The Information Matrix Test with Bootstrap-Based Covariance Matrix Estimation

Geert Dhaene* K.U.Leuven
Dirk Hoorelbeke K.U.Leuven

August 8, 2002

Abstract

We propose an information matrix test in which the covariance matrix of the vector of indicators is estimated using the parametric bootstrap. Monte Carlo results and heuristic arguments show that its small sample performance is comparable with that of the efficient score form.

JEL classification: C12, C15
Key-words: information matrix test, parametric bootstrap

*Corresponding author. Address: K.U.Leuven, Department of Economics, Naamsestraat 69, 3000 Leuven, Belgium. Telephone: +32-(0)16-32 67 98. Fax: +32-(0)16-32 67 96. E-mail: geert.dhaene@econ.kuleuven.ac.be

1 Introduction

While the information matrix (IM) test introduced by White (1982) is well known as a general test for misspecification of a parametric likelihood function, its use in applied econometric research is still limited. A major drawback of the IM test is that the asymptotic χ^2 distribution is a very poor approximation to the finite sample distribution of the test statistic. This seriously limits its usefulness in practice. Large deviations from the asymptotic distribution are typical even in relatively large samples, as evidenced by the Monte Carlo experiments reported in Taylor (1987), Orme (1990), Chesher and Spady (1991), Davidson and MacKinnon (1992, 1998), and Horowitz (1994). Several approaches have been proposed to overcome this problem. Chesher and Spady (1991) derive, for specific models, critical values for the IM test statistic that are based on a higher order Edgeworth expansion. Davidson and MacKinnon (1992) propose a variant of the IM test based on double-length artificial regressions. Their method, however, cannot be applied to models for discrete, censored, or truncated data. Horowitz (1994) proposes bootstrap-based critical values for the IM test. Despite these efforts, computing the correct critical value of an IM test statistic for an arbitrary model is still not particularly easy.

Several versions of the IM test have been proposed that differ only in the way the covariance matrix of the vector of indicators is estimated. In Section 2, we propose a new form of the IM test, making use of the parametric bootstrap to estimate the covariance matrix. Although this approach is based on simulations, the computational demands are very modest and there are no analytical requirements at all. The proposed method eliminates the approximation errors that result in other IM tests from the use of asymptotic covariance matrix formulae and from approximating expectations by sample averages. When a fixed number of bootstrap simulations is used, the proposed test statistic has an asymptotic T^2 distribution.

In Section 3, we report Monte Carlo results for the linear model and the probit model, showing that the new test performs very similarly to the efficient score form of the IM test. Hence, it is a promising alternative to existing IM tests in models where the efficient score form is not available. Section 4 concludes.

2 Estimating the covariance matrix of the IM test

Consider a parametric model with log-density $F(y;\theta)$, where θ is a $p \times 1$ vector of parameters. Let $F_i = [\partial F/\partial\theta_i]_{\theta=\theta_0}$ and $F_{ij} = [\partial^2 F/\partial\theta_i\partial\theta_j]_{\theta=\theta_0}$, where θ_0 maximizes $E[F(y;\theta)]$ with respect to θ and $E[\cdot]$ denotes expectation. The null hypothesis underlying the class of IM tests is

$$H_0 : E\left[F_i F_j + F_{ij}\right] = 0 \qquad (1 \leq i, j \leq p). \tag{1}$$

Given a sample of observations $y_1, \ldots, y_n$, define the indicators

$$\hat{D}_{ij} = n^{-1/2} \sum_{t=1}^{n} (\hat{F}_i \hat{F}_j + \hat{F}_{ij}), \tag{2}$$

where a hat indicates evaluation at y_t and $\hat{\theta}$, the MLE of θ_0. Most existing IM tests are based on an asymptotically χ^2_q distributed statistic of the form

$$\omega = \hat{D}'\hat{V}^{-1}\hat{D}, \tag{3}$$

where $\hat{D}$ is a $q \times 1$ vector of appropriately selected indicators $\hat{D}_{ij}$ and $\hat{V}$ is a consistent estimate of its covariance matrix under H_0 (alternatively, under the stronger assumption that $F(y;\theta)$ is the correct log-density).

