On improving the robustness and reliability of Rao’s score test

L.G. Godfrey\textsuperscript{a}, *, C.D. Orme\textsuperscript{b}

\textsuperscript{a}Department of Economics and Related Studies, University of York, Heslington, York YO1 5DD, UK
\textsuperscript{b}School of Economic Studies, University of Manchester, Manchester M13 9PL, UK

Abstract

The results of misspecification tests, based on Rao’s score principle, are now routinely reported in applied econometric work. This paper draws together some important recent results which are designed to improve: (a) the robustness of standard score tests; and (b) the reliability of the asymptotic approximations used for inferential purposes. The discussion of robustness includes (i) parametric, (ii) distributional, and (iii) higher-order moment robustness. The issue of finite sample reliability focuses on controlling the size of the score test using (i) different variance estimators in conjunction with standard asymptotic theory, (ii) finite sample corrections obtainable from higher-order asymptotic analysis, and (iii) bootstrap procedures. © 2001 Elsevier Science B.V. All rights reserved.

MSC: 62P20

Keywords: Bootstrap; Finite sample corrections; Robustness; Score tests

1. Introduction

Professor Rao’s (1948) paper on the score test provides researchers with an extremely powerful tool for inference. It is, therefore, not surprising that the score test approach has been adopted widely and, in particular, many of the checks for misspecification that are now routinely used in econometrics and time-series analysis have been derived or justified by using Rao’s results. (The score test is sometimes called the Lagrange multiplier test by econometricians.) There is an enormous literature dealing with the derivation and implementation of many different score statistics, the reinterpretation of other approaches to testing as variants of the score method, and the properties of score tests in nonstandard situations. Fortunately there are several books...
and survey articles available, e.g. Bera and Ullah (1991), Breusch and Pagan (1980), Engle (1984), Godfrey (1988), Godfrey and Tremayne (1988), Kramer and Sonnberger (1986), Maddala (1995) and White (1994). This paper focuses on recent work on modifications of the score test procedure that increase robustness and reliability in finite samples. In order to keep this paper to a manageable length, we do not attempt to cover all related areas. For example, the modified profile likelihood approach of Cox and Reid (1987) is not discussed; see Simonoff and Tsai (1994) for an application to tests for heteroskedasticity in regression models.

Any score test is designed to be (at least asymptotically) valid under a set of assumptions. If these assumptions are not satisfied, the original form of the score test can produce misleading results with a true null hypothesis being rejected far too frequently. A degree of robustness (size resistance) is often desirable; e.g., if each test being applied were robust to misspecifications for which it was not designed, there would be some scope for isolating and identifying causes of model inadequacy. Three types of robustness are discussed below: (i) parametric robustness; (ii) distribution robustness; and (iii) higher-order moment robustness. This discussion includes consideration of models in which there is not full specification of the likelihood and so score-type statistics based upon other types of estimation functions are required for testing.

Robustification is intended to give score-type procedures asymptotic validity over a wider range of situations, but asymptotic theory may provide inaccurate predictions about behaviour in finite samples. The quality of the performance of score tests in finite samples is clearly of great practical importance. Various approaches have been taken to the derivation of modifications that improve finite sample properties of score tests and some of these are examined below, along with the summaries of relevant Monte Carlo evidence.

The plan of this paper is as follows. Section 2 contains details of notation and preliminary results. The scope for making score-type tests more robust is discussed in Section 3. Small sample corrections are considered in Section 4. Some concluding remarks are made in Section 5.

2. Notation and preliminary results

To save space, details of regularity conditions are omitted and it is simply assumed that standard results of asymptotic theory are available; see Amemiya (1985, Chapter 4) for a statement of suitable conditions.

First consider the case in which a likelihood framework is used. The $p$-dimensional parameter vector for the likelihood function of the true data process is denoted by $\theta$. A sample of $n$ observations provides the data for estimation and testing. Let the log of the likelihood function be denoted by $L(\theta)$. (Notation is simplified by not showing the dependence of $L(\theta)$ and other functions on $n$.) The first- and second-order partial derivatives of $L(\theta)$ with respect to $\theta$ are the elements of

$$d(\theta) = \partial L(\theta)/\partial \theta, \quad (p \times 1)$$  

(2.1)
respectively. It is assumed that appeal can be made to a suitable Central Limit Theorem with

\[ n^{-1/2} d(\theta^*) \sim N(0, \mathcal{I}), \]  

(2.3)

where \( \theta^* \) is the true value of \( \theta \), “\( \sim \)” is read as “is asymptotically distributed as” and \( \mathcal{I} \) is the average information matrix, i.e. \( \mathcal{I} = p \lim \frac{1}{n} \mathcal{D}(\theta) \).

Following Godfrey (1996a), \( \theta \) is partitioned as \( \theta' = (\theta'_1, \theta'_2, \theta'_3) \) with the \( \theta_i \) being functionally independent and of dimension \( p_i, i = 1, 2, 3 \). The corresponding subvectors of \( d(\theta) \) and submatrices of \( \mathcal{D}(\theta) \) and \( \mathcal{I} \) are denoted by \( d_i(\theta), D_{ij}(\theta) \) and \( \mathcal{I}_{ij} \) for \( i, j = 1, 2, 3 \). The model to be tested by means of a score test is assumed to be the special case derived by imposing \( \theta'_2 = 0 \) and \( \theta'_3 = 0 \). For the purposes of asymptotic analysis, it is assumed that the actual data process is characterized by a sequence of local alternatives with

\[ \theta'_i = n^{-1/2} \delta_i, \quad 0 < \delta'_i \delta_i < \infty, \quad i = 2, 3. \]

These local alternatives are used to ensure that, in general, the power function of a test tends to a limit between the nominal size and unity as \( n \to \infty \).

The maximum-likelihood estimator (MLE) for the model under test is denoted by \( \hat{\theta}_1 \), so that \( \hat{\theta}' = (\hat{\theta}'_1, 0', 0') \) can be regarded as the restricted MLE for the more general model with parameter vector \( \theta \). The unrestricted MLE for the latter model is denoted by \( \tilde{\theta}' = (\tilde{\theta}'_1, \tilde{\theta}'_2, \tilde{\theta}'_3) \). The first-order conditions for MLE imply that \( d_1(\hat{\theta}) = 0 \) and \( d(\tilde{\theta}) = 0 \). The MLE is, of course, a special case of the extremum estimator (\( M \)-estimator, see Huber, 1981) and score-type tests are often used with the latter method. Consider then the situation in which the likelihood is not available because the precise form of the underlying probability law is not part of the model specification and estimation is instead based upon maximization of an objective function \( Q(\theta) \). To economize on notation, \( \hat{\theta} \) and \( \tilde{\theta} \) will again denote the restricted and unrestricted estimators. Under conventional regularity conditions with \( Q(\theta) \) being \( O_p(n) \), it can be shown that

\[ g(\theta) = \frac{\partial Q(\theta)}{\partial \theta}, \quad (p \times 1) \]  

(2.4)

which is the analogue of the score \( d(\theta) \), is such that

\[ n^{-1/2} g(\theta^*) \sim N(0, B(\theta^*)), \]  

(2.5)

in which

\[ B(\theta^*) = \lim n^{-1} E[g(\theta^*)g(\theta^*)'], \quad (p \times p). \]  

(2.6)

It can also be shown that

\[ n^{1/2}(\hat{\theta} - \theta^*) \sim N(0, A(\theta^*)^{-1}B(\theta^*)A(\theta^*)^{-1}), \]  

(2.7)

in which

\[ A(\theta^*) = \lim n^{-1} E(\partial^2 Q(\theta^*)/\partial \theta \partial \theta'), \quad (p \times p). \]  

(2.8)
It is important to note that when $Q(\theta)$ is the log of a likelihood function, i.e. $Q(\theta) = L(\theta)$, the information matrix equality applies so that $A(\theta^*) + B(\theta^*) = 0$, but that this relationship will not hold for other choices of estimation function.

