



## A Goodness of Fit Test for the Multilevel Logistic Model

A. A. P. N. M. Perera, M. R. Sooriyachchi & S. L. Wickramasuriya

To cite this article: A. A. P. N. M. Perera, M. R. Sooriyachchi & S. L. Wickramasuriya (2014): A Goodness of Fit Test for the Multilevel Logistic Model, Communications in Statistics - Simulation and Computation, DOI: [10.1080/03610918.2013.868906](https://doi.org/10.1080/03610918.2013.868906)


To link to this article: <http://dx.doi.org/10.1080/03610918.2013.868906>

 View supplementary material 

 Accepted author version posted online: 05 Aug 2014.  
Published online: 05 Aug 2014.

 Submit your article to this journal 

 Article views: 75

 View related articles 

 View Crossmark data 

# A Goodness of Fit Test for the Multilevel Logistic Model

A. A. P. N. M. PERERA, M. R. SOORIYARACHCHI,  
AND S. L. WICKRAMASURIYA

Department of Statistics, University of Colombo, Colombo, Sri Lanka

*It is crucial to test the goodness of fit of a model before it is used to make statistical inferences. However, no satisfactory goodness of fit test is available for the case of categorical multilevel data which occur when categorical data are clustered or hierarchical in nature. Hence the aim of this paper is to develop a new goodness of fit test for multilevel binary data based on Hosmer and Lemeshow and Lipsitz, et.al. In order to identify the properties of the developed test, simulation studies were carried out to assess the Type I error and the power.*

**Keywords** Binary data; Hosmer–Lemeshow test; Multilevel models; Power; Simulation; Type I error.

**Mathematics Subject Classification** 62F05

## 1. Introduction

### 1.1 Background

Multilevel data structures consist of data measured at multiple levels. For example, one may have survey data collected on individuals, indexed by  $i$ , while the individuals in turn may reside in distinct geographical units, indexed by  $j$ . The  $j$  units could be census tracts, counties, states, or countries. Data structured in this way are implicitly ‘hierarchical’ in so far as there is a clear nesting of ‘lower level’ units ( $i$ ) within ‘higher’ level units ( $j$ ). It is not always possible to deal with single level data structures and there are many instances in practice where the data are clustered or hierarchical, resulting in multilevel data structures. Usually in data modeling, stratification or clustering is ignored. If, however, stratification is indeed effective, the ignoring of stratification usually increases the resulting variance of the estimators (Parson, 1992) as the independence of the observations is violated due to the correlation within strata, so that the usual maximum likelihood method for estimating the standard error is not valid. In the literature, there are several methods to handle this within cluster correlation in multilevel data. These methods include replicated sampling techniques (Lee et al., 1989), sandwich estimation of the standard errors (Huber, 1967), generalized estimating equations (Liang and Zeger, 1986; Zorn, 2001), and multilevel modeling (Steenbergen and Jones, 2002). In this research, the multilevel modeling approach

Received May 20, 2013; Accepted November 20, 2013

Address correspondence to Prof. M.R. Sooriyarachchi, Department of Statistics, University of Colombo, Colombo, Sri Lanka. E-mail: roshini@stat.cmb.ac.lk; roshinis@hotmail.com

is considered for analyzing clustered data. Multilevel modeling has become very important for categorical responses and occurs in many areas such as Biology, Medicine, Social Science, Education, Environmental Science etc. Since multilevel modeling is a newly introduced technique, no really satisfactory measure to assess the fit of the model with discrete response variables is currently available. Thus this paper aims to propose a new technique which is more efficient and simpler than the available techniques.

## 1.2 Objectives

The primary objective of this paper is to develop a suitable goodness of fit test for the clustered binary logistic multilevel model, based on the Hosmer and Lemeshow (1980) test.

Secondary objectives of this research are to apply the developed test to real life data and to identify the properties of the goodness of fit test for varying numbers of clusters, cluster sizes, and Intra Cluster Correlation (ICC) for the binary multilevel model by using simulated data.

## 1.3 Data for the Study

In the simulation study, 1,000 datasets were generated under the four scenarios, namely, a large number of clusters with a large cluster size, a small number of clusters with a small cluster size, a large number of clusters with a small cluster size and a small number of clusters with a large cluster size. The above four conditions were varied within three standard deviation values of random effect. Therefore altogether 12,000 datasets were considered in the simulation study. Each dataset consisted of one explanatory variable and each observation consisted of a two level hierarchical structure where the first level referred to an individual observation and the second level to the cluster the observation belongs to. In addition, to illustrate the developed test for the binary logistic model, a dataset was taken from the inbuilt datasets of MLwiN software. The dataset used was a sub-sample from the 1989 Bangladesh Fertility Survey (Huq and Cleland, 1990). The response variable of interest refers to whether a woman was practicing contraception at the time of the survey or not and it is a binary response variable. The data has a two level hierarchical structure, with 2,867 women nested within 61 districts.

## 1.4 Brief Description of the Theory

In order to develop the goodness of fit test for clustered binary data, initially the Hosmer and Lemeshow test (1980) for a single level was considered. The primary focus of the research was to extend the theory behind Hosmer's and Lemeshow's method to the multilevel binary case.

In the ordinary logistic regression model, it is considered that the responses on each observation are independent of each other. However, this assumption is violated in multilevel studies because when the data are clustered, responses are correlated with each other within the cluster. When responses are correlated the ordinary logistic regression model is unsuitable as the standard errors will be biased. Therefore the ordinary logistic model should be adjusted suitably for the clustered effect. In order to adjust for the cluster effect, model-based goodness of fit testing methods described in Lipsitz et al. (1996) and Abeysekera and Sooriyarachchi (2008) are incorporated within the multilevel model.

## 1.5 Outline of the Paper

The following is an overview of the sections included in the paper. Section 1 consists of an introduction giving an insight into the research. Section 2 gives a literature review which supports in organizing the structure and domain of this research. Section 3 presents the methodology on which this research is based. Section 4 consists of a simulation study for determining the properties of the developed goodness of fit test. Section 5 illustrates the developed test by applying it to an example. The final section, section 6 consists of a general discussion and conclusion.

