Constrained Statistical Inference in Generalized Linear, and Mixed Models with Incomplete Data

by

Karelyn Alexandrea Davis

A thesis submitted to the Faculty of Graduate Studies and Postdoctoral Affairs

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in

Statistics

Carleton University

Ottawa, Ontario, Canada

June 2011

Your file Votre référence
ISBN: 978-0-494-83250-9
Our file Notre référence
ISBN: 978-0-494-83250-9

Abstract

For many years, statisticians have recognized the improvement in efficiency of many inference problems as a result of implementing the prior ordering of parameters or restrictions in the analysis. These restrictions often arise naturally, and have applications to many scientific and non-scientific disciplines. Moreover, as it is often the case that observations are not normally distributed and are sometimes observed in a cluster, Generalized Linear Models (GLMs) or Generalized Linear Mixed Models (GLMMs) are employed. Furthermore, many studies involve analysis of data for which some observations are unobserved, or missing. Previous research has indicated that omitting these incomplete values may lead to inefficient, and often biased results. A full unrestricted maximum likelihood estimation based on the likelihood of the responses has been well studied for estimation in clustered and missing-data situations. The present thesis will extend maximum likelihood estimation and likelihood ratio hypothesis testing

techniques for these popular models under linear inequality constraints on the parameters of interest. Such methods will improve upon previous results to incorporate general linear comparisons to nonlinear models, and allow for a wider variety of hypothesis tests. The innovative procedures avail of the gradient projection technique for maximum likelihood estimation, and chi-bar-square statistics for likelihood ratio tests. Theoretical and empirical results demonstrate the effectiveness of the maximum likelihood estimators and likelihood ratio tests under parameter constraints. The research is motivated by applications to clustered smoking data among Canadian youth, and an examination of missing variables in relation to contaminant detection of pregnant women for the Canadian government's Northern Contaminants Programme.

Acknowledgments

"Statistics...the most important science in the whole world: for upon it depends the practical application of every other science and of every art: the one science essential to all political and social administration, all education, all organization based on experience, for it only gives results of our experience."

- Florence Nightingale

"Please be good enough to put your conclusions and recommendations on one sheet of paper in the very beginning of your report, so I can even consider reading it."

- Sir Winston Churchill

I would like to gratefully acknowledge the encouragement and advice of my thesis supervisors, Dr. Chul Gyu Park and Dr. Sanjoy Sinha of the School of Mathematics and Statistics, Carleton University, throughout the preparation of

this thesis. I was always challenged to work harder and delve deeper, which has undoubtedly helped me to become a better researcher. A wonderful sense of collaboration and discovery was felt at every thesis meeting over the past four years, which I shall miss now that my research is completed.

I would also like to thank the members of my academic committee, Dr. Bhaskar Bhattacharya of Southern Illinois University, Dr. Craig Leth-Steevenson of Carleton University, Dr. J.N.K. Rao of Carleton University and Dr. Mayer Alvo of the University of Ottawa for their valuable comments on my thesis research.

The present research was supported by graduate student fellowships from the Natural Sciences and Engineering Research Council of Canada and the Faculty of Graduate Studies and Research, Carleton University.

As the majority of my thesis was completed as a part-time student, I would like to thank my government colleagues for their support throughout my research. In particular, thanks to Martin St-Pierre of Statistics Canada, and Mike Walker and Dr. Sheryl Bartlett of Health Canada for their encouragement. Also, special thanks to my former colleagues in the Methodology Branch and Special Surveys Division of Statistics Canada for their technical advice with respect to

the Youth Smoking Survey dataset. Similarly, thanks to Dr. Jay van Oostdam and members of the Northern Contaminants Programme at Health Canada, the Government of the Northwest Territories and Nunavut for their valuable discussions and assistance.

I would like to dedicate the present thesis to the memory of my teacher and mentor, Dr. Chu-In Charles Lee of the Department of Mathematics and Statistics, Memorial University of Newfoundland. Dr. Lee introduced me to the field of order-restricted statistical inference, and encouraged me to pursue doctoral studies. His high standards, both personally and professionally, were a positive influence in the early stages of my career, and will continue to impact my future ventures.

Lastly and most importantly, I would like to thank my family for their love and unwavering support in all my endeavours. To my parents, for instilling in my brother and I a love of learning, as well as the discipline, integrity and tenacity needed to succeed in research, as well as in life. Also, thanks to my brother Mark for his positive attitude, sense of humour and for always listening, especially during the more intense moments.

Table of Contents

List of Tables

List of Illustrations

List of Appendices

Chapter 1

Introduction

Statistical inference has roots in mathematical statistics and branches in applications to almost all areas of life and science. Modeling and analysis techniques for observational and experimental data often require methods to incorporate constraints in either the space of the unknown parameters or the sample space of observable random effects. While implementation of constraints in statistical analysis may complicate the analysis techniques, such constraints incorporate statistical information which may lead to improvements in efficiency over counterparts for which the constraints are ignored.

Natural orderings often occur in many interdisciplinary problems: A pharmaceutical researcher may desire a method to only declare a lower dose to be

efficacious if a higher dose is first found to be efficacious in a dose-response study; an economist may wish to test the validity of the Liquidity Preference Hypothesis, for which the term premium on a bill is a nondecreasing function of time to maturity; and a National Hockey League (NHL) owner may be interested in determining whether selecting players with a high ranking in the Entry Draft will lead to improved team performance (Ruberg 1989, Peng et. al., 2008, Richardson et. al., 1992, and Dawson and Magee, 2001).

The implementation of constraints in statistical analysis has been studied under various names, including one-sided testing, isotonic regression or restricted analysis. Advantages of using such constraints are that the restrictions are often natural, and allow for additional estimation and hypothesis tests; inference with constraints is often more efficient than unrestricted methods which ignore the constraints; and restricted maximum likelihood (ML) estimation has also been shown to obtain consistent estimates of parameters in most cases. On the other hand, such constraints require additional algorithms which may be complex or inefficient in terms of computing time to implement; and algorithms are usually proposed only for specific cases. The reader is referred to the books by Silvapulle and Sen (2005) and Robertson, Wright and Dykstra (1988) for further details.

In many instances, data follow a normal distribution, and statistical research

often involves linear modeling techniques. However, in many statistical applications, it is not the case that the mean of an observation is a linear combination of parameters nor that data are normally distributed. Moreover, a researcher may wish to accommodate overdispersion and correlation in the model by incorporating random effects. Such extensions to linear models are named Generalized Linear Models (GLMs) and Generalized Linear Mixed Models (GLMMs). These two models will be of primary interest in the present thesis.

Estimation and hypothesis testing for GLMs and GLMMs have been widely studied, and statistical software is well developed. The familiar logistic regression, loglinear regression and probit regression have been applied in many disciplines. These models are characterized by a nonlinear link function which relates the parameters of interest to various explanatory variables in a linear fashion. Inference methods, including maximum likelihood, quasi-likelihood and generalized estimating equations, utilize the link function in expressions for both parameter estimation and associated hypothesis tests. The texts by McCulloch, Searle and Neuhaus (2008) and McCullagh and Nelder (1989) review the available methods.

In particular, GLMMs are of importance in many statistical problems. Breslow and Clayton (1993) describe numerous applications including modeling of longitudinal data, overdispersion, spatial aggregation, etc. In effect, models are

built to accommodate correlated data or to consider levels of a factor as selected from a population of levels in order to make inference to that population (McCulloch, Searle and Neuhaus 2008). Nevertheless, the addition of random effects into a generalized linear model complicates procedures for estimating the model parameters. In essence, maximum likelihood estimation of parameters is preferred, however it is difficult and computationally intensive. Many alternatives to exact likelihood are available, such as integral approximations (quasi-likelihood), stochastic methods involving Monte Carlo methods, or other procedures such as generalized estimating equations. However, inefficiency and inconsistency of these methods in certain situations have overshadowed their computational advantages. As a result, the present thesis research will focus on maximum likelihood inference techniques.

Constrained inferences have been considered in many papers, since the early works of Eeden (1956) considered maximum likelihood estimation techniques, while Bartholomew (1959a, 1959b) developed a likelihood ratio test for equality of several normal means against ordered alternatives. Kudo (1963), Dykstra (1983), El Barmi and Dykstra (1994, 1995), El Barmi and Johnson (2006) and Dardanoni and Forcina (1998) have focused on inferences under the normal or multinomial setting. Gourieroux et. al. (1982) and Shapiro (1988) considered

estimation and testing under linear inequality restrictions in Gaussian linear models.

In the context of constrained inference for generalized linear models, important papers by Piergosch (1990), Silvapulle (1994) and Fahrmeir and Kligner (1994) detailed the properties of a one-sided or linear inequality hypothesis test of the regression parameters. The asymptotic null distribution for the ordered hypothesis was found to be chi-bar-squared, which is a mixture of chi-squared distributions. However, these authors did not consider the case with random effects or missing observations in their model setup. For generalized linear mixed models, very little research regarding constrained inference has been conducted. Lin (1997) proposed a score test for homogeneity in the GLMM, which was further improved by Hall and Praestgaard (2001) by the introduction of an order-restricted version. The latter test accounts for the assumption that covariance parameters corresponding to random effects must form a positive semidefinite matrix. Park et. al. (1998) and Pilla et. al. (2006) considered constrained estimation of parameters in a longitudinal setup. Dunson and Neelon (2003) developed a Bayesian approach for the order restricted test on the regression parameters in a generalized linear model. By way of theory and simulations, authors have demonstrated increases in testing power once constraints of this

type are correctly considered.

Furthermore, a common issue in modern applied statistics is that of incomplete data. Such problems include missing data, censored data, truncated data or grouped data, as investigated by Dempster, Laird and Rubin (1977) and Little and Rubin (2002). For maximum likelihood estimation with incomplete data, a standard method employs the Expectation-Maximization (EM) algorithm, with well-established convergence properties (Wu, 1983). The missing data mechanism is classified as being either ignorable, for which the missingness depends only on the observed values; or nonignorable, which implies the missingness depends on observed and missing data. The full likelihood is implemented in ignorable and nonignorable models, however the latter requires an additional model for the missing data mechanism.

In the context of linear modeling, estimation of order-restricted model parameters with incomplete data has been studied previously (Kim and Taylor 1995; Shi, Zheng and Guo 2005; Zheng, Shi and Guo, 2005). For linear mixture models, Jamshidian (2004) used a somewhat different approach, and proposed a globally convergent algorithm based on the gradient projection (GP) method which may be employed as part of the EM algorithm. The GP procedure considers both equality and inequality constraints and an algorithm may be derived.

The GP method was determined to be less tedious to implement and exhibit more stable convergence than competing procedures. Finally, Nettleton (1999) discusses various theoretical issues relating to the convergence properties of the EM procedure under order-restrictions.

1.1 Motivation and Statement of Problem

As discussed earlier, this research is motivated by the analysis of actual datasets collected from two different government studies.

First, in the recent Canadian Youth Smoking Survey (Health Canada, 2005), data were collected from students in grades 5 to 9, where schools were considered the primary sampling units. Observations from the same school are possibly dependent and may form a cluster. Of primary interest is the proportion of youth smokers, which tends to monotonically increase or decrease according to the student's age or family income level respectively (see Figure 1.1).

In the second problem, we consider data from the Canadian Northern Contaminants Programme (NCP) (Indian and Northern Affairs Canada, 2009), which collects data from pregnant women in Canada's northern territories. Maternal and umbilical cord blood samples were analyzed for levels of harmful contami-

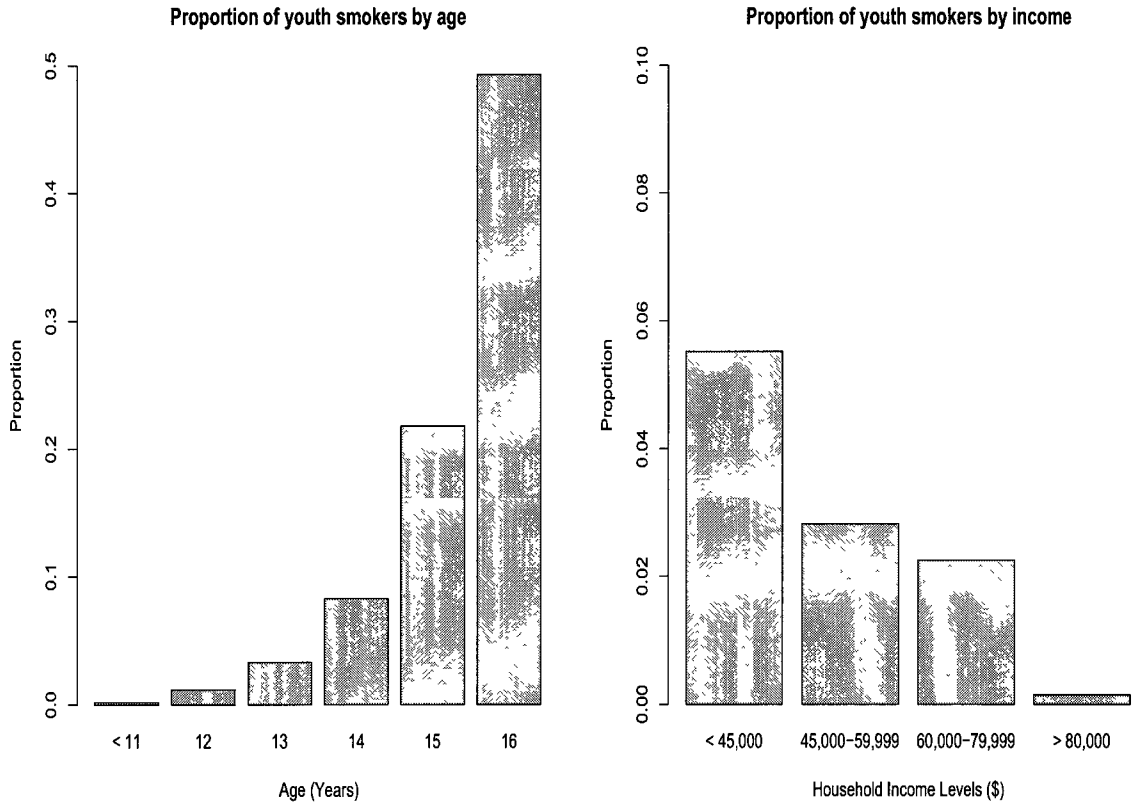

Figure 1.1: Proportion of youth smokers by age group and household income level (Statistics Canada: Youth Smoking Survey 2002)

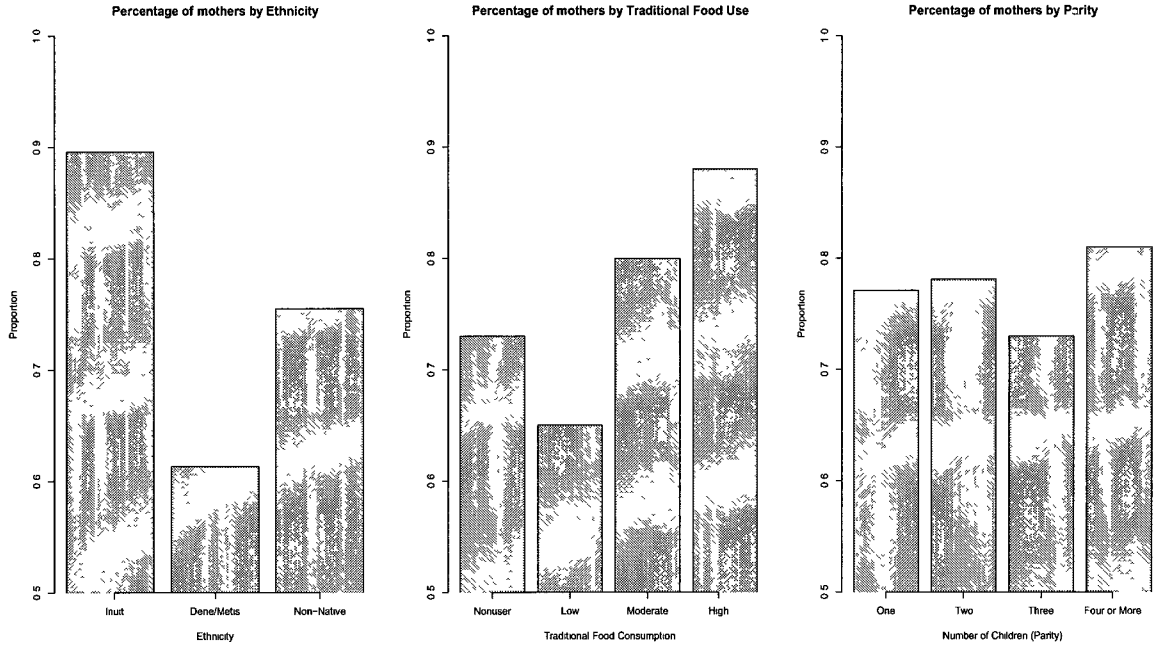

Figure 1.2: Proportion of mothers with *p'p*-DDT detected in maternal blood sample by covariates (Government of Canada: Northern Contaminants Programme)

nants such as metals (e.g. cadmium, lead, mercury) or other persistent organic pollutants such as pesticides and industrial polychlorinated biphenyls (PCBs). Researchers are interested in the probability of detecting these contaminants, which are believed to monotonically increase with consumption of traditional foods and smoking, and monotonically decrease with the number of children born to the mother, as per Figure 1.2 for the contaminant *p'p*-DDT. The smoking covariate exhibits many missing values, thus it is necessary to incorporate the missingness mechanism when analyzing this data.

For both applications, tendencies of certain regression parameters might be naturally, or empirically, presumed in a modeling stage and may be represented in terms of parameter constraints.

As there are no available methods to perform maximum likelihood inference for data displaying the above characteristics, we are motivated to develop constrained estimation techniques for nonlinear models, in the presence of clustering or incomplete data. Furthermore, the problem of extending the estimation of complete and incomplete-data problems for GLM and GLMM under constraints has not received much attention in the literature and is worthy of statistical consideration. In particular, as commented by Agresti and Coull (1998), "a variety of applications may benefit from order-restricted analysis in a generalized linear

mixed model format."

The present Ph.D. thesis develops maximum likelihood inference under constraints for two important cases: regression parameters of a generalized linear mixed model, and regression parameters of a generalized linear model with missing covariate data. The aforementioned gradient projection algorithm will be implemented in maximum likelihood estimation, which will subsequently be incorporated in constrained likelihood ratio tests. Asymptotic null distributions for constrained likelihood ratio tests will also be derived. Empirical results will be obtained to compare methods of estimation and testing, and finally, associated applications will be discussed.

1.2 Outline

The following chapters provide a further review of the literature and outline the proposed research in detail. In Chapter 2, some technical results concerning the models under study are discussed. A brief review of constrained inference and optimization techniques used throughout the thesis is provided in Chapter 3. The case of missing covariate data for generalized linear models with inequality constraints is detailed in Chapter 4, and an approach to maximum likelihood in-

ference is explored. Simulation studies and power comparisons are conducted to assess the performance of the constrained estimates and hypothesis tests, respectively. Chapter 5 extends previous research to the case of constrained generalized linear mixed models, by deriving an algorithm for maximum likelihood estimation as well as appropriate null distributions of hypothesis tests. In Chapter 6, the new constrained methods are applied to data from the Youth Smoking Survey and Northern Contaminants Programme. Chapter 7 summarizes the results and suggests areas for future research. Finally, a list of pertinent references and an appendix of relevant R code are offered at the end of the thesis.

Chapter 2

Generalized Linear and Mixed Models

In this section, we briefly review unconstrained inference in generalized linear and mixed models for cluster correlated data.

2.1 Generalized Linear Models

We first describe the generalized linear model (GLM). Consider the vector $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, which is assumed to have independent measurements from a distribution with density from the exponential family or similar distribution. In

other words,

$$\begin{aligned} y_i &\sim \text{indep. } f_{y_i}(y_i), \\ f_{y_i}(y_i) &= \exp\{[y_i\theta_i - b(\theta_i)]/\tau^2 - c(y_i,\tau)\}, \end{aligned} \tag{2.1}$$

where for convenience, the above distribution is in canonical form. An important aspect of these models is the link function, $g(\boldsymbol{\mu})$, which relates the mean of the distribution, $E[y_i] = \mu_i$, to the parameters by $g(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, with $\mathbf{x}_i^{\mathrm{T}}$ being the ith row of the model matrix for the fixed effects, and $\boldsymbol{\beta}$ the parameter vector of the linear predictor.

The log-likelihood function for these models is then defined as:

$$l(\boldsymbol{\beta},\tau) = \sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)]/\tau^2 - \sum_{i=1}^{n} c(y_i,\tau). \tag{2.2}$$

The expected value of the response variable is $E(y_i) = \mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i}$. In addition, the variance of the response variable may be written as

$$\mathrm{var}(y_i) = -E\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \theta_i^2}\right] = \tau^2\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} = \tau^2 v(\mu_i).$$

The score function, $s(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ may be written as:

$$s(\boldsymbol{\beta}) = \frac{1}{\tau^2}\mathbf{X}^{\mathrm{T}}V^*\boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu})$$

where $v_i^*(\mu_i) = [v(\mu_i)g_\mu^2(\mu_i)]^{-1}$, with $\boldsymbol{\Delta}$ as the diagonal matrix of $\{g_\mu(\mu_i)\}$ (derivative of $g(\mu_i)$ with respect to μ_i), and V^* as the diagonal matrix of $\{v_i^*(\mu_i)\}$.

Hence, with $\boldsymbol{\gamma} = (\boldsymbol{\beta}^{\mathrm{T}}, \tau^2)^{\mathrm{T}}$, the information matrix may be written as:

$$\mathcal{I}(\boldsymbol{\gamma}) = \begin{bmatrix} \mathcal{I}(\boldsymbol{\beta}) & \mathcal{I}(\boldsymbol{\beta}, \tau^2) \\ \mathcal{I}(\tau^2, \boldsymbol{\beta}) & \mathcal{I}(\tau^2) \end{bmatrix} = \begin{bmatrix} \mathcal{I}(\boldsymbol{\beta}) & \mathbf{0} \\ \mathbf{0} & \mathcal{I}(\tau^2) \end{bmatrix}$$

where

$$\begin{aligned} \mathcal{I}(\boldsymbol{\beta}) &= -E\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}\right] \\ &= -E\left[\frac{1}{\tau^2}\mathbf{X}^{\mathrm{T}}\frac{\partial V^* \boldsymbol{\Delta}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\mathbf{y} - \boldsymbol{\mu}) - \frac{1}{\tau^2}\mathbf{X}^{\mathrm{T}} V^* \boldsymbol{\Delta}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right] \\ &= \frac{1}{\tau^2}\mathbf{X}^{\mathrm{T}} V^* \mathbf{X}. \end{aligned} \tag{2.3}$$

The dispersion parameter, τ is present in the above expressions involving $\boldsymbol{\beta}$. In practice, such as for the binary and Poisson regression models, the dispersion parameter τ is fixed at unity. However, in other situations, such as the gamma distribution, an estimating equation for τ would be required, and may be found using a similar technique. Unless otherwise noted, we will assume $\tau = 1$.

From equation (2.2), the maximum likelihood estimating equations for $\boldsymbol{\beta}$ in the GLM are as follows:

$$\frac{1}{\tau^2}\mathbf{X}^{\mathrm{T}} V^* \boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}. \tag{2.4}$$

2.1.1 GLM Inference

There are two popular techniques to solve (2.4) in the generalized linear model setting: Newton–Raphson techniques and quasi-likelihood methods. The Newton–Raphson algorithm with Fisher scoring maximizes the likelihood function directly, while the quasi-likelihood method specifies only the mean and variance relationship, rather than the full likelihood. Further details on these methods are described in Section 2.2, for generalized linear mixed models.

Likelihood ratio tests for the $\boldsymbol{\beta}$ parameter may also be derived. Define $\boldsymbol{\beta}$ to be $\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})$ where interest lies in $\boldsymbol{\beta}_1$ while $\boldsymbol{\beta}_2$ is left unspecified. Suppose the hypothesis is of the form $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1,0}$, where $\boldsymbol{\beta}_{1,0}$ is a specified value of $\boldsymbol{\beta}_1$, and let $\hat{\boldsymbol{\beta}}_{2,0}$ be the MLE of $\boldsymbol{\beta}_2$ under the restriction that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1,0}$. The likelihood ratio test statistic is given by

$$-2\log\Lambda = -2\left[l(\boldsymbol{\beta}_{1,0}, \hat{\boldsymbol{\beta}}_{2,0}) - l(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)\right] \tag{2.5}$$

where $\hat{\boldsymbol{\beta}}^{\mathrm{T}} = (\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_2^{\mathrm{T}})$. The large sample critical region of the test is to reject H_0 in favor of the alternative when

$$-2\log\Lambda > \chi^2_{\nu,1-\alpha} \tag{2.6}$$

where ν is the dimension of $\boldsymbol{\beta}_1$ and $\chi^2_{\nu,1-\alpha}$ is the 100(1 - α)% percentile of

the χ^2 distribution with ν degrees of freedom. Further details are provided in McCulloch, Searle and Neuhaus (2008) and McCullagh and Nelder (1989).

2.2 Generalized Linear Mixed Models

In the generalized linear mixed model (GLMM) case, we may augment the link function $g(\cdot)$ to account for the clustering and/or longitudinal aspect of the data. We define $g(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}$, where $\mathbf{u}$ is the vector of random effects which accounts for such correlation or overdispersion; $\mathbf{x}_i^{\mathrm{T}}$ is the ith row vector of the predictor corresponding to y_i, $\mathbf{z}_i^{\mathrm{T}}$ is the ith row vector for the random effects, and $\boldsymbol{\beta}$ is the parameter vector of the linear predictor. A typical procedure begins with the conditional distribution of the response vector $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_k)^{\mathrm{T}}$, with density from the exponential family. We note that

$$\begin{aligned} \mathbf{y}_i|\mathbf{u} &\sim \text{indep. } f_{\boldsymbol{y}_i|\mathbf{u}}(\mathbf{y}_i|\mathbf{u}), \\ f_{\mathbf{y}_i|\mathbf{u}} &= \exp\{(y_i\theta_i - b(\theta_i))/\tau^2 - c(y_i,\tau)\} \quad \text{and} \\ \mathbf{u} &\sim f_{\mathbf{u}}(\mathbf{u}|\boldsymbol{\Sigma}). \end{aligned}$$

In general, we consider the marginal likelihood function, obtained by integrating over the random effects,

$$L(\boldsymbol{\beta}, \tau, \boldsymbol{\Sigma}|\mathbf{y}) = \int \prod_{i=1}^{k} f_{\mathbf{y}_i|\mathbf{u}}(\mathbf{y}_i|\mathbf{u}, \boldsymbol{\beta}, \tau) f_{\mathbf{u}}(\mathbf{u}|\boldsymbol{\Sigma}) d\mathbf{u}. \tag{2.7}$$

As before, we may assume the dispersion parameter τ equals 1 for most nonlinear models of interest such as the binary and Poisson regression models. Methods to compute the maximum likelihood estimates, and their inherent difficulties, are provided in the next section.

2.2.1 Estimation Techniques for GLMM

The ML estimators in GLMMs are obtained by maximizing the likelihood (2.7). The estimation involves integration with respect to $\mathbf{u}$, the vector of random effects. Such integration is problematic when the dimension of $\mathbf{u}$ is large, as these high-dimensional integrals are often intractable (see McCulloch, Searle and Neuhaus 2008). Numerous methods have been proposed in the literature to alleviate these computational difficulties. Most methods involve either approximating the integral, such as using quadrature methods (i.e. Gauss-Hermite or adaptive) or Monte Carlo methods. Other methods involve approximating the integrands, such as penalized or marginal quasi-likelihood methods. Finally, gen-

eralized estimating equations (Liang and Zeger 1986) are an alternative method of estimation which has many practical advantages. These procedures are described in more detail in the following sections. Note that the generalized linear model is a special case of the generalized linear mixed model without random effects. As a result, simpler procedures are appropriate for maximum likelihood estimates, since we do not need to integrate over the vector of random effects in the GLM setting.

In the unconstrained case, differentiation of the GLMM marginal likelihood function (2.7) leads to the following ML estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$,

$$E\left[\frac{\partial \ln f_{\mathbf{y}|\mathbf{u}}(\mathbf{y}|\mathbf{u},\boldsymbol{\beta},\tau)}{\partial \boldsymbol{\beta}}\,|\mathbf{y}\right] = \mathbf{0} \tag{2.8}$$

$$E\left[\frac{\partial \ln f_{\mathbf{u}}(\mathbf{u}|\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}}\,|\mathbf{y}\right] = \mathbf{0}, \tag{2.9}$$

where the expectation is with respect to the vector of random effects $\mathbf{u}$ given the response vector $\boldsymbol{y}$. Note that for the ith observation,

$$\frac{\partial \ln f_{y_i|\mathbf{u}}(y_i|\mathbf{u},\boldsymbol{\beta},\tau)}{\partial \boldsymbol{\beta}} = \{y_i - \mu_i(\boldsymbol{\beta},\mathbf{u})\}\,\mathbf{x_i}$$

with $\mu_i(\boldsymbol{\beta},\mathbf{u}) = E[y_i|\mathbf{u}] = b'(\theta_i)$.

2.2.2 Approximation of the Integral - Deterministic Methods

First, we consider numerical maximization methods from the optimization literature to determine the ML estimates. Assuming $\mathbf{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, if we let $\boldsymbol{\delta}_i = \boldsymbol{\Sigma}^{-1/2}\mathbf{u}_i$, then $\boldsymbol{\delta}_i$ follows a normal distribution with mean $\mathbf{0}$ and covariance matrix I and the linear predictor becomes $g(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}_i$. The variance components of $\boldsymbol{\Sigma}$ are now given in the linear predictor, and we write the likelihood as

$$\begin{aligned} L(\boldsymbol{\beta}, \tau, \boldsymbol{\Sigma}|\mathbf{y}) &= \int \prod_{i=1}^{k} f_{\mathbf{y}_i|\mathbf{u}}(\mathbf{y}_i|\mathbf{u}, \boldsymbol{\beta}, \tau) f_{\mathbf{u}}(\mathbf{u}|\boldsymbol{\Sigma}) d\mathbf{u} \\ &= \int \prod_{i=1}^{k} f_{\mathbf{y}_i|\boldsymbol{\delta}}(\mathbf{y}_i|\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \tau) f_{\boldsymbol{\delta}}(\boldsymbol{\delta}) d\boldsymbol{\delta}. \end{aligned} \tag{2.10}$$

In Gaussian quadrature, an integral is approximated by the weighted sum

$$\int f(z)\phi(z)dz \approx \sum_{q=1}^{Q} w_q f(z_q)$$

where Q is the order of approximation, with a higher Q indicating a more accurate approximation. The nodes or quadrature points, z_q, are solutions to the Qth order Hermite polynomial, while w_q are appropriately chosen weights. Both z_q and w_q are tabulated for various values (see Abramowitz and Stegun 1964),

or may be calculated via an algorithm for any value of Q (see McCulloch, Searle and Neuhaus 2008, page 328).

An alternative method known as adaptive Gaussian quadrature, forms quadrature points that are centered and scaled as though the function $f(z)\phi(z)$ originated from a normal distribution. While adaptive quadrature requires fewer quadrature points than classical Gaussian quadrature, the method is more time consuming due to the calculation of $\hat{z}$. Further, quadrature methods are limited to integrations involving products of functions of the form e^{-x^2}, i.e. for normally distributed random effects. McCulloch, Searle and Neuhaus (2008) note other modifications to incorporate exponential integrals or lognormally distributed random effects, however only in special cases. Also, quadrature is not always appropriate for models with crossed random effects or higher levels of nesting.

2.2.3 Approximation to the Integral - Stochastic methods

As an alternative to non-stochastic Gaussian quadrature methods, Markov Chain Monte Carlo (MCMC) integration methods may be used to simulate the likelihood rather than computing it directly. The algorithms begin with the notion

that the maximum likelihood integration is an expectation as follows:

$$L(\boldsymbol{\beta}, \tau, \boldsymbol{\Sigma}|\mathbf{y}) = \int f_{\mathbf{y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \tau) f_{\mathbf{u}}(\mathbf{u}|\boldsymbol{\Sigma}) d\mathbf{u} = E_{\mathbf{u}}[f_{\mathbf{y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \tau)]. \quad (2.11)$$

where $E_{\mathbf{u}}$ represents the expectation with respect to $\mathbf{u}$. A finite-sample approximation to this expectation is calculated by sampling m independent realizations $\mathbf{u}^1, \ldots, \mathbf{u}^m$ from $N(\mathbf{u}|\mathbf{0}, \boldsymbol{\Sigma})$ and then computing the sample average

$$L(\boldsymbol{\beta}, \tau, \boldsymbol{\Sigma}|\mathbf{y}) \approx \frac{1}{m} \sum_{h=1}^{m} f_{\mathbf{y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}^h, \boldsymbol{\beta}, \tau). \quad (2.12)$$

As m approaches infinity, the sample average converges to the true likelihood. If both k and m go to infinity, then the maximum likelihood estimators converge to their true values under appropriate regularity conditions. McCulloch (1997) and McCulloch, Searle and Neuhaus (2008) described two main Monte Carlo (MC) methods which may be applied to generalized linear mixed model estimation. The algorithms, denoted Monte Carlo Expectation Maximization (MCEM) and Monte Carlo Newton Raphson (MCNR) are each described in turn. In all cases, the MCMC methods require selection of random draws from the conditional distribution of $\mathbf{u}|\mathbf{y}$ to compute the aforementioned sum. A general method proposed in the literature for GLMM is the Metropolis-Hastings algorithm, which generates a Markov chain sequence of values that eventually stabilizes to draws from the candidate distribution.

2.2.4 Monte Carlo EM Algorithm

In the EM algorithm, the random effects are considered to be missing data, and the complete data vector $\mathbf{R} = (\mathbf{Y}, \mathbf{u})$ is formed. Then, the complete data log-likelihood is written as

$$\ln L_R = \sum_i \ln f_{y_i|u}(y_i|\mathbf{u}, \boldsymbol{\beta}, \tau) + \ln f_u(\mathbf{u}, \boldsymbol{\Sigma}). \tag{2.13}$$

Since $\boldsymbol{\beta}$ and τ are found only in the first term, the M step with respect to these parameters uses only $f_{\mathbf{y}|\mathbf{u}}$ (i.e. the generalized linear model portion of the likelihood) and is similar to a standard GLM computation with $\mathbf{u}$ known. In addition, maximizing with respect to $\boldsymbol{\Sigma}$ is the MLE of the distribution of $\mathbf{u}$ after replacement of sufficient statistics with their conditional expected values.

Further details and properties of the EM algorithm as it relates to maximum likelihood estimation with incomplete data is provided in Chapter 4.

2.2.5 Monte Carlo Newton-Raphson Algorithm

As a Newton-Raphson or scoring algorithm is often used in the absence of random effects, we may use Monte Carlo methods to extend these procedures to the generalized linear mixed model case. McCulloch (1997) showed that the term on the left side of (2.8) (without the expectation) may be expanded around the

value of $\boldsymbol{\beta}$ to obtain

$$\frac{\partial \ln f_{\boldsymbol{y}|\mathbf{u}}(y|\mathbf{U},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}} \cong \frac{\partial \ln f_{\boldsymbol{y}|\mathbf{u}}(\mathbf{y}|\mathbf{U},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0} + \frac{\partial^2 \ln f_{\boldsymbol{y}|\mathbf{u}}(\boldsymbol{y}|\mathbf{U},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0} (\boldsymbol{\beta}-\boldsymbol{\beta}_0), \tag{2.14}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\beta}^{\mathrm{T}}, \tau)^{\mathrm{T}}$. Then from (2.8) and (2.14), the iteration equation for estimating $\boldsymbol{\beta}$ may be expressed in the form

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{X}^{\mathrm{T}} E[V^*|\mathbf{y}]\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} E[V^*\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})|\mathbf{y}], \tag{2.15}$$

where $v_i^* = [v(\mu_i,\boldsymbol{\gamma})g_\mu^2(\mu_i,\boldsymbol{\gamma})]^{-1}$, with $\boldsymbol{\Delta}$ as the diagonal matrix of $\{g_\mu(\mu_i)\}$ (derivative of $g(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\mathbf{u}$ with respect to $\mu_i(\boldsymbol{\beta},\mathbf{u}) = E[\mathbf{y}|\mathbf{u}]$), and $V^*(\boldsymbol{\gamma},\mathbf{u})$ as the diagonal matrix with entries $\{v_i^*(\boldsymbol{\gamma},\mathbf{u})\}$.

