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

Göran Kauermann * Raymond J. Carroll †

December 16, 1999

**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. 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; Quasilikelihood; Robust covariance estimator; Sandwich estimator.

**Short Title:** The Sandwich Estimator

*Institute of Statistics; Ludwig-Maximilians-University Munich; Akademiestr. 1; 80796 Munich, Germany

†Department of Statistics, Texas A & M University, College Station, TX 77843-3143, USA. 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).

# 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. Diggle, Liang & Zeger (1994), Liang & Zeger (1986), and Liang, Zeger & Qaqish (1992. 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. The robust 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 the first point 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 inefficiency equal to the inverse of the sample kurtosis of the design values. This inefficiency still holds in generalized linear models. 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 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 apply an Edgeworth expansion and concentrate on the coverage probability of confidence intervals. 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 $\hat{\sigma}^2 = \sum_{i=1}^n \widehat{\epsilon}_i^2/(n-p)$ with $\widehat{\epsilon}_i = Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ as fitted residuals. A major assumption implicitly used in the calculation of $V_{model}$ is that the errors $\epsilon_i$ 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

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

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, $E(\widehat{\epsilon}_i^2) = \sigma^2(1 - h_{ii})$. It then follows that

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

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. To demonstrate this let the first point

be a *leverage* point such that $h_{11} = \max_{1 \le i \le n} h_{ii}$ and set $\mathbf{z} = \mathbf{x}_1 / \left\{ \mathbf{x}_1 (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{x}_1^{\mathrm{T}} \right\}^{1/2}$. Then the bias of the sandwich estimator satisfies

$$\max_{\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}=1} |\mathrm{bias}(V_{sand})| \ge \sigma^2 \max_{1 \le i \le n} h_{ii}^2.$$

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

Bias problems can be avoided by replacing $\widehat{\epsilon}_i$ in (1) by $\widetilde{\epsilon}_i = \widehat{\epsilon}_i / (1 - h_{ii})^{1/2}$. The resulting estimator is refered to as unbiased sandwich variance estimator in the following denoted by $V_{sand,u}$ (see also Wu, 1986, equation 2.6). It is easily seen that $E(V_{sand,u}) = \mathrm{var}(\mathbf{z}\widehat{\boldsymbol{\beta}})$. 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 $i \ne j$, where $\widetilde{h}_{ij} = h_{ij} / \{(1 - h_{ii})(1 - h_{jj})\}^{1/2}$, it follows that

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

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

**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:

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

If in addition $\max(h_{ii}) = o(n^{-1/2})$, then the middle term in (4) gives the asymptotic relative efficiency.

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. The loss of efficiency is thereby inversely proportional to the kurtosis of the design points as the following two examples show.

*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 $\widehat{\beta}_1 = \mathbf{z}\widehat{\boldsymbol{\beta}}$ 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 $\kappa_n^{-1}$, where $\kappa_n = n^{-1}\sum u_i^4/(n^{-1}\sum_{i=1}^n u_i^2)^2 \geq 1$. Note that $\kappa_n$ 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 non linear 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. $\mathrm{var}(Y|\mathbf{X}) = \sigma^2 V(\mu_i) =: \sigma^2 \mathbf{V}_i$ where $\mu_i$ abbreviates $\mu_i = h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$, $V(\cdot)$ is a known variance function and $\sigma^2$ 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 (see Wedderburn, 1974) or generalized linear models (see McCullagh & Nelder, 1989). The parameter $\boldsymbol{\beta}$ can be estimated using the generalized estimating equation (see 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 $\mathbf{V}_{model} := \hat{\sigma}^2 \mathbf{z}^{\mathrm{T}} \widehat{\boldsymbol{\Omega}}^{-1} \mathbf{z}$ where $\widehat{\boldsymbol{\Omega}}$ is a simple plug in estimate of $\boldsymbol{\Omega}$ and $\sigma^2$ 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 $\mathrm{var}(\mathbf{z}^{\mathrm{T}} \boldsymbol{\beta})$ can be estimated by the sandwich formula

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

where $\widehat{\boldsymbol{\epsilon}}_i = Y_i - \widehat{\mu}_i = Y_i - h(\mathbf{X}_i \widehat{\boldsymbol{\beta}})$ are the fitted residuals and $\widehat{\mathbf{V}}_i$ as plug in estimate of $\mathbf{V}_i$. 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.

We rewrite the sandwich estimate (7) using a matrix notation. 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

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

with $\mathrm{diag}_m(\mathbf{V}_i^{-1})$ denoting the block diagonal matrix having $\mathbf{V}_i^{-1}$, $i = 1,\ldots n$ on its diagonal. 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 now $\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \widehat{\boldsymbol{\mu}} = \mathbf{P}(\mathbf{Y} - \boldsymbol{\mu})\{1 + O_p(n^{-1/2})\}$ be the fitted residual where $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1^{\mathrm{T}},\ldots,\widehat{\mu}_n^{\mathrm{T}})^{\mathrm{T}}$ with $\widehat{\mu}_i = h(\mathbf{X}_i\widehat{\boldsymbol{\beta}})$. With $\mathbf{H}_{ii}$ denoting the $i$-th $m\times m$ diagonal block of $\mathbf{H}$ we define the leverage–adjusted residual $\widetilde{\boldsymbol{\epsilon}}_i = (I - \mathbf{H}_{ii})^{-\frac{1}{2}}\widehat{\boldsymbol{\epsilon}}_i$. Replacing now $\widehat{\boldsymbol{\epsilon}}$ in (7) by $\widetilde{\boldsymbol{\epsilon}}$ gives the bias reduced sandwich estimate $\mathbf{V}_{sand,u}$ which fulfills $E(\mathbf{V}_{sand,u}) = \mathrm{var}(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\{1 + O(n^{-1})\}$.

