Tests With Correct Size When Instruments Can Be Arbitrarily Weak

Marcelo J. Moreira*

September 17, 2001

Abstract

Classical exponential-family statistical theory is employed to characterize the class of exactly similar tests for a structural coefficient in a simultaneous equations model with normal errors and known reduced-form covariance matrix. A score test and a variant of the Anderson-Rubin test are shown to be members of the class. When identification is strong, the power surface for the score test is generally close to the power envelope for the class of similar tests. However, when identification is weak, the score test no longer has power close to the envelope. Dropping the restrictive assumptions of normality and known covariance matrix, the results are shown to remain valid in large samples even in the presence of weak instruments.

JEL Classification: C12, C31.

Keywords: structural model, power envelope, similar test, score test, confidence regions, LIML estimator.

---

*Lengthy discussions with Thomas Rothenberg were extremely important for this work and I am deeply indebted for all his help and support. For comments and advice, I also would like to thank Peter Bickel, David Card, Kenneth Chay, John Dinardo, Michael Jansson, David Lee, Daniel McFadden, Whitney Newey, James Powell, Tiago Ribeiro, Paul Ruud, Samuel Thompson, Arnold Zellner and seminar participants of the Econometrics seminar and Labor lunch at the University of California at Berkeley.

1 Introduction

Applied researchers are often interested in making inferences about the parameters of endogenous variables in a structural equation. Identification is achieved by assuming the existence of instrumental variables uncorrelated with the structural error but correlated with the endogenous regressors. If the instruments are strongly correlated with the regressors, standard asymptotic theory can be employed to develop reliable inference methods. However, as emphasized in recent work by Bound, Jaeger and Baker (1995), Dufour (1997), and Staiger and Stock (1997), these methods are not satisfactory when instruments are only weakly correlated with the regressors. In particular, the usual tests and confidence regions do not have correct size in the weak instrument case.

This paper develops a methodology for finding tests for structural parameters having correct size even when instruments are weak. The class of tests with this property turns out to be quite large and includes a familiar score test. An upper bound for the power of these tests is derived. The score test is shown to be essentially optimal when instruments are strong, but it may have relatively poor power when instruments are weak. In particular, it can sometimes have lower power than the test proposed by Anderson and Rubin (1949).

The paper is organized as follows. Sections 2 and 3 develop exact results for the special case of a two-equation model with normal errors and known reduced-form covariance matrix. In Section 4 and 5 the results are extended to more realistic cases, although at the cost of introducing some asymptotic approximations. Monte Carlo simulations suggest that the asymptotic approximations are quite accurate. Section 6 indicates how confidence regions can be constructed from the score test. Section 7 contains concluding remarks. All proofs are given in an appendix.

2 Testing with Known Covariance Matrix

2.1 The Model

To avoid tedious notation and asymptotic approximations, it will be useful to begin with a simple special case. Consider the structural equation

$$y_1 = \beta y_2 + u \tag{1}$$

where y_1 and y_2 are $n \times 1$ vectors of observations on two endogenous variables, u is an $n \times 1$ unobserved error vector, and β is an unknown scalar parameter. This equation is assumed to be part of a larger linear simultaneous equations system which implies that y_2 is correlated with u. However, the complete system contains exogenous variables which can be used as instruments for conducting inference on β. Specifically, it is assumed that the reduced form for $Y = [y_1, y_2]$ can be written as

$$\begin{aligned} y_1 &= Z\pi\beta + v_1 \\ y_2 &= Z\pi + v_2 \end{aligned} \tag{2}$$

where Z is an $n \times k$ matrix of nonrandom exogenous variables having full column rank k; π is a $k \times 1$ vector and the n rows of the $n \times 2$ matrix of reduced form errors $V = (v_1, v_2)$ are i.i.d. with mean zero and *known* 2×2 covariance matrix

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{pmatrix}.$$

The restrictions on the reduced form regression coefficients are implied by the identifying assumption that the exogenous variables do not appear in (1).

The goal is to test the hypothesis $H_0 : \beta = \beta_0$, treating π as a nuisance parameter. A test is said to be of size α if the probability of rejecting the null hypothesis when it is true does not exceed α. That is, if P is the subset of k-dimensional Euclidian space in which π is known to lie,

$$\sup_{\pi \in P} \text{prob}\,(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) = \alpha.$$

Since π is unknown, finding a test with correct size is nontrivial. The task is simplified if one can find tests whose null rejection probability does not depend on the nuisance parameters at all. These tests are called *similar tests*. If, for example, one rejects if some test statistic $\mathcal{T}$ is greater than a given constant, the test will be similar if the distribution of $\mathcal{T}$ under the null hypothesis does not depend on π. Such test statistics are said to be *pivotal*. If $\mathcal{T}$ has null distribution depending on π but it can be bounded by a pivotal statistic, then $\mathcal{T}$ is said to be boundedly pivotal.

In practice, one often uses test statistics that are only asymptotically pivotal:

$$\lim_{n\to\infty} \text{prob}\,(\mathcal{T} > c_\alpha) = \alpha.$$

These tests may be satisfactory when the convergence is uniform and the sup and lim operators can be interchanged. However, if the converge is not uniform, the actual size of the test may differ substantially from the size based on the asymptotic distribution of $\mathcal{T}$. In fact, based on earlier work by Gleser and Hwang (1987), Dufour (1997) has shown that the true levels of the usual $Wald$-type tests deviate arbitrarily from their nominal levels if π cannot be bounded away from the origin.[1] Since weak instruments are common in empirical research, it would be desirable to find tests with approximately correct size α even when π cannot be bounded away from the origin.

One such test was proposed by Anderson and Rubin (1949). Let $b = (1, -\beta_0)'$. Define $u_0 \equiv y_1 - y_2\beta_0 = Yb$ and $\sigma_0^2 = b'\Omega b$. In the case where Ω is known, the Anderson-Rubin approach rejects the null hypothesis if

$$AR_0 = u_0' Z(Z'Z)^{-1}Z'u_0/\sigma_0^2 \tag{3}$$

is large. Since $u_0 = Vb$ under H_0, the test statistic is pivotal having a chi square distribution with k degrees of freedom no matter what the value of the nuisance parameter π. When $k = 1$ so the model is exactly identified, this test is uniformly most powerful among the class of unbiased tests. However, if $k > 1$, the test has no particular optimum properties. Indeed, if k is large, the power of the Anderson-Rubin test can be quite low. To improve power,

[1] His analysis remains valid in the special case Ω is unknown.

Wang and Zivot (1998) studied alternative boundedly pivotal test statistics and found critical values to insure the tests have correct size no matter how weak the instruments. Unfortunately, when π is not near the origin, these tests have null rejection probabilities much lower than α and waste power. The tests proposed by Wang and Zivot turn out to be no better than the Anderson-Rubin test. A more fruitful approach is to find other tests that are, like the Anderson-Rubin test, based on pivotal statistics.

2.2 A Family of Similar Tests

Under the normality assumption, the probability model is a member of the curved exponential family. We can therefore adapt the extensive set of results summarized in Lehmann (1986) for testing H_0. Note that, for any non-singular 2×2 matrix D, the two columns of $Z'YD$ are a pair of sufficient statistics for the unknown parameters (β, π). A convenient choice is the pair

$$S = Z'Yb \qquad \text{and} \qquad T = Z'Y\Omega^{-1}a \tag{4}$$

where $a = (\beta_0, 1)'$. Although the null distribution of the statistic $S = Z'u_0$ does not depend on the nuisance parameter π, the null distribution of T is very sensitive to the value of π. Indeed, T is a sufficient statistic for π under the null hypothesis. A little algebra shows that

$$T = a'\Omega^{-1}a \cdot Z'Z\widehat{\pi}$$

where $\widehat{\pi}$ is the maximum likelihood estimate of π when β is constrained to take the null value β_0. The vectors S and T are independent and normally distributed under both the null and alternative hypotheses. Specifically,

$$S \sim N(Z'Z\pi(\beta - \beta_0), Z'Z \cdot b'\Omega b)$$

$$T \sim N(Z'Z\pi \cdot \overline{a}'\Omega^{-1}a, Z'Z \cdot a'\Omega^{-1}a)$$

where $\overline{a} = (\beta, 1)$.

Because they are sufficient statistics, all tests can be written as (possibly randomized) functions of S and T. Specifically, let ϕ be a critical function such that $0 \leq \phi \leq 1$. For each S and T the test rejects or accepts the null with probabilities $\phi(S,T)$ and $1-\phi(S,T)$, respectively. Let E_0 represent expectation over the distribution of S when the null hypothesis is true. In the appendix we show:

THEOREM 1 (SIMILARITY CONDITION): *A test is similar at size α if and only if it can be written as $\phi(S,T)$ such that $E_0\phi(S,t)=\alpha$ for almost every t.*

Since the Anderson-Rubin test depends only on S, it satisfies the similarity condition as long as the appropriate chi-square critical value is used. But there are also similar tests that use the additional information contained in T. Consider, for example, the family of pivotal test statistics

$$\mathcal{T}_g = \frac{g(T)' S}{\sigma_0 \sqrt{g(T)' Z'Z g(T)}} \tag{5}$$

where g is a measurable mapping from $\mathbf{R}^k$ onto $\mathbf{R}^k$. These statistics are $N(0,1)$ under H_0. For the one-sided alternative $\beta > \beta_0$, a similar test at significance level α rejects H_0 if $\mathcal{T}_g > z_\alpha$, the $(1-\alpha)$ standard normal quantile. For a two sided-hypothesis, a similar test at significance level α rejects if $|\mathcal{T}_g| > z_{\alpha/2}$.

