Small-Sample Adjustments for Wald-type Tests using Sandwich Estimators

Michael P. Fay* and Barry I. Graubard#

(This is the same a Biometrics, 2001, 57: 1198-1206, except for the formatting and the contact information, which has been updated on October 19, 2007)

* National Institute of Allergy and Infectious Diseases

Email: mfay@niaid.nih.gov

# National Cancer Institute

Email: graubarb@exchange.nih.gov

August 13, 2001

SUMMARY

The sandwich estimator of variance may be used to create robust Wald-type tests from estimating equations that are sums of K independent or approximately independent terms. For example, for repeated measures data on K individuals, each term relates to a different individual. These tests applied to a parameter may have greater than nominal size if either K is small, or more generally if the parameter to be tested is essentially estimated from a small number of terms in the estimating equation. We offer some practical modifications to these robust Wald-type tests which asymptotically approach the usual robust Wald-type tests. We show that one of these modifications provides exact coverage for a simple case, and examine by simulation the modifications applied to the generalized estimating equations of Liang and Zeger (1986), conditional logistic regression, and the Cox proportional hazard model.

Keywords: Conditional logistic regression; Cox proportional hazards model; Generalized estimating equations; Robust Wald statistics; Sandwich estimator; Small sample size

1 Introduction

The sandwich estimator of variance has been used with many different types of data to provide inferences robust to certain model misspecifications. We mention three examples, Liang and Zeger (1986), using their generalized estimating equations (GEEs), applied sandwich estimators of variances to repeated measures data, Lin and Wei (1989) applied them to the Cox proportional hazards model, and Fay, et al. (1998) applied them to conditional logistic regression. Each of these applications use estimating equations with K independent or approximately independent terms. The Wald-type tests used with the sandwich estimator are valid as K goes to infinity. For finite samples simulations have shown that these (unadjusted) Wald sandwich tests tend to be liberal (see Emrich and Piedmonte, 1992; Fay et al. 1998; Gunsolley, Getchell, and Chinchilli, 1995; Lin and Wei, 1989; and Mancl and DeRouen, 2001). In this paper we consider finite sample adjustments for Wald sandwich tests and apply them to quite general estimating equations which include the three applications mentioned above.

Consider the data that motivated this research. Fay, et al. (1997) performed a meta analysis to determine the effect of different types of dietary fat on the development of mammary tumors in rodents. The analysis combined 146 experiments ("sets" in the terminology of Fay et al., 1997). Each experiment has two or more groups of animals, each group receives the same intervention except for diets, and the response is the number of animals in the group that develop tumors. Although within an experiment there were only diet differences between the groups, between experiments there were many differences. For example, the type and/or dose of carcinogen and the amount of follow-up time varied between experiments. Thus, as detailed in Fay, et al. (1998), a conditional logistic regression was used to condition out intercept differences between experiments, and a sandwich estimator of variance was used to account for heterogeneity of the diet parameters. A random effects model for the diet parameters was not tractable because each of the diet effects was not varied within each experiment. For example, Fay et al. (1997) measured effects of 4 types of fat, but one of the types of fats, n-3 polyunsaturated fatty acid (n-3 PUFA), was rarely present except in very small portions. Only 22 of the experiments had any dietary groups with greater than 1% of the dietary fat equal to n-3 PUFA. Thus, it is difficult to estimate a random effect for the n-3 PUFA parameter in all but these 22 experiments. The important aspect of these data for this paper is that even though there is a large sample size of independent observations, i.e., 146 experiments that make up the terms in the estimating equation, there is an effectively small sample size for testing the n-3 PUFA parameter. When a parameter is primarily estimated from a small proportion of the terms in the estimating equations (e.g., the n-3 PUFA parameter in the diet data), we call the estimating equations unbalanced for that parameter. We reanalyze the diet data in section 3.3.

We propose two types of adjustments. First, we use Taylor series approximations to adjust for the bias of the sandwich estimator of variance. Mancl and DeRouen (2001) provide such an adjustment for GEEs and compare it to many earlier bias corrections (see that paper for earlier references). In this paper we propose a bias correction similar to that of Mancl and DeRouen (2001), but our correction may be applied to more general estimating equations than the GEE. Our second adjustment is to use an F (or t) distribution instead of a chi square (or a normal) distribution to calculate significance. Others have used this adjustment with the canonical use of 1 and $K - p$ degrees of freedom, where p is the number of parameters describing the mean (see Lipsitz, et al. 1994 and Mancl and DeRouen, 2001, for the GEE case). As seen for the diet example above, these canonical degrees of freedom may not be appropriate for estimating equations that are unbalanced for certain parameters. This problem was noted by Chesher and Austin (1991) for the simple linear model. We propose an estimator of degrees of freedom that produces different degrees of freedom depending on which parameter (or combination of parameters) is being tested. These estimators can more properly adjust for situations such as the diet example. Similar but not equivalent small sample adjustments have been proposed for some special cases (see Fay, et al., 1998 for conditional logistic regression; Lipsitz and Ibrahim (1999) for linear regression; and Fai and Cornelius (1996) for unbalanced split-plot experiments).

Because of the way we motivate our degrees of freedom estimators, we restrict ourselves to tests of linear combinations of parameters. In other words if β is the parameter vector, then we are restricted to tests with null hypotheses equal to $C^T\beta = C^T\beta_0$, where β_0 is known, and C is a vector of constants (i.e., C is not allowed to be a matrix).

In section 2.1 we review the use of the sandwich estimator of variance and its associated Wald test. In section 2.2 we propose our bias correction and compare it to that of Mancl and DeRouen (2001), and in section 2.3 we propose some degrees of freedom estimators. In section 3 we examine one special case where one of the modifications produces an exact test, and we perform simulations to test these modifications when used with GEEs, conditional logistic regression, and Cox proportional hazards models.

