A Comparison of Methods for Multivariate Familial Binary Responses

by

Abu Hena M. Mahbub-ul Latif

M.Sc., Dhaka University, 1991

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

we accept this thesis as conforming
to the required standard

The University of British Columbia

September 2001

Department of STATISTICS

The University of British Columbia
Vancouver, Canada

Date September 25, 2001

Abstract

Among the existing methods for analysing the multivariate familial binary response, we discuss latent variable models and the estimating equations based methods. A brief description of the multivariate Plackett distribution is given and the role of this distribution in developing the estimating equations based methods is pointed out. The maximum likelihood and estimating equations based methods for estimating the parameters of the multivariate logistic model are compared. For this comparison, a simulation study examines the effects of the sample sizes, dependence structures, the within–family dependence, etc. in estimating the parameters. The data are generated from the multivariate probit models. The multivariate logistic and probit models are compared for estimating conditional probabilities of interest in a genetics context and the respective standard errors. Numerical methods are used to estimate the parameters of the models considered. Because the original GEE2 code cannot handle multivariate binary data for arbitrary family structures, we have a new implementation of the GEE2 method for familial data; this routine used automatic differentiation for computing the Hessian matrix.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr. Harry Joe for his guidance and supervision throughout the development of this thesis. I would also like to thank Dr. John Petkau for his careful reading of the manuscript and subsequent helpful suggestions. Many thanks to Dr. Lionel Pandolfo for his generous support.

ABU HENA M. MAHBUB-UL LATIF

The University of British Columbia
September 2001

To my Parents.

Chapter 1

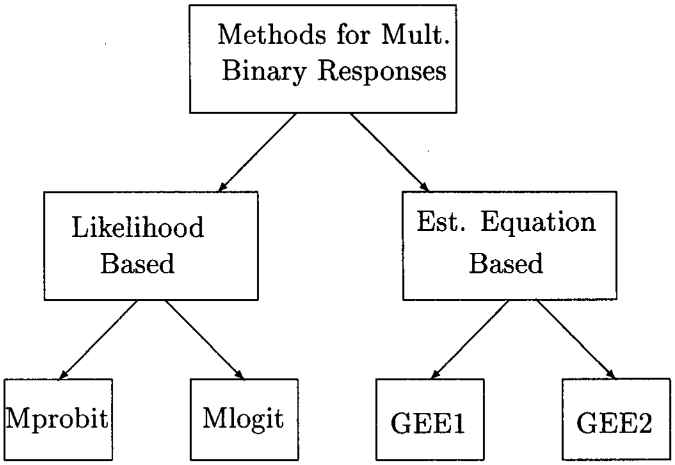
Introduction

In quantitative genetics, researchers are often interested in identifying important variables related to the occurrence of a genetic disease of interest. Since relatives are genetically more alike, the association between the members of a family (familial aggregation) for the occurrence of the disease is also of interest. Accurate quantification of the familial aggregation leads to more sophisticated genetic studies. Depending on the type of the disease and the objective of the study, the response corresponding to the disease status can be continuous or discrete. For continuous responses (e.g. blood pressure, cholesterol level, etc.), statistical models are well developed (see [1]) to meet the objective of this type of genetic study. But for discrete responses (e.g. presence/absence of disease, levels of disease status, etc.), there is still active research in development of statistical methods. In this thesis, we focus on methods available for binary response.

The same data structure also arises in the analysis of repeated measurements, where each individual is examined over time to record the presence/absence of a specific event. Covariates are also recorded for all the individuals over time. Since the measurements are made from the members of a family or from the same individual over time, the responses are usually positively correlated. Responses of this type are known as multivariate or correlated binary responses. The methods available for analyzing multivariate binary response can be classified into two broad classes of methods: likelihood based and estimating equation based methods. The likelihood based methods require complete specification of the joint distribution of the multivariate responses. On the other hand, the estimating equation based methods can be employed when the joint distribution is not fully specified. The likelihood based models can further be divided into several categories: (i) latent variable models, (ii) random effect models, (iii) transition or Markov models, and (iv) conditional logistic regression models. In this thesis, we are mainly interested in latent variable models because of the attractive interpretation of the parameters of this models.

The latent variable approach assumes that there is an underlying continuous variable which is categorized to give the observed discrete response. Here the regression models express the parameters of the distribution of the latent variable as functions of the explanatory variables. Equivalently, this is a model of the marginal distribution of the observed response directly as a function of the explanatory variables. For estimating equation based methods, only the univariate marginal probabilities and some form of the associations are specified. Figure 1.1 shows the classification of the methods for analyzing multivariate binary responses that are considered in this thesis.

Figure 1.1: Classifications of the methods for multivariate binary responses.



A common latent variable model is the probit model (Finney [8]), which is widely used in the dose-response studies. In such studies, different dosages of a stimulus are applied to randomly selected subjects and binary responses such as the presence/absence of a specific event of interest (e.g. death) are observed on the subjects. It is assumed that the binary responses depend upon the tolerance of the stimulus and for each subject there is a certain level of tolerance below (above) which the response occurs and above (below) which the response does not occur. In probit analysis, it is assumed that the unobserved tolerance variable follows normal distribution; hence the probability of the occurrence of the observed response can be expressed as a function of the cumulative distribution function of the standard normal distribution (see §2.1). In case of multivariate binary responses, by assuming the tolerance distribution is multivariate normal, Ashford and Sowden [2] used a multivariate probit model (Mprobit) to analyze the data from a dose-response study.

In quantitative genetics, when the character or trait is measured on an individual, the observed value is known as the phenotypic value; this depends on the genotypic value and the environmental deviation. The genotypic value is the aggregation of one or several genes possessed by the individual and all other sources of variation in the phenotypic value are assumed to arise from environmental deviation. The phenotypic value can be continuous or discrete. In this thesis, we consider only binary phenotypic values such as presence/absence of a specific trait. Falconer [7] proposed that there is an underlying variable liability and the observed phenotypic value depends on some threshold values of the distribution of the liability. The liability variable related to the phenotypic value is known as the phenotypic liability which is the sum of two independent random variables: the genetic liability and the environmental liability. Each of these two liability variables may be considered as the sum of many small effects so that a normal distribution is reasonable for liability variables, based on the central limit theorem (CLT). The liability variable is analogous to the tolerance in dose-response studies. If the distribution of the liability variable is assumed to be normal, the probit model can be used to analyze the binary response regarding the qualitative trait (e.g. presence/absence of the disease) of interest. In genetic study, the individual who is the first member of the family diagnosed with the disease is known as the proband and usually subjects are selected from the family members of the probands. Since the family members are genetically alike, responses of this type of study are usually correlated. Mendell and Elston [30] considered multivariate probit models for studying multifactorial qualitative traits. Lesaffre and Molenberghs [22] used multivariate probit model to study the relationship between drinking and smoking behavior among Belgians.

Another important latent variable model for univariate binary responses is the linear logistic model. This model assumes the underlying distribution is logistic and can be also used for identifying important covariates for occurrence of a specific event of interest. The standard logistic density is bell–shaped like the normal density and has a variance of $\pi^2/3$ (SD=1.81). Logistic regression for a binary response is common in biostatistics and epidemiology. This is mainly due to the convenient closed form of the logistic density and the ready interpretation of resulting regression coefficients as log odds–ratios; however it has not been derived from physical principles.

The multivariate version of the logistic model is not as advanced as the multivariate probit model because there is no natural multivariate logistic distribution. Bivariate Plackett distributions (see Plackett [34], Mardia [28]), can provide a legitimate joint bivariate distribution function corresponding to a given pair of the univariate margins with the cross–product ratio or odds–ratio as the dependence parameter. The bivariate logistic distribution with cross–product ratio as the dependence parameter can be constructed by using univariate logistic margins in the bivariate Plackett distribution. Molenberghs and Lesaffre [31] proposed a procedure for constructing multivariate Plackett distributions for the given univariate margins and two– and higher–order cross–product ratios as the dependence parameters. Joe [17] also discussed different properties and applications of bivariate and multivariate Plackett distributions, as well as a more general view of the construction. The joint distribution function of the multivariate Plackett construction satisfies the Fréchet bounds (see Joe [17]), which is one of the necessary conditions for a multivariate joint distribution function. To be a proper distribution function the joint distribution function of the three– and higher–order Plackett distribution must satisfy some necessary conditions; the analytic proof of these conditions are still open problems. Joe [17] has numerically shown for many combinations of parameters that the multivariate Plackett construction is a proper distribution if the third– and higher–order parameters are not too large or too small.

Dale [5] used bivariate Plackett distribution for analyzing bivariate categorical responses. She considered regression models corresponding to the univariate margins as well as the dependence parameter the global cross–product ratio. Molenberghs and Lesaffre [31] extended Dale's model for multivariate ordinal categorical responses. Dale's models are defined for any continuous univariate margins and different link functions can be considered for modeling the univariate margins and the cross–product ratios (so that there is not always a latent variable interpretation). A multivariate logistic model (Mlogit) can be obtained by using the logit link for the univariate margins and the log link for the cross–product ratios. Glonek and McCullagh [11] considered a different approach to define a multivariate logistic model. In their approach, the regression models are defined for both the univariate margins and the dependence parameters (cross–product ratios). No distributional assumption corresponding to the univariate margins are required to estimate the parameters of the model.

Beside these likelihood based methods for analyzing the multivariate binary responses, a common approach is generalized estimating equations (GEE). Liang and Zeger [23] proposed a estimating equation based method (known as GEE1) which can be used for analyzing both continuous and discrete correlated responses within the generalized linear model framework. This method provides inferences only for the regression coefficients and considers the dependence among the observations as a nuisance. This method can provide consistent estimators of the regression parameters if the specification of the marginal means is correct. They introduce the "working" correlation matrix in which a larger value of the working correlation parameter is used if there is more dependence in the data.

In 1990, Zhao and Prentice [38] proposed a method for correlated binary regression by using a quadratic exponential model. The quadratic exponential family is a special case of Cox's [3] log–linear representation of the joint distribution of multivariate binary responses. Zhao and Prentice called this method a "pseudo–likelihood" approach which can estimate the regression parameters corresponding to the marginal means and the dependence parameters of the model. The difference of the "pseudo–likelihood" approach to the classical likelihood approach is that the former does not require to estimate all the parameters which are necessary to fully define the likelihood function of interest. Zhao and Prentice [38] consider one–to–one transformations of the canonical parameters of the quadratic exponential family to the first two moments of the marginal responses. If the regression models corresponding to the marginal means and the pairwise marginal correlations are correctly specified then this pseudo–likelihood approach can consistently estimate the regression parameters of the model. By using the Fréchet bounds it can be shown that the range of the pairwise correlation coefficient depends on the univariate margins (see Joe [17], Chapter 7). Lipsitz, Laird and Harrington [27] considered a simulation study to show the advantages of odds–ratio over the correlation coefficient as a dependence parameter for analyzing multivariate binary responses.

Fitzmaurice and Laird [9] considered conditional log odds–ratios (canonical parameters of the Cox's log–linear representation) as the dependence parameters to model multivariate binary responses by using the quadratic exponential family. This model is explained more clearly in Joe and Liu [16]. This model is not appropriate to use for familial data of various sizes, because it is not closed under margins (not reproducible). In 1992, Liang, Zeger and Qaqish [25] proposed a method (known as GEE2) for analyzing multivariate binary responses which is based on estimating equations for regression and dependence parameters. They considered odds–ratios as the dependence parameters. Given the correct model specification of the mean function and the dependence structure, the GEE2 can provide consistent estimators for the parameters defined for the mean function and the dependence structure simultaneously. Liang and Beaty [24] used the GEE2 method for examining the degree of familial aggregation for binary responses. Because the GEE2 method considers odds–ratio as the dependence parameter, it turns out that it implicitly uses the multivariate logistic model of Molenberghs and Lesaffre, but are using estimating equations that are easier to compute compared with the maximum likelihood equations.

The objective of this thesis is to compare the likelihood based and estimating equation based procedures for the multivariate logistic model based on cross–product ratios. For this comparison, we first review the theoretical development of these models and then consider a simulation study to examine their performance in analyzing data. In comparing these models, the effect of the family sizes and the strength of the within–family dependence are also examined. In the subsequent three chapters the multivariate probit model, the multivariate logistic model, and the estimation equation based methods are described. In Chapter 5, the results of the simulation study is given. A brief description of the numerical optimization methods used for fitting these models is given in Chapter 5.

Chapter 2

The Multivariate Probit Model

Probit analysis (Finney [8]) is a technique which is commonly used to study the dose-response relationship in a population of biological organisms. In dose-response study, different levels of a stimulus (e.g. a vitamin, a drug, etc.) are applied to a randomly selected group of subjects and the action of a particular level of stimulus is assessed in terms of a quantal responses which depend on the intensity of the stimulus. For any subject, there will be a certain level of intensity below which the response does not occur and above which the response occurs. This level of intensity is known as tolerance or threshold which will vary from subject to subject in the population.

The main objective of the dose-response study is to assess the relationship between the levels of stimulus and the probability of occurrence of the response. If the distribution of the tolerance is assumed to be normal normal, a regression technique of probit analysis can assess this relationship after controlling for important covariates.

In a dose-response study the response can also be multivariate. In some studies for a specific level of the stimulus, in addition to the responses regarding the main effect responses about some side effects might also be of interest in some studies. In this situation to assess the efficacy of the stimulus, analyzing the responses simultaneously would be the most efficient procedure. Ashford and Sowden [2] used the multivariate probit model in the dose-response context.

The multivariate probit model is also considered in quantitative genetics. Mendell and Elston [30] used the multivariate probit model to analyze multi–factorial qualitative traits. In genetics, the term liability is used analogously to tolerance in the dose-response study. For any subject the trait will show up if the liability is smaller (or greater depending on the relationship between the trait and the liability) than a specific threshold value. It is assumed that there is an underlying normal distribution of the liability and different thresholds of this distribution provides the response in terms of qualitative traits. The normal distribution is reasonable because the liability variable may be considered as the sum of many small effects in a polygenic model (response is influenced by many genes).

Ochi and Prentice [33] considered multivariate probit models with exchangeable correlation structure. Lesaffre and Molenberghs [22] used multivariate probit models to analyze multivariate ordinal categorical response. A detailed discussion of the general class of multivariate probit models can be found in Joe [17]. In the following sections, probit models are described for univariate, bivariate, and multivariate binary responses respectively.

2.1 Univariate Probit Model

Let us consider a study where K subjects are randomly selected from a population and each subject is examined to collect information about the presence of a specific qualitative trait of interest. The focus of the study is to identify important covariates for the occurrence of the trait (we are assuming the selected subjects are independent for the time being). The response y_i ($i = 1, 2, \ldots, K$) corresponding to the i^{th} subject is a binary variable, where $y_i = 1$ if the trait is present and otherwise $y_i = 0$. Let $\boldsymbol{X}$ be the design matrix of order $K \times (p+1)$, with the first column of $\boldsymbol{X}$ being $\mathbf{1}$ to accommodate an intercept term.

Suppose the random variable Z_i, which denotes the latent variable for the i^{th} subject, follows a normal distribution with mean 0 and variance 1. Let the response y_i be the realization of the random variable Y_i, such that

$$Y_i = I(Z_i \leq \boldsymbol{X}_i \boldsymbol{\beta}), \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1 \ldots, \beta_p)^T$ is the vector of regression parameters corresponding to the design matrix $\boldsymbol{X}$, $\boldsymbol{X}_i$ is the i^{th} row of $\boldsymbol{X}$, and $I(\cdot)$ is an indicator function such that

$$I(A) = \begin{cases} 1, & \text{if } A \text{ is true,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

The probability that the i^{th} subject has the trait is

$$\pi_i(\boldsymbol{\beta}) = \Pr(Y_i = 1) = \Pr(Z_i \leq \boldsymbol{X}_i \boldsymbol{\beta}) = \Phi(\boldsymbol{X}_i \boldsymbol{\beta}), \tag{2.3}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. For the given sample $y_1, y_2, \ldots, y_K$, the log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{K} \{y_i \log \pi_i(\boldsymbol{\beta}) + (1 - y_i) \log(1 - \pi_i(\boldsymbol{\beta}))\}. \tag{2.4}$$

The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is a solution of the score equations $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$, where the j^{th} $(j = 1, 2, \ldots, p)$ element of the score vector $\boldsymbol{U}(\boldsymbol{\beta})$ is

$$\begin{aligned} U_j(\boldsymbol{\beta}) &= \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} \\ &= \sum_{i=1}^{K} \left\{ \frac{y_i}{\Phi(\boldsymbol{X}_i \boldsymbol{\beta})} - \frac{1 - y_i}{1 - \Phi(\boldsymbol{X}_i \boldsymbol{\beta})} \right\} \phi(\boldsymbol{X}_i \boldsymbol{\beta}) X_{ij}, \end{aligned} \tag{2.5}$$

and $\phi(\cdot)$ is the probability density function of the standard normal distribution. The covariance matrix of $\hat{\boldsymbol{\beta}}$ can be obtained from the Fisher information matrix. The (j, k) element of the observed Fisher Information matrix $I(\boldsymbol{\beta})$ is

$$I_{jk}(\boldsymbol{\beta}) = \frac{-\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \, \partial \beta_k}$$

$$
\begin{aligned}
= & \sum_{i=1}^{K}\left\{\frac{y_i}{\Phi^2(\boldsymbol{X}_i\boldsymbol{\beta})}+\frac{1-y_i}{\{1-\Phi(\boldsymbol{X}_i\boldsymbol{\beta})\}^2}\right\}\phi^2(\boldsymbol{X}_i\boldsymbol{\beta})X_{ij}X_{ik} \\
& -\sum_{i=1}^{K}\left\{\frac{y_i}{\Phi(\boldsymbol{X}_i\boldsymbol{\beta})}-\frac{1-y_i}{1-\Phi(\boldsymbol{X}_i\boldsymbol{\beta})}\right\}\phi(\boldsymbol{X}_i\boldsymbol{\beta})X_{ij}X_{ik}. \qquad (2.6)
\end{aligned}
$$

The estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ can be obtained as $I(\hat{\boldsymbol{\beta}})^{-1}$. In the univariate probit model the score function and the elements of the Fisher information matrix can be written in simple convenient form, so an iterative procedure such as the Newton-Raphson procedure can be used to estimate the parameters of the model.

2.2 Bivariate Probit Model

In the previous section to describe univariate probit model, we considered a hypothetical study where each of the K independent units provide binary responses about a specific qualitative trait of interest. But there might be a situation where binary responses are available for the members of a family; for example, we could consider a situation where the responses are available for each father–son pair of K randomly selected families from a population. Besides identifying the important covariates for the occurrence of the trait, the objective of this study is to estimate the association of the occurrence of the trait between the father and the son in the father-son pairs. In this case, the responses between families are independent but within–family responses are correlated. So the univariate probit model is not appropriate for this problem.

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2})^T$ and $\boldsymbol{X}_i$ be the response vector and and covariate matrix of order $2 \times (p+1)$ corresponding to the i^{th} $(i = 1, 2, \ldots, K)$ family. As the binary responses are bivariate, the corresponding distribution of the latent variable is also bivariate. Suppose the latent vector $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2})^T$ follows a bivariate standard normal distribution with correlation coefficient ρ. That means the marginal distribution of Z_{ij} $(j = 1, 2)$ corresponding to the i^{th} family follows standard normal distribution and $\text{corr}(Z_{i1}, Z_{i2}) = \rho$.