EXAMPLE 1: Let $g(T) = (Z'Z)^{-1}T/a'\Omega^{-1}a = \hat{\pi}$ so

$$\mathcal{T}_g = \frac{\hat{\pi}' S}{\sigma_0 \sqrt{\hat{\pi}' Z'Z \hat{\pi}}}$$

The power properties of this test and its extensions when Ω is unknown will be studied in Sections 3 and 4.

EXAMPLE 2: Let $g(x) \equiv d$ for any $x \in \mathbf{R}^k$. In this case, T_g is given by

$$\mathcal{T}_g = \frac{d' S}{\sigma_0 \sqrt{d' Z'Z d}}$$

In Section 3, it will be shown that tests based on this pivotal statistical are optimal when d *happens* to be a positive multiple of π. This test can be used in practice if there is strong prior belief about the coefficients on the instruments.

Example 3: Let the jth element of g be one if $|\hat{\pi}_j|$ is the largest among $|\hat{\pi}_1|, ..., |\hat{\pi}_k|$ and zero otherwise. Tests based on this statistic should have good power properties when only one instrument is valid, but it is not known which one.

Not all pivotal statistics are members of the $\mathcal{T}_g$ family. For example, Staiger and Stock (1997) suggest the possibility of splitting the sample into two half samples to construct similar tests, each sub-sample consisting of $\left\{ y_1^{(j)}, y_2^{(j)}, Z^{(j)} \right\}$ for $j = 1, 2$. Consider the statistic

$$\mathcal{J}_{12} = \frac{\hat{\pi}_{ols}^{(1)'} S^{(2)}}{\sigma_0 \sqrt{\hat{\pi}_{ols}^{(1)\prime} Z^{(2)\prime} Z^{(2)} \hat{\pi}_{ols}^{(1)}}}$$

where $\hat{\pi}_{ols}^{(1)} \equiv \left(Z^{(1)'} Z^{(1)} \right)^{-1} Z^{(1)'} y_2^{(1)}$ is the OLS estimator of π using the first sub-sample and $S^{(2)} \equiv Z^{(2)\prime}(y_1^{(2)} - y_2^{(2)} \beta_0)$. An important feature is that $\hat{\pi}_{ols}^{(1)}$ is independent of $S^{(2)}$, as long as the observations from the two sub-samples are independent. Therefore, $\mathcal{J}_{12}$ is pivotal and tests based on $\mathcal{J}_{12}$ are similar at level α if the appropriate normal critical value is used. However, this test wastes power since $\mathcal{J}_{12}$ requires randomization to be expressed as a function of S and T.

2.3 Pre-testing Procedures

Example 3 can be interpreted as a pre-test procedure: one first decides on the basis of $\widehat{\pi}$ which instrument is best and then tests H_0 using a one-instrument version of the Anderson-Rubin test. Although pre-test procedures are commonly used in econometrics, the fact that the first-step typically affects the size of the second-step test is usually ignored. Pre-testing on the basis of $\widehat{\pi}$ however does not cause any difficulties with tests based on the pivotal statistics $\mathcal{T}_g$. More generally, we have the following implication of Theorem 1:

PROPOSITION 1: *Let $h(T)$ be a measurable real valued function and let $\phi_1(S,T)$ and $\phi_2(S,T)$ be two similar tests at level α. Finally, let $\phi_3 = I[h(T) > c]\phi_1 + I[h(T) \leq c]\phi_2$ where I is the indicator function taking the value one if the argument is true and zero otherwise. Then ϕ_3 is also a similar test at level α.*

For example, one might decide to use the Anderson-Rubin test if $\widehat{\pi}$ is near the origin and use the test of Example 1 if $\widehat{\pi}$ is far from the origin. If the decision is based on the reduced-form "F-statistic" $a'\Omega^{-1}a \cdot \hat{\pi}'Z'Z\hat{\pi}$, the procedure is valid. That is, choosing which similar test to be used after testing if π is significantly different from zero does not affect the final test's size as long as the preliminary test is based on the *constrained* maximum likelihood estimate.

3 Power Functions

3.1 Power Envelope for Similar Tests

When π is far from the origin and the sample size is large, the standard likelihood ratio, Wald, and Lagrange multiplier two-sided tests of the hypothesis $\beta = \beta_0$ are approximately best unbiased and have approximate power

$$1 - G\left(c_\alpha\ ;\ \frac{\pi'Z'Z\pi\,(\beta-\beta_0)^2}{\sigma_0^2}\right) \tag{6}$$

where c_α is the $1-\alpha$ quantile of a central $\chi^2(1)$ distribution and $G(\cdot;\mu)$ is the noncentral $\chi^2(1)$ distribution function with noncentrality parameter μ. However, these tests are not generally similar and the power approximation is unreliable when π is near the origin. Only in the case $k=1$ where the model is exactly identified, do we have an exact optimality result. Then, as shown in the appendix, the Anderson-Rubin AR_0 test is uniformly most powerful unbiased and has exact power function given by (6). Therefore, it is not surprising why Monte Carlo simulations ran by Wang and Zivot (1998) and Zivot, Startz and Nelson (1998) suggests that no test dominates the one proposed by Anderson and Rubin (1949) when $k = 1$. Its power is very close to the power of the Anderson-Rubin AR_0 test, which is itself the optimal test when Ω is known.

When $k > 1$, there exists no uniformly most powerful test. To assess the power properties of the similar tests described in Section 2, it is useful to find the *power envelope*, the upper bound for the rejection probability for each alternative $\beta \neq \beta_0$ and value of π. Moreover, one could find the power upper bound *within* the class of similar tests. In the appendix we show:

THEOREM 2: *For testing $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ when Ω is known and $\pi \neq 0$, we have*

a. *If the model is just identified, the uniformly most power unbiased test has a power function given by*

$$P_{\beta,\pi}(AR_0 > c_\alpha) = 1 - G\left(c_\alpha\ ; \frac{\pi' Z' Z \pi\, (\beta - \beta_0)^2}{\sigma_0^2}\right) \tag{7}$$

b. *If π is known, the uniformly most powerful unbiased test has a power function is given by*

$$P_{\beta,\pi}(\mathcal{R} > c_\alpha) = 1 - G\left(c_\alpha ; \frac{\pi' Z' Z \pi\, (\beta - \beta_0)^2}{\omega_{11} - \omega_{12}^2/\omega_{22}}\right) \tag{8}$$

where $\mathcal{R}$ is defined in equation A.3.

c. *If π unknown and P contains a k-dimensional rectangle, the two-sided power envelope for the class of exactly similar tests is given by*

$$P_{\beta,\pi}\left(\frac{(\pi' S)^2}{\sigma_0^2 \pi' Z' Z \pi} > \ c_\alpha\right) = 1 - G\left(c_\alpha ; \frac{\pi' Z' Z \pi\, (\beta - \beta_0)^2}{\sigma_0^2}\right) \tag{9}$$

Note that (8) is an upper bound for the power of any two-sided test with correct size. Since σ_0^2 necessarily is less than $\omega_{11} - \omega_{12}^2/\omega_{22}$ when u is correlated with v_2, insisting on similarity lowers the attainable power of the test. The optimal test for known π can be understood as the optimal similar test when the nuisance-parameter set contains only one element; the loss in power is then due to increase in the nuisance parameter space.

3.2 Score Test

Theorem 2 suggests that replacing π in (9) by an estimate might lead to a reasonable test. Using the OLS estimator $(Z'Z)^{-1}Z'y_2$ does not produce a similar test. But, as already suggested in our Example 1, using the constrained maximum likelihood estimator does. It is shown in the appendix that the gradient of the log likelihood function with respect to β, when evaluated at $(\beta_0, \widehat{\pi})$, is proportional to $\widehat{\pi}'S$. Hence, we have the following result:

THEOREM 3: *The test that rejects the null if*

$$LM_0 = \frac{(\widehat{\pi}'S)^2}{\sigma_0^2 \widehat{\pi}' Z' Z \widehat{\pi}} \tag{10}$$

is larger than c_α is a Lagrange multiplier (or score) test based on the normal likelihood with Ω known.

Although $\hat{\pi}$ is an unbiased estimator of π if the null hypothesis is true, it is biased under the alternatives $\beta \neq \beta_0$. If fact, $E(\hat{\pi}) = \pi d$, where the scalar d is given by

$$d = \frac{\omega_{11} - \omega_{12}(\beta + \beta_0) + \omega_{22}\beta\beta_0}{\omega_{11} - 2\omega_{12}\beta_0 + \omega_{12}\beta_0^2}. \tag{11}$$

The fact that $\widehat{\pi}$ is a biased estimator of π under the alternative hypothesis does not necessarily imply bad power properties for the LM_0 test. The LM_0 test fails to have good power only when the direction of π, thought as a vector, is not estimated accurately. In the extreme case $d = 0$, $(Z'Z)^{-1/2}\hat{\pi}$ will randomly pick equally likely directions, regardless of the true value of the nuisance parameter π. This suggests that the test will have poor power properties whenever d is near zero.

