

Efficient Split-Sample Anderson-Rubin Tests in IV and GMM Models

Jonathan H. Wright *

This Version: December 2008.

Abstract

In the linear IV model with possibly weak instruments, one approach to inference is the split-sample Anderson-Rubin test, discussed by Dufour and Jasiak (2001) and Kleibergen (2002). This splits the sample into two pieces; the first is used to find the optimal instruments and the second is used to check for orthogonality between the errors and the chosen linear combination of the instruments. Because the two subsamples are independent, the resulting test statistic has a χ^2 null limiting distribution. However, this test uses only part of the sample to check for correlation between structural errors and instruments, and as such it seems likely to waste power. In this paper I propose a test that takes two different split-sample Anderson-Rubin tests, in which the two subsamples are interchanged, and then combines these into a single test statistic. With strong instruments, this test statistic has a χ^2 distribution. With weak instruments, it has a nonstandard distribution, but one that can be bounded—and the bound is quite close to being χ^2 . This allows an asymptotically conservative identification-robust test to be conducted. The test is nonetheless reasonably powerful, and nearly always more powerful than the standard split-sample test. It can also be extended to a GMM framework.

* Department of Economics, Johns Hopkins University, 3400 N. Charles St., Baltimore MD 21218; e-mail: wrightj@jhu.edu; phone (410) 516-5728.

1. Introduction

Problems of weak instruments and weak identification are enormously common in applied econometric work and have motivated a search over the past 15 years for methods of inference that are reliable even when identification is weak. The profession has learned two key things from this work. First, with weak identification, the search for point estimates is ill-motivated as structural parameters are not consistently estimable; researchers should instead focus on forming testing hypotheses and forming confidence sets from the inversion of the acceptance region of pivotal test statistics. These confidence sets will have infinite expected volume if the identification is weak; but that is an appropriate statement of our knowledge under these circumstances (Dufour (1997)). Second, in the presence of overidentifying restrictions, efficiency gains are available by a judicious choice of the pivotal test statistic that effectively reduces the degrees of freedom from the number of moment conditions to the number of parameters. Reviews of the recent literature on weak instruments include Dufour (2003) and Stock, Wright and Yogo (2002).

Over half a century ago, Anderson and Rubin (1949) proposed the first approach to inference in the linear instrumental variables (IV) model that is robust to weak instruments. It remains widely used today. The idea is to test a hypothesis about the structural parameter by projecting the implied structural error term on the set of instruments. A confidence set can be formed by inverting the acceptance region of this test statistic. An analog in the context of the generalized method of moments (GMM) was proposed by Stock and Wright (2000). Under the null, the test statistic has a χ^2 limiting distribution on degrees of freedom equal to the number of instruments or moment conditions. This distribution applies regardless of the strength of identification. However, the approach wastes power in the sense that it projects the structural

error term onto the whole space of instruments, whereas more recently-developed methods instead project the structural error term onto an optimally chosen subspace of instruments.

Dufour and Jasiak (2001) and Kleibergen (2002) consider a simple yet very appealing approach to inference that is robust to weak identification and may be more powerful than the Anderson and Rubin (1949) approach. The idea is to split the sample into two. The first subsample is used to regress the right-hand-side endogenous variables onto the instruments. The second subsample is then used to test the correlation between the error term in the structural equation and the optimal instruments—using the coefficient estimates from the first subsample. Under the null hypothesis, the test statistic has a χ^2 limiting distribution on degrees of freedom equal to the number of parameters.

Appealing as this method is, it too seems to waste power in that it uses only one part of the sample for checking orthogonality of the instruments and structural error. This paper proposes a variant on the split-sample instrumental variable technique that combines two statistics. The first statistic uses the first subsample to estimate the optimal instrument coefficients and the second subsample to check if the optimal instruments and error term are correlated. The second statistic simply reverses the role of the two subsamples. Although each statistic has a marginal χ^2 null limiting distribution, the combined statistic does not necessarily have this distribution. But I can derive its distribution; with strong instruments, it has a χ^2 distribution, but with weak instruments, it has a nonstandard distribution. However, even this nonstandard distribution can be bounded—and the bound is quite close to being χ^2 . This allows an asymptotically conservative identification-robust test to be conducted. The test is nonetheless reasonably powerful. It can also be extended to a GMM framework.

The plan for the remainder of this paper is as follows. In section 2, I describe the model and the proposed test statistic. In section 3, I report some Monte-Carlo evidence on its performance. In section 4, I give some empirical applications, in which the proposed test statistic has a different and smaller acceptance region than other alternatives. Section 5 concludes.

2. The Model and the Proposed Test Statistic

Consider the standard linear IV model

$$y = X\beta + u$$

$$X = Z\pi + v$$

where y is a $T \times 1$ vector of endogenous variables, X is a $T \times p$ matrix of endogenous variables, Z is a $T \times k$ matrix of instruments, $k \geq p$, and u and v are matrices of shocks of order $T \times 1$ and $T \times p$, respectively such that the rows of $[u \ v]$ are iid with mean zero and variance-covariance matrix Ω which can be partitioned conformably as $\begin{pmatrix} \sigma_{uu}^2 & \sigma_{uv} \\ \sigma_{vu} & \Sigma_{vv} \end{pmatrix}$. I make no assumptions about the rank of the $k \times p$ matrix π and so the model is not necessarily well identified.

The objective is to test the hypothesis $\beta = \beta_0$. The Anderson-Rubin test statistic is

$$AR(\beta_0) = \frac{(y - X\beta_0)' P_Z (y - X\beta_0) / k}{(y - X\beta_0)' M_Z (y - X\beta_0) / (T - k)}$$

where, throughout, for any matrix A , $P_A = A(A'A)^{-1}A'$ and $M_A = I - P_A$. Under the null, if u is normal, this has an exact F distribution on k and $T - k$ degrees of freedom. Without assuming normality it is asymptotically $\chi^2(k)/k$ distributed.

Dufour and Jasiak (2001) and Kleibergen (2002) propose a split-sample Anderson-Rubin statistic. Suppose that there are two subsamples, with the data partitioned conformably, such that $y = (y_1', y_2')'$, $X = (X_1' \ X_2')'$ and $Z = (Z_1' \ Z_2')'$. Define the statistic

$$SS(\beta_0) = \frac{(y_2 - X_2\beta_0)' P_{Z_2\hat{\pi}_1} (y_2 - X_2\beta_0)}{(y - X\beta_0)' M_Z (y - X\beta_0) / (T - k)}$$

where $\hat{\pi}_1 = (Z_1'Z_1)^{-1}Z_1'X_1$. This is effectively the Anderson-Rubin statistic applied to the second subsample, but using a $p \times 1$ set of instruments, obtained by weighting the k original instruments by the projection coefficients from the first subsample. The test statistic is asymptotically $\chi^2(p)$ conditional on $\hat{\pi}_1$ and is independent of $\hat{\pi}_1$ and so it is unconditionally $\chi^2(p)$ distributed as well.¹

However, the split-sample Anderson-Rubin test seems to waste power, because it only uses one of the subsamples to check for correlation between the instruments and the structural error term. In this paper, I propose a variant on this method that combines two split-sample statistics that reverses the roles of the first and second subsamples. Let the first subsample cover observations $1, 2, \dots, T_1 = \lfloor T/2 \rfloor$ and observations $T_1 + 1, T_1 + 2, \dots, T$, respectively. The proposed test statistic is

$$CSS(\beta_0) = \frac{\{P_{Z_1\hat{\pi}_2}^{1/2} (y_1 - X_1\beta_0) + P_{Z_2\hat{\pi}_1}^{1/2} (y_2 - X_2\beta_0)\} \{P_{Z_1\hat{\pi}_2}^{1/2} (y_1 - X_1\beta_0) + P_{Z_2\hat{\pi}_1}^{1/2} (y_2 - X_2\beta_0)\}}{2(y - X\beta_0)' M_Z (y - X\beta_0) / (T - k)}$$

where $\hat{\pi}_2 = (Z_2'Z_2)^{-1}Z_2'X_2$. I consider the null limiting distribution of this test statistic under two alternative asymptotics. The first is the “strong instrument” case in which π is a fixed full-rank matrix. The second is the “weak instrument” case in which, following Staiger and Stock (1997),

¹ If we did *not* use disjoint subsamples, then the estimate of π would be correlated with the structural error term if the instruments are weak, as discussed, for example, in Hall, Rudbebusch and Wilcox (1996).

