STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4081–4098
Published online 3 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2502

Improved hypothesis testing for coefficients in generalized estimating equations with small samples of clusters

Daniel F. McCaffrey[1,*,†] and Robert M. Bell[2]

[1]*The RAND Corporation, 201 North Craig Street, Suite 202, Pittsburgh, PA 15213, U.S.A.*
[2]*Statistics Research Department, AT&T Labs-Research, 180 Park Ave., Florham Park, NJ 07932, U.S.A.*

SUMMARY

The sandwich standard error estimator is commonly used for making inferences about parameter estimates found as solutions to generalized estimating equations (GEE) for clustered data. The sandwich tends to underestimate the variability in the parameter estimates when the number of clusters is small, and reference distributions commonly used for hypothesis testing poorly approximate the distribution of Wald test statistics. Consequently, tests have greater than nominal type I error rates. We propose tests that use bias-reduced linearization, BRL, to adjust the sandwich estimator and Satterthwaite or saddlepoint approximations for the reference distribution of resulting Wald t-tests. We conducted a large simulation study of tests using a variety of estimators (traditional sandwich, BRL, Mancl and DeRouen's BC estimator, and a modification of an estimator proposed by Kott) and approximations to reference distributions under diverse settings that varied the distribution of the explanatory variables, the values of coefficients, and the degree of intra-cluster correlation (ICC). Our new method generally worked well, providing accurate estimates of the variability of fitted coefficients and tests with near-nominal type I error rates when the ICC is small. Our method works less well when the ICC is large, but it continues to out-perform the traditional sandwich and other alternatives. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: logistic regression; generalized linear models; linearization; sandwich estimator; saddlepoint approximation; complex samples

1. INTRODUCTION

Clustered designs complicate production of accurate standard errors, tests, and confidence intervals for coefficients of logistic regression and other generalized linear models. These designs, common in many applications, are likely to result in positive correlation among

*Correspondence to: D. F. McCaffrey, The RAND Corporation, 201 North Craig Street, Suite 202, Pittsburgh, PA 15213, U.S.A.
†E-mail: daniel_mccaffrey@rand.org

Contract/grant sponsor: National Science Foundation; contract/grant number: 00017630

Received 16 December 2003
Accepted 5 December 2005

outcomes from the same cluster. Many analysts utilize generalized estimating equations (GEE), which require specification for only the first two moments of the outcomes [1], to try to improve the efficiency of estimated regression coefficients. Often, the empirical sandwich estimator of the covariance matrix of the regression coefficients is used to provide standard error estimates that are consistent even if the working covariance model is mis-specified [1]. Inferences about individual coefficients or linear combinations of the coefficients use Wald-like t-statistics (i.e. the ratio of a coefficient to its estimated standard error) for hypothesis testing or as pivotal values for constructing confidence intervals. The distributions of the statistics are approximated by either the standard normal distribution or a t distribution with degrees of freedom equal to the number of clusters less one [1, 2].

However, these procedures perform poorly when the number of clusters is small. Type I errors for tests generally exceed the nominal values and the coverage rates for confidence intervals are less than the nominal values [3–8]. For linear models fit to data with 20 clusters, we found that type I error rates typically exceeded 1.4 times the nominal values and could even exceed three times the nominal values for coefficients for highly clustered explanatory variables [7]. Similarly for logistic regression, Mancl and DeRouen [6] found that with 20 clusters of data, type I errors for tests exceeded nominal levels by at least 1.4 times. Gunsolley *et al.* [5] found that type I errors for GEE exceeded twice the nominal rate and were extremely high when the true value of the coefficient was far from zero, the intra-cluster correlation (ICC) was high and the sample had 20 clusters of 100 observations each.

Our work with linear models [7] suggests that the poor performance of these procedures is due both to bias in the sandwich estimator and to underestimation of the estimator's variability. We developed an alternative estimator that is unbiased when the errors are i.i.d. and greatly reduced bias for correlated errors in a simulation study of linear models. We also developed an approximation to the distribution of the test statistic based on the true distribution of the variance estimator. Our proposed test procedure, called bias-reduced linearization (BRL), had near-nominal type I error rates in our simulation study even in situations when the traditional sandwich procedures have error rates exceeding 7 per cent for a nominal 0.05-level test [7].

For nonlinear models such as binary logistic regression, Lipsitz *et al.* [3] proposed a one-step jackknife procedure for estimating a coefficient and its standard error. Mancl and DeRouen [6] proposed an alternative aimed at correcting the bias in the sandwich estimator for the GEE framework. In contrast, Pan and Wall [8] ignored the bias and derived reference distributions for test statistics that account for the variability of the sandwich estimator.

In this paper, we extend our BRL methods for linear models to GEE for generalized linear models. Unlike the work just referenced, we develop both an alternative standard error estimator and approximations for the distribution of Wald-like t-test statistics under the null hypothesis. Given that estimators for GEE are derived from large sample approximations rather than exact small sample results, we conduct an extensive simulation study of our test procedure and some alternative methods when the number of clusters is small.

Section 2 contains a brief description of the traditional sandwich estimator and Wald and score tests based on this estimator. Section 3 describes our BRL method for linear models and extends the method to GEE including: development of the variance estimator and Satterthwaite and saddlepoint approximations for the reference distributions of Wald and score tests. The design and results of our simulation study follow, and the paper ends with discussion.

2. TESTS BASED ON THE SANDWICH ESTIMATOR

We consider generalized linear models [9] fit to clustered data. For clusters $i=1,\ldots,K$, and individuals $j=1,\ldots,n_i$ in cluster i, the data contain outcome measurements y_{ij} and p-dimensional vectors $\mathbf{x}_{ij}$ of explanatory variables, which may be either individual or cluster-level. The marginal mean $\mu_{ij}=E(y_{ij})$ is related to the explanatory variables $\mathbf{x}_{ij}$ by $g(\mu_{ij})=\mathbf{x}'_{ij}\boldsymbol{\beta}$, for a known link function g. We assume that the model for the mean is specified correctly. $\mathrm{Var}(y_{ij})=\phi v(\mu_{ij})$, where v is a known function of the mean and ϕ is a possibly unknown scale parameter that might be estimated from the data. We let $\mathbf{y}_i=(y_{i1},\ldots,y_{in_i})'$, $\boldsymbol{\mu}_i=(\mu_{i1},\ldots,\mu_{in_i})'$ and $\mathrm{Var}(\mathbf{y}_i)=\mathbf{V}_i=\phi\mathbf{A}_i^{1/2}\mathbf{R}_{0i}\mathbf{A}_i^{1/2}$, where $\mathbf{A}_i=\mathrm{diag}(v(\mu_{i1}),\ldots,v(\mu_{in_i}))$ and $\mathbf{R}_{0i}$ is a correlation matrix whose structure is in general unknown.

Under weak assumptions [1], consistent estimates of $\boldsymbol{\beta}$ obtain for a model fit with working correlation matrices $\mathbf{R}_{Wi}$, even if the working correlation is mis-specified. The estimate, $\widehat{\boldsymbol{\beta}}$, solves the equations:

$$\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{U}_i^{-1}(\mathbf{y}_i-\widehat{\boldsymbol{\mu}}_i)=0 \tag{1}$$

where $\widehat{\boldsymbol{\mu}}_i$ is the vector of estimated means for observations in the ith cluster with $\widehat{\mu}_{ij}=g^{-1}(\mathbf{x}'_{ij}\widehat{\boldsymbol{\beta}})$, $\mathbf{D}_i=\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}'|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$ and $\mathbf{U}_i=\phi\mathbf{A}_i^{1/2}\mathbf{R}_{Wi}\mathbf{A}_i^{1/2}$ with $\mathbf{A}_i$ evaluated at $\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}$. Optionally, $\mathbf{R}_{Wi}$ may depend on parameters $\boldsymbol{\alpha}$, which might be specified or estimated from the data.

If the working correlation matrix is taken as correct, then the covariance of $\widehat{\boldsymbol{\beta}}$ is estimated by $\mathbf{V}_{\mathrm{GEE}}=(\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{U}_i^{-1}\mathbf{D}_i)^{-1}$, with $\mathbf{D}_i$'s evaluated at $\widehat{\boldsymbol{\beta}}$. However, in the usual situation where that assumption is unwarranted, the covariance matrix for $\widehat{\boldsymbol{\beta}}$ can be estimated consistently by the sandwich (also called the robust, linearization, Huber or empirical) covariance estimator:

$$\mathbf{V}_{\mathrm{SAND}}=\left(\frac{K}{K-1}\right)\left(\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{U}_i^{-1}\mathbf{D}_i\right)^{-1}\left\{\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{U}_i^{-1}\mathbf{s}_i\mathbf{s}_i'\mathbf{U}_i^{-1}\mathbf{D}_i\right\}\left(\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{U}_i^{-1}\mathbf{D}_i\right)^{-1} \tag{2}$$

with $\mathbf{s}_i=\mathbf{y}_i-\widehat{\boldsymbol{\mu}}_i$ and the $\mathbf{D}_i$ and $\mathbf{U}_i$ evaluated at $\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}$ for $i=1,\ldots,K$ [10]. Under mild regularity conditions, $\mathbf{V}_{\mathrm{SAND}}$ converges to the variance of the asymptotic distribution of $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ as K, the number of clusters, grows to infinity.

