The Stata Journal (2009)
9, Number 3, pp. 1–24

Implementing weak-instrument robust tests for a general class of instrumental-variables models

Keith Finlay
Tulane University
New Orleans, LA
kfinlay@tulane.edu

Leandro M. Magnusson
Tulane University
New Orleans, LA
lmagnuss@tulane.edu

Abstract. We present a minimum distance approach for conducting hypothesis testing in the presence of potentially weak instruments. Under this approach, we propose size-correct tests for limited dependent variable models with endogenous explanatory variables such as endogenous tobit and probit models. Additionally, we extend weak-instrument tests for the linear instrumental-variables model by allowing for variance–covariance estimation that is robust to arbitrary heteroskedasticity or intracluster dependence. We invert these tests to construct confidence intervals on the coefficient of the endogenous variable. We also provide a postestimation command for Stata, called `rivtest`, for computing the tests and estimating confidence intervals.

Keywords: st0001, rivtest, ivregress, ivprobit, ivtobit, condivreg, ivreg2, weak instruments, endogenous tobit, endogenous probit, two-stage least squares, hypothesis testing, confidence intervals

1 Introduction

In this article, we present an indirect method for performing hypothesis testing based on the classical minimum distance approach. This method allows us to develop two extensions to the current set of weak-instrument robust tests that are available for linear instrumental-variables (IV) models. The first extension allows one to perform size-correct inference for a class of limited dependent variable (LDV) models that includes the endogenous tobit and probit models. The second extension allows size-correct inference with the linear IV model when dealing with covariance matrices with arbitrary heteroskedasticity or intracluster dependence.

There exists vast literature dealing with inference in the linear IV model when instruments are weak (see Stock, Wright, and Yogo [2002] for a review). When instruments are weak, point estimators are biased and Wald tests are unreliable. There are several tests available for linear IV models that have the correct size even when instruments are weak. These include the Anderson–Rubin (AR) statistic (Anderson and Rubin 1949), the Kleibergen–Moreira Lagrange multiplier (LM) test (Moreira 2003; Kleibergen 2007), the overidentification (J) test, and the conditional likelihood-ratio (CLR) test.

Concern about weak identification is not isolated to linear IV models. Identification issues also arise in the popular class of LDV models with endogenous explanatory variables. The endogenous tobit and endogenous probit models are two examples of these models (the `ivtobit` and `ivprobit` commands in Stata).

Extending the weak-instrument robust tests from the linear IV case to the LDV models is not straightforward. In the LDV models, the untested (nuisance) parameters are not separable from the structural parameters. As such, the orthogonal transformation that projects nuisance parameters out from the tests in the linear IV is not valid in the LDV case.

Fortunately, for this particular class of LDV models, the structural model also has a reduced-form representation. Consequently, inference on the structural parameter can be conducted indirectly by testing the restrictions on the reduced-form coefficients imposed by the underlying relationship between the structural and reduced-form parameters. Magnusson (2008a) describes this method of conducting inference under weak identification as the minimum distance approach. Our proposed tests for the endogenous variable coefficient have the correct size regardless of whether the identification condition holds.

Working with the reduced-form models also allows us to relax the homoskedastic assumption used in other implementations of the tests (e.g., the `condivreg` command of Moreira and Poi [2003] and Mikusheva and Poi [2006]). This is possible because the asymptotic behavior of our tests is derived from the reduced-form parameters estimator. In the linear IV model, this property allows us to use the heteroskedastic-robust variance–covariance matrix estimate as the reduced-form parameters covariance matrix. The same method allows us to deal with covariance matrices with cluster dependence. Some of these tests are asymptotically equivalent to those proposed by Chernozhukov and Hansen (2008), who also use a reduced-form approach.