3.3 Monte Carlo Simulations

To evaluate the power of the LM_0 test, a 10,000 replication Monte Carlo experiment was performed based on design I of Staiger and Stock (1997). The hypothesized value β_0 is zero. The elements of the 100×4 matrix Z are drawn as independent standard normal and then hold fixed over the replications. Three different values of the π vector are used so (in the notation of Staiger and Stock) $\lambda'\lambda/k = \pi' Z' Z \pi / (\omega_{22} k)$ takes the values 0.25 (poor instruments), 1.00 (weak instruments) and 10 (good instruments). The rows of (u, v_2) are i.i.d. normal random vectors with unit variances and correlation ρ. Results are reported for ρ taking the values 0.00, 0.50 and 0.99.

Figures 1-3 graph, for a fixed value of π, the LM_0 and AR_0 rejection probability as a function of the true value of β.[2] The power envelopes for known and unknown π are also included. In each figure, all four power curves are at their minimum of 5% level when $\beta - \beta_0$ is zero. This reflects the fact that each test is unbiased. As expected, the power curves become steeper as the instruments improve.

The power curve of the AR_0 test is asymmetric about $\beta = 0$ for $\rho \neq 0$ with the degree of asymmetry decreasing as ρ approaches zero. More importantly, it is substantially below the power envelope for similar tests even for a small number of instruments (four). The power curve of the LM_0 test is very close to the power envelope for similar tests when the instruments are good and β is close enough to β_0 (Figure 3). However, for some values of β and β_0, the LM_0 test does not have good power.

4 Testing When Ω Is Unknown

4.1 Score Tests

When Ω is unknown, it seems reasonable to construct tests based on (4) after replacing Ω in T with some consistent estimate. Two alternative estimates are available: the OLS estimate $\widehat{\Omega} = Y'(I - Z(Z'Z)^{-1}Z'Y)/(n-k)$ and the estimate that maximizes the likelihood function when β is constrained to equal the hypothesized value β_0. In particular, the score test statistic LM_0

[2] As β varies, ω_{11} and ω_{12} change to keep the structural error variance and the correlation between u and v_2 constant.

can be modified as

$$LM_1 = \frac{(\widetilde{\pi}'S)^2}{\widetilde{\sigma}^2 \widetilde{\pi}' Z' Z \widetilde{\pi}} \tag{12}$$

where $\widetilde{\pi} = (Z'Z)^{-1} Z'Y \widehat{\Omega}^{-1} b$ and $\widetilde{\sigma}^2 = b'\widehat{\Omega} b$; or alternatively as

$$LM_2 = \frac{(\pi_0' S)^2}{n^{-1} u_0' u_0 \cdot \pi_0' Z' Z \pi_0} \tag{13}$$

where $\pi_0 = (Z'M_0 Z)^{-1} Z'M_0 y_2$ is the constrained MLE for π and $M_0 = I - u_0(u_0'u_0)^{-1} u_0$. The LM_1 test has been independently proposed by Kleibergen (2000).

The LM_1 and LM_2 tests can both be interpreted as score tests. Let $V = (y_1 - Z\pi\beta, y_2 - Z\pi)$. Then the term $\widetilde{\pi}'S$ appearing in the numerator of (12) is just the gradient with respect to β of the objective function

$$Q(\beta, \pi) = tr(\widehat{\Omega}^{-1} V'V)$$

evaluated at the constrained maximizing value $(\beta_0, \widetilde{\pi})$. The term $\pi_0' S$ appearing in the numerator of (13) is just the gradient with respect to β of the log likelihood function

$$L(\beta_0, \pi, \Omega) = -n \ln(2\pi) - \frac{n}{2} \ln |\Omega| - \frac{1}{2} tr\left(\Omega^{-1} V'V\right).$$

evaluated at the constrained maximum likelihood estimates.

The fact that the LM_2 test (which is asymptotically similar even with weak instruments) can be interpreted as a score test for the log-likelihood function is somewhat surprising since Zivot, Startz and Nelson (1998) show that their version of the likelihood function score test has poor size properties. The difference arises not from the score itself, but from the estimate used for the variance of the score. The statistic LM_2 uses the asymptotic variance of the score, evaluated at the constrained MLE. The statistic analyzed by Zivot, Startz and Nelson uses instead the Hessian of the concentrated log likelihood function. Although the two tests are asymptotically equivalent, they have different size properties when π is near the origin. In particular, under the weak-instrument asymptotics employed by Staiger and Stock (1997), the Zivot, Startz and Nelson test is not asymptotically similar whereas the LM_2 test is.

4.2 Monte Carlo Simulations

Although the LM_1 and LM_2 tests are not exactly similar, they have good size properties even when the instruments may be weak. To evaluate the rejection probability under H_0, the design I proposed by Staiger and Stock (1997) is once more replicated. Results are reported for the same parameter values used in Section 3 except for sample size. Chi-square critical values with nominal significance levels of 5% were used.

Tables I and II present the rejection probability under H_0 for the LM_1 test, the LM_2 test, the Lagrange multiplier test described in Zivot, Startz and Nelson (1998), hereinafter called LM_H test, and the AR test. For the AR test, a $\chi^2(k)$ critical value was used.

The Zivot, Startz and Nelson test does not have good size properties. For example, when $\rho = 0.99$ and $\lambda'\lambda/k = 0$, the rejection probability of the LM_H test under the null is about .42-.45, far larger than the significance level of .05. Although the AR test rejects the null slightly more often than its nominal size, its rejection probability under H_0 is not sensitive to the values of ρ and $\lambda'\lambda$. For example, Table II shows that the AR test rejects the null about 6% of the time even when the instruments are invalid ($\lambda'\lambda/k = 0$) and y_2 is highly correlated with u ($\rho = 0.99$).

Like the AR test, both LM_1 and LM_2 tests present good size properties. When the number of observations is 20, the null rejection probabilities of the LM_1 test ranges from .065 to .091 and those of the LM_2 test range from .046 to .069. When the number of observations is 80, the LM_1 test rejects about 6% of the time and the LM_2 test rejects about 5% of the time. As the sample size increases, the distribution of the LM_1 and LM_2 statistics approaches the distribution of the LM_0 statistic. This effect is similar to approximating the distribution of the $LIML$ estimator by the distribution of the $LIMK$ estimator, cf. Anderson, Kunitomo and Sawa (1982).

Results for power are analogous. When Staiger and Stock's design I with 100 observations are used as in Section 3, the power curves of the LM_1 and LM_2 tests are very close to the power curve of the LM_0 test. Tables III-V compare the power of the LM_0 test with that of the LM_2 test. The difference between the two power curves is small, which suggests that the power comparison in Section 3.1 for the LM_0 test holds also for the LM_2 test.

Finally, Tables VI and VII show the rejection probabilities of some nominal 5% tests when Staiger and Stock's design II is used. The structural disturbances, u and v_2, are serially uncorrelated with $u_t = (\xi_{1t}^2 - 1)/\sqrt{2}$ and $v_{2t} = (\xi_{2t}^2 - 1)/\sqrt{2}$ where ξ_{1t} and ξ_{2t} are normal with unit variance and correlation $\sqrt{\rho}$. The k instruments are indicator variables with equal number of observations in each cell. Even though the disturbances are non-normal, the rejection probabilities under H_0 of the LM_1 and LM_2 tests still remain approximately equal to 5% for all values of $\lambda'\lambda/k$ and ρ.

4.3 Large-Sample Properties Under Weak-Instrument Asymptotics

Monte Carlo simulations suggest that, even for unknown Ω and non-normal disturbances, the LM_1 and LM_2 tests still have good size properties. This claim is supported by the fact that both tests are asymptotically similar even under the weak-instrument asymptotics suggested by Staiger and Stock (1997). Kleibergen (2000) presents a proof for the LM_1 test. The following result holds for the LM_2 test:

PROPOSITION : *Suppose $\pi = n^{-1/2}c$, where c is a fixed k-dimensional vector, and the following limits hold jointly:.*

(i) $\left(\frac{u'u}{n}, \frac{v_2'u}{n}, \frac{v_2'v_2}{n}\right) \xrightarrow{p} (\sigma_u^2, \sigma_u\sigma_2 p, \sigma_2^2)$
(ii) $\frac{Z'Z}{n} \xrightarrow{p} Q$
(iii) $\left(\frac{Z'u}{n^{1/2}}, \frac{Z'v_2}{n^{1/2}}\right) \Rightarrow (\Psi_{Zu}, \Psi_{Zv})$ where $(\Psi'_{Zu}, \Psi'_{Zv}) \sim N(0, \Omega \otimes Q)$

Then the following convergence results hold jointly as $n \to \infty$:

