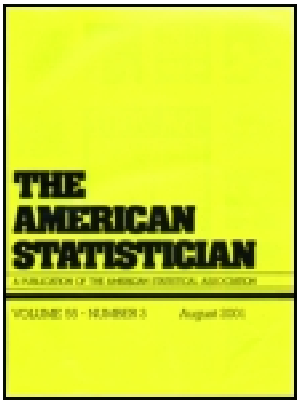

The American Statistician
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

On the Boundedness and Nonmonotonicity of Generalized Score Statistics

C. A. Field [a], Zhen Pang [b] & A. H. Welsh [c]

[a] Department of Mathematics and Statistics, Dalhousie University, Halifax N.S., Canada, B3H 3J5

[b] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371

[c] Centre for Mathematics and Its Applications, The Australian National University, Canberra ACT, 0200, Australia

Version of record first published: 15 Aug 2012.

To cite this article: C. A. Field, Zhen Pang & A. H. Welsh (2012): On the Boundedness and Nonmonotonicity of Generalized Score Statistics, The American Statistician, 66:2, 92-98

To link to this article: http://dx.doi.org/10.1080/00031305.2012.703888

On the Boundedness and Nonmonotonicity of Generalized Score Statistics

C. A. FIELD, Zhen PANG, and A. H. WELSH

We show in the context of the linear regression model fitted by Gaussian quasi-likelihood estimation that the generalized score statistics of Boos and Hu and Kalbfleisch for individual parameters can be bounded and nonmonotone in the parameter, making it difficult to make inferences from the generalized score statistic. The phenomenon is due to the form of the functional dependence of the estimators on the parameter being held fixed and the way this affects the score function and/or the estimator of the asymptotic variance. We note that in some settings, the score statistic can be bounded and nonmonotone.

KEY WORDS: Confidence intervals; Estimating equations; Quasi-likelihood estimation; Score test.

1. INTRODUCTION

The score test was introduced in the context of likelihood inference by Rao (1948) and Aitchison and Silvey (1958). It is widely used in econometrics where it is called the Lagrange multiplier test (Breusch and Pagan 1980). The test has the attractive property that the score statistic on which it is based uses only the restricted parameter estimate under the null hypothesis and avoids calculating the unrestricted parameter estimate. The score test is widely used with likelihood models and has a limiting chi-squared distribution under general regularity conditions.

Once we are no longer in the likelihood setting, the generalized score statistic becomes appropriate. One natural setting is that of estimating equations that are developed in general by Godambe (1960), in robustness with M-estimation by Huber (1964) and in longitudinal models with generalized estimating equations or GEE by Liang and Zeger (1986). This widely used approach of estimating equations provides a useful, general way to define estimators in complicated situations where the model is misspecified or incompletely specified. In particular, estimating equations do not assume (1) that the estimating equation is necessarily obtained by differentiating the log-likelihood or indeed any objective function, and (2) any model used to generate the estimating equations is not necessarily the underlying model that generated the data. In such situations, it is natural to base inference on generalized score-test statistics that are appropriate for these general frameworks. General constructions have been given by Kent (1982), Engle (1984), Breslow (1990), Roznitzky and Jewell (1990), Boos (1992), and Hu and Kalbfleisch (2000). An important feature is that unlike generalized versions of likelihood ratio tests, the distribution of generalized score statistics is asymptotically chi-squared even when the estimating equations are obtained from a model different from that that generated the data. In addition, some versions of the generalized score statistic are invariant to smooth transformations of the parameters.

The score statistic and the generalized score statistic are usually presented and discussed from the point of view of carrying out hypothesis tests. Like all tests, those based on these statistics can in principle be inverted to produce confidence intervals for the parameters. This is most straightforward when the statistic is a pivotal quantity that is monotone in the parameter of interest. We discovered that generalized score statistics can be bounded and nonmonotone in a practical context when we actually tried to use them to compute confidence intervals for variance components in mixed models and failed. For a test, boundedness means that the test may have zero power; for a confidence interval, boundedness means that the equations defining the endpoints of the interval may have no solutions and nonmonotonicity means that they may have more than one solution, making it difficult to solve the endpoint equations for confidence intervals. Thus, boundedness and nonmonotonicity have real, practical consequences for making inference in small samples. The purpose of this note is to point out that the score statistic and the generalized score statistic can be bounded and nonmonotone even in such simple problems as simple regression.

The score statistic and the generalized score statistics of Boos (1992) and Hu and Kalbfleisch (2000) are described in detail in Section 2 and applied to the slope parameter in the linear regression model with the parameters estimated by Gaussian quasi-likelihood estimation in Section 3. We show that the score statistic is bounded and monotone and the generalized score statistics can be bounded and nonmonotone even in this simple situation. In Section 4, we show that in longitudinal models the generalized score statistic is often bounded and nonmonotone and that in other simple models the score statistic itself can be bounded and nonmonotone.

C. A. Field is Professor, Department of Mathematics and Statistics, Dalhousie University, Halifax N.S., Canada B3H 3J5 (E-mail: *field@mathstat.dal.ca*). Zhen Pang is Assistant Professor, Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371 (E-mail: *zpang@ntu.edu.sg*). A. H. Welsh is E.J. Hannan Professor of Statistics, Centre for Mathematics and Its Applications, The Australian National University, Canberra ACT 0200, Australia (E-mail: *Alan.Welsh@anu.edu.au*). This research was supported by the Australian Research Council, Natural Sciences and Engineering Research Council of Canada, and Nanyang Technological University.

  DOI: 10.1080/00031305.2012.703888

2. SCORE AND GENERALIZED SCORE STATISTICS

Suppose that $\boldsymbol{\theta}_0$ is a p-vector parameter and that we are interested in making inference about a component of $\boldsymbol{\theta}_0$ that, by reordering the elements of $\boldsymbol{\theta}_0$ if necessary, we can take to be the first element θ_{01}. Let $\boldsymbol{\theta}_{02}$ be the $(p-1)$-vector of the remaining elements of $\boldsymbol{\theta}_0$ so that $\boldsymbol{\theta}_0 = (\theta_{01}, \boldsymbol{\theta}_{02}^T)^T$. Suppose we have independent observations $y_1, \ldots, y_n$ with common probability density function f that we want to use to make inferences about $\boldsymbol{\theta}_0$.