For a given p-vector, ℓ, $v_{\mathrm{SAND}}=\ell'\mathbf{V}_{\mathrm{SAND}}\ell$ is the sandwich estimator of $\mathrm{Var}(\ell'\widehat{\boldsymbol{\beta}})$. Tests of the null hypothesis H_0: $\ell'\boldsymbol{\beta}=c$ are usually conducted by comparing Wald-like t-statistics of the form $(\ell'\widehat{\boldsymbol{\beta}}-c)/\sqrt{v_{\mathrm{SAND}}}$ to the standard normal distribution or a t-distribution with $K-1$ degrees of freedom [1, 2].

Because Wald tests can be unstable for some measurement scales, for example when a coefficient approaches infinity in a model for binary data, Rotnitzky and Jewell [11] propose a score test based on the sandwich estimator. Let $\boldsymbol{\gamma}$ denote an r-vector subset of the parameter vector $\boldsymbol{\beta}$ and consider testing H_0: $\boldsymbol{\gamma}=\boldsymbol{\gamma}_0$. Let $\boldsymbol{\beta}=(\boldsymbol{\gamma}',\boldsymbol{\delta}')'$ and let $\widetilde{\boldsymbol{\delta}}(\gamma_0)$ equal the GEE estimator of $\boldsymbol{\delta}$ in the restricted parameter space under H_0. Let

$$M_{\boldsymbol{\gamma}}\{\boldsymbol{\gamma}_0,\widetilde{\boldsymbol{\delta}}(\gamma_0)\}=\sum_{i=1}^{K}\left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\gamma}}\right)'\mathbf{U}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)|_{\{\boldsymbol{\gamma}_0,\widetilde{\boldsymbol{\delta}}(\gamma_0)\}}$$

Finally, let $\widetilde{\mathbf{V}}_{\mathrm{SAND},\gamma}$ and $\widetilde{\mathbf{V}}_{\mathrm{GEE},\gamma}$ be the $r \times r$ principal submatrices corresponding to γ of the matrices $\mathbf{V}_{\mathrm{SAND}}$ and $\mathbf{V}_{\mathrm{GEE}}$ resulting from the estimates $\{\gamma_0, \widetilde{\boldsymbol{\delta}}(\gamma_0)\}$ and let $\widetilde{\Sigma}_\gamma = ((K-1)/K)\widetilde{\mathbf{V}}^{-1}_{\mathrm{GEE},\gamma}\widetilde{\mathbf{V}}_{\mathrm{SAND},\gamma}\widetilde{\mathbf{V}}^{-1}_{\mathrm{GEE},\gamma}$. The score test statistic is $T_{\mathrm{S}} = M_\gamma'\widetilde{\Sigma}_\gamma^{-1}M_\gamma$. Under H_0, T_{S} converges to a χ^2_r random variable as the number of clusters gets large [11].

3. BRL METHODOLOGY

3.1. Linear models

We [7] studied linear models like the model of Section 2 with the identity link and working covariance matrix $\mathbf{U}_i \propto \mathbf{I}$. We use $\mathbf{I}$ to denote identity matrices of conforming dimensions and we let $\mathbf{V}$ and $\mathbf{U}$ denote the block diagonal true and working covariance matrices for the entire sample with blocks $\mathbf{V}_i$ and $\mathbf{U}_i$, respectively. For this linear model, $\mathbf{D}_i = \mathbf{X}_i$, the design matrix of explanatory variables for cluster i, and the sandwich estimator given in (2) simplifies to

$$\mathbf{V}_{\mathrm{S}} = \frac{K}{K-1}(\mathbf{X}'\mathbf{X})^{-1}\left\{\sum_{i=1}^{K}\mathbf{X}_i'\mathbf{s}_i\mathbf{s}_i'\mathbf{X}_i\right\}(\mathbf{X}'\mathbf{X})^{-1} \tag{3}$$

where $\mathbf{X}$ denotes the design matrix for the entire sample. The sandwich tends to be biased because regardless of the true covariance matrix,

$$E(\mathbf{s}_i\mathbf{s}_i') = (\mathbf{I}-\mathbf{H}_X)_i\mathbf{V}(\mathbf{I}-\mathbf{H}_X)_i' = \mathbf{Q}_{0i} \tag{4}$$

where $(\mathbf{I}-\mathbf{H}_X)_i$ denotes the rows of $(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ corresponding to the ith cluster, and as discussed in Reference [7], the unknown $\mathbf{Q}_{0i}$ generally will not equal $((K-1)/K)\mathbf{V}_i$.

To correct for bias in the sandwich estimator, we suggested ignoring the factor of $K/(K-1)$ in (2) and replacing $\mathbf{s}_i$, the raw residuals for cluster i, with adjusted residuals $\mathbf{B}_i\mathbf{s}_i$ designed to have the same covariance as the unknown errors, $\mathbf{y}_i - \boldsymbol{\mu}_i$, when the unknown covariance $\mathbf{V}$ is proportional to the working covariance $\mathbf{U}$. When $\mathbf{V} \propto \mathbf{U}$, equation (4) yields that $E(\mathbf{s}_i\mathbf{s}_i') = (\mathbf{I}-\mathbf{H}_X)_i\mathbf{U}(\mathbf{I}-\mathbf{H}_X)_i'$ and the desired adjustment matrix $\mathbf{B}_i$ solves

$$\mathbf{B}_i(\mathbf{I}-\mathbf{H}_X)_i\mathbf{U}(\mathbf{I}-\mathbf{H}_X)_i'\mathbf{B}_i' = \mathbf{U}_i \tag{5}$$

That is, $\mathbf{B}_i$ solves, $\mathbf{B}_i\mathbf{Q}_i\mathbf{B}_i' = \mathbf{U}_i$ where $\mathbf{Q}_i$ equals $\mathbf{Q}_{0i}$ with the unknown covariance matrix $\mathbf{V}$ replaced with the working covariance matrix $\mathbf{U}$. The BRL estimator for the covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\mathbf{V}_{\mathrm{BRL}} = (\mathbf{X}'\mathbf{X})^{-1}\left\{\sum_{i=1}^{K}\mathbf{X}_i'\mathbf{B}_i\mathbf{s}_i\mathbf{s}_i'\mathbf{B}_i\mathbf{X}_i\right\}(\mathbf{X}'\mathbf{X})^{-1} \tag{6}$$

If the working covariance matrix is proportional to the true covariance matrix, then the BRL estimator is unbiased [7, 12]. Although there is no unique solution to $\mathbf{B}_i\mathbf{Q}_i\mathbf{B}_i' = \mathbf{U}_i$, there is a unique symmetric solution, which worked well in a simulation study [7]. McCaffrey *et al.* [12], extended BRL to linear generalized least squares where the working covariance matrix is not restricted to be the identity or even diagonal, but is assumed to be known.

If the data are normally distributed conditional on the explanatory variables, then $v_{\mathrm{BRL}} = \ell' V_{\mathrm{BRL}}\ell$ is distributed $\sum_{i=1}^{K}\lambda_i^* z_i$ where the z_i are i.i.d. χ^2_1 random variables and λ_i^* are the eigenvalues of the $K \times K$ matrix $\mathbf{G}_{\mathbf{V}}$ with elements $\mathbf{G}_{\mathbf{V}}[i,j] = \mathbf{g}_i'\mathbf{V}\mathbf{g}_j$, where

$\mathbf{g}_i = \ell'(\mathbf{X}'\mathbf{X})\mathbf{X}_i'\mathbf{B}_i(\mathbf{I} - \mathbf{H}_X)_i'$. Generally, the BRL estimator is relatively more variable than a χ^2 random variable with $K-1$ degrees of freedom. We used Satterthwaite's approximation [13], for the distribution of v_{BRL} based on the eigenvalues of the matrix $\mathbf{G}_{\mathbf{U}}$, where the working covariance matrix substitutes for the unknown $\mathbf{V}$.

3.2. BRL for GEE

BRL has two components. The first adjusts the residuals for each cluster so that the expected value of their outer-products will better approximate the true covariance matrices and uses these to modify the sandwich estimator. The second develops improved approximations for the distributions of Wald t-test statistics based on the alternative variance estimator.

3.2.1. The BRL modified variance estimator for GEE. We begin with formulas for the expected value of the outer-product of the GEE residuals. Unlike linear models where exact formulas exist, nonlinear models require approximations.

For $i = 1, \ldots, K$, let $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ denote the cluster means evaluated at $\boldsymbol{\beta}$, $\boldsymbol{\beta}^*$ denote the true value of the parameter and $\mathbf{e}_i = \mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}^*)$; then $\mathbf{s}_i = \mathbf{e}_i + \boldsymbol{\mu}_i(\boldsymbol{\beta}^*) - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\beta}})$ and a first-order expansion of $\boldsymbol{\mu}_i$ around $\boldsymbol{\beta}^*$ yields that $\mathbf{s}_i(\widehat{\boldsymbol{\beta}}) \doteq \mathbf{e}_i - \mathbf{D}_i^*(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, where $\mathbf{D}_i^*$ is evaluated at $\boldsymbol{\beta}^*$. It follows from Reference [1] that $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \doteq (\sum_{j=1}^K \mathbf{D}_j^{*\prime}\mathbf{U}_j^{*-1}\mathbf{D}_j^*)^{-1}\sum_{j=1}^K \mathbf{D}_j^{*\prime}\mathbf{U}_j^{*-1}\mathbf{e}_j = \mathbf{V}_{\mathrm{GEE}}^*\sum_{j=1}^K \mathbf{D}_j^{*\prime}\mathbf{U}_j^{*-1}\mathbf{e}_j$, where $\mathbf{U}_j^*$ and $\mathbf{V}_{\mathrm{GEE}}^*$ correspond to $\mathbf{U}_j$ and $\mathbf{V}_{\mathrm{GEE}}$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. Thus, $\mathbf{s}_i \doteq \mathbf{e}_i - \mathbf{D}_i^*\mathbf{V}_{\mathrm{GEE}}^*\sum_j \mathbf{D}_j^{*\prime}\mathbf{U}_j^{*-1}\mathbf{e}_j$.

