribution. Given this skewness, sampled points in the right tail produce wider confidence intervals, thus inflating the coverage. This of course is no theoretical justification for the observed result, and it can only be described as fortuitous. The coverage of the confidence intervals for N_3 is not as accurate, tending to over cover the true value, which is at least conservative. The cause of this effective overestimate of variability is not entirely obvious. A potential candidate is the small sample size in this case leading to the asymptotics being not as useful an approximation.

With regards to the choice of method 1 or method 2 for the construction of confidence intervals little can be determined. For small n the coverage of the method 1 interval is closer to the nominal level, but this result is somewhat spurious due to the increased coverage being caused by an increased proportion of very wide intervals caused by the high variability in the estimates of Σ. The intervals are constructed as if Σ is known, which is reasonable as N increases, but is obviously inappropriate with small samples. The method 1 intervals perform extremely poorly in this case. The effect can be seen by referring to Figure 4.4.

Note that the coverage with small sample sizes is poor. This is brought about by the skewed and biased distribution of the convergent estimates. In addition, in these cases the value of d defined in Section 4.6.1 is large, and significant undercoverage

results. Note that the magnitude of d can be assessed by examining

$$\frac{\hat{VAR}(\hat{N})_1}{N}.$$

Note also that the two variance estimates are approximately equal as the sample size increases and the level of truncation decreases. This is expected as they are asymptotically equivalent, and equal to zero, when the probability of being observed is 1.

The coverage in the situations with extreme truncation and hence small samples is not unexpected. In these cases a significant proportion of the data sets have no information regarding N except that it is greater than n. In this case there is very little that can be done to improve the procedure. The only real option is to impose additional information on the problem, by using a prior distribution to obtain either a penalised likelihood estimate or a modal Bayes estimate (depending on one's taste and perspective). This obviously has the problem of removing the objectivity inherent in the frequentist approach, but there must surely be times when this is appropriate.

The figures reinforce these points. First, when the level of truncation is extreme the methods perform poorly. Examining Figures 4.4 and 4.7 we see that the behaviour of the confidence intervals is poor, the lower bound often being below the sample size n. In this case, without any extra information, it is hard to think of an approach that will give practical confidence intervals for N. The data is too sparse to draw strong conclusions from. The approximations in this case are inappropriate. Note that the strange patterns in the upper bound of the CI's is due to the categorical nature of the data and the small sample sizes and the negative bias in Figure 4.4 is due to the unusual nature of the convergent estimates. The situation does not appear to be improving with N in this case. It would appear that N would need to be very large in this case, for the approximations to be adequate.

Examining Figures 4.5 and 4.8 we see that the methods work better in these cases. Note that when $N = 600$ the only intervals that do not contain N are those fully below the true value. This effect decays as the sample size increases. Figures 4.5 and 4.8 appear wholly satisfactory. In this case the sampling distribution is approximately normal. The lower bounds remain away from n and the length of the intervals remains fairly constant over the samples.

4.6.3 Road Safety Example

In this example we investigate the road safety data presented in Section 2.6. We estimate the number of potentially fatal accidents implied by the fatal accidents observed. In this case the problem is ill posed. This is due to the removal of accidents that contained missing data. Hence the estimate of N is conditional on the complete data accidents. In addition the interpretation of the number of unobserved accidents estimated is problematic. How does one define an accident that was potentially fatal? In this case we view the analysis in an exploratory way. The estimate of N produced is that which is most consistent with the data and the assumed model. Thus it gives insight into the pattern of the data that is not available by simple perusal of the raw figures, and allows evaluation of the amount of implied truncation.

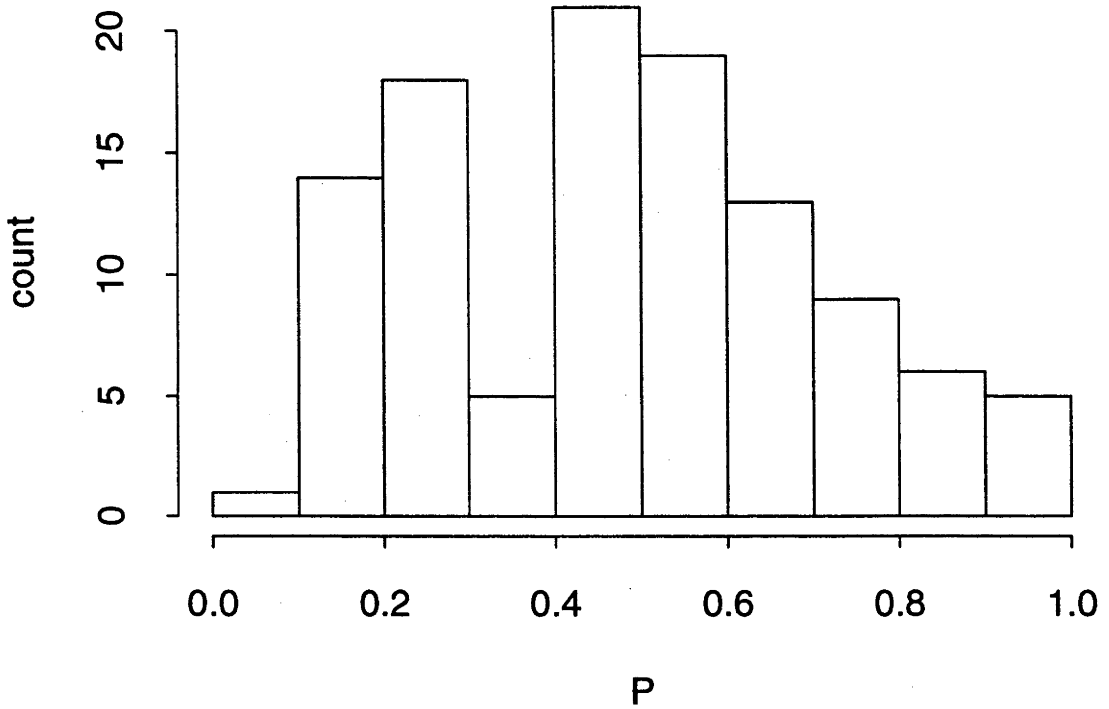


Figure 4.10: Histogram of $P(\hat{\beta}, x_i)$ for 88/90 frontal impact collisions

Recall that the data consisted of records from 111 single vehicle accidents involving 306 individuals. The estimate of N from this data is

$$\hat{N} = 323$$

with the estimated standard error of the estimate being 58.97. This leads to an estimated 90% confidence interval for N of [226,420]. The observed distribution of $P(\hat{\beta}, x_i)$ is given in figure 4.10.

From the fit of the ordinal model presented in Chapter 3 we obtain an estimate of

$$\hat{N} = 284$$

with an estimated standard error of 43.76. This leads to an estimated 90% confidence interval for N of [213,356]. The interpretation of the rough agreement between the estimates must be modified by the knowledge that they are essentially the same fit to the same data. The reduced standard error of the second estimate can be explained by the higher marginal probability of being observed estimated by the ordinal model (which is manifested in the smaller estimate of N), and by the increased accuracy (possibly spurious) of the treatment parameter estimates derived from the additional information in the ordinal model.

4.7 Discussion and Conclusions

In this chapter the frequentist estimation of N has been discussed. The results of Sanathanan (1972b) have been extended to deal with the general truncated regression case with either categorical or continuous covariates. Results have been developed which allow the estimation of N in a variety of regression situations, and allows for the estimation of N for various sub classes of the population. A credible simulation study has demonstrated the efficacy of these techniques. Under extreme truncation problems arise but this is the result of these samples containing little information about the parameters. The problem of making robust inference with little information is an enduring one in frequentist inference. Except in specialised cases the problem remains largely unresolved.

If it is possible to control the intensity of truncation the results suggest that truncation rates greater than than around 80% are undesirable, and should thus be avoided. This result of course depends on the sample sizes and covariate configurations. If the intensity of truncation is beyond the control of the experimenter, then the obvious way to proceed in the small sample/extreme truncation case is to use a penalised approach. This is easily incorporated by following Blumenthal (1977). For credible penalty functions the asymptotic results considered here can be easily modified.

The results in this chapter allow us to examine the joint distribution of β and N. In certain cases the distribution of the covariates in the untruncated population may be of interest. In Section 4.4 we have alluded to an approach to this problem using the empirical distribution to approximate the covariate distribution. In the next chapter we investigate an alternative Bayesian approach which extends this idea.

Chapter 5

Bayesian Analysis of Truncated Regression Models

5.1 Introduction

This chapter examines the Bayesian regression analysis of data in the presence of truncation. The example considered will be truncated Poisson data. An example of such data is the records of fishing trips presented in Grogger & Carson (1991). This data set is based on a sample of households and records the number of fishing trips undertaken as well as a range of covariates. Inclusion in the sample is dependent on at least one fishing trip being undertaken. This is an example of truncation where the response of the unobserved members is known, in this case zero. This can be contrasted with truncated Gaussian regression, where the value of the unobserved responses is unknown, only that observations were observed outside a certain level.

The analysis in this chapter differs from that presented in the previous chapters due to its use of Bayesian methods. Bayesian methods are considered for several reasons. First, the usual arguments for the use of Bayesian methods apply here. For example they allow the introduction of prior information into the analysis. As has been previously discussed, this can be an important consideration with truncated data, with divergent estimates occuring with high probability under extreme truncation. Of course, the usual arguments against Bayesian statistics also apply.

The second reason for the interest in applying Bayesian arguments to this problem is that different components of interest can be treated in a symmetric fashion. In a truncated regression analysis with observationally truncated data interest can centre on the distribution of the response given the covariates, the distribution of the covariates, and the size of the unobserved sample.

The use of Bayesian methods in such a complicated situation as this would have been computationally restrictive in previous years. The use of data augmentation(Tanner & Wong, 1987), the development of Markov Chain Monte Carlo (MCMC) methods and the advancement in computing power has lead to the feasibility of the approach, and much progress and research has gone into the use of these techniques in missing data models such as censored regression and grouped data. In addition the use of data augmentation has opened up avenues of analysis for traditionally difficult models, for example the random effects models discussed in Zeger & Karim (1991).

The developments in this chapter will concentrate on the regression analysis of Poisson data. All methods presented can be potentially extended to more gen-

eral truncated problems, provided the truncated likelihood can be appropriately defined. Section 5.2 develops the likelihood for the truncated model. In Section 5.3 the priors are discussed, posteriors are derived, and a Gibbs sampling algorithm to explore them is presented. Section 5.4 considers possible alternative augmentations, Section 5.5 presents some numerical examples and the method is discussed in Section 5.6.

5.2 Bayesian Model

We will begin by considering the likelihood of obtaining a particular truncated sample. Consider a sample of size N from the untruncated distribution. For clarity we will use the notation found in Gelfand, Smith, & Lee (1992) for the densities. This notation denotes the density of a random variable K as $[K]$ instead of the usual functional notation $f(k)$.

As before we consider a truncated sample of observations Y. Assume that there exist covariates x which are realisations of the random variables X with density $[X]$, independent of the response, and that $[Y|X]$ is known. Consider the sample with n untruncated observations. We assume that the response of the unobserved members is known to be a certain value, denoted by zero. This covers cases such as group truncation of the zero class, and the Poisson distribution truncated at zero. This restrictive assumption can be easily extended and this will be discussed later. By permutation if necessary we consider the likelihood of

$$(Y_1, X_1), \ldots, (Y_n, X_n), (0, X_{n+1}), \ldots, (0, X_N)$$

where n is the sample size of the truncated sample. The likelihood of the complete sample is then

$$[Y, X, n|\beta, N] = \binom{N}{n} \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i] \prod_{i=n+1}^{N} [Y = 0|X_i, \beta][X_i].$$

Given a prior $[\beta]$ the posterior is thus

$$[\beta|Y, X] \propto [Y, X|\beta][\beta].$$

This is of no immediate use, as the covariate values for the truncated observations are unknown. To avoid this problem we calculate the marginal distribution of n and $(Y_1, X_1) \ldots (Y_n, X_n)$. This is the likelihood of the observed data and is

$$[Y, X^{(n)}, n|\beta, N] = \binom{N}{n} \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i][\beta] \int \cdots \int \prod_{i=n+1}^{N} [Y = 0|X_i, \beta][X_i] \prod_{i=n+1}^{N} dx_i,$$

where $X^{(n)}$ is the set of covariates associated with the observed response Y. As the distribution of the X_i's are identical this reduces to

$$[Y, X^{(n)}, n|\beta, N] = \binom{N}{n} \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i][\beta]\{1 - P(\beta)\}^{N-n} \tag{5.1}$$

where

$$P(\beta) = 1 - \int [Y = 0|X, \beta][X] dx, \tag{5.2}$$

the unconditional probability of observing a unit randomly chosen from $[X]$.

If N and $[X]$ are known Equation 5.1 could be used to form inferences about β. Alternatively if suitable priors for N and $[X]$ can be produced then the analysis can proceed. The use of MCMC methods may be of use in this case depending on the specification of $[X]$. Unfortunately, many statisticians would have serious reservations about specifying a prior for $[X]$ which is possibly multidimensional. This is due to the difficulty in defining a prior that is both flexible enough and tractable, when no obvious parametric information is available. To avoid this problem we will consider an alternative approach.

With truncated data, it is not usual for N to be known. In this case we assume that N is a random variable and postulate a distribution for N. We could proceed by specifying a prior for N directly, but for added flexibility we will consider the case where N has a distribution depending on another parameter. We will denote this by $[N|\lambda]$ where λ is some parameter. The joint distribution of $[Y, X^{(n)}, n, N|\beta, \lambda]$ is then

$$[Y, X^{(n)}, n, N|\beta, \lambda] = \binom{N}{n} \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i]\{1 - P(\beta)\}^{N-n}[N|\lambda] \tag{5.3}$$

and the marginal for $[Y, X^{(n)}, n|\beta, \lambda]$ is

$$\sum_{j=n}^{\infty} \binom{j}{n} \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i]\{1 - P(\beta)\}^{j-n} Pr(N = j|\lambda). \tag{5.4}$$

The choice of $[N|\lambda]$ will depend on the particular problem being considered. For example if it is assumed that N has a Poisson distribution with mean λ then Equation 5.4 simplifies to

$$[Y, X^{(n)}, n|\beta, \lambda] \propto \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i]\lambda^n e^{-\lambda P(\beta)}.$$

The final issue is the specification of the distribution of the X's. It is possibly multi-dimensional and with no knowledge of the distribution, the specification of a form that is flexible enough is difficult. Arguing that the observed covariates provide the only information regarding the distribution of X we propose to approximate the covariate density by using the 'empirical' support of the observed covariates. By this we mean that we shall let the support of the distribution be defined by the observed covariates. We propose using a multinomial distribution over this empirical support. We use the parameter η in this case to represent the vector of length n with components the cell probabilities of the multinomial distribution. This is loosely analogous to the use of the multinomial simplification to derive empirical likelihoods from the intractable nonparametric likelihoods (see Efron & Tibshirani (1993)).

The posterior is then

$$[Y, X^{(n)}, n|\beta, \lambda, \eta] \propto \prod_{i=1}^{n} [Y_i|X_i, \beta][X_i|\eta]\lambda^n e^{-\lambda P(\beta|\eta)} \tag{5.5}$$

where $P(\beta|\eta) = 1 - \sum_{i=1}^{n} [Y = 0|x_i, \beta]\eta_i$.

5.3 Priors and Posteriors

5.3.1 Prior Specification

5.3.1.1 β, λ priors

Where there is no prior information we propose using a uniform prior for λ and β. These priors are attractive as they are dominated by the likelihoods, and produce modal estimates which maximise the likelihood function in Equation 5.5. Following the arguments in Jeffreys (1961) we could possibly consider the use of the prior $1/\lambda$, as λ is restricted to the positive axis. This is considered in 5.4. We will now consider the prior for η.

5.3.1.2 η prior

The specification of the prior is complicated by the truncation. It is important to distinguish between priors at the untruncated and truncated levels of the likelihood. The correspondence between a prior $[\beta, X]$ in the untruncated space and the induced prior $[\beta, X]_T$ in the truncated space is

$$[\beta, X]_T = \frac{[\beta, X]P(\beta, X)}{\int\int [\beta, X]P(\beta, X)d\beta dX}, \tag{5.6}$$

where $P(\beta, X)$ defines the probability a unit is observed which is a function of both β and X. So for example in situations where β and X are assumed to be apriori independent and the non-informative priors $[\beta]$ and $[X]$ are constants, then the non-informative truncated prior will be

$$[\beta, X]_T = \frac{P(\beta, X)}{\int\int P(\beta, X)d\beta dX}$$

which is not of the usual form. Note that Equation 5.6 can be inverted to give

$$[\beta, X] = \frac{[\beta, X]_T P(\beta, X)^{-1}}{\int\int [\beta, X]_T P(\beta, X)^{-1} d\beta dX}. \tag{5.7}$$

Equation 5.7 suggests that priors in the truncated situation should be tilted by $P(\beta, X)^{-1}$ to obtain the corresponding prior in the untruncated situation. Note that if β and X are assumed to be apriori independent in the untruncated situation, then $[\beta, X] = [\beta][X]$ and from Equation 5.7,

$$[X] = \frac{[X \mid \beta]_T [\beta]_T [\beta]^{-1} P(\beta, X)^{-1}}{\int\int [\beta, X]_T P(\beta, X)^{-1} d\beta dX},$$

or

$$[X] = \frac{[X \mid \beta]_T P(\beta, X)^{-1}}{\int [X \mid \beta]_T P(\beta, X)^{-1} dX}. \tag{5.8}$$

So the prior $[X]$ is equal to the prior $[X \mid \beta]_T$ tilted by $P(\beta, X)^{-1}$.

In the case where X is known to be discrete and have a finite set of values, the prior $[X]$ can be specified with support the observed truncated x. Since ultimately all of the possible values of x will appear in the truncated sample, this will ultimately be equivalent to knowing the support of $[X]$ apriori and specifying the

prior on that support. The nonparametric approach would assume that $[X \mid \beta]_T$ is Multinomial with support the observed x and parameters η which have a Dirichlet distribution. If a Multinomial has k points, then a non-informative self-consistent prior for the parameters of the Multinomial is Dirichlet$(2/k, 2/k, \ldots, 2/k)$. Self consistent means that when the k points are aggregated in k_0 equal sized sets, the marginal prior for the k_0 equal sized sets obtained from the k dimensional prior is Dirichlet$(2/k_0, 2/k_0, \ldots, 2/k_0)$. Note that for $k = 1$, Dirichlet$(1,1)$ is $U(0,1)$ which is the non-informative prior for two points. Equation 5.8 suggests that the location of the truncated prior for X should be tilted by $P(\beta, X)^{-1}$. This can be achieved in the nonparametric setting by taking the 'non-informative' Empirical Dirichlet Prior (EDP) to be

$$Dirichlet\,(\gamma_i, i = 1, \ldots, n)$$

where

$$\gamma_i = \frac{2P(\beta, x_i)^{-1}}{\sum_j P(\beta, x_j)^{-1}}.$$

It is worth examining in detail the outcome if EDP is applied to a situation where X is actually discrete. For simplicity of presentation we will study the case where X has only two values, a and b with probabilities π_0 and π_1 respectively. Using the knowledge that X is discrete and assuming a uniform prior for π_1, we would obtain posteriors

$$[\beta \mid \lambda, \pi_1, Y, X, n] \propto [Y \mid X] \exp(-\lambda P(\beta \mid \pi_1))[\beta], \tag{5.9}$$

$$[\lambda \mid \beta, \pi_1, Y, X, n] \propto \lambda^n \exp(-\lambda P(\beta \mid \pi_1))[\lambda], \tag{5.10}$$

$$[\pi_1 \mid \lambda, \beta, Y, X, n] \propto \exp(-\lambda P(\beta \mid \pi_1))\pi_1^{n_1}\pi_0^{n_0}, \tag{5.11}$$

where

$$P(\beta \mid \pi_1) = \pi_0 P(\beta, a) + \pi_1 P(\beta, b),$$

n_1 is the number of b observations in the sample and $n_0 = n - n_1$. By contrast, if EDP is used and we define

$$\pi_1 = \sum_{b \text{ observations}} \eta_i$$

and $\pi_0 = 1 - \pi_1$, then the posteriors for β and λ are once again given by Equations 5.9 and 5.10. Assuming without loss of generality that the b observations are labelled $1, \ldots, n_1$ and letting

$$y_i = \begin{cases} \eta_i/\pi_1 & , \quad i = 1, \ldots, n_1 \\ \eta_i/\pi_0 & , \quad i = n_1 + 1, \ldots, n \end{cases},$$

the posterior of $\pi_1, y_1, \ldots, y_{n_1-1}, y_{n_1+1}, \ldots, y_{n-1}$ is

$$\begin{aligned} [\pi_1, y_1, \ldots, y_{n_1-1}, y_{n_1+1}, \ldots, y_{n-1} \mid \lambda, \beta, Y, X, n] \;&\propto\; \exp(-\lambda P(\beta \mid \pi_1)) \\ &\times\; \pi_1^{n_1+r_1-1}\pi_0^{n_0+r_0-1}\prod_1^n y_i^{c_i} \end{aligned}$$

where

$$r_1 = \frac{2n_1 P(\beta, b)^{-1}}{n_1 P(\beta, b)^{-1} + n_0 P(\beta, a)^{-1}},$$

$$r_0 = \frac{2n_0 P(\beta, a)^{-1}}{n_1 P(\beta, b)^{-1} + n_0 P(\beta, a)^{-1}}$$

and

$$c_i = \frac{2P(\beta, x_i)^{-1}}{n_1 P(\beta, b)^{-1} + n_0 P(\beta, a)^{-1}}.$$

So π_1 and $y_1, \ldots, y_{n_1-1}, y_{n_1+1}, \ldots, y_{n-1}$ are independent in the posterior and the posterior of π_1 is

$$[\pi_1 \mid \lambda, \beta, Y, X, n] \propto \exp(-\lambda P(\beta \mid \pi_1))\pi_1^{n_1+r_1-1}\pi_0^{n_0+r_0-1}. \tag{5.12}$$

Note that $y_1, \ldots, y_{n_1-1}, y_{n_1+1}, \ldots, y_{n-1}$ do not appear in the posteriors for β and λ and so can be ignored unless specifically required. Consequently the only difference between the posteriors are the coefficients of π_1 and π_0 in Equations 5.11 and 5.12. Since the difference in the coefficients is at most one and converges to zero as the sample size increases, the difference in the posteriors will be negligible as the sample size increases and EDP will be essentially equivalent to the known two point case. An identical argument applies to establish the essential equivalence of EDP and k point support for X. Finally since any continuous density for X can be arbitrarily well approximated by a discrete k point support density with k sufficiently large, it follows that any non truncated prior for X can be arbitrarily well approximated by EDP for sufficiently large sample sizes.

5.3.2 Gibbs Sampling Algorithm

We demonstrate the results of the previous sections using a Poisson model for the untruncated data. We thus briefly review the standard approach. Following McCullagh & Nelder (1989), consider a sample, $(Y_1, X_1) \ldots (Y_N, X_N)$ where Y_i is the response for the ith individual and X_i are covariates measured on the ith individual. It is assumed that Y_i follows a Poisson distribution

$$Pr(Y_i = y) = \frac{e^{-\mu_i}\mu_i^y}{y!}; \quad y = 0, 1, 2, \ldots$$

where μ_i is the mean of the process. If the log link is used the mean is modelled by

$$g(\mu) = \log \mu_i = x_i'\beta$$

where β is a vector of parameters and $g()$ is the link function.

Thus with this model for $[Y|X]$ and the above priors the posterior becomes,

$$[\beta, \lambda, \eta | Y, X^{(n)}, n] \propto \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \prod_{i=1}^{n} \eta_i e^{-\lambda P(\beta|\eta)} \prod_{i=1}^{n} \eta_i^{\gamma_i - 1}. \tag{5.13}$$

To produce inference about β, λ and η requires the numerical integration of the form given in Equation 5.13 to obtain the required posteriors. This is computationally prohibitive in all but the most simple cases.

Instead we propose using a Gibbs sampling algorithm (Geman & Geman, 1984) to approximate the posterior 5.13. The use of the Gibbs sampler in statistical problems is described in numerous articles, for example see Tanner (1993). The implementation of the Gibbs sampler requires the decomposition of Equation 5.13 into a set of convenient conditional distributions. We have chosen the following.

$$[\beta|\lambda,\eta,X^{(n)},Y,n] \propto \prod_{i=1}^{n}\frac{e^{-\mu_i}\mu^{y_i}}{y_i!}\prod_{i=1}^{n}\eta_i\lambda^n e^{-\lambda P(\beta|\eta)}\prod_{i=1}^{n}\eta_i^{\gamma_i-1} \tag{5.14}$$

$$[\lambda|\beta,\eta,Y,X^{(n)},n] \propto \lambda^n e^{-\lambda P(\beta|\eta)} \tag{5.15}$$

$$[\eta|\beta,\lambda,Y,X^{(n)},n] \propto \prod_{i=1}^{n}\eta_i e^{-\lambda P(\beta|\eta)}\prod_{i=1}^{n}\eta_i^{\gamma_i-1} \tag{5.16}$$

To sample from Equation 5.14, note that it consists of the usual likelihood used in the non truncated analysis, but tilted by the remaining terms. Following Zeger & Karim (1991) we sample from Equation 5.14 by rejection sampling. This is done by maximising the function for β, and finding the curvature at the maximum point. A Gaussian density with appropriately deflated curvature is then used as the rejection envelope. Although we have not proved the log concavity of Equation 5.14, no problems with incorrect acceptances have arisen in a range of simulations carried out. Heuristic arguments suggest it should not present difficulties. The choice of the deflation factor depends on the problem, but no difficulties were encountered in choosing a value that achieved acceptable rejection rates, while still being a valid envelope. Alternatively the distribution can be decomposed into the univariate conditional distributions and a technique such as Adaptive Rejection Sampling (Gilks & Wild, 1992) implemented. This will not affect the convergence of the chain but may increase the number of iterations required to adequately explore the posterior.

Examining Equation 5.15 it is immediately apparent that it is proportional to a Gamma density with shape and scale parameters, $n+1$ and $P(\beta|\eta)$ respectively. It is thus simple to generate deviates from this distribution. Of course in some applications it is of interest to make inference regarding N. This can be done by noting that the Gibbs sampling algorithm produces approximate samples from the marginal distribution $[\beta,\lambda,\eta|Y,X,n]$. Thus points sufficiently separated in the sequence can be treated as independent realizations and the conditional density for N, $[N|\beta,\lambda,\eta,Y,X,n]$ can be derived from Equation 5.3 and used to produce a sample from the full posterior. Alternatively, the conditional for $[N|.]$ can be included explicitly in the sampler.

Sampling from Equation 5.16 presents greater difficulties. One approach is to consider the reduced vector $[\eta_1,\ldots,\eta_{n-1}]$. Observing that the conditional distribution of η_i, given the other η's in the reduced vector is

$$[\eta_i|\beta,\lambda,n,\eta_{-i},Y,X^{(n)}] \propto e^{-d_i\eta_i}\eta_i^{\gamma_i}\left(1-\sum_{j\neq n}\eta_j\right)^{\gamma_n} \tag{5.17}$$

where

$$d_i = \lambda P(\beta,x_i) - \lambda P(\beta,x_n),$$

$$\eta_{-i} = \eta_1,\ldots\eta_{i-1},\eta_{i+1},\ldots,\eta_{n-1}$$

and

$$0 \leq \eta_i \leq 1 - \sum_{j\neq i,n}\eta_j,$$

it is therefore possible to use a rejection sampling algorithm to sample η_i. This is done using a uniform distribution over $[0, 1-\sum_{j\neq i,n}\eta_j]$ scaled to the maximum

of Equation 5.17. This maximum is the solution of a quadratic equation and is thus easily found. To produce a sample from the full vector a Gibbs sampling sub-chain can be used, by repeating this process until approximate convergence is reached. This sub-chain is of course not necessary for the Gibbs algorithm to converge, but limited practical experience suggests there is strong dependence in the η sequence and the chain mixes slowly. The time consuming step of the algorithm is the rejection sampling of the conditional distribution for β, and it is efficient to perform this as few times as possible.

A concern with the Gibbs sub-chain is that convergence may still be slow due to the dependence between the components of η. An alternative is to consider a rejection sampling algorithm to simulate from the full η vector. This can be used to either assess the convergence of the Gibbs approximation, or if efficient enough, to directly sample η. This is detailed in Appendix 2. The conclusion reached was that the Gibbs algorithm appeared to give a good approximation to a sample from $[\eta|\beta, \lambda, n, \eta_{-i}, y, x]$ if the following conditions were met. Firstly, if it was run for n full iterations. Secondly if, by permutation if necessary, η_n was set to the η with the largest expectation. This second condition allows the easiest traversal of the simplex and thus speeds convergence from arbitrary starting values.

5.4 Alternative Augmentations

In the case of censored regression, progress has been made by augmenting the data with the censored responses and implementing a Gibbs algorithm. This exploits the simple structure of the resulting likelihood. Truncated data has some similarities to this situation. In the truncated case considered above the response is known for the unobserved members, but the covariate values are not. Thus it is attractive to consider the augmentation of the data with the unobserved covariates. Complications arise though due to the unknown value of N. Thus we must consider augmenting by a random number of observations, a situation that has not been addressed in the literature. It has been considered in the implementation of the EM algorithm for truncated data (McLachlan & Jones, 1988), but in that case the E-step constrains the size of the problem.

This augmentation approach is attractive as it reduces the truncated case to the untruncated case, with the obvious simplifications in the sampling algorithms. It can thus potentially allow a richer class of models, incorporating for example random effects, to be implemented without unduly complicating the basic scheme. Problems do arise though. The augmentation considered can result in computational problems, causing slow convergence of the Gibbs sequence. To examine this behaviour we consider the following simple situation.

5.4.1 Data Augmentation

We consider a sample of size N drawn from a $N(\theta, 1)$ distribution, where all points are truncated below some constant c, producing n observations. In this case the unconditional likelihood of the data is found from Equation 4.2 with $k = 1$ and

$$L_1(n; \theta, N) = \binom{N}{n} (1 - \Phi(c - \theta))^n \Phi(c - \theta)^{N-n}$$

$$L_2(y|n;\theta,N) = \prod_i^n \frac{\phi(y-\theta)}{1-\Phi(c-\theta))},$$

where $\phi()$ and $\Phi()$ are the density and distribution function of the standard Gaussian distribution. We propose using the improper uniform prior for θ over the real line, and the improper prior $[N] = 1/N$ for $N = 1, 2, \ldots$. This choice of prior is made for pragmatic reasons, although may also be justified by the arguments of Jeffreys (1961). In this case it is justified as follows. First, with heavy truncation improper posteriors can arise if the uniform prior is used for N as it effectively puts too much weight on extreme values of N for which there is no prior belief. Secondly this prior leads to tractable forms for the conditional distributions. This second justification, though indefencible in applied work, is sufficient here as we only wish to demonstrate the deficiency of the approach, and choosing this prior does not impact on this deficiency.

With these priors the posterior for θ and N is thus

$$[\theta, N|n, y] \propto \binom{N-1}{n-1} (1-\Phi(c-\theta))^n \Phi(c-\theta)^{N-n} \prod_i^n \frac{\phi(y-\theta)}{1-\Phi(c-\theta)} \qquad (5.18)$$

As θ contains a single parameter this likelihood can be handled by analytic techniques to produce marginal posteriors and other quantities of interest. For arguments sake we will consider the use of data augmentation and Gibbs sampling to explore the posterior.

If we attempt to implement the Gibbs algorithm directly we find that $[N|n, y, \theta]$ has a negative binomial distribution and is thus easy to sample from, but $[\theta|N, n, y]$ does not have a simple form and thus presents difficulties for efficient sampling. Hence we attempt to think of a data augmentation scheme that can simplify the problem. The obvious candidate is to augment the data with the unobserved $N-n$ values, labelled y_N say. Now the y_N values have density

$$[y_N|y, N, n, \theta] = \prod_{i=n+1}^N \frac{\phi(y_i-\theta)}{1-\Phi(c-\theta)}$$

where $y_i < c$ for all i. A problem at this point is that the distribution of $[N|y_N, y, \theta]$ is degenerate, as the number of components in y_N and y uniquely determines N. This problem is removed if analogous to the case with the η's we consider the joint distribution of N and y_N, where conditional on N the y_N are independent. With this augmentation we have the series of conditionals to draw from:

$$\begin{aligned}
[N|n,\theta] &\propto \text{Negative Binomial}(n, 1-\Phi(c-\theta)) \\
[y_N|N,\theta] &\propto \text{Truncated Normal with mean } \theta \text{ truncated at } c \\
[\theta|y, y_N, N, \theta] &\propto \text{Normal with mean } \bar{y} \text{ and variance } 1/N
\end{aligned}$$

where $\bar{y}$ is the mean over the observed y and augmented, y_N responses.

