8-2013

Adjusted variance components for unbalanced clustered binary data models

Intesar N. Abdoslam

UNIVERSITY OF NORTHERN COLORADO
Greeley, Colorado
The Graduate School

ADJUSTED VARIANCE COMPONENTS FOR UNBALANCED CLUSTERED BINARY DATA MODELS

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

Intesar N. Abdoslam

College of Education & Behavioral Science
School of Educational Research, Leadership, & Technology
Department of Applied Statistics & Research Methods

August 2013

This Dissertation by: Intesar N. Abdoslam

Entitled: *Adjusted Variance Components for Unbalanced Clustered Binary Data Models*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in the College of Education and Behavioral Sciences in the School of Educational Research, Leadership, and Technology.

Accepted by the Doctoral Committee

Jay Schaffer, Ph.D., Co-Research Advisor

Trent Lalonde, Ph.D., Co-Research Advisor

Robert Pearson, Ph.D., Committee Member

Robert Heiny, Ph.D., Faculty Representative

Date of Dissertation Defense _____

Accepted by the Graduate School

Linda L. Black, Ed.D., LPC
Acting Dean of the Graduate School and International Admissions

ABSTRACT

Abdoslam, Intesar N. *Adjusted Variance Components for Unbalanced Clustered Binary Data Models*. Published Doctor of Philosophy dissertation, University of Northern Colorado, 2013.

In practice, it is very common to have clustered binary responses, where binary data are naturally grouped by sampling technique or some property of the sampling units. Often these clusters are unbalanced. The preferred class of models for clustered binary data is the Hierarchical Generalized Linear Model (HGLM), where random effects are used to account for the overdispersion known to exist for clustered binary data. There are many methods to estimate the parameters in Hierarchical Generalized Linear Models, but none of the current methods allowed the overdispersion to vary from cluster to cluster. As clustered binary data led to overdispersion, it was reasonable to conclude that unbalanced clustered binary data may have been different overdispersion for different cluster sizes. By ignoring possible changes in overdispersion across clusters, test statistics tended to show inflated Type I error rates. In this research, two HGLM methods were adjusted to account for different overdispersion across different cluster sizes. The first new method was the Extended Restricted Pseudo Likelihood (EREPL), an adjustment of Restricted Pseudo Likelihood. Extended Restricted Pseudo Likelihood allowed for different dispersion adjustments for each cluster. The new second method was Adjusted Scale Binomial Beta (ASBB), an extension of the classical Binomial Beta model.

This method allowed the Beta distributed random effect to have different scale parameters for each cluster. Through simulation, these extensions were compared to the original methods in terms of power, Type I error rate, and estimator standard errors. Adjusted Scale Binomial Beta h-likelihood was comparable to existing methods, as it gave us a low standard error and acceptable Type I error. Moreover, Binomial Beta h-likelihood had inflated Type I error. The Restricted Pseudo Likelihood could also be applied to unbalanced clustered binary data.

ACKNOWLEDGMENTS

I feel very fortunate to have had the opportunity to work with my advisor, Dr. Trent Lalonde. I would like to express my sincere appreciation and admiration for his guidance, being a good support during the course of this dissertation, and for his extensive expertise which he shared with me in spite of his numerous responsibilities and other commitments. I also want to thank my committee members, Jay Schaffer, Robert Person, and Robert Heiny for their review of these chapters and for their insights and comments. Special thanks to Dr. Jay Schaffer for being kind and gentle with me. I want to thank and offer my appreciation to Keyleigh Gurney for all her help and support. I am sincerely grateful to Dr. Khalil Shaife for helping me figure out LaTeX typing and R ®programming.

My most sincere appreciation goes to my husband, Ali, for enduring a years during my study, and for his patience, encouragement, and gentleness. Without him, I would not have been able to achieve my goal to get a PhD. Many thanks to my parents for their love, encouragement, support, and generosity. Many thanks to my siblings for their love and support. Last, but not least, my beloved kids Rahaf, Raghd, and Yusuf - with them, I was able to be strong. Thanks to all my friends who wished me good luck and encouraged me to do it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SYMBOLS

$\mathbf{Y}$	Vector or response observations
$\mathbf{X}$	Fixed effect design matrix
$\boldsymbol{\beta}$	Vector of response fixed effect parameters
$\mathbf{Z}$	Random effect design matrix
$\boldsymbol{u}$	Vector of random effect parameters.
$\boldsymbol{\mu}$	Response mean vector
$\boldsymbol{V_R}$	Variance covariance matrix for random component
D	Any distribution from exponential family such as normal, poisson, binomial,.. etc.
θ	The canonical parameter
ϕ	The dispersion parameter
$\boldsymbol{V_P}$	Variance covariance matrix for Pseudo response
$\boldsymbol{P}$	Pseudo response observation vector
Q	Quasi likelihood function
Q^+	Extended quasi likelihood function
Q^{++}	Double Extended Quasi Likelihood function
h	h-Likelihood function
h_A	Adjusted profile h-likelihood function
ϕ_i	Adjusted dispersion term in Extended Pseudo Likelihood method

$\boldsymbol{\gamma}$	The mean parameter in Beta distribution
$\boldsymbol{\lambda}$	The scale parameter in Beta distribution
λ_i	Adjusted scale parameter in Beta distribution

ABBREVIATIONS

LM	Linear Model
ML	Maximum Likelihood
REML	Restricted Maximum Likelihood
QL	Quasi Likelihood
EQL	Extended Quasi Likelihood
PQL	Penalized Quasi Likelihood
HL	Hierarchical Likelihood
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
HGLM	Hierarchical Generalized Linear Model
EREPL	Extended Restricted Pseudo Likelihood
ASBB	Adjusted Scale Binomial Beta
DEQL	Double Extended Quasi Likelihood
LMM	Linear Mixed Model
IWLS	Iteratively Weighted Least Squares
MHLE	Maximum h-likelihood Estimates
H	Hessian matrix
GEE	Generalized Estimating Equations

CHAPTER I

INTRODUCTION

Models for clustered binary data are important in many areas such as medical research, education, and finance. Binary data, where the outcome has taken only two possible values, is often represented as success and failure; more generally, binary data represents the presence or absence of an attribute of interest. For example, in health services research where patients are clustered within hospitals, the binary outcome could be whether the patient dies or lives. Also, in educational studies where students are nested within schools, the binary outcome could be whether the student passes or fails.

Clustered data or nested design is an experimental design in which the variables have an implicit hierarchy. For example, a hospital has two wings (I and II). Patients in wing I are randomly assigned to either consultant A or consultant B. Patients in wing II are randomly assigned to either consultant C or consultant D. Thus, consultants A and B are nested within wing I and consultants C and D are nested within wing II. The clusters may be balanced or unbalanced, i.e., the number of observations in a cluster (the size of the cluster) is equal or unequal. Unbalanced clusters may result from sub-sampling unequal numbers of observations from each cluster. Unbalanced clusters may also occur when there are randomly

missing vector elements for a clustered multivariate outcome or if subjects differ in the number of relevant vector elements for the analysis. The different cluster size could lead to different dispersions for each cluster. For a nested model with a binary response, there are two sources of variation. The first source of variation is the between-cluster variation that represents the variation from cluster to cluster. The second source of variation is the within-cluster variation that represents the random variation among responses in each cluster. For binary data that are clustered with variation in each stage, instead of using a linear model, which assumes normality of the dependent variable, it is more appropriate to use the extension of the linear model the generalized linear model. The generalized linear model (GLM) is an extension of the general linear model, which includes response variables that follow any probability distribution in the exponential family of distributions. The exponential family includes useful distributions such as normal, binomial, poisson, multinomial, gamma, negative binomial, and others. Hypothesis tests applied to the GLM do not require normality of the response variable, nor do they require homogeneity of variances. Hence, GLMs could be used when response variables follow distributions other than the normal distribution and when variances are not constant.

The nested design with a binary outcome is popular in many research areas, especially in medical studies. The nested design with unequal cluster size could lead to more variation between the clusters. To account for the extra variation due to different cluster sizes, the hierarchical generalized linear model (HGLM) method is used. The most common methods, such as quasi-likelihood, penalized

quasi-likelihood, and extended quasi-likelihood, allow for overdispersion; however, these methods deal with overdispersion as constant for all clusters. It is common to not apply these methods for changing overdispersion. Unbalanced clustered binary data may have different dispersions for different clusters. It was reasonable to think that unbalanced clustered binary data may have had different dispersion for different clusters, but current methods ignored this possibility. By neglecting to account for different dispersion in binary data with unbalanced clusters, Type I error rate may have been inflated, efficiency may have been lost and power may have been low. To solve this problem, two modified methods were explained. The purpose of this study was to evaluate whether the two presented methods, Extended Restricted Pseudo Likelihood (EREPL) and Adjusted Scale Binomial Beta (ASBB), accounted for overdispersion in unbalanced clustered binary data better then existing methods. These two new methods were compared to REPL and Binomial Beta h-likelihood in terms of power, Type I error rate, and standard error through computer simulation. These new methods were expected to have smaller Type I errors and more power in the case of unbalanced binary clustered data. The goal of this dissertation is to present two methods of estimation for hierarchical generalized linear models (HGLM) for unequal cluster size with binary response to account for overdispersion: (a) The first adjusted method was the Extended Restricted Pseudo Likelihood (EREPL) which allowed for different dispersion adjustments for each cluster. The EREPL used different dispersions denoted by $\boldsymbol{\phi_i}$ in estimating a mixed

effect model for binary outcomes with unequal cluster size. The HGLM formula for ERPL is

$$\boldsymbol{Y_i}|\boldsymbol{u} \sim \boldsymbol{D}(\boldsymbol{\mu}, \boldsymbol{\phi_i}\boldsymbol{\mu}(\mathbf{1}-\boldsymbol{\mu})),$$

$$\boldsymbol{u} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{V_R}),$$

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},$$

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}),$$

where Y is the dependent variable following a binomial distribution with parameters n, and p, D is the binomial distribution from exponential family distribution, $i = 1, 2, ..., K$ denotes the cluster. The parameter $\boldsymbol{u}$ is the random effect following the normal distribution with mean equal to zero and variance covariance matrix V_R. X, and Z are explanatory matrices for fixed and random effects respectively, and g is the link function which is logit for binomial distribution, and μ is the mean.

(b) The second adjusted method was an Adjusted Scale Binomial Beta h-likelihood that allowed for a different scale parameter for the Beta distribution for each cluster to account for overdispersion. The HGLM formula for an Adjusted Scale Binomial Beta h-likelihood is

$$Y_{ij}|u_i \sim Bin(n, p_{ij}),$$

$$u_i \sim \quad Beta\ (\gamma, \lambda_i),$$

$$\eta_{ij} = \ \boldsymbol{x_{ij}}\boldsymbol{\beta} + v(u_i),$$

$$\eta_{ij} = \ logit\ (p_{ij}),$$

where Y is dependent variable follow binomial distribution with parameters n, and variance covariance $\phi\mu(1-\mu)$. The parameter u_i is the random effect following the beta distribution with mean equal to γ, and λ_i is the varying scale from cluster to cluster. The systematic component is η_{ij}, and v is the transformation of u_i to occur linearly with $x_{ij}\beta$. β is the fixed parameter, x_{ij} is explanatory variable for fixed effects j^{th} observation in i^{th} cluster, and g is the link function which is logit for binomial distribution.

The intention of each method was to allow dispersion to differ in clusters of different sizes. In Chapter II, the following methods of parameter estimation for mixed logistic models are reviewed: the methods for the linear model (LM), which are maximum likelihood (ML) for fixed linear models and restricted maximum likelihood (REML) for mixed linear models; and the methods for the generalized linear model (GLM), which are maximum likelihood (ML), quasi-likelihood (QL), and extended quasi-likelihood (EQL). Moreover, a random effect for the GLM is incorporated and then extended to the hierarchical generalized linear model (HGLM). For hierarchical generalized linear models (HGLM), the restricted pseudo likelihood (REPL) method, penalized quasi-likelihood (PQL) method, hierarchical likelihood (HL) method, and double extended quasi-likelihood (DEQL) methods were reviewed. In Chapter III, two modified methods for estimating model parameters are presented and developed, allowing the dispersion to vary to account for unequal cluster sizes in a nested design with binary outcomes. In Chapter IV, computer simulations are presented to investigate the methodology, and comparisons

the adjusted methods with methods are made. Chapter V contains the summary, discussion, and directions for future research.

The Research Questions to be Studied

Q1 Does Extended Restricted Pseudo Likelihood account for different dispersion for different clusters size?

Q2 Does Adjusted Scale Binomial Beta h-likelihood account for different dispersions for different cluster size?

Q3 Is Extended Restricted Pseudo Likelihood more powerful than Restricted Pseudo Likelihood in the case of unbalanced binary clustered data?

Q4 Is Adjusted Scale Binomial Beta h-likelihood more powerful than Binomial Beta h-likelihood in the case of unbalanced binary clustered data?

Q5 Does Extended Restricted Pseudo Likelihood method improve efficiency?

Q6 Does Adjusted Scale Binomial Beta h-likelihood method improve efficiency?

The Limitations of This Study

1. All methods are likelihood based estimation methods.

2. The dependent variable is binary.

3. The number of cluster, and sample sizes are not small.

CHAPTER II

REVIEW OF LITERATURE

Estimation of Mixed Logistic Model Parameters

In many applications, data have hierarchical or clustered structures, e.g, medical and health services research where patients are clustered within hospitals, or educational studies where students are nested within schools. These studies often involve the analysis of data with complex patterns of variability. Mixed models are often the most appropriate models to use in practice, as they contain fixed effects of interest and random effects to account for the clustering. The random effects reflect multiple error structures. As for data that are clustered, we have variation in each cluster as well as variation between clusters.

For mixed models which contain both fixed and random effects, we have the equation

$$\boldsymbol{E}[\boldsymbol{Y}|\boldsymbol{u}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},$$

where $\mathbf{X}$ is the fixed effect design matrix, $\mathbf{Z}$ is the random effect design matrix, $\boldsymbol{\beta}$ is the vector of fixed effect parameters, and $\boldsymbol{u}$ is the vector of random effect parameters. We need to estimate the parameters $\boldsymbol{\beta} = [\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_p}]^{\boldsymbol{T}}$, and predict the random effects $\boldsymbol{u} = [\boldsymbol{u_1}, \boldsymbol{u_2}, ...\boldsymbol{u_q}]^{\boldsymbol{T}}$.

Instead of using linear models assuming normality of the dependent variable, we used the extension of linear models to the more appropriate generalized linear models when the dependent variable in mixed model is a binary variable. The mixed models equation was in the form

$$\boldsymbol{g(E[Y|u]) = X\beta + Zu.}$$

For mixed effects models, a variance component procedure, estimates the contribution of each random effect to the variance of the dependent variable. This procedure is particularly interesting for analysis of mixed models. The overriding problem with estimating variance components from clustered data is that many methods of estimation are available and choosing a method is dependent on one's model and what components the model includes. Here we briefly summarize some methods that were used for estimating the parameters in two models: the Linear Model and the Generalized Linear Model. Then methods for clustered data are presented and current methods for unbalanced cluster data are examined.

The Linear Model

The Linear Model (LM) is either a statistical model with fixed effects only, called a fixed model, or with random effects only, called a random model.

The Fixed Effects Linear Model

The linear model (LM) is a statistical model with fixed effects. In matrix notation, a fixed Model could have been represented as

$$\boldsymbol{Y = X\beta + \epsilon,}$$

where $\boldsymbol{Y}$ is a response variable (vector of observations), $\boldsymbol{\beta}$ is a parameter vector of fixed effects $\boldsymbol{\beta} = [\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_p}]^{\boldsymbol{T}}$, and $\boldsymbol{\epsilon}$ is a vector of IID random error terms with mean $\boldsymbol{E(\epsilon) = 0}$ and variance $\boldsymbol{var(\epsilon) = \sigma^2\, I}$. Then $\boldsymbol{Y}$ follow

$$\boldsymbol{Y \sim \mathcal{N}(X\beta, \sigma^2 I).}$$

For the linear model, there are a variety of methods to estimate the parameter. Here, we explain the maximum likelihood estimation (ML) method to estimate the parameters in the fixed Linear Model.

Maximum likelihood estimation: Estimation by maximum likelihood (ML) is a well-established method of estimation, originating with Fisher (1925). Hartley and Rao (1967) first applied it to the general linear mixed model. Assuming that the error terms are normally distributed, the maximum likelihood (ML) method could have been used to estimate both the variance components and the fixed parameters. The pdf function of the fixed model is

$$\boldsymbol{f(Y) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\, exp\, \left[\frac{-1}{2}\,(Y - X\beta)^T\, (\sigma^2)^{-1}\,(Y - X\beta)\right].}$$

Then the method of maximum likelihood could have been applied to the complete likelihood function, denoted by

$$\boldsymbol{L\left(\beta, \sigma^2 | Y\right) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\, exp\, \left[\frac{-1}{2}\,(Y - X\beta)^T\, (\sigma^2)^{-1}\,(Y - X\beta)\right],} \quad (1)$$

and so the ln likelihood is

$$\boldsymbol{l = \ln L = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2}\,(Y - X\beta)^T\,(\sigma^2)^{-1}\,(Y - X\beta).} \quad (2)$$

Maximizing l with respect to elements of $\boldsymbol{\beta}$ and the variance $\boldsymbol{\sigma^2}$ leads to equations that have to be solved to yield the ML estimators of $\boldsymbol{\beta}$ and for the variance $\boldsymbol{\sigma^2}$. The solution for estimating the fixed parameters $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T}\ \boldsymbol{X})^{-1}\boldsymbol{X^T}\ \boldsymbol{Y}, \tag{3}$$

and for the variance parameter $\boldsymbol{\sigma^2}$ is

$$\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\boldsymbol{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\boldsymbol{N}}. \tag{4}$$

The Linear Mixed Effects Model

The linear mixed model (LMM) is a statistical model combining fixed effects and random effects. In matrix notation, a linear mixed model could have been represented as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon},$$

$$\boldsymbol{\epsilon} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{0}, \boldsymbol{R}), \quad \boldsymbol{u} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{0}, \boldsymbol{G}).$$

where $\boldsymbol{Y}$ is a response variable (vector of observations), $\boldsymbol{\beta}$ is a parameter vector of fixed effects $\boldsymbol{\beta} = [\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_p}]^{\boldsymbol{T}}$, and $\boldsymbol{u}$ is a vector of independent and identically distributed (IID) predicted random effects $\boldsymbol{u} = [\boldsymbol{u_1}, \boldsymbol{u_2}, ...\boldsymbol{u_q}]^{\boldsymbol{T}}$ with mean $\boldsymbol{E(u)} = \boldsymbol{0}$ and variance-covariance matrix $\mathbf{var}(\boldsymbol{u}) = \boldsymbol{G}$, and $\boldsymbol{\epsilon}$ is a vector of IID random error terms with mean $\boldsymbol{E(\epsilon)} = \boldsymbol{0}$ and variance $\mathbf{var}(\boldsymbol{\epsilon}) = \boldsymbol{R}$. Then $\boldsymbol{Y}$ followed the normal distribution, with mean $\boldsymbol{E(Y)} = \boldsymbol{X\beta}$, and variance covariance

$$\begin{aligned}\boldsymbol{cov(Y)} &= \boldsymbol{cov}\ (\boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{ZGZ^T} + \boldsymbol{R} \\ &= \boldsymbol{V}.\end{aligned}$$

For the linear mixed model we used maximum likelihood estimation (ML) in the same way as in the last section or restricted maximum likelihood estimators (REML) to estimate the parameters in linear mixed model.