Let $\mathbf{S}(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}(y_i, \boldsymbol{\theta})$ be a set of p estimating functions for $\boldsymbol{\theta}_0$ in the sense that the estimating equations $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{0}$ have a unique solution $\hat{\boldsymbol{\theta}}$, which is a Fisher consistent estimator of $\boldsymbol{\theta}_0$. Let $\mathbf{S}(\boldsymbol{\theta})$ be partitioned as $\mathbf{S}(\boldsymbol{\theta}) = (S_1(\boldsymbol{\theta}), \mathbf{S}_2(\boldsymbol{\theta}))$ so that $S_1(\boldsymbol{\theta})$ is the (real) estimating function associated with estimating the (scalar) component θ_{01} and $\mathbf{S}_2(\boldsymbol{\theta})$ is the set of $p-1$ estimating functions associated with estimating $\boldsymbol{\theta}_{02}$. For each fixed θ_1, let $\tilde{\boldsymbol{\theta}}_2(\theta_1) \in R^{p-1}$ be the solution in $\boldsymbol{\theta}_2$ of the equations

$$\mathbf{0} = \mathbf{S}_2(\theta_1, \boldsymbol{\theta}_2).$$

We call $\tilde{\boldsymbol{\theta}}(\theta_1) = (\theta_1, \tilde{\boldsymbol{\theta}}_2(\theta_1)) \in R^p$ the profile estimator of $\boldsymbol{\theta}_0$ when the null hypothesis $\theta_{01} = \theta_1$ holds, and $S_1\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$, the first component of the estimating equation evaluated at the profile estimator, the profile estimating function for θ_{01}. We consider basing inference about θ_{01} on statistics of the form

$$Q(\theta_1) = S_1\{\tilde{\boldsymbol{\theta}}(\theta_1)\}/\widehat{U}\{\tilde{\boldsymbol{\theta}}(\theta_1)\}^{1/2}, \tag{1}$$

where $\widehat{U}\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$ is an estimator of the asymptotic variance of the profile estimating function when $\theta_{01} = \theta_1$.

When f is parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\psi}(y_i, \boldsymbol{\theta}) = f'(y_i, \boldsymbol{\theta})/f(y, \boldsymbol{\theta})$ is the derivative of $\log f$ with respect to $\boldsymbol{\theta}$, we call the p-vector $\mathbf{S}(\boldsymbol{\theta})$ the score function and the $p \times p$ matrix

$$\mathbf{W}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \mathrm{E}\{\boldsymbol{\psi}'(y_i, \boldsymbol{\theta})\}$$

the Fisher information. If $\mathbf{W}(\boldsymbol{\theta})$ is partitioned conformably with $\mathbf{S}(\boldsymbol{\theta})$, we can use the large sample theory for maximum likelihood estimators to show that the asymptotic variance of $S_1\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$ when $\theta_{01} = \theta_1$ is

$$U(\boldsymbol{\theta}_0) = \frac{1}{W(\boldsymbol{\theta}_0)^{11}} = W_{11}(\boldsymbol{\theta}_0) - \mathbf{W}_{12}(\boldsymbol{\theta}_0)\mathbf{W}_{22}(\boldsymbol{\theta}_0)^{-1}\mathbf{W}_{21}(\boldsymbol{\theta}_0). \tag{2}$$

The statistic $Q(\theta_1)$ in Equation (1) with $\widehat{U}\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$ estimating Equation (2) is called the score statistic (Rao 1948; Aitchison and Silvey 1958). More generally, in the theories of estimating equations, M-estimation and GEE, we consider estimating functions in which $\boldsymbol{\psi}(y_i, \boldsymbol{\theta})$ is typically not the derivative of $\log f$ and this needs to be taken into account in the asymptotic variance of $S_1\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$ when $\theta_{01} = \theta_1$. Let the variance matrix

$$\mathbf{V}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \mathrm{var}\{\boldsymbol{\psi}(y_i, \boldsymbol{\theta})\}$$

also be partitioned conformably with $\mathbf{S}(\boldsymbol{\theta})$. Then, with $\mathbf{b}(\boldsymbol{\theta}) = \mathbf{W}_{22}(\boldsymbol{\theta})^{-1}\mathbf{W}_{21}(\boldsymbol{\theta})$, the asymptotic variance of $S_1\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$ when $\theta_{01} = \theta_1$ is

$$U(\boldsymbol{\theta}_0) = V_{11}(\boldsymbol{\theta}_0) - 2\mathbf{b}(\boldsymbol{\theta}_0)^T\mathbf{V}_{21}(\boldsymbol{\theta}_0) + \mathbf{b}(\boldsymbol{\theta}_0)^T\mathbf{V}_{22}(\boldsymbol{\theta}_0)\mathbf{b}(\boldsymbol{\theta}_0). \tag{3}$$

As noted by Boos (1992), there are other ways to write $U(\boldsymbol{\theta}_0)$ but we only need to consider the above form. In this case, the statistic $Q(\theta_1)$ in Equation (1) with $\widehat{U}\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$ estimating Equation (3) is called a generalized score statistic (Boos 1992; Hu and Kalbfleisch 2000). When $\boldsymbol{\psi}$ is the derivative of $\log f$, we find that $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{W}(\boldsymbol{\theta})$ and $U(\boldsymbol{\theta}_0)$ in Equation (3) reduces to $U(\boldsymbol{\theta}_0)$ in Equation (2). This means that the score statistic is a particular generalized score statistic and we can treat the two types of statistic together.

An important issue for $Q(\theta_1)$ in Equation (1) is how we estimate the asymptotic variance. We need to estimate

- $\boldsymbol{\theta}_0$ when $\theta_{01} = \theta_1$, and
- $\mathbf{W}(\boldsymbol{\theta})$ and $\mathbf{V}(\boldsymbol{\theta})$.