The conditions given in Smith & Roberts (1993) are satisfied and so the stationary distribution of the algorithm will be the same as Equation 5.18, and by the ergodic property of the chain, histograms of the Gibbs sequence will converge to the true posterior density. We consider a sample with heavy truncation. We sample from a distribution with $N = 50$, $\theta = 0$ and c=1. The sample path for a sample is shown in Figure 5.1. Note the heavy dependence in the sequence especially as θ becomes small, and hence the number of augmented values becomes large. Figure 5.2

presents the estimated density of the marginal posterior for θ for various sequence lengths, with the true marginal posterior calculated from Equation 5.18 overlaid. Note that there is still significant error in the empirical marginal posterior after 80000 iterations of the algorithm. This slow mixing means the Gibbs sampler must be run for large number of iterations to ensure that the estimated posteriors are accurate with high probability.

This need for a large number of iterations presents problems when the number of augmented parameters is also large. In the present case the number of augmented observations becomes large. The truncated normal observations are drawn using the obvious rejection algorithm, which can become very inefficient, when the current estimate of θ implies extreme truncation. This sampling can made more efficient by modifying the Box-Muller algorithm (see Morgan (1989)) so that it only generates observations close to the region of interest (this is done by generating uniform random numbers on a restricted range of $(0, 1)$). In more complex sampling situations these efficiency gains may not be available, and the method becomes computationally intractable.

These observations are in agreement with general discussion in the literature. The general understanding is that excessive augmentation can lead to problems of slow convergence, as the set being conditioned on becomes too large (Liu, Wong, & Kong, 1994). Thus in the present case, where the augmentation is not producing major simplifications in the sampling algorithm the use of this augmentation is not advised. Unpublished results have shown that in cases with less extreme truncation these problems are not as severe and convergence of the posterior estimates is much more rapid. Thus this type of augmentation may find use in these cases.

5.5 Examples

5.5.1 Example 1

As a simple example consider a Poisson regression model with a single intercept term. In this case X is 1 with probability 1. Thus the posterior 5.13 reduces to

$$[\beta, \lambda, \eta | n, y, x] \propto \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \prod_{i=1}^{n} \eta_i \lambda^n e^{-\lambda P(\beta)} \prod_{i=1}^{n} \eta_i^{\gamma_i - 1}. \tag{5.19}$$

For any η vector this function is maximised at the posterior mode by $\lambda = n/P(\beta)$ and the solution of

$$\frac{\sum_n y_i}{n} = \frac{\mu}{P(\beta)} \tag{5.20}$$

which is the same as the maximum likelihood estimate based on the conditional model.

5.5.2 Example 2

In this section a small scale simulation experiment is presented. The Gibbs sampling algorithm was implemented using Splus, with C code used to gain efficiency when sampling from the η vector. The simulations consisted of the following steps:

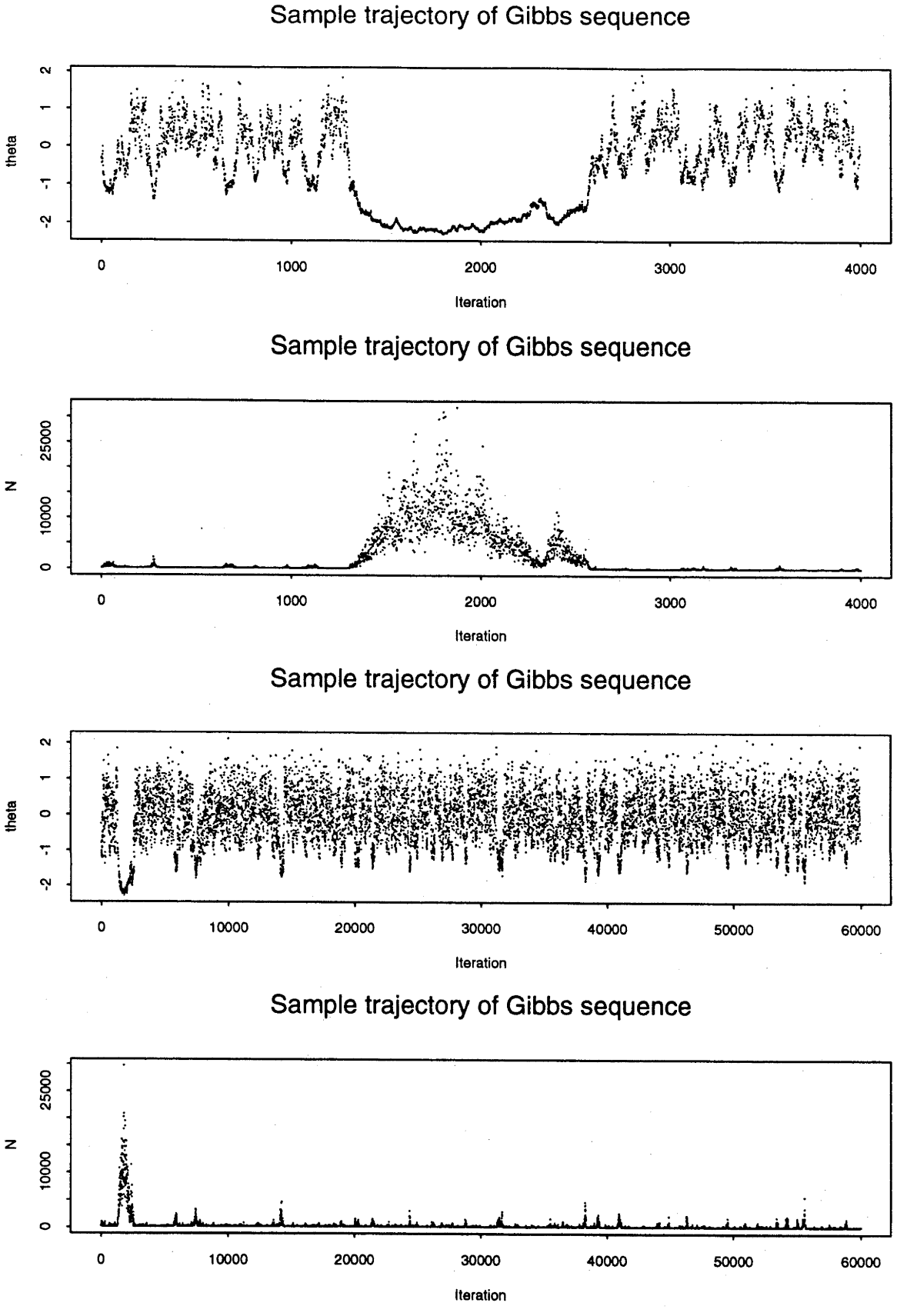
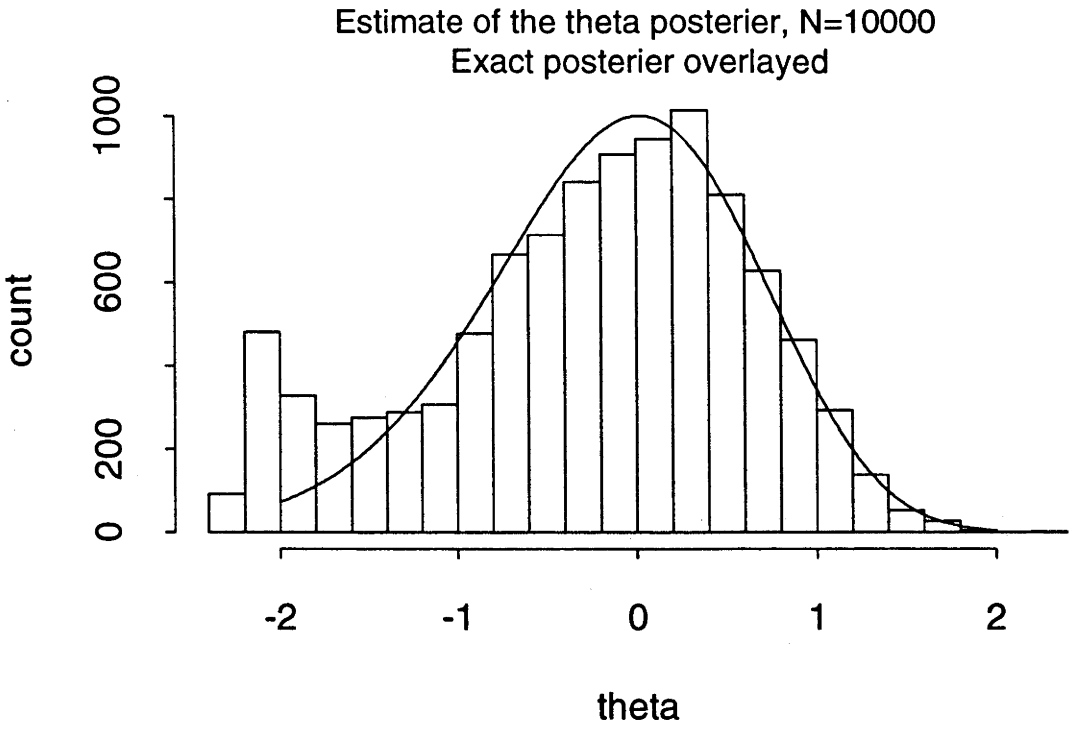


Figure 5.1: Observed sample trajectories for θ and N, from a typical sequence. Top panel plots the first 4000 iterates from the bottom panel sequence.



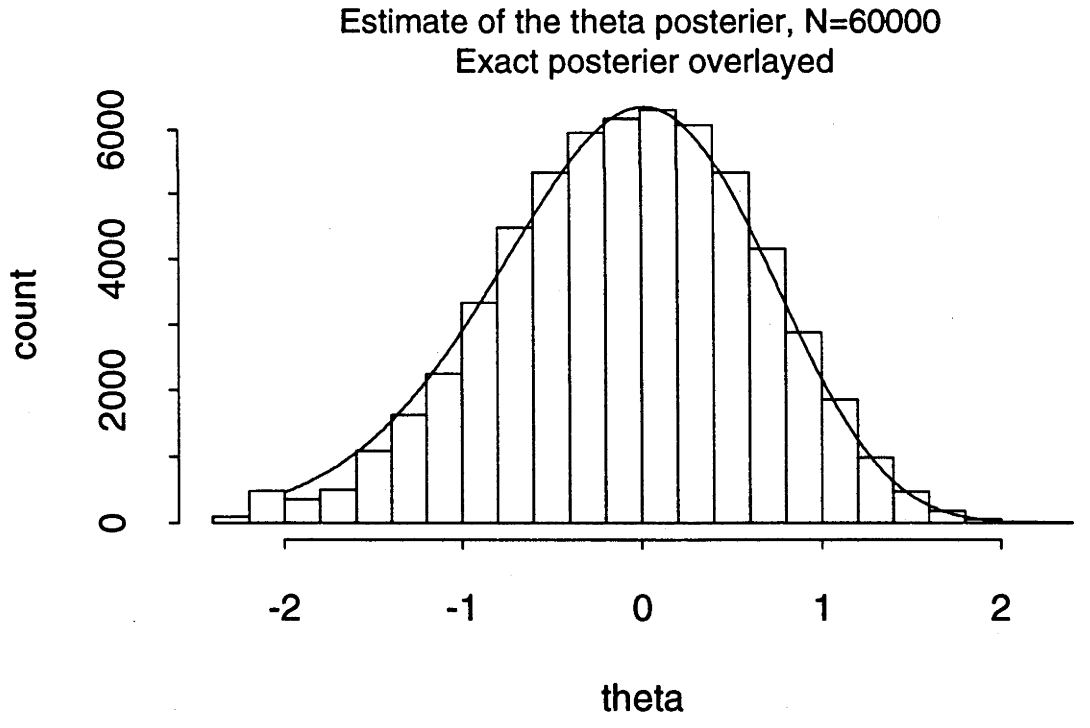


Figure 5.2: Estimated posterior from θ calculated for 2 sequence lengths. Note the same sequence was used in each calculation. The true posterior is overlaid.

1. A covariate matrix was generated with 100 rows and ith row

$$(1, x_{i1}, x_{i2})$$

where x_{i1} had a uniform distribution over [-1,1] and x_{i2} was Bernoulli with probability .5. From this the expected values, μ were generated via the link function $g(x) = log(x)$ and the linear predictor $X\beta$, where $\beta = (\beta_0, \beta_1, \beta_2) = (1, -2, 2)$. In the simulations the marginal probability of being observed was .598, so each data set had approximately 60 observations.

2. The following steps were iterated two hundred times.

   - A sample y was generated from μ.
   - The sample was truncated to form a sample y^*, x^*.
   - The Gibbs algorithm was run for 2000 iterations, and the sample path for each parameter was saved.

The use of simulation to assess the behaviour of a Bayesian procedure is worthy of comment. The Bayesian approach does not require the frequentist justification over repeated samples. The Bayesian could simulate from their priors and then the densities, but this would be purely an exercise in verifying the calculus. The simulations carried out here are justified in the following ways. Firstly, the use of the non informative priors for β and λ means that the posterior is approximating the likelihood function and so the posterior modes are interpretable as maximum likelihood estimates. Secondly, the use of the procedure with the Empirical Dirichlet Prior in any practical problem requires confidence in its behaviour.

Each iteration of the Gibbs sampler consisted of drawing samples directly from the conditional densities for η and λ and using the Gibbs sub-chain described in Section 5.3.2 to gain an approximate sample from the η's. The number of iterations used in the Gibbs sub-chain was equal to the length of the observed y.

The Gibbs sampler was started by using the known true parameter values. It was run for 50 iterations to eliminate the effect of the initial conditions. In the preliminary investigations the posterior densities appeared unimodal and well behaved so it was considered unnecessary to use a sequence of different starting values. The simulations performed in Section 5.5.3 confirm this view.

For each run, the posterior mean, mode and selected quantiles were estimated for each of the λ and β marginals. The mode was estimated using kernel smoothing. The use of the mean of the sequence as an estimate of the mode exhibited some bias due to the skewed posterior distributions. This was most pronounced with respect to λ, which possessed a long right tail. For the η's the calculation of the mode could only sensibly be done by considering the joint mode. As the dimension of this is approximately 60 it was not undertaken. Instead the mean of each component of η was calculated.

The results of the simulation for the regression coefficients are presented in Figure 5.3. This figure shows the distribution of the estimated modes from the simulations. For comparison the distribution of the maximum likelihood estimates (MLE) based on the true truncated model and those produced by the Poisson model, ignoring the truncation, are also presented. These were produced from the same data sets generated during the simulation. Examining the figure it is seen that the estimate of the intercept term exhibits a slight negative bias over the

parameter	β_0	β_1	β_2	λ
coverage	0.93	0.92	0.86	0.86

Table 5.1: Estimated coverage probabilities for simulation 1.

simulations, while the estimate for the continuous covariate has a small positive bias. The exact extent of these biases would require extensive simulation. This is currently not computationally feasible. The outlying terms in these plots are produced by simulations where the response for units with $x_{i2} = 1$ were all 1. In this case β_2 can equally plausibly have a range of negative values (implying heavy to severe truncation), and this is reflected in the Gibbs sampler, which drifts over this parameter. This also causes λ to inflate as there is no information regarding the extent of the truncation. The distribution of the estimated mode of the marginal posterior for λ is presented in Figure 5.4.

Note that the comparison between the Poisson MLE, ignoring the truncation, and the other estimates is not simple. This is due to the Poisson MLE attempting to model the mean of the observed responses, whereas the other techniques attempt to model the mean of the underlying process. Thus they are attempting to estimate different quantities. With this in mind, it is still informative to see the effect of the tilting in Equation 5.5 on the usual Poisson likelihood. In addition it also shows the serious errors in location and hence precision that can occur if the truncation is ignored, and the resulting estimates are given a truncated interpretation.

Figure 5.5 compares the distribution of the x_{i1}, which are drawn from a uniform distribution, with the empirical distribution estimated using the approximate expected value of the η vector. As can be seen, the empirical approach effectively recovers the untruncated marginal distribution of the covariates. Note that the similarities in behaviour between the curves is due to the fixed design matrix leading to common points between simulations.

Table 5.1 gives the estimated coverage of the 90% Bayesian intervals calculated from the estimated 5% and 95% quantiles. These coverages are approximately correct.

The simulation was repeated using the same covariate design but now with a population size of 200. This repetition was performed in an attempt to detect any pathological behaviour. The first repetition used parameter values $\beta = (0, 1, .5)$, which gave a marginal probability of being observed of approximately .71. Thus in this simulation there is less truncation of the covariate distribution, due to the lower truncation rate. The results of the simulation for the regression parameters are given in Figure 5.6, while the distribution of the Bayes modal estimates are given in Figure 5.7. The estimated marginal distribution of the uniform covariate is given in Figure 5.8. The coverage of the approximate 90% Bayesian intervals calculated from the sample quantiles of the marginal posterior distributions are given in Table 5.2. There is some evidence of over coverage but it is mild. The sampling distribution of the confidence intervals is given in Figure 5.9, and it can be seen that the procedure is well behaved, with no extreme variation in interval lengths.

The last repetition used a population size of 200 and parameter values $\beta = (0, -2, 3)$, which gave a marginal probability of being observed of approximately

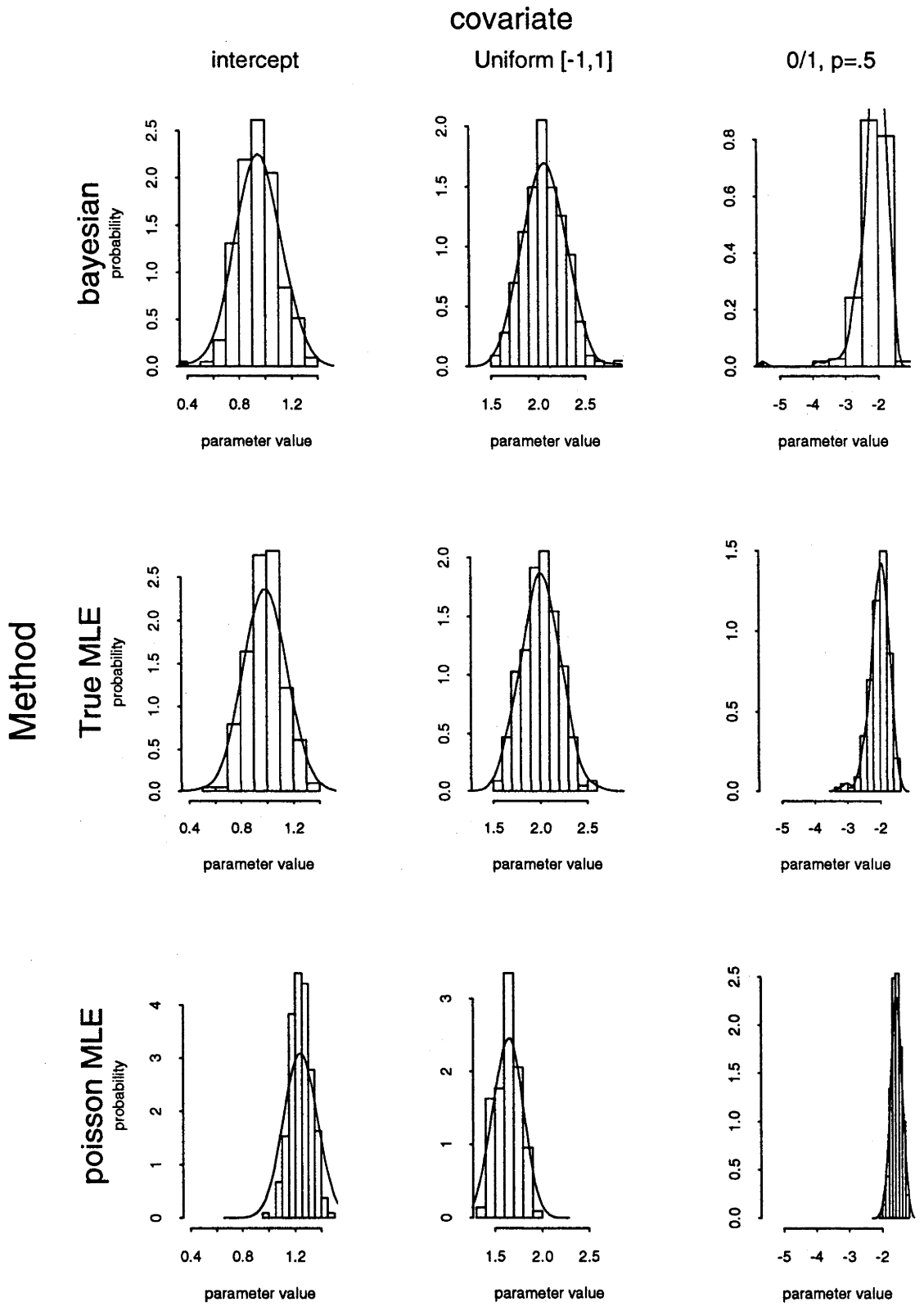


Figure 5.3: The estimated sampling distributions for the estimators with $\beta = (1, -2, 2)$ over 214 simulations. See text for details.

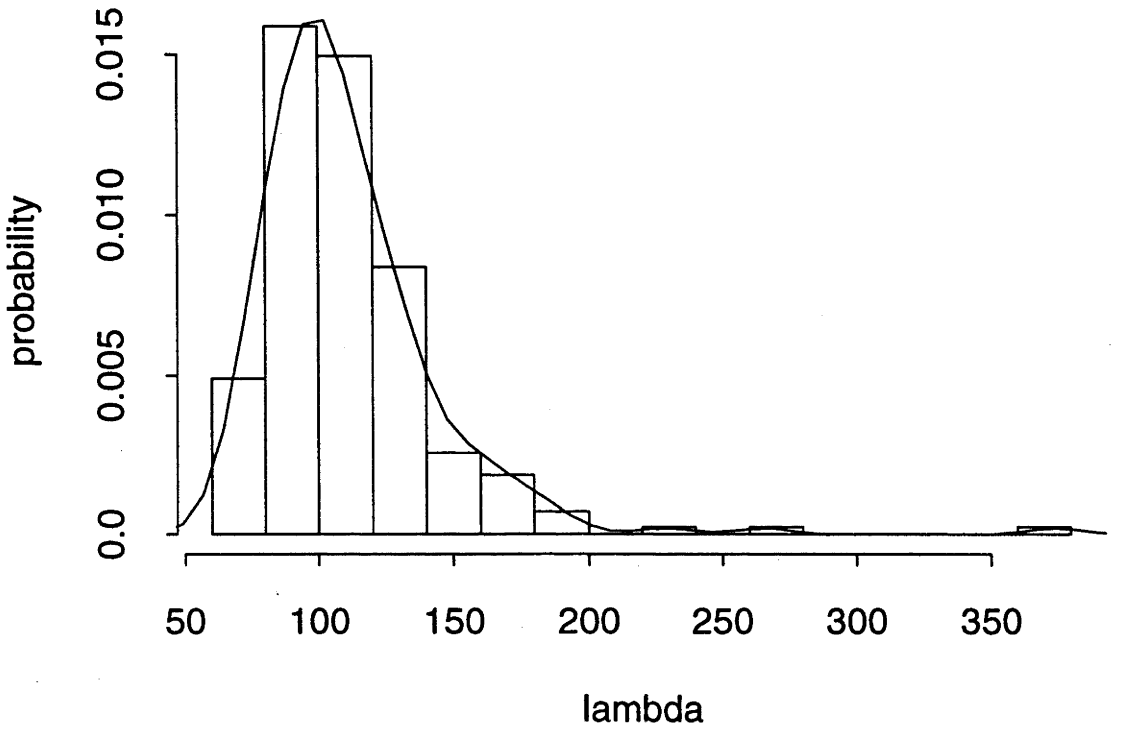


Figure 5.4: Distribution of estimated modes for λ over 214 simulations.

parameter	β_0	β_1	β_2	λ
coverage	0.91	0.92	0.94	0.95

Table 5.2: Estimated coverage probabilities for simulation 2.

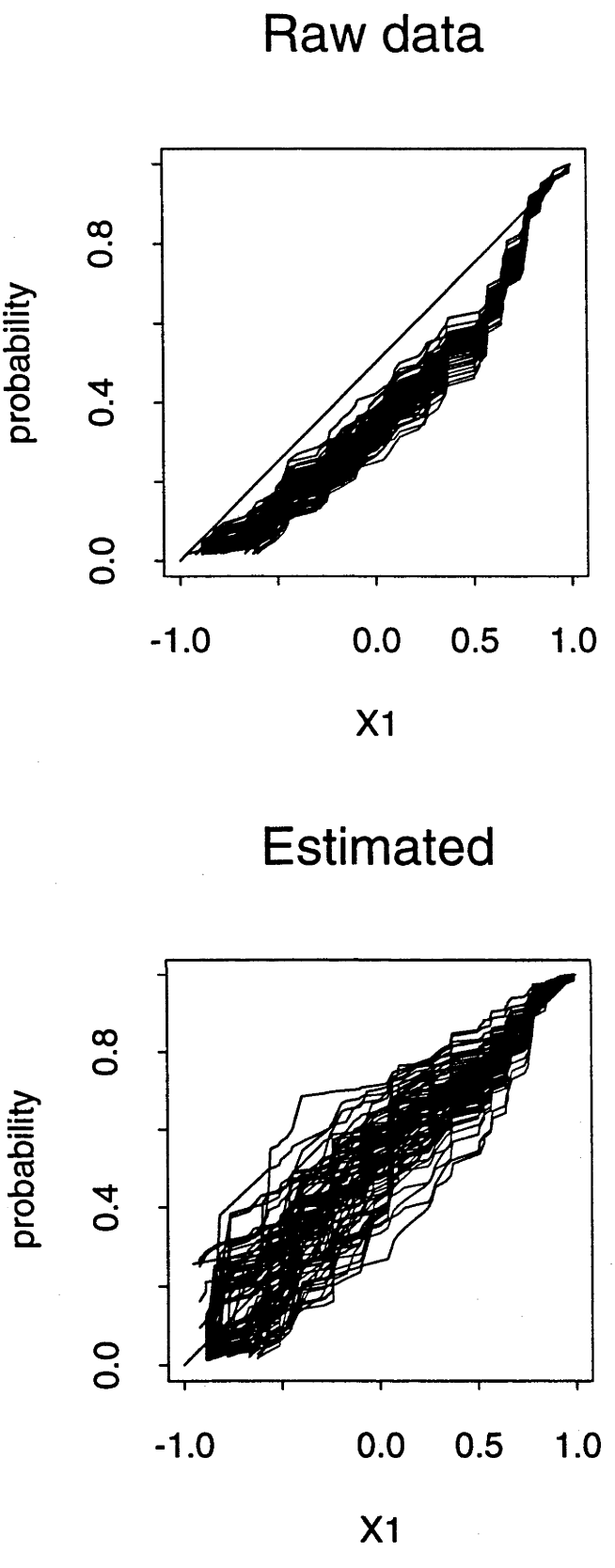


Figure 5.5: Observed and estimated marginal distribution of the uniform covariate in simulation 1. See text for details.

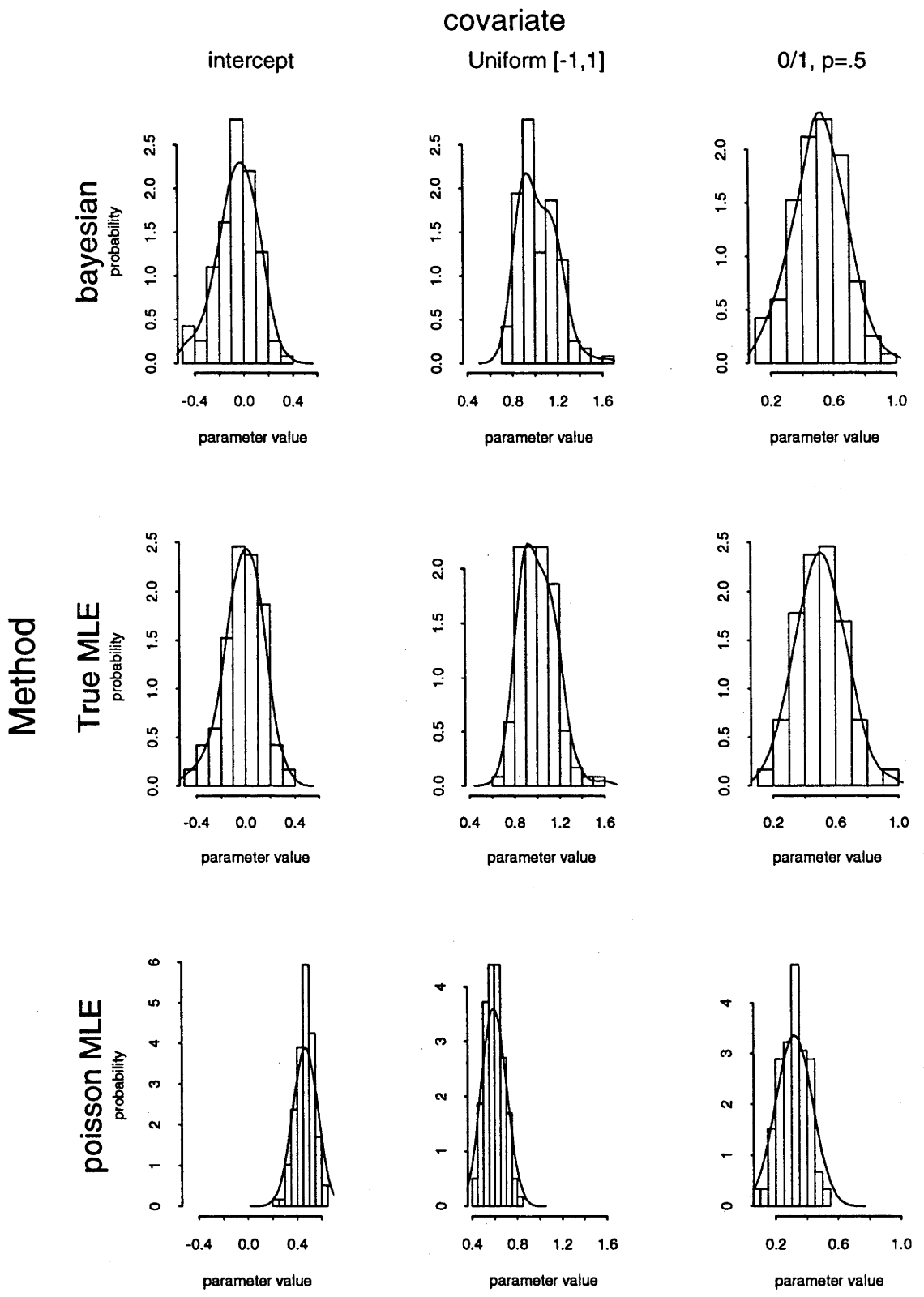


Figure 5.6: The estimated sampling distributions for the estimators with $\beta = (0, 1, .5)$ over 118 simulations. See text for details.

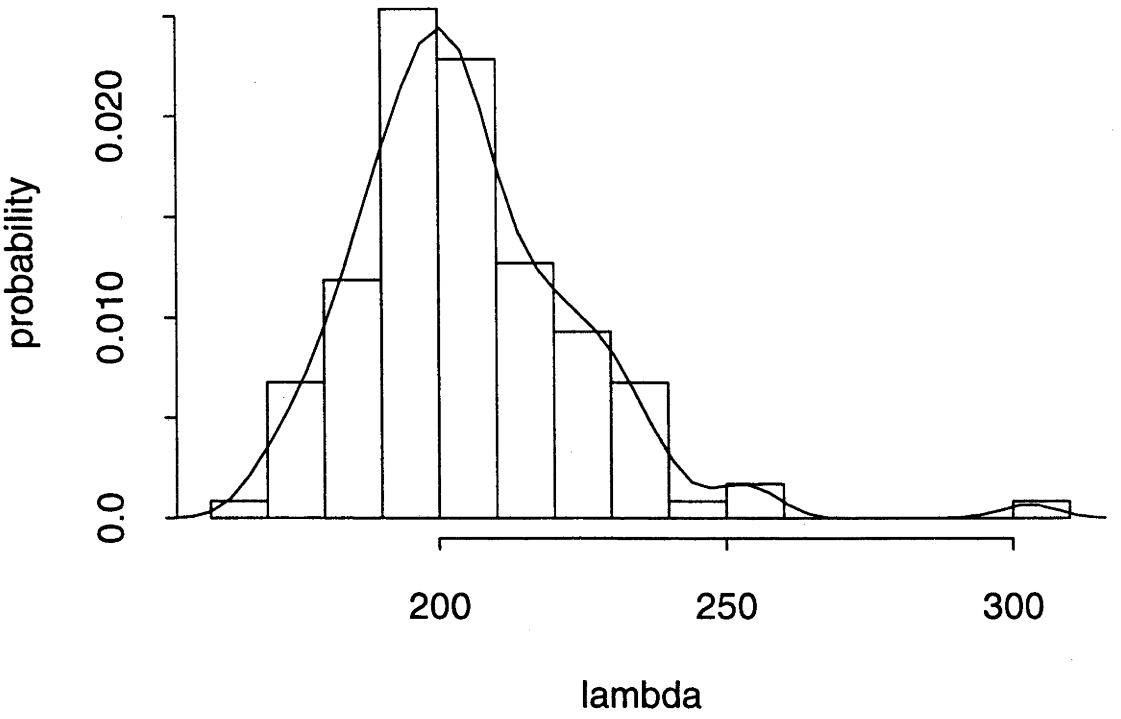


Figure 5.7: Distribution of estimated modes for λ over 118 simulations, $\beta = (0, 1, .5)$, $N = 200$.

parameter	β_0	β_1	β_2	λ
coverage	0.96	0.93	0.92	0.94

Table 5.3: Estimated coverage probabilities for simulation 3.

