TOPICS IN ORDINAL LOGISTIC REGRESSION AND ITS APPLICATIONS

A Dissertation

by

HYUN SUN KIM

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2004

Major Subject: Statistics

TOPICS IN ORDINAL LOGISTIC REGRESSION AND ITS APPLICATIONS

A Dissertation

by

HYUN SUN KIM

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Suojin Wang
(Chair of Committee)

Michael T. Longnecker
(Member)

Marina Vannucci
(Member)

Steven Taliaferro
(Member)

Michael T. Longnecker
(Head of Department)

August 2004

Major Subject: Statistics

ABSTRACT

Topics in Ordinal Logistic Regression and Its Applications. (August 2004)

Hyun Sun Kim, B.S., Dongguk University, Seoul, Korea;

M.S., Dongguk University, Seoul, Korea

Chair of Advisory Committee: Dr. Suojin Wang

Sample size calculation methods for ordinal logistic regression are proposed to test statistical hypotheses. The author was motivated to do this work by the need for statistical analysis of the red imported fire ants data. The proposed methods use the concept of approximation by the moment-generating function. Some correction methods are also suggested. When a prior data set is available, an empirical method is explored. Application of the proposed methodology to the fire ant mating flight data is demonstrated. The proposed sample size and power calculation methods are applied in the hypothesis testing problems. Simulation studies are also conducted to illustrate their performance and to compare them with existing methods.

To my parents

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge the excellent guidance and constant encouragement provided by Dr. Suojin Wang during the past few years. I also wish to thank Dr. Michael T. Longnecker, Dr. Marina Vannucci, and Dr. Steven Taliaferro for their willingness to serve on my advisory committee and to provide me with valuable suggestions.

I would like to thank Monisha Dey and Xian Zhou for their assistance in this research. This research was supported in part by a grant from the Texas Advanced Research Program.

Finally, I have to say 'thank-you' to: all my friends and family, wherever they are, particularly my parents who were a constant source of moral support. I am forever indebted to my parents for their understanding, endless patience and encouragement when it was most required.

TABLE OF CONTENTS

Page

ABSTRACT iii

DEDICATION iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vi

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER

I INTRODUCTION 1

II LITERATURE REVIEW 5

2.1 Introduction 5
2.2 Model and Likelihood Function 5
2.3 Sample Size Approximations 7
2.4 Power and Sample Size Calculations for Generalized Linear Models 10

III ORDINAL LOGISTIC REGRESSION 14

3.1 Introduction 14
3.2 Model 14
3.3 Likelihood Function 15
3.4 The Fisher Information Matrix 17
3.5 Sample Size Calculations 22

IV METHODS FOR SAMPLE SIZE CALCULATIONS 25

4.1 Introduction 25
4.2 Sample Size Calculations with Small Response Probabilities 26
4.3 A Bias Correction 32
4.4 Sample Size Calculations with General Response Probabilities 34
4.5 Empirical Methods 35
4.6 Simulation Study 37

CHAPTER Page

V ANALYSIS OF THE FIRE ANT DATA 61
5.1 Introduction 61
5.2 Background 61
5.3 Methodology for the Observational Study 63
5.4 Statistical Modeling 68
5.5 Data Analysis 70
5.6 Sample Size and Power for the Fire Ant Data 76
5.7 Conclusions 78
VI CONCLUSIONS 80
6.1 Summary 80
6.2 Possible Extensions 81
REFERENCES 82
APPENDIX A 85
APPENDIX B 88
VITA 98

LIST OF TABLES

TABLE Page

1 Sample sizes and powers, where sample sizes are for one-tailed test ($k = 2, s = 1, p = 1,\ e^{\theta_1} = 0.05$) 47

2 Sample sizes and powers, where sample sizes are for one-tailed test ($k = 2, s = 1, p = 1,\ e^{\theta_1} = 0.25$) 48

3 Sample sizes and powers, where sample sizes are for one-tailed test ($k = 2, s = 1, p = 1,\ e^{\theta_1} = 0.50$) 49

4 Sample sizes and powers, where sample sizes are for one-tailed test ($k = 2, s = 1, p = 1,\ e^{\theta_1} = 0.75$) 50

5 Sample sizes and power values in cases where the values of the test parameter are not small (one-tailed test, $k = 2, s = 1, p = 1,\ e^{\theta_1} = 0.25$) 51

6 Sample sizes and powers, where sample sizes are for one-tailed test ($k = 3, s = 1, p = 1,\ e^{\theta_1} = 0.05$, and $e^{\theta_1} = 0.15$) 54

7 Sample sizes and powers, where sample sizes are for one-tailed test ($k = 3, s = 1, p = 1,\ e^{\theta_1} = 0.5$, and $e^{\theta_2} = 7$) 57

8 Example of observations every half hour: From 9:30 a.m. to 5 p.m. on May 1, 2001 in College Station. 65

9 Frequencies of the combined variable MF. 66

10 Parameter estimates for the combined variable MF with adjusted predictor variables. 73

11 Example of estimated π's for the combined variable MF. The case $X_2^* = X_4^* = X_6^* = X_7 = 0$ implies that temperature is 82.1 ^oF, barometric pressure is 29.6, wind speed is 1.61, and humidity is 54.1%. Variable X_8 is for change of barometric pressure and X_9 is for rain. 75

12 Sample size for the combined variable MF. 77

TABLE Page

13 Sample size for the combined variable MF with adjusted predictor variables. 78

LIST OF FIGURES

FIGURE Page

1 Sample Size for fixed $e^{\theta_1} = 0.05$. 39

2 Sample Size for fixed $e^{\theta_1} = 0.25$. 40

3 Sample Size for fixed $e^{\theta_1} = 0.50$. 41

4 Sample Size for fixed $e^{\theta_1} = 0.75$. 42

5 Power for fixed $e^{\theta_1} = 0.05$. 43

6 Power for fixed $e^{\theta_1} = 0.25$. 44

7 Power for fixed $e^{\theta_1} = 0.50$. 45

8 Power for fixed $e^{\theta_1} = 0.75$. 46

9 Power values in cases where the values of the test parameter are not small ($e^{\theta_1} = 0.25$). 52

10 Sample Size for fixed $e^{\theta_1} = 0.05$ and $e^{\theta_2} = 0.15$ 55

11 Power for fixed $e^{\theta_1} = 0.05$ and $e^{\theta_2} = 0.15$ 56

12 Power for fixed $e^{\theta_1} = 0.5$ and $e^{\theta_2} = 7$ 59

13 Power for fixed $e^{\theta_1} = 0.5$ and $e^{\theta_2} = 7$ 60

14 Scatter plots of fire ant activities against temperature, humidity, barometric pressure, and wind speed. 71

CHAPTER I

INTRODUCTION

Logistic regression is broadly used in many scientific fields, such as biostatistics and epidemiology. It is a simple and effective method to describe the effects of some explanatory variables on a categorical response variable. There are many examples where the association between a binary response, such as “healthy” or “unhealthy”, and some covariates is desired. Standard logistic regression techniques play an important role in such cases. When the response variable has an ordinal nature, ordinal logistic regression is often a natural extension of standard logistic regression. A common ordinal logit model using cumulative logits considers a natural ordering of response categories. This model assumes a variable’s effect on the odds of response below category i is the same for all i. The odds ratio of cumulative probabilities in the expression is called a cumulative odds ratio. The log of the cumulative odds ratio is proportional to the distance between the values of the explanatory variables, with the same proportionality constant applying to each cut-point. Because of this property, this model is called a proportional odds model (McCullagh (1980); Agresti (1996)).

McCullagh (1980) developed and discussed a general class of regression models for ordinal data. The purpose of his paper was to investigate structural models appropriate to measurements on an ordinal scale. In particular, he introduced the most commonly used ordered logit model: the proportional odds. The proportional

The format and style follow that of Journal of the American Statistical Association.

odds model is widely useful in practice because its interpretation is simple and it has potentially greater power than multinomial logit models for ordered response variables.

Much literature exists on approximations to the power and sample size of different statistical tests within logistic regression model (Mehta, Patel, and Tsiatis (1984); Hilton and Mehta (1993); Lui (1993)). Whittemore (1981) considered sample size approximations in the case of standard logistic regression with small response probability. At present, sample size issues in ordinal logistic regression setting do not appear to have been studied in depth in the literature.

Our research was mainly motivated by the need of statistical analysis of our fire ant study. A study of the red imported fire ant mating flights is considered because of its structure and environmental importance (Callcott and Collins (1996); Greenberg, Vinson, and Ellison (1992)). The red imported fire ant is one of the most destructive insects in the US due particularly to its ability to disperse through mating flights (Porter, Bhatkar, Mulder, Vinson, and Clair (1991); Li and Heinz (1998)). Understanding of environmental cues triggering mating flights would be vital to a comprehensive control solution. The method of ordinal logistic regression modeling is employed to identify environmental factors associated with mating flights. Our statistical analyses indicate that ambient temperature, relative humidity, wind speed, barometric pressure, change in the barometric pressure, and recent rain are significant factors that trigger or influence fire ant mating flights. Rainfall one or two days prior is almost a necessity for mating flights to take place. A drop in barometric pressure, mild wind, temperature between 72-92 ^oF (22-33 ^oC) and humidity 30-69% are associated with high activity. Ordinal logistic regression is effective in identifying environmental cues triggering fire ant mating flights. Identification of these factors provides some important information about the reproduction process of the fire ant and will help us devise a lasting control strategy.

There are two objectives in this dissertation. The first one is to suggest sample size and power calculation methods for ordinal logistic regression to test statistical hypotheses. The second one is to fit a suitable model and check the reliability of the model using the red imported fire ant mating flights dataset. We will also apply the sample size calculation techniques to the fire ant mating flights data set.

One reason for our sample size consideration is that the proportional odds model is not a member of standard generalized linear models because of the multivariate nature of the response variable (Halekoh (2004)). Thus, the previous approaches (Self and Mauritsen (1988); Self, Mauritsen, and Ohara (1992)) for estimating power and sample size calculation within the framework of generalized linear models may not be suitable for ordinal logistic models. Therefore, we propose to develop sample size calculation methods within the proportional odds model structure (McCullagh (1980)). Such a sample size is needed to construct a test of hypotheses in ordinal logistic regression having a desired power. Whittemore (1981) considered a test for a single parameter with other parameters treated as nuisance parameters. Our approach first extends Whittemore's method in two directions within the proportional odds model: (a) when the probabilities of response categories are small and (b) testing for multiple parameters.

Another contribution of this research is to propose a method for the case where at least one of cumulative response probabilities is not small. Furthermore, we suggest an empirical method to calculate Fisher's information matrix without using the assumption of small response probabilities when a prior data set is available. The results are compared with those obtained by the Monte Carlo method and those in the previous literature.

The rest of the dissertation is organized as follows. In Chapter II, we introduce the logistic regression model and give a brief background of sample size calculation methods for the binary logistic regression. The concept of approximation by the moment generating function is briefly reviewed. The sample size calculation technique for the ordinal logistic regression is developed in Chapter III. We discuss the proposed method used in hypotheses testing problems. In Chapter IV, a modification of the sample size calculation for the logistic regression is introduced. We approximate the sample size calculation method by using the moment generating function. Some correction methods are also discussed to improve the performance. Furthermore, some results of simulation studies are illustrated. The results are compared with those obtained by Monte Carlo method and the previous literature. The design of the fire ants mating flights data is described in Chapter V. Some concluding remarks are given in Chapter VI. Some mathematical details are given in the Appendix.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

In this chapter we review the model and sample size calculation methods to test hypothesis about a parameter. A brief review of the model and likelihood function is given in Section 2.2. Sample size approximations and the closely related Fisher information matrix for the estimated parameters in a multiple covariate binary logistic regression are discussed in Section 2.3. That can be approximated by the augmented Hessian matrix of the moment-generating function for the covariates. The power and sample size calculations for generalized linear model are briefly reviewed in Section 2.4.

2.2 Model and Likelihood Function

For simplicity, in this research we consider the following proportional odds model to a response variable which has the ordinal nature.

2.2.1 The Proportional Odds Model

Suppose that the k ordered categories of the response have probabilities $\pi_1(\underline{x})$, $\pi_2(\underline{x})$, $\cdots$, $\pi_k(\underline{x})$ when the covariates have the value $\underline{x}$. Let Y be the response which takes values in the range $1, \cdots, k$ with the probabilities given above, and let $\kappa_j(\underline{x})$ be the odds that $Y \leq j$ given the covariate values $\underline{x}$, that is $\kappa_j(\underline{x}) = Pr(Y \leq j|\underline{x})/\{1 - Pr(Y \leq j|\underline{x})\}$. Then the proportional odds model is defined in McCullagh (1980) as

$$\kappa_j(\underline{x}) = \kappa_j \exp(\underline{\eta}'\underline{x}) \quad (1 \leq j < k),$$

where $\underline{\eta}$ is a vector of unknown parameters. The ratio of corresponding odds

$$\kappa_j(\underline{x}_1)/\kappa_j(\underline{x}_2) = \exp\{\underline{\eta}'(\underline{x}_1 - \underline{x}_2)\}, \quad (1 \leq j < k),$$

is independent of j and depends only on the difference between the covariate values, $\underline{x}_2 - \underline{x}_1$.

Since the odds for the event $Y \leq j$ is the ratio $\gamma_j(\underline{x})/\{1-\gamma_j(\underline{x})\}$, where $\gamma_j(\underline{X}) = \pi_1(\underline{X}) + \cdots + \pi_j(\underline{X})$, the proportional odds model is identical to the linear logistic model

$$\log[\gamma_j(\underline{x})/\{1 - \gamma_j(\underline{x})\}] = \theta_j + \underline{\eta}'\underline{x} \quad (1 \leq j < k),$$

with $\theta_j = \log\kappa_j$, so that the difference between corresponding cumulative logits is independent of the category involved (McCullagh (1980)).

2.2.2 Likelihood Function

Let the response cell counts be $\{n_{ij}\}$ with row totals n_1, n_2 and column or category totals $\{n_{.j}\}$. Let R_{ij} be the cumulative row sums, therefore $n_i = R_{ik}$ is the ith row total. Under the assumption of multinomial sampling in each row, the marginal distribution of R_{ij} conditional only on the row total n_i is binomial with index n_i and parameter γ_{ij}.

The contribution from a single multinomial observation $(n_1, \cdots, n_k)$ to the likelihood function is $\pi_1^{n_1} \cdots \pi_k^{n_k}$ with the probabilities π_j. Since we are dealing with cumulative probabilities, we define

$$\begin{aligned} R_1 &= n_1, \\ R_2 &= n_1 + n_2, \\ &\vdots \\ R_k &= \sum_{j=1}^{k} n_j = n. \end{aligned}$$

In terms of the parameters of the cumulative transformation, the likelihood can be written as the product of $k-1$ quantities (McCullagh (1980))

$$\left\{\left(\frac{\gamma_1}{\gamma_2}\right)^{R_1}\left(\frac{\gamma_2-\gamma_1}{\gamma_2}\right)^{R_2-R_1}\right\}\left\{\left(\frac{\gamma_2}{\gamma_3}\right)^{R_2}\left(\frac{\gamma_3-\gamma_2}{\gamma_3}\right)^{R_3-R_2}\right\}$$
$$\ldots\times\left\{\left(\frac{\gamma_{k-1}}{\gamma_k}\right)^{R_{k-1}}\left(\frac{\gamma_k-\gamma_{k-1}}{\gamma_k}\right)^{R_k-R_{k-1}}\right\}.$$

These factors are respectively the probability given R_2 that the first two cells divide in the ratio $R_1 : R_2-R_1$; the probability given R_3 that the proportion in cell 3 relative to cell 1 and 2 combined is $R_2 : R_3 - R_2$ and so on for the other components.

2.3 Sample Size Approximations

In this section, the approximation method of sample sizes needed to test hypotheses about η_1 with specified significance and power is given against given alternatives in the case when the probability of response is small in standard logistic regression (Whittemore (1981)). The calculations are based on a simple closed-form approximation to the asymptotic covariance matrix of the maximum likelihood estimates. We deal with studies in which a random sample is drawn from the joint distribution of (Y, X), where

- Y is a binary response,
- $X' = (X_1, \cdots, X_s)$ is a vector of covariates.

Assume that

- $Pr(\underline{X})$ is the conditional probability of response given $\underline{X} = \underline{x}$, that is, $Pr[Y = 1|\underline{X} = \underline{x}]$,
- $\text{logit}Pr(\underline{x}) = \theta_1 + \underline{\eta}'\underline{x}$.

For a given sample size n the likelihood of the observations y_v, $\underline{x}^{(v)}$, $v = 1, \cdots, n$ is

$$L(\theta_1, \underline{\eta}) = \prod_{v=1}^{n} f(\underline{x}^{(v)}) p(\underline{x}^{(v)})^{y_v} [1 - p(\underline{x}^{(v)})]^{1-y_v},$$

where $f(\underline{x})$ is the joint p.d.f. of $\underline{x}$ and it depends on none of the unknown parameters, $\theta_1, \underline{\eta}$. If the logistic model is valid, the maximum likelihood estimates $\hat{\theta}_1, \hat{\underline{\eta}}$ satisfy

$$(\hat{\theta}_1, \hat{\underline{\eta}}') \sim N((\theta_1, \underline{\eta}'), I^{-1}(\theta_1, \underline{\eta}'))$$

approximately. The (i, j)th entry of I is

$$\begin{aligned} I_{ij} &\equiv -E\left[\frac{\partial^2 logL}{\partial\eta_i \partial\eta_j}\right] \\ &= nE\left[X_i X_j \frac{e^{\theta_1 + \underline{\eta}'\underline{X}}}{(1 + e^{\theta_1 + \underline{\eta}'\underline{X}})^2}\right], \end{aligned} \tag{2.1}$$

where $i, j = 0, 1, \ldots, s$, $\eta_0 \equiv \theta_1$, and $X_0 \equiv 1$ and $X' = (X_1, \cdots, X_s)$. When the conditional response probability $p = Pr(\underline{X})$ is small, use of the binomial expansion $(1 - p)^{-1} = 1 + p + O(p^2)$ in (2.1) gives

$$I_{ij} \cong ne^{\theta_1} E[X_i X_j e^{\underline{\eta}'\underline{X}}]. \tag{2.2}$$

Let $m(\underline{\eta}) = E(e^{\underline{\eta}'\underline{X}})$ denote the moment-generating function of $\underline{X}$, with $m_i \equiv \partial m/\partial\eta_i$, $i = 1, \cdots, s$ and $m_{ij} = \partial^2 m/\partial\eta_i\partial\eta_j$, $i, j = 1, \cdots, s$. We extend this notation by defining $m_0 = m_{0,0} \equiv m$, and $m_{0,i} = m_{i,0} \equiv m_i$, $i = 1, \cdots, s$. Then (2.2) can be written

$$I_{ij} \cong ne^{\theta_1} m_{ij}(\underline{\eta}), \qquad i, j = 0, 1, \cdots, s. \tag{2.3}$$

To express (2.3) in matrix form, let $\mathbf{m^{(1)}}$ denote the s-dimensional column vector of first partials of m, and let $\mathbf{m^{(2)}}$ be the $s \times s$ Hessian matrix of second partials of m. We define the augmented Hessian of m to be the $(s + 1) \times (s + 1)$ matrix H defined by

$$H(\underline{\eta}) \equiv \begin{bmatrix} m & \mathbf{m^{(1)'}} \\ \mathbf{m^{(1)}} & \mathbf{m^{(2)}} \end{bmatrix}.$$

This enables us to write (2.3) as

$$I(\theta_1, \underline{\eta}) \cong ne^{\theta_1} H(\underline{\eta}), \tag{2.4}$$

Thus, the asymptotic covariance matrix of the estimates $\hat{\theta}_1$, $\underline{\hat{\eta}}'$ is approximately $[ne^{\theta_1} H(\underline{\eta})]^{-1}$ and the asymptotic variance of $\hat{\eta}_1$ is

$$\text{var}(\hat{\eta}_1) \simeq (ne^{\theta_1})^{-1} v(\underline{\eta}), \tag{2.5}$$

where $v(\underline{\eta})$ is the second diagonal entry of $H^{-1}(\underline{\eta})$.

To estimate the sample size needed to test at level α, with power$\geq 1 - \beta$, the hypothesis $\eta_1 = 0$ against the alternative $\eta_1 = \tilde{\eta}_1$, We use the approximation (2.5).

$$ne^{\theta_1} \geq [v^{1/2}(\underline{\eta}^0) z_\alpha + v^{1/2}(\underline{\tilde{\eta}}) z_\beta]^2 / \tilde{\eta}_1^2, \tag{2.6}$$

where $\underline{\eta}^0 = (0, \eta_2, \ldots, \eta_s)'$, $\underline{\tilde{\eta}} = (\tilde{\eta}_1, \eta_2, \ldots, \eta_s)'$, and z_c is the $100(1-c)$th percentile of the standard normal distribution.

2.3.1 The Multivariate Case

From the definition of $H(\underline{\eta})$, we can derive the second diagonal entry of $H(\underline{\eta})^{-1}$. When the distribution for $\underline{X}$ is of a general multivariate exponential type, we can estimate the sample size needed to achive a given power and significance level for tests about η_1. The moment-generating function for $\underline{X}$ is of the form (Bildikar and Patil (1968))

$$m(\underline{t}) = e^{q(\underline{r}+\underline{t}) - q(\underline{r})}, \tag{2.7}$$

where $\underline{r}$ is a vector of parameters and q is a real-valued function of s variables whose Hessian matrix of second derivatives exists and is positive definite. The mean of $\underline{X}$ is given by the vector $\mathbf{q}^{(1)}(\underline{r})$ of first partials of q, evaluated at $\underline{r}$, and the variance of $\underline{X}$ is given by Hessian $\mathbf{q}^{(2)}(\underline{r})$. (A.1) (in the Appendix A) shows that $v(\underline{\eta})$ is $e^{q(\underline{r}) - q(\underline{r}+\underline{\eta})}$

times the first diagonal entry of the inverse of $\mathbf{q}^{(2)}$, evaluated $\underline{r} + \underline{\eta}$:

$$v(\underline{\eta}) = e^{q(r)-q(r+\underline{\eta})}[\mathbf{q}^{(\mathbf{2})}(\mathbf{r} + \underline{\eta})]^{-1}_{11}. \tag{2.8}$$

2.4 Power and Sample Size Calculations for Generalized Linear Models

A sample size and power estimation is described within the framework of generalized linear models (Self and Mauritsen (1988)). This approach is based on the score test under contiguous alternatives and is applicable to tests of composite null hypotheses. Commonly used test statistics can be derived as score statistics from within the framework of generalized linear models (GLM). In the GLM independent scalar response variables, $Y_1, \cdots, Y_n$, are assumed to have probability density functions of the form (Gart and Tarone (1983))

$$\exp[Y_i\theta_i - b(\theta_i) + c(Y_i)], \tag{2.9}$$

where the canonical parameter θ_i has a relationship with the expected value of Y_i, $\mu_i = b'(\theta_i)$, where b' denotes the first derivative of b. The μ_i is related to linear predictors, θ_i, by the link function g in the expression $\eta_i = g(\mu_i)$. The θ_i are assumed to have the following linearly parameterized form

$$\theta_i = Z_i'\psi + X_i'\lambda, \tag{2.10}$$

where Z_i is a p-vector of covariates, X_i is a q-vector of covariates, and ψ and λ represent the associated unknown regression coefficients. We want to test the hypothesis $\psi = \psi_0$ while treating λ as a nuisance parameter. The score test is based on the statistic $S_{n\psi}(\psi_0, \hat{\lambda}_0)$, where λ_0 is a solution to the equation $S_{n\lambda}(\psi_0, \lambda) = 0$ and $S_{n\psi}$ and $S_{n\lambda}$ represent derivatives of the log-likelihood function with respect to ψ and λ, respectively. The test statistic is a quadratic form in $S_{n\psi}(\psi_0, \hat{\lambda}_0)$ and is referred to its asymptotic distribution under the null hypothesis, which is a central chi-square distribution on p degrees of freedom.

2.4.1 An Approximation to the Power of the Score Test

The score test statistic is computed as the quadratic form

$$T_n = S'_{n\psi}(\psi_0, \hat{\lambda}_0) V_n^{-1} S_{n\psi}(\psi_0, \hat{\lambda}_0), \tag{2.11}$$

where V_n represents an estimate of the covariance matrix of $S_{n\psi}(\psi_0, \hat{\lambda}_0)$. In order to approximate the distribution T_n under alternative hypotheses, the limiting destribution of $S_{n\psi}(\psi_0, \hat{\lambda}_0)$ is described. Note that $\hat{\lambda}_0$ is generally not a consistent estimator of λ. It converges to some value λ_0^* which is defined as the solution to the equation

$$lim_{n\to\infty} n^{-1} E[S_{n\psi}(\psi_0, \lambda)] = 0. \tag{2.12}$$

Taylor series arguments are used to obtain an approximation to $S_{n\psi}(\psi_0, \hat{\lambda}_0)$ for which the error is $O_p(n^{-1})$. This approximation is given by

$$S_{n\psi}(\psi_0, \hat{\lambda}_0) \approx S_{n\psi}(\psi_0, \lambda_0^*) - I^*_{n\psi\lambda} {I^*_{n\lambda\lambda}}^{-1} S_{n\psi}(\psi_0, \lambda_0^*), \tag{2.13}$$

where $I^*_{n\psi\lambda}$ and $I^*_{n\lambda\lambda}$ represent elements of the expected information matrix evaluated at ψ_0 and λ_0^*. Let ξ_n and Σ_n denote the mean and covariance matrix of (2.13), respectively.

Based on the limiting normality of $S_{n\psi}(\psi_0, \hat{\lambda}_0)$ the distribution T_n is approximate by a chi-square distribution on p degrees of freedom with noncentrality parameter, γ_n, given by

$$\xi'_n \Sigma_n^{-1} \xi_n. \tag{2.14}$$

In the special case of generalized linear models, the form of $S_{n\psi}$ and $S_{n\lambda}$ implied by (2.11) may be used to write expression (2.13) as

$$\sum_{i=1}^{n} [Y_i - \mu_i(\psi_0, \lambda_0^*)] \Delta_i Z_i^*, \tag{2.15}$$

where Δ_i is the first derivative of θ_i with respect to η_i, Z_i^* represents $Z_i - I_{n\psi\lambda}^* {I_{n\lambda\lambda}^*}^{-1} X_i$, and μ_i is written to show explicitly its dependence on ψ and λ. It follows that the expected value, ξ_n, and covariance matrix, Σ_n, of (2.13) are given by the expressions

$$\sum_{i=1}^{n} [\mu_i(\psi, \lambda) - \mu_i(\psi_0, \lambda_0^*)] \Delta_i Z_i^*, \tag{2.16}$$

and

$$\sum_{i=1}^{n} v_i(\psi, \lambda) \Delta_i^2 Z_i^* Z_i^{*\prime}, \tag{2.17}$$

where $v_i(\psi, \lambda) = \text{var}(y_i)$.

2.4.2 An Implementation

Self and Mauritsen (1988) made the simplifying assumption that all of the covariates in the model are categorical. As the first step in the calculations, values for the parameters ψ, λ, and $\{\Pi_i; i = 1, \cdots, m\}$ are specified. If the average response is specified in addition to the regression parameters, then the intercept parameter in the model is found as the solution to the equation

$$\bar{\mu} = \sum_{i=1}^{m} \Pi_i \mu_i(\psi, \lambda), \tag{2.18}$$

where $\bar{\mu}$ denotes the population mean response. If parameter values for ψ and λ are chosen, λ_0^* is computed as the solution to the equation

$$\sum_{i=1}^{m} \Pi_i [\mu_i(\psi, \lambda) - \mu_i(\psi_0, \lambda_0^*)] \Delta_i X_i = 0. \tag{2.19}$$

Solving equation (2.15) is equivalent to fitting the null model in a weighted analysis where the data are taken to be $\mu_i(\psi, \lambda); i = 1, \cdots, m\}$. In a case of logistic regression, the response variable, Y_i, follows a Bernoulli distribution with probability of response μ_i. The link function is the logistic function $\log \frac{(\mu_i)}{(1 - \mu_i)}$. In order to compute the intercept parameter given values for the other parameters and a value for the overall probability of response, equation (2.15) is solved with the iteration scheme

$$\begin{aligned} \bar{\mu} &= \sum_{i=1}^{m} \Pi_i \mu_i(\psi, \lambda) \\ &= \sum_{i=1}^{m} \Pi_i \frac{\exp(\alpha + Z_i'\psi + X_i'\lambda)}{1 + \exp(\alpha + Z_i'\psi + X_i'\lambda)} \\ &= \exp(\alpha) \sum_{i=1}^{m} \Pi_i \frac{\exp(Z_i'\psi + X_i'\lambda)}{1 + \exp(\alpha + Z_i'\psi + X_i'\lambda)}, \end{aligned}$$

and

$$\alpha^{(k)} = \log(\bar{\mu}) - \log \left\{ \sum_{i=1}^{m} \Pi_i \frac{\exp(Z_i'\psi + X_i'\lambda)}{1 + \exp(\alpha^{(k-1)} + Z_i'\psi + X_i'\lambda)} \right\}.$$

The main emphasis of the paper (Self and Mauritsen (1988)) in this section is estimating sample size within the framework of GLM. However, as mentioned in Chapter I our ordinal regression study does not appear to be suitable, but the score test method can be applied to improve our approximation methods by contiguous alternatives.

CHAPTER III

ORDINAL LOGISTIC REGRESSION

3.1 Introduction

As discussed in Chapter I, sample size issues in ordinal logistic regression setting do not appear to have been investigated. In this chapter we introduce a sample size calculation method for ordinal logistic regression by the large sample properties of the maximum likelihood estimate. In Section 3.2, the ordinal logistic model is mentioned again. The likelihood function of the ordinal logistic regression and a brief review of the maximum likelihood estimate are given in Section 3.3. To obtain the asymptotic variance, the derivation of the Fisher information matrix is given in Section 3.4. The sample size calculation for the ordinal logistic regression to test the parameters is given in Section 3.5.

3.2 Model

We deal with studies in which a random sample is drawn from the joint distribution of $(Y, \underline{X})$, where Y is an ordinal response and $\underline{X}' = (X_1, \ldots, X_s)$ is a vector of covariates. Let $\pi_j(\underline{X})$ denote the classification probabilities $Pr(Y = j|\underline{X})$ of response variable Y, $j = 1, 2, \ldots, k$ at value $\underline{X}' = (X_1, X_2, \ldots, X_s)$ for a set of explanatory variables $X_1, X_2, \ldots, X_s$. Here our interest is centered on the problem of relating $\underline{\pi}' = (\pi_1(\underline{X}), \pi_2(\underline{X}), \ldots, \pi_k(\underline{X}))$ to the predictor $\underline{X}$.

Since our response categories have a natural ordering, logit models should utilize that ordering. We use the proportional-odds model that is described below. The ordered multiple response models assume the relationship

$$\text{logit}[Pr(Y \leq j|\underline{X})] = \theta_j + \underline{\eta}'\underline{X}, \qquad j = 1, 2, \ldots, k-1,$$

where $\underline{\theta}$ is a vector of the intercept parameters and $\underline{\eta}' = (\eta_1, \eta_2, \ldots, \eta_s)$ is the slope parameter vector not including the intercept term. By construction, $\theta_j < \theta_{j+1}$ holds. This model fits a common slopes cumulative model that is a parallel lines regression model based on the cumulative probabilities of the response categories.