For GLMMs, the expectations in (2.15) do not have closed form expressions and may be calculated using a stochastic procedure known as the Metropolis algorithm. This method produces random observations from the conditional distribution $\mathbf{u}|\boldsymbol{y}$, where the specification of the density $f_{\boldsymbol{y}}$ is not required. Then, Monte Carlo approximations to these expectations may be found, as in (2.12). The Metropolis algorithm begins by choosing $f_{\boldsymbol{u}}$ as the candidate distribution from which potential new draws originate. The acceptance function is then specified which provides the probability of accepting the new value as opposed to retaining the previous one. Suppose $\mathbf{u}$ is an outcome of the previous draw from

the conditional distribution of $\mathbf{u}|\boldsymbol{y}$. Generate a new value u_j^* for the jth element of $\mathbf{u}^* = (u_1, \ldots, u_{j-1}, u_j^*, u_{j+1}, \ldots, u_k)$ by using the candidate distribution $f_{\boldsymbol{u}}$. As suggested by McCulloch (1997), with probability

$$\alpha_j(\mathbf{u}, \mathbf{u}^*) = \min\left\{1, \frac{f_{\boldsymbol{u}|\boldsymbol{y}}(\mathbf{u}^*|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) f_{\boldsymbol{u}}(\mathbf{u}|\boldsymbol{\Sigma})}{f_{\boldsymbol{u}|\boldsymbol{y}}(\mathbf{u}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) f_{\boldsymbol{u}}(\mathbf{u}^*|\boldsymbol{\Sigma})}\right\}, \tag{2.16}$$

we accept the candidate value $\mathbf{u}^*$, otherwise reject and retain the previous value $\mathbf{u}$. The second term on the right hand side of (2.16) may be simplified to

$$\begin{aligned} \frac{f_{\boldsymbol{u}|\boldsymbol{y}}(\mathbf{u}^*|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) f_{\boldsymbol{u}}(\mathbf{u}|\boldsymbol{\Sigma})}{f_{\boldsymbol{u}|\boldsymbol{y}}(\mathbf{u}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) f_{\boldsymbol{u}}(\mathbf{u}^*|\boldsymbol{\Sigma})} &= \frac{f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}|\mathbf{u}^*, \boldsymbol{\beta})}{f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}|\mathbf{u}, \boldsymbol{\beta})} \\ &= \frac{\prod_{i=1}^k f_{\boldsymbol{y}_i|\boldsymbol{u}}(\boldsymbol{y}_i|\mathbf{u}^*, \boldsymbol{\beta})}{\prod_{i=1}^k f_{\boldsymbol{y}_i|\boldsymbol{u}}(\boldsymbol{y}_i|\mathbf{u}, \boldsymbol{\beta})}. \end{aligned} \tag{2.17}$$

In the above case, calculation of the acceptance function $\alpha_j(\mathbf{u}, \mathbf{u}^*)$ involves only the specification of the conditional distribution of $\boldsymbol{y}|\mathbf{u}$. The Metropolis step is then incorporated into the Newton–Raphson iterative equation (2.15) for the Monte Carlo estimates of the expected values, as outlined in Sinha (2004).

Other Monte Carlo methods have been proposed in the literature, including stochastic approximation methods (Gu and Kong, 1998) and simulated maximum likelihood (Geyer and Thompson, 1992 and Gelfand and Carlin, 1993). As these methods have been shown to exhibit convergence problems not seen in the other methods, recent literature suggests the MCEM and MCNR methods are

the most appropriate of the proposed stochastic procedures (McCulloch, 1997, McCulloch, Searle and Neuhaus, 2008).

2.2.6 Quasi-Likelihood Method

As the beneficial features of the quasi-likelihood method have been demonstrated for the generalized linear model, recent research has extended these methods to GLMMs. The most implemented approximation is that of Laplace, which begins with a second order Taylor series expansion of the form (Breslow and Clayton, 1993):

$$\log \int_{\mathbb{R}^q} e^{h(\mathbf{u})} d\mathbf{u} \doteq h(\mathbf{u}_0) + \frac{q}{2}\log 2\pi - \frac{1}{2}\log\left|-\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u}\partial \mathbf{u}^{\mathrm{T}}}\right|, \tag{2.18}$$

where $\mathbf{u}_0$ is the solution to

$$\left.\frac{\partial h(\mathbf{u})}{\partial \mathbf{u}}\right|_{\mathbf{u}=\mathbf{u}_0} = 0. \tag{2.19}$$

We then approximate the log likelihood of the GLMM as

$$\begin{aligned} l &= \log \int f_{\boldsymbol{y}|\mathbf{u}} f_{\mathbf{u}} d\mathbf{u} \\ &= \log \int e^{\log f_{\boldsymbol{y}|\mathbf{u}} + \log f_{\mathbf{u}}} d\mathbf{u} = \log \int e^{h(\mathbf{u})} d\mathbf{u} \end{aligned}$$

with $h(\mathbf{u}) = \log f_{\boldsymbol{y}|\mathbf{u}} + \log f_{\mathbf{u}}$. In developing the Laplace approximation, (2.19) must be solved and an expression for $\partial^2 h(\mathbf{u})/\partial \mathbf{u}\partial \mathbf{u}^{\mathrm{T}}$ is required. If we assume

$\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then

$$h(\mathbf{u}) = \log f_{\boldsymbol{y}|\mathbf{u}} + \log f_{\mathbf{u}} = \log f_{\boldsymbol{y}|\mathbf{u}} - \frac{1}{2}\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{u} - \frac{q}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}|.$$

Differentiating with respect to $\mathbf{u}$ gives

$$\frac{\partial h(\mathbf{u})}{\partial \mathbf{u}} = \frac{\partial \log f_{\boldsymbol{y}|\mathbf{u}}}{\partial \mathbf{u}} - \boldsymbol{\Sigma}^{-1}\mathbf{u} = \frac{1}{\tau^2}\mathbf{Z}^{\mathrm{T}}V^*\boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu}) - \boldsymbol{\Sigma}^{-1}\mathbf{u},$$

where V^* and $\boldsymbol{\Delta}$ were defined previously. Some algebra is required to obtain the second derivative $\partial^2 h(\mathbf{u})/\partial\mathbf{u}\partial\mathbf{u}^{\mathrm{T}}$, which is then substituted into equation (2.18). The entire approximate likelihood is then differentiated with respect to $\boldsymbol{\beta}$. The approximate score function for $\boldsymbol{\beta}$ also requires the assumption that the matrix V^* changes negligibly as a function of $\boldsymbol{\beta}$.

The estimating equations for $\boldsymbol{\beta}$ and $\mathbf{u}$ respectively, are to be solved simultaneously as:

$$\frac{1}{\tau^2}\mathbf{X}^{\mathrm{T}}V^*\boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0} \tag{2.20}$$

$$\frac{1}{\tau^2}\mathbf{Z}^{\mathrm{T}}V^*\boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}\mathbf{u}. \tag{2.21}$$

However, the above methods provide an estimate of $\boldsymbol{\beta}$ only, while another method must be implemented to estimate the variance matrix $\boldsymbol{\Sigma}$. Note that the Laplace approximation is a special case of adaptive Gaussian quadrature with one node (i.e. $Q = 1$).

Many authors have also shown that the two estimating equations (2.20) and (2.21) may also result from jointly maximizing the following equation, with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$:

$$PQL = \sum_i Q_i - \frac{1}{2}\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{u} = \log f_{Y|\mathbf{u}} - \frac{1}{2}\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{u}$$

which is similar to a quasi-likelihood term with a "penalty" function. Such methods are termed penalised quasi-likelihood (PQL) methods.

Despite the variety of derivations, implementation of the PQL method has met with difficulties. Research has demonstrated that the numerous approximations made throughout the procedure lead to estimators which are inconsistent, with paired binary data as a particular case (Breslow and Lin 1995, Lin and Breslow 1996). While attempts have been made to increase the order of the Taylor expansion to obtain better approximations, such as the sixth-order Laplace expansion (Laplace6) developed by Raudenbush et. al. (2000), the implementation may be more computationally involved than the Monte Carlo or direct maximization techniques, and exhibit reduced performance. Hence, in general, approximations to the integrand in the generalized linear mixed model case are not recommended (McCulloch, Searle and Neuhaus 2008).

2.2.7 Other Estimation Methods

Often implemented for longitudinal data, the generalized estimating equations (GEE) method (Liang and Zeger, 1986) assumes a marginal generalized linear model for the mean of the response $\mathbf{y}$ as a function of the predictors. While previous estimation techniques model the correlation between repeated measurements through the inclusion of random effects, conditional on which repeated measures are assumed independent; in the GEE approach, the association is modeled through a marginal working correlation matrix. Beginning with a working assumption of independence of all the elements of the response $\mathbf{y}$, the method leads to unbiased estimating equations. However, while parameter consistency is desirable, other authors have shown that the working independence assumption may lead to inefficient estimates and other working variance-covariance structures have been developed. Typically, we model $\mathbf{W}_i^{-1}$ as the working variance for $\mathbf{y}_i$ and update the estimating equations for $\boldsymbol{\beta}$ as

$$\sum \mathbf{X}_i \mathbf{W}_i \mathbf{y}_i = \sum \mathbf{X}_i \mathbf{W}_i E(\mathbf{y}_i).$$

While penalized quasi-likelihood and generalized estimating equations have many practical advantages, these methods do not have the optimal properties of statistical efficiency as those of maximum likelihood estimation and may lead

to inconsistent estimates in some cases. Thus, for both constrained and unconstrained problems, we will implement the exact likelihood method for finding the ML estimators of the parameters in GLMMs.

2.2.8 Hypothesis Testing for GLMM

Given the additional complexity of generalized linear mixed models, large sample tools are the preferred techniques available for statistical inference without constraints. It has been well documented that the likelihood ratio test (LRT) for such nested models may be performed by comparing $-2\log\Lambda$ to a chi-squared distribution, where Λ is the ratio of the likelihood under the null hypothesis values versus the alternative hypothesis of an unconstrained parameter vector.

Moreover, we may also perform a test for the random effect variance. As a simple case, consider the null hypothesis that a single variance component is equal to zero. McCulloch, Searle and Neuhaus (2008) note that the large-sample distribution is a 50/50 mixture of the constant 0 and a χ^2_1 distribution. It may be shown that the critical values of a level-α test are given by $\chi^2_{1,1-2\alpha}$, where $\chi^2_{1,1-2\alpha}$ is the $100(1-2\alpha)\%$ percentile of the χ^2 distribution with 1 degree of freedom. As with the likelihood equations, the LRT may only be computed using numerical techniques, which may be complicated in certain settings.

Chapter 3

A Brief Review of Constrained Statistical Inference and Optimization

As mentioned previously, constrained statistical inference has been proposed under various names in the literature, including order restricted analysis, isotonic regression, one-sided testing, etc. Review books by Robertson, Wright and Dykstra (1988) and Silvapulle and Sen (2005) highlight the main achievements in estimation and hypothesis testing. In this chapter we present some important

results which are utilized in later chapters. We begin with maximum likelihood inference for a one-way analysis of variance and progress to the multivariate setting.

3.1 Order Restricted Tests for Several Normal Means

We illustrate the beginnings of constrained inference by considering order restrictions for an analysis of variance (ANOVA) model. Denote a set of increasing dose levels by 1, 2, ..., k where 1 corresponds to the zero or control dose level. A one-way model is discussed, in which n_i experimental units are tested at the ith dose level, $i = 1, \ldots, k$. Let observations y_{ij} be mutually independent with $y_{ij} \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. Then $\bar{y}_i \sim N(\mu_i, \sigma^2/n_i)$, $i = 1, \ldots, k$ are the sample means of the dose groups and let $S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2/\nu$ be an unbiased estimate of the common variance σ^2, with $\nu = \sum_{i=1}^{k} n_i - k > 0$. Then S^2 is distributed as $\sigma^2 \chi^2_\nu/\nu$, independently of $\bar{y}_1, \ldots, \bar{y}_k$. The parameter space for this problem is defined as $\Omega = \{\boldsymbol{\mu} \in \mathbb{R}^k : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k\}$, with σ^2 as a nuisance parameter. The space Ω is known as the *simple order* in the constrained literature, since it denotes a

non-decreasing tendency among group means.

We next describe maximum likelihood inference for the one-way ANOVA model in the presence of parameter constraints.

3.1.1 Isotonic Regression for Monotone Trend

The restricted maximum likelihood estimator of $\boldsymbol{\mu}$ subject to Ω is denoted by $\boldsymbol{\mu}^\star = (\mu_1^\star, \ldots, \mu_k^\star)$ and is defined as the isotonic regression of $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_k)$ under Ω with sample sizes n_1, ..., n_k. As the observations are assumed to be normally distributed, the maximum likelihood estimate (MLE) is the solution to the following constrained weighted least squares problem:

$$\min_{\mu \in \Omega} \sum_{i=1}^{k} n_i (\bar{y}_i - \mu_i)^2. \tag{3.1}$$

The MLE is readily calculated using the Pool-Adjacent-Violators Algorithm (PAVA) (see Robertson, Wright and Dykstra, 1988). The process is essentially a successive averaging of $\bar{y}_i$'s until a sequence of non–decreasing values is obtained, and the MLE is represented as

$$\begin{aligned} \mu_j^\star &= \max_{i \le j} \min_{l \ge j} A(i,l), \quad j = 1, \ldots, k \quad \text{where} \\ A(i,l) &= \frac{\sum_{m=i}^{l} n_m y_m}{\sum_{m=i}^{l} n_m}. \end{aligned}$$

The MLE of μ may then be partitioned into consecutive sequences of equal–valued $\mu_j^\star$'s such that

$$\mu_1^\star = \cdots = \mu_{i_1}^\star < \mu_{i_1+1}^\star = \cdots = \mu_{i_2}^\star < \cdots < \mu_{i_{l-1}+1}^\star = \cdots = \mu_k^\star. \tag{3.2}$$

With the previous representation, the following results provide useful properties of the constrained MLEs in the simple ordering setting (see Robertson, Wright and Dykstra, 1988).

Lemma 3.1. The vector $\boldsymbol{\mu}^\star$ is the MLE of $\boldsymbol{\mu}$ if and only if

$$\begin{aligned} \sum_{i=1}^{k} n_i(\bar{y}_i - \mu_i^\star)\mu_i^\star &= 0 \qquad \text{and} \\ \sum_{i=1}^{k} n_i(\bar{y}_i - \mu_i^\star)\nu_i &\leq 0 \qquad \text{for all vectors } \nu_i \in \Omega. \end{aligned}$$

Lemma 3.2. If $\boldsymbol{\mu}^\star$ is the MLE of $\boldsymbol{\mu}$ given in (3.2), then for $r = 1, \ldots, l$,

$$\mu_{i_{r-1}+1}^\star = \cdots = \mu_{i_r}^\star = A(i_{r-1}+1,\ i_r). \tag{3.3}$$

Lemma 3.3. With $\boldsymbol{\mu}^\star$ as the MLE of $\boldsymbol{\mu}$, if $\mu_{i_r}^\star < \mu_{i_r+1}^\star$, it follows that $A(j, i_r) < A(i_{r-1}+1,\ i_r)$, $j = 1, \ldots, i_r$.

3.1.2 Likelihood Ratio Test under Order Restrictions

As is often the case in applications, a researcher may believe that the response means are monotone increasing, *a priori*, thus likelihood ratio tests (LRTs) for

homogeneity of normal means with simple order restrictions are introduced. We wish to derive such tests under the monotonicity assumption $\mu_1 \leq \cdots \leq \mu_k$. The LRT for ordered alternatives was introduced by Bartholomew (1959a,b, 1961a,b) and further discussed by Robertson, Wright and Dykstra (1988) as follows:

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_1: \quad \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$$

$$H_2: \quad \text{No restrictions on } \mu_i\text{'s.}$$

The LRT rejects H_0 in favour of $H_1 - H_0$ for large values of the test statistic

$$S_{01} = \frac{\sum_{i=1}^{k} n_i(\mu_i^\star - \bar{\mu})^2}{\sum_{i=1}^{k} n_i(\bar{y}_i - \mu_i^\star)^2/\nu + S^2},$$

where $\bar{\mu} = \sum_{i=1}^{k} n_i\bar{y}_i / \sum_{i=1}^{k} n_i$, the overall sample mean. When σ^2 is known, the test statistic is given by

$$\bar{\chi}^2_{01} = \frac{\sum_{i=1}^{k} n_i(\mu_i^\star - \bar{\mu})^2}{\sigma^2}. \tag{3.4}$$

As shown in Robertson, Wright and Dykstra (1988), as $\nu \to \infty$, the distribution of S_{01} approaches that of $\bar{\chi}^2_{01}$. Similarly, test statistics for testing H_1 against

$H_2 - H_1$ are given by

$$S_{12} = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \mu_i^\star)^2}{S^2},$$

and

$$\bar{\chi}^2_{12} = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \mu_i^\star)^2}{\sigma^2}. \tag{3.5}$$

The null distributions of S_{01}, S_{12}, $\bar{\chi}^2_{01}$ and $\bar{\chi}^2_{12}$ are

$$\begin{aligned}
P[S_{01} \geq s] &= \sum_{j=2}^{k} P_s(j,k;\mathbf{n}) P[F_{j-1,N-j} \geq \frac{s(N-j)}{\nu(j-1)}] \\
P[S_{12} \geq s] &= \sum_{j=2}^{k} P_s(j,k;\mathbf{n}) P[F_{k-j+1,N-k} \geq \frac{s(N-k)}{\nu(k-j+1)}] \\
P[\bar{\chi}^2_{01} \geq s] &= \sum_{j=2}^{k} P_s(j,k;\mathbf{n}) P[\chi^2_{j-1} \geq s] \qquad (3.6) \\
P[\bar{\chi}^2_{12} \geq s] &= \sum_{j=2}^{k} P_s(j,k;\mathbf{n}) P[\chi^2_{k-j+1} \geq s] \qquad (3.7)
\end{aligned}$$

for any $s > 0$, where $N = \sum_{i=1}^{k} n_i$, $\mathbf{n} = (n_1, \ldots, n_k)$, $P_s(j,k;\mathbf{n})$ is the level probability under H_0 that $\boldsymbol{\mu}^\star$ takes j distinct values, $F_{a,b}$ is an F-distribution with a and b degrees of freedom, and χ^2_c is a chi–squared variable with c degrees of freedom. For the case of equal weights, the level probabilities and the critical values are tabled in Robertson, Wright and Dykstra (1988). The null distributions of $\bar{\chi}^2_{01}$ and $\bar{\chi}^2_{12}$ are essentially weighted averages of chi-squared distributions. Hence, $\bar{\chi}^2_{01}$ and $\bar{\chi}^2_{12}$ are denoted the *chi-bar-square* distributions and

play a prominent role in constrained inference. We now discuss the calculation of level probabilities or chi-bar-square weights in more detail.

For the simply ordered case, i.e. with $\mu_1 \leq \cdots \leq \mu_k$, the level probabilities are denoted $P_s(l, k; \mathbf{n})$. When $k = 2$, the level probabilities are $P_s(1, 2; \mathbf{n}) = P_s(2, 2; \mathbf{n}) = \frac{1}{2}$. In general, no closed form of level probabilities are available for arbitrary sample sizes, however, if the sample sizes are equal, the level probabilities are more readily obtained. For this case, we simplify the notation and denote the level probabilities as $P_s(l, k)$. Robertson, Wright and Dykstra (1988) demonstrate that the $P_s(l, k)$ are distribution free over the collection of independent, identically distributed continuous random variables, i.e. the probability that the isotonic regression of y_1, y_2, ..., y_k with a simple order and equal weights has l level sets does not depend in the distribution of the y_i, provided they are independent with a common continuous distribution. Furthermore, an expression for the probability generating function of $\{P_s(l, k)\}$ is obtained and used to derive a recurrence relationship for the equal–weights level probabilities. In particular, the equal weight level probabilities are

$$P_s(1, k) = \frac{1}{k} \qquad \text{and} \qquad P_s(k, k) = \frac{1}{k!},$$

and

$$P_s(l,k) = \frac{1}{k}P_s(l-1,k-1) + \frac{k-1}{k}P_s(l,k-1)$$

for $l = 2, 3, \ldots, k-1$.

Similar results are available for testing problems with other orderings including simple tree and unimodal orders. In those cases, level probabilities are completely unknown unless k is very small even if the weights are equal.

In the next section, we review constrained inferences on the multivariate normal distribution with more general restricted parameter spaces.

3.2 Constrained Inference for Multivariate Normal Mean Vector

For multivariate analysis, constraints imposed on model parameters are often defined in terms of linear equality constraints (i.e. $A\boldsymbol{\theta} = \mathbf{c}$) and inequality constraints ($A\boldsymbol{\theta} \leq \mathbf{c}$), for a given A, $\boldsymbol{\theta}$ and $\mathbf{c}$. In this section, we discuss procedures for estimation and testing of these equality or inequality constraints.

3.2.1 Concepts and Definitions

In this section, we define some important technical terms which will be used throughout the paper. Let $\mathbb{R}^p$ denote the p-dimensional Euclidean space.

Definition 3.1. A set $\mathcal{A} \subset \mathbb{R}^p$ is said to be *convex* if $\{\lambda \mathbf{x} + (1-\lambda)\mathbf{y}\} \in \mathcal{A}$ whenever $\mathbf{x}$, $\mathbf{y} \in \mathcal{A}$ and $0 < \lambda < 1$. Therefore, $\mathcal{A}$ is a *convex set* if the line segment joining $\mathbf{x}$ and $\mathbf{y}$ is in $\mathcal{A}$ whenever the points $\mathbf{x}$ and $\mathbf{y}$ are in $\mathcal{A}$.

Definition 3.2. A set $\mathcal{A}$ is said to be a *cone with vertex* $\mathbf{x}_0$ if $\mathbf{x}_0 + k(\mathbf{x} - \mathbf{x}_0) \in \mathcal{A}$ for every $\mathbf{x} \in \mathcal{A}$ and $k \geq 0$. If the vertex is the origin, then we refer to the set $\mathcal{A}$ as a *cone*.

Definition 3.3. Let $\boldsymbol{V}$ be a $p \times p$ symmetric positive definite matrix, $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^p$. Then $< \mathbf{x}, \mathbf{y} >_{\boldsymbol{V}} = \mathbf{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \mathbf{y}$ defines an *inner product* on $\mathbb{R}^p$. This induces the corresponding *norm* $||\mathbf{x}||_{\boldsymbol{V}} =< \mathbf{x}, \mathbf{x} >_{\boldsymbol{V}}^{1/2}$. The corresponding *distance* between $\mathbf{x}$ and $\mathbf{y}$ is $||\mathbf{x} - \mathbf{y}||_{\boldsymbol{V}}$.

Definition 3.4. Let C be a closed convex set in $\mathbb{R}^p$ and $\mathbf{x} \in \mathbb{R}^p$. Let $\tilde{\mathbf{x}}$ in C be the point in C that is closest to $\mathbf{x}$ with respect to the distance $||\cdot||_{\boldsymbol{V}}$, i.e.

$$||\mathbf{x} - \tilde{\mathbf{x}}||_{\boldsymbol{V}} = \min_{\boldsymbol{\theta} \in C} (\mathbf{x} - \boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{V}^{-1} (\mathbf{x} - \boldsymbol{\theta}).$$

The vector $\tilde{\mathbf{x}}$ is the *projection* of $\mathbf{x}$ onto C and is denoted by $P_{\boldsymbol{V}}(\mathbf{x}|C)$, thus

$$\tilde{\mathbf{x}} = P_{\boldsymbol{V}}(\mathbf{x}|C) = \min_{\boldsymbol{\theta}\in C}(\mathbf{x}-\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\mathbf{x}-\boldsymbol{\theta}). \tag{3.8}$$

Definition 3.5. For any set $S \in \mathbb{R}^p$, we define the *orthogonal complement of S with respect to $\boldsymbol{V}$* as $\{\mathbf{y} \in \mathbb{R}^p : \mathbf{y}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{x} = 0 \text{ for all } \mathbf{x} \in S\}$. We also define the *polar or dual cone of S with respect to $\boldsymbol{V}$* as $\{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{y} \leq 0 \text{ for all } \mathbf{y} \in S\}$, which is denoted S^o.

3.2.2 Maximum Likelihood Inference

Let $\mathbf{y} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ be a $p \times 1$ normal random vector, C a closed convex cone in $\mathbb{R}^p$. Analogous to (3.1) the constrained maximum likelihood estimate of $\boldsymbol{\theta}$ under the cone C is the least squares projection of $\mathbf{y}$, that is, $\boldsymbol{\theta}^* = P_{\boldsymbol{V}}(\mathbf{y}|C)$. One possible approach to determine $\boldsymbol{\theta}^*$ is to use the gradient projection method, which is outlined in the next section. The advantage of this method is that it is directly applicable even when C is a translated cone with vertex other than the origin. We now present distributional properties of the likelihood ratio tests under linear inequality constraints.

Consider a testing problem for $H_0 : \boldsymbol{\theta} \in C_0 = \{\mathbf{z} \in \mathbb{R}^p : \mathbf{z} = \mathbf{0}\}$ versus

$H_1 - H_0$ where $H_1 : \boldsymbol{\theta} \in C$. Then, the likelihood ratio test (LRT) is given by

$$\bar{\chi}^2_{01}(\boldsymbol{V}, C) = \mathbf{y}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{y} - \min_{\boldsymbol{\theta}\in C}(\mathbf{y}-\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\mathbf{y}-\boldsymbol{\theta}) \tag{3.9}$$

$$= ||P(\mathbf{y}|C)||^2_{\boldsymbol{V}} \tag{3.10}$$

where the projection is taken with respect to the matrix $\boldsymbol{V}^{-1}$. When testing $H_1 : \boldsymbol{\theta} \in C$ against $H_2 - H_1$, where H_2 imposes no restriction on $\boldsymbol{\theta}$, the test statistic is

$$\bar{\chi}^2_{12}(\boldsymbol{V}, C) = \min_{\boldsymbol{\theta}\in C}(\mathbf{y}-\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\mathbf{y}-\boldsymbol{\theta}) \tag{3.11}$$

$$= ||\mathbf{y} - P(\mathbf{y}|C)||^2_{\boldsymbol{V}}. \tag{3.12}$$

Hence, the chi-bar-squared test statistics are expressed in terms of the distance between the origin or $\mathbf{y}$ and its projection onto a closed convex cone. Many distributional results concerning these models may be stated under the assumption of normality. These results are well summarized in Theorems 3.4 and 3.5 (see Silvapulle and Sen (2005) for further details).

Theorem 3.4. Let C be a closed convex cone in $\mathbb{R}^p$ and $\boldsymbol{V}$ be a $p \times p$ positive definite matrix. Then under H_0 we have

$$pr\{\bar{\chi}^2_{01}(\boldsymbol{V}, C) \leq c\} = \sum_{i=0}^{p} w_i(p, \boldsymbol{V}, C) pr(\chi^2_i \leq c), \tag{3.13}$$

$$pr\{\bar{\chi}^2_{12}(\boldsymbol{V}, C) \leq c\} \;=\; \sum_{i=0}^{p} w_{p-i}(p, \boldsymbol{V}, C) pr(\chi^2_i \leq c), \tag{3.14}$$

where $w_i(p, \boldsymbol{V}, C)$, $i = 0, \ldots, p$ are some nonnegative numbers and $\sum_{i=0}^{p} w_i(p, \boldsymbol{V}, C) = 1$.

The right hand side of equation (3.13) is the *chi-bar-square distribution*, and is a weighted mean of several tail probabilities of χ^2 distributions. The set $\{w_i(p, \boldsymbol{V}, C)\}$ is known as the *chi-bar-square weights* or simply *weights*.

We can also derive similar results even when the null parameter space $\{\mathbf{0}\}$ is replaced by a linear space contained in C. In particular, if the constraints are a linear inequality, then we have the following results:

Theorem 3.5. Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ where $\boldsymbol{V}$ is a positive definite matrix, $\boldsymbol{R}$ be a matrix of order $r \times p$, $\text{rank}(\boldsymbol{R}) = r \leq p$, and let $\boldsymbol{R}_1$ be a submatrix of $\boldsymbol{R}$ of order $q \times p$. Let the hypotheses be $H_0 : \boldsymbol{R\theta} = \mathbf{0}$, $H_1 : \boldsymbol{R_1\theta} \geq \mathbf{0}$ and H_2: no restrictions on $\boldsymbol{\theta}$, respectively. Also, let $\bar{\chi}^2_{01}$ and $\bar{\chi}^2_{12}$ denote the LRT statistics for testing H_0 versus $H_1 - H_0$ and H_1 versus $H_2 - H_1$ respectively. Then, under H_0 we have

$$pr[\bar{\chi}^2_{01} \leq c] \;=\; \sum_{i=0}^{q} w_i(q, \boldsymbol{R}_1 \boldsymbol{V} \boldsymbol{R}_1^{\mathrm{T}}, C) pr(\chi^2_{r-q+i} \leq c), \tag{3.15}$$

$$pr[\bar{\chi}^2_{12} \leq c] \;=\; \sum_{i=0}^{q} w_{q-i}(q, \boldsymbol{R}_1 \boldsymbol{V} \boldsymbol{R}_1^{\mathrm{T}}, C) pr(\chi^2_i \leq c). \tag{3.16}$$

where $C = \{\mathbf{z} \in \mathbb{R}^q : z_i \geq 0, i = 1, \ldots, q\}$. Note from the above expression that if there are no inequality constraints in the alternative hypothesis, then $q = 0$ and hence the classical chi-square test with r degrees of freedom is obtained.

One distinguishing factor of the $\bar{\chi}^2_{12}$ test is that the null hypothesis involves inequalities. Hence, the p-value depends on the underlying parameter $\boldsymbol{\theta}$, which may be anywhere in the null parameter space $\{\boldsymbol{\theta} : \boldsymbol{R}_1\boldsymbol{\theta} \geq \mathbf{0}\}$, for example. However, in order to obtain the critical value, c, which assures size α, we must solve $\sup_{\boldsymbol{R}_1\boldsymbol{\theta}\geq\mathbf{0}} P_{\boldsymbol{\theta}}[\bar{\chi}^2_{12} > c] = \alpha$. As explained in Silvapulle and Sen (2005), the supremum occurs at any $\boldsymbol{\theta}_0$ with $\boldsymbol{R}_1\boldsymbol{\theta}_0 = \mathbf{0}$, and hence $\boldsymbol{\theta} = \mathbf{0}$ is one such case. This particular null distribution is denoted the least favorable distribution.

Furthermore, note that closed form expressions for w_i exist only when the number of parameters is small (i.e. $p \leq 4$). Such closed-form weight expressions were given by Kudo (1963) and Shapiro (1988). If $p \geq 5$, simulated weights may be used as the LRT p-value is not sensitive to the weights. A standard approach to simulate the chi-bar-square weights $w_i(p, \boldsymbol{B}, \mathbb{R}^{+p})$, $i = 0, \ldots, p$, is given below:

Algorithm 3.1. Simulation Algorithm

(1) Generate $\boldsymbol{Z}$ from $N(\mathbf{0}, \boldsymbol{B})$. (2) Compute $\tilde{\boldsymbol{Z}}$, the point at which $(\boldsymbol{Z} - \boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{B}^{-1}(\boldsymbol{Z} - \boldsymbol{\theta})$ is a minimum over $\boldsymbol{\theta} \geq \mathbf{0}$. (3) Count the number of posi-

tive components of $\tilde{\boldsymbol{Z}}$. (4) Repeat the previous steps N times, (say $N = 10000$). (5) Estimate $w_i(p, \boldsymbol{B}, \mathbb{R}^{+p})$ by the proportion of times $\tilde{\boldsymbol{Z}}$ had exactly i positive components, $i = 0, 1, \ldots, p$.

Note that whenever the cone C is the positive orthant, i.e. $C = \mathbb{R}^{+p}$, then we write

$$w_i(p, \boldsymbol{V}) = w_i(p, \boldsymbol{V}, C). \tag{3.17}$$

The following theorem provides some theoretical results concerning the chi-bar-square weights, that may be applied when computing or simulating values.

Theorem 3.6. Let C be a closed convex cone in $\mathbb{R}^p$ and $\boldsymbol{V}$ be a $p \times p$ nonsingular covariance matrix. Then we have the following:

1. Let $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{V})$ and C be the nonnegative orthant. Then,

   $w_i(p, \boldsymbol{V}, C) = pr\{P_{\boldsymbol{V}}(\boldsymbol{Z}|C)$ *has exactly* i *positive components*$\}$.

2. $\sum_{i=0}^{p}(-1)^i w_i(p, \boldsymbol{V}, C) = 0$.

3. $0 \leq w_i(p, \boldsymbol{V}, C) \leq 0.5$.

4. Let C^o denote the polar cone, $\{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{y} \leq 0$ for every $\mathbf{y} \in C\}$, of C with respect to the inner product $< \mathbf{x}, \mathbf{y} >= \mathbf{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{y}$. Then, $w_i(p, \boldsymbol{V}, C^o) = w_{p-i}(p, \boldsymbol{V}, C)$.

5. Let $C = \{\boldsymbol{\theta} \in \mathbb{R}^p : \boldsymbol{R}\boldsymbol{\theta} \geq \mathbf{0}\}$ where $\boldsymbol{R}$ is a $p \times p$ nonsingular matrix. Then $\bar{\chi}^2(\boldsymbol{V}, C) = \bar{\chi}^2(\boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^{\mathrm{T}}, \mathbb{R}^p)$ and $w_i(p, \boldsymbol{V}, C) = w_i(p, \boldsymbol{R}\boldsymbol{V}\boldsymbol{R}^{\mathrm{T}})$.

6. $w_i(p, \boldsymbol{V}) = w_{p-i}(p, \boldsymbol{V}^{-1})$.

7. Let $C = \{\boldsymbol{\theta} \in \mathbb{R}^p : A\boldsymbol{\theta} \geq \mathbf{0}\}$ where A is a $r \times p$ matrix of rank $r \leq p$. Then

$$w_{p-r+i}(p, \boldsymbol{V}, C) = \begin{cases} w_i(r, A\boldsymbol{V}A^{\mathrm{T}}) & \text{for } i = 0, \ldots, r \\ 0 & \text{otherwise} \end{cases} \tag{3.18}$$

8. Let $C = \{\boldsymbol{\theta} \in \mathbb{R}^p : A_1\boldsymbol{\theta} \geq \mathbf{0}, A_2\boldsymbol{\theta} = \mathbf{0}\}$ where A_1 is $s \times p$, A_2 is $t \times p$, $s + t \leq p$, $[A_1^{\mathrm{T}}, A_2^{\mathrm{T}}]$ is of full rank, and

$$\boldsymbol{V}_{new} = A_1\boldsymbol{V}A_1^{\mathrm{T}} - (A_1\boldsymbol{V}A_2^{\mathrm{T}})(A_2\boldsymbol{V}A_2^{\mathrm{T}})^{-1}(A_2\boldsymbol{V}A_1^{\mathrm{T}}).$$

$$w_{p-s-t+j}(p, \boldsymbol{V}, C) = \begin{cases} w_j(s, \boldsymbol{V}_{new}) & \text{for } j = 0, \ldots, s \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

Other theoretical results regarding the mixing weights are given in Proposition 3.6.1 of Silvapulle and Sen (2005), as well as Kudo (1963) and Shapiro (1988).

3.3 Optimization Method

For equality and inequality constraints, we review the Kuhn-Tucker conditions for optimality and the Gradient Projection method to be implemented in estimation procedures for the models under consideration.

