LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK
SONDERFORSCHUNGSBEREICH 386

Kauermann, Carroll:

The Sandwich Variance Estimator: Efficiency Properties and Coverage Probability of Confidence Intervals

Sonderforschungsbereich 386, Paper 189 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

gsf

MAX-PLANCK-GESELLSCHAFT

TUM

The Sandwich Variance Estimator: Efficiency Properties and Coverage Probability of Confidence Intervals

Göran Kauermann Raymond J. Carroll *

February 28, 2000

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 the fitted 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. Under certain circumstances we show that when the quasilikelihood 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 one pays to obtain consistency even when the parametric model fails. 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 estimates combined with t-distribution quantiles gives confidence intervals with coverage probability falling below the nominal value. We propose a simple adjustment to compensate this defect, where the adjustment explicitly considers the variance of the sandwich estimate.

Keywords: Coverage probability; Generalized estimating equations; Generalized linear models; Heteroscedasticity; Linear regression; Marginal Models; Quasilikelihood; Robust covariance estimator; Sandwich estimator.

Short Title: The Sandwich Estimator

*Göran Kauermann (kauerman@stat.uni-muenchen.de), Institute of Statistics; Ludwig-Maximilians-University Munich; Akademiestr. 1; 80796 Munich, Germany. Raymond J. Carroll (carroll@stat.tamu.edu), Department of Statistics, Texas A & M University, College Station, TX 77843-3143, USA. Carroll's research was supported by a grant from National Cancer Institute (CA–57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ESO9106). The authors also acknowledge the support received from the Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 386). Parts of this paper contain material in an unpublished technical report written with Douglas Simpson, Arny Stromberg, Suojin Wang and David Ruppert, whose help is gratefully acknowledged.

1 Introduction

Sandwich variance estimators are a common tool used for variance estimation of parameter estimates. Originally introduced by Huber (1967) and White (1982), the method is now widely used in the context of generalized estimating equations, see e.g. Liang & Zeger (1986), Liang, Zeger & Qaqish (1992) and Diggle, Liang & Zeger (1994). Efficient estimation of parameters in this setting requires the specification of a 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 without making distributional assumptions; and even if the assumed model underlying the parameter estimates is incorrect. Because of this desirable model–robustness property, the sandwich estimator is often called the *robust covariance matrix* estimator, or the *empirical covariance matrix* estimator. The argument in favor of the sandwich estimate is that asymptotic normality and proper coverage confidence intervals require only a consistent variance estimate, so there is no great need to construct a highly accurate covariance matrix estimate. This consistency however has its price in an increase of the variability, i.e. sandwich variance estimators generally have a larger variance than model based classical variance estimates. In his discussion of the paper by 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”. An earlier discussion about small sample improvements for the sandwich estimate is found in MacKinnon & White (1985), who propose jackknife sandwich estimates.

The objectives of this paper are twofold, first we investigate the sandwich estimate in terms of efficiency and secondly we analyze the effect of the increased variability of the sandwich estimate on the coverage probability of confidence intervals. For efficiency we derive asymptotic as well as 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 holds in generalized linear models as well. For example, in simple linear logistic regression, at the null value where 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 probability of confidence intervals built from sandwich variance estimates is discussed in the second part of the paper. Simulation studies given by Wu (1986) and Breslow (1990) report somewhat elevated levels of Wald–type tests based on the sandwich estimator. Rothenberg (1988) derives 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, i.e. the coverage probability falls below the nominal value. We show that undercoverage is mainly determined by the variance of the variance estimate. To correct this deficit we present a simple adjustment which depends on normal distribution quantiles and the variance of the sandwich variance estimate.

The paper is organized as follows. In Section 2 we compare the sandwich estimator with the usual parametric regression estimator in the homoscedastic linear regression model. Section 3 gives a discussion of the sandwich estimate for quasilikelihood and generalized estimating equations (GEE). Some simulations are presented in Section 4 where we suggest a simple adjustment which improves coverage probability. Section 5 contains concluding remarks. Proofs and general statements are given in the Appendix.

2 Linear Regression

2.1 The Sandwich Estimator

Consider the simple linear regression model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$, $i = 1, \ldots, n$ where $\mathbf{x}_i^{\mathrm{T}}$ are $1 \times p$ dimensional vectors of covariates and $\epsilon_i \sim N(0, \sigma^2)$. Let $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ 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 $\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$, where $\mathbf{z}^{\mathrm{T}}$ is a $1 \times p$ dimensional contrast vector of unit length, i.e. $\mathbf{z}^{\mathrm{T}}\mathbf{z} = 1$. The variance of $\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ is given by $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}$ which can be estimated by the classical model based variance estimator $V_{model} = \hat{\sigma}^2\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}$ where $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\epsilon}_i^2/(n-p)$ with $\hat{\epsilon}_i = Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ as fitted residuals. A major assumption implicitly used in the calculation of V_{model} is that the errors ϵ_i are homoscedastic. If this assumption is violated V_{model} does not provide a consistent variance estimate. In contrast even if the errors are not homoscedastic the sandwich variance estimate

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

consistently estimates $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$, where $a_i = \mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_i$. In linear regression, (1) is often multiplied by $n/(n-p)$ (Hinkley, 1977) to reduce the bias.