.81. Hence in this simulation there was even less truncation than the previous two. The simulation was repeated only 114 time due to computing constraints. The results of the simulation for the regression parameters are given in Figure 5.10, while the distribution of the Bayes modal estimates are given in Figure 5.11. The estimated marginal distribution of the uniform covariate is given in Figure 5.12 with the estimated coverage of the 90% Bayesian intervals given in Table 5.3. There is again some evidence of over coverage, but the sample sizes are small.

5.5.3 Convergence Behaviour

There has been considerable discussion in the literature regarding techniques for assessing the convergence of the Gibbs sampler. This issue is critical for assessing the length of sequence necessary to adequately approximate the posterior distribution. As is well recognised, while there is presently a reassuring literature providing results that ensure the convergence of the Gibbs sampler to the target distribution, there is a paucity of analytic results regarding the rate of convergence of Markov Chain Monte Carlo Methods in practical problems (Smith & Roberts, 1993). For example Roberts & Polson (1994) present results regarding the geometric con-

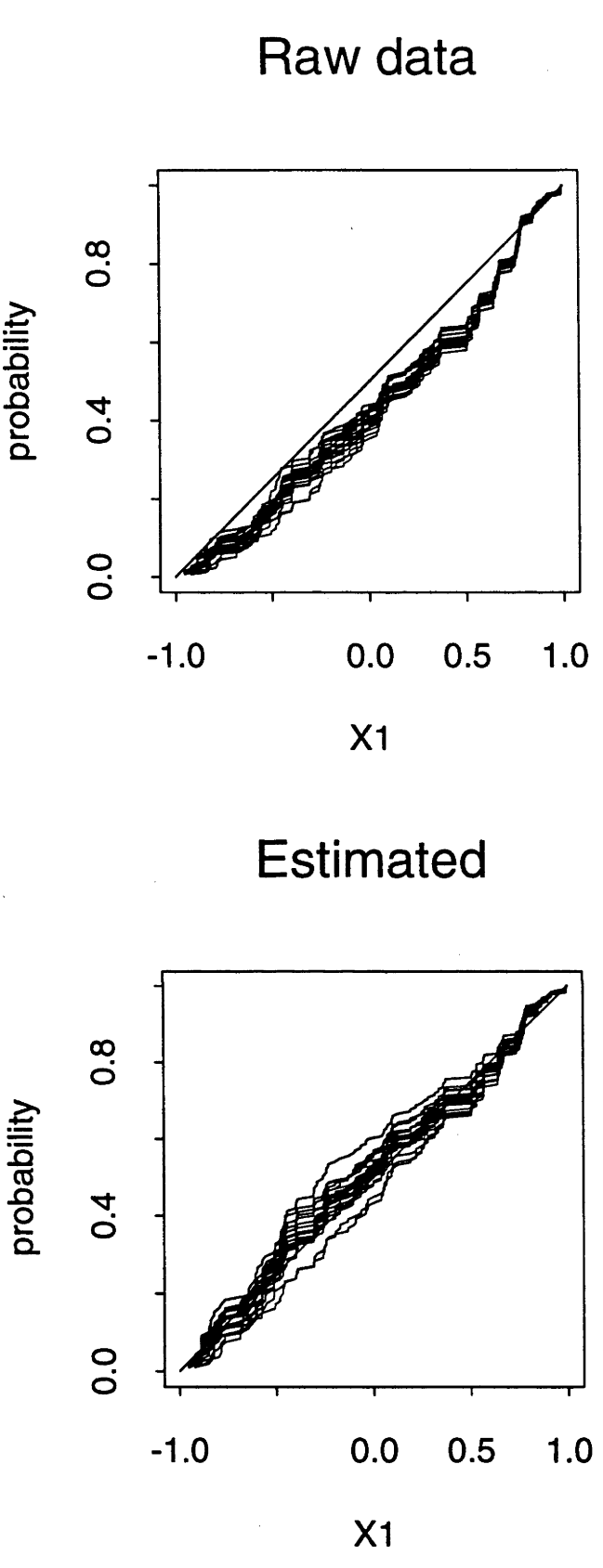


Figure 5.8: The empirical and estimated marginal distribution for the uniform covariate with $\beta = (0, 1, .5)$ over 20 randomly selected simulations. See text for details.

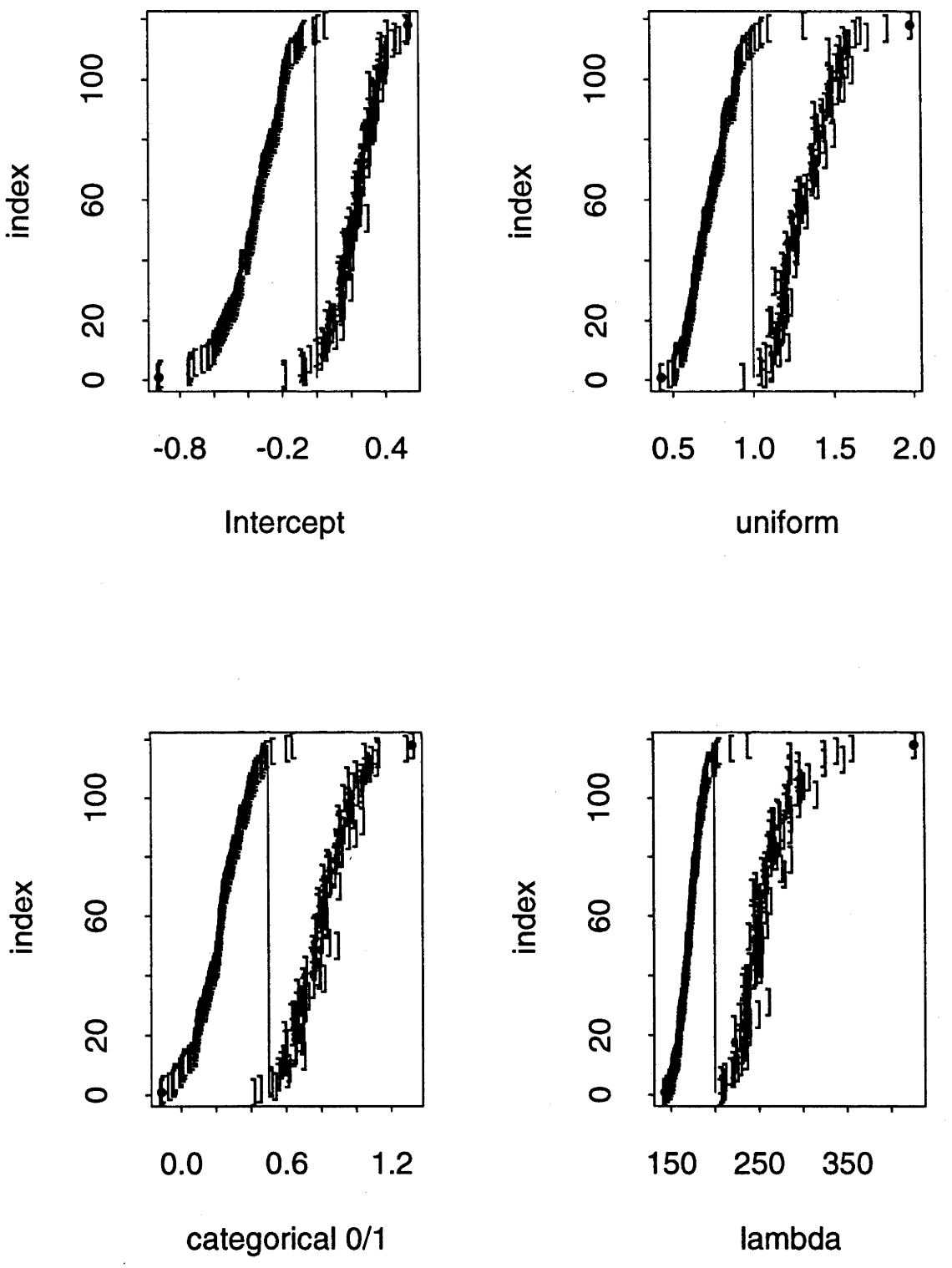


Figure 5.9: Sampling distribution of approximate 90% Bayesian intervals. [=lower,]=upper.

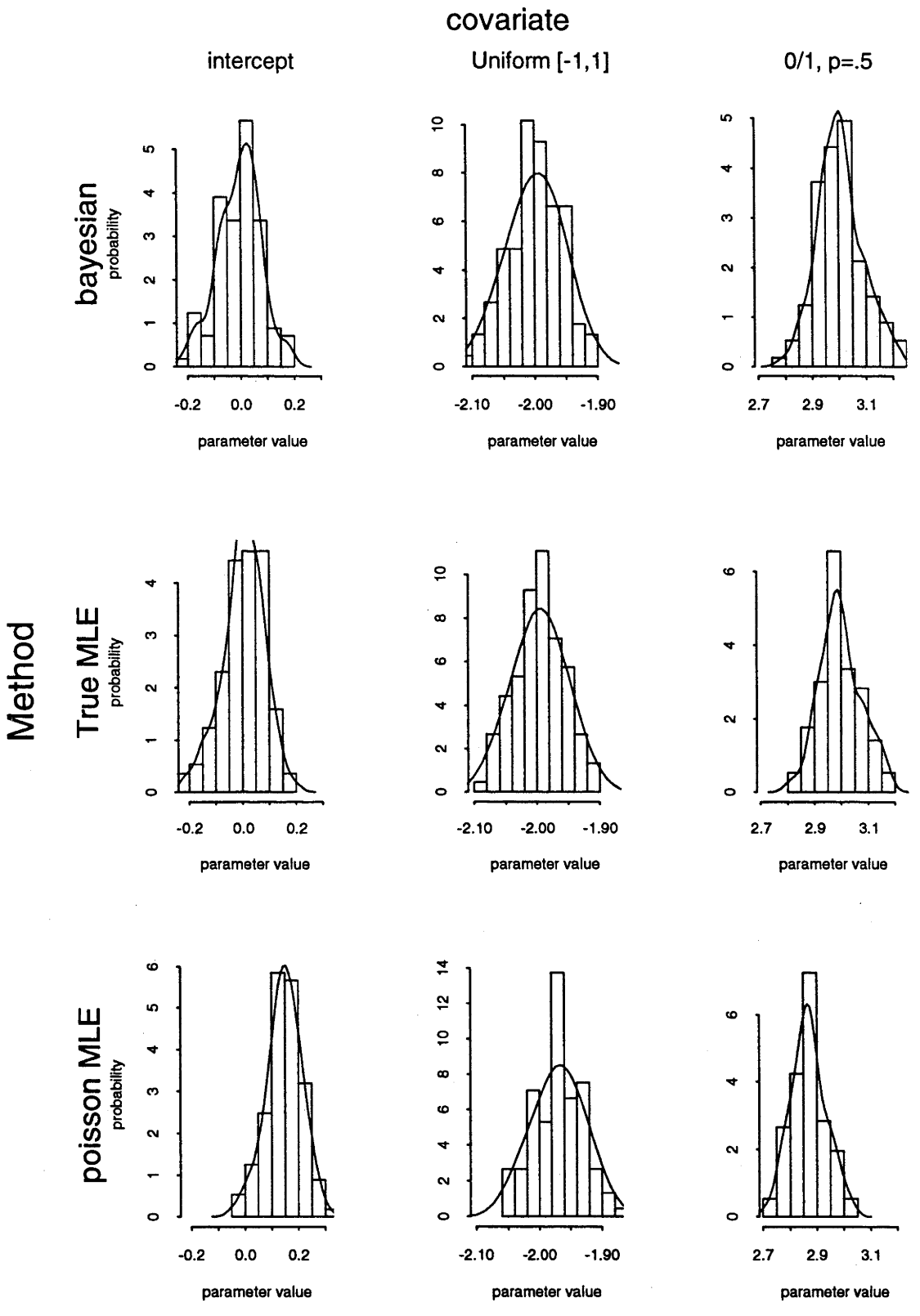


Figure 5.10: The estimated sampling distributions for the estimators with $\beta = (0, -2, 3)$ over 114 simulations. See text for details.

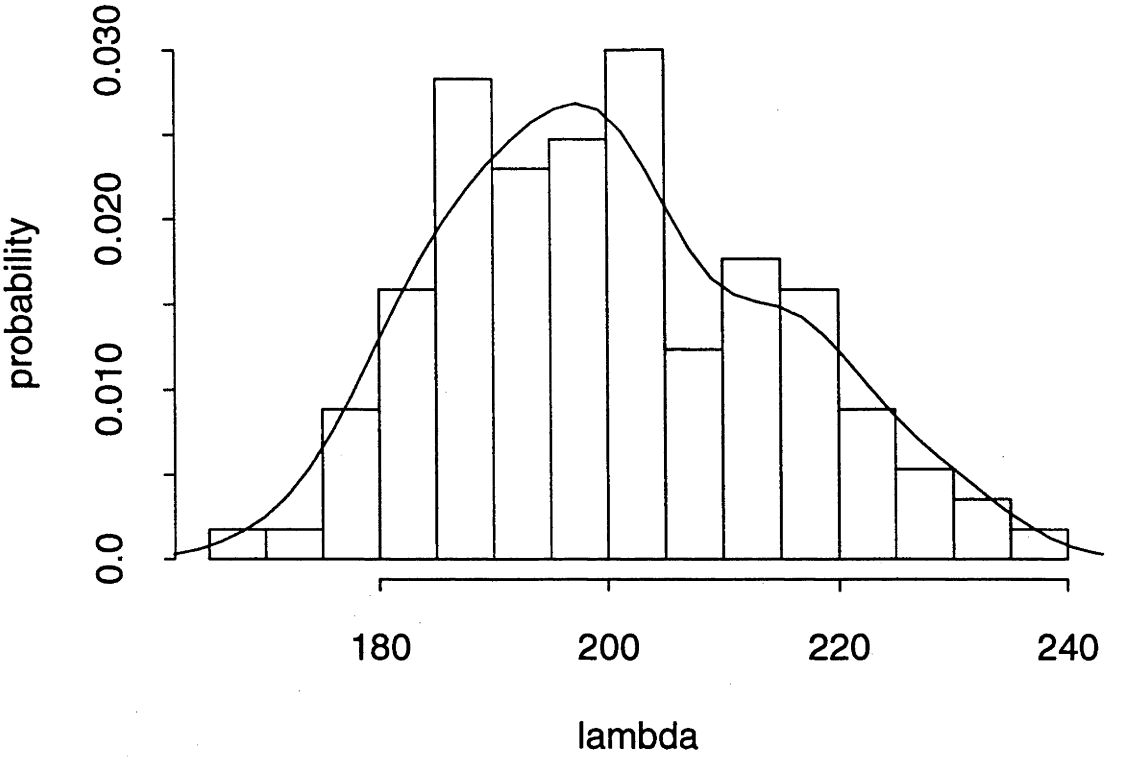


Figure 5.11: Distribution of estimated modes for λ over 114 simulations, $\beta = (0, -2, 3)$, $N = 200$.

vergence of the Gibbs sampler, but as they note, workable estimates for rates of convergence are only available in certain stylised problems. As another example, Liu et al. (1994) present technical results regarding the efficiency of different augmentation schemes and estimators of the posterior.

Given the lack of theoretical results, it is not surprising that there are numerous suggestions in the literature proposing techniques for assessing convergence. A useful discussion of the practical implementation of algorithms and assessment of convergence can be found in Gelman & Rubin (1992a) and Geyer (1992) and the discussion of these papers, and also in Smith & Roberts (1993). It is beyond the scope of this thesis to attempt a general treatment of the convergence properties of the Gibbs sampling algorithm discussed. For one, the convergence behaviour will depend critically on the design space chosen, with the additional complication of the empirical distribution of the covariates. Instead we will present limited empirical evidence from the examples given to justify our confidence in the sampler's behaviour.

Raftery & Lewis (1992a) and Raftery & Lewis (1992b) present arguments to allow the estimation of the length of a chain that is required to produce estimates of quantiles of the posterior distribution to a desired accuracy. This technique is implemented in the Splus function *gibbsit* available from statlib. This software was run on a random sequence from the second simulation. The inputs were chosen so that the algorithm estimated the run length needed to estimate the $Pr(U < u)$ of the posterior to within .01 with 95% confidence, when the true value is .05 . The parameter *eta* was set at the recommended default of .01, which determines the length of the burn-in period required for the N step probability of the 0/1 process

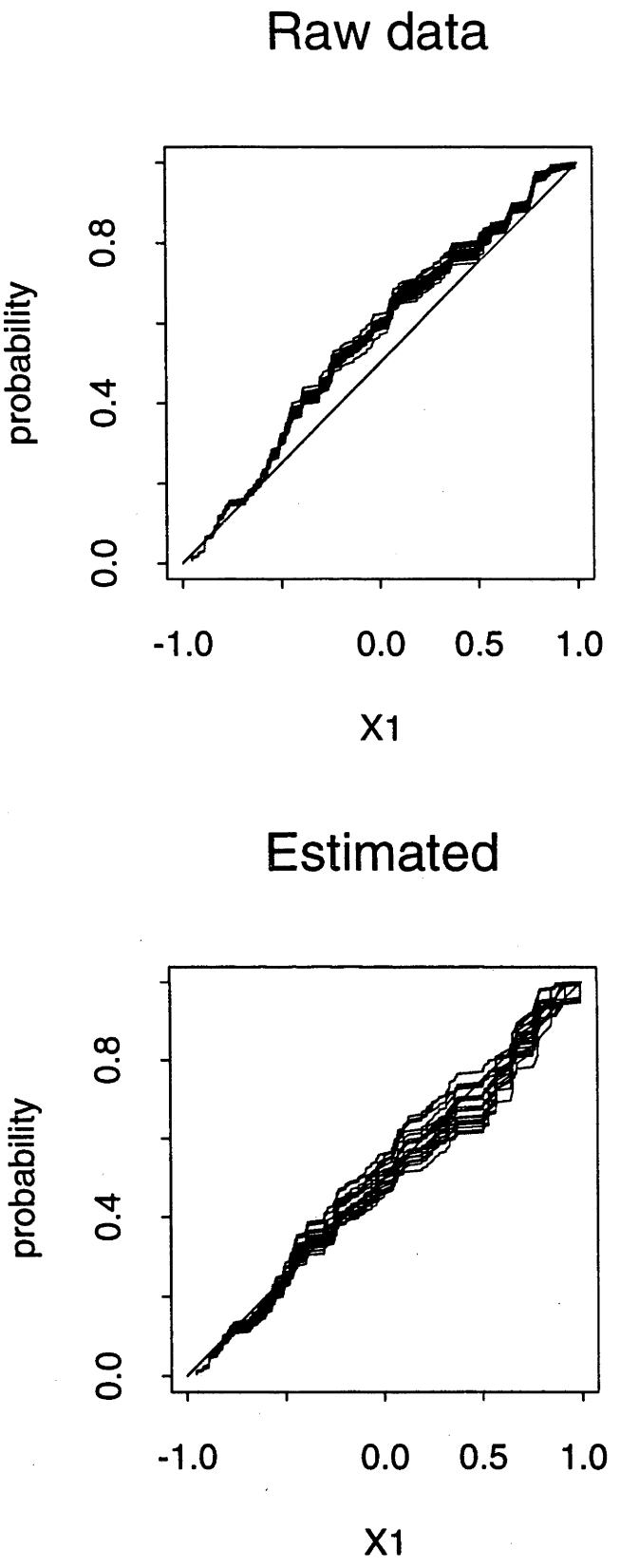


Figure 5.12: The empirical and estimated marginal distribution for the uniform covariate with $\beta = (0, -2, 3)$ over 20 randomly selected simulations. See text for details.

parameter	β_0	β_1	β_2	λ
Nburn	4	3	3	4
Nmin	2457	2029	2029	2029

Table 5.4: Results of the *gibbsit* analysis. Nburn=Number of burn-in iterations required, Nmin= minimum chain length required to obtain specified accuracy.

associated with the quantiles to be within η of the stationary distribution. The results are shown in Table 5.4

The results show that the chain is rapidly mixing and that the 2000 iterations performed should provide reasonably accurate estimates of the quantiles of the distribution. This is also shown in the sample autocovariance functions which are not shown. In addition the burn-in period required is small. We note though that Gelman & Rubin (1992b) have been very critical of this approach and presented a number of examples where it has failed.

The second technique used to assess the behaviour of the chain is given in Figure 5.13. This plots the convergence of the mean of the distribution and the running mean, for a number of simulations commencing from widely dispersed initial values, for a data set from simulation 3. The result for λ is shown. Figure 5.13 is in agreement with the previous results, and demonstrates the lack of long term dependence in the sequence. This is not unexpected as the joint posterior for the parameters is unimodal and there are no extreme dependencies in the set of conditional distributions.

5.5.4 Example 3

The Gibbs sampling algorithm was used to fit the Poisson model to some truncated count data on the abundance of Leadbeaters Possum, an Australian marsupial. This data consists of counts of possums and habitat variables from a sample of sites. This data has been analysed in the context of zero inflation. With zero inflation the number of zero responses is greater than would be expected from the Poisson model. Welsh, Cunningham, Donnelly, & Lindenmeyer (1994) analyse this data by assuming that the data is contaminated by zero responses. They proceed by modelling the probability of a positive response. For the positive responses they modelled the effect of the covariates on the response, conditional on a positive response, via the truncated model.

The data set consists of counts of possums on 151 three hectare sites. The environmental variables measured were *stags* which is a count of suitable habitat trees, *bark* which is a score for the amount of decorticating bark, *no_s* which measures the number of shrubs at the site, and *slope* which is a measure of the slope of the site. After sites with no possums were truncated 55 observations remained.

The Gibbs sampling algorithm allows for the fitting of the truncated model, where we may consider λ and the η's as auxiliary variables. In this case the technique replaces maximum likelihood for the estimation of the regression effects. Alternatively we can consider the following approach. In this analysis we assume that there is an unobserved covariate with two levels. If the covariate is at the first level then the habitat is unsuitable for Leadbeaters Possums and none will

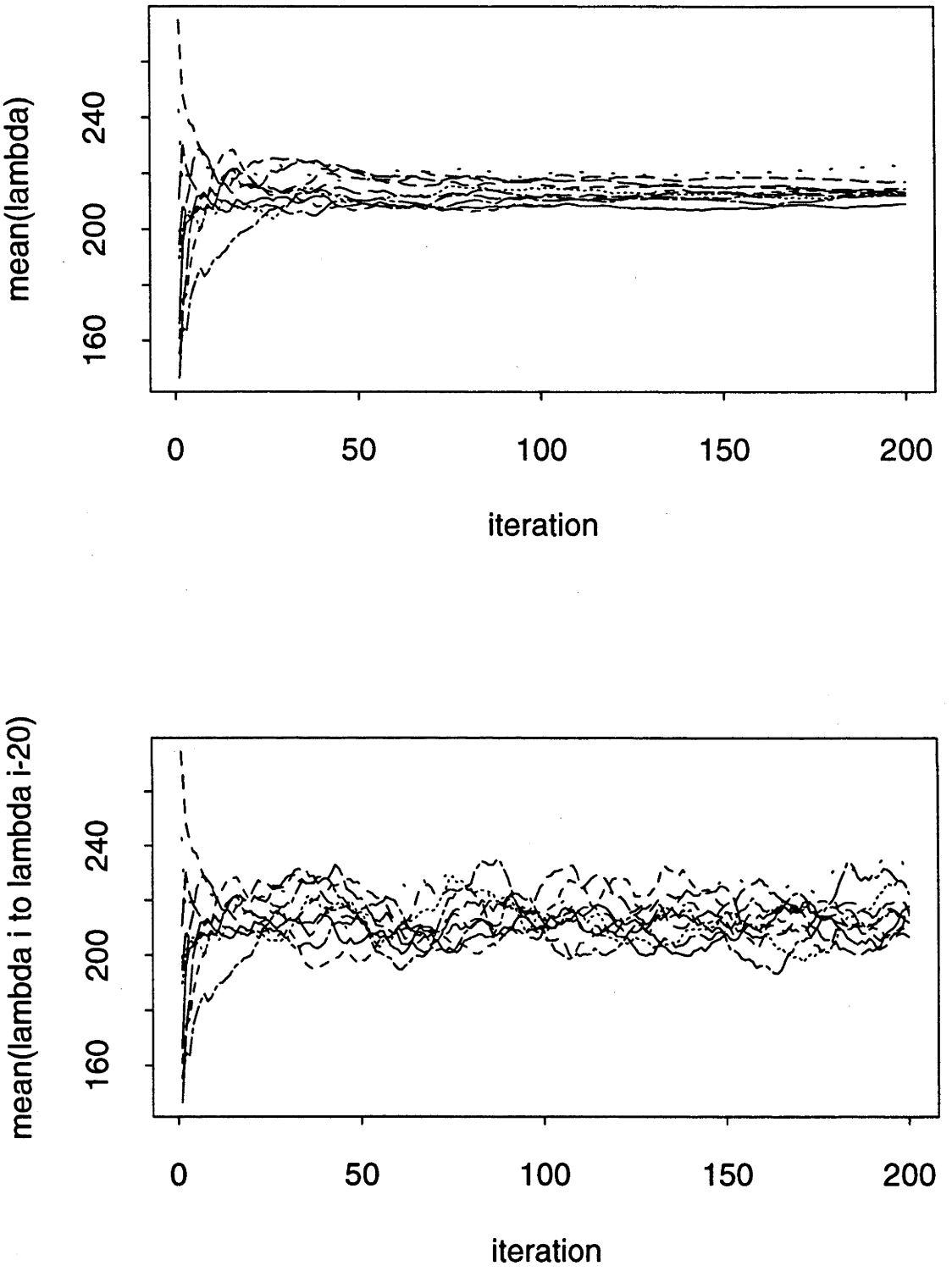


Figure 5.13: Top panel plots $(\sum_1^i \lambda_i)/i$, bottom panel plots $(\sum_{i-20}^i \lambda_i)/20$ for multiple sequences, conditional on a data set from simulation 3.

variable	estimated posterior mode	estimated variance	Truncated MLE	estimated variance
Intercept	1.08	0.17	1.13	0.11
log(stags+1)	0.247	0.019	0.246	0.012
bark	0.040	0.0003	0.037	0.0002
no.s	-.08	0.001	-0.09	0.0008
slope	-0.036	0.0002	-0.031	0.00016

Table 5.5: Parameter estimates and standard errors for possum data.

be found. If the covariate is at the second level the habitat is suitable and the density of the possums follows a Poisson distribution with mean depending on the habitat variables via the log link. The analysis will then enable inference to be made regarding the number of suitable sites in the sample.

The variables chosen were the same as those used by Welsh et al. (1994). These were chosen by the standard analysis of deviance used in fitting generalised linear models.

The Gibbs sampler was run for 20000 iterations, after an initial 100 iterations to limit the effect of the initial conditions. As in the previous example a Gibbs sub-chain was used to sample η. The maximum likelihood estimates were used as the initial parameter values. Other starting points were tried. These had no effect on the results, with the exception of initially upsetting some of the tuning in the rejection sampling algorithm.

The results from the analysis are presented in Table 5.5, along with the maximum likelihood estimates. The approximate marginal posterior distribution for N is given in Figure 5.14. This was generated by taking samples at intervals of 50 from the Gibbs sequence and drawing from the conditional distribution of N. Note that the estimated mode is approximately 60 and thus the intensity of truncation is low.

5.6 Discussion

The techniques presented in this chapter provide a novel approach to the estimation of regression coefficients from observationally truncated data. They also allow for inference to be drawn regarding the truncation process. This is an extension of the usual conditional approach which considers the likelihood of the response *within* the observed sample. The results of the simulation study and example provide encouragement about the stability and interpretability of the estimators. Of course, if real prior information is incorporated the sampling schemes become more complicated. This should not present as many problems here as it does in other situations, where the augmentation and prior are chosen carefully to produce conditional distributions which are easy to sample from. In this case most of the sampling is performed via brute force and modifications would be more easily implemented.

When the data are distributionally truncated the new method offers no advantage over the usual maximum likelihood estimates for a frequentist statistician.

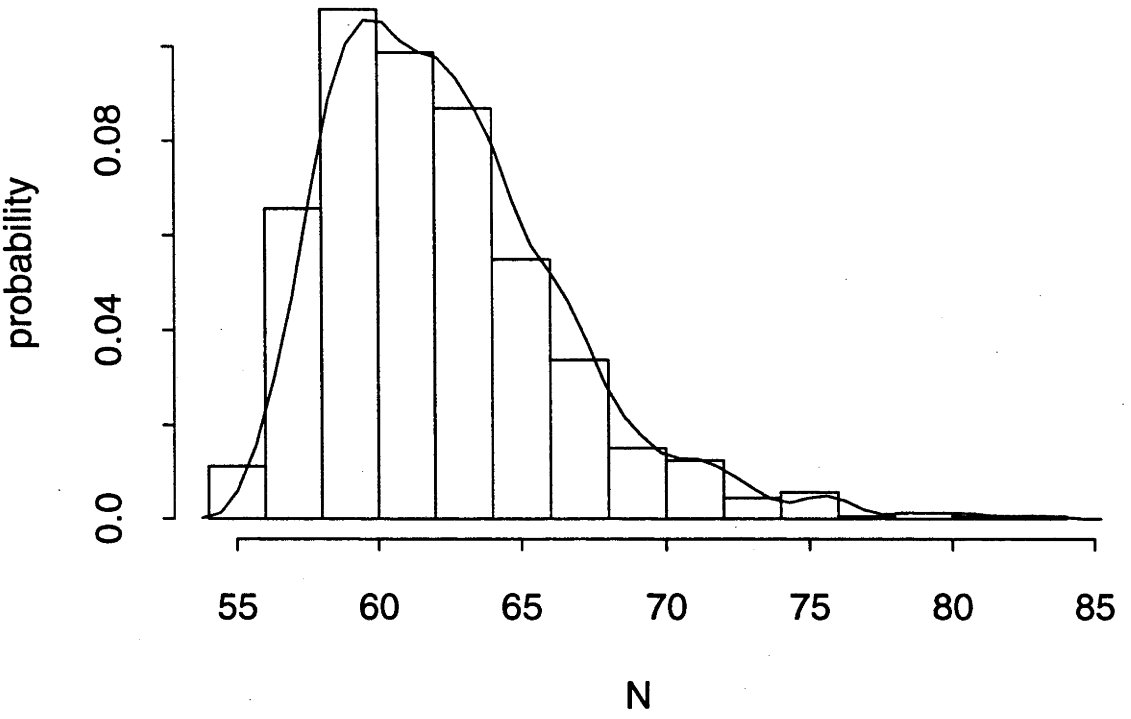


Figure 5.14: Marginal Posterior distribution for N from possum data.

This is because it is only the regression coefficients that are of interest, and the truncation is only used to produce a distribution consistent with the data. Indeed, for certain truncated data the latent structure may have no interpretation. Alternatively, for observationally truncated data, the Gibbs algorithm provides a method for inferring the observational process. This can be important in attempting to estimate the size and distribution of covariates across the unobserved population. This is sometimes a side interest, but can also be the primary focus of the study.

The method is facilitated by the use of the empirical distribution of the covariates. Research is still required into the effect of this approximation on the procedure. It is obvious that in very sparse data sets the empirical distribution may provide a poor approximation of the true distribution, but in this case the empirical prior will represent this uncertainty. Whether there are any deeper pathological problems is unknown at this time, although various unpublished simulations have not highlighted any. Another area of interest is how well the Dirichlet specification reflects the uncertainty in η, and how partial information regarding the covariate distribution can be incorporated into the prior. The work of Ferguson (1973) on nonparametric Bayesian inference may have potential application here.

The use of 'Bayesian' methods in this context is worthy of discussion. One objection to the present explosion in the use of Bayesian methods via the Gibbs sampler is the continued need for tractable priors to make the sampling process simple. Thus the methodology has allowed a wider class of models to be routinely considered, but the choice of prior is still effectively restricted if simple sampling schemes are to be used. There of course exists many specialised techniques for random number generation, but their derivation and implementation can be time consuming. In addition there are alternative Markov Chain Monte Carlo methods

that do not rely on tractable distributions. The questions regarding the validity of expressing subjective prior information in prior distributions remain. If one has a philosophical objection to Bayesian inference with subjective or arbitrary priors the situation has not changed.