3.3.1 Kuhn-Tucker Conditions

Let $\mathbf{x}$ be an $n \times 1$ vector and $H(\mathbf{x})$ be an $m \times 1$ vector whose components $h_1(\mathbf{x}), \ldots, h_m(\mathbf{x})$ are differentiable concave functions of $\mathbf{x} \geq \mathbf{0}$. In addition, let $g(\mathbf{x})$ be a differentiable concave function of $\mathbf{x} \geq \mathbf{0}$. The Kuhn-Tucker equivalence theorem will determine an $\mathbf{x}^o$ that maximizes $g(\mathbf{x})$ constrained by $H(\mathbf{x}) \geq \mathbf{0}$ and $\mathbf{x} \geq \mathbf{0}$. A vector $\mathbf{x}$ is said to be feasible if it satisfies all given constraints. The optimal value of the problem is the maximum of $g(\mathbf{x})$ over the sets of feasible points. Those feasible points which attain the optimal value are called optimal solutions.

An inequality constraint $h_1(\boldsymbol{x}) \geq 0$ is *active* at a feasible point $\boldsymbol{x}$ if $h_1(\boldsymbol{x}) = 0$ and inactive if $h_1(\boldsymbol{x}) > 0$. Let $\left[\frac{\partial \phi}{\partial x_i}\right]^o$ and $\left[\frac{\partial \phi}{\partial \lambda_j}\right]^o$ denote the partial derivatives evaluated at a particular point $\mathbf{x}^o$ and $\boldsymbol{\lambda}^o$ respectively.

Theorem 3.7. (Equivalence Theorem). Let $h_1(\mathbf{x}), \ldots, h_m(\mathbf{x}), g(\mathbf{x})$ be concave as well as differentiable for $\mathbf{x} \geq \mathbf{0}$. Let $\phi(\mathbf{x}, \boldsymbol{\lambda}) = g(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}} H(\mathbf{x})$, where $H = (h_1, \ldots h_m)^{\mathrm{T}}$. Then $\mathbf{x}^o$ is a solution that maximizes $g(\mathbf{x})$ constrained by $H(\mathbf{x}) \geq \mathbf{0}$ and $\mathbf{x} \geq \mathbf{0}$ if and only if $\mathbf{x}^o$ and some $\boldsymbol{\lambda}^o$ satisfy the following conditions:

$$(1) \quad \left[\frac{\partial \phi}{\partial x_i}\right]^o \leq 0, \quad \left[\frac{\partial \phi}{\partial x_i}\right]^{o\mathrm{T}} x^o = 0, \quad \mathbf{x^o} \geq \mathbf{0};$$

$$(2) \quad \left[\frac{\partial \phi}{\partial \lambda_j}\right]^o \geq 0, \quad \left[\frac{\partial \phi}{\partial \lambda_j}\right]^{o\mathrm{T}} \lambda^o = 0, \quad \boldsymbol{\lambda}^o \geq \mathbf{0}.$$

(Theorem 3 Kuhn-Tucker 1951)

Simple modifications are made when the constraints $H(\mathbf{x}) \geq \mathbf{0}$, $\mathbf{x} \geq \mathbf{0}$ are changed to the following three cases:

Case 1. $H(\mathbf{x}) \geq \mathbf{0}$.

Here, using $\phi(\mathbf{x}, \boldsymbol{\lambda}) = g(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}} H(\mathbf{x})$ defined for all $\mathbf{x}$ and constrained only by $\boldsymbol{\lambda} \geq \mathbf{0}$, one must replace condition (1) by

$$(1^\star) \qquad \left[\frac{\partial \phi}{\partial x_i}\right]^o = 0.$$

Case 2. $H(\mathbf{x}) = \mathbf{0}$, $\mathbf{x} \geq \mathbf{0}$.

In this case, using $\phi(\mathbf{x}, \boldsymbol{\lambda}) = g(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}} H(\mathbf{x})$ defined for all $\boldsymbol{\lambda}$ and constrained

only by $\mathbf{x} \geq \mathbf{0}$, one must replace condition (2) by

$$(2^\star) \qquad \left[\frac{\partial \phi}{\partial \lambda_j}\right]^o = 0.$$

Case 3. $H(\mathbf{x}) = \mathbf{0}$.

Here, using $\phi(\mathbf{x}, \boldsymbol{\lambda}) = g(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}} H(\mathbf{x})$ defined for all $\mathbf{x}$ and $\boldsymbol{\lambda}$ without constraints, one must replace conditions (1) and (2) by (1*) and (2*). This corresponds to the familiar method of Lagrange multipliers.

As stated by Luenberger (2003), if it were known *a priori* which constraints were active at the solution to the optimization problem, the solution would be a local maximum point of the problem defined by ignoring the inactive constraints and treating all active constraints as equality constraints. Hence, with respect to local or relative solutions, the problem could be regarded as having equality constraints only. This observation suggests that the majority of theory applicable to the optimization problem may be derived by considering the equality constraints alone. In particular, if we define $\mathcal{A}$ to be the index set of active constraints, i.e. $\mathcal{A}$ is the set of i such that $H(\boldsymbol{x}^*) = \mathbf{0}$, then the necessary conditions from Case 1 are:

$$g'(\boldsymbol{x}) + \sum_{i \in \mathcal{A}} h_i'(\boldsymbol{x}) = \mathbf{0}$$

$$\begin{aligned} h_i(\boldsymbol{x}) &= 0 \quad i \in \mathcal{A} \\ h_i(\boldsymbol{x}) &> 0 \quad i \notin \mathcal{A} \\ \lambda_i &\geq 0 \quad i \in \mathcal{A} \\ \lambda_i &= 0 \quad i \notin \mathcal{A}. \end{aligned} \tag{3.20}$$

While the Kuhn-Tucker Theorem provides the first-order necessary conditions for optimality, it does not suggest an algorithm to find such an optimal point. The optimization literature discusses a gradient projection method, which determines the optimal value under inequality constraints satisfying (3.20).

3.3.2 Gradient Projection Theory

In essence, the gradient projection (GP) method involves projecting the gradient onto the working surface of active constraints to define the direction of movement. Jamshidian (2004) applied the method to optimize a likelihood function with linear equality and inequality constraints. More specifically, we begin with equality constraints of the form $A\boldsymbol{\beta} = \mathbf{c}$, where A is a $r \times p$ matrix of full rank ($r \leq p$), and $\mathbf{c}$ is an $r \times 1$ vector of known constants. The corresponding constrained parameter space is then $\Omega = \{\boldsymbol{\beta} : A\boldsymbol{\beta} = \mathbf{c}\}$. We define $\mathbb{R}^p$ to be a p-dimensional Euclidean space with the inner product $< \mathbf{x}, \mathbf{y} >_W = \mathbf{x}^{\mathrm{T}} W^{-1} \mathbf{y}$,

where W is a positive definite symmetric matrix. Let $s(\boldsymbol{\beta})$ denote the gradient of the log-likelihood function $l(\boldsymbol{\beta})$. The generalized gradient in the metric of W is then $\widetilde{s}(\boldsymbol{\beta}) = W^{-1}s(\boldsymbol{\beta})$. Note the objective function to maximize is $\phi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = l(\boldsymbol{\beta}) + \boldsymbol{\lambda}^{\mathrm{T}}(A\boldsymbol{\beta} - \mathbf{c})$.

Further, we let $\mathbf{d}$ denote a direction along which a move is made from a feasible point $\boldsymbol{\beta}_r \in \Omega$ to a new point $\widetilde{\boldsymbol{\beta}}_r = \boldsymbol{\beta}_r + \mathbf{d}$. We note that the new point $\widetilde{\boldsymbol{\beta}}_r$ will be in Ω if and only if $\mathbf{d}$ is in the null space of A_q, denoted $\mathcal{N} = \{\mathbf{d} \in \mathbb{R}^p : A_q\mathbf{d} = \mathbf{0}\}$. Here, A_q represents the rows of working constraints, such that A_q is an $m \times p$ matrix of rank m $(< p)$. The set $\mathcal{N}$ is also referred to as the space of feasible directions. Beginning with a point in Ω, the GP algorithm produces a sequence of feasible points by moving along feasible directions that converge to a solution satisfying the equality constraints, say $\hat{\boldsymbol{\beta}}_r$. We obtain the feasible direction at a point $\boldsymbol{\beta}_r \in \Omega$ by projecting $\widetilde{s}(\boldsymbol{\beta}_r)$ onto $\mathcal{N}$ in the matrix of W.

A formula for the direction $\mathbf{d}$ may be found by considering the space $\mathcal{O} = \{\mathbf{u} \in \mathbb{R}^p : \mathbf{u} = W^{-1}A_q^{\mathrm{T}}\boldsymbol{\lambda} \text{ for some } \boldsymbol{\lambda} \in \mathbb{R}^m\}$. The space $\mathcal{O}$ is conjugate to $\mathcal{N}$ in the metric of W defined by the aforementioned inner product. Since $\mathcal{O}$ and $\mathcal{N}$ are conjugate, there exists a $\mathbf{d} \in \mathcal{N}$ and $\boldsymbol{\lambda} \in \mathbb{R}^m$ such that

$$\widetilde{s}(\boldsymbol{\beta}) = \mathbf{d} + W^{-1}A_q^{\mathrm{T}}\boldsymbol{\lambda}. \tag{3.21}$$

If we multiply both sides of (3.21) by A_q and since $A_q\mathbf{d} = \mathbf{0}$ and $\text{rank}(A_q) = m$, we have

$$\boldsymbol{\lambda} = (A_q W^{-1} A_q^{\mathrm{T}})^{-1} A_q \widetilde{s}(\boldsymbol{\beta}). \tag{3.22}$$

If we then substitute (3.22) into equation (3.21) and solve for $\mathbf{d}$, we have

$$\mathbf{d} = P_W \widetilde{s}(\boldsymbol{\beta}), \quad \text{with} \tag{3.23}$$

$$P_W = I - W^{-1} A_q^{\mathrm{T}} (A_q W^{-1} A_q^{\mathrm{T}})^{-1} A_q, \tag{3.24}$$

where I is the identity matrix. As a result, the direction $\mathbf{d}$ is the projection of $\widetilde{s}(\boldsymbol{\beta})$ onto $\mathcal{N}$ in the metric of W. Furthermore, Jamshidian (2004) notes the following facts:

Theorem 3.8. (Jamshidian, 2004). Let $\mathbf{d}$ be defined as in equation (3.23). Then,

(1) $\mathbf{d}$ is an ascent direction with respect to the log-likelihood function $l(\cdot)$, and

(2) $\mathbf{d}$ is a generalized gradient of the log-likelihood function $l(\cdot)$ in $\mathcal{N}$ in the metric of W, with the aforementioned inner product.

Jamshidian (2004) further notes that by the Global Convergence Theorem of Luenberger (2003, page 187), since the Gradient Projection algorithm is a

generalized steepest ascent algorithm, it is globally convergent. Further, since $\mathbf{d}$ is an ascent and feasible direction, it is guaranteed that a small enough step from $\boldsymbol{\beta}_r$ in the direction of $\mathbf{d}$ results in a new feasible point $\tilde{\boldsymbol{\beta}}_r$ such that $l(\tilde{\boldsymbol{\beta}}_r) > l(\boldsymbol{\beta}_r)$. If the components of $\boldsymbol{\lambda}$ corresponding to the active constraints are all non-negative and with $s(\boldsymbol{\beta}_r) + \boldsymbol{\lambda}^{\mathrm{T}} A_q = \mathbf{0}$, then the Kuhn-Tucker conditions for the original problem are satisfied at $\boldsymbol{\beta}_r$ and the process terminates. Conversely, if at least one of the components of $\boldsymbol{\lambda}$ is negative, Luenberger (2003) shows that it is possible to move in a new direction to an improved point, by relaxing the corresponding inequality. A formal definition of the algorithm for inequality constraints is provided in the following section.

3.3.3 Gradient Projection Algorithm for Inequality Constraints

Consider inequality constraints of the form $A\boldsymbol{\beta} \leq \mathbf{c}$, where A is a $r \times p$ matrix of full rank ($r \leq p$), thus the constrained parameter space is $\Omega = \{\boldsymbol{\beta} : A\boldsymbol{\beta} \leq \mathbf{c}\}$. Jamshidian (2004) proposes the following gradient projection algorithm to find a solution to maximize the log-likelihood function $l(\boldsymbol{\beta})$ subject to

$$a_i^{\mathrm{T}}\boldsymbol{\beta} \;=\; c_i \quad i \in I_1,$$

$$a_i^{\mathrm{T}}\boldsymbol{\beta} \leq c_i \quad i \in I_2,$$

where the likelihood is assumed to be sufficiently smooth. The algorithm begins with an initial working set of active constraints, denoted $\mathcal{W}$. This set includes indexes of the constraints in I_1, if any, and may include indexes from I_2. Let $\overline{A}$ be an $\overline{m} \times p$ matrix whose rows consist of a_i^{T} for all $i \in \mathcal{W}$ and let $\overline{\mathbf{c}}$ be the corresponding vector of c_i's.

Beginning with an initial point $\boldsymbol{\beta}_r$ that satisfies $\overline{A}\boldsymbol{\beta}_r = \overline{\mathbf{c}}$, the algorithm proceeds as follows:

1. Compute $\mathbf{d} = P_W \widetilde{s}(\boldsymbol{\beta}_r)$ where $P_W = I - W^{-1}\overline{A}^{\mathrm{T}}(\overline{A}W^{-1}\overline{A}^{\mathrm{T}})^{-1}\overline{A}$.

2. If $\mathbf{d} = \mathbf{0}$, compute the Lagrange multipliers $\boldsymbol{\lambda} = (\overline{A}W^{-1}\overline{A}^{\mathrm{T}})^{-1}\overline{A}\widetilde{s}(\boldsymbol{\beta}_r)$.

   a) If $\lambda_i \geq 0$ for all $i \in \mathcal{W} \cap I_2$, stop. The current point satisfies the Kuhn-Tucker necessary conditions.

   b) If there is at least one $\lambda_i < 0$ for $i \in \mathcal{W} \cap I_2$, determine the index corresponding to the smallest such λ_i and delete the index from $\mathcal{W}$. Modify $\overline{A}$ and $\overline{\mathbf{c}}$ by dropping a row from each accordingly and go to Step 1.

3. If $\mathbf{d} \neq \mathbf{0}$, obtain $\alpha_1 = \max_\alpha\{\alpha : \boldsymbol{\beta} + \alpha\mathbf{d} \text{ is feasible}\}$. Then search for

$\alpha_2 = \max_\alpha\{l(\boldsymbol{\beta} + \alpha\mathbf{d}) : 0 \leq \alpha \leq \alpha_1\}$. Set $\widetilde{\boldsymbol{\beta}_r} = \boldsymbol{\beta}_r + \alpha_2\mathbf{d}$. Add indexes of new coordinates, if any, of $\widetilde{\boldsymbol{\beta}_r}$ that are newly on the boundary to the working set $\mathcal{W}$. Modify $\overline{A}$ and $\overline{\mathbf{c}}$ by adding additional rows.

4. Replace $\boldsymbol{\beta}$ by $\widetilde{\boldsymbol{\beta}_r}$ and go to Step 1, continuing until convergence.

We define the matrix W to be the estimate of the variance matrix at the current value of $\widetilde{\boldsymbol{\beta}_r}$. As will be demonstrated in Chapters 4 and 5, the observed information matrix is used to estimate W, and we extend the GP algorithm to constrained GLM with missing covariates and GLMM problems.

3.3.4 Illustration of the Gradient Projection Algorithm

To illustrate the GP Algorithm and its relation to the Kuhn–Tucker optimality conditions, we consider a simple example. Let $\boldsymbol{y} = (y_1, y_2)^{\mathrm{T}} \sim N(\boldsymbol{\mu} = (\mu_1, \mu_2)^{\mathrm{T}}, I_2)$ with the equality constraint $\mu_1 = 2\mu_2$. In this case, the objective function is $\phi(\boldsymbol{\mu}, \lambda) = l(\boldsymbol{\mu}) + \lambda(\mu_1 - 2\mu_2)$, where $l(\boldsymbol{\mu})$ is log-likelihood function of $\boldsymbol{y}$. From Theorem 3.7 Case (3), the necessary conditions for optimality are:

$$\begin{aligned} \frac{\partial\phi}{\partial\boldsymbol{\mu}} &= \begin{bmatrix} (y_1 - \mu_1) + \lambda \\ (y_2 - \mu_2) - 2\lambda \end{bmatrix} = \mathbf{0}, \\ \frac{\partial\phi}{\partial\lambda} &= \mu_1 - 2\mu_2 = 0. \end{aligned} \tag{3.25}$$

Hence, the equations to solve are $\lambda = -(y_1 - \mu_1) = (y_2 - \mu_2)/2$ and $\mu_1 = 2\mu_2$. After some algebra, we obtain the optimal solutions, $\boldsymbol{\mu}^*$ to be

$$\boldsymbol{\mu}^* = \frac{1}{5}\begin{bmatrix} 4y_1 + 2y_2 \\ 2y_1 + y_2 \end{bmatrix}, \tag{3.26}$$

which satisfy the constraints.

We next implement the GP algorithm and compare to the solutions obtained in equation (3.26). Since we have only one constraint, $A_q = A_1 = [1 \quad -2]$, and $\boldsymbol{\lambda} = \lambda$. Also, choose an initial point $\boldsymbol{\mu}^r = (\mu_1^r, \mu_2^r)^{\mathrm{T}} = ((y_1 - y_2), (y_1 - y_2)/2)^{\mathrm{T}}$ which satisfies the constraint.

Then, with $W = I_2$, the distance vector is calculated as $\boldsymbol{d} = P_W s(\boldsymbol{\mu})$ where

$$\begin{aligned} P_W &= I - A_1^{\mathrm{T}}(A_1 A_1^{\mathrm{T}})^{-1} A_1 \\ &= I - \frac{1}{5}\begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix} \\ &= \frac{1}{5}\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}, \end{aligned}$$

and $s(\boldsymbol{\mu}) = (y_1 - \mu_1, y_2 - \mu_2)^{\mathrm{T}}$ is the score vector. Hence

$$\boldsymbol{d} = P_W s(\boldsymbol{\mu}) = \frac{1}{5}\begin{bmatrix} 4(y_1 - \mu_1) + 2(y_2 - \mu_2) \\ 2(y_1 - \mu_1) + (y_2 - \mu_2) \end{bmatrix}. \tag{3.27}$$

Note also that

$$A_1\boldsymbol{d} = [1 \quad -2]\,\frac{1}{5}\begin{bmatrix} 4(y_1-\mu_1)+2(y_2-\mu_2) \\ 2(y_1-\mu_1)+(y_2-\mu_2) \end{bmatrix} = 0.$$

According to Step 1 of the GP algorithm, we evaluate $\boldsymbol{d} = (d_1, d_2)^{\mathrm{T}}$ at the initial point $\boldsymbol{\mu}^r$. We have

$$\begin{aligned} d_1 &= \frac{1}{5}(4(y_1-\mu_1)+2(y_2-\mu_2)) \\ &= -\frac{1}{5}(y_1-7y_2), \quad \text{and} \\ d_2 &= \frac{1}{5}(2(y_1-\mu_1)+(y_2-\mu_2)) \\ &= -\frac{1}{10}(y_1-7y_2). \end{aligned}$$

Since $\boldsymbol{d} \neq \boldsymbol{0}$ at $\boldsymbol{\mu}^r$, then according to Step 3, we must find a new point $\boldsymbol{\mu}^{r+1} = \boldsymbol{\mu}^r + \alpha\boldsymbol{d}$ which is feasible and maximizes $l(\boldsymbol{\mu}^{r+1})$. To find α, we maximize

$$l(\alpha) = -\frac{(y_1-\mu_1^{r+1})^2}{2} - \frac{(y_2-\mu_2^{r+1})^2}{2} - \frac{1}{2}\log(2\pi)$$

where $\mu_1^{r+1} = \mu_1 + \alpha d_1$ and $\mu_2^{r+1} = \mu_2 + \alpha d_2$. Then,

$$\begin{aligned} \frac{\partial l(\alpha)}{\partial \alpha} &= (y_1-\mu_1^{r+1})d_1 + (y_2-\mu_2^{r+1})d_2 \\ &= (y_1-\mu_1)d_1 + (y_2-\mu_2)d_2 - \alpha(d_1^2+d_2^2). \end{aligned} \tag{3.28}$$

Setting (3.28) to zero and solving for α, we obtain

$$\alpha = \frac{(y_1-\mu_1)d_1 + (y_2-\mu_2)d_2}{d_1^2+d_2^2}. \tag{3.29}$$

Next, at $\boldsymbol{\mu}^r$ we substitute and find

$$\begin{aligned} d_1^2 + d_2^2 &= \frac{(y_1 - 7y_2)^2}{20} \\ (y_1 - \mu_1)d_1 &= -\frac{y_2}{5}(y_1 - 7y_2) \quad \text{and} \\ (y_2 - \mu_2)d_2 &= \frac{(y_1 - 3y_2)(y_1 - 7y_2)}{20}. \end{aligned}$$

After some algebra, we have $\alpha = 1$. Therefore the updated solution vector is

$$\begin{aligned} \boldsymbol{\mu}^{r+1} &= \boldsymbol{\mu}^r + \alpha \boldsymbol{d} \\ &= \begin{bmatrix} y_1 - y_2 \\ (y_1 - y_2)/2 \end{bmatrix} + 1 \cdot \begin{bmatrix} -1/5(y_1 - 7y_2) \\ -1/10(y_1 - 7y_2) \end{bmatrix} \\ &= \frac{1}{5}\begin{bmatrix} 4y_1 + 2y_2 \\ 2y_1 + y_2 \end{bmatrix}, \end{aligned} \tag{3.30}$$

and we return to Step 1. We then re-calculate $\boldsymbol{d}$ at the new value $\boldsymbol{\mu}^{r+1}$, for which $\boldsymbol{d} = (0, 0)^{\mathrm{T}}$. Hence, from Step 2, we stop and declare that the GP algorithm has converged. We note that the solution to the GP algorithm (3.30) is identical to the one obtained from the Kuhn–Tucker conditions, namely (3.26).

Chapter 4

Inference for GLM with Incomplete Covariate Data under Inequality Constraints

There are diverse situations for which missing or incomplete data is prevalent in a statistical analysis. For example, attrition in longitudinal studies, nonresponse in survey sampling, and latent variable patterns for which certain variables are never observed, constitute frequent situations for which missing data methods are required. However, as mentioned in the Introduction, few methods have in-

corporated constrained inference for modeling techniques with incomplete data. Moreover, while methods have been proposed for constraints in linear models and linear mixed models, extensions to GLMs have received little attention, particularly for the case of missing observations. The following sections summarize important aspects of incomplete data problems and propose extensions to constrained inference in GLMs with missing data.

4.1 Statistical Inference with Incomplete Data

Conventional statistical methods require all covariates to be observed, thus when subjects with incomplete covariate information differ from those with complete data with respect to the characteristic under study, a traditional analysis using only the observed data may be biased and invalid. A common approach is to analyze only the complete cases (CC), or those subjects to which all covariates are observed. However, it is well established that a CC analysis may be biased when the data are not missing completely at random (MCAR), i.e. when the observed data may not be considered as a random sample of the complete data (Ibrahim et. al., 2005). Any method yielding valid inferences in the absence of missing data will also yield valid inferences when data are missing completely at

random and the analysis is based on all data, or even when restricted to those complete cases with no missing data (Sinha, 2008).

A less restrictive assumption than MCAR assumes the missingness depends only on the observed data, and not on the incomplete data. In this case, the mechanism is referred to as missing at random (MAR). Rubin (1976) showed that if the data are MAR, then the likelihood-based inference does not depend on the missing-data mechanism. On the other hand, if missingness is associated with values of the missing variables, then the mechanism is said to be not missing at random (NMAR). For NMAR data, an additional model for the missing data mechanism is required, as discussed by Little (1995).

The seminal textbook of Little and Rubin (2002) and the review paper by Ibrahim et. al. (2005) summarize the many available procedures to analyze incomplete data problems. For unconstrained generalized linear models, many authors have focused on maximum likelihood techniques, however other procedures such as multiple imputation, fully Bayesian and weighted estimating equations (Robins, Rotnitzky and Zhao, 1994) are also available. Maximum likelihood methods depend upon the structure of the missing data, and include factorization techniques, Newton-Raphson or quasi-Newton-Raphson methods to directly maximize the observed data likelihood, as well as the EM (Expectation-

Maximization) algorithm of Dempster, Laird and Rubin (1977) to obtain maximum likelihood estimates from the complete-data likelihood. The EM algorithm is a popular iterative method to obtain maximum likelihood estimates, both for missing data problems, as well as estimation of other non-incomplete data problems such as variance components and factor analysis (Becker, Yang and Lange, 1997). The E step of EM calculates the conditional expectation of the missing data given the observed data and current estimates of the parameters, and then substitutes these expectations for the missing data or some functions therein. The M step performs maximum likelihood estimation as if there were no missing data, i.e. using the estimates obtained from the E step. The algorithm iterates between these two steps until convergence is achieved.

4.2 Missing Data Mechanisms

The literature on missing value problems has distinguished between what is referred to as the missing-value *pattern*, which describes which values are observed or missing in the data matrix; and the missing-value *mechanism*, or mechanisms, which concerns the relationship between missingness and the values of variables in the data matrix.

To illustrate these mechanisms, we define the complete data $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ with y_i the ith response variable and $\boldsymbol{x}_i$ the ith random vector of p covariates. The missing-data indicator matrix $\boldsymbol{R} = (r_{ij})$ is such that $r_{ij} = 1$ if covariate x_{ij} is missing and $r_{ij} = 0$ if covariate x_{ij} is observed. Thus, the matrix $\boldsymbol{R}$ defines the pattern of missing data. We further assume that the response y_i is fully observed, however missing responses require only minor modifications to the results presented in this chapter.

The missing data mechanism is defined by the conditional distribution of $\boldsymbol{R}$ given $(\boldsymbol{y}, \boldsymbol{X})$, i.e. $f(\boldsymbol{R}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ denotes unknown parameters. The data are defined as missing completely at random (MCAR) if missingness does not depend on the values of the data $(\boldsymbol{y}, \boldsymbol{X})$, missing or observed, that is

$$f(\boldsymbol{R}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\phi}) = f(\boldsymbol{R}|\boldsymbol{\phi}) \text{ for all } (\boldsymbol{y}, \boldsymbol{X}), \boldsymbol{\phi}. \tag{4.1}$$

Otherwise, a less restrictive assumption than MCAR occurs when the missingness depends only on the components $\boldsymbol{X}_{obs}$ of $\boldsymbol{X}$ that are observed, and not on the missing components. In this case, the mechanism is denoted missing at random (MAR), and then

$$f(\boldsymbol{R}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\phi}) = f(\boldsymbol{R}|\boldsymbol{y}, \boldsymbol{X}_{obs}, \boldsymbol{\phi}) \text{ for all } \boldsymbol{y}, \boldsymbol{X}_{mis}, \boldsymbol{\phi}. \tag{4.2}$$

The mechanism is called not missing at random (NMAR) if the distribution of

$\boldsymbol{R}$ depends on the missing values in the data matrix $\boldsymbol{X}$.

The literature on incomplete data techniques also describes the situation when the missing-data mechanism is ignorable. To illustrate, consider the covariate vector $\boldsymbol{x}_i = (\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i})$, where $\boldsymbol{x}_{obs,i}$ denotes the observed values and $\boldsymbol{x}_{mis,i}$ denotes the missing values, and let $f(y_i, \boldsymbol{x}_i|\boldsymbol{\theta}) \equiv f(y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\theta})$ be the density of the joint distribution. The marginal probability density may then be found by integrating out the missing-data $\boldsymbol{x}_{mis,i}$ as follows:

$$f(y_i, \boldsymbol{x}_{obs,i}|\boldsymbol{\theta}) = \int f(y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\theta}) d\boldsymbol{x}_{mis,i}. \tag{4.3}$$

Then, the likelihood of $\boldsymbol{\theta}$ based on the observed data $(y_i, \boldsymbol{x}_{obs,i})$ *ignoring the missing-data mechanism* is found to be any function of $\boldsymbol{\theta}$ proportional to the above density, $f(y_i, \boldsymbol{x}_{obs,i}|\boldsymbol{\theta})$.

Moreover, we may include the missing-data matrix $\boldsymbol{R}$ in order to represent a general model which indicates which component of $\boldsymbol{X}$ is observed or missing. This full model treats $\boldsymbol{R}$ as a random variable and specifies the joint distribution of $\boldsymbol{R}$ and $(\boldsymbol{y}, \boldsymbol{X})$, as follows:

$$f(y_i, \boldsymbol{x}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(y_i, \boldsymbol{x}_i|\boldsymbol{\theta}) f(\boldsymbol{r}_i|y_i, \boldsymbol{x}_i, \boldsymbol{\psi}), \ (\boldsymbol{\theta}, \boldsymbol{\psi}) \in \Omega_{\boldsymbol{\theta},\boldsymbol{\psi}}, \tag{4.4}$$

where $\boldsymbol{\psi}$ is an unknown parameter of the conditional (missing-data mechanism) distribution of $\boldsymbol{r}_i$ given $(y_i, \boldsymbol{x}_i)$, and $\Omega_{\boldsymbol{\theta},\boldsymbol{\psi}}$ is the parameter space of $(\boldsymbol{\theta}, \boldsymbol{\psi})$. The

following definition may now be noted:

Definition 4.1. The missing-data mechanism is *ignorable* for likelihood inference if

(a) MAR: The missing data are missing at random and,

(b) Distinctness: The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct, in the sense that the joint parameter space of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ is the product of the parameter space of $\boldsymbol{\theta}$ and the parameter space of $\boldsymbol{\psi}$.

When Definition 4.1 is not satisfied, the missing-data mechanism is said to be *nonignorable*, and maximum likelihood estimation requires a model for the missing-data mechanism and maximization of the full-likelihood, based on equation (4.4). For either case, the Expectation-Maximization (EM) algorithm may be used to obtain maximum likelihood estimates of the parameters. We will discuss ignorable and nonignorable estimation in the following sections.

4.3 Expectation–Maximization Algorithm

4.3.1 The EM Algorithm - Ignorable Mechanism

Conditional on $\boldsymbol{x}_i$, assume y_i follows a distribution in the exponential family:

$$f(y_i|\boldsymbol{x}_i, \theta_i, \tau) = \exp[\{y_i\theta_i - b(\theta_i)\}/a(\tau) + c(y_i, \tau)] \tag{4.5}$$

for some functions a, b, and c, with dispersion parameter τ. The mean response function is modeled by $g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ where $\mu_i = E(y_i) = b'(\theta_i)$, g is a link function, and $\boldsymbol{x}_i$ represents the ith row of the design matrix $\boldsymbol{X}$. We assume $\boldsymbol{X}$ is an $n \times p$ full rank matrix, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of regression coefficients. While (4.5) represents a large class of models, we note that for many popular cases, such as binary and Poisson regression, the dispersion parameter τ is fixed at unity. Hence, without loss of generality, we assume $\tau = 1$.

The complete-data log-likelihood for (4.5) based on all observations is

$$l(\boldsymbol{\beta}, \tau|\boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n}[\{y_i\theta_i - b(\theta_i)\}/a(\tau) + c(y_i, \tau)]. \tag{4.6}$$

Since we assume $\tau = 1$, then $f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \tau) \equiv f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})$.

Now, assume that some covariates in the vector $\boldsymbol{x}_i$ are missing at random (MAR). That is, we assume the missingness depends only on the components $\boldsymbol{x}_{obs,i}$ of $\boldsymbol{x}_i$ which are observed, and not on the missing components. Then, the

observed data likelihood is obtained by integrating or summing (4.6) over the missing values, with respect to its distribution. Here, the random vector $\boldsymbol{x}_i$ is assumed to follow density $f(\boldsymbol{x}_i|\boldsymbol{\alpha})$. If we let $\boldsymbol{\gamma}^T = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)$, the complete-data log-likelihood may be decomposed as

$$l(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_i) = \sum_{i=1}^{n} \{\log[f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})] + \log[f(\boldsymbol{x}_i|\boldsymbol{\alpha})]\}. \tag{4.7}$$

Suppose the covariates can be rearranged in the form $\boldsymbol{x}_i = (\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i},)$. Then, we may write $f(y_i, \boldsymbol{x}_i|\boldsymbol{\gamma}) \equiv f(y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\gamma})$. For continuous covariates, the marginal probability density is found by integrating out the missing-data $\boldsymbol{x}_{mis,i}$ as follows:

$$\begin{aligned} f(y_i, \boldsymbol{x}_{obs,i}|\boldsymbol{\gamma}) &= \int f(y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\gamma}) d\boldsymbol{x}_{mis,i} \\ &= \int f(y_i|\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \boldsymbol{\beta}) f(\boldsymbol{x}_i|\boldsymbol{\alpha}) d\boldsymbol{x}_{mis,i}. \end{aligned} \tag{4.8}$$

Since $\boldsymbol{x}_i = \boldsymbol{x}_{obs,i}$ if there are no missing covariates in $\boldsymbol{x}_i$, we will implicitly use $f(y_i, \boldsymbol{x}_{obs,i}|\boldsymbol{\gamma})$ to indicate both joint and marginal densities.

Note that for categorical missing covariates, we sum over all possible values of $\boldsymbol{x}_{mis,i}$. Then, the log-likelihood of $\boldsymbol{\gamma}$ based on observed data may be expressed as

$$l(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{x}_{obs}) = \sum_{i=1}^{n} l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_{obs,i}) = \sum_{i=1}^{n} \ln f(y_i, \boldsymbol{x}_{obs,i}|\boldsymbol{\gamma}). \tag{4.9}$$

However, the observed log-likelihood in (4.9) is not in closed-form, and cannot be maximized explicitly. Variants of the Newton-Raphson algorithm may be applied, however these methods require second derivatives that are approximated only at a high computational cost. An alternative method for incomplete-data problems in the literature which avoids such calculations is called the Expectation-Maximization (EM) algorithm.

Each iteration of the EM algorithm consists of an E (expectation) step and an M (maximization) step. The E step determines the conditional expectation of the complete-data likelihood given the observed data and current estimated parameters. Then, in the M step, the conditional expected likelihood is maximized in the same way as the usual maximum likelihood estimation procedures.

The log-likelihood (4.7) may be further decomposed as

$$l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_i) = l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}) = l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_{obs,i}) + \ln f(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{\gamma}).$$

Then, the observed log-likelihood is written as

$$l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_{obs,i}) = l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_i) - \ln f(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{\gamma}). \tag{4.10}$$

Let $\boldsymbol{\gamma}^{(t)}$ be the current estimate of the parameter $\boldsymbol{\gamma}$. The E step of EM determines the expected complete-data log-likelihood from the conditional density

of $\boldsymbol{x}_{mis}$ given $(\boldsymbol{y}, \boldsymbol{x}_{obs})$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t)}$:

$$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) = \sum_{i=1}^{n} E\left[l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_i)|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t)}\right]. \tag{4.11}$$

For continuous covariates, the E-step evaluates

$$\begin{aligned} Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) &= \sum_{i\ \boldsymbol{x}_i = \boldsymbol{x}_{obs,i}} l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_i) \\ &\quad + \sum_{i\ \boldsymbol{x}_i \neq \boldsymbol{x}_{obs,i}} \int l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_i) f(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t)}) d\boldsymbol{x}_{mis,i} \end{aligned} \tag{4.12}$$

while for categorical covariates, $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ is obtained by similarly summing over the support of $\boldsymbol{x}_{mis,i}$.