Let $\gamma_j(\underline{X}) = \pi_1(\underline{X}) + \ldots + \pi_j(\underline{X})$. Then $\gamma_1(\underline{X}) = \pi_1(\underline{X})$, $\gamma_2(\underline{X}) = \pi_1(\underline{X}) + \pi_2(\underline{X})$, and $\gamma_k(\underline{X}) = \pi_1(\underline{X}) + \ldots + \pi_k(\underline{X}) = 1$. The ordinal logistic regression model in our setting is given as follows:

$$\begin{aligned}
\text{logit}(\gamma_1) &= \log\left(\frac{\gamma_1}{1-\gamma_1}\right) = \theta_1 + \eta_1 X_1 + \eta_2 X_2 + \ldots + \eta_s X_s, \\
\text{logit}(\gamma_2) &= \log\left(\frac{\gamma_2}{1-\gamma_2}\right) = \theta_2 + \eta_1 X_1 + \eta_2 X_2 + \ldots + \eta_s X_s, \\
&\ldots \\
\text{logit}(\gamma_{k-1}) &= \log\left(\frac{\gamma_{k-1}}{1-\gamma_{k-1}}\right) = \theta_{k-1} + \eta_1 X_1 + \eta_2 X_2 + \ldots + \eta_s X_s,
\end{aligned}$$

where

$$\gamma_j(\underline{X}) = \pi_1(\underline{X}) + \pi_2(\underline{X}) + \ldots + \pi_j(\underline{X}) = \frac{e^{\theta_j + \underline{\eta}'\underline{X}}}{1 + e^{\theta_j + \underline{\eta}'\underline{X}}}, \quad j = 1, \ldots, k-1, \tag{3.1}$$

and $\gamma_k = 1$. This model is known as the proportional-odds model because the odds-ratio of the event $(Y \leq j)$ is independent of the category indicator.

3.3 Likelihood Function

When more than one observation on Y occurs at a fixed $x^{(v)}$ value, it is sufficient to record the number of observations $n_j^{(v)}$ and the number of "j" outcomes, for $j = 1, \ldots, k$. Thus we let $Y^{(v)}$ refer to these counts rather than to individual binary responses. The $\{Y^{(v)}, v = 1, \ldots, n\}$ are independent multinomial random variables $Y^{(v)} \sim \textit{multinomial}(n_1^{(v)}, \ldots, n_k^{(v)})$ with $E(Y^{(v)}) = n_j^{(v)} \gamma_j(\underline{x}^{(v)})$, where $n_1^{(v)} + \ldots +$

$n_k^{(v)} = 1$. We define

$$\begin{aligned}
R_1^{(v)} &= n_1^{(v)}, \\
R_2^{(v)} &= n_1^{(v)} + n_2^{(v)}, \\
&\vdots \\
R_k^{(v)} &= 1.
\end{aligned}$$

Since we are dealing with cumulative probabilities, in terms of the parameters of the cumulative transformation, the likelihood can be written as the product of $k-1$ quantities. The joint probability mass function of $(Y_1, \ldots, Y_n)$ is proportional to the product of n multinomial functions.

3.3.1 Likelihood Function

For a given sample size n the likelihood of the observations $y^{(v)}, \underline{x}^{(v)}$, $v = 1, \ldots, n$ is

$$\begin{aligned}
L(\underline{\theta}, \underline{\eta}) &= \prod_{v=1}^{n} f(\underline{x}^{(v)}) f(y^{(v)} | \underline{x}^{(v)}) \\
&\propto \prod_{v=1}^{n} f(y^{(v)} | \underline{x}^{(v)}) \\
&= \prod_{v=1}^{n} \left\{ \left(\frac{\gamma_1^{(v)}}{\gamma_2^{(v)}} \right)^{R_1^{(v)}} \left(\frac{\gamma_2^{(v)} - \gamma_1^{(v)}}{\gamma_2^{(v)}} \right)^{R_2^{(v)} - R_1^{(v)}} \right\} \\
&\quad \times \left\{ \left(\frac{\gamma_2^{(v)}}{\gamma_3^{(v)}} \right)^{R_2^{(v)}} \left(\frac{\gamma_3^{(v)} - \gamma_2^{(v)}}{\gamma_3^{(v)}} \right)^{R_3^{(v)} - R_2^{(v)}} \right\} \cdots \\
&\quad \times \left\{ \left(\frac{\gamma_{k-1}^{(v)}}{\gamma_k^{(v)}} \right)^{R_{k-1}^{(v)}} \left(\frac{\gamma_k^{(v)} - \gamma_{k-1}^{(v)}}{\gamma_k^{(v)}} \right)^{R_k^{(v)} - R_{k-1}^{(v)}} \right\}, \qquad (3.2)
\end{aligned}$$

where $f(\underline{x})$ is the joint p.d.f. of $\underline{x}$. It is assumed that $f(\underline{x})$ does not depend on unknown parameters, $(\underline{\theta}', \underline{\eta}')$.

If the logistic model is valid, the maximum likelihood estimates $\hat{\underline{\theta}}', \hat{\underline{\eta}}'$ satisfy

$$\hat{\underline{\theta}}', \hat{\underline{\eta}}' \sim N((\underline{\theta}', \underline{\eta}'), I^{-1}(\underline{\theta}', \underline{\eta}'))$$

approximately.

3.4 The Fisher Information Matrix

To derive the information matrix, we have the following derivations. First, the log-likelihood is

$$\begin{aligned}
\log \mathrm{L} &= \sum_{v=1}^{n} \Big\{ R_{v1}\log(\gamma_{v1}) + (R_{v2}-R_{v1})\log(\gamma_{v2}-\gamma_{v1}) - R_{v2}\log(\gamma_{v2}) \\
&\quad + R_{v2}\log(\gamma_{v2}) + (R_{v3}-R_{v2})\log(\gamma_{v3}-\gamma_{v2}) - R_{v3}\log(\gamma_{v3}) + \ldots \\
&\quad + R_{v,k-1}\log(\gamma_{v,k-1}) + (R_{v,k}-R_{v,k-1})\log(\gamma_{v,k}-\gamma_{v,k-1}) - R_{v,k}\log(\gamma_{v,k}) \Big\} \\
&= \sum_{v=1}^{n} \Big\{ R_{v1}\log(\gamma_{v1}) + (R_{v2}-R_{v1})\log(\gamma_{v2}-\gamma_{v1}) \\
&\quad + (R_{v3}-R_{v2})\log(\gamma_{v3}-\gamma_{v2}) + \ldots + (R_{v,k}-R_{v,k-1})\log(\gamma_{v,k}-\gamma_{v,k-1}) \Big\} \\
&= \sum_{v=1}^{n} \Bigg\{ R_{v1}\log\bigg(\frac{e^{\theta_1+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_1+\underline{\eta}'\underline{x}_v}}\bigg) \\
&\quad + (R_{v2}-R_{v1})\log\bigg(\frac{e^{\theta_2+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_2+\underline{\eta}'\underline{x}_v}} - \frac{e^{\theta_1+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_1+\underline{\eta}'\underline{x}_v}}\bigg) \\
&\quad + (R_{v3}-R_{v2})\log\bigg(\frac{e^{\theta_3+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_3+\underline{\eta}'\underline{x}_v}} - \frac{e^{\theta_2+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_2+\underline{\eta}'\underline{x}_v}}\bigg) \\
&\quad + \ldots + (R_{v,k}-R_{v,k-1})\log\bigg(1 - \frac{e^{\theta_{k-1}+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_{k-1}+\underline{\eta}'\underline{x}_v}}\bigg) \Bigg\}.
\end{aligned}$$

Since

$$\begin{aligned}
\log\bigg(\frac{e^{\theta_2+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_2+\underline{\eta}'\underline{x}_v}} - \frac{e^{\theta_1+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_1+\underline{\eta}'\underline{x}_v}}\bigg) &= \log\frac{e^{\theta_2+\underline{\eta}'\underline{x}_v} - e^{\theta_1+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}_v})(1+e^{\theta_1+\underline{\eta}'\underline{x}_v})} \\
&= \log\frac{e^{\underline{\eta}'\underline{x}_v}(e^{\theta_2-\theta_1})}{(1+e^{\theta_2+\underline{\eta}'\underline{x}_v})(1+e^{\theta_1+\underline{\eta}'\underline{x}_v})},
\end{aligned}$$

the log-likelihood function is

$$\begin{aligned}
\log \mathrm{L} = \sum_{v=1}^{n} \Bigg\{ & R_{v1}\Bigg(\theta_1 + \underline{\eta}'\underline{x}_v - \log[1+e^{\theta_1+\underline{\eta}'\underline{x}_v}]\Bigg) \\
& +(R_{v2}-R_{v1})\Bigg(\underline{\eta}'\underline{x}_v + \log(e^{\theta_2}-e^{\theta_1}) - \log(1+e^{\theta_2+\underline{\eta}'\underline{x}_v}) - \log(1+e^{\theta_1+\underline{\eta}'\underline{x}_v})\Bigg) \\
& +(R_{v3}-R_{v2})\Bigg(\underline{\eta}'\underline{x}_v + \log(e^{\theta_3}-e^{\theta_2}) - \log(1+e^{\theta_3+\underline{\eta}'\underline{x}_v}) - \log(1+e^{\theta_2+\underline{\eta}'\underline{x}_v})\Bigg) \\
& +\ldots-(1-R_{v,k-1})\log(1+e^{\theta_{k-1}+\underline{\eta}'\underline{x}_v})\Bigg\}.
\end{aligned}$$

From this log-likelihood, we calculate derivatives

$$\begin{aligned}
\frac{\partial \log \mathrm{L}}{\partial \theta_1} &= \sum_{v=1}^{n}\Bigg\{R_{v1}\Bigg(1-\frac{e^{\theta_1+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_1+\underline{\eta}'\underline{x}_v}}\Bigg) \\
&\qquad +(R_{v2}-R_{v1})\Bigg(-\frac{e^{\theta_1}}{e^{\theta_2}-e^{\theta_1}}-\frac{e^{\theta_1+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_1+\underline{\eta}'\underline{x}_v}}\Bigg)\Bigg\},
\end{aligned}$$

$$\frac{\partial^2 \log \mathrm{L}}{\partial \theta_1^2} = \sum_{v=1}^{n}\Bigg\{-R_{v2}\frac{e^{\theta_1+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}_v})^2}-(R_{v2}-R_{v1})\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2}\Bigg\},$$

$$\frac{\partial^2 \log \mathrm{L}}{\partial \theta_1 \partial \theta_2} = \sum_{v=1}^{n}\Bigg\{(R_{v2}-R_{v1})\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2}\Bigg\},$$

$$\frac{\partial^2 \log \mathrm{L}}{\partial \theta_1 \partial \theta_j} = 0, \quad j=3,\ldots,k-1,$$

$$\frac{\partial^2 \log \mathrm{L}}{\partial \theta_1 \partial \eta_j} = \sum_{v=1}^{n}\Bigg\{-R_{v2}\frac{X_{vj}e^{\theta_1+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}_v})^2}\Bigg\}, \quad j=2,\ldots,k-2,$$

$$\begin{aligned}
\frac{\partial \log \mathrm{L}}{\partial \theta_i} &= \sum_{v=1}^{n}\Bigg\{(R_{v,i}-R_{v,i-1})\Bigg(\frac{e^{\theta_i}}{e^{\theta_i}-e^{\theta_{i-1}}}-\frac{e^{\theta_i+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_i+\underline{\eta}'\underline{x}_v}}\Bigg) \\
&\qquad +(R_{v,i+1}-R_{v,i})\Bigg(-\frac{e^{\theta_i}}{e^{\theta_{i+1}}-e^{\theta_i}}-\frac{e^{\theta_i+\underline{\eta}'\underline{x}_v}}{1+e^{\theta_i+\underline{\eta}'\underline{x}_v}}\Bigg)\Bigg\},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \log \mathrm{L}}{\partial \theta_i^2} &= \sum_{v=1}^{n}\Bigg\{-(R_{v,i+1}-R_{v,i-1})\frac{e^{\theta_i+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_i+\underline{\eta}'\underline{x}_v})^2} \\
&\qquad -(R_{v,i}-R_{v,i-1})\frac{e^{\theta_i+\theta_{i-1}}}{(e^{\theta_i}-e^{\theta_{i-1}})^2}-(R_{v,i+1}-R_{v,i})\frac{e^{\theta_i+\theta_{i+1}}}{(e^{\theta_{i+1}}-e^{\theta_i})^2}\Bigg\}, \\
&\qquad i=2,\ldots,k-2,
\end{aligned}$$

$$\frac{\partial^2 \text{log L}}{\partial\theta_i \partial\theta_{i-1}} = \sum_{v=1}^{n} \left\{ (R_{v,i} - R_{v,i-1}) \frac{e^{\theta_i + \theta_{i-1}}}{(e^{\theta_i} - e^{\theta_{i-1}})^2} \right\}, \quad i = 2, \ldots, k-1,$$

$$\frac{\partial^2 \text{log L}}{\partial\theta_i \partial\theta_j} = 0, \quad |i - j| \geq 2,$$

$$\frac{\partial^2 \text{log L}}{\partial\theta_i \eta_j} = \sum_{v=1}^{n} \left\{ -(R_{v,i+1} - R_{v,i-1}) \frac{X_{vj} e^{\theta_i + \underline{\eta}'\underline{x}_v}}{(1 + e^{\theta_i + \underline{\eta}'\underline{x}_v})^2} \right\}, \quad i = 2, \ldots, k-2,$$

$$\frac{\partial \text{log L}}{\partial\theta_{k-1}} = \sum_{v=1}^{n} \left\{ (R_{v,k-1} - R_{v,k-2}) \left(\frac{e^{\theta_{k-1}}}{e^{\theta_{k-1}} - e^{\theta_{k-2}}} - \frac{e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v}}{1 + e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v}} \right) \right.$$
$$\left. - (1 - R_{v,k-1}) \left(\frac{e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v}}{1 + e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v}} \right) \right\},$$

$$\frac{\partial^2 \text{log L}}{\partial\theta_{k-1}^2} = \sum_{v=1}^{n} \left\{ -(R_{v,k-1} - R_{v,k-2}) \frac{e^{\theta_{k-1} + \theta_{k-2}}}{(e^{\theta_{k-1}} - e^{\theta_{k-2}})^2} \right.$$
$$\left. - (1 - R_{v,k-2}) \frac{e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v}}{(1 + e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v})^2} \right\},$$

$$\frac{\partial^2 \text{log L}}{\partial\theta_{k-1} \partial\eta_j} = \sum_{v=1}^{n} \left\{ -(1 - R_{v,k-2}) \frac{X_{vj} e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v}}{(1 + e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v})^2} \right\},$$

$$\begin{aligned} \frac{\partial \text{log L}}{\partial\eta_i} &= \sum_{v=1}^{n} \left\{ R_{v1} \left(X_{vi} - \frac{X_{vi} e^{\theta_1 + \underline{\eta}'\underline{x}_v}}{1 + e^{\theta_1 + \underline{\eta}'\underline{x}_v}} \right) + (R_{v2} - R_{v1}) \right. \\ & \times \left(X_{vi} - \frac{X_{vi} e^{\theta_1 + \underline{\eta}'\underline{x}_v}}{1 + e^{\theta_1 + \underline{\eta}'\underline{x}_v}} - \frac{X_{vi} e^{\theta_2 + \underline{\eta}'\underline{x}_v}}{1 + e^{\theta_2 + \underline{\eta}'\underline{x}_v}} \right) + \ldots \\ & \left. + (1 - R_{v,k-1}) \left(- \frac{X_{vi} e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v}}{1 + e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v}} \right) \right\}, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 \text{log L}}{\partial\eta_i \partial\eta_j} &= \sum_{v=1}^{n} \left\{ R_{v1} \left(- \frac{X_{vi} X_{vj} e^{\theta_1 + \underline{\eta}'\underline{x}_v}}{(1 + e^{\theta_1 + \underline{\eta}'\underline{x}_v})^2} \right) \right. \\ & + (R_{v2} - R_{v1}) \left(- \frac{X_{vi} X_{vj} e^{\theta_1 + \underline{\eta}'\underline{x}_v}}{(1 + e^{\theta_1 + \underline{\eta}'\underline{x}_v})^2} - \frac{X_{vi} X_{vj} e^{\theta_2 + \underline{\eta}'\underline{x}_v}}{(1 + e^{\theta_2 + \underline{\eta}'\underline{x}_v})^2} \right) \\ & \left. + \ldots + (1 - R_{v,k-1}) \left(- \frac{X_{vi} X_{vj} e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v}}{(1 + e^{\theta_{k-1} + \underline{\eta}'\underline{x}_v})^2} \right) \right\}. \end{aligned}$$

Thus, the (i,j)th entry of the $(k+s-1)\times(k+s-1)$ Fisher information matrix I is

$$\begin{aligned} I_{11} &= -\mathrm{E}\Bigg[\frac{\partial^2 \log \mathrm{L}}{\partial\theta_1^2}\Bigg] \\ &= n\mathrm{E}\Bigg\{R_2\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}+(R_2-R_1)\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2}\Bigg\}, \end{aligned}$$

where

$$f(x,R)=f(R|x)f(x),$$

$$\begin{aligned} E(g(X)h(R)) &= \int_x\int_R g(X)h(R)f(x,R)dRdx \\ &= \int_x\int_R g(X)h(R)f(R|x)f(x)dRdx \\ &= \int_x g(X)f(x)\left\{\int_R h(R)f(R|x)dR\right\}dx, \end{aligned}$$

$E(R_1|x)=\gamma_1=\dfrac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})}$, $E(R_2|x)=\gamma_2=\dfrac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})}$, and

$$\begin{aligned} \mathrm{E}\Bigg\{R_2\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}\Bigg\} &= \int_x \frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}f(x)\left\{\int_R R_2 f(R|x)dR\right\}dx \\ &= \int_x \frac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})}\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}f(x)dx. \end{aligned}$$

By using these

$$\begin{aligned} I_{11} &= n\mathrm{E}\Bigg\{R_2\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}+(R_2-R_1)\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2}\Bigg\} \\ &= n\mathrm{E}\Bigg\{\frac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})}\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}\Bigg\} \\ &\quad +n\mathrm{E}\Bigg\{\frac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})}-\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})}\Bigg\}\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2}, \end{aligned}$$

$$I_{12}=-n\mathrm{E}\Bigg\{\frac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})}-\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})}\Bigg\}\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2},$$

$$I_{1,j+k-1}=n\mathrm{E}\Bigg\{X_j\frac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})}\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}\Bigg\},\quad j=1,\ldots,s,$$

$$\begin{aligned}
I_{ii} &= -\mathrm{E}\left\{\frac{\partial^2 \log \mathrm{L}}{\partial\theta_i^2}\right\} \\
&= n\mathrm{E}\Bigg\{\left(\frac{e^{\theta_{i+1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{i+1}+\underline{\eta}'\underline{x}})}-\frac{e^{\theta_{i-1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{i-1}+\underline{\eta}'\underline{x}})}\right)\frac{e^{\theta_i+\underline{\eta}'\underline{x}}}{(1+e^{\theta_i+\underline{\eta}'\underline{x}})^2} \\
&\quad+\left(\frac{e^{\theta_i+\underline{\eta}'\underline{x}}}{(1+e^{\theta_i+\underline{\eta}'\underline{x}})}-\frac{e^{\theta_{i-1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{i-1}+\underline{\eta}'\underline{x}})}\right)\frac{e^{\theta_i+\theta_{i-1}}}{(e^{\theta_i}-e^{\theta_{i-1}})^2} \\
&\quad+\left(\frac{e^{\theta_{i+1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{i+1}+\underline{\eta}'\underline{x}})}-\frac{e^{\theta_i+\underline{\eta}'\underline{x}}}{(1+e^{\theta_i+\underline{\eta}'\underline{x}})}\right)\frac{e^{\theta_i+\theta_{i+1}}}{(e^{\theta_{i+1}}-e^{\theta_i})^2}\Bigg\}, \quad 2 \le i \le k-2,
\end{aligned}$$

$$I_{i,i-1} = -n\mathrm{E}\left(\frac{e^{\theta_i+\underline{\eta}'\underline{x}}}{(1+e^{\theta_i+\underline{\eta}'\underline{x}})}-\frac{e^{\theta_{i-1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{i-1}+\underline{\eta}'\underline{x}})}\right)\frac{e^{\theta_i+\theta_{i-1}}}{(e^{\theta_i}-e^{\theta_{i-1}})^2}, \quad 2 \le i \le k-1,$$

$$I_{ij} = 0, \quad \text{if } |i-j| \ge 2 \text{ and } i,j = 1,\ldots k-1,$$

$$\begin{aligned}
I_{i,j+k-1} &= n\mathrm{E}\left\{X_j\left(\frac{e^{\theta_{i+1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{i+1}+\underline{\eta}'\underline{x}})}-\frac{e^{\theta_{i-1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{i-1}+\underline{\eta}'\underline{x}})}\right)\frac{e^{\theta_i+\underline{\eta}'\underline{x}}}{(1+e^{\theta_i+\underline{\eta}'\underline{x}})^2}\right\}, \\
&\quad \text{if } i = 2,\ldots,k-2 \text{ and } j = 1,\ldots,s.
\end{aligned}$$

$$\begin{aligned}
I_{k-1,k-1} &= n\mathrm{E}\Bigg\{\left(\frac{e^{\theta_{k-1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{k-1}+\underline{\eta}'\underline{x}})}-\frac{e^{\theta_{k-2}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{k-2}+\underline{\eta}'\underline{x}})}\right)\frac{e^{\theta_{k-1}+\theta_{k-2}}}{(e^{\theta_{k-1}}-e^{\theta_{k-2}})^2} \\
&\quad+\left(1-\frac{e^{\theta_{k-2}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{k-2}+\underline{\eta}'\underline{x}})}\right)\frac{e^{\theta_{k-1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{k-1}+\underline{\eta}'\underline{x}})^2}\Bigg\},
\end{aligned}$$

$$I_{k-1,j+k-1} = n\mathrm{E}\left\{X_j\left(1-\frac{e^{\theta_{k-2}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{k-2}+\underline{\eta}'\underline{x}})}\right)\frac{e^{\theta_{k-1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{k-1}+\underline{\eta}'\underline{x}})^2}\right\}, \quad j = 1,\ldots,s,$$

and

$$\begin{aligned}
I_{i+k-1,j+k-1} &= -\mathrm{E}\left[\frac{\partial^2 \log \mathrm{L}}{\partial\eta_i\partial\eta_j}\right] \\
&= n\mathrm{E}\Bigg\{X_iX_j\Bigg[\gamma_2\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}+(\gamma_3-\gamma_1)\frac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})^2} \\
&\quad+\ldots+(1-\gamma_{k-2})\frac{e^{\theta_{k-1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{k-1}+\underline{\eta}'\underline{x}})^2}\Bigg]\Bigg\} \\
&= n\mathrm{E}\Bigg\{X_iX_j\Bigg[\frac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})}\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2} \\
&\quad+\left(\frac{e^{\theta_3+\underline{\eta}'\underline{x}}}{1+e^{\theta_3+\underline{\eta}'\underline{x}}}-\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{1+e^{\theta_1+\underline{\eta}'\underline{x}}}\right)\frac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})^2} \\
&\quad+\ldots+\left(1-\frac{e^{\theta_{k-2}+\underline{\eta}'\underline{x}}}{1+e^{\theta_{k-2}+\underline{\eta}'\underline{x}}}\right)\frac{e^{\theta_{k-1}+\underline{\eta}'\underline{x}}}{(1+e^{\theta_{k-1}+\underline{\eta}'\underline{x}})^2}\Bigg]\Bigg\}, \\
&\quad i,j = 1,\ldots,s.
\end{aligned}$$

The estimated asymptotic covariance matrix of the estimates $(\underline{\hat{\theta}}', \underline{\hat{\eta}}')$ is $I^{-1}(\underline{\hat{\theta}}', \underline{\hat{\eta}}')$.

3.5 Sample Size Calculations

The goal of a sample size calculation is to obtain a sample that is just sufficiently large enough to be confident of being able to obtain an inference with required precision. It is very important because it is directly related to the cost and time involved in a survey or data collection.

Consider the problem of testing the null hypothesis $H_0 : \eta = 0$ against the one-sided alternative $H_A : \eta = \tilde{\eta}$ to test at level α with power $\geq 1 - \beta$. When the distribution of $\hat{\eta}$ is treated as normal with mean η and variance σ^2. The critical region is

$$\begin{cases} \hat{\eta} > z_\alpha \left(\dfrac{\sigma_0}{\sqrt{n}} \right), & \text{if } \tilde{\eta} > 0, \\ \hat{\eta} < -z_\alpha \left(\dfrac{\sigma_0}{\sqrt{n}} \right), & \text{if } \tilde{\eta} < 0, \end{cases}$$

where z_α is the $100(1 - \alpha)$th percentile of the standard normal distribution. The sample size n will be found so that the test has a specified power $1 - \beta$ at the alternative $H_A : \eta = \tilde{\eta}$. The sample size n is thus chosen so that

$$Pr(\hat{\eta} > z_\alpha \left(\frac{\sigma_0}{\sqrt{n}} \right) |\eta = \tilde{\eta} > 0, \sigma = \sigma_a) = 1 - \beta$$

or

$$Pr(\hat{\eta} < -z_\alpha \left(\frac{\sigma_0}{\sqrt{n}} \right) |\eta = \tilde{\eta} < 0, \sigma = \sigma_a) = 1 - \beta.$$

This can be rewritten as

$$1 - \Phi \left(z_\alpha \frac{\sigma_0}{\sigma_a} - \frac{\tilde{\eta}}{\sigma_a / \sqrt{n}} \right) = 1 - \beta, \quad \text{if } \tilde{\eta} > 0,$$

and

$$\Phi \left(-z_\alpha \frac{\sigma_0}{\sigma_a} - \frac{\tilde{\eta}}{\sigma_a / \sqrt{n}} \right) = 1 - \beta, \quad \text{if } \tilde{\eta} < 0.$$

If $\tilde{\eta} > 0$ and an approximate n, a solution of the above formula satisfies

$$z_\alpha \frac{\sigma_0}{\sigma_a} - \frac{\tilde{\eta}}{\sigma_a/\sqrt{n}} = -z_\beta.$$

On the other hand, if $\tilde{\eta} < 0$, a solution of the above formula satisfies

$$-z_\alpha \frac{\sigma_0}{\sigma_a} - \frac{\tilde{\eta}}{\sigma_a/\sqrt{n}} = z_\beta.$$

Then

$$n = \frac{(z_\alpha \sigma_0 + z_\beta \sigma_a)^2}{\tilde{\eta}^2},$$

for both cases ($\tilde{\eta} > 0$ and $\tilde{\eta} < 0$).

Now, let us consider testing for multiple parameters. We wish to test the hypothesis $H_0 : \underline{\eta}_1 = 0$ against $H_1 : \underline{\eta}_1 = \underline{\tilde{\eta}}_1$. Let $\underline{\eta}' = (\eta_1, \ldots, \eta_s) = (\underline{\eta}_1', \underline{\eta}_2')$, $\underline{\eta}_1' = (\eta_1, \ldots, \eta_p)$, and $\underline{\eta}_2' = (\eta_{p+1}, \ldots, \eta_s)$. It is equal to test the hypothesis $H_0 : A\underline{\eta} = 0$ against $H_1 : A\underline{\eta} = A\underline{\tilde{\eta}}$. Since the maximum likelihood estimates satisfy

$$\underline{\hat{\phi}} = \begin{bmatrix} \underline{\hat{\theta}} \\ \underline{\hat{\eta}} \end{bmatrix} \sim N_{k+s-1}\left(\begin{bmatrix} \underline{\theta} \\ \underline{\eta} \end{bmatrix}, I^{-1}(\underline{\phi}) \right)$$

asymptotically, where

$$I(\underline{\phi}) = \begin{bmatrix} I(\underline{\theta\theta}) & I(\underline{\theta\eta}) \\ I(\underline{\eta\theta}) & I(\underline{\eta\eta}) \end{bmatrix}.$$

Then the maximum likelihood estimates of $\underline{\eta}$ satisfy

$$\underline{\hat{\eta}} \sim N_s(\underline{\eta}, \{I(\underline{\eta\eta}) - I(\underline{\eta\theta})I(\underline{\theta\theta})^{-1}I(\underline{\theta\eta})\}^{-1})$$

and

$$A\underline{\hat{\eta}} = \underline{\hat{\eta}}_1 \sim N_p(\underline{\eta}_1, A\{I(\underline{\eta\eta}) - I(\underline{\eta\theta})I(\underline{\theta\theta})^{-1}I(\underline{\theta\eta})\}^{-1}A'),$$

where

$$A = (I_{p\times p}, 0_{p\times(s-p)}) = \begin{bmatrix} 1 & 0 & \dots & 0 & | & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & | & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & | & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & | & 0 & \dots & 0 \end{bmatrix}_{p\times s} .$$

Let $I^{(\underline{\eta\eta})} = n^{-1}\{I(\underline{\eta\eta}) - I(\underline{\eta}\underline{\theta})I(\underline{\theta\theta})^{-1}I(\underline{\theta}\underline{\eta})\}^{-1}$. Then, the sample size needed to test at level α with power $\geq 1-\beta$ is

$$n \geq \left\{ \frac{z_{\alpha/2}}{\sqrt{\underline{\tilde{\eta}}' A'(AI^{(\underline{00})}A')^{-1}A\underline{\tilde{\eta}}}} + \frac{z_{\beta}}{\sqrt{\underline{\tilde{\eta}}' A'(AI^{(\underline{\tilde{\eta}\tilde{\eta}})}A')^{-1}A\underline{\tilde{\eta}}}} \right\}^2, \qquad (3.3)$$

where z_c is the $100(1-c)$th percentile of the standard normal distribution.

CHAPTER IV

METHODS FOR SAMPLE SIZE CALCULATIONS

4.1 Introduction

To calculate sample size, we used the Fisher information matrix as a covariance matrix in Chapter III. There are many different ways to calculate the Fisher information matrix for the estimated parameters in an ordinal logistic regression. The main issue in this chapter is the integration for each component of the Fisher information matrix.

As one of the calculation methods, the Fisher information matrix can be approximated by the moment-generating function. The approximation is valid when the probabilities of response categories are small. In section 4.2 we use the approximation method to calculate the required sample size to test at level α, with power $1-\beta$, the null hypothesis $\underline{\eta}_1 = \underline{0}$. That approximation uses the first term of Taylor's expansion and it contains some sources of error. To reduce the error, we consider corrections and such methods are provided in Section 4.3 to Section 4.5. When the probabilities of response categories are small, the first two terms of Taylor's expansion can be considered as a correction method and it is provided in Section 4.3. If the probabilities of response categories are not small, the error of the approximations above are not negligible. For that reason, another correction method is proposed in Section 4.4. In Section 4.5 we use the empirical method to improve small sample problems. In addition, simulation results for suggested methods are presented in Section 4.6.