Irrespective of whether estimation is by MLE or by some more general extremum method, it is useful, when discussing certain forms of robustness of score tests, to appeal to a result on two stage estimators. This general result concerns the construction, in a single Newton–Raphson iteration from a first stage root-$n$ consistent estimator, of an estimator that is asymptotically equivalent to $\tilde{b}_{DC2}$. Let $DY!\left[ b_{DC2} \right]$ denote the Dqqrst-stage estimator, so that $(DY!\left[ b_{DC2} \right] - b_{DC2})$ is $O_p(n^{-1/2})$. The two-stage estimator $\hat{\theta}$, defined by

$$\hat{\theta} = \tilde{\theta} - \left[ \partial^2 Q(\tilde{\theta})/\partial \theta \partial \theta' \right]^{-1} \partial Q(\tilde{\theta})/\partial \theta,$$

is asymptotically equivalent to $\tilde{\theta}$, i.e. $(\hat{\theta} - \tilde{\theta})$ is $O_p(n^{-1})$. Eq. (2.9) is derived by linearization of the first-order conditions for $\tilde{\theta}$ and $Q(.)$ should be replaced by $L(.)$ when it is applied to MLE; see, for example, Rothenberg and Leenders (1964, p. 69). The usefulness of (2.9) for this paper derives from the fact that both score and Neyman’s (1959) $C(x)$ tests can be interpreted as Wald-type checks based upon an appropriate subvector of a two-stage estimator; see Pagan (1986) and Godfrey (1988, pp. 25–28). Further details are given when relevant in the next section.

3. Robustness and score tests

This section is divided into three parts, each corresponding to a different type of robustness. In the first part, it is assumed that robustification of a test of $\theta_2 = 0$ is based upon explicit specification of an estimation objective function that depends upon $\theta_3$, as well as $\theta_1$ and $\theta_2$. (Thus the intention is to derive a test that is robust to a specified departure from standard assumptions.) In the other two parts, there is no specification of a function that depends upon $\theta_3$ and robustification applies to quite general departures from the assumptions underpinning the usual tests of $\theta_2 = 0$. For each part, results from asymptotic analysis and Monte Carlo studies are reported.

3.1. Parametric robustness

Suppose that, as allowed by the notation of Section 2, the parameter vector for the family of alternatives which includes the true data process can be divided into three subvectors. Consider using a score test of $\theta_2 = 0$, conditional upon $\theta_3 = 0$. The score test is said to have the property of parametric robustness if local departures from the untested restrictions $\theta_3 = 0$ do not affect the asymptotic distribution of the score statistic under $\theta_2 = 0$. (Some readers may prefer to refer to such tests as being insensitive, rather than robust.) This type of robustness is helpful when the researcher is trying to use outcomes of separate tests to guide respecification after the null model has been found to be inconsistent with the sample data; see Jaggia and Trivedi’s (1994, pp. 274–275) discussion of three approaches to testing. The concept of parametric robustness is
also relevant to the important practical problem of testing when the alternative model used to derive a score test is underspecified. Bera and Yoon (1993) refer to such misspecification of the alternative as a type-III error. The consequences of such errors and the modifications required to achieve robustness can be outlined using the models and results of Section 2. Detailed general analyses of the asymptotic behaviour of tests when both the null and selected alternative models are incorrect are provided by Davidson and MacKinnon (1987) and Saikkonen (1989).

In a classical likelihood-based framework, the score test of \( \theta_2 = 0 \) is as a check of the significance of \( d_2(\hat{\theta}) \). Godfrey (1996a) shows that

\[
\begin{align*}
n^{-1/2}d_2(\hat{\theta}) &= n^{-1/2}d_2(\theta^0) - J_{22}J_{11}^{-1}n^{-1/2}d_1(\theta^0) + \lambda_{23} + o_p(1), \\
\lambda_{23} &= (J_{22} - J_{21}J_{11}^{-1}J_{12})\hat{\delta}_2 + (J_{23} - J_{21}J_{11}^{-1}J_{13})\hat{\delta}_3.
\end{align*}
\]

in which

\[
\lambda_{23}^* = (J_{23} - J_{21}J_{11}^{-1}J_{13})\hat{\delta}_3.
\]

If the restrictions of \( \theta_2 = 0 \) are true, then \( \hat{\delta}_2 = 0 \) and \( \lambda_{23} \) reduces to

\[
\lambda_{23}^* = (J_{23} - J_{21}J_{11}^{-1}J_{13})\hat{\delta}_3.
\]

The first two terms of the right-hand side of (3.1) are asymptotically normally distributed with zero mean vectors. Hence \( \lambda_{23}^* \) must be a vector with every element equal to zero for the standard score test of \( \theta_2 = 0 \) to be asymptotically valid. If \( \lambda_{23}^* \neq 0 \), the usual score statistic is asymptotically distributed as a noncentral, rather than central, \( \chi^2(p_2) \) variable and the large sample probability of rejecting the true restrictions \( \theta_2 = 0 \) is greater than the selected significance level.

The score test of \( \theta_2 = 0 \), therefore, has the property of parametric robustness if \( \lambda_{23}^* = 0 \) for all nonzero values of \( \hat{\delta}_3 \). This condition is equivalent to

\[
J_{23} - J_{21}J_{11}^{-1}J_{13} = 0.
\]

In the context of maximum-likelihood estimation and testing, the information matrix equality applies and the covariance matrix in (2.7) simplifies to \( J^{-1} \). Eq. (3.4) can then be interpreted as the condition that the unrestricted MLE \( \hat{\theta}_2 \) and \( \hat{\theta}_3 \) are asymptotically independent. In many important cases, this condition does not hold, and so there is a need to consider modifications of the score test that are robust.

The Wald test of the significance of \( \hat{\theta}_2 \) is clearly robust to nonzero values of \( \hat{\delta}_3 \) because it uses the results of unrestricted estimation. The test statistic for this test is given by

\[
W_2 = \hat{\theta}_2'[J^{22}]^{-1}\hat{\theta}_2,
\]

in which \( J^{22} \) is the appropriate submatrix of an estimate of \( J^{-1} \) derived from \( \hat{\theta} \). A modification of the standard score test that is asymptotically equivalent to \( W_2 \), and is, therefore, robust, can be obtained by using a two-stage estimate of the type defined by (2.9). When \( \theta_2 \) and \( \theta_3 \) are both at most \( O(n^{-1/2}) \), \( \hat{\theta} \) is a root-\( n \) consistent estimator and so can be used as the first-stage estimate. Using (2.9) with \( \hat{\theta} = \hat{\theta} \) yields

\[
\hat{\theta}_2 = -(\hat{\delta}_2 + \hat{\delta}_3),
\]

(3.6)
in which \( \hat{\theta} \) denotes evaluation of partial derivatives at \( \theta = \hat{\theta} \). Eq. (3.6) provides an estimator which is asymptotically equivalent to \( \hat{\theta}_2 \), under the assumptions of the analysis. Standard results on the irrelevance of premultiplying a score vector by a nonsingular matrix imply that the test of the significance of \( \hat{\theta}_2 \) is equivalent to a test of the significance of

\[
\hat{d}_2^* = \hat{d}_2 + (\hat{D}^{-22})^{-1}\hat{D}^{-23}\hat{d}_3,
\]  

which reveals the modification of the score \( \hat{d}_2 \) that gives a robust test when \( \delta_3 \neq 0 \). If \(-n^{-1}\hat{D}\) is used to estimate \( \hat{\theta} \), and hence for the purposes of replacing \( \hat{\theta}^{-22} \) in (3.5), the modified score vector can be written as

\[
\hat{d}_2^* = \hat{d}_2 - [\hat{\theta}_{23} - \hat{\theta}_{21}\hat{\theta}_{11}^{-1}\hat{\theta}_{13}] [\hat{\theta}_{33} - \hat{\theta}_{31}\hat{\theta}_{11}^{-1}\hat{\theta}_{13}]^{-1}\hat{d}_3,
\]  

and this expression makes it clear that the adjustment to the score is asymptotically negligible when (3.4) is satisfied and the conventional score test is robust.