## 2. Literature Review

### 2.1 Hosmer and Lemeshow Goodness of Fit Test for Binary Single Level Data

In binary data analysis, likelihood ratio deviance and Pearson chi-square statistics (Agresti, 1984) can easily be used to assess the goodness of fit of the logistic regression model for binary responses. However, these two statistics are highly inflated for small sample sizes and the  $p$ -values associated with these two statistics based on the chi-square distribution are incorrect (Hosmer and Lemeshow, 2000). According to Hosmer and Lemeshow (2000), one way to avoid this problem is to collapse the columns into a fixed number of groups and then obtain the observed and expected frequencies.

Hosmer and Lemeshow (1980) and Lemeshow and Hosmer (1982) proposed grouping strategies based on the values of the estimated probabilities of the logistic regression model. Suppose that the fitted logistic model contains  $p$  independent variables denoted by  $\underline{x}' = (x_1, x_2, x_3, \dots, x_p)$  and the set of unique patterns from the  $p$  set of covariates is denoted by a  $Q$  row by  $p$  column ( $Q \times p$ ) data frame where the  $Q$  rows represent unique observations and the  $p$  columns denote covariates, then  $Q \leq n$ . Therefore, denote the number of subjects with  $\underline{x}_q' = (x_{q1}, x_{q2}, x_{q3}, \dots, x_{qp})$  by  $m_q$ ,  $q = 1, 2, \dots, Q$ . It follows that  $\sum m_q = n$ .

Let  $y_q$  denote the number of positive responses,  $y = 1$ , among the  $m_q$  subjects with  $\underline{x} = \underline{x}_q$ . It follows that,  $\sum y_q = n_1$ , the total number of subjects with  $y = 1$ . Suppose for consideration of discussion, that  $Q = n$ . Consider that the  $n$  estimated probabilities are sorted in ascending order. Hosmer and Lemeshow (1980) grouped these sorted estimated probabilities into  $G$  groups based on the following grouping strategies.

- (1) Collapse the table into  $G$  groups based on the percentiles of the estimated probabilities
- (2) Collapse the table in to  $G$  groups based on fixed values of the estimated probabilities.

Hosmer and Lemeshow (1980) test statistic,  $\hat{C}$  is given by  $\hat{C} = \sum_{g=1}^G \frac{(O_g - n'_g \bar{\pi}_g)^2}{n'_g \bar{\pi}_g (1 - \bar{\pi}_g)}$  where  $n'_g$  is the total number of subjects in the  $g^{\text{th}}$  group,  $c_g$  denotes the number of covariate patterns in the  $g^{\text{th}}$  group where  $O_g = \sum_{j=1}^{c_g} y_j$  is the number of responses among the  $c_g$  covariate patterns, and  $\bar{\pi}_g = \sum_{j=1}^{c_g} \frac{m_j \hat{\pi}_j}{n'_g}$  is the average estimated probability, where  $\hat{\pi}_j = \frac{\exp[\widehat{h}(x)]}{1 + \exp[\widehat{h}(x)]}$  where  $\widehat{h}$  is the estimated logit.

Though  $G = 10$  groups is the most popular, there is a range of values that can be used to define  $G$ . As discussed by Hosmer and Lemeshow (1989), as a general rule,  $G$  should be chosen such that  $6 \leq G < n/5r$  where  $r$  is the number of response levels (in this case  $r = 2$  as the response is binary). Using simulations, Hosmer and Lemeshow (1980) demonstrated that, when  $Q \cong n$  and the fitted logistic regression model is the correct

model, the distribution of the statistic  $\hat{C}$  by Hosmer et al. (1988) have shown that the grouping method based on percentiles of the estimated probabilities is preferable to the one based on fixed cut points in the sense of better adherence is well approximated by the chi-square distribution with  $G-2$  degrees of freedom,  $\chi^2_{(G-2)}$ . Additional research to the  $\chi^2_{(G-2)}$  distribution, especially when many of the estimated probabilities are small (i.e. less than 0.2). Lipsitz et al. (1996) extended the method of Hosmer and Lemeshow (1980) to a model based approach.

## 2.2. Current Goodness of Fit Tests for Multilevel Studies with Binary Responses

Hosmer and Lemeshow (2000) suggested goodness of fit tests based on the design of the study. Archer and Lemeshow (2006) studied a method for assessing goodness of fit for clustered binary data and Archer et al. (2006) additionally proposed alternative design-based goodness of fit tests for logistic regression models. Graubard et al. (1997) proposed an alternative grouping method for establishing deciles of risk for the Hosmer and Lemeshow goodness of fit test. Sturdivant and Hosmer (2007) extended the goodness of fit measures used in the standard logistic setting to the hierarchical case. Sturdivant and Hosmer (2007) also developed Kernel smoothed versions of the statistics and applied a bias correction method to the uncorrected sums of squares (USS) and Pearson statistics. Pardoe (2004) extended the Bayes Marginal Model Plot (BMMP) assessment technique from a traditional logistic regression setting to a multilevel application in the area of criminal justice.

## 3. Methodology

### 3.1 Novel Goodness of Fit Test

According to the literature, there is no satisfactory goodness of fit test to check the fit of multilevel binary logistic models. Therefore the main concern of this research is to develop a goodness of fit test for multilevel binary data based on the single level Hosmer and Lemeshow (1980) goodness of fit test (Here the approach is to collapse the table according to the percentiles of the estimated probabilities based on the multilevel logistic regression model) and on the Lipsitz et al. (1996) method to include indicator variables to the multilevel binary logistic regression model. The test that will be developed has some features similar to the  $F$ -adjusted Wald test developed by Archer et al. (2006). However, the  $F$ -adjusted Wald test has resulted in an inflated Type I error rate, while the power of this particular test is small.

**3.1.1 Development of the New Goodness of Fit Test.** Multilevel data are implicitly 'hierarchical' in so far as there is a clear nesting of 'lower level' units ( $i$ ) within 'higher' level units ( $j$ ). Let  $y_{ij}$  denote the binary response of the  $i^{\text{th}}$  lower level individual unit within the  $j^{\text{th}}$  higher level (second level) to which that unit belongs and  $\pi_{ij}$  denotes the probability corresponding to  $y_{ij}$  where  $\pi_{ij} = \Pr(y_{ij} = 1)$ .