The minimal requirement is that the estimators are consistent but, as we will see, this still leaves us with considerable freedom.

The estimator $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$ but it is more usual to use the profile estimator $\tilde{\boldsymbol{\theta}}(\theta_1)$ in score statistics. The profile estimator is used by both Boos (1992) and Hu and Kalbfleisch (2000); so, to simplify notation, we have incorporated it into the definition (1) from the start. However, it is useful to also explore the effect of using other estimators, such as $\hat{\boldsymbol{\theta}}$.

We may be able to compute $\mathbf{W}(\boldsymbol{\theta})$ and $\mathbf{V}(\boldsymbol{\theta})$ directly and avoid the need to estimate them. However, in the context of GEE, the model often only describes the first two moments of y_i and in M-estimation we often specify a vague neighborhood of models, so we do not have a fully specified parametric model and may not be able to compute $\mathbf{W}(\boldsymbol{\theta})$ and $\mathbf{V}(\boldsymbol{\theta})$. In this case, we can use the estimators

$$\mathbf{W}_y(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \boldsymbol{\psi}'(y_i, \boldsymbol{\theta})$$

and

$$\mathbf{V}_y(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \boldsymbol{\psi}(y_i, \boldsymbol{\theta})\boldsymbol{\psi}(y_i, \boldsymbol{\theta})^T - n^{-1}\sum_{j=1}^n \boldsymbol{\psi}(y_j, \boldsymbol{\theta}) n^{-1}\sum_{j=1}^n \boldsymbol{\psi}(y_j, \boldsymbol{\theta})^T,$$

which we call the observed versions of $\mathbf{W}(\boldsymbol{\theta})$ and $\mathbf{V}(\boldsymbol{\theta})$. When $\mathbf{W}(\boldsymbol{\theta})$ can be computed, Boos (1992) argued that it is more appealing than $\mathbf{W}_y(\boldsymbol{\theta})$. In the Gaussian location-scale problem, $\mathbf{W}_y(\boldsymbol{\theta})$ depends on estimates of the third and fourth moments, which Boos argues are unnecessary in Gaussian inference for the mean, whereas $\mathbf{W}(\boldsymbol{\theta})$ is diagonal and these terms do not appear. There are also hybrid schemes that use the estimates from $\mathbf{W}_y(\boldsymbol{\theta})$ of nonzero elements in $\mathbf{W}(\boldsymbol{\theta})$. In contrast, Boos observed $\mathbf{V}_y(\boldsymbol{\theta})$ seems generally to be preferred to $\mathbf{V}(\boldsymbol{\theta})$. When we compute $\mathbf{V}_y(\boldsymbol{\theta})$ to estimate $\mathbf{V}(\boldsymbol{\theta}_0)$, we have $\mathrm{E}\boldsymbol{\psi}(y_j, \boldsymbol{\theta}_0) = 0$ so we can consider the simpler estimator

$$\mathbf{V}_y^*(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \boldsymbol{\psi}(y_i, \boldsymbol{\theta})\boldsymbol{\psi}(y_j, \boldsymbol{\theta})^T.$$

The observed $\mathbf{V}_y(\boldsymbol{\theta})$ is used by Hu and Kalbfleisch (2000) and its simpler version $\mathbf{V}_y^*(\boldsymbol{\theta})$ by Boos (1992). The score statistic uses $\mathbf{W}(\boldsymbol{\theta})$ in place of $\mathbf{V}(\boldsymbol{\theta})$.

To summarize, we are considering statistics $Q(\theta_1)$ of the form (1) with different estimators $\widehat{U}\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$ of $U(\theta_0)$ in Equation (3). For the score statistic, $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{W}(\boldsymbol{\theta}_0)$ so $\widehat{U}_S\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$ has $\mathbf{V}(\boldsymbol{\theta}_0)$ estimated by $\mathbf{W}\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$; for the generalized score statistic by Hu and Kalbfleisch (2000), $\widehat{U}_{HK}\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$ has $\mathbf{V}(\boldsymbol{\theta}_0)$ estimated by $\mathbf{V}_y\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$; for the generalized score statistic by Boos (1992), $\widehat{U}_B\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$ has $\mathbf{V}(\boldsymbol{\theta}_0)$ estimated by $\mathbf{V}_y^*\{\tilde{\boldsymbol{\theta}}(\theta_1)\}$. These estimators differ only in how they estimate $\mathbf{V}(\boldsymbol{\theta}_0)$. From the definition of $\tilde{\boldsymbol{\theta}}(\theta_1)$, only the first component of $n^{-1}\sum_{i=1}^n \boldsymbol{\psi}\{y_i, \tilde{\boldsymbol{\theta}}(\theta_1)\}$ can be nonzero, so $V_{y11}\{\tilde{\boldsymbol{\theta}}(\theta_1)\} = V_{y11}^*\{\tilde{\boldsymbol{\theta}}(\theta_1)\} - n^{-1}S_1\{\tilde{\boldsymbol{\theta}}(\theta_1)\}^2$ and we have $\widehat{U}_{HK}\{\tilde{\boldsymbol{\theta}}(\theta_1)\} = \widehat{U}_B\{\tilde{\boldsymbol{\theta}}(\theta_1)\} - n^{-1}S_1\{\tilde{\boldsymbol{\theta}}(\theta_1)\}^2$. The fact that the difference between the estimators depends on θ_1 means that the generalized score statistics using these estimators are different functions of θ_1 and hence can have different shapes. If the estimators are evaluated at the estimator $\hat{\boldsymbol{\theta}}$ instead of the profile estimator $\tilde{\boldsymbol{\theta}}(\theta_1)$, then $\mathbf{V}_y(\hat{\boldsymbol{\theta}}) = \mathbf{V}_y^*(\hat{\boldsymbol{\theta}})$ and they both reduce to

$$\widehat{U}(\hat{\boldsymbol{\theta}}) = V_{y11}^*(\hat{\boldsymbol{\theta}}) - 2\mathbf{b}(\hat{\boldsymbol{\theta}})^T\mathbf{V}_{y21}^*(\hat{\boldsymbol{\theta}}) + \mathbf{b}(\hat{\boldsymbol{\theta}})^T\mathbf{V}_{y22}^*(\hat{\boldsymbol{\theta}})\mathbf{b}(\hat{\boldsymbol{\theta}}).$$

Both Boos (1992) and Hu and Kalbfleisch (2000) argued in favor of their estimates over $\widehat{U}(\hat{\boldsymbol{\theta}})$ on the grounds that their estimates make the generalized score statistic invariant to reparameterization. The converse view is that using $\widehat{U}(\hat{\boldsymbol{\theta}})$ makes the generalized score statistic for the location parameter in the location-scale problem estimated by Gaussian quasi-likelihood reduce to Student's *t*-statistic.