Maximum Likelihood Estimation

To estimate both the variance components and the fixed parameters in Mixed Effects Model, the pdf function of the mixed model is

$$\boldsymbol{f(Y) = \frac{1}{(2\pi|V|)^{\frac{N}{2}}}\ exp\ \left[\frac{-1}{2}\,(Y - X\beta)^T\ V^{-1}\,(Y - X\beta)\ \right],}$$

The method of maximum likelihood could have been applied to the complete likelihood function, denoted by

$$\boldsymbol{L = (2\pi)^{-\frac{N}{2}}\ |V|^{-\frac{N}{2}}\ exp\left[\frac{-1}{2}(Y - X\beta)^T\ V^{-1}\ (Y - X\beta)\right],} \tag{5}$$

so the ln likelihood is

$$\boldsymbol{l = \ln L = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln|V| - \frac{1}{2}\,(Y - X\beta)^T\,V^{-1}\,(Y - X\beta),} \tag{6}$$

Maximizing l with respect to elements of $\boldsymbol{\beta}$ and the variance components

($\boldsymbol{\tau = (\sigma_1^2, \sigma_2^2, ..., \sigma_l^2)^T}$'s that occur in V) leads to equations that have to be solved to yield the ML estimators of $\boldsymbol{\beta}$ and of $\boldsymbol{\tau}$. The equation is

$$\boldsymbol{X^T\ \hat{V}^{-1}\ X\ \hat{\beta} = X^T\ \hat{V}^{-1}\ Y,} \tag{7}$$

and for the random parameters components $\boldsymbol{V}$ is

$$\boldsymbol{tr\ (\hat{V}^{-1}\ Z_i\ Z_i^T) = (Y - X\hat{\beta})^T\ \hat{V}^{-1}\ Z_i\ Z_i^T\ \hat{V}^{-1}(Y - X\ \hat{\beta}).} \tag{8}$$

For $i = 1, 2, ..., r$; equations (7) and (8) have to be solved for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\tau}}$, the elements of $\hat{\boldsymbol{\tau}}$ being implicit in $\hat{V}$. So they have to be solved numerically, by iteration. For convenience, write

$$\boldsymbol{P = V^{-1} - V^{-1}\ X(X^T\ V^{-1}\ X)^-\ X^T\ V^{-1},}$$

and with

$$\boldsymbol{I = V^{-1}\ V}$$

and $\boldsymbol{V = Z_i\ Z_i^T \tau}$, McCullagh and Searle (2001) used the trace operation inside on the left-hand side of (8), so set of r equations could have been written as

$$\boldsymbol{tr\ (\hat{V}^{-1}\ Z_i\ Z_i^T\ \hat{V}^{-1}\ Z_i\ Z_i^T)\hat{\tau} = Y^T \hat{P} Z_i\ Z_i^T\ \hat{P}\ Y.} \tag{9}$$

for $i = 1, 2, 3, ..., r$, r^{th} equation. These provide easier visualization of an iterative process than do (7) and (8); in (9) we could use a starting value for $\hat{\boldsymbol{\tau}}$ in $\hat{V}$ and $\hat{P}$ to solve (9) and repeat the process. There are several problems associated with solving either (7) and (8) or (9). Briefly, the choice of a starting value for $\hat{\boldsymbol{\tau}}$ affects the final result. In fact, the final result obtained for $\hat{\boldsymbol{\tau}}$ is given a global maximum of l or only a local maximum.

The maximum likelihood method of estimation is well-defined and the resulting estimators have attractive, well-known large-sample properties: they are normally distributed and their sampling variances are known, e.g, Searle (1987).

Restricted Maximum Likelihood Estimation

In general, the ML for the variance components do not take into account the loss in degrees of freedom resulting from the estimation of the fixed effects, and hence they become biased (McCullagh & Searle, 2001). A variant of maximum likelihood estimation in the mixed model is restricted (residual) maximum likelihood (REML). Restricted maximum likelihood estimators are obtained from maximizing the part of the likelihood which is invariant to the location parameter, in terms of the mixed model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}$, invariant to $\boldsymbol{X\beta}$. To estimate only the variance components, we allowed the fixed part equal to zero. Suppose $\boldsymbol{K^T Y}$ for vector $\boldsymbol{K}$, so that $\boldsymbol{K^T Y}$ which contains none of the fixed effects in $\boldsymbol{\beta}$. This means having $\boldsymbol{k^T}$ such that $\boldsymbol{k^T X} = \boldsymbol{0}$. For optimality using the maximum number, $N - rx$, of linearly independent vectors $\boldsymbol{k^T}$ and write $\boldsymbol{K} = [\boldsymbol{k_1\ k_2\ .\ .\ .k_{v-rx}}]$. This results in doing maximum likelihood on $\boldsymbol{K^T Y}$ instead of Y, where $\boldsymbol{k^T X} = \boldsymbol{0}$ and $\boldsymbol{K^T}$ has full row rank $N - rx$. Then the vector

$$\boldsymbol{K^T Y} \sim \boldsymbol{\mathcal{N}(0, K^T\ V\ K).}$$

ML equations for $K^T\ Y$ was therefore, derived from those for

$$\boldsymbol{Y} \sim \boldsymbol{\mathcal{N}(X\beta, V),}$$

by replacing

$\boldsymbol{Y}$ with $\boldsymbol{K^T Y}$; $\boldsymbol{X}$ with $\boldsymbol{K^T X}$

$\boldsymbol{Z}$ with $\boldsymbol{K^T\ Z}$; and $\boldsymbol{V}$ with $\boldsymbol{K^T V K.}$

On using

$$\boldsymbol{P = K(K^T\ V\ K)^{-1}K^T,}$$

the ML equations for $\boldsymbol{K^T\ Y}$ reduce to

$$\boldsymbol{tr\ (\hat{P}\ Z_i\ Z_i^T\ \hat{P}\ Z_i\ Z_i^T)\hat{\tau} = Y^T\hat{P}Z_i\ Z_i^T\ \hat{P}\ Y.} \tag{10}$$

These are the REML equations, to be solved for $\hat{\boldsymbol{\tau}}$ which occurs in $\hat{\boldsymbol{P}}$. It is easily seen that they are the same as the ML equations (9) except for $\hat{\boldsymbol{V}}$ on the left-hand side being replaced by $\hat{\boldsymbol{P}}$ in (10). The basic idea behind both REML and ML estimation is to find the set of weights for the random effects in the model (McCullagh & Searle, 2001). The relative advantage of ML is that it provides estimation of fixed effects, while REML does not. The REML takes account of the degrees of freedom involved in estimating the fixed effects, whereas ML estimators do not (Searle, 1987).

The Generalized Linear Model

The generalized linear model (GLM) is an extension of the linear model to include response variables that follow any probability distribution in the exponential family of distributions. The exponential family includes useful distributions, e.g, the normal, binomial, poisson, multinomial, gamma, negative binomial, and others. Hypothesis tests applied to the Generalized Linear Model do not require normality of the response variable, nor do they require homogeneity of variances. Hence, generalized linear models could have been used when response variables follow distributions other than the normal distribution, and when variances are not

constant. For example, binary data would be appropriately analyzed as a binomial random variable within the context of the generalized linear model. The GLM was specified in three pieces (GLM structure):

1. Response Distribution

$$\boldsymbol{Y} \sim \boldsymbol{D}(\boldsymbol{\mu}, \boldsymbol{a}(\boldsymbol{\phi})\boldsymbol{V}(\boldsymbol{\mu})).$$

The vector $\mathbf{y}$ is assumed to consist of independent measurements from a distribution with density from the exponential family :

$$f_{Y_i}(y_i) = e^{\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right)},$$

where, for convenience, we have written the distribution in what is called canonical form. For example, for binary response data, the data would be independent Bernoulli so that

$$f_{Y_i}(y_i) = \binom{n}{y_i} p_i^{y_i}(1-p_i)^{1-y_i},$$

where p_i is the probability of a success and $\theta_i = \ln[p_i/(1-p_i)]$, (McCullagh & Searle, 2001).

2. Linear Systematic Component $\boldsymbol{\eta} = \boldsymbol{X\beta}$

The linear component is the quantity which incorporates the information about the independent variables into the model. The symbol $\boldsymbol{\eta}$ is typically used to denote a linear predictor, and is expressed as linear combinations (thus, "linear") of unknown parameters $\boldsymbol{\beta}$. The coefficients of the linear combination are represented as the matrix of independent variables $\mathbf{X}$.

3. Link Function $\boldsymbol{\eta = g(\mu)}$

To relate the parameters of the distribution to various predictors, we model a transformation of the mean, μ_i, which would be some function of θ_i, as a linear model in the predictors:

$$g(\mu_i) = x_i^T \beta,$$

where $g(.)$ is a known function, called the link function (since it links together the mean of y_i and the linear form of predictors), x_i^T is the i^{th} row of the model matrix, and β is the parameter vector in the linear predictor. Some examples of $g(.)$ are given in Table 1.

Table 1

Canonical Link Functions

Distribution	Link Name	$g(.)$
Binomial	Logit	$ln(p/1-p)$
Poisson	Log	$ln(\mu)$
Normal	Identical	μ
Gamma	Inverse	μ^{-1}

This GLM structure is appropriate for any response distribution from the *Exponential Family*. The pdf for the exponential family is

$$\boldsymbol{f(y;\theta,\phi) = e^{\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right)}},$$

θ is canonical parameter, and ϕ is the dispersion parameter. Where

$$\boldsymbol{E(y) = b'(\theta),}$$

where $b'(\boldsymbol{\theta})$ is the first derivative of $b(\theta)$ and

$$\boldsymbol{Var(y) = a(\phi)b''(\theta),}$$

where $b''(\boldsymbol{\theta})$ is the second derivative of $b(\theta)$.

For example; the exponential family for Binomial distribution: the binomial distribution function,

$$f(y; n, p) \quad = \quad \binom{n}{y} p^y (1 \ - \ p)^{(n-y)}.$$

The Binomial distribution in the form of the exponential family of distributions is

$$\begin{aligned} f(y;p) &= \binom{n}{y} p^y (1-p)^{n-y} \\ &= e^{\left[\ln\left(\binom{n}{y} p^y (1-p)^{n-y}\right)\right]} \\ &= e^{\left[\ln\binom{n}{y} + y\ln p + (n-y)\ln(1-p)\right]} \\ &= e^{\left[\ln\binom{n}{y} + y\ln p + n\ln(1-p) - y\ln(1-p)\right]} \\ &= e^{\left[\ln\binom{n}{y} + y\ln\left(\frac{p}{1-p}\right) - n\ln\left(\frac{1-p+p}{1-p}\right)\right]} \\ &= e^{\left[\ln\binom{n}{y} + y\ln\left(\frac{p}{1-p}\right) - n\ln\left(1+\frac{p}{1-p}\right)\right]} \\ &= e^{\left[\ln\binom{n}{y} + y\ln\left(\frac{p}{1-p}\right) - n\ln\left(1+\exp\left(\ln\left(\frac{p}{1-p}\right)\right)\right)\right]}. \end{aligned}$$

For $\theta = \ln\left(\frac{p}{1-p}\right)$, $a(\phi) = 1$, $b = \ln\left(1 + e^{\theta}\right)^n$, and $c(y, \phi) = \ln\binom{n}{y}$

There are several methods for estimating the parameters of a generalized linear model, e.g, maximum likelihood, quasi-likelihood, and extended quasi-likelihood, which are summarized here.

Maximum Likelihood Estimation

The likelihood function of the exponential family is

$$l_i(\theta_i; y_i; \phi) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi).$$

The maximum likelihood method is used to estimate the mean model parameters. When $l_i(\theta_i; y_i; \phi)$ is differentiable, the goal is to maximize l_i with respect to the parameter β_j, producing the likelihood estimating equation:

$$\frac{\partial l_i}{\partial \beta_j} = 0.$$

By applying the chain rule to get the estimation of mean model:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\,\frac{\partial \theta_i}{\partial \mu_i}\,\frac{\partial \mu_i}{\partial \tau_i}\,\frac{\partial \tau_i}{\partial \beta_i},$$

where the conical parameter is

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b(\theta_i)}{a(\phi)},$$

because $\mu_i = E(y_i) = b'(\theta_i)$,

$$\frac{\partial}{\partial \mu_i}\left(\mu_i = b'(\theta_i)\right),$$

then by differentiating both sides with respect to the mean, we get

$$1 = b''(\theta_i)\frac{\partial \theta_i}{\partial \mu_i}.$$

Solving the equation we get

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)},$$

and $var(Y) = a(\phi)b''(\theta)$, where $a(\phi) = 1$, (Nelder & Lee, 1992). Thus, $var(Y) = b''(\theta)$,

and we could write

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{var(y_i)} = \frac{1}{var(\mu_i)}.$$

$$\frac{\partial \mu_i}{\partial \tau_i} \frac{\partial \tau_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \beta_j}.$$

The maximum likelihood estimating equations for N independent responses are

$$\sum_{i=1}^{N} (y_i - \mu_i) \frac{1}{var(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0,$$

for each j=1,2,...,p. The equations above depend on first and second moments.

In matrix notation,

$$\boldsymbol{D^T\ V^{-1}\ (Y - \mu) = 0,}$$

where $D_{ij} = \frac{\partial \mu_i}{\partial \beta_j}$, and $\boldsymbol{V^{-1}}$ is the covariance structure of the response. Maximum likelihood estimations are asymptotically normal (Nelder & Lee, 1992). The maximum likelihood estimates could have been found using an iteratively weighted least squares (IWLS) using either a Newton Raphson method or a Fisher's scoring method, (Gu, 2008).

The maximum likelihood estimation requires a fully specified response distribution. When we could not specify the full response distribution but could determine the mean variance relationship from the data, we could apply quasi-likelihood. If we recognize the relationship between the mean and the variance, then the quasi-likelihood estimation is appropriate.

Quasi Likelihood Estimation

The quasi-likelihood (QL) method specifies the first two moments only, without completely specifying the distribution of the data. The main purpose of many analyses is to show how the mean response is affected by several covariates. Sometimes there is insufficient information about the data for us to specify a full distribution for the data. However, we may be able to specify some of the features of the data.

From McCullagh and Nelder (1989), we summarized the method of quasi-likelihood (Q-L): suppose we have a vector of responses $\mathbf{Y} = [y_1, y_2, ..., y_n]^T$ which are independent with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{a(\phi)V(\mu)}$, where $\boldsymbol{a(\phi)}$ may have been unknown and $\boldsymbol{V(\mu)}$ is a matrix of known functions. We assume that $\boldsymbol{\mu}$ is a function of covariates $\boldsymbol{X}$, and some parameters $\boldsymbol{\beta}$. We did not need to limit the nature of this relationship. Quasi-likelihood assumes variance $\boldsymbol{a(\phi)}$ is given, and $\boldsymbol{V(\mu)}$is made up of known functions. As it is assumed that the components of $\mathbf{Y}$ are independent, $\boldsymbol{V(\mu)}$ has to be diagonal. Thus, they write

$$\boldsymbol{V(\mu) = diag\,(\,V_1(\mu),\, V_2(\mu), ...,\, V_n(\mu))\,.}$$

It is also necessary to assume that $\boldsymbol{V_i(\mu)}$ only depends on the i^{th} component of $\boldsymbol{\mu}$. This seems to be a reasonable assumption, as it is difficult to see why the variance of an observation would depend on other mean components, even if the mean does not. In most applications, the functions $V_1(.), V_2(.), ..., V_n(.)$, may be the same, although their arguments could have been different. To construct the quasi-

likelihood, we start by looking at a single component y_i of $\boldsymbol{Y}$. Now suppose we have independent responses $y_1, y_2, ..., y_n$ with means $E(y_i) = \mu_i$ and variance $\boldsymbol{var(y_i) = \phi V(\mu_i)}$.

Wedderburn (1974) defined the quasi-likelihood as a function $Q_i(\mu_i; y_i)$ satisfying

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\boldsymbol{\phi V(t)}} dt,$$

and for the complete data is the sum of the individual contributions, the total quasi-likelihood is

$$Q(\mu; Y) = \sum_{i=1}^{n} Q_i(\mu_i; y_i),$$

To estimate the mean model parameters $\hat{\boldsymbol{\beta}}$, maximizing the $\mathbf{Q}$ with respect to the parameter $\boldsymbol{\beta}$ and equal to zero

$$\frac{\partial Q_i}{\partial \beta_j} = 0,$$

Similar to maximum likelihood to get estimation of mean model, the equations for N independent responses are

$$\sum_{i=1}^{N} (y_i - \mu_i) \ \frac{1}{\phi \ V(\mu_i)} \ \frac{\partial \eta_i}{\partial \beta_j} = 0,$$

for each j=1,2,...,p. The equations above depend on first and second moments. The matrix notation

$$\boldsymbol{D^T \ V^{-1} \ (Y - \mu) = 0.}$$

Wedderburn (1974) derived some properties of QL, but his theory assumes ϕ is known; in the following it is set to unity. With this assumption, the QL is a true likelihood if and only if the response y_i comes from a one parameter exponential family model (GLM with $\phi = 1$). The quasi-likelihood does not specify a distribution, only the mean-to-variance relationship. This is not a sufficient basis on which to estimate the variance covariance structure.

Extended Quasi Likelihood Estimation

The quasi-likelihood method (QL) estimates the mean parameter, and it does not estimate the dispersion part. The quasi-likelihood method assumes $\boldsymbol{\phi}$ is known. An extended quasi-likelihood method, Pregibon (1987) estimated the mean and dispersion parameters for fixed effects in the generalized linear model. The extended quasi-likelihood method supposed the relationship between $\boldsymbol{\mu_i}$ and x_i is $\boldsymbol{g(\mu_i) = x_i\beta}$, and defines the function Q^+ for a single observation y with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\phi V(\mu)}$ by

$$Q^+(\boldsymbol{y;\mu}) = -\frac{\mathbf{1}}{\mathbf{2}}\ \ln\left\{\mathbf{2\pi}\ \boldsymbol{\phi V(y)} - \frac{\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{D(y;\mu)}}{\boldsymbol{\phi}}\right\},$$

where Q^+, like quasi-likelihood method, did not presuppose a full distributional assumption, but just the first and second moments. This estimates the $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ by maximizing Q^+ for the mean and for the dispersion parameters respectively. This method estimated the parameters for the fixed effects model only; it did not deal with random effects.

To incorporate random effects, a mixed generalized linear model was used. The model included the random component and the fixed effect as well. The ex-

tension of the generalized linear model (GLM) to include random effects was the generalized linear mixed model (GLMM), also named the hierarchical generalized linear model (HGLM).

The Hierarchical Generalized Linear Model

In generalized linear models (GLM), when the model contains both fixed effects and random effects, it is named the generalized linear mixed models (GLMM) or hierarchical generalized linear models (HGLM),(Lee & Nelder, 1996). Hierarchical generalized linear models allow extra error components in the linear predictors of generalized linear models. The distribution of these components is not required to be normal, allowing a broader class of models. In hierarchical generalized linear models, the response and random effects are allowed to follow any distribution in the exponential family. As such, the HGLM is more appropriate for clustered data than the GLM. Specify a HGLM in three pieces:

1. Response Distribution:

$$\boldsymbol{Y}|\boldsymbol{u} \sim \boldsymbol{D}(\boldsymbol{\mu}, \boldsymbol{a}(\boldsymbol{\phi})\boldsymbol{V}(\boldsymbol{\mu})),$$

$$\boldsymbol{u} \sim \boldsymbol{D_R}(\boldsymbol{\mu_R}, \boldsymbol{V_R}(\boldsymbol{\mu_R})).$$

2. Linear Systematic Component: $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zu}.$

3. Link Function: $\boldsymbol{\eta} = g(\boldsymbol{\mu}).$

where $\mathbf{X}$ is the design matrix for the fixed effect, $\boldsymbol{\beta}$ is the vector of fixed parameter, $\mathbf{Z}$ is the design matrix for the random effect, and $\boldsymbol{u}$ is the vector of the

random parameter. We need to estimate the fixed effect and predict the random parameters $\boldsymbol{\beta} = [\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_p}]^{\boldsymbol{T}}$, and $\boldsymbol{u} = [\boldsymbol{u_1}, \boldsymbol{u_2}, ..., \boldsymbol{u_q}]^{\boldsymbol{T}}$.

There are several methods for estimating the parameters of a hierarchical generalized linear model, e.g, Restricted Pseudo Likelihood estimation, Penalized Quasi-Likelihood, and $h-$likelihood.

For generalized linear models, we used the maximum likelihood (ML) to estimate the mean component. An extension to ML in HGLM is Restricted Pseudo Likelihood estimation (REPL). Geys, Molenberghs, & Ryan (1997) showed ML and REPL have parameter estimates that agree fairly closely.