Bera and Yoon (1993) provide an explicit form of the \( \chi^2 \) statistic for carrying out a joint test of the significance of the elements of the modified score \( \hat{d}_2^* \). Their approach involves adjustments of the mean and variance of the standard score statistic and yields a test that is asymptotically equivalent to a Wald-type test using a two stage estimate, i.e. to a \( C(bVT) \) test.

The modified score test using \( \hat{d}_2^* \) is more widely asymptotically valid than the original score test based upon \( \hat{d}_2 \) because the latter is, in general, invalid when \( \delta_3 \neq 0 \). There is, however, a cost as well as a benefit associated with the modification. By construction, the modified score test is asymptotically equivalent to a Wald test after unrestricted estimation of all subvectors of \( \theta \). If \( \theta_3 = 0 \), there is asymptotic inefficiency relative to estimators that impose these restrictions. Consequently, when \( \delta_2 \neq 0 \) and \( \delta_3 = 0 \), the Wald statistic \( W_2 \) will be asymptotically distributed as noncentral \( \chi^2 \) with a noncentrality parameter that is smaller than that of the Wald test that uses estimates of \( \theta_1 \) and \( \theta_2 \) with \( \theta_3 = 0 \) imposed. It follows that when modification is not required the modified test has smaller asymptotic local power than the unadjusted form; see Bera and Yoon (1993) for detailed expressions for the relevant noncentrality parameters.

The small sample performance of modified score tests has received attention in recent papers by Anselin et al. (1996) and Bera et al. (1996). The former study deals with tests for spatial dependence and the latter with tests for error component models. Encouraging Monte Carlo results are contained in both papers. Modified score tests are found to have good finite sample properties and seem useful in determining the direction of departure from the null hypothesis when it is false.

The extension of the results obtained for MLE to be relevant to general extremum estimation is straightforward. Eq. (3.1) is replaced by

\[
n^{-1/2}\hat{g}_2 = n^{-1/2}g_2^o - A_{21}^o[A_{11}^o]^{-1}n^{-1/2}g_1^o + \lambda_{23} + o_p(1),
\]

in which \( ^{\sim} \) and \( ^{o} \) denote evaluation at \( \theta = \hat{\theta} \) and \( \theta = \theta^o \), respectively, and

\[
\lambda_{23} = -\{(A_{22}^o - A_{21}^o[A_{11}^o]^{-1}A_{12}^o)\delta_2 + (A_{23}^o - A_{21}^o[A_{11}^o]^{-1}A_{13}^o)\delta_3\}.
\]
It follows from (3.9) that, when \( b_{SO}^2 = 0 \) and \( b_{SO}^3 = 0 \), \( n^{-1/2} \hat{g}_2 \) is asymptotically normally distributed with zero mean vector. Combining (3.9) with the results quoted in Section 2 yields

\[
\hat{V}(n^{-1/2} \hat{g}_2) = \hat{B}_{22} - \hat{A}_{21}[\hat{A}_{11}]^{-1}\hat{B}_{12} - \hat{B}_{21}[\hat{A}_{11}]^{-1}\hat{A}_{12} + \hat{A}_{21}[\hat{A}_{11}]^{-1}\hat{B}_{11}[\hat{A}_{11}]^{-1}\hat{A}_{12},
\]

(3.11)
as a consistent estimator of the covariance matrix required to derive the \( b_{US}^2 \) test of the significance of \( n^{-1/2} \hat{g}_2 \). Expression (3.11) is relatively cumbersome because, in contrast to MLE, it cannot be assumed that the information matrix type equality \( A(\theta^\circ) + B(\theta^\circ) = 0 \) holds.

Eq. (3.10) implies that the score test using extremum estimates is robust to nonzero values of \( \delta_3 \) if

\[
(A_{23}^\circ - A_{21}^\circ[A_{11}^\circ]^{-1}A_{13}^\circ) = 0.
\]

(3.12)

This condition does not involve \( B(\theta^\circ) \) of (2.6) and so, in general, has no implication for the covariance matrix of the asymptotic distribution of the unrestricted estimators, as given in (2.7). Godfrey and Orme (1996) discuss the absence of a link between robustness and independence of test statistics for extremum estimation and provide examples.

In terms of developing a robust score-type test for extremum estimation, the approach is very similar to that for MLE. Eq. (3.5) is modified by replacing the submatrix of \( \mathcal{J}^{-1} \) by the corresponding submatrix of \( [A^\circ]^{-1}B^\circ[A^\circ]^{-1} \). In (3.6)–(3.8), \( d, D \) and \( \mathcal{J} \) are replaced by \( g, \hat{c}^2 Q/\partial \theta \partial \theta^\prime \) and \(-A\), respectively.

The use of \( \hat{\theta} \) as the first-stage estimator to derive robust tests is acceptable when, as in this paper, only local departures from the null are considered in asymptotic analyses. The combination of asymptotic theory and local departures is intended to provide a guide to actual behaviour when the evidence of misspecification is not overwhelming. When there are very large departures from the untested restrictions of \( \theta_3 = 0 \), an assumption of fixed alternatives may be more appropriate. If \( \theta_3^\circ \) were \( O(1) \), then \( \hat{\theta} - \theta^\circ \) would not be at most \( O_p(n^{-1/2}) \), \( \hat{\theta} - \hat{\theta} \) would not be at most \( O_p(n^{-1}) \), and the modified score (3.8) would not be robust. Jaggia and Trivedi (1994) use a \( C(z) \) test that is valid when \( \theta_2^\circ = 0 \) and \( \theta_3^\circ \) is \( O(1) \). This procedure, which Jaggia and Trivedi call the conditional score test with proper conditioning, has, as the implicit first stage estimator, \( \hat{\theta} = (\hat{\theta}_1, \hat{\theta}^\prime, \hat{\theta}_3) \) with \( (\hat{\theta}_1 - \theta_1^\circ) \) and \( (\hat{\theta}_3 - \theta_3^\circ) \) both being \( O_p(n^{-1/2}) \), so any subvector of \( \theta \) that is not under test must be estimated consistently.

It seems reasonable to conjecture that the small sample performance of the Jaggia–Trivedi procedure will be better than that of the test based upon the modified score (3.8) when there is a really substantial departure from \( \theta_3 = 0 \). If there is reason, on a priori grounds, to suspect a departure of this type, a score test of \( \theta_2 = 0 \) using the maximizers of \( L(\theta_1, 0, \theta_3) \) or \( Q(\theta_1, 0, \theta_3) \) seems appropriate. Indeed this test is a special case of the general class proposed by Jaggia and Trivedi.

Jaggia and Trivedi (1994) discuss conditional score tests, separate (non-robust) score tests, and a joint test in the context of checking for heterogeneity and state dependence.
in duration models. They carry out a number of Monte Carlo experiments and summarize their findings as follows: the non-robust score tests can be very misleading, e.g. the separate test for heterogeneity has rejection rates between 95% and 100% when the nominal size is 5%, heterogeneity is absent but there is duration dependence; modified score tests using criteria like (3.8) perform much better but sometimes have rejection rates that are rather larger than the nominal size; a version of their $C(x)$ test works better than the modified score test; and the joint test also performs well. Bera and Yoon (1991) show that Saikkonen’s (1989) asymptotically valid results are useful in understanding the Monte Carlo evidence obtained by Jaggia and Trivedi.