Orme (1990) reviews many alternative choices of $\hat{V}$, including those leading up to White's (1982) form, the Chesher (1983) and Lancaster (1984) form, and the efficient score form of the IM test [1]. All these choices of $\hat{V}$ are based on

[1] White's (1982) estimator $\hat{V}$ is consistent under H_0; the other estimators $\hat{V}$ mentioned here are consistent under somewhat stronger assumptions.

equivalent analytical formulae for the asymptotic covariance matrix of $\hat{D}$, the differences arising essentially from replacing expectations with sample averages in different parts of those formulae. Available Monte Carlo evidence, in settings where $F(y;\theta)$ *is* the correct log-density, shows that the ensuing IM test statistics have finite sample distributions that are poorly approximated by the χ^2_q distribution. Four sources of possible error may be involved in the approximation:

(i) the finite sample distribution of $\hat{D}$ may be non-normal;

(ii) the finite sample covariance matrix of $\hat{D}$, say V_n, may differ from its asymptotic covariance matrix, V_∞;

(iii) $\hat{\theta}$ is used in place of θ_0 in formulae for V_∞;

(iv) sample averages replace expectations in parts of formulae for V_∞.

In most circumstances, the error sources (i)-(iii) effectively apply to the IM tests discussed so far. Moreover, the efficient score form is the only one not vulnerable to (iv).

Rather than relying on an asymptotic covariance matrix formula, one may choose $\hat{V}$ to estimate the *exact* finite sample covariance matrix of $\hat{D}$, denoted $V_n(\theta_0)$, since it typically depends on θ_0. Although it is simple enough to write $V_n(\theta_0)$ as an integral, working out the integral analytically is bound to be impossible in all but the simplest models. A simple and feasible alternative is to estimate $V_n(\theta_0)$ by the parametric bootstrap, which involves the following steps:

0. compute the MLE $\hat{\theta}$;

1. for $b = 1, \ldots, B$:

   - generate an i.i.d. sample $y_{1b}, \ldots, y_{nb}$ from the density $\exp F(\cdot;\hat{\theta})$;
   - for this sample, compute the MLE $\hat{\theta}_b$ and the vector of selected indicators $\hat{D}_b$;

2. compute

$$\hat{V}_B = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{D}_b - \bar{D})(\hat{D}_b - \bar{D})', \tag{4}$$

where $\bar{D} = B^{-1}\sum_{b=1}^{B}\hat{D}_b$.

It is obvious that $E[\hat{V}_B \mid \hat{\theta}] = V_n(\hat{\theta})$ and that for fixed n, $\hat{V}_B \overset{\text{a.s.}}{\to} V_n(\hat{\theta})$ as $B \to \infty$. Thus, through the choice of the number of bootstrap replications B, $\hat{V}_B$ approximates $V_n(\hat{\theta})$ to any desired accuracy. Taking $\hat{V} = \hat{V}_B$ in (3) yields the IM test statistic

$$\omega_B = \hat{D}'\hat{V}_B^{-1}\hat{D}. \tag{5}$$

Under the assumption that $F(y;\theta)$ is the correct log-density, ω_B has the following limit behavior. As $n \to \infty$ and $B \to \infty$, $\omega_B \overset{\text{d}}{\to} \chi^2_q$, where $\overset{\text{d}}{\to}$ denotes convergence in distribution. For fixed $B \geq q+1$ and $n \to \infty$,

$$\omega_B \overset{\text{d}}{\to} T^2_{q,B-1} \tag{6}$$

(Hotelling's T^2), since $\hat{D} \overset{\text{d}}{\to} N(0, V_\infty(\theta_0))$, $(B-1)\hat{V}_B \overset{\text{d}}{\to} W(V_\infty(\theta_0), B-1)$ (central Wishart), and $\hat{D}$ and $\hat{V}_B$ are asymptotically independent. Note that (6) may also be stated as $\frac{B-q}{(B-1)q}\omega_B \overset{\text{d}}{\to} F_{q,B-q}$. Using the IM test statistic ω_B and critical values from the $T^2_{q,B-1}$ distribution, (ii) is eliminated as a source of approximation error. This IM test is closest in spirit to the efficient score form of the IM test as it replaces V_∞ with V_n in the latter. With finite B, V_n is estimated with some noise, but the T^2 critical values correct for this. Since the test based on ω_B has less sources of error, we expect it in general to exhibit smaller errors in rejection probability (ERP) [2] than the IM tests based on the χ^2_q approximation to ω already discussed. We note, however, that exceptions to

[2] The ERP of a test is the actual minus the nominal (i.e. chosen) probability of rejecting the null hypothesis when it is true. The ERP often depends on the parameter that indexes the distributions constituting the null hypothesis.

this rule will almost certainly exist, because different errors may to some extent counterbalance each other.