4.2 Sample Size Calculations with Small Response Probabilities

4.2.1 Expansion of functions in power series: Taylor series

Suppose that $f(x)$ and its derivatives $f'(x)$, $f''(x)$, ..., $f^{(n)}(x)$ exist and are continuous in the closed interval $a \leq x \leq b$ and that $f^{(n+1)}(x)$ exists in the open interval $a < x < b$. Then

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n, \qquad (4.1)$$

where R_n, the remainder, is given in either of the forms

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1}, \qquad \text{(Lagrange's form)}$$

$$R_n = \frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n(x-a), \qquad \text{(Cauchy's form)}$$

where ξ, which lies between a and x, is in general different in the two forms (Spiegel (1974)). As n changes, ξ also changes in general. If for all x and ξ in $[a,b]$ we have $\lim_{n\to\infty} R_n = 0$, then (4.1) can be written

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots, \qquad (4.2)$$

which is called the *Taylor series or expansion* of $f(x)$. In case $a = 0$, it is often called the *Maclaurin series or expansion* of $f(x)$. We use the following expansions:

- *Maclaurin expansion* of $f(z) = z/(1+z)$:

$$\begin{aligned} f(z) &= f(0) + f'(0)z + \frac{f''(0)}{2!}z^2 + \frac{f^{(3)}(0)}{3!}z^3 + \ldots \\ &= 0 + z - z^2 + z^3 + \ldots \\ &= \sum_{v=1}^{\infty}(-1)^{v+1}z^v, \end{aligned}$$

- *Maclaurin expansion* of $g(z) = z/(1+z)^2$:

$$\begin{aligned} g(z) &= g(0) + g'(0)z + \frac{g''(0)}{2!}z^2 + \frac{g^{(3)}(0)}{3!}z^3 + \ldots \\ &= 0 + z - 2z^2 + 3z^3 + \ldots \\ &= \sum_{v=1}^{\infty} (-1)^{v+1} v z^v. \end{aligned}$$

Useing the *Maclaurin expansion* in power of $e^{\theta_j + \underline{\eta}'\underline{X}}$ leads to

$$\frac{e^{\theta_j+\underline{\eta}'\underline{X}}}{(1+e^{\theta_j+\underline{\eta}'\underline{X}})} = \sum_{v=1}^{\infty}(-1)^{v+1}\{e^{\theta_j+\underline{\eta}'\underline{X}}\}^v$$

and

$$\frac{e^{\theta_j+\underline{\eta}'\underline{X}}}{(1+e^{\theta_j+\underline{\eta}'\underline{X}})^2} = \sum_{w=1}^{\infty}(-1)^{w+1}w\{e^{\theta_j+\underline{\eta}'\underline{X}}\}^w.$$

Let $m(\underline{\eta}) = E[e^{\underline{\eta}'\underline{X}}]$ denote the moment-generating function of $\underline{X}$, with $m_i \equiv \partial m/\partial\eta_i$, $i = k, \ldots, k+s-1$ and $m_{ij} = \partial^2 m/\partial\eta_i\partial\eta_j$, $i, j = k, \ldots, k+s-1$. We extend this notation by defining $m_0 = m_{0,0} = m$, and $m_{0,i} = m_{i,0} = m_i$, $i = k, \ldots, k+s-1$. When the probabilities of $k-1$ categories of response are small for likely $\underline{x}$,

$$\frac{e^{\theta_j+\underline{\eta}'\underline{X}}}{(1+e^{\theta_j+\underline{\eta}'\underline{X}})} = e^{\theta_j+\underline{\eta}'\underline{X}} + O(e^{2\theta_j}).$$

Then,

$$\begin{aligned} I_{11} &= -\mathrm{E}\left[\frac{\partial^2 \log \mathrm{L}}{\partial\theta_1^2}\right] \\ &= n\mathrm{E}\left\{\frac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})}\frac{e^{\theta_1+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}_v})^2}\right\} \\ &\quad +n\mathrm{E}\left\{\frac{e^{\theta_2+\underline{\eta}'\underline{x}}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}})} - \frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})}\right\}\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2} \\ &= n\{\mathrm{E}(e^{\theta_2+\underline{\eta}'\underline{x}}) - \mathrm{E}(e^{\theta_1+\underline{\eta}'\underline{x}})\}\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2} + O(e^{2\theta_2}) \\ &= n\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})}m(\underline{\eta}) + O(e^{2\theta_2}), \end{aligned}$$

$$I_{12} = -n\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})}m(\underline{\eta}) + O(e^{2\theta_2}),$$

$$I_{1j} = O(e^{2\theta_1}), \quad 3 \leq j \leq k+s-1,$$

$$I_{ii} = ne^{\theta_i}\left\{\frac{e^{\theta_{i-1}}}{(e^{\theta_i} - e^{\theta_{i-1}})} + \frac{e^{\theta_{i+1}}}{(e^{\theta_{i+1}} - e^{\theta_i})}\right\} m(\underline{\eta}) + O(e^{2\theta_{i+1}}), \quad 2 \leq i < k-2,$$

$$I_{i,i-1} = -n\left\{\frac{e^{\theta_i+\theta_{i-1}}}{(e^{\theta_i} - e^{\theta_{i-1}})}\right\} m(\underline{\eta}) + O(e^{2\theta_i}), \quad 2 \leq i < k-1,$$

$$I_{i,j} = O(e^{2\theta_j}), \quad \text{if } j - i \geq 2 \text{ and } i, j = 2, \ldots, k-1,$$

$$I_{i,j} = O(e^{2\theta_i}), \quad \text{if } i = 2, \ldots, k-2 \text{ and } j = k, \ldots, k+s-1,$$

$$I_{k-1,k-1} = n\left\{\frac{e^{2\theta_{k-1}}}{(e^{\theta_{k-1}} - e^{\theta_{k-2}})}\right\} m(\underline{\eta}) + O(e^{2\theta_{k-1}}),$$

$$I_{k-1,j+k-1} = ne^{\theta_{k-1}} m_j(\underline{\eta}) + O(e^{2\theta_{k-1}}), \quad j = 1, \ldots, s,$$

and

$$I_{i+k-1,j+k-1} = ne^{\theta_{k-1}} m_{ij}(\underline{\eta}) + O(e^{2\theta_{k-1}}), \quad i, j = 1 \ldots, s.$$

To express I_{ij} in a matrix form, let $\mathbf{m^{(1)}}$ denote the s -dimensional vector of first partials of m, and let $\mathbf{m^{(2)}}$ be the $s \times s$ Hessian matrix of second partials of m. We define the augmented Hessian of m to be the $(k+s-1) \times (k+s-1)$ matrix H defined by

$$H(\underline{\theta}, \underline{\eta}) \equiv \begin{bmatrix} m(\underline{\eta})\mathbf{C_{11}} & \mathbf{C_{21}}\mathbf{m^{(1)}}'(\underline{\eta}) \\ \mathbf{m^{(1)}}(\underline{\eta})\mathbf{C'_{21}} & \mathbf{m^{(2)}}(\underline{\eta}) \end{bmatrix},$$

where $\mathbf{C'_{21}} = (0, \ldots, 0, 1)_{1\times(k-1)}$,

$$\mathbf{C_{11}} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1,k-1} \\ c_{21} & c_{22} & \cdots & c_{2,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k-1,1} & c_{k-1,2} & \cdots & c_{k-1,k-1} \end{bmatrix},$$

$c_{i,j} = 0$ (if $|i-j| \geq 2$), $c_{11} = \dfrac{(e^{\theta_1+\theta_2})}{e^{\theta_{k-1}}(e^{\theta_2} - e^{\theta_1})}$, $c_{i,i-1} = -\dfrac{(e^{\theta_i+\theta_{i-1}})}{e^{\theta_{k-1}}(e^{\theta_i} - e^{\theta_{i-1}})}$, $c_{ii} = \dfrac{(e^{\theta_i+\theta_{i-1}})}{e^{\theta_{k-1}}(e^{\theta_i} - e^{\theta_{i-1}})} + \dfrac{(e^{\theta_i+\theta_{i+1}})}{e^{\theta_{k-1}}(e^{\theta_{i+1}} - e^{\theta_i})}$ $(i = 2, \ldots, k-2)$, and $c_{k-1,k-1} = \dfrac{e^{\theta_{k-1}}}{e^{\theta_{k-1}} - e^{\theta_{k-2}}}$.

If $k = 2$, then $c_{11} = 1$. This enables us to write I_{ij} as

$$I(\underline{\theta}, \underline{\eta}) \cong ne^{\theta_{k-1}} H(\underline{\theta}, \underline{\eta}). \tag{4.3}$$

Thus, the asymptotic covariance matrix of the estimates $\underline{\hat{\phi}}' = (\underline{\hat{\theta}}', \underline{\hat{\eta}}')$ is approximately $[ne^{\theta_{k-1}} H(\underline{\theta}, \underline{\eta})]^{-1}$. We first consider testing for one parameter. In particular, the asymptotic variance of $\hat{\eta}_1$ is

$$\text{var}(\hat{\eta}_1) \simeq (ne^{\theta_{k-1}})^{-1} v(\eta_1), \tag{4.4}$$

where $v(\eta_1)$ is the k^{th} diagonal entry of $H^{-1}(\underline{\theta}, \underline{\eta})$. We use the approximation (4.4) to estimate the sample size needed to test at level α, with power$\geq 1 - \beta$, the hypothesis $\eta_1 = 0$ against the alternative $\eta_1 = \tilde{\eta}_1 > 0$. The approximated sample size, n, satisfies

$$ne^{\theta_{k-1}} \geq [v^{1/2}(0) z_\alpha + v^{1/2}(\tilde{\eta}_1) z_\beta]^2 / \tilde{\eta}_1^2, \tag{4.5}$$

where $v(0)$ is the k^{th} diagonal entry of $H^{-1}(\theta_1, \ldots, \theta_{k-1}, 0, \eta_2, \ldots, \eta_s)'$, $v(\tilde{\eta}_1)$ is the k^{th} diagonal entry of $H^{-1}(\theta_1, \ldots, \theta_{k-1}, \tilde{\eta}_1, \eta_2, \ldots, \eta_s)'$, and z_c is the $100(1-c)$th percentile of the standard normal distribution. In particular, when there are only two response categories, the Hessian of m to be the $(s+1) \times (s+1)$ matrix H defined by

$$H(\underline{\theta}, \underline{\eta}) \equiv \begin{bmatrix} m & \mathbf{m}^{(\mathbf{1})'} \\ \mathbf{m}^{(\mathbf{1})} & \mathbf{m}^{(\mathbf{2})} \end{bmatrix},$$

and (4.5) reduces to the binary logistic case (Whittemore, 1981), as expected.

When the distribution for $\underline{X}$ is of a general multivariate exponential type, the moment-generating function for $\underline{X}$ is of the form

$$m(\underline{t}) = \exp\{q(\underline{\gamma} + \underline{t}) - q(\underline{\gamma})\}, \tag{4.6}$$

where $\underline{\gamma}$ is a vector of parameters and q is a real-valued function of $\underline{X}$ variables whose Hessian matrix of second derivatives exists and is positive definite. Let $\mathbf{q}^{(\mathbf{1})}$ denote the s-dimensional vector of the first partials of q, and let $\mathbf{q^{(2)}}$ be the $s \times s$ Hessian matrix of the second partials of q. Then

$$\mathbf{m^{(1)}} = \mathbf{q^{(1)}}(\underline{\gamma} + \underline{t}) \times m(\underline{\theta})$$

and

$$\mathbf{m^{(2)}} = [\mathbf{q^{(1)}}(\underline{\gamma} + \underline{t})\mathbf{q^{(1)}}'(\underline{\gamma} + \underline{\theta}) + \mathbf{q^{(2)}}(\underline{\gamma} + \underline{t})]m(\underline{t}).$$

Then the asymptotic covariance matrix of $\hat{\underline{\eta}}_1' = (\hat{\eta}_1, \ldots, \hat{\eta}_s)$ is approximately

$$(ne^{k-1})^{-1}[H(\underline{\theta}, \underline{\eta})]_{22}^{-1},$$

where

$$H^{-1} = \begin{bmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{bmatrix},$$

and $[H(\underline{\theta}, \underline{\eta})]_{22}^{-1} = H^{22}$ (See Appendix A).

For binary response variable with one covariate which has a standard normal distribution, the approximated sample size for desired power $1 - \beta$ is

$$n \geq \frac{[z_\alpha + z_\beta e^{-\tilde{\eta}_1^2/4}]^2}{\tilde{\eta}_1^2 e^{\theta_1}}.$$

For a response variable with three categories and one covariate which has a standard normal distribution, the approximated sample size for desired power $1 - \beta$ is

$$n \geq \frac{[z_\alpha + z_\beta e^{-\tilde{\eta}_1^2/4}]^2}{\tilde{\eta}_1^2 e^{\theta_2}}.$$

Let us now consider testing for multiple parameters. We wish to test the hypothesis $H_0 : \underline{\eta}_1 = 0$ against $H_1 : \underline{\eta}_1 = \tilde{\underline{\eta}}_1$. Let $\underline{\eta}' = (\eta_1, \ldots, \eta_s) = (\underline{\eta}_1', \underline{\eta}_2')$, $\underline{\eta}_1' = (\eta_1, \ldots, \eta_p)$, and $\underline{\eta}_2' = (\eta_{p+1}, \ldots, \eta_s)$. It is equal to test the hypothesis $H_0 : A\underline{\eta} = 0$ against $H_1 : A\underline{\eta} = A\tilde{\underline{\eta}}$. Then the maximum likelihood estimates satisfy

$$A\hat{\underline{\eta}} = \hat{\underline{\eta}}_1 \sim N_p(\underline{\eta}_1, (ne^{k-1})^{-1} A H^{22} A')$$

asymptotically, where

$$A = (I_{p\times p}, 0_{p\times(s-p)}) = \begin{bmatrix} 1 & 0 & \dots & 0 & | & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & | & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & | & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & | & 0 & \dots & 0 \end{bmatrix}_{p\times s}.$$

The estimated sample size needed to test at level α with power $\geq 1-\beta$ is

$$ne^{k-1} \geq \left\{ \frac{z_{\alpha/2}}{\sqrt{\underline{\tilde{\eta}}' A'(AH_0^{22}A')^{-1}A\underline{\tilde{\eta}}}} + \frac{z_\beta}{\sqrt{\underline{\tilde{\eta}}' A'(AH_1^{22}A')^{-1}A\underline{\tilde{\eta}}}} \right\}^2, \tag{4.7}$$

where z_c is the $100(1-c)$th percentile of the standard normal distribution.

As an example, for a response variable with three categories and two covariates which have a joint normal distribution,

$$\underline{X} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right),$$

the approximated sample size to test the hypothesis $\underline{\eta}' = (\eta_1, \eta_2) = (0, 0)$ against the alternative $\underline{\eta}' = (\eta_1, \eta_2) = (\tilde{\eta}_1, \tilde{\eta}_2)$ is

$$\begin{aligned} ne^{\theta_2} &\geq \left[z_\alpha \Big/ \sqrt{\tilde{\eta}_1^2 + \tilde{\eta}_2^2} + z_\beta \Big/ \sqrt{\exp\left(\frac{\tilde{\eta}_1^2 + \tilde{\eta}_2^2}{2}\right)(\tilde{\eta}_1^2 + \tilde{\eta}_2^2)} \right]^2 \\ &= \frac{1}{\tilde{\eta}_1^2 + \tilde{\eta}_2^2} \left[z_\alpha + \exp\left(-\frac{\tilde{\eta}_1^2 + \tilde{\eta}_2^2}{4}\right) z_\beta \right]^2. \end{aligned}$$

For a response variable with three categories and s covariates which have a joint normal distribution,

$$\underline{X} \sim N_s\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \right),$$

the approximated sample size to test the hypothesis

$$\underline{\eta}' = (\eta_1, \eta_2, \ldots, \eta_s) = (0, 0, \ldots, 0, \eta_{p+1}, \ldots, \eta_s)$$

against the alternative

$$\underline{\eta}' = (\eta_1, \eta_2, \ldots, \eta_s) = (\tilde{\eta}_1, \tilde{\eta}_2, \ldots, \tilde{\eta}_p, \eta_{p+1}, \ldots, \eta_s)$$

is

$$ne^{\theta_2} \geq \frac{e^{-\sum_{i=p+1}^{s} \eta_i^2/2}}{\sum_{j=1}^{p} \tilde{\eta}_j^2} \left[z_\alpha + z_\beta \text{exp} \left(-\frac{1}{4} \sum_{j=1}^{p} \tilde{\eta}_j^2 \right) \right]^2 . \tag{4.8}$$

since $H^{22} = e^{-\{\underline{\eta}'\Sigma\underline{\eta}/2 + \underline{\mu}'\underline{\eta}\}}\Sigma^{-1}$, (Appendix A),

$$\sqrt{\underline{\tilde{\eta}}' A'(AH_0^{22}A')^{-1}A\underline{\tilde{\eta}}} = \sqrt{e^{(\sum_{i=p+1}^{s} \eta_i^2)/2} \sum_{j=1}^{p} \tilde{\eta}_j^2},$$

and

$$\sqrt{\underline{\tilde{\eta}}' A'(AH_1^{22}A')^{-1}A\underline{\tilde{\eta}}} = \sqrt{e^{(\sum_{i=1}^{p} \tilde{\eta}_i^2 + \sum_{i=p+1}^{s} \eta_i^2)/2} \sum_{j=1}^{p} \tilde{\eta}_j^2}.$$

Thus, the approximated sample size to test the hypothesis $\underline{\eta}' = (\eta_1, \eta_2, \eta_3) = (0, 0, \eta_3)$ against the alternative $\underline{\eta}' = (\eta_1, \eta_2, \eta_3) = (\tilde{\eta}_1, \tilde{\eta}_2, \eta_3)$ is

$$ne^{\theta_2} \geq \frac{e^{-\eta_3^2/2}}{\tilde{\eta}_1^2 + \tilde{\eta}_2^2} \left[z_\alpha + \text{exp} \left(-\frac{\tilde{\eta}_1^2 + \tilde{\eta}_2^2}{4} \right) z_\beta \right]^2 .$$

4.3 A Bias Correction

In this section we will consider the error in the sample size approximation and a correction is presented for situations when the approximating is not good. If we add one more term for the approximation,

$$\frac{e^{\theta_j + \underline{\eta}'\underline{X}}}{(1 + e^{\theta_j + \underline{\eta}'\underline{X}})} = e^{\theta_j + \underline{\eta}'\underline{X}} + e^{2\theta_j + 2\underline{\eta}'\underline{X}} + O(e^{3\theta_j}).$$

It follows that

$$I_{11} = n\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2} - e^{\theta_1})}\{m(\underline{\eta}) - 2e^{\theta_1}m(2\underline{\eta})\} + O(e^{3\theta_2}),$$

$$I_{i,i-1} = -n\frac{e^{\theta_i+\theta_{i-1}}}{(e^{\theta_i} - e^{\theta_{i-1}})}\{m(\underline{\eta}) - (e^{\theta_i} + e^{\theta_{i-1}})m(2\underline{\eta})\} + O(e^{3\theta_i}), \quad i = 2, \ldots, k-1,$$

$$I_{ij} = 0, \quad \text{if } |i-j| \geq 2 \text{ and } i, j = 1, \ldots, k-1,$$

$$I_{1,j+k-1} = ne^{\theta_1+\theta_2}m_j(2\underline{\eta}) + O(e^{3\theta_2}), \quad j = 1, \ldots, s,$$

$$\begin{aligned} I_{ii} &= n\left[\frac{e^{\theta_i+\theta_{i-1}}}{(e^{\theta_i} - e^{\theta_{i-1}})}\{m(\underline{\eta}) - 2e^{\theta_i}m(2\underline{\eta})\}\right] \\ &\quad +n\left[\frac{e^{\theta_i+\theta_{i+1}}}{(e^{\theta_{i+1}} - e^{\theta_i})}\{m(\underline{\eta}) - 2e^{\theta_i}m(2\underline{\eta})\}\right] + O(e^{3\theta_{i+1}}), \quad i = 2, \ldots, k-2, \end{aligned}$$

$$I_{i,j+k-1} = n(e^{\theta_i+\theta_{i+1}} - e^{\theta_i+\theta_{i-1}})m_j(2\underline{\eta}) + O(e^{3\theta_{i+1}}) \quad \text{if } i = 2, \ldots, k-2 \text{ and } j = 1, \ldots, s,$$

$$I_{k-1,k-1} = n\frac{e^{2\theta_{k-1}}}{(e^{\theta_{k-1}} - e^{\theta_{k-2}})}\left[m(\underline{\eta}) - 2e^{\theta_{k-1}}m(2\underline{\eta})\}\right] + O(e^{3\theta_{k-1}}),$$

$$I_{k-1,j+k-1} = ne^{\theta_{k-1}}\{m_j(\underline{\eta}) - (2e^{\theta_{k-1}} + e^{\theta_{k-2}})m_j(2\underline{\eta})\} + O(e^{3\theta_{k-1}}), \quad j = 1, \ldots, s,$$

and

$$\begin{aligned} I_{i+k-1,j+k-1} &= -\mathrm{E}\left[\frac{\partial^2 \log \mathrm{L}}{\partial\eta_i\partial\eta_j}\right] \\ &= ne^{\theta_{k-1}}\{m_{ij}(\underline{\eta}) - 2e^{\theta_{k-1}}m_{ij}(2\underline{\eta})\} + O(e^{3\theta_{k-1}}), \quad i, j = 1, \ldots, s. \end{aligned}$$

To express I_{ij} in a matrix form, we define the augmented Hessian of m to be the $(k+s-1) \times (k+s-1)$ matrix H^* defined by

$$H^*(\underline{\phi}) \equiv H(\underline{\phi}) - \begin{bmatrix} m(2\underline{\eta})\boldsymbol{\Delta_{11}} & \boldsymbol{\Delta_{21}}\mathbf{m}^{\mathbf{(1)}'}(\mathbf{2}\underline{\eta}) \\ \mathbf{m}^{\mathbf{(1)}}(\mathbf{2}\underline{\eta})\boldsymbol{\Delta}'_{\mathbf{21}} & 2e^{\theta_{k-1}}\mathbf{m}^{\mathbf{(2)}}(\mathbf{2}\underline{\eta}) \end{bmatrix},$$

where

$$\boldsymbol{\Delta}'_{\mathbf{21}} = -\frac{1}{e^{\theta_{k-1}}}\left(e^{\theta_1+\theta_2}, \ldots, e^{\theta_i}(e^{\theta_{i+1}} - e^{\theta_{i-1}}), \ldots, e^{\theta_{k-1}}(e^{\theta_{k-1}} + e^{\theta_{k-2}})\right)_{1\times(k-1)},$$

$$\boldsymbol{\Delta_{11}} = \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1,k-1} \\ d_{21} & d_{22} & \ldots & d_{2,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ d_{k-1,1} & d_{k-1,2} & \ldots & d_{k-1,k-1} \end{bmatrix},$$

$d_{i,j} = 0$ (if $|i-j| \geq 2$),

$$d_{ii} = 2\frac{e^{2\theta_i}}{e^{\theta_{k-1}}}\left\{\frac{e^{\theta_{i-1}}}{e^{\theta_i} - e^{\theta_{i-1}}} + \frac{e^{\theta_{i+1}}}{e^{\theta_{i+1}} - e^{\theta_i}}\right\}, \quad i = 1, \ldots, k-2,$$

$e^{\theta_0} = 0$,

$$d_{i,i-1} = -\frac{e^{\theta_i + \theta_{i-1}}(e^{\theta_i} + e^{\theta_{i-1}})}{e^{\theta_{k-1}}(e^{\theta_i} - e^{\theta_{i-1}})}, \quad i = 2, \ldots, k-1,$$

and $d_{k-1,k-1} = 2\dfrac{e^{2\theta_{k-1}}}{e^{\theta_{k-1}} - e^{\theta_{k-2}}}$.

4.4 Sample Size Calculations with General Response Probabilities

In many cases with more than two response categories, we have several response probabilities which can be small or large. In this case,

$$f(z) = \frac{z}{(1+z)} = \begin{cases} \sum_{l=1}^{\infty}(-1)^{l+1}z^l, & \text{if } 0 < z \leq 1, \\ \sum_{l=1}^{\infty}(-1)^{l+1}z^{-l}, & \text{if } 1 < z < \infty, \end{cases}$$

and it is simply approximated by

$$\frac{e^{\theta_i + \eta x}}{1 + e^{\theta_i + \eta x}} = \begin{cases} e^{\theta_i + \eta x} + O(e^{2\theta_i}), & \text{if } \theta_i + \eta x \leq 0, \\ e^{-(\theta_i + \eta x)} + O(e^{-2\theta_i}), & \text{if } \theta_i + \eta x \geq 0, \end{cases}$$

where e^{θ_i} is small when $\theta_i + \eta x \leq 0$ (or $e^{-\theta_i}$ is small when $\theta_i + \eta x \geq 0$). For example, there is a response variable with three categories which has small response probability for the first category and large probability for the sum of two categories, i.e., $\dfrac{e^{\theta_1 + \eta x}}{(1 + e^{\theta_1 + \eta x})} < 0.5$ and $\dfrac{e^{\theta_2 + \eta x}}{(1 + e^{\theta_2 + \eta x})} > 0.5$. Then

$$\frac{e^{\theta_1 + \eta x}}{(1 + e^{\theta_1 + \eta x})^2} = e^{\theta_1 + \eta x} + O(e^{2\theta_1}),$$

and

$$\frac{e^{\theta_2 + \eta x}}{(1 + e^{\theta_2 + \eta x})^2} = e^{-(\theta_2 + \eta x)} + O(e^{-2\theta_2}).$$

$$\begin{aligned} I_{11} &\approx n\{\mathrm{E}(e^{-\theta_2-\eta x})-\mathrm{E}(e^{\theta_1+\eta x})\}\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2} \\ &= n\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2}\{e^{-\theta_2}m(-\eta)-e^{\theta_1}m(\eta)\} \end{aligned}$$

$I_{12} \approx -I_{11}$, $I_{13} = 0$, $I_{22} \approx I_{11} + ne^{-\theta_2}m(-\eta)$, $I_{23} = ne^{-\theta_2}m_1(-\eta)$, and $I_{33} = ne^{-\theta_2}m_{11}(-\eta)$. Then

$$\mathrm{var}(\eta) \approx \frac{m(-\eta)}{ne^{-\theta_2}\{m(-\eta)m_{11}(-\eta)-m_1(-\eta)\}}.$$

If $X \sim N(0,1)$, then $\mathrm{var}(\eta) = e^{-\eta_2^2}/ne^{-\theta_2}$ and

$$n \geq e^{\theta_2}\frac{[z_\alpha + z_\beta e^{-\tilde{\eta}_1^2/4}]^2}{\tilde{\eta}_1^2}.$$

4.5 Empirical Methods

When we have the information, such as pilot study, we can get the required sample size using the given data. From the given data, we can calculate the sample size empirically.

An important role is played in nonparametric analysis by the empirical distribution which puts equal probabilities n^{-1} at each sample value y_j (Davison and Hinkley (1997)). The corresponding estimate of F is the empirical distribution function (EDF) $\hat{F}$, which is defined as the sample proportion

$$\hat{F}(y) = \frac{\sharp\{y_j \leq y\}}{n}.$$

where $\sharp\{A\}$ means the number of times the event A occurs.

Suppose that we have no parametric model, but that it is sensible to assume that $Y_1, \ldots, Y_n$ are independent and identically distributed according to an unknown distribution function F. We use the EDF $\hat{F}$ to estimate the unknown CDF F. We shall use $\hat{F}$ just as we would a parametric model: theoretical calculation if possible, otherwise simulation of datasets and empirical calculation of required properties. In the case of the average, exact moments under sampling from the EDF are easily found. We shall use Y^* to denote the random variable distributed according to the fitted model $\hat{F}$, and the superscript * will be used when the moments are calculated according to the fitted distribution. For example,

$$E^*(\bar{Y}^*) = E^*(Y^*) = \sum_{j=1}^{n} \frac{1}{n} y_j = \bar{y}.$$

Using the EDF leads to

$$I^*_{11} = n\frac{1}{m}\sum_{v=1}^{m}\left\{R_{v2}\frac{e^{\theta_1+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}_v})^2} + (R_{v2}-R_{v1})\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2}\right\},$$

$$I^*_{12} = -n\frac{1}{m}\sum_{v=1}^{m}\left\{(R_{v2}-R_{v1})\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2}\right\},$$

$$I^*_{ij} = 0, \quad \text{if } |i-j| \leq 2 \text{ and } i,j = 1,\ldots,k-1,$$

$$I^*_{1,j+k-1} = n\frac{1}{m}\sum_{v=1}^{m}\left\{R_{v2}\frac{X_{vi}e^{\theta_1+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}_v})^2}\right\}, \quad j = 1,\ldots,s,$$

$$\begin{aligned} I^*_{ii} \quad &= \quad n\frac{1}{m}\sum_{v=1}^{m}\left\{(R_{v,i+1}-R_{v,i-1})\frac{e^{\theta_i+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_i+\underline{\eta}'\underline{x}_v})^2}\right. \\ &\quad \left.(R_{v,i}-R_{v,i-1})\frac{e^{\theta_i+\theta_{i-1}}}{(e^{\theta_i}-e^{\theta_{i-1}})^2}(R_{v,i+1}-R_{v,i})\frac{e^{\theta_i+\theta_{i+1}}}{(e^{\theta_{i+1}}-e^{\theta_i})^2}\right\}, \\ &\quad i = 2,\ldots,k-2, \end{aligned}$$

$$I^*_{i,i-1} = -\sum_{v=1}^{n}\left\{(R_{v,i}-R_{v,i-1})\frac{e^{\theta_i+\theta_{i-1}}}{(e^{\theta_i}-e^{\theta_{i-1}})^2}\right\}, \quad i = 2,\ldots,k-1,$$

$$\begin{aligned} I^*_{i,j+k-1} \quad &= \quad n\frac{1}{m}\sum_{v=1}^{m}\left\{(R_{v,i+1}-R_{v,i-1})\frac{X_{vj}e^{\theta_i+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_i+\underline{\eta}'\underline{x}_v})^2}\right\}, \\ &\quad \text{if } i = 2,\ldots,k-2 \text{ and } j = 1,\ldots,s, \end{aligned}$$

$$\begin{aligned} I^*_{k-1,k-1} = n\frac{1}{m}\sum_{v=1}^{m}&\left\{(R_{v,k-1}-R_{v,k-2})\frac{e^{\theta_{k-1}+\theta_{k-2}}}{(e^{\theta_{k-1}}-e^{\theta_{k-2}})^2}\right. \\ &\left.+(1-R_{v,k-2})\frac{e^{\theta_{k-1}+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_{k-1}+\underline{\eta}'\underline{x}_v})^2}\right\}, \end{aligned}$$

$$I^*_{k-1,j+k-1} = n\frac{1}{m}\sum_{v=1}^{m}\left\{(1-R_{v,k-2})\frac{X_{vj}e^{\theta_{k-1}+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_{k-1}+\underline{\eta}'\underline{x}_v})^2}\right\}, \quad j=1,\ldots,s,$$

and

$$\begin{aligned}
I^*_{i+k-1,j+k-1} \quad = \quad & n\frac{1}{m}\sum_{v=1}^{m}\left\{R_{v1}\left(\frac{X_{vi}X_{vj}e^{\theta_1+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}_v})^2}\right)\right. \\
& +(R_{v2}-R_{v1})\left(\frac{X_{vi}X_{vj}e^{\theta_1+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}_v})^2}\frac{X_{vi}X_{vj}e^{\theta_2+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_2+\underline{\eta}'\underline{x}_v})^2}\right) \\
& \left. +\ldots+(1-R_{v,k-1})\left(\frac{X_{vi}X_{vj}e^{\theta_{k-1}+\underline{\eta}'\underline{x}_v}}{(1+e^{\theta_{k-1}+\underline{\eta}'\underline{x}_v})^2}\right)\right\}, \quad i,j=1,\ldots,s.
\end{aligned}$$

The asymptotic covariance matrix of the estimates $\hat{\underline{\phi}}'=(\hat{\underline{\theta}}',\hat{\underline{\eta}}')$ is $I^{-1}(\underline{\phi})$.