(i) $\frac{Z'M_uZ}{n} \xrightarrow{p} Q$;
(ii) $(Z'M_uZ)^{-1/2} Z'M_uy_2 \Longrightarrow \sigma_2 \{\lambda + \eta\}$;
(iii) $(Z'M_uZ)^{-1/2} Z'u \Longrightarrow \sigma_u z_u$,

where $\lambda \equiv \sigma_2^{-1} Q^{1/2} C$*,* $z_v \equiv \sigma_2^{-1} Q^{-1/2} \Psi_{Zv}$*,* $z_u \equiv \sigma_u^{-1} Q^{-1/2} \Psi_{Zu}$*,* $z_v \equiv \rho z_u + (1-\rho^2)^{1/2} \xi \equiv \rho z_u + \eta$ *and* ξ *is a* k*-dimensional normal random vector independent of* z_u*. Furthermore,*

$$LM_2 \Longrightarrow \frac{[(\lambda + \eta) z_u]^2}{(\lambda + \eta)' (\lambda + \eta)}$$

Since z_u *and* η *are independent, under the null* LM_2 *is asymptotically distributed as chi-squared with one degree of freedom.*

5 Extensions

The previous theory can easily be extended to a structural equation with more than two endogenous variables and with additional exogenous variables as long as inference is to be conducted on *all* the endogenous coefficients. Consider the structural equation

$$y_1 = Y_2 \beta + X \gamma + u$$

where Y_2 is the $n \times l$ matrix of observations on the l explanatory endogenous variable and X is the $n \times r$ matrix of observations on r exogenous variables. This equation is part of a larger linear system containing the additional exogenous variables Z. The reduced form for $Y = [y_1, Y_2]$ is

$$\begin{aligned} y_1 &= Z\Pi\beta + X\delta + v_1 \\ Y_2 &= Z\Pi + X\Gamma + V_2 \end{aligned}$$

where $\delta = \Gamma\beta + \gamma$. The rows of $V = [v_1, V_2]$ are i.i.d. with mean zero and covariance matrix Ω. It is assumed that X and Z have full column rank. The problem is to test the joint hypothesis $H_0 : \beta = \beta_0$ treating Π, Γ, δ, and Ω as nuisance parameters.

The unknown parameters associated with X can be eliminated by taking orthogonal projections. Define the $n \times n$ idempotent projection matrix $M = I - X(X'X)^{-1}X'$ and the $l+1$ component column vector $b = (1, -\beta_0')'$.

Let A be any $(l+1) \times l$ matrix whose columns are orthogonal to b. Then defining $u_0 = y_1 - Y_2\beta_0 = Yb$, the statistics

$$S = Z'Mu_0 \qquad \text{and} \qquad T = Z'MY\Omega^{-1}A$$

are independent and normally distributed. For any $k \times l$ matrix $G(T)$ that is a measurable function of T with rank l, the test statistic

$$S'G(T)'[G(T)'Z'MZG(T)]^{-1}G(T)'S/\sigma_0^2$$

has a null $\chi^2(l)$ distribution. Again, Ω can be replaced with a consistent estimate without affecting the results asymptotically. For example, the LM_2 test generalizes to

$$\frac{u_0'MZ\Pi_0(\Pi_0'Z'MZ\Pi_0)^{-1}\Pi_0'Z'Mu_0}{u_0'u_0/n}$$

where the constrained maximum likelihood estimator Π_0 is given by $\Pi_0 = (Z'M^*Z)^{-1} ZM^*Y_2$ and $M^* = M - Mu_0(u_0'Mu_0)^{-1}u_0'M$.

6 Confidence Regions

Valid confidence regions for β can be constructed by inverting similar tests. For example, let C be the set of all values β_0 that cannot be rejected using a similar test of level α. Then C is a confidence set with coverage probability $1-\alpha$. If the score test statistic is used, the resulting confidence region will be valid in large samples no matter how weak the instruments. The score test's confidence region will necessarily contain the limited-information maximum-likelihood estimator of β.

To illustrate how informative the confidence regions based on the score test are, design I of Staiger and Stock (1997) is once more used. One sample was drawn where the true value of β is zero. Figures 4-7 plot the LM_2 statistic as a function of β_0. The region in which the LM_2 statistic is below the horizontal critical value line is the corresponding confidence set. The figures show that the confidence regions can have complicated shape. In all the examples, the true parameter $\beta = 0$ is inside of the confidence region.

When the instruments are invalid (Figure 4), the confidence regions cover the real line. This is expected to happen about $(1-\alpha)\cdot 100\%$ of time since the confidence regions have correct coverage probability and in this case the parameter β is unidentified. As the quality of the instruments increases, the confidence regions become narrower. For example, when $\lambda'\lambda/k = 10$ and $\rho = 0.5$, the confidence region is the set $[-0.1, 0.3] \cup [2.7, 3.5]$ (Figure 7). While some values of β_0 very close to the true β are excluded, values as large as 3.5 are included in this confidence region. This suggests that other similar tests might have better power and, consequently, more accurate confidence regions.

7 Conclusions

Previous authors have noted that the simultaneous equations model with known reduced-form covariance matrix has a simpler mathematical structure than the model with unknown covariance matrix, but inference procedures for the two models behave very much alike in moderate sized samples. Exploiting this fact, classical statistical theory has been employed in this paper to characterize the class of similar tests in the simpler model. Replacing the reduced-form covariance matrix with an estimate appears to have little effect on size and power.

Unlike alternative procedures using the bootstrap or higher-order asymptotics, the methods discussed in this paper behave well even in the extreme case where there is no identification at all. Confidence regions based on the score statistics LM_1 and LM_2 have coverage probabilities close to their nominal level no matter how weak the instruments; and they are informative when the instruments are good. Improved inference in the weak instrument case might be possible by exploring the properties of other tests that satisfy the similarity condition of Theorem 1.

The results derived here cover only the case where inference is desired for the complete set of endogenous-variable coefficients. Inference on the coefficient of one endogenous variable when the structural equation contains additional endogenous explanatory variables is not allowed. Dufour (1997) shows how this limitation can be overcome in the context of the Anderson-Rubin test and the same projection technique can be applied to the similar tests discussed here. However, this may entail considerable loss of power.

Department of Economics, University of California at Berkeley, 549 Evans Hall #3880, Berkeley, CA 94720-3880 USA; jovita@econ.berkeley.edu; http://socrates.berkeley.edu/˜jovita

8 References

Anderson, T., N. Kunitomo, and T. Sawa (1982): "Evaluation of the Distribution Function of the Limited Information Maximum Likelihood Estimator," *Econometrica*, 50, 1009-28.

Anderson, T., and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20, 46-63.

Bound, J., D. Jaeger and R. Baker (1995): "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak," *Journal of American Statistical Association*, 90, 443-50.

Dufour, J-M. (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65, 1365-88.

Gleser, L., and J. Hwang (1987): "The non-existence of 100(1-α)% confidence sets of finite expected diameter in errors-in-variables and related models," *Annals of Statistics*, 15, 1351-62.

Kleibergen, F. (2000): "Pivotal Statistics for testing Structural Parameters in Instrumental Variables Regression," Tinbergen Institute Discussion Paper TI 2000-055/4.

Lehmann, E. (1986): *Testing Statistical Hypothesis.* 2nd edition, Wiley Series in Probability and Mathematical Statistics.

Staiger, D., and J. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65, 557-86.

Wang, J., and E. Zivot (1998): "Inference on a structural parameter in instrumental variables regression with weak instruments," *Econometrica*, 66, 1389-404.

Zivot, E., R. Startz, and C. Nelson (1998): "Valid confidence intervals and inference in the presence of weak instruments," *International Economic Review*, 39, 1119-44.

Table i

Percent Rejected under H_0 at Nominal Level of 5% (20 obs.)

ρ	$\lambda'\lambda/k$	AR	LM_1	LM_2	LM_H
0.00	0.00	9.3	9.0	6.6	4.3
0.00	0.25	9.7	9.0	6.5	4.4
0.00	1.00	9.6	7.7	5.5	4.5
0.00	10.00	9.4	6.9	4.9	4.9
0.50	0.00	9.4	9.1	6.8	13.1
0.50	0.25	9.9	8.8	6.3	11.8
0.50	1.00	9.9	7.5	5.2	7.8
0.50	10.00	9.5	6.6	4.6	5.0
0.99	0.00	9.9	9.1	6.9	45.4
0.99	0.25	9.7	7.2	4.9	37.5
0.99	1.00	9.7	6.6	4.6	21.1
0.99	10.00	9.7	6.5	4.6	6.8

Table ii

Percent Rejected under H_0 at Nominal Level of 5% (80 obs.)