Restricted Pseudo Likelihood Estimation

In maximum likelihood estimation, we estimated the fixed effects of the mean model. Estimating both the fixed and random effects in HGLM means that we have to consider the dispersion components and correlated errors. To handle this situation, Wolfinger and O'Connell (1993) use Restricted Pseudo Likelihood estimation. The response and random components in the HGLM could have been written

$$\begin{aligned} \textbf{1.} \quad & \boldsymbol{Y}|\boldsymbol{u} \sim \boldsymbol{D}(\boldsymbol{\mu}, \boldsymbol{a}(\boldsymbol{\phi})\boldsymbol{V}(\boldsymbol{\mu})), \\ & \boldsymbol{u} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{0}, \boldsymbol{V_R}), \\ \textbf{2.} \quad & \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \\ \textbf{3.} \quad & \boldsymbol{\eta} = g(\boldsymbol{\mu}), \end{aligned}$$

where $E[\boldsymbol{y}|\boldsymbol{u}] = \boldsymbol{\mu}$, $\boldsymbol{V_R}$ is unknown.

First, write the mean in terms of the link function

$$\boldsymbol{\mu} = \boldsymbol{g}^{-1}\left(\boldsymbol{X\beta} + \boldsymbol{Zu}\right).$$

Apply a Taylor approximation of $\boldsymbol{g(\mu)}$ about the initial estimate $\boldsymbol{\mu_0}$,

$$\boldsymbol{g(\mu)} = \boldsymbol{g(\mu_0)} + \hat{\boldsymbol{D}}(\boldsymbol{\mu} - \boldsymbol{\mu_0}) + \boldsymbol{k}\left((\boldsymbol{\mu} - \boldsymbol{\mu_0})^{\boldsymbol{T}}(\boldsymbol{\mu} - \boldsymbol{\mu_0})\right),$$

where $\boldsymbol{k}\left((\boldsymbol{\mu} - \boldsymbol{\mu_0})^{\boldsymbol{T}}(\boldsymbol{\mu} - \boldsymbol{\mu_0})\right)$ is the quadratic and higher- order terms for the Taylor Polynomial, and

$$\hat{\boldsymbol{D}} = \frac{\boldsymbol{\partial g(\mu)}}{\boldsymbol{\partial \mu}} \mid_{\boldsymbol{\mu=\mu_0}}.$$

Dropping the higher-order terms

$$\boldsymbol{g(\mu)} \approx \boldsymbol{g(\mu_0)} + \hat{\boldsymbol{D}}(\boldsymbol{\mu} - \boldsymbol{\mu_0}).$$

After we get the linearizion, we redefine the Pseudo response

$$\boldsymbol{P} = g(\boldsymbol{\mu_0}) + \hat{\boldsymbol{D}}(\boldsymbol{Y} - \boldsymbol{\mu_0}).$$

For the linearizion $\mathbf{P}$, we have

$$E(\boldsymbol{P}|\boldsymbol{u}) = \boldsymbol{X\beta} + \boldsymbol{Zu},$$

and

$$\boldsymbol{var}\left(\boldsymbol{P}|\boldsymbol{u}\right) = \hat{\boldsymbol{D}}\boldsymbol{cov}(\boldsymbol{Y})\hat{\boldsymbol{D}}^{\boldsymbol{T}}.$$

The redefined model is

$$\boldsymbol{P} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}.$$

Now we have a linear relationship between the pseudo response and the predictors. The pseudo response variable is assumed to follow a normal distribution

$$\boldsymbol{P \sim \mathcal{N}(X\beta, ZV_R Z^T + \hat{D}cov(Y)\hat{D}^T).}$$

Let $\boldsymbol{V_P = ZV_R Z^T + \hat{D}cov(Y)\hat{D}^T}$. Assuming normality, the likelihood for the linear mixed model for the new pseudo response $\mathbf{P}$ is

$$\boldsymbol{f\left(P;\beta\right) = \frac{1}{\sqrt{2\pi}\ \ |\phi V_P|^{\frac{1}{2}}}\ \ e^{\frac{1}{2}(P-X\beta)^T\ V_P^{-1}(P-X\beta)},}$$

and the ln likelihood is

$$\boldsymbol{l\left(\beta;P\right) = \frac{-1}{2}\ \ln\ |\phi V_P| - \frac{1}{2}\ \phi^{-1}\left(P - X\beta\right)^T\ V_P^{-1}\left(P - X\beta\right).}$$

To estimate the parameter $\boldsymbol{\beta}$, we maximize l with respect to the parameter vector $\boldsymbol{\beta}$.

In the above discussion, Wolfinger and O'Connell (1993) assume $\boldsymbol{\phi} = 1$, but to allow $\boldsymbol{\phi} \neq 1$, we make use of the profile ln likelihood to estimate the additional dispersion parameter. To estimate the additional dispersion parameter, using the profile ln likelihood in Wolfinger and O'Connell (1993),

$$\boldsymbol{l(\tau;P) = -\frac{1}{2}\ln|V_P| - \frac{n}{2}\ \ln\left(r^T\ V_P^{-1}\ r\right) - \frac{n}{2}\left[1 + \ln\left(\frac{2\pi}{n}\right)\right],}$$

where $\boldsymbol{r = P - X\left(X^T\ V_P^{-1}\ X\right)^{-1}\ \ X^T\ V_P^{-1}P}$ is the residual pseudo response "$\boldsymbol{r = P - \hat{P}}$", the vector parameters $\boldsymbol{\tau}$ is the parameter that allows the data to have more dispersion, and the restricted marginal ln likelihood is given by

$$\boldsymbol{l_R(\tau;P) = -\frac{1}{2}\ln|V_P|\ -\left(\frac{n-p}{2}\right)\ln\ \left(r^T\ V_P^{-1}\ r\right)\frac{1}{2}\ln|\ X^T\ V_P^{-1}\ X| - \frac{n-p}{2}\left[1 + \ln\left(\frac{2\pi}{n-p}\right)\right].}$$

Numerical methods are generally required to maximize l and l_R over the parameters in $\boldsymbol{\tau}$. The resulting equations could by solved using the Newton Raphson procedure.

The parameter estimates are:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X^T}\hat{\boldsymbol{V_P}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X^T}\hat{\boldsymbol{V_P}}^{-1}\boldsymbol{P},$$

$$\hat{\boldsymbol{u}} = \hat{\boldsymbol{V_R}}\boldsymbol{Z^T}\hat{\boldsymbol{V_P}}^{-1}\,\hat{\boldsymbol{r}},$$

$$\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{r}}^{\boldsymbol{T}}\,\hat{\boldsymbol{V_P}}^{-1}\,\hat{\boldsymbol{r}}\,/\boldsymbol{n^*},$$

where (n^*) equals (n) for PL and $(n-p)$ for $REPL$.

Notice that the method of Wolfinger and O'Connell (1993) applied a linearizion, and that their method assumed the normality of pseudo response to estimate the parameters by using ML. Restricted Pseudo Likelihood Estimation was shown to be a very useful alternative for Maximum likelihood Estimation in clustered data with non-continuous response (Geys et al., 1997).

There is an another method which does not need to apply a linearizion, called the Penalized Quasi-Likelihood. Penalized Quasi-Likelihood (PQL) adds a random part to the quasi-likelihood method. In PQL, we need to determine the first two moments.

Penalized Quasi-Likelihood Estimation

The penalized quasi-likelihood (PQL) approach is the estimation procedure for the HGLM. PQL is used for inference on parameters in the hierarchical models. To remedy biased estimates for variance-covariance dispersion, Green and

Silverman (1994b) suggested adding a penalty function to the quasi-likelihood, referred to as the penalized quasi-likelihood (PQL). To estimate the parameters for a (HGLM) model by using the penalized quasi-likelihood (PQL), add a random part $\mathbf{u}$ to the quasi-likelihood of the form $\frac{1}{2}\boldsymbol{u^T}\ \boldsymbol{V_R^{-1}u}$, assuming that $\mathbf{u}$ has a normal distribution with mean zero and variance covariance matrix $\boldsymbol{V_R}$. The PQL is

$$\boldsymbol{PQL} = \sum Q_i - \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{u^T}\ \boldsymbol{V_R^{-1}u},$$

where

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)}\ dt.$$

Green (1987) presented PQL as

$$-\frac{\mathbf{1}}{\mathbf{2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} Q_i - \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{u^T}\ \boldsymbol{V_R^{-1}u}$$

and differentiation with respect to fixed parameters $\boldsymbol{\beta}$ and predict random parameter $\boldsymbol{u}$ leads to the score equations for the mean parameters

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{\phi V(\mu_i)g`(\mu_i)} = 0, \tag{11}$$

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)z_i}{\phi V(\mu_i)g`(\mu_i)} = V_R^{-1}u. \tag{12}$$

where observations on the i^{th} of n units consist of a univariate response variable y_i together with vectors x_i and z_i of explanatory variables associated with the fixed and random effects. Green (1987) developed the Fisher scoring algorithm for the solution of equations (11) and (12) as an iterated weighted least squares (IWLS).

The estimators for fixed parameters and random predictor parameters, respectively, are

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\ \hat{\boldsymbol{V}}^{-\mathbf{1}}\ \boldsymbol{X}\ )^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\ \hat{\boldsymbol{V}}^{-\mathbf{1}}\ \boldsymbol{Y}, \tag{13}$$

and

$$\hat{\boldsymbol{u}} = \boldsymbol{V_R}\ \boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{V}^{-\mathbf{1}}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right). \tag{14}$$

Breslow and Lin (1995), and Green and Silverman (1994a) mentioned that PQL has not been found to work well in practice, especially for binary data in small clusters. McCullagh and Searle (2001) showed that PQL methods for binary data in small clusters led to estimators which were asymptotically biased and hence inconsistent. Of course, inconsistency by itself may not fave been worrisome if the asymptotic bias was small. Unfortunately, for situations like paired binary data, the PQL estimator could perform quite badly. McCullagh and Searle (2001) recommend that unmodified penalized quasi-likelihood not be used in practice.

The penalized quasi-likelihood (PQL) approach is one of the most common estimation procedure for the HGLM. Jang and Lim (2006) proved that the PQL tended to underestimate the variance components and (in absolute value) fixed effects when applied to clustered binary data. There is another method that may have been used for HGLM with binary outcome, which is hierarchical likelihood estimation (HL). HL may well have been a more appropriate method for HGLM with binary response then PQL.

Hierarchical Likelihood Estimation

The normality assumption used in restricted pseudo likelihood (REPL) and penalized quasi-likelihood (PQL) methods are not appropriate all the time (Gu, 2008). Moreover, REPL and PQL both ignore the estimation of the dispersion parameters, and usually estimate the mean parameters only. To estimate the mean parameters and dispersion parameters, we use hierarchical likelihood estimation (HL). In HL the distribution of random components does not need to be normal; this allows for a broader class of models (Lee & Nelder, 1996).

Lee and Nelder (1996) defined the hierarchical likelihood for $\mathbf{y}$

$$h = \ln\left(\boldsymbol{f}\left(\boldsymbol{y}|\boldsymbol{v};\boldsymbol{\beta},\boldsymbol{\phi}\right)\right) + \ln\ \left(\boldsymbol{f}\left(\boldsymbol{v};\boldsymbol{\alpha}\right)\right) \tag{15}$$

$$\equiv l\ \ \left(\boldsymbol{\beta},\boldsymbol{\phi};\boldsymbol{y}|\boldsymbol{v}\right) + l\ \ \left(\boldsymbol{\alpha};\boldsymbol{v}\right), \tag{16}$$

where $f(\boldsymbol{y}|\boldsymbol{v};\boldsymbol{\beta})$ and $f(\boldsymbol{v};\boldsymbol{\alpha})$ denote the condition density function of $\mathbf{y}$ given random effect $\mathbf{v}$, and the density function of v, respectively. One reason for developing an algorithm for the v-scale rather than for the u-scale is that v could often assume any real value whereas u usually has range restrictions, which may cause problems in convergence (Lee & Nelder, 1996). The random component $\mathbf{v}$ is the scale on which the random effect u occurs linearly in the linear predictor, $v = v(u)$, where $\boldsymbol{\beta}$ are fixed effects, $\boldsymbol{\phi}$ are the dispersion parameters for the conditional distribution of $y|v$, and $\boldsymbol{\alpha}$ are the parameters for the random effects.

Call estimates are derived from maximizing the h-likelihood and the maximum h-likelihood estimates (MHLEs); these are obtained by solving:

$$\frac{\partial h}{\partial \boldsymbol{\beta}} = 0,$$

$$\frac{\partial h}{\partial \boldsymbol{v}} = 0.$$

Unfortunately, the estimation of random parameters and dispersion parameters are biased estimators when using h-likelihood. The dispersion components are estimated by maximizing the adjusted profile h-likelihood, which is restricted likelihood for the dispersion parameters.

An adjusted profile h-likelihood leads to reliable and useful estimators (Lee & Nelder, 2001). To estimate the dispersion parameters, Lee and Nelder (1996) proposed an adjusted h-likelihood,

$$h_A = \left(h + \frac{1}{2} \ln |2\pi \boldsymbol{\phi} \boldsymbol{H^{-1}}| \right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{v}=\hat{\boldsymbol{v}}},$$

where H is the Hessian matrix of the h-likelihood,

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H_1} & \boldsymbol{H_2} \\ \boldsymbol{H_2^T} & \boldsymbol{H_3} \end{bmatrix},$$

where

$$H_{1ij} = -\frac{\partial^2 h}{\partial \beta_i \, \partial \beta_j},$$

$$H_{2ik} = -\frac{\partial^2 h}{\partial \beta_i \, \partial v_k},$$

and

$$H_{3kl} = -\frac{\partial^2 h}{\partial v_k \, \partial v_l}.$$

The maximum adjusted profile h-likelihood estimators for random effect parameter $\boldsymbol{\alpha}$ and dispersion parameters $\boldsymbol{\phi}$ are obtaining by solving

$$\frac{\partial h_A}{\partial \boldsymbol{\alpha}} = 0,$$
$$\frac{\partial h_A}{\partial \boldsymbol{\phi}} = 0.$$

As an example to explain the HGLM, we focus on the binary outcome in this work. According to (Lee & Nelder, 1996), the appropriate distribution for the dependent variable is binomial (since the outcome is binary) and the appropriate distribution for the random effect is a beta distribution, example for binary data outcome with beta distribution for random effects by (Lalonde, 2009) and (Lee & Nelder, 1996). The HGLM pieces are as follows:

The response distribution is

$$Y_{ij}|u_i \sim \quad Bin(\mu, \mu(1-\mu)),$$

the random distribution is

$$u_i \sim \quad Beta\ (\gamma, \lambda),$$

the linear component is

$$\eta_{ij} = \ x_{ij}\beta + v(u_i),$$

the link function is

$$\eta_{ij} = \ logit\ (p),$$

the $h-$likelihood for binomial-beta model (Lee & Nelder, 1996)

$$h = l \ \ (\boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{y}|\boldsymbol{v}) + l \ \ (\boldsymbol{\alpha}; \boldsymbol{v}).$$

where the pdf of the binomial distribution

$$f_Y(y_{ij}|v_i; p) = \binom{n_i}{y_{ij}} p_{ij}^{y} (1-p)^{(n_i - y_{ij})},$$

The canonical and dispersion parameters are

$$\theta = \ln\left(\tfrac{p}{1-p}\right), \quad a(\phi) = 1, \quad b = \ln\left(1 + e^{\theta}\right)^{n}, \text{ and } c(y, \phi) = \ln\tbinom{n}{y},$$

and the ln-likelihood for p,

$$l \ \ (\phi; y_{ij}|v_i) = y_{ij}\theta - \ln\left(1 + e^{\theta}\right).$$

The linear component is $\theta = x_{ij}\boldsymbol{\beta} + v(u_i)$, and by summing over all observations, the ln-likelihood

$$l(\boldsymbol{\beta}, \boldsymbol{v}; \boldsymbol{y}|\boldsymbol{v}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n_i}} [y_{ij} \left(x_{ij}\boldsymbol{\beta} + v_i\right) - \ln\left(1 + e^{(x_{ij}\boldsymbol{\beta} + v_i)}\right)].$$

The pdf for random component (beta distribution) is

$$f_{u_i}\left(u_i; \gamma, \lambda\right) = \frac{\Gamma(\gamma)\Gamma(\lambda)}{\Gamma(\gamma + \lambda)} \ u_i^{(\gamma - 1)} \ \left(1 - u_i\right)^{(\lambda - 1)}.$$

and the beta function

$$B(\gamma, \lambda) = \frac{\Gamma(\gamma)\Gamma(\lambda)}{\Gamma(\gamma + \lambda)}$$

and the relationship $v_i = v(u_i) = \ln\left(u_i\right)$, the ln likelihood for parameters γ and λ from Lee and Nelder (2006) are

$$l(\gamma, \lambda; v_i) = \gamma \ v_i - (\gamma + \lambda) \ ln \ (1 + e^{v_i}) - \ n_i \ ln \left(B(\gamma + \lambda)\right)$$

Summing over all observation u_i

$$l(\gamma, \lambda; v) = \sum_{i=1}^{k} [\gamma \ v_i - (\gamma + \lambda) ln(1 + e^{v_i})] - n_i \ ln \left(B(\gamma + \lambda) \right).$$

As such, the $h-$likelihood estimation equation for the fixed part $\boldsymbol{\beta}$ and random component $\mathbf{v}$ respectively are

$$\frac{\partial h}{\partial \beta_k} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n_i}} \left[x_{ijk} y_{ij} - n_i x_{ijk} \frac{e^{(x_{ij}\boldsymbol{\beta} + v_i)}}{1 + e^{(x_{ij}\boldsymbol{\beta} + v_i)}} \right] = 0, \tag{17}$$

Thus,

$$\hat{\beta}_k = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n_i}} [(y_{ij} - n_i \ p_i) x_{ijk}] = 0,$$

where

$$p_i = \frac{e^{(x_{ij}\boldsymbol{\beta} + v_i)}}{1 + e^{(x_{ij}\boldsymbol{\beta} + v_i)}},$$

and

$$\hat{v_i} = \frac{\partial h}{\partial v_i} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{n_i}} \left[y_{ij} - \frac{e^{(x_{ij}\boldsymbol{\beta} + v_i)}}{1 + e^{(x_{ij}\boldsymbol{\beta} + v_i)}} \right] + \gamma - (\gamma + \lambda) \frac{e^{(v_i)}}{1 + e^{(v_i)}} = 0. \tag{18}$$

Thus, equating $\dfrac{\partial h}{\partial v_i}$ to zero gives an estimate of the random effect

$$\hat{u_i} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{y_{ij} - n_i \ p_i + \lambda}{\lambda + \gamma}.$$

Then we could solve equations (17) and (18) by using either a Newton Raphson method or a Fisher's scoring method (Gu, 2008).

Double Extended Quasi Likelihood

In the last section, we saw that h-likelihood estimation required us to specify the full distribution of the response variable and any random effects. Extended

quasi-likelihood is an extension to h-likelihood, which is more convenient to use. A less restrictive estimation method is double extended quasi-likelihood (DEQL). Double extended quasi-likelihood (DEQL) requires us to specify the first and second moments for the response variable and random effects. DEQL pertains to hierarchical generalized linear model. Lee and Nelder (2001) proposed using the double extended quasi-likelihood for inference from quasi-likelihood models. From Lee et al., (2006), we summarized the Q^{++} as

$$Q^{++} \quad = \quad Q(\theta(\mu), \phi; y|u) \;\; + \;\; Q_R(u; v_R),$$

where

$$2Q(\theta(\mu), \phi; y|u) = -\sum_{i=1}^{N} \left[\frac{d_i}{\phi_i} + \ln 2 \; \pi \; \phi_i \; V_(y_i) \right],$$

and

$$2Q_R(u; v_R) = -\sum_{j=1}^{M} \left[\frac{d_{Rj}}{u_j} + \ln 2 \; \pi u_j V_R(v_j) \right].$$

The deviance components of $y|u$ are

$$d_i = 2 \int_{y_i}^{\mu_i} \frac{(y_i - s)}{V_{(s)}} ds,$$

and the deviance components of u similarly are

$$d_{Rj} = 2 \int_{v_j}^{u_j} \frac{(v_{Rj} - s)}{V_{1j}(s)}.$$

Estimate the fixed parameters $\boldsymbol{\beta}$ and random effects $\mathbf{v}$ by equating first derivatives of Q^{++} to zero.

The Hierarchical (Nested) Model

Agresti (2007) defined a hierarchical model as one which is appropriate for observations that have a nested structure. In this type of data, units at one level are contained within units of another level. Hierarchical data are common in certain application areas, e.g, in educational, agricultural, genetic, industrial, medical and other types of research. For example, a study of factors that affect student performance might measure, for each student and each exam in a battery of exams, whether the student passed. Students are nested within schools, and the model could study variability among students as well as variability among schools.