The Gibbs sampler coupled with data augmentation is being used to solve problems that are not purely Bayesian. In a pure Bayesian approach, once the prior has been specified and the data observed, the analysis is essentially over. The posterior summarises all information regarding the unknown parameters. It could be argued that some of the literature has drifted away from a purely Bayesian approach. For example, Zeger & Karim (1991) use non informative priors to essentially investigate the likelihood surface. The Bayesian intervals and point estimates produced are invested with an approximate frequentist interpretation, with examination of issues such as bias and coverage. In the simulation given in the paper, the Bayesian analysis is justified in frequentist terms. Asymptotically the duality of the approaches may be justified, if the log likelihood surface converges to a quadratic (Cox & Hinkley, 1974) and the prior is dominated. In finite samples, assumed generated from some distribution with fixed parameters, and uninformative prior distributions, there is no guarantee that the coverage of the 'Bayesian' intervals will be correct. This is of course true for all frequentist inference based on asymptotic results. Still, the simulation results of this chapter, and those of Zeger & Karim (1991) provide encouraging evidence regarding the behaviour of the methods.

In conclusion, the Bayesian analysis is attractive to either a true Bayesian, or to a statistician with relevant prior information. It provides posterior information on all parameters in the model, and thus allows Bayesian intervals for these quantities to be set. In addition it allows the incorporation of additional structure into the model in a natural way. The method has application to other truncation problems such as group truncated ordinal data (O'Neill & Barry (1993a),O'Neill & Barry (1993b)), and data from multiple capture/recapture experiments. It can be potentially extended to include random effects. It could also be used in systems with more general truncation patterns by replacing the term $[Y = 0|x, \beta]$ with the appropriate integral. The sampling algorithms should continue to hold providing the likelihood is approximately quadratic. Alternatively, if interest centres on the regression parameters and the estimation of N then the the frequentist analysis should be considered superior as it has well developed asymptotic theory and is computationally simpler to perform. With uniform priors there has been observed a very high agreement between the estimates produced by each technique in the Poisson case considered here.

Chapter 6

Over-dispersion and the Group Truncated Model

6.1 Introduction

6.1.1 The analysis of correlated binary data

This chapter considers the modelling of group truncated data when we relax the assumption of independence of response within groups, which has been explicitly assumed in previous chapters. This extension is important as it is widely recognised that a large proportion of grouped binary data observed does not exhibit variation consistent with the independence model. To ignore this over-dispersion is to risk underestimating the variability of any estimated quantity. In the truncated case the potential errors are compounded, due to the data being used to essentially extrapolate to the untruncated model.

There has been a proliferation of techniques to model clustered binary data since the early 1980's. This diversity is fuelled by the discrete nature of the response increasing the number of plausible associations that may be of interest, as well as introducing a range of potential measures of association. Thus there are a richer number of associations with potential interest to the researcher than is usually entertained in the Gaussian case. In addition, the marginal and conditional interpretation of mean models, such as those including random effects, do not usually coincide as they do in the standard Gaussian analysis. This is due to the non linearity of the mean functions commonly utilised for non Gaussian responses.

We begin by reviewing techniques available for the analysis of grouped binary data. We then review the techniques that have been used to model over-dispersion in the presence of truncation.

6.1.2 Random Effects Models

Some early attempts at modelling clustered binary data were based on the direct extension of the random effects models used in clustered Gaussian regression (see Harville (1977), Ware (1985)) to the analysis of generalized linear models. This extension was based on models of the form

$$h(\mu_{ij}) = x'_{ij}\beta + b_j$$

where μ_{ij} is the expected value for the ith unit in the jth group, x_{ij} is a vector of associated covariates, β is a vector of fixed effects, b_j is the random effect associated with the jth group and $h()$ is the link function. Thus the random effect is introduced on the scale of the linear predictor, and the regression effects maintain their usual interpretation, conditional on the b_j. The model is largely accepted, due to its success in the Gaussian case, but its use with generalized linear models is restricted since the regression parameters of the conditional and marginal specification have different interpretations as discussed by Zeger et al. (1988). This arises because of the non-linearity of the link function. This is not a feature in the case of a conditionally Gaussian response if the link function is the identity. Thus the conditional and marginal regression parameters, β, are the same in the Gaussian case.

There have been numerous articles on the incorporation of random effects in generalized linear models. Although there are limited examples prior to the 1980's the lack of efficient computing resources seriously limited developments. An early example of a simple model is Sanathanan (1972a). More recent examples considering the random effects model in greater generality is the work of Striatelli, Laird, & Ware (1984) who examines the use of the random effect models in longitudinal binary models, Gilmour, Anderson, & Rae (1985) who consider binomial data from an animal breeding experiment, and Anderson & Aitken (1985) who considers the analysis of binomial data from a clustered survey design. Harville & Mee (1984) and Jansen (1990) have considered the incorporation of random effects in the analysis of ordinal data. A useful discussion of the advantages and disadvantages of the random effect model can be found in Zeger et al. (1988) and Diggle, Liang, & Zeger (1994).

These models are notoriously difficult to fit, as in general no closed form solution exists for the marginal distribution over the random effects. Several approaches have been considered. For binomial models Striatelli et al. (1984) consider empirical Bayes procedures applied via the EM algorithm. Several authors discuss approximation of the integrals by low order Gaussian quadrature such as Anderson & Aitken (1985). Follmann & Lambert (1989) consider the nonparametric estimation of the random effect distribution via the use of nonparametric maximum likelihood.

Various estimation approaches have been developed based on linearization approximations such as Schall (1991) and McGilchrist (1994). Breslow & Clayton (1993) terms these approaches approximate quasi likelihood, or equivalently penalised quasi likelihood. Goldstein (1991), by using a different linearization method, produced what Breslow & Clayton (1993) refers to as marginal quasi likelihood, as the algorithm produces regression parameters which model the marginal expectation of the response. In the case of binomial regression this leads to parameter attenuation with respect to the conditional model. These last methods rely heavily on heuristic arguments and do not have a soundly developed asymptotic theory. In addition, the linearization arguments are inaccurate with binary data and small group sizes. This has been observed in limited simulation studies by Rodríguez & Goldman (1995) and Breslow & Clayton (1993). This is still an active area of research.

6.1.3 Estimating Equations

An alternative line of research that did not directly consider the use of random effects was that popularised by Zeger & Liang (1986) and Liang & Zeger (1986). In these papers the authors recognised that by correctly modelling the marginal mean response they could consistently estimate any regression parameters, even if the covariance of the observations was misspecified. The technique proceeds by specifying the mean function, through an appropriate link, for the marginal mean, and then specifying a 'working' covariance matrix. Any misspecification of the covariance matrix results in loss of efficiency but not consistency. Thus the problem of specifying the full multivariate density is neatly finessed. They recommended using the methods of Royall (1985), to obtain robust estimates of variance, by using sandwich type estimators to provide consistent estimates of the variance of the parameters (but note Drum & McCullagh (1993) for a discussion of this). The consistency of the method depends only on the mean function being correctly specified.

This was achieved by using their Generalized Estimating Equations methodology. The theory of these estimating equations was not new, being closely related to quasi likelihood (Wedderburn (1974), McCullagh (1983)) and the theory of M-estimation (Huber, 1967). It extends the quasi likelihood ideas to multivariate responses, but does not in general produce equations that can be integrated to produce quasi likelihoods (Liang, Zeger, & Qaqish, 1992). In addition, nuisance parameters appear in the estimating equations. The advantage of this approach is that it models the data directly, that is, via the marginal expectation. This is not a feature of the random effects models, which are specified conditionally on the random effects which are a latent structure. The marginal model can be extended in a variety of ways to handle a wide range of correlated responses, for example nested correlation structures such as in Qaqish & Liang (1992). A useful review and discussion is found in Liang et al. (1992) and Diggle et al. (1994).

Concerned with the efficiency of these methods, Prentice (1988) extended the estimating equations to allow joint estimation of the marginal mean and association parameters. This unfortunately leads to inconsistent estimates if the model of association is misspecified. This estimation scheme is further considered by Zhao & Prentice (1990) and Prentice & Zhao (1991) who use a likelihood based approach, and show the equivalence of the resulting score equations to a particular GEE. Further discussion is given in Zhao, Prentice, & Self (1992).

The use of the likelihood of Zhao & Prentice (1990) has been further considered by Fitzmaurice & Laird (1993) for the analysis of longitudinal binary data. They consider a special case of the partly quadratic exponential model described in Zhao et al. (1992) and show that by modelling the marginal mean structure and the conditional odds ratios they can produce estimates that are uncorrelated, thus avoiding the problems with inconsistent estimates. The approaches are discussed in Fitzmaurice, Laird, & Rotnitzky (1993).

6.1.4 Conditional Models

An alternative to the techniques presented above are the conditional models, where the response of each unit in the cluster is modelled conditional on the response of the other units in the cluster. These models have been proposed to handle

specific inferential problems, and the interpretation of the parameters relates to specialised conditional comparisons. Examples of this type of approach are Rosner (1984) who modelled data on the status of people's eyes. This model was extended by Qu, Williams, Beck, & Goormastic (1987) and Connolly & Liang (1988) to arbitrary grouping. Rosner (1989) and Rosner (1992) considered extensions to clustered binary data with more than one level of nesting. Another example is the longitudinal models of Bonney (1987). These models are useful when the conditional means are important as well as estimating the pattern of dependence. Their disadvantages are their complexity and unsuitability for estimating marginal effects due to the complicated marginal distributions they imply.

6.1.5 Truncated models

The incorporation of over-dispersion in truncated regression models has been considered in a number of specialised cases. In the econometrics literature Gurmu (1991) and Grogger & Carson (1991) have considered the truncated negative binomial distribution as an over-dispersed alternative to the truncated Poisson model. Their results are a direct extension of the non-truncated model to the truncated case, and involve the usual likelihood based approach.

In the capture-recapture literature there have been several developments for the analysis of correlated data. Recall that the usual multiple capture-recapture data from static(closed) populations can be considered group truncated data with fixed group size and a distinct ordering of capture times, and hence responses within each group. Sanathanan (1972a) considered a simple random effect model with truncated binary data and covariates fixed across groups, but did not develop it. Huggins (1989) considered several models. In his trap response model, the probability of capture for an individual changes if the individual has been previously captured. The probability is modelled via the logistic link and is in effect a regression on previous observations. In this way the probability model for the response of an individual over the experiment can be considered as a special case of Bonney (1987). These dependence models are appropriate if there is a distinct ordering in the responses within a group, which is clearly not true in the road safety data. Huggins (1989) model could be extended by the consideration of the general models presented by Bonney (1987) but this extension is trivial and will not be considered here.

Huggins (1991) also considers a 'group level' effect model where a term is included in the linear predictor for each group. As it is well known that estimates of these effects are not consistent for finite group sizes, Huggins (1991) does not estimate these. Instead, he derives a score test for the null hypotheses that these terms are all zero.

With paired group truncated ordinal data, Weiss (1993) has developed a latent correlated response model. Weiss considers a paired ordinal response, severity of head injury and body injury, for individuals involved in motorcycle accidents. The severity of these injuries is assumed to follow a bivariate normal distribution with the means and variances being modelled by covariates. The accident severity score is generated via grouping. The accident is truncated if neither injury requires hospitalisation. This model relies on the Gaussian assumption for its tractability, and does not extend easily to larger group sizes. Also, by allowing so many components of the model to vary, the full interpretation of the model is difficult, and

the assessment of the goodness of fit is hampered by the innate flexibility of the hypothesised model allowing a wide range of outcomes to be approximately likely. Also, the data set that the model was applied to exhibited very mild truncation. It would be interesting to examine its performance under more extreme truncation. The model of Weiss is still an important contribution to the statistical literature.

This chapter considers the incorporation of random effects at the group level in the group truncated binary model. Section 6.2 discusses the potential extension of current techniques to the truncated model, develops the truncated random effect model and discusses the novel aspects of the approach, Section 6.3 considers the estimation of the unknown parameters in these models, Section 6.4 considers testing for the presence of random effects, while Section 6.5 presents an example. The technique is discussed in Section 6.6.

6.2 Random Effects Model

6.2.1 Group Truncation

In considering a method to model dependence with group truncated binary data, a number of issues arise. First, we must decide if we are interested in the marginal response after truncation. As an example, we consider groups of fixed size with the only covariate an intercept. The marginal mean in this case is the probability that a unit has response 1 given the group is observed. We can then use the GEE approach of Zeger & Liang (1986) as an estimation technique, to find a consistent estimate of the mean, given a 'working' covariance matrix V^* that we specify in some way. This is

$$U(\beta) = X'(V^*)^{-1}(y - \mu)$$

where X is a column vector of 1's and y is the vector of observed responses. We could then use the method of Royall (1985) to find consistent estimates of the variance. Note that in this case we are modelling the *marginal* response.

The specification of V^* is complicated by the truncation. We note that the truncation causes the marginal responses to be correlated, even under the independence model. Thus, to specify plausible working values for V^*, we must specify both the nature of the dependence in the untruncated sample, and then modify this for the effect of truncation. The technique used by Zeger & Liang (1986) of specifying a simple, plausible form for the second order moments does not work in this case. This is due to the need to determine the effect of truncation on the specification of dependence. Of course, the consistency of the procedure does not rely on the form of V^*, but assigning an arbitrary value will produce extremely inefficient estimates.

The effect of truncation requires the specification of the full likelihood for the untruncated case. Attempts to relax this need for the full likelihood have not been successful. The likelihood is at the centre of the analysis, and our understanding of the data flows directly through it. To see this note that in the binary case the interpretation of truncated data relies on the specification of the probability of the null response. In all practical models considered this leads to specification of the full likelihood model. The case of independent responses within a group is an obvious example.

With arbitrary covariates present the situation becomes impossible. If we believe that truncation is occuring then it is impossible to specify simple models for the marginal means, using the standard link functions and linear predictor. To do so would imply a truncation pattern inconsistent with our belief. In addition, the presence of truncation implies that linearity of treatment effects would rarely occur at the marginal level, due to the marginal mean depending on the covariates of the entire group. Thus the sensible interpretation of the parameters from any such marginal fit is not possible without knowledge of the entire population.

Of course we can use estimating equations, if we specify the marginal mean and variance correctly by using our knowledge of the truncated likelihood and the conditional mean. In this case the contribution to the quasi score equation by the ith group is

$$U(\beta) = \sum_i \frac{\partial \mu_i}{\partial \beta'} V_i^{-1}(y_i - \mu_i)$$

where μ_i is the vector of marginal means implied by the truncated model for the ith group, and V_i is the marginal variance of the response for the ith group. If we assume the responses are independent this reduces to

$$U(\beta) = \sum_i X_i'(y_i - \mu_i)$$

where X_i are the covariates for the ith group. Note that this is just the score equations derived in Chapter 2, as expected. Thus the GEE approach is a useful estimation method, but its use relies on the specification of the likelihood of the data to determine the marginal quantities. Thus it offers no attraction over the full likelihood specification. We will therefore now consider full specification of the likelihood model.

We consider two potential candidates for the likelihood modelling of grouped binary data. The first candidate is the log linear specification (Cox, 1972):

$$f(y_i, \Psi_i, \Omega_i) = exp\{\Psi_i' y_i + \Omega_i w_i - A(\Psi_i, \Omega_i)\}$$

where $w_i = [y_{i1}y_{i2}, \ldots, y_{in_i-1}y_{in_i}, \ldots, y_{i1}y_{i2}\ldots y_{in_i}]$ is the $2^{n_i} - n_i - 1$ vector of two and higher order cross products of y_i, and $\Psi_i = [\Psi_{i1}, \ldots, \Psi_{in_i}]$ and $\Omega_i = [\omega_{i12}, \ldots, \omega_{i12\ldots n_i}]$ are vectors of canonical probabilities. This model allows arbitrary patterns of dependence. The canonical parameters have interpretations in terms of log conditional odds for Ω_i and log conditional odds ratios for Ω_i. This is discussed in detail in Laird (1991) and Fitzmaurice et al. (1993). This model is attractive in that it allows us to fully specify the joint distribution of the response vector. Unfortunately, if we attempt to model the canonical parameters directly by some linear function, the interpretation of the parameters is conditional on specific configurations of the other responses in the group. For example

$$\Psi_{i1} = \log\left[\frac{Pr(y_{i1} = 1 | y_{ij} = 0, j \neq 1)}{Pr(y_{i1} = 0 | y_{ij} = 0, j \neq 1)}\right].$$

In addition, the interpretation of Ω_i is dependent on the group size. Another way of saying this is that the distribution of each unit in the group depends on all the parameters associated with the group. Thus the model is not reproducible, and any analysis performed is dependent on the groups sizes being equal or different parameter sets have to be fitted to different group sizes. The use of this model in

the literature has relied on a transformation of the canonical parameters $[\Psi_i, \Omega_i]$ to a set $[\mu_i, \Phi_i]$ where μ_i are the marginal means, and Φ_i model the association in some way. The marginal mean is then modelled by the usual link function. In these models interest is centred on modelling the marginal mean, with the association parameters considered to be nuisance parameters.

Parametrising the model in terms of the marginal means is not as attractive in the truncated case. In the non-truncated case the marginal means summarise directly the observed data and are thus of considerable practical interest. In the truncated case the choice of model provides a specification of the likelihood, but the specification has no simple relationship to what is actually observed after truncation. Thus this choice of parametrisation is less compelling.

In considering road traffic data it is reasonable to expect that there will be positive association between individuals within a group if we assume that there are missing group level covariates. A method of introducing this positive association is to incorporate random effects at the group level. This is consistent with viewing the random effect as approximating any missing covariates. The model incorporating the random effect is reproducible, and the parameter estimates retain their usual interpretation conditional on the random effect. In the definition of Zeger et al. (1988) this is a subject specific model, as the interpretation of the parameters is relative to the individual units, and not population averages.

We consider the binary case here. The extension to group truncated ordinal data follows naturally from the results presented.

Consider a sample of N groups, each of size n_j, $j = 1 \ldots N$, subject to truncation, where the response for each unit in the group is binary. We consider modelling the mean, conditional on the random effect, of the response for each individual by

$$logit(\mu_{ij}(b_j)) = log(\frac{\mu_{ij}(b_j)}{1 - \mu_{ij}(b_j)}) = x'_{ij}\beta + b_j \tag{6.1}$$

where x_{ij} is a $p \times 1$ column vector of covariates for individual i in accident j, β is a $p \times 1$ column vector of regression parameters, b_j is a random variable with distribution $g(b_j; \sigma^2)$, and σ^2 parametrises the variance. Without loss of generality the mean is assumed zero, as it is absorbed into the intercept term which will always be fitted. Let $\mu_{ij}(b_j)$ be the expected value of y_{ij} conditional on the random effect b_j, with its dependence on x_{ij} and β suppressed for clarity.

If the b_j were known then estimation could proceed following O'Neill & Barry (1993a). In this case the b_j are simple offsets and can be incorporated directly into the design matrix. In the usual case the b_j are unknown and alternative methods of estimation must be considered. There have been several approaches to fitting models with the form of Equation 6.1. Conditional logistic techniques (Breslow & Day, 1980) can be used. These approaches treat the group level effects as nuisance parameters and the conditioning eliminates them from the likelihood. As discussed previously, this approach is unattractive in this context due to the interest in group level parameters, and the complications that arise as the group sizes vary. The second approach is to consider the marginal distribution of the y_{ij} found by integrating over the b_j distribution. This approach has been considered by a number of authors, as discussed in Section 6.1.

Consider the response of the jth untruncated group, $y_j = [y_{1j}, \ldots, y_{n_j,j}]$ which consists of n_j units. The distribution of y_j in the absence of truncation, conditional

on b_j and the x_{ij}, is

$$L(y_j|b_j) = \prod_{i=1}^{n_j} \mu_{ij}(b_j)^{y_{ij}}(1-\mu_{ij}(b_j))^{1-y_{ij}}. \tag{6.2}$$

The truncated likelihood is therefore

$$L_T(y_j|b_j, \text{observed}) = \frac{\prod_{i=1}^{n_j} \mu_{ij}(b_j)^{y_{ij}}(1-\mu_{ij}(b_j))^{1-y_{ij}}}{P_j(b_j)}, \tag{6.3}$$

where $P_j(b_j) = 1 - \prod_{i=1}^{n_j}(1-\mu_{ij}(b_j))$ is the probability that the group was observed, conditional on b_j. The random effect distribution is specified with respect to the latent model 6.1 so it is also truncated by the observational process. The distribution of the observed random effects, b_j, is then given by

$$g_T(b_j|\text{observed}) = \frac{g(b_j;\sigma^2)P_j(b_j)}{\int g(b_j;\sigma^2)P_j(b_j)db_j} \tag{6.4}$$

where the integration is over the support of b_j. The joint distribution of y_j and b_j given that the group is observed is thus

$$L(y_j, b_j|\text{observed}) = L_T(y_j|b_j, \text{observed})g_T(b_j|\text{observed}). \tag{6.5}$$

We proceed by calculating the marginal distribution of y_j which is

$$\begin{aligned} L(y_j|\text{observed}) &= \int L_T(y_j|b_j)g_T(b_j|\text{observed})db_j \\ &= \frac{\int \prod_{i=1}^{n_j} \mu_{ij}(b_j)^{y_{ij}}(1-\mu_{ij}(b_j))^{1-y_{ij}} g(b_j;\sigma^2)db_j}{\int g(b_j;\sigma^2)P_j(b_j)db_j} \end{aligned} \tag{6.6}$$

where the integration is over the support of $g()$.

6.3 Estimation

As is widely known, the integrals in Equations 6.4 and 6.6 are not guaranteed to have explicit solutions. In fact, unless $g()$ is discrete it is usually necessary to estimate the integrals by numerical techniques. We will consider the case where $g(b_j;\sigma^2)$ is a Gaussian distribution with variance σ^2. This can be justified by assuming that there is a number of random, missing covariates, with small effects for each group, and thus a central limit theorem applies. The alternative justification is that this is the most common case presented in the literature, and is thus most well understood. Whichever justification is implicitly or explicitly given the model still allows exploration of the dependence structure of the data. For other random effects distributions the obvious modifications can made to the approach.

For convenience we can rewrite model 6.1 as

$$logit(\mu_{ij}(b_j^*)) = \log(\frac{\mu_{ij}^*(b_j^*)}{1-\mu_{ij}^*(b_j^*)}) = x_{ij}^*\beta^* \tag{6.7}$$

where $x_{ij}^* = (x_{ij}, b_j^*)$, $\beta^* = (\beta, \sigma)$ and $g(b_j^*;1)$ is the standard Gaussian distribution. This form is convenient as it expresses the random effects symmetrically with the covariates, and ensures all integrals are performed across a standard distribution. To simplify notation we will drop the *, and assume the new parametrisation holds. There are several approaches to estimation which will now be considered.

6.3.1 EM algorithm

As the random effects can be considered missing data, the EM algorithm (Dempster et al., 1978) coupled with Gaussian Quadrature can be used to analyse mixed random effect models. Examples of its use are Anderson & Aitken (1985), and Im & Gianola (1988) who considered binary data, Jansen (1990) who considered ordinal data, and Jansen (1993) for Poisson data. The popularity of the algorithm in this case is for several reasons. First, the algorithm exhibits monotonic convergence in the exponential family and thus some of the difficulties inherent in the full maximum likelihood approach do not arise. Second, the M-step can be expressed as an iteratively re-weighted least squares problem and thus performed by statistical packages such as GENSTAT(Genstat 5 Committee). Third, the E-step is performed with a small number of quadrature points and so the size of the computational problem is constrained.

The EM algorithm can potentially be applied here. For simplicity consider a single truncated group and suppress the j subscript. In this case, in the terminology of Tanner (1993), the posterior density for the jth group is

$$f(\beta|y,x,b,\text{observed}) = \frac{\prod_{i=1}^{n_j}\mu_i(b)^{y_i}(1-\mu_i(b))^{1-y_i}}{P(b)}$$

and the predictive density is

$$f(b|y,x,\beta^{(i)},\text{observed}) = \frac{\phi(b)f(\beta^{(i)}|y,x,b,\text{observed})}{\int_{-\infty}^{+\infty}\phi(b)f(\beta^{(i)}|y,x,b,\text{observed})db} \tag{6.8}$$

where $\beta^{(i)}$ is the estimate of β at the ith iteration. Using Gaussian Quadrature we can approximate Equation 6.8 by

$$f(b|y,x,\beta^{(i)}) \approx \frac{\phi(b)f(\beta^{(i)}|y,x,b)}{\sum_{q=1}^{Q}w_q f(\beta^{(i)}|y,x,d_q)} \tag{6.9}$$

where $\phi()$ is the standard Gaussian distribution, Q is the number of quadrature points, w_q are the quadrature weights and d_q are the quadrature nodes. These can be found from Abramowitz & Stegun (1972). The E-step is thus

$$Q(\beta,\beta^{(i)}) = \int_{-\infty}^{+\infty} log(f(\beta|y,x,b,\text{observed}))\frac{\phi(b)f(\beta^{(i)}|y,x,b,\text{observed})}{\sum_{q=1}^{Q}w_q f(\beta^{(i)}|y,x,d_q)}db_j. \tag{6.10}$$

As the density remains the same we can use the same quadrature points to obtain

$$Q(\beta,\beta^{(i)}) \approx \sum_{q=1}^{Q}\frac{w_q f(\beta^{(i)}|y,x,d_q,\text{observed})}{\sum_{r=1}^{Q}w_r f(\beta^{(i)}|y,x,d_r,\text{observed})}log(f(\beta|y,x,d_q,\text{observed})).$$

The M-step involves maximising $Q(\beta,\beta^{(i)})$ with respect to β. $Q(\beta,\beta^{(i)})$ is just a weighted likelihood of simple form and can be maximised by augmenting the design matrix by the quadrature points and applying a Fisher scoring or an equivalent iteratively re-weighted least squares algorithm as described in Chapter 2. Thus the parameters are estimated by the following algorithm:

1. Obtain an initial estimate $\beta^{(1)}$ from the truncated fit assuming no random effects.

2. Set $\sigma = 1$.

3. Calculate $f(\beta^{(i)}|y, x, d_q, \text{observed})$ and recalculate weights.

4. Iteratively maximise $Q(\beta, \beta^{(i)})$ to obtain $\beta^{(i+1)}$.

5. Iterate 3 and 4 until convergence.

The use of the EM algorithm in this context has several disadvantages. First, the algorithm can converge slowly, especially when the number of missing values is large, as in this case. Second, if the random effects are large the initial estimate may be very poor. Third, no estimate of the information is readily available.

An additional problem is that the use of small numbers of quadrature points leads to inaccurate parameter estimates, while the use of large numbers of quadrature points greatly increases the size of the problem. We note though that with nested random effects this approach may be the only one computationally feasible. A further problem is that the convergence of the algorithm is not ensured outside the exponential family (Wu, 1982), although no practical problems have been reported in the random effects literature for a variety of models. It is noted that the use of these approximate techniques was originally driven by the lack of efficient computing resources. This is no longer the case.

Given these points, we will not demonstrate the algorithm. The interested reader is instead directed to the relevant literature. Instead, in the next section we will consider the direct maximisation of the marginal likelihood using accurate numerical approximations to the required integrals.

6.3.2 Maximum Likelihood estimation

The approach we will take in this section is to maximise the likelihood Equation 6.6 directly, without resorting to the EM machinery above. We will maximise the likelihood by using a modified Newton Method. The basis of the technique is that due to the boundedness and smoothness of the conditional likelihood, we can differentiate under the integral sign. Thus if we can perform numerical integration to a high degree of accuracy we can calculate the gradient and curvature of the likelihood at any point in the parameter space. We can then use the Newton-Raphson method if the likelihood is sufficiently quadratic, or a modified Newton method if the likelihood surface is sufficiently irregular to cause the breakdown of the Newton-Raphson method. This approach is of course only computationally feasible with small numbers of nested effects.

Consider the jth group. The derivatives of this group's contribution to the likelihood, $\ell_j = \log L(\beta, \sigma; y_j|\text{observed})$ are found from Equation 6.6 to be

$$\begin{aligned}
\frac{\partial \ell_j}{\partial \beta_k} &= \frac{A(k)}{B} - \frac{C(k)}{D} \\
\frac{\partial \ell_j}{\partial \beta_k \partial \beta_l} &= \frac{E(k,l)B - A(k)A(l)}{B^2} - \frac{F(k,l)D - C(k)C(l)}{D^2}
\end{aligned}$$

where

$$
\begin{aligned}
A(k) &= \int_{-\infty}^{\infty} \sum_{i=1}^{n_j} x_{ijk}[y_{ij} - \mu_{ij}(b_j)]L(y_j|b_j)g(b_j;1)db_j \\
B &= \int_{-\infty}^{\infty} L(y_j|b_j)g(b_j;1)db_j \\
C(k) &= \int_{-\infty}^{\infty} \sum_{i=1}^{n_j} x_{ijk}\mu_{ij}(b_j)[1 - P_j(b_j)]g(b_j;1)db_j \\
D &= \int_{-\infty}^{\infty} P_j(b_j)g(b_j;1)db_j \\
E(k,l) &= \int_{-\infty}^{\infty} \sum_{i=1}^{n_j} x_{ijk}L(y_j|b_j)\{\sum_{m=1}^{n_j} x_{mjl}[y_{mj} - \mu_{mj}(b_j)][y_{ij} - \mu_{ij}(b_j)] \\
&\quad -x_{ijl}\mu_{ij}(b_j)[1 - \mu_{ij}(b_j)]\}g(b_j;1)db_j \\
F(k,l) &= \int_{-\infty}^{\infty} \sum_{i=1}^{n_j} x_{ijk}[1 - P_j(b_j)]\{x_{ijl}\mu_{ij}(b_j)[1 - \mu_{ij}(b_j)] \\
&\quad -\mu_{ij}(b_j) \sum_{m=1}^{n_j} x_{mjl}\mu_{mj}(b_j)\}g(b_j;1)db_j
\end{aligned}
$$

where $L(y_j|b_j)$ is the untruncated likelihood of the data defined in 6.2 and x_{ijk} is the kth covariate for individual i in accident j.

As the responses between groups are independent the score and observed information are

$$
\begin{aligned}
U(\beta) &= \sum_{j=1}^{N} \frac{\partial \ell_j}{\partial \beta_k} \\
I(\beta) &= \sum_{j=1}^{N} \frac{\partial \ell_j}{\partial \beta_k \partial \beta_l},
\end{aligned}
$$

where in this case N refers to the number of truncated groups observed.

Several issues must be addressed in fitting these models. Firstly the integrals used in calculating the score and information must be approximated. The numerical integration used in this chapter was performed using routines from the NAG Fortran library Mark 16. Routines D01BAF which used a 64 point Gaussian quadrature and D01AJF which uses an adaptive routine to produce integrals of the desired accuracy, were used interchangeably in a wide variety of simulated examples and the results were not dependent on the choice of routine. The use of 64 point quadrature was perhaps inefficient, but the algorithm converged rapidly so the inefficiency was not a severe handicap. For applied work, and for problems involving nested random effects, more careful consideration of efficiency considerations may be appropriate. Note that Crouch & Spiegelman (1990) has considered the problem of accurate numerical integration of the type of integrals considered here. Unfortunately their approach is less efficient than Gaussian quadrature (20 point), and the gains in accuracy are not an issue here given the high order numeric integrations used. A second issue is the choice of maximisation technique. A Newton-Raphson algorithm was used, and provided that the initial estimate was sufficiently close to the maximum, convergence was rapid. Unfortunately, poor choice of initial estimate can lead to non convergence and convergence to non maxima. This is discussed further below.

The next issue to be considered is the restriction on the parameter space. The likelihood function is symmetric about the line $\sigma = 0$, and maximisation must be performed only over the parameter space $\sigma = [0, \infty)$. This symmetry at $\sigma = 0$ produces a turning point which a simple Newton-Raphson procedure will converge to. This restriction can be enforced by reparametrising as $\tilde{\sigma} = \log(\sigma)$ and making the obvious modifications to the derivatives. Unfortunately, if the maximum is attained at $\sigma = 0$ then $\tilde{\sigma}$ diverges to $-\infty$. This restriction can also be enforced by using an appropriate maximisation routine. Routines from the NAG library used were E04HAF, which uses a modified Newton algorithm and calculates the derivatives numerically. This numerical differentiation did not introduce major inefficiencies as the number of integrations remains the same provided the likelihood surface is sufficiently well behaved. Any maximum produced by the NAG routine was then checked via a Newton-Raphson algorithm to confirm the result, and also estimate the information matrix. An alternative is to produce a Newton-Raphson algorithm with a step length checker. Discussion of this can be found in Seber & Wild (1989). It is simple to write an algorithm that will converge. The NAG routines were efficient in this case.

With group truncated binary data questions regarding the identifiability of parameters arises. Consider a sample of truncated observations of groups of size two. If we only fit an intercept term then $x_{ij} = (1, b_j)\, \forall i, j$, and $\beta = (\beta_1, \sigma)$. If we observe a sample of N groups and tabulate n_{11}, the number of groups with response of both members 1, and n_{01}, the number of groups with disparate response, it is clear that, conditional on N, n_{11} is a sufficient statistic. It is thus obvious that for any σ a value of β_1 can be found to maximise the likelihood. This is a direct effect of the dependence of the mean on the random effect, and the truncation. The identifiability problem does not occur in the absence of truncation, as in this case conditional on N the minimal sufficient statistic has dimension two. The lack of identifiability obviously extends to truncated paired data with only categorical covariates as explanatory variables. Continuous covariates alleviate the problem, as do group sizes greater than 2. Of course if only a few groups in the sample have sizes greater than two then there will be strong correlation between the parameter estimates.

