# Smoothed Residual Based Goodness-of-Fit Statistics for Logistic Hierarchical Regression Models

Rodney X. Sturdivant
Department of Mathematics Sciences
United States Military Academy, West Point, New York

David W. Hosmer, Jr.
Department of Biostatistics and Epidemiology
University of Massachusetts – Amherst, Amherst, MA

## 1 Executive Summary

We extend goodness-of-fit measures used in the standard logistic setting to the hierarchical case. We develop theoretical asymptotic distributions for a number of statistics using residuals at the lowest level. Using simulation studies we examine the performance of statistics extended from the standard logistic regression setting: the Unweighted Sums of Squares (USS), Pearson residual and Hosmer-Lemeshow statistics. Our results suggest such statistics do not offer reasonable performance in the hierarchical logistic model in terms of Type I error rates. We also develop Kernel smoothed versions of the statistics and apply a bias correction method to the USS and Pearson statistics. Our simulations demonstrate satisfactory performance of the Kernel smoothed USS statistic, using Type I error rates, in small sample settings. Finally, we discuss issues of bandwidth selection for using our proposed statistic in practice.

## 2 Introduction

The logistic regression model is a widely used and accepted method of analyzing data with binary outcome variables. The standard logistic model does not easily address the situation, common in practice, in which the data is clustered or has a natural hierarchy. . For example, in education students are grouped by teachers, schools and districts. In medicine, patients may have the same doctor or use the same clinic or hospital. In recent years, statistical research has led to development of models that explicitly account for the hierarchical nature of the data. Many of these models are now available in commonly used software packages. While use of the models has increased, the development of methods to assess model adequacy and fit has not been commensurate with their popularity.

## 3 The Hierarchical Logistic Regression Model

The standard logistic approach models the probability that $Y$ takes on the value one, denoted $\pi = \Pr(Y = 1)$. For simplicity, first consider the case where there are two levels in the hierarchy. Further, suppose in this situation there is a single predictor variable. Ignoring the second level, the standard logistic regression model is:

$$Y_{ij} = \pi_{ij} + \varepsilon_{ij},$$
$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{ij}, \quad (1)$$

where $i = 1,...,n_j$ is the subject or level one indicator and $j = 1,...,J$ is the group or level two indicator. The model assumes the distribution of the outcome variable is binomial: $Y_{ij} \sim \mathrm{B}(1, \pi_{ij})$. The standard assumptions about the error structure are then that the errors are independent with moments:

$$\text{put in text line } \mathrm{E}(\varepsilon_{ij}) = 0 \text{ and } \mathrm{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2 = \pi_{ij}(1-\pi_{ij}).$$

.

The hierarchical logistic regression model accounts for the structure of the data by introducing random effects to model (1). In this case, with two levels, we might suppose that either or both coefficients (intercept and slope of the linear logit expression) vary randomly across level two groups. Assuming both are random the hierarchical logistic model is written:

$$\mathrm{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \beta_{1j}x_{ij}, \qquad (2)$$

with $\beta_{0j} = \beta_0 + \mu_{0j}$, and $\beta_{1j} = \beta_1 + \mu_{1j}$. The random effects are typically assumed to have a normal distribution so that $\mu_{0j} \sim \mathrm{N}(0, \sigma_0^2)$ and $\mu_{1j} \sim \mathrm{N}(0, \sigma_1^2)$. Further, the random effects need not be uncorrelated so we have $\mathrm{Cov}(\mu_{0j}, \mu_{1j}) = \sigma_{01}$. Assumptions about $\varepsilon_{ij}$ (the level one errors) remain the same as in the standard logistic model.

Substituting the random effects into expression (2) and rearranging terms, the model is:

$$\mathrm{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = (\beta_0 + \beta_1 x_{ij}) + (\mu_{0j} + \mu_{1j}x_{ij}). \qquad (3)$$

In this version of the model, we see a separation of fixed and random components which suggests a general matrix expression for the hierarchical logistic regression model given by:

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\pi} + \boldsymbol{\varepsilon} \\ \mathrm{logit}(\boldsymbol{\pi}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu}, \end{aligned} \qquad (4)$$

where $\mathbf{y}$ is an $N \times 1$ vector of the binary outcomes; $\boldsymbol{\pi}$ the vector of probabilities; $\mathbf{X}$ is a design matrix for the fixed effects; and $\boldsymbol{\beta}$ a $p \times 1$ vector a parameters for the fixed portion of the model. The level one errors have mean zero and variance given by the diagonal matrix of binomial variances:

$$\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{W} = \mathrm{diag}[\pi_{ij}(1-\pi_{ij})].$$

Choppy These quantities, then, are the same as in standard logistic models. The quantities added to the model to introduce the random effects are the design matrix for the random effects, $\mathbf{Z}$, and the vector of random parameters, $\boldsymbol{\mu}$. This latter vector has assumed distribution $\boldsymbol{\mu} \square N(\mathbf{0}, \boldsymbol{\Omega})$ with a block diagonal covariance matrix.