In the treatment structure, which consists of the various treatments or treatment combinations that the experimenter wishes to study, nesting occurs when the levels of one factor occur with only one level of a second factor. In that case, the levels of the first factor are said to be nested within the level of the second factor. For example, suppose an animal scientist wants to study the growth rate of lambs. She has 4 males (sires, Factor (A)) and 12 females (dams, factor (B)). The breeding structure is shown in Table 2 (an "X" denotes a mating). For this example, each sire is mated to three dams, the three dams being different for each sire. Thus, dam is called a nested effect, where dam is nested within sire, we write this as "B(A)". When nesting occurs in the treatment structure, the treatment structure consist of at least two factors, according to McCullagh and Searle (2001).

For a nested model in which the dependent variable $\mathbf{Y}$ is a binary outcome, each component Y_i is assumed to follow a Binomial distribution,

Table 2

Breeding structure showing dams nested within sires

	DAMS											
Sire	1	2	3	4	5	6	7	8	9	10	11	12
1	X	X	X									
2				X	X	X						
3							X	X	X			
4										X	X	X

$$\boldsymbol{Y_i \sim Bin(n, P).}$$

Nested (or hierarchical) classifications are usually analyzed using mixed models. Most of the time, the nested factor is random effect from the population under study, and the nested factor is a fixed or a random effect. If there is another fixed factor, then the mixed effects model is the most appropriate in the nested design (Searle, 1987). The nested model (or hierarchical model) is a particular technique for representing a nested design. For example, we could have factor A represent hospitals as a "random effect", and factor B represent the patient as a fixed effect. We randomly chose the number of hospitals in a specific area and observe the patient in each hospital, i.e., patients had surgery and whether the patient lived or died. Given this, B is a fixed effect nested within the random effect A. Here, factor A has different dispersions that reflect the different hospitals chosen.

Nested Design Models

Two stage nested design model, In the treatment structure, each level of factor B occurs with only one level of factor A. For the mixed model structure, we considered that factor A is fixed and factor B is random.

The Hierarchical Generalized Linear Model of two-stage nested designs is given by:

$$y_{ijk} \sim Bin(\mu,\ \mu\ (1-\mu)),$$

$$u_i \sim \mathcal{D}(\mu_R, V_R),$$

$$\eta_{ij} = X\beta + Zu_i,$$

$$\eta_{ij} = logit(p_{ij}).$$

Where Y_{ijK} is the dependent variable following binomial distribution with parameters n and p. The parameter u_i is the random effect that follows any distribution from the exponential family distribution with mean equal to μ_R and V_R is the variance covariance matrices. X and Z are the explanatory variables for the fixed and random effects respectively, and g is the link function which is logit for binomial distribution.

$i = 1, 2, ..., K; \quad j = 1, 2, ..., n_i$, and k the number of observations $k = 1, 2, ..., n_{ij}$. The parameters $\boldsymbol{\beta}$ is the vector parameter for the fixed effect, u_i is the parameter of the random effect.

Clustered Data Models

Experimental designs with hierarchical (nested) classifications are frequently used in agricultural, genetic, industrial, medical, biological, and even in social sci-

ence field experiments. Clustered data or nested design is an experimental design in which the data have an implicit hierarchy. The clusters may be balanced or unbalanced, i.e., the number of observations in a cluster (the size of the cluster) is equal or unequal. The unbalanced clustered data bring up the problem of heterogeneous models which require different variance components, as had been addressed in previous studies for continuous response (Abdoslam, 2004). In the case of unbalanced clustered data with continuous outcomes in the linear model, Abdoslam (2004) found that there was a different dispersions for different clusters sizes. Accounting for the different dispersions led to the minimization of mean square error, which was shown through two examples. In this study, the researcher focused on the binary outcomes. When using mixed effects for clustered data with binary outcomes, a preferred model is Hierarchical Generalized Linear Model (HGLM).

Clustered Binary Data Models

Models for clustered binary data are important in many areas, e.g, medical, education, finance, and many other research areas where the outcome has only two possible values.

For the nested model with binary response, there are two sources of variation. The first source of variation is the between-cluster variation, which represents the variation from cluster to cluster. The second source of variation is the within-cluster variation which represents the variation inside each cluster, and it is a constant $\frac{\pi^2}{3}$ for the logistic distribution, (Bauer, 2009).

Dai (2006) explained the use of the GLIMMIX package in SAS as an example of model fitting and testing hypotheses of clustered binary data. The authors

considered two-level models, which were the patient-level and the hospital-level effects. The two level data structure is shown in the figure below.

Hospital	H_1	H_2	....	H_I
Patients	1 2 ... n_1	1 2 ... n_2	....	1 2 ... n_i

Here n_i is the number of patients; $i = 1, 2, ..., I$ the patient level indicator in the i^{th} hospital. The model is

$$logit(p_{ij}) = \ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta x_{ij} + u_i.$$

where β is the vector of fixed parameter, x_{ij} is the patient j in the hospital i, and u_i is the random variable here to represent the hospital effect. The authors use SAS code to analyze this data and suggested that the SAS GLIMMIX procedure is a highly useful tool for hierarchical modelling with binary responses. The GLIMMIX procedure in SAS uses Restricted Pseudo Likelihood (REPL) to estimate the parameters, which assumes constant dispersion from cluster to cluster. Alternatively, in HGLM, we could use penalized quasi-likelihood (PQL) or h-likelihood (HL) to estimate the parameter, fit the models, and test hypotheses.

Balanced Clustered Binary Data Models

The equal cluster size with binary outcomes means each cluster consists of the same number of subjects with two possible outcomes. To estimate the parameters in balanced clustered binary data models, it is possible to use generalized estimating equations (GEE) or the hierarchical generalized linear model (HGLM); both methods pertain to HGLM and many books mentioned that these methods may be used to obtain good estimates for parameters and fitting the model. Wang (2010)

used a GEE for analysis of clustered binary data with a large number of covariates, and he found it worked well even when the number of covariates grew to infinity. To estimate the parameters by using one of the methods in HGLM, suppose the dispersion equals one, that the dispersions across clusters are not different, and that the variance for the random effect is constant (Fitzmaurice & Ware, 2004).

Unbalanced Clustered Binary Data Models

An unequal cluster size with a binary outcome is common in many areas of application, especially in medical research. Sample size formulas for cluster randomized trials were based on the assumption of equal cluster sizes, but in practice this assumption would rarely be met. Many designs evaluating the effect of an intervention are characterized by a nesting of subjects within clusters. Owing to variation in actual cluster sizes, but also due to non-response or drop-out, unequal cluster sizes are rather common. There were many research studies for unequal cluster size with continuous outcome, but few applied to binary outcomes. Here we discussed some authors who studied unequal cluster size with binary data and their method to estimate parameters.

Unequal Cluster Size Using Maximum Likelihood

For unequal cluster sizes with binary outcomes, suppose the random effect follows a normal distribution, then the model is the Binomial-Normal HGLM (Lee & Nelder, 1996).

The Binomial-Normal distribution could have been written,

$$\mathbf{1.}\ \boldsymbol{Y|u \sim Bin(n, P)},$$

$$\boldsymbol{u \sim \mathcal{N}(0, \sigma^2 I)},$$

$$\mathbf{2.}\ \boldsymbol{\eta = X\beta + Zu},$$

$$\mathbf{3.}\ \boldsymbol{\eta} = logit(\boldsymbol{P}).$$

Since the random term follows a normal distribution, we could use the maximum likelihood estimation (MLE). Heo and Leon (2005) and Neuhaus and Lesperance (1996) studied performance of a mixed effects logistic regression model for binary outcomes with unequal cluster size. Both authors consider the following probability model for the clustered binary outcomes with an intervention indicator x_{ij}

$$logit(p_{ij}) = \beta x_{ij} + u_i,$$

where Y is a binary outcomes variable (e.g, the patient survived or died after a surgery), $logit(p) = ln\left(\frac{p}{(1-p)}\right)$, $p_{ij} = E\left(y_{ij}|x_{ij}, u_i\right)$, x_{ij} is a patient-level predictor. The random variable u_i reflects a random effect specific to the i^{th} cluster and the variance of u reflects a degree of heterogeneity in "frailty" across the clusters. Here, u_i is assumed to be normally distributed with mean zero and unknown variance σ^2, $\boldsymbol{u \sim N(0, \sigma^2 I)}$. Moreover, they assumed u_i and y_{ij} are conditionally independent over j. The first variation is the between cluster variation. The second variation is the within cluster variation which represent the variation inside each cluster, and the authors used the constant $(\frac{\pi^2}{3})$ (Hedeker & Gibbons, 1994).

To compare the performance of the mixed effects logistic regression model for binary outcomes with unequal cluster size, Heo and Leon (2005) used maximum

likelihood estimation since they assumed normality of the random effect. Their simulation study compared the performance of maximum likelihood estimation in a mixed effects logistic regression model for equal and unequal cluster size. These simulation results applied where the cluster size n is as small as 5. Overall, the results were insensitive to variability in cluster size across the clusters. Neuhaus and Lesperance (1996) investigated the efficiency of conditional likelihood, which eliminates the random intercept terms and likelihood generated from the marginal distribution of the data where the random intercepts are integrated out. By using simulation and example data, they showed the asymptotic relative efficiency of conditional likelihood estimators relative to parametric estimators was a decreasing function of within-cluster covariate correlation. Also, their simulation results showed, for fixed covariate correlation, the asymptotic relative efficiency of the parametric versus the conditional increases as cluster sizes increase. The normality of the random effects distribution was assumed, but it was not the best method because this assumption did not always hold (Lee & Nelder, 1996).

Unequal Cluster Size Using Penalized Quasi Likelihood

In unequal cluster sizes with binary outcomes, without knowing the distribution of the random component, we could use any distribution for the random

component. Since the dependent variable follows a binomial distribution, then the HGLM component is

$$\mathbf{1.}\ \boldsymbol{Y|u \sim Bin(n, P)},$$

$$\boldsymbol{u \sim \mathcal{N}(0, V_R)},$$

$$\mathbf{2.}\ \boldsymbol{\eta = X\beta + Zu},$$

$$\mathbf{3.}\ \boldsymbol{\eta = logit(P)}.$$

Using the penalized quasi-likelihood method to estimate the generalized linear mixed model's parameter, Candel and Breukelen (2010) handled the unequal cluster size with binary outcomes to estimate the efficiency loss due to unequal cluster size for a mixed effects model. Their model was

$$ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta x_{ij} + u_i$$

Their model and their assumption for normality of the random effect was the same as Heo and Leon (2005), and they also used the same constant $\frac{\pi^2}{3}$ for within cluster variation. Candel and Breukelen (2010) found 14 percent more observation within cluster is sufficient to repair the efficiency loss due to varying cluster size. As mentioned, they used the PQL method with binary outcome, but there are many authors who do not agree with using PQL when the outcome is binary because PQL could underestimate parameters (Jang & Lim, 2006).

Comparing the Estimation of Models for Unbalanced Clustered Binary Data

Comparing estimation of parameters for Unbalanced Clustered Binary Data is not easy, and the results are not the same as when the outcome is continuous.

For continuous outcomes, we compared two estimation methods for models according to the standard deviation or power. Here, when the response variables were binary, it was hard to make comparisons. Previous research had compared methods of estimation for fitting models to binary data.

Bauer (2009) studied the use of linear models for binary outcomes. When fitting models for binary outcomes, comparisons between such models were impeded by the implicit rescaling of the model estimates that took place with the inclusion of random effects. He presented an approach for putting the estimates on a common scale to facilitate relative comparisons between model fit to binary outcomes. He compared two methods for binary outcomes: generalized estimating equations (GEE) and hierarchical generalized linear model (HGLM). These models were referred to as marginal and conditional models, respectively. Bauer (2009) found that the rescaled estimates are intended to be used primarily for making relative comparisons between models. Lee and Nelder (2009) did not agree with using generalized estimating equations (GEE) and generalized linear mixed model (GLMM) to compare the models. They argued that the use of an estimation method without a probabilistic term was problematic and the GEE method was not probabilistic.

Bauer and Sterba (2011) compared two generalized linear estimation methods to employ when instead fitting multilevel cumulative logit models to ordinal data: maximum likelihood (ML) or penalized quasi-likelihood (PQL). ML and PQL were compared across variations in sample size, magnitude of variance components, number of outcome categories, and distribution shape. Fitting a multilevel linear model to ordinal outcomes is shown to be inferior in virtually all circumstances.

PQL performance improves markedly with the number of ordinal categories, regardless of distribution shape. In contrast to binary data, PQL often performs as well as ML when used with ordinal data. Further, the performance of PQL is typically superior to ML when the data include a small to moderate number of clusters. Even Bauer and Sterba (2011)'s updated article, he still used the PQL method with binary outcomes. There are many authors who do not agree with using PQL with binary outcomes because it has been shown the underestimate parameter (Jang & Lim, 2006).

None of the accepted methods reviewed in Chapter II allowed overdispersion terms to be different from cluster to cluster. To handle the varying dispersion from cluster to cluster, we needed to correct one of the hierarchical generalized linear model (HGLM) estimation methods to estimate the mean and dispersion parameters. In the next chapter, two methods were presented to handle this difference in variation across clusters. The first method was an extension of REPL using ML to estimate the parameter, and the second method was an adjustment the binomial beta model using h-likelihood.

CHAPTER III

UNBALANCED CLUSTERED BINARY DATA MODELS

Many research studies in health, finance, education, and social sciences have involved collecting binary data clustered into groups, such as the smoking status of students sampled from different schools or disease status of animals from different farms. Such data would be expected to be correlated within clusters, as students from the same school would tend to be more similar than those from different schools, and animals from the same farm would tend to be more similar than those from different farms. When designing such studies, a choice need to be made regarding the number of groups to sample from. A larger number of groups or schools resulted in less dependence in the data and more precision in estimating the effects of explanatory variables. In some experiments, the clusters were unbalanced; that is, the number of observations in a cluster (the size of the cluster), differs among the clusters.

Unbalanced clusters resulted from sub-sampling unequal numbers of observations from each cluster. Unbalanced clusters also occurred when there were randomly missing vector elements for a clustered multivariate outcome or if subjects differed in the number of relevant vector elements for the analysis. The different cluster size could lead to different dispersions for each cluster. This unbalanced

data in each cluster brought up the problem of heterogeneous models which required different variance components, as had been addressed in previous studies for continues response (Abdoslam, 2004). In this study, the researcher used a nested design. The mixed model was used in this study because it was the most appropriate model to use in practice, as it contained fixed and random factors.

In this chapter, the researcher aims to quantify the effect of varying cluster sizes in parameter estimation for nested binary data with unbalanced clusters. Some authors have studied the efficiency in a binary mixed effect model when applied to unbalanced clustered binary data. They found losses in efficiency because of the unbalance.

Breukelen and Candel (2012) pointed out that there were many publications that discussed losses of efficiency for treatment evaluation that were due to cluster size variation in cluster randomized trials. These studies focused on how to increase the efficiency by increasing sample size or by adjusting the number of cell by using the hierarchical generalized linear model. There was no study that tried to adjust the method or address efficiency directly to the problems that were created by having different sizes for each cluster.

By adjusting two methods, and investigating the methods through computer simulation, we answered the research questions:

Q1 Does Extended Restricted Pseudo Likelihood account for different dispersion for different clusters size?

Q2 Does Adjusted Scale Binomial Beta h-likelihood account for different dispersions for different cluster size?

Q3 Is Extended Restricted Pseudo Likelihood more powerful than Restricted Pseudo Likelihood in the case of unbalanced binary clustered data?

Q4 Is Adjusted Scale Binomial Beta h-likelihood more powerful than Binomial Beta h-likelihood in the case of unbalanced binary clustered data?

Q5 Does Extended Restricted Pseudo Likelihood method improve efficiency?

Q6 Does Adjusted Scale Binomial Beta h-likelihood method improve efficiency?

In this chapter, the researcher presented two methods of accounting for different dispersions across clusters as a result of unequal cluster size. The researcher expected to get more efficiency and low Type I error rate using the two adjusted HGLM methods. The first method was an Extension of Restricted Pseudo Likelihood (EREPL) estimation that allowed the dispersion parameter ϕ to be different from cluster to cluster $\boldsymbol{\phi_i}$. The second method was an Adjusted Scale Binomial Beta model in which the dependent variable followed a binomial distribution and the random effect followed beta distribution with the same mean and different scale parameter from cluster to cluster.

Extended Restricted Pseudo Likelihood for Unequal Cluster Size

In Chapter II under the heading Restricted Pseudo Likelihood Estimation, a marginal pseudo model was described according to Wofinger and O'Connell (1993).

The marginal pseudo response variable is distributed as

$$\boldsymbol{P \sim \mathcal{N}(X\beta, ZV_R Z^T + \hat{D}cov(Y)\hat{D}^T),}$$

where

$$\boldsymbol{\hat{D} = \frac{\partial g(\mu)}{\partial \mu} \mid_{\mu=\mu_0}.}$$

For the overdispersion parameter $\boldsymbol{\phi}$, Wolfinger and O'Connell (1993) suggested assuming an equal dispersion parameter and assuming it is equal to one, $\boldsymbol{\phi = 1}$. The dispersion is equal from cluster to cluster. If the dispersion parameter $\boldsymbol{\phi}$ is constant across clusters, but it does not equal one, the estimator of parameter $\boldsymbol{\phi}$ is

$$\boldsymbol{\hat{\phi} = \hat{r}^T \ \hat{V}^{-1} \ \hat{r} \ /n.}$$

In the Restricted Pseudo Likelihood method, the dispersion parameter $\boldsymbol{\phi}$ is constant, and it does not account for different variation across clusters.

The researcher proposed the Extended Restricted Pseudo Likelihood and the Pseudo Likelihood with different dispersion $\boldsymbol{\phi_i}$, where $i = 1, 2, ..., K$ with K clusters. The vector of dispersion is $\boldsymbol{\phi = [\phi_1, \phi_2, ..., \phi_K]^T}$. Using Extended Restricted Pseudo Likelihood (EREPL) $\boldsymbol{\phi_i}$ to fit a mixed effect model for binary outcomes with unequal cluster size, the HGLM was considered,

$$\textbf{1.} \quad \boldsymbol{Y|u \sim D(\mu, \phi_i \ \mu(1-\mu))),}$$

$$\boldsymbol{u \sim \mathcal{N}(0, V_R),}$$

$$\textbf{2.} \quad \boldsymbol{\eta = X\beta + Zu,}$$

$$\textbf{3.} \quad \boldsymbol{\eta} = g(\boldsymbol{\mu}).$$

The estimate of the mean parameter vector remained unchanged,

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X^T}\hat{\boldsymbol{V_P}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X^T}\hat{\boldsymbol{V_P}}^{-1}\boldsymbol{P}.$$

where $\boldsymbol{P}$ is the vector of pseudo response, and $\boldsymbol{V_P}$ is the variance covariance matrix for pseudo response. The estimate of the random effect parameter vector remained unchanged as well,

$$\hat{\boldsymbol{u}} = \hat{\boldsymbol{V_R}}\boldsymbol{Z^T}\hat{\boldsymbol{V_P}}^{-1}\,\hat{\boldsymbol{r}},$$

where $\boldsymbol{V_R}$ is the variance covariance matrix for random effect u, and r is the residual $r = P - \hat{P}$. The estimation of dispersion constants,

$$\hat{\boldsymbol{\phi}}_{\boldsymbol{i}} = \hat{\boldsymbol{r}}_{\boldsymbol{i}}^{\boldsymbol{T}}\,\hat{\boldsymbol{V}}_{\boldsymbol{i}}^{-1}\,\hat{\boldsymbol{r}}_{\boldsymbol{i}}\,/\boldsymbol{n_i},$$

where n_i is the number of observations in each cluster, $i = 1, 2, ..., K$ the cluster from 1 to K^{th}, and $\hat{r}$ is the residual for each cluster, the residual being different in each cluster. $\hat{\boldsymbol{V}}_{\boldsymbol{i}}$ is the variance covariance matrix which has diagonal entries that represent variances for each cluster and zeros in the off diagonal, assuming clusters are independent.