2.2 Properties of Sandwich Estimator

Let h_{ii} be the i-th digonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} = (h_{ij})$. Under homoscedasticity, $E(\hat{\epsilon}_i^2) = \sigma^2(1-h_{ii})$. It then follows that

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

where $b_n = \sum_{i=1}^n h_{ii}a_i^2/\sum_{i=1}^n a_i^2 \leq \max_{1\leq i\leq n} h_{ii}$. Since $b_n \geq 0$ one obtains that in general the sandwich estimator is biased *downward.* The bias thereby depends on the design of $\mathbf{x}_i$ and it can be substantial when there are leverage points. Bias problems can be avoided by replacing $\hat{\epsilon}_i$ in (1) by $\tilde{\epsilon}_i = \hat{\epsilon}_i/(1-h_{ii})^{1/2}$. The resulting estimator is refered to as the unbiased sandwich variance estimator and denoted by $V_{sand,u}$ (Wu, 1986, equation 2.6). It is easily seen that $E(V_{sand,u}) = \mathrm{var}(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$. Since $\mathrm{var}(\widetilde{\epsilon}_i^2) = 2\sigma^4$ and $\mathrm{cov}(\widetilde{\epsilon}_i^2, \widetilde{\epsilon}_j^2) = 2\widetilde{h}_{ij}^2\sigma^4$ for $i \neq j$, where $\widetilde{h}_{ij} = h_{ij}/\{(1-h_{ii})(1-h_{jj})\}^{1/2}$, it follows that

$$\mathrm{var}(V_{sand,u}) = \sum_{i=1}^{n} a_i^4 \mathrm{var}(\widetilde{\epsilon}^2) + \sum_{i \neq j} a_i^2 a_j^2 \mathrm{cov}(\widetilde{\epsilon}_i^2, \widetilde{\epsilon}_j^2) = 2\sigma^4 \sum_{i=1}^{n} a_i^4 + 2\sigma^4 \sum_{i \neq j} a_i^2 a_j^2 \widetilde{h}_{ij}^2. \qquad (3)$$

We now compare the variance (3) to the variance of the model based variance estimator $\mathbf{V}_{model}$ which equals $\mathrm{var}(V_{model}) \approx 2\sigma^4\{\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}\}^2/n = 2\sigma^4(\sum a_i^2)^2/n$.

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

$$\frac{\mathrm{var}(\mathbf{V}_{sand,u})}{\mathrm{var}(\mathbf{V}_{model})} \quad \geq \quad \{n^{-1}\sum_{i=1}^{n} a_i^4\}\{n^{-1}\sum_{i=1}^{n} a_i^2\}^{-2} \geq 1. \qquad (4)$$

The proof follows directly from the Cauchy Schwarz inequality. Theorem 1 states that the sandwich estimate is less efficient when the model is correct, i.e. when the errors are homescedastic.

Example 1 (the intercept): Suppose the first column of $\mathbf{X}$ is a vector of ones, the other columns have means of zero, and $\mathbf{z}^{\mathrm{T}} = (1, 0, \ldots, 0)$. 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 $\mathbf{x}_i^{\mathrm{T}} = (1, u_i)$ where $\sum u_i = 0$. Suppose $\mathbf{z} = (0, 1)$ so that $\widehat{\beta}_1 = \mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ 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 κ_n^{-1}, where $\kappa_n = n^{-1}\sum u_i^4/(n^{-1}\sum_{i=1}^n u_i^2)^2 \geq 1$. Note that κ_n is the sample kurtosis of the design points u_i. For instance if the design points $(u_1, ..., u_n)$ were realizations of a normal distribution, $\kappa_n \to 3$ and hence the sandwich estimator $V_{sand,u}$ has 3 times the variability of the usual model based estimator V_{model}. If the design points were generated from a Laplace distribution, the usual sandwich estimator is 6 times more variable.

The examples above show that the use of sandwich variance estimates in linear models can lead to a substantial loss of efficiency. A similar phenomena occurs in nonlinear models as discussed in the next section.