The M step of EM determines $\boldsymbol{\gamma}^{(t+1)}$ by maximizing this expected complete-data log-likelihood such that

$$Q(\boldsymbol{\gamma}^{(t+1)}|\boldsymbol{\gamma}^{(t)}) \geq Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}), \text{ for all } \boldsymbol{\gamma}. \tag{4.13}$$

If we define a sequence of iterates $\boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\gamma}^{(1)}$, ..., where $\boldsymbol{\gamma}^{(t+1)} = M(\boldsymbol{\gamma}^{(t)})$ for some function M, then from (4.10), the difference in values of the log-likelihood $l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_{obs,i})$ at successive iterates is given by

$$\begin{aligned} \sum_{i=1}^{n} \left[l(\boldsymbol{\gamma}^{(t+1)}|y_i, \boldsymbol{x}_{obs,i}) - l(\boldsymbol{\gamma}^{(t)}|y_i, \boldsymbol{x}_{obs,i})\right] &= [Q(\boldsymbol{\gamma}^{(t+1)}|\boldsymbol{\gamma}^{(t)}) - Q(\boldsymbol{\gamma}^{(t)}|\boldsymbol{\gamma}^{(t)})] \\ &\quad - [H(\boldsymbol{\gamma}^{(t+1)}|\boldsymbol{\gamma}^{(t)}) - H(\boldsymbol{\gamma}^{(t)}|\boldsymbol{\gamma}^{(t)})]. \end{aligned} \tag{4.14}$$

where

$$H(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) = \sum_{i=1}^{n} \int [\ln f(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{\gamma})] f(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{\gamma}^{(t)}) d\boldsymbol{x}_{mis,i}. \tag{4.15}$$

An EM algorithm chooses $\boldsymbol{\gamma}^{(t+1)}$ to maximize $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ with respect to $\boldsymbol{\gamma}$. Further, a *generalized EM algorithm (GEM)* chooses $\boldsymbol{\gamma}^{(t+1)}$ so that $Q(\boldsymbol{\gamma}^{(t+1)}|\boldsymbol{\gamma}^{(t)})$ is greater than $Q(\boldsymbol{\gamma}^{(t)}|\boldsymbol{\gamma}^{(t)})$. Since the difference in the H functions is negative, then any change from $\boldsymbol{\gamma}^{(t)}$ to $\boldsymbol{\gamma}^{(t+1)}$ increases the likelihood for any EM or GEM algorithm.

In order to maximize $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$, we must solve

$$\sum_{i=1}^{n} E\left[\frac{\partial \log f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|y_i,\boldsymbol{x}_{obs,i},\boldsymbol{\gamma}^{(t)}\right] = \mathbf{0}, \tag{4.16}$$

$$\sum_{i=1}^{n} E\left[\frac{\partial \log f(\boldsymbol{x}_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}|y_i,\boldsymbol{x}_{obs,i},\boldsymbol{\gamma}^{(t)}\right] = \mathbf{0}, \tag{4.17}$$

where the conditional expectations are taken with respect to $\boldsymbol{x}_{mis,i}$, given the actual observed data $(y_i,\boldsymbol{x}_{obs,i})$ and current estimate $\boldsymbol{\gamma}^{(t)}$. To solve equation (4.16), we may use a Newton-Raphson or scoring algorithm, similar to the approach often used in the generalized linear model for complete data as in Section 2.1. Furthermore, for the exponential family (4.5), the maximum likelihood estimating equation for $\boldsymbol{\beta}$ takes the form

$$\sum_{i=1}^{n} E\left[\{y_i-\mu_i(\boldsymbol{\beta},\boldsymbol{x}_i)\}\,\boldsymbol{x}_i\,|y_i,\boldsymbol{x}_{obs,i}\right]=0 \tag{4.18}$$

with $\boldsymbol{x}_i=(\boldsymbol{x}_{obs,i},\boldsymbol{x}_{mis,i})$.

In addition, the asymptotic variance of the unconstrained maximum likeli-

hood estimate of $\boldsymbol{\gamma}$ may be obtained from the observed Fisher information matrix (see Louis (1982) for details):

$$\begin{aligned}\mathcal{I}_o(\boldsymbol{\gamma}) &= -\sum_{i=1}^{n} E\left[\partial U_i(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^T | y_i, \boldsymbol{x}_{obs,i}\right] \\ &\quad -\sum_{i=1}^{n} E\left[U_i(\boldsymbol{\gamma})U_i(\boldsymbol{\gamma})^T | y_i, \boldsymbol{x}_{obs,i}\right] \\ &\quad +\sum_{i=1}^{n} E\left[U_i(\boldsymbol{\gamma})|y_i, \boldsymbol{x}_{obs,i}\right] E\left[U_i(\boldsymbol{\gamma})|y_i, \boldsymbol{x}_{obs,i}\right]^T \end{aligned} \tag{4.19}$$

where $U_i(\boldsymbol{\gamma}) = \partial l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_i)/\partial\boldsymbol{\gamma}$. The observed information may be decomposed into the form

$$\mathcal{I}_o(\boldsymbol{\gamma}) = \begin{bmatrix} \mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\alpha}) \\ \mathcal{I}_o(\boldsymbol{\alpha},\boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\alpha},\boldsymbol{\alpha}) \end{bmatrix}, \tag{4.20}$$

which will be utilized in the constrained estimation techniques.

4.3.2 The EM Algorithm - Nonignorable Mechanism

In many practical situations, the missing data mechanism depends on the values of $\boldsymbol{x}_{mis,i}$, and hence is nonignorable. To incorporate the missing data mechanism into the likelihood, we assume that given the data $(y_i, \boldsymbol{x}_i)$, the conditional distribution of $\boldsymbol{r}_i$ is a multinomial distribution $f(\boldsymbol{r}_i|y_i, \boldsymbol{x}_i, \boldsymbol{\psi})$ with 2^p cell probabilities depending on some parameters $\boldsymbol{\psi}$.

Consequently, the actual observed data consists of the values of $(y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i)$.

The distribution of the observed data is obtained by integrating $\boldsymbol{x}_{mis,i}$ out of the joint density of $(y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)$, as follows:

$$\begin{aligned} f(y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i|\boldsymbol{\gamma}) &= \int f(y_i|\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \boldsymbol{\beta}) f(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}|\boldsymbol{\alpha}) \\ &\quad \times f(\boldsymbol{r}_i|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \boldsymbol{\psi}) d\boldsymbol{x}_{mis,i}. \end{aligned} \tag{4.21}$$

The full likelihood of $\boldsymbol{\gamma} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\psi}^{\mathrm{T}})^{\mathrm{T}}$ is the product of (4.21) for all n observations. The complete-data log-likelihood of $\boldsymbol{\gamma}$ for given $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{r})$ may be expressed in the form

$$\begin{aligned} l(\boldsymbol{\gamma}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{r}) &= \sum_{i=1}^{n} l(\boldsymbol{\gamma}|\boldsymbol{x}_i, y_i, \boldsymbol{r}_i) \\ &\equiv \sum_{i=1}^{n} \{\log[f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})] + \log[f(\boldsymbol{x}_i|\boldsymbol{\alpha})] \\ &\quad + \log[f(\boldsymbol{r}_i|\boldsymbol{x}_i, y_i, \boldsymbol{\psi})]\}. \end{aligned} \tag{4.22}$$

Analogous to the ignorable setting, for continuous covariates, the E-step under the nonignorable model is written as

$$\begin{aligned} Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}) &= \sum_{\{i \;\; \boldsymbol{x}_i = \boldsymbol{x}_{obs,i}\}} l(\boldsymbol{\gamma}|y_i, \boldsymbol{x}_i, \boldsymbol{r}_i) \\ &\quad + \sum_{\{i \;\; \boldsymbol{x}_i \neq \boldsymbol{x}_{obs,i}\}} \Bigg[\int \log[f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})] f(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t)}) d\boldsymbol{x}_{mis,i} \\ &\quad + \int \log[f(\boldsymbol{x}_i|\boldsymbol{\alpha})] f(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t)}) d\boldsymbol{x}_{mis,i}, \\ &\quad + \int \log[f(\boldsymbol{r}_i|y_i, \boldsymbol{x}_i, \boldsymbol{\psi})] f(\boldsymbol{x}_{mis,i}|y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t)}) d\boldsymbol{x}_{mis,i} \Bigg]. \end{aligned} \tag{4.23}$$

For categorical covariates, we find $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$ by similarly taking the summation over the discrete case.

The updated estimate, $\boldsymbol{\gamma}^{(t+1)}$ is obtained by solving the following equations:

$$\sum_{i=1}^{n} E\left[\frac{\partial \log f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|y_i,\boldsymbol{x}_{obs,i},\boldsymbol{r}_i,\boldsymbol{\gamma}^{(t)}\right] = \mathbf{0}, \tag{4.24}$$

$$\sum_{i=1}^{n} E\left[\frac{\partial \log f(\boldsymbol{x}_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}|y_i,\boldsymbol{x}_{obs,i},\boldsymbol{r}_i,\boldsymbol{\gamma}^{(t)}\right] = \mathbf{0}, \tag{4.25}$$

$$\sum_{i=1}^{n} E\left[\frac{\partial \log f(\boldsymbol{r}_i|y_i,\boldsymbol{x}_i,\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}|y_i,\boldsymbol{x}_{obs,i},\boldsymbol{r}_i,\boldsymbol{\gamma}^{(t)}\right] = \mathbf{0}, \tag{4.26}$$

where the conditional expectations are taken with respect to $\boldsymbol{x}_{mis,i}$, given the observed data $(\boldsymbol{\beta}^{(t)},\boldsymbol{\alpha}^{(t)},\boldsymbol{\psi}^{(t)})$ and $\boldsymbol{\gamma}^{(t)}$.

We next define the quantity

$$\begin{aligned}\mathcal{I}_{oi}(\boldsymbol{\gamma}) = & -E\left[\partial U_i(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^T|y_i,\boldsymbol{x}_{obs,i},\boldsymbol{r}_i\right] \\ & -E\left[U_i(\boldsymbol{\gamma})U_i(\boldsymbol{\gamma})^T|y_i,\boldsymbol{x}_{obs,i},\boldsymbol{r}_i\right] \\ & +E\left[U_i(\boldsymbol{\gamma})|y_i,\boldsymbol{x}_{obs,i},\boldsymbol{r}_i\right]E\left[U_i(\boldsymbol{\gamma})|y_i,\boldsymbol{x}_{obs,i},\boldsymbol{r}_i\right]^T\end{aligned} \tag{4.27}$$

where $U_i(\boldsymbol{\gamma}) = \partial l(\boldsymbol{\gamma}|\boldsymbol{x}_i,y_i,\boldsymbol{r}_i)/\partial \boldsymbol{\gamma}$. As the primary focus is the estimation of $\boldsymbol{\beta}$ with $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ viewed as nuisance parameters, we may also derive the estimated asymptotic variance of $\boldsymbol{\beta}$, obtained from the observed Fisher information, $\mathcal{I}_o$, where $\mathcal{I}_o = \sum_{i=1}^{n}\mathcal{I}_{oi}$ from (4.27).

Note that a suitable model for the missing data mechanism is required in the nonignorable setting. As discussed in Ibrahim et. al. (2005) and Sinha (2008), one possibility is a joint log-linear model for $f(\boldsymbol{r}_i|y_i, \boldsymbol{x}_i, \boldsymbol{\psi})$. The missing-data mechanism may also be modeled by a sequence of one-dimensional conditional distributions:

$$\begin{aligned} f(r_{i1}, \ldots, r_{ip}|y_i, \boldsymbol{x}_i, \boldsymbol{\psi}) &= f(r_{ip}|r_{i1}, \ldots, r_{i,p-1}, y_i, \boldsymbol{x}_i, \boldsymbol{\psi}_p) \\ &\quad \times f(r_{i,p-1}|r_{i1}, \ldots, r_{i,p-2}, y_i, \boldsymbol{x}_i, \boldsymbol{\psi}_{p-1}) \\ &\quad \ldots \\ &\quad \times f(r_{i2}|r_{i1}, y_i, \boldsymbol{x}_i, \boldsymbol{\psi}_2) f(r_{i1}|y_i, \boldsymbol{x}_i, \boldsymbol{\psi}_1), \quad (4.28) \end{aligned}$$

where $\boldsymbol{\psi}_k$ is a vector of indexing parameters for the kth conditional distribution, and $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p)$.

As noted by many authors (Baker and Laird, 1988; Ibrahim et. al., 2005; Sinha, 2008), the modeling strategy of (4.28) permits flexibility in the missing-data model specification and facilitates an intuitive method of determining the joint distribution of the missing-data indicators when knowledge concerning the missingness of one variable influences the probability of missingness of others. Random sampling from the conditional distributions of missing covariates given the observed data is also permissible, which is subsequently applied in the ap-

proximation of the conditional expectations for the estimating equations. As each of the univariate distributions in (4.28) is a logistic regression, each objective function is log-concave in the parameters. Consequently, the computation of the maximum likelihood estimates is improved.

However, as indicated by Baker and Laird (1988) and Ibrahim, Lipsitz and Chen (1999), it is possible to build a missing-data model which is too large, and becomes unidentifiable due to overparameterization. To compare various candidate models empirically, the likelihood ratio or Akaike information criterion may be used. As stated by Ibrahim et. al. (2005), the main-effects model will typically be an adequate approximation to the missing-data mechanism, and interaction or higher-order terms should be interpreted with caution. On the other hand, in some applications, the data may not adequately distinguish between missing data models, and a sensitivity analyses may be utilized to assess the appropriate nonignorable model. Research is ongoing to determine the best method to incorporate nonignorability in missing data methods.

4.3.3 Convergence of the EM Algorithm

The important paper written by Wu (1983) describes conditions which ensure the convergence of a sequence of likelihood values $\{l(\boldsymbol{\gamma}^{(t)}|\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r})\}$ to a stationary

value of $l(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r})$. However, these methods assume that the sequence of values generated by the EM algorithm lie in the interior of the parameter space. A more recent paper by Nettleton (1999) provided convergence results for the EM algorithm when this assumption may be relaxed to include cases when the maximizing parameter value is on the boundary of the parameter space, as is common in constrained statistical inference. Both theorems are now presented.

Theorem 4.1. (Wu, 1983) Suppose that the incomplete data likelihood function $L(\boldsymbol{\gamma})$ is unimodal in its parameter space Ω with $\boldsymbol{\gamma}^*$ being the only stationary point, and that $\partial Q(\boldsymbol{\gamma}, \boldsymbol{\psi})/\partial \boldsymbol{\psi}$ is continuous in $\boldsymbol{\gamma}$ and $\boldsymbol{\psi}$. Then, any EM sequence $\{\boldsymbol{\gamma}^{(k)}\}$ converges to the unique maximizer $\boldsymbol{\gamma}^*$ of $L(\boldsymbol{\gamma})$. That is, it converges to the unique MLE of $\boldsymbol{\gamma}$.

Theorem 4.2. (Nettleton, 1999) Suppose that $\boldsymbol{\gamma}^*$ is the unique maximizer of $L(\boldsymbol{\gamma})$ over the constrained parameter space $\Theta \subset \Omega$ and that $\boldsymbol{\gamma}^*$ is the only element of $S = \{\boldsymbol{\gamma}^* \in \Theta : \frac{d}{d\lambda} L\{(1-\lambda)\boldsymbol{\gamma}^* + \lambda\boldsymbol{\gamma}\}|_{\lambda=0} \leq 0 \text{ for all } \boldsymbol{\gamma} \in C\}$, where C is the union of all convex subsets of Θ containing $\boldsymbol{\gamma}^*$. If $G(\boldsymbol{\psi}|\boldsymbol{\gamma}) = \sup\{\frac{d}{d\lambda} Q\{(1-\lambda)\boldsymbol{\psi} + \lambda\boldsymbol{\gamma}^*|\boldsymbol{\gamma}\}|_{\lambda=0}; \boldsymbol{\gamma}^* \in C\}$ is continuous in $\boldsymbol{\psi}$ and $\boldsymbol{\gamma}$ then for any EM sequence $\{\boldsymbol{\gamma}_p\}$, $\boldsymbol{\gamma}_p$ converges to $\boldsymbol{\gamma}^*$.

4.4 ML Estimation under Linear Inequalities with Incomplete Data

As noted by Little and Rubin (2002, p. 171), interest often lies in computing ML estimates for models with incomplete-data that place constraints on the parameters. However, when the EM algorithm is applied to fit such models, the constraints do not affect the E step, which is the missing-data part of the problem. The M step sequentially maximizes the expected complete-data likelihood function given the observed data and current parameter estimates, *subject to the parameter constraints.*

With no missing data, the problem of constrained inference for generalized linear models is well-developed and outlined in Silvapulle and Sen (2005). Dunson and Neelon (2003) discussed Bayesian inference for generalized linear models under the simple order constraint, $\{\boldsymbol{\beta} : \beta_1 \leq \beta_2 \leq \cdots \leq \beta_p\}$. The authors focused on deriving a posterior distribution for the means under the order restriction, and performing inferences.

As noted in the Introduction, many authors have considered constrained estimation methods for linear models in the presence of incomplete data. In particular, Kim and Taylor (1995) developed a restricted EM algorithm for the general

hypothesis of $A\boldsymbol{\beta} \leq \mathbf{c}$ for variance component and bivariate normal models with missing values of the response vector, $\boldsymbol{y}$. The method was later improved by Jamshidian (2004) to the general likelihood and constraint function with missing data, as discussed in Chapter 3. Shi, Zheng and Guo (2005) provided EM algorithms for estimating the mean vector in a multivariate normal model with known and unknown covariance matrix and missing response data under the restriction $A\boldsymbol{\beta} \leq 0$. A subsequent paper (Zheng, Shi and Guo, 2005) extended to the case when $A\boldsymbol{\beta} \leq \mathbf{c}$. Tian, Ng and Tan (2008) developed an EM algorithm to estimate mean and variance-covariance parameters under constraints for the multivariate t-distribution with missing data. In a slightly different problem, Nettleton and Praestgaard (1998) considered interval mapping procedures for genetics studies using order-restricted inference in combination with an EM algorithm. Constrained likelihood ratio tests were also derived, and shown to follow a chi-bar-square distribution.

In this section, a modification of the GP algorithm is proposed to obtain constrained parameter estimates in the GLM with missing covariate data.

4.4.1 GP-EM Algorithm for GLM with Incomplete Data

In the ignorable case, let the constrained space of parameters be denoted $C_{1I} = \{(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T : A\boldsymbol{\beta} \leq \mathbf{c}\}$, while in the nonignorable case we have $C_{1N} = \{(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\psi}^T)^T : A\boldsymbol{\beta} \leq \mathbf{c}\}$, where A is a $r \times p$ matrix of full row rank $(r \leq p)$. Recall that the log-likelihood based on the observed data is given by

$$l(\boldsymbol{\gamma}^*) = \begin{cases} \sum_{i=1}^{n} \ln f(y_i, \boldsymbol{x}_{obs,i}|\boldsymbol{\gamma}) & \text{if missingness is ignorable} \\ \sum_{i=1}^{n} \ln f(y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i|\boldsymbol{\gamma}) & \text{if missingness is nonignorable,} \end{cases} \tag{4.29}$$

where the densities are those in (4.7) and (4.22) respectively. To implement the EM algorithm for estimation over C_{1I} or C_{1N}, we must calculate the appropriate form of $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)})$, and then utilize a constrained maximization procedure.

Let $\boldsymbol{\gamma}^{(k)}$ denote the result of the kth EM iteration, with $\boldsymbol{\gamma}^{(0)}$ representing an initial value satisfying the constraints. For the constrained problem, $\boldsymbol{\gamma}^{(k+1)}$ is determined from $\boldsymbol{\gamma}^{(k)}$ in the following manner:

E-step: Determine $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(k)})$ using (4.12) for the ignorable mechanism and (4.23) for the nonignorable model.

M-step: Compute $\boldsymbol{\gamma}^{(k+1)} = \arg\max_{\boldsymbol{\gamma} \in C_1} Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(k)})$.

Note that the unconstrained estimate is obtained as a special case of the M-step by replacing C_{1I} or C_{1N} by $\mathbb{R}^p$.

For generalized linear models with missing data, the M step indicates that we maximize $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(k)})$ under linear inequality constraints on the regression parameters. To apply the gradient projection algorithm, we require an estimate of the observed information matrix and score vector. From the EM algorithm, we define the score vector to be

$$s^{EM}(\boldsymbol{\gamma}) = \partial Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(k)})/\partial\boldsymbol{\gamma}. \tag{4.30}$$

In the ignorable setting, since the parameter vector is composed of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we partition the inverse of the observed information matrix as follows:

$$\mathcal{I}_o^{-1}(\boldsymbol{\gamma}) \equiv W(\boldsymbol{\gamma}) = \begin{bmatrix} W_{11}(\boldsymbol{\gamma}) & W_{12}(\boldsymbol{\gamma}) \\ W_{21}(\boldsymbol{\gamma}) & W_{22}(\boldsymbol{\gamma}) \end{bmatrix} \tag{4.31}$$

where $W_{11}(\boldsymbol{\gamma}) = [\mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\beta}) - \mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\alpha})\mathcal{I}_o^{-1}(\boldsymbol{\alpha},\boldsymbol{\alpha})\mathcal{I}_o(\boldsymbol{\alpha},\boldsymbol{\beta})]^{-1}$,

$W_{12}(\boldsymbol{\gamma}) = -\mathcal{I}_o^{-1}(\boldsymbol{\beta},\boldsymbol{\beta})\mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\alpha})[\mathcal{I}_o(\boldsymbol{\alpha},\boldsymbol{\alpha}) - \mathcal{I}_o(\boldsymbol{\alpha},\boldsymbol{\beta})\mathcal{I}_o^{-1}(\boldsymbol{\beta},\boldsymbol{\beta})\mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\alpha})]^{-1}$, $W_{22}(\boldsymbol{\gamma}) =$

$[\mathcal{I}_o(\boldsymbol{\alpha},\boldsymbol{\alpha}) - \mathcal{I}_o(\boldsymbol{\alpha},\boldsymbol{\beta})\mathcal{I}_o^{-1}(\boldsymbol{\beta},\boldsymbol{\beta})\mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\alpha})]^{-1}$, and $W_{21}(\boldsymbol{\gamma}) = W_{12}^T(\gamma)$. Then, the generalized score vector can be expressed as $s^*(\boldsymbol{\gamma}) = \mathcal{I}_o^{-1}(\boldsymbol{\gamma})s_{\boldsymbol{\gamma}}^{EM}(\boldsymbol{\gamma}) =$

$(s_1^{EM*}(\boldsymbol{\gamma})^T, s_2^{EM*}(\boldsymbol{\gamma})^T)^T$ where $s_1^{EM*}(\boldsymbol{\gamma}) = W_{11}(\boldsymbol{\gamma})s_{\boldsymbol{\beta}}^{EM}(\boldsymbol{\gamma}) + W_{12}(\boldsymbol{\gamma})s_{\boldsymbol{\alpha}}^{EM}(\boldsymbol{\gamma})$ and

$s_2^{EM*}(\boldsymbol{\gamma}) = W_{21}(\boldsymbol{\gamma})s_{\boldsymbol{\beta}}^{EM}(\boldsymbol{\gamma}) + W_{22}(\boldsymbol{\gamma})s_{\boldsymbol{\alpha}}^{EM}(\boldsymbol{\gamma})$.

If the unconstrained estimate satisfies the constraints so that $\hat{\boldsymbol{\gamma}} \in C_1$, then the constrained estimate is identical to $\hat{\boldsymbol{\gamma}}$. Otherwise, we proceed with the fol-

lowing gradient projection expectation-maximization (GP-EM) algorithm with an initial value, $\boldsymbol{\gamma}_r$, chosen from C_1. Based on the constraints which hold with equality at $\boldsymbol{\gamma}_r$, we form the active constraint set $\mathcal{W}$, and the corresponding coefficient matrix $\tilde{A}$ and vector $\tilde{\mathbf{c}}$.

Step 1: Using the current value $\boldsymbol{\gamma}_r$ of $\boldsymbol{\gamma}$, evaluate $W(\boldsymbol{\gamma}_r)$.

Step 2: Compute $B(\boldsymbol{\gamma}_r) = \tilde{A}^T[\tilde{A}W_{11}(\boldsymbol{\gamma}_r)\tilde{A}^T]^{-1}\tilde{A}$, and $\boldsymbol{d} = (\boldsymbol{d}_1^T, \boldsymbol{d}_2^T)^T$ where $\boldsymbol{d}_1 = [I - W_{11}(\boldsymbol{\gamma}_r)B(\boldsymbol{\gamma}_r)]s_1^{EM*}(\boldsymbol{\gamma}_r)$ and $\boldsymbol{d}_2 = -W_{21}(\boldsymbol{\gamma}_r)B(\boldsymbol{\gamma}_r)s_1^{EM*}(\boldsymbol{\gamma}_r) + s_2^{EM*}(\boldsymbol{\gamma}_r)$.

Step 3: If $\boldsymbol{d} = \mathbf{0}$, compute the Lagrange multiplier vector $\boldsymbol{\lambda} = [\tilde{A}W_{11}(\boldsymbol{\gamma}_r)\tilde{A}^T]^{-1}\tilde{A}s_1^{EM*}(\boldsymbol{\gamma}_r)$.

a) If $\lambda_i \geq 0$ for all $i \in \mathcal{W}$, the current point is the constrained estimate. Stop.

b) If $\lambda_i < 0$ for some $i \in \mathcal{W}$, drop the index corresponding to the smallest λ_i from $\mathcal{W}$. Form new $\tilde{A}$ and $\tilde{\mathbf{c}}$ based on the constraints remaining in $\mathcal{W}$. Go to Step 2.

Step 4: If $\boldsymbol{d} \neq \mathbf{0}$, find $\delta_1 = \arg\max_\delta\{\delta : \boldsymbol{\gamma}_r + \delta\boldsymbol{d} \in C_1\}$ and then determine δ_2 such that $Q(\boldsymbol{\gamma}_r + \delta_2\boldsymbol{d} \mid \boldsymbol{\gamma}^{(k)}) \geq Q(\boldsymbol{\gamma}_r + \delta\boldsymbol{d} \mid \boldsymbol{\gamma}^{(k)})$ for $0 \leq \delta \leq \delta_1$. Let

$\boldsymbol{\gamma}_r^* = \boldsymbol{\gamma}_r + \delta_2 \boldsymbol{d}$ and determine which constraints newly hold with equality at $\boldsymbol{\gamma}_r^*$. Add their indexes, if any, to $\mathcal{W}$ and update $\tilde{A}$ and $\tilde{\mathbf{c}}$ accordingly.

Step 5: Replace $\boldsymbol{\gamma}_r$ by $\boldsymbol{\gamma}_r^*$ and go to Step 1.

At each stage of the above algorithm, the calculation of the observed information matrix, score vector or $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(k)})$ requires evaluation of conditional expectations of certain functions of $\boldsymbol{x}_{mis,i}$ given $\{y_i, \boldsymbol{x}_{obs,i}\}$ with intermediate parameter value $\boldsymbol{\gamma}_r$. We rely on approximation of these expectations by numerical methods, similar to those outlined in Chapter 2.

In the nonignorable setting, we have the additional vector $\boldsymbol{\psi}$, corresponding to the missing data mechanism. Since there are no constraints placed on $\boldsymbol{\psi}$ as part of C_{1N}, we may implement the preceding GP-EM algorithm by replacing $\boldsymbol{\alpha}$ by $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}^T, \boldsymbol{\psi}^T)^T$ and proceed accordingly.

Additional analysis of constrained GLMs with incomplete data may be found in Section 6.2 with respect to the Northern Contaminants Programme analysis.

4.5 Likelihood Ratio Tests for Constrained GLM with Incomplete Data

We extend hypothesis testing for generalized linear models with missing data to the case of linear inequality constraints on the regression parameters. We begin with the constrained set $\Omega = \{\boldsymbol{\gamma} : A\boldsymbol{\beta} \leq \mathbf{c}\}$, with A an $r \times p$ matrix of full rank, and $\mathbf{c}$ an $r \times 1$ vector defined previously. The constrained tests are associated with the hypotheses:

$$H_0 : A\boldsymbol{\beta} = \mathbf{c}, \quad H_1 : A\boldsymbol{\beta} \leq \mathbf{c}, \quad H_2 : \text{ no restriction on } \boldsymbol{\beta}. \tag{4.32}$$

We utilize the constrained maximum likelihood estimators obtained from the GP-EM algorithm, as described in Section 4.4. Based on the maximum likelihood estimators, $\boldsymbol{\gamma}^0$ for H_0, $\boldsymbol{\gamma}^*$ for H_1 and $\hat{\boldsymbol{\gamma}}$ under H_2, we may construct the likelihood ratio tests for the three sets of hypotheses in (4.32), using the log-likelihood function $l^*(\boldsymbol{\gamma})$ given in (4.29) . If $T_{02} = 2[l^*(\hat{\boldsymbol{\gamma}}) - l^*(\boldsymbol{\gamma}^0)]$ is large, then the unconstrained test rejects H_0 in favor of $H_2 - H_0$. In the GLM case with incomplete data, T_{02} asymptotically follows $\chi^2(r)$ under H_0. When the parameter space is restricted by H_1, we test H_0 against $H_1 - H_0$ with the statistic $T_{01} = 2[l^*(\boldsymbol{\gamma}^*) - l^*(\boldsymbol{\gamma}^0)]$. We also confirm the usefulness of the test corresponding

to T_{01} as only when H_1 is true, using the goodness-of-fit test which rejects H_1 for large values of $T_{12} = 2[l^*(\hat{\gamma}) - l^*(\gamma^*)]$.

The asymptotic distributions of T_{01} and T_{12} are demonstrated to be chi-bar-square, as outlined in the following theorem.

4.5.1 Derivation of Asymptotic Results under Inequality Constraints with Incomplete Data

We next present the following main result for constrained likelihood ratio tests in generalized linear models with incomplete covariate data.

Theorem 4.3. Under appropriate regularity assumptions, the asymptotic distributions of the likelihood ratio test statistics T_{01} and T_{12} under H_0, are given as follows:

$$\lim_{n\to\infty} P_{\gamma_0}[T_{01} > x] = \sum_{i=0}^{r} w_i(r, AV(\gamma_0)A^T)P[\chi_i^2 > x], \tag{4.33}$$

$$\lim_{n\to\infty} P_{\gamma_0}[T_{12} > x] = \sum_{i=0}^{r} w_{r-i}(r, AV(\gamma_0)A^T)P[\chi_i^2 > x] \tag{4.34}$$

for any $x \geq 0$. Here, r is the rank of A, $\gamma_0 = (\beta_0^T, \alpha_0^T, \psi_0^T)^T = (\beta_0^T, \alpha_0^{*T})^T$ is a value of γ under H_0, and $V(\gamma_0)$ equals

$[\mathcal{I}_U(\beta_0, \beta_0) - \mathcal{I}_U(\beta_0, \alpha_0^*)\mathcal{I}_U^{-1}(\alpha_0^*, \alpha_0^*)\mathcal{I}_U(\alpha_0^*, \beta_0)]^{-1}$ where $\mathcal{I}_U(\cdot,\cdot) = E[\mathcal{I}_{oi}(\cdot,\cdot)]$,

and $\mathcal{I}_{oi}(\cdot,\cdot)$ was defined by (4.27). Note that the unit sample Fisher information matrix $\mathcal{I}_U(\cdot,\cdot) = E[\mathcal{I}_{oi}(\cdot,\cdot)]$ does not depend on i since $(y_i, \boldsymbol{x}_{obs,i}, \boldsymbol{r}_i)$, $i = 1,\ldots,n$ are independent and identically distributed.

The null distributions in Theorem 4.3 depend on the unknown parameter vector $\boldsymbol{\gamma}_0$, which may be approximated by replacing the parameter vector by its estimate. We may use $\boldsymbol{\gamma}^*$ for T_{01} and $\hat{\boldsymbol{\gamma}}$ for T_{12}, analogous to a Wald statistic.

4.5.2 Calculation of Chi-bar-square weights

The chi-bar-square weights, $w_i(r, \boldsymbol{D})$, represent the probability that the least squares projection of an r-dimensional multivariate normal observation from $N(\boldsymbol{0}, \boldsymbol{D})$ onto the positive orthant cone has exactly i positive component values. Kudo (1963) expressed the weights $w_i(r, \boldsymbol{D})$ as the sum of multivariate normal orthant probabilities. To illustrate, suppose $\boldsymbol{Z} = (Z_1, \cdots, Z_r)$ follows $N(\boldsymbol{0}, \boldsymbol{D})$. For any subset ζ of $\mathcal{K} = \{1, \cdots, r\}$, define $\boldsymbol{Z}(\zeta)$ as the vector of all Z_i, $i \in \zeta$ so that $\boldsymbol{Z}(\zeta)$ follows $N(\boldsymbol{0}, \boldsymbol{D}_\zeta)$. Also let $\boldsymbol{Z}^*(\zeta)$ denote a random vector following $N(\boldsymbol{0}, \boldsymbol{D}_\zeta^{-1})$. Then,

$$w_i(r, \boldsymbol{D}) = \sum_{\zeta \subset \mathcal{K}:\ |\zeta| = i} P[\boldsymbol{Z}^*(\zeta^c) > 0] P[\boldsymbol{Z}(\zeta) > 0 \mid \boldsymbol{Z}(\zeta^c) = 0]; \qquad (4.35)$$

where $\mid \zeta \mid$ denotes the cardinality of ζ and ζ^c is the complement of ζ. If I is the identity matrix, then $w_i(r, I)$ are simply binomial $(r, 1/2)$ probabilities. In general, for a small number of variables, formula (4.35) may be calculated explicitly, leading to a closed form expression of $w_i(r, \boldsymbol{D})$ when $r \leq 4$. For moderate values of r, the orthant probabilities may be found using the algorithm of Genz (1992).

Alternatively, a simulation-based approach may be used to approximate the chi-bar-square weights which is applicable for any value of r. The method proceeds by i) generating an r-dimensional random vector $\boldsymbol{Z}$ from $N(\boldsymbol{0}, \boldsymbol{D})$, ii) finding $\boldsymbol{\delta}^* = (\delta_1^*, \cdots, \delta_r^*)^T$ that minimizes $q(\boldsymbol{\delta}) = (\boldsymbol{Z} - \boldsymbol{\delta})^T \boldsymbol{D}^{-1} (\boldsymbol{Z} - \boldsymbol{\delta})$ subject to $\boldsymbol{\delta} \geq \boldsymbol{0}$, and then iii) counting the number of strictly positive δ_i^*'s in $\boldsymbol{\delta}^*$, which we denote by M. If these procedures are repeated a sufficient number of times, say $N = 10,000$, then the chi-bar-square weights are approximated by $w_i(r, \boldsymbol{D}) \doteq N_i/N$, $i = 0, 1, \cdots, r$ where N_i is the number of times that $M = i$.

An R program to compute the weights using the simulation approach is outlined in the Appendix. The proof of Theorem 4.3 is provided in the next section.

4.5.3 Proof of Theorem 4.3.

The proof is provided for the nonignorable case, however may be simplified to apply in the ignorable setting. First, recall that the observed-data log-likelihood function is given by

$$l^*(\boldsymbol{\gamma}) \equiv l^*(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{x}_{obs,i}, \boldsymbol{r}) = \sum_{i=1}^{n} \ln f(y_i, \boldsymbol{x}_{obs,i}, r_i|\boldsymbol{\gamma})$$

as in (4.22). Then, the maximum likelihood estimation equations are

$$\sum_{i=1}^{n} E\left[\frac{\partial \log f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|y_i, \boldsymbol{x}_{obs,i}, r_i\right] = \mathbf{0}, \tag{4.36}$$

$$\sum_{i=1}^{n} E\left[\frac{\partial \log f(\boldsymbol{x}_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}|y_i, \boldsymbol{x}_{obs,i}, r_i\right] = \mathbf{0}, \tag{4.37}$$

$$\sum_{i=1}^{n} E\left[\frac{\partial \log f(\boldsymbol{r}_i|y_i, \boldsymbol{x}_i, \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}|y_i, \boldsymbol{x}_{obs,i}, r_i\right] = \mathbf{0}, \tag{4.38}$$

where the conditional expectations are with respect to the density of $\boldsymbol{x}_{mis,i}$ given the observed data. Note that these equations are equivalently solved using the EM algorithm discussed earlier. Next, define $m'(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma}) = (m_1'(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})^T, m_2'(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})^T, m_3'(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})^T)^T$, whose components are the negative of the ith terms in (4.36), (4.37) and (4.38), respectively. As the estimating equations are

$$\sum_{i=1}^{n} m'(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma}) = \mathbf{0},$$

the unconstrained maximum likelihood estimator, $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\alpha}}^{\mathrm{T}}, \hat{\boldsymbol{\psi}}^{\mathrm{T}})^{\mathrm{T}}$, of $\boldsymbol{\gamma}$ is the same as the one that minimizes, with respect to $\boldsymbol{\gamma}$,

$$M_n(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^{n} m(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma}).$$

Further, $l^*(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}) = -nM_n(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma}) + d(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r})$ for some $d(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r})$ which does not depend on $\boldsymbol{\gamma}$.

