5-1-2013

Goodness of fit statistics for mixed effect logistic regression models

Jalal Abdalla Saaid

University of Northern Colorado

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

GOODNESS OF FIT STATISTICS FOR MIXED EFFECT LOGISTIC REGRESSION MODELS

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

Jalal Abdalla Saaid

College of Education and Behavioral Sciences
Department of Applied Statistics and Research Methods

May, 2013

This Dissertation by: Jalal Abdalla Saaid

Entitled: *Goodness of Fit Statistics for Mixed Effect Logistic Regression Models*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in College of Education and Behavioral Sciences in Department of Applied Statistics and Research Methods

Accepted by the Doctoral Committee

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

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

Robert Pearson, Ph.D., Committee Member

Robert L. 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

Saaid, Jalal Abdalla. *Goodness of Fit Statistics for Mixed Effect Logistic Regression Models*. Published Doctor of Philosophy dissertation, University of Northern Colorado, 2013.

Mixed effects logistic regression models have become widely used statistical models to model clustered binary responses. However, assessing the goodness of fit (GOF) in these models when the cluster sizes and the number of clusters are small is not clear. In this research, three GOF statistics were proposed and their performance in terms of Type I error rate and power was examined via simulation study. The proposed GOF statistics were the logit residual, log-transformed residual, and the absolute residual GOF statistics. The simulation study was applied on different cases of number of clusters, cluster sizes, and types of predictors. The simulation results showed the performance of the logit residual and the log-transformed residual GOF statistics was poor. The absolute residual GOF statistic performed well over most cases of the simulation. It gave proper Type I error rates and high power for most cases. It is recommended for use in mixed effects logistic regression models as long as the number of clusters is at least 10 and the cluster sizes are 10 or more. However, the absolute residual GOF statistic can be affected by extremely small or large estimated probabilities and further research is recommended to avoid or reduce this restriction.

ACKNOWLEDGEMENTS

I would like to thank the Department of Applied Statistics and Research Methods for supporting me during my graduate study.

I am fully grateful to my advisors, Dr. Trent Lalonde and Dr. Jay Schaffer, for their kind help, quick feedback, and support to achieve this work.

Finally, I would like to think Dr. Rodney Sturdivant and Dr. Scott Evans for their support in SAS programs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The theory of linear models in statistics was extended by incorporating a random effect to the mixed effects linear model. For example in ANOVA models, a predictor factor is a random variable with a specific distribution; in this case, including the random effect in the model is necessary. In other cases, when we are dealing with longitudinal or repeated measures data sets, the response variable may be correlated within cases or subjects. Incorporating the random effect is very important to avoid the effect of autocorrelation in the response variable. Suppose we are interested in modeling the cholesterol levels on some independent variables for individuals over time. In that case, the measurements of the cholesterol on an individual would be correlated. Therefore, we might include a random effect in the model to account for the individual source of variability.

Further extensions of models are the class of generalized linear mixed effects models. In generalized linear mixed effects models, the response variables might be binary, categorical, continuous, or counts, i.e., the distribution of the response might not be a normal distribution. The word “generalized” refers to the distribution of the response. For example, a researcher is interested to determine whether an experimental teaching method is effective at improving math scores. In that case, the response variable “effectiveness” might be a categorical variable. Also, students from the same classroom should be correlated since they are taught by the same teacher. Classrooms within the same school might also be correlated. Therefore, we could include random effects at school and class levels to account for different sources of variability. This study focused on a special case of generalized linear mixed effects models where the response variable was a binary variable and its distribution was Bernoulli. Suppose we are treating a specific disease of some patients over a period of time and interested in modeling the existence of this disease on some predictors of those patients. In that case, the response variable would be binary and would take the value 1 if the disease exists and 0 if not. Also, the response variable of the same patient would be correlated. Therefore, we should include the random effect (patients) in the model to account for this source of variability. This type of model is known as mixed effects logistic regression model. It is very commonly used model in analyzing clustered binary responses. The mixed effects linear logistic regression model can be written,

$$\boldsymbol{y}|\boldsymbol{v} \sim \ Bernoulli\ (\boldsymbol{p}),$$
$$\boldsymbol{v} \sim \ i.i.d.D_n(\mathbf{0}, \boldsymbol{D}),$$
$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zv},$$
$$\boldsymbol{\eta} = logit\ (\boldsymbol{p}),$$

where $\boldsymbol{v}$ is a vector of the random effect parameter with covariance matrix $\boldsymbol{D}$, and D_n is an distribution from the exponential family for the random effect vector. The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the design matrices for the fixed and random effects parameters, respectively, and $\boldsymbol{\beta}$ is a vector of the fixed effects parameters. The systematic component $\boldsymbol{\eta}$ is equated to the "$logit$" transformation of the probability vector $\boldsymbol{p}$, $logit\ (\boldsymbol{p}) = \ln(\boldsymbol{p}/(1 - \boldsymbol{p})$.

Some methods of estimation developed in recent years can be used to estimate the parameters in such models. These methods of estimation use different likelihood functions such as pseudo-likelihood (Wolfinger & O'Connell, 1993), penalized quasi-likelihood (Breslow & Clayton, 1993), and hierarchical likelihood (Lee & Nelder, 1996). Using one of the likelihood functions, statistical inference for the estimated fixed and random effect parameters is typically based on asymptotic normality of the estimators using the estimated fisher information matrix of the used estimation technique.

To assess the goodness of fit for these models, some methods have been developed in recent years. However, there are advantages and disadvantages of using these methods. One of the present goodness of fit statistics is the Hosmer and Lemeshow (1980) test statistic. This test statistic groups the response variable into subgroups based on the estimated probabilities. That means instead of dealing with single observations, we deal with a frequency of each group so the residuals of the model represent the differences between the actual frequencies and the estimated frequencies. However, a recent study showed this test statistic is slightly conservative; it is not recommended for use in mixed effects logistic models (Evans & Hosmer, 2004). Furthermore, it might have low power to detect departure of model fit because it is only based on grouping the response while it ignores the predictors region (Hosmer, Hosmer, le Cessie, & Lemeshow, 1997; le Cessie & van Houwelingen, 1991).

Another goodness of fit statistic proposed by Evans and Hosmer (2004) is based on estimating the moments of the Pearson chi square statistic and unweighted sum of squares. Evans and Hosmer applied some Taylor series approximations to write the estimated residuals in terms of the actual residuals. The approximation for the estimated residuals' moments can be obtained by taking the expectation and variance for the approximated expression to the estimated residuals. These approximated moments of residuals are used to approximate the moments of both Pearson chi square and unweighted sum of squares statistics. Finally, Evans and Hosmer approximated the distributions of these statistics as chi square and normal distributions. Evans and Hosmer concluded that using the chi square approximation for the unweighted sum of squares and Pearson chi square statistics had good results in terms of type I error rate and were recommended for use in mixed effects logistic models. However, this recommendation was only for cluster sizes of 100 or greater and the model should have at least one continuous predictor (Evans & Hosmer, 2004).

Pan and Lin (2005) proposed graphical and numerical goodness of fit statistics for generalized linear mixed effects models. These statistics used the cumulative sum of the residuals, which have a distribution that can be approximated as a zero-mean Gaussian process under the true model. They generated the realizations of these processes by using Monte Carlo simulation and then compared the observed process of the model visually and analytically to the simulated realization. For large samples, their test statistics gave good type I error rate and power.

A recent study was conducted by Sturdivant and Hosmer (2007) to develop a new test statistic. This test statistic could be considered as an extension of Evans and Hosmer's (2004) work. They used the same idea of estimating the moments of the unweighted sum of squares and Pearson chi square statistics after smoothing the residuals by using cubic, uniform, and normal kernel functions. The smoothed residuals test statistic gave an appropriate type I error rate and good power for cluster sizes of 20 or more. They addressed a problem of selecting the optimal bandwidth for the kernel functions used in the smoothing and recommended further study for the optimal bandwidth selection.

The estimated residuals in the logistic models are the differences between the actual response value, which is 0 or 1, and the estimated probabilities of the model. The estimated residuals are fractions between -1 and 1. Therefore, the sum of squares of the residuals will be negligible and cannot be approximated as a chi square distribution with degrees of freedom related to the whole summation (Agresti, 2002). This situation is not an easy way to develop a goodness of fit test for mixed effects logistic models.

Hosmer and Lemeshow's (1980) test statistic can handle this situation; however, as mentioned previously, it is slightly conservative. Also the test statistic proposed by Evans and Hosmer (2004) is recommended for large samples only. Sturdivant and Hosmer's (2007) test statistic can be useful for cluster sizes of 20 or more but it needs to smooth the residuals before using a chi square or a normal distribution approximation. Furthermore, selecting the bandwidth for the kernel function used in the smoothing is not clear and needs further study.

In this dissertation, three goodness of fit statistics that could be used to test the model fit in mixed effects logistic regression models or usual logistic regression models are proposed. Estimates of the moments of these statistics are given so their distributions could be approximated as a normal or a chi square distribution. These test statistics could be valid for small cluster sizes.

The first test statistic is based on the residuals of the "$logit$" of the probabilities instead of the actual residuals,

$$e_i = logit(y_i) - logit(\hat{p}_i) \ ; \quad i = 1, \dots, n.$$

Since the "$logit$" of the probabilities is a continuous random variable that has values in the interval $(-\infty, \infty)$, the unweighted sum of squares and Pearson statistics are approximated using a chi square distribution. Taylor series approximations to the above residuals expression are applied to estimate the moments of these residuals. The estimated moments are used to estimate the moments of the unweighted sum of squares and Pearson statistics. However, the distributions of these statistics are approximated as a normal or a chi square distribution; the details for this test statistic are provided in Chapter III.

The second test statistic is based on transforming the actual residuals of the probabilities such that the new residuals would be a continuous variable in the interval $(0, \infty)$, with the same variability of the actual residuals. The new transformed residuals are

$$\varepsilon_i = -\ln(1 - |e_i|) \qquad ; \quad i = 1, \dots, n,$$

where

$$e_i = y_i - p_i.$$

The moments of the transformed residuals are estimated using a Taylor series approximation. The same idea of the previous proposed test statistics is applied to estimate the moments of the unweighted sum of squares and Pearson statistics and approximate their distributions.

The third test statistic is based on the absolute residuals instead of the actual residuals, and it is simply the sum of the absolute residuals. Assuming the residuals of the logistic models are approximately normally distributed, the moments of this fit statistic can be derived using a folded normal approximation, and then its distribution can be approximated.

A simulation study is conducted in Chapter IV to examine the proposed test statistics and answer the following research questions:

Q1 What is the sampling distribution of the logit residual goodness of fit statistic?

Q2 What is the sampling distribution of the log-transformed residual goodness of fit statistic?

Q3 What is the sampling distribution of the absolute residual goodness of fit statistic?

Q4 Do these proposed goodness of fit statistics have greater power than existing goodness of fit statistics for small cluster sizes?

Q5 Do these proposed goodness of fit statistics have proper type I error rate?

In Chapter II, logistic regression models, the method of maximum likelihood to estimate the parameters in these models, quasi-likelihood, assessing the goodness of fit in such models, and the overdispersion problem are presented. Also, the mixed effects logistic regression models, methods of estimation including pseudo likelihood, penalized quasi-likelihood, and hierarchical likelihood are introduced. Some goodness of fit statistics that have been developed in recent years for mixed effects logistic regression models are presented. In Chapter III, the proposed new test statistics, the approximations for their moments, and how to approximate their distributions are introduced. In Chapter IV, a simulation study to compare the proposed test statistics with Sturdivant and Hosmer's (2007) test statistics is conducted. The type I error rate and the power of each test statistic are considered in the comparison over some cases of cluster sizes and number of clusters. In Chapter V, some conclusions about this work and some recommendations for the future work are presented.

CHAPTER II

REVIEW OF LITERATURE

Logistic Regression Models

Many practical studies in the medical sciences, social sciences, and other fields need to model binary response variables for which the response outcomes are success or failure. For example, one might be interested in modeling the results of admission into graduate school on some observed variables of a sample of students, e.g., grade point average, GRE score, etc. In this case, the response variable would take the value of 1 if a student is admitted and 0 if not. One of the statistical models that could be used to deal with binary response data is the logistic regression model. The binary random response can be defined as

$$y = \begin{Bmatrix} 1 & if\ the\ outcome\ is\ success \\ 0 & if\ the\ outcome\ is\ failure \end{Bmatrix}.$$

The above binary random response could be considered as a Bernoulli random variable with probability of success p and probability of failure $(1 - p)$. Similarly, the sum of the responses over a sample n would have a binomial distribution. The general form for the logistic regression model can be written as

$$logit\ (\boldsymbol{p}) = \log\left(\frac{\boldsymbol{p}}{1 - \boldsymbol{p}}\right) = \boldsymbol{X\beta}. \qquad 2.1$$

The right hand side of the above equation is called the systematic component, where $\boldsymbol{X}$ is a (n x k) design matrix and $\boldsymbol{\beta}$ is a (k x 1) vector of parameters.

The term $logit\ (\boldsymbol{p}) = \log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right)$ is a logit transformation from probabilities to a continuous random response; it is called the link function. From equation 2.1, we can write the probability of success vector as

$$\boldsymbol{p} = \frac{e^{\boldsymbol{X\beta}}}{1+e^{\boldsymbol{X\beta}}}.$$

Estimation

Maximum likelihood. Let $y_1, \dots, y_m$ be independent random variables such that y_i is the number of successes in the group or class i, n_i is the number of trials in the ith class, and p_i is the probability of success in the ith class. In that case, y_i would have the binomial distribution with parameters (n_i, p_i). The likelihood function for the ith observation could be written as

$$l(p_i; y_i) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}.$$

For the independent observations, the likelihood function would be

$$l(\boldsymbol{p}; \boldsymbol{y}) = \prod_{i=1}^{m} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}.$$

Therefore, the log likelihood function $L(.)$ could be written as

$$L(\boldsymbol{p}; \boldsymbol{y}) = \sum_{i=1}^{m} \left[log \binom{n_i}{y_i} + y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i(1-p_i)\right]. \qquad 2.2$$

Using Equation 2.1, the log likelihood function in terms of the $X_{i's}$ and $\beta_{i's}$ could be written as follows:

$$L(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{m} \left[log \binom{n_i}{y_i} + y_i \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{X_{ij}} \beta_{\mathrm{j}} - n_i \log\left(1 + \exp \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{X_{ij}} \beta_{\mathrm{j}}\right)\right].$$

To find the estimators for the coefficients $\boldsymbol{\beta}$, we would derive the log likelihood function with respect to $\boldsymbol{\beta}$ and maximize this function. First, the derivative of the log likelihood function 2.2 with respect to p_i is

$$\frac{\partial L}{\partial p_i} = \sum_{i=1}^{m} \frac{y_i - n_i p_i}{p_i(1-p_i)}.$$

Using the relation $p_i = \frac{e^{\beta_0+\,\beta_1 X_{i1}+\cdots+\beta_k X_{ik}}}{1+e^{\beta_0+\,\beta_1 X_{i1}+\cdots+\beta_k X_{ik}}}$, we can find $\frac{\partial p_i}{\partial \beta_j}$.

By applying the chain rule, we can find the derivative of the log likelihood function with respect to β_j as follows:

$$\frac{\partial L}{\partial \beta_j} = \frac{\partial L}{\partial p_i}\ \frac{\partial p_i}{\partial \beta_j},$$

where

$$\frac{\partial L}{\partial p_i} = \sum_{i=1}^{m} \frac{y_i - n_i p_i}{p_i(1-p_i)},$$

and

$$\frac{\partial p_i}{\partial \beta_j} = \frac{e^{\beta_0+\,\beta_1 X_{i1}+\cdots+\beta_k X_{ik}}}{[1+e^{\beta_0+\,\beta_1 X_{i1}+\cdots+\beta_k X_{ik}}]^2}\ X_{ij}.$$

Therefore, the maximum likelihood estimator for the jth coefficient could be obtained by solving the following score equations:

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{m} \frac{y_i - n_i p_i}{p_i(1-p_i)}\ \ \frac{e^{\beta_0+\,\beta_1 X_{i1}+\cdots+\beta_k X_{ik}}}{[1+e^{\beta_0+\,\beta_1 X_{i1}+\cdots+\beta_k X_{ik}}]^2}\ X_{ij} = 0.$$

These equations could be solved by applying the iterative method using a computer program. The same procedure would be applied for all entire parameters.

For the vector of parameters $\boldsymbol{\beta}$, the iterative equations could be derived by using Newton-Raphson method, which is derived by using Taylor expansions around initial parameter for a function $f'(\theta)$ of parameter θ. This procedure is based on the general iterative equation,

$$\theta_{t+1} = \theta_t - \frac{f'(\theta)}{f''(\theta)},$$

where $\theta_{t+1} = \theta_t$ (converges) when $f'(\theta) = 0$.

Using this method, we can derive the iterative equations for the parameters vector as

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{new}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{old}} + \left[-L''\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{old}}\right)\right]^{-1}.\, L'\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{old}}\right).$$

The maximum likelihood estimators for the parameters would be efficient, consistent, and asymptotically normally distributed.

Quasi-likelihood. It is known that when generalized linear models are applied using a distribution such as Binomial or Poisson, there is a specific relationship between the mean and the variance of the distribution. In the case of binary response data, which follow a Bernoulli distribution, there is the mean-variance relationship,

$$V(p) = p(1-p).$$

Normally the likelihood function is constructed using the assumption that the response distribution is fully specified, but unusual relationships between the mean and the variance of a response could occur. In most real data situations, the above mean-variance relationship does not hold. Extra variation usually exists with practical observations. The quasi-likelihood approach could deal with such problems; it needs only the specification of the mean-variance relationship rather than specifying the full distribution of the response. In general, the quasi-likelihood function is constructed as follows (Wedderburn, 1974).

Let $y_1, \dots, y_n$ be independent responses with mean $E(y_i) = \mu_i$ and variance $var(y_i) = a(\emptyset)V(\mu_i)$. Assume μ_i is a function of the unknown parameters $\beta_1, \dots, \beta_k$, V is a known function, and $a(\emptyset)$ is the dispersion parameter. This method of estimation needs only a model for the mean with respect to the relationship between the mean and the variance of the data and does not need the full distribution of the data. The quasi-likelihood is defined as a function $q(\mu_i, y_i)$ such that

$$\frac{\partial q(\mu_i; y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{a(\emptyset)V(\mu_i)}.$$

This function is defined for one observation(y_i). For n independent observations, the quasi-likelihood function could be written as

$$Q = \sum_{i=1}^{n} q(\mu_i, y_i) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\emptyset)V(\mu_i)}.$$

As μ_i is a function of the regression coefficients, estimators for the regression coefficients could be obtained by solving the following equations:

$$\begin{aligned} \frac{\partial Q}{\partial \beta_j} &= \frac{\partial Q}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \\ &= \sum_{i}^{n} \frac{\partial q(\mu_i, y_i)}{\partial \beta_j} \\ &= \sum_{i}^{n} \frac{\partial \mu_i}{\partial \beta_j} \frac{y_i - \mu_i}{a(\emptyset)V(\mu_i)}, \end{aligned}$$

for $j = 1, \dots, k$. The above equations are full-data, quasi-likelihood functions for estimating β_j and can be written in matrix notation as follows:

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = \boldsymbol{D}' \left[\frac{1}{a(\emptyset)\boldsymbol{V}}\right]^{-1} (\boldsymbol{y} - \boldsymbol{\mu}), \qquad 2.3$$

where $\frac{\partial Q}{\partial \boldsymbol{\beta}}$ is a (n x1) vector of the elements $\frac{\partial Q}{\partial \beta_j}$, $\boldsymbol{D}$ is a (n x p) matrix of the elements $\frac{\partial \mu_i}{\partial \beta_j}$, $\boldsymbol{V}$ is a (n x n) diagonal matrix with elements $V(\mu_i)$, $a(\emptyset)$ is the dispersion parameter, $\boldsymbol{y}$ is a (n x 1) vector of responses, and $\boldsymbol{\mu}$ is a (n x 1) vector of associated means. Equation 2.3 is called the quasi-score function and it has the general form of

$$U(\boldsymbol{\beta}) = \frac{\partial Q}{\partial \boldsymbol{\beta}} = \boldsymbol{D}' \left[\frac{1}{a(\emptyset)\boldsymbol{V}}\right]^{-1} (\boldsymbol{y} - \boldsymbol{\mu}).$$

Therefore, the mean of the quasi-score function is

$$E[U(\boldsymbol{\beta})] = \boldsymbol{D}' \left[\frac{1}{a(\emptyset)\boldsymbol{V}}\right]^{-1} E(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}.$$

We find $\widehat{\boldsymbol{\beta}}$ by solving $E[U(\boldsymbol{\beta})] = 0$. The covariance matrix of $U(\boldsymbol{\beta})$ can be obtained as follows:

$$\begin{aligned} Var[U(\boldsymbol{\beta})] = &\quad -E\left[\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] \\ = &\quad -E\left[\frac{\partial^2 Q}{\partial \boldsymbol{\beta}^2}\right] \\ = &\quad \frac{\boldsymbol{D}^{\boldsymbol{T}}\boldsymbol{V}^{-\boldsymbol{1}}\boldsymbol{D}}{a(\emptyset)} \\ = &\quad I(\boldsymbol{\beta}). \end{aligned}$$

McCullagh and Nelder (1989) showed that this matrix approximately played the same role as the Fisher information matrix of the ordinary likelihood functions. Under some limitations on the eigenvalues of $I(\boldsymbol{\beta})$, the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ could be written as

$$Cov(\widehat{\boldsymbol{\beta}}) \approx I^{-1}(\boldsymbol{\beta}) = a(\emptyset)[\boldsymbol{D}'\boldsymbol{V}^{-\boldsymbol{1}}\boldsymbol{D}]^{-1}.$$

One estimation procedure for the quasi-likelihood equations is called the iterative quasi-scoring procedure; it is performed using the iteration approach on the quasi-score functions.