In the Extended Restricted Pseudo Likelihood method, the random effect is assumed to be normally distributed, and maximum likelihood is applied to the pseudo response. For a more appropriate method, when normality for the random effect does not hold, we suggested to adjust the scale parameter of the Binomial Beta HGLM and use h-likelihood to get the estimated value of parameters.

Adjusted Scale Binomial Beta for Unequal Cluster Size

Our goal in this chapter is to estimate the parameters for unequal cluster sizes in a nested model with binary outcomes. Since we focus on the binary outcomes as the dependent variable, the appropriate distribution for the random effects is the beta distribution. Assuming a normal distribution is convenient, but it is not always the best choice in a HGLM (Lee & Nelder, 1996). By assuming the conditional dependent variable $Y|u$ is binomial, and by assuming a beta distribution for the random effect, the distribution of conditional response and random effect are fully specified. In this case the appropriate estimation method is $h-$likelihood (Lee & Nelder, 1996). Assume the model

$$1.\ Y_{ij}|u_i \sim \quad Bin(n, p_{ij})),$$

$$u_i \sim \quad Beta\ (\gamma, \lambda_i),$$

$$2.\ \eta_{ij} = \ \boldsymbol{x_{ij}\beta} + v(u_i),$$

$$3.\ \eta_{ij} = \ logit\ (p_{ij}).$$

where λ_i is the scale parameter for the beta distribution. It varied from cluster to cluster, where i is the number of clusters $i = 1, 2, ..., K$. The $h-$likelihood for the Binomial- Beta model (Lee and Nelder, 1996)

$$\boldsymbol{h = l\ (\beta, \phi; y|v) + l\ (\alpha; v)}.$$

The binomial pdf for the dependent variable is

$$f_Y(y_{ij}|v_i; p) = \binom{n_i}{y_{ij}} p_{ij}^{y} (1-p)^{(n_i - y_{ij})},$$

and the pdf for the random effect is

$$f_{u_i}\left(u_i;\gamma,\lambda_i\right)=\frac{\Gamma(\gamma)\Gamma(\lambda_i)}{\Gamma(\gamma+\lambda_i)}\ u_i^{(\gamma-1)}\ \left(1-u_i\right)^{(\lambda_i-1)}.$$

$$\boldsymbol{h\left(\beta,\gamma,\lambda_i;y|v\right)=l\ \left(\beta,v;y|v\right)+l\ \left(\gamma,\lambda_i;v\right)}.$$

where $l\ (\beta,v;y|v)$ was unchanged from Chapter II,

$$l\ (\beta,v;y|v)=\boldsymbol{l\left(\beta,v;y|v\right)}=\sum_{\boldsymbol{i=1}}^{\boldsymbol{K}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{n_i}}\left[y_{ij}\left(x_{ij}\boldsymbol{\beta}+v_i\right)-\ln\left(1+e^{(x_{ij}\boldsymbol{\beta}+v_i)}\right)\right],$$

and $l\ (\gamma,\lambda_i;v)$ would be

$$l(\gamma,\lambda_i;v_i)=\gamma\ v_i-(\gamma+\lambda_i)\ ln\ (1+e^{v_i})-n_i\ ln(B(\gamma+\lambda_i))$$

Summing over all observations u_i

$$l(\gamma,\lambda_i;v_i)=\sum_{i=1}^{K}[\gamma\ v_i-(\gamma+\lambda_i)ln(1+e^{v_i})-n_i\ ln(B(\gamma+\lambda_i)).$$

Then the $h-$likelihood estimating equation for fixed parameters $\boldsymbol{\beta}$ and random components $\mathbf{v}$ are

$$\frac{\partial h}{\partial\beta_k}=\sum_{\boldsymbol{i=1}}^{\boldsymbol{K}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{n_i}}\left[x_{ijk}y_{ij}-n_ix_{ijk}\frac{e^{(x_{ij}\boldsymbol{\beta}+v_i)}}{1+e^{(x_{ij}\boldsymbol{\beta}+v_i)}}\right]=0,$$

and

$$\frac{\partial h}{\partial v_i}=\sum_{\boldsymbol{j=1}}^{\boldsymbol{n_i}}\left[y_{ij}-n_l\frac{e^{(x_{ij}\boldsymbol{\beta}+v_i)}}{1+e^{(x_{ij}\boldsymbol{\beta}+v_i)}}\right]+\gamma+\frac{\left(e^{(v_i(1-\gamma-\lambda_i))}\right)-1}{1+e^{(v_i)}}=0.$$

Thus, equating $\dfrac{\partial h}{\partial v_i}$ to zero gives an estimate of the random effect

$$\hat{u_i}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{y_{ij}-n_i\ p_i+\lambda}{\lambda+\gamma_i}.$$

The estimation of random parameters and dispersion parameters are biased estimators when using h-likelihood. The dispersion components are estimated by maximizing the adjusted profile h-likelihood, which is the Restricted likelihood for the dispersion parameters.

An adjusted profile h-likelihood leads to reliable and useful estimators (Lee & Nelder, 2001). Estimating the random parameters and the dispersion parameter remained the same as In Chapter II under the heading Restricted Pseudo Likelihood Estimation,

$$h_A = \left(h + \frac{1}{2} log|2\pi \boldsymbol{\phi} \boldsymbol{H^{-1}}| \right)_{\beta=\hat{\beta}, v=\hat{v}},$$

where H is the Hessian matrix of the h-likelihood.

The maximum adjusted profile h-likelihood estimators for the random effect parameter $\boldsymbol{\gamma}, \boldsymbol{\lambda_i}$ and dispersion parameters $\boldsymbol{\phi}$ are obtaining by solving the equations (Lee & Nelder, 1996) deriving the equations for general random effect and dispersion effect:

$$\frac{\partial h_A}{\partial \boldsymbol{\gamma}} = 0,$$
$$\frac{\partial h_A}{\partial \boldsymbol{\lambda_i}} = 0,$$
$$\frac{\partial h_A}{\partial \boldsymbol{\phi}} = 0.$$

Because of varying variation from cluster to cluster, adjusting the parameter scale for Binomial Beta distribution allows the h-likelihood to have inter-cluster correlation.

The two methods presented in this chapter led to higher efficiency and lower Type I error rate of the design. To investigate whether or not the presented two methods were more appropriate for dealing with different variance components for unbalanced cluster binary data models, a computer simulations was presented in the next chapter to investigate the methodology by comparing the two presented methods to REPL and h-likelihood in terms of power, Type I error, and standard error. These new methods were expected to have more power and small Type I errors in the case of unbalanced binary clustered data. In the next chapter, a simulation for comparing the performance of the four methods was presented.

CHAPTER IV

SIMULATION

Unbalanced cluster size has lead to different dispersions for each cluster. The unbalanced data in each cluster brought up the problem of heterogeneous models, which required different variance components. In this study, the researcher studied the unbalanced cluster size for binary outcomes. In this chapter, the researcher explained the generating data and simulation steps to find the performance of the adjusted methods that dealt with unbalanced cluster size for binary outcomes. The results for each simulation step were explained for each method and comparisons made.

The simulation for comparing the performance of each of the four methods presented were:

1. Restricted Pseudo Likelihood.
2. Extended Restricted Pseudo Likelihood.
3. Binomial Beta h-likelihood.
4. Adjusted Scale Binomial Beta h-likelihood.

These four models were evaluated in terms of their power, Type I error rate, and standard error for parameter estimates through computer simulations of the

number of clusters, number of observations in each cluster, and fixed values for parameters. In the next sections, the estimation methods and their results were discussed. The first section describes the data generation for each method and simulation steps. The second section explains the Restricted Pseudo Likelihood method and simulation results with figures. The third section explores the Extended Restricted Pseudo Likelihood method and showed the process that allowed for this adjusted method. The fourth section explains the Binomial Beta h-likelihood method and its results with figures. The next section explores the Adjusted Scale Binomial Beta h-likelihood method and simulation results with figures. The last section compares all estimation methods.

Steps of Simulation

For generating data, in which the researcher defined the values for parameters and generated the X values, random effect variable, and calculated the probability p of the dependent variable y. First was generated an unequal number of subjects n_i per cluster from the Poisson distribution for unequal cluster size. The mean from the Poisson distribution was the mean for the number of observations for each cluster. By choosing three different varying mean cluster sizes ($\bar{n} = 10, 25, 100$), the researcher showed the difference in statistical performance for various sample sizes.

The next step was to generate a normally distributed continuous variable X_{ij} with mean = 3 and a known variance = 20; $x_{1ij} \sim \mathcal{N}(3, 20)$. Thus, the researcher generated a beta distributed random variable u_i with a parameter $\gamma = 2$ and $\lambda = 3$ for each cluster i; $u_i \sim Beta(2, 3)$. Finally, Y_{ij} was generated for

each data unit randomly from a Bernoulli distribution with a success probability p_{ij},where

$$p_{ij} = \frac{e^{\beta_0+\beta_1\ x_{1ij}+u_i}}{1+e^{\beta_0+\beta_1\ x_{1ij}+u_i}},$$

and $\beta_0 = 1,\ \beta_1 = 0.2$. Parameter estimates were obtained using Restricted Pseudo Likelihood, Extended Restricted Pseudo Likelihood, Binomial Beta h-likelihood, and Adjusted Scale Binomial Beta h-likelihood (Heo & Leon, 2005).

The project defined K to be the number of clusters [$K = 20, 50$] and $\bar{n}$ to be the mean number of observations per cluster [$\bar{n} = 10, 25, 100$]. For each combination of K and $\bar{n}$, $1,000$ data sets were generated to calculate the power, Type I error, and standard errors. To calculate the power, Type I error rate, and standard error, data were generated according to the model with the systematic component $\eta_{ij} = \beta_0 + \beta_1\ x_{1ij} + v_i$, with one affected treatment of β_1. Thus, the model was fitted with the systematic component $\eta_{ij} = \beta_0 + \beta_1\ x_{1ij} + \beta_2\ x_{2ij} + v_i$, where β_0 was the intercept, β_1 was the treatment effect, x_1 was generated from normal distribution, β_2 was an extra parameter, and x_2 was the second treatment effect generated from the Poisson distribution with mean = 3, $x_2 \sim \mathcal{P}(\lambda = 3)$. Power was estimated as proportion of correct detection of significance for β_1, while Type I error rate was estimated as proportion of incorrect detection of significance for β_2.

Restricted Pseudo Likelihood

The REPL HGLM in Chapter II under the heading Restricted Pseudo Likelihood Estimation was described

$$\textbf{1. } \boldsymbol{Y_{ij}|u \sim D(\mu, \phi\, V(\mu))},$$

$$\boldsymbol{u_i \sim \mathcal{N}(0, V_R)},$$

$$\textbf{2. } \boldsymbol{\eta = X\beta + Zu},$$

$$\textbf{3. } \boldsymbol{\eta} = ln(\boldsymbol{\mu}).$$

The systematic component applied for generating data was

$$\eta_{ij} = 1 + 0.2 \times x_{1ij} + v_i,$$

and the systematic component for the fit model was

$$\eta_{ij} = \beta_0 + \beta_1\ x_{1ij} + \beta_2\ x_{2ij} + v_i,$$

where $v_i \sim Beta(2, 3)$. For Restricted Pseudo Likelihood, the researcher wrote code in R to produce the iterative weighted least squares (IWLS) algorithm to estimate the mean parameters β and v, and the dispersion parameter ϕ. R code was in Appendix B and Appendix E, section Restricted Pseudo Likelihood. Table 3 summarized the averages of β_1 and β_2, power of the hypothesis test for β_1, Type I error rate of the hypothesis test for β_2, and standard error for β_1 for the REPL method.

Table 3 demonstrated that REPL was a good estimate method, since the average of 1,000 replications gave estimates that were very close to actual value, which was 0.2, and $\hat{\beta}_2$ was close to zero. The power of the hypothesis test for β_1

Table 3

Restricted Pseudo Likelihood

Clusters	$\bar{n_i}$	$\hat{\beta}_1$	$\hat{\beta}_2$	Power	Type I error	$S.E_{\hat{\beta}_1}$
	10	0.2075701	-0.005259478	0.972	0.049	0.05519608
K = 20	25	0.2036177	-0.003936949	1	0.055	0.03315632
	100	0.2016445	0.0005931241	1	0.038	0.01646315
	10	0.2041978	0.00357477	1	0.016	0.02605315
K = 50	25	0.2024797	0.006654026	1	0.045	0.01623582
	100	0.2002964	0.001345378	1	0.034	0.008043962

was high since the sample size was large for each of the combinations, and the Type I error rate for the hypothesis test for β_2 was acceptable because it was close to 0.05. The standard error for β_1 was small and fits in the range from 0.0080 to 0.055.

From Figures 1 (for power), 2 (for Type I error rate), and 3 (for the standard error), the REPL method was shown to work better for a large number of clusters. Figures showed that, for $K = 50$, REPL had smaller values for Type I error rate and standard error. As such, REPL method for $K = 50$ was better than $K = 20$ for an unbalanced cluster size with binary outcomes. A comparison of REPL method with others was made in the last section.

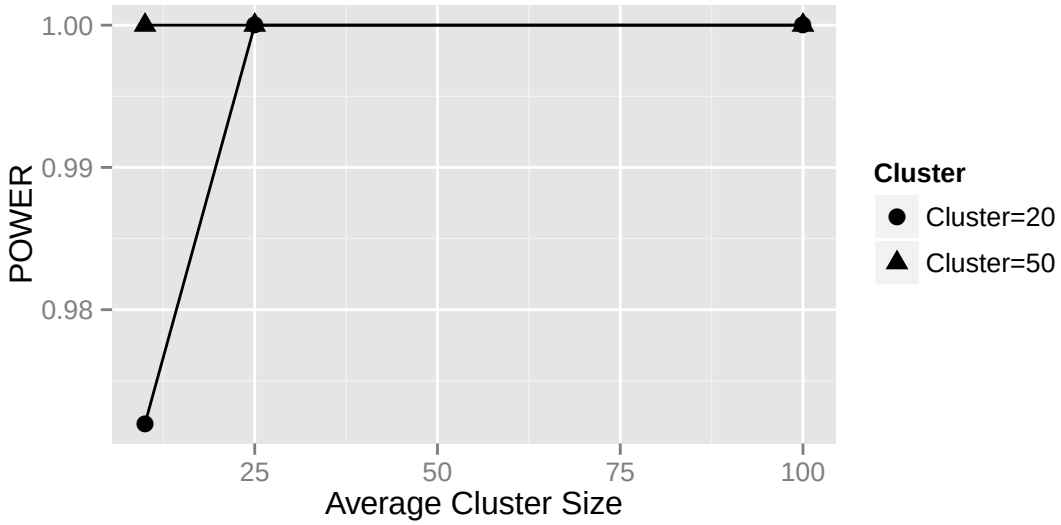

Figure 1. Power for $\hat{\beta}_1$

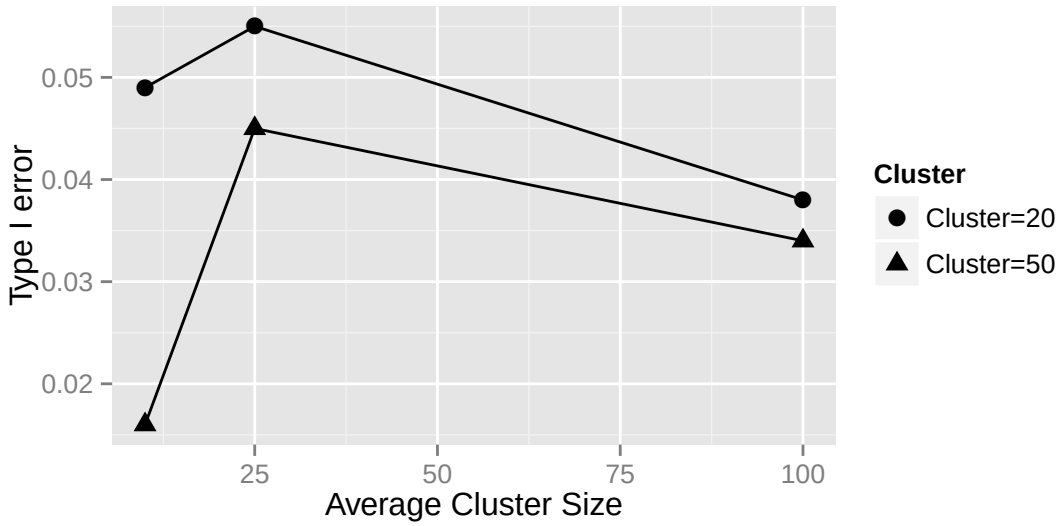

Figure 2. Type I error for $\hat{\beta}_2$

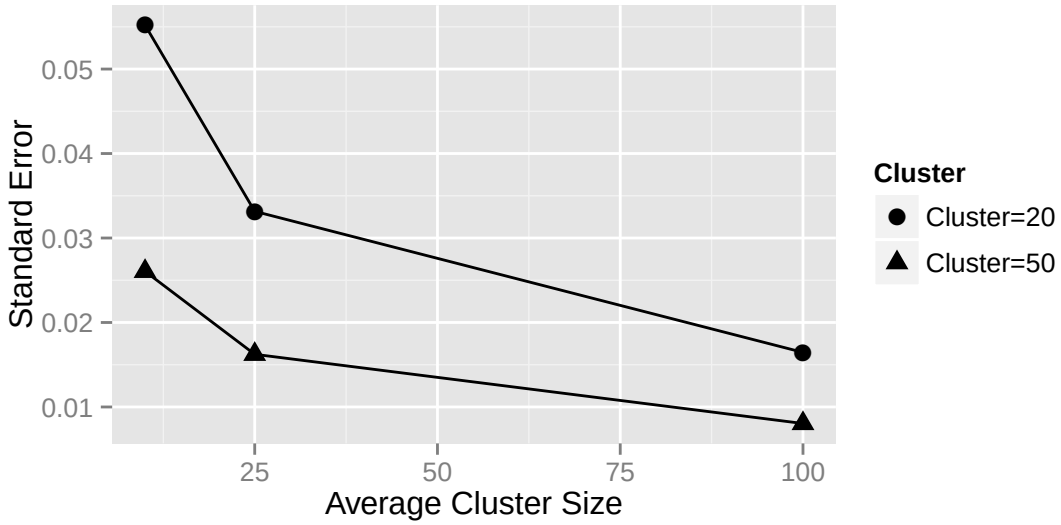

Figure 3. Standard error for $\hat{\beta}_1$

Extended Restricted Pseudo Likelihood

The HGLM for EREPL in Chapter III under the heading extended restricted pseudo likelihood was described

$$\textbf{1. } \boldsymbol{Y_{ij}|u \sim D(\mu, \phi_i\ V(\mu))},$$

$$\boldsymbol{u_i \sim \mathcal{N}(0, V_R)},$$

$$\textbf{2. } \boldsymbol{\eta = X\beta + Zu},$$

$$\textbf{3. } \boldsymbol{\eta} = ln(\boldsymbol{\mu}).$$

The systematic component applied for generating data was

$$\eta_{ij} = 1 + 0.2 \times x_{1ij} + v_i,$$

and the systematic component for the fit model was

$$\eta_{ij} = \beta_0 + \beta_1\ x_{1ij} + \beta_2\ x_{2ij} + v_i.$$

where $v_i \sim Beta(2,3)$. The extended REPL method allowed for different dispersion multipliers ϕ_i for each cluster. Unfortunately, the program did not converge, because the values of $\hat{\beta}_1$ oscillated.

Figures 4 to 7 showed the divergence of the $\hat{\beta}_1$ value. Figures 4 and 5 showed the case of $K = 20$ clusters, with an average cluster size of $\bar{n} = 100$. Figure 4 showed the oscillating values of $\hat{\beta}_1$ before it reached the divergence point, and Figure 5 showed the oscillating values of $\hat{\beta}_1$ as it diverged. Figures 6 and 7 showed the case of $K = 50$ clusters, with size of $\bar{n} = 10$. Figure 6 showed the oscillating values of $\hat{\beta}_1$ before it reaches the divergence point, and Figure 7 showed $\hat{\beta}_1$ at divergence.

