THE SANDWICH (ROBUST COVARIANCE MATRIX) ESTIMATOR

R. J. Carroll, Suojin Wang, D. G. Simpson, A. J. Stromberg and D. Ruppert *

January 26, 1998

Abstract

The sandwich estimator, often known as the robust covariance matrix estimator or the empirical covariance matrix estimator, has achieved increasing use 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 a parametric model fails to hold, or is not even specified. Surprisingly though, there has been little discussion of the properties of the sandwich method other than consistency. We investigate the sandwich estimator in quasilikelihood models asymptotically, and in the linear case analytically. We show that when the quasilikelihood model is correct, the sandwich covariance matrix estimate is often far more variable than the usual parametric variance estimate, and its coverage probabilities can be abysmal. The increased variance is a fixed feature of the method, and the price one pays to obtain consistency even when the parametric model fails. We make some simple suggestions for modifying the method which improve coverage probabilities.

Key words and phrases: Estimating equations; Generalized estimating equations; Generalized linear models; Heteroscedasticity; Linear regression; Quasilikelihood; Robust covariance estimator; Sandwich estimator.

Short title. The Sandwich Estimator

*R. J. Carroll and Suojin Wang are with the Department of Statistics, Texas A&M University, College Station, TX 77843–3143. Arnold J. Stromberg is with the Department of Statistics, University of Kentucky, Lexington KY 40506–0027. Douglas G. Simpson is Professor of Statistics, Department of Statistics, University of Illinois, Champaign IL 61820. David Ruppert is Professor, School of Operations Research & Industrial Engineering, Cornell University, Ithaca, New York 14853-3801. Carroll and Wang's research was supported by a grant from the National Cancer Institute (CA–57030). Wang's research was also supported by a grant from the National Security Agency (MDA 904-96-1-0029). Stromberg's research was supported by a grant from the National Security Agency (97–051). Simpson's research was supported by grants from the National Science Foundation (DMS 92-07730) and the National Security Agency (MDA 904-92-H-3058). Ruppert's research was supported by a grant from the National Science Foundation (DMS-9306196).

1 INTRODUCTION

The sandwich estimation procedure is a general method for estimating the covariance matrix of parameter estimates. Traceable back at least to Huber (1967) and White (1982), the method yields asymptotically consistent covariance matrix estimates

- without making distributional assumptions; and
- even if the assumed model underlying the parameter estimates is incorrect.

Because of these two desirable model–robustness properties, the sandwich estimator is often called the *robust covariance matrix* estimator, or the *empirical covariance matrix* estimator. The sandwich method is widely employed in marginal methods such as generalized estimating equations (Diggle, Liang & Zeger, 1994; Liang & Zeger, 1986; Liang, Zeger & Qaqish, 1992), and has achieved increasing popularity to the point that estimation of the covariance matrix of parameter estimates is no longer considered much of an issue.

For example, consider linear regression. The usual mean squared error based covariance matrix estimate of the least squares regression parameter estimates is used almost uniformly, but it is inconsistent if the errors are heteroscedastic. In contrast, the sandwich estimate is consistent even under heteroscedasticity, under some reasonable regularity conditions. The argument in favor of the sandwich estimate is that it is only estimating the variance of an estimator, and asymptotic normality and proper coverage confidence intervals only require a consistent variance estimate, so that there really is no great need to construct a highly accurate covariance matrix estimate.

There have been, however, some intimations that the sandwich estimator might not be a very good estimator. In his discussion of a paper of Wu (1986), Efron (1986) gives simulation evidence of this phenomenon. Breslow (1990) demonstrated this in a simulation study of overdispersed Poisson regression. Firth (1992) and McCullagh (1992) both raise concerns that the sandwich estimator may be particularly inefficient. Diggle, et al. (1994, page 77) suggest that it is best used when the data come from "many experimental units".

The purposes of this paper are to investigate two questions:

- How bad is the sandwich estimate when compared to a parametric estimate, when the parametric model is true; and
- Does inefficiency have any impact on inference for samples of moderate size?

The first question turns out to have a fairly precise asymptotic answer, and sometimes a fairly precise small sample answer, neither of which appear to have been quantified precisely before. For example, the sandwich method in simple linear regression when estimating the slope has an asymptotic inefficiency equal to the inverse of the sample kurtosis of the design values, so that for example if the predictors were generated according to the Laplace distribution, the sandwich method has asymptotic efficiency $1/6$ compared to the usual estimate when the linear model holds. In fact we show that the sandwich estimate is much more affected by leverage than is the usual estimate. This inefficiency still holds in generalized linear models. For example, in simple linear logistic regression, at the null value that there is no effect due to 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 coverage is far more difficult to answer, because there is no precise definition of "samples of moderate size", and because the results vary according to the situation. The major evidence of which we are aware consists of the simulation studies of Wu (1986) and Breslow (1990), both of whom find somewhat elevated levels of Wald–type tests based on the sandwich estimator.

We have run our own simulations of linear and logistic regression. In simple linear regression, with samples of size 20, with the predictors generated from a Laplace distribution, the simulated coverage probability of a nominal 95% confidence interval using the usual implementation of the sandwich method has only 88.1% coverage, and even with leverage adjustments and using $n-p$ degrees of freedom the coverage is only 91.9%. Coverage is closer to the nominal as the distribution of the predictor becomes lighter tailed, and worse if the predictor is heavier tailed.

In logistic regression, the important part of sample size considerations is usually thought to be the number of events. When the number of events is small, undercoverage of Wald–type tests using the sandwich estimator also can be a problem, again especially for heavy–tailed design distributions. With samples of size $200, 300, 400$ and a response rate of 5%, with Laplace distributed predictors, at the null model the coverage of the usual sandwich method based on $5,000$ simulations is only 88.7%, 91.0% and 91.4%. With the same sample sizes, but with the response rate chosen so that the expected number of events is 15, the coverage probabilities are 91.7%, 91.0% and 90.6%.