3 Quasilikelihood and Generalized Estimating Equations

3.1 The Sandwich Estimate

In the following section we consider the sandwich variance estimate in generalized estimating equations (GEE). Let $Y_i = (y_{i1}, \ldots, y_{im})^{\mathrm{T}}$, be a random vector taken at the i-th unit, for $i = 1, \ldots, n$. The components of Y_i are allowed to be correlated while observations taken at two different units are independent. 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. We assume that the variance matrix of Y_i depends on the mean of Y_i, i.e. $\mathrm{var}(Y_i|\mathbf{X}_i) = \sigma^2 V(\mu_i) =: \sigma^2 \mathbf{V}_i$ where $\mu_i = h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$, $V(\cdot)$ is a known variance function and σ^2 is a dispersion scalar which 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 referred to as marginal models, see e.g. Diggle et al. (1994) and references given there. If Y_i is a scalar, i.e. if $m = 1$, models of this type are also known as quasilikelihood models (Wedderburn, 1974) or generalized linear models (McCullagh & Nelder, 1989). The parameter $\boldsymbol{\beta}$ can be estimated using the generalized estimating equation (e.g. Liang & Zeger, 1986 or Gourieroux, Monfort and Trognon. 1984)

$$0 = \sum_i \frac{\partial \mu_i^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} (Y_i - \mu_i). \tag{5}$$

In the previous section, we were able to perform exact calculations. In quasilikelihood models, such exact calculations are not feasible, and asymptotics are required. 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 $h(\cdot)$. Under such conditions a Taylor expansion of (5) about the true parameter $\boldsymbol{\beta}$ provides the first order approximation

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;=\; \boldsymbol{\Omega}^{-1} \sum_i \frac{\partial \mu_i^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(Y_i - \mu_i) + O_p(n^{-1}), \tag{6}$$

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, i.e. $\sigma^2 \mathbf{V}_i = \mathrm{var}(Y_i|\mathbf{X}_i)$, one gets $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{z}\sigma^2$ in first order approximation. Hence we can estimate $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$ by $\mathbf{V}_{model} := \widehat{\sigma}^2 \mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{z}$ where $\widehat{\boldsymbol{\Omega}}$ is a simple plug in estimate of $\boldsymbol{\Omega}$ and $\widehat{\sigma}^2$ an estimate of the dispersion parameter, if this is unknown. However in practice the covariance $\mathrm{var}(Y_i|\mathbf{X}_i)$ may not be known so that $\mathbf{V}_i$ serves as prior estimate of the covariance in (5). In this case $\mathbf{V}_i$ is called the working covariance and the variance $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$ can be estimated by the sandwich formula

$$\mathbf{V}_{sand} = \mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}^{-1} \left(\sum_i \frac{\partial \widehat{\mu}_i^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \widehat{\mathbf{V}}_i^{-1} \widehat{\boldsymbol{\epsilon}}_i \widehat{\boldsymbol{\epsilon}}_i^{\mathrm{T}} \widehat{\mathbf{V}}_i^{-1} \frac{\partial \widehat{\mu}_i}{\partial \boldsymbol{\beta}} \right) \widehat{\boldsymbol{\Omega}}^{-1}\mathbf{z} \tag{7}$$

where $\widehat{\boldsymbol{\epsilon}}_i = Y_i - \widehat{\mu}_i = Y_i - h(\mathbf{X}_i\widehat{\boldsymbol{\beta}})$ are the fitted residuals and the hat notation refers to simple plug in estimates. The fitted residuals can be expanded by $\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{\epsilon}_i - \partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\{1 + O_p(n^{-1/2})\}$ which gives with (6) $E(\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{\epsilon}}_i^{\mathrm{T}}) = \sigma^2\mathbf{V}_i - \sigma^2 \partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}})\, \boldsymbol{\Omega}^{-1}\, \partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta})\{1 + O(n^{-1})\}$, assuming that $\mathbf{V}_i$ correctly specifies the covariance, i.e. $E(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}}) = \sigma^2\mathbf{V}_i$. Since $\partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}})\, \boldsymbol{\Omega}^{-1}\, \partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta})$ is positive definite one finds the sandwich estimate $\mathbf{V}_{sand}$ to be biased downward. As in the previous section the bias can be corrected. With $\mathbf{H}_{ii} = \partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}})\boldsymbol{\Omega}^{-1}\partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta})\mathbf{V}_i^{-1}$ we define the the leverage–adjusted residual $\widetilde{\boldsymbol{\epsilon}}_i = (\mathbf{I} - \mathbf{H}_{ii})^{-\frac{1}{2}}\widehat{\boldsymbol{\epsilon}}_i$ with $\mathbf{I}$ as identity matrix. Replacing now $\widehat{\boldsymbol{\epsilon}}$ in (7) by $\widetilde{\boldsymbol{\epsilon}}$ gives the bias reduced sandwich estimate $\mathbf{V}_{sand,u}$ which fulfills $E(\mathbf{V}_{sand,u}) = \mathrm{var}(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\{1 + O(n^{-1})\}$, assuming that the variance is correctly specified.