Once we compute the statistical tests, we derive confidence intervals by inverting them. This guarantees that our confidence intervals have the correct coverage probability despite the instruments' strength or weakness. For the linear IV model under homoskedasticity, the existence of a closed-form solution for confidence intervals has been shown by Dufour (2003) for the AR test and by Mikusheva (2005) for the LM and CLR tests. However, their methods do not extend to nonlinear models or models with nonspherical residuals, so we use a grid search for estimating confidence intervals for the other models.

Because our tests are not model specific, we propose just one postestimation command for Stata, called `rivtest`. The command tests the simple composite hypothesis $H_0 : \beta = \beta_0$ against the alternative $H_a : \beta \neq \beta_0$ using five statistics: AR, LM, J, the combination of LM and J, and CLR. The command will also compute the confidence intervals based on these statistics. `rivtest` can be used after running `ivregress`, `ivreg2`, `ivprobit`, or `ivtobit` in Stata with one endogenous variable.

In the next section, we present a brief description of our tests. Then we present a general algorithm for implementing them. Next we discuss the command syntax of our postestimation command, `rivtest`, and provide examples of its use. Finally, we show results from Monte Carlo simulations we performed using the `rivtest` command.

2 Weak-instrument robust tests in LDV models: A minimum distance approach

2.1 Setup

We start by considering a class of models that includes both typical two-stage least-squares models and LDV models. Suppose there exists a model that satisfies the following structural form representation:

$$\begin{cases} y_i^* = x_i\beta + w_i\gamma + u_i \\ x_i = z_i\pi_z + w_i\pi_w + v_i \end{cases} \qquad \text{for} \quad i = 1, \ldots, n \tag{1}$$

where y_i^* is a latent endogenous variable and x_i is a continuously observed endogenous explanatory variable; w_i and z_i are, respectively, vectors of included and excluded instruments with dimensions $1 \times k_w$ and $1 \times k_z$; and the residuals u_i and v_i are independently distributed. Rather than observing y_i^*, we observe

$$y_i = f\left(y_i^*\right)$$

where f is a known function. This representation is compatible with the class of LDV models in this study. For the endogenous tobit model, let d_{lb} and d_{ub} be, respectively, the lower and the upper bound. So, we have $y_i = d_{lb}$ if $y_i^* \leq d_{lb}$; $y_i = y_i^*$ if $d_{lb} < y_i^* < d_{ub}$; and $y_i = d_{ub}$ if $y_i^* \geq d_{ub}$. For the endogenous probit, we have $y_i = 0$ if $y_i^* \leq 0$ and $y_i = 1$ if $y_i^* > 0$. In particular, when $y_i = y_i^*$ we have the well-known linear IV model.

Our goal is to test $H_0 \colon \beta = \beta_0$ against $H_a \colon \beta \neq \beta_0$. However, whereas the coefficient γ can be concentrated out of the linear IV model, this is not possible under a more general specification, so the available tests are inappropriate.

An unrestricted reduced-form model derived from (1) is

$$\begin{cases} y_i^* = z_i\delta_z + w_i\delta_w + \epsilon_i \\ x_i = z_i\pi_z + w_i\pi_w + v_i \end{cases} \tag{2}$$

where $\epsilon_i = v_i\beta + u_i$. The restrictions imposed by the structural model over the reduced-form parameters give us the following relation:

$$\delta_z = \pi_z\beta \tag{3}$$

We use (3) to develop our tests on the structural parameter, β, based on the unrestricted model (2). In this representation, the global identification of β requires that $\| \pi_z \| \neq 0$. So as π_z approaches zero, the instruments become weaker.