Following Reference [6], and using the fact that $E(\mathbf{e}_j\mathbf{e}_l') = \mathbf{0}$ for $j \neq l$, this first-order expansion yields

$$
\begin{aligned}
\mathbf{Q}_i = E(\mathbf{s}_i\mathbf{s}_i') &\doteq E(\mathbf{e}_i\mathbf{e}_i') - \mathbf{D}_i^*\mathbf{V}_{\mathrm{GEE}}^*\sum_j \mathbf{D}_j^{*\prime}\mathbf{U}_j^{*-1}E(\mathbf{e}_j\mathbf{e}_i') - \sum_l E(\mathbf{e}_i\mathbf{e}_l')\mathbf{U}_l^{*-1}\mathbf{D}_l^*\mathbf{V}_{\mathrm{GEE}}^*\mathbf{D}_i^{*\prime} \\
&\quad + \mathbf{D}_i^*\mathbf{V}_{\mathrm{GEE}}^* E\left(\sum_j \mathbf{D}_j^{*\prime}\mathbf{U}_j^{*-1}\mathbf{e}_j\sum_l \mathbf{e}_l'\mathbf{U}_l^{*-1}\mathbf{D}_l^*\right)\mathbf{V}_{\mathrm{GEE}}^*\mathbf{D}_i^{*\prime} \\
&\doteq \mathbf{V}_i - \mathbf{D}_i^*\mathbf{V}_{\mathrm{GEE}}^*\mathbf{D}_i^{*\prime}\mathbf{U}_i^{*-1}\mathbf{V}_i - \mathbf{V}_i\mathbf{U}_i^{*-1}\mathbf{D}_i^*\mathbf{V}_{\mathrm{GEE}}^*\mathbf{D}_i^{*\prime} \\
&\quad + \mathbf{D}_i^*\mathbf{V}_{\mathrm{GEE}}^*\left\{\sum_j \mathbf{D}_j^{*\prime}\mathbf{U}_j^{*-1}\mathbf{V}_j\mathbf{U}_j^{*-1}\mathbf{D}_j^*\right\}\mathbf{V}_{\mathrm{GEE}}^*\mathbf{D}_i^{*\prime} \qquad (7)
\end{aligned}
$$

By replacing the unknown covariance matrices $\mathbf{V}_j$ with the working covariance matrices and $\boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\beta}}$, we estimate $E(\mathbf{s}_i\mathbf{s}_i')$ by $\mathbf{Q}_i = \mathbf{U}_i - \mathbf{D}_i\mathbf{V}_{\mathrm{GEE}}\mathbf{D}_i'$. Although we defined $\mathbf{U}_i$ to be scaled by ϕ, the derivation of the BRL estimator is invariant to the inclusion of this scaling factor and in practice it is ignored. The accuracy of the approximations used to derive $\mathbf{Q}_i$ depends on many factors; in particular, the approximations hold only as $\widehat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}^*$, which generally requires the number of clusters to be large. As such, the derivations above should be viewed as motivation, and not evidence for the accuracy of the BRL estimator used with GEE.

Having obtained an approximation to $\mathbf{Q}_i$, we solve $\mathbf{B}_i\mathbf{Q}_i\mathbf{B}_i = \mathbf{U}_i$ by $\mathbf{B}_i = \mathbf{U}_i^{1/2}\{\mathbf{U}_i^{1/2}\mathbf{Q}_i\mathbf{U}_i^{1/2}\}^{-1/2}\mathbf{U}_i^{1/2}$, where all matrix roots are symmetric, and obtain the BRL

variance estimator:

$$\mathbf{V}_{\mathrm{BRL}} = \mathbf{V}_{\mathrm{GEE}} \left\{ \sum_{i=1}^{K} \mathbf{D}_i' \mathbf{U}_i^{-1} \mathbf{B}_i \mathbf{s}_i \mathbf{s}_i' \mathbf{B}_i \mathbf{U}_i^{-1} \mathbf{D}_i \right\} \mathbf{V}_{\mathrm{GEE}} \tag{8}$$

As with the linear model we use the unique symmetric solution to $\mathbf{B}_i \mathbf{Q}_i \mathbf{B}_i = \mathbf{U}_i$ to ensure the estimator is well-defined.

Unlike linear models, the expected value of the standard sandwich estimator for GEE is not determined solely by the expected values of the outer-products of the residuals because the other terms in the estimator (i.e. $\mathbf{D}_i$ and $\mathbf{U}_i$) involve the estimated regression coefficients. Although our proposed estimator does not account for this fact, we anticipate that the expected value of the outer-products of the residuals often dominates other sources of bias. However, there is no guarantee that the estimator reduces bias as there is in linear models. We retain the name BRL because of the analogy to our previous work.

3.2.2. Hypothesis testing using BRL for GEE. The final step to obtaining BRL inference is to approximate the distribution of the test statistic for testing the null hypothesis H_0: $\ell'\boldsymbol{\beta} = c$. To simplify presentation we set $c = 0$ for the remainder of this section but the methods generalize to other values of c.

We write $\mathbf{s}_i \doteq \mathbf{e}_i - \mathbf{D}_i \mathbf{V}_{\mathrm{GEE}} \sum_{j=1}^{K} \mathbf{D}_j' \mathbf{U}_j^{-1} \mathbf{e}_j$ and $v_{\mathrm{BRL}} \doteq \mathbf{e}'(\sum_i \mathbf{g}_i \mathbf{g}_i')\mathbf{e}$ where $\mathbf{g}_i' = \ell' \mathbf{V}_{\mathrm{GEE}} \mathbf{D}_i \mathbf{U}_i^{-1} \mathbf{B}_i (\mathbf{I} - \mathbf{D}\mathbf{V}_{\mathrm{GEE}} \mathbf{D}'\mathbf{U}^{-1})_i$, $\mathbf{D} = [\mathbf{D}_1' | \mathbf{D}_2' | \cdots | \mathbf{D}_K']'$ and $(\mathbf{I} - \mathbf{D}\mathbf{V}_{\mathrm{GEE}} \mathbf{D}'\mathbf{U}^{-1})_i$ denotes the rows of $(\mathbf{I} - \mathbf{D}\mathbf{V}_{\mathrm{GEE}} \mathbf{D}'\mathbf{U}^{-1})$ corresponding to the units from cluster i. Consequently, we can approximate the distribution of v_{BRL} as $\sum_i \lambda_i z_i$ where z_i are i.i.d. χ_1^2 random variables, and the λ's are the eigenvalues of the matrix $\mathbf{G}_{\mathbf{U}}[ij] = \mathbf{g}_i' \mathbf{U} \mathbf{g}_j$. Satterthwaite [13] provides an approximation to the distribution of $a v_{\mathrm{BRL}}$ by a χ_f^2 where $a = \sum \lambda_i / \sum \lambda_i^2$ and $f = a \sum \lambda_i$. Because $\sum \lambda_i \doteq v = \mathrm{Var}(\ell'\widehat{\boldsymbol{\beta}})$ and $\ell'\widehat{\boldsymbol{\beta}}/\sqrt{v}$ is approximately distributed standard normal under $\mathbf{H}_0 : \ell'\boldsymbol{\beta} = 0$, we can approximate the distribution of $\ell'\widehat{\boldsymbol{\beta}}/\sqrt{v_{\mathrm{BRL}}}$ by a t-distribution with f degrees of freedom.

However, reference t-distributions based on the Satterthwaite approximation tend to overestimate tail probabilities and yield conservative tests when the approximate degrees of freedom are small [7], because the distribution of v_{BRL} tends to have less mass in the lower tail than the Satterthwaite approximation. Hence, we also consider the following saddlepoint approximations for calculating p-values and critical values.

Let t_{obs} equal the observed test statistic, so that under $\mathbf{H}_0$

$$\Pr\left(\left| \frac{\ell'\widehat{\boldsymbol{\beta}}}{\sqrt{v_{\mathrm{BRL}}}} \right| > |t_{\mathrm{obs}}| \right) = \Pr\left(\left(\frac{\ell'\widehat{\boldsymbol{\beta}}}{\sqrt{v}} \right)^2 - t_{\mathrm{obs}}^2 v_{\mathrm{BRL}}/v > 0 \right) = \Pr\left(\sum_{i=0}^{K} \gamma_i z_i > 0 \right)$$

where $z_0 = (\ell'\widehat{\boldsymbol{\beta}}/\sqrt{v})^2$, $\gamma_0 = 1$ and $\gamma_i = -t_{\mathrm{obs}}^2 \lambda_i / \sum \lambda_j$ for $i \geqslant 1$. Because $z_0 \dot{\sim} \chi_1^2$ and $v_{\mathrm{BRL}} \dot{\sim} \sum_i \lambda_i z_i$, we treat the random variable $z = \sum_{i=0}^{K} \gamma_i z_i$ as a weighted sum of independent χ_1^2 random variables and use a saddlepoint approximation of its CDF evaluated at zero to calculate a p-value.