The observed responses are realizations of the random variables Y_{ij}, where

$$Y_{ij} = I(Z_{ij} \leq \boldsymbol{X}_{ij}\boldsymbol{\beta}),\ j = 1, 2.$$

The marginal probabilities corresponding to the occurrence of the trait are

$$\begin{aligned} \pi_{i1\cdot}(\boldsymbol{\beta}) = \Pr(y_{i1} = 1) &= \Pr(Z_{i1} \leq \boldsymbol{X}_{i1}\boldsymbol{\beta}) = \Phi(\boldsymbol{X}_{i1}\boldsymbol{\beta}), \\ \pi_{i\cdot 1}(\boldsymbol{\beta}) = \Pr(y_{i2} = 1) &= \Pr(Z_{i2} \leq \boldsymbol{X}_{i2}\boldsymbol{\beta}) = \Phi(\boldsymbol{X}_{i2}\boldsymbol{\beta}), \end{aligned} \tag{2.7}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The other marginal probabilities are $\pi_{i0\cdot} = 1 - \pi_{i1\cdot}$ and $\pi_{i\cdot 0} = 1 - \pi_{i\cdot 1}$. The bivariate probability of both the members of the family has the trait is

$$\begin{aligned} \pi_{i11}(\boldsymbol{\beta}, \rho) &= \Pr(Y_{i1} = 1, Y_{i2} = 1) \\ &= \Pr(Z_{i1} \leq \boldsymbol{X}_{i1}\boldsymbol{\beta}, Z_{i2} \leq \boldsymbol{X}_{i2}\boldsymbol{\beta}) \\ &= \Phi_2(\boldsymbol{X}_{i1}\boldsymbol{\beta}, \boldsymbol{X}_{i2}\boldsymbol{\beta}; \rho), \end{aligned} \tag{2.8}$$

where $\Phi_2(\cdot, \cdot; \cdot)$ is the cumulative distribution function of the bivariate standard normal distribution. Similarly, we can write the other bivariate probabilities as

$$\begin{aligned} \pi_{i10}(\boldsymbol{\beta}, \rho) &= \Pr(Y_{i1} = 1, Y_{i2} = 0) \\ &= \Pr(Z_{i1} \leq \boldsymbol{X}_{i1}\boldsymbol{\beta}, Z_{i2} > \boldsymbol{X}_{i2}\boldsymbol{\beta}) \\ &= \Pr(Z_{i1} \leq \boldsymbol{X}_{i1}\boldsymbol{\beta}) - \Pr(Z_{i1} \leq \boldsymbol{X}_{i1}\boldsymbol{\beta}, Z_{i2} \leq \boldsymbol{X}_{i2}\boldsymbol{\beta}) \\ &= \pi_{i1\cdot}(\boldsymbol{\beta}) - \pi_{i11}(\boldsymbol{\beta}, \rho), \\ \pi_{i01}(\boldsymbol{\beta}, \rho) &= \pi_{i\cdot 1}(\boldsymbol{\beta}) - \pi_{i11}(\boldsymbol{\beta}, \rho), \\ \pi_{i00}(\boldsymbol{\beta}, \rho) &= 1 - \pi_{i11}(\boldsymbol{\beta}, \rho) - \pi_{i10}(\boldsymbol{\beta}, \rho) - \pi_{i01}(\boldsymbol{\beta}, \rho). \end{aligned}$$

To construct the likelihood function, let us consider a 2×2 contingency table M_i for the i^{th} response vector $\boldsymbol{y}_i$, with m_{ijk} $(j, k = 0, 1)$ be the number of observations in the (j, k) cell. If there is only one observation per subject within any family then only one cell of M_i will take the value 1 and the remaining cells take the value 0. Throughout this thesis we assume that there is only one observation per member of any family. The likelihood function for the sample of size K can be written as

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{K} \sum_{j_1=0}^{1} \sum_{j_2=0}^{1} m_{ij_1j_2} \log \pi_{ij_1j_2}(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{K} \sum_{\mathbf{j}} m_{i\mathbf{j}} \log \pi_{i\mathbf{j}}(\boldsymbol{\theta}), \end{aligned} \tag{2.9}$$

where $\boldsymbol{j}$ indicates a multi-index $\boldsymbol{j} = \{(j_1, j_2) : j_1, j_2 = 0, 1\}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \rho)^T$. The vector of score function for the parameters can be written as

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial\, l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{K} \sum_{\mathbf{j}} \frac{m_{i\mathbf{j}}}{\pi_{i\mathbf{j}}(\boldsymbol{\theta})} \frac{\partial \pi_{i\mathbf{j}}(\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}. \tag{2.10}$$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ can be obtained from the solution of the equation $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$. The elements of the score vector and Fisher information matrix can be expressed in terms of probability density function (pdf) and cumulative distribution function (cdf) of the standardized univariate and bivariate normal distributions.

2.3 Multivariate Probit Model

The multivariate probit model is an extension of the bivariate probit model, considering the underlying distribution of the latent vector as multivariate normal. When binary responses are available for more than two members of a family, multivariate probit models can be used. In practice the family sizes can be unequal but for notational simplicity throughout this thesis, we will consider equal family sizes for deriving multivariate methods. Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{id})^T$ be the binary response vector corresponding to the i^{th} $(i = 1, 2, \ldots, K)$ family and $\boldsymbol{X}_{ij}$ be the $1 \times (p+1)$ covariate vector corresponding to the j^{th} $(j = 1, 2, \ldots, d)$ member of the i^{th} family. Let $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{id})^T$ be the latent vector corresponding to the i^{th} family which is assumed to follow a multivariate standard normal distribution with correlation matrix R.

Let us assume the observed responses are the realizations of the random variables

$$Y_{ij} \;=\; I(Z_{ij} \le \boldsymbol{X}_{ij}\boldsymbol{\beta}).$$

Assume that there are q different types of relative pairs (e.g. father-offspring, sibling-sibling, etc.) within a family, $q \le d(d-1)/2$. Then we can express the correlation matrix as $R(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_q)^T$ is the vector of correlation parameters corresponding to the different types of relative pairs within families.

For a response vector of dimension d, 2^1 univariate, 2^2 bivariate, 2^3 3-variate, ..., 2^d d-variate probabilities can be obtained. These probabilities can be expressed in terms of the cumulative distribution function of the multivariate normal distribution. We can express the marginal probabilities

$$\begin{aligned}\pi_{i1\cdot\cdot\;\cdot}(\boldsymbol{\beta}) &= \Pr(Y_{ij_1} = 1)\\ &= \Phi_1(\boldsymbol{X}_{ij_1}\boldsymbol{\beta}),\end{aligned} \tag{2.11}$$

the bivariate probabilities

$$\begin{aligned}\pi_{i11\cdot\cdot\;\cdot}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \Pr(Y_{ij_1} = 1, Y_{ij_2} = 1), \quad j_1 \neq j_2\\ &= \Phi_2(\boldsymbol{X}_{ij_1}\boldsymbol{\beta}, \boldsymbol{X}_{ij_2}\boldsymbol{\beta}; R_{j_1j_2}(\boldsymbol{\alpha})),\end{aligned} \tag{2.12}$$

the 3-variate probabilities

$$\begin{aligned}\pi_{i111\cdot\cdot\;\cdot}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \Pr(Y_{ij_1} = 1, Y_{ij_2} = 1, Y_{ij_3} = 1), \quad j_1 \neq j_2 \neq j_3\\ &= \Phi_3(\boldsymbol{X}_{ij_1}\boldsymbol{\beta},\, \boldsymbol{X}_{ij_2}\boldsymbol{\beta},\, \boldsymbol{X}_{ij_3}\boldsymbol{\beta}; R_{j_1j_2j_3}(\boldsymbol{\alpha})),\end{aligned} \tag{2.13}$$

and in general the d-variate probabilities

$$\begin{aligned}\pi_{i11\cdots1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \Pr(Y_{ij_1} = 1, Y_{ij_2} = 1, \ldots, Y_{ij_d} = 1),\, j_1 \neq j_2 \neq \cdots \neq j_d\\ &= \Phi_d(\boldsymbol{X}_{ij_1}\boldsymbol{\beta}, \boldsymbol{X}_{ij_2}\boldsymbol{\beta}, \ldots, \boldsymbol{X}_{ij_d}\boldsymbol{\beta}; R_{j_1\cdots j_d}(\boldsymbol{\alpha})),\end{aligned} \tag{2.14}$$

where $\Phi_j(\cdot;\Sigma)$ is the cumulative distribution function corresponding to the j-variate $(j = 1, 2, \ldots, d)$ standard normal distribution with correlation matrix Σ. Once we have these probabilities, we can compute the remaining orthant probabilities.

As in the bivariate case, let M_i be a $2 \times 2 \times \cdots \times 2$ $(= 2^d)$ contingency table corresponding to the $d-$dimensional response vector $\boldsymbol{y}_i$. Let $m_{ij_1 \ldots j_d}$ be the number of observations corresponding to the $(j_1, j_2, \ldots, j_d)$ cell of M_i. The log-likelihood function can be written as

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{K} \sum_{j_1=0}^{1} \cdots \sum_{j_d=0}^{1} m_{ij_1 \ldots j_d} \log \pi_{ij_1 \ldots j_d}(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{K} \sum_{\mathbf{j}} m_{i\mathbf{j}} \log \pi_{i\mathbf{j}}(\boldsymbol{\theta}), \end{aligned} \tag{2.15}$$

where $\boldsymbol{j}$ indicates a multi-index $\boldsymbol{j} = \{(j_1, j_2, \ldots, j_d) : j_r \in \{0, 1\},\ r = 1, 2, \ldots, d\}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is the solution of the equations $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$, where

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial\, l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{K} \sum_{\mathbf{j}} \frac{m_{i\mathbf{j}}}{\pi_{i\mathbf{j}}(\boldsymbol{\theta})} \frac{\partial \pi_{i\mathbf{j}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{2.16}$$

is the vector of the score functions. The equations (2.15) and (2.16) are the general form of the corresponding equations (2.10) and (2.9) of the bivariate case.

As the score function $\boldsymbol{U}(\boldsymbol{\theta})$ contains multi–dimensional integrals, the second derivative of the score function (i.e. elements of Fisher information matrix) is very difficult to write down. Numerical integration can be used to approximate these high–dimensional integrals; we used the approximations proposed in Joe [15].

Since for the multivariate probit model, it is very difficult to evaluate the Hessian matrix analytically; the classical Newton-Raphson method is no longer useful as an optimization method for estimating the parameters of the model. Numerical methods, which can apply without having the Hessian matrix analytically, are required for this case. A brief description of these methods will be given in Chapter 5.

2.4 Conditional Probabilities

In the preceding sections, we discussed the multivariate probit model which can be used for identifying important covariates for the occurrence of a disease of interest. This model can be used to estimate the association of the occurrence of the disease between the relatives. The multivariate probit model can also be used to predict the disease status of an individual given the disease status and the covariate values of that individual's relatives. Predicting future disease status has significant importance in genetics.

In this section, we discuss a procedure to predict an individual's future disease status given the information about his relatives. Let us define

$$\begin{aligned} p(d \mid 1, 2, \ldots, d-1) &= \Pr(Y_d = 1 \mid Y_1 = y_1, Y_2 = y_2, \ldots, Y_{d-1} = y_{d-1}, \boldsymbol{X}) \\ &= \frac{\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_d = 1, \boldsymbol{X})}{\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_{d-1} = y_{d-1}, \boldsymbol{X})}, \end{aligned} \tag{2.17}$$

where $p(d \mid 1, 2, \ldots, d-1)$ is the conditional probability that the d^{th} member of the family will have the disease given the current disease status of the other members of the family and the covariate values. This conditional probability is the ratio of orthant probabilities of order d and $d-1$ respectively. The equation (2.14) shows that these orthant probabilities are functions of $\boldsymbol{\theta} = (\beta^T, \boldsymbol{\alpha}^T)^T$, where $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the regression and association parameters defined for the multivariate probit model. So for the multivariate probit model, the conditional probability can be written as

$$\begin{aligned} p(d \mid 1, 2, \ldots, d-1) &= \frac{\Phi_d(\boldsymbol{X}_1 \boldsymbol{\beta}, \ldots, \boldsymbol{X}_d \boldsymbol{\beta}; R_{1 \cdots d}(\boldsymbol{\alpha}))}{\Phi_{d-1}(\boldsymbol{X}_1 \boldsymbol{\beta}, \ldots, \boldsymbol{X}_{d-1} \boldsymbol{\beta}; R_{1 \cdots d-1}(\boldsymbol{\alpha}))} \\ &= q(\boldsymbol{\theta}). \end{aligned} \tag{2.18}$$

Given the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ and the covariate values, the maximum likelihood estimate of the conditional probability $p(d \mid 1, 2, \ldots, d-1)$ can be obtained from (2.18) as $q(\hat{\boldsymbol{\theta}})$.

The conditional probability is a function of the parameters $\boldsymbol{\theta}$, we can approximate the variance of the maximum likelihood estimate of the conditional probability by using delta method. The approximate expression for the variance of the estimate of $p(d \mid 1, 2, \ldots, d-1)$ can be written as

$$\text{var}\{q(\hat{\boldsymbol{\theta}})\} = \frac{\partial q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} V(\boldsymbol{\theta}) \frac{\partial q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \qquad (2.19)$$

where $V(\boldsymbol{\theta})$ is the covariance matrix of $\hat{\boldsymbol{\theta}}$. The estimate of the variance can be obtained by replacing $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ in (2.19). Because of the complex form of $q(\boldsymbol{\theta})$, the partial derivatives of the equation (2.19) is very difficult to be obtained analytically. We have used numerical methods to compute the conditional probabilities and the respective standard errors, which are shown in Chapter 5.

2.5 Summary

In this chapter different types of probit models are described which can be used to analyze binary responses. The multivariate probit models are very useful in analyzing correlated binary responses. Maximum likelihood estimates for the regression parameters of the model can be obtained. The estimates of the association between the members of the family in terms of the latent correlation coefficient, also known as the tetrachloric correlation, can also be obtained. The multivariate probit model can estimate the conditional probabilities that a particular member of a family has the disease given the current status of the other members of the family and the covariate values, which has importance in genetics.

Chapter 3

The Multivariate Logistic Model

In the previous chapter, the multivariate probit model was described. This chapter contains the description of an analogous multivariate logistic model. This model is also a latent variable model and has univariate margins that regress each response on the covariate vector. The multivariate logistic model models the dependence parameter in terms of the cross–product ratio which has an attractive interpretation. However there is no physical or stochastic model that leads naturally to the cross–product ratio as the natural dependence parameter. For the multivariate logistic and probit models the underlying univariate margins are assumed to be logistic and normal respectively.

This approach of modeling multivariate binary data is based on a class of bivariate distributions proposed by Plackett [34]. For given univariate margins and cross–product ratios as dependence parameter, the Plackett distribution can completely specify the joint latent distribution of bivariate binary responses. Mardia [28] studied the properties of the bivariate Plackett distribution.

Dale [5] first considered the bivariate Plackett distribution to model bivariate categorical responses; this model is known as bivariate Dale model in the literature. Molenberghs and Lesaffre [31] introduced the multivariate Plackett distribution and extended the Dale model to multivariate ordinal categorical responses. In the Dale model, the global cross–product ratio is considered as the dependence parameter which is equivalent to the local cross–product ratio for the binary case. The multivariate version of the Dale model is defined for any underlying univariate continuous distribution. For this model, different link functions can be considered for modeling the univariate margins and the cross–product ratios. For the multivariate logistic distribution, the underlying univariate margins are logistic and the multivariate Plackett distribution is used to construct higher order margins with the univariate logistic margins. Joe [17] studied the multivariate Plackett construction within a more general theoretical framework and considered different data sets to show the applications of the multivariate logistic model. Besides the Plackett distribution based approach to multivariate logistic models, McCullagh and Nelder [29] described this model differently based on the logistic transformation of linear combinations of joint probabilities. Their approach does not consider any latent variables. Glonek and McCullagh [11] also studied this approach of modelling the multivariate logistic model.

The main focus of this chapter is to describe the multivariate logistic models for analyzing multivariate binary response. The Plackett distribution plays a vital role in defining the multivariate logistic models that we consider, so we first discuss the bivariate and multivariate Plackett distribution in Section 3.1. The subsequent two sections contain the model description and methods of estimating parameters of this multivariate logistic model. In Section 3.3, a brief description of the McCullagh–Nelder–Glonek approach is given.

3.1 Plackett Distributions

In this section, a brief introduction of bivariate and multivariate Plackett distribution is given. Throughout this chapter we will consider the binary responses as taking either 1 (diseased) or 2 (non–diseased) instead of 1/0 which we considered in the previous chapters. This will help us to derive general expression of cumulative distribution function of the multivariate Plackett construction.

3.1.1 The Bivariate Plackett Distribution

In 1965, Plackett proposed a procedure for constructing a class of bivariate distributions. Suppose Z_1 and Z_2 are two continuous random variables and $F_1(z_1) = \Pr(Z_1 \leq z_1)$ and $F_2(z_2) = \Pr(Z_2 \leq z_2)$ are the univariate margins of Z_1 and Z_2 respectively. Plackett [34] considered $F_{12}(z_1, z_2) = \Pr(Z_1 \leq z_1, Z_2 \leq z_2)$, the possible joint bivariate distribution function of Z_1 and Z_2, as the solution of the equation

$$\gamma = \frac{F_{12}(F_{12} - F_1 - F_2 + 1)}{(F_1 - F_{12})(F_2 - F_{12})}, \tag{3.1}$$

where $F_{12} = F_{12}(z_1, z_2)$, $F_1 = F_1(z_1)$, $F_2 = F_2(z_2)$ and the cross–product ratio or odds–ratio γ is a positive constant for all (z_1, z_2) for which neither F_1 nor F_2 assumes the value 0 or 1. The equation (3.1) is known as defining equation. Mardia [28] showed that only the following solution of the defining equation (3.1)

$$F_{12} = \begin{cases} (1/2)(\gamma - 1)^{-1}\{1 - (F_1 + F_2)(1 - \gamma) - S(F_1, F_2, \gamma)\}, & \text{if } \gamma \neq 1 \\ F_1 F_2, & \text{if } \gamma = 1, \end{cases} \tag{3.2}$$

where

$$S(F_1, F_2, \gamma) = \left[\{1 - (F_1 + F_2)(1 - \gamma)\}^2 + 4\gamma(1 - \gamma)F_1 F_2\right]^{1/2} \tag{3.3}$$

is the only root leading to a proper bivariate distribution. The corresponding joint density function of Z_1 and Z_2 can be written as

$$f_{12}(z_1, z_2) = \frac{\partial^2 F_{12}(z_1, z_2)}{\partial z_1 \, \partial z_2} = \frac{\gamma f_1 f_2 \{(\gamma - 1)(F_1 + F_2 - 2F_1 F_2) + 1\}}{S^3}, \quad \gamma > 0, \tag{3.4}$$

where f_1 and f_2 are the univariate marginal density functions of Z_1 and Z_2 respectively and S is given in (3.3).

For given univariate margins F_1 and F_2, the dependence parameter of the bivariate Plackett distribution (the cross-product ratio γ) is a monotonic increasing function in F_{12}, i.e. $\gamma = 0$ when $F_{12} = F_L$ and $\gamma = \infty$ when $F_{12} = F_U$, where

$$F_U(z_1, z_2) = \min\{F_1(z_1), F_2(z_2)\}, \quad \text{and} \quad F_L(z_1, z_2) = \max\{F_1(z_1) + F_2(z_2) - 1, 0\},$$

are known as upper and lower Fréchet bounds (see [17]) respectively. From equation (3.2) the following can also be seen

(i) If F_1 (F_2) tend to 1 then F_{12} tends to F_2 (F_1).

(ii) If F_1 and F_2 tend to 1 then F_{12} also tends to 1 which indicates that F_1 and F_2 are marginal distribution functions.

(iii) For fixed F_1 (F_2) and γ, F_{12} increases with F_2 (F_1).

Bivariate Plackett–normal vs Bivariate Normal

The bivariate Plackett distribution is defined for continuous random variables with arbitrary univariate margins. The most widely used bivariate distribution is the bivariate normal, so it is of interest to examine how the bivariate Plackett distribution with univariate standard normal margins (which is known as bivariate Plackett–normal distribution) resembles the usual bivariate standard normal distribution.

To compare the bivariate standard normal and the bivariate Plackett–normal distributions, the relationship between the cross–product ratio (γ) and the correlation coefficient (ρ), the dependence parameters of these two distributions respectively, is required. For a specific cut–off point (z_1, z_2) the cross–product ratio can be written from equation (3.1) as

$$\gamma(z_1, z_2; \rho) = \frac{\Phi_2(z_1, z_2; \rho)\Phi_2(-z_1, -z_2; \rho)}{\{\Phi(z_1) - \Phi_2(z_1, z_2; \rho)\}\{\Phi(z_2) - \Phi_2(z_1, z_2; \rho)\}}. \qquad (3.5)$$

The quantity on the right hand side of equation (3.5) depends on the bivariate normal orthant probabilities corresponding to the cut-off point (z_1, z_2). Kepner *et al.* [21] derived the following expression for the orthant probability for the cut–off point $(0,0)$

$$\Phi_2(0,0;\rho) \quad = \quad 1/4 + (2\pi)^{-1}\sin^{-1}\rho.$$

Using this result in equation (3.5), we get the following relationship between the cross–product ratio and the correlation coefficient for the cut–off point $(0,0)$:

$$\gamma(0,0;\rho) \quad = \quad \left\{\frac{1+(2/\pi)\arcsin\rho}{1-(2/\pi)\arcsin\rho}\right\}^2. \tag{3.6}$$

We use this relationship to obtain the cross–product ratio at the cut–off point $(0,0)$ from a given value of the correlation coefficient.

For a given correlation coefficient ρ, different cross–product ratios can be obtained from the equation (3.5) for different cut-off points (z_1, z_2). Numerically it can be shown that

$$\gamma(0,0;\rho) \quad = \quad \min_{z_1,z_2} \quad \gamma(z_1,z_2;\rho), \tag{3.7}$$

i.e. the lower bound of the cross–product ratios is attained at the cut-off point $(0,0)$. The theoretical proof of this result is still an open problem.

Figure 3.1 shows contour plots of the bivariate standard normal density for selected values of correlation coefficient ρ. The corresponding plot for the Plackett–normal distribution is shown in Figure 3.2, where the identical cross–product ratios are obtained from the correlation coefficient values (used to generate plots for bivariate normal) by using the equation (3.6). These sets of plots are almost similar which indicates that the bivariate Plackett distribution with univariate standard normal margins is similar to the bivariate normal distribution.

3.1.2 The 3-variate Plackett Construction

Let Z_1, Z_2, and Z_3 be three continuous random variables with univariate margins F_1, F_2, and F_3 respectively. Let us consider $F_{j_1j_2}$ $(1 \leq j_1 < j_2 \leq 3)$ as the bivariate Plackett margin corresponding to

(Z_{j_1}, Z_{j_2}). Given the univariate and bivariate margins, the 3-variate margin $F_{123} = F_{123}(z_1, z_2, z_3)$ is a solution of the defining equation

$$\gamma_{123} \;=\; \frac{F_{123}(F_{123} - a_1)(F_{123} - a_2)(F_{123} - a_3)}{(b_1 - F_{123})(b_2 - F_{123})(b_3 - F_{123})(b_4 - F_{123})}, \tag{3.8}$$

where $a_1 = F_{12} + F_{13} - F_1$, $a_2 = F_{12} + F_{13} - F_2$, $a_3 = F_{13} + F_{23} - F_3$, $b_1 = F_{12}$, $b_2 = F_{13}$, $b_3 = F_{23}$, $b_4 = 1 - \sum_i F_i + \sum_{i<j} F_{ij}$. The function F_{123} satisfies the Fréchet bounds (Joe, 1997),

$$F_U = \min\{b_1, b_2, b_3, b_4\}, \quad \text{and} \quad F_L = \max\{a_1, a_2, a_3, 0\}.$$

Because $F_{123}(z_1, z_2, z_3)$ is defined implicitly for each (z_1, z_2, z_3), it is difficult to check if F_{123} satisfies the rectangle condition necessary for a proper cumulative distribution function.