Section on estimation

Several methods are available for estimating the parameters of this model. By conditioning on the random effects and then integrating them out, an expression for the maximum likelihood estimates is available. This integral is difficult to evaluate, but recently estimation techniques using numerical integration, such as adaptive Gaussian quadrature, have been implemented in software packages. This method is computational intensive and suffers from instability. In some packages, the ability to handle larger models is lacking.

Don't expand on EM

A second closely related method uses the E-M algorithm to maximize the conditional likelihood function [1]. In this case, the random effects are treated as "missing data" and the algorithm, in the "E" (Expectation) step estimates these parameters by obtaining their conditional (on the data and current estimates of the fixed parameters) expected values. Then, with random parameter estimates in place, the "M" or Maximization step is invoked in which standard generalized least squares (GLS) estimates of the fixed parameters are calculated. The algorithm alternates between E and M steps until some convergence criteria is met. The E-M algorithm also involves heavy computation and is not available in most commercial software packages.

Bayesian methods of estimation have increased in popularity although they have not been implemented in the more popular software packages. Gibbs Sampling [3] and Metropolis-Hastings (M-H) are Markov Chain Monte Carlo simulation techniques [4] typically used to produce parameter estimates under this approach. Again, the techniques involve heavy computation.

The most readily available methods in software packages involve quasi-likelihood estimation [5]. For the logistic hierarchical model the idea is generally to use a Taylor approximation to "linearize" the model. The estimation is then iterative between fixed and random parameters. These procedures suffer from known bias in parameter estimates [6]. However, there are methods to reduce this bias available [7]. Further, the methods are easily implemented and generally converge with less computational effort than other methods. Throughout this study we use the SAS GLMMIX macro which implements a version of quasilikelihood estimation SAS refers to as PL or "pseudo-likelihood" [8].

# 4 Theoretical Asymptotic Distribution Development of ?

better start of section

Copas [9] proposed the unweighted sum of squares (USS) statistic as a goodness-of-fit measure for the standard logistic model. If consistent number of subscripts $y_i$ is the observed response for the $i^{\text{th}}$ subject (here we are only concerned with indexing at level

one) and $\hat{\pi}_i$ the model predicted value based on the estimated parameters, the USS statistic is the sum of the squared residuals, $\hat{e}_i = y_i - \hat{\pi}_i$, or:

$$\hat{S} = \hat{\mathbf{e}}^t\hat{\mathbf{e}} = \sum (y_i - \hat{\pi}_i)^2$$

Hosmer et. al. give the asymptotic moments of $\hat{S}$ for the standard logistic case as:

$$\mathrm{E}(\hat{S}) \cong \mathrm{trace}(\mathbf{W})$$

and

$$\mathrm{Var}[\hat{S} - \mathrm{trace}(\mathbf{W})] \cong \mathbf{d}'\mathbf{(I - M}_1\mathbf{)Wd},$$

where $\mathbf{d}$ is the vector with general element $d_i = 1 - 2\pi_i$, $\mathbf{W}$ is the covariance matrix in standard logistic regression given by $\mathbf{W} = \mathrm{diag}[w_i = \pi_i(1-\pi_i)]$, and $\mathbf{M}_1 = \mathbf{WX(X'WX)}^{-1}\mathbf{X}'$ is the logistic regression version of the "hat" matrix.

Model fit is then assessed forming a standardized version of the statistic for comparison to the standard normal distribution:

$$\frac{\hat{S} - \mathrm{trace}(\hat{\mathbf{W}})}{\sqrt{\hat{\mathrm{Var}}[\hat{S} - \mathrm{trace}(\hat{\mathbf{W}})]}}.$$

Evans follows a similar procedure to produce a standardized statistic for a logistic 2-level mixed model with random intercept only. Our simulation studies of a version of this statistic in models with random slopes suggest that the theoretical normal distribution under the null hypothesis of a correctly specified model does not hold in smaller samples typically encountered in practice. The statistic itself is inflated due to either large or small observations in the covariance matrix.

le Cessie and van Houwelingen [12] note similar problems with goodness-of-fit measures in certain standard logistic regression settings. They observe a shrinkage effect when considering the approximation for the test statistic. The estimated test statistic may be written as the statistic with true values minus a quantity that is always positive. In order to control the problem in the standard logistic case they use kernel smoothing of Pearson residuals. We similarly develop a USS statistic in the hierarchical logistic model using kernel smoothed residuals.