The saddlepoint approximation to the CDF of z, F_z, depends on the cumulant generating function of z, $C(s) = -(1/2)\sum_{i=0}^{K} \log(1 - 2\gamma_i s)$, and its first and second derivatives $\dot{C}(s)$ and $\ddot{C}(s)$. The saddlepoint s_{sp} solves $\dot{C}(s_{\mathrm{sp}}) = \sum_{i=0}^{K} \gamma_i / (1 - 2\gamma_i s_{\mathrm{sp}}) = 0$. The saddlepoint

approximation [14, 15] to F_z, is

$$\widehat{F}_z(z) = \Phi(r) + \phi(r)\left\{\frac{1}{r} - \frac{1}{q}\right\} \tag{9}$$

where Φ and ϕ denote the standard normal CDF and probability density function, $r = \text{sign}(s_{\text{sp}})[2\{s_{\text{sp}}z + C(s_{\text{sp}})\}]^{1/2}$ and $q = s_{\text{sp}}\ddot{C}(s_{\text{sp}})^{1/2}$, where $\text{sign}(s_{\text{sp}}) = 1$ if $s_{\text{sp}} \geqslant 0$ and -1 otherwise. The saddlepoint is found numerically. Butler and Paolella [16] use a similar approach to estimate the CDF of ratios of sums of squares.

The p-values for values of t_{obs} follow from using equation (9) to approximate $\widehat{F_z(0)}$. Critical values t_α satisfying $\Pr(|\ell'\widehat{\boldsymbol{\beta}}/\sqrt{v_{\text{BRL}}}| > t_\alpha) = \alpha$ are found by iteratively solving for t_α such that $\widehat{F}_{z(t_\alpha)}(0) = \alpha$, where $z(t)$ denotes the random variable that results from replacing t_{obs} by t in the definition of z.

As yet another alternative, the BRL estimator can be used in the score test of Rotnitzky and Jewell as a means of testing the null hypothesis. To obtain the test statistic, $\widetilde{\mathbf{V}}_{\text{SAND},\gamma}$ in $\widetilde{\Sigma}$ is replaced by $\widetilde{\mathbf{V}}_{\text{BRL},\gamma}$.

4. ALTERNATIVES TO BRL

Mancl and DeRouen [6] take a similar approach to deriving an alternative sandwich estimator. They also try adjusting the residuals so that the second moment of the adjusted residuals approximately equals $\mathbf{V}_i$, using the approximation:

$$E(\mathbf{s}_i\mathbf{s}_i') = (\mathbf{I} - \mathbf{D}_i\mathbf{V}_{\text{GEE}}\mathbf{D}_i'\mathbf{U}_i^{-1})\mathbf{V}(\mathbf{I} - \mathbf{D}_i\mathbf{V}_{\text{GEE}}\mathbf{D}_i'\mathbf{U}_i^{-1})'$$

Their resulting variance estimator, which they call the bias-corrected estimator, is

$$\mathbf{V}_{\text{MD}} \doteq \mathbf{V}_{\text{GEE}}\left\{\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{U}_i^{-1}\mathbf{B}_{\text{MD}i}\mathbf{s}_i\mathbf{s}_i'\mathbf{B}_{\text{MD}i}\mathbf{U}_i^{-1}\mathbf{D}_i\right\}\mathbf{V}_{\text{GEE}} \tag{10}$$

where $\mathbf{B}_{\text{MD}i} = (\mathbf{I} - \mathbf{D}_i\mathbf{V}_{\text{GEE}}\mathbf{D}_i'\mathbf{U}_i^{-1})^{-1}$.

The Mancl and DeRouen approximation ignores cross-product terms in the first-order approximation to the expected value of the outer-product of residuals, thereby avoiding the need to calculate matrix roots. For linear models, this yields an estimator that is equivalent to a form of jackknife estimator, which tends to be biased too large [7]. BRL accounts for these cross-products terms. Mancl and DeRouen use Wald test statistics based on their alternative estimator and approximate the distribution of the statistic under the null with a standard normal distribution; however, Satterthwaite and saddlepoint approximations follow directly from the methods for BRL using the corresponding $\mathbf{G}_{\mathbf{U}}$ matrix.

Kott [17] proposed an alternative to the sandwich estimator for linear models using the ratio of $\text{Var}(\ell'\widehat{\boldsymbol{\beta}})$ to the expected value of v_{SAND}, where both quantities are calculated under the assumption of independent errors. The standard sandwich estimator is multiplied by this ratio to produce Kott's variance estimator. The method extends to GEE using the approximations for the expected values of the outer-products of residuals described above. To our knowledge this method has not been tested empirically in this context. Unlike BRL, Kott's method does not require calculating matrix roots for each cluster. Both Satterthwaite and saddlepoint

approximations for Wald tests follow directly from the methods we developed for BRL. Because the Kott variance estimator is proportional to the traditional sandwich, both approximations are the same for Kott as for the traditional sandwich method.

5. SIMULATION METHODS

We conducted a large-scale simulation study of the properties of the BRL and the alternative variance estimators and approximations for references distributions of test statistics. We modelled correlated dichotomous outcomes using a balanced two-stage cluster sample with $K = 20$ clusters and a constant $n = 51$ observations in each cluster. All logistic regression coefficients were estimated by maximum likelihood, equivalent to GEE with $\mathbf{R}_{Wi} = \mathbf{I}$ for all i.

5.1. Experimental design

All simulation replications use a common design matrix $\mathbf{X}$ with an intercept and four explanatory variables chosen to represent, in terms of their impact on nonparametric variance estimation, the principal types of explanatory variables that might be encountered in practice. Results from this design matrix should therefore generalize to many other situations. The first two explanatory variables, x_1 and x_2, are dichotomous and constant within cluster. The variable x_1 is 0.50 in the odd numbered clusters and -0.50 in the rest; x_2 is 0.85 in just three clusters: 9, 10, and 11 and -0.15 in the rest. Both x_3 and x_4 were generated from standard normal distributions: x_3 was generated from a multivariate normal with ICC of 0.5 within cluster; x_4 was generated from independent normal distributions. Observed ICC are $1.00, 1.00, 0.55$ and 0.005, respectively. Observed correlations among the explanatory variables are all very small with the exception of $\mathrm{Corr}(x_1, x_2) = 0.14$, $\mathrm{Corr}(x_1, x_3) = 0.20$ and $\mathrm{Corr}(x_2, x_3) = 0.32$.

Because the bias and skew of estimated logistic regression coefficients vary with the true values of the parameters, we selected a range of values for each coefficient. The values are:

- Intercept, β_0: the log odds of $0.1, 0.2, 0.3$ and 0.5.
- β_1: $0, 0.3, 0.6$, and 0.9.
- β_2: $-0.9, -0.6, -0.3, 0, 0.3, 0.6$, and 0.9.
- β_3: $0, 0.15, 0.30$, and 0.45.
- β_4: $0, 0.15, 0.30$, and 0.45.

Variation in β_0 affects the proportion of $y_{ij} = 1$, and consequently the precision of all the estimated coefficients. We consider only positive values for β_1, β_3 and β_4. Unlike x_2, the explanatory variables x_1, x_3 and x_4 are roughly symmetric about zero, so that bias of the variance estimators should be approximately invariant to the sign of the coefficients. We chose a smaller range for β_3 and β_4 because those variables had greater variance than x_1 and x_2.

Using the methods of Oman and Zucker [18], we generated correlated binary variables y_{ij} with mean $\mu = 1/(1 + e^{-\eta})$, for $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$. Outcomes from the same cluster had ICC of $\rho = 0, 0.02, 0.06$ or 0.12. For clusters of size 51, these values correspond to design effects for the mean of 1, 2, 4 and 7. The small values (0.02 and 0.06) are consistent with values we find in health outcomes studies of patients and of school-based interventions

on health behaviours such as drug or violence prevention [19, 20]. The value of $\rho = 0.12$ tests the methods under fairly challenging conditions for clusters of this size.

We varied six design parameters (β_0 to β_4 and ρ), five with four levels and one with seven. We selected a fraction of the possible design points. We first restricted β_2 to positive values and zero yielding 4^6 possible design points from which we selected 4^5 design points for the simulation experiment. We then restricted β_2 to negative values and zero and again choose a 1/4 fraction of the 4^6 design points. The resulting simulation experiment had 2048 design points with each level of every factor appearing in 512 points except β_2, for which values other than zero occurred in 256 points. The design allows for estimates of all main effect and two-way interactions without confounding. At each design point, we generated 200 independent samples of 1020 correlated binary outcome variables. All computations were conducted in SAS.