Interpretation of the parameters

Let Y_1, Y_2, and Y_3 are three independent Bernoulli variables. The third–order dependence parameter of the 3-variate Plackett distribution can be expressed as

$$\begin{aligned} \gamma_{123} &= CR_3(Y_1, Y_2, Y_3) \\ &= \frac{CR_2(Y_1, Y_2 \mid Y_3 = 1)}{CR_2(Y_1, Y_2 \mid Y_3 = 2)} \\ &= \frac{\pi_{111}\,\pi_{221}\,\pi_{122}\,\pi_{212}}{\pi_{112}\,\pi_{222}\,\pi_{121}\,\pi_{211}}. \end{aligned}$$

Now by comparing this with the equation (3.8), the orthant probabilities can be expressed in terms of the 3-variate joint distribution function as

$$\pi_{111} = F_{123},\ \pi_{122} = F_{123} - a_1,\ \pi_{212} = F_{123} - a_2,\ \pi_{221} = F_{123} - a_3,$$

$$\pi_{112} = b_1 - F_{123},\ \pi_{121} = b_2 - F_{123},\ \pi_{211} = b_3 - F_{123},\ \pi_{222} = b_4 - F_{123}.$$

Given the dependence parameters γ_{12}, γ_{123} and the univariate and bivariate margins, these orthant probabilities can be computed from the 3-variate Plackett distribution function.

3.1.3 The d–variate Plackett Construction

In this section, we describe the multivariate Plackett distribution for arbitrary dimension d ($d > 1$). To express the general form of the joint distribution and the cross–product ratios, we use the notation of Molenberghs and Lesaffre [31]. Let $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_d)^T$ be the d–dimensional random vector and F_{j_1} $(1 \leq j_1 \leq 2)$, $F_{j_1 j_2}$ $(1 \leq j_1 < j_2 \leq 2)$, $\ldots$, $F_{j_1 j_2 \cdots j_{d-1}}$ $(1 \leq j_1 < \cdots < j_{d-1} \leq 2)$ be the univariate, bivariate,$\ldots$, $(d-1)$–variate Plackett distribution functions of $\boldsymbol{Z}$ respectively. Let $\pi_{j_1 j_2 \cdots j_d} = \Pr(Y_1 = j_1, Y_2 = j_2, \ldots, Y_d = j_d)$ be the d–dimensional orthant probabilities.

The general expression for the d–dimensional cross–product ratio can be written as

$$\gamma_{12\cdots d} = \frac{\prod_{(j_1,\ldots,j_d)\in A_d^+} \pi_{j_1\cdots j_d}}{\prod_{(j_1,\ldots,j_d)\in A_d^-} \pi_{j_1\cdots j_d}}, \tag{3.9}$$

where

$$\begin{aligned} A_d^+ &= \left\{(j_1, j_2, \ldots, j_d) \in \{1,2\}^d \ : \ \sum_{l=1}^{d} j_l - d \ \text{ is even}\right\}, \\ A_d^- &= \{1,2\}^d \backslash A_d^+. \end{aligned}$$

For example,

$$\begin{aligned} &\text{when} \quad d = 2, \quad A_2^+ = \{(1,1),(2,2)\}, \\ &\text{when} \quad d = 3, \quad A_3^+ = \{(1,1,1),(1,2,2),(2,1,2),(2,2,1)\}, \\ &\text{when} \quad d = 4, \quad A_4^+ = \{(1,1,1,1),(1,1,2,2),(1,2,1,2),(1,2,2,1), \\ &\qquad\qquad\qquad\qquad (2,1,2,1),(2,2,1,1),(2,1,1,2),(2,2,2,2)\}, \\ &\qquad \vdots \end{aligned}$$

Now by using these elements of the sets A_d^+ and A_d^- in equation (3.9), we can get the cross–product ratios

$$\gamma_{12} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}, \quad \gamma_{123} = \frac{\pi_{111}\pi_{122}\pi_{212}\pi_{221}}{\pi_{112}\pi_{121}\pi_{211}\pi_{222}}, \ \cdots.$$

To get the general expression of the defining equation corresponding to the d-variate Plackett distribution, the orthant probabilities in equation (3.9) need to be expressed in terms of the marginal distributions of different dimensions. Suppose the multi–index $\boldsymbol{j}$ represents a d–dimensional vector of 1's and 2's, i.e. $\boldsymbol{j} \in \{1,2\}^d$. Let $\kappa(\boldsymbol{j}) = \kappa(j_1, j_2, \ldots, j_d)$ be the set of dimensions for which $j_l = 1$ $(1 \leq l \leq d)$, i.e.

$$\kappa(\boldsymbol{j}) \quad = \quad \{l : 1 \leq l \leq d, j_l = 1\}.$$

For example, $\kappa(1,2) = \{1\}$, $\kappa(1,2,1) = \{1,3\}$, $\kappa(2,1,2,1) = \{2,4\}$, etc. Let us introduce another notation $\boldsymbol{s}(\boldsymbol{j})$: the set $\boldsymbol{s}(\boldsymbol{j})$ contains all possible subsets of $\{1, 2, \ldots, d\}$ which contain $\kappa(\boldsymbol{j})$, with the elements of $\boldsymbol{s}(\boldsymbol{j})$ are arranged in lexicographical order. For example,

$$\begin{array}{ll} d = 2, & \kappa(1,2) = \{1\} \Rightarrow \boldsymbol{s}(1,2) = \{1\}, \{1,2\}, \\ d = 3, & \kappa(2,2,1) = \{3\} \Rightarrow \boldsymbol{s}(2,2,1) = \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}, \\ d = 4, & \kappa(2,1,1,2) = \{2,3\} \Rightarrow \boldsymbol{s}(2,1,1,2) = \{2,3\}, \{1,2,3\}, \{2,3,4\}, \{1,2,3,4\}, \\ \vdots & \end{array}$$

Let N be the difference between the number of elements of the sets $\boldsymbol{s}(\boldsymbol{j})$ and $\kappa(\boldsymbol{j})$. Using the inclusion and exclusion probability law, we can define the general form of the orthant probabilities as

$$\pi_{j_1 j_2 \cdots j_d} \quad = \quad \sum_{\boldsymbol{s}(\mathbf{j})} \operatorname{sgn}(\boldsymbol{s}(\boldsymbol{j})) F_{\boldsymbol{s}(\mathbf{j})}, \tag{3.10}$$

where

$$\operatorname{sgn}(\boldsymbol{s}(\boldsymbol{j})) = \begin{cases} 1 & \text{if } N \text{ is even,} \\ -1 & \text{otherwise.} \end{cases}$$

For instance, we can write down the expressions for the orthant probabilities as

$$\pi_{12} \quad = \quad F_1 - F_{12},$$

$$
\begin{aligned}
\pi_{221} &= F_3 - F_{13} - F_{23} + F_{123}, \\
\pi_{2112} &= F_{23} - F_{123} - F_{234} + F_{1234}, \\
&\vdots
\end{aligned}
$$

Using equation (3.10), we can rewrite the equation (3.9) as

$$
\begin{aligned}
\gamma_{12\cdots d} &= \frac{\prod_{\mathbf{j}\in A_d^+} \pi_{\mathbf{j}}}{\prod_{\mathbf{j}\in A_d^-} \pi_{\mathbf{j}}} = \frac{\prod_{\mathbf{j}\in A_d^+} \sum_{\boldsymbol{s}(\mathbf{j})} \operatorname{sgn}(\boldsymbol{s}(\boldsymbol{j})) F_{\boldsymbol{s}(\mathbf{j})}}{\prod_{\mathbf{j}\in A_d^-} \sum_{\boldsymbol{s}(\mathbf{j})} \operatorname{sgn}(\boldsymbol{s}(\boldsymbol{j})) F_{\boldsymbol{s}(\mathbf{j})}} \\
&= \frac{\prod_{\mathbf{j}\in A_d^+} (F - b_{\mathbf{j}})}{\prod_{\mathbf{j}\in A_d^-} (F - a_{\mathbf{j}})}, \qquad (3.11)
\end{aligned}
$$

where $F = F_{12\cdots d}$, $a_{\boldsymbol{j}}$ and $b_{\boldsymbol{j}}$ are the functions of the margins of order d' $(< d)$ which can be obtained from equation (3.10). The d–dimensional Plackett distribution function is the solution of equation (3.11) which satisfies the Fréchet bounds

$$
\left(\max_{\mathbf{j}} \{a_{\mathbf{j}}, 0\}, \min_{\mathbf{j}} \{b_{\mathbf{j}}\} \right).
$$

For third– and higher–order Plackett distributions, an open problem is whether the function F is a proper distribution function, i.e. whether the mixed derivatives are non-negative.

3.2 The Regression Model

The main objective of this study is to compare the existing estimation methods for analyzing multivariate binary responses. We wish to examine the performance of the multivariate logistic model in the field of genetics. As already discussed in the previous chapter, in genetics studies the response is sometimes binary (e.g. presence/absence of a qualitative trait of interest) and the responses are not independent within families. It is assumed that for any individual, the occurrence of the qualitative trait depends on the underlying distribution of a latent variable. Let us assume that the trait occurs if the liability is less than a predefined threshold value; otherwise it does not occur.

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{id})^T$ be the response vector corresponding to the i^{th} $(i = 1, 2, \ldots, K)$ family, where the y_{ij}'s are binary variables representing the presence $(y_{ij} = 1)$ or absence $(y_{ij} = 2)$ of the qualitative trait of interest in the j^{th} member of the family. Suppose the random vector $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{id})^T$ denotes the latent vector corresponding to the i^{th} family. Let us assume Z_{ij} follows standard logistic distribution, i.e. location parameter is 0 and scale parameter is 1. Let $\boldsymbol{X}_i$ be the covariate matrix of order $d \times (p+1)$ corresponding to the i^{th} family with a first column of $\mathbf{1}$, to accommodate an intercept term. Suppose the observed response y_{ij} is the realization of the random variable Y_{ij}, where

$$Y_{ij} \quad = \quad I(Z_{ij} \leq \boldsymbol{X}_{ij}\boldsymbol{\beta}),$$

where $\boldsymbol{X}_{ij}$ is the j^{th} row of $\boldsymbol{X}_i$ and $I(\cdot)$ is the indicator function such that

$$I(A) = \begin{cases} 1, & \text{if } A \text{ is true,} \\ 2, & \text{otherwise.} \end{cases}$$

Let us assume only the univariate margins depend on the covariates. The univariate margin corresponding to the j^{th} member of the i^{th} family $\pi_{ij} = \Pr(Y_{ij} = 1)$ can be written as a function of the unknown regression parameters $\boldsymbol{\beta}$ as

$$\pi_{ij}(\boldsymbol{\beta}) \quad = \quad \Pr(Y_{ij} = 1) = \Pr(Z_{ij} \leq \boldsymbol{X}_{ij}\,\boldsymbol{\beta}) = F_0(\boldsymbol{X}_{ij}\,\boldsymbol{\beta}), \quad 1 \leq j \leq d, \tag{3.12}$$

where $F_0(x) = 1/\{1 + e^{-x}\}$ is the distribution function of the standard logistic distribution. From equation (3.12) the linear predictor can be written as

$$\eta_{ij}(\boldsymbol{\beta}) \quad = \quad \boldsymbol{X}_{ij}\,\boldsymbol{\beta} = F_0^{-1}\left(\pi_{ij}(\boldsymbol{\beta})\right) = \log \frac{\pi_{ij}(\boldsymbol{\beta})}{1 - \pi_{ij}(\boldsymbol{\beta})}.$$

In GLM terminology the function $F_0^{-1}(\cdot)$ is known as the link function; beside this logit link, depending on the distribution of the latent variable other link functions can also be considered.

These univariate margins do not fully determine the joint distribution of $\boldsymbol{y}_i$ because the elements of $\boldsymbol{y}_i$ are not independent. This mean dependence parameters are needed to describe the association between the elements of $\boldsymbol{y}_i$. Let us assume the third– and higher–order cross–product ratios, the dependence parameters of the Plackett distribution, are constant at a fixed value $\gamma_0 = 1$. The choice of $\gamma_0 = 1$ leads to a multivariate logistic distribution that is analogous to the multivariate normal distribution (see Joe [17] for an entropy interpretation). Suppose q different types of pairs (e.g. parent–offspring, sib–sib, grandfather–grandchild, etc.) are possible in the selected families. Suppose the (j, k) pair of the i^{th} family is of the l^{th} $(1 \leq l \leq q)$ type of the pair. We can define the bivariate cross–product ratios corresponding to (j, k) pair of the i^{th} family as

$$\gamma_{ijk}(\boldsymbol{\alpha}) \quad = \quad g^{-1}(\alpha_l), \quad 1 \leq l \leq q, \tag{3.13}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_q)^T$ is the vector of the parameters corresponding to the q different types of the association and $g(\cdot)$ is the link function for the cross–product ratios.

Having values of the parameters $\boldsymbol{\theta} = (\beta^T, \boldsymbol{\alpha}^T)^T$, the univariate margins and the cross–product ratios can be computed from equations (3.12) and (3.13) respectively. Given the univariate margins and cross–product ratios corresponding to a pair of members within a family, the bivariate margins $\pi_{ij_1j_2}(\boldsymbol{\theta}) = \Pr(Y_{ij_1} = 1, Y_{ij_2} = 1)$ $(1 \leq j_1 < j_2 \leq d)$ can be computed from the equation (3.2). As we already assumed the third– and higher–order cross–product ratios are fixed at $\gamma_0 = 1$ (which is known), the 3-variate margins are the solution of equation (3.8) provided the univariate and bivariate margins are known. Similarly, higher–order margins can also be obtained by using the equation (3.9). Once all the margins are known, the orthant probabilities can be obtained from the equation (3.10).

3.2.1 The Likelihood Function

Let M_i $(i = 1, 2, \ldots, K)$ denote a $2 \times 2 \times \cdots \times 2$ $(= 2^d)$ contingency table which can be constructed from the observed response vector $\boldsymbol{y}_i$. Let $m_{i\mathbf{j}}$ be the number of observations in the $\boldsymbol{j}$th cell of the table M_i, where $\boldsymbol{j}$ indicates a multi-index $\boldsymbol{j} = (j_1, j_2, \ldots, j_d)$. Let $\pi_{i\mathbf{j}}(\boldsymbol{\theta}) = \Pr(Y_1 = j_1, Y_2, =$

$j_2, \ldots, Y_d = j_d)$ $(j_l \in \{1,2\}, 1 \leq l \leq d)$ be the orthant probabilities corresponding to the $\boldsymbol{j}$th cell of M_i.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ be the vector of parameters of the model. As for the multivariate probit model, for the given sample $\boldsymbol{y}_i$ $(i = 1, 2, \ldots, K)$, the log likelihood function can be written as

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{K} \sum_{j_1=1}^{2} \cdots \sum_{j_d=1}^{2} m_{ij_1j_2\ldots j_d} \log \pi_{ij_1j_2\ldots j_d}(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{K} \sum_{\mathbf{j}=1}^{2} m_{i\mathbf{j}} \log \pi_{i\mathbf{j}}(\boldsymbol{\theta}). \end{aligned} \tag{3.14}$$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is the solution of the score equation $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$, where the score function

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial\, l(\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} = \sum_{i=1}^{K} \sum_{\mathbf{j}=1}^{2} \frac{m_{i\mathbf{j}}}{\pi_{i\mathbf{j}}(\boldsymbol{\theta})} \frac{\partial \pi_{i\mathbf{j}}(\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}. \tag{3.15}$$

Using the chain rule we can write

$$\begin{aligned} \frac{\partial \pi_{i\mathbf{j}}}{\partial \boldsymbol{\beta}} &= \frac{\partial \pi_{i\mathbf{j}}}{\partial \boldsymbol{\eta}_i} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} = \sum_{k=1}^{d} \frac{\partial \pi_{i\mathbf{j}}}{\partial \eta_{ik}} \frac{\partial \eta_{ik}}{\partial \boldsymbol{\beta}} \\ \frac{\partial \pi_{i\mathbf{j}}}{\partial \boldsymbol{\alpha}} &= \frac{\partial \pi_{i\mathbf{j}}}{\partial \boldsymbol{\gamma}_i} \frac{\partial \boldsymbol{\gamma}_i}{\partial \boldsymbol{\alpha}} = \sum_{j=1}^{d} \sum_{k=j+1}^{d} \frac{\partial \pi_{i\mathbf{j}}}{\partial \gamma_{ijk}} \frac{\partial \gamma_{ijk}}{\partial \boldsymbol{\alpha}}. \end{aligned}$$

For the general dimension d, it is impossible to obtain the algebraic expression of the terms $(\partial \pi_{i\mathbf{j}}/\partial \eta_{ik})$ and $(\partial \pi_{i\mathbf{j}}/\partial \gamma_{ijk})$ because $\pi_{i\mathbf{j}}$ contains higher dimension cumulative distribution functions of the multivariate Plackett construction which has no closed form. So the expressions for the elements of the Fisher information matrix cannot be shown. Numerical methods are needed to get the maximum likelihood estimates of the regression parameters $\boldsymbol{\theta}$. In Chapter 5, two such numerical methods will be described.

The conditional probabilities which we discussed for the multivariate probit model (see §2.4) can also be defined for the multivariate logistic model. The conditional probability $p(d \mid 1, 2, \ldots, d-1)$, which is the ratio of two orthant probabilities, is a function of the parameters of the model. So, given the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of the multivariate logistic model, the estimate of $p(d \mid 1, 2, \ldots, d-1)$ and corresponding standard error can be obtained for this model.

3.3 McCullagh–Nelder–Glonek Approach

McCullagh and Nelder [29] introduced a multivariate logistic transformation and used this to define a class of regression models which can be applied for analyzing multivariate categorical responses. These models are also known as multivariate logistic models and are systematically studied by Glonek and McCullagh [11]. This approach of defining multivariate logistic models does not assume any underlying distribution of the univariate margins. The likelihood construction is similar for both the logistic models, but the estimating procedure of the orthant probabilities is different. In the previous sections, we developed how the multivariate Plackett distribution can be used to estimate the orthant probabilities; McCullagh and Nelder used the multivariate logistic transformation to estimate these probabilities. We have numerically checked that these two approaches give identical orthant probabilities. In the following section, inference procedure of the McCullagh–Nelder–Glonek approach is briefly described.

3.3.1 The Model and Parameter Estimation

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{id})^T$ be the response vector corresponding to the i^{th} $(i = 1, 2, \ldots, K)$ family, where the response corresponding to the j^{th} $(j = 1, 2, \ldots, d)$ member of the i^{th} family, y_{ij} is binary indicates the presence $(y_{ij} = 1)$ or absence $(y_{ij} = 2)$ of a qualitative trait of interest. As before, the objective is to estimate the effect of the covariate for the occurrence of the trait. The estimate of the dependence parameter of the occurrence of the trait between the members within a family is also of interest. Let $\boldsymbol{X}_i$ be the covariate matrix of order $d \times (p+1)$ corresponding to the members of the i^{th} family.

Let us define

$$\pi_{i\mathrm{j}}(\boldsymbol{X}_i) \;=\; \Pr(Y_{i1}=y_{i1}, Y_{i2}=y_{i2}, \ldots, Y_{id}=y_{id} \mid \boldsymbol{X}_i), \quad y_{il}=1,2, \quad l=1,2,\ldots d,$$

as the probability of observing the response vector $\boldsymbol{y}_i$ given the covariate matrix $\boldsymbol{X}_i$ and $\boldsymbol{j}$ indicates a multi–index. Let $\boldsymbol{\pi}$ be the vector of all possible 2^d probabilities. For example, for bivariate case $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^T$. If $\boldsymbol{\gamma}$ is the vector of $\binom{d}{1} \times 2^1$ univariate, $\binom{d}{2} \times 2^2$ bivariate, …, $\binom{d}{d} \times 2^d$ d–variate margins, let us make a linear transformation $\boldsymbol{\pi} \to \boldsymbol{\gamma}$ by

$$\boldsymbol{\gamma} \;=\; L\boldsymbol{\pi} \tag{3.16}$$

where L is a matrix of zeros and ones. For example, when $d = 3$, $\boldsymbol{\gamma}$ would be the vector of 6 univariate, 12 bivariate and 8 trivariate margins.

Let us define $\boldsymbol{\eta} = (\eta_0, \eta_1, \ldots, \eta_d, \eta_{12}, \ldots, \eta_{d-1,d}, \ldots, \eta_{12\cdots d})^T$ as the vector of the logistic factorial contrasts which can be obtained from $\boldsymbol{\gamma}$ by

$$\boldsymbol{\eta} \;=\; C\log(L\boldsymbol{\pi}), \tag{3.17}$$

where C is an appropriately chosen contrast matrix. The transformation $\boldsymbol{\pi} \to \boldsymbol{\eta}$ defined in equation (3.17) is called a multivariate logistic transformation by McCullagh and Nelder [29]. A latent multivariate logistic distribution obtains only for a suitable choice of L. The first element η_0 of the vector $\boldsymbol{\eta}$ is for ensuring the requirement $\sum_{\mathrm{j}} \pi_{\mathrm{j}} = 1$; η_0 also ensures the transformation $\boldsymbol{\pi}$ to $\boldsymbol{\gamma}$ is of full rank. For the bivariate case the elements of the vector $\boldsymbol{\eta}$ would be

$$\begin{aligned}
\eta_0 &= \log(\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22}), \\
\eta_1 &= \log \pi_{1.} - \log \pi_{2.} = \mathrm{logit}(\pi_{1.}), \\
\eta_2 &= \log \pi_{.1} - \log \pi_{.2} = \mathrm{logit}(\pi_{.1}), \\
\eta_{12} &= \log \pi_{11} + \log \pi_{22} - \log \pi_{12} - \log \pi_{21} = \log OR_{12},
\end{aligned}$$

where OR_{12} is the ratio of the odds of having the trait corresponding to the first and second member of the family, i.e. $OR_{12} = \{\Pr(Y_1 = 1)/\Pr(Y_1 = 2)\}/\{\Pr(Y_2 = 1)/\Pr(Y_2 = 2)\}$.