3.2 Examples

In quasilikelihood and generalized estimating equations, variance estimates have two different sources of stochastic variation. The first occurs due to the estimation of the dispersion parameter σ^2, if this is unknown. The second occurs due to the use of plug in estimates, which applies if the variance function $V(\mu)$ depends on the mean or if the link function $h(\cdot)$ is not canonical. In the following two examples we investigate how the latter source affects the loss of efficiency for the sandwich variance estimate. We consider a Poisson and Binomial model where $\sigma^2 \equiv 1$ is fixed, but variability of the variance estimate occurs due to plug in estimation. A general theoretical derivation is given in the appendix.

Example 3 (Poisson loglinear 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 as scalar, $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ and y_i being 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, $\mathrm{var}(\mathbf{V}_{sand})/\mathrm{var}(\mathbf{V}_{model}) = \kappa_n\{1 + 2\exp(\beta_0)\}$ 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 2 above. 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 zero.

Example 4 (Logistic Regression): Let y_i be binary with $E(y_i|\mathbf{x}_i) = \mathrm{logit}^{-1}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ with $\mathbf{x}_i$ as described above. Again, the slope β_1 is the parameter of interest. We vary β_1 while choosing β_0 so that marginally $E(y|\mathbf{x}) = 0.10$. With $\beta_1 = 0.0, 0.5, 1.0, 1.5$, the asymptotic relative efficiency $\mathrm{var}(\mathbf{V}_{sand})/\mathrm{var}(\mathbf{V}_{model})$ varies for u_i standard normally distributed as $3.00, 2.59, 1.92, 1.62$, respectively. When u_i comes from a Laplace distribution (with unit variance), the corresponding efficiencies are $6.00, 4.36, 3.31, 2.57$. 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 above two examples show that the loss of efficiency of the sandwich variance estimate in non-normal models differs from and can be worse than compared to normal models.

4 Confidence Intervals based on Sandwich Variance Estimates

4.1 The Property of Undercoverage

In the following section we investigate the effect of the additional variability of the sandwich variance estimate on the coverage probability of confidence intervals. As one would expect, the excess variability of the sandwich estimate is directly reflected in undercoverage of confidence intervals. We concentrate on normal response models. Let $\theta = \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}$ be the unknown parameter of interest with $\widehat{\theta} = \mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \sim N(\theta, \sigma^2/n)$ as unbiased estimate of θ based on a random sample of size n. We consider the symmetric $1 - \alpha$ confidence intervals $CI(\sigma^2, \alpha) := [\widehat{\theta} \pm z_p \sigma/\sqrt{n}]$ 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 $\widehat{\sigma}^2$ it is well known that the confidence interval $CI(\widehat{\sigma}^2, \alpha)$ shows undercoverage and typically t-distribution quantiles are used instead of normal quantiles. The following theorem shows how the variance of the variance estimate $\widehat{\sigma}^2$ directly affects affect the undercoverage.

Theorem 2: *Under the assumptions above and assuming that $\widehat{\sigma}^2$ and $\widehat{\theta} - \theta$ are independent, the coverage probability of the $1 - \alpha$ confidence interval $CI(\widehat{\sigma}^2, \alpha)$ equals*

$$P\{\theta \in CI(\widehat{\sigma}^2, \alpha)\} \quad = \quad 1 - \alpha - \phi(z_p)\mathrm{var}(\widehat{\sigma}^2)\left(\frac{z_p^3 + z_p}{8\sigma^4}\right) + O(n^{-3/2}) \qquad (8)$$

where $\phi(\cdot)$ is the standard normal distribution density.

The proof of the theorem is given in the appendix. One should note that the postulated assumption of independence of $\widehat{\sigma}^2$ and $\widehat{\theta} - \theta$ holds in a normal regression model if $\widehat{\sigma}^2$ is calculated from fitted residuals. Hence it holds for sandwich variance estimates. It is seen from (8) that the coverage probability of the confidence interval falls below the the nominal value. Moreover, the undercoverage increases linearly with the variance of the variance estimate $\widehat{\sigma}^2$. Using the results of Theorem 1 we therefore find that the sandwich variance estimator can be expected to have lower coverage probability for confidence intervals than for model based variance estimates. Moreover, $t-$distribution quantiles instead of normal quantiles do not correct the undercoverage. The result in Theorem 2 resembles that given in Rothenberg (1988, p. 1005) who derives an adjustment for the distribution function of the t statistic based on sandwich variance estimates. In contrast to Rothenberg, Theorem 2 points out the distinct role of the variance of the variance estimate which is used in the following section to correct the undercoverage.

4.2 A Simple Coverage Adjustment

Formula (8) can be employed to construct a simple coverage correction for confidence intervals. Instead of using the quantile z_p directly we suggest choosing $\widetilde{p} > p$ and make use of the $z_{\widetilde{p}}$ quantile. We thereby select $\widetilde{p}$ such that $P(\theta \in [\hat{\theta} \pm z_{\widetilde{p}}\widehat{\sigma}/\sqrt{n}]) = p$ holds, i.e. using (8) we solve

$$p \;=\; \widetilde{p} - \phi(z_{\widetilde{p}})\mathrm{var}(\widehat{\sigma}^2)\frac{z_{\widetilde{p}}^3 + z_{\widetilde{p}}}{8\sigma^4} \tag{9}$$

for $\widetilde{p}$ iteratively which is easily carried out numerically.