The smoothed residuals are weighted average of the residuals which controls for issues with extremely large or small values. One can perform kernel smoothing of the residuals in either the "y-space" or "x-space" [10]. In the "x-space" all covariates are used in developing the weights. In the "y-space", the weights are produced using relative distances of the model predicted probabilities of the outcome given by:

$$\hat{\boldsymbol{\pi}} = \begin{pmatrix} \hat{\pi}_1 \\ \vdots \\ \hat{\pi}_n \end{pmatrix}.$$

We use Kernel smoothing of the residuals in the "y-space" in this research. In the standard logistic setting the difference between the two approaches was negligible [10]. The "y-space" smoothing is somewhat simpler and, as demonstrated in the next section, produces reasonable results.

The vector of smoothed residuals is given by:

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

where $\boldsymbol{\Lambda}$ is the matrix of smoothing weights:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & & \lambda_{1n} \\ & \ddots & \\ \lambda_{n1} & & \lambda_{nn} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \\ \boldsymbol{\lambda}_n \end{bmatrix}.$$

The weights, $\lambda_{ij}$, are produced using the kernel density by:

$$\lambda_{ij} = \frac{\mathrm{K}\left(\frac{|\hat{\pi}_i - \hat{\pi}_j|}{h}\right)}{\sum_j \mathrm{K}\left(\frac{|\hat{\pi}_i - \hat{\pi}_j|}{h}\right)}. \tag{5}$$

where $\mathrm{K}(\xi)$ is the Kernel density function and *h* is the bandwidth.

We explore three choices used in other studies for the Kernel density function. The first was the uniform density used in a study of a goodness-of-fit measure in standard logistic regression [12] defined as:

$$\mathrm{K}(\xi) = \begin{cases} 1 \text{ if } |\xi| < 0.5 \\ 0 \text{ otherwise} \end{cases}.$$

A second choice used in standard logistic studies involving smoothing in the "y-space" ([10] and [13]) was the cubic kernel given by:

$$K(\xi) = \begin{cases} 1 - |\xi|^3 \text{ if } |\xi| < 1 \\ 0 \text{ otherwise} \end{cases}.$$

Finally, we tested the Gaussian Kernel density [14] defined:

$$K(\xi) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\xi^2) .$$

The bandwidth, *h*, controls the number of observations weighted in the case of the uniform and cubic densities. For the Gaussian Kernel, all observations are weighted. However, observations outside of two or three standard deviations of the mean effectively receive zero weight. The bandwidth then determines how many residuals are effectively given zero weight in the Gaussian case.

The choice of Kernel function is considered less critical than that of the bandwidth [15]. There are several methods available (plug-in, cross-validation etc.) for selecting the "optimal" bandwidth. Here we are more concerned with the efficacy of smoothing as an approach. Thus, we examine several bandwidth choices.

Simulations suggest that, using the uniform Kernel in the Pearson statistic for standard logistic models, a bandwidth in which approximately $\sqrt{n}$ of the observations have non-zero weights is best and the weighting too many observations is too conservative [12]. The same criteria worked well with the cubic Kernel in the standard logistic case [10].

Some preliminary work suggested that the use of fewer observations is preferred in the hierarchical setting. Sentence makes no sense ->This is not surprising as the shrinkage effect in the statistic appears even more pronounced. We thus test the bandwidth weighting $\sqrt{n}$ of the residuals for the uniform and cubic kernel for each $\hat{\pi}_i$, as well as smaller bandwidths so that $0.5\sqrt{n}$ or $0.25\sqrt{n}$ of the kernel values were not zero. For the Gaussian kernel, the chosen bandwidth places the selected number of observations within two standard deviations of the mean of the $N(0,1)$ density used in the kernel estimation (somewhat analogous to the other kernels as outside of 2 standard deviations the weights are extremely small in the normal density).

Regardless of the bandwidth criteria, we choose a different bandwidth $h_i$ for each $\hat{\pi}_i$ (in the fashion of Fowlkes, [13]). The weights are then standardized so that they sum to one for each $\hat{\pi}_i$ by dividing by the total weights for the observation as shown in expression (5).