Under standard conditions, the distribution of $Q(\theta_1)$ in Equation (1) under $H_0: \theta_{01} = \theta_1$ is asymptotically a standard Gaussian distribution but we can also use other methods such as the bootstrap to approximate the distribution. For example, Hu and Kalbfleisch (2000) proposed using the estimating function bootstrap as an alternative way to approximate the distribution of $Q(\theta_1)$. With either distribution, it is in principle straightforward to use $Q(\theta_1)$ (or its square) to test the simple null hypothesis $H_0: \theta_{01} = \theta_1$ (Boos 1992) and the test can be inverted to construct confidence intervals for θ_{01}. Specifically, for a $100(1-\gamma)\%$ confidence interval, we need to find the upper and lower endpoints for the interval by solving equations of the form $Q(\theta_1) = q_{1-\gamma/2}$ and $Q(\theta_1) = q_{\gamma/2}$, where q_u is the uth quantile of the distribution of $Q(\theta_1)$ (Hu and Kalbfleisch 2000). This is straightforward when $Q(\theta_1)$ is monotone and unbounded in θ_1. However, $Q(\theta_1)$ can be bounded and nonmonotone in θ_1. As noted in the Introduction, for a test, boundedness means that the test may have zero power, and for a confidence interval, boundedness means that $Q(\theta_1) = q$ may have no solutions and nonmonotonicity means that $Q(\theta_1) = q$ may have more than one solution. It might be tempting to argue that we should use $\widehat{U}(\hat{\boldsymbol{\theta}})$ instead so that the denominator does not depend on θ_1. However, with general estimating equations for arbitrary models, the profile estimating function can be bounded and/or nonmonotone either from the functional dependence of $\tilde{\boldsymbol{\theta}}(\theta_1)$ on θ_1 or because the estimating function $S_1(\boldsymbol{\theta})$ is nonmonotone in θ_1 and, in the latter case, using $\widehat{U}(\hat{\boldsymbol{\theta}})$ does not resolve the problem. This raises difficulties when trying to use the $Q(\theta_1)$ to make inference for θ_1 from small samples.

3. THE SLOPE IN REGRESSION-SCALE PROBLEMS

To demonstrate and gain more insight into the potential boundedness and nonmonotonicity of $Q(\theta_1)$, we compute $Q(\theta_1)$ for the slope parameter in a regression-scale model and explore its properties. Let

$$y_i = \mu_i + e_i, \text{ with } \mu_i = \alpha + x_i\beta, \quad i = 1, \ldots, n,$$

where $\{e_i\}$ are independent and identically distributed random variables with mean zero and variance σ^2. Suppose we are interested in the slope parameter β. Write $\boldsymbol{\theta} = (\beta, \alpha, \sigma^2)^T$ so the parameter of interest β is the first element of $\boldsymbol{\theta}$. The Gaussian quasi-likelihood estimating function (Wedderburn 1974) for $\boldsymbol{\theta}$ is $\mathbf{S}(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}(y_i, \boldsymbol{\theta})$, where

$$\boldsymbol{\psi}(y_i, \boldsymbol{\theta}) = \{x_i(y_i - \mu_i)/\sigma^2,\ (y_i - \mu_i)/\sigma^2,\\ -1/2\sigma^2 + (y_i - \mu_i)^2/2\sigma^4\}^T.$$

The resulting estimating equations have a unique, explicit solution that includes the familiar least squares estimators of α and β, and the method of moments estimator of σ^2.

To simplify notation, let $\bar{x} = n^{-1}\sum_{i=1}^n x_i$, $\overline{xy} = n^{-1}\sum_{i=1}^n x_i y_i$, $\overline{x\mu} = n^{-1}\sum_{i=1}^n x_i\mu_i$, etc. Then the derivative matrix for $\mathbf{S}(\boldsymbol{\theta})$ is

$$\mathbf{W}_y(\boldsymbol{\theta}) = -\begin{pmatrix} n^{-1}\sum_{i=1}^n x_i^2/\sigma^2 & \bar{x}/\sigma^2 & (\overline{xy} - \overline{x\mu})/\sigma^4 \\ \bar{x}/\sigma^2 & 1/\sigma^2 & (\bar{y} - \bar{\mu})/\sigma^4 \\ (\overline{xy} - \overline{x\mu})/\sigma^4 & (\bar{y} - \bar{\mu})/\sigma^4 & -1/2\sigma^4 + n^{-1} \times \sum_{i=1}^n (y_i - \mu_i)^2/\sigma^6 \end{pmatrix},$$

so the expected derivative matrix for $\mathbf{S}(\boldsymbol{\theta})$ (under the model) is

$$\mathbf{W}(\boldsymbol{\theta}) = -\begin{pmatrix} n^{-1}\sum_{i=1}^n x_i^2/\sigma^2 & \bar{x}/\sigma^2 & 0 \\ \bar{x}/\sigma^2 & 1/\sigma^2 & 0 \\ 0 & 0 & 1/2\sigma^4 \end{pmatrix}.$$

If we regress y_i on $x_i - \bar{x}$ instead of x_i, then $\mathbf{W}(\boldsymbol{\theta})$ is diagonal and all three parameters are orthogonal. These calculations require only the first two moments of the model $\mathrm{E}(y|x_i) = \mu_i$ and $\mathrm{var}(y|x_i) = \sigma^2$ to be correctly specified, essentially because the estimating equations are linear. However, for nonlinear and other M-estimation equations, the off-diagonal terms are not necessarily zero, so using $\mathbf{W}(\boldsymbol{\theta})$ may not greatly simplify $\mathbf{W}_y(\boldsymbol{\theta})$. This point is often obscured by the popular but misleading traditions of assuming either that the scale parameter σ is known or that the e_i have a symmetric distribution. Explicit expressions for $\mathbf{V}_y(\boldsymbol{\theta})$ and $\mathbf{V}_y^*(\boldsymbol{\theta})$ are given in the Appendix.