The multilevel binary logistic model has the following form when there is a single explanatory variable,  $x_{ij}$ , measured at the lower level. In this case, the single-level model can be extended to a two-level random intercept model (Goldstein, 2003)

$$\text{logit}(\pi_{ij}) = \beta_{0j} + \beta_1 x_{ij} \text{ where } \beta_{0j} = \beta_0 + u_{0j} \text{ and } u_{0j} \sim N(0, \sigma_{u_0}^2) \quad (3.1)$$

$i = 1, 2, \dots, n_j$  and  $j = 1, 2, \dots, k$  where  $k$  is the number of clusters.

The novel goodness of fit test is developed using the following steps.

**Step 1:** The multilevel logistic regression model for binary response data as in Eq. (3.1) is initially fitted and the model parameters are estimated by using the second order Penalized Quasi Likelihood (PQL) method (if there exists a convergence problem an intermediate choice of estimation procedure can be applied) (Browne, 2004).

**Step 2:** The  $\pi_{ij}$  for the  $i^{\text{th}}$  observation in the  $j^{\text{th}}$  cluster is estimated from the fitted model.

**Step 3:** In the Hosmer and Lemeshow test, the **are** assumed to be independent and not correlated as in clustered data. Thus in this test, the estimated and sorted probabilities of the entire dataset are collapsed into 'G' groups. The multilevel data structures consist of data measured at multiple levels. Data structured in this way are implicitly 'clustered' or 'hierarchical' in so far as there is a clear nesting of 'lower level' units ( $i$ ) within 'higher' level units ( $j$ ). Such data are correlated within clusters. Thus it is not possible to directly rank the estimated probabilities within each cluster. In order to overcome this problem the method of Rosner, Glynn, and Tinglee (2003) which is an asymptotic approach of ranking clustered data is used. Here the ranking of the estimated probabilities is done among all units over all clusters. The estimated probabilities are sorted and ranked in ascending order. It is important to note that the overall ranking system will be preserved within each cluster too. As the ranking system is thus preserved within each cluster, and the observations in different clusters are independent of each other (no between cluster correlation), the Hosmer and Lemeshow test (1980) approach can now be applied within cluster. Accordingly, using the Hosmer and Lemeshow (1980) method, the estimated and sorted probabilities are collapsed in to 'G' groups within each cluster, where G is a positive integer. Generally, probabilities are partitioned into 10 groups. Within a cluster the estimated probabilities are allocated into regions (groups) such that the first group contains observations with the smallest predicted probabilities and the last group contains observations with the largest probabilities. According to the partition of the data, the goodness of fit test is formulated by defining (G-1) group indicator variables for each cluster (where G is the number of groups into which each cluster is partitioned).

Then the indicator variable,

$$I_{g_{ij}} = 1; \text{ if } \pi_{ij} \text{ is in region } g \\ = 0; \text{ otherwise}$$

where  $g = 2, 3, \dots, G$ .

**Step 4:** The probabilities in each cluster ( $\pi_{ij}$ ) are sorted with respect to the observation index (the dataset arranged in the same order before the data **are** sorted).

**Step 5:** Then to assess the goodness of fit of model (3.1) it is compared with the alternative model (3.2) that has to be constructed by using indicator variables. As observations between clusters are independent, all indicator variable values are pooled to form a single indicator variable for all the clusters. Therefore the newly developed method can also be used in the same way as the Hosmer and Lemeshow test (1980).

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + \sum_{g=2}^n \gamma_g I_{g_{ij}} \quad (3.2)$$

where  $\beta_{0j} = \beta_0 + u_{0j}$  and  $u_{0j} \sim N(0, \sigma_{u_0}^2)$

$$\sum_{g=2}^{10} \gamma_g I_{gij} = \gamma_2 I_{2ij} + \gamma_3 I_{3ij} \dots \dots \dots \gamma_{10} I_{10ij}$$

$i = 1, 2, \dots, n_j$  and  $j = 1, 2, \dots, k$  where  $k$  is the number of clusters.

**Step 6:** The model (3.2) is fitted as mentioned in step 1.

**Step 7:** The joint Wald statistic (Liao, 2004) is calculated by using MLwiN software for model (3.2) to check the hypothesis :

**H<sub>0</sub> :**  $\gamma_2 = \gamma_3 = \dots = \gamma_n = 0$ ; that is, all the coefficients of the indicator variables are equal to zero, and **H<sub>1</sub> :** Not all the coefficients of indicator variables are equal zero.

**If all the indicator variables are simultaneously not significantly different from zero, it indicates that there is no evidence for lack of fit in the model (3.1)** and if not, it implies that the model under consideration (3.1) has a questionable fit. If **H<sub>0</sub>** is not rejected, it implies that there is no evidence for lack of fit of the model (3.1), and if **H<sub>0</sub>** is rejected it implies that model (3.1) does not fit the data adequately. This is the basic concept behind this novel goodness of fit test.

**Step 8:** The test is performed at  $\alpha\%$  significance level. If the calculated joint Wald statistic value is greater than the  $\chi^2_{(n-1)\alpha\%}$  value, then the null hypothesis is rejected at  $\alpha\%$  significance level and therefore it can be concluded that model (3.1) does not fit the data well. If the calculated joint Wald statistic value is less than the  $\chi^2_{(n-1)\alpha\%}$  value, then the null hypothesis is not rejected at  $\alpha\%$  significance level and it can be concluded that **there is no evidence for the lack of fit of the model (3.1)**.

The eight steps given above describe the goodness of fit testing procedure for assessing model adequacy for multilevel binary data.

#### 4. Simulation Study to Determine the Properties of the Novel Goodness of Fit Test for Clustered Data

In this section a binary response variable  $y_{ij}$  which represents the  $i$ th individual in the  $j$ th cluster and a continuous explanatory variable  $x_{ij}$  which represents the value of the explanatory variable for the  $i$ th observation in the  $j$ th cluster are considered in order to handle the multilevel nature in the simulated data.

##### 4.1 Introduction to the Simulation Study

There are no restrictions on the selection of a distribution to simulate the explanatory variable from for performing the novel goodness of fit test. Archer et al. (2007) suggest the Bernoulli distribution, normal distribution, and uniform distribution. Hdet (1999) suggests the normal and uniform distributions. In our dataset, the explanatory variable is simulated from a normal distribution.

After simulating the data, this study uses MLwiN v2.19 to fit the multilevel models. Once the type of model has been specified, it is necessary to determine the estimation procedure and the linearization. The PQL method with second order terms of the Taylor

series expansion will yield more precise estimates than the other available quasi-likelihood estimation methods in MLwiN (Browne, 2004). Therefore the PQL procedure is used as the estimation procedure in the simulations. According to Browne (2004), the second order PQL method is less stable and convergence problems may be encountered. Therefore the intermediate choice, the first-order PQL was used whenever such problems arose.

