INFERENCE ON LOGISTIC REGRESSION MODELS

Mamunur Rashid

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2008

Committee:

John T. Chen, Advisor

Ray Kresman
Graduate Faculty Representative

Hanfeng Chen

Maria Rizzo

ABSTRACT

John T. Chen, Advisor

The logistic regression model is one of the popular mathematical models for the analysis of binary data with applications in physical, biomedical, and behavioral sciences, among others. The feature of this model is to quantify the effects of several explanatory variables on one dichotomous outcome variable. Normally, the asymptotic properties of the maximum likelihood estimates in the model parameters are used for statistical inference. However, logistic regression models have serious numerical problems if zero cells occur in the contingency table. For this scenario, this dissertation proposed a new approach to investigate the asymptotic properties of maximum likelihood estimators for the logistic regression models. In this dissertation, a generalization of the hybrid logistic regression model was introduced, which was originally proposed by Chen et al. (2003). These models deal with situations in which risk factors associated with the outcome are exceedingly rare in the control group. In principle, a two-stage hybrid procedure models the risks due to the rare factors in the first stage and models the residual risks due to the other factors in the second stage using the standard logistic regression model.

Another highlight of this dissertation is on the multinomial logistic regression model, which handles the categorical dependent outcome variable with more than two levels. It extended the hybrid logistic regression model to the multinomial hybrid logistic regression model when the case group of the outcome variable has mutually exclusive and exhaustive subgroups. In the last part of the dissertation, we studied the bootstrap method to estimate the variances for the parameter estimates in the logistic regression model.

ACKNOWLEDGMENTS

I would like to thank my advisor Professor John T. Chen for his support and many thoughtful suggestions throughout this research. I also want to extend my gratitude to the other members of my committee, Dr. Ray Kresman, Dr. Hanfeng Chen, Dr. Maria Rizzo, and Dr. Mike Earley for their time and advice. I extend special thanks to all my professors in the Department of Mathematics and Statistics.

I am also grateful to the Department of Mathematics and Statistics at Bowling Green State University for the teaching assistantship and non-service fellowship awarded to me during my study. My very sincere thanks go to Marcia Seubert, Cyndi Patterson, and Mary Busdeker for their assistance.

Finally, my deepest gratitude goes to my parents and uncles for their prayers and encouragement, and to my family Naima, Maieasha, and Naiar for their love, patience, and sacrifice during my study.

Bowling Green, Ohio Mamunur Rashid

August, 2008

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Regression analysis is one of the most useful and the most frequently used statistical methods (Efron and Tibsirani, 1993). The aim of the regression methods is to describe the relationship between a response variable and one or more explanatory variables. Among the different regression models, logistic regression plays a particular role. The basic concept, however, is universal. The linear regression model is, under certain conditions, in many circumstances a valuable tool for quantifying the effects of several explanatory variables on one dependent continuous variable. For situations where the dependent variable is qualitative, however, other methods have been developed. One of these is the logistic regression model, which specifically covers the case of a binary (dichotomous) response. Cramer (2003) discussed an overview of the development of the logistic regression model. He identifies three sources that had a profound impact on the model: applied mathematics, experimental statistics, and economic theory. Agresti (2002) also provided details of the development on logistic regression in different areas. He states, "Sir David R. Cox introduced many statisticians to logistic regression through his 1958 article and 1970 book, *The Analysis of Binary Data*." However, logistic regression is widely used as a popular model for the analysis of binary data with the areas of applications including physical, biomedical, and behavioral sciences. For example, Cornfield (1962) presented the preliminary results from the Framingham Study. The purpose of the study was to find the roles of risk factors of cholesterol levels (low versus high values) and blood pressure (low versus high values) in the development of coronary heart disease (yes or no) in the population of the town.

The logistic regression model can be easily modified to handle the case in which the

outcome variable is nominal with more than two levels (Hosmer and Lameshow, 2000). An extension of the logistic regression model is called the multinomial logistic regression model, when the categorical dependent outcome variable has more than two levels (Chan, 2004). For example, Zocchi and Atkinson (1999) note that in their multinomial logistic regression model on the dose level experiment to measure the influence of gamma radiation on the emergence of house flies, three disjoint outcomes occurred: death before the pupae opened, death during emergence, and life after emergence. A modification of the logistic regression model, known as the discrete choice model, was first proposed by McFadden (1974). The model is also known as multinomial, polytomous, or polychotomous logistic regression in the health sciences and as the discrete choice model in econometrics (Breslow and Powers, 1978). The maximum likelihood estimation (MLE) is the most widely-used general method of estimation procedures and is treated as a standard approach to parameter estimation and inference in statistics (van der Vaart, 1998).

In this dissertation, the logistic and multinomial logistic regression models, as well as the maximum likelihood procedure for the estimation of their parameters, are introduced in detail. Based on two real data sets, an attempt has been made to illustrate the application of the logistic and multinomial logistic regression models.

The MLE has good asymptotic (large sample) properties for the estimates. Under very general conditions, maximum likelihood estimates are consistent, asymptotically efficient, and asymptotically-normally distributed. Notice that this normality allows one to compute the confidence interval and perform statistical tests in a manner analogous to the analysis of linear multiple regression models, provided the sample size is large. However, asymptotic properties of the maximum likelihood (ML) estimator in logistic models had been studied earlier, see, for

example, Gourieroux and Monfort (1981) and Amemiya (1985), and different results have been established. For example, different proofs of consistency can be found in the literature such as Beer (2005), Gourieroux and Monfort (1981), and Amemiya (1985). All of them are based upon the fact that the probability of the existence of the estimators approaches one as sample size tends to infinity. Furthermore, they proceed on the assumption that the number of explanatory variables is fixed. In other words, the number of variables is compelled to remain constant while the sample size increases. Another result presented by Beer (2001) enables us to relax the former condition. It allows for any number of variables, but depends on sample size, and examines the relationship between the number of variables and sample size that is necessary to preserve the consistency of the estimators.

This dissertation focuses on a completely different approach to investigate the asymptotic properties of maximum likelihood estimators for logistic regression models. More precisely, we are going to show that the maximum likelihood estimators converge under certain conditions to the real value of the parameters if the number of observations tends to infinity. To show this, we follow the theorem described by Lehman and Casella (1998) in which the asymptotic properties of maximum likelihood estimators hold if certain regularity conditions are satisfied. It needs to be pointed out that none of the authors cited above verified their work via the Monte Carlo simulation study. Gourieroux and Monfort (1981) note, "it should be stressed that all these asymptotic results give little indication on the properties of the estimators in finite sample, and it would be interesting to clarify this point by means of Monte Carlo studies." In this dissertation, we also provide an extensive standard Monte Carlo simulation study to show the consistency and asymptotic normality of the ML estimators of the logistic and multinomial logistic regression models.

Logistic regression encounters serious numerical calculation problems, especially for overestimating parameter coefficients and their standard errors for both the outcome variable and individual covariates if there are zero frequencies in the contingency tables (Agresti, 2002 and Hosmer and Lameshow, 2000). The analysis of the contingency table with few or zero cell counts can have two types of problems. One class of problems associated with such contingency tables is related to the goodness of fit testing since the asymptotic approximations of the standard chi-squared statistics tend to be poor for these tables. Another class of problems is related to the non-existence of the ML estimates and the asymptotic standard errors for logit models. More precisely, sometimes parameter estimates take on values plus or minus infinity. In such a situation, the Newton-Raphson algorithm may not converge (Clogg et al., 1991). Even if the ML estimates exist, they may be biased. However, there are some strategies to remove the zeroes in the contingency table and then apply the logistic regression model. Among many strategies, one general approach is adding 0.5 to each cell to perform statistical analysis. Haldane (1956) suggested a correction term of 0.5 to add to all four cells prior to analysis of the 2×2 contingency table. Goodman (1970,1971) recommends this procedure for a saturated model only. Agresti (1996, 2002) suggests adding a very small constant to cell counts and doing a sensitivity analysis to determine the smallest such number to add to the zero cells. Hosmer and Lemeshow (2000) recommend three strategies: collapsing the categories of the covariates in a sensible way, eliminating a category completely, or modeling ordinal variables as if they are continuous variables.

There is another approach discussed by Agresti (2002) for smoothing contingency tables called the pseudo-Bayes approach. This approach provides a way of smoothing the data in a less *ad hoc* manner than adding an arbitrary constant to cells. The pseudo-Bayes approach, originally

proposed by Bishop et al. (1975), postulates that their method is superior to the generally-accepted practice of adding 0.5 to the count in each cell of a large, sparse table. The details of this method are discussed in Chapter 3.

Each of the remedial approaches generates positive adjusted counts for all cells; the adding 0.5 approach and the pesudo-Bayes approach both generate cell counts that are equal across both response groups. On the other hand, the pseudo-Bayes estimation approach is of limited usefulness because of the difficulty in setting values for the λ parameter. The effect of setting $\lambda = (T^{-1}, \mathrm{K}\ , T^{-1})$ is essentially to smooth each cell by the same constant. Hence, cells having equal counts prior to smoothing will have different counts but become equal after smoothing (Dillon et al., 1981).

1.1 The Hybrid Logistic Regression Model

Avoiding such problems, Chen et al. (2003) proposed another method called the hybrid logistic regression model for use in case-control studies. It is worthwhile to mention that the odds ratio estimators are the same for both cohort and case-control studies (see, for example, Cornfield, 1951 and Prentice, 1976). Prentice and Pyke (1979) showed that the odds ratio estimators and their asymptotic covariance matrices may be obtained by applying the prospective (cohort) logistic regression model to the case-control study as if the data had been obtained in a prospective (cohort) study. Others with significant contributions to the logistic regression for case-control studies include Breslow (1996), Anderson (1972), Fears and Brown (1986), Breslow and Cain (1988), Scott and Wild (1986,1991), Farewell (1979), Zhang (2006), Breslow and Powers (1978), and Mantel (1973).

The basic idea of the hybrid logistic regression model (Chen et al., 2003) for case-control studies is that if there is a rare disease in the control group for some risk factors, then the

estimation of the parameters for those risk factors is difficult to achieve. To avoid such troublesome risk factors, in practice, investigators (for example, Shaffer et al., 1996 and Brent et.al., 1999) usually do not include such risk factors and consider instead the other risk factors. However, their approach spreads the contribution of the troublesome risk factors among the remaining factors in the model and may consequently result in an overestimate of the odds ratio of the later in the model. In sum, Chen et al. (2003) proposed the hybrid logistic regression model because previous work on backward and forward logistic regression models do not account for the proper handling of troublesome risk factors. In the hybrid logistic regression model, Chen et al. (2003) adjust the troublesome risk factor first and then model the rest of the risk factors by using regular logistic regression. They note that the rare risk factor is considered as univariate.

One contribution of this dissertation is that we develop a theoretical generalization of Chen et al. (2003) procedure for *k*-variate rare risk factors. In addition, we also extend the hybrid logistic regression model to a multinomial hybrid logistic regression model under a case-control study. In this case, we assume that there are *k* mutually exclusive and exhaustive disease groups existing in the case group. In the generalization of the hybrid logistic regression model, when we estimate the troublesome risk factors, we adjust them with possible combinations of all risk factors. In this case, we assume that proportions of having diseases are equal across all strata. As a result, it would not be convenient to deal with it in practice if the risk factors having more levels are included in the model. In addition, in the model fitting strategies, we consider all risk factors as well as their possible interaction terms. In the hybrid logistic model, it would be intricate not only to consider the interaction terms in the model, but also to estimate and interpret the parameters for those terms.

Efron (1979) introduced a very general resampling procedure called the bootstrap for estimating the distributions of statistics based on independent observations. The term 'bootstrapping,' is an allusion to the expression 'pulling oneself up by one's bootstraps' – in this case, using the sample data as a population from which repeated samples are drawn. As a result, by making use of numerous samples drawn from the initial observation, these techniques require fewer assumptions and offer greater accuracy and insight than do standard methods (Stine, 1989). The use of this technique plays a central role in statistics, especially when the estimators of interest do not have an explicit formula. The first approach in the development of bootstrap methods is the non-parametric bootstrap, followed by parametric and the Bayesian approaches. Efron (1979) considered a variety of statistical problems and showed how easy it is to apply this simulation method. The bootstrap has been the object of much research in statistics since its introduction. For the linear regression model, Freedman (1981) and Wu (1986) discussed the asymptotic properties using the bootstrap method. Moulton and Zeger (1991) used a bootstrap technique for generalized linear models (GLMs).

In this dissertation, we study the bootstrap strategies to evaluate its performance in estimating the variances for the logistic regression model. Friedl and Tilg (1995) used the one-step bootstrap procedure for the variance estimates in the logistic regression model based on the residual resampling, which was introduced by Moulton and Zeger (1991) for the whole class of GLMs. To use the vector resampling method for the GLMs, Moulton and Zeger (1991) mentioned two problems: first, if the sample size is large, several iterations might be needed for each bootstrap replication. As a result, the computational cost may be quite high. Another problem they pointed out was that obtaining extreme data replications would result in the parameter estimates failing to converge. However, Carroll et al. (2006) mentioned that as a

general-purpose technique, the vector resampling procedure can be used for the logistic regression model. In this dissertation, we implement the vector resampling procedure for the variance estimation of the parameters in the logistic regression model.

1.2 Chapter Outline

The dissertation is structured as follows: Chapter 1 introduces the motivation of this study. Chapter 2 discusses the estimation procedures and the interpretation of the parameters in the logistic regression model and relates the model to prospective and case-control studies. Chapter 2 also provides the asymptotic properties of the model and presents results based on an extensive simulation study. In addition, an application of the logistic regression model based on a real data set is given at the end of this chapter. Chapter 3 discusses theoretical aspects of the pseudo-Bayes approach, introduces the hybrid logistic regression model and its extension to the *k*-rare risk factors, and provides the estimation procedure for the model. Chapter 4 discusses the multinomial distribution and its parameter estimation procedure, introduces the multinomial logistic regression model and the estimation of parameters for the models, and provides a simulation study to show the consistency and normality of the ML estimators. In Chapter 4, an application of the multinomial logistic regression model is illustrated. Furthermore, this chapter introduces the multinomial hybrid logistic regression model, and the estimation procedures of the model parameters are discussed. Chapter 5 discusses the bootstrap method and its consistency, applies the bootstrap method to the regression model, and employs the vector resampling procedure to the logistic regression model to estimate the variances. Finally, Chapter 6 presents concluding remarks and suggestions for future research.

CHAPTER 2

PROPERTIES OF ESTIMATES FOR THE PARAMETERS IN THE LOGISTIC REGRESSION MODEL

2.1 The Logistic Regression Model

Suppose a binary random variable *y* follows a Bernoulli distribution, that is, *y* takes either the value 1 or the value 0 with probabilities $\pi(x)$ or $1-\pi(x)$ respectively, where $x=(x_1,x_2,\ldots,x_p)\in\Re^p$ is a vector of *p* explanatory variables. In fact, $\pi(x)$ represents the conditional probability P(*y* =1|*x*) of *y* =1, given *x*. Based on the binary outcome variable, we use the logistic distribution (see, for example, Cox and Snell, 1989; Hosmer and Lameshow, 2000). The specific form of the logistic regression model with unknown parameters $\beta_0,\beta_1,\ldots,\beta_p$ is

$$\pi(x)=\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_px_p}}$$

At times, it is convenient to change the notation slightly by writing $x_0=1$, thus the above model becomes

$$\pi(x)=\frac{e^{x^T\beta}}{1+e^{x^T\beta}} \tag{2.1.1}$$

where $x=(x_0,x_1,\ldots,x_p)^T$ and $\beta=(\beta_0,\beta_1,\ldots,\beta_p)^T$.

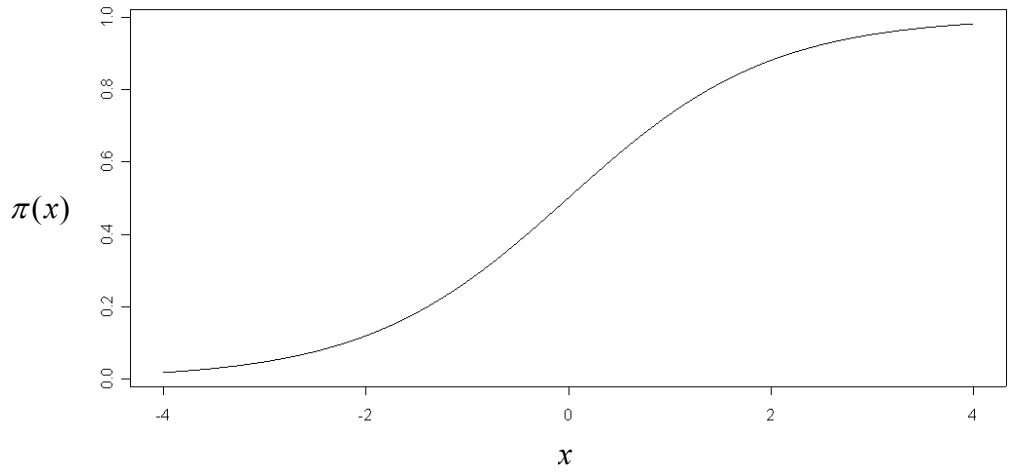
A transformation of $\pi(x)$ is called the *logit transformation,* and is given by

$$\operatorname{logit}\pi(\mathrm{x})=\ln\frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})} \tag{2.1.2}$$

Under the above transformation, we can write the regression model (2.1.1) as

$$\operatorname{logit}\pi(\mathrm{x})=\mathrm{x}^{\mathrm{T}}\beta \tag{2.1.3}$$

Figure 2.1.1: Logistic regression function, $\pi(x) = \dfrac{\exp(x)}{1+\exp(x)}$



2.2 Maximum Likelihood (ML) Estimation of the Parameters

Suppose we have a sample of *n* independent observations $\{(y_i, x_i)\}_{i\in\{1,2,\ldots,n)}$ $\in (\{0,1\}\times\Re^{p+1})^n$, where y_i denotes the value of a dichotomous outcome variable, and x_i is the value of the explanatory variables for the *i*th subject. Assume

$$y_i \sim Bernoulli(1, \pi(x_i)),\ i = 1,2,\ldots,n$$

Based on a set of data, we estimate the parameter vector $\beta = (\beta_0, \beta_1, ..., \beta_p)^T \in \Re^{p+1}$ to fit the logistic regression model in equation (2.1.1). To find the ML estimator of β, we define the likelihood function as follows

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1-\pi(x_i))^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)^{y_i} (1-\pi(x_i))$$

$$= \prod_{i=1}^{n} \left(\frac{\frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}}{\frac{1+e^{x_i^T \beta} - e^{x_i^T \beta}}{1+e^{x_i^T \beta}}} \right)^{y_i} \left(\frac{1+e^{x_i^T \beta} - e^{x_i^T \beta}}{1+e^{x_i^T \beta}} \right)$$

$$= \prod_{i=1}^{n} \left(\frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}} \cdot \frac{1+e^{x_i^T \beta}}{1} \right)^{y_i} \cdot \frac{1}{1+e^{x_i^T \beta}}$$

$$= \prod_{i=1}^{n} \frac{(e^{x_i^T \beta})^{y_i}}{1+e^{x_i^T \beta}}$$

$$= \prod_{i=1}^{n} \frac{e^{y_i x_i^T \beta}}{1+e^{x_i^T \beta}}$$

Now, we find the ML estimates, $\hat{\beta}$, of β by maximizing the log-likelihood function for the observed values of y_i and x_i. Since maximizing the log of a function is equivalent to maximizing the function, we often work with the log-likelihood because it is generally less cumbersome to use for mathematical operations, such as differentiation. Therefore, the log-likelihood function yields,

$$\ell(\beta) = \sum_{i=1}^{n} y_i x_i{}^T \beta - \sum_{i=1}^{n} \ln(1+e^{x_i{}^T \beta}) \tag{2.2.1}$$

The first derivative of the log-likelihood function gives the gradient.

We have the first derivative of $x_i{}^T \beta$ with respect to β_j is x_{ij}, so

$$\frac{\delta\, \ell(\beta)}{\delta \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{e^{x_i{}^T \beta}}{1+e^{x_i{}^T \beta}} \cdot x_{ij}$$

$$= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \pi_i x_{ij}$$

$$= \sum_{i=1}^{n} (y_i - \mu_i) x_{ij}, \text{ where } \mu_i = E(y_i) = \pi_i \tag{2.2.2}$$

The second derivatives are

$$\frac{\delta^2 \ell(\beta)}{\delta\beta_j \, \delta\beta_k} = -\sum_{i=1}^{n} x_{ij} \frac{\delta}{\delta\beta_k}\left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)$$

$$= -\sum_{i=1}^{n} x_{ij}\left[\frac{(1+e^{x_i^T\beta})\, e^{x_i^T\beta}\, x_{ik} - e^{x_i^T\beta}\, e^{x_i^T\beta}\, x_{ik}}{(1+e^{x_i^T\beta})^2}\right]$$

$$= -\sum_{i=1}^{n} x_{ij}x_{ik}\left[\frac{e^{x_i^T\beta}\,(1+e^{x_i^T\beta} - e^{x_i^T\beta})}{(1+e^{x_i^T\beta})^2}\right]$$

$$= -\sum_{i=1}^{n} \pi_i(1-\pi_i)x_{ij}x_{ik} \qquad (2.2.3)$$

The third derivatives are

$$\frac{\delta^3 \ell(\beta)}{\delta\beta_j \, \delta\beta_k \, \delta\beta_r} = -\sum_{i=1}^{n} x_{ij}x_{ik} \frac{\delta}{\delta\beta_r}\left(\frac{e^{x_i^T\beta}}{(1+e^{x_i^T\beta})^2}\right)$$

$$= -\sum_{i=1}^{n} x_{ij}x_{ik}\left(\frac{(1+e^{x_i^T\beta})^2 e^{x_i^T\beta}x_{ir} - e^{x_i^T\beta}\cdot 2(1+e^{x_i^T\beta})e^{x_i^T\beta}x_{ir}}{(1+e^{x_i^T\beta})^4}\right)$$

$$= -\sum_{i=1}^{n} x_{ij}x_{ik}x_{ir}\left(\frac{(1+e^{x_i^T\beta})e^{x_i^T\beta}(1+e^{x_i^T\beta} - 2\, e^{x_i^T\beta})}{(1+e^{x_i^T\beta})^4}\right)$$

$$= -\sum_{i=1}^{n} x_{ij}x_{ik}x_{ir}\left(\frac{e^{x_i^T\beta}(1-e^{x_i^T\beta})}{(1+e^{x_i^T\beta})^3}\right)$$

$$= -\sum_{i=1}^{n} \pi_i(1-\pi_i)(1-2\pi_i)x_{ij}x_{ik}x_{ir} \qquad (2.2.4)$$

Let $y = \begin{bmatrix} y_1 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$, $X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$, and $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$.

Notice that y and μ are $n \times 1$, X is $n \times (p+1)$, and the elements of μ are non-linear functions of an assumed value for β. Also, we define

$$W = diag\,(\pi_i(1-\pi_i))$$

which is $n \times n$. Then, we can write the gradient

$$\ell'(\beta) = \frac{\delta\,\ell(\beta)}{\delta\,\beta_j} = X^T(y-\mu)$$

and the Hessian matrix

$$\ell''(\beta) = \frac{\delta^2\,\ell(\beta)}{\delta\beta_j\,\delta\beta_k} = -X^T W X\,.$$

Now, we are going to show that $l''(\beta)$ is negative semi-definite for any $\beta \in \Re^{p+1}$.

We have, $u^T \ell''(\beta)u = -u^T X^T W X u = -\sum_{i=1}^{n}(x_i^T u)^2\, diag(\pi_i(1-\pi_i))$. As $diag\,(\pi_i(1-\pi_i))$ is always positive, we can see that $u^T \ell''(\beta)u \le 0$ for all $u \in \Re^{p+1}$ and all $\beta \in \Re^{p+1}$.

Since $\ell''(\beta)$ is negative semi-definite, the log-likelihood, ℓ, is a concave function of the parameter β; several optimization techniques are available for finding the maximizing parameters (see, for example, Mak, 1993; Givens and Hoeting, 2005). We use the Newton-Raphson algorithm for maximizing ℓ. For one step of the Newton-Raphson, we use $\beta^{(t)}$, the current estimate of β, to calculate $\mu^{(t)}$ and $W^{(t)}$. The new estimate of β is then

$$\beta^{(t+1)} = \beta^{(t)} + (X^T W^{(t)} X)^{-1} X^T (y - \mu^{(t)})\,.$$

This process is repeated until the estimates stop changing, that is, until $\beta^{(t+1)}$ is sufficiently close to $\beta^{(t)}$, then we say the Newton-Raphson method converges. To better understand what ensures convergence, we must carefully analyze the errors at successive steps. This can be shown by

using the following theorem, a notation and terminology that differs slightly from that of the theorem discussed by Givens and Hoeting (2005).

Theorem 2.2.1. If $l'''(\beta)$ is continuous and β^* is a simple root of $\ell'(\beta)$, then there exists a neighborhood of β^* for which Newton-Raphson method converges to β^* when started from any $\beta^{(t)}$, $t = 0,1,2\ldots$ in that neighborhood.

Proof: Suppose $\ell'(\beta)$ has two continuous derivatives and $\ell''(\beta^*) \neq 0$. Since $\ell''(\beta^*) \neq 0$ and $\ell''(\beta)$ is continuous at β^*, there exists a neighborhood of β^* within which $\ell''(\beta) \neq 0$ for all β. Let us confine interest to this neighborhood, and define $\varepsilon^{(t)} = \beta^{(t)} - \beta^*$. A Taylor expansion yields

$$0 = \ell'(\beta^*) = \ell'(\beta^{(t)}) + (\beta^* - \beta^{(t)})\,\ell''(\beta^{(t)}) + (\beta^* - \beta^{(t)})^2\,\ell'''(q^{(t)})/2$$

for some $q^{(t)}$ between $\beta^{(t)}$ and β^*. Rearranging terms, we find

$$\beta^{(t)} - \frac{\ell'(\beta^{(t)})}{\ell''(\beta^{(t)})} - \beta^* = (\beta^* - \beta^{(t)})^2\,\frac{\ell'''(q^{(t)})}{2\ell''(\beta^{(t)})} \tag{2.2.5}$$

$$\text{or,} \qquad \varepsilon^{(t)}\ell''(\beta^{(t)}) - \ell'(\beta^{(t)}) = \frac{1}{2}\ell'''(q^{(t)})(\varepsilon^{(t)})^2 \tag{2.2.6}$$

Since the left hand side of equation (2.2.5) equals $\beta^{(t+1)} - \beta^*$, we conclude

$$\varepsilon^{(t+1)} = \frac{\ell'''(q^{(t)})}{2\,\ell''(\beta^{(t)})}(\varepsilon^{(t)})^2 \approx \frac{1}{2}\frac{\ell'''(\beta^*)}{\ell''(\beta^*)}(\varepsilon^{(t)})^2 = C(\varepsilon^{(t)})^2 \tag{2.2.7}$$

This implies that the rate of convergence of Newton-Raphson method is quadratic.

Now, consider a neighborhood of β^*, $N_\delta(\beta^*) = [\beta^* - \delta,\ \beta^* + \delta]$, for $\delta > 0$. Let

$$c(\delta) = \max_{\beta_1,\beta_2 \in N_\delta(\beta^*)} \frac{1}{2}\left|\frac{\ell'''(\beta_1)}{\ell''(\beta_2)}\right|.$$

Since $c(\delta) \to \frac{1}{2}\left|\frac{\ell'''(\beta^*)}{\ell''(\beta^*)}\right|$ as $\delta \to 0$, it follows that $\delta c(\delta) \to 0$ as $\delta \to 0$. Let us choose δ such that $\delta c(\delta) < 1$.

Having fixed δ, set $\rho = \delta c(\delta)$.

Suppose start Newton-Raphson method with $\beta^{(0)}$ satisfying $\left|\beta^{(0)} - \beta^*\right| \le \delta$. Then

$$\left|\varepsilon^{(0)}\right| \le \delta \text{ and } \left|q^{(0)} - \beta^*\right| \le \delta$$

This implies by definition of $c(\delta)$

$$\frac{1}{2}\frac{\left|\ell'''(q^{(0)})\right|}{\left|\ell''(\beta^{(0)})\right|} \le c(\delta)$$

By (2.2.6),

$$\left|\beta^{(1)} - \beta^*\right| = \left|\varepsilon^{(1)}\right| \le (\varepsilon^{(0)})^2 c(\delta) = \left|\varepsilon^{(0)}\right| \left|\varepsilon^{(0)}\right| c(\delta) \le \left|\varepsilon^{(0)}\right| \delta c(\delta) = \left|\varepsilon^{(0)}\right| \rho < \left|\varepsilon^{(0)}\right| \le \delta$$

$\therefore$ $\beta^{(1)}$ lies within δ distance to β^*. Repeating,

$$\left|\varepsilon^{(1)}\right| \le \rho \left|\varepsilon^{(0)}\right|$$

$$\left|\varepsilon^{(2)}\right| \le \rho \left|\varepsilon^{(1)}\right| \le \rho^2 \left|\varepsilon^{(0)}\right|$$

$$\left|\varepsilon^{(3)}\right| \le \rho^3 \left|\varepsilon^{(0)}\right|$$

$$\vdots$$

$$\left|\varepsilon^{(t)}\right| \le \rho^t \left|\varepsilon^{(0)}\right|$$

Since $0 \le \rho \le 1$ $\lim_{t\to\infty} \rho^t = 0 \quad \therefore \quad \lim_{t\to\infty} \varepsilon^{(t)} = 0$

So, $\beta^{(t)} \to \beta^*$.

Hence, the proof follows.

Therefore, the value at which the Newton-Raphson method converges is the estimate of parameter vector $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$.

2.3 Odds and Odds Ratio

The odds ratio is a measure of association, which quantifies the relationship between an exposure and health outcome from a comparative study. It is the ratio of the odds in favor of getting the disease, if exposed, to the odds in favor of getting the disease, if not exposed. Cox (1970) discussed some general advantages of the odds ratio as a measure of association for binary responses. Bland and Douglas (2000) mentioned that there are mainly three reasons to use the odds ratio. Firstly, they provide an estimate (with confidence interval) for the relationship between two binary variables. Secondly, they enable us to examine the effects of other variables on that relationship, using logistic regression. Thirdly, they have a special and very convenient interpretation.

Therefore, it is essential to introduce the terms *odds* and *odds ratio* in order to discuss binary data and to interpret the logistic regression coefficients. For a probability π of success, the odds are defined to be

$$odds = \frac{\pi}{1-\pi}$$

The odds are nonnegative, with $odds > 1.0$ when a success is more likely than a failure. In a 2×2 table, the probability of success is π_1 in row 1 and π_2 in row 2. Within row 1, the *odds* of success are defined to be

$$odds_1 = \frac{\pi_1}{1-\pi_1}$$

and within row 2, the *odds* of success are defined to be

$$odds_2 = \frac{\pi_2}{1-\pi_2}$$

The ratio of odds from the two rows is called the *odds ratio,* which is given by

$$odds\ ratio = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} \tag{2.3.1}$$

To illustrate the odds ratio, we consider the following table where it reports on the relationship between aspirin use and myocardial infarction (heart attacks) by the Physicians' Health Study Research Group at Harvard Medical School (Agresti, 1996). The Physicians' Health Study was a five-year randomized study testing whether regular intake of aspirin reduces mortality from cardiovascular disease. Every other day, physicians participating in the study took either one aspirin tablet or a placebo. The study was blind: the physicians in the study did not know which type of pill they were taking.

Table 2.3.1: Cross-classification of Aspirin Use and Myocardial Infarction (MI).

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Source: *An introduction to categorical data analysis (Agresti, 1996)*

For the physicians taking the placebo, the estimated odds of MI equal 189/10,845=0.0174 and the estimated odds for those taking aspirin equal 104/10,933=0.0095. Thus, the sample odds

ratio equals 0.0174/0.0095=1.832. This implies that the estimated odds of MI for physicians taking the placebo equal 1.832 times the estimated odds for physicians taking aspirin.