$\pi = CT^{-1/2}$ and so is local to zero. Of course this nests the case of completely irrelevant instruments as a special case. Theorem 1 gives the limiting distributions in these two cases.

Theorem 1: Under the null $\beta = \beta_0$, (i) with strong instruments, $CSS(\beta_0) \rightarrow_d \chi^2(p)$, while (ii) with weak instruments

$$CSS(\beta_0) \rightarrow_d \mu' \mu / 2$$

where

$$\mu = [\Sigma_{vv}^{1/2}(\lambda + \Phi_2)'(\lambda + \Phi_2)\Sigma_{vv}^{1/2}]^{-0.5}(\lambda + \Phi_2)\Sigma_{vv}^{1/2}\phi_1 + [\Sigma_{vv}^{1/2}(\lambda + \Phi_1)'(\lambda + \Phi_1)\Sigma_{vv}^{1/2}]^{-0.5}(\lambda + \Phi_1)\Sigma_{vv}^{1/2}\phi_2$$

$\lambda = Q_{zz}^{1/2}c\Sigma_{vv}^{-1/2}$, $Q_{zz} = E(Z_i Z_i')$, ϕ_1 and ϕ_2 are $k \times 1$ vectors and Φ_1 and Φ_2 are $k \times p$ matrices such that $(\phi_1', \text{vec}(\Phi_1)')$ and $(\phi_2', \text{vec}(\Phi_2)')$ are mutually independent $N(0, (I_{(p+1)} \otimes \Omega))$ vectors,

Proof: Under the strong instruments case, $\hat{\pi}_1$ and $\hat{\pi}_2$ are both consistent for π and so

$$CSS(\beta_0) = \frac{\{P_{Z_1\pi}^{1/2}(y_1 - X_1\beta_0) + P_{Z_2\pi}^{1/2}(y_2 - X_2\beta_0)\}'\{P_{Z_1\pi}^{1/2}(y_1 - X_1\beta_0) + P_{Z_2\pi}^{1/2}(y_2 - X_2\beta_0)\}}{2(y - X\beta_0)'M_Z(y - X\beta_0)/(T - k)} + o_p(1)$$

But,

$$P_{Z_1\pi}^{1/2}(y_1 - X_1\beta_0) = (\pi' Z_1' Z_1 \pi)^{-1/2} \pi' Z_1' u \rightarrow_d N(0, \sigma_u^2 I_p)$$

$$P_{Z_2\pi}^{1/2}(y_2 - X_2\beta_0) = (\pi' Z_2' Z_2 \pi)^{-1/2} \pi' Z_2' u_2 \rightarrow_d N(0, \sigma_u^2 I_p)$$

and these are mutually independent. So

$$P_{Z_1\pi}^{1/2}(y_1 - X_1\beta_0) + P_{Z_2\pi}^{1/2}(y_2 - X_2\beta_0) \rightarrow_d N(0, 2\sigma_u^2 I_p)$$

Meanwhile, $(y - X\beta_0)'M_Z(y - X\beta_0)/(T - k) \rightarrow_p \sigma_u^2$. Combining these, $CSS(\beta_0) \rightarrow_d \chi^2(p)$.

Under the weak instruments case

$$P_{Z_1\hat{\pi}_2}^{1/2}(y_1 - X_1\beta_0) = (\hat{\pi}_2' Z_1' Z_1 \hat{\pi}_2)^{-1/2} \hat{\pi}_2' Z_1' u_1 = (\hat{\pi}_2' Z_1' Z_1 \hat{\pi}_2)^{-1/2} \hat{\pi}_2' (Z_1' Z_1)^{1/2} (Z_1' Z_1)^{-1/2} Z_1' u_1$$

$$(Z_1'Z_1)^{1/2} \hat{\pi}_2 = (Z_1'Z_1)^{1/2} \pi + (Z_1'Z_1)^{1/2} (Z_2'Z_2)^{-1} Z_2'v_2 = (T_1^{-1}Z_1'Z_1)^{1/2} \left(\frac{T}{T_1}\right)^{1/2} c + (T_1^{-1}Z_1'Z_1)^{1/2} (T_1^{-1}Z_2'Z_2)^{-1} T_1^{-1/2} Z_2'v_2$$

$$(Z_1'Z_1)^{1/2} \hat{\pi}_2 \rightarrow_d \lambda \Sigma_{vv}^{1/2} + \Phi_2 \Sigma_{vv}^{1/2}$$

$$(Z_1'Z_1)^{-1/2} Z_1'u_1 \rightarrow_d \sigma_{uu} \phi_1$$

$$\therefore P_{Z_1\hat{\pi}_2}^{1/2} (y_1 - X_1\beta_0) \rightarrow_d [\Sigma_{vv}^{1/2} (\lambda + \Phi_2)' (\lambda + \Phi_2) \Sigma_{vv}^{1/2}]^{-1/2} (\lambda + \Phi_2) \Sigma_{vv}^{1/2} \sigma_{uu} \phi_1$$

Similarly

$$P_{Z_2\hat{\pi}_1}^{1/2} (y_2 - X_2\beta_0) \rightarrow_d [\Sigma_{vv}^{1/2} (\lambda + \Phi_1)' (\lambda + \Phi_1) \Sigma_{vv}^{1/2}]^{-1/2} (\lambda + \Phi_1) \Sigma_{vv}^{1/2} \sigma_{uu} \phi_2$$

and these are *not* independent. Combining these

$$P_{Z_1\hat{\pi}_2}^{1/2} (y_1 - X_1\beta_0) + P_{Z_2\hat{\pi}_1}^{1/2} (y_2 - X_2\beta_0) \rightarrow_d \sigma_{uu} \mu$$

Meanwhile, $(y - X\beta_0)' M_Z (y - X\beta_0) / (T - k) \rightarrow_p \sigma_{uu}^2$. Combining these,

$$CSS(\beta_0) \rightarrow_d \mu' \mu / 2$$

as required.

The expression in the second part of Theorem 1 simplifies somewhat in the case $p = 1$. In this case, under the weak instrument asymptotics, $CSS(\beta_0) \rightarrow_d \mu' \mu / 2$ where

$$\mu = [(\lambda + \Phi_2)' (\lambda + \Phi_2)]^{-0.5} (\lambda + \Phi_2) \phi_1 + [(\lambda + \Phi_1)' (\lambda + \Phi_1)]^{-0.5} (\lambda + \Phi_1) \phi_2 \quad (1)$$

Under the weak instrument asymptotics, the limiting distribution in (1) depends on λ , ρ and k .

The dependence on ρ and k is not problematic, but the dependence on λ is because this parameter is not consistently estimable. The distribution can however be simulated for any choice of λ , ρ and k . Let $F(\lambda, \rho, k; \alpha)$ denote the upper 100α percentile of $\mu' \mu / 2$. The test that I propose in this paper rejects if

$$CSS(\beta_0) > \sup_{\lambda} F(\lambda, \hat{\rho}, k; \alpha) \quad (2)$$

where $\hat{\rho}$ is any consistent estimator of ρ . Clearly this is a conservative test with asymptotic size bounded above by α , uniformly in λ . The parameter ρ can be consistently estimated as the correlation between $y - X\beta_0$ and the residuals in a regression of X on Z .

For the case $k = 2$, I simulated $F(\lambda, \rho, k; 0.95)$ for a grid of values of λ for two different values of ρ . The percentiles are plotted against λ in Figure 1, and are maximized at $\lambda = 0$, while for larger values of λ , the percentiles are close to 3.84, the upper 5th percentile of a $\chi^2(1)$ distribution, which is to be expected from the first part of Theorem 1, since the higher is λ , the stronger are the instruments. Using sparser grids of values of λ for $k=3$ and 4, I also find that $F(\lambda, \rho, k; 0.95)$ is maximized at $\lambda = 0$.

Table 1 gives a lookup-table of the 5 percent critical values, $\sup_{\lambda} F(\lambda, \rho, k; 0.05)$ for different values of ρ and k . In constructing this table, for $k > 4$, I assume that $\sup_{\lambda} F(\lambda, \rho, k; 0.05) = F(0, \rho, k; 0.05)$.