Another problem that occurs with severe truncation is that all groups have only one positive response. As mentioned previously this is a sparse data problem and will occur with probability zero as the size of the truncated sample increases, provided the probability model is correct. Care must still be taken in maximising the likelihood in finite samples as particular configurations of the response can very occasionally lead to multiple turning points and it is prudent to initiate any iterative procedure from a variety of starting points.

6.4 Inference

We now consider the question of inference in the truncated random effects model. This is of interest for two reasons. First, there is the usual wish to assess the precision of conclusions that are drawn from a statistical model. The second reason is that tests of this model can potentially be used as a partial test of goodness of fit, and a more general test of over-dispersion. This can be seen by noting that with $\sigma = 0$ the responses become independent within each untruncated group. Thus if

we can construct a test of $\sigma = 0$ we can perform a partial test of goodness of fit. The term partial is used as the test is constructed against the random effect model, although modifications below slightly relax this constraint.

As the finite sample behaviour of the maximum likelihood estimates is intractable we will, as is usual, resort to using approximations based on their asymptotic distributions. If the true parameter vector is an interior point of the parameter space, the problem is entirely regular and the maximum likelihood estimates, $\hat{\beta}$ are normal with mean β and approximate variance $I(\hat{\beta})^{-1}$. Thus the usual Wald, score and likelihood ratio tests can be constructed to test hypotheses about subsets of the parameters, subject to the usual mild regularity conditions.

If the parameter σ is on the boundary of the parameter space the problem is non regular and difficulties arise. There has been interest in such cases in the literature. Chernoff (1954) has considered the distribution of likelihood ratio tests for hypothesis of the form $H_0 : \sigma = c, H_a : \sigma > c$, although this work does not specifically address the effect of a parameter on the boundary. In a related series of papers Moran (1971b), Moran (1971a) and Chant (1974) have considered the problem of determining the asymptotic distribution of the MLE when a parameter is on the boundary of the parameter space. The results of Chant (1974) show that under mild regularity conditions the asymptotic distribution is a mixture of truncated normals (see Self & Liang (1987) for corrections to some errors in this paper). Self & Liang (1987) also provide general results about the distribution of the likelihood ratio test, in range of situations with parameters on the boundary of the parameter space. These papers present results which show that under suitable regularity conditions, only the score test retains its usual asymptotic properties.

If we attempt to construct a score test for the hypothesis

$$H_0 : \sigma = 0$$

problems arise due to the score being identically zero on the line $\sigma = 0$ for all values of β. Thus higher order derivatives of the likelihood surface would be required to construct a score type test. Instead we construct the test based on the score for σ^2 which it will be found has desirable properties and interpretation.

Consider a score test of

$$H_0 : \sigma^2 = 0$$

against

$$H_a : \sigma^2 > 0.$$

The Log Likelihood for the jth group can be written from Equation 6.6 as

$$\ell = \log \int L(y_j|b_j)g(b_j;1) - \log \int P(\beta)g(b_j;1)db_j.$$

where $logit(p_{ij}) = x'_{ij}\beta + \sqrt{\sigma^2}b_j$. In the following we will suppress the j (group) subscript. The derivative of ℓ with respect to σ^2 is found by calculating the limit of the derivative of ℓ as $\sigma^2 \downarrow 0$. We find, after differentiation, and with the help of l'Hopitals rule that the limit of the first derivative as $\sigma^2 \downarrow 0$ is

$$\lim_{\sigma^2 \to 0+} \frac{\partial L}{\partial \sigma^2} = \frac{(Y-\mu)^2 - V - \frac{VQ}{P} + \frac{\mu^2 Q}{P}}{2} \tag{6.11}$$

where

$$\begin{aligned} Y &= \sum_i y_i \\ \mu &= \sum_i p_i \\ V &= \sum_i p_i q_i. \end{aligned}$$

This derivative is thus the difference between the observed covariance of the response and that implied by the truncated model under the assumption of independence. The second derivative is found by further differentiation and repeated applications of l'Hopitals rule to be

$$\begin{aligned} &\lim_{\sigma^2 \to 0+} \frac{\partial^2 L}{\partial \sigma^{2 2}} = \\ &\quad [-4V(Y-\mu)^2 + (8W - 4V)(Y-\mu) - 2V^2 - V - 6W + 6W_1]/4 \\ &\quad -[-Q\mu^4 + 6VQ\mu^2 - 4(V-2W)Q\mu - 3V^2Q - 2Q(2W - 3W_1) \\ &\quad +V - 2W]/4P - (V^2Q^2 + \mu^4Q^2 - 2\mu^2Q^2V)/4P^2 \end{aligned} \tag{6.12}$$

with

$$\begin{aligned} W &= \sum_i p_i^2 q_i \\ W_1 &= \sum_i p_i^3 q_i. \end{aligned}$$

To use the standard results we need to find the score and information in the sample evaluated at the restricted MLE estimate, $(\hat{\beta}, \sigma_0^2)$ where $\sigma_0^2 = 0$. Given H_0 the likelihood is maximised by the maximum likelihood estimates assuming independence derived in Chapter 2. The score vector is thus

$$U(\hat{\beta}, 0) = (0, \lim_{\sigma^2 \to 0+} \tfrac{\partial L}{\partial \sigma^2})$$

The information matrix evaluated at the restricted estimate can be partitioned as

$$I(\hat{\beta}, 0) = \left(\begin{array}{cc} I_{\beta,\beta} & I_{\sigma^2,\beta} \\ I_{\beta,\sigma^2} & I_{\sigma^2,\sigma^2} \end{array} \right) \Bigg|_{(\hat{\beta},0)}$$

where $I_{\beta,\beta}$ is the information matrix found from the truncated logistic fit and I_{σ^2,σ^2} is found from Equation 6.12. The elements I_{β,σ^2} can be found by differentiation of Equation 6.11 to be

$$\begin{aligned} \lim_{\sigma^2 \to 0+} \frac{\partial^2 L}{\partial \beta_p \partial \sigma^2} &= -V_x(Y-\mu) - (V_x - 2W_x)/2 \\ &\quad [-2\mu_x PQ(V-\mu^2) \\ &\quad +2PQ(V_x - 2W_x - 2\mu_x V_x) - 2\mu_x Q^2(V-\mu^2)]/4P^2 \end{aligned}$$

with

$$\begin{aligned} \mu_x &= \sum_i x_{ip} p_i \\ V_x &= \sum_i x_{ip} p_i q_i \\ W_x &= \sum_i x_{ip} p_i^2 q_i \end{aligned}$$

where x_{ip} is the covariate associated with β_p for the ith individual. The test statistic is then

$$X = U(\hat{\beta}, 0)' I(\hat{\beta}, 0)^{-1} U(\hat{\beta}, 0)$$

which under the usual regularity condition for regression models which ensures the expected information converges, has asymptotically, as the number of observed groups goes to infinity, a χ_1^2 distribution.

As σ^2 contains a single parameter we can consider the equivalent test

$$Z = \frac{\lim_{\sigma^2 \to 0+} \frac{\partial L}{\partial \sigma^2}}{(I_{\sigma^2,\sigma^2} - I_{\sigma^2,\beta} I_{\beta,\beta}^{-1} I_{\beta,\sigma^2})^{-1}}$$

which has asymptotically a standard normal distribution. Retention of the sign in this case can be used as an empirical diagnostic to detect over or under dispersion. Note that the random effect model can only imply positive association. Hence if $Z < 0$ there is no evidence to support H_0. Thus Z can be used to construct one sided tests, and the tests will therefore be more powerful than those constructed via X which takes no account of the direction of the deviation. In addition negative values of Z provide evidence of under-dispersion.

6.5 Examples

6.5.1 Example 1

6.5.1.1 Simulation design

In this section results from a simulation study are presented. The simulation study is designed to compare the performance of the truncated random effects model(TRE) introduced in Section 6.2, the truncated model assuming independence(TLR), the conditional logistic model(CLR) and the use of ordinary logistic regression(OLR) ignoring the truncation. As the independence truncated model is conceptually and computationally simpler we wish to investigate the gains produced by incorporating the random effects. We will use the bias and standard error of the estimates to assess their behaviour. The design of simulation experiments investigating regression models is always complicated by the choice of the design matrix. With truncated data further difficulties arise.

One approach is to simulate data from the non-truncated distribution, then truncate the data set and apply the estimation techniques. This will be referred to as marginal simulation and is the technique used in Shaw (1988) and Grogger & Carson (1991). Alternatively, truncated data can be generated from a fixed design matrix. This is a conditional simulation and an example is the simulations

performed in Tsui et al. (1988). Both options have merits. The marginal option is attractive if we view the data generation process in this way and wish to examine the sampling distribution of our estimates with respect to this process. This allows the effect of the observational process to be assessed. The conditional option is attractive as it estimates the finite sampling distribution conditional on the design space which is how the truncated estimators are derived, and hence the simulated behaviour of the estimates can be compared to their asymptotic approximations. In addition, as the sample size is conditioned upon, problems of sparse data sets do not arise as often when there is extreme truncation if n is chosen to be moderate. Also, the results from the simulation are not confounded with the truncation process. We choose to use the second method, though accepting that there are situations where the results of the first technique are of interest. An example is in the estimation of N previously discussed, where the sampling distribution of $\hat{N}$ relates directly to the marginal distribution. Of course, a hybrid between the two techniques may be applied, where we draw from the marginal distribution, and then replicate the conditional distribution, but this soon becomes computationally prohibitive.

We consider two simulations. The first is generated assuming 200 groups of size two. Groups of size two possess the least information regarding the random effect. This group size also allows the easiest investigation of the conditional logistic technique, although this is complicated by the number of death discordant pairs being random, and thus we are investigating the marginal behaviour of the conditional technique over the fixed, truncated, sample.

The second simulation assumes 100 groups of size 4. Hence this simulation uses the same number of individuals as the first simulation, but with larger group sizes. The larger group size is used as it provides more information about the specific random effects, thus allowing assessment of the stability of the estimates in this case.

The covariate design for the two simulations was identical. For each individual an intercept term, a 0/1 variable and a continuous U(0,1) covariate at the level of the group was generated. Using the conditional simulation method we will assume that the covariates are independent, justifying this by arguing that as we are only considering a particular slice of the marginal distribution, the choice of covariates is not of primary interest. The covariates for each individual i in the jth group, were generated as

$$[1, x_{j1}, x_{ij2}]$$

where x_{j1} had a uniform distribution over [0,1] at the group level and x_{ij2} was Bernoulli with probability .5. The probability of response for each individual in the jth group, p_{ij}, was modelled at the untruncated level by

$$logit(p_{ij}) = \beta_0 + x_{j1}\beta_1 + x_{i2}\beta_2 + b_j\sigma$$

with the marginal distribution of y found from the integral 6.6 used to produce the required sample.

The simulation was performed over a grid of parameter values, which were chosen to imply a range of truncation intensities and levels of random effects. The coefficient values for the intercept, β_0 was chosen to be -2, 0, the coefficients for x_{i1} to be $0, 2$ and the coefficient for x_{i2} was fixed at .5. The random effect standard deviation was taken to be σ either 0, 1, or 4, which correspond to a range of random effects from none to extreme.

For each parameter combination the simulation was iterated 500 times.

6.5.1.2 Simulation Results

The simulation results for the regression parameters are presented in Tables 6.1 to 6.18. These tables summarise the sampling distribution of the regression parameters for the various assumed models, and compares them with the asymptotic approximations derived previously. The confidence intervals were set assuming that the parameter vector was an interior point of the parameter space. This is true except for the simulations with $\sigma = 0$. In this case the asymptotic sampling distribution is a mixture of distributions, resulting from the conditional components relating to $\hat{\sigma} = 0$ and $\hat{\sigma} \neq 0$. This was not done for the sake of simplicity. In the Tables POW denotes the estimated power of the test that the random effect variance was zero, while SC is the estimated type I error rate when the null hypothesis is correct.

The coverages are all calculated with respect to the parameters of the underlying model. This is in some sense an unfair comparison, as the interpretation of the parameters in each model is potentially different. For example, TRE and CLR estimate the underlying parameters, while TLR approximates the marginal distribution over the random effects, and OLR attempts to model the marginal distribution of the observed response. These last two methods are of course biased estimates due to the effect of the truncation. In addition the reassuring results of Neuhaus, Kalbfleisch, & Hauck (1991) and Zeger et al. (1988) showing that the approximate linearity of regression effects in the conditional model implies approximate linearity of the regression effects on the marginal scale, albeit attenuated, do not hold as the truncation introduces complicated non linearities. Nevertheless, it can still be argued that the results present a useful comparison of the techniques, and highlight the problems that can arise if an inappropriate technique is used to estimate the parameters of the latent model. Of course in specific circumstances other measures may be more appropriate, such as assessing how well the TLR model approximates the marginal distribution of the non-truncated response. Questions such as these are most easily answered via unconditional simulation and are thus not presented here. The sampling distributions of the variance estimates are given in Figure 6.1 for groups of size two and Figure 6.2 for groups of size four. The final point to note is that comparison of the information in each sample is done conditional on the fixed truncated design space, and is thus not dependent on the level of truncation affecting the size of the sample observed.

The simulations produce several main conclusions. First, in the case of paired data, there is very little information regarding the random effects in samples of the size considered here. This has serious implications for capture-recapture estimation, as it effectively means that there is little power in the test of significance of the random effects term. Thus the independence model will be nearly always accepted, leading to potential errors in the estimation of the population size. Increasing the group size to 4 increases the power considerably, even though there are in fact half the observations on the random effects. This increase is due to the greatly increased information regarding each unobserved random effect.

Second, as expected there are signs of the attenuation of coefficients in the essentially marginal models, TLR and OLR. This can be observed in all simulations with $\sigma > 0$. This is identical to that seen in the non-truncated model, but as noted these estimates will be biased estimates of the true marginal relationship. The third point to note is that the parameter estimates for the group level effects are

			TRE			TLR			OLR			CLR
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2
-2	.5	.5	-9.36	1.51	.53	-2.07	.58	.51	-.08	.056	.48	.51
-2	2	.5	-7.86	4.81	.52	-2.07	2.10	.50	-.05	.29	.47	.49
0	.5	.5	-1.8	1.3	.55	.01	.51	.49	.59	.27	.48	.50
0	2	.5	-1.7	4.42	.56	0.01	2.03	.51	.63	1.10	.51	.50

Table 6.1: Estimated mean of sampling distributions for groups of size 2 with $\sigma = 0$.

extremely variable when there are random effects present. This is due both to the added variability in the system, and to the strong dependence between the group level estimates and the random effects variance. This can be seen in Figures 6.3 to Figure 6.6 which plot the pairwise and marginal sampling distributions for the parameter estimates when the true parameter values were [0,.5,.5, 1] for each of the two group sizes. It must be remembered that the comparison in this case is complicated by the competing effects of number of groups and probability of observation. In these figures note the highly skewed sampling distribution in the simulations with group size two presented in Figure 6.5. Also note in Figure 6.5 the strong dependence between the intercept term and the estimate of the random effect variance. This simulation has been checked closely to ensure that the algorithm was in fact converging to a unique maximum. No problems were detected. This strong correlation is negligible in the simulation with group sizes four shown in Figures 6.3 and 6.4, and the estimates appear well behaved.

A point to note is the skewed sampling distributions which impact on the bias and variability of the estimates. In addition there is positive biases in the marginal estimates produced by both TLR and OLR due to the greater prevalence of groups with fully positive response when the random effects variance is large. This is a novel effect of the truncation highlighting the meaningless interpretation of the associated parameter estimates in the presence of truncation and extra binomial variation.

As expected TLR performed well when $\sigma = 0$ and hence the independence model holds. In this case the TRE model performs adequately provided the group sizes are 4, and the level of truncation is moderate. If the level of truncation is high there are a large number of groups with only one or two positive responses. These groups have little information regarding the random effects that cannot also be explained via the truncation, and this leads to instability in the parameter estimates. This has been observed in a number of simulations, where the likelihood surface has a very flat ridge across the group level parameter and random effect variance dimensions. As observed previously, in the case of paired data with only an intercept term, the likelihood is flat along this ridge.

The CLR approach worked well in all cases.

The individual level effect was estimated effectively by all techniques provided the random effects variance and truncation were not extreme. The estimation techniques were effectively unbiased for this parameter except when the level of truncation is high. In this case the sampling distribution is highly skewed introducing biases into the estimates.

			TRE			TLR		
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
-2	.5	.5	414(99)	12.89(10.16)	.047(.046)	.092(.098)	.84(.93)	.044(.043)
-2	2	.5	100(85)	33(24.6)	.047(.050)	.11(.12)	.95(1.02)	.044(.047)
0	.5	.5	21.7(11.1)	5.42(5.37)	.062(.060)	.033(.037)	.23(.25)	.050(.049)
0	2	.5	16.7(9.3)	32.7(17.47)	.063(.072)	.035(.039)	.27(.25)	.051(.060)

Table 6.2: Mean of estimated variances for groups of size 2 with $\sigma = 0$. Observed variance given in parentheses.

			TRE			TLR			OLR			CLR	SC
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2	$\alpha = .05$
-2	.5	.5	.83	.97	.956	.96	.954	.948	0	.942	.958	.952	.003
-2	2	.5	.85	.925	.948	.960	.958	.942	0	0	.944	.948	.01
0	.5	.5	.904	.946	.952	.932	.944	.958	.006	.968	.958	.942	0
0	2	.5	.896	.924	.930	.938	.958	.920	.016	.250	.918	.948	0

Table 6.3: Estimated coverage of approximate 95% confidence intervals, and coverage of $\alpha = .05$ score test of $\sigma = 0$ of sampling for groups of size 2 with $\sigma = 0$.

			TRE			TLR			OLR			CLR
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2
-2	.5	.5	-7.21	1.18	.52	-1.14	.34	.47	.12	.08	.45	.50
-2	2	.5	-6.7	4.33	.53	-1.16	1.46	.49	-.12	.40	.47	.52
0	.5	.5	-1.76	1.08	.54	.29	.42	.45	.77	.25	.45	.50
0	2	.5	-1.61	3.95	.54	.31	1.48	.45	.80	.91	.45	.51

Table 6.4: Estimated mean of sampling distributions for groups of size 2 with $\sigma = 1$.

			TRE			TLR		
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
-2	.5	.5	80.0(75.5)	9.45(8.05)	.050(.051)	.048(.45)	.40(.38)	.044(.043)
-2	2	.5	66.6(58.5)	26.9(19.7)	.051(.049)	.052(.051)	.44(.46)	.044(.043)
0	.5	.5	23.1(8.06)	6.69(3.19)	.071(.069)	.032(.033)	.23(.21)	.053(.053)
0	2	.5	19.43(7.04)	29.1(12.7)	.072(.073)	.034(.034)	.24(.22)	.054(.055)

Table 6.5: Mean of estimated variances for groups of size 2 with $\sigma = 1$. Observed variance given in parentheses.

			TRE			TLR			OLR			CLR	POW
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2	$\alpha = .05$
-2	.5	.5	.426	.944	.956	.038	.942	.946	0	.904	.950	.960	0
-2	2	.5	.488	.866	.962	.084	.834	.962	0	0	.964	.970	0
0	.5	.5	.800	.970	.956	.628	.946	.956	0	.960	.950	.960	.008
0	2	.5	.790	.870	.956	.604	.814	.950	0	.134	.950	.954	.027

Table 6.6: Estimated coverage of approximate 95% confidence intervals and power of test of $\sigma = 0$ with $\alpha = .05$ for groups of size 2 with $\sigma = 1$.

			TRE			TLR			OLR			CLR
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2
-2	.5	.5	-1.79	.56	.50	.96	.12	.33	-1.25	.09	.32	.51
-2	2	.5	-1.61	1.68	.51	.96	.37	.34	1.25	.27	.34	.51
0	.5	.5	-.024	.51	.53	1.40	.11	.33	1.59	.087	.33	.52
0	2	.5	-.036	1.97	.52	1.42	.47	.34	1.60	.38	.33	.52

Table 6.7: Estimated mean of sampling distributions for groups of size 2 with $\sigma = 4$.

			TRE			TLR		
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
-2	.5	.5	12.48(5.21)	6.90(6.26)	.11(.11)	.036(.038)	.23(.22)	.065(.062)
-2	2	.5	8.81(4.15)	6.98(7.47)	.11(.12)	.036(.042)	.23(.25)	.066(.077)
0	.5	.5	6.11(2.04)	6.63(6.95)	.14(.15)	.042(.043)	.27(.29)	.082(.087)
0	2	.5	5.69(1.92)	7.71(6.90)	.14(.17)	.042(.044)	.27(.26)	.082(.090)

Table 6.8: Mean of estimated variances for groups of size 2 with $\sigma = 4$. Observed variance given in parentheses. Note that the large disparities are due to the highly skewed sampling distribution.

			TRE			TLR			OLR			CLR	POW
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2	$\alpha = .05$
-2	.5	.5	.710	.948	.952	0	.880	.902	0	.872	.902	.958	.027
-2	2	.5	.702	.722	.928	0	.082	.890	0	.004	.888	.952	.028
0	.5	.5	.646	.940	.940	0	.874	.906	0	.876	.904	.954	.084
0	2	.5	.680	.798	.940	0	.162	.910	0	.062	.910	.954	.096

Table 6.9: Estimated coverage of approximate 95% confidence intervals and power of test for $\sigma = 0$ with $\alpha = .05$ for groups of size 2 with $\sigma = 4$.

			TRE			TLR			OLR			CLR
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2
-2	.5	.5	-2.27	.58	.50	-2.03	.54	.49	-1.04	.13	.49	.49
-2	2	.5	-2.22	2.20	.48	-2.01	2.03	.48	-1.00	.58	.48	.48
0	.5	.5	-0.01	.54	.50	-0.02	.52	.49	.10	.43	.49	.50
0	2	.5	-.02	2.13	.51	0.00	2.05	.51	.12	1.68	.51	.52

Table 6.10: Estimated mean of sampling distributions for groups of size 4 with $\sigma = 0$.

			TRE			TLR		
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
-2	.5	.5	.68(.38)	.64(.62)	.049(.047)	.06(.06)	.48(.52)	.048(.046)
-2	2	.5	.54(.36)	.87(.72)	.042(.045)	.068(.062)	.53(.51)	.047(.043)
0	.5	.5	.025(.029)	.16(.15)	.042(.045)	.023(.026)	.14(.13)	.042(.043)
0	2	.5	.030(.027)	.23(.21)	.046(.046)	.025(.025)	.17(.17)	.045(.044)

Table 6.11: Mean of estimated variances for groups of size 4 with $\sigma = 0$. Observed variance given in parentheses.

			TRE			TLR			OLR			CLR	SC
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2	$\alpha = .05$
-2	.5	.5	.972	.956	.970	.964	.948	.970	0	.954	970	.97	.002
-2	2	.5	.971	.963	..947	.968	.958	.948	.000	.002	.948	.946	.006
0	.5	.5	.939	.963	.949	.930	.952	.950	.898	.966	.950	.960	.012
0	2	.5	.956	.952	.946	.952	.936	.942	.872	.872	.936	.968	.010

Table 6.12: Estimated coverage of approximate 95% confidence intervals, and coverage of $\alpha = .05$ score test of $\sigma = 0$ of sampling for groups of size 4 with $\sigma = 0$.

			TRE			TLR			OLR			CLR
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2
-2	.5	.5	-2.15	.58	.50	-1.26	.38	.47	-.75	.17	.47	.51
-2	2	.5	-2.18	2.08	.49	-1.26	1.41	.46	-.72	.66	.46	.51
0	.5	.5	-.037	.54	.50	-.15	.37	.43	.23	.32	.43	.51
0	2	.5	-.018	2.04	.50	.17	1.44	.44	.26	1.23	.44	.50

Table 6.13: Estimated mean of sampling distributions for groups of size 4 with $\sigma = 1$.

			TRE			TLR		
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
-2	.5	.5	1.20(.94)	.82(.64)	.049(.049)	.036(.42)	.25(.28)	.044(.045)
-2	2	.5	1.47(.97)	1.39(1.16)	.049(.052)	.039(.045)	.27(.35)	.043(.046)
0	.5	.5	.079(.077)	.40(.38)	..054(.062)	.023(.027)	.14(.18)	.043(.048)
0	2	.5	.085(.079)	.56(.52)	.056(.056)	.024(.029)	.15(.21)	.045(.044)

Table 6.14: Mean of estimated variances for groups of size 4 with $\sigma = 1$. Observed variance given in parentheses.

			TRE			TLR			OLR			CLR	POW
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2	$\alpha = .05$
-2	.5	.5	.880	.958	.960	.07	.748	.941	0	0	.945	.935	.192
-2	2	.5	.886	.929	.953	.07	.745	.941	0	.004	.945	.935	.229
0	.5	.5	.968	.964	.932	.786	.904	.918	.654	.90	.918	.934	.60
0	2	.5	.944	.944	.952	.784	.672	.934	.598	.414	.934	.962	.58

Table 6.15: Estimated coverage of approximate 95% confidence intervals and power of test of $\sigma = 0$ with $\alpha = .05$ for groups of size 4 with $\sigma = 1$.

			TRE			TLR			OLR			CLR
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2
-2	.5	.5	-2.01	.51	.53	-.55	.07	.33	.59	.06	.33	.54
-2	2	.5	-1.95	2.25	.50	.56	.42	.31	.60	.39	.31	.51
0	.5	.5	-.029	.46	.51	.99	.08	.31	1.00	.08	.31	.52
0	2	.5	.027	1.99	.50	.99	.43	.30	1.01	.41	.30	.52

Table 6.16: Estimated mean of sampling distributions for groups of size 4 with $\sigma = 4$.

			TRE			TLR		
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
-2	.5	.5	1.34(2.43)	4.15(8.52)	.084(.083)	.024(.036)	.14(.31)	.047(.047)
-2	2	.5	1.43(2.36)	4.41(8.62)	.083(.083)	.024(.033)	.14(.25)	.047(.048)
0	.5	.5	.71(1.31)	4.13(7.63)	.11(.11)	.028(.041)	.16(.31)	.056(.057)
0	2	.5	.65(1.23)	3.88(7.74)	.104(.097)	.028(.036)	.16(.264)	.056(.054)

Table 6.17: Mean of estimated variances for groups of size 4 with $\sigma = 4$. Observed variance given in parentheses. Note that the large disparities are due to the highly skewed sampling distribution.

			TRE			TLR			OLR			CLR	POW
β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_2	$\alpha = .05$
-2	.5	.5	.776	.838	.960	0	.706	.876	0	.708	.876	.974	1
-2	2	.5	.774	.860	.958	0	.054	.856	0	.026	.858	.960	1
0	.5	.5	.754	.864	.958	0	.748	.878	0	.746	.878	.944	1
0	2	.5	.758	.862	.962	0	.048	.882	0	.040	.882	.952	1

Table 6.18: Estimated coverage of approximate 95% confidence intervals and power of test for $\sigma = 0$ with $\alpha = .05$ for groups of size 4 with $\sigma = 4$.

6.5.2 Road Traffic Data

This example will be based on data from the Fatal Accident Reporting System. The data set considered in Chapters 2 and 3 is revisited. We recall that the data consists of frontal impact collisions involving passenger cars. It consists of 306 individuals involved in 111 accident for the years 1988 and 1990, and contains 138 fatalities. The distribution of the group sizes is given in Figure 6.7. We note that there are 45 groups with greater than 2 individuals, and hope that this will allow reasonable identification of the intercept and random effects variance.

For illustration purposes, we first consider the fitting of the model with individual effects but no group level effects. If we fit the independence model with *age*, *sex*, *restraint use*, and *seating location* we obtain a log likelihood of -159.93 and an estimate of the intercept of 1.76. The score test for the hypothesis of no random effect gives a score statistic of $Z = 2.32$ so there is good evidence to reject the independence model. We thus fit the random effects model which gives an estimate $\hat{\sigma} = 3.46$ with an estimate of the intercept term of 11.1 and a log likelihood of -152.66. Thus the model implies a large random effect, which is not surprising given the absence of group level effects. In theoretical terms the estimates produced by the independence model are attenuated, but practically it is hard to interpret this result. What the data implies is that there is a large effect not taken into account by the model. It provides evidence that the marginal probability of dying in an accident relies heavily on unmeasured factors. Before accepting this conclusion, we also note that the simulation results demonstrate that the distribution of parameter estimates can be highly skewed.

We now consider the addition of group level factors, and predict that this should lower the random effect variance, as the omission of significant group level effects will cause unexplained clustering of like responses. We include the variables *damage* and *estimated speed*. The independence model produces a likelihood of -158.28 so although there is some evidence that these variables affect fatality the change in deviance is not significant at the 5% level. The score test gives a statistic of $Z = 2.05$ so there is still significant lack of fit in the model. If we fit the random effects model we obtain an estimate of $\hat{\sigma} = 2.98$. As expected this is reduced from the previous model but is still extreme.

We next consider the addition of the variable *speedcat* instead of the estimated speed. Fitting the independence model we get a log likelihood of -153.98, and the score test is $Z = 1.62$. Thus this new variable is highly significant. There is still strong evidence of a random effect, but it is not significant at the 5% level. If we fit the random effect model we obtain a log likelihood of -150.01 with an estimate of $\hat{\sigma} = 2.27$. We present this series of models in Table 6.19.

To appreciate the impact of the truncation on the random effects we can investigate the distribution of the observed random effects. To do this we note that the conditional distribution of the random effects, given a groups covariates X and that the group is observed is, from Equation 6.4,

$$g_T(b_j|\text{observed}) = \frac{g(b_j;\sigma^2)P_j(b_j)}{\int g(b_j;\sigma^2)P_j(b_j)db_j}.$$

If we assume that the untruncated covariates x are derived from a density $h(x)$ we can then consider the joint distribution of the random effects and covariates, given that they are observed. Thus the observed covariates are a sample from

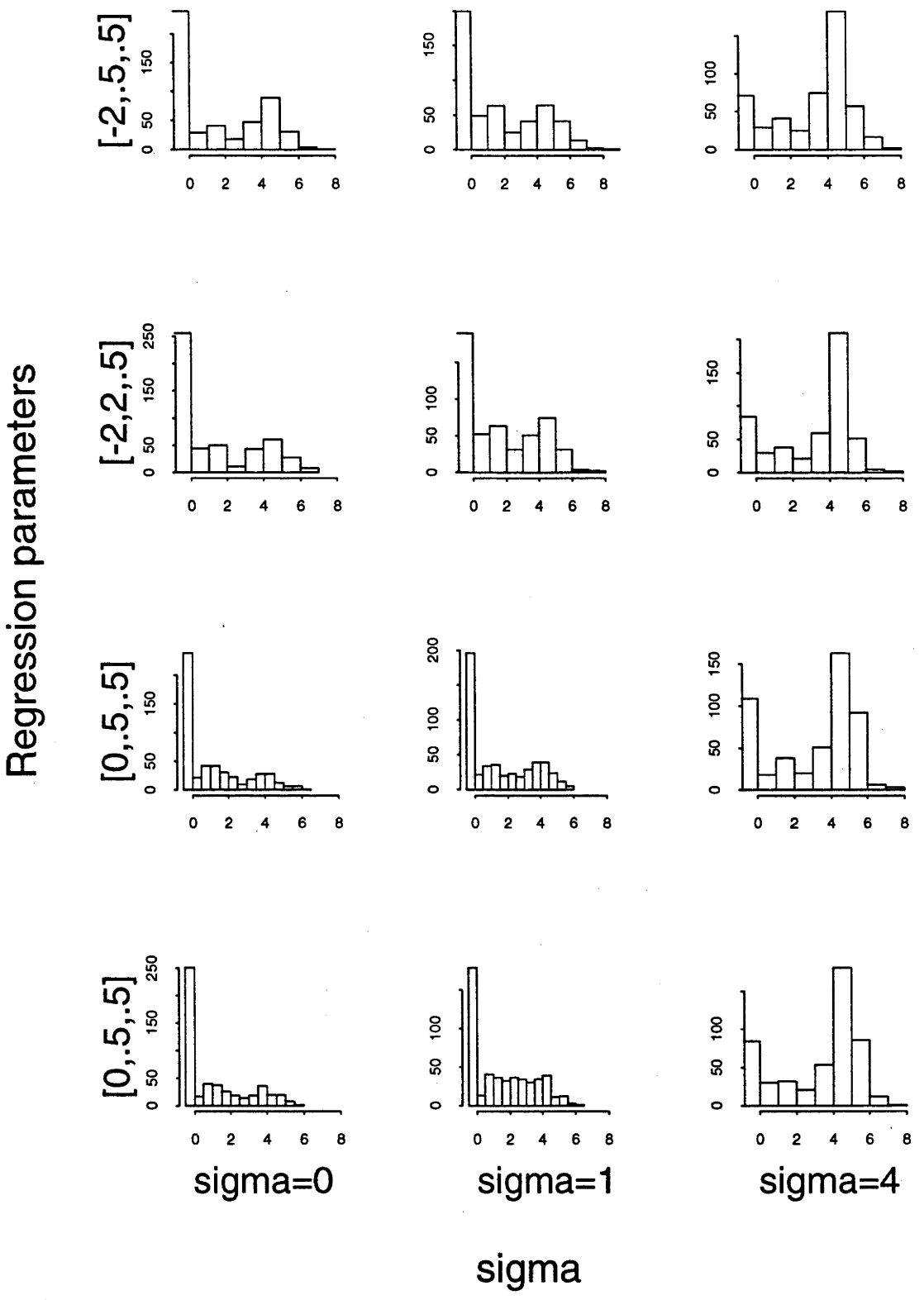


Figure 6.1: The approximate sampling distributions for the estimates of σ for simulations with group size two. $[a, b, c]$ is the simulation with the intercept equal to a, the group level uniform covariate parameter equal to b, and the 0/1 treatment effect equal to c.

