A Note on the Efficiency of Sandwich Covariance Matrix Estimation

Göran KAUERMANN and Raymond J. CARROLL

The sandwich estimator, also known as robust covariance matrix estimator, heteroscedasticity-consistent covariance matrix estimate, or empirical covariance matrix estimator, has achieved increasing use in the econometric literature as well as with the growing popularity of generalized estimating equations. Its virtue is that it provides consistent estimates of the covariance matrix for parameter estimates even when the fitted parametric model fails to hold or is not even specified. Surprisingly though, there has been little discussion of properties of the sandwich method other than consistency. We investigate the sandwich estimator in quasi-likelihood models asymptotically, and in the linear case analytically. We show that under certain circumstances when the quasi-likelihood model is correct, the sandwich estimate is often far more variable than the usual parametric variance estimate. The increased variance is a fixed feature of the method and the price that one pays to obtain consistency even when the parametric model fails or when there is heteroscedasticity. We show that the additional variability directly affects the coverage probability of confidence intervals constructed from sandwich variance estimates. In fact, the use of sandwich variance estimates combined with t-distribution quantiles gives confidence intervals with coverage probability falling below the nominal value. We propose an adjustment to compensate for this fact.

KEY WORDS: Coverage probability; Generalized estimating equation; Generalized linear model; Heteroscedasticity; Linear regression; Marginal model; Quasi-likelihood; Robust covariance estimator; Sandwich estimator.

1. INTRODUCTION

The *heteroscedasticity-consistent covariance matrix estimator* is a common tool used for variance estimation of parameter estimates. Originally introduced by Huber (1967), Eicker (1967), and White (1980), the estimate has become popular in the econometric literature. In the last decade, the method has also been widely used in the context of generalized estimating equations (see, e.g., Diggle, Liang, and Zeger 1994; Liang and Zeger 1986; Liang, Zeger and Qaqish 1992, where it was introduced as the *sandwich variance estimator*. Whereas in econometric models the estimate is used to cope for heteroscedastic errors, in generalized estimating equations its objective is consistent variance estimation for dependent data. In the latter setting, efficient estimation of parameters requires specification of the correlation structure among the observations—which, however, typically is unknown. Therefore, a so-called working covariance matrix is used in the estimation step, which for variance estimation is combined with its corresponding empirical version in a sandwich form. This approach yields consistent estimates of the covariance matrix under misspecified working covariances as well as under heteroscedastic errors. Because of this desirable model-robustness property, the sandwich estimator is also sometimes called the *robust covariance matrix estimator* or the *empirical covariance matrix estimator*. We use the term *sandwich variance estimator* throughout the article.

The argument in favor of the sandwich estimate is that asymptotic normality and asymptotic coverage of confidence intervals require only a consistent variance estimate, so there is no direct need to construct a highly accurate covariance matrix estimate. But the consistency of the sandwich variance estimate has its price in increased variability; that is, sandwich variance estimators generally have a larger variance than model-based classical variance estimates. In his discussion of the article by Wu (1986), Efron (1986) gave simulation evidence of this phenomenon. Breslow (1990) demonstrated this in a simulation study of overdispersed Poisson regression. Firth (1992) and McCullagh (1992) both raised concerns that the sandwich estimator may be particularly inefficient. Diggle et al. (1994, p. 77) suggest that it is best used when the data come from "many experimental units." We clarify and refine these statements. An earlier discussion about small sample improvements for the sandwich estimate in the econometric literature was given by MacKinnon and White (1985), who proposed jackknife sandwich estimates. The performance of this estimate compared to other approaches was recently investigated by means of simulations by Long and Ervin (2000).

The objectives of this article are twofold. First we investigate the sandwich estimate in terms of efficiency; and second, we analyze the effect of the increased variability of the sandwich estimate on the coverage probability of confidence intervals. For efficiency, we derive asymptotic and fairly precise small-sample properties, neither of which appear to have been quantified before. For example, the sandwich method in simple linear regression when estimating the slope has an asymptotic efficiency equal to the inverse of the sample kurtosis of the design values. This inefficiency also holds for quasi-likelihood estimation and in generalized linear models. For example, in simple linear logistic regression, at the null value where there is no effect from the predictor, the sandwich method's asymptotic relative efficiency is again the inverse of the kurtosis of the predictors. In Poisson regression, the sandwich method has even less efficiency. The problem of undercoverage of confidence intervals was shown through simulation studies by Wu (1986) and by Breslow (1990), who reported somewhat

Göran Kauermann is Lecturer, Department of Statistics & Robertson Centre, University of Glasgow, University Gardens, Glasgow G12 8QW, Scotland (E-mail: *goeran@stats.gla.ac.uk*). Raymond J. Carroll is Distinguished Professor, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *carroll@stat.tamu.edu*). Carroll's research was supported by National Cancer Institute Grant CA-57030, and by the Texas A&M Center for Environmental and Rural Health National Institute of Environmental Health Sciences Grant P30-ESO9106. The authors also acknowledge the support received from the Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 386 at the Ludwig-Maximilans-Universitiät München). Parts of this article contain material in an unpublished technical report written with Douglas Simpson, Arny Stromberg, Suojin Wang, and David Ruppert, whose help is gratefully acknowledged.

elevated levels of Wald-type tests based on the sandwich estimator. Rothenberg (1988) derived an adjusted distribution function for the t-statistic calculated from sandwich variance estimates. We give a different theoretical justification for the empirical fact that confidence intervals calculated from sandwich variance estimates and t-distribution quantiles are generally too small; that is, the coverage probability falls below the nominal value. We show that undercoverage is determined mainly by the variance of the variance estimate. To correct this deficit, we present an adjustment that depends on normal distribution quantiles and the variance of the sandwich variance estimate.

The article is organized as follows. In Section 2 we compare the sandwich estimator with the usual parametric regression estimator in the linear regression model. In Section 3 we discuss the sandwich estimate for quasi-likelihood and generalized estimating equations (GEEs). We provide proofs and other general statements in the Appendix.

2. LINEAR REGRESSION

2.1 Properties of the Sandwich Estimator

First, consider the simple homoscedastic linear regression model

$$Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i \text{ with } \epsilon_i \sim N(0, \sigma^2), \tag{1}$$

where $\mathbf{x}_i^{\mathrm{T}}$ are $1 \times p$-dimensional vectors of covariates and $i = 1, \ldots, n$. Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ be the ordinary least squares estimator of $\boldsymbol{\beta}$, where $\mathbf{Y}^{\mathrm{T}} = (Y_1, \ldots Y_n)$ and $\mathbf{X}^{\mathrm{T}} = (\mathbf{x}_1, \ldots \mathbf{x}_n)$. Assume now that we are interested in inference about the linear combination $\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, where $\mathbf{z}^{\mathrm{T}}$ is a $1 \times p$-dimensional contrast vector of unit length, that is, $\mathbf{z}^{\mathrm{T}}\mathbf{z} = 1$. The variance of $\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is given by $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}$, which can be estimated by the classical model-based variance estimator $V_{\mathrm{model}} = \hat{\sigma}^2\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}$, where $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\epsilon}_i^2/(n-p)$ with $\hat{\epsilon}_i = Y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ as fitted residuals. A major assumption used implicitly in the calculation of V_{model} is that the errors ϵ_i are homoscedastic. This assumption is often not very plausible, particularly in econometric models, where one is faced with heteroscedasticity, so that the model

$$Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i \quad \text{with } \epsilon_i \sim N(0, \sigma_i^2) \tag{2}$$

holds. In this case V_{model} does not provide a consistent variance estimate. In contrast, the sandwich variance estimate,

$$V_{\mathrm{sand}} = \mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\Big(\sum_i \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\hat{\epsilon}_i^2\Big)(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z} = \sum_{i=1}^n a_i^2\hat{\epsilon}_i^2, \tag{3}$$

consistently estimates $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$, where $a_i = \mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_i$. Estimate (3) is called the *sandwich variance estimator* because of its sandwich structure, even though the terms *robust variance estimator, heteroscedasticity-consistent covariance estimator*, and *empirical covariance estimator* are more common in the econometric literature.

In linear regression, (3) is often multiplied by $n/(n-p)$ (Hinkley 1977) to reduce the bias. Let h_{ii} be the ith diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} = (h_{ij})$. Under homoscedasticity, one finds $E(\hat{\epsilon}_i^2) = \sigma^2(1-h_{ii})$, so that

$$E(V_{\mathrm{sand}}) = \sigma^2\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}(1-b_n), \tag{4}$$

where $b_n = \sum_{i=1}^n h_{ii}a_i^2/\sum_{i=1}^n a_i^2 \le \max_{1\le i\le n} h_{ii}$. Because $b_n \ge 0$, one obtains that in general the sandwich estimator is biased *downward*, as was shown by Chesher and Jewitt (1987), (see also MacKinnon and White 1985). The bias therefore depends on the design of $\mathbf{x}_i$ and can be substantial when there are leverage points. Bias problems can be avoided by replacing $\hat{\epsilon}_i$ in (3) by $\tilde{\epsilon}_i = \hat{\epsilon}_i/(1-h_{ii})^{1/2}$. The resulting estimator is referred to as the *unbiased* sandwich variance estimator and is denoted by $V_{\mathrm{sand},u}$ (Wu 1986, eq. 2.6). It is easily seen that $E(V_{\mathrm{sand},u}) = \mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$, whereas under heteroscedasticity of model (2), the estimate is still consistent but with an asymptotic of order $O(n^{-1})$. Because $\mathrm{var}(\tilde{\epsilon}_i^2) = 2\sigma^4$ and $\mathrm{cov}(\tilde{\epsilon}_i^2, \tilde{\epsilon}_j^2) = 2\tilde{h}_{ij}^2\sigma^4$ for $i \ne j$, where $\tilde{h}_{ij} = h_{ij}/\{(1-h_{ii})(1-h_{jj})\}^{1/2}$, it follows that

$$\begin{aligned}\mathrm{var}(V_{\mathrm{sand},u}) &= \sum_{i=1}^n a_i^4\mathrm{var}(\tilde{\epsilon}^2) + \sum_{i\ne j} a_i^2a_j^2\mathrm{cov}(\tilde{\epsilon}_i^2, \tilde{\epsilon}_j^2)\\ &= 2\sigma^4\Big(\sum_{i=1}^n a_i^4 + \sum_{i\ne j} a_i^2a_j^2\tilde{h}_{ij}^2\Big).\end{aligned} \tag{5}$$

We now compare the variance (5) to the variance of the model-based variance estimator $\mathbf{V}_{\mathrm{model}}$, which equals $\mathrm{var}(V_{\mathrm{model}}) \approx 2\sigma^4\{\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}\}^2/n = 2\sigma^4(\sum a_i^2)^2/n$.