2 Main Result

2.1 Background

Consider estimating equations of the form, $\sum_{i=1}^{K} U_i(\beta) = 0$, where β is a $p \times 1$ parameter vector. Let $\widehat{\beta}$ be the solution to the estimating equations. Let $U_i = U_i(\beta)$ and let a hat over any function denote evaluation at $\widehat{\beta}$ (e.g., $\widehat{U}_i = U_i(\widehat{\beta})$). Assume $E\left\{\sum_{i=1}^{K} U_i(\beta_0)\right\} = 0$ for some β_0, and both $\text{cov}\left\{U_i(\beta_0), U_j(\beta_0)\right\} \to 0$ for $i \neq j$, and $\widehat{\beta} - \beta_0 \xrightarrow{P} 0$ as $K \to \infty$. We use a Taylor series approximation about $\hat{\beta}$ but replace the derivative by an estimator,

$$U_i \approx \widehat{U}_i - \widehat{\Omega}_i\left(\beta - \widehat{\beta}\right), \tag{1}$$

where $\widehat{\Omega}_i$ is an estimator of $-\partial U_i/\partial\beta$ evaluated at $\widehat{\beta}$. Summing over all clusters and rearranging terms gives,

$$\widehat{\beta} - \beta \approx V_m\left(\sum_{i=1}^{K} U_i\right), \tag{2}$$

where $V_m = \left(\sum_{i=1}^K \widehat{\Omega}_i\right)^{-1}$ and will be called the "model-based" variance. If V_m is approximately constant with small changes in $\widehat{\beta}$, then the variance of $\widehat{\beta} - \beta$ may be estimated with the sandwich estimator, $V_s = V_m \left(\sum_{i=1}^K \widehat{U}_i \widehat{U}_i^T\right) V_m$.

We consider hypotheses of the form, Null: $C^T\beta = C^T\beta_0$ vs. Alternative: $C^T\beta \neq C^T\beta_0$, where C is a $p \times 1$ vector of constants. For example, we test whether $\beta_j = 0$ or not by letting C be all zeros except for a one in the jth row and $\beta_0 = 0$. The unadjusted Wald sandwich test rejects when $T_s^2 = \left\{C^T(\widehat{\beta} - \beta_0)\right\}^2 / C^T V_s C > (\chi_1^2)^{-1}(1-\alpha)$, where $(\chi_1^2)^{-1}(q)$ is the qth quantile of the chi-square distribution with one degree of freedom.

2.2 A Bias-Corrected Sandwich Covariance Estimator

Mancl and DeRouen (2001) motivated a bias correction of V_s for the GEE case by using a first-order Taylor series expansion of the ith residual vector together with approximation (2), to approximate the expected value of the ith squared residuals. Analogously, we use a first-order Taylor series expansion of U_i (approximation (1)) together with approximation (2), to obtain,

$$E\left(\widehat{U}_i \widehat{U}_i^T\right) \approx \left(I_p - \widehat{\Omega}_i V_m\right) \Psi_i \left(I_p - \widehat{\Omega}_i V_m\right)^T + \widehat{\Omega}_i V_m \left(\sum_{j \neq i} \Psi_j\right) V_m \widehat{\Omega}_i, \quad (3)$$

where $\Psi_i = \mathrm{cov}(U_i)$ and I_p is a $p \times p$ identity matrix. For tractability, Mancl and DeRouen (2001) motivated their bias correction by assuming the the last term in their expression (4) is small (note there is a typo in expression (4) of Mancl and DeRouen (2001); the $\mathrm{cov}(y_i)$ in last term should be $\mathrm{cov}(y_j)$), which is analogous to assuming here that the last term in expression (3) is small.

We take a different approach to tractability. We consider a correction that is reasonable when the working variance model is approximately within a scale factor of the true variance, i.e., when $\Psi_i \approx c\widehat{\Omega}_i$ for all i and some constant c and the model-based variance is consistent. For example, in the GEE case this would occur when the correlation model is correctly specified. When $\Psi_i \approx c\widehat{\Omega}_i$, approximation (3) simplifies to $E\left(\widehat{U}_i \widehat{U}_i^T\right) \approx \left(I_p - \widehat{\Omega}_i V_m\right) \Psi_i \approx \Psi_i \left(I_p - V_m \widehat{\Omega}_i\right)$. To partially correct for this bias (i.e., $E\left(\widehat{U}_i\widehat{U}_i^T\right) \neq \Psi_i$) we estimate Ψ_i with $\widehat{\Psi}_i = H_i\widehat{U}_i\widehat{U}_i^T H_i^T$, where H_i is given below and the form of the correction ensures that $\widehat{\Psi}_i$ is a symmetric nonnegative definite matrix. In general, $\left(I_p - \widehat{\Omega}_i V_m\right)$ is not symmetric, so that the choice of $H_i = \left(I_p - \widehat{\Omega}_i V_m\right)^{-1/2}$ may not exist. Instead, we propose the simple bias correction of letting H_i be a $p \times p$ diagonal matrix with jjth element equal to $\left\{1 - \min\left(b, \{\widehat{\Omega}_i V_m\}_{jj}\right)\right\}^{-1/2}$, where $b < 1$ is a constant defined by the user. Setting a bound, b, is a practical necessity to prevent extreme adjustments when the jjth element of $\widehat{\Omega}_i V_m$ is very close to 1. (In fact it is possible for the jjth element of $\widehat{\Omega}_i V_m$ to be greater than 1, see http://srab.cancer.gov/sandwich.) We arbitrarily use $b = .75$ for our simulations which ensures that each diagonal element of H_i is less than or equal to 2. We write this adjusted sandwich estimator as $V_a = V_m\left(\sum_{i=1}^K H_i\widehat{U}_i\widehat{U}_i^T H_i\right)V_m$. Although the bound .75 is arbitrary, the bound of $b = .75$ is rarely reached. For example, the simulations in section 3 gave almost exactly the same results (results not shown) when run without any bound (i.e., the bound is infinity).

2.3 An Approximate F-Distribution

Let T_a^2 be the Wald test statistic using V_a instead of V_s. The main result motivated in this section is

$$T_a^2 = \frac{\left\{C^T(\widehat{\beta} - \beta_0)\right\}^2}{C^T V_a C} \approx \frac{U^T B_0 U}{U^T B_1 U} \mathrel{\dot\sim} F_{1,d} \tag{4}$$

where $\dot\sim$ denotes approximate distribution under the null hypothesis, $U^T = [U_1^T \ \cdots \ U_k^T]$, B_0 and B_1 are $pK \times pK$ matrices given in appendix A, $F_{1,d}$ is an F distribution with 1 and d degrees of freedom, and we give estimators of d later. Following a standard derivation of the F distribution, our F distribution approximation may be motivated by the following 3 approximate conditions: (1) $\sigma^{-2}U^T B_0 U \mathrel{\dot\sim} \chi_1^2$, (2) $\sigma^{-2}dU^T B_1 U \mathrel{\dot\sim} \chi_d^2$, and (3) $U^T B_0 U$ and $U^T B_1 U$ are approximately independent, where $\sigma^2 = \text{var}\left\{C^T\left(\widehat{\beta} - \beta_0\right)\right\}$. We discuss each of these approximations in appendix A.

