

## ABSTRACT

Title of dissertation: ASYMPTOTIC NORMALITY  
IN GENERALIZED  
LINEAR MIXED MODELS

Min Min  
Doctor of Philosophy, 2007

Dissertation directed by: Professor Paul J. Smith  
Statistics Program  
Department of Mathematics

Generalized Linear Mixed Models (GLMMs) extend the framework of Generalized Linear Models (GLMs) by including random effects into the linear predictor. This will achieve two main goals of incorporating correlation and allowing broader inference. This thesis investigates estimation of fixed effects as the number of random effects grows large. This model describes cluster analysis with many clusters and also meta-analysis.

After reviewing currently available methods, especially the penalized likelihood and conditional likelihood estimators of Jiang [14], we focus on the random intercept problem. We propose a new estimator  $\hat{\beta}_w$  of regression coefficient and prove that when  $m$ , the number of random effects, grows to infinity at a slower rate than the smallest cluster sample size,  $\hat{\beta}_w$  is consistent and given the realization of random effects, is asymptotically normal. We also show how to estimate the standard errors of our estimators. We also study the asymptotic distribution of Jiang's [14] penalized likelihood estimators. In the absence of regression coefficients,

the normalized estimated intercept  $\sqrt{m}(\hat{a} - a_0)$  converges to a normal distribution. Difficulties arise in establishing the conditional asymptotic normality of Jiang's [14] penalized likelihood estimator  $\hat{\beta}$  of regression coefficients for fixed effects in a general GLMM.

In Chapter 4, we make an extended analysis of the  $2 \times 2 \times m$  table to show how to verify the general conditions in Chapter 3. We compare our estimator to the Mantel-Haenszel estimator. Simulation studies and real data analysis results validate our theoretical results.

In Chapter 5, asymptotic normality of joint fixed effect estimate and scale parameter estimate is proved for the case as  $m/N \rightarrow 0$ . An example was used to verify the general conditions in this case.

Simulation studies were performed to validate the theoretical results as well as to investigate conjectures that are not covered in the theoretical proofs. The asymptotic theory for  $\hat{\beta}_w$  describes the finite sample behavior of  $\hat{\beta}_w$  very accurately. We find that in the case as  $m/N \rightarrow 0$ , in the random logistic and Poisson intercept models, consistency and conditional asymptotic normality results appear to hold for the penalized regression coefficient estimates  $\hat{\beta}$ .

ASYMPTOTIC NORMALITY IN GENERALIZED  
LINEAR MIXED MODELS

by

Min Min

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2007

Advisory Committee:  
Dr. Paul Smith, Chair/Advisor  
Dr. C. Mitchell Dayton  
Dr. Abram Kagan  
Dr. Partha Lahiri  
Dr. Eric Slud

© Copyright by  
Min Min  
2007

## DEDICATION

In the memory of my grandparents,  
Houzhang and Yunxuan

## ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Dr. Paul Smith for his teaching, guiding and encouraging me through all these years. He has always made himself available for help and provided valuable insights. Without his academic advising and moral support, it is impossible to complete my dissertation. I am honored to work with such an extraordinary scholar.

I would like to thank the committee members: Drs. Abram Kagan and Eric Slud for reading my manuscript and making valuable comments and corrections which enhanced the quality of this work. Drs. C. Mitchell Dayton and Partha Lahiri for agreeing to serve on my dissertation committee.

I would also like to thank Chip Denman for his support when I was working at stat lab.

I thank all of my friends at UMCP, especially Guanhua Lu, Huaizhong Ren, Hsiao-Hui Tsou, Shihua Wen, Chin-Fang Weng and Zhihui Yang for their support and loyal friendship.

I owe my deepest appreciation to my family, especially my parents for their unconditional love and always believing in me.

Finally, I want to thank my husband for his love, understanding and sharing the happiness of my achievement. I also want to thank my best friend Yuehan Wang for always being there for me. A special thank you goes to my son Chance for his patience, for bringing me great joy and enriching my life.

## Table of Contents

List of Tables	v
List of Figures	viii
1 Introduction	1
1.1 Notations . . . . .	4
2 Literature Review	7
2.1 Generalized Linear Models . . . . .	7
2.1.1 Definitions . . . . .	7
2.1.2 Asymptotics . . . . .	10
2.2 Quasilikelihood . . . . .	13
2.3 Generalized Estimating Equations . . . . .	14
2.3.1 Definition . . . . .	15
2.3.2 Asymptotics . . . . .	17
2.4 Generalized Linear Mixed Models . . . . .	22
2.4.1 Definitions . . . . .	22
2.4.2 Estimation approaches . . . . .	23
2.4.3 Random intercept model (canonical link) . . . . .	26
2.5 Generalized GLMM (canonical link function case) . . . . .	27
2.5.1 Definitions . . . . .	28
2.5.2 Penalized Generalized Weighted Least Squares in Case 1 . . . . .	28
2.5.3 Maximum Conditional Likelihood Estimates in Case 2 . . . . .	33
3 Case 1 random intercept model results.	36
3.1 Case 1 random intercept (combine $a$ and $v_i$ ) . . . . .	37
3.2 Penalized likelihood estimator $\hat{a}$ in the case 1 random intercept model when $\beta=0$ . . . . .	42
3.3 Discussion of Jiang [14] penalized likelihood estimator $\hat{\beta}$ in case 1 random intercept model ( $\beta \neq 0$ ) . . . . .	44
3.4 Proofs. . . . .	48
3.4.1 Proof of Lemma 3.1.1 . . . . .	48
3.4.2 Proof of Theorem 3.1.2 . . . . .	51
3.4.3 Proof of Theorem 3.1.3 . . . . .	55
3.4.4 Proof of Corollary 3.1.4 . . . . .	59
3.4.5 Proof of Theorem 3.2.1 . . . . .	60
4 Logistic $2 \times 2 \times m$ table	64
4.1 Logistic $2 \times 2 \times m$ table . . . . .	64
4.2 Simulation results for $2 \times 2 \times m$ table. . . . .	73
4.2.1 Unconditional convergence in distribution . . . . .	74
4.2.2 Conditional Convergence in Distribution . . . . .	84
4.3 Analysis of real data for a $2 \times 2 \times 22$ table . . . . .	89

5	Case 2 Results.	91
5.1	Case 2 consistency results of Jiang [14]. . . . .	91
5.2	The Simple case ( $\alpha = 0$ ). . . . .	95
5.3	The logistic model $\text{logit}P(y_{ijk} = 1 b_{ij}) = \mu + b_{ij}$ . . . . .	99
6	Simulation	106
6.1	Case 1 simulations for $\hat{\beta}_w$ . . . . .	107
6.1.1	Case 1 combining $a$ with $v_i$ . . . . .	107
6.2	Case 1 logistic random intercept simulation for fixed realizations of random effects . . . . .	117
6.3	Case 1 random intercept simulations for penalized likelihood estimators	121
6.3.1	Case 1 logistic random intercept combining $a$ with $v_i$ . . . . .	122
6.4	Case 2 logistic simple example . . . . .	126
7	Conclusions and Future Work.	130
7.1	Theoretical Conclusions . . . . .	130
7.2	Conclusions from the simulation studies and real data analysis . . . . .	131
7.3	Future work. . . . .	132
A	Simulation Results	133
A.1	Case 1 random intercept model for penalized likelihood estimator . .	133
A.1.1	Case 1 logistic random intercept combining $a$ with $v_i$ . . . . .	133
A.1.2	Case 1 logistic random intercept when $a$ and $v_i$ are estimated separately. . . . .	146
A.1.3	Case 1 Poisson random intercept combining $a$ and $v_i$ . . . . .	151
A.1.4	Case 1 Poisson random intercept with $a$ and $v_i$ estimated separately . . . . .	158
	Bibliography	159

## List of Tables

1.1 Notations used throughout the Thesis . . . . .	5
1.2 Table 1.1 (Continued): Notation used throughout the Thesis . . . . .	6
4.1 $2 \times 2 \times m$ table . . . . .	64
4.2 Simulated estimates of logistic odds ratio in $2 \times 2 \times m$ balanced table. . . . .	76
4.3 Simulated standardized estimates of log odds ratio (standardized by true and estimated conditional standardized error) in $2 \times 2 \times m$ balanced table . . . . .	77
4.4 Simulated estimates of log odds ratio (standardized by true and estimated conditional standardized error) in $2 \times 2 \times m$ unbalanced table. . . . .	81
4.5 Simulated standardized estimates of log odds ratio (standardized by true and estimated conditional standardized error) in $2 \times 2 \times m$ unbalanced table . . . . .	82
4.6 Simulated estimates of logistic odds ratio for fixed realizations of uniformly distributed random effects ( $m = 10, n_1 = 200$ ) and $\beta_0 = 0.5$ . . . . .	85
4.7 Simulated standardized estimate of logistic odds ratio with fixed realizations of uniformly distributed random effects ( $m = 10, n_1 = 200$ ) and $\beta_0 = 0.5$ . . . . .	86
4.8 Simulated estimates of logistic odds ratio with fixed realizations of uniformly distributed random effects ( $m = 10, n_1 = 200$ ) and $\beta_0 = 1$ . . . . .	87
4.9 Simulated standardized estimate of logistic odds ratio with fixed realizations of uniformly distributed random effects ( $m = 10, n_1 = 200$ ) and $\beta_0 = 1$ . . . . .	88
4.10 Summary of simulation of our estimator and Mantel-Haenszel estimator of log odds ratio. . . . .	90
6.1 Simulated estimates of logistic and Poisson regression coefficient combining fixed intercept with uniformly distributed random effects. . . . .	110
6.2 Simulated standardized estimate of logistic and Poisson regression coefficient with uniformly distributed random effects. . . . .	111

6.3	Simulated estimates of logistic and Poisson regression coefficient combining fixed intercept with normally distributed random effects. . . . .	114
6.4	Simulated standardized estimate of logistic and Poisson regression coefficient with normally distributed random effects. . . . .	115
6.5	Simulated estimates of logistic regression coefficient combining fixed effect and fixed realizations of uniformly distributed random effects ( $m = 20, n_1 = 200$ ). . . . .	118
6.6	Simulated standardized estimate of logistic regression coefficient with fixed realizations of uniformly distributed random effects ( $m = 20, n_1 = 200$ ). . . . .	119
6.7	Simulated estimates of logistic regression coefficient combining fixed effect and fixed realizations of uniformly distributed random effects ( $m = 40, n_1 = 240$ ). . . . .	120
6.8	Simulated standardized estimate of logistic regression coefficient with fixed realizations of uniformly distributed random effects ( $m = 40, n_1 = 240$ ). . . . .	121
6.9	Simulated standardized estimates of logistic regression coefficient (standardized by true and estimated conditional standardized error) combining fixed intercept with random effects. . . . .	124
6.10	Simulated standardized estimate of logistic regression coefficient for various initial values. . . . .	124
6.11	Simulated standardized estimate of logistic regression coefficient for various $\lambda_1$ values. . . . .	125
6.12	Simulated standardized estimate of $\mu$ and $\tau$ in case 2 simple example for various $(m, n_1)$ values . . . . .	128
A.1	Simulated standardized estimates of logistic regression coefficient (standardized by true and estimated conditional standardized error) combining fixed intercept with random effects. . . . .	136
A.2	Simulated standardized estimate of logistic regression coefficient for various initial values. . . . .	137
A.3	Simulated standardized estimate of logistic regression coefficient for various $\lambda_1$ values. . . . .	139

A.4 Simulated standardized estimates of logistic regression coefficient (standardized by true and estimated conditional standardized error) with $a$ and $v_i$ estimated separately. . . . .	147
A.5 Simulated standardized estimate of Poisson regression coefficient for various $(m, n_1)$ values. . . . .	152
A.6 Simulated standardized estimate of logistic regression coefficient for various $\lambda_1$ values. . . . .	153

## List of Figures

4.1 Q-Q plots for $(\hat{\beta}_w - \beta_0)$ and $(\hat{\beta}_w - \beta_0)$ standardized by true and estimated conditional standard error, for various values of $(m, n)$ in logistic $2 \times 2 \times m$ balanced setup. . . . .	78
4.2 Q-Q plots for $(\hat{\beta}_w - \beta_0)$ and $(\hat{\beta}_w - \beta_0)$ standardized by true and estimated conditional standard error, for various values of $(m, n)$ in logistic $2 \times 2 \times m$ unbalanced setup where $n1=600, 400$ for $(1/3, 2/3)$ , $(1/4, 3/4)$ setups respectively . . . . .	83
6.1 Q-Q plots for $(\hat{\beta}_w - \beta_0)$ and $(\hat{\beta}_w - \beta_0)$ standardized by true and estimated standard error, for various values of $(m, n)$ in case 1 random intercept model with uniformly distributed random effects. . . . .	112
6.2 Q-Q plots for $(\hat{\beta}_w - \beta_0)$ and $(\hat{\beta}_w - \beta_0)$ standardized by true and estimated standard error, for various values of $(m, n)$ in case 1 random intercept model with normally distributed random effects. . . . .	116
6.3 Q-Q plots for standardized $(\hat{\mu} - \mu_0)$ and $(\hat{\tau} - \tau_0)$ by true standard error, for various $(m, n_1)$ in case 2 logistic simple example. . . . .	129
A.1 Q-Q plots for $(\hat{\beta}_1 - \beta_{10})$ standardized by true conditional standard error combining $a$ with $v_i$ , for various $(m, n_1)$ in logistic random intercept case 1. . . . .	140
A.2 Q-Q plots for $(\hat{\beta}_1 - \beta_{10})$ standardized by estimated conditional standard error combining $a$ with $v_i$ , for various $(m, n_1)$ in logistic random intercept case 1. . . . .	141
A.3 Q-Q plots for $(\hat{\beta}_1 - \beta_{10})$ standardized by true conditional standard error combining $a$ with $v_i$ , for various initial values of $\hat{\beta}_1$ in logistic random intercept case 1. . . . .	142
A.4 Q-Q plots for $(\hat{\beta}_1 - \beta_{10})$ standardized by estimated standard error combining $a$ with $v_i$ , for various initial values of $\hat{\beta}_1$ in logistic random intercept case 1. . . . .	143
A.5 Q-Q plots for $(\hat{\beta}_1 - \beta_{10})$ standardized by true conditional standard error combining $a$ with $v_i$ , for various values of $\lambda_1$ in logistic random intercept case 1. . . . .	144

A.6 Q-Q plots for $(\hat{\beta}_1 - \beta_{10})$ standardized by estimated conditional standard error combining $a$ with $v_i$ , for various values of $\lambda_1$ in logistic random intercept case 1. . . . .	145
A.7 Q-Q plot for $(\hat{\beta}_1 - \beta_{10})$ standardized by true conditional standard error not combining $a$ with $v_i$ , for various $(m, n_1)$ in logistic random intercept case 1. . . . .	149
A.8 Q-Q plot for $(\hat{\beta}_1 - \beta_{10})$ standardized by estimated conditional standard error with estimated $a$ with $v_i$ separately, for various $(m, n_1)$ in logistic random intercept case 1. . . . .	150
A.9 Q-Q plots for $(\hat{\beta}_1 - \beta_{10})$ standardized by true conditional standard error with $a$ and $v_i$ estimated separately in Poisson random intercept case 1. . . . .	154
A.10 Q-Q plots for $(\hat{\beta}_1 - \beta_{10})$ standardized by estimated conditional standard error with $a$ and $v_i$ estimated separately in Poisson random intercept case 1. . . . .	155
A.11 Q-Q plots for $(\hat{\beta}_1 - \beta_{10})$ for standardized by true conditional standard error with $a$ and $v_i$ estimated separately, for various values of $\lambda_1$ in Poisson random intercept case 1. . . . .	156
A.12 Q-Q plots for $(\hat{\beta}_1 - \beta_{10})$ standardized by estimated conditional standard error with $a$ and $v_i$ estimated separately, for various values of $\lambda_1$ in Poisson random intercept case 1. . . . .	157

# Chapter 1

## Introduction

This thesis is focused on asymptotic normality of fixed effect estimates in Generalized Linear Mixed Models (GLMMs) for the canonical link function case when the number of random effects is large. Due to the wide range of applications of GLMMs, these models have received substantial attention during the last decade. Although asymptotic behavior of fixed effect estimates has been well studied in linear regression models, Generalized Linear Models (GLMs) and Generalized Estimating Equations (GEEs), for GLMMs there is still a lot of work to do. These models and the main idea of this thesis are addressed briefly below.

Generalized Linear Models (GLMs), originally introduced by Nelder and Wedderburn [23], provide a unified family of models that is widely used for regression analysis. These models are intended to describe non-normal responses. In particular, they avoid having to select a single transformation of the data to achieve the possibly conflicting objectives of normality, linearity and homogeneity of variance. Important examples include binary and count data. They can be applied to a wide array of discrete, continuous and censored outcomes. They are most commonly used when the outcomes are independent. These models are described in detail in Section 2.1. Another important extension of GLM is the Quasilikelihood model approach which can be defined by specifying only the relation between the mean and the

linear predictors and between the mean and variance of observations. This model is illustrated in Section 2.2.

However, in many applications, independence of outcomes is not a reasonable assumption. This is particularly obvious in longitudinal studies, where multiple measurements made on the same individual are likely to be correlated. One technique for the analysis of such general correlated data is the generalized estimating equations (GEE) approach introduced by Liang and Zeger [37]. This approach has the desired quality of independence between subjects while adjusting for the correlation structure within subjects. These models are described in detail in Section 2.3.

We may wish to build a model that accommodates correlated data, or to consider the levels of a factor as selected from a population of levels in order to make inference about that population. Generalized Linear Mixed Models (GLMMs) extend the framework of Linear Mixed Models (LMMs) and of Generalized Linear Models (GLMs) by allowing for non-Gaussian data, nonlinear link functions and inclusion of random effects and of correlated error. These models are described in detail in Section 2.4.

Asymptotic normality and consistency results for fixed effect estimates have been proved for GLMs ( e.g., Habermann [13] and Fahrmeir and Kaufmann [11]) and also for GEEs ( e.g., Zeger and Liang [37] and Xie and Yang [35]). For GLMMs, most work about asymptotic properties of fixed effect estimates is based on eliminating random effects either by integrating them out (e.g., Sinha [30]) or by conditioning on minimal sufficient statistics of random effects (e.g, Sartori and Severini [28] ). Simply

eliminating random effects may discard information. In addition, integrating out all the random effects requires knowledge about the distribution of all the random effects.

Jiang [14] extended GLMMs to generalized GLMMs by making only an expectation assumption for random effects instead of a normal distribution assumption. He divided generalized GLMMs into two cases:  $m/N \rightarrow 0$  (case 1) and  $m/N \not\rightarrow 0$  (case 2) where  $m$  is the number of levels of random effects and  $N$  is the sample size. Both  $m$  and  $N$  go to infinity. For case 1 and 2, he proposed two methods: Penalized Generalized Weighted Least Squares (PGWLS) and Maximum Conditional Likelihood (MCL). Under reasonable conditions, consistency of both fixed and random effect estimates was proved rigorously. Jiang [14] used three examples to illustrate those reasonable conditions. The concepts and results mentioned above are discussed in Section 2.5.

In order to find approximate tests and confidence regions, an asymptotic normality result is needed. In Chapter 3, focusing on the random intercept problem, we propose a new estimator  $\hat{\beta}_w$  of regression coefficients and prove that when  $m$ , the number of random effects, grows to infinity at a slower rate than the smallest cluster sample size,  $\hat{\beta}_w$  is consistent and given the realization of random effects, is asymptotically normal. We also show how to estimate the normalizer and weight matrices of our estimators. We also study the asymptotic distribution of Jiang's [14] penalized likelihood estimators. In the absence of regression coefficients, the normalized estimated intercept  $\sqrt{m}(\hat{a} - a_0)$  converges to a normal distribution. Difficulties arise in establishing the conditional asymptotic normality of the penalized likelihood

estimator  $\hat{\beta}$  of regression coefficients for fixed effects in a general GLMM.

In Chapter 4, we make an extended analysis of the  $2 \times 2 \times m$  table to show how to verify the general conditions in Chapter 3. We compare our estimator to the Mantel-Haenszel estimator. Simulation studies and real data analysis results validate our theoretical results. In Chapter 5, asymptotic normality of joint fixed effect estimate and scale parameter estimate is proved for the case as  $m/N \rightarrow 0$ . An example was used to verify the general conditions in this case. In Chapter 6, in order to check the asymptotic results of the theorems in Chapter 3 and 5, logistic and Poisson random intercept models are simulated in case 1. For case 2, a simple model is simulated. The Splus built-in function `nlsinb` is used to compute the estimates. The contents of Chapters 3-6 are new.

In Chapter 7, we summarize our theoretical and simulation results, and discuss the potential direction of future work.

## 1.1 Notations.

Table 1.1: Notations used throughout the Thesis

Symbol	Meaning
$B_{ij}$	$b''_{ij}(\eta_{ij0}) + \frac{1}{2}b_{ij}^{(3)}(\eta_{ij}^*)(\hat{\eta}_{ij} - \eta_{ij0})$
$r_{ij}$	$\frac{1}{2}b_{ij}^{(3)}(\eta_{ij}^*)(\hat{\eta}_{ij} - \eta_{ij0})$
$\tilde{\mathbf{B}}$	$\text{diag}(B_{ij})_{1 \leq i \leq m, 1 \leq j \leq n_i}$
$E_i^*$	$\sum_j B_{ij}$
$n_*$	$\min_{1 \leq i \leq m} n_i$
$C_{\lambda_1}^*$	$-\lambda_1 / (\lambda_1 \sum_i (E_i^*)^{-1} + m)$
$\hat{\eta}_{ij}$	$\hat{a}_i + \mathbf{x}_{ij}^t \hat{\boldsymbol{\beta}}_i$
$\hat{\hat{\eta}}_{ij}$	$\hat{a}_i + \mathbf{x}_{ij}^t \hat{\hat{\boldsymbol{\beta}}}_w$

Table 1.2: Table 1.1 (Continued): Notation used throughout the Thesis

Symbol	Meaning
$\mathbf{D}_N(\gamma)$	$-\partial^2 l_P(\gamma)/\partial^2 \gamma$
$\mathbf{g}_N(\gamma)$	$\partial l_P(\gamma)/\partial \gamma$
$\mathbf{1}_{n_i}$	the $n_i \times 1$ vector whose elements are all 1's
$\ \mathbf{u}\  = (\sum_{i=1}^n u_i^2)^{1/2}$	vector norm for vector $\mathbf{u} \in \mathbf{R}^n$
$\ \mathbf{A}\  = \sup_{\ \mathbf{u}\ =1}  \mathbf{u}^t \mathbf{A} \mathbf{u}  = \max_j( \lambda_j )$	matrix norm for a symmetric matrix $\mathbf{A}$
$\lambda_{\max}(\mathbf{A}) = \sup_{\ \mathbf{u}\ =1} \mathbf{u}^t \mathbf{A} \mathbf{u}$	largest eigenvalue of $\mathbf{A}$
$\lambda_{\min}(\mathbf{A}) = \inf_{\ \mathbf{u}=1\ } \mathbf{u}^t \mathbf{A} \mathbf{u}$	smallest eigenvalue of $\mathbf{A}$
$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$	Partitioned matrix $\mathbf{A}$
$\mathbf{A}_{11.2} = \mathbf{A}_{11} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21}$	Schur component of $\mathbf{A}_{11}$

For notational simplicity, we let  $b''_{ij} = b''(\eta_{ij0})$ ,  $b'_{ij} = b'(\eta_{ij0})$ ,  $\mathbf{D}_N = \mathbf{D}_N(\gamma_0)$  and  $\mathbf{g}_N = \mathbf{g}_N(\gamma_0)$ , ect. where  $\eta_{ij0}$  and  $\gamma_0$  are the true values of  $\eta_{ij}$  and  $\gamma$  respectively.

$\mathbf{A}^{1/2}$  or  $\mathbf{A}^{T/2}$  represents a left or right square root.  $\mathbf{A}^-$  represents generalized inverse.

## Chapter 2

### Literature Review

#### 2.1 Generalized Linear Models

Nelder and Wedderburn [23] introduced Generalized Linear Models (GLMs) as a unifying class of models which are widely used in regression analysis. Section 2.1.1 presents their original definition and the extended definition from Fahrmeir and Kaufmann [11]. Asymptotic properties of estimates are described in Section 2.1.2.

##### 2.1.1 Definitions

GLMs were originally described as follows: A vector of observations  $\mathbf{y}$  having  $n$  components is assumed to be a realization of a random variable  $\mathbf{Y}$  whose components are independently distributed with means  $\mu_i$ ,  $i = 1, \dots, n$ . Assume that the components  $Y_i$  are independent and have a distribution in the exponential family, taking the form

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \quad (2.1)$$

for some specific functions  $a(\cdot)$ ,  $b(\cdot)$  and  $c(\cdot)$ . Then

$$E(Y_i) = \mu_i = b'(\theta_i),$$

$$\text{Var}(Y_i) = b''(\theta_i)a(\phi).$$

Define  $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i$ , where  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is a vector of unknown parameters,  $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$  is a vector of covariates and  $g$  is a known link function (since it links together the mean of  $y_i$  and the linear form of predictors).

The large sample theory of GLM's is derived by Fahrmeir and Kaufmann [11]. They consider GLM's with the following structure and notations.

(i) The  $\{\mathbf{y}_i\}$  are independent  $q$ -dimensional random variables with densities

$$f(\mathbf{y}_i | \boldsymbol{\theta}_i) = c(\mathbf{y}_i) \exp(\boldsymbol{\theta}_i^T \mathbf{y}_i - \mathbf{b}(\boldsymbol{\theta}_i)) \quad (2.2)$$

of the natural exponential type,  $\boldsymbol{\theta}_i \in \Theta^0$  (assume  $\Theta \in \Re^2$  to be the natural parameter space, that is, the set of all  $\boldsymbol{\theta}$  satisfying  $0 < \int c(\mathbf{y}) \exp(\boldsymbol{\theta}^T \mathbf{y}) d\mathbf{y} < \infty$ ). Then  $\Theta$  is convex, and in the interior  $\Theta^0$  of  $\Theta$ , all derivatives of  $b(\boldsymbol{\theta})$  and all moments of  $\mathbf{y}$  exist (assume  $\Theta^0 \neq \emptyset$ ). In particular we have  $E(\mathbf{y}) = \partial b(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \equiv \boldsymbol{\mu}(\boldsymbol{\theta}) \in \Re^2$  and  $\text{Cov}(\mathbf{y}) = \partial^2 b(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T \equiv \Sigma(\boldsymbol{\theta})$ , a  $q \times q$  matrix.

(ii) The deterministic matrix  $\mathbf{x}_i$  influences  $\mathbf{y}_i$  in the form of a linear combination

$$\boldsymbol{\eta}_i = \mathbf{x}_i^T \boldsymbol{\beta}, \text{ where } \boldsymbol{\beta} \text{ is a } p\text{-dimensional parameter.}$$

(iii) The linear combination  $\boldsymbol{\eta}_i$  is related to the mean  $\boldsymbol{\mu}(\boldsymbol{\theta}_i)$  of  $\mathbf{y}_i$  by the injective link function  $\mathbf{g} : M \rightarrow \Re^q$  where  $M \in \Re^q$  is the range of the function  $\boldsymbol{\mu}(\boldsymbol{\theta})$ ,  $\boldsymbol{\eta}_i = \mathbf{g}(\boldsymbol{\mu}(\boldsymbol{\theta}_i))$ . One can write  $\boldsymbol{\theta}_i = \mathbf{u}(\mathbf{x}_i^T \boldsymbol{\beta})$ , where  $\mathbf{u}$  is the injective function  $(\mathbf{g} \circ \boldsymbol{\mu})^{-1}$ , mapping  $\Re$  into  $\Theta$ .

Here (2.2) is different from (2.1) because the response variable may be a vector rather than a scalar and (2.1) has an additional nuisance parameter  $\phi$ . The maximum likelihood estimating equations are unchanged by the presence of  $a(\phi)$ ,

but the information matrix has to be multiplied by an unknown scale factor  $a(\phi)$ , which can be estimated consistently. Thus, without loss of generality, Fahrmeir and Kaufmann [11] define GLM's using the simpler form (2.2). By allowing  $\mathbf{y}$  to be a vector, Fahrmeir and Kaufmann [11] can accommodate multivariate responses, such as multinomial data.

Then the loglikelihood of a sample  $y_1, \dots, y_n$  is given by

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^n (\boldsymbol{\theta}_i^T \mathbf{y}_i - b(\boldsymbol{\theta}_i)) - C_l \quad (2.3)$$

where  $\boldsymbol{\theta}_i = \mathbf{u}(\mathbf{x}_i^T \boldsymbol{\beta})$ ,  $i = 1, \dots, n$  and  $C_l$  does not depend on  $\boldsymbol{\beta}$ .

Setting  $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = \mu(\mathbf{u}(\mathbf{x}_i^T \boldsymbol{\beta}))$ ,  $\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) = \boldsymbol{\Sigma}(\mathbf{u}(\mathbf{x}_i^T \boldsymbol{\beta}))$ ,  $\mathbf{U}_i(\boldsymbol{\beta}) = [\partial \mathbf{u}(\mathbf{x}_i^T \boldsymbol{\beta}) / \partial \boldsymbol{\eta}]^T$  and differentiating  $l_n(\boldsymbol{\beta})$ , we find the score function  $\mathbf{s}_n(\boldsymbol{\beta})$  and the information matrix  $\mathbf{F}_n(\boldsymbol{\beta})$  to be

$$\mathbf{s}_n(\boldsymbol{\beta}) = \partial l_n(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \sum_{i=1}^n \mathbf{x}_i \mathbf{U}_i(\boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) \quad (2.4)$$

$$\mathbf{F}_n(\boldsymbol{\beta}) = \text{Cov}[\mathbf{s}_n(\boldsymbol{\beta})] = \sum_{i=1}^n \mathbf{x}_i \mathbf{U}_i(\boldsymbol{\beta}) \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \mathbf{U}_i^T(\boldsymbol{\beta}) \mathbf{x}_i^T \quad (2.5)$$

Further differentiation yields

$$\mathbf{H}_n(\boldsymbol{\beta}) = -\partial^2 l_n(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T. \quad (2.6)$$

It is easy to see that  $E(\mathbf{s}_n(\boldsymbol{\beta})) = 0$  and  $E(\mathbf{H}_n(\boldsymbol{\beta})) = \mathbf{F}_n(\boldsymbol{\beta})$ .

For canonical link functions, these expressions simplify considerably:

$$\mathbf{s}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{x}_i (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})), \quad \mathbf{F}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{x}_i \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \mathbf{x}_i^T, \quad \mathbf{H}_n(\boldsymbol{\beta}) = \mathbf{F}_n(\boldsymbol{\beta}).$$

The true parameter is denoted by  $\boldsymbol{\beta}_0$ , and  $\mathbf{F}_n(\boldsymbol{\beta}_0)$  is written as  $\mathbf{F}_n$  for simplicity.

Similarly we write  $\mathbf{s}_n$  and  $\mathbf{H}_n$ , etc.

### 2.1.2 Asymptotics

Based on maximum likelihood, under reasonable conditions, asymptotic existence, consistency and asymptotic normality of fixed effect estimates are proved by Haberman [13] and Fahrmeir and Kaufmann [11]. The details of their work are presented in the following.

Haberman [13] provides conditions for the asymptotic existence of the MLE by use of the general theory for exponential models derived by Berk [2] and Barndorff-Nielsen [4]. Here asymptotic properties of MLE are considered for canonical link functions when the dimension of  $\beta$  grows to infinity. The main idea is that the parameter space grows with the sample size, but one only wants to estimate a linear functional of the natural parameter which is determined by a finite dimensional subspace of the parameter space. Suppose  $\kappa = \kappa(\theta)$  is the linear functional and  $\hat{\kappa}_n$  is the unique MLE of  $\kappa(\theta)$ , and  $\sigma_n(\kappa)$  is the asymptotic standard deviation of  $\hat{\kappa}_n$ . Under some technical conditions, consistency results like  $\sigma_n(\kappa) \rightarrow_p 0$  and  $\hat{\kappa}_n \rightarrow_p \kappa(\theta)$  as  $n \rightarrow \infty$  are established. Based on Newton's method,  $\hat{\kappa}_n$  is proved to be asymptotically normal with asymptotic mean  $\kappa(\theta)$ , meaning that  $(\hat{\kappa}_n - \kappa(\theta))/\sigma_n(\kappa) \rightarrow_d N(0, 1)$ . Asymptotic confidence intervals are also considered since the MLE  $\hat{\sigma}_n(\kappa)$  of  $\sigma_n(\kappa)$  is a consistent estimate of  $\sigma_n(\kappa)$  (e.g.,  $\hat{\sigma}_n(\kappa)/\sigma_n(\kappa) \rightarrow_p 1$ ). There are about six different inner products appearing back and forth in Haberman's [13] paper which makes it intractable and not easily understood. Fahrmeir and Kaufmann [11] present mild general conditions which, respectively, assure weak or strong convergence or asymptotic normality of the MLE for both canonical and noncanonical link function

cases, where the dimension of  $\beta$  is fixed.

In the canonical link function case, the normality condition, though obtained by a different approach, is closely related to one of Haberman's [13] conditions. The assumptions of Fahrmeir and Kaufmann [11] are simpler to interpret and check and they seem slightly weaker. In addition to regularity conditions on the parameter space, link function and covariates, in order to derive conditions for consistency and asymptotic normality of MLE, they define a sequence  $N_n(\delta)$ ,  $\delta > 0$ , of neighborhoods of  $\beta_0$  (the true value of  $\beta$ ) as

$$N_n(\delta) = \{\beta : \|\mathbf{F}_n^{T/2}(\beta - \beta_0)\| \leq \delta\}$$

where  $n = 1, 2, \dots$  and  $\mathbf{A}^{1/2}$  is a left square root of the positive definite matrix  $\mathbf{A}$  and  $\mathbf{A}^{T/2}$  denotes  $(\mathbf{A}^{1/2})^T$  so that  $\mathbf{A}^{1/2}\mathbf{A}^{T/2} = \mathbf{A}$ . Their conditions are:

(D) Divergence:  $\lambda_{\min}(\mathbf{F}_n) \rightarrow \infty$ .

(C) Boundedness from below: for all  $\delta > 0$ ,  $\mathbf{F}_n(\beta) - c\mathbf{F}_n$  is positive semidefinite,  $\beta \in N_n(\delta)$ ,  $n \geq n_1$  with some constants  $n_1 = n_1(\delta)$ ,  $c > 0$  independent of  $\delta$ .

(N) Convergence and continuity: for all  $\delta > 0$ ,  $\max_{\beta \in N_n(\delta)} \|\mathbf{V}_n(\beta) - \mathbf{I}\| \rightarrow 0$ , where  $\mathbf{V}_n(\beta) = \mathbf{F}_n^{-1/2}\mathbf{F}_n(\beta)\mathbf{F}_n^{-T/2}$  is the normed information matrix.

( $S_\delta$ ) Boundedness of the eigenvalue ratio: there is a neighborhood  $N \subset B$  of  $\beta_0$  such that

$$\lambda_{\min}(\mathbf{F}(\beta)) \geq c(\lambda_{\max}(\mathbf{F}_n))^{1/2+\delta}, \quad \beta \in N, \quad n \geq n_1.$$

where  $B$ , the set of admissible values of  $\beta$ , is open in  $\Re^p$  and additionally, convex for canonical link functions and  $c > 0$ ,  $\delta > 0$ ,  $n_1$  are some constants.

Fahrmeir and Kaufmann's [11] main theorems for the canonical link function case are the following:

**Theorem 2.1.1.** *Under (D) and (C), there is a sequence  $\{\hat{\beta}_n\}$  of random variables with*

- (i)  $P[\mathbf{s}_n(\hat{\beta}_n) = 0] \rightarrow 1$  (asymptotic existence),
- (ii)  $\hat{\beta}_n \rightarrow_p \beta_0$  (weak consistency).

**Theorem 2.1.2.** *Under (D) and  $(S_\delta)$  with a  $\delta > 0$ , there is a sequence  $\{\hat{\beta}_n\}$  of random variables and a random number  $n_2$  with*

- (i)  $P(\mathbf{s}_n(\hat{\beta}_n) = \mathbf{0} \text{ for all } n \geq n_2) = 1$ ,
- (ii)  $\hat{\beta}_n \rightarrow_{a.s.} \beta_0$  (strong consistency).

**Lemma 2.1.3.** *Under (D) and (N), the normed score function is asymptotically normal:*

$$\mathbf{F}_n^{-1/2} \mathbf{s}_n \rightarrow_d N(\mathbf{0}, \mathbf{I}).$$

**Theorem 2.1.4.** *Under (D) and (N), the normed MLE is asymptotically normal:*

$$\mathbf{F}_n^{T/2} (\hat{\beta}_n - \beta_0) \rightarrow_d N(\mathbf{0}, \mathbf{I}).$$

Remark: Because  $B$  is convex and  $\mathbf{F}_n(\beta)$  is positive definite, there is at most one zero of the score function. Lemma 2.1.3 and Theorem 2.1.4 hold for any version of the matrix square root.

Remark: In practice the normalizing matrix  $\mathbf{F}_n^{T/2}$  must be replaced by  $\mathbf{F}_n^{T/2}(\hat{\beta}_n)$ . Fahrmeir and Kaufmann [11] state that  $\mathbf{F}_n^{T/2}(\hat{\beta}_n)(\hat{\beta}_n - \beta_0) \rightarrow_d N(\mathbf{0}, \mathbf{I})$  if the following condition holds

(Q) For all  $\delta > 0$ ,  $\sup_{\beta \in N_n(\delta)} \|\mathbf{F}_n^{1/2} \mathbf{F}_n^{1/2}(\beta) - \mathbf{I}\| \rightarrow 0$

They state that  $(Q) \implies (N)$ . If  $\mathbf{F}_n^{1/2}$  is the Choleski square root, then  $(N) \implies (Q)$ .

Likewise,  $(N) \implies (Q)$  if  $\lambda_{\max}(\mathbf{F}_n)/\lambda_{\min}(\mathbf{F}_n) \leq c < \infty$ ,  $n \geq n_0$  for some constants  $c$  and  $n_0$ .

They also verify these general conditions in several examples, including Poisson model, response with a bounded range, regressor with compact range and stochastic regressor.

Noncanonical link functions enlarge the class of GLMs but they cause additional difficulties in establishing consistency and asymptotic normality, mainly because of the existence of the MLE can not be guaranteed except in special cases (for a number of important examples see Wedderburn [33]). A local maximum of the likelihood in a neighborhood of the true  $\beta_0$  does not necessarily define a global maximum. Here consistency and asymptotic normality results only apply to a sequence  $\{\hat{\beta}_n\}$  of solutions of maximum likelihood equations  $\mathbf{s}_n(\beta) = \mathbf{0}$ . The same line of arguments in the canonical link function case is followed in the proof, but only for conditions (C), (D), (N),  $(S_\delta)$  with  $\mathbf{F}_n(\beta)$  replaced by  $\mathbf{H}_n(\beta)$ .

## 2.2 Quasilikelihood

Another important extension of GLMs is the Quasilikelihood model approach which was introduced by Wedderburn [34]. To define a likelihood one has to specify the form of distribution of the observations. It may be difficult to decide what distribution the observations follow, but to define a quasilikelihood function one

needs only to specify a model for the mean and a relation between the mean and variance of the observations.

This is what makes quasi-likelihood useful. By using the chain rule we can rewrite Equation (2.4) in the following form

$$\mathbf{s}_n(\boldsymbol{\beta}) = \partial l_n(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \sum_{i=1}^n \mathbf{x}_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}} \Big|_{\boldsymbol{\eta}=\mathbf{x}_i^T \boldsymbol{\beta}} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})). \quad (2.7)$$

We can see that the score function in (2.7) depends on the parameters only through the mean  $\boldsymbol{\mu}_i(\boldsymbol{\beta})$  and  $\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) = \boldsymbol{\Sigma}_i(\boldsymbol{\mu}_i(\boldsymbol{\beta}))$ , the variance function.

Fahrmeir [12] extended the analysis of correctly specified GLM's to misspecified GLM's, where misspecification occurs if the true density of response variable is not of the assumed exponential type, and/or the true expectation  $\mathbf{y}_i$  can not be represented by  $\boldsymbol{\mu}(\mathbf{u}(\mathbf{x}_i^T \boldsymbol{\beta}))$ .

He followed the same line argument as Fahrmeir and Kaufmann [11] and proved consistency and asymptotic normality of Maximum Quasileikelihood Estimators (MQLE) under very similar conditions. Fahrmeir [12] also verified these conditions in several examples like regressors with a compact range, response  $\mathbf{y}$  with bounded range and univariate models, etc.

### 2.3 Generalized Estimating Equations

The class of GLM's (Nelder and Wedderburn [23]) plays a central role in regression problems with discrete or nonnegative responses. This class of regression models was extended by Liang and Zeger [37] to analyze longitudinal or batch correlated data. The Liang and Zeger approach is known as Generalized Estimating

Equations (GEE). The definition of GEE is presented in Section 2.3.1. Several approaches to the asymptotic properties of GEE are summarized in Section 2.3.2.

### 2.3.1 Definition

The GEE model is the following: Suppose  $(y_{ij}, \mathbf{x}_{ij})$  are observations for the  $j$ th measurement on the  $i$ th subject,  $j = 1, 2, \dots, n_i$  and  $i = 1, 2, \dots, m$ , where  $y_{ij}$  is a scalar response,  $\mathbf{x}_{ij}$  is a  $p \times 1$  covariate vector, and  $n_i$  is the cluster size. Assume that the observations on different subjects are independent and the observations on the same subjects are correlated. For  $i = 1, \dots, m$ , let  $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^T$  and  $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})^T$ . Liang and Zeger [37] used a generalized linear model to model the marginal density of  $y_{ij}$  (with respect to a  $\sigma$ -finite measure  $\xi$ ):

$$f(y_{ij} | \mathbf{x}_{ij}, \boldsymbol{\beta}, \phi) = \exp[\{y_{ij}\theta_{ij} - b(\theta_{ij}) + c(y_{ij})\}/\phi]. \quad (2.8)$$

As in Section 2.1,  $\theta_{ij} = u(\eta_{ij})$ ,  $u$  is a known injective function mapping  $\Re$  into  $\Theta$  and  $\eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$ . The vector  $\boldsymbol{\beta}$  contains the regression parameters of interest, and  $\phi$  is a nuisance scale parameter. Under such a model specification, the first two moments of  $y_{ij}$  are given by

$$\mu_{ij}(\boldsymbol{\beta}) = E(y_{ij} | \mathbf{x}_{ij}, \boldsymbol{\beta}, \phi) = b'(\theta_{ij}), \quad \sigma^2(\boldsymbol{\beta}) = Cov(y_{ij} | \mathbf{x}_{ij}, \boldsymbol{\beta}, \phi) = b''(\theta_{ij})\phi. \quad (2.9)$$

Let  $g(t) = (b' \circ u)^{-1}(t)$ ; then  $g(\mu_{ij}(\boldsymbol{\beta})) = \mathbf{x}_{ij}^t \boldsymbol{\beta}$ . The function  $g(t)$  is the link function and its inverse function  $h(s) = (b' \circ u)(s)$  is called the inverse link function. Of importance are the canonical link functions, where  $u(s) = s$ , so  $g(t) = (b')^{-1}(t)$  and  $h(s) = b'(s)$ .

Let  $\mu_i(\boldsymbol{\beta}) = E(\mathbf{y}_i) = (\mu_{i1}(\boldsymbol{\beta}), \dots, \mu_{in_i}(\boldsymbol{\beta}))^T$  and  $\Sigma_i(\boldsymbol{\beta}) = \text{Cov}(\mathbf{y}_i)$ . We write  $\mathbf{A}_i(\boldsymbol{\beta}) = \text{diag}(\sigma_{i1}^2(\boldsymbol{\beta}), \dots, \sigma_{in_i}^2(\boldsymbol{\beta}))$  and  $\Delta_i(\boldsymbol{\beta}) = \text{diag}(u'(\mathbf{x}_{i1}^T \boldsymbol{\beta}), \dots, u'(\mathbf{x}_{in_i}^T \boldsymbol{\beta}))$ , where, for any vector  $\mathbf{v}$ ,  $\text{diag}(\mathbf{v})$  represents a diagonal matrix whose diagonal elements are the elements of  $\mathbf{v}$ . Let  $\mathbf{D}_i(\boldsymbol{\beta}) = \mathbf{A}_i(\boldsymbol{\beta})\Delta_i(\boldsymbol{\beta})\mathbf{X}_i$  and  $\mathbf{V}_i(\boldsymbol{\beta}, \alpha) = \mathbf{A}_i^{\frac{1}{2}}(\boldsymbol{\beta})\mathbf{R}_i(\alpha)\mathbf{A}_i^{\frac{1}{2}}(\boldsymbol{\beta})$ . Here  $\mathbf{R}_i(\alpha)$  is the “working” correlation matrix which one can choose freely and which may possibly contain a nuisance parameter (or parameter vector)  $\boldsymbol{\alpha}$ . If  $\mathbf{R}_i(\alpha)$  is equal to the true (often unspecified) correlation matrix  $\bar{\mathbf{R}}_i$ , then  $\mathbf{V}_i(\boldsymbol{\beta}_0, \alpha) = \Sigma_i(\boldsymbol{\beta}_0)$ . Liang and Zeger [37] proposed solving the following equations:

$$\mathbf{g}_{nm}(\boldsymbol{\beta}) = \sum_{i=1}^m \mathbf{g}_{n_i,i} = \sum_{i=1}^m \mathbf{D}_i(\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\boldsymbol{\beta}, \alpha) (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0, \quad (2.10)$$

which they called “generalized estimating equations.” Let  $\lambda_{\min}(\mathbf{T})$  ( $\lambda_{\max}(\mathbf{T})$ ) denote the smallest (largest) eigenvalue of the matrix  $\mathbf{T}$ , and define

$$\begin{aligned} \mathbf{M}_{nm}(\boldsymbol{\beta}) &= \text{Cov}(\mathbf{g}_{nm}(\boldsymbol{\beta})) \\ &= \sum_{i=1}^m \mathbf{D}_i^T(\boldsymbol{\beta}) \mathbf{V}_i^{-1}(\boldsymbol{\beta}, \alpha) \Sigma_i(\boldsymbol{\beta}) \mathbf{V}_i^{-1}(\boldsymbol{\beta}, \alpha) \mathbf{D}_i(\boldsymbol{\beta}), \end{aligned} \quad (2.11)$$

$$\mathcal{D}_{nm}(\boldsymbol{\beta}) = -\frac{\partial \mathbf{g}_{nm}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}, \quad (2.12)$$

$$\mathbf{H}_{nm}(\boldsymbol{\beta}) = \sum_{i=1}^m \mathbf{D}_i^T(\boldsymbol{\beta}) \mathbf{V}_i^{-1}(\boldsymbol{\beta}, \alpha) \mathbf{D}_i(\boldsymbol{\beta}). \quad (2.13)$$

Let

$$\mathbf{g}_{nm} = \mathbf{g}_{nm}(\boldsymbol{\beta}_0), \quad \mathbf{H}_{nm} = \mathbf{H}_{nm}(\boldsymbol{\beta}_0),$$

and

$$\mathbf{M}_{nm} = \mathbf{M}_{nm}(\boldsymbol{\beta}_0), \quad \mathbf{F}_{nm} = \mathbf{H}_{nm} \mathbf{M}_{nm}^{-1} \mathbf{H}_{nm}.$$

### 2.3.2 Asymptotics

The GEE method allows the nuisance parameter to be determined from the sample, which extends the flexibility and applicability of the method. Furthermore, the covariance matrices  $\Sigma_i(\beta)$  can be estimated from the data, so that large-sample testing and interval estimation is possible. In the past, most research in GEE has been directed to methodological development and modeling issues. Most of the work relies on the asymptotic results presented by Liang and Zeger [37], in which exact conditions are not specified. Xie and Yang [35] developed a set of (information matrix based) general conditions, which leads to the proof of weak and strong consistency as well as asymptotic normality. In both papers, the dimension of  $\beta$  is fixed.

Liang and Zeger [37] claim consistency and asymptotic normality of GEE estimator under regularity conditions when the number of independent subjects goes to infinity and the number of observations on each subject stays bounded. Exact conditions are not specified. Efficiency is improved if the “working” correlation is correct or close to correct.

Xie and Yang [35] present asymptotic properties of the GEE estimator  $\hat{\beta}_{nm}$  in each of three distinct large sample settings:

- (i)  $m \rightarrow \infty$  and  $n = n(m) = \max_{1 \leq i \leq m} n_i$  is uniformly bounded for all  $m$ ,
- (ii)  $m$  is bounded but  $n \rightarrow \infty$ ,
- (iii)  $n \rightarrow \infty$  as  $m \rightarrow \infty$ ,

where  $m$  is the number of independent clusters and  $n_i$  is the cluster size.