Theorem 1. Under the homoscedastic linear model (1), the efficiency of the unbiased sandwich estimate $\mathbf{V}_{\mathrm{sand},u}$ compared to the classical variance estimate $\mathbf{V}_{\mathrm{model}}$ for $\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ satisfies

$$\frac{\mathrm{var}(\mathbf{V}_{\mathrm{sand},u})}{\mathrm{var}(\mathbf{V}_{\mathrm{model}})} \ge \Big\{n^{-1}\sum_{i=1}^n a_i^4\Big\}\Big\{n^{-1}\sum_{i=1}^n a_i^2\Big\}^{-2} \ge 1. \tag{6}$$

The proof follows directly from the Cauchy–Schwarz inequality. Theorem 1 states that the sandwich estimate is less efficient when the model is correct, that is, when the errors are homoscedastic. Because of the vector $\mathbf{z}$, the loss of efficiency basically depends on the design of the covariates, as the following example shows.

Example 1 (The Intercept and the Slope in Simple Linear Regression). Assume that $\mathbf{x}_i^{\mathrm{T}} = (1, u_i)$, where $\sum u_i = 0$. Suppose that we are interested in the intercept, that is, $\mathbf{z}^{\mathrm{T}} = (1, 0)$. We then have $a_i = n^{-1}$, and the asymptotic relative efficiency in (6) is 1. Suppose now that $\mathbf{z} = (0, 1)$, so that $\hat{\beta}_1 = \mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is the slope estimate. Assuming $\max(|u_i|) = o(n^{1/4})$ for technical purposes, the asymptotic relative efficiency is κ_n^{-1}, where $\kappa_n = n^{-1}\sum u_i^4/(n^{-1}\sum_{i=1}^n u_i^2)^2 \ge 1$. Note that κ_n is the sample kurtosis of the design points u_i. For instance, if the design points $(u_1, \ldots, u_n)$ are realizations of a normal distribution, then $\kappa_n \to 3$, and hence the sandwich estimator $V_{\mathrm{sand},u}$ has three times the variability of the usual model-based estimator V_{model}. If the design points are generated from a Laplace distribution, then the usual sandwich estimator is six times more variable.

The foregoing example shows that using sandwich variance estimates in linear models can lead to a substantial loss of efficiency. A similar phenomena occurs for quasi-likelihood estimation, as discussed in the next section. Note that Theorem 1 is formulated under the assumption of homoscedasticity. Even though it might be interesting to weaken this assumption and analyze the efficiency of $V_{\text{sand},u}$ under heteroscedasticity theoretically, one should keep in mind that V_{model} is not a consistent estimate under heteroscedasticity, so its bias also must be taken into account. Instead, we investigate the behavior of the estimate under heteroscedastic errors empirically in the simulations studies of the following subsection.

2.2 Coverage Probability of Confidence Intervals

In this section we investigate how the additional variability affects the coverage probability of confidence intervals obtained from sandwich variance estimates. As one would expect, the excess variability of the sandwich estimate is directly reflected in undercoverage of confidence intervals. Let $\theta = \mathbf{z}^T\boldsymbol{\beta}$ be the unknown parameter of interest with $\hat{\theta} = \mathbf{z}^T\hat{\boldsymbol{\beta}} \sim N(\theta, \sigma^2/n)$ an unbiased estimate of θ based on a random sample of size n. The symmetric $1-\alpha$ confidence interval is given by $CI(\sigma^2, \alpha) := [\hat{\theta} \pm z_p\sigma/\sqrt{n}]$, where z_p is the $p = 1-\alpha/2$ quantile of the standard normal distribution. If σ^2 is estimated by an unbiased variance estimate $\hat{\sigma}^2$, it is well known that the confidence interval $CI(\hat{\sigma}^2, \alpha)$ shows undercoverage, and typically t-distribution quantiles are used instead of normal quantiles. The following theorem shows explicitly how the variance of $\hat{\sigma}^2$ affects the undercoverage.

Theorem 2. Let $\hat{\theta} \sim N(\theta, \sigma^2/n)$ and let $\hat{\sigma}^2$ be an unbiased estimate of σ^2 independent of $\hat{\theta}$. The coverage probability of the $1-\alpha$ confidence interval $CI(\hat{\sigma}^2, \alpha)$ equals

$$\Pr\{\theta \in CI(\hat{\sigma}^2, \alpha)\} = 1 - \alpha - c_p \frac{\operatorname{var}(\hat{\sigma}^2)}{\sigma^4} + O(n^{-2}), \qquad (7)$$

where $c_p = \phi(z_p)(z_p^3 + z_p)/8$, with $\phi(\cdot)$ the standard normal distribution density.

The proof of Theorem 2 is given in the Appendix. Note that the assumption of independence of $\hat{\sigma}^2$ and $\hat{\theta} - \theta$ holds in a normal homoscedastic regression model if $\hat{\sigma}^2$ is calculated from fitted residuals; that is, it holds for sandwich variance estimates. Because $c_p > 0$ (for $p > 1/2$), undercoverage becomes obvious. In particular, the undercoverage increases linearly with the variance of the variance estimate $\hat{\sigma}^2$. Using the results of Theorem 1, we therefore conclude that confidence intervals based on sandwich variance estimators have lower coverage probability than confidence intervals based on model-based variance estimates. This also implies that t-distribution quantiles do not correct the undercoverage. The result stated in Theorem 2 resembles that given by Rothenberg (1988, p. 1005). He derived an adjustment for the distribution function of the t-statistic based on sandwich variance estimates. In contrast to Rothenberg, however, Theorem 2 points out the distinct role of the variance of $\hat{\sigma}^2$.

Coverage Adjustment. In normal linear regression models, formula (7) can be used directly to construct a coverage correction for confidence intervals. Instead of using quantile z_p, we suggest choosing $\tilde{p} > p$ and make use of the quantile $z_{\tilde{p}}$. The increased $\tilde{p}$ is then selected such that $\Pr(\theta \in [\hat{\theta} \pm z_{\tilde{p}}\hat{\sigma}/\sqrt{n}]) = p$ holds; that is, with (7), $\tilde{p}$ is defined as the numerical solution to

$$p = \tilde{p} - \phi(z_{\tilde{p}})\operatorname{var}(\hat{\sigma}^2)\frac{z_{\tilde{p}}^3 + z_{\tilde{p}}}{8\sigma^4}. \qquad (8)$$

Example 2 (t-Distribution Quantiles). Before applying the correction to the sandwich variance estimate, we demonstrate the use of (8) in a setting where an exact solution is available. Let the random sample $Y_i \sim N(\mu, \sigma^2)$ be drawn from an univariate normal distribution. The centered mean estimate $n^{1/2}(\hat{\mu} - \mu)$ is distributed as a normal $(0, \sigma^2)$, with $\hat{\mu} = \sum_i^n Y_i/n$ and variance σ^2 estimated by $\hat{\sigma}^2 = \sum_i^n (Y_i - \hat{\mu})^2/(n-1)$. Exact quantiles for confidence intervals based on the estimates $\hat{\mu}$ and $\hat{\sigma}^2$ are available from t-distribution quantiles with $n-1$ degrees of freedom. Approximative quantiles $z_{\tilde{p}}$ follow from solving (8) using $\operatorname{var}(\hat{\sigma}^2) = 2\sigma^4/(n-1)$. It is a special feature of the normal distribution that the unknown variance component σ^4 in (8) cancels out and is not required for the calculation of $z_{\tilde{p}}$. In Table 1 we compare the exact quantiles based on a t-distribution with the corrected versions based on (8). Even for small sample sizes, the corrected quantiles $z_{\tilde{p}}$ are distinctly close to the exact t-distribution quantiles. This is also seen in the true one-sided coverage probability $\Pr(\hat{\theta} \le \theta + z_{\tilde{p}}\hat{\sigma}/\sqrt{n})$ of the confidence intervals, and demonstrates that the adjustment applied in a standard setting behaves quite well.