Regularity assumptions are required to justify the following theoretical results. Note that $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ refer to the true parameter values when H_0 and H_1 are true, respectively.

1. The sequence $m'(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})$, $i = 1, \ldots, n$ are continuous functions of $\boldsymbol{\gamma}$ for each y_i, and measurable functions of y_i for each $\boldsymbol{\gamma} \in \Theta$, a compact subset of a finite-dimensional Euclidean space.

2. The random components $\{y_i\}$ are either (a) ϕ-mixing with $\phi(m)$ of size $r_1/(2r_1 - 1)$, $r \geq 1$; or (b) α-mixing with $\alpha(m)$ of size $r_1/(r_1 - 1)$, $r_1 > 1$.

3. The sequence $m(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})$ is dominated by uniformly $(r_1+\delta)$-integrable functions, $r_1 \geq 1$, $0 < \delta \leq r_1$.

4. The function $\bar{M}_n(\boldsymbol{\gamma}) = E[M_n(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma})]$ has an identifiably unique minimizer $\boldsymbol{\gamma}_0$ (in the sense of Definition 2.1 of Domowitz and White, 1982).

5. The functions $m(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})$ are twice continuously differentiable in $\boldsymbol{\gamma}$, uniformly in i, a.s. - P.

6. For $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p+q})^{\mathrm{T}}$, $\{m'_j(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})^2\}$, $j = 1, \ldots, p+q$ are dominated by uniformly r_2-integrable functions, $r_2 > 1$, where $m'_j(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma}) = \partial m(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})/\partial\gamma_j$.

7. Define $\boldsymbol{Q}_{a,n} = \mathrm{var}[n^{-1/2} \sum_{i=a+1}^{a+n} m'(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma}_0)]$. Assume that there exists a positive definite matrix $\boldsymbol{Q}$ such that $\boldsymbol{\lambda}'\boldsymbol{Q}_{a,n}\boldsymbol{\lambda} - \boldsymbol{\lambda}'\boldsymbol{Q}\boldsymbol{\lambda} \to \mathbf{0}$ as $n \to \infty$, uniformly in a, for any nonzero vector $\boldsymbol{\lambda}$.

8. For $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p+q})'$, $\{m''_{jl}(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})\}$, $j, l = 1, \ldots, p+q$ are dominated by uniformly $(r_1 + \delta)$-integrable functions, $0 < \delta \leq r_1$, where $m''_{jl}(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma}) = \partial^2 m(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})/\partial\gamma_j\partial\gamma_l$.

9. The matrix $\bar{M}''_n(\boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^{n} E[m''(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})]$ has constant rank $p+q$ in some ϵ-neighbourhood of $\boldsymbol{\gamma}_0$, for all sufficiently large n, uniformly in n.

As noted in Sinha (2004), the mixing condition in assumption 2 restricts the memory of the process $\{y_i\}$ in a manner analogous to the rule ergodicity for a stationary stochastic process. More specifically, mixing implies that values of

the process that are far apart in time (or space) are statistically independent. For a precise definition of mixing, the reader is referred to Domowitz and White (1982). Assumption 3 restricts the moments of the process $m(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})$ and assumption 4 gives an identification condition. Assumption 5 requires that the sequence $m(y_i, \boldsymbol{x}_{obs,i}, r_i, \boldsymbol{\gamma})$ is continuously differentiable of order 2 on Θ. Assumption 8 is used to ensure the convergence of the sample Hessian and, together with assumption 6, allows for the calculation of the gradient and Hessian of $\bar{M}_n(\boldsymbol{\gamma})$ by interchanging differentiation with integration. Assumption 9 is used to ensure that $\bar{M}_n''(\boldsymbol{\gamma})$ is positive definite for sufficiently large n. These assumptions together also ensure the nonsingularity of $\bar{M}_n''(\hat{\boldsymbol{\gamma}}_n)$.

Since the likelihood is the product of functions in the exponential family which are known to be concave and bounded, the assumptions of uniqueness, differentiable and finite variance are satisfied. The mixing condition required for establishing the asymptotic results is the most contentious issue. As White (1984) mentioned, such a condition is impossible to verify empirically. However, the extent to which the asymptotic results hold in simulation studies may be assessed empirically. From simulations results, it was noted that the unconstrained values of $\boldsymbol{\gamma}$ were normally distributed under the models considered.

The following strengthening of Assumption 2 is required to establish a version

of the central limit theorem, as per Domowitz and White (1982):

2a. Assumption 2 holds and either $\phi(m)$ is of size $r_2/(r_2-1)$ or $\alpha(m)$ is of size $\max[r_1/(r_1-1), r_2/(r_2-1)]$, $r_1, r_2 > 1$.

We then let $M_n'(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma})$ and $M_n''(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma})$ denote the first and second derivatives of $M_n(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$, respectively. Also let $\bar{M}_n(\boldsymbol{\gamma}) = E[M_n(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma})]$. Since $M_n'(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma})$ is the sum of independent stochastic terms, the multivariate central limit theorem holds under appropriate regularity conditions established previously. These assumptions ensure $M_n(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma})$, $M_n'(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma})$, and $M_n''(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \boldsymbol{\gamma})$ converge almost surely to the limits of $\bar{M}_n(\boldsymbol{\gamma})$, $\bar{M}_n'(\boldsymbol{\gamma})$ and $\bar{M}_n''(\boldsymbol{\gamma})$, respectively. If we denote $\boldsymbol{\gamma}_0$ as the true model parameter value, then $\bar{M}_n''(\boldsymbol{\gamma}_0) = \mathcal{I}_U(\boldsymbol{\gamma}_0)$ since $(y_\imath, \boldsymbol{x}_{obs,\imath}, r_\imath)$, $\imath = 1, \ldots, n$ are independent and identically distributed. Further, $\bar{M}_n'(\boldsymbol{\gamma}_0) = \mathbf{0}$ since $\bar{M}_n(\cdot)$ assumes its minimum value at the true value of model parameter vector. The following lemma summarizes the large sample properties of $\hat{\boldsymbol{\gamma}}$ under missing covariates.

Lemma 1. Let $\boldsymbol{\gamma}_0$ denote the true model parameter value. Then, under the regularity conditions, the following results hold:

(a) $\hat{\gamma} \to \gamma_0$ almost surely as $n \to \infty$,

(b) $\mid M_n(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \hat{\gamma}) - \bar{M}_n(\gamma_0) \mid \to \mathbf{0}$ almost surely as $n \to \infty$,

(c) $\sqrt{n} M_n'(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \hat{\gamma}) \to N(\mathbf{0}, R^*(\gamma_0))$ where $R^*(\gamma_0) = \mathcal{I}_U(\gamma_0)$,

(d) $\sqrt{n}(\hat{\gamma} - \gamma_0) \to N(\mathbf{0}, R^*(\gamma_0)^{-1})$.

Proof of Theorem. The theorem is first proven for the case where the right hand side of inequality constraint is $\mathbf{c} = \mathbf{0}$. Since $M_n'(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \hat{\gamma}) = \mathbf{0}$, the second order Taylor expansion of $M_n(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \gamma)$ about $\hat{\gamma}$ gives

$$M_n(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \gamma) = M_n(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \hat{\gamma}) + \frac{1}{2}(\hat{\gamma} - \gamma)^T M_n''(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \bar{\gamma})(\hat{\gamma} - \gamma); \quad (4.39)$$

where $\bar{\gamma}$ is a point on the segment between γ and $\hat{\gamma}$. By (a) of Lemma 1, $\bar{\gamma}$ also converges to γ_0 almost surely, and hence it follows that $M_n''(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \bar{\gamma}) \to R^*(\gamma_0)$ and $M_n''(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \bar{\gamma}) - M_n''(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \hat{\gamma}) \to \mathbf{0}$ (null matrix) almost surely.

Let $q(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \gamma) = n(\hat{\gamma} - \gamma)^T M_n''(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \bar{\gamma})(\hat{\gamma} - \gamma)$ and $\dot{q}(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \gamma) = n(\hat{\gamma} - \gamma)^T M_n''(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \hat{\gamma})(\hat{\gamma} - \gamma)$. Then, it follows that $\mid q(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \gamma) - \dot{q}(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \gamma) \mid = o_p(1)$. Since $q(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \gamma) = 2[l^*(\hat{\gamma} \mid \boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}) - l^*(\gamma \mid \boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r})]$ by (4.39), the likelihood ratio test statistics are written as

$$T_{01} = q(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \gamma^0) - q(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \gamma^*) \quad \text{and} \quad T_{12} = q(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{r}, \gamma^*) \quad (4.40)$$

where $\boldsymbol{\gamma}^0$ and $\boldsymbol{\gamma}^*$ are estimators of $\boldsymbol{\gamma}$ under H_0 and H_1, respectively.

Next let A^* be the $r \times (p+q+l)$ matrix obtained by augmenting the matrix A by $q+l$ columns of zeros. The parameter spaces under H_0 and H_1 may be redefined as $C_0 = \{\boldsymbol{\gamma} \in \mathbb{R}^{p+q+l} : A^*\boldsymbol{\gamma} = \mathbf{c}\}$ and $C_1 = \{\boldsymbol{\gamma} \in \mathbb{R}^{p+q+l} : A^*\boldsymbol{\gamma} \leq \mathbf{c}\}$. Then, the least squares projections $P_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_0)$ and $P_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_1)$ are the constrained estimators obtained by minimizing $\dot{q}(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\boldsymbol{\gamma}) \equiv \| \sqrt{n}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}) \|^2_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}$ under H_0 and H_1, respectively. Thus, we may approximate T_{01} in (4.40) with $o_p(1)$ error as

$$\begin{aligned} T_{01} &= \| \sqrt{n}(\hat{\boldsymbol{\gamma}} - P_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_0)) \|^2_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})} \\ &\quad - \| \sqrt{n}(\hat{\boldsymbol{\gamma}} - P_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_1)) \|^2_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})} + o_p(1) \\ &= \| \sqrt{n}(P_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_1) - P_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_0)) \|^2_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})} + o_p(1). \end{aligned}$$

Due to properties of the least squares projection, for any $\boldsymbol{\gamma}_0 \in C_0$ we have that

$$\sqrt{n}(P_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_i) - \boldsymbol{\gamma}_0) = P_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}(\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \mid C_i), \; i = 0, 1.$$

Thus, if $\boldsymbol{\gamma}_0$ is the underlying model parameter under H_0, it follows by Lemma 1 and the continuity property of the projection operator that

$$\begin{aligned} T_{01} &= \| P_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}(\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \mid C_1) \\ &\quad - P_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})}(\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \mid C_0) \|^2_{M_n''(\boldsymbol{y},\boldsymbol{x}_{obs},\boldsymbol{r},\hat{\boldsymbol{\gamma}})} + o_p(1) \\ &\to^D \| P_{R^*(\boldsymbol{\gamma}_0)}(Z \mid C_1) - P_{R^*(\boldsymbol{\gamma}_0)}(Z \mid C_0) \|^2_{R^*(\boldsymbol{\gamma}_0)} \end{aligned}$$

where $Z \sim N(\mathbf{0}, R^*(\boldsymbol{\gamma}_0)^{-1})$. From Theorem 3.5, the asymptotic null distribution is chi-bar-square with weights $w_i(r, A^* R^*(\boldsymbol{\gamma}_0)^{-1} A^{*T})$. Since A^* is the matrix formed by augmenting the matrix A by columns of zeros, we can express $A^* R^*(\boldsymbol{\gamma}_0)^{-1} A^{*T}$ in terms of A and the submatrix of $R^*(\boldsymbol{\gamma}_0)^{-1}$ corresponding to $\boldsymbol{\beta}$ and achieve the desired result, as $R^*(\boldsymbol{\gamma}_0) = \mathcal{I}_U(\boldsymbol{\gamma}_0)$ by Lemma 1 (c).

The asymptotic null distributions of T_{12} under H_0 may also be derived. Applying a large sampling approximation to T_{12} in (4.40), we have

$$T_{12} \rightarrow^D \| Z - P_{R^*(\boldsymbol{\gamma}_0)}(Z \mid C_1) \|^2_{R^*(\boldsymbol{\gamma}_0)}$$

for $\boldsymbol{\gamma}_0 \in C_0$. The asymptotic distribution of T_{12} with a parameter $\boldsymbol{\gamma}_0$ in H_0 may be immediately obtained by Theorem 3.5.

When $\mathbf{c} \neq \mathbf{0}$, we reparameterize by $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} - \dot{\boldsymbol{\beta}}$ where $\dot{\boldsymbol{\beta}}$ is a solution to $A\boldsymbol{\beta} = \mathbf{c}$. Then the new parameter spaces are the same as those with $\mathbf{c} = \mathbf{0}$. Using the invariance property of maximum likelihood estimators, the estimators of $\tilde{\boldsymbol{\beta}}$ are those of their original counterparts minus $\dot{\boldsymbol{\beta}}$. Hence, reparameterization does not affect the test statistics in (4.40). Consequently, for any parameter $\boldsymbol{\gamma}_0$ with $A\boldsymbol{\beta}_0 = \mathbf{c}$, the asymptotic null distributions provided in Theorem 4.3 follow. □.

4.6 Empirical Results for Constrained GLM with Incomplete Data

4.6.1 Simulation Study

To investigate the performance of the proposed GP-EM algorithm for generalized linear models with missing covariates, simulation studies were performed for the binary and Poisson regression models. The statistical package R was used for analysis.

A series of 1000 datasets were generated with sample sizes of $n = 100$ and $n = 250$ for each parameter setting. The Bernoulli (μ_i) model considered $\mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))$ while the Poisson (μ_i) model was of the form $\mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$. The covariates $\boldsymbol{x}_i^T = (x_{1i}, x_{2i})$ were assumed to be independent normal with mean $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{\Sigma}_x$. We let x_{1i} be completely observed ($r_{1i} = 0$ for all i) and let some values of x_{2i} be missing at random (MAR) with missing data mechanism

$$\text{logit}(\pi_i) = \text{logit}(P(r_{2i} = 1|y_i, x_{1i})) = \psi_0 + \psi_1 x_{1i} + \psi_2 y_i.$$

For the Bernoulli model, the covariates were generated from a bivariate normal

distribution with $\boldsymbol{\mu}_x = (2, 1)$ and covariance matrix

$$\boldsymbol{\Sigma}_x = \begin{pmatrix} 1 & .2 \\ .2 & 1 \end{pmatrix}.$$

The missing data mechanism was defined as $\text{logit}(\pi_i) = -5 + 1.5x_{1i} + y_i$. For the Poisson model, $\boldsymbol{\mu}_x = (0.5, 0.5)$ and covariance matrix

$$\boldsymbol{\Sigma}_x = \begin{pmatrix} .5 & .1 \\ .1 & .5 \end{pmatrix},$$

while the missing data mechanism was defined as $\text{logit}(\pi_i) = -2+0.5x_{1i}+0.25y_i$. With these parameters, the simulated data had, on average, 32% missing values on x_{2i} for the binary case, compared with 29% for the Poisson model.

We then impose constraints on the regression parameters, as below

$$\begin{array}{ccccccc} \beta_0 & + & \beta_1 & + & \beta_2 & \leq & c_1 \\ \beta_0 & - & \beta_1 & + & \beta_2 & \geq & -c_2 \end{array} \implies A = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

In the Bernoulli setting, $c_1 = 0$, $c_2 = -5$, while for Poisson $c_1 = 1.5, c_2 = 0$. We examine the performance of the simulated bias and mean square error (MSE) of parameter vector $\boldsymbol{\gamma}^T = (\beta_0, \beta_1, \beta_2, \mu_1, \mu_2, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{12})$ using constrained and unconstrained methods. We consider a variety of types of $\boldsymbol{\beta}$, namely (a) both constraints hold with equality (i.e. both are active), (b) at least one constraint

is a strict inequality (i.e. at least one is inactive) and (c) both constraints are inactive (i.e. both are within the constraint cone). The results for each model are presented in the following tables for sample sizes of 100 and 250 observations.

Intuitively, we expect the constrained estimators to have larger bias and smaller mean square error, especially for values of β which lie close to the boundary. In such cases, the unconstrained estimators are more likely to be outside of the cone. Then, the estimators are projected onto the closed convex cone, which results in a bias since the estimates are always forced towards the direction of the cone.

Table 4.1. Simulated Bias and MSE of Constrained and Unconstrained Estimates for Bernoulli Model with Missing Data ($n = 100$)

Case	Parameter	True Value	Unconstrained		Constrained	
			Bias	MSE	Bias	MSE
(a)	β_0	-2.00	-0.2983	1.1047	0.0267	0.1355
	β_1	2.50	0.3096	0.7206	-0.1136	0.0430
	β_2	-0.50	-0.0667	0.1837	0.0189	0.1066
	μ_{x1}	2.00	-0.0027	0.0096	-0.0027	0.0096
	μ_{x2}	1.00	-0.0024	0.0180	-0.0045	0.0180
	σ_1^2	1.00	-0.0090	0.0204	-0.0090	0.0204
	σ_2^2	1.00	-0.0022	0.0302	-0.0034	0.0300
	σ_{12}	0.20	-0.0056	0.0209	-0.0073	0.0209
(b)	β_0	-2.00	-0.1751	0.6602	-0.3531	0.5431
	β_1	1.00	0.0695	0.1405	0.1412	0.1171
	β_2	1.00	0.0922	0.1697	0.0607	0.1493
	μ_{x1}	2.00	0.0000	0.0099	0.0000	0.0099
	μ_{x2}	1.00	0.0039	0.0161	0.0035	0.0160
	σ_1^2	1.00	-0.0121	0.0182	-0.0121	0.0182
	σ_2^2	1.00	-0.0078	0.0281	-0.0080	0.0280
	σ_{12}	0.20	-0.0037	0.0189	-0.0056	0.0187
(b)	β_0	-2.50	-0.1972	0.7477	0.2143	0.2556
	β_1	1.25	0.1122	0.1919	-0.1085	0.0462
	β_2	-1.25	-0.1034	0.2498	0.0518	0.1262
	μ_{x1}	2.00	0.0014	0.0099	0.0014	0.0099
	μ_{x2}	1.00	0.0059	0.0130	0.0042	0.0129
	σ_1^2	1.00	-0.0094	0.0200	-0.0094	0.0200
	σ_2^2	1.00	-0.0080	0.0248	-0.0117	0.0244
	σ_{12}	0.20	0.0009	0.0157	-0.0046	0.0156
(c)	β_0	-2.00	-0.1337	0.5418	-0.2160	0.4738
	β_1	0.90	0.0461	0.1089	0.0794	0.0949
	β_2	0.90	0.0717	0.1399	0.0589	0.1334
	μ_{x1}	2.00	-0.0022	0.0095	-0.0022	0.0095
	μ_{x2}	1.00	-0.0019	0.0155	-0.0022	0.0155
	σ_1^2	1.00	-0.0089	0.0204	-0.0089	0.0204
	σ_2^2	1.00	-0.0106	0.0282	-0.0107	0.0282
	σ_{12}	0.20	-0.0031	0.0203	-0.0041	0.0201

Table 4.2. Simulated Bias and MSE of Constrained and Unconstrained Estimates for Bernoulli Model with Missing Data ($n = 250$)

Case	Parameter	True Value	Unconstrained		Constrained	
			Bias	MSE	Bias	MSE
(a)	β_0	-2.00	-0.0984	0.3047	0.0094	0.0561
	β_1	2.50	0.0879	0.1650	-0.0894	0.0244
	β_2	-0.50	-0.0027	0.0613	0.0334	0.0469
	μ_{x1}	2.00	0.0016	0.0039	0.0016	0.0039
	μ_{x2}	1.00	-0.0003	0.0065	-0.0016	0.0065
	σ_1^2	1.00	-0.0031	0.0078	-0.0031	0.0078
	σ_2^2	1.00	-0.0020	0.0122	-0.0027	0.0121
	σ_{12}	0.20	0.0014	0.0077	0.0005	0.0077
(b)	β_0	-2.00	-0.0472	0.1991	-0.1768	0.1659
	β_1	1.00	0.0283	0.0433	0.0798	0.0394
	β_2	1.00	0.0251	0.0463	0.0089	0.0444
	μ_{x1}	2.00	0.0017	0.0040	0.0017	0.0040
	μ_{x2}	1.00	0.0004	0.0061	0.0003	0.0060
	σ_1^2	1.00	-0.0099	0.0081	-0.0099	0.0081
	σ_2^2	1.00	-0.0071	0.0112	-0.0072	0.0112
	σ_{12}	0.20	-0.0018	0.0074	-0.0030	0.0073
(b)	β_0	-2.50	-0.0491	0.2578	0.1644	0.1210
	β_1	1.25	0.0308	0.0618	-0.0822	0.0236
	β_2	-1.25	-0.0419	0.0793	0.0353	0.0519
	μ_{x1}	2.00	0.0009	0.0037	0.0010	0.0037
	μ_{x2}	1.00	-0.0031	0.0057	-0.0041	0.0056
	σ_1^2	1.00	-0.0042	0.0078	-0.0042	0.0078
	σ_2^2	1.00	-0.0032	0.0104	-0.0052	0.0103
	σ_{12}	0.20	-0.0040	0.0060	-0.0035	0.0060
(c)	β_0	-2.00	-0.0568	0.1853	-0.0859	0.1662
	β_1	0.90	0.0272	0.0374	0.0386	0.0345
	β_2	0.90	0.0312	0.0464	0.0278	0.0456
	μ_{x1}	2.00	-0.0009	0.0037	-0.0009	0.0037
	μ_{x2}	1.00	0.0020	0.0061	0.0019	0.0060
	σ_1^2	1.00	-0.0051	0.0079	-0.0051	0.0079
	σ_2^2	1.00	-0.0045	0.0123	-0.0045	0.0123
	σ_{12}	0.20	-0.0028	0.0072	-0.0031	0.0071

Table 4.3. Simulated Bias and MSE of Constrained and Unconstrained Estimates for Poisson Model with Missing Data ($n = 100$)

			Unconstrained		Constrained	
Case	**Parameter**	**True Value**	Bias	MSE	Bias	MSE
(a)	β_0	0.50	-0.0078	0.0117	0.0155	0.0089
	β_1	0.75	0.0044	0.0082	-0.0435	0.0047
	β_2	0.25	-0.0018	0.0116	0.0045	0.0092
	μ_{x1}	0.50	-0.0023	0.0051	-0.0020	0.0051
	μ_{x2}	0.50	-0.0039	0.0077	0.0031	0.0075
	σ_1^2	0.50	-0.0066	0.0050	-0.0064	0.0051
	σ_2^2	0.50	-0.0071	0.0077	-0.0013	0.0075
	σ_{12}	0.10	-0.0031	0.0040	0.0047	0.0040
(b)	β_0	0.50	-0.0083	0.0125	-0.0133	0.0120
	β_1	0.50	0.0032	0.0091	-0.0098	0.0081
	β_2	0.50	0.0026	0.0090	-0.0058	0.0084
	μ_{x1}	0.50	0.0008	0.0050	0.0008	0.0049
	μ_{x2}	0.50	-0.0025	0.0072	0.0041	0.0069
	σ_1^2	0.50	-0.0079	0.0047	-0.0079	0.0046
	σ_2^2	0.50	-0.0062	0.0066	0.0003	0.0063
	σ_{12}	0.10	-0.0027	0.0033	0.0024	0.0030
(b)	β_0	1.00	-0.0091	0.0082	0.0153	0.0055
	β_1	0.45	0.0021	0.0095	-0.0323	0.0053
	β_2	-0.55	0.0050	0.0098	0.0291	0.0084
	μ_{x1}	0.50	0.0038	0.0051	0.0039	0.0050
	μ_{x2}	0.50	0.0030	0.0066	0.0045	0.0063
	σ_1^2	0.50	-0.0049	0.0049	-0.0050	0.0048
	σ_2^2	0.50	-0.0059	0.0068	-0.0048	0.0065
	σ_{12}	0.10	0.0023	0.0032	0.0006	0.0030
(c)	β_0	1.00	-0.0044	0.0076	0.0076	0.0061
	β_1	0.45	0.0027	0.0097	-0.0152	0.0062
	β_2	-0.45	0.0026	0.0103	0.0139	0.0090
	μ_{x1}	0.50	-0.0012	0.0048	-0.0012	0.0047
	μ_{x2}	0.50	0.0051	0.0067	0.0052	0.0065
	σ_1^2	0.50	-0.0100	0.0048	-0.0099	0.0047
	σ_2^2	0.50	-0.0030	0.0070	-0.0027	0.0067
	σ_{12}	0.10	0.0035	0.0038	0.0022	0.0036

Table 4.4. Simulated Bias and MSE of Constrained and Unconstrained Estimates for Poisson Model with Missing Data ($n = 250$)

Case	Parameter	True Value	Unconstrained		Constrained	
			Bias	MSE	Bias	MSE
(a)	β_0	0.50	-0.0023	0.0045	0.0104	0.0037
	β_1	0.75	0.0001	0.0035	-0.0283	0.0020
	β_2	0.25	0.0019	0.0043	0.0040	0.0037
	μ_{x1}	0.50	-0.0014	0.0020	-0.0014	0.0020
	μ_{x2}	0.50	0.0016	0.0029	0.0070	0.0029
	σ_1^2	0.50	-0.0003	0.0021	-0.0003	0.0021
	σ_2^2	0.50	-0.0044	0.0030	-0.0003	0.0030
	σ_{12}	0.10	-0.0007	0.0015	0.0051	0.0015
(b)	β_0	0.50	-0.0015	0.0049	-0.0055	0.0049
	β_1	0.50	0.0004	0.0036	-0.0082	0.0034
	β_2	0.50	0.0000	0.0038	-0.0053	0.0037
	μ_{x1}	0.50	0.0008	0.0019	0.0008	0.0019
	μ_{x2}	0.50	-0.0006	0.0027	0.0037	0.0026
	σ_1^2	0.50	-0.0027	0.0019	-0.0027	0.0019
	σ_2^2	0.50	-0.0045	0.0028	-0.0003	0.0028
	σ_{12}	0.10	-0.0005	0.0014	0.0027	0.0014
(b)	β_0	1.00	-0.0024	0.0030	0.0133	0.0021
	β_1	0.45	0.0021	0.0039	-0.0199	0.0022
	β_2	-0.55	0.0017	0.0037	0.0159	0.0032
	μ_{x1}	0.50	0.0002	0.0020	0.0002	0.0020
	μ_{x2}	0.50	0.0033	0.0027	0.0039	0.0027
	σ_1^2	0.50	-0.0028	0.0019	-0.0028	0.0019
	σ_2^2	0.50	-0.0044	0.0028	-0.0041	0.0028
	σ_{12}	0.10	-0.0004	0.0013	-0.0019	0.0013
(c)	β_0	1.00	-0.0011	0.0032	0.0036	0.0027
	β_1	0.45	0.0000	0.0036	-0.0065	0.0027
	β_2	-0.45	-0.0002	0.0036	0.0039	0.0033
	μ_{x1}	0.50	0.0001	0.0021	0.0001	0.0021
	μ_{x2}	0.50	-0.0007	0.0025	-0.0005	0.0025
	σ_1^2	0.50	-0.0008	0.0019	-0.0008	0.0019
	σ_2^2	0.50	-0.0030	0.0028	-0.0030	0.0028
	σ_{12}	0.10	0.0004	0.0014	-0.0001	0.0014

We note from the previous tables that in most cases, the constrained estimates have larger bias with a smaller mean square error (MSE) than the unconstrained counterparts. This finding is consistent with order-constrained inference for normal models. For both Bernoulli and Poisson models, the differences are less pronounced in case (c), when the generating point is within the constraint cone and a larger percentage of the unconstrained estimates satisfy the constraints. Also, the parameters pertaining to $\boldsymbol{x}$ are relatively unaffected by the constrained estimation, since no restrictions were imposed. The exception is μ_{x_2}, which has larger bias, however a mean square error equal to or less than the unconstrained estimate. Further, note the asymptotic bias and MSE diminish as the sample size increases from $n = 100$ to $n = 250$, for both unconstrained and constrained settings.

We next consider a comparative analysis of constrained and unconstrained maximum likelihood estimation in the presence of missing data, full data and complete cases (cases with missing data removed). The binary and Poisson models are displayed in Tables 4.5 and 4.6 respectively, for the $n = 250$ case. Values of $\boldsymbol{\beta}$ on the boundary and within the constraint space are considered.

The results show that the constrained estimates tend to have larger bias and smaller MSE regardless of the estimation procedure used, i.e. for the GP-EM

algorithm, full data or complete cases. Estimation with the full data is the most efficient, yet the pattern of larger bias and smaller mean square error continues for estimation with constraints. In most settings, the GP-EM algorithm performs well, with MSE values only slightly larger than the full data counterparts. Since the missing data mechanism depends on the response y_i, the complete case is highly inefficient and exhibits much larger bias in most cases. In particular, since larger values of x_1 are likely to be missing, a large negative bias is exhibited for the estimation of μ_1 with only the complete cases.

Thus, as in the case of full data, the GP-EM algorithm for maximum likelihood estimation with missing data under parameter constraints performs well, and provides more efficient estimators as compared to the analysis based on complete cases. While estimation with the full data is preferred, the GP-EM algorithm offers an efficient alternative to both unconstrained and constrained estimation, and outperforms analysis with only complete cases.

4.6.2 LRT Power Comparisons

Powers of constrained and unconstrained likelihood ratio tests were compared for Bernoulli and Poisson distributions in a small empirical study. The statistical package R was used for analysis.

Table 4.5. Simulated Bias and Mean Squared Error of Constrained and Unconstrained Estimates for Bernoulli Model: Comparison of Full Data, GP-EM Algorithm and Complete Cases

Parameter	True Value	Full Data				GP-EM Algorithm				Complete Cases			
		Unconstrained		Constrained		Unconstrained		Constrained		Unconstrained		Constrained	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
β_0	-2 00	-0 0358	0 1757	-0 1590	0 1417	-0 0472	0 1991	-0 1768	0 1659	0 0336	0 2183	-0 0424	0 1722
β_1	1 00	0 0165	0 0394	0 0647	0 0347	0 0283	0 0433	0 0798	0 0394	-0 1348	0 0725	-0 1038	0 0576
β_2	1 00	0 0239	0 0382	0 0102	0 0366	0 0251	0 0463	0 0089	0 0444	0 0269	0 0516	0 0165	0 0502
μ_{x1}	2 00	-0 0010	0 0037	-0 0011	0 0037	0 0017	0 0040	0 0017	0 0040	-0 3358	0 1171	-0 3357	0 1170
μ_{x2}	1 00	-0 0010	0 0040	-0 0008	0 0040	0 0004	0 0061	0 0003	0 0060	-0 0953	0 0146	-0 0953	0 0146
σ_1^2	1 00	-0 0012	0 0072	-0 0010	0 0072	-0 0099	0 0081	-0 0099	0 0081	-0 2432	0 0660	-0 2432	0 0660
σ_2^2	1 00	-0 0041	0 0073	-0 0038	0 0073	-0 0071	0 0112	-0 0072	0 0112	-0 0051	0 0118	-0 0051	0 0118
σ_{12}	0 20	0 0002	0 0040	-0 0001	0 0040	-0 0018	0 0074	-0 0030	0 0073	-0 0637	0 0086	-0 0637	0 0086
β_0	-2 25	-0 0675	0 1964	-0 0839	0 1825	-0 0444	0 2035	-0 0663	0 1868	0 0330	0 2314	0 0223	0 2182
β_1	0 95	0 0272	0 0379	0 0335	0 0359	0 0202	0 0381	0 0288	0 0358	-0 1435	0 0722	-0 1391	0 0688
β_2	1 05	0 0243	0 0380	0 0227	0 0377	0 0238	0 0463	0 0212	0 0458	0 0300	0 0519	0 0286	0 0515
μ_{x1}	2 00	-0 0010	0 0037	-0 0010	0 0037	0 0017	0 0040	0 0017	0 0040	-0 3300	0 1131	-0 3300	0 1131
μ_{x2}	1 00	-0 0010	0 0039	-0 0010	0 0039	0 0008	0 0059	0 0007	0 0059	-0 0977	0 0149	-0 0977	0 0149
σ_1^2	1 00	-0 0012	0 0072	-0 0012	0 0072	-0 0099	0 0081	-0 0099	0 0081	-0 2418	0 0654	-0 2418	0 0654
σ_2^2	1 00	-0 0041	0 0073	-0 0041	0 0073	-0 0072	0 0111	-0 0072	0 0111	-0 0076	0 0118	-0 0075	0 0118
σ_{12}	0 20	0 0002	0 0040	0 0002	0 0040	-0 0017	0 0071	-0 0020	0 0071	-0 0670	0 0090	-0 0670	0 0090

Table 4.6. Simulated Bias and Mean Squared Error of Constrained and Unconstrained Estimates for Poisson Model: Comparison of Full Data, GP-EM Algorithm and Complete Cases

Parameter	True Value	Full Data				GP-EM Algorithm				Complete Cases			
		Unconstrained		Constrained		Unconstrained		Constrained		Unconstrained		Constrained	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
β_0	0 50	-0 0016	0 0040	-0 0076	0 0040	-0 0015	0 0049	-0 0055	0 0049	-0 0472	0 0083	-0 0473	0 0082
β_1	0 50	0 0010	0 0027	-0 0036	0 0027	0 0004	0 0036	-0 0082	0 0034	-0 0404	0 0067	-0 0406	0 0066
β_2	0 50	0 0003	0 0027	-0 0042	0 0026	0 0000	0 0038	-0 0053	0 0037	-0 0221	0 0057	-0 0223	0 0057
μ_{x1}	0 50	-0 0001	0 0020	-0 0001	0 0020	0 0008	0 0019	0 0008	0 0019	-0 1324	0 0202	-0 1324	0 0202
μ_{x2}	0 50	-0 0014	0 0020	-0 0014	0 0020	-0 0006	0 0027	0 0037	0 0026	-0 0780	0 0087	-0 0780	0 0087
σ_1^2	0 50	-0 0004	0 0019	-0 0004	0 0019	-0 0027	0 0019	-0 0027	0 0019	-0 0522	0 0049	-0 0522	0 0049
σ_2^2	0 50	-0 0011	0 0018	-0 0010	0 0018	-0 0045	0 0028	-0 0003	0 0028	-0 0291	0 0034	-0 0291	0 0034
σ_{12}	0 10	-0 0010	0 0011	-0 0010	0 0011	-0 0005	0 0014	0 0027	0 0014	-0 0351	0 0025	-0 0351	0 0025
β_0	0 50	0 0001	0 0039	0 0130	0 0027	-0 0023	0 0045	0 0104	0 0037	-0 0499	0 0087	-0 0289	0 0052
β_1	0 75	-0 0003	0 0026	-0 0247	0 0014	0 0001	0 0035	-0 0283	0 0020	-0 0495	0 0085	-0 0797	0 0089
β_2	0 25	0 0000	0 0026	-0 0014	0 0022	0 0019	0 0043	0 0040	0 0037	-0 0047	0 0051	0 0039	0 0047
μ_{x1}	0 50	0 0010	0 0020	0 0010	0 0020	-0 0014	0 0020	-0 0014	0 0020	-0 1541	0 0263	-0 1541	0 0263
μ_{x2}	0 50	0 0006	0 0020	0 0006	0 0020	0 0016	0 0029	0 0070	0 0029	-0 0574	0 0060	-0 0574	0 0060
σ_1^2	0 50	0 0036	0 0020	-0 0036	0 0020	-0 0003	0 0021	-0 0003	0 0021	-0 0732	0 0076	-0 0733	0 0076
σ_2^2	0 50	0 0009	0 0021	0 0009	0 0021	-0 0044	0 0030	-0 0003	0 0030	-0 0152	0 0030	-0 0152	0 0030
σ_{12}	0 10	-0 0007	0 0010	-0 0007	0 0010	-0 0007	0 0015	0 0051	0 0015	-0 0318	0 0022	-0 0319	0 0022

As before, a series of 500 datasets were generated with sample size of $n = 250$ per replication. The Bernoulli (μ_i) model considered $\mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))$ while the Poisson (μ_i) model was of the form $\mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$. The covariates $\boldsymbol{x}_i^T = (x_{1i}, x_{2i})$ were assumed to be independent normal with mean $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{\Sigma}_x$. We let x_{1i} be completely observed ($r_{1i} = 0$ for all i) and let some values of x_{2i} be missing at random (MAR) with missing data mechanism

$$\text{logit}(\pi_i) = \text{logit}(P(r_{2i} = 1|y_i, x_{1i})) = \psi_0 + \psi_1 x_{1i} + \psi_2 y_i.$$

For the Bernoulli model, the covariates were generated from a bivariate normal distribution with $\boldsymbol{\mu}_x = (2, 1)$ and covariance matrix

$$\boldsymbol{\Sigma}_x = \begin{pmatrix} 1 & .2 \\ .2 & 1 \end{pmatrix}.$$

The missing data mechanism was defined as $\text{logit}(\pi_i) = -5 + 1.5x_{1i} + y_i$. For the Poisson model, $\boldsymbol{\mu}_x = (0.5, 0.5)$ and covariance matrix

$$\boldsymbol{\Sigma}_x = \begin{pmatrix} .5 & .1 \\ .1 & .5 \end{pmatrix},$$

while the missing data mechanism was defined as $\text{logit}(\pi_i) = -2 + 0.5x_{1i} + 0.25y_i$. With these parameters, the simulated data had, on average, 32% missing values on x_{2i} for the binary case, compared with 29% for the Poisson model.