## 3.2 Properties of Sandwich Estimator

For the calculation of the variance of $\mathbf{V}_{sand,u}$ it is helpful to write (7) as quadratic form. Let therefore $\mathbf{W}$ be the block diagonal matrix $\mathbf{W} = \mathrm{diag}_m(\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i)$ where $\mathbf{a}_i = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\dfrac{\partial\mu_i^{\mathrm{T}}}{\partial\boldsymbol{\beta}}V_i^{-1}$ and let $\mathbf{D} = \mathrm{diag}_m(I - H_{ii})^{-1/2}$. As above We use the hat notation to denote plug in estimates, for instance $\widehat{\mathbf{W}}$ or $\widehat{\mathbf{D}}$ are plug in estimates $\mathbf{W}$ or $\mathbf{D}$, respectively. This allows to write

$$\mathbf{V}_{sand,u} \;=\; \widetilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\widehat{\mathbf{W}}\widetilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{P}\,\widehat{\mathbf{D}}\,\widehat{\mathbf{W}}\,\widehat{\mathbf{D}}\,\mathbf{P})\boldsymbol{\epsilon} = \dot{\boldsymbol{\epsilon}}^{\mathrm{T}}\widehat{\mathbf{M}}\dot{\boldsymbol{\epsilon}}\,\{1 + O(n^{-1})\}, \tag{8}$$

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 to calculate 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

$$\text{var}(\mathbf{V}_{sand,u}) \;=\; 2\text{tr}(\mathbf{MM}) + \sum_k \{E(\dot{\epsilon}_k^4) - 3\} m_{kk}^2 \tag{9}$$

with $\text{tr}(\cdot)$ denoting the trace of a matrix. If $\dot{\epsilon}_k$ are standard normal (9) simplifies due to $E(\dot{\epsilon}_k^4) = 3$ to $\text{var}(\mathbf{V}_{sand,u}) = 2\text{tr}(\mathbf{MM})$. 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 following example shows that this variance has a lower bound which equals the variance of a standard variance estimate for $n$ independent, univariate normally distributed variables. In particular the lower bound does neither depend on $m$, the number of replicates per unit, nor on $p$, the dimension of $\boldsymbol{\beta}$.

*Example 3 (lower bound of the variance):* Consider the multivariate normal model $Y_i \sim N(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, $i = 1,\ldots,n$, with $\mathbf{X}_i^{\mathrm{T}}$ as $m \times p$ design matrix and $\mathbf{I}$ as $m \times m$ identity matrix. For simplicity 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

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

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 $i$. In particular this is the case if the individual design $\mathbf{X}_i$ does not differ among the individuals. In this case one gets the lower bound $\text{var}(\mathbf{V}_{sand,u}) = 2\sigma^4/\{n^2(n-1)\}\{1 + O(n^{-1})\}$. One should note that the property $\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} = \mathbf{I}$ either requires $m > 1$, i.e. one observes more than one observation for each unit, or if $m = 1$ it implies that $\mathbf{X}_i \equiv 1$, because otherwise the model would not be identifiable. For $m = 1$ the simple

univariate normal model $y_i \sim N(\mu, \sigma^2)$ results with $\mu$ as unknown constant mean and the resulting variance estimate $\widehat{\sigma}^2 = \sum_i (y_i - \bar{y})^2/(n-1)$. This in turn implies a lower bound for the quantiles used for the calculation of confidence intervals, as will be discussed in the next section.

For the calculation of $\mathrm{var}(\mathbf{V}_{sand,u})$ in (9) we neglected the variability occuring due to plug-in estimation of $\hat{\beta}$. If however the variance function $V(\mu)$ is not constant both estimates $\mathbf{V}_{sand,u}$ and $\mathbf{V}_{model}$ have a variance greater than zero. The next two examples show how the additional variance occuring from the plug in estimates effects the relative efficiency $\mathrm{var}(\mathbf{V}_{sand})/\mathrm{var}(\mathbf{V}_{model})$. We consider univariate Poisson and Logistic regression models, a general discussion is given in the appendix.