The USS statistic based upon these smoothed residuals is then given by:

$$\hat{S}_s = \sum_{i=1}^{n} \hat{e}_{si}^2 = \hat{\mathbf{e}}_\mathbf{s}' \hat{\mathbf{e}}_\mathbf{s}$$

The distribution (moments?) of this statistic under the null hypothesis that the model is correctly specified is extremely complicated. However, we can produce expressions to approximate the moments of the statistic. We make first approximate the residuals in terms of the level one errors [16]:

$$\hat{\mathbf{e}} \approx (\mathbf{I} - \mathbf{M})\mathbf{e} + \mathbf{g} , \qquad (6)$$

where take out of in line equations $\mathbf{M} = \mathbf{WQ}\left[\mathbf{Q}'\mathbf{WQ} + \mathbf{R}\right]^{-1}\mathbf{Q}'$ and $\mathbf{g} = \mathbf{WQ}\left[\mathbf{Q}'\mathbf{WQ} + \mathbf{R}\right]^{-1}\mathbf{R}\boldsymbol{\delta}$. In these expressions, $\mathbf{Q} = \left[\mathbf{X} \quad \mathbf{Z}\right]$ is the design matrix for both fixed and random effects, and $\hat{\boldsymbol{\delta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\mu}} \end{pmatrix}$ the vector of estimated fixed and random effects. The other matrix in the expression involves the estimated random parameter covariances and is defined: $\mathbf{R} = \begin{bmatrix} 0 & 0 \\ 0 & \hat{\boldsymbol{\Omega}}^{-1} \end{bmatrix}$.

Under the null hypothesis of correct model specification, the errors have known moments allowing us to produce the approximate mean and variance for the statistic. We first write the statistic using the approximation of (6):

$$\begin{aligned} \hat{S}_s &= \hat{\mathbf{e}}_s'\hat{\mathbf{e}}_s \\ &= \hat{\mathbf{e}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\hat{\mathbf{e}} \\ &\approx [(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + \hat{\mathbf{g}}]'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}[(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + \hat{\mathbf{g}}] \\ &= \mathbf{e}'(\mathbf{I} - \hat{\mathbf{M}})'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + 2\hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + \hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\hat{\mathbf{g}}. \end{aligned}$$

Standard methods to calculate the expected value of a quadratic form (for example, [17]) allow us to express the first moment as:

$$\begin{aligned} \mathrm{E}(\hat{S}_s) &= \mathrm{E}[\mathbf{e}'(\mathbf{I} - \hat{\mathbf{M}})'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + 2\hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + \hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\hat{\mathbf{g}}] \\ &= \mathrm{trace}[(\mathbf{I} - \hat{\mathbf{M}})'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{W}] + \hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\hat{\mathbf{g}} \ . \end{aligned} \tag{7}$$

The variance is expressed as:

$$\begin{aligned} \mathrm{Var}(\hat{S}_s) &= \mathrm{Var}[\mathbf{e}'(\mathbf{I} - \hat{\mathbf{M}})'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + 2\hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})\mathbf{e} + \hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\hat{\mathbf{g}}] \\ &= \mathrm{Var}(\mathbf{e}'\mathbf{A}_4\mathbf{e}) + \mathrm{Var}(\mathbf{b}_4'\mathbf{e}) + 2\,\mathrm{Cov}(\mathbf{e}'\mathbf{A}_4\mathbf{e}, \mathbf{b}_4'\mathbf{e}) \end{aligned}$$

where no 4: $\mathbf{A}_4 = (\mathbf{I} - \hat{\mathbf{M}})'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})$ and $\mathbf{b}_4' = 2\hat{\mathbf{g}}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}(\mathbf{I} - \hat{\mathbf{M}})$. To evaluate this expression we use a lesser known result [18] so that the final expression becomes:

$$\begin{aligned} \mathrm{Var}(\hat{S}_s) &= \mathrm{Var}(\mathbf{e}'\mathbf{A}_4\mathbf{e}) + \mathbf{b}_4'\hat{\mathbf{W}}\mathbf{b}_4 + 2\,\mathrm{Cov}(\mathbf{e}'\mathbf{A}_4\mathbf{e}, \mathbf{b}_4'\mathbf{e}) \\ &= \sum_{i=1}^{n}[a_{4ii}^2 w_i(1-6w_i)] + 2\,\mathrm{trace}(\mathbf{A}_4\hat{\mathbf{W}}\mathbf{A}_4\hat{\mathbf{W}}) + \mathbf{b}_4'\hat{\mathbf{W}}\mathbf{b}_4 \\ &\quad +2\sum_i a_{4ii}b_{4i}\pi_i(1-\pi_i)(1-2\pi_i) \ . \end{aligned} \tag{8}$$

The moment expressions are then used to create a standardized statistic:

$$\frac{\hat{S}_s - \mathrm{E}(\hat{S}_s)}{\sqrt{\mathrm{Var}(\hat{S}_s)}}. \qquad (9)$$

Under the null hypothesis of correct model fit this statistic has don't use "theory here…reword: theoretical asymptotic standard normal distribution. In order to test model fit, the moments are evaluated using the model estimated quantities where necessary in expressions (7) and (8).