The paper is organized as follows. In Section 2 we compare the sandwich estimator to the usual parametric regression estimator in the homoscedastic linear regression model. Section 3 makes an asymptotic comparison of the methods for quasilikelihood models. Some simulations are discussed in Section 4, in which we suggest some simple modifications to the sandwich method which improve coverage probabilities. Section 5 contains concluding remarks. Proofs are given in the appendix.

2 LINEAR REGRESSION

2.1 The Estimators

Consider the linear regression model $Y_i = \mathbf{X}_i^t\boldsymbol{\beta} + \epsilon_i$, where the random errors ϵ_i are normally distributed with mean zero and common variance σ^2. Let $\widehat{\boldsymbol{\beta}}$ be the ordinary least squares estimator of $\boldsymbol{\beta}$, and define $\widetilde{\boldsymbol{\mathcal{Z}}} = (\mathbf{X}_1, ..., \mathbf{X}_n)^t$. Let the hat matrix $\mathbf{H} = \widetilde{\boldsymbol{\mathcal{Z}}}(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\widetilde{\boldsymbol{\mathcal{Z}}}^t = (h_{ij})$.

We are interested in estimating the linear combination $\mathbf{L}^t\boldsymbol{\beta}$. Our particular interest here is in estimating the variance of $\mathbf{L}^t\widehat{\boldsymbol{\beta}}$, this variance given by $\sigma^2\mathbf{L}^t(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\mathbf{L}$. The classical estimator is given by $V_{ols} = s^2\mathbf{L}^t(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\mathbf{L}$, where $s^2 = (n-p)^{-1}\sum_{i=1}^n r_i^2$, p is the dimension of $\boldsymbol{\beta}$ and the residuals are $r_i = Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}$.

We will distinguish between two versions of the sandwich variance estimator. The sandwich estimator as commonly employed, which we call V_{sand}, is defined as follows. Let $\mathbf{R} = (r_1, ..., r_n)^t(r_1, ..., r_n)$ and $\mathbf{D}_r = \mathrm{diag}(\mathbf{R})$. Define $a_i = \mathbf{L}^t(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\mathbf{X}_i$. Then

$$V_{sand} = \mathbf{L}^t(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\widetilde{\boldsymbol{\mathcal{Z}}}^t\mathbf{D}_r\widetilde{\boldsymbol{\mathcal{Z}}}(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\mathbf{L} = \sum_{i=1}^n a_i^2 r_i^2. \tag{1}$$

In linear regression, (1) is often multiplied by $n/(n-p)$ (Hinkley, 1977). While V_{sand} is most commonly applied, it is a biased estimator because $E(r_i^2) = \sigma^2(1-h_{ii})$. This suggests replacing r_i in (1) by $t_i = r_i/(1-h_{ii})^{1/2}$, an estimator we refer to as $V_{sand,u}$ (Wu, 1986, equation 2.6).

2.2 Properties of the Ordinary Sandwich Estimator

We have that $E(V_{sand}) = \sigma^2\sum_{i=1}^n(1-h_{ii})a_i^2$, for $i = 1, \ldots, n$. Observing that $\sum_{i=1}^n a_i^2 = \mathbf{L}^t(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\mathbf{L}$, the expectation becomes

$$E(V_{sand}) = \sigma^2\mathbf{L}^t(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\mathbf{L}(1-b_n), \tag{2}$$

where $0 \leq b_n = \sum_{i=1}^n h_{ii}a_i^2/\sum_{i=1}^n a_i^2 \leq \max_{1\leq i\leq n} h_{ii}$. In general the sandwich estimator is biased *downward*.

We have observed that in simulations, the bias of the usual sandwich estimator tends to be substantial when there are leverage points. The following result confirms this and is easy to show by making the first point a *leverage* point such that $h_{11} = \max_{1\leq i\leq n} h_{ii}$ and setting $\mathbf{L} = \mathbf{X}_1/\sqrt{\mathbf{X}_1^t(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\mathbf{X}_1}$.

Theorem 1: The sandwich estimator has

$$\max_{\mathrm{var}(\mathbf{L}^t\widehat{\boldsymbol{\beta}})=\sigma^2} |\mathrm{bias}(V_{sand})| \geq \max_{1\leq i\leq n} h_{ii}^2.$$

Thus, if there is a large leverage point, the usual sandwich estimator can be expected to have poor behavior relative to the classical formula.

Even in problems without leverage points, the usual sandwich estimator is typically inefficient. Using well–known results on higher moments of the multivariate normal distribution we obtain $\mathrm{var}(r_i^2) = 2(1-h_{ii})^2\sigma^4$, and $\mathrm{cov}(r_i^2, r_j^2) = 2h_{ij}^2\sigma^4$ $(i \neq j)$. It follows that the variance of the sandwich estimator is given by

$$\mathrm{var}(V_{sand}) = \sum_{i=1}^n a_i^4 \mathrm{var}(r_i^2) + \sum_{i\neq j} a_i^2 a_j^2 \mathrm{cov}(r_i^2, r_j^2) = 2\sigma^4 \sum_{i=1}^n a_i^4 (1-h_{ii})^2 + 2\sigma^4 \sum_{i\neq j} a_i^2 a_j^2 h_{ij}^2, \tag{3}$$

We combine this calculation with the the result that $\mathrm{var}(V_{ols}) \approx 2\sigma^4\{\mathbf{L}^t(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\mathbf{L}\}^2/n$ to obtain the asymptotic relative efficiency versus the classical estimate for regular designs.

Theorem 2: If the design sequence is regular, i.e., if $\max_{1\leq i\leq n} h_{ii} = o(n^{-1/2})$, then

$$\mathrm{ARE}(\mathrm{Sandwich}\ |\mathrm{Classical}) \sim \left\{ n^{-1}\sum_{i=1}^n a_i^2 \right\}^2 \left\{ n^{-1}\sum_{i=1}^n a_i^4 \right\}^{-1} \leq 1. \tag{4}$$

Example 1 (the intercept): Suppose the first column of $\widetilde{\boldsymbol{\mathcal{Z}}}$ is a vector of ones, the other columns have means of zero, and $\mathbf{L}^t = (1, 0, \ldots, 0)$. We then have $a_i = n^{-1}$ and the asymptotic relative efficiency in (4) is 1.