In order to estimate d we need estimators of Ψ_i for $i = 1, \ldots, K$. We estimate Ψ_i in one of two ways. First, we simply use $\widehat{\Psi}_i$ as in Section 2.2; this estimator tends to overestimate the heterogeneity of the Ψ_i. One can see this by noting that even when all Ψ_i are equal, the $\widehat{\Psi}_i$ vary. The associated estimator of d is

$$\widehat{d}_H = \frac{\left\{\text{trace}\left(\widehat{\Psi}B_1\right)\right\}^2}{\text{trace}\left(\widehat{\Psi}B_1\widehat{\Psi}B_1\right)}, \tag{5}$$

where $\widehat{\Psi}$ = block diagonal($\widehat{\Psi}_1, \ \cdots \ \widehat{\Psi}_K$) (see Appendix A).

Alternatively, we propose the more complicated $\tilde{\Psi}_i = w_i\left(\sum_{\ell=1}^K w_\ell\right)^{-1}\left(\sum_{j=1}^K \widehat{\Psi}_j\right)$ where $w_i = C^T\left\{\left(\sum_{j\neq i}\widehat{\Omega}_j\right)^{-1} - V_m\right\}C$ and $w_i\left(\sum w_\ell\right)^{-1}$ represents the proportional reduction in the model-based variance of $C^T(\widehat{\beta} - \beta)$ due to adding the ith cluster. The idea behind the $\tilde{\Psi}_i$ is that it smooths the extremely variable estimates $\widehat{\Psi}_j$, $j = 1, \ldots, K$, yet unlike the unweighted sum (i.e., using $K^{-1}\sum\widehat{\Psi}_j$ to estimate each Ψ_i) it still accounts for some of the differences in cluster variability associated with the working variance model. We estimate $\tilde{d}_H$ in a similar manner to $\hat{d}_H$ by replacing $\widehat{\Psi}_i$ with $\tilde{\Psi}_i$ in equation 5.

To see if the adjustment of section 2.2 is necessary, we estimate the distribution of T_s^2 with an F distribution with 1 and d (equal to either $\widehat{d}$ or $\tilde{d}$) degrees of freedom. Here $\widehat{d}$ is calculated the same as $\widehat{d}_H$ except replace H_i with an identity matrix for all i. Similarly define $\tilde{d}$. We compare the different methods in section 3.

The adjustments of this section do not affect the asymptotic properties of the unadjusted Wald sandwich test. Under the assumption that the data are sufficiently regular such that $\widehat{\Omega}_i\left(\sum\widehat{\Omega}_j\right)^{-1} \xrightarrow{P} 0$ for all i and $d \to \infty$ as $K \to \infty$, then as $K \to \infty$, $\left(I_p - \widehat{\Omega}_i V_m\right)^{-1} \xrightarrow{P} I_p$ and $V_a - V_s \xrightarrow{P} 0$. Further, if $d \to \infty$ then the $F_{1,d}$ distribution approaches a chi square distribution with 1 degree of freedom. Thus, even if the assumptions and approximations that motivate (4) are tenuous in some situations, the adjustments of this section may likely be an improvement over the unadjusted Wald sandwich test.

3 Simulations and Special Cases

3.1 Normal Model with the Same Design for Each Cluster

Consider the (true) model $Y_i \sim N(X\beta, \Sigma)$, where Y_i is an $n \times 1$ vector of responses, X is an $n \times p$ full-rank design matrix, $n \geq p$, and Σ is a $p \times p$ covariance matrix. We use independence estimating equations (see Liang and Zeger, 1986), i.e., $U_i = \widehat{\phi} X^T (Y_i - X\beta)$, where $\widehat{\phi}^{-1}$ is a scalar dispersion estimator. Under the null model for any non-degenerate Σ, X, and C, then $T_a^2 = \{(K-1)/K\}\, T_s^2 \sim F_{1,K-1}$. Further, here the degrees of freedom estimators, $\tilde{d}_H$ and $\tilde{d}$, give $K-1$, so that the test, say δ_5, that compares T_a^2 with $F_{1,\tilde{d}_H}$ produces an exact test for any K (see Appendix B). In contrast, for $K = 20$, the standard sandwich test, δ_1 (T_s^2 compared to χ_1^2), has size .071, δ_2 (T_s^2 compared to $F_{1,\widehat{d}}$) has simulated size .034, δ_3 (T_s^2 compared to $F_{1,\tilde{d}}$) has size .055, δ_4 (T_a^2 compared to $F_{1,\widehat{d}_H}$) has simulated size .031, where the simulations had 100,000 replications. The associated test statistic of Mancl and DeRouen (2001) is $T_{MD}^2 = \{(K-1)/K\}^2\, T_s^2$. For $K = 20$ comparing T_{MD}^2 to χ_1^2 has size .059, while comparing it to $F_{1,K-p}$ with $p = 1$ has size .045, and comparing it to $F_{1,K-p}$ with $p = 2$ has size .044. Thus, in this case δ_5 performs best, but δ_3 and the tests of Mancl and DeRouen (2001) have close to nominal size.

3.2 Generalized Estimating Equations

Consider GEEs of the form (see Liang and Zeger, 1986), $\sum_{i=1}^K U_i(\beta) = \sum_{i=1}^K D_i^T(\beta) V_i^{-1}(\beta)\, \{Y_i - \mu_i(\beta)\}$, where Y_i is a $n_i \times 1$ vector of responses, μ_i is the model of $E(Y_i)$, $D_i = \partial \mu_i(\beta)/\partial \beta$ and $V_i(\beta)$ estimates the "working variance" of Y_i. The function $V_i(\beta)$ has a complicated form. (It is denoted $\tilde{V}_i(\beta)$ in Liang and Zeger, 1986. See that paper for details). Let our estimate of $-\partial U_i/\partial \beta$ be $\widehat{\Omega}_i = \widehat{D}_i^T \widehat{V}_i^{-1} \widehat{D}_i$, and V_s is the sandwich estimator proposed by Liang and Zeger (1986). To test the GEE models we simulated both Poisson and binomial data and tested the GEE model with both independence and exchangeable working correlation within cluster.