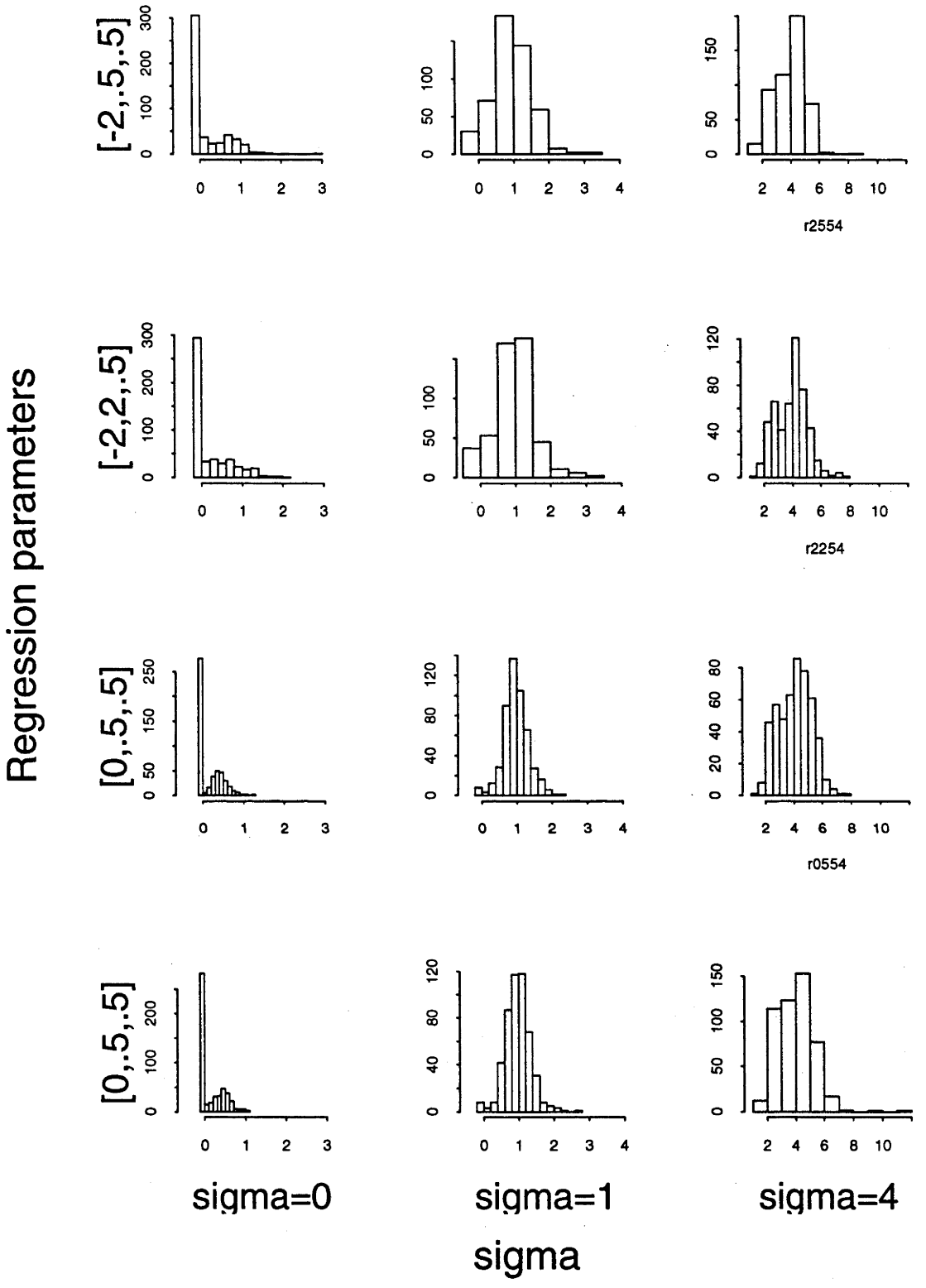


Figure 6.2: The approximate sampling distributions for the estimates of σ for simulations with group size four. $[a, b, c]$ is the simulation with the intercept equal to a, the group level uniform covariate parameter equal to b, and the 0/1 treatment effect equal to c

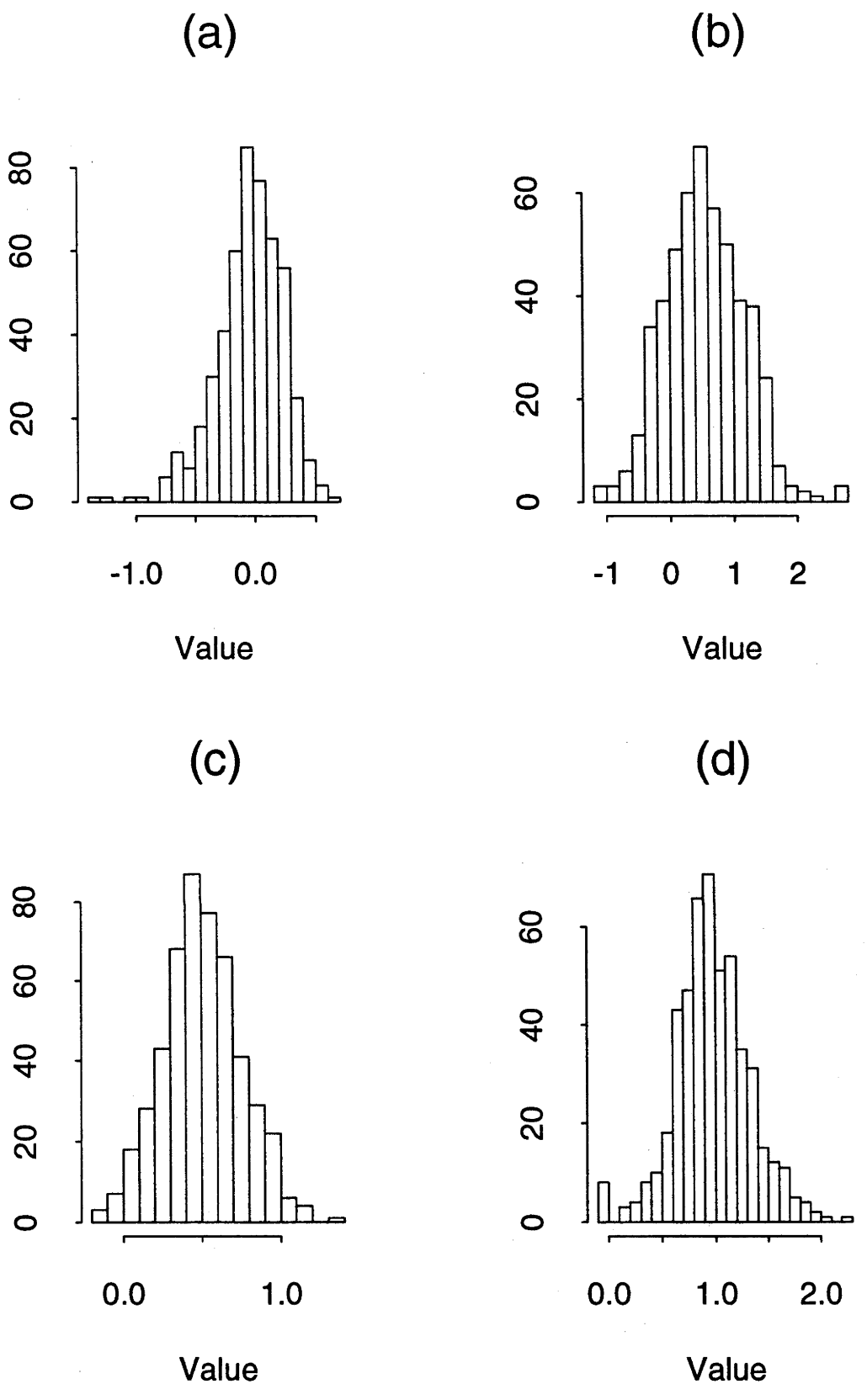


Figure 6.3: The approximate sampling distributions of the estimates for simulations with group size four, and parameter vector [0,.5,.5,1]. (a)=β_0, (b)=β_1, (c)=β_2 and (d)=σ^2.

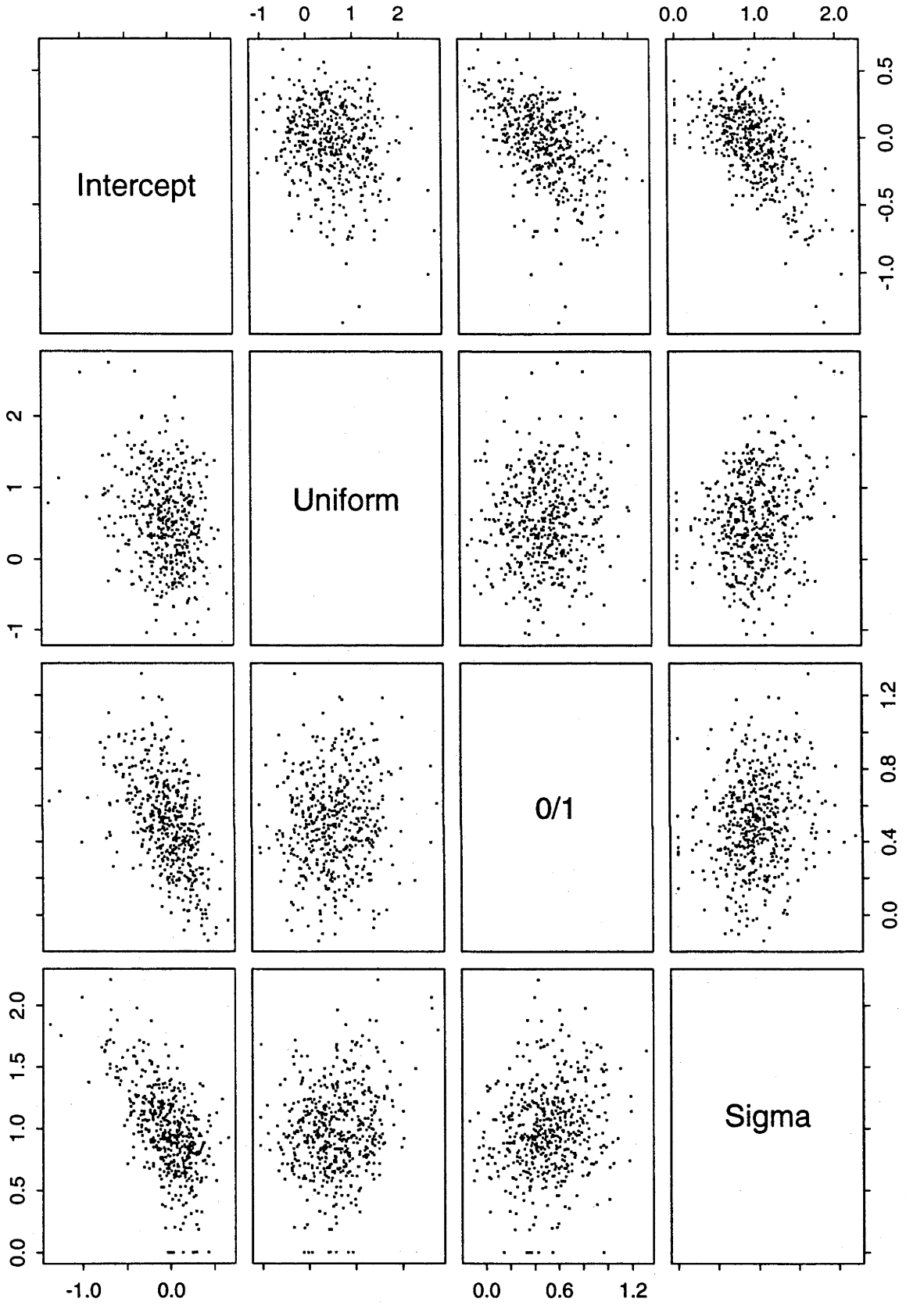


Figure 6.4: The pairwise sampling distributions of the estimates for simulations with group size four, and parameter vector [0,.5,.5,1].

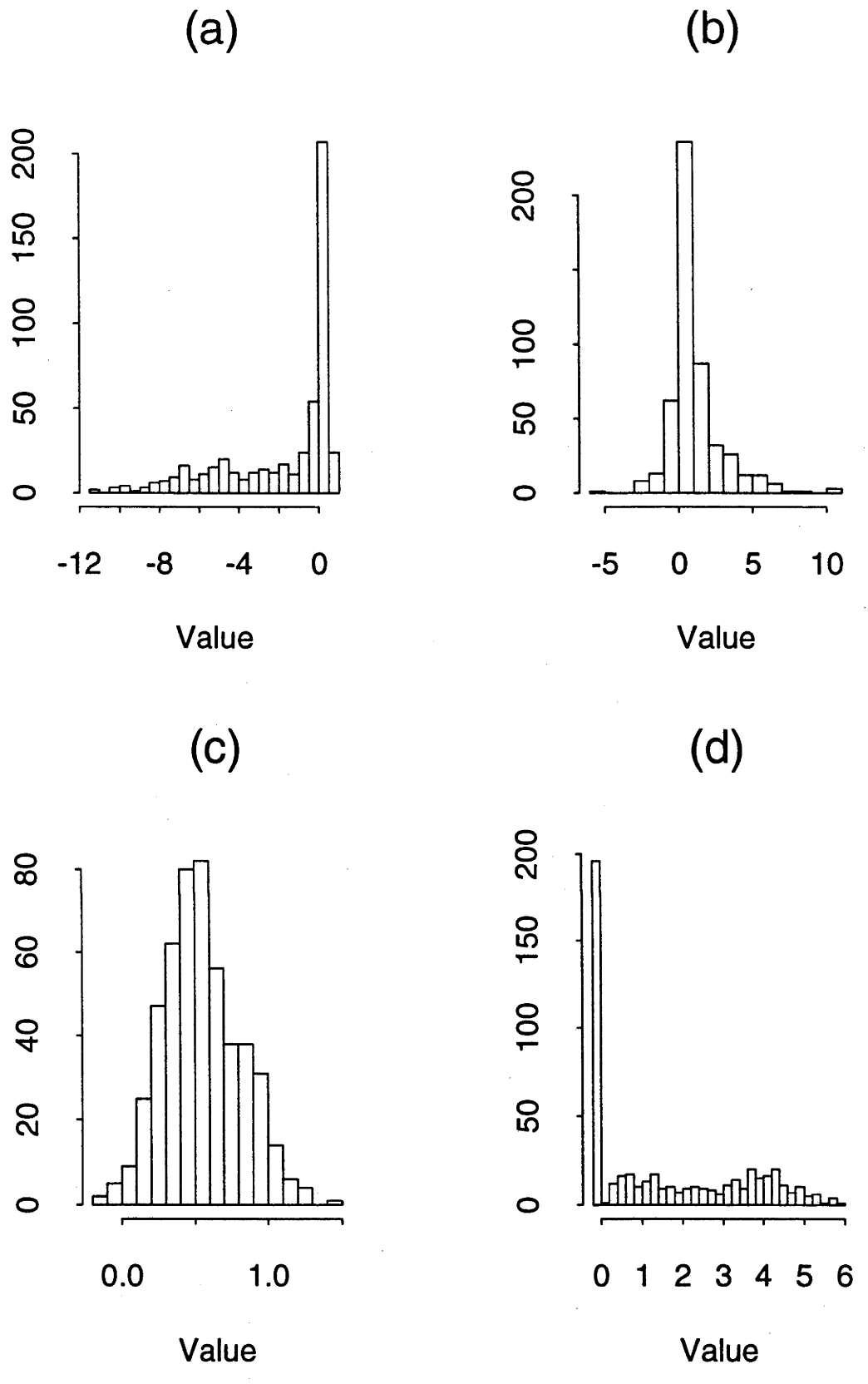


Figure 6.5: The approximate sampling distributions of the estimates for simulations with group size two, and parameter vector [0,.5,.5,1]. (a)=β_0, (b)=β_1, (c)=β_2 and (d)=σ^2.

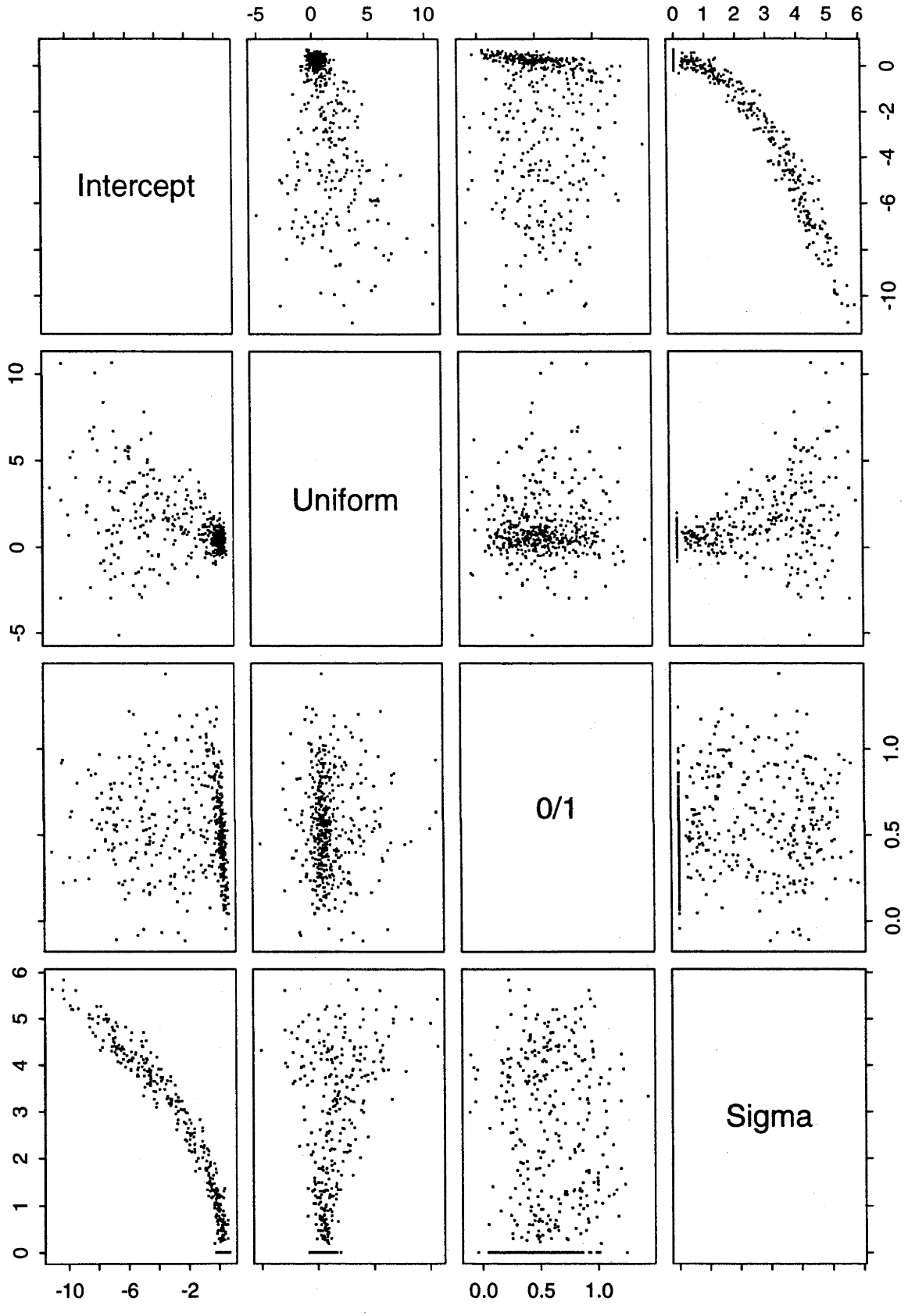


Figure 6.6: The pairwise sampling distributions of the estimates for simulations with group size two, and parameter vector [0,.5,.5,1].

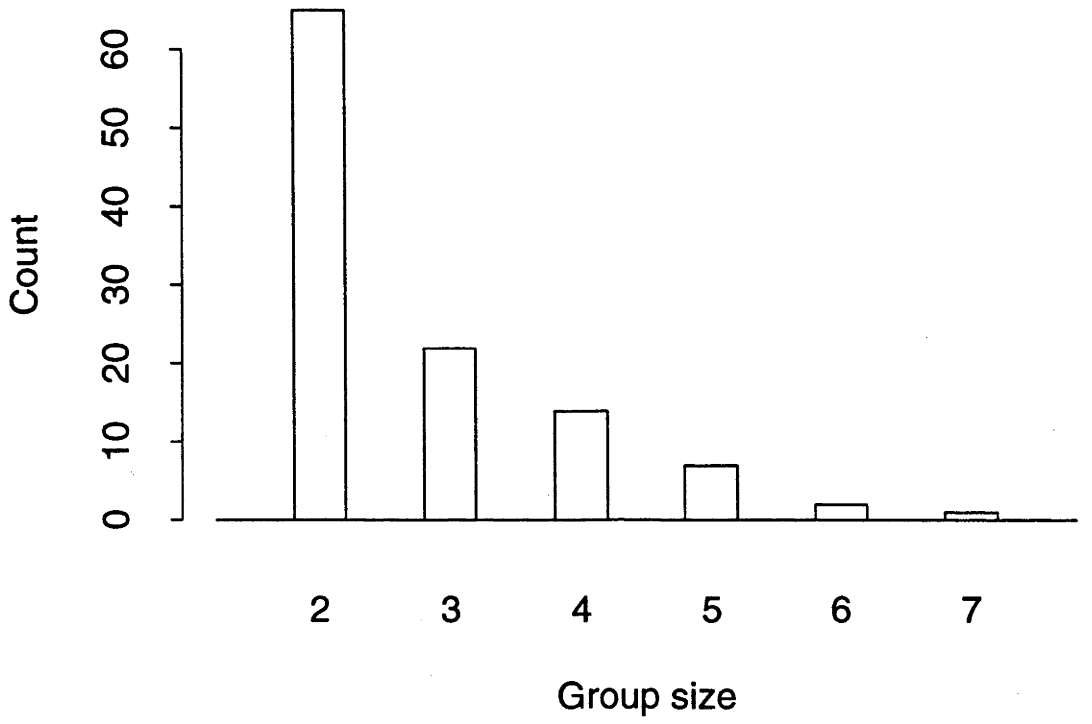


Figure 6.7: Distribution of group sizes for 88/90 FORS data.

Variable	Level	Ind1	RE1	ind2	RE2
Intercept	1	1.76	11.10	3.50(.88)	9.08(5.62)
Damage	1	NA	NA	0.53(.61)	.87(1.53)
Resavl	1	.93	1.64	1.10(.35)	1.58(.47)
Sex	1	.27	.40	0.37(.26)	.41(.28)
Speedlim	1	NA	NA	0.62(.51)	1.22(1.49)
	2	NA	NA	0.83(.83)	2.16(2.29)
Speedcat	1	NA	NA	1.59(.49)	3.19(1.95)
Age	1	.18	-.12	0.02(.50)	-.10(.55)
	2	.22	.38	0.25(.54)	.42(.61)
	3	1.00	1.62	1.21(.71)	1.69(.97)
Perloc	1	-.26	-.51	-0.33(.26)	-.48(.34)
σ		NA	3.46	NA	2.20(1.16)

Table 6.19: Log odds ratio estimates for single vehicle, frontal impact collisions, using the combined 1988 and 1990 FORS data. (Standard errors are in parenthesis, NA=not applicable.)

the marginal density of the observed covariates. Under mild regularity conditions about the covariate distribution we can thus form a consistent estimate (Gelfand & Smith, 1990) of the observed random effects distribution as

$$g_T(b) = \frac{1}{n}\sum_j g_T(b_j|j\text{th observed}),$$

where any unknown parameters are replaced by their estimates, provided they are consistent. The estimated distribution for the observed random effects is presented in Figures 6.8 and 6.9. Figure 6.8 compares $g(b_j;\hat{\sigma}^2)\hat{P}_j(b_j)$ for a sample of the truncated groups to the untruncated random effects density. It thus displays the bias and intensity of the truncation on the random effects distribution. Note that the relative areas of the truncated densities with respect to the untruncated density is equal to

$$\int g(b_j;\sigma^2)P_j(b_j)db_j$$

which is the marginal probability that the jth group is observed. Figure 6.9 plots $g_T(b)$ and the untruncated density. The figure demonstrates the estimated intensity of truncation is high, and if the random effects model is accurate, that there is a large proportion of the population unobserved. This is reflected in the estimate of N obtained from the truncated model via the marginal probabilities of being observed, which is

$$\hat{N} = 3398$$

an order of magnitude increase over the estimate assuming independence. This result highlights the problems involved in the analysis of such data. The form of the random effect distribution is essentially an untestable assumption in the model. For instance, a random effects distribution with a heavy left tail cannot be distinguished in this case due to the extreme truncation of random effects from this portion of the distribution. While it is possible to use theoretical arguments to finesse this problem, for example showing that asymptotically information will increase regarding the random effect distribution, these results are of no use in the finite sample case. This is an enduring problem with missing data and must be taken into account when interpreting the results of these models.

6.6 Discussion

It is well known that maximum likelihood estimates of the random effects, though asymptotically unbiased under feasible assumptions, are biased in finite samples (see Harville (1977)). This should not present a serious problem in this case due to the single random effect used in the model, provided the number of groups is large with respect to the number of parameters estimated.

The use of random effects in this context must be considered with some care. In the untruncated Gaussian model the inference is hopefully fairly robust to the distributional specification of the random effects. They can thus be considered a convenient device for the introduction of nested variation. In the truncated model with binary data this is no longer the case as they have a large impact on all aspects of the model.

A further problem with the analysis of this data is the high correlation between the intercept and random effects term. This is a feature of the model, and is

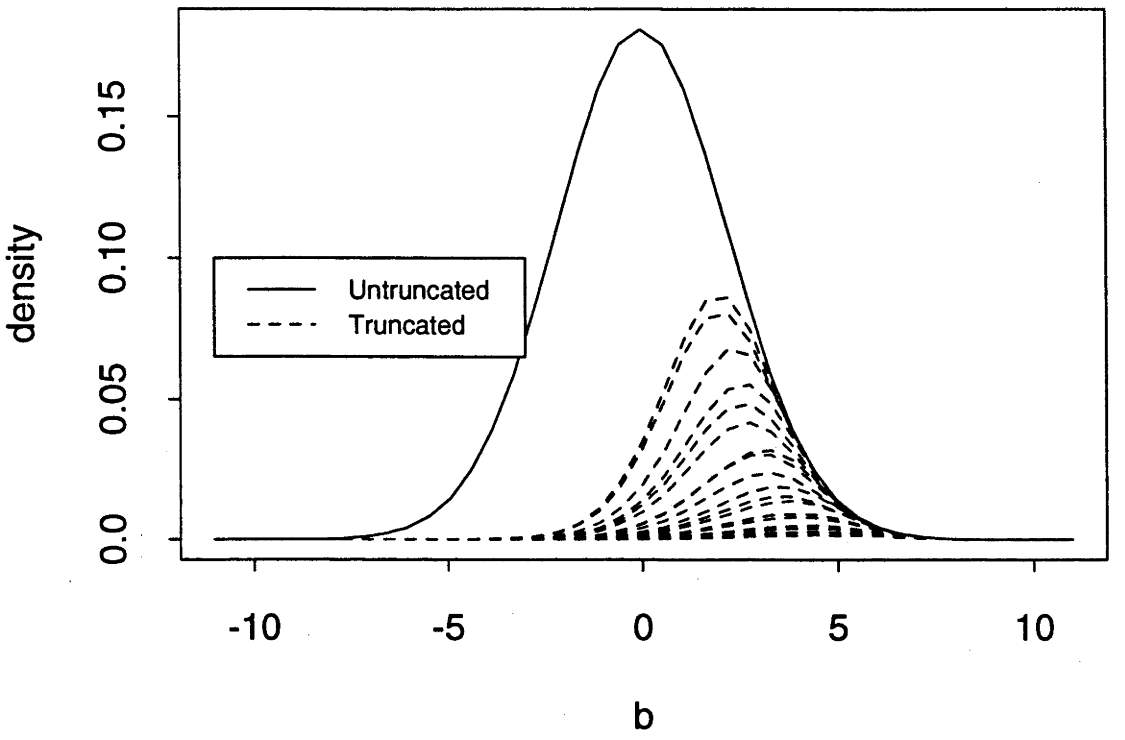


Figure 6.8: Estimated observed random effects distribution for a sample of groups from the 88/90 FORS data.

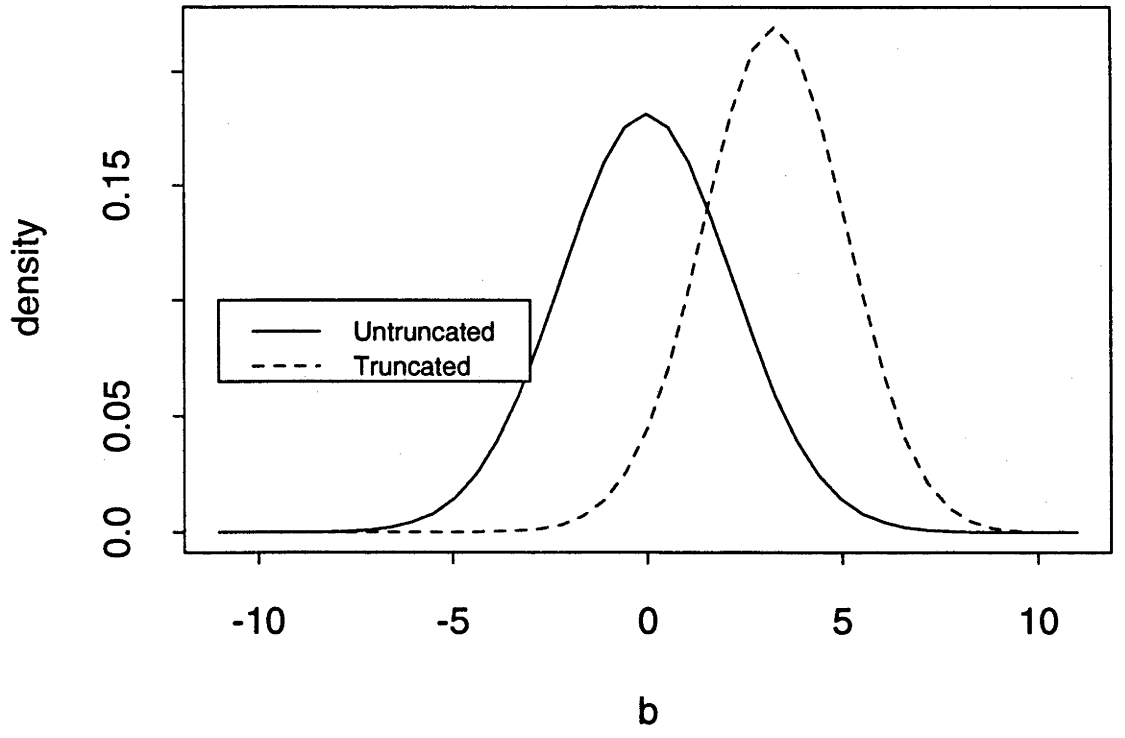


Figure 6.9: Estimated marginal distribution with respect to the covariate distribution of the random effects for the 88/90 FORS data.

a simple reflection of our lack of knowledge about the unobserved cases. With truncation of the null class, truncation of the random effects distribution effectively allows the data to be fitted by the right hand tail of the random effect distribution, leading to negative biasing of the intercept term. These fits correspond to extremely ridged likelihoods that are very flat along the ridge. This is a sparseness problem, but as the simulations demonstrate, binary data can be sparse for comparatively large sample sizes. This is due to the binary response providing little information regarding the random effect, and the truncation further disguising this information.