2.4 Interpretation of the Parameter β

To understand the interpretation of the logistic coefficients, we consider here a single explanatory variable coded as either 0 or 1. The *odds* of the outcome being present among individuals with $x = 1$ is defined as $\frac{\pi(1)}{1-\pi(1)}$. Similarly, the *odds* of the outcome being present among individuals with $x = 0$ is defined as $\frac{\pi(0)}{1-\pi(0)}$. The possible values of the logistic probabilities may be displayed in the following table.

Table 2.4.1: Values of the logistic regression model when the independent variable is binary.

Explanatory variable	Outcome variable $y=1$	Outcome variable $y=0$	Total
$x=1$	$\pi(1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$1-\pi(1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	1.0
$x=0$	$\pi(0)=\frac{e^{\beta_{01}}}{1+e^{\beta_{01}}}$	$1-\pi(0)=\frac{1}{1+e^{\beta_{01}}}$	1.0

Therefore, the *odds ratio* is defined as the ratio of the odds for x=1 to the odds for x=0, and is given by

$$\text{Odds ratio} = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}$$

$$= \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \Big/ \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} \Big/ \frac{1}{1+e^{\beta_0}}}$$

$$= \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \cdot \frac{1+e^{\beta_0+\beta_1}}{1}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} \cdot \frac{1+e^{\beta_0}}{1}}$$

$$= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}$$

$$= e^{\beta_1}$$

This implies that the relationship between the odds ratio and an independent dichotomous variable for the logistic regression coefficient is

$$Odds\ ratio = e^{\beta_1} \tag{2.4.1}$$

This tells that the odds on the event that *y* equals 1 increase (or decrease) by the factor e^{β_1} among those with $x = 1$ than among those $x = 0$. Hosmer and Lameshow (2000) state, "this fact concerning the interpretability of the coefficients is the fundamental reason why logistic regression has proven to be such a powerful analytic tool for epidemiologic research."

2.5 Odds Ratio: Prospective versus Retrospective Studies

Cornfield (1951) first studied the invariance of the odds ratio under prospective (cohort) and retrospective (case-control) study designs. Mantel and Haenszel (1959), Mantel (1973), Siegel and Greenhouse (1973), Prentice and Breslow (1978), Santnner and Duffy (1989), Christensen (1997) and so on emphasized the relationship between prospective and retrospective studies. Consider, for instance, an experiment in which 300 people of an arbitrary population are sampled. A binary response "diseased" (*D*) or "non-diseased" ($\overline{D}$) is observed for each person. Then, there is an explanatory variable "exposed" (*E*) or "non-exposed" ($\overline{E}$). This kind of study is called prospective or cohort study. According to Farewell (1979), "In a prospective study of the incidence of a disease, a sample of individuals is drawn from the population of interest, and

risk factors under study are recorded and regarded as fixed variables. The sample is then followed through time to determine disease incidence, viewed as a random event." Let OR_P denote the (prospective) ratio of odds of disease for the exposed group to odds of disease for the non-exposed group as

$$OR_P = \frac{\frac{P(D \mid E)}{1 - P(D \mid E)}}{\frac{P(D \mid \bar{E})}{1 - P(D \mid \bar{E})}}$$

According to the nature of the study, diseased individuals may be very rare in a random sample of 300 people. So, most of the collected data is about non-diseased persons. It is, therefore, sometimes useful to fix the sample size, in the rare event category, by design. In our example, one could possibly study a separate sample of 150 diseased and 150 non-diseased individuals while determining for every person whether he or she had been exposed or not. This procedure is called retrospective or case-control study and leads directly to information about the probability of exposure among the diseased and among the healthy groups. According to Farewell (1979), "The retrospective study consists of separate samples of individuals with the disease under study, termed cases, and of individuals who do not have the disease, termed controls. In this particular study, risk factors are treated as random variables, and the occurrence of disease is regarded as a fixed variable." Let OR_R denote the (retrospective) ratio of odds of disease for the exposed group to odds of disease for the non-exposed group as

$$OR_R = \frac{\frac{P(E \mid D)}{1 - P(E \mid D)}}{\frac{P(E \mid \bar{D})}{1 - P(E \mid \bar{D})}}$$

However, we obtain by Bayes's rule that

$$OR_P = \frac{\frac{P(D \mid E)}{P(\bar{D} \mid E)}}{\frac{P(D \mid \bar{E})}{P(\bar{D} \mid \bar{E})}} = \frac{\frac{P(E \mid D)\,P(D)}{P(E \mid \bar{D})\,P(\bar{D})}}{\frac{P(\bar{E} \mid D)\,P(D)}{P(\bar{E} \mid \bar{D})\,P(\bar{D})}} = \frac{\frac{P(E \mid D)}{P(E \mid \bar{D})}}{\frac{P(\bar{E} \mid D)}{P(\bar{E} \mid \bar{D})}} = \frac{\frac{P(E \mid D)}{1 - P(E \mid D)}}{\frac{P(E \mid \bar{D})}{1 - P(E \mid \bar{D})}} = OR_R \tag{2.5.1}$$

so that we are able to make inferences about OR_P even from a retrospective study.

McCullough and Nelder (1989) pointed out that the logistic function applies for both prospective (cohort) and retrospective (case-control) studies. Therefore, one can easily come up with the model for case-control data and can estimate the parameters.

2.6 Logistic Regression Model Under Case-Control Study

Let the variable s denote the selection $(s = 1)$ or non-selection $(s = 0)$ of a subject. Let $\tau_1 = P(s = 1 \mid y = 1)$ denote the probability of sampling a case, and let $\tau_0 = P(s = 1 \mid y = 0)$ denote the probability of sampling a control. According to Hosmer and Lameshow (2000), the full likelihood for a sample of size n_1 cases ($y = 1$) and n_0 controls ($y = 0$) is

$$\prod_{i=0}^{n_1} P(x_i \mid y_i = 1, s_i = 1) \prod_{i=1}^{n_0} P(x_i \mid y_i = 0, s_i = 1) \tag{2.6.1}$$

For an individual term in the likelihood function shown in equation (2.6.1) yields

$$P(x \mid y, s = 1) = \frac{P(y, s = 1, x)}{P(y, s = 1)}$$

$$= \frac{P(y \mid x, s = 1) \cdot P(x, s = 1)}{P(y, s = 1)}$$

$$= \frac{P(y \mid x, s = 1) \cdot P(x \mid s = 1) \cdot P(s = 1)}{P(y \mid s = 1) \cdot P(s = 1)}$$

$$= \frac{P(y \mid x, s = 1) \cdot P(x \mid s = 1)}{P(y \mid s = 1)}$$

That is, $P(x \mid y, s = 1) = \dfrac{P(y \mid x, s = 1)\,P(x \mid s = 1)}{P(y \mid s = 1)}$ (2.6.2)

The first term in the numerator of equation (2.6.2) for $y = 1$ yields

$$P(y=1 \mid x, s=1) = \frac{P(x, s=1, y=1)}{P(x, s=1)}$$

$$= \frac{P(s=1 \mid x, y=1) \cdot P(x, y=1)}{P(x, s=1)}$$

$$= \frac{P(s=1 \mid x, y=1) \cdot P(y=1 \mid x) \cdot P(x)}{P(s=1 \mid x) \cdot P(x)}$$

$$= \frac{P(s=1 \mid x, y=1) \cdot P(y=1 \mid x)}{P(s=1 \mid x)}$$

$$= \frac{P(s=1 \mid x, y=1) \cdot P(y=1 \mid x)}{P(s=1 \mid x, y=0) \cdot P(y=0 \mid x) + P(s=1 \mid x, y=1) \cdot P(y=1 \mid x)}$$

That is, $P(y=1 \mid x, s=1) = \dfrac{P(s=1 \mid x, y=1) \cdot P(y=1 \mid x)}{P(s=1 \mid x, y=0) \cdot P(y=0 \mid x) + P(s=1 \mid x, y=1) \cdot P(y=1 \mid x)}$ (2.6.3)

Assume that the selection of cases and controls is independent of the covariates with respective probabilities

$$\tau_1 = P(s=1 \mid y=1, x) = P(s=1 \mid y=1),$$

and $$\tau_0 = P(s=1 \mid y=0, x) = P(s=1 \mid y=0).$$

Now, we rewrite the logistic model $\pi(x) = \dfrac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}$ as $\pi(x) = \dfrac{e^{\beta_0 + \beta' x}}{1 + e^{\beta_0 + \beta' x}}$, and then substitution of τ_1, τ_0 and the logistic regression model, $\pi(x)$, for $P(y=1 \mid x)$, into equation (2.6.3) yields

$$P(y=1 \mid x, s=1) = \frac{P(y=1 \mid x) \cdot \tau_1}{P(y=0 \mid x) \cdot \tau_0 + P(y=1 \mid x) \cdot \tau_1}$$

$$= \frac{\tau_1 \cdot \pi(x)}{\tau_0 (1-\pi(x)) + \tau_1 \cdot \pi(x)}$$

Dividing numerator and denominator by $\tau_0 \cdot (1-\pi(x))$, we get

$$P(y=1 \mid x, s=1) = \frac{\frac{\tau_1 \cdot \pi(x)}{\tau_0 \cdot (1-\pi(x))}}{1 + \frac{\tau_1 \cdot \pi(x)}{\tau_0 \cdot (1-\pi(x))}}$$

$$= \frac{\frac{\tau_1}{\tau_0} \cdot \frac{\frac{e^{\beta_0+\beta' x}}{1+e^{\beta_0+\beta' x}}}{1-\frac{e^{\beta_0+\beta' \underline{x}}}{1+e^{\beta_0+\beta' x}}}}{1 + \frac{\tau_1}{\tau_0} \cdot \frac{\frac{e^{\beta_0+\beta' x}}{1+e^{\beta_0+\beta' x}}}{1-\frac{e^{\beta_0+\beta' x}}{1+e^{\beta_0+\beta' x}}}}$$

$$= \frac{\frac{\tau_1}{\tau_0} \cdot \frac{e^{\beta_0+\beta' x}}{1+e^{\beta_0+\beta' x}} \cdot \frac{1+e^{\beta_0+\beta' x}}{1}}{1 + \frac{\tau_1}{\tau_0} \cdot \frac{e^{\beta_0+\beta' x}}{1+e^{\beta_0+\beta' x}} \cdot \frac{1+e^{\beta_0+\beta' x}}{1}}$$

$$= \frac{\frac{\tau_1}{\tau_0} \cdot e^{\beta_0+\beta' \underline{x}}}{1 + \frac{\tau_1}{\tau_0} \cdot e^{\beta_0+\beta' \underline{x}}}$$

$$= \frac{e^{\ln(\frac{\tau_1}{\tau_0})+\beta_0+\beta' \underline{x}}}{1 + e^{\ln(\frac{\tau_1}{\tau_0})+\beta_0+\beta' \underline{x}}}$$

$$= \frac{e^{\beta_0^*+\beta' x}}{1+e^{\beta_0^*+\beta' x}}, \text{ where } \beta_0^* = \ln(\frac{\tau_1}{\tau_0}) + \beta_0$$

Thus, let $\pi^*(x) = P(y=1 \mid x, s=1) = \frac{e^{\beta_0^*+\beta' x}}{1+e^{\beta_0^*+\beta' x}}$

Therefore, the equation (2.6.2) becomes, for $y = 1$,

$$P(x \mid y = 1, s = 1) = \pi^*(x) \cdot \frac{P(x \mid s = 1)}{P(y = 1 \mid s = 1)}$$

Since we assume that sampling is carried out independent of covariate values,

$P(x \mid s = 1) = P(x)$, where $P(x)$ denotes the probability distribution of the covariates.

Thus, $P(x \mid y = 1, s = 1) = \pi^*(x) \cdot \frac{P(x)}{P(y = 1 \mid s = 1)}$

Similarly, $P(x \mid y = 0, s = 1) = (1 - \pi^*(x)) \cdot \frac{P(x)}{P(y = 0 \mid s = 1)}$

If we let $L^*(\beta) = \prod_{i=0}^{n} \pi^*(x_i)^{y_i} [1 - \pi^*(x_i)]^{1-y_i}$ the likelihood function (2.6.1) becomes

$$L^*(\beta) \cdot \prod_{i=1}^{n} \left[\frac{P(x_i)}{P(y_i \mid s_i = 1)} \right] \tag{2.6.4}$$

where, $L^*(\beta)$ is the likelihood obtained where we pretend that the case-control data were collected in a cohort study, with the outcome of interest modeled as the dependent variable. Notice that the estimates of the parameters and variance-covariance matrix can be obtained by any standard computer statistical packages such as SAS and SPSS.

2.7 Asymptotic Properties of the ML Estimators

In the 1920s, R.A. Fisher originally developed the principle of maximum likelihood estimation (MLE) and established optimum properties of estimates by maximizing the likelihood function (see, for example, Aldrich, 1997 and Myung, 2003). The optimal properties in estimation are: consistency (true parameter value that generated the data recovered asymptotically, that is, for data of sufficiently large samples); sufficiency (complete information about the parameter of interest contained in its MLE estimator); efficiency (lowest-possible

variance of parameter estimates achieved asymptotically); and parameterization invariance (same MLE solution obtained independent of the parameterization used). Under certain regularity conditions, the MLE exhibits several characteristics that can be interpreted to mean that it is "asymptotically optimal." Lehmann and Casella (1998) provided the following results in Theorem 2.7.1 of the MLE under some regularity conditions. These conditions are:

(A0) The distributions P_θ of the observations are distinct (otherwise, θ cannot be estimated consistently).

(A1) The distributions P_θ have common support.

(A2) The random variables are $X_i = (X_{i1}, \ldots, X_{ip})$, $i = 1, 2, \ldots, n$ where the X_i are independent and identically distributed (iid) with probability density $f(x_i \mid \theta)$ with respect to probability measure μ.

(A3) There exists an open subset ω of Ω containing the true parameter point θ^0 such that for almost all x the density $f(x \mid \theta)$ admits all third derivatives $\frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} f(x \mid \theta)$ for all $\theta \in \omega$.

(A4) The first and second derivatives of log f satisfy the equations

$$E_\theta\left[\frac{\partial}{\partial\theta_j}\log f(X \mid \theta)\right] = 0 \text{ for } j = 1, \ldots, p, \text{ and}$$

$$I_{jk} = E_\theta\left[\frac{\partial}{\partial\theta_j}\log f(X \mid \theta)\cdot\frac{\partial}{\partial\theta_k}\log f(X \mid \theta)\right] = E_\theta\left[-\frac{\partial^2}{\partial\theta_j \partial\theta_k}\log f(X \mid \theta)\right].$$

(A5) Since the $p \times p$ matrix $I(\theta)$ is a covariance matrix, it is positive semidefinite. We will assume that $I_{jk}(\theta)$ are finite and that the matrix $I(\theta)$ is positive definite for all θ in ω, and the p statistics

$$\frac{\partial}{\partial\theta_1}\log f(x\mid\theta),\ldots,\frac{\partial}{\partial\theta_p}\log f(x\mid\theta)$$

are affinely independent with probability 1.

(A6) Finally, we will assume that there exists function M_{jkl} such that

$$\left|\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l}\log f(x\mid\theta)\right|\le M_{jkl}(x)\ \text{ for all }\ \theta\in\omega$$

where $m_{jkl}=E_{\theta^0}[M_{jkl}(X)]<\infty$ for all $j,k,l.$

Theorem 2.7.1. Let $X_1,\ldots,X_n$ be iid each with a density $f(x\mid\theta)$ (with respect to μ) which satisfies (A0)-(A6) above. Then, with probability tending to 1 as $n\to\infty$, there exist solutions $\hat\theta_n=\hat\theta_n(X_1,\ldots,X_n)$ of the likelihood equations

$$\frac{\partial}{\partial\theta_j}\left[f(x_1\mid\theta)\ldots f(x_n\mid\theta)\right]=0,\ j=1,\ldots,p,$$

or, equivalently,

$$\frac{\partial}{\partial\theta_j}\left[\log L(\theta)\right]=0,\ \ j=1,\ldots,p,$$

such that

(a) $\hat\theta_{jn}$ is consistent for estimating θ_j.

(b) $\sqrt{n}(\hat\theta_n-\theta)$ is asymptotically normal with mean (vector) zero and covariance matrix $[\,I(\theta)^{-1}\,]$, and

(c) $\hat\theta_{jn}$ is asymptotically efficient in the sense that

$$\sqrt{n}(\hat\theta_{jn}-\theta_j)\xrightarrow{L}N\left\{0,\left[I(\theta)\right]^{-1}_{jj}\right\}.$$

2.8 Asymptotic Properties of the ML Estimators in Logistic Regression Model

In this section, we verify all the regularity conditions under the logistic regression model discussed in section 2.7 and then we apply Theorem 2.7.1 to show the asymptotic properties of ML estimators for the logistic regression model.

Assumption (A0): Let $\theta_1 = (\beta_0^{(1)}, \beta_1^{(1)}, ..., \beta_p^{(1)})$ and $\theta_2 = (\beta_0^{(2)}, \beta_1^{(2)}, ..., \beta_p^{(2)})$. We define the models as

$$P_{\theta_1}(Y=1 \mid X=x) = \frac{e^{\beta_0^{(1)}x_0+\beta_1^{(1)}x_1+\cdots+\beta_p^{(1)}x_p}}{1+e^{\beta_0^{(1)}x_0+\beta_1^{(1)}x_1+\cdots+\beta_p^{(1)}x_p}} = \frac{e^{\beta^{(1)^T}x}}{1+e^{\beta^{(1)^T}x}} \tag{2.8.1}$$

$$P_{\theta_2}(Y=1 \mid X=x) = \frac{e^{\beta_0^{(2)}x_0+\beta_1^{(2)}x_1+\cdots+\beta_p^{(2)}x_p}}{1+e^{\beta_0^{(2)}x_0+\beta_1^{(2)}x_1+\cdots+\beta_p^{(2)}x_p}} = \frac{e^{\beta^{(2)^T}x}}{1+e^{\beta^{(2)^T}x}} \tag{2.8.2}$$

where $\beta^{(1)^T} = (\beta_0^{(1)}, \beta_1^{(1)}, ..., \beta_p^{(1)})$, $\beta^{(2)^T} = (\beta_0^{(2)}, \beta_1^{(2)}, ..., \beta_p^{(2)})$ and $x^T = (x_0, x_1, ..., x_p)$

If $\beta^{(1)} = \beta^{(2)}$, then the equations (2.8.1) and (2.8.2) are the same. On the contrary, we are going to show that if the equations (2.8.1) and (2.8.2) are equal, then $\beta^{(1)} = \beta^{(2)}$.

We have, $\frac{e^{\beta^{(1)^T}x}}{1+e^{\beta^{(1)^T}x}} = \frac{e^{\beta^{(2)^T}x}}{1+e^{\beta^{(2)^T}x}}$

That is, $e^{\beta^{(1)^T}x}\left(1+e^{\beta^{(2)^T}x}\right) = e^{\beta^{(2)^T}x}\left(1+e^{\beta^{(1)^T}x}\right)$

That is, $e^{\beta^{(1)^T}x} + e^{\beta^{(1)^T}x} \cdot e^{\beta^{(2)^T}x} = e^{\beta^{(2)^T}x} + e^{\beta^{(2)^T}x} \cdot e^{\beta^{(1)^T}x}$

That is, $e^{\beta^{(1)^T}x} = e^{\beta^{(2)^T}x}$

That is, $\beta^{(1)^T}x = \beta^{(2)^T}x$

That is, $(\beta_0^{(1)}, \beta_1^{(1)}, ..., \beta_p^{(1)})\,x = (\beta_0^{(2)}, \beta_1^{(2)}, ..., \beta_p^{(2)})\,x$

That is, $\left[(\beta_0^{(1)}, \beta_1^{(1)}, ..., \beta_p^{(1)}) - (\beta_0^{(2)}, \beta_1^{(2)}, ..., \beta_p^{(2)})\right] x = 0$

That is, $\left[(\beta_0^{(1)} - \beta_0^{(2)}), (\beta_1^{(1)} - \beta_1^{(2)}), ..., (\beta_p^{(1)} - \beta_p^{(2)})\right] x = 0$

That is, $(\beta_0^{(1)} - \beta_0^{(2)})x_0 + (\beta_1^{(1)} - \beta_1^{(2)})x_1 + , ..., + (\beta_p^{(1)} - \beta_p^{(2)})x_p = 0$

That is, $a_0 x_0 + a_1 x_1 + , ..., + a_p x_p = 0$, where $a_i = \beta_i^{(1)} - \beta_i^{(2)}, \quad i = 0, 1, ..., p$

Since *x*'s are independent, so $a_0 = a_1 = ... = a_p = 0$

This implies that, $\beta^{(1)} = \beta^{(2)}$

This indicates that the distributions are unique, therefore, if $\theta_1 \neq \theta_2$, then the distributions P_θ of the observations are distinct.

Assumption (A1): The variables in the model are $x_1, x_2, ..., x_p$, let $x = (x_1, x_2, ..., x_p)$ where $x \in \Re^p$ and the parameter β takes values $-\infty < \beta_j < \infty$, $j = 1, 2, \ldots, p$. This is true for each model stated in the assumption (A0). Therefore, the distributions P_θ have common support.

Assumption (A2): In this case, we consider the observations of the form $x_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, where the x_i are iid with probability density $P(x|.)$.

Assumption (A3): When *Y*=1, we define $f(x \mid \beta) := \dfrac{e^{\beta_0 + \beta^T x_i}}{1 + e^{\beta_0 + \beta^T x_i}}$ have the likelihood for the logistic regression model is proportional to

$$L = \prod_{i=1}^{n} \frac{e^{\beta_0 + \beta^T x_i}}{1 + e^{\beta_0 + \beta^T x_i}}$$

Taking log on both sides and we get,

$$\log L = \sum_{i=1}^{n} \left[\beta_0 + \beta^T x_i - \log(1 + e^{\beta_0 + \beta^T x_i})\right]$$

Now, taking derivative with respect to β_j, we have

$$\frac{\delta \log L}{\delta \beta_j} = \sum_{i=1}^{n}\left[x_{ij} - \frac{x_{ij}\, e^{\beta_0+\beta^T x_i}}{1+e^{\beta_0+\beta^T x_i}} \right] = \sum_{i=1}^{n}\left[\frac{x_{ij}}{1+e^{\beta_0+\beta^T x_i}} \right]$$

The above derivative comes to the form $\frac{x_{ij}}{1+e^{\beta_0+\beta^T x_i}}$ and if we take the derivative of *k*th order,

then the derivative continues to the form $\frac{x_{ij}}{(1+e^{\beta_0+\beta^T x_i})^k}$, which can be proved by the

mathematical induction. Therefore, not only does the derivative of $f(x \mid \beta)$ third order exist, but the derivatives of all orders also exist.

Assumption (A4): The condition (A4) can be proved, in general, for the density $f(x \mid \beta)$ under the condition that the differentiation under the integral sign is allowed. The only thing we need to check for the logistic model is whether it permits the differentiation under the integral sign. To show that part we consider the following theorem, available in real analysis or probability books (see, for example, Durrett, 2004), which allows the differentiation under the integral sign.

Theorem 2.8.1. Suppose we are given the following:

- An open interval $I \subset \Re$.
- A measurable subset $X \subset \Re$.
- A function $H : I \times X \to \Re$
- A function $g : X \to [0, \infty]$

Assume the following:

- $\left|\frac{\partial H}{\partial t}(t, x)\right| \le g(x)$ for every $t \in I$ and $x \in X$.
- g is integrable.

- $t \to H(t,x)$ is a differentiable function of $t \in I$ for every $x \in X$.
- $x \to H(t,x)$ is an integrable function of $x \in X$ for every $t \in I$.

Then the following hold:

- $x \to \frac{\partial H}{\partial t}(t,x)$ is an integrable function of $x \in X$ for every $t \in I$.
- $t \to \int_X H(t,x)dx$ is a differentiable function of $t \in I$.
- $\frac{d}{dt}\int_X H(t,x)dx = \int_X \frac{\partial}{\partial t} H(t,x)dx$ for every $t \in I$.

To verify the above assumptions of Theorem 2.8.1 for logistic regression model, we consider the following function when y =1.

$$H(\beta,x) = \frac{e^{\beta_0+\beta^T x}}{1+e^{\beta_0+\beta^T x}}$$

$$\frac{\partial}{\partial \beta} H(\beta,x) = \frac{x e^{\beta_0+\beta^T x}}{(1+e^{\beta_0+\beta^T x})^2}$$

$$\therefore \quad \left| \frac{\partial}{\partial \beta} H(\beta,x) \right| = \left| \frac{x_i e^{\beta_0+\beta^T x}}{(1+e^{\beta_0+\beta^T x})^2} \right| \le |x| \left| \frac{e^{\beta_0+\beta^T x}}{(1+e^{\beta_0+\beta^T x})^2} \right| = g(x) \text{ as } \left| \frac{e^{\beta_0+\beta^T x}}{(1+e^{\beta_0+\beta^T x})^2} \right| < 1$$

Similarly, this can be shown for y =0.

Since $H(\beta,x)$ is a differentiable function of $x \in X$ for every $\beta \in \Re^p$ and integrable for $x \in X$ for every $\beta \in \Re^p$. Thus, the results of Theorem 2.8.1 hold.

Assumption (A5): We take the derivative of $\log f(x \mid \beta)$ with respect to $\beta_1, \beta_2, \cdots, \beta_p$, we have

$$\frac{\partial \log f}{\partial \beta_j} = x_j - \frac{x_j e^{\beta_0+\beta^T x}}{1+e^{\beta_0+\beta^T x}}, j = 1,2,\ldots,p$$

Now, we write the vectors in the form so that they are linearly dependent in the following way,

$$x_p - \frac{x_p e^{\beta_0 + \beta^T x}}{1 + e^{\beta_0 + \beta^T x}} = \sum_{j=1}^{p-1} \alpha_j x_j - \sum_{j=1}^{p-1} \alpha_j \frac{x_j e^{\beta_0 + \beta^T x}}{1 + e^{\beta_0 + \beta^T x}}$$

That is, $$\left(x_p - \sum_{j=1}^{p-1} \alpha_j x_j \right) - \frac{e^{\beta_0 + \beta^T x}}{1 + e^{\beta_0 + \beta^T x}} \left(x_p - \sum_{j=1}^{p-1} \alpha_j x_j \right) = 0$$

That is, $$\left(x_p - \sum_{j=1}^{p-1} \alpha_j x_j \right) \left(1 - \frac{e^{\beta_0 + \beta^T x}}{1 + e^{\beta_0 + \beta^T x}} \right) = 0 \quad \forall\ \beta$$

Since $\left(1 - \frac{e^{\beta_0 + \beta^T x}}{1 + e^{\beta_0 + \beta^T x}} \right) \neq 0$, So $\left(x_p - \sum_{j=1}^{p-1} \alpha_j x_j \right) = 0$

Thus, $x_p = \sum_{j=1}^{p-1} \alpha_j x_j$

We have,

$$P_\beta \left[x_p = \sum_{j=1}^{p-1} \alpha_j x_j \right] = 0,\ \forall\ \beta$$

since the joint distribution of $x_1, x_2, ..., x_p$ is continuous on $\Re^p$.

Thus, $P_\beta \left[x_p \neq \sum_{j=1}^{p-1} \alpha_j x_j \right] = 1,\ \forall\ \beta$

This implies that the statistics are affinely independent.

Assumption (A6): We have, $\frac{\delta \log L}{\delta \beta_j} = \sum_{i=1}^{n} \left[x_{ij} - \frac{x_{ij}\ e^{\beta_0 + \beta^T x}}{1 + e^{\beta_0 + \beta^T x}} \right] = \sum_{i=1}^{n} \left[\frac{x_{ij}}{1 + e^{\beta_0 + \beta^T x}} \right]$

and $$\frac{\delta^2 \log L}{\delta \beta_j \delta \beta_k} = -\sum_{i=1}^{n} x_{ij} x_{ik} \left[\frac{e^{\beta_0 + \beta^T x}}{\left(1 + e^{\beta_0 + \beta^T x} \right)^2} \right]$$

So, $$\frac{\delta^3 \log L}{\delta \beta_j\ \delta \beta_k\ \delta \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{ik} x_{il} \left(\frac{e^{\beta_0 + \beta^T x} (1 - e^{\beta_0 + \beta^T x})}{(1 + e^{\beta_0 + \beta^T x})^3} \right)$$

$$\therefore \quad \left| \frac{\delta^3 \log L}{\delta\beta_j \, \delta\beta_k \, \delta\beta_l} \right| = \left| -\sum_{i=1}^{n} x_{ij} x_{ik} x_{il} \left(\frac{e^{\beta_0 + \beta^T x}(1 - e^{\beta_0 + \beta^T x})}{(1 + e^{\beta_0 + \beta^T x})^3} \right) \right|$$

$$\leq \sum_{i=1}^{n} \left| x_{ij} x_{ik} x_{il} \left(\frac{e^{\beta_0 + \beta^T x}(1 - e^{\beta_0 + \beta^T x})}{(1 + e^{\beta_0 + \beta^T x})^3} \right) \right|$$

$$\leq \sum_{i=1}^{n} \left| x_{ij} x_{ik} x_{il} \right| \left| \frac{e^{\beta_0 + \beta^T x}(1 - e^{\beta_0 + \beta^T x})}{(1 + e^{\beta_0 + \beta^T x})^3} \right|$$

$$\leq \sum_{i=1}^{n} \left| x_{ij} x_{ik} x_{il} \right| \quad \text{since} \left| \frac{e^{\beta_0 + \beta^T x}(1 - e^{\beta_0 + \beta^T x})}{(1 + e^{\beta_0 + \beta^T x})^3} \right| < 1$$

$$\leq \sum_{i=1}^{n} \left| x_{ij} \right| \left| x_{ik} \right| \left| x_{il} \right|$$

$$\leq M_{jkl}(x)$$

Where $m_{jkl} = E\left[M_{jkl}(X)\right]$

$$= E\left[\sum_{i=1}^{n} \left| X_{ij} \right| \left| X_{ik} \right| \left| X_{il} \right| \right]$$

$$= \sum_{i=1}^{n} E\left| X_{ij} X_{ik} X_{il} \right| \text{ , which is finite.}$$

Since the logistic regression model satisfies all the regularity conditions (A0)-(A6), therefore, $\hat{\beta}$ satisfies (a) – (c) of Theorem 2.7.1.

2.9 A Simulation Study

2.9.1 Consistency of the ML Estimators

We now assess, via standard Monte Carlo simulation, the finite sample performance of consistency of the maximum likelihood estimators of the logistic regression model. In our

simulation study, we consider four explanatory variables x_1, x_2, x_3, and x_4, which are fixed and the binary response variable *y,* which is treated as a random variable in the logistic model. For fixed values of the intercept parameter β_0 and four other parameters β_1, β_2, β_3, and β_4, our aim is to compare the performance of the values of parameters and their standard errors when sample size increases. For fixed values of $\beta_0 = 0.7$, $\beta_1 = 1.0$, $\beta_2 = 1.3$, $\beta_3 = 0.25$, and $\beta_4 = 0.05$, the logistic regression model becomes

$$\pi(x) = \frac{e^{0.7+1.0\,x_i+1.3\,x_2+0.25\,x_3+0.05x_4}}{1+e^{0.7+1.0\,x_i+1.3\,x_2+0.25\,x_3+0.05x_4}}$$

In the simulation, we consider sample sizes of $n = 50$, 100, 150, and 200 and generate 1,000 independent sets of random samples for each different sample size. For each set of random sample with a particular sample size, we estimate β_0, β_1, β_2, β_3, and β_4 and their standard errors based on the logistic regression model. The final estimates and standard errors of β_0, β_1, β_2, β_3, and β_4 are the average of 1,000 estimates of β_0, β_1, β_2, β_3, and β_4 for that particular sample size. The following table gives the results of the simulation study for different sample sizes.