As Horowitz (1994) has shown, IM tests with smaller ERP do not necessarily have better power properties, even when bootstrap-corrected critical values are used. As an intuitive example of how this may occur, consider adding noise to a severely biased estimate $\hat{V}$ (i.e. biased for $V_n(\theta_0)$): this is likely to reduce the ERP of the corresponding test, while it obviously also reduces the power when appropriate critical values are used. As for $\hat{V}_B$, the only source of bias is the fact that $\hat{\theta} \neq \theta_0$, and the only noise stems from taking B finite. Thus, for large enough B, we expect the power of the IM test based on ω_B to be no less than the power of the efficient score test, which, given the available evidence, appears to be the most powerful of existing IM tests (Davidson and MacKinnon, 1998). The efficient score test, however, requires calculating certain expectations analytically and hence is only available for models where this has proven feasible. For more complicated models, the test proposed here offers a feasible, and presumably powerful, alternative.

We have two final remarks. First, the only computational requirement to obtain ω_B is that observations can be generated from the density $\exp F$ and that the vector of indicators can be computed. The latter can often be extracted without effort from econometric software packages, either as the difference between two information matrix estimates, or as the difference between the inverses of two estimates of the covariance matrix of the MLE. Thus, no analytical work is required before the test can be applied. Second, although Monte Carlo results show that the ERP of the newly proposed test is moderate, it may be advisable in situations with few observations to use bootstrap-based critical values, as suggested by Horowitz (1994) in the context of the IM test. Although this requires a nested bootstrap – the inner bootstrap serves to calculate $\hat{V}_B$ – this is nowadays quite feasible: 50 inner and 99 outer bootstrap replications will often suffice.

3 ERP and power: Monte Carlo evidence

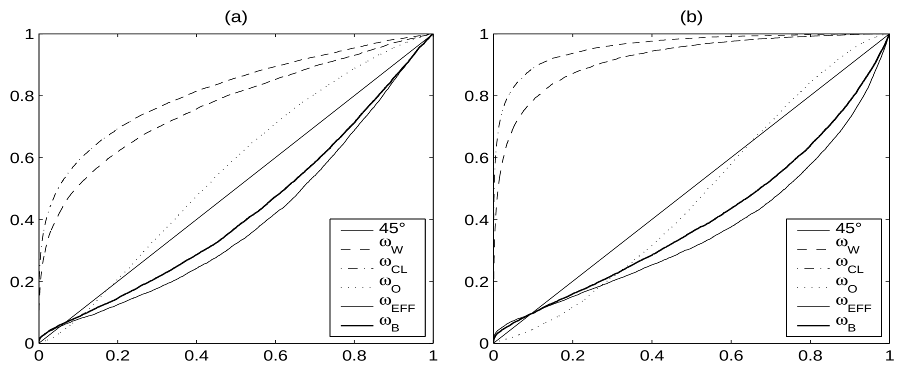
Here we report comparative Monte Carlo results on the finite sample properties of the statistic ω_B, White's (1982) IM test statistic ω_W, Chesher (1983) and Lancaster's (1984) IM test statistic ω_{CL}, Orme's (1990) ω_3, here ω_O, and the efficient score IM test statistic ω_{EFF}. Without ambiguity, we refer to ω_B, ω_W, etc. as IM test statistics and IM tests, with the understanding that ω_B is used with $T^2_{q,B-1}$ critical values, and the other statistics with χ^2_q critical values. We study the ERP under the null of correct specification as well as the power against a heteroskedastic alternative, both in the linear model and in the probit model. Throughout, the IM tests are based on the maximum number of linearly independent indicators. Published Monte Carlo results show that ω_W and ω_{CL}, and to a lesser extent ω_O and ω_{EFF}, suffer from substantial ERPs in these models and that, after bootstrap-correcting the critical values, ω_{EFF} is the most powerful. Our results confirm this and add ω_B to the picture. It turns out that ω_B offers an improvement on ω_{EFF} in terms of smaller ERP, and that it is very close to ω_{EFF} in terms of power. These findings support the intuitive arguments advanced in Section 2.