For now, let's assume that δ_z and π_z are consistently estimated by $\widehat{\delta}_z$ and $\widehat{\pi}_z$. Let's also assume that Λ, the asymptotic variance–covariance of $\sqrt{n}\left[(\widehat{\delta}_z - \delta_z)', \quad (\widehat{\pi}_z - \pi_z)'\right]'$, is also consistently estimated by

$$\widehat{\Lambda} = \begin{bmatrix} \widehat{\Lambda}_{\delta_z\delta_z} & \widehat{\Lambda}_{\delta_z\pi_z} \\ \widehat{\Lambda}_{\pi_z\delta_z} & \widehat{\Lambda}_{\pi_z\pi_z} \end{bmatrix}$$

Let's introduce two more statistics:

$$\begin{aligned}\widehat{\Psi}_\beta &= \widehat{\Lambda}_{\delta_z\delta_z} - \beta\widehat{\Lambda}_{\delta_z\pi_z} - \beta\widehat{\Lambda}_{\pi_z\delta_z} + (\beta)^2\widehat{\Lambda}_{\pi_z\pi_z} \\ \widehat{\overline{\pi}}_\beta &= \widehat{\pi}_z - \left(\widehat{\Lambda}_{\pi_z\delta_z} - \beta\widehat{\Lambda}_{\pi_z\pi_z}\right)\widehat{\Psi}_\beta^{-1}\left(\widehat{\delta}_z - \widehat{\pi}_z\beta\right)\end{aligned}$$

The first statistic is an estimate of the asymptotic covariance matrix of $\sqrt{n}(\widehat{\delta}_z - \widehat{\pi}_z\beta)$. The second statistic is an estimate of π_z, whose properties are discussed in Magnusson (2008a).

2.2 Weak-instrument robust tests

Under $H_0\colon \beta = \beta_0$, our version of the AR test is

$$\begin{aligned}\text{AR}_{\text{MD}}(\beta_0) &= n\left(\widehat{\delta}_z - \widehat{\pi}_z\beta_0\right)'\widehat{\Psi}_{\beta_0}^{-1}\left(\widehat{\delta}_z - \widehat{\pi}_z\beta_0\right) \\ &\xrightarrow{d} \chi^2(k_z)\end{aligned}$$

where the value inside the parentheses indicates the chi-squared distribution degrees of freedom. Then we reject H_0 at significance level α if $\text{AR}_{\text{MD}}(\beta_0)$ is greater than the $1-\alpha$ percentile of the $\chi^2(k_z)$ distribution.

The AR_{MD} statistic simultaneously tests the value of the structural parameter and the overidentification restriction. We can make an orthogonal decomposition of the AR_{MD} test into two statistics, namely, the LM_{MD} and J_{MD} tests. Under the null hypothesis, the LM_{MD} statistic tests the value of the structural parameter given that the overidentification condition holds, while the J_{MD} statistic tests the overidentification restriction given the value of β_0. They are

$$\text{LM}_{\text{MD}}(\beta_0) = n\left\{\widehat{\Psi}_{\beta_0}^{-\frac{1}{2}}\left(\widehat{\delta}_z - \widehat{\pi}_z\beta_0\right)\right\}'\widehat{P}_{\beta_0}\left\{\widehat{\Psi}_{\beta_0}^{-\frac{1}{2}}\left(\widehat{\delta}_z - \widehat{\pi}_z\beta_0\right)\right\} \tag{4}$$

$$J_{\text{MD}}(\beta_0) = n\left\{\widehat{\Psi}_{\beta_0}^{-\frac{1}{2}}\left(\widehat{\delta}_z - \widehat{\pi}_z\beta_0\right)\right\}'\widehat{M}_{\beta_0}\left\{\widehat{\Psi}_{\beta_0}^{-\frac{1}{2}}\left(\widehat{\delta}_z - \widehat{\pi}_z\beta_0\right)\right\} \tag{5}$$

where

$$\begin{aligned}\widehat{P}_{\beta_0} &= \frac{\left(\widehat{\Psi}_{\beta_0}^{-\frac{1}{2}}\widehat{\overline{\pi}}_{\beta_0}\right)\left(\widehat{\Psi}_{\beta_0}^{-\frac{1}{2}}\widehat{\overline{\pi}}_{\beta_0}\right)'}{\left(\widehat{\overline{\pi}}_{\beta_0}'\widehat{\Psi}_{\beta_0}^{-1}\widehat{\overline{\pi}}_{\beta_0}\right)} \\ \widehat{M}_{\beta_0} &= I_{k_z} - \widehat{P}_{\beta_0}\end{aligned}$$