5.2. *Procedures evaluated and performance measures*

We evaluated bias of the alternative variance estimators and type I error rates for proposed test procedures. For each replication of the simulation, we obtained the traditional sandwich (SAND), BRL, Mancl and DeRouen bias-corrected (MD) and extended Kott (KOTT) estimators of the variance for each coefficient. For each estimator, we used the eigenvalues of the corresponding $\mathbf{G}_U$ matrix to calculate the Satterthwaite and saddlepoint approximations for use with Wald t-statistic for testing the null hypotheses $H_0 : \beta_k = c_k^*$, $k = 0, \ldots, 4$, where c_k^* is the value of β_k used to generate the data. All tests were two-sided nominal 0.05-level tests. In addition, we sometimes use the standard normal and t-distribution with $K - 1$ degrees of freedom to approximate the reference distribution. Finally, we also calculated score statistics for each estimator and all coefficients other than β_0 and tested the hypothesis by comparing the resulting test statistic to a χ_1^2 distribution.

For each of the 2048 design points we estimate per cent relative bias, $\mathrm{PRB} = 100 \times (\bar{\hat{v}}/\widehat{\mathrm{Var}(\hat{\beta})} - 1)$, where for a given variance estimator $\hat{v}$ and coefficient β, $\bar{\hat{v}}$ denotes the average across the 200 Monte Carlo replicates and $\widehat{\mathrm{Var}(\hat{\beta})}$ equals the sample variance of the estimated coefficient across the replicates. We use the jackknife method [21] to correct PRB for ratio bias that results from the use of a relatively small sample to estimate the denominator of the ratio. Let $r = \bar{\hat{v}}/\widehat{\mathrm{Var}(\hat{\beta})}$ and $\tilde{r}$ equal the average of the 200 estimated ratios from the jackknife pseudo-replicates created by leaving out single simulation replicates. The bias-corrected ratio, which equals $200 \times (r - 199\tilde{r}/200)$, typically decreased the ratio by about 1–2 per cent.

For evaluating test procedures, we consider the overall type I error rate and its positive and negative components. The positive (negative) error rate equals the proportion of times we incorrectly reject the null hypothesis when the estimated coefficient is greater (less) than the true value.

6. SIMULATION RESULTS

6.1. *'Bias' in variance estimators*

Figure 1 summarizes the findings on PRB. Each panel in the matrix of plots presents the PRB for the sandwich estimator (denoted by 'S'), the MD estimator ('M'), our BRL estimator

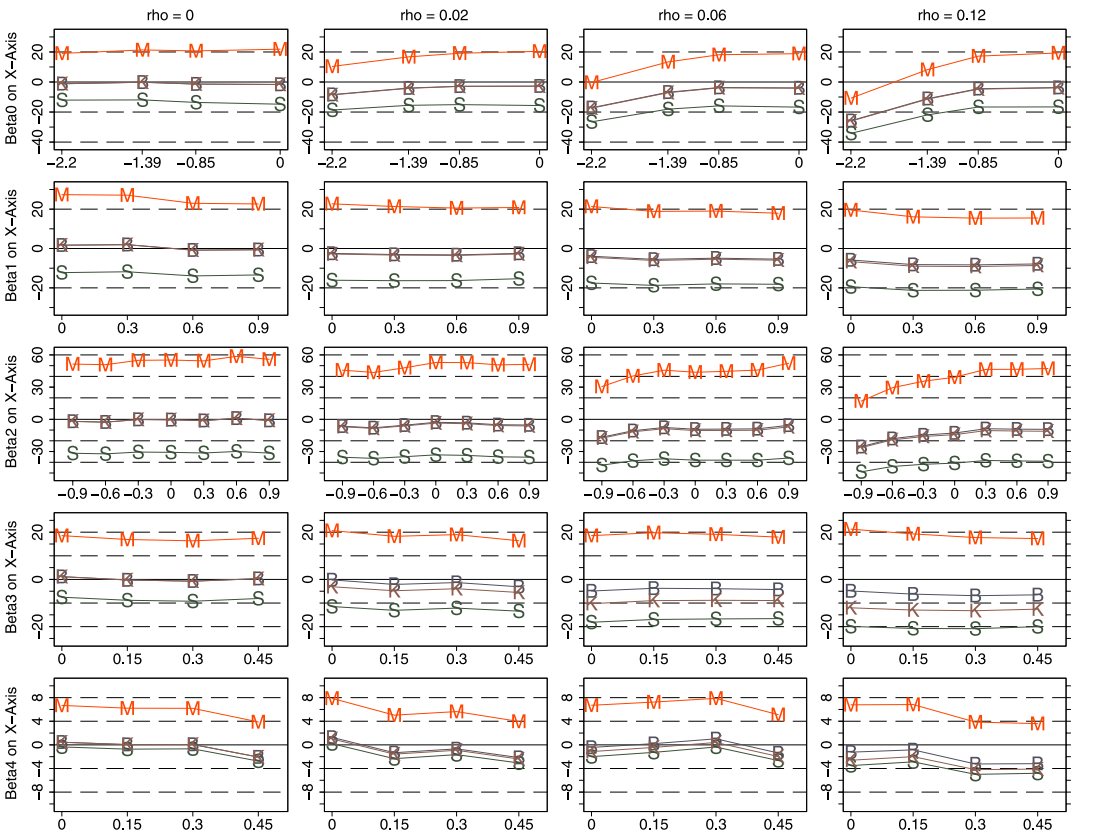

Figure 1. PRB for various estimators: traditional sandwich (S); MD (M); BRL (B); and Kott (K) by values of the ICC (rho) and coefficients.

('B') and the modified Kott estimator ('K') for a given coefficient and a given value of ρ. Within each panel, PRB values for each estimator are plotted against values of the coefficient, averaged across all the design points with the given value of ρ and the coefficient. For β_0, β_1, β_3 and β_4 there are 128 design points for each combination of ρ and the coefficient values. For $\beta_2 \neq 0$ there are 64 design points for each value of the coefficient and each value of ρ, while there are 128 design points for $\beta_2 = 0$ at each value of ρ.

The traditional sandwich estimator has appreciable negative bias for all values of ρ, including zero. Biases vary greatly across the explanatory variables, but are relatively invariant to values of ρ and the true coefficient. The bias is most pronounced, nearly 40 per cent too low, for the estimated variance of β_2. This occurs because $\widehat{\beta}_2$ is determined to a great extent by the outcomes in just three of the 20 clusters, where x_2 is positive. Even for the well-balanced cluster-level variable, x_1, the estimator ranges from about 13 to 20 per cent smaller than the empirical variability as ρ ranges from 0 to 0.12. However, the bias is relatively small for β_4—no more than 5 per cent too low.

The MD estimator is consistently too large, with the bias sometimes reaching nearly 60 per cent for β_2 and typically near 20 per cent for x_1 and x_3. For each variable except the intercept, the bias is relatively invariant to the values of the coefficient and the values of ρ.

The BRL estimator has no appreciable bias when $\rho = 0$ for any coefficients at any value, but it is consistently too small when $\rho > 0$, with the absolute value of the bias generally increasing with ρ and dependent on the values of the coefficients. For β_0 and β_1, BRL has the most bias for the values of the coefficient farthest from zero. For $\beta_2 = -0.90$, the relative bias is about 17 per cent when $\rho = 0.06$ and it grows to just over 25 per cent when $\rho = 0.12$.

Our extension of Kott's estimator performs very much like the BRL estimator for the coefficients of the three cluster-level variables (β_0, β_1 and β_2) and the relative bias is indistinguishable from that of BRL in the figure. For variables that vary within cluster, Kott's estimator performs worse than BRL when $\rho > 0$. For example, for β_3 the relative bias of Kott is about twice as large as it is for BRL when $\rho > 0$.

6.2. *Type I errors*

6.2.1. Wald tests using the traditional sandwich estimator when $\rho = 0$. Figure 2 presents type I error rates, when $\rho = 0$, for Wald tests based on the standard sandwich estimator using four alternatives for approximating the distribution of the test statistic under the null hypothesis: normal; t with $K - 1$ degrees of freedom; Satterthwaite; and saddlepoint. The figure contains one row of plots for each coefficient and one column for each approximation. Each panel of the figure plots the overall type I error rate (diamonds), the positive error rate (triangles) and the negative error rate (circles) against values of the coefficient.

Because of the bias in the sandwich estimator, almost all tests exceed the nominal 0.05 error rate. However, the sizes of the excesses depend heavily on the reference distribution. Overall type I error rates for tests using the normal approximation, which underestimates the variability of the test statistic, exceed 0.08 for β_0 and β_1, 0.07 for β_3 and 0.15 for β_2. Except for β_2, the other three approximations provide notable improvements over the normal, with type I error rates typically less than 0.07 and very close to nominal levels for β_4.

For β_2, the distribution of the sandwich variance estimator is approximated poorly by a χ^2_{K-1} distribution. As a result, the t-distribution with $K - 1$ degrees of freedom underestimates the variability of the test statistic and the overall type I error rates still exceed 0.10. The Satterthwaite t approximation yields tests with overall type I error rates that are much closer to the nominal rate. This occurs even though the Satterthwaite approximation overstates the mass in the lower tail of the distribution of the variance estimator. If the estimator were unbiased, the Satterthwaite degrees of freedom would yield an approximation to the test statistic with tails that were too heavy. Instead, this overcorrection of the reference distribution counterbalances the bias in the sandwich estimator of variance to produce the best type I error rates for the sandwich. The saddlepoint approximation, which relies on the variance estimator being unbiased, gives type I error rates that substantially exceed the nominal levels.