*Example 4 (Poisson loglinear regression):* We consider the 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 $\beta_1$ 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 already occured in Example 2 above. The additional variability in the Poisson case is a 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 5 (Logistic Regression):* Let now $y_i$ be binary with $E(y_i|\mathbf{x}) = \mathrm{logit}^{-1}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ where $\mathbf{x}_i$ as above. Again, the slope $\beta_1$ is the parameter of interest. We varied $\beta_1$ while choosing $\beta_0$ so that marginally $E(y|\mathbf{x}) = 0.10$. With $\beta_1 = 0.0, 0.5, 1.0, 1.5$, the asymptotic relative efficiency $\mathrm{var}(\mathbf{V}_{sand})/\mathrm{var}(\mathbf{V}_{model})$ varied for $u_i$ standard normally distributed over $3.00, 2.59, 1.92, 1.62$,

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 example show that the loss of efficiency of the sandwich variance estimate in non normal models differs and can be worse compared to normal models. In the appendix we discuss this point in more generality.

# 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. Let $\theta = \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}$ be the unknown parameter of interest and let $\widehat{\theta} = \mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ be an unbiased, $n^{1/2}$–consistent estimate of $\theta$. We consider confidence intervals based on the (asymptotic) normality of $\hat{\theta}$, i.e. we investigate the symmetrical $1-\alpha$ confidence intervals $CI(\sigma^2, \alpha) := [\hat{\theta} \pm z_p \sigma/\sqrt{n}]$ where $\sigma^2/n = \mathrm{var}(\hat{\theta})$ and $z_p$ as $p = 1 - \alpha/2$ quantile of the standard normal distribution. If $\sigma^2$ is estimated by an unbiased variance estimate $\hat{\sigma}^2$ it is well known that the confidence interval $CI(\hat{\sigma}^2, \alpha)$ shows undercoverage and typically $t$-distribution quantiles are used instead of normal quantiles. The following theorem shows how the variance of the variance estimate $\widehat{\sigma}^2$ directly effects the undercoverage.

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

equals

$$P\{\theta \in CI(\hat{\sigma}^2, \alpha)\} \quad = \quad 1 - \alpha - \phi(z_p)\mathrm{var}(\hat{\sigma}^2)\left(\frac{z_p^3 + z_p}{8\sigma^4}\right) + O(n^{-3/2}) \tag{10}$$

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

The proof of the theorem is obtained by Edgeworth expansion and given in the appendix. One should note that the postulated assumption of independence of $\widehat{\sigma}^2$ and $\widehat{\theta}$ holds in a normal regression model if $\widehat{\sigma}^2$ is calculated from fitted residuals. Hence it holds for sandwich variance estimates. In the non-normal case we have to rely on asymptotic independence. It is seen from (10) that for $\alpha$ small, i.e. $p$ large such that $z_p > 1$, 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 estimator can be expected to have lower coverage probability than the model based variance. Moreover, $t-$distribution quantiles instead of normal quantiles do not correct the undercoverage, as seen below.

## 4.2 A Simple Coverage Adjustment

Formula (10) can be employed to construct a simple coverage correction for confidence intervals. Instead of using the quantile $z_p$ directly we suggest to choose $\widetilde{p} > p$ and make use of the $z_{\widetilde{p}}$ quantile. We thereby select $\widetilde{p}$ such that $P(\theta \in [\hat{\theta} \pm z_{\widetilde{p}}\widehat{\sigma}]) = p$ holds, i.e. with (10) we solve

$$p \quad = \quad \widetilde{p} - \phi(z_{\widetilde{p}})\mathrm{var}(\widehat{\sigma}^2)\frac{z_{\widetilde{p}}^3 + z_{\widetilde{p}}}{8\sigma^4} \tag{11}$$

for $\tilde{p}$. Though equation (11) does not allow for an analytical solution for $\widetilde{p}$, a numerical solution is easily calculated by iteration. It should be noted that $\tilde{p}$ depends on $p$ which is however suppressed in the notation.

*Example 6 (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 $\sigma^2$ 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 (11) using $\mathrm{var}(\widehat{\sigma}^2) = 2\sigma^4/(n-1)$. One should note that the unknown variance in (11) cancels out so that estimation of $\sigma^2$ is not required for the calculation of $z_{\widetilde{p}}$. In Table 1 we compare the exact quantiles with the corrected version $z_{\widetilde{p}}$. The corrected quantiles $z_{\widetilde{p}}$ are distinctly close to the exact $t$-distribution quantiles, even for small sample sizes. This also shows in the true coverage probability $P(\hat{\theta} \leq \theta + z_{\widetilde{p}}\hat{\sigma}/\sqrt{n})$ of the confidence intervals and demonstrates that the adjustment applied in a standard setting behaves convincingly well.

The adjustment (11) can now easily be adopted to sandwich variance estimates. Assume that $\sigma^2/n = \mathrm{var}(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$ is estimated by the sandwich estimate $\widehat{\sigma}^2/n = V_{sand,u}$. The variance $\mathrm{var}(\widehat{\sigma}^2) = n^2\mathrm{var}(V_{sand,u})$ is calculated form (9). Inserting this into (11) directly gives the adjusted quantile $z_{\widetilde{p}}$ which is used to get the $(1-\alpha)$ confidence interval $[\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \pm z_{\widetilde{p}} V_{sand,u}^{1/2}]$ with $\alpha = 2(1-p)$. We present some simulations below to demonstrate the benefits of the adjustment. We also compare our approach to jackknife sandwich estimates as suggested in MacKinnon & White (1985, formula 13). Assuming working independence, for simplicity, and considering the multivariate normal model $Y_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, I)$ 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 $\widehat{\gamma} = \mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\epsilon}}$.