Using the Newton-Raphson method,

$$\widehat{\boldsymbol{\beta}}_{new} = \widehat{\boldsymbol{\beta}}_{old} + [-Q''(\boldsymbol{\beta})]^{-1}Q'(\boldsymbol{\beta}),$$

where

$$Q'(\boldsymbol{\beta}) = \frac{\partial Q}{\partial \boldsymbol{\beta}} = \boldsymbol{D}'\left[\frac{1}{a(\emptyset)\boldsymbol{V}}\right]^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) = \boldsymbol{D}'\boldsymbol{V}^{-\mathbf{1}}(\boldsymbol{y}-\boldsymbol{\mu})$$

and

$$Q''(\boldsymbol{\beta}) = \frac{\partial^2 Q}{\partial \boldsymbol{\beta}^{\mathbf{2}}} = -\boldsymbol{D}'\left[\frac{1}{a(\emptyset)\boldsymbol{V}}\right]^{-1}\boldsymbol{D} = -\boldsymbol{D}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{D}.$$

Thus, the iterative procedure for quasi-likelihood estimators would be

$$\widehat{\boldsymbol{\beta}}_{new} = \widehat{\boldsymbol{\beta}}_{old} + [\boldsymbol{D}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{D}]^{-\mathbf{1}}\boldsymbol{D}'\boldsymbol{V}^{-\mathbf{1}}(\boldsymbol{y}-\boldsymbol{\mu}),$$

where **D,** $\boldsymbol{V}^{-\mathbf{1}}$ and $\boldsymbol{\mu}$ are calculated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{old}$.

This estimation approach assumes that the variance function $a(\emptyset)V(\mu_i)$ is correctly specified. Under this assumption, the estimates of the regression coefficients using this procedure are consistent, asymptotically unbiased, and asymptotically, normally distributed (Lee, Nelder, & Pawitan, 2006). If the variance function is not correctly specified, the regression coefficients will not be efficient. In other words, the estimated variance $a(\emptyset)[\boldsymbol{D}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{D}]^{-1}$will not be a consistent estimator of $Cov(\widehat{\boldsymbol{\beta}})$.

Huber (1967) and White (1980) proposed a new estimate for the covariance of parameter estimators. This estimate is called the robust or “sandwich” estimator; it can be used for any specified variance function. The sandwich estimator can be defined as

$$Cov(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{D}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{D})^{-\mathbf{1}}(\boldsymbol{D}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{\Sigma}\,\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{D})(\boldsymbol{D}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{D})^{-\mathbf{1}},$$

where Σ is a diagonal matrix consists of the elements

$$\sigma_{ii} = [D_{ii}]^2\, Var(y_i).$$

Assessing Model Fit

In this section, I discuss some fit statistics that could be used to test the model fit in the logistic regression models.

Deviance. The deviance is simply the twice difference between the log likelihood function under the actual sample and the log likelihood under the fitted observations of the model. Therefore, the deviance function can be written as

$$D(\boldsymbol{y}; \widehat{\boldsymbol{p}}) = 2\{l(\boldsymbol{p}; \boldsymbol{y}) - l(\widehat{\boldsymbol{p}}; \boldsymbol{y})\}.$$

That is,

$$\begin{aligned} D(\boldsymbol{y}; \widehat{\boldsymbol{p}}) &= 2\left\{\sum_{i=1}^{m}\left[\left[log\binom{n_i}{y_i} + y_i log\, p_i + (n_i - \mathrm{y_i})\log(1-p_i)\right]\right.\right. \\ &\qquad \left.\left. - \left[log\binom{n_i}{y_i} + y_i log\, \hat{p}_i\ + (n_i - \mathrm{y_i})\log(1-\hat{p}_i)\right]\right]\right\} \\ &= 2\sum_{i=1}^{m}\left\{y_i log\left(\frac{p_i}{\hat{p}_i}\right) + (n_i - \mathrm{y_i})\log(\frac{1-p_i}{1-\hat{p}_i})\right\}, \end{aligned}$$

where $p_i = \frac{y_i}{n_i}$ is the probability of success calculated under the entire observations for group or category i and $\hat{p}_i$ is that probability estimated under the fitted model. The deviance function can be written as

$$D(\boldsymbol{y}; \widehat{\boldsymbol{p}}) = 2\sum_{i=1}^{m}\left\{y_i log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - \mathrm{y_i})\log(\frac{n_i - y_i}{n_i - \hat{y}_i})\right\} \quad ; \quad y_i = n_i p_i \quad \& \quad \hat{y}_i = n_i \hat{p}_i,$$

where y_i is the number of successes and $\hat{y}_i$ is the predicted number of successes in the ith group using the fitted model.

In most situations, the deviance function has a behavior similar to the residual sum of squares or the weighted sum of squares in the usual linear models. This test statistic is very useful as long as we have categorical or binary predictors in the model, i.e., this test statistic is designed for grouped responses and will not be useful if we have only continuous predictors. Under the assumption of independence of the groups, the deviance statistic can be used to test the hypothesis that the model is fit or not by comparing the deviance statistic with the tabulated chi square distribution with degrees of freedom (*m*-*p*), where *m* is the number of groups and *p* is the number of parameters in the model.

Pearson sum of squares statistic. The Pearson goodness of fit statistic is the sum of squares of the Pearson residuals; it can be written as

$$X^2 = \sum_{i=1}^{n} \hat{r}_i^2,$$

where

$$\hat{r}_i = \frac{(p_i - \hat{p}_i)}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

is the ith “studentized” Pearson residual.

This statistic is assumed to have asymptotic chi square distribution with (*n*-*p*) degrees of freedom. Hosmer et al. (1997) showed that the *p*-value of this statistic is usually conservative using a chi square distribution. In other words, the value of this test statistic is usually negligible when compared with a chi square value, which leads us to fail to reject a model with poor fit.

Unweighted sum of squares statistic. The unweighted sum of squares test statistic is simply the sum of squares of the nonstandardized residuals, which can be defined as

$$\hat{e}_i = (p_i - \hat{p}_i).$$

Thus, the unweighted sum of squares statistic is

$$S = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (p_i - \hat{p}_i)^2.$$

This test statistic is assumed to have asymptotic chi square distribution with *(n-p)* degrees of freedom and can be used to test the hypothesis that the model is fit or not. However, if the response variable is not grouped over the predictors, this test statistic is not useful and could be conservative.

Hosmer and Lemeshow's test statistic. Suppose we have a model with only continuous predictors. In that case, the response observations could not be grouped into classes based on a categorical or binary predictor. Therefore, the previous test statistics would not fit in such a situation because the calculated chi square would be negligible and then the test would be conservative. Hosmer and Lemeshow's (1980) test statistic grouped the response observations into subgroups based on percentiles of the estimated probabilities. Usually 10 groups are chosen with 10 in each group; however, the number of groups is subjective to the researcher and depends on the sample size. Each group has two pairs of counts: one for the observed counts of data falling into the group and the other for the predicted counts. In the first group, the first pairs of counts have the highest decile estimated probabilities; the next second pairs have second decile estimated probabilities, and so forth. After grouping the data, Hosmer and Lemeshow used a

Pearson test statistic to compare the observed counts with the fitted counts. Therefore, their test statistic can be written as

$$\hat{C} = \sum_{j=1}^{10} (O_j - \hat{f}_j)^2 / \hat{f}_j,$$

where O_j and $\hat{f}_j$ are the observed and the predicted number of successes in the j$^{\text{th}}$ group, respectively. Depending on a simulation study, Hosmer and Lemeshow showed that this test statistic had approximately chi square distributed with degrees of freedom equal to number of groups minus 2. This is approximately true when the model is a good fit and the estimated expected frequencies are large.

Smoothed residual-based tests. Le Cessie and van Houwelingen (1991) noted that the Hosmer and Lemeshow's (1980) test depended on a grouping technique in the response space and it ignored the "x" space. They pointed that the Hosmer and Lemeshow test might lack power to detect departures from the model in regions of the "x" space, which might give the same predicted probabilities, i.e., the Hosmer and Lemeshow test might not be an appropriate fit statistic in detecting departures from linearity in the "x" space.

Le Cessie and van Houwelingen (1991) proposed a class of tests based on the smoothing of residuals with respect to "x" space. The idea of using smoothed residuals was proposed by Copas (1980) and Azzalini, Bowman, and Hardle (1989) who applied it in the non-parametric regression. They computed a smoothing value of the outcome variable for each subject, which is a weighted average of the response values for subjects near a subject, and compared it with the corresponding fitted probability.

Hosmer et al. (1997) employed weight functions, using the uniform kernel for the "x" space as applied by le Cessie and van Houwelingen (1991) and a cubic weight in the "y" space. To introduce these test statistics, the "x" space weight, which defines the distance between subject i and j, could be written as

$$w_{ij}^{x} = \prod_{k=1}^{p} u(x_{ik}, x_{jk}),$$

where $u(x_{ik}, x_{jk}) = 1$ if $|x_{ik} - x_{jk}|/s_k \leq c_u$, and zero otherwise, and s_k is the sample standard deviation of the k^th^ predictor. Based on a simulation study, le Cessie and van Houwelingen recommended that the cut point c_u should be chosen such that $\sqrt{n}$ of the subjects had non-zero weights.

For the "y" space, the cubic weights were used and defined as

$$w_{ij}^{y} = 1 - (|\hat{p}_i - \hat{p}_j|)^3$$

if $|\hat{p}_i - \hat{p}_j| \leq c_i$ and zero otherwise. The cut point c_i depended on *i* and was chosen such that $\sqrt{n}$ of the subjects had non-zero weights.

The smoothed standardized residuals could be written as

$$\hat{r}_{si} = \sum_{j=1}^{n} w_{ij}^{y} \hat{r}_j \quad ; \quad \hat{r}_j = (y_j - \hat{p}_j)/[\hat{p}_j(1 - \hat{p}_j)]$$

for the "y" space and

$$\hat{r}_{si} = \sum_{j=1}^{n} w_{ij}^{x} \hat{r}_j \quad ; \quad \hat{r}_j = (y_j - \hat{p}_j)/[\hat{p}_j(1 - \hat{p}_j)]$$

for the "x" space.

Using the above smoothed residuals, the smoothed residuals-based test statistic could be defined as

$$\hat{T}_r = \sum\nolimits_{i=1}^{n} \frac{\hat{r}_{si}^2}{\widehat{Var(\hat{r}_{si}^2)}}.$$

They denoted the statistic as $\hat{T}_{ru}$ in case of using the uniform kernel weights in the "x" space and as $\hat{T}_{rc}$ in case of using the cubic space in the "y" space.

To apply this test statistic, we needed to estimate the mean and variance for each test statistic under the assumption that the fitted model was true. Using the estimates of the moments, we used the normal approximation or chi square approximation by estimating its degrees of freedom.

Now to estimate the moments, Hosmer et al. (1997) assumed $\boldsymbol{W}$ was a (n x n) matrix of weights with i^th^ row $\boldsymbol{w_i}$, which consisted of the weights for the distance of subject i to subjects 1 to n.

Let $\widehat{\boldsymbol{R}} = diag[\hat{p}_{ij}(1 - \hat{p}_{ij})]$ be an n x n diagonal matrix. Thus, in matrix notation, the standardized residuals and the smoothed standardized residuals could be written as

$$\hat{\boldsymbol{r}} = \widehat{\boldsymbol{R}}^{\boldsymbol{1/2}}\hat{\boldsymbol{e}}$$

and

$$\hat{\boldsymbol{r}}_{\boldsymbol{s}} = \boldsymbol{W}\hat{\boldsymbol{r}} = \boldsymbol{W}\widehat{\boldsymbol{R}}^{\boldsymbol{1/2}}\hat{\boldsymbol{e}}.$$

Therefore, in matrix notation, the smoothed residuals based test could be expressed as

$$\hat{T}_r = \hat{\boldsymbol{r}}'\boldsymbol{W}'\boldsymbol{D}_{\boldsymbol{r}}^{-\boldsymbol{1}}\boldsymbol{W}\hat{\boldsymbol{r}} = \hat{\boldsymbol{e}}'\widehat{\boldsymbol{R}}^{\boldsymbol{1/2}}(\boldsymbol{W}'\boldsymbol{D}_{\boldsymbol{r}}^{-\boldsymbol{1}}\boldsymbol{W})\widehat{\boldsymbol{R}}^{\boldsymbol{1/2}}\hat{\boldsymbol{e}},$$

where $\boldsymbol{D_r}$ is a (n x n) diagonal matrix of the diagonal elements of the matrix $\boldsymbol{WW'}$and represents a diagonal matrix of the variances of the smoothed residuals.

This quadratic form cannot be simplified as a Pearson chi-square statistic because the matrix $\boldsymbol{D_r^{-1}}$ in the middle includes only diagonal elements of the variances of the smoothed residuals; it is not a full covariance structure.

Using the first order Taylor's approximation, le Cessie and van Houwelingen (1991) derived

$$\boldsymbol{\widehat{p}} \cong \boldsymbol{p} + \boldsymbol{Me},$$

$$\boldsymbol{\hat{e}} \cong (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{e},$$

$$\boldsymbol{\widehat{R}} = \boldsymbol{R},$$

where $\boldsymbol{M} = \boldsymbol{RX}(\boldsymbol{X^T RX})^{-\mathbf{1}}\boldsymbol{X^T}$.

By substituting $\boldsymbol{\hat{e}} \cong (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{e}$ and $\boldsymbol{\widehat{R}} = \boldsymbol{R}$, the test statistic could be written as

$$\hat{T}_r = \boldsymbol{e}'(\boldsymbol{I} - \boldsymbol{M})'\boldsymbol{R^{1/2}}(\boldsymbol{W'D_r^{-1}W})\boldsymbol{R^{1/2}}(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{e} = \boldsymbol{e'A_r e},$$

where $\boldsymbol{A_r} = (\boldsymbol{I} - \boldsymbol{M})^{\boldsymbol{T}}\boldsymbol{R^{1/2}}(\boldsymbol{W^T D_r^{-1} W})\boldsymbol{R^{1/2}}(\boldsymbol{I} - \boldsymbol{M})$.

Now, using the results for the moments of quadratic forms by Seber (1977), le Cessie and van Houwelingen (1991) derived the moments of the test statistic as

$$E(\hat{T}_r) = trace(\boldsymbol{A_r R}),$$

$$var(\hat{T}_r) = \sum_{i=1}^{n} a_{rii}^2 r_i(1 - 6r_i) + 2\ trace(\boldsymbol{A_r R}\ \boldsymbol{A_r R}).$$

Estimates of the above moments could be obtained by substituting $\hat{R}$ in the formulas.

Based on a simulation study, Hosmer et al. (1997) concluded that for small samples, it was better to approximate the distribution of the smoothed residual based test as a scaled chi square distribution $c\chi_d^2$. The constant c and the degrees of freedom d depended on the estimated mean and variance and could be estimated as

$$\hat{T}_r \dot{\sim} c\chi_d^2,$$

$$E(\hat{T}_r) = cE(\chi_d^2) = cd,$$

and

$$Var(\hat{T}_r) = c^2 Var(\chi_d^2) = 2c^2 d.$$

Solving the above moment equations, we could get

$$c = \frac{Var(\hat{T}_r)}{2E(\hat{T}_r)}, \qquad d = \frac{2[E(\hat{T}_r)]^2}{Var(\hat{T}_r)},$$

$$\frac{2E(\hat{T}_r)}{Var(\hat{T}_r)}\ \hat{T}_r \ \dot{\sim}\ \chi_d^2 \ ; \quad d = \frac{2[E(\hat{T}_r)]^2}{Var(\hat{T}_r)}.$$

Therefore, we could accept the hypothesis that the model is fit if

$$\hat{T}_r < \tfrac{Var(\hat{T}_r)}{2E(\hat{T}_r)}\ \chi_d^2 \ ; \quad d = \tfrac{2[E(\hat{T}_r)]^2}{Var(\hat{T}_r)}.$$

Depending on a simulation study, Hosmer et al. (1997) found that these tests had power exceeding 90% to detect moderate departures from the model linearity when the sample size was 500 and over 50% when the sample size was 100.

From the previous review of the test statistics for the logistic regression models, each test statistic was not always appropriate to use. The deviance, Pearson chi square, and the unweighted sum of squares statistics would not be useful if the response could not be grouped over the predictors, i.e., they would not be appropriate for models with only continuous predictors. Hosmer and Lemeshow's (1980) test statistic could be used for any situation of the model but a recent study by Evans and Hosmer (2004) showed it was a slightly conservative test. Another study noted that it might lack power to detect departures from the model in regions of the "x" space. Smoothed residual-based tests are appropriate for use with any model to detect a moderate departure of linearity since they have power exceeding 50% with cluster sizes that are at least 100 and exceeding 90% with cluster sizes that are at least 500.

Overdispersion

Overdispersion in binomial responses is the property of variance in the response y_i being larger than the variance indicated by the binomial model. The overdispersion problem commonly occurs in practical applications when the responses are correlated or clustered as in the case of longitudinal data. However, extra variation in the data causes both the Pearson and deviance statistics to be too large, which leads to a false conclusion of poor fit. In ordinary linear models, there is no existence of such a problem because in a linear regression model,

$$y_i \sim N(\boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta}, \sigma^2),$$

the variance σ^2 is estimated separately of the mean function $\boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta}$. However, with discrete response variables, the variance is estimated by the mean. The reason is that the Binomial and Poisson distributions specify particular relationships between the variance and the mean. However, overdispersion is an undesired problem because it inflates the type 1 error rate in the model.

There are some approaches to deal with this problem. One way is to specify a more dispersed distribution than the usual distribution I used. For example, a binomial model could be changed to the beta-binomial model. A more popular method for adjusting for overdispersion came from the theory of quasi-likelihood and different estimating equation techniques.

Mixed Effects Logistic Regression Models

In normal linear mixed models, we learned that we should incorporate a random effect into the model when we have a random sample of a grouping factor as a predictor. This adjusts the response variance to account for another source of variation. Furthermore, in dealing with repeated measures and longitudinal studies, the response observations are usually clustered or correlated within each subject. Therefore, including the random effect in the model becomes important to account for different sources of variability. The same situations might occur when we deal with generalized linear models. The mixed effects logistic regression model is a special case of the generalized linear mixed model when the response variable is binary. To introduce this model, let $\boldsymbol{y}$ be a vector of Bernoulli observations with a vector of corresponding probabilities $\boldsymbol{p}$. The mixed effects logistic regression model probability could be defined as

$$\boldsymbol{p} = \frac{e^{\boldsymbol{X\beta}+\boldsymbol{Zv}}}{1 + e^{\boldsymbol{X\beta}+\boldsymbol{Zv}}},$$

with the components of

$$\boldsymbol{y}|\boldsymbol{v} \sim \ Bernoulli\ (\boldsymbol{p}),$$

$$\boldsymbol{v} \sim \ i.i.d. D_n(\boldsymbol{0}, \boldsymbol{D}),$$

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zv},$$

$$\boldsymbol{\eta} = logit\ (\boldsymbol{p}),$$

where $\boldsymbol{v}$ is a vector of the random effect with covariance matrix $\boldsymbol{D}$. The distribution of the random effect D_n is an arbitrary distribution from an exponential family. The link $\boldsymbol{\eta}$ is the "$logit$" transformation of the probabilities vector $\boldsymbol{p}$. In subsequent sections, some methods of estimation for the parameters in such models are introduced.

Estimation of Parameters

Pseudo likelihood. The pseudo likelihood was proposed by Wolfinger and O'Connell (1993); it is one of the methods that can be used to estimate the parameters in generalized linear mixed effects models. The idea of this approach is to transform a nonlinear mixed model to a regular linear mixed effects model by using the first order Taylor approximation on the inverse of the link function. Assume that we have the generalized mixed logistic effects model,

$$\boldsymbol{y}|\boldsymbol{u} \sim D_1[\boldsymbol{\mu}, \phi\, \boldsymbol{V}(\boldsymbol{\mu})],$$

$$\boldsymbol{u} \sim D_2[\boldsymbol{\mu_r}, \lambda\, \boldsymbol{V_r}(\boldsymbol{\mu_r})],$$

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zv},$$

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

where D_1 and D_2 are the distributions for the conditional response $\boldsymbol{y}|\boldsymbol{u}$ and the random effect $\boldsymbol{u}$, respectively. The random effect $\boldsymbol{v}$ is a link transformation of $\boldsymbol{u}$, $\boldsymbol{v} = g(\boldsymbol{u})$ and its distribution is assumed, $\boldsymbol{v} \sim D_3(\boldsymbol{0}, \boldsymbol{D})$. However, the random effect under the pseudo-likelihood approach is assumed to be approximately normally distributed.