# 5 Simulation Study Results

The standardized statistic of expression (9) has a dittotheoretical standard normal distribution asymptotically. In a large enough sample, one would expect these statistics to appropriately reject the null hypothesis for a given rejection rate. In a hierarchical model "large enough sample" has two implications. First, the total sample must be large. Further, the number of subjects in each group should also be large. In practice, both conditions may not always be met. Usually the total sample size for hierarchical data is reasonably large, but the cluster sizes might still cause us to question the validity of asymptotic results. We used simulations to examine the performance of the statistics in settings with small sample and cluster sizes likely to occur in practice.

The simulation study consisted of 28 different settings involving four factors: dimension, number of covariates, Intra-class or Intra-cluster Correlation (ICC), and random effects.

The first factor, dimension, involved the number of levels in the hierarchy as well as cluster sizes. Noting that hierarchical models of more than three levels are rare in practice, we defined four levels for this factor:

1A: 2-level model with 20 groups of 20 subjects (400 subjects)
1B: 2-level model with 50 groups of 4 subjects (200 subjects)
1C: 2-level model with 25 groups of 4 subjects (100 subjects)
1D: 3-level model with 10 groups at level three, each with 5 subgroups of 4 subjects (200 subjects)

The second factor, , had 2 levels defined as:

2A: A single continuous covariate at each of level one and level two
2B: A single continuous covariate at level two and 5 covariates at level one (three continuous and two dichotomous)

The ICC was also broken into two levels. We used several measures to calculate the ICC and experimented to determine what values of each constituted high and low ICC values. One of these, $\rho_{hl}$ [19], is sufficient to give an idea of the factor levels:

3A: Moderately low ICC; this corresponds to $\rho_{hl}$ of roughly 0.20 to 0.24.

3B: Moderately high ICC; this corresponds to $\rho_{hl}$ of roughly 0.50 to 0.57.

The final factor involves the number of random effects in the models and was broken into three levels, again based on the most likely scenarios in previous studies:

4A: Random intercept (only in the three-level model).

4B: Random intercept and one random slope for a level one continuous covariate.

4C: Random intercept and two random slopes (level one continuous and dichotomous variables); available for factor 2 level 2B only.

The resulting 28 simulations are shown in Table 1.

**Table 1: Four Factor Simulation Study Design**

| SIMULATION | FACTOR 1 | FACTOR 2 | FACTOR 3 | FACTOR 4 |
|---|---|---|---|---|
| **1** | 1A | 2A | 3A | 4B |
| **2** | 1A | 2A | 3B | 4B |
| **3** | 1A | 2B | 3A | 4B |
| **4** | 1A | 2B | 3B | 4B |
| **5** | 1A | 2B | 3A | 4C |
| **6** | 1A | 2B | 3B | 4C |
| **7** | 1B | 2A | 3A | 4B |
| **8** | 1B | 2A | 3B | 4B |
| **9** | 1B | 2B | 3A | 4B |
| **10** | 1B | 2B | 3B | 4B |
| **11** | 1B | 2B | 3A | 4C |
| **12** | 1B | 2B | 3B | 4C |
| **13** | 1C | 2A | 3A | 4B |
| **14** | 1C | 2A | 3B | 4B |
| **15** | 1C | 2B | 3A | 4B |
| **16** | 1C | 2B | 3B | 4B |
| **17** | 1C | 2B | 3A | 4C |
| **18** | 1C | 2B | 3B | 4C |
| **19** | 1D | 2A | 3A | 4A |
| **20** | 1D | 2A | 3A | 4B |
| **21** | 1D | 2A | 3B | 4A |
| **22** | 1D | 2A | 3B | 4B |
| **23** | 1D | 2B | 3A | 4A |
| **24** | 1D | 2B | 3A | 4B |
| **25** | 1D | 2B | 3A | 4C |
| **26** | 1D | 2B | 3B | 4A |
| **27** | 1D | 2B | 3B | 4B |
| **28** | 1D | 2B | 3B | 4C |

We generated 1000 data sets for each of the 28 simulations outlined in the previous section. We then fit the appropriate hierarchical logistic model using the SAS Glimmix macro (PQL estimation). Finally, proposed kernel smoothed USS goodness-of-fit statistic was computed using the model output. In this study, we were concerned with rejection rates for the statistic when the correct model was fit to the data.