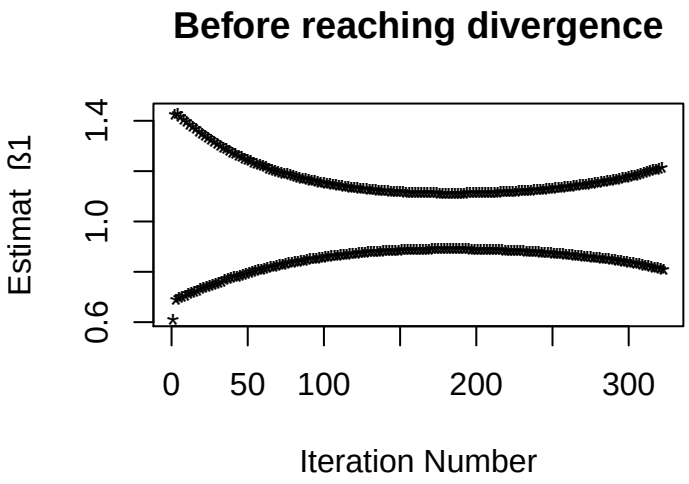

Figure 4. $\hat{\beta}_1$ before reach divergent point for K = 20 and $\bar{n} = 100$

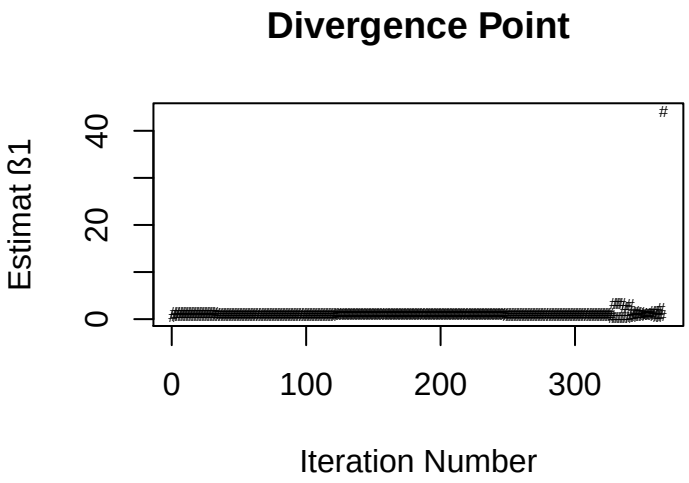

Figure 5. $\hat{\beta}_1$ at divergence point for K = 20 and $\bar{n} = 100$

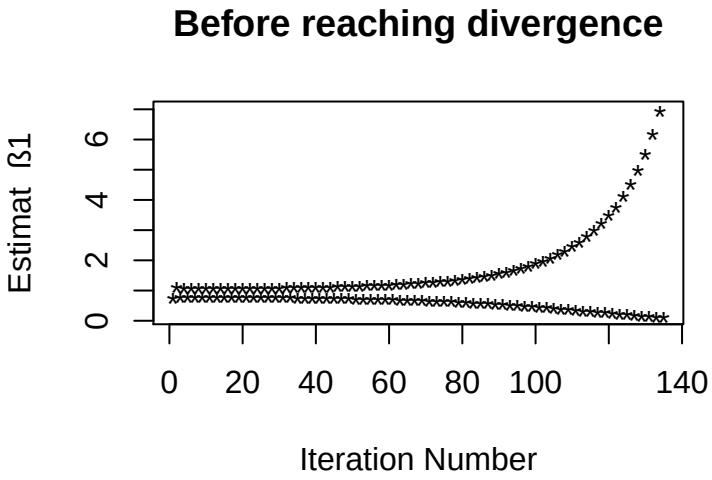

Figure 6. $\hat{\beta}_1$ at divergence point for number in cluster = 50 and $\bar{n} = 10$

From the Figures, it was clear that $\hat{\beta}_1$ oscillates, dramatically increasing then suddenly jumping to a very far single point, which was shown in the Figures 5, and 7. The process does not converge. R code was in Appendix C.

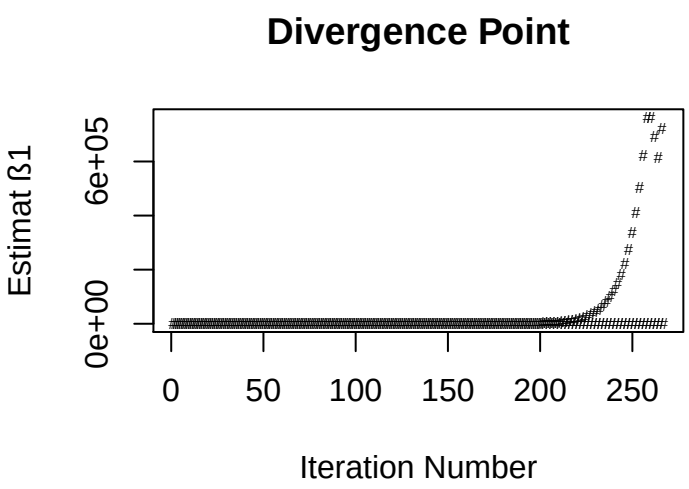

Figure 7. $\hat{\beta}_1$ at divergence point for K = 50 and $\bar{n} = 10$

Binomial Beta h-likelihood

The HGLM for the Binomial Beta model in Chapter II under section Hierarchical Likelihood Estimation, was described

$$\textbf{1. } \boldsymbol{Y_{ij}|u \sim Bin(\mu,\ \mu(1-\mu))},$$

$$\boldsymbol{u_i \sim Beta(\gamma, \lambda)},$$

$$\textbf{2. } \boldsymbol{\eta = X\beta + Zu},$$

$$\textbf{3. } \boldsymbol{\eta} = ln(\boldsymbol{\mu}).$$

The systematic component applied for generating data was

$$\eta_{ij} = 1 + 0.2 \times x_{1ij} + v_i,$$

and the systematic component for the fit model was

$$\eta_{ij} = \beta_0 + \beta_1\ x_{1ij} + \beta_2\ x_{2ij} + v_i,$$

where $v_i \sim Beta(2,3)$. For the Binomial Beta h-likelihood, the researcher used the HGLM function in the HGLM package in R. Using the hglm function got the estimation for parameters β and t-statistics with the p-values. Through simulation, the average of 1,000 estimates was calculated for β_1, β_2, power of the hypothesis test for β_1, Type I error of the hypothesis test for β_2, and standard error for β_1. (See R code was in Appendix E, section Binomial Beta h-Likelihood). Table 4 showed that Binomial Beta h-likelihood was a good estimate method, with estimated values close to true parameters. The power of β_1 was high, the Type I error rate for β_2 was somewhat high, with values ranging from 0.07 to 0.143. This may have been due to ignoring overdispersion caused by different cluster sizes. The standard error for β_1 had small values for largest sample sizes, the standard error values ranging from 0.009 to 0.047. A comparison of the Binomial Beta h-likelihood method with others made in the last section.

Table 4

Binomial Beta h-likelihood

Clusters	$\bar{n_i}$	$\hat{\beta}_1$	$\hat{\beta}_2$	Power	Type I error	$S.E_{\hat{\beta}_1}$
	10	0.2113867	-0.009203517	1	0.143	0.04729659
K = 20	25	0.2020606	0.005317432	1	0.096	0.02872977
	100	0.2010578	0.003415107	1	0.107	0.01431681
	10	0.2084046	0.007679551	1	0.092	0.02909505
K = 50	25	0.2031552	0.004931511	1	0.07	0.01813028
	100	0.1988225	0.002102863	1	0.091	0.009000959

Figures 8 to 10 showed, respectively, the power of the hypothesis test for β_1, the Type I error rate of the hypothesis test for β_2, and the standard error for β_1 for the Binomial Beta h-likelihood method for different cluster sizes. From the Figures when $K = 50$, Binomial Beta had smaller values for the Type I error rate and smaller values for standard error. A comparison of the Binomial Beta method with others made in the last section.

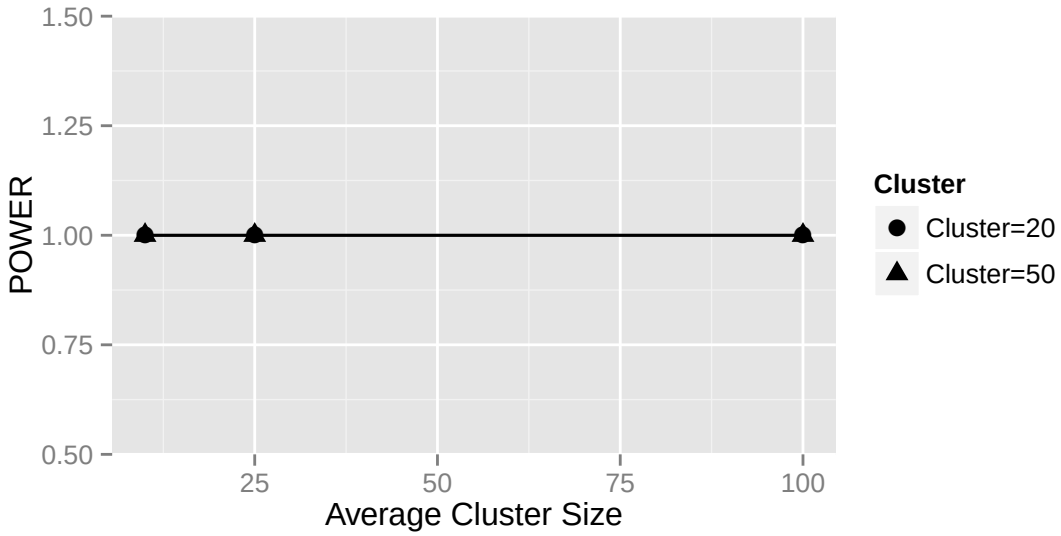

Figure 8. Power for $\hat{\beta}_1$

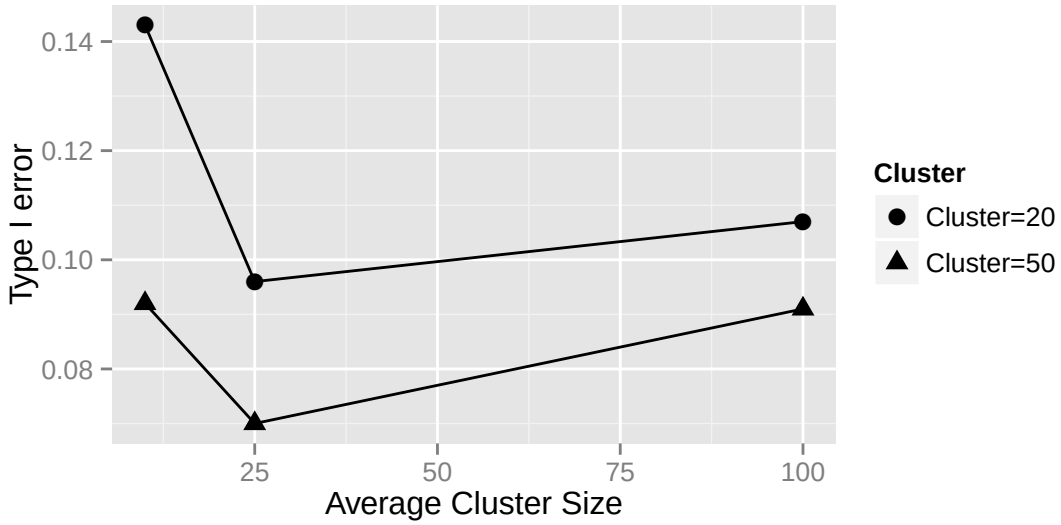

Figure 9. Type I error for $\hat{\beta}_2$

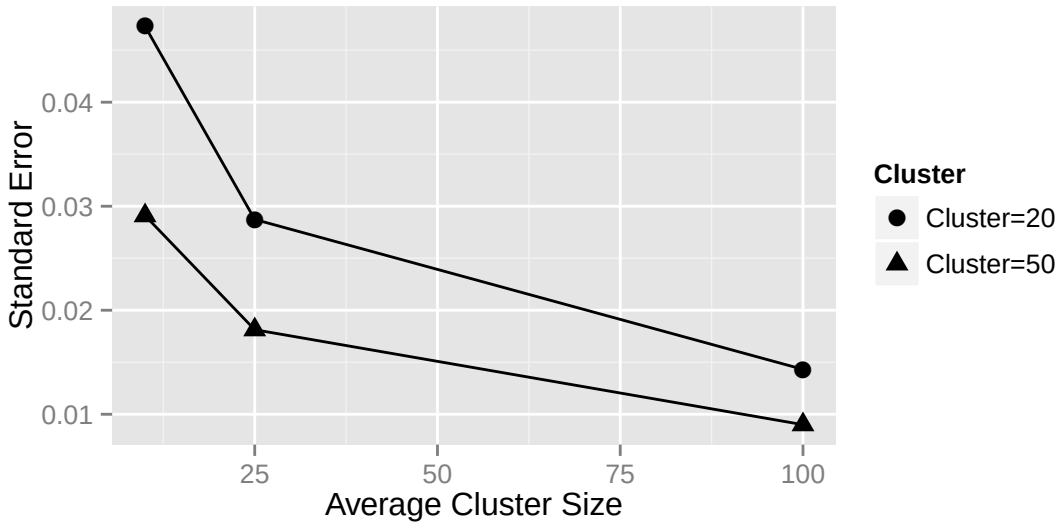

Figure 10. Standard error for $\hat{\beta}_1$

Adjusted Scale Binomial Beta h-likelihood

The HGLM in Chapter III under section Adjusted Scale Binomial Beta, was described

$$\textbf{1. } \boldsymbol{Y_{ij}|u \sim Bin(\mu,\ \mu(1-\mu))},$$

$$\boldsymbol{u_i \sim Beta(\gamma, \lambda_i)},$$

$$\textbf{2. } \boldsymbol{\eta = X\beta + Zu},$$

$$\textbf{3. } \boldsymbol{\eta} = ln(\boldsymbol{\mu}).$$

The systematic component applied for generating data was

$$\eta_{ij} = 1 + 0.2 \times x_{1ij} + v_i,$$

and the systematic component for the fit model was

$$\eta_{ij} = \beta_0 + \beta_1\ x_{1ij} + \beta_2\ x_{2ij} + v_i,$$

where $v_i \sim Beta(2, 3)$. The adjusted h-likelihood used to obtain estimates using a random effect with beta distributions with different scale parameters to account for overdispersion due to differing cluster sizes. For Adjusted Scale Binomial Beta h-likelihood, the researcher wrote the adjusted h-likelihood function. The estimates for the mean parameters $\boldsymbol{\beta}$, along with the t-test statistics and p-values, were obtained through maximum h-likelihood estimation using the maxLik function in the maxLik package in R (Henningsen & Toomet, 2011). The code was in the Appendix D and Appendix E, section Adjusted Scale Binomial Beta h-Likelihood. Table 5 demonstrated that Adjusted Scale Binomial Beta h-likelihood was a good estimate method, with estimated values close to true parameter values. The power of

the hypothesis test for β_1 was high with value equal to one, Type I error of the hypothesis test for β_2 was acceptable with value ranging from 0.054 to 0.085. In fact Adjusted Scale Binomial Beta was better than Binomial Beta h-likelihood because it accounted for overdispersion due to different cluster sizes. The standard error for β_1 showed that there was small variability of the parameter estimates, with values from 0.01 to 0.05, which were small values for the large sample sizes.

Table 5

Adjusted Scale Binomial Beta h-likelihood

Clusters	$\bar{n_i}$	$\hat{\beta}_1$	$\hat{\beta}_2$	Power	type I error	S.E
	10	0.2173841	0.004827131	0.992	0.058	0.05579434
K = 20	25	0.21352	0.001662735	1	0.054	0.03393393
	100	0.2136255	0.003884209	1	0.071	0.0169782
	10	0.217621	0.01406111	1	0.057	0.03438107
K = 50	25	0.2182764	0.006173511	1	0.063	0.02149756
	100	0.2134524	0.002064118	1	0.085	0.01066414

Figures 11 to 13 showed the power, Type I error rate, and standard error for Adjusted Scale Binomial Beta for different cluster sizes.

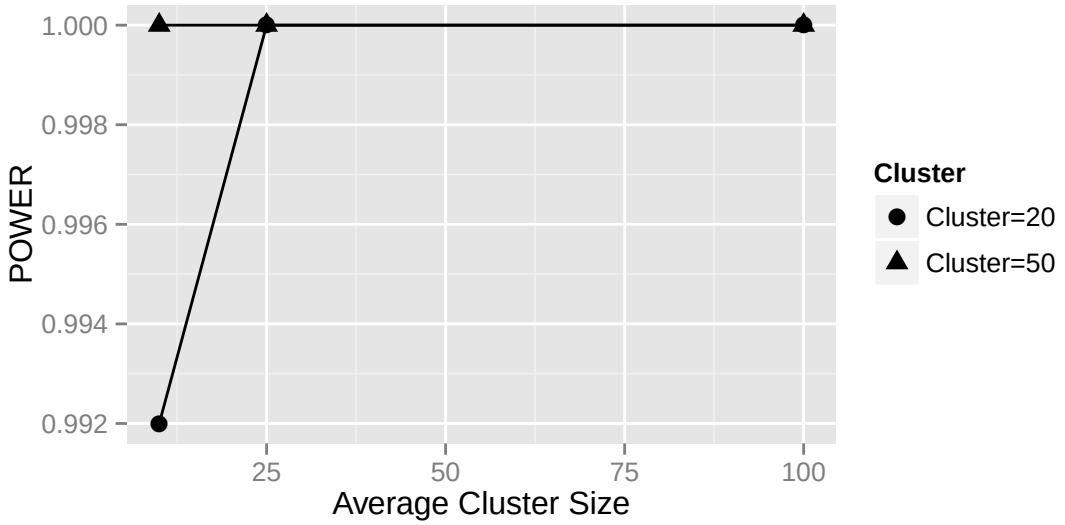

Figure 11. Power for $\hat{\beta}_1$

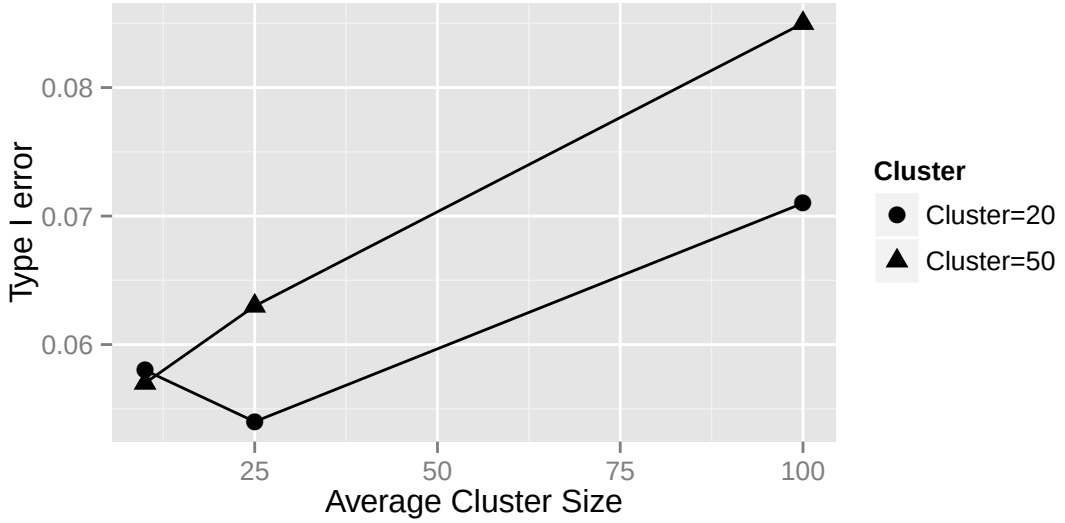

Figure 12. Type I error for $\hat{\beta}_2$

Figures showed that Type I error rate was small, and the standard error was large when cluster size was equal to 20. The Adjusted Scale Binomial Beta h-likelihood worked well, especially since Type I errors occurred at an acceptable rate. This means that the Adjusted Scale Binomial Beta h-likelihood accounted for

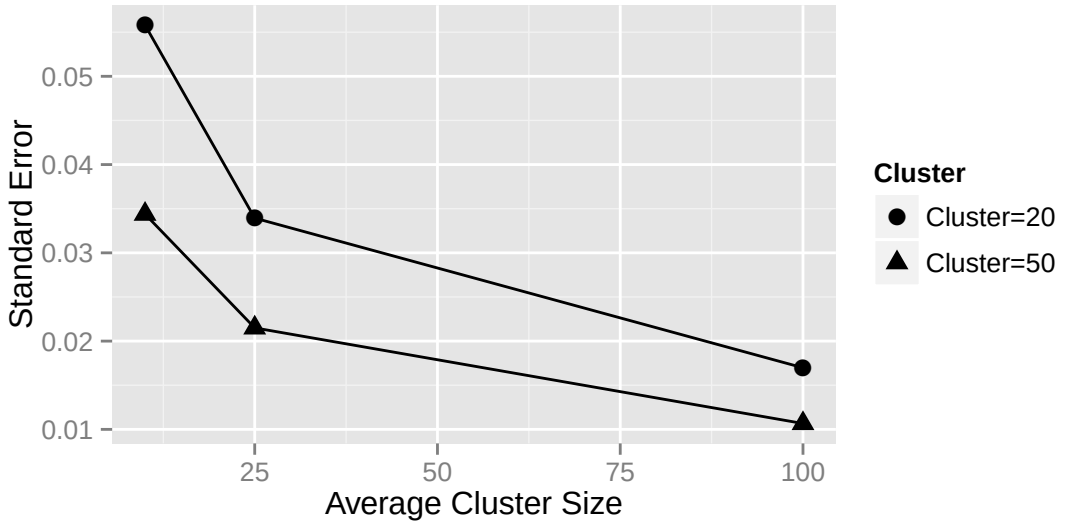

Figure 13. Standard error for $\hat{\beta}_1$

different dispersions across clusters. As such, it was a suitable method for unbalanced cluster sizes with binary outcomes.