Using the link function of the above model information, the response mean can be written as

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X\beta} + \boldsymbol{Zv}).$$

Now if we expand the response mean in a first order Taylor series approximation about estimators of the fixed effect parameters $\boldsymbol{\beta}$ and the random effects $\boldsymbol{v}$,

$$\begin{aligned}\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) \quad &\approx \; g^{-1}(\widehat{\boldsymbol{\eta}}) + diag\left(\left.\frac{\partial g^{-1}(\widehat{\boldsymbol{\eta}})}{\partial \widehat{\boldsymbol{\eta}}}\right|_{\eta=\widehat{\eta}}\right)(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}) \\ &\approx \; g^{-1}(\widehat{\boldsymbol{\eta}}) + \widehat{\boldsymbol{W}}\left(\boldsymbol{X\beta} + \boldsymbol{Zv} \;\; - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{v}} \;\right), \qquad 2.4\end{aligned}$$

where

$$\widehat{\boldsymbol{W}} = diag\left(\left.\frac{\partial g^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right|_{\eta=\widehat{\eta}}\right)$$

and

$$\widehat{\boldsymbol{\eta}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{v}}.$$

From equation 2.4, we can write

$$\boldsymbol{X\beta} + \boldsymbol{Zv} \approx \widehat{\boldsymbol{W}}^{-\mathbf{1}}[\boldsymbol{\mu} - g^{-1}(\widehat{\boldsymbol{\eta}})] + \widehat{\boldsymbol{\eta}}.$$

Thus, we can write the pseudo response of

$$\boldsymbol{P} = \widehat{\boldsymbol{W}}^{-\mathbf{1}}[\boldsymbol{\mu} - g^{-1}(\widehat{\boldsymbol{\eta}})] + \widehat{\boldsymbol{\eta}}$$

with a conditional mean of

$$\begin{aligned}E(\boldsymbol{P}|\boldsymbol{v}) \quad &= \widehat{\boldsymbol{W}}^{-\mathbf{1}}[E(\boldsymbol{y}|\boldsymbol{v}) - g^{-1}(\widehat{\boldsymbol{\eta}})] + \widehat{\boldsymbol{\eta}} \\ &= \widehat{\boldsymbol{W}}^{-\mathbf{1}}[\boldsymbol{\mu} - g^{-1}(\widehat{\boldsymbol{\eta}})] + \widehat{\boldsymbol{\eta}} \approx \boldsymbol{X\beta} + \boldsymbol{Zv},\end{aligned}$$

and conditional variance of

$$\begin{aligned}V(\boldsymbol{P}|\boldsymbol{v}) \quad &= Var(\widehat{\boldsymbol{W}}^{-\mathbf{1}}\boldsymbol{y}|\boldsymbol{v}) \\ &= \widehat{\boldsymbol{W}}^{-\mathbf{1}}Var(\boldsymbol{y}|\boldsymbol{v})\widehat{\boldsymbol{W}}^{-\mathbf{1}} \\ &= \widehat{\boldsymbol{W}}^{-\mathbf{1}}\boldsymbol{V}\widehat{\boldsymbol{W}}^{-\mathbf{1}}.\end{aligned}$$

Using the conditional variance rule, we can derive the marginal variance and the marginal mean for the pseudo response:

$$\begin{aligned} E(\boldsymbol{P}) &= E_{\boldsymbol{v}}[E(\boldsymbol{P}/\boldsymbol{v}) \\ &= E_{\boldsymbol{v}}[\boldsymbol{X\beta} + \boldsymbol{Zv}] \\ &= \boldsymbol{X\beta} + \boldsymbol{Z}\, E_{v}(\boldsymbol{v}) \\ &= \boldsymbol{X\beta}, \end{aligned}$$

and

$$\begin{aligned} V(\boldsymbol{P}) &= Var_{\boldsymbol{v}}[E(\boldsymbol{P}/\boldsymbol{v})] + E_{\boldsymbol{v}}[Var(\boldsymbol{P}/\boldsymbol{v})] \\ &= Var_{\boldsymbol{v}}[\boldsymbol{X\beta} + \boldsymbol{Zv}] + E_{\boldsymbol{v}}[\boldsymbol{\widehat{W}^{-1}V\widehat{W}^{-1}}] \\ &= \boldsymbol{ZDZ}' + \boldsymbol{\widehat{W}^{-1}V\widehat{W}^{-1}}. \end{aligned}$$

Now, we can write the linear pseudo model with an unobserved error term of

$$\boldsymbol{P} = \boldsymbol{X\beta} + \boldsymbol{Zv} + \boldsymbol{e},$$

where

$$E(\boldsymbol{e}|\boldsymbol{v}) = 0$$

and

$$Var(\boldsymbol{e}|\boldsymbol{v}) = V(\boldsymbol{P}|\boldsymbol{v}) = \boldsymbol{\widehat{W}^{-1}V\widehat{W}^{-1}}.$$

Wolfinger and O’Connell (1993) assumed

$$\boldsymbol{V} = \boldsymbol{V_{\mu}^{1/2}V^{*}V_{\mu}^{1/2}},$$

where $\boldsymbol{V_{\mu}^{1/2}}$ was a diagonal matrix of the variance function of $\boldsymbol{\mu}$ for a specific generalized linear model under the study and $\boldsymbol{V}^{*}$ was unknown. If we assume we are dealing with a logistic regression model, we have

$$
\begin{aligned}
\widehat{\boldsymbol{W}}^{\boldsymbol{-1}} &= diag\left(\left.\frac{\partial g^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}}\right)^{-1} \\
&= diag(g'[g^{-1}(\widehat{\boldsymbol{\eta}})]) \\
&= diag(g'[\widehat{\boldsymbol{p}}]) \\
&= diag\left(\frac{1}{\widehat{\boldsymbol{p}}(1-\widehat{\boldsymbol{p}})}\right),
\end{aligned}
$$

and

$$\boldsymbol{V}_{\boldsymbol{\mu}} = diag[\boldsymbol{p}(1-\boldsymbol{p})].$$

Using $\widehat{\boldsymbol{\mu}}$ as an initial approximation of $\boldsymbol{\mu}$, we can approximate the error conditional variance of

$$\widehat{\boldsymbol{V}} = \boldsymbol{V}_{\widehat{\boldsymbol{\mu}}}^{\boldsymbol{1/2}}\boldsymbol{V}^{*}\boldsymbol{V}_{\widehat{\boldsymbol{\mu}}}^{\boldsymbol{1/2}}.$$

Assuming that the error term has a normal distribution, we can specify the conditional distribution of the pseudo response of

$$\boldsymbol{P} \mid \boldsymbol{\beta}, \boldsymbol{v} \sim N\left[\boldsymbol{X\beta} + \boldsymbol{Zv},\ g'(\widehat{\boldsymbol{\mu}})\boldsymbol{V}_{\widehat{\boldsymbol{\mu}}}^{\boldsymbol{1/2}}\boldsymbol{V}^{*}\boldsymbol{V}_{\widehat{\boldsymbol{\mu}}}^{\boldsymbol{1/2}}g'(\widehat{\boldsymbol{\mu}})\right].$$

Now, suppose that the random effects are normally distributed $\boldsymbol{v} \sim N(\boldsymbol{0}, \boldsymbol{D})$. We can treat the pseudo model as a normal linear mixed effects model. Therefore, the marginal log likelihood function for the pseudo response could be written as

$$L(\boldsymbol{\beta}, \boldsymbol{V_P}, \boldsymbol{D}; \boldsymbol{P}) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{V_P}| - \frac{1}{2}(\boldsymbol{P} - X\boldsymbol{\beta})'\boldsymbol{V_P}^{\boldsymbol{-1}}(\boldsymbol{P} - X\boldsymbol{\beta}),$$

which could be maximized for $\boldsymbol{\beta}$, where

$$\boldsymbol{V_P} = \boldsymbol{ZDZ'} + \widehat{\boldsymbol{W}}^{\boldsymbol{-1}}\boldsymbol{V}\widehat{\boldsymbol{W}}^{\boldsymbol{-1}}.$$

If the covariance matrices for the conditional response and the random effect had dispersion parameters $\emptyset$, they could be estimated using the marginal variance $\boldsymbol{V_P}$. To estimate the dispersion parameters, Wolfinger and O’Connell (1993) derived the profile likelihood of

$$L(\emptyset;\boldsymbol{P}) = (-1/2)N\ln\left(\boldsymbol{r}'\boldsymbol{V}_{\boldsymbol{P}}^{-\mathbf{1}}\boldsymbol{r}\right) - (1/2)\ln|\boldsymbol{V}_{\boldsymbol{P}}| - (1/2)N[1+\ln(2\pi/N)],$$

where

$$\boldsymbol{r} = \boldsymbol{P} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{P}.$$

The estimates of the dispersion parameter $\emptyset$ could be done using numerical methods to estimate $\boldsymbol{V}$ and $\boldsymbol{D}$ from the profile likelihood through $\boldsymbol{V}_{\boldsymbol{P}}$, which gives

$$\widehat{\emptyset} = \hat{\boldsymbol{r}}'\,\widehat{\boldsymbol{V}}_{\boldsymbol{P}}\,\hat{\boldsymbol{r}}/n.$$

Using the profile likelihood estimates for $\widehat{\boldsymbol{V}}$ and $\widehat{\boldsymbol{D}}$, we could get simultaneous estimates for the parameters $\boldsymbol{\beta}$ and $\boldsymbol{v}$ by using the hierarchical joint log likelihood of the pseudo response and the random effect (Henderson, Kempthorne, Searle, & Krosigk,1959):

$$L(\boldsymbol{v},\boldsymbol{P}) \propto -\frac{1}{2}(\boldsymbol{P}-\boldsymbol{X\beta}-\boldsymbol{Zv})'\,\widehat{\boldsymbol{V}}^{-\mathbf{1}}(\boldsymbol{P}-\boldsymbol{X\beta}-\boldsymbol{Zv}) - \frac{1}{2}\boldsymbol{v}'\widehat{\boldsymbol{D}}^{-\mathbf{1}}\boldsymbol{v}.$$

The derivatives would give two score equations that could be solved using iterative least squares of

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{v}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{X} & \boldsymbol{X}'\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{Z}\\ \boldsymbol{Z}'\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{X} & \boldsymbol{Z}'\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{Z}+\widehat{\boldsymbol{D}}^{-\mathbf{1}}\end{bmatrix}^{-\mathbf{1}}\begin{bmatrix}\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{P}\\ \boldsymbol{Z}'\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{P}\end{bmatrix}.$$

Furthermore, the variance covariance matrix for the estimated fixed and random effects parameters could be estimated by the inverse of the Fisher information matrix:

$$\boldsymbol{H}^{-\mathbf{1}} = \begin{bmatrix}\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{X} & \boldsymbol{X}'\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{Z}\\ \boldsymbol{Z}'\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{X} & \boldsymbol{Z}'\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{Z}+\widehat{\boldsymbol{D}}^{-\mathbf{1}}\end{bmatrix}^{-1}.$$

The estimators of parameters using pseudo-likelihood are asymptotically consistent and normally distributed.

Penalized quasi-likelihood. In generalized linear mixed models, to estimate the parameters using maximum likelihood, we need to evaluate the full likelihood function under a hierarchical model. This function usually includes high order integration, which is hard to evaluate in closed form. The penalized quasi-likelihood is a method that approximates the likelihood function, which was proposed as a method of estimation in generalized linear mixed models by Breslow and Clayton (1993). To introduce this method of estimation, consider the generalized linear mixed model of

$$\boldsymbol{y}/\boldsymbol{u} \sim D_1[\boldsymbol{\mu}, \phi\, \boldsymbol{V}(\boldsymbol{\mu})],$$

$$\boldsymbol{u} \sim D_2[\boldsymbol{\mu_r}, \lambda\, \boldsymbol{V_r}(\boldsymbol{\mu_r})],$$

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zv},$$

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

where D_1 and D_2 are the distributions for the conditional response $\boldsymbol{y}/\boldsymbol{u}$ and the random effect $\boldsymbol{u}$, respectively, and $\boldsymbol{v} = g(\boldsymbol{u})$, has a distribution $\boldsymbol{v}{\sim}D_3(\boldsymbol{0}, \boldsymbol{D})$. For the above model, the integrated quasi likelihood function could be written as

$$e^{q(\boldsymbol{\beta},\boldsymbol{\theta})} \propto |\boldsymbol{D}|^{-1/2} \int exp\left[-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i\,, \mu_i) - \frac{1}{2}\boldsymbol{v}'\boldsymbol{D}^{-\mathbf{1}}\boldsymbol{v}\right] dv,$$

where $\boldsymbol{D}$ is the covariance matrix of the random effect $\boldsymbol{v}$, $\boldsymbol{\theta}$ is the canonical parameter, and $d(\boldsymbol{y}, \boldsymbol{\mu}) = -2\int_{\boldsymbol{y}}^{\boldsymbol{\mu}} \frac{\boldsymbol{y}-\boldsymbol{u}}{\phi v(\boldsymbol{\mu})} d\boldsymbol{u}$**.** Now, after writing the above integrated quasi likelihood function as $c|\boldsymbol{D}|^{-1/2} \int e^{-k(\boldsymbol{v})} d\boldsymbol{v}$, Breslow and Clayton applied Laplace's method for integral approximation to derive

$$e^{q(\boldsymbol{\beta},\boldsymbol{\theta})} \approx |\boldsymbol{D}|^{-1/2} e^{-k(\tilde{\boldsymbol{v}})} \frac{1}{\sqrt{k''(\tilde{\boldsymbol{v}})}},$$

where

$$k(\boldsymbol{v}) = \frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i\,,\mu_i) - \frac{1}{2}\boldsymbol{v}'\boldsymbol{D}^{-\mathbf{1}}\boldsymbol{v}$$

and $\tilde{\boldsymbol{v}}$ is the solution to $k'(\boldsymbol{v}) = 0$. From the previous definitions, we have

$$d_i(y_i, \mu_i) = \frac{(y_i - \mu_i)}{\phi\, V(\mu_i)}$$

and

$$\frac{\partial d_i(y_i, \mu_i)}{\partial v_i} = -\frac{(y_i - \mu_i)}{\phi\, V(\mu_i)}\ \frac{\partial \mu_i}{\partial v_i}.$$

It follows that

$$\mu_i = g^{-1}(\eta_i)$$

and

$$\frac{\partial \mu_i}{\partial v_i} = \frac{1}{g'[g^{-1}(\eta_i)]}\ \frac{\partial \eta_i}{\partial v_i} = \frac{1}{g'(\mu_i)}\ \frac{\partial \eta_i}{\partial v_i} = \frac{z_i}{g'(\mu_i)}.$$

Thus,

$$q(\boldsymbol{\beta}, \boldsymbol{\theta}) \approx -\tfrac{1}{2}log|\boldsymbol{D}| - \tfrac{1}{2}log|k''(\tilde{\boldsymbol{v}})| - k(\tilde{\boldsymbol{v}}),$$

where

$$k'(\boldsymbol{v}) = \frac{\partial\left[\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i\,,\mu_i) - \frac{1}{2}\boldsymbol{v}'\boldsymbol{D}^{-\mathbf{1}}\boldsymbol{v}\right]}{\partial \boldsymbol{v}} = -\sum\nolimits_{i=1}^{n} \frac{(y_i - \mu_i)z_i}{\phi v(\mu_i)g'(\mu_i)} + \boldsymbol{D}^{-\mathbf{1}}\boldsymbol{v}$$

and

$$k''(\boldsymbol{v}) = \quad -\sum\nolimits_{i=1}^{n}\{\left[\frac{1}{\phi V(\mu_i)g'(\mu_i)}\ \frac{\partial}{\partial \boldsymbol{v_i}}\ (y_i - \mu_i)z_i\right] + \left[(y_i - \mu_i)z_i\ \frac{\partial}{\partial \boldsymbol{v_i}}\ \frac{1}{\phi V(\mu_i)g'(\mu_i)}\right]\}$$
$$+ \boldsymbol{D}^{-\mathbf{1}}$$

$$= \quad \sum\nolimits_{1=1}^{n} \frac{z_i z_i'}{\phi V(\mu_i)[g'(\mu_i)]^2} + \boldsymbol{D^{-1}} - \sum\nolimits_{i=1}^{n} (y_i - \mu_i) z_i \frac{\partial}{\partial \boldsymbol{v_i}} \frac{1}{\phi V(\mu_i) g'(\mu_i)}$$

$$\approx \quad \boldsymbol{Z'WZ} + \boldsymbol{D^{-1}}.$$

The matrix $\boldsymbol{W}$ is an *n* x *n* diagonal matrix of the elements

$$w_i = \{\phi V(\mu_i)[g'(\mu_i)]^2\}^{-1}.$$

McCullagh and Nelder (1989) showed that for the canonical link function, $g'(\mu) = [V(\mu)]^{-1}$ was satisfied:

$$E\left[-\sum\nolimits_{i=1}^{n} (y_i - \mu_i) z_i \frac{\partial}{\partial \boldsymbol{v_i}} \frac{1}{\phi V(\mu_i) g'(\mu_i)}\right] = 0.$$

Now from the previous discussion and derivatives, we can write the approximation for the quasi likelihood function as

$$q(\boldsymbol{\beta}, \boldsymbol{\theta}) \approx -\tfrac{1}{2} log|\boldsymbol{D}| - \tfrac{1}{2} log\left|\boldsymbol{Z'WZ} + \boldsymbol{D^{-1}}\right| - \tfrac{1}{2\phi} \textstyle\sum_{i=1}^{n} d_i(y_i, \mu_i) - \tfrac{1}{2} \boldsymbol{v'D^{-1}v}$$

$$\approx -\frac{1}{2} log|\boldsymbol{I} + \boldsymbol{Z'WZ}| - \frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_i) - \frac{1}{2} \boldsymbol{\tilde{v}'D^{-1}\tilde{v}},$$

where $\boldsymbol{\tilde{v}}$ maximizes the sum of the last two terms $k(\boldsymbol{v})$. Breslow and Clayton (1993) assumed that the iterative weights varied slowly or were not a function of the mean at all so they ignored the first term: $-\frac{1}{2} log|\boldsymbol{I} + \boldsymbol{Z^T WZ}|$. Now choosing $\boldsymbol{\beta}$ that maximizes $\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_i) - \frac{1}{2} \boldsymbol{\tilde{v}'D^{-1}\tilde{v}}$, we have $\left(\boldsymbol{\hat{\beta}}, \boldsymbol{\hat{v}}\right) = \left(\boldsymbol{\hat{\beta}(\theta)}, \boldsymbol{\hat{v}(\theta)}\right)$, where $\boldsymbol{\hat{v}}(\theta) = \boldsymbol{\tilde{v}}(\boldsymbol{\hat{\beta}(\theta)})$.

These estimates jointly maximized the penalized quasi likelihood of

$$\tilde{q}(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_i) - \frac{1}{2} \boldsymbol{v'D^{-1}v}.$$

To estimate the parameters $(\boldsymbol{\beta}, \boldsymbol{v})$, we take the first derivative with respect to $\boldsymbol{\beta}$ and $\boldsymbol{v}$, respectively, which yield the following score equations:

$$\sum\nolimits_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{\emptyset V(\mu_i) g'(\mu_i)} = 0, \qquad 2.5$$

and

$$\sum\nolimits_{i=1}^{n} \frac{(y_i - \mu_i)z_i}{\emptyset V(\mu_i) g'(\mu_i)} - \boldsymbol{D}^{-\mathbf{1}}\boldsymbol{v} = 0. \qquad 2.6$$

As a solution for the above score equations, Green (1987) derived the Fisher scoring algorithm as a weighted least squares solution. Using first order Taylor's approximation for the link function at μ_i, Green defined the pseudo response vector:

$$\boldsymbol{P} = \boldsymbol{\eta} + (\boldsymbol{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu}),$$

with variance

$$\begin{aligned} \boldsymbol{V} = \quad & Var(\boldsymbol{P}) = Var(\boldsymbol{\eta}) + g'(\boldsymbol{\mu})\, Var(\boldsymbol{y})\, g'(\boldsymbol{\mu}) \\ = \quad & \boldsymbol{Z}\, var(\boldsymbol{v})\boldsymbol{Z}' + \boldsymbol{W}^{-\mathbf{1}} = \boldsymbol{ZDZ}' + \boldsymbol{W}^{-\mathbf{1}}, \end{aligned}$$

where $\boldsymbol{W}^{-\mathbf{1}}$ is a diagonal matrix of the terms, $\phi V(\mu_i)[g'(\mu_i)]^2$ and $Var(y_i) = \emptyset V(\mu_i)$.

Using the Fisher scoring algorithm, Breslow and Clayton (1993) derived the simultaneous solution to equations 2.5 and 2.6 as an iterative solution to the equations of

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{X} & \boldsymbol{D}^{-\mathbf{1}} + \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{W}\boldsymbol{P} \\ \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{P} \end{bmatrix}.$$

The estimates of parameters using this approach were asymptotically consistent and normally distributed.