Table 2.9.1: Estimated parameter values and their standard errors using the logistic regression model for different sample sizes of 50, 100, 150, and 200.

Parameters	$n = 50$		$n = 100$		$n = 150$		$n = 200$	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_0	1.236	0.132	0.864	0.043	0.736	0.017	0.747	0.015
β_1	2.644	0.184	1.263	0.058	1.084	0.026	1.082	0.025
β_2	4.143	0.225	1.759	0.081	1.461	0.041	1.382	0.025
β_3	1.030	0.159	0.320	0.041	0.252	0.017	0.263	0.015
β_4	0.380	0.147	0.016	0.044	0.060	0.017	0.045	0.015

SE=Simulation standard error

As seen in the above table, for sample size $n = 50$, the estimated values of parameters are different from the true values ($\beta_2 = 0.7$, β_1=1.0, β_2=1.3, β_3=0.25, and β_4=0.05), and also the standard errors become larger. However, when the sample size increases from $n = 50$ to $n = 200$, the estimated values of the parameters β_0, β_1, β_2, β_3, and β_4 are very close to the true values, and the standard errors of the estimates are noticeably smaller. This indicates that this simulation study performs well in showing the consistency of the maximum likelihood estimators for parameters of the logistic regression model.

2.9.2 Normality of ML the Estimators

In this section, we illustrate the large sample behavior of the estimated parameters $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)^T$. Specifically, we want to show

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{L} N\left(0,\ [I(\beta)]^{-1}\right) \qquad (2.9.1)$$

Where,

$$I(\beta) = -E_{\beta}\begin{bmatrix} \frac{\partial^2 \log L}{\partial \beta_0^2} & \frac{\partial^2 \log L}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 \log L}{\partial \beta_0 \partial \beta_2} & \frac{\partial^2 \log L}{\partial \beta_0 \partial \beta_3} & \frac{\partial^2 \log L}{\partial \beta_0 \partial \beta_4} \\ \frac{\partial^2 \log L}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \log L}{\partial \beta_1^2} & \frac{\partial^2 \log L}{\partial \beta_1 \partial \beta_2} & \frac{\partial^2 \log L}{\partial \beta_1 \partial \beta_3} & \frac{\partial^2 \log L}{\partial \beta_1 \partial \beta_4} \\ \frac{\partial^2 \log L}{\partial \beta_2 \partial \beta_0} & \frac{\partial^2 \log L}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 \log L}{\partial \beta_2^2} & \frac{\partial^2 \log L}{\partial \beta_2 \partial \beta_3} & \frac{\partial^2 \log L}{\partial \beta_2 \partial \beta_4} \\ \frac{\partial^2 \log L}{\partial \beta_3 \partial \beta_0} & \frac{\partial^2 \log L}{\partial \beta_3 \partial \beta_1} & \frac{\partial^2 \log L}{\partial \beta_3 \partial \beta_2} & \frac{\partial^2 \log L}{\partial \beta_3^2} & \frac{\partial^2 \log L}{\partial \beta_3 \partial \beta_4} \\ \frac{\partial^2 \log L}{\partial \beta_4 \partial \beta_0} & \frac{\partial^2 \log L}{\partial \beta_4 \partial \beta_1} & \frac{\partial^2 \log L}{\partial \beta_4 \partial \beta_2} & \frac{\partial^2 \log L}{\partial \beta_4 \partial \beta_3} & \frac{\partial^2 \log L}{\partial \beta_4^2} \end{bmatrix}$$

For different sample sizes of n = 100, 250, 500, we calculate the equation (2.9.1) and repeat it 1,000 times. The results are presented below (Figures 2.9.1 – 2.9.4) through the quantile-normal graphs of $\hat{\beta}$. A quantile-normal graph plots the quantiles of the data set against the theoretical quantiles of the standard normal distribution. If the data set appears to be a sample from a normal population, then the points will fall roughly along a line. The computation results indicate that the distribution of parameters approximates normal distribution as sample size, n, increases.

Figure 2.9.1: Monte Carlo simulation of finite sample behavior for normality of the parameter $\hat{\beta}_1$
(Simulation size = 1,000)

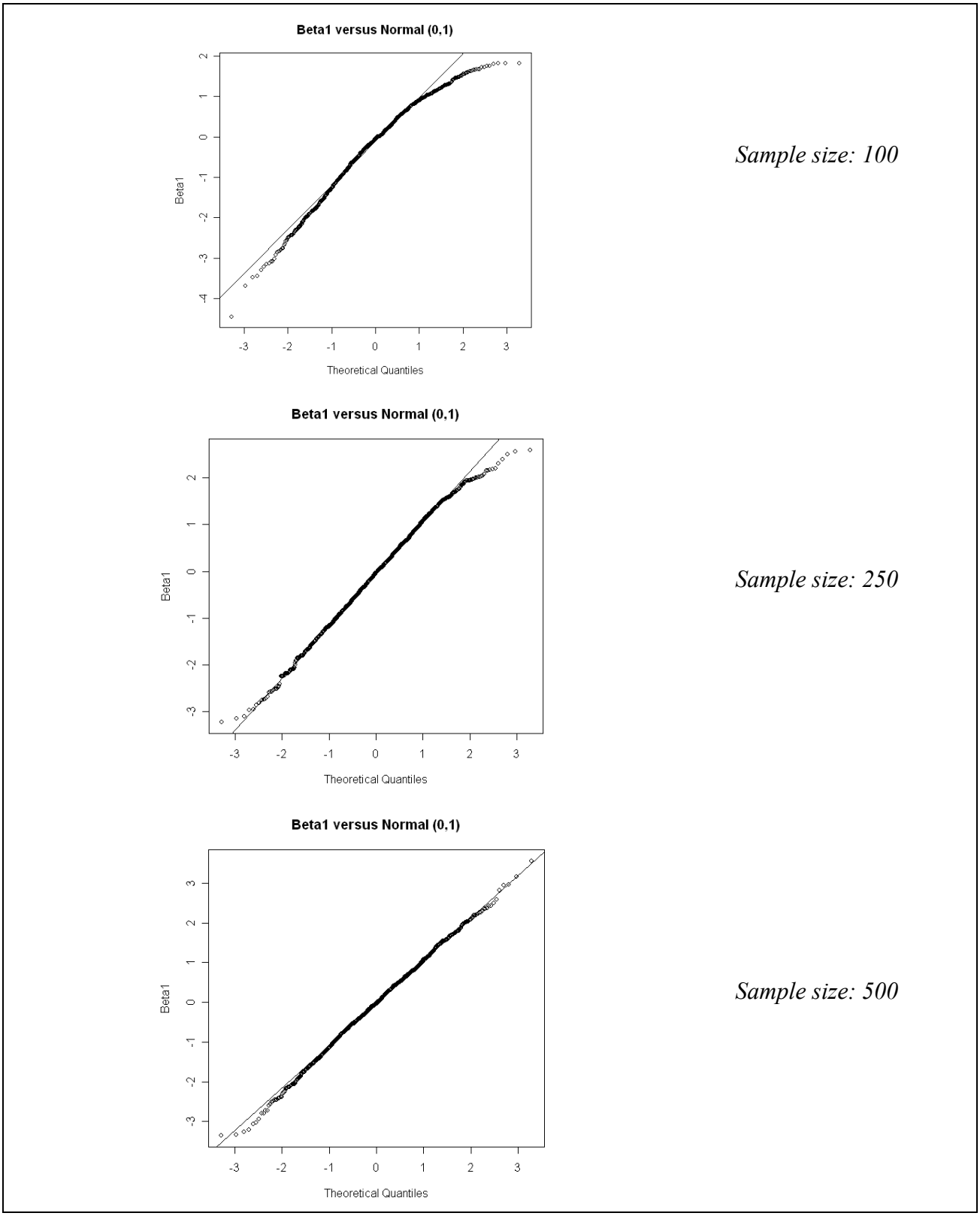
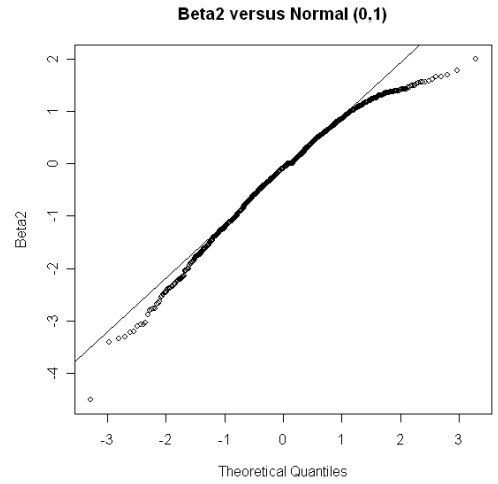
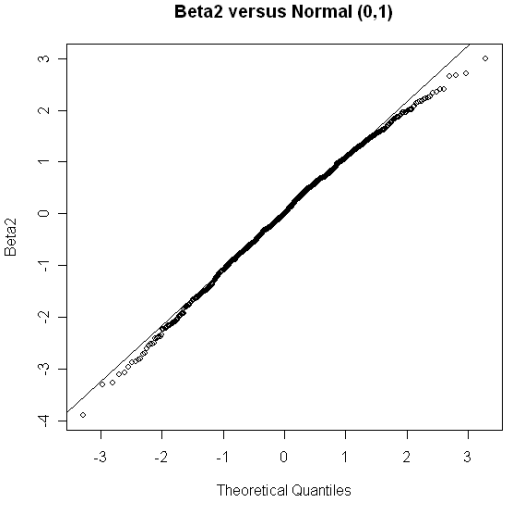


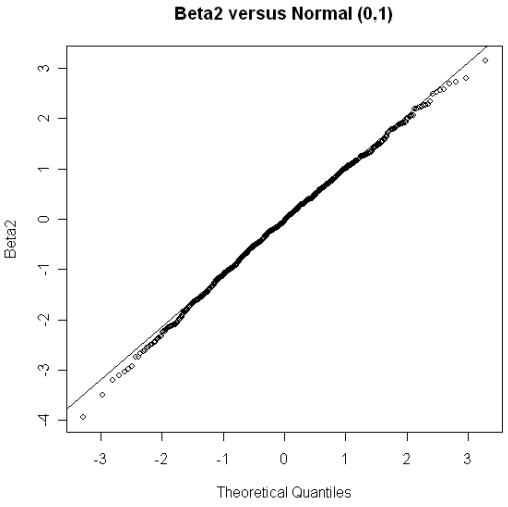
Figure 2.9.2**:** Monte Carlo simulation of finite sample behavior for normality of the parameters $\hat{\beta}_2$ *(Simulation size = 1,000)*



Sample size: 100

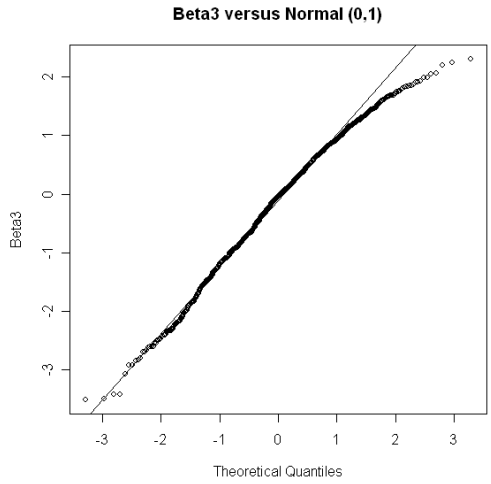


Sample size: 250

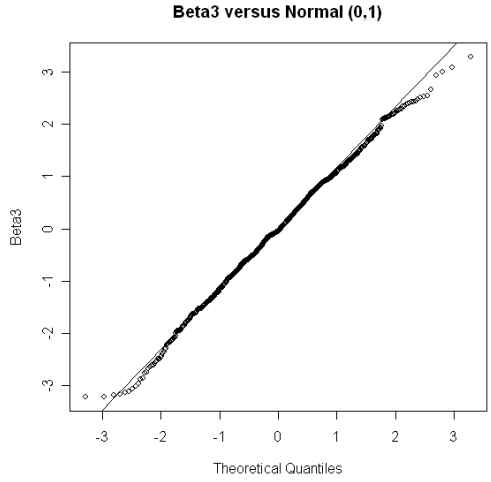


Sample size: 500

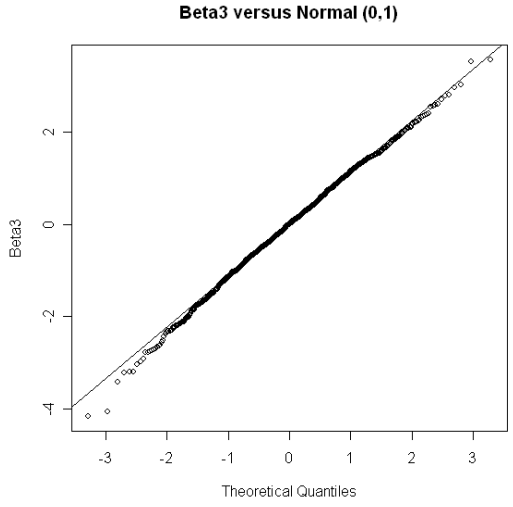
Figure 2.9.3: Monte Carlo simulation of finite sample behavior for normality of the parameters $\hat{\beta}_3$ *(Simulation size = 1,000)*



Sample size: 100

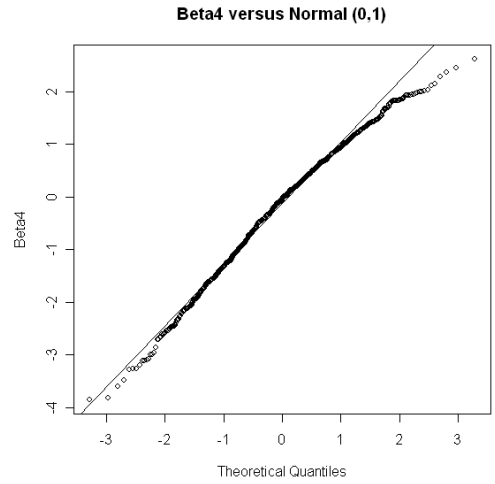


Sample size: 250

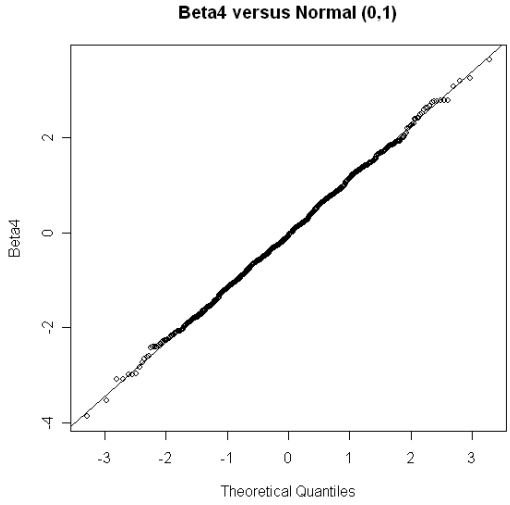


Sample size: 500

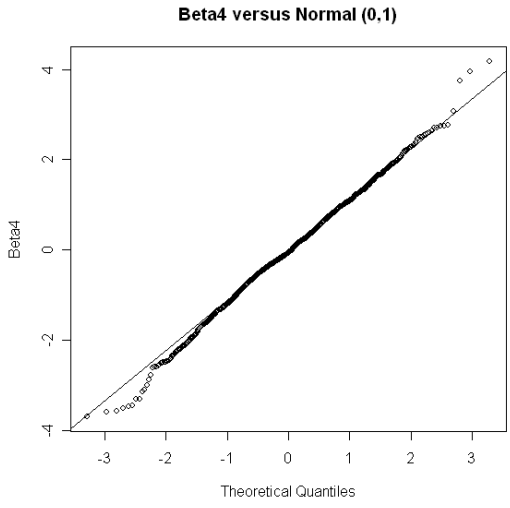
Figure 2.9.4: Monte Carlo simulation of finite sample behavior for normality of the parameters $\hat{\beta}_4$ *(Simulation size = 1,000)*



Sample size: 100



Sample size: 250



Sample size: 500

2.10 Application of the Logistic Regression Model in a Real Data Set

Rashid and Ahmed (2002) studied the correlates of timing of induced abortion in a rural area of Bangladesh. Abortion is not permitted in Bangladesh unless it is done to save the life of a woman. Khan et al. (1986) showed that the highest number of induced abortions occurred at three or more months of gestation, causing unavoidable morbidity and mortality. Other studies showed that the contribution of unsafe induced abortion is related to maternal morbidity and mortality (Sai and Nassim, 1989). Here, we concentrate on the application of the logistic regression model in the area of public health to identify important risk factors associated with the timing of induced abortion. The data for this study relates to the period 1991-1998 during which 2,247 abortion cases were obtained. The data were extracted from a longitudinal Health and Demographic Surveillance Unit (HDSU), which has been maintained by the ICDDR,B since 1966 (see, www.icddrb.org for details). The HDSU has collected information for both ICDDR,B area and Comparison area on pregnancy outcomes (live births, stillbirths, spontaneous abortions, and induced abortions) and other demographic events such as deaths, migrations, and marriages. Community health workers who make routine visits to every household monthly register all of these events and are strongly supervised for accurate completion of the vital events.

In this study, the outcome variable is

$$y = \begin{cases} 1, & \text{if woman sought abortion three or later months of gestation.} \\ 0, & \text{otherwise.} \end{cases}$$

and a reduced set of categorical explanatory variables is the following

x_1 : maternal age,

x_2 : number of living children,

x_3 : women's education,

x_4 : study area, and

x_5 : women's occupation.

To investigate the association between the explanatory variables and the binary response variable, we express the logistic regression model as the following

$$\ln\left[\frac{P(y=1\mid x)}{1-P(y=1\mid x)}\right]=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5$$

The parameters of the model can be estimated using standard statistical software, and thus, the results of the fitted model can be organized in the following tabulated form.

Table 2.10.1: Logistic regression of three months or later months of gestation of abortions by selected characteristics.

Characteristics	Odds ratio	Confidence interval
Number of living children		
None (ref.)	1.00	–
1-2	0.60**	0.35–0.75
3+	0.50**	0.31–0.73
Women's education		
None (ref.)	1.00	–
Some	0.65**	0.56–0.82
Study area		
ICDDR,B area (ref.)	1.00	–
Comparison area	2.28**	1.94–2.88

**p<0.01 and **p<0.001; ref. indicates reference category*

As can be seen, the odds of women who had sought abortion three months or later having three or more children were expected to be 0.50 times the odds of women having no child. This indicates that the practice of abortion in the later gestational period was higher among women who had no children. The odds of educated women who had sought an abortion three months or later were 0.65 times the odds of women who had no education. This means that educated women wanted an abortion earlier in the pregnancy compared to women who had no education.

This model also shows the odds of women living in the Comparison area who had sought abortion three months or later were estimated to be 2.30 times the odds of women living in the ICDDR,B area. This implies that women residing in the Comparison area sought abortion in the third or later months of gestation. Our study found that the differences in the timing of induced abortion depend on a variety of factors. Factors such as residing in the Comparison area, having no living children, and having no education significantly increased the risk of having an induced abortion in or after the third month of pregnancy.

CHAPTER 3

THE HYBRID LOGISTIC REGRESSION MODELS FOR MORE THAN ONE RARE RISK FACTOR

3.1 Definition: Zero Cells Count

When analyzing sample tables of counts, we encounter two types of empty cells: sampling zeroes and structural zeroes. Sampling zero cells occur in situations where one or more cases exist in the population of interest, but such zero cell counts arise because of sampling variation, especially the use of small sample sizes for a contingency table composed of a large number of cells. Such zero cells will tend to disappear if sample size is sufficiently large (Agresti, 1996). The classic example used by Fienberg (1980) to illustrate sampling zeroes is the observed zero cell count for Iowa Jewish farmers; such individuals exist, but small simple random samples of Iowa farmers will often not include these people because of their small population size. On the other hand, structural zero cells occur in situations where a cell is empty due to the impossibility of observing positive cell counts for specific combinations of various categories. An example by Agresti (1996) to illustrate structural zeroes, "suppose that professors employed in a given department at the university of Rochester for at least five years were cross-classified on their current rank (assistant professor, associate professor, professor) and their rank five years ago. Professors cannot be demoted in rank, so three of the nine cells in the table contain structural zeroes. One of these is the cell corresponding to the rank of professor five years ago and assistant professor now; it cannot contain any observations."

The sampling zeroes are the part of the observed data set that are much more common than the structural zeroes and have the contributions to the likelihood function and the model-fitting process. Our discussions based on sampling zeroes, which can affect the ML estimation

of the parameters in the logistic regression model, often by reporting infinite estimates for the parameters. A value of ∞ (or $-\infty$) for a parameter estimate means that the likelihood function keeps increasing as the parameter moves toward ∞ (or $-\infty$). Such results imply that ML fitted values equal to 0 in some cells, and some odd ratio estimates have values of ∞ or 0. The consequence of the sampling zeroes has a severe bias in estimators of odds ratios and poor chi-squared approximations for goodness-of-fit statistics. However, different ideas appeared in the literature are discussed below to smooth the data before fitting the model.

3.2 Methods for Smoothing the Data

3.2.1 "Add-a-Constant" Approach

Adding a small constant, generally 0.5, to every cell of the table has been a common recommendation in some standard references; for example, Haldane (1956) suggested a correction term 0.5 to add to all four cells prior to analysis of the 2×2 contingency table. If a, b, c, and d are the cells count of the 2×2 contingency table, then the odds ratio estimate is given by

$$OR = \left(a + \frac{1}{2}\right)\left(d + \frac{1}{2}\right) \Big/ \left(b + \frac{1}{2}\right)\left(c + \frac{1}{2}\right),$$

(see, for example, Walter and Cook, 1991). For a multi-way table, Goodman (1970,1971) recommended this procedure for the saturated models only. An example of the beneficial effect of this for a saturated model is bias reduction for estimating as odds ratio in a 2×2 table (Gart, 1966; Gart and Zweiful, 1967). A different approach proposed by Clogg et al., (1991) is to preserve the marginal distribution of the dependent variable when prior observations are divided among cells of the contingency table. Agresti (2002) mentioned that adding 0.5 to the unsaturated model smoothes the data too much that causes an influence on estimated effects and

test statistics. However, this practice helps to find the asymptotic variances of the parameters for which the confidence intervals are calculated as well as point estimate for the unknown parameters. Hosmer and Lemeshow (2000) and Agresti (1996, 2002) suggested certain ways to deal with zero cells:

Hosmer and Lemeshow (2000) recommend three strategies as *ad hoc* solutions for eliminating the problems caused by zero cells for sampling zeroes:

(i) collapse the categories of a nominal variable by combining a zero cell with a non-zero cell, thus eliminating the zero cell by reducing the number of variable categories (levels) by pooling two or more cell counts;

(ii) simply eliminate the zero cell by discarding the variable category in which it appears;

(iii) treat the variable as intervally measured, if the variable with a zero cell in one of its categories is an ordinal measure.

In addition, Agresti (1996, 2002) recommends the following approach:

> add a very small constant (such as 10^{-8}) to cell counts and perform a sensitivity analysis by adding constants of varying sizes to determine the effect on the parameter estimates and goodness-of-fit statistics. The total amount added should be very small in comparison to the total sample size.

3.2.2 "Pseudo-Bayes" Approach

Bayesian methods, an alternative approach to ML estimation, provide a way of smoothing the data in a less *ad hoc* manner than adding arbitrary constant to cells (Agresti,

2002). Bishop et al. (1975) proposed pseudo-Bayes estimators which provide an all purpose method for removing the zeros in an observed frequency distribution or contingency table, so that other analyses, can be made that were previously hampered by the presence of zero cell counts, can be made. In Bayes and pseudo-Bayes estimation approaches (see, Bishop et al., 1975; Dillon et al., 1981; and Agresti, 2002), the parameters of the multiway table (that is, the cell probabilities) themselves are assumed to have a probability distribution that can be characterized by a smaller set of "hyperparameters". In the process of estimating the hyperparameters, estimators for the original set of parameters are obtained which often have more superior properties (that is, smaller risk) than the estimators not based on the hyperparametric structure.

Let $X = (X_1, X_2, \ldots, X_t)$ have the multinomial distribution with parameters $N = \sum_{i=1}^{t} X_i$ and $p = (p_1, p_2, \ldots, p_t)$. We observe a vector of values $x = (x_1, x_2, \ldots, x_t)$, where x_i is the observed count in the *i*th category and $\sum_{i=1}^{t} x_i = N$. The vector *p* takes values in the parameter space ζ_t, where

$\zeta_t = \left\{ p = (p_1, p_2, \ldots, p_t) : p_i \geq 0 \text{ and } \sum_{i=1}^{t} p_i = 1 \right\}$ and we denote the "center" of ζ_t by $c = \left(t^{-1}, \ldots, t^{-1} \right)$.

The kernel of the likelihood function for this multinomial distribution is

$$l(p \mid x) = \prod_{i=1}^{t} p_i^{x_i}$$

Assuming the natural conjugate family of prior distributions for this likelihood is the Dirichlet, whose densities have the form

$$f(p \mid \beta) = \frac{\Gamma\left(\sum_{i=1}^{t} \beta_i\right)}{\prod_{i=1}^{t} \Gamma(\beta_i)} \prod_{i=1}^{t} p_i^{\beta_i - 1}$$

where $\beta_i > 0$ for all *i* and $\Gamma(y)$ is the gamma function given by $\Gamma(y) = \int_0^{\infty} e^{-z} z^{y-1} dz$ and the mean and variance of the Dirichlet distribution are

$E(p_i \mid \beta) = \frac{\beta_i}{\sum_{i=1}^{t} \beta_i}$ and $Var(p_i \mid \beta) = \frac{\beta_i \left(\sum_{i=1}^{t} \beta_i - \beta_i \right)}{\left(\sum_{i=1}^{t} \beta_i \right)^2 \left(\sum_{i=1}^{t} \beta_i + 1 \right)}$ respectively.

The posterior distribution can be obtained from the likelihood and the prior, that is,

Posterior $\propto$ Prior $\times$ Likelihood

That is, $\pi(p \mid \beta) \propto f(p \mid \beta) \times l(p \mid x)$

That is, $\pi(p \mid \beta) \propto \frac{\Gamma\left(\sum_{i=1}^{t} \beta_i\right)}{\prod_{i=1}^{t} \Gamma(\beta_i)} \prod_{i=1}^{t} p_i^{(\beta_i + x_i) - 1}$

If we set $K = \sum_{i=1}^{t} \beta_i$ and $\lambda_i = \frac{\beta_i}{K}$, then the prior and posterior means of p_i are given by

Prior mean, $E(p_i \mid K, \lambda) = \lambda_i$

Posterior mean, $E(p_i \mid K, \lambda, x) = \frac{x_i + K\lambda_i}{N + K}$

Now, we can rewrite in vector notation the mean of the posterior distribution as

$$E(p \mid K, \lambda, x) = \frac{N}{N+K} \cdot \frac{x}{N} + \frac{K}{N+K} \cdot \lambda \qquad (3.2.1)$$

In this case, a Bayesian would specify *K* and λ on the basis of his prior information. According to the Bayesian interpretation, the right–hand side of (3.2.1) illustrates a well-known method for

"smoothing" multinomial data. The data, x/N, is shrunk towards a "smooth" probability vector, λ, by a convex weight, $N/(N+K)$. This is the same as adding $K\lambda_i$ "pseudo-counts" to x_i and normalizing by the new total $N+K$. Various choices of K in (3.2.1) have appeared in the literature (see, for example, Fienberg and Holland, 1972). A popular choice of parameters in this situation is $\lambda = c$ and $K = \frac{1}{2}t$, which corresponds to adding a fake count of $\frac{1}{2}$ to each cell. Here, K is called the *smoothing constant* and λ is regarded as a device for allocating a fraction of K to each cell of the multinomial. Adding $\frac{1}{2}$ is an example of a *data-independent* smoothing constant.

Next, we discuss another standard device called "pseudo-Bayes approach" which removes zero counts in contingency tables. Pseudo-Bayes estimates are obtained by using data-dependent values of both K and λ. Note that the smoothing constants K and λ are functions of x. Now, we denote the random variable version of the Bayes estimator given in (3.2.1) by

$$\hat{q} = \hat{q}(K, \lambda) = \frac{N}{N+K} \cdot \frac{X}{N} + \frac{K}{N+K} \cdot \lambda \tag{3.2.2}$$

Since K and λ are constants, the risk function of $\hat{q}$ is given by:

$$R(\hat{q}, p) = \left(\frac{N}{N+K}\right)^2 \left(1 - \| p \|^2\right) + \left(\frac{K}{N+K}\right)^2 N \| p - \lambda \|^2 \tag{3.2.3}$$

The risk function of $\hat{q}\left(\frac{1}{2}t, c\right)$ is obtained by substituting the appropriate values into equation (3.2.3). This yields,

$$R\left(\hat{q}\left(\frac{1}{2}t, c\right), p\right) = \left(\frac{2\delta}{2\delta+1}\right)^2 \left(1 - \| p \|^2\right) + \left(\frac{1}{2\delta+1}\right)^2 N \left\| p - \frac{1}{t} \right\|^2 \tag{3.2.4}$$

where $\delta = \frac{N}{t}$.

In order to use Bayesian estimator given in (3.2.1), we need to know the values of *K* and λ. Typically, the assessment of these prior parameters is a difficult task. The following is discussed a way of choosing *K* so that it depends on the data and the choice of λ.

If λ is regarded as fixed and particular value of *p*, then we can find the value of *K* that minimizes the risk $R\left(\hat{q}(K,\lambda),p\right)$ by differentiating (3.2.3) in *K* and solving the resulting equation. This yields,

$$K = K(p,\lambda) = \frac{1-\| p \|^2}{\| p-\lambda \|^2} \tag{3.2.5}$$

The optimal value of *K* depends on the unknown value of *p*. We may obtain an estimate of this unknown optimal value of *K* by replacing *p* by $\hat{p} = \frac{X}{N}$, yielding

$$\hat{K} = K(\hat{p},\lambda) = \frac{1-\| \hat{p} \|^2}{\| \hat{p}-\lambda \|^2} \tag{3.2.6}$$

or in terms of *x*, the observed value of the random variable X,

$$\hat{K} = \frac{N^2 - \sum_{i=1}^{t} x_i^2}{\sum_{i=1}^{t} x_i^2 - 2N\sum_{i=1}^{t} x_i\lambda_i + N^2\sum_{i=1}^{t}\lambda_i^2} \tag{3.2.7}$$

A pseudo-Bayes estimator of *p* is then

$$p^* = \hat{q}(\hat{K},\lambda) = \left(\frac{N}{N+\hat{K}}\right)\cdot\hat{p} + \frac{\hat{K}}{N+\hat{K}}\cdot\lambda \tag{3.2.8}$$

where $\hat{K}$ is given in equation (3.2.7).

Pseudo-Bayes estimates are found by using data-dependent values for both *K* and λ. However, the problem here is that there are no good rules for choosing the λ parameters; the problem exacerbated in large sparse tables. When $\lambda = (t^{-1},\ldots,t^{-1})$, (3.2.7) may be written as

$$\hat{K} = \frac{N^2 - \sum_{i=1}^{t} x_i^2}{\sum_{i=1}^{t} x_i^2 - \frac{N^2}{t}} \tag{3.2.9}$$

(see, for example, Dillon et al., 1981).

3.3 A Hybrid Logistic Regression Procedure

3.3.1 A Hybrid Logistic Regression Model for the Case-Control Study

Chen et al. (2003) proposed a hybrid logistic regression model for case-control studies to deal with the zero cells. In case-control studies, if there tends to be rare disease in the control group for the risk factors, then the estimation of the parameters of those risk factors is difficult. The following table provides an example of the rare risk factor for case-control study.