4.1.1 *Parameters Used in the Simulation Study.* The simulation procedure was carried out by varying three main conditions in order to generate datasets with various properties. These main conditions are,

**Condition 1** : Number of clusters (15 and 61)

**Condition 2** : Observations per cluster (20 and 50)

**Condition 3** : Intra cluster correlation or standard deviation (1, 1.5, and 2)

The number of clusters and the cluster size were selected on the basis of the guidelines set out by Maas et al. (2005) , Van et al. (1997) and Kreft et al. (1998) . The number of clusters was chosen so that, the smaller cluster size is 15 and the larger cluster size is 61. The choice of this larger cluster size is dependent on the inbuilt dataset in MLwiN software. As the simulation study initially used parameters from this inbuilt dataset which was a sub-sample from the 1989 Bangladesh Fertility Survey (Huq and Cleland, 1990), the same cluster size was used in our study. The observations per cluster were selected such that the number of observations per cluster in the smaller cluster is 20 and the number of observations per cluster in the larger cluster is 50. Each of these four combinations was simulated under three different ICC values by considering three different standard deviations. The combinations were named as follows.

**Dataset A:** 61 clusters with 50 observations in each cluster (Altogether 3050 observations).

**Dataset B:** 15 clusters with 20 observations in each cluster (Altogether 300 observations).

**Dataset C:** 61 clusters with 20 observations in each cluster (Altogether 1220 observations).

**Dataset D:** 15 clusters with 50 observations in each cluster (Altogether 750 observations).

For each of the 12 ( $2 \times 2 \times 3$ ) conditions, a thousand datasets were generated. An MLwiN macro was used in order to generate the datasets and perform the developed goodness of fit test. This is included in the supplementary material.

4.1.2 *Data Generation Procedure.* As discussed in section 4.1.1, by using 12 conditions, 12,000 datasets were generated to determine whether the type I error holds for the developed test. Another 12,000 datasets were generated to determine the power of the developed test. The null and alternative hypotheses are,

$H_0$  : The multilevel binary logistic model fits the data well;

$H_1$  : The multilevel binary logistic model does not fit the data well

In order to generate the datasets, MLwiN macros were used with selected parameter estimates.

Before directly selecting parameters for the study, a small trial and error analysis was conducted and then the following parameters were selected.

$\beta_0 = -0.686$  and  $\beta_1 = 0.707$

4.1.2.1 *Generation of Datasets Under the Null Hypothesis.* In the data generation procedure, first the explanatory variable was generated from the normal distribution with selected parameters. Here the normal distribution with mean 2 and variance 1 ( $x_{ij} \sim N(2, 1)$ ) was used. Then according to the value of standard deviation, the random effect  $u_{0j}$  was generated from the normal distribution because, according to the theory of multilevel



**Table 1**  
Observed Type I error rates for the simulation study

|                                  |                                    | Number of<br>significant<br>datasets | Rejection<br>proportion | Result             |
|----------------------------------|------------------------------------|--------------------------------------|-------------------------|--------------------|
| <i>standard deviation is 1.0</i> | <b><math>k = 61, n = 50</math></b> | 58                                   | 0.058                   | Within the limits  |
|                                  | <b><math>k = 15, n = 20</math></b> | 37                                   | 0.037                   | Just within limits |
|                                  | <b><math>k = 61, n = 20</math></b> | 51                                   | 0.051                   | Within the limits  |
|                                  | <b><math>k = 15, n = 50</math></b> | 47                                   | 0.047                   | Within the limits  |
| <i>standard deviation is 1.5</i> | <b><math>k = 61, n = 50</math></b> | 53                                   | 0.053                   | Within the limits  |
|                                  | <b><math>k = 15, n = 20</math></b> | 38                                   | 0.038                   | Just within limits |
|                                  | <b><math>k = 61, n = 20</math></b> | 44                                   | 0.044                   | Within the limits  |
|                                  | <b><math>k = 15, n = 50</math></b> | 51                                   | 0.051                   | Within the limits  |
| <i>standard deviation is 2.0</i> | <b><math>k = 61, n = 50</math></b> | 46                                   | 0.046                   | Within the limits  |
|                                  | <b><math>k = 15, n = 20</math></b> | 13                                   | 0.013                   | Outside the limits |
|                                  | <b><math>k = 61, n = 20</math></b> | 49                                   | 0.049                   | Within the limits  |
|                                  | <b><math>k = 15, n = 50</math></b> | 48                                   | 0.048                   | Within the limits  |

Note: 5% significance level was considered,  $k$  represents the number of clusters, and  $n$  represents the number of observations per cluster.

data,  $u_{0j} \sim N(0, \sigma_{u0}^2)$ . After  $x_{ij}$  and  $u_{0j}$  were generated from the normal distribution, the respective probabilities of the fitted logistic models were estimated for the selected  $\beta_0$  and  $\beta_1$  parameter values.

The fitted model under the null hypothesis can be represented by

$$\text{logit}(\pi_{ij}) = \beta_{0j} + 0.707x_{ij} \text{ where } \beta_{0j} = -0.686 + u_{0j} \text{ and } u_{0j} \sim N(0, \sigma_{u0}^2)$$

$i = 1, 2, \dots, n_j$  and  $j = 1, 2, \dots, k$  where  $k$  is the number of clusters.

Table 1 gives the results of the simulation study for the Type I error for each simulation condition.

For the developed goodness of fit test, the Type I error rate clearly holds under conditions A, C, and D, irrespective of the standard deviation values.

Under condition B, the Type I error rate is marginal, being just on the lower border of the 95% probability interval when the standard deviation is 1.0 and 1.5. The reason for this can be explained by the fact that small number of clusters and small cluster sizes result in poor estimation of the fixed and random coefficients leading to bias in the joint Wald statistics and hence marginal convergence probabilities (Maas et al., 2005). When the standard deviation is 2.0 under condition B, the developed goodness of fit has a conservative Type I error rate.

Also the marginal Type I error rate for small sample size arising from a small number of clusters of small size (condition B), can be attributed to the fact that in the developed goodness of fit test, the method of indicator variable allocation works only for at least moderately large samples as explained by Rosner et al. (2003).