4.6 Simulation Study

In order to evaluate the adequacy of the approximation, we used the method described above to calculate sample sizes required to achieve selected size and power for tests within logistic regression models. We performed computer simulations to estimate the power when the calculated sample sizes were used. Each Monte Carlo estimate of sample size was based on 10,000 generated data sets.

For convenience, we refer to the sample size calculation method with the first term of Taylor series as a M1 and the sample size calculation method with the first and second terms of Taylor series as a M2.

4.6.1 Binary Response Case with One Covariate

When $k=2$, the elements of Fisher information matrix are

$$I_{ij} = nE\left(\frac{X_iX_je^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}\right), \quad i,j=1,\ldots,s,$$

where $X_1=1$. We wish to test the hypothesis $H_0 : \underline{\eta}_1 = 0$ against $H_1 : \underline{\eta}_1 = \tilde{\underline{\eta}}_1$. Suppose we have one covariate, consider the problem of testing the null hypothesis

$H_0 : \eta = 0$ against the one-sided alternative $H_A : \eta = \tilde{\eta}$ to test at level α with power $1 - \beta$. Then

$$\begin{aligned} \mathrm{Var}(\eta) &= \frac{nE\left(\dfrac{e^{\theta_1+\eta' x}}{(1+e^{\theta_1+\eta' x})^2}\right)}{n^2E\left(\dfrac{e^{\theta_1+\eta' x}}{(1+e^{\theta_1+\eta' x})^2}\right)E\left(\dfrac{X_iX_je^{\theta_1+\eta' x}}{(1+e^{\theta_1+\eta' x})^2}\right) - n^2E\left(\dfrac{X_ie^{\theta_1+\eta' x}}{(1+e^{\theta_1+\eta' x})^2}\right)^2} \\ &= \frac{1}{n}v(\eta) \end{aligned}$$

and

$$n \geq \left\{ \frac{z_\alpha \sqrt{v(0)} + z_\beta \sqrt{v(\tilde{\eta})}}{\tilde{\eta}^2} \right\}^2.$$

When $X \sim N(0,1)$, the approximated sample size by the small response probabilities (M1) method is

$$v_1(\eta) = e^{-\theta_1} \frac{m(\eta)}{m(\eta)m_{11}(\eta) - m_1(\eta)^2},$$

where $m(\eta) = e^{\eta^2/2}$, $v_1(\eta) = e^{-\eta^2/2}$, and the approximated sample size, n, satisfies

$$n_1 \geq e^{-\theta_1}[z_\alpha + e^{-\tilde{\eta}^2/4} z_\beta]^2 / \tilde{\eta}^2. \tag{4.9}$$

If we consider the approximation error and add one more term (M2),

$$v(\eta)^* = v_1(\eta)[1 + 2e^{\theta_1}R(\eta)] + O(\epsilon^2),$$

where

$$\begin{aligned} R(\eta) &= v_1(\eta)[m_{11}(2\eta) + m^{-2}(\eta)m(2\eta)m_1^2(\eta) - 2m^{-1}(\eta)m_1(\eta)m_1(2\eta)] \\ &= (\eta^2 + 1)e^{3\eta^2/2}. \end{aligned}$$

For the standard normal distribution

$$v^*(\eta) \approx e^{-\eta^2/2}[1 + 2e^{\theta_1}(\eta^2 + 1)e^{3\eta^2/2}]$$

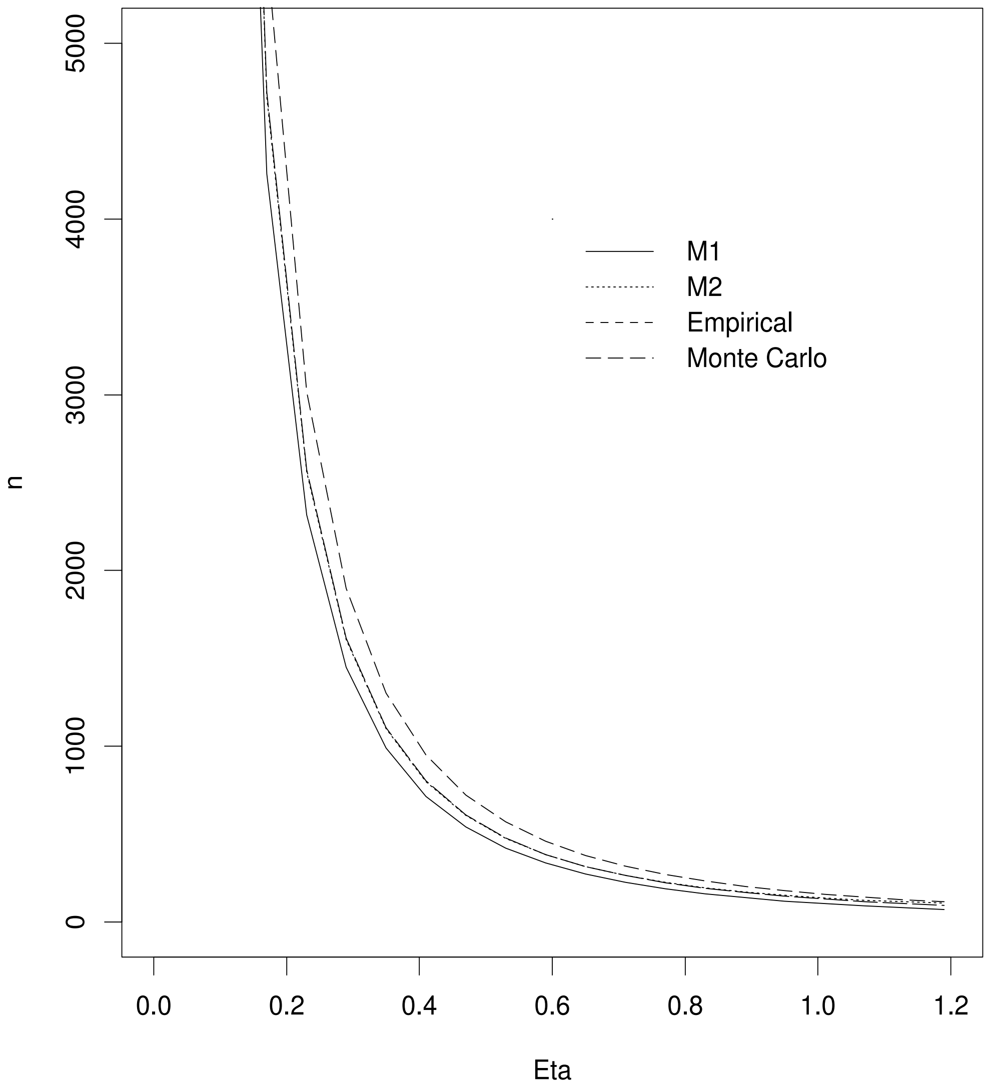


Figure 1: Sample Size for fixed $e^{\theta_1} = 0.05$.

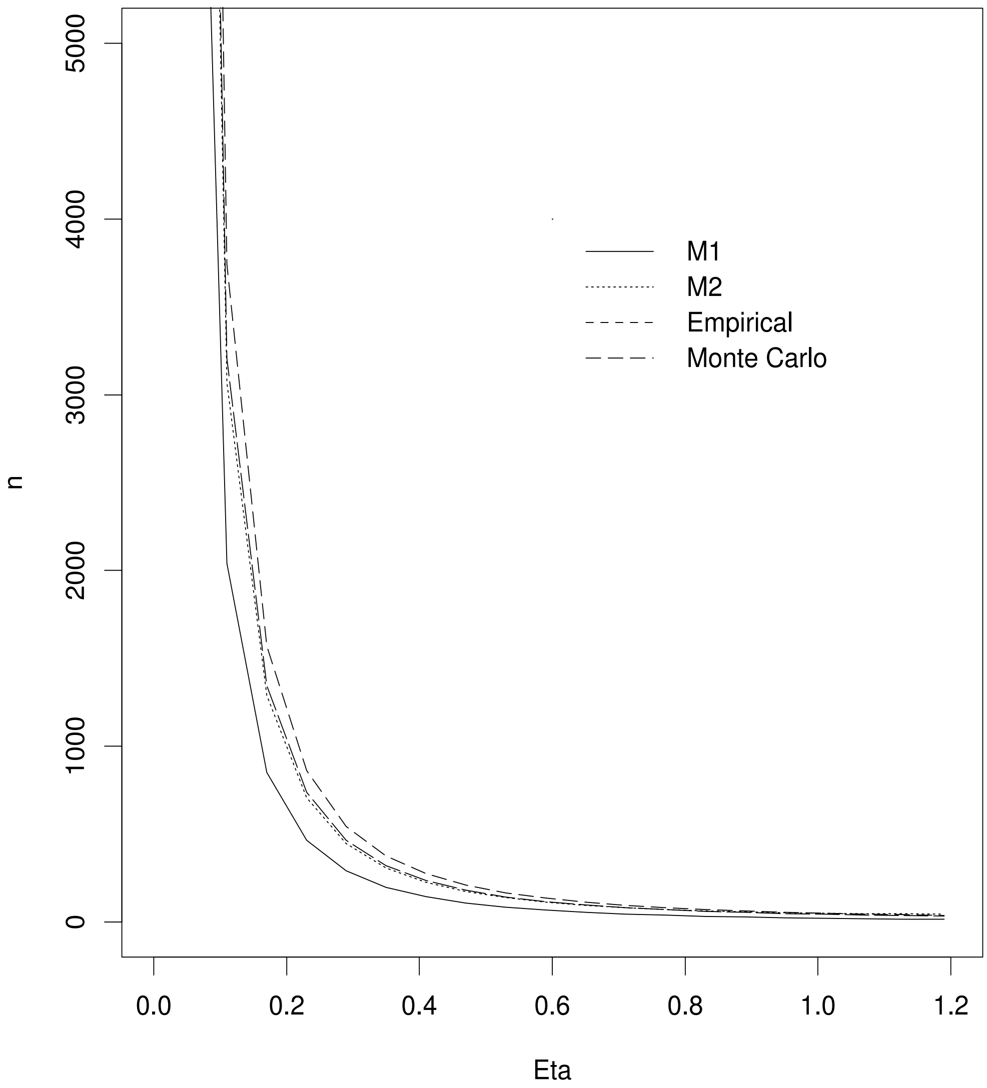


Figure 2: Sample Size for fixed $e^{\theta_1} = 0.25$.

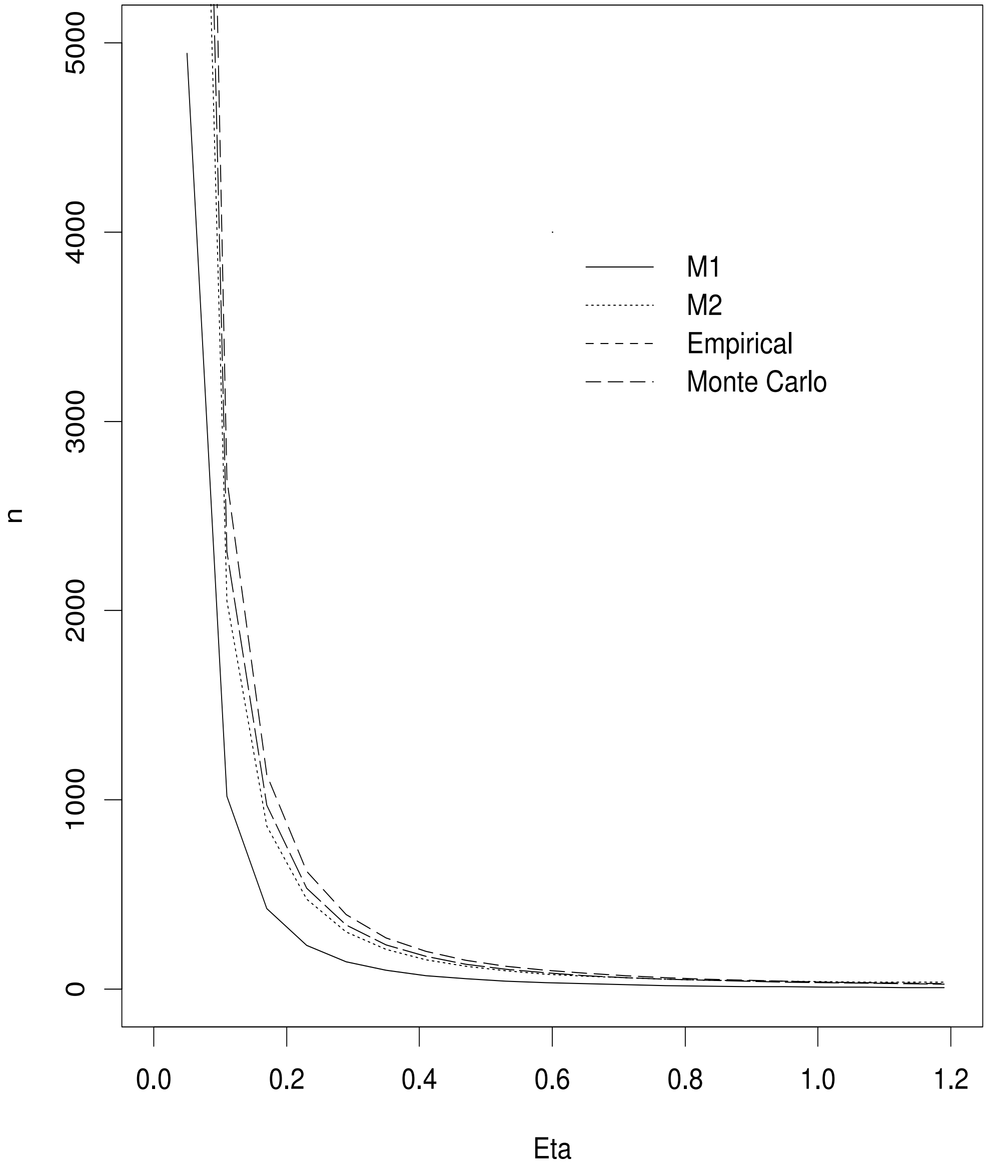


Figure 3: Sample Size for fixed $e^{\theta_1} = 0.50$.

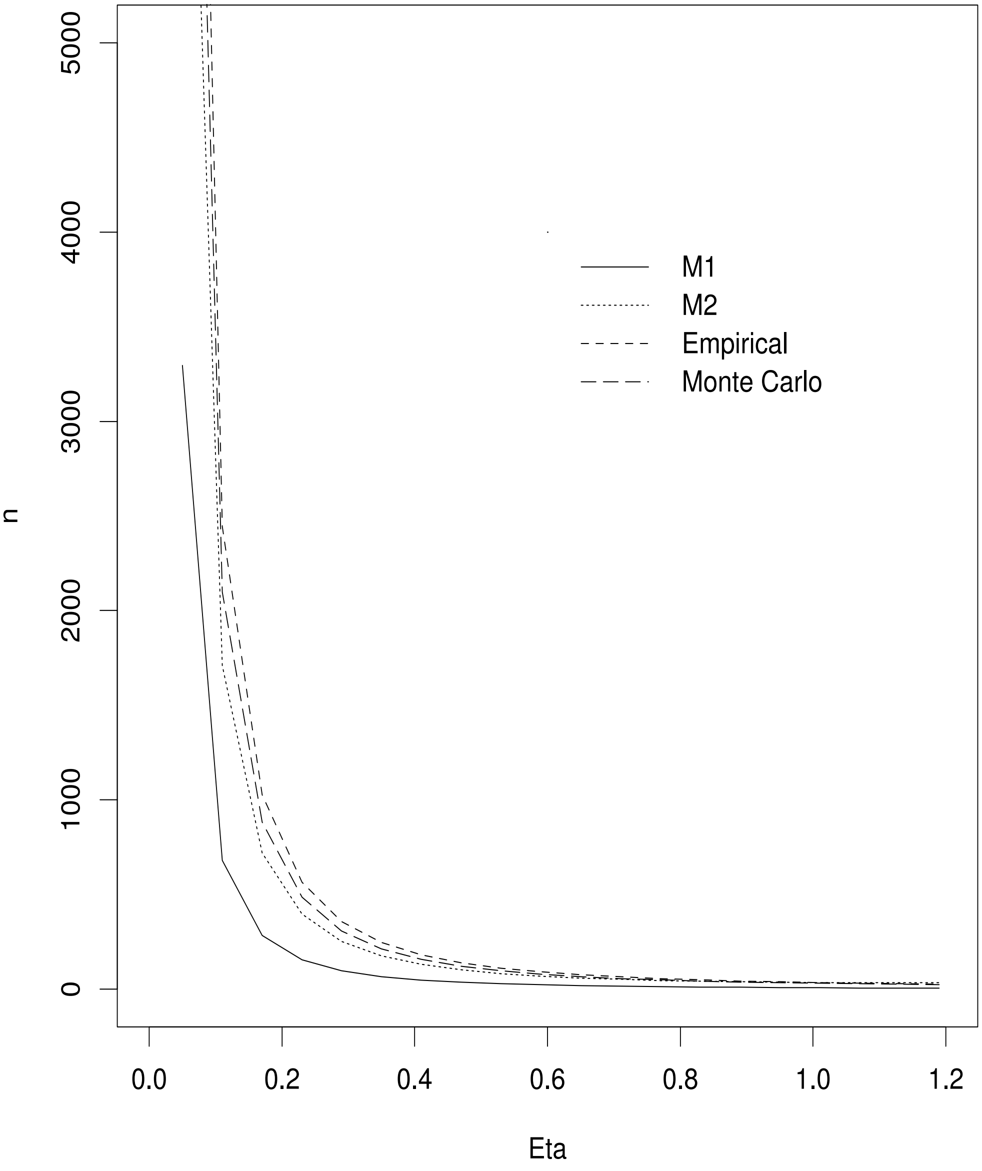


Figure 4: Sample Size for fixed $e^{\theta_1} = 0.75$.

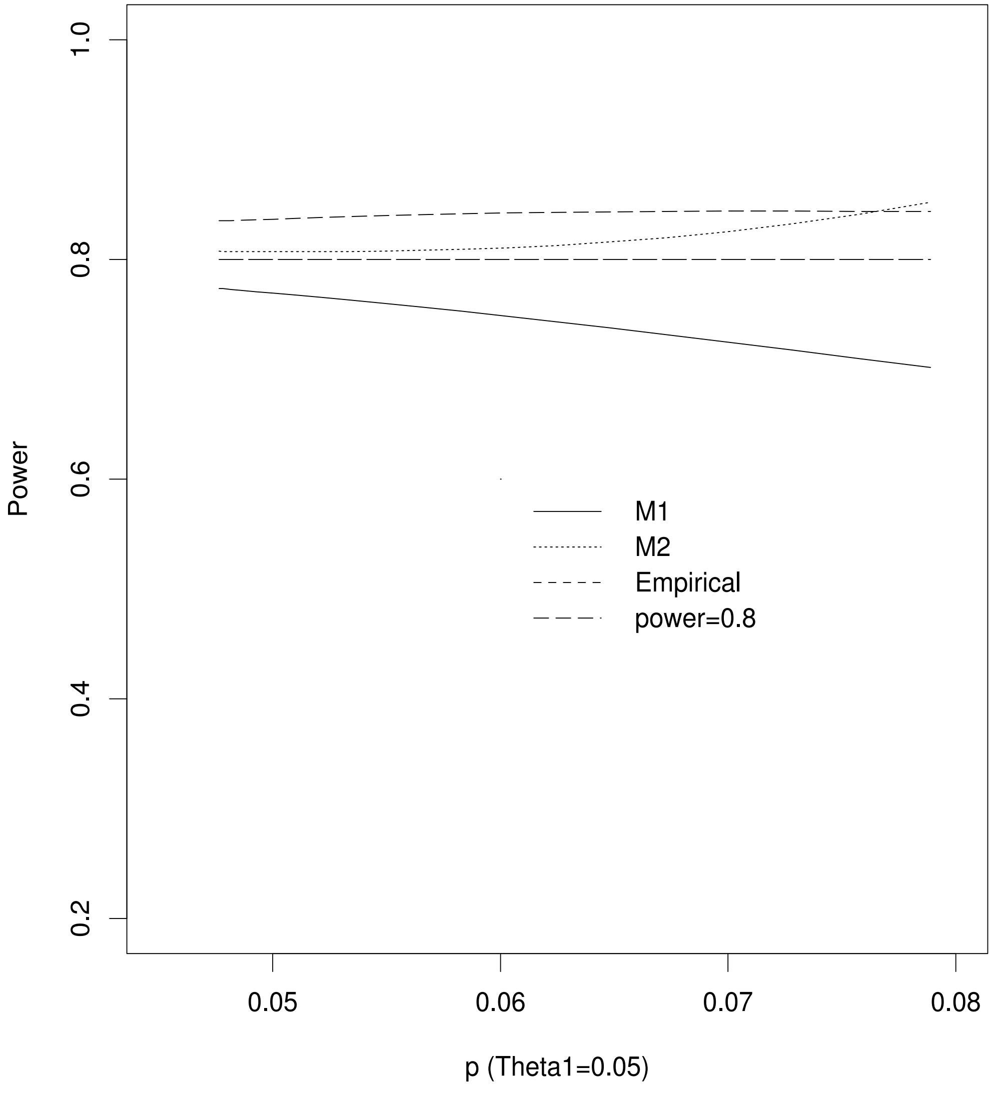


Figure 5: Power for fixed $e^{\theta_1} = 0.05$.

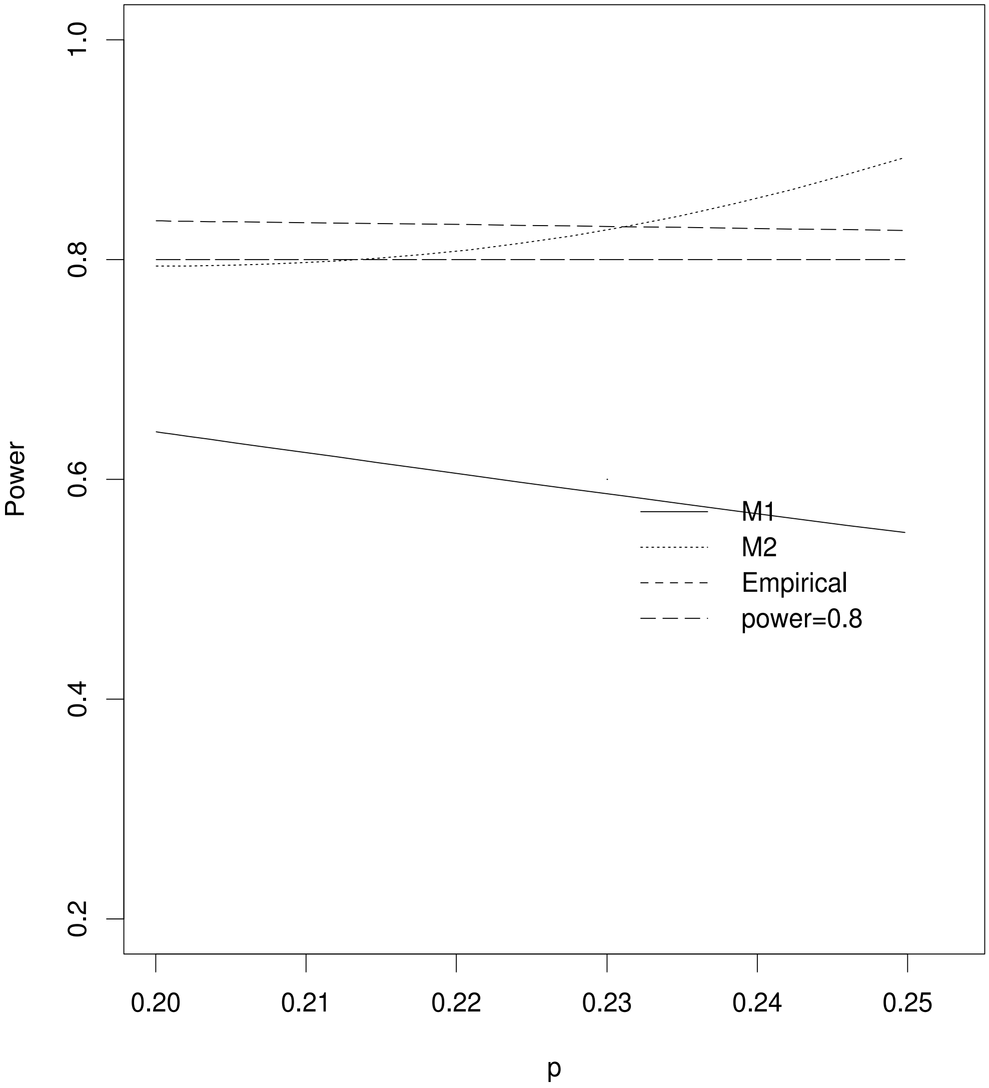


Figure 6: Power for fixed $e^{\theta_1} = 0.25$.

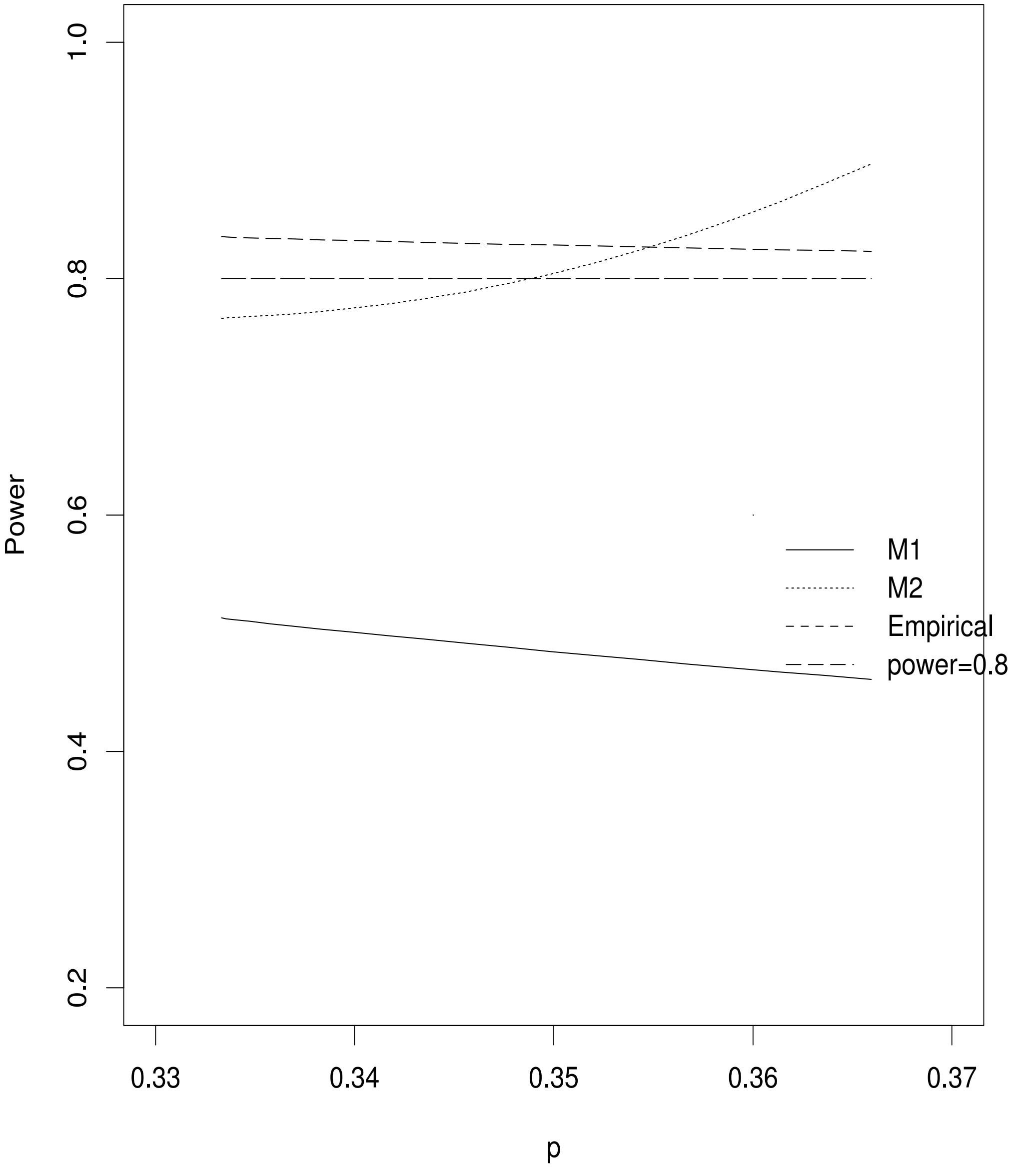


Figure 7: Power for fixed $e^{\theta_1} = 0.50$.