Liang and Zeger [37] only consider large sample setting (i). Most of the conditions Xie and Yang [35] derived for consistency and asymptotic normality of GEE estimator parallel the elegant conditions presented by Fahrmeir and Kaufmann [11]. They also ignore the nuisance parameter  $\phi$  in their equations (2.10) and (2.11), following (2.4) and (2.5) of Fahrmeir and Kaufmann [11]. Also for simplicity, they do not study the effect of estimating the nuisance parameter  $\alpha$  that appears in the working correlation matrix  $R_i(\alpha)$ .

In addition to regularity conditions on the parameter space, link function and covariates, Xie and Yang [?] propose the following conditions for asymptotic behaviors of GEE estimation:

$$(I_w) \quad \lambda_{\min}(\mathbf{F}_{nm}) \rightarrow \infty.$$

$(L_w)$  There exists a constant  $c_0 > 0$ , for any  $r > 0$  such that

$$P(\mathcal{D}_{nm}^T(\boldsymbol{\beta}) \mathbf{M}_{nm}^{-1} \mathcal{D}_{nm}(\boldsymbol{\beta}) \geq c_0 \mathbf{F}_{nm} \text{ and}$$

$$\mathcal{D}_{nm}(\boldsymbol{\beta}) \text{ is nonsingular , for all } \boldsymbol{\beta} \in \mathbf{B}_{nm}(r)) \rightarrow 1$$

$$\text{where } \mathbf{B}_{nm}(r) = \{\boldsymbol{\beta} : \|\mathbf{M}_{nm}^{-\frac{1}{2}} \mathbf{H}_{nm}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\| \leq r\}.$$

$$(I_w^*) \quad (\tau_{nm})^{-1} \lambda_{\min}(\mathbf{H}_{nm}) \rightarrow \infty, \quad \text{where } \tau_{nm} = \max_{1 \leq i \leq m} \{\lambda_{\max}(\mathbf{R}_i^{-1}(\alpha) \bar{\mathbf{R}}_i)\}.$$

$(L_w^*)$  There exists a constant  $c_0$ , for any  $\delta > 0$  and  $r > 0$ , such that

$$P(\mathcal{D}_{nm}(\boldsymbol{\beta}) \geq c_0 \mathbf{H}_{nm} \text{ and}$$

$$\mathcal{D}_{nm}(\boldsymbol{\beta}) \text{ is nonsingular, for } \boldsymbol{\beta} \in \mathbf{B}_{nm}^*(r)) \rightarrow 1$$

where  $\mathbf{B}_{nm}^*(r) = \{\boldsymbol{\beta} : \|\mathbf{H}_{nm}^{\frac{1}{2}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\| \leq (\tau_{nm})^{\frac{1}{2}}r\}$ .

(CC) For any given  $r > 0$  and  $\delta > 0$

$$P \left( \sup_{\boldsymbol{\beta} \in \mathbf{B}_{nm}^*(r)} \|\mathbf{H}_{nm}^{-\frac{1}{2}} \mathcal{D}_{nm}(\boldsymbol{\beta}) \mathbf{H}_{nm}^{-\frac{1}{2}} - \mathbf{I}\| < \delta \right) \rightarrow 1.$$

The matrix norm is the Euclidean matrix norm and their main theorems are the following:

**Theorem 2.3.1.** *Under conditions  $(I_w)$  and  $(L_w)$ , there exists a sequence of random variables  $\hat{\boldsymbol{\beta}}_{nm}$ , such that*

$$P(\mathbf{g}_{nm}(\hat{\boldsymbol{\beta}}_{nm}) = 0) \rightarrow 1$$

and

$$\hat{\boldsymbol{\beta}}_{nm} \rightarrow \boldsymbol{\beta}_0 \text{ in probability.}$$

**Theorem 2.3.2.** *The results of Theorem 2.3.1 hold if  $(I_w)$  and  $(L_w)$  are replaced by  $(I_w^*)$  and  $(L_w^*)$ , respectively.*

**Theorem 2.3.3.** *Suppose that conditions  $(I_w)$ ,  $(L_w)$  and (CC) hold, or that conditions  $(I_w^*)$  and (CC) hold. Then there exists a sequence of solutions  $\hat{\boldsymbol{\beta}}_{nm}$  to the GEE equation in  $\mathbf{B}_{nm}^*(r)$  such that  $\mathbf{M}_{nm}^{-\frac{1}{2}} \mathbf{H}_{nm}(\hat{\boldsymbol{\beta}}_{nm} - \boldsymbol{\beta}_0)$  and  $\mathbf{M}_{nm}^{-\frac{1}{2}} \mathbf{g}_{nm}$  are asymptotically identically distributed.*

For  $t > 0$ , let  $\psi(t)$  be a positive nondecreasing function such that  $\lim_{t \rightarrow \infty} \psi(t) = \infty$  and  $t\psi(t)$  is a convex function. Xie and Yang [35] use  $\psi(t)$  to establish Lindeberg conditions. Examples include  $\psi(t) = t^{1/\delta}$ ,  $\delta > 0$  and  $\psi(t) = \exp(t)$ .

**Lemma 2.3.4.** Under the GEE setting, suppose there exist a constant  $K$  (independent of  $n$ ) and an integer  $n_0$  such that, for  $j = 1, 2, \dots, n_i$  and  $i = 1, 2, \dots, m$ , when  $n > n_0$

$$E[y_{ij}^{*2}\psi(y_{ij}^{*2})] \leq K$$

where  $\mathbf{y}_i^* = (y_{i1}^*, \dots, y_{im_i}^*)^T = \mathbf{A}_i^{-\frac{1}{2}}(\mathbf{y}_i - \boldsymbol{\mu}_i)$ . In addition, for any  $\varepsilon > 0$ ,

$$c_{nm}\tilde{\lambda}_{nm}n \left[ \psi \left( \frac{\varepsilon^2}{c_{nm}\tilde{\lambda}_{nm}n\gamma_{nm}^{(D)}} \right) \right]^{-1} \rightarrow 0$$

Then, when  $m \rightarrow \infty$ , we have

$$\mathbf{M}_{nm}^{-\frac{1}{2}}\mathbf{g}_{nm} \rightarrow N(\mathbf{0}, \mathbf{I}) \text{ in distribution}$$

where  $c_{nm} = \lambda_{\max}(\mathbf{M}_{nm}^{-1}\mathbf{H}_{nm})$ ,

$$\gamma_{nm}^{(D)} = \max_{1 \leq i \leq m} \lambda_{\max}(\mathbf{H}_{nm}^{-\frac{1}{2}}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i\mathbf{H}_{nm}^{-\frac{1}{2}}).$$

Strong consistency of the GEE estimator can be proven under the following condition.

( $L_s$ ) In a neighborhood of  $\boldsymbol{\beta}_0$ , say  $N$ , there exists a constant  $c_0 > 0$  (independent of  $m$ ) and  $\delta > 0$  such that when  $m \rightarrow \infty$ .

$$\lambda_{\min}(\mathcal{D}_{nm}(\boldsymbol{\beta})^T\mathbf{M}_{nm}^{-1}\mathcal{D}_{nm}(\boldsymbol{\beta})) \geq c_0(\log m)^{2(1+\delta)}$$

and  $\mathcal{D}_{nm}(\boldsymbol{\beta})$  is nonsingular a.s. for  $\boldsymbol{\beta} \in N$ .

**Theorem 2.3.5.** Suppose  $\mathbf{g}_{n,i,i}$ ,  $i = 1, \dots, m$ , the summands of the GEE score function  $\mathbf{g}_{nm}$ , form a infinitesimal double array sequence. Under condition ( $L_s$ ), there exist a sequence of random variables  $\hat{\boldsymbol{\beta}}_{nm}$  and a random number  $n_0$ , such that

$$P(\mathbf{g}_{nm}(\hat{\boldsymbol{\beta}}_{nm}) = 0, \text{ for all } n \geq n_0) = 1$$

and when  $m \rightarrow \infty$

$$\hat{\beta}_{nm} \rightarrow \beta_0 \quad a.s.$$

Also they stated that when  $m$  is bounded or  $n$  goes to infinity too fast, asymptotic normality of  $\mathbf{g}_{nm}$  or  $\hat{\beta}_{nm}$  does not hold without specifying the dependence structure on each subject. They even point out that if we put  $R_i = \mathbf{I}$ ,  $\mathbf{g}_{nm}(\beta)$  is asymptotically normally distributed if and only if  $n/m \rightarrow 0$ . Also for settings (i) and (iii) when  $n$  is bounded above or tends to infinity at a limited rate as  $m \rightarrow \infty$ , they present a set of sufficient conditions to ensure asymptotic normality of  $\mathbf{g}_{nm}$  and  $\hat{\beta}_{nm}$ .

They verify these general conditions for some cases of practical importance, such as marginal GLM with compact covariate set, marginal Poisson regression model and marginal GLMs with bounded responses (binomial or polytomous regression models).

A drawback to the GEE approach is that it assumes all subjects have the same covariance structure. The “working” correlation matrix is not necessarily an essential feature of GEE. Jiang [16] proposed a nonparametric quasi-likelihood approach for getting a nonparametric estimator of the unknown covariance matrices. Chiou and Müller [7] proposed Estimated Estimating Equations (EEE) whose covariance structure is modeled nonparametrically as a function of the mean and therefore is an essential component that is part of the model fitting.

The difference between the approaches of Jiang [16] and Chiou and Müller [7] is that the mean function is correctly specified in Jiang [16] but unknown in Chiou

and Müller [7]. It has been shown by Jiang [16] that the estimator obtained from the nonparametric quasi-likelihood approach has an asymptotic normal distribution, the same as the estimator obtained from quasi-likelihood approach with true covariance matrix. Moreover, the rate of convergence has been established. Chiou and Müller [7] gave a sketchy proof on consistency and asymptotic normality of their EEE estimator along the lines of McCullagh and Nelder [19].

## 2.4 Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMM's) are a natural extension of GLM by including random effects. It is usually assumed that the random effects have a multivariate normal distribution whose variance components are to be estimated from the data. In Section 2.4.1 the definition of GLMM is presented and some asymptotic results are discussed in Section 2.4.2.

### 2.4.1 Definitions

The structure of GLMM's (McCulloch and Searle [22]) is the following:

The response vector  $\mathbf{y}$  is typically, but not necessarily, assumed to consist of conditionally independent elements, each with a distribution with density from the exponential family:

$$y_i | \boldsymbol{\alpha} \sim \text{independent}, \quad f_{Y_i|\boldsymbol{\alpha}}(y_i | \boldsymbol{\alpha})$$

$$f_{Y_i|\boldsymbol{\alpha}}(y_i | \boldsymbol{\alpha}) = \exp\{[y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}.$$

$$E[y_i|\boldsymbol{\alpha}] = \mu_i$$

$$g(\mu_i) = \mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\alpha} = \eta_i.$$

Here,  $g(\cdot)$  is a known function, called the link function,  $\mathbf{x}_i^t$  is the  $i$ th row of the model matrix for the fixed effects, and  $\boldsymbol{\beta}$  is the fixed effects parameter vector. To that specification we have added  $\mathbf{z}_i^t$ , which is the  $i$ th row of the model matrix for the random effects, and  $\boldsymbol{\alpha}$ , the random effects vector. To complete the specification we assign a distribution to the random effects:

$$\boldsymbol{\alpha} \sim f_{\alpha}(\boldsymbol{\alpha}|\mathbf{D})$$

$$E(\boldsymbol{\alpha}) = \mathbf{0}$$

where  $\mathbf{D}$  represents the parameters governing the distribution of  $\boldsymbol{\alpha}$ . Often  $f(\boldsymbol{\alpha}|\mathbf{D})$  is multivariate normal and  $\mathbf{D}$  is the covariance matrix.

In GLMM, the canonical parameter  $\theta_i$  is related to the covariate by  $\theta_i = \theta(\eta_i)$ . When  $\theta_i = \eta_i$ , the link is said to be the canonical link.

#### 2.4.2 Estimation approaches

According to Jiang [14], the GLMM setup is divided into two cases: the case where there is enough information about the random effects and the case where there is not. The first case is characterized by  $m/N \rightarrow 0$  (case 1), while the second by  $m/N \not\rightarrow 0$  (case 2), where  $m$  is the dimension of the random effects and  $N$  is the sample size.

In case 1 when the dimension of fixed effects is fixed, Sartori and Severini [28] extend Davison's [9] conditional likelihood approach for GLMs to GLMMs. The

Conditional Maximum Likelihood Estimate (CMLE) is defined by Andersen [1] in the following way:

To describe the situation Neyman and Scott [24] introduced the concept of structural and incidental parameters as follows. Consider a sequence of independent random variables  $X_1, X_2, X_3, \dots$ . The distribution of  $X_i$  depends on the parameters  $\beta$  and  $\tau_i$ , where the value of  $\beta$  is the same, independent of  $i$ , while the value of  $\tau_i$  changes with  $i$ . Then  $\beta$  is called a structural parameter and the  $\tau$ 's incidental parameters.

Andersen [1] discussed a general method for obtaining consistent estimates for a structural parameters  $\beta$  of fixed dimension in the presence of an increasing number of incidental parameters. He eliminated the incidental parameters by considering the conditional distribution given minimal sufficient statistics for the  $\tau$ 's. The value of  $\beta$  that maximizes this conditional distribution is then called the Conditional Maximum Likelihood Estimate (CMLE) for  $\beta$ .

Sartori and Severini [28] showed that the conditional likelihood function is valid for any distribution of the random effects. Hence, the inferences about the fixed effects are insensitive to misspecification of the random effect distribution. Furthermore, Andersen [1] showed that the convergence of the normalized  $\hat{\beta}$  to a normal distribution holds given the random effects. Hence, the asymptotic normality of  $\hat{\beta}$  is valid for any random effect distribution.

Li, Lindsay and Waterman [18] considered a rectangular array asymptotic embedding for multistratum datasets, in which both the number of strata and the number of within-stratum replications increase, and at the same rate. They pointed

that under this embedding the MLE is consistent but may not be efficient owing to a non-zero mean in its asymptotic normal distribution. By using a projection operator on the score function, an adjusted MLE can be obtained that is asymptotically unbiased and has a variance that attains the Cramer-Rao lower bound. The adjusted MLE can be viewed as an approximation to the conditional MLE.

In case 2 with fixed dimension of fixed effects, Sinha [30] develops a technique for finding a Robust Maximum Likelihood (RML) estimate of the model parameters in GLMM's by using Huber's  $\psi$  function and the Mahalanobis distance, which appears to be useful in downweighting the influential data points when estimating the parameter. The asymptotic properties of RMLE are investigated under regularity conditions. Sinha [30] also proposed a Robust Monte Carlo Newton-Raphson (RMCNR) algorithm for fitting GLMM's to avoid the computational problems involving high-dimensional integrals. RMCNR can be considered as a modification of the Monte Carlo Newton-Raphson (MCNR) method of McCulloch [21].

Breslow and Clayton [8] considered two closely related approximate methods of inference in GLMM's: Penalized Quasi-likelihood (PQL), which is based on integrated quasi-likelihood for integral approximation, and Marginal Quasi-likelihood (MQL). The major difference between those two is that PQL has  $E(\mathbf{y}|\boldsymbol{\alpha}) = h(\mathbf{X}^T \boldsymbol{\beta} + \mathbf{Z}^T \boldsymbol{\alpha})$  in which  $\boldsymbol{\alpha}$  is a vector of random effects and MQL only has  $E(\mathbf{y}) = h(\mathbf{X}^T \boldsymbol{\beta})$ .

Lee and Nelder [17] developed a joint likelihood, called h-likelihood, for Hierarchical Generalized Linear Models (HGLMs) which allows extra error components in the linear predictors of GLMM's. In order to get a marginal likelihood, one has to integrate out the random effects from the joint likelihood. However, this

integration is often quite intractable and at the same time it makes random effects nonestimable. By contrast the h-likelihood is easily available and avoids the need for burdensome integration. Under appropriate conditions, Lee and Nelder [17] showed that a random effect MHLE (Maximum h-likelihood estimator) is an asymptotically best unbiased predictor and that a fixed effect MHLE is asymptotically efficient as the marginal MLE. With the h-likelihood, the scaled deviance test and test statistics for fixed and random effects offer a simple unified framework of analysis. They also proposed an extended quasi-h-likelihood and several algorithms.

#### 2.4.3 Random intercept model (canonical link)

The penalized methods proposed by Jiang [14] are discussed in detail in Section 2.5 and we have the following important case of GLMM.

Jiang [14] considers a special case of GLMM in which the responses are clustered into groups with each group associated with a single random effect (possibly vector valued). Suppose that given unobservable random vectors  $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_m$  satisfying  $E(\boldsymbol{\alpha}_i) = 0$  the responses  $y_{ij}$ ,  $1 \leq i \leq m$ ,  $1 \leq j \leq n_i$ , ( $n_i \geq 1$ ) are independent with  $E(y_{ij}|\boldsymbol{\alpha}) = b'_{ij}(\eta_{ij})$ , where  $b_{ij}(\cdot)$  is differentiable. Write,

$$\eta_{ij} = a + \mathbf{x}_{ij}^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\alpha}_i$$

where  $a$  is an unknown intercept,  $\boldsymbol{\beta} = (\beta_k)_{1 \leq k \leq s}$  ( $s$  is fixed) is an unknown vector of regression coefficients, and  $\mathbf{x}_{ij} = (x_{ijk})_{1 \leq k \leq s}$  and  $\mathbf{z}_i$  are known vectors. Such models are useful, for example, in the context of small-area estimation in which

$\alpha_i$  represents a random effect associated with the  $i$ th selected area. Here we are interested in the estimation of the fixed effect  $a$ ,  $\beta_k$ ,  $1 \leq k \leq s$ , and the “area-specific” random effects  $v_i = \mathbf{z}_i^t \boldsymbol{\alpha}_i$ ,  $1 \leq i \leq m$ . Therefore, we may assume that in the above model  $\eta_{ij}$  has the following expression:

$$\eta_{ij} = a + \mathbf{x}_{ij}^t \boldsymbol{\beta} + v_i$$

where  $v_1, \dots, v_m$  are random variables with  $E(v_i) = 0$ . Note that here we regard  $a_i = a + v_i$  as a random intercept.

Logistic  $2 \times 2 \times m$  table is an important example in this case and it is modeled as  $\text{logit}P(y_{ij} = 1 | \alpha_i) = \alpha_i + x_{ij}\beta$  where  $\alpha_i$  is the random effect,  $\beta$  is the common log odds ratio and  $x_{ij}=0$  or 1.

## 2.5 Generalized GLMM (canonical link function case)

In order to apply GLMM, one has to know the distribution of random effects. In fact, in many problems little is known about the distribution of random effects. Therefore, it is of practical interest to develop methods and models that do not require strong distributional assumptions. In Section 2.5.1 the definition of generalized GLMM is given. Jiang [14] proposed methods called Penalized Generalized Weighted Least Squares (PGWLS) and Maximum Conditional Likelihood Estimate (MCLE) which are discussed in Sections 2.5.2 and 2.5.3 respectively. Jiang’s [14] consistency results are summarized in Section 2.5.4.

### 2.5.1 Definitions

Jiang [14] generalized the definition of GLMM's without assuming distributions of random effects and  $y_i$  by conditioning on random effects in the following way.

Suppose that, given a vector  $\boldsymbol{\alpha} = (\alpha_k)_{1 \leq k \leq m}$  of unobservable random variables (the random effects) satisfying

$$E(\boldsymbol{\alpha}) = 0, \quad (2.14)$$

the responses  $y_1, \dots, y_N$  are independent with conditional expectation

$$E(y_i | \boldsymbol{\alpha}) = b'_i(\eta_i) \quad (2.15)$$

where  $b_i(\cdot)$  is a differentiable function. Let suppose

$$\eta_i = \mathbf{x}_i^t \boldsymbol{\beta} + \mathbf{z}_i^t \boldsymbol{\alpha} \quad (2.16)$$

where  $\boldsymbol{\beta} = (\beta_j)_{1 \leq j \leq p}$  is a vector of unknown constants (the fixed effects), and  $\mathbf{x}_i = (x_{ij})_{1 \leq j \leq p}$ ,  $\mathbf{z}_i = (z_{ik})_{1 \leq k \leq m}$  are known vectors,  $1 \leq i \leq N$ . In vector notation,  $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}$ .

### 2.5.2 Penalized Generalized Weighted Least Squares in Case 1

The method of maximum likelihood is widely used for analyzing GLMMs. A full maximum likelihood analysis requires numerical integration techniques to calculate the log-likelihood and also the distribution of random effects needs to be known. Jiang [14] proposed a method of inference which in many ways resembles

the method of Least Squares (LS) in linear models and relies on weak distributional assumptions about random effects.

Assume without loss of generality that  $\text{rank}(\mathbf{X}) = p$  and no column of  $\mathbf{Z}$  is 0. In linear models (LMs), which correspond to (2.15) and (2.16) with  $b_i(\eta_i) = \eta_i^2/2$  and  $m = 0$  (i.e., there are no random effects), a well-known method is weighted least squares (WLS), which defines the estimate of  $\boldsymbol{\beta}$  as the minimizer of

$$\sum_{i=1}^N w_i(y_i - \eta_i)^2, \quad (2.17)$$

where  $w_i$ ,  $1 \leq i \leq N$ , are weights, or equivalently, the maximizer of

$$\sum_{i=1}^N w_i \left( y_i \eta_i - \frac{\eta_i^2}{2} \right). \quad (2.18)$$

A straightforward generalization of this method to the case of GLMM would suggest the maximizer of the following function as the estimates of  $\boldsymbol{\beta}$  and  $\boldsymbol{\alpha}$ :

$$\sum_{i=1}^N w_i(y_i \eta_i - b_i(\eta_i)). \quad (2.19)$$

However, conditionally, the individual fixed and random effects may not be identifiable. In LM there are two remedies when the identifiability problem arises, namely, reparameterization and constraints. We shall, for now, focus on the latter. A set of linear constraints on  $\boldsymbol{\alpha}$  may be expressed as  $\mathbf{P}\boldsymbol{\alpha} = \mathbf{0}$  for some matrix  $\mathbf{P}$ . By Lagrange's method of multipliers, maximizing (2.19) subject to  $\mathbf{P}\boldsymbol{\alpha} = \mathbf{0}$  is equivalent to maximizing

$$\sum_{i=1}^N w_i(y_i \eta_i - b_i(\eta_i)) - \frac{\lambda_1}{2} \|\mathbf{P}\boldsymbol{\alpha}\|^2 \quad (2.20)$$

without constraint, where  $\lambda_1$  is an additional variable. On the other hand, for fixed  $\lambda_1$  the last term in (2.20) may be regarded as a penalizer. The only thing that

needs to be specified is the matrix  $\mathbf{P}$ . For any matrix  $\mathbf{M}$  and vector space  $\mathbf{V}$ , let  $\mathcal{B}(\mathbf{V}) = \{\mathbf{B} : \mathbf{B} \text{ is a matrix whose columns constitute a base for } \mathbf{V}\}$ ;  $\mathcal{N}(\mathbf{M})$  = the null-space of  $\mathbf{M} = \{\mathbf{v} : \mathbf{M}\mathbf{v} = 0\}$ ;  $\mathbf{P}_{\mathbf{M}} = \mathbf{M}(\mathbf{M}^t\mathbf{M})^{-1}\mathbf{M}^t$ , and  $\mathbf{P}_{\mathbf{M}^\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{M}}$ . Let  $\mathbf{A} \in \mathcal{B}(\mathcal{N}(P_{\mathbf{X}^\perp}\mathbf{Z}))$  so that  $\mathbf{P}_{\mathbf{X}^\perp}\mathbf{Z}\mathbf{A} = \mathbf{0}$ . We define the penalized generalized WLS (PGWLS) estimate of  $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$  as the maximizer of

$$l_P(\boldsymbol{\gamma}) = \sum_{i=1}^N w_i(y_i\eta_i - b_i(\eta_i)) - \frac{\lambda_1}{2} \|\mathbf{P}_{\mathbf{A}}\boldsymbol{\alpha}\|^2. \quad (2.21)$$

where  $\lambda_1$  is a positive constant. The notation  $l_P$  is used because (2.21) may also be viewed as a penalized conditional quasi-log-likelihood.

Consider the expression (2.21). The reason that one needs a penalizer here is because the first term,  $l_C(\boldsymbol{\gamma}) = \sum_{i=1}^N w_i(y_i\eta_i - b_i(\eta_i))$ , depends on  $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$  only through  $\boldsymbol{\eta}$ . However,  $\boldsymbol{\gamma}$  can not be identified by  $\boldsymbol{\eta}$ , so there may be many vectors  $\boldsymbol{\gamma}$  for which  $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}$  is the same. The idea is therefore to consider a restricted space  $S = \{\boldsymbol{\gamma} : \mathbf{P}_{\mathbf{A}}\boldsymbol{\alpha} = 0\}$  such that within this subspace,  $\boldsymbol{\gamma}$  is uniquely determined by  $\boldsymbol{\eta}$ .

Case 1 Consistency Results.

Jiang [14] uses the following notations:

Let  $\mathbf{B} = (b_{ij})_{1 \leq i \leq k, 1 \leq j \leq l}$  be a matrix,  $\mathbf{v} = (v_i)_{1 \leq i \leq k}$  a vector and  $\mathbf{V}$  a vector space. Define  $\|\mathbf{v}\| = \max_{1 \leq i \leq k} |v_i|$ ;  $\|\mathbf{B}\| = \lambda_{\max}^{\frac{1}{2}}(\mathbf{B}^t\mathbf{B})$ ,  $\|\mathbf{B}\|_R = (\text{tr}(\mathbf{B}^t\mathbf{B}))^{\frac{1}{2}}$ ,  $\|\mathbf{B}\|_\infty = \max_{1 \leq i \leq k} \sum_{j=1}^l |b_{ij}|$ ;  $\mathbf{BV} = \{\mathbf{B}\mathbf{v} : \mathbf{v} \in \mathbf{V}\}$ ,  $\lambda_{\min}(\mathbf{B})|\mathbf{v} = \inf_{\mathbf{v} \in \mathbf{V} \setminus \{0\}} (\mathbf{v}^t\mathbf{B}\mathbf{v}/\mathbf{v}^t\mathbf{v})$ .

Jiang [14] presents consistency results for case 1 in the following Theorem 2.5.2 and 2.5.4 and Corollary 2.5.3.

**Theorem 2.5.1.** *Let  $b_i''(\cdot)$  be continuous, let  $\max_{1 \leq i \leq N} \{w_i^2 E\text{var}(y_i|\boldsymbol{\alpha}_0)\}$  be bounded*

and let

$$\frac{1}{N} \left[ (\max_{1 \leq u \leq p} |\mathbf{X}_u|^2) \|(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Z}\|^2 + (\max_{1 \leq k \leq m} |\mathbf{Z}_k|^2) \right] |\mathbf{P}_{\mathbf{A}} \boldsymbol{\alpha}_0|^2 \rightarrow_p 0. \quad (2.22)$$

Let  $c_N, d_N > 0$  be any sequences such that  $\limsup \|\boldsymbol{\beta}_0\|/c_N < 1$  and  $P(\|\boldsymbol{\alpha}_0\|/d_N < 1) \rightarrow 1$ ,  $M_i \geq c_N \sum_{u=1}^p |\mathbf{x}_{iu}| + d_N \sum_{k=1}^m |\mathbf{z}_{ik}|$ ,  $1 \leq i \leq N$  and  $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$  be the maximizer of  $l_P$  over  $\Gamma(M) = \{\boldsymbol{\gamma} : |\eta_i| \leq M_i, 1 \leq i \leq N\}$ . Then

$$\frac{1}{N} \left( \sum_{u=1}^p |\mathbf{X}_u|^2 (\hat{\beta}_u - \beta_{0u})^2 + \sum_{k=1}^m |Z_k|^2 (\hat{\alpha}_k - \alpha_{0k})^2 \right) \rightarrow_p 0 \quad (2.23)$$

provided that

$$\frac{p+m}{N} = o(\omega^2) \quad (2.24)$$

where

$$\omega = \lambda_{\min}(\mathbf{W}^{-1} \mathbf{H} \mathbf{W}^{-1}) \min_{1 \leq i \leq N} \{w_i \inf_{|h| \leq M_i} b_i''(h)\}$$

with

$$\mathbf{W} = \text{diag}(|\mathbf{X}_1|, \dots, |\mathbf{X}_p|, |\mathbf{Z}_1|, \dots, |\mathbf{Z}_m|).$$

**Corollary 2.5.2.** Let the conditions of Theorem 2.5.2 [including (2.24)] hold.

(i) Suppose  $p$  is fixed, and

$$\liminf \lambda_{\min}(\mathbf{X}^t \mathbf{X})/N > 0 \quad (2.25)$$

Then  $\hat{\boldsymbol{\beta}} \rightarrow_p \boldsymbol{\beta}_0$ .

(ii) Suppose  $\mathbf{Z} = (\mathbf{Z}_{(1)} \cdots \mathbf{Z}_{(q)})$  and correspondingly,  $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_q)$ , where  $\boldsymbol{\alpha}_u = (\alpha_{uv})_{1 \leq v \leq m_u}$ , and each  $\mathbf{Z}_{(u)}$  is a standard design matrix in the same sense as for  $\mathbf{U}$  defined below Lemma 2.5.1,  $1 \leq u \leq q$ . Let  $\mathbf{Z}_{uv}$  be the  $v$ th column of  $\mathbf{Z}_{(u)}$  and  $n_{uv} = |\mathbf{Z}_{uv}|^2$  = the number of appearances of the  $v$ th component of  $\boldsymbol{\alpha}_u$ .

Then

$$\left( \sum_{v=1}^{m_u} n_{uv} \right)^{-1} \sum_{v=1}^{m_u} n_{uv} (\hat{\alpha}_{uv} - \alpha_{0uv})^2 \rightarrow_p 0, \quad 1 \leq u \leq q, \quad (2.26)$$

where  $\hat{\alpha}_{uv}$  and  $\alpha_{0uv}$ ,  $1 \leq v \leq m_u$ ,  $1 \leq u \leq q$  are the corresponding components of  $\hat{\alpha}$  and  $\alpha_0$ , respectively.

For the special case of GLMM described in Section 2.4.3 as a random intercept model, the following theorem presents the consistency results:

**Theorem 2.5.3.** Let  $b''_{ij}(\cdot)$  be continuous; let  $w_{ij}^2 E\text{var}(y_{ij}|v_0)$ ,  $|x_{ij}|$  be bounded, let  $\liminf(\lambda_N/N) > 0$  where  $\lambda_N = \lambda_{\min}(\sum_{i=1}^m \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^t)$  with  $\bar{\mathbf{x}}_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$ . and let  $\bar{v}_0 \rightarrow_p 0$ . Let  $c_N, d_N > 0$  be such that  $\limsup |a_0| \vee |\beta_0|/c_N < 1$  and  $P(\|v_0\|/d_N < 1) \rightarrow 1$ ,  $M_{ij} \geq c_N(1 + |x_{ij}|) + d_N$  and  $\hat{\gamma} = (\hat{a}, \hat{\beta}, \hat{v})$  be the maximizer of  $l_P$  over  $\Gamma(M) = \{\gamma : |\eta_{ij}| \leq M_{ij}, \text{all } i, j\}$  and  $\delta_N = \min_{ij} \inf_{|h| \leq M_{ij}} b''(h)$ .

Then  $\hat{\beta} \rightarrow_p \beta_0$  and

$$\frac{1}{N} \sum_{i=1}^m n_i (\hat{a}_i - a_{0i})^2 \rightarrow_p 0, \quad (2.27)$$

where  $\hat{a}_i = \hat{a} + \hat{v}_i$  and  $a_{0i} = a_0 + v_{0i}$ , provided that  $m/N = o(\delta_N^2)$ . If the latter is strengthened to  $(\min_{1 \leq i \leq m} n_i)^{-1} = o(\delta_N^2)$ , then, in addition,  $\hat{a} \rightarrow_p a_0$ , and

$$\frac{1}{N} \sum_{i=1}^m n_i (\hat{v}_i - v_{0i})^2 \rightarrow_p 0, \quad \frac{1}{m} \sum_{i=1}^m (\hat{v}_i - v_{0i})^2 \rightarrow_p 0. \quad (2.28)$$

Note. It can be shown, by simple example, that  $\hat{\alpha} \rightarrow_p \alpha_0$  and (2.28) may not hold without  $\min_{1 \leq i \leq m} n_i \rightarrow \infty$ , even if  $m/N \rightarrow 0$ .

### 2.5.3 Maximum Conditional Likelihood Estimates in Case 2

In case 2 since  $m/N \not\rightarrow 0$ , we do not have enough information to estimate all random effects. Jiang [14] states that it is often possible to estimate with adequacy a subset of the random effects, and the ones which are not estimable will be integrated out. As in case 1, conditionally, the individual effects may not be identifiable. A basic technique here is reparameterization which is a map from  $(\beta, \alpha)$  to  $(\tilde{\beta}, \tilde{\alpha})$ .

Jiang's [14] reparameterization is stated in the following lemma:

**Lemma 2.5.4.** *There is a map  $\beta \mapsto \tilde{\beta}$ ,  $\alpha \mapsto \tilde{\alpha}$  such that*

(i)  $\mathbf{X}\beta + \mathbf{Z}\alpha = \tilde{\mathbf{X}}\tilde{\beta} + \tilde{\mathbf{Z}}\tilde{\alpha}$ , where  $(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}})$  is a known matrix of full column rank.

(ii)  $\tilde{\mathbf{z}}_i = \tilde{z}_{*j}$ ,  $i \in S_j$  for some known vector  $\tilde{z}_{*j}$ , where  $\tilde{z}_j^t$  is the  $i$ th row of  $\tilde{\mathbf{Z}}$

and where  $S_j$  is defined below.

Let  $U$  be a standard design matrix in the sense that it consists of 0's and 1's and there is exactly one 1 in each row and at least one 1 in each column. The vector  $u_i^t$  is the  $i$ th row of  $U$  and  $e_{M,j}$  is the  $M$ -dimensional vector whose  $j$ th component is 1 and whose other components are 0. Let  $S_j = \{1 \leq i \leq N : u_i = e_{M,j}\}$ , and  $y^{(j)} = (y_i)_{i \in S_j}$ ,  $1 \leq j \leq M$ . Suppose that  $\zeta_1, \dots, \zeta_M$  are independent with common distribution  $\nu(\cdot/\tau)/\tau$ , where  $\nu(\cdot)$  is a known density function and  $\tau > 0$  is an unknown scale parameter. Furthermore, we assume that there are no random effects nested within  $\zeta$ . In notation, this means that  $z_i = z_{*j}$ ,  $i \in S_j$ ,  $1 \leq j \leq M$ , where  $z_{*j} = (z_{*jk})_{1 \leq k \leq l}$ .

Let  $\varphi = (\tilde{\beta}, \tau)$ ,  $\psi = (\tilde{\alpha}, \varphi)$ . Then we have

$$f(y_i | \alpha, \zeta) = f(y_i | \eta_i), \quad 1 \leq i \leq N \quad (2.29)$$

where  $f(\xi_2|\xi_1)$  denotes the conditional density of  $\xi_2$  given  $\xi_1$ .

By Lemma 2.5.1, we have

$$\boldsymbol{\eta} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{Z}}\tilde{\boldsymbol{\alpha}} + \mathbf{U}\boldsymbol{\zeta}.$$

Let  $\boldsymbol{\varphi} = (\tilde{\boldsymbol{\beta}}, \tau)$ ,  $\boldsymbol{\psi} = (\tilde{\boldsymbol{\alpha}}, \boldsymbol{\varphi})$ . By (2.29) we have

$$f(y|\boldsymbol{\psi}) = \prod_{j=1}^M f(y^{(j)}|\boldsymbol{\psi})$$

and it is easy to show that  $f(y^{(j)}|\boldsymbol{\psi}) = g_j(z_{*j}^t \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tau)$ , where

$$g_j(s) = E \left[ \prod_{i \in S_j} f(y_i|s_1 + x_i^t s_{(2)} + s_{r+2} \zeta) \right],$$

$s_{(2)} = (s_2, \dots, s_{r+1})$  and  $r$  is the dimension of  $\tilde{\boldsymbol{\beta}}$ . Note that  $r \leq p$ . Let  $n$  be the dimension of  $\tilde{\boldsymbol{\alpha}}$ ,  $h_j(s) = \log g_j(s)$ ,  $l_C(\boldsymbol{\psi}) = \log f(y|\boldsymbol{\psi})$  and  $l_{C,j}(\boldsymbol{\psi}) = \log f(y^{(j)}|\boldsymbol{\psi}) = h_j(z_{*j}^t \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tau)$ . Then

$$l_c(\boldsymbol{\psi}) = \sum_{j=1}^M l_{c,j}(\boldsymbol{\psi})$$

Let  $\mathbf{Z}_*$  be the matrix whose  $j$ th row is  $z_{*j}^t$ ,  $1 \leq j \leq M$ . Let  $\boldsymbol{\varphi}_0$  and  $\boldsymbol{\psi}_0$  be the vectors corresponding to the true parameters and realization of random effects.

### Case 2 Consistency Results.

Under some regularity conditions on derivatives, smoothness, integrability of densities  $f(y^{(j)}|\boldsymbol{\psi})$ , boundness of  $h_j(s)$ 's second, third derivatives, etc., Jiang [14] proved consistency results for estimates of reparameterized parameters  $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}})$  but not for the original  $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ , and one may never recover a consistency result for the original parameter  $(\boldsymbol{\beta}, \boldsymbol{\alpha})$  from the reparameterized  $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}})$ .

Jiang [14] verifies the general conditions for consistency in three examples: for case 1, the two way crossed logistic model  $\text{logit}(P(y_{ij} = 1|a, b)) = \mu + a_i + b_j$ , the

other one is random intercept logistic regression model are analyzed. For case 2 the only example is two stage nested logistical model  $\text{logit}(P(y_{ijk} = 1|a, b)) = \mu + a_i + b_{ij}$ .

## Chapter 3

Case 1 random intercept model results.

This Chapter proposes a new estimator  $\sum_{i=1}^m \hat{\mathbf{w}}_i \hat{\boldsymbol{\beta}}_i$  and establishes its conditional asymptotic normality in the random intercept problem in case 1. Here  $\hat{\boldsymbol{\beta}}_i$  is the maximum likelihood estimator based on the  $i$ th group and  $\hat{\mathbf{w}}_i$  is a matrix weight proportional to the inverse estimated covariance of  $\hat{\boldsymbol{\beta}}_i$ . Also this Chapter extends the consistency results of Jiang [14] to establishing asymptotic normality of the penalized likelihood estimators for fixed intercept in the case 1 random intercept problem when there are no regression coefficients. Furthermore, this Chapter discusses the difficulties of establishing asymptotic normality of Jiang's [14] penalized likelihood estimators in general.

Throughout this chapter we consider the canonical GLMM with

$$f(y_{ij}|v_i) = \exp[\eta_{ij}y_{ij} - b(\eta_{ij}) + c(y_{ij})]$$

and  $\eta_{ij} = a + v_i + \mathbf{x}_i^t \boldsymbol{\beta} = a_i + \mathbf{x}_i^t \boldsymbol{\beta}$ .

Assumptions:

- (i)  $v_{i0}$  are iid,  $E(v_{i0}) = 0$ ,  $\text{Var}(v_{i0}) = \sigma_v^2$  where  $\sigma_v^2$  is a constant.  $\exists M$  such that  $P[|v_{i0}| \leq M] = 1$ .
- (ii)  $\exists K'$  such that  $\|\mathbf{x}_{ij}\| \leq K'$  for all  $i, j$ .

### 3.1 Case 1 random intercept (combine $a$ and $v_i$ )

Recalling the Estimating Equation (2.3), for the  $i$ th group in the random intercept model we have the loglikelihood function :

$$l_n(\boldsymbol{\gamma}_i) = \sum_{j=1}^{n_i} (y_{ij}\eta_{ij} - b(\eta_{ij}) - C(y_{ij})). \quad (3.1)$$

where  $\boldsymbol{\gamma}_i = (a_i, \beta_i^t)^t$ . Take Taylor expansion for partial derivative of  $l_n(\boldsymbol{\gamma}_i)$  around  $\boldsymbol{\gamma}_{i0}$ :

$$\begin{aligned} U_r(\hat{\boldsymbol{\gamma}}_i) &= \frac{\partial l_n(\boldsymbol{\gamma}_{ri})}{\partial \boldsymbol{\gamma}_i}(\hat{\boldsymbol{\gamma}}_i) = U_r(\boldsymbol{\gamma}_{i0}) + (\nabla U_r(\boldsymbol{\gamma}_i^*))(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i0}) \\ &= U_r(\boldsymbol{\gamma}_{i0}) + (-\mathbf{F}_{n_i})(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i0}) + [\mathbf{F}_{n_i} - \mathbf{F}_{n_i}(\boldsymbol{\gamma}^*)](\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i0}) \end{aligned} \quad (3.2)$$

where  $\boldsymbol{\gamma}_i^*$  is between  $\hat{\boldsymbol{\gamma}}_i$  and  $\boldsymbol{\gamma}_{i0}$  and  $-\mathbf{F}_{n_i}(\boldsymbol{\gamma}^*) = \nabla U_r(\boldsymbol{\gamma}_i^*)$ .

Because the random effects  $v_{i0}$  are independently and identically distributed with mean 0 and bounded and  $\|\mathbf{x}_{ij}\|$  are uniformly bounded, for all  $i, j$ ,  $b'_{ij}$  is bounded and  $b''_{ij}$  is bounded below and above by positive constants  $b_l$  and  $b_u$ , Furthermore, the consistency of  $\hat{\boldsymbol{\gamma}}_i$  implies that the  $b_{ij}^{(3)}(\tilde{\eta}_{ij})$  are bounded with high probability where  $\tilde{\eta}_{ij}$  is between  $\hat{\eta}_{ij}$  and  $\eta_{ij0}$ .

Define

$$(\mathbf{F}_i^{\beta\beta})^{-1} = \mathbf{W}_{bi} = (W_{kl}^{bi})_{1 \leq kl \leq s, 1 \leq l \leq s}$$

where

$$W_{kl}^{bi} = \sum_j b''_{ij} \left( x_{ijk} - \frac{\sum_j x_{ijk} b''_{ij}}{\sum_j b''_{ij}} \right) \left( x_{ijl} - \frac{\sum_j x_{ijl} b''_{ij}}{\sum_j b''_{ij}} \right). \quad (3.3)$$

**Lemma 3.1.1.** Let  $\liminf(\lambda_{n_i}/n_i) > 0$ , where  $\lambda_{n_i} = \lambda_{\min}(\sum_j (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T)$  with  $\bar{\mathbf{x}}_i = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$ . Then there exists a positive constant  $c_2$  such that for any  $h > 0$ ,

$$\inf_{\|\mathbf{u}\|=1} \frac{\mathbf{u}^T \mathbf{W}_{bi} \mathbf{u}}{n_i} > \frac{1}{c_2}. \quad (3.4)$$

for all  $n_i$  sufficiently large.

By considering each group separately, we construct a new estimator  $\hat{\boldsymbol{\beta}}_w = \sum_{i=1}^m \hat{\mathbf{w}}_i \hat{\boldsymbol{\beta}}_i$ . We can treat each group as a conditional GLM, as in Fahrmeir and Kaufmann [11]. We make the following assumptions: First we assume that  $(\mathbf{Z}_i, \mathbf{X}_i)$  is a full-column rank matrix and define a sequence  $N_{n_i}(\delta)$ ,  $\delta > 0$ , of neighborhoods of  $\boldsymbol{\gamma}_{i0}$  (the true value of  $\boldsymbol{\gamma}_i$ ) by

$$N_{n_i}(\delta) = \{\boldsymbol{\gamma}_i : \|\mathbf{F}_{n_i}^{T/2}(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_{i0})\| \leq \delta\}$$

where  $n_i = 1, 2, \dots$  and  $\mathbf{F}_{n_i}^{T/2}$  any right square root of the positive definite matrix  $\mathbf{F}_{n_i}$ ; that is,  $\mathbf{F}_{n_i}^{1/2} \mathbf{F}_{n_i}^{T/2} = \mathbf{F}_{n_i}$ . Conditions (D), (C) and (N) of Fahrmeir and Kaufmann [11] in our setting become:

- (1) Divergence:  $\lambda_{\min} \mathbf{F}_{n_i} \rightarrow \infty$ .
- (2) Boundedness from below: for all  $\delta > 0$ ,  $\mathbf{F}_{n_i}(\boldsymbol{\gamma}_i) - c \mathbf{F}_{n_i}$  is positive semidefinite, for all  $\boldsymbol{\gamma}_i \in N_{n_i}(\delta)$ ,  $n_i \geq n_1$  with some constants  $n_1 = n_1(\delta)$ ,  $c > 0$  independent of  $\delta$ .
- (3) Convergence and continuity: for all  $\delta > 0$ ,  $\max_{\boldsymbol{\gamma}_i \in N_{n_i}(\delta)} \|\mathbf{V}_{n_i}(\boldsymbol{\gamma}_i) - \mathbf{I}\| \rightarrow 0$ , where  $\mathbf{V}_{n_i}(\boldsymbol{\gamma}_i) = \mathbf{F}_{n_i}^{-1/2} \mathbf{F}_{n_i}(\boldsymbol{\gamma}_i) \mathbf{F}_{n_i}^{-T/2}$  is the normed information matrix.

Let  $(\hat{a}_i, \hat{\beta}_i)$  be the solution of the score equation in the  $i$ th group.

Then, given  $v_{i0}$ ,

$$(\mathbf{F}_{n_i})^{T/2} \begin{pmatrix} \hat{a}_i - a_{i0} \\ \hat{\beta}_i - \beta_{i0} \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \mathbf{I}). \quad (3.5)$$

Informally, given  $v_{i0}$ ,

$$\begin{pmatrix} \hat{a}_i \\ \hat{\beta}_i \end{pmatrix} \sim AN \left( \begin{pmatrix} a_{i0} \\ \beta_0 \end{pmatrix}, \mathbf{F}_{n_i}^{-1} \right), \quad (3.6)$$

where

$$\mathbf{F}_{n_i}^{-1} = \begin{pmatrix} \mathbf{Z}_i^T \mathbf{D}_i \mathbf{Z}_i & \mathbf{Z}_i^T \mathbf{D}_i \mathbf{X}_i \\ \mathbf{X}_i^T \mathbf{D}_i \mathbf{Z}_i & \mathbf{X}_i^T \mathbf{D}_i \mathbf{X}_i \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{F}_i^{aa} & \mathbf{F}_i^{a\beta} \\ \mathbf{F}_i^{\beta a} & \mathbf{F}_i^{\beta\beta} \end{pmatrix}. \quad (3.7)$$

Here

$$\mathbf{F}_i^{\beta\beta} = \left( \mathbf{X}_i^T \mathbf{D}_i \mathbf{X}_i - \mathbf{X}_i^T \mathbf{D}_i \mathbf{Z}_i (\mathbf{Z}_i^T \mathbf{D}_i \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \mathbf{D}_i \mathbf{X}_i \right)^{-1}$$

and

$$\mathbf{F}_i^{aa} = \left( \mathbf{Z}_i^T \mathbf{D}_i \mathbf{Z}_i - \mathbf{Z}_i^T \mathbf{D}_i \mathbf{X}_i (\mathbf{X}_i^T \mathbf{D}_i \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{D}_i \mathbf{Z}_i \right)^{-1}$$

where  $\mathbf{Z}_i = \mathbf{1}_{n_i}$  and  $\mathbf{D}_i = \text{diag}(b''_{ij})_{1 \leq j \leq n_i}$ .

We choose  $\hat{\mathbf{w}}_i = (\sum_i (\hat{\mathbf{F}}_i^{\beta\beta})^{-1})^{-1} (\hat{\mathbf{F}}_i^{\beta\beta})^{-1}$  and  $\hat{\beta}_w = \sum_i (\sum_i (\hat{\mathbf{F}}_i^{\beta\beta})^{-1})^{-1} (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \hat{\beta}_i$ .

Note that a scalar version of  $\hat{\beta}_w$  was proposed by Woolf (1955) to estimate the common log odds ratio in a  $2 \times 2 \times m$  contingency table.

**Theorem 3.1.2.** Suppose that Assumptions (i) and (ii) hold and for each  $i$ ,  $\liminf[n_i^{-1} \lambda_{n_i}] > 0$ ,  $n_i(\mathbf{F}_{n_i})^{-1} \rightarrow_p \mathbf{F}_{0i}$ ,  $P(\inf_i \lambda_{\min}(\mathbf{F}_{0i}) > \delta > 0) > 1 - h$ ,  $P(\sup_i \lambda_{\max}(\mathbf{F}_{0i}) < M < \infty) > 1 - h$ , where  $\mathbf{F}_{0i}$  is a positive definite matrix and that the following asymptotic relations are true:

(1) There exist positive numbers  $K_1 < K_2$  such that

$$K_1 < \min_i \lambda_{\min}(\mathbf{F}_i^{\beta\beta}) / \max_i \lambda_{\max}(\mathbf{F}_i^{\beta\beta}) < K_2,$$

(2)

$$m / \min_i n_i = o(1), \quad \max_i n_i / \min_i n_i = O(1),$$

(3)

$$\max_i (\lambda_{\max}(\mathbf{F}_{n_i})^{-1} / \lambda_{\min}(\mathbf{F}_{n_i})^{-1}) = O(1), \quad \sum_i \|(\mathbf{F}_{n_i})^{-1}\| = o(1),$$

(4)

$$\max_i E \left[ (\hat{\boldsymbol{\beta}}_i^*)^t \hat{\boldsymbol{\beta}}_i^* \psi((\hat{\boldsymbol{\beta}}_i^*)^t \hat{\boldsymbol{\beta}}_i^*) | v_{i0} \right] = O_p(1),$$

where  $\hat{\boldsymbol{\beta}}_i^* = (\mathbf{F}_i^{\beta\beta})^{-\frac{1}{2}} (\hat{\boldsymbol{\beta}}_i - E(\hat{\boldsymbol{\beta}}_i | v_{i0}))$  and  $\psi(t)$  is a positive nondecreasing function on  $[0, \infty)$  such that  $\lim_{t \rightarrow \infty} \psi(t) = \infty$  and  $t\psi(t)$  is a convex function.