In GLM terminology these η's are known as linear predictors. The dependence of the π's on the covariates can be described through these η's. Let us assume that only the univariate margins depend on the covariates, so we can define the following regression models for the univariate margins

$$\eta_j = \boldsymbol{X}_j\boldsymbol{\beta}, \quad j = 1, 2, \ldots, d,$$

where $\boldsymbol{\beta}$ is the vector of regression parameters. To define the joint distribution of $\boldsymbol{y}_i$ fully, we need to define regression models corresponding to the two– and higher–order margins. Let us assume there are at most q different types of pairs (e.g. father-son, sib-sib, etc.) are in a family; the bivariate margins can be expressed as

$$\eta_{j_1j_2} = \boldsymbol{\alpha}, \quad 1 \leq j_1 < j_2 \leq d,$$

where the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T$ represents the dependence parameter. Let us assume third– and higher–order margins are constant at α_0, i.e.

$$\eta_{123} = \eta_{124} = \cdots = \eta_{1234} = \cdots = \eta_{12\cdots d} = \alpha_0.$$

The likelihood construction is similar to the model described in previous sections. The score function (3.15) can be written as

$$\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{i=1}^{K}\sum_{\mathrm{j}=1}^{2} \frac{m_{i\mathrm{j}}}{\pi_{i\mathrm{j}}} \frac{\partial \pi_{i\mathrm{j}}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}}. \tag{3.18}$$

From equation (3.17), we can write

$$\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}} = (CD^{-1}L)^{-1},$$

where $D = \text{diag}(L\boldsymbol{\pi})$. The matrix D can be written in terms of the elements of $\boldsymbol{\gamma}$, as an example which is shown in Appendix (A.2) for 3-variate model. McCullagh and Nelder [29] provide the detailed expressions for the elements of the score function and the Fisher information matrix.

3.4 Summary

In this chapter, we have described multivariate logistic models for analyzing multivariate binary responses. The multivariate Plackett distribution, which is described in Section 3.1.3, has been used to construct a multivariate logistic distribution. The bivariate Plackett distribution (see Section 3.1.1) is flexible to consider any continuous univariate margin to construct corresponding bivariate margin. Using two univariate logistic margins, we have constructed bivariate logistic margins from the Plackett distribution with a constant bivariate cross–product ratio. A comparison between the bivariate Plackett–normal distribution and the bivariate normal distribution with a set of comparable dependence parameters is shown in Section 3.1.1. This comparison reveals that the bivariate Plackett distribution with standard normal margin is similar to the bivariate normal distribution. In Section 3.3, McCullagh–Nelder approach of multivariate logistic model is briefly described.

Figure 3.1: Contour plots for the bivariate normal distribution.

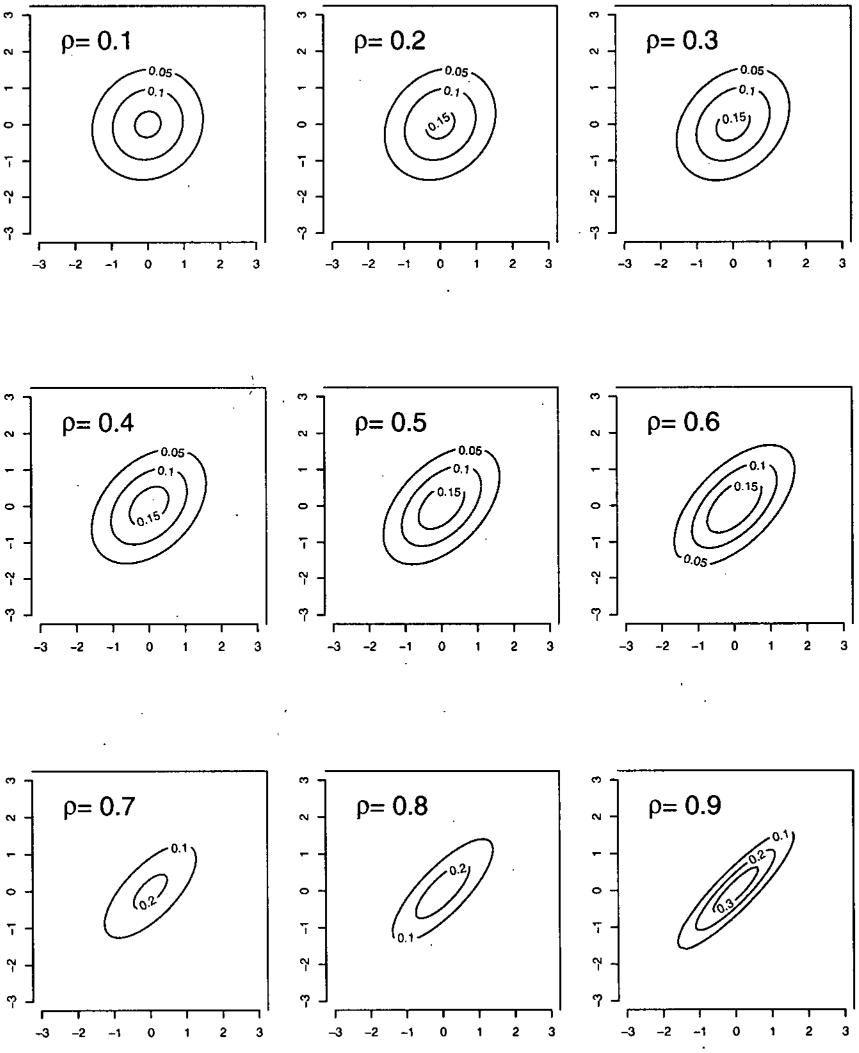
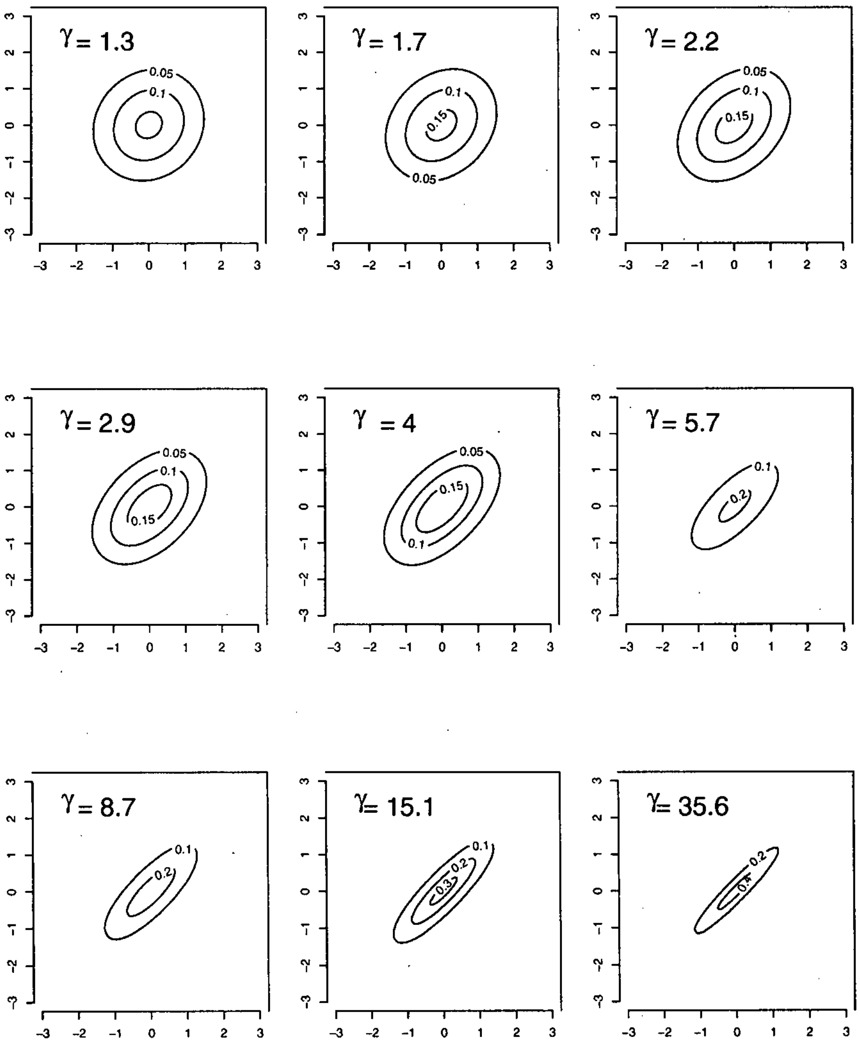


Figure 3.2: Contour plots for the bivariate Plackett–normal distribution.



Chapter 4

Generalized Estimating Equations

The objective of this thesis is to compare the available regression models and methods for multivariate binary responses. Generalized linear models (GLM) [29] are a general class of regression methods for univariate discrete and continuous responses, in which the density has a certain exponential form. Logistic regression models for binary responses, linear regression models for continuous responses, and log-linear models for count data are special cases of generalized linear models. In this chapter, we start our discussion from generalized linear models because the construction of a class of multivariate regression methods (estimating equation based) are closely related to that of GLM.

The main focus of this chapter is to describe generalized estimating equations (GEE) methods for analyzing multivariate binary responses. In 1986, Liang and Zeger [23] and Zeger and Liang [37] proposed the first version of the GEE method (GEE1) which can be used for analyzing both multivariate continuous and discrete responses in which the univariate margins are GLM. By considering the dependence parameters as a nuisance, the GEE1 method focus on estimation of the regression parameters defined for the mean function of the model. Later Liang, Zeger, and Qaqish [25] proposed a second version of the GEE method (GEE2) which can estimate both the regression and dependence parameters (in the form of the log odds–ratios) simultaneously. GEE methods are not likelihood based other than in the case of the normal distribution; estimating equations are defined for estimating the parameters. These estimating equations are similar to the score equations of a multivariate normal model. For analyzing multivariate binary data, models based on the quadratic exponential family, which can provide pseudo–likelihood estimators, are also available in the literature, e.g. Zhao and Prentice [38], Fitzmaurice and Laird [9]. A detailed discussion of these models can be found in Fitzmaurice *et al.* [10].

In this chapter, first we describe generalized linear models for analyzing univariate and multivariate binary responses. The GEE1 method and quadratic exponential family based methods are described in §4.2 and §4.3 respectively. Section 4.4 contains a description of the GEE2 method.

4.1 Generalized Linear Models

The GLM is the unified class of regression models for univariate continuous and discrete responses. Though our main focus is binary responses, in this section, we describe the GLM for general responses. The GLM has two important components: systematic and random component. The random component specifies the distribution of the response. The systematic component specifies the linear predictor which is a linear function of the known explanatory variables. The systematic component can be expressed as a known function of the mean parameter of the distribution of the response. This function is known as the link function which is a monotonic and differentiable function, with an appropriate domain.

4.1.1 The Model

Let $y_1, y_2, \ldots, y_K$ be a random sample from a distribution in the exponential family (see Lindsey [26], p. 11) having mean parameter μ_i $(i = 1, 2, \ldots, K)$ and constant dispersion parameter ϕ. The density function of y_i is of the form

$$f_y(y_i; \theta_i, \phi) \quad = \quad \exp\left\{(y_i \theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)\right\}, \tag{4.1}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are specified functions and θ_i is a function of μ_i, known as canonical parameter of the exponential family. It can be shown that

$$\mu_i = E(Y_i) = b'(\theta_i) \quad \text{and} \quad \text{var}(Y_i) = b''(\theta_i)\, a(\phi).$$

The variance of y_i is product of two terms: one, $V(\mu_i) \overset{def}{=} b''(\theta_i)$, a function of μ_i, is known as variance function and the other, $a(\phi)$, is a function of only the dispersion parameter ϕ. Thus, the second moment of y_i is a function of its first moment.

To define the systematic component of the model, let us consider the matrix $\boldsymbol{X}$ of order $K \times (p+1)$ as the design matrix and to accommodate an intercept term, the first column of $\boldsymbol{X}$ is $\mathbf{1}$. The linear predictor corresponding to the i^{th} $(i = 1, 2, \ldots, K)$ observation can be written as

$$\eta_i(\boldsymbol{\beta}) \quad = \quad \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where β_0 is the intercept term and β_j $(j = 1, 2, \ldots, p)$ is the regression coefficient corresponding to the j^{th} explanatory variable.

Let $h(\cdot)$ be any monotonic differentiable function such that

$$\eta_i(\boldsymbol{\beta}) = h(\mu_i), \quad i = 1, 2, \ldots, K;$$

$h(\cdot)$ is known as the link function in generalized linear model terminology. The link function relates the systematic component to the random component of the model.

4.1.2 Parameter Estimation

Given the sample $y_1, y_2, \ldots, y_K$, the log-likelihood function can be written as

$$l(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{K} \log f_y(y_i; \theta_i, \phi)$$

$$= \sum_{i=1}^{K} \{(y_i \theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)\} \overset{def}{=} \sum_{i=1}^{K} l_i(\boldsymbol{\theta}).$$

To estimate the parameters $\boldsymbol{\beta}$ of the model, we must solve the score equations

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \mathbf{0}. \tag{4.2}$$

By using chain rule, we can write

$$\begin{aligned} \boldsymbol{U}(\boldsymbol{\beta}) &= \sum_{i=1}^{K} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{K} \frac{(y_i - \mu_i)}{a(\phi)} \frac{1}{V(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{K} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} V_i^{-1} (y_i - \mu_i) \\ &= \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}\right)^T V^{-1} (\boldsymbol{y} - \boldsymbol{\mu}), \end{aligned} \tag{4.3}$$

where $V = \text{diag}\{V_1, V_2, \ldots, V_K\}$ and $V_i = V(\mu_i)a(\phi) = \text{var}(y_i)$. From the expression of the score equation (4.3), it is evident that the dispersion parameter ϕ can be ignored for estimating the regression parameters $\boldsymbol{\beta}$. But the estimate of the dispersion parameter is required to compute the standard error of the estimate of regression parameter.

The equation (4.2) can be solved by iteratively reweighted least squares (McCullagh and Nelder [29], p.41) to estimate the parameters of interest $\boldsymbol{\beta}$. Replacing $\boldsymbol{\beta}$ by the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$, equation (4.3) becomes a function of the dispersion parameter only which can be solved to obtain the maximum likelihood estimator $\hat{\phi}$. The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ has an asymptotic multivariate normal distribution with mean vector $\boldsymbol{\beta}$ and covariance matrix

$$V_U = \left[\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}\right)^T V^{-1} \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}\right)\right]^{-1}. \tag{4.4}$$

The variance can be estimated by $\hat{V}_U$ which is obtained by replacing $\boldsymbol{\beta}$ and ϕ by the corresponding maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ respectively in expression (4.4).

4.1.3 Multivariate Binary Response

In this section, the GLM procedure is described for multivariate binary responses. Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{id})^T$ be the binary response vector corresponding to the i^{th} $(i = 1, 2, \ldots, K)$ family. Let $\boldsymbol{X}_{ij}$ be the p-dimensional covariate vector corresponding to the j^{th} $(j = 1, 2, \ldots, d)$ member of the i^{th} family.

To apply the GLM procedure to multivariate binary responses, let us naively assume the responses are "independent" within each family. That means there are $d \times K$ independent observations in the sample. Assuming the marginal distribution of y_{ij} $(i = 1, 2, \ldots K;\ j = 1, 2, \ldots, d)$ is a member of the exponential family, the log–likelihood contribution of the j^{th} member of the i^{th} family is

$$l_{ij}(\theta_{ij}) \;=\; \exp\{(y_{ij}\theta_{ij} - b(\theta_{ij}))/a(\phi) + c(y_{ij}, \phi)\}, \tag{4.5}$$

where θ_{ij} is the canonical parameter which is a function of the corresponding mean function $\mu_{ij} = E(Y_{ij})$. Let $h(\cdot)$ be the link function which relates the mean parameter μ_{ij} to the linear predictor $\eta_{ij}(\boldsymbol{\beta}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}$ as $\mu_{ij} = h^{-1}(\boldsymbol{X}_{ij}\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$ is the vector of parameters of interest. The score function for the i^{th} family (similar to equation (4.3)) can be written as

$$\boldsymbol{U}_i(\boldsymbol{\beta}) \;=\; \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i), \tag{4.6}$$

where $V_i = \text{diag}\{\text{var}(y_{i1}), \text{var}(y_{i2}), \ldots, \text{var}(y_{id})\}$. The maximum likelihood estimator $\hat{\boldsymbol{\beta}}_I$ is the solution of the equation $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$, where

$$\begin{aligned}\boldsymbol{U}(\boldsymbol{\beta}) &= \sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\ &= \sum_{i=1}^{K} C_i V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i). \end{aligned} \tag{4.7}$$

If the binary responses within each family were independent, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_I$ is consistent and has an asymptotic multivariate normal distribution with mean vector $\boldsymbol{\beta}$ and covariance matrix

$$V_{I_0} = \left(\sum_{i=1}^{K} C_i V_i^{-1} C_i^T\right)^{-1}.$$

Replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_I$, equation (4.5) becomes a function of the dispersion parameter ϕ only. The corresponding likelihood function can be maximized to obtain the estimate ($\hat{\phi}$) of the dispersion parameter.

In spite of the fact that the responses within a family are correlated, the pseudo maximum likelihood estimator $\hat{\boldsymbol{\beta}}_I$ (obtained by assuming within family responses as independent) is consistent but $\hat{V}_{I_0}$ can be inconsistent. To obtain the consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}_I$, Liang and Zeger [23] used the inverse Godambe information matrix (see Godambe [12]) as

$$V_I = V_{I_0} H_2(\boldsymbol{\beta}_I) V_{I_0}, \tag{4.8}$$

where

$$H_2(\boldsymbol{\beta}) = \sum_{i=1}^{K} C_i V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T V_i^{-1} C_i^T.$$

The estimated covariance matrix $\hat{V}_I$ of $\hat{\boldsymbol{\beta}}_I$ can be obtained by using the estimator $\hat{\boldsymbol{\beta}}_I$ and $\hat{\phi}$ in the equation (4.8). The main disadvantage of this approach is that it sometimes provides less efficient estimates of the regression parameters in some cases when the intra–familial association is high.

4.2 Generalized Estimating Equations I

In the previous section, we have seen that the GLM procedure can be used for analyzing multivariate binary responses. Although this approach does not consider the within–family dependence in the analysis, with the correct specification of the univariate margins this approach can provide consistent estimators for both the regression parameters and the respective variances. However, ignoring the within–family dependence in the analysis costs in the efficiency of the estimators corresponding to the regression parameters.

To incorporate within–family dependence in the analysis, Liang and Zeger [23] proposed the generalized estimating equations procedure (GEE1). The main focus of the GEE1 method is to examine the dependence of the response on the covariate set which are measured for each observation. In the GEE1 method, regression models are defined for the mean function of each observation. Instead of specifying the joint distribution of the responses, Liang and Zeger [23] defined estimating equations for the regression parameters only, which they called generalized estimating equations. They introduce a "working" correlation or weight matrix to avoid the specification of a joint distribution of the responses. When there is stronger dependence in the data, then one should use larger correlations in the weight matrix. Crowder [4] showed some examples where the estimators corresponding to the parameters of the "working" correlation matrix do not converge in probability to a value in $[-1, 1]$ or $[0, 1]$.

4.2.1 The Method and Parameter Estimation

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{id})^T$ be the response vector corresponding to the i^{th} $(i = 1, 2, \ldots, K)$ family, where y_{ij}, the binary response corresponding to the j^{th} $(j = 1, 2, \ldots, d)$ member of the i^{th} family, representing the presence/absence of a specific trait of interest. Let $\boldsymbol{X}_{ij}$ be the p-dimensional vector of covariate values for the j^{th} member of the i^{th} family. As in Section (4.1.3), let us assume that y_{ij} follows a distribution from exponential family and the dependence of the mean function $\mu_{ij} = \Pr(Y_{ij} = 1)$ on the covariate set can be expressed by the link function $h(\cdot)$ as $\mu_{ij} = h^{-1}(\boldsymbol{X}_{ij}\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$ is the parameter of interest.

Liang and Zeger [23] used a "working" correlation matrix $R_i(\boldsymbol{\alpha})$ $(i = 1, 2, \ldots, K)$ of order $d \times d$ for specifying the within–family dependence. In the case of unequal family sizes, the dimension of the "working" correlation matrix is different for different families. The form of the "working" correlation matrix is assumed to be fully specified by the parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_q)^T$. Common correlation structures such as "independence" and "exchangeable" correlation structures can be obtained by considering $R_i(\boldsymbol{\alpha}) = I_d$ and $R_i(\boldsymbol{\alpha}) = (1-\rho)I_d + \rho\boldsymbol{J}_d$ respectively, where $\rho = \mathrm{corr}(Y_{ij}, Y_{ik})$, $j, k = 1, 2, \ldots, d\,(j \neq k)$, where I_d is an unit matrix of order $d \times d$ and $\boldsymbol{J}_d$ is a $d \times d$ matrix with all elements are one. The estimation equations corresponding to the "independence" correlation structure is known as independence estimating equations (IEE) which is similar to the procedure described in 4.1.3.