*Simulation 1 (normal response):* Let $E(Y_i|\mathbf{X}_i) = h\{\mathbf{X}_i\boldsymbol{\beta}\}$ with $\mathbf{X_i} = (\mathbf{1_m}, \mathbf{U_i})$ where $1_m$ 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 consider the following designs for the covariates, let $\mathbf{U}_i = 1_m u_i$ 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. The response $y$ is drawn from a normal distribution with identity link function. Working independence is used for fitting $\boldsymbol{\beta}$ but $Y_i$ is simulated from two setting, (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 setting are shown as slanted numbers. For comparison we also report coverage probabilities if $t$-distribution quantiles with $n-2$ degrees of freedom are used. Moreover we give the coverage rate for the jackknife estimate $\mathbf{V}_{jack}$ combined with $t$-distribution quantiles. For all three designs our proposed adjustment shows a very 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 where the jackknife approach behaves more accurate, as already mentioned in MacKinnon & White (1985).

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

*Example 3 (continued):* We pick up Example 3 from above again. We showed there that $\mathrm{var}(\mathbf{V}_{sand,u}) \geq \mathrm{var}(\hat{\sigma}^2/n)$ with $\hat{\sigma}^2 = \sum_i (y_i - \bar{y})^2/(n-1)$ and $y_i \sim N(\mu, \sigma)$. For the latter case confidence intervals for $\mu$ are constructed from $\hat{\sigma}/n^{1/2}$ and $t$-distribution quantiles. As seen from Example 5 the corrected quantiles $z_{\widetilde{p}}$ is approximately equal to $t$-distribution quantiles when applied in a standard setting. This shows, with Example 3, that if $\sigma^2$ is estimated by a

sandwich estimate one obtains $z_{\tilde{p}} \geq t_{p,n-1}$, at least approximately with $t_{p,n-1}$ as $p$ quantile of the $t$-distribution with $n-1$ degrees of freedom. If the design is individually balanced we saw above that the lower bound of the variance is reached. This implies whenever the individual design $\mathbf{X}_i$ does not differ among the individuals, i.e. does not depend on $i$, one should use $t$ distribution quantiles with $n-1$ degrees of freedom when constructing confidence intervals based on sandwich variance estimates. Consequently, undercoverage is not an issue in this case.

In the next simulation we show how the adjustment behaves for binomial and Poisson data.

*Simulation 2 (Logistic and Poisson regression):* It should be noted that the adjustment for normal data depends only in the design but not on $\mu$ or $\sigma^2$. This property does not hold for non-normal data since $\mathbf{V}_i$ typically depends on the mean $\mu_i$. The calculation of $\text{var}(\mathbf{V}_{sand,u})$ therefore requires some plug-in estimates. We neglect the effect of the plug in estimates on the variance and calculate the variance of the sandwich variance estimate from (9). We make use of the adjusted quantiles $z_{\tilde{p}}$ from (11), where again plug in estimates are used to calculate $z_{\tilde{p}}$. We simulate binomial data with predictor $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ and $\boldsymbol{\beta} = (0.5, 0.5)^T$ while we chose $\boldsymbol{\beta} = (1, 1)^{\mathrm{T}}$ for the Poisson simulation. The covariates $\mathbf{x}_i$ are distributed as in Simulation 1 and we are interested in the slope parameter $\beta_1$. For comparison we again compare our proposed correction with a 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 comparable to our approach.

# 5 Discussion

We showed above that sandwich variance estimates carry the problem of being less efficient than model based variance estimates. The loss of efficiency thereby depends on the design and for standard cases it is directly 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 effects the coverage probability of confidence intervals and undercoverage is implied if the design differs among the independent individuals. A simple adjustment which depends on the design is suggested which allows to correct the deficit of undercoverage.

# A Technical Details

## A.1 Proof of Examples 4 and 5

Below we derive the relative efficiency in quasi likelihood models in the following. 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, $\sigma^2$ is estimated, which we indicate by setting $\xi = 1$, while when $\sigma^2$ 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 $\mathrm{var}(n^{1/2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}) = \sigma^2 \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n(\boldsymbol{\beta})\mathbf{z}$ where $\boldsymbol{\Omega}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} Q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ with $Q(\eta) = \{\mu^{(1)}(\eta)\}^2/V(\eta)$. The model based variance estimator is $V_{model} = \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}})\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{z}$, 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 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 $\boldsymbol{\Omega}(\boldsymbol{\beta}) = E\{\boldsymbol{\Omega}_n(\boldsymbol{\beta})\}$, $q = \mathbf{x}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\boldsymbol{\beta})\mathbf{z}$, $a = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\boldsymbol{\beta})\mathbf{z}$, $\mathbf{C} = E\{q^2 Q^{(1)}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}\}$, etc., i.e. if the subscript $n$ is neglected we refer to asymptotic moments.

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

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

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,

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

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

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

as claimed.