3.2. Distributional robustness

The assumption that the data come from the general model specified for estimation and testing is now relaxed so that the assumed log likelihood, denoted by $L_a(\theta)$, is not the true function $L(\theta)$. (To avoid overlap with the analysis of the previous section, it is assumed that $L(\theta)$ does not contain $L_a(\theta)$ as a special case derived by imposing $\theta_3 = 0$.) It might be thought that, in such situations, robust tests could be developed by viewing $L_a(\theta)$ as a $Q$-type objective function and then making appeal to results on testing after extremum estimation. Unfortunately, when the distribution is misspecified, consistent estimation of parameters and covariance matrices required for asymptotically valid tests is not usually possible; see White (1983, 1994, pp. 194, 195). If the data are independently and identically distributed (iid), consistent estimation of relevant covariance matrices is possible; see White (1982). In forming such consistent estimators, it must be recognized that the information matrix equality cannot be expected to hold and so, for example, the covariance matrix of the unrestricted maximizer of $L_a(\theta)$ can be estimated using $A(\hat{\theta})^{-1}B(\hat{\theta})A(\hat{\theta})^{-1}$, but not $-A(\hat{\theta})^{-1}$. Discussions of pseudo maximum-likelihood theory and applications are provided by Gourieroux et al. (1984a, b). Kent (1982) studies robust properties of likelihood ratio tests and makes the point that the vector $\theta$ is best viewed as a convenient tool for summarizing the data in his framework. Similar remarks apply when discussing score tests at the same level of generality.

Some useful results on robust tests are available if the level of generality is reduced by focusing on particular types of model. In particular, many robust tests for linear regression models can be obtained. Suppose that the assumed null model is

$$y = X_1 \beta_1 + u, \quad u \sim N(0, \sigma^2 I_n), \quad (3.13)$$

in which $X_1$ is a $n$ by $(p_1 - 1)$ matrix and $\theta'_1 = (\beta'_1, \sigma^2)$, but that the true error distribution is not normal. A large number of score and other tests can be implemented by the method of variable addition, that is by testing $\tau = 0$ in the expanded regression model

$$y = X_1 \beta_1 + T\tau + u, \quad (3.14)$$

for some suitable matrix of test variables $T$; see Godfrey (1988, Chapter 4) and Pagan (1984). Under the false assumption of normality, the MLE are obtained by ordinary
least squares (OLS) and the scaled score for testing \( \tau = 0 \) is proportional to \( n^{-1/2} T' \hat{u} \), where \( \hat{u} \) is the residual vector from the OLS estimation of (3.13). It is simple to show that \( n^{-1/2} T' \hat{u} = n^{-1/2} \hat{T}' u \), where \( \hat{T} \) is the matrix of residuals from the multivariate OLS regression of \( T \) on \( X_1 \). The distributional robustness of the score test of \( \tau = 0 \) derived using the incorrect normal family of distributions, therefore, requires only that

\[
\sum_{i=1}^{n} u_i^2 \sim N(0, \sigma^2 \lim n^{-1} \hat{T}' \hat{T}),
\]

and this is true for many nonnormal distributions. Thus, provided appeal can be made to a Central Limit Theorem to justify (3.15), score tests that can be calculated by variable addition are asymptotically valid. (These tests, being based upon an incorrect likelihood, cannot, however, be assumed to be asymptotically optimal.)

Godfrey and Orme (1994) use Monte Carlo experiments to examine the impact of nonnormality on several general checks for regression equations. They find that the predictions of asymptotic theory are reflected by estimates of finite sample significance levels that are reasonably close to the nominal size. Tests of the type considered by Godfrey and Orme, along with other variable addition tests, are often carried out as (asymptotically valid) \( F \) tests. There is a large literature on the effects of nonnormality on regression \( F \) tests; for a recent example, see Ali and Sharma (1996), who provide many useful references. The magnitude of such effects depends, in part, upon features of the data on regressors, e.g. the extent of leveraged observations; see Ali and Sharma (1996) for a detailed analysis. If the regressor observations lead to concern about the small sample robustness of variable addition \( F \) tests, then it may be worth considering the use of nonparametric bootstrap estimates of critical values.

There is a score test for linear regression models that is widely used and is not conveniently calculated by variable addition. This is the check for heteroskedasticity proposed by Breusch and Pagan (1979) and Cook and Weisberg (1983). These authors assume normality and make use of the properties of the normal distribution when deriving the score test statistic. Suppose that the alternative can be written, at least under local departures from homoskedasticity, as

\[
\sigma_i^2 = \sigma^2 h(\eta_i), \quad i = 1, \ldots, n,
\]

in which \( h(0) = 1 \) and the vector of scedastic variables \( z_i \) satisfies regularity conditions; see Breusch and Pagan (1979) and Cook and Weisberg (1983) for discussions of the choice of scedastic functions and variables. Let the OLS residuals for the null model (3.13) be denoted by \( \hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n \) and \( \hat{\sigma}^2 = n^{-1} \sum_i \hat{u}_i^2 \). The score statistic for testing \( \eta = 0 \) is the explained sum of squares from the OLS regression of \( (\hat{u}_i^2 - \hat{\sigma}^2) \) on the \( z_i \) and an intercept term, divided by \( 2\hat{\sigma}^4 \). The division of the explained sum of squares from the auxiliary regression by \( 2\hat{\sigma}^4 \) is appropriate because the variance of \( u_i^2 \) is \( 2\sigma^4 \) when the errors are normally distributed and the null hypothesis is true. If the errors are not normally distributed, the conventional score test is, in general, asymptotically invalid. This lack of robustness is observed in the simulation results reported by Evans (1992).
Koenker (1981) provides a simple adjustment to the score test for heteroskedasticity that makes it asymptotically valid for nonnormal distributions with first four moments finite. Koenker’s robust score statistic is \( nR^2_K \), where \( R^2_K \) is the \( R^2 \) statistic from the auxiliary regression described above, or equivalently from the regression of \( \hat{u}_i^2 \) on the \( z_i \) and an intercept term. Godfrey (1996b) finds that this modification is not always successful in controlling finite sample significance levels when there are nonnormal errors, with some estimated rejection rates being about twice the nominal value in the presence of skewness.

Godfrey also examines Glejser’s (1969) test for heteroskedasticity. Glejser’s statistic, which is calculated from the auxiliary regression of \( |\hat{u}_i| \) on \( z_i \) and an intercept term, can be interpreted as the score test when the errors have the double exponential distribution. The asymptotic results given by Godfrey indicate that it has the property of distributional robustness, provided that \( \text{Prob}(u_i > 0) = 0.5 \) for the true distribution. This condition is satisfied when the error distribution is symmetric. Machado and Santos-Silva (2000) modify the Glejser test to achieve asymptotic validity whether or not the errors have a symmetric distribution and obtain encouraging Monte Carlo results.

3.3. Higher-order moment robustness

It is often the case that, when testing some hypothesis concerning the parameters that characterize one moment of a distribution, auxiliary assumptions about higher-order moments are made. For example, a test relevant to the mean function will usually be derived under assumptions about variances and possibly even higher-order moments. If, when testing a true null hypothesis, the asymptotic distribution of a test statistic is not detected when higher-order moments do not satisfy the assumptions that underpin standard score tests, the test is said to have the property of higher-order moment robustness.

It was noted above that many score-type criteria for linear regression models can be computed by the method of variable addition. Conventional significance tests of additional regressors, e.g. an \( F \) test, while not requiring normality for asymptotic validity, are based upon the assumptions that the errors are uncorrelated and homoskedastic. If the covariance matrix of the errors is a symmetric positive-definite matrix \( \Omega \), which is not proportional to \( I_n \), standard tests are invalid.

Consider testing (3.13) against (3.14). In this case, the score-type test is a check of the joint significance of the elements of the OLS estimator of \( \tau \), i.e. of \( \hat{\tau} = (\hat{T}'\hat{T})^{-1}\hat{T}'y \), where \( \hat{T} \) is as defined in Section 3.2. Under the null hypothesis and quite general conditions, it can be shown that, when \( E(\omega\omega') = \Omega \),

\[
n^{1/2}\hat{\tau} \overset{d}{\sim} N(0, V_H), \tag{3.17}
\]

in which

\[
V_H = p\lim n(\hat{T}'\hat{T})^{-1}(\hat{T}'\Omega\hat{T})(\hat{T}'\hat{T})^{-1}. \tag{3.18}
\]
If $p \lim n^{-1}(\tilde{T}' \Omega \tilde{T})$, and hence $V_{11}$, can be estimated consistently in the presence of unspecified autocorrelation and/or heteroskedasticity, then robust tests of $\tau = 0$ can be obtained. The problem of consistent covariance matrix estimation has been the subject of considerable interest in econometrics. Newey and West (1987) describe a simple estimator; also see Newey and West (1994) and West (1997). Hansen (1992) provides a consistency proof for a general kernel-based estimator. Andrews and Monahan (1992) and Passos (1994) report Monte Carlo evidence on the finite sample usefulness of various covariance matrix estimators. This evidence suggests that asymptotic theory will certainly not always provide a reliable guide to finite sample behaviour. As with nonnormality, leverage points are relevant to the performance of tests using autocorrelation and heteroskedasticity robust covariance matrix estimates. It appears that some care must be taken when interpreting the outcomes of such tests in applied work.