Then, given  $\mathbf{v}_0$ ,

$$\left( \sum_{i=1}^m (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \sum_{i=1}^m (\mathbf{F}_i^{\beta\beta})^{-1} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathbf{I}). \quad (3.8)$$

Note that  $\mathbf{F}_i^{\beta\beta}$  depends on the unknown  $(a_i, \boldsymbol{\beta})$  through  $\mathbf{D}_i$ .

Let  $\hat{\mathbf{D}}_i = \text{diag}(b_{ij}''(\hat{\eta}_{ij})_{1 \leq j \leq n_i})$ , where  $\hat{\eta}_{ij} = \hat{a}_i + \mathbf{x}_{ij}^t \hat{\boldsymbol{\beta}}_i$ ,  $\hat{\mathbf{D}}_i = \text{diag}(b_{ij}''(\hat{\eta}_{ij})_{1 \leq j \leq n_i})$

where  $\hat{\eta}_{ij} = \hat{a}_i + \mathbf{x}_{ij}^t \hat{\boldsymbol{\beta}}_w$ .

Define

$$\hat{\mathbf{F}}_i^{\beta\beta} = \left( \mathbf{X}_i^T \hat{\mathbf{D}}_i \mathbf{X}_i - \mathbf{X}_i^T \hat{\mathbf{D}}_i \mathbf{Z}_i \left( \mathbf{Z}_i^T \hat{\mathbf{D}}_i \mathbf{Z}_i \right)^{-1} \mathbf{Z}_i^T \hat{\mathbf{D}}_i \mathbf{X}_i \right)^{-1}.$$

Then  $\hat{\mathbf{F}}_i^{\beta\beta}$  estimates  $\mathbf{F}_i^{\beta\beta}$  and  $\hat{\boldsymbol{\beta}}_w = \sum_{i=1}^m \left( \sum_{i=1}^m (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \right)^{-1} (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0)$ .

Define

$$\hat{\mathbf{F}}^{\beta\beta}_i = \left( \mathbf{X}_i^T \hat{\mathbf{D}}_i \mathbf{X}_i - \mathbf{X}_i^T \hat{\mathbf{D}}_i \mathbf{Z}_i \left( \mathbf{Z}_i^T \hat{\mathbf{D}}_i \mathbf{Z}_i \right)^{-1} \mathbf{Z}_i^T \hat{\mathbf{D}}_i \mathbf{X}_i \right)^{-1}.$$

Then  $\hat{\mathbf{F}}^{\beta\beta}_i$  estimates  $\mathbf{F}_i^{\beta\beta}$ .

**Theorem 3.1.3.** *Let the hypotheses of Theorem 3.1.2 hold. Then we have*

(1)

$$\left\| \left( \sum_{i=1}^m (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{\frac{1}{2}} \left( \sum_{i=1}^m (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} - \mathbf{I} \right\| = o_p(1),$$

(2)

$$\|(\hat{\mathbf{F}}_i^{\beta\beta})^{-1} - (\mathbf{F}_i^{\beta\beta})^{-1}\| = O_p(\sqrt{n_i}),$$

(3)

$$\left\| \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \right\| = O_p(1/\sqrt{N}), \quad \|(\mathbf{F}_i^{\beta\beta})^{1/2}\| = O_p(1/\sqrt{n_i}).$$

Furthermore, given  $\mathbf{v}_0$ , we have

$$\left( \sum_{i=1}^m (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \sum_{i=1}^m (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathbf{I}). \quad (3.9)$$

Under the same hypothesis with (1), (2) and (3) in Theorem 3.1.4 modified by replacing  $\hat{\mathbf{F}}_i^{\beta\beta}$  by  $\hat{\mathbf{F}}_i^{\beta\beta}$ , given  $\mathbf{v}_0$ ,

$$\left( \sum_{i=1}^m (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \sum_{i=1}^m (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{I}). \quad (3.10)$$

According to Theorem 3.1.2 and 3.1.3, given the random effects  $\mathbf{v}_0$ , the normalized versions of  $\hat{\boldsymbol{\beta}}_w$  and  $\hat{\boldsymbol{\beta}}_w$  converge in distribution to  $N(\mathbf{0}, \mathbf{I})$ . Since this convergence holds for almost all realizations of  $\mathbf{v}_0$ , we can also claim that the convergence in distribution holds unconditionally. This is formalized in the following Corollary.

**Corollary 3.1.4.** *Under the hypotheses of Theorem 3.1.2, the convergence statements (3.8), (3.9), and (3.10) hold unconditionally (except on a set of  $\mathbf{v}_0$ -probability zero).*

### 3.2 Penalized likelihood estimator $\hat{a}$ in the case 1 random intercept model when $\beta=0$

From Jiang [14] we can write  $l_P(\gamma)$  as follows:

$$l_P(\gamma) = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij}\eta_{ij} - b(\eta_{ij})) - \frac{\lambda_1}{2}m\bar{v}^2 \quad (3.11)$$

where  $\eta_{ij} = a + v_i$  and there are no regression coefficient. Here the penalty term  $(\lambda_1/2)m\bar{v}^2$  is used to take care of the identifiability problem and we can estimate  $a$  and  $v_i$  separately.

We will write

$$B_{ij} = b''_{ij} + \frac{1}{2}b_{ij}^{(3)}(\eta_{ij}^*)(\hat{\eta}_{ij} - \eta_{ij0}) = b''_{ij} + r_{ij} \quad (3.12)$$

where

$$r_{ij} = \frac{1}{2}b_{ij}^{(3)}(\eta_{ij}^*)(\hat{\eta}_{ij} - \eta_{ij0}) \quad (3.13)$$

and

$$\tilde{\mathbf{B}} = \text{diag}(B_{ij})_{1 \leq i \leq m, 1 \leq j \leq n_i}. \quad (3.14)$$

Taylor expansion of  $\partial l_P(\gamma)/\partial\gamma$  around  $\gamma_0$  takes the following matrix form:

$$\begin{pmatrix} \hat{a} - a_0 \\ \hat{\mathbf{v}} - \mathbf{v}_0 \end{pmatrix} = (\mathbf{Q}^*)^{-1} \begin{pmatrix} \mathbf{1}_N^T(\mathbf{y} - \boldsymbol{\mu}) \\ \mathbf{Z}^T(\mathbf{y} - \boldsymbol{\mu}) - \lambda_1 \bar{v}_0 \mathbf{1}_m^t \end{pmatrix} \quad (3.15)$$

where

$$\mathbf{Q}^* = \begin{pmatrix} \mathbf{1}_N^T \tilde{\mathbf{B}} \mathbf{1}_N & \mathbf{1}_N^T \tilde{\mathbf{B}} \mathbf{Z} \\ \mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{1}_N & \mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z} + \lambda_1/m \mathbf{J} \end{pmatrix} = \begin{pmatrix} Q_{11}^* & \mathbf{Q}_{12}^* \\ \mathbf{Q}_{21}^* & \mathbf{Q}_{22}^* \end{pmatrix}. \quad (3.16)$$

Let  $\mathbf{Z}_{new} = (\mathbf{1}_N, \mathbf{Z})$  and note that  $\mathbf{Q}^*$  is a nonsingular matrix since

$$\mathbf{Q}^* = \tilde{\mathbf{A}}^t \tilde{\mathbf{A}} + \tilde{\mathbf{H}}^t \tilde{\mathbf{H}}.$$

where

$$\tilde{\mathbf{A}} = \tilde{\mathbf{B}}^{\frac{1}{2}} \mathbf{Z}_{new} \quad \tilde{\mathbf{H}} = (0 \quad \frac{\sqrt{\lambda_1}}{\sqrt{m}} \mathbf{1}_m^t)$$

Here  $\tilde{\mathbf{A}}$  is an  $N \times (m+1)$  matrix with rank  $m$ , since  $\mathbf{Z}_{new}$  is not of full column rank matrix, and  $\tilde{\mathbf{H}}$  is a  $1 \times (m+1)$  matrix with rank 1. However,  $\tilde{\mathbf{A}}$  and  $\tilde{\mathbf{H}}$  are complementary matrices since  $\text{Sp}(\tilde{\mathbf{A}}^t) \cap \text{Sp}(\tilde{\mathbf{H}}^t) = \{0\}$ , we have  $\mathbf{Q}^*$  is a full-rank matrix with rank  $m+1$ .

Then

$$(\mathbf{Q}^*)^{-1} = \begin{pmatrix} (Q_{11.2}^*)^{-1} & -(Q_{11.2}^*)^{-1} \mathbf{Q}_{12}^* (\mathbf{Q}_{22}^*)^{-1} \\ -(\mathbf{Q}_{22}^*)^{-1} \mathbf{Q}_{21}^* (Q_{11.2}^*)^{-1} & (\mathbf{Q}_{22.1}^*)^{-1} \end{pmatrix}. \quad (3.17)$$

Since we're only interested in the  $\hat{a}$  terms,

$$\begin{aligned} (\hat{a} - a_0) &= (\mathbf{Q}_{11.2}^*)^{-1} (\mathbf{1}_N^T (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{Q}_{12}^* (\mathbf{Q}_{22}^*)^{-1} \mathbf{Z}^T (\mathbf{y} - \boldsymbol{\mu})) \\ &\quad + (\mathbf{Q}_{11.2}^*)^{-1} \mathbf{Q}_{12}^* (\mathbf{Q}_{22}^*)^{-1} \lambda_1 \bar{v}_0 \mathbf{1}_m. \end{aligned} \quad (3.18)$$

Here we choose  $\sqrt{m}$  as normalizer. Then

$$\begin{aligned} \sqrt{m}(\hat{a} - a_0) &= \sqrt{m}(\mathbf{Q}_{11.2}^*)^{-1} (\mathbf{1}_N^T (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{Q}_{12}^* (\mathbf{Q}_{22}^*)^{-1} \mathbf{Z}^T (\mathbf{y} - \boldsymbol{\mu})) \\ &\quad + \sqrt{m}(\mathbf{Q}_{11.2}^*)^{-1} \mathbf{Q}_{12}^* (\mathbf{Q}_{22}^*)^{-1} \lambda_1 \bar{v}_0 \mathbf{1}_m. \end{aligned} \quad (3.19)$$

**Theorem 3.2.1.** *We assume the following conditions, which are equivalent to the hypotheses of Theorem 2.5.4 (Jiang [14], Theorem 2.2).*

(J1)  $E(\text{Var}(y_{ij}|v_0))$  and  $\sum_k x_{ijk}^2$  are bounded.

(J2) Let  $\lambda_N = \lambda_{\min}(\sum_i \sum_j (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^t)$  with  $\bar{\mathbf{x}}_i = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$  and  $\delta_N = \min_{i,j} \inf_{|h| \leq M_{ij}} b''_{ij}(h)$ ,  $\liminf(\lambda_N/N) > 0$ ,  $m/N = o(\delta_N^2)$ ,  $(\min_{1 \leq i \leq m} n_i)^{-1} = o(\delta_N^2)$  and  $\bar{v}_0 \rightarrow_p 0$ .

(J3) Let  $c_N, d_N > 0$  be such that  $\limsup(|a_0| \vee |\boldsymbol{\beta}_0|)/c_N < 1$  and  $P(\|\mathbf{v}_0\|/d_N < 1) \rightarrow 1$ ,  $M_{ij} \geq c_N(1 + (\sum_k x_{ijk}^2)^{\frac{1}{2}}) + d_N$  and let  $\hat{\boldsymbol{\gamma}} = (\hat{a}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$  be the maximizer of  $l_P$  over  $\Gamma(\mathbf{M}) = \{\boldsymbol{\gamma} : |\eta_{ij}| \leq M_{ij}, \text{ all } i, j\}$ .

Then

$$\sqrt{m}(\hat{a} - a_0) \rightarrow_d N(0, \sigma_v^2).$$

### 3.3 Discussion of Jiang [14] penalized likelihood estimator $\hat{\boldsymbol{\beta}}$ in case 1 random intercept model ( $\boldsymbol{\beta} \neq 0$ )

According to Jiang [14], for the random intercept model the penalized loglikelihood is stated  $l_P(\boldsymbol{\gamma})$  as equation (??) in Section 2.5.2. Here for simplicity, we let  $w_{ij} = 1$  so that  $l_P(\boldsymbol{\gamma})$  becomes the following:

$$l_P(\boldsymbol{\gamma}) = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij}\eta_{ij} - b(\eta_{ij})) - \frac{\lambda_1}{2} m \bar{v}^2. \quad (3.20)$$

By Theorem 2.5.4 in Section 2.5.4 (Jiang [14], Theorem 2.2),  $\hat{\boldsymbol{\gamma}}$  is the maximizer of  $l_P(\boldsymbol{\gamma})$  over  $\Gamma(M)$ . The penalizer here takes care of the identifiability problem and

when combining  $a$  with  $v_i$ ,  $(\mathbf{X}, \mathbf{Z})$  is a full-column rank matrix we do not need a penalizer term.

Based on Jiang's [14] consistency results ( $\hat{a} \rightarrow_p a_0$ ,  $\hat{\beta} \rightarrow_p \beta_0$ ,  $(1/m) \sum_{i=1}^m (\hat{v}_i - v_{i0})^2 \rightarrow_p (1/N) \sum_{i=1}^m n_i (\hat{v}_i - v_{i0})^2 \rightarrow_p 0$ ) and (J1), (J2) and (J3) in Section 3.2, we tried to establish conditional asymptotic normality for  $(\hat{\beta} - \beta_0)$ , given  $\mathbf{v}_0$ , under certain conditions. Taylor expansion was used to solve  $l_p(\gamma)$  to write the estimators as

$$\begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{\mathbf{a}} - \mathbf{a}_0 \end{pmatrix} = \mathbf{Q}^{-1} \mathbf{g}_N \quad (3.21)$$

where

$$\mathbf{g}_N = \begin{pmatrix} \mathbf{X}^T(\mathbf{y} - \mu) \\ \mathbf{Z}^T(\mathbf{y} - \mu) \end{pmatrix}$$

and

$$\mathbf{Q} = \begin{pmatrix} \mathbf{X}^T \tilde{\mathbf{B}} \mathbf{X} & \mathbf{X}^T \tilde{\mathbf{B}} \mathbf{Z} \\ \mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{X} & \mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z} \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix}. \quad (3.22)$$

Let

$$\mathbf{Q}_0 = \begin{pmatrix} \mathbf{X}^T \mathbf{D} \mathbf{X} & \mathbf{X}^T \mathbf{D} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{D} \mathbf{X} & \mathbf{Z}^T \mathbf{D} \mathbf{Z} \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_{110} & \mathbf{Q}_{120} \\ \mathbf{Q}_{210} & \mathbf{Q}_{220} \end{pmatrix}. \quad (3.23)$$

We will write

$$B_{ij} = b''_{ij} + \frac{1}{2} b^{(3)}(\eta_{ij}^*)(\hat{\eta}_{ij} - \eta_{ij0}) \quad (3.24)$$

and

$$\tilde{\mathbf{B}} = \text{diag}(B_{ij})_{1 \leq i \leq m, 1 \leq j \leq n_i} \quad \mathbf{D} = \text{diag}(b''_{ij}(\eta_{ij0}))_{1 \leq i \leq m, 1 \leq j \leq n_i} \quad (3.25)$$

Recall that  $\mathbf{Z}$  is a block diagonal matrix,  $\mathbf{Z} = \text{diag}(\mathbf{1}_{n_i})_{1 \leq i \leq m}$ , and  $\mathbf{X}$  is  $(\mathbf{X}_1, \dots, \mathbf{X}_s)$  without the first column  $\mathbf{1}_N$ , after combining fixed intercept  $a$  with the random

effects  $v_i$  to obtain  $a_i = a + v_i$ . Note that  $\mathbf{Q}$  is a nonsingular matrix since  $(\mathbf{X}, \mathbf{Z})$  is a full column rank matrix (not all elements of  $\mathbf{X}$  can be equal).

Then

$$\mathbf{Q}^{-1} = \begin{pmatrix} \mathbf{Q}_{11.2}^{-1} & -\mathbf{Q}_{11.2}^{-1}\mathbf{Q}_{12}\mathbf{Q}_{22}^{-1} \\ -\mathbf{Q}_{22}^{-1}\mathbf{Q}_{21}\mathbf{Q}_{11.2}^{-1} & \mathbf{Q}_{22.1}^{-1} \end{pmatrix} \quad (3.26)$$

where

$$\mathbf{Q}_{11.2} = \mathbf{Q}_{11} - \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{21}$$

$$\mathbf{Q}_{22.1} = \mathbf{Q}_{22} - \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}.$$

Since we are only interested in the  $\beta$  terms, we find that

$$\hat{\beta} - \beta_0 = \mathbf{Q}_{11.2}^{-1} \left\{ \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}) - \mathbf{Q}_{12}(\mathbf{Q}_{22})^{-1}\mathbf{Z}^T(\mathbf{y} - \boldsymbol{\mu}) \right\}. \quad (3.27)$$

Let

$$\mathbf{W}^* = \mathbf{X}^T \mathbf{D} \mathbf{X} - (\mathbf{X}^T \mathbf{D} \mathbf{Z})(\mathbf{Z}^T \mathbf{D} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{D} \mathbf{X}. \quad (3.28)$$

Here we choose  $(\mathbf{W}^*)^{-\frac{1}{2}}\mathbf{Q}_{11.20}$  as the normalizer. Then

$$\begin{aligned} & (\mathbf{W}^*)^{-\frac{1}{2}}\mathbf{Q}_{11.20}(\hat{\beta} - \beta_0) \\ &= (\mathbf{W}^*)^{-\frac{1}{2}}\mathbf{Q}_{11.20}\mathbf{Q}_{11.2}^{-1} \left\{ \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}) - \mathbf{Q}_{12}(\mathbf{Q}_{22})^{-1}\mathbf{Z}^T(\mathbf{y} - \boldsymbol{\mu}) \right\} \\ &= I + II + III \end{aligned} \quad (3.29)$$

where

$$I = (\mathbf{W}^*)^{-\frac{1}{2}} \left\{ \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}) - \mathbf{Q}_{120}(\mathbf{Q}_{220})^{-1}\mathbf{Z}^T(\mathbf{y} - \boldsymbol{\mu}) \right\}, \quad (3.30)$$

$$II = (\mathbf{W}^*)^{-\frac{1}{2}} (\mathbf{Q}_{11.20}\mathbf{Q}_{11.2}^{-1} - \mathbf{I}) \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}), \quad (3.31)$$

$$III = (\mathbf{W}^*)^{-\frac{1}{2}} \left\{ \mathbf{Q}_{120}\mathbf{Q}_{220}^{-1} - \mathbf{Q}_{11.20}\mathbf{Q}_{11.2}^{-1}\mathbf{Q}_{12}\mathbf{Q}_{22}^{-1} \right\} \mathbf{Z}^t(\mathbf{y} - \boldsymbol{\mu}). \quad (3.32)$$

We can prove that under regularity conditions, given  $\mathbf{v}_0$ ,  $I \rightarrow_d N(\mathbf{0}, \mathbf{I})$  and  $II \rightarrow_p \mathbf{0}$ .

Consider  $III$ , We can write  $III$  as

$$III = (\mathbf{W}^*)^{-\frac{1}{2}} \left\{ \mathbf{I} - \mathbf{Q}_{11.20} \mathbf{Q}_{11.2}^{-1} \right\} \mathbf{Q}_{120} \mathbf{Q}_{220}^{-1} \mathbf{Z}^t (\mathbf{y} - \boldsymbol{\mu}) \quad (3.33)$$

$$+ (\mathbf{W}^*)^{-\frac{1}{2}} \mathbf{Q}_{11.20} \mathbf{Q}_{11.2}^{-1} \left\{ \mathbf{Q}_{120} \mathbf{Q}_{220}^{-1} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \right\} \mathbf{Z}^t (\mathbf{y} - \boldsymbol{\mu}). \quad (3.34)$$

we can also prove that  $\|(\mathbf{W}^*)^{-\frac{1}{2}}\sqrt{N}\| = O_p(1)$ ,  $\|\mathbf{I} - \mathbf{Q}_{11.20} \mathbf{Q}_{11.2}^{-1}\| = o_p(1)$  and the norm of (3.33) goes to 0. where

$$\mathbf{Q}_{120} \mathbf{Q}_{220}^{-1} \mathbf{Z}^t (\mathbf{y} - \boldsymbol{\mu}) = \left( \sum_i \left\{ \frac{\sum_j x_{ijk} b''_{ij} \sum_j (y_{ij} - b'_{ij})}{\sum_j b''_{ij}} \right\} \right)_{1 \leq k \leq s}$$

and we have

$$\begin{aligned} & (\mathbf{Q}_{120} \mathbf{Q}_{220}^{-1} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1}) \mathbf{Z}^t (\mathbf{y} - \boldsymbol{\mu}) \\ &= \left( \sum_i \left\{ \frac{\sum_j x_{ijk} b''_{ij}}{\sum_j b''_{ij}} - \frac{\sum_j x_{ijk} b''_{ij} (\eta_{ij}^*)}{\sum_j b''_{ij} (\eta_{ij}^*)} \right\} \sum_j (y_{ij} - b'_{ij}) \right)_{1 \leq k \leq s} \end{aligned}$$

Consider the norm of (3.34)

$$\begin{aligned} & \|(\mathbf{W}^*)^{-\frac{1}{2}} \mathbf{Q}_{11.20} \mathbf{Q}_{11.2}^{-1} \left\{ \mathbf{Q}_{120} \mathbf{Q}_{220}^{-1} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \right\} \mathbf{Z}^t (\mathbf{y} - \boldsymbol{\mu})\| \\ & \leq \|\mathbf{Q}_{11.20} \mathbf{Q}_{11.2}^{-1}\| \|(\mathbf{W}^*)^{-1/2}\| \| \left\{ \mathbf{Q}_{120} \mathbf{Q}_{220}^{-1} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \right\} \mathbf{Z}^t (\mathbf{y} - \boldsymbol{\mu}) \| \\ & \leq \frac{O_p(1)}{\sqrt{N}} \| \left\{ \mathbf{Q}_{120} \mathbf{Q}_{220}^{-1} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \right\} \mathbf{Z}^t (\mathbf{y} - \boldsymbol{\mu}) \| \\ & = O_p(1) \left( \sum_k \left\{ \sum_{i=1}^m n_i \left\{ \frac{\sum_j x_{ijk} b''_{ij}}{\sum_j b''_{ij}} - \frac{\sum_j x_{ijk} b''_{ij} (\eta_{ij}^*)}{\sum_j b''_{ij} (\eta_{ij}^*)} \right\} \frac{1}{n_i} \sum_j (y_{ij} - b'_{ij}) \right\}^2 \right)^{\frac{1}{2}} \end{aligned} \quad (3.35)$$

We conjecture  $1/n_i \sum_j (y_{ij} - b'_{ij}) = O_p(1/\sqrt{n_i})$  and that  $\hat{\eta}_{ij} - \eta_{ij0} = O_p(1/\sqrt{n_i})$ . If so, inside of square of (3.35) would be  $O_p(m/\sqrt{N}) \leq O_p(\sqrt{m/\min_i n_i})$ . However, we are unable to calculate the convergence rate of  $(\hat{\eta}_{ij} - \eta_{ij0})$  because the number

of random effects goes to infinity. Li, Lindsay and Waterman [18] examined the asymptotic behavior of MLE's of a scalar  $\beta$  as the number of nuisance parameters goes to infinity. They found that under rectangular asymptotics ( $m = cn$ ,  $c$  fixed,  $m \rightarrow \infty$ )  $\sqrt{N}(\hat{\beta}_1 - \beta_0) \rightarrow_d N(\tau, \mathcal{I}^{-1}(\beta))$  when  $\tau$  is possibly nonzero and  $\mathcal{I}^{-1}(\beta)$  is the Fisher information. They proposed an alternative estimator (based on projected scores) which satisfied  $\sqrt{N}(\hat{\beta}_2 - \beta_0) \rightarrow_d N(0, \mathcal{I}^{-1}(\beta))$ . In the Chapter 6 and Appendix A simulation studies, we find that in Logistic and Poisson random intercept with combined fixed intercept  $a$  and random effects  $v_i$ , consistency and conditional asymptotic normality results appear to hold.

### 3.4 Proofs.

#### 3.4.1 Proof of Lemma 3.1.1

Recall the definition of  $\mathbf{W}_{bi}$  in (3.3). We have

$$\mathbf{u}^T \mathbf{W}_{bi} \mathbf{u} = \sum_{j=1}^{n_i} b_{ij}'' \left( \sum_{k=1}^s u_k \left\{ x_{ijk} - \frac{\sum_{l=1}^{n_i} x_{ilk} b_{il}''}{\sum_{l=1}^{n_i} b_{il}''} \right\} \right)^2.$$

Let

$$\tilde{\mathbf{W}}_i = \text{diag}(\pi_i) - \pi_i \pi_i^T$$

where

$$\pi_i = \begin{bmatrix} b_{i1}'' / \sum_j b_{ij}'' \\ \vdots \\ b_{in_i}'' / \sum_j b_{ij}'' \end{bmatrix}_{n_i \times 1}.$$

Then

$$\begin{aligned}\frac{1}{n_i} \mathbf{u}^T \mathbf{W}_{bi} \mathbf{u} &= \frac{1}{n_i} \mathbf{u}^T \sum_j b_{ij}'' \mathbf{X}_i^T \tilde{\mathbf{W}}_i \mathbf{X}_i \mathbf{u} \\ &= \frac{1}{n_i} \mathbf{u}^T \sum_j b_{ij}'' (\mathbf{X}_i - \bar{\mathbf{X}}_i)^T \tilde{\mathbf{W}}_i (\mathbf{X}_i - \bar{\mathbf{X}}_i) \mathbf{u}\end{aligned}$$

where

$$\mathbf{X}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{is}),$$

$$\mathbf{X}_{ik} = \begin{bmatrix} x_{i1k} \\ \vdots \\ x_{in_i k} \end{bmatrix}_{n_i \times 1},$$

$$\bar{\mathbf{X}}_i = (\bar{x}_{i1} \mathbf{1}_{n_i}, \dots, \bar{x}_{is} \mathbf{1}_{n_i}),$$

and  $\bar{x}_{ik} = n_i^{-1} \sum_j x_{ijk}$ .

Therefore

$$\begin{aligned}&\inf_{\|\mathbf{u}\|=1} \frac{\mathbf{u}^T \left( \sum_j b_{ij}'' (\mathbf{X}_i - \bar{\mathbf{X}}_i)^T \tilde{\mathbf{W}}_i (\mathbf{X}_i - \bar{\mathbf{X}}_i) \right) \mathbf{u}}{n_i} \\ &= \inf_{\|\mathbf{u}\|=1} \frac{\sum_j b_{ij}'' \left( \mathbf{u}^T (\mathbf{X}_i - \bar{\mathbf{X}}_i)^T \tilde{\mathbf{W}}_i (\mathbf{X}_i - \bar{\mathbf{X}}_i) \mathbf{u} \right)}{n_i} \\ &\geq \inf_{\|\mathbf{u}\|=1} \frac{\sum_j b_{ij}'' \lambda_2(\tilde{\mathbf{W}}_i) (\mathbf{u}^T (\mathbf{X}_i - \bar{\mathbf{X}}_i)^T (\mathbf{X}_i - \bar{\mathbf{X}}_i) \mathbf{u})}{n_i}.\end{aligned}$$

The bound above  $\lambda_2(\tilde{\mathbf{W}}_i)$  is the second smallest eigenvalue of  $\tilde{\mathbf{W}}_i$ , because  $(\mathbf{X}_i - \bar{\mathbf{X}}_i) \mathbf{u}$  is orthogonal to  $\mathbf{1}_{n_i}$  which is the eigenvector corresponding to the unique eigenvalue 0.

Here we use the definition of the second smallest eigenvalue of  $\mathbf{W}_i$ :

$$\begin{aligned}
& \frac{\mathbf{u}^T(\mathbf{X}_i - \bar{\mathbf{X}}_i)^T \tilde{\mathbf{W}}_i (\mathbf{X}_i - \bar{\mathbf{X}}_i) \mathbf{u}}{\mathbf{u}^T(\mathbf{X}_i - \bar{\mathbf{X}}_i)^T (\mathbf{X}_i - \bar{\mathbf{X}}_i) \mathbf{u}} \\
& \geq \inf_{\|\mathbf{u}\|=1} \frac{\mathbf{u}^T(\mathbf{X}_i - \bar{\mathbf{X}}_i)^T \tilde{\mathbf{W}}_i (\mathbf{X}_i - \bar{\mathbf{X}}_i) \mathbf{u}}{\mathbf{u}^T(\mathbf{X}_i - \bar{\mathbf{X}}_i)^T (\mathbf{X}_i - \bar{\mathbf{X}}_i) \mathbf{u}} = \lambda_2(\tilde{\mathbf{W}}_i).
\end{aligned}$$

Now we need to find a lower bound of  $\lambda_2(\tilde{\mathbf{W}}_i)$ . Stewart's [31] Theorem 5.1 states the following result:

Let  $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n$ ,  $\beta_1 \leq \beta_2 \leq \dots \leq \beta_n$  and  $\gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_n$  be the eigenvalues of the real Hermitian matrices  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C} = \mathbf{A} + \mathbf{B}$ .

Then

$$\alpha_i + \beta_1 \leq \gamma_i \leq \alpha_i + \beta_n.$$

In our case we set  $\mathbf{A} = -\pi_i \pi_i^T$  and  $\mathbf{B} = \text{diag}(\pi_i)$ , and by the special structure of  $\mathbf{A}$  we have  $\alpha_2 = \dots = \alpha_{n_i} = 0$  and  $\alpha_1 = -\pi_i^T \pi_i = -\sum_j \pi_j^2$ .

Then

$$\min_j \frac{b_{ij}''}{\sum_j b_{ij}''} \leq \lambda_2(\tilde{\mathbf{W}}_i). \quad (3.36)$$

Consider

$$\begin{aligned}
& \inf_{\|\mathbf{u}\|=1} \frac{\mathbf{u}^T \mathbf{W}_{bi} \mathbf{u}}{n_i} \\
& = \inf_{\|\mathbf{u}\|=1} \frac{\mathbf{u}^T \{ \sum_j b_{ij}'' (\eta_{ij0}) (\mathbf{X}_i - \bar{\mathbf{X}}_i)^T \tilde{\mathbf{W}}_i (\mathbf{X}_i - \bar{\mathbf{X}}_i) \} \mathbf{u}}{n_i} \\
& \geq \inf_{\|\mathbf{u}\|=1} \frac{\mathbf{u}^T \{ \sum_j b_{ij}'' (\eta_{ij0}) (\mathbf{X}_i - \bar{\mathbf{X}}_i)^T \lambda_2(\tilde{\mathbf{W}}_i) (\mathbf{X}_i - \bar{\mathbf{X}}_i) \} \mathbf{u}}{n_i} \\
& \geq \inf_{\|\mathbf{u}\|=1} \frac{\mathbf{u}^T \{ \min_j b_{ij}'' (\eta_{ij0}) (\mathbf{X}_i - \bar{\mathbf{X}}_i)^T (\mathbf{X}_i - \bar{\mathbf{X}}_i) \} \mathbf{u}}{n_i} > \frac{1}{c_2}.
\end{aligned}$$

where  $c_2$  is a suitably large constant. Equivalently  $\mathbf{W}_{bi}$  is a positive definite matrix.

### 3.4.2 Proof of Theorem 3.1.2

It is easy to prove that of  $\hat{\beta}_w - \beta_0$ , where  $\hat{\beta}_w = \sum_i \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-1} (\mathbf{F}_i^{\beta\beta})^{-1} \hat{\beta}_i$ ,

then  $\hat{\beta}_w \rightarrow_p \beta_0$ . by using the result of Lemma 3.1.1 and conditions of Theorem 3.1.2

because we have

$$\begin{aligned} & \| \hat{\beta}_w - \beta_0 \| \\ & \leq \sum_i \left\| \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-1} \right\| \| (\mathbf{F}_i^{\beta\beta})^{-1/2} \| \| (\mathbf{F}_i^{\beta\beta})^{-1/2} (\hat{\beta}_i - \beta_0) \| \\ & \leq \frac{O_p(1)}{K_1 \min_i \sqrt{\lambda_{\max}(\mathbf{F}_i^{\beta\beta})^{-1}}} \leq \frac{O_p(1)/K_1}{\min_i \sqrt{n_i} \sqrt{b_i} \liminf(\lambda_{n_i}/n_i)} = o_p(1). \end{aligned}$$

Here we use the fact that  $(\mathbf{F}_i^{\beta\beta})^{-1/2} (\hat{\beta}_i - \beta_0)$  is  $O_p(1)$ , as proved by Fahrmeir and Kaufmann [11].

Since

$$\begin{aligned} & \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} (\hat{\beta}_i - \beta_0) \\ & = \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} (\hat{\beta}_i - E(\hat{\beta}_i | v_{i0}) + E(\hat{\beta}_i | v_{i0}) - \beta_0), \end{aligned}$$

let  $\mathbf{u}$  be a unit vector and

$$T_i = \mathbf{u}^t \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} (\mathbf{F}_i^{\beta\beta})^{-1} (\hat{\beta}_i - E(\hat{\beta}_i | v_{i0})) \quad (3.37)$$

$$= \mathbf{u}^t \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} (\mathbf{F}_i^{\beta\beta})^{-\frac{1}{2}} (\mathbf{F}_i^{\beta\beta})^{-\frac{1}{2}} (\hat{\beta}_i - E(\hat{\beta}_i | v_{i0})). \quad (3.38)$$

Recall  $\hat{\beta}_i^* = (\mathbf{F}_i^{\beta\beta})^{-\frac{1}{2}} (\hat{\beta}_i - E(\hat{\beta}_i | v_{i0}))$ . By the Cauchy-Schwartz inequality we have

$$T_i^2 \leq \left\{ \mathbf{u}^t \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} (\mathbf{F}_i^{\beta\beta})^{-1} \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \mathbf{u} \right\} (\hat{\beta}_i^*)^t \hat{\beta}_i^* \quad (3.39)$$

$$\leq \lambda_{\max} \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-1} \lambda_{\max} (\mathbf{F}_i^{\beta\beta})^{-1} (\hat{\beta}_i^*)^t \hat{\beta}_i^* \quad (3.40)$$

By Lemma 3.1.1 we have  $n_i \|\mathbf{W}_{bi}^{-1}\| = n_i \lambda_{\max} \mathbf{F}_i^{\beta\beta} = O(1)$ , and because of boundedness of  $\sum_k x_{ijk}^2$  and  $b''_{ij}$

$$\frac{1}{n_i} \mathbf{u}^t \mathbf{W}_{bi} \mathbf{u} = \frac{1}{n_i} \sum_j b''_{ij} \left\{ \sum_k u_k (x_{ijk} - \frac{\sum_j x_{ijk} b''_{ij}}{\sum_j b''_{ij}}) \right\}^2 = O(1).$$

Equivalently we have  $(\lambda_{\min} \mathbf{F}_i^{\beta\beta})^{-1} = \|\mathbf{W}_{bi}\| = O_p(n_i)$  and (1) of Theorem 3.1.3 follows. Now we check the Lindeberg condition:

$$\begin{aligned} \sum_{i=1}^m E(T_i^2 I(|T_i| > \varepsilon) | v_{i0}) &= \sum_{i=1}^m E(T_i^2 I(T_i^2 > \varepsilon^2) | v_{i0}) \\ &\leq \left( \sum_i \lambda_{\max}^{-1} \mathbf{F}_i^{\beta\beta} \min_i \lambda_{\min} \mathbf{F}_i^{\beta\beta} \right)^{-1} \\ &\quad \times \sum_{i=1}^m E \left[ (\hat{\boldsymbol{\beta}}_i^*)^t \hat{\boldsymbol{\beta}}_i^* I \left( (\hat{\boldsymbol{\beta}}_i^*)^t \hat{\boldsymbol{\beta}}_i^* > \varepsilon^2 \lambda_{\min}(\mathbf{F}_i^{\beta\beta}) \lambda_{\min} \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right) \right) | v_{i0} \right] \\ &\leq \frac{\max_i \lambda_{\max} \mathbf{F}_i^{\beta\beta}}{m \min_i \lambda_{\min} \mathbf{F}_i^{\beta\beta}} \sum_{i=1}^m E[(\hat{\boldsymbol{\beta}}_i^*)^t \hat{\boldsymbol{\beta}}_i^* \psi((\hat{\boldsymbol{\beta}}_i^*)^t \hat{\boldsymbol{\beta}}_i^*) | v_{i0}] \\ &\quad \times \left\{ \psi \left( \varepsilon^2 \lambda_{\min}(\mathbf{F}_i^{\beta\beta}) \left( \sum_i \lambda_{\max}^{-1} \mathbf{F}_i^{\beta\beta} \right) \right) \right\}^{-1} \\ &\leq O_p(1) \frac{\max_i \lambda_{\max} \mathbf{F}_i^{\beta\beta}}{\min_i \lambda_{\min} \mathbf{F}_i^{\beta\beta}} \left\{ \psi \left( \frac{m \min_i \lambda_{\min} \mathbf{F}_i^{\beta\beta}}{\max_i \lambda_{\max} \mathbf{F}_i^{\beta\beta}} \right) \right\}^{-1} = o_p(1) \end{aligned}$$

since by condition (1) and (4), the argument of  $\psi$  goes to  $\infty$ .

According to equation (3.2) we have

$$l'_n(\hat{\boldsymbol{\gamma}}_i) = l'_n(\boldsymbol{\gamma}_{i0}) - \mathbf{F}_{n_i}(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i0}) + (\mathbf{F}_{n_i} - \mathbf{F}_{n_i}(\boldsymbol{\gamma}^*))(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i0})$$

where  $\boldsymbol{\gamma}_i^*$  is between  $\boldsymbol{\gamma}_{i0}$  and  $\hat{\boldsymbol{\gamma}}_i$ .

Taking the expectation on both sides and conditioning on  $v_{i0}$  we have

$$\begin{aligned} E[(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i0}) | v_{i0}] &= (\mathbf{F}_{n_i})^{-1} E \left[ (\mathbf{F}_{n_i} - \mathbf{F}_{n_i}(\boldsymbol{\gamma}^*)) \mathbf{F}_{n_i}^{-T/2} \mathbf{F}_{n_i}^{T/2} (\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i0}) | v_{i0} \right] \\ &= (\mathbf{F}_{n_i})^{-1} E \left[ (1/\sqrt{n_i}) (\mathbf{F}_{n_i} - \mathbf{F}_{n_i}(\boldsymbol{\gamma}^*)) \sqrt{n_i} \mathbf{F}_{n_i}^{-T/2} \mathbf{F}_{n_i}^{T/2} (\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{i0}) | v_{i0} \right] \end{aligned}$$

Let  $\mathbf{u}$  be a unit vector and consider

$$\begin{aligned}
& (1/\sqrt{n_i}) \|\mathbf{F}_{n_i} - \mathbf{F}_{n_i}(\boldsymbol{\gamma}_i^*)\| \\
&= (1/\sqrt{n_i}) \sup_{\|\mathbf{u}=1\|} \left| \sum_j (b''_{ij} - b''_{ij}(\eta_{ij}^*)) \left( u_1 + \sum_{k=1}^s u_{k+1} x_{ijk} \right)^2 \right| \\
&\leq (2sK' + 2)(1/n_i) \sum_j \sqrt{n_i} |b''_{ij} - b''_{ij}(\eta_{ij}^*)| \\
&\leq (2sK' + 2)(1/2n_i) \sum_j \sqrt{n_i} |b_{ij}^{(3)}(\tilde{\eta}_{ij})| |\eta_{ij} - \hat{\eta}_{ij}| \tag{3.41}
\end{aligned}$$

$$= O_p(1). \tag{3.42}$$

Since we have  $b_{ij}^{(3)}(\tilde{\eta}_{ij})$  bounded and according to Fahrmeir and Kaufmann [11], we have, given  $v_{i0}$ ,

$$\sqrt{n_i}(\hat{\eta}_{ij} - \eta_{ij0}) \sim AN(0, n_i(\mathbf{F}_i^{aa} + \mathbf{F}_i^{a\beta} \mathbf{x}_{ij} + \mathbf{x}_{ij}^t \mathbf{F}_i^{\beta a} + \mathbf{x}_{ij}^t \mathbf{F}_i^{\beta\beta} \mathbf{x}_{ij})). \tag{3.43}$$

In addition, by the assumptions of Theorem 3.1.3 we can get from (3.41) to (3.42) and  $\sqrt{n_i} \|(\mathbf{F}_{n_i})^{-T/2}\| = O_p(1)$ .

Then

$$\|E[(\hat{\beta}_i - \beta_0)|v_{i0}]\| \leq \|E[(\hat{\gamma}_i - \gamma_{i0})|v_{i0}]\| \leq \|(\mathbf{F}_{n_i})^{-1}\| \delta O_p(1).$$

Now we consider the bias term

$$\left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} (E(\hat{\beta}_i|v_{i0}) - \beta_0). \tag{3.44}$$

Then

$$\begin{aligned}
& \left( \mathbf{u}^t \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} (E(\hat{\boldsymbol{\beta}}_i | v_{i0}) - \boldsymbol{\beta}_0) \right)^2 \\
& \leq m \sum_{i=1}^m \left( \mathbf{u}^t \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} (\mathbf{F}_i^{\beta\beta})^{-1} (E(\hat{\boldsymbol{\beta}}_i | v_{i0}) - \boldsymbol{\beta}_0) \right)^2 \\
& \leq m \sum_{i=1}^m \mathbf{u}^t \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} (\mathbf{F}_i^{\beta\beta})^{-1} \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \mathbf{u} \\
& \quad \times (E(\hat{\boldsymbol{\beta}}_i | v_{i0}) - \boldsymbol{\beta}_0)^t (\mathbf{F}_i^{\beta\beta})^{-1} (E(\hat{\boldsymbol{\beta}}_i | v_{i0}) - \boldsymbol{\beta}_0) \\
& \leq m \left\{ \max_i \lambda_{\max}(\mathbf{F}_i^{\beta\beta})^{-1} \lambda_{\max} \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-1} \right\} \\
& \quad \times \sum_{i=1}^m (E(\hat{\boldsymbol{\beta}}_i | v_{i0}) - \boldsymbol{\beta}_0)^t (\mathbf{F}_i^{\beta\beta})^{-1} (E(\hat{\boldsymbol{\beta}}_i | v_{i0}) - \boldsymbol{\beta}_0) \\
& \leq \frac{\max_i \lambda_{\max} \mathbf{F}_i^{\beta\beta}}{\min_i \lambda_{\min} \mathbf{F}_i^{\beta\beta}} \sum_{i=1}^m (E(\hat{\boldsymbol{\beta}}_i | v_{i0}) - \boldsymbol{\beta}_0)^t (\mathbf{F}_i^{\beta\beta})^{-1} (E(\hat{\boldsymbol{\beta}}_i | v_{i0}) - \boldsymbol{\beta}_0) \\
& \leq O_p(1) \sum_{i=1}^m \|E(\hat{\boldsymbol{\beta}}_i | v_{i0}) - \boldsymbol{\beta}_0\|^2 \|(\mathbf{F}_i^{\beta\beta})^{-1}\| \\
& \leq \delta^2 O_p(1) \sum_{i=1}^m \|(\mathbf{F}_{n_i})^{-1}\|^2 \|(\mathbf{F}_i^{\beta\beta})^{-1}\| \tag{3.45}
\end{aligned}$$

$$\leq \delta^2 O_p(1) \max_i \frac{\lambda_{\max}(\mathbf{F}_{n_i})^{-1}}{\lambda_{\min}(\mathbf{F}_{n_i})^{-1}} \sum_{i=1}^m \|(\mathbf{F}_{n_i})^{-1}\| \tag{3.46}$$

$$= o_p(1)$$

The step from (3.45) to (3.46) follows Schott's [29] Theorem 3.20:

If  $\mathbf{A}$  is a  $m \times m$  symmetric matrix and  $\mathbf{A}_k$  is a leading  $k \times k$  principal submatrix we have the following inequality:

$$\lambda_{m-i+1}(\mathbf{A}) \leq \lambda_{k-i+1}(\mathbf{A}_k) \leq \lambda_{k-i+1}(\mathbf{A})$$

where  $i = 1, \dots, k$  and  $\lambda_1$  is the maximum eigenvalue of  $\mathbf{A}$ . Here we chose  $\mathbf{A} = (\mathbf{F}_{n_i})^{-1}$  and  $\mathbf{A}_k = \mathbf{F}_i^{\beta\beta}$ .

By the Central Limit Theorem and Slutsky's Theorem, we have, given  $\mathbf{v}_0$ ,

$$\left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) \rightarrow_d N(\mathbf{0}, \mathbf{I}).$$

### 3.4.3 Proof of Theorem 3.1.3

If we let  $\mathbf{V} = \sum_i (\mathbf{F}_i^{\beta\beta})^{-1}$  and  $\hat{\mathbf{V}} = \sum_i (\hat{\mathbf{F}}_i^{\beta\beta})^{-1}$ , then

$$\begin{aligned} & \| \hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_0 \| \\ & \leq \| \hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_w \| + \| \hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_0 \| \\ & = \left\| \sum_{i=1}^m \left( \hat{\mathbf{V}}^{-1} \left\{ [(\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \mathbf{F}_i^{\beta\beta} - \mathbf{I}] \mathbf{V} + \mathbf{V} - \hat{\mathbf{V}} \right\} \right) \mathbf{w}_i (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) \right\| + o_p(1) \\ & \leq \sum_i \left\| \left( \hat{\mathbf{V}}^{-1} \left\{ [(\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \mathbf{F}_i^{\beta\beta} - \mathbf{I}] \mathbf{V} + \mathbf{V} - \hat{\mathbf{V}} \right\} \right) \right\| \| \mathbf{w}_i (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) \| + o_p(1) \\ & = o_p(1) \end{aligned}$$

Since we can prove

$$\begin{aligned} & \left\| \left( \hat{\mathbf{V}}^{-1} \left\{ [(\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \mathbf{F}_i^{\beta\beta} - \mathbf{I}] \mathbf{V} + \mathbf{V} - \hat{\mathbf{V}} \right\} \right) \right\| \\ & \leq \| \hat{\mathbf{V}}^{-1} \| \| [(\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \mathbf{F}_i^{\beta\beta} - \mathbf{I}] \| \| \mathbf{V} \| + \| \hat{\mathbf{V}}^{-1} \| \| \mathbf{V} - \hat{\mathbf{V}} \| \end{aligned} \quad (3.47)$$

$$= o_p(1) \quad (3.48)$$

Consider

$$\begin{aligned} & \left\| \sum_{i=1}^m \left[ \left( \sum_i (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} - \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-\frac{1}{2}} (\mathbf{F}_i^{\beta\beta})^{-1} \right] (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) \right\| \\ & = \left\| \sum_{i=1}^m \left( \hat{\mathbf{V}}^{-\frac{1}{2}} (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} - \mathbf{V}^{-\frac{1}{2}} (\mathbf{F}_i^{\beta\beta})^{-1} \right) (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) \right\| \\ & = \left\| \sum_{i=1}^m \mathbf{V}^{-\frac{1}{2}} \left( \mathbf{V}^{\frac{1}{2}} \hat{\mathbf{V}}^{-\frac{1}{2}} (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} - (\mathbf{F}_i^{\beta\beta})^{-1} \right) (\mathbf{F}_i^{\beta\beta})^{\frac{1}{2}} (\mathbf{F}_i^{\beta\beta})^{-\frac{1}{2}} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) \right\| \\ & \leq \sum_{i=1}^m \| \mathbf{V}^{-\frac{1}{2}} \| \| \mathbf{V}^{\frac{1}{2}} \hat{\mathbf{V}}^{-\frac{1}{2}} (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} - (\mathbf{F}_i^{\beta\beta})^{-1} \| \| (\mathbf{F}_i^{\beta\beta})^{\frac{1}{2}} \| \| (\mathbf{F}_i^{\beta\beta})^{-\frac{1}{2}} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) \| \\ & \leq I_1 + I_2 \end{aligned}$$

where

$$\begin{aligned} I_1 &= \sum_{i=1}^m \|\mathbf{V}^{-\frac{1}{2}}\| \|\mathbf{V}^{\frac{1}{2}} \hat{\mathbf{V}}^{-\frac{1}{2}} - \mathbf{I}\| \|(\hat{\mathbf{F}}_i^{\beta\beta})^{-1}\| \|(\mathbf{F}_i^{\beta\beta})^{\frac{1}{2}}\| \|(\mathbf{F}_i^{\beta\beta})^{-\frac{1}{2}} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0)\| \\ I_2 &= \sum_{i=1}^m \|\mathbf{V}^{-\frac{1}{2}}\| \|(\hat{\mathbf{F}}_i^{\beta\beta})^{-1} - (\mathbf{F}_i^{\beta\beta})^{-1}\| \|(\mathbf{F}_i^{\beta\beta})^{\frac{1}{2}}\| \|(\mathbf{F}_i^{\beta\beta})^{-\frac{1}{2}} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0)\| \end{aligned}$$

Consider

$$\|\mathbf{V}^{\frac{1}{2}} \hat{\mathbf{V}}^{-\frac{1}{2}} - \mathbf{I}\| = \|(\mathbf{V}^{\frac{1}{2}} - \hat{\mathbf{V}}^{\frac{1}{2}}) \hat{\mathbf{V}}^{-\frac{1}{2}}\| \quad (3.49)$$

$$\leq \|\mathbf{V} - \hat{\mathbf{V}}\|^{\frac{1}{2}} \|\hat{\mathbf{V}}^{-\frac{1}{2}}\| \quad (3.50)$$

The step from (3.49) to (3.50) follows from Theorem X.1.1 of Bhatia's [3]. A function  $f$  is said to be matrix monotone of order  $n$  if it is monotone with respect to this order  $n \times n$  Hermitian matrices, i.e., if  $\mathbf{A} \leq \mathbf{B}$  implies  $f(\mathbf{A}) \leq f(\mathbf{B})$ . If  $f$  is matrix monotone of order  $n$  for all  $n$  then we say  $f$  is matrix monotone or operator monotone.