We considered, in particular, how often the null hypothesis was rejected at three commonly used significance levels ($\hat{\alpha}$): 0.01, 0.05 and 0.1. A statistic for which the asymptotic distribution continues to hold in the smaller samples rejects at the same rate as $\hat{\alpha}$ in the 1000 simulations. Using 1000 replications in each simulation, approximate

95% confidence intervals are within 0.6%, 1.4% and 1.9% of the respective values of $\hat{\alpha}$ used.

We simulated the statistic for three different choices of kernel density and bandwidth. In most simulations, the cubic Kernel density was best. In general, the three kernels appear similar, for their optimal bandwidth choice. In our study, the cubic might appear better due to having an optimal bandwidth nearest one of the three bandwidths we chose. In practice the density chosen appears to be much less important than the bandwidth.

The three simulated bandwidths ranged from the smallest which weighted the fewest subjects among the three choices (roughly $\frac{1}{4}\sqrt{n}$ subjects). The other two bandwidths weight more of the subjects: roughly $\frac{1}{2}\sqrt{n}$ subjects and $\sqrt{n}$ subjects respectively. After reviewing the results for these three bandwidth choices, the optimal choice appeared to weight fewer subjects than the smallest bandwidth in five of the simulation settings (simulations 7, 12, 13, 20 and 28). In those cases, we ran additional simulations to find the approximately optimal bandwidth choice.

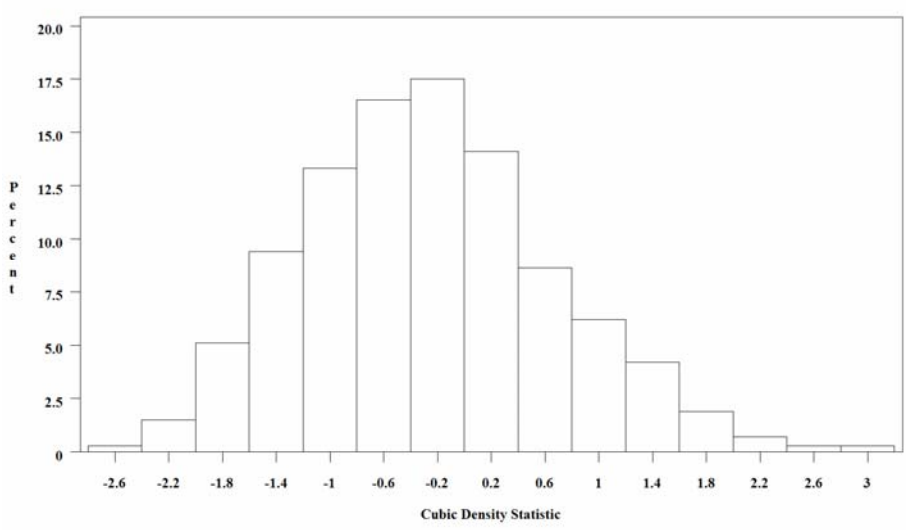
Results for the estimated optimal choice are shown in Table 2 for all 28 simulation settings using the cubic kernel density. In each case, the simulated rejection rate based on 1000 replications is displayed for each of the three significance levels ($\hat{\alpha}$): 0.01, 0.05 and 0.1. Shaded cells in the table are simulation runs in which the 95% confidence interval for the estimated rejection level includes the desired value.