In general, type I errors are split nearly equally between the positive and negative errors. The exceptions occur for the smallest values of β_0, where the negative type I error rate exceeds the positive rate. That is, tests for this coefficient are more likely to reject the null when the estimate is less than the true value than when it is greater than the true value. This imbalance in the type I errors is due to asymmetry in the tails of $(\widehat{\beta}_0 - \beta_0)$. For example,

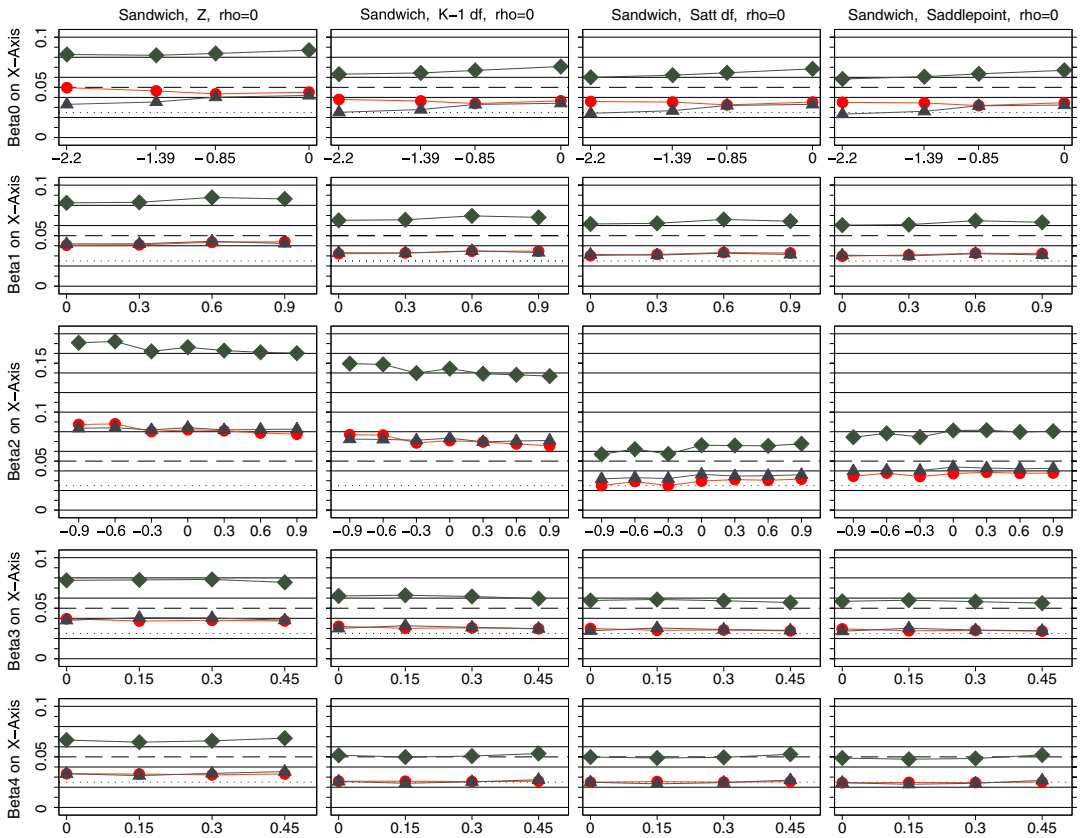

Figure 2. Type I error rates for Wald tests based on the traditional sandwich estimator when $\rho = 0$. Columns of plots correspond to alternative approximations for the distribution under the null: standard normal, t with $K - 1$ degrees of freedom, Satterthwaite approximation, and saddlepoint. Overall type I error rates (diamonds), negative errors (circles), and positive error (triangles) are plotted by the values of the coefficients. Vertical axis ranges from 0.00 to 0.10 for all panels. The dashed line is at 0.05 and the dotted line is at 0.025.

when $\beta_0 = -1.39$, about two-thirds of the top 5 per cent of absolute deviations, $|\widehat{\beta}_0 - \beta_0|$, occur for $\widehat{\beta}_0 < \beta_0$.

6.2.2. Wald tests using alternative variance estimators when $\rho = 0$. Figure 3 presents the type I error rates, again with $\rho = 0$, for Wald tests based on MD with the normal reference distribution, BRL and Kott estimators with the saddlepoint approximation, and for comparison, the sandwich with the Satterthwaite approximation. For each variance estimator, we present the reference distribution that tended to give type I error rates closest to the nominal value.

Wald tests using the MD estimator provide near-nominal type I error rates for all coefficients other than β_2, even though the estimator generally over estimates the variability in the estimated coefficient. For these coefficients the type I error rate is relatively invariant to the

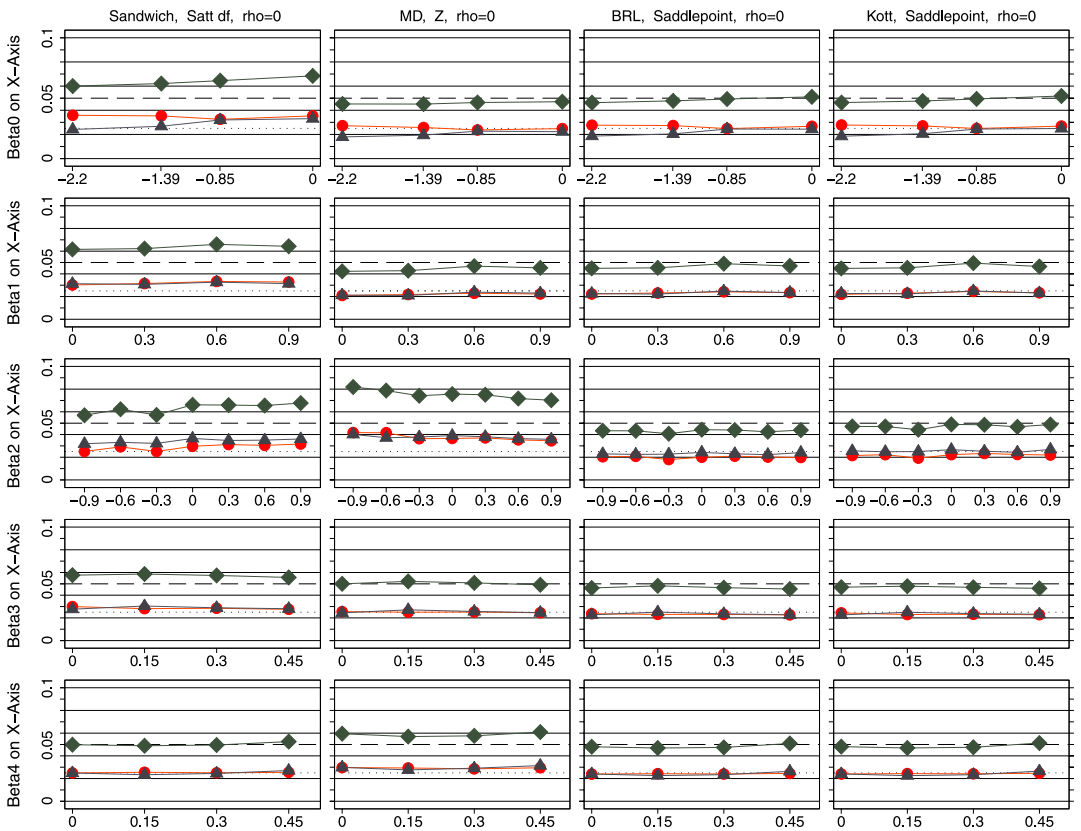

Figure 3. Type I error rates when $\rho = 0$ for Wald tests based on alternative variance estimators and distributions for the test statistics under the null: the traditional sandwich and t with $K - 1$ degrees of freedom; MD and the standard normal; BRL and the saddlepoint approximation; and Kott and the saddlepoint approximation. Overall type I error rates (diamonds), negative errors (circles), and positive error (triangles) are plotted by the values of the coefficients. The dashed line is at 0.05 and the dotted line is at 0.025.

value of the coefficient. For β_2, overall type I errors exceed 0.07 or 0.08 and are largest for the most negative values of the coefficient (-0.90 and -0.60).

Wald tests based on BRL variance estimators have near-nominal type I error rates for coefficients other than β_2. For β_2 the test is a little bit conservative with type I error rates of 0.043 on average across the values of the coefficient. However, the test does considerably better than tests based on the Satterthwaite approximation for β_2 (not shown in the figure). The Satterthwaite degrees of freedom overstate the tails of the test distribution, resulting in conservative tests with large critical values and type I error rates of about 0.031.

Wald tests based on the Kott estimator perform very similarly to the tests based on BRL. With the saddlepoint approximation the overall type I error rates are very close to nominal when $\rho = 0$ for all values of the various coefficients. For β_2, Kott-based tests slightly outperform the test using BRL with an average type I error rate of 0.047 rather than 0.043.