Table 3.3.1: Female adolescent suicides and controls by PAS

		Case	Control
Past attempt of suicide (PAS)	Yes	13	0
	No	8	40

Source: Chen et al. (2003)

In this situation, previous investigations (for example, Shaffer et al., 1996) do not include such risk factors and consider the other risk factors instead. Avoiding the former risk factors may overestimate the odds ratio of the remaining risk factors in the model (Brent et al., 1999). However, if all risk factors are included in the model, the model may not converge. As noted in Chapter 1, the hybrid logistic model (Chen et al., 2003) overcomes these limitations by adjusting troublesome risk factor first, and then models the remaining risk factors using the logistic regression. The specific form of the hybrid logistic model for case-control studies is expressed for one rare risk factor z and the other risk factors $x^T = (x_1, x_2, ..., x_p)$ as

$$P(x, Z = z \mid Y = y, s = 1) = \alpha^{zy}(1-\alpha)^{(1-z)y}\left(\frac{e^{\beta_0^* + \beta' x}}{1+e^{\beta_0^* + \beta' x}}\right)^{y}\left(\frac{1-z}{1+e^{\beta_0^* + \beta' x}}\right)^{1-y}\frac{P(x \mid s=1)}{P(Y=y \mid s=1)} \qquad (3.3.1)$$

where, α is the proportion of the covariate $z = 1$ in the case group

$P(z = 1 \mid y = 1, x, s = 1) = \alpha$, α depends on *x* and $P(z = 0 \mid y = 1, x, s = 1) = 1 - \alpha$

The parameters in the model α and $\beta^T = (\beta_1, \beta_2, ..., \beta_p)$ are estimated using the maximum likelihood estimation procedure.

3.3.2 A Hybrid Logistic Model: Bivariate Case

3.3.2.1 Consider the case when the rare risk factors z_1 and z_2 are independent

Suppose $Y_1, Y_2, ..., Y_n$ are a family of mutually independent {0,1} valued indicator random variables representing the cases (*Y*=1) or controls (*Y*=0) for *n* individuals in a case control study. The set of risk factors (z^T, x^T) where $z^T = (z_1, z_2)$ is the rare risk factors and $x^T = (x_1, x_2, ..., x_p)$ is the other risk factors. These risk factors for subject *i* take the values $(z_{i1}, z_{i2}, x_{i1,} x_{i2,}..., x_{ip})$. In this case we consider the rare risk factor z_i, *i* = 1,2 takes two possible values with 1 (occurrence of the event) and 0 (not occurrence). The structure of the covariate z_i, *i* = 1,2 has the pattern with the outcome variable, *Y* for a sample of size n_1 cases ($y = 1$) and n_0 controls ($y = 0$) as shown in the following table.

Table 3.3.2: Cross-classification between the variables $z_i, i = 1,2$ versus y

		$y=1$ (Case)	$y=0$ (Control)
z_1	1	$n_{11}^{(1)}$	$n_{01}^{(1)} = 0$
	0	$n_{10}^{(1)}$	$n_{00}^{(1)}$
Total		$n_1^{(1)}$	$n_0^{(1)}$

		$y=1$ (Case)	$y=0$ (Control)
z_2	1	$n_{11}^{(2)}$	$n_{01}^{(2)} = 0$
	0	$n_{10}^{(2)}$	$n_{00}^{(2)}$
Total		$n_1^{(2)}$	$n_0^{(2)}$

The full likelihood for a sample of n_1 cases $(y=1)$ and n_0 controls $(y=0)$ is,

$$\prod_{i=0}^{n_1} P(x_i, z_{i1}, z_{i2} \mid y_i = 1, s_i = 1) \prod_{i=1}^{n_0} P(x_i, z_{i1}, z_{i2} \mid y_i = 0, s_i = 1) \tag{3.3.2}$$

For an individual term in the likelihood function shown in equation (3.3.2), the simplification is given using the Bayes theorem.

$$P(x, z_1, z_2 \mid y, s = 1) = \frac{P(x, z_1, z_2 \mid s = 1) \cdot P(y \mid x, z_1, z_2, s = 1)}{P(y \mid s = 1)} \tag{3.3.3}$$

The first term in the numerator of equation (3.3.3) yields,

$$P(x, z_1, z_2 \mid s = 1) = \frac{P(x, z_1, z_2, s = 1)}{P(s = 1)}$$

$$= \frac{P(z_1, z_2 \mid x, s = 1) \cdot P(x, s = 1)}{P(s = 1)}$$

$$= \frac{P(z_1, z_2 \mid x, s = 1) \cdot P(x \mid s = 1) \cdot P(s = 1)}{P(s = 1)}$$

$$= P(z_1, z_2 \mid x, s = 1) \cdot P(x \mid s = 1) \tag{3.3.4}$$

The second term in the numerator of equation (3.3.3) yields,

$$P(y \mid x, z_1, z_2, s = 1) = \frac{P(y \mid x, s = 1) \cdot P(z_1, z_2 \mid y, x, s = 1)}{P(z_1, z_2 \mid x, s = 1)}$$

$$= \frac{P(y \mid x, s=1) \cdot \dfrac{P(z_1, z_2, y, x, s=1)}{P(y, x, s=1)}}{P(z_1, z_2 \mid x, s=1)}$$

$$= \frac{P(y \mid x, s=1) \cdot \dfrac{P(z_1 \mid z_2, y, x, s=1) \cdot P(z_2, y, x, s=1)}{P(y, x, s=1)}}{P(z_1, z_2 \mid x, s=1)}$$

$$= \frac{P(y \mid x, s=1) \cdot \dfrac{P(z_1 \mid z_2, y, x, s=1) \cdot P(z_2 \mid y, x, s=1) \cdot P(y, x, s=1)}{P(y, x, s=1)}}{P(z_1, z_2 \mid x, s=1)}$$

$$= \frac{P(y \mid x, s=1) \cdot P(z_1 \mid z_2, y, x, s=1) \cdot P(z_2 \mid y, x, s=1)}{P(z_1, z_2 \mid x, s=1)} \tag{3.3.5}$$

Substituting (3.3.4) and (3.3.5) in (3.3.3) we get,

$$P(x, z_1, z_2 \mid y, s=1) = \frac{P(z_1, z_2 \mid x, s=1) \cdot P(x \mid s=1) \cdot \dfrac{P(y \mid x, s=1) \cdot P(z_1 \mid z_2, y, x, s=1) \cdot P(z_2 \mid y, x, s=1)}{P(z_1, z_2 \mid x, s=1)}}{P(y \mid s=1)}$$

$$= P(y \mid x, s=1) \cdot P(z_1 \mid z_2, y, x, s=1) \cdot P(z_2 \mid y, x, s=1) \cdot \frac{P(x \mid s=1)}{P(y \mid s=1)}$$

$$= P(y \mid x, s=1) \cdot P(z_1 \mid y, x, s=1) \cdot P(z_2 \mid y, x, s=1) \cdot \frac{P(x \mid s=1)}{P(y \mid s=1)}, \text{ as } z_1 \text{ and}$$

z_2 are independent. (3.3.6)

Let α_1 be the proportion of the covariate $z_1 = 1$ and α_2 be the proportion of the covariate $z_2 = 1$ in the case group

$$P(z_1 = 1 \mid y = 1, x, s = 1) = \alpha_1$$

and

$$P(z_2 = 1 \mid y = 1, x, s = 1) = \alpha_2, \ \alpha_1 \text{ and } \alpha_2 \text{ depends on } x$$

In the case when $z_1 = 0$ and $z_2 = 0$, we have

$$P(z_1 = 0 \mid y = 1, x, s = 1) = 1 - \alpha_1$$

and

$$P(z_2 = 0 \mid y = 1, x, s = 1) = 1 - \alpha_2$$

The following model is obtained for the joint distribution of risk factors

$P(x, z_1, z_2 \mid y, s = 1) = P(x, Z_1 = z_1, Z_2 = z_2 \mid Y = y, s = 1)$, $z_1 = z_2 = 0,1$, $y = 0,1$ in the case-control study,

$P(x, Z_1 = z_1, Z_2 = z_2 \mid Y = y, s = 1) =$

$$\alpha_1^{z_1 y}(1-\alpha_1)^{(1-z_1)y} \cdot \alpha_2^{z_2 y}(1-\alpha_2)^{(1-z_2)y} \cdot \left(\frac{e^{\beta_0^* + \beta' x}}{1+e^{\beta_0^* + \beta' x}}\right)^y \left(\frac{(1-z_1)(1-z_2)}{1+e^{\beta_0^* + \beta' x}}\right)^{1-y} \cdot \frac{P(x \mid s=1)}{P(Y = y \mid s = 1)} \quad (3.3.7)$$

To find the MLE of α_1, α_2, and β, we substitute (3.3.7) in (3.3.2) for the n_1 cases and n_0 controls, we have the likelihood is proportional to

$$L = \prod_{i=1}^{n} \alpha_1^{z_{i1} y_i}(1-\alpha_1)^{(1-z_{i1})y_i} \cdot \alpha_2^{z_{i2} y_i}(1-\alpha_2)^{(1-z_{i2})y_i} \cdot \left(\frac{e^{\beta_0^* + \beta' x_i}}{1+e^{\beta_0^* + \beta' x_i}}\right)^{y_i} \left(\frac{(1-z_{i1})(1-z_{i2})}{1+e^{\beta_0^* + \beta' x_i}}\right)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\prod_{k=1}^{2} \alpha_k^{z_{ik} y_i}(1-\alpha_k)^{(1-z_{ik})y_i} \right) \cdot \left(\frac{e^{\beta_0^* + \beta' x_i}}{1+e^{\beta_0^* + \beta' x_i}}\right)^{y_i} \left(\frac{(1-z_{i1})(1-z_{i2})}{1+e^{\beta_0^* + \beta' x_i}}\right)^{1-y_i} \quad (3.3.8)$$

Taking ln on both sides of the equation (3.3.8) and we get,

$$\ln L = \sum_{i=1}^{n} \Big[z_{i1} y_i \ln \alpha_1 + (1-z_{i1}) y_i \ln(1-\alpha_1) + z_{i2} y_i \ln \alpha_2 + (1-z_{i2}) y_i \ln(1-\alpha_2)$$
$$+ y_i \ln\left(\frac{e^{\beta_0^* + \beta' x_i}}{1+e^{\beta_0^* + \beta' x_i}}\right) + (1-y_i) \ln\left(\frac{(1-z_{i1})(1-z_{i2})}{1+e^{\beta_0^* + \beta' x_i}}\right) \Big]$$

Setting $\frac{d(\ln L)}{d\alpha_k} = 0$, $k = 1,2$ we obtain,

$$\sum_{i=1}^{n}\left[\frac{z_{i1}y_i}{\alpha_1}-\frac{(1-z_{i1})y_i}{1-\alpha_1}\right]=0$$

$$\sum_{i=1}^{n}\left[\frac{z_{i2}y_i}{\alpha_2}-\frac{(1-z_{i2})y_i}{1-\alpha_2}\right]=0$$

Since α_1 and α_2 depend on *x*, we let $n_{11}^{(ki)}, n_{10}^{(ki)}$, $k=1,2$ and $i=1,2,...,I$ represent the number of cases when $Z_k=1,0$ for *k*=1,2 for all permissible strata of the covariates respectively, and let α_{ki}, *k*=1,2 be the proportion of $Z_k=1$, *k*=1,2 in stratum numbered *i*. Therefore, we have

$$\frac{n_{11}^{(1i)}}{\alpha_{1i}}-\frac{n_{10}^{(1i)}}{1-\alpha_{1i}}=0 \text{ and } \frac{n_{11}^{(2i)}}{\alpha_{2i}}-\frac{n_{10}^{(2i)}}{1-\alpha_{2i}}=0$$

This implies, $\hat{\alpha}_{1i}=\frac{n_{11}^{(1i)}}{n_{11}^{(1i)}+n_{10}^{(1i)}}=\frac{n_{11}^{(1i)}}{n_1^{(1i)}}$ and $\hat{\alpha}_{2i}=\frac{n_{11}^{(2i)}}{n_{11}^{(2i)}+n_{10}^{(2i)}}=\frac{n_{11}^{(2i)}}{n_1^{(2i)}}$

To obtain the variance of $\hat{\alpha}_k\equiv\hat{\alpha}_{ki}, k=1,2$, we take the second derivative and we get

$$\frac{d^2(\ln L)}{d\alpha_1^2}=\sum_{i=1}^{n}\left[-\frac{z_{i1}y_i}{\alpha_1^2}+\frac{(1-z_{i1})y_i}{(1-\alpha_1)^2}\right]$$

$$=\frac{-n_{11}^{(1i)}}{\alpha_1^2}+\frac{-n_{10}^{(1i)}}{(1-\alpha_1)^2}$$

We have, $\widehat{Var}(\hat{\alpha}_1)=\dfrac{1}{-E\left(\dfrac{d^2(\ln L)}{d\hat{\alpha}_1^2}\right)}$

$$=\frac{1}{-\left(\dfrac{-n_{11}^{(1i)}}{\hat{\alpha}_1^2}+\dfrac{-n_{10}^{(1i)}}{(1-\hat{\alpha}_1)^2}\right)}$$

$$=\frac{1}{\dfrac{n_{11}^{(1i)}}{\hat{\alpha}_1^2}+\dfrac{n_{10}^{(1i)}}{(1-\hat{\alpha}_1)^2}}$$

$$= \frac{1}{\frac{n_{11}^{(1i)}}{\left(\frac{n_{11}^{(1i)}}{n_1^{(1i)}}\right)^2} + \frac{n_{10}^{(1i)}}{\left(1-\frac{n_{11}^{(1i)}}{n_1^{(1i)}}\right)^2}}$$

$$= \frac{1}{\frac{\left(n_1^{(1i)}\right)^2}{n_{11}^{(1i)}} + \frac{\left(n_1^{(1i)}\right)^2}{n_{10}^{(1i)}}}$$

$$= \frac{1}{\frac{\left(n_1^{(1i)}\right)^2\left[n_{10}^{(1i)} + n_{11}^{(1i)}\right]}{n_{11}^{(1i)} n_{10}^{(1i)}}}$$

That is, $\widehat{Var}(\hat{\alpha}_1) = \frac{n_{11}^{(1i)} n_{10}^{(1i)}}{(n_1^{(1i)})^3}$

Similarly, $\widehat{Var}(\hat{\alpha}_2) = \frac{n_{11}^{(2i)} n_{10}^{(2i)}}{(n_1^{(2i)})^3}$

We consider the expression $\hat{\alpha}_k, k = 1,2$ can be simplified in the case where $\alpha_{ki}, k = 1,2$ is the same across all permissible strata. Summarizing the above analysis we have the following theorem.

Result 3.3.1: The ML estimates of α_k, $\hat{\alpha}_k$, k =1,2 can be obtained based on the outcome variable y and the risk factors $z_1, z_2, x_{1,} x_{2,}..., x_p$ under case-control sampling design in the model (3.3.7) ,

$$\hat{\alpha}_k = \frac{n_{11}^{(k)}}{n_1^{(k)}} \ , \ k = 1,2$$

with the variance $\widehat{Var}(\hat{\alpha}_k) = \frac{n_{11}^{(k)} n_{10}^{(k)}}{(n_1^{(k)})^3}$, where $n_{10}^{(k)}, n_{11}^{(k)}$ are the number of cases in the $z_k = 0$ and $z_k = 1$, $k = 1,2$ groups, respectively and $n_1^{(k)} = n_{10}^{(k)} + n_{11}^{(k)}$, $k = 1,2$, the total number of cases.

For the other parameters involved in the model, Model (3.3.7) can be expressed in the following forms:

$$P(x, z_1 = 1, z_2 = 1 \mid y = 1, s = 1) = \alpha_1 \cdot \alpha_2 \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.9}$$

$$P(x, z_1 = 1, z_2 = 0 \mid y = 1, s = 1) = \alpha_1 \cdot (1 - \alpha_2) \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.10}$$

$$P(x, z_1 = 0, z_2 = 1 \mid y = 1, s = 1) = (1 - \alpha_1) \cdot \alpha_2 \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.11}$$

$$P(x, z_1 = 0, z_2 = 0 \mid y = 1, s = 1) = (1 - \alpha_1) \cdot (1 - \alpha_2) \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.12}$$

$$P(x, z_1 = 1, z_2 = 1 \mid y = 0, s = 1) = 0 \tag{3.3.13}$$

$$P(x, z_1 = 1, z_2 = 0 \mid y = 0, s = 1) = 0 \tag{3.3.14}$$

$$P(x, z_1 = 0, z_2 = 1 \mid y = 0, s = 1) = 0 \tag{3.3.15}$$

$$P(x, z_1 = 0, z_2 = 0 \mid y = 0, s = 1) = \frac{1}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.16}$$

By combining equation (3.3.9)-(3.3.12), we have

Result 3.3.2: The ML estimates of $\beta_1^*, ..., \beta_p^*$ for model (3.3.7) is the same as the estimates from

$$P(x \mid Y = 1, s = 1) = \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \quad \frac{P(x \mid s = 1)}{P(Y = 1 \mid s = 1)}$$

and

$$P(x \mid Y = 0, s = 1) = \frac{1}{1 + e^{\beta_0^* + \beta' x}} \quad \frac{P(x \mid s = 1)}{P(Y = 1 \mid s = 1)}$$

3.3.2.2 Consider the case when the rare risk factors z_1 and z_2 are not independent

Suppose the variables z_1 and z_2 are not independent and the following table gives the cross-classification between these two variables

Table 3.3.3: Cross-classification between z_1 and z_2

		Z_2	
		1	0
Z_1	1	α_{11}	α_{10}
	0	α_{01}	α_{00}

Suppose that $P(z_1 = 1, z_2 = 1 \mid y = 1, x) = \alpha_{11}$

$P(z_1 = 1, z_2 = 0 \mid y = 1, x) = \alpha_{10}$

$P(z_1 = 0, z_2 = 1 \mid y = 1, x) = \alpha_{01}$

and $P(z_1 = 0, z_2 = 0 \mid y = 1, x) = \alpha_{00}$

where, $\alpha_{11}, \alpha_{10}, \alpha_{01}$, and α_{00} depend on the covariate *x*.

The following model is proposed for the joint distribution of risk factors where the variables z_1 and z_2 are not independent

$P(x, z_1, z_2 \mid y, s = 1) = P(x, Z_1 = z_1, Z_2 = z_2 \mid Y = y, s = 1)$, $z_1 = z_2 = 0,1$, $y = 0,1$ in the case-control study:

$$P(x, z_1, z_2 \mid y) = \alpha_{11}^{z_1 z_2 y} \alpha_{10}^{z_1 (1-z_2) y} \alpha_{01}^{(1-z_1) z_2 y} \alpha_{00}^{(1-z_1)(1-z_2) y} \left(\frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \right)^{y} \left(\frac{(1-z_1)(1-z_2)}{1 + e^{\beta_0^* + \beta' x}} \right)^{1-y} \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.17}$$

Theorem 3.3.3. If $\alpha_{11} = \alpha_1 \alpha_2$, then the model defined in equation (3.3.17) is similar to the model in equation (3.3.7).

Proof: Suppose $\alpha_{11} + \alpha_{10} = \alpha_1$

$\alpha_{11} + \alpha_{01} = \alpha_2$

and $\alpha_{00} = 1 - \alpha_1 - \alpha_2 + \alpha_{11}$

then the model (3.3.17) becomes

$$P(x, z_1, z_2 \mid y) = (\alpha_1\alpha_2)^{z_1z_2y}(\alpha_1 - \alpha_{11})^{z_1(1-z_2)y}(\alpha_2 - \alpha_{11})^{(1-z_1)z_2y}(1 - \alpha_1 - \alpha_2 - \alpha_{11})^{(1-z_1)(1-z_2)y}$$

$$\left(\frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}}\right)^y \left(\frac{(1-z_1)(1-z_2)}{1 + e^{\beta_0^* + \beta' x}}\right)^{1-y} \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)}$$

That is,

$$P(x, z_1, z_2 \mid y) = (\alpha_1\alpha_2)^{z_1z_2y}(\alpha_1 - \alpha_1\alpha_2)^{z_1(1-z_2)y}(\alpha_2 - \alpha_1\alpha_2)^{(1-z_1)z_2y}(1 - \alpha_1 - \alpha_2 - \alpha_1\alpha_2)^{(1-z_1)(1-z_2)y}$$

$$\left(\frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}}\right)^y \left(\frac{(1-z_1)(1-z_2)}{1 + e^{\beta_0^* + \beta' x}}\right)^{1-y} \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)}$$

That is,

$$P(x, z_1, z_2 \mid y) = \alpha_1^{z_1z_2y}\alpha_2^{z_1z_2y}\alpha_1^{z_1(1-z_2)y}(1 - \alpha_2)^{z_1(1-z_2)y}\alpha_2^{(1-z_1)z_2y}(1 - \alpha_1)^{(1-z_1)z_2y}(1 - \alpha_1)^{(1-z_1)(1-z_2)y}(1 - \alpha_2)^{(1-z_1)(1-z_2)y}$$

$$\left(\frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}}\right)^y \left(\frac{(1-z_1)(1-z_2)}{1 + e^{\beta_0^* + \beta' x}}\right)^{1-y} \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)}$$

That is,

$$P(x, z_1, z_2 \mid y) = \alpha_1^{z_1z_2y}\alpha_2^{z_1z_2y}\alpha_1^{z_1y}\alpha_1^{-z_1z_2y}(1 - \alpha_2)^{-z_1(1-z_2)y}\alpha_2^{z_2y}\alpha_2^{-z_1z_2y}(1 - \alpha_1)^{-(1-z_1)z_2y}(1 - \alpha_1)^{(1-z_1)y}(1 - \alpha_1)^{-(1-z_1)z_2y}$$

$$(1 - \alpha_2)^{(1-z_2)y}(1 - \alpha_2)^{-z_1(1-z_2)y}\left(\frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}}\right)^y \left(\frac{(1-z_1)(1-z_2)}{1 + e^{\beta_0^* + \beta' x}}\right)^{1-y} \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)}$$

That is,

$$P(x, z_1, z_2 \mid y) = \alpha_1^{z_1y}(1 - \alpha_1)^{(1-z_1)y}\alpha_2^{z_2y}(1 - \alpha_2)^{(1-z_2)y}\left(\frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}}\right)^y \left(\frac{(1-z_1)(1-z_2)}{1 + e^{\beta_0^* + \beta' x}}\right)^{1-y} \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)},$$

which is the model shown in (3.3.7).

Hence, the proof follows.

Now, to find the MLE of $\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00}$, and β, we substitute equation (3.3.17) in (3.3.2) for the n_1 cases and n_0 controls and we have the likelihood is proportional to

$$L=\prod_{i=1}^{n}\alpha_{11}^{z_{i1}z_{i2}y_i}\alpha_{10}^{z_{i1}(1-z_{i2})y_i}\alpha_{01}^{(1-z_{i1})z_{i2}y_i}\alpha_{00}^{(1-z_{i1})(1-z_{i2})y_i}\cdot\left(\frac{e^{\beta_0^*+\beta' x_i}}{1+e^{\beta_0^*+\beta' x_i}}\right)^{y_i}\left(\frac{(1-z_{i1})(1-z_{i2})}{1+e^{\beta_0^*+\beta' x_i}}\right)^{1-y_i} \tag{3.3.18}$$

Taking ln on both sides of the equation (3.3.18) and we get,

$$\ln L=\sum_{i=1}^{n}\left[z_{i1}z_{i2}y_i\ln\alpha_{11}+z_{i1}(1-z_{i2})y_i\ln\alpha_{10}+(1-z_{i1})z_{i2}y_i\ln\alpha_{01}+(1-z_{i1})(1-z_{i2})y_i\ln\alpha_{00}\right.$$
$$\left.+y_i\ln\left(\frac{e^{\beta_0^*+\beta' x_i}}{1+e^{\beta_0^*+\beta' x_i}}\right)+(1-y_i)\ln\left(\frac{(1-z_{i1})(1-z_{i2})}{1+e^{\beta_0^*+\beta' x_i}}\right)\right]$$

Now, we take the derivatives with respect to $\alpha_{11},\alpha_{10},$ and α_{01} respectively. This yields,

$$\frac{d\ln L}{d\alpha_{11}}=\sum_{i=1}^{n}\left[\frac{z_{i1}z_{i2}y_i}{\alpha_{11}}-\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}}\right]=0$$

That is, $$\sum_{i=1}^{n}\frac{z_{i1}z_{i2}y_i}{\alpha_{11}}=\sum_{i=1}^{n}\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}}$$

$$\frac{d\ln L}{d\alpha_{10}}=\sum_{i=1}^{n}\left[\frac{z_{i1}(1-z_{i2})y_i}{\alpha_{10}}-\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}}\right]=0$$

That is, $$\sum_{i=1}^{n}\frac{z_{i1}(1-z_{i2})y_i}{\alpha_{10}}=\sum_{i=1}^{n}\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}}$$

$$\frac{d\ln L}{d\alpha_{01}}=\sum_{i=1}^{n}\left[\frac{(1-z_{i1})z_{i2}y_i}{\alpha_{01}}-\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}}\right]=0$$

That is, $$\sum_{i=1}^{n}\frac{(1-z_{i1})z_{i2}y_i}{\alpha_{01}}=\sum_{i=1}^{n}\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}}$$

Thus,

$$\sum_{i=1}^{n}\frac{z_{i1}z_{i2}y_i}{\alpha_{11}}=\sum_{i=1}^{n}\frac{z_{i1}(1-z_{i2})y_i}{\alpha_{10}}=\sum_{i=1}^{n}\frac{(1-z_{i1})z_{i2}y_i}{\alpha_{01}}=\sum_{i=1}^{n}\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}}$$

$$=\frac{\sum_{i=1}^{n}z_{i1}z_{i2}y_i+\sum_{i=1}^{n}z_{i1}y_i-\sum_{i=1}^{n}z_{i1}z_{i2}y_i+\sum_{i=1}^{n}z_{i2}y_i-\sum_{i=1}^{n}z_{i1}z_{i2}y_i+\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}z_{i1}y_i-\sum_{i=1}^{n}z_{i2}y_i+\sum_{i=1}^{n}z_{i1}z_{i2}y_i}{\alpha_{11}+\alpha_{10}+\alpha_{01}+\alpha_{00}}$$

$$=\sum_{i=1}^{n}y_i \text{ as } \alpha_{11}+\alpha_{10}+\alpha_{01}+\alpha_{00}=1$$

Since α_{11}, α_{10}, α_{01}, and α_{00} depend on x, we let $n_{11}^{(i)}, n_{10}^{(i)}, n_{01}^{(i)}$, and $n_{00}^{(i)}$, $i=1,2,...,I$ represent the number of cases of the combination of the variables $z_1, z_2 = 1,0$ for all permissible strata of the covariates respectively, and let α_{11i}, α_{10i}, α_{01i}, and α_{00i} be the proportion of the combination of $z_1, z_2 = 1,0$ in stratum numbered *i*. Therefore, we have

$$\frac{n_{11}^{(i)}}{\alpha_{11i}}=\sum_{i=1}^{n}y_i=n_1^{(i)},\ \frac{n_{10}^{(i)}}{\alpha_{10i}}=n_1^{(i)},\ \frac{n_{01}^{(i)}}{\alpha_{01i}}=n_1^{(i)}, \text{ and } \frac{n_{00}^{(i)}}{\alpha_{00i}}=n_1^{(i)}$$

This implies, $\hat{\alpha}_{11i}=\frac{n_{11}^{(i)}}{n_1^{(i)}}$, $\hat{\alpha}_{10i}=\frac{n_{10}^{(i)}}{n_1^{(i)}}, \hat{\alpha}_{01i}=\frac{n_{01}^{(i)}}{n_1^{(i)}}$, and $\hat{\alpha}_{00i}=\frac{n_{00}^{(i)}}{n_1^{(i)}}$

We consider the proportions are the same across all permissible strata. Therefore, summarizing the above we have the following statement.

Result 3.3.4: The ML estimates of $\alpha_{11}, \alpha_{10}, \alpha_{11}$, and α_{00} for Model (3.3.17) under case-control data can be obtained by,

$$\hat{\alpha}_{11}=\frac{n_{11}}{n_1},\ \hat{\alpha}_{10}=\frac{n_{10}}{n_1}, \hat{\alpha}_{01i}=\frac{n_{01}}{n_1}, \text{ and } \hat{\alpha}_{00}=\frac{n_{00}}{n_1}$$

and estimated variances are obtained by

$\hat{V}ar(\hat{\alpha}_{11}) = \frac{n_{11}(n_1 - n_{11})}{n_1^3}$, $\hat{V}ar(\hat{\alpha}_{10}) = \frac{n_{10}(n_1 - n_{10})}{n_1^3}$, $\hat{V}ar(\hat{\alpha}_{01}) = \frac{n_{01}(n_1 - n_{01})}{n_1^3}$, and

$\hat{V}ar(\hat{\alpha}_{00}) = \frac{n_{10}(n_1 - n_{10})}{n_1^3}$

where n_{ij}, $i, j = 1,0$ is the number of cases of the combination of variables $z_1, z_2 = 1,0$ and n_1 is the total number of cases.

For the other parameters involved in the model, Model (3.3.17) can be expressed in the following forms:

$$P(x, z_1 = 1, z_2 = 1 \mid y = 1, s = 1) = \alpha_{11} \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.19}$$

$$P(x, z_1 = 1, z_2 = 0 \mid y = 1, s = 1) = \alpha_{10} \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.20}$$

$$P(x, z_1 = 0, z_2 = 1 \mid y = 1, s = 1) = \alpha_{01} \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.21}$$

$$P(x, z_1 = 0, z_2 = 0 \mid y = 1, s = 1) = \alpha_{00} \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.22}$$

$$P(x, z_1 = 1, z_2 = 1 \mid y = 0, s = 1) = 0 \tag{3.3.23}$$

$$P(x, z_1 = 1, z_2 = 0 \mid y = 0, s = 1) = 0 \tag{3.3.24}$$

$$P(x, z_1 = 0, z_2 = 1 \mid y = 0, s = 1) = 0 \tag{3.3.25}$$

$$P(x, z_1 = 0, z_2 = 0 \mid y = 0, s = 1) = \frac{1}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{3.3.26}$$

By combining equation (3.3.19)-(3.3.22), we have

Result 3.3.5: The ML estimates of $\beta_1^*, ..., \beta_p^*$ for model (3.3.17) are the same as the estimates from

$$P(x \mid Y=1, s=1) = \frac{e^{\beta_0^* + \beta' x}}{1+e^{\beta_0^* + \beta' x}} \quad \frac{P(x \mid s=1)}{P(Y=1 \mid s=1)}$$

and

$$P(x \mid Y=0, s=1) = \frac{1}{1+e^{\beta_0^* + \beta' x}} \quad \frac{P(x \mid s=1)}{P(Y=1 \mid s=1)}$$

3.3.3 A Hybrid Logistic Model: *k-V*ariate Case

3.3.3.1 When the rare risk factors $z_1, z_2, ..., z_k$ are independent

In this case, we consider *k* covariates, $z_1, z_2, z_3, ..., z_{k-1}$, and z_k have no event in the control group. Consider (z^T, x^T) is a set of explanatory variables in the model, where $z^T = (z_1, z_2, z_3, ..., z_k)$ represents a rare risk factors and $x^T = (x_1, x_2, ..., x_p)$ represents the other risk factors. In the case, assuming $z_1, z_2, \ldots, z_{k-1}$, and z_k are independent and each variable consists two groups 1 and 0, we propose the following model,

$$P(x, Z_1 = z_1, Z_2 = z_2, ..., Z_k = z_k \mid Y = y, s=1) = \prod_{j=1}^{k} \alpha_j^{z_{ji} y_i} (1-\alpha_j)^{(1-z_{ji}) y_i} \cdot \left(\frac{e^{\beta_0^* + \beta' x}}{1+e^{\beta_0^* + \beta' x}} \right)^y$$
$$\left(\frac{(1-z_1)(1-z_2)...(1-z_k)}{1+e^{\beta_0^* + \beta' x}} \right)^{1-y} \cdot \frac{P(x \mid s=1)}{P(Y=y \mid s=1)} \qquad (3.3.27)$$

where α_j, j = 1,2,…,k be the proportion of the covariate $z_j = 1$ in the case group

$$P(z_j = 1 \mid y=1, x, s=1) = \alpha_j \ , j=1,2,\ldots,k$$

The estimates for the case-control data can be obtained by finding the MLE of α_j, j=1,2,…,k which is similar to described in Theorem 3.3.1. The estimates of other parameters can be obtained by applying Theorem 3.3.2.