The special case in which the errors of a linear model are uncorrelated and heteroskedastic has been discussed by several authors; see, for example, Chesher and Austin (1991), Chesher and Jewitt (1987), Eicker (1963), Messer and White (1984), MacKinnon and White (1985), and White (1980). The Monte Carlo results on different approaches to estimating the heteroskedasticity robust covariance matrix point to the importance of the regressor matrix and reveal that substantial finite sample biases can be present.

In the absence of a covariance matrix estimate that can be relied upon to yield accurate inferences, it may be useful to investigate the application of an appropriate form of the bootstrap to estimate the finite sample critical values of tests that are designed to be robust to autocorrelation and heteroskedasticity. Jeong and Maddala (1993, Section 4) discuss the application of bootstrap methods in the context of linear regression models with non-iid errors.

Given that Koenker’s (1981) widely used Studentized score test for heteroskedasticity can be computed as a test for significance in a linear auxiliary regression, it is not surprising that it has been modified to be based upon a robust covariance matrix estimate. Hsieh (1983) derives a variant of the score test for heteroskedasticity that is robust to heterokurticity. (Hsieh generalizes a test due to White (1980) which is a special case of Koenker’s procedure, but the extension of Hsieh’s results to the latter test is straightforward.)

Results on tests that have the property of higher-order moment robustness are not limited to diagnostic checks for linear regression models. Davidson and MacKinnon (1985) discuss heteroskedasticity robust tests for nonlinear regression models. Wooldridge (1990) considers a class of nonlinear models and proposes a general approach to obtaining regression-based tests that are robust to the failure of some assumptions about higher-order moments that are made in the literature on score tests. An interesting feature of Wooldridge’s approach is that he constructs test variables in such a way that estimation effects do not have to be considered when deriving the asymptotic distribution of test statistics, provided that the estimator for the null model is root-$n$ consistent.
The main features of Wooldridge-type tests can be illustrated by considering the simple example of testing (3.13) using the test variables of $T$. Suppose that, rather than using OLS, some other estimator of $\beta_1$, denoted by $\hat{\beta}_1$, is obtained with associated residual vector $\hat{u}$. Like $\hat{\beta}_1$, $\hat{\beta}_1$ is consistent when (3.13) gives the correct mean function. The standard score test checks the significance of $n^{-1/2}T'\hat{u}$; consider the corresponding criterion derived from $\hat{\beta}_1$, i.e.

$$n^{-1/2}T'\hat{u} = n^{-1/2}T'u - \left[n^{-1}T'X_1\right][n^{1/2}(\hat{\beta}_1 - \beta_1)].$$

Eq. (3.19) implies that, unless $n^{-1}T'X_1$ tends to a matrix with every element equal to zero, the asymptotic distribution of $n^{-1/2}T'\hat{u}$ will depend upon that of $n^{1/2}(\hat{\beta}_1 - \beta_1)$. In order to remove the dependence on estimators of nuisance parameters, Wooldridge (1990) proposes that test variables be constructed as residuals from auxiliary regressions. In this simple example, rather than using $T$ with $\hat{\beta}_1$, Wooldridge’s approach would lead to the use of $\hat{T}$ which is the matrix of OLS residuals from regressing $T$ on $X_1$. With this choice of test variables, the scaled vector of covariances with residuals is

$$n^{-1/2}\hat{T}'\hat{u} = n^{-1/2}\hat{T}'u,$$

(3.20)
since $\hat{T}'X_1 = 0$. This result implies an exact equivalence with the score test of $\tau = 0$ in (3.14). In more general nonlinear models, there is an asymptotic equivalence between score and Wooldridge-type tests in many, but not all, cases; see Wooldridge (1990, p. 26). Non-linear specifications may lead to $\hat{\theta}_1$-estimators having high computational cost relative to some other root-$n$ consistent estimator $\hat{\theta}_1$ and so tests based upon the latter procedure may be more convenient.

The second distinguishing feature of Wooldridge’s approach is the calculation of a robust test of the significance of the criterion of (3.20). In the standard score test, a test is derived under the assumptions that the errors are uncorrelated and homoskedastic. Wooldridge relaxes a higher-order moment condition by allowing for heteroskedasticity of unspecified form. More precisely, the Wooldridge-type statistic can be calculated as $n$ times the uncentred $R^2$ from the OLS regression of $i$, a vector with every element equal to one, on $\hat{S}\hat{T}$, where $\hat{S}$ is a diagonal matrix with $\hat{s}_{ii} = \hat{u}_i$, $i = 1, \ldots, n$. This artificial regression leads to the test being based upon a heteroskedasticity consistent covariance matrix of the type discussed by Eicker (1963) and White (1980). The corresponding regression for nonlinear models is discussed by Davidson and MacKinnon (1993, Section 11.6), who term it the heteroskedasticity robust Gauss–Newton regression.

Wooldridge (1990) provides a very general discussion of his approach and several examples to illustrate its application. Given the existing results on the finite sample behaviour of tests of linear regression models based upon heteroskedasticity consistent covariance matrices, it would be very useful to have Monte Carlo evidence to assist in the evaluation of Wooldridge-type tests.
4. Finite sample problems

The justification for all the test procedures discussed thus far has been first-order, O(1), theory which provides a relatively simple approximation to the true sampling distribution of the test statistic under consideration, usually \(\chi^2\). In practice, however, this approximation can be extremely poor, bestowing on a test procedure an actual significance level that is far from the nominal value. This section reports on three areas of research which have addressed the potential failings of first-order theory.

First, standard O(1) asymptotic theory does not distinguish between asymptotically equivalent test statistics generated, for example, by alternative consistent variance matrix estimators. It is well known that asymptotically equivalent statistics need not be numerically similar, when confronted with exactly the same data. There is now a large and growing literature on the effect of different variance estimators on the sampling distribution of score test statistics; see, for example, Davidson and MacKinnon (1983, 1984b, 1992), Bera and MacKenzie (1986), Orme (1990) and Chesher and Spady (1991). This is discussed in Section 4.1. Second, even if a test statistic is judged “best” behaved, the reliability of the approximate O(1) sampling distribution can remain very poor, even for quite large sample sizes. Refined approximations applicable in the classical likelihood framework are reviewed in Section 4.2. Finally, Section 4.3 outlines the bootstrap as an alternative to the derivation and application of analytical finite sample corrections. The nonparametric bootstrap can be applied outside the classical likelihood framework and recent evidence suggests that this approach works well.

4.1. Score tests and variance estimators

Within the classical likelihood framework, and the notation of Section 2, consider the score test of \(b_2 = 0\) conditional upon \(b_3 = 0\). The generic form of the score test statistic is

\[
T_2 = n^{-1}d_2(\hat{\theta})[\hat{J}_{22} - \hat{J}_{21}(\hat{J}_{11})^{-1}\hat{J}_{12}]^{-1}d_2(\hat{\theta}),
\]

(4.1)

where \(\hat{J}_{ij}\) denotes any consistent estimator of \(\mathcal{J}_{ij}\) under the null hypothesis, \(i,j = 1,2\). Under the null hypothesis, the sampling distribution of \(T_2\) is, to O(1), approximately \(\chi^2(p_2)\) whatever the choice of \(\hat{J}_{ij}\). However, different choices for \(\hat{J}_{ij}\) lead to numerically different values of \(T_2\) with markedly different, and in some cases quite unappealing, sampling behaviour; e.g., certain choices of \(\hat{J}_{ij}\) can lead to negative values of \(T_2\) in finite samples, although the probability of this diminishes as \(n \to \infty\).