ρ	$\lambda'\lambda/k$	AR	LM_1	LM_2	LM_H
0.00	0.00	6.3	5.8	5.4	4.9
0.00	0.25	5.7	5.1	4.8	4.4
0.00	1.00	6.3	5.5	5.0	4.6
0.00	10.00	6.1	5.6	5.1	5.1
0.50	0.00	5.9	5.5	5.1	13.7
0.50	0.25	5.9	5.8	5.4	11.9
0.50	1.00	5.9	5.9	5.4	8.9
0.50	10.00	5.9	5.4	5.0	5.2
0.99	0.00	6.0	5.6	5.1	42.2
0.99	0.25	5.8	5.5	5.1	33.2
0.99	1.00	6.3	5.6	5.1	18.8
0.99	10.00	5.9	4.9	4.5	6.2

Table iii

Percent Rejected at Nominal Level of 5%

Poor Instruments

	$\rho = 0.00$		$\rho = 0.50$		$\rho = 0.99$	
β	LM_0	LM_2	LM_0	LM_2	LM_0	LM_2
-2.00	7.3	7.6	9.3	9.8	49.1	48.1
-1.80	7.0	7.3	9.7	10.1	57.4	55.9
-1.60	7.2	7.4	8.8	9.2	70.4	69.4
-1.40	6.9	7.3	9.8	9.8	88.1	87.1
-1.20	6.9	7.3	9.2	9.5	99.5	99.2
-1.00	6.5	6.9	8.7	9.1	69.7	68.5
-0.80	6.2	6.3	8.0	8.5	90.8	90.0
-0.60	5.7	5.9	7.2	7.3	29.5	28.8
-0.40	5.5	5.7	6.0	6.1	9.7	9.9
-0.20	5.3	5.4	5.1	5.2	5.7	5.7
0.00	5.1	5.2	5.2	5.4	4.6	4.7
0.20	5.3	5.6	5.4	5.8	5.4	5.1
0.40	5.5	5.8	5.3	5.6	5.9	5.8
0.60	5.7	6.1	5.9	6.1	6.6	6.5
0.80	6.3	6.6	6.1	6.3	7.1	7.3
1.00	6.8	7.1	6.6	6.7	7.9	7.7
1.20	6.9	7.4	6.9	7.2	8.7	8.5
1.40	7.1	7.5	7.4	7.5	8.4	8.2
1.60	7.1	7.3	7.1	7.6	9.3	9.4
1.80	7.6	8.0	7.8	7.9	9.6	9.4
2.00	7.5	7.7	7.3	7.7	10.0	9.7

Table IV
Percent Rejected at Nominal Level of 5%
Weak Instruments

	$\rho = 0.00$		$\rho = 0.50$		$\rho = 0.99$	
β	LM_0	LM_2	LM_0	LM_2	LM_0	LM_2
-2.00	17.4	17.5	21.1	21.5	99.2	99.1
-1.80	16.6	16.6	21.7	22.4	99.8	99.8
-1.60	16.8	16.7	22.5	22.9	100.0	100.0
-1.40	16.7	16.7	23.0	22.6	100.0	100.0
-1.20	15.8	15.5	24.0	23.9	100.0	100.0
-1.00	14.8	14.8	25.5	24.6	92.0	86.8
-0.80	13.6	13.5	23.5	22.8	100.0	100.0
-0.60	11.2	10.8	18.0	17.8	89.0	88.2
-0.40	8.3	8.0	11.1	10.8	30.8	30.2
-0.20	5.9	6.1	6.9	6.8	8.2	8.0
0.00	4.7	4.7	4.9	4.9	4.9	4.7
0.20	6.3	6.4	6.3	6.4	6.3	6.3
0.40	8.0	8.0	8.4	8.5	10.5	10.3
0.60	11.1	11.0	10.6	10.7	12.9	12.6
0.80	13.1	13.6	12.7	12.5	16.9	16.4
1.00	15.4	15.1	14.5	14.2	19.5	18.8
1.20	15.9	16.2	16.4	16.2	22.9	22.5
1.40	17.1	16.6	17.6	17.1	25.6	24.7
1.60	17.5	16.9	19.2	18.7	27.6	26.9
1.80	16.9	16.9	19.8	19.2	29.7	28.7
2.00	17.6	17.3	20.2	19.5	32.6	31.5

Table v

Percent Rejected at Nominal Level of 5%

Good Instruments

	$\rho = 0.00$		$\rho = 0.50$		$\rho = 0.99$	
β	LM_0	LM_2	LM_0	LM_2	LM_0	LM_2
-2.00	97.6	95.0	62.8	62.1	100.0	100.0
-1.80	98.5	96.4	65.0	63.9	100.0	100.0
-1.60	99.1	97.7	72.7	69.4	100.0	100.0
-1.40	99.2	98.0	84.5	78.9	100.0	100.0
-1.20	99.2	98.5	94.6	89.7	100.0	100.0
-1.00	98.5	97.9	98.6	96.6	99.9	95.6
-0.80	96.7	95.9	99.0	98.3	100.0	100.0
-0.60	88.5	87.5	95.2	94.1	100.0	100.0
-0.40	64.6	62.8	68.7	67.8	97.1	96.8
-0.20	23.7	22.9	22.0	21.5	31.4	31.1
0.00	5.2	5.1	5.3	5.3	4.8	4.7
0.20	22.9	22.1	15.9	15.5	16.5	16.1
0.40	64.1	62.5	40.7	39.8	38.6	37.6
0.60	88.7	87.6	63.1	61.3	59.2	58.4
0.80	96.5	95.7	78.7	77.5	73.3	72.1
1.00	98.7	98.1	86.3	85.5	83.6	82.8
1.20	99.1	98.4	91.3	90.3	88.9	88.2
1.40	99.2	98.3	94.0	92.9	92.9	92.6
1.60	98.9	97.5	95.5	94.7	95.3	94.8
1.80	98.4	96.4	96.7	95.9	96.3	95.9
2.00	97.9	95.4	97.3	96.6	97.3	97.1

Table vi

Percent Rejected Under H_0 at Nominal Level of 5% (20 obs.)

non-normal disturbances and binary instruments

ρ	$\lambda'\lambda/k$	AR	LM_1	LM_2	LM_H
0.00	0.00	11.59	10.08	7.34	5.30
0.00	0.25	11.63	10.59	7.84	5.46
0.00	1.00	11.81	10.17	7.40	5.89
0.00	10.00	11.41	11.36	9.01	8.82
0.50	0.00	11.26	10.33	7.65	7.73
0.50	0.25	11.86	10.41	7.71	7.81
0.50	1.00	11.18	10.49	7.77	6.95
0.50	10.00	11.66	11.50	9.30	9.66
0.99	0.00	11.32	11.19	8.45	46.48
0.99	0.25	12.12	11.67	9.17	42.49
0.99	1.00	11.70	12.01	9.98	30.16
0.99	10.00	11.33	11.36	9.41	12.18

Table vii

Percent Rejected Under H_0 at Nominal Level of 5% (80 obs.)

non-normal disturbances and binary instruments

ρ	$\lambda'\lambda/k$	AR	LM_1	LM_2	LM_H
0.00	0.00	6.62	6.47	5.92	5.13
0.00	0.25	6.82	6.57	6.14	5.50
0.00	1.00	6.61	6.63	6.17	5.69
0.00	10.00	6.37	6.90	6.37	6.26
0.50	0.00	6.20	6.55	5.99	7.74
0.50	0.25	6.23	5.80	5.34	6.87
0.50	1.00	6.67	6.70	6.14	7.13
0.50	10.00	6.48	6.95	6.53	6.62
0.99	0.00	6.43	6.56	6.00	43.69
0.99	0.25	7.22	7.41	7.03	40.74
0.99	1.00	6.30	6.60	6.30	31.47
0.99	10.00	6.58	6.88	6.37	10.36