For the Poisson case we simulated 4 types of data sets all with $K = 20$ clusters. Each type of data set is denoted by one level of each of 2 descriptors, the variance (either Poisson or overdispersed Poisson), and the treatment assignment (either changed within cluster or fixed within cluster). For the models that changed treatment assignments within cluster we used the working independence variance and for those that did not we used the working exchangeable variance. For the model-based variance we included a scalar overdispersion term as described in Liang and Zeger (1986); it is the default in the yags software which we modified to use for this simulation (the yags function written in Splus by V. Carey, is reviewed in Horton and Lipsitz, 1999, and can be found at http://biosun1.harvard.edu/~carey/index.ssoft.html). The details of the data generation are listed with the results in Table 1. We see that δ_1 is liberal in all 4 cases, δ_3 is less liberal, while δ_5 appears to have values closer to the nomial level. The test δ_6 (T^2_{MD} compared to $F_{1,K-p}$) appears to perform well, although perhaps slightly conservatively. The tests δ_2 and δ_4 can be very conservative especially in the cases with both treatments in each cluster. We see in the last column of the first row that δ_m may perform very poorly when the model is misspecified. All of the modified tests ($\delta_2, \delta_3, \delta_4$, δ_5, and δ_6) appear to have sizes that come closer to maintaining the nominal level than the standard sandwich test, and the average length of the confidence intervals give a crude measure of the price paid to achieve those sizes.

We modeled the simulation for logistic regression after real data. Preisser and Qaqish (1999) have made available data from the Guidelines for Urinary Incontinence Discussion and Evaluation (GUIDE) study at http://www.phs.wfubmc.edu/data/uipreiss.html. Preisser and Qaqish (1999) use 5 covariates measured from a baseline survey to predict whether individual patients will respond yes or no that they are bothered by accidental loss of urine. There are 137 patients in the study from 38 practices, and 54 patients responded that they were bothered by their urinary incontinence. We fit a logistic model without each of the five covariates in the model and used each of the five sets of fitted values as the true probability of response. Then simulated binary responses using those probabilities. Thus each of the five sets of simulated data come from a logistic model with one of the parameters equal to zero. We then test whether that parameter equals zero. For this simulation we let the working correlation be the independence model. The results are presented in Table 2. In this case the model based variance is correctly specified and δ_m performs well, δ_1 is appears consistently liberal, while δ_4 and δ_5, and appear quite conservative, with δ_2 and δ_6 slightly less conservative. In this case, δ_3 appears to be the best of the sandwich tests.

In Table 3 we list the sample variance of the parameter estimate and compare it to the mean of the variance estimates for the four variance estimates, V_m, V_s, V_{MD} (Mancl and DeRouen's, 2001, adjustment), and V_a. We bold the variance estimator with mean closest to the sample variance of the parameter estimate. When the model is misspecified (see the fourth row), V_m underestimates the variance, but otherwise V_m does well. The estimator V_s tends to underestimate the variance. The neither of the corrected sandwich variance estimates, V_{MD} and V_a, appears to regularly outperform the other. In many situations both V_{MD} and V_a appear to overcorrect for bias.

3.3 Conditional Logistic Regression

For conditional logistic regression, we have binary responses grouped into clusters, and within each cluster we condition on the total number of positive responses. The estimating equations can be written in terms of the sufficient statistic for β, t_i, i.e., $U_i(\beta) = t_i - E(t_i|\beta)$, where $E(t_i|\beta)$ is the modeled value for t_i given β. Then $\Omega_i(\beta) = -\partial U_i / \partial \beta$. For details see e.g., Fay, et al. (1998).

We repeat 4 sets of the simulations from Fay et al. (1998) which were motivated by a meta-analysis. We briefly describe the simulations using the same terminology as Fay et al. (1998); see that paper for more details. Each simulation estimates $\beta = [\beta_1 \;\; \beta_2]^T$ from K clusters of 60 binary responses, and here we test whether β_1 is equal to zero or not. Each of the 4 sets of simulations is described by both the "Design" (either A or D) and the "Case" (either 1 or 4), and consists of 1000 simulations. Design A has $K = 20$ clusters and by design 1/4 of the clusters do not contribute the estimation of β_1 because of conditioning, while similarly Design D has $K = 40$ clusters but 37/40 of the clusters do not contribute the estimation of β_1. The responses are all generated with a random intercept term for each cluster. For Case 1 the effects for β are fixed, while for Case 4 these effects are random. Thus, the conditional logistic model is correctly specified for Case 1 but not for Case 4. All simulations are under the null hypothesis. The results are presented in Table 4. Note that Design D shows that δ_1 can be very liberal in extreme cases.

We return to the meta analysis of Fay, et al. (1997) mentioned in the introduction. The 6 different tests that the effect of the n-3 PUFA is different from zero give two-sided p-values of $p = 0.0134$ for δ_m, $p = 0.3377$ for δ_1, $p = 0.4136$ for δ_2, $p = 0.3590$ for δ_3, $p = 0.4320$ for δ_4, and $p = 0.3824$ for δ_5. If there is a random effect of the n-3 PUFA the usual sandwich test, δ_1, gives better coverage than the model based test, δ_m, but judging from our simulations may be slightly liberal. Based on our simulations, the modified sandwich tests ($\delta_2, \delta_3, \delta_4$, and δ_5) appear more likely to have nominal coverage. We recommend the use of either δ_3 or δ_5 for having better coverage than δ_1 while (usually) not being overly conservative. For details of the biological issues and the full model see Fay, et al. (1997).