Example 3 (Sandwich Variance Estimate). We now apply the corrected quantile $z_{\tilde{p}}$ to confidence intervals based on sandwich variance estimates. Inserting (5) in (8) shows again that the variance component σ^4 cancels out, so that the correction depends exclusively on the design of the covariates. We ran a small simulation study to demonstrate the behavior of the correction. Let $Y_i = \beta_0 + x_i\beta_x + \varepsilon_i$ with $\beta_0 = 0$, $\beta_x = 1$, and $\varepsilon_i \sim N(0, \sigma_i^2)$. The errors are drawn from the homoscedastic model (1), that is, σ_i constant with value .2 (model 1), as well as from the heteroscedastic model (2) with $\sigma_i = .2 + \exp(x_i/2)/2$ (model 2) and $\sigma_i = \sqrt{(.1 + x_i^2)}$ (model 3). The covariates x_i are chosen to be (a) uniformly, (b)

Table 1. Comparison of Coverage Probability Based on $z_{\tilde{p}}$ and t-Distribution Quantiles $t_{p,n-1}$ for $n-1$ Degrees of Freedom

p	$t_{p,n-1}$	$z_{\tilde{p}}$	$P(\hat{\theta} \le \theta + z_{\tilde{p}}\hat{\sigma}/\sqrt{n})$
$n = 5$			
.90	1.533	1.551	.902
.95	2.132	2.095	.948
.975	2.776	2.543	.968
$n = 15$			
.90	1.345	1.346	.900
.95	1.761	1.761	.950
.975	2.145	2.137	.975

Table 2. Corrected Quantiles $z_{\tilde{p}}$ for Different Designs

Design	$p = .90$		$p = .95$	
	$t_{p,n-2}$	$z_{\tilde{p}}$	$t_{p,n-2}$	$z_{\tilde{p}}$
		$n = 20$		
(a)		1.81		2.21
(b)	1.73	1.86	2.10	2.27
(c)		1.94		2.36
		$n = 40$		
(a)		1.72		2.07
(b)	1.68	1.75	2.02	2.12
(c)		1.80		2.19

normally, and (c) Laplace distributed. The corrected quantiles, $z_{\tilde{p}}$, are listed in Table 2. The results shown in Figure 1 give the empirical coverage probability based on 2000 replicates with $n = 20$ (upper row) and $n = 40$ (bottom row) observations. For comparison, we also show the coverage probability for confidence intervals calculated from $\mathbf{V}_{\text{sand},u}$ and t-distribution quantiles. Moreover, we calculate confidence intervals based on the jackknife estimate as suggested by MacKinnon and White (1985, form. 13). This has the form

$$\mathbf{V}_{\text{jack}} = \frac{n-1}{n}\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\left(\sum_i \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i\hat{\epsilon}_i^{*2}\right) \times (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z} - \frac{n-1}{n^2}\hat{\gamma}^{\mathrm{T}}\hat{\gamma}, \quad (9)$$

where $\hat{\gamma} = \mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\hat{\epsilon}^*$ and $\hat{\epsilon}_i^* = \hat{\epsilon}_i(1 - h_{ii})$.

It appears that uncorrected intervals clearly suffer from undercoverage. This is corrected to a large extent by both the corrected quantiles and the jackknife estimate. In general, however, the corrected quantiles behave slightly better in all three models, for heteroscedastic as well as homoscedastic errors.

The foregoing examples show the benefits of correction (8). For practical purposes, it might be cumbersome to solve (8) explicitly, however. Instead, an approximate solution of (8) based on the relative efficiency $\text{var}(\mathbf{V}_{\text{sand},u})/\text{var}(\mathbf{V}_{\text{model}})$ given in (6) in Theorem 1 can be used. As shown in the Appendix, one easily gets

$$z_{\tilde{p}} = t_{p,n-p} + d_{p,n-p}\left(\frac{\text{var}(\mathbf{V}_{\text{sand},u})}{\text{var}(\mathbf{V}_{\text{model}})} - 1\right) + O(n^{-2}), \quad (10)$$

where t_p is the t-distribution quantile with $n - p$ degrees of freedom and $d_{p,n-p} = \text{var}(\mathbf{V}_{\text{model}})(z_p^3 + z_p)/(8\sigma^4)$. As before, the variance term σ^4 cancels out when $\text{var}(\mathbf{V}_{\text{model}})$ is inserted. Formula (10) shows that the corrected quantiles depend linearly on the relative efficiency. The slope parameter is thereby decreasing with increasing sample size. Relation (10) is visualized in Figure 2, where we plot $z_{\tilde{p}}$ against the relative efficiency $\text{var}(\mathbf{V}_{\text{sand},u})/\text{var}(\mathbf{V}_{\text{model}})$ for $p = .95$ and $p = .975$. The linear shape is obvious, and it appears that the correction is substantial if the relative efficiency is large and/or the sample size is small. Figure 2 and (10) can also be used to provide confidence intervals with appropriate coverage probability by calculating the relative efficiency.

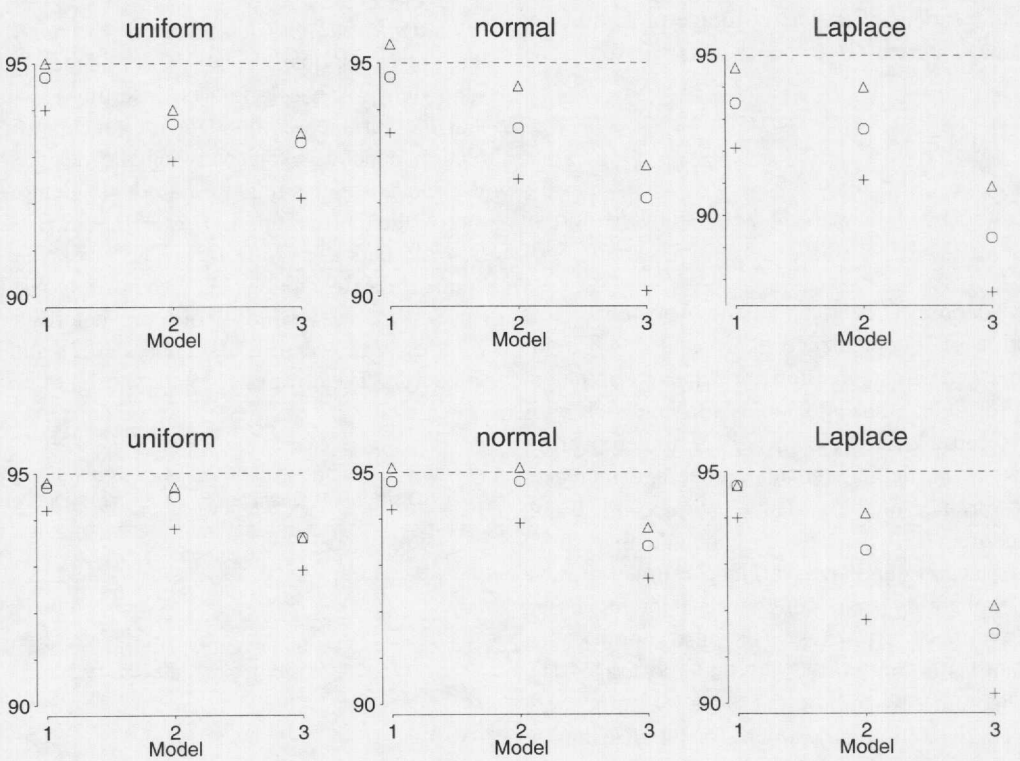

Figure 1. Coverage Probability of Confidence Intervals Based on $\mathbf{V}_{\text{sand},u}$ With Corrected Quantiles $z_{\tilde{p}}$ (△) as Well as With t-Distribution Quantiles $t_{p,n-1}$ (+) and Based on the Jackknife Estimate $\mathbf{V}_{\text{jack}}$ (○). The upper row is for $n = 20$; the bottom row, for $n = 40$.

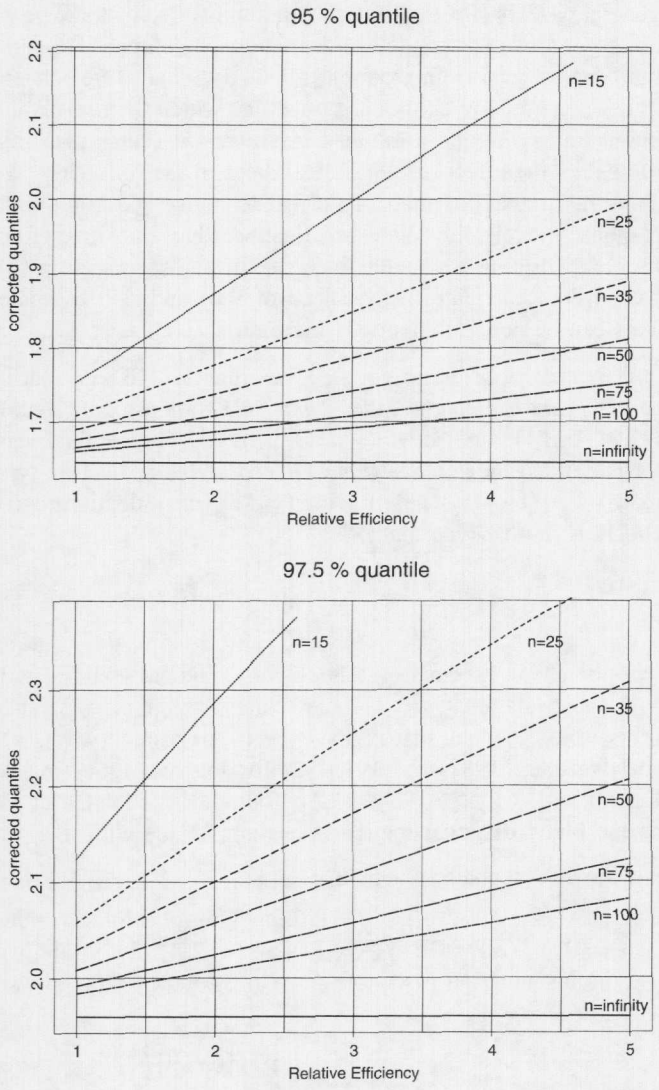

Figure 2. Corrected Quantiles $z_{\tilde{p}}$ in Dependence of Sample Size n and Relative Efficiency $\text{var}(\mathbf{V}_{\text{sand},u})/\text{var}(\mathbf{V}_{\text{model}})$.