Example 2 (the slope in simple linear regression): Assume $\mathbf{X}_i^t = (1, U_i)$ where $\sum U_i = 0$. Suppose $\mathbf{L}^t = (0, 1)$ so $\mathbf{L}^t\widehat{\boldsymbol{\beta}}$ is the slope estimate. Because $h_{ii} = n^{-1}(1 + U_i^2)$, the design sequence is regular as long as $\max(|U_i|) = o(n^{1/4})$, in which case the asymptotic relative efficiency is M_n^{-1}, where $M_n = n^{-1}\sum U_i^4/(n^{-1}\sum_{i=1}^n U_i^2)^2 \geq 1$.

Example 2 shows that the distribution of the design points has an important role to play in the properties of the sandwich estimator. The asymptotic efficiency is inversely proportional to the kurtosis of the design points, where the kurtosis equals 3 for normally distributed observations. Thus in particular, if the design points $(U_1, ..., U_n)$ were realizations of a normal distribution, the sandwich estimator V_{sand} has 3 times the variability of the usual estimator V_{ols}. If the design points were generated from a Laplace distribution, the usual sandwich estimator is 6 times more variable.

2.3 The Unbiased Sandwich Estimator

Similar calculations can be performed for the unbiased sandwich estimator $V_{sand,u}$. The computation of the variance is similar to the computation for the ordinary sandwich estimator V_{sand}.

The calculations are based on the facts $\text{var}(t_i^2) = 2\sigma^4$ and $\text{cov}(t_i^2, t_j^2) = 2h_{ij}^2\sigma^4/\{(1-h_{ii})(1-h_{jj})\}$ for $i \neq j$. We can bound the relative efficiency without design regularity conditions. Regularity conditions allow evaluation of the asymptotic relative efficiency, which we state without proof.

Theorem 3: Under the homoscedastic linear model the unbiased sandwich and classical variance estimates for $\mathbf{L}^t\widehat{\boldsymbol{\beta}}$ satisfy:

$$\text{Efficiency(Sandwich} \mid \text{Classical)} \leq \{n^{-1}\sum_{i=1}^{n} a_i^2\}^2\{n^{-1}\sum_{i=1}^{n} a_i^4\}^{-1} \leq 1.$$

If in addition $\max(h_{ii}) = o(n^{-1/2})$, then the middle term is the asymptotic relative efficiency, which is of course the same as for the usual sandwich estimator.

3 QUASILIKELIHOOD

We now derive an asymptotic comparison between the sandwich and usual estimators in a quasilikelihood model. The mean of Y given $\mathbf{X}$ is $\mu(\mathbf{X}^t\boldsymbol{\beta})$ and its variance is $\sigma^2\mathcal{V}(\mathbf{X}^t\boldsymbol{\beta})$, where the functions $\mu(\cdot)$ and $\mathcal{V}(\cdot)$ are known. In some problems, σ^2 is estimated, which we indicate by setting $\xi = 1$, while when σ^2 is known we set $\xi = 0$. The quasilikelihood estimate of $\boldsymbol{\beta}$ is the solution $\widehat{\boldsymbol{\beta}}$ to

$$0 = \sum_{i=1}^{n}\{Y_i - \mu(\mathbf{X}_i^t\widehat{\boldsymbol{\beta}})\}\mathbf{X}_i\mu^{(1)}(\mathbf{X}_i^t\widehat{\boldsymbol{\beta}})/\mathcal{V}(\mathbf{X}_i^t\widehat{\boldsymbol{\beta}}),$$

where in general the jth derivative of a function $f(x)$ is denoted by $f^{(j)}(x)$.

The usual estimator of the covariance matrix of $n^{1/2}\mathbf{L}^t(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is

$$V_{ql} = \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}})\mathbf{L}^t\mathbf{A}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{L},$$

where $\mathbf{A}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^t Q(\mathbf{X}_i^t\boldsymbol{\beta})$; $Q(x) = \{\mu^{(1)}(x)\}^2/\mathcal{V}(x)$, and

$$\widehat{\sigma}^2(\boldsymbol{\beta}) = \xi n^{-1}\sum_{i=1}^{n}\{Y_i - \mu(\mathbf{X}_i^t\boldsymbol{\beta})\}^2/\mathcal{V}(\mathbf{X}_i^t\boldsymbol{\beta}) + \sigma^2(1-\xi).$$

Defining $\mathbf{B}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^t M(\mathbf{X}_i^t\boldsymbol{\beta})\{Y_i - \mu(\mathbf{X}_i^t\boldsymbol{\beta})\}^2$ and $M(x) = \{\mu^{(1)}(x)/\mathcal{V}(x)\}^2$, the usual sandwich estimator is

$$V_{sand} = \mathbf{L}^t\mathbf{A}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{B}_n(\widehat{\boldsymbol{\beta}})\mathbf{A}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{L}. \tag{5}$$

There is a version of the hat matrix for quasilikelihood models, see Cook & Weisberg (1982, pages 191–192) for logistic regression and Carroll & Ruppert (1987, page 74) for other models. In either case, we use the notation $\mathbf{H} = (h_{ij})$ for the hat matrix. In particular, let $\widehat{W} = \text{diag}\{Q(\mathbf{X}_i^t\widehat{\boldsymbol{\beta}})\}$. Then the hat matrix is $H = (\widehat{W}^{1/2}\widetilde{\boldsymbol{\mathcal{Z}}})(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widehat{W}\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widehat{W}^{1/2})$. Thus, $h_{ii} = Q(\mathbf{X}_i^t\widehat{\boldsymbol{\beta}})\,\mathbf{X}_i^t\mathbf{A}_n^{-1}(\widehat{\beta})\mathbf{X}_i$.

The "unbiased" sandwich estimator is defined similarly to (5) but with the term $\{Y_i - \mu(\mathbf{X}_i^t\boldsymbol{\beta})\}^2$ in the definition of $\mathbf{B}_n(\boldsymbol{\beta})$ replaced by $\{Y_i - \mu(\mathbf{X}_i^t\boldsymbol{\beta})\}^2/(1-h_{ii})$.