Suppose that we fix β and profile out the other parameters. We obtain the estimators

$$\tilde{\alpha}(\beta) = \bar{y} - \bar{x}\beta$$
$$\tilde{\mu}_i(\beta) = \bar{y} + (x_i - \bar{x})\beta$$
$$\tilde{\sigma}(\beta)^2 = n^{-1}\sum_{i=1}^{n}\{y_i - \tilde{\mu}_i(\beta)\}^2.$$

The profile estimating function for β is $S_1\{\boldsymbol{\theta}(\beta)\} = n^{1/2}(\overline{xy} - \overline{x\tilde{\mu}}(\beta))/\tilde{\sigma}(\beta)^2 = n^{1/2}(s_{xy} - s_{xx}\beta)/\tilde{\sigma}(\beta)^2$, where $s_{xy} = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and $s_{xx} = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, so $Q(\beta)$ given by Equation (1) is

$$Q(\beta) = n^{1/2}(s_{xy} - s_{xx}\beta)/\tilde{\sigma}(\beta)^2\widehat{U}(\beta)^{1/2}, \tag{4}$$

where $\widehat{U}(\beta) = \widehat{U}\{\tilde{\boldsymbol{\theta}}(\beta)\}$. The score statistic for β uses

$$\tilde{\sigma}(\beta)^4\widehat{U}_S(\beta) = s_{xx}n^{-1}\sum_{i=1}^{n}\{y_i - \tilde{\mu}_i(\beta)\}^2,$$

the Hu and Kalbfleisch statistic uses

$$\begin{aligned}\tilde{\sigma}(\beta)^4\widehat{U}_{HK}(\beta) = n^{-1}&\sum_{i=1}^{n}\{x_i y_i - \overline{xy} - (x_i\mu_i - \overline{x\mu})\}^2\\ &- 2\bar{x}n^{-1}\sum_{i=1}^{n}\{y_i - \bar{y} - (\mu_i - \bar{\mu})\}\\ &\times \{x_i y_i - \overline{xy} - (x_i\mu_i - \overline{x\mu})\}\\ &+ \bar{x}^2 n^{-1}\sum_{i=1}^{n}\{y_i - \bar{y} - (\mu_i - \bar{\mu})\}^2,\end{aligned}$$

and the Boos statistic uses

$$\begin{aligned}\tilde{\sigma}(\beta)^4\widehat{U}_B(\beta) = n^{-1}\sum_{i=1}^{n}x_i^2(y_i - \mu_i)^2 - 2\bar{x}n^{-1}\sum_{i=1}^{n}x_i(y_i - \mu_i)^2\\ + \bar{x}^2 n^{-1}\sum_{i=1}^{n}(y_i - \mu_i)^2.\end{aligned}$$

If we regress on $x_i - \bar{x}$ instead of x_i, the score statistic is unchanged but $\mathbf{b}(\beta) = \mathbf{0}$, so the Hu and Kalbfleisch statistic uses

$$\begin{aligned}\tilde{\sigma}(\beta)^4\widehat{U}_{HK}(\beta) = n^{-1}\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})\\ - s_{xy} - \{(x_i - \bar{x})^2 - s_{xx}\}\beta]^2\end{aligned}$$

and the Boos statistic uses

$$\tilde{\sigma}(\beta)^4\widehat{U}_B(\beta) = n^{-1}\sum_{i=1}^{n}\{(x_i - \bar{x})(y_i - \bar{y}) - (x_i - \bar{x})^2\beta\}^2.$$

The score, Hu and Kalbfleisch, and Boos statistics for regression on x_i are plotted in the subfigures in Figure 1 for $n = 10$ observations generated with the x_i independent uniform (0, 1) random variables, the e_i independent standard Gaussian random variables, and $\boldsymbol{\theta} = (0.5, 1, 1)^T$. The score statistic is monotone and bounded while the Hu and Kalbfleisch and Boos statistics are nonmonotone and bounded for these data. All three subfigures have two outer dashed boundaries, while the second and the third have also a shaded gray strip above the lower boundary. The equation $Q(\beta) = q$ has two solutions for q in the gray strip, one solution for q between the upper edge of the gray strip and the upper dashed boundary, and no solutions for q either above the upper dashed boundary or below the gray strip. The Boos statistic runs into difficulties solving $Q(\beta) = q$ for $q = \pm 1.96$, the value large sample theory would suggest for 95% confidence intervals. The Hu and Kalbfleisch statistic runs into difficulties for q slightly larger than 2 and between -2 and -3 and the score statistic for q slightly larger than 3 or slightly smaller than -3. These values are not the usual values we encounter when we use the Gaussian approximation, but they can arise when we use bootstrap values. For the slope parameter β, the plots from regression on $x_i - \bar{x}$ are very similar to those for regression on x_i and hence are not included.

Figure 1 is a useful illustration but it does raise some questions such as when and how often does boundedness and/or nonmonotonicity happen? What are the values of $Q(\beta)$ at the asymptotes and of $Q(\beta)$ at the turning point when it is nonmonotone? What is the turning point? We can obtain precise answers to these questions by studying $Q(\beta)$ as a function of β.