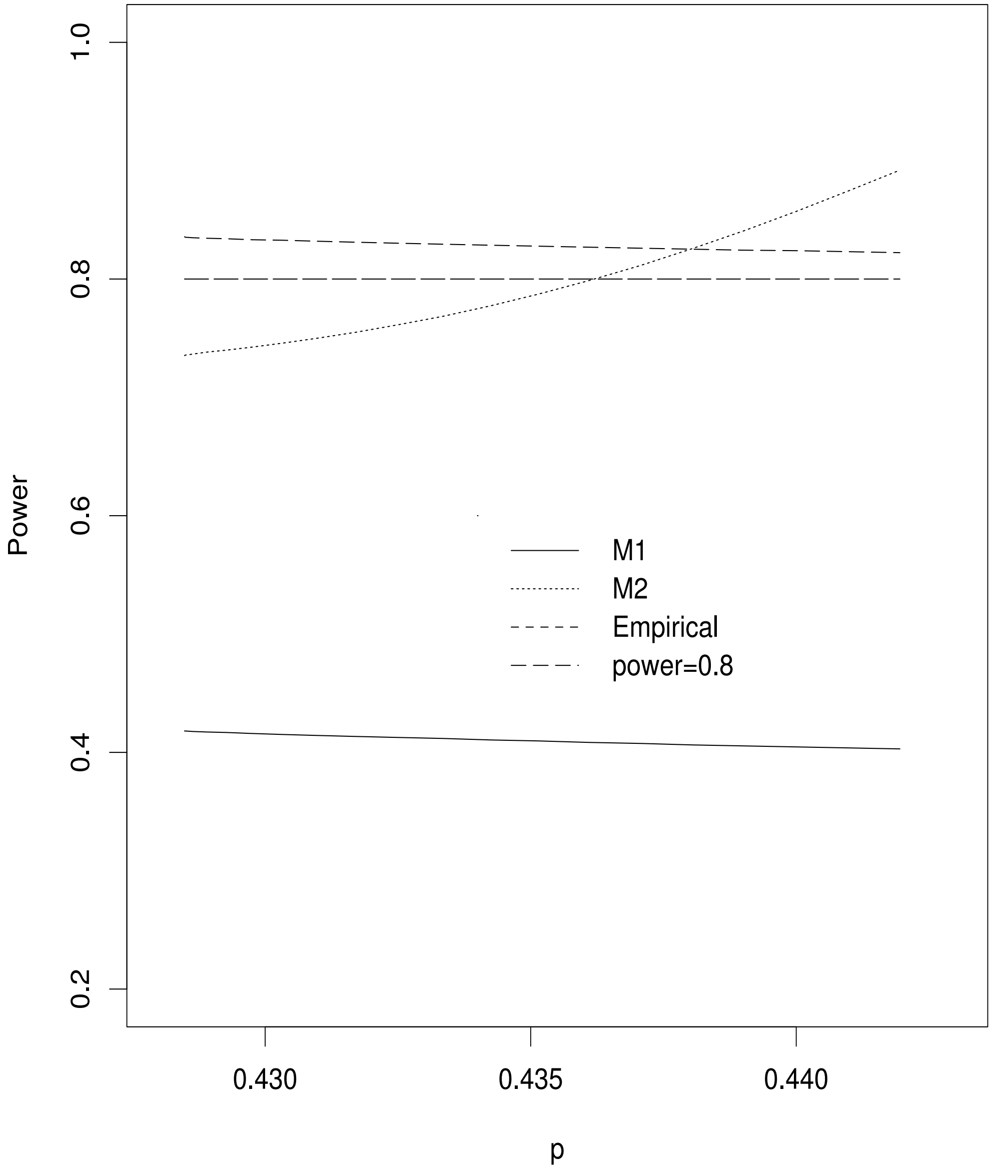


Figure 8: Power for fixed $e^{\theta_1} = 0.75$.

Table 1: Sample sizes and powers, where sample sizes are for one-tailed test ($k = 2, s = 1, p = 1,\ e^{\theta_1} = 0.05$)

	M1		M2		Empirical	method	Monte
$\tilde{\eta}$	n	Power	n	Power	n	Power	n
0.05	49439.54	0.77	54393.97	0.81	59153.82	0.84	53241.57
0.11	10198.22	0.77	11228.59	0.81	12211.31	0.84	10993.75
0.17	4257.76	0.77	4694.21	0.81	5107.46	0.84	4597.26
0.23	2316.71	0.77	2559.23	0.81	2787.06	0.84	2506.99
0.29	1449.65	0.77	1605.70	0.81	1750.90	0.84	1573.25
0.35	988.89	0.77	1099.17	0.81	1200.48	0.84	1077.11
0.41	715.23	0.77	798.51	0.81	873.68	0.84	782.50
0.47	539.59	0.77	605.75	0.81	664.01	0.84	593.50
0.53	420.22	0.76	474.98	0.81	521.56	0.84	465.16
0.59	335.46	0.76	382.38	0.81	420.43	0.84	374.11
0.65	273.14	0.76	314.58	0.81	346.10	0.84	307.27
0.71	226.02	0.75	263.62	0.81	289.90	0.84	256.82
0.77	189.56	0.75	224.53	0.81	246.40	0.84	217.85
0.83	160.78	0.74	194.07	0.81	212.07	0.84	187.18
0.89	137.70	0.74	170.05	0.82	184.51	0.84	162.63
0.95	118.92	0.73	150.99	0.82	162.07	0.84	142.71
1.01	103.45	0.72	135.81	0.83	143.57	0.84	126.34
1.07	90.57	0.72	123.75	0.83	128.14	0.84	112.74
1.13	79.75	0.71	114.27	0.84	115.16	0.84	101.33
1.19	70.58	0.70	106.95	0.85	104.13	0.84	91.67

and

$$n_2 e^{\theta_1} \geq \left[\sqrt{(1+2e^{\theta_1})}\,z_\alpha + \sqrt{e^{-\tilde{\eta}^2/2}[1+2e^{\theta_1}(\eta^2+1)e^{3\tilde{\eta}^2/2}]}\,z_\beta\right]^2 / \tilde{\eta}^2. \qquad (4.10)$$

The graphical representation of the sample sizes for the simulation are given in Figures 1-4. Since the sample sizes depended on the two parameters, θ_1 and η, simultaneously, we fixed one parameter. If we change two parameters, the estimated sample sizes fluctuated too much. To show the performance of the approximation methods according to the response probability, Figures 5-8 are given. In those figures, power depends on η.

Table 1 presents sample sizes computed by (4.9), (4.10), empirical method, and

Table 2: Sample sizes and powers, where sample sizes are for one-tailed test ($k = 2, s = 1, p = 1$, $e^{\theta_1} = 0.25$)

	M1		M2		Empirical	method	Monte
$\tilde{\eta}$	n	Power	n	Power	n	Power	n
0.05	9887.91	0.64	14842.34	0.79	16775.73	0.84	15097.37
0.11	2039.64	0.64	3070.00	0.79	3468.43	0.84	3123.01
0.17	851.55	0.64	1287.97	0.79	1454.39	0.84	1310.12
0.23	463.34	0.64	705.80	0.79	796.43	0.83	717.71
0.29	289.93	0.64	445.88	0.79	502.59	0.83	453.07
0.35	197.78	0.63	307.90	0.79	346.46	0.83	312.44
0.41	143.05	0.63	226.10	0.80	253.75	0.83	228.92
0.47	107.92	0.63	173.77	0.80	194.24	0.83	175.32
0.53	84.04	0.62	138.38	0.80	153.79	0.83	138.89
0.59	67.09	0.62	113.45	0.80	125.07	0.83	113.02
0.65	54.63	0.61	95.32	0.80	103.94	0.83	94.00
0.71	45.20	0.60	81.83	0.81	87.95	0.83	79.61
0.77	37.91	0.60	71.63	0.81	75.56	0.83	68.46
0.83	32.16	0.59	63.84	0.82	65.76	0.83	59.65
0.89	27.54	0.58	57.86	0.83	57.89	0.83	52.57
0.95	23.78	0.58	53.30	0.84	51.47	0.83	46.79
1.01	20.69	0.57	49.89	0.85	46.16	0.83	42.02
1.07	18.11	0.56	47.41	0.86	41.73	0.83	38.03
1.13	15.95	0.56	45.75	0.88	37.98	0.83	34.66
1.19	14.12	0.55	44.81	0.89	34.79	0.83	31.79

Monte Carlo method and estimated powers for selected values of $\alpha = 0.05$, $\beta = 0.2$, $e^{\theta_1} = 0.05$, and $\eta = \tilde{\eta} > 0$ when the explanatory variable has the standard normal distribution. For the empirical method, we generated pseudo standard normal random numbers. In this simulation, we assumed the size of empirical data set is 30. The estimated sample sizes from the empirical method are the mean value of 100 simulation runs in each case. Table 2 presents sample sizes under the same conditions as Table 1 except $e^{\theta_1} = 0.25$. Tables 1-4 and Figures 1-8 show that the approximation method M1 and M2 performed very well under the small response probability condition.

Table 3: Sample sizes and powers, where sample sizes are for one-tailed test ($k = 2, s = 1, p = 1,\ e^{\theta_1} = 0.50$)

	SRP		SRP2		Empirical	method	Monte
$\tilde{\eta}$	n	Power	n	Power	n	Power	n
0.05	4943.95	0.51	9898.38	0.77	12082.65	0.84	10872.24
0.11	1019.82	0.51	2050.17	0.77	2500.38	0.84	2251.11
0.17	425.78	0.51	862.17	0.77	1049.95	0.83	945.92
0.23	231.67	0.51	474.09	0.77	576.07	0.83	519.41
0.29	144.97	0.51	300.85	0.77	364.39	0.83	328.85
0.35	98.89	0.51	208.92	0.77	251.90	0.83	227.56
0.41	71.52	0.50	154.44	0.77	185.07	0.83	167.38
0.47	53.96	0.50	119.63	0.78	142.16	0.83	128.72
0.53	42.02	0.50	96.12	0.78	112.98	0.83	102.43
0.59	33.55	0.49	79.58	0.78	92.24	0.83	83.73
0.65	27.31	0.49	67.60	0.79	76.97	0.83	69.96
0.71	22.60	0.49	58.73	0.80	65.40	0.83	59.53
0.77	18.96	0.48	52.06	0.80	56.43	0.83	51.43
0.83	16.08	0.48	47.02	0.81	49.32	0.83	45.02
0.89	13.77	0.48	43.22	0.82	43.60	0.83	39.86
0.95	11.89	0.47	40.38	0.84	38.93	0.83	35.63
1.01	10.34	0.47	38.34	0.85	35.06	0.83	32.14
1.07	9.06	0.47	36.97	0.86	31.83	0.82	29.21
1.13	7.98	0.46	36.20	0.88	29.09	0.82	26.73
1.19	7.06	0.46	35.97	0.90	26.75	0.82	24.61

The results in Tables 1-4 show us that the approximation (4.9) is suitable when the response probabilities are small but it always under estimates. The approximation with corrected term (4.10) performs better than the approximation (4.9) when the response probabilities are small, but it highly over estimates when the response probabilities are large.

Table 5 shows power values in cases where the values of the test parameter are not small. In this case, the first and second methods are far from our objective value. So, here we suggest one more method. In this method we consider the mean value theorem and the Taylor approximation. The Taylor polynomials can be a

Table 4: Sample sizes and powers, where sample sizes are for one-tailed test ($k = 2, s = 1, p = 1,\ e^{\theta_1} = 0.75$)

	SRP		SRP2		Empirical	method	Monte
$\tilde{\eta}$	n	Power	n	Power	n	Power	n
0.05	3295.97	0.42	8250.40	0.74	10965.84	0.84	9866.02
0.11	679.88	0.42	1710.23	0.74	2270.15	0.84	2043.40
0.17	283.85	0.42	720.23	0.74	953.82	0.84	859.14
0.23	154.45	0.42	396.84	0.74	523.71	0.83	472.14
0.29	96.64	0.42	252.49	0.74	331.57	0.83	299.23
0.35	65.93	0.42	175.90	0.74	229.44	0.83	207.31
0.41	47.68	0.42	130.53	0.75	168.76	0.83	152.68
0.47	35.97	0.41	101.54	0.75	129.79	0.83	117.58
0.53	28.01	0.41	81.97	0.76	103.28	0.83	93.70
0.59	22.36	0.41	68.23	0.76	84.43	0.83	76.71
0.65	18.21	0.41	58.29	0.77	70.54	0.83	64.19
0.71	15.07	0.41	50.94	0.78	60.02	0.83	54.70
0.77	12.64	0.41	45.44	0.79	51.85	0.83	47.32
0.83	10.72	0.41	41.30	0.80	45.38	0.83	41.48
0.89	9.18	0.41	38.21	0.81	40.17	0.83	36.77
0.95	7.93	0.41	35.94	0.83	35.91	0.83	32.92
0.01	6.90	0.41	34.35	0.84	32.38	0.82	29.73
1.07	6.04	0.40	33.34	0.86	29.43	0.82	27.05
1.13	5.32	0.40	32.85	0.87	26.93	0.82	24.79
1.19	4.71	0.40	32.85	0.89	24.79	0.82	22.85

method to form polynomial approximation of complicated functions. We can make approximations based on the mean value and extended value theorems. The extended mean value theorem concludes

$$f(z) = f(a) + f'(a)(z-a) + f''(c)(z-a)^2/2$$

for some c between a and z, provided $f''(z)$ is continuous on $[a, b]$. Here $f(z) = z(1+z)^{-2}$ is approximated by a second order polynomial.

$$f(z) = \frac{z}{(1+z)^2} = \begin{cases} z + \dfrac{f''(c_1)}{2} z^2, & \text{if } 0 < z \leq 1, \\ \dfrac{1}{z} + \dfrac{f''(c_2)}{2} \dfrac{1}{z^2}, & \text{if } 1 < z < \infty. \end{cases}$$

Table 5: Sample sizes and power values in cases where the values of the test parameter are not small (one-tailed test, $k = 2, s = 1, p = 1$, $e^{\theta_1} = 0.25$)

	M1		M2		M3		Empirical	method	Monte
$\tilde{\eta}$	n	Power	n	Power	n	Power	n	Power	n
0.5	94.91	0.60	154.46	0.78	184.10	0.84	172.00	0.81	164.78
0.6	64.75	0.59	110.03	0.78	128.50	0.83	121.10	0.81	115.92
0.7	46.61	0.59	83.82	0.79	94.99	0.83	90.42	0.81	86.50
0.8	34.87	0.58	67.47	0.80	73.27	0.83	70.51	0.81	67.44
0.9	26.86	0.57	57.01	0.81	58.41	0.82	56.87	0.81	54.38
1.0	21.17	0.56	50.39	0.83	47.83	0.82	47.11	0.81	45.06
1.1	16.99	0.54	46.49	0.86	40.07	0.81	39.90	0.81	38.17
1.2	13.84	0.53	44.72	0.88	34.27	0.81	34.42	0.81	32.93
1.3	11.42	0.52	44.84	0.91	29.86	0.81	30.15	0.81	28.86
1.4	9.53	0.52	46.84	0.94	26.48	0.81	26.76	0.81	25.63
1.5	8.02	0.51	50.96	0.97	23.86	0.81	24.03	0.81	23.02
1.6	6.82	0.50	57.71	0.98	21.81	0.81	21.80	0.81	20.89
1.7	5.84	0.49	68.02	0.99	20.16	0.81	19.95	0.81	19.13
1.8	5.03	0.49	83.38	1.00	18.80	0.82	18.40	0.81	17.65
1.9	4.37	0.48	106.22	1.00	17.64	0.82	17.09	0.81	16.39
2.0	3.82	0.47	140.47	1.00	16.63	0.82	15.97	0.81	15.33
2.1	3.36	0.47	192.62	1.00	15.74	0.82	15.01	0.81	14.41
2.2	2.97	0.47	273.45	1.00	14.93	0.82	14.18	0.81	13.61
2.3	2.64	0.46	401.25	1.00	14.20	0.82	13.46	0.81	12.92
2.4	2.36	0.46	607.59	1.00	13.54	0.82	12.83	0.81	12.31

To find $C_1 = \dfrac{f''(c_1)}{2}$ and $C_2 = \dfrac{f''(c_2)}{2}$, we simply assume $C_1^* = 1/z^* - 1/(1+z^*)^2$. Then, $z = e^{\theta_1+\eta x}$, $z^* = e^{\theta_1}m(\eta) = e^{\theta_1}e^{\eta^2/2}$, and $C_1 = 1/e^{\theta_1+\eta^2/2} - 1/(e^{\theta_1+\eta^2/2})^2$. By this method (M3)

$$v^{**}(\eta) \approx e^{-\eta^2/2}[1 + C_1 e^{\theta_1}(\eta^2+1)e^{3\eta^2/2}]$$

$$n_3 e^{\theta_1} \geq \left[\sqrt{(1+2e^{\theta_1})}z_\alpha + \sqrt{e^{-\tilde{\eta}^2/2}[1 + C_1 e^{\theta_1}(\eta^2+1)e^{3\tilde{\eta}^2/2}]}z_\beta\right]^2 / \tilde{\eta}^2. \qquad (4.11)$$

The conclusions drawn from Table 5 are demonstrated graphically in Figure 9. From Figure 9, it seems that the new bias correction method are suitable when the test parameter η is large under the alternative hypothesis. The empirical simulation results show us that every calculated sample sizes are a little bit over estimated but

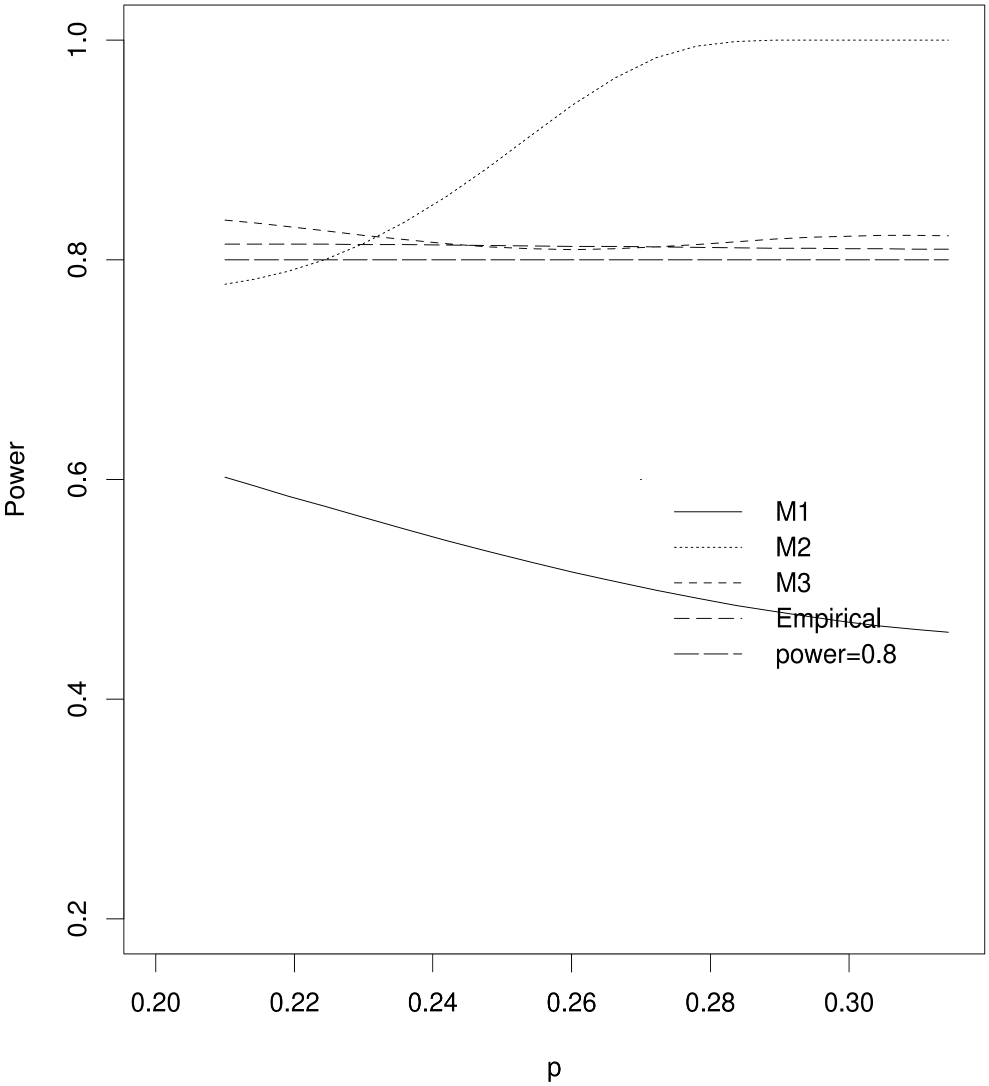


Figure 9: Power values in cases where the values of the test parameter are not small ($e^{\theta_1} = 0.25$).

it performed well when the test parameter η is greater than 1 under the alternative hypothesis.

4.6.2 Three Response Categories Case 1.

Now we consider three response categories with one covariate case. When we have one covariate, the variance of η is simply calculated using the elements of the Fisher information matrix

$$\begin{aligned}\mathrm{Var}(\eta) &= \frac{I_{11}I_{22} - I_{12}^2}{I_{11}I_{22}I_{33} + 2I_{12}I_{23}I_{13} - I_{13}^2 I_{22} - I_{23}^2 I_{11} - I_{12}^2 I_{33}} \\ &= \frac{1}{n} v(\eta).\end{aligned}$$

Assume that the cumulative probability of the first two categories are small. Then the third diagonal entry of $H^{-1}(\underline{\phi})$ is simply

$$v(\eta) = m/\{mm_{11} - m_1^2\}(\eta).$$

For a response variable with three categories and one covariate which has a standard normal distribution, the approximated sample size for desired power $1 - \beta$ is

$$n_1 e^{\theta_2} \geq [z_\alpha + e^{-\tilde{\eta}^2/4} z_\beta]^2 / \tilde{\eta}^2, \tag{4.12}$$

by the small response probabilities (M1) method. If we consider the approximation error and apply a correction term (M2). Then the corrected sample size is

$$n_2 e^{\theta_2} \geq \left[z_\alpha \sqrt{1 + 2e^{\theta_2}} + z_\beta \sqrt{e^{-\tilde{\eta}^2/2} \left[1 + 2(\eta^2 e^{\theta_1} + e^{\theta_2}) e^{3\tilde{\eta}^2/2} \right]} \right]^2 / \tilde{\eta}^2. \tag{4.13}$$

Table 6 presents sample sizes computed by (4.12), (4.13), and Monte Carlo method and estimated powers for selected values of $\alpha = 0.05$, $\beta = 0.2$, $e^{\theta_1} = 0.05$, $e^{\theta_2} = 0.15$, and $\eta = \tilde{\eta} > 0$ when the explanatory variable has the standard normal distribution. The results in Table 6 show us that the approximation (4.12) is suitable when the response probabilities are small but it always under estimates. The

Table 6: Sample sizes and powers, where sample sizes are for one-tailed test ($k = 3, s = 1, p = 1$, $e^{\theta_1} = 0.05$, and $e^{\theta_1} = 0.15$)

	M1		M2		Empirical	method	Monte
$\tilde{\eta}$	n	Power	n	Power	n	Power	n
0.05	16479.85	0.71	21437.79	0.80	24258.37	0.85	21131.89
0.11	3399.41	0.71	4433.38	0.80	4939.78	0.84	4376.60
0.17	1419.25	0.70	1859.47	0.80	2107.91	0.84	1847.24
0.23	772.24	0.70	1018.77	0.80	1160.99	0.85	1006.30
0.29	483.22	0.70	643.63	0.81	720.36	0.84	631.61
0.35	329.63	0.70	444.72	0.81	498.19	0.84	435.01
0.41	238.41	0.70	327.09	0.81	363.45	0.85	315.14
0.47	179.86	0.70	252.17	0.82	292.81	0.85	239.68
0.53	140.07	0.69	201.92	0.82	219.71	0.85	189.60
0.59	111.82	0.69	167.02	0.83	176.32	0.85	152.27
0.65	91.05	0.68	142.26	0.84	149.96	0.85	127.16
0.71	75.34	0.68	124.60	0.85	122.84	0.85	105.74
0.77	63.19	0.67	112.20	0.87	106.45	0.85	90.82
0.83	53.59	0.66	103.92	0.89	93.57	0.86	78.17
0.89	45.90	0.66	99.13	0.91	80.20	0.86	78.17
0.95	39.64	0.65	97.51	0.93	70.61	0.85	60.17
1.01	34.48	0.65	99.08	0.96	62.86	0.85	53.34
1.07	30.19	0.64	104.12	0.98	57.49	0.86	47.84
1.13	26.58	0.64	113.33	0.99	51.53	0.86	42.39
1.19	23.53	0.63	127.89	1.00	45.97	0.85	38.95

approximation with corrected term (4.13) performs better than the approximation (4.12) when the response probabilities are small, but it highly over estimates when the response probabilities are large the same as in the binary response case. Figures 10-11 graphically show the results of Table 6.

4.6.3 Three Response Categories Case 2.

When there are three response categories with one explanatory variable and the cumulative probability of the first two response categories are greater than 0.5, the third diagonal entry of $H^{-1}(\underline{\phi})$ is simply

$$v(\eta) = m(-\eta)/\{m(-\eta)m_{11}(-\eta) - m_1^2(-\eta)\}(\eta).$$

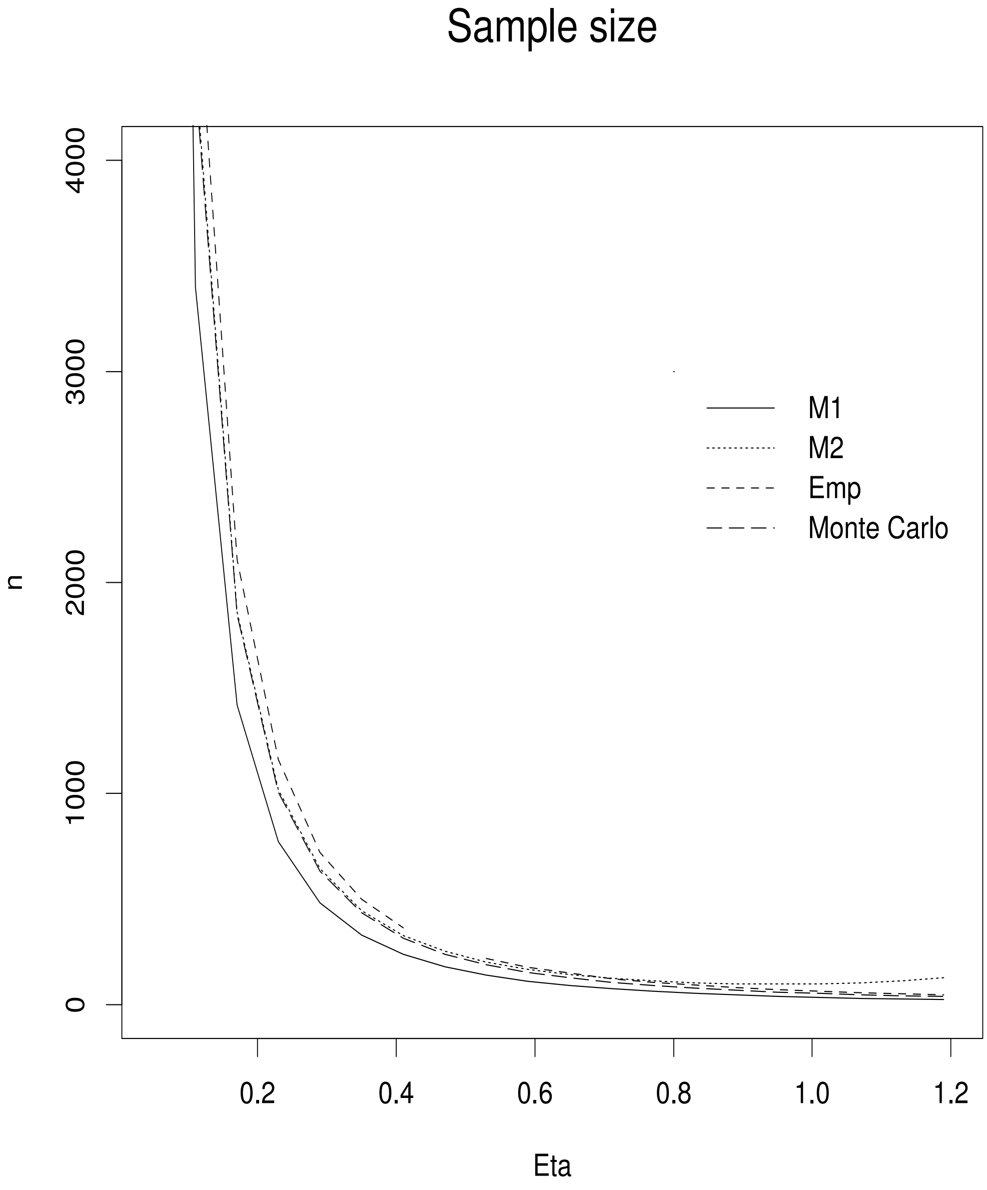


Figure 10: Sample Size for fixed $e^{\theta_1} = 0.05$ and $e^{\theta_2} = 0.15$

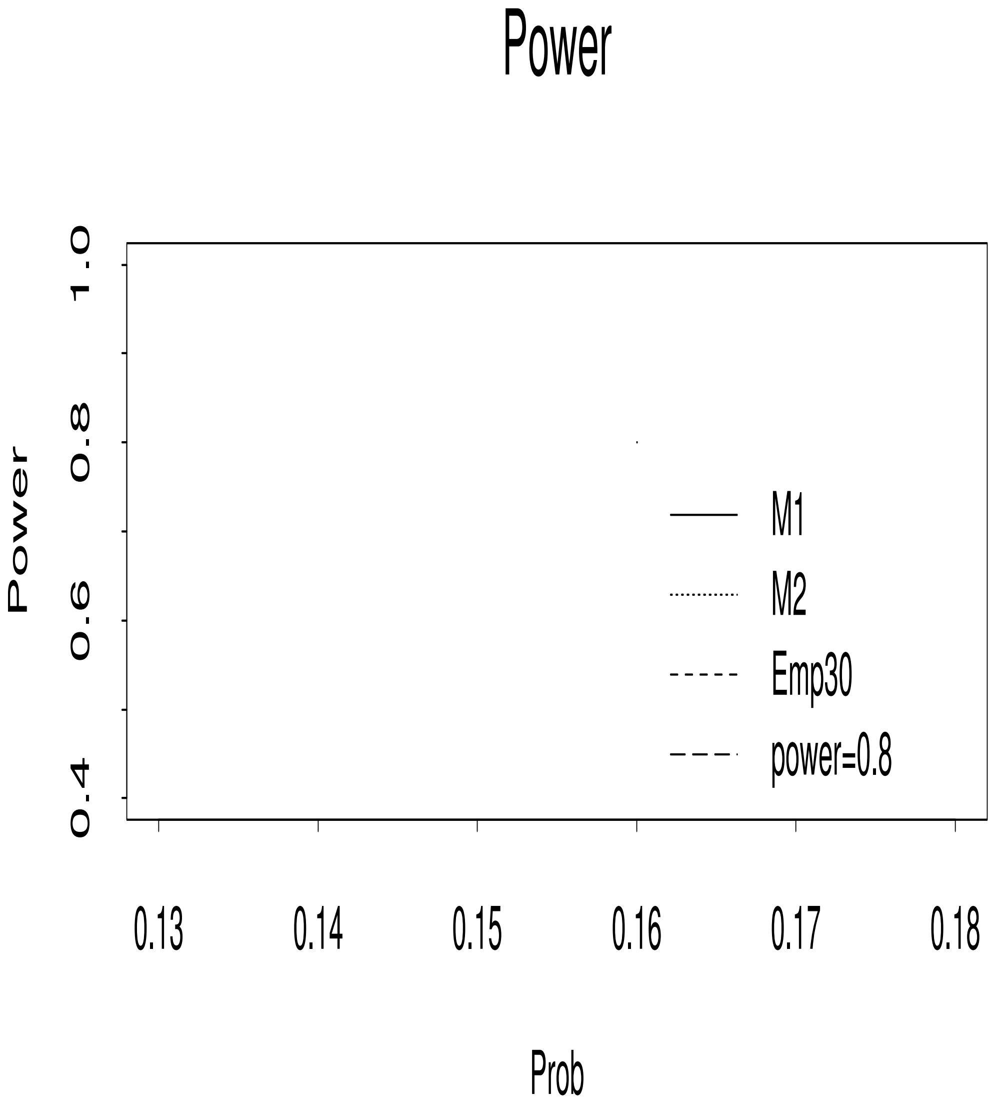


Figure 11: Power for fixed $e^{\theta_1} = 0.05$ and $e^{\theta_2} = 0.15$

Table 7: Sample sizes and powers, where sample sizes are for one-tailed test ($k = 3, s = 1, p = 1$, $e^{\theta_1} = 0.5$, and $e^{\theta_2} = 7$)

	M1		M2		Empirical	method	Monte
$\tilde{\eta}$	n	Power	n	Power	n	Power	n
0.5	4.75	0.16	118.63	0.91	92.83	0.84	82.01
0.6	3.24	0.16	80.94	0.90	65.68	0.84	57.51
0.7	2.33	0.16	58.26	0.89	49.35	0.84	43.76
0.8	1.74	0.16	43.59	0.88	37.92	0.83	34.02
0.9	1.34	0.16	33.57	0.86	30.64	0.83	27.93
1.0	1.06	0.17	26.46	0.84	25.42	0.83	23.00
1.1	0.85	0.17	21.23	0.82	21.83	0.83	19.81
1.2	0.69	0.17	17.30	0.80	18.72	0.83	17.08
1.3	0.57	0.18	14.27	0.78	16.57	0.83	15.12
1.4	0.48	0.18	11.91	0.76	14.86	0.83	13.52
1.5	0.40	0.19	10.03	0.74	13.43	0.83	12.14
1.6	0.34	0.19	8.52	0.72	11.95	0.82	11.06
1.7	0.29	0.20	7.30	0.70	11.02	0.82	10.16
1.8	0.25	0.20	6.29	0.68	10.10	0.82	9.52
1.9	0.22	0.21	5.46	0.67	9.52	0.82	8.73
2.0	0.19	0.21	4.77	0.65	8.89	0.82	8.22
2.1	0.17	0.22	4.20	0.64	8.33	0.82	7.67
2.2	0.15	0.22	3.71	0.62	7.93	0.82	7.31
2.3	0.13	0.23	3.30	0.61	7.46	0.82	6.91
2.4	0.12	0.23	2.95	0.59	7.06	0.82	6.64

To compare this formula with small probability formula, we use

- M1

$$n_1 e^{\theta_2} \geq [z_\alpha + e^{-\tilde{\eta}^2/4} z_\beta]^2/\tilde{\eta}^2. \tag{4.14}$$

  M2 method is our suggested method.