**Table 2**  
Observed results for the simulation study under power analysis I

|                                  |                       | Rejection Proportion                    |   |   |   |
|----------------------------------|-----------------------|---|---|---|---|
|                                  |                       | $\mu = 2,$<br>$\sigma = 1,$<br>CV = 0.5 | $\mu = 2,$<br>$\sigma = 2,$<br>CV = 1.0 | $\mu = 2,$<br>$\sigma = 3,$<br>CV = 1.5 | $\mu = 3,$<br>$\sigma = 3,$<br>CV = 1.0 |
| <i>standard deviation is 1.0</i> | <b>K = 61, n = 50</b> | 0.622                                   | 1.000                                   | 1.000                                   | 1.000                                   |
|                                  | <b>K = 15, n = 20</b> | 0.06                                    | 0.234                                   | 0.493                                   | 0.232                                   |
|                                  | <b>K = 61, n = 20</b> | 0.210                                   | 0.857                                   | 1.000                                   | 0.881                                   |
|                                  | <b>K = 15, n = 50</b> | 0.157                                   | 0.811                                   | 0.991                                   | 0.848                                   |
| <i>standard deviation is 1.5</i> | <b>K = 61, n = 50</b> | 0.571                                   | 1.000                                   | 1.000                                   | 1.000                                   |
|                                  | <b>K = 15, n = 20</b> | 0.067                                   | 0.188                                   | 0.435                                   | 0.202                                   |
|                                  | <b>K = 61, n = 20</b> | 0.194                                   | 0.845                                   | 0.993                                   | 0.832                                   |
|                                  | <b>K = 15, n = 50</b> | 0.162                                   | 0.766                                   | 0.982                                   | 0.800                                   |
| <i>standard deviation is 2.0</i> | <b>K = 61, n = 50</b> | 0.522                                   | 1.000                                   | 1.000                                   | 1.000                                   |
|                                  | <b>K = 15, n = 20</b> | 0.041                                   | 0.169                                   | 0.377                                   | 0.176                                   |
|                                  | <b>K = 61, n = 20</b> | 0.163                                   | 0.769                                   | 0.991                                   | 0.772                                   |
|                                  | <b>K = 15, n = 50</b> | 0.153                                   | 0.712                                   | 0.958                                   | 0.744                                   |

Note: 5% significance level was considered,  $k$  represents the number of clusters, and  $n$  represents the number of observations per cluster.

#### 4.2 Study of Power

It is important to discuss the power of the developed goodness of test by using a simulation study. The power of the test is associated with Type II error,  $\beta$ , where,

$$\text{Power} = 1 - \beta = \Pr(\text{reject } H_0 | H_1 \text{ is true})$$

The data were generated from the model under the alternative hypothesis. Here, in the power analysis, the model fitted to the data is mis-specified in the sense that the model fitted uses an incorrect form of the explanatory variable to the model generating the data. This mis-specification took the following form. The simulated data used a transformation of  $x$ , namely,  $\log X^2$ , as the explanatory variable whereas the fitted model used the raw form of  $X$ . The  $\beta_0$  and  $\beta_1$  of this model was taken to be  $-0.686$  and  $0.3535$ . For each combination of number of clusters, cluster size and variance of random effect term used previously the rejection proportion of the null hypothesis out of 1,000 was obtained. Initially, the variance of the predictor variable ( $X$ ) was taken to be 1 as  $X$  was simulated from a  $N(2,1)$ . Therefore the random effect was larger than the covariate effect. This attributed to poor power. Thus more simulations where the covariate ( $X$ ) was taken to have more explanatory power were done and all this is given in Table 2.

As the standard deviation of the random effect increases, the power of the test decreases. For a given standard deviation the power increases with increasing number of first level

**Table 3**  
Observed results for the simulation study under power analysis II

| Standard Deviation of Random Effect | Combination                             | No. of significant datasets | rejection proportion |
|-------------------------------------|---|-----------------------------|----------------------|
| <i>standard deviation is 1.0</i>    | <b>15 clusters with 20 observations</b> | 413                         | 0.413                |
|                                     | <b>61 clusters with 20 observations</b> | 716                         | 0.716                |
|                                     | <b>15 clusters with 50 observations</b> | 559                         | 0.559                |
|                                     | <b>61 clusters with 50 observations</b> | 998                         | 0.998                |
| <i>standard deviation is 1.5</i>    | <b>15 clusters with 20 observations</b> | 114                         | 0.114                |
|                                     | <b>61 clusters with 20 observations</b> | 553                         | 0.553                |
|                                     | <b>15 clusters with 50 observations</b> | 401                         | 0.401                |
|                                     | <b>61 clusters with 50 observations</b> | 966                         | 0.966                |
| <i>standard deviation is 2.0</i>    | <b>15 clusters with 20 observations</b> | 110                         | 0.110                |
|                                     | <b>61 clusters with 20 observations</b> | 286                         | 0.286                |
|                                     | <b>15 clusters with 50 observations</b> | 259                         | 0.259                |
|                                     | <b>61 clusters with 50 observations</b> | 804                         | 0.804                |

units. It is also clearly seen that the power increases dramatically with increasing coefficient of variation of the covariate where the covariate has more explanatory power.

Another scenario was examined for power where the data were generated from the model under the alternative hypothesis which used two covariates, the first from a uniform[-3, 3] and the second from a Bernoulli(0.5). The  $\beta_0$ ,  $\beta_1$  and  $\beta_2$  of this model were taken to be -0.3, 1.8, and 10, respectively. The parameters and distributions were selected using both Liu(2007) and trial and error methods. The model fitted to the data was taken to have a misspecified linear predictor in the sense that the model fitted uses only the first explanatory variable. The results of this study are given in Table 3.

The results of Table 3 shows the same pattern of power as Table 2.

## 5. Application to an Example

### 5.1 Description of the Example

The example dataset was taken from the inbuilt datasets of MLwiN software. This dataset was a sub-sample from the 1989 Bangladesh Fertility Survey (Huq and Cleland, 1990). The

**Table 4**  
Description of potential analytical factors

| Variable | Variable description   | Base category |
|----------|--|---------------|
| Lc       | Number of living children at time of survey.                           | None          |
| Age      | Age of woman at time of survey centered on the sample mean of 30 years | —             |
| Urban    | Type of region of residence  | Rural         |
| Educ     | Woman's level of education   | None          |
| Hindu    | Woman's religion   | Muslim        |
| d_lit    | Proportion of women in the district who are literate                   | —             |
| d_pray   | Proportion of Muslim women in the district who pray every day          | —             |

response of interest is binary and refers to whether a woman was practicing contraception at the time of the survey or not. The dataset consisted of 7 explanatory variables spread across two main levels. The second level unit of the dataset can be identified as the district to which each woman belongs, while the first level unit comprises of an individual woman. Table 4 shows the potential analytical factors/covariates and their base categories.