Make the following definitions: $V_{asymp} = \sigma^2\mathbf{L}^t\mathbf{A}_n^{-1}(\boldsymbol{\beta})\mathbf{L}$; $\mathbf{R}_n = \xi n^{-1}\sum_{i=1}^n g(\mathbf{X}_i^t\boldsymbol{\beta})\mathbf{X}_i$; $g(x) = (\partial/\partial x)\log\{\mathcal{V}(x)\}$; $\epsilon_i = \{Y_i - \mu(\mathbf{X}_i^t\boldsymbol{\beta})\}/\mathcal{V}^{1/2}(\mathbf{X}_i^t\boldsymbol{\beta})$; $q_{in} = \mathbf{X}_i^t\mathbf{A}_n^{-1}(\boldsymbol{\beta})\mathbf{L}$; $a_n = \mathbf{L}^t\mathbf{A}_n^{-1}(\boldsymbol{\beta})\mathbf{L}$; $\mathbf{C}_n = n^{-1}\sum_{i=1}^n q_{in}^2 Q^{(1)}(\mathbf{X}_i^t\boldsymbol{\beta})\mathbf{X}_i$ and

$$\begin{aligned}
\boldsymbol{\ell}_{in} &= \mathbf{A}_n^{-1}(\boldsymbol{\beta})\mathbf{X}_i\mu^{(1)}(\mathbf{X}_i^t\boldsymbol{\beta})/\mathcal{V}^{1/2}(\mathbf{X}_i^t\boldsymbol{\beta}); \\
v_i &= \{Y_i - \mu(\mathbf{X}_i^t\boldsymbol{\beta})\}^2 M(\mathbf{X}_i^t\boldsymbol{\beta}) - \sigma^2 Q(\mathbf{X}_i^t\boldsymbol{\beta}); \\
\mathbf{W}_n &= n^{-1}\sum_{i=1}^n q_{in}^2\mathcal{V}(\mathbf{X}_i^t\boldsymbol{\beta})M^{(1)}(\mathbf{X}_i^t\boldsymbol{\beta})\mathbf{X}_i.
\end{aligned}$$

In linear regression, we were able to perform exact calculations, and we did not rely on asymptotics. In quasilikelihood models, such exact calculations are not feasible, and asymptotics are required. In what follows, we will treat the $\mathbf{X}$'s as a sample from a distribution, and terms without the subscript n will refer to probability limits. We will not write down formal regularity conditions, but essentially what is necessary is that sufficient moments of the components of $\mathbf{X}$ and Y exist, as well as sufficient smoothness of $\mu(\cdot)$. Under such conditions, at least asymptotically there will be no leverage points, so that the usual and unbiased sandwich estimators will have similar asymptotic behavior. Thus $\mathbf{A}(\boldsymbol{\beta}) = E\{\mathbf{A}_n(\boldsymbol{\beta})\}$, $q = \mathbf{X}^t\mathbf{A}^{-1}(\boldsymbol{\beta})\mathbf{L}$, $a = \mathbf{L}^t\mathbf{A}^{-1}(\boldsymbol{\beta})\mathbf{L}$, $\mathbf{C} = E\{q^2 Q^{(1)}(\mathbf{X}^t\boldsymbol{\beta})\mathbf{X}\}$, etc.

Theorem 4: As $n \to \infty$,

$$\begin{aligned}
n^{1/2}(V_{ql} - V_{asymp}) &\Rightarrow \text{Normal}[0, \Sigma_{ql} = E\{a\xi(\epsilon^2 - \sigma^2) - \sigma^2(a\mathbf{R} + \mathbf{C})^t\boldsymbol{\ell}\epsilon\}^2]; \\
n^{1/2}(V_{sand} - V_{asymp}) &\Rightarrow \text{Normal}[0, \Sigma_{sand} = E\{q^2 v + (\mathbf{W} - 2\sigma^2\mathbf{C})^t\boldsymbol{\ell}\epsilon\}^2].
\end{aligned}$$

The terms V_{ql} and V_{sand} can be computed and compared in a few special cases with a scalar predictor where the slope is of interest, so that $\mathbf{X} = (1, U)^t$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^t$.

- In linear homoscedastic regression, $\mu(x) = x$, $\mathcal{V}(x) = 1$. When U has a symmetric distribution, then simple calculations show that $\Sigma_{sand}/\Sigma_{ql} = \kappa$, the kurtosis of U, i.e., $\kappa = E(U^4)/\{E(U^2)\}^2$. This is the asymptotic version of Theorems 2 and 3.

- In logistic regression, $\mathcal{V}(x) = \mu^{(1)}(x) = Q(x) = \mu(x)\{1-\mu(x)\}$, $\sigma^2 = 1$, $\xi = 0$, $\mathbf{R}_n = 0$, $Q^{(1)}(x) = \mu^{(1)}(x)\{1 - 2\mu(x)\}$. All the terms in Theorem 4 can be computed by numerical integration. We have evaluated the expressions when U has a normal or Laplace distribution, both with variance 1. We varied β_1 while choosing β_0 so that marginally $\text{pr}(Y = 1) = 0.10$.

With $\beta_1 = 0.0, 0.5, 1.0, 1.5$, the asymptotic relative efficiency of the usual information covariance matrix estimate compared to the sandwich estimate when the predictors are normally distributed is $3.00, 2.59, 1.92, 1.62$, respectively. When the predictors have a Laplace distribution, the corresponding efficiencies are $6.00, 4.36, 3.31, 2.57$.

The interesting feature occurs at the null case $\beta_1 = 0$, in which case 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.

- In Poisson loglinear regression, $\mu(x) = \mathcal{V}(x) = \exp(x)$, $\sigma^2 = 1$, $\xi = 0$ and $\mathbf{R}_n = 0$. Here we consider only the null case, so that $\beta_1 = 0$. Then, as sketched in the appendix, if U has a symmetric distribution,

$$\Sigma_{sand}/\Sigma_{ql} = \kappa + 2\kappa \exp(\beta_0).$$

  This is a somewhat surprising result, namely that as the background event rate $\exp(\beta_0)$ increases, at the null case the sandwich estimator has efficiency decreasing to zero.