3. QUASI-LIKELIHOOD AND GENERALIZED ESTIMATING EQUATIONS

3.1 Properties of the Sandwich Estimate

In this section we consider the sandwich variance estimate for quasi-likelihood estimates from GEEs. Let $Y_i = (Y_{i1}, \ldots, Y_{im})^{\mathrm{T}}$, be a random vector taken at the ith unit for $i = 1, \ldots, n$ and $m \geq 1$. For $m > 1$, the components of Y_i are allowed to be correlated while observations taken at two different units are independent. Although in principle the number of observations per unit may vary from unit to unit, for ease of notation we take m as constant here. The case $m = 1$ is of course a special case in the formulas. The mean of Y_i given the $m \times p$-dimensional design matrix $\mathbf{X}_i^{\mathrm{T}}$ is given by the generalized linear model $E(Y_i|\mathbf{X}_i) = h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$, where $h(\cdot)$ is an invertible m-dimensional link function. Efficient estimation of $\boldsymbol{\beta}$ requires knowledge of the covariance matrix of Y_i. This is typically unknown, and thus one specifies $\sigma^2 V(\mu_i) =: \sigma^2 \mathbf{V}_i$ as the so-called working covariance matrix, where $\mu_i = h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$, $V(\cdot)$ is a specified covariance variance function, and σ^2 is a dispersion scalar that is either unknown (e.g., for normal response) or a known constant, (e.g., $\sigma^2 \equiv 1$ for Poisson data). Models of this type are also called marginal models (see Diggle et al. 1994 and references therein). If Y_i is a scalar, (i.e., if $m = 1$), models of this type are better known as quasi-likelihood models (Wedderburn 1974) or generalized linear models (McCullagh and Nelder 1989). The parameter $\boldsymbol{\beta}$ can be estimated using the GEE (e.g., Gourieroux, Monfort, and Trognon 1984; Liang and Zeger 1986)

$$0 = \sum_i \frac{\partial \mu_i^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(Y_i - \mu_i). \tag{11}$$

In the previous section, we were able to perform exact calculations. In quasi-likelihood models, such exact calculations are not feasible, and asymptotics are required. We do not write down formal regularity conditions, but essentially what is necessary is that sufficient moments of the components of $\mathbf{X}$ and Y exist. We also require sufficient smoothness of $h(\cdot)$. Under such conditions, a Taylor expansion of (11) about the true parameter $\boldsymbol{\beta}$ provides the first-order approximation

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{\Omega}^{-1} \sum_i \frac{\partial \mu_i^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(Y_i - \mu_i) + O_p(n^{-1}), \tag{12}$$

where $\boldsymbol{\Omega} = \sum_i \partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta}) \mathbf{V}_i^{-1} \partial \mu_i/(\partial \boldsymbol{\beta})$. Assume that we are interested in inference about $\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}$. If $\mathbf{V}_i$ is correctly specified, that is, if $\sigma^2 \mathbf{V}_i = \text{var}(Y_i|\mathbf{X}_i)$, then one gets $\text{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{z}\sigma^2$ to a first-order approximation. Hence we can estimate $\text{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$ by $\mathbf{V}_{\text{model}} := \hat{\sigma}^2 \mathbf{z}^{\mathrm{T}} \widehat{\boldsymbol{\Omega}}^{-1} \mathbf{z}$, where $\widehat{\boldsymbol{\Omega}}$ is a simple plug-in estimate of $\boldsymbol{\Omega}$ and $\hat{\sigma}^2$ is an estimate of the dispersion parameter if this is unknown. But in practice the covariance matrix may not be known—that is, $\mathbf{V}_i$ in (11) can be misspecified—which means that $\sigma^2 \mathbf{V}_i \neq \text{var}(Y_i|\mathbf{X}_i)$ holds. In this case the variance $\text{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$ can be estimated consistently by the sandwich formula

$$\mathbf{V}_{\text{sand}} = \mathbf{z}^{\mathrm{T}} \widehat{\boldsymbol{\Omega}}^{-1} \left(\sum_i \frac{\partial \hat{\mu}_i^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \widehat{\mathbf{V}}_i^{-1} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^{\mathrm{T}} \widehat{\mathbf{V}}_i^{-1} \frac{\partial \hat{\mu}_i}{\partial \boldsymbol{\beta}} \right) \widehat{\boldsymbol{\Omega}}^{-1} \mathbf{z}, \tag{13}$$

where $\hat{\boldsymbol{\epsilon}}_i = Y_i - \hat{\mu}_i = Y_i - h(\mathbf{X}_i\hat{\boldsymbol{\beta}})$ are the fitted residuals and the hat notation refers to simple plug-in estimates. The fitted residuals can be expanded as $\hat{\boldsymbol{\epsilon}}_i = \boldsymbol{\epsilon}_i - \partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\{1 + O_p(n^{-1/2})\}$, and assuming for the moment that $\mathbf{V}_i$ correctly specifies the covariance, that is, $E(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^{\mathrm{T}}) = \sigma^2 \mathbf{V}_i$, one finds via (12) that $E(\hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^{\mathrm{T}}) = \sigma^2 \mathbf{V}_i - \sigma^2 \partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}}) \boldsymbol{\Omega}^{-1} \partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta})\{1 + O(n^{-1})\}$. Because $\partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}}) \boldsymbol{\Omega}^{-1} \partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta})$ is positive definite, the sandwich estimate $\mathbf{V}_{\text{sand}}$ appears to be biased downward with order $O(n^{-1})$, and, as in the previous section, the bias can be corrected. Thus let $\tilde{\boldsymbol{\epsilon}}_i = (\mathbf{I} - \mathbf{H}_{ii})^{-\frac{1}{2}} \hat{\boldsymbol{\epsilon}}_i$ define the leverage-adjusted residuals with $\mathbf{I}$ as identity matrix and $\mathbf{H}_{ii} = \partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}}) \boldsymbol{\Omega}^{-1} \partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta}) \mathbf{V}_i^{-1}$. Replacing $\hat{\boldsymbol{\epsilon}}$ in (13) by $\tilde{\boldsymbol{\epsilon}}$ gives the bias-reduced sandwich estimate $\mathbf{V}_{\text{sand},u}$ that satisfies $E(\mathbf{V}_{\text{sand},u}) = \text{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})\{1 + O(n^{-2})\}$, assuming that the variance is correctly specified. If in contrast the variance is not correctly specified, that is, if $\mathbf{V}_i \sigma^2 \neq E(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^{T})$ holds, then the first-order asymptotic bias remains, so that $E(\mathbf{V}_{\text{sand},u}) = \text{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})\{1 + O(n^{-1})\}$. This means that the first-order bias reduction holds only if the variance is known.

In practice, however, it seems to be a plausible strategy to work with $\mathbf{V}_{\text{sand},u}$ instead of $\mathbf{V}_{\text{sand}}$, even if $\mathbf{V}_i$ is a working covariance and the true variance structure is unknown.

3.2 Examples

Theorem A.1 in Appendix Section A.2 extends Theorem 1 to the quasi-likelihood setting. The formulation and presentation of the result is cumbersome, however, and thus is deferred to the Appendix. The major reason for the additional complications is that in quasi-likelihood equations and GEEs, variance estimates have two different sources of stochastic variation. The first source is estimation of the dispersion parameter σ^2, if this is unknown; the second is the use of plug-in estimates, which are used if the variance function $V(\mu)$ depends on the mean. We demonstrate the loss of efficiency from the second source with Poisson and binomial data where the dispersion parameter is known. The variability of the model-based variance estimate occurs here solely from plug-in estimation.

Example 4 (Poisson Log-Linear Regression). We consider the univariate model $E(Y_i|\mathbf{x}) = \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ where $\mathbf{x}_i = (1, u_i)$ with u_i scalar, $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$, and Y_i Poisson distributed. The slope β_1 is the parameter of interest, and we investigate the null case $\boldsymbol{\beta} = (1, 0)^{\mathrm{T}}$. Then, as seen in the Appendix, if u has a symmetric distribution, then in limit as $n \to \infty$, $\text{var}(\mathbf{V}_{\text{sand}})/\text{var}(\mathbf{V}_{\text{model}}) = \kappa_n\{1 + 2\exp(\beta_0)\}$, where $\kappa_n = n^{-1}\sum_i u_i^4/(n^{-1}\sum_i u_i^2)^2$ is the sample kurtosis as in Example 3. The additional variability in the Poisson case is somewhat surprising—namely, that as the background event rate $\exp(\beta_0)$ increases, at the null case the sandwich estimator has efficiency decreasing to 0.