Comparison of Methods

Unbalanced data could have been defined as an unequal number of data units within K clusters. Cluster sizes were randomly generated from poisson distributions with means of 10, 25, and 100. This meant that the number of observations for each combination was large, with approximately 200 responses for each combination. The sample size may have effected the power for each method, since greater sample sizes causes higher power.

All the simulations were conducted as specified previously. The Gaussian quadrature approximation algorithm successfully converged in three methods: Restricted Pseudo Likelihood, Binomial Beta h-likelihood, and Adjusted Scale Bino-

mial Beta h-likelihood; however, the algorithm did not converge for Extended Restricted Pseudo Likelihood.

For the three methods, statistical power, Type I error rate, and standard error were displayed in Tables 6 to 8, respectively, and the results summarized.

Statistical Power

Statistical power was computed as the proportion of correct rejections of the hypothesis $H_0 : \beta_1 = 0$. Through simulation, the test was conducted 1,000 times to see how often the test was significant. The power was the proportion of those 1,000 tests rejected correctly. As shown in Table 6, it was hard to decide which method performed better since the power was one and was high for all methods because the sample size was large for each combination. There were no differences among the three methods in power, so all methods worked well using power as a criterion. Figures 14, and 15 compare the three methods with $K = (20, 50)$, and they showed the close results for the three methods. Figures demonstrate that the power was high (very close to one) because of large sample sizes.

Table 6

Statistical Power for β_1

Clusters	$\bar{n}_i$	REPL	Binomial-Beta	Adjusted Scale Binomial-Beta
	10	0.972	1	0.992
K = 20	25	1	1	1
	100	1	1	1
	10	1	1	1
K = 50	25	1	1	1
	100	1	1	1

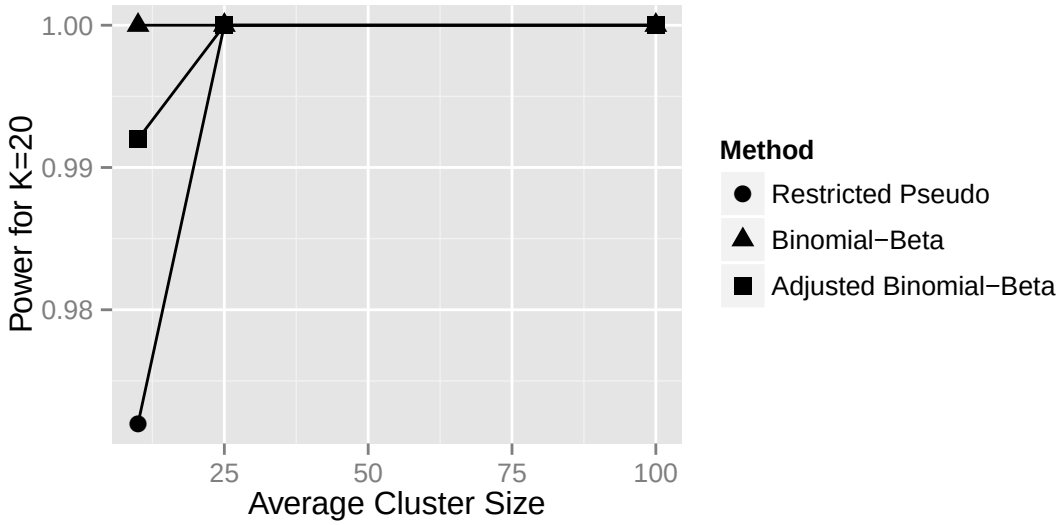

Figure 14. Power for all methods with K = 20

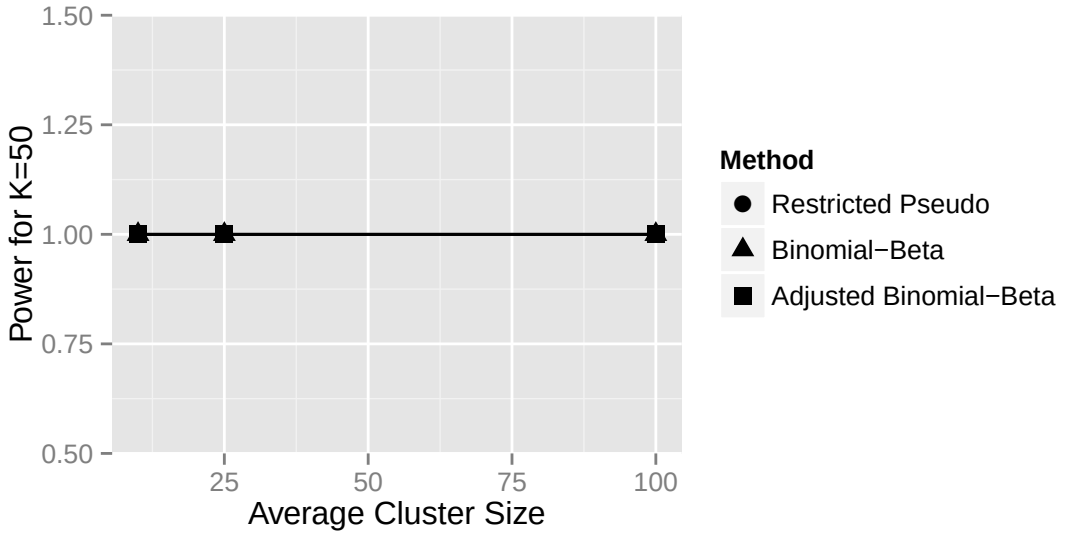

Figure 15. Power for all methods with K = 50

Type I Error Rate

Type I error rates were computed as the proportion of p values less than 0.05 under a null hypothesis $H_0 : \beta_2 = 0$. Ideally, Type I error rate should be close to 0.05. As shown in Table 7, the Adjusted Scale Binomial Beta was better than the Binomial Beta h-likelihood, in the sense that Type I error rate was closer to 0.05. Because Adjusted Scale Binomial Beta h-likelihood accounted for the overdispersion caused by unequal cluster sizes, it showed better results than h-likelihood with regard to Type I error. For REPL, the method seemed to have acceptable Type I error rate and fit in the range from 0.016 to 0.055.

Figures 16, and 17 display the difference between the three methods with $K = 20$, $K = 50$ for Type I error.

Table 7

Type I Error Rate

Clusters	$\bar{n}_i$	REPL	Binomial-Beta	Adjusted Scale Binomial-Beta
	10	0.049	0.143	0.058
K = 20	25	0.055	0.096	0.054
	100	0.038	0.107	0.071
	10	0.016	0.092	0.057
K = 50	25	0.045	0.07	0.063
	100	0.034	0.091	0.085

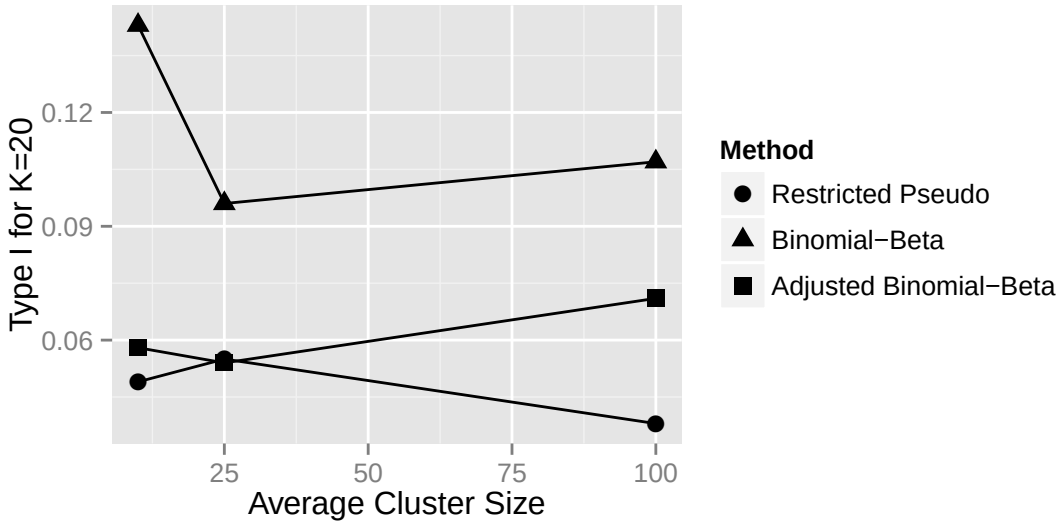

Figure 16. Type I error rate for all methods with K = 20

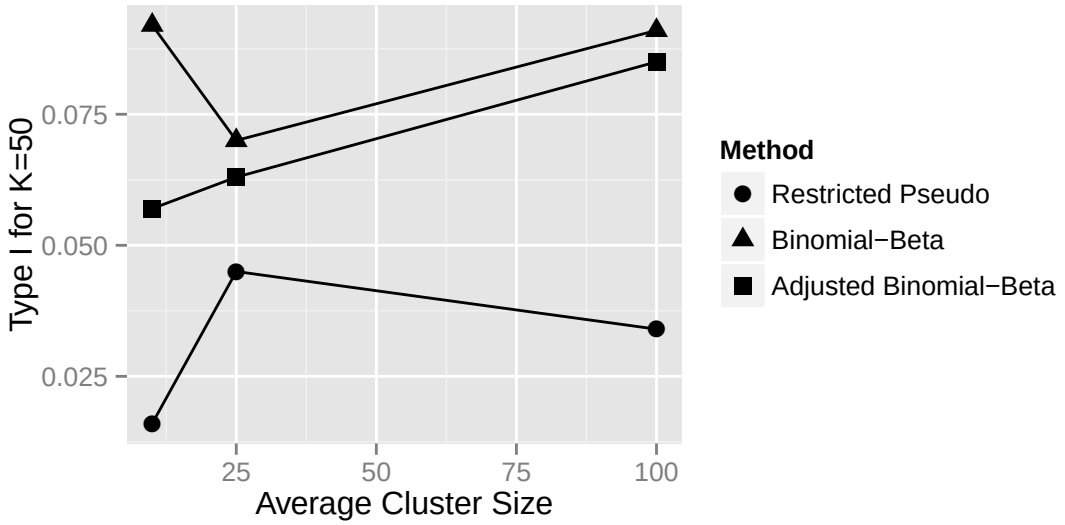

Figure 17. Type I error rate for all methods with K = 50

The REPL method had a smaller Type I error rate but that did not mean it was the best method. A Type I error rate smaller than 0.05 typically means lower power, since as the Type I error rate decreases power also decreases. In our study, because the sample sizes were large, power was universally high.

Standard Error

The $\bar{SE}$ was computed as the average of 1,000 SE of the estimates of β_1. Smaller $\bar{SE}$ represented smaller estimated variability, or greater precision, of the parameter estimates,(Heo & Leon, 2005). The standard error for $\hat{\boldsymbol{\beta}}$ indicated whether or not the efficiency improved. From Table 8, the Binomial Beta h-likelihood showed the smallest standard errors as compared to the other methods in all combinations. However, Binomial Beta also showed the highest Type I error rate as a consequence due to ignoring to account for different dispersions.

Table 8

Standard Error for β_1

Clusters	$\bar{n_i}$	REPL	Binomial-Beta	Adjusted scale Binomial-Beta
	10	0.05519608	0.04729659	0.05579434
K = 20	25	0.03315632	0.02872977	0.03393393
	100	0.01646315	0.01431681	0.0169782
	10	0.02605315	0.02909505	0.03438107
K = 50	25	0.01623582	0.01813028	0.02149756
	100	0.008043962	0.009000959	0.01066414

Figures 18, and 19 compare the standard errors for the three methods for different cluster sizes.

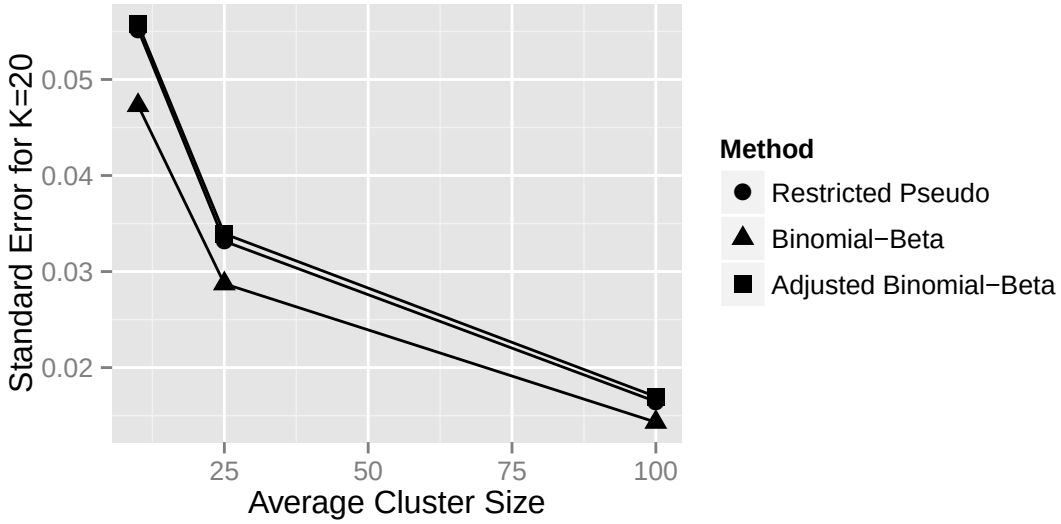

Figure 18. Standard Error for all methods with K = 20

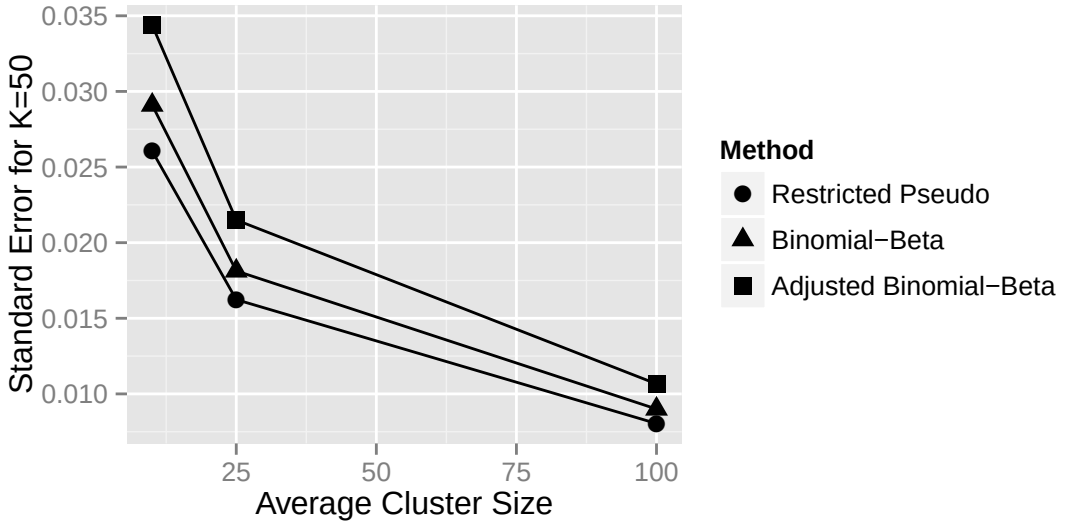

Figure 19. Standard Error for all methods with K = 50

From Figures the results were somewhat close, with acceptable standard error value for all three methods.

Overall Comparison

From the previous sections, all three methods were good estimate methods for mean parameters with estimate values close to actual parameters, and all showed improvement for large sample sizes. It was good to know that the Adjusted Scale Binomial Beta h-likelihood was a suitable method for binary outcomes because it had a small Type I error rate. Also Adjusted Scale Binomial Beta appears to be a good estimate method and showed power and standard error close to other methods. The Type I error rate for Adjusted Scale Binomial Beta h-likelihood increased as sample size increased because large sample size led to small standard error, which caused an increase in Type I error rate. My suggestion would be to try this study with small sample sizes to see if the power and Type I error rate

changed. The Extended Restricted Pseudo Likelihood failed to converge since $\hat{\beta}_1$ oscillated. It was not able to get the estimation parameter values and other statistics measured to compare it with other methods. The Binomial Beta h-likelihood had a high Type I error rate since it did not account for different dispersions due to different cluster sizes. The Type I error rate had inflated in the Binomial Beta method. It may be that with data that had overdispersion or had variability, the Adjusted Scale Binomial Beta would give a better estimate than other methods.

CHAPTER V

SUMMARY AND FURTHER RESEARCH

Summary

Unbalanced data with binary outcomes were quite common in practice. Unbalanced data suggested the use of heterogeneous models, as demonstrated in previous studies with continuous outcomes. In this study, the researcher used a mixed effects generalized linear model containing fixed and random factors with binary outcomes, or a Hierarchical Generalized Linear Model (HGLM). The researcher used the Adjusted Scale Binomial Beta h-likelihood to account for overdispersion caused by different cluster sizes.

In this work, the researcher evaluated the performance of estimation methods using power, Type I error rate, and standard error. High power was required in methods, at the same time with acceptable Type I error. Without accounting for overdispersion, Type I error rate could be inflated. The standard error was a measure efficiency. Smaller standard error represented smaller variability, or greater precision (Heo & Leon, 2005). The conclusions from methods discussed in Chapter IV follow.

Restricted Pseudo Likelihood was a good estimate method, since the average of 1,000 replications gave estimates that were very close to actual values.

The power of the hypothesis test for regression parameters was close to one, and the Type I error rate for the hypothesis test for regression parameters was acceptable because it was close to 0.05. The standard error for regression parameters was small and fits in the range from 0.0080 to 0.055. The REPL show a good estimation for binary data with unbalanced clusters, (Geys et al., 1997) showed the Restricted Pseudo Likelihood estimation was a very useful estimation in clustered data with non-continuous response.

For Extended Restricted Pseudo Likelihood, the process does not converge when $\hat{\beta}_1$ oscillates, dramatically increasing then suddenly jumping to a very far single point. By trying to understand the $\hat{\beta}_1$ behavior, it would be a good extension to this method.

Binomial Beta h-likelihood was a good estimate method, with estimated values close to true parameters. The power of β_1 was close to one, the Type I error rate for β_2 was somewhat high. This may be due to ignoring overdispersion caused by different cluster sizes. The standard error for β_1 had small values ranging from 0.009 to 0.047. Even Binomial Beta h-likelihood method had a small values of standard error, did not mean it was a correct values. It may not have been appropriate.

Adjusted Scale Binomial Beta h-likelihood was a good estimate method, with estimated values close to true parameter values. The power of the hypothesis test for β_1 was equal to one, Type I error of the hypothesis test for β_2 was acceptable with value ranging from 0.054 to 0.085. In fact Adjusted Scale Binomial Beta was better than Binomial Beta h-likelihood because it accounts for overdispersion

due to different cluster sizes. The standard error for β_1 shows that there was small variability of the parameter estimates, with values from 0.01 to 0.05.
From the graphs in the figures the conclusions for comparing the converging methods follow.