3.3.3.2 When the rare risk factors $z_1, z_2, ..., z_k$ are not independent

Suppose the rare risk factors $z_1, z_2, \ldots, z_k$ are not independent and $x^T = (x_1, x_2, ..., x_p)$ represents the other risk factors. The following model is proposed for the joint distribution of risk factors

$P(x, z_1, z_2, ..., z_k \mid y, s = 1) = P(x, Z_1 = z_1, ..., Z_k = z_k \mid Y = y, s = 1),\ z_1 = ... = z_k = 0,1,\ y = 0,1$ in the case-control study:

$$P(x, z_1, z_2, ..., z_k \mid y) = \prod_{\substack{(i_1 i_2 \ldots i_k): \\ i_1, i_2, ..., i_k = 0,1}} \alpha_{i_1 i_2 \cdots i_k}^{\prod_{j=1}^{k} z_j^{i_j} (1-z_j)^{1-i_j}} \left(\frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \right)^y \left(\frac{(1-z_1)...(1-z_k)}{1 + e^{\beta_0^* + \beta' x}} \right)^{1-y} \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \quad (3.3.28)$$

The estimates for the case-control data can be obtained by finding the MLE of $\prod_{\substack{(i_1 i_2 \ldots i_k): \\ i_1, i_2, ..., i_k = 0,1}} \hat{\alpha}_{i_1 i_2 \cdots i_k}$ which is similar to described in Theorem 3.3.4. The estimates of other parameters can be obtained by applying Theorem 3.3.5.

CHAPTER 4

THE MULTINOMIAL HYBRID LOGISTIC REGRESSION MODEL

4.1 The Multinomial Distribution

4.1.1 The Distribution

Multinomial distribution is the generalization of the binomial distribution. In the case of binomial distribution, each trial consists of two outcomes with probabilities *p* and *q*. Let each trial consists of *k* mutually exclusive outcomes $E_1, E_2, ..., E_k$ with probabilities $p_1, p_2, ..., p_k$ such that $\sum_{i=1}^{k} p_i = 1$. If this experiment is repeated *n* times, then the probability that E_1 occurs x_1 times, E_2 occurs x_2 times, … , E_k occurs x_k times is

$$f(x_1, x_2, ..., x_k; n, p_1, p_2, ..., p_k) = \frac{n!}{x_1! x_2! ... x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k} \tag{4.1.1}$$

such that $n = \sum_{i=1}^{k} x_i$ and $\sum_{i=1}^{k} p_i = 1$.

The probability function defined in (4.1.1) is known as multinomial distribution, since the probability function is the general term of the multinomial expansion $(p_1 + p_2 + ... + p_k)^n$.

To estimate the parameters of multinomial distribution, we consider the kernel of the probability mass function defined in (4.1.1). Thus, the log-likelihood function becomes

$$\ell(p) = \log \prod_{i=1}^{k} p_i^{x_i} = \sum_{i=1}^{k} x_i \log p_i$$

$$= x_1 \log p_1 + x_2 \log p_2 + ... + x_k \log p_k$$

$$= x_1 \log p_1 + x_2 \log p_2 + ... + x_k \log(1 - p_1 - p_2 ... - p_{k-1})$$

Taking the partial derivatives with respect to $p_1, p_2, ..., p_{k-1}$ and then setting them equal to 0, we have

$$\frac{\partial \ell(p)}{\partial p_1} = \frac{x_1}{p_1} - \frac{x_k}{1-p_1-p_2...-p_{k-1}} = 0 \quad \Rightarrow \frac{x_1}{p_1} = \frac{x_k}{1-p_1-p_2...-p_{k-1}}$$

$$\frac{\partial \ell(p)}{\partial p_2} = \frac{x_2}{p_2} - \frac{x_k}{1-p_1-p_2...-p_{k-1}} = 0 \quad \Rightarrow \frac{x_2}{p_2} = \frac{x_k}{1-p_1-p_2...-p_{k-1}}$$

$$\vdots$$
$$\vdots$$
$$\vdots$$

$$\frac{\partial \ell(p)}{\partial p_{k-1}} = \frac{x_{k-1}}{p_{k-1}} - \frac{x_k}{1-p_1-p_2...-p_{k-1}} = 0 \quad \Rightarrow \frac{x_{k-1}}{p_{k-1}} = \frac{x_k}{1-p_1-p_2...-p_{k-1}}$$

This implies that

$$\frac{x_1}{p_1} = \frac{x_2}{p_2} = ... = \frac{x_{k-1}}{p_{k-1}} = \frac{x_k}{1-p_1-p_2...-p_{k-1}} = \frac{x_1+x_2+...+x_{k-1}+x_k}{p_1+p_2+...+p_{k-1}+1-p_1-p_2...-p_{k-1}} = n$$

Therefore,

$$\hat{p}_1 = \frac{x_1}{n}, \hat{p}_2 = \frac{x_2}{n}, ..., \hat{p}_{k-1} = \frac{x_{k-1}}{n}, \hat{p}_k = \frac{x_k}{n}$$

Theorem 4.1.1. Let $X_1, X_2, ..., X_k$ be *k* discrete random variables which follows multinomial distribution with probability function defined in (4.1.1) then

$$M_{X_1, X_2, ..., X_k}(t_1, t_2, ..., t_k) = (p_1 e^{t_1} + p_2 e^{t_2} + ... + p_k e^{t_k})^n$$

where $M_{X_1, X_2, ..., X_k}(t_1, t_2, ..., t_k)$ is the moment generating function of $X_1, X_2, ..., X_k$.

$E(X_i) = np_i$ for all $i = 1,2,...,k$.

$Var(X_i) = np_i(1-p_i)$, and

$Cov(X_i, X_j) = -np_i p_j$ for $i \neq j$.

Proof: By definition,

$$M_{X_1, X_2, ..., X_k}(t_1, t_2, ..., t_k) = E\left(e^{t_1 x_1 + t_2 x_2 + ... + t_k x_k}\right)$$

$$= \sum_{x_1 x_2 \ldots x_k} \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} e^{\sum_{i=1}^{k} t_i x_i}$$

$$= \sum_{x_1 x_2 \ldots x_k} \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} e^{\sum_{i=1}^{k} t_i x_i}$$

$$= (p_1 e^{t_1} + p_2 e^{t_2} + \ldots + p_k e^{t_k})^n$$

Now, $E(X_i) = \left(\frac{\partial M(t_1, t_1 \ldots, t_k)}{\partial t_i} \right)_{t_1 = t_2 = \ldots = t_k = 0}$

$$= \left(np_i e^{t_i} (p_1 e^{t_1} + p_2 e^{t_2} + \ldots + p_k e^{t_k}) \right)_{t_1 = t_2 = \ldots = t_k = 0}^{n-1}$$

$$= np_i$$

$$E(X_i^2) = \left(\frac{\partial M^2(t_1, t_1 \ldots, t_k)}{\partial t_i^2} \right)_{t_1 = t_2 = \ldots = t_k = 0}$$

$$= \left(n(n-1) e^{t_i} p_i^2 e^{t_i} (p_1 e^{t_1} + p_2 e^{t_2} + \ldots + p_k e^{t_k})^{n-2} + np_i e^{t_i} (p_1 e^{t_1} + p_2 e^{t_2} + \ldots + p_k e^{t_k})^{n-1} \right)_{t_1 = t_2 = \ldots = t_k = 0}$$

$$= n(n-1) p_i^2 np_i$$

Hence, $Var(X_i) = E(X_i^2) - (E(X_i))^2$

$$= n(n-1) p_i^2 np_i - n^2 p_i^2)$$

$$= np_i (1 - p_i)$$

Now, $Cov(X_i, X_j) = E(X_i X_j) - E(X_i) E(X_j)$

$$E(X_i X_j) = \left(\frac{\partial M^2(t_1, t_1 \ldots, t_k)}{\partial t_i \, \partial t_j} \right)_{t_1 = t_2 = \ldots = t_k = 0}$$

$$= \left(np_i e^{t_i} (n-1) p_j (p_1 e^{t_1} + p_2 e^{t_2} + \ldots + p_k e^{t_k}) \right)_{t_1 = t_2 = \ldots = t_k = 0}^{n-2}$$

$$= n(n-1)p_i\,p_j$$

Thus, $Cov(X_i, X_j) = n(n-1)p_i\,p_j - n^2 p_i\,p_j$

$$= -np_i\,p_j$$

Hence, the proof follows.

4.1.2 The Asymptotic Distribution

Let $\hat{p} = (\hat{p}_1, \hat{p}_2, ..., \hat{p}_k)^T$, where $\hat{p}_i = \frac{x_i}{n}, i = 1,2,...,k$

We have, $E(\hat{P}) = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{pmatrix}$

$$Cov(\hat{P}) = \frac{1}{n}\begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \dots & -p_2p_k \\ \vdots & & & \\ -p_kp_1 & -p_kp_2 & \dots & p_k(1-p_k) \end{bmatrix}$$

$$= \frac{1}{n}\left[diag(p) - pp^T\right]$$

where $diag(p)$ is the diagonal matrix with the elements of p on the main diagonal.

This covariance matrix is singular because of the linear dependence, $\sum_{i=1}^{k} p_i = 1$.

Using the multivariate central limit theorem (Rao, 1973), we have

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N\left[0, diag(p) - pp^T\right]$$

By the delta method, functions of $\hat{p}$ having nonzero differential at p are also asymptotically normal. Let $g(t_1, t_2, ..., t_k)$ be a differentiable function, and let

$$\phi_i = \frac{\partial g}{\partial p_i}, \quad i = 1,2,...,k$$

denote $\frac{\partial g}{\partial t_i}$ evaluated at $t = p$.

By the delta method,

$$\sqrt{n}\left[g(\hat{p}) - g(p)\right] \xrightarrow{d} N\left(0, \phi^T [diag(p) - pp^T]\phi\right)$$

where the asymptotic variance equals $\phi^T diag(p)\phi - (\phi^T p)^2 = \sum_{i=1}^{k} p_i \phi_i^2 - \left(\sum_{i=1}^{k} p_i \phi_i\right)^2$.

4.2 The Multinomial Logistic Regression Model

Multinomial logistic regression model, a generalization of logistic model (binary response), can handle multiple category responses. At each combination of levels of the explanatory variables, the model assumes that the response counts for the categories of outcome variable have multinomial distribution. According to Hosmer and Lameshow (2000), the multinomial logistic model could be extended by any number of levels (or categories) of the outcome variable, but the details of the model would be most understandable if the outcome variable has three categories. This is because the generalization to more than three categories is a problem more of notation than of concept. Following Hosmer and Lameshow (2000), in this chapter we consider only the situation where the outcome variable has three levels.

4.2.1 The Model and Estimation of the Parameters

Let *Y* be a categorical response variable with three categories, codes as 1, 2, or 3. Since the outcome variable has three categories, we need two logit models as the logistic regression

model uses for a binary outcome variable which parameterizes in terms of the logit $Y = 1$ versus $Y = 0$. We assume there are p explanatory variables, $x = (x_1, x_2, ..., x_p)$, in the model.

The logit models for nominal responses pair each response category to a baseline category and the choice is arbitrary. If the last category is the baseline, then the baseline-category logits are

$$\ln\left[\frac{P(Y=1 \mid x)}{P(Y=3 \mid x)}\right] = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p = \beta_1' x$$

$$\ln\left[\frac{P(Y=2 \mid x)}{P(Y=3 \mid x)}\right] = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p = \beta_2' x$$

Under this model, the response probabilities are

$$P(Y=1 \mid x) = \frac{e^{\beta_1' x}}{1 + e^{\beta_1' x} + e^{\beta_2' x}}$$

$$P(Y=2 \mid x) = \frac{e^{\beta_2' x}}{1 + e^{\beta_1' x} + e^{\beta_2' x}}$$

$$P(Y=3 \mid x) = \frac{1}{1 + e^{\beta_1' x} + e^{\beta_2' x}}$$

with unknown parameters $\beta = (\beta_1, \beta_2)$

Now, we recode the outcome variables as the following

$Y_1 = 1, \quad Y_2 = 0, \quad Y_3 = 0$ for $Y = 1$

$Y_1 = 0, \quad Y_2 = 1, \quad Y_3 = 0$ for $Y = 2$

$Y_1 = 0, \quad Y_2 = 0, \quad Y_3 = 1$ for $Y = 3$

We note that no matter what value Y takes on, the sum of these variables is $\sum_{j=1}^{3} y_j = 1$.

The conditional likelihood function given the covariates for sample of *n* independent observations is

$$L(\beta) = \prod_{i=1}^{n} \left[\left(\frac{e^{\beta_1' x_i}}{1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}} \right)^{y_{1i}} \left(\frac{e^{\beta_2' x_i}}{1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}} \right)^{y_{2i}} \left(\frac{1}{1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}} \right)^{y_{3i}} \right] \quad (4.2.1)$$

Taking *log* on both sides we have,

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_{1i} \left(\beta_1' x_i - \ln(1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}) \right) + y_{2i} \left(\beta_2' x_i - \ln(1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}) \right) - y_{3i} \ln(1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}) \right]$$

or, $$\ell(\beta) = \sum_{i=1}^{n} \left[y_{1i}\, \beta_1' x_i + y_{2i}\, \beta_2' x_i - (y_{1i} + y_{2i} + y_{3i}) \ln(1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}) \right]$$

or, $$\ell(\beta) = \sum_{i=1}^{n} \left[y_{1i}\, \beta_1' x_i + y_{2i}\, \beta_2' x_i - \ln(1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}) \right] \text{ as } \sum_{j=1}^{3} y_{ji} = 1 \text{ for each } i,$$

The maximum likelihood estimators are obtained by taking the first partial derivatives of $\ell(\beta)$ with respect to each of the unknown parameters and setting these equations equal to zero. As nonlinear equations, we use similar iterative procedures like Newton-Raphson method. The Hessian matrix is calculated to obtain the estimator of the covariance matrix of the ML estimator, which is the inverse of the observed information matrix. Again, the estimates of the parameters and variance covariance matrix can be obtained by any standard statistical computer packages like SAS, SPSS, and R (nnet package).

4.2.2 Odds Ratio: Prospective Versus Case-Control Studies (When the outcome variable is more than two categories)

Prentice and Pyke (1979) showed that odds ratios are the same for both the prospective (cohort) and case-control studies when there are more than two categories of the outcome variable. Suppose that *k* mutually exclusive and exhaustive disease groups are defined and let $Y = i$ denote the development of the *i*th disease during the defined study period, and $Y = 0$ denote

the disease-free state at the end of the study period. Suppose that a regression vector $x = (x_1, x_2, ..., x_p)$ is to be related to disease incidence. The odds ratio for disease $Y = i$ for an individual with characteristics x, relative to that for an individual with some standard regression vector x_0 is

$$Odds\ Ratio_{(cohort)} = \frac{P(Y = i \mid x)/P(Y = 0 \mid x)}{P(Y = i \mid x_0)/P(Y = 0 \mid x_0)},\ i=1,2,\ldots,k \tag{4.2.2}$$

Let $P(Y)$ and $P(x)$ represent marginal probability functions or probability density functions in the population as a whole. We have

$$P(Y \mid x) = \frac{P(x \mid Y)P(Y)}{P(x)} \tag{4.2.3}$$

Substituting (4.2.3) in (4.2.2) we get,

$$Odds\ Ratio_{(cohort)} = \frac{\dfrac{P(x \mid Y = i)P(Y = i)/P(x)}{P(x \mid Y = 0)P(Y = 0)/P(x)}}{\dfrac{P(x_0 \mid Y = i)P(Y = i)/P(x_0)}{P(x_0 \mid Y = 0)P(Y = 0)/P(x_0)}}$$

$$= \frac{P(x \mid Y = i)/P(x_0 \mid Y = i)}{P(x \mid Y = 0)/P(x_0 \mid Y = 0)}$$

$$= Odds\ Ratio_{(case-control)}$$

4.2.3 Asymptotic Properties of Multinomial Logistic Regression Model

4.2.3.1 Consistency of ML estimators

In this section, we show the consistency of the ML estimators for the multinomial logistic regression model via standard Monte Carlo simulation. In this case, we consider the outcome variable Y is random and has three categories, that is, Y takes values coded as 1, 2, and 3. We

assume that there are four explanatory variables x_1, x_2, x_3, and x_4 in the model, where each of them is a vector and takes two possible values coded as 0 or 1. If we treat the last category of the outcome variable as the baseline, then the multinomial logistic regression model can be written as

$$\ln\left[\frac{P(Y=1\mid x_1,x_2,x_3,x_4)}{P(Y=3\mid x_1,x_2,x_3,x_4)}\right]=\beta_{01}+\beta_{11}x_1+\beta_{12}x_2+\beta_{13}x_3+\beta_{14}x_4$$

$$\ln\left[\frac{P(Y=2\mid x_1,x_2,x_3,x_4)}{P(Y=3\mid x_1,x_2,x_3,x_4)}\right]=\beta_{02}+\beta_{21}x_1+\beta_{22}x_2+\beta_{23}x_3+\beta_{24}x_4$$

Under these models, the response probabilities are

$$P(Y=1\mid x_1,x_2,x_3,x_4)=\frac{e^{\beta_{01}+\beta_{11}x_1+\beta_{12}x_2+\beta_{13}x_3+\beta_{14}x_4}}{1+e^{\beta_{01}+\beta_{11}x_1+\beta_{12}x_2+\beta_{13}x_3+\beta_{14}x_4}+e^{\beta_{02}+\beta_{21}x_1+\beta_{22}x_2+\beta_{23}x_3+\beta_{24}x_4}} \tag{4.2.4}$$

$$P(Y=2\mid x_1,x_2,x_3,x_4)=\frac{e^{\beta_{02}+\beta_{21}x_1+\beta_{22}x_2+\beta_{23}x_3+\beta_{24}x_4}}{1+e^{\beta_{01}+\beta_{11}x_1+\beta_{12}x_2+\beta_{13}x_3+\beta_{14}x_4}+e^{\beta_{02}+\beta_{21}x_1+\beta_{22}x_2+\beta_{23}x_3+\beta_{24}x_4}} \tag{4.2.5}$$

$$P(Y=3\mid x_1,x_2,x_3,x_4)=\frac{1}{1+e^{\beta_{01}+\beta_{11}x_1+\beta_{12}x_2+\beta_{13}x_3+\beta_{14}x_4}+e^{\beta_{02}+\beta_{21}x_1+\beta_{22}x_2+\beta_{23}x_3+\beta_{24}x_4}} \tag{4.2.6}$$

For the above models, we estimate the unknown parameters $\beta_{01},\beta_{11},\beta_{12},\beta_{13},\beta_{14},\beta_{02},\beta_{21},\beta_{22}$, β_{23} and β_{24}. The purpose is to show that if the number of observations $(y_i,x_{1i},x_{2i},x_{3i},x_{4i})$, $i=1,2,\ldots,n$ increases, then the estimates of the parameters converge to their true values. Now, we simulate the values of the outcome and explanatory variables. As the explanatory variables are fixed, the variables x_1,x_2, x_3, and x_4 are created based on the binomial distribution for arbitrary number of sample size. Once the variables x_1,x_2, x_3, and x_4 are in hand, we calculate probabilities for the outcome variable based on the above equations (4.2.4), (4.2.5), and (4.2.6).

These probabilities are used to simulate the data for *Y* from the multinomial distribution as *Y* exceeds more than two categories (actually, in this case it would be trinomial since *Y* has only there categories). The results of the simulation study are provided in the following table.

Table 4.2.1: Estimated parameter values and their standard errors using the multinomial logistic regression model for different sample sizes of 200, 500, and 1,000.

Estimated parameter	$n = 200$		$n = 500$		$n = 1{,}000$	
	Estimate	SE	Estimate	SE	Estimate	SE
$\hat{\beta}_{10}$	0.462	0.026	0.437	0.013	0.420	0.010
$\hat{\beta}_{11}$	0.884	0.027	0.804	0.013	0.811	0.009
$\hat{\beta}_{12}$	2.403	0.095	1.365	0.018	1.344	0.011
$\hat{\beta}_{13}$	-0.494	0.026	-0.497	0.0130	-0.505	0.010
$\hat{\beta}_{14}$	1.171	0.028	1.113	0.013	1.109	0.010
$\hat{\beta}_{20}$	1.263	0.024	1.235	0.012	1.231	0.010
$\hat{\beta}_{21}$	1.637	0.026	1.520	0.013	1.507	0.009
$\hat{\beta}_{22}$	2.012	0.095	0.967	0.017	0.942	0.012
$\hat{\beta}_{23}$	0.233	0.025	0.204	0.012	0.200	0.009
$\hat{\beta}_{24}$	-0.475	0.027	-0.506	0.013	-0.504	0.009

SE: Simulated standard error

For standard Monte Carlo simulation, we consider sample sizes of $n = 200$, 500, and 1,000. For the arbitrary fixed values of $\beta_{10} = 0.4$, $\beta_{11} = 0.8$, $\beta_{12} = 1.3$, $\beta_{13} = -0.5$, $\beta_{14} = 1.1$, $\beta_{20} = 1.2$, $\beta_{21} = 1.5$, $\beta_{22} = 0.9$, $\beta_{23} = 0.2$, and $\beta_{24} = -0.5$, we generate 1,000 independent sets of random samples for each different sample sizes. Then we estimate $\beta_{01}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{02}, \beta_{21}, \beta_{22}, \beta_{23}$, and β_{24} based on the average of 1,000 estimates of $\beta_{01}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{02}, \beta_{21}, \beta_{22}, \beta_{23}$, and β_{24}, which are estimated from the simultaneously fitted multinomial logistic regression model, and so are their standard errors for each estimated

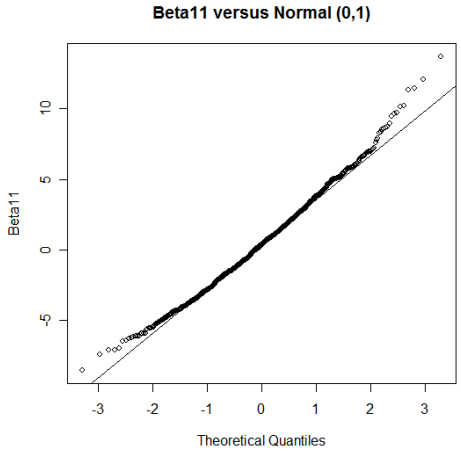
parameter. The results in Table 4.2.1, however, reveal that the estimated parameters converge to their true value when sample size increases, and also the simulated standard errors decrease with the increase of sample size.

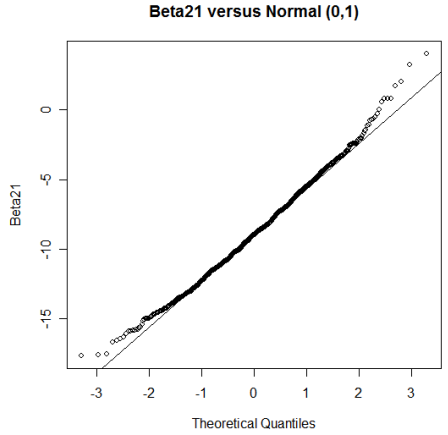
4.2.3.2 Normality of the ML estimators

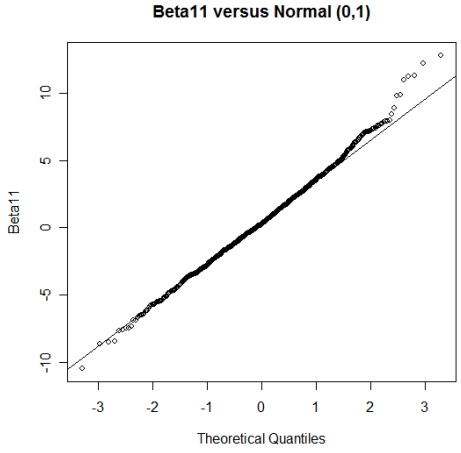
In this section, we show the large sample behavior of ML estimators of the parameters for the multinomial logistic regression model; that is, we show that the ML estimators of the parameters follow approximately normal distribution as sample size increases. This idea is similar to what we demonstrated in Chapter 2, Section 2.9.2. The result of the simulation study is provided below for the sample of sizes 750, 1,500, and 3,000. For each of the sample sizes, we replicate 1,000 times, and then results of the estimated parameters are provided through Q-Q plots. The result indicates that as the sample size increases, the parameters of multinomial distribution approximately follow normal distributions.

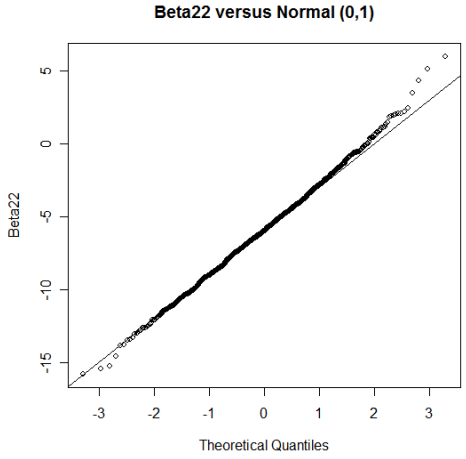
Figure 4.2.1: Monte Carlo simulation of finite sample behavior for normality of the parameters *(Simulation size = 1,000)*

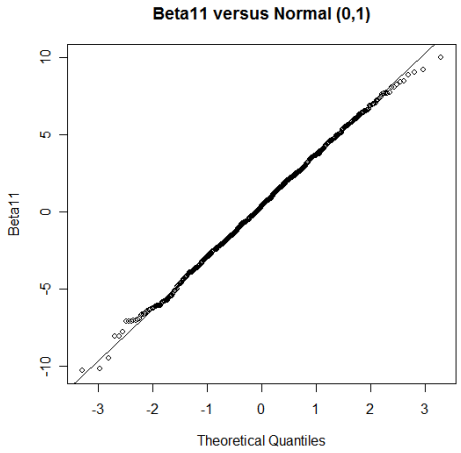
$(\hat{\beta}_{11})$ *Sample size: 750, 1,500, 3,000 respectively*	$(\hat{\beta}_{21})$ *Sample size: 750, 1,500, 3,000 respectively*











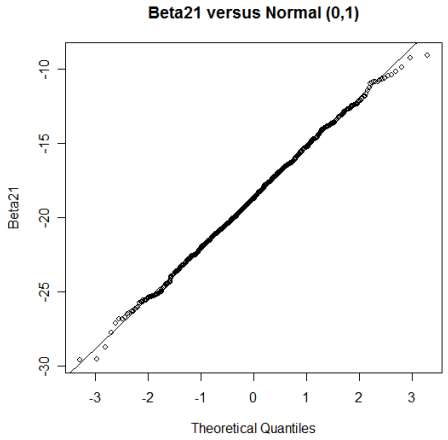


Figure 4.2.2: Monte Carlo simulation of finite sample behavior for normality of the parameters
(Simulation size = 1,000)

$(\hat{\beta}_{12})$ *Sample size: 750, 1,500, 3,000 respectively*	$(\hat{\beta}_{22})$ *Sample size: 750, 1,500, 3,000 respectively*

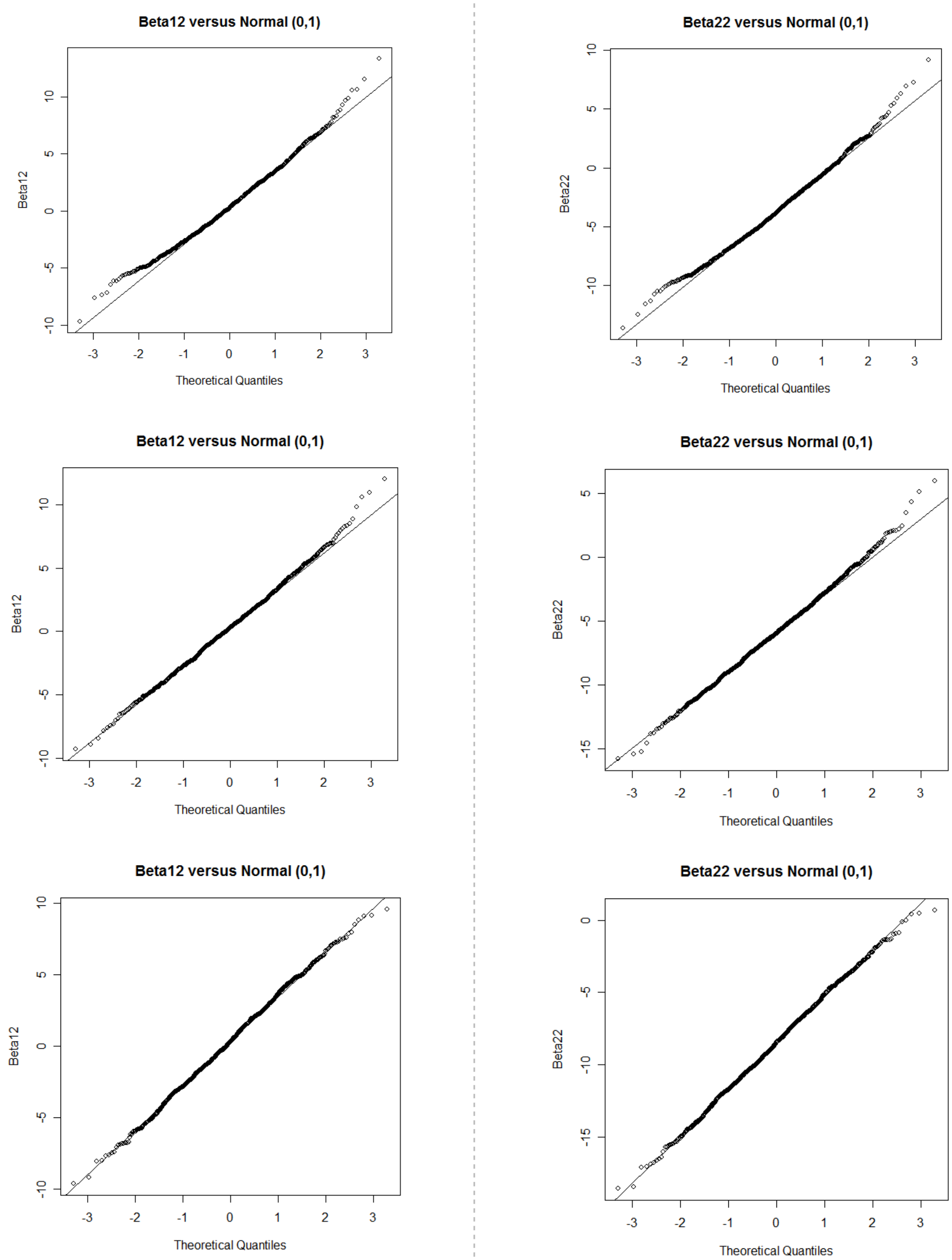


Figure 4.2.3: Monte Carlo simulation of finite sample behavior for normality of the parameters
(Simulation size = 1,000)

$(\hat{\beta}_{13})$ *Sample size: 750, 1,500, 3,000 respectively*	$(\hat{\beta}_{23})$ *Sample size: 750, 1,500, 3,000 respectively*