and I_{k_z} is a $k_z \times k_z$ identity matrix. Assuming that some regularity conditions hold under the null hypothesis, we have

$$\begin{aligned}\text{LM}_{\text{MD}}(\beta_0) &\xrightarrow{d} \chi^2(1) \\ J_{\text{MD}}(\beta_0) &\xrightarrow{d} \chi^2(k_z - 1)\end{aligned}$$

independent of whether the instruments are weak (see Magnusson [2008a] for more details). From (4) and (5), we have

$$\text{AR}_{\text{MD}} = \text{LM}_{\text{MD}} + J_{\text{MD}}$$

It is well-known that the LM_{MD} test suffers from a spurious decline of power at some regions of the parameter space. In those regions, the J_{MD} test approximates the AR_{MD} test, which always has discriminatory power. We combine the LM_{MD} and J_{MD} tests to rule out regions where the LM_{MD} test behaves spuriously. For example, testing $H_0 : \beta = \beta_0$ at the 5% significance level could be performed by testing the null at the 4% significance level with the LM_{MD} test and at the 1% significance level with the J_{MD} test. We reject the null if either K_{MD} or J_{MD} is rejected. We call this combination test the LM-J_{MD} test.

The minimum distance version of Moreira's (2003) conditional likelihood-ratio test is

$$\begin{aligned}\text{CLR}_{\text{MD}}(\beta_0) =& \frac{1}{2}\Big[\text{AR}_{\text{MD}}(\beta_0) - \text{rk}(\beta_0) + \\ & \sqrt{\{\text{AR}_{\text{MD}}(\beta_0) + \text{rk}(\beta_0)\}^2 - 4J_{\text{MD}}(\beta_0)\text{rk}(\beta_0)}\Big]\end{aligned}$$

where

$$\begin{aligned}\text{rk}(\beta_0) =& \; n\left(\widehat{\overline{\pi}}'_{\beta_0}\widehat{\Xi}^{-1}_{\beta_0}\widehat{\overline{\pi}}_{\beta_0}\right) \\ \widehat{\Xi}_{\beta_0} =& \; \widehat{\Lambda}_{\pi_z\pi_z} - \left(\widehat{\Lambda}_{\pi_z\delta_z} - \beta_0\widehat{\Lambda}_{\pi_z\pi_z}\right)\widehat{\Psi}^{-1}_{\beta_0}\left(\widehat{\Lambda}_{\delta_z\pi_z} - \beta_0\widehat{\Lambda}_{\pi_z\pi_z}\right)\end{aligned}$$

The asymptotic distribution of the CLR_{MD} is not pivotal and depends on $\text{rk}(\beta_0)$. The critical values of this test are calculated by simulating independent values of $\chi^2(1)$ and $\chi^2(k_z-1)$ for a given value of $\text{rk}(\beta_0)$. This approach is not satisfactory because accuracy demands many simulations, which can be computationally intensive. For linear IV models under homoskedasticity, Andrews, Moreira, and Stock (2007) provide a formula for computing the p-value function of the CLR test (which is embedded in the `condivreg` command). Although this is not the correct p-value function when homoskedasticity is violated, our simulations indicate that it provides a good approximation.

Two Stata packages currently provide some functionality to perform these tests. For the linear IV case under homoskedastic residuals, the `condivreg` command in Stata provides a set of weak-instrument robust tests (Moreira and Poi 2003; Mikusheva and Poi 2006). Our command, `rivtest`, complements `condivreg` by offering weak-instrument robust tests for a larger class of models. For nonhomoskedastic residuals, Baum, Schaffer, and Stillman (2007) provide the AR test in the `ivreg2` package. The degrees of freedom of the AR test depends on the number of instruments and not on the number of endogenous variables, so its power decreases as one increases the number of instruments. We complement this package by offering a set of tests that are valid even with many instruments.