Then the following theorem holds:

Let  $f$  be an operator monotone function on  $[0, \infty]$  such that  $f(0) = 0$ . Then for all positive operators  $A, B$ ,

$$\|f(A) - f(B)\| \leq f(\|A - B\|).$$

Here our operator monotone function is the square root function.

In addition, we have

$$\begin{aligned}
\frac{1}{N} \|\mathbf{V} - \hat{\mathbf{V}}\| &= \frac{1}{N} \left\| \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} - \sum_i (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \right\| \\
&\leq \frac{1}{N} \sum_i \|(\mathbf{F}_i^{\beta\beta})^{-1} - (\hat{\mathbf{F}}_i^{\beta\beta})^{-1}\| \\
&= \frac{1}{N} \sum_i n_i \left\| \frac{1}{n_i} (\mathbf{F}_i^{\beta\beta})^{-1} - \frac{1}{n_i} (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \right\|.
\end{aligned}$$

Since here we have the assumption that  $|v_{i0}|$  is bounded, the assumption of Theorem 3.1.3 and the result of Lemma 3.1.2, considering condition (3) of Theorem 3.1.4 we have

$$\begin{aligned}
N \|\mathbf{V}^{-\frac{1}{2}}\|^2 &= N \left\| \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-1} \right\| = N \left( \lambda_{\min} \left( \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} \right) \right)^{-1} \\
&\leq N \left( \sum_i \lambda_{\min} (\mathbf{F}_i^{\beta\beta})^{-1} \right)^{-1} = N \left( \sum_i n_i \frac{1}{n_i} \lambda_{\max}^{-1} \mathbf{F}_i^{\beta\beta} \right)^{-1} \\
&= O(1).
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n_i} \left( (\mathbf{F}_i^{\beta\beta})^{-1} - (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \right) \\
&= \frac{1}{n_i} \left( \sum_j x_{ijk} x_{ijl} (b''_{ij} - b''(\hat{\eta}_{ij})) + \frac{(\sum_j x_{ijk} b''_{ij})(\sum_j x_{ijl} b''_{ij})}{\sum_j b''_{ij}} \right. \\
&\quad \left. - \frac{(\sum_j x_{ijk} b''_{ij}(\hat{\eta}_{ij}))(\sum_j x_{ijl} b''_{ij}(\hat{\eta}_{ij}))}{\sum_j b''_{ij}(\hat{\eta}_{ij})} \right)_{1 \leq k \leq s, 1 \leq l \leq s}
\end{aligned}$$

Therefore,

$$\frac{1}{n_i} \|((\mathbf{F}_i^{\beta\beta})^{-1} - (\hat{\mathbf{F}}_i^{\beta\beta})^{-1})\| \leq II_1 + II_2$$

where

$$II_1 = \frac{1}{n_i} \sup_{\|\mathbf{u}\|=1} \left| \sum_j (b''_{ij}(\hat{\eta}_{ij}) - b''_{ij})(\sum_k u_k x_{ijk})^2 \right| \quad (3.51)$$

$$II_2 = \frac{1}{n_i} \sup_{\|\mathbf{u}\|=1} \left| \left( \frac{(\sum_k u_k \sum_j x_{ijk} b''_{ij})^2}{\sum_j b''_{ij}} - \frac{(\sum_k u_k \sum_j x_{ijk} b''_{ij}(\hat{\eta}_{ij}))^2}{\sum_j b''_{ij}(\hat{\eta}_{ij})} \right) \right|. \quad (3.52)$$

Let  $r_{ij} = b''_{ij}(\hat{\eta}_{ij}) - b''_{ij}(\eta_{ij0})$  and consider

$$\begin{aligned} & \left| \frac{(\sum_k u_k \sum_j x_{ijk} b''_{ij})^2}{\sum_j b''_{ij}} - \frac{(\sum_k u_k \sum_j x_{ijk} b''_{ij}(\hat{\eta}_{ij}))^2}{\sum_j b''_{ij}(\hat{\eta}_{ij})} \right| \\ &= \left| \frac{\sum_j b''_{ij}(\hat{\eta}_{ij})(\sum_k u_k \sum_j x_{ijk} b''_{ij})^2 - \sum_j b''_{ij}(\sum_k u_k \sum_j x_{ijk} b''_{ij}(\hat{\eta}_{ij}))^2}{\sum_j b''_{ij} \sum_j b''_{ij}(\hat{\eta}_{ij})} \right| \\ &\leq \left| \frac{(\sum_k u_k \sum_j x_{ijk} r_{ij})^2}{\sum_j b''_{ij}} \right| + \left| \frac{2(\sum_k u_k \sum_j x_{ijk} b''_{ij}(\hat{\eta}_{ij}))(\sum_k u_k \sum_j x_{ijk} r_{ij})}{\sum_j b''_{ij}} \right| \\ &\quad + \left| \frac{\sum_j r_{ij} (\sum_k u_k \sum_j x_{ijk} b''_{ij}(\hat{\eta}_{ij}))^2}{\sum_j b''_{ij} \sum_j b''_{ij}(\hat{\eta}_{ij})} \right|, \end{aligned}$$

By the Cauchy-Schwartz inequality

$$\begin{aligned} II_1 &\leq \frac{1}{n_i} \sup_{\|\mathbf{u}\|=1} \left| \sum_j (b''_{ij}(\hat{\eta}_{ij}) - b''_{ij}) \sum_k u_k^2 \sum_k x_{ijk}^2 \right| \\ &\leq \frac{K'^2}{n_i} \sum_j |\frac{1}{2} r_{ij}|, \end{aligned}$$

and

$$\begin{aligned} II_2 &\leq \frac{1}{n_i} \frac{(\sum_k \sum_j x_{ijk}^2)(\sum_j r_{ij}^2)}{\sum_j b''_{ij}} + \frac{1}{n_i} \frac{(\sum_j |r_{ij}|) \sum_k (\sum_j x_{ijk} b''_{ij}(\hat{\eta}_{ij}))^2}{\sum_j b''_{ij}(\hat{\eta}_{ij}) \sum_j b''_{ij}} \\ &\quad + \frac{1}{n_i} \sup_{\|\mathbf{u}\|=1} \frac{(\sum_k |u_k| \sum_j |x_{ijk}| b''_{ij}(\hat{\eta}_{ij})) (\sum_k |u_k| \sum_j |x_{ijk}| |r_{ij}|)}{\sum_j b''_{ij}} \\ &\leq \left( \frac{K'^2 + 2sK'^2}{4b_l} \right) \frac{1}{n_i} \sum_j |\frac{1}{2} r_{ij}| + \left( \frac{K'^2}{b_l} \right) \frac{1}{n_i} \sum_j \left( \frac{1}{2} r_{ij} \right)^2. \end{aligned}$$

Since  $\sum_{k=1}^s x_{ijk}^2 < K'$  where  $K'$  is a positive constant and according to Fahrmeir and Kaufmann [11], given  $v_{i0}$ , we have (3.43).

By the delta method and assumptions of Theorem 3.1.3, we obtain

$$\left\| \frac{1}{n_i} \left( (\mathbf{F}_i^{\beta\beta})^{-1} - (\hat{\mathbf{F}}_i^{\beta\beta})^{-1} \right) \right\| = O_p(1/\sqrt{n_i}).$$

By the assumption of Theorem 3.1.2, we have,  $(1/N)\|\mathbf{V} - \hat{\mathbf{V}}\| \leq O_p(1)/N \sum_i \sqrt{n_i} c_i = o_p(1)$  and conditions (1) and (2) of Theorem 3.1.3 are satisfied.

We can prove  $(1/N)\|\mathbf{V} - \hat{\mathbf{V}}\| = o_p(1)$  and  $N\|\mathbf{V}^{-1}\| = O(1)$  which leads the result  $N\|\hat{\mathbf{V}}^{-1}\| = O_p(1)$ . Also it is easy to get  $(\sqrt{n_i}/N)\|\mathbf{V} - \hat{\mathbf{V}}\| = O_p(1)$  and  $(1/n_i)\|(\hat{\mathbf{F}}_i^{\beta\beta})^{-1}\| = O_p(1)$ , then we can get from (3.47) to (3.48) which leads the conclusion that  $\hat{\boldsymbol{\beta}}_w$  is consistent. By (2) of Theorem 3.1.2, we have  $I_1 \rightarrow_p 0$  and  $I_2 \rightarrow_p 0$ . The result of Theorem 3.1.3 follows.

The similar argument follows for replacing  $\hat{\mathbf{F}}_i^{\beta\beta}$  by  $\hat{\mathbf{F}}_{\mathbf{i}}^{\beta\beta}$  since by the conclusion of Theorem 3.1.3, we have, according to Fahrmeir and Kaufmann [11] and Slutsky's Theorem, given  $v_{i0}$ ,

$$\sqrt{n_i}(\hat{\eta}_{ij} - \eta_{ij0}) \sim AN(0, n_i \mathbf{F}_i^{aa}).$$

By the delta method and assumptions of Theorem 3.1.2, we obtain

$$\left\| \frac{1}{n_i} \left( (\mathbf{F}_i^{\beta\beta})^{-1} - (\hat{\mathbf{F}}_{\mathbf{i}}^{\beta\beta})^{-1} \right) \right\| = O_p(1/\sqrt{n_i}).$$

The asymptotic normality result follows.

### 3.4.4 Proof of Corollary 3.1.4

We prove the unconditional convergence in (3.8); the same method applies to (3.9) and (3.10).

Write  $\mathbf{Z}_{m, N} = \left( \sum_{i=1}^m \{\mathbf{F}_i^{\beta\beta}\}^{-1} \right)^{-1/2} (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0)$ , Let  $\mathbf{B}$  be any  $p$ -dimensional rectangle with rational coordinates for all vertices. We know from Theorem 3.1.2

that

$$P[\mathbf{Z}_{m, N} \in \mathbf{B} | \mathbf{v}_0] \rightarrow P(\mathbf{Z} \in \mathbf{B})$$

where  $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$ .

From properties of conditional expectation,  $E[P(\mathbf{Z}_{m, N} \in \mathbf{B} | \mathbf{v}_0)] = P[\mathbf{Z}_{m, N} \in \mathbf{B}]$ . where the expectation is taken over the distribution of  $\mathbf{v}_0$ . Almost surely,  $P[\mathbf{Z}_{m, N} \in \mathbf{B} | \mathbf{v}_0] \leq 1$ , so we can use the Dominated Convergence Theorem to write

$$\begin{aligned} \lim P[\mathbf{Z}_{m, N} \in \mathbf{B}] &= \lim E[P(\mathbf{Z}_{m, N} \in \mathbf{B} | \mathbf{v}_0)] \\ &= E[\lim P(\mathbf{Z}_{m, N} \in \mathbf{B} | \mathbf{v}_0)] \\ &= E[P(\mathbf{Z} \in \mathbf{B})] = P(\mathbf{Z} \in \mathbf{B}). \end{aligned}$$

In fact, since the rationals are countable, the above convergence holds simultaneously over all rectangles whose vertices have rational components, and hence for all Borel sets.

### 3.4.5 Proof of Theorem 3.2.1

Consider  $(\mathbf{Q}_{22}^*)^{-1}$ :

$$(\mathbf{Q}_{22}^*)^{-1} = (\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z} + \frac{\lambda_1}{m} \mathbf{J})^{-1} = (\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-\frac{1}{2}} (\mathbf{I} + \mathbf{A})^{-1} (\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-\frac{1}{2}} \quad (3.53)$$

where

$$\mathbf{A} = (\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-\frac{1}{2}} \frac{\lambda_1}{m} \mathbf{J} (\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-\frac{1}{2}}.$$

Since  $\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z}$  is a symmetric matrix, using the  $l_2$  norm of a symmetric matrix we have:

$$\begin{aligned}\|(\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-\frac{1}{2}}\|^2 &= \sup_{\|\mathbf{u}\|=1} |\mathbf{u}^T (\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-\frac{1}{2}} (\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-\frac{1}{2}} \mathbf{u}| \\ &= \|(\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-1}\|.\end{aligned}$$

Then

$$\|\mathbf{A}\| \leq \|(\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-\frac{1}{2}}\|^2 \left\| \frac{\lambda_1}{m} \mathbf{J} \right\| = \frac{|\lambda_1|}{\min_i |E_i^*|}.$$

Recall that  $B_{ij}$  defined in (3.24) is positive and lies between  $b''_{ij}$  and  $b''_{ij}(\hat{\eta}_{ij})$ .

By Jiang's [14] condition (J2), we have the following:

$$\|\mathbf{A}\| \leq \frac{|\lambda_1|}{\min_i |E_i^*|} \leq 1/(n_* \delta_N) = o(1/\sqrt{n_*}).$$

where  $n_* = \min_i n_i$ .

By applying the Neumann series we have

$$(\mathbf{Q}_{22}^*)^{-1} = (\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-\frac{1}{2}} (\mathbf{I} - \mathbf{A} + \mathbf{A}^2 - \mathbf{A}^3 + \cdots) (\mathbf{Z}^T \tilde{\mathbf{B}} \mathbf{Z})^{-\frac{1}{2}}$$

After the calculation we can get a closed form of  $(\mathbf{Q}_{22}^*)^{-1}$  as the following:

$$(\mathbf{Q}_{22}^*)^{-1} = \begin{bmatrix} (E_1^*)^{-1} + (E_1^*)^{-2} C_{\lambda_1}^* & \cdots & (E_1^*)^{-1} (E_m^*)^{-1} C_{\lambda_1}^* \\ \vdots & \ddots & \vdots \\ (E_1^*)^{-1} (E_m^*)^{-1} C_{\lambda_1}^* & \cdots & (E_m^*)^{-1} + (E_m^*)^{-2} C_{\lambda_1}^* \end{bmatrix}_{m \times m} \quad (3.54)$$

where

$$C_{\lambda_1}^* = \sum_{i=1}^{\infty} \left( \frac{\lambda_1}{m} \right)^i \left( \sum_{k=1}^m (E_k^*)^{-1} \right)^{i-1} (-1)^i = -\frac{\lambda_1}{\lambda_1 \sum_{k=1}^m (E_k^*)^{-1} + m}. \quad (3.55)$$

and  $E_i^* = \sum_j B_{ij}$ .

Recall (3.19). We have

$$\begin{aligned}\sqrt{m}(\hat{a} - a_0) &= \sqrt{m}(Q_{11.2}^*)^{-1} (\mathbf{1}_N^T - \mathbf{Q}_{12}^*(\mathbf{Q}_{22}^*)^{-1} \mathbf{Z}^T) (\mathbf{y} - \boldsymbol{\mu}) \\ &\quad + \sqrt{m}(Q_{11.2}^*)^{-1} \mathbf{Q}_{12}^*(\mathbf{Q}_{22}^*)^{-1} \lambda_1 \bar{v}_0 \mathbf{1}_m\end{aligned}$$

where  $\mathbf{Q}_{12}^* = (E_1^*, \dots, E_m^*)$  and  $\mathbf{Q}_{11.2}^* = -m^2 C_{\lambda_1}^*$ . Then,

$$\begin{aligned}&|\sqrt{m}(\mathbf{Q}_{11.2}^*)^{-1} (\mathbf{1}_N^T(\mathbf{y} - \boldsymbol{\mu}) - \mathbf{Q}_{12}^*(\mathbf{Q}_{22}^*)^{-1} \mathbf{Z}^T(\mathbf{y} - \boldsymbol{\mu}))| \\ &= \left\{ \sqrt{m}|(Q_{11.2}^*)^{-1} (\mathbf{1}_N^T - \mathbf{Q}_{12}^*(\mathbf{Q}_{22}^*)^{-1} \mathbf{Z}^T)(\mathbf{y} - \boldsymbol{\mu})| \right\} \\ &= \sqrt{m} \left| \frac{1}{m} \sum_i (E_i^*)^{-1} \sum_j (y_{ij} - b'_{ij}) \right| \\ &\leq \frac{o_p(1)}{\sqrt{mn_*}} \left| \sum_i \sum_j (y_{ij} - b'_{ij}) \right| \\ &= o_p(1).\end{aligned}\tag{3.56}$$

Since by (J1) we have  $1/\sqrt{N} |\sum_i \sum_j (y_{ij} - b'_{ij})| = O_p(1)$ .

Then  $\sqrt{m}(\hat{a} - a_0)$  and  $\sqrt{m}(Q_{11.2}^*)^{-1} \mathbf{Q}_{12}^*(\mathbf{Q}_{22}^*)^{-1} \lambda_1 \bar{v}_0 \mathbf{1}_m$  have the same asymptotic distribution. Furthermore,

$$\begin{aligned}&\sqrt{m}(Q_{11.2}^*)^{-1} \mathbf{Q}_{12}^*(\mathbf{Q}_{22}^*)^{-1} \lambda_1 \bar{v}_0 \mathbf{1}_m \\ &= \sqrt{m} \frac{-1}{m^2 C_{\lambda_1}^*} \left( m \bar{v}_0 \lambda_1 + \bar{v}_0 \lambda_1 m C_{\lambda_1}^* \sum_i (E_i^*)^{-1} \right) \\ &= -\frac{\lambda_1}{m C_{\lambda_1}^*} \sqrt{m} \bar{v}_0 - \frac{\bar{v}_0 \lambda_1}{\sqrt{m}} \sum_i (E_i^*)^{-1}.\end{aligned}$$

Since by (J1), condition (2) of Theorem 3.1.3 and  $\bar{v}_0 \rightarrow_p 0$ ,

$$-\frac{\lambda_1}{m C_{\lambda_1}^*} = \left( 1 + \frac{\lambda_1 \sum_i (E_i^*)^{-1}}{m} \right) \rightarrow_p 1\tag{3.57}$$

$$\bar{v}_0 \lambda_1 \frac{\sum_i (E_i^*)^{-1}}{\sqrt{m}} \rightarrow_p 0\tag{3.58}$$

and the fact that the  $v_{i0}$  are iid with mean 0 and variance  $\sigma_v^2$ , by the Central Limit Theorem

$$\sqrt{m}v_0 \rightarrow_D N(0, \sigma_v^2).$$

By Slutsky's theorem, asymptotic normality follows.

## Chapter 4

### Logistic $2 \times 2 \times m$ table

This Chapter illustrates conditional asymptotic normality results of Theorem 3.1.3 and 3.1.4. In order to check the asymptotic results, simulations are performed to explore the asymptotic properties of our estimator. We also apply the estimator to a real data set to compare with Mantel-Haenszel estimator.

#### 4.1 Logistic $2 \times 2 \times m$ table

We use this example to illustrate Theorem 3.1.2 and 3.1.3. The  $2 \times 2 \times m$  table example is set up as the following:

$$\text{logit}P(y_{ij} = 1|x) = \alpha_i + \beta x_{ij}$$

where  $x_{ij}=1$  or 0 and the table is the following:

Table 4.1:  $2 \times 2 \times m$  table

	$y = 0$	$y = 1$
$x = 0$	$a_1$	$b_1$
$x = 1$	$c_1$	$d_1$

...

	$y = 0$	$y = 1$
$x = 0$	$a_m$	$b_m$
$x = 1$	$c_m$	$d_m$

For  $2 \times 2 \times m$  tables, there are two types of models:

Model I : the number of tables  $m$  remains fixed but individual cell sizes increase without bound.

Model II : the number of tables  $m$  increases but the cell sizes remained bounded.

Breslow [5] studied the properties of four commonly used estimators of the odds ratio in Model II:

Consider a series of  $m$  pairs of independent binomial observations  $(d_i, b_i)$  with denominators  $(n_i = d_i + c_i, m_i = a_i + b_i)$  and success probabilities  $(p_{1i}, p_{0i})$  for  $i = 1, \dots, m$ . Its assumed throughout that the odds ratio  $\psi = (p_{1i}q_{0i})/(p_{0i}q_{1i})$  remains constant from table to table.

One of the earliest estimators (Woolf, 1955) of the common odds ratio  $\psi$  is the empirical logit estimator defined by

$$\log(\hat{\psi}_G) = \left[ \sum_{i=1}^m w_i \log \left\{ \frac{(a_i + \Delta)(d_i + \Delta)}{(c_i + \Delta)(b_i + \Delta)} \right\} \right] \Bigg/ \left( \sum_i w_i \right)$$

where the weights are

$$w_i = (1/(a_i + \Delta) + 1/(b_i + \Delta) + 1/(c_i + \Delta) + 1/(d_i + \Delta))^{-1}$$

and  $\Delta > 0$  is a constant added to each cell to avoid zero denominators. The choice  $\Delta = 1/2$  is the most popular because it is thought to reduce the bias in small examples (Anscombe, 1956). There are unconditional MLE and conditional MLE (the details omitted here). The fourth and final estimator is given by famous formula of Mantel and Haenszel (1959):

$$\hat{\psi}_{MH} = \frac{\sum_i (a_i d_i / N_i)}{\sum_i (c_i b_i / N_i)}$$

where  $N_i = a_i + b_i + c_i + d_i = m_i + n_i$ . Due largely to its simplicity,  $\hat{\psi}_{MH}$  has been widely used by practicing statisticians and epidemiologists. Breslow [5] showed that

in the Model II setting, the empirical logit estimator does not converge to the true odds ratio. The Mantel-Haenszel estimator is consistent and retains good efficiency even for moderately large odds ratios in the Model II setting (sparse data).

Let  $R_i = a_i d_i / N_i$  and  $S_i = c_i b_i / N_i$ , Breslow [5] proposed an empirical estimate of  $\text{Var}[\hat{\psi}_{MH}]$  as

$$m\widehat{\text{Var}}_E(\hat{\psi}_{MH}) = \frac{\sum_i (R_i - \hat{\psi}_{MH} S_i) / m}{(\sum_i S_i / m)^2}$$

Robins, Breslow and Greenland [25] proposed a new estimator of  $\text{Var}[\hat{\psi}_{MH}]$  as:

$$m\widehat{\text{Var}}_{US}(\hat{\psi}_{MH}) = m \left[ \frac{\sum_i P_i R_i}{2R_+^2} + \frac{\sum_i (P_i S_i + Q_i R_i)}{2R_+ S_+} + \frac{\sum_i Q_i S_i}{2S_+^2} \right] (\hat{\psi}_{MH})^2$$

where  $P_i = (a_i + d_i) / N_i$ ,  $Q_i = (c_i + b_i) / N_i$ ,  $R_+ = \sum_i R_i$  and  $S_+ = \sum_i S_i$ . The corresponding estimators of  $m\text{Var}(\log \hat{\psi}_{MH})$  are

$$mV_i = m\widehat{\text{Var}}_i(\hat{\psi}_{MH}) / (\hat{\psi}_{MH})^2$$

where  $i \in US$ ,  $E$  means using  $\widehat{\text{Var}}_{US}$ ,  $\widehat{\text{Var}}_E$  respectively. The Robins-Breslow-Greenland [25] estimator is consistent under both Model I and Model II.

The estimating equation  $\mathbf{x}_i^t(\mathbf{y}_i - \mu_i)$  is

$$\begin{aligned} 0 &= b_i + d_i - (a_i + b_i) \exp(\alpha_i) / (1 + \exp(\alpha_i)) \\ &\quad - (c_i + d_i) \exp(\alpha_i + \beta) / (1 + \exp(\alpha_i + \beta)) \end{aligned} \tag{4.1}$$

$$0 = d_i - (c_i + d_i) \exp(\alpha_i + \beta) / (1 + \exp(\alpha_i + \beta)) \tag{4.2}$$

and we can get the MLE of  $\beta$  as  $\log(a_i d_i / b_i c_i) = \log a_i + \log d_i - \log b_i - \log c_i$ .

Let

$$Z_{i0} = (a_i - E(a_i | \alpha_{i0})) / \sqrt{\text{Var}(a_i | \alpha_{i0})}$$

and

$$Z_{i1} = (c_i - E(c_i|\alpha_{i0})) / \sqrt{\text{Var}(c_i|\alpha_{i0})}$$

For simplicity we assume  $a_i + b_i = n_i$  and  $c_i + d_i = n_i$ . Since given  $\alpha_{i0}$ ,  $a_i$  belongs to  $\text{Binomial}(n_i, 1/(1 + \exp(\alpha_{i0})))$  we have

$$\begin{aligned} a_i &= \frac{n_i}{1 + \exp(\alpha_{i0})} + \sqrt{\text{Var}(a_i|\alpha_{i0})} \left( \frac{a_i - E(a_i|\alpha_{i0})}{\sqrt{\text{Var}(a_i|\alpha_{i0})}} \right) \\ &= \frac{n_i}{1 + \exp(\alpha_{i0})} + Z_{i0} \frac{\exp(\alpha_{i0}/2) \sqrt{n_i}}{1 + \exp(\alpha_{i0})} \\ &= \frac{n_i}{1 + \exp(\alpha_{i0})} \left( 1 + Z_{i0} \frac{\exp(\alpha_{i0}/2)}{\sqrt{n_i}} \right), \\ \log a_i &\approx \log \left( \frac{n_i}{1 + \exp(\alpha_{i0})} \right) + Z_{i0} \frac{\exp(\alpha_{i0}/2)}{\sqrt{n_i}} - Z_{i0}^2 \frac{\exp(\alpha_{i0})}{2n_i} + o\left(\frac{1}{n_i}\right). \end{aligned}$$

By the same argument we have

$$\begin{aligned} -\log b_i &\approx -\log \left( \frac{n_i \exp(\alpha_{i0})}{1 + \exp(\alpha_{i0})} \right) + Z_{i0} \frac{\exp(-\alpha_{i0}/2)}{\sqrt{n_i}} + Z_{i0}^2 \frac{\exp(-\alpha_{i0})}{2n_i} + o\left(\frac{1}{n_i}\right), \\ -\log c_i &\approx -\log \left( \frac{n_i}{1 + \exp(\alpha_{i0} + \beta_0)} \right) - Z_{i1} \frac{\exp((\alpha_{i0} + \beta_0)/2)}{\sqrt{n_i}}, \\ &\quad + Z_{i1}^2 \frac{\exp(\alpha_{i0} + \beta_0)}{2n_i} + o\left(\frac{1}{n_i}\right) \\ \log d_i &\approx \log \left( \frac{n_i \exp(\alpha_{i0} + \beta_0)}{1 + \exp(\alpha_{i0} + \beta_0)} \right) - Z_{i1} \frac{\exp(-(\alpha_{i0} + \beta_0)/2)}{\sqrt{n_i}} \\ &\quad - Z_{i1}^2 \frac{\exp(-(\alpha_{i0} + \beta_0))}{2n_i} + o\left(\frac{1}{n_i}\right). \end{aligned}$$

Then

$$\begin{aligned} \hat{\beta}_i &= \beta_0 + \frac{Z_{i0}}{\sqrt{n_i}} (\exp(\alpha_{i0}/2) + \exp(-\alpha_{i0}/2)) \\ &\quad - \frac{Z_{i1}}{\sqrt{n_i}} (\exp((\alpha_{i0} + \beta_0)/2) + \exp(-(\alpha_{i0} + \beta_0)/2)) \\ &\quad - \frac{Z_{i0}^2}{2n_i} (\exp(\alpha_{i0}) - \exp(-\alpha_{i0})) \\ &\quad + \frac{Z_{i1}^2}{2n_i} (\exp(\alpha_{i0} + \beta_0) - \exp(-(\alpha_{i0} + \beta_0))) + o_p\left(\frac{1}{n_i}\right). \end{aligned}$$

The bias of  $\hat{\beta}_i$  is the following:

$$\begin{aligned} & E((\hat{\beta}_i - \beta_0) | \alpha_{i0}) \\ &= \frac{\exp(\alpha_{i0})(\exp(\beta_0) - 1) - \exp(-\alpha_{i0})(\exp(-\beta_0) - 1)}{2n_i} + o\left(\frac{1}{n_i}\right). \end{aligned}$$

Let

$$G(x) = (4 + 4\exp(-x) + 4\exp(x))^{-1}, \quad (4.3)$$

in order to check the hypotheseses of Theorem 3.1.2, first we consider  $\frac{1}{n_i}(F_i^{\beta\beta})^{-1} = G(\eta_{ij})$  where  $\eta_{ij} = \alpha_i + x_{ij}\beta$  and  $x_{ij} = 1$  or  $0$ . Since we assume  $|\alpha_{i0}|$  is bounded, it is obvious that  $\inf_i G(\eta_{ij0}) > \delta_1 > 0$  and  $\sup_i G(\eta_{ij0}) < M_1 < \infty$ , for some positive constants  $\delta_1$  and  $M_1$ . Similar arguments follow for  $n_i(\mathbf{F}_{n_i})^{-1}$ .

We need to verify condition (1) of Theorem 3.1.3. In this case we have the log-likelihood function for  $i$ th group as the following:

$$l = b_i\alpha_i + d_i(\alpha_i + \beta) - n_i \log(1 + \exp(\alpha_i)) - n_i \log(1 + \exp(\alpha_i + \beta)).$$

Then

$$-\frac{1}{n_i} l''(\alpha_{i0}, \beta_0) = \begin{pmatrix} I_{i1} + I_{i2} & I_{i2} \\ I_{i2} & I_{i2} \end{pmatrix} \quad (4.4)$$

where

$$\begin{aligned} I_{i1} &= \frac{\exp(\alpha_{i0})}{(1 + \exp(\alpha_{i0}))^2} \\ I_{i2} &= \frac{\exp(\alpha_{i0} + \beta_0)}{(1 + \exp(\alpha_{i0} + \beta_0))^2} \end{aligned}$$

Since in this special example  $F_i^{\beta\beta}$  as a scalar instead of a matrix, the Lindeberg condition verification can be simplified and for condition (1) of Theorem 3.1.2 we

only need to verify

$$K_1 < \frac{\min_i F_i^{\beta\beta}}{\max_i F_i^{\beta\beta}} < K_2$$

where  $K_1$  and  $K_2$  are positive constants. Obviously here  $K_2 = 1$  since here we have

$$F_i^{\beta\beta} = (1/n_i) ((1 + \exp(-\beta_0)) \exp(-\alpha_{i0}) + (1 + \exp(\beta_0)) \exp(\alpha_{i0}) + 4).$$

Consider condition (3) of Theorem 3.1.2: Let

$$\frac{\lambda_{\max} \mathbf{F}_{n_i}^{-1}}{\lambda_{\min} \mathbf{F}_{n_i}^{-1}} = \frac{\sup_{\|\mathbf{u}\|=1} f_i(\mathbf{u})}{\inf_{\|\mathbf{u}\|=1} f_i(\mathbf{u})}$$

where

$$f_i(\mathbf{u}) = \frac{(1 + \exp(\alpha_{i0} + \beta_0))^2}{\exp(\alpha_{i0} + \beta_0)} u_2^2 + (u_1 - u_2)^2 \frac{(1 + \exp(\alpha_{i0}))^2}{\exp(\alpha_{i0})}.$$

For the further calculation we write

$$\begin{aligned} f_i(\mathbf{u}) &= \exp(\alpha_{i0}) \left[ \frac{(\exp(-\alpha_{i0}) + \exp(\beta_0))^2}{\exp(\beta_0)} u_2^2 + (u_1 - u_2)^2 (1 + \exp(-\alpha_{i0}))^2 \right] \\ &= \exp(-\alpha_{i0}) \left[ \frac{(1 + \exp(\alpha_{i0} + \beta_0))^2}{\exp(\beta_0)} u_2^2 + (u_1 - u_2)^2 (1 + \exp(\alpha_{i0}))^2 \right] \end{aligned}$$

Let

$$B_1 = \frac{(\exp(-\alpha_{i0}) + \exp(\beta_0))^2}{\exp(\beta_0)}, \quad A_1 = (1 + \exp(-\alpha_{i0}))^2.$$

and

$$B_2 = \frac{(1 + \exp(\alpha_{i0} + \beta_0))^2}{\exp(\beta_0)}, \quad A_2 = (1 + \exp(\alpha_{i0}))^2.$$

Here we have  $u_1 = \sqrt{1 - u_2^2}$  for  $A_1, B_1$  and  $A_2, B_2$  as the above, we have

$$\begin{aligned} (u_2^2)_1 &= \frac{1}{2} \pm \frac{1}{2} \sqrt{1 - 4A_1^2/(4A_1^2 + B_1^2)} \\ (u_2^2)_2 &= \frac{1}{2} \pm \frac{1}{2} \sqrt{1 - 4A_2^2/(4A_2^2 + B_2^2)} \end{aligned}$$

So we have

$$\begin{aligned} & \max_i \frac{\sup_{\|\mathbf{u}\|=1} f_i(\mathbf{u})}{\inf_{\|\mathbf{u}\|=1} f_i(\mathbf{u})} \\ &= \max_i \frac{A_1(1 - 2\sqrt{A_1^2/(4A_1^2 + B_1^2)}) + B_1(\frac{1}{2} + \frac{1}{2}\sqrt{1 - 4A_1^2/(4A_1^2 + B_1^2)})}{A_1(1 - 2\sqrt{A_1^2/(4A_1^2 + B_1^2)}) + B_1(\frac{1}{2} - \frac{1}{2}\sqrt{1 - 4A_1^2/(4A_1^2 + B_1^2)})} \end{aligned}$$

and we can see that when  $\alpha_{i0} \rightarrow +\infty$  we have  $\max_i (\sup_{\|\mathbf{u}\|=1} f_i(\mathbf{u}) / \inf_{\|\mathbf{u}\|=1} f_i(\mathbf{u})) = O(1)$ . Also

$$\begin{aligned} & \max_i \frac{\sup_{\|\mathbf{u}\|=1} f_i(\mathbf{u})}{\inf_{\|\mathbf{u}=1\|} f_i(\mathbf{u})} \\ &= \max_i \frac{A_2(1 - 2\sqrt{A_2^2/(4A_2^2 + B_2^2)}) + B_2(\frac{1}{2} + \frac{1}{2}\sqrt{1 - 4A_2^2/(4A_2^2 + B_2^2)})}{A_2(1 - 2\sqrt{A_2^2/(4A_2^2 + B_2^2)}) + B_2(\frac{1}{2} - \frac{1}{2}\sqrt{1 - 4A_2^2/(4A_2^2 + B_2^2)})} \end{aligned}$$

we can see that when  $\alpha_{i0} \rightarrow -\infty$  we have  $\max_i (\sup_{\|\mathbf{u}\|=1} f_i(\mathbf{u}) / \inf_{\|\mathbf{u}\|=1} f_i(\mathbf{u})) = O(1)$ . Consider

$$\mathbf{F}_{n_i}^{-1} = \left( -l''(\alpha_{i0}, \beta_0) \right)^{-1} = 1/n_i \begin{pmatrix} I_{i1}^{-1} & -I_{i1}^{-1} \\ -I_{i1}^{-1} & I_{i1}^{-1} + I_{i2}^{-1} \end{pmatrix}. \quad (4.5)$$

Here  $\|\mathbf{u}\|^2 = u_1^2 + u_2^2 = 1$  so that

$$\begin{aligned} \sum_i^m \|\mathbf{F}_{n_i}^{-1}\| &= \sum_i \frac{1}{n_i} \sup_{\|\mathbf{u}\|=1} \left( \frac{(1 + \exp(\alpha_{i0} + \beta_0))^2}{\exp(\alpha_{i0} + \beta_0)} u_2^2 + (u_1 - u_2)^2 \frac{(1 + \exp(\alpha_{i0}))^2}{\exp(\alpha_{i0})} \right) \\ &\leq \sum_i^m \frac{1}{n_i} (6 + (2 + \exp(-\beta_0)) \exp(-\alpha_{i0}) + (2 + \exp(\beta_0)) \exp(\alpha_{i0})) \\ &\leq \frac{6m}{\min_i n_i} + \sum_{i=1}^m c_1 \exp(|\alpha_{i0}|) \frac{1}{n_i} \\ &\leq \frac{6m}{\min_i n_i} + \frac{mc_1}{\min_i n_i m} \sum_{i=1}^m \exp(|\alpha_{i0}|). \end{aligned} \quad (4.6)$$

If we have condition the  $(1/m) \sum_{i=1}^m \exp(|\alpha_{i0}|) \rightarrow_{a.s.} E(\exp(|\alpha_{i0}|))$ , then by condition (2) of Theorem 3.1.2,  $(4.6) = 6m/\min_i n_i + O_p(m/\min_i n_i) = o_p(1)$ , the asymptotic

relation  $\|(1/\sqrt{n_i})[\mathbf{F}_{n_i} - \mathbf{F}_{n_i}(\boldsymbol{\gamma}^*)]\| = O_p(1)$  can be verified following the same argument in the proof of Theorem 3.1.3, we verified assumptions of Theorem 3.1.2 which gives  $\|\sqrt{n_i}(\mathbf{F}_{n_i})^{-T/2}\| = O(1)$ , so condition (3) of Theorem 3.1.3 is satisfied.

Consider condition (4) of Theorem 3.1.2 and here we let  $\psi(t) = t$ :

We have:

$$\begin{aligned} & \hat{\beta}_i - E(\hat{\beta}_i | \alpha_{i0}) \\ &= \frac{Z_{i0}}{\sqrt{n_i}} (\exp(\alpha_{i0}/2) + \exp(-\alpha_{i0}/2)) \\ &\quad - \frac{Z_{i1}}{\sqrt{n_i}} (\exp((\alpha_{i0} + \beta_0)/2) + \exp(-(α_{i0} + β_0)/2)) \\ &\quad - \frac{1 + Z_{i0}^2}{2n_i} (\exp(\alpha_{i0}) - \exp(-\alpha_{i0})) \\ &\quad + \frac{1 + Z_{i1}^2}{2n_i} (\exp(\alpha_{i0} + \beta_0) - \exp(-(α_{i0} + β_0))) + o\left(\frac{1}{n_i}\right), \end{aligned}$$

and

$$\begin{aligned} & \left[ \frac{K_{i1}Z_{i0}}{\sqrt{n_i}} - \frac{K_{i3}(1 + Z_{i0}^2)}{2n_i} + \frac{K_{i4}(1 + Z_{i1}^2)}{2n_i} - \frac{K_{i2}Z_{i1}}{\sqrt{n_i}} + o\left(\frac{1}{n_i}\right) \right]^4 \\ & \leq \left[ \left| \frac{K_{i1}Z_{i0}}{\sqrt{n_i}} - \frac{K_{i2}Z_{i1}}{\sqrt{n_i}} \right| + \left| \frac{K_{i4}(1 + Z_{i1}^2)}{2n_i} - \frac{K_{i3}(1 + Z_{i0}^2)}{2n_i} + o\left(\frac{1}{n_i}\right) \right| \right]^4 \\ & \leq 8 \left[ \left| \frac{K_{i1}Z_{i0}}{\sqrt{n_i}} - \frac{K_{i2}Z_{i1}}{\sqrt{n_i}} \right|^4 + \left| \frac{K_{i4}(1 + Z_{i1}^2)}{2n_i} - \frac{K_{i3}(1 + Z_{i0}^2)}{2n_i} + o\left(\frac{1}{n_i}\right) \right|^4 \right] \\ & \leq 64 \left[ \left| \frac{K_{i1}Z_{i0}}{\sqrt{n_i}} \right|^4 + \left| \frac{K_{i2}Z_{i1}}{\sqrt{n_i}} \right|^4 + \left| \frac{K_{i4}(1 + Z_{i1}^2)}{2n_i} \right|^4 + \left| o\left(\frac{1}{n_i}\right) - \frac{K_{i3}(1 + Z_{i0}^2)}{2n_i} \right|^4 \right] \\ & \leq 64 \left[ \frac{K_{i1}^4 Z_{i0}^4}{n_i^2} + \frac{K_{i2}^4 Z_{i1}^4}{n_i^2} + \frac{K_{i4}^4 (1 + Z_{i1}^2)^4}{16n_i^4} + o\left(\frac{8}{n_i^4}\right) + \frac{8K_{i3}^4 (1 + Z_{i0}^2)^4}{16n_i^4} \right] \\ & \leq 64 \left[ \frac{K_{i1}^4 Z_{i0}^4}{n_i^2} + \frac{K_{i2}^4 Z_{i1}^4}{n_i^2} + \frac{K_{i4}^4 (1 + Z_{i1}^8)}{2n_i^4} + o\left(\frac{8}{n_i^4}\right) + \frac{4K_{i3}^4 (1 + Z_{i0}^8)}{n_i^4} \right]. \end{aligned}$$

Then we have

$$\begin{aligned}
& E((\hat{\beta}_i^{*t} \hat{\beta}_i^*)^2 | \alpha_{i0}) \\
&= (F_i^{\beta\beta})^{-2} E \left\{ \left[ \frac{K_{i1} Z_{i0}}{\sqrt{n_i}} - \frac{K_{i3}(1+Z_{i0}^2)}{2n_i} + \frac{K_{i4}(1+Z_{i1}^2)}{2n_i} - \frac{K_{i2} Z_{i1}}{\sqrt{n_i}} + o\left(\frac{1}{n_i}\right) \right]^4 | \alpha_{i0} \right\} \\
&\leq 64(F_i^{\beta\beta})^{-2} E \left\{ \left[ \frac{K_{i1}^4 Z_{i0}^4}{n_i^2} + \frac{K_{i2}^4 Z_{i1}^4}{n_i^2} + \frac{K_{i4}^4(1+Z_{i1}^8)}{2n_i^4} + o\left(\frac{8}{n_i^4}\right) + \frac{4K_{i3}^4(1+Z_{i0}^8)}{n_i^4} \right] | \alpha_{i0} \right\}
\end{aligned}$$

where  $K_{i1} = \exp(\alpha_{i0}/2) + \exp(-\alpha_{i0}/2)$ ,  $K_{i2} = \exp((\alpha_{i0} + \beta_0)/2) + \exp(-(\alpha_{i0} + \beta_0)/2)$ ,  $K_{i3} = \exp(\alpha_{i0}) - \exp(-\alpha_{i0})$ ,  $K_{i4} = \exp(\alpha_{i0} + \beta_0) - \exp(-(\alpha_{i0} + \beta_0))$ .

Since here  $Z_{i0}$  and  $Z_{i1}$  are normalized binomial random variables for  $n_i$  with different means  $n_i/(1+\exp(\alpha_{i0}))$ ,  $n_i/(1+\exp(\alpha_{i0}+\beta_0))$  and variances  $n_i \exp(\alpha_{i0})/(1+\exp(\alpha_{i0}))$ ,  $n_i \exp(\alpha_0 + \beta_0)/(1+\exp(\alpha_{i0} + \beta_0))$  respectively, which are both in order of  $n_i$ , by M. Znidaric [38] we have  $E(Z_{i0}^4 | \alpha_{i0}) = (n_i^2 + O(n_i^{-1}))O_p(n_i^{-2}) = O_p(1)$ . By the same argument we have  $E(Z_{i0}^8 | \alpha_{i0}) = O_p(1)$ ,  $E(Z_{i1}^4 | \alpha_{i0}) = O_p(1)$  and  $E(Z_{i1}^8 | \alpha_{i0}) = O_p(1)$ , so it is easy to show that condition (4) of Theorem 3.1.2 satisfied.

Now we consider the following:

$$\begin{aligned}
& \left( (\sum_i (F_i^{\beta\beta})^{-1})^{-\frac{1}{2}} \sum_i (\mathbf{F}_i^{\beta\beta})^{-1} (E(\hat{\beta}_i | v_{i0}) - \beta_0) \right)^2 \\
&= \left\{ \sum_i (F_i^{\beta\beta})^{-1} \left( \frac{K_{i4} - K_{i3}}{2n_i} + o_p\left(\frac{1}{n_i}\right) \right) \right\}^2 \times \left\{ \sum_i n_i (I_{i1}^{-1} + I_{i2}^{-1})^{-1} \right\}^{-1} \\
&\leq m \sum_i n_i^2 \left( \frac{I_{i1} I_{i2}}{I_{i1} + I_{i2}} \right)^2 \left( \frac{K_{i4} - K_{i3}}{2n_i} + o_p\left(\frac{1}{n_i}\right) \right)^2 \left\{ \sum_i n_i (I_{i1}^{-1} + I_{i2}^{-1})^{-1} \right\}^{-1} \quad (4.7)
\end{aligned}$$

$$\leq 2m \sum_i (I_{i1}^{-1} + I_{i2}^{-1})^{-2} ((K_{i4}^2 + K_{i3}^2) + o_p(1)) \left\{ \sum_i n_i (I_{i1}^{-1} + I_{i2}^{-1})^{-1} \right\}^{-1} \quad (4.8)$$

$$\leq O_p(1) \frac{m}{\min_i n_i} \quad (4.9)$$

The step from (4.7) to (4.8) follows because of (1) and (2) of Theorem 3.1.2, the fact that  $|\alpha_{i0}|$  is bounded and  $\max_i (K_{i3}^2 + K_{i4}^2) / \min_i (I_{i1}^{-1} + I_{i2}^{-1}) = O_p(1)$ . The result

of Theorem 3.1.2 follows.

Consider (1) of Theorem 3.1.3. We have

$$\begin{aligned}\frac{1}{N} \|\hat{\mathbf{V}} - \mathbf{V}\| &= \frac{1}{N} \left\| \sum_i n_i \{G(\hat{\eta}_{ij}) - G(\eta_{ij0})\} \right\| \\ &\leq \frac{1}{N} \sum_{i=1}^m n_i \|G(\hat{\eta}_{ij}) - G(\eta_{ij0})\| = \frac{1}{N} \sum_i \sqrt{n_i} O_p(1) = o_p(1).\end{aligned}$$

The above conclusion follows from the delta method since, given  $\alpha_{i0}$ ,

$$\sqrt{n_i}(G(\hat{\eta}_{ij}) - G(\eta_{ij0})) \sim AN(0, n_i(G'(\eta_{ij0}))^2(F_i^{aa} + 2F_i^{a\beta} + F_i^{\beta\beta})).$$

Also  $N|V^{-1}| = N|(\sum_i n_i G(\eta_{ij0})| = O(1)$  and condition (1) of Theorem 3.1.3 is satisfied. For condition (2) of Theorem 3.1.4,  $|(\hat{F}_i^{\beta\beta})^{-1} - (F_i^{\beta\beta})^{-1}| = n_i|G(\hat{\eta}_{ij}) - G(\eta_{ij0})| = O_p(\sqrt{n_i})$ , the condition (3) is easy to verified. The result of Theorem 3.1.3 follows. Similar argument follow if we replace  $\hat{\beta}_{ij}$  by  $\hat{\eta}_{ij}$  since ,given  $\alpha_{i0}$ ,

$$\sqrt{n_i}(G(\hat{\eta}_{ij}) - G(\eta_{ij0})) \sim AN(0, n_i(G'(\eta_{ij0}))^2 F_i^{aa}).$$

## 4.2 Simulation results for $2 \times 2 \times m$ table.

We consider the  $2 \times 2 \times m$  table as the following set up:

$$\text{logit}P(y_{ij} = 1 | \alpha_i, x_{ij}) = \alpha_i + x_{ij}\beta$$

where  $\alpha_i$  is the random effect and  $\beta$  is fixed effect,  $x_{ij}=0$  or 1. The  $\alpha_i$ s are uniformly distributed between -0.8 and 0.8 since from the Section 4.1 we assume  $|\alpha_{i0}|$  bounded. We maximize the following likelihood function to get estimator of  $\hat{\beta}_i$  and  $\hat{\alpha}_i$  for each group according to Table 4.1.

$$L = \left( \frac{1}{1 + \exp(\alpha_i)} \right)^{a_i} \left( \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)} \right)^{b_i} \left( \frac{1}{1 + \exp(\alpha_i + \beta)} \right)^{c_i} \left( \frac{\exp(\alpha_i)}{1 + \exp(\beta + \alpha_i)} \right)^{d_i}$$

The weighted sum of the  $\hat{\beta}$  (using both true and estimated weights) are simulated. It is easy to solve the above equation and get  $\hat{\beta}_i = \log(a_i d_i / c_i b_i)$  and  $\hat{\alpha}_i = \log b_i / a_i$  for the  $i$ th table.