A key simplifying step is to notice that after we square the denominators and gather the resulting terms, we can write all three functions of $Q(\beta)$ in the form

$$Q(\beta) = n^{1/2}\frac{c\beta + d}{(e\beta^2 - 2f\beta + g)^{1/2}}$$

with different values of c, d, e, f, and g. For $Q(\beta)$ to be well defined, we require the quadratic function $e\beta^2 - 2f\beta + g > 0$ for all β. This implies that $e\beta^2 - 2f\beta + g$ has no real roots and hence $eg - f^2 > 0$. This in turn implies that $e \neq 0$. Letting $\beta \to \pm\infty$, we see that $Q(\beta) \to \pm n^{1/2}c/e^{1/2}$, giving the values of the asymptotes. These depend on n and are further apart for larger sample sizes. We can study the monotonicity properties of $Q(\beta)$ by computing its derivative with respect to β and checking whether it changes sign or not. We find

$$\begin{aligned}Q'(\beta) &= n^{1/2}\frac{c}{(e\beta^2 - 2f\beta + g)^{1/2}} - n^{1/2}\frac{(c\beta + d)(2e\beta - 2f)}{2(e\beta^2 - 2f\beta + g)^{3/2}}\\ &= n^{1/2}\frac{c(e\beta^2 - 2f\beta + g)}{(e\beta^2 - 2f\beta + g)^{3/2}} - \frac{(c\beta + d)(e\beta - f)}{(e\beta^2 - 2f\beta + g)^{3/2}}\\ &= n^{1/2}\frac{-(cf + de)\beta + cg + df}{(e\beta^2 - 2f\beta + g)^{3/2}}.\end{aligned}$$

The function Q is monotone if and only if the numerator is constant, that is, $cf + de = 0$. If $cf + de \neq 0$, there is a turning point at

$$\beta_t = \frac{cg + df}{cf + de}$$

at which

$$Q(\beta_t) = \frac{n^{1/2}(c^2 g + 2cdf + d^2 e)}{(c^2 eg^2 - c^2 f^2 g - 2cdf^3 - d^2 ef^2 + 2cdefg + d^2 e^2 g)^{1/2}}.$$

To apply these calculations to the different versions of $Q(\beta)$, we compute the coefficients $c, \ldots, g$ and then the asymptotes $\pm n^{1/2}c/e^{1/2}$, the monotonicity condition $cf + de = 0$, and the turning point $\{\beta_t, Q(\beta_t)\}$. We can summarize these results as follows. The score statistic is monotone and bounded with asymptotes $\mp n^{1/2}$ as $\beta \to \pm\infty$. The Hu and Kalbfleisch and Boos

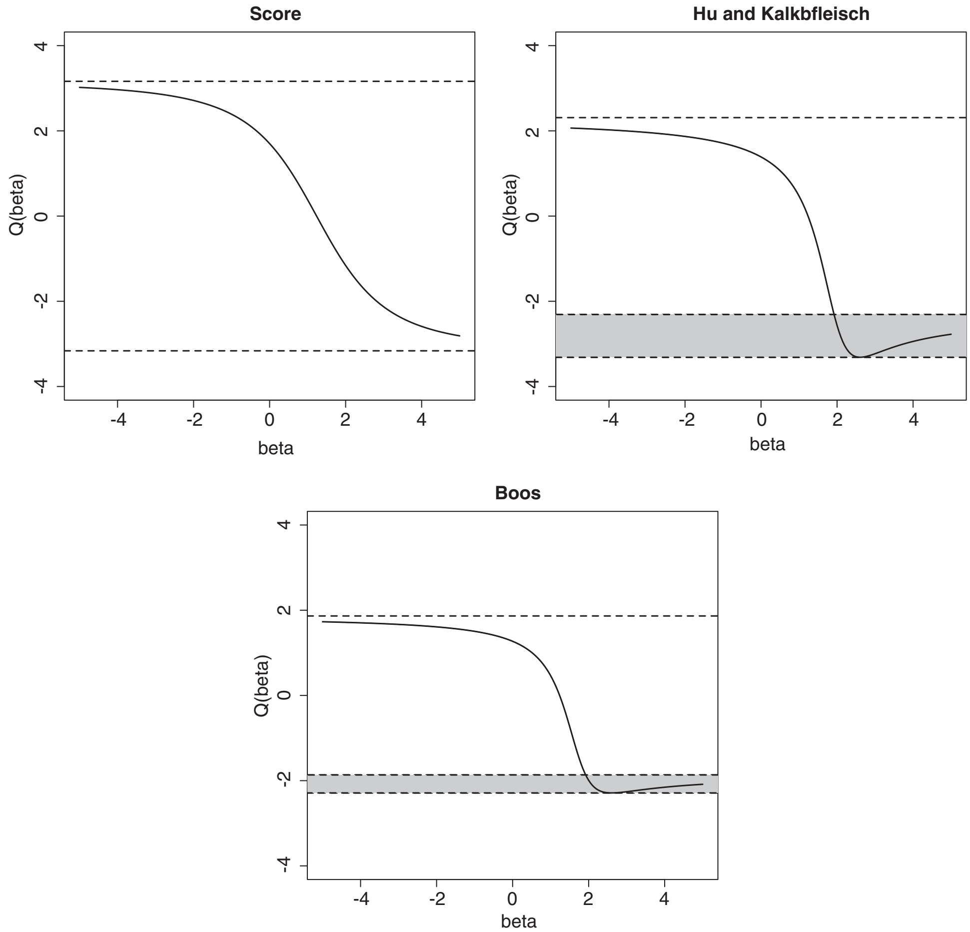

Figure 1. Plots of the score, Hu and Kalkbfleisch, and Boos score statistics for $n = 10$ observations generated from a simple linear model with uniformly distributed covariates and normally distributed errors. All three subfigures have two outer dashed boundaries, while the second and the third have also a shaded gray strip above the lower boundary. The equation $Q(\beta) = q$ has two solutions for q in the gray strip, one solution for q between the upper edge of the gray strip and the upper dashed boundary, and no solutions for q either above the upper dashed boundary or below the gray strip.

statistics for β are bounded: they asymptote to

$$\mp n^{1/2} s_{xx} \Big/ \left\{ n^{-1} \sum_i (x_i - \bar{x})^4 - s_{xx}^2 \right\}^{1/2} \quad \text{and}$$

$$\mp n^{1/2} s_{xx} \Big/ \left\{ n^{-1} \sum_i (x_i - \bar{x})^4 \right\}^{1/2},$$

respectively, as $\beta \to \pm\infty$. They are monotone in β if

$$H_\beta = s_{xx} n^{-1} \sum_i (x_i - \bar{x})^3 (y_i - \bar{y}) - s_{xy} n^{-1} \sum_i (x_i - \bar{x})^4 = 0;$$

otherwise they have a single turning point at

$$\left[s_{xy} n^{-1} \sum_i (x_i - \bar{x})^3 (y_i - \bar{y}) - s_{xx} n^{-1} \sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2 \right] \Big/ H_\beta$$

and hence are nonmonotone. These results are illustrated in Figure 1; when the Hu and Kalbfleisch and Boos statistics are nonmonotone, both asymptotes are approached from below.