Hierarchical likelihood. Hierarchical likelihood is a method of estimation that can be used to estimate the parameters in generalized linear mixed models. This technique was proposed by Lee and Nelder (1996) as an extension of the joint h-likelihood approach to estimate the parameters in the normal linear models with random effects (Henderson et al., 1959).

To introduce this method of estimation, consider the following hierarchical generalized linear mixed model:

$$\boldsymbol{y}/\boldsymbol{u} \sim D_1[\boldsymbol{\mu}, \phi\, V(\boldsymbol{\mu})],$$

$$\boldsymbol{u} \sim D_2[\boldsymbol{\mu_r}, \lambda\, V_r(\boldsymbol{\mu_r})],$$

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zv},$$

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

where D_1 is the distribution for the response, D_2 is an arbitrary distribution from the exponential family for the random component, and $\boldsymbol{v} = g(\boldsymbol{u})$ are the random effects, which is a function of $\boldsymbol{u}$ related to the response canonical link. Under this estimation technique, the distribution of the random effect is not necessarily normal distribution. Lee and Nelder (1996) defined the log hierarchical likelihood function as

$$h = L(\boldsymbol{\theta}, \phi; \boldsymbol{y}|\boldsymbol{v}) + L(\boldsymbol{\alpha}; \boldsymbol{v}),$$

where $\boldsymbol{\theta}$ and ϕ are the canonical and dispersion parameters, respectively, and $\boldsymbol{\alpha}$ is the parameters vector of the distribution of $\boldsymbol{v}$**,** considered as dispersion parameters.

To estimate the fixed and random effects parameters simultaneously, we can take the first derivative of h with respect to both fixed and random effects parameters:

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

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

The solution for these equations can be obtained using an iterative weighted least square approach. For example, the response variable has a binomial distribution and the random effects have a Beta distribution with parameters α_1 and α_2. In this case, we have

$$\boldsymbol{y}|\boldsymbol{u} \sim binomial\ (n\ ,\boldsymbol{p}),$$

$$\boldsymbol{u} \sim beta\ (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2),$$

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + g(\boldsymbol{u}),$$

$$\boldsymbol{\eta} = logit(\boldsymbol{p}).$$

For this model, the hierarchical log likelihood function can be written as

$$h = L(\boldsymbol{\theta}, \phi, \boldsymbol{\alpha}; \boldsymbol{y}, \boldsymbol{v}) = L(\boldsymbol{\theta}, \phi; \boldsymbol{y}|\boldsymbol{v}) + L(\boldsymbol{\alpha}; \boldsymbol{v}),$$

where $\boldsymbol{\theta} = \ln(\frac{\boldsymbol{p}}{1-\boldsymbol{p}})$ is the canonical parameter and $\boldsymbol{\phi}$ is the dispersion parameter for the binomial distribution.

Now, since the conditional distribution of the response given $\boldsymbol{v}$ is a binomial, the log likelihood function for the conditional response distribution is

$$L(\boldsymbol{\theta}, \phi; \boldsymbol{y}|\boldsymbol{v}) = \ln \binom{n}{y} + \boldsymbol{y} \ln\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) + n\ ln(1-\boldsymbol{p}).$$

Using the canonical parameter $\boldsymbol{\theta} = \ln(\frac{\boldsymbol{p}}{1-\boldsymbol{p}})$ and ignoring the constant ln $\binom{n}{\boldsymbol{y}}$ yield,

$$L(\boldsymbol{\theta}, \phi; \boldsymbol{y}|\boldsymbol{v}) = \boldsymbol{y}\ \boldsymbol{\theta} + n \text{ lin}\left(\frac{1}{1+e^{\boldsymbol{\theta}}}\right).$$

Also, by substituting the relation $\theta = x_{ij}\beta + v_i$ and taking the sum over all sample observation across the groups for the above function, we can rewrite the log likelihood function as

$$L(\boldsymbol{\theta}, \phi; \boldsymbol{y}|\boldsymbol{v}) = \sum_{i}^{g}\sum_{j}^{n_i}\{y_{ij}(x_{ij}\beta + v_i) + m_i \ln\left[\frac{1}{1+e^{x_{ij}\beta+v_i}}\right]\}.$$

Now, the random effect $\boldsymbol{u}$ has a beta distribution with parameters α_1 and α_2; thus, its log likelihood function for one observation can be written as

$$L(\alpha_1, \alpha_2; u_i) = -\ln B(\alpha_1, \alpha_2) + (\alpha_1 - 1)\ln u_i + (\alpha_2 - 1)\ln(1 - u_i).$$

Again, by using the relation $v_i = logit(u_i)$, we can write the log likelihood function over all observations as

$$L(\alpha_1, \alpha_2; v_i) = \sum_i^g \{-\ln B(\alpha_1, \alpha_2) + \alpha_1 v_i - (\alpha_1 + \alpha_1)\ln(1 + e^{v_i}) - \ln\left[\frac{e^{v_i}}{(1 + e^{v_i})^2}\right]\}.$$

Therefore, the log h-likelihood function for this model can be expressed as

$$L(\boldsymbol{\theta}, \phi, \boldsymbol{\alpha}; \boldsymbol{y}, \boldsymbol{v}) = \sum_i^g \sum_j^{n_i} \{ y_{ij}\left(x_{ij}\beta + v_i\right) + n_i \ln\left[\frac{1}{1+e^{x_{ij}\beta+v_i}}\right]\} +$$

$$\sum_i^g \{-\ln B(\alpha_1, \alpha_2) + \alpha_1 v_i - (\alpha_1 + \alpha_1)\ln(1 + e^{v_i}) - \ln\left[\frac{e^{v_i}}{(1 + e^{v_i})^2}\right]\}.$$

Using the log h-likelihood, we can get the estimates of the parameters vectors, $\boldsymbol{\beta}$ and $\boldsymbol{v}$, by solving the following score equations simultaneously:

$$\frac{\partial h}{\partial \beta_k} = \sum_i^g \sum_j^{n_i} \{y_{ij} x_{ijk} + n_i\, x_{ijk} \left[\frac{e^{x_{ij}\beta+v_i}}{1+e^{x_{ij}\beta+v_i}}\right]\} = 0 \quad ; \qquad k = 1, \ldots, p,$$

$$\frac{\partial h}{\partial v_l} = \sum_j^{n_l} \{y_{lj} + n_l \left[\frac{e^{x_{lj}\beta+v_l}}{1+e^{x_{lj}\beta+v_l}}\right] + \alpha_1 - \frac{1}{e^{v_l}} + \frac{(2 - \alpha_1 - \alpha_2)e^{v_l}}{(1 + e^{v_l})}\} \;; \quad l = 1, \ldots, q.$$

The above equations can be solved simultaneously by using the Fisher scoring algorithm. Furthermore, the augmented model can be used as an alternative procedure to estimate these parameters using iterative weighted least squares. The h-likelihood estimators are asymptotically efficient, consistent, and normally distributed.

Goodness of Fit Statistics

Hosmer and Lemeshow's test statistic. Hosmer and Lemeshow's (1980) test, which was discussed as a goodness of fit statistic for the logistic models, could also be used in mixed effects logistic models. It is a straightforward test to conduct in such models but choosing the number of groups is very subjective. Some studies showed that this test was very sensitive to the number of groups since the cut point for this test statistic depended on the number of groups. Furthermore, some studies indicated that the Hosmer and Lemeshow test might have low power for detecting a model departure of linearity because it only depended on grouping the response region and ignored the predictors region (Hosmer et al., 1997; le Cessie & van Houwelingen, 1991).

Statistics based on the estimated moments of Pearson and unweighted sum of squares statistics. Evans and Hosmer (2004) developed a goodness of fit test statistic based on tests used in the usual logistic regression models. They used first order Taylor approximations to estimate the mean and variance for both the Pearson and unweighted sum of squares statistics for the mixed effects logistic models. For the estimation procedure, they used the pseudo likelihood approach to estimate the parameters (Wolfinger & O'Connell, 1993). They considered the following mixed effects logistic regression model:

$$\boldsymbol{y}|\boldsymbol{v} \sim Bernoulli\ (\boldsymbol{p}),$$

$$\boldsymbol{v} \sim N\ (\boldsymbol{0}, \boldsymbol{D}),$$

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zv},$$

$$\boldsymbol{\eta} = logit(\boldsymbol{p}).$$

Under this approach, they assumed that the random effect $\boldsymbol{v}$ had $E(\boldsymbol{v}) = \mathbf{0}$ and $Cov(\boldsymbol{v}) = \boldsymbol{D}$, where $\boldsymbol{D}$ was assumed to be unknown. Also for the unobserved error vector $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{p}$, they assumed that $E(\boldsymbol{e}|\boldsymbol{v}) = \mathbf{0}$ and $Cov(\boldsymbol{e}|\boldsymbol{v}) = \boldsymbol{R}_{\boldsymbol{p}}^{\mathbf{1/2}}\boldsymbol{R}\,\boldsymbol{R}_{\boldsymbol{p}}^{\mathbf{1/2}}$, where $\boldsymbol{R}_{\boldsymbol{p}} = diag\left[p_{ij}\left(1 - p_{ij}\right)\right]$ and $\boldsymbol{R}$ is a matrix of unknown correlation parameters. Using this estimation technique, the pseudo response vector for the iterative procedure was

$$\boldsymbol{P} = (\boldsymbol{X\beta} + \boldsymbol{Zv}) + \boldsymbol{R}_{\widehat{\boldsymbol{p}}}^{-\mathbf{1}}(\boldsymbol{y} - \widehat{\boldsymbol{p}})$$

with a variance matrix of

$$\boldsymbol{V}_{\boldsymbol{P}} = \boldsymbol{R}_{\boldsymbol{p}}^{-\mathbf{1/2}}\boldsymbol{R}\,\boldsymbol{R}_{\boldsymbol{p}}^{-\mathbf{1/2}} + \boldsymbol{ZDZ}'.$$

The idea of their work was that both the Pearson and unweighted sum of squares statistics could be written in terms of the estimated error vector so the moments for these statistics could be obtained using the moments of the estimated error vector. To estimate the moments of the estimated error vector, they used a first order Taylor approximation to write the estimated error vector in terms of the actual error. To explain their work, they first used a first order approximation to write the estimated probabilities about the true mixed parameters:

$$\widehat{\boldsymbol{p}}(\boldsymbol{\gamma}) \approx \boldsymbol{p}(\boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}} + \frac{\partial}{\partial\boldsymbol{\gamma}}\boldsymbol{p}(\boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}), \qquad 2.7$$

where

$$\boldsymbol{\gamma} = \begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{v}\end{bmatrix}$$

is the true vector of fixed and random effects parameters. Let $\boldsymbol{U} = [\boldsymbol{X}\ \boldsymbol{Z}]$ be the design matrix for both fixed and random effects and assuming that $\boldsymbol{R} = \boldsymbol{I}$, we could write the estimation iterative equations under this estimation approach as

$$\boldsymbol{H}^{*}\,\widehat{\boldsymbol{\gamma}} = \boldsymbol{C},$$

where

$$\boldsymbol{H}^{*} \quad = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}_{\widehat{\boldsymbol{p}}}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}_{\widehat{\boldsymbol{p}}}\boldsymbol{X} \\ \boldsymbol{Z}'\boldsymbol{R}_{\widehat{\boldsymbol{p}}}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}_{\widehat{\boldsymbol{p}}}\boldsymbol{Z} + \widehat{\boldsymbol{D}}^{-\boldsymbol{1}} \end{bmatrix}$$

$$= \boldsymbol{U}'\boldsymbol{R}_{\widehat{\boldsymbol{p}}}\boldsymbol{U} + \boldsymbol{D}^{*},$$

where

$$\boldsymbol{D}^{*} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \widehat{\boldsymbol{D}}^{-\boldsymbol{1}} \end{bmatrix},$$

and

$$\boldsymbol{C} \quad = \begin{bmatrix} \boldsymbol{X}'\,\boldsymbol{R}_{\widehat{\boldsymbol{p}}}\,\boldsymbol{P} \\ \boldsymbol{Z}'\,\boldsymbol{R}_{\widehat{\boldsymbol{p}}}\,\boldsymbol{P} \end{bmatrix}$$

$$= \boldsymbol{U}'\,\boldsymbol{R}_{\widehat{\boldsymbol{p}}}\boldsymbol{P}.$$

Now, assume $S(\widehat{\boldsymbol{\gamma}}) = \boldsymbol{H}^{*}\widehat{\boldsymbol{\gamma}} - \boldsymbol{C} = 0$ so the first order Taylor approximation of S about the true parameter vector $\boldsymbol{\gamma}$ is

$$S(\widehat{\boldsymbol{\gamma}}) \approx S(\boldsymbol{\gamma})|_{\gamma=\gamma} + \frac{\partial}{\partial\boldsymbol{\gamma}}S(\boldsymbol{\gamma})|_{\gamma=\gamma}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}), \qquad 2.8$$

where

$$\begin{aligned} S(\boldsymbol{\gamma})|_{\gamma=\gamma} = \quad S(\boldsymbol{\gamma}) \quad &= \boldsymbol{H}^{*}\boldsymbol{\gamma} - \boldsymbol{C}(\boldsymbol{\gamma}) \\ &= \left(\boldsymbol{U}'\boldsymbol{R}_{\boldsymbol{p}}\boldsymbol{U} + \boldsymbol{D}^{*}\right)\boldsymbol{\gamma} - \boldsymbol{U}'\,\boldsymbol{R}_{\boldsymbol{p}}\left[\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{R}_{\boldsymbol{p}}^{-\boldsymbol{1}}(\boldsymbol{y} - \boldsymbol{p})\right] \\ &= \boldsymbol{D}^{*}\boldsymbol{\gamma} - \boldsymbol{U}'(\boldsymbol{y} - \boldsymbol{p}). \end{aligned}$$

Thus,

$$\frac{\partial}{\partial\boldsymbol{\gamma}}S(\boldsymbol{\gamma})|_{\gamma=\gamma} \quad = \quad \frac{\partial}{\partial\boldsymbol{\gamma}}[\boldsymbol{D}^{*}\boldsymbol{\gamma} - \boldsymbol{U}'(\boldsymbol{y} - \boldsymbol{p})]|_{\gamma=\gamma}$$

$$= \boldsymbol{D}^{*} - \boldsymbol{U}' \frac{\partial}{\partial \boldsymbol{\gamma}} (\boldsymbol{y} - \boldsymbol{p})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}}$$

$$= \boldsymbol{D}^{*} + \boldsymbol{U}' \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U}.$$

Putting the above expressions in equation 2.8, we can get

$$(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \approx [\boldsymbol{D}^{*} + \boldsymbol{U}' \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U}]^{-\mathbf{1}} [\boldsymbol{U}'(\boldsymbol{y} - \boldsymbol{p}) - \boldsymbol{D}^{*}\boldsymbol{\gamma}]. \qquad 2.9$$

Also, it can be shown as

$$\frac{\partial}{\partial \boldsymbol{\gamma}} \boldsymbol{p}(\boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}} = \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U}.$$

Thus, equation 2.7 could be written as

$$\widehat{\boldsymbol{p}} \approx \boldsymbol{p} + \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U} \left[\boldsymbol{D}^{*} + \boldsymbol{U}' \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U}\right]^{-\mathbf{1}} [\boldsymbol{U}'(\boldsymbol{y} - \boldsymbol{p}) - \boldsymbol{D}^{*}\boldsymbol{\gamma}].$$

According to these approximations, the estimated error vector could be written as

$$\begin{aligned} \widehat{\boldsymbol{e}} &= \boldsymbol{y} - \widehat{\boldsymbol{p}} \\ &\approx \boldsymbol{y} - \boldsymbol{p} + \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U} \left[\boldsymbol{D}^{*} + \boldsymbol{U}' \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U}\right]^{-\mathbf{1}} [\boldsymbol{U}'(\boldsymbol{y} - \boldsymbol{p}) - \boldsymbol{D}^{*}\boldsymbol{\gamma}] \\ &\approx (\boldsymbol{I} - \boldsymbol{H}_{\mathbf{1}})\boldsymbol{e} + \boldsymbol{k}, \end{aligned}$$

where

$$\boldsymbol{H}_{\mathbf{1}} = \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U} \left[\boldsymbol{D}^{*} + \boldsymbol{U}' \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U}\right]^{-\mathbf{1}} \boldsymbol{U}'$$

and

$$\boldsymbol{k} = \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U} \left[\boldsymbol{D}^{*} + \boldsymbol{U}' \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{U}\right]^{-\mathbf{1}} \boldsymbol{D}^{*}\boldsymbol{\gamma}.$$

To derive the moments for the Pearson statistic, this statistic could be written as

$$X^{2} = (\mathbf{1} - 2\widehat{\boldsymbol{p}})' \boldsymbol{R}_{\widehat{\boldsymbol{p}}}^{-\mathbf{1}} \widehat{\boldsymbol{e}} + n.$$

After substituting the above approximations for the estimated probabilities and errors and then taking the expected value and variance for the Pearson statistic expression, Evans and Hosmer (2004) approximated the moments

$$E(X^2) \approx \mathbf{1}'\boldsymbol{R}_{\hat{\boldsymbol{p}}}^{-\mathbf{1}}\boldsymbol{k} - \mathbf{2}\boldsymbol{p}'\boldsymbol{R}_{\hat{\boldsymbol{p}}}^{-\mathbf{1}}\boldsymbol{k} + \mathbf{2}\boldsymbol{k}'\boldsymbol{R}_{\hat{\boldsymbol{p}}}^{-\mathbf{1}}\boldsymbol{k} - 2\,trace\left[\boldsymbol{H}_{\mathbf{1}}'\boldsymbol{R}_{\hat{\boldsymbol{p}}}^{-\mathbf{1}}(\mathbf{1} - \boldsymbol{H}_{\mathbf{1}})\boldsymbol{R}_{\boldsymbol{p}}\right] + n,$$

and

$$Var(X^2) \approx (\mathbf{1} - \mathbf{2}\hat{\boldsymbol{p}})'\boldsymbol{R}_{\hat{\boldsymbol{p}}}^{-\mathbf{1}}(\mathbf{1} - \boldsymbol{H}_{\mathbf{1}})\boldsymbol{R}_{\boldsymbol{p}}(\mathbf{1} - \boldsymbol{H}_{\mathbf{1}})'\boldsymbol{R}_{\hat{\boldsymbol{p}}}^{-\mathbf{1}'}(\mathbf{1} - \mathbf{2}\hat{\boldsymbol{p}}).$$

Using the same approximations, they derived the moments for the unweighted sum of squares statistic. They approximated the distribution of Pearson statistic using two approximations: a normal distribution and a scaled chi square distribution:

$$X^2 \sim c\,\chi^2(d),$$

where

$$c = Var(X^2)/2\,E(X^2)$$

and

$$d = \; 2\,[E(X^2)]^2/Var(X^2).$$

Using the chi square distribution approximation, the hypothesis that the model was fit would be accepted if

$$X^2 \; < \; [Var(X^2)/2\,E(X^2)]\;\; \chi^2(d) \quad ; \quad d = \; 2\,[E(X^2)]^2/Var(X^2).$$

For the distribution of unweighted sum of squares statistic, Evans and Hosmer approximated it as a normal distribution based on a simulated distribution. Using a normal distribution approximation and the approximated moments of unweighted sum of squares statistic S, the test statistic could be written as

$$z = \frac{S - E(S)}{\sqrt{Var(S)}} \;\; \overset{\cdot}{\sim} \;\; N(0,1).$$

Therefore, the hypothesis that the model is fit would be accepted if

$$\frac{S - E(S)}{\sqrt{Var(S)}} \quad < \quad Z_{1-\alpha}.$$

Based on an extensive simulation study, Evans and Hosmer (2004) concluded that the unweighted sum of squares statistic and Pearson statistic using a scaled chi square distribution had proper type I error rates. They were recommended for use in such models when the cluster size was 100 or more observations. For models with only discrete covariates, they noted that these statistics would not give good results with respect to type I errors. Also, they recommended further research to determine the effect of the estimation techniques on these statistics because some estimation methods such as penalized quasi-likelihood, which was used in this study, could give biased estimates when cluster sizes were too small (Lin & Breslow, 1996).

A goodness of fit statistic based on smoothing the residuals. Recent work was done by Sturdivant and Hosmer (2007) to develop a new goodness of fit statistic for mixed effects logistic regression models. The idea of this test statistic was to apply the unweighted sum of squares statistic to the kernel smoothed residuals instead of the actual residuals of a fitted model. They used the SAS GLMMIX macro for the estimation procedure, which used the pseudo-likelihood approach.