#### 4.2.1 Unconditional convergence in distribution

We simulated both balanced and unbalanced  $m$  tables with  $n$  observations per table. We generated product binomial data for the first and second rows with success probabilities  $\exp(\alpha_{i0})/(1+\exp(\alpha_{i0}))$  and  $\exp(\beta_0+\alpha_{i0})/(1+\exp(\beta_0+\alpha_{i0}))$  respectively. Here  $a_i + b_i + c_i + d_i = n$  and for the balanced case we have  $a_i + b_i = c_i + d_i = n/2$ . For the unbalanced case we either have  $a_i + b_i = n/3$  or  $a_i + b_i = n/4$ . We simulated with either  $\beta_0 = 1$  or  $= 0.5$ . Various combinations of  $(m, n)$  and true values of  $\beta$  were used for both balanced and unbalanced settings.

For each combination, 1000 replications of random effects  $\alpha_i$  and  $m$  groups of  $a_i$ ,  $b_i$ ,  $c_i$  and  $d_i$  were generated. Estimated regression coefficients for various choices of  $(m, n)$  and  $\beta_0$  with  $\alpha_i \sim Unif(-0.8, 0.8)$  are summarized in Table 4.2 and Table 4.3.

For the balanced setup, the tables display the means and standard errors of the simulated values of  $(\hat{\beta}_w - \beta_0)/s.e.$ ,  $(\hat{\beta}_w - \beta_0)/\widehat{s.e.}$ ,  $\hat{\beta}_w - \beta_0$ ,  $\hat{\beta}_w - \beta_0$ , 95% confidence bounds for  $(\hat{\beta}_w - \beta_0)$  and  $(\hat{\beta}_w - \beta_0)$  based on Student's  $t$ . From Table 4.2 and Table 4.3 under the balanced setup, we can see that for  $\beta_0 = 0.5$  in all the combinations  $\hat{\beta}_w$  is approximately unbiased and for all combinations with  $\beta_0 = 1$  except for  $(100, 50)$   $\hat{\beta}_w$  is slightly biased. For  $\beta_0 = 0.5$  and  $\beta_0 = 1$   $\hat{\beta}_w$  is approximately biased in all

combinations except the combination  $\beta_0 = 1$  or  $0.5$ ,  $m = 100$  and  $n = 50$ .

For all of the combinations in both tables, Shapiro-Wilk test  $p$  values are above 0.24 except for  $\beta_0 = 0.5$  (20, 400) and  $\beta_0 = 1$  and (40, 800) which means in those combinations,  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  given  $\alpha_0$  are normal. Likewise, Kolmogorov-Smirnov test  $p$  values are larger than 0.06.

From Table 4.2 and Table 4.3, we can see that bias/se < 0.208 except for extreme case ( $m = 100, n = 50$ ) so the normal inference is not greatly affected in the cases where consistency results do not hold. For the combination (100, 50), in order to avoid 0 observations in the cells we use (Woolf, 1955)'s adjusted method to have

$$\hat{\beta}_i = \log \{[(a_i + 1/2)(d_i + 1/2)] / [(c_i + 1/2)(b_i + 1/2)]\} \text{ and } \hat{\alpha}_i = \log (\{b_i + 1/2\} / \{a_i + 1/2\}).$$

We compared our estimators  $\hat{\beta}_w$  with true weights and  $\hat{\hat{\beta}}_w$  with estimated weights to the Mantel-Haenszel estimator  $\hat{\beta}_{MH}$ . To see whether our estimators are more efficient, we compared the standard deviation of our estimators from 1000 replications with that of the Mantel-Haenszel estimator. From the results of different combinations of  $(m, n)$  and  $\beta_0$ , our estimators are almost as efficient as the Mantel-Haenszel estimator. Since  $\hat{\hat{\beta}}_w$  is constructed by plugging the MLE  $\hat{\beta}_i$  for  $i$ th group into the weight formula, we can also construct another empirical estimator of  $\beta$  by plugging in  $\hat{\hat{\beta}}_w$ . We ran similar simulations of this new empirical estimator, which still introduced more positive bias. But the normality results hold for this new empirical estimator. It suggests that in practice, one can stay with the simple empirical estimator by plugging MLE  $\hat{\beta}_i$  from each group. The plots show that  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  are approximately normal, with departures from normality in the extreme tails in this balanced setup.

Table 4.2: Simulated estimates of logistic odds ratio in  $2 \times 2 \times m$  balanced table.

$(m, n)$	$(\hat{\beta}_w - \beta_0)$		$(\hat{\hat{\beta}}_w - \beta_0)$		
	mean	std	95% CI for $(\hat{\beta}_w - \beta_0)$	mean	std
<u><math>\beta_0 = 0.5, \text{ balanced set up}</math></u>					
(10,200)	0.006	0.095	(0.000, 0.001)	-0.0001	0.093
(10,300)	0.002	0.077	(-0.003, 0.007)	-0.002	0.076
(20,400)	0.002	0.067	(-0.002, 0.007)	-0.004	0.066
(30,600)	0.001	0.032	(-0.002, 0.001)	-0.002	0.032
(40,800)	-0.002	0.023	(0.000, 0.003)	0.000	0.023
(100,50)	-0.001	0.060	(-0.005, 0.002)	-0.028	0.056
<u><math>\beta_0 = 1, \text{ balanced setup}</math></u>					
(10,200)	0.013	0.100	(0.006, 0.019)	-0.002	0.098
(10,300)	0.012	0.080	(0.008, 0.017)	0.003	0.079
(20,400)	0.007	0.050	(0.004, 0.010)	-0.001	0.050
(30,600)	0.004	0.033	(0.002, 0.006)	-0.001	0.032
(40,800)	0.005	0.024	(0.003, 0.006)	0.001	0.024
(100,50)	0.000	0.063	(-0.004, 0.004)	0.061	0.057

Table 4.3: Simulated standardized estimates of log odds ratio (standardized by true and estimated conditional standardized error) in  $2 \times 2 \times m$  balanced table .

$(\hat{\beta}_w - \beta_0)/s.e.$			$(\hat{\hat{\beta}}_w - \beta_0)/\widehat{s.e.}$		
$(m, n)$	mean	std	95% CI for $(\hat{\hat{\beta}}_w - \beta_0)$	mean	std
<u><math>\beta_0 = 0.5, \text{ balanced set up}</math></u>					
(10,200)	0.064	1.016	(-0.006, 0.006)	-0.005	0.995
(10,300)	0.026	1.011	(-0.007, 0.003)	-0.029	0.997
(20,400)	0.037	1.017	(-0.008, 0.0003)	-0.060	0.993
(30,600)	0.016	1.019	(-0.001, 0.002)	-0.054	1.014
(40,800)	0.064	0.990	(-0.002, 0.001)	-0.006	0.986
(100,50)	-0.023	1.032	(-0.031, -0.024)	-0.475	0.958
<u><math>\beta_0 = 1, \text{ balanced set up}</math></u>					
(10,200)	0.131	1.029	(-0.008, 0.004)	-0.026	1.000
(10,300)	0.156	1.005	(-0.002, 0.008)	0.026	0.985
(20,400)	0.144	1.030	(-0.004, 0.003)	-0.013	1.021
(30,600)	0.120	1.002	(-0.003, 0.001)	-0.037	1.000
(40,800)	0.190	0.979	(-0.001, 0.002)	0.035	0.972
(100,50)	-0.001	1.043	(-0.064, -0.057)	-0.994	0.933

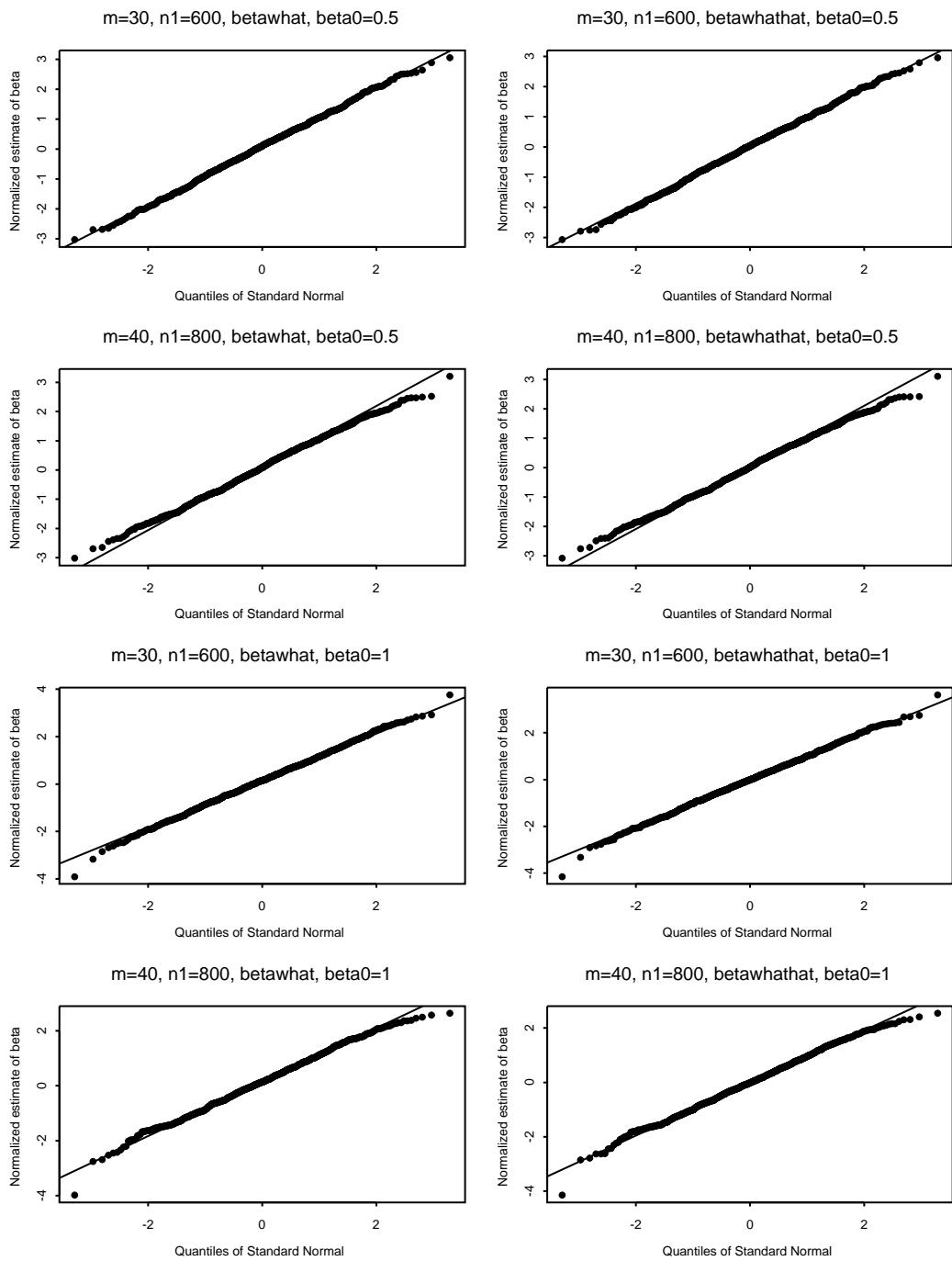


Figure 4.1: Q-Q plots for  $(\hat{\beta}_w - \beta_0)$  and  $(\hat{\hat{\beta}}_w - \beta_0)$  standardized by true and estimated conditional standard error, for various values of  $(m, n)$  in logistic  $2 \times 2 \times m$  balanced setup.

For each combination, 1000 replications of random effects  $\alpha_i$  and  $m$  groups of  $a_i, b_i, c_i$  and  $d_i$ s were generated. Estimated regression coefficients for various choices of  $(m, n)$ ,  $\beta_0$  with  $\alpha_i \sim Unif(-0.8, 0.8)$  are summarized in Table 4.4 and Table 4.5. For the unbalanced setup, the tables display the means and standard errors of the simulated values of  $(\hat{\beta}_w - \beta_0)/s.e.$ ,  $(\hat{\beta}_w - \beta_0)/\widehat{s.e.}$ ,  $(\hat{\beta}_w - \beta_0)$ ,  $(\hat{\beta}_w - \beta_0)$ , 95% confidence bounds for  $(\hat{\beta}_w - \beta_0)$  and  $(\hat{\beta}_w - \beta_0)$  based on Student's  $t$ . From Table 4.4 and Table 4.5 for  $\beta_0 = 0.5$  or 1 with unbalanced setup  $(1/3, 2/3)$  (meaning  $a_i + b_i = n/3$ ) and  $(1/4, 3/4)$  (meaning  $a_i + b_i = n/4$ ), we can see that in all the combinations  $\hat{\beta}_w$  is approximately unbiased hold except for the combination with  $\beta_0 = 1$ ,  $(1/4, 3/4)$ ,  $m = 10$ ,  $n = 200$  or  $\beta_0 = 0.5$ ,  $(1/3, 2/3)$ ,  $m = 10$ ,  $n = 300$  and  $\beta_0 = 0.5$  with  $m = 30$ ,  $n = 600$  or  $m = 20$ ,  $n = 400$ .

But their Shapiro-Wilk test  $p$  values are above 0.19 except  $\beta_0 = 0.5$ ,  $(20, 400)$  which means in those combinations,  $\hat{\beta}_w$  given  $\alpha_0$  are conditionally normal and Kolmogorov-Smirnov test  $p$  values are larger than 0.10.

For  $\hat{\beta}_w$  in all the combinations except for  $\beta_0 = 1$  or 0.5,  $(1/4, 3/4)$  or  $(1/3, 2/3)$ ,  $m = 100$ ,  $n = 60$  the estimator is approximately unbiased. From Table 4.4 and Table 4.5, we can see that bias/se < 0.109 except for extreme combinations like  $(m = 100, n = 60)$ , so the normal inference is not greatly affected in these cases because the bias is small.

For the combination  $(100, 60)$ , in order to avoid 0 observations in the cells we use (Woolf, 1955)'s adjusted method to have

$$\hat{\beta}_i = \log((a_i + 1/2)(d_i + 1/2)/(c_i + 1/2)(b_i + 1/2))$$

and  $\hat{\alpha}_i = \log(b_i + 1/2/a_i + 1/2)$ .

In order to compare our estimators  $\hat{\beta}_w$  with true weights and  $\hat{\hat{\beta}}_w$  with estimated weights to the Mantel-Haenszel estimator  $\hat{\beta}_{MH}$  and to see whether our estimators are more efficient, we compared the standard deviation of our estimators from 1000 replication with of Mantel-Haenszel estimator. From the results of different combinations of  $(m, n)$  and  $\beta_0$ , our estimators are almost the same efficient as Mantel-Haenszel estimator. Since  $\hat{\beta}_w$  is constructed by plugging MLE  $\hat{\beta}_i$  for  $i$ th group, we can also construct another empirical estimator of  $\beta$  by plugging in  $\hat{\beta}_w$ , we run the similar simulations under this new empirical estimator which introduced more positive bias. But the normality results hold for this new empirical estimator. It suggests that in practice, one can stay with the simpler empirical estimator by obtained by plugging MLE  $\hat{\beta}_i$  into the formula for the weight matrix. The plots show that  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  are approximately normal, with departures from normality in the extreme tails in logistic  $2 \times 2 \times m$  unbalanced setup.

Since when we introduce 1/2 adjustment into estimator we reduce bias, we tried simulation for  $(10, 200)$  in balanced setup and unbalanced setup  $(1/4, 3/4)$  and  $(10, 300)$  in unbalanced setup  $(1/3, 2/3)$ . Adding the 1/2 adjustment can reduce bias for  $\hat{\beta}_w - \beta_0$  and  $\hat{\hat{\beta}}_w - \beta_0$  a lot, especially for the combination  $(10, 200)$ .

Table 4.4: Simulated estimates of log odds ratio (standardized by true and estimated conditional standardized error) in  $2 \times 2 \times m$  unbalanced table.

	$(\hat{\beta}_w - \beta_0)$			$(\hat{\hat{\beta}}_w - \beta_0)$		
$(m, n)$	mean	std	95% CI for $(\hat{\beta}_w - \beta_0)$	mean	std	
<u><math>\beta_0 = 1</math>, unbalanced setup (1/4,3/4)</u>						
(10,200)	0.012	0.110	(0.005, 0.019)	0.005	0.107	
(20,400)	0.003	0.054	(-0.0003, 0.006)	-0.0002	0.053	
(100,60)	0.001	0.063	(-0.003, 0.004)	0.023	0.057	
<u><math>\beta_0 = 1</math>, unbalanced setup (1/3,2/3)</u>						
(10,300)	0.010	0.084	(0.005, 0.015)	0.004	0.083	
(30,600)	0.002	0.034	(-0.0003, 0.004)	-0.001	0.034	
(100,60)	-0.001	0.059	(-0.004, 0.003)	-0.032	0.055	
<u><math>\beta_0 = 0.5</math>, unbalanced setup (1/3,2/3)</u>						
(10,300)	0.005	0.082	(-0.0004, 0.010)	0.002	0.081	
(30,600)	0.002	0.033	(0.0004, 0.004)	0.0003	0.032	
(100,60)	-0.003	0.056	(-0.007, 0.0004)	-0.016	0.052	
<u><math>\beta_0 = 0.5</math>, unbalanced setup (1/4,3/4)</u>						
(10,200)	0.006	0.109	(-0.0005, 0.013)	0.003	0.106	
(20,400)	0.005	0.054	(0.002, 0.008)	0.004	0.054	
(100,60)	-0.001	0.064	(-0.003, 0.005)	-0.009	0.058	

Table 4.5: Simulated standardized estimates of log odds ratio (standardized by true and estimated conditional standardized error) in  $2 \times 2 \times m$  unbalanced table .

$(\hat{\beta}_w - \beta_0)/s.e.$				$(\hat{\beta}_w - \beta_0)/\widehat{s.e.}$			
$(m, n)$	mean	std	95% CI for $(\hat{\beta}_w - \beta_0)$	mean	std		
<u><math>\beta_0 = 1</math>, unbalanced setup (1/4,3/4)</u>							
(10,200)	0.110	1.005	(-0.002, 0.012)	0.041	0.969		
(20,400)	0.056	0.985	(-0.004, 0.003)	-0.006	0.965		
(100,60)	0.008	1.015	(-0.027, -0.020)	-0.372	0.899		
<u><math>\beta_0 = 1</math>, unbalanced setup (1/3,2/3)</u>							
(10,300)	0.122	1.011	(-0.001, 0.009)	0.047	0.992		
(30,600)	0.054	1.004	(-0.003, 0.001)	-0.035	0.996		
(100,60)	-0.012	1.025	(-0.035, -0.028)	-0.542	0.939		
<u><math>\beta_0 = 0.5</math>, unbalanced setup (1/3,2/3)</u>							
(10,300)	0.058	1.018	(-0.003, 0.007)	0.025	1.001		
(30,600)	-0.049	0.996	(-0.002, 0.002)	0.010	0.986		
(100,60)	-0.055	1.005	(-0.019, -0.013)	-0.283	0.924		
<u><math>\beta_0 = 0.5</math>, unbalanced setup (1/4,3/4)</u>							
(10,200)	0.059	1.018	(-0.003, 0.010)	0.029	0.980		
(20,400)	0.094	1.020	(0.0002, 0.007)	0.064	1.002		
(100,60)	0.017	1.066	(-0.013, -0.006)	-0.149	0.947		

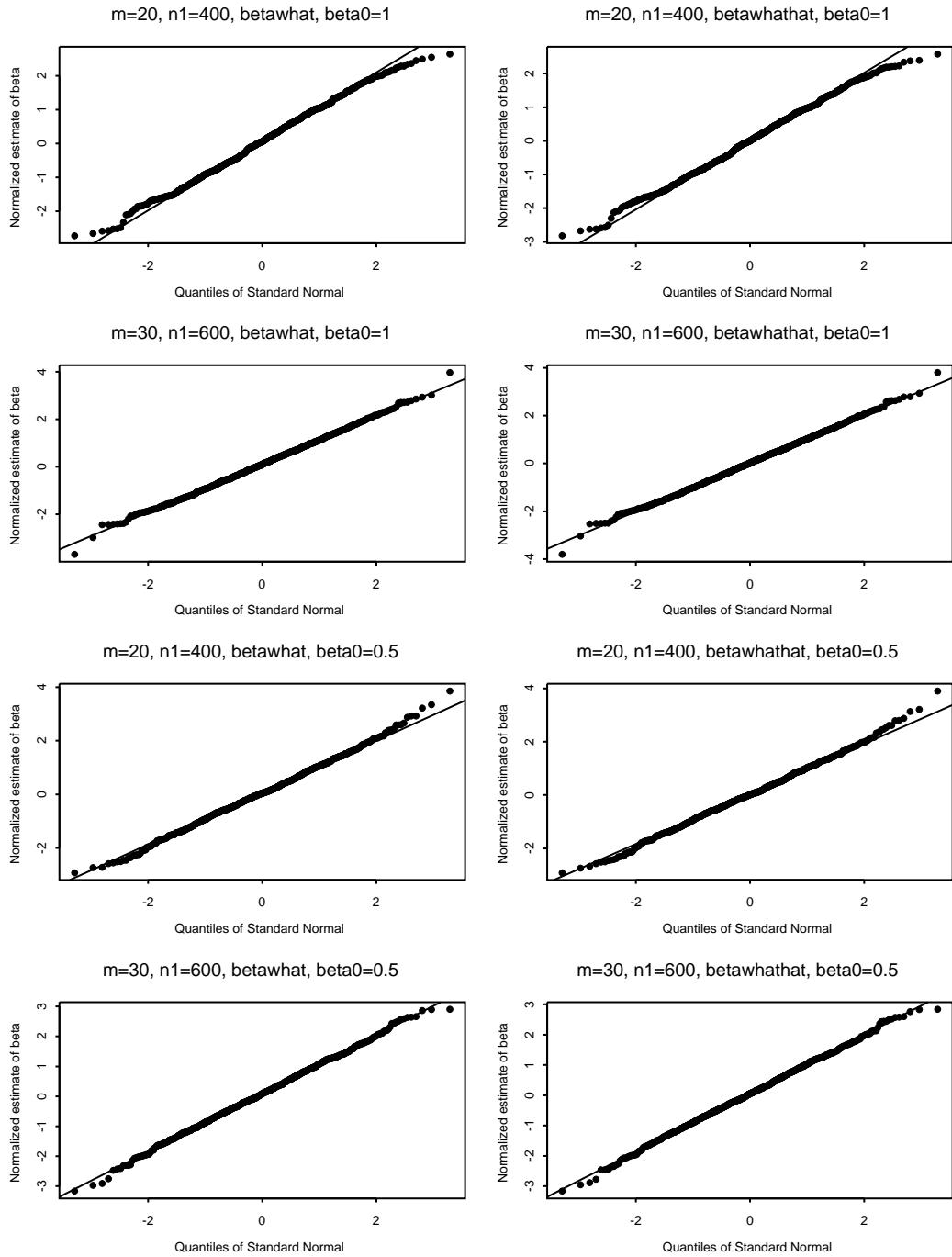


Figure 4.2: Q-Q plots for  $(\hat{\beta}_w - \beta_0)$  and  $(\hat{\hat{\beta}}_w - \beta_0)$  standardized by true and estimated conditional standard error, for various values of  $(m, n)$  in logistic  $2 \times 2 \times m$  unbalanced setup where  $n1=600, 400$  for  $(1/3, 2/3), (1/4, 3/4)$  setups respectively .

### 4.2.2 Conditional Convergence in Distribution

Recall that Theorem 3.1.2 and Theorem 3.1.3 gave conditional convergence in distribution of the normalized  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  to the  $N(\mathbf{0}, \mathbf{I})$  distribution. We performed a limited study to examine this conditional convergence.

We repeated the simulations designed for the logistic  $2 \times 2 \times m$  table, but with the following modification: we generated 10 realizations of i.i.d  $\text{Unif}(-0.8, 0.8)$  of  $\alpha_0$ . For each realization of  $\alpha_0$  we generated 1000 replications of  $\mathbf{y}$  with the same  $\alpha_0$ . This was performed only for sample sizes  $m = 10$ ,  $n = 200$  for  $\beta_0=1$  or 0.5.

From the following tables Table 4.6, Table 4.7, Table 4.8 and Table 4.9, we found that each realization of  $\alpha_0$ , the normalized  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  had Monte Carlo mean zero and variances near 1 in the case  $m = 10$ ,  $n_1 = 200$ . The Kolmogorov-Smirnov and Shapiro-Wilk tests all indicated no significant departures from normality. These findings are very similar to those which describe the unconditional distribution.

Table 4.6: Simulated estimates of logistic odds ratio for fixed realizations of uniformly distributed random effects ( $m = 10, n_1 = 200$ ) and  $\beta_0 = 0.5$ .

Realization	$(\hat{\beta}_w - \beta_0)$			$\hat{\beta}_w - \beta_0$		
	mean	std	95% CI for $\hat{\beta}_w - \beta_0$	mean	std	
1	0.007	0.097	(0.001, 0.013)	0.000	0.096	
2	0.002	0.091	(-0.005, 0.006)	-0.005	0.090	
3	0.004	0.100	(-0.002, 0.010)	-0.001	0.100	
4	0.003	0.093	(-0.003, 0.009)	-0.004	0.092	
5	0.011	0.095	(0.005, 0.017)	0.004	0.093	
6	0.002	0.092	(-0.004, 0.008)	-0.003	0.090	
7	0.003	0.092	(-0.003, 0.009)	-0.003	0.091	
8	0.006	0.098	(0.000, 0.012)	-0.001	0.097	
9	0.006	0.092	(0.000, 0.012)	0.005	0.092	
10	0.003	0.095	(-0.003, 0.009)	-0.003	0.093	

Table 4.7: Simulated standardized estimate of logistic odds ratio with fixed realizations of uniformly distributed random effects ( $m = 10$ ,  $n_1 = 200$ ) and  $\beta_0 = 0.5$ .

Realization	$(\hat{\beta}_w - \beta_0)/s.e$			$(\hat{\beta}_w - \beta_0)/\widehat{s.e.}$		
	mean	std	95% CI for $(\hat{\beta}_w - \beta_0)$	mean	std	
1	0.071	1.030	(-0.005, 0.006)	-0.003	1.000	
2	0.002	0.980	(-0.011, 0.000)	-0.062	0.964	
3	0.045	1.089	(-0.008, 0.005)	-0.019	1.069	
4	0.030	1.000	(-0.009, 0.002)	-0.041	0.970	
5	0.116	1.000	(-0.002, 0.010)	0.037	0.975	
6	0.024	1.000	(-0.009, 0.002)	-0.039	0.980	
7	0.032	0.991	(-0.008, 0.003)	-0.033	0.974	
8	0.066	1.040	(-0.006, 0.006)	-0.009	1.016	
9	0.068	1.007	(-0.005, 0.006)	0.002	0.993	
10	0.033	1.018	(-0.009, 0.003)	-0.035	0.993	

Table 4.8: Simulated estimates of logistic odds ratio with fixed realizations of uniformly distributed random effects ( $m = 10, n = 200$ ) and  $\beta_0 = 1$ .

Realization	$(\hat{\beta}_w - \beta_0)$			$\hat{\beta}_w - \beta_0$		
	mean	std	95% CI for $\hat{\beta}_w - \beta_0$	mean	std	
1	0.015	0.097	(0.009, 0.021)	0.002	0.095	
2	0.013	0.102	(0.007, 0.020)	-0.003	0.099	
3	0.010	0.100	(0.003, 0.016)	-0.004	0.099	
4	0.013	0.103	(0.008, 0.020)	0.000	0.102	
5	0.014	0.094	(0.009, 0.020)	0.000	0.092	
6	0.011	0.094	(0.005, 0.017)	0.000	0.093	
7	0.013	0.099	(0.007, 0.020)	-0.001	0.097	
8	0.011	0.097	(0.005, 0.017)	-0.003	0.095	
9	0.009	0.099	(0.003, 0.015)	-0.004	0.098	
10	0.011	0.098	(0.005, 0.017)	-0.005	0.096	

Table 4.9: Simulated standardized estimate of logistic odds ratio with fixed realizations of uniformly distributed random effects ( $m = 10$ ,  $n_1 = 200$ ) and  $\beta_0 = 1$ .

Realization	$(\hat{\beta}_w - \beta_0)/s.e$			$(\hat{\hat{\beta}}_w - \beta_0)/\widehat{s.e.}$		
	mean	std	95% CI for $(\hat{\hat{\beta}}_w - \beta_0)$	mean	std	
1	0.153	1.021	(-0.003, 0.008)	0.019	0.996	
2	0.138	1.022	(-0.009, 0.003)	-0.039	0.987	
3	0.099	1.033	(-0.010, 0.002)	-0.050	1.008	
4	0.144	1.064	(-0.007, 0.006)	-0.011	1.038	
5	0.149	0.968	(-0.005, 0.006)	-0.003	0.940	
6	0.119	1.001	(-0.006, 0.006)	-0.008	0.982	
7	0.139	1.025	(-0.007, 0.006)	-0.013	0.999	
8	0.109	0.994	(-0.009, 0.003)	-0.039	0.969	
9	0.095	1.031	(-0.010, 0.002)	-0.051	1.011	
10	0.111	1.003	(-0.010, 0.001)	-0.053	0.977	

### 4.3 Analysis of real data for a $2 \times 2 \times 22$ table

We used the data from Yusuf et al. [36] which has 22 clinical trials of beta-blockers for reducing mortality after myocardial infarction. The data structure is the following: Clinical trial  $j$ , where  $1 \leq j \leq 22$  (in the series to be considered for meta-analysis), involves the use of  $n_{0j}$  subjects in the control group and  $n_{1j}$  in the treatment group, giving rise to  $b_j$  and  $d_j$  deaths in the control and treatment groups, respectively. Then the usual sampling models involve two independent binomial distributions with probability of death  $p_{0j}$  and  $p_{1j}$ , respectively. We concentrate on estimating the common log odds ratio which we label  $\beta$ . Here we assume the model  $\text{logit}(y_{ij} = 1|x_{ij}, \alpha_i) = \alpha_i + x_{ij}\beta$  where  $\alpha_i$  is a random effect which may represent variation between clinical trials. We use this real data to calculate our estimator, the Mantel-Haenszel estimator and their variance estimators, respectively. Here we need to point out that the difference between our asymptotic setting and that Woolf (1955) is that we allow  $m \rightarrow \infty$ , but in Woolf's setting,  $m$  is fixed. The following table summarizes the result.

Table 4.10: Summary of simulation of our estimator and Mantel-Haenszel estimator of log odds ratio.

$\hat{\beta}_w$	$\hat{\beta}_{MH}$	95% CI for $\hat{\beta}_w$	95% CI for $\hat{\beta}_{MH}$	$\text{Var}(\hat{\beta}_w)$	$\text{Var}(\hat{\beta}_{MH})$
-0.260	-0.261	(-0.359, -0.161 )	(-0.372, -0.150)	0.00253	0.00249

First, we use Woolf's test in Splus, Breslow-Day test and Likelihood test in SAS to test homogeneity of odds ratios whose p-values are 0.3595, 0.3149 and 0.3118 respectively. They all suggest that the common odds ratio model is appropriate for combining the 22 clinical trials.

The above empirical variance estimator of  $\text{Var}[\hat{\beta}_{MH}]$  is based on Breslow, Greenland and Robins [25]. We bootstrapped the data as follows: for each  $i = 1, \dots, 22$  and  $j = 0, 1$ , we generated  $Y_{ij}^* \sim \text{Binomial}(n_{ij}, p_{ij})$ , where  $p_{ij} = Y_{ij}/n_{ij}$  is the sample proportion of successes in table  $i$  with  $x = j$ . Based on this bootstrap sample, new estimators  $\hat{\beta}_w^*$  and  $\hat{\beta}_{MH}^*$  were calculated. This process were repeated 1000 times, and the sample variance of these bootstrap replicates was used to estimate  $\text{Var}[\hat{\beta}_w]$  and  $\text{Var}[\hat{\beta}_{MH}]$ .

We obtained 0.002476 for  $\hat{\beta}_w$  and 0.002603 for  $\hat{\beta}_{MH}$ . These results agree with Table 4.10 and we can see that our estimator  $\hat{\beta}_w$  is very close to the Mantel-Haenszel estimator. The Mantel-Haenszel estimator is consistent in the sparse data case, unlike our estimator. The logic of our estimator can be extended to other types of GLMM where  $m \rightarrow \infty$  and  $m/\min_i n_i \rightarrow 0$ . No corresponding extension for Mantel-Haenszel seems available in the literature.

## Chapter 5

### Case 2 Results.

This Chapter establishes asymptotic normality results for parameter estimates in certain versions of Case 2, when  $m/n \not\rightarrow 0$ . In Section 5.1 we review Jiang's [14] results on consistency. In Section 5.2 we state and prove our asymptotic normality theorem, and in Section 5.3 we focus on a random effect logistic regression model.

Our results focus on the Maximum Conditional Likelihood Estimates (MCLEs). These estimates are based on maximizing the likelihood function conditional on the estimable random effects after integrating out the unestimable random effects. The estimates derived from this likelihood function were called MCLEs by Jiang [14].

A basic technique here is reparameterization, because, conditionally, the individual effects may not be identifiable. To illustrate this method, a special case is considered. The analysis of the more general setting will be similar. In case 2 the dimension of  $\beta$  is fixed.

#### 5.1 Case 2 consistency results of Jiang [14].

This section reviews the results on MCLE's obtained by Jiang [14]. A key tool in this analysis is to reparameterize the model to address identifiability.

**Lemma 5.1.1.** *There is a map  $\beta \mapsto \tilde{\beta}$ ,  $\alpha \mapsto \tilde{\alpha}$  such that*

(i)  $\mathbf{X}\beta + \mathbf{Z}\alpha = \tilde{\mathbf{X}}\tilde{\beta} + \tilde{\mathbf{Z}}\tilde{\alpha}$ , where  $(\tilde{\mathbf{X}} \ \tilde{\mathbf{Z}})$  is a known matrix of full column rank.

(ii)  $\tilde{z}_i = \tilde{z}_{*j}$ ,  $i \in S_j$  for some known vector  $\tilde{z}_{*j}$ , where  $\tilde{z}_j^t$  is the  $i$ th row of  $\tilde{\mathbf{Z}}$  and  $S_j$  is defined as below.

Vector  $u_i^t$  is the  $i$ th row of  $\mathbf{U}$  where  $\mathbf{U}$  is standard in the sense that consists of 0's and 1's and there is at least one 1 in each column and exactly one 1 in each row. Vector  $e_{M,j}$  is the  $M$ -dimensional vector whose  $j$ th component is 1 and other components are 0. Let  $S_j = \{1 \leq i \leq N : u_i = e_{M,j}\}$ , and  $y^{(j)} = (y_i)_{i \in S_j}$ ,  $1 \leq j \leq M$ . Suppose that  $\zeta_1, \dots, \zeta_M$  are independent with common distribution  $\nu(\cdot/\tau)/\tau$ , where  $\nu(\cdot)$  is a known density function and  $\tau > 0$  is an unknown scale parameter. Furthermore, we assume that there are no random effects nested within  $\zeta$ . In notation, this means that  $z_i = z_{*j}$ ,  $i \in S_j$ ,  $1 \leq j \leq M$ , where  $z_{*j} = (z_{*jk})_{1 \leq k \leq l}$ . By Jiang's [14] Lemma 4.1.1, we have

$$\boldsymbol{\eta} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{Z}}\tilde{\boldsymbol{\alpha}} + \mathbf{U}\zeta \quad (5.1)$$

Let  $\boldsymbol{\varphi} = (\tilde{\boldsymbol{\beta}}, \tau)$ ,  $\boldsymbol{\theta} = (\tilde{\boldsymbol{\alpha}}, \boldsymbol{\varphi})$ . By (2.23) distribution of  $\mathbf{y}$  given only  $\boldsymbol{\alpha}$  is

$$f(y|\boldsymbol{\theta}) = \prod_{j=1}^M f(y^{(j)}|\boldsymbol{\theta}) \quad (5.2)$$

and it is easy to show that  $f(y^{(j)}|\boldsymbol{\theta}) = g_j(z_{*j}^t \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tau)$ , where

$$g_j(s) = E \left[ \prod_{i \in S_j} f(y_i | s_1 + x_i^t s_{(2)} + s_{r+2} \xi) \right]$$

with  $s_{(2)} = (s_2, \dots, s_{r+1})$  and  $r$  is the dimension of  $\tilde{\boldsymbol{\beta}}$ . Note that  $r \leq p$ . Let  $n$  be the dimension of  $\tilde{\boldsymbol{\alpha}}$  and the expectation is with respect to the distribution of  $\xi$ ,

$$h_j(s) = \log g_j(s), l_C(\boldsymbol{\theta}) = \log f(y|\boldsymbol{\theta}) \text{ and } l_{C,j}(\boldsymbol{\theta}) = \log f(y^{(j)}|\boldsymbol{\theta}) = h_j(z_{*j}^t \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tau).$$

Then

$$l_c(\boldsymbol{\theta}) = \sum_{j=1}^M l_{c,j}(\boldsymbol{\theta}). \quad (5.3)$$

Let  $\mathbf{Z}_*$  be the matrix whose  $j$ th row is  $z_{*j}^t$ ,  $1 \leq j \leq M$ . Let  $\boldsymbol{\varphi}_0$  and  $\boldsymbol{\theta}_0$  be the vectors corresponding to the true parameters and realization of random effects.

Define  $s_{M,k}^{(l)} = \sum_{j=1}^M |\tilde{z}_{*jk}|^l$ ,  $l = 1, 2, \dots$ ,  $t_{M,k} = \sum_{j=1}^M \sum_{l \neq k} |\tilde{z}_{*jk} \tilde{z}_{*jl}|$ ,

$$\mathbf{H}_j = (\partial^2 h_j / \partial s^2) \Big|_{s_1=\tilde{z}_{*j}^t \tilde{\alpha}, s_{(2)}=\tilde{\beta}, s_{r+2}=\tau}, \quad (5.4)$$

and

$$\mathbf{A}_2 = \sum_{j=1}^M \begin{pmatrix} \tilde{z}_{*j} & 0 \\ 0 & \mathbf{I}_{r+1} \end{pmatrix} (\mathbf{H}_j(\boldsymbol{\theta}_0) - E(\mathbf{H}_j(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_0)) \begin{pmatrix} \tilde{z}_{*j}^t & 0 \\ 0 & \mathbf{I}_{r+1} \end{pmatrix} \quad (5.5)$$

where  $\mathbf{I}_l$  represents the  $l$ -dimensional identity matrix. Define

$$\mathbf{G} = \begin{pmatrix} \tilde{\mathbf{Z}}_*^t \tilde{\mathbf{Z}}_* & 0 \\ 0 & M \mathbf{I}_{r+1} \end{pmatrix} = \sum_{j=1}^M \begin{pmatrix} \tilde{z}_{*j} \tilde{z}_{*j}^t & 0 \\ 0 & \mathbf{I}_{r+1} \end{pmatrix}, \quad (5.6)$$

$$\lambda_M(\boldsymbol{\theta}) = \min_{1 \leq j \leq M} \lambda_{\min} \left( \text{Var} \left( (\partial h_j / \partial s) |_{(\tilde{z}_{*j}^t \tilde{\alpha}, \tilde{\beta}, \tau)} | \boldsymbol{\theta} \right) \right), \quad (5.7)$$

and  $\lambda_M = \lambda_M(\boldsymbol{\theta}_0)$ . Let  $\xi_j^{(l)}(\boldsymbol{\theta}) = (\partial^l h_j / \partial^l s_1)(\tilde{z}_{*j}^t \tilde{\alpha}, \tilde{\beta}, \tau)$ ,  $l = 1, 2$ ,

$$\begin{aligned} V_k^{(1)}(\varepsilon) &= \max_{1 \leq j \leq M} E \left( |\xi_j^{(1)}(\boldsymbol{\theta}_0)| 1_{(\tilde{z}_{*jk} \xi_j^{(1)}(\boldsymbol{\theta}_0)) > 1/2\varepsilon} | \boldsymbol{\theta}_0 \right), \\ V_k^{(2)}(\varepsilon) &= \max_{1 \leq j \leq M} E \left( |\xi_j^{(2)}(\boldsymbol{\theta}_0) - E(\xi_j^{(2)}(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_0)| 1_{(\tilde{z}_{*jk}^2 | \dots | > 1/2\varepsilon)} | \psi_0 \right). \end{aligned}$$

**Theorem 5.1.2.** Suppose:

(i) the conditional densities  $f(y^{(j)}|\boldsymbol{\theta})$ ,  $1 \leq j \leq M$ , are with respect to a common measure  $\mu$  and have common support, and the first and second partial derivatives of  $\int f(y^j|\boldsymbol{\theta}) d\mu$  with respect to components of  $\boldsymbol{\theta}$  exist and can be taken under the integral sign.

(ii)  $h_j(s)$ ,  $1 \leq j \leq M$ , are three times differentiable and there exist  $\delta, B > 0$

such that

$$\max_{1 \leq j \leq M} \left\{ \left( \max_{1 \leq u \leq r+2} |(\partial^2 h_j / \partial s_1 \partial s_u)| \right) \vee \left( \max_{1 \leq u, v, w \leq r+2} |(\partial^3 h_j / \partial s_u \partial s_v \partial s_w)| \right) \right\} \leq B \quad (5.8)$$

for all  $\boldsymbol{\theta}$  such that  $\|\varphi - \varphi_0\| < \delta$ .

(iii)  $\tilde{\mathbf{Z}}_{*k} \neq 0$ ,  $1 \leq k \leq n$ , where  $\tilde{\mathbf{Z}}_{*k}$  is the  $k$ th column of  $\tilde{\mathbf{Z}}_*$ , and the following are bounded:

$$\|\tilde{\mathbf{Z}}_*\|_\infty, \quad \max_{1 \leq k \leq n} \left( \frac{s_{M, k}^{(1)}}{s_{M, k}^{(2)}} \right), \quad \max_{1 \leq k \leq n} \left( \frac{s_{M, k}^{(2)}}{s_{M, k}^{(4)}} \right), \quad \max_{1 \leq k \leq n} \left( \frac{s_{M, k}^{(4)}}{s_{M, k}^{(2)}} \right)$$

and

$$\max_{1 \leq j \leq M} \left( |\tilde{z}_{*j}|^2 E \left( \left( \frac{\partial h_j}{\partial s_1} \Big|_{s_0} \right)^2 \right) \vee \left( \max_{2 \leq u \leq r+2} E \left( \left( \frac{\partial h_j}{\partial s_u} \Big|_{s_0} \right)^2 \right) \right), \quad (5.9)$$

(iv)  $\lambda_M > 0$ , and there is a sequence  $\rho_M$  such that  $0 < \rho_M \leq \lambda_M \wedge 1$ , and the following  $\rightarrow 0$  in probability:

$$\lambda_{\max}(\mathbf{G}^{-1/2} \mathbf{A}_2 \mathbf{G}^{-1/2}) / \rho_M, \quad \max_{1 \leq k \leq n} \left( \frac{t_{M, k}}{s_{M, k}^{(2)}} \right) / \rho_M, \quad \left( \frac{n}{M} \right) / \rho_M^4$$

and

$$\max_{l=1, 2} (\log n / \rho_M^{2l}) \min_{1 \leq k \leq n} s_{M, k}^{(6-2l)} \vee \left( n \max_{1 \leq k \leq n} V_k^{(3-l)}(\rho_M^l) / \rho_M^l \right).$$

Then, with probability approaching 1, there is a sequence  $\hat{\boldsymbol{\theta}}$  satisfying  $(\partial l_C / \partial \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}) = 0$  and  $\max_i |\hat{\theta}_i - \theta_{i0}| = o_p(\rho_M)$ .

Note. In fact, it is seen from the proof of the theorem that  $\max_i |\hat{\varphi}_i - \varphi_{i0}| = o_p(\rho_M^2)$ .

Consider a special case in which there is only one random effect factor. In such a case, one may integrate out all the random effects, if necessary. The resulting MCL estimates are the maximum likelihood estimates for the fixed parameters. We have the following.

**Corollary 5.1.3.** *Suppose that in (5.1)  $\alpha = 0$  (i.e., there are no random effects besides  $\zeta$ ), and that:*

- (1) *Part (i) of Theorem 4.1.2 holds with  $\boldsymbol{\theta}$  replaced by  $\boldsymbol{\varphi}$ .*
- (2)  *$h_j(\boldsymbol{\varphi})$ ,  $1 \leq j \leq M$  are three times differentiable and there exists  $\delta$ ,  $B > 0$  such that*

$$\max_{1 \leq j \leq M} \sup_{\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_0\| \leq \delta} |(\text{any third derivative of } h_j)(\boldsymbol{\varphi})| \leq B.$$

- (3)  $\lambda_M = \min_{1 \leq j \leq M} \lambda_{\min}(Var((\partial h_j / \partial \boldsymbol{\varphi})(\boldsymbol{\varphi}_0)) | \boldsymbol{\varphi}_0) > 0$  and

$$\frac{1}{M^2(\lambda_M \wedge 1)^2} \sum_{j=1}^M E(\|\mathbf{H}_j(\boldsymbol{\varphi}_0) - E\mathbf{H}_j(\boldsymbol{\varphi}_0)\|_R^2) \rightarrow 0,$$

where  $\mathbf{H}_j(\boldsymbol{\varphi}) = \partial^2 h_j / \partial \boldsymbol{\varphi}^2$ . Then, with probability approaching 1, there is a sequence

## 5.2 The Simple case ( $\boldsymbol{\alpha} = 0$ ).

In this case, we have the classical likelihood function after integrating out the unestimable random effects and do not need reparameterization, since we only have fixed effects and the dispersion parameter of the random effect distribution.

Recall the likelihood function is defined between (5.1) and (5.4), Then

$$\begin{aligned}
& \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\hat{\boldsymbol{\varphi}}} \\ (\partial l_C(\boldsymbol{\varphi})/\partial \tau)|_{\hat{\boldsymbol{\varphi}}} \end{pmatrix} \\
&= \begin{pmatrix} (\partial^2 l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t)|_{\boldsymbol{\varphi}^*} & (\partial^2 l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta} \partial \tau)|_{\boldsymbol{\varphi}^*} \\ (\partial^2 l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta}^t \partial \tau)|_{\boldsymbol{\varphi}^*} & (\partial^2 l_C(\boldsymbol{\varphi})/\partial \tau^2)|_{\boldsymbol{\varphi}^*} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\tau} - \tau_0 \end{pmatrix} \\
&+ \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi})/\partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix}
\end{aligned}$$

where  $\boldsymbol{\varphi}^*$  is between  $\hat{\boldsymbol{\varphi}}$  and  $\boldsymbol{\varphi}_0$ .