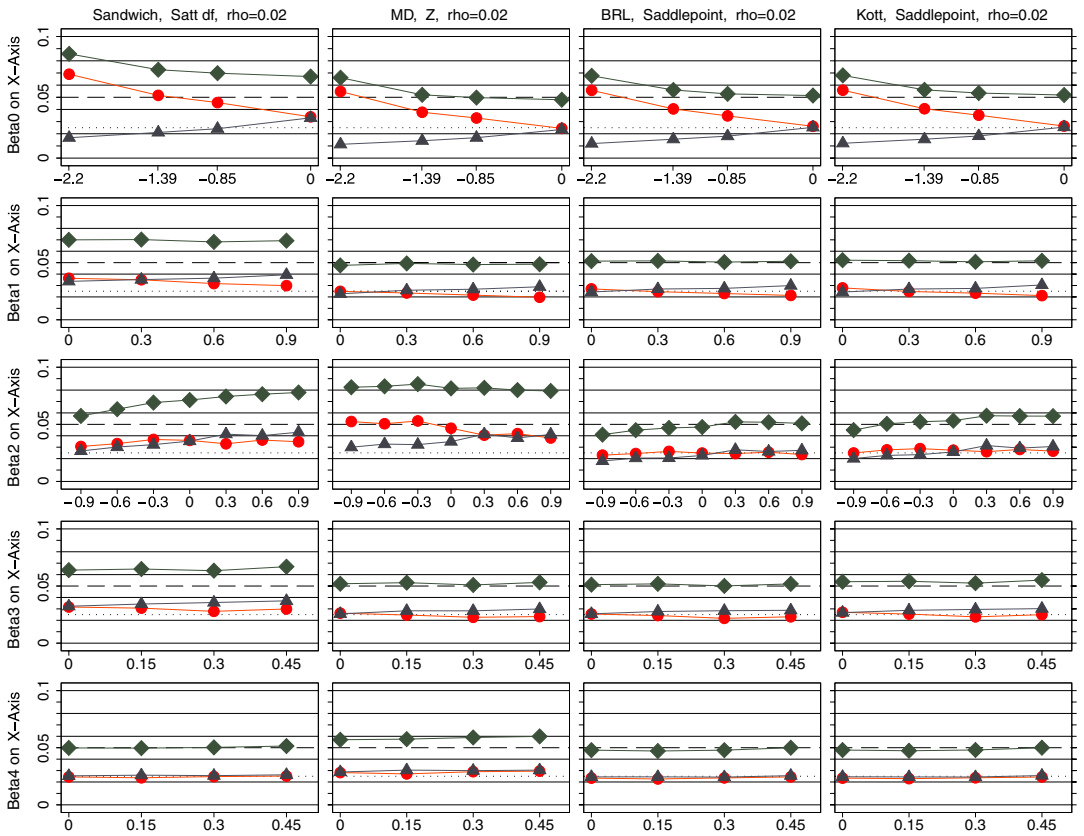

Figure 4. Type I error rates when $\rho = 0.02$ for Wald tests based on alternative variance estimators and distributions for the test statistics under the null: the traditional sandwich and t with $K - 1$ degrees of freedom; MD and the standard normal; BRL and the saddlepoint approximation; and Kott and the saddlepoint approximation. Overall type I error rates (diamonds), negative errors (circles), and positive error (triangles) are plotted by the values of the coefficients. The dashed line is at 0.05 and the dotted line is at 0.025.

All the alternatives outperform even the best sandwich-based Wald test. The type I error rates for the sandwich estimator always exceed the nominal value and those of the other estimators with the exception of the MD-based test for β_2 and β_4. Performance of all the tests is generally invariant to the values of the coefficients, and, except for β_0, positive and negative errors are equally likely.

6.2.3. Wald tests when $\rho > 0$. Figure 4, replicates Figure 3 with $\rho = 0.02$. The relative performance among the estimators is very similar to their performance when $\rho = 0$. The Wald tests based on BRL and Kott with the saddlepoint approximation have near-nominal type I error rates except for $\beta_0 = -2.2$. The MD-based tests are again inferior to BRL and Kott, with type I error rates for β_2 and β_4 exceeding the nominal values and averaging a little over 0.08 and nearly 0.06, respectively, for these two coefficients. The sandwich-based tests again generally perform worst among all the methods.

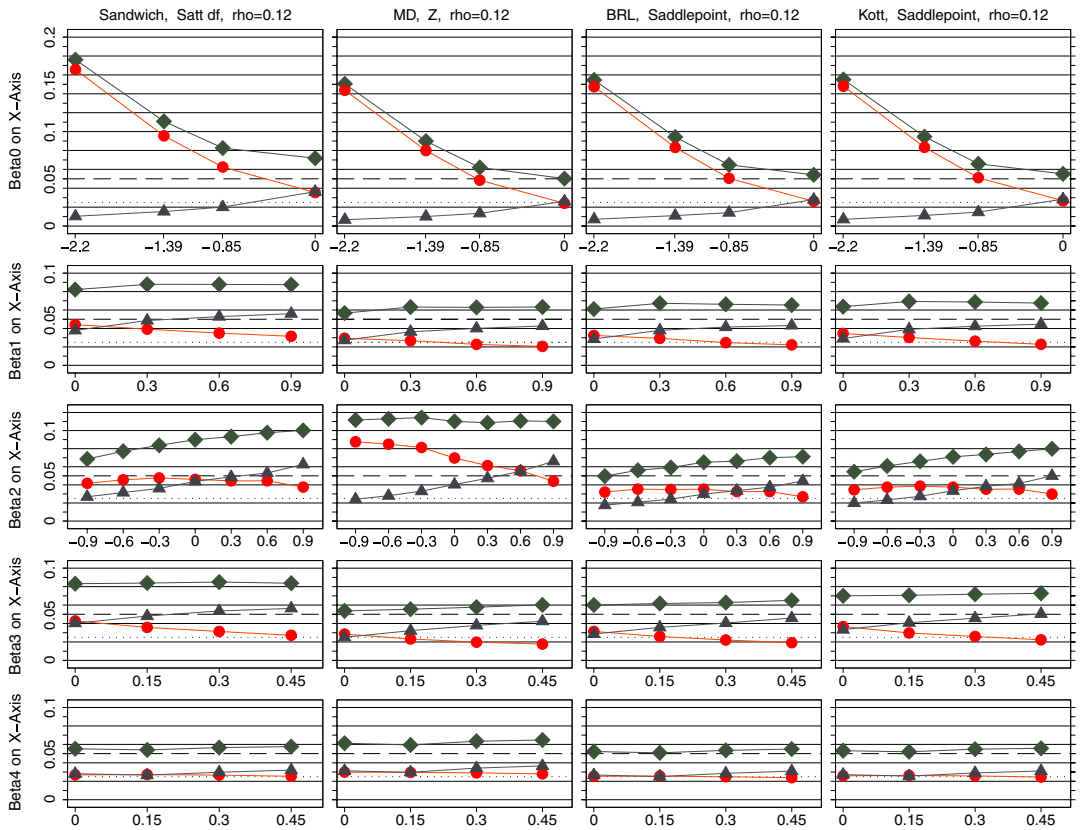

Figure 5. Type I error rates when $\rho = 0.12$ for Wald tests based on alternative variance estimators and distributions for the test statistics under the null: the traditional sandwich and t with $K - 1$ degrees of freedom; MD and the standard normal; BRL and the saddlepoint approximation; and Kott and the saddlepoint approximation. Overall type I error rates (diamonds), negative errors (circles), and positive error (triangles) are plotted by the values of the coefficients. The dashed line is at 0.05 and the dotted line is at 0.025.

Except for β_4, the type I errors for all tests are generally higher when $\rho = 0.02$ than when $\rho = 0$. All the tests have type I errors that greatly exceed the nominal value for very low values of β_0. The skew in the errors in β_0 is exacerbated by the correlation in the outcomes and by near- or quasi-complete separations, which are rare but frequent enough to affect the negative type I error rates. Also, asymmetry of negative and positive errors occurs for some other coefficients. For example, positive errors are more frequent for all the tests for the highest values of β_1, but negative errors are more frequent for the MD-based test at low values of β_2.

Roughly speaking, the performance of the tests smoothly degrades as ρ increases from 0.02 to 0.06 (not shown) and then to 0.12. As shown in Figure 5, when $\rho = 0.12$ almost all the tests have type I error rates that substantially exceed the nominal values. Tests for β_4 are generally an exception with error rates of about 0.055, 0.062, 0.053 and 0.054 for the

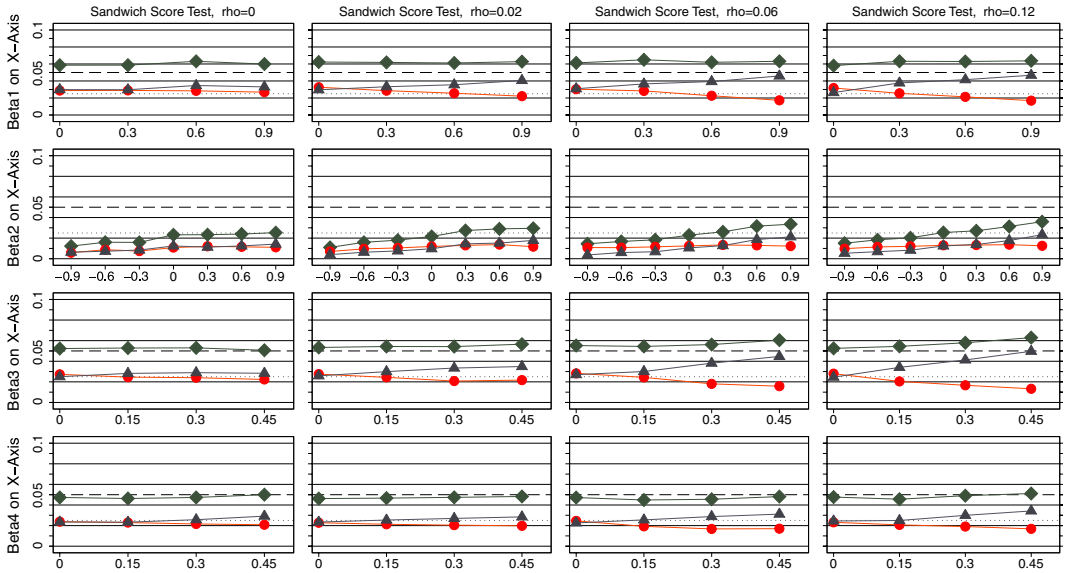

Figure 6. Type I error rates for score tests based on traditional sandwich estimators and χ^2 approximation for $\rho = 0.00$, 0.02, 0.06 and 0.12. Overall type I error rates (diamonds), negative errors (circles), and positive errors (triangles) are plotted by the values of the coefficients. Vertical axis ranges from 0.00 to 0.10 for all panels. The dashed line is at 0.05 and the dotted line is at 0.025.