For estimating the regression parameters, Liang and Zeger [23] proposed the following set of estimating equations

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i), \tag{4.9}$$

where V_i is the "working" covariance matrix considered for i^{th} family, which can be expressed as a function of "working" correlation matrix as

$$V_i = A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2}, \tag{4.10}$$

where $A_i = \mathrm{diag}\{\mathrm{var}(y_{i1}), \mathrm{var}(y_{i2}), \ldots, \mathrm{var}(y_{id})\}$ and $\mathrm{var}(y_{ij}) = \mu_{ij}\, a(\phi)$, is a function of the known mean function and the dispersion parameter. So the "working" covariance matrix V_i is a function of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and ϕ.

Thus the estimating equations defined in equation (4.9) are a function of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ. Since we are only interested in estimating the regression parameters, the score equation (4.9) can be reduced as a function of $\boldsymbol{\beta}$ only by replacing $\boldsymbol{\alpha}$ and ϕ by $\hat{\boldsymbol{\alpha}}(\boldsymbol{Y}, \boldsymbol{\beta}, \phi)$ and $\hat{\phi}(\boldsymbol{Y}, \boldsymbol{\beta})$ respectively. So equation (4.9) can be written as

$$\begin{aligned} \boldsymbol{U}[\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}\{\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})\}] &= \sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \left[V_i\left(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \hat{\phi})\right)\right]^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\ &= \sum_{i=1}^{K} C_i B_i A_i. \end{aligned} \tag{4.11}$$

Let $\hat{\boldsymbol{\beta}}_{G_1}$ be the solution of the equation $\boldsymbol{U}[\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}\{\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})\}] = \boldsymbol{0}$. According to Liang and Zeger [23] given the estimators of $\boldsymbol{\alpha}$ and ϕ that converge in probability, the estimator of the regression parameter $\hat{\boldsymbol{\beta}}_{G_1}$ is consistent and asymptotically multivariate normal with mean vector $\boldsymbol{\beta}$ and covariance matrix

$$V_{G_1} = \left(\sum_{i=1}^{K} C_i B_i C_i^T\right)^{-1} \left\{\sum_{i=1}^{K} C_i B_i (\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T B_i C_i^T\right\} \left(\sum_{i=1}^{K} C_i B_i C_i^T\right)^{-1} . \quad (4.12)$$

The variance estimate $\hat{V}_{G_1}$ of $\hat{\boldsymbol{\beta}}_{G_1}$ can be obtained by replacing $\boldsymbol{\beta}$, ϕ, and $\boldsymbol{\alpha}$ in the expression for V_{G_1} by their estimates. Liang and Zeger [23] showed that the consistency of the regression parameters and their variances does not depend on the choice of the "working" correlation matrix as long as the estimates of the parameters of this matrix converge in probability. This is an error of Liang and Zeger pointed out by Crowder [4] and Sutradhar and Das [35]; one cannot talk about the consistency of the parameter that does not have real interpretation.

There are also some other drawbacks of the maximum quasi–likelihood estimator $\hat{\boldsymbol{\beta}}_{G_1}$. Crowder [4] also indicated that there may not exist any solution for $\hat{\boldsymbol{\alpha}}$; in that situation GEE1 cannot estimate the regression parameters. Even if $\hat{\boldsymbol{\alpha}}$ exist and it converges to a specific value, its limiting value depend on the form chosen for "working" correlation matrix. Sutradhar and Das [35] showed some results for multivariate binary data where the estimators corresponding to IEE are found as efficient as the estimators from GEE1.

4.3 Quadratic Exponential Family

As we have seen in the Section (4.2), GEE1 focus on estimation of the regression parameters corresponding to the marginal mean functions. The GEE1 procedure is not sufficient when the objective of the study is to estimate the parameters corresponding to both the mean function and within–family dependence structure. Regression methods based on a quadratic exponential family can estimate both kind of parameters simultaneously. In this section, we first describe the quadratic exponential family and then review a regression method based on this family of distribution.

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{id})^T$ be the binary response vector corresponding to the i^{th} $(i = 1, 2, \ldots, K)$ family. According to Cox [3], the joint distribution of $\boldsymbol{y}_i$ can be written in the following saturated log-linear form:

$$f(\boldsymbol{y}_i;\Theta_i,\Omega_i)=\exp\{\Theta_i^T\boldsymbol{y}_i+\Omega_i^T\boldsymbol{z}_i-A(\Psi_i,\Omega_i)\}, \tag{4.13}$$

where $\boldsymbol{z}_i=(y_{i1}y_{i2},\ldots,y_{id-1}y_{id},\ldots,y_{i1}y_{i2}\cdots y_{id})^T$ contains the second– and higher–order cross-products of $\boldsymbol{y}_i$, $\Theta_i=(\theta_{i1},\theta_{i2},\ldots,\theta_{id})^T$ and $\Omega_i=(\omega_{i12},\ldots,\omega_{id-1,d},\ldots,\omega_{i12\ldots d})^T$ are vectors of canonical parameters, and $A(\Theta_i,\Omega_i)$ is a normalizing constant such that $\sum\exp\{A(\Theta_i,\Omega_i)\}=\sum\exp\{\Theta_i^T\boldsymbol{y}_i+\Omega_i^T\boldsymbol{z}_i\}$ where sum is over all 2^d possible values of $\boldsymbol{y_i}$.

The parameters of Θ_i have interpretations in terms of conditional probabilities as

$$\theta_{ij} \quad = \quad \operatorname{logit}\{\Pr(Y_{ij}=1\mid Y_{ik}=0,\ j\neq k)\}$$

and the parameters of Ω_i can be interpreted in terms of log conditional odds–ratios and contrasts of log conditional odds–ratios as

$$\begin{aligned}
\omega_{ij_1j_2} &= \log OR(y_{ij_1},y_{ij_2}\mid y_{ij_3}=0,\ j_3\neq j_1,j_2),\\
\omega_{ij_1j_2j_3} &= \log\frac{OR(y_{ij_1},y_{ij_2}\mid y_{ij_3}=1,y_{ij_4}=0,\ j_4\neq j_1,j_2,j_3)}{OR(y_{ij_1},y_{ij_2}\mid y_{ij_3}=0,y_{ij_4}=0,\ j_4\neq j_1,j_2,j_3)},\\
&\cdots
\end{aligned}$$

where

$$OR(y_1,y_2)=\frac{\Pr(Y_1=1,Y_2=1)\ \Pr(Y_1=0,Y_2=0)}{\Pr(Y_1=1,Y_2=0)\ \Pr(Y_1=0,Y_2=1)},$$

is the odds–ratio between between two binary variables Y_1 and Y_2.

A special case of Cox's log–linear representation is the quadratic exponential family (Zhao and Prentice [38]) obtained by setting three– and higher–order dependence parameters in Ω_i at some fixed values in (4.13). This special case of the Cox's log–linear representation is equivalent to that obtained from conditionally logistic regressions (Joe and Liu [16]). Gourieroux *et al.* [13] consider an exponential quadratic model, parameterized by the mean vector and covariance matrix, for a general response vector.

However, one drawback of this representation is that it is not reproducible or closed under margins, i.e. if $\boldsymbol{y}_i^*$ is a subset of $\boldsymbol{y}_i$ of order $d^* \times 1$ $(d^* < d)$, then

$$\begin{aligned} f^*(\boldsymbol{y}_i^*) &= \sum_{\boldsymbol{y}_{i'}} f(\boldsymbol{y}_i; \Theta_i, \Omega_i), \quad \boldsymbol{y}_{i'} = \boldsymbol{y}_i \setminus \boldsymbol{y}_i^* \\ &\neq \exp\{\Theta_i^{*T} \boldsymbol{y}_i^* + \Omega_i^{*T} \boldsymbol{z}_i^* - A(\Theta_i^*, \Omega_i^*)\}, \end{aligned}$$

where Θ_i^* is the corresponding subset of Θ_i of order $d^* \times 1$, and Ω_i^* and $\boldsymbol{z}_i^*$ are the corresponding subsets of Ω_i and $\boldsymbol{z}_i$ respectively.

The canonical parameters in Ω_i are defined in terms of conditional odds–ratios which have limited use for the studies with unequal family sizes because conditional odds–ratios are specific to the number of subjects in a family. For example, since $OR(y_{i1}, y_{i2} \mid y_{i3} = 0) \neq OR(y_{i1}, y_{i2} \mid y_{i3} = y_{i4} = 0)$, the same parameters cannot be used to measure the association between y_{i1} and y_{i2} if some families have three subjects and others have four subjects.

4.3.1 Zhao-Prentice Method

Zhao and Prentice [38] used the quadratic exponential family for analyzing multivariate binary data by making a one–to–one transformation from the canonical parameters (Θ_i, Ω_i) to the marginal parameters $(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$, where $\boldsymbol{\mu}_i = \boldsymbol{\theta}_i$ and $\boldsymbol{\sigma}_i$ is the vector of the marginal covariances for the i^{th} family. Let us define the regression model for marginal parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ as

$$\boldsymbol{\mu}_i(\boldsymbol{\beta}) = h^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}), \ \boldsymbol{\sigma}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = g^{-1}(\boldsymbol{\alpha}, \boldsymbol{X}_i\boldsymbol{\beta}),$$

respectively, where $\boldsymbol{X}_i$ be the matrix of covariates for the i^{th} $(i = 1, 2, \ldots, K)$ family, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the vectors of parameters to be estimated, and $h(\cdot)$ and $g(\cdot)$ are the link functions. Let $\boldsymbol{\xi}_i = E(\boldsymbol{z}_i)$ and for (j, k) pair of the i^{th} family we can write $\xi_{ijk} = \sigma_{ijk} + \mu_{ij}\mu_{ik}$. That is, $\boldsymbol{\xi}_i$ is the function of both the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

The score function for the Zhao-Prentice model is

$$\begin{aligned}\boldsymbol{U}(\boldsymbol{\beta},\boldsymbol{\alpha}) &= \sum_{i=1}^{K}\begin{pmatrix}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}} & \mathbf{0}\\ \frac{\partial\boldsymbol{\xi}_i}{\partial\boldsymbol{\beta}} & \frac{\partial\boldsymbol{\xi}_i}{\partial\boldsymbol{\alpha}}\end{pmatrix}^T\begin{pmatrix}\operatorname{cov}(\boldsymbol{y}_i) & \operatorname{cov}(\boldsymbol{y}_i,\boldsymbol{z}_i)\\ \operatorname{cov}(\boldsymbol{z}_i,\boldsymbol{y}_i) & \operatorname{cov}(\boldsymbol{z}_i)\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{y}_i-\boldsymbol{\mu}_i\\ \boldsymbol{z}_i-\boldsymbol{\xi}_i\end{pmatrix}\\ &= \sum_{i=1}^{K} C_iB_iA_i. \end{aligned}\tag{4.14}$$

The solution of $\boldsymbol{U}(\boldsymbol{\beta},\boldsymbol{\alpha})=0$ provides pseudo-maximum likelihood estimators $(\hat{\boldsymbol{\beta}}^T,\hat{\boldsymbol{\alpha}}^T)^T$.

Pseudo-maximum likelihood estimation of $(\boldsymbol{\beta},\boldsymbol{\alpha})$ requires direct calculation of third and fourth order moments of $\boldsymbol{y}_i$ which involves summation over 2^d possible values of $\boldsymbol{y}_i$. This computation is tedious when the family size is large. In this context, Zhao and Prentice [38] suggested the use of any convenient "working" covariance matrices in equation (4.14). In this case, the estimator $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}})$ is no longer a pseudo-maximum likelihood estimator but the estimator is consistent and is asymptotically normally distribution provided the model specification of $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ is correct.

Both the GEE1 and Zhao–Prentice approaches considered the correlation coefficient for specifying the dependence among the observations within a family. But as a dependence parameter, the correlation coefficient is not the best choice because its range depends on the univariate margins. Fitzmaurice and Laird [9] used quadratic exponential family with conditional log odds–ratio as the dependence parameter. The main drawback of this approach is that the joint distribution of the observations are not reproducible. Thus this approach has limited application for studies with unequal family sizes. The interpretation of the conditional log odds–ratio is not as attractive as the unconditional association parameters. In the following section we will describe a GEE2 method which overcomes some of the drawbacks of the GEE1, the Zhao–Prentice, and the Fitzmaurice–Laird methods.

4.4 Generalized Estimating Equations II

The focus of the GEE1 method is to describe the dependence of the mean function of the response on the explanatory variables. By considering the dependence as a nuisance, the GEE1 procedure provides consistent estimators of the regression parameters given the correct specification of the mean function. If the objective of the study is to describe both the dependence of the mean response on the explanatory variables and the dependence structures among the responses, the GEE1 method is not sufficient. The quadratic exponential model is one way to deal with such problems but has several drawbacks. Liang, Zeger, and Qaqish [25] extended the GEE1 procedure for estimating the parameters defined in the mean function and the dependence structure simultaneously. They call this procedure GEE2 and considered the bivariate log–odds ratio as the dependence parameter to illustrate this procedure.

4.4.1 The Method

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \dots, y_{id})^T$ be the binary response vector corresponding to the i^{th} $(i = 1, 2, \dots, K)$ family. Let us also define $\boldsymbol{z}_i = (y_{i1}y_{i2}, y_{i1}y_{i3}, \dots, y_{id-1}y_{id})^T$, a vector of order m, where $m = \binom{d}{2}$. Let $\boldsymbol{X}_{ij}$ be the covariate vector of order $1 \times (p+1)$ corresponding to the j^{th} $(j = 1, 2, \dots, d)$ member of the i^{th} family. Let $h(\cdot)$ be the link function such that

$$\mu_{ij}(\boldsymbol{\beta}) = E(Y_{ij}) = h^{-1}(\boldsymbol{X}_{ij}\boldsymbol{\beta}).$$

The pairwise dependence is expressed in terms of the odds–ratios, for the (j, k) pair of the i^{th} family the odds–ratio is defined as

$$\gamma_{ijk} \quad = \quad \frac{\Pr(y_{ij} = 1, y_{ik} = 1)\ \Pr(y_{ij} = 0,\ y_{ik} = 0)}{\Pr(y_{ij} = 1,\ y_{ik} = 0)\ \Pr(y_{ij} = 0,\ y_{ik} = 1)}.$$

Let us assume the vector $\boldsymbol{\gamma}$ of order $m \times 1$, where m is the number of the pairs in the i^{th} family, can be expressed as a function of the $q \times 1$ $(q \leq m)$ vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)^T$, i.e. $\boldsymbol{\gamma} = g^{-1}(\boldsymbol{\alpha})$, where

$g(\cdot)$ is any suitable link function. From the bivariate Plackett distribution (see §3.1), for any pair of responses we can write

$$\xi_{ijk}(\boldsymbol{\beta},\boldsymbol{\alpha}) = E(Y_{ij}\,Y_{ik}) = \begin{cases} (1/2)(\gamma_{ijk}-1)^{-1}\left\{S(\mu_{ij},\mu_{ik},\gamma_{ijk}) - [\{S(\mu_{ij},\mu_{ik},\gamma_{ijk})\}^2 \right. & \\ \quad \left. -\ 4(\gamma_{ijk}-1)\,\gamma_{ijk}\,\mu_{ij}\,\mu_{ik}]^{1/2}\right\}, & \text{if } \gamma_{ijk} \neq 1, \\ \mu_{ij}\,\mu_{ik}, & \text{if } \gamma_{ijk} = 1, \end{cases} \tag{4.15}$$

where $S(\mu_1,\mu_2,\gamma) \overset{def}{=} 1-(\mu_1+\mu_2)(1-\gamma)$.

4.4.2 Estimating Equations

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T,\boldsymbol{\alpha}^T)^T$ be the parameters of the model which we wish to estimate. Since the joint distribution of the response vector $\boldsymbol{y}_i$ is not fully specified, Liang *et al.* [25] consider the following estimating equations for $\boldsymbol{\theta}$:

$$\begin{aligned} U(\boldsymbol{\theta}) &= \sum_{i=1}^{K} \frac{\partial(\boldsymbol{\mu}_i,\boldsymbol{\xi}_i)}{\partial\boldsymbol{\theta}} \operatorname{cov}^{-1}\begin{pmatrix}\boldsymbol{y}_i\\ \boldsymbol{z}_i\end{pmatrix}\begin{pmatrix}\boldsymbol{y}_i-\boldsymbol{\mu}_i\\ \boldsymbol{z}_i-\boldsymbol{\xi}_i\end{pmatrix} \\ &= \sum_{i=1}^{K}\begin{pmatrix}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}} & \mathbf{0}\\ \frac{\partial\boldsymbol{\xi}_i}{\partial\boldsymbol{\beta}} & \frac{\partial\boldsymbol{\xi}_i}{\partial\boldsymbol{\alpha}}\end{pmatrix}^T \operatorname{cov}^{-1}\begin{pmatrix}\boldsymbol{y}_i\\ \boldsymbol{z}_i\end{pmatrix}\begin{pmatrix}\boldsymbol{y}_i-\boldsymbol{\mu}_i\\ \boldsymbol{z}_i-\boldsymbol{\xi}_i\end{pmatrix} \\ &= \sum_{i=1}^{K} C_iB_iA_i = \mathbf{0}. \end{aligned} \tag{4.16}$$

For the i^{th} family the inverse of the covariance matrix is

$$B_i = \begin{pmatrix}\operatorname{cov}(\boldsymbol{y}_i) & \operatorname{cov}(\boldsymbol{y}_i,\boldsymbol{z}_i)\\ \operatorname{cov}(\boldsymbol{z}_i,\boldsymbol{y}_i) & \operatorname{cov}(\boldsymbol{z}_i)\end{pmatrix}^{-1}.$$

The components of this matrix can be expressed in terms of the first four moments of $\boldsymbol{y}_i$, which are shown in Appendix A.1. The estimating equations (4.16) are similar to the pseudo-score functions of the Zhao–Prentice method with the odds–ratio as the marginal dependence parameter and a specific choice of "working" covariance matrix.

The solution $\hat{\boldsymbol{\theta}}^T = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\alpha}}^T)$ of $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$ follows an asymptotic multivariate normal distribution with mean vector $\boldsymbol{\theta}$ and covariance matrix

$$V_{G_2} \;=\; H_1^{-1}(\boldsymbol{\theta})\, H_2(\boldsymbol{\theta})\, H_1^{-1}(\boldsymbol{\theta}), \tag{4.17}$$

where

$$\begin{aligned} H_1(\boldsymbol{\theta}) &= \sum_{i=1}^{K} C_i B_i C_i^T, \\ H_2(\boldsymbol{\theta}) &= \sum_{i=1}^{K} C_i B_i \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{\mu}_i \\ \boldsymbol{z}_i - \boldsymbol{\xi}_i \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{\mu}_i \\ \boldsymbol{z}_i - \boldsymbol{\xi}_i \end{pmatrix}^T B_i\, C_i^T. \end{aligned}$$

The estimator $\hat{\boldsymbol{\theta}}$ is consistent if the model for both the mean parameters $h(\boldsymbol{\mu})$ and the association parameters $g(\boldsymbol{\gamma})$ are correctly specified.

The main advantage of the GEE2 method over the regression methods based on the quadratic exponential family lies in the interpretation of the dependence parameters. The GEE2 models the bivariate odds–ratios for the within–family dependence which has a straightforward interpretation regarding the magnitude of the association between any pair of members within a family. The quadratic exponential family methods consider conditional models for association parameters which are of limited use in the case of unequal family sizes.

4.5 Summary

In this chapter chronological developments of some estimating equation approaches which are applicable to multivariate binary responses have been shown. Starting from the GLM, which is used for analyzing univariate binary responses, we have covered regression methods based on generalized estimating equations as well as quadratic exponential family based methods.

By naively assuming the correlated binary responses as "independent", one can apply GLM to estimate the regression parameters defined for the mean function. These estimators are consistent but less efficient if the within–family association is high. The GEE1 approach considers within family dependence in the analysis as a nuisance and can estimate regression parameters corresponding to the marginal mean function. Provided the model specification of the marginal means is correct, the GEE1 approach provides consistent estimators of the regression parameters.

The GEE1 approach does not estimate the parameters corresponding to the within family correlation structure which might be of interest of some studies. Some models based on quadratic exponential family (e.g Zhao–Prentice, Fitzmaurice–Laird) and the GEE2 can estimate regression parameters and parameters corresponding to dependence structure simultaneously. These three models considered three different types of association parameters in the respective models. Pairwise correlations, conditional odds–ratios and bivariate odds–ratios are considered as dependence parameters in the Zhao-Prentice, the Fitzmaurice-Laird and the GEE2 methods respectively. Liang *et al.* [25] proposed the GEE2 method without mentioning that the probabilistic assumptions are consistent with the multivariate logistic model of Molenberghs and Lesaffre [31] (which came later).

Chapter 5

Simulation Study

The main objective of this thesis is to compare maximum likelihood and estimation equation based estimators of the multivariate logistic model for analyzing multivariate binary responses. For this comparison, a simulation study is considered with different family structures. The multivariate probit and logistic models are also compared for estimating the conditional probabilities. Numerical methods have been used for estimating the parameters of the methods that are considered in this thesis.

This chapter contains results of the simulation study and the description of the numerical methods that are used in this thesis for estimating the parameters of the models. In the following section, a brief description of these numerical methods is given.

5.1 Numerical Optimization Methods

Generally numerical optimization methods are used to optimize a function of the independent variables which can also consider restrictions on the independent variables; the function to be optimized is known as the objective function. In statistical applications, the negative of the log–likelihood function $l(\boldsymbol{\theta})$ is the objective function to be minimized to estimate the unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. A common method for estimating the unknown parameters is the Newton–Raphson method which requires analytic evaluation of the gradient and the Hessian matrix. If the objective function has a complex form, often the Hessian matrix cannot be evaluated analytically; in such situations quasi–Newton methods which only require analytic evaluation of the objective function, can be used for minimizing the negative of the log likelihood function. Numerical differentiation techniques are also useful for estimating the Hessian and possibly the gradient at fixed values of the parameters.