This feature is pronounced in the paired case. In this case, as noted before, the parameters are not identifiable when only categorical covariates are present, and with continuous covariates the amount of information about σ that is orthogonal to that for the intercept term is very small. Thus the use of the truncated model with paired data is potentially misleading, as there is no power in the test between the two plausible models, the independence model and the random effects model, unless the sample sizes are extremely large. The impact of this on the estimation of N would be potentially large, especially if the random effect variance was large. In this case, the assumption of the independence model would lead to significant negative biasing of the estimate of N. If the random effect model was chosen when the independence model was in fact true the bias would be reversed. Thus the paired data situation is potentially a source of practical problems.

The use of post enumeration surveys (PES) to adjust census counts in many countries is thus a source of concern. Alho et al. (1993) has considered the use of a paired binary truncated model to perform this adjustment. This model assumes that conditional on the covariate values the untruncated responses are independent. Note that in this case the situation is slightly different to that considered here. The intercept terms are different at each collection period, due to the probability of being included in the census being much higher than the probability of being included in the PES. In addition, the primary focus of the exercise is to adjust the census totals so that relative estimates of different classifications truly reflect the population. Thus bias is acceptable in the totals if this bias is consistent across all classifications reported. Thus further research is required to investigate this case.

Darroch, Fienberg, Glonek, & Junker (1993) recognised some of these problems and considered methods of estimation of N that relied on the use of three separate samples. The method is based on the analysis of incomplete contingency tables, and only allows the use of covariates for stratification, and not for explicitly modelling the response probabilities. The stratification/log linear approach has the drawback of requiring the assumption of no second order interaction to identify the parameters of the missing cell corresponding to the individuals never observed. This assumption of no second order interaction is only plausible for some covariate distributions. This is not a feature of the analysis presented in this chapter, due to the probability structure being directly modelled.

The specification of random effects in the traffic accident data could potentially be extended by allowing random effects for particular makes of car, perhaps, to assess the variability in the safety between the cars. Thus the car level effects are nested within the make of car. This presents difficulties. Consider a truncated sample of n accidents involving a particular make of car. Let the random effect associated with the jth accident be denoted by b_j and the random effect for the particular make be b_m, all being mutually independent, with associated distributions $g_b()$ and $g_m()$. We model the probabilities of response for the ith unit in the

jth car of make m as

$$logit(\mu_{ij}(b_j, b_m)) = log(\frac{\mu_{ij}(b_j, b_m)}{1 - \mu_{ij}(b_j, b_m)}) = x_{ij}\beta + b_j + b_m.$$

Now if we assume that conditional on the random effects the untruncated responses are independent the likelihood is

$$\prod_j L_T(y_j|b_j, b_m)$$

where $L_T(y_j|b_j, b_m)$ is the probability of the response for the jth group, given b_j, b_m, and that it was observed. Again we wish to form the joint density of the y_j, b_j and b_m. Now, the joint distribution of the b_j given b_m and that they were observed is

$$\prod_j \frac{g(b_j)P_j}{\int g(b_j)P_j db_j}$$

due to the independence of the random effects conditional on b_m, and $P_j = P(\beta, b_j, b_m)$ is the probability the jth group was observed. When we attempt to calculate the marginal distribution of b_m it is a complicated expression which depends on the unobserved groups. This comes about because the random effect, b_m, is truncated if none of a particular make of car is observed. To see this note that the conditional distribution of b_m given the b_j, and that the group was observed is proportional to

$$g(b_m)P_m$$

where $P_m = 1 - \prod_j^N (1 - P_j)$ and N is the total number of accidents involving make m, observed and unobserved. Thus the truncation occurs at a higher level and depends on unobserved responses. This can be interpreted as saying that potentially, very safe cars will not be observed in the sample. This again draws attention to the heavy dependence of the form of $g()$ on the result of the analysis. At this point it is perhaps prudent to consider the use of fixed effects to model the effect of car make, but for arguments sake we will continue.

We note that if the P_j are strictly positive, i.e. there is a finite chance that any group is observed, then P_m converges to 1 with increasing N. Thus for large N we may reasonably approximate the marginal distribution of b_m given that it is observed by its unconditional marginal distribution $g(b_m)$. Hence the approximate joint distribution of the response and random effects given that the group is observed is

$$\prod_j L_T(y_j) \frac{g(b_j)P_j}{\int g(b_j)P_j db_j} g(b_m).$$

Thus we can calculate the marginal distribution of the y's as before, by integrating over b_j and b_m. The model can thus be fitted provided an efficient computer is available, but the increase in computation time and complexity of the expressions is dramatic. For instance, for the sample of n cars from a particular make, calculation of the likelihood itself would require the integral of the product of n integrals. Thus if k is the number of quadrature points the computational problem rises from $O(k)$ to $O(k^2)$. This rapidly becomes unmanageable as the number of random effects increases. Crossed random effects present even greater computational difficulties. With these problems it is unlikely that the use of numeric integration will be

feasible with these models in the near future. While more efficient computers speed the calculations the cumulative errors of high order integration will restrict its application. Given these difficulties approximations must be considered.

Lowering the order of the quadrature is possible but leads to the same problem of inaccurate estimation of the likelihood surface, with no theoretical results to ensure that these inaccuracies will be uniform across the parameter space. Thus this is not a real option. An alternative is to consider the use of analytic approximations currently being investigated with generalized linear mixed models such as Breslow & Clayton (1993) and McGilchrist (1994). While these potentially have application here with substantial modifications, the quality of the central approximations to the likelihood surface are extremely poor in any conceivable applied problem. Thus the technique is essentially ad hoc, and has no proven consistency properties. Recently Kuk (1995) has presented a Monte Carlo method for producing asymptotically unbiased estimates in the Generalized Linear Mixed Model. This has potential application in this case.

An alternative is to consider the Bayesian approach of Zeger & Karim (1991). In this case sampling from the conditional distribution of the random effects is complicated but could be efficiently performed via adaptive rejection sampling (Gilks & Wild, 1992) provided the density is regular enough. Sampling from the conditional distribution of the random effects variances, and the regression parameters, is also complicated due to the evaluation of N one dimensional integrals at each step, although luckily the number of integrals required does not increase with increasing nesting of random effects. A rejection sampling algorithm could be applied, or perhaps a griddy Gibbs sampler, or a Gaussian approximation (appropriately truncated). Alternatively we could use Laplace's approximation to the integral (Tierney & Kadane, 1985). The direct application of the Gibbs algorithm involves the formulation of the prior distributions after truncation, which presents logical difficulties. Another avenue is to consider the use of alternative Markov Chain Monte Carlo Methods, such as the Metropolis algorithm.

Another issue to be discussed is model selection. To develop a strategy for model selection difficulties arise. If one is willing to suppress concern with practical issues then there is little problem asymptotically. To test simple hypotheses the usual score tests can be constructed easily from the results above. For composite hypothesis score tests can again be considered or likelihood ration tests based on Self & Liang (1987) can be considered if the parameter is on the boundary. For non-nested models the use of information theoretic measure such as the AIC may be called for.

When confronted with numerous covariates and a model that is of course only approximately correct the approach is complicated. As is usual, the model selection process is by necessity recursive, and due to the lack of orthogonality the significance of particular terms depends on the order they are fitted. If one is willing to assume that a random effect term is always present, then the score tests and likelihood ratio tests mentioned above can be performed in the usual step wise manner outlined in McCullagh & Nelder (1989). These tests are asymptotically correct, but in finite samples, the problem of obtaining estimates on the boundaries arises, which further complicates their finite sample properties. There appears to be no direct way of correcting for this problem in finite samples. If the asymptotic distribution is considered the problem disappears due to the consistency of the MLE's.

The assumption of the presence of random effects/over-dispersion is made in much practical work. It stems from the lack of power of tests for the significance of the variance terms, and from the legitimate belief that some over-dispersion will always be present and thus should be accounted for. The question of the exact form of this unknown dependence is not addressed, and practically it is not possible to specify. Of course, as terms are added and deleted $\hat{\sigma}^2$ can be on the boundary at different times. Due to the consistency of the MLE's this will not occur in infinite samples if $\hat{\sigma}^2 > 0$ but will occur in small samples. Thus the usual asymptotic results will be potentially worse approximations than is usually the case.

As the conclusions that can be drawn from a truncated analysis are so heavily dependent on the assumptions it appears sensible to view all analyses in an exploratory way. It would seem appealing to search for a set of covariates that reduce the effect of the random effects, but this search is outside the usual paradigm of inference i.e. hypothesis, sample, test, and thus has reduced validity. The problems inherent in overfitting arise in this case. The addition of significant group level covariates in the clustered model will in most cases serve to lower the estimate of σ. This effect carries over to the truncated model, although complications obviously arise due to the asymmetric marginal truncation of the model.

A possible approach is to consider the following:

1. Fit the truncated model assuming independence using the likelihood ratio tests to determine significant covariates.

2. Test for random effects at the group level.

3. If a significant random effect is present, add dropped group level effects and repeat the test.

Item 3 is important as the random effects operate directly as the group level covariates. Note that it is not possible to test for random effects at the beginning of the analysis, as their significance will change as the model changes. This procedure's main advantage is its simplicity. Although this issue should not be too high on a statisticians list of priorities, the fitting of models to group truncated data is an exercise in extrapolation. Though theoretically possible to perform rigourous hypothesis testing, in practice the aim is to produce a model roughly consistent with the data. Thus this last procedure has something to commend it. If necessary it may be appropriate to present a number of plausible models.

A final point to discuss is the assessment of goodness of fit. The score test constructed previously also serves as a goodness of fit test, as it tests the independence model against correlated alternatives. Of course, the power of the test against specific alternatives is open to question, and the alternative considered here is that there is positive correlation within the groups. Thus, the technique will lack power against alternatives that imply disparate response within groups, which is the analogue of under-dispersion in the usual logistic regression model, with constant covariates within the 'group'. In addition the truncation can lower the effective power of the test compared to samples from the non-truncated distributions. This effect is pronounced when the truncation removes an extreme (with respect to the random effect) response from observation. For example, the groups with no positive response contain much information regarding the random effects, and if these groups are truncated, estimation of the random effects is hindered.

Chapter 7

Discussion

There has been some discussion of the techniques presented in this thesis at the end of each chapter and this will not be repeated. Instead, there are two main issues that will be discussed in this chapter. The first issue is the contribution of this thesis to the understanding of truncated data, and the remaining problems that need to be addressed in this area. This is essentially a mathematical discussion and will consider what theoretical developments need to be produced to give flexibility in the application of these models. The second part of the discussion will concern itself with the validity of using the truncated model. During the course of this study there has been some scepticism regarding the usefulness of these models. For this reason some discussion of this point is appropriate.

The results of this thesis have provided several original contributions to statistical science. The first contribution is the general treatment of grouped truncated data. Where data of this type has been considered in the literature, for example the capture-recapture analysis of Huggins (1989), it has been considered as a particular type of conditional analysis and has not been approached from the general viewpoint of truncation. The truncated approach is useful as it allows the particular analysis to be viewed as an example of a more general class of procedures. This development adds insight into the nature and behaviour of the analysis. It also provides direct insight into the extension of models defined on the untruncated sample space to the truncated sample space. This can be seen in Chapters 2 and 3 where the standard analysis is extended to truncated data in a natural way.

The work on the frequentist estimation of N has generalised the results of Sanathanan to general truncated regression models. The results of Huggins arise as a special case, with the present formulation being more flexible, considering the estimation of subpopulations in the presence of covariates. The simulation results demonstrate that the procedure works well for moderate truncation and sufficient truncated sample to provide stable estimates of the regression parameters. Research could potentially be carried out into the more extreme cases, although it must be recognised that the problem is fundamentally due to a lack of information, as opposed to any pathological deficiency in the problems formulation. The obvious candidate is to consider the use of prior information, and either pursue a Bayesian approach or a penalized likelihood approach, the choice depending on our inclination. These extensions would be relatively straightforward, but the usefulness of obtaining general results for these cases is doubtful. The analysis and interpretation of observationally truncated data is complicated at present. The addition of another layer of complexity into the analysis should be done only in

specialised circumstances, which will depend crucially on the problem at hand. For example, the naive truncated analysis of the road safety data in this thesis does not provide any insight into the potential application of these extended methods, although it may be possible that a road safety researcher could sensibly provide acceptable priors.

The Bayesian approach considered in Chapter 5 is novel, and has rich potential for further research. The results presented are encouraging in the insights that they provide. Further work should be carried out comparing this approach with the standard analyses in the literature. In particular there is a need to explore the links between the current approach and empirical Bayes techniques and non-parametric Bayesian techniques. In addition, the Bayesian formulation and the use of Markov Chain Monte Carlo methods opens up the possibility of incorporating additional complexity into the model in a logically appealing and computationally feasible way. The use of data augmentation has already provided new avenues for the statistical analysis of complicated statistical models. Preliminary research has indicated that the truncation may negate some of the simplifications that are generated by the augmentation approach. Still, there are a number of promising directions, such as incorporating dependence structure in the grouped response.

The random effects model of Chapter 6 is interesting for a number of reasons. First, it obviously allows for dependence in the binary response of the clusters. The existence of this dependence, typically positive, in grouped binary data is well recognised in applied work. Techniques to model this have been an extremely active area of research in recent years. The random effect specification is a plausible model of dependence, and extends techniques presently applied in the untruncated case. Importantly, this approach allows an assessment of the independence models' goodness of fit via the score test developed in this chapter.

The issue of dependence in group truncated binary data deserves further study. The random effects model considered in Chapter 6 provides a useful tool for the modelling of positive association, but does not extend easily to other forms of association. The use of the exponential model of Cox, as used by Fitzmaurice & Laird (1993), provides a mechanism for the introduction of arbitrary dependence structure. Unfortunately the modelling of the canonical parameters in this case is not obvious, and the modelling of the marginal mean via a linear predictor, a great attraction of the model in untruncated case, appears somewhat arbitrary in the presence of truncation.

This thesis has been concerned with the analysis of observationally truncated data. The discussion in the previous chapters has considered methods of analysis that perform well when the model is correctly specified. In applied work it is well understood that any model is only an approximation to the 'truth', and thought must be given to the broader interpretation of the results of any analysis. We now discuss these philosophical issues.

The analysis of truncated data is problematic in applied statistics. The method of analysis essentially involves the extrapolation of the probability model to the unobserved class. This is seen most clearly when we consider the estimation of N in the truncated case, but also lies behind the conditional analysis of the regression parameters considered in Chapters 2 and 3. The use of extrapolation is rightfully considered a dangerous practice in applied statistics.

But surely it can be argued that data such as the road safety examples contains information regarding the group level effects. The use of Conditional Logistic Re-

gression, where the number of positive responses is conditioned on, is appealing due to the reduced set of assumptions that are needed. It can be viewed as one extreme of the useful available conditional analyses. The opposite extreme is the truncated logistic case model developed here, which directly specifies the observed data's distribution. It is interesting to ponder whether a conditional analysis between these two extremes could retain some of the desirable properties of each analysis, while being more resistant to the effects of misspecification than the truncated model. This could be an interesting area of future research

Truncated data can be considered as an extreme in missing data problems. From the full data situation we move through the type I censored data, and then finally we reach the truncated data. At each step information is lost, but it is useful to consider that the analysis of censored data shares many features with the analysis of truncated data in that a probability model is effectively used to extrapolate from the observed data to the unobserved population. The situation with truncated data is just more extreme.

The analysis of censored data proceeds under the proviso that the results of the analysis depend critically on the assumed distributional form. For example, in the analysis of censored survival data the available information may support a variety of possible models, and on the basis of the available information it is not possible to distinguish between particular models. Nonetheless, the results of the various analyses are still of scientific interest as they still provide valid descriptions of the observed data and thus allow interpretation of plausible mechanisms for the generation of the data. This point is subtle. Any analysis performed is conditional on the assumptions used. In the analysis of complete data the analysis can direct attention to the observed features of the data, and the procedures can be justified with direct reference to this data. An example of this is the problem of inferring properties of the mean of an unknown distribution.

Modern analysis has directed attention to performing valid inference under weak conditions about the assumed distribution. These procedures only rely on the mapping of the observations to a density. For many inferential problems the form of the density is not critical for the validity of a technique. For example, the mean provides a weakly consistent estimate of the mean of any distribution with finite mean, and is both a direct summary of the observed data as well as providing a theoretically useful statistic.

With missing data structure must be imposed for estimation to proceed. In this case we have a mapping from the density to a theoretical sample which is then mapped to the reduced sample. Under current thinking this mapping must be explicit, and generally requires an explicit, parametric formulation of the problem. Examples such as Tsui et al. (1988) exist which relax this assumption, but in this case strong assumptions must still be made regarding the density and the truncation process. In many circumstances the validity of these parametric assumptions is either fundamentally untestable or the specification of plausible alternatives produces tests that are either not easily constructed or exhibit so little power that they are of no assistance in determining the 'true' pattern of the data. Examples can be found in Little & Rubin (1987).

It could be argued that in cases where consistent inference cannot be achieved statistics has no valid place. By consistent we mean that a procedure cannot be developed that, with increasing information, will produce valid inference. Censored and truncated models fail this test. Though they are consistent if the probabil-

ity model is correctly specified, no information is gathered regarding the missing observations and thus the procedures cannot be consistent with respect to general alternatives.

This view is extreme. Although the production of valid inference is the primary aim of statistical science, there are other important objectives. Statistics allows complicated and expansive data sets to be summarised in a form that allows clearer understanding. In the case of group truncated ordinal data it is doubtful that any person could, by simple study of the expansive data, unpick the competing effects of the covariates and the truncation to arrive at general conclusions that truly reflect the data, let alone assess the statistical strength of these conclusions. The truncated analysis gives this capability, and its use can thus provide insight into the structure of the data that could not be achieved by any other means. Its interpretation must of course be carried out with respect to the assumptions that were used in the construction of the estimates, but again, if the model is approximately correct we can expect answers that are also approximately correct.

Appendix A

Bootstrapping truncated regression models

A.1 Introduction

Group truncated binary data occurs when a cluster of ordinal responses is observed only if the response of at least one of the individuals in the cluster exceeds a certain value. As an example, consider a sample of individuals involved in car accidents. The data is group truncated if only accidents involving fatalities are observed. O'Neill & Barry (1993a) developed a maximum likelihood approach to the estimation of regression parameters in the binary group truncated case. This was extended in O'Neill & Barry (1993b) to group truncated ordinal responses. As is commonly found, the exact variance of the maximum likelihood estimates is not explicitly available, and the usual asymptotic approximations must be used.

The bootstrap has been found to be a useful tool in estimating the variance of statistical estimators and setting approximate confidence intervals (Efron & Tibshirani, 1993). This chapter considers the use of the bootstrap to set confidence intervals, and thus better approximate the finite sample distributions of the maximum likelihood estimates. The use of the bootstrap in Gaussian regression was first considered by Efron (1979) and its behaviour further considered by many others (see the references in (Efron & Tibshirani, 1993)). In the case of regression models, Wu (1986) has explored the behaviour of the bootstrap based on resampling residuals versus the resampling of complete observation vectors. Moulton & Zeger (1991) applied the Gaussian regression arguments to generalised linear models and used a simulation study to assess their behaviour. As will be seen the use of the bootstrap in the truncated case is not feasible and a parametric or Monte Carlo bootstrap is defined.

Section A.2 examines the approaches available for bootstrapping truncated ordinal data. Section A.3 presents a simulation study to assess the behaviour of the bootstrap. Section A.4 presents conclusions.

A.2 Bootstrapping Regression models

A.2.1 Group Truncated Data

Recall from Chapter 2 that group truncated binary model is as follows. Suppose that the truncation group is an accident and the binary variable is survival. Let

$$
\begin{aligned}
y_{ij} &= \begin{cases} 1 & \text{if the } i\text{th individual in the } j\text{th accident dies,} \\ 0 & \text{otherwise,} \end{cases} \\
x_{ij} &= \text{the } p \times 1 \text{ vector of covariates associated with the } i\text{th individual in accident } j, \\
\mathcal{R}_j &= \text{the set of individuals in accident } j, \\
\beta &= \text{the } p \times 1 \text{ vector of regression parameters.} \\
X_j &= \text{the matrix with } i\text{th row } x_{ij}
\end{aligned}
$$

We suppose that a logistic model holds for the individual probabilities of death,

$$Pr(y_{ij} = 1) = \frac{exp(\beta' x_{ij})}{1 + exp(\beta' x_{ij})} = p(\beta, x_{ij}) = 1 - q(\beta, x_{ij}).$$

The Truncated Logistic Regression (TLR) approach conditions on the probability that an accident is observed which is the probability that it results in at least one fatality. This has the effect of introducing a divisor to a conventional logistic regression likelihood. The TLR estimator $\hat{\beta}$ is the maximiser of

$$\prod_{j=1}^{N} \frac{\prod_{i \in \mathcal{R}_j} p(\beta, x_{ij})^{y_{ij}} q(\beta, x_{ij})^{1-y_{ij}}}{P_j(\beta)} \tag{A.1}$$

where

$$
\begin{aligned}
P_j(\beta) = P(\beta, X_j) = 1 - Q_j(\beta) &= 1 - Q(\beta, X_j) \\
&= 1 - \prod_{j \in \mathcal{R}_j} q(\beta, x_{ij}),
\end{aligned}
$$

is the probability of being observed in the sample, and N is the number of observed accidents.

A.2.2 Bootstrapping Techniques

We will first examine the potential of the bootstrap to estimate variability in truncated data situations. We will assume the aim of the analysis is to make inference regarding the parameters of the underlying distribution, instead of the truncated distribution. As a simple example consider data from a single parameter exponential distribution,

$$f(x) = \exp[x\theta - b(\theta) + c(x)]$$

where θ is an unknown parameter. If we assume that the distribution is truncated below a point k then the truncated density $f_T(x; \theta)$ is

$$f_T(x; \theta) = \exp[x\theta - b(\theta) + c(x) - \log(k(\theta))]$$

where $\log(k(\theta)) = \int_{-\infty}^{k} f(x;\theta)dx$. We assume that the parameter θ is still identifiable. Consider a truncated sample X of size n. The score is thus

$$U(\theta) = \sum_{i=1}^{n}[x_i - b'(\theta) - \frac{k'(\theta)}{k(\theta)}]$$

which is equivalent to

$$U(\theta) = \sum_{i=1}^{n}[x_i - \mu_T(\theta)]$$

where $\mu_T(\theta)$ is the mean of the observed data. We consider the alternative parametrisation in terms of the mean, assuming that the mean is a one to one function of the canonical parameter θ. Let μ be the mean of $f(x;\theta)$ and μ_T the mean of $f_T(x;\theta)$. We can define a one to one function

$$h(\mu) = \int_{k}^{\infty} x f_T(x;\mu)dx = \mu_T$$

which maps the mean of the untruncated distribution to the truncated mean. Thus as the function is one to one we may write $\mu = h^{-1}(\mu_T)$. The MLE estimate, $\hat{\mu}$ of μ is thus

$$\hat{\mu} = h^{-1}(\hat{\mu}_T) = h^{-1}(\bar{X}).$$

Thus for large n we may write the approximation

$$\hat{\mu} - \mu \sim Normal(0, (h^{-1}(\mu_T)')^2\sigma^2/n) \tag{A.2}$$

where σ^2 is the variance of x calculated over the truncated distribution and $h^{-1}(\mu_T)'$ is the derivative of $h^{-1}(x)$ with respect to x, evaluated at μ_T. This derivation allows us to clearly identify the strengths and weaknesses of the bootstrap approach when analysing truncated data. Note that we can sample from the observations to form a bootstrap distribution, and if the model is correct it will asymptotically(as $n \to \infty$) produce results identical to the asymptotic form A.2. Now if the model is incorrect, the bootstrap will still produce a nonparametric estimate of the variability of the estimate $\hat{\mu}_T$. Unfortunately, no such estimate exists for $h^{-1}()$. As the form of $h^{-1}()$ impacts critically on both the bias and variability of the estimate, it is a source of some concern that in practical situations the choice of the function $h^{-1}()$ is generally by assumption. Thus bootstrap methods cannot provide nonparametric inference in the truncated case. They may of course provide improvements on the asymptotic approximations used in finite samples, but the magnitude of these improvements depends on the case at hand.

In summary we confront two possible cases. Either the model is correct and then the bootstrap distribution and approximation A.2 are equivalent in large samples. Alternatively, if the model is incorrect, the bootstrap distribution of $\hat{\mu}_T$ and $\hat{\mu}$ will reflect the true sampling distribution of our estimator with respect to the observed population, but not the unobserved population. Unfortunately this is of limited interest due to the bias of the estimator being unknown, and depending explicitly on both the true, and assumed density, with only the second being known.

Recognising these limitations, we now consider the use of the bootstrap to perform inference regarding group truncated regression models. Consider the regression model of data, Y, derived from the following model:

$$Y = X\beta + e, \tag{A.3}$$

where X is an $n \times p$ matrix of explanatory variables, and β is a p vector of covariates and e is an n vector of independently and identically distributed random variables with variance σ^2. There is a quantity $f(\beta)$ of interest in the experiment.

There are two approaches commonly used in the bootstrapping of models such as model A.3 (Moulton & Zeger, 1991). In the first approach, ordinary least squares is used to produce a parameter estimate $\hat{\beta}$ and a vector of residuals $r = [I - X^T(X^TX)^{-1}X^T]Y$. The following steps are then performed:

1. Sample with replacement the residuals from the least squares fit, to form a vector e^*.

2. Create a new data set, $Y^* = X\hat{\beta} + e^*$.

3. Estimate $\hat{\beta}^*$ from Y^*, X and calculate $f(\hat{\beta}^*)$

The distribution of $f(\hat{\beta}^*)$ generated by this procedure can be found either explicitly or via Monte Carlo. This distribution is then used to approximate the sampling distribution of $f(\beta)$.

Various modifications can be used to remove small biases from the procedure but steps 1-3 remain essentially the same. The use of the bootstrapping algorithm relies on the exchangability of the error terms, e. This occurs due to the assumption that the error terms are independent and identically distributed. With generalized linear models the assumption of constant variance is untenable and Moulton & Zeger (1991) present a one step procedure for estimating the bootstrap distribution. This is based on using standardised residuals to arrive at quantities that are more nearly exchangeable.

The use of residual resampling is not in general feasible with group truncated data. With group truncated data the residuals are neither independent or identically distributed. The lack of independence arises due to the clustered and truncated nature of the sample. As an example consider a cluster of size 2. If the response of the first unit is zero, the response of the second unit must be one, otherwise the sample would have been truncated. Thus the definition of an exchangeable residual quantity is difficult. In addition, as the responses are binary the residuals have a two point distribution and the addition of simple quantities such as e^* to fitted values is not sensible except for specific e^*. We could of course consider resampling of residuals within identical truncation groups, but this is of course equivalent to resampling from the covariate/response pairs which will be discussed shortly.

For these reasons residual resampling is not considered feasible with group truncated data. Some progress could possibly be made with aggregated data sets, where the binary nature of the response would be removed, and transformations used to make the residuals more exchangeable, but its domain of application is small.

The second approach is to consider resampling from the y_i, x_i pairs. This attempts to approximate the unconditional sampling distribution of the data, instead of estimating it conditional on the covariates. This is done as follows:

1. Sample with replacement n times from the observation pairs (y_i, x_i)

2. Create a new data set, Y^*, X^*, from the sampled observations.

3. Estimate $\hat{\beta}^*$ from Y^*, X^* and calculate $f(\hat{\beta}^*)$

Again, the distribution of $f(\hat{\beta}^*)$ generated by this procedure can be found either explicitly (Moulton & Zeger, 1991) or via Monte Carlo. This approach potentially extends to group truncated data, which will be demonstrated heuristically below.

As a simple example consider a population of size N made up of 2 distinct categories, of size N_1 and N_2. Let the N_i groups in the ith category have identical covariate configuration and P_i be the probability that a group from configuration i of the population is observed. Let n_i be the (random) sample size of the ith category observed. We assume that

$$\frac{N_i}{N_1 + N_2} \rightarrow \pi_i$$

as $N \rightarrow \infty$. We consider estimation based on a truncated sample of size $n = n_1 + n_2$ from this population. Standard arguments can be used to show that the asymptotic variance of $n^{1/2}(\hat{\beta} - \beta)$ is

$$\frac{\pi_1 P_1}{\pi_1 P_1 + \pi_2 P_2} I_1 + \frac{\pi_2 P_2}{\pi_1 P_1 + \pi_2 P_2} I_2$$

where I_i is the expected information in a truncated group from the ith category of the population.

Consider a truncated sample from this population. If n_1 and n_2 are the sizes of the observed samples for each category, it is easy to show that

$$\frac{n_i}{n_1 + n_2} \rightarrow \frac{P_i \pi_i}{P_1 \pi_1 + P_2 \pi_2}.$$

Now if we produce bootstrap samples of size $n_1 + n_2$ from this truncated sample we find that the variance of the bootstrap estimates $n^{1/2}(\hat{\beta}^* - \beta)$ is asymptotically

$$\frac{P_1 \pi_1}{P_1 \pi_1 + P_2 \pi_2} I_1 + \frac{P_1 \pi_1}{P_1 \pi_1 + P_2 \pi_2} I_2.$$

Thus the resampling of groups from the observed population produces asymptotically valid inference. Unfortunately, it still requires the correctness of the parametric model.

Given the need to assume the correctness of the parametric model for viable inference to proceed we consider using the parametric bootstrap. The parametric bootstrap estimate for a function$f(\beta)$ is as follows (Efron & Tibshirani, 1993):

1. Estimate $\hat{\beta}$ using Equation A.1.
2. Sample from Y^* from $Y|X, \hat{\beta}$ conditional on the group being observed.
3. Calculate $\hat{\beta}^*$ from Y^* and X using Equation A.1 to produce $f(\hat{\beta}^*)$.

Items 2 and 3 are iterated to produce a Monte Carlo approximation to the sampling distribution of $f(\hat{\beta})$. The variance can then be calculated or bootstrap confidence intervals set. Step 2 is easy to to perform once it is noted that the distribution of deaths conditional on the covariates and that the group was observed is a multinomial distribution. Step 3 is easy to perform provided efficient fitting software is available to ensure that enough replications can be carried out to provide an adequate approximation to the bootstrap distribution.

A.3 Examples

The performance of the bootstrap was examined using two simulation experiments. In each experiment a population of accidents was simulated. For the ith individual in the jth accident the probability of death was calculated via

$$logit[Pr(y_{ij} = 1] = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} \tag{A.4}$$

where x_{ij1} was drawn from a Bernoulli distribution with probability .5, and x_{ij2} was drawn from a Uniform distribution over $[0, 1]$. The covariates remained constant over each simulation. For each set of parameter values the following steps were iterated 200 times.

1. A data set was generated using the probability model A.4.
2. The data set was truncated to produce a group truncated dataset.
3. Maximum likelihood was used to produce an estimate, $\hat{\beta}$ of the regression coefficients, and the approximate variance matrix of the parameters found from the information matrix.
4. The parametric bootstrap described in A.2.2 was applied to set percentile intervals for β_0, β_1, and β_2. In addition a percentile interval was set for the probability of death of an individual, given that they were in a vehicle with one other person and observed. This interval is a non linear function of the parameters,

   $$p = Pr(y_{1j} = 1|\text{observed}) = \frac{Pr(y_{i1} = 1)}{1 - Pr(y_{i1} = 0)Pr(y_{i2} = 0)}, \tag{A.5}$$

   with the probabilities being given by

   $$logit(Pr(y_{1j} = 1)) = \beta_1 + \beta_2,$$

   and

   $$logit(Pr(y_{2j} = 1)) = \beta_1.$$

   This non linearity could potentially present difficulties for the Gaussian approximation. The Monte Carlo simulation was run for 300 replications. If used on a real data set this should be larger to minimise the stochastic effects, but as will be seen it is adequate for assessing the coverage rates.
5. The Fisher information was used to set approximate confidence intervals for $\beta_0, \beta_1, \beta_2$ and $p = Pr(y_{i1} = 1|\text{observed})$.

The number of replications in the simulation was enforced by the high computational burden imposed by the bootstrap. Although more replications would have allowed better accuracy in estimating the coverage rates of the procedures, the number is still sufficient to both limit sampling errors while showing significant trends. With a sample of size 200 the standard error in estimated coverage if the true coverage is .9 is .02.

In running the simulations a number of issues arose. Firstly, for some extreme data sets, the likelihood converged but the parameters estimates did not. This is a

		β_0 coverage		β_1 coverage		β_2 coverage		p coverage	
		BS	Fisher	BS	Fisher	BS	Fisher	BS	Fisher
$\beta_1 = -2$	$\beta_2 = -2$	0.92	0.95	0.87	0.89	0.84	0.92	0.88	0.83
	$\beta_2 = -1$	0.89	0.91	0.89	0.90	0.88	0.88	0.87	0.85
	$\beta_2 = 0$	0.88	0.90	0.88	0.91	0.91	0.92	0.91	0.89
$\beta_1 = -1$	$\beta_2 = -2$	0.90	0.91	0.89	0.91	0.88	0.91	0.89	0.86
	$\beta_2 = -1$	0.88	0.90	0.90	0.88	0.88	0.88	0.88	0.85
	$\beta_2 = 0$	0.93	0.94	0.91	0.90	0.90	0.90	0.93	0.89
$\beta_1 = 0$	$\beta_2 = -2$	0.89	0.89	0.89	0.91	0.88	0.91	0.90	0.85
	$\beta_2 = -1$	0.89	0.90	0.88	0.92	0.94	0.92	0.95	0.91
	$\beta_2 = 0$	0.87	0.88	0.91	0.94	0.89	0.90	0.94	0.88

Table A.1: Estimated coverage probabilities of approximate 90% confidence intervals for simulation 1. BS=Bootstrap interval, Fisher= Information based interval.

feature of logistic regression in general and occurs when the fitted values converge to zero or one, and hence one of the linear components in Equation A.4 becomes infinite. This is most likely to occur in the presence of categorical covariates. In this case the estimates will not converge if all the responses within one level of the covariate are either 0 or 1. This will not present a problem in large samples as the probability of this occuring goes to zero as the sample size increases.