Under fairly general conditions it is possible to express the (likelihood) score functions as \(d_i(\theta) = \sum_{i=1}^{n} s_i(\theta), i = 1,2,3\), in which case the linearised expression for \(n^{-1/2}d_2(\hat{\theta})\) given in (3.1), under the null hypothesis, suggests a class of variance estimators guaranteed to avoid negative values of \(T_2\). These variance estimators are obtained by considering different forms of the following \((n \times p_2)\) matrix:

\[
\hat{W} = \hat{S}_2 - \hat{S}_1(\hat{J}_{11})^{-1}\hat{J}_{12},
\]

(4.2)
where $\hat{S}_i$ is an $(n \times p_i)$ matrix with rows $\hat{s}_{it} = s_{it}(\hat{\theta})'$, $i = 1, 2$. Again under fairly general conditions, $p \lim n^{-1} \hat{S}_i' \hat{S}_i = \mathcal{F}_{ij}$ so that $n^{-1} \hat{W}' \hat{W}$ is a consistent variance estimator, guaranteed to be positive definite (unless there exist linear dependencies between the columns of $\hat{S}_2$ and $\hat{S}_1$). Since $\hat{S}_1' = 0$, so that $\hat{W}' = d_3(\hat{\theta})$, the resultant score statistic, $T_2$, emerges as ‘the sample size minus the residual sum of squares’ $(n - \text{RSS})$ from running the following artificial regression

$$i = \hat{W}b + \text{residual}. \tag{4.3}$$

The information matrix equality affords many possibilities for $\hat{W}$ leading to differing values of $T_2$. Employing $(\hat{\mathcal{F}}_{11})^{-1} \hat{\mathcal{F}}_{12} = \hat{\mathcal{D}}_{11}^{-1} \hat{\mathcal{D}}_{12}$ gives a robust test statistic, in the spirit of White (1982), Kent (1982) and as discussed in Section 3.2. Another variant, often referred to as the outer product of the gradient (OPG) test statistic, is obtained when $(\hat{\mathcal{F}}_{11})^{-1} \hat{\mathcal{F}}_{12} = (\hat{S}_1' \hat{S}_1)^{-1} \hat{S}_1' \hat{S}_2$. In this case the appropriate statistic can then simply be calculated as $n - \text{RSS}$ from

$$i = \hat{S}_1b_1 + \hat{S}_2b_2 + \text{residual}. \tag{4.4}$$

Denoted by $T_2^0$, this formulation has been advocated a number of times in the literature; see Godfrey and Wickens (1982), Chesher (1983b, 1984), Lancaster (1984), Newey (1985). (One variant of $T_2^0$ is the $F$-test of $b_2 = 0$ in (4.4), another is discussed by Tauchen (1985).) All test statistics derived from (4.3) can also be obtained from restricted estimation of (4.4). Thus the OPG variant will be numerically the largest in the class of statistics generated by (4.3). (It can also be shown that Tauchen’s (1985) variant of the OPG test statistic is larger still; see Chesher, 1983a.)

Due to its simplicity, the OPG variant has proved to be popular in applied work; see, for example, Blundell et al. (1987), Machin and Stewart (1990). Furthermore, if (4.4) is augmented to become $i = \hat{S}_1b_1 + \hat{S}_2b_2 + \hat{S}_3b_3 + \text{residual}$, where $\hat{S}_3$ has rows $s_{3t}(\hat{\theta})'$, then an asymptotically valid Bera and Yoon (1993) score test can be obtained as a test of $b_2 = 0$; see Section 3.1. Unfortunately many researchers have reported that O(1) theory provides an abysmal approximation to the true finite sampling behaviour of OPG test statistics, even in what are considered to be quite large samples. Chesher and Spady (1991) show that the OPG variance estimator tends to be downward biased and highly variable, leading to a test statistic which can be disastrously over-sized and Davidzon and MacKinnon (1983, 1984b), Bera and MacKenzie (1986), Taylor (1987) and Kennan and Neumann (1988) provide early Monte Carlo evidence on this lack of reliability.

Since the OPG variant is numerically the largest variant generated by (4.3) or (4.4), another choice of test statistic in this class is guaranteed, at the very least, to reduce the extent of over-rejection. In sampling experiments, Orme (1990) found that the statistic $T_2^*$, which uses $\hat{\mathcal{F}}_{ij} = n^{-1} \sum_{t=1}^n \hat{E}[s_{it}(\hat{\theta})' s_{jt}(\hat{\theta})']$ in the construction of $\hat{W}$, often exhibited the least size-distortion. Here, $\hat{E}[.]$ denotes expectation taken with respect to the null model (conditionally upon any regressors) but with unknown parameters replaced by their MLE. Unfortunately, $T_2^*$ can still behave poorly relative to O(1) asymptotic theory; see Orme (1990) and Horowitz (1994, 1997).
In the search for a test statistic with acceptable finite sample size properties, many alternative artificial regressions have been proposed; Davidson and MacKinnon (1990) discuss the general properties required of artificial regressions and MacKinnon (1992) provides a useful survey of applications. Perhaps two worthy of mention are the double length regression (DLR) approaches of Davidson and MacKinnon (1983, 1984a, 1988, 1992) and Orme (1995a), both of which have been found to perform well, relative to the OPG regression (4.4). The DLR proposed by Orme is specifically designed to yield the efficient variant of the score test (whereas the DLR of Davidson and MacKinnon is not) and is applicable for many commonly used microeconometric models (unlike the DLR of Davidson and MacKinnon, which is restricted to nonlinear models with normal disturbances). This efficient variant, denoted $T^E_2$, is given by (4.1) but with $\mathcal{J}_{ij}$ replaced by $n^{-1} \sum_{t=1}^{n} \hat{E}[s_t(\hat{\theta}) s_t'(\hat{\theta})']$ throughout $\mathcal{J}_{22} - \mathcal{J}_{21}(\mathcal{J}_{11})^{-1} \mathcal{J}_{12}$. In certain cases, the artificial regression proposed by Orme could be triple or even quadruple length, but the required data manipulations become overly involved at this stage.

The available evidence (e.g., Davidson and MacKinnon, 1984b; Orme, 1990; Chesher and Spady, 1991) suggests that $T^E_2$ is more likely to have sampling behaviour that is in closest agreement with $O(1)$ theory. For example, Orme (1990) considers various score tests for heteroskedasticity in the context of a binary probit model where the asymptotically efficient version is no harder to obtain than the OPG; see Davidson and MacKinnon (1984b). Using a nominal size of 5%, simulations of $T^E_2$ from a correctly specified (homoskedastic) probit model with three regressors yielded empirical sizes in the range of 6.0–7.4% for sample sizes 50–1000; the same experiment yielded empirical sizes in the range 1.2–11.0% for $T^*_2$ and 29.6–87.6% for $T^0_2$!

The regression-based method of Orme (1995a) does not preclude the, often tedious, calculations necessary in forming the efficient variance estimator. However, simulation methods can often be used to approximate this estimator, without recourse to analytical derivations. Suppose the model is $f(y|x_t; \theta), \theta \in \Theta, t = 1, \ldots, n$, and $\hat{\theta}$ is the null MLE. For each $t$, taking the value of $x_t$ and $\hat{\theta}$ as fixed, $R$ independent pseudo-random draws can be made from $f(y|x_t; \hat{\theta})$ yielding $y^{(r)}_t, r = 1, \ldots, R$. The triplet $(y^{(r)}_t, x_t, \hat{\theta})$ is used to construct $s^{(r)}_t(\hat{\theta}), i = 1, 2$, and the relevant expectations are then simulated as $R^{-1} \sum_{r=1}^{R} s^{(r)}_t(\hat{\theta}) s^{(r)}_t(\hat{\theta})' = \hat{V}_{ij}$, say. In (4.1), $\mathcal{J}_{ij}$ is replaced by $n^{-1} \sum_{t=1}^{n} \hat{V}_{ij}$, yielding a simulated efficient score test denoted $T^S_2$. Under fairly general conditions, once estimation had been performed and $\hat{\theta}$ obtained, $\lim_{R \to \infty} (T^E_2 - T^S_2) = 0$ as $R \to \infty$, for fixed $n$; good approximations can be obtained using smaller $R$ as $n$ increases. Note that the actual components that are simulated are simply those used in the calculation of $T^0_2$.