The usefulness of the test in (2) depends on the sensitivity of $F(\lambda, \rho, k; \alpha)$ to λ . If $F(\lambda, \rho, k; \alpha)$ did not depend on λ , then the test would not be conservative at all. On the other hand, if it is highly sensitive to λ , then there must be regions in the parameter space in which it is highly conservative and hence inefficient. Numerically, as shown in Table 1, it seems that the $\sup_{\lambda} F(\lambda, \rho, k; \alpha)$ is fairly close to the corresponding percentile of a $\chi^2(1)$ distribution, implying that the test will not be too conservative and so will not waste too much power.² This is particularly true if ρ is not too big. For example, with 10 instruments and $\rho = 0.5$, the 5

² A natural variant on the test proposed here is to form a $100(1-\alpha)$ percent confidence set for λ and then to compare the test statistic with the sup of $F(\lambda, \rho, k; \alpha)$ over all λ in this confidence set. By the Bonferroni inequality, the size of this test is at most $100(1-2\alpha)$ percent. But this would considerably more computationally costly, and not necessarily more powerful.

percent critical value is 4.67 in Table 1, versus 3.84 for the $\chi^2(1)$ distribution. Meanwhile, the Anderson-Rubin statistic which has 10 degrees of freedom would have a critical value of 18.31.

The fact that $P_{Z_1\hat{\sigma}_2}^{1/2}(y_1 - X_1\beta_0)$ and $P_{Z_2\hat{\sigma}_1}^{1/2}(y_2 - X_2\beta_0)$ both have marginal normal distributions, means that the null limiting distribution of the proposed test statistic has to be bounded by twice a $\chi^2(p)$ distribution. But the critical values in Table 1 are all a good bit smaller than 7.68 (which is twice the upper 5th percentile of a $\chi^2(1)$ distribution) so this bound is not as sharp as can be attained.

Although the proposed test statistic is not pivotal with weak instruments, it is *boundedly pivotal*. There are indeed many statistics testing the hypothesis $\beta = \beta_0$ that have non-pivotal distributions that are a function of the nuisance parameter λ but that can however be bounded, implying an asymptotically conservative test. The usefulness of such a testing procedure depends crucially on the extent to which the distribution of the test statistic depends on λ —if it is very sensitive to λ , then the test cannot control size uniformly in λ without being very conservative. The merit of the proposed test statistic that $F(\lambda, \rho, k; \alpha)$ does not seem very dependent on λ .

2.1 GMM version

An attractive feature of the proposed test is that it is applicable more generally, in a GMM model that could have heteroskedastic and/or autocorrelated shocks. Consider the model $E(f(Y_t, \theta_0)) = 0$ where Y_t is a vector of data, θ is a $p \times 1$ vector of parameters, θ_0 is the true parameter value and $f()$ is a $k \times 1$ vector of moment conditions. I make the following “high level” assumption:

$$T^{-1/2} \sum_{t=1}^{\lfloor Tr \rfloor} \begin{pmatrix} f(Y_t, \theta_0) \\ \text{vec} \left(\frac{\partial f(Y_t, \theta_0)}{\partial \theta} - J(\theta_0) \right) \end{pmatrix} \rightarrow_d G^{1/2} B(r) \quad (3)$$

where $B(r)$ is a standard Brownian motion and $J(\theta) = E \left(\frac{\partial f(Y_t, \theta_0)}{\partial \theta} \right)$. Let G be partitioned as

$$\begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} \text{ where } G_{11} \text{ is of order } k \times k. \text{ Partition the sample into the first } T_1 = \lfloor T/2 \rfloor$$

observations and the remaining observations. Define $f_1(\theta) = \sum_{t=1}^{T_1} f(Y_t, \theta)$, $f_2(\theta) = \sum_{t=T_1+1}^T f(Y_t, \theta)$,

$$\hat{D}_{(1)} = \frac{1}{T_1} \sum_{t=1}^{T_1} \frac{\partial f(Y_t, \theta_0)}{\partial \theta}, \quad \hat{D}_{(2)} = \frac{1}{T_1} \sum_{t=T_1+1}^T \frac{\partial f(Y_t, \theta_0)}{\partial \theta}, \quad \hat{G}_{11}^{(1)} = \frac{1}{T_1} \sum_{t=1}^{T_1} f(Y_t, \theta_0) f(Y_t, \theta_0)' \quad \text{and}$$

$$\hat{G}_{11}^{(2)} = \frac{1}{T_1} \sum_{t=T_1+1}^T f(Y_t, \theta_0) f(Y_t, \theta_0)'. \text{ Finally, let } \phi_1, \phi_2, \text{vec}(\Phi_1) \text{ and } \text{vec}(\Phi_2) \text{ denote the limiting}$$

distributions of $T^{1/2} f_1(\theta_0)$, $T^{1/2} f_2(\theta_0)$, $T^{1/2} (\hat{D}_{(1)} - J(\theta_0))$ and $T^{1/2} (\hat{D}_{(2)} - J(\theta_0))$, respectively,

implied by the assumption in equation (3). Where appropriate, the estimates of G can be

replaced by their heteroskedasticity-and-autocorrelation robust counterparts. Note that the

assumption in equation (3) implies that the two subsamples are asymptotically independent.

For the GMM case, the analog of the Anderson-Rubin test, discussed in Stock and Wright (2000) is

$$(\sum_{t=1}^T f(Y_t, \theta_0))' [T^{-1} \sum_{t=1}^T f(Y_t, \theta_0) f(Y_t, \theta_0)']^{-1} (\sum_{t=1}^T f(Y_t, \theta_0)) / T \quad (4)$$

which has a $\chi^2(k)$ null limiting distribution. The analog of the split-sample Anderson-Rubin statistic is

$$SS(\theta_0) = f_2(\theta_0)' \hat{G}_{11}^{(1)-1/2} P_{\hat{G}_{11}^{(1)-1/2} \hat{D}_{(1)}} \hat{G}_{11}^{(1)-1/2} f_2(\theta_0) / T_1 \quad (5)$$

which has a $\chi^2(p)$ null limiting distribution. Correspondingly the proposed new test statistic in

the GMM case is

$$\begin{aligned}
CSS(\theta_0) &= \{P_{\hat{G}_{11}^{(2)-1/2}\hat{D}_{(2)}}^{1/2} \hat{G}_{11}^{(2)-1/2} f_1(\theta) + P_{\hat{G}_{11}^{(1)-1/2}\hat{D}_{(1)}}^{1/2} \hat{G}_{11}^{(1)-1/2} f_2(\theta)\}' \\
&\{P_{\hat{G}_{11}^{(2)-1/2}\hat{D}_{(2)}}^{1/2} \hat{G}_{11}^{(2)-1/2} f_1(\theta) + P_{\hat{G}_{11}^{(1)-1/2}\hat{D}_{(1)}}^{1/2} \hat{G}_{11}^{(1)-1/2} f_2(\theta)\} / T
\end{aligned} \tag{6}$$

Again, I consider the null limiting distribution of this test statistic under two alternative asymptotics. The first is the ‘‘strong instrument’’ case in which the Jacobian $J(\theta_0)$ is a fixed full-rank matrix. The second is the ‘‘weak identification’’ case, in which, following Stock and Wright, $J(\theta_0) = C(\theta_0)T^{-1/2}$ and so is local to zero. Theorem 2 gives the limiting distribution in these two cases.

Theorem 2: Under the null $\theta = \theta_0$, (i) with strong instruments, $CSS(\theta_0) \rightarrow_d \chi^2(p)$, while (ii) with weak instruments $CSS(\theta_0) \rightarrow_d \mu' \mu / 2$ where

$$\begin{aligned}
\mu &= [(\Lambda(\theta_0) + \Phi_1)' G_{11}^{-1} (\Lambda(\theta_0) + \Phi_1)]^{-1/2} (\Lambda(\theta_0) + \Phi_1) G_{11}^{-1} \phi_2 \\
&+ [(\Lambda(\theta_0) + \Phi_2)' G_{11}^{-1} (\Lambda(\theta_0) + \Phi_2)]^{-1/2} (\Lambda(\theta_0) + \Phi_2) G_{11}^{-1} \phi_1
\end{aligned}$$

where $\Lambda(\theta_0) = C(\theta_0)G_{11}^{-1/2}$.