Theorem 3 can now be used to prove the statements listed in Examples 4 and 5. For the logistic case we have $V(\eta) = \mu^{(1)}(\eta) = Q(\eta) = \mu(\eta)\{1-\mu(\eta)\}$, $\sigma^2 = 1$, $\xi = 0$, $\mathbf{R}_n = 0$, $Q^{(1)}(\eta) = \mu^{(1)}(\eta)\{1-2\mu(\eta)\}$. All the terms in Theorem 3 can then be computed by numerical integration which gives the numbers presented in Example 5.

For the Poisson case it is easily verified that $\boldsymbol{\Omega}(\boldsymbol{\beta}) = \exp(\beta_0)I_2$, where $I_2$ is the identity matrix. Also, $q = U\exp(-\beta_0)$, $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = \beta_0$, $Q^{(1)}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \exp(\beta_0)$, $\mathbf{C} = \exp(-\beta_0)(1,0)^{\mathrm{T}}$, $\boldsymbol{\ell} = \exp(-\beta_0/2)(1,U)^{\mathrm{T}}$, $\epsilon = \{Y - \exp(\beta_0)\}/\exp(\beta_0/2)$ 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, $M(\eta) = 1$, $M^{(1)}(\eta) = 0$, $\mathbf{K} = 0$. 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 4.

## A.2 Proof of Theorem 2

We give the proof of Theorem 2 in a rather general fashion. Let $F(\cdot)$ denote the distribution of $\hat{\theta}-\theta$ and $\widehat{F}(\cdot)$ be a corresponding estimate of $F(\cdot)$. Let $\kappa_l$ and $\widehat{\kappa}_l$ be the $l$-th cumulant of $F(\cdot)$ and $\widehat{F}(\cdot)$ respectively, $l=1,2,\ldots$. We assume that $\kappa_1=\widehat{\kappa}_1=0$ and $\kappa_l=O(n^{-(l-2)/2})$ which mirrors standard $n^{1/2}$–asymptotics. We also postulate that $\widehat{\kappa}_l$ is a consistent estimate of $\kappa_l$, i.e. $\widehat{\kappa}_l-\kappa_l=O_p(n^{-\frac{1}{2}})O(\kappa_l)$ and we assume that the cumulants $\widehat{\kappa}_l$ are independent of $\hat{\theta}$ for $l=2,3,\ldots$. Let $\widehat{v}_p$ denote the empirical $p$ quantile, i.e. $\widehat{F}(\widehat{v}_p)=p$. As seen below, since $\widehat{\kappa}_l$ and $\widehat{\theta}$ are assumed to be independent one gets $\widehat{v}_p$ and $\widehat{\theta}$ independent. Our intention is to calculate the true coverage probability $P\{(\hat{\theta}-\theta)\leq\hat{v}_p\}$. Let $H_{\hat{v}_p}(\cdot)$ denote the distribution function of $\hat{v}_p$. One finds, by making use of the above independence assumption

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

Hence, we have to calculate the expectation of $F(\hat{v}_p)$ to obtain the coverage probability. Simple expansion about the true quantile $v_p$ yields

$$F(\widehat{v}_p) = F(v_p)+F^{(1)}(v_p)(\widehat{v}_p-v_p)+\frac{1}{2}F^{(2)}(v_p)(\widehat{v}_p-v_p)^2+\ldots \tag{13}$$

where $F^{(l)}(\cdot)$ denotes the $l$-th derivative of $F(\cdot)$. The calculation of the expectation requires an expansion for $\widehat{v}_p$ which is derived in the following (see also Hall, 1992, for a similar expansion). We first expand $\widehat{F}(\cdot)$ by Edgeworth expansion about $F(\cdot)$. Following McCullagh (1987, page 144–47) this yields $\widehat{F}(v)=F(v)+\widehat{\Delta}_F(v)$ where the correction term equals

$$\widehat{\Delta}_F(v)=\sum_{l\geq 2}(-1)^l\frac{F^{(l)}(v)}{l!}\hat{\delta}_l \tag{14}$$

with $\hat{\delta}_2=\hat{\kappa}_2-\kappa_2$; $\hat{\delta}_3=\hat{\kappa}_3-\kappa_3$ and $\hat{\delta}_4=\hat{\kappa}_4-\kappa_4+3\hat{\delta}_2^2$ and so on. Let $G(p)=F^{-1}(p)$ denote the inverse distribution function. The empirical quantile $p=\widehat{F}(\widehat{v}_p)=F(\widehat{v}_p)+\widehat{\Delta}_F(\widehat{v}_p)$ is then expanded by

$$\widehat{v}_p = G\{p-\widehat{\Delta}_F(\widehat{v}_p)\}$$

$$
\begin{aligned}
&= G(p) - G^{(1)}(p)\{\widehat{\Delta}_F(v_p) + \widehat{\Delta}_F^{(1)}(v_p)(\widehat{v}_p - v_p) + \ldots\} \\
&\quad + \frac{1}{2}G^{(2)}(p)\{\widehat{\Delta}_F(v_p) + \widehat{\Delta}_F^{(1)}(v_p)(\widehat{v}_p - v_p) + \ldots\}^2 + \ldots \qquad (15)
\end{aligned}
$$

where $\widehat{\Delta}_F^{(l)}(v) = \partial^l \widehat{\Delta}_F(v)/(\partial v)^l$. With $v_p = G(p)$ as true quantile we can solve (15) for $(\widehat{v}_p - v_p)$ by series inversion. This permits the first order approximation