Sturdivant and Hosmer (2007) derived the moments of the unweighted sum of squares of the smoothed residuals test statistic depending on some approximations; they assumed its distribution was a normal distribution. To introduce this test statistic, we know that the unweighted sum of squares test statistic could be written as

$$S = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i}^{n} (y_i - \hat{p}_i)^2.$$

Their test statistic was the same as the above statistic but instead of using the actual residuals, they used the smoothed residuals:

$$\hat{\boldsymbol{e}}_{\boldsymbol{s}} = \boldsymbol{\Lambda}\,\hat{\boldsymbol{e}},$$

where

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & & \lambda_{1n} \\ & \ddots & \\ \lambda_{n1} & & \lambda_{nn} \end{bmatrix}$$

was the matrix of the weights. Sturdivant and Hosmer used three kernel functions in smoothing the residuals: uniform, normal, and cubic kernel functions. For the uniform and the cubic kernel functions, the kernel weights λ_{ij} for the above matrix could obtained using the same formulas in the smoothed residual-based tests section. The kernel weights under the normal kernel function could be calculated as

$$\lambda_{ij} = \frac{K(\frac{\hat{p}_i - \hat{p}_j}{b})}{\sum_{j=1}^{n} K(\frac{\hat{p}_i - \hat{p}_j}{b})},$$

where b is the bandwidth and $K(.)$ is the normal kernel function:

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp(t^2/2) \quad ; \qquad -\infty < t < \infty.$$

In their study, Sturdivant and Hosmer used the kernel weights depending on the "y-space" because a recent study on the standard logistic models by Hosmer et al. (1997) showed no significant difference between using the "x-space" or the "y-space" to calculate the weights. However, they suggested that further research might investigate the difference between using "x-space" or the "y-space" for this statistic in hierarchical logistic models.

To derive the moments of the proposed test statistic, they used the same approximations as in the Evans and Hosmer (2004) study to write the estimated residuals in terms of the actual residuals. Using the same notation of the Evans and Hosmer study, the estimated residuals could be approximated as

$$\hat{\boldsymbol{e}} \approx (\boldsymbol{I} - \boldsymbol{H_1})\boldsymbol{e} + \boldsymbol{k},$$

where

$$\boldsymbol{H_1} = \boldsymbol{R_p U} \left[\boldsymbol{D}^* + \boldsymbol{U}' \boldsymbol{R_p U}\right]^{-\boldsymbol{1}} \boldsymbol{U}'$$

and

$$\boldsymbol{k} = \boldsymbol{R_p U} \left[\boldsymbol{D}^* + \boldsymbol{U}' \boldsymbol{R_p U}\right]^{-\boldsymbol{1}} \boldsymbol{D}^* \boldsymbol{\gamma_0}\,.$$

The proposed test statistic could be written as

$$S_s = \sum_{i=1}^{n} \hat{e}_{si}^2 = \hat{\boldsymbol{e}}_{\boldsymbol{s}}' \hat{\boldsymbol{e}}_{\boldsymbol{s}} = \hat{\boldsymbol{e}}' \boldsymbol{\Lambda}' \boldsymbol{\Lambda}\, \hat{\boldsymbol{e}}.$$

Substituting the above approximation of the estimated residuals yields,

$$\begin{aligned} S_s \;&\approx\; [(\boldsymbol{I} - \boldsymbol{H_1})\boldsymbol{e} + \boldsymbol{k}]' \,\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\, [(\boldsymbol{I} - \boldsymbol{H_1})\boldsymbol{e} + \boldsymbol{k}] \\ &\approx\; \boldsymbol{e}'(\boldsymbol{I} - \boldsymbol{H_1})' \,\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{H_1})\boldsymbol{e} + \boldsymbol{2k}' \,\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\, (\boldsymbol{I} - \boldsymbol{H_1})\boldsymbol{e} + \boldsymbol{k}' \,\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\, \boldsymbol{k}. \end{aligned}$$

Now by utilizing the properties of the quadratic forms and the linear combinations, the following approximations for the moments were derived:

$$\begin{aligned} E(S_s) \;&\approx E[(\boldsymbol{I} - \boldsymbol{H_1})\boldsymbol{e} + \boldsymbol{k}]' \,\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\, [(\boldsymbol{I} - \boldsymbol{H_1})\boldsymbol{e} + \boldsymbol{k}] \\ &\approx trace\left[(\boldsymbol{I} - \boldsymbol{H_1})' \,\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{H_1})\, \boldsymbol{R_p}\right] + \boldsymbol{k}' \,\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\, \boldsymbol{k}, \end{aligned}$$

and

$$\begin{aligned} Var(S_s) \;&\approx Var\; [(\boldsymbol{I} - \boldsymbol{H_1})\boldsymbol{e} + \boldsymbol{k}]' \,\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\, [(\boldsymbol{I} - \boldsymbol{H_1})\boldsymbol{e} + \boldsymbol{k}] \\ &\approx Var(\boldsymbol{e}'\boldsymbol{A}\,\boldsymbol{e}) + Var(\boldsymbol{b}'\boldsymbol{e}) + 2Cov(\boldsymbol{e}'\boldsymbol{A}\,\boldsymbol{e}, \boldsymbol{b}'\boldsymbol{e}), \end{aligned}$$

where

$$\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{H_1})'\,\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{H_1})$$

and

$$\boldsymbol{b} = \boldsymbol{2k}'\,\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\,(\boldsymbol{I} - \boldsymbol{H_1}).$$

Using the above approximations for the moments with the normal approximation for the test statistic, the final test statistic could be written as

$$Z_{S_s} = \frac{S_s - E(S_s)}{\sqrt{Var(S_s)}} \quad \dot{\sim} \quad N(0,1).$$

The estimators of the above approximated moments could be obtained by substituting the matrix $\boldsymbol{R_{\hat{p}}}$ instead of $\boldsymbol{R_p}$ in the above expressions. The hypothesis that the model was fit would be accepted if

$$\frac{S_s - E(S_s)}{\sqrt{Var(S_s)}} \quad < \quad Z_{1-\alpha}.$$

An extensive simulation study was done by Sturdivant and Hosmer (2007) to estimate the power of this test statistic along with the type I error rate. They concluded that this test statistic was recommended to check the goodness of fit in hierarchical logistic regression models because it gave very good rates of type I error. Also, it gave good power to detect departures in fixed effects with cluster sizes of 20 subjects per cluster. For the power to detect the departures in random effects, further study was needed. Also, they noted that the choice of the kernel density would not have any effect on the test statistic, while the choice of the bandwidth would have. The best bandwidth choice was not clear and further research was recommended. However, without further study, they recommended approximately ($\frac{1}{2}\sqrt{n}$) as a bandwidth when the cluster sizes were reasonable (20) and approximately ($\frac{1}{4}\sqrt{n}$) when the cluster sizes were too small. Furthermore, Sturdivant and Hosmer mentioned that smoothing the residuals over "x-space" might give good results in the context of power and type I error rate and improve the test statistic in some cases of cluster sizes and numbers.

From the previous review of the test statistics, I concluded that Hosmer and Lemeshow's (1980) test statistic is a very simple test to use but might not be appropriate for use in mixed effects logistic regression models because it is a slightly conservative test (Evans & Hosmer, 2004). The test statistic proposed by Evans and Hosmer (2004) is appropriate for use in such models but it needs a large sample of at least 100 observations within each cluster. The test statistic developed by Sturdivant and Hosmer (2007), which used the same idea of Evans and Hosmer on the smoothed residuals, is a very good test statistic because it gives good results for cluster sizes of 20 or more. However, it requires a smoothing method; thus, it is tedious to conduct without computer packages. Also, it has an issue of selecting the optimal bandwidth for the kernel function, which requires additional research. However, goodness of fit statistics that can be applied for small cluster sizes are not well developed for mixed effects logistic regression models.

CHAPTER III

GOODNESS OF FIT STATISTICS FOR MIXED EFFECTS LOGISTIC REGRESSION MODELS

In this chapter, goodness of fit statistics to test the model fit in the mixed effects logistic regression models are presented. The idea of estimating the moments of unweighted sum of squares and Pearson chi square statistics and then approximate their distributions as a normal or as a chi square distribution are used (Evans & Hosmer, 2004). The first test statistic utilized the residuals of the logit variable and the second test statistic was based on a transformation of the actual residuals. These test statistics used the pseudo-likelihood estimation technique.

Logit Residual Goodness of Fit Statistic

Most goodness of fit statistics that have been developed for mixed effects logistic models utilized the residuals of the estimated probabilities. These test statistics were designed to detect any change or departure in the model systematic component but this component was connected with these probabilities through the inverse of the link function. Some approximations on the inverse link are needed to find an expression to the model residuals and then evaluate their distribution. It might be more powerful if we connect the estimated residuals to the model equation without any approximations to evaluate the inverse link of the model equation. To introduce the idea of this test statistic, consider the following mixed effects logistic model:

$$\boldsymbol{y}|\boldsymbol{v}\sim\ Bernoulli\ (\boldsymbol{p}),$$

$$\boldsymbol{v}\sim\ i.i.d.N(\boldsymbol{0},\boldsymbol{D}),$$

$$\boldsymbol{\eta}=\boldsymbol{X\beta}+\boldsymbol{Zv},$$

$$\boldsymbol{\eta}=logit\ (\boldsymbol{p}), \qquad 3.1$$

where $\boldsymbol{v}$ is a vector of the random effect parameter and has a normal distribution with a covariance matrix $\boldsymbol{D}$. The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the design matrices for the fixed and random effects parameters, respectively, and $\boldsymbol{\beta}$ is a vector of the fixed effect parameter. The systematic component $\boldsymbol{\eta}$ is equated to the "$logit$" transformation of the probability vector $\boldsymbol{p}$.

The logit of the binary response is assumed to be a continuous variable with the following assumptions:

$$logit(E[\boldsymbol{y}|\boldsymbol{v}])=\boldsymbol{X\beta}+\boldsymbol{Zv}$$

and

$$Cov[logit(E[\boldsymbol{y}|\boldsymbol{v}])]=\boldsymbol{R}_{\boldsymbol{logit}(\boldsymbol{E}[\boldsymbol{y}|\boldsymbol{v}])}.$$

Now we can approximate the conditional distribution of the logit variable as

$$logit(E[\boldsymbol{y}|\boldsymbol{v}])\ \sim\ N\big(\ \boldsymbol{X\beta}+\boldsymbol{Zv}\ ,\boldsymbol{R}_{\boldsymbol{logit}(\boldsymbol{E}[\boldsymbol{y}|\boldsymbol{v}])}\ \big).$$

Thus, the conditional residuals of the logit variable could be written as

$$\begin{aligned}\boldsymbol{e}|\boldsymbol{v}\quad &=\quad logit(E[\boldsymbol{y}|\boldsymbol{v}])-E[logit(E[\boldsymbol{y}|\boldsymbol{v}])]\\ &=\quad logit(E[\boldsymbol{y}|\boldsymbol{v}])-\boldsymbol{U\gamma},\end{aligned}$$

where $\boldsymbol{U}=[\boldsymbol{X}\ \boldsymbol{Z}]$ and $\boldsymbol{\gamma}=\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{v}\end{bmatrix}$.

Accordingly, the distribution of the conditional logit residuals might be approximated as

$$\boldsymbol{e}|\boldsymbol{v}\ \ \dot{\sim}\ \ \boldsymbol{N}\big(\ \boldsymbol{0}\ ,\boldsymbol{R}_{\boldsymbol{logit}(\boldsymbol{y})}\ \big).$$

To derive the variance of the logit conditional residuals, let us write

$$\boldsymbol{e}|\boldsymbol{v} = \ln(\boldsymbol{y}) - \ln(1-\boldsymbol{y}) - \boldsymbol{U\gamma}.$$

Using the first order Taylor series approximations for both $\ln(\boldsymbol{y})$ and $\ln(1-\boldsymbol{y})$ around the true probability vector $\boldsymbol{p}$, we can write

$$\begin{aligned}\ln(\boldsymbol{y}) &\approx \ln(\boldsymbol{p}) + \frac{\partial \ln(\boldsymbol{y})}{\partial\, \boldsymbol{y}}|_{\boldsymbol{y}=\boldsymbol{p}}\ (\boldsymbol{y}-\boldsymbol{p}) \\ &\approx \ln(\boldsymbol{p}) + diag\left(\frac{1}{\boldsymbol{p}}\right)\ (\boldsymbol{y}-\boldsymbol{p}),\end{aligned}$$

and

$$\ln(1-\boldsymbol{y}) \approx \ln(1-\boldsymbol{p}) + \frac{\partial \ln(1-\boldsymbol{y})}{\partial\, \boldsymbol{y}}|_{\boldsymbol{y}=\boldsymbol{p}}\ (\boldsymbol{y}-\boldsymbol{p}).$$

Therefore, the approximated conditional residuals can be written as

$$\begin{aligned}\boldsymbol{e}|\boldsymbol{v} &\approx \ln(\boldsymbol{p}) - \ln(1-\boldsymbol{p}) + diag\left(\frac{1}{\boldsymbol{p}(1-\boldsymbol{p})}\right)\ (\boldsymbol{y}-\boldsymbol{p}) - \boldsymbol{U\gamma} \\ &\approx \ln(\boldsymbol{p}) - \ln(1-\boldsymbol{p}) + \boldsymbol{R}_{\boldsymbol{p}}^{-\boldsymbol{1}}(\boldsymbol{y}-\boldsymbol{p}) - \boldsymbol{U\gamma},\end{aligned}$$

where

$$\boldsymbol{R}_{\boldsymbol{p}} = diag[p_i(1-p_i)],$$

which is the weight matrix for the estimation procedure under this model.

Substituting $\widehat{\boldsymbol{\gamma}}$ instead of $\boldsymbol{\gamma}$ in the above conditional residuals expression, we can write

$$\hat{\boldsymbol{e}}|\boldsymbol{v} \quad \approx \ln(\boldsymbol{p}) - \ln(1-\boldsymbol{p}) + \boldsymbol{R}_{\boldsymbol{p}}^{-\boldsymbol{1}}(\boldsymbol{y}-\boldsymbol{p}) - \boldsymbol{U}\widehat{\boldsymbol{\gamma}}.$$

The conditional expected value for the above estimated residual vector is assumed to be zero. We can approximate its covariance matrix as

$$\begin{aligned} Cov(\hat{\boldsymbol{e}}|\boldsymbol{v}) &\approx \boldsymbol{R}_{\boldsymbol{p}}^{-\boldsymbol{1}} Cov(\boldsymbol{y})\, \boldsymbol{R}_{\boldsymbol{p}}^{-\boldsymbol{1}\prime} \\ &\approx \boldsymbol{R}_{\boldsymbol{p}}^{-\boldsymbol{1}} \boldsymbol{R}_{\boldsymbol{p}} \boldsymbol{R}_{\boldsymbol{p}}^{-\boldsymbol{1}\prime} \\ &\approx \boldsymbol{R}_{\boldsymbol{p}}^{-\boldsymbol{1}}. \end{aligned}$$

To estimate the above covariance matrix of the estimated conditional residual vector, we can substitute $\boldsymbol{R}_{\hat{\boldsymbol{p}}}$ instead of $\boldsymbol{R}_{\boldsymbol{p}}$ in the above result:

$$\widehat{Cov(\hat{\boldsymbol{e}}|\boldsymbol{v})} \approx \boldsymbol{R}_{\hat{\boldsymbol{p}}}^{-\boldsymbol{1}}.$$

Using the approximated estimated residuals, we can write the following unweighted sum of squares statistic:

$$S = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}.$$

Using the estimated moments of the estimated residuals, the estimated moments of this statistic could be derived from

$$\begin{aligned} E(S|\boldsymbol{v}) &= E(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}|\boldsymbol{v}) \\ &= trace[\widehat{Cov(\hat{\boldsymbol{e}}|\boldsymbol{v})}] + [E(\hat{\boldsymbol{e}}|\boldsymbol{v})]'[E(\hat{\boldsymbol{e}}|\boldsymbol{v})] \\ &= trace\left[\boldsymbol{R}_{\hat{\boldsymbol{p}}}^{-\boldsymbol{1}}\right] \end{aligned}$$

and

$$\begin{aligned} Var(S|\boldsymbol{v}) &= Var(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}|\boldsymbol{v}\ ) \\ &= 2\ trace[\widehat{Cov(\hat{\boldsymbol{e}}|\boldsymbol{v})}]^2 + 4\ [E(\hat{\boldsymbol{e}}|\boldsymbol{v})]'\ [\widehat{Cov(\hat{\boldsymbol{e}}|\boldsymbol{v})}][E(\hat{\boldsymbol{e}}|\boldsymbol{v})] \\ &= 2\ trace\left[\boldsymbol{R}_{\hat{\boldsymbol{p}}}^{-\boldsymbol{2}}\right]. \end{aligned}$$

Now if we use the normal distribution approximation for the approximated conditional residuals,

$$\boldsymbol{e}|\boldsymbol{v} \ \dot{\sim}\ N\left(\boldsymbol{0}\,,\ \boldsymbol{R}_{\boldsymbol{p}}^{-\boldsymbol{1}}\right),$$

we can approximate the distribution of this statistic as follows:

$$S \; \dot{\sim} \; c\,\chi^2_{(V)},$$

where

$$c = \frac{Var(S|\boldsymbol{v})}{2E(S|\boldsymbol{v})},$$

$$V = \frac{2[E(S|\boldsymbol{v})]^2}{Var(S|\boldsymbol{v})}.$$

According to this approximated test statistic, the hypothesis that the model is fit would be accepted if

$$S \; < \; \frac{Var(S|\boldsymbol{v})}{2E(S|\boldsymbol{v})} \; \chi^2_{(1-\alpha,V)} \quad ; \quad V = \frac{2[E(S|\boldsymbol{v})]^2}{Var(S|\boldsymbol{v})}.$$

Log-Transformed Residual Goodness of Fit Statistic

The estimated residuals of the mixed effects logistic regression model have values in the interval (-1,1). Therefore, the sum of squares of theses residuals would be small and could not be approximated as a chi square distribution with degrees of freedom equal to (n - # of parameters). In this section, a new fit statistic based on a transformation of the actual residuals of the model is introduced. The transformed residual variable was a continuous random variable on the interval (0,∞) with similar variability to the actual residuals. To introduce this test statistic, let us consider the mixed effects logistic regression model in Equation 3.1 and propose the following transformation of the estimated residual vector:

$$\boldsymbol{\varepsilon} = -\ln(1 - |\boldsymbol{e}|),$$

where

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{p} \qquad ; \qquad -1 < \boldsymbol{e} < 1.$$

Based on this transformation, the new residuals would be a continuous random variable on the interval (0,∞).

To derive the moments of the transformed residuals, a first order Taylor series approximation of the transformed residuals around the expected value of the absolute residuals was used. Using a Taylor series approximation, we can write

$$\boldsymbol{\varepsilon} \approx \ln\left(\frac{1}{1-E(|\boldsymbol{e}|)}\right) + \frac{\partial}{\partial|e|}\ln\left(\frac{1}{1-|\boldsymbol{e}|}\right)|_{|\boldsymbol{e}|=E(|\boldsymbol{e}|)}\ [\ |\boldsymbol{e}| - E(|\boldsymbol{e}|)\ ]$$

$$\approx \ln\left(\frac{1}{1-E(|\boldsymbol{e}|)}\right) + diag\left[\frac{1}{1-E(|\boldsymbol{e}|)}\right]\ [|\boldsymbol{e}| - E(|\boldsymbol{e}|)\ ].$$

Now if we substitute $\hat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{p}}$ instead of $\boldsymbol{e}$ in the above expression, we can write

$$\hat{\boldsymbol{\varepsilon}} \approx \ln\left(\frac{1}{1-E(|\hat{\boldsymbol{e}}|)}\right) + diag\left[\frac{1}{1-E(|\hat{\boldsymbol{e}}|)}\right]\ [|\hat{\boldsymbol{e}}| - E(|\hat{\boldsymbol{e}}|)\ ].$$

Thus, we can derive the moments of the estimated transformed residuals as

$$E(\hat{\boldsymbol{\varepsilon}}) \approx \ln\left(\frac{1}{1-E(|\hat{\boldsymbol{e}}|)}\right)$$

and

$$Cov(\hat{\boldsymbol{\varepsilon}}) \approx \left(diag\left[\frac{1}{1-E(|\hat{\boldsymbol{e}}|)}\right]\right)\ Var(|\hat{\boldsymbol{e}}|)\ \left(diag\left[\frac{1}{1-E(|\hat{\boldsymbol{e}}|)}\right]\right)'.$$

The above approximated moments will depend on the probability vector $\boldsymbol{p}$ and the weight matrix $\boldsymbol{R_p}$. Therefore, we can get estimates of the above moments by substituting $\widehat{\boldsymbol{p}}$ and $\boldsymbol{R}_{\widehat{\boldsymbol{p}}}$ in their expressions. The unweighted sum of squares of the estimated transformed residuals would be

$$S = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}.$$

If we use the estimates of the transformed residuals moments, we can derive the estimates of the unweighted sum of squares statistic as

$$\begin{aligned} E(S) &= E(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) \\ &= trace[\widehat{Var(\hat{\boldsymbol{\varepsilon}})}] + [\widehat{E(\hat{\boldsymbol{\varepsilon}})}]'\big[\widehat{E(\hat{\boldsymbol{\varepsilon}})}\big] \end{aligned}$$

and

$$\begin{aligned} Var(S) &= Var(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}\ ) \\ &= 2\ trace[\widehat{Var(\hat{\boldsymbol{\varepsilon}})}]^2 + 4\ [\widehat{E(\hat{\boldsymbol{\varepsilon}})}]'\ [\widehat{Var(\hat{\boldsymbol{\varepsilon}})}]\big[\widehat{E(\hat{\boldsymbol{\varepsilon}})}\big]. \end{aligned}$$

The distribution of the unweighted sum of squares statistic might be approximated as a normal or as a chi square distribution. However, initial simulation studies showed that the distribution of the unweighted sum of squares statistic, using this transformation, was approximated as a normal distribution. Using the normal distribution approximation, we can write

$$Z = \frac{S - E(S)}{Var(S)} \quad \dot{\sim} \quad N(0,1).$$

Therefore, the hypothesis that the model is fit would be accepted if

$$\frac{S - E(S)}{Var(S)} \quad < \quad Z_{1-\alpha}.$$

For small sample sizes, we could use the *t* distribution.