3.1 The normal linear regression model

The conditional density in this model is $\phi\big((y_t - x_t'\beta)/\sigma\big)$, with ϕ the standard normal density, x_t a $k \times 1$ vector of given regressors, and parameters β ($k \times 1$) and $\sigma > 0$. Hall (1987) shows that the IM test is a combined test against heteroskedasticity (White, 1980), conditional skewness, and non-normal kurtosis.

We use the following Monte Carlo design. The regressor matrix X, which is kept fixed across Monte Carlo replications, consists of a vector of ones and independent drawings from $N(0, 1)$ elsewhere. We note that all IM test statistics

Figure 1: LINEAR MODEL: p-VALUE PLOTS FOR (a) $n = 100$, $k = 2$; (b) $n = 100$, $k = 4$



Horizontal axis: nominal RP. Vertical axis: actual RP.

are invariant under non-singular linear transformations of the rows of X (i.e. transformations $X \to XA$ with A non-singular). Therefore the results extend to any case where the $k-1$ non-constant regressors are generated from a non-singular $(k-1)$-variate normal distribution. We set σ and all elements of β equal to one. However, since all statistics considered are pivotal under the null, the results concerning ERP are valid for any β and σ. We implement a full factorial design with $k = 2, 3, 4, 5$ and $n = 50, 100, 250, 500, 1000$. Throughout, $B = 50$. All results are based on 10000 Monte Carlo replications.

The ERP is displayed using p-value plots (Davidson and MacKinnon, 1998). A p-value plot graphs the empirical distribution function of the p-values of the test statistics generated under the null by Monte Carlo experimentation. The p-values are derived from the approximating distribution, $T^2_{q,B-1}$ or χ^2_q in our case. Thus, a p-value plot gives the (estimated) actual rejection probability (RP) of a test as a function of the nominal RP. On the 45° line actual and nominal RP agree, so one would hope to see a p-value plot close to the 45° line.

In Figure 1 the p-value plots for $n = 100$ and $k = 2, 4$ are given. The

ERP is largest for ω_W and ω_{CL}. This is in fact true for all design points [3]. The performance of ω_B is in general comparable with that of ω_{EFF}, although overall ω_B has the smallest ERP of all tests considered. The behavior of ω_O is better than that of ω_W and ω_{CL}, although its convergence to the 45° line as n grows is remarkably slow. The IM test is a good example of how bad a first-order asymptotic approximation can work: even for $n = 1000$ the ERP of ω_W, ω_{CL}, ω_O and even ω_{EFF} is too large to use these tests in practice. The $T^2_{q,B-1}$ approximation to the distribution of ω_B, on the other hand, works fine for larger sample sizes. Since the statistics are pivotal, all the tests considered can be turned into exact tests using bootstrap-based critical values.

We investigated the power of the IM tests against a heteroskedastic alternative with density $\phi\big((y_t - x_t'\beta)/\mid x_t'\beta\mid^{1/2}\big)$. In order to correct the power for ERP, we plot power as a function of actual RP under the null [4] (Davidson and MacKinnon, 1998). For a test based on pivotal statistics, a power curve is obtained as follows. Run R Monte Carlo replications under the null and under the alternative hypothesis. Order the R test statistics obtained under the null from high to low to obtain $\omega_1^0 \geq \ldots \geq \omega_R^0$. The power at actual RP $k/(R+1)$ is then estimated as the fraction of test statistics generated under the alternative that are larger than ω_k^0. To reduce experimental error, we took the same streams of standard normal drawings under the null and under the alternative.

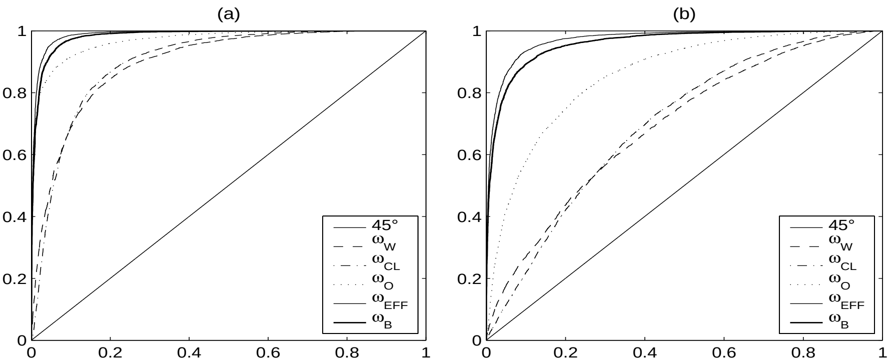
Figure 2 gives the power curves for $n = 100$ and $k = 2, 4$. From the results at these and other design points, the following patterns emerge. The tests ω_W and ω_{CL} have similar power, and are in most cases dominated by ω_O [5]. The tests ω_{EFF} and ω_B have very similar power, with ω_{EFF} being slightly better. These tests always outperform the others. We note that taking B larger would

[3] Detailed results are available from the authors.