2.3 Confidence intervals

Confidence intervals for the proposed tests are derived by inverting the statistical tests. By definition, confidence intervals derived from the AR_{MD}, LM_{MD}, LM-J_{MD}, and CLR_{MD} tests are, respectively,

$$\begin{aligned}
C^{\text{AR}_{\text{MD}}}_{(1-\tau)} &= \left\{\beta_0 : \text{AR}_{\text{MD}}(\beta_0) < \chi^2_{1-\tau,k_z}\right\} \\
C^{\text{LM}_{\text{MD}}}_{(1-\tau)} &= \left\{\beta_0 : \text{LM}_{\text{MD}}(\beta_0) < \chi^2_{1-\tau,1}\right\} \\
C^{\text{LM}_{\text{MD}}-J_{\text{MD}}}_{(1-\tau)} &= \left[\beta_0 : \left\{\text{LM}_{\text{MD}}(\beta_0) < \chi^2_{1-w_1\tau,1}\right\} \cap \right. \\
&\qquad \left. \left\{J_{\text{MD}}(\beta_0) < \chi^2_{1-w_2\tau,k_z-1}\right\}\right] \\
C^{\text{CLR}_{\text{MD}}}_{(1-\tau)} &= \left[\beta_0 : \text{CLR}_{\text{MD}}(\beta_0) < c\left\{\text{rk}(\beta_0)\right\}\right]
\end{aligned}$$

where τ denotes the significance level, $w_1 + w_2 = 1$, and $c\{\text{rk}(\beta_0)\}$ is the 95th percentile of the distribution of the CLR_{MD} tests conditional on the value of $\text{rk}(\beta_0)$.

The weak instrument robust confidence intervals are not necessarily convex or symmetric as is the usual Wald-type confidence interval, which includes points two standard deviations from the estimated coefficient. For example, they can be a union of disjoint intervals or the real line when the instruments are completely irrelevant. The AR_{MD} confidence interval can be empty. This occurs when the overidentifying restriction is rejected for any value of β. However, the LM_{MD} and CLR_{MD} confidence intervals are never empty because the continuous updating minimum distance estimate always belongs to them.[1]

Dufour (2003) and Mikusheva (2005) provide closed-form solutions for obtaining confidence intervals in the homoskedastic linear IV model. In particular, Mikusheva, by solving quadratic inequalities, proposes a numerically simple algorithm for estimating confidence intervals derived from the LM_{MD} and CLR_{MD} tests. However, their methods are not generalized to either nonspherical residuals or models with LDV. We employ their solutions for the homoskedastic linear IV model. In the other models, we use the grid search method for generating the confidence intervals by testing points in the parameter space. Points $\overline{\beta}$ for which $H_0 : \beta = \overline{\beta}$ is not rejected belong in the confidence interval. The user has the option to choose the interval and the number of points in the grid search. For the LM-J_{MD} test, the user can select the weight, w_1, given to the LM_{MD}. The default option is $w_1 = 0.8$.

1. The continuous updating minimum distance estimate is the value that minimizes the AR_{MD} test. It is not numerically equal to the generalized method of moments continuous updating estimate of Hansen, Heaton, and Yaron (1996).

3 Generic algorithm for implementing minimum distance weak-instrument robust tests

The implementation of our weak-instrument robust tests takes advantage of several built-in functions of Stata. We separate our implementation into two cases: one in which residuals are homoskedastic and another in which residuals have either arbitrary heteroskedasticity or intracluster dependence.