1. For the statistical power graphs, all methods showed a high power since the sample size was large for each simulation.

2. For the Type I error rates graphs, there was a strange trend behavior. The type I error rate was first decreasing with increasing sample size, then was increasing with increasing sample size.

3. The Standard Error graphs showed decreasing average of standard error with increasing sample size.

The results from the simulation demonstrated that the capability of the Adjusted Scale Binomial Beta h-likelihood was comparable to existing methods, as it gave us a low standard error and acceptable Type I error. Moreover, Binomial Beta h-likelihood had inflated Type I error. Therefore, the results suggested that the Adjusted Scale Binomial Beta h-likelihood method should be an option in computer statistical programs to analyze unbalanced clustered data with binary outcomes. The Restricted Pseudo Likelihood can also be applied to unbalanced clustered binary data

Directions for Further Research

Below are some suggestions for future studies based on what was obtained from this project.

1. Since the Extended Restricted Pseudo Likelihood did not converge, it would be a good idea to adjust this method or apply this method in another program.

2. It would be a good idea to repeat this study with small sample sizes and compare the results with this study's results.

3. According to previous studies, unbalanced clustered data may have led to loss of efficiency. In this dissertation, the researcher focused on unbalanced clustered data with binary outcomes, which followed a binomial distribution. Instead of using binary outcomes, future research should include another type of dependent variable with unbalanced clustered data.

4. Since the Type I error rate graphs showed strange trend behavior, it would be informative to try small numbers of clusters, $K = 5$ and $K = 30$, with the same average cluster size ($\bar{n} = 10, 25, 100$) and evaluate the Type I error rate. It would be worth try a large number of clusters to see the difference.

5. Finally, it may be worthwhile to apply the double extended quasi-likelihood with binary outcomes.

LIST OF REFERENCES

Abdoslam, I. N. (2004). Messy data in heteroscedastic models case study: Mixed nested design. M.Sc. THESIS.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Bauer, J. J. (2009). A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes. *PSYCHOMETRIKA, 74*, 97-105.

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods, 16*(4), 373-390.

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

Breukelen, G. J., & Candel, M. J. (2012). Comments on effciency loss because of varying cluster size in cluster randomized trials is smaller than literature suggests. *Statistics in Medicine, 31*, 397-400.

Candel, M. J. J. M., & Breukelen, G. J. P. V. (2010). Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Statistics in Medicine, 29*, 1488-1501.

Dai, J., L. Z. R. D. (2006). Hierarchical logistic regression modeling with SAS glimmix. *Proceedings of the 14th Annual Conference, WUSS*, Irvine, CA.

Fisher, R. (1925). *Statistical methods for research workers*(1st ed.). London: Oliver and Boyd.

Fitzmaurice, G., N. L. & Ware, J. (2004). *Applied longitudinal analysis*. NY: Wiley.

Green, P., & Silverman, B. (1994a). Bias correction in generalized linear mixed models with multiple components of dispersion.*the American Statistical Association, 91*, 1007-1016.

Green, P., & Silverman, B. (1994b). *Nonparametric regression and generalized linear models*. London: Chapman & Hall.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review, 55*, 245-259.

Gu, Z. (2008). Model diagnostics for generalized linear mixed models. Dissertations.

Hartley, H., & Rao, J. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika, 54*, 93-108.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50*, 933-944.

Geys, H., Molenberghs, G., & Ryan, L. (1997). Pseudo-likelihood inference for clustered binary data. *COMMUN STATIST-THEORY METH, 26*(11), 2743-2767.

Henningsen, A., & Toomet, O. (2011). Maxlik: A package for maximum likelihood-estimation in R. *Comput Stat, 26*, 443-458.

Heo, M., & Leon, A. (2005). Performance of a mixed effects logistic regression model for binary outcomes with unequal cluster size. *Biopharmaceutical Statistics,15*, 513-526.

Jang, W., & Lim, J. (2006). PQL estimation biases in generalized linear mixed models. *Institute of Statistics and Decision Sciences, Duke University, Durham, NC, USA.*, 05-21.

Lalonde, T. L. (2009). Components of overdispersion in hierarchical generalized linear models. Dissertations.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B (Methodological), 58*(4), 619-678.

Lee, Y., & Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika, 88*(4), 987-1006.

Lee, Y., & Nelder, J. A. (2006). Double hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B (Methodological), 55*, 139-185.

Lee, Y., & Nelder, J. A. (2009). Likelihood inference for models with unobservables: Another view. *Statistical Science, 24*(3), 255-269.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2006). *Generalized linear models with random effects*(1st ed.). Boca Raton: Chapman & Hall / CRC.

McCullagh, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models.* NY: John Wiley & Sons, Inc.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*(2nd ed.). London: Chapman & Hall.

Nelder, J. A., & Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood: Some comparisons. *Journal of the Royal Statistical Society, Series B (Methodological), 54*(1), 273-284.

Neuhaus, J. M., & Lesperance, M. L. (1996). Estimation effciency in a binary mixed effects model setting. *Biometrika, 83*, 441-446.

Pregibon, J. A. N. D. (1987). An extended quasi-likelihood function. *Biometrika, 74*(2), 221-232.

Searle, S. R. (1987). *Linear models for unbalanced data*. NY: John Wiley & Sons, Inc.

Wang, J. L. (2010). Supplement to gee analysis of clustered binary data with diverging number of covariates. Submitted to the Annals of Statistics.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika, 61*(3), 439-447.

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation,48*(3), 233-243.

APPENDIX A

R-CODE FOR GENERATING DATA

```
mydata=function(seed){
set.seed(seed)
  k            < -  20
  beta0     <-  1
  beta1     <-  0.2
  beta2     <-  3.1
sigma2     <-  sqrt(20)
n             <- rpois(k,10)
z=matrix(0,sum(n),k)
y=matrix(0,sum(n),1)
x=matrix(c(rep(1,sum(n)),rep(0,sum(n)),rep(0,sum(n))),sum(n),3)

     u1=as.matrix(rep(rbeta(k,2,3),n),n,1)
     u=as.matrix(rep(rbeta(k,2,3),1),n,1)
index=1
 for (i in 1:k)
{
    z[(index:(index+n[i]-1)),i]= rep(1,n[i])
    index=index+n[i]
}
   id=as.matrix(rep(1:k,n), n,1)

 ## GENERATE X-VALUES ##

       x[,2] =rnorm(sum(n),3,sigma1)
       x[,3]=rpois(sum(n),3)

         linpred=beta0+beta1*x[,2]+z%*%u
         expit=exp(linpred)/(1+exp(linpred))

   ## GENERATE RESPONSE VALUES ##
       y[,1]= rbinom(sum(n),1,expit)
     dat=list( x=x, y=y, z=z,id=id)

}
```

APPENDIX B

RESTRICTED PSEUDO LIKELIHOOD FUNCTION

```
Pseude=function (x,y,z, beta0.cell=0 , phi=1,conv_crit=1e-8,n_maxiter=1000)
{
    N          <-   length(y)
    A          <-   diag(ncol(z)) ##random variance
    beta10   <-   rep(beta0.cell,ncol(x))
    u10       <-   rep(beta0.cell,ncol(z))
    phi10    <-   phi   ## phi for all desin###

  eta        <-   x%*%beta10+z%*%u10
  mu         <-   exp(eta)/(1+exp(eta))

  D          <- diag(as.numeric(1/(mu*(1-mu))))  ## Partial derivative for eta ###
  V          <- diag(as.numeric(mu*(1-mu)))  ### V(m)=m(1-m)
  Vp        <- (z%*%A%*%t(z))+phi10 * D  ### Variance Psuedo
  P          <-  t(D )%*%(y-mu)+  eta  ### Linearization "Psudeo" ###

 betaHat     <-   ginv(t(x) %*% ginv(Vp) %*% x) %*% t(x)%*%ginv(Vp)%*%P
 r              <-   P-(x%*%betaHat)  ### residuale
 uHat          <-   A%*%t(z)%*%ginv(Vp)%*%r
 phiHat       <-  (1/N)*(t(r)%*%ginv(Vp)%*%r)

  beta1       <- betaHat[1:ncol(x)]
  u1           <- uHat[1:ncol(z)]
  phi1         <- phiHat

  d      <-  max(abs(beta1-beta10), abs(u1 - u10 ),abs( phi1 -  phi10 ) )

  if(d<conv_crit) {conv<-T} else{conv <- F}

 n <- 1

while(n<=n_maxiter & d>=conv_crit){

  beta10    <- as.numeric(beta1)
  u10       <- as.numeric(u1)
  phi10     <- as.numeric(phi1)
```

```
eta   <-  x%*%beta10+z%*%u10
 mu    <-  exp(eta)/(1+exp(eta))
 D     <-  diag(as.numeric(1/(mu*(1-mu))))  ## Partial derivative for eta ###
 V     <-  diag(as.numeric(mu*(1-mu)))  ### V(m)=m(1-m)
 Vp    <-  (z%*%A%*%t(z))+ phi10 *D   ### Variance Psuedo
 P     <-  t(D )%*%(y-mu)+  eta  ### Linearization "Psudeo" ###

 betaHat <-ginv(t(x) %*% ginv(Vp) %*% x) %*% t(x)%*%ginv(Vp)%*%P
 r     <-  P-(x%*%betaHat)  ### residuale
 uHat   <-  A%*%t(z)%*%ginv(Vp)%*%r
 phiHat  <- (1/N)*(t(r)%*%ginv(Vp)%*%r)

 beta1  <- betaHat[1:ncol(x)]
 u1    <- uHat[1:ncol(z)]
 phi1   <-  phiHat

 d     <-max(abs(beta1-beta10), abs(u1 -u10 ),abs( phi1 -  phi10 ))

 n     <- n+1

}

if(d<conv_crit) {conv<-T} else{conv <- F}
if(conv==T){

d2beta11 <-  (1/phiHat)* (t(x)%*%ginv(D)%*%x)
d2beta22 <-  (1/phiHat)* (t(z)%*%ginv(D)%*%z)+ A
d2beta12 <-  (1/phiHat)* (t(x)%*%ginv(D)%*%z)
d2beta21 <-  (1/phiHat)* (t(z)%*%ginv(D)%*%x)
H       <- rbind(cbind(d2beta11,d2beta12),cbind(d2beta21,d2beta22))
H       <- as.matrix(H)
Se      <- sqrt(diag(ginv(H)))
num.iteration  <- paste("Iterations converged after", n, "times")
list(betaHat=beta1, uHat=u1, phiHat = phi1 ,Iteration=num.iteration, Se=Se )
}
else {print("Iterations did NOT converge!")}
}
```

APPENDIX C

EXTENDED RESTRICTED PSEUDO LIKELIHOOD FUNCTION

```
Pseude=function (x,y,z, beta0.cell=0 , phi=1,conv_crit=1e-8,n_maxiter=1000)
{
    N     <-   length(y)
    A     <-   diag(ncol(z)) ##random variance
    beta10 <-   as.matrix(rep(beta0.cell,ncol(x)),ncol(x),1)
    u10   <-   as.matrix(rep(beta0.cell,ncol(z)),ncol(z),1)

    phi100  <-   rep(phi,ncol(z))

    phi10  <-   diag(rep(phi100,n))


   eta   <-   x%*%beta10+z%*%u10
   mu    <-   exp(eta)/(1+exp(eta))

   D     <- diag(as.numeric(1/(mu*(1-mu))))  ## Partial derivative for eta ###
   V     <- diag(as.numeric(mu*(1-mu)))  ### V(m)=m(1-m)
   Vp    <- (z%*%A%*%t(z))+ phi10%*%D%*%V%*%D  ### Variance Psuedo
   P     <-  t(D )%*%(y-mu)+  eta  ### Linearization "Psudeo" ###

  betaHat <-   ginv(t(x) %*% ginv(Vp) %*% x) %*% t(x)%*%ginv(Vp)%*%P
  r      <-   P-(x%*%betaHat)  ### residuale
  uHat   <-   A%*%t(z)%*%ginv(Vp)%*%r

for (i in  1:k )
{

  VP <- Vp[(sum(n[1:(i-1)])+1):sum(n[1:i]),(sum(n[1:(i-1)])+1):sum(n[1:i])]

  R  <- r[(sum(n[1:(i-1)]+1):sum(n[1:i]))]

  phiHat  <- (1/n[i])*(t(R)%*%ginv(VP)%*%R)


}
   beta1  <- betaHat[1:ncol(x)]
   u1     <- uHat[1:ncol(z)]
   phi1   <- diag(rep(phiHat,n))
```

```
d   <-  max(abs(beta1-beta10), abs(u1 - u10 ), abs(det(phi1)- det(phi10)))


  if(d<conv_crit) {conv<-T} else{conv <- F}

 n <- 1
while(n<=n_maxiter & d>=conv_crit){

   beta10   <-  beta1
   u10      <-  u1
   phi10    <-  as.matrix(phi1)


   eta   <-   x%*%beta10+z%*%u10
   mu    <-   exp(eta)/(1+exp(eta))

   D     <- diag(as.numeric(1/(mu*(1-mu))))
   V     <- diag(as.numeric(mu*(1-mu)))
   Vp    <- (z%*%A%*%t(z))+ phi10%*%D%*%V%*%D
   P     <-  t(D )%*%(y-mu)+  eta

  betaHat <-  ginv(t(x) %*% ginv(Vp) %*% x) %*% t(x)%*%ginv(Vp)%*%P
  r      <-   P-(x%*%betaHat)
  uHat   <-   A%*%t(z)%*%ginv(Vp)%*%r

for (i in  1:k )
{

  VP <- Vp[((sum(n[1:(i-1)])+1):sum(n[1:i])),(sum(n[1:(i-1)])+1):sum(n[1:i])]

  R  <- r[((sum(n[1:(i-1)])+1):sum(n[1:i]))]

  phiHat  <- (1/n[i])*(t(R)%*%ginv(VP)%*%R)

}
```

```
 beta1  <- betaHat[1:ncol(x)]
  u1     <- uHat[1:ncol(z)]
  phi1   <- diag(rep(phiHat,n))



 d    <-  max(abs(beta1-beta10), abs(u1 - u10 ), abs(phi1- phi10))

 n      <- n+1

}

if(d<conv_crit) {conv<-T} else{conv <- F}
if(conv==T){

d2beta11 <-  t(x)%*%ginv(phi10%*%D%*%V%*%t(D))%*%x
d2beta22 <-  t(z)%*%ginv(phi10%*%D%*%V%*%t(D))%*%z+ginv(A)
d2beta12 <-  t(x)%*%ginv(phi10%*%D%*%V%*%t(D))%*%z
d2beta21 <-  t(z)%*%ginv(phi10%*%D%*%V%*%t(D))%*%x
H        <-  rbind(cbind(d2beta11,d2beta12),cbind(d2beta21,d2beta22))
H        <-  as.matrix(H)
Se       <-  sqrt(diag(ginv(H)))


num.iteration  <- paste("Iterations converged after", n, "times")

list(betaHat=beta1, uHat=u1, phiHat=phi1 , Iteration=num.iteration, Se=Se)
}
else {print("Iterations did NOT converge!")}
}
```

APPENDIX D

ADJUSTED SCALE BINOMIAL BETA h-LIKELIHOOD FUNCTION

```
k=20
Term2   <- matrix(0,k,1)

hA.lilihood=function(x,y,u,id)
{
ha.likA=function(param)

{
beta  <- param[1:3]

Term1  <-  t(y)%*%(x%*%beta+v)-(t(id)%*%log(1+exp(x%*%beta+v)))

for(i in  1:k)
{
Term2  <- sum((a*v)-((a+b[i])*log(1+exp(v)))-log(gamma(a))- log(gamma(b[i]))
         + log(gamma(a+b[i]))-log(exp(v)/(1+exp(v))^2))

}

Term3 = sum(Term2 )

fn1 <- sum(Term1 +Term3)

return(fn1)

}

a=2
b=rep(5,k)
v=u/1-u

hlA=maxLik(ha.likA, start =c(.1,.5,-.4),grad=NULL, hess= NULL)
}
```

APPENDIX E

POWER, TYPE I ERROR RATE AND STANDARD ERROR

Restricted Pseudo Likelihood

```
source("G:/dissertation/R-program/Diss.R/data.100.1000.txt")


simA= function (N1){

set.seed(1234)

  k      =  100

beta_1=matrix(c(beta0,beta1,beta2), 3,1)

alpha      <-  0.05

b11count    <-  0
b12count    <-  0
S.E1        <-  matrix(0,nrow=N1, ncol=1)


seeds=rnorm(N1,0,50)
set.seed(seeds)
for(i  in 1:N1)
{

 datta= mydata(seeds[i])
 x=datta$x
 y=datta$y
 z=datta$z
 u=datta$u
 id=datta$id
```

```
glmmres=lmer(y~x[,2]+x[,3]+ (1|id), family=binomial(link="logit"))
Vcov <- vcov(glmmres, useScale = FALSE)
betas <- fixef(glmmres)
se <- sqrt(diag(Vcov))
zval <- betas / se
pval <- 2 * pnorm(abs(zval), lower.tail = FALSE)
S.E1[i,] <- se[2]

       p11 = pval[2]
    if(p11 < alpha){b11count = b11count+1}


       p12 = pval[3]
    if (p12 < alpha){b12count = b12count+1}

}

typeI1=b12count/N1
power1=b11count/N1
se1 <- sum(S.E1)/N1

list(se1=se1,power1=power1,typeI1=typeI1)
 }
```

```
source("G:/h.Ah.power/data.h.txt")

simA= function (N1){

  k      =  50


alpha      <-  0.05

b21count    <-  0
b22count    <-  0
S.E2        <-  matrix(0,nrow=N1, ncol=1)
b.E21        <-  matrix(0,nrow=N1, ncol=1)
b.E22        <-  matrix(0,nrow=N1, ncol=1)



seeds=rnorm(N1,0,50)
set.seed(seeds)
for(i  in 1:N1)
{

 datta= mydata(seeds[i])
 x=datta$x
 y=datta$y
 id=datta$id
 z=datta$z
```

```
R <- hglm(X =x, y = y, Z = z,
family = binomial(link = logit))

betas <- R$fixef
se    <- R$SeFe
zval  <- betas / se
pval   <- 2 * pnorm(abs(zval), lower.tail = FALSE)

S.E2[i,] <- se[2]
b.E21[i,] <- betas[2]
b.E22[i,] <- betas[3]

       p21 = pval[2]
    if(p21 < alpha){b21count = b21count+1}


       p22 = pval[3]
    if (p22 < alpha){b22count = b22count+1}


}

typeI2=b22count/N1
power2=b21count/N1
se2 <- sum(S.E2)/N1
be21 <- sum(b.E21)/N1
be22 <- sum(b.E22)/N1

list(se2=se2,power2=power2,typeI2=typeI2,be21=be21,be22=be22)
}
```

```
 source("E:/R.diss/Diss.R/20.10.txt")
source("E:R.diss/hA.lik.max.txt")


simA= function (N1){

set.seed(1234)

  k      =  20


alpha      <-  0.05

b31count    <-  0
b32count    <-  0
S.E3       <-  matrix(0,nrow=N1, ncol=1)
b.E31       <-  matrix(0,nrow=N1, ncol=1)
b.E32       <-  matrix(0,nrow=N1, ncol=1)



seeds=rnorm(N1,0,50)
set.seed(seeds)
for(i  in 1:N1)
{

 datta= mydata(seeds[i])
 x=datta$x
 y=datta$y
 u=datta$u1
 id=datta$id
```

```
tt=hA.lilihood(x,y,u,id)

betas    <-  coef(tt)      # to find coeffecient
se       <-  stdEr(tt)     # standred erroe
zval     <-  betas / se
pval     <- 2 * pnorm(abs(zval), lower.tail = FALSE)
S.E3[i,] <- se[2]
b.E31[i,] <- betas[2]
b.E32[i,] <- betas[3]

        p31 = pval[2]
    if(p31 < alpha){b31count = b31count+1}


        p32 = pval[3]
    if (p32 < alpha){b32count = b32count+1}

}

typeI3=b32count/N1
power3=b31count/N1
se3 <- sum(S.E3)/N1
be31 <- sum(b.E31)/N1
be32 <- sum(b.E32)/N1

list(se3=se3,power3=power3,typeI3=typeI3,be31=be31,be32=be32)
}
```