$$
\begin{aligned}
\widehat{v}_p - v_p &= \{1 + G^{(1)}(p)\widehat{\Delta}_F^{(1)}(v_p)\}^{-1}\{-G^{(1)}(p)\widehat{\Delta}_F(v_p) + \frac{1}{2}G^{(2)}(p)\widehat{\Delta}_F(v_p)^2\} + O(\widehat{\Delta}_F^3) \\
&= -G^{(1)}(p)\widehat{\Delta}_F(v_p) + \frac{1}{2}G^{(2)}(p)\widehat{\Delta}_F(v_p)^2 \qquad (16) \\
&\quad + {G^{(1)}}^2(p)\widehat{\Delta}^{(1)}(v_p)\widehat{\Delta}_F(v_p) + O(\widehat{\Delta}_F^3).
\end{aligned}
$$

In $O(\widehat{\Delta}_F^3)$ we collect components of the third power of $\widehat{\Delta}_F(v_p)$ or its derivatives, e.g. $\widehat{\Delta}_F(v_p)^3$ is a representative. Reflecting that $\widehat{\Delta}_F(v_p)$ is dominated by $\widehat{\delta}_2 = \widehat{\kappa}_2 - \kappa_2$ provides $O(\widehat{\Delta}_F^3) = O_p(n^{-\frac{3}{2}})O(\kappa_2)$ so that $O(\widehat{\Delta}_F^3)$ collects components of negligible order. Inserting (16) in (13) yields

$$
\begin{aligned}
F(\widehat{v}_p) &= F(v_p) \\
&\quad + F^{(1)}(v_p)\{-G^{(1)}(p)\widehat{\Delta}_F(v_p) + \frac{1}{2}G^{(2)}(p)\widehat{\Delta}_F(v_p)^2 + {G^{(1)}}^2(p)\widehat{\Delta}_F^{(1)}(v_p)\widehat{\Delta}_F(v_p)\} \\
&\quad + \frac{1}{2}F^{(2)}(v_p){G^{(1)}}^2(p)\widehat{\Delta}_F(v_p)^2 + O(\widehat{\Delta}_F^3)
\end{aligned}
$$

which simplifies with $G^{(1)}(p) = 1/F^{(1)}(v_p)$ and $G^{(2)}(p) = -F^{(2)}(v_p)/F^{(1)}(v_p)^3$ to

$$
F(\widehat{v}_p) = p - \widehat{\Delta}_F(v_p) + \frac{1}{F^{(1)}(v_p)}\widehat{\Delta}^{(1)}(v_p)\widehat{\Delta}_F(v_p) + O(\widehat{\Delta}_F^3).
$$

Assuming unbiased estimates for the cumulants we find with (14) $E\{\widehat{\Delta}_p(v_p)\} = 1/24F^{(4)}(v_p)\mathrm{var}(\widehat{\delta}_2)\{1 + O(n^{-1/2})\}$ and $E(\widehat{\Delta}^{(1)}(v_p)\widehat{\Delta}_F(v_p)) = 1/4\, F^{(2)}(v_p)F^{(3)}(v_p)\mathrm{var}(\widehat{\delta}_2)\{1 + O(n^{-1/2})\}$ which finally yields

$$
E\{F(\widehat{v}_p)\} = p + \mathrm{var}(\hat{\kappa}_2)\left\{-\frac{1}{8}F^{(4)}(v_p) + \frac{1}{4}\frac{F^{(2)}(v_p)F^{(3)}(v_p)}{F^{(1)}(v_p)}\right\}\{1 + O(n^{-1/2})\}. \qquad (17)
$$

Taking now $F(v) = \Phi(v/\sigma)$ with $\Phi(\cdot)$ as standard normal distribution function and $\hat{F}(v) = \Phi(v/\hat{\sigma})$, where $\hat{\sigma}^2 = \hat{\kappa}_2$ is an estimate of the second order cumulant, gives with $v_p = z_p\sigma$, $\hat{v}_p = z_p\hat{\sigma}$ and (17) formula (10) 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.

| 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 1: Comparison of coverage probability based on $z_{\tilde{p}}$ and $t$-distribution quantiles $t_{p,n-1}$ for $n-1$ degrees of freedom

| | | | 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*) |

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)

| design | $t_{p,n-2}$ | $z_{\widetilde{p}}$ | coverage based on $\mathbf{V}_{sand,u}$ $z_{\widetilde{p}}$ | $\mathbf{V}_{sand,u}$ $t_{p,n-2}$ | $\mathbf{V}_{jack}$ $t_{p,n-2}$ |
|---|---|---|---|---|---|
| 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 $z_{\widetilde{p}}$ calculated with true and fitted parameters and $t$ distribution quantiles $t_{p,n-1}$

This is the original Section 3.2. I incorporated the results given here in the previous sections. I put the Theorem in the appendix. Please see this as a proposal !!