Example 5 (t-distribution quantiles): We demonstrate the above correction 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 mean is estimated by $\widehat{\mu} = \sum_i^n y_i/n$ so that $n^{1/2}(\widehat{\mu} - \mu)$ is $N(0, \sigma^2)$ distributed. The variance σ^2 in turn is estimated by $\widehat{\sigma}^2 = \sum_i^n (y_i - \widehat{\mu})^2/(n-1)$. Exact quantiles for confidence intervals based on the estimates $\widehat{\mu}$ and $\widehat{\sigma}^2$ are available from t-distribution quantiles with $n-1$ degrees of freedom. Approximative quantiles $z_{\widetilde{p}}$ follow from solving (9) using $\mathrm{var}(\widehat{\sigma}^2) = 2\sigma^4/(n-1)$. One should note that the unknown variance in (9) cancels out so that estimation of σ^2 is not required for the calculation of $z_{\widetilde{p}}$. In Table 1 we compare the exact quantiles with the corrected version $z_{\widetilde{p}}$. Even for small sample sizes the corrected quantiles $z_{\widetilde{p}}$ are distinctly close to the exact t-distribution quantiles. This also shows in the true coverage probability $P(\hat{\theta} \leq \theta + z_{\tilde{p}}\hat{\sigma}/\sqrt{n})$ of the confidence intervals and demonstrates that the adjustment applied in a standard setting behaves convincingly well.

We now return to the quasilikelihood and generalized estimating equation approach. We will neglect the effect of plug in estimates in the following and concentrate on normal response models first. For the calculation of the variance of $\mathbf{V}_{sand,u}$ it is helpful to write (7) 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 $\boldsymbol{\epsilon} = \mathbf{Y} - \boldsymbol{\mu}$ and with $\mathbf{P}$ we 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

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

with $\mathrm{diag}_m(\mathbf{V}_i^{-1})$ denoting the block diagonal matrix having $\mathbf{V}_i^{-1}$ on its diagonal, $i = 1, \ldots n$. Note that for $m \equiv 1$ other versions of the hat matrix have been suggested (see Cook & Weisberg, 1982, pages 191–192, for logistic regression or Carroll & Ruppert, 1987, page 74, for other models). Let $\mathbf{W}$ be the block diagonal matrix $\mathbf{W} = \mathrm{diag}_m(\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i)$ 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 $\mathbf{D} = \mathrm{diag}_m(\mathbf{I} - \mathbf{H}_{ii})^{-1/2}$ we get the leverage adjusted fitted residuals $\widetilde{\boldsymbol{\epsilon}} = \mathbf{D}\{\mathbf{Y} - \widehat{\boldsymbol{\mu}}\} = \mathbf{DP}(\mathbf{Y} - \boldsymbol{\mu})\{1 + O_p(n^{-1/2})\}$. As above we use the hat notation to denote plug in estimates. This allows us to write

$$\mathbf{V}_{sand,u} \quad = \quad \widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}} \widehat{\mathbf{W}} \widetilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}^{\mathrm{T}} (\mathbf{P}\, \widehat{\mathbf{D}}\, \widehat{\mathbf{W}}\, \widehat{\mathbf{D}}\, \mathbf{P}) \boldsymbol{\epsilon} = \dot{\boldsymbol{\epsilon}}^{\mathrm{T}} \widehat{\mathbf{M}} \dot{\boldsymbol{\epsilon}}\; \{1 + O(n^{-1})\}, \tag{10}$$

where $\mathbf{M} = \sigma^2 \mathrm{diag}_m(\mathbf{V}_i^{1/2})\, \mathbf{P}\, \mathbf{D}\, \mathbf{W}\, \mathbf{D}\, \mathbf{P}\, \mathrm{diag}_m(\mathbf{V}_i^{1/2})$ and $\dot{\boldsymbol{\epsilon}}^{\mathrm{T}} = (\dot{\boldsymbol{\epsilon}}_1^{\mathrm{T}}, \ldots, \dot{\boldsymbol{\epsilon}}_n^{\mathrm{T}})$ independent, homoscedastic residuals defined by $\dot{\boldsymbol{\epsilon}}_i = \mathbf{V}_i^{-1/2} \boldsymbol{\epsilon}_i / \sigma$, where we assumed 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, l-th 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

$$\mathrm{var}(\mathbf{V}_{sand,u}) \quad = \quad 2\mathrm{trace}(\mathbf{MM}) + \sum_k \{E(\dot{\epsilon}_k^4) - 3\} m_{kk}^2. \tag{11}$$

If the $(\dot{\epsilon}_k)$ are standard normal, (11) simplifies to $\mathrm{var}(\mathbf{V}_{sand,u}) = 2\mathrm{tr}(\mathbf{MM})$. The variance of the sandwich variance estimate again depends distinctly on the design of the covariates due to $\partial\mu_i^{\mathrm{T}}/\partial\boldsymbol{\beta} = \mathbf{X}_i\partial h(\eta)/\partial\eta$ with $\eta = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$.