5.1.1 *Data Preparation.* The dataset consisted of 61 clusters. However, for our example, only 49 clusters were considered, omitting very sparse clusters with less than 20 observations. In the original example, 12 clusters had observations numbering less than 20. In order to get proper parameter estimates, these 12 clusters were ignored and the remaining clusters were renamed. Thus our example consists of 49 clusters having a total of 2711 observations. The data are included with the supplementary material.

## 5.2 Model Fitting

Initially it is necessary to identify the variables that have a significant impact on the response variable. The forward selection method with significance level of 5% (0.05) was used to select a suitable model. (Blanchet et al., 2008).

When selecting important variables to the model, the Wald statistic associated with the variable was used instead of the likelihood ratio (or deviance) test statistic, which is the standard statistic used in the basic logistic regression model. The reason for this is that for discrete response multilevel models the likelihood test is not available in MLwiN. Parameter estimates are obtained from the quasi-likelihood method PQL2.

5.2.1 *Final Main Effects Model.* The final main effects model consists of six factors/covariates. The logistic regression model selected by the forward selection procedure is

$$\begin{aligned}
 \text{logit}(p_{ij}) = & \beta_{0j} + 1.197lc\_1ij + 1.504lc\_2ij + 1.546lc\_3ij + 0.239educ\_2ij \\
 & + 0.697educ\_3ij + 1.99educ\_4ij + 0.539urbanij + 0.423hindu_{ij} \\
 & - 0.019age_{ij} - 1.039d\_pray_{ij} \\
 \beta_{0j} = & -1.626 + u_{0j}
 \end{aligned} \tag{5.1}$$

**Table 5**Parameter estimates, standard errors, Wald statistics and  $p$ -values of the main effects model

| Factor/Covariate | Category           | $\hat{\beta}$ (SE ( $\hat{\beta}$ )) | Wald statistic | $p$ -value |
|------------------|--------------------|--------------------------------------|----------------|------------|
| Lc               | Child              | 1.197(0.138)                         | 75.237         | < 0.0001   |
|                  | 2 children         | 1.504(0.151)                         | 99.207         | < 0.0001   |
|                  | 3 or more children | 1.546(0.157)                         | 96.966         | < 0.0001   |
| Age              | —                  | -0.019(0.007)                        | 7.367          | 0.0033     |
| urban            | Urban              | 0.539(0.107)                         | 25.375         | < 0.0001   |
| Educ             | Lower primary      | 0.239(0.132)                         | 3.278          | 0.0702     |
|                  | Upper primary      | 0.697(0.147)                         | 22.482         | < 0.0001   |
|                  | Secondary+         | 1.199(0.131)                         | 83.771         | < 0.0001   |
| hindu            | Hindu              | 0.423(0.131)                         | 10.426         | 0.0012     |
| d_pray           | Pray               | -1.039(0.519)                        | 4.008          | 0.0453     |

where  $p_{ij}$  is the probability of practicing contraception at the time of the survey for the  $i^{\text{th}}$  observation in the  $j^{\text{th}}$  cluster. The terms within parentheses are the standard errors of the estimated parameters. Table 5 gives the parameter estimates, standard errors, Wald statistics, and fitted values of the main effects model.

Results of Table 5 show that with increasing number of children, the odds of Bangladeshi women using contraceptives increases. With increasing age, the odds of Bangladeshi women using contraceptives decreases. The odds of Bangladeshi women using contraceptives are higher in urban areas compared to rural areas. As the education level increases, the odds of Bangladeshi women using contraceptives increases. Hindus have higher odds of contraceptive use than Muslims. Women who pray have lower odds of contraceptive use than those who do not pray.

Adhering to the principal of parsimony, the main effects model is preferred over the interaction model, because the main effects model is less complex and more comprehensible than the interaction model.

### 5.3 Goodness of the Fitted Model

In this section, the developed goodness of fit test for the multilevel clustered binary logistic model is applied to the final model for the example dataset as explained in section 3.

In the example dataset, most of the clusters consisted of observations that were not divisible by 10. Therefore according to Abeysekara and Sooriyarachchi (2009), indicator variables were defined for this multilevel data as follows.

$$g_i = \frac{\text{Number of observations in the } i\text{th cluster}}{10}$$

If  $i \leq a^*g_i$  then  $I = a$  for  $a = 1, 2, 3, \dots, 10$ ; where  $i = 1, 2, \dots, 49$  and  $I$  represents the indicator variable.

Therefore each cluster had a different count of indicator variables. In this goodness of fit test the null hypothesis is;

$H_0$ : The model fits the data well.

versus the alternative hypothesis;

$H_1$ : The model does not fit the data well.

**Table 6**

Parameter estimates, standard errors, Wald statistics and *p*-values of the Goodness of fit model

| Factor/Covariate | Category           | $\hat{\beta}$ (SE ( $\hat{\beta}$ )) | Wald statistic | <i>p</i> -Value |
|------------------|--------------------|--------------------------------------|----------------|-----------------|
| Lc               | Child              | 1.129(0.255)                         | 19.602         | 0.000010        |
|                  | 2 children         | 1.424(0.321)                         | 19.679         | 0.000009        |
|                  | 3 or more children | 1.453(0.337)                         | 17.830         | 0.000024        |
| Age              | —                  | -0.018(0.009)                        | 4.000          | 0.045500        |
| Urban            | Urban              | 0.562(0.140)                         | 16.114         | 0.000060        |
| Educ             | Lower primary      | 0.262(0.145)                         | 3.265          | 0.070773        |
|                  | Upper primary      | 0.774(0.199)                         | 15.127         | 0.000101        |
|                  | Secondary+         | 1.220(0.241)                         | 25.626         | < 0.000001      |
| Hindu            | Hindu              | 0.459(0.149)                         | 9.490          | 0.002066        |
| d_pray           | Pray               | -1.022(0.518)                        | 3.893          | 0.048488        |
| I_g              | I.2                | -0.333(0.258)                        | 1.665          | 0.196930        |
|                  | I.3                | -0.160(0.283)                        | 0.320          | 0.571608        |
|                  | I.4                | 0.071(0.310)                         | 0.052          | 0.819619        |
|                  | I.5                | 0.114(0.327)                         | 0.122          | 0.726875        |
|                  | I.6                | 0.091(0.345)                         | 0.070          | 0.791337        |
|                  | I.7                | 0.033(0.365)                         | 0.008          | 0.928730        |
|                  | I.8                | -0.056(0.393)                        | 0.020          | 0.887537        |
|                  | I.9                | -0.065(0.439)                        | 0.022          | 0.882087        |
|                  | I.10               | -0.137(0.523)                        | 0.069          | 0.792798        |