Sum of Absolute Residuals Goodness of Fit Statistic

In this section, a new goodness of fit statistic for the mixed effect logistic models is introduced. The statistic is simply the sum of the absolute residuals of the logistic model. Assume that the estimated probabilities are a uniform random variable over the interval (0,1). Under this circumstance, simulation studies showed that the distribution of the residuals of a logistic model could be approximated as a normal distribution:

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R}),$$

where $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{p}$ and $\boldsymbol{R} = diag[p_i(1 - p_i)]$. Using these assumptions, the absolute value of the residuals would have a half or folded normal distribution with mean and variance:

$$E(|\boldsymbol{e}|) = \boldsymbol{R}^{1/2}(\boldsymbol{2}/\boldsymbol{\pi})^{1/2},$$

$$Var(|\boldsymbol{e}|) = \boldsymbol{R}\,[\boldsymbol{I} - diag(\boldsymbol{2}/\boldsymbol{\pi})].$$

If use the Pearson residuals

$$\boldsymbol{e}^* = \boldsymbol{R}^{-1/2}\,\boldsymbol{e},$$

the expected value and the variance for the absolute Pearson's residuals could be derived:

$$E(|\boldsymbol{e}^*|) = (\boldsymbol{2}/\boldsymbol{\pi})^{1/2},$$

$$Var(|\boldsymbol{e}^*|) = [\boldsymbol{I} - diag(\boldsymbol{2}/\boldsymbol{\pi})].$$

Using the Pearson residuals, the sum of absolute residuals goodness of fit statistic could be defined as

$$S = \sum_{i=1}^{n} |\boldsymbol{e}^*|.$$

Thus, the mean and variance of this statistic could be written as

$$\widehat{E(S)} = \boldsymbol{1}'(\boldsymbol{2}/\boldsymbol{\pi})^{1/2} = n\sqrt{(2/\pi)},$$

$$\widehat{Var(S)} = trace[\boldsymbol{I} - diag(2/\pi)] = n\,[1 - (2/\pi)].$$

Therefore, we could write the distribution of this statistic as

$$S \sim D[E(S), Var(S)].$$

However, simulations studies showed that the distribution of this test statistic could be approximated as a normal distribution. Using this approximation, we can write

$$S \sim N[E(S), Var(S)].$$

To use this test statistic to test the goodness of fit of a logistic model, we can use the standardized version:

$$Z_{cal} = \frac{S - E(S)}{\sqrt{Var(S)}}.$$

Using this test statistic, the hypothesis that a logistic model is fit would be accepted if

$$Z_{above} < Z_{1-\alpha}.$$

In Chapter IV, a simulation study is conducted to investigate the performance of these test statistics and answer the following research questions:

Q1 What is the sampling distribution of the logit residual goodness of fit statistic?

Q2 What is the sampling distribution of the log-transformed residual goodness of fit statistic?

Q3 What is the sampling distribution of the absolute residual goodness of fit statistic?

Q4 Do the proposed goodness of fit statistics have greater power than existing goodness of fit statistics for small cluster sizes?

Q5 Do the proposed goodness of fit statistics have a proper type I error rate?

The performance was evaluated according to the type I error rate and the power of each test statistic. Also, the proposed test statistics were compared with the test statistics proposed by Sturdivant and Hosmer (2007).

Data for the mixed effects logistic models were generated using the following fixed effect predictors:

- One continuous predictor
- One continuous and one categorical predictors
- Two continuous and one categorical predictors

- Two continuous and one binary predictors

The random effects variable was generated to follow a normal distribution. However, the data were generated over different cluster sizes and number of clusters for each test statistic and model's equation.

To fit the generated data over all cases of cluster sizes, number of clusters, and model's systematic component, the pseudo-likelihood approach was used (Wolfinger & O'Connell, 1993).

The power of these test statistics was investigated with respect to the fixed effects component. To examine the power of these statistics, one or two predictors were generated and were used in fitting a wrong model over the replications. For example, data were generated under the following model,

$$logit(p_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + v_i,$$

and the following wrong models were fitted:

$$logit(p_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + v_i,$$

$$logit(p_{ij}) = \beta_0 + \beta_1 X_{1ij} + v_i.$$

The power represented the proportion of rejecting a wrong model over the replications.

The results of type I error rate and the power are presented for each model and test statistic.

CHAPTER IV

SIMULATION AND DATA ANALYSIS

In this chapter, the results of simulation studies applied to the proposed goodness of fit statistics are presented. These studies were conducted to examine the performance of the proposed goodness of fit statistics and then answer the following research questions:

Q1 What is the sampling distribution of the logit residual goodness of fit statistic?

Q2 What is the sampling distribution of the log-transformed residual goodness of fit statistic?

Q3 What is the sampling distribution of the absolute residual goodness of fit statistic?

Q4 Do these proposed goodness of fit statistics have greater power than existing goodness of fit statistics for small cluster sizes?

Q5 Do these proposed goodness of fit statistics have proper type I error rate?

Simulation Study

Simulation studies were conducted to investigate the performance of the proposed goodness of fit statistics. Data for mixed effect logistic models were generated according to different model equations. The suggested model equations included one continuous predictor, one continuous and one categorical predictor, two continuous and one categorical predictor, and a model of two continuous and one binary predictor.

1.

$$y_{ij}|v_i \sim Bernoulli(p_{ij})$$

$$v_i \sim N(0,1)$$

$$logit(p_{ij}) = x_{2ij} + v_i,$$

where x_{2ij} is a continuous predictor and v_i is the random effect.

2.

$$y_{ij}|v_i \sim Bernoulli(p_{ij})$$

$$v_i \sim N(0,1)$$

$$logit(p_{ij}) = x_{2ij} + 0.5\, x_{3ij} + v_i,$$

where x_{2ij} is a continuous predictor, x_{3ij} is a categorical predictor, and v_i is the random effect.

3.

$$y_{ij}|v_i \sim Bernoulli(p_{ij})$$

$$v_i \sim N(0,1)$$

$$logit(p_{ij}) = x_{1ij} + x_{2ij} + 0.5\, x_{3ij} + v_i,$$

where x_{1ij} is a continuous predictor, x_{2ij} is a continuous predictor, x_{3ij} is a categorical predictor, and v_i is the random effect.

4.

$$y_{ij}|v_i \sim Bernoulli(p_{ij})$$

$$v_i \sim N(0,1)$$

$$logit(p_{ij}) = x_{1ij} + x_{2ij} - \ x_{4ij} + v_i,$$

where x_{1ij} is a continuous predictor, x_{2ij} is a continuous predictor, x_{4ij} is a binary predictor, and v_i is the random effect.

First, the random effect's variable (v) is generate to follow a normal distribution; the conditional response mean, which is probabilities (p_{ij}), was generated to be related to the generated random effect's variable. According to initial simulation studies, the residuals of the mixed effect logistic models can have an approximately normal distribution if the aggregation of the generated probabilities over the clusters or the subjects have an approximately uniform distribution on the interval (0,1). Thus, to make the assumption of normality hold for some of the proposed test statistics, the probabilities were generated to have an approximated uniform distribution. Second, a categorical predictor's variable (x_3)--takes values 1, 2 and 3--was generated using the random uniform variable

$$x_3 = \mathrm{int}\big((\mathrm{ranuni(seed)} * \mathbf{3}) + \mathbf{1}\big).$$

Also, a binary predictor's variable (x_4)--takes the values 0 or 1--was generated to follow a $Bernoulli(1/2)$. In addition, a continuous predictor (x_1) was generated to follow a normal distribution with zero mean and variance equal to 2. To generate the second continuous variable (x_2), the "$logit$" transformation of the generated probabilities was calculated. Thus x_2 is generated for each model equation using the following expressions,

1. $x_{2ij} = logit\big(p_{ij}\big) - v_i.$

2. $x_{2ij} = logit\big(p_{ij}\big) - [0.5\, x_{3ij} + v_i].$

3. $x_{2ij} = logit(p_{ij}) - [\, x_{1ij} + 0.5\, x_{3ij} + v_i]$.

4. $x_{2ij} = logit(p_{ij}) - [\, x_{1ij} - x_{4ij} + v_i]$.

The response variable (y), which takes the value 0 or 1, was generated using a random Bernoulli variable with respect to the generated probabilities (p). The generated data were generated over different cluster sizes (4, 10, 20, 40, and 80) and different numbers of clusters (10, 20, 25, and 50).

The pseudo-likelihood approach (Wolfinger & O'Connell, 1993) was used to fit the above correct models over all cases and the proportion of type I error was calculated for each proposed test statistic for all cases of models, number of clusters, and cluster sizes.

To examine the power of the proposed goodness of fit statistics over all cases of models, number of clusters, and cluster sizes, some incorrect models for the generated data of the above models were fitted and the proportions of rejected incorrect models were calculated for all cases. The fitted wrong models included the following systematic components:

1. $logit(p_{ij}) = x_{2ij}$ {To detect $[v]$ out of the model (1)}.
2. $logit(p_{ij}) = 0.5\, x_{3ij} + v_i$ {To detect $[x_2]$ out of the model (2)}.
3. $logit(p_{ij}) = x_{2ij} + v_i$ {To detect $[x_1]$ and $[x_3]$ out of the model (3)}.
4. $logit(p_{ij}) = x_{2ij} - x_{4ij} + v_i$ {To detect $[x_1]$ out of the model (4)}.

Logit Residual Goodness of Fit Statistic

In this section, the results of some simulation studies applied to the logit residual goodness of fit statistic are presented. For the logit residual goodness of fit statistic that was introduced in Chapter III,

$$S = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$

$$\hat{\boldsymbol{e}} \approx \ln(\boldsymbol{p}) - \ln(1 - \boldsymbol{p}) + \boldsymbol{R}_{\boldsymbol{p}}^{-\mathbf{1}}(\boldsymbol{y} - \boldsymbol{p}) - \boldsymbol{U}\widehat{\boldsymbol{\gamma}},$$

where

$$\widehat{E(S)} \quad = trace\left[\boldsymbol{R}_{\widehat{\boldsymbol{p}}}^{-\mathbf{1}}\right] \qquad 4.1$$

and

$$\widehat{Var(S)} \quad = 2\ trace\left[\boldsymbol{R}_{\widehat{\boldsymbol{p}}}^{-\mathbf{2}}\right]. \qquad 4.2$$

The simulation results showed that the sampling distribution of this fit statistic could be approximated as a chi square distribution. However, the empirical variance did not match the estimated variance using the above expression to estimate the variance.

Table 1 shows some simulated values of p-value of this test statistic, observed statistic, mean, and variance calculated according to equations 4.1 and 4.2 of this fit statistic under the correct fitted model:

$$logit(p_{ij}) = x_{1ij} + x_{2ij} + 0.5\ x_{3ij} + v_i.$$

Table 1

A Sample of Simulated Values of p-Value, Observed Statistic, and Moments of the Logit Residual Goodness of Fit Statistic under the True Model, $logit(p_{ij}) = x_{1ij} + x_{2ij} + 0.5\ x_{3ij} + v_i$

Obs.	P-Value	Statistic	Mean	Variance
1	0.70182	5336.05	5497.79	88933.73
2	0.98143	4873.41	5447.95	80434.81
3	0.86572	5220.23	5546.09	87457.15
4	0.94490	4577.87	4966.50	61193.38
5	0.99297	4558.80	5181.70	69053.84
6	0.95497	4608.54	5023.98	62366.04
7	0.96905	5182.35	5734.12	91629.65
8	0.98969	4618.78	5208.40	69214.60
9	0.94489	4510.59	4886.00	57069.33
10	0.99706	4347.21	5016.52	64085.81
11	0.97087	4858.04	5389.70	82841.28
12	0.99031	4480.25	5073.66	68900.38
13	0.88161	4941.71	5251.26	69295.03
14	0.99998	5125.09	6852.27	70587.41
15	0.93829	4468.40	4830.13	56821.11

Using 10000 replications and the true model, $logit(p_{ij}) = x_{1ij} + x_{2ij} + 0.5\, x_{3ij} + v_i$, the empirical mean for this test statistic was 4817.59 and the empirical variance was 151731.88, which is larger than the simulated values of the variance in Table 1. Therefore, equation 4.2 always gives small estimates of the variance. However, this expression was derived by using first order Taylor series approximations. Further research might be needed to approximate or adjust this expression to estimate the variance of this proposed goodness of fit statistic using alternative or higher order approximation.

Log-Transformed Residual Goodness of Fit Statistic

The log-transformed residual test statistic was introduced in Chapter III,

$$S = \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}$$

$$\hat{\boldsymbol{\varepsilon}} = -\ln(1 - |\hat{\boldsymbol{e}}|),$$

with estimated moments

$$\widehat{E(S)} = trace[\widehat{Var(\hat{\boldsymbol{\varepsilon}})}] + [\widehat{E(\hat{\boldsymbol{\varepsilon}})}]'[\widehat{E(\hat{\boldsymbol{\varepsilon}})}] \qquad 4.3$$

and

$$\widehat{Var(S)} = 2\, trace[\widehat{Var(\hat{\boldsymbol{\varepsilon}})}]^2 + 4\, [\widehat{E(\hat{\boldsymbol{\varepsilon}})}]'\, [\widehat{Var(\hat{\boldsymbol{\varepsilon}})}][\widehat{E(\hat{\boldsymbol{\varepsilon}})}]. \qquad 4.4$$

The simulation results for this goodness of fit statistic showed that the sampling distribution of this statistic could be approximated as a normal distribution. However, the empirical mean of this fit statistic is always larger than the estimated mean using equations 4.3. The reason might be the first order Taylor series approximation that was applied to derive this expression.

Table 2 shows some simulated values of p-value for the fit statistic, observed statistic, and moments of log-transformed residual test statistic under the true fitted model:

$$logit(p_{ij}) = x_{1ij} + x_{2ij} + 0.5\ x_{3ij} + v_i.$$

Table 2

A Sample of Simulated Values of P-Value, Observed Statistic, and Moments of the Log-Residual Goodness of Fit Statistic under the True Model, $logit(p_{ij}) = x_{1ij} + x_{2ij} + 0.5\ x_{3ij} + v_i$

Obs.	P-Value	Statistic	Mean	Variance
1	2.844E-11	438.201	344.479	204.634
2	1.602E-10	435.747	345.329	206.734
3	1.676E-14	435.092	329.940	192.238
4	0	432.407	288.820	155.041
5	0	432.761	313.881	177.686
6	2.0586E-8	429.739	349.718	212.781
7	0	436.444	312.721	174.949
8	0	432.055	313.387	179.884
9	8.966E-13	434.070	334.335	200.151
10	0	433.474	311.136	177.098
11	1.31E-14	433.182	327.523	192.441
12	3.331E-16	434.032	324.345	184.365
13	0	424.678	306.723	174.408
14	3.829E-10	431.892	343.416	206.842
15	1.11E-16	436.630	324.986	186.047

Using the same correct model $logit\left(p_{ij}\right) = x_{1ij} + x_{2ij} + 0.5\, x_{3ij} + v_i$ and 10000 replications, the empirical mean and variance of this test statistic were 400.005 and 190.85, respectively. That means the simulated means in Table 2, using equation 4.3 to estimate the mean, is always smaller than the empirical mean of this fit statistic. However, approximations or adjustments might solve this issue in future research.

Absolute Residual Goodness of Fit Statistic

The absolute residual goodness of fit statistic was introduced in Chapter III:

$$S = \sum_{i=1}^{n} |\boldsymbol{e}^{*}|$$

$$\boldsymbol{e}^{*} = \boldsymbol{R}^{-1/2}\, \boldsymbol{e},$$

with estimated moments of

$$\widehat{E(S)} = n\,\sqrt{(2/\pi)}$$

$$\widehat{Var(S)} = n[1 - (2/\pi)].$$

The simulation results of the absolute residual goodness of fit statistic showed the distribution of this fit statistic could be consistently approximated as a normal distribution for most cases of cluster sizes, number of clusters, and model equations. However, for very small cluster sizes of four observations per cluster and 10 clusters, the sampling distribution of this statistic could not be approximated as a normal distribution. The simulation was applied over all cases. It gave good results in terms of type I error rate and the power of this goodness of fit statistic in most cases.

Tables 3 and 4 represent the empirical type I error rate and power for this fit statistic, respectively, using generated data under model 1. The empirical power values were calculated by omitting the random effect v of the model statement while generating the data under true model 1. The results of type I error rates in Table 3 and Figure 1 are good for some cases of cluster sizes and number of clusters, especially when the cluster size is 20 or more observations and the number of clusters is 10 or more. Also, the empirical power values for this test statistic in Table 4 and Figure 2 demonstrate this test statistic has power of 100% to detect omitting the random effect variable (v) of the model equation for number of clusters of 20 or more and cluster sizes of four or more. However, when the number of clusters is 10 each of four observations, the distribution of this goodness of fit statistic cannot be approximated as a normal.

Table 3

Type I Error Rate Under the Correct Model, $logit(p_{ij}) = x_{2ij} + v_i$

		Number of Clusters			
		10	**20**	**25**	**50**
Cluster sizes	**4**	0.0822	0.0523	0.0605	0.0552
	10	0.0603	0.0542	0.0529	0.0535
	20	0.0565	0.0535	0.0525	0.0533
	40	0.0533	0.0521	0.0516	0.0527
	80	0.0540	0.0513	0.0510	0.0509

Table 4

Power for Detecting the Random Effect (v) Out of the Model, $logit(p_{ij}) = x_{2ij} + v_i$

		Number of Clusters			
		10	**20**	**25**	**50**
Cluster sizes	**4**	0.5130	0.9810	1.0000	1.0000
	10	1.0000	1.0000	1.0000	1.0000
	20	1.0000	1.0000	1.0000	1.0000
	40	1.0000	1.0000	1.0000	1.0000
	80	1.0000	1.0000	1.0000	1.0000

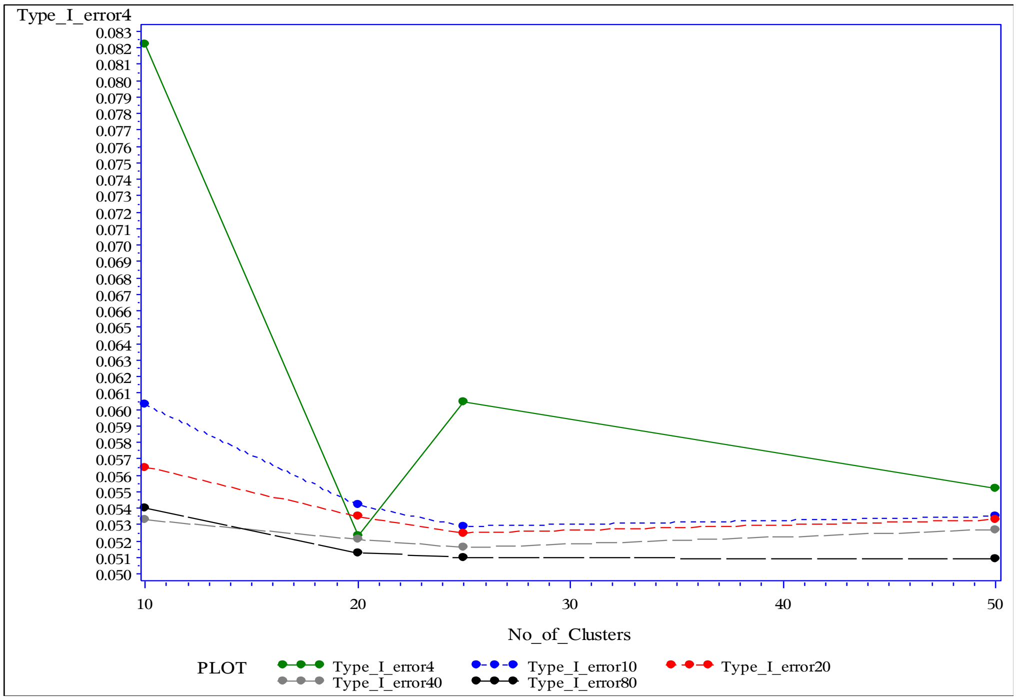

Figure 1. Plot of type I error for different cluster sizes against the number of clusters using the model equation, $logit(p_{ij}) = x_{2ij} + v_i$.