**Table 2: USS Kernel Statistic (Cubic Kernel Density Function) Simulation Study Results**

| | Significance | | | | Significance | | |
|---|---|---|---|---|---|---|---|
| **Simulation** | **0.01** | **0.05** | **0.1** | **Simulation** | **0.01** | **0.05** | **0.1** |
| **1** | 0.02 | 0.062 | 0.103 | **15** | 0.011 | 0.049 | 0.093 |
| **2** | 0.009 | 0.032 | 0.085 | **16** | 0.006 | 0.042 | 0.076 |
| **3** | 0.011 | 0.047 | 0.084 | **17** | 0.014 | 0.039 | 0.094 |
| **4** | 0.011 | 0.039 | 0.083 | **18** | 0.013 | 0.038 | 0.08 |
| **5** | 0.017 | 0.062 | 0.1 | **19** | 0.016 | 0.052 | 0.091 |
| **6** | 0.016 | 0.042 | 0.084 | **20** | 0.035 | 0.03 | 0.06 |
| **7** | 0.01 | 0.04 | 0.09 | **21** | 0.011 | 0.052 | 0.089 |
| **8** | 0.018 | 0.055 | 0.097 | **22** | 0.014 | 0.064 | 0.115 |
| **9** | 0.008 | 0.049 | 0.102 | **23** | 0.007 | 0.035 | 0.074 |
| **10** | 0.009 | 0.048 | 0.087 | **24** | 0.014 | 0.062 | 0.11 |
| **11** | 0.014 | 0.051 | 0.099 | **25** | 0.017 | 0.06 | 0.112 |
| **12** | 0.01 | 0.04 | 0.07 | **26** | 0.012 | 0.037 | 0.074 |
| **13** | 0.01 | 0.04 | 0.08 | **27** | 0.016 | 0.053 | 0.11 |
| **14** | 0.009 | 0.031 | 0.07 | **28** | 0.01 | 0.04 | 0.08 |

reverse shading – maybe not shade
footnote to table indicating shading
As shown, the statistic rejects appropriately in nearly all simulation runs. The simulations suggest that the USS Kernel statistic is appropriate for use in logistic hierarchical models. The study includes a variety of small sample settings and the use of our theoretical asymptotic distribution performs admirably.

We do note that tests of normality typically reject (in all but five settings) the normal distribution in the simulation runs. However, we believe that this is in part due to the power to detect departures from normality with 1000 replications. Examination of the histogram (Figure 1) and QQ Plot (Figure 2) in a typical simulation setting suggests that the assumption of normality for the standardized statistic holds even in the small sample setting. We note a slight skew in the statistic but, coupled with rejection rates at various




significance levels, believe the statistic is appropriate for use in practice.

Remove figures and verbally describe

**Figure 1: Histogram of USS Kernel Smoothed Standardized Statistic Values in Simulation 2 (1000 replications)**

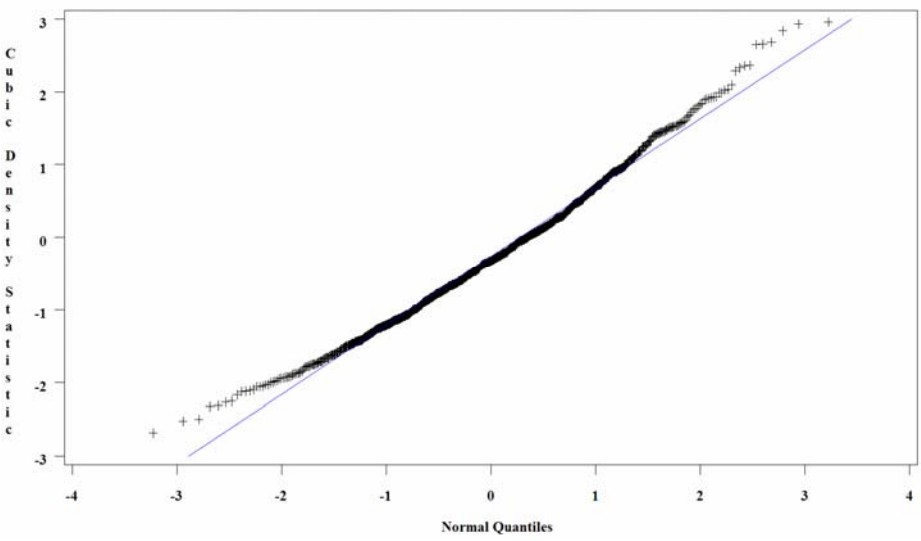


**Figure 2: QQ Plot of USS Kernel Smoothed Standardized Statistic Values in Simulation 2 (1000 replications)**

# 6 Discussion

end with what we didn't start with what we did We did not include results for several other versions of goodness-of-fit statistics that were explored in this study. These included a USS statistic using the residuals without smoothing, a statistic using the Pearson residuals (both with and without smoothing) and a version of the Hosmer-Lemeshow statistic. In each case, the theoretical asymptotic distribution did not hold for the small sample settings of our simulation study [16]. We do not recommend these statistics for use in practice.