Proof: Under the strong instruments case, $\hat{D}_{(1)}$ and $\hat{D}_{(2)}$ are both consistent for $J(\theta_0)$ while $\hat{G}_{11}^{(1)}$

and $\hat{G}_{11}^{(2)}$ are both consistent for G_{11} and so

$$CSS(\theta_0) = \{P_{G_{11}^{-1/2}J(\theta_0)}^{1/2} G_{11}^{-1/2} (f_1(\theta) + f_2(\theta))\}' \{P_{G_{11}^{-1/2}J(\theta_0)}^{1/2} G_{11}^{-1/2} (f_1(\theta) + f_2(\theta))\} / 2T_1 + o_p(1)$$

But

$$T_1^{-1/2} P_{G_{11}J(\theta_0)}^{1/2} f_2(\theta) = (J(\theta_0)' G_{11}^{-1} J(\theta_0))^{-1/2} J(\theta_0)' G_{11}^{-1} T_1^{-1/2} f_2(\theta) \rightarrow_d N(0, I_p)$$

since $T_1^{-1/2} f_2(\theta) \rightarrow_d N(0, G_{11})$. Similarly, $T_1^{-1/2} P_{G_{11}J(\theta_0)}^{1/2} f_1(\theta) \rightarrow_d N(0, I_p)$. Hence

$$T_1^{-1/2} \{P_{G_{11}^{-1/2}J(\theta_0)}^{1/2} G_{11}^{-1/2} (f_1(\theta) + f_2(\theta))\} \rightarrow_d N(0, 2I_p)$$

and so $CSS(\theta_0) \rightarrow_d \chi^2(p)$

Under the weak identification case

$$P_{\hat{G}_{11}^{(1)-1/2}\hat{D}_{(1)}}^{1/2} = (\hat{D}_{(1)}' \hat{G}_{11}^{(1)-1} \hat{D}_{(1)})^{-1/2} \hat{D}_{(1)}' \hat{G}_{11}^{(1)-1/2}$$

But $T_1^{1/2} \hat{D}_{(1)} \rightarrow_d C(\theta_0) + \Phi_1$ and $\hat{G}_{11}^{(1)} \rightarrow_p G_{11}$. Hence,

$$P_{\hat{G}_{11}^{(1)-1/2}\hat{D}_{(1)}}^{1/2} \rightarrow_d [(\Lambda(\theta_0) + \Phi_1 G_{11}^{-1/2})' (\Lambda(\theta_0) + \Phi_1 G_{11}^{-1/2})]^{-1/2} (\Lambda(\theta_0) + \Phi_1 G_{11}^{-1/2})$$

Meanwhile, $T^{-1/2} f_1 \rightarrow_d \phi_1$. Hence $CSS(\theta_0) \rightarrow_d \mu' \mu / 2$ as required.

Letting $F(\Lambda(\theta_0), G, k; \alpha)$ denote the upper α percentile of $\mu' \mu / 2$, the proposed test rejects if

$$CSS(\theta_0) > \sup_{\Lambda(\theta_0)} F(\Lambda(\theta_0), \hat{G}, k; \alpha) \quad (7)$$

where \hat{G} is any consistent estimator of G .

3. Monte-Carlo Evidence

In this section, I present some Monte-Carlo evidence on the small-sample properties of the proposed test and other alternatives in the context of the linear IV model. The design follows Staiger and Stock (1997), Kleibergen (2002) and others. There is a single endogenous explanatory variable. The data generating process is

$$y_i = \beta x_i + u_i$$

$$x_i = \pi' z_i + v_i$$

where (u_i, v_i) is iid $N(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$, z_i is a $k \times 1$ vector that is independent of these errors and is

$$N(0, I_k), \pi = (\sqrt{\frac{R^2}{1-R^2}}, 0, 0, \dots, 0)' \text{ and } R^2 \text{ is the population R-squared from the regression of } x_i$$

on z_i . I test the hypothesis that $\beta = 0$ where the true value of this parameter ranges from -1 to

+1. Figure 2 shows the power curves with a sample size $T=100$ and some values of k and ρ using the following tests: the Anderson-Rubin, split-sample Anderson-Rubin (using 25 percent of the data to estimate the π matrix and the balance for checking correlation of instruments and the error term)³, the new proposed test given by equation (2), and Kleibergen's test.

The proposed test is conservative, and as such it is to be expected that it has lower power than the Anderson-Rubin and split-sample Anderson-Rubin tests when the hypothesized and true values of β are very close ($\beta \approx 0$). But for most values of β , the proposed test is generally more powerful than the Anderson-Rubin and the split-sample Anderson-Rubin tests. The ordinary split-sample test wastes a lot of power in some cases. The improvement relative to the Anderson-Rubin test is typically modest. However, all three tests are typically less powerful than the test of Kleibergen (2002). The power properties of the Kleibergen test can be quirky⁴ (a hint of which can be seen in Figure 2), in the sense that power is not always monotonic in $|\beta|$.

3.1 Non-normal shocks

I also did a version of the Monte-Carlo simulation in which the data-generating process is exactly as before except that the shocks are non-normal, such that $u_t = (\xi_{u,t}^2 - 1) / \sqrt{2}$,

$v_t = (\xi_{v,t}^2 - 1) / \sqrt{2}$ and $(\xi_{u,t}, \xi_{v,t})'$ is iid $N(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$. The power functions for this case are

shown in Figure 3. As in the case with normal errors, the proposed test is generally more powerful than the Anderson-Rubin and the split-sample Anderson-Rubin tests, while being less powerful than the Kleibergen (2002) test.

³ This choice of the sample-split was found by Kleibergen (2002) to maximize power in most simulations.

⁴ Kleibergen (2005) points out that the LM test is not consistent against inflection points of the continuous-updating GMM objective function, and proposes a device that effectively combines this test statistic with the Anderson-Rubin statistic to obtain power against these alternatives.

3.2 Robustness to Omitted Instruments

Dufour and Taamouti (2007) point out that an appealing feature of the Anderson-Rubin test is that it is robust to omitting relevant instruments. Meanwhile the tests of Kleibergen (2002) and Moreira (2003) are not—they will reject too often, even asymptotically, when relevant instruments are omitted. This is because these methods are exploiting the joint distribution of the moment condition and the scores, and this is not robust to model misspecification. This seems a worrying problem, given that it is rarely, if ever, possible to be confident that we have obtained all possible instruments. The split-sample Anderson-Rubin test is also robust to weak instruments (Dufour (2008)); the proposed test is not, although it is still bounded by twice a $\chi^2(p)$ distribution, even with omitted instruments.

To compare the performance of different tests with omitted instruments, I considered the following variant on the Monte-Carlo experiment. The data generating process follows Doko and Dufour (2007) and is

$$y_t = \beta x_t + u_t$$

$$x_t = \pi' z_t + v_t$$

where (u_t, v_t) is iid $N(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$, z_t is a $(k+1) \times 1$ vector that is independent of these errors and

is $N(0, I_{k+1})$, $\pi = (\sqrt{\frac{R^2}{1-R^2}}, 0, 0, \dots, \gamma)'$. The $k+1$ th instrument is a relevant one, but the researcher

does not know about this instrument and instead conducts inference using only the first k instruments.

Table 2 shows the effective sizes of the different tests in this case, for a number of choices of ρ , R^2 and k , setting $\gamma = 3$ where the sample size is $T = 100$. The nominal size of all tests is 5 percent. The Kleibergen test has severe size distortions in this case, especially when the degree of overidentification is large. The effective size of the test can exceed 50 percent, when the number of instruments is large, ρ is high and the instruments are weak. Doko and Dufour (2007) show even more severe size distortions for the Kleibergen test with other parameter values. The conditional likelihood ratio test also rejects too often, but the size distortions are much milder. The Anderson-Rubin, ordinary split-sample Anderson-Rubin and the proposed test all have effective sizes that are no greater than 6 percent, and so are all in practice quite robust to omitted instruments.

4. Empirical Applications.

In this section, I report the results of applying the proposed test in three different applications. In each case, I use the inversion of the acceptance region of the proposed test—along with some alternatives—to form confidence sets for structural parameters of interest.

4.1 Returns to Schooling

The problem of estimating the returns to schooling sparked much of the recent resurgence in the interest on small-sample properties of IV estimators when instruments are weak. Angrist and Krueger (1991) considered the regression of log wages on years of schooling using quarter-of-birth dummies as instruments. Angrist and Krueger (1995) did this too, but with split-sample point estimates. Neither paper considered weak-instrument robust tests or confidence sets.