- M2

$$n_1 e^{-\theta_2} \geq [z_\alpha + e^{-\tilde{\eta}^2/4} z_\beta]^2/\tilde{\eta}^2. \tag{4.15}$$

The results in Table 7 show us that the approximation (4.14) is not suitable when the response probabilities are not small and it always under estimates. The suggested approximation (4.15) performs much better than the approximation (4.14) when the response probabilities are not small, and it over estimates when the response probabilities are large. When we evaluate a large alternative value, small sample size is required to obtain a certain power, it contains another error associated with using the standard normal approximation to the distribution of maximum likelihood estimate $(\hat{\eta}-\tilde{\eta})/\sqrt{var(\hat{\eta})}$. Thus a small value of approximated sample size yields serious error.

4.6.4 Remarks on Simulation Results

From these simulation results, we have the following summary.

- The suggested method (M1) works well when the response categories have small probabilities.
- For the binary response cases, we can use two correction methods. If the response probability is small, the formula (4.10) works very well. Alternatively, in cases where the response probability is not small, (4.11) has better performance.
- For the three response categories case, (4.13) works well if the response probabilities are small.
- If the probabilities of response categories are not small, we can apply (4.15) except in the small sample size cases. To reduce the approximation error, another consideration is discussed in Chapter VI.
- If a data set is available, we can apply an empirical method to estimate the sample size. It is suitable when the response categories do not have small probabilities.

Note that all our computations were conducted in Splus.

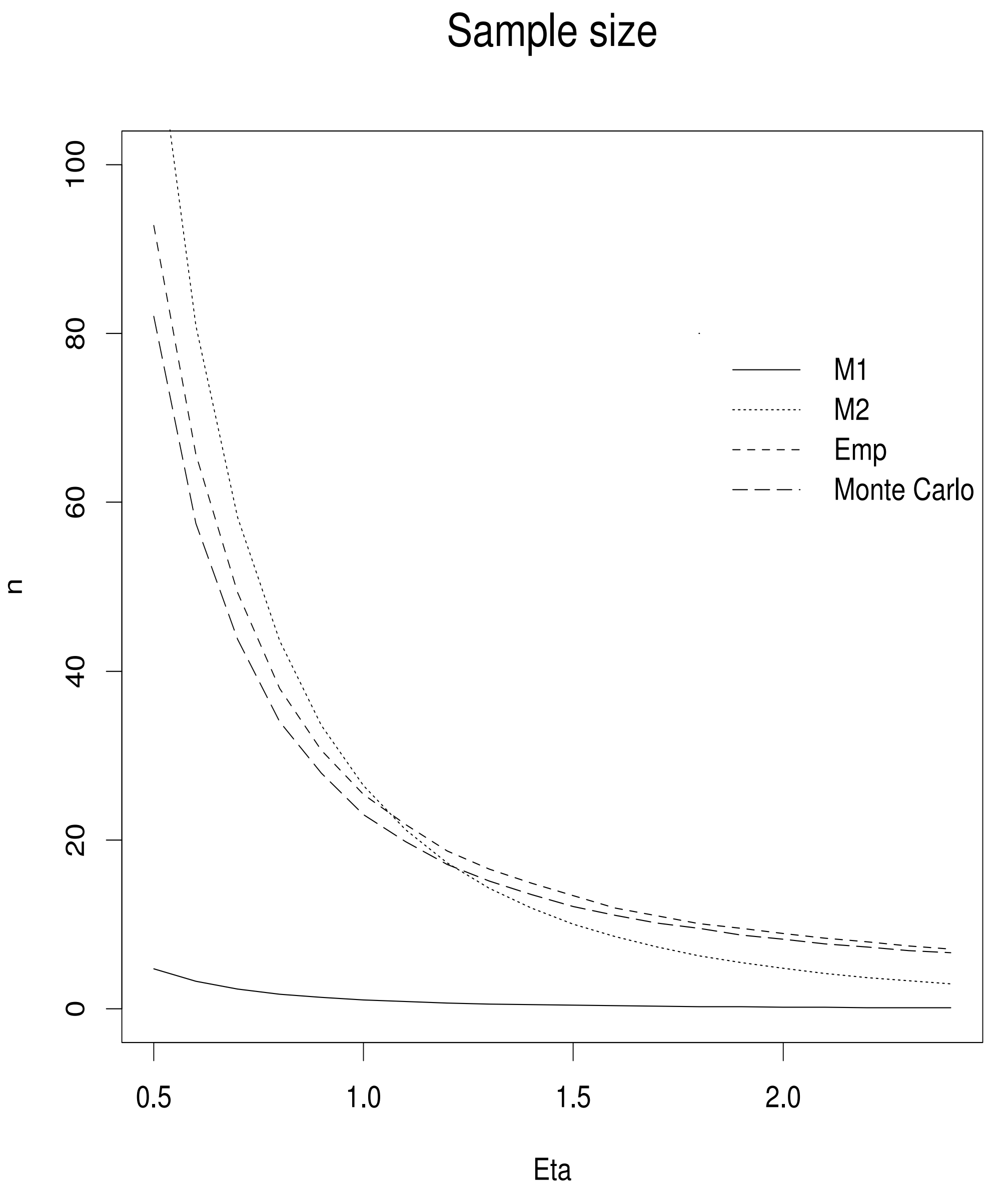


Figure 12: Power for fixed $e^{\theta_1} = 0.5$ and $e^{\theta_2} = 7$

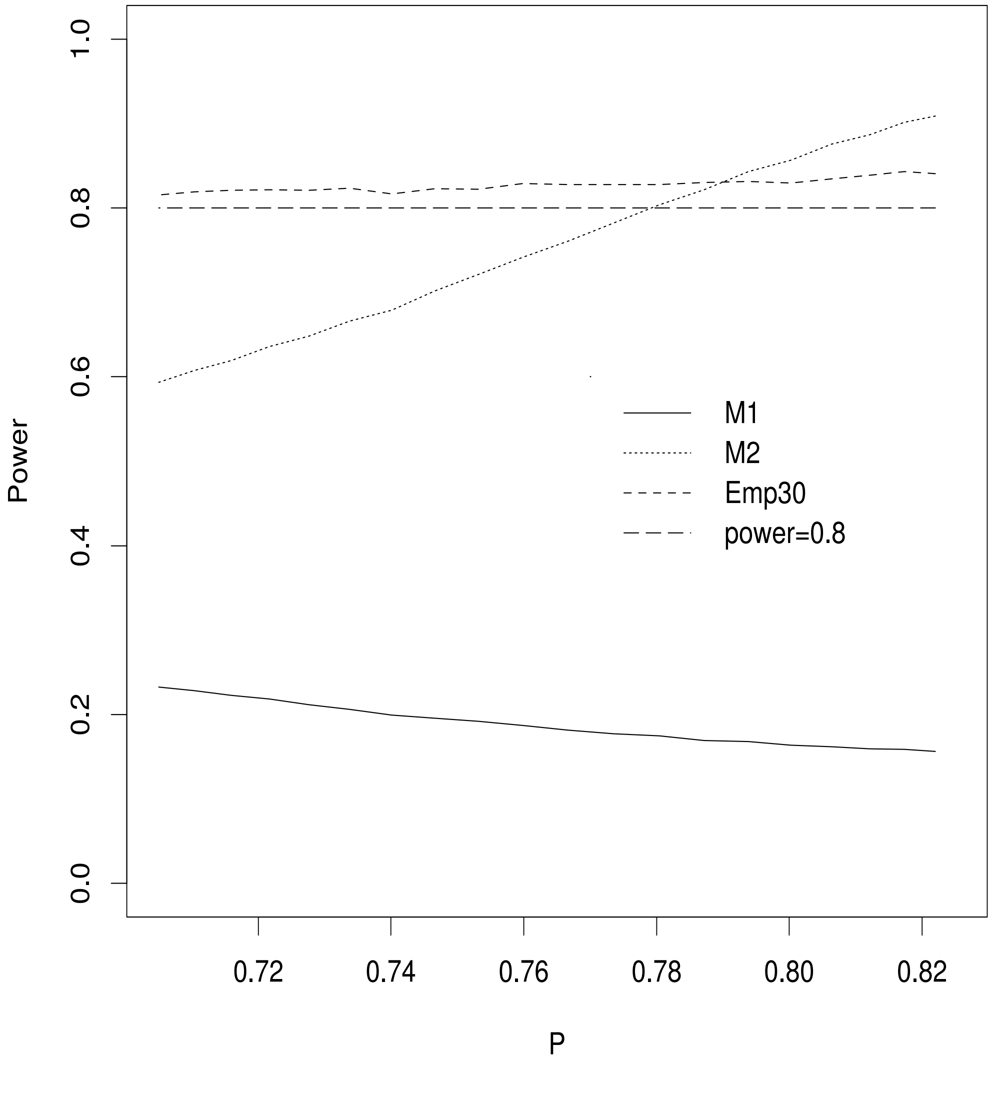


Figure 13: Power for fixed $e^{\theta_1} = 0.5$ and $e^{\theta_2} = 7$

CHAPTER V

ANALYSIS OF THE FIRE ANT DATA

5.1 Introduction

In this chapter we use ordinal logistic regression to carry out the data analysis for the fire ant data set. In Section 5.2 we briefly give the background of fire ants. Next we describe the methodology of the observational study. In Section 5.4 we discuss the statistical model. The results of the analysis of the fire ant data is given in Section 5.5. Short implementations of the sample size calculation methods are included in Section 5.6. The conclusions of the analysis are given in Section 5.7.

5.2 Background

The red imported fire ant Solenopsis invicta Buren is a soil nesting social insect native to South America. This species was accidentally introduced into the United States in the beginning of the 20th century and has become arguably one of the most destructive pests ever to invade the US. The fire ant has been spreading in all directions from Alabama, where it first landed, at an estimated rate of over 1 million hectares per year, and now occupies all the southern states (Callcott and Collins (1996)), including the recently conquered New Mexico and California. This notorious invader reached Texas in the late 1950's and currently causes an estimated annual loss of 300 million dollars to the Texas economy alone (Porter, Bhatkar, Mulder, Vinson, and Clair (1991); Li and Heinz (1998)). Hundreds of million dollars have been spent, primarily on pesticides, in fighting what is now obviously a losing war against the fire ant. It is clear that there is no easy answer to the fire ant problem. We must learn more about the pest's biology, ecology, and genetics in search of an effective and environmental safe control solution.

The understanding of the genetic basis for the fire ant's great success as an invading species and a major pest is vital to the development of an effective and sustainable control strategy. At present, however, experimental studies that require controlled mating cannot be done readily due to the many unknown environmental cues that trigger fire ant mating flights. In the natural environment, a typical mature colony has tens of thousands of workers, all sterile females, hundreds of winged males, queen-to-be, and one or more egg-laying queens (Greenberg, Vinson, and Ellison (1992); Vinson (1997)). Any time of the year when a fire ant colony matures, queens-to-be and males will, under the right environmental conditions, emerge from the home nest to embark on mating flights (Vinson (1997)). After mating on the wings a male will drop dead and a fertilized female, now a new queen, can land within a radius of 2 kilometers, sheds her wings and starts producing hundreds of eggs a day for up to 7 years (Vinson (1997)). It would take just a few mating flights to infest an area that would cost million dollars to eradicate the fire ant temporarily. Obviously, the ability to manipulate mating flights has significant value in the development of effective and environmentally sound control measures.

Weather conditions are known to influence many behaviors in humans, animals, and insects, including the fire ant. In humans, for example, fluctuations in some meteorological factors are known to influence the onset of childbirth (Driscoll (1995)); and arthritis patients can painfully feel the approach of a weather front (Aikman (1997)). Numerous winged and wingless arthropods actively take advantage of weather changes in search of abundant food sources. For example, a wingless spider mite from a depleted food source may raise its front body and become airborne when uplifting air movement is available (Li and Margolies (1994)). Another example is that a black-fly can sense the change in air pressure as little as 0.25 millibar generated by passing clouds and disperse accordingly (Wellington (1974)). Weather conditions play a major role in mating flights in some ant species, although the time of day at which mating flights occur is apparently programmed by a species-specific duel rhythm in some others. For the fire ant, early observations indicated that mating flights take place between 10 a.m. to 4 p.m. in a sunny and warm day ($> 24\ ^oC$ or $75.2\ ^oF$) with gentle wind, and often 1-2 days after a rain (Vinson (1997)). These observations suggest that fire ant mating flights are related to meteorological factors or change in the factors. We hypothesize that some meteorological factors are responsible for triggering a mating flight, while other weather conditions are merely prerequisites for the activity to occur.

In this dissertation we investigate the relationships between a large set of meteorological factors and fire ant mating flight activities, aiming at identifying cues that trigger these events. Specifically, we report our experimental procedure that is designed efficiently to capture all relevant behaviors and activities related to mating flights in field observations. We also develop a statistical model for analyzing mating flight activities in relation to a variety of weather conditions, such as, temperature, barometric pressure, the change of the pressure, relative humidity, wind speed, time of day, and rain.

5.3 Methodology for the Observational Study

5.3.1 General Characteristics of Fire Ant Mating Flights

Vinson (1997) described some results of observational study on fire ant mating flights. According to Vinson (1997), such mating flights may occur from mature colonies during any part of a year when proper conditions occur, usually in the spring and fall. These colonies produce large numbers of winged females and males referred to as reproductives, sexuals, or alates. The winged females are about 1 cm in length, brownish-red in color, and have a head just slightly smaller than the thorax. On the other hand, winged males are black and slightly smaller, and their heads are distinctly smaller than the thorax. These reproductives accumulate in the colony until induced by environmental conditions to initiate a mating flight. Vinson (1997) observed that just prior to the mating flight, workers come out through small openings and become very active on and around the surface of the nest. Males first emerge from these openings and fly away or climb surrounding vegetation to facilitate flight. Females begin to emerge an hour or so later and join the males in flight. We collected our own extensive data on fire ant mating flights to be described as follows. Our data confirm some of the conclusions in the previous study, but differ in others.

5.3.2 Fire Ant Data Collection

We collected data of the mating flights daily from April 1, 2001 to June 30, 2002. Our study covers this period of time. The data collection was setup with 3 response variables viz. Worker, Winged Male, and Winged Female. Each of these response variables had 3 levels: 0 for 'not active', 1 for 'moderately active', and 2 for 'very active', respectively. For the males and females, 1 indicates their appearance outside of mounds and 2 indicates the occurrence of mating flights. We also recorded other traits of workers, males, and females for the study.

In this research, we usually started observation from 10 a.m. to 5 p.m., the necessary duration of a day that would cover all possible flight activities. The observations were taken every 30 minutes. The data were collected by observing several mounds near the Texas A&M campus, usually 7 days a week at 30 a minutes interval. Only the nests that participated in mating flights were studied.

Table 8: Example of observations every half hour: From 9:30 a.m. to 5 p.m. on May 1, 2001 in College Station.

Date	Time	Dry Bulb Temp (F)	% Relative Humidity	Wind Speed (KT)	Barometric Pressure	Sea Level Pressure
5/1	0923	76	70	8	29.71	174
5/1	0953	78	64	9	29.71	173
5/1	1023	80	61	7	29.71	172
5/1	1053	81	58	6	29.70	170
5/1	1123	82	55	6	29.69	166
5/1	1153	84	53	6	29.67	161
5/1	1223	84	54	6	29.66	155
5/1	1253	83	55	7	29.64	150
5/1	1323	83	54	7	29.63	146
5/1	1353	88	53	6	29.62	142
5/1	1423	84	52	6	29.60	138
5/1	1453	84	51	5	29.59	134
5/1	1523	84	51	7	29.58	130
5/1	1553	84	51	9	29.57	126
5/1	1623	84	51	9	29.57	125
5/1	1653	84	51	8	29.56	125
5/1	1723	83	54	8	29.56	124

5.3.3 Weather Data Collection

We can use a number of predictor variables that describe weather conditions. The weather data were obtained at the website 'National Virtual Data System' with weather station's name 'College Station, TX' (they were recorded within 1–2 miles of our observation sites). The weather data are available every 30 minutes or one hour for 24 hours a day. There are eight meteorological variables available for analysis. Table 1 illustrates such weather data for May 1, 2001.

5.3.4 Organizing Data

The activities measured by 0, 1, or 2 are used as the response variable and meteorological factors are used as predictor variables. Rain is added as one of predictor

Table 9: Frequencies of the combined variable MF.

Activity	MF
0	6612
1	132
2	96

variables because it is an important factor to induce mating flights (Vinson (1997)). The rain variable was recorded as 0 for no rain on the previous two days and 1 for rain on either of the previous two days. In this paper, by the "previous" day we mean the duration from the previous day (including today). In addition, we expect the change of barometric pressure to be an important predictor variable. We simply define the change of barometric pressure to be $+1$ (-1) if the current pressure is higher (lower) than the pressure half hour ago, and 0 for no change. Therefore, predictor variables include dry bulb temperature, % relative humidity, wind speed, wind direction, velocity for gusts, barometric pressure, pressure tendency, sea level pressure, change of barometric pressure, and rain for a preliminary analysis. In our large data set with thousands of observations, there are more than 98% of zeros for winged male and female variables. In our statistical analysis all the observations from 10 a.m. to 5 p.m. were used. Our data set shows that only one appearance of males (at 10:30 a.m.) and no mating flights were observed between 10 a.m. to 12 p.m. throughout our observation period of 456 days. This is in contrast to other observational studies in the literature; see Vinson(1997).

The MF in the table is the combined variable of Male and Female defined to be the maximum value of the male and female's activities, since it is a good measure of overall mating flight activities. The resulting combined variable has the frequencies given in Table 2. In this paper we concentrate on studying this combined characteristic only.

5.3.5 Observed Traits

Most mating flights were observed in the springs of 2001 and 2002, one or two days after a rain. According to our observations, the winged females have a brownish-red head and thorax and dark-brown abdomen that is bigger than the head and thorax. In contrast, winged males are black and their body size is small. Under the proper conditions, workers and males usually come out of the nest first followed by females within 30 minutes. This is slightly shorter period of time than one hour or so suggested in the literature (Vinson (1997)). It appears that the temperature plays a role in the timing of female appearances. Most of them come out within 5 minutes after males in the spring and summer and in about 30 minutes in the other seasons. Workers become excited and they run continuously outside of the nest, while winged males and females climb up nearby grass or objects and take off onto mating flights. In general, males can fly away right after they come out, while females need more time to start flying. In the spring and fall, workers are generally very active all day long on the days when the mating flights occur. Around the time of flight, workers are very excited and run around the mound, while during other times they walk around. In the summer, they come out right before the males are out. Workers are even more excited when females are out. After females fly away, most workers go back into the nest. while both winged males and females can be found in the same mound, one form is usually dominant. Thus, it is likely that males and females from different colonies mate in the air. According to the literature, mating flights may occur from mature colonies during any part of a year when proper conditions occur, usually in the spring and fall. In contrast, our mating flight data were collected mostly during the spring and summer. According to our observations, most mating flights occurred when the temperature was around 28 ^oC (82 ^oF) and one or two days after a rain during time period of 1 p.m. to 5 p.m. It is somewhat different from Vinson's study that indicated that mating flights usually take place between 10 a.m. to 4 p.m. When mating flights occur, it is generally partially cloudy with mild wind. In the summer, the females look very strong and can fly far away but many males look weak and die near the nests. It is interesting to note that some fire ants make their mounds under roadways, probably because they are cool in the summer and warm in the winter.

5.4 Statistical Modeling

Let $\pi_j(\underline{X})$ denote the classification probabilities $Pr(Y = j - 1|\underline{X})$ of response variable Y, $j = 1, 2, 3$ at value $\underline{X}^T = (X_1, X_2, \cdots, X_k)$ for a set of explanatory variables $X_1, X_2, \cdots, X_k$. Here our interest is centered on the problem of relating $\underline{\pi}^T = (\pi_1(\underline{X}), \pi_2(\underline{X}), \pi_3(\underline{X}))$ to the predictor $\underline{X}$. We propose to use a form of logistic regression model for the combined variable MF. Since the combined variable MF has 3 levels of response, we consider ordinal logistic regression.

Since our response categories have a natural ordering, logit models should utilize that ordering. For this purpose we will use the proportional-odds model that is described below: the ordered multiple response model assumes the relationship

$$\text{logit}[Pr(Y \leq j - 1|\underline{X})] = \theta_j + \underline{\eta}^T \underline{X}, \qquad j = 1, 2,$$

where θ_j are two intercept parameters and $\underline{\eta}^T = (\eta_1, \eta_2, \cdots, \eta_k)$ is the slope parameter vector not including the intercept terms. By construction, $\theta_1 < \theta_2$ holds. This model fits a common slopes cumulative model that is a parallel lines regression model based on the cumulative probabilities of the response categories.

The ordinal logistic regression model in our setting is given as follows:

$$\text{logit}(\pi_1) = \log\left(\frac{\pi_1}{1 - \pi_1}\right) = \theta_1 + \eta_1 X_1 + \eta_2 X_2 + \cdots + \eta_k X_k, \tag{5.1}$$

$$\text{logit}(\pi_1 + \pi_2) = \log\left(\frac{\pi_1 + \pi_2}{1 - \pi_1 - \pi_2}\right) = \theta_2 + \eta_1 X_1 + \eta_2 X_2 + \cdots + \eta_k X_k, \tag{5.2}$$

where

$$\begin{aligned} \pi_1(\underline{X}) &= \frac{e^{\theta_1+\underline{\eta}^T\underline{X}}}{1+e^{\theta_1+\underline{\eta}^T\underline{X}}}, & (5.3) \\ \pi_1(\underline{X})+\pi_2(\underline{X}) &= \frac{e^{\theta_2+\underline{\eta}^T\underline{X}}}{1+e^{\theta_2+\underline{\eta}^T\underline{X}}}, & (5.4) \end{aligned}$$

and

$$\pi_1+\pi_2+\pi_3=1.$$

This model is known as the proportional-odds model because the odds-ratio of the event $(Y \leq j-1)$ is independent of the category indicator. We can compute the estimated odds for X_1, X_2, $\cdots$, X_k, respectively, through any standard statistical software, such as SAS. Through a formal statistical test, the empirical result given in the next section suggests that this model appears to be reasonable for analyzing our data.

Here we take $\underline{X}^T = (X_1, X_2, \cdots, X_k)$ to be the standardized meteorological variables mentioned earlier, such as the dry bulb temperature, % relative humidity, wind speed, and barometric pressure, and their squared terms, change of barometric pressure, and rain. Then, we apply the method of maximum likelihood to obtain estimates, $\hat{\theta}$'s and $\hat{\eta}$'s, for θ's and η's, respectively. Thus we can find the following estimates

$$\begin{aligned} \hat{\pi}_1(\underline{X}) &= \frac{e^{\hat{\theta}_1+\underline{\hat{\eta}}^T\underline{X}}}{1+e^{\hat{\theta}_1+\underline{\hat{\eta}}^T\underline{X}}}, & (5.5) \\ \hat{\pi}_1(\underline{X})+\hat{\pi}_2(\underline{X}) &= \frac{e^{\hat{\theta}_2+\underline{\hat{\eta}}^T\underline{X}}}{1+e^{\hat{\theta}_2+\underline{\hat{\eta}}^T\underline{X}}}. & (5.6) \end{aligned}$$

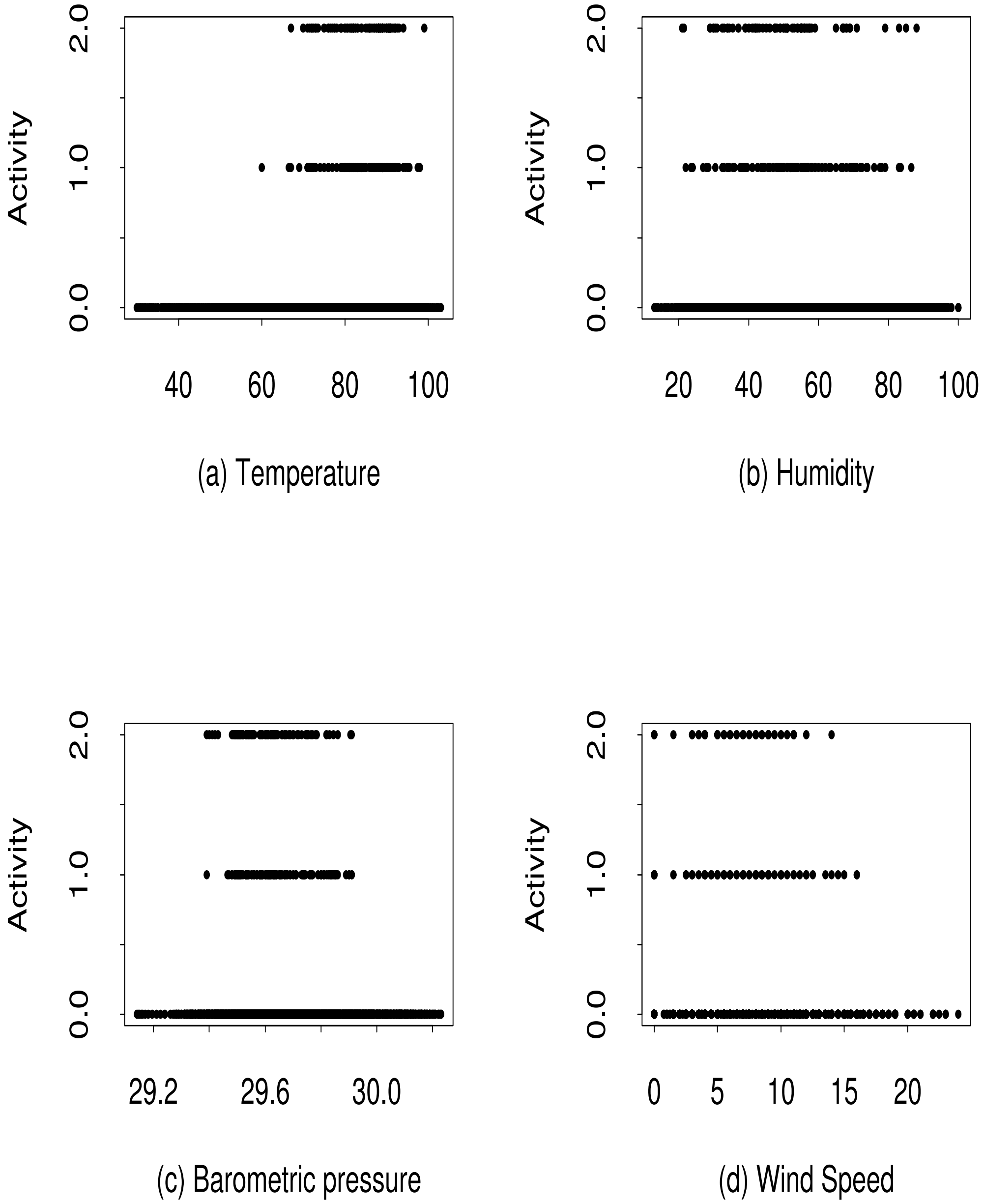
For each $\underline{X}$, $\pi_j(\underline{X}) = Pr(Y = j-1|\underline{X})$ can then be estimated by $\hat{\pi}_j$.

After a preliminary inspection and model fitting, we eliminated variables, wind direction, velocity for gusts, pressure tendency, and sea level pressure from our further analyses since they are either scientifically irrelevant or closely correlated to other independent variables in our model.

5.5 Data Analysis

We propose to apply the ordinal logistic regression approach because it is best suited for response variable of ordinal values. We wish to identify best independent variables for predicting the response variable of combined MF. Since each meteorological measure has a different scale, the standard score can be used to help analyze such a multiple logistic regression. The standardization is carried out by subtracting the sample mean from each observed variable and dividing the difference by the sample standard deviation. From a practical point of view, it is sensible to include the quadratic term for each meteorological measure in our model, since we know that the best chance of mating flight occurs somewhat in the middle of the range of each measure. Figure 1 appears to support this argument. Since it is possible that the linear and quadratic terms are strongly correlated we included the linear terms only when such a term is significant in fitting the model that has the linear and quadratic terms for that meteorological measure as the only independent variables. After fitting each of the reduced logistic regression models with only one meteorological variable and its square term each time, we concluded that the linear terms of temperature, barometric pressure, and wind speed are significant since the Wald χ^2 statistic gave a p-value less than 0.01 for each of the cases. On the other hand, for the linear term of humidity, the Wald χ^2 statistic gave a p-value of 0.054. Therefore, we opted to eliminate this term from further analyses.

Now our new model has the following independent variables: the standardized temperature (X_1), barometric pressure (X_3), and wind speed (X_5) and their squared terms, the squared standardized humidity (X_7), change of barometric pressure (X_8), and rain (X_9). Define $X_2 = X_1^2$, $X_4 = X_3^2$, and $X_6 = X_5^2$ and $\underline{X}^T = (X_1, \cdots, X_9)$. Then we can use the maximum likelihood estimation procedure to obtain parameter



(a) Temperature

(b) Humidity

(c) Barometric pressure

(d) Wind Speed

Figure 14: Scatter plots of fire ant activities against temperature, humidity, barometric pressure, and wind speed.

estimates for θ_1, θ_2, and $\underline{\eta}^T = (\eta_1, \cdots, \eta_9)$ in model (1)-(2).