We do recommend use of the USS Kernel smooth statistic but the choice of bandwidth deserves some discussion. The "optimal" bandwidth choice is not entirely clear and is a subject for further research. Without further study, we offer only a general rule for practice. For reasonably large cluster sizes (20) and number of groups (20) the bandwidth weighting approximately $\frac{1}{2}\sqrt{n}$ of the residuals works well. For smaller cluster or sample sizes, we recommend a smaller bandwidth ($\frac{1}{4}\sqrt{n}$). The scope of our study prevents us from speculating for other data schemes.

Based on our study, these are conservative bandwidth choices; if anything, the USS kernel statistic will reject a bit too often. In fact, we observed that the statistic will generally reject too often when the bandwidth chosen is too large. This suggests that a quick "sensitivity analysis" to bandwidth choice can help. If a larger choice does not reject the analyst can be reasonable certain the selected model is reasonable.

A summary goodness-of-fit statistic is not the only criterion to determine whether a model is acceptable. Rather, it is used to alert the analyst to a potential problem. The possibility of the statistic rejecting too often in isolated cases is not problematic. This result should prompt the model builder to look more closely at the model and data. If no unacceptable problems are discovered upon further research the model will generally still be useful.

The tendency of the statistic to reject too often in some simulation study settings using our recommended bandwidth choices is also somewhat mitigated by the ability to produce significant parameter estimates for those data schemes. We found that the amount of "over rejection" is greater in simulation runs in which one or the other of the random parameters is not "significant". Here, significance is based on the ratio of the estimate to its standard error (to form a z-statistic). In the five settings where are proposed bandwidths would reject too often, one of the random effects is often not significant. In such a situation, the analyst might choose a model excluding the random effect. The kernel smoothed statistic in settings with fewer random effects appears to perform well (in fact, even the unadjusted statistics may be useful in such cases [11]).

## References


[1] Guo, G. and Zhao, H. (2000), "Multilevel Modeling for Binary Data", Annual Reviews of Sociology, 26, 441-462.

[2] Bryk, A. and Raudenbush, S. (1992), Hierarchical Linear Models, Sage Publications, Newbury Park.

[3] Zeger, S. and Karim, M. (1991), "Generalized Linear Models with Random Effects; a Gibbs Sampling Approach", Journal of the American Statistical Society, 86, 79-102.

[4] Gilks, W., Richardson, S., and Spiegelhalter, D. (1996), Markov Chain Monte Carlo in Practice, Chapman and Hall, London.

[5] Breslow, N.E. and Clayton, D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models", Journal of the American Statistical Association, 88, 421, 9-25.

[6] Rodriguez, G. and Goldman, N. (1995), "An Assessment of Estimation Procedures for Multilevel Models with Binary Responses", Journal of the Royal Statistical Society A, 158, 1, 73-89.

[7] Goldstein, H. and Rasbash, J. (1996), "Improved Approximations for Multilevel Models with Binary Responses", Journal of the Royal Statistical Society A, 159, 3, 505-513.

[8] Wolfinger, R. and O'Connell, M. (1993), "Generalized Linear Mixed Models: A Pseudo-likelihood Approach", Journal Statistical Computation and Simulation, 48, 233-243.

[9] Copas, J. (1989), "Unweighted Sum of Squares Test for Proportions", Applied Statistics, 38, 1, 71-80.

[10] Hosmer, D., Hosmer, T., le Cessie, S., and Lemeshow, S. (1997), "A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model", Statistics in Medicine, 16, 965-980.

[11] Evans, S. (1998), "Goodness-of-Fit in Two Models for Clustered Binary Data", Ph.D. Dissertation, University of Massachusetts Amherst, Ann Arbor: University Microfilms International.

[12] le Cessie, S., and van Houwelingen, J. (1991), "A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods", Biometrics, 47, 1267-1282.

[13] Fowlkes, E. (1987), "Some Diagnostics for Binary Logistic Regression via Smoothing Methods", Biometrika, 74, 503-515.

[14] Wand, M. and Jones, M. (1995), Kernel Smoothing, Chapman & Hall/CRC, Boca Raton.

[15] Hardle, W. (1990), Applied Nonparametric Regression, Cambridge University Press, Cambridge.

[16] Sturdivant, R. (2005), "Goodness-of-Fit in Hierarchical Logistic Regression Models", Ph.D. Dissertation, University of Massachusetts Amherst, Ann Arbor: University Microfilms International.

[17] Searle, S. (1982), Matrix Algebra Useful for Statistics, John Wiley and Sons, New York.

[18] Seber, G. (1977), Linear Regression Analysis, John Wiley and Sons, New York.

[19] Hosmer, D. and Lemeshow, S. (2000), Applied Logistic Regression 2nd Edition, John Wiley & Sons, Inc., New York.