The different statistics of $Q(\beta)$ differ only in the denominator. The common numerator $S_1\{\boldsymbol{\theta}(\beta)\}$ is bounded, behaving like $-n^{1/2}/\beta$ as $\beta \to \pm\infty$, and has two turning points at $\{s_{xy} \pm (s_{xx}s_{yy} - s_{xy}^2)^{1/2}\}/s_{xx}$, so is nonmonotone. The score, Hu and Kalbfleisch, and Boos statistics for β have a different shape from $S_1\{\boldsymbol{\theta}(\beta)\}$, so the properties of the $Q(\beta)$ statistics are due to both $S_1\{\boldsymbol{\theta}(\beta)\}$ and the way we choose to estimate the variance of $S_1\{\boldsymbol{\theta}(\beta)\}$.

The monotonicity condition for the generalized score statistics can be rearranged to show that $Q(\beta)$ is monotone if

$$\frac{s_{xy}}{s_{xx}} = \frac{n^{-1} \sum_i (x_i - \bar{x})^3 (y_i - \bar{y})}{n^{-1} \sum_i (x_i - \bar{x})^4}.$$

One interpretation is that, for regression of y_i on $x_i - \bar{x}$, the statistic $Q(\beta)$ is monotone if the least squares estimator of β is the same as the weighted least squares estimator of β with weights $(x_i - \bar{x})^2$. According to condition Z8 in the article by Puntanen and Styan (1989), this occurs if and only if for any b_1 and b_2 we can find c_1 and c_2 such that $0 = b_1 + b_2(x_i - \bar{x}) - c_1(x_i - \bar{x})^2 - c_2(x_i - \bar{x})^3$, for all $i = 1, \ldots, n$. This is possible when x_i takes on just two distinct values but usually not otherwise, so $Q(\beta)$ is often nonmonotone. The results depend on the estimation method (through the estimating equations) and the model being fitted but not on the actual underlying distribution. Thus, they apply even when the distribution is misspecified. Finally, the results hold for any fixed sample size n. They show that the boundedness and nonmonotonicity have decreasing impact as the sample size increases in the sense that the asymptotes and the value of the statistic evaluated at a turning point diverges at the rate $n^{1/2}$. This shows that the problem does not conflict with standard asymptotic results, although this is little consolation in practice with small samples.

4. DISCUSSION

We have shown in the context of the linear regression model fitted by Gaussian quasi-likelihood estimation that the score statistic is bounded and the generalized score statistics of Boos (1992) and Hu and Kalbfleisch (2000) for individual parameters can be bounded and nonmonotone in the parameter of interest. We obtained simple theoretical results that characterize the boundedness and nonmonotonicity of these generalized score statistics and illustrated their applicability on simulated datasets. As shown in the theoretical results, the phenomenon depends on the particular configuration of the data. It occurs in simulation runs holding the covariate x_i fixed and varying the errors e_i and also in simulation runs holding the errors e_i fixed and varying the covariate x_i. It occurs even when the parameters are orthogonal. The problem can be avoided in the regression context by treating the variance as known (actually a common strategy in illustrative examples) or by using the unrestricted estimator $\hat{\boldsymbol\theta}$ instead of $\tilde{\boldsymbol\theta}(\theta_1)$ in estimating the asymptotic variance. This latter approach is somewhat against the spirit of score-based inference and sacrifices the property of invariance under parameter transformation. More seriously, it does not work in general because the profile estimating function can be a complicated function of θ_1 through $\tilde{\boldsymbol\theta}(\theta_1)$ that is not easily made unbounded or monotone. These subtleties can make it difficult to carry out inference based on generalized score statistics.

When we fit the simple linear regression model with Gaussian quasi-likelihood estimation, the score statistic treating the errors as actually having Gaussian distributions is bounded but monotone. However, in general, score statistics can also be nonmonotone. A simple example to show this is when the data follow the simple regression model with Student t-distributed errors. The estimating function derived from the Student t-likelihood is bounded and nonmonotone and the Fisher information depends on the variance but not the slope or intercept so, if the variance is known, the score statistic inherits nonmonotonicity from the estimating function. This example illustrates that boundedness and nonmonotonicity can occur with the score statistic as well as the generalized score statistic and that it can occur even when there are no nuisance parameters.

We chose to examine the slope in the simple linear regression model to illustrate the issues in the simplest possible case. As there are alternatives to basing inference on score or generalized score statistics for this problem, its use as the example in the article risks giving the impression that the difficulties are easily avoided by using other methods. However, in the estimating equation or GEE context, there are fewer alternative choices: Likelihood ratio-like statistics are often not available and Wald statistics are usually thought to perform less well, so the problem cannot be so easily avoided. We have computed the Hu and Kalbfleisch and Boos generalized score statistics for the parameters of longitudinal data models (regression with correlated errors) used in the GEE context and shown that the generalized score statistics can be bounded and nonmonotone under both the exchangeable correlation structure and autoregression of order 1 (AR(1)) correlation structure. For example, with the AR(1) correlation structure, the generalized score functions for the intercept and slope are bounded and nonmonotone most of the time, the generalized score function for correlation is generally bounded above but monotone, and, for the variance, the Hu and Kalbfleisch statistic seems to always be unbounded and monotone while the Boos statistic seems to be always bounded below and nonmonotone. As we showed in our theoretical results, boundedness and nonmonotonicity also occur when the model is misspecified. This means that boundedness and nonmonotone behavior also occur when, as often happens with GEE, the working or fitted correlation matrix is different from the actual correlation matrix. Incidentally, while $n = 10$ is small in a simple regression problem, it is less unusual to have 10 or fewer independent groups in more structured problems such as arise with longitudinal or clustered data.