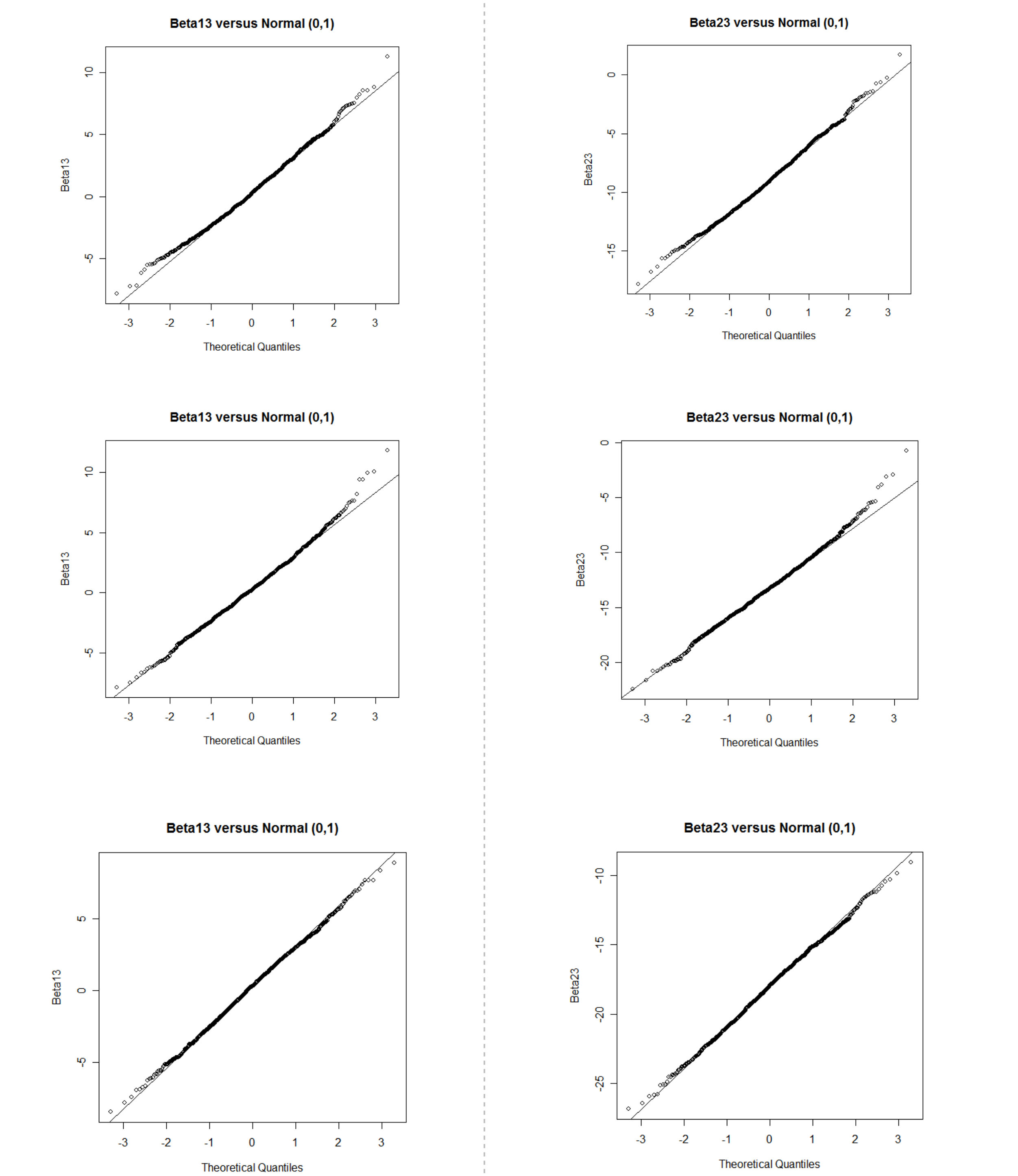
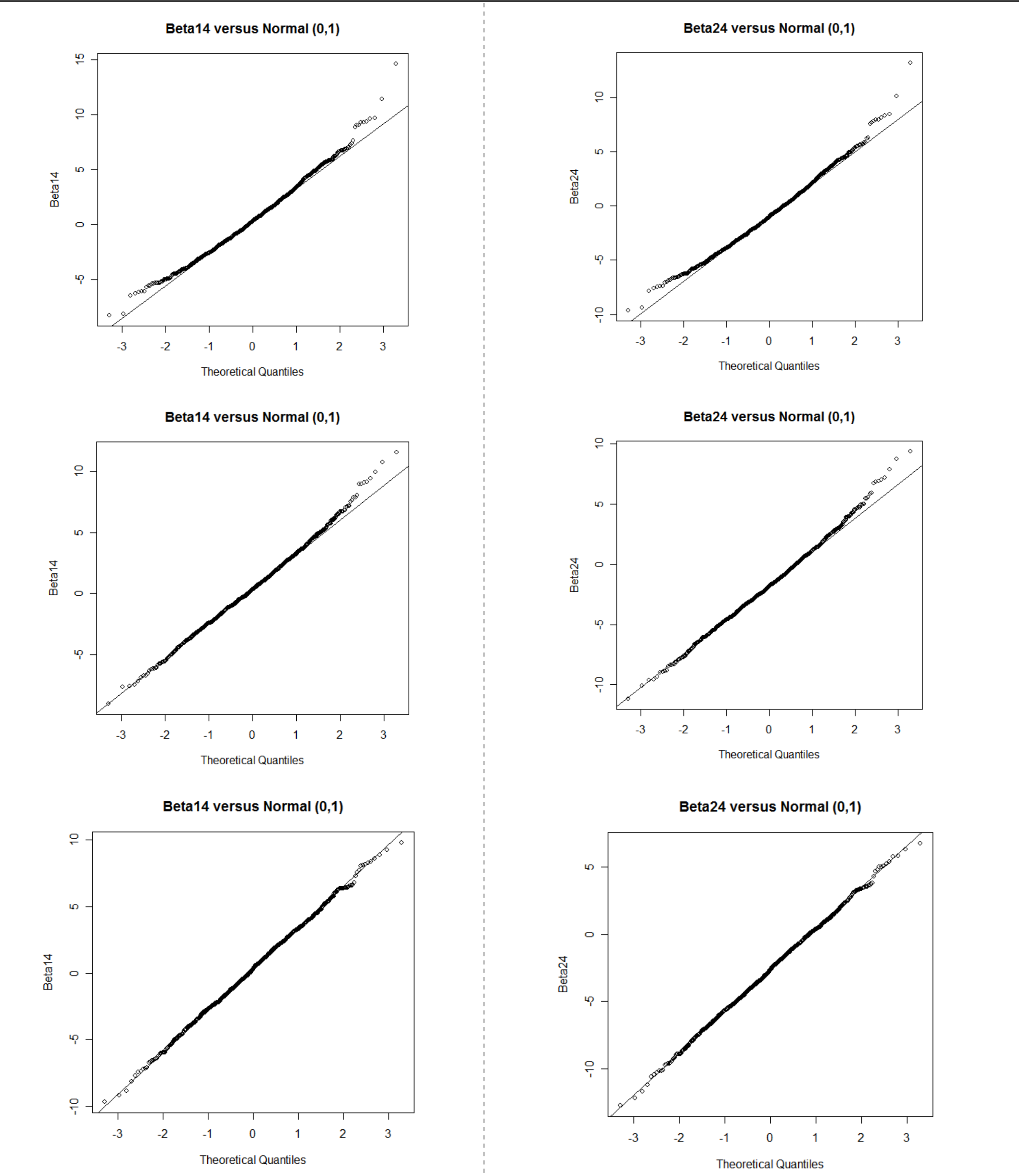


Figure 4.2.4: Monte Carlo simulation of finite sample behavior for normality of the parameters
(Simulation size = 1,000)

$(\hat{\beta}_{14})$ *Sample size: 750, 1,500, 3,000 respectively*	$(\hat{\beta}_{24})$ *Sample size: 750, 1,500, 3,000 respectively*



4.2.4 An Application Based on a Real Dataset

Rashid and Shifa (2007) conducted a study to investigate the important risk factors associated with women's unintended pregnancy. The data for this study were extracted from the Bangladesh Demographic and Health Survey (BDHS) conducted during 2004. This study considers women whose most recent pregnancy occurred five years preceding the date of the interview or who were currently pregnant. The pregnancy data were extracted using the question: Are you pregnant now? The answer was coded as either yes, no, or not sure. If she answered yes, she was considered as pregnant. Further, she was asked the question, "At the time of becoming pregnant, did you want this pregnancy now, or later, or did not want to have any (more) children at all?" The women who wanted the pregnancy 'now' were considered under the wanted group, the women who desired pregnancy 'later' were considered under the mistimed group, and the women who did not want to have any (more) children were considered under the unwanted group. The BDHS 2004 covered a nationally representative sample of 11,440 ever-married women from the ages of 10-49 years. The analysis is based on 5,817 women who had a pregnancy five years preceding the survey or who were currently pregnant. To analyze the data, the study considers pregnancy intention status (wanted, mistimed, and unwanted) as a response variable (Y), that is,

$$\text{Response variable, } Y: \begin{cases} 1 = Unwanted \\ 2 = Mistimed \\ 3 = Wanted \end{cases}$$

and a set of explanatory variables considered as risk factors of the pregnancy intentions which is given below.

$$\text{Explanatory variables, } X: \begin{cases} \text{age of respondents } (x_1) \\ \text{access to media } (x_2) \\ \text{education of respondents } (x_3) \\ \text{religion } (x_4) \\ \text{number of living children } (x_5) \\ \text{age at first marriage } (x_6) \\ \text{used modern method of FP prior to pregnancy } (x_7) \\ \text{respondent's working status } (x_8) \\ \text{wealth index } (x_9). \end{cases}$$

The mathematical form of the multinomial logistic regression model with three categories outcome variable and explanatory variables is

$$\ln\left[\frac{P(unwanted \mid x_1,\ldots,x_9)}{P(wanted \mid x_1,\ldots,x_9)}\right] = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{19}x_9$$

$$\ln\left[\frac{P(mistimed \mid x_1,\ldots,x_9)}{P(wanted \mid x_1,\ldots,x_9)}\right] = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{29}x_9$$

$$\ln\left[\frac{P(unwanted \mid x_1,\ldots,x_9)}{P(mistimed \mid x_1,\ldots,x_9)}\right] = \ln\left[\frac{P(unwanted \mid x_1,\ldots,x_9)/P(wanted \mid x_1,\ldots,x_9)}{P(mistimed \mid x_1,\ldots,x_9)/P(wanted \mid x_1,\ldots,x_9)}\right]$$

$$= \ln\left[\frac{P(unwanted \mid x_1,\ldots,x_9)}{P(wanted \mid x_1,\ldots,x_9)}\right] - \ln\left[\frac{P(mistimed \mid x_1,\ldots,x_9)}{P(wanted \mid x_1,\ldots,x_9)}\right]$$

$$= (\beta_{10} - \beta_{20}) + (\beta_{11} - \beta_{21})x_1 + \ldots + (\beta_{19} - \beta_{29})x_9$$

In the above models, $\beta' = (\beta_{i0}, \beta_{i1} \ldots, \beta_{i9}), i = 1,2$ is a vector of regression parameters corresponding to vector of covariates $X = (x_1,\ldots,x_9)$. The parameters of the models are estimated using standard statistical package SPSS. Results of the multivariate analysis are presented in the following table.

Table 4.2.2: Odds ratios from multinomial logistic regression model showing likelihood that a woman's pregnancy was unwanted or mistimed by selected characteristics, Bangladesh, 2004

Characteristics	Unwanted versus Wanted	Mistimed versus Wanted	Unwanted versus Mistimed
Age of respondents			
<19	0.04***	2.62***	0.02***
20-29	0.58***	1.92***	0.30***
30+	1.00	1.00	1.00
Access to media			
No	0.90	1.10	0.81*
Yes	1.00	1.00	1.00
Education of respondents			
No education	2.24**	0.88	2.55**
Primary	2.12**	1.05	2.02**
Secondary	2.20**	1.08	2.04**
Higher	1.00	1.00	1.00
Religion***			
Muslim	1.36*	1.41**	0.96
Non-Muslim	1.00	1.00	1.00
Number of living children***			
None	0.01***	1.32	0.01***
1-2	0.05***	1.33	0.03***
3-4	0.54***	1.63*	0.33***
5+	1.00	1.00	1.00
Age at first marriage***			
<15	1.86***	0.88	2.10**
15-19	1.65**	1.04	1.59
20+	1.00	1.00	1.00
Used modern method of FP prior to pregnancy***			
No	0.36***	0.70***	0.51***
Yes	1.00	1.00	1.00
Respondent's working status***			
No	0.80**	1.13	0.71**
Yes	1.00	1.00	1.00
Wealth index***			
Poorest	1.02	1.09	0.94
Poorer	1.04	0.97	1.06
Middle	1.16	1.27**	0.91
Richer	0.80	1.03	0.77
Richest	1.00	1.00	1.00

P-value: ***p<0.01, **p<0.05, *p<0.10

In the following paragraphs, we provide some explanations of the results in Table 4.2.2.

Unwanted vs. wanted: Compared to women age 30 and above, those women younger than 19 and those between the ages of 20 and 29 were about 96 and 42 percent less likely to say that the pregnancy was unwanted than wanted. The result indicates that women who had higher education had fewer tendencies for unwanted pregnancy; for example, the odds of women with less than higher education were 2 times more likely than women with higher education to say that their most recent birth or current pregnancy was unwanted as opposed to wanted. The practice of unwanted pregnancy was 36 percent higher for Muslim women compared to non-Muslim women. With regard to the number of living children, women having no child, having 1 to 2 children, having 3 to 4 children were 99 percent, 95 percent, 46 percent respectively were less likely than the women with more than 5 living children to report that a pregnancy was unwanted as opposed to wanted. Results showed that unwanted pregnancy was higher for those women who had married before 20 years of age. Women who never used modern contraception were 64 percent less likely to say that their most recent pregnancy was unwanted as opposed to wanted. Results also found that employed women had 20 percent higher unwanted pregnancies compared to women who were not employed.

Mistimed vs. wanted: The relationship between women's age and mistimed pregnancy, as opposed to wanted pregnancy, was negative. Compared to women age 30 and above, those women younger than 19 and those between the ages 20 and 29 were about 3 and 2 times more likely to say that the pregnancy was mistimed than wanted. Similar to findings in pattern showed that the previous paragraph, practice of unwanted pregnancy was 41 percent higher for Muslim women compared to non-Muslim women. Women with 3 to 4 living children were 63 percent more likely than the women with more than 5 living children to report that a pregnancy was

mistimed as opposed to wanted. Women who never used modern contraception were 30 percent less likely to say that their most recent pregnancy was mistimed as opposed to wanted.

Unwanted vs. mistimed: The relationship of women's age to planning status of the index birth is such that the youngest women (less than 19) were about 98 percent less likely than the oldest women to say that their most recent pregnancy was unwanted as opposed to mistimed. Unwanted pregnancy was 19 percent lower for women who did not have access to media compared to women who had. Education increased the odds that a pregnancy was unwanted rather than mistimed. For example, women who had primary or secondary education were about 2 times more likely than those who had higher education to have experienced an unwanted pregnancy rather than a mistimed pregnancy. With regard to the number of living children, women having no child, having 1 to 2 children, having 3 to 4 children were 99 percent, 97 percent, 67 percent were less likely than the women with more than 5 living children to report that a pregnancy was unwanted as opposed to mistimed. Women who never used modern contraception and were not employed were 49 percent and 29 percent respectively less likely to say that their most recent pregnancy was unwanted as opposed to mistimed.

4.3 The Multinomial Hybrid Logistic Model for Case-Control Study

In this section, we generalize the hybrid logistic model when the outcome variable is more than two categories. We assume that the case group has several different diseases, and the control group is disease free.

4.3.1 The Model

Consider the following table where the rare risk factor *z* has the zero event in the control group. In this table, we assume the outcome variable, $Y = i$, $i = 1,2$ treated as cases (two diseases) and $Y = 0$ treated as control.

Table 4.3.1: Cross-classification between the outcome variable (Y) and the factor, z

		$Y = i$, $i = 1,2$ (disease)		$Y = 0$ (disease free)
		1	2	3
Z	1	a	b	0
	0	c	d	e

** a-e indicate positive integers like Table 3.3.1*

Let $P(Z = 1 \mid Y = i, x, s = 1) = \alpha_i$

and $P(Z = 0 \mid Y = i, x, s = 1) = 1 - \alpha_i$, where α_i, $i = 1,2$ depend on *x*.

We propose the following model for the joint distribution of risk factors in the case-control study:

$$P(x, Z = z \mid Y = y, s = 1) = \alpha_1^{zy_1}(1-\alpha_1)^{(1-z)y_1}\alpha_2^{zy_2}(1-\alpha_2)^{(1-z)y_2}$$

$$\left(\frac{e^{\beta_1' x}}{1+e^{\beta_1' x}+e^{\beta_2' x}}\right)^{y_1}\left(\frac{e^{\beta_2' x}}{1+e^{\beta_1' x}+e^{\beta_2' x}}\right)^{y_2}\left(\frac{1-z}{1+e^{\beta_1' x}+e^{\beta_2' x}}\right)^{1-y_1-y_2}\cdot\frac{P(x \mid s=1)}{P(Y=y \mid s=1)} \qquad (4.3.1)$$

4.3.2 Estimation of the Parameters

To estimate the MLE of α and β, we consider the conditional likelihood function for a sample of *n* independent observations ($n = n_1 + n_0$, n_1 cases, n_0 controls) is proportional to

$$L(\beta) = \prod_{i=1}^{n} \Big[\alpha_1^{z y_{1i}} (1-\alpha_1)^{(1-z_i) y_{1i}} \alpha_2^{z_i y_{2i}} (1-\alpha_2)^{(1-z_i) y_{2i}} \left(\frac{e^{\beta_1' x_i}}{1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}} \right)^{y_{1i}} \left(\frac{e^{\beta_2' x_i}}{1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}} \right)^{y_{2i}} \left(\frac{1-z}{1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}} \right)^{1-y_{1i}-y_{2i}} \Big]$$

Taking ln on both sides, we have,

$$\ln L(\beta) = \sum_{i=1}^{n} \big[z_i y_{1i} \ln \alpha_1 + (1-z_i) y_{1i} \ln(1-\alpha_1) + z_i y_{2i} \ln \alpha_2 + (1-z_i) y_{2i} \ln(1-\alpha_2)$$

$$+ y_{1i} \beta_1' x_i - y_{1i} \ln(1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}) + y_{2i} \beta_{21}' x_i - y_{2i} \ln(1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}) + (1 - y_{1i} - y_{2i}) \ln(1 - z_i)$$

$$+ (1 - y_{1i} - y_{2i}) \ln(1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}) \big]$$

That is,

$$\ln L(\beta) = \sum_{i=1}^{n} \big[z_i y_{1i} \ln \alpha_1 + (1-z_i) y_{1i} \ln(1-\alpha_1) + z_i y_{2i} \ln \alpha_2 + (1-z_i) y_{2i} \ln(1-\alpha_2)$$

$$+ y_{1i} \beta_1' x_i + y_{2i} \beta_{21}' x_i + (1 - y_{1i} - y_{2i}) \ln(1 - z_i) - \ln(1 + e^{\beta_1' x_i} + e^{\beta_2' x_i}) \big]$$

The partial derivatives with respect to α_1 and α_2 respectively, we have,

$$\frac{\partial \ln L(\beta)}{\partial \alpha_1} = \sum_{i=1}^{n} \left[\frac{z_i y_{1i}}{\alpha_1} - \frac{(1-z_i) y_{1i}}{1-\alpha_1} \right]$$

$$\frac{\partial \ln L(\beta)}{\partial \alpha_2} = \sum_{i=1}^{n} \left[\frac{z_i y_{2i}}{\alpha_2} - \frac{(1-z_i) y_{2i}}{1-\alpha_2} \right]$$

Since y_{ji}, $j = 1,2$ depends on x, so α_i, i=1,2 depend on x. After setting each above equation equal to zero, we have

$$\alpha_1 = \frac{\sum_i z_i y_{1i}}{\sum_i y_{1i}} \text{ and } \alpha_2 = \frac{\sum_i z_i y_{2i}}{\sum_i y_{2i}}$$

Now, the model (4.3.1) can be written as,

$$P(x, Z = 1 \mid Y_1 = 1, s = 1) = \alpha_1 \cdot \frac{e^{\beta_1' x}}{1 + e^{\beta_1' x} + e^{\beta_2' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = 1 \mid s = 1)}$$

$$P(x, Z = 0 \mid Y_1 = 1, s = 1) = (1 - \alpha_1) \cdot \frac{e^{\beta_1' x}}{1 + e^{\beta_1' x} + e^{\beta_2' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = 1 \mid s = 1)}$$

$$P(x, Z = 1 \mid Y_2 = 1, s = 1) = \alpha_2 \cdot \frac{e^{\beta_2' x}}{1 + e^{\beta_1' x} + e^{\beta_2' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = 2 \mid s = 1)}$$

$$P(x, Z = 0 \mid Y_2 = 1, s = 1) = (1 - \alpha_2) \cdot \frac{e^{\beta_2' x}}{1 + e^{\beta_1' x} + e^{\beta_2' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = 2 \mid s = 1)}$$

$$P(x, Z = 1 \mid Y_3 = 1, s = 1) = 0$$

$$P(x, Z = 0 \mid Y_3 = 1, s = 1) = \frac{1}{1 + e^{\beta_1' x} + e^{\beta_2' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = 3 \mid s = 1)}$$

The remaining parameters involved in the model (4.3.1) can be obtained by pairwise combining the above, we get

$$P(x \mid Y_1 = 1, s = 1) = \frac{e^{\beta_1' x}}{1 + e^{\beta_1' x} + e^{\beta_2' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = 1 \mid s = 1)}$$

$$P(x \mid Y_2 = 1, s = 1) = \frac{e^{\beta_2' x}}{1 + e^{\beta_1' x} + e^{\beta_2' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = 2 \mid s = 1)}$$

$$P(x \mid Y_3 = 1, s = 1) = \frac{1}{1 + e^{\beta_1' x} + e^{\beta_2' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = 3 \mid s = 1)}$$

The likelihood function is set as a regular multinomial regression, and thus, we get the estimates of β's. Hence, the odds ratios can be estimated.

CHAPTER 5

VARIANCE ESTIMATION IN LOGISTIC REGRESSION MODEL USING THE BOOTSRAP

5.1 The Bootstrap Method

5.1.1 The Basic Idea

The use of resampling techniques plays a central role in statistics, especially when the estimators of interest do not have an explicit formula. A very general resampling technique, called the bootstrap method, was introduced by Efron (1979) for estimating unknown quantities associated with the statistical models. The bootstrap method is often used to find standard errors for estimators, confidence intervals for unknown parameters, or *p* values for test statistics under a null hypothesis (Boos, 2003). This method consists of approximating the distribution of a function of the observations based on independent observations. Freedman et al. (1981) state that "this distribution is obtained by replacing the unknown distribution by the empirical distribution of the data in the definition of the statistical function, and then resampling the data to obtain a Monte Carlo distribution for the resulting random variable." A formal description of the bootstrap method follows.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size *n* from a population with distribution *F* and let $\tau(X_1, X_2, \ldots, X_n; F)$ be the specified random variable of interest, possibly depending upon the unknown distribution *F*. Let F_n denote the empirical distribution function (EDF) of $X_1, X_2, \ldots, X_n$, that is, the distribution that puts mass $1/n$ at each of the points of $X_1, X_2, \ldots, X_n$. The bootstrap method is to approximate the distribution of $\tau(X_1, X_2, \ldots, X_n; F)$ under *F* by that

of $\tau\,(X_1^*, X_2^*,,\ldots, X_n^*;\ F_n)$ under F_n, where $X_1^*, X_2^*,,\ldots, X_n^*$ denotes a random sample of size *n* from F_n.

Usually, the distribution of $\tau\,(X_1^*, X_2^*,,\ldots, X_n^*;\ F_n)$ under F_n cannot be evaluated analytically. It can, however, be estimated with arbitrary accuracy by carrying out a Monte Carlo simulation in which random samples are drawn from F_n. In fact, the bootstrap is usually implemented by the Monte Carlo simulation study. The Monte Carlo procedure for estimating the distribution of $\tau\,(X_1, X_2,\ldots, X_n; F_n)$ is as follows

Step 1: Generate a bootstrap sample of size *n*, say, $\left(X_1^*, X_2^*,\ldots, X_n^*\right)$ from F_n randomly with replacement.

Step 2: Compute $\tau(X_1^*, X_2^*,\ldots, X_n^*)$

Step 3: Use the results of many repetitions, say, *B* times of steps 1 and 2 to construct the bootstrap EDF of $\hat{T} = \tau\,(X_1^*, X_2^*,,\ldots, X_n^*;\ F_n)$. Suppose that the sequence $(\hat{T}^{(1)}, \hat{T}^{(2)},\ldots,\hat{T}^{(B)})$ represents the set of estimates obtained by repeating steps (1) and (2) and then the bootstrap EDF can be obtained by $\hat{G}(t) = \dfrac{\#\left\{\hat{T}^{(b)} \le t\right\}}{B}$, $b = 1,2,\ldots, B,$ where $\hat{G}(\cdot)$ is a bootstrap empirical distribution function and *t* is some specific value of $\hat{T}$.

Based on $\hat{T}^{(1)}, \hat{T}^{(2)},\ldots,\hat{T}^{(B)}$, the bootstrap estimate of $\hat{T} = \tau\,(X_1^*, X_2^*,,\ldots, X_n^*;\ F_n)$ is defined as the average of the *B* bootstrap estimates:

$$\hat{T}^{(boot)} = \frac{1}{B}\sum_{b=1}^{B}\hat{T}^{(b)} \qquad (5.1.1)$$

and the variance of the bootstrap estimate $\hat{T}^{(boot)}$ is

$$\hat{V}^{(boot)} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{T}^{(b)} - \hat{T}^{(boot)}\right)^2 \tag{5.1.2}$$

Large-sample distribution can be derived from the bootstrap distribution of $\hat{T}^{(1)}, \hat{T}^{(2)}, \ldots, \hat{T}^{(B)}$, if the bootstrap distribution is approximately normal, a $100(1-\alpha)\%$ bootstrap confidence interval for a scalar $T = \tau(X_1, X_2, \ldots, X_n; F)$ can be computed as

$$CI_{norm}(T) = \hat{T}^{(boot)} \pm z_{1-\alpha/2} \cdot \sqrt{\hat{V}^{(boot)}} \tag{5.1.3}$$

Alternatively, if the bootstrap distribution is non-normal, a $100(1-\alpha)\%$ bootstrap confidence interval can be computed empirically as

$$CI_{emp}(T) = (\hat{T}^{(b,l)}, \hat{T}^{(b,u)}) \tag{5.1.4}$$

where $\hat{T}^{(b,l)}$ and $\hat{T}^{(b,u)}$ are the $\alpha/2$ and $1-\alpha/2$ percentiles of the empirical bootstrap distribution of *T*. Stable intervals based on equation (5.1.3) requires bootstrap sample of the order of *B* = 200 and equation (5.1.4) requires larger samples, for example, B = 2000 or more (Efron, 1994).

5.1.2 The Bootstrap Consistency

Most of the mathematical results regarding the bootstrap describe how it performs as the sample size increases. Asymptotic theory of the Efron's bootstrap were discussed by many authors; see, for example, Bickel and Freedman (1981), Singh (1981), Beran and Dunharme (1991), and Mammen (1992), among others. We now discuss the consistency of the bootstrap.

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with distribution function F. Assume that *F* has finite mean μ and variance σ^2, both unknown. Consider the parameter function

$$\tau(F) = \int h(x)\, dF(x) = E_F\left[h(X)\right]$$

Now, the plug-in estimate of an expectation $E_F\left[h(X)\right]$ is

$$\hat{\tau}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i) = \tau(F_n)$$

(see, for example, Efron and Tibshirani, 1993), where F_n is the empirical distribution of the $X_i's$. The empirical distribution function (EDF) is defined by

$$F_n(X) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x).$$

Let $G_n(F, y)$ be the distribution function for $\hat{\tau}_n - \tau(F)$, that is,

$$G_n(F, y) = P_F(\hat{\tau}_n - \tau(F) \le y)$$

Now, if $\hat{\tau}_n - \tau$ is a pivot, then $G_n(F, \cdot)$ does not depend on *F* at all. But as strict pivotness is too much to hope for, it is sensible to examine just to see how much G_n varies with *F*. Now, we examine this variability in terms of the uniform distance between the distribution functions,

$$\left\| G_n(F, \cdot) - G_n(F', \cdot) \right\| = \sup_{y \in \Re} \left| G_n(F, y) - G_n(F', y) \right|$$

Theorem 5.1.1. (Berry-Esseen): Let $X_1, X_2, \ldots$ be independent and identically distributed real stochastic variables with

$$E(X_i) = 0, \quad V(X_i) = 1, \quad E\left|X\right|^3 < \infty.$$

Let G_n be the distribution function for $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$ and Φ be the distribution function for the standard normal distribution. It holds that

$$\left\| G_n - \Phi \right\| \le C_{BE} \frac{E\left|X_i\right|^3}{\sqrt{n}}$$

for a global constant C_{BE} (Berry-Esseen constant).

See, for example, Durret (2004) for details of the proof.

Theorem 5.1.2. It holds that

$$\| G_n(F, \cdot) - G_n(F', \cdot) \| \leq C_{BE} \left(\frac{\gamma(F)}{\sigma(F)^2} + \frac{\gamma(F')}{\sigma(F')^2} \right) \frac{1}{\sqrt{n}} + \left\| \Phi\left(\frac{\sigma(F')}{\sigma(F)} y \right) - \Phi(y) \right\| \quad (5.1.5)$$

where Φ is the distribution function for the standard normal distribution, where

$$\sigma(F)^2 = V_F(X_i),$$

$$\sigma(F')^2 = V_{F'}(X_i),$$

$$\gamma(F) = E_F \left| X_i - \tau(F) \right|^3,$$

and $$\gamma(F') = E_{F'} \left| X_i - \tau(F') \right|^3$$

Proof: Standard use of the triangle inequality shows that

$$\| G_n(F, \cdot) - G_n(F', \cdot) \| \leq \left\| G_n(F, y) - \Phi\left(\sqrt{\frac{n}{\sigma(F)^2}}\, y \right) \right\| + \left\| \Phi\left(\sqrt{\frac{n}{\sigma(F)^2}}\, y \right) - \Phi\left(\sqrt{\frac{n}{\sigma(F')^2}}\, y \right) \right\|$$

$$+ \left\| G_n(F', y) - \Phi\left(\sqrt{\frac{n}{\sigma(F')^2}}\, y \right) \right\|$$

In this case, we are hoping that the specific choice of intermediate normal distributions gives rise to sensible estimates, since the central limit theorem (CLT) implies that

$$\hat{\tau}_n - \tau(F) \overset{as}{\sim} N\left(0, \frac{\sigma(F)^2}{n} \right)$$

and as $y \mapsto \Phi\left(\sqrt{\frac{n}{\sigma(F)^2}}\, y\right)$ is the distribution function for this approximating normal

distribution. Using the fact

$$\sup_{y\in\Re} \left| G(y) \right| = \sup_{y\in\Re} \left| G(a\, y) \right|$$

for any function $G : \Re \to \Re$ and any $a \neq 0$, we transform the inequality to

$$\begin{aligned} \| G_n(F, \cdot) - G_n(F', \cdot) \| \;\leq\; & \left\| G_n\left(F, \sqrt{\frac{\sigma(F)^2}{n}}\, y\right) - \Phi(y) \right\| + \left\| \Phi\left(\sqrt{\frac{\sigma(F')^2}{\sigma(F)^2}}\, y\right) - \Phi(y) \right\| \\ & + \left\| G_n\left(F', \sqrt{\frac{\sigma(F')^2}{n}}\, y\right) - \Phi(y) \right\| \end{aligned}$$

$$\begin{aligned} = \; & \left\| P_F\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \tau(F) \leq, \sqrt{\frac{\sigma(F)^2}{n}}\, y\right) - \Phi(y) \right\| + \left\| \Phi\left(\sqrt{\frac{\sigma(F')^2}{\sigma(F)^2}}\, y\right) - \Phi(y) \right\| \\ & + \left\| P_{F'}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \tau(F') \leq, \sqrt{\frac{\sigma(F')^2}{n}}\, y\right) - \Phi(y) \right\| \end{aligned}$$

$$\begin{aligned} = \; & \left\| P_F\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{X_i - \tau(F)}{\sigma(F)} \leq y\right) - \Phi(y) \right\| + \left\| \Phi\left(\sqrt{\frac{\sigma(F')^2}{\sigma(F)^2}}\, y\right) - \Phi(y) \right\| \\ & + \left\| P_{F'}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{X_i - \tau(F')}{\sigma(F')} \leq y\right) - \Phi(y) \right\| \end{aligned}$$

As the variables $\frac{X_i - \tau(F)}{\sigma(F)}$ are iid with mean zero and variance 1 under P_F and so does

$\frac{X_i - \tau(F')}{\sigma(F')}$ under $P_{F'}$, it follows from Berry-Esseen theorem that

$$\| G_n(F,\cdot) - G_n(F',\cdot) \| \leq C_{BE}\, E_F \left| \frac{X_i - \tau(F)}{\sigma(F)} \right|^3 \frac{1}{\sqrt{n}} + \left\| \Phi\left(\sqrt{\frac{\sigma(F')^2}{\sigma(F)^2}}\, y \right) - \Phi(y) \right\|$$

$$+ C_{BE}\, E_{F'} \left| \frac{X_i - \tau(F')}{\sigma(F')} \right|^3 \frac{1}{\sqrt{n}}$$

That is,

$$\| G_n(F,\cdot) - G_n(F',\cdot) \| \leq C_{BE} \left(\frac{\gamma(F)}{\sigma(F)^3} + \frac{\gamma(F')}{\sigma(F')^3} \right) \frac{1}{\sqrt{n}} + \left\| \Phi\left(\frac{\sigma(F')}{\sigma(F)}\, y \right) - \Phi(y) \right\|$$

Hence, the proof follows.