Then

$$(-\mathbf{H}) \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\tau} - \tau_0 \end{pmatrix} = \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi})/\partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix}$$

where

$$\mathbf{H} = \begin{pmatrix} (\partial^2 l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t)|_{\boldsymbol{\varphi}^*} & (\partial^2 l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta} \partial \tau)|_{\boldsymbol{\varphi}^*} \\ (\partial^2 l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta}^t \partial \tau)|_{\boldsymbol{\varphi}^*} & (\partial^2 l_C(\boldsymbol{\varphi})/\partial \tau^2)|_{\boldsymbol{\varphi}^*} \end{pmatrix}. \quad (5.10)$$

Let

$$\mathbf{C} = \text{Cov} \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi})/\partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix} \quad (5.11)$$

Then

$$\begin{aligned}
& \mathbf{C}^{-\frac{1}{2}} (-\mathbf{H}) \mathbf{C}^{-\frac{1}{2}} \mathbf{C}^{\frac{1}{2}} \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\tau} - \tau_0 \end{pmatrix} \\
&= \mathbf{C}^{-\frac{1}{2}} \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi})/\partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix}.
\end{aligned} \quad (5.12)$$

**Theorem 5.2.1.** Suppose we have the following conditions:

(1) For any  $\delta > 0$ , assume  $\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H})\mathbf{C}^{-\frac{1}{2}}$  is a positive definite matrix and

$$P(\|\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H})\mathbf{C}^{-\frac{1}{2}} - \mathbf{I}\| < \delta) \rightarrow 1$$

(2)

$$M\|\mathbf{C}^{-1}\| = O(1)$$

(3)

$$E\{\|\mathbf{G}_j\|\psi(\|\mathbf{G}_j\|)\} \leq K_2$$

where  $K_2$  is a positive constant,  $\psi(t)$  is a positive nondecreasing function mapping from  $[0, \infty)$  to  $[0, \infty)$  such that  $\lim_{t \rightarrow \infty} \psi(t) = \infty$  and  $t\psi(t)$  is a convex function, and

$$\mathbf{G}_j = \begin{pmatrix} ((\partial h_j / \partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0}) ((\partial h_j / \partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0})^T & ((\partial h_j / \partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0}) ((\partial h_j / \partial \tau)|_{\boldsymbol{\varphi}_0}) \\ ((\partial h_j / \partial \tau)|_{\boldsymbol{\varphi}_0}) ((\partial h_j / \partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0})^T & ((\partial h_j / \partial \tau)|_{\boldsymbol{\varphi}_0})^2 \end{pmatrix}. \quad (5.13)$$

Then

$$\mathbf{C}^{\frac{1}{2}} \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\tau} - \tau_0 \end{pmatrix} \xrightarrow{D} N(\mathbf{0}, \mathbf{I}). \quad (5.14)$$

Proof:

Recall (5.11), (5.12). By condition (1), we have

$$\begin{aligned} & \mathbf{C}^{\frac{1}{2}} \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\tau} - \tau_0 \end{pmatrix} \\ &= (\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H})\mathbf{C}^{-\frac{1}{2}})^{-1} \mathbf{C}^{-\frac{1}{2}} \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi}) / \partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi}) / \partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix} \end{aligned}$$

Then

$$\left\| \mathbf{C}^{-\frac{1}{2}} \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi})/\partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix} - \mathbf{C}^{\frac{1}{2}} \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\tau} - \tau_0 \end{pmatrix} \right\| \quad (5.15)$$

$$\leq \|(\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H})\mathbf{C}^{-\frac{1}{2}})^{-1} - \mathbf{I}\| \left\| \mathbf{C}^{-\frac{1}{2}} \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi})\partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix} \right\| \quad (5.16)$$

$$= \|(\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H})\mathbf{C}^{-\frac{1}{2}})^{-1} - \mathbf{I}\| O_p(1). \quad (5.17)$$

The second factor of (5.16) is  $O_p(1)$  from the following argument:

$$\begin{aligned} & \mathbf{P} \left( \left\| \mathbf{C}^{-\frac{1}{2}} \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi})\partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix} \right\| < c_4 \right) \\ & \geq 1 - (1/c_4^2) E \left( \left\| \mathbf{C}^{-\frac{1}{2}} \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi})\partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix} \right\|^2 \right)^2 \\ & = 1 - (1/c_4^2) \text{tr} \left( \mathbf{C}^{-1} \text{Cov} \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi})\partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix} \right) = 1 - (p+1)/c_4^2 \end{aligned}$$

By condition (1) of Theorem 5.2.1, We have

$$\begin{aligned} & \|(\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H})\mathbf{C}^{-\frac{1}{2}})^{-1} - \mathbf{I}\| \\ & \leq \|(\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H})\mathbf{C}^{-\frac{1}{2}})^{-1}\| \|\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H})\mathbf{C}^{-\frac{1}{2}} - \mathbf{I}\| \\ & = o_p(1). \end{aligned}$$

So  $\mathbf{C}^{-\frac{1}{2}} \begin{pmatrix} (\partial l_C(\boldsymbol{\varphi})/\partial \boldsymbol{\beta})|_{\boldsymbol{\varphi}_0} \\ (\partial l_C(\boldsymbol{\varphi})/\partial \tau)|_{\boldsymbol{\varphi}_0} \end{pmatrix}$  and  $\mathbf{C}^{\frac{1}{2}} \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\tau} - \tau_0 \end{pmatrix}$  have the same asymptotic distribution.

Let  $\mathbf{u}$  be a unit vector and

$$T_j = \mathbf{u}^T \mathbf{C}^{-\frac{1}{2}} \begin{pmatrix} (\partial h_j / \partial \beta)|_{\varphi_0} \\ (\partial h_j / \partial \tau)|_{\varphi_0} \end{pmatrix}$$

Recall (5.13). Then we have

$$T_j^2 = \mathbf{u}^T \mathbf{C}^{-\frac{1}{2}} \mathbf{G}_j \mathbf{C}^{-\frac{1}{2}} \mathbf{u}. \quad (5.18)$$

By conditions (1), (2) and (3)

$$\begin{aligned} & \sum_{j=1}^M E(T_j^2 I(|T_j| > \varepsilon)) \\ &= \sum_{j=1}^M E(T_j^2 I(T_j^2 > \varepsilon^2)) \\ &\leq \|\mathbf{C}^{-1}\| \sum_{j=1}^M E \left\{ \|\mathbf{G}_j\| I \left( \|\mathbf{G}_j\| > \frac{\varepsilon^2}{\|\mathbf{C}^{-1}\|} \right) \right\} \\ &\leq \|\mathbf{C}^{-1}\| \sum_{j=1}^M E(\|\mathbf{G}_j\| \psi\{\|\mathbf{G}_j\|\}) \left\{ \psi \left( \frac{\varepsilon^2}{\|\mathbf{C}^{-1}\|} \right) \right\}^{-1} \\ &\leq M \|\mathbf{C}^{-1}\| K_2 \left\{ \psi \left( \frac{\varepsilon^2}{\|\mathbf{C}^{-1}\|} \right) \right\}^{-1} = o_p(1). \end{aligned}$$

The Lindeberg condition is satisfied and by Slutsky's theorem, the conclusion of Theorem 5.2.1 follows.

### 5.3 The logistic model $\text{logit}P(y_{ijk} = 1 | b_{ij}) = \mu + b_{ij}$ .

We use the logistic model  $\text{logit}P(y_{ijk} = 1 | b_{ij}) = \mu + b_{ij}$  example to illustrate Theorem 5.2.1.

Suppose  $a_i = 0$  and  $b_{ij}$  are iid normal with  $1 \leq i \leq m_1$ ,  $1 \leq j \leq n$ . The binary responses  $y_{ijk}$  are conditionally independent with

$$\text{logit}P(y_{ijk} = 1|b) = \mu + b_{ij}, \quad 1 \leq k \leq r \quad (5.19)$$

where  $r$  is fixed.

Recall  $h_j(s)$  above (5.3). Then

$$h_{ij}(s) = \log E \exp\{(s_1 + s_2 \xi_{ij}) \sum_{k=1}^r y_{ijk} - r \log(1 + \exp(s_1 + s_2 \xi_{ij}))\},$$

where the  $\xi_{ij}$  are independently identically distributed with a standard normal distribution.

Recall (5.11). In this example we have

$$\mathbf{C} = \sum_i \sum_j \text{Var}\left(\frac{\partial h_{ij}}{\partial \varphi}|_{\varphi_0}\right). \quad (5.20)$$

By condition (iii) of Corollary 5.1.3,  $\mathbf{C}$  is a positive definite matrix. We need to verify  $mn_1\|\mathbf{C}^{-1}\| = O(1)$ , equivalently as

$$\inf_{\|\mathbf{u}\|=1} (\mathbf{u}^t \mathbf{C} \mathbf{u}) / mn_1 > 1/c_{mn_1}$$

where  $c_{mn_1}$  is a positive constant and  $\mathbf{u}$  is a unit vector.

By condition (iii) of Corollary 5.1.3 we have

$$\inf_{\|\mathbf{u}\|=1} (\mathbf{u}^t \mathbf{C} \mathbf{u}) / mn_1 > \lambda_M$$

where  $\lambda_M > 0$ . Then condition (2) of Theorem 5.2.1 is verified.

Recall (5.10), (5.11) and consider condition (1) of Theorem 5.2.1. In this

example we have

$$\begin{aligned} -\mathbf{H} &= \begin{pmatrix} -(\partial^2 l_C(\varphi)/\partial\mu^2)|_{\varphi^*} & -(\partial^2 l_C(\varphi)/\partial\mu\partial\tau)|_{\varphi^*} \\ -(\partial^2 l_C(\varphi)/\partial\mu\partial\tau)|_{\varphi^*} & -(\partial^2 l_C(\varphi)/\partial\tau^2)|_{\varphi^*} \end{pmatrix} \\ &= \mathbf{C} - \mathbf{C} - \mathbf{H}. \end{aligned} \quad (5.21)$$

Then

$$\begin{aligned} &\|\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H})\mathbf{C}^{-\frac{1}{2}} - \mathbf{I}\| \\ &= \|\mathbf{C}^{-\frac{1}{2}}(\mathbf{C} - \mathbf{H} - \mathbf{C})\mathbf{C}^{-\frac{1}{2}} - \mathbf{I}\| \\ &= \|\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H} - \mathbf{C})\mathbf{C}^{-\frac{1}{2}}\|. \end{aligned}$$

Let

$$-\mathbf{C} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

where

$$\begin{aligned} C_{11} &= -\sum_i \sum_j E\left(\frac{\partial h_{ij}}{\partial \mu}|_{\varphi_0}\right)^2, \\ C_{12} = C_{21} &= -\sum_i \sum_j E\left(\frac{\partial h_{ij}}{\partial \mu}|_{\varphi_0}\right)\left(\frac{\partial h_{ij}}{\partial \tau}|_{\varphi_0}\right), \\ C_{22} &= -\sum_i \sum_j E\left(\frac{\partial h_{ij}}{\partial \tau}|_{\varphi_0}\right)^2. \end{aligned}$$

Let

$$w_{ij}(\varphi_0, \xi_{ij}) = (\mu_0 + \tau_0 \xi_{ij}) \sum_k y_{ijk} - r \log(1 + \exp(\mu_0 + \tau_0 \xi_{ij})) \quad (5.22)$$

and

$$v_{ij}(\varphi_0, \xi_{ij}) = \sum_k y_{ijk} - \frac{r \exp(\mu_0 + \tau_0 \xi_{ij})}{1 + \exp(\mu_0 + \tau_0 \xi_{ij})}. \quad (5.23)$$

Since we have  $P(y_{ijk} = 1 | \mu, \tau \xi_{ij}) = \exp(\mu + \tau \xi_{ij}) / (1 + \exp(\mu + \tau \xi_{ij}))$ , then

$$\sum_k y_{ijk} | \mu_0, \tau_0 \xi_{ij} \sim \text{Binomial} \left( r, \frac{\exp(\mu_0 + \tau_0 \xi_{ij})}{1 + \exp(\mu_0 + \tau_0 \xi_{ij})} \right).$$

Thus

$$\frac{\partial h_{ij}}{\partial \mu} |_{\varphi_0} = \frac{E(\exp(w_{ij}(\varphi_0, \xi_{ij}))) v_{ij}(\varphi_0, \xi_{ij})}{E(\exp(w_{ij}(\varphi_0, \xi_{ij})))} \quad (5.24)$$

$$\frac{\partial h_{ij}}{\partial \tau} |_{\varphi_0} = \frac{E \{ \xi_{ij}(\exp(w_{ij}(\varphi_0, \xi_{ij}))) v_{ij}(\varphi_0, \xi_{ij}) \}}{E(\exp(w_{ij}(\varphi_0, \xi_{ij})))} \quad (5.25)$$

$$-\frac{\partial^2 h_{ij}}{\partial \mu^2} |_{\varphi_0} = -\frac{1}{E(\exp(w_{ij}(\varphi_0, \xi_{ij})))} \left\{ E(\exp(w_{ij}(\varphi_0, \xi_{ij}))) (v_{ij}^2(\varphi_0, \xi_{ij})) \right.$$

$$- \text{Var} \left( \sum_k y_{ijk} | \mu_0, \tau_0 \xi_{ij} \right) \Bigg\}$$

$$+ \left\{ \frac{E(\exp(w_{ij}(\varphi_0, \xi_{ij}))) v_{ij}(\varphi_0, \xi_{ij})}{E(\exp(w_{ij}(\varphi_0, \xi_{ij})))} \right\}^2$$

$$-\frac{\partial^2 h_{ij}}{\partial \tau^2} |_{\varphi_0} = -\frac{1}{E(\exp(w_{ij}(\varphi_0, \xi_{ij})))} \left\{ E(\xi_{ij}^2 \exp(w_{ij}(\varphi_0, \xi_{ij}))) v_{ij}^2(\varphi_0, \xi_{ij}) \right.$$

$$- \text{Var} \left( \sum_k y_{ijk} | \mu_0, \tau_0 \xi_{ij} \right) \Bigg\}$$

$$+ \left\{ \frac{E \{ \xi_{ij}(\exp(w_{ij}(\varphi_0, \xi_{ij}))) v_{ij}(\varphi_0, \xi_{ij}) \}}{E(\exp(w_{ij}(\varphi_0, \xi_{ij})))} \right\}^2$$

$$-\frac{\partial^2 h_{ij}}{\partial \tau \partial \mu} |_{\varphi_0} = -\frac{1}{E(\exp(w_{ij}(\varphi_0, \xi_{ij})))} \left\{ E(\xi_{ij}(\exp(w_{ij}(\varphi_0, \xi_{ij}))) v_{ij}^2(\varphi_0, \xi_{ij}) \right.$$

$$- \text{Var} \left( \sum_k y_{ijk} | \mu_0, \tau_0 \xi_{ij} \right) \Bigg\}$$

$$+ \left( \frac{E(\exp(w_{ij}(\varphi_0, \xi_{ij}))) v_{ij}(\varphi_0, \xi_{ij})}{\{E(\exp(w_{ij}(\varphi_0, \xi_{ij})))\}^2} \right)$$

$$\times (E \{ \xi_{ij}(\exp(w_{ij}(\varphi_0, \xi_{ij}))) v_{ij}(\varphi_0, \xi_{ij}) \}).$$

Since we have the classical likelihood function in this example, then

$$-E(\frac{\partial h_{ij}}{\partial \mu} |_{\varphi_0})^2 - E(\frac{\partial^2 h_{ij}}{\partial \mu^2} |_{\varphi_0}) = 0, \quad -E(\frac{\partial h_{ij}}{\partial \mu} |_{\varphi_0})(\frac{\partial h_{ij}}{\partial \tau} |_{\varphi_0}) - E(\frac{\partial^2 h_{ij}}{\partial \mu \partial \tau} |_{\varphi_0}) = 0,$$

$$-E(\frac{\partial h_{ij}}{\partial \tau} |_{\varphi_0})^2 - E(\frac{\partial^2 h_{ij}}{\partial \tau^2} |_{\varphi_0}) = 0.$$

So now we have

$$-\mathbf{C} = \begin{pmatrix} A_{11}^* & A_{12}^* \\ A_{21}^* & A_{22}^* \end{pmatrix}$$

where

$$\begin{aligned} A_{11}^* &= \sum_i \sum_j E\left(\frac{\partial^2 h_{ij}}{\partial \mu^2}|_{\varphi_0}\right), & A_{12}^* = A_{21}^* &= \sum_i \sum_j E\left(\frac{\partial^2 h_{ij}}{\partial \mu \partial \tau}|_{\varphi_0}\right), \\ A_{22}^* &= \sum_i \sum_j E\left(\frac{\partial^2 h_{ij}}{\partial \tau^2}|_{\varphi_0}\right). \end{aligned}$$

Let

$$-\mathbf{C} - \mathbf{H} = \begin{pmatrix} H_{11}^* & H_{12}^* \\ H_{21}^* & H_{22}^* \end{pmatrix}$$

where

$$\begin{aligned} H_{11}^* &= \sum_i \sum_j \left\{ E\left(\frac{\partial^2 h_{ij}}{\partial \mu^2}|_{\varphi_0}\right) - \frac{\partial^2 h_{ij}}{\partial \mu^2}|_{\varphi_0} - \frac{\partial^2 h_{ij}}{\partial \mu^2}|_{\varphi^*} + \frac{\partial^2 h_{ij}}{\partial \mu^2}|_{\varphi_0} \right\}, \\ H_{21}^* = H_{12}^* &= \sum_i \sum_j \left\{ E\left(\frac{\partial^2 h_{ij}}{\partial \mu \partial \tau}|_{\varphi_0}\right) - \frac{\partial^2 h_{ij}}{\partial \mu \partial \tau}|_{\varphi_0} - \frac{\partial^2 h_{ij}}{\partial \mu \partial \tau}|_{\varphi^*} + \frac{\partial^2 h_{ij}}{\partial \mu \partial \tau}|_{\varphi_0} \right\}, \\ H_{22}^* &= \sum_i \sum_j \left\{ E\left(\frac{\partial^2 h_{ij}}{\partial \tau^2}|_{\varphi_0}\right) - \frac{\partial^2 h_{ij}}{\partial \tau^2}|_{\varphi_0} - \frac{\partial^2 h_{ij}}{\partial \tau^2}|_{\varphi^*} + \frac{\partial^2 h_{ij}}{\partial \tau^2}|_{\varphi_0} \right\}. \end{aligned}$$

By Jiang's [14] Corollary 4.1.3 (ii) and Taylor expansion

$$\max_{ij} \left| \frac{\partial^2 h_{ij}}{\partial \mu^2}|_{\varphi^*} - \frac{\partial^2 h_{ij}}{\partial \mu^2}|_{\varphi_0} \right| \leq |\hat{\varphi} - \varphi_0| \max_{ij} \left| \frac{\partial^3 h_{ij}}{\partial \mu^3}|_{\tilde{\varphi}} \right| \leq B |\hat{\varphi} - \varphi_0| \quad (5.26)$$

$$\max_{ij} \left| \frac{\partial^2 h_{ij}}{\partial \tau^2}|_{\varphi^*} - \frac{\partial^2 h_{ij}}{\partial \tau^2}|_{\varphi_0} \right| \leq |\hat{\varphi} - \varphi_0| \max_{ij} \left| \frac{\partial^3 h_{ij}}{\partial \tau^3}|_{\tilde{\varphi}} \right| \leq B |\hat{\varphi} - \varphi_0| \quad (5.27)$$

$$\max_{ij} \left| \frac{\partial^2 h_{ij}}{\partial \tau \partial \mu}|_{\varphi^*} - \frac{\partial^2 h_{ij}}{\partial \tau \partial \mu}|_{\varphi_0} \right| \leq |\hat{\varphi} - \varphi_0| \max_{ij} \left| \frac{\partial^3 h_{ij}}{\partial \varphi^3}|_{\tilde{\varphi}} \right| \leq B |\hat{\varphi} - \varphi_0| \quad (5.28)$$

where  $\tilde{\varphi}$  is between  $\varphi^*$  and  $\varphi_0$  and  $B|\hat{\varphi} - \varphi_0| \rightarrow_p 0$ .

Since  $-\mathbf{H} - \mathbf{C}$  is a symmetric matrix, then

$$\begin{aligned}
& \| -\mathbf{H} - \mathbf{C} \| \\
&= \sup_{\|\mathbf{u}\|=1} |\mathbf{u}^t(-\mathbf{H} - \mathbf{C})\mathbf{u}| \\
&\leq 2 \left\{ \left( m_1 n B |\hat{\varphi} - \varphi_0| + \left| \sum_i \sum_j (E \frac{\partial^2 h_{ij}}{\partial \mu^2} |_{\varphi_0} - \frac{\partial^2 h_{ij}}{\partial \mu^2} |_{\varphi_0}) \right| \right) \right. \\
&\quad \vee \left( m_1 n B |\hat{\varphi} - \varphi_0| + \left| \sum_i \sum_j (E \frac{\partial^2 h_{ij}}{\partial \tau^2} |_{\varphi_0} - \frac{\partial^2 h_{ij}}{\partial \tau^2} |_{\varphi_0}) \right| \right) \\
&\quad \left. \vee \left( m_1 n B |\hat{\varphi} - \varphi_0| + \left| \sum_i \sum_j (E \frac{\partial^2 h_{ij}}{\partial \tau \partial \mu} |_{\varphi_0} - \frac{\partial^2 h_{ij}}{\partial \tau \partial \mu} |_{\varphi_0}) \right| \right) \right\}
\end{aligned}$$

Since  $(\partial^2 h_{ij}/\partial \mu^2)|_{\varphi_0}$  are iid,  $(\partial^2 h_{ij}/\partial \tau^2)|_{\varphi_0}$  are iid, and  $(\partial^2 h_{ij}/\partial \tau \partial \mu)|_{\varphi_0}$  are iid.

Suppose we have  $0 < \tau_0 < b$  where  $b$  is a positive constant. By Jiang's [14]

Lemma 3.3, the second derivatives of  $h_{ij}$  are uniformly bounded. Then

$$E \left| \frac{\partial^2 h_{ij}}{\partial \mu^2} |_{\varphi_0} \right| < \infty \quad E \left| \frac{\partial^2 h_{ij}}{\partial \tau^2} |_{\varphi_0} \right| < \infty \quad E \left| \frac{\partial^2 h_{ij}}{\partial \tau \partial \mu} |_{\varphi_0} \right| < \infty. \quad (5.29)$$

By the Law of Large numbers

$$\frac{1}{mn_1} \left| \sum_i \sum_j \left( E \frac{\partial^2 h_{ij}}{\partial \mu^2} |_{\varphi_0} - \frac{\partial^2 h_{ij}}{\partial \mu^2} |_{\varphi_0} \right) \right| = o_p(1), \quad (5.30)$$

$$\frac{1}{mn_1} \left| \sum_i \sum_j \left( E \frac{\partial^2 h_{ij}}{\partial \tau^2} |_{\varphi_0} - \frac{\partial^2 h_{ij}}{\partial \tau^2} |_{\varphi_0} \right) \right| = o_p(1), \quad (5.31)$$

$$\frac{1}{mn_1} \left| \sum_i \sum_j \left( E \frac{\partial^2 h_{ij}}{\partial \tau \partial \mu} |_{\varphi_0} - \frac{\partial^2 h_{ij}}{\partial \tau \partial \mu} |_{\varphi_0} \right) \right| = o_p(1). \quad (5.32)$$

Then

$$\begin{aligned}
& \|\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H})\mathbf{C}^{-\frac{1}{2}} - \mathbf{I}\| \\
&= \|\mathbf{C}^{-\frac{1}{2}}(-\mathbf{H} - \mathbf{C})\mathbf{C}^{-\frac{1}{2}}\| \\
&\leq \|\mathbf{C}^{-1}\| \|-\mathbf{H} - \mathbf{C}\| \\
&= O_p(1) \frac{\|-\mathbf{H} - \mathbf{C}\|}{mn_1} = o_p(1).
\end{aligned}$$

Consider condition (3) of Theorem 5.2.1 and recall (5.13). We have

$$\mathbf{G}_{ij} = \begin{pmatrix} ((\partial h_{ij}/\partial \mu)|_{\varphi_0})^2 & ((\partial h_{ij}/\partial \mu)|_{\varphi_0}) ((\partial h_{ij}/\partial \tau)|_{\varphi_0}) \\ ((\partial h_{ij}/\partial \tau)|_{\varphi_0}) ((\partial h_{ij}/\partial \mu)|_{\varphi_0}) & ((\partial h_{ij}/\partial \tau)|_{\varphi_0})^2 \end{pmatrix}.$$

Since  $\mathbf{G}_{ij}$  is a symmetric matrix, we have

$$\|\mathbf{G}_{ij}\| \leq 2 \max \left( \left( \frac{\partial h_{ij}}{\partial \mu}|_{\varphi_0} \right)^2, \left| \left( \frac{\partial h_{ij}}{\partial \tau}|_{\varphi_0} \right) \left( \frac{\partial h_{ij}}{\partial \mu}|_{\varphi_0} \right) \right|, \left( \frac{\partial h_{ij}}{\partial \tau}|_{\varphi_0} \right)^2 \right).$$

Recall (5.24), (5.23). It is easy to show

$$\left( \frac{\partial h_{ij}}{\partial \mu}|_{\varphi_0} \right)^2 \leq 4r^2$$

since  $\xi_{ij}$  are independent identically distributed standard normal. Suppose we have

$0 < \tau_0 < b$  where  $b$  is a positive constant, By Jiang's [14] Lemma 3.3 that first derivatives of  $h_{ij}$  are uniformly bounded. We have  $E(\partial h_{ij}/\partial \tau|_{\varphi_0})^2$  bounded and

$$\left| \left( \frac{\partial h_{ij}}{\partial \tau}|_{\varphi_0} \right) \left( \frac{\partial h_{ij}}{\partial \mu}|_{\varphi_0} \right) \right|^2 \leq \left( \frac{\partial h_{ij}}{\partial \mu}|_{\varphi_0} \right)^2 \left( \frac{\partial h_{ij}}{\partial \tau}|_{\varphi_0} \right)^2.$$

Condition (1), (2) and (3) of Theorem 5.2.1 are verified, so we have

$$\mathbf{C}^{\frac{1}{2}} \begin{pmatrix} \hat{\mu} - \mu_0 \\ \hat{\tau} - \tau_0 \end{pmatrix} \xrightarrow{D} N(\mathbf{0}, \mathbf{I}) \tag{5.33}$$

## Chapter 6

### Simulation

In this chapter, in order to check the asymptotic results of Chapters 3 and 5, logistic and Poisson random intercept models are simulated under case 1 for both our new estimator (linear combination of weighted MLE) and PGWLE (Penalized Generalized Weighted Least Square Estimate) from Jiang [14]. For case 2, one simple model was simulated. The asymptotic behavior of the estimates is investigated in samples generated by Splus 2000 or R 2.6 for various sample sizes and parameter configurations. The built-in function `nlminb` was used to compute the estimates. Both statistical and computational questions were examined in the course of the simulations.

In our simulations we compared Monte Carlo average of estimators to the known true values of parameters, and we compared Monte Carlo averages of approximate variance formulas to the Monte Carlo sample variances. We also used both the Kolmogorov-Smirnov and Shapiro-Wilk tests to assess agreement of standardized estimators with the  $N(0, 1)$  distribution.

The Kolmogorov-Smirnov test is based on  $D_n = \sup_x |\hat{F}_n(x) - \Phi(x)|$ , where  $\hat{F}_n(x)$  is the empirical cdf and  $\Phi(x)$  is the  $N(0, 1)$  cdf. The Shapiro-Wilk test [27] is implemented in R using the algorithm of Royston. Intuitively, the Shapiro-Wilk test is based on the observed correlation between an ordered sample and expected

values of  $N(0, 1)$  order statistics. The actual calculation of the statistic and its p-value relies on various approximations of means and variances of normal order statistics. See Royston [26] and references there in for details.

## 6.1 Case 1 simulations for $\hat{\beta}_w$

We simulate logistic and Poisson random intercept regression to investigate consistency and asymptotic normality of the estimated regression coefficients  $\hat{\beta}_w$  and  $\hat{\beta}_w$ . The results of Chapter 3 established the conditional and unconditional asymptotic behavior of regression coefficients, given the random effects  $\mathbf{v}_0$ .

### 6.1.1 Case 1 combining $a$ with $v_i$ .

We consider:

$$E(y_{ij}|v_i) = b'_{ij}(a + v_i + \beta x_{ij}) \quad (6.1)$$

where  $a$  and  $\beta$  are fixed parameters,  $x_{ij}$  is a scalar valued predictor, and  $v_i$  are iid Unif  $(-0.8, 0.8)$  or  $N(0, 0.25)$ . For our simulations we let  $a_0 = 1$ . We used the **nlinb** minimization function in Splus to get estimators which minimize  $-l_{n_i}(\gamma_i)$  in each cluster defined as

$$-l_{n_i}(\gamma_i) = \sum_j (-y_{ij}\eta_{ij} + b_{ij}(\eta_{ij}))$$

where  $\eta_{ij} = a + v_i + x_{ij}\beta$  and  $\gamma = (a, \beta, v_1, \dots, v_m)^t$ . We do not attempt to estimate  $a$  and  $v_i$  separately but only  $a + v_i$ . For simplicity, we simulated the balanced model with  $m$  clusters and  $n_1$  observations per cluster. We generated  $m \times n_1$  covariates  $x_{ij}$  uniformly spaced between  $-1$  and  $1$ , so that the ranges of  $\{x_{ij}\}$  and  $\{x_{lj}\}$ ,  $i \neq l$ ,

did not overlap. For example if we have  $m = 3$  clusters and  $n_1 = 2$  observations per cluster, first we divided interval  $[-1, 1]$  into  $m = 3$  equal sized intervals  $[-1, -1/3]$ ,  $[-1/3, 1/3]$  and  $[1/3, 1]$ , and then divide each interval into  $n_1 = 2$  subintervals to get  $x_{11} = -1$ ,  $x_{12} = -2/3$ ,  $x_{21} = -1/3$ ,  $x_{22} = 0$ , and  $x_{31} = 1/3$ ,  $x_{32} = 2/3$  for clusters 1, 2, 3 respectively. We also generate  $m$  independent random effects  $v_i$  from the Uniform distribution  $(-0.8, 0.8)$  or  $N(0, 0.25)$  according to  $(x_{ij}, v_i)$  with regression coefficient  $\beta_0 = 1$ ,  $m$  samples of  $n_1$  binary random variables  $Y_{ij}$  was generated with  $E(Y_{ij}|v_i, x_{ij}) = b'_{ij}(\eta_{ij})$  where for Poisson model with  $b'_{ij}(\eta_{ij}) = \exp(\eta_{ij})$  and for logistic model with  $b'_{ij}(\eta_{ij}) = \exp(\eta_{ij})/(1 + \exp(\eta_{ij}))$ . Various combinations of  $(m, n_1)$  are used in the simulations.

For each combination, 1000 replications of random effects  $v_i$  and responses  $Y_{ij}$  were generated. The covariates  $x_{ij}$  were the same for all the simulations. Estimated regression coefficients for various choices of  $(m, n_1)$  are summarized in Table 6.1 and Table 6.2. The tables display the Monte Carlo means and standard errors of the simulated values of  $\hat{\beta}_w - \beta_0$ ,  $\hat{\hat{\beta}}_w - \beta_0$ ,  $(\hat{\beta}_w - \beta_0)/s.e.$ ,  $(\hat{\hat{\beta}}_w - \beta_0)/\widehat{s.e.}$  and 95% confidence bounds for  $(\hat{\beta}_w - \beta_0)$  and  $\hat{\hat{\beta}}_w - \beta_0$  based on Student's  $t$ . From Table 6.1 and Table 6.2, we can see that in combinations  $(30, 120)$ ,  $(40, 240)$  in logistic model for  $\hat{\beta}_w$ ,  $\hat{\hat{\beta}}_w$ , some bias is present (based on confidence intervals). But for all the combinations from both tables, their Shapiro-Wilk test  $p$  values are above 0.18 and Kolmogorov-Smirnov test  $p$  values are larger than 0.1, which means in those combinations,  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  given  $\mathbf{v}_0$  seem normal. From Table 6.1 and Table 6.2, we can see that bias/se < 0.165 so the normal inference is not greatly affected in the cases where bias is present. The plots show that  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  are approximately

normal, with departures from normality in the extreme tails when random effects are uniformly distributed.

Table 6.1: Simulated estimates of logistic and Poisson regression coefficient combining fixed intercept with uniformly distributed random effects.

<u><math>(\hat{\beta}_w - \beta_0)</math></u>			<u><math>\hat{\beta}_w - \beta_0</math></u>		
(m, n <sub>1</sub> )	mean	std	95% CI for $\hat{\beta}_w - \beta_0$	mean	std
<u>Logistic <math>\beta_0 = 1</math></u>					
(10,200)	0.020	0.808	(-0.029, 0.071)	0.002	0.794
(20,200)	-0.020	1.190	(-0.089, 0.050)	-0.033	1.100
(30,120)	-0.224	1.432	(-0.313, -0.135)	-0.229	1.421
(30,210)	-0.055	1.200	(-0.130, 0.019)	-0.063	1.191
(40,240)	-0.197	1.226	(-0.273, -0.120)	-0.202	1.221
<u>Poisson <math>\beta_0 = 1</math></u>					
(10,200)	0.007	0.355	(-0.015, 0.029)	0.002	0.352
(20,200)	0.028	0.510	(-0.004, 0.059)	0.023	0.504
(30,210)	-0.033	0.590	(-0.069, 0.004)	-0.036	0.588
(40,240)	0.003	0.648	(-0.038, 0.043)	-0.001	0.644

Table 6.2: Simulated standardized estimate of logistic and Poisson regression coefficient with uniformly distributed random effects.

$(\hat{\beta}_w - \beta_0)/s.e$				$(\hat{\beta}_w - \beta_0)/\widehat{s.e.}$			
$(m, n_1)$	mean	std	95% CI for $(\hat{\beta}_w - \beta_0)$	mean	std		
<u>Logistic <math>\beta_0 = 1</math></u>							
(10,200)	0.026	1.014	(-0.047, 0.051)	0.001	0.989		
(20,200)	-0.018	0.995	(-0.101, 0.035)	-0.030	0.971		
(30,120)	0.126	0.806	(-0.317, -0.140)	-0.127	0.792		
(30,210)	-0.041	0.893	(-0.137, 0.010)	-0.047	0.881		
(40,240)	-0.135	0.845	(-0.278, -0.126)	-0.139	0.837		
<u>Poisson <math>\beta_0 = 1</math></u>							
(10,200)	0.014	1.006	(-0.020, 0.024)	0.001	1.000		
(20,200)	0.053	0.995	(-0.008, 0.054)	0.044	0.982		
(30,210)	-0.051	0.963	(-0.073, 0.0002)	-0.057	0.959		
(40,240)	0.005	0.976	(-0.041, 0.039)	-0.0005	0.968		

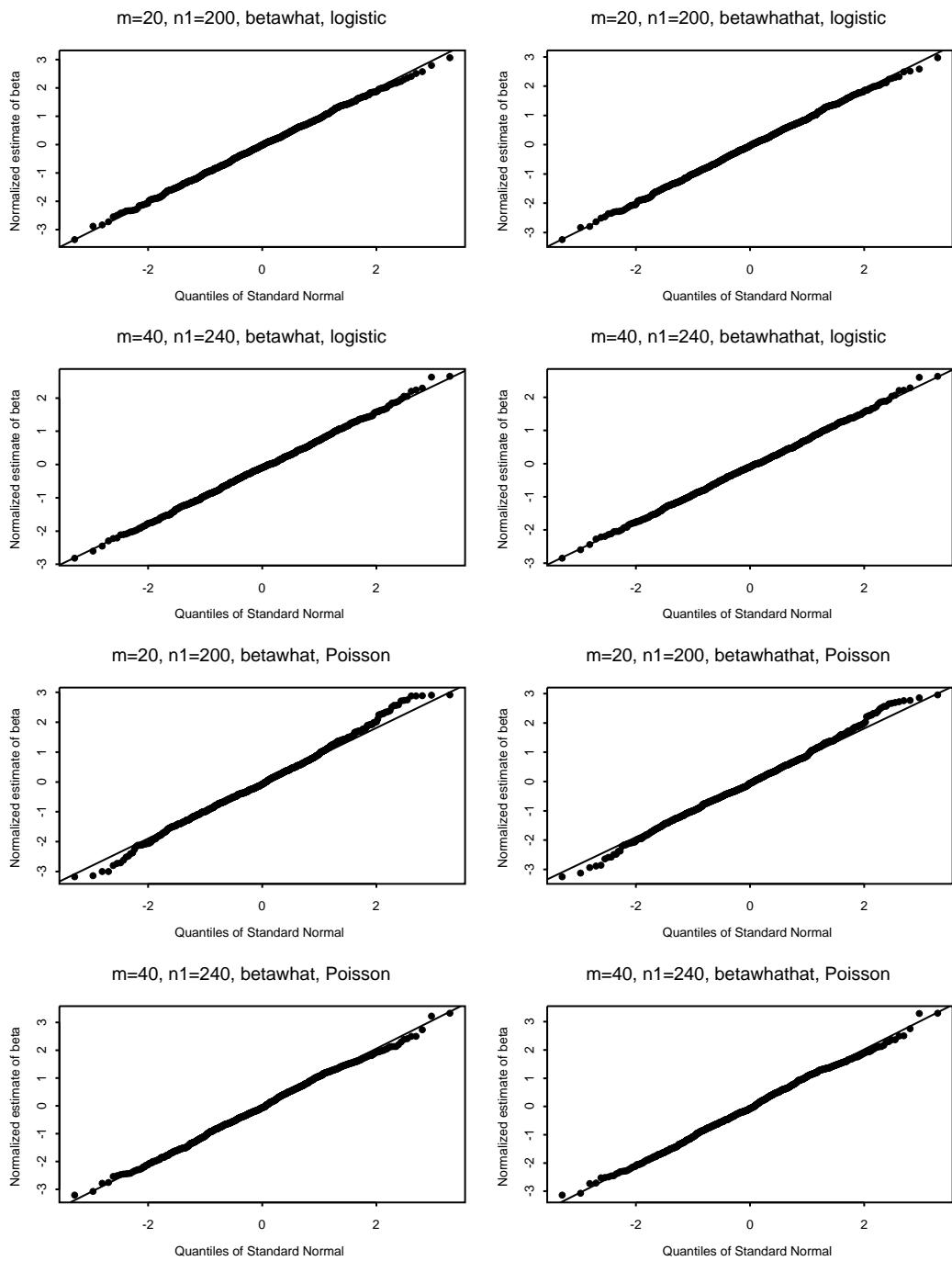


Figure 6.1: Q-Q plots for  $(\hat{\beta}_w - \beta_0)$  and  $(\hat{\hat{\beta}}_w - \beta_0)$  standardized by true and estimated standard error, for various values of  $(m, n)$  in case 1 random intercept model with uniformly distributed random effects.

Our theoretical results assumed bounded random effects, but we investigated unbounded normal random effects by simulation studies to those with uniform random effects. For each combination  $(m, n)$ , 1000 replications of random  $N(0, 0.25)$  effects  $v_i$  and responses  $Y_{ij}$  were generated. The covariates  $x_{ij}$  were the same for all the simulations. Estimated regression coefficients for various choices of  $(m, n_1)$  are summarized in Table 6.3 and Table 6.4. The tables display the means and standard errors of the simulated values of  $\hat{\beta}_w - \beta_0$ ,  $\hat{\hat{\beta}}_w - \beta_0$ ,  $(\hat{\beta}_w - \beta_0)/s.e.$ ,  $(\hat{\hat{\beta}}_w - \beta_0)/\widehat{s.e.}$  and 95% confidence bounds for  $(\hat{\beta}_w - \beta_0)$ ,  $\hat{\hat{\beta}}_w - \beta_0$  based on Student's  $t$ . From Table 6.3 and Table 6.4, we can see that in combinations  $(30, 210)$ ,  $(40, 240)$  in logistic model for  $\hat{\beta}_w$ ,  $\hat{\hat{\beta}}_w$ , some bias is present. But for all the combinations from both tables, the Shapiro-Wilk test  $p$  values are above 0.463 and the Kolmogorov-Smirnov test  $p$  values are larger than 0.1, which means in those combinations,  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  given  $\mathbf{v}_0$  seem normal. From Table 6.3 and Table 6.4, we can see that  $bias/se < 0.177$ , so the normal inference is not greatly affected in the cases where bias is present. The plots show that  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  are approximately normal, with departures from normality in the extreme tails when random effects are normally distributed. Here we need to point out that we have not proven asymptotic normality of our estimator in case 1 random intercept model for normally distributed random effects, but the simulation studies show that normality appears to hold.

Table 6.3: Simulated estimates of logistic and Poisson regression coefficient combining fixed intercept with normally distributed random effects.

$(\hat{\beta}_w - \beta_0)$			$\hat{\beta}_w - \beta_0$		
(m, n <sub>1</sub> )	mean	std	95% CI for $\hat{\beta}_w - \beta_0$	mean	std
<u>Logistic <math>\beta_0 = 1</math></u>					
(20,200)	0.012	1.132	(-0.058, 0.082)	-0.002	1.117
(30,210)	-0.206	1.219	(-0.282, -0.130)	-0.214	1.209
(40,240)	-0.202	1.227	(-0.278, -0.126)	-0.203	1.221
<u>Poisson <math>\beta_0 = 1</math></u>					
(20,200)	-0.001	0.512	(-0.033, 0.031)	-0.005	0.510
(30,210)	0.011	0.587	(-0.026, 0.047)	0.005	0.585
(40,240)	0.024	0.662	(-0.017, 0.065)	0.021	0.659

Table 6.4: Simulated standardized estimate of logistic and Poisson regression coefficient with normally distributed random effects.

$(\hat{\beta}_w - \beta_0)/s.e$	$(\hat{\beta}_w - \beta_0)/\widehat{s.e.}$					
$(m, n_1)$	mean	std	95% CI for $(\hat{\beta}_w - \beta_0)$	mean	std	
<u>Logistic <math>\beta_0 = 1</math></u>						
(20,200)	0.011	1.004	(-0.071, 0.068)	-0.002	0.983	
(30,210)	-0.153	0.904	(-0.289, -0.139)	-0.158	0.891	
(40,240)	-0.139	0.843	(-0.279, -0.127)	-0.139	0.834	
<u>Poisson <math>\beta_0 = 1</math></u>						
(20,200)	-0.003	1.003	(-0.037, 0.026)	-0.011	0.997	
(30,210)	0.019	0.959	(-0.031, 0.414)	0.009	0.954	
(40,240)	0.037	0.994	(-0.019, 0.062)	0.033	0.987	

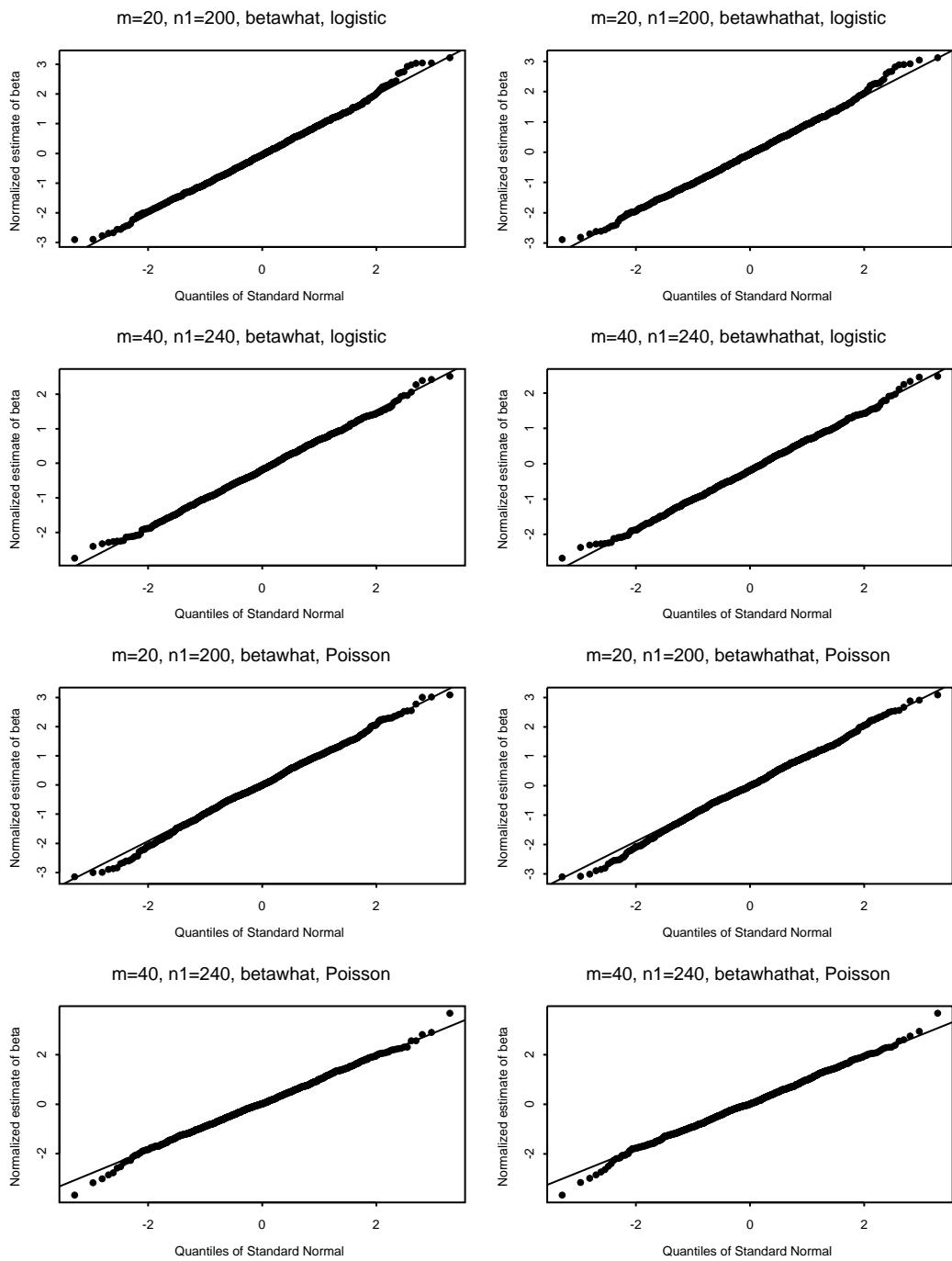


Figure 6.2: Q-Q plots for  $(\hat{\beta}_w - \beta_0)$  and  $(\hat{\hat{\beta}}_w - \beta_0)$  standardized by true and estimated standard error, for various values of  $(m, n)$  in case 1 random intercept model with normally distributed random effects.

## 6.2 Case 1 logistic random intercept simulation for fixed realizations of random effects

Recall that Theorem 3.1.2 and Theorem 3.1.3 gave conditional convergence in distribution of the normalized  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  to the  $N(\mathbf{0}, \mathbf{I})$  distribution. We performed a limited study to examine this conditional convergence.

We repeated the simulations designed in Section 6.1.1 for the random intercept problem, but with the following modification: we generated 10 realizations of i.i.d  $\text{Unif}(-0.8, 0.8)$  of  $\mathbf{v}_0$ . For each realization of  $\mathbf{v}_0$  we generated 1000 replications of  $\mathbf{y}$  with the same  $\mathbf{v}_0$  and  $\mathbf{x}$ 's as described in Section 6.1.1. This was performed only for sample sizes  $m = 20$ ,  $n_1 = 200$  and  $m = 40$ ,  $n_1 = 240$ , and for the logistic model (6.1).

From the following tables Table 6.5, Table 6.6, Table 6.7 and Table 6.8, we found that each realizations of  $\mathbf{v}_0$ , the normalized  $\hat{\beta}_w$  and  $\hat{\hat{\beta}}_w$  had Monte Carlo mean zero and variances near 1 in the case  $m = 20$ ,  $n_1 = 200$ . For the case  $m = 40$ ,  $n_1 = 240$ , the Monte Carlo mean and variance show some departures from the desired 0 and 1. The Kolmogorov-Smirnov and Shapiro-Wilk tests all indicated no significant departures from normality. These findings are very similar to those of Section 6.1.1, which describe the unconditional distribution.

Table 6.5: Simulated estimates of logistic regression coefficient combining fixed effect and fixed realizations of uniformly distributed random effects ( $m = 20, n_1 = 200$ ).

Realization	$(\hat{\beta}_w - \beta_0)$			$\hat{\beta}_w - \beta_0$		
	mean	std	95% CI for $\hat{\beta}_w - \beta_0$	mean	std	
1	0.004	1.063	(-0.062, 0.070)	-0.012	1.047	
2	-0.208	1.133	(-0.098, 0.043)	-0.042	1.117	
3	-0.012	1.127	(-0.082, 0.058)	-0.028	1.109	
4	-0.044	1.105	(-0.113, 0.024)	-0.058	1.087	
5	0.028	1.066	(-0.038, 0.094)	0.012	1.048	
6	0.019	1.141	(-0.052, 0.089)	0.005	1.127	
7	-0.024	1.075	(-0.090, 0.043)	-0.039	1.059	
8	0.026	1.110	(-0.043, 0.095)	0.010	1.095	
9	-0.035	1.079	(-0.102, 0.032)	-0.048	1.061	
10	-0.026	1.110	(-0.095, 0.042)	-0.040	1.092	

Table 6.6: Simulated standardized estimate of logistic regression coefficient with fixed realizations of uniformly distributed random effects ( $m = 20$ ,  $n_1 = 200$ ).

Realization	$(\hat{\beta}_w - \beta_0)/s.e$			$(\hat{\hat{\beta}}_w - \beta_0)/\widehat{s.e.}$		
	mean	std	95% CI for $(\hat{\hat{\beta}}_w - \beta_0)$	mean	std	
1	0.003	0.948	(-0.077, 0.052)	-0.011	0.926	
2	-0.025	1.005	(-0.111, 0.027)	-0.037	0.982	
3	0.010	0.998	(-0.097, 0.041)	-0.025	0.975	
4	-0.040	0.988	(-0.125, 0.010)	-0.052	0.975	
5	0.025	0.946	(-0.053, 0.077)	0.011	0.924	
6	0.017	1.016	(-0.065, 0.075)	0.004	0.996	
7	-0.021	0.959	(-0.105, 0.027)	-0.035	0.937	
8	0.023	0.991	(-0.058, 0.078)	0.008	0.970	
9	-0.031	0.961	(-0.114, 0.018)	-0.043	0.938	
10	-0.024	0.992	(-0.108, 0.027)	-0.036	0.969	

Table 6.7: Simulated estimates of logistic regression coefficient combining fixed effect and fixed realizations of uniformly distributed random effects ( $m = 40, n_1 = 240$ ).