I consider the regression

$$\log(\text{wage}_i) = \alpha + \beta \text{school}_i + \gamma' x_i + v_i \quad (8)$$

where wage_i and school_i denote the weekly wage and years of education of the i th individual and x_i denotes the included exogenous variables—year-of-birth and place-of-birth dummies. The instruments for years of education are the quarter-of-birth dummies interacted with year-of-birth dummies. The data are for men born in 1930-1939, in the 1980 Census, so that there is a total of 30 instrumental variables.⁵

I form confidence intervals for the structural parameter β by inverting the acceptance region of four different tests—the Anderson-Rubin, Kleibergen, split-sample Anderson-Rubin and the proposed test. The confidence intervals are shown in Table 3. The Anderson-Rubin confidence interval turns out to be empty; the other three confidence intervals are quite wide. However, the proposed confidence interval is shorter than either the Kleibergen or conventional split-sample Anderson-Rubin confidence interval, and implies that the returns to an extra year of schooling are at least 6 percent.

4.2 The Consumption CAPM Euler Equation

This application is a nonlinear GMM problem; inference on the parameters in the familiar moment condition implied by the Euler equation CAPM

$$E\left([\delta \begin{pmatrix} R_{t+1} \\ R_{t+1}^{RF} \end{pmatrix} (C_{t+1} / C_t)^{-\gamma} - 1] \otimes Z_t\right) = 0 \quad (9)$$

where δ is the discount factor, γ is the coefficient of risk-aversion, C_t is real per-capita consumption, R_{t+1} is the real stock return, R_{t+1}^{RF} is the real short-term interest rate and

⁵ I am grateful to Josh Angrist for providing me with these data. This being a cross-sectional application, there is no natural ordering of the data. In performing the split-sample tests, I ordered the data as in the dataset on Josh Angrist's website.

$Z_t = (1, C_t / C_{t-1}, R_t, R_t^{RF})'$ is a vector of instruments in the information set at time t . Thus the number of parameters is 2, while the number of moment conditions is 8. For this application, I use U.S. annual data from Campbell and Shiller (1987), updated to cover the years 1889-2004.

Figure 4 shows four alternative confidence sets for the parameters δ and γ in equation (9). These are:

- (i) The S-set of Stock and Wright (2000), formed by inverting the acceptance region of the test statistic in (4),
- (ii) The GMM confidence set proposed by Kleibergen (2005),
- (iii) The GMM version of the split-sample S-set, based on inverting the acceptance region of the test statistic in (5), and
- (iv) The GMM version of the proposed confidence set, based on equation (7).

The first three confidence sets include only extremely high values of the risk-aversion parameter above about 40. The proposed confidence set has a different shape and indicates a value of the coefficient of risk-aversion around 20, but extremely high values of γ are excluded from this confidence set.

4.3 The New Keynesian Phillips Curve

Estimation of forward-looking macroeconomic models imposing rational expectations is commonly implemented by instrumental variables methods, assuming rational expectations. The hybrid New-Keynesian Phillips curve is a leading example, in a literature starting with the paper of Galí and Gertler (1999), and discussed recently with special attention to issues of weak instruments by Kleibergen and Mavroeidis (2008). The model is

$$\pi_t = \alpha + \gamma_f E_t(\pi_{t+1}) + \gamma_b \pi_{t-1} + \lambda s_t + \varepsilon_t \quad (10)$$

where π_t denotes inflation and s_t is a measure of marginal cost. The pure New-Keynesian Phillips curve would have only expectations of inflation on the right-hand-side ($\gamma_b = 0$), but it is common to include lagged inflation to allow for the possibility of some backward-looking price-setters. Replacing the expectation with its realized value, assuming rational expectations, and imposing the constraint that $\gamma_f + \gamma_b = 1$, equation (10) can be rewritten as

$$\pi_t - \pi_{t-1} = \alpha + \gamma_f (\pi_{t+1} - \pi_{t-1}) + \lambda s_t + u_t \quad (11)$$

where $E(z_t u_t) = 0$ for any z_t in the information set at time t . For inflation data, I used the quarterly real GDP deflator; for marginal cost I used the labor share, scaled as in Gali and Gertler (1999). The sample period is 1960Q1 to 2007Q4. I constructed confidence sets for the parameters $(\gamma_f, \lambda)'$ in (11) by the same four methods as in the previous application, using a total of 8 instruments: $s_{t-1}, s_{t-2}, s_{t-3}, s_{t-4}, \Delta\pi_{t-1}, \Delta\pi_{t-2}, \Delta\pi_{t-3}$ and $\Delta\pi_{t-4}$. The results are shown in Figure 5. Although the model is linear, I use the GMM formulation of the test statistics so as to allow for heteroskedasticity and autocorrelation in the errors for equation (11), using a Newey-West estimator with lag length 4 for the estimation of G in equation (3).

In this application, the ordinary split-sample and Kleibergen (2005) confidence sets both have quite odd shapes and go to the edge of the parameter space considered. The S-set is much smaller, but the proposed confidence set has the smallest volume of all. The proposed confidence set includes values of γ_f that span from about 0.5 to a little over 1, indicating that the pure New-Keynesian Phillips curve model is not rejected. For λ , some of the included values are negative, which is inconsistent with economic theory, as it seems unreasonable to think that greater economic slack would be associated with less inflation. However, the

confidence set also includes positive values of λ , but only very small ones, consistent with the view that economic slack has little or no effect on inflation.

5. Conclusion

Split-sample methods are potentially attractive approaches to inference that are robust to weak instruments and weak identification in IV and GMM models. However, they waste power by using only part of the sample to check for correlation between structural errors and instruments. In this paper, I have proposed a test that effectively takes two different split-sample Anderson-Rubin tests, in which the two subsamples are interchanged, and then combines these into a single test statistic. With strong instruments, this test statistic has a pivotal χ^2 distribution on degrees of freedom equal to the number of parameters. With weak instruments, it has a boundedly pivotal distribution. This allows an asymptotically conservative test to be conducted. The test is however generally more powerful than either the Anderson-Rubin or the conventional split-sample Anderson-Rubin tests. The test is also applicable in a GMM model with general heteroskedasticity and autocorrelation. A researcher who wishes to use the conventional split-sample Anderson-Rubin test has nothing to lose, and something to gain, by instead using the test proposed in this paper.

References

- Anderson, Theodore W. and Herman Rubin (1949): Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations, *Annals of Mathematical Statistics*, 20, pp.46-63.
- Campbell, John Y. and Robert J. Shiller (1987): Cointegration and Tests of Present Value Models, *Journal of Political Economy*, 95, pp.1062-1088.
- Doko, Firmin and Jean-Marie Dufour (2007): Impact of Instrument Endogeneity on Some Test Statistics, working paper.
- Dufour, Jean-Marie (1997): Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models, *Econometrica*, 65, pp.1365-1388.
- Dufour, Jean-Marie (2003): Identification, Weak Instruments and Statistical Inference in Econometrics, *Canadian Journal of Economics*, 36, pp.767-808.
- Dufour, Jean-Marie (2008): Discussion of "Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve" by Frank Kleibergen and Sophocles Mavroeidis, *Journal of Business and Economic Statistics*, forthcoming.
- Dufour, Jean-Marie and Joanna Jasiak (2001): Finite Sample Limited Information Inference Methods for Structural Equations and Models with Generated Regressors, *International Economic Review*, 42, pp.815-843.
- Dufour, Jean-Marie and Mohamed Taamouti (2007): Further Results on Projection-based Inference in IV regressions with Weak, Collinear or Missing Instruments, *Journal of Econometrics*, 139, pp.133-153.
- Gali, Jordi and Mark Gertler (1999): Inflation Dynamics: A Structural Econometric Analysis, *Journal of Monetary Economics*, 44, pp.195-222.
- Hall, Alastair R., Glenn D. Rudebusch and David W. Wilcox (1996): Judging Instrument Relevance in Instrumental Variables Estimation, *International Economic Review*, 37, pp.283-289.
- Kleibergen, Frank (2002): Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression, *Econometrica*, 70, pp.1781-1803.
- Kleibergen, Frank (2005): Testing Parameters in GMM Without Assuming that they are Identified, *Econometrica*, 73, pp.1103-1124.