Then to assess the goodness of fit of the final main effects model, i.e., model (5.1), the alternative model (5.2) is constructed.

$$\begin{aligned}
 \text{logit}(p_{ij}) &= \beta_{0j} + 1.129lc_{.1ij} + 1.424lc_{.2ij} + 1.453lc_{.3ij} + 0.262educ_{.2ij} \\
 &\quad + 0.774educ_{.3ij} + 1.220educ_{.4ij} + 0.562urban_{ij} + 0.459hindu_{ij} \\
 &\quad - 0.018age_{ij} - 1.022d\_pray_{ij} + \sum_{g=2}^{10} \gamma_g I_{-gij} \\
 \beta_{0j} &= -1.557 + u_{0j}
 \end{aligned}
 \tag{5.2}$$

where  $I_{-gij}$  is the indicator variable of the  $g^{\text{th}}$  group for the  $i^{\text{th}}$  observation in the  $j^{\text{th}}$  cluster.

$$\begin{aligned}
 \sum_{g=2}^{10} \gamma_g I_{-gij} &= -0.333I_{.2ij} - 0.160I_{.3ij} + 0.071I_{.4ij} + 0.114I_{.5ij} + 0.091I_{.6ij} \\
 &\quad + 0.033I_{.7ij} - 0.056I_{.8ij} - 0.065I_{.9ij} - 0.137I_{.10ij}
 \end{aligned}$$

If model (5.1) is correctly specified, then the  $H_0$ : The model (5.1) fits correctly to the data is not rejected and it indicates that,  $\gamma_2 = \gamma_3 = \dots = \gamma_{10} = 0$ .

Table 6 gives the parameter estimates, standard errors, Wald statistics and  $p$ -values of the goodness of fit model (model 5.2)

By using the MLwiN software, the joint Wald statistic was calculated for model (5.2) in order to check the hypothesis,

$H_0 : \gamma_2 = \gamma_3 = \dots = \gamma_{10} = 0$ , that all the coefficients of the indicator variables are equal to zero.

$H_1$  : Not all the coefficients of the indicator variables are equal to zero.

The joint Wald statistic obtained was 7.308 on 9 degrees of freedom giving a  $p$ -value of 0.6051. As the  $p$ -value is much larger than 0.05, it can be concluded that  $\gamma_2 = \gamma_3 = \dots = \gamma_{10} = 0$  is not rejected at 5% significant level and hence, it can be concluded that model (5.1) that is the model without indicator variables fits the data well. One might argue based on the results of Table 2 that if the coefficient of variance in the covariates are small for this example (e.g.: around 0.5) this corresponds to a maximum power in Table 2 for 61 clusters of 0.622 and the number of clusters in this example dataset is only 49 and thus the power of rejecting the null hypothesis of a well-fitting model when the model does not in reality fit well could be small for this example. However, Hoenig and Heisey (2001) explain that that this is a misconception about the relationship between observed power and  $p$ -value which often appears in the applied literature.

## 6. Discussion and Conclusions

### 6.1 Discussion

Many of the tests proposed for goodness of fit of binary data can be used only under the assumption that the observations in the sample are independent, that is under the single level case (Liu, 2007). It is not always possible to deal with single level data structures and there are many instances in practice where the dataset is clustered or hierarchical, resulting in multilevel data structures. One of the methods of handling the within cluster correlation in multilevel data is by fitting a multilevel model (Steenbergen and Jones, 2002). Although multilevel modeling can be used, for discrete responses, there are very few satisfactory techniques to assess the goodness of fit of the fitted multilevel models in specialized packages like MLwiN (Browne, 2004).

Therefore, the main objective of this research was to develop a goodness of fit test to assess model adequacy of multilevel binary logistic models. The secondary objective was to identify the properties of the developed goodness of fit test under different scenarios; that is, to test the effect of different numbers of clusters, different numbers of observations per cluster and different intra cluster correlations on the power of the test and Type I error rates for the developed novel technique.

In order to achieve the above-mentioned prime objective, this research extends the method of Hosmer and Lemeshow (1980) to the multilevel binary logistic regression model. To identify the properties of the developed test, a simulation study was used. This was subdivided into two main sections. These are simulations to determine the Type I error rate of the developed test and simulation to identify the power of the developed test. In the case of the simulation study for Type I error, the twelve different sets of simulations cover twelve different scenarios. These different simulation studies provide evidence that the Type I error holds for the novel goodness of fit test for eleven out of twelve conditions. In this simulation study, it is evident that when the number of clusters and the cluster size are both small, the convergence probability is marginal. The reason for this can be explained by the fact that a small number of clusters and a small cluster size results in poor

estimation of the fixed and random coefficients leading to bias in the joint Wald statistics and hence marginal convergence probabilities (Maas et al., 2005). The study of Moineddin et al. (2007) explains the reason behind this phenomena well. The marginal Type I error rate for small sample size arising from a small number of clusters of small size can also be attributed to the fact that in the developed goodness of fit test, the method of indicator variable allocation works only for at least moderately large samples, as explained by Rosner et al. (2003). In the simulation studies for power, the developed goodness of fit test gave highest power against the mis-specified functional form when both the number of clusters and the number of observations per cluster were large, the standard deviation of the random effect was low and the coefficient of variation of the covariate was high. As the number of clusters and the number of observations per cluster decreases, the developed goodness of fit test tends to have smaller power. When both the number of clusters and the cluster size are small, the developed test shows lowest power for any given ICC value. This phenomenon of low sample size resulting in low power was also seen in Hosmer and Lemeshow (1980) original test. It was observed that as the standard deviation of the random effect increased from 1.0 to 2.0, the power of the test decreased by small amounts for all combinations. It was seen that as the coefficient of variation of the covariate increased, the power of the test also increased.