Figure 1
Empirical Power of Tests: Poor Instruments

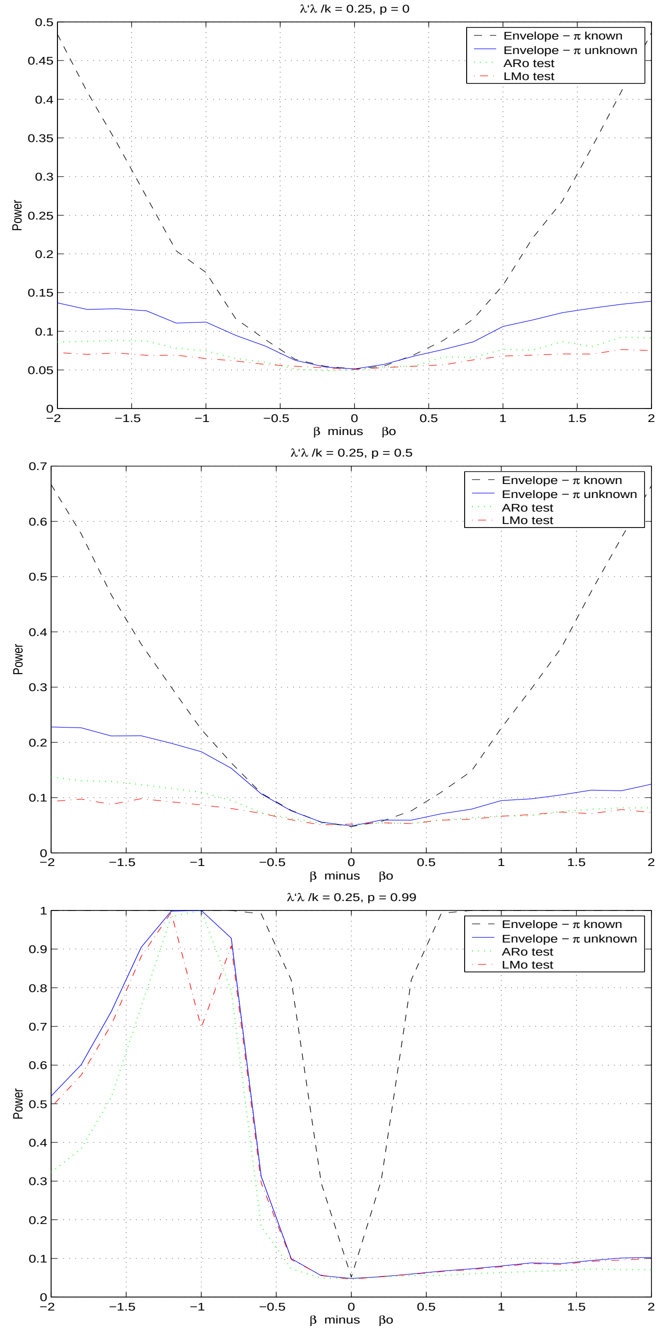


Figure 2
Empirical Power of Tests: Weak Instruments

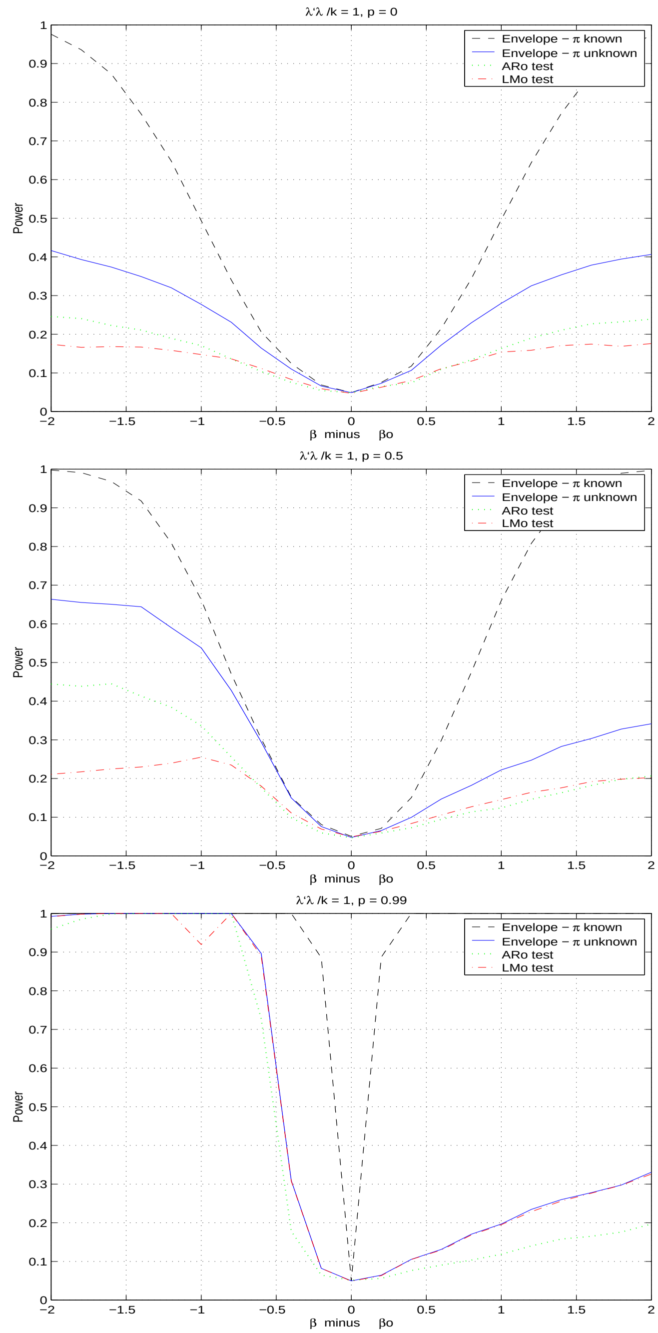


Figure 3
Empirical Power of Tests: Good Instruments

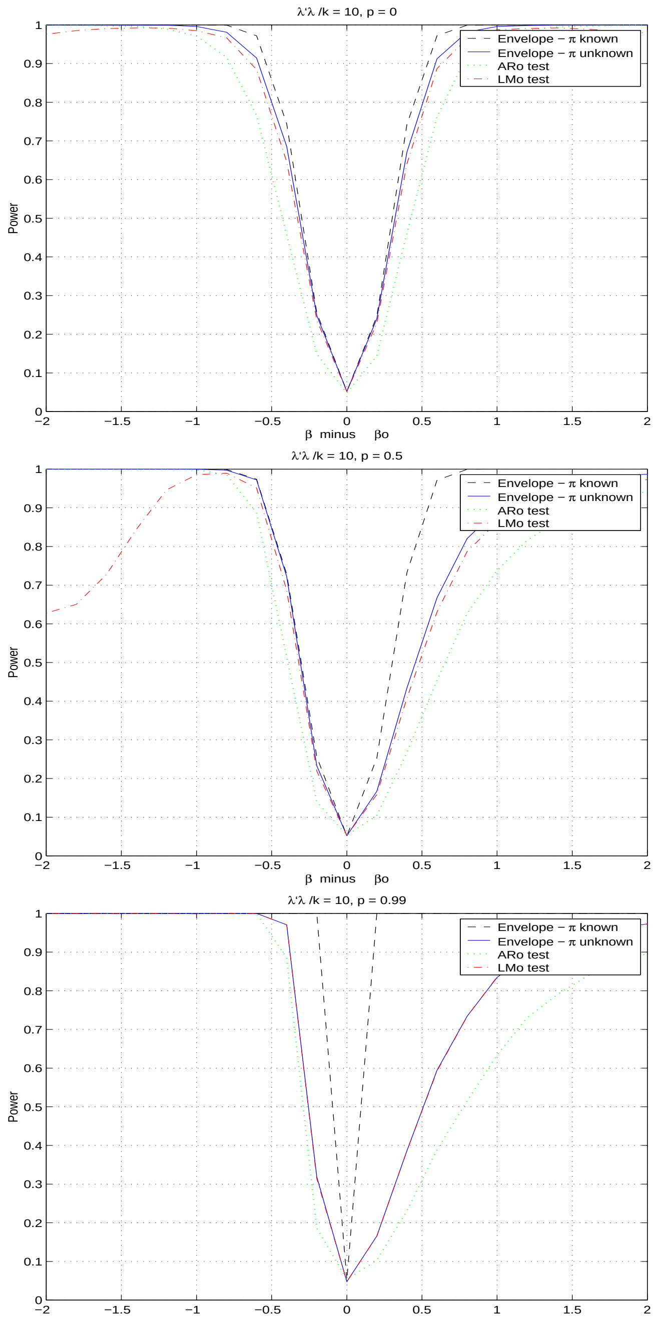


Figure 4

confidence regions: Invalid Instruments

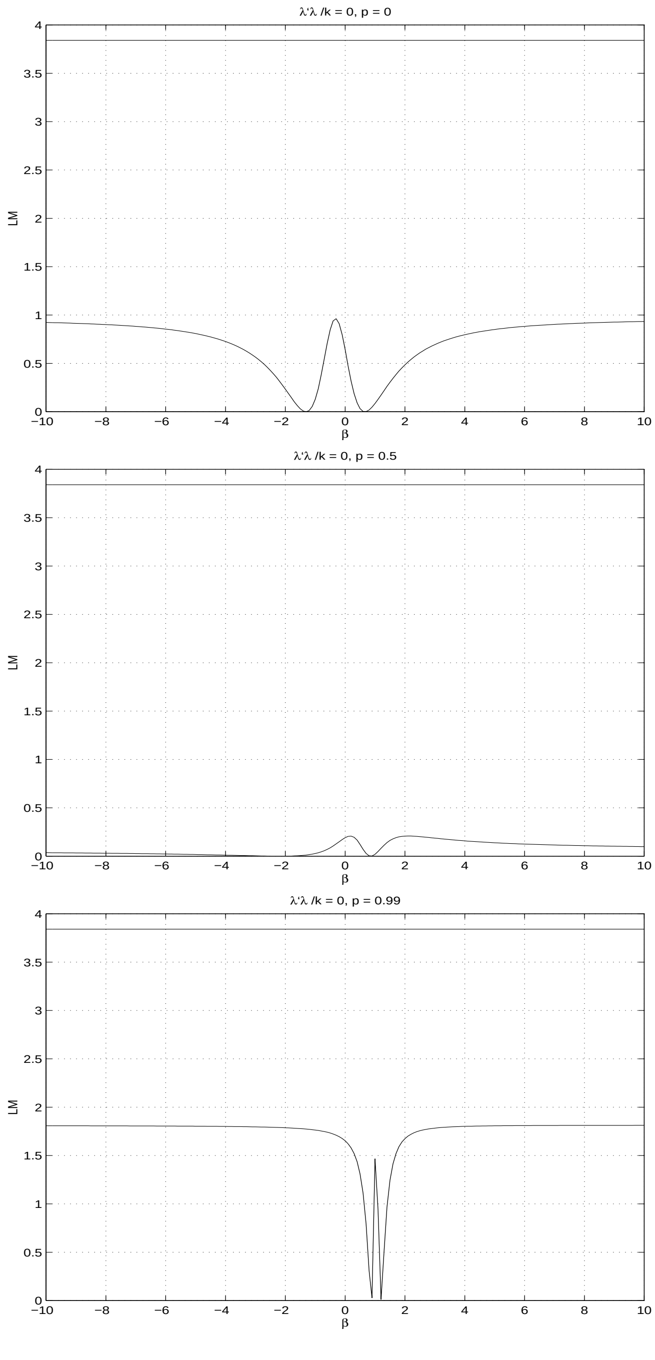


Figure 5

confidence regions: Poor Instruments

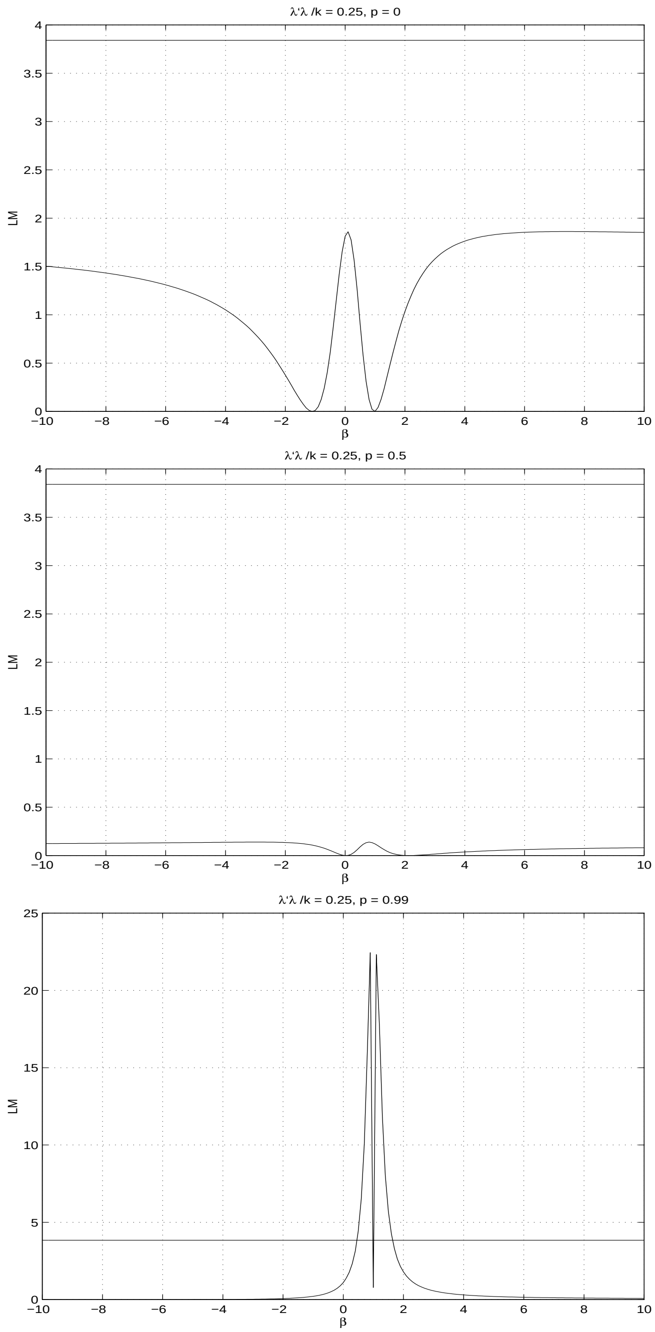


Figure 6

confidence regions: Weak Instruments

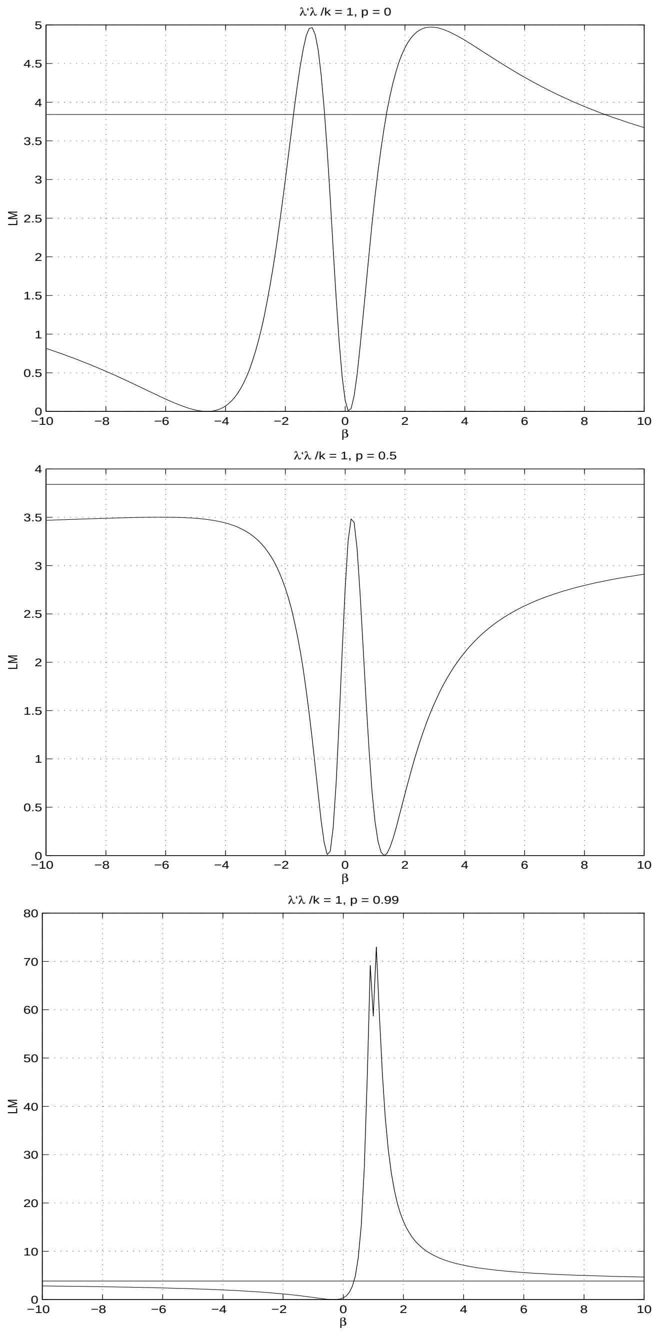