Kleibergen, Frank and Sophocles Mavroeidis (2008): Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve, *Journal of Business and Economic Statistics*, forthcoming.

Moreira, Marcelo (2003): A Conditional Likelihood Ratio Test for Structural Models, *Econometrica*, 71, pp.1027-1048.

Staiger, Douglas and James H. Stock (1997): Instrumental Variables Regression with Weak Instruments, *Econometrica*, 65, pp.557-586.

Stock, James H., Jonathan H. Wright and Motohiro Yogo (2002): A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments, *Journal of Business and Economic Statistics*, 20, pp.518-529.

Stock, James H. and Jonathan H. Wright (2000): GMM with Weak Identification, *Econometrica*, 68, pp.1055-1096.

Table 1: 5 percent critical values for the proposed test

	k=2	k=3	k=4	k=5	k=10	k=15	k=20	k=25
$\rho=0$	3.84	3.84	3.84	3.84	3.84	3.84	3.84	3.84
$\rho=0.1$	3.87	3.88	3.90	3.97	3.88	3.87	3.82	3.87
$\rho=0.2$	3.98	3.97	3.93	4.00	3.99	4.00	3.96	3.98
$\rho=0.3$	4.13	4.15	4.08	4.19	4.16	4.21	4.21	4.17
$\rho=0.4$	4.31	4.40	4.29	4.39	4.37	4.47	4.50	4.46
$\rho=0.5$	4.58	4.64	4.58	4.68	4.67	4.83	4.82	4.79
$\rho=0.6$	4.82	4.94	4.89	5.02	5.07	5.29	5.26	5.18
$\rho=0.7$	5.14	5.32	5.27	5.48	5.46	5.79	5.81	5.66
$\rho=0.8$	5.39	5.71	5.64	5.94	5.98	6.31	6.39	6.18
$\rho=0.9$	5.62	6.03	6.16	6.40	6.61	6.90	7.00	6.75

Notes: This table gives 5 percent critical values for the proposed test in the linear IV model with $p=1$. They are constructed by simulation as $\sup_{\lambda} F(\lambda, \rho, k; 0.05)$, using a grid of values of λ for $k = 2, 3, 4$. In each of these cases, the critical value was maximized at $\lambda = 0$. For higher values of k , I simply assume that the critical values is maximized at $\lambda = 0$.

Table 2: Empirical Sizes for Alternative Tests in the Presence of Missing Instruments

			AR	K	CLR	SS	CSS
$\rho=0.5$	$R^2=0.01$	k=5	4.9	9.4	6.0	4.9	5.0
		k=10	4.6	16.1	6.5	4.1	4.7
		k=20	5.0	29.6	8.9	2.6	4.8
		k=30	4.9	40.6	10.8	--	4.5
	$R^2=0.1$	k=5	4.9	6.3	6.2	5.0	4.7
		k=10	4.6	10.6	6.3	4.2	4.8
		k=20	5.0	22.4	9.2	2.7	4.9
		k=30	4.9	33.9	10.8	--	4.6
	$R^2=0.3$	k=5	4.9	5.5	6.0	5.0	4.3
		k=10	4.6	7.4	6.5	4.2	5.0
		k=20	5.0	14.3	9.2	2.8	5.0
		k=30	4.9	24.3	10.8	--	4.8
$\rho=0.9$	$R^2=0.01$	k=5	4.9	12.7	5.9	5.0	5.6
		k=10	4.6	26.6	6.5	4.0	5.0
		k=20	5.0	50.0	8.7	2.6	4.6
		k=30	4.9	62.3	10.6	--	4.3
	$R^2=0.1$	k=5	4.9	6.5	6.2	4.8	5.0
		k=10	4.6	11.7	6.4	4.0	5.1
		k=20	5.0	29.2	8.8	2.7	4.8
		k=30	4.9	45.8	10.7	--	4.5
	$R^2=0.3$	k=5	4.9	5.5	6.0	5.0	4.6
		k=10	4.6	7.3	6.6	4.1	5.0
		k=20	5.0	15.5	9.2	2.6	4.8
		k=30	4.9	27.2	10.8	--	4.5

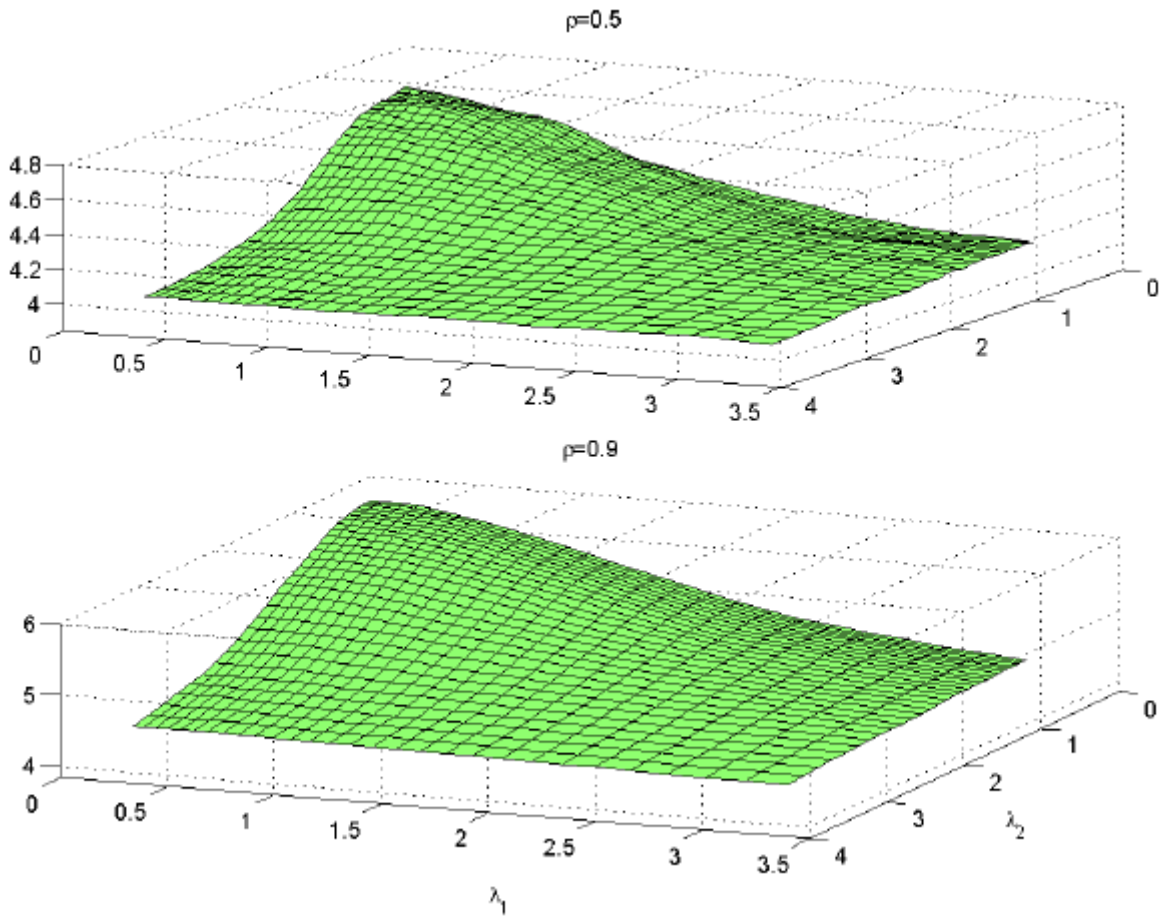
Notes: This table gives the effective size of the following tests: Anderson-Rubin (AR), the Kleibergen (2002) test (K), the conditional likelihood-ratio test (CLR), the split-sample Anderson-Rubin test (SS) and the proposed combined split-sample Anderson-Rubin test (CSS). The data generating process is as described in subsection 3.2; one instrument is missing from the set of instruments that are used for any of these tests. The sample size is 100 in all cases, and the nominal size of the test is 5 percent. The Anderson-Rubin test uses the F critical values and so is exact; the deviations of effective size from 5 percent reflect sampling error. The ordinary split-sample Anderson-Rubin test uses the first 25 percent of the data to form the optimal instruments, and as such is not applicable when the number of instruments is 30.