- More generally, at the null case the role of the kurtosis of the design becomes clear. Let $\mathbf{L}^t\mathbf{L} = 1$ and $\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}} = nI$. Then $Q(x) = Q(\beta_0) = Q$, $A = QI$, $M = Q/\mathcal{V}$, $g = (\partial/\partial\beta_0)\log\{\mathcal{V}(\beta_0)\}$, $\mathbf{R} = \xi g(1,0)^t$, $q = U/Q$, $a = 1/Q$, $\mathbf{C} = (Q^{(1)}/Q^2)(1,0)^t$, $\boldsymbol{\ell} = Q^{-1/2}(1,U)$, $v = (\epsilon^2 - \sigma^2)Q$, $\mathbf{W} = (M^{(1)}\mathcal{V}/Q^2)(1,0)^t$ and thus

$$\begin{aligned}
\Sigma_{ql} &= E\left[\xi(\epsilon^2 - \sigma^2)/Q - \sigma^2\{(\xi g/Q) + (Q^{(1)}/Q^2)\}Q^{-1/2}\epsilon\right]^2; \\
\Sigma_{sand} &= E\left[U^2(\epsilon^2 - \sigma^2)/Q + \{(\mathcal{V}M^{(1)}/Q^2) - (2\sigma^2 Q^{(1)}/Q^2)\}Q^{-1/2}\epsilon\right]^2.
\end{aligned}$$

  The kurtosis of U arises because of fourth moments of U appear in the expression for Σ_{sand}.

4 COVERAGE PROBABILITIES AND ALTERNATIVES IN LINEAR AND LOGISTIC REGRESSION

One would expect that the excess variability of the sandwich estimate would be reflected in undercoverage of confidence intervals. Here we investigate this problem via simulation.

4.1 Linear Models

In most applications of the sandwich method, the formula (1) is used directly and the resulting "studentized" statistic is compared to the normal distribution. In the types of sample sizes considered here, it is fairly clear that this practice makes little sense, since it is the analogue of using the maximum likelihood estimate of the variance and normal instead of t–percentiles. Hence, some adjustment for degrees of freedom is necessary. In what follows, we adjusted the usual sandwich estimator V_{sand} and its associated confidence intervals for degrees of freedom in the standard way. Specifically, if the degrees of freedom are $\mathrm{df} = n - p$ where p is the number of parameters in $\boldsymbol{\beta}$, then the covariance matrix estimator is multiplied by n/df, an ad hoc unbiasing measure, and the percentiles of the "studentized" statistic are compared to a t–distribution with df degrees of freedom. For the unbiased sandwich estimator, we simply used the t–percentiles with df degrees of freedom.

We simulated the simple linear model with homoscedastic normal errors and with an intercept, so that $\mathbf{X}^t = (1, U)$, where U was given three distributions: Normal, Laplace, and $t(3)$, the t–distribution with 3 degrees of freedom. The actual design matrix $\widetilde{\boldsymbol{\mathcal{Z}}}$ thus varied with each simulation. Coverage probabilities were based on $10,000$ simulations, and are of course independent of the parameters in the model. The confidence intervals formed had nominal 95% coverage.

According to the theory, the sandwich estimator becomes increasingly less efficient with a more dispersed design. In Table 1, we see that this has negative consequences. For example, in a sample of size $n = 20$ with the design generated from the Laplace distribution, the usual sandwich method even with $n - p$ degrees of freedom has a coverage probability of only 89.9%, while the unbiased sandwich with $n - p$ degrees of freedom has a coverage of only 91.9%.

There are many possibilities to get better coverage from confidence intervals, while still retaining asymptotically correct coverage in case of heteroscedasticity. A few of these are examined in Table 1.

- A simple method of moments method can be constructed for the linear case. The suggestion is to use the unbiased sandwich formula but to compare it to the t–distribution with $(n-p)/\widehat{\kappa}$ degrees of freedom, where $\widehat{\kappa}$ is the sample kurtosis of the terms $a_i = \mathbf{L}(\widetilde{\boldsymbol{\mathcal{Z}}}^t\widetilde{\boldsymbol{\mathcal{Z}}})^{-1}\mathbf{X}_i^t$. Here is the motivation for this method. We know that $E(V_{ols}) = \sigma^2/n$ and that

$$n(n-p)V_{ols}/\sigma^2 \sim \chi^2(n-p). \tag{6}$$

Also, $E(V_{sand,u}) = E(V_{ols})$. Now, suppose that for some k we have approximately that

$$nkV_{sand,u}/\sigma^2 \sim \chi^2(k). \tag{7}$$

In this case, what would k be? By (6) and (7) $\mathrm{var}(V_{ols}) = 2\sigma^4/\{n^2(n-p)\}$ and $\mathrm{var}(V_{sand,u}) = 2\sigma^4/(n^2k)$. Therefore, the asymptotic relative efficiency of the unbiased sandwich estimate to the ordinary estimate is $\mathrm{ARE}(V_{sand,u}|V_{ols}) = \mathrm{var}(V_{ols})/\mathrm{var}(V_{sand,u}) = k/(n-p) = 1/\kappa$, where

κ is the kurtosis of $\mathbf{X}^t\mathbf{L}$. Since $V_{sand,u}$ is a function of the residuals, it is independent of $\mathbf{L}^t\widehat{\beta}$. Thus $\mathbf{L}^t(\widehat{\beta}-\beta)/V_{sand,u}^{1/2}$ is approximately $t_{(n-p)/\kappa}$, as claimed.

There is an interesting special case where this adjustment to the degrees of freedom gives exact inference. Consider the one-way layout with I populations and equal sample sizes of N, and suppose that we want a confidence interval for the difference between the means of the first two populations. The unbiased sandwich estimator is the pooled estimate using the first two samples only whereas V_{ols} is the pooled estimate using all of the samples. Using $V_{sand,u}$ with the adjusted degrees of freedom, which is $2(N-1)$, gives the usual two-sample t-interval and so will be exact under homoscedasticity. If I is large and N is small, then using $V_{sand,u}$ with $n-p = I(N-1)$ degrees of freedom will give considerable under coverage.