Example 5 (Logistic Regression). Let Y_i be binary with $E(Y_i|\mathbf{x}_i) = \text{logit}^{-1}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ with $\mathbf{x}_i$ as described before. Again, the slope β_1 is the parameter of interest. We vary β_1 while choosing β_0 so that marginally $E(y|\mathbf{x}) = .10$. With $\beta_1 = 0$, .5, 1.0, and 1.5, the asymptotic relative efficiency $\text{var}(\mathbf{V}_{\text{sand}})/\text{var}(\mathbf{V}_{\text{model}})$ varies for u_i standard normally distributed as 3.00, 2.59, 1.92, and 1.62. When u_i comes from a Laplace distribution (with unit variance), the corresponding efficiencies are 6.00, 4.36, 3.31, and 2.57. Note that in both cases, at the null case $\beta_1 = 0$ the efficiency of the sandwich estimator is exactly the same as the linear regression problem. This is no numerical fluke, and in fact can be shown to hold generally when u has a symmetric distribution.

The previous two examples show that the loss of efficiency of the sandwich variance estimate in nonnormal models differs from and can be worse than that occurring in normal models.

3.3 Coverage Probability

Undercoverage as pointed out in (7) of Theorem 2 extends asymptotically to quasi-likelihood or generalized linear models. For multivariate normal response models with correctly specified covariance matrices $\mathbf{V}_i$, $i = 1, \ldots, n$, Theorem 2 still holds exactly because variance estimates are independent of parameter estimates. But even if covariance matrices are misspecified, correction (8) derived from Theorem 2 can provide improved coverage, as demonstrated in our simulations. In the general case, however, exact calculation of the coverage probability and of the variance of the sandwich estimate are cumbersome, as seen from Examples 4 and 5. Because we concentrate on symmetric confidence intervals, which themselves are based on asymptotic normality arguments, it seems plausible to neglect the effect of plug-in estimates in the following. We show in simulations that for normal response and nonnormal response models, the coverage adjustment has a positive effect by compensating for undercoverage. To apply correction (8), we have to calculate the variance of the sandwich estimate. This can be done efficiently using matrix algebra.

Calculation of $\text{var}(\mathbf{V}_{\text{sand},u})$. We rewrite (13) in matrix form. Let $\mathbf{Y}$ denote the $(mn) \times 1$-dimensional vector $(Y_1^{\mathrm{T}}, \ldots, Y_n^{\mathrm{T}})^{\mathrm{T}}$ and set $\boldsymbol{\mu} = (\mu_1^{\mathrm{T}}, \ldots, \mu_n^{\mathrm{T}})^{\mathrm{T}}$. The residual vector is defined by $\boldsymbol{\epsilon} = \mathbf{Y} - \boldsymbol{\mu}$. Let $\mathbf{P}$ denote the projection-type matrix $\mathbf{P} = (\mathbf{I} - \mathbf{H})$, where $\mathbf{I}$ is the $(nm) \times (nm)$ identity matrix and $\mathbf{H}$ is the hat-type matrix

$$\mathbf{H} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \boldsymbol{\Omega}^{-1} \frac{\partial \boldsymbol{\mu}^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \text{diag}_m(\mathbf{V}_i^{-1}),$$

with $\text{diag}_m(\mathbf{V}_i^{-1})$ denoting the block diagonal matrix with $\mathbf{V}_i^{-1}$ on its diagonal, $i = 1, \ldots n$. Note that for $m \equiv 1$, other versions of the hat matrix have been suggested (see Cook and Weisberg 1982, pp. 191–192, for logistic regression or Carroll and Ruppert 1988, p. 74, for other models). Let $\mathbf{W}$ be the block diagonal matrix $\mathbf{W} = \text{diag}_m(\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i)$ with $\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i$ on the block diagonals, where $\mathbf{a}_i = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\frac{\partial \mu_i^{\mathrm{T}}}{\partial \boldsymbol{\beta}} V_i^{-1}$. With $\mathbf{D} = \text{diag}_m(\mathbf{I} - \mathbf{H}_{ii})^{-1/2}$, we get the leverage-adjusted fitted residuals $\tilde{\boldsymbol{\epsilon}} = \mathbf{D}\{\mathbf{Y} - \hat{\boldsymbol{\mu}}\} = \mathbf{DP}(\mathbf{Y} - \boldsymbol{\mu})\{1 + O_p(n^{-1/2})\}$. As before, we use the hat notation to denote plug-in estimates. This allows us to write

$$\begin{aligned}\mathbf{V}_{\text{sand},u} &= \tilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\widehat{\mathbf{W}}\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{P}\widehat{\mathbf{D}}\widehat{\mathbf{W}}\widehat{\mathbf{D}}\mathbf{P})\boldsymbol{\epsilon} \\ &= \sigma^2\dot{\boldsymbol{\epsilon}}^{\mathrm{T}}\widehat{\mathbf{M}}\dot{\boldsymbol{\epsilon}}\{1 + O(n^{-1})\}, \qquad (14)\end{aligned}$$

where $\mathbf{M} = \text{diag}_m(\mathbf{V}_i^{1/2})\mathbf{PDWDP}\,\text{diag}_m(\mathbf{V}_i^{1/2})$ and $\dot{\boldsymbol{\epsilon}}^{\mathrm{T}} = (\dot{\boldsymbol{\epsilon}}_1^{\mathrm{T}}, \ldots, \dot{\boldsymbol{\epsilon}}_n^{\mathrm{T}})$ independent, homoscedastic residuals defined by $\dot{\boldsymbol{\epsilon}}_i = \mathbf{V}_i^{-1/2}\boldsymbol{\epsilon}_i$, where we assume again that $\sigma^2\mathbf{V}_i$ correctly specifies the variance of Y_i. The quadratic form now easily allows calculation of the variance of the sandwich variance. Let m_{kl} denote the k, lth element of $\mathbf{M}$ and let $\dot{\epsilon}_k$ be the elements of $\dot{\boldsymbol{\epsilon}}$, where $k, l = 1, 2, \ldots mn$. Neglecting the effect of plug-in estimates, we find

$$\text{var}(\mathbf{V}_{\text{sand},u}) = 2\sigma^4\text{trace}(\mathbf{MM}) + \sigma^4\sum_k\{E(\dot{\epsilon}_k^4) - 3\}m_{kk}^2. \qquad (15)$$

If the $(\dot{\epsilon}_k)$ are standard normal, then (15) simplifies to $\text{var}(\mathbf{V}_{\text{sand},u}) = 2\sigma^4\text{tr}(\mathbf{MM})$. The variance of the sandwich variance estimate again depends distinctly on the design of the covariates because of $\partial\mu_i^{\mathrm{T}}/\partial\boldsymbol{\beta} = \mathbf{X}_i\partial h(\eta)/\partial\eta$ with $\eta = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$.

The foregoing calculation of the variance depends on the covariance structure $\mathbf{V}_i$ used for fitting. In the calculations we implicitly assumed that $\mathbf{V}_i$ was specified correctly. Even though this appears to be a conceptional restriction, we demonstrate in simulations that correction (8) actually is rather robust against misspecified covariances. This means that even if $\mathbf{V}_i$ is misspecified, the corrected profiles $z_{\tilde{p}}$ show a positive effect.

Example 6 (Multivariate Normal Response). Let $Y_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\mathbf{I})$ with $\mathbf{X}_i = (\mathbf{1}_m, \mathbf{U}_i)$, where $\mathbf{1}_m$ is the $m \times 1$-dimensional unit vector and $\mathbf{U}_i$ is an $m \times 1$ covariate vector. We set $\boldsymbol{\beta} = (.5, .5)^{\mathrm{T}}$ and consider $\beta_1 = (0, 1)\boldsymbol{\beta}$ to be the parameter of interest. We simulate from the following designs for the covariates: Let $\mathbf{U}_i = \mathbf{1}_m u_i$ with scalar $u_i \in \Re$ chosen (a) uniformly, (b) normally, or (c) from a Laplace distribution. Inserting $\mathrm{var}(\mathbf{V}_{\mathrm{sand},u}) = 2\sigma^4\mathrm{tr}(\mathbf{MM})$ in (8) shows that σ^4 cancels out as before, so that the correction depends only on the design and the working covariance $\mathbf{V}_i$. We assume working independence (i.e., $\mathbf{V}_i = \mathbf{I}$) and simulate Y_i from three different settings: with correctly specified working covariance matrix, that is, $\mathrm{var}(Y_i) = \sigma^2\mathbf{I}$ (model 1); with misspecified working covariances, that is, $\mathrm{var}(Y_i) = \sigma^2(3/4\mathbf{I} + 1/4\,\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}})$ (model 2), and with autocorrelated errors $\mathrm{var}(Y_i)_{rs} = \sigma^2\rho^{|r-s|}$ with $\rho = .5$ (model 3). The corrected quantiles are listed in Table 3. Figure 3 shows simulated coverage probabilities for 2000 simulations for the $p = .9$ confidence interval. For comparison, we again report the coverage probabilities for t-distribution quantiles with $n - 2$ degrees of freedom and for the multivariate jackknife estimate. The proposed adjustment shows satisfactory behavior for all three designs. The misspecification of the covariance has only a small effect on the coverage probability, so the adjustment appears to work for misspecified models as well. In contrast, both $t_{p,n-2}$ distribution quantiles and jackknife estimates show undercoverage, although the jackknife approach behaves more accurately.