3.4 Cox Proportional Hazards

The Cox model uses the partial likelihood for right censored failure time data. Let the associated efficient score statistic for β be written as $\sum_{i=1}^{K} U_i^{\bullet}(\beta)$, where the summation is over the K individuals whose failure time may or may not be right censored (see Appendix C for details of the notation). The terms $U_i^{\bullet}$ are not independent, but Lin and Wei (1989) showed that the estimating equation may be written as $\sum_{i=1}^{K} U_i^{\bullet}(\beta) = \sum_{i=1}^{K} U_i(\beta)$, where the U_i are asymptotically independent, $U_i = U_i^{\bullet} + U_i^{\circ}$, and where $\sum_{i=1}^{K} U_i^{\circ}(\beta) = 0$ for all β. Then $V_m = -\left\{\sum_{i=1}^{K} \partial U_i^{\bullet}/\partial\beta + \partial U_i^{\circ}/\partial\beta\right\}^{-1} = \left(\sum_{i=1}^{K} -\partial U_i^{\bullet}/\partial\beta\right)^{-1}$ and the sandwich estimator of Lin and Wei (1989) is V_s.

Although we do not need to define Ω_i, our estimator of $-\partial U_i/\partial\beta$, for the calculation of V_m and V_s, for our adjustments we do need to define Ω_i for $i = 1, \ldots, K$. To do this we use numerical derivatives, similar to the method that Gail, Lubin, and Rubinstein (1981) used to calculate $\partial U_i^{\bullet} / \partial \beta$. We let the abth element of $\widehat{\Omega}_i$ be $\left\{ \Omega_i(\widehat{\beta}) \right\}_{ab} = (4h)^{-1} \left\{ U_{ia}(\widehat{\beta} + h_b) - U_{ia}(\widehat{\beta} - h_b) + U_{ib}(\widehat{\beta} + h_a) - U_{ib}(\widehat{\beta} - h_a) \right\}$, where U_{ia} is the ath element of U_i, h_a is a vector of length p with all zeros except the ath row which has a value of h, and h is some small number. We use $h = 0.0001$ for our simulations.

We performed simulations on 3 true models, (1) a proportional hazards model, (2) model 10 from Table 1 of Lin and Wei (1989) in which the model-based test (δ_m) performed worst for $K = 50$, and (3) model 11 from the same table in which the sandwich test (δ_1) performed worst for $K = 50$. We give the details of the models and the results in Table 5. Again the modified sandwich tests ($\delta_2, \delta_3, \delta_4$, and δ_5) appear closer to the nominal level than the standard sandwich test, δ_1, although δ_3 appears quite liberal for $K = 20$ for models 10 and 11.

4 Discussion

We have examined 4 new modifications to the standard Wald test with a sandwich estimator of variance. These modifications were necessary because the standard sandwich test is liberal when the parameter to be tested is essentially estimated from a small number of terms in the estimating equation. This liberalness can occur when either K, the number of terms in the estimating equation, is small, or when primarily a small proportion of terms are used to estimate the parameter (or linear combination of parameters). We have referred to the latter type of estimating equations as unbalanced for that parameter.

In all of our simulations the 4 new modifications had estimated size less than the liberal size of the standard sandwich test. In the simple balanced normal model, δ_5 is exact, and δ_5 did reasonably well in the other balanced cases (see Tables 1 and 5, and Design A from Table 4). However, in more unbalanced cases (see Table 2 and Design D from Table 4), δ_5 appeared to produce a quite conservative test. For this reason, unless the data are fairly well balanced, we favor the use of δ_3, which appears to be less liberal than δ_1 but not overly conservative even for unbalanced equations. The difference between δ_5 and δ_3 is that δ_5 includes a bias correction for the variance, while δ_3 does not. In our simulations on the GEE case, the standard sandwich estimator regularly underestimated the variance, while both our adjusted variance estimator, V_a, and that proposed by Mancl and DeRouen (2001), V_{MD}, appeared less biased in the balanced cases (see top section of Table 3). In the more unbalanced cases (see bottom section of Table 3), both V_a and V_{MD} appeared to overcorrect for bias in most cases, and this overcorrection in V_a may be the source of the conservativeness of δ_5. Further work is required on correcting the bias of the sandwich variance estimator in unbalanced data. For the balanced cases the advantage of V_a over V_{MD} is that it may be applied to other cases besides the GEE case.

Although not listed in the tables, we included in our simulations a test comparing T_s^2 to $F_{1,K-p}$ as has been used in the literature (see e.g., Lipsitz,et al. 1994). This test gave estimated sizes that were always less than the standard sandwich test (δ_1), but at least as large as the 4 new tests ($\delta_2, \delta_3, \delta_4$, and δ_5). As expected this test does reasonably well with balanced data but not with unbalanced data. For example, for Table 4 the 4 estimated sizes (to test $\beta_1 = 0$) were .078 (Design A, Case 1), .073 (Design A, Case 4), .245 (Design D, Case 1), and .248 (Design D, Case 4). For insight into this, note the failure of that method to address the degrees of freedom differences in the parameters in Design D, where β_1 is estimated from only 3 clusters, while β_2 is estimated from 39 clusters (see Fay, et al., 1998). The degrees of freedom estimate, $K - p$, does not properly account for this data structure and compares both the T_s^2 statistic for testing $\beta_1 = 0$ and the T_s^2 statistic for testing $\beta_2 = 0$ against the same distribution, $F_{1,K-p}$. The complete simulation results of the tests comparing T_s^2 to $F_{1,K-p}$ along with Splus functions that perform our adjustments and the Splus programs used to perform all the simulations are given at http://srab.cancer.gov/sandwich.

Acknowledgements

We thank William Davis, Edward Korn, and the reviewers and editors for helpful comments on earlier versions of this paper.

References

Chesher, A., and Austin, G. (1991). The finite-sample distributions of heteroscedasticity robust Wald statistics. *Journal of Econometrics* **47,** 153-173.

Emrich, L.J., and Piedmonte, M.R. (1992). On Some Small Sample Properties of Generalized Estimating Equation Estimates for Multivariate Dichotomous Outcomes. *Journal of Statistical Computation and Simulation,***41,** 19-29.

Fai, A.H-T., and Cornelius, P.L. (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation* **54**, 363-378.

Fay, M.P., Freedman, L.S., Clifford, C.K., and Midthune, D.N. (1997). Effect of different types and amounts of fat on the development of mammary tumors in rodents: a review. *Cancer Research* **57**, 3979-3988.

Fay, M.P., Graubard, B.I., Freedman, L.S., and Midthune, D.N. (1998). Conditional logistic regression with sandwich estimators: application to a meta-analysis. *Biometrics*, **54**, 195-208.

Gail, M.H., Lubin, J.H., and Rubinstein, L.V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika*, **68**, 703-707.