From the relationship of the standardized term and the squared standardized term, we can rewrite model (1)-(2) as follows:

$$\begin{aligned}
\text{logit}(\pi_1) &= \theta_1 + \eta_1 X_1 + \eta_2 X_2 + \eta_3 X_3 + \eta_4 X_4 + \eta_5 X_5 \\
&\quad + \eta_6 X_6 + \eta_7 X_7 + \eta_8 X_8 + \eta_9 X_9 \\
&= \theta_1^* + \eta_1^* X_1 + \eta_2^* X_2^* + \eta_3^* X_3 + \eta_4^* X_4^* + \eta_5^* X_5 \\
&\quad + \eta_6^* X_6^* + \eta_7^* X_7 + \eta_8^* X_8 + \eta_9^* X_9,
\end{aligned}$$

$$\begin{aligned}
\text{logit}(\pi_1 + \pi_2) &= \theta_2 + \eta_1 X_1 + \eta_2 X_2 + \eta_3 X_3 + \eta_4 X_4 + \eta_5 X_5 \\
&\quad + \eta_6 X_6 + \eta_7 X_7 + \eta_8 X_8 + \eta_9 X_9 \\
&= \theta_2^* + \eta_1^* X_1 + \eta_2^* X_2^* + \eta_3^* X_3 + \eta_4^* X_4^* + \eta_5^* X_5 \\
&\quad + \eta_6^* X_6^* + \eta_7^* X_7 + \eta_8^* X_8 + \eta_9^* X_9,
\end{aligned}$$

where $X_2^* = (X_1 - c_1)^2$, $X_4^* = (X_3 - c_3)^2$, and $X_6^* = (X_5 - c_5)^2$ with selected c_k, $k = 1, 3, 5$. In our data analyses, we used their maximum likelihood estimates $\hat{c}_k = \hat{\eta}_k/(2\hat{\eta}_{k+1})$, so that $\hat{c}_1 = 0.34$, $\hat{c}_3 = -0.71$, and $\hat{c}_5 = -1.66$. Therefore, it is of interest to test the hypothesis that fire ant mating flights are most likely to occur when the standardized temperature, barometric pressure, and wind speed are 0.34, -0.71, and -1.66, respectively. That is, the temperature is 82.1 ^oF (27.8 ^oC), barometric pressure is 29.6 and wind speed is 1.6 KT, as is explained in the next paragraph. We call these values estimated optimal values. Then, we wish to test the equivalent hypotheses

$$H_0 : \eta_1^* = \eta_3^* = \eta_5^* = 0 \ \text{ versus } \ \mathrm{H}_1 : \text{Not } \ \mathrm{H}_0.$$

Since the p-value for the hypothesis testing result is highly insignificant, the null

Table 10: Parameter estimates for the combined variable MF with adjusted predictor variables.

Parameter	θ_1^*	θ_2^*	η_2^*	η_4^*	η_6^*	η_7^*	η_8^*	η_9^*
Estimate	3.02	3.92	1.16	0.21	0.12	0.18	0.55	-0.99

hypothesis is not rejected and the properly reduced model is

$$\text{logit}(\pi_1) = \theta_1^* + \eta_2^* X_2^* + \eta_4^* X_4^* + \eta_6^* X_6^* + \eta_7^* X_7 + \eta_8^* X_8 + \eta_9^* X_9, \tag{5.7}$$

and

$$\text{logit}(\pi_1 + \pi_2) = \theta_2^* + \eta_2^* X_2^* + \eta_4^* X_4^* + \eta_6^* X_6^* + \eta_7^* X_7 + \eta_8^* X_8 + \eta_9^* X_9.$$

This model also appears to be reasonable intuitively in light of Figure 10, which indicates that mating flights occur most likely around certain value of temperature, humidity, barometric pressure, and wind speed. Since model (1)-(2) above assumes identical effects of $\underline{X}$ for the first two categories of the response, we can use the score statistic to test such an assumption of parallel lines. The p-value of the test for the proportional odds assumption was 0.20, suggesting that the assumption is reasonable.

After running the statistical analysis, the significant predictor variables were obtained to be the re-centered squared standardized temperature, barometric pressure, wind speed, squared standardized humidity, change of barometric pressure, and rain, respectively. The estimates are shown in Table 3. The analysis quantifies the influence of each variable on mating flights. For instance, if $X_2^* = X_4^* = X_6^* = X_7 = X_8 = 0$, and $X_9 = 0$ then the estimated logit values are $\text{logit}(\hat{\pi}_1) = 3.02$ and $\text{logit}(\hat{\pi}_1 + \hat{\pi}_2) = 3.92$, and thus $\hat{\pi}_1 = 0.953$, $\hat{\pi}_2 = 0.028$, and $\hat{\pi}_3 = 0.019$. Here, $X_2^* = 0$ implies that the temperature in the actual scale is 0.34 times one sample standard deviation above the sample mean of 77.28. Likewise, $X_4^* = 0$ implies that the barometric pressure is 0.71 times one sample standard deviation below the sample mean of 29.69, $X_6^* = 0$ implies that the wind speed is 1.66 times one sample standard deviation below the sample mean of 7.72, $X_7 = 0$ implies that the % relative humidity is the sample mean of 54.14, $X_8 = 0$ implies that the barometric pressure does not change from half hour ago, and $X_9 = 0$ implies that there was no rain on the previous two days. The sample standard deviations of temperature, humidity, barometric pressure, and wind speed are 14.21, 17.51, 0.17, and 3.68, respectively. Similarly, if $X_2^* = X_4^* = X_6^* = X_7 = X_8 = 0$, and $X_9 = 1$ then $\text{logit}(\hat{\pi}_1) = 3.02 - 0.99 = 2.03$ and $\text{logit}(\hat{\pi}_1 + \hat{\pi}_2) = 3.92 - 0.99 = 2.93$ and $\hat{\pi}_1 = 0.884$, $\hat{\pi}_2 = 0.065$, and $\hat{\pi}_3 = 0.051$, where $X_9 = 1$ means that there was a rain on either (or both) of the previous two days. The estimates $\hat{\pi}_2$ (=0.100) and $\hat{\pi}_3$ (=0.085) increase dramatically when $X_8 = -1$, that is, when the barometric pressure drops. Since females and males come out of the nest in preparation of flight, the probability of mating flight activity is quite high at .185 ($\hat{\pi}_2 + \hat{\pi}_3$) under these conditions. See Table 4 for the estimated π's for selected values of $\underline{X}$.

Judging from c-statistics available in SAS our statistical analysis shows that aside from the temperature condition rain exerts the most influence for mating flights, followed by dropping in barometric pressure. That is, for males and females to embark on a mating flight, a rain in the previous two days is nearly a necessary condition and dropping in barometric pressure facilitates the timing of flight. This notion is supported by our observations that 80.2% of mating flights occurred when there was a rain on the previous day and 10.4% when there was a rain two days before and that 89.6% of mating flights occurred when barometric pressure was dropping and 7.3% of mating flights occurred when barometric pressure was unchanged. Without a rain mating flights are very unlikely to occur. For example, when $X_2^* = X_4^* = X_6^* = X_7 = X_8 = 1$, and $X_9 = 0$, the temperature is lower than 67.9 ^oF (19.9 ^oC) or higher than 96.3 ^oF (35.7 ^oC), the humidity is lower than 37% or higher than 72%, the barometric pressure is lower than 29.4 or higher than 29.7, the wind speed is higher than 5.3 KT or no wind, rising barometric pressure, and there was a no

Table 11: Example of estimated π's for the combined variable MF. The case $X_2^* = X_4^* = X_6^* = X_7 = 0$ implies that temperature is 82.1 ^oF, barometric pressure is 29.6, wind speed is 1.61, and humidity is 54.1%. Variable X_8 is for change of barometric pressure and X_9 is for rain.

X_2^*	X_4^*	X_6^*	X_7	X_8	X_9	π_1	π_2	π_3
0	0	0	0	-1	0	0.922	0.045	0.033
0	0	0	0	-1	1	0.815	0.100	0.085
0	0	0	0	0	0	0.953	0.028	0.019
0	0	0	0	0	1	0.884	0.065	0.051
0	0	0	0	1	0	0.973	0.016	0.011
0	0	0	0	1	1	0.930	0.040	0.030
0	0	0	1	-1	0	0.934	0.038	0.028
0	0	0	1	-1	1	0.840	0.088	0.072
0	0	0	1	0	0	0.961	0.023	0.016
0	0	0	1	0	1	0.901	0.056	0.043
1	1	1	0	-1	1	0.951	0.029	0.020
1	1	1	0	0	0	0.989	0.007	0.004
1	1	1	0	0	1	0.971	0.017	0.012
1	1	1	0	1	0	0.994	0.003	0.003
1	1	1	0	1	1	0.983	0.010	0.007
1	1	1	1	-1	0	0.984	0.010	0.006
1	1	1	1	-1	1	0.959	0.024	0.017
1	1	1	1	0	0	0.991	0.005	0.004
1	1	1	1	0	1	0.976	0.014	0.010
1	1	1	1	1	0	0.995	0.003	0.002
1	1	1	1	1	1	0.986	0.008	0.006

rain, the estimated probability function ($\hat{\pi}_3$) of mating flights is as low as 0.002. The table also indicates that an increase or decrease in temperature, humidity, barometric pressure, or wind speed from their estimated optimal values would lead to a decrease of $\hat{\pi}_3$ for mating flights.

Considering each meteorological variables marginally, we can obtain their ranges for high activities of mating flights. Specifically, about 90% of flights occurred when temperature was between 72-92 ^oF (22-33 ^oC). The same percentage of flights occurred when barometric pressure was between 29.5-29.9, humidity was between 30-

69%, and wind speed was between 0.7-11 KT, respectively.

5.6 Sample Size and Power for the Fire Ant Data

When the distribution for $\underline{X}$ is of a general multivariate exponential type with density

$$f(\underline{x}, \underline{\eta}) = h(\underline{x})\text{exp}\{\underline{x}'\underline{\eta} - q(\underline{\eta})\},$$

the moment-generating function for $\underline{X}$ is of the form

$$m(\underline{s}) = e^{q(\underline{\eta}+\underline{s})-q(\underline{\eta})}, \tag{5.8}$$

where $\underline{\eta}$ is a vector of parameters and q is a real-valued function of p variables whose Hessian matrix of second derivatives exists and is positive definite. The mean of $\underline{X}$ is given by the vector $\mathbf{q}^{(1)}(\underline{\eta})$ of first partials of q, evaluated at $\underline{\eta}$, and the variance of $\underline{X}$ is given by Hessian $\mathbf{q}^{(2)}(\underline{\eta})$.

First, consider the case where X_1 is temperature. We only want to test the coeffient parameter for temperature. We simply apply the formula

$$n \geq [z_\alpha + e^{-\tilde{\eta}^2/4} z_\beta]^2 / \{e^{-\theta_2}\tilde{\eta}^2\}.$$

If $\theta_1 = 3$, $\theta_2 = 4$, and $\tilde{\eta} = 1$. Then to test $H_0 : \eta = 0$ against $H_1 : \eta = -1$, the sample size can be calculated from the univariate Normal distribution and the value is $n = 107$. The adjusting for the normal covariates by corrected term is $n = 178$. If we want to consider other variables and assume that temperature, humidity, pressure, and wind speed are normally distributed and those variables are independent. Then, by the formula (4.8) approximated sample sizes are in Table 10.

Second, consider the case where X_1 is Bernoulli variable with parameter π, independent of $X_2, \ldots, X_p$, and $\tilde{X} = (X_2, \ldots, X_s) \sim N_{s-1}(\mu, \Sigma)$. This applies to our model to the rain $X_1 = 1$, or no rain $X_1 = 0$. Then if $\eta = (\eta_1, \ldots, \eta_s)'$, where

Table 12: Sample size for the combined variable MF.

Test Variables	$\tilde{\eta}$	Nuiance Variables	Nuisance Parameters	Sample Size
Temp	1		$\theta_2 = 3$	107
Temp	1	Humidity	$\theta_2 = 3$, $\eta = 0.2$	109
Temp	1	Humidity,Pressure	$\theta_2 = 3$, $\eta = (0.2, 0.2)$	111
Temp	1	Humidity,Wind Speed	$\theta_2 = 3$, $\eta = (0.2, 0.1)$	109
Temp	1	Humidity, Wind Speed Press	$\theta_2 = 3$, $\eta = (0.2, 0.1)$ 0.2	112
Temp, Humid	1, 0.2		$\theta_2 = 3$	102
Temp, Humid	1, 0.2	Pressure	$\theta_2 = 3$, $\eta = (0.2)$	104
Temp, Humid Press, Wind	1, 0.2, 0.2, 0.1		$\theta_2 = 3$	97

$\eta_1 = \text{logit}(\pi)$ and $\tilde{\eta} = (\eta_2, \ldots, \eta_s)' = \Sigma^{-1}\mu$,

$$\begin{aligned} f(\underline{x}) &= h(\underline{x})\text{exp}\{\underline{x}'\underline{\eta} - q(\underline{\eta})\} \\ &= \pi^{x_1}(1-\pi)^{1-x_1}h(\tilde{x})\text{exp}\{\tilde{x}'\Sigma^{-1}\underline{\mu} - \frac{1}{2}\underline{\mu}\Sigma^{-1}\underline{\mu}\} \\ &= h(\underline{x})\text{exp}\{x_1\text{log}\frac{\pi}{1-\pi} + \tilde{x}'\Sigma^{-1}\underline{\mu} + \log(1-\pi) - \frac{1}{2}\underline{\mu}\Sigma^{-1}\underline{\mu}\}, \end{aligned}$$

where $\underline{\eta} = (\text{log}\frac{\pi}{1-\pi}, \Sigma^{-1}\underline{\mu})$, $q(\underline{\eta}) = (-\text{log}(1-\pi) + \frac{1}{2}\underline{\mu}\Sigma^{-1}\underline{\mu}) = \left\{\log(1+e^{\eta_1}) + \frac{1}{2}\tilde{\eta}\Sigma\tilde{\eta}\right\}$, and $\underline{0}'$ is a $s-1$ vector. Thus,

$$[\mathbf{q}^{(\mathbf{2})}(\underline{\gamma} + \underline{\eta})] = \begin{bmatrix} \frac{e^{\eta_1}}{(1+e^{\eta_1})^2} & \underline{0}' \\ \underline{0} & \Sigma \end{bmatrix},$$

and

$$[\mathbf{q}^{(\mathbf{2})}(\underline{\gamma} + \underline{\eta})]_{11}^{-1} = \left[\frac{e^{\eta_1}}{(1+e^{\eta_1})^2}\right]^{-1}.$$

Hence using the independence of X_1 and $X_2, \cdots, X_s$ and $\eta = (\eta_1, \cdots, \eta_s)'$ and $\tilde{\eta} = (\eta_2, \ldots, \eta_s)'$

$$\begin{aligned} v(\eta_1) &= \left\{\frac{1}{(1-\pi) + \pi e^{\eta_1}}\right\}\text{exp}(-\tilde{\eta}'\mu - \frac{1}{2}\tilde{\eta}'\Sigma\tilde{\eta})\left(\frac{e^{\eta_1}}{(1+e^{\eta_1})^2}\right)^{-1} \\ &= \left\{\frac{1}{1-\pi} + \frac{1}{\pi e^{\eta_1}}\right\}\text{exp}(-\tilde{\eta}'\mu - \frac{1}{2}\tilde{\eta}'\Sigma\tilde{\eta}) \qquad (5.9) \end{aligned}$$

Table 13: Sample size for the combined variable MF with adjusted predictor variables.

Test variable	Nuisance variables	Sample size
Rain		313
Rain	Temp	190
Rain	Humidity	307
Rain	Wind Speed	312
Rain	Temp, Humid	186
Rain	Temp, Humid, Pressure	183

Thus to test $H_0 : \beta = (0, \eta_2, \ldots, \eta_s)$ against $H_1 : \beta = (\beta_1, \cdots, \eta_s)$, the sample size can be calculated from the univariate Bernoulli case, adjusting for the normal covariates by multiplying by a factor of $\exp(-\tilde{\eta}'\mu - 1/2\tilde{\eta}'\Sigma\tilde{\eta})$. where the test parameter under alternative hypothesis are $\eta_1 = -1$ for rain, $\eta_2 = 1$ for temperature, $\eta_3 = 0.2$ for humidity, $\eta_4 = 0.2$ for pressure, and $\eta_5 = 0.1$ for wind speed.

5.7 Conclusions

Our data analysis and results show that ordinal logistic regression appears to be a useful approach for modeling fire ants mating flights. The fitting model indicates that weather conditions, such as rain, changing in barometric pressure, time of day when mating flights occur, temperature, and wind speed are important factors that influence the initiation of fire ant mating flights. According to our analysis, the best chance for fire ant mating flights to occur is when there is mild wind (2 KT), temperature is around 82 ^oF (28 ^oC), humidity is around 54%, barometric pressure is around 29.6 and dropping, and most importantly there was a rain on the previous day or within two days. The model-identified weather conditions are in good agreement with previous observations (Vinson (1997)) that fire ant mating flights occur in warm, sunny and calm weather after a rain. The model also reveals the significance of dropping in barometric pressure in the initiation of mating flights. Similar effects of dropping barometric pressure were reported in other animals (e.g., Li and Margolies (1994); Wellington (1974)). It is likely that the effect of barometric pressure on a broad range of biological functions may be largely explained through the change of the pressure. Identification of these meteorological factors on fire ant mating flights will add our ability in monitoring population dynamics and movement of the species in a region. Knowledge of the significance of these conditions may enable scientists in creating a laboratory-based micro-environment to study behaviors and genetics that are associated with or conditioned upon mating flights. It is conceivable that the statistical modeling, analyses and reasoning found useful in our investigation of fire ant mating flights may be employed for broader use in studying the behavior of other insects and animals.

CHAPTER VI

CONCLUSIONS

6.1 Summary

In this dissertation, we have proposed sample size calculation methods for ordinal logistic regression to tests for statistical hypotheses. We have also considered to test the multiple parameters. We gave a simple closed-form formula for approximated sample sizes when the probabilities of the response categories are small. This method has been approximated by the moment generating function. It was discussed in Whittemore (1981) that the sample size for logistic regression with small response probability. We extended that approach to the ordinal response case. According to our simulation results, suggested sample size calculation methods appear to be suitable under the small response probabilities assumption. The results have been verified in Chapter IV. Since the suggested methods were derived within the limit of small response probabilities assumption, adjustments of the sample size calculation methods were needed. We have considered bias correction steps. For the binary response case, we could apply the simple closed form by the moment generating function when the response variable had small probability. If the test parameter had a large value under alternative hypothesis, our suggested bias correction method by the mean value theorem has been successfully implemented to approximate the sample size.

Furthermore, we have also considered the general case with no assumption about small probabilities of response categories. To calculate Fisher's information matrix without using some assumptions, we have employed the empirical estimation method when some data are available. Additional simulation studies have also been carried out.

6.2 Possible Extensions

We have developed the sample size calculation methods in the parametric case of covariate variables. It is conceivable that the methodology can be extended to non-parametric case of covariates. We have used a normal approximation to test statistical hypothesis. Alternatively, we can consider the score test or the likelihood ratio test. We will also consider how to control the nuisance parameters to reduce the approximation errors. Furthermore, it is possible to use a plug-in method in approximating the intractable integrals needed in the spirit of Jensen's inequality. The objective function is neither a convex nor concave function which seems to help reduce the approximation errors. Preliminary numerical results show that this simple plug-in approach appears to be a quite accurate approximation to the true sample sizes. All this will be explored in the future.

REFERENCES

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley.

Aikman, H. (1997). "The association between arthritis and the weather." *International Journal of Biometeorology*, 40, 192–199.

Bildikar, S. and Ratil, G. B. (1968). "Multivariate exponential type distributions." *American Journal of Public Health*, 55, 1993–1996.

Callcott, A. M. A. and Collins, H. L. (1996). "Invasion and range expansion of imported fire ants (Hymenoptera: Formicidae) in North America from 1918-1995." *Florida Entomologist*, 79, 240–251.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Driscoll, D. (1995). "Weather and birth - A further search for relationships." *International Journal of Biometeorology*, 38, 152–155.

Gart, J. and Tarone, R. (1983). "The relationship between score tests and approximate UMPU tests in exponential models common in biometry." *Biometrics*, 39, 781–786.

Greenberg, L., Vinson, S. B., and Ellison, S. (1992). "Nine-year study of a field containing both monogyne and polygyne red imported fire ants (Hymenoptera: Formicidae)." *Annals of the Entomological Society of America*, 85, 686–695.

Halekoh, U. (2004). "Polytomous responses - proportional odds model." 18. Biometry Research Unit, Danish Institute of Agricultural Science.

Hilton, F. J. and Mehta, R. C. (1993). "Power and sample size calcuations for exact conditional tests with ordered categorical data." *Biometrics*, 49, 609–616.

Li, J. and Heinz, K. M. (1998). "Genetic variation in desiccation resistance and adaptability of the red fire ant (Hymenoptera: Formicidae) to arid regions." *Annals of the Entomological Society of America*, 91, 726–729.

Li, J. and Margolies, D. (1994). "Barometric-pressure influences initiation of aerial dispersal in the 2-spotted spider mite." *Journal of the Kansas Entomological Society*, 67, 386–393.

Lui, K. J. (1993). "Sample size determination for multiple continuous risk factors in case-control studies." *Biometrics*, 49, 873–876.

McCullagh, P. (1980). "Regression models for ordinal data." *Journal of the Royal Statistical Society. Series B*, 42, 109–142.

Mehta, R. C., Patel, R. N., and Tsiatis, A. A. (1984). "Exact significance testing to establish treatment equivalence with ordered categorical data." *Biometrics*, 40, 819–825.

Porter, S., Bhatkar, A., Mulder, R., Vinson, S. B., and Clair, D. (1991). "Distribution and density of polygyne fire ants (Hymenoptera: Formicidae) in Texas." *Economic Entomology*, 84, 866–874.

Self, S. G. and Mauritsen, R. (1988). "Power/sample size calculations for generalized linear models." *Biometrics*, 44, 79–86.

Self, S. G., Mauritsen, R., and Ohara, J. (1992). "Power calculations for likelihood ratio tests in generalized linear models." *Biometrics*, 48, 31–39.

Spiegel, M. R. (1974). *Theory and Problems of Advanced Calculus*. Schaum's Outline Series. London: McGraw-Hill.

Vinson, S. B. (1997). "Invasion of the red imported fire ant (Hymenoptera: Formicidae). Spread, biology, and impact." *American Entomologist*, 43, 23–39.

Wellington, W. G. (1974). "Black-fly activity during cumulus-induced pressure fluctuations." *Environmental Entomology*, 3, 351–353.

Whittemore, A. S. (1981). "Sample size for logistic regression with small response probability." *Journal of American Statistical Association*, 76, 27–32.

APPENDIX A

A.1 Derivation of Covariance Matrix

This theorem is can be found in standard textbook such as Graybill (p.19).

Theorem:

If a matrix B is has submatrices such as

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

then the inverse matrix is

$$B^{-1} = \begin{bmatrix} \{B_{11} - B_{12}B_{22}^{-1}B_{21}\}^{-1} & -B_{11}^{-1}B_{12}\{B_{22} - B_{21}B_{11}^{-1}B_{12}\}^{-1} \\ -B_{22}^{-1}B_{21}\{B_{11} - B_{12}B_{22}^{-1}B_{21}\}^{-1} & \{B_{22} - B_{21}B_{11}^{-1}B_{12}\}^{-1} \end{bmatrix}.$$

By this theorem, we derive the inverse matrix of the following Hessian matrix. Since our matrix has submatrices such as

$$H(\underline{\phi}) \equiv \begin{bmatrix} m\mathbf{C_{11}} & \mathbf{C_{21}m^{(1)\prime}} \\ \mathbf{m^{(1)}C_{21}'} & \mathbf{m^{(2)}} \end{bmatrix},$$

the inverse matrix also has submatrices.

$$H^{-1} = \begin{bmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{bmatrix},$$

where

$$\begin{aligned} H^{22} &= \{\mathbf{m^{(2)}} - \mathbf{m^{(1)}C_{21}'}(m\mathbf{C_{11}})^{-1}\mathbf{C_{21}m^{(1)\prime}}\}^{-1} \\ &= \{\mathbf{m^{(2)}} - m^{-1}\mathbf{C_{21}'}\mathbf{C_{11}}^{-1}\mathbf{C_{21}m^{(1)}m^{(1)\prime}}\}^{-1}, \end{aligned}$$

and

$$\mathbf{C_{21}'}\mathbf{C_{11}}^{-1}\mathbf{C_{21}} = (\mathbf{C_{11}}^{-1})_{k-1,k-1}.$$

Therefore, if $k=2$, $(\mathbf{C_{11}}^{-1})_{k-1,k-1} = 1$. If $k=3$, $(\mathbf{C_{11}}^{-1})_{k-1,k-1} = c_{11}/\{c_{11}c_{22} - c_{12}^2\} = 1$. If $k=4$, $(\mathbf{C_{11}}^{-1})_{k-1,k-1} = \{c_{11}c_{22} - c_{12}^2\}/\{c_{11}c_{22}c_{33} - c_{23}^2c_{11} - c_{12}^2c_{33}\}$. If $k=5$, $(\mathbf{C_{11}}^{-1})_{k-1,k-1} = \{c_{11}c_{22}c_{33} - c_{23}^2c_{11} - c_{12}^2c_{33}\}/\{c_{11}c_{22}c_{33}c_{44}\}$. If $k \geq 6$,

$$\mathbf{C_{21}'}\mathbf{C_{11}}^{-1}\mathbf{C_{21}} = (\mathbf{C_{11}}^{-1})_{k-1,k-1} = \frac{1}{c_{k-1,k-1}}.$$

Example:

When the distribution for $\underline{X}$ is of a general multivariate exponential type, the moment-generating function for $\underline{X}$ is of the form

$$f(\underline{X}) = h(\underline{X})\text{exp}\{\underline{X}'\underline{\gamma} - q(\underline{\gamma})\}$$

the moment-generating function for $\underline{X}$ is of the form (Bildikar and Patil (1968))

$$\begin{aligned} m(\underline{\eta}) &= E(e^{\underline{x}'\underline{\eta}}) \\ &= \int e^{\underline{x}'\underline{\eta}} h(\underline{x})\text{exp}\{\underline{x}'\underline{\gamma} - q(\underline{\gamma})\}d\underline{x} \\ &= e^{-q(\underline{\gamma})}\int h(\underline{x})\text{exp}\{\underline{x}'\underline{\gamma} + \underline{x}'\underline{\eta}\}d\underline{x} \\ &= \text{exp}\{q(\underline{\gamma}+\underline{\eta}) - q(\underline{\gamma})\}, \end{aligned}$$

$$\mathbf{m^{(1)}}(\underline{\eta}) = \mathbf{q^{(1)}}(\underline{\gamma}+\underline{\eta}) \times m(\underline{\eta}),$$

and

$$\mathbf{m^{(2)}}(\underline{\eta}) = [\mathbf{q^{(1)}}(\underline{\gamma}+\underline{\eta})\mathbf{q^{(1)}}'(\underline{\gamma}+\underline{\eta}) + \mathbf{q^{(2)}}(\underline{\gamma}+\underline{\eta})]m(\underline{\eta}).$$

Then

$$\begin{aligned} H^{22} &= \{\mathbf{m^{(2)}}(\underline{\eta}) - m^{-1}\mathbf{m^{(1)}}(\underline{\eta})\mathbf{m^{(1)}}'(\underline{\eta})\}^{-1} \\ &= e^{-\{q(\underline{\gamma}+\underline{\eta})-q(\underline{\gamma})\}}[\mathbf{q^{(2)}}(\underline{\gamma}+\underline{\eta})]^{-1}, \end{aligned}$$

for $k = 2, 3$.