sandwich-, MD-, BRL- and Kott-based tests. Asymmetry in errors continues to increase with the value of ρ. For all coefficients (except β_4) and all tests the rates of negative and positive type I errors are unequal. The asymmetry continues to be most pronounced for the lowest values of β_0 but it is now substantial for the largest values of β_1 and β_3 and the extreme values of β_2.

Except for β_1, where MD has slightly smaller type I error rates, BRL and Kott have the lowest error rates among all the alternatives when $\rho = 0.12$, with BRL often doing somewhat better. In particular for β_3, the type I error rates for BRL average about 0.06 and the error rates for the Kott-based tests are about 14 per cent larger. The superiority of BRL- and Kott-based tests holds up when ρ is substantially greater than 0, even though the bias adjustments used by BRL and Kott, in this simulation, rely on the working independence assumption.

6.2.4. Score tests. Figure 6 presents results for the score test based on the traditional sandwich estimator (without the $K/(K-1)$ factor). For the coefficients other than β_2 the score test generally out performs Wald tests based on the traditional sandwich and is very competitive with the alternatives shown in Figures 3–5, especially for $\rho > 0$. For β_1 the overall type I errors are typically less than 0.06 and always less than 0.07. For β_3 error rates are between 0.05 and 0.06 and for β_4 the type I errors are between 0.045 and 0.05. However, for β_2 the type I error rates are too low, often less than 0.03. Type I error rates for the score test are nearly invariant to the value of ρ, except for β_3, where performance degrades somewhat as ρ increases. However, asymmetry in the positive and negative type I error rates increases with ρ as the distribution of estimated coefficients becomes more skewed.

We do not present results for score tests based on MD, BRL or Kott because these test statistics tend to asymptote as the estimated coefficient deviates from the true value. As such, these tests rarely reject the null hypothesis and the type I error rates are substantially less than the nominal 0.05 level. For BRL and Kott, the largest type I error rates are just above 0.03, while for MD the rates are even lower, often less than 0.02 and equal to zero for β_2.

7. DISCUSSION

The simulation results confirm for logistic regression our previous major findings for linear models [7]. Both the sandwich and the jackknife-like MD estimators can exhibit very large biases for all values of ρ, especially for coefficients of cluster-level explanatory variables. Wald tests using the sandwich estimator and a t with $K-1$ degrees of freedom as the reference distribution, produce an excess of type I errors. In contrast, Wald tests using the MD estimator and a normal distribution often, but not always, produce near-nominal type I error rates. Both the BRL and Kott methods perform very well for $\rho=0$, with nearly unbiased estimation and nominal type I error rates for Wald tests using a saddlepoint approximation. Although both methods tend to underestimate the variability of coefficients as ρ increases above 0, the biases are never as large (in absolute value) as they are for the sandwich or MD, and the BRL and Kott tests outperform the commonly used sandwich-based Wald tests.

The Kott estimator performs nearly as well as BRL for cluster-level variables. However, for variables that vary within cluster, Kott has larger negative bias than BRL when $\rho>0$. The Kott estimator applies an overall adjustment to the sandwich estimator for a coefficient. In contrast, BRL applies distinct adjustments to the residuals within each cluster, which may explain its superior robustness against misspecification of the working covariance matrix.

For the traditional sandwich estimator, the score test had smaller type I error rates than Wald tests regardless of the reference distribution used with the Wald test. Although superior to Wald tests based on the sandwich estimator, these score tests generally did not perform as well as the preferred Wald tests for BRL, Kott, and MD. However, because the score test is nearly invariant to ρ in the range we considered, it typically performed as well as or better than Wald tests based on any of the estimators and any reference distribution when $\rho=0.12$.

When the true coefficients are far from zero and $\rho\geqslant 0.06$, type I errors for all tests, including BRL, are more common than the nominal levels and asymmetric about zero with more errors in the same direction as the sign of the true coefficient. This is due primarily to the distribution of the estimated coefficients, so that asymmetric error rates might be an inherent characteristic of Wald tests for logistic regression. Using the test statistic as a pivotal value for confidence intervals might yield incorrect inference if the coefficient is truly far from zero. Alternatives such as the BC_a method for correcting bootstrap confidence intervals [22] might be preferable, although bootstrapping with clustered data poses challenges of its own.

Our simulation study does not address the performance of BRL or the alternative methods when ρ is negative or larger than 0.12. Also the study uses a single design matrix with only four variables and fixed design of 20 clusters of 51 observations. Although this limits the generalizability of our results, the conditions we consider are relevant to health applications where study participants are clustered within healthcare units, such as hospitals or providers. The distributions of x_1–x_4 and the design parameters were chosen to cover settings that are likely to affect the performance of the estimators and test procedures. Care should be taken,

however, in extrapolating our findings to analyses with substantially fewer or more than 20 clusters. Any of the methods could perform poorly with fewer than 20 clusters, while results could be expected to improve as the number of clusters grows. Any such relationship may depend on other factors, as illustrated by differences in the performance of tests across the explanatory variables in the simulation. For variables like x_4, we expect BRL tests to work well with most sample sizes, whereas we expect coefficients that are determined by few clusters, like the one for x_2, to be problematic even as the number of clusters increases above 20.

Our simulation study clearly demonstrates that the BRL methodology can greatly improve tests compared with more commonly used methods. It also suggests that BRL is likely to provide the greatest improvement when the number of clusters is small and the independent variables have large ICC, especially if observations from a small subset of clusters are influential for determining estimated coefficients.

REFERENCES

1. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
2. Fellegi I. Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association* 1980; **75**:261–268.
3. Lipsitz S, Laird N, Harrington D. Using the jackknife to estimate the variance of regression estimators from repeated measures studies. *Communications in Statistics—Theory and Methods* 1990; **19**:821–845.
4. Qu Y, Piedmonte M, Williams G. Small sample validity of latent variable models for correlated binary data. *Communications in Statistics—Simulation* 1994; **23**:243–269.
5. Gunsolley J, Getchell C, Chinchilli V. Small sample characteristics of generalized estimating equations. *Communications in Statistics—Simulation* 1995; **24**:869–878.
6. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**:126–134.
7. Bell RM, McCaffrey DF. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology* 2002; **28**:169–179.
8. Pan W, Wall M. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine* 2001; **21**:1429–1441.
9. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989.
10. Shah BV, Barnwell BG, Bieler GS. *SUDAAN User's Manual*, *Release 7.5*. Research Triangle Institute: Research Triangle Park, NC, 1997.
11. Rotnitzky A, Jewell N. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 1990; **77**:485–497.
12. McCaffrey DF, Bell RM, Botts CH. Generalizations of bias reduced linearization. *Proceedings of the Annual Meeting of the American Statistical Association*, 5–9 August 2001.
13. Satterthwaite F. An approximate distribution of estimates of variance components. *Biometrics* 1946; **2**:110–114.
14. Lugannani R, Rice S. Saddlepoint approximations for the distribution of the sum of independent random variables. *Advances in Applied Probability* 1980; **12**:475–490.
15. Huzurbazar S. Practical saddlepoint approximations. *The American Statistician* 1999; **53**:225–232.
16. Butler RW, Paolella MS. Saddlepoint approximation and bootstrap inference for the Satterthwaite class of ratios. *Journal of the American Statistical Association* 2002; **97**:836–846.
17. Kott PS. Linear regression in the face of specification error: model-based exploration of randomization-based techniques. *Statistical Society of Canada Proceedings of the Survey Methods Section* 1996; 39–47.
18. Oman SD, Zucker DM. Modelling and generating correlated binary variables. *Biometrika* 2001; **88**:287–290.
19. Wells KB, Sherbourne C, Schoenbaum M *et al*. Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *Journal of the American Medical Association* 2000; **283**:212–220.
20. Elliot M, Golinelli D, Hambarsoomians K, Perlman J, Wenzel S. *Sampling with Field Burden Constraints*: *An Application to Sheltered Homeless and Low-income Housed Women DRU-3057-NIDA*. RAND: Santa Monica, CA, 2003.
21. Miller RG. The jackknife—a review. *Biometrika* 1974; **61**:1–15.
22. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.