The application to the dataset illustrated that this test can be easily generalized to the case of an unequal number of observations per cluster.

## 6.2 Conclusions

The developed goodness of fit test is a direct generalization of the Hosmer and Lemeshow test (1980) for a single level logistic regression model. In order to apply the newly proposed goodness of fit test, explanatory variables can be categorical or continuous and from any distribution. The developed test can be applied with unequal cluster sizes. The developed test is not based on complex theories and anyone who can understand the single level Hosmer and Lemeshow test can easily understand this test. For the multilevel binary logistic model, the Type I error holds for the developed test for a large number of clusters with a large cluster size, a large number of clusters with a small cluster size and for a small number of clusters with a large cluster size. The developed goodness of fit test has low power when there is a small number of clusters with a small number of observations per cluster and high power for a large number of clusters with a large number of observations per cluster. The same phenomenon was observed in Hosmer and Lemeshow (1980) test for a single level binary logistic regression model and since this novel technique is an extension of the Hosmer and Lemeshow (1980) method it is so for this test also. The developed goodness of fit test is superior to all the goodness of fit tests considered by Archer et al. (2006) in terms of the Type I error rate. It is comparable to the goodness of fit test of Sturdivant and Hosmer (2007) in terms of Type I error but is far simpler and easier to understand. Also, Sturdivant and Hosmer (2007) have not studied the power of their test.

## 6.3 Further Work

We are working towards applying this test to the random slope model and extending it to the ordinal categorical case and the proportional odds model.



## Acknowledgments

The authors are grateful to Dr. Romaine Jayawardene of the Department of Mathematics, University of Colombo, Sri Lanka, for editing this manuscript.

## Supplemental Material

Supplemental data for this article can be accessed on the publisher's website at <http://dx.doi.org/10.1080/>

## References

- Abeysekera, W. W. M., Sooriyachchi, R. (2008). A novel method for testing goodness of fit of a proportional odds model: An application to AIDS study. *Journal of National Science Foundation Sri Lanka* 36(2):125–135.
- Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York; John Wiley & Sons.
- Archer K. J., Lemeshow S. (2006). Goodness-of-fit test for a logistic regression model fitted using survey sample data. *The Stata Journal* 6(1):97–105.
- Archer, K. J., Lemeshow, S., Hosmer, D. W. (2007). Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics & Data Analysis* 51:4450–4464.
- Blanchet, F. G., Legendre, P., Borcard, D. (2008). Forward selection of explanatory variables. *Ecological Society of America* 89(9):2623–2632.
- Browne, W. J. (2004). *A User's Guide to MLwiN*. Center for Multilevel Modeling, University of Bristol, Bristol, UK.
- Goldstein, H. (2003). *Multilevel Statistical Models*. London: Arnold.
- Graubard, B. I., Korn, E. L., Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. In *Proceeding of the Section on Survey Research Methods*. Alexandria, VA: *American Statistical Association*, pp. 170–174.
- Hdett, D. C., (1999) *Goodness of Fit Tests in Logistic Regression*. Ph.D. thesis, University of Toronto, Toronto, Canada.
- Hoenig, J. M., Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician* 2001, 55(1):1–6
- Hosmer, D. W., Lemeshow, S. (1980). A goodness-of-fit test for multiple logistic regression model. *Communications in Statistics, Theory and Methods A* 9(10):1043–1069.
- Hosmer, D. W., Lemeshow, S. (1989). *Applied Logistic Regression*. New York: John Wiley & Sons.
- Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley & Sons.
- Hosmer, D. W., Lemeshow, S., Klar, J. (1988). Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. *Biometrical Journal* 30(8):911–924.
- Huber, P. J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 221–233.
- Huq, N. M., Cleland, J. (1990). *Bangladesh Fertility Survey, 1980*. Dhaka: National Institute of Population Research and Training (NIPORT).
- Kreft, I. G. G., De Leeuw, J. (1998). *Introducing Multilevel Modeling*. Newbury Park, CA: Sage.
- Lee, E. S., Forthofer, R. N., Lorimor, R. J. (1989). *Analyzing Complex Survey Data*. Newbury Park: Sage.
- Lemeshow, S., Hosmer, W. (1982). A review of goodness-of-fit statistics for use in the development of logistic regression models. *The American Journal of Epidemiology* 115:92–106.
- Liang, K., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22.
- Liao, T.F. (2004). Comparing social groups: Wald statistics for testing equality among multiple logit models *International Journal of Comparative Sociology* 45:3–16,

- Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G. (1996). Goodness-of-fit test for ordinal response regression models. *Journal of the Royal Statistical Society* 45(2):175–190.
- Liu, Y. (2007). *On Goodness of Fit of Logistic Regression Models*. Kansas State University. Manhattan, Kansas.
- Maas, C. J., Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* 86–92.
- Moineddin, R., Matheson, F. I., Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology* 7:34.
- Pardoe, I. (2004). Model assessment plots for multilevel logistic regression. *Computational Statistics & Data Analysis* 46(2):295–307.
- Parson, V. L. (1992). *Using the Sampling Design in Logistic Regression Analysis of NCHS Survey Data Some Applications*. Alexandria, USA: American Statistical Association.
- Rosner, B., Glynn, R. J., Tinglee, M. L. (2003). Incorporation of clustering effects for the Wilcoxon rank sum test: A large-sample approach. *BIOMETRICS* 59:1089–1098.
- Steenbergen, M. R., Jones, B. S. (2002). Modeling multilevel data structures. *American Journal of Political Science* 46(1):218–237.
- Sturdivant, R. X., Hosmer, D. W. (2007). Smoothed residual based goodness-of-fit statistics for logistic hierarchical regression models. *Computational Statistics & Data Analysis* 51(8):3898–3912.
- Tsiatis, A. A. (1980). A note on a goodness-of-fit for the logistic regression model. *Biometrika* 67:250–251.
- Van, D. L. R., Busing, F., Meijer, E. (1997). Applications of bootstrap methods for two-level models. In *Paper presented at Multilevel Conference*. Amsterdam.
- Zorn, C. J. (2001). Generalized equation models for correlated data: A review with applications. *American Journal of political Science* 45:470–490.