## A.3 Asymptotic Comparisons in the Univariate Case

*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 $\sigma^2 V(\mathbf{X}^T\boldsymbol{\beta})$, where the functions $\mu(\cdot)$ and $V(\cdot)$ are known. In some problems, $\sigma^2$ is estimated, which we indicate by setting $\xi = 1$, while when $\sigma^2$ is known we set $\xi = 0$. The quasilikelihood estimate of $\boldsymbol{\beta}$ is the solution $\widehat{\boldsymbol{\beta}}$ to*

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

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

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

$$V_{ql} = \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}})\mathbf{z}^T\boldsymbol{\Omega}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{z},$$

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

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

*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 $M(x) = \{\mu^{(1)}(x)/V(x)\}^2$, the usual sandwich estimator is*

$$V_{sand} = \mathbf{z}^T\boldsymbol{\Omega}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{B}_n(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{z}. \tag{18}$$

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

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

*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* $\boldsymbol{\Omega}(\boldsymbol{\beta}) = E\{\boldsymbol{\Omega}_n(\boldsymbol{\beta})\}$, $q = \mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\boldsymbol{\beta})\mathbf{z}$, $a = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\boldsymbol{\beta})\mathbf{z}$, $\mathbf{C} = E\{q^2 Q^{(1)}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{X}\}$, *etc.*

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

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

*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* $\mathbf{X} = (1, U)^{\mathrm{T}}$ *and* $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$.

- *In linear homoscedastic regression,* $\mu(x) = x$, $V(x) = 1$. *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,* $V(x) = \mu^{(1)}(x) = Q(x) = \mu(x)\{1 - \mu(x)\}$, $\sigma^2 = 1$, $\xi = 0$, $\mathbf{R}_n = 0$, $Q^{(1)}(x) = \mu^{(1)}(x)\{1 - 2\mu(x)\}$. *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* $\beta_1$ *while choosing* $\beta_0$ *so that marginally* $pr(Y = 1) = 0.10$. *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* $6.00, 4.36, 3.31, 2.57$.

  *Note that in both these situations, 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.*

- *In Poisson loglinear regression,* $\mu(x) = V(x) = \exp(x)$, $\sigma^2 = 1$, $\xi = 0$ *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* $\mathbf{z}^{\mathrm{T}}\mathbf{z} = 1$ *and* $\widetilde{\boldsymbol{\mathcal{Z}}}^{\mathrm{T}}\widetilde{\boldsymbol{\mathcal{Z}}} = nI$. *Then* $Q(x) = Q(\beta_0) = Q$, $A = QI$, $M = Q/V$, $g = (\partial/\partial\beta_0)\log\{V(\beta_0)\}$, $\mathbf{R} = \xi g(1, 0)^{\mathrm{T}}$, $q = U/Q$, $a = 1/Q$, $\mathbf{C} = (Q^{(1)}/Q^2)(1, 0)^{\mathrm{T}}$, $\boldsymbol{\ell} = Q^{-1/2}(1, U)$, $v = (\epsilon^2 - \sigma^2)Q$, $\mathbf{K} = (M^{(1)}V/Q^2)(1, 0)^{\mathrm{T}}$ *and thus*

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

  *The kurtosis of* $U$ *arises because of fourth moments of* $U$ *appear in the expression for* $\Sigma_{sand}$.

This is the remaining abstract. I also put the multivariate case here and left the univariate in the paper (appendix). Again I stress that I see this as a proposal and if you have a different opinion please change the file as you like.

## B Asymptotic Expression for Göran's Multivariate Case

This assumes that the working covaraince matrix is asymptotically correct.

Let $\mathbf{C}_i(\boldsymbol{\beta}) = (\partial\mu_i^{\mathrm{T}})/(\partial\boldsymbol{\beta})$, $\boldsymbol{\epsilon}_i(\boldsymbol{\beta}) = Y_i - \mu(\mathbf{X}_i\boldsymbol{\beta})$, $\mathbf{L}_{in}(\boldsymbol{\beta}) = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{C}_i^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{V}_i^{-1}(\boldsymbol{\beta})$, $\boldsymbol{\Omega}_n(\boldsymbol{\beta}) = n^{-1}\sum_i \mathbf{C}_i^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{V}_i^{-1}(\boldsymbol{\beta})\mathbf{C}_i(\boldsymbol{\beta})$, $\mathbf{M}_{in}(\boldsymbol{\beta}) = (\partial\mathbf{L}_{in}(\boldsymbol{\beta})/(\partial\boldsymbol{\beta}^{\mathrm{T}})$, $\mathbf{q}_i(\boldsymbol{\beta}) = \mathbf{L}_{in}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\epsilon}_i(\boldsymbol{\beta})$, $\mathbf{Q}_n^{\mathrm{T}}(\boldsymbol{\beta}) = n^{-1}\sum_i \mathbf{L}_{in}^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{V}_i(\boldsymbol{\beta})\mathbf{M}_{in}(\boldsymbol{\beta})$. The sandwich estimator and the asymptotic variance are, respectively,

$$\begin{aligned} V_{sand} &= \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\widehat{\boldsymbol{\beta}})n^{-1}\sum_{i=1}^{n}\mathbf{C}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\mathbf{V}_i^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{\epsilon}_i(\widehat{\boldsymbol{\beta}})\boldsymbol{\epsilon}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\mathbf{V}_i^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{C}_i(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{z}; \\ V_{asymp} &= \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\widehat{\boldsymbol{\beta}})n^{-1}\sum_{i=1}^{n}\mathbf{C}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\mathbf{V}_i^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{C}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Omega}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{z}. \end{aligned}$$