The distance estimate (5.1.5) holds for any two measures F and F', as long as the relevant moments exists. If we plug in the true measure F and the empirical measure F_n based on the observations of $X_1, X_2, \ldots, X_n$, we obtain the distance between the distribution function we would like to know and the distribution function that comes from bootstrapping (in the limit of infinitely many replications). We can bind this distance as

$$\| G_n(F,\cdot) - G_n(F_n,\cdot) \| \leq C_{BE} \left(\frac{\gamma(F)}{\sigma(F)^3} + \frac{\gamma_n}{\sigma_n{}^3} \right) \frac{1}{\sqrt{n}} + \left\| \Phi\left(\frac{\sigma_n}{\sigma(F)}\, y \right) - \Phi(y) \right\| \qquad (5.1.6)$$

where $\tau_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \tau_n)^2$, $\gamma_n = \frac{1}{n}\sum_{i=1}^{n} \left| X_i - \tau_n \right|^3$

The left-hand side of the equation (5.1.6) is the distance between a deterministic-but unknown-sequence of distribution functions and a random-but observable-sequence of distribution functions. It is random in the sense that it depends on the observations $X_1, X_2, \ldots, X_n$. Note also that the bound on the right hand side is random.

Theorem 5.1.3. (Bootstrap consistency) It holds that

$$\| G_n(F, \cdot) - G_n(F_n, \cdot) \| \rightarrow 0 \qquad \text{for } n \rightarrow \infty \text{ almost surely}$$

Proof: By the law of large numbers we have that

$$\tau_n \rightarrow \tau(F), \ \sigma_n^2 \rightarrow \sigma(F)^2, \ \gamma_n \rightarrow \gamma(F) \qquad \text{for } n \rightarrow \infty \text{ almost surely}$$

This implies that

$$C_{BE} \left(\frac{\gamma(F)}{\sigma(F)^3} + \frac{\gamma_n}{\sigma_n^3} \right) \frac{1}{\sqrt{n}} \rightarrow 0 \text{ for } n \rightarrow \infty \text{ almost surely}$$

It also implies that

$$\frac{\sigma_n}{\sigma(F)} \rightarrow 1 \quad \text{for } n \rightarrow \infty \text{ almost surely}$$

So the Theorem 5.1.3 will follow if we can show that for any sequence of scalars $(\lambda_n)_{n \in \mathrm{N}}$ with the property that $\lambda_n \rightarrow 1$ for $n \rightarrow \infty$, it holds that

$$\| \Phi(\lambda_n y) - \Phi(y) \| \rightarrow 0 \qquad \text{for } n \rightarrow \infty$$

Recall Scheffés lemma: if f_n is a sequence of probability densities, eg. with respect to *m*, if f is a probability density and if $f_n \rightarrow f$ almost surely, then $f_n \rightarrow f$ in L^1.

We observe that $y \mapsto \Phi(\lambda_n y)$ is the distribution function for the normal distribution with mean 0 and variance $\frac{1}{\lambda_n^2}$. Let ϕ_n be the density for this normal distribution, and ϕ be the density for the standard normal distribution. As

$$\phi_n(y) = \lambda_n \phi_n(\lambda_n y)$$

and as ϕ is continuous, we see that $\phi_n(y) \rightarrow \phi(y)$ for $n \rightarrow \infty$ for every *y*. Hence $\phi_n \rightarrow \phi$ in L^1 according to the Scheffés lemma. And thus

$$\| \Phi(\lambda_n y) - \Phi(y) \| = \sup_{y \in \Re} \left| \int_{-\infty}^{y} \phi_n(y)\, dy - \int_{-\infty}^{y} \phi(y)\, dy \right| \le \int_{-\infty}^{\infty} | \phi_n(y) - \phi(y) |\, dy \quad \to \quad 0 .$$

Therefore,

$$\| G_n(F, \cdot) - G_n(F_n, \cdot) \| \quad \to \quad 0 \qquad \text{for } n \to \infty \text{ almost surely.}$$

Hence the proof follows.

5.1.3 An Example

Suppose a random sample $X_1, X_2, \ldots, X_n$ of size *n* is observed from a completely unspecified probability distribution *F*. Assume *F* has finite mean μ and variance σ^2, both unknown. The sample average $\overline{X}$ and sample standard deviation s^2 are the conventional estimate for μ and variance σ^2 respectively. By the Central Limit Theorem, the distribution of the pivotal quantity

$$Q = \frac{\sqrt{n}\ (\overline{X} - \mu)}{s}$$

tends weakly to $N(0,1)$. So the asymptotics are known in this situation.

Let F_n be the empirical distribution of $X_1, X_2, \ldots, X_n$, putting mass $1/n$ on each X_i, $i = 1,\ldots,n$. Now, we draw a random sample $X_1^*, X_2^*, \ldots, X_n^*$, called bootstrap sample, of size *n* from F_n with replacement and estimate the distribution of the bootstrap pivotal quantity

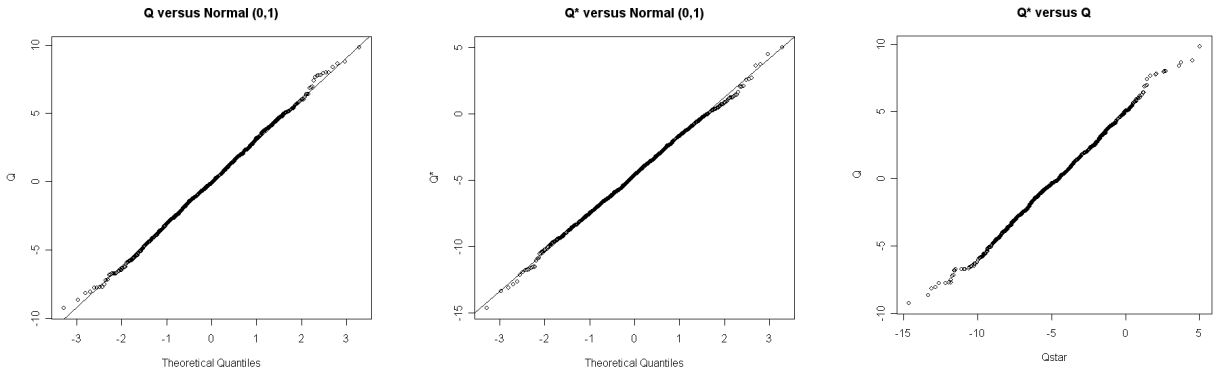
$$Q^* = \frac{\sqrt{n}\ (\overline{X}^* - \overline{X})}{s^*}, \text{ where } \overline{X}^* = \frac{1}{n}\sum_{i=1}^{n} X_i^* \text{ and } s^{*2} = \frac{1}{n-1}\sum_{i=1}^{n} (X_i^* - \overline{X}^*)^2 .$$

In the bootstrap technique, the random sample $X_1, X_2, \ldots, X_n$ are treated as a population, with distribution function F_n and mean $\overline{X}$; and $\overline{X}^*$ is considered as an estimator of $\overline{X}$. The idea is that the behavior of the bootstrap pivotal quantity Q^* mimics that of Q. Thus, the distribution of

Q^* could be computed from the data and used to approximate the unknown distribution of Q. In other words, the bootstrap distribution of $\sqrt{n}(\overline{X}^* - \overline{X})$ could be used to approximate the sampling distribution of $\sqrt{n}(\overline{X} - \mu)$.

A Monte Carlo evaluation is performed of the above ideas. The quantile-normal graphs of Q and Q^* are provided below. It is clear from the graphs (a) and (b) that they follow normal distribution. The Q-Q plot of Q and Q^* is given in the graphs (c) which allows to compare two sample distributions with one another. As most of the values of Q and Q^* fall on a straight line, it can be said that the two data sets have the same parent distribution. Therefore, it is concluded that the bootstrap method is used for estimating the distribution of an estimator by resampling one's data.

Figure 5.1.1: Q-Q plot of Q and Q^* for sample size= 50 and Monte Carlo sample size= 1,000



(a) (b) (c)

5.2 Bootstrapping a Linear Regression Model

In the linear regression model, $Y = (y_1, y_2, \ldots, y_n)^T$ denotes the $n \times 1$ vector of the response, and $X = (x_1, x_2, \ldots, x_n)^T$ is the matrix of regressors with $n \times p$ dimension including the intercept, p is the number of parameters. The usual linear regression model is then

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \text{ or } y_i = x_i^T \beta + \varepsilon_i \ , i = 1, 2, \ldots, n$$

where, ε is an $n \times 1$ vector of uncorrelated error terms having mean zero and identical variance σ^2, usually unknown. The $p \times 1$ vector β holds the unknown parameters, for which the ordinary least squares (OLS) estimator is $\hat{\beta} = (X^T X)^{-1} X^T Y$ and has variance-covariance matrix $\sigma^2 (X^T X)^{-1}$.

Theorem 5.2.1. Under the linear regression model, if the *Y*-vector is treated to be the observed value of the random vector $X\beta + \varepsilon$, then $E(\hat{\beta}) = \beta$ and $Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$. Suppose that $\frac{1}{n} X'X \to V$, which is positive definite and also suppose that the elements of X are uniformly small by comparison with $\sqrt{n}$, then

$$\sqrt{n}(\hat{\beta} - \beta) \overset{asympt.}{\sim} N\left(0, \sigma^2 V^{-1}\right)$$

and the distribution of the pivotal quantity

$$\frac{(X'X)^{1/2}(\hat{\beta} - \beta)}{\sigma} \overset{asympt.}{\sim} N\left(0, I\right)$$

where, I is the $p \times p$ identity matrix (see, for details, Freedman, 1981).

Traditional approaches, like ordinary least squares, rely very much on some major modeling assumptions, for example, normal random errors with constant variances. But for generalizations to non-normal errors and non-constant variance, exact methods rarely exist, and

we are faced with approximate methods based on linear approximations to estimators and central limit theorem. As a result, the ordinary sampling techniques use some assumptions related to the form of the estimator distribution, but resampling methods do not need these assumptions because the sample is thought as population. Therefore, resampling methods have the potential to provide more accurate analysis. See, references for bootstrapping regression, Efron and Tibshirani (1993), Davison and Hinkley (1997), Wu (1986), Freedman (1981), Stine (1985), and Peters and Freedman (1984).

There are two approaches for bootstrapping the regression model, and the choice of either methods depends upon the regressors being fixed or random. If the regressors are fixed, the bootstrap uses resampling of the error term. If the regressors are random, the bootstrap uses resampling of observations (Stine, 1989).

(a) Bootstrap based on the resampling observations (or vector resampling)

This approach is usually applied when the regression models built from data have regressors that are as random as the response. Let the $(p+1)\times 1$ vector $z_i = (y_i, x_i^T)^T$ denote the values associated with *i*th observation and assume that z_i's are drawn independently and identically from a distribution of *F*. In this case, the set of observations are the vectors $(z_1, z_2, \ldots, z_n)$. The bootstrap procedure based on the resampling observations is as follows.

1. Draw a *n* sized bootstrap sample $(z_1^*, z_2^*, \ldots, z_n^*)$ from the observations with replacement giving 1/*n* probability each z_i values and label the elements of each vector

   $z_i^* = (y_i^*, x_i^{*T})^T$, $i =$ 1, 2, …, *n,*

   and then form the vector $Y^* = (y_1^*, y_2^*, \ldots, y_n^*)^T$ and the matrix $X^* = (x_1^*, x_2^*, \ldots, x_n^*)^T$.
2. Calculate the OLS coefficients from the bootstrap sample

$$\hat{\beta}^* = (X^{*T} X^*)^{-1} X^{*T} Y^*$$

3. Repeat steps 1 and 2 for *B* times, where *B* is the number of repetition and then use the resulting bootstrap estimates $\hat{\beta}^{*(1)}, \hat{\beta}^{*(2)}, \ldots, \hat{\beta}^{*(B)}$ to estimate variances or confidence intervals. The bootstrap estimate of the covariance matrix of $\hat{\beta}$ is

$$Var(\hat{\beta}^*) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}_b^* - \overline{\hat{\beta}}^* \right)^2, \text{ where } \overline{\hat{\beta}}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_b^*$$

(b) Bootstrap based on the resampling errors (or residual resampling)

The bootstrap procedure based on the resampling errors is as follows.

1. Fit the least squares regression equation for full sample to obtain the fitted responses $\hat{y}_i$ and residuals $\hat{\varepsilon}_i$, where $\hat{\varepsilon}_i = y_i - \hat{y}_i$.
2. Draw a *n* sized bootstrap set of residuals $\{\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, \ldots, \hat{\varepsilon}_n^*\}$ completely at random with replacement from the set of fitted residuals $\{\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n\}$, giving 1/*n* probability each $\hat{\varepsilon}_i$ values.
3. Create a bootstrap set of pseudo-responses, $Y^* = X\hat{\beta} + \hat{\varepsilon}^*$, where $\hat{\varepsilon}^* = (\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, \ldots, \hat{\varepsilon}_n^*)^T$ is the $n \times 1$ vector.
4. Regress Y^* on X to obtain a bootstrap parameter estimate by

$$\hat{\beta}^* = (X^T X)^{-1} X^T Y^*$$

5. Repeat steps 2-4 for *B* times, where *B* is the number of repetition and use the resulting bootstrap estimates $\hat{\beta}^{*(1)}, \hat{\beta}^{*(2)}, \ldots, \hat{\beta}^{*(B)}$ to estimate variances or confidence intervals. The bootstrap estimate of the covariance matrix of $\hat{\beta}$ is

$$Var(\hat{\beta}^*) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_b^* - \bar{\hat{\beta}}^*\right)^2, \text{ where } \bar{\hat{\beta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b^*$$

It can be shown that the bootstrap will give the same asymptotic results as the classical methods if the simulations are performed. Freedman (1981) discussed the asymptotic theory for bootstrapping multiple regression models. The theoretical results of the asymptotic properties are summarized by the following theorem.

Theorem 5.2.2. Assume the linear regression model defined above with assumptions given by Theorem 6.1. Along almost all sample sequences, given $Y_1, Y_2, \ldots, Y_n$, as *n* tends to ∞,

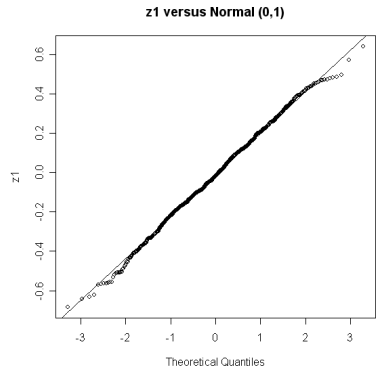
a) the conditional distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ converges weakly to normal with mean 0 and variance-covariance matrix $\sigma^2 V^{-1}$.

b) the conditional distribution of $\hat{\sigma}^*$ converges to point mass at σ.

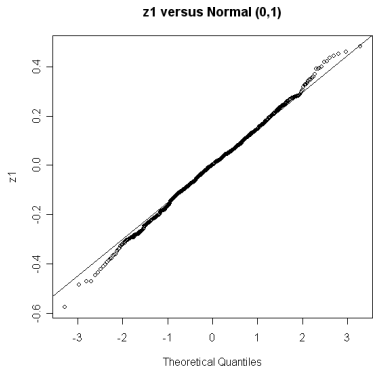
c) the conditional distribution of the pivot $\frac{(X^{*'}X^*)^{1/2}(\hat{\beta}^* - \beta^*)}{\hat{\sigma}^*}$ converges to standard normal in $\Re^p$.

To verify the above theorem with the bootstrap method, a Monte Carlo simulation study is carried out. We start with the sample of size 50 and then we draw a bootstrap sample of the same size with replacement. Using the regression model, we calculate the quantities $z_1 = \sqrt{n}(\hat{\beta}^* - \hat{\beta})$ and $z_2 = \frac{(X^{*'}X^*)^{1/2}(\hat{\beta}^* - \beta^*)}{\hat{\sigma}^*}$, and then we replicate the procedure for 1000 times. To see whether the calculated statistics follow normal distribution, we represent the results by the following Q-Q plots.

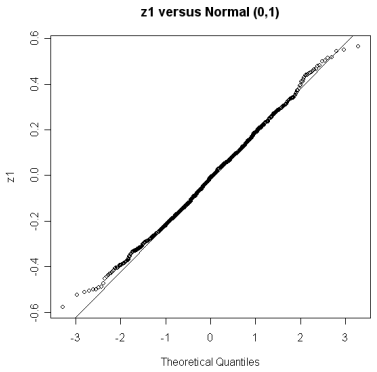
Figure 5.2.1: Q-Q plot of the quantity $z_1 = \sqrt{n}(\hat{\beta}^* - \hat{\beta})$ for different sample sizes



Sample size = 50
Mean =-0.01328 (~ 0)
Variance = 0.04679 (~ 0.03394)

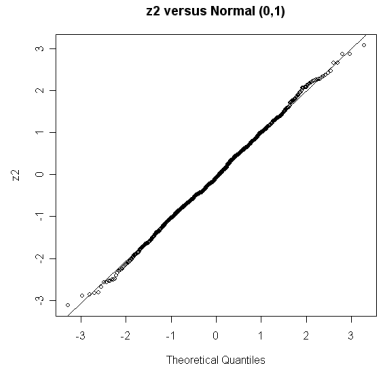


Sample size = 100
Mean = -0.00443 (~ 0)
Variance = 0.02510 (~ 0.03394)

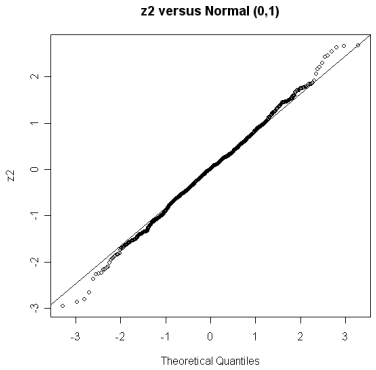


Sample size = 500
Mean = -0.01433 (~ 0)
Variance = 0.03883 (~ 0.03394)

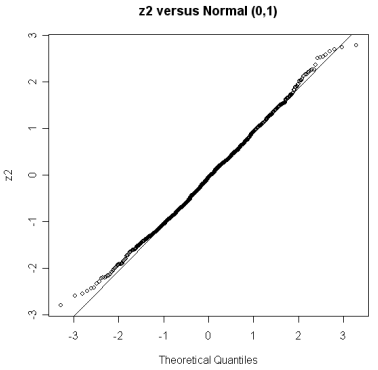
Figure 5.2.2: Q-Q plot of the quantity $z_2 = \dfrac{(X^{*'}X^*)^{1/2}(\hat{\beta}^* - \beta^*)}{\hat{\sigma}^*}$ for different sample sizes.



Sample size = 50
Mean = -0.04748 (~ 0)
Variance = 1.06303 (~ 1)



Sample size = 100
Mean = -0.01527 (~ 0)
Variance = 0.86465 (~ 1)

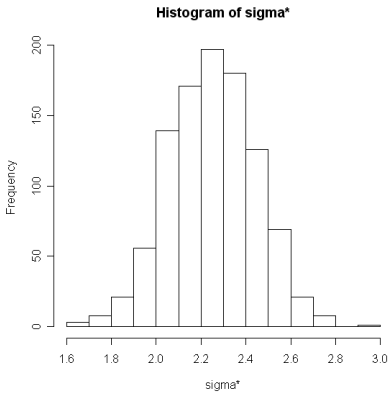


Sample size = 500
Mean = -0.06711 (~ 0)
Variance = 0.93794 (~1)

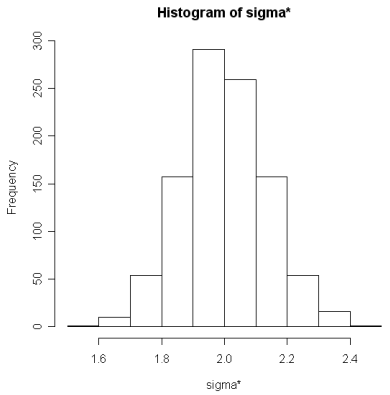
It appears from the Q-Q plots that both statistics approximately follow normal distribution. Figure 5.2.2 indicates that with the increases of sample size, mean and variances converge to zero and true variances respectively. However, Figure 5.2.2 shows that though mean of z_2 converges to zero as sample size increases, there are slight variations observed among the variances. Freedman (1981) discussed about the choice of bootstrap sample size. For instance, a

sample of size *n* can be bootstrapped to see what would happen with a sample of size is n^2, or $\sqrt{n}$, or others. In our case, we consider same bootstrap sample size as the sample size. The result in Figure 5.2.3 provides that the square root of the estimated variance of the random error for bootstrap converges to true square root of the variance of the random error when the sample sizes are increased.

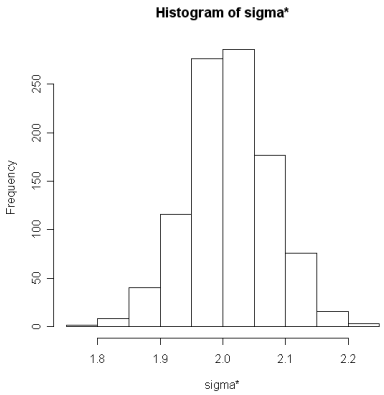
Figure 5.2.3: Histogram of the distribution of σ^* for different sample sizes.



Sample size = 50
Mean = 2.25350 (~2)
Variance = 0.03663



Sample size = 100
Mean = 2.00085
Variance = 0.01768



Sample size = 500
Mean = 2.010431
Variance = 0.00440

5.3 Logistic Regression Model Using the Bootstrap Method

For the generalized linear models (GLMs), Moulton and Zeger (1991) used bootstrap methods to estimate the functions of the estimated parameters. In that paper, they adopt bootstrapping techniques for GLM analogous to those used for ordinary linear models. Furthermore, they proposed a one-step procedure to estimate the parameters for each bootstrap replication though iteration to convergence cannot generally be expected. The reason behind the

one-step procedure is that they are efficient in terms of computation time. Like the linear regression models, we do not assume additivity of the error for the logistic regression models, and hence, the exchangeability of the error terms is typically no longer valid for these models. Nevertheless, Friedl and Tilg (1995) used the residual resampling method for the variance estimation in the logistic regression model. However, in this section, we use the vector resampling algorithm for the logistic regression model to estimate the variances. The algorithm of vector resampling method for the logistic regression model is as follows:

Step 1: Create the following pseudo-data set by resampling from the original data as mentioned in section 5.2, we get

$$(y_i^*, x_i^{*T})^T,\ i = 1, 2, \ldots, n$$

Step 2: Carry out the Newton-Raphson method as discussed in chapter 2,

$$\beta^{*(t+1)} = \beta^{*(t)} + (X^{*T} W^{*(t)} X^*)^{-1} X^{*T} (y^* - \mu^{*(t)})$$

to estimate β^* (notations and terminologies are the same as discussed in chapter 2)

Step 3: Repeat steps 1 and 2 for B times, where B is the number of repetition, one could use the resulting bootstrap estimates $\hat{\beta}^{*(1)}, \hat{\beta}^{*(2)}, \ldots, \hat{\beta}^{*(B)}$ to estimate variances or confidence intervals. The bootstrap estimate of the covariance matrix of $\hat{\beta}$ is

$$Var(\hat{\beta}^*) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}_b^* - \bar{\hat{\beta}}^* \right) \left(\hat{\beta}_b^* - \bar{\hat{\beta}}^* \right)^T, \text{ where } \bar{\hat{\beta}}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_b^* \qquad (5.3.1)$$

We use two datasets the Low Birth Weight Study (see, Hosmer and Lameshow, 2000) and Timing of Induced Abortion Study discussed in Chapter 2 to illustrate this technique. Then the comparative pictures of the variances are discussed.

5.3.1 Applications of Bootstrapping Logistic Regression Model

Study I (sample size=189):

The following example explains the method relating to a sample data of 189 subjects obtained from the Baystate Medical Center in Springfield, Massachusetts (Hosmer and Lameshow, 2000). The data set contains information on births in which 59 were low birth weight to women seen in the obstetrics clinic. Low birth weight, defined as birth weight less than 2500 grams, is an outcome that has been of concern to physicians for years because infant mortality rates and birth defect rates are very high due to the low birth weight. There are several factors, such as mother's age and smoking habits, that greatly affect the delivery of a baby of normal birth weight. Here, we discuss the estimates of $Var(\hat{\beta})$ in the logistic regression model using the vector resampling method. The description of the variables is given below:

Table 5.3.1: Code sheet for the selected variables in the low birth weight data

Variables	Codes/Values	Varable's Name
Low Birth Weight	0 = ≥ 2500 gm 1 = < 2500 gm	LOW
Smoking Status During Pregnancy	1 = Yes 0 = None	SMOKE
Age of Mother	0 = ≤ 25 Years 1 = > 25 Years	AGEYR

The cross-classification of the variables is shown in the following table. As can be seen from the table, the data are distributed evenly, that is, without any sparse cell.

Table 5.3.2: Cross-classification of Low Birth Weight × Age of Mother × Smoking Status

Age of Mother			Low Birth Weight ≥ 2500 gm	Low Birth Weight < 2500 gm	Total
≤ 25 years	Smoking Status	No	53	19	72
		Yes	27	21	48
	Total		80	40	120
> 25 years	Smoking Status	No	33	10	43
		Yes	17	9	26
	Total		50	19	69

For the bootstrapping logistic regression model using vector resampling, we take the bootstrap sample of the same size from the original sample data and then estimate the parameter using the logistic regression model. We replicate the procedure 1,000 times, and then estimates of the parameters are the average of parameters' values obtained from 1,000 replication. The variances of the parameters are obtained by the equation (5.3.1). The results of Table 5.3.3 are provided based on both the classical and bootstrapping logistic regression models. It indicates that both models provide almost similar estimates of the parameters, but the standard errors are slightly higher for the bootstrap method.

Table 5.3.3: Comparative results of the estimated parameters and their standard errors based on the classical logistic and bootstrapping logistic regression models

Classical logistic regression model			Bootstrapping logistic regression model		
Variables	Coefficients	Std. Errors	Variables	Coefficients	Std. Errors
SMOKE	0.701	0.320	SMOKE	0.716	0.338
AGEYR	-0.265	0.336	AGEYR	-0.264	0.356

Study II (sample size=2,247):

We consider here the part of the data in the timing of the induced abortion study discussed in the last section of Chapter 2. Here, we have included only the outcome variable of the women's abortion in or after the 3 months of gestation and two other explanatory variables, study area and women's education. The purpose of the study is to see the effects of study area and women's education of the women who had sought abortion in or after the third month of gestation. The variable description is provided below.

Table 5.3.4: Code sheet for the selected variables in the timing of induced abortion study

Variables	Codes/Values	Variable's Name
Gestational age	0 = less than 3 months 1 = 3+ months	GEST
Study area	1 = ICDDR,B area 2 = Comparison area	AREA
Women's education	0 = No education 1 = Some education	EDUYR

To see the cell frequencies of the contingency table, the results of the cross-classification of the variables are displayed in the following table. Again, the cross-classification shows that the data are distributed without any sparse cell.

Table 5.3.5: Cross-classification of Area × Gestational age × Mother's education

			Gestational age		Total
Women's education			<3 months	3+ months	
No education	Area	ICDDR,B area	111	166	277
		Comparison area	158	516	674
	Total		269	682	951
Some education	Area	ICDDR,B area	200	192	392
		Comparison area	269	635	904
	Total		469	827	1296

Table 5.3.6: Comparative results of the estimated parameters and their standard errors based on the classical logistic and bootstrapping logistic regression models

Classical logistic regression model			Bootstrapping logistic regression model		
Variables	Coefficients	Std. Errors	Variables	Coefficients	Std. Errors
AREA	0.852	0.096	AREA	0.851	0.098
EDUYR	-0.366	0.094	EDUYR	-0.367	0.097

After performing the classical logistic and bootstrapping logistic regression models, it is seen that parameter estimates are quite similar for both models, but the standard errors are little bit higher for bootstrap model.

Therefore, based on the two studies of different sample sizes, results of the simulation suggest that the bootstrap method provides slightly high variances compared to that of the classical logistic regression method. It would be interesting to see how the bootstrap method performs if the contingency table contains sparse data in one more cells.

CHAPTER 6

SUMMARY AND CONCLUSIONS

This dissertation dealt with logistic regression models and their variations. In Chapter 2, the details of the logistic regression model were demonstrated, followed by the MLE procedure to estimate the unknown parameters of the model. The main emphasis of this chapter was to prove the asymptotic properties of the MLE for the logistic regression model using a completely different approach. In particular, the logistic regression models have serious numerical problems if zero cells occur in the contingency table, and for this scenario, the different approach was motivated. In addition, the simulation study was carried out to assess the finite sample behavior of the consistency and normality of the MLE. The results of the simulation studies were provided through the tabular form and graphical display to get a clear pictures of the consistency and normality of the MLE for different sample sizes. In the last section of the second chapter, an application based on the real dataset was illustrated. This application identifies several risk factors, such as women residing in the Comparison area, having no living children, and having no education which significantly increased the risk of having an induced abortion in or after the third month of pregnancy.

In Chapter 3, the generalization of the hybrid logistic regression model under case-control study was discussed. The hybrid logistic model was originally proposed by Chen et al. (2003) which deals with situations in which risk factors associated with the outcome are exceedingly rare in the control group. In principle, a two-stage hybrid procedure models the risks due to the rare factors in the first stage and models the residual risks due to the other factors in the second stage using the standard logistic regression model. In the case of the

generalization, the rare risk factors were considered both independent and not independent, and we discussed the relevant estimation procedures for the parameters.