For nonnormal data, $\mathrm{var}(\mathbf{V}_{\mathrm{sand},u})$ depends not only on the design and the working covariance, but also on the unknown parameter $\boldsymbol{\beta}$. This implies that in practice matrix $\mathbf{M}$ must to be estimated by plug-in estimates. Moreover, the latter term in (15) does not vanish, and the kurtosis must be estimated. Even though at first glance this appears cumbersome, estimation is usually not too complicated when assuming an underlying probability model. We demonstrate this using a binomial model.

Table 3. Corrected Quantiles $z_{\tilde{p}}$ for Different Designs

	$p = .90$		$p = .95$	
Design	$t_{p,n-2}$	$z_{\tilde{p}}$	$t_{p,n-2}$	$z_{\tilde{p}}$
	$n = 10$ ($m = 4$)		$n = 20$ ($m = 4$)	
(a)		2.03		1.81
(b)	1.86	2.10	1.71	1.86
(c)		2.18		1.94

Example 7 (Logistic Regression). We simulate (independent) binomial data with predictor $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (0, .5)^T$ (model 1) and $\boldsymbol{\beta} = (1, 1)^{\mathrm{T}}$ (model 2). The covariates $\mathbf{X}_i$ are distributed as in Example 6, and we are interested in the slope parameter β_1. For comparison, we again compare our proposed correction with the jackknife estimate, which in this case is a weighted and multivariate version of (9). The results are given in Table 4. The general positive appearance of the corrected quantiles carries over to binomial data, even if the distribution is rather skew, as in model 2. A similar behavior was also observed for simulations with Poisson data, not reported here.

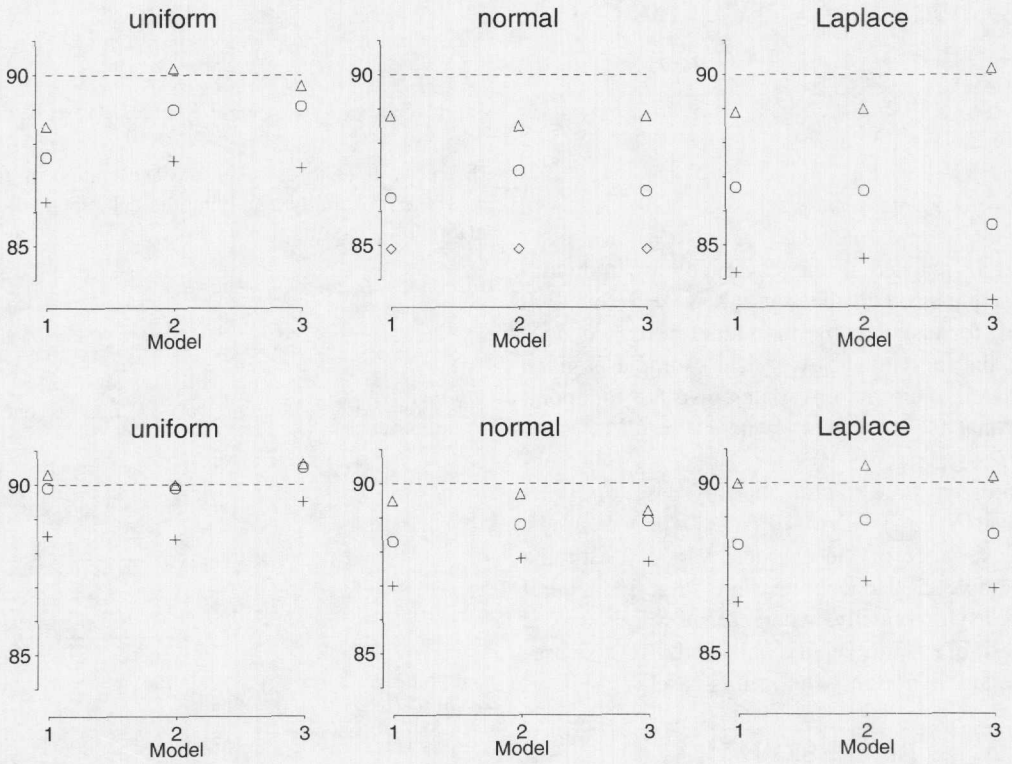

Figure 3. Coverage Probability of confidence Intervals Based on $\mathbf{V}_{\mathrm{sand},u}$ With Corrected Quantiles $z_{\tilde{p}}$ (△) as Well as With t-Distribution Quantiles $t_{p,n-1}P(+)$ and Based on the Jackknife Estimate $\mathbf{V}_{\mathrm{jack}}$ (○). The upper Row is for $n = 10$, $m = 4$; the bottom row is for $n = 20$, $m = 4$.

Table 4. Coverage Probability of Confidence Based on $\mathbf{V}_{\text{sand},u}$ With $z_{\tilde{p}}$ Calculated With True and Fitted Parameters and t-Distribution Quantiles $t_{p,n-1}$

			Coverage based on		
Design	$t_{p,n-2}$	$z_{\tilde{p}}$	$\mathbf{V}_{\text{sand},u}$ $z_{\tilde{p}}$	$\mathbf{V}_{\text{sand},u}$ $t_{p,n-2}$	$\mathbf{V}_{\text{jack}}$ $t_{p,n-2}$
Logistic regression $n = 30$ ($m = 4$), $p = .9$					
(a)		1.74 (1.74)	89.9 (90.6)	87.3 (87.7)	89.8 (90.4)
(b)	1.70	1.77 (1.78)	89.5 (90.1)	85.3 (84.6)	88.5 (89.0)
(c)		1.82 (1.83)	91.1 (91.8)	85.6 (85.1)	89.6 (90.5)
Logistic regression $n = 30$ ($m = 4$), $p = .95$					
(a)		2.08 (2.11)	95.4 (95.5)	93.4 (92.0)	95.4 (95.3)
(b)	2.04	2.12 (2.16)	95.4 (95.6)	92.1 (91.1)	94.9 (95.1)
(c)		2.19 (2.22)	95.8 (96.0)	91.7 (89.6)	95.2 (94.7)

3.4 Balanced Design

Finally, we revisit the design issue. So far we have focused on undercoverage properties with sandwich estimates. This undercoverage basically occurs if the covariates differ between the units, as in the foregoing simulations. In contrast, as we shown later, if the covariate design is the same for all units, then undercoverage may not occur.

Example 8 (Balance Design). Consider again the multivariate normal model $Y_i \sim N(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, with $\mathbf{X}_i^{\mathrm{T}}$ as a $m \times p$ design matrix. We assume that the covariates are scaled and orthogonal such that $\boldsymbol{\Omega} = \sum_i \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} = n\mathbf{I}$. This gives $\sum_i \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i = n$, and the variance is obtained from

$$\begin{aligned}\operatorname{var}\{\mathbf{V}_{\text{sand},u}\} &= 2\sigma^4\operatorname{tr}(\mathbf{MM}) = 2\operatorname{tr}(\mathbf{WW})\{1+O(n^{-1})\}\\ &= 2n^{-4}\sigma^4\sum_i(\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i)^2\{1+O(n^{-1})\}\\ &\geq 2n^{-5}\sigma^4\Big(\sum_i \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i\Big)^2\{1+O(n^{-1})\}\\ &= 2n^{-3}\sigma^4\{1+O(n^{-1})\}.\end{aligned}$$

The lower bound is reached if the covariates are individually orthogonal or balanced in the sense $\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} = \mathbf{I}$ for all i. This is the case if, for instance, the individual design $\mathbf{X}_i$ does not differ among the individuals. A typical example is given by longitudinal data, where the covariates give the timepoint of measurement, that is, $\mathbf{X}_i = (\mathbf{1}, \mathbf{t})$, where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}$ and $\mathbf{t} = (-T, -T+1, \ldots, T-1, T)^{\mathrm{T}}/(\sum_{t=-T}^{T} t^2)$ is a centered and standardized time vector. In this case one gets the lower bound $\operatorname{var}(\mathbf{V}_{\text{sand},u}) = 2\sigma^4/\{n^2(n-1)\}\{1+O(n^{-1})\}$, which equals the variance of the classical variance estimate discussed in Example 2. Hence one finds that in general $z_{\tilde{p}} \geq t_{p,n-1}$ holds asymptotically, where the lower bound is reached if the design is individually balanced. As a consequence undercoverage is not an issue in this case.

4. DISCUSSION

We have shown that sandwich variance estimates are typically less efficient than model-based variance estimates. The loss of efficiency depends mainly on the design; for standard cases, it is proportional to the inverse of the design kurtosis of the design points, and for nonnormal data, additional components beside the kurtosis influence the loss of efficiency. The variance of the sandwich variance estimate directly affects the coverage probability of confidence intervals. An adjustment has been suggested that depends on the design. The adjustment has shown promising behavior, although we expect it to be possible to break down the method.

Basically, the use of the sandwich variance estimate leads to undercoverage of confidence intervals if the covariates differ between the units. For individually balanced designs, as may occur in dynamic data, undercoverage does not occur. Therefore, we can refine the statement of Diggle et al. (1994, p. 77) that the sandwich variance estimate should be used with care if the data come from a small number of "experimental units" and the covariates differ between the units. In this case, the suggested corrected quantiles provide a small-sample adjustment for the confidence intervals.

APPENDIX: TECHNICAL DETAILS

A.1 Proof of Theorem 2

In general, the result can be proved by applying an Edgeworth series to $\hat{\theta} - \theta$ (see, e.g., Hall 1992, pp. 46–68). But we pursue a more direct proof here, which makes the result accessible for readers not too familiar with Edgeworth series.