Realization	$(\hat{\beta}_w - \beta_0)$			$\hat{\beta}_w - \beta_0$		
	mean	std	95% CI for $\hat{\beta}_w - \beta_0$	mean	std	
1	-0.210	1.207	(-0.285, -0.135)	-0.214	1.202	
2	-0.225	1.226	(-0.301, -0.149)	-0.228	1.223	
3	-0.117	1.125	(-0.194, -0.041)	-0.120	1.229	
4	-0.201	1.261	(-0.279, -0.122)	-0.204	1.257	
5	-0.164	1.261	(-0.238, -0.089)	-0.204	1.225	
6	-0.211	1.204	(-0.285, -0.136)	-0.214	1.198	
7	-0.185	1.211	(-0.260, -0.110)	-0.188	1.206	
8	-0.206	1.189	(-0.280, -0.132)	-0.209	1.184	
9	-0.205	1.203	(-0.279, -0.130)	-0.207	1.200	
10	-0.249	1.223	(-0.325, -0.173)	-0.252	1.219	

Table 6.8: Simulated standardized estimate of logistic regression coefficient with fixed realizations of uniformly distributed random effects ( $m = 40$ ,  $n_1 = 240$ ).

Realization	$(\hat{\beta}_w - \beta_0)/s.e$			$(\hat{\beta}_w - \beta_0)/\widehat{s.e.}$		
	mean	std	95% CI for $(\hat{\beta}_w - \beta_0)$	mean	std	
1	-0.145	0.836	(-0.288, -0.139)	-0.147	0.829	
2	-0.150	0.845	(-0.304, -0.152)	-0.157	0.839	
3	-0.081	0.853	(-0.200, -0.044)	-0.083	0.845	
4	-0.138	0.868	(-0.282, -0.126)	-0.140	0.861	
5	-0.112	0.822	(-0.242, -0.093)	-0.114	0.815	
6	-0.145	0.831	(-0.288, -0.140)	-0.147	0.823	
7	-0.128	0.836	(-0.263, -0.113)	-0.129	0.829	
8	-0.142	0.821	(-0.283, -0.136)	-0.144	0.813	
9	-0.142	0.833	(-0.281, -0.133)	-0.143	0.824	
10	-0.172	0.845	(-0.328, -0.176)	-0.174	0.839	

### 6.3 Case 1 random intercept simulations for penalized likelihood estimators

We simulate logistic and Poisson random intercept regression to investigate consistency and asymptotic normality of penalized regression coefficient estimates by using Jiang's [14] PGWLE (Penalized Generalized Weighted Least Squares) method.

### 6.3.1 Case 1 logistic random intercept combining $a$ with $v_i$ .

We consider Example 3.2 from Jiang [14]:

$$\text{logit}P(y_{ij} = 1|v_i) = a + v_i + \beta_1 x_{ij} \quad (6.2)$$

where  $a$  and  $\beta_1$  are fixed parameters,  $x_{ij}$  is a scalar valued predictor, and  $v_i$  are iid  $N(0, V_a)$ . For our simulations we let  $a = 0$ . We used the PGWLS method by Jiang [14] and the `nlinnb` minimization function in Splus to get estimators which minimize  $-l_P(\gamma)$  derived by Jiang [14] as (??), which is defined as

$$-l_P(\gamma) = \sum_i \sum_j (-y_{ij}\eta_{ij} + \log(1 + \exp(\eta_{ij}))) + \frac{\lambda_1}{2}m(\bar{v})^2$$

where  $\eta_{ij} = a + v_i + x_{ij}\beta_1$ ,  $\bar{v} = \sum_i^m v_i/m$ , and  $\gamma = (a, \beta_1, v_1, \dots, v_m)^t$ . We do not attempt to estimate  $a$  and  $v_i$  separately but only  $a + v_i$ . For simplicity, we simulated the balanced model with  $m$  clusters and  $n_1$  observations per cluster. We generated  $m \times n_1$  covariates  $x_{ij}$  uniformly spaced between -1 and 1, so that the ranges of  $\{x_{ij}\}$  and  $\{x_{lj}\}$ ,  $i \neq l$ , did not overlap. We also generate  $m$  independent random effects  $v_i$  from the normal distribution with mean 0 and variance  $V_a$ . Conditionally on  $(x_{ij}, v_i)$  with regression coefficient  $\beta_{10} = 1$ , a sample of  $mn_1$  binary random variables  $Y_{ij}$  was generated with  $E(Y_{ij}|v_i, x_{ij}) = \exp(x_{ij} + v_i)/(1 + \exp(v_i + x_{ij}))$ .

Various combinations of  $(m, n_1)$ ,  $V_a$ ,  $\lambda_1$  and initial values of estimated  $\beta_1$ ,  $v_1, \dots, v_m$  were used.

For each combination, 1000 replications of random effects  $v_i$  and responses  $Y_{ij}$  were generated. The covariates  $x_{ij}$  were the same for all the simulations. Estimated regression coefficients for various choices of  $(m, n_1)$  with  $V_a = 4$ , initial estimates set

to 0, and  $\lambda_1 = 1$ , are summarized in Table 6.9. The tables display the means and standard errors of the simulated values of  $(\hat{\beta}_1 - \beta_{10})/s.e.$ ,  $(\hat{\beta}_1 - \beta_{10})/\widehat{s.e.}$  and 95% confidence bounds for  $\hat{\beta}_1 - \beta_{10}$  based on Student's  $t$ . From Table 6.9, we can see that in the combination (20, 20) with  $V_a = 4$ , bias is present. But its Kolmogorov-Smirnov test  $p$  values are larger than 0.1. This can be explained in terms of the consistency condition  $\log m / (\log n_1)^2 \rightarrow 0$  [Jiang ([14], Ex.3.2)]. Among all the pairs of sample sizes for simulation, (20, 20) has the highest values of  $\log m / (\log n_1)^2$  as 0.334. For all combinations from Table 6.9 whose 95% confidence intervals including 0.

In order to check whether the computations of  $\hat{\beta}_1$  are sensitive to initial values, we ran various combinations of  $m$ ,  $n_1$ ,  $V_a$  and  $\lambda_1$  with initial values set to 0. For the same  $v_i$  and  $y_{ij}$ , we initialized  $\hat{\beta}_1$  at 3 and at  $\text{logit}(\bar{y}_{..}/(1 - \bar{y}_{..}))$ . In each case the  $\hat{v}_i$  were initialized at 0. This comparison was repeated 1000 times. The results are summarized in Table 6.10.

Table 6.9: Simulated standardized estimates of logistic regression coefficient (standardized by true and estimated conditional standardized error) combining fixed intercept with random effects.

$(\hat{\beta}_1 - \beta_{10})/s.e.$				$(\hat{\beta}_1 - \beta_{10})/\widehat{s.e.}$			
(m, n <sub>1</sub> )	mean	std	95% CI for $\hat{\beta}_1 - \beta_{10}$	mean	std		
<u><math>V_a = 4</math>, initials set to 0, <math>\lambda_1=1</math></u>							
(10,20)	-0.017	1.021	(-0.127, 0.293)	-0.015	1.010		
(20,20)	0.077	1.012	(0.082, 0.665)	0.076	1.011		
(20,40)	0.054	0.995	(-0.019, 0.384)	0.053	0.991		
<u><math>V_a = 3</math>, initials set to 0, <math>\lambda_1 = 1</math></u>							
(20,20)	0.054	0.938	(-0.008, 0.509)	0.053	0.932		

Table 6.10: Simulated standardized estimate of logistic regression coefficient for various initial values.

$(\hat{\beta}_1 - \beta_{10})/s.e.$				$(\hat{\beta}_1 - \beta_{10})/\widehat{s.e.}$			
initial $\hat{\beta}_1$	mean	std	95% CI for $\hat{\beta}_1 - \beta_{10}$	mean	std		
<u><math>V_a = 4</math>, (m, n<sub>1</sub>) = (10, 20), <math>\lambda_1=1</math></u>							
0	-0.017	1.021	(-0.127, 0.293)	0.015	1.011		
3	-0.014	0.878	(-0.150, 0.277)	0.009	1.022		
logit	-0.020	0.896	(-0.150, 0.277)	0.015	1.011		

As we can see from Table 6.10, the estimates are biased for larger  $\log m / (\log n_1)^2$ .

Moreover, the variance of  $(\hat{\beta}_1 - \beta_{10})/s.e.$  is sensitive to the initial guess of  $\hat{\beta}_1$ . Also in the combination (20, 20) with  $V_a = 4$  and initial  $\hat{\beta}_1$  equal to logit, their Kolmogorov-Smirnov test rejects the hypothesis of normality with  $p$  values 0.043 and 0.019. For all the other combinations from Table 6.10, their goodness-fit tests suggest normality.

Since in reality we do not know how to choose  $\lambda_1$ , we tried various values of  $\lambda_1$  to assess the effect of  $\lambda_1$  on normality and consistency. The Monte Carlo averages and standard deviations of  $(\hat{\beta}_1 - \beta_{10})/s.e.$  are displayed in Table 6.11 for various  $(m, n_1)$ ,  $\lambda_1$ , initial values and  $V_a$ . The estimates are computed for  $\lambda_1 = 0.1, 1$  and  $5$  for the same data, and this comparison was repeated 1000 times. Table 6.11 shows that extremely large values of the penalty parameter ( $\lambda_1 = 5$ ) did affect the bias.

Table 6.11: Simulated standardized estimate of logistic regression coefficient for various  $\lambda_1$  values.

$\lambda_1$	$(\hat{\beta}_1 - \beta_{10})/s.e.$			$(\hat{\beta}_1 - \beta_{10})/\widehat{s.e.}$		
	mean	std	95% CI for $\hat{\beta}_1 - \beta_0$	mean	std	
$V_a = 4, (m, n_1) = (20, 50), \text{ initials set to 0}$						
1	0.048	1.012	(-0.034, 0.330)	0.048	1.009	
0.1	0.022	0.858	(-0.034, 0.330)	0.048	1.009	
5	0.022	0.858	(0.034, 0.400)	0.070	1.017	

## 6.4 Case 2 logistic simple example

We consider Example 3.3 from Jiang [14], with  $\alpha_i = 0$  for simplicity. We have

$$\text{logit}P(y_{ijk} = 1|b_{ij}) = \mu + b_{ij} = \mu + \tau\xi_{ij} \quad (6.3)$$

where the  $\xi_{ij}$  are iid  $N(0, 1)$ . We estimate  $\mu$  and  $\tau$  by Jiang's [14] MCLE method based on integrating  $\xi_{ij}$  out of the likelihood. This gives us an unconditional log-likelihood function in this special example. We also use the `nlminb` function in Splus to get estimates which minimize the negative log likelihood function. In Case 2, reparameterization is used to take care of the identifiability problem. Here since we integrate out  $\xi_{ij}$ , we only have the fixed effect  $\mu$  and scale parameter  $\tau$ . We do not need to worry about the identifiability problem, so the negative log-likelihood function is the following:

$$-l_C(\psi) = -\sum_i \sum_j h_{ij} = -\sum_i \sum_j E \exp[(\mu + \tau\xi_{ij})y_{ij+} - r \log(1 + \exp(\mu + \tau\xi_{ij}))] \quad (6.4)$$

where  $y_{ij+} = \sum_k y_{ijk}$ .

More precisely, here we have  $m$  rows and  $n$  columns for each cell we have  $r$  observations. We generate  $m \times n_1$  random effects  $b_{ij}$  from  $N(0, \tau_0^2)$ , where  $\tau_0$  is a scale parameter and set  $r = 2$ . Then we can get  $\eta_{ij1} = \eta_{ij2} = \mu + b_{ij}$ . Conditionally on  $\eta_{ijk}$  with  $k=1$  or  $2$ , a sample of  $m \times n_1 \times 2$  random variables  $Y_{ijk}$  is generated from the Bernoulli distribution with  $E(Y_{ijk}|\mu, \tau\xi_{ij}) = \exp(\mu + \tau\xi_{ij})/(1 + \exp(\mu + \tau\xi_{ij}))$ .

Various choices of  $(m, n_1)$  were simulated. For each choice, 500 replications of  $b_{ij}$ ,  $Y_{ijk}$  were generated. We used the Splus `nlminb` function to calculate estimates of  $\mu$  and  $\tau$ . In this simulation,  $\tau_0 = 1$ , initials for  $\hat{\mu}$  and  $\hat{\tau}$  are 0 and 1, respectively.

Asymptotic normality results for standardized  $\hat{\mu} - \mu_0$  and  $\hat{\tau} - \tau_0$  are summarized in Table 6.12.

For the combinations of Table (6.12), the only pair  $(m, n_1) = (20, 20)$  has Kolmogorov-Smirnov test  $p$  values larger than 0.1. For  $(m, n_1) = (10, 10)$ , the Kolmogorov-Smirnov test rejects the normality hypothesis with  $p$  values 0.01 and 0 for  $(\hat{\mu} - \mu_0)/s.e(\mu)$  and  $(\hat{\tau} - \tau_0)/s.e.(\tau)$ , respectively. For combinations  $(m, n_1) = (10, 40)$  and  $(m, n_1) = (10, 30)$ , normality only appear to hold for standardized  $(\hat{\tau} - \tau_0)$  with Kolmogorov-Smirnov test  $p$  values larger than 0.1. Moreover, the Kolmogorov-Smirnov test rejects the normality hypothesis with  $p$  value less than 0.008.

The 95% confidence intervals for  $(\hat{\mu} - \mu_0)$  for all the pairs  $(m, n_1)$  include 0, but 95% confidence intervals for  $(\hat{\tau} - \tau_0)$  for  $(10, 10)$  and  $(10, 30)$  do not include 0. In the combinations  $(m, n_1) = (20, 20)$  and  $(m, n_1) = (10, 40)$ , both  $(\hat{\mu} - \mu_0)$  and  $(\hat{\tau} - \tau_0)$  seem unbiased.

As a partial check on the joint normality of  $(\hat{\mu}, \hat{\tau})$ , we also examined  $[(\hat{\mu} - \mu_0) + (\hat{\tau} - \tau_0)]/\sqrt{2}$ . Similar results were obtained.

Table 6.12: Simulated standardized estimate of  $\mu$  and  $\tau$  in case 2 simple example  
for various  $(m, n_1)$  values

$(m, n_1)$	$(\hat{\mu} - \mu_0)/s.e(\mu)$		$(\hat{\tau} - \tau_0)/s.e.(\tau)$	
	mean	std	mean	std
$\tau_0 = 1, \mu_0 = 1, \text{ initials set to } (0,1)$				
(10,10)	-0.024	0.641	-0.236	1.550
(10,30)	0.079	0.835	-0.025	0.475
(20,20)	0.062	0.816	-0.015	0.523
(10,40)	0.036	0.800	-0.024	0.531

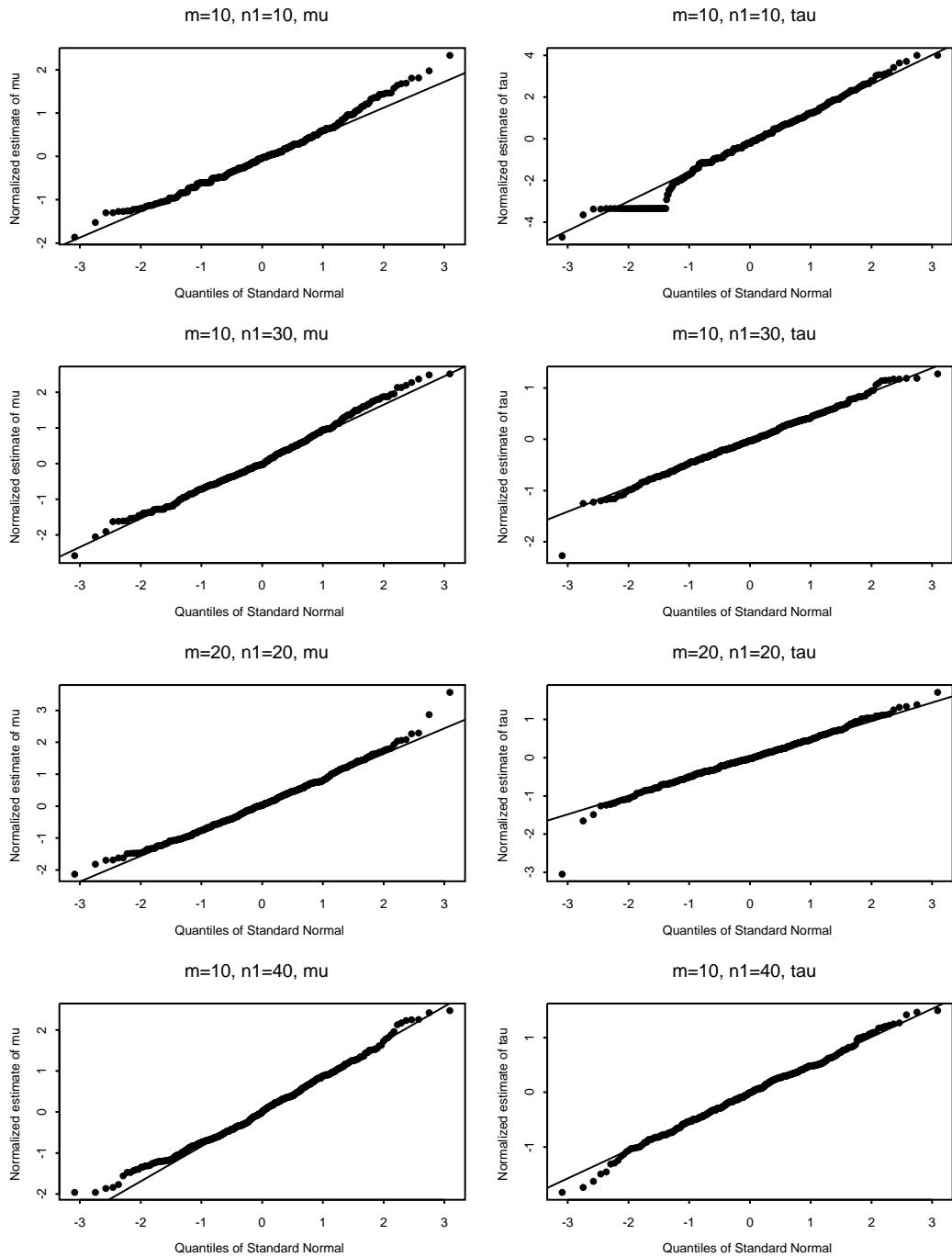


Figure 6.3: Q-Q plots for standardized  $(\hat{\mu} - \mu_0)$  and  $(\hat{\tau} - \tau_0)$  by true standard error, for various  $(m, n_1)$  in case 2 logistic simple example.

## Chapter 7

### Conclusions and Future Work.

#### 7.1 Theoretical Conclusions

We have proposed a new estimator  $\hat{\beta}_w = \sum_i \hat{\mathbf{w}}_i \hat{\beta}_i$  and have proven its conditional asymptotic normality for the random intercept problem when  $m$ , the number of random effects goes to infinity with the sample size  $N$  but at a slower rate, characterized as  $m/N \rightarrow 0$ . Given the random effects, the linear combinations of weighted MLE denoted  $\hat{\beta}_w$  and  $\hat{\beta}_w$  are asymptotically normal. The weight matrices and asymptotic conditional covariance matrix of  $\hat{\beta}_w$  can be estimated, and the standardized regression estimates, standardized by either the true or estimated covariance, have a limiting  $N(\mathbf{0}, \mathbf{I})$  distribution. We have proven that in the absence of regression coefficients, the normalized Jiang's [14] penalized likelihood estimator of fixed intercept  $\sqrt{m}(\hat{a} - a_0)$  converges to a normal distribution. Difficulties arise in establishing the conditional asymptotic normality of the penalized likelihood estimator  $\hat{\beta}$  of regression coefficients for fixed effects in a general GLMM.

For the case  $m/N \not\rightarrow 0$ , joint asymptotic normality is proved for regression coefficient and scale parameter estimates after suitable standardization. A logistic example with  $\text{logit}P(y_{ijk}|b_{ij}) = \mu + b_{ij}$  is used to illustrate how to verify the general conditions in this case.

## 7.2 Conclusions from the simulation studies and real data analysis

We focused on investigating asymptotic behavior of our new estimator  $\hat{\beta}_w$ , theoretic estimator  $\hat{\beta}_w$  and simulated balanced models for simplicity. In the case  $m/N \rightarrow 0$ , logistic and Poisson random intercept models were simulated. For both models, we considered normally and uniformly distributed random effects. When  $m/n \geq 1/7$ , our estimators had some bias under logistic model. But for Poisson model this problem did not occur except for  $(30, 210)$  with  $\beta_0 = 1$ . In all the cases, normality appears to hold and the ratio bias/s.e is less than 0.218, which means normal inference is not greatly affected. In the simulation studies for Jiang's [14] penalized estimator  $\hat{a}$  of fixed intercept, asymptotic consistency and normality does not hold for  $(30, 60)$  with  $V_a = 4$  and  $(30, 60), (40, 60), (40, 80)$  with  $V_a = 2$ . Estimates of  $a_0$  and standard deviation shows considerable bias.

We simulated logistic  $2 \times 2 \times m$  tables for both balanced and unbalanced setups with uniformly distributed random effects. In both setups,  $\hat{\beta}_w$  has better consistency results than  $\hat{\beta}_w$ . In all the cases, normality holds and the ratio bias/s.e is less than 0.194 except in extreme cases  $(100, 50)$  and  $(100, 60)$ , which means normal inference is not greatly affected.

In the case  $m/N \not\rightarrow 0$  for the logistic model  $\text{logit}P(y_{ijk} = 1|b_{ij}) = \mu + b_{ij}$ , among the combinations  $(20, 20)$  with  $r = 2$ , unbiasedness and approximate normality hold for the intercept and scale parameter estimates. Scale parameter estimates are biased if the number of clusters is too small ( $< 300$ ). Normality holds for standardized  $\hat{\tau} - \tau_0$  but not for standardized  $\hat{\mu} - \mu_0$  in  $(10, 40)$  and  $(10, 30)$

combinations.

We analized a real data set of 22 clinical trials of beta-blocker for heart attach treatment (Yusuf et al. (1985)). We first test homogeneity of odds ratios by using Woolf's test, the Breslow-Day test and the likelihood ratio test. None of them reject the null hypothesis, which suggest common odds ratio model is appropriate. Our estimator and Mantel-Haenszel estimator are very close. Mantel-Haenszel estimator is consistent in sparse data case unlike our estimator. The logic of our estimator can be extended to other types of GLMM where  $m \rightarrow \infty$  and  $m / \min_i n_i \rightarrow 0$ . No corresponding extension for Mantel-Haenszel is available.

### 7.3 Future work.

We can try to use projected score methods to overcome the difficulties in proving conditional asymptotic normality for penalized likelihood estimates of fixed effects in the case 1 random intercept problem. Also we can consider conditional logistic regression estimate of random intercept model.

We can investigate the asymptotic behavior of Penalized Generalized Weighted Least Square (PGWLS) estimate in a more complicated random effects models such as two way crossed random effects model.

## Chapter A

### Simulation Results

#### A.1 Case 1 random intercept model for penalized likelihood estimator

##### A.1.1 Case 1 logistic random intercept combining $a$ with $v_i$ .

We consider Example 3.2 from Jiang [14]:

$$\text{logit}P(y_{ij} = 1|v_i) = a + v_i + \beta_1 x_{ij} \quad (\text{A.1})$$

where  $a$  and  $\beta_1$  are fixed parameters,  $x_{ij}$  is a scalar valued predictor, and  $v_i$  are iid  $N(0, V_a)$ . For our simulations we let  $a = 0$ . We used the PGWLS method by Jiang [14] and **nlminb** minimization function in Splus to get estimators which minimize  $-l_P(\gamma)$  derived by Jiang [14] as (??), which is defined as

$$-l_P(\gamma) = \sum_i \sum_j (-y_{ij}\eta_{ij} + \log(1 + \exp(\eta_{ij}))) + \frac{\lambda_1}{2}m(\bar{v})^2$$

where  $\eta_{ij} = a + v_i + x_{ij}\beta_1$ ,  $\bar{v} = \sum_i^m v_i/m$ , and  $\gamma = (a, \beta_1, v_1, \dots, v_m)^t$ . We do not attempt to estimate  $a$  and  $v_i$  separately but only  $a + v_i$ . For simplicity, we simulated the balanced model with  $m$  clusters and  $n_1$  observations per cluster. We generated  $m \times n_1$  covariates  $x_{ij}$  uniformly spaced between -1 and 1, so that the ranges of  $\{x_{ij}\}$  and  $\{x_{lj}\}$ ,  $i \neq l$ , did not overlap. For example if we have  $m = 3$  clusters and  $n_1 = 2$  observations per cluster, first we divided interval  $[-1, 1]$  into

$m = 3$  equal sized intervals  $[-1, -1/3]$ ,  $[-1/3, 1/3]$  and  $[1/3, 1]$ , and then divide each interval into  $n_1 = 2$  subintervals to get  $x_{11} = -1$ ,  $x_{12} = -2/3$ ,  $x_{21} = -1/3$ ,  $x_{22} = 0$ , and  $x_{31} = 1/3$ ,  $x_{32} = 2/3$  for clusters 1, 2, 3 respectively. We also generate  $m$  independent random effects  $v_i$  from the normal distribution with mean 0 and variance  $V_a$ . Conditionally on  $(x_{ij}, v_i)$  with regression coefficient  $\beta_{10} = 1$ , a sample of  $mn_1$  binary random variables  $Y_{ij}$  was generated with  $E(Y_{ij}|v_i, x_{ij}) = \exp(x_{ij} + v_i)/(1 + \exp(v_i + x_{ij}))$ .

Various combinations of  $(m, n_1)$ ,  $V_a$ ,  $\lambda_1$  and initial values of estimated  $\beta_1$ ,  $v_1, \dots, v_m$  were used.

For each combination, 1000 replications of random effects  $v_i$  and responses  $Y_{ij}$  were generated. The covariates  $x_{ij}$  were the same for all the simulations. Estimated regression coefficients for various choices of  $(m, n_1)$  with  $V_a = 4$ , initial estimates set to 0, and  $\lambda_1 = 1$ , are summarized in Table A.1. The tables display the means and standard errors of the simulated values of  $(\hat{\beta}_1 - \beta_{10})/s.e.$  and 95% confidence bounds for  $E(\hat{\beta}_1 - \beta_{10})$  based on Student's  $t$ . From Table A.1, we can see that in combinations  $(20, 20)$ ,  $(20, 30)$  with  $V_a = 4$  and even for  $V_a = 3$  with  $(20, 20)$ , the bias is present. But their Kolmogorov-Smirnov test  $p$  values are larger than 0.1. This can be explained in terms of the consistency condition  $\log m/(\log n_1)^2 \rightarrow 0$  [Jiang ([14], Ex.3.2)]. Among all the pairs of sample sizes for simulation,  $(20, 20)$  and  $(20, 30)$  have the highest values of  $\log m/(\log n_1)^2$ , 0.334 and 0.259 respectively. For all combinations from Table A.1 whose 95% confidence intervals including 0.

In order to check whether the computations of  $\hat{\beta}_1$  are sensitive to initial values, we ran various combinations of  $m$ ,  $n_1$ ,  $V_a$  and  $\lambda_1$  with initial values set to 0. For

the same  $v_i$  and  $y_{ij}$ , we initialized  $\hat{\beta}_1$  at 3 and at  $\text{logit}(\bar{y}_{..}/(1 - \bar{y}_{..}))$ . In each case the  $\hat{v}_i$  were initialized at 0. This comparison was repeated 1000 times. The estimated standardized regression coefficients, standardized by conditional standard error and estimated conditional standard error, are summarized in Table A.2.

Table A.1: Simulated standardized estimates of logistic regression coefficient (standardized by true and estimated conditional standardized error) combining fixed intercept with random effects.

$(\hat{\beta}_1 - \beta_{10})/s.e.$				$(\hat{\beta}_1 - \beta_{10})/\widehat{s.e.}$			
$(m, n_1)$	mean	std	95% CI for $\hat{\beta}_1 - \beta_{10}$	mean	std		
<u><math>V_a = 4</math>, initials set to 0, <math>\lambda_1=1</math></u>							
(10,20)	-0.017	1.021	(-0.127, 0.293)	-0.015	1.010		
(10,30)	0.025	1.009	(-0.075, 0.260)	0.024	1.002		
(10,40)	0.021	0.980	(-0.086, 0.195)	0.021	0.977		
(20,20)	0.077	1.012	(0.082, 0.665)	0.076	1.011		
(20,30)	0.075	0.995	(0.059, 0.523)	0.075	0.990		
(20,40)	0.054	0.995	(-0.019, 0.384)	0.053	0.991		
(20,50)	0.049	1.012	(-0.034, 0.330)	0.048	1.009		
<u><math>V_a = 3</math>, initials set to 0, <math>\lambda_1 = 1</math></u>							
(20,20)	0.054	0.938	(-0.008, 0.509)	0.053	0.932		

Table A.2: Simulated standardized estimate of logistic regression coefficient for various initial values.

	$(\hat{\beta}_1 - \beta_{10})/s.e.$			$(\hat{\beta}_1 - \beta_{10})/\widehat{s.e.}$		
initial $\hat{\beta}_1$	mean	std	95% CI for $\hat{\beta}_1 - \beta_{10}$	mean	std	
<u><math>V_a = 4, (m, n_1) = (10, 20), \lambda_1 = 1</math></u>						
0	-0.017	1.021	(-0.127, 0.293)	0.015	1.011	
3	-0.014	0.878	(-0.150, 0.277)	0.009	1.022	
logit	-0.020	0.896	(-0.150, 0.277)	0.015	1.011	
<u><math>V_a = 4, (m, n_1) = (20, 50), \lambda_1 = 1</math></u>						
0	0.048	1.011	(-0.034, 0.330)	0.048	1.009	
logit	0.022	0.058	(-0.034, 0.330)	0.048	1.009	
<u><math>V_a = 4, (m, n_1) = (20, 20), \lambda_1 = 1</math></u>						
0	0.077	1.012	(0.082, 0.665)	0.076	1.001	
logit	0.074	0.839	(0.152, 0.696)	0.087	0.940	
<u><math>V_a = 4, (m, n_1) = (20, 30), \lambda_1 = 1</math></u>						
0	0.075	0.995	(0.059, 0.523)	0.075	0.990	
logit	0.062	0.845	(0.114, 0.562)	0.087	0.958	

As we can see from Table A.2, the estimates are biased for larger  $\log m / (\log n_1)^2$ .

Moreover, the variance of  $(\hat{\beta}_1 - \beta_{10})/s.e.$  is sensitive to the initial guess of  $\hat{\beta}_1$ . Also in the combination (20, 20) with  $V_a = 4$  and initial  $\hat{\beta}_1$  equal to logit, their Kolmogorov-Smirnov test rejects the hypothesis of normality with  $p$  values 0.043 and 0.019. For all the other combinations from Table A.2, their Kolmogorov-Smirnov test  $p$  values are larger than 0.1.

Since in reality we do not know how to choose  $\lambda_1$ , we tried various values of  $\lambda_1$  to assess the effect of  $\lambda_1$  on normality and consistency. The Monte Carlo averages and standard deviations of  $(\hat{\beta}_1 - \beta_{10})/s.e.$  are displayed in Table A.3 for various  $(m, n_1)$ ,  $\lambda_1$ , initial values and  $V_a$ . The estimates are computed for  $\lambda_1 = 0.1$ , 1 and 5 for the same data, and this comparison was repeated 1000 times. Table A.3 shows that extremely large values of the penalty parameter ( $\lambda_1 = 5$ ) did affect the consistency. For Table A.3, the combinations (20, 20) with  $V_a = 4$  have their the Kolmogorov-Smirnov test rejects the hypothesis of normality with  $p$  values between 0.019 and 0.043, except for  $\lambda_1 = 5$  whose  $p$  value is bigger than 0.1. For all the other combinations from Table A.3, their Kolmogorov-Smirnov test  $p$  values are larger than 0.1. Generally speaking, the plots show that  $\hat{\beta}_1$  is approximately normal, with departures from normality in the extreme tails.

Table A.3: Simulated standardized estimate of logistic regression coefficient for various  $\lambda_1$  values.

						$(\hat{\beta}_1 - \beta_{10})/s.e.$	$(\hat{\beta}_1 - \beta_{10})/\widehat{s.e.}$
$\lambda_1$	mean	std	95% CI for $\hat{\beta}_1 - \beta_{10}$	mean	std		
<u><math>V_a = 4, (m, n_1) = (10, 20), \text{ initials set to 0}</math></u>							
1	-0.017	0.957	(-0.229, 0.164)	-0.018	0.947		
0.1	-0.020	0.896	(-0.150, 0.277)	-0.009	1.002		
5	-0.020	0.914	(-0.164, 0.258)	0.003	1.020		
<u><math>V_a = 3, (m, n_1) = (20, 20), \text{ initials set to 0}</math></u>							
1	0.054	0.938	(-0.008, 0.509)	0.053	0.932		
0.1	0.037	0.889	(-0.021, 0.520)	0.052	0.975		
5	0.037	0.889	(-0.021, 0.520)	0.052	0.975		
<u><math>V_a = 4, (m, n_1) = (20, 50), \text{ initials set to 0}</math></u>							
1	0.048	1.012	(-0.034, 0.330)	0.048	1.009		
0.1	0.022	0.858	(-0.034, 0.330)	0.048	1.009		
5	0.022	0.858	(0.034, 0.400)	0.070	1.017		
<u><math>V_a = 4, (m, n_1) = (20, 20), \text{ initials set to 0}</math></u>							
1	0.077	1.012	(0.082, 0.665)	0.076	1.001		
0.1	0.074	0.839	(0.250, 0.800)	0.087	0.940		
5	0.078	0.855	(0.250, 0.800)	0.108	0.959		

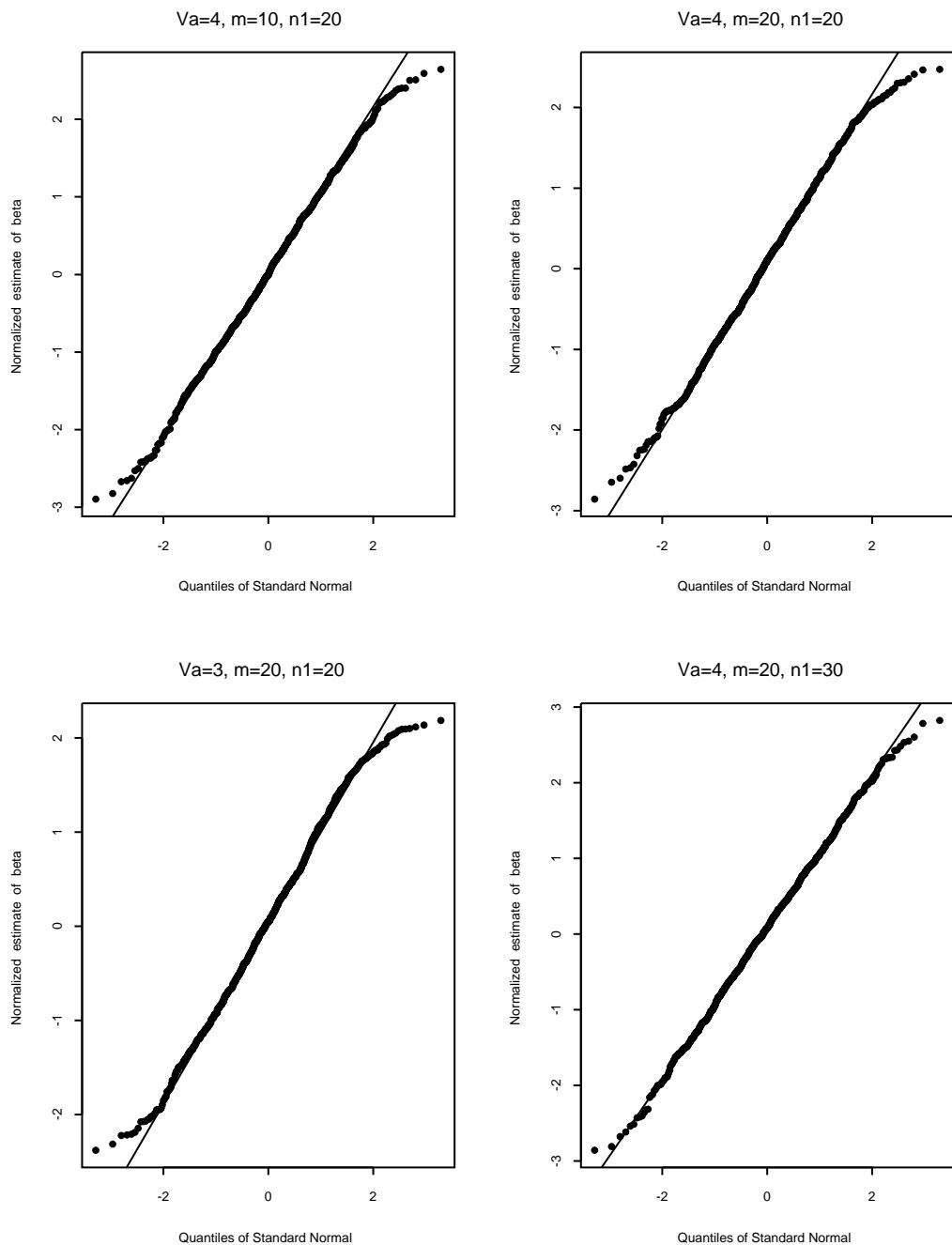


Figure A.1: Q-Q plots for  $(\hat{\beta}_1 - \beta_{10})$  standardized by true conditional standard error combining  $a$  with  $v_i$ , for various  $(m, n_1)$  in logistic random intercept case 1.

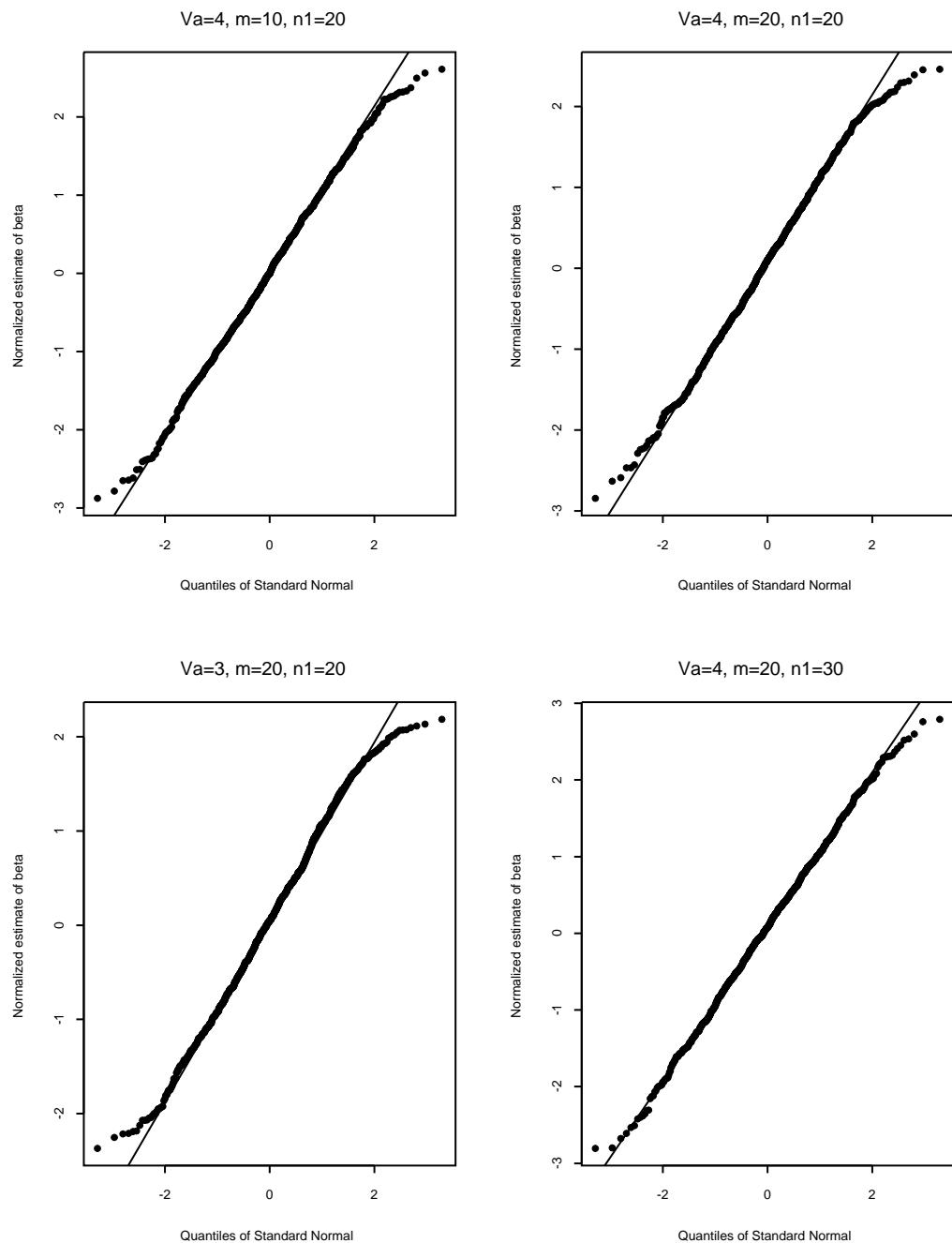


Figure A.2: Q-Q plots for  $(\hat{\beta}_1 - \beta_{10})$  standardized by estimated conditional standard error combining  $a$  with  $v_i$ , for various  $(m, n_1)$  in logistic random intercept case 1.

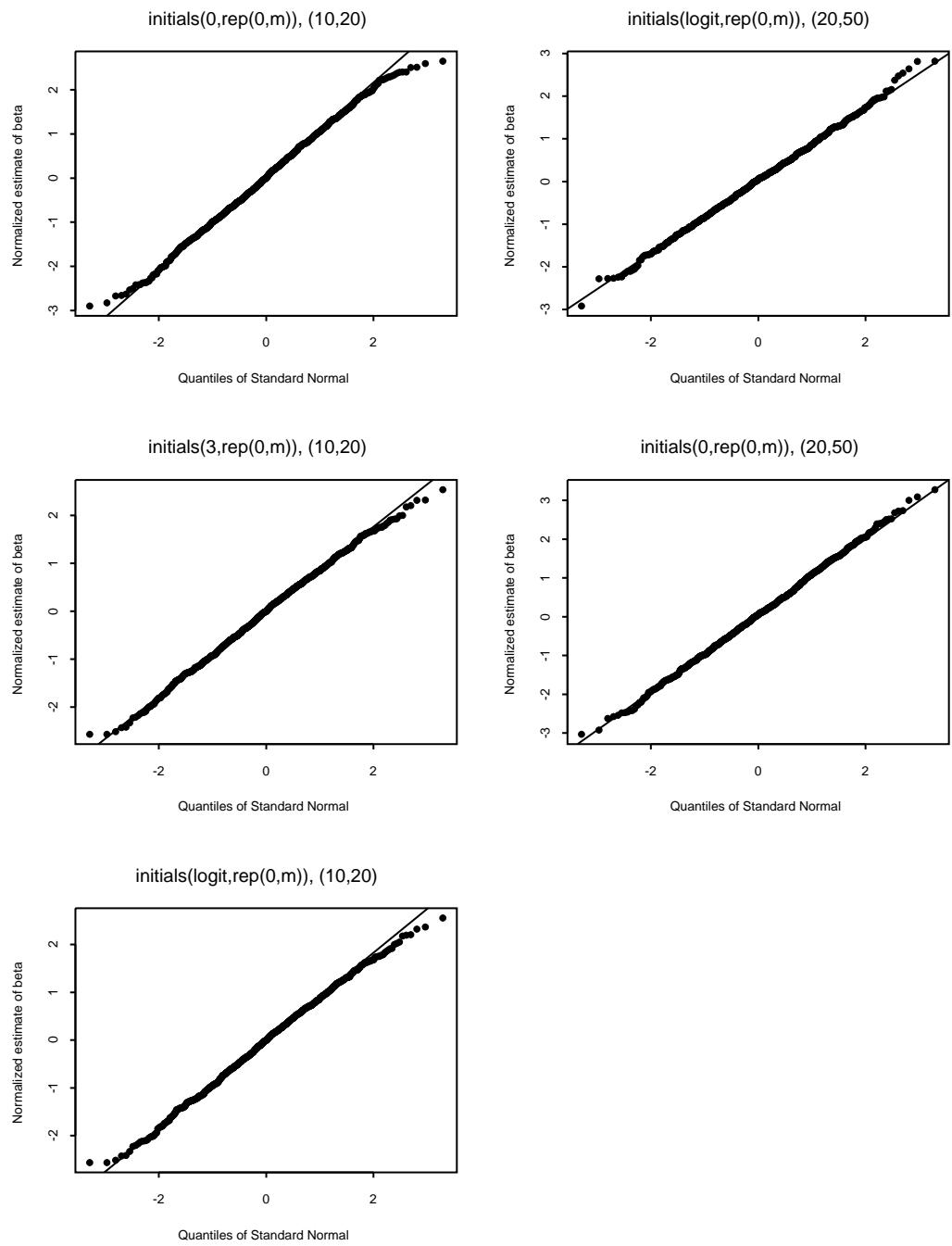


Figure A.3: Q-Q plots for  $(\hat{\beta}_1 - \beta_{10})$  standardized by true conditional standard error combining  $a$  with  $v_i$ , for various initial values of  $\hat{\beta}_1$  in logistic random intercept case 1.

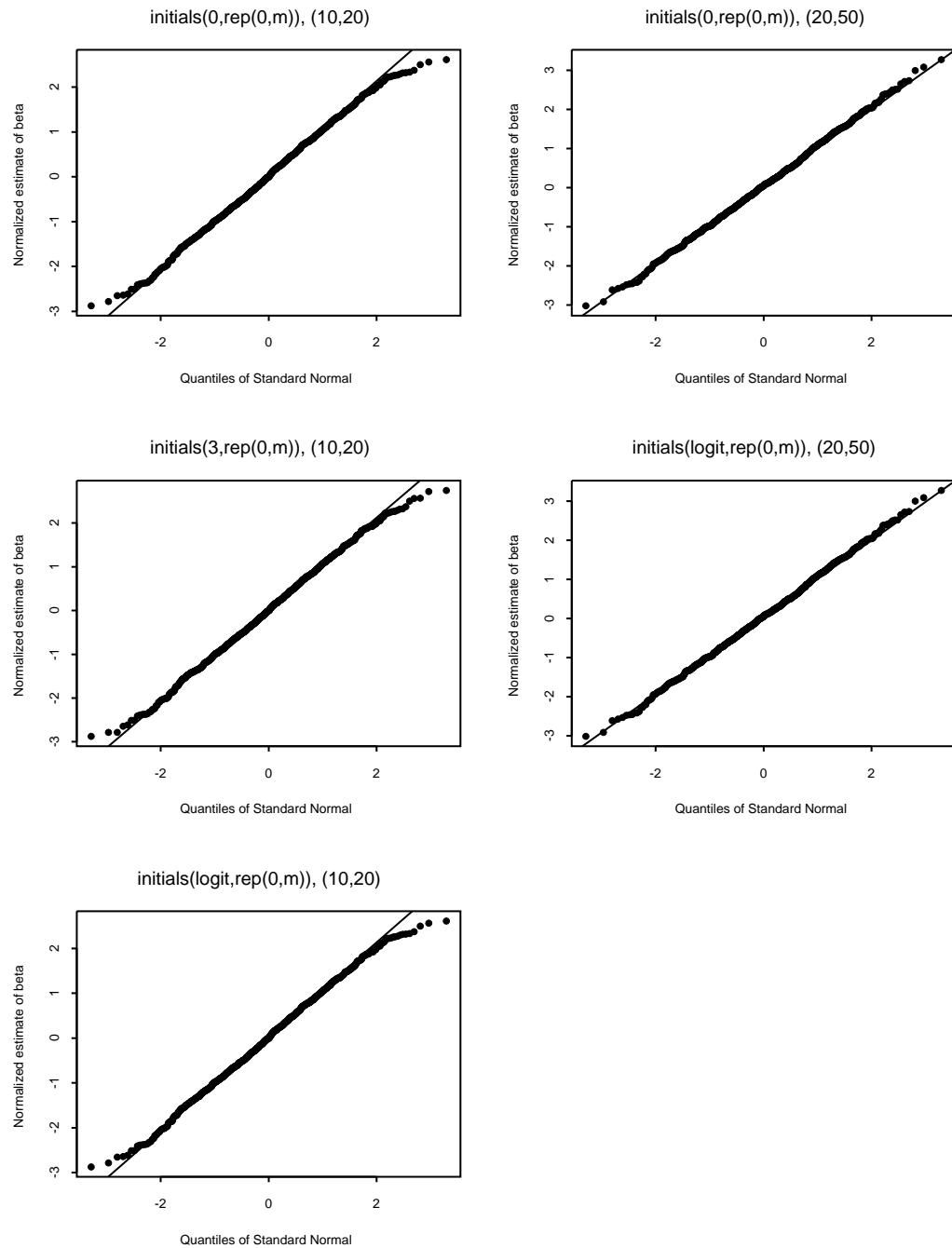


Figure A.4: Q-Q plots for  $(\hat{\beta}_1 - \beta_{10})$  standardized by estimated standard error combining  $a$  with  $v_i$ , for various initial values of  $\hat{\beta}_1$  in logistic random intercept case 1.