Recall that we discovered that generalized score statistics can be bounded and nonmonotone in a practical context when we actually tried to use them to compute confidence intervals for variance components in mixed models and failed. We have shown that this is a real problem with practical consequences for making inference in small samples and developed some understanding of why it occurs.

APPENDIX

To find the elements of the estimator $\sigma^4\mathbf{V}_y(\boldsymbol\theta) = (c_{ij}(\boldsymbol\theta))$ used by Hu and Kalbfleisch (2000), note that $\boldsymbol\psi(y_i, \boldsymbol\theta) - n^{-1/2}\mathbf{S}(\boldsymbol\theta)$ has elements $\{y_i - \bar y - (\mu_i - \bar\mu)\}/\sigma^2$, $\{x_i y_i - \overline{xy} - (x_i\mu_i - \overline{x\mu})\}/\sigma^2$, and $\{(y_i - \mu_i)^2 - n^{-1}\sum_{i=1}^n (y_i - \mu_i)^2\}/2\sigma^4$, respectively. So,

$$c_{11}(\boldsymbol\theta) = n^{-1}\sum_{i=1}^n \{y_i - \bar y - (\mu_i - \bar\mu)\}^2$$

$$c_{12}(\boldsymbol\theta) = n^{-1}\sum_{i=1}^n \{y_i - \bar y - (\mu_i - \bar\mu)\}\{x_i y_i - \overline{xy} - (x_i\mu_i - \overline{x\mu})\}$$

$$c_{22}(\boldsymbol\theta) = n^{-1}\sum_{i=1}^n \{x_i y_i - \overline{xy} - (x_i\mu_i - \overline{x\mu})\}^2$$

$$c_{13}(\boldsymbol\theta) = n^{-1}\sum_{i=1}^n \{y_i - \bar y - (\mu_i - \bar\mu)\} \times \left\{(y_i - \mu_i)^2 - n^{-1}\sum_{i=1}^n (y_i - \mu_i)^2\right\} \Big/ 2\sigma^2$$

$$c_{23}(\boldsymbol\theta) = n^{-1}\sum_{i=1}^n \{x_i y_i - \overline{xy} - (x_i\mu_i - \overline{x\mu})\} \times \left\{(y_i - \mu_i)^2 - n^{-1}\sum_{i=1}^n (y_i - \mu_i)^2\right\} \Big/ 2\sigma^2$$

$$c_{33}(\boldsymbol\theta) = n^{-1}\sum_{i=1}^n \left\{(y_i - \mu_i)^2 - n^{-1}\sum_{i=1}^n (y_i - \mu_i)^2\right\}^2 \Big/ 4\sigma^4.$$

For the simplified observed estimator $\sigma^4 V_y^*(\theta) = (c_{ij}^*(\theta))$ used by Boos (1992), we find

$$c_{11}^*(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}(y_i - \mu_i)^2, \quad c_{12}^*(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}x_i(y_i - \mu_i)^2$$

$$c_{22}^*(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}x_i^2(y_i - \mu_i)^2,$$

$$c_{13}^*(\boldsymbol{\theta}) = -(\bar{y} - \bar{\mu})/2 + n^{-1}\sum_{i=1}^{n}(y_i - \mu_i)^3/2\sigma^2$$

$$c_{23}^*(\boldsymbol{\theta}) = -(\overline{xy} - \overline{x\mu})/2 + n^{-1}\sum_{i=1}^{n}x_i(y_i - \mu_i)^3/2\sigma^2,$$

$$c_{33}^*(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\{(y_i - \mu_i)^2 - \sigma^2\}^2/4\sigma^4.$$

Regressing on $x_i - \bar{x}$ instead of x_i simplifies some terms but does not eliminate the third and fourth power terms.

[Received May 2011. Revised May 2012.]

REFERENCES

Aitchison, J., and Silvey, S. D. (1958), "Maximum-Likelihood Estimation of Parameters Subject to Restraints," *The Annals of Mathematical Statistics*, 19, 813–828. [92,93]

Boos, D. D. (1992), "On Generalized Score Tests," *The American Statistician*, 46, 327–333. [92,93,94,97]

Breslow, N. E. (1990), "Tests of Hypotheses in Overdispersed Poisson and Other Quasi-Likelihood Models," *Journal of the American Statistical Association*, 85, 565–571. [92]

Breusch, T. S., and Pagan, A. R. (1980), "The Lagrange Multiplier Test and its Application to Model Specification in Econometrics," *Review of Economic Studies*, 47, 239–254. [92]

Engle, R. R. (1984), "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," in *Handbook of Econometrics* (Vol. II), eds. Z. Griliches and M. Intriligator, Amsterdam: North-Holland, 775–820. [92]

Godambe, V. P. (1960), "An Optimal Property of Regular Maximum Likelihood Estimation," *The Annals of Mathematical Statistics*, 31, 1208–1212. [92]

Hu, F., and Kalbfleisch, J. D. (2000), "The Estimating Function Bootstrap," *Canadian Journal of Statistics*, 28, 449–481. [92,93,94,97]

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101. [92]

Kent, J. T. (1982), "Robust Properties of Likelihood Ratio Tests," *Biometrika*, 69, 19–27. [92]

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22. [92]

Puntanen, S., and Styan, G. P. H. (1989), "The Equality of the Ordinary Least Squares Estimator and the Best Linear Unbiased Estimator," *The American Statistician*, 43, 153–161. [96]

Rao, C. R. (1948), "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters With Applications to Problems of Estimation," *Proceedings of the Cambridge Philosophical Society*, 44, 50–57. [92,93]

Roznitzky, A., and Jewell, N. P. (1990), "Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Cluster Correlated Data," *Biometrika*, 77, 485–497. [92]

Wedderburn, R. W. M. (1974), "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, 61, 439–447. [94]