Let $n^{1/2}(\hat{\theta} - \theta) \sim \text{normal}(0, \sigma^2)$ and $z_p = \Phi^{-1}(p)$, where $\Phi(\cdot)$ is the standard normal distribution function. We define $v_p = \sigma z_p$ and $\hat{v}_p = \hat{\sigma} z_p$ such that $F(v_p) = \Pr\{n^{1/2}(\hat{\theta} - \theta) \leq v_p\} = p$ with $F(v_p) = \Phi(z_p)$. The intention is to calculate $\Pr\{n^{1/2}(\hat{\theta} - \theta) \leq \hat{v}_p\}$. Let $H_{\hat{v}_p}(\cdot)$ denote the distribution function of $\hat{v}_p$, and take $\hat{\sigma}^2$ as the $\sqrt{n}$-consistent variance estimate independent of $\hat{\theta} - \theta$. This gives

$$\begin{aligned}\Pr\{(\hat{\theta} - \theta) \leq \hat{v}_p\} &= \int \Pr\{(\hat{\theta} - \theta) \leq (v|\hat{v}_p = v)dH_{\hat{v}_p}(v)\\ &= \int F(v)dH_{\hat{v}_p}(v) = E\{F(\hat{v}_p)\}.\end{aligned}$$

Hence we have to calculate the expectation of $F(\hat{v}_p)$ to obtain the coverage probability. Applying the delta method to the root function $g(v) = v^{1/2}$, we find that

$$\begin{aligned}\hat{\sigma} - \sigma &= g(\hat{\sigma}^2) - g(\sigma^2)\\ &= \frac{\hat{\sigma}^2 - \sigma^2}{2\sigma} - \frac{(\hat{\sigma}^2 - \sigma^2)^2}{8\sigma^3} + O_p(n^{-3/2}).\end{aligned}$$

This along with $\hat{v}_p = v_p + z_p(\hat{\sigma} - \sigma)$ implies that

$$\begin{aligned}F(\hat{v}_p) &= F\left\{v_p + z_p\frac{\hat{\sigma}^2 - \sigma^2}{2\sigma^2} - z_p\frac{(\hat{\sigma}^2 - \sigma^2)^2}{8\sigma^4}\right\} + O_p(n^{-3/2})\\ &= F(v_p) + F^{(1)}(v_p)\left\{z_p\frac{\hat{\sigma}^2 - \sigma^2}{2\sigma^2} - z_p\frac{(\hat{\sigma}^2 - \sigma^2)^2}{8\sigma^4}\right\}\\ &\quad + \frac{1}{2}F^{(2)}(v_p)\left\{z_p\frac{\hat{\sigma}^2 - \sigma^2}{2\sigma^2}\right\}^2 + O_p(n^{-3/2}).\end{aligned}$$

Because $F(v_p) = p$, this yields

$$\begin{aligned}E\{F(\hat{v}_p)\} &= p + \operatorname{var}(\hat{\sigma}^2)\left\{\frac{z_p^2F^{(2)}(z_p)}{8\sigma^4} - \frac{z_pF^{(1)}(z_p)}{8\sigma^4}\right\}\\ &\quad + O(n^{-3/2}).\end{aligned}$$

Inserting the derivatives for $F(v) = \Phi(v/\sigma)$ gives formula (7) in Theorem 2.

Note that $\tilde{p} - p = O(n^{-1})$, so that by simple Taylor expansion,

$$\begin{aligned} z_p &= \Phi^{-1}(p) = \Phi^{-1}(\tilde{p} + p - \tilde{p}) \\ &= z_{\tilde{p}} + (p - \tilde{p})/\phi(z_{\tilde{p}}) + O(n^{-2}), \end{aligned}$$

which provides $z_p - z_{\tilde{p}} = O(n^{-1})$. Let $\hat{\sigma}_t^2$ be the variance estimate such that $(\hat{\theta} - \theta)/\hat{\sigma}_t$ is t-distributed with the usual degrees of freedom. Denoting by t_p the t-distribution quantiles, calculations similar to the foregoing show that $t_p = z_p + (p - p_t)/\phi(t_p) + O(n^{-2})$, where p_t is defined through $z_{p_t} = t_p$. Using these equations and applying (8) provides

$$\begin{aligned} z_{\tilde{p}} &= t_p + \frac{z_{\tilde{p}}^3 + z_{\tilde{p}}}{8\sigma^4}\operatorname{var}(\hat{\sigma}^2) - \frac{t_p^3 + t_p}{8\sigma^4}\operatorname{var}(\hat{\sigma}_t^2) + O(n^{-2}) \\ &= t_p + \frac{z_p^3 - z_p}{8}\frac{\operatorname{var}(\hat{\sigma}_t^2)}{\sigma^4}\left(\frac{\operatorname{var}(\hat{\sigma}^2)}{\operatorname{var}(\hat{\sigma}_t^2)} - 1\right) + O(n^{-2}), \end{aligned}$$

as claimed in Section 2.2.

A.2 Sandwich Estimates in Quasi-Likelihood and Generalized Estimating Equations

Here we derive the relative efficiency in quasi-likelihood models. For simplicity of notation, we consider univariate regression models of the form $E(Y_i|\mathbf{x}_i) = \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) = h(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ with $\mathbf{x}_i^{\mathrm{T}}$ as a $1 \times p$ vector. The variance of Y_i is given by $\operatorname{var}(Y_i|\mathbf{x}_i) = \sigma^2 V\{\mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}$, where $V(\cdot)$ is a known variance function. In some problems σ^2 is estimated, which we indicate by setting $\xi = 1$, whereas when σ^2 is known, we set $\xi = 0$. We denote the derivatives of functions by superscripts, for example, $\mu^{(l)}(\eta) = \partial^l \mu(\eta)/(\partial\eta)^l$. Let us assume that the variance is correctly specified, that is, $\operatorname{var}(Y_i|\mathbf{x}_i) = \sigma^2 V\{\mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}$, so that with expansion (12) we get $\operatorname{var}(n^{1/2}\hat{\mathbf{z}}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{V}_{\mathrm{asymp}}\{1 + O(n^{-1})\}$, where $\mathbf{V}_{\mathrm{asymp}} = \sigma^2\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z}$ and $\boldsymbol{\Omega}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} Q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ with $Q(\eta) = \{\mu^{(1)}(\eta)\}^2/V(\eta)$. The model-based variance estimator for $n^{1/2}\hat{\mathbf{z}}^{\mathrm{T}}\boldsymbol{\beta}$ is $\mathbf{V}_{\mathrm{model}} = \hat{\sigma}^2(\hat{\boldsymbol{\beta}})\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\hat{\boldsymbol{\beta}})\mathbf{z}$, where

$$\hat{\sigma}^2(\boldsymbol{\beta}) = \xi n^{-1}\sum_{i=1}^n \{Y_i - \mu(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}^2/V(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) + \sigma^2(1 - \xi).$$

Defining $\mathbf{B}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} M(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\{Y_i - \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^2$ and $M(\eta) = \{\mu^{(1)}(\eta)/V(\eta)\}^2$, the sandwich estimator for $n^{1/2}\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is written as $V_{\mathrm{sand}} = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\hat{\boldsymbol{\beta}})\mathbf{B}_n(\hat{\boldsymbol{\beta}})\boldsymbol{\Omega}_n^{-1}(\hat{\boldsymbol{\beta}})\mathbf{z}$.

To derive the following theorem, we need some additional notation. Let $\mathbf{R}_n = \xi n^{-1}\sum_{i=1}^n g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}_i^{\mathrm{T}}$, where $g(\eta) = (\partial/\partial\eta)\log\{V(\eta)\}$, $\epsilon_i = \{Y_i - \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}/V^{1/2}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, $q_{in} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z}$, $a_n = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z}$, $\mathbf{C}_n = n^{-1}\sum_{i=1}^n q_{in}^2 Q^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}_i$,

$$\begin{aligned} \boldsymbol{\ell}_{in} &= \boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{x}_i\mu^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})/V^{1/2}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \\ v_i &= \{Y_i - \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^2 M(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \sigma^2 Q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \end{aligned}$$

and

$$\mathbf{K}_n = n^{-1}\sum_{i=1}^n q_{in}^2 V(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})M^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}_i.$$

In what follows, we treat $\mathbf{x}_i$ as a sample from a distribution. We assume that sufficient moments of the components of $\mathbf{x}$ and y exist, as does sufficient smoothness of $\mu(\cdot)$. Under the foregoing conditions, at least asymptotically there are no leverage points, so that the usual and unbiased sandwich estimators will have similar asymptotic behavior. We write $\overline{\boldsymbol{\Omega}}(\boldsymbol{\beta}) = E\{\boldsymbol{\Omega}_n(\boldsymbol{\beta})\}$, $q = \mathbf{x}^{\mathrm{T}}\overline{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta})\mathbf{z}$, $a = \mathbf{z}^{\mathrm{T}}\overline{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta})\mathbf{z}$, $\overline{\mathbf{C}} = E\{q^2 Q^{(1)}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}\}$, and so on—that is, the bar notation refers to asymptotic moments.