For fitting the multivariate logistic and probit models we have used the C routines of Joe ([18], [19]) respectively. The quasi–Newton method with numerically approximated gradient has been used in these routines. For solving the estimating equations, we have used the C++ code of Joe [20]. This routine is based on quasi–Newton method with the gradient and the Hessian matrix computed by a differentiation package FADBAD (`http://www.imm.dtu.dk/~os/fadbad.html`) which uses automatic differentiation techniques to compute the derivatives of a function written in a high level language such as FORTRAN, C, or C++.

The public domain version of the GEE2 code, originally written in Pascal[1] and converted to C with *p2c*, cannot handle general familial data. It can handle familial data with a simple interclass/intraclass structure. It is also restrictive in computer memory requirements. It would have taken more time to modify the original GEE2 code to suit our purposes than to start from scratch. The advantage of the automatic differentiation approach is that we need to code only the likelihood function. For simulated familial data with a structure such that the original GEE2 program can be used, we checked that our program and the original program gave the same (or nearly the same) estimates. The main difference between the public domain version of GEE2 and our implementation is the computational procedure for the second- and higher–order moments. We used the multivariate Plackett construction for computing these moments, whereas the pub-

[1] `http://statlab.uni-heidelberg.de/statlib/GEE/GEE2/`

lic domain version uses the McCullagh–Nelder–Glonek approach. For estimating the parameters, the public domain version uses the classical Newton–Raphson method and we have used a FADBAD based Newton–Raphson method. Because of the operator overloading, the FADBAD based Newton–Raphson method is bit slower than the classical Newton–Raphson method.

In the subsequent two sections, we first describe the classical Newton–Raphson method and then the quasi–Newton methods are also briefly described.

5.1.1 The Newton-Raphson Method

The Newton-Raphson method is based on approximating the objective function $l(\boldsymbol{\theta})$ locally by a quadratic model and then minimizing that function. For the value of the parameters $\boldsymbol{\theta_k}$ at the kth iteration, the objective function can be approximated as

$$l(\boldsymbol{\theta_k} + \boldsymbol{d}) \quad \approx \quad l(\boldsymbol{\theta_k}) + U^T(\boldsymbol{\theta_k})\boldsymbol{d} + \frac{1}{2}\boldsymbol{d^T} H(\boldsymbol{\theta_k})\boldsymbol{d}, \tag{5.1}$$

where

$$U(\boldsymbol{\theta_k}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta^T}}\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta_k}} \quad \text{and} \quad H(\boldsymbol{\theta_k}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta^T}}\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta_k}}$$

are the gradient and the Hessian matrix of the objective function $l(\boldsymbol{\theta})$ respectively. The Hessian matrix is the Jacobian of the gradient which is symmetric and assumed to be positive definite. The minimum of the right–hand side of (5.1) is achieved when $\boldsymbol{d_k}$ is the minimum of the quadratic function

$$Q(\boldsymbol{d}) \quad = \quad U^T(\boldsymbol{\theta_k})\boldsymbol{d} + \frac{1}{2}\boldsymbol{d^T} H(\boldsymbol{\theta_k})\boldsymbol{d}.$$

The expression for $\boldsymbol{d_k}$ can be obtained by solving the equation $\{\partial\, Q(\boldsymbol{d})/\partial \boldsymbol{d}\} = 0$, which leads to

$$\boldsymbol{d_k} = -H^{-1}(\boldsymbol{\theta_k})U(\boldsymbol{\theta_k}).$$

The quantity $\boldsymbol{d_k}$ is known as the Newton direction and is used to update the current parameter value $\boldsymbol{\theta_k}$ by

$$\begin{aligned}\boldsymbol{\theta_{k+1}} &= \boldsymbol{\theta_k} + \boldsymbol{d_k}, \quad k = 0, 1, 2\ldots \\ &= \boldsymbol{\theta_k} - H^{-1}(\boldsymbol{\theta_k})U(\boldsymbol{\theta_k}) \\ &= \boldsymbol{\theta_k} - H_k^{-1}\boldsymbol{g_k}, \qquad (5.2)\end{aligned}$$

where $H_k = H(\boldsymbol{\theta_k})$ and $\boldsymbol{g_k} = U(\boldsymbol{\theta_k})$ are introduced to simplify the notation. Starting with a suitable initial value $\boldsymbol{\theta}_0$, the equation (5.2) is successively evaluated until the parameter vector $\boldsymbol{\theta}$ has converged.

This method requires the computation of the gradient and inverse of the Hessian matrix at each iteration. To evaluate the Hessian matrix, $p(p+1)/2$ partial derivatives must be calculated analytically and for inverting the Hessian matrix a system of linear equations must be solved. But in many cases the second derivative of the objective function is not available analytically, so the Hessian matrix cannot be computed. The Newton method also breaks down if the Hessian matrix is singular at some iteration. Otherwise Newton's method works very well if the initial value is sufficiently close to the true value. These difficulties of the Newton's method lead researchers to search for other optimization methods, such as quasi–Newton methods which do not require computing the Hessian matrix or the gradient analytically.

5.1.2 The Quasi-Newton Method

The quasi–Newton method, also known as a variable metric method (see Nash [32]), approximates the inverse of the Hessian matrix by some modification of the previously constructed matrix. This method does not require computation or inversion of the Hessian matrix for minimizing the objective function $l(\boldsymbol{\theta})$. For implementing the quasi–Newton method, assume that either the gradient is available analytically or a computer routine is available for numerically computing the gradient.

For example, at the k^{th} iteration, from the currently available quantities such as $\boldsymbol{\theta_k}$, $\boldsymbol{g_k}$, $\boldsymbol{\theta_{k-1}}$, and $\boldsymbol{g_{k-1}}$ the quasi–Newton method provide an approximation B_k to H_k^{-1}. This approximation can be used to update the current value of the parameter by

$$\boldsymbol{\theta_{k+1}} = \boldsymbol{\theta_k} - B_k\,\boldsymbol{g_k}. \tag{5.3}$$

By using Taylor's theorem, the gradient at k^{th} iteration can approximately be written as

$$\begin{aligned} \boldsymbol{g_{k-1}} &\simeq \boldsymbol{g_k} + H_k(\boldsymbol{\theta_{k-1}} - \boldsymbol{\theta_k}) \\ \Rightarrow \boldsymbol{y_k} &\simeq H_k\boldsymbol{s_k} \end{aligned} \tag{5.4}$$

where $\boldsymbol{s_k} = \boldsymbol{\theta_k} - \boldsymbol{\theta_{k-1}}$ and $\boldsymbol{y_k} = \boldsymbol{g_k} - \boldsymbol{g_{k-1}}$. If the objective function is quadratic, the Hessian matrix is constant, i.e. $H_k = H$, $\forall\, k \geq 0$ and the relationship of (5.4) is exact. If the objective function is not quadratic but is strictly convex and has continuous second partial derivatives in a neighborhood of $\boldsymbol{\theta}^*$ then the objective function is well approximated by a quadratic function with Hessian matrix $H(\boldsymbol{\theta}^*)$ in a sufficiently small neighborhood of $\boldsymbol{\theta}^*$ (Wolfe [36]). For such objective functions the relation (5.4) becomes exact as $\boldsymbol{\theta}_k$ approaches $\boldsymbol{\theta}^*$. So, it seems desirable that the approximation B_k to H_k^{-1} should satisfy the relation

$$B_k\boldsymbol{y_k} = \boldsymbol{s_k}. \tag{5.5}$$

This equation is known as the quasi-Newton equation. It is desirable that B_k can be computed from B_{k-1}, $\boldsymbol{s_k}$, and $\boldsymbol{y_k}$. Consider

$$B_k = B_{k-1} + C_k(\boldsymbol{s_k}, \boldsymbol{y_k}, B_{k-1}), \tag{5.6}$$

where the correction matrix C_k has the same properties as the matrix B_k. This construction of the sequence $\{B_k\}$ is not unique; a number of methods are available for computing B_k from the current values of the parameters. One such method is described in the following subsection.

Davidson-Fletcher-Powell Method

The Davidson-Fletcher-Powell method (see Wolfe [36]) is one of the popular methods for constructing the sequence $\{B_k\}$. For this method the correction matrix is written as

$$C_k(\boldsymbol{s_k}, \boldsymbol{y_k}, B_{k-1}) \quad = \quad \boldsymbol{s_k}\boldsymbol{u_k^T} - B_{k-1}\boldsymbol{y_k}\,\boldsymbol{v_k^T},$$

where $\boldsymbol{u_k}$ and $\boldsymbol{v_k}$ are nonzero vectors such that

$$\boldsymbol{u_k^T}\boldsymbol{y_k} = 1, \quad \boldsymbol{v_k^T}\boldsymbol{y_k} = 1. \tag{5.7}$$

By using the this form of correction matrix C_k in equation (5.6) we get

$$\begin{aligned} B_k\boldsymbol{y_k} &= B_{k-1}\boldsymbol{y_k} + \boldsymbol{s_k}\boldsymbol{u_k^T}\boldsymbol{y_k} - B_{k-1}\boldsymbol{y_k}\boldsymbol{v_k^T}\boldsymbol{y_k} \\ &= B_{k-1}\boldsymbol{y_k} + \boldsymbol{s_k} - B_{k-1}\boldsymbol{y_k} \\ &= \boldsymbol{s_k}. \end{aligned}$$

This shows that the Davidson-Fletcher-Powell method satisfies the quasi–Newton equation (5.5). The choice

$$\boldsymbol{u_k} = \frac{\boldsymbol{s_k}}{\boldsymbol{s_k^T}\boldsymbol{y_k}}, \quad \boldsymbol{v_k} = \frac{B_{k-1}\boldsymbol{y_k}}{\boldsymbol{y_k^T}B_{k-1}\boldsymbol{y_k}}.$$

satisfies the equation (5.7). Therefore, we can write the expression of the matrix B_{k+1} as

$$B_k \quad = \quad B_{k-1} + \frac{\boldsymbol{s_k}\boldsymbol{s_k^T}}{\boldsymbol{s_k^T}\boldsymbol{y_k}} - \frac{(B_{k-1}\boldsymbol{y_k})(B_{k-1}\boldsymbol{y_k})^T}{\boldsymbol{y_k^T}B_{k-1}\boldsymbol{y_k}}.$$

After getting the value of B_k we can update the parameter by using the equation (5.2).

5.1.3 Newton method with Automatic Differentiation

As we already mentioned that the Newton–Raphson/quasi–Newton method can be used with the gradient and the Hessian matrix obtained from differentiation packages. A differentiation package FADBAD, which is based on automatic differentiation, is used for coding the routine of Joe [20].

Automatic differentiation has become a very popular tool for numerical differentiation in the last twenty years. The advantages of this method over other existing procedures (e.g. divided difference, symbolic differentiation) have been discussed by Griewank [14]. There are two versions of automatic differentiation, known as forward and backward modes. A number of software packages for automatic differentiation written in high–level languages such as FORTRAN or C++ are available. The user only needs to provide the subroutine to evaluate the underlying objective function provided an interface routine has been written for Newton–Raphson. We have used the forward mode of the automatic differentiation method to evaluate the Hessian matrix. This Hessian matrix used in Newton–Raphson routine for estimating the parameters of the GEE2 method.

Automatic differentiation introduced a data type `doublet` (see Dixon [6]). A `doublet` variable U, consists of $p+1$ values $(u, \bigtriangledown u_i)$, where $\bigtriangledown u_i = \partial u/\partial x_i$, $i = 1, 2, \ldots, p$. For `doublet` variable the usual operators (e.g. $+$, $-$, $*, \ldots$) are overloaded; operator overloading is a nice feature of the computer languages such as C++/Fortran which helps to redefine the meaning of the elementary operators. For an assignment $W = U * V$ where U and V are `doublets`, the multiplication operator $*$ is defined in such a way that it not only computes the product of U and V but also update the associated gradient object by $\bigtriangledown w_i = u \bigtriangledown v_i + v \bigtriangledown u_i$ $(i = 1, 2, \ldots, p)$. Similarly all other elementary operations can also be defined; for example

$$\begin{aligned} W = U + V &\quad\to\quad (u + v, \bigtriangledown u_i + \bigtriangledown v_i,\;\; i = 1, 2, \ldots, p), \\ W = 1/U &\quad\to\quad \left(\frac{1}{u}, \frac{-1}{u^2} \bigtriangledown u_i,\; i = 1, 2, \ldots, p\right), \\ W = \log u &\quad\to\quad \left(\log u, \frac{1}{u} \bigtriangledown u_i\; i = 1, 2, \ldots, p\right). \end{aligned}$$

Automatic differentiation packages convert any given functions to these elementary functions and then use the chain rule to compute the derivatives of the given function.

5.2 Comparison of the Methods

In this section, a comparison of maximum likelihood (ML) and GEE2 approaches for estimating the parameters of the multivariate logistic model is shown by considering a simulation study. Four different types of the family structures namely Pedigree A, B, C, and D are considered, the description and the graphical representation of these families are given in the following sections. Because the multivariate logistic distribution is defined based on implicit equations for dimensions $d \geq 3$, simulation from it would be very difficult. To simulate from the multivariate logit binary model requires the computation of 2^d (d is the family size) multivariate probabilities and then the 2^d possible binary d–dimensional vectors can be considered as outcomes of a multivariate discrete random variable with these probabilities. This is much more difficult than simulation from the multivariate probit model. Therefore, the comparison is based on multivariate binary data simulated from the multivariate probit model.

A sample of 200 families (family size depends on the pedigree types) is generated for each pedigree. For simplicity of the comparison, all 200 families have the same family structure. A mixture of pedigrees A, B, and C is also considered for the simulation study because for real data pedigrees for different families will typically be different. To examine the effect of the sample size for the comparison of ML and GEE2 approaches, we also analyzed a data set from Pedigree A with 600 families. The only covariate Age is assumed uniform on (l, u), where l and u depend on the member of the family. The response vector is generated by using the multivariate probit model; i.e., given the specified correlation structure and the values of the intercept (β_0) and regression coefficient (β_1) for the covariate Age the response vector is obtained by using the equation (2.1). To simulate binary response vectors by the multivariate probit models, two types of correlation structures (exchangeable and familial) are considered for each pedigree. For an exchangeable correlation structure, the correlation is the same for all pairs of members of a family; on the other hand, for a familial correlation structure, the correlations differ for different types of pairs. Three values (0.9, 0.5, 0.1) of the correlation coefficients and two sets of the correlation coefficients are considered for the exchangeable and familial correlation structures respectively. We also used different values of the regression constant and coefficient in different analyses.

For this analysis, we assume that only the univariate margins depend on the covariate Age. The multivariate logistic model has equations

$$\begin{aligned} \text{logit}\, \boldsymbol{\pi} &= \beta_0 + \beta_1 * \boldsymbol{Age}, \\ \log \boldsymbol{\gamma} &= \boldsymbol{\alpha}, \end{aligned}$$

for the univariate margins and the dependence structure respectively. Here, $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ are the vector of univariate margins and cross–product ratios respectively. For the multivariate logistic model the third– and higher–order dependence parameters are assumed fixed at one.

For each pedigree, the results of the simulation analysis for the exchangeable and familial correlation structures are shown in two separate tables. The maximum likelihood and GEE2 methods are compared with respect to the mean estimates of the regression parameters and log OR for dependence parameters, and their corresponding standard errors. The average of the absolute difference of the parameter estimates and their standard errors corresponding to these two methods are also shown in these tables. Empirical standard deviations of the parameter estimates corresponding to these two approaches are also shown in these tables. For each of the analysis 500 repetitions were considered to obtain the results of these tables. The true parameter values which are used to generate the data for different pedigree are also listed in these tables.

5.2.1 Pedigree A

Pedigree A is the simplest family structure considered for the simulation study. This pedigree consists of four members who are from two different relative classes and the graphical diagram of this pedigree is shown in Figure 5.1. The first member is one of the parents and the others are offspring, so only two types of dependence namely, sib–sib (SS) (2:3, 2:4, 3:4) and parent–offspring (PO) (1:2, 1:3, 1:4) are considered for this pedigree.

Figure 5.1: Graphical diagram of Pedigree A.

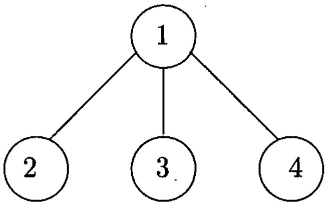


Table 5.2 shows the simulation results for the Pedigree A, where the responses are generated by a multivariate probit model with an exchangeable correlation structure, i.e. $\rho_{PO} = \rho_{SS}$. The average absolute difference of the estimates show that the maximum likelihood and GEE2 parameter estimates are equal up to two decimal places for the correlation coefficient $\rho = 0.1$ and $\rho = 0.5$. For $\rho = 0.9$, the parameter estimates corresponding to the covariate Age ($\hat{\beta}_1$) and corresponding to parent–offspring pair ($\log \widehat{OR}_{PO}$) are equal up to one decimal place and others are equal up to two decimal places. The differences among the parameter estimates tend to increase as the within–family dependence increases. The averages of the maximum likelihood and the GEE2 estimates of the standard errors are equal up to one decimal place for all the values of ρ. For the extreme values of correlation coefficients (i.e. $\rho = 0.1, 0.9$) the absolute differences between the corresponding standard errors are larger than at $\rho = 0.5$. The average standard errors of the estimates of the dependence parameters ($\log \widehat{OR}_{PO}$ and $\log \widehat{OR}_{SS}$) increase as ρ increases. The average of the intercept ($\hat{\beta}_0$) and the estimates corresponding to the covariate Age tend to decrease as ρ increases. For Pedigree A, on average the GEE2 estimates are marginally more efficient than maximum likelihood estimates with exchangeable correlation structure.

Table 5.3 shows the simulation results for Pedigree A, where the responses are generated by a multivariate probit model with a familial correlation structure with sib–sib and parent–offspring correlations. Two sets of correlation coefficients $\{\rho_{SS} = 0.8, \rho_{PO} = 0.6\}$ and $\{\rho_{SS} = 0.9, \rho_{PO} = 0.4\}$ are considered for this case. The ML estimates are found marginally less efficient than the corresponding estimates from the GEE2.

Table 5.4 shows the simulation results for Pedigree A with a sample of size 600. This results show that the average of the difference of the ML and GEE2 estimates tend to decrease as sample size increases and standard errors behave as expected.

For the Pedigree A, the empirical standard deviations (SDs) of the parameter estimates are also shown in the Tables 5.2–5.4. The SDs are quite similar for the estimates corresponding to the ML and GEE2 approaches. As expected these empirical standard deviations are approximately equal to the average of the corresponding standard errors.

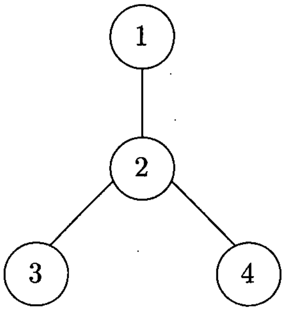
Note that as expected, most of the regression estimates are approximately 1.6–1.8 times the probit regression coefficients. Also the odds–ratio parameter estimates are roughly satisfied the equations (3.6) and (3.7). For example, for Pedigree A with exchangeable correlation structure ($\rho = 0.5$), the regression parameter estimates from the logistic model of $\hat{\beta}_0 = 1.318$ and $\hat{\beta}_1 = 0.362$ correspond to the regression parameters of the probit model of $\beta_0 = 0.8$ and $\beta_1 = 0.2$. That is, the logistic estimates are 1.65 (=1.318/0.8) and 1.81 (=0.362/0.2) times the respective probit parameters.

5.2.2 Pedigree B

Figure 5.2 shows the graphical diagram of Pedigree B, which has four members from three different relative classes. The members are one grandparent, one parent and two offspring; three types of dependence namely sib–sib (3:4), parent–offspring (1:2, 2:3, 2:4), and second–degree relationship (D2) (1:3, 1:4) can be considered for this family structure.

Table 5.5 shows the simulation results of Pedigree B, where the responses are generated by a multivariate probit model with an exchangeable correlation structure. The averages of the

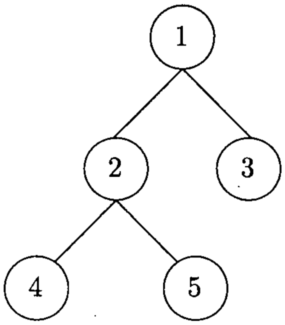
Figure 5.2: Graphical diagram of Pedigree B.



parameter estimates and the corresponding standard errors for the maximum likelihood and the GEE2 methods are very close. For $\rho = 0.1$, the average absolute differences show that the maximum likelihood and GEE2 estimates of the regression parameters are similar up to two decimal places. For $\rho = 0.5$, $\hat{\beta}_1$ and $\log \widehat{OR}_{PO}$ are similar up to two decimal places and other estimators are similar up to one decimal place. For $\rho = 0.9$, all the estimators are similar up to one decimal place. The average of the difference between the parameter estimates increases as the within–family dependence increases. The difference between the standard errors of the parameter estimates corresponding to the maximum likelihood and the GEE2 are equal up to one decimal points for all the values of the correlation coefficients. The average standard errors agree more closely for $\rho = 0.5$ than for the other values of ρ. For Pedigree B with exchangeable correlation structure, the estimates of β_1 and $\log OR$ are found to be marginally less efficient for the maximum likelihood in most cases.