- A second simple method which we have found to work reasonably well is to use the unbiased sandwich estimator but instead of the $t(n-p)$–distribution as a reference, use the $c_n t(n-2-p)$–distribution, where $c_n = \{n/(n-2-p)\}^{1/2}$.

- One can also use the bootstrap where the pairs (Y, U) are resampled; the reason for resampling pairs is that at least theoretically bootstrapping the residuals will not work under heteroscedasticity, although it is the obvious approach for homoscedastic errors. When resampling pairs, our simulations indicate that some variants of the bootstrap do not work very well, e.g., bootstrapping the usual t–statistic assuming homoscedasticity has fairly unsatisfactory coverage even when homoscedasticity holds. Other methods do have reasonable coverage in our simulations, particularly the t–statistic with $n-2$ degrees of freedom when the standard error of the least squares estimate is estimated by the bootstrap standard deviation.

- There is also a theory of likelihood ratio tests modified to be consistent for heteroscedasticity, see Schrader & Hettmansperger (1980) and Kent (1982). These methods have improved coverage probabilities in our simulations by reference to a $c_{n*}^2 \times F(1, n-p)$–distribution, where $c_{n*} = \{n/(n-p)\}^{1/2}$.

Table 1 also gives the average lengths of some selected confidence intervals. The usual sandwich intervals have lower than nominal coverage because they are too short, especially for $n = 10$.

The results we have described are averages with the predictors generated by a particular design. In some cases, however, coverage of unmodified sandwich intervals can be very low. For example, when $n = 10$, consider the design

$$(U_1, ..., U_n) = (0.030, -0.015, 0.006, 0.507, -0.173, 0.526, 0.753, 0.514, 0.554, -2.702).$$

The last observation is a leverage point (it has leverage 0.91), and it is just these sorts of situations where our theory predicts problems. Based on $20,000$ simulations, the exact percentile for the usual sandwich test statistic is 5.44, with Hinkley's adjustment it is 4.87, and for the unbiased sandwich estimator it is 3.54. This is in contrast with the t–percentile with 8 degrees of freedom of 2.31. Our simple adjustment for the unbiased sandwich which uses $c_n t(n-2-p)$–distribution, where $c_n = \{n/(n-2-p)\}^{1/2}$, proposes a percentile of 3.16, not perfect but reasonably close. Comparing the unbiased sandwich to the t–distribution with $(n-p)/\widehat{\kappa}$ degrees of freedom is conservative, with approximately 1.5 degrees of freedom and a percentile of more than 6.00. The usual sandwich with $n-p$ degrees of freedom has actual coverage probability, based on $5,000$ simulations, of only 72.1%, the unbiased sandwich has coverage 86.7%, and the simple adjustment to the unbiased sandwich, namely using $\{n/(n-4)\}^{1/2}t(n-4)$ as the reference distribution, has coverage 93.4%, and the use of $(n-p)/\widehat{\kappa}$ degrees of freedom has coverage 99.9%.

4.2 Logistic Regression

We also simulated the null case of simple linear logistic regression, for samples of size $200, 300, 400$, in two scenarios: (a) when the response rate was 0.05, and (b) when the expected number of events was 15. In all cases, the predictors U were generated to have a Laplace distribution.

We defined an "unbiased" sandwich estimate, which is just (5) *except* that $B_n(\widehat{\boldsymbol{\beta}})$ is adjusted for leverage, i.e.,

$$\mathbf{B}_n(\widehat{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^2 M(\mathbf{X}_i^t \widehat{\boldsymbol{\beta}})\{Y_i - \mu(\mathbf{X}_i^t \widehat{\boldsymbol{\beta}})\}^2/(1-h_{ii}),$$

where the hat matrix for logistic regression is defined for example in Cook & Weisberg (1982, pages 191–192). Because of the old convention that the effective sample size in the binary case is the minimum n_y of the number of events and non–events, we also compared the unbiased sandwich Wald test to the $\{n_y/(n_y-2)\}^{1/2}t(n_y-2)$ distribution, and the unbiased sandwich likelihood ratio test was compared to the square of this distribution.

The results are displayed in Table 2, based on $5,000$ simulations. For comparison purposes, we also display the usual logistic regression Wald test and likelihood ratio test for a null effect due to the predictor, both of which have nearly nominal coverage. Once again we see that the usual sandwich estimate has lower than nominal coverage, "unbiasing" the sandwich helps somewhat, and making a further adjustment for degrees of freedom helps even more.

We reran these simulations when the predictors U were normally distributed. As expected, with the smaller leverage, the coverage probabilities of the sandwich estimators improved considerably, just as happens in the linear regression simulation.

5 DISCUSSION

As an estimator of variance, the sandwich estimator is inefficient compared to a parametric estimator computed at the correct model. We have shown that in linear, logistic and Poisson regression, the sandwich estimator becomes increasingly inefficient as the values of the predictors become heavier–tailed. The sandwich estimator will typically be inefficient if the design includes leverage values.

Simulations show that the inefficiency of sandwich methods, caused largely by leverage values, carries over to lower than nominal coverages for heavier–tailed designs, again when the parametric model is correct.

We do believe that the sandwich estimator is useful in practice, especially for larger sample sizes and/or when a hypothesized parametric model underlying (say) the generalized estimating equation approach is seriously questionable. With small samples, and particularly when the predictors are generated from a heavier–tailed distribution, the increased variability of the sandwich estimator often leads to coverage probabilities lower than the nominal. The solution is perhaps not to discard the sandwich estimator, but on a case–by–case basis two obvious strategies are (a) to change the reference distribution based on simulation experience (as we have done in linear and logistic regression); and (b) test out the many forms of bootstraps consistent against model deviations to determine one which works best in practice.

REFERENCES

Breslow, N. (1990). Test of hypotheses in overdispersion regression and other quasilikelihood models. *Journal of the American Statistical Association*, 85, 565–571.

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

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

Efron, B. (1986). Discussion of the paper by C. F. J. Wu "Jackknife, bootstrap and other resampling methods in statistics". *Annals of Statistics*, 14, 1301–1304.