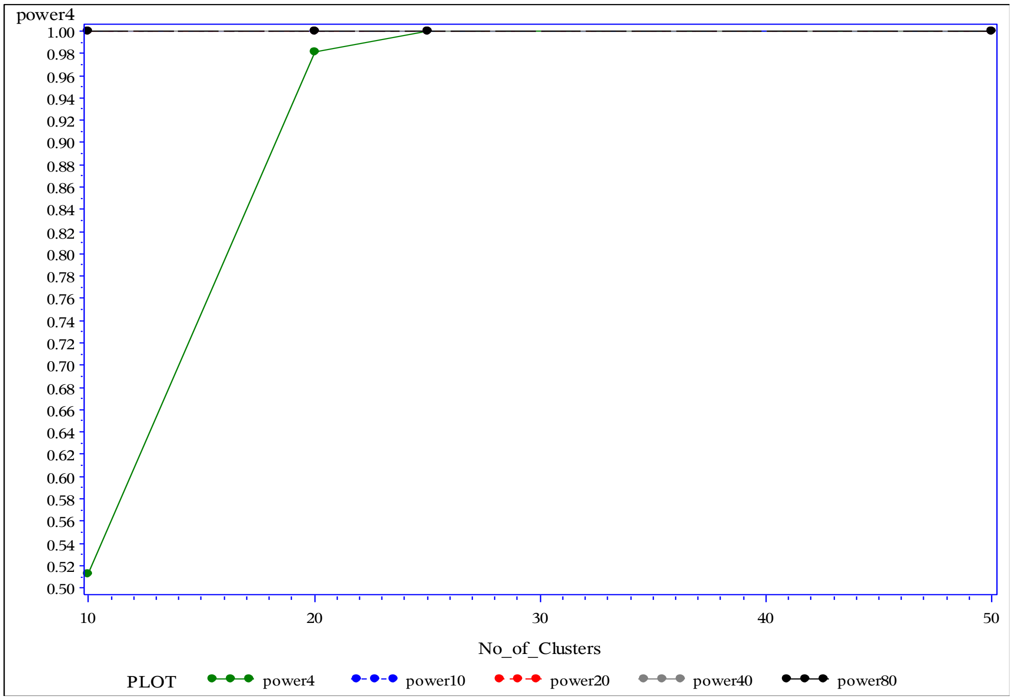

Figure 2. Plot of power to detect v, for different cluster sizes against the number of clusters using the model equation, $logit(p_{ij}) = x_{2ij} + v_i$

Tables 5 and 6 show the empirical type I error rate and power for this fit statistic, respectively, using generated data under model 2. Using model 2, the results of the empirical type I error rates in Table 5 and Figure 3 are very proper as long as the cluster size is 10 observations or more. Also, the results of power in Table 6 and Figure 4 to detect the fixed effect predictor (x_2) demonstrate this fit statistic has good power of about 60% for moderate sizes of number of clusters and cluster sizes. For cluster sizes and number of clusters of 20 or more, the results of power are very good (0.801 – 1.000) and tend to be 100% for large samples. The empirical power values in Table 6 were calculated by omitting the fixed effect predictor (x_2) of the model statement with generating the data under true model 2.

Table 5

Type I Error Rate Under the True Model, $logit(p_{ij}) = x_{2ij} + 0.5\ x_{3ij} + v_i$

		Number of Clusters			
		10	**20**	**25**	**50**
Cluster sizes	**4**	0.0752	0.0515	0.0521	0.0476
	10	0.0520	0.0510	0.0518	0.0531
	20	0.0484	0.0495	0.0523	0.0519
	40	0.0520	0.0516	0.0512	0.0504
	80	0.0505	0.0511	0.0499	0.0521

Table 6

Power for Detecting x_2 *Out of the Model,* $logit(p_{ij}) = x_{2ij} + 0.5\ x_{3ij} + v_i$

		Number of Clusters			
		10	**20**	**25**	**50**
Cluster sizes	**4**	0.1773	0.3344	0.3460	0.5021
	10	0.6234	0.7967	0.8954	0.9903
	20	0.5948	0.8010	0.8413	0.9972
	40	0.8147	0.9924	0.9985	1.0000
	80	1.0000	1.0000	1.0000	1.0000

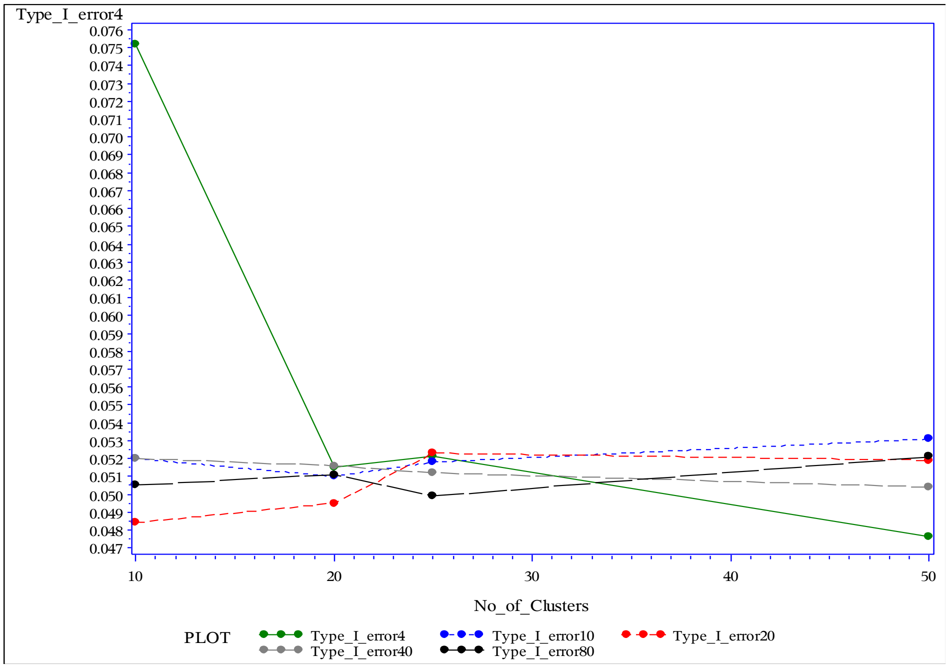

Figure 3. Plot of type I error for different cluster sizes against the number of clusters, using the model equation, $logit(p_{ij}) = x_{2ij} + 0.5\ x_{3ij} + v_i$.

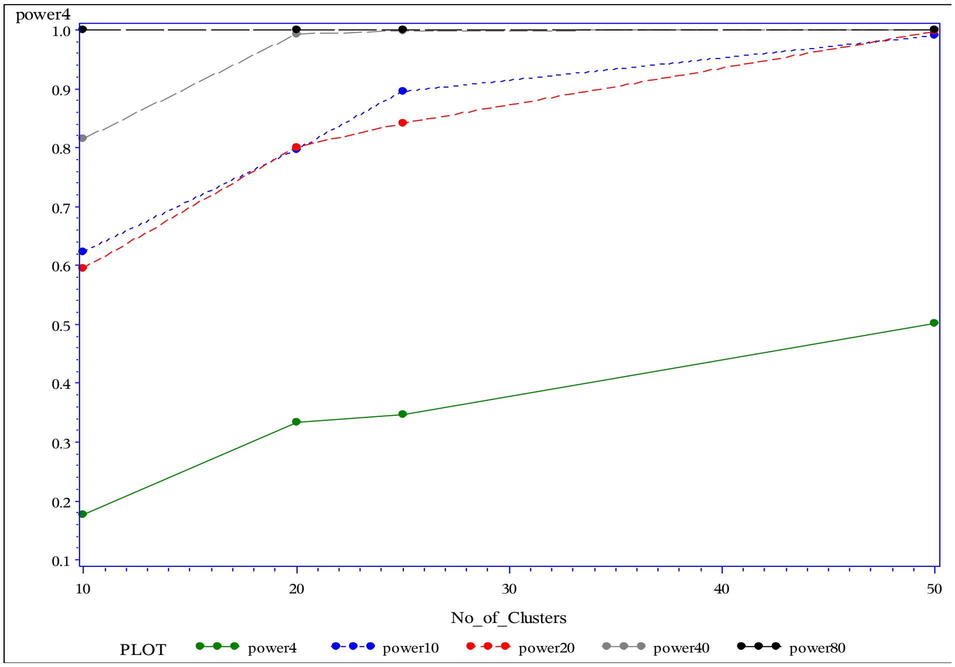

Figure 4. Plot of power to detect x_2, for different cluster sizes against the number of clusters, using the model equation,$logit(p_{ij}) = x_{2ij} + 0.5\ x_{3ij} + v_i$.

Tables 7 and 8 represent type I error rates and power for the absolute residual goodness of fit statistic, respectively, using generated data under model 3. The empirical power values were calculated by omitting the fixed effect predictors (x_2 and x_3) of the model statement and generating the data under true model 3. The results of the empirical type I error rates in Table 7 and Figure 5 are good and very close to the theoretical type I error ($\alpha = 0.05$) for number of clusters and cluster sizes of 10 or more. The empirical power results in Table 8 and Figure 6 to detect the fixed effect predictors x_1 and x_3 demonstrate the absolute residual GOF statistic has good power (0.5541 – 0.7526) for moderate sizes (10-20) of number of clusters and cluster sizes. For large cluster sizes and number of clusters, the results of power demonstrate, this statistic is powerful to detect one continuous fixed effect predictor.

Table 7

Type I Error Rate Under the True Model, $logit(p_{ij}) = x_{1ij} + x_{2ij} + 0.5\ x_{3ij} + v_i$

		Number of Clusters			
		10	**20**	**25**	**50**
Cluster sizes	**4**	0.0706	0.0479	0.0486	0.0499
	10	0. 0529	0.0502	0.0515	0.0501
	20	0. 0489	0.0520	0.0504	0.0487
	40	0.0509	0.0506	0.0496	0.0511
	80	0.0511	0.0509	0.0519	0.0505

Table 8

Power for Detecting x_1 and x_3 Out of the Model, $logit(p_{ij}) = x_{1ij} + x_{2ij} + 0.5\ x_{3ij} + v_i$

		Number of Clusters			
		10	**20**	**25**	**50**
Cluster sizes	**4**	0.1344	0.2601	0.2776	0.4290
	10	0.5541	0.7503	0.7797	0.9540
	20	0.5946	0.7526	0.8215	0.9843
	40	0.7570	0.9387	0.9769	0.9995
	80	0.9360	0.9971	1.0000	1.0000

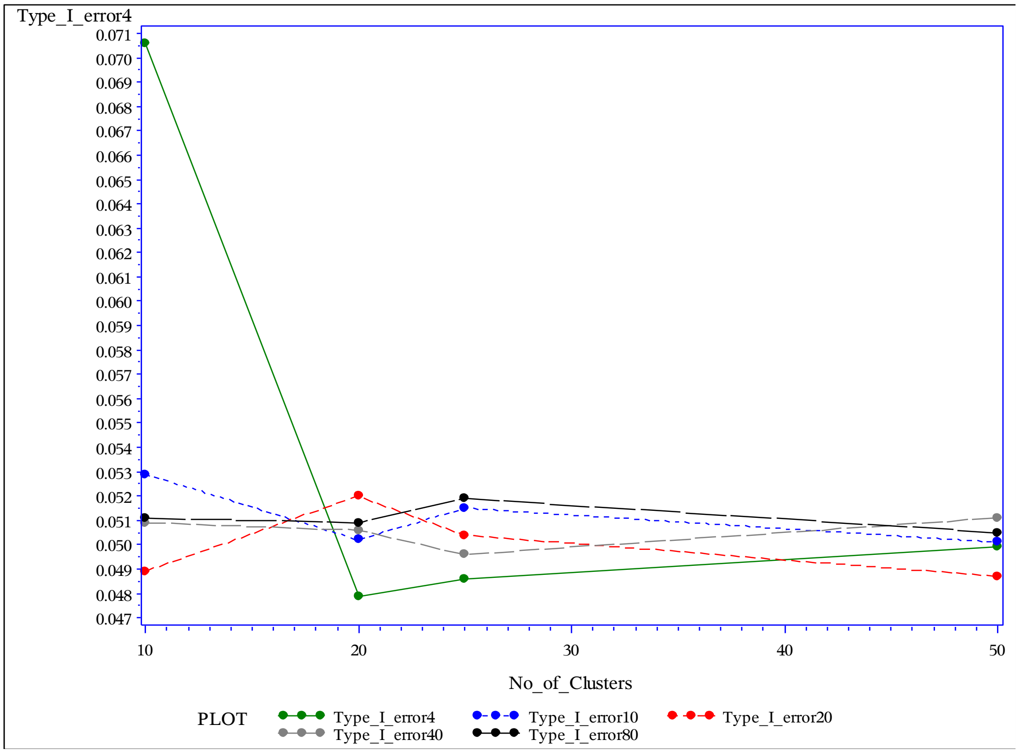

Figure 5. Plot of type I error for different cluster sizes against the number of clusters using the model equation, $logit(p_{ij}) = x_{1ij} + x_{2ij} + 0.5\ x_{3ij} + v_{i.}$.

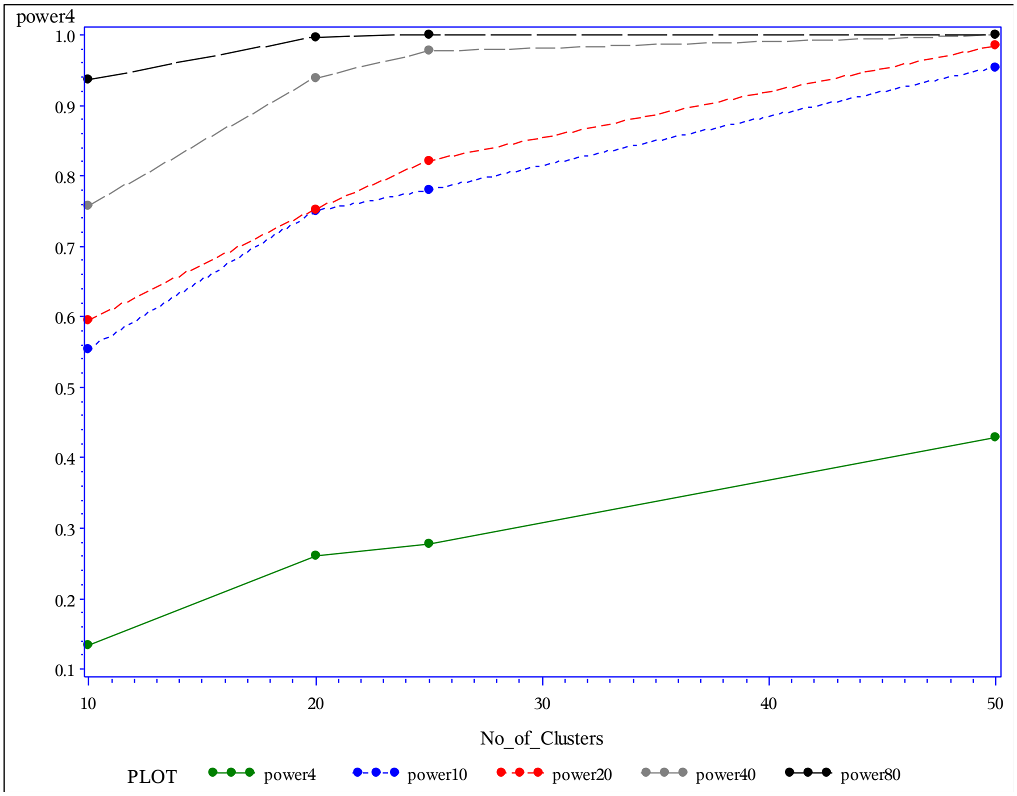

Figure 6. Plot power to detect x_1 and x_3, for different cluster sizes against the number of clusters using the model equation, $logit(p_{ij}) = x_{1ij} + x_{2ij} + 0.5\ x_{3ij} + v_{i.}$

Tables 9 and 10 represent the empirical type I error rates and power percentages for this goodness of fit statistic, respectively, using generated data under model 4. The empirical power values were calculated by omitting the fixed effect predictor (x_1) of the model statement and generating the data under true model 4. The results of type I error rates in Table 9 and Figure 7 are proper rates under the theoretical type I error ($\alpha = 0.05$) for cluster sizes of 20 observations or more over all applied number of clusters. The empirical power percentages in Table 10 and Figure 8 are very good (0.4910 – 1.000) for cluster sizes of 20 observations or more over all applied number of clusters; the power tends to be 100% for large samples.

Table 9

Type I Error Rate Under the True Model, $logit\left(p_{ij}\right) = x_{1ij} + x_{2ij} -\ x_{4ij} + v_i$

		Number of Clusters			
		10	**20**	**25**	**50**
Cluster sizes	**4**	0.0443	0.488	0.0532	0.0524
	10	0.0454	0.0467	0.0488	0.0518
	20	0.0479	0.0483	0.0481	0.0491
	40	0.0489	0.0504	0.0497	0.0512
	80	0.0507	0.0491	0.0511	0.0482

Table 10

Power for Detecting x_1 *Out of the Model,* $logit\left(p_{ij}\right) = x_{1ij} + x_{2ij} -\ x_{4ij} + v_i$

		Number of Clusters			
		10	**20**	**25**	**50**
Cluster sizes	**4**	0.1170	0.2191	0.2386	0.3875
	10	0.4469	0.7138	0.7830	0.9474
	20	0.4910	0.7200	0.7994	0.9795
	40	0.7161	0.9321	0.9838	1.0000
	80	0.9481	0.9959	1.0000	1.0000

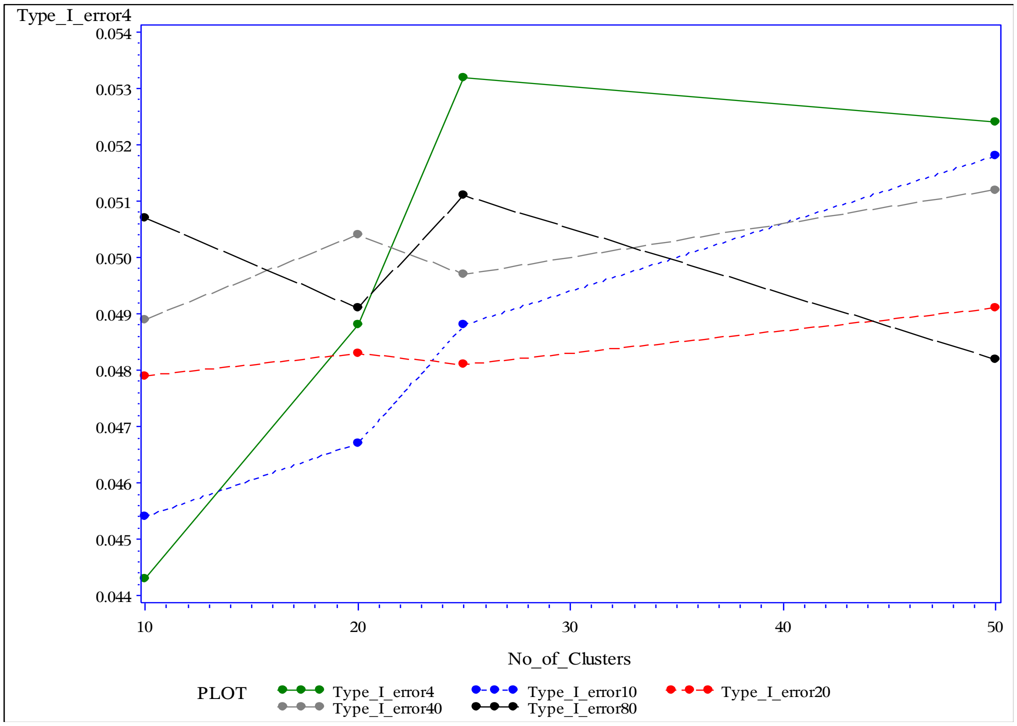

Figure 7. Plot of type I error for different cluster sizes against the number of clusters using the model equation, $logit(p_{ij}) = x_{1ij} + x_{2ij} - x_{4ij} + v_i$.

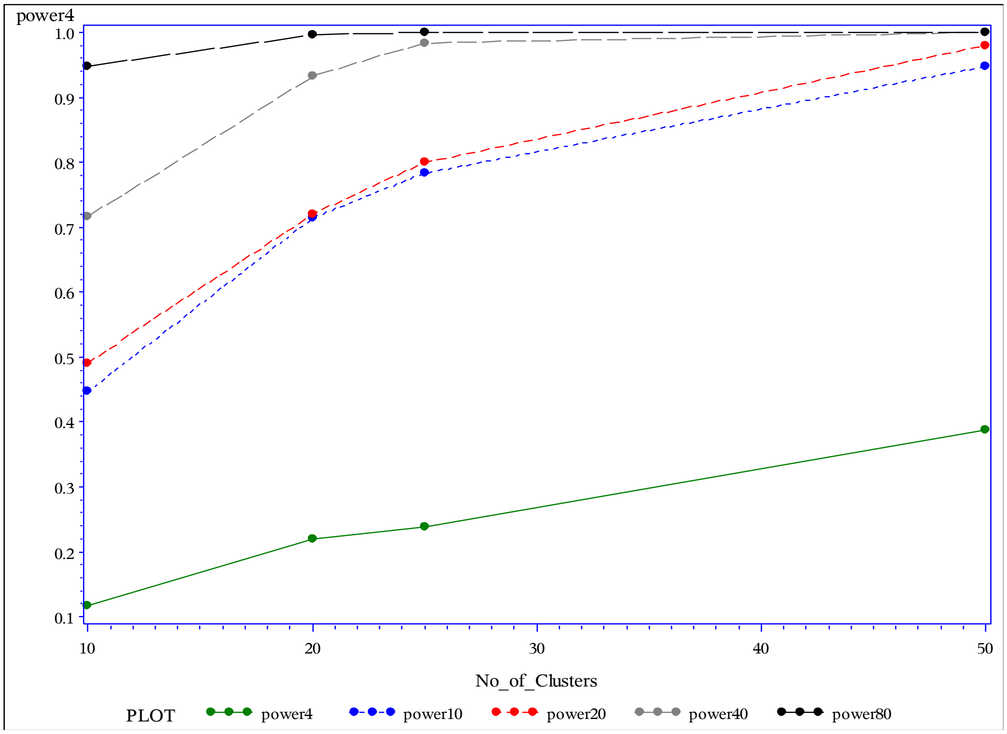

Figure 8. Plot of power to detect x_1 for different cluster sizes against the number of clusters using the model equation,$logit(p_{ij}) = x_{1ij} + x_{2ij} - x_{4ij} + v_i$.