Table 5.6 shows the simulation result for Pedigree B, where the responses are generated by a multivariate probit model with a familial correlation structure. Two sets of correlation coefficients $\{\rho_{SS} = 0.8,\ \rho_{PO} = 0.6, \text{ and } \rho_{D2} = 0.5\}$ and $\{\rho_{SS} = 0.9,\ \rho_{PO} = 0.5, \text{ and } \rho_{D2} = 0.1\}$ are considered for this case. The average of the absolute differences show that except for $\log \widehat{OR}_{D2}$, all other estimators from the maximum likelihood and the GEE2 are similar up to two decimal places. For this case the parameter estimates corresponding to the GEE2 are found to be marginally more

Figure 5.3: Graphical diagram of Pedigree C.



efficient than maximum likelihood estimator.

The empirical SDs of the corresponding parameter estimates of the ML and GEE2 approaches are very close and are consistent with the average of the corresponding standard errors.

5.2.3 Pedigree C

Pedigree C has five members: one grandparent, one parent, one uncle, and two offspring; the graphical diagram of this pedigree is in Figure 5.3. Similar to Pedigree B, for this pedigree three types of dependence can be considered: sib–sib (2:3, 4:5), parent–offspring (1:2, 1:3, 2:4, 2:5), and second–degree relationship (1:4, 1:5, 3:4, 3:5).

Table 5.7 shows the results for Pedigree C, where the responses are generated by a multivariate probit model with an exchangeable correlation structure. For correlation coefficient $\rho = 0.1$, the average of the absolute differences show that the parameter estimates corresponding to the maximum likelihood and GEE2 are similar up to two decimal places. For $\rho = 0.5$ and $\rho = 0.9$ all the estimates corresponding to the dependence parameters are similar up to one decimal place. The absolute difference between the respective parameter estimates increases as the within–family dependence increases. The average of the absolute differences also show that the standard errors are similar up to one decimal place for all values of the correlation coefficients. For $\rho = 0.5$ these standard errors are more similar than for the other values of the correlation coefficient. For Pedigree C, the maximum likelihood method provides marginally more efficient estimators than the GEE2 only for the case in which within–family dependence is high.

Table 5.8 shows the simulation results for Pedigree C, where the responses are generated by a multivariate probit model with a familial correlation structure. As before two sets of correlation coefficients are considered for this case. The averages of the absolute differences show that both the parameter estimates and the corresponding standard errors for the maximum likelihood and the GEE2 are similar up to one decimal place. The estimates corresponding to the GEE2 are found to be marginally efficient than the estimates corresponding to maximum likelihood.

The empirical SDs of the corresponding parameter estimates of the ML and GEE2 approaches are very close and are consistent with the average of the corresponding standard errors.

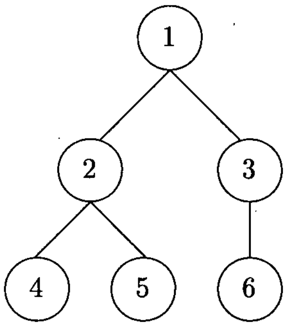
Though the family sizes are different, the same dependence structures were used for Pedigree B and C in the simulation study. The comparison of the results of these two pedigrees would provide the effect of family size for comparing the parameter estimates of the maximum likelihood and the GEE2. This comparison reveals that the standard errors of all the parameter estimates decreases as the pedigree size increases and the absolute differences between the parameters and the standard errors also decreases as pedigree size increases.

5.2.4 Pedigree D

Figure 5.4 shows the graphical diagram of the Pedigree D which has six members: one grandparent, one parent, one uncle, two offsprings, and a cousin. This pedigree can consider four types of dependences: sib–sib (2:3, 4:5), parent–offsprings (1:2, 1:3, 2:4, 2:5, 3:6), second–degree relationship (1:4, 1:5, 1:6, 2:6, 3:4, 3:5) and third–degree relationship (D3) (4:6, 5:6).

Table 5.9 shows the simulation results for Pedigree D, where the responses are generated by a multivariate probit model with exchangeable correlation structure. The average absolute difference

Figure 5.4: Graphical diagram of Pedigree D.



show that the parameter estimates corresponding to the dependence structure (i.e. log OR) are similar up to one decimal place for all correlation coefficients we have considered. The parameter estimate corresponding to the covariate Age is similar up to two decimal place except for $\rho = 0.9$. The average of the absolute difference also show that the standard errors corresponding to log ORs are similar up to one decimal place for $\rho = 0.1$ and 0.5. For correlation coefficient $\rho = 0.9$ the average differences between these standard errors are greater than 0.1. The maximum likelihood estimates corresponding to the log ORs are found marginally efficient than the corresponding GEE2 estimates for $\rho = 0.9$. For $\rho = 0.5$ both the parameter estimates and the standard errors are found to be closer than the extreme cases.

Table 5.10 shows the simulation results for Pedigree D, where the responses are generated by a multivariate probit model with a familial correlation structure. The averages of the absolute differences show that the maximum likelihood and GEE2 parameter estimates and standard errors are similar up to one decimal place. The maximum likelihood estimates are found marginally more efficient than the GEE2 only for log OR_{PO} and log OR_{D2}.

The empirical SDs of the corresponding parameter estimates of the ML and GEE2 approaches are very close and are consistent with the average of the corresponding standard errors.

5.2.5 Mixture of pedigrees A, B, and C

For the mixture of the pedigrees A, B, and C, a sample of 100 families from each of these pedigrees are considered. The comparison of ML and GEE2 for this mixture pedigree is very important because in practice, the families are of different sizes. For this case, we can consider three types of dependences: sib–sib, parent–offspring, second–degree relationship.

Table 5.11 shows the simulation results for the mixture of pedigrees with exchangeable correlation structure. The averages of the absolute differences show that the maximum likelihood and GEE2 parameter estimates are similar up to two decimal places for $\rho = 0.1, 0.5$; for $\rho = 0.9$ these estimates are similar up to one decimal place. The absolute differences also show that the standard errors corresponding to maximum likelihood and GEE2 are similar up to one decimal place for all values of the correlation coefficients. The estimators corresponding to maximum likelihood are found to be marginally more efficient than the corresponding estimators of the GEE2 method only when the within–family dependence is high.

Table 5.12 shows the results for mixture pedigrees with a familial correlation structure. The absolute differences show that the maximum likelihood and GEE2 estimators are similar up to one decimal place. The estimators corresponding to the GEE2 are found to be marginally more efficient than the corresponding parameters of the maximum likelihood.

5.2.6 Estimating Conditional Probabilities

In Chapter 2, we have discussed the procedure of estimating conditional probabilities and their standard errors for the multivariate probit model (see §2.4). A similar procedure can also be derived for the multivariate logistic model, but not for the GEE2 method. The GEE2 method does not specify the joint distribution, so expressions for the orthant probabilities which enter into the conditional probabilities are not available. Given estimates of the parameters of the multivariate probit and logistic models, these orthant probabilities can be computed from equations (2.14) and

(3.10) respectively. We have used C routines of Joe ([18] and [19]) for computing these orthant probabilities; the conditional probabilities and corresponding standard errors are obtained by using numerical differentiation methods.

To compare the multivariate probit and logistic models for the estimation of these conditional probabilities, we have considered samples of size 200 from the pedigrees A, B, and C. As before these samples are also generated from the multivariate probit models. Tables 5.13–5.15 show the estimates of the conditional probabilities and their standard errors for these pedigrees. As an example, five different conditional probabilities are arbitrarily chosen for each of these cases. Given the parameter estimates and the covariate values similar estimates of other probabilities can also be obtained. For example, given a diseased individual of age 67 from Pedigree B the estimate of the probability that his younger offspring of age 14 has the disease is about 0.85, with a standard error 0.02. These tables show that most of these estimates are similar for the multivariate probit and logistic models.

The empirical SDs of the corresponding parameter estimates of the ML and GEE2 approaches are very close and are consistent with the average of the corresponding standard errors.

5.2.7 Comments on Computing Time

This research was motivated by the observation that maximum likelihood estimation for the multivariate logistic model became very time consuming when there were pedigrees in the data set of sizes 7 or more. The computing time for the multivariate Plackett distribution is exponentially increasing in the dimension, because of the equations that must be solved for dimension 3 and higher. Table 5.1 shows the average computing time needed for the maximum likelihood and the GEE2 methods. The GEE2 equations are computationally more efficient in dimensions ≥ 6 because they require only computation of the multivariate Plackett distribution up to dimension 4. In practice, if the familial data has many pedigrees of dimension 6 or more, we recommend the GEE2 approach over maximum likelihood.

5.3 Summary

In this chapter, a simulation study for comparing the maximum likelihood and GEE2 method for the multivariate logistic model is discussed. In Section 5.1, a brief description of the numerical optimization methods which are used to fit these models is given. Because of the complex form of the gradient and Hessian matrix, the classical Newton–Raphson method is not be used for these models. The quasi–Newton method is used to estimate the parameters of the models considered in this chapter. The results of the simulation study indicate that the maximum likelihood and GEE2 estimates of the regression parameters and their respective standard errors are usually equal at least up to one decimal place. This result is consistent for all the pedigrees we have considered for this study. The empirical SDs of the parameter estimates corresponding to the ML and GEE2 approaches are found quite close and are consistent with the corresponding average of the standard errors. The estimates of the conditional probabilities corresponding to the multivariate probit and logistic models are also found to be similar.

Table 5.1: Average computing time (in mins.) by pedigrees and methods.

	Correlation coefficient					
	0.1		0.5		0.9	
	ML	GEE2	ML	GEE2	ML	GEE2
A	0.216	0.465	0.215	0.460	0.217	0.507
B	0.294	0.460	0.283	0.466	0.277	0.542
C	3.755	2.037	3.628	1.911	3.653	2.190
D	36.852	4.204	31.120	3.681	40.951	4.914
Mixed	2.288	1.479	2.324	1.494	2.266	1.704

Table 5.2: Simulation results for Pedigree A with exchangeable correlation structure.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.8	1.307 (0.155)	1.308 (0.155)	0.001	0.155	0.152	0.004
Age	0.2	0.383 (0.532)	0.381 (0.532)	0.002	0.517	0.504	0.017
SS	0.1	0.282 (0.249)	0.283 (0.249)	0.001	0.253	0.250	0.019
PO	0.1	0.291 (0.273)	0.292 (0.272)	0.002	0.265	0.263	0.015
Const.	0.8	1.318 (0.149)	1.318 (0.149)	0.002	0.158	0.159	0.004
Age	0.2	0.362 (0.442)	0.360 (0.442)	0.008	0.444	0.442	0.014
SS	0.5	1.547 (0.266)	1.545 (0.266)	0.005	0.274	0.272	0.010
PO	0.5	1.578 (0.287)	1.575 (0.287)	0.008	0.284	0.281	0.009
Const.	0.8	1.316 (0.175)	1.312 (0.175)	0.009	0.170	0.169	0.003
Age	0.2	0.376 (0.311)	0.378 (0.305)	0.032	0.305	0.307	0.014
SS	0.9	3.827 (0.379)	3.824 (0.374)	0.008	0.390	0.376	0.016
PO	0.9	3.849 (0.393)	3.839 (0.378)	0.032	0.409	0.388	0.023

†empirical SDs are in the parenthesis

Table 5.3: Simulation results for Pedigree A with familial correlation structure.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.8	1.314 (0.181)	1.312 (0.181)	0.005	0.183	0.180	0.004
Age	0.2	0.369 (0.442)	0.371 (0.443)	0.012	0.450	0.440	0.015
SS	0.8	2.960 (0.346)	2.959 (0.347)	0.003	0.339	0.327	0.012
PO	0.6	1.927 (0.325)	1.921 (0.325)	0.009	0.341	0.330	0.012
Const.	0.5	0.810 (0.185)	0.810 (0.185)	0.003	0.184	0.179	0.006
Age	1.0	1.731 (0.532)	1.731 (0.533)	0.010	0.505	0.483	0.025
SS	0.9	3.717 (0.364)	3.718 (0.364)	0.006	0.368	0.357	0.017
PO	0.4	1.221 (0.359)	1.219 (0.359)	0.010	0.376	0.364	0.015

†empirical SDs are in the parenthesis

Table 5.4: Simulation results for Pedigree A with exchangeable correlation structure and sample of size 600.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.8	1.311 (0.084)	1.312 (0.084)	0.001	0.090	0.090	0.001
Age	0.2	0.361 (0.292)	0.359 (0.291)	0.002	0.300	0.299	0.006
SS	0.1	0.298 (0.146)	0.298 (0.146)	0.001	0.146	0.146	0.007
PO	0.1	0.306 (0.154)	0.306 (0.154)	0.001	0.152	0.151	0.005
Const.	0.8	1.316 (0.090)	1.316 (0.089)	0.001	0.093	0.093	0.001
Age	0.2	0.352 (0.252)	0.352 (0.252)	0.004	0.260	0.260	0.005
SS	0.5	1.548 (0.164)	1.546 (0.164)	0.003	0.160	0.158	0.004
PO	0.5	1.558 (0.165)	1.556 (0.165)	0.005	0.165	0.162	0.004
Const.	0.8	1.324 (0.099)	1.320 (0.098)	0.006	0.098	0.097	0.002
Age	0.2	0.344 (0.168)	0.346 (0.167)	0.018	0.173	0.176	0.006
SS	0.9	3.812 (0.221)	3.809 (0.222)	0.004	0.223	0.220	0.005
PO	0.9	3.840 (0.229)	3.834 (0.227)	0.016	0.231	0.226	0.007

†empirical SDs are in the parenthesis

Table 5.5: Simulation results for Pedigree B with exchangeable correlation structure.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.8	1.309 (0.154)	1.310 (0.154)	0.001	0.158	0.156	0.003
Age	0.2	0.372 (0.395)	0.370 (0.395)	0.001	0.400	0.393	0.011
SS	0.1	0.261 (0.432)	0.261 (0.432)	0.002	0.422	0.417	0.015
PO	0.1	0.287 (0.265)	0.287 (0.265)	0.002	0.258	0.262	0.019
D2	0.1	0.288 (0.340)	0.288 (0.340)	0.002	0.321	0.317	0.024
Const.	0.8	1.316 (0.148)	1.317 (0.149)	0.003	0.159	0.162	0.004
Age	0.2	0.357 (0.321)	0.354 (0.326)	0.008	0.346	0.343	0.010
SS	0.5	1.527 (0.397)	1.563 (0.920)	0.047	0.385	0.388	0.015
PO	0.5	1.574 (0.273)	1.572 (0.274)	0.008	0.282	0.281	0.012
D2	0.5	1.580 (0.324)	1.581 (0.328)	0.011	0.317	0.316	0.018
Const.	0.8	1.323 (0.177)	1.314 (0.172)	0.020	0.171	0.171	0.006
Age	0.2	0.356 (0.246)	0.367 (0.236)	0.067	0.229	0.243	0.019
SS	0.9	3.822 (0.493)	3.857 (0.478)	0.035	0.482	0.475	0.030
PO	0.9	3.862 (0.386)	3.838 (0.373)	0.052	0.397	0.392	0.035
D2	0.9	3.897 (0.414)	3.878 (0.409)	0.086	0.432	0.432	0.041

†empirical SDs are in the parenthesis

Table 5.6: Simulation results for Pedigree B with familial correlation structure.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.5	0.797 (0.163)	0.797 (0.163)	0.002	0.171	0.167	0.005
Age	1.0	1.768 (0.391)	1.767 (0.392)	0.005	0.412	0.397	0.018
SS	0.8	2.861 (0.411)	2.861 (0.412)	0.003	0.399	0.393	0.017
PO	0.6	1.971 (0.280)	1.972 (0.278)	0.009	0.294	0.288	0.013
D2	0.4	1.249 (0.409)	1.245 (0.406)	0.013	0.393	0.381	0.027
Const.	0.5	0.796 (0.183)	0.797 (0.183)	0.001	0.180	0.175	0.005
Age	1.0	1.765 (0.445)	1.764 (0.446)	0.003	0.430	0.415	0.016
SS	0.9	3.764 (0.478)	3.765 (0.477)	0.003	0.469	0.452	0.023
PO	0.5	1.581 (0.272)	1.581 (0.270)	0.006	0.295	0.289	0.013
D2	0.3	0.905 (0.443)	0.904 (0.442)	0.010	0.419	0.400	0.031

†empirical SDs are in the parenthesis

Table 5.7: Simulation results for Pedigree C with exchangeable correlation structure.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.8	1.313 (0.157)	1.314 (0.158)	0.001	0.156	0.153	0.006
Age	0.2	0.361 (0.405)	0.359 (0.405)	0.002	0.400	0.391	0.015
SS	0.1	0.281 (0.324)	0.281 (0.325)	0.003	0.306	0.303	0.013
PO	0.1	0.264 (0.231)	0.264 (0.232)	0.003	0.231	0.229	0.016
D2	0.1	0.283 (0.222)	0.284 (0.222)	0.002	0.227	0.224	0.021
Const.	0.8	1.313 (0.159)	1.311 (0.160)	0.004	0.161	0.159	0.006
Age	0.2	0.370 (0.339)	0.370 (0.339)	0.008	0.343	0.341	0.014
SS	0.5	1.563 (0.289)	1.557 (0.292)	0.012	0.290	0.287	0.014
PO	0.5	1.553 (0.246)	1.546 (0.248)	0.012	0.253	0.247	0.014
D2	0.5	1.579 (0.240)	1.573 (0.239)	0.010	0.249	0.244	0.017
Const.	0.8	1.317 (0.174)	1.309 (0.175)	0.015	0.171	0.168	0.007
Age	0.2	0.364 (0.235)	0.369 (0.230)	0.032	0.229	0.234	0.020
SS	0.9	3.835 (0.393)	3.817 (0.407)	0.024	0.364	0.370	0.034
PO	0.9	3.811 (0.375)	3.793 (0.374)	0.027	0.342	0.342	0.030
D2	0.9	3.835 (0.364)	3.818 (0.348)	0.035	0.341	0.342	0.032

†empirical SDs are in the parenthesis

Table 5.8: Simulation results for Pedigree C with familial correlation structure.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.5	0.787 (0.157)	0.787 (0.157)	0.003	0.165	0.160	0.007
Age	1.0	1.796 (0.395)	1.798 (0.394)	0.011	0.406	0.394	0.021
SS	0.8	2.885 (0.295)	2.888 (0.293)	0.011	0.294	0.294	0.014
PO	0.6	1.944 (0.272)	1.947 (0.271)	0.021	0.273	0.271	0.018
D2	0.4	1.223 (0.272)	1.212 (0.273)	0.019	0.281	0.279	0.022
Const.	0.8	1.310 (0.178)	1.310 (0.178)	0.004	0.181	0.178	0.008
Age	0.2	0.374 (0.394)	0.375 (0.393)	0.011	0.399	0.393	0.021
SS	0.9	3.815 (0.353)	3.821 (0.351)	0.015	0.343	0.342	0.020
PO	0.5	1.527 (0.272)	1.534 (0.275)	0.018	0.267	0.263	0.020
D2	0.3	0.914 (0.293)	0.909 (0.293)	0.020	0.287	0.279	0.025

†empirical SDs are in the parenthesis

Table 5.9: Simulation results for Pedigree D with exchangeable correlation structure.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.8	1.303 (0.131)	1.306 (0.139)	0.003	0.137	0.131	0.009
Age	0.2	0.385 (0.317)	0.388 (0.325)	0.004	0.321	0.315	0.020
SS	0.1	0.317 (0.301)	0.294 (0.590)	0.026	0.309	0.307	0.030
PO	0.1	0.320 (0.197)	0.306 (0.363)	0.017	0.216	0.212	0.032
D2	0.1	0.309 (0.190)	0.286 (0.542)	0.026	0.200	0.196	0.036
D3	0.1	0.272 (0.315)	0.249 (0.584)	0.026	0.311	0.306	0.035
Const.	0.8	1.326 (0.147)	1.323 (0.146)	0.005	0.146	0.144	0.007
Age	0.2	0.354 (0.264)	0.355 (0.264)	0.008	0.275	0.271	0.017
SS	0.5	1.558 (0.285)	1.544 (0.289)	0.019	0.287	0.286	0.023
PO	0.5	1.577 (0.233)	1.564 (0.234)	0.017	0.231	0.226	0.021
D2	0.5	1.572 (0.221)	1.559 (0.223)	0.017	0.220	0.215	0.020
D3	0.5	1.548 (0.290)	1.538 (0.294)	0.016	0.298	0.294	0.025
Const.	0.8	1.342 (0.181)	1.325 (0.174)	0.019	0.190	0.162	0.040
Age	0.2	0.327 (0.181)	0.342 (0.175)	0.030	0.193	0.187	0.048
SS	0.9	3.855 (0.364)	3.817 (0.379)	0.046	0.326	0.366	0.115
PO	0.9	3.825 (0.317)	3.787 (0.322)	0.046	0.285	0.319	0.106
D2	0.9	3.840 (0.314)	3.802 (0.316)	0.052	0.274	0.311	0.107
D3	0.9	3.813 (0.389)	3.779 (0.407)	0.046	0.346	0.378	0.108

†empirical SDs are in the parenthesis

Table 5.10: Simulation results for Pedigree D with familial correlation structure.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.8	1.321 (0.145)	1.318 (0.140)	0.008	0.156	0.142	0.022
Age	0.2	0.354 (0.303)	0.353 (0.310)	0.017	0.318	0.303	0.041
SS	0.9	3.802 (0.347)	3.812 (0.336)	0.025	0.333	0.333	0.059
PO	0.5	1.529 (0.228)	1.546 (0.232)	0.024	0.212	0.222	0.053
D2	0.3	0.911 (0.232)	0.907 (0.232)	0.022	0.212	0.221	0.048
D3	0.1	0.264 (0.409)	0.258 (0.406)	0.027	0.388	0.378	0.063

†empirical SDs are in the parenthesis

Table 5.11: Simulation results for mixed Pedigree with exchangeable correlation structures.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.8	1.318 (0.125)	1.319 (0.125)	0.001	0.127	0.125	0.003
Age	0.2	0.352 (0.338)	0.351 (0.338)	0.002	0.346	0.341	0.010
SS	0.1	0.286 (0.263)	0.287 (0.263)	0.002	0.253	0.247	0.018
PO	0.1	0.304 (0.201)	0.304 (0.201)	0.002	0.205	0.203	0.012
D2	0.1	0.321 (0.263)	0.321 (0.263)	0.002	0.261	0.257	0.031
Const.	0.8	1.316 (0.129)	1.315 (0.128)	0.002	0.131	0.130	0.003
Age	0.2	0.358 (0.302)	0.359 (0.301)	0.006	0.300	0.300	0.009
SS	0.5	1.547 (0.256)	1.544 (0.258)	0.007	0.250	0.247	0.011
PO	0.5	1.561 (0.206)	1.557 (0.206)	0.008	0.221	0.218	0.010
D2	0.5	1.580 (0.267)	1.576 (0.268)	0.009	0.259	0.258	0.024
Const.	0.8	1.318 (0.136)	1.313 (0.136)	0.010	0.138	0.137	0.004
Age	0.2	0.360 (0.206)	0.363 (0.204)	0.024	0.199	0.206	0.013
SS	0.9	3.822 (0.330)	3.813 (0.335)	0.016	0.324	0.324	0.024
PO	0.9	3.830 (0.295)	3.821 (0.308)	0.022	0.303	0.302	0.021
D2	0.9	3.850 (0.323)	3.843 (0.331)	0.035	0.331	0.340	0.038

†empirical SDs are in the parenthesis

Table 5.12: Simulation results for mixed Pedigree with familial correlation structures.