Remembering that $\boldsymbol{\epsilon}_i(\boldsymbol{\beta}) = \boldsymbol{\epsilon}$, Simple algebra shows that

$$\begin{aligned} n^{1/2}(V_{sand} - V_{asymp}) &= n^{-1/2}\sum_{i=1}^{n}\{\mathbf{L}_{in}(\widehat{\boldsymbol{\beta}}) - \mathbf{L}_{in}(\boldsymbol{\beta})\}^{\mathrm{T}}\boldsymbol{\epsilon}_i(\widehat{\boldsymbol{\beta}})\boldsymbol{\epsilon}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\mathbf{L}_{in}(\widehat{\boldsymbol{\beta}}) \\ &\quad + n^{-1/2}\sum_{i=1}^{n}\mathbf{L}_{in}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\{\boldsymbol{\epsilon}_i(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\epsilon}_i\}\boldsymbol{\epsilon}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\mathbf{L}_{in}(\widehat{\boldsymbol{\beta}}) \\ &\quad + n^{-1/2}\sum_{i=1}^{n}\mathbf{L}_{in}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\epsilon}_i\{\boldsymbol{\epsilon}_i(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\epsilon}_i\}^{\mathrm{T}}\mathbf{L}_{in}(\widehat{\boldsymbol{\beta}}) \\ &\quad + n^{-1/2}\sum_{i=1}^{n}\mathbf{L}_{in}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}}\{\mathbf{L}_{in}(\widehat{\boldsymbol{\beta}}) - \mathbf{L}_{in}(\boldsymbol{\beta})\} \\ &\quad + n^{-1/2}\sum_{i=1}^{n}\mathbf{L}_{in}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\{\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}} - \mathbf{V}_i(\boldsymbol{\beta})\}\mathbf{L}_{in}(\widehat{\boldsymbol{\beta}}). \end{aligned}$$

A simple expansion shows that

$$\begin{aligned} n^{1/2}(V_{sand} - V_{asymp}) &= n^{-1/2}\sum_{i=1}^{n}\mathbf{L}_{in}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\{\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}} - \mathbf{V}_i(\boldsymbol{\beta})\}\mathbf{L}_{in}(\widehat{\boldsymbol{\beta}}) \\ &\quad + 2n^{-1}\sum_{i=1}^{n}\mathbf{L}_{in}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}}\mathbf{M}_{in}(\boldsymbol{\beta})n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\} \\ &\quad + 2n^{-1}\sum_{i=1}^{n}\mathbf{L}_{in}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\epsilon}_i\mathbf{L}_{in}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\{\boldsymbol{\epsilon}_i(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\epsilon}_i\} + o_p(1). \end{aligned}$$

Since $\boldsymbol{\epsilon}_i(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\epsilon}_i \approx -\mathbf{C}_i(\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, it is easily seen that the last term in the previous expression is $o_p(1)$. Hence we have shown that

$$n^{1/2}(V_{sand} - V_{asymp}) = n^{-1/2}\sum_{i=1}^{n}\mathbf{L}_{in}^{\mathrm{T}}(\boldsymbol{\beta})\{\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}} - \mathbf{V}_i(\boldsymbol{\beta})\}\mathbf{L}_{in}(\boldsymbol{\beta})$$

$$+2n^{-1}\sum_{i=1}^{n}\mathbf{L}_{in}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\mathbf{V}_i(\boldsymbol{\beta})\mathbf{M}_{in}(\boldsymbol{\beta})n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\}+o_p(1).$$

Define $\mathbf{G}_n(\boldsymbol{\beta})=n^{-1}\sum_i\mathbf{L}_{in}^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{V}_i(\boldsymbol{\beta})\mathbf{M}_{in}(\boldsymbol{\beta})$. Since we have that

$$n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \;=\; n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\Omega}^{-1}(\boldsymbol{\beta})\mathbf{C}_i(\boldsymbol{\beta})\mathbf{V}_i^{-1}(\boldsymbol{\beta})\epsilon_i,$$

we have thus show that

$$\begin{aligned}n^{1/2}(V_{sand}-V_{asymp}) &= n^{-1/2}\sum_{i=1}^{n}\mathbf{L}_{in}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}})\{\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}}-\mathbf{V}_i(\boldsymbol{\beta})\}\mathbf{L}_{in}(\widehat{\boldsymbol{\beta}})\\ &\quad +2n^{-1/2}\sum_{i=1}^{n}\mathbf{G}_n(\boldsymbol{\beta})\boldsymbol{\Omega}^{-1}(\boldsymbol{\beta})\mathbf{C}_i(\boldsymbol{\beta})\mathbf{V}_i^{-1}(\boldsymbol{\beta})\epsilon_i,\end{aligned}$$

as claimed.

## B.1 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

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

Thus,

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

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

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

as claimed.

## B.2 Calculations in the Poisson Case

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

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