Table 3: Alternative Confidence Sets for the Slope Coefficient in an IV regression of log wages on years of schooling

	Lower Bound	Upper Bound
Anderson-Rubin		Empty Set
Kleibergen	0.026	0.184
Split-Sample Anderson-Rubin	0.028	0.160
Proposed	0.060	0.160

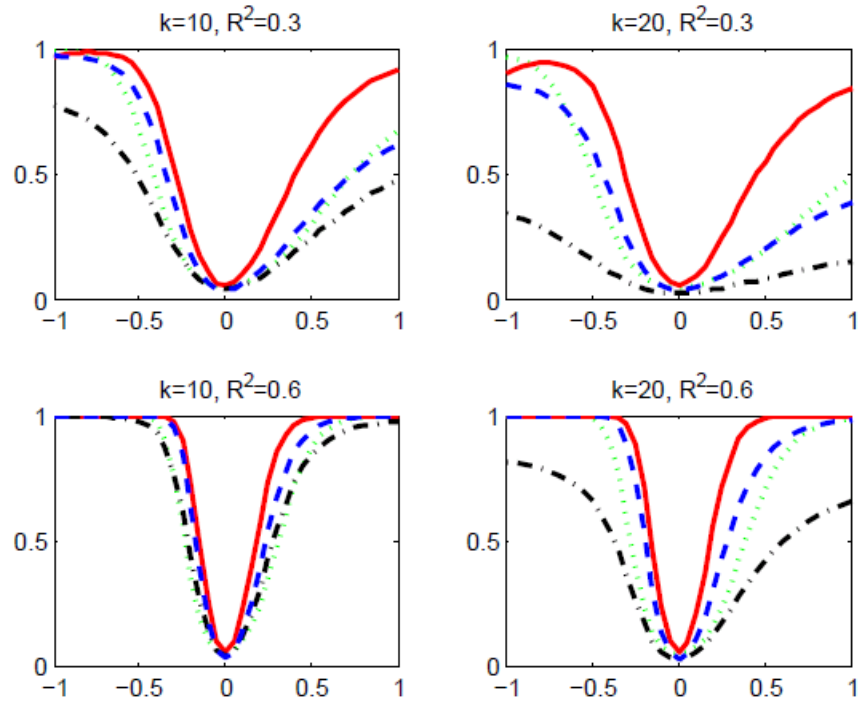
Notes: This table gives confidence sets for the slope coefficient in the IV regression of log wages on years of schooling given by equation (8). Included exogenous variables are year-of-birth and place-of-birth dummies; there are 30 instruments given by the interactions between year-of-birth and quarter-of-birth dummies. The sample size is 329,509. The data are for men born from 1930 to 1939, as recorded in the 1980 census. The confidence sets have 95 percent nominal coverage, and are given by the inversion of the acceptance regions of 5 percent tests using the different test statistics.

Figure 1: Plot of $F(\lambda, \rho, 2; 0.05)$ against λ

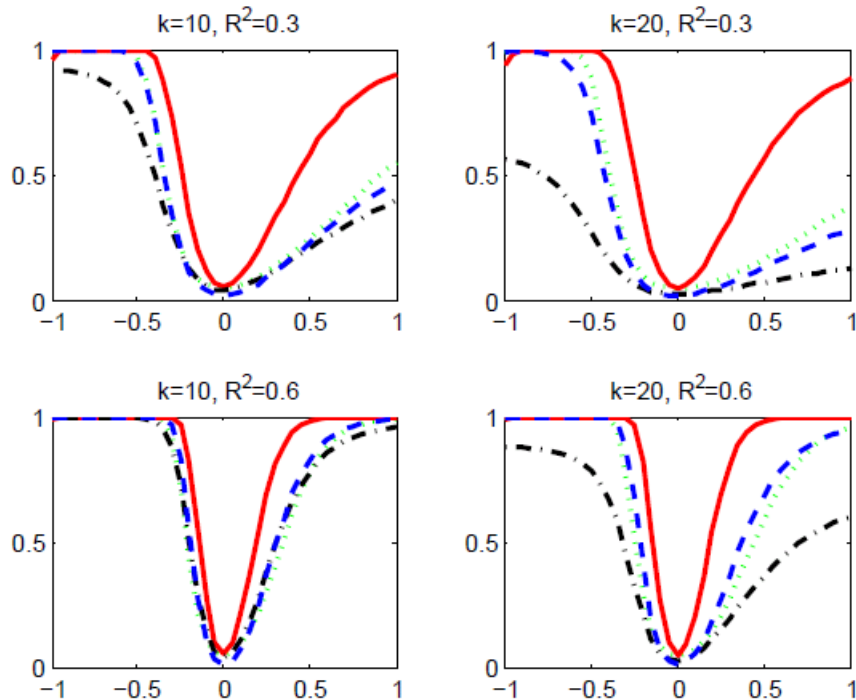


Notes: This figure plots the simulated upper 5th percentile of the weak-instrument limiting distribution in Theorem 1 against λ in the case $k=2$ and $\rho=0.5$ (upper panel) or $\rho=0.9$ (lower panel). The critical values were obtained by simulation methods, with 100,000 simulations in each case.

Figure 2. Power curves for alternative test statistics
Panel A: $\rho=0.5$

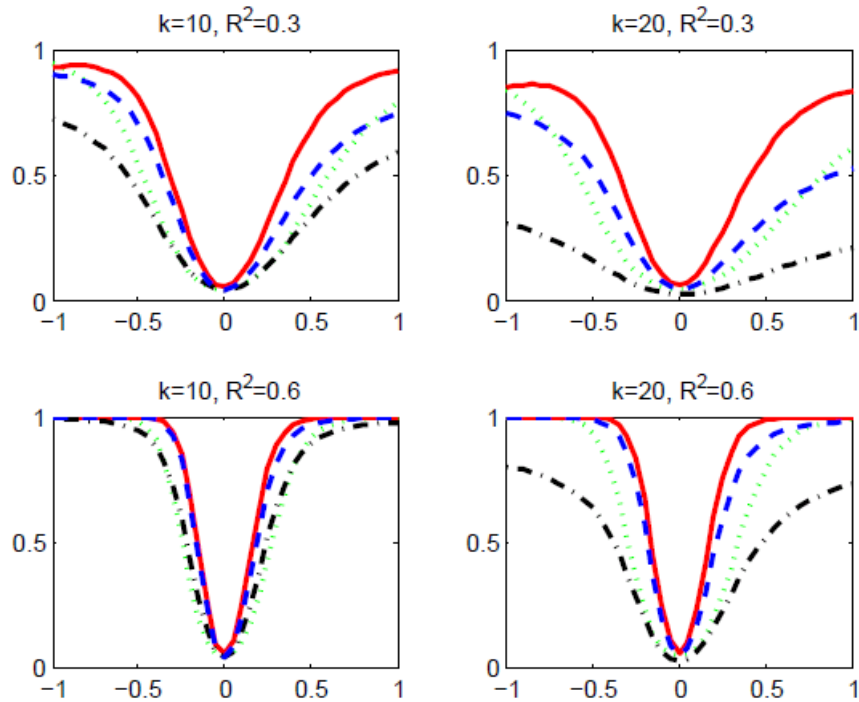


Panel B: $\rho=0.9$

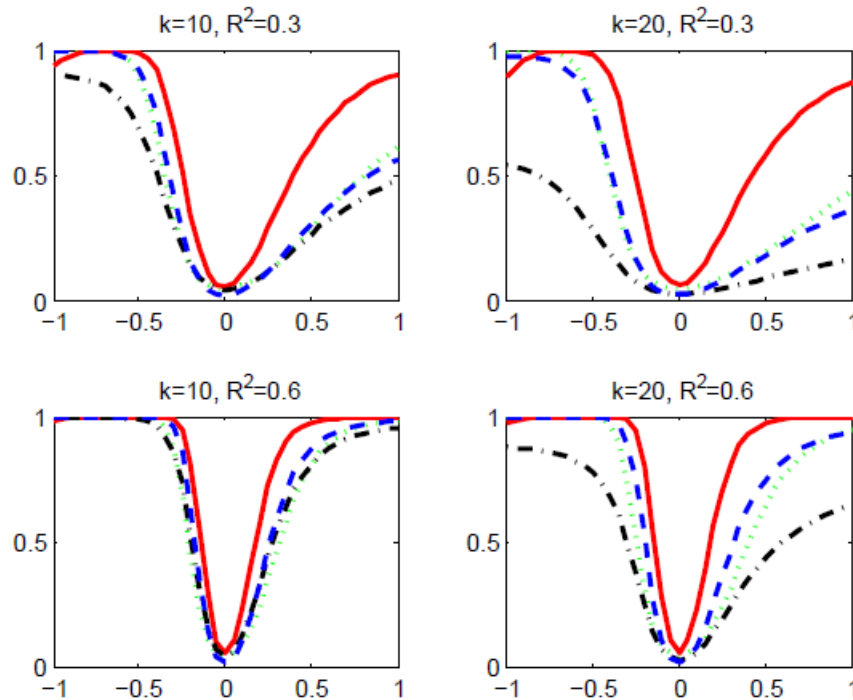


Notes: This plots the power curve in the Monte-Carlo simulation described in section 3 with normal errors for the following tests: Kleibergen (red solid line), Anderson-Rubin (green dashed-dotted line), split-sample Anderson-Rubin (black dotted line) and the proposed test (blue dashed line). The sample size is $n=100$.