Firth, D. (1992). Discussion of the paper by Liang, Zeger & Qaqish "Multivariate regression analysis for categorical data". *Journal of the Royal Statistical Society, Series B*, 54, 24–26.

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.

Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, LeCam, L. M. and Neyman, J. editors. University of California Press, pp. 221–233.

Kent, J.T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69, 19-27.

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

Liang, K. Y., Zeger, S. L. & Qaqish, B. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B*, 54, 3–40.

McCullagh, P. (1992). Discussion of the paper by Liang, Zeger & Qaqish "Multivariate regression analysis for categorical data". *Journal of the Royal Statistical Society, Series B*, 54, 24–26.

Schrader, R.M. and Hettmansperger, T.P. (1980). Robust analysis of variance based upon a likelihood criterion. *Biometrika*, 67, 93-101.

Simpson, D. G., Guth. D., Zhou, H. & Carroll, R. J. (1996). Interval censoring and marginal analysis in ordinal regression. *Journal of Agricultural, Biological and Environmental Statistics*, 1, 354–376.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in statistics. *Annals of Statistics*, 14, 1261–1350.

APPENDIX

5.1 Proof of Theorem 2

The inequality in (4) follows from the Cauchy-Schwarz inequality because

$$\sum_{i=1}^{n} a_i^2 \leq n^{1/2} (\sum_{i=1}^{n} a_i^4)^{1/2},$$

which implies that $\sum a_i^4 \geq (\sum a_i^2)^2/n$. Next, we approximate the variance given in (3). Observe that

$$\begin{aligned} \sum_{i=1}^{n} a_i^4 (1-h_{ii})^2 &= \{1+O(\max(h_{jj})\} \sum_{i=1}^{n} a_i^4; \\ \sum_{i \neq j} a_i^2 a_j^2 h_{ij}^2 &\leq \max(h_{jj}^2)(\sum_{i=1}^{n} a_i^2)^2 = o(n^{-1}) O(n \sum_{i=1}^{n} a_i^4) = o(\sum_{i=1}^{n} a_i^4). \end{aligned}$$

It follows that the variance of the sandwich estimator is asymptotically equivalent to $2\sigma^4 \sum a_i^4$.

It remains to show that the bias of the sandwich estimator is asymptotically negligible for regular design sequences. Using (2) we have

$$\sigma^{-4} \text{bias}^2(\text{Sandwich}) \leq \max(h_{ii}^2)(\sum_{i=1}^{n} a_i^2)^2 = o(\sum_{i=1}^{n} a_i^4).$$

Combining these results with the approximation mse(*classical*) $\sim 2\sigma^4 (\sum a_i^2)^2/n$ completes the proof.

5.2 Calculations in the Poisson Case

It is easily verified that $\mathbf{A}(\boldsymbol{\beta}) = \exp(\beta_0)I_2$, where I_2 is the identity matrix. Also, $q = U\exp(-\beta_0)$, $\mathbf{X}^t\boldsymbol{\beta} = \beta_0$, $Q^{(1)}(\mathbf{X}^t\boldsymbol{\beta}) = \exp(\beta_0)$, $\mathbf{C} = \exp(-\beta_0)(1,0)^t$, $\boldsymbol{\ell} = \exp(-\beta_0/2)(1,U)^t$, $\epsilon = \{Y - \exp(\beta_0)\}/\exp(\beta_0/2)$ and hence $\Sigma_{ql} = \exp(-3\beta_0)$.

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, $M(x) = 1$, $M^{(1)}(x) = 0$, $\mathbf{W} = 0$. A detailed calculation then shows that $\Sigma_{sand} = 2\kappa\exp(-2\beta_0)+\kappa\exp(-3\beta_0)$, as claimed.

5.3 Proof of Theorem 4

A standard quasilikelihood expansion gives $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx n^{-1/2}\sum_{i=1}^n \boldsymbol{\ell}_{in}\epsilon_i$, where $\approx$ means that the difference is of order $o_p(1)$. A simple delta–method calculation yields

$$\xi n^{1/2}\{\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}) - \sigma^2\} \approx n^{-1/2}\sum_{i=1}^n \xi(\epsilon_i^2 - \sigma^2) - \sigma^2\mathbf{R}_n^t n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Thus,

$$\begin{aligned}
n^{1/2}(V_{ql} - V_{asymp}) &\approx \xi n^{1/2}\{\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}) - \sigma^2\}a_n + n^{1/2}\sigma^2\mathbf{L}^t\{\mathbf{A}_n^{-1}(\widehat{\boldsymbol{\beta}}) - \mathbf{A}_n^{-1}(\boldsymbol{\beta})\}\mathbf{L} \\
&\approx \xi n^{1/2}\{\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}) - \sigma^2\}a_n - \sigma^2 n^{1/2}\mathbf{L}^t\mathbf{A}_n^{-1}(\boldsymbol{\beta})\{\mathbf{A}_n(\widehat{\boldsymbol{\beta}}) - \mathbf{A}_n(\boldsymbol{\beta})\}\mathbf{A}_n^{-1}(\boldsymbol{\beta})\mathbf{L} \\
&\approx \xi n^{1/2}\{\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}) - \sigma^2\}a_n - \sigma^2\mathbf{C}_n^t n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&\approx n^{-1/2}\sum_{i=1}^n\{a_n\xi(\epsilon_i^2 - \sigma^2) - \sigma^2(a_n\mathbf{R}_n + \mathbf{C}_n)^t\boldsymbol{\ell}_{in}\epsilon_i\},
\end{aligned}$$

which shows the first part of Theorem 4.