[4] Any serious power comparison should of course only compare powers of tests with (nearly) correct RP under the null.

[5] In some cases (e.g. $n = 50$ and $k = 5$), however, ω_O has power smaller than actual RP. Horowitz (1994) also observed this for ω_O in the tobit model.

Figure 2: LINEAR MODEL: POWER CURVES FOR (a) $n = 100$, $k = 2$; (b) $n = 100$, $k = 4$



Horizontal axis: actual RP. Vertical axis: power.

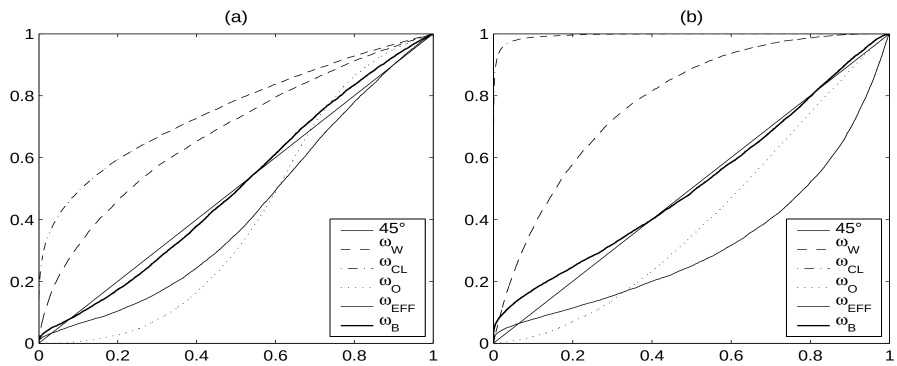
increase the power of ω_B, but a relatively small B already yields a powerful test.

3.2 The probit model

In this model y_t is binary with conditional mean $\Pr[y_t = 1] = \Phi(x_t'\beta)$, where Φ is the standard normal distribution function. The IM test is sensitive to misspecification of $\Pr[y_t = 1]$ as a function of x_t. Orme (1988) gives the efficient score form of the IM test. In the Monte Carlo experiment, we set $\beta = (0.5, 1, \ldots, 1)'$ and choose X, n, k, B, and the number of Monte Carlo replications as in the linear model. None of the IM statistics is pivotal, so the results are specific to the choice of β. A further consequence is that the bootstrap-based critical values are not exact anymore, although their use ensures that the ERP vanishes quickly as the sample size gets larger (Horowitz, 1994).

Figure 3 gives the p-value plots for $n = 100$ and $k = 2, 4$. More or less the same patterns are observed as in the linear model: ω_W and ω_{CL} severely overreject; ω_O, ω_{EFF}, and ω_B have much smaller ERP, although for small n and

Figure 3: PROBIT MODEL: p-VALUE PLOTS FOR (a) $n = 100$, $k = 2$; (b) $n = 100$, $k = 4$



Horizontal axis: nominal RP. Vertical axis: actual RP.

large k it may still be substantial; and most of the time ω_B has the smallest ERP.

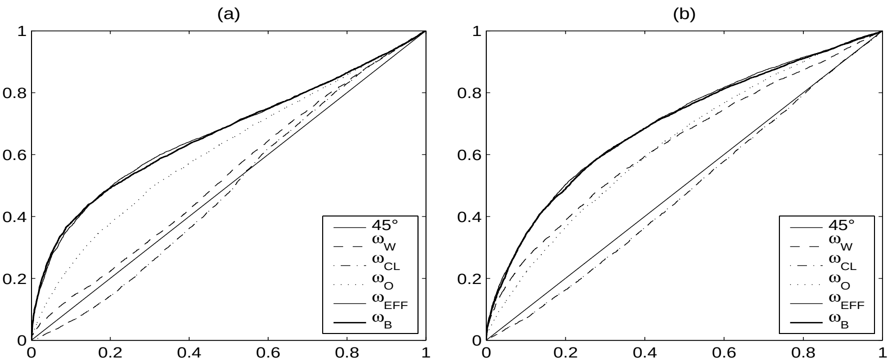
We studied the power of the IM tests against the heteroskedastic probit $\Pr[y_t = 1] = \Phi\big(x_t'\beta / \mid x_t'\beta \mid^{1/2}\big)$. We have noted already that the statistics are not pivotal. This has an implication for the construction of the power curve, because now it matters from which null distribution the statistics are generated. We follow Horowitz (1994), Horowitz and Savin (2000), and Davidson and MacKinnon (1996), and generate the test statistics under the pseudo-true null (rather than any null, as in the linear model) and then proceed to construct power curves as outlined above. The pseudo-true values are computed as the solution of