Sturdivant and Hosmer (2007) applied the smoothed residual goodness of fit statistic, which was introduced in Chapter II, in some cases of model equations, cluster sizes, and number of clusters. Some of their results of type I error using six suggested model equations and sample sizes similar to our work are presented in Table 11. However, the model equations they used were not clearly specified in their work; some were models of one continuous predictor with random intercept and slope, three continuous and two binary predictors with random intercept, and two random slopes for one of the continuous predictors and one of the binary predictors.

Table 11

Empirical Type I Error Rates of Some Cases of Hosmer and Sturdivant's Work

Sample Sizes	Model Equations					
	1	2	3	4	5	6
20 clusters and 20 observations in each cluster	0.062	0.032	0.047	0.039	0.062	0.042
50 clusters and 4 observations in each cluster	0.04	0.055	0.049	0.048	0.051	0.04
25 clusters and 4 observations in each cluster	0.04	0.031	0.049	0.042	0.039	0.038

Table 12 shows the results of power of Sturdivant and Hosmer's (2007) work. These results are for two cases of the model equations and some cases of cluster sizes and number of clusters similar to our work. However, these results of power were to detect moderate and significant quadratic terms in the models. For example, consider the following model:

$$y_{ij}|v_i \sim Bernoulli(p_{ij})$$

$$v_i \sim N(\mu_v, \sigma_v^2)$$

$$logit(p_{ij}) = \beta_q x_{qij} + v_i,$$

where x_q is a continuous variable. Hosmer and Sturdivant use the quadratic terms $1.13x_q^2$ and $2.13x_q^2$ for the moderate and significant quadratic detecting, respectively.

Table 12

Power Results of Some Cases of Sturdivant and Hosmer's Work

Sample Sizes	Moderate Quadratic		Significant Quadratic	
	Case 1	Case 2	Case 1	Case 2
20 clusters and 20 observations in each cluster	0.156	0.238	0.830	0.722
50 clusters and 4 observations in each cluster	0.102	0.036	0.100	0.030
25 clusters and 4 observations in each cluster	0.046	0.078	0.098	0.066

The results of type I error rates in Table 11 were good for only some cases of the model equations. The results of power in Table 12 demonstrated that Sturdivant and Hosmer's (2007) test statistic did not have high power to detect a moderate quadratic term for the cases of model equations, cluster sizes, and number of cluster. To detect significant quadratic terms, the results of power were very good (0.722 – 0.83) for 20 clusters and 20 observations per cluster. However, the exact model equations Sturdivant and Hosmer used were not used in our work because they were not clearly specified in their paper.

In general, the absolute residual goodness of fit statistic gave good results in terms of type I error rates and power for all cases of cluster sizes of 10 observations or more and number of cluster of 10 or more. The only restriction on this test statistic was that it could be affected by extremely small or large estimated probabilities. This might be affected by the standardized Pearson's residuals we used. However, the estimated covariance matrix of the response depended on the estimated probabilities and could have affected the Pearson residual distributions. The optimal interval of the estimated probabilities, for which this goodness of fit statistic was valid, was not very clear. Also, our simulation studies showed that choosing this optimal interval could be affected by the number of clusters and the cluster sizes. However, depending on simulation studies, the previous results of our simulation were conducted by using the estimated probability intervals in Table 13.

Table 13

Probability Intervals for Different Number of Clusters and Cluster Sizes

Cases	Number of Clusters	Cluster Sizes	Probability Interval Limits
1	10	4	$0.0100 < \hat{p} < 0.9900$
2	20	4	$0.0125 < \hat{p} < 0.9875$
3	25	4	$0.0180 < \hat{p} < 0.9820$
4	50	4	$0.0250 < \hat{p} < 0.9750$
5	10	10	$0.0380 < \hat{p} < 0.9620$
6	20	10	$0.0495 < \hat{p} < 0.9505$
7	25	10	$0.0505 < \hat{p} < 0.9495$
8	50	10	$0.0405 < \hat{p} < 0.9595$
9	10	20	$0.0320 < \hat{p} < 0.9680$
10	20	20	$0.0330 < \hat{p} < 0.9670$
11	25	20	$0.0335 < \hat{p} < 0.9665$
12	50	20	$0.0370 < \hat{p} < 0.9630$
13	10	40	$0.0418 < \hat{p} < 0.9582$
14	20	40	$0.0318 < \hat{p} < 0.9682$
15	25	40	$0.0250 < \hat{p} < 0.9750$
16	50	40	$0.0225 < \hat{p} < 0.9775$
17	10	80	$0.0206 < \hat{p} < 0.9794$
18	20	80	$0.0200 < \hat{p} < 0.9800$
19	25	80	$0.0200 < \hat{p} < 0.9800$
20	50	80	$0.0170 < \hat{p} < 0.9830$

These probability intervals were chosen depending on some simulation studies for each case of our simulation. For example, using one of the suggested models to generate data of 20 clusters and cluster size of 20 observations and based on simulation studies, the best interval of the estimated probabilities that made the absolute residual goodness of fit statistic give proper type I error size was $(0.0330 < \hat{p} < 0.9670)$. The simulation was applied to 1000 replications of data. This interval was chosen by first generating probabilities without restriction $(0 < \hat{p} < 1)$ and then narrowing the predicted probabilities interval until the Type I error rate is within the interval (0.045 - 0.055). After choosing the predicted probabilities interval, this interval is applied to the rest of the simulation cases. That is, in practical application if a mixed effect logistic model was fitted for data of 20 clusters and cluster size of 20 observations, the absolute residual goodness of fit statistic would be appropriate to use as long as this restriction was typically satisfied. Furthermore, the structure of the model equation had a slight effect on choosing the optimal interval. That is, these intervals would be subject to adjustment or changed in future research and applications.

Further research is recommended to make adjustments or approximations on the estimated covariance matrix of the response such that the distribution of Pearson's residuals could always be approximated as a standardized normal. Thus, this goodness of fit statistic could be valid for all values of the estimated probabilities.

CHAPTER V

CONCLUSIONS AND FUTURE STUDIES

Conclusions

In this research, three goodness of fit statistics for the mixed effect logistic regression model were proposed; their performance in terms of type I error rates and power was examined via simulation studies. These simulation studies were applied to answer the following research questions and compare in general the results with Sturtevant and Hosmer's (2007) work:

Q1 What is the sampling distribution of the logit residual goodness of fit statistic?

Q2 What is the sampling distribution of the log-transformed residual goodness of fit statistic?

Q3 What is the sampling distribution of the absolute residual goodness of fit statistic?

Q4 Do these proposed goodness of fit statistics have greater power than existing goodness of fit statistics for small cluster sizes?

Q5 Do these proposed goodness of fit statistics have a proper type I error rate?

The three proposed test statistics were the logit residual, log-transformed residual, and absolute residual goodness of fit statistics, which were introduced in Chapter III. According to the simulation studies, it was found that the sampling distributions of the logit residual, log-transformed residual, and absolute residual goodness of fit statistics could be approximated as a chi square distribution, a normal distribution, and a normal distribution, respectively. Also, it was noted that these approximated distributions were sensitive to the change in the model's equation. However, the absolute residual goodness of fit statistic was more sensitive than the other test statistics for any slight change in the model's equation.

Using the derived expression to estimate the variance of the logit residual goodness of fit statistic, the estimated variance was always smaller than the empirical variance; also, using the derived expression to estimate the mean of the log-transformed residual goodness of fit statistic, the estimated mean was always smaller than the empirical mean of this statistic. Therefore, under these issues, the statistics could not give either proper type I error rates or good power (0.50 or more). However, the expressions used to estimate the moments of both goodness of fit statistics were based on first order Taylor series approximations. Further research might solve the issues of these goodness of fit statistics by using higher order Taylor series approximations. This might help increase the precision of estimating the moments of these fit statistics; specifically, a second order approximation might be investigated.

The estimated moments of the absolute residual goodness of fit statistic using the standardized Pearson residuals and a folded normal approximation were very close to the empirical moments of this fit statistic over all simulation cases. Accordingly, this fit statistic gave proper size of type I error rates and very good power (0.72 – 1.00) over most cases of our simulation. In general, the results were proper in terms of power and type I error rates for small sample size of 10 clusters and 10 observations per cluster. For sample sizes of 20 clusters and 20 observations per cluster or more, the results of type I error rates were very close to the theoretical Type I error rate (5%). The results of type I error rates were proper for all model equations and sample sizes cases except the case of 10 clusters each of four observations. Based on our simulation studies, the distribution of the absolute residual goodness of fit statistic could not be approximated as a normal distribution. Also, there were slight differences across various model equations results. Some models gave empirical type I error rates very close to 0.05 and others gave type I error rates slightly larger or smaller than the theoretical type I error of 0.05.

The power to detect the random effect was 100% for small sample sizes of 10 clusters and 10 observations per cluster or more. Also, it was very good (0.981 - 1.000) for sample sizes of 20 or more clusters each of four observations. The power to detect one or two fixed effects predictors was very good (0.4910 - 1.000) for sample sizes of 20 clusters and 10 observations per cluster or more; it tended to be 100% as the sample size got larger. Generally, unless the number of clusters was 10 and the observations per cluster were four, the power results of detecting one or more fixed effect predictor were good (0.22 - 1.00) for all other cases.

The absolute residual goodness of fit statistic was a straight-forward test; the derivative of its moments did not depend on any Taylor series approximations and any smoothing methods. The only assumption this statistic needed was the normality assumption for the model's residuals. Also, the log-transformed goodness of fit statistic needed the assumption of normality the residuals held and the logit residual goodness of fit statistic assumed that the logit of the residuals was approximately normally distributed. These two statistics had the additional restriction of a first order Taylor series approximation.

Furthermore, the simulation results of the absolute residual goodness of fit statistic were better than the results of Sturdivant and Hosmer's (2007) work. Most of the selected cases of their results of type I error in Table 11 were not consistent with the expected nominal type I error ($\alpha = 0.05$). Also, from Table 12, the power to detect a moderate quadratic term in the model was low. However, most of the results of type I error rates were close to the theoretical type I error ($\alpha = 0.05$) even if the sample sizes were small. Also, if one considered detecting the fixed effect (x_1) close to detecting the moderate or significant quadratic terms ($1.13x_1^2$ or $2.13x_1^2$) in the model equation, most of the results of power were higher than Sturdivant and Hosmer's results even if the sample sizes were small. However, according to the simulation, the absolute residual goodness of fit statistic is recommended to use in the mixed effect logistic models as long as the sample sizes are 10 clusters and 10 observations per cluster or more. Also, it is recommended for use in the ordinary logistic models with same restriction on the sample sizes.

The only restriction of the absolute residual goodness of fit statistic is that it could be affected by extremely small or large estimated probabilities of the model. However, the results were presented under specific intervals of the estimated probabilities for each case of our simulation but it seemed that this issue could be solved in future work.

Future Studies

The goodness of fit statistics for the logistic model could be affected if the estimated probabilities were extremely large or small. These probabilities would affect the estimated covariance matrix of the response. However, except for Hosmer and

Lemeshow's (1980) test statistic, all the present goodness of fit statistics had estimated moments based on the estimated covariance matrix of the response.

The absolute residual goodness of fit statistic was based on standardized Pearson's residuals. The problem with this test statistic arose from extreme values of the estimated probabilities (close to 1 or close to 0). However, the Pearson's residuals could be affected by these extreme probabilities. To reduce or avoid this problem in future research, the following techniques are recommended:

1. Using scaled Pearson residuals instead of the standardized Pearson residuals. The scaled Pearson residuals are similar to the standardized residuals in usual cases but they could control the overdispersion problem that might occur with the binary response. However, overdispersion might arise due to extremely small or large probabilities.
2. Applying some approximations on the estimated covariance matrix of the response such that the approximations have been used to estimate the confidence interval of the Binomial proportion in case of extreme probabilities. These approximations might help to adjust the standardized residuals to be always approximately normally distributed.
3. Smoothing the standardized Pearson residuals using the normal kernel smoothing function.

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: John Wiley and Sons.

Azzalini, A., Bowman, A. W., & Hardle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika*, *76*, 1-11.

Breslow, N. E., & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9-25.

Copas, J. B. (1980). Plotting *p* against *x*'. *Applied Statistics*, *32*, 25-31.

Evans, S. R. & Hosmer, D. W. (2004) Goodness of fit tests for mixed effects logistic models characterized by clustering. *Communications in Statistics: Theory and Methods, 33*(5), 1139–1155.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review, 55*, 245-260.

Henderson. C. R, Kempthorne, O., Searle, S. R., & Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, *15*, 192-218.

Hosmer, D., Hosmer, T., le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine, 16*, 965–980.

Hosmer, D. W., & Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics, A10,* 1043-1069.

Huber, P. J. (1967). *The behavior of maximum likelihood estimates under nonstandard conditions*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, CA.

le Cessie, S., & van Houwelingen, J. (1991). A goodness-of-fit test for binary regression models based on smoothing methods. *Biometrics 47,* 1267–1282.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, 58*(4). 629-678.

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

Lin, X., & Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of American Statistical Society, 91(435), 1007–1016.

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

Pan, Z, & Lin, D. Y. (2005). Goodness-of-fit methods for generalized linear mixed models. *Biometrics, 61,* 1000-1009.

Seber, G. A. F. (1977). *Linear regression analysis*. New York: Wiley.

Sturdivant, R. X., & Hosmer, D. W. (2007). A smoothed residual based goodness-of-fit statistic for logistic hierarchical regression models. *Computational Statistics & Data Analysis, 51,* 3898– 912.

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

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48,* 817–38.

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

APPENDIX

SAS PROGRAMS

The following SAS program was used to generate data for the mixed effect logistic regression models. It included two programs: one to fit the generated data using GLIMMIX Macro and other to calculate the proposed fit statistic, its moments, and *p*-value.

```
Dm 'log' clear;
Dm 'output' clear;

PROC PRINTTO  new
LOG='C:\log.txt'
print='C:\out.txt';
RUN;

%Macro GOF(rep=, btwn=, wthn=, L= );

ODS RESULTS = off;

*** Generate data for the mixed effect logistic regression models ***;

%Do i=1 %TO &rep;

data Test;

do case=1 to &btwn;

seed=5*case;
seed1=3*seed;
seed2=2*seed1;
seed3=5*seed2;

u1=ranuni(seed);              ** Generate the random component
                                (Inverse link) **;

Do withincase=1 to &wthn;

cat=int((ranuni(seed1)*3)+1); ** Generate categorical predictor from
                                 1-3 **;

x0=1;

x1= RAND ('NORMAL',0,2);       ** Generate continuous predictor **;

bin= RAND('BERNOULLI',.5);     ** Generate binary predictor **;



U=1-&L;            ** Generate probabilities related to the
                      random component on an interval (L,U) **;


if 0 < u1 <= 0.5 then p = &L + (0.5-&L)*ranuni(seed2);

else p = 0.5 + (U-0.5)*ranuni(seed3);
```

```
rand=log(u1/(1-u1));               ** Logit and the random effect
                                      variables **;

logit=log(p/(1-p));


x2=(logit-( x1 + (0.5*cat) +rand)); ** Generate other continuous
                                        variable **;

y=RAND('BERNOULLI',p);        ** Generate the binary response according
                                 to the probabilities **;

keep case withincase rand x0 x1 x2 bin cat  y p logit;

output;

end;

end;

*** Fit the generated data using GLIMMIX Macro ***;

%inc 'C:\glmm800.sas' / nosource;

            %glimmix(data=Test,

                  procopt=info cl mmeq mmeqsol absolute
                                  covtest,out=Pearson,
                  stmts=%str(

                        class case  ;

                        model  y = x1 x2  cat / solution;

                        random rand /s g gi solution sub=case;

                        ods output InvG=ginv MMEqSol=MMSol G=gdat;
                            ),
                        error=binomial, options=noprint;
                            );

*** Macro caculate fit statistic, moments and p-value **;

%inc 'C:\fit3.sas' / nosource;
%fit3

%END;

%mend;

%GOF(rep=1000, btwn=50, wthn=20, L=0.0337);
```

The following SAS program is used to calculate the logit residual GOF statistic, its moments and the p-value, using the output data set (_ds) of the GLIMMIX Macro.

```
Dm 'log' clear;
Dm 'output' clear;

%MACRO fit1;

ODS RESULTS = off;

PROC IML;

      USE _ds;
            read all var {_y} into yvec;
            read all var {_w} into wvec;
            read all var {mu} into probhat;
      CLOSE _ds;

*** Logit residuals ***;

      ehat = yvec-probhat;
      what = diag(wvec);
      winv = inv(what);
      n = nrow(probhat);
      a=log(probhat);
      b=log(1-probhat);


elogit=a-b+(invwmat*ehat);


*** Logit fit statistic and its moments **;

pearson=t(elogit)*elogit;


mean=trace(invwmat);
variance=2*trace(invwmat*invwmat);


*** P-value using chi approximation ***;

C=VARIANCE/(2*MEAN);
V=(2*(MEAN**2))/VARIANCE;


stat=pearson/c;

pvalue = 1-probchi(stat,v);


*** Save p-value, statistic and its moments in file fit1 ***;
```

```
filename fit1 'C:\fit1.txt' mod;
      file fit1;
      put @1 pvalue @25 pearson  @45 mean @60 variance;
      run;
quit ; run ;

%MEND ;
```

The following SAS program was used to calculate the log-transformed residual goodness of fit statistic, its moments, and the *p*-value using the output data set (_ds) of the GLIMMIX Macro.

```
Dm 'log' clear;
Dm 'output' clear;

%MACRO fit2;

ods results=off;


PROC IML;

      USE _ds ;

            read all var {_y} into yvec ;
            read all var {_w} into wvec ;
            read all var {mu} into probhat ;


      CLOSE _ds ;

*** Calculate absolute residuals and their moments ***;

ehat = yvec-probhat;

what = diag(wvec);

winv = inv(what);

n = nrow(probhat) ;

a=-log(1-abs(ehat));

sqrtwtmat=sqrt(what);

constvec= j(n,1,sqrt(14/22));

constvec1= j(n,1,(14/22));


diagconst1=diag(constvec1);

meanabs=sqrtwtmat*constvec;



    varabs=wtmat-(wtmat*diagconst1);


    cont1=(1/(1-meanabs));


    meane=log(cont1);
    matrix=diag(cont1);


    vare=matrix*varabs*t(matrix);

** Moments, the statistic and p-value using normal approx. ***;

means=trace(vare)+(t(meane)*meane);

variances=2*trace(vare*vare)+(4*(t(meane)*vare*meane));

stat=t(a)*a;

stat1 = (stat-means)/(variances**0.5);

pvalue = (1-probnorm(stat1));


*** Save p-value, statistic and its moments in file fit2 ***;


filename fit2 'C:\fit2.txt' mod;
      file fit2;
      put  @1 pvalue  @20 stat  @40 means @60 variances;
      run;

quit;
run;

%MEND ;
```

The following SAS program was used to calculate the absolute residual goodness of fit statistic, its moments, and the *p*-value using the output data set (_ds) of the GLIMMIX Macro.

```
Dm 'log' clear;

Dm 'output' clear;

%MACRO fit3;

ods results=off;

PROC IML;

      USE _ds ;

            read all var {_y} into yvec ;
            read all var {_w} into wvec ;
            read all var {mu} into probhat ;

      CLOSE _ds ;

*** Standardized residuals and moments of absolute std. residuals ***;

    n = nrow(probhat);

    wtmat=diag(wvec);

    invwt=inv(wtmat);

    ehat = (yvec-probhat);


    stdehat=sqrt(invwt)*ehat;


    meanabs=sqrt(14/22);

    varabs=(1-(14/22));


*** Sum of absolute residual statistic ***;

    r=abs(stdehat);

    unitvec= j(n,1,1);

    stat=t(unitvec)*r;


*** Test statistic's moments and p-value ***;

mean=n*meanabs;
```

```
var=n*varabs;


stat = (stat-mean)/(var**0.5);


pvalue =(1-probt(stat, n-1 ));


*** Put p-value, statistic and moments in file fit3 ***;


filename fit3 'C:\fit3.txt' mod;

    file fit3;

      put  @1 pvalue  @15  stat @30 mean @45 var ;

    run;

quit;
run;

%MEND ;
```