The coverage adjustment (9) can now easily be adopted to sandwich variance estimates. Assume $n^{1/2}\mathbf{z}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(0, \sigma^2)$ with $\widehat{\sigma}^2 = n\mathbf{V}_{sand,u}$ and variance $\mathrm{var}(\widehat{\sigma}^2) = n^2\mathrm{var}(\mathbf{V}_{sand,u})$ calculated from (11). Inserting this into (9) directly gives the adjusted quantile $z_{\widetilde{p}}$ which is used to get the $(1-\alpha)$ confidence interval $[\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \pm z_{\widetilde{p}}\mathbf{V}_{sand,u}^{1/2}]$ with $\alpha = 2(1-p)$.

Simulation 1 (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 $\boldsymbol{\beta} = (0.5, 0.5)^{\mathrm{T}}$ and consider $\beta_1 = (0,1)\boldsymbol{\beta}$ as 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) normally, (b) uniformly or (c) from a Laplace distribution. Table 2 shows simulated coverage probabilities for 2000 simulations for the $p = 0.9$ confidence interval. Working independence is used for fitting $\boldsymbol{\beta}$, but Y_i is simulated from two settings, (i) with covariance $\mathrm{var}(Y_i) = \sigma^2\mathbf{I}$, i.e. correctly specified working covariance, and (ii) with $\mathrm{var}(Y_i) = \sigma^2(3/4\,\mathbf{I} + 1/4\,1_m 1_m^{\mathrm{T}})$, i.e. correlated observations. Drawings from the latter settings are shown as slanted numbers. For comparison we report the coverage probabilities if t-distribution quantiles with $n-2$ degrees of freedom are used. Moreover we also report the coverage probability based on t-distribution quantiles and the jackknife estimate as suggested by MacKinnon & White (1985, formula 13). Assuming working independence and normality their jackknife estimate becomes

$$\mathbf{V}_{jack} = \frac{n-1}{n}\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\sum_i \mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}_i\widetilde{\boldsymbol{\epsilon}}_i^{\mathrm{T}}\mathbf{X}_i)\ (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z} - \frac{n-1}{n^2}\widehat{\gamma}^{\mathrm{T}}\widehat{\gamma}, \tag{12}$$

where $\widehat{\gamma} = \mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}$.

For all three designs our proposed adjustment shows satisfactory behavior. The misspecification of the covariance thereby hardly has an effect on the coverage probability so that the adjustment appears promising also for misspecified models. In contrast, both $t_{p,n-2}$ distribution quantiles and jackknife estimates show undercoverage, although the jackknife approach behaves more accurately, as already described in MacKinnon & White (1985).

The simulation above shows that undercoverage can be severe and should be corrected if covariates vary between units. For individually balanced covariates on the other hand undercoverage is not an issue as seen from the following example.

Example 6 (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 $m \times p$ 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

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

The lower bound is thereby 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 for instance if the individual design $\mathbf{X}_i$ does not differ among the individuals. In this case one gets the lower bound $\mathrm{var}(\mathbf{V}_{sand,u}) = 2\sigma^4/\{n^2(n-1)\}\{1+O(n^{-1})\}$ which equals the variance of the classical variance estimate discussed in Example 5. Hence, one finds that in general $z_{\widetilde{p}} \geq t_{p,n-1}$ holds asymptotically, where the lower bound is reached if the design is individually balanced. As consequence, undercoverage is not an issue in this case.

In a final simulation we show how the adjustment behaves for non-normal data.

Simulation 2 (Logistic and Poisson regression): It should be noted that the adjustment for normal data depends only on the design but not on μ or σ^2. This property does not hold for non-normal data since $\mathbf{V}_i$ typically depends on the mean μ_i. The calculation of $\mathrm{var}(\mathbf{V}_{sand,u})$

therefore requires some plug-in estimates. In contrast to Section 3 we here neglect the effect of plug in estimates and approximate the variance of the sandwich variance estimate by (11). We make use of the adjusted quantiles $z_{\tilde{p}}$ from (9), where again plug in estimates are used to calculate $z_{\tilde{p}}$. We simulate (independent) binomial data with predictor $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ $\boldsymbol{\beta} = (0.5, 0.5)^T$ and Poisson data with $\boldsymbol{\beta} = (1, 1)^{\mathrm{T}}$. The covariates $\mathbf{X}_i$ are distributed as in Simulation 1 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 version of (12). The results are given in Table 3. The corrected adjustment shows slight overcoverage which results from neglecting the effect of the plug-in estimates. For Poisson response and Laplace distributed covariates the adjustment can not entirely compensate the undercoverage. The use of t-distribution quantiles in all cases clearly implies undercoverage. The jackknife estimate behaves comparably to our approach in this example.