Gunsolley, Getchell, and Chinchilli (1995) Small Sample Characteristics of Generalized Estimating Equations. *Communications in Statistics: Simulation and Computation* **24,** 869-878.

Horton, N.J., and Lipsitz, S.R. (1999). Review of Software to Fit Generalized Estimating Equation Regression Models. *American Statistician* **53,** 160-169.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika,* **73,** 13-22.

Lin, D.Y., and Wei, L.J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84**, 1074-1078.

Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J., and Laird,N.M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics* **50,** 270-278.

Lipsitz, S.R., and Ibrahim, J.G. (1999). A degrees-of-freedom approximation for a t-statistic with heterogeneous variance. *The Statistician* **48**, 495-506.

Mancl, L.A., and DeRouen, T.A. (2001). A covariance estimator for GEE with improved small sample properties. *Biometrics* **57,** 126-134.

Preisser, J.S., and Qaqish, B.F. (1999). Robust regression for clustered data with application to binary responses. *Biometrics,* **55,** 574-579.

Searle, S.R. (1982). *Matrix Algebra useful for Statistics* John Wiley and Sons, New York.

A Details of Degrees of Freedom Approximation

Using approximation (2), we write $\left\{C^T\left(\widehat{\beta}-\beta_0\right)\right\}^2 \approx U^T B_0 U$, where $B_0 = FV_m CC^T V_m F^T$, and $F^T = [I_p \ \cdots \ I_p]$. Rewrite $C^T V_a C$ as

$$C^T V_a C = \sum_{i=1}^{K} C^T V_m \left(H_i \widehat{U}_i \widehat{U}_i^T H_i\right) V_m C = \sum_{i=1}^{K} \widehat{U}_i^T \left(H_i V_m C C^T V_m H_i\right) \widehat{U}_i,$$

where the second equality uses the fact that $C^T V_m H_i \widehat{U}_i$ is a scalar; this is why we require C to be a vector. We rewrite this as $C^T V_a C = \widehat{U}^T M \widehat{U}$, where M is a $pK \times pK$ block diagonal matrix with ith block equal to $H_i V_m C C^T V_m H_i$. We combine the Taylor series approximations for the $\widehat{U}_i$ to get $\widehat{U} \approx GU$, where $G = I_{pK} - \widehat{\Omega} V_m F^T$ and $\widehat{\Omega}^T = [\widehat{\Omega}_1 \quad \cdots \quad \widehat{\Omega}_K]$. Thus, $C^T V_a C \approx U^T B_1 U$, where $B_1 = G^T M G$.

Now consider the 3 conditions:

(1) $\sigma^{-2} U^T B_0 U \dot{\sim} \chi^2_1$**:** Assume $C^T V_m U_i$ has mean 0 and unknown variance σ_i^2 (where we allow $\sigma_i^2 = 0$ for some i), that the $C^T V_m U_i$ are independent, and that the approximation 2 holds. Provided some regularity conditions are true, by the Liapunouv central limit theorem $\sigma^{-1} C^T \left(\widehat{\beta} - \beta_0\right) \rightarrow N(0,1)$ as $K \rightarrow \infty$, where $\sigma^2 = \sum_{i=1}^K \sigma_i^2$. Thus, $\sigma^{-2} \left\{C^T \left(\widehat{\beta} - \beta_0\right)\right\}^2$ is asymptotically chi-squared with 1 degree of freedom.

(2) $\sigma^{-2} d U^T B_1 U \dot{\sim} \chi^2_d$**:** Assume U is distributed normal with mean 0 and variance $\Psi =$ block diagonal $(\Psi_1, \ldots, \Psi_K)$. Under these assumptions, $U^T B_1 U$ has a distribution described by a weighted sum of chi square random variables; however, we approximate its distribution with a (much simpler) Gamma distribution with the same mean and variance, i.e., with mean equal to trace(ΨB_1) and variance equal to 2trace$(\Psi B_1 \Psi B_1)$ (see Searle, 1982, p. 355). Since $U^T B_1 U$ is an approximately unbiased estimator of σ^2 we estimate σ^2 with trace(ΨB_1), so that $\sigma^{-2} U^T B_1 U$ is approximately Gamma with mean 1 and variance 2trace$(\Psi B_1 \Psi B_1) / \{\text{trace}(\Psi B_1)\}^{-2}$ or equivalently $\sigma^{-2} d U^T B_1 U \dot{\sim} \chi^2_d$ where $d = \{\text{trace}(\Psi B_1)\}^2 / \text{trace}(\Psi B_1 \Psi B_1)$.

(3) $U^T B_0 U$ **and** $U^T B_1 U$ **are independent:** Assume that U is normal with mean 0 and variance Ψ, with rank$(\Psi) = pK$. Under correct model specification, i.e., $\Psi_i \approx c\widehat{\Omega}_i$ for all i. Then $B_0 \Psi B_1 \approx 0$ and $U^T B_0 U$ and $U^T B_1 U$ are approximately independent. (See Searle, 1982, p.356).