$$\max_{\beta_*} \sum_{t=1}^{n} \left[\Phi\left(\frac{x_t'\beta}{|x_t'\beta|^{1/2}}\right) \log \Phi(x_t'\beta_*) + \left(1 - \Phi\left(\frac{x_t'\beta}{|x_t'\beta|^{1/2}}\right)\right) \log\big(1 - \Phi(x_t'\beta_*)\big)\right], \tag{7}$$

which is solved by numerical optimization.

Figure 4 displays the power curves for $n = 100$ and $k = 2, 4$. The powers of ω_B and ω_{EFF} are in all design points extremely close to each other, and

Figure 4: PROBIT MODEL: POWER CURVES FOR (a) $n = 100,\ k = 2$; (b) $n = 100,\ k = 4$



Horizontal axis: actual RP. Vertical axis: power.

well above the power of ω_W, ω_{CL} and ω_O. The power of ω_{CL} is in many cases smaller than its actual RP, a fact also observed by Stomberg and White (2000) in a somewhat different setting. We also find that ω_O has smaller power than actual RP in small samples.

4 Conclusion

We have introduced an alternative IM test, which uses the parametric bootstrap to estimate the covariance matrix of the indicator vector. The new test is easy to compute using standard econometric software and requires no analytical derivations. Its analytical simplicity comes of course at a cost, namely it requires a limited number of simulations. When one wants to use bootstrap-based critical values, a nested bootstrap becomes necessary. In the models analyzed here, its performance was found to be comparable to that of the efficient score form of the IM test. Therefore, the IM test proposed here offers a valuable alternative to existing IM tests in more complex models for which the efficient score form of the IM test is not available.

Acknowledgements

Financial support from the Flemish Fund for Scientific Research (grant G.0366.01) is gratefully acknowledged.

References

Chesher, A., 1983, The information matrix test: Simplified calculation via a score test interpretation, Economics Letters 13, 45-48.

Chesher, A. and R. Spady, 1991, Asymptotic expansions of the information matrix test statistic, Econometrica 59, 787-815.

Davidson, R. and J.G. MacKinnon, 1992, A new form of the information matrix test, Econometrica 60, 145-157.

Davidson, R. and J.G. MacKinnon, 1996, The power of bootstrap tests, Queen's University at Kingston, Queen's Institute for Economic Research Working Paper, n° 937.

Davidson, R. and J.G. MacKinnon, 1998, Graphical methods for investigating the size and power of hypothesis tests, The Manchester School 66, 1-26.

Hall, A., 1987, The information matrix test for the linear model, The Review of Economic Studies 54, 257-263.

Horowitz, J.L., 1994, Bootstrap-based critical values for the information matrix test, Journal of Econometrics 61, 395-411.

Horowitz, J.L. and N.E. Savin, 2000, Empirically relevant critical values for hypothesis tests: A bootstrap approach, Journal of Econometrics 95, 375-389.

Lancaster, T., 1984, The covariance matrix of the information matrix test, Econometrica 52, 1051-1053.

Orme, C., 1988, The calculation of the information matrix test for binary data models, The Manchester School 56, 370-376.

Orme, C., 1990, The small-sample performance of the information-matrix test, Journal of Econometrics 46, 309-331.

Stomberg C. and H. White, 2000, Bootstrapping the information matrix test, University of California, San Diego, Department of Economics Discussion Paper, n° 2000-4.

Taylor, L.W., 1987, The size bias of the information matrix test, Economics Letters 24, 63-67.

White, H., 1980, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica 48, 817-838.

White, H., 1982, Maximum likelihood estimation of misspecified models, Econometrica 50, 1-26.