Theorem A.1. As $n \to \infty$, under the foregoing conditions we have

$$\begin{aligned} &n^{1/2}(\mathbf{V}_{\mathrm{model}} - \mathbf{V}_{\mathrm{asymp}}) \\ &\quad \Rightarrow \operatorname{normal}\left[0, \Sigma_{\mathrm{model}} := E\{a\xi(\epsilon^2 - \sigma^2) - \sigma^2(a\overline{\mathbf{R}} + \overline{\mathbf{C}})^{\mathrm{T}}\boldsymbol{\ell}\epsilon\}^2\right] \end{aligned}$$

and

$$\begin{aligned} &n^{1/2}(\mathbf{V}_{\mathrm{sand}} - \mathbf{V}_{\mathrm{asymp}}) \\ &\quad \Rightarrow \operatorname{normal}\left[0, \Sigma_{\mathrm{sand}} := E\{q^2 v + (\overline{\mathbf{K}} - 2\sigma^2\overline{\mathbf{C}})^{\mathrm{T}}\boldsymbol{\ell}\epsilon\}^2\right]. \end{aligned}$$

For the proof, recall that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx n^{-1/2}\sum_{i=1}^n \boldsymbol{\ell}_{in}\epsilon_i$, where $\approx$ means that the difference is of order $o_p(1)$. By a simple delta-method calculation we get $\xi n^{1/2}\{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}) - \sigma^2\} \approx n^{-1/2}\sum_{i=1}^n \xi(\epsilon_i^2 - \sigma^2) - \sigma^2\mathbf{R}_n^{\mathrm{T}} n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Thus

$$\begin{aligned} &n^{1/2}\{\mathbf{V}_{\mathrm{model}} - \mathbf{V}_{\mathrm{asymp}}\} \\ &\quad \approx \xi n^{1/2}\{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}) - \sigma^2\}a_n + n^{1/2}\sigma^2\mathbf{z}^{\mathrm{T}}\{\boldsymbol{\Omega}_n^{-1}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\}\mathbf{z} \\ &\quad \approx \xi n^{1/2}\{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}) - \sigma^2\}a_n - \sigma^2 n^{1/2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\{\boldsymbol{\Omega}_n(\hat{\boldsymbol{\beta}}) \\ &\qquad - \boldsymbol{\Omega}_n(\boldsymbol{\beta})\}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z} \\ &\quad \approx \xi n^{1/2}\{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}) - \sigma^2\}a_n - \sigma^2\mathbf{C}_n^{\mathrm{T}} n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad \approx n^{-1/2}\sum_{i=1}^n \{a_n\xi(\epsilon_i^2 - \sigma^2) - \sigma^2(a_n\mathbf{R}_n + \mathbf{C}_n)^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i\}, \end{aligned}$$

which shows the first part of Theorem A.1.

We now turn to the sandwich estimator and note that $\mathbf{B}_n(\boldsymbol{\beta}) - \sigma^2\boldsymbol{\Omega}_n(\boldsymbol{\beta}) = O_p(n^{-1/2})$. Because of this, we have that

$$\begin{aligned} &n^{1/2}\{\mathbf{V}_{\mathrm{sand}} - \mathbf{V}_{\mathrm{asymp}}\} \\ &\quad \approx -2\sigma^2 n^{1/2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\{\boldsymbol{\Omega}_n(\hat{\boldsymbol{\beta}}) - \boldsymbol{\Omega}_n(\boldsymbol{\beta})\}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z} \\ &\qquad + n^{1/2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\{\mathbf{B}_n(\hat{\boldsymbol{\beta}}) - \sigma^2\boldsymbol{\Omega}_n(\boldsymbol{\beta})\}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z} \\ &\quad \approx -2\sigma^2 n^{-1/2}\sum_{i=1}^n \mathbf{C}_n^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i + n^{-1/2}\sum_{i=1}^n q_{in}^2[M(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}) \\ &\qquad \times \{Y_i - \mu(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})\}^2 - \sigma^2 Q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})] \\ &\quad \approx -2\sigma^2 n^{-1/2}\sum_{i=1}^n \mathbf{C}_n^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i + n^{-1/2}\sum_{i=1}^n q_{in}^2 v_i \\ &\qquad + n^{-1}\sum_{i=1}^n q_i^2 M^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{X}_i\{Y_i - \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^2 n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad \approx -2\sigma^2 n^{-1/2}\sum_{i=1}^n \mathbf{C}_n^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i + n^{-1/2}\sum_{i=1}^n q_{in}^2 v_i \\ &\qquad + n^{-1}\sum_{i=1}^n q_i^2 M^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{X}_i V(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad \approx n^{-1/2}\sum_{i=1}^n (-2\sigma^2\mathbf{C}_n^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i + q_i^2 v_i + \mathbf{K}_n^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i), \end{aligned}$$

as claimed.

Theorem A.1 can now be used to prove the statements listed in Examples 4 and 5. For the logistic case, we have $V(\eta) = \mu^{(1)}(\eta) = Q(\eta) = \mu(\eta)\{1 - \mu(\eta)\}$, $\sigma^2 = 1, \xi = 0, \mathbf{R}_n = 0$, and $Q^{(1)}(\eta) = \mu^{(1)}(\eta)\{1 - 2\mu(\eta)\}$. All of the terms in Theorem A.1 can then be computed by numerical integration, which gives the numbers presented in Example 5.

For the Poisson case, it is easily verified that $\overline{\boldsymbol{\Omega}}(\boldsymbol{\beta}) = \exp(\beta_0)\mathbf{I}_2$, where $\mathbf{I}_2$ is the 2×2 identity matrix. Also, $q = U\exp(-\beta_0)$, $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = \beta_0$, $Q^{(1)}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \exp(\beta_0)$, $\overline{\mathbf{C}} = \exp(-\beta_0)(1, 0)^{\mathrm{T}}$,

$\ell = \exp(-\beta_0/2)(1, U)^{\mathrm{T}}$, $\epsilon = \{Y - \exp(\beta_0)\}/\exp(\beta_0/2)$, and hence $\Sigma_{\text{model}} = \exp(-3\beta_0)$.

Let $\theta = \exp(\beta_0)$. Then $E(Y^2) = \theta + \theta^2$, $E(Y^3) = \theta^3 + 3\theta^2 + \theta$, and $E(Y^4) = \theta^4 + 6\theta^3 + 7\theta^2 + \theta$. If we define $Z = Y - \theta$, then $E(Z) = 0$, $E(Z^2) = E(Z^3) = \theta$, and $E(Z^4) = 3\theta^2 + \theta$. Further, $M(\eta) = 1$, $M^{(1)}(\eta) = 0$, and $\overline{\mathbf{K}} = 0$. A detailed calculation then gives that $\Sigma_{\text{sand}} = 2\kappa\exp(-2\beta_0) + \kappa\exp(-3\beta_0)$, which shows the relative efficiency given in Example 4.

[Received March 2000. Revised April 2001.]

REFERENCES

Breslow, N. (1990), "Test of Hypotheses in Overdispersion Regression and Other Quasi-Likelihood Models." *Journal of the American Statistical Association*, 85, 565–571.

Carroll, R. J., and Ruppert, D. (1998), *Transformation and Weighting in Regression*, London: Chapman and Hall.

Chesher, A., and Jewitt, I. (1987), "The Bias of a Heteroscedasticity Consistent Covariance Matrix Estimator," *Econometrica*, 55, 1217–1222.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman and Hall.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford, U.K.: Clarendon Press.

Efron, B. (1986), Discussion of "Jackknife, Bootstrap and Other Resampling Methods in Statistics," by C. F. J. Wu, *The Annals of Statistics*, 14, 1301–1304.

Eicker, F. (1963), "Asymptotic Normality and Consistency of the Least Squares Estimator for Families of Linear Regression," *Annals of Mathematical Statistics*, 34, 447–456.

Firth, D. (1992), Discussion of "Multivariate Regression Analysis for Categorical Data," by Liang, Zeger, and Qaqish, *Journal of the Royal Statistical Society*, Ser. B, 54, 24–26.

Gourieroux, C., Monfort, A., and Trognon, A. (1984), "Pseudo Maximum Likelihood Methods: Applications to Poisson Models," *Econometrica*, 52, 701–720.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

Hinkley, D. V. (1977), "Jackknifing in Unbalanced Situations," *Technometrics*, 19, 285–292.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimation Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, eds. L. M. LeCam and J. Neyman, Berkeley: University of California Press, pp. 221–233.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Liang, K. Y., Zeger, S. L., and Qaqish, B. (1992), "Multivariate Regression Analysis for Categorical Data," *Journal of the Royal Statistical Society*, Ser. B, 54, 3–40.

Long, J. S., and Ervin, L. H. (2000), "Using Heteroscedasticity-Consistent Standard Errors in the Linear Regression Model," *The American Statistician*, 54, 217–223.

MacKinnon, J. G., and White, H. (1985), "Some Heteroscedasticity-Consistent Covariance Matrix Estimators With Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325.

McCullagh, P. (1987), *Tensor Methods in Statistics*, London: Chapman and Hall.

———. (1992). Discussion of "Multivariate Regression Analysis for Categorical Data," by Liang, Zeger, and Qaqish, *Journal of the Royal Statistical Society*, Ser. B, 54, 24–26.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman and Hall.

Rothenberg, T. J. (1988), "Approximative Power Functions for Some Robust Tests of Regression Coefficients," *Econometrica*, 56, 997–1019.

Wedderburn, R. W. M. (1974), "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss–Newton Method," *Biometrika*, 61, 439–447.

White, H. (1980), "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity," *Econometrica*, 48, 817–838.

Wu, C. F. J. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Statistics," *The Annals of Statistics*, 14, 1261–1350.