For further details, together with Monte Carlo evidence, see Orme (1995b).

Although the distribution of $T^E_2$ is generally found to be more closely approximated by its $O(1) \chi^2$ distribution, large discrepancies can still arise. Building on the work of Harris (1985, 1987), Chesher and Spady (1991) find that, to $O(n^{-1})$, the mean of the efficient score statistic is lower than predicted by its reference $O(1) \chi^2$ distribution, its variance is higher and its sampling distribution is positively skewed relative to $\chi^2$. The results of Davidson and MacKinnon (1984b) and Orme (1990) support these findings.
Thus users of the test will frequently observe unusually low values, but occasionally overly large values (even under the null).

The discussion thus far has focused specifically on asymptotically equivalent variants of a score test statistic using MLE. Clearly, though, the general point about differing variance estimators leading to differing finite sample properties of the resultant test statistic remains valid in the framework of extremum estimation as well. In particular, the possibility of generating a negative value of a statistic is not precluded. However, provided the ‘scores’ can be expressed in the form

\[ g_i(\theta) = \sum_{t=1}^{n} s_{it}(\theta), \quad i = 1, 2, 3, \]

a straightforward extension of the MLE formulae yields a class of variance estimators which guarantees non-negative test statistics. As before, an inspection of the linearised expression for \( n^{-1/2} g_2(\hat{\theta}) \) given in (3.9), under the null, implies that the appropriate \( T_2 \) score-type test statistic can be obtained as \( n – RSS \) from (4.3) but where the ‘regressor’ matrix is

\[ \tilde{W} = \hat{S}_2 - \hat{S}_1(\tilde{A}_{11})^{-1}\tilde{A}_{12} \]

in which \( p \lim \tilde{A}_{ij} = A_{ij}^0 \). Differing values of \( T_2 \) are generated by the choice of \( (\tilde{A}_{11})^{-1}\tilde{A}_{12} \) in the above, although the fact that the information matrix equality fails to hold in this case limits the number of possible variants. As in the MLE case, the use of \( \tilde{A}_{ij} = n^{-1}\partial^2 Q(\hat{\theta})/\partial \theta_i \partial \theta_j \) in the construction of \( \tilde{W} \) can imply certain robustness properties for resulting test statistic. For example, using this variant for testing (3.13) against (3.14) is essentially the heteroskedasticity robust statistic described at the end of Section 3.3. Interestingly, and where expectations are defined, the \( T_2^* \) variant for the extremum estimation case, which employs \( \hat{A}_{ij} = n^{-1}\hat{E}[\partial^2 Q(\hat{\theta})/\partial \theta_i \partial \theta_j] \) in the construction of \( \tilde{W} \), can also exhibit higher order moment robustness. In particular, Wooldridge’s (1990) higher order moment robust test for heteroskedasticity could be interpreted as a \( T_2^* \)-type test statistic as could Godfrey and Orme’s (1991) non-normality robust test for symmetry of regression disturbances. As pointed out in Section 3.3, it would be useful to have some Monte Carlo results on the behaviour of the robust (Wooldridge-type) variants since the finite sample properties of the corresponding statistics in the MLE case has not been uniformly encouraging.

4.2. Finite sample corrections to the efficient score test

Edgeworth expansions have been used to obtain approximations to the distributions of various test statistics. The validity of such formal expansions is discussed, for example, by Bhattacharya and Ghosh (1978, 1989); also see Hall (1992, Chapter 2). Using Edgeworth-type expansions, Harris (1985, 1987) develops the \( O(n^{-1}) \) approximation to the null sampling distribution of the efficient score test statistic, in a likelihood framework for the case of independently but not necessarily identically distributed data. (Taniguchi, 1991 derives expressions for special cases involving dependent observations.) Recent research has investigated the application and value of such higher-order expansions for the efficient score test; see, for example, Chesher and Spady (1991), Cordeiro and Ferrari (1991), Cordeiro et al. (1993), Cribari-Neto and Ferrari

(1995a, b), Peters (1996) and Cribari–Neto (1997). Smith et al. (1997) provide the corresponding expansion for the OPG form of the score test. Although the appeal of such refinements has yet to be seen in applied work several important theoretical findings have emerged. Apart from indicating when O(1) theory is likely to be a poor guide, the O(n^{-1}) approximation contains information concerning the effect of regression design on the sampling behaviour of the test statistic; a matter on which standard asymptotic analysis has nothing to say.

In general, let the efficient score test statistic under consideration be denoted $S$ and let $q$ be the number of restrictions under test. The O(1) approximation provides quantiles $u_j$ satisfying $\Pr(\chi^2(q) \leq u_j) = \lambda$. Harris (1985, 1987) derives the following approximation to the sampling distribution of $S$, which is correct to O(n^{-1}):

$$\Pr(S \leq u) = P_q + \frac{1}{24n} \sum_{s=0}^{3} A_{2s} P_{q+2s} + o(n^{-1}),$$

(4.5)

and a Cornish–Fisher-type inverse expansion yields quantiles correct to O(n^{-1}):

$$u^c_j = u_j + \frac{1}{12n} \left( \frac{A_6 u_j \{u_j^2 + (q+4)u_j + (q+2)(q+4)\}}{q(q+2)(q+4)} + \frac{A_4 u_j (u_j + q+2)}{q(q+2)} + \frac{A_2 u_j}{q} \right),$$

(4.6)

where $P_q = \Pr(\chi^2_q \leq u)$, $A_0 = x_2 - x_1 - x_3$, $A_2 = 3x_3 - 2x_2 + x_1$, $A_4 = x_2 - 3x_3$, $A_6 = x_3$ and the $x$’s, which define the O(n^{-1}) corrections, are determined by null cumulants, up to order four, of the first and second derivatives of the alternative model’s log-likelihood function. As noted by a referee, however, there is a possibility that (4.6) can lead to a negative critical value.

Chesher and Spady (1991) give the required coefficients in tensor notation and consider tests for heteroskedasticity and non-normality in the normal linear model. Such statistics are pivotal (their exact sampling distributions do not depend upon unknown parameters; see Breusch, 1980) so that the obtained $x$’s are parameter free. Moreover, the corrections for the non-normality (skewness and kurtosis) test statistics are also invariant to the regression design and thus applicable in complete generality. The corrections for the heteroskedasticity test, however, are sensitive to regression design; see also Honda (1988) and Cribari-Neto and Ferrari (1995a). For the cases considered, Chesher and Spady find that for sample sizes in the range 100–500, (4.5) is often substantially better than the O(1) $\chi^2$ approximation.

The O(n^{-1}) corrections have also been re-produced in matrix notation: see Cordeiro et al. (1993), for application in generalised linear models; Cribari-Neto and Ferrari (1995a) and Cribari-Neto and Ferrari (1995b), when testing for, respectively, heteroskedasticity and linear restrictions, in the normal linear model. Ferrari and Cordeiro (1994) provide generally applicable matrix expressions, as do Orme and Peters (2000) in a different manner.

Cordeiro et al. (1993) show that, for score tests of the dispersion parameter in generalised linear models, the O(n^{-1}) corrections depend upon the regression design...
only through the number, \( k \), of covariates present in the null model: \( x_1 \) is quadratic in \( k \), \( x_2 \) is linear in \( k \) and \( x_3 \) is a constant. Orme and Peters (2000) obtain exactly the same property when considering score tests of distributional failure in location-scale models. Although, in general, different both results cover the situation of testing the exponential regression model against a Gamma alternative and it is encouraging to observe that the general formulae obtained by the two sets of authors yield the same finite sample corrections.