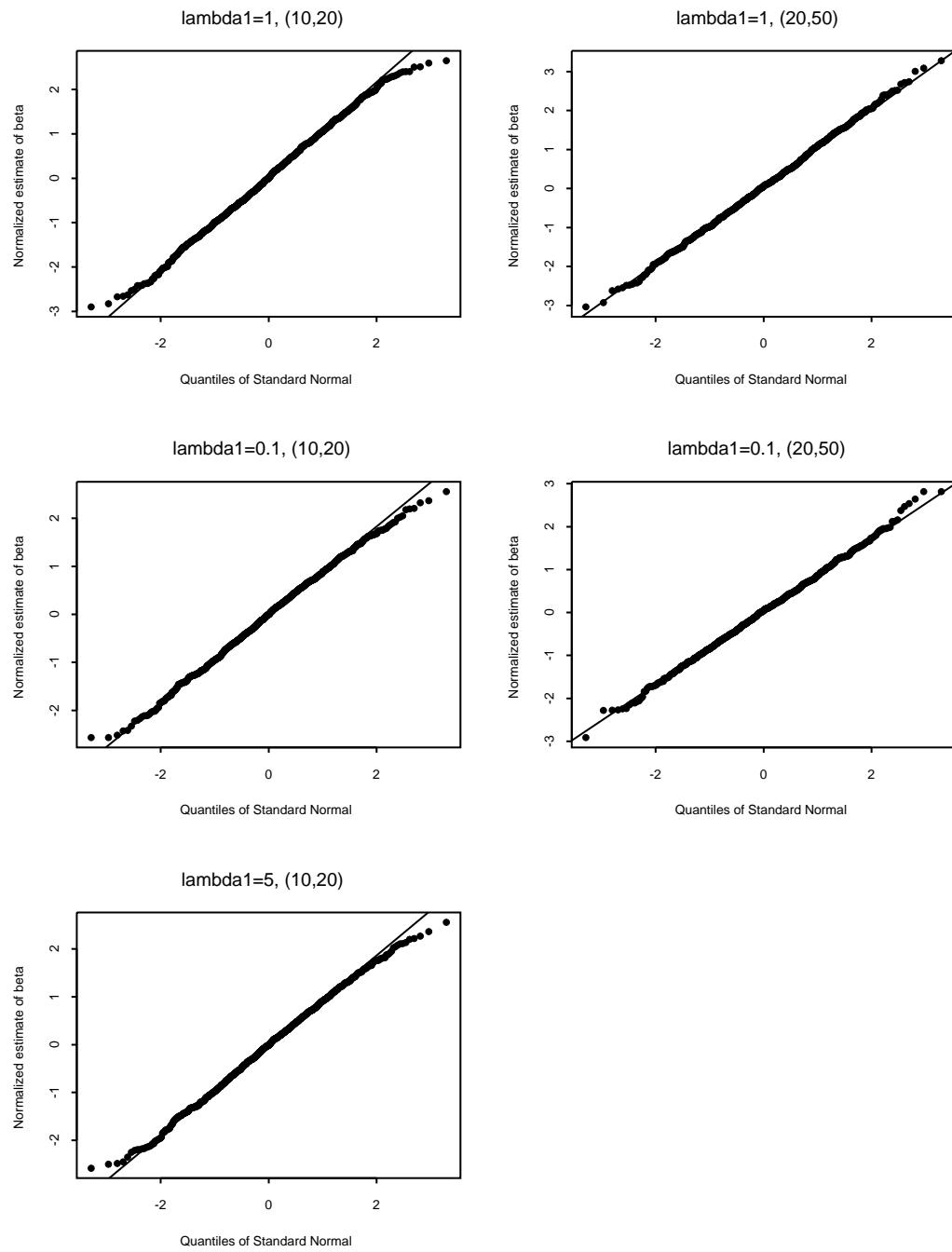


Figure A.5: Q-Q plots for  $(\hat{\beta}_1 - \beta_{10})$  standardized by true conditional standard error combining  $a$  with  $v_i$ , for various values of  $\lambda_1$  in logistic random intercept case 1.

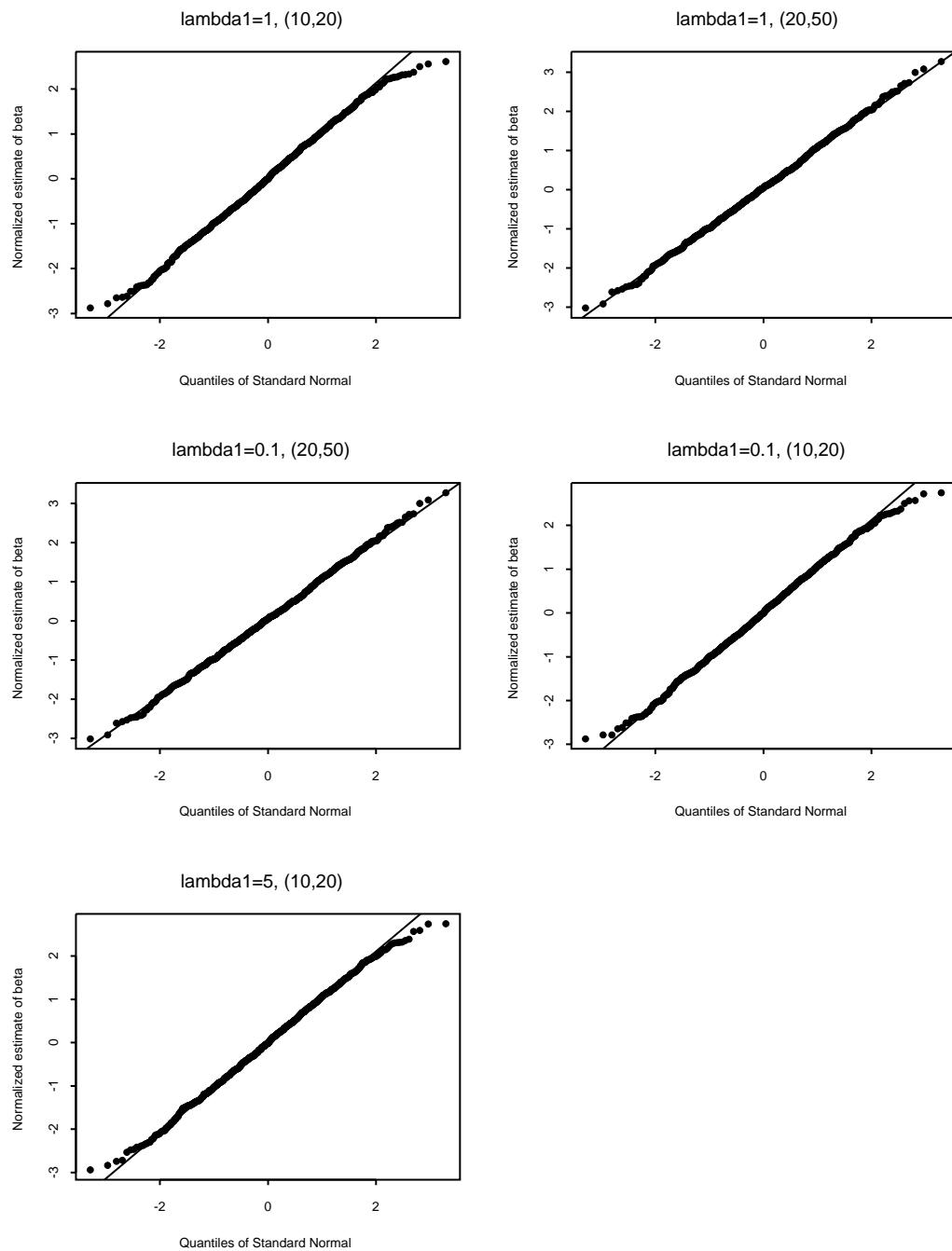


Figure A.6: Q-Q plots for  $(\hat{\beta}_1 - \beta_{10})$  standardized by estimated conditional standard error combining  $a$  with  $v_i$ , for various values of  $\lambda_1$  in logistic random intercept case 1.

### A.1.2 Case 1 logistic random intercept when $a$ and $v_i$ are estimated separately.

We consider the following example from Jiang [14]:

$$\text{logit}P(y_{ij} = 1|v) = a + v_i + x_{ij}\beta_1. \quad (\text{A.2})$$

Since  $\hat{a}_0$  and  $\hat{\beta}_1$  have different convergence rates, we do not consider joint asymptotic distribution. We are interested in  $\hat{\beta}_1$  and use the estimating equation (??) and the Splus function `nlinnb` to investigate the asymptotic behavior of  $\hat{\beta}_1$ .

We use a balanced setup with  $m$  clusters and  $n_1$  observations per cluster. We generate  $m$  random effects  $v_i$  from the normal distribution with mean 0 and variance  $V_a$ . Conditionally on  $(x_{ij}, v_i)$  with  $\beta_{10} = a_0 = 1$ , a sample of  $mn_1$  random variables  $Y_{ij}$  is generated from Bernoulli distribution with  $E(Y_{ij}|v_i, x_{ij}) = \exp(x_{ij}\beta_1 + a + v_i)/(1 + \exp(x_{ij}\beta_1 + a + v_i))$ .

Various combinations of  $(m, n_1)$ ,  $V_a$ ,  $\lambda_1$  were simulated and all combinations initialized  $\hat{a}$ ,  $\hat{\beta}_1$  and  $\hat{v}_i$  at zero.

For each combination, 1000 random replications of random effects  $v_i$  and the responses  $Y_{ij}$  are generated. Estimated standardized regression coefficients  $\hat{\beta}_1$  standardized by true conditional standardized error and estimated conditional standardized error are summarized in Table A.4.

Table A.4: Simulated standardized estimates of logistic regression coefficient (standardized by true and estimated conditional standardized error) with  $a$  and  $v_i$  estimated separately.

$(m, n_1)$	$(\hat{\beta}_1 - \beta_{10})/s.e.$			$(\hat{\beta}_1 - \beta_{10})/\widehat{s.e.}$		
	mean	std	95% CI for $\hat{\beta}_1 - \beta_{10}$	mean	std	
<u><math>V_a = 4</math>, initials set to 0, <math>\lambda_1=1</math></u>						
(10,20)	-0.007	0.999	(-0.207, 0.223)	-0.009	0.976	
(20,20)	0.018	0.903	(-0.344, 0.194)	-0.019	0.890	
(20,30)	0.011	0.970	(-0.190, 0.280)	0.010	0.963	
(20,40)	0.054	1.030	(-0.110, 0.276)	0.053	0.956	
<u><math>V_a = 3</math>, initials set to 0, <math>\lambda_1 = 1</math></u>						
(20,20)	0.020	0.924	(-0.160, 0.370)	0.018	0.910	
(20,30)	0.020	0.975	(-0.159, 0.290)	0.020	0.967	
<u><math>V_a = 2</math>, initials set to 0, <math>\lambda_1 = 1</math></u>						
(20,20)	0.014	0.963	(-0.190, 0.340)	0.012	0.950	

For the above Table (A.4), the 95% confidence intervals for  $(\hat{\beta}_1 - \beta_{10})$  include 0. Normality holds for  $(\hat{\beta}_1 - \beta_{10})$  standardized by true conditional standard error given  $\mathbf{v}_0$ . The mean and standard error estimates are close to 0 and 1 respectively, except for the pairs as (20, 20) and (20, 30) with  $V_a = 4$ . In these cases, their the Kolmogorov-Smirnov test rejects the hypothesis of normality with p-values of 0.0179 and 0.0171 respectively. Even for  $V_a = 3$ ,  $(m, n_1) = (20, 20)$ , the p-value of 0.06 is

almost significant.

Normality holds for  $(\hat{\beta}_1 - \beta_{10})$  standardized by estimated conditional standard error given  $\mathbf{v}_0$ . The mean and standard error estimates are close to 0 and 1 respectively, except for the pairs as (20, 20), (20, 30) with  $V_a = 4$  and (20, 20) with  $V_a = 3$ . In these cases, the Kolmogorov-Smirnov test rejects the hypothesis of normality with p-values of 0.014, 0.011, 0.04 respectively.

For all the other combinations for Table A.4, their Kolmogorov-Smirnov test  $p$  values are larger than 0.1.

The 95% confidence intervals for  $(\hat{a} - a_0)$  do not include 0 (the intervals are not listed in the table but all were approximately (0.1, 0.4)). The Kolmogorov-Smirnov test applied to Monte Carlo distribution of  $\hat{a}$  significantly rejects the hypothesis of normality. From the theoretical results of Chapter 3, we know the convergence rates differ between estimated logistic regression coefficient and intercept, with rates approximately  $1/N$  and  $1/m$ , respectively. Here we have sample sizes from 200 to 1000, but number of clusters 10 or 20. Even in a small simulation with  $(m, n_1) = (40, 60)$  with 200 replications, there is bias present for estimated logistic regression intercept.

The Q-Q plots show that the standardized  $(\hat{\beta}_1 - \beta_{10})$  is approximately normal, with departures from normality in the extreme tails.

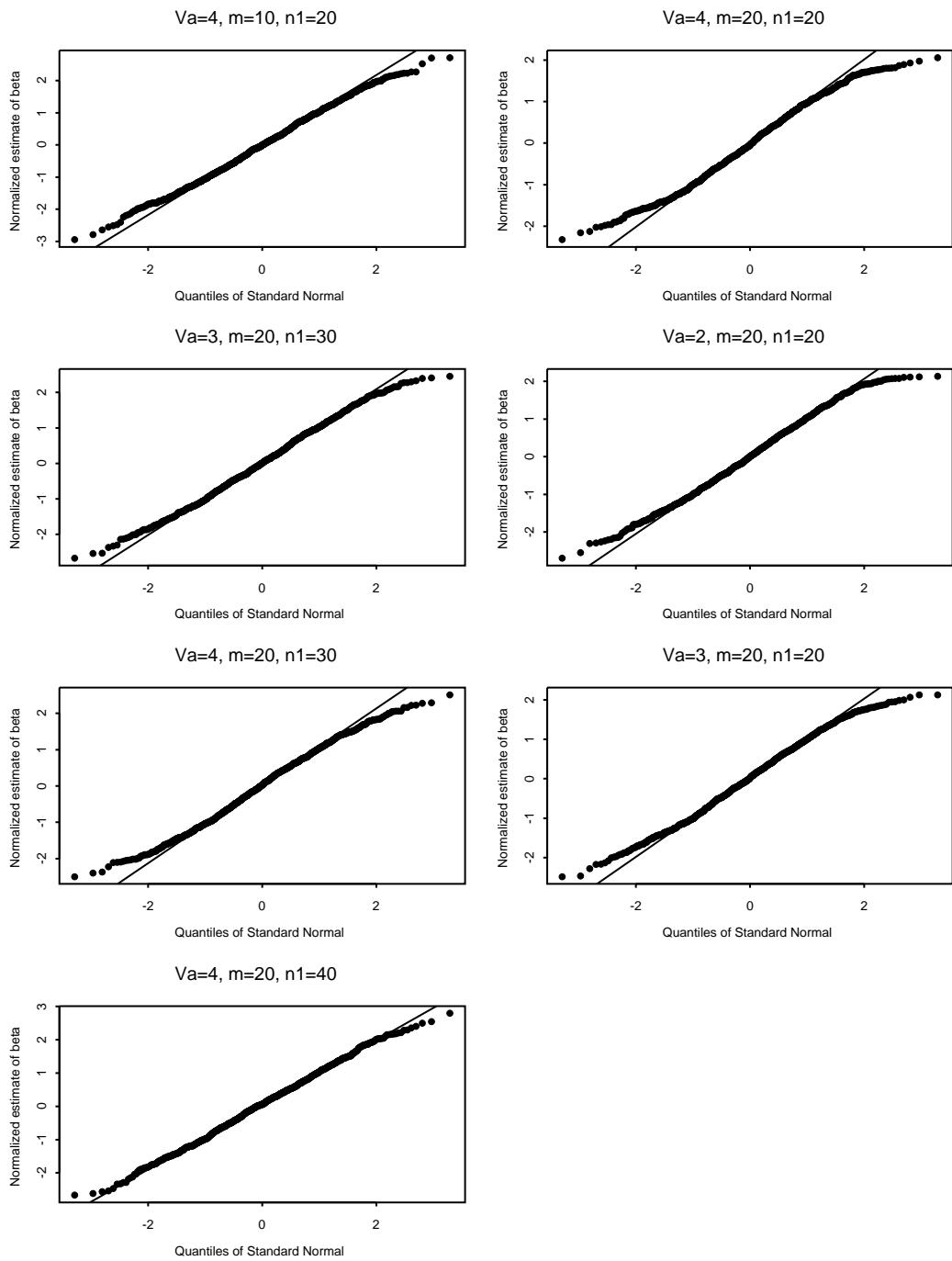


Figure A.7: Q-Q plot for  $(\hat{\beta}_1 - \beta_{10})$  standardized by true conditional standard error not combining  $a$  with  $v_i$ , for various  $(m, n_1)$  in logistic random intercept case 1.

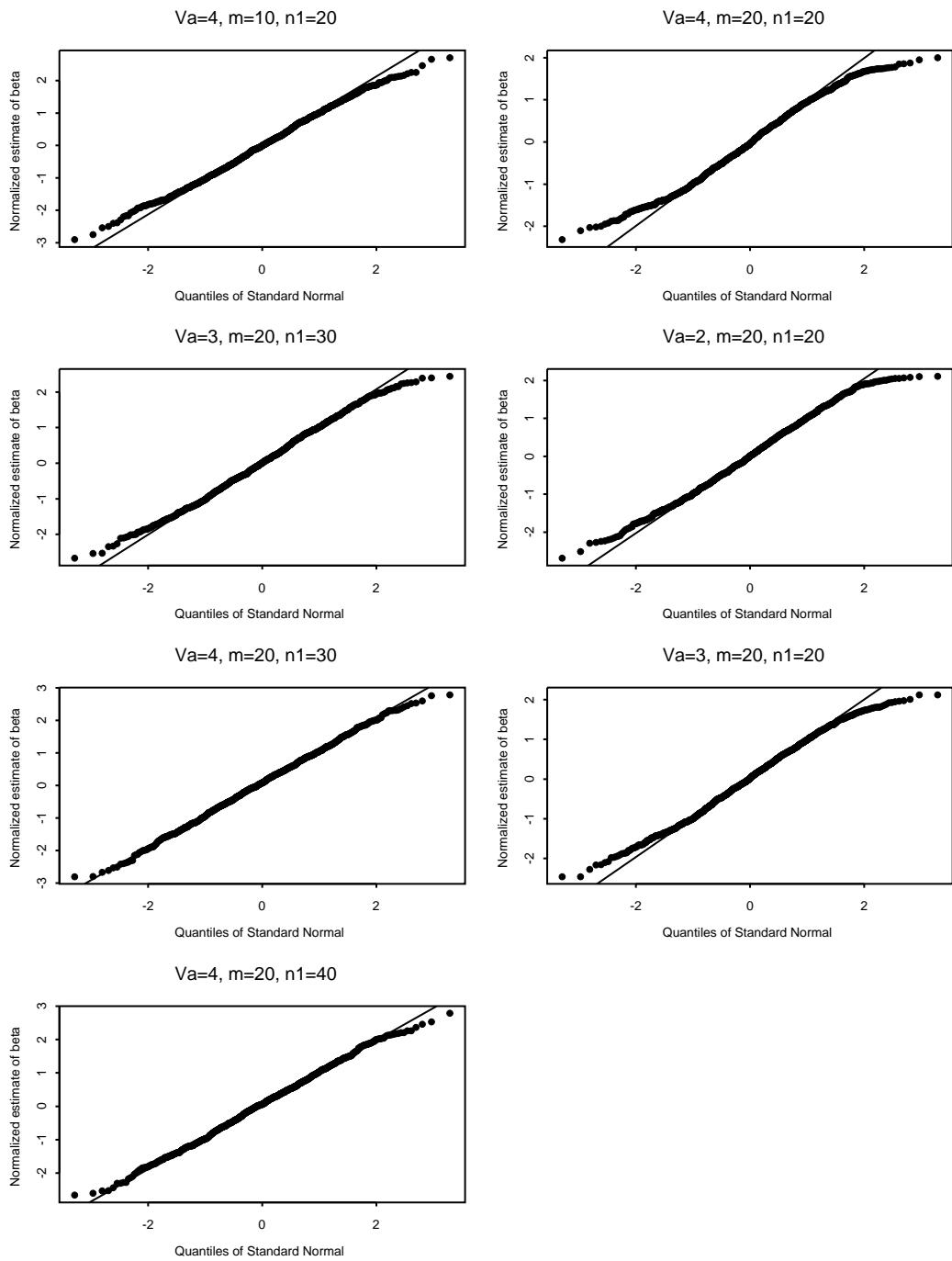


Figure A.8: Q-Q plot for  $(\hat{\beta}_1 - \beta_{10})$  standardized by estimated conditional standard error with estimated  $a$  with  $v_i$  separately, for various  $(m, n_1)$  in logistic random intercept case 1.

### A.1.3 Case 1 Poisson random intercept combining $a$ and $v_i$ .

We use the PGWLS method by Jiang [14] and `nlinb` minimization function in Splus to get estimators which minimize negative  $l_P(\gamma)$  as (??), Here we have

$$-l_P(\gamma) = \sum_i \sum_j (-y_{ij}\eta_{ij} + \exp(\eta_{ij})) + \frac{\lambda_1}{2}m(\bar{v})^2$$

where  $\eta_{ij} = a_i + \beta_1 x_{ij}$ ,  $a_i = a + v_i = v_i$  (in our simulation we let  $a = 0$ ) and  $\bar{v} = \sum_i^m v_i/m$ ,  $\gamma = (\beta_1, v_1, \dots, v_m)$ . Here  $a$  and  $\beta_1$  are fixed parameters,  $x_{ij}$  is a scalar valued predictor, and  $v_i$  are iid  $N(0, V_a)$ . We do not attempt to estimate  $a$  and  $v_i$  separately. For simplicity we simulated the balanced model with  $m$  clusters and  $n_1$  observations per cluster. We generate  $m$  random effects  $v_i$  from normal distribution with mean 0 and variance  $V_a$ . Covariates  $x_{ij}$  are the same here as in logistic Case 1. Conditionally on  $(x_{ij}, v_i)$  with  $\beta_{10} = 1$ , a sample of  $mn_1$  random variables  $Y_{ij}$  were generated from Poisson distribution with  $E(Y_{ij}|v_i, x_{ij}) = \exp(x_{ij} + v_i)$ .

Various combinations of  $(m, n_1)$ ,  $V_a$  and  $\lambda_1$  values are investigated.

For each combination, 1000 replications of random effects  $v_i$  and responses  $Y_{ij}$  were generated. Estimated regression coefficients, for various choices of  $(m, n_1)$  with  $V_a = 4$ ,  $\lambda_1 = 1$  and initial estimates set to 0, are summarized in Table A.5. The table displays the means and standard errors of the simulated values of  $(\hat{\beta}_1 - \beta_{10})/s.e.$  and 95% confidence bounds for  $E(\hat{\beta}_1 - \beta_{10})$  based on Student's t.

In order to check sensitivity of  $\lambda_1$  values, the estimates are computed for  $\lambda_1 = 0.1, 1$  and  $5$  for the same data, and this comparison was repeated 1000 times. From Table A.5 and Table A.6, consistency and normality results hold and are not sensitive to the values of  $\lambda_1$ . In these cases, their Kolmogorov-Smirnov test  $p$  values are larger than 0.1 except for the combination (20,20) with  $V_a = 4$ . In this case, the Kolmogorov-Smirnov test almost rejects the normality hypothesis with  $p$  value 0.06. The Q-Q plot shows that standardized  $(\hat{\beta}_1 - \beta_{10})$  is approximately normal.

Table A.5: Simulated standardized estimate of Poisson regression coefficient for various  $(m, n_1)$  values.

$(m, n_1)$	$(\hat{\beta}_1 - \beta_{10})/s.e.$			$(\hat{\beta}_1 - \beta_{10})/\widehat{s.e.}$		
	mean	std	95% CI for $\hat{\beta}_1 - \beta_{10}$	mean	std	
$V_a = 4, \lambda_1 = 1, \text{ initials set to 0}$						
(10,20)	0.018	0.970	(-0.022, 0.064)	0.018	0.969	
(20,20)	-0.005	1.008	(-0.050, 0.054)	0.005	1.008	
(20,30)	-0.0006	1.007	(-0.042, 0.039)	0.006	1.007	

Table A.6: Simulated standardized estimate of logistic regression coefficient for various  $\lambda_1$  values.

$\lambda_1$	$(\hat{\beta}_1 - \beta_{10})/s.e.$			$(\hat{\beta}_1 - \beta_{10})/\widehat{s.e.}$		
	mean	std	95% CI for $\hat{\beta}_1 - \beta_{10}$	mean	std	
<u><math>V_a = 4, (m, n_1) = (10, 20), \text{ initials set to 0}</math></u>						
1	0.018	0.970	(-0.022, 0.064)	0.018	0.969	
0.1	0.015	1.100	(-0.026, 0.064)	0.013	1.024	
5	0.028	1.021	(-0.018, 0.064)	0.022	0.965	
<u><math>V_a = 4, (m, n_1) = (20, 30), \text{ initials set to 0}</math></u>						
1	0.006	1.007	(-0.042, 0.039)	0.006	1.007	
0.1	0.001	1.025	(-0.044, 0.033)	-0.002	1.008	
5	0.009	1.012	(-0.038, 0.043)	0.006	1.003	

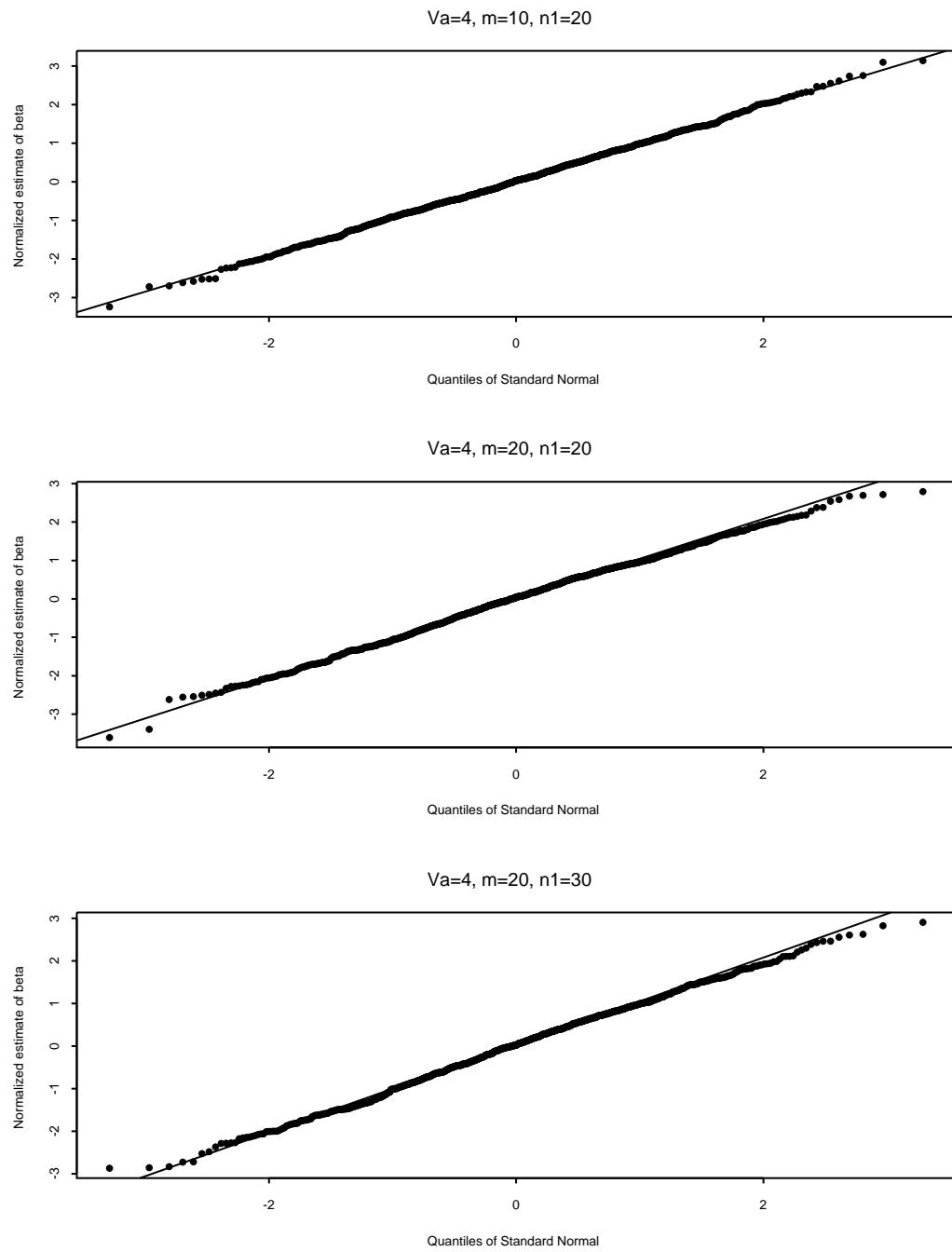


Figure A.9: Q-Q plots for  $(\hat{\beta}_1 - \beta_{10})$  standardized by true conditional standard error with  $a$  and  $v_i$  estimated separately in Poisson random intercept case 1.

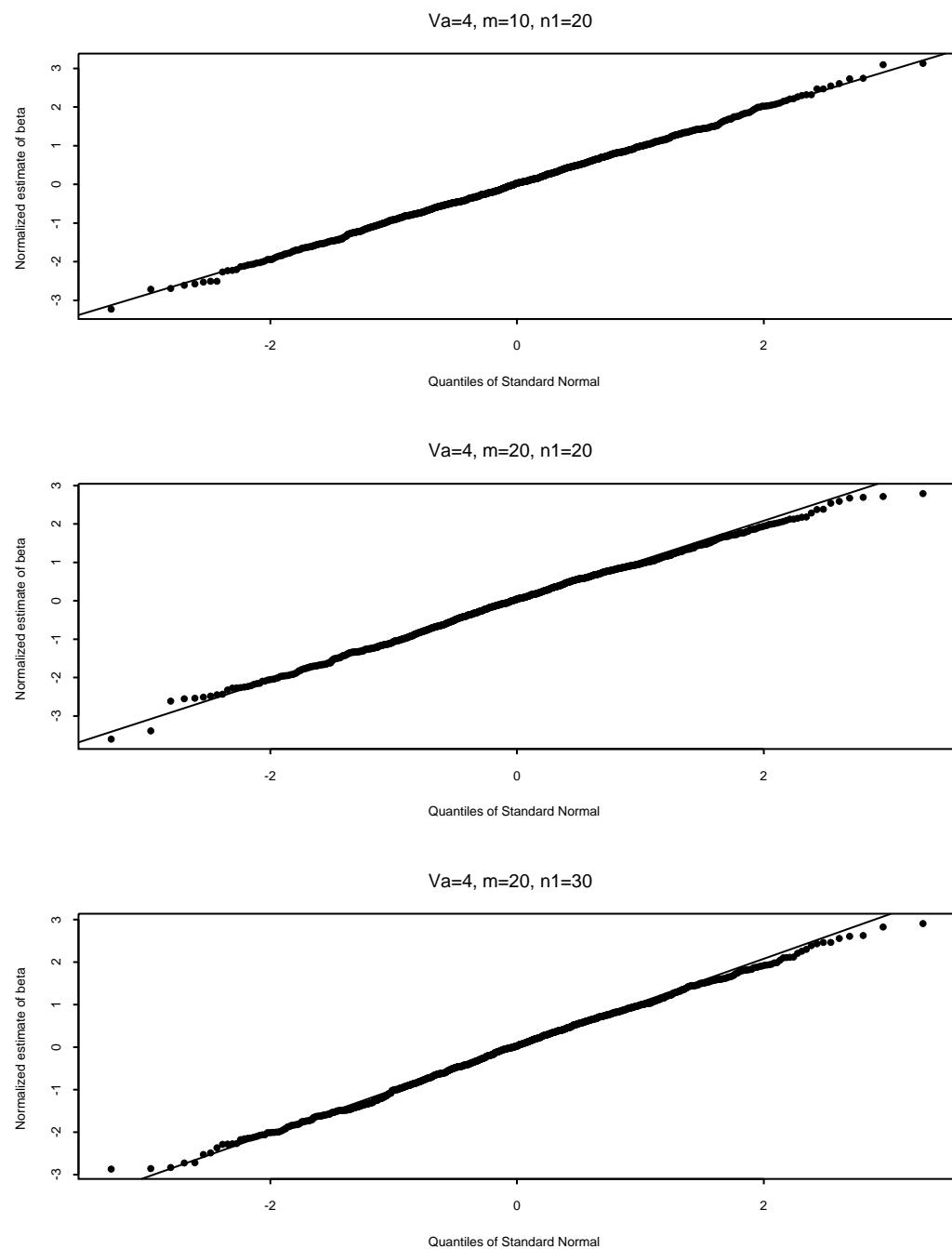


Figure A.10: Q-Q plots for  $(\hat{\beta}_1 - \beta_{10})$  standardized by estimated conditional standard error with  $a$  and  $v_i$  estimated separately in Poisson random intercept case 1.

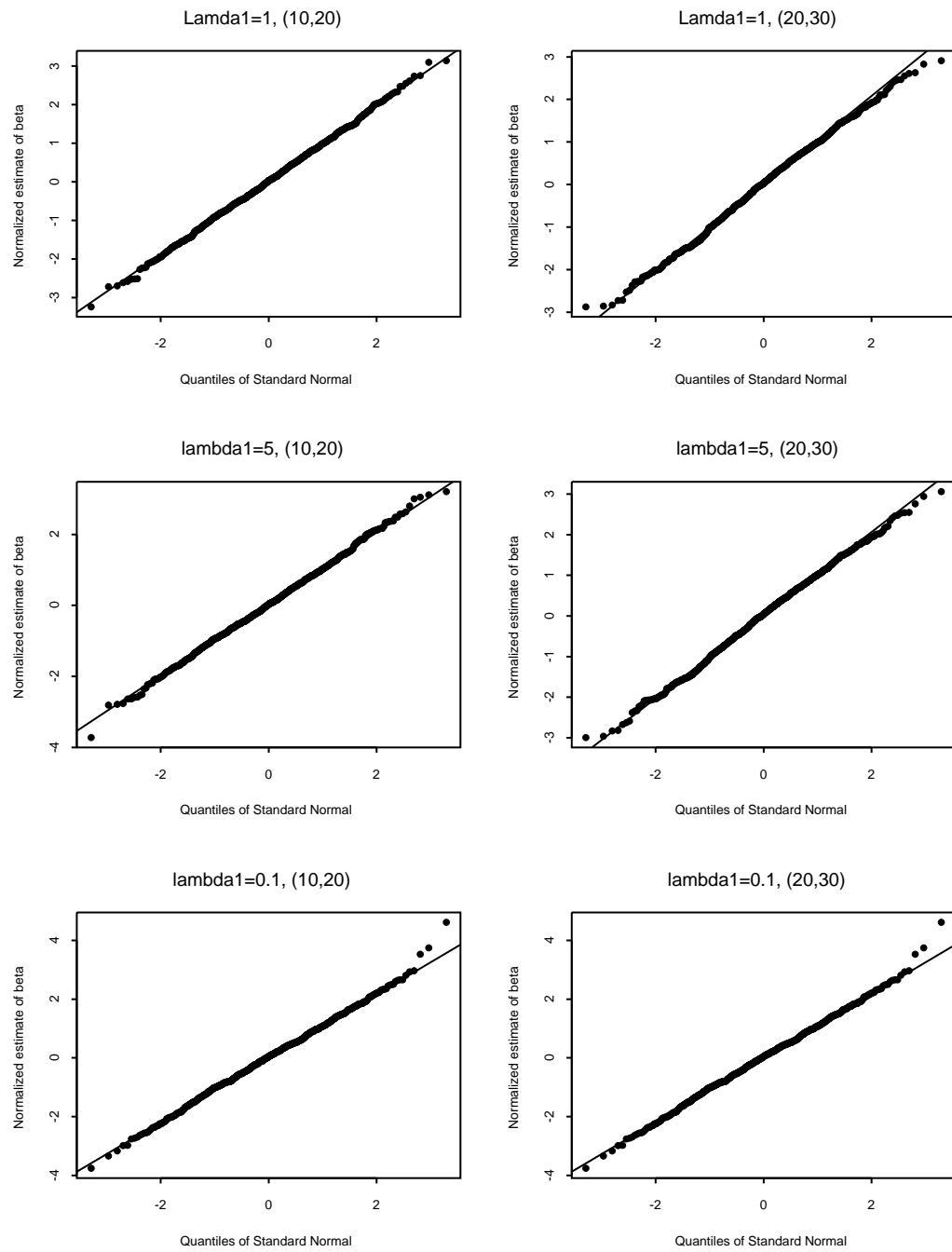


Figure A.11: Q-Q plots for  $(\hat{\beta}_1 - \beta_{10})$  for standardized by true conditional standard error with  $a$  and  $v_i$  estimated separately, for various values of  $\lambda_1$  in Poisson random intercept case 1.

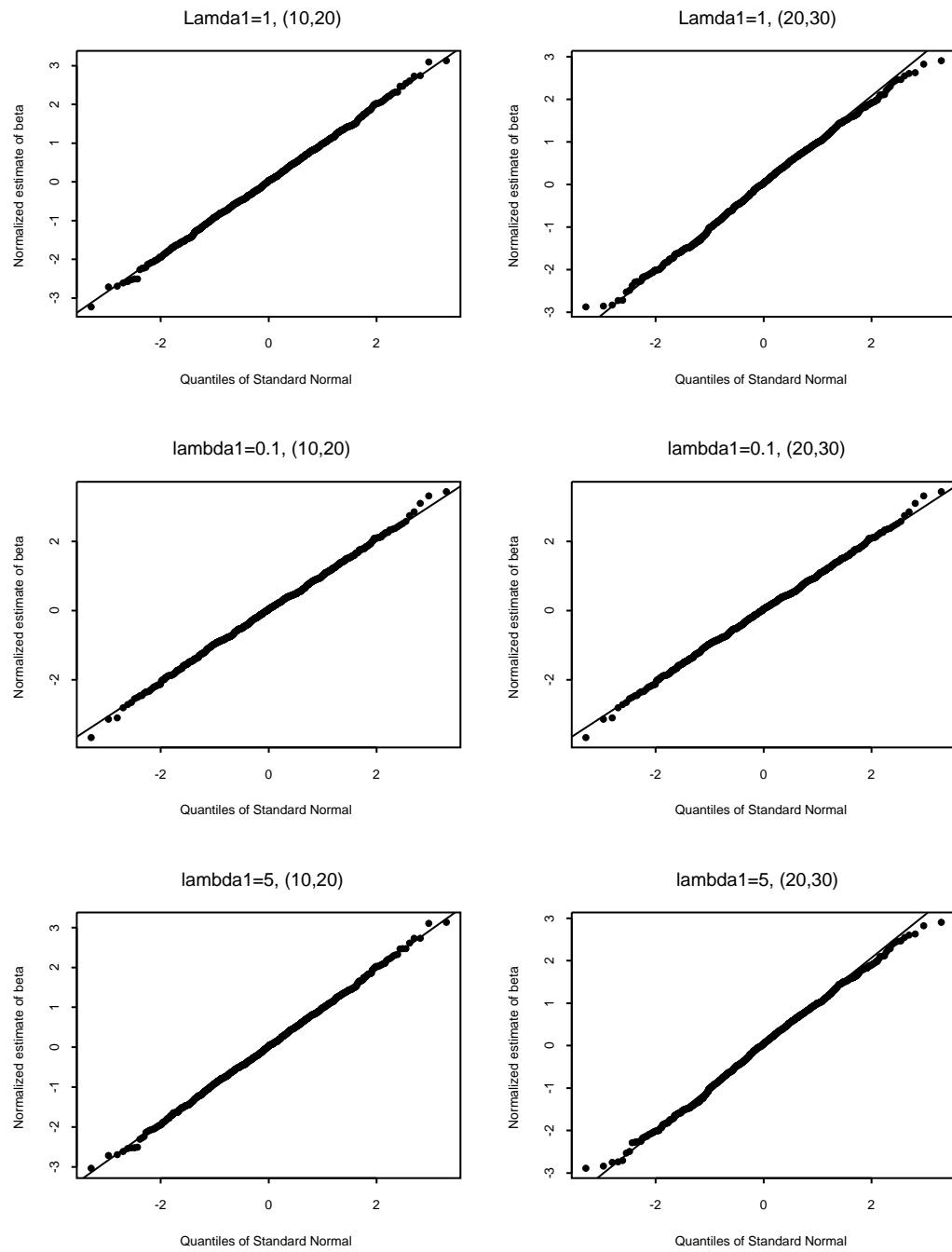


Figure A.12: Q-Q plots for  $(\hat{\beta}_1 - \beta_{10})$  standardized by estimated conditional standard error with  $a$  and  $v_i$  estimated separately, for various values of  $\lambda_1$  in Poisson random intercept case 1.

#### A.1.4 Case 1 Poisson random intercept with $a$ and $v_i$ estimated separately

We use the PGWLS method by Jiang [14] and `nlminb` minimization function in Splus to get estimators which minimize negative  $l_P(\gamma)$  as (??). Here we have

$$-l_P(\gamma) = \sum_i \sum_j (-y_{ij}\eta_{ij} + \exp(\eta_{ij})) + \frac{\lambda_1}{2}m(\bar{v})^2$$

where  $\eta_{ij} = a + v_i + \beta_1 x_{ij}$  and  $\bar{v} = \sum_i^m v_i/m$ ,  $\gamma = (\beta_1, a, v_1, \dots, v_m)$ . Here  $a$  and  $\beta_1$  are fixed parameters,  $x_{ij}$  is a scalar valued predictor, and  $v_i$  are iid  $N(0, V_a)$ . We attempt to estimate  $a$  and  $v_i$  separately. For simplicity we simulated the balanced model with  $m$  clusters and  $n_1$  observations per cluster. We generate  $m$  random effects  $v_i$  from normal distribution with mean 0 and variance  $V_a$ . Covariates  $x_{ij}$  are the same here as in logistic Case 1. Conditionally on  $(x_{ij}, v_i)$  with  $\beta_{10} = a_0 = 1$ , a sample of  $mn_1$  random variables  $Y_{ij}$  were generated from Poisson distribution with  $E(Y_{ij}|v_i, x_{ij}) = \exp(\beta_1 x_{ij} + v_i + a)$ .

The same choices of  $(m, n_1)$  and  $V_a$  values as in the previous subsection were simulated.

Unlike the results when only  $v_i + a$  were estimated, as in the previous subsection, when we attempted to estimate  $a$  and  $v_i$  separately, none of the desired asymptotic results for  $\hat{\beta}_1$  were observed. The estimate  $\hat{\beta}_1$  was biased and its Monte Carlo distribution failed tests of normality.

## Bibliography

- [1] E. B. Andersen, “Asymptotic properties of conditional maximum-Likelihood estimators”, *J. Roy. Statist. Soc. Ser. B* **32**, 283-301 (1970).
- [2] R. H. Berk, “Consistency and asymptotic normality of MLE’s for exponential models,” *The Annals of Mathematical Statistics*, **43**, 193-204 (1972).
- [3] R. Bhatia, *Matrix Analysis*, Springer (1991).
- [4] O. Barndorff-Nielsen, “On a formula for the distribution of the maximum likelihood estimate,” *Biometrika*, **70**, 343-365 (1983).
- [5] N. Breslow “Odds ratio estimators when the data are sparse,” *Biometrika*, **68**, 73-84 (1981).
- [6] E. Cantoni, “Robust inference for generalized linear models,” *J. Amer. Statist. Assoc* **96**, 1022-1030 (2001).
- [7] J. M. Chiou and H. G. Muller, “Estimated estimating equations: semiparametric inference for clustered and longitudinal data,” *J. Roy. Statist. Soc. Ser.B* **67**, 531-553 (2005).
- [8] N. E. Breslow and D. G. Clayton, “Approximate inference in generalized linear mixed models,” *J. Amer. Statist. Assoc.* **88**, 9-25 (1993).
- [9] A. C. Davison, “Approximate conditional inference in generalized linear models,” *J. Roy. Statist. Soc. Ser. B*, **50**, 445-461 (1988).
- [10] I. Domowitz, “Misspecified models with dependent observations,” *J. of Econometrics* **20**, 35-38 (1982).
- [11] L. Fahrmeir and H. Kaufmann, “Consistency and asymptotic normality of the maximum likelihood estimators in generalized linear models,” *The Annals of Statistics* **13**, 342-368 (1985).
- [12] L. Fahrmeir, “Maximum likelihood estimation in misspecified generalized linear models,” *Akademie-Verlag Berlin statistics* **21** 4, 487-502 (1990).
- [13] S. J. Haberman, “Maximum likelihood estimates in exponential response models,” *Annals of Statistics* **5**, 815-841 (1977).

- [14] J. Jiang, “Conditional inference about generalized linear mixed models,” *The Annals of Statistics* **27**, 1974-2007 (1999).
- [15] J. Jiang, H. Jia and H. Chen, “Maximum posterior estimates of random effects in generalized linear mixed models,” *Statistica Sinica* **11**, 97-120 (2001).
- [16] X. P. Jiang, Ph.D. thesis, “Nonparametric quasi-likelihood in longitudinal data analysis,” University of Maryland, College Park, (2004).
- [17] Y. Lee and J. A. Nelder, “Hierarchical generalized linear models,” *J. Roy. Statist. Soc. Ser. B* **58**, 619-678 (1996).
- [18] H. Li, B. G. Lindsay and R. P. Waterman, “Efficiency of projected score methods in rectangular array asymptotics,” *J. Roy. Statist. Soc. Ser. B*, **65**, 191-208 (2003)
- [19] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, (2nd ed. Chapman and Hall, New York, 1989).
- [20] P. McCullagh, “Quasi-likelihood functions,” *Annals of Statistics* **11**, 59-67 (1983).
- [21] C. E. McCulloch, “Maximum likelihood algorithms for generalized linear mixed models,” *J. Amer. Statist. Assoc.* **92**, 162-170 (1997).
- [22] C. E. McCulloch and S. R. Searle, *Generalized, Linear, and Mixed Models* (Wiley-Interscience Publication, 2001).
- [23] J. A. Nelder and R. W. M. Wedderburn, “Generalized linear models,” *J. Roy. Statist. Soc. Ser.A* **135**, 370-384 (1972).
- [24] J. Neyman and E. L. Scott, “Consistent estimates based on partially consistent observations,” *Econometrica*. **16**, 1-32 (1948).
- [25] J. Robins, N. Breslow and S. Greenland, “Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models,” *Biometrics* **42**, 311-323 (1986).
- [26] P. Royston “A remark on algorithm AS 181: The W test for normality” *Applied Statistics* **44** 547-551 (1995).
- [27] S.S.Shapiro and M.B.Wilk “An analysis of variance test for normality (complete samples)” *Biometrika* **52** 591-611 (1965)

- [28] N. Sartori and T. A. Severini, “Conditional likelihood inference in generalized linear mixed models,” *Statistica Sinica* **14**, 349-360 (2004).
- [29] J. R. Schott, *Matrix Analysis for Statistics*
- [30] S. K. Sinha, “Robust analysis of generalized linear mixed models,” *J. Amer. Statist. Assoc.* **99**, 451-460 (2004).
- [31] G. W. Stewart, *Introduction to Matrix Computations* (Academic Press, 1973).
- [32] A. W. Van der Vaart, *Asymptotic Statistics* (Cambridge, 1998).
- [33] R. W. M. Wedderburn, “On the existence and uniqueness of the maximum likelihood estimates for generalized linear models,” *Biometrika* **63**, 27-32 (1976).
- [34] R. W. M. Wedderburn, “Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method,” *Biometrika*, **61**, 439-447 (1974).
- [35] X. Xie and Y. Yang, “Asymptotics for generalized estimating equations with large cluster sizes,” *The Annals of Statistics* **31**, 310-347 (2003).
- [36] S. Yusuf, R. Peto, J. Lewis, R. Collins and P. Sleight “Beta blockade during and after myocardial infarction: an overview of the randomized trials” *Progress in Cardiovascular Diseases* **27** 335-371 (1985)
- [37] S. L. Zeger and K. -Y. Liang, “Longitudinal data analysis for discrete and continuous outcomes,” *Biometrics*, **42**, 121-130 (1986).
- [38] M. Znidaric, “Asymptotic expansion for inverse moments of Binomial and Poisson Distributions’.