5 Discussion

We have shown that sandwich variance estimates are less efficient than model based variance estimates. The loss of efficiency depends on the design and for standard cases it is proportional to the inverse of the kurtosis of the design points. For non-normal 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. A simple adjustment which depends on the design was suggested, which appears to have promising behavior.

A Technical Details

A.1 Sandwich Estimates in Quasilikelihood and Generalized Estimating Equations

Below we derive the relative efficiency in quasilikelihood 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 $1 \times p$ vector. The variance of y_i is given by $\mathrm{var}(y_i|\mathbf{x}_i) = \sigma^2 V\{\mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}$ where $V(\cdot)$ is a known variance function. In some problems, σ^2 is estimated, which we indicate by setting $\xi = 1$, while when σ^2 is known we set $\xi = 0$. We denote the derivatives of functions by superscripts, e.g. $\mu^{(l)}(\eta) = \partial^l \mu(\eta)/(\partial\eta)^l$. Let us assume that the variance is correctly specified, i.e. $\mathrm{var}(y_i|\mathbf{x}_i) = \sigma^2 V\{\mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}$, so that with expansion (6) we get $\mathrm{var}(n^{1/2}\widehat{\mathbf{z}}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{V}_{asymp}\{1 + O(n^{-1})\}$ where $\mathbf{V}_{asymp} = \sigma^2 \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z}$ 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 $n^{1/2}\widehat{\mathbf{z}}^{\mathrm{T}}\boldsymbol{\beta}$ is $\mathbf{V}_{model} = \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}})\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{z}$, where

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

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 $n^{1/2}\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ is written as $V_{sand} = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{B}_n(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{z}$.

For the derivation of 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 $g(\eta) = (\partial/\partial\eta)\log\{V(\eta)\}$; $\epsilon_i = \{y_i - \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}/V^{1/2}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$; $q_{in} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z}$; $a_n = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z}$; $\mathbf{C}_n = n^{-1}\sum_{i=1}^n q_{in}^2 Q^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}_i$ and

$$\begin{aligned}
\boldsymbol{\ell}_{in} &= \boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{x}_i\mu^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})/V^{1/2}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta});\\
v_i &= \{y_i - \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^2 M(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \sigma^2 Q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta});\\
\mathbf{K}_n &= n^{-1}\sum_{i=1}^n q_{in}^2 V(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})M^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}_i.
\end{aligned}$$

In what follows, we will 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 well as sufficient smoothness of $\mu(\cdot)$. Under the conditions from above, at least asymptotically there will be no leverage points, so that the usual and unbiased sandwich estimators will have similar asymptotic behavior. We write $\bar{\boldsymbol{\Omega}}(\boldsymbol{\beta}) = E\{\boldsymbol{\Omega}_n(\boldsymbol{\beta})\}$, $q = \mathbf{x}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta})\mathbf{z}$, $a = \mathbf{z}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta})\mathbf{z}$, $\bar{\mathbf{C}} = E\{q^2 Q^{(1)}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}\}$, etc., i.e. the bar notation refers to asymptotic moments.

Theorem 3 As $n \to \infty$, under the conditions above we have

$$\begin{aligned}
n^{1/2}(\mathbf{V}_{model} - \mathbf{V}_{asymp}) &\Rightarrow \text{Normal}[0, \Sigma_{model} := E\{a\xi(\epsilon^2 - \sigma^2) - \sigma^2(a\bar{\mathbf{R}} + \bar{\mathbf{C}})^{\mathrm{T}}\boldsymbol{\ell}\epsilon\}^2]; \\
n^{1/2}(\mathbf{V}_{sand} - \mathbf{V}_{asymp}) &\Rightarrow \text{Normal}[0, \Sigma_{sand} := E\{q^2 v + (\bar{\mathbf{K}} - 2\sigma^2\bar{\mathbf{C}})^{\mathrm{T}}\boldsymbol{\ell}\epsilon\}^2].
\end{aligned}$$

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

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

which shows the first part of Theorem 4.

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

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

as claimed.

Theorem 3 can now be used to prove the statements listed in Examples 3 and 4. For the logistic case we have $V(\eta) = \mu^{(1)}(\eta) = Q(\eta) = \mu(\eta)\{1-\mu(\eta)\}$, $\sigma^2 = 1$, $\xi = 0$, $\mathbf{R}_n = 0$, $Q^{(1)}(\eta) = \mu^{(1)}(\eta)\{1-2\mu(\eta)\}$. All the terms in Theorem 3 can then be computed by numerical integration which gives the numbers presented in Example 4.