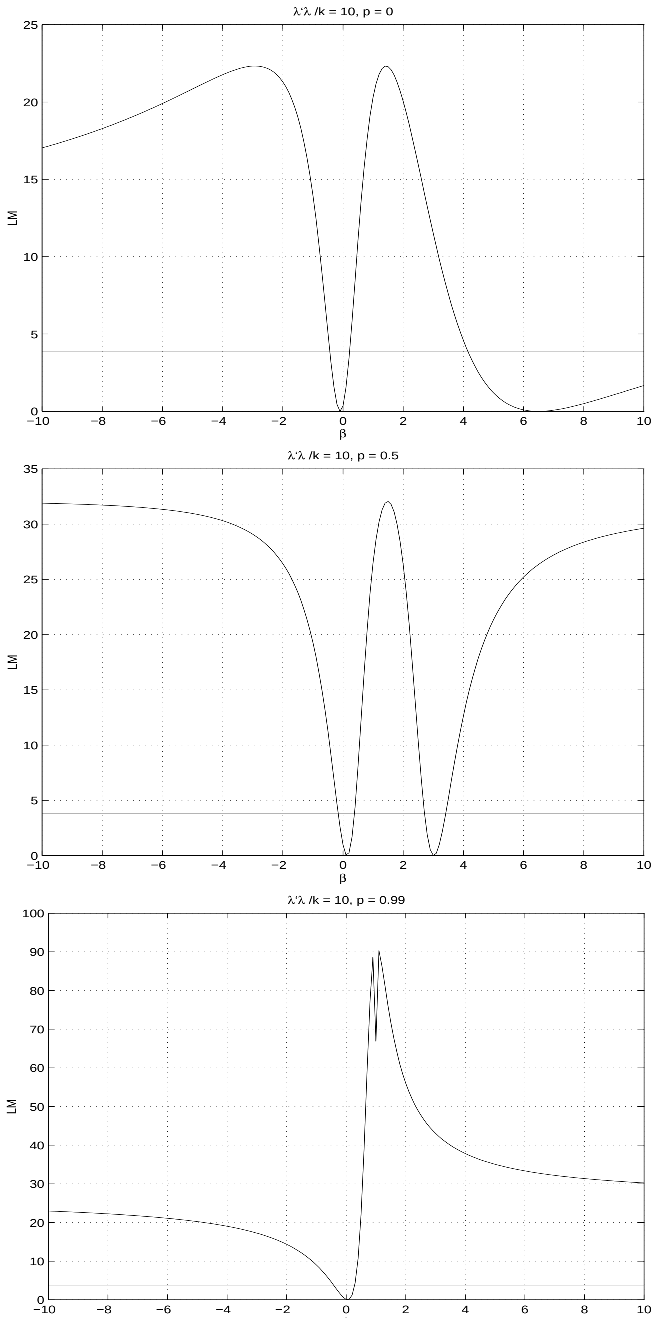


Figure 7

confidence regions: Good Instruments



9 Appendix

The results stated in Sections 2 and 3 are based on the following two lemmas proved in Lehmann (1986), pp. 142-3:

LEMMA A.1: *Let X be a random vector with probability distribution*

$$dP_\theta(x) = C(\theta) \exp\left[\sum_{j=1}^{k} T_j(x)\,\theta_j\right] d\mu(x)$$

and let P^T be the family of distributions of $T = (T_1(X), ..., T_k(X))$ as θ ranges over the set W. Then P^T is complete provided W contains a k-dimensional rectangle.