B Details: Normal Model with Identical Designs

Refer to structure defined in section 3.1. Let $Z_i = K^{-1/2}\sigma^{-1}\left\{C^T(X^TX)^{-1}X^TY_i - C^T\beta_0\right\}$ where $\sigma^2 = \text{var}\left\{C^T\left(\widehat{\beta}-\beta_0\right)\right\} = K^{-1}C^T(X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}C$. Under the null hypothesis assumptions of the model, the Z_i are independent standard normal random variables. Then, $K^{-1}\sum_{i=1}^K Z_i = \bar{Z} = K^{-1/2}\sigma^{-1}C^T(\widehat{\beta}-\beta_0)$, and $\sum_{i=1}^K(Z_i-\bar{Z})^2 = \sigma^{-2}KC^TV_SC = \sigma^{-2}(K-1)C^TV_aC$. The last step comes because $H_i = I_p\left\{K/(K-1)\right\}^{1/2}$ for all i. Thus, under the null hypothesis, $T_a = \sqrt{K}\bar{Z}\left\{(K-1)^{-1}\sum_{i=1}^K(Z_i-\bar{Z})^2\right\}^{-1/2}$ and is distributed t_{K-1} by the standard derivation of the t-test, and T_a^2 is distributed $F_{1,K-1}$. All proposed estimators of d use equation 5 with different estimators of Ψ_i and different values of M (for $\widehat{d}$ and $\tilde{d}$ the value H_i is set to I_p in the definition of M). Let $\dot{\Psi}_i$ ($\dot{\Psi}$) be an arbitrary estimator of Ψ_i (Ψ) and M_i be an arbitrary block diagonal element of M. Then because the $\widehat{\Omega}_i = \widehat{\phi}X^TX$ are all equal, $\text{trace}(\dot{\Psi}B_1) = \left\{(K-1)/K\right\}\sum_{i=1}^K\text{trace}(\dot{\Psi}_iM_i)$ and $\text{trace}(\dot{\Psi}B_1\dot{\Psi}B_1) = \left\{(K-2)/K\right\}\sum_{i=1}^K\text{trace}(\dot{\Psi}_iM_i\dot{\Psi}_iM_i) + K^{-2}\sum_{i=1}^K\sum_{j=1}^K\text{trace}(\dot{\Psi}_iM_i\dot{\Psi}_jM_j)$. To simplify these expressions we use the theorem that $\text{trace}(AB) = \text{trace}(BA)$ for any matrices A and B for which AB and BA are defined, so that the traces may be written as scalars. Then we can show equation 5 gives $K-1$ for both $\tilde{d}$ and $\tilde{d}_H$, while $\widehat{d}$ and $\widehat{d}_H$ may both be written as

$$\frac{R}{1+\frac{R-1}{(K-1)^2}} \qquad \text{where} \qquad R = \frac{\left\{\sum_{i=1}^K(Z_i-\bar{Z})^2\right\}^2}{\sum_{i=1}^K(Z_i-\bar{Z})^4},$$

and the Z_i are the same independent standard normal random variables as in the previous expression for T_a.

C Notation for Cox Proportional Hazards Model

We use standard notation for counting processes, so that the notation Y_i, Z_i, X_i, and δ_i is different in this section than in the rest of the paper. Let $X_1, \ldots, X_K$ be the time until failure or right-censoring, with $\delta_i = 1$ when X_i represents a failure and $\delta_i = 0$ otherwise. Let $Z_i(t)$ be the covariate vector for the ith observation observed at time t. Then $U_i^{\bullet} = \delta_i\left\{Z_i(X_i) - \bar{Z}(\beta, X_i)\right\}$ where

$\bar{Z}(\beta, t) = \left[\sum_{i=1}^{K} Y_i(t) Z_i(t) \exp\left\{\beta' Z_i(t)\right\}\right] / \left[\sum_{j=1}^{K} Y_j(t) \exp\left\{\beta' Z_j(t)\right\}\right]$ and here $Y_i(t) = 1$ if $X_i \geq t$ and 0 otherwise. Further,

$$U_i^{\circ} = -\sum_{j=1}^{K} \left[\delta_j Y_i(X_j) \exp\left\{\beta' Z_i(X_j)\right\} \left\{Z_i(X_j) - \bar{Z}(\beta, X_j)\right\}\right] / \left[\sum_{h=1}^{K} Y_h(X_j) \exp\left\{\beta' Z_h(X_j)\right\}\right], \text{and}$$

$$\frac{-\partial U_i^{\bullet}}{\partial \beta} = \delta_i \left(\left[\sum_{j=1}^{K} Y_j(X_i) \exp\left\{\beta' Z_j(X_i)\right\} \left\{Z_j(X_i) - \bar{Z}(\beta, X_i)\right\}^{\otimes 2}\right] / \left[\sum_{h=1}^{K} Y_h(X_i) \exp\left\{\beta' Z_h(X_i)\right\}\right] \right),$$

where $A^{\otimes 2} = AA^T$, for any vector, A.

Table 1: Simulations for Poisson GEE: Proportion Rejected at $\alpha = .05$ level (average length of Confidence Intervals) for 1000 simulations

			One Treatment		Both Treatments	
		Test	Per Cluster		for each Cluster	
		Distribution	R_i =Exchangeable		R_i =Independence	
Test	Variance	(df)	$\tau = 0^*$	$\tau = 1/2$	$\tau = 0$	$\tau = 1/2$
δ_m	V_m	$\chi^2(1)$	.090 (.079)	.091 (.424)	.043 (.060)	.330 (.115)
δ_1	V_s	$\chi^2(1)$	.103 (.079)	.089 (.421)	.075 (.058)	.074 (.227)
δ_2	V_s	$F(1, \widehat{d})$	.052 (.097)	.040 (.555)	.028 (.071)	.029 (.325)
δ_3	V_s	$F(1, \tilde{d})$	.066 (.087)	.067 (.451)	.053 (.062)	.055 (.244)
δ_4	V_a	$F(1, \widehat{d}_H)$	.042 (.101)	.039 (.570)	.022 (.073)	.024 (.335)
δ_5	V_a	$F(1, \tilde{d}_H)$	.059 (.091)	.064 (.463)	.046 (.064)	.048 (.251)
δ_6	V_{MD}	$F(1, K-p)$	.041 (.098)	.047 (.501)	.043 (.066)	.041 (.258)

* For this column, 991 out of 1000 converged (see http://srab.cancer.gov/sandwich for a non-converging data set); results relate to those 991 simulations. All other columns had 1000 converged simulations.

Data are simulated by $Y_{ij} \sim$ *Poisson*$(\mu_{ij}), i = 1, \ldots, 20; j = 1, \ldots, n_{i1} + n_{i2}$ *where* $\mu_{ij} = \exp(\log(10) + x_{ij} b_{ij})$, $x_{ij} = 1$ *for* $j \leq n_{i1}$, $x_{ij} = -1$ *for* $j > n_{i1}$, *and* b_{ij} *are independent pseudo-normal random variates with mean 0 and standard deviation,* τ. *All models of means are* $\mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij})$. *We test whether* $\beta_1 = 0$ *and the average length of confidence intervals are for* β_1 *only. "Both treatments for each cluster" is* $n_{ia} =$ *ceiling*(N_{ia}), *for* $a = 1, 2$, *where* N_{ia} *is distributed pseudo-Gamma with mean 10 and variance 20, and ceiling*(X) *gives the smallest integer greater than or equal to* X. *"One treatment per cluster" uses the same method for generating the* n_{ia} *then sets* $n_{i2} = 0$ *for* $i \leq 10$ *and sets* $n_{i1} = 0$ *for* $i > 10$.