	True	Parameter Estimates			Standard Errors		
	values	ML	GEE2	Diff.	ML	GEE2	Diff.
Const.	0.8	1.322 (0.136)	1.323 (0.137)	0.004	0.156	0.153	0.007
Age	0.2	0.342 (0.358)	0.343 (0.355)	0.009	0.350	0.344	0.016
SS	0.9	3.781 (0.251)	3.783 (0.245)	0.012	0.322	0.316	0.017
PO	0.5	1.539 (0.212)	1.541 (0.215)	0.012	0.243	0.238	0.015
D2	0.3	0.917 (0.279)	0.908 (0.282)	0.025	0.306	0.294	0.033
Const.	0.5	0.809 (0.149)	0.807 (0.150)	0.006	0.138	0.135	0.005
Age	1.0	1.734 (0.347)	1.737 (0.348)	0.017	0.359	0.351	0.016
SS	0.8	2.809 (0.341)	2.816 (0.337)	0.018	0.254	0.253	0.017
PO	0.6	1.931 (0.221)	1.942 (0.223)	0.023	0.230	0.226	0.016
D2	0.3	0.611 (0.291)	0.602 (0.294)	0.035	0.298	0.290	0.033

†empirical SDs are in the parenthesis

Table 5.13: Conditional probabilities for a family of Pedigree A with $\rho_{SS} = 0.9$, $\rho_{PO} = 0.4$, $X_1 = 53$, $X_2 = 20$, $X_3 = 16$, and $X_4 = 14$.

Conditional	Mlogit		Mprobit	
Probabilities	Est.	SE	Est.	SE
$\Pr(Y_4 = 1 \mid Y_1 = 1, Y_2 = 1, Y_3 = 1, \boldsymbol{X})$	0.952	0.009	0.959	0.008
$\Pr(Y_4 = 1 \mid Y_1 = 1, Y_2 = 1, Y_3 = 0, \boldsymbol{X})$	0.638	0.028	0.551	0.022
$\Pr(Y_4 = 1 \mid Y_1 = 1, Y_2 = 0, Y_3 = 0, \boldsymbol{X})$	0.140	0.033	0.160	0.033
$\Pr(Y_4 = 1 \mid Y_1 = 0, Y_2 = 1, Y_3 = 1, \boldsymbol{X})$	0.932	0.016	0.931	0.018
$\Pr(Y_4 = 1 \mid Y_1 = 1, Y_2 = 1, \boldsymbol{X})$	0.930	0.013	0.932	0.012

Table 5.14: Conditional probabilities for a family of Pedigree B with $\rho_{SS} = 0.9$, $\rho_{PO} = 0.5$, $\rho_{D2} = 0.3$, $X_1 = 68$, $X_2 = 46$, $X_3 = 14$ and $X_4 = 10$.

Conditional	Mlogit		Mprobit	
Probabilities	Est.	SE	Est.	SE
$\Pr(Y_4 = 1 \mid Y_1 = 1, Y_2 = 1, Y_3 = 1, \boldsymbol{X})$	0.950	0.012	0.953	0.011
$\Pr(Y_4 = 1 \mid Y_1 = 1, Y_2 = 0, Y_3 = 0, \boldsymbol{X})$	0.156	0.046	0.178	0.041
$\Pr(Y_4 = 1 \mid Y_1 = 0, Y_2 = 1, Y_3 = 0, \boldsymbol{X})$	0.321	0.093	0.259	0.060
$\Pr(Y_4 = 1 \mid Y_1 = 0, Y_2 = 0, Y_3 = 1, \boldsymbol{X})$	0.912	0.029	0.895	0.032
$\Pr(Y_4 = 1 \mid Y_2 = 1, \boldsymbol{X})$	0.825	0.021	0.826	0.021

Table 5.15: Conditional probabilities for a family of Pedigree C with $\rho_{SS} = 0.9$, $\rho_{PO} = 0.5$, $\rho_{D2} = 0.3$, $X_1 = 68$, $X_2 = 46$, $X_3 = 14$, and $X_4 = 10$.

Conditional	Mlogit		Mprobit	
Probabilities	Est.	SE	Est.	SE
$\Pr(Y_5 = 1 \mid Y_1 = 1, Y_2 = 1, Y_3 = 1, Y_4 = 1, \boldsymbol{X})$	0.957	0.009	0.962	0.008
$\Pr(Y_5 = 1 \mid Y_1 = 0, Y_2 = 1, Y_3 = 0, Y_4 = 0, \boldsymbol{X})$	0.277	0.094	0.227	0.053
$\Pr(Y_5 = 1 \mid Y_1 = 1, Y_2 = 0, Y_3 = 0, Y_4 = 0, \boldsymbol{X})$	0.171	0.035	0.186	0.034
$\Pr(Y_5 = 1 \mid Y_1 = 0, Y_2 = 0, Y_3 = 1, Y_4 = 0, \boldsymbol{X})$	0.094	0.035	0.139	0.033
$\Pr(Y_5 = 1 \mid Y_1 = 1, Y_2 = 1, Y_3 = 1, \boldsymbol{X})$	0.870	0.022	0.875	0.023

Chapter 6

Discussion

The objective of this thesis was to compare existing methods for analyzing multivariate binary responses. We are mainly interested to study the performance of these methods for analyzing the binary response (e.g. presence/absence of a genetic disease) in the context of data structures typically arising in genetics. In genetics, it is assumed that these binary responses are the quantal values of an underlying continuous distribution. This assumption leads us to consider latent variable models for analyzing binary data in genetics. These latent variable models are likelihood based and have been widely used since the early 1930's. Recently, estimating equation based methods were introduced which can also be used to analyze multivariate binary responses. Among the existing methods, we have compared latent variable and estimating equation based methods for multivariate binary responses.

We have reviewed the latent variable models (multivariate probit and logistic) in Chapters 2 and 3 respectively. The multivariate probit model is based on the assumption that the underlying latent variable follows multivariate normal distribution. In genetics, this assumption has physical meaning because the observed phenotypic value is assumed to be the sum of many small genetic effects and the environmental deviations. By the central limit theorem, the normality assumption for this phenotypic value is reasonable. The multivariate probit model is useful to obtain maximum likelihood estimator of the parameters corresponding to the regression model defined for the univariate margins. The maximum likelihood estimator of the association between the members of the family can also be obtained in terms of the latent correlation coefficient, which is known as tetrachloric correlation. In biostatistics and epidemiology, the odds–ratio has a more attractive interpretation than the correlation coefficient as a dependence parameter. Because of that linear logistic model is widely used though there is no physical reasoning of considering this model.

In Chapter 3, we discussed the multivariate logistic model which is based on the assumption that the underlying univariate margins are logistic. Given the univariate logistic margins and the two– and higher–order cross–product ratios as the dependence parameters, we have used the multivariate Plackett construction to obtain the higher–order margins. Besides estimating the regression parameters corresponding to the univariate margins, the multivariate logistic model parameterizes the dependence among the members of the family in terms of the log odds–ratio. We also described the McCullagh–Nelder–Glonek approach to the multivariate logistic model. Instead of using the multivariate Plackett construction, to obtain the orthant probabilities, they defined univariate margins and the two– and higher–order dependence parameters in terms of the linear combinations of the joint probabilities. Using the logit and logarithmic links corresponding to the univariate margins and the dependence parameters respectively, the McCullagh–Nelder–Glonek approach provide similar results as the multivariate Plackett construction based model.

In Chapter 4, estimating equation based methods are reviewed. These methods do not require full specification of the joint distribution of the responses. Among the estimating equation based methods, the GEE2 method can estimate the regression parameters and the dependence parameters (in terms of logORs) simultaneously. Instead of using likelihood based score functions, for the GEE2 method, a set of estimating equations are considered to estimate the parameters corresponding to the univariate margins and to the dependence structure. These estimating equations are similar to the score function of a multivariate normal model. In GEE2, the bivariate dependence parameters are defined according to the Plackett distribution. These estimating equations require the computation of third– and fourth–order moments, for which McCullagh–Nelder–Glonek's approach of multivariate logistic model has been used. Besides using estimating equations in place of likelihood score functions for estimating the parameters of the model, the GEE2 method is similar to the multivariate logistic model. That means, the GEE2 method is an estimating equation approach for estimating the parameters of the multivariate logistic model.

For analyzing binary responses, besides identifying important covariates and estimating the association between the family members for the occurrence of the disease, the estimate of the probability that an individual has the disease given the disease status of the other family members is also of interest in genetics. This conditional probability is the ratio of the two orthant probabilities. The GEE2 method cannot estimate the orthant probabilities. In this thesis we have compared the multivariate logistic and probit models for estimating these conditional probabilities and the respective standard errors.

Since the form of the likelihood function of the multivariate logistic model is very complex, the elements of the gradient and the Hessian matrix cannot be computed analytically. We have used quasi–Newton method to solve the system of equations for estimating the parameters and the corresponding standard errors of the multivariate logistic model. For estimating the parameters of the GEE2 method, we have used a Newton–Raphson routine with the gradient and Hessian matrix which are computed by FADBAD, a differentiation package. FADBAD is relatively a new differentiation package which uses automatic differentiation to compute the derivatives of a given function. Automatic differentiation could be useful elsewhere in statistical applications for solving system of non–linear equations or optimizing a function where analytic derivatives are too difficult to obtain with symbolic manipulation software.

To compare the maximum likelihood and estimating equation based approaches, we have carried out a simulation study with different pedigrees. The multivariate Plackett construction does not have a closed form expression of the joint cumulative distribution function, so generating multivariate binary data from the multivariate Plackett construction is difficult. For our simulation study, we used the multivariate probit model to generate multivariate binary data. The effects of the pedigree sizes, the types of pedigree, the dependence structure, and the sample sizes have been studied for the comparison of these two approaches. To define the regression model corresponding to the univariate margins, only the covariate Age is used. No covariates have been used for the model corresponding to the dependence structure.

Chapter 5 contains the simulation results and the description of the estimation procedures that are used for the methods considered in this thesis. The maximum likelihood and estimating equation based estimates of the parameters defined for the univariate margins and dependence structures are found to be quite similar for all the pedigrees we have considered. For samples of size 200, the estimating equation based estimates are marginally more efficient than the maximum likelihood based estimates. But for samples of size 600, there is no difference between the efficiency of the corresponding estimates of these two approaches. The estimates of the conditional probabilities and the respective standard errors are similar for the multivariate logistic and probit models. The estimating equation based method GEE2 requires the computation only to the fourth–order moments, whereas the maximum likelihood method based on the multivariate Plackett construction requires the computation of all the moments up to the dth–order (d is the family size). As a consequence, this of this maximum likelihood based method requires more computation time for estimating the parameters than the estimating equation based approach if the family size ≥ 6.

Appendix A

A.1

$$\operatorname{cov}(y_{ij},\, y_{ik}) = \begin{cases} E(y_{ij})\{1 - E(y_{ij})\} & \text{if } j = k, \\ E(y_{ij}\, y_{ik}) - E(y_{ij})E(y_{ik}) & \text{if } j \neq k. \end{cases}$$

$$\operatorname{cov}(y_{ij},\, y_{ik}\, y_{il}) = \begin{cases} E(y_{ij}\, y_{ik}\, y_{il}) - E(y_{ij})E(y_{ik}\, y_{il}) & \text{if } j \neq k \neq l \\ E(y_{ij}\, y_{ik}) - E(y_{ij})E(y_{ik}\, y_{il}) & \text{if } j = l \neq k \\ E(y_{ij}\, y_{il}) - E(y_{ij})E(y_{ik}\, y_{il}) & \text{if } j = k \neq l \end{cases}$$

$$\operatorname{cov}(y_{ij}\, y_{ik},\, y_{il}\, y_{is}) = \begin{cases} E(y_{ij}\, y_{ik}\, y_{il}\, y_{is}) - E(y_{ij}\, y_{ik})E(y_{il}\, y_{is}) & \text{if } j \neq k \neq l \neq s \\ E(y_{ij}\, y_{ik}\, y_{il}) - E(y_{ij}\, y_{ik})E(y_{il}\, y_{is}) & \text{if } j = s \text{ or } k = s \\ E(y_{ij}\, y_{ik}\, y_{is}) - E(y_{ij}\, y_{ik})E(y_{il}\, y_{is}) & \text{if } j = l \text{ or } k = l \\ E(y_{ij}\, y_{ik})\{1 - E(y_{ij}\, y_{ik})\} & \text{if } (j = l\ \&\ k = s) \text{ or} \\ & \quad (j = s\ \&\ k = l) \end{cases}$$

A.2

$$\left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\pi}}\right)^{-1} = \begin{array}{c|cccccccc}
 & \pi_{111} & \pi_{121} & \pi_{211} & \pi_{221} & \pi_{112} & \pi_{122} & \pi_{212} & \pi_{222} \\
\eta_0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\eta_1 & \pi_{1\cdot\cdot}^{-1} & \pi_{1\cdot\cdot}^{-1} & -\pi_{2\cdot\cdot}^{-1} & -\pi_{2\cdot\cdot}^{-1} & \pi_{1\cdot\cdot}^{-1} & \pi_{1\cdot\cdot}^{-1} & -\pi_{2\cdot\cdot}^{-1} & -\pi_{2\cdot\cdot}^{-1} \\
\eta_2 & \pi_{\cdot1\cdot}^{-1} & -\pi_{\cdot2\cdot}^{-1} & \pi_{\cdot1\cdot}^{-1} & -\pi_{\cdot2\cdot}^{-1} & \pi_{\cdot1\cdot}^{-1} & -\pi_{\cdot2\cdot}^{-1} & \pi_{\cdot1\cdot}^{-1} & -\pi_{\cdot2\cdot}^{-1} \\
\eta_3 & \pi_{\cdot\cdot1}^{-1} & \pi_{\cdot\cdot1}^{-1} & \pi_{\cdot\cdot1}^{-1} & \pi_{\cdot\cdot1}^{-1} & -\pi_{\cdot\cdot2}^{-1} & -\pi_{\cdot\cdot2}^{-1} & -\pi_{\cdot\cdot2}^{-1} & -\pi_{\cdot\cdot2}^{-1} \\
\eta_{12} & \pi_{11\cdot}^{-1} & -\pi_{12\cdot}^{-1} & -\pi_{21\cdot}^{-1} & \pi_{22\cdot}^{-1} & \pi_{11\cdot}^{-1} & -\pi_{12\cdot}^{-1} & -\pi_{21\cdot}^{-1} & \pi_{22\cdot}^{-1} \\
\eta_{13} & \pi_{1\cdot1}^{-1} & \pi_{1\cdot1}^{-1} & -\pi_{2\cdot1}^{-1} & -\pi_{2\cdot1}^{-1} & -\pi_{1\cdot2}^{-1} & -\pi_{1\cdot2}^{-1} & \pi_{2\cdot2}^{-1} & \pi_{2\cdot2}^{-1} \\
\eta_{23} & \pi_{\cdot11}^{-1} & -\pi_{\cdot21}^{-1} & \pi_{\cdot11}^{-1} & -\pi_{\cdot21}^{-1} & -\pi_{\cdot12}^{-1} & \pi_{\cdot22}^{-1} & -\pi_{\cdot12}^{-1} & \pi_{\cdot22}^{-1} \\
\eta_{123} & \pi_{111}^{-1} & -\pi_{121}^{-1} & -\pi_{211}^{-1} & \pi_{221}^{-1} & \pi_{112}^{-1} & -\pi_{122}^{-1} & -\pi_{212}^{-1} & \pi_{222}^{-1}
\end{array}$$

Bibliography

[1] Andrade, M. de, Amos, C. I. and Thiel, T. J. (1999). Methods to estimate genetic components of variance for quantitative traits in family studies. *Genetic Epidemiology*, **17**, 64-76.

[2] Ashford, J. R. and Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics*, **26**, 535–46.

[3] Cox, D. R. (1972). Analysis of multivariate binary data. *Applied Statistics*, **21**, 113–20.

[4] Crowder, M. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika*, **82**, 407–10.

[5] Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete and ordered responses. *Biometrics*, **42**, 909–17.

[6] Dixon, L. C. W. (1991). On the impact of automatic differentiation on the relative performance of parallel truncated Newton and variable metric algorithms. *SIAM J. Opt.*, **1**, 475–86.

[7] Falconer, D. S. (1960). *Introduction to Quantitative Genetics.* The Ronald Press, New York.

[8] Finney, D. J. (1971). *Probit Analysis, Third edition.* Cambridge University Press, London.

[9] Fitzmaurice, G. M. and Laird N. M. (1993). A likelihood–based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–51.

[10] Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science*, **8**, 284-99.

[11] Glonek, G. F. V. and McCullagh, P. (1995). Multivariate logistic models. *J. Roy. Stat. Soc. B*, **57**, 533–46.

[12] Godambe, V. P. (ed.) (1991). *Estimating Functions.* Oxford University Press, Oxford.

[13] Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo maximum likelihood methods: theory. *Econometrica*, **52**, 681–700.

[14] Griewank, A. (1989). On automatic differentiation. *Mathematical Programming*, Kluwer Academic Publishers, Japan.

[15] Joe, H. (1995). Approximations to multivariate normal rectangle probabilities based on conditional expectations. *J. Amer. Statist. Assoc.*, **90**, 957-64.

[16] Joe, H. and Liu, Y. (1996). A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions. *Statistics & Probability Letters*, **31**, 113–20.

[17] Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman & Hall, London.

[18] Joe, H. (1999). C routine for multivariate Plackett distribution based on Molenberghs–Lesaffre construction. `ftp://ftp.stat.ubc.ca/pub/hjoe/famil/`.

[19] Joe, H. (1999a). C routine for multivariate probit model. `ftp://ftp.stat.ubc.ca/pub/hjoe/famil/`.

[20] Joe, H. (2001). C++ rountine for modified Newton–Raphson using FADBAD. `ftp://ftp.stat.ubc.ca/pub/hjoe/famil/`.

[21] Kepner, J. L., Harper, J. D., and Keith, S. Z. (1989). A note on evaluating a certain orthant probability. *Amer. Statist.*, **43**, 48–9.

[22] Lesaffre, E. and Molenberghs, G. (1991). Multivariate probit analysis: A neglected procedure in medical statistics. *Statistics in Medicine*, **10**, 1391–1401.

[23] Liang, K.–Y. and Zeger, S. C. (1986). Longitudinal data analysis with generalized linear models. *Biometrika*, **73**, 13–22.

[24] Liang, K.–Y. and Beaty, T. H. (1992). Measuring familial aggregation by using odds–ratio regression models. *Genetic Epidemiology*, **8**, 361–70.

[25] Liang, K.–Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analysis for categorical data (with discussion). *J. Roy. Stat. Soc. B*, **54**, 3–40.

[26] Lindsey, J. K. (1997). *Applying Generalized Linear Models.* Springer–Verlag New York, Inc.

[27] Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika*, **78**, 153–60.

[28] Mardia K. V. (1970). *Families of Bivariate Distributions.* Griffin, London.

[29] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Second Edition.* Chapman & Hall, London.

[30] Mendell, N. R. and Elston, R. C. (1974). Multifactorial qualitative traits: Genetic analysis and prediction of recurrence risks. *Biometrics*, **30**, 41–57.

[31] Molengerghs, G. and Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *J. Amer. Statist. Assoc.*, **89**, 633–44.

[32] Nash, J. C. (1979). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation.* Adam Hilger Ltd., Bristol.

[33] Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression analysis. *Biometrika*, **71**, 531–43.

[34] Plackett, R. L. (1965). A class of bivariate distributions. *J. Amer. Statist. Assoc.*, **60**, 516–22.

[35] Sutradhar, B. C. and Das, K. (1999). On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika*, **86**, 459–65.

[36] Wolfe, M. A. (1978). *Numerical Methods for Unconstrained Optimization : an introduction.* Van Nostrand Reinhold Company Ltd., England.

[37] Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.

[38] Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–48.