LEMMA A.2: *Suppose that the distribution of X is given by*

$$dP_{\theta,\mathcal{V}}(x) = C(\theta,\mathcal{V}) \exp\left[\theta R(x) + \sum_{j=1}^{k} \mathcal{V}_j T_j(x)\right] d\mu(x)$$

where the V_j are the nuisance parameters and μ is absolutely continuous with respect to the Lebesgue measure. Suppose that $S = h(R,T)$ is independent of T when $\theta = \theta_0$ and that

$$h(r,t) = a(t)r + b(t) \qquad \text{with } a(t) > 0.$$

Then the uniformly most powerful unbiased (UMPU) test ϕ for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is given by

$$\phi(s) = \begin{cases} 1 & \text{if } s < C_1 \, or \; s > C_2 \\ 0 & \text{otherwise} \end{cases}$$

where C_1 and C_2 are determined by $E_0\{\phi(S)\} = \alpha$ and $E_0\{S\phi(S)\} = \alpha E_0\{S\}$.

PROOF OF THEOREM 1: Since randomization is allowed, any test can be written as $\phi(S,T)$. Since the test is similar at size α, it must be the case that:

$$E_0\phi(S,T) = \alpha \qquad , \forall \pi \in P \tag{A.1}$$

By Lemma A.1, the family of distributions of T when the null hypothesis is true, $\mathcal{P}^T = \left\{P^T_{\beta_0,\pi}; \pi \in P\right\}$, is complete. Consequently, the following holds:

$$E_0\left\{\phi\left(S,T\right)|t\right\} = \alpha \qquad , a.e.\ \mathcal{P}^T$$

Note that the distribution of S does not depend on π under the null hypothesis and that S is independent of T. Therefore, using also the fact that ϕ is integrable:

$$E_0\phi\left(S,t\right) = \alpha \qquad , a.e.\ \mathcal{P}^T \tag{A.2}$$

Conversely, if the test is such that (A.2) holds, then (A.1) is trivially true. Therefore, the test is similar at size α.

Q.E.D.

PROOF OF THEOREM 2: The following is true:
a. For some measure $\mu\left(y\right)$, the probability distribution of Y can be written as:

$$dP_{\theta,\pi}\left(y\right) = C\left(\theta,\pi\right)\exp\left[\theta R\left(y\right) + \pi T\left(y\right)\right]d\mu\left(y\right)$$

where $R\left(Y\right)$ is the first column of $Z'Y\Omega^{-1}$ and $\theta = \pi\left(\beta - \beta_0\right)$. Since P does not contain the origin and the model is just identified, testing $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ is equivalent to testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Let

$$\bar{S} = \frac{\left(Z'Z\right)^{-1/2} Z'\left(y_1 - y_2\beta_0\right)}{\sigma_0}.$$

Notice that $\bar{S} = \delta_1 R + \delta_2 T$ where

$$\delta_1 = \sigma_0\left(Z'Z\right)^{-1/2} \qquad \text{and} \qquad \delta_2 = \frac{-\omega_{22}\beta_0 + \omega_{12}}{\sigma_0}\left(Z'Z\right)^{-1/2}$$

Now Lemma A.2 can be applied. Since $\bar{S} \sim N\left(0,1\right)$ under H_0 and, in particular, it is symmetric around zero, it is straightforward to show that the optimal test rejects the null if $AR_0 > c_\alpha$. Under the alternative β,

$$AR_0 \sim \chi^2\left(1, \frac{\pi' Z' Z \pi \left(\beta - \beta_0\right)^2}{\sigma_0^2}\right).$$

Consequently, the power of the optimal test is given by (7).

b. Since π is known, for some measure $\mu(y)$, the probability distribution can be written as:

$$dP_{\beta,\pi}(y) = C(\beta,\pi)\exp\left[R(y)'\pi\beta\right] d\mu(y)$$

Since this distribution is a one-parameter exponential family, the $UMPU$ test rejects the null hypothesis if

$$\mathcal{R} = \frac{\left\{\pi' Z'\left[\left(y_1 - Z\pi\beta_0\right) - \omega_{12}\omega_{22}^{-1}\left(y_2 - Z\pi\right)\right]\right\}^2}{\left(\omega_{11} - \omega_{12}\omega_{22}^{-1}\omega_{12}\right)\pi' Z' Z\pi} \tag{A.3}$$

is larger than c_α. Under the alternative β,

$$\mathcal{R} \sim \chi^2\left(1, \frac{\pi' Z' Z\pi}{\left(\omega_{11} - \omega_{12}\omega_{22}^{-1}\omega_{12}\right)}\left(\beta - \beta_0\right)^2\right)$$

Consequently, the power of the optimal test is given by (8).

c. The power of the test ϕ is given by $E_{\beta,\pi}\phi(S,T)$. Since S and T are independent, then:

$$E_{\beta,\pi}\phi(S,T) = \int\left[\int \phi(s,t) f(s,\beta,\pi)\, ds\right] g(t,\beta,\pi)\, dt$$

where $f(s,\beta,\pi)$ and $g(t,\beta,\pi)$ are the density functions associated to S and T, respectively. Notice that the power conditioned on $T = t$ is

$$\int \phi(s,t) f(s,\beta,\pi)\, ds.$$

Consider the test $\phi^*(S)$ that assigns 1 if $f(s,\beta,\pi) > kf(s,\beta_0)$ and 0 otherwise, where k is chosen such that $E_{\beta_0,\pi}\phi^*(S) = \alpha$. The claim is that the test $\phi^*(S)$ is most powerful among all similar tests at the significance level α.

Let S^+ and S^- be the sets in the sample space where $\phi^*(s) - \phi(s,t) > 0$ and $\phi^*(s) - \phi(s,t) < 0$, respectively. Notice that, if s is in S^+, $\phi^*(s) = 1$ and $f(s,\beta,\pi) > kf(s,\beta_0)$. Analogously, if s is in S^-, $\phi^*(s) = 0$ and $f(s,\beta,\pi) \leq kf(s,\beta_0)$. Therefore:

$$\int [\phi^*(s) - \phi(s,t)] [f(s,\beta,\pi) - kf(s,\beta_0)] \, ds \geq 0$$

The difference in power satisfies

$$\int [\phi^*(s) - \phi(s,t)] f(s,\beta,\pi) \, dv \geq k \int [\phi^*(s) - \phi(s,t)] f(v,\beta_0) \, ds$$

By Theorem 1, if the test $\phi(S,T)$ is similar then $E_0\phi(S,t) = \alpha$, *a.e.* $\mathcal{P}^T$. Without loss of generality, it can be considered that $E_0\phi(S,t) = \alpha$, $\forall t$. That is:

$$\int \phi(s,t) f(s,\beta,\pi) \, ds = \alpha \qquad , \forall t$$

Therefore, the following holds:

$$\int [\phi^*(s) - \phi(s,t)] f(s,\beta,\pi) \, ds \geq 0$$

Since the test that maximizes the conditional power does not depend on t, then this test itself maximizes power, as was to be proved. Since S is normally distributed, $f(s,\beta,\pi) > kf(s,\beta_0)$ for some k such that $E\phi^*(S) = \alpha$ if and only if the following holds. If $\beta > \beta_0$ then the test rejects the null if $\pi' S > z_\alpha \sqrt{\sigma_0^2 \pi' Z' Z \pi}$. If $\beta < \beta_0$ then the test rejects the null if $\pi' S < -z_\alpha \sqrt{\sigma_0^2 \pi' Z' Z \pi}$, where z_α is the critical value of a $\mathcal{N}(0,1)$ distribution for the significance level α. For two-sided alternative, the optimal test rejects H_0 if

$$\frac{(\pi' S)^2}{\sigma_0^2 \pi' Z' Z \pi} > c_\alpha$$

Under the alternative β,

$$\frac{(\pi' S)^2}{\sigma_0^2 \pi' Z' Z \pi} \sim \chi^2 \left(1, \frac{\pi' Z' Z \pi \, (\beta - \beta_0)^2}{\sigma_0^2} \right)$$

Consequently, the power envelope is given by (9).

Q.E.D.

PROOF OF THEOREM 3: The derivative of the log likelihood function with respect to β evaluated at β_0 and at $\hat{\pi}$ is given by:

$$\frac{\omega_{12} \hat{\pi}' Z' (y_2 - Z\hat{\pi}) - \omega_{22} \hat{\pi}' Z' (y_1 - Z\hat{\pi}\beta_0)}{\omega_{11}\omega_{22} - \omega_{12}^2}$$

Tedious algebraic manipulations show that the latter term is equal to:

$$-\frac{\hat{\pi}' Z' (y_1 - y_2 \beta_0)}{\omega_{11} + \omega_{22}\beta_0^2 - 2\omega_{12}\beta_0}$$

Q.E.D.