Table 2: Simulations for Logistic Independence Estimating Equations derived from GUIDE data: Proportion Rejected at $\alpha = .05$ level (average length of Confidence Intervals) for 1000 simulations

		Test	Model	Model	Model	Model	Model
		Distn	without	without	without	without	without
Test	Var	(df)	FEMALE	AGE	DAYACC	SEVERE	TOILET
δ_m	V_m	$\chi^2(1)$	.048 (2.58)	.052 (2.38)	.049 (.258)	.051 (1.42)	.055 (.350)
δ_1	V_s	$\chi^2(1)$	.078 (2.41)	.067 (2.32)	.076 (.242)	.062 (1.36)	.080 (.330)
δ_2	V_s	$F(1, \widehat{d})$	.038 (3.06)	.039 (2.68)	.044 (.286)	.029 (1.62)	.040 (.405)
δ_3	V_s	$F(1, \tilde{d})$	.051 (2.70)	.057 (2.46)	.058 (.269)	.049 (1.46)	.056 (.375)
δ_4	V_a	$F(1, \widehat{d}_H)$	.018 (3.47)	.018 (3.06)	.031 (.329)	.010 (1.87)	.016 (.636)
δ_5	V_a	$F(1, \tilde{d}_H)$	.026 (3.02)	.029 (2.74)	.037 (.296)	.017 (1.69)	.028 (.551)
δ_6	V_{MD}	$F(1, K-p)$	.042 (2.87)	.039 (2.63)	.040 (.289)	.033 (1.56)	.042 (.408)

Table 3: Comparison of Simulated Variance Estimators

	Sample Variance of $\widehat{\beta}_i$	Mean of V_m	Mean of V_s	Mean of V_{MD}	Mean of V_a
	Poisson GEE (Table 1)				
R_i=Exchangeable, $\tau = 0^*$	0.00051	0.00043	0.00042	**0.00056**	0.00045
R_i=Exchangeable, $\tau = 1/2$	0.01410	0.01237	0.01201	**0.01482**	0.01264
R_i=Independence, $\tau = 0$	0.00023	0.00024	**0.00023**	0.00025	0.00024
R_i=Independence, $\tau = 1/2$	0.00405	0.00087	0.00376	0.00424	**0.00398**
	Logistic Independence Estimating Equations (Table 2)				
Model without FEMALE	0.47065	0.43737	0.39234	0.51774	**0.47954**
Model without AGE	0.40968	0.37511	0.35659	**0.42700**	0.44090
Model without DAYACC	0.00466	0.00434	0.00390	0.00521	**0.00455**
Model without SEVERE	0.12409	0.13186	**0.12274**	0.15042	0.16223
Model without TOILET	0.00898	**0.00806**	0.00732	0.01050	0.01101

* For this row, 991 out of 1000 converged; statistics calculated from only those 991 simulations. All other rows had 1000 converged simulations.

The bold value in each row represents the closest mean to the sample variance of $\widehat{\beta}_i$.

Table 4: Simulations for Conditional Logistic Regression: Proportion Rejected at $\alpha = .05$ level (average length of Confidence Intervals) for 1000 simulations

Test	Var	Test Distn (df)	Design A Case 1	Design A Case 4	Design D Case 1	Design D Case 4
δ_m	V_m	$\chi^2(1)$	.049 (.577)	.430 (.601)	.027 (2.60)	.345 (2.80)
δ_1	V_s	$\chi^2(1)$	.098 (.534)	.090 (1.36)	.252 (1.85)	.253 (4.02)
δ_2	V_s	$F(1, \widehat{d})$	.042 (.681)	.045 (1.74)	.044 (5.35)	.054 (12.00)
δ_3	V_s	$F(1, \tilde{d})$	.072 (.590)	.065 (1.51)	.047 (4.45)	.038 (10.19)
δ_4	V_a	$F(1, \widehat{d}_H)$	.033 (.708)	.038 (1.82)	.025 (6.92)	.030 (16.33)
δ_5	V_a	$F(1, \tilde{d}_H)$	.066 (.613)	.056 (1.57)	.021 (5.77)	.019 (13.79)

All simulations test whether $\beta_1 = 0$, and average confidence interval lengths are for β_1 only. In Design A β_1 is estimated from 15 clusters, while in Design D β_1 is estimated from 3 clusters. For Case 1 the effects for β_1 are fixed, while for Case 4 these effects are random.

Table 5: Simulations for Cox Regression: Proportion Rejected at $\alpha = .05$ level (average length of Confidence Intervals) for 1000 simulations

	Proportional Hazards		Model 10		Model 11	
Test	$K = 20$	$K = 50$	$K = 20$	$K = 50$	$K = 20$	$K = 50$
δ_m	.051 (1.13)	.042 (.614)	.180 (1.71)	.195 (1.02)	.104 (1.13)	.073 (.614)
δ_1	.103 (0.99)	.063 (.567)	.119 (2.26)	.079 (1.47)	.137 (1.01)	.091 (.593)
δ_2	.055 (1.33)	.037 (.672)	.047 (3.29)	.049 (1.67)	.076 (1.38)	.049 (.729)
δ_3	.051 (1.25)	.043 (.625)	.080 (2.89)	.063 (1.59)	.089 (1.28)	.066 (.656)
δ_4	.038 (1.48)	.032 (.715)	.039 (3.69)	.044 (1.72)	.059 (1.55)	.039 (.780)
δ_5	.043 (1.37)	.037 (.652)	.068 (3.13)	.059 (1.63)	.073 (1.40)	.054 (.686)

All simulations modeled the hazard proportional to $\exp(\beta_1 z_1 + \beta_2 z_2)$ *and tested whether* $\beta_1 = 0$ *or not. True models with* t *=time to event: (proportional hazards)* $t = \exp(z_2)\epsilon_2$*; (model 10)* $t = \exp(-.5z_2 - z_1^2 + .5\epsilon_1)$*; (model 11)* $t = \exp(-.5z_2) + .5\epsilon_1$*; where* z_1, z_2*, and* ϵ_1 *are independent standard normal pseudo-random variables (p-r.v.), with* z_1 *and* z_2 *truncated at* ± 5*, and* ϵ_2 *is a standard exponential p-r.v.*