An outline of the multinomial logistic regression model was given in Chapter 4, followed by the simulation study for the consistency and normality of the MLE for this model. This simulation study ensured that when sample size increases, the estimated parameters converge to their true values and follow approximately normal distributions. For the three categories' outcome variable of certain data, a set of important risk factors was identified by applying the multinomial logistic regression model. In addition, this chapter extends the hybrid logistic regression model to the multinomial hybrid logistic regression model, and this can be employed when the case group of the outcome variable has mutually exclusive and exhaustive subgroups. Based on the three categories' outcome variable with a rare risk factor, the estimation procedure of the parameters was discussed at the end of this chapter.

In the last part of the dissertation, the bootstrap method to estimate the variances for the parameter estimates in the logistic regression model was studied. Two examples of different sample sizes were applied to the classical logistic and bootstrapping logistic regression models. The results of the simulation suggested that the bootstrap method provides slightly high variances compared to that of the classical logistic regression method.

Elsewhere in the dissertation, we identified some follow on work that others could pursue. Those are summarized below.

(i) The hybrid logistic regression models were developed for the case-control studies only; one could propose a new model for the cohort or prospective studies and discuss the estimation procedures of the unknown parameters for the model.

(ii) The hybrid logistic regression models can be fitted in stratified case-control studies if separate samples of cases and controls might be taken within each stratum.

(iii) The hybrid logistic regression model and its generalization considered the rare risk factors are categorical, one could model the rare risks measured on a continuous scale.

(iv) It would be interesting to perform a simulation study to show the asymptotic properties of the hybrid logistic and multinomial hybrid regression models under case-control study.

(v) If the data are available, then one could provide applications of the generalization of the hybrid logistic regression models and the multinomial hybrid logistic model.

(vi) It would be interesting to see how the bootstrapping logistic regression model performs if the contingency table contains sparse data.

REFERENCES

[1] Agresti, A. (1996). *An Introduction to Categorical Data analysis.* John Wiley & Sons, Inc.

[2] Agresti, A. (2002). *Categorical Data Analysis.* John Wiley & Sons, Inc.

[3] Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922. *Statistical Science,* **12**, 162-176.

[4] Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, Harvard University Press.

[5] Anderson, J.A. (1972). Separate Sample Logistic Discrimination. *Biometrika*, **59**, 19-35.

[6] Beer, M. (2001). Asymptotic Properties of the Maximum Likelihood Estimator in [1] Dichotomous Logistic Regression Models. Diploma Thesis, *University of Fribourg Switzerland*.

[7] Beran, R. and Ducharme G.R. (1991). *Asymptotic Theory for Bootstrap Methods in Statistics.* Les Publications CRM, Canada.

[8] Bickel, P.J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, **9,** 1196-1217.

[9] Bishop, Y.M.M., Feinberg, S.E., and Holland, P.W. (1975). *Discrete multivariate analysis: Theory and Practice,* MIT Press, Boston.

[10] Bland J. Martin and Altman G. Douglas (2000). Statistics Notes: The odds ratio. *British Medical Journal,* 320, 1468.

[11] Boos, D. D. (2003). Introduction to the Bootstrap World. *Statistical Science*, **18**(2), 168-174.

[12] Brent, D.A., Baugher,M, Bridge,J., Chen, T., and Chiappetta, L. (1999). Age- and Sex-related factors for adolescent suicide. *Journal of the American Academy of Child and*

Adolescent Psychiatrry, **38**, 1497-1505.

[13] Breslow, N. and Powers, W. (1978). Are There Two Logistic Regressions for Retrospective Studies? *Biometrics*, **34**, 100-105.

[14] Breslow, N. and Powers, W. (1978). Are There Two Logistic Regressions for Retrospective Studies? *Biometrics*, **34**, 100-105.

[15] Breslow, N. E. (1996). Statistics in Epidemiology: The Case-Control Study. *Journal of the American Statistical Association*, **91**, 14-28.

[16] Breslow, N. E. and Cain, K. C (1988). Logistic regression for two-stage case-control data. *Biometrika*, **75,** 11-20.

[17] Carroll R. J., Ruppert, D. and Stefanski, L.A. (2006). *Measurement Error in Nonlinear Models.* Chapman & Hall, CRC Press, London.

[18] Chan, Y. H. (2004). Multinomial logistic regression. *Singapore Medical Journal,* **46**, 259-269.

[19] Chen, T, Hoppe, F.M., Iyengar, S., and Brent, D. (2003). A Hybrid Logistic Model for Case-Control Studies. *Methodology and Computing in Applied Probability,* **5**, 419-2003.

[20] Christensen, R. (1997). *Log-Linear Models and Logistic Regression.* 2nd Ed., Springer, New York.

[21] Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression. *Journal of the American Statistical Association*, **86**, 68-78.

[22] Cornfield, J. (1951). A Method of Estimating Comparative Rates from Clinical Data. Applications to cancer of the Lung, Breast, and Cervix. *Journal of the National cancer Institute,* **11**, 1269-1275.

[23] Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure. *Federation Proc*., **21**, 58-61

[24] Cox, D. R. (1970). *Analysis of Binary Data.* London: Chapman & Hall.

[25] Cox, D. R., and Snell, E. J. (1989). *Analysis of Binary Data.* London: Chapman and Hall.

[26] Cramer, J.S. (2003). *Logit Models from Economics and Other Fields*. Cambridge University Press.

[27] Davison, A.C. and Hinckley, D.V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press, New York.

[28] Dillon, W. R., Goldstein, M., and Lement, L. (1981). Analyzing Qualitative Predictors with Too Few Data: An Alternative Approach to Handling Sparse-Cell Values. *Journal of Marketing Resear*ch, **18**, 63-72.

[29] Durrett, R. (2004). *Probability: Theory and Examples*. Third Ed., Thomson, Brooks/Cole.

[30] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, **7**, 1-26.

[31] Efron, B. (1994). Missing Data, Imputation, and the Bootstrap. *Journal of the American Statistical Association*, **89**, 463-478.

[32] Efron, B. and Tsibirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[33] Farewell, V. T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika*, **66**, 27-32.

[34] Fears, 'T. R. and Brown. C. C. (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics*, **42**, 955-960.

[35] Fienberg, S. (1980). *The analysis of cross-classified categorical data*, 2nd ed. Boston, MA:

MIT Press.

[36] Fienberg, S. and P. Holland (1972). On the choice of flattening constants for estimating multinomial Probabilities. *Journal of Multivariate Analysis*, 2, 127—134.

[37] Freedman, D.A. (1981). Bootstrapping regression models. *Annals of Statistics*, **9**, 1218-1228.

[38] Friedl, H. and Tilg, N. (1994). Variance estimates in logistic regression using the bootstrap. *Communications in Statistics - Theory and Methods*, **24**, 473-486.

[39] Gart, J. J. (1966). Alternative analyses of contingency tables. *Journal of the Royal Statistical Society*, Series B **28**, 164-179.

[40] Gart, J. J. and Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika*, **54**, 181-187.

[41] Givens, G.H. and Hoeting, J.A. (2005). *Computational Statistics*, John Wiley & Sons, Inc.

[42] Goodman, L. A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *Journal of the American Statistical Association,* **65**, 226-256.

[43] Goodman, L. A. (1971). The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics*, **13**, 33-61.

[44] Gourieroux, C. and Monfort, A. (1981). Asymptotic Properties of the Maximum Likelihood Estimator in Dichotomous Logit Models. *Journal of Econometrics*, **17**, 83-97.

[45] Haldane, J. B. S. (1956). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, **20**, 309-311.

[46] Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*. Second Edition, John Wiley & Sons Inc., New York.

[47] Khan, Rochat RW, Jahan FA, Begum SF, (1988). Induced abortion in a rural area of Bangladesh. *Studies in Family Planning,* **17**, 95-99

[48] Lehman, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York.

[49] Mak, K. Tak (1993). Solving Non-Linear Estimation Equations**.** *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**, 945-955.

[50] Mammen, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulations*. Springer-Verlag, New York.

[51] Mantel, N. (1973). Synthetic Retrospective Studies and Related Topics. *Biometrics*, **29**, 479-486.

[52] Mantel, N. and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *Journal of the National Cancer Institute,* **22**, 719-748.

[53] McCullagh, P. and J. A. Nelder (1989). *Generalized linear model*s. 2nd Ed. Chapman and Hall, New York, USA.

[54] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, Edited by Paul Zarembka. New York: Academic Press.

[55] Moulton, L.H., and Zeger, S.L. (1991). Bootstrapping generalized linear models. *Computational Statistics & Data Analysis,* **11**, 53-63.

[56] Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical. Psychology*. **47**, 90-100.

[57] Peters, S. C. and Freedman, D. A. (1984). Some Notes on the Bootstrap in Regression Problems. *Journal of Business & Economic Statistics*, **2**, 406-409.

[58] Prentice, R. (1976). Use of the Logistic Model in Retrospective Studies. *Biometrics*, **32**, 599-606

[59] Prentice, R.L. and Breslow, N.E. (1978). Retrospective Studies and Failure Time Models. *Biometrika*, **66**, 153-158.

[60] Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika,* **66**, 403-411.

[61] Rao, C. R. (1973). Linear Statistical Inference and Its Applications. 2nd Ed. New York: Wiley.

[62] Rashid, M and Ahmed, K (2002). Correlates of timing of induced abortion in rural Bangladesh. *Paper presented at the 2002 Asia Pacific Social Science and Medical Conference,* Kuming, China.

[63] Rashid, M. and Shifa, N. (2007). Mistimed and Unwanted Pregnancies in Bangladesh: Trends and Determinants. *Paper presented at the Population Association of America (PAA) Annual Conference*, NY, USA.

[64] Sai, Fred T. and Nassim, J. (1989). The Need for a Reproductive Health Approach. *International Journal of Gyneocology's Obsterics,* **3**, 103-113.

[65] Santnner, T. J. and Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data*. Springer, New York.

[66] Siegel, D.G. and Greenhouse, S. W. (1973). Multiple Relative risk Functions in Case-Control Studies. *American Journal of Epidemiology*, **97**, 324-331.

[67] Scott, A. J. and Wild, C. J. (1986). Fitting. logistic models under case-control or choice-based sampling. Journal of the Royal Statistical Society, Serial B **48**, 170-182,

[68] Scott, A.J and Wild, C. J.(1991). Fitting Logistic Regression Models in Stratified Case-Control Studies. *Biometrics*, **47**, 497-510.

[69] Shaffer, D.,Gould, M.S., Trautman, P., Moreau, D., Kleinman,M., and Flory,M. (1996).

Psychiatric Diagnosos in Child and Adolescent Suicide. *Archives of General Psychiatry*, **55**, 339-348.

[70] Singh, K. (1981). On the Asymptotic Accuracy of Efron's Bootstrap. *Annals of Statistics,* **9**, 1187-1196

[71] Stine, R. (1989). An Introduction to Bootstrap Methods**.** *Sociological Methods and Research***, 18**(2 and 3), 243-291.

[71] Stine, R. A. (1985). Bootstrap Prediction Intervals for Regression. *Journal of the American Statistical Association*, **80**, 1026-1031.

[72] van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

[73] Walter, S. D. and Cook, R. J. (1991). A Comparison of Several Point Estimators of the Odds ratio in a Single 2×2 Contingency Table. *Biometrics*, **47**, 792-811.

[74] Wu, C. F. J. (1986). Jackknife, bootstrap and the resampling methods in regression analysis. *Annals of Statistics***, 14**, 1261-135.

[75] Zhang, B. (2006). Prospective and retrospective analyses under logistic regression models. *Journal of Multivariate Analysis*, **97**, 211-230.

[76] Zocchi, S. S. and Atkinson, A. C. (1999). Optimum Experimental Designs for Multinomial Logistic Models. *Biometrics*, **55**, 437-444.

APPENDIX: R AND SPSS CODES

```
#----------------------------CHAPTER 2-------------------------------------
# Figure 2.1.1: logistic regression function

x=seq(-4,4,.0001)
p=exp(x)/(1+exp(x))
plot(x,p,"l")

# Consistency and Normality:

data1=rbinom(200,1,.5); x1=data1
data2=rbinom(200,1,.5); x2=data2
data3=rbinom(200,1,.5); x3=data3
data4=rbinom(200,1,.5); x4=data4

N=1000; a=0; b1=0; b2=0; b3=0; b4=0
for (j in 1:N)
{
   alpha=0.7; beta1=1; beta2=1.3; beta3=0.25; beta4=0.05
   p=0
   y=0
   for (i in 1:length(x1))
   {
   p[i]=exp(alpha+beta1*x1[i]+beta2*x2[i]+beta3*x3[i]+beta4*x4[i])/
   (1+exp(alpha+beta1*x1[i]+beta2*x2[i]+beta3*x3[i]+beta4*x4[i]))
   y[i]=rbinom(1,1,p[i])
   }

# Newton-Raphson algorithm
 lmodel = function(x, y, maxits=20, eps=1e-10)
 {
   # use a starting value of beta=0
   newbeta = rep(0, ncol(x) )
   iter = 0
   converged = F
   while( (!converged) & (iter<maxits) )
   {
   iter = iter+1
   cat(iter); cat("...")
   beta = newbeta
   tmp = exp(x%*%beta)
   pi = tmp/(1+tmp)
   mu = pi
   w = as.vector(pi*(1-pi)) # this is a vector, not a matrix
   xtwx = t(w*x)%*%x
   xtwxinv = solve(xtwx)
   newbeta = beta + xtwxinv%*%t(x)%*%(y-mu)
   converged = all(abs(newbeta-beta)<eps)
   }
 cat("\n")
 tmp = exp(x%*%newbeta)
 pi = tmp/(1+tmp)
 loglik = sum( y*log(pi) + (1-y)*log(1-pi) )
 # add names to coefficients
```

```
 names(newbeta) = dimnames(x)[[2]]
 result = list(beta=newbeta, cov.beta=xtwxinv, iter=iter,
 converged=converged, loglik=loglik)
 result
 }

x = cbind(int=1,x1,x2,x3,x4)
 res = lmodel(x,y)

# Consistency
   a[j]=res$beta[1,]
   b1[j]=res$beta[2,]
   b2[j]=res$beta[3,]
   b3[j]=res$beta[4,]
   b4[j]=res$beta[5,]


# Normality
 n = length(x1)
 betao = c(alpha,beta1,beta2,beta3,beta4)
 betao = matrix(betao,5,1)

 betae = c(res$beta[1,],res$beta[2,],res$beta[3,],res$beta[4,],res$beta[5,])
 betae = matrix(betae,5,1)

 d1 =(betae-betao)
 d2 = sqrt(solve(matrix(res$cov.beta[1:25],nrow=5,ncol=5)))
 d=d2%*%d1
 dd1[j]=d[1,];dd2[j]=d[2,];dd3[j]=d[3,];dd4[j]=d[4,];dd5[j]=d[5,]

}

# Output for consistency
mean(a); mean(b1); mean(b2); mean(b3); mean(b4)
sd(a)/sqrt(N); sd(b1)/sqrt(N); sd(b2)/sqrt(N); sd(b3)/sqrt(N); sd(b4)/sqrt(N)

# Output for normality
qqnorm(dd2, ylab='Beta1', main="Beta1 versus Normal (0,1)"); qqline(dd2)
mean(dd2);  sd(dd2)^2;sd(dd2)/sqrt(1000)

qqnorm(dd3, ylab='Beta2', main="Beta2 versus Normal (0,1)"); qqline(dd3)
mean(dd3);  sd(dd3)^2;sd(dd3)/sqrt(1000)

qqnorm(dd4, ylab='Beta3', main="Beta3 versus Normal (0,1)"); qqline(dd4)
mean(dd4);  sd(dd4)^2;sd(dd4)/sqrt(1000)

qqnorm(dd5, ylab='Beta4', main="Beta4 versus Normal (0,1)"); qqline(dd5)
mean(dd5);  sd(dd5)^2;sd(dd5)/sqrt(1000)


****Application : SPSS syntax code****

get translate file='c:\Mamun\Research_Paper\kapil\Time\DABR8998.DBF'.

recode gest (1,2=0)(else=1).
value labels gest 0'<3' 1'3+'.
```

```
*recode gest (1=1)(2=2)(else=3).
*value labels gest 1'1' 2'2' 3'3+'.

recode mage (lo thru 19=1)(20 thru 29=2)(30 thru 39=3)(40 thru hi=4).
value labels mage 1'<20' 2'20-29' 3'30-39' 4'40+'.

recode livch (0=0)(1 thru 2=1)(else=2).
value labels livch 0'0' 1'1-2' 2'3+'.

recode meduyr (0=0)(1 thru 16=1)(99=sysmis).
value labels meduyr 0'No education' 1'Some education'.

recode dwell (low thru 349=1)(else=2).
value labels dwell 1'<350' 2'350+'.

value labels area 1' MCH-FP' 2'Comparison'.

value labels religion 1'Muslim' 2'Non-muslim'.

*recode occu (999=sysmis)(50,106=1)(104,45=2)(else=3).
*value labels occu  2'Students' 3'Others'.

recode occu (999=sysmis)(50,106,104,45=1)(else=2).
value labels occu  1'Not working' 2'Working'.

recode marr_age (0 thru 11=sysmis)(12 thru 16=1)( 17 thru 20=2)(21 thru
hi=3).
value labels marr_age 1'12-16' 2'17-20' 3'21+'.

*recode age_m_a (0 thru 4=1)(5 thru 900=2)(999=sysmis).
*value labels age_m_a 1'<5' 2'5+'.

*recode con (1 thru 9=1) (44=2) (else=sysmis).
*value labels con 1'User' 2'Non-user'.

*recode con (1=1) (2=2) (3=3) (5=5) (6 thru 9=6) (44=7) (else=sysmis).
*value labels con 1'Pill' 2'IUD' 3'Injection' 5'Condom' 6'Others' 7'Non-
user'.

*recode con (1 thru 5=1) (6 thru 9=2) (else=sysmis).
*value labels con 1'Modern' 2'Traditional' .

cross mage livch meduyr dwell area religion occu marr_age by gest/cells=count
row/stat=chisq.

LOGISTIC REGRESSION VAR=gest
  /METHOD=BSTEP(LR) mage livch meduyr area occu
  /CONTRAST (mage)=Indicator(1)  /CONTRAST (livch)=Indicator(1)  /CONTRAST
  (meduyr)=Indicator(1)  /CONTRAST (area)=Indicator(1)  /CONTRAST
  (occu)=Indicator(1)
  /PRINT=GOODFIT
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

```
#----------------------------CHAPTER 4-------------------------------------
#Consistency and Normality:

data1=rbinom(1500,1,.8); x1=data1
data2=rbinom(1500,1,.4); x2=data2
data3=rbinom(1500,1,.5); x3=data3
data4=rbinom(1500,1,.5); x4=data4

beta01=0.4; beta11=0.8; beta12=1.3; beta13=-0.5; beta14=1.1

beta02=1.2; beta21=1.5; beta22=0.9; beta23=0.2; beta24=-0.5

p1=exp(beta01+beta11*x1+beta12*x2+beta13*x3+beta14*x4)/(1+exp(beta01+beta11*x
1+beta12*x2+beta13*x3+beta14*x4)+exp(beta02+beta21*x1+beta22*x2+beta23*x3+bet
a24*x4))

p2=exp(beta02+beta21*x1+beta22*x2+beta23*x3+beta24*x4)/(1+exp(beta01+beta11*x
1+beta12*x2+beta13*x3+beta14*x4)+exp(beta02+beta21*x1+beta22*x2+beta23*x3+bet
a24*x4))

p0=1-p1-p2

N=1000
b01=0;b11=0;b12=0;b13=0;b14=0
b02=0;b21=0;b22=0;b23=0;b24=0

d1=0;dd1=0;dd2=0;dd3=0;dd4=0;dd5=0
d2=0;g1=0;g2=0;g3=0;g4=0;g5=0

for (j in 1:N)
{
  n=length(x1)
  y=0
  for (i in 1:n)
 {
 x=rmultinom(1,1,prob=c(p0[i],p1[i],p2[i]))
 if (x[1,]==1)
      {
      y[i]=0
      }
            if (x[2,]==1)
            {
            y[i]=1
            }
                  if (x[3,]==1)
                  {
                  y[i]=2
                  }

}

cbind(y,x1,x2,x3,x4)

# nnet package for multinomial distribution

library(nnet)
 out=summary(multinom(y~x1+x2+x3+x4, Hess=T))
```

```
# Consistency
b01[j]=out$coefficients[1]
b11[j]=out$coefficients[3]
b12[j]=out$coefficients[5]
b13[j]=out$coefficients[7]
b14[j]=out$coefficients[9]

b02[j]=out$coefficients[2]
b21[j]=out$coefficients[4]
b22[j]=out$coefficients[6]
b23[j]=out$coefficients[8]
b24[j]=out$coefficients[10]

# Normality

 n = length(x1)
 betao1 = c(beta01,beta11,beta12,beta13,beta14)
 betao1 = matrix(betao1,5,1)

 betae1 =
 c(out$coefficients[1,1],out$coefficients[1,2],out$coefficients[1,3],
 out$coefficients[1,4],out$coefficients[1,5])
 betae1 = matrix(betae1,5,1)

 d11 =(betae1-betao1)
 d21 = sqrt(solve(solve(out$Hessian)))[1:5,1:5]
 d1=d21%*%d11
 dd1[j]=d1[1,];dd2[j]=d1[2,];dd3[j]=d1[3,];dd4[j]=d1[4,];dd5[j]=d1[5,]

#-----------------------------------------

 betao2 = c(beta02,beta21,beta22,beta23,beta24)
 betao2 = matrix(betao2,5,1)

 betae2 =
 c(out$coefficients[2,1],out$coefficients[2,2],out$coefficients[2,3],
 out$coefficients[2,4],out$coefficients[2,5])
 betae2 = matrix(betae1,5,1)

 d12 =(betae2-betao2)
 d22 = sqrt(solve(solve(out$Hessian)))[6:10,6:10]
 d2=d22%*%d12
 g1[j]=d2[1,];g2[j]=d2[2,];g3[j]=d2[3,];g4[j]=d2[4,];g5[j]=d2[5,]
}

# Output for consistency

mean(b01); mean(b11); mean(b12); mean(b13); mean(b14)

mean(b02); mean(b21); mean(b22); mean(b23); mean(b24)

sd(b01)/sqrt(N); sd(b11)/sqrt(N); sd(b12)/sqrt(N); sd(b13)/sqrt(N);
sd(b14)/sqrt(N)

sd(b02)/sqrt(N); sd(b21)/sqrt(N); sd(b22)/sqrt(N); sd(b23)/sqrt(N)
sd(b24)/sqrt(N)
```

```
# Output for normality
# For 1st set of beta

qqnorm(dd2, ylab='Beta11', main="Beta11 versus Normal (0,1)"); qqline(dd2)
mean(dd2);  sd(dd2)^2;sd(dd2)/sqrt(1000)

qqnorm(dd3, ylab='Beta12', main="Beta12 versus Normal (0,1)"); qqline(dd3)
mean(dd3);  sd(dd3)^2;sd(dd3)/sqrt(1000)

qqnorm(dd4, ylab='Beta13', main="Beta13 versus Normal (0,1)"); qqline(dd4)
mean(dd4);  sd(dd4)^2;sd(dd4)/sqrt(1000)

qqnorm(dd5, ylab='Beta14', main="Beta14 versus Normal (0,1)"); qqline(dd5)
mean(dd5);  sd(dd5)^2;sd(dd5)/sqrt(1000)

# For 2nd set of beta

qqnorm(g2, ylab='Beta21', main="Beta21 versus Normal (0,1)"); qqline(g2)
mean(g2);  sd(g2)^2;sd(g2)/sqrt(1000)

qqnorm(g3, ylab='Beta22', main="Beta22 versus Normal (0,1)"); qqline(g3)
mean(dd3);  sd(g3)^2;sd(g3)/sqrt(1000)

qqnorm(g4, ylab='Beta23', main="Beta23 versus Normal (0,1)"); qqline(g4)
mean(g4);  sd(g4)^2;sd(g4)/sqrt(1000)

qqnorm(g5, ylab='Beta24', main="Beta24 versus Normal (0,1)"); qqline(g5)
mean(g5);  sd(g5)^2;sd(g5)/sqrt(1000)


******Application: SPSS Code**********

get file='c:\mamun\paa2007\unintended\women_2004.sav'.

compute wtvar=v005/1000000.
weight by wtvar.

*Dependent variable.
recode v225 v367(9=sysmis).
recode v225 v367(sysmis=0).

compute v3677=v367.
if v225<>0  v3677=0.
compute unintd=v225+ v3677.
value labels unintd 1'planned' 2' mistimed' 3'unwanted'.
select if unintd<>0.

*freq unintd.
*Explanatory variables.

*v024=region of residence.
*v025=urban/rural.
*v012=current age.
recode v012 ( lo thru 19=1) (20 thru 29=2) (30 thru hi=3).
value label v012 1 '<19' 2'20-29' 3'30+'.
*v106=education.
*v302=ever use modern method.
```

```
recode v302 (3=1)(else=0).
value labels v302 1'yes' 0'no'.
*v511=age at first marriage.
recode  v511(lo thru 14=1)(15 thru 19=2)(20 thru hi=3).
value labels v511 1'<15' 2'15-19' 3'20+'.
*v714=employment.
recode v714(9=sysmis).
*v190=wealth.
**v218=lchild.
recode v218 (0=0) (1 thru 2=1) (3 thru 4=2)( 5 thru hi=3).
value labels v218 0'None' 1'1-2' 2'3-4' 3'5+'.
*access to media.
compute media=1.
if (v157=0 and v158=0 and v159=0) media=0.
*v611=dicuss FP.
recode v611 (9=sysmis).
*v130=religion.
recode v130 (1=1) (2 thru 8=2) (9=sysmis).
value labels v130 1'Muslim' 2'Non-Muslim'.

*freq unintd v024 v025 v012 v106 v302  v511 v714 v190 v218 media v611 v130.

*CROSSTABS
/TABLES=v024 v025 v012 v106 v302  v511 v714 v190 v218 media v611 v130 by
unintd/FORMAT= AVALUE TABLES/STATISTIC=CHISQ CORR/CELLS= COUNT ROW .

*Unwanted VS wanted**Mistimed VS wanted
*recode unintd (1=3)(2=2)(3=1).
*value labels unintd 1'unwanted' 2'mistimed' 3'planned'.

*freq unintd.

*NOMREG unintd  BY v012  media v106 v130 v218 v511 v302 v714 v190
/CRITERIA = CIN(95) DELTA(0) MXITER(100) MXSTEP(5) LCONVERGE(0)
PCONVERGE(1.0E-6) SINGULAR(1.0E-8)/MODEL/INTERCEPT = INCLUDE
/PRINT = FIT PARAMETER SUMMARY LRT.

*Unwanted VS Mistimed

recode unintd (1=1)(2=3)(3=2).
value labels unintd 1'wanted' 2'unwanted' 3'mistimed'.

freq unintd.

NOMREG unintd  BY v012  media v106 v130 v218 v511 v302 v714 v190
/CRITERIA = CIN(95) DELTA(0) MXITER(100) MXSTEP(5) LCONVERGE(0)
PCONVERGE(1.0E-6) SINGULAR(1.0E-8)/MODEL/INTERCEPT = INCLUDE
/PRINT = FIT PARAMETER SUMMARY LRT .



#----------------------------CHAPTER 5-------------------------------------

# An Example: Bootstrap consistency

n=50
mu=20; sigma=3
```

```
Q=0
for ( i in 1:1000)
{
x=rnorm(n,mu,sigma)
x.bar=mean(x)
Q[i]=sqrt(n)*(x.bar-mu)
}
qqnorm(Q, ylab='Q', main="Q versus Normal (0,1)");qqline(Q)

#-bootstrap---------

x1=rnorm(n,mu,sigma)
xx=x1
x1.bar=mean(xx)

Qstar=0

for ( i in 1:1000)
{
xstar=sample(xx,replace=T)
xstar.bar=mean(xstar)
Qstar[i]=sqrt(n)*(xstar.bar-x.bar)
}

qqnorm(Qstar, ylab='Q*', main="Q* versus Normal (0,1)");qqline(Qstar)

qqplot(Qstar,Q, main="Q* versus Q")

# Linear regression consistency and normality
rm(list=ls())
set.seed(12345)

n=100       # original sample size

dat=rnorm(n,19,10)
x=dat

mean=0;sd=2
b0=0.9; b1=1.5
b01=as.matrix(c(b0,b1))

datax=matrix(c(rep.int(1, n),x),c(n,2))
v=(1/n)*t(datax)%*%datax

sd^2*solve(v)

error=rnorm(n,mean,sd)
y=b0+b1*x+error
fit=lm(y~x)
b0hat=summary(fit)$coef[1]
b1hat=summary(fit)$coef[2]

a1=data.frame(x,y)
a=data.matrix(a1)

b=1000      # number of bootstrap replication
```

```
x0=0; x1=0; sigma.boot=0; x00=0; x11=0

for (i in 1:b)
{
      m=n         # bootstrap sample size
      v=sample(1:length(x),m,replace=TRUE)
      indep=a[v,1]
      dep=a[v,2]

      res=lm(dep~indep)
      bet0=summary(res)$coef[1]
      bet1=summary(res)$coef[2]


x0[i]=sqrt(m)*(bet0-b0hat)
x1[i]=sqrt(m)*(bet1-b1hat)

sigma.boot[i]=summary(res)$sigma

x00[i]=(bet0-b0hat)/sqrt(vcov(res)[1,1])
x11[i]=(bet1-b1hat)/sqrt(vcov(res)[2,2])

}

mean(x0); var(x0)
mean(x1); var(x1)

mean(sigma.boot)

mean(x00); var(x00)
mean(x11); var(x11)

qqnorm(x1, ylab='z1', main="z1 versus Normal (0,1)");qqline(x1)
qqnorm(x11, ylab='z2', main="z2 versus Normal (0,1)");qqline(x11)



########### Bootstrapping logistic model#############
# Study I

data.lbwt=read.table("lbwt.dat",header=TRUE)

data.lbwt

names(data.lbwt)
attach(data.lbwt)

logis.fit=glm(LOW~ SMOKE + AGEYR, family=binomial(link = logit))

summary(logis.fit)

detach(data.lbwt)

#Bootstrap method for study I

data.lbwt=read.table("lbwt.dat",header=T)
attach(data.lbwt)
```

```
n=length(SMOKE)
detach(data.lbwt)

boot0=0; boot1=0; boot2=0
for (i in 1:1000)
{
      m=n         # bootstrap sample size
      v=sample(1:n, m, replace=TRUE)

      LOWstar=data.lbwt[v,1]
      SMOKEstar=data.lbwt[v,2]
      AGEYRstar=data.lbwt[v,3]

      boot.fit=glm(LOWstar~SMOKEstar+AGEYRstar, family=binomial(link =
logit))

      boot0[i]=summary(boot.fit)$coef[1]
      boot1[i]=summary(boot.fit)$coef[2]
      boot2[i]=summary(boot.fit)$coef[3]
}

mean(boot0); sd(boot0)
mean(boot1); sd(boot1)
mean(boot2); sd(boot2)


#Study II
data.t=read.table("data3.dat", header=T)
data.t


names(data.t)
attach(data.t)

logis.fit=glm(GEST~ AREA + MEDUYR, family=binomial(link = logit))

summary(logis.fit)

detach(data.t)

#Bootstrap method for study II

data.t=read.table("data3.dat", header=T)
attach(data.t)
n=length(GEST)
detach(data.t)

boot0=0; boot1=0; boot2=0
for (i in 1:1000)
{
      m=n         # bootstrap sample size
      v=sample(1:n, m, replace=TRUE)

      AREAstar=data.t[v,1]
      GESTstar=data.t[v,2]
      MEDUYRstar=data.t[v,3]
```

```
      boot.fit=glm(GESTstar~AREAstar+MEDUYRstar, family=binomial(link =
logit))

      boot0[i]=summary(boot.fit)$coef[1]
      boot1[i]=summary(boot.fit)$coef[2]
      boot2[i]=summary(boot.fit)$coef[3]
}

mean(boot0); sd(boot0)
mean(boot1); sd(boot1)
mean(boot2); sd(boot2)
```