Rather than correct the critical values, an alternative strategy is to correct the test statistic multiplicatively by a term which is \( 1 - O_p(n^{-1}) \). The form of such a corrected score test statistic is

\[
S^c = S \left\{ 1 - \frac{1}{12n} \sum_{j=1}^{3} \gamma_j S^j \right\}, \tag{4.7}
\]

where \( \gamma_1 = (x_1 - x_2 + x_3)/q \), \( \gamma_2 = (x_2 - 2x_3)/(q(q + 2)) \), \( \gamma_3 = x_3/(q(q + 2)(q + 4)) \); see Cordeiro and Ferrari (1991). It is interesting to note that corrected critical values (4.6), for use with \( S \), can be expressed as \( u^c = u_S(1 + (12n)^{-1} \sum_{j=1}^{3} \gamma_j u^j_S \left\{ 1 - \frac{1}{12n} \sum_{j=1}^{3} \gamma_j S^j \right\} \). On the other hand, critical values for \( S^c \) are obtained from the usual \( \chi^2(q) \) distribution since, to \( O(n^{-1}) \), \( S^c \) is constructed to be \( \chi^2(q) \) under the null. \( S^c \) has become known as a Bartlett-corrected score statistic; see Cribari-Neto (1996) and links therein. The correction, however, is stochastic unlike the familiar Bartlett correction to the likelihood-ratio test statistic. Clearly, the correction factor in (4.7) need not be of the form \( (1 - B/n) \), where \( B \) is quadratic in \( S \); other correction factors, which are equivalent to \( O(n^{-1}) \), include \( (1 + B/n)^{-1} \) and \( \exp(-B/n) \), with the latter guaranteeing a positive outcome for \( S^c \).

Comparing the three forms of Bartlett-corrected statistic, Cribari-Neto and Cordeiro (1996) report that the second form, using \( (1 + B/n)^{-1} \), exhibited least-size distortion when conducting tests of dispersion parameters in the normal, inverse Gaussian and gamma models; see, also Rao and Mukerjee (1995) for a comparison of corrections. Cribari-Neto and Cordeiro also note that, in some cases, size adjustment comes at the expense of some loss of power; Horowitz (1994) has also emphasised this point.

Bartlett-corrected score tests have been developed and studied extensively by Cordeiro, Cribari-Neto, Ferrari and Paula, as detailed in the references. Cribari-Neto and Cordeiro (1996) give a useful survey. Mukerjee (1993) also provides several useful results and, in this issue, Ghosh and Mukerjee (2001) extend the analysis for the usual likelihood by considering Bartlett adjustments for tests derived from a quasi likelihood.

One point which is not emphasised in some studies is the applicability of the underlying Edgeworth expansion. In a given context, there is a minimum sample size \( (n_{\min}) \) below which the approximating distribution function (4.5) is not necessarily positive and monotonically increasing over its domain. For sample sizes below \( n_{\min} \) using (4.5) may well lead to over-corrections. Chesher and Spady (1991) do provide such minimum sample sizes for the cases they consider as does Peters (1996), who analyses finite sample corrections to the efficient score test in Probit and Tobit models. The slightly discouraging result to emerge from their work is that the requisite minimum sample size can be quite large.
4.3. Bootstrap methods

The bootstrap is an alternative way of improving on the quality of $O(1)$ theory. It can be used to obtain improved approximations to the sampling distribution of any variant of the score test statistic discussed in Section 4.1 and is also applicable outside the classical likelihood framework. A full review of relevant bootstrap methods is beyond the scope of this paper and, in this section, we simply mention some useful surveys and recent developments in econometric applications. The basic idea of bootstrapping is due to Efron (1979). Surveys of applications in econometrics can be found in Jeong and Maddala (1993) and Vinod (1993) whilst Young (1994) provides a critical review of recent research. Originally, most applications of the bootstrap focused on the sampling distribution of estimators and confidence intervals; see, for example, Efron and Tibshirani (1993). More recently, ‘bootstrapping’ test statistics has received growing attention in the econometric literature; see Davidson and MacKinnon (1996, 1999) and Horowitz (1994). Such approaches are simplest to explain and apply when dealing with independent observations; see Horowitz (1994, 1997). Lahiri (1992), Li and Maddala (1996) and Hall and Horowitz (1996) consider the more demanding case of dependent data.

The practical value of the bootstrap lies in the fact that, when, as is very often the case, score-type tests are asymptotically pivotal, bootstrap critical values possess at least the same level of accuracy as those obtained using Edgeworth expansion methodology. For further discussion and details see Beran (1988), Hall (1992), Horowitz (1997) and references therein. In a few exceptional cases, the statistics may be exactly pivotal; e.g., the test statistics considered by Chesher and Spady (1991) and the tests for distributional failure in location scale models examined by Orme and Peters (2000), both in the likelihood framework. In such cases bootstrap samples can be generated using any pair $(F, \theta^*)$, where $F$ is the cdf specified by the null hypothesis and $\theta^*$ is an admissible value of $\theta$. Whether or not the score-type statistic is exactly pivotal, the parametric bootstrap can be used when a parametric distribution is specified under the null hypothesis. Horowitz (1994) reports the parametric bootstrap achieves remarkable reductions in size distortion for OPG forms of test statistics in Probit and Tobit Models, even with as few as 100 bootstrap samples. For example, using a nominal level of 5% the use of $O(1)$ critical values yields empirical sizes in the range 30–74% for an OPG test statistic; using the bootstrapped critical values the empirical sizes were reduced to 3.9–5.9%.

If a fully parametric form for the model under the null is not specified, then the so-called non parametric bootstrap can be adopted. If the null model is a linear regression with iid$(0, \sigma^2)$ errors, the common error distribution being unspecified, the nonparametric bootstrap can be implemented by various forms of residual resampling; see Davidson and MacKinnon (1999) for some Monte Carlo evidence on the small sample impact of the form of residual resampling. The non parametric bootstrap can also accommodate heteroskedasticity of unknown form in the errors of regression models. Thus, test statistics based on robust covariance matrix estimators may be employed but,
as pointed out by Horowitz (1997), the bootstrap cannot be implemented by simply resampling OLS residuals independently of $X_1$ (if one wishes to allow for the heteroskedasticity of unknown form to depend upon $X_1$). One could, however, randomly resample from the $n$ rows of the data matrix $(y,X_1)$; this can be further improved upon by employing the ‘wild’ (weighted) bootstrap, developed by Wu (1986) and Liu (1988). See, also, Mammen (1993), Jeong and Maddala (1993) and Horowitz (1997), who finds that using asymptotic critical values and heteroskedasticity robust variance estimation produces large discrepancies between actual and nominal levels whilst the wild bootstrap “essentially removes [these] distortions of level”; see also Hu and Zidek (1995).

The above illustrations have all used the simplest case of resampling independent observations. Jeong and Maddala (1993, Section 4.2) give some discussion of resampling in regression models with serially correlated disturbances. In general, methods for resampling dependent data must take into account such dependency. The bootstrap is implemented by dividing the sample into blocks and sampling the blocks independently with replacement. More details are given in Lahiri (1992), Li and Maddala (1996) and Hall and Horowitz (1996).

5. Conclusions

A common feature of much applied work in econometrics is the use of a number of separate score tests to check for different types of departure from the null model under scrutiny. There is evidence that such separate tests are not robust to the presence of misspecifications that they were not designed to detect: true null hypotheses may be rejected too frequently, false null hypotheses may be accepted frequently. Some “robustification” can be achieved by modifying the score test as described in Section 3. Initial Monte Carlo evidence indicates that a progression from the original score to a modified score may be worthwhile in empirical work, but there is scope for further investigations.

Whether or not the researcher wishes to use a modified score test, it is clearly important to make use of procedures that improve reliability. Bartlett-type corrections offer a theory-based approach, but they have not become popular. A computer-intensive alternative to theory-based corrections is provided by the bootstrap in its various forms. Again what evidence we have suggests that the bootstrap can lead to much improved agreement between finite sample significance levels and the nominal size.

An interesting area for future research is the small sample behaviour of modified/robust score tests with bootstrap critical values.

References