This feature of the experiment was dealt with in the following way. If non convergence of the parameter estimates was observed in Step 3 of the procedure, Steps 1-3 were repeated. If non convergence occured in the bootstrap replications, two options were considered.

1. The divergent parameter estimates can be retained in the bootstrap sample and the percentile interval calculated as in Efron & Tibshirani (1993). This procedure produces sensible intervals provided the level of divergence is lower than the level of the interval. Otherwise one or both end points of the interval is infinite. If this occurs it is necessary to lower the level of the interval to produce a finite interval.

2. The second procedure deletes the bootstrap replications where the parameter estimates are divergent. This estimates the bootstrap distribution conditional on convergence of the parameters. This is an ad hoc method to avoid the problem and has little to recommend itself except convenience.

A.3.1 Example 1

Two simulation experiments were performed. The first experiment consisted of 50 accidents involving four individuals in each accident. The population was estimated using model A.4. This simulation was performed with $\beta_0 = 0$ and over a grid of $\beta_1, \beta_2 = -2, -1, 0$. This experimental design examined the behaviour of the intervals over varying levels of truncation. The truncated data sets were rarely sparse, due to the large group sizes and did not produce large (>20) numbers of resamples with parameter estimates diverging. Percentile intervals were calculated using method 1 above. The coverage of the percentile intervals are given in Table A.1.

As can be seen from the results the coverage of both the bootstrap percentile intervals and those found from the Gaussian approximation are adequate for the

		β_0 coverage		β_1 coverage		β_2 coverage		p coverage	
		BS	Fisher	BS	Fisher	BS	Fisher	BS	Fisher
$\beta_1 = -2$	$\beta_2 = -2$	0.75	1.00	0.51	0.65	1.00	1.00	1.00	0.85
	$\beta_2 = -1$	0.78	1.00	0.74	0.93	0.99	0.99	0.97	0.88
	$\beta_2 = 0$	0.83	0.96	0.83	0.94	0.83	0.97	0.89	0.81
	$\beta_2 = 1$	0.82	0.98	0.81	0.89	0.84	0.93	0.83	0.83
$\beta_1 = -1$	$\beta_2 = -2$	0.87	1.00	0.82	0.99	0.96	0.91	0.84	0.87
	$\beta_2 = -1$	0.75	0.92	0.77	0.92	0.92	0.97	0.89	0.79
	$\beta_2 = 0$	0.86	0.92	0.81	0.92	0.90	0.98	0.90	0.81
	$\beta_2 = 1$	0.91	0.94	0.89	0.94	0.85	0.90	0.84	0.79
$\beta_1 = 0$	$\beta_2 = -2$	0.80	0.94	0.82	0.90	0.94	0.93	0.83	0.85
	$\beta_2 = -1$	0.80	0.88	0.82	0.90	0.88	0.93	0.81	0.76
	$\beta_2 = 0$	0.86	0.92	0.87	0.91	0.89	0.90	0.88	0.82
	$\beta_2 = 1$	0.85	0.91	0.84	0.89	0.82	0.89	0.83	0.81
$\beta_1 = 1$	$\beta_2 = -2$	0.80	0.91	0.83	0.94	0.91	0.95	0.85	0.77
	$\beta_2 = -1$	0.86	0.90	0.86	0.93	0.84	0.90	0.86	0.76
	$\beta_2 = 0$	0.86	0.92	0.87	0.93	0.87	0.91	0.83	0.73
	$\beta_2 = 1$	0.88	0.95	0.87	0.92	0.90	0.93	0.89	0.79

Table A.2: Estimated coverage probabilities of approximate 90% confidence intervals for simulation 2. BS=Bootstrap interval, Fisher= Information based interval.

untransformed parameters. This is not surprising as the sample sizes at the highest level of truncation should have still been around 100. Any variation from the nominal level can be plausibly explained by chance. The results for p show evidence of slight undercoverage by the Gaussian interval, while the percentile interval is still accurate.

A.3.2 Example 2

The second experiment consisted of 25 accidents involving two people in each accident. Again the population was simulated using model A.4. This simulation was performed with $\beta_0 = 0$ and over a grid of $\beta_1, \beta_2 = -2, -1, 0, 1$. This range of parameter values ensures that extreme data sets are very likely. As this was the case method 2 was used to calculate the percentile intervals. The coverage of these intervals is given in Table A.2.

The results in Table A.2 have several features. Firstly the percentile interval consistently undercovers the true parameter value, while the information based interval still performs adequately. This would appear to be a direct consequence of the method. The deletion of the infinite points in the bootstrap distribution can only shorten the percentile interval and hence lower the coverage rate. Again the coverage of the delta method variance estimate for p performs poorly. The percentile interval is an improvement, but still undercovers.

A.3.3 Example 3

The final example illustrates the use of the method on a real data set. The data set was extracted from the 1990 FARS fatal file compiled in the USA. It consists of frontal impact motorcycle accidents that involved two individuals, at least one of

	Bootstrap		Fisher	
parameter	lower	upper	lower	upper
β_0	-.43	3.1	-.511	1.48
β_1	-.51	1.31	-.41	.984
β_2	-.91	.35	-.86	.36
β_3	-1.64	1.11	-1.47	.97

Table A.3: 90 % confidence intervals for the parameters in the motorcycle data.

whom died. The extracted data set contained information on 37 accidents involving 74 individuals. The following model was fitted:

$$logit(Pr(die)) = \beta_0 + I_{\text{speed}>80\text{kmh}}\beta_1 + I_{\text{helmet worn}}\beta_2 + I_{\text{pillion}}\beta_3$$

where I_X is one if X is true, zero otherwise. The parametric bootstrap was iterated for 5000 replications, and percentile intervals were set including the divergent resamples. The estimated 90% percentile intervals for the parameters are shown in Table A.3, along with the intervals set using the Fisher information.

A.4 Discussion

The results of the simulations suggest several tentative conclusions. First, the asymptotic variance approximations for the maximum likelihood estimates appear to work well for quite small sample sizes. Thus we can have confidence in the coverage of the confidence intervals calculated using them. Note that the coverage of these intervals is only calculated conditional on samples that produced convergent estimates. This should not present a problem as this set of samples is the set that would be used to estimate β. If the data set produced divergent estimates alternative estimates would be attempted.

The second point to note is that the percentile method produced reasonable estimates, provided the divergent estimates were retained in the bootstrap resamples. This presents some problems, in that confidence intervals covering the whole real line can be constructed. This problem could be removed by limiting the number of replications of the Fisher scoring algorithm, so that the convergent estimates attain approximately the correct value, while the divergent estimates effectively maximise the likelihood without becoming infinite.

Finally in small sample sizes the delta method approximation to the variance of some function $g(\beta)$ performs worse than the percentile method. Thus in setting confidence intervals for functions of parameters the percentile method should be preferred.

Appendix B

Rejection Sampling from Exponential Tilted Dirichlet Random Variables

We consider the problem of generating random variables from the density,

$$f(\eta) \propto \prod_{i=1}^{n} \eta_i^{\alpha_i} \exp(-c_i \eta_i), \sum_{i=1}^{n} \eta_i = 1, c_i \geq 1. \tag{B.1}$$

Since $\sum_{i=1}^{n} \eta_i = 1$ we may take in equation B.1

$$f(\eta) \propto \prod_{i=1}^{n} \eta_i^{\alpha_i} \exp(-d_i \eta_i) \tag{B.2}$$

where

$$d_i = c_i - c_{\min},$$

and

$$c_{\min} = \text{minimum}(c_1, \ldots, c_n).$$

Without loss of generality, we may assume by permutation if necessary, that $d_n = 0$ in equation B.2. Then letting x be the solution of

$$x \sum_{i=1}^{n} \frac{\alpha_i}{(\alpha_n + d_i x)^{-1}} = 1, \tag{B.3}$$

it follows that $\prod_{i=1}^{n} \eta_i^{\alpha_i} \exp(-d_i \eta_i)$ is a strongly unimodal function with maximum at

$$v = \left(\frac{\alpha_i x}{\alpha_n + d_i x} \right)_{i=1,\ldots,n}. \tag{B.4}$$

We wish to use a combination of rejection sampling and Gibbs sampling to generate observations from equation B.1. To do this we require the curvature of f at v. But since v is the location of the maximum of f, the curvature of f at v is $f(v)$ by the curvature of $\log f$ at v. So

$$\left. \frac{\partial^2 f}{\partial \eta \partial \eta'} \right|_v = f(v) \left(\left. \frac{\partial^2 \log f}{\partial \eta \partial \eta'} \right|_v \right)$$

where we are differentiating with respect to

$$\eta' = (\eta_1, \ldots, \eta_{n-1}).$$

Now

$$-\left.\frac{\partial^2 f}{\partial\eta\partial\eta'}\right|_v = \text{diag}\left(\frac{1}{v_i^2}\right) + \frac{1}{v_n^2}E. \tag{B.5}$$

There appear to be two main options for approximating f.

The first alternative is to use an approximating Gaussian distribution for the density of $\eta' = (\eta_1, \ldots, \eta_{n-1})$. This method has the advantage of being able to exactly match the curvature of density B.1 at v and therefore potentially lower the rate of rejection. The drawback is that unlike the tilted Dirichlet where the observations are constrained to lie in the simplex, the Gaussian random variables can have individual entries which are less than zero or they can sum to greater than one. Either of these will cause the random vector to be rejected. However this will not become a serious problem until $n \approx 10$. The problem with this method is that the ratios obtained can be greater than one and so the rejection sampling is not valid. A similar approach using suitably chosen independent Beta random variables was also tried but suffered from similar problems.

The other alternative is to try to approximate the tilted Dirichlet by an ordinary Dirichlet. In order to get the maximum to occur at the same location, it is necessary to take a Dirichlet with density $\propto \prod_{i=1}^n \eta_i^{c_0 v_i}$ where c_0 is a constant that should be chosen to get the curvatures to match as closely as possible at v. The best possible choice turns out to be $c_0 = 1/x$. Unfortunately when this Dirichlet is used in a rejection sampling scheme, the rate of rejection can be so large that it makes the method unusable. This situation occurs when some of the d_i are very large, say fifty or more. This can happen for $n = 2$.

In order to overcome this difficulty, the n units are split into two groups, those with d_i less than some value, c say, and those with d_i greater than c. Without loss of generality we will assume that $d_n = 0$, $d_i \leq c, i = k+1, \ldots, n$ and $d_i > c, i = 1, \ldots, k$. Then since $1 - x \leq \exp(-x)$ for $x \in (0,1)$ it follows that

$$\begin{aligned}
\prod_{i=1}^n \eta_i^{\alpha_i} \exp(-d_i\eta_i) &= \left\{\prod_{i=1}^k \eta_i^{\alpha_i} \exp(-d_i\eta_i)\right\}\left\{\prod_{i=k+1}^{n-1} \eta_i^{\alpha_i} \exp(-d_i\eta_i)\right\} \times \\
&\quad (1 - \eta_{k+1} - \ldots - \eta_{n-1})^{\alpha_n}\left(1 - \frac{\eta_1 + \ldots + \eta_k}{1 - \eta_{k+1} - \ldots - \eta_{n-1}}\right)^{\alpha_n} \\
&\leq \prod_{i=1}^k \eta_i^{\alpha_i} \exp\left\{-\left(d_i + \frac{\alpha_i}{1 - \eta_{k+1} - \ldots - \eta_{n-1}}\right)\eta_i\right\} \times \\
&\quad \left\{\prod_{i=k+1}^{n-1} \eta_i^{\alpha_i} \exp(-d_i\eta_i)\right\}(1 - \eta_{k+1} - \ldots - \eta_{n-1})^{\alpha_n}. \\
&\leq \prod_{i=1}^k \eta_i^{\alpha_i} \exp\left\{-\left(d_i + \alpha_i\right)\eta_i\right\} \times \qquad \text{(B.6)} \\
&\quad \left\{\prod_{i=k+1}^{n-1} \eta_i^{\alpha_n} \exp(-d_i\eta_i)\right\}(1 - \eta_{k+1} - \ldots - \eta_{n-1})^{\alpha_n}. \qquad \text{(B.7)}
\end{aligned}$$

The recommended carrying density is to approximate the density B.7 by a suitably chosen Dirichlet as discussed above and the density B.6 by independent Gammas with shape α_i and scale $(d_i + \alpha_i)^{-1}$. It is reasonable to restrict the range of the Gamma random variables and generate another Gamma if this does not hold. It may also be tempting to generate another suite of Gammas if their sum is either

greater than one or even η_n say. However this has the effect of repeatedly generating Gammas for unfavourable $\eta_{k+1}, \ldots, \eta_n$ and so it is necessary to generate a complete new η vector in this case. A potential η should be accepted if

$$U \leq \prod_{i=1}^{k} exp(\frac{\alpha_n \eta_i}{1 - \sum_{j=k+1}^{n-1} \eta_j}) \left\{ \prod_{i=k+1}^{n-1} \left(\frac{\eta_i}{v_i^*} \right)^{\alpha_i - c_0 \alpha_n} \exp\{-d_i(\eta_i - v_i^*)\} \right\} \times \left(1 - \frac{\eta_1 + \ldots + \eta_k}{1 - \eta_{k+1} - \ldots - \eta_{n-1}} \right)^{\alpha_n}, \quad \text{(B.8)}$$

where $U \sim U(0,1)$ and v^* is calculated from $d_{k+1}, \ldots, d_n$ only using equations B.3 and B.4.

There are two reasons for a vector η to be rejected:

- The generated deviate may not lie in the simplex which automatically means that it cannot be a candidate for a tilted Dirichlet variable.
- The vector η may not satisfy the rejection sampling inequality given in equation B.8.

For large n, the combination of these two reasons can possibly lead to high probabilities of rejection and different methods are required. The method that we propose generates the vector η by Gibbs sampling where we generate the components of η in pieces of length k by rejection sampling based on equations B.6, B.7 and B.8. It is well known (Smith & Roberts, 1993) that by using multivariate sections of the vector rather than univariate sections, convergence can be greatly increased in situations where there is correlation between the components.

The comparison of the rejection sampling approach to the Gibbs sampling approximation considered previously is complicated by the dependence of the results on the particular situation considered. The convergence of the Gibbs subchain was examined by running parallel chains and examining the behaviour of the output over iterations, comparing this via Q-Q plots to a sample from the true distribution obtained via the rejection sampling algorithm. The results of this for $n = 30$ and a plausible set of λ and $P(\beta, x_i)$ showed that approximate convergence was obtained after n iterations, provided that η_n was taken to have $d_n = 0$. In this case the expected value of η_n is larger than that of the other η's, and the sampler can traverse the simplex more freely. In addition, the Gibbs subchain sampling of η is only a component of the full sampler, and approximate convergence is adequate for the convergence of the chain.

Bibliography

Abramowitz, M., & Stegun, I. (1972). *Handbook of Mathematical Functions.* Dover Publications, New York.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71*, 1–10.

Alho, J. (1991). Logistic regression in capture-recapture models. *Biometrics, 46*, 623–635.

Alho, J., Mulry, M., Wurdeman, K., & Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association, 88*, 1130–1136.

Amemiya, T. (1979). *Advanced Econometrics.* Chapman and Hall, New York.

Amemiya, T. (1984). Tobit models: a review. *Journal of Econometrics, 4*, 3–61.

Anderson, D. A., & Aitken, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society, Series B, Methodological, 47*, 203–210.

Bhattacharya, P., Chernoff, H., & Yang, S. (1983). Nonparametric estimation of the slope of a truncated regression. *The Annals of Statistics, 11*, 505–514.

Blumenthal, S. (1977). Estimating population size with truncated sampling. *Communications in Statistics: Part A - Theory and Methods, 6*, 297–308.

Blumenthal, S. (1985). Asymptotic expansions for modified maximum likelihood estimators with percentile truncated data. *Communications in Statistics: Part A - Theory and Methods, 14*, 905–925.

Blumenthal, S., Dahiya, R., & Gross, A. (1978). Estimating the complete sample size from an incomplete Poisson sample. *Journal of the American Statistical Association, 73*, 182–187.

Blumenthal, S., & Marcus, R. (1975). Estimating population size with Exponential failure. *Journal of the American Statistical Association, 70*, 913–922.

Bonney, G. (1987). Logistic regression for dependent binary observations. *Biometrics, 43*, 951–973.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*, 9–25.

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research. 1: The analysis of case-control studies.* 32. International Agency for Research on Cancer, Lyon.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., & Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology, 108*, 299–307.

Chambers, M. C., & Hastie, T. J. (1992). *Statistical models in S.* Wadsworth & Brooks/Cole, California.

Chant, D. (1974). On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika, 61*, 291–298.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics, 25*, 573–578.

Clogg, C. C., Rubin, D., Schenker, N., & Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association, 86*, 68–78.

Cohen, A. (1991). *Truncated and Censored samples: theory an applications.* Marcel Dekker, New York.

Connolly, M., & Liang, K.-Y. (1988). Conditional logistic regression for correlated binary data. *Biometrika, 75*, 501–506.

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics, 21*, 113–120.

Cox, D. R., & Hinkley, D. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Crouch, A. C., & Spiegelman, E. (1990). The evaluation of integrals of the form $\int f(t)exp(-t^2)dt$: application to logistic-normal models. *Journal of the American Statistical Association, 85*, 464–469.

Darroch, J. N. (1958). The multiple recapture census. 1: Estimation of a closed population. *Biometrika, 45*, 343–359.

Darroch, J. N., Fienberg, S., Glonek, G., & Junker, B. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association, 88*, 1137–1148.

Dempster, A. P., Laird, N., & Rubin, D. (1978). Maximum likelihood estimation from incomplete data via the EM algorithm(with discussion). *Journal of the Royal Statistical Society, Series B, Methodological, 39*, 1–38.

Diggle, P. J., Liang, K., & Zeger, S. (1994). *Analysis of longitudinal data.* Clarendon Press, Oxford.

Drum, M., & McCullagh, P. (1993). Discussion of "regression models for discrete longitudinal response". *Statistical Science, 8*, 284–309.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Evans, L. (1985). Double pair comparisons - a new method to determine how occupant characteristics affect fatality risk in traffic crashes. *Accident Analysis and Prevention*, *18*, 217–227.

Evans, L. (1991). *Traffic Safety and the Driver.* Van Nostrand Reinhold, New York.

Ferguson, T. S. (1973). A Bayesian analysis of some non parametric problems. *The Annals of Statistics*, *1*, 209–222.

Fienberg, S. E. (1972). The multiple-recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, *59*, 591–603.

Fitzmaurice, G., & Laird, N. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, *80*, 141–151.

Fitzmaurice, G., Laird, N., & Rotnitzky, A. G. (1993). Regression models for discrete longitudinal data. *Statistical Science*, *8*, 284–309.

Follmann, D. A., & Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, *84*, 295–300.

Fowlkes, W. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika*, *74*, 503–515.

Gelfand, A., & Smith, A. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelfand, A., Smith, A., & Lee, T. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*, 523–532.

Gelman, A., & Rubin, D. (1992a). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Gelman, A., & Rubin, D. (1992b). Rejoinder: Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Geman, S., & Geman, D. (1984). Stochastic relaxation:, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern analysis and Machine Intelligence*, *6*, 721–741.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, *7*, 473–511.

Gilks, W., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, *41*, 337–348.

Gilmour, A., Anderson, R., & Rae, A. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika*, *72*, 593–599.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, *78*, 45–51.

Greenland, S. (1994). Models risk ratios from matched cohort data: an estimating equation approach. *Applied Statistics*, *43*, 223–232.

Grogger, J. T., & Carson, R. (1991). Models for truncated counts. *Journal of Econometrics*, *6*, 225–238.

Gurmu, S. (1991). Tests for detecting overdispersion in the positive poisson regression model. *Journal of Business and Economic Statistics*, *9*, 215–222.

Gurmu, S., & Trivedi, P. (1992). Overdispersion tests for truncated Poisson regression models. *Journal of Econometrics*, *54*, 347–370.

Hartely, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, *14*, 174–194.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*, 320–340.

Harville, D. A., & Mee, R. (1984). A mixed-model procedure for analysing ordered categorical data. *Biometrics*, *40*, 393–408.

Hodeshima, J. (1988). The effect of truncation on the identifiability of regression coefficients. *Journal of the American Statistical Association*, *83*, 1055–1056.

Holford, T. R., White, C., & Kelsey, J. L. (1978). Multivariate analysis for matched case-control studies. *American Journal of Epidemiology*, *107*, 245–256.

Huber, P. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley Symposium in Mathematical Statistics and Probability*. Berkeley:University of California Press.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, *76*, 133–140.

Huggins, R. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, *47*, 725–732.

Im, S., & Gianola, D. (1988). Mixed models for binomial data with an application to lamb mortality. *Applied Statistics*, *37*, 196–204.

Jansen, J. (1990). On the statistical analysis of ordinal data when extra-variation is present. *Applied Statistics*, *39*, 75–84.

Jansen, J. (1993). Analysis of counts involving random effects with applications in experimental biology. *Biometrical Journal*, *35*, 395–447.

Jeffreys, H. (1961). *Theory of Probability* (third edition). Clarendon Press, Oxford.

Jennings, D. (1986). Outliers and residual distributions in logistic regression. *Journal of the American Statistical Association, 81*, 987–990.

Kuk, A. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society, Series B, Methodological, 56*, 395–407.

Laird, N. M. (1991). Topics in likelihood-based methods for longitudinal data analysis. *Statistica Sinica, 1*, 33–50.

Landwehr, J., Pregibon, D., & Shoemaker (1984). Graphical methods for assessing the logistic regression models. *Journal of the American Statistical Association, 79*, 61–71.

Liang, K. Y., & Zeger, S. (1986). Longitudinal data analysis using generalised linear models. *Biometrika, 73*, 13–22.

Liang, K. Y., Zeger, S., & Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B, Methodological, 54*, 3–40.

Little, R., & Rubin, D. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Liu, J. S., Wong, W., & Kong, A. (1994). Covariance stucture of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika, 81*, 27–40.

Lui, K. J., McGee, D., Rhodes, P., & Pollack, D. (1988). An application of conditional logistic regression to study the effects the effects of safety belts, principal impact points and car weights on drivers fatalities. *Journal of Safety Research, 19*, 197–203.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, Methodological, 42*, 109–142.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics, 11*, 59–67.

McCullagh, P. (1984). On the elimination of nuisance parameters in the proportional odds model. *Journal of the Royal Statistical Society, Series B, Methodological, 46*, 250–256.

McCullagh, P. (1985). On the asymptotic distribution of Pearsons statistic in linear exponential-family models. *International Statistical Review, 53*, 61–67.

McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *Journal of the American Statistical Association, 81*, 104–107.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models.* Chapman and Hall, New York.

McGilchrist, C. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Series B, Methodological, 56*, 61–69.

McLachlan, G., & Jones, P. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, *44*, 571–578.

Miller, M., & Landis, J. (1991). Generalized variance component models for clustered categorical response variables. *Biometrics*, *47*, 33–44.

Moran, P. (1971a). Maximum-likelihood estimation in non-standard situations. *Proceedings of the Cambridge Philosophical Society*, *70*, 441–450.

Moran, P. (1971b). The uniform consistency of maximum-likelihood estimators. *Proceedings of the Cambridge Philosophical Society*, *70*, 435–439.

Morgan, B. (1989). *Elements of simulation.* Chapman and Hall, New York.

Moulton, L. H., & Zeger, S. (1991). Bootstrapping generalized linear models. *Computational Statistics and Data Analysis*, *11*, 53–63.

Neuhaus, J., Kalbfleisch, J., & Hauck, W. (1991). A comparison of cluster-specific and and population averaged approaches for analysing correlated binary data. *International Statistical Review*, *59*, 25–36.

O'Neill, T. J., & Barry, S. C. (1993a). Truncated logistic regression. To appear in Biometrics.

O'Neill, T. J., & Barry, S. C. (1993b). Truncated ordinal regression. To appear in Statistics and Probability Letters.

Orme, C. (1989). On the uniqueness of the maximum likelihood estimator in truncated regression models. *Econometric Reviews*, *8*, 217–222.

Pierce, D., & Schafer, D. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, *81*, 977–986.

Pollock, K., Hines, J., & Nichols, J. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, *40*, 329–340.

Powell, J. (1986). Symmetrically trimmed least squares estimation for Tobit models. *Econometrica*, *54*, 1435–1460.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, *9*, 705–724.

Prentice, R. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, *44*, 1033–1048.

Prentice, R., & Zhao, L. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, *47*, 825–839.

Qaqish, B., & Liang, K.-Y. (1992). Marginal models for correlated binary responses with multiple classes an multiple levels of nesting. *Biometrics*, *48*, 939–950.

Qu, Y., Williams, G., Beck, G., & Goormastic, M. (1987). A generalized model of logistic regression for correlated binary data. *Communications in Statistics: Part A - Theory and Methods*, *16*, 3447–3477.

Raftery, A., & Lewis, S. (1992a). Comment: One long run with diagnostics: implementation strategies for Markov Chain Monte Carlo. *Statistical Science*, *7*, 493–497.

Raftery, A., & Lewis, S. (1992b). How many iterations in the Gibbs sampler?. In Bernardo, J., Berger, J., A.P., D., & Smith, A. (Eds.), *Bayesian Statistics*, Vol. 4, pp. 763–773. Oxford, U.K.: Oxford University Press.

Roberts, G., & Polson, N. (1994). On the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society, Series B, Methodological*, *56*, 377–384.

Rodríguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multi level models with binary responses. *Journal of the Royal Statistical Society: Series A*, *158*, 73–89.

Rosner, B. (1984). Multivariate methods in opthalmology with applications to other paired data situations. *Biometrics*, *40*, 1025–1034.

Rosner, B. (1989). Multivariate methods for clustered binary data with more than one level of nesting. *Journal of the American Statistical Association*, *40*, 373–380.

Rosner, B. (1992). Multivariate methods for clustered binary data with multiple subclasses, with application to binary longitudinal data. *Biometrics*, *48*, 721–731.

Royall, R. (1985). Model robust inference using Maximum Likelihood Estimators. *International Statistical Review*, *54*, 221–226.

Sanathanan, L. (1972a). A comparison of some models in visual scanning experiments. *Technometrics*, *14*, 813–830.

Sanathanan, L. (1972b). Estimating the size of a multinomial sample. *The Annals of Mathematical Statistics*, *43*, 142–152.

Sanathanan, L. (1972c). Models and estimation methods in visual scanning experiments. *Technometrics*, *14*, 813–830.

Sanathanan, L. (1977). Estimating the size of a truncated sample. *Journal of the American Statistical Association*, *72*, 669–672.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, *78*, 719–727.

Seber, G. (1982). *The Estimation of Animal Abundance.* Charles Griffin and Company, London.

Seber, G., & Wild, C. (1989). *Nonlinear regression.* Wiley, New York.

Self, S. G., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605–610.

Serfling, R. (1980). *Approximation theorems of mathematical statistics.* John Wiley and Sons, New York.

Shaw, D. (1988). On site samples' regression: Problems of non negative integers, truncation, and exogenous stratification. *Journal of Econometrics, 37*, 211–223.

Sleeper, L., & Harrington, D. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association, 85*, 941–949.

Smith, A., & Roberts, O. R. (1993). Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, Methodological, 55*, 3–23.

Stone, C., & Koo, C, Y. (1985). Additive splines in statistics. In *Proceedings of the Statistical Computing Section*, pp. 45–48. American Statistical Association.

Striatelli, R., Laird, N., & Ware, J. (1984). Random effects models for serial observations and binary responses. *Biometrics, 40*, 961–971.

Tanner, M. (1993). *Tools for Statistical Inference.* Springer-Verlag, New York.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*, 528–550.

Thompson, J. (1991). Conditional logistic regression in Genstat. *Genstat Newsletter, 26*, 6–10.

Tierney, L., & Kadane, J. (1985). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association, 81*, 82–86.

Tsui, K.-L., Jewell, N., & Wu, C. (1988). A nonparametric approach to the truncated regression problem. *Journal of the American Statistical Association, 83*, 785–792.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B, Methodological, 38*, 290–295.

Ware, J. (1985). Linear models for the analysis of serial measurements in longitudinal studies. *American Statistician, 39*, 95–101.

Watson, D., & Blumenthal, S. (1980a). Estimating the size of a truncated sample. *Communications in Statistics: Part A - Theory and Methods, 15*, 1535–1550.

Watson, D., & Blumenthal, S. (1980b). A two stage procedure for estimating the size of a truncated exponential sample. *Australian Journal of Statistics, 22*, 317–327.

Wedderburn, R. (1974). Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. *Biometrika, 61*, 439–447.

Weiss, A. A. (1992). The effects of helmet use on the severity of head injuries in motorcycle accidents. *Journal of the American Statistical Association, 87*, 48–56.

Weiss, A. A. (1993). A bivariate ordered Probit model with truncation: Helmet use and motorcycle injuries. *Applied Statistics, 42*, 487–499.

Welsh, A. H., Cunningham, R., Donnelly, C., & Lindenmeyer, D. (1994). Modelling the abundance of rare species: Statistical models for counts with extra zeroes. submitted.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1–25.

Wu, C. F. J. (1982). On the convergence properties of the EM algorithm. *Annals of Statistics, 11*, 95–103.

Wu, C. F. J. (1986). Jacknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics, 14*, 1261–1350.

Zeger, S., & Karim, M. R. (1991). Generalized linear models with random effects:a Gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79–86.

Zeger, S., Liang, K., & Albert, P. (1988). Models for longitudinal data: A generalised estimating equation approach. *Biometrics, 44*, 1049–1060.

Zeger, S., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42*, 121–130.

Zhao, L. P., & Prentice, R. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika, 77*, 642–648.

Zhao, L. P., Prentice, R., & Self, S. (1992). Multivariate mean parameter estimation using a partly exponential model. *Journal of the Royal Statistical Society, Series B, Methodological, 54*, 805–811.

1 Corrections

- Page 3, Line 28: *peoples* should be *people's.*
- Page 3, Line 31: *populations* should be *population's.*
- Page 10, Line 20 *occuring* should be *occurring.*
- Page 21, Line 21: *one to one* should be *one-to-one.*
- Page 25, Line 1 *occuring* should be *occurring.*
- Page 25, Line 7 Redundancy in *being being.*
- Page 26, line 14: Sentence should end with "?".
- Page 25, Line 3 from bottom: *occured* should be *occurred.*
- Page 37, Line 1: *occured* should be *occurred.*
- Page 37, Line 1: *were* should be *where.*
- Page 39, Line 6: *foundered* should be *floundered.*
- Page 39, Line 11: Redundancy in *in in.*
- Page 39, Line 29: *there* should be *their.*
- Page 51, Line 6 *occuring* should be *occurring.*
- Page 60, Line 2 from bottom: Redundancy in *was was.*
- Page 64, Line 32 *were divergent* should be *where divergent.*
- Page 64, Line 33: *occured* should be *occurred.*
- Page 76, Line 13 from bottom: Redundancy in *than than.*
- Page 77, Line 19 *occuring* should be *occurring.*
- Page 111, Line 2 from bottom: *groups* should be *group.*
- Page 113, Line 8 from bottom: change *be* to *can be made.*
- Page 121, Line 3 from bottom: change σ *either 0, 1 or 4* to σ *= 0,1, or 4.*
- Page 141, Line 29: *statisticians* should be *statistician's.*
- Page 141, Line 34: *goodness of fit* should be *goodness-of-fit.*
- Page 143, Line 26: *goodness of fit* should be *goodness-of-fit.*
- Page 148, Line 9: *one to one* should be *one-to-one.*
- An additional reference on Conditional Logistic Regression is

  Breslow, N.E. & Day, N.E. (1980) *Statistical Methods in Cance Research: Volume 2. The Design and Analysis of Cohort Studies.* IARC, London

- Additional references on MCMC methods that do not require tractable priors can be found in

  Besag, J., Green, P., Higdon, D., & Mengerson, K. (1995) Bayesian Computation and Stochastic Systems. *Statistical Science 10,*1-58

- Updated references from this thesis are

  O'Neill, T. J. & Barry, S. C. (1995) Group Truncated Ordinal Regression *Statistics and Probability Letters 22,*195-203

  O'Neill, T. J. & Barry, S. C. (1995) Truncated Logistic Regression *Biometrics 51,*533-542