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

$$\begin{aligned}
& n^{1/2}(V_{sand} - V_{asymp}) \approx -2\sigma^2 n^{1/2}\mathbf{L}^t\mathbf{A}_n^{-1}(\boldsymbol{\beta})\{\mathbf{A}_n(\widehat{\boldsymbol{\beta}}) - \mathbf{A}_n(\boldsymbol{\beta})\}\mathbf{A}_n^{-1}(\boldsymbol{\beta})\mathbf{L} \\
&\qquad + n^{1/2}\mathbf{L}^t\mathbf{A}_n^{-1}(\boldsymbol{\beta})\{\mathbf{B}_n(\widehat{\boldsymbol{\beta}}) - \sigma^2\mathbf{A}_n(\boldsymbol{\beta})\}\mathbf{A}_n^{-1}(\boldsymbol{\beta})\mathbf{L} \\
\approx\ & -2\sigma^2 n^{-1/2}\sum_{i=1}^n \mathbf{C}_n^t\boldsymbol{\ell}_{in}\epsilon_i + n^{-1/2}\sum_{i=1}^n q_{in}^2[M(\mathbf{X}_i^t\widehat{\boldsymbol{\beta}})\{Y_i - \mu(\mathbf{X}_i^t\widehat{\boldsymbol{\beta}})\}^2 - \sigma^2 Q(\mathbf{X}_i^t\boldsymbol{\beta})] \\
\approx\ & -2\sigma^2 n^{-1/2}\sum_{i=1}^n \mathbf{C}_n^t\boldsymbol{\ell}_{in}\epsilon_i + n^{-1/2}\sum_{i=1}^n q_{in}^2 v_i + n^{-1}\sum_{i=1}^n q_i^2 M^{(1)}(\mathbf{X}_i^t\boldsymbol{\beta})\mathbf{X}_i\{Y_i - \mu(\mathbf{X}_i^t\boldsymbol{\beta})\}^2 n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
\approx\ & -2\sigma^2 n^{-1/2}\sum_{i=1}^n \mathbf{C}_n^t\boldsymbol{\ell}_{in}\epsilon_i + n^{-1/2}\sum_{i=1}^n q_{in}^2 v_i + n^{-1}\sum_{i=1}^n q_i^2 M^{(1)}(\mathbf{X}_i^t\boldsymbol{\beta})\mathbf{X}_i\mathcal{V}(\mathbf{X}_i^t\boldsymbol{\beta})n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
\approx\ & n^{-1/2}\sum_{i=1}^n(-2\sigma^2\mathbf{C}_n^t\boldsymbol{\ell}_{in}\epsilon_i + q_i^2 v_i + \mathbf{W}_n^t\boldsymbol{\ell}_{in}\epsilon_i),
\end{aligned}$$

as claimed.

Method	Normal n=10	Normal n=20	Normal n=30	Laplace n=10	Laplace n=20	Laplace n=30	t(3) n=10	t(3) n=20	t(3) n=30
Usual Sandwich, df=n	0.856	0.906	0.912	0.830	0.881	0.899	0.829	0.865	0.889
Usual sandwich, df=$n-2$	0.896	0.923	0.923	0.878 (1.05)	0.899 (0.66)	0.910 (0.52)	0.870	0.883	0.893
Unbiased sandwich, df=$n-2$	0.914	0.930	0.931	0.905 (1.20)	0.919 (0.71)	0.922 (0.55)	0.907	0.909	0.914
Usual Sandwich, df=$n-6$	0.981	0.949	0.944	0.971	0.939	0.930	0.969	0.920	0.916
Unbiased Sandwich, df=$n-4$ multiplicative adjustment	0.968	0.951	0.947	0.958 (1.64)	0.946 (0.80)	0.938 (0.60)	0.956	0.933	0.931
Unbiased Sandwich, df uses kurtosis	0.951	0.954	0.945	0.956	0.954	0.956	0.956	0.963	0.954
Bootstrap–t usual sandwich	0.937	0.940	0.940	0.916 (1.75)	0.919 (0.85)	0.923 (0.62)	0.916	0.915	0.905
Bootstrap–t unbiased sandwich	0.929	0.933	0.937	0.928 (1.99)	0.927 (0.92)	0.931 (0.66)	0.930	0.915	0.919
Bootstrap–t parametric se	0.942	0.930	0.930	0.931 (1.40)	0.920 (0.71)	0.920 (0.54)	0.932	0.909	0.909
Usual t–test with bootstrap se	0.965	0.944	0.937	0.971 (1.62)	0.953 (0.78)	0.943 (0.58)	0.971	0.952	0.948
Usual LR Test	0.955	0.952	0.947	0.955	0.954	0.949	0.955	0.951	0.948
Sandwich LR test, df=$n-2$	0.946	0.946	0.942	0.930	0.928	0.928	0.923	0.911	0.912
Unbiased sandwich LR test, df=$n-2$	0.952	0.946	0.941	0.941	0.934	0.933	0.944	0.925	0.925
Unbiased Sandwich LR test df=$n-2$, and with multiplicative adjustment	0.981	0.957	0.951	0.953	0.940	0.937	0.970	0.940	0.933

Table 1: *Results of a simulation study for simple linear regression. Tabulated values are coverage probabilities in based on* $10,000$ *simulations. Distributions are for the predictor variable* U*. Average lengths of some selected confidence intervals are in parentheses.*

Method	p=.05 n=200 np=10	p=.05 n=300 np=15	p=.05 n=400 np=20	p=.075 n=200 np=15	p=.05 n=300 np=15	p=.0375 n=400 np=15
Logistic Wald Test, df=n	.945	.949	.943	.948	.949	.951
Logistic LR test, df=n	.951	.955	.947	.949	.955	.956
Usual sandwich Wald Test, df=n	.887	.910	.914	.917	.910	.906
"Unbiased" sandwich Wald Test	.905	.921	.920	.928	.921	.914
"Unbiased" sandwich Wald Test, df=$n_y - 2$	.980	.961	.954	.970	.961	.955
Sandwich LR test, df=n	.895	.919	.922	.919	.919	.914
"Unbiased" Sandwich LR test	.912	.926	.928	.929	.926	.923
"Unbiased" Sandwich LR test, df=$n_y - 2$	.971	.955	.953	.962	.955	.956

Table 2: *Results of a simulation study for simple logistic regression. Tabulated values are coverage probabilities in based on* $5,000$ *simulations. The predictor variable* U *has a Laplace distribution. Here "Unbiased" refers to a leverage adjustment to the residuals, while* n_y *is the number of events.*