For the Poisson case it is easily verified that $\bar{\boldsymbol{\Omega}}(\boldsymbol{\beta}) = \exp(\beta_0)\mathbf{I}_2$, where $\mathbf{I}_2$ is the 2×2 identity matrix. Also, $q = U\exp(-\beta_0)$, $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = \beta_0$, $Q^{(1)}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \exp(\beta_0)$, $\bar{\mathbf{C}} = \exp(-\beta_0)(1,0)^{\mathrm{T}}$, $\boldsymbol{\ell} = \exp(-\beta_0/2)(1,U)^{\mathrm{T}}$, $\epsilon = \{Y - \exp(\beta_0)\}/\exp(\beta_0/2)$ and hence $\Sigma_{model} = \exp(-3\beta_0)$.

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, $M(\eta) = 1$, $M^{(1)}(\eta) = 0$, $\bar{\mathbf{K}} = 0$. A detailed calculation then shows that $\Sigma_{sand} = 2\kappa\exp(-2\beta_0) + \kappa\exp(-3\beta_0)$ which shows the relative efficiency given in Example 3.

A.2 Proof of Theorem 2

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

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

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

$$\widehat{\sigma} - \sigma = g(\widehat{\sigma}^2) - g(\sigma^2) = \frac{\widehat{\sigma}^2-\sigma^2}{2\sigma} - \frac{(\widehat{\sigma}^2-\sigma^2)^2}{8\sigma^3} + O_p(n^{-3/2}).$$

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

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

Since $F(v_p) = p$ this yields

$$E\left\{F(\widehat{v}_p)\right\} = p + \mathrm{var}(\widehat{\sigma}^2)\left\{\frac{z_p^2 F^{(2)}(z_p)}{8\sigma^4} - \frac{z_p F^{(1)}(z_p)}{8\sigma^4}\right\} + O(n^{-3/2}).$$

Inserting now the derivatives for $F(v) = \Phi(v/\sigma)$ gives formula (8) in Theorem 2.

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.

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. Springer Verlag*, Berlin,New York.

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.

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.

MacKinnon, J. G., and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 305–325.

McCullagh, P. (1987). Tensor methods in statistics. *Chapman & Hall*, London.

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.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. *Chapman and Hal*, New York.

Rothenberg, T.J. (1988). Approximative power functions for some robust tests of regression coefficients. *Econometrica*, 56, 997–1019.

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.

Wedderburn, R. W. M (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biomtrika*, 61, 439–447.

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	$t_{p,n-1}$	$z_{\tilde{p}}$	$P(\hat{\theta} \leq \theta + z_{\tilde{p}}\hat{\sigma}/\sqrt{n})$
		$n=5$	
.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: Coverage probability based on $\mathbf{V}_{sand,u}$ with $z_{\tilde{p}}$ and t distribution quantiles $t_{p,n-1}$ and jackknife estimate $\mathbf{V}_{jack}$ (Slanted numbers show simulations for correlated responses)

			coverage based on		
			$\mathbf{V}_{sand,u}$	$\mathbf{V}_{sand,u}$	$\mathbf{V}_{jack}$
design	$t_{p,n-2}$	$z_{\tilde{p}}$	$z_{\tilde{p}}$	$t_{p,n-2}$	$t_{p,n-2}$
			$n=10$ $(m=4)$		
(a)		2.10	88.8 (*88.5*)	84.9 (*84.9*)	86.4 (*87.2*)
(b)	1.86	2.03	88.5 (*90.2*)	86.3 (*87.5*)	87.6 (*89.0*)
(c)		2.18	88.9 (*89.0*)	84.2 (*84.6*)	86.7 (*86.8*)
			$n=20$ $(m=4)$		
(a)		1.86	89.5 (*89.7*)	87.0 (*87.8*)	88.3 (*88.8*)
(b)	1.71	1.81	90.3 (*90.0*)	88.5 (*88.4*)	89.8 (*89.9*)
(c)		1.94	90.0 (*90.5*)	86.5 (*87.1*)	88.2 (*88.9*)

			coverage based on		
			$\mathbf{V}_{sand,u}$	$\mathbf{V}_{sand,u}$	$\mathbf{V}_{jack}$
design	$t_{p,n-2}$	$z_{\tilde{p}}$	$z_{\tilde{p}}$	$t_{p,n-2}$	$t_{p,n-2}$
Logistic regression $n = 15$ $(m = 4)$					
(a)		1.89	91.0	83.8	90.2
(b)	1.77	1.84	89.7	85.6	89.2
(c)		1.96	91.1	83.3	89.9
Poisson regression $n = 15$ $(m = 4)$					
(a)		1.90	90.1	86.9	89.3
(b)	1.77	1.87	90.4	88.1	89.8
(c)		1.90	87.5	84.2	86.3

Table 3: Coverage probability of confidence based on $\mathbf{V}_{sand,u}$ with $z_{\tilde{p}}$ calculated with true and fitted parameters and t distribution quantiles $t_{p,n-1}$