We then impose constraints on the regression parameters, as below

$$\begin{array}{rcrcrcr} \beta_0 & + & \beta_1 & + & \beta_2 & \leq & c_1 \\ \beta_0 & - & \beta_1 & + & \beta_2 & \geq & -c_2 \end{array} \implies A = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

In the Bernoulli setting, $c_1 = 0$, $c_2 = -5$, while for Poisson $c_1 = 1.5, c_2 = 0$. We evaluated values of $\boldsymbol{\beta}$ under (a) H_0 where both constraints are active and (b) H_1 where at least one constraint is inactive. Table 4.7 lists the percentages of rejecting the null hypothesis in 500 replications at the 5% level of significance.

As per previous notation, T_{01} represents the constrained hypothesis test of H_0 versus $H_1 - H_0$, T_{12} represents the constrained goodness-of-fit test H_1 versus $H_2 - H_1$, and T_{02} represents the unconstrained likelihood ratio test H_0 versus $H_2 - H_0$. Constrained estimates were obtained using the Gradient-Projection EM algorithm, as discussed previously.

Table 4.7. LRT power comparisons for constrained and unconstrained models with incomplete data for Bernoulli and Poisson models at 5% significance level

	Bernoulli Model				Poisson Model			
Case	$\boldsymbol{\beta}^{\mathrm{T}}$	T_{01}	T_{02}	T_{12}	$\boldsymbol{\beta}^{\mathrm{T}}$	T_{01}	T_{02}	T_{12}
	$(-2.00, 2.50, -0.50)$	4.0	3.8	5.4	$(0.40, 0.75, 0.35)$	5.2	4.2	3.0
(a)	$(-2.20, 2.50, -0.30)$	4.2	3.6	5.2	$(0.50, 0.75, 0.25)$	5.8	4.6	4.4
	$(-2.25, 2.50, -0.25)$	4.0	6.6	6.0	$(0.60, 0.75, 0.15)$	5.2	4.2	3.0
	$(-1.75, 2.25, -0.50)$	19.6	11.2	2.0	$(0.45, 0.70, 0.35)$	16.6	11.2	2.2
(b)	$(-1.75, 2.00, -0.25)$	48.8	31.4	1.4	$(0.45, 0.65, 0.40)$	46.4	33.8	1.4
	$(-2.25, 2.30, -0.45)$	74.2	63.6	1.8	$(0.45, 0.60, 0.45)$	76.4	68.6	1.2
	$(-2.25, 2.20, -0.55)$	95.8	91.8	1.0	$(0.40, 0.60, 0.40)$	87.6	75.8	0.0

In case (a), we anticipate the empirical sizes to have values close to the nominal value of 5%, which occurs for both the Bernoulli and Poisson models. Powers of the tests when the first and/or second constraints are active, are described in case (b). Values corresponding to T_{01} and T_{02} are empirical powers, with the constrained test T_{01} exhibiting improved performance in all settings. These results are consistent with the fact that the constrained test incorporates information pertaining to the constrained parameter space in the hypothesis test. In addition, values of T_{12} in case (b) represent sizes as opposed to powers as the values of $\boldsymbol{\beta}$ are within H_1. These values are smaller than the nominal level of 5% since the particular null value is not the least favourable value. Hence, for both

cases, sizes of T_{12} are consistent with previous results for linear models. With incomplete covariate data, the constrained test T_{01} exhibits reasonable empirical size, and significantly better power performance than T_{02} when the constraints are satisfied.

Chapter 5

Inference for GLMMs under Inequality Constraints

In this chapter we detail the derivation of constrained ML estimators and hypothesis tests in the generalized linear mixed model case. We briefly review the GLMM, and discuss limitations of previous research before extending such results.

5.1 Generalized Linear Mixed Models

For many applications, data are often clustered or observed longitudinally. Let the observed response vector for the ith individual or cluster be $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})^T$. Conditional on the random effects $\boldsymbol{u}_i = (u_{i1}, \cdots, u_{im})^T$, the y_{ij}'s are independent, and follow a distribution in the exponential family:

$$f_{y_{ij}|\boldsymbol{u}_i}(y_{ij} \mid \boldsymbol{u}_i, \boldsymbol{\beta}, \tau) = \exp\{(y_{ij}\eta_{ij} - b(\eta_{ij}))/a(\tau) + c(y_{ij}, \tau)\} \tag{5.1}$$

with $i = 1, \ldots, k, j = 1 \ldots n_i$ and where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are specified functions, η_{ij}'s are canonical parameters, and τ is the dispersion parameter. The overall response vector is $\boldsymbol{y} = (\boldsymbol{y}_1^{\mathrm{T}}, \ldots, \boldsymbol{y}_k^{\mathrm{T}})^{\mathrm{T}}$. The conditional mean and variance of y_{ij} are $E[y_{ij}|\boldsymbol{u}_i] = \mu_{ij} = b'(\eta_{ij})$ and $Var[y_{ij}|\boldsymbol{u}_i] = a(\tau)b''(\eta_{ij})$, respectively. Canonical parameters are systematically expressed as $\eta_{ij} = g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{u}_i$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients, $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are vectors of explanatory variables associated with the fixed and random effects respectively, and $g(\cdot)$ is the canonical link function. Further, assume the vector of random effects $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)^{\mathrm{T}}$ follows a distribution

$$\boldsymbol{u} \sim f_{\boldsymbol{u}}(\boldsymbol{u} \mid \boldsymbol{\theta}) \tag{5.2}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$ represents the variance components. Typically, $f_{\boldsymbol{u}}(\boldsymbol{u} \mid \boldsymbol{\theta})$ is assumed to have no relation with $\boldsymbol{\beta}$. In particular, we assume the $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ are mutually independent with a common distribution having mean $\mathbf{0}$ and vector of variance components $\boldsymbol{\theta}$.

From models (5.1) and (5.2), the marginal likelihood of $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \tau^T, \boldsymbol{\theta}^T)^T$ given data $\boldsymbol{y}$ may be expressed as

$$L(\boldsymbol{\beta}, \tau, \boldsymbol{\theta} \mid \boldsymbol{y}) = \prod_{i=1}^{k} \int \prod_{j=1}^{n_i} f_{y_{ij}|\boldsymbol{u}_i}(y_{ij} \mid \boldsymbol{u}_i, \boldsymbol{\beta}, \tau) f_{\boldsymbol{u}_i}(\boldsymbol{u}_i \mid \boldsymbol{\theta}) d\boldsymbol{u}_i. \tag{5.3}$$

In the unconstrained case, differentiation of $l^*(\boldsymbol{\gamma}|\boldsymbol{y}) \equiv \ln L(\boldsymbol{\beta}, \tau, \boldsymbol{\theta} \mid \boldsymbol{y})$ leads to the following estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$,

$$s_{\boldsymbol{\beta}}(\boldsymbol{\gamma}) \equiv \sum_{i=1}^{k} E\left[\frac{\partial \log f_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i \mid \boldsymbol{u}_i, \boldsymbol{\beta}, \tau)}{\partial \boldsymbol{\beta}} | \boldsymbol{y}_i\right] = \mathbf{0}, \tag{5.4}$$

$$s_{\boldsymbol{\theta}}(\boldsymbol{\gamma}) \equiv \sum_{i=1}^{k} E\left[\frac{\partial \log f_{\boldsymbol{u}_i}(\boldsymbol{u}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \boldsymbol{y}_i\right] = \mathbf{0}, \tag{5.5}$$

where the expectation is taken with respect to the conditional distribution of $\boldsymbol{u}_i$ given $\boldsymbol{y}_i$. Treating $(\boldsymbol{y}, \boldsymbol{u})$ as the "complete data," the log-likelihood of $\boldsymbol{\gamma}$ given $(\boldsymbol{y}, \boldsymbol{u})$ may be expressed as $l(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{u}) = \sum_{i=1}^{k} \log f_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i \mid \boldsymbol{u}_i, \boldsymbol{\beta}, \tau) + \sum_{i=1}^{k} \log f_{\boldsymbol{u}_i}(\boldsymbol{u}_i \mid \boldsymbol{\theta})$. Analogous to Louis (1982), the observed information matrix for model (5.1) is

$$\mathcal{I}_o(\boldsymbol{\gamma}) = -\sum_{i=1}^{k} E\left[\partial t_i(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^T \mid \boldsymbol{y}_i\right] - \sum_{i=1}^{k} E\left[t_i(\boldsymbol{\gamma}) t_i(\boldsymbol{\gamma})^T \mid \boldsymbol{y}_i\right]$$

$$+\sum_{i=1}^{k} E\left[t_i(\boldsymbol{\gamma}) \mid \boldsymbol{y}_i\right] E\left[t_i(\boldsymbol{\gamma}) \mid \boldsymbol{y}_i\right]^T \tag{5.6}$$

where $t_i(\boldsymbol{\gamma}) = \partial l(\boldsymbol{\gamma}|\boldsymbol{y}_i, \boldsymbol{u}_i)/\partial\boldsymbol{\gamma}$ and the expectations are taken with respect to the conditional distribution of $\boldsymbol{u}_i$ given $\boldsymbol{y}_i$. In many practical situations such as the binary and Poisson regression models, the dispersion parameter function is fixed at unity. Since it is possible to extend to the general case with a minor change, we assume $a(\tau) = 1$ and hence $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ for simplicity. Then, the observed information, $\mathcal{I}_o(\boldsymbol{\gamma})$, may be decomposed into the form

$$\mathcal{I}_o(\boldsymbol{\gamma}) = \begin{bmatrix} \mathcal{I}_o(\boldsymbol{\beta}, \boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\beta}, \boldsymbol{\theta}) \\ \mathcal{I}_o(\boldsymbol{\theta}, \boldsymbol{\beta}) & \mathcal{I}_o(\boldsymbol{\theta}, \boldsymbol{\theta}) \end{bmatrix}. \tag{5.7}$$

In subsequent sections, we will denote the unconstrained estimators of $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ by $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\theta}}^T)^T$.

5.2 Previous Research

In the context of linear mixed models, only a few authors have proposed algorithms for maximum likelihood estimation under inequality or equality constraints (Shi et al.(2005), Zheng et al.(2005), Kim and Taylor(1995)). Nevertheless, most of these algorithms concern a specific cone type of constraint, such as $A\boldsymbol{\beta} \leq \mathbf{0}$. In contrast, Jamshidian (2004) used the Gradient Projection (GP)

algorithm for equality and inequality constraints of a general likelihood function with missing values. Given the general nature of this procedure, both in terms of linear inequality constraints and general likelihood function, we are motivated to extend these results to the nonlinear mixed model context.

For a slightly different problem, Hall and Praestgaard (2001) developed an order restricted score test for homogeneity of the random effects variance of the GLMMs. The test is defined as: $H_0 : \boldsymbol{\Sigma}(\mathbf{0}) = \mathbf{0}$ against the alternative that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive semidefinite, where $\boldsymbol{\theta}$ represents a vector of variance components of $\boldsymbol{\Sigma}$. The proposed score test is a constrained version of Lin (1997)'s test statistic. The authors also derive the limiting distribution as chi-bar-square, and demonstrate through simulation studies that their methods have higher statistical power than unconstrained score tests.

Recent papers have demonstrated improvements in efficiency for testing homogeneity of the random effects variance. Fitzmaurice et. al. (2007) developed a permutation test which was seen to outperform the above score test in small samples. Sinha (2009) extended these results to consider a bootstrap test for variance components in GLMM, which also performed better than the constrained score test based on chi-square mixtures in small samples. However, for the remainder of this chapter, we will discuss order-restricted estimation and testing

for regression parameters.

In later sections, we specify the GLMM for clustered or longitudinal data, and derive constrained maximum likelihood estimators and likelihood ratio tests. Simulation studies to assess the bias and mean square error, as well as power comparisons for the likelihood ratio test demonstrate improvements in efficiency and statistical power of these constrained procedures.

5.3 Constrained Maximum Likelihood Inference for GLMMs

In this section, we consider restricted ML estimation in GLMMs using the well-studied gradient projection method.

5.3.1 Gradient Projection Algorithm for GLMMs

Let $C_1 = \{(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T : A\boldsymbol{\beta} \leq \boldsymbol{c}\}$ denote the constrained parameter space, where A is a $r \times p$ matrix of full row rank, $r \leq p$. In order to maximize the log likelihood function under such inequality constraints, we implement a modified version of the gradient projection algorithm of Jamshidian (2004) which searches active

constraint sets to determine the optimal solution, as discussed in Chapter 3. An active constraint set signifies the set of constraints which hold with equality.

In generalized linear mixed models, unlike the case in Jamshidian (2004), we maximize the marginal log-likelihood $l^*(\boldsymbol{\gamma}|\boldsymbol{y})$ under linear inequality constraints on the regression parameters. Since the parameter vector is composed of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, we partition the inverse of the observed information matrix (5.7) as follows:

$$\mathcal{I}_o^{-1}(\boldsymbol{\gamma}) \equiv W(\boldsymbol{\gamma}) = \begin{bmatrix} W_{11}(\boldsymbol{\gamma}) & W_{12}(\boldsymbol{\gamma}) \\ W_{21}(\boldsymbol{\gamma}) & W_{22}(\boldsymbol{\gamma}) \end{bmatrix}$$

where $W_{11}(\boldsymbol{\gamma}) = [\mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\beta}) - \mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\theta})\mathcal{I}_o^{-1}(\boldsymbol{\theta},\boldsymbol{\theta})\mathcal{I}_o(\boldsymbol{\theta},\boldsymbol{\beta})]^{-1}$,

$W_{12}(\boldsymbol{\gamma}) = -\mathcal{I}_o^{-1}(\boldsymbol{\beta},\boldsymbol{\beta})\mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\theta})[\mathcal{I}_o(\boldsymbol{\theta},\boldsymbol{\theta}) - \mathcal{I}_o(\boldsymbol{\theta},\boldsymbol{\beta})\mathcal{I}_o^{-1}(\boldsymbol{\beta},\boldsymbol{\beta})\mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\theta})]^{-1}$,

$W_{22}(\boldsymbol{\gamma}) = [\mathcal{I}_o(\boldsymbol{\theta},\boldsymbol{\theta}) - \mathcal{I}_o(\boldsymbol{\theta},\boldsymbol{\beta})\mathcal{I}_o^{-1}(\boldsymbol{\beta},\boldsymbol{\beta})\mathcal{I}_o(\boldsymbol{\beta},\boldsymbol{\theta})]^{-1}$, and $W_{21}(\boldsymbol{\gamma}) = W_{12}^T(\boldsymbol{\gamma})$. Then, the generalized score vector can be expressed as $s^*(\boldsymbol{\gamma}) = \mathcal{I}_o^{-1}(\boldsymbol{\gamma})s_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) =$

$(s_1^*(\boldsymbol{\gamma})^T, s_2^*(\boldsymbol{\gamma})^T)^T$ where $s_1^*(\boldsymbol{\gamma}) = W_{11}(\boldsymbol{\gamma})s_{\boldsymbol{\beta}}(\boldsymbol{\gamma}) + W_{12}(\boldsymbol{\gamma})s_{\boldsymbol{\theta}}(\boldsymbol{\gamma})$ and

$s_2^*(\boldsymbol{\gamma}) = W_{21}(\boldsymbol{\gamma})s_{\boldsymbol{\beta}}(\boldsymbol{\gamma}) + W_{22}(\boldsymbol{\gamma})s_{\boldsymbol{\theta}}(\boldsymbol{\gamma})$.

If the unconstrained estimate satisfies the constraints so that $\hat{\boldsymbol{\gamma}} \in C_1$, then the constrained estimate is identical to $\hat{\boldsymbol{\gamma}}$. Otherwise, we proceed with the following algorithm with an initial value, $\boldsymbol{\gamma}_r$, chosen from C_1. Based on the constraints that hold with equality at $\boldsymbol{\gamma}_r$, we form the active constraint set $\mathcal{W}$, as well as coefficient matrix $\tilde{A}$ and constraint vector $\tilde{\mathbf{c}}$.

Step 1: Using the current value $\boldsymbol{\gamma}_r$ of $\boldsymbol{\gamma}$, evaluate $W(\boldsymbol{\gamma}_r)$.

Step 2: Compute $B(\boldsymbol{\gamma}_r) = \tilde{A}^T[\tilde{A}W_{11}(\boldsymbol{\gamma}_r)\tilde{A}^T]^{-1}\tilde{A}$, and $\boldsymbol{d} = (\boldsymbol{d}_1^T, \boldsymbol{d}_2^T)^T$ where $\boldsymbol{d}_1 = [I - W_{11}(\boldsymbol{\gamma}_r)B(\boldsymbol{\gamma}_r)]s_1^*(\boldsymbol{\gamma}_r)$ and $\boldsymbol{d}_2 = -W_{21}(\boldsymbol{\gamma}_r)B(\boldsymbol{\gamma}_r)s_1^*(\boldsymbol{\gamma}_r) + s_2^*(\boldsymbol{\gamma}_r)$.

Step 3: If $\boldsymbol{d} = \mathbf{0}$, compute the Lagrange multiplier vector $\boldsymbol{\lambda} = [\tilde{A}W_{11}(\boldsymbol{\gamma}_r)\tilde{A}^T]^{-1}\tilde{A}s_1^*(\boldsymbol{\gamma}_r)$.

a) If $\lambda_i \geq 0$ for all $i \in \mathcal{W}$, the current point is the constrained estimate. Stop.

b) If $\lambda_i < 0$ for some $i \in \mathcal{W}$, drop the index corresponding to the smallest λ_i from $\mathcal{W}$. Form new $\tilde{A}$ and $\tilde{\mathbf{c}}$ based on the constraints remaining in $\mathcal{W}$. Go to Step 2.

Step 4: If $\boldsymbol{d} \neq \mathbf{0}$, find $\delta_1 = \arg\max_\delta\{\delta : \boldsymbol{\gamma}_r + \delta\boldsymbol{d} \in C_1\}$ and then determine δ_2 such that $l^*(\boldsymbol{\gamma}_r + \delta_2\boldsymbol{d} \mid \boldsymbol{y}) \geq l^*(\boldsymbol{\gamma}_r + \delta\boldsymbol{d} \mid \boldsymbol{y})$ for $0 \leq \delta \leq \delta_1$. Let $\boldsymbol{\gamma}_r^* = \boldsymbol{\gamma}_r + \delta_2\boldsymbol{d}$ and determine which constraints newly hold with equality at $\boldsymbol{\gamma}_r^*$. Add their indexes, if any, to $\mathcal{W}$ and update $\tilde{A}$ and $\tilde{\mathbf{c}}$ accordingly.

Step 5: Replace $\boldsymbol{\gamma}_r$ by $\boldsymbol{\gamma}_r^*$ and go to Step 1.

At each stage of the above algorithm, the calculation of the observed information matrix, score vector or marginal likelihood requires evaluation of conditional

expectations of certain functions of $\boldsymbol{u}$ given $\boldsymbol{y}$ with intermediate parameter value $\boldsymbol{\gamma}_r$. We rely on approximation of these expectations by numerical methods. Although this is a rather computationally intensive procedure, the high power of modern computers have mitigated these difficulties.

While the preceding algorithm was designed for the estimation of parameters under $H_1 : A\boldsymbol{\gamma} \leq \mathbf{c}$, it may also be used, with a slight modification, to estimate parameters under $H_0 : A\boldsymbol{\gamma} = \mathbf{c}$. If we retain $\mathcal{W} = \{1, \cdots, r\}$, $\tilde{A} = A$ and $\tilde{\mathbf{c}} = \mathbf{c}$ throughout the algorithm and modify Step 3 so that it stops if $\boldsymbol{d} = \mathbf{0}$, then the algorithm provides the estimates under H_0. We will let $\boldsymbol{\gamma}^*$ and $\boldsymbol{\gamma}^0$ denote estimates constrained by H_1 and H_0, respectively.

5.3.2 Numerical Example

To illustrate the preceding algorithm, we generated data from a Poisson GLMM model for analysis. More specifically, the data follow a Poisson(λ_{ij}) distribution with $\lambda_{ij} = \exp(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + u_i)$, with $\boldsymbol{\beta} = (1.25, 0.75)^{\mathrm{T}}$ and $\mathbf{x}_{ij}^{\mathrm{T}}$ simulated from $U(-2.5, 2.5)$. We let the random effects $u_i \sim N(0, \sigma^2 = 0.25)$, with $k = 100$ clusters and $n_i = 4$ observations per cluster.

Overall, the generated y_{ij} variables have the following summary statistics:

Table 5.1. Summary Statistics for Generated Poisson Dataset

N	Minimum	Q_1	Median	$\bar{y}$	Q_3	Maximum
400	0	1	3.5	6.94	10	42

Then, using y_{ij} and the cluster identifiers, we first fit an exact unconstrained GLMM using the R function `glmmML()`, which employs adaptive Gaussian quadrature methods. Then, the link function becomes $g(\lambda_{ij}) = \log(\lambda_{ij}) = \beta_0 + \beta_1 x_{ij}$ with the following parameter estimates:

Table 5.2. Unconstrained GLMM Estimates for Generated Poisson Dataset

Coefficient	Estimate	Standard Error	Z statistic	P-value
β_0	1.3426	0.05736	23.41	< 0.0001
β_1	0.7000	0.03922	17.85	< 0.0001
σ	0.4715	0.03711		

We next impose the following parameter constraints:

$$\begin{array}{rcrcl} \beta_0 & + & \beta_1 & \leq & 2 \\ \beta_0 & - & \beta_1 & \geq & 0.5 \end{array} \quad \Longrightarrow \quad A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 2 \\ -0.5 \end{bmatrix}.$$

Thus, since the unconstrained MLE $\hat{\boldsymbol{\beta}} = (1.3426, 0.7000)^{\mathrm{T}}$ does not satisfy the constraints, we may use the GP algorithm to compute the constrained parameter estimates. Note that only one constraint is violated, namely the first constraint. Hence, this constraint becomes active, and to implement the GP algorithm we consider the initial working set to be $\mathcal{W} = \{1\}$. The GP Algorithm proceeds as follows:

Iteration (0): Set $m = 1$, $\overline{A} = [1\ 1]$, and $\overline{\mathbf{c}} = [2]$. Choose as a starting value: $\boldsymbol{\gamma}_0 = (1.45, 0.55, 0.23)^{\mathrm{T}}$.

Iteration (1): Compute the projection matrix and distance vector as follows:

$$B(\boldsymbol{\gamma}_0) = \begin{bmatrix} 0.31724462 & -0.68275538 \\ -0.31724462 & 0.68275538 \end{bmatrix}, \quad \mathbf{d}_0 = \begin{bmatrix} -0.13827557 \\ 0.13827557 \\ 0.00524472 \end{bmatrix}.$$

Using a tolerance level of 0.001 for convergence, we note that $\mathbf{d}_0 > 0.001 \times \mathbf{1}$, so we move to Step 3. We find $\delta_1 = 1.446$, and $\delta_2 = 1.0096$, so then the updated value is

$$\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_0 + \delta_2 \mathbf{d}_0 = \begin{bmatrix} 1.3104 \\ 0.6896 \\ 0.2353 \end{bmatrix}.$$

Iteration (2): As before, we compute the revised projection matrix and distance vector as

$$B(\boldsymbol{\gamma}_1) = \begin{bmatrix} 0.32383110 & -0.67616890 \\ -0.32383110 & 0.67616890 \end{bmatrix}, \quad \mathbf{d}_1 = \begin{bmatrix} -0.0001018 \\ 0.0001018 \\ -0.0018812 \end{bmatrix}.$$

While the distance vector values for $\boldsymbol{\beta}$ satisfy the tolerance level, the value for σ^2 does not. Hence, we must continue with Step 3 and compute a new $\delta_1 = 592.9$, and $\delta_2 = 4.465$. Hence the updated parameter vector is then calculated as

$$\boldsymbol{\gamma}_2 = \boldsymbol{\gamma}_1 + \delta_2 \mathbf{d}_1 = \begin{bmatrix} 1.3099 \\ 0.6900 \\ 0.2269 \end{bmatrix}.$$

Iteration (3): In this case, using $\boldsymbol{\gamma}_2$, we find the new distance vector and projection matrix to be:

$$B(\boldsymbol{\gamma}_2) = \begin{bmatrix} 0.3113223 & -0.6886777 \\ -0.3113223 & 0.6886777 \end{bmatrix}, \quad \mathbf{d}_2 = \begin{bmatrix} 0.0008514 \\ -0.0008514 \\ -0.0001407 \end{bmatrix}.$$

Since all three values of $\mathbf{d}_2$ are less than 0.001, then convergence is achieved, and $\boldsymbol{\gamma}_2$ becomes the constrained estimator. The Lagrange Multiplier for the

active constraint in $\mathcal{W}$ is $\lambda_1 = 11.69 \neq 0$, hence the Kuhn-Tucker conditions for optimality are satisfied. The results of the analysis are summarized in Table 5.3:

Table 5.3. Constrained and Unconstrained GLMM Estimates for Generated Poisson Dataset

Coefficient	Constrained	Unconstrained	Standard Error*
β_0	1.3099	1.3426	0.05736
β_1	0.6900	0.7000	0.03922
σ	0.4763	0.4715	0.03711

* where the standard error refers to the unconstrained value.

Empirically, the restricted maximum likelihood estimates were found to converge to unique solutions, which is consistent with the global convergence property of the GP algorithm. We noted the property held for missing and clustered data, for all subsequent simulations and numerical examples. Additional examples of the implementation of the GP algorithm for GLMMs are presented in Section 6.1 in connection with the Youth Smoking Survey 2002 analysis.

5.4 Constrained Hypothesis Tests for GLMMs

Once maximum likelihood estimates have been obtained for the parameters of interest, constrained hypothesis testing may be performed. As in the generalized linear model, we consider the constrained set $\Omega = \{\boldsymbol{\beta} : A\boldsymbol{\beta} \leq \mathbf{c}\}$, with A an $r \times p$ matrix of full rank, and define the hypotheses

$$H_0 : A\boldsymbol{\beta} = \mathbf{c}, \quad H_1 : A\boldsymbol{\beta} \leq \mathbf{c}, \quad H_2 : \text{ no restriction on } \boldsymbol{\beta}, \tag{5.8}$$

with A and $\mathbf{c}$ as defined previously.

Based on the maximum likelihood estimators, $\boldsymbol{\gamma}^0$, $\boldsymbol{\gamma}^*$ and $\hat{\boldsymbol{\gamma}}$ under H_0, H_1, and H_2 respectively, we may construct the likelihood ratio tests in the same manner as the previous chapter. That is, the unconstrained test rejects H_0 in favor of $H_2 - H_0$ if $T_{02} = 2[l^*(\hat{\boldsymbol{\gamma}}|\boldsymbol{y}) - l^*(\boldsymbol{\gamma}^0|\boldsymbol{y})]$ is large where $l^*(\boldsymbol{\gamma}|\boldsymbol{y})$ is the logarithm of the marginal likelihood as defined by (5.3). As is well known, T_{02} asymptotically follows $\chi^2(r)$ under H_0. However, if the parameter space is restricted by H_1, then we test H_0 against $H_1 - H_0$ using the statistic $T_{01} = 2[l^*(\boldsymbol{\gamma}^*|\boldsymbol{y}) - l^*(\boldsymbol{\gamma}^0|\boldsymbol{y})]$. Since the test based on T_{01} is meaningful only when H_1 is true, we may need to confirm H_1 by a goodness-of-fit test which rejects H_1 for large values of $T_{12} = 2[l^*(\hat{\boldsymbol{\gamma}}|\boldsymbol{y}) - l^*(\boldsymbol{\gamma}^*|\boldsymbol{y})]$.

Once the restricted hypothesis H_1 is included in the testing procedures, the

asymptotic null distribution of the likelihood ratio test statistic is no longer of a chi-square type. Instead, the distribution is a mixture of chi-square distributions. In the following sections, we derive a theorem which details the form of the asymptotic distributions, and outlines the required regularity assumptions. We note that such a theorem holds only in the asymptotic sense. For small samples, empirical results of other authors such as Fitzmaurice et. al. (2007) have demonstrated lower sizes and power for the likelihood ratio or score tests. Nevertheless, the differences vanish once the sample size increases. Hence, a bootstrap procedures analogous to Sinha (2009) may be used for small samples, whereas the chi-bar square distribution may be implemented when the sample size is moderately large.

5.4.1 Derivation of Asymptotic Results under Inequality Constraints

We next present the following main result for constrained likelihood ratio tests in generalized linear mixed models. The underlying models are assumed to satisfy regularity conditions which will be presented in the following section.

Theorem 5.1. Under appropriate regularity assumptions, the asymptotic null distributions of likelihood ratio test statistics T_{01} and T_{12}, are given as follows:

$$\lim_{k\to\infty} P_{\boldsymbol{\gamma}_0}[T_{01} > x] = \sum_{i=0}^{r} w_i(r, AV(\boldsymbol{\gamma}_0)A^T)P[\chi_i^2 > x], \tag{5.9}$$

$$\lim_{k\to\infty} P_{\boldsymbol{\gamma}_0}[T_{12} > x] = \sum_{i=0}^{r} w_{r-i}(r, AV(\boldsymbol{\gamma}_0)A^T)P[\chi_i^2 > x] \tag{5.10}$$

for any $x \geq 0$. Here, r is the rank of A, $\boldsymbol{\gamma}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\theta}_0^T)^T$ is a value of $\boldsymbol{\gamma}$ under H_0, and $V(\boldsymbol{\gamma}_0)$ is the limit of $k[\mathcal{I}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0) - \mathcal{I}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)\mathcal{I}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\mathcal{I}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)]^{-1}$ where $\mathcal{I}(\cdot,\cdot) = E[\mathcal{I}_o(\cdot,\cdot)]$, and $\mathcal{I}_o(\cdot,\cdot)$ are elements of the partitioned observed information $\mathcal{I}_o$ defined in (5.7).

The null distributions in Theorem 5.1 depend on the unknown parameter vector $\boldsymbol{\gamma}_0$, which may be approximated by replacing the parameter vector by its estimate. As was mentioned earlier, we may use $\boldsymbol{\gamma}^*$ for T_{01} and $\hat{\boldsymbol{\gamma}}$ for T_{12}, analogous to a Wald statistic.

The calculation of the chi-bar-square weights needed in (5.9) and (5.10) may be found in a similar manner as that described in Section 4.5.2. The simulation algorithm may be adapted from the GLM case with missing values by simply replacing the matrix $\boldsymbol{D}$ by the variance matrix in Theorem 5.1. The R program to compute these weights is provided in the Appendix.

5.4.2 Proof of Theorem 5.1

Let $g'(y_i, \gamma) = (g_1'(y_i, \gamma)^T, g_2'(y_i, \gamma)^T)^T$ denote the derivative of $g(y_i, \gamma)$ with respect to γ where $g_1'(y_i, \gamma)$ and $g_2'(y_i, \gamma)$ are the negative values of the i-th terms in (5.4) and (5.5), respectively. Since the estimating equations are

$$\sum_{i=1}^{k} g'(y_i, \gamma) = \mathbf{0},$$

the unconstrained maximum likelihood estimator, $\hat{\gamma} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}$, of γ is the same as the one that minimizes, with respect to γ,

$$G_k(\boldsymbol{y}, \gamma) = k^{-1} \sum_{i=1}^{k} g(y_i, \gamma).$$

Further, $l^*(\gamma \mid \boldsymbol{y}) = -kG_k(\boldsymbol{y}, \gamma) + d(\boldsymbol{y})$ for some $d(\boldsymbol{y})$ which does not depend on γ.

We require the following regularity assumptions for further theoretical development. Note that γ_0 refers to the true parameter value when H_0 is true.

1. The sequence $g'(\boldsymbol{y}_i, \gamma)$, $i = 1, \ldots, k$ are continuous functions of γ for each $\boldsymbol{y}_i$, and measurable functions of $\boldsymbol{y}_i$ for each $\gamma \in \Theta$, a compact subset of a finite-dimensional Euclidean space.

2. The random components $\{\boldsymbol{y}_i\}$ are either (a) ϕ-mixing with $\phi(m)$ of size $r_1/(2r_1 - 1)$, $r \geq 1$; or (b) α-mixing with $\alpha(m)$ of size $r_1/(r_1 - 1)$, $r_1 > 1$.

3. The sequence $g(\boldsymbol{y}_i, \boldsymbol{\gamma})$ is dominated by uniformly $(r_1 + \delta)$-integrable functions, $r_1 \geq 1$, $0 < \delta \leq r_1$.

4. The function $\bar{G}_k(\boldsymbol{\beta}) \equiv E[G_k(\boldsymbol{y}, \boldsymbol{\gamma})]$ has an identifiably unique minimizer $\boldsymbol{\gamma}_0$ (in the sense of Definition 2.1 of Domowitz and White, 1982).

5. The functions $g(\boldsymbol{y}_i, \boldsymbol{\gamma})$ are twice continuously differentiable in $\boldsymbol{\gamma}$, uniformly in i, a.s. - P.

6. For $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p+q})^{\mathrm{T}}$, $\{g'_j(\boldsymbol{y}_i, \boldsymbol{\gamma})^2\}$, $j = 1, \ldots, p+q$ are dominated by uniformly r_2-integrable functions, $r_2 > 1$, where $g'_j(\boldsymbol{y}_i, \boldsymbol{\gamma}) = \partial g(\boldsymbol{y}_i, \boldsymbol{\gamma})/\partial \gamma_j$.

7. Define $\boldsymbol{Q}_{a,k} = \mathrm{var}[k^{-1/2} \sum_{i=a+1}^{a+k} g'(\boldsymbol{y}_i, \boldsymbol{\gamma}_0)]$. Assume that there exists a positive definite matrix $\boldsymbol{Q}$ such that $\boldsymbol{\lambda}' \boldsymbol{Q}_{a,k} \boldsymbol{\lambda} - \boldsymbol{\lambda}' \boldsymbol{Q} \boldsymbol{\lambda} \to \mathbf{0}$ as $k \to \infty$, uniformly in a, for any nonzero vector $\boldsymbol{\lambda}$.