If $\underline{X} \sim N_s(\underline{\mu}, \Sigma)$,

$$\begin{aligned} f(\underline{x}) &= (2\pi)^{s/2}|\Sigma|^{-1/2}e^{-\dfrac{1}{2}(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})} \\ &= h(\underline{x})\exp\{\underline{x}'\Sigma^{-1}\underline{\mu} - \frac{1}{2}\underline{\mu}'\Sigma^{-1}\underline{\mu}\}, \end{aligned}$$

where $\underline{\gamma} = \Sigma^{-1}\underline{\mu}$, $q(\underline{\gamma}) = \underline{\mu}'\Sigma^{-1}\underline{\mu} = \underline{\gamma}'\Sigma^{-1}\underline{\gamma}$,

$$m(\underline{\eta}) = \exp\left\{\frac{1}{2}\underline{\eta}'\Sigma\underline{\eta} + \underline{\mu}'\underline{\eta}\right\},$$

$\mathbf{q^{(1)}}(\underline{\gamma} + \underline{\eta}) = \Sigma\underline{\eta} + \underline{\mu}$, $\mathbf{q^{(2)}}(\underline{\gamma} + \underline{\eta}) = \Sigma$, and

$$H^{22} = e^{-\{\underline{\eta}'\Sigma\underline{\eta}/2+\underline{\mu}'\underline{\eta}\}}\Sigma^{-1},$$

for $k = 2$ or 3. When $k \geq 6$,

$$\begin{aligned} H^{22} &= \{\mathbf{m^{(2)}}(\underline{\eta}) - [mc_{k-1,k-1}]^{-1}\mathbf{m^{(1)}}(\underline{\eta})\mathbf{m^{(1)}}'(\underline{\eta})\}^{-1} \\ &= e^{-\{q(\underline{\gamma}+\underline{\eta})-q(\underline{\gamma})\}}[\mathbf{q^{(2)}}(\underline{\gamma} + \underline{\eta}) + \frac{e^{k-2}}{e^{k-1}}\mathbf{q^{(1)}}(\underline{\gamma} + \underline{\eta})\mathbf{q^{(1)}}(\underline{\gamma} + \underline{\eta})']^{-1} \end{aligned}$$

and

$$H^{22} = e^{-\{\underline{\eta}'\Sigma\underline{\eta}/2+\underline{\mu}'\underline{\eta}\}}\left[\Sigma + \frac{e^{\theta_{k-2}}}{e^{\theta_{k-1}}}(\Sigma\underline{\eta} + \underline{\mu})(\Sigma\underline{\eta} + \underline{\mu})'\right]^{-1}.$$

APPENDIX B

B.1 Calculations of Corrected Terms

B.1.1 Binary Response Case with One Covariate

When we have a binary response variable ($k = 2$), the elements of Fisher information matrix is

$$I_{ij} = nE\left(\frac{X_i X_j e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}\right), \quad i, j = 1, \ldots, s,$$

where $X_1 = 1$. When we have one covariate, consider the problem of testing the null hypothesis $H_0 : \eta = 0$ against the one-sided alternative $H_A : \eta = \tilde{\eta}$ to test at level α with power $1 - \beta$. Then

$$\begin{aligned}
\mathrm{Var}(\eta) &= \frac{nE\left(\dfrac{e^{\theta_1+\eta' x}}{(1+e^{\theta_1+\eta' x})^2}\right)}{n^2 E\left(\dfrac{e^{\theta_1+\eta' x}}{(1+e^{\theta_1+\eta' x})^2}\right) E\left(\dfrac{X_i X_j e^{\theta_1+\eta' x}}{(1+e^{\theta_1+\eta' x})^2}\right) - n^2 E\left(\dfrac{X_i e^{\theta_1+\eta' x}}{(1+e^{\theta_1+\eta' x})^2}\right)^2} \\
&= \frac{1}{n} v(\eta).
\end{aligned}$$

The full expansion for the right-hand side of $\dfrac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2}$ in power of $e^{\theta_1+\eta' x}$ is

$$\frac{e^{\theta_1+\underline{\eta}'\underline{x}}}{(1+e^{\theta_1+\underline{\eta}'\underline{x}})^2} = \sum_{l=1}^{\infty} (-1)^{l+1} l (e^{\theta_1+\underline{\eta}'\underline{x}})^l.$$

Then

$$v_\infty(\eta) = \frac{\sum(-1)^{l+1} l e^{l\theta_1} m(l\eta)}{[\sum(-1)^{l+1} l e^{l\theta_1} m(l\eta)][\sum(-1)^{l+1} l e^{l\theta_1} m_{11}(l\eta)] - [\sum(-1)^{l+1} l e^{l\theta_1} m_1(l\eta)]^2}$$

and

$$v_1(\eta) = \frac{m(\eta)}{m(\eta) m_{11}(\eta) - m_1(\eta)^2}.$$

Letting $\epsilon = e^{\theta_1}$, we have

$$v(\eta)^* = \frac{m(\eta) - 2\epsilon m(2\eta)}{D} + O(\epsilon^2),$$

where

$$\begin{aligned}
D &= [m(\eta) - 2\epsilon m(2\eta) + O(\epsilon^2)][m_{11}(\eta) - 2\epsilon m_{11}(2\eta) + O(\epsilon^2)] \\
&\quad -[m_1(\eta) - 2\epsilon m_1(2\eta) + O(\epsilon^2)]^2 \\
&= m(\eta)m_{11}(\eta) - m_1(\eta)^2 \\
&\quad -2\epsilon\{m(\eta)m_{11}(2\eta) + m(2\eta)m_{11}(\eta) - 2m_1(\eta)m_1(2\eta)\} + O(\epsilon^2) \\
&= \{m(\eta)m_{11}(\eta) - m_1(\eta)^2\} \\
&\quad \times\left\{1 - \frac{2\epsilon[m(2\eta)m_{11}(\eta) + m(\eta)m_{11}(2\eta) - 2m_1(\eta)m_1(2\eta)]}{m(\eta)m_{11}(\eta) - m_1(\eta)^2}\right\} + O(\epsilon^2) \\
&= \frac{m(\eta)}{v_1(\eta)}\left\{1 - 2\epsilon\frac{v_1(\eta)}{m(\eta)}[m(2\eta)m_{11}(\eta) + m(\eta)m_{11}(2\eta) - 2m_1(\eta)m_1(2\eta)]\right\} \\
&\quad +O(\epsilon^2)
\end{aligned}$$

and

$$\begin{aligned}
v(\eta)^* &= \frac{m(\eta) - 2\epsilon m(2\eta)}{\frac{m(\eta)}{v_1(\eta)}\left\{1 - 2\epsilon\frac{v_1(\eta)}{m(\eta)}[m(2\eta)m_{11}(\eta) + m(\eta)m_{11}(2\eta) - 2m_1(\eta)m_1(2\eta)]\right\}} \\
&\quad +O(\epsilon^2) \\
&= v(\eta)\frac{[1 - 2\epsilon m(2\eta)m^{-1}(\eta)]}{1 - 2\epsilon v_1(\eta)[m_{11}(2\eta) + m^{-1}(\eta)m(2\eta)m_{11}(\eta) - 2m^{-1}(\eta)m_1(\eta)m_1(2\eta)]} \\
&\quad +O(\epsilon^2).
\end{aligned}$$

Use of the binomial expansion $(1-x)^{-1} = 1 + x + O(x^2)$,

$$\begin{aligned}
v(\eta)^* &= v(\eta)[1 - 2\epsilon m(2\eta)m^{-1}(\eta)][1 + 2\epsilon v(\eta)\{m_{11}(2\eta) \\
&\quad +m^{-1}(\eta)m(2\eta)m_{11}(\eta) - 2m^{-1}(\eta)m_1(\eta)m_1(2\eta)\}] + O(\epsilon^2) \\
&= v_1(\eta)[1 + 2\epsilon v_1(\eta)\{m_{11}(2\eta) + m^{-2}(\eta)m(2\eta)m_1^2(\eta) \\
&\quad -2m^{-1}(\eta)m_1(\eta)m_1(2\eta)\}] + O(\epsilon^2),
\end{aligned}$$

or

$$v(\eta)^* = v_1(\eta)[1 + 2\epsilon R(\eta)] + O(\epsilon^2),$$

where $R(\eta) = v_1(\eta)[m_{11}(2\eta) + m^{-2}(\eta)m(2\eta)m_1^2(\eta) - 2m^{-1}(\eta)m_1(\eta)m_1(2\eta)]$ is of the required form.

B.1.2 The Binomial Expansion

The binomial expansion is also can be found in standard textbook. If we have an expression of the form $(1+x)^n$, where $x^2 < 1$, then we can replace the expression using the binomial expansion. Here n can be an integer or half-integer, positive or negative. When the magnitude of x is less than one, we can write the following power series in x:

$$(1+x)^n = 1 + nx + \frac{1}{2}n(n-1)x^2 + \cdots$$

Notice that if the magnitude of x is less than one, higher powers of x are smaller than x. Therefore, the series expansion can be terminated with a finite number of terms where the remaining terms are negligible. For example, we may make the following replacement:

$$(1+x)^n = 1 + nx + \frac{1}{2}n(n-1)x^2$$

Note that the additional higher power terms are now missing - they were ”negligible.”

B.1.3 Three Response Categories Case with One Covariate with Small Response Probabilities

It is likewise the binary case, we can derive the information matrix. The log-likelihood has following form:

$$\begin{aligned}
\log \mathrm{L} &= \sum_{v=1}^{n} \Bigg\{ R_{v1} \log\bigg(\frac{e^{\theta_1+\eta x_v}}{1+e^{\theta_1+\eta x_v}} \bigg) \\
&\quad + (R_{v2} - R_{v1}) \log\bigg(\frac{e^{\theta_2+\eta x_v}}{1+e^{\theta_2+\eta x_v}} - \frac{e^{\theta_1+\eta x_v}}{1+e^{\theta_1+\eta x_v}} \bigg) \\
&\quad + (1 - R_{v2}) \log\bigg(1 - \frac{e^{\theta_2+\eta x_v}}{1+e^{\theta_2+\eta x_v}} \bigg) \\
&= \sum_{v=1}^{n} \Bigg\{ R_{v1} \bigg(\theta_1 + \eta x_v - \log[1 + e^{\theta_1+\eta x_v}] \bigg) \\
&\quad + (R_{v2} - R_{v1}) \bigg(\eta x_v + \log(e^{\theta_2} - e^{\theta_1}) - \log(1 + e^{\theta_2+\eta x_v}) - \log(1 + e^{\theta_1+\eta x_v}) \bigg) \\
&\quad - (1 - R_{v2}) \log(1 + e^{\theta_2+\eta x_v}) \Bigg\}
\end{aligned}$$

and the derivatives are

$$\frac{\partial \log \mathrm{L}}{\partial \theta_1} = \sum_{v=1}^{n} \Bigg\{ R_{v1} \bigg(1 - \frac{e^{\theta_1+\eta x_v}}{1+e^{\theta_1+\eta x_v}} \bigg) + (R_{v2} - R_{v1}) \bigg(-\frac{e^{\theta_1}}{e^{\theta_2} - e^{\theta_1}} - \frac{e^{\theta_1+\eta x_v}}{1+e^{\theta_1+\eta x_v}} \bigg) \Bigg\},$$

$$\frac{\partial \log \mathrm{L}}{\partial \theta_2} = \sum_{v=1}^{n} \Bigg\{ (R_{v2} - R_{v1}) \bigg(\frac{e^{\theta_2}}{e^{\theta_2} - e^{\theta_1}} - \frac{e^{\theta_2+\eta x_v}}{1+e^{\theta_2+\eta x_v}} \bigg) - (1 - R_{v2}) \frac{e^{\theta_2+\eta x_v}}{1+e^{\theta_2+\eta x_v}} \bigg) \Bigg\},$$

$$\begin{aligned}
\frac{\partial \log \mathrm{L}}{\partial \eta} &= \sum_{v=1}^{n} \Bigg\{ R_{v1} \bigg(x_v - \frac{x_v e^{\theta_1+\eta x_v}}{1+e^{\theta_1+\eta x_v}} \bigg) + (R_{v2} - R_{v1}) \bigg(x_v - \frac{x_v e^{\theta_1+\eta x_v}}{1+e^{\theta_1+\eta x_v}} \\
&\quad - \frac{x_v e^{\theta_2+\eta x_v}}{1+e^{\theta_2+\eta x_v}} \bigg) - (1 - R_{v2}) \bigg(\frac{x_v e^{\theta_2+\eta x_v}}{1+e^{\theta_2+\eta x_v}} \bigg) \Bigg\},
\end{aligned}$$

$$\frac{\partial^2 \log \mathrm{L}}{\partial \theta_1^2} = \sum_{v=1}^{n} \Bigg\{ - R_{v1} \frac{e^{\theta_1+\eta x_v}}{(1+e^{\theta_1+\eta x_v})^2} - (R_{v2} - R_{v1}) \bigg(\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2} - e^{\theta_1})^2} + \frac{e^{\theta_1+\eta x_v}}{(1+e^{\theta_1+\eta x_v})^2} \bigg) \Bigg\},$$

$$\begin{aligned}
\frac{\partial^2 \log \mathrm{L}}{\partial \theta_2^2} &= \sum_{v=1}^{n} \Bigg\{ - (R_{v2} - R_{v1}) \bigg(\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2} - e^{\theta_1})^2} + \frac{e^{\theta_2+\eta x_v}}{(1+e^{\theta_2+\eta x_v})^2} \bigg) \\
&\quad - (1 - R_{v2}) \frac{e^{\theta_2+\eta x_v}}{(1+e^{\theta_2+\eta x_v})^2} \bigg) \Bigg\},
\end{aligned}$$

$$\frac{\partial^2 \text{log L}}{\partial\theta_1\partial\theta_2} = \sum_{v=1}^{n}\bigg\{(R_{v2}-R_{v1})\frac{e^{\theta_1+\theta_2}}{(e^{\theta_2}-e^{\theta_1})^2}\bigg\},$$

$$\frac{\partial^2 \text{log L}}{\partial\theta_1\partial\eta} = \sum_{v=1}^{n}\bigg\{-R_{v2}x_v\frac{e^{\theta_1+\eta x_v}}{(1+e^{\theta_1+\eta x_v})^2}\bigg\},$$

$$\frac{\partial^2 \text{log L}}{\partial\theta_2\partial\eta} = \sum_{v=1}^{n}\bigg\{-(1-R_{v1})x_v\frac{e^{\theta_2+\eta x_v}}{(1+e^{\theta_2+\eta x_v})^2}\bigg\},$$

and

$$\frac{\partial^2 \text{log L}}{\partial\eta^2} = \sum_{v=1}^{n}\bigg\{-R_{v2}\frac{x_v^2 e^{\theta_1+\eta x_v}}{(1+e^{\theta_1+\eta x_v})^2}-(1-R_{v1})\frac{x_v^2 e^{\theta_2+\eta x_v}}{(1+e^{\theta_2+\eta x_v})^2}\bigg)\bigg\}.$$

The elements of the information matrix are

$$I_{11} = -\text{E}\bigg[\frac{\partial^2 \text{log L}}{\partial\theta_1^2}\bigg],$$

$$I_{12} = -\text{E}\bigg[\frac{\partial^2 \text{log L}}{\partial\theta_1\partial\theta_2}\bigg],$$

$$I_{13} = -\text{E}\bigg[\frac{\partial^2 \text{log L}}{\partial\theta_1\partial\eta}\bigg],$$

$$I_{22} = -\text{E}\bigg[\frac{\partial^2 \text{log L}}{\partial\theta_2^2}\bigg],$$

$$I_{23} = -\text{E}\bigg[\frac{\partial^2 \text{log L}}{\partial\theta_2\partial\eta}\bigg],$$

and

$$I_{33} = -\text{E}\bigg[\frac{\partial^2 \text{log L}}{\partial\eta^2}\bigg].$$

Use of Taylor's expansion in power of $e^{\theta_j+\eta X}$

$$\frac{e^{\theta_j+\eta X}}{(1+e^{\theta_j+\eta X})} = \sum_{v=1}^{\infty}\frac{1}{v!}\{e^{\theta_j+\eta X}\}^v g^{(v)}(0) = \sum_{v=1}^{\infty}\frac{1}{v!}g^{(v)}(0)\{e^{\theta_j}\}^v\{e^{\eta X}\}^v = \sum_{v=1}^{\infty}(-1)^{v+1}\epsilon_j^v e^{v\eta X}$$

where $\epsilon_j = e^{\theta_j} << 1$, $g^{(v)}(z) = \dfrac{d^v}{dz^v}\left(\dfrac{z}{1+z}\right)$, $g^{(v)}(0) = (-1)^{v+1}v!$, $g(0) = 0$, $\dfrac{1}{v!}g^{(v)}(0) = (-1)^{v+1}$, $v = 1, \cdots, \infty$.

$$\frac{e^{\theta_j+\eta X}}{(1+e^{\theta_j+\eta X})^2} = \sum_{w=1}^{\infty}\frac{1}{w!}\{e^{\theta_j+\eta X}\}^w f^{(w)}(0) = \sum_{w=1}^{\infty}(-1)^{w+1}w\epsilon_j^w e^{w\eta X},$$

where $\epsilon_j = e^{\theta_j} << 1$, $f^{(w)}(z) = \dfrac{d^w}{dz^w}\left(\dfrac{z}{(1+z)^2}\right)$, $f^{(w)}(0) = (-1)^w w! + (-1)^{w+1}(w+1)!$, $f(0) = 0$, and $1/w! f^{(w)}(0) = (-1)^{w+1} w$. Then the full expansion for the right-hand side of I_{ij} in powers of $e^{\theta_j + \eta X}$ is

$$
\begin{aligned}
I_{11} &= nE\Bigg[\sum_{v=1}^{\infty}(-1)^{v+1}(\epsilon_2^v - \epsilon_1^v)\frac{\epsilon_1\epsilon_2}{(\epsilon_2-\epsilon_1)^2}e^{v\eta X} \\
&\quad + \sum_{v=1}^{\infty}(-1)^{v+1}\epsilon_2^v e^{v\eta X}\sum_{w=1}^{\infty}(-1)^{w+1} w\epsilon_1^w e^{w\eta X}\Bigg] \\
&= n\epsilon_1\epsilon_2 E\Bigg[\frac{1}{(\epsilon_2-\epsilon_1)}e^{\eta X} - \frac{(\epsilon_1+\epsilon_2)}{(\epsilon_2-\epsilon_1)}e^{2\eta X} + O(\epsilon_2^2) \\
&\quad + \{e^{\eta X} - \epsilon_2 e^{2\eta X} + O(\epsilon_2^2)\}\{e^{\eta X} - 2\epsilon_1 e^{2\eta X} + O(\epsilon_1^2)\}\Bigg] \\
&= n\epsilon_1\epsilon_2 E\Bigg[\frac{1}{(\epsilon_2-\epsilon_1)}e^{\eta X} - \frac{2\epsilon_1}{(\epsilon_2-\epsilon_1)}e^{2\eta X} + O(\epsilon_2^2)\Bigg] \\
&= n\frac{\epsilon_1\epsilon_2}{(\epsilon_2-\epsilon_1)}\Bigg[m(\eta) - 2\epsilon_1 m(2\eta) + O(\epsilon_2^2)\Bigg],
\end{aligned}
$$

$$
\begin{aligned}
I_{12} &= -nE\Bigg[\sum_{v=1}^{\infty}(-1)^{v+1}(\epsilon_2^v - \epsilon_1^v)\frac{\epsilon_1\epsilon_2}{(\epsilon_2-\epsilon_1)^2}e^{v\eta X}\Bigg] \\
&= -n\frac{\epsilon_1\epsilon_2}{(\epsilon_2-\epsilon_1)}\Bigg[m(\eta) - (\epsilon_1+\epsilon_2)m(2\eta) + O(\epsilon_2^2)\Bigg],
\end{aligned}
$$

$$
\begin{aligned}
I_{13} &= nE\Bigg[X_v\sum_{v=1}^{\infty}(-1)^{v+1}\epsilon_2^v e^{v\eta X}\sum_{w=1}^{\infty}(-1)^{w+1} w\epsilon_1^w e^{w\eta X}\Bigg] \\
&= n\epsilon_2\Bigg[\epsilon_1 m_1(2\eta) + O(\epsilon_2^2)\Bigg],
\end{aligned}
$$

$$\begin{aligned}
I_{22} &= nE\Big[\sum_{v=1}^{\infty}(-1)^{v+1}(\epsilon_2^v - \epsilon_1^v)\frac{\epsilon_1\epsilon_2}{(\epsilon_2-\epsilon_1)^2}e^{v\eta X} \\
&\quad +(1-\sum_{v=1}^{\infty}(-1)^{v+1}\epsilon_1^v e^{v\eta X})\sum_{w=1}^{\infty}(-1)^{w+1}w\epsilon_2^w e^{w\eta X}\Big] \\
&= n\epsilon_1\epsilon_2 E\Big[\frac{1}{(\epsilon_2-\epsilon_1)}e^{\eta X} - \frac{(\epsilon_1+\epsilon_2)}{(\epsilon_2-\epsilon_1)}e^{2\eta X} + O(\epsilon_2^2) \\
&\quad +\frac{1}{\epsilon_1}\{e^{\eta X} - 2\epsilon_2 e^{2\eta X} + O(\epsilon_2^2)\} \\
&\quad -\{e^{\eta X} - \epsilon_1 e^{2\eta X} + O(\epsilon_1^2)\}\{e^{\eta X} - 2\epsilon_2 e^{2\eta X} + O(\epsilon_2^2)\}\Big] \\
&= n\frac{\epsilon_2^2}{\epsilon_2-\epsilon_1}\Big[m(\eta) - 2\epsilon_2 m(2\eta) + O(\epsilon_2^2)\Big],
\end{aligned}$$

$$\begin{aligned}
I_{23} &= nE\Big[X_v(1-\sum_{v=1}^{\infty}(-1)^{v+1}\epsilon_1^v e^{v\eta X})\sum_{w=1}^{\infty}(-1)^{w+1}w\epsilon_2^w e^{w\eta X}\Big] \\
&= n\epsilon_1\epsilon_2 E\Big[\frac{X_v}{\epsilon_1}\{e^{\eta X} - 2\epsilon_2 e^{2\eta X} + O(\epsilon_2^2)\} \\
&\quad -X_v\{e^{\eta X} - \epsilon_1 e^{2\eta X} + O(\epsilon_1^2)\}\{e^{\eta X} - 2\epsilon_2 e^{2\eta X} + O(\epsilon_2^2)\}\Big] \\
&= n\epsilon_2\Big[m_1(\eta) - (\epsilon_1 + 2\epsilon_2)m_1(2\eta) + O(\epsilon_2^2)\Big],
\end{aligned}$$

and

$$\begin{aligned}
I_{33} &= nE\Big[X_v^2\sum_{v=1}^{\infty}(-1)^{v+1}\epsilon_2^v e^{v\eta X}\sum_{w=1}^{\infty}(-1)^{w+1}w\epsilon_1^w e^{w\eta X} \\
&\quad +X_v^2(1-\sum_{v=1}^{\infty}(-1)^{v+1}\epsilon_1^v e^{v\eta X})\sum_{w=1}^{\infty}(-1)^{w+1}w\epsilon_2^w e^{w\eta X}\Big] \\
&= n\epsilon_1\epsilon_2 E\Big[X_v^2\{e^{\eta X} - \epsilon_2 e^{2\eta X} + O(\epsilon_2^2)\}\{e^{\eta X} - 2\epsilon_1 e^{2\eta X} + O(\epsilon_1^2)\} \\
&\quad +\frac{X_v^2}{\epsilon_1}\{e^{\eta X} - 2\epsilon_2 e^{2\eta X} + O(\epsilon_2^2)\} \\
&\quad -X_v^2\{e^{\eta X} - \epsilon_1 e^{2\eta X} + O(\epsilon_1^2)\}\{e^{\eta X} - 2\epsilon_2 e^{2\eta X} + O(\epsilon_2^2)\}\Big] \\
&= n\epsilon_2\Big[m_{11}(\eta) - 2\epsilon_2 m_{11}(2\eta) + O(\epsilon_2^2)\Big].
\end{aligned}$$

Use of the information matrix, variance of $\hat{\eta}$ is calculated as follows:

$$\text{var}(\hat{\eta}) = \frac{I_{11}I_{22} - I_{12}^2}{I_{11}I_{22}I_{33} + 2I_{12}I_{23}I_{13} - I_{13}^2I_{22} - I_{12}^2I_{33} - I_{23}^2I_{11}},$$

where $I_{11}I_{22} = n^2 \frac{\epsilon_1\epsilon_2^3}{(\epsilon_2-\epsilon_1)^2}\left[m^2(\eta) - 2(\epsilon_1+\epsilon_2)m(\eta)m(2\eta) + O(\epsilon_2^2)\right],$

$$I_{12}^2 = n^2 \frac{\epsilon_1^2\epsilon_2^2}{(\epsilon_2-\epsilon_1)^2}\left[m^2(\eta) - 2(\epsilon_1+\epsilon_2)m(\eta)m(2\eta) + O(\epsilon_2^2)\right],$$

$$I_{11}I_{22} - I_{12}^2 = n^2 \frac{\epsilon_1\epsilon_2^2}{(\epsilon_2-\epsilon_1)}\left[m^2(\eta) - 2(\epsilon_1+\epsilon_2)m(\eta)m(2\eta) + O(\epsilon_2^2)\right],$$

$$\begin{aligned}
I_{11}I_{22}I_{33} &= n^3 \frac{\epsilon_1\epsilon_2^4}{(\epsilon_2-\epsilon_1)^2} m(\eta)\Big[m(\eta)m_{11}(\eta) \\
&\quad -2(\epsilon_1+\epsilon_2)m(2\eta)m_{11}(\eta) - 2\epsilon_2 m(\eta)m_{11}(2\eta) + O(\epsilon_2^2)\Big] \\
&= n^3 \frac{\epsilon_1\epsilon_2^3}{(\epsilon_2-\epsilon_1)} m(\eta)\Big[\frac{\epsilon_2}{(\epsilon_2-\epsilon_1)} m(\eta)m_{11}(\eta) \\
&\quad -2\frac{\epsilon_2}{(\epsilon_2-\epsilon_1)}(\epsilon_1+\epsilon_2)m(2\eta)m_{11}(\eta) - 2\frac{\epsilon_2}{(\epsilon_2-\epsilon_1)}\epsilon_2 m(\eta)m_{11}(2\eta) + O(\epsilon_2^2)\Big],
\end{aligned}$$

$$I_{12}I_{23}I_{13} = -n^3 \frac{\epsilon_1\epsilon_2^3}{(\epsilon_2-\epsilon_1)} m(\eta)\left[\epsilon_1 m_1(\eta)m_1(2\eta) + O(\epsilon_2^2)\right],$$

$$I_{13}^2I_{22} = n^3 \frac{\epsilon_1^2\epsilon_2^2}{(\epsilon_2-\epsilon_1)}\left[O(\epsilon_2^2)\right],$$

$$\begin{aligned}
I_{12}^2I_{33} &= n^3 \frac{\epsilon_1^2\epsilon_2^3}{(\epsilon_2-\epsilon_1)^2}\Big[m^2(\eta)m_{11}(\eta) - 2(\epsilon_1+\epsilon_2)m(\eta)m(2\eta)m_{11}(\eta) \\
&\quad -2\epsilon_2 m^2(\eta)m_{11}(2\eta) + O(\epsilon_2^2)\Big] \\
&= n^3 \frac{\epsilon_1\epsilon_2^3}{(\epsilon_2-\epsilon_1)} m(\eta)\Big[\frac{\epsilon_1}{(\epsilon_2-\epsilon_1)} m(\eta)m_{11}(\eta) - 2\frac{\epsilon_1}{(\epsilon_2-\epsilon_1)}(\epsilon_1+\epsilon_2)m(2\eta)m_{11}(\eta) \\
&\quad -2\frac{\epsilon_1}{(\epsilon_2-\epsilon_1)}\epsilon_2 m(\eta)m_{11}(2\eta) + O(\epsilon_2^2)\Big],
\end{aligned}$$

$$\begin{aligned}
I_{23}^2I_{11} &= n^3 \frac{\epsilon_1\epsilon_2^3}{(\epsilon_2-\epsilon_1)} m(\eta)\Big[m_1^2(\eta) - 2(\epsilon_1+2\epsilon_2)m_1(\eta)m_1(2\eta) \\
&\quad -2\epsilon_2 m(2\eta)m_1^2(\eta)/m(\eta) + O(\epsilon_2^2)\Big],
\end{aligned}$$

and

$$\begin{aligned}
& I_{11}I_{22}I_{33} + 2I_{12}I_{23}I_{13} - I_{13}^2 I_{22} - I_{12}^2 I_{33} - I_{23}^2 I_{11} \\
= \; & n^3 \frac{\epsilon_1 \epsilon_2^3}{(\epsilon_2 - \epsilon_1)} m(\eta) \Big[\{m(\eta)m_{11}(\eta) - m_1^2(\eta)\} - 2(\epsilon_1 + \epsilon_2) m(2\eta) m_{11}(\eta) \\
& +2\epsilon_2 \{2m_1(\eta)m_1(2\eta) - m(\eta)m_{11}(2\eta) + m(2\eta)m_1^2(\eta)/m(\eta)\} \Big] \\
= \; & n^3 \frac{\epsilon_1 \epsilon_2^3}{(\epsilon_2 - \epsilon_1)} m(\eta) \Big[\{m(\eta)m_{11}(\eta) - m_1^2(\eta)\} - 2\epsilon_1 A_1 - 2\epsilon_2 A_2 \Big],
\end{aligned}$$

where $A_1 = m(2\eta)m_{11}(\eta)$ and $A_2 = m(2\eta)m_{11}(\eta) - 2m_1(\eta)m_1(2\eta) + m(\eta)m_{11}(2\eta) - m(2\eta)m_1^2(\eta)/m(\eta)$. Then

$$\begin{aligned}
var(\hat{\eta}) &= \frac{\Big[m(\eta) - 2(\epsilon_1 + \epsilon_2) m(2\eta) + O(\epsilon_2^2) \Big]}{n\epsilon_2 \Big[\{m(\eta)m_{11}(\eta) - m_1^2(\eta)\} - 2\epsilon_1 A_1 - 2\epsilon_2 A_2 + O(\epsilon_2^2) \Big]} \\
&= v_1(\eta) \frac{\Big[1 - 2(\epsilon_1 + \epsilon_2) m(2\eta)/m(\eta) + O(\epsilon_2^2) \Big]}{n\epsilon_2 \Big[1 - 2\epsilon_1 A_1 v_1(\eta)/m(\eta) - 2\epsilon_2 A_2 v_1(\eta)/m(\eta) O(\epsilon_2^2) \Big]}
\end{aligned}$$

where $v_1(\eta) = m(\eta)/\{m(\eta)m_{11}(\eta) - m_1^2(\eta)\}$. By the Binomial theorem

$$\begin{aligned}
& \frac{\Big[1 - 2(\epsilon_1 + \epsilon_2) m(2\eta)/m(\eta) + O(\epsilon_2^2), \Big]}{\Big[1 - 2\epsilon_1 A_1 v_1(\eta)/m(\eta) - 2\epsilon_2 A_2 v_1(\eta)/m(\eta) + O(\epsilon_2^2) \Big]} \\
& = \Big[1 - 2(\epsilon_1 + \epsilon_2) m(2\eta)/m(\eta) + O(\epsilon_2^2) \Big] \\
& \times \Big[1 + 2\epsilon_1 A_1 v_1(\eta)/m(\eta) + 2\epsilon_2 A_2 v_1(\eta)/m(\eta) + O(\epsilon_2^2) \Big] \\
& = 1 - 2(\epsilon_1 + \epsilon_2) m(2\eta)/m(\eta) + 2\epsilon_1 A_1 v_1(\eta)/m(\eta) + 2\epsilon_2 A_2 v_1(\eta)/m(\eta) \\
& + O(\epsilon_2^2).
\end{aligned}$$

Therefore,

$$var(\hat{\eta}) = \frac{v_1(\eta)}{n\epsilon_2} \Big[1 + 2v_1(\eta) \frac{R_1(\eta)}{m(\eta)} - 2\frac{R_2(\eta)}{m(\eta)} \Big] + O(\epsilon_2^2)$$

where

$$\begin{aligned} R_1(\eta) &= \epsilon_1 m(2\eta)m_{11}(\eta) + \epsilon_2\{m(2\eta)m_{11}(\eta) \\ &\quad -2m_1(\eta)m_1(2\eta) + m(\eta)m_{11}(2\eta) - m(2\eta)m_1^2(\eta)/m(\eta)\} \end{aligned}$$

and

$$R_2(\eta) = (\epsilon_1 + \epsilon_2)m(2\eta).$$

When $X \sim N(0,1)$, $m(\eta) = e^{\eta^2/2}$, $m_1(\eta) = \eta e^{\eta^2/2}$, $m_{11}(\eta) = (1+\eta^2)e^{\eta^2/2}$, $m(2\eta) = e^{2\eta^2}$, $m_1(2\eta) = 2\eta e^{2\eta^2}$, $m_{11}(2\eta) = (1+4\eta^2)e^{2\eta^2}$. Thus

$$v_1(\eta) = e^{\eta^2/2}/\{e^{\eta^2/2}(1+\eta^2)e^{\eta^2/2} - \eta^2 e^{\eta^2}\} = e^{-\eta^2/2},$$

$$R_1(\eta) = \epsilon_1(1+\eta^2)e^{5\eta^2/2} + 2\epsilon_2 e^{5\eta^2/2},$$

$$R_2(\eta) = (\epsilon_1 + \epsilon_2)e^{2\eta^2},$$

and

$$var(\hat{\eta}) \approx \frac{e^{-\eta^2/2}}{n\epsilon_2}\left[1 + (2\eta^2\epsilon_1 + 2\epsilon_2)e^{3\eta^2/2}\right].$$

VITA

Hyun Sun Kim was born in Seoul, Korea. She received a Bachelor of Science degree in statistics in 1995 from Dongguk University, Korea. She received a Master of Science degree in statistics in 1997 from Dongguk University, Korea. In 1999, she was admitted to the Ph.D. program in the Department of Statistics, Texas A&M University. She received her Ph.D. degree in August 2004. Her permanent address is:

1807-1 Eunhaeng 2 Dong, Jungwon-Gu
Sungnam, 462-152
Republic of Korea.