Figure 3. Power curves for alternative test statistics
Panel A: $\rho=0.5$

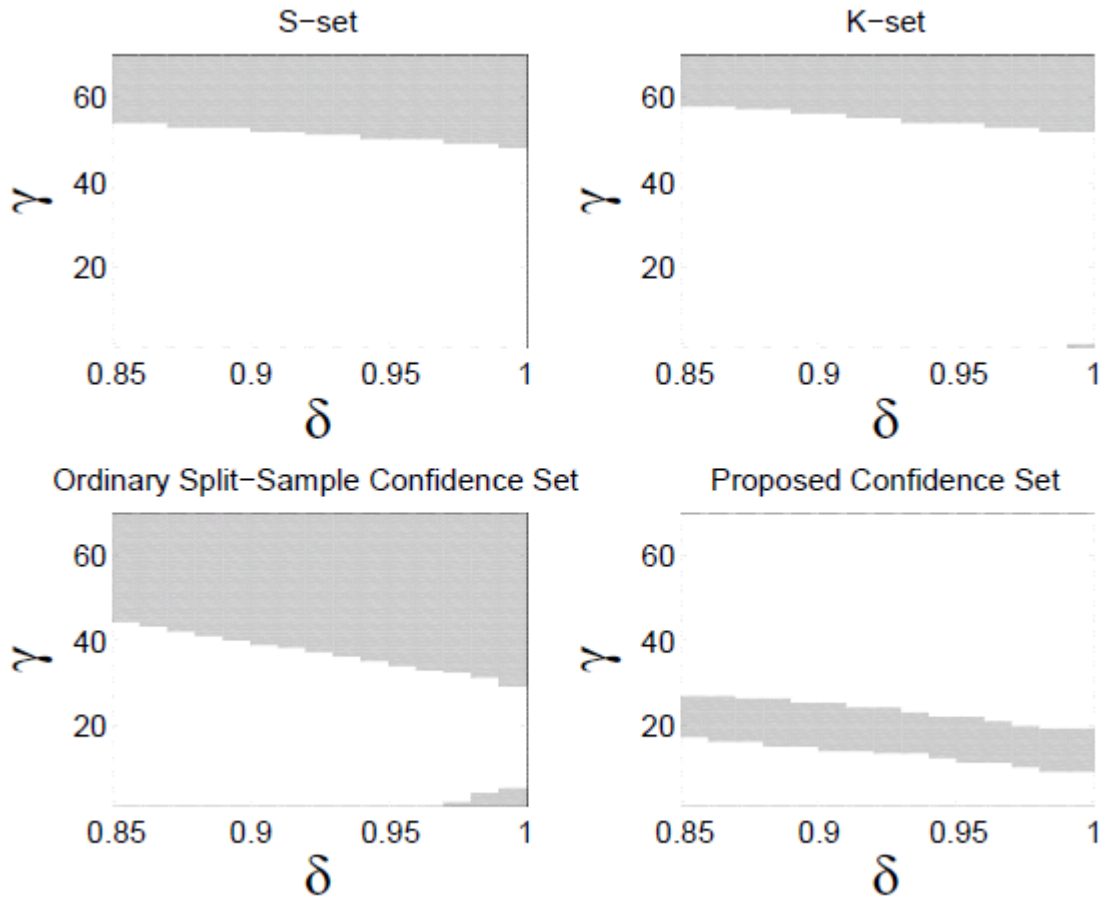


Panel B: $\rho=0.9$



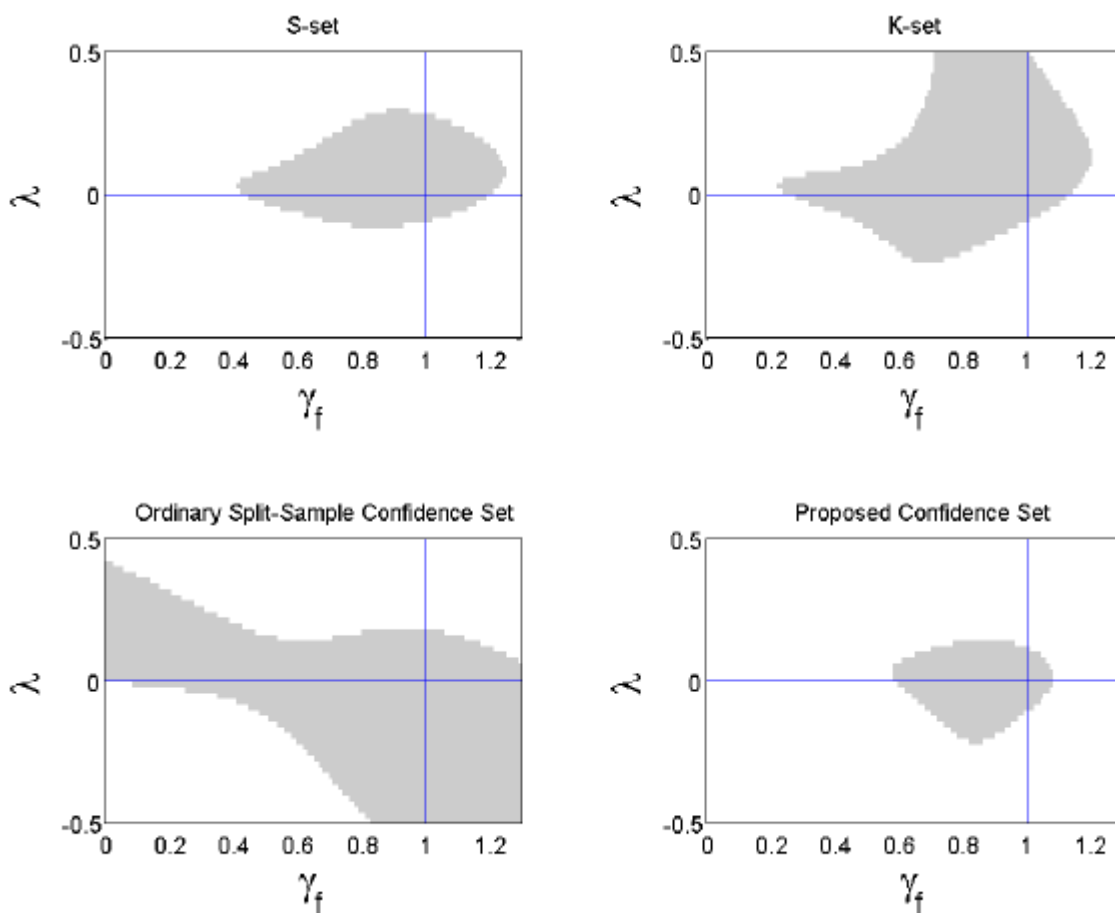
Notes: This plots the power curves in the Monte-Carlo simulations as in Figure 2, except with non-normal errors.

Figure 4. Alternative Confidence Sets in the Euler Equation Consumption CAPM



Notes: This figure gives the 95 percent confidence sets based on the acceptance regions of alternative tests for the parameters of the Euler equation consumption CAPM.

Figure 5. Alternative Confidence Sets for the New-Keynesian Phillips Curve



Notes: This figure gives the 95 percent confidence sets based on the acceptance regions of alternative tests for the parameters of the new Keynesian Phillips curve.