8. For $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p+q})'$, $\{g''_{jl}(\boldsymbol{y}_i, \boldsymbol{\gamma})\}$, $j, l = 1, \ldots, p+q$ are dominated by uniformly $(r_1 + \delta)$-integrable functions, $0 < \delta \leq r_1$, where $g''_{jl}(\boldsymbol{y}_i, \boldsymbol{\gamma}) = \partial^2 g(\boldsymbol{y}_i, \boldsymbol{\gamma})/\partial \gamma_j \partial \gamma_l$.

9. The matrix $\bar{G}''_k(\boldsymbol{\gamma}) = k^{-1} \sum_{i=1}^{k} E[g''(\boldsymbol{y}_i, \boldsymbol{\gamma})]$ has constant rank $p+q$ in some ϵ-neighbourhood of $\boldsymbol{\gamma}_0$, for all sufficiently large k, uniformly in k.

Interpretations of the previous regularity assumptions are similar to those outlined in Section 4.5.3. We also require the following strengthening of Assumption 2 to prove a version of the central limit theorem with dependent observations:

2a. Assumption 2 holds and either $\phi(m)$ is of size $r_2/(r_2-1)$ or $\alpha(m)$ is of size $\max[r_1/(r_1-1), r_2/(r_2-1)]$, $r_1, r_2 > 1$.

Then, let $G'_k(\boldsymbol{y}, \boldsymbol{\gamma})$ and $G''_k(\boldsymbol{\gamma})$ denote the first and second derivatives of $G_k(\boldsymbol{y}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$, respectively. Since $G'_k(\boldsymbol{y}, \boldsymbol{\gamma})$ is the sum of independent stochastic terms, the multivariate central limit theorem holds under the regularity conditions given previously. These conditions are required to ensure $G_k(\boldsymbol{y}, \boldsymbol{\gamma})$, $G'_k(\boldsymbol{y}, \boldsymbol{\gamma})$, and $G''_k(\boldsymbol{y}, \boldsymbol{\gamma})$ converge almost surely to the limits of $\bar{G}_k(\boldsymbol{\gamma})$, $\bar{G}'_k(\boldsymbol{\gamma})$ and $\bar{G}''_k(\boldsymbol{\gamma})$, respectively. If $\boldsymbol{\gamma}_0$ is the true model parameter value, then $\bar{G}''_k(\boldsymbol{\gamma}_0) = \frac{1}{k}\mathcal{I}(\boldsymbol{\gamma}_0)$ where $\mathcal{I}(\boldsymbol{\gamma}_0) \equiv E[\mathcal{I}_o(\boldsymbol{\gamma}_0)]$. Also, we have $\bar{G}'_k(\boldsymbol{\gamma}_0) = \mathbf{0}$ since $\bar{G}_k(\cdot)$ assumes its minimum value at the true value of model parameter vector. The following lemma summarizes the large sample properties of $\hat{\boldsymbol{\gamma}}$.

Lemma 2. Let $\boldsymbol{\gamma}_0$ denote the true model parameter value. Then, under the regularity assumptions, we have that

(a) $\hat{\boldsymbol{\gamma}} \rightarrow \boldsymbol{\gamma}_0$ almost surely as $k \rightarrow \infty$,

(b) $\mid G_k(\boldsymbol{y}, \hat{\boldsymbol{\gamma}}) - \bar{G}_k(\boldsymbol{\gamma}_0) \mid \rightarrow \mathbf{0}$ almost surely as $k \rightarrow \infty$,

(c) $\sqrt{k} G_k'(\boldsymbol{y}, \hat{\boldsymbol{\gamma}}) \rightarrow N(\mathbf{0}, R(\boldsymbol{\gamma}_0))$ where $R(\boldsymbol{\gamma}_0) = \lim_{k \rightarrow \infty} \frac{1}{k} \mathcal{I}(\boldsymbol{\gamma}_0)$,

(d) $\sqrt{k}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \rightarrow N(\mathbf{0}, R(\boldsymbol{\gamma}_0)^{-1})$.

Proof of Theorem 5.1. We first prove the theorem for the case where the right hand side of the inequality constraint in (5.8) is $\mathbf{c} = \mathbf{0}$. Since $G_k'(\boldsymbol{y}, \hat{\boldsymbol{\gamma}}) = \mathbf{0}$, the second order Taylor expansion of $G_k(\boldsymbol{y}, \boldsymbol{\gamma})$ about $\hat{\boldsymbol{\gamma}}$ gives

$$G_k(\boldsymbol{y}, \boldsymbol{\gamma}) = G_k(\boldsymbol{y}, \hat{\boldsymbol{\gamma}}) + \frac{1}{2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T G_k''(\boldsymbol{y}, \bar{\boldsymbol{\gamma}})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}); \tag{5.11}$$

where $\bar{\boldsymbol{\gamma}}$ is a point on the segment between $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}$. By (a) of Lemma 2, $\bar{\boldsymbol{\gamma}}$ also converges to $\boldsymbol{\gamma}_0$ almost surely, and hence it follows that $G_k''(\boldsymbol{y}, \bar{\boldsymbol{\gamma}}) \rightarrow R(\boldsymbol{\gamma}_0)$ and $G_k''(\boldsymbol{y}, \bar{\boldsymbol{\gamma}}) - G_k''(\boldsymbol{y}, \hat{\boldsymbol{\gamma}}) \rightarrow \mathbf{0}$ (null matrix) almost surely.

Let $Q(\boldsymbol{y}, \boldsymbol{\gamma}) = k(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T G_k''(\boldsymbol{y}, \bar{\boldsymbol{\gamma}})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ and $\dot{Q}(\boldsymbol{y}, \boldsymbol{\gamma}) = k(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T G_k''(\boldsymbol{y}, \hat{\boldsymbol{\gamma}})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$. Then, we have that $\mid Q(\boldsymbol{y}, \boldsymbol{\gamma}) - \dot{Q}(\boldsymbol{y}, \boldsymbol{\gamma}) \mid = o_p(1)$. Since $Q(\boldsymbol{y}, \boldsymbol{\gamma}) = 2[l^*(\hat{\boldsymbol{\gamma}} \mid \boldsymbol{y}) - l^*(\boldsymbol{\gamma} \mid \boldsymbol{y})]$ by (5.11), the likelihood ratio test statistics may be expressed as

$$T_{01} = Q(\boldsymbol{y}, \boldsymbol{\gamma}^0) - Q(\boldsymbol{y}, \boldsymbol{\gamma}^*) \quad \text{and} \quad T_{12} = Q(\boldsymbol{y}, \boldsymbol{\gamma}^*) \tag{5.12}$$

where $\boldsymbol{\gamma}^0$ and $\boldsymbol{\gamma}^*$ are estimators of $\boldsymbol{\gamma}$ under H_0 and H_1, respectively.

Next let A^* be the $r \times (p+q)$ matrix obtained by augmenting the matrix A by q columns of zeros. The parameter spaces under H_0 and H_1 may be redefined as $C_0 = \{\boldsymbol{\gamma} \in \mathbb{R}^{p+q} : A^*\boldsymbol{\gamma} = \mathbf{c}\}$ and $C_1 = \{\boldsymbol{\gamma} \in \mathbb{R}^{p+q} : A^*\boldsymbol{\gamma} \leq \mathbf{c}\}$. Then, the least squares projections $P_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_0)$ and $P_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_1)$ are the constrained estimators obtained by minimizing $\dot{Q}(\boldsymbol{y},\boldsymbol{\gamma}) \equiv \| \sqrt{k}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \|^2_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}$ under H_0 and H_1, respectively. Thus, we may approximate T_{01} in (5.12) with $o_p(1)$ error as

$$\begin{aligned} T_{01} &= \| \sqrt{k}(\hat{\boldsymbol{\gamma}} - P_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_0)) \|^2_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})} \\ &\quad - \| \sqrt{k}(\hat{\boldsymbol{\gamma}} - P_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_1)) \|^2_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})} + o_p(1) \\ &= \| \sqrt{k}(P_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_1) - P_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_0)) \|^2_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})} + o_p(1). \end{aligned}$$

From the property of least squares projection, we have for any $\boldsymbol{\gamma}_0 \in C_0$ that

$$\sqrt{k}(P_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}(\hat{\boldsymbol{\gamma}} \mid C_i) - \boldsymbol{\gamma}_0) = P_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}(\sqrt{k}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \mid C_i), \ i = 0, 1.$$

Thus, if $\boldsymbol{\gamma}_0$ is the underlying model parameter under H_0, it follows by Lemma 2 and the continuity property of the projection operator that

$$\begin{aligned} T_{01} &= \| P_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}(\sqrt{k}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \mid C_1) - P_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})}(\sqrt{k}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \mid C_0) \|^2_{G_k''(\boldsymbol{y},\hat{\boldsymbol{\gamma}})} + o_p(1) \\ &\to^D \| P_{R(\boldsymbol{\gamma}_0)}(Z \mid C_1) - P_{R(\boldsymbol{\gamma}_0)}(Z \mid C_0) \|^2_{R(\boldsymbol{\gamma}_0)} \end{aligned}$$

where $Z \sim N(\mathbf{0}, R(\boldsymbol{\gamma}_0)^{-1})$. From Theorem 3.5, the asymptotic null distribution is chi-bar-square with weights $w_i(r, A^*R(\boldsymbol{\gamma}_0)^{-1}A^{*T})$. Since A^* is the matrix formed by augmenting the matrix A by columns of zeros, we can express

$A^* R(\gamma_0)^{-1} A^{*T}$ in terms of A and the submatrix of $R(\gamma_0)^{-1}$ corresponding to β and achieve the desired result, as $R(\gamma_0) = \lim_{k\to\infty} \frac{1}{k}\mathcal{I}(\gamma_0)$ by Lemma 2 (c).

The asymptotic null distributions of T_{12} under H_1 may also be derived similarly. Applying large sampling approximation to T_{12} in (5.12), we have

$$T_{12} \ \to^D \ \| \ Z - P_{R(\gamma_0)}(Z \mid C_1) \ \|^2_{R(\gamma_0)}$$

for $\gamma_0 \in C_0$. The asymptotic distribution of T_{12} at $\gamma = \gamma_0 \in H_0$ is immediately obtained by Theorem 3.5.

In the case where $\mathbf{c} \neq \mathbf{0}$, we may apply the identical argument as outlined in the proof of Theorem 4.3, and the result follows. □.

5.5 Empirical Results for Constrained GLMMs

5.5.1 Simulation Study

A simulation study was conducted to assess the bias and mean square error (MSE) of the constrained and unconstrained estimates, for generalized linear mixed models using the aforementioned methods. The constraints of interest and associated matrix are as follows:

$$\begin{array}{rcrcl} \beta_0 & + & \beta_1 & \leq & c_1 \\ \beta_0 & - & \beta_1 & \geq & -c_2 \end{array} \quad \Longrightarrow \quad A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

Both Bernoulli and Poisson models for the GLMM are considered. We assumed $u_i \sim$ ind. $N(0, \sigma^2)$. In the Bernoulli GLMM, $\sigma^2 = 1$, and $k = 200$ clusters each of size $n_i = 4$ were generated from Bernoulli(π_{ij}) with $\pi_{ij} = \frac{\exp(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}+u_i)}{1+\exp(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}+u_i)}$. For the Poisson GLMM, $k = 200$ clusters each with size $n_i = 4$ were generated from Poisson (λ_{ij}) where $\lambda_{ij} = \exp(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + u_i)$ and $\sigma^2 = 0.25$.

Also, $X \sim U(a_1, a_2)$, where for Bernoulli we have $a_1 = -4.5$, $a_2 = -2$ and for Poisson, $a_1 = 1$, $a_2 = 5$. In addition, for Bernoulli we consider $c_1 = 5$, $c_2 = -2$ while for Poisson, $c_1 = 1$, $c_2 = 0.5$. Estimates were calculated over $N = 1000$ replications.

The unconstrained estimation was performed using the software package R and associated function `glmmML()`. For each simulation, the empirical bias and MSE were calculated for both the constrained and unconstrained cases. In the constrained setting, three constraint cases were considered: (a) those beta values on the vertex point (e.g. $\boldsymbol{\beta} = (3.50, 1.50)^{\mathrm{T}}$), (b) values on the boundary of the constraints or (c) values within the constraint cone. The results are displayed in Tables 5.4 and 5.5, respectively.

Table 5.4 - Bias and Mean Squared Error for Unconstrained and Constrained MLE for Binary Mixed Models

		Constrained		Unconstrained	
Case	**Parameter**	Bias	MSE	Bias	MSE
(a)	$\beta_0 = 3.50$	-0.0102	0.0016	-0.0554	0.3509
	$\beta_1 = 1.50$	-0.0216	0.0017	-0.0261	0.0849
	$\sigma^2 = 1$	0.0009	0.0036	-0.0317	0.0699
(b)	$\beta_0 = 2.5$	0.1309	0.0687	-0.0450	0.1555
	$\beta_1 = 0.5$	0.0321	0.0062	-0.0167	0.0134
	$\sigma^2 = 1$	0.0256	0.0387	-0.0195	0.0811
(b)	$\beta_0 = 4.00$	-0.2118	0.1202	-0.0684	0.1989
	$\beta_1 = 1.00$	-0.0613	0.0103	-0.0201	0.0169
	$\sigma^2 = 1$	-0.0561	0.0457	-0.0259	0.0828
(c)	$\beta_0 = 2.65$	0.1224	0.0382	-0.0558	0.1513
	$\beta_1 = 0.60$	0.0317	0.0052	-0.0172	0.0126
	$\sigma^2 = 1$	-0.0587	0.0382	-0.1735	0.0911
(c)	$\beta_0 = 4.00$	-0.2184	0.1296	-0.0677	0.2299
	$\beta_1 = 0.90$	-0.0629	0.0107	-0.0208	0.0182
	$\sigma^2 = 1$	-0.0627	0.0593	-0.0271	0.0906

Table 5.5 - Bias and Mean Squared Error for Unconstrained and Constrained Poisson Mixed Models

Case	Parameter	Constrained		Unconstrained	
		Bias	MSE	Bias	MSE
(a)	$\beta_0 = 0.25$	0.0401	0.0028	0.0541	0.0041
	$\beta_1 = 0.75$	-0.0682	0.0071	-0.0689	0.0073
	$\sigma^2 = 0.25$	0.0065	0.0008	0.0056	0.0011
(b)	$\beta_0 = 2.00$	-0.0397	0.0033	0.0681	0.0158
	$\beta_1 = -1.00$	0.0335	0.0023	-0.0074	0.0029
	$\sigma^2 = 0.25$	0.0101	0.0010	0.0485	0.0050
(b)	$\beta_0 = -0.75$	0.1433	0.0367	0.1135	0.0417
	$\beta_1 = -0.25$	-0.0349	0.0037	-0.0291	0.0042
	$\sigma^2 = 0.25$	0.0315	0.0105	0.0569	0.0160
(c)	$\beta_0 = -0.65$	0.1276	0.0329	0.0987	0.0384
	$\beta_1 = -0.25$	-0.0288	0.0030	-0.0216	0.0036
	$\sigma^2 = 0.25$	0.0386	0.0096	0.0530	0.0125
(c)	$\beta_0 = -0.55$	0.1169	0.0326	0.1064	0.0352
	$\beta_1 = -0.25$	-0.0285	0.0032	-0.0257	0.0035
	$\sigma^2 = 0.25$	0.0488	0.0106	0.0527	0.0112

From these results, it is evident that in many cases, the constrained estimates have larger bias than their unconstrained counterparts. However, the MSE for the constrained estimates is considerably less for all cases under study. In particular, when a given value of beta moves closer to the boundary points, the bias increases and MSE decreases for the constrained over the unconstrained estimates. Moreover, when comparing the two GLMM cases, we note a similar pattern of smaller MSE in the variance component analysis, despite having no

a priori constraints. Hence, estimation of the variance component is affected by the constraints on the regression parameters. These results are consistent with empirical results from constrained linear models.

5.5.2 LRT Power Comparisons

We next consider powers of constrained and unconstrained likelihood ratio tests for two discrete probability distributions: Bernoulli and Poisson. The constraints under study are

$$\begin{array}{rcccccc} \beta_0 & + & \beta_1 & + & \beta_2 & \leq & c_1 \\ \beta_0 & - & \beta_1 & + & \beta_2 & \geq & -c_2 \end{array} \implies A = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \end{bmatrix}, \quad c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

We assume the random effects $\boldsymbol{u}_i$ are independent $N(0, \sigma^2)$, for $i = 1, \ldots, k$ clusters with $\sigma^2 = 1.0$ for Bernoulli and $\sigma^2 = 0.25$ for Poisson. We generated $k = 100$, $k = 200$, and $k = 300$ clusters each of size $n_i = 4$ from Bernoulli(π_{ij}), where $\pi_{ij} = (\exp(\mathbf{x}_{ij}^{\mathrm{T}}\beta + u_i))/(1 + \exp(\mathbf{x}_{ij}^{\mathrm{T}}\beta + u_i))$, and from Poisson($\lambda_{ij}$) where $\lambda_{ij} = \exp(\mathbf{x}_{ij}^{\mathrm{T}}\beta + u_i)$. The design matrix consists of vectors $\mathbf{x}_{ij}^{\mathrm{T}} = (1, x_{1ij}, x_{2ij})^{\mathrm{T}}$. For the Bernoulli model, we considered $x_{1ij} \sim U(-1, 1)$, and $x_{2ij} = x_{1ij} + r_{1ij}$, where r_{1ij} is a number generated randomly from N(0, 0.25^2). We chose constraints, c_1 = 0.1 and $c_2 = 0.2$. For the Poisson model, we generated $x_{1ij} \sim U(-5, 5)$ and $x_{2ij} = x_{1ij} + r_{2ij}$ where r_{2ij} is randomly generated from N(0, 0.5^2). The selected

constraints were $c_1 = c_2 = 0.3$.

We generated $N = 1000$ replicates of data sets for each combination of the parameter values considered in the two binary and Poisson regression models. We considered (a) values of $\boldsymbol{\beta}$ under H_0, where both constraints are active and (b) values under H_1, where at least one constraint is inactive. In Tables 5.6 and 5.7, percentages of rejecting the null hypothesis in 1000 replications are reported at the 5% significance level.

Table 5.6 - Power comparisons (%) for likelihood ratio tests for Binary Mixed Models with 5% significance level

		$k = 100$			$k = 200$			$k = 300$		
Case	$\boldsymbol{\beta}^{\mathrm{T}}$	T_{01}	T_{02}	T_{12}	T_{01}	T_{02}	T_{12}	T_{01}	T_{02}	T_{12}
	(0.050, 0.150, -0.100)	4.0	8.5	10.8	6.9	6.4	5.8	5.6	5.6	6.0
(a)	(0.100, 0.150, -0.150)	3.8	9.1	9.8	4.9	6.0	5.7	4.9	5.1	6.3
	(-0.075, 0.150, 0.025)	7.2	6.9	6.0	6.9	6.9	5.2	6.1	5.9	5.0
	(0.150, -0.200, 0.150)	17.9	13.5	3.1	21.7	15.4	1.5	37.2	27.9	1.0
(b)	(-0.050, 0.050, -0.100)	21.7	15.9	2.8	22.4	20.1	5.7	40.7	30.6	3.7
	(0.150, -0.350, 0.250)	18.6	15.9	3.9	41.3	33.8	1.9	56.4	45.8	0.5

Table 5.7 - Power comparisons (%) for likelihood ratio tests for Poisson Mixed Models with 5% significance level

		$k = 100$			$k = 200$			$k = 300$		
Case	β^T	T_{01}	T_{02}	T_{12}	T_{01}	T_{02}	T_{12}	T_{01}	T_{02}	T_{12}
	(0.350, 0.300, -0.350)	6.8	6.7	6.0	5.3	7.8	6.4	4.6	4.3	4.0
(a)	(0.400, 0.300, -0.400)	3.1	5.8	8.6	4.1	3.0	5.0	4.7	4.7	4.8
	(0.450, 0.300, -0.450)	6.6	6.5	5.5	4.7	5.6	5.9	4.1	4.2	4.7
	(0.300, 0.200, -0.200)	12.4	11.8	5.1	46.1	44.8	4.3	64.2	51.5	0.2
(b)	(0.150, 0.250, -0.200)	19.7	16.5	4.5	48.8	40.7	3.4	86.3	80.2	2.7
	(0.300, 0.200, -0.300)	40.8	29.0	2.0	70.0	58.1	0.8	96.2	92.4	0.6

For case (a), we expect the empirical sizes to converge to the nominal level of 5% which occurs as the number of clusters increases from $k = 100$ to $k = 300$. For some parameter values, the sizes of T_{12} for $k = 100$ are slightly larger than the nominal level, however improvements are noted as the number of clusters increases. Case (b) describes powers of the tests when the first or second constraint is inactive, or when both constraints are inactive. In this setting, the values corresponding to T_{01} and T_{02} are empirical powers, with T_{01} displaying improved performance, even in the $k = 100$ situation. The results are consistent with the fact that T_{01} incorporates information pertaining to the constrained parameter space in the hypothesis test. Further, in case (b), the values of T_{12} represent sizes as opposed to powers since the values of $\boldsymbol{\beta}$ are within H_1. The

values are smaller than the nominal level of 5%, since the particular null value is not the least favourable value. Hence for cases (a) and (b) combined, sizes of T_{12} are within reason. The proposed constrained test T_{01} exhibits reasonable empirical size when data are observed with a moderate number of clusters, and significantly better power performance than T_{02} when the constraints are satisfied.

Chapter 6

Applications

In this chapter, we revisit the two applications discussed in the Introduction: the Canadian Youth Smoking Survey 2002 and the Northern Contaminants Programme. A more detailed explanation of each dataset is provided, including descriptive statistics of important variables. Subsequently, implementation of the constrained inferential methods are described, and results are summarized.

6.1 Canadian Youth Smoking Survey

6.1.1 Description of the Data

To illustrate the preceding unconstrained and constrained methods, we consider an analysis from the Youth Smoking Survey (YSS) 2002. Conducted by Statistics Canada on behalf of Health Canada, the main objective of the YSS was to provide up-to-date information on the smoking behaviour of students in grades 5 to 9 (in Quebec primary school grades 5 and 6 and secondary school grades 1 to 3). The target population consisted of all young Canadian residents attending private or public schools in grades 5 to 9 inclusively. Specifically excluded from the survey's coverage are residents of the Yukon, Northwest Territories and Nunavut, persons living on Indian Reserves, inmates of institutions, students attending special schools (e.g. schools for the blind) and other geographically inaccessible areas.

The YSS was a voluntary, cross-sectional survey administered by sampling schools containing grades 5 to 9 and then subsampling students within each school. The final sample included 19,018 students selected from 982 schools across Canada. Initial results were released by Statistics Canada in June, 2004. Parents of sampled students were also administered a questionnaire pertaining to the parents' smoking status and other household characteristics, such as highest

level of education, household income, etc.

For the purposes of model fitting, we may consider each school as a cluster and perform a generalized linear mixed model analysis. Hence, students within the same school are assumed to exhibit similar smoking habits and these habits are independent from those students in other schools. In particular, interest lies in determining the effects of certain variables on the binary smoking status (Ever Smoker or Never Smoker). According to the survey, an "Ever smoker" is defined as either a current or former smoker who has tried smoking, even just a few puffs; while a "Never smoker" defines a youth who has never tried a cigarette, even just a few puffs.

An assessment of important covariates and their effect on smoking status using unweighted counts is given in Table 6.1. The number of missing values and relative percentage per covariate is also given. Table 6.1 demonstrates that approximately 20% of the students had at least one missing covariate, while the smoking status response was fully observed. The student's age and household income have directional effects on the probability of being a smoker. In particular, as the age increases, the students were more likely to be a smoker; while as the household income increases, the students were less likely to be a smoker. The YSS Technical Report (Health Canada, 2005) did not include an analysis

of the child's age and its effect on smoking status. Thus, the present analysis improves upon previously conducted governmental research.

Table 6.1. Summary Statistics for Youth Smoking Survey 2002

		Response:	Smoking Status
Covariate	**Level**	**Never Smoker**	**Ever Smoker**
Age	16	50.70%	49.30%
	15	78.20%	21.80%
	14	91.72%	8.28%
	13	96.67%	3.33%
	12	98.84%	1.16%
	11 or less	$\geq$99.4%	$\leq$0.6%
Sex	Female	96.55%	3.45%
	Male	96.43%	3.57%
Aboriginal	Yes	91.24%	8.76%
	No	96.80%	3.20%
Household Member(s)	Yes	92.31%	7.69%
Smokes	No	97.99%	2.01%
Parent's Highest	High School or less	94.70%	5.30%
Education	College	97.27%	2.73%
	University	98.37%	1.63%
Friends Smoke	Yes	87.57%	12.43%
	No	99.74%	0.26%
Household Income	Less than $45,000	94.48%	5.52%
	$45,000 - $ 59,999	97.18%	2.82%
	$60,000 - $79,999	97.75%	2.25%
	$80,000 or more	98.52%	1.48%
No. Extracurricular		Mean = 2.21	Mean = 1.56
Activities		SD = 1.22	SD = 1.21
		Range = (0,4)	Range = (0,4)
Total	**Complete Cases**	**96.49% (14567)**	**3.51% (530)**
	Missing Response	***0% (0)***	***0% (0)***
	Missing Covariates	***20.0% (3765)***	***22.7% (156)***

6.1.2 Data Analysis

As no current methods exist to analyze constrained generalized linear mixed models with survey weights representing the multi-stage sampling design, we assume the design is non-informative and the sample drawn is representative of the target population of Canadian youth in grades 5 to 9. We analyze only the complete cases for all preceding covariates from which the sample size reduces from 19,018 in 982 schools to 15,097 complete cases in 768 schools. The binary response variable of interest y_{ij} is 1 of the jth student in the ith school is identified as an "ever smoker" and 0 otherwise. To describe the effects of covariates on the response variable, we consider the logistic model

$$\begin{aligned}\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_0 + \beta_1 AGE16_{ij} + \beta_2 AGE15_{ij} + \beta_3 AGE14_{ij} + \beta_4 AGE13_{ij} \\ &\quad + \beta_5 AGE12_{ij} + \beta_6 ABOR_{ij} + \beta_7 SEX_{ij} + \beta_8 ACT_{ij} \\ &\quad + \beta_9 HSMOK_{ij} + \beta_{10} INC1_{ij} + \beta_{11} INC2_{ij} + \beta_{12} INC3_{ij} \\ &\quad + \beta_{13} FSMOK_{ij} + u_i; \qquad (6.1)\end{aligned}$$

where the binary covariate $AGE\ell$ is 1 if the student is ℓ years old and 0 otherwise, for $\ell = 12, 13, 14, 15, 16$; $ABOR$ is 1 if the student is aboriginal and 0 otherwise; SEX is 1 if the student is female and 0 otherwise. Also, ACT is a quantitative variable indicating the number of extracurricular activities for which the student

is involved; $HSMOK$ is 1 if at least one member of the household is a smoker and 0 otherwise; $INC1$ is 1 if the total household income is less than \$45,000 and 0 otherwise; $INC2$ has a value of 1 if the total household income is between \$45,000 and \$59,999 and 0 otherwise; $INC3$ has a value of 1 if the total household income is between \$60,000 and \$79,999 and 0 otherwise; $FSMOK$ is 1 if the student has at least one close friend who is a smoker and 0 otherwise; u_i is the random effect, which represents the school.

The estimates of the parameters in (6.1) were obtained using the R function `glmmML()`, and are presented in Table 6.2. The unconstrained estimates are in fact the constrained estimates relative to the parameter space $C_1 = \{\boldsymbol{\gamma} : \beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4 \geq \beta_5\}$. We assume the school effects u_i are independently distributed $N(0, \sigma^2)$ and $\boldsymbol{\gamma} = [\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2]^{\mathrm{T}}$. A likelihood ratio test of the random effects variance results in a p-value of $0.5\mathrm{pr}(\chi^2_1 \geq 16.81) = 0.00002$. Thus we reject the null hypothesis of $H_0 : \sigma^2 = 0$ versus $H_1 : \sigma^2 > 0$, and conclude the school effect is significant.

We simultaneously constrain the age groups to have an increasing effect on the smoking probability and income levels to have a non-increasing effect. More precisely, under H_0, $C_{Both0} = \{\boldsymbol{\gamma} : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5, \beta_{10} = \beta_{11} = \beta_{12}\}$ while under H_1, $C_{Both1} = \{\boldsymbol{\gamma} : \beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4 \geq \beta_5, \beta_{10} \geq \beta_{11} = \beta_{12}\}$.

Constrained estimates for all models were found using the algorithm described in Section 5.3, and are presented in Table 6.2. The gradient projection algorithm converged within two iterations for C_{Both1} and four iterations for C_{Both0}. If we rewrite C_1 in the form of pairwise contrasts $\beta_l - \beta_m$, the constraint matrix becomes

$$A_{Both} = \begin{bmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \end{bmatrix}. \quad (6.2)$$

The hypotheses of interest are $H_0 : A_{Both}\boldsymbol{\gamma} = \mathbf{c}$, $H_1 : A_{Both}\boldsymbol{\gamma} \leq \mathbf{c}$ and $H_2 : \boldsymbol{\gamma}$ unconstrained, with $r = 6$, and $\mathbf{c} = \mathbf{0}$, of length r. The goodness-of-fit test statistic is $t_{12}^B = 0.26$, with p-value = 0.907 from (5.10), implying the constraints are significant. Since the constrained parameter space involves an equality, chi-bar-square weights used to calculate the p-value must be modified slightly to avoid redundancy. We write $C_{Both1} = \{\boldsymbol{\gamma} : A_{Both,1}\boldsymbol{\gamma} \leq 0, A_{Both,2}\boldsymbol{\gamma} = \mathbf{0}\}$ where $A_{Both,1}$ represents the inequality constraints with dimension $s = 5$, i.e. the first five rows of A_{Both}. Also, $A_{Both,2}$ represents the equality constraint with dimension

$t = 1$, from the last row of A_{Both}, and in total $r = s + t$. From Theorem 3.6 equation 8, the adjusted weight vector is found with $N = 50,000$ simulations to be

$$w^B(s, V^B_{new}(\hat{\gamma})) = (0.0630, 0.2328, 0.3430, 0.2561, 0.0925, 0.0126)^{\mathrm{T}},$$

with adjusted variance-covariance matrix

$$\begin{aligned} V^B_{new} &= A_{Both,1} V(\hat{\gamma}) A^{\mathrm{T}}_{Both,1} \\ &\quad -(A_{Both,1} V(\hat{\gamma}) A^{\mathrm{T}}_{Both,2})(A_{Both,2} V(\hat{\gamma}) A^{\mathrm{T}}_{Both,2})^{-1}(A_{Both,2} V(\hat{\gamma}) A^{\mathrm{T}}_{Both,1}). \end{aligned}$$

The test statistic for H_0 versus $H_1 - H_0$ becomes $t^B_{01} = 185.78$ with p-value < 0.0001, so using (5.9) we reject H_0 and conclude the increasing age and decreasing income levels have significant and directional effects on youth smoking status, while the $\$45,000 - \$59,999$ and $\$60,000 - \$79,999$ income groups do not differ. Moreover the unconstrained test, H_0 versus $H_2 - H_0$ has test statistic $t_{02} = 186.04$ with p-value < 0.0001, suggesting age and income effects are significant, however no directional effects are indicated. Thus, the constrained likelihood ratio tests provide additional information which would not otherwise be detected using unconstrained hypothesis tests.

Table 6.2. Parameter estimates for Youth Smoking Survey 2002

	Unconstrained		**Constrained**	
Covariate	$\{\gamma : \gamma \in \mathbb{R}^{p+q}\}$	Standard Error*	C_{Both0}	C_{Both1}
Intercept	-8.1744	0.4170	-8.2627	-8.1897
$AGE16$	4.7020	0.4499	2.6117	4.7028
$AGE15$	3.5485	0.3628	2.6117	3.5511
$AGE14$	3.0613	0.3445	2.6117	3.0656
$AGE13$	2.1791	0.3522	2.6117	2.1810
$AGE12$	1.5057	0.3746	2.6117	1.5079
$ABOR$	0.6053	0.1638	0.6121	0.6102
SEX	-0.0101	0.1034	-0.0668	-0.0106
ACT	-0.2171	0.0426	-0.2694	-0.2162
$HSMOK$	0.7598	0.1041	0.8380	0.7611
$INC1$	0.5813	0.1752	0.4256	0.5753
$INC2$	0.1872	0.1958	0.4256	0.1535
$INC3$	0.1124	0.2130	0.4256	0.1535
$FSMOK$	3.1594	0.1920	3.4069	3.1611
σ	0.5197	0.0733	0.5474	0.5479

*where the standard error refers to the unconstrained value.

6.2 Northern Contaminants Programme

6.2.1 Description of the Data

The Canadian government's Northern Contaminants Programme (NCP) (Indian and Northern Affairs Canada, 2009) was established in 1991 to address issues pertaining to contaminants in Arctic Canada, especially those related to human health. Recent studies have shown that environmental contaminants are present in many animal species traditionally consumed by northerners and often such contaminants are transported over long distances to the Arctic atmospherically and in waters flowing northward (Butler Walker et. al., 2003). Studies prior to 1991 found substances, many with no Arctic or Canadian sources, exhibiting high levels in the Arctic ecosystem. The contaminants under study include metals, radionuclides and organochlorines. Organochlorines, or persistent organic pollutants (POPs), include pesticides and industrial compounds such as polychlorinated biphenyls (PCBs), dioxins and furans. The compounds are resistant to degradation, are lipophilic (fat soluble) and biomagnify in the food chain (Butler Walker et. al., 2003).

Since the early 1990s, the NCP has collected contaminant data from blood of new mothers and their infants in the Northwest territories and Nunavut (see

Figure 6.1). Mothers and their newborns are the target population since previous evidence has indicated the fetus is particularly vulnerable to exposure to various contaminants (Butler Walker et. al., 2006). In addition to contaminant analysis, a lifestyle survey was administered to participants to ascertain general demographic characteristics and the amount of traditional foods consumed. The ethnic composition of Canada's northern territories is composed of Inuit, Dene/Métis and non-aboriginals (primarily Caucasians), with different cultural and dietary customs. Research has demonstrated that traditional foods consumption such as marine and terrestrial mammals, fish and birds, may be responsible for elevated levels of certain contaminants (Butler Walker et. al., 2006). Other potential factors, such as a mother's age, and tobacco use may contribute to elevated exposure. Furthermore, the number of prior children born to the mother (parity) is expected to reduce blood contaminant levels. Information pertaining to such variables were also collected through the lifestyle questionnaire.

Between 1994-2003, the NCP included a voluntary survey of 384 mothers of various ethnic backgrounds and ages. In addition to estimating average contaminant levels (Butler Walker et. al., 2003, 2006), researchers are interested in the probability of a contaminant being detected in a mother's blood sample, and

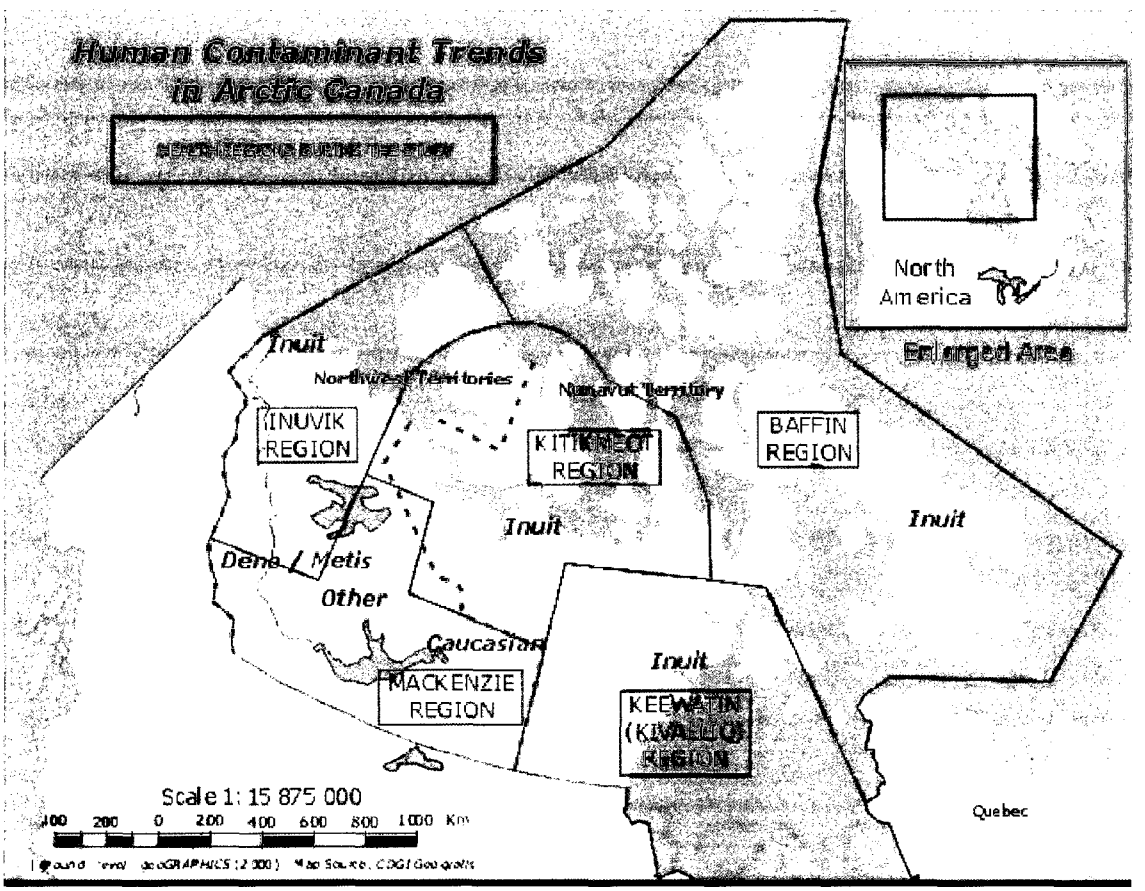

Figure 6.1: Major ethnic groups and health regions of the NCP monitoring program, with the Northwest Territories/Nunavut boundary post-1999 included (Government of Canada: Northern Contaminants Programme)

important factors therein. In Table 6.3, a summary of demographic variables is presented for the pesticide *p'p*-DDT (DDT). To represent tobacco consumption, a variable cigarette-years is used, which is the product of number of cigarettes smoked per day and the number of years smoked. We note that women with DDT levels detected tended to be Inuit, were older on average and exhibited a larger value of cigarette-years. A mother's parity does not seem to influence whether DDT was detected, since the percentages decrease then increase with a rise in the number of children born. Traditional food usage exhibits an increasing effect, as 72.7% of mothers who were non-users had values of DDT above the detection limit, compared with 88.7% of high consumers.

Furthermore, there were 44 mothers, or 11%, for which the cigarette-years information was missing. In the following section, we explore the use of constrained statistical techniques with incomplete data to determine whether these observed effects are significant.

Table 6.3. Summary Statistics for Northern Contaminants Programme Data - *p'p*-DDT

			p'p-DDT Detected	
Covariate	**Level**	**Counts**	**Yes (n=297)**	**No (n=87)**
Parity	One child	157	77.07%	22.93%
	Two children	114	78.07%	21.93%
	Three children	60	73.33%	26.67%
	Four or more children	53	81.13%	18.87%
Ethnicity	Inuit	144	89.58%	10.42%
	Dene/Métis	93	61.29%	38.71%
	Other	147	75.51%	24.49%
Traditional	High	115	88.70%	11.30%
Food Use	Moderate	120	80.00%	20.00%
	Low	116	64.66%	35.34%
	Non-User	33	72.73%	27.27%
Mother's Age	Mean	384	28.17	27.15
	Standard deviation		6.14	5.84
	Range		(15,45)	(15,44)
CIGYEARS	Mean	340	81.34	37.05
(Complete Cases)	Standard deviation		121.05	52.46
	Range		(0,800)	(0,275)
	Percent Missing		13.47%	4.60%

6.2.2 Data Analysis

For the NCP analysis, the binary response variable of interest y_i is 1 if the ith mother had *p'p*-DDT detected, and 0 otherwise. The predictors outlined in Table 6.3 were originally considered in the analysis. Due to small sample sizes when crossed with other variables, the parity groups "Three Children" and "Four or

more children" were combined, as were the traditional food usage groups "Low" and "Nonuser." All other categorical variables remained as in Table 6.3.

Given that the smoking covariate cigarette-years displayed missing values, both a complete case analysis and a nonignorable model were considered in parameter estimation. The nonignorable model was fit since it is believed the missingness of cigarette-years is associated with the values, in other words smokers may have been less likely to report a value compared with nonsmokers.

After a preliminary analysis of the unconstrained model, it was noted that the Parity variables were not significant ($\chi^2 = 2.246$ with p-value $= 0.3253$ on 2 degrees of freedom). Thus, the number of children born to the mother is not a significant predictor of whether DDT is detected. However, since parity is of biological interest, these variables were retained in the model. Hence, for the response variable, y_i, with $\pi_i = P(y_i = 1|\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i}, \boldsymbol{\beta})$, the predictors considered were

$$\begin{aligned}\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \beta_1 Parity_{1i} + \beta_2 Parity_{2i} + \beta_3 Inuit_i + \beta_4 DeneMetis_i \\ &\quad + \beta_5 FoodHigh_i + \beta_6 FoodModerate_i + \beta_7 Age_i + \beta_8 LogCigYears_i\end{aligned}$$

where $Parity_{1i}$, $Parity_{2i}$, $Inuit_i$, $DeneMetis_i$, $FoodHigh_i$, $FoodModerate_i$, are indicator variables, and Age_i and $LogCigYears_i$ are continuous variables. The logarithm of cigarette-years was considered to reduce the variability of the values.

We model the joint density of these two variables as $f(Age_i, LogCigYears_i|\boldsymbol{\alpha})$, assumed to follow a multivariate normal distribution with $\boldsymbol{\alpha} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12})^{\mathrm{T}}$. Here, μ_1, σ_1^2 represent the mean and variance of the mother's age, whereas μ_2, σ_2^2 represent the mean and variance of cigarette-years, and σ_{12} represents the covariance between age and cigarette-years, expected to be nonzero.

For the nonignorable case, a model for the missing data mechanism is also required. If we let $r_i = 1$ of the ith observation of cigarette-years is missing and 0 otherwise, then

$$\begin{aligned}\text{logit}(P(r_i = 1)) \;=\; & \psi_0 + \psi_1 Parity_{1i} + \psi_2 Parity_{2i} + \psi_3 Inuit_i + \psi_4 DeneMetis_i \\ & + \psi_5 Age_i + \psi_6 LogCigYears_i + \psi_7 y_i.\end{aligned}$$

The traditional food usage variables were not included in the missing data model since a Wald test showed these terms to be insignificant. As a type of sensitivity analysis, we fit various missing data models and noted no substantial changes in the estimates of the regression parameters.

In the MAR situation, assuming the missing values of cigarette-years do not depend on the values themselves, the unconstrained estimates are presented in Table 6.4. We note that the standard errors are smaller than the complete case counterparts, and certain parameter estimates are different between the two

procedures. For instance, the effect of being an Inuit has a higher value for the MAR versus complete cases, whereas the effect of cigarette-years is lower for MAR than for complete cases.

The unconstrained nonignorable model was fit using the EM Algorithm, and was compared to the complete case analysis in Table 6.5. The parameter estimates and standard errors are also provided. We note that the complete case displays higher standard errors for all parameters as compared with the nonignorable model, which is indicative of a loss of efficiency. In addition, the parameter $\hat{\psi}_6 = -1.1755$ is significant (SE = 0.1439), indicating the nonignorable assumption is reasonable. The negative sign is interesting, since it indicates that the higher the cigarette-years, the lower the probability of a missing value, which is opposite to what is expected. In consultation with subject-matter experts, it was suggested that since most northern mothers are smokers, there is little stigma associated with expressing the smoking status. As a result, the question was likely left blank for non-smokers, since no additional information could be gathered.

Other parameter estimates, such as β_3 (Inuit), β_7 (age), and β_8 (cigarette-years), exhibit differences under the nonignorable model as opposed to the complete cases. As expected, since values of cigarette-years are missing, estimates

of $\boldsymbol{\alpha}$ vary between the two models. Estimates of $\hat{\mu}_1$ and $\hat{\mu}_2$ are smaller for the nonignorable EM, whereas variability of σ_1^2 and σ_2^2 are estimated to be larger under the missing data model. The covariance σ_{12} is estimated to be closer to zero based on the nonignorable model.

Next, we impose constraints on the parity groups and traditional food usage variables, to test whether decreased parity and increased usage leads to a higher probability of *p'p*-DDT detected. We define the constrained parameter space to be $C = \{\boldsymbol{\gamma} : \beta_1 \geq \beta_2, \beta_5 \geq \beta_6 \geq 0\}$. Since the complete case, ignorable and nonignorable unconstrained models satisfy C, they are also the constrained estimates relative to this parameter space.

Based on the previous graphical analysis and biological relevance, we also test whether a difference exists between the two parity groups, assuming the high and moderate traditional food categories retain a natural ordering. More specifically, under H_0, define $C_0 = \{\boldsymbol{\gamma} : \beta_1 = \beta_2, \beta_5 = \beta_6 = 0\}$, while under H_1, we have $C_1 = \{\boldsymbol{\gamma} : \beta_1 = \beta_2, \beta_5 \geq \beta_6 \geq 0\}$. Constrained estimates for both complete cases, ignorable and nonignorable models were found using the gradient projection algorithm described in Section 4.2. As in the unconstrained setting, differences are noted with the complete cases. In particular, values of age are higher, whereas cigarette-years values are lower.

The hypotheses of interest may be written as $H_0 : A\boldsymbol{\gamma} = \mathbf{c}$, $H_1 : A\boldsymbol{\gamma} \leq \mathbf{c}$, and $H_2 : \boldsymbol{\gamma}$ unconstrained, with $r = 3$, $\mathbf{c} = \mathbf{0}$ of length r. As in the youth smoking setting, since the constraint space involves an equality, chi-bar-square weights must be modified slightly to avoid redundancy, and we re-write $C_1 = \{\boldsymbol{\gamma} : A_1\boldsymbol{\gamma} \leq \mathbf{0}, A_2\boldsymbol{\gamma} = \mathbf{0}\}$ where A_1 represents the inequality constraints with dimension $s = 2$, i.e. the second and third rows of A. Then, A_2 represents the equality constraint with dimension $t = 1$, from the first row of A. Using Theorem 3.6 equation 8, the adjusted weight vector is found with $N = 50,000$ simulations to be

$$w(s, V_{new}(\hat{\boldsymbol{\gamma}})) = (0.31540, 0.49818, 0.18642)^{\mathrm{T}}$$

with adjusted variance-covariance matrix

$$V_{new} = A_1 V(\hat{\boldsymbol{\gamma}}) A_1^{\mathrm{T}} - (A_1 V(\hat{\boldsymbol{\gamma}}) A_2^{\mathrm{T}})(A_2 V(\hat{\boldsymbol{\gamma}}) A_2^{\mathrm{T}})^{-1}(A_2 V(\hat{\boldsymbol{\gamma}}) A_1^{\mathrm{T}}).$$

In the MAR setting, the goodness-of-fit test statistic is $t_{12} = 0.403$ with p-value $= 0.7062$, thus the constraints are significant. The test statistic for H_0 versus $H_1 - H_0$ becomes $t_{01} = 5.990$ with p-value $= 0.0034$, so we conclude that traditional food usage has an increasing effect on the probability of DDT detected, and there is no significant difference between low and non-consumers. Furthermore, the unconstrained test H_0 versus $H_2 - H_0$ has test statistic $t_{02} =$

9.393 and p-value = 0.0091. Hence, there is evidence of an effect, but we are unable to determine the direction using unconstrained hypothesis tests.

Furthermore, in the nonignorable case, the results are similar. The adjusted weight vector is

$$w(s, V_{new}(\hat{\gamma})) = (0.23714, 0.49912, 0.26374)^{\mathrm{T}}$$

while the test statistic $t_{12} = 0.314$ with p-value = 0.7535. As before, the constrains are significant. The test statistic $t_{01} = 9.816$ with p-value = 0.0028, so we conclude the two parity groups have a similar effect, and traditional usage has an increasing effect on whether DDT is detected.

As a comparison, we repeat the above analysis for the complete case, and find the goodness-of-fit test statistic $t_{12CC} = 0.006$ with p-value = 0.9668, thus the constraints are also significant. Moreover, the test statistic $t_{01CC} = 7.258$ with p-value = 0.0085, and the unconstrained test statistic $t_{02CC} = 7.264$ has p-value = 0.0264. Hence, the conclusions are identical, however we have lost precision in that the p-values are larger in the complete case setting.

Table 6.4. Parameter estimates for Northern Contaminants Programme - Complete Cases and MAR

	Complete Cases				MAR			
	Unconstrained		*Constrained*		*Unconstrained*		*Constrained*	
	$\{\gamma \in \mathbb{R}^{p+q}\}$	SE*	C_0	C_1	$\{\gamma \in \mathbb{R}^{p+q}\}$	SE*	C_0	C_1
β_0	-1.3292	1.0459	-1.4666	-1.3103	-1.6640	1.0133	-1.6666	-1.5181
β_1	0.5097	0.3676	0.4925	0.4967	0.5446	0.3608	0.4556	0.4431
β_2	0.4826	0.3740	0.4925	0.4967	0.3371	0.3606	0.4556	0.4431
β_3	0.5556	0.4803	1.0755	0.5529	0.8485	0.4643	1.4454	0.8346
β_4	-0.6795	0.3769	-0.3793	-0.6830	-0.6437	0.3694	-0.3311	-0.6685
β_5	1.0793	0.4258	0.0000	1.0766	1.1520	0.4028	0.0000	1.1309
β_6	0.4778	0.3112	0.0000	0.4760	0.5792	0.3096	0.0000	0.5692
β_7	0.0536	0.0289	0.0636	0.0531	0.0641	0.0280	0.0711	0.0602
β_8	0.1333	0.0589	0.1332	0.1328	0.1241	0.0555	0.1186	0.1201
μ_1	28.1118	0.3217	28.1118	28.1119	27.9375	0.3101	27.9375	27.9375
μ_2	2.5509	0.1315	2.5509	2.5507	2.5571	0.1173	2.5594	2.5575
σ_1^2	35.1875	2.6988	35.1874	35.1857	36.9180	2.6643	36.9178	36.9179
σ_2^2	5.8816	0.4511	5.8817	5.8824	5.8711	0.4027	5.8705	5.8698
σ_{12}	-0.9655	0.7819	-0.9655	-0.9655	-1.0050	0.7022	-1.0136	-1.0071

*where the standard error refers to the unconstrained value.

Table 6.5. Parameter estimates for Northern Contaminants Programme - Complete Cases and Nonignorable

	Complete Cases				Nonignorable			
	Unconstrained		*Constrained*		*Unconstrained*		*Constrained*	
	$\{\gamma \in \mathbb{R}^{p+q}\}$	SE*	C_0	C_1	$\{\gamma \in \mathbb{R}^{p+q+l}\}$	SE*	C_0	C_1
β_0	-1.3292	1.0459	-1.4666	-1.3103	-1.6326	1.0082	-1.6327	-1.5181
β_1	0.5097	0.3676	0.4925	0.4967	0.5305	0.3603	0.4452	0.4511
β_2	0.4826	0.3740	0.4925	0.4967	0.3381	0.3608	0.4452	0.4511
β_3	0.5556	0.4803	1.0755	0.5529	0.9158	0.4597	1.5548	0.9028
β_4	-0.6795	0.3769	-0.3793	-0.6830	-0.6310	0.3681	-0.2923	-0.6519
β_5	1.0793	0.4258	0.0000	1.0766	1.2142	0.4053	0.0000	1.1957
β_6	0.4778	0.3112	0.0000	0.4760	0.5844	0.3095	0.0000	0.5769
β_7	0.0536	0.0289	0.0636	0.0531	0.0641	0.0278	0.0717	0.0611
β_8	0.1333	0.0589	0.1332	0.1328	0.1093	0.0522	0.0935	0.1070
μ_1	28.1118	0.3217	28.1118	28.1119	27.9375	0.3101	27.9375	27.9375
μ_2	2.5509	0.1315	2.5509	2.5507	2.1914	0.1294	2.1936	2.1920
σ_1^2	35.1875	2.6988	35.1874	35.1857	36.9180	2.6643	36.9178	36.9173
σ_2^2	5.8816	0.4511	5.8817	5.8824	6.6572	0.4484	6.6444	6.6526
σ_{12}	-0.9655	0.7819	-0.9655	-0.9655	-0.4535	0.7784	-0.4595	-0.4547
ψ_0					-9.9275	2.0150	-9.8470	-9.8993
ψ_1					0.4071	0.6031	0.3952	0.4021
ψ_2					1.3313	0.6504	1.3115	1.3188
ψ_3					6.5334	1.0113	6.4992	6.5200
ψ_4					2.6986	0.9150	2.6884	2.6937
ψ_5					0.1017	0.0434	0.1008	0.1015
ψ_6					-1.1755	0.1439	-1.1664	-1.1709
ψ_7					1.5408	0.7511	1.5229	1.5354

*where the standard error refers to the unconstrained value.

Chapter 7

Summary and Future Work

We have demonstrated that incorporating inequality constraints in nonlinear models results in increased efficiency at the expense of additional theoretical and computational development. We derived constrained maximum likelihood estimators for generalized linear mixed models and generalized linear models with missing data, utilizing a modification of the gradient projection algorithm technique. The new algorithms were seen to converge within a reasonable number of steps under the constraints considered. Constrained likelihood ratio test statistics were also derived for the models under study and were shown to follow a chi-bar-square distribution asymptotically, under the null hypothesis. Practical implementation of these distributions necessitates calculation of chi-bar weights,

for which a computational algorithm was established and R code written.

Simulation studies and power comparisons demonstrated efficiency gains for both models when constrained techniques were implemented. Theoretical results were applied to two government datasets, which demonstrated increased analytic capability over previous methods. The analysis produced results which are beneficial to government policymakers, and would not have otherwise been detected using previous methods.

Contrary to unconstrained models, hypothesis testing under the chi-bar-square distribution is the preferred method of inference as confidence intervals are difficult to obtain in constrained environments. The distribution of the constrained estimator $\tilde{\boldsymbol{\beta}}$ is dependent upon the proximity of the unconstrained estimator to the boundary of the constraint, and has been derived only for linear models under simple orderings, as per Hwang and Peddada (1994). Additional theoretical advances would be needed to obtain confidence intervals for more complicated models in constrained settings.

The wide applicability of constrained inference suggests numerous possible areas of future research. Methods similar to those of Binder, Kovacevic and Roberts (2005) suggest further improvements to parameter estimates by incorporating survey weights and information pertaining to the complex survey design in

the model analysis. Additional technical details are needed to incorporate such weights into the likelihood function, which will vary depending on the survey design. Such methods would allow for large government surveys using complex weighting schemes to implement advanced constrained inference techniques.

The paper by Dunson and Neelon (2003) regarding Bayesian constrained methods for GLMs, highlights the usefulness of Bayesian constrained procedures. With continual advances in computational methods, Bayesian constrained techniques are a timely and useful area of future research. The authors noted that sampling from the constrained posterior distribution is obtained by transforming draws from the unconstrained posterior density. As a result, existing Gibbs sampling algorithms for posterior computation of generalized linear models apply directly.

Alvo (2008) considers nonparametric tests for umbrella orderings, for which the dosage mean values increase and then decrease after a certain peak. Other approaches to maximum likelihood which require fewer assumptions, are also relevant in constrained settings, and could be extended to nonparametric methods for other parameter orderings.

Another area of future work would be to consider constraints on variance-covariance parameters. Calvin and Dykstra (1995) developed a REML estima-

tion scheme for covariance matrices, with both balanced and unbalanced data. Such an extension would be particularly useful in generalized linear mixed models, for which tests for increasing or decreasing trends in variance components could be developed.

In addition, if a researcher is interested solely in regression parameters and not in variance components, constrained inference for marginal models may be developed. In the unconstrained case, such methods are known to produce consistent, if less efficient, estimators then maximum likelihood, however are easier to compute.

Furthermore, some of the criticisms of the EM algorithm include its problem of slow convergence, compared to other maximization techniques. This issue would be further complicated in the gradient projection algorithm if a large number of constraints were present. A vast body of literature details modifications to either the E- or M- steps to accelerate convergence. Unconstrained extensions such as the ECM algorithm (Meng and Rubin, 1993) which replaces each M step of EM by a sequence of conditional maximization steps which monotonically increase the likelihood function. Liu and Rubin (1994) proposed the Expectation/Conditional Maximization Either (ECME) algorithm which replaces some of the conditional maximization steps of the ECM algorithm with steps that

maximize the corresponding actual likelihood function. The algorithm maintains the stable monotone convergence and simplicity of EM and ECM, and has the advantage of a faster rate of convergence since the actual likelihood is maximized in earlier stages, rather than an approximation. The Alternating Expectation Conditional Maximization (AECM) algorithm of Meng and van Dyk (1997) extends the ECME by maximizing functions other than the likelihood or Q, which correspond to various definitions of missing data, in an effort to improve computing time. Other hybrid maximization methods are outlined in Little and Rubin (2002) and include the gradient EM of Lange (1995 a,b) and the accelerated EM (AEM) of Jamshidian and Jenrich (1993). Nettleton (1999) outlined a theorem for constrained estimators from the ECM and AECM algorithms, which suggests extensions to constrained versions of other EM-type algorithms are useful.

In addition, rather than apply the EM algorithm, other missing value methods such as multiple imputation could be used. Such methods may simplify iterative procedures, especially when the number of constraints is large.

More recently, authors have investigated likelihood ratio tests for multivariate responses. Sinha, Laird and Fitzmaurice (2010) analyzed a multivariate logistic regression model with multiple binary outcomes in the presence of incomplete

data and auxiliary information. Multivariate analysis is preferred over marginal modeling of each variable, since efficiency gains are noted when there are strong associations among the multiple outcomes. In addition, joint likelihood analysis and tests of heterogeneity may be performed. Farrell and Park (2007) proposed a constrained likelihood ratio tests for ordered group effects with one binary and one continuous response. Their method was shown to follow a chi-bar-square distribution asymptotically. Further research would extend such results to more complicated models and constrained parameter spaces.

References

Abramowitz M., and Stegun, I. (eds.) (1964). *Handbook of Mathematical Functions.* National Bureau of Standards, Washington D.C.

Agresti, A. and Coull, B. A. (1998). An empirical comparison of inference using order-restricted and linear logit models for binary response. *Communications in Statistics - Simulations.* 27: 147-166.

Alvo, M. (2008) Nonparametric tests of hypotheses for umbrella orderings. *The Canadian Journal of Statistics*, 36: 143-156.

Baker, S.G. and Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83: 62-69.

Bartholomew, D. J.(1959a). A test of homogeneity for ordered alternatives. *Biometrika*, 46: 36-48.

————————(1959b). A test of homogeneity for ordered alternatives II. *Biometrika*, 46: 328-335.

————————(1961a) A Test of Homogeneity of Means Under Restricted Alternatives. *Journal of the Royal Statistical Society, Series B*, 23 (2) 239-281.

————————(1961b) Ordered Tests in the Analysis of Variance. *Biometrika*, 48(3/4) 325-332.

Becker, M.P., Yang, I., and Lange, K. (1997). EM algorithms without missing data. *Statistical Methods in Medical Research*, 6: 38-54.

Binder, D. A., Kovacevic, M. S. and Roberts, G. (2005). How important is the informativeness of the sample design? *Proceedings of the Survey Methods Sections, Statistical Society of Canada.*

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association,* 88: 9-25.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* 82: 81-91.

Butler Walker, J., Seddon, L., McMullen, E., Houseman, J., Tofflemire, K., Corriveau, A., Weber, J.P., Mills, C., Smith, S. and Van Oostdam, J., (2003) Organochlorine levels in maternal and umbilical cord blood plasma in Arctic Canada. *The Science of the Total Environment.* 302: 27-52.

Butler Walker, J., Houseman, J., Seddon, L., McMullen, E., Tofflemire, K., Mills, C., Corriveau, A., Weber, J.P., LeBlanc, A., Walker, M., Donaldson, S.G. and Van Oostdam, J., (2006) Maternal and umbilical cord blood levels of mercury, lead, cadmium, and essential trace elements in Arctic Canada. *Environmental Research.* 100: 295-318.

Calvin, J.A. and Dykstra, R.L. (1995) REML Estimation of Covariance Matrices with Restricted Parameter Spaces. *Journal of the American Statistical Association*, 90: 321-329.

Dardanoni, V. and Forcina, A. (1998). A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *Journal of the American Statistical Association*, 93: 1112-1123.

Davis, K. A., Park, C. G. and Sinha, S. K. (2008) Constrained inference in generalized linear and mixed models. *Proceedings of the Survey Methods Section, Statistical Society of Canada.*

Dawson, D. and Magee, L. (2001) The National Hockey League Entry Draft, 1969-1995: An Application of a Weighted Pool-Adjacent Violators Algorithm. *The American Statistician*, 55(3) 194-199.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39: 1-38.

Domowitz, I. and White, H. (1982). Misspecified Models with Dependent Observations. *Journal of Econometrics.* 20: 35-58.

Dunson, D. B. and Neelon, B. (2003). Bayesian inference on order–constrained parameters in generalized linear models. *Biometrics* 59: 286-295.

Dykstra, R. (1983). An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78: 837-842.

Eeden, C. van (1956). Maximum likelihood estimation or ordered probabilities. *Proc. K. ned. Akad. Wet (A)*, **59**/*Indag. Math.*, **18**, 444-455.

El Barmi, H. and Dykstra, R. (1994). Restricted multinomial maximum likelihood estimation based upon Fenchel duality. *Statistics and Probability Letters.* 21: 121-130.

————————(1995). Testing for and against a set of linear inequality constraints in a multinomial setting. *The Canadian Journal of Statistics*, 23: 131-143.

El Barmi, H. and Johnson, M. (2006). A unified approach to testing for and against a set of linear inequality constraints in the product multinomial setting. *Journal of Multivariate Analysis*, 97: 1894-1912.

Fahrmeir, L., and Klinger, J. (1994). Estimating and testing generalized linear models under inequality restrictions, *Statistical Papers*, 35: 211-229

Farrell, P.J. and Park, C.G. (2007) Testing for ordered group effects in binary and continuous outcomes. *Biometrical Journal* 49: 585-598.

Fitzmaurice, G.M., Lipsitz, S.R. and Ibrahim, J.G. (2007) A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics* 63: 942-946.

Gelfand, A., and Carlin, B. (1993). Maximum-likelihood estimation for constrained- or missing-data models. *Canadian Journal of Statistics* 21: 303-311.

Genz, A. (1992) Numerical Computation of Multivariate Normal Probabilities. *Journal of Computational and Graphical Statistics* 1: 141-149.

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistics Society, Series B* 54: 657-699.

Gourieroux, C., Holly, A., and Monfort, A. (1982) Likelihood ratio test, Wald test, and Kuhn-Tucker-test in linear models with inequality constraints on the regression parameters. *Econometrica* 50: 63-80.

Gu, M.G. and Kong, F.H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences* 95: 7270-7274.

Hall, D.B., Praestgaard, J.T. (2001). Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika* 88: 739-751.

Health Canada (2005). 2002 Youth Smoking Survey Technical Report. Health Canada Catalogue no. H46-1/44-2002E. Ottawa, Ontario. http://www.hc-sc.gc.ca/hc-ps/pubs/tobac-tabac/yss-etj-2002/index-eng.php (accessed February 19, 2011)

Hwang, J. T. G., and Peddada, S. D. (1994) Confidence interval estimation subject to order restrictions. *The Annals of Statistics.* 22: 67-93.

Ibrahim, J.G., Chen, M.H.. and Lipsitz, S.R. (2001) Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika* 88: 551-564.

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R. and Herring, A.H. (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. *Journal of the American Statistical Association* 100: 332-346.

Ibrahim, J.G., Lipsitz, S.R. and Chen, M.H. (1999) Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society, Series B* 61: 173-190.

Indian and Northern Affairs Canada (2009). Northern Contaminants Programme: Canadian Arctic Contaminants and Health Assessment Report: Human Health 2009. http://www.ainc-inac.gc.ca/nth/ct/ncp/pubs/har/har-eng.pdf (accessed February 19, 2011).

Jamshidian, M. (2004). On algorithms for restricted maximum likelihood estimation. *Computational Statistics & Data Analysis* 45: 137-157.

Jamshidian, M. and Jenrich, R.I. (1993). Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association* 88: 221-228.

Kim, D.K., and Taylor, J.M.G. (1995). The Restricted EM Algorithm for Maximum Likelihood Estimation Under Linear Restrictions on the Parameters. *Journal of the American Statistical Association* 90: 708-716.

Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika* 50: 403-418.

Kuhn, H.W. and Tucker, A.W. (1951). Nonlinear Programming. *2nd Berkeley Symposium, Math. Statist. Probab. (University of California Press, Berkeley)*, 481–492.

Lange, K. (1995a). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B* 57: 425-437.

————————(1995b). A quasi-Newtonian acceleration of the EM algorithm. *Statistica Sinica* 5:1-18.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika.* 73: 13-22.

Lin, X. (1997). Variance component testing in generalized linear models with random effects. *Biometrika* 84: 309-326.

Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* 91: 1007-1016.

Little, R.J.A. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association* 90: 1112-1121.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd Ed.* New York: Wiley.

Liu, C.H. and Rubin, D.B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* 81: 633-648.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society - Series B.* 44: 226-233.

Luenberger, D.G. (2003). *Linear and Nonlinear Programming, 2nd Edition.* New York: Springer.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, 2nd Ed.* London: Chapman & Hall.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models *Journal of the American Statistical Association* 92: 162-170.

McCulloch, C.E., Searle, S.E. and Neuhaus, J. (2008). *Generalized, Linear, and Mixed Models, 2nd edition* New York, USA: John Wiley.

McLachlan, J. and Krishnan, T. (1997). *The EM Algorithm and Extensions* New York: John Wiley.

Meng, X.-L. and Rubin, D.B. (1993). Maximization likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80: 267-278.

Meng, X.-L. and van Dyk, D.A. (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B* 59: 511-567.

Molenberghs, G. and Verbeke, G. (2006). *Models for Discrete Longitudinal Data.* New York: Springer.

Nettleton, D. (1999). Convergence properties of the EM algorithm in constrained parameter spaces. *The Canadian Journal of Statistics* 27: 639-648.

Nettleton, D. and Praestgaard, J. (1998). Interval mapping of quantitative trait loci through order-restricted inference. *Biometrics* 54: 74-87.

Park, T., Shin, D. W., and Park, C. G. (1998). A generalized estimating equations approach for testing ordered group effects with repeated measurements. *Biometrics*, 54: 1645-1653.

Peng, J., Lee, C. I. C., Davis, K. A., and Wang, W. (2008) Stepwise confidence intervals for monotone dose–response studies. *Biometrics.* 64: 877-885.

Piergosch, W.W. (1990). One-sided significance tests for generalized linear models under dichotomous response. *Biometrics* 46: 309-316.

Pilla, R.S., Qu, A., Loader, C. (2006). Testing for order-restricted hypotheses in longitudinal data. *Journal of the Royal Statistical Society - Series B.* 68: 437-455.

Raudenbush, S., Yang, M., and Y.M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation, *Journal of Computational and Graphical Statistics* 9: 141-157.

Richardson, M., Richardson, P., Smith, T. (1992). The monotonicity of the term premium. *Journal of Financial Economics* 31: 97-105.

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). *Order Restricted Statistical Inference,* Chichester, U.K.: John Wiley.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89: 846-866.

Ruberg, S.J. (1989) Contrasts for identifying the minimum effective dose. *Journal of the American Statistical Association* 84: 816-822.

Rubin, D.B. (1976) Inference and missing data. *Biometrika* 63: 581-592.

Shapiro, A. (1988) Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review* 56: 49-62.

Shi, N.Z., Zheng, S.R., Guo, J. (2005). The restricted EM algorithm under inequality restrictions on the parameters. *Journal of Multivariate Analysis* 92: 53-76.

Shin, D. W., Park, C. G., and Park, T. (1996). Testing ordered group effects with repeated measurements. *Biometrika*, 83: 688-694.

Silvapulle, M.J. (1994). On tests against one-sided hypotheses in some generalized linear models. *Biometrics*, 50: 853-858.

Silvapulle, M.J. and Sen, P.K. (2005) *Constrained Statistical Inference,* Hoboken, New Jersey: John Wiley.

Sinha, S.K. (2004). Robust Analysis of Generalized Linear Mixed Models. *Journal of the American Statistical Association.* 99: 451-460.

————————(2006). Robust inference in generalized linear models for longitudinal data. *The Canadian Journal of Statistics.* 34: 261-278.

————————(2008). Robust methods for generalized linear models with nonignorable missing covariates. *The Canadian Journal of Statistics.* 36: 277-299.

————————(2009). Bootstrap tests for variance components in generalized linear mixed models. *The Canadian Journal of Statistics.* 37: 219-234.

Sinha, S.K., Laird, N.M. and Fitzmaurice, G.M. (2010). Multivariate logistic regression with incomplete covariate and auxiliary information. *Journal of Multivariate Analysis* 101: 2389-2397.

Tian, G.L., Ng, K.W., and Tan, M. (2008) EM-type algorithms for computing restricted MLEs in multivariate normal distributions and multivariate t-distributions. *Computational Statistics and Data Analysis* 52: 4768-4778.

White, H. (1984) *Asymptotic Theory for Econometricians* London: Academic Press.

Wu, C.F.J. (1983) On the convergence properties of the EM Algorithm. *The Annals of Statistics* 11: 95-103.

Zheng, S.R., Shi, N.Z., Guo, J. (2005). The restricted EM algorithm under linear inequalities in a linear model with missing data. *Science in China Series A - Mathematics* 48: 819-828.

Appendix I - R Algorithm to Compute Chi-Bar-Square Weights

```
############################################################
# R Program to Compute Chi-bar-square weight vector #
# for given variance-covariance matrix #
# #
# Author: Karelyn Davis #
# Date: August 7, 2008 #
############################################################

# Define number of iterations and length of weight vector
# Example below illustrates for r = 4 constraints

N1 <- 50000
a0 <- 0; a1 <- 0; a2 <- 0; a3 <- 0; a4 <- 0

# Need as inputs: constraint matrix Astar_r and
# variance-covariance matrix var.mat
```

```
for(i in 1:N1) {

var.matA <- (Astar_r) %*% var.mat %*% t(Astar_r)

sig_beta <- var.matA


beta.1 <- mvrnorm(n=1, mu=rep(0,4), Sigma = sig_beta)

Dmat <- solve(sig_beta)
Z1 <- as.matrix(beta.1, ncol = 1)


del.fun <- function(delta, Z, D) {
quad.del <- (Z - delta)
t(quad.del) %*% D %*% quad.del
}

comp <- nlminb(start = c(0,0), del.fun, lower = 0,
upper = Inf, Z = Z1, D = sig_beta)$par

sum_comp <- sum(comp > 0)

if (sum_comp == 0) {a0 <- a0 + 1}
if (sum_comp == 1) {a1 <- a1 + 1}
if (sum_comp == 2) {a2 <- a2 + 1}
if (sum_comp == 3) {a3 <- a3 + 1}
if (sum_comp == 4) {a4 <- a4 + 1}


}


weights.vec <- c(a0/N1, a1/N1, a2/N1, a3/N1, a4/N1)
```