3.1 Homoskedastic residuals

Under a homoskedastic assumption, we use the fact that $u_i = v_i\alpha + \varepsilon_i$, where $\alpha = \sigma_{vu}^{-1}\sigma_{vv}$. This condition is suitable, for example, if residuals are jointly normally distributed. Moreover, the assumption allows the computation of the tests by using built-in functions available in Stata (Magnusson 2008b). The reduced-form (2) becomes

$$\begin{cases} y_i^* = z_i\delta_z + w_i\delta_w + v_i\delta_v + \varepsilon_i \\ x_i = z_i\pi_z + w_i\pi_w + v_i \end{cases} \tag{6}$$

In the above representation, ε_i and v_i are independent by construction. The test algorithm has the following steps:

1. Estimate π_z and $\Lambda_{\pi_z\pi_z}$ by ordinary least squares (OLS). Denote the estimated values as $\widehat{\pi}_z$ and $\widehat{\Lambda}_{\pi_z\pi_z}$. Also compute the OLS estimated residuals:

   $$\widehat{v}_i = x_i - z_i\widehat{\pi}_z - w_i\widehat{\pi}_w$$

2. Estimate δ_z and δ_w by using the following equation:

   $$y_i^* = z_i\delta_z + w_i\delta_w + \widehat{v}_i\delta_v + \widetilde{\varepsilon}_i$$

   where $\widetilde{\varepsilon}_i = \varepsilon_i - (\widehat{v}_i - v_i)\delta_v$. Denote the estimated values of δ_z, δ_w, and δ_v as $\widehat{\delta}_z$, $\widehat{\delta}_w$, and $\widehat{\delta}_v$, respectively. For the endogenous probit model, our algorithm fixes $\sigma_{\varepsilon\varepsilon} = 1$ for normalization, which is a different normalization than the default option in Stata ($\sigma_{uu} = 1$) but the same as the Newey two-step estimator (see [R] **ivprobit**).

3. Save $\widehat{\Gamma}_{\delta_z\delta_z}$, the output of the variance–covariance matrix estimate of $\widehat{\delta}_z$. This is not the "correct" variance–covariance of $\widehat{\delta}_z$ because we are not adjusting for the presence of $\widehat{v}_i$.

Using the same notation as in the body of the text, we have

$$\begin{aligned} \widehat{\Psi}_\beta &= \widehat{\Gamma}_{\delta_z\delta_z} + (\widehat{\delta}_v - \beta)^2\widehat{\Lambda}_{\pi_z\pi_z} \\ \widehat{\overline{\pi}}_\beta &= \widehat{\pi}_z - (\widehat{\delta}_v - \beta)^2\widehat{\Psi}_\beta^{-1}\widehat{\Lambda}_{\pi_z\pi_z} \\ \widehat{\Xi}_{\beta_0} &= \widehat{\Lambda}_{\pi_z\pi_z} - (\widehat{\delta}_v - \beta)^2\widehat{\Lambda}_{\pi_z\pi_z}\widehat{\Psi}_\beta^{-1}\widehat{\Lambda}_{\pi_z\pi_z} \end{aligned}$$

3.2 Heteroskedastic/clustered residuals

For heteroskedasticity or cluster dependence in the distribution of errors, we consider just the linear model. Baum, Schaffer, and Stillman (2007) provide an option using a generalized method of moments approach for autocorrelation- and heteroskedasticity-robust AR tests in the `ivreg2` command. We extend this functionality for the LM_{MD}, LM-J_{MD}, and CLR_{MD} tests.

The implementation is similar to the homoskedastic case. The reduced-form model is

$$\begin{cases} y_i = z_i\delta_z + w_i\delta_w + e_i \\ x_i = z_i\pi_z + w_i\pi_w + v_i \end{cases}$$

We estimate the δ_z, π_z, $\Lambda_{\delta_z\delta_z}$, and $\Lambda_{\pi_z\pi_z}$ by running two separate regressions with the appropriate robust or cluster options. The covariance term $\Lambda_{\pi_z\delta_z}$ has the general sandwich formula

$$\widehat{\Lambda}_{\pi_z\delta_z} = A\ B\ A'$$

where $A = (Z^{\perp\prime}Z^{\perp})^{-1}$ is a $k_z \times k_z$ matrix, $Z^{\perp} = M_W Z$, and $M_W = I_n - W(W'W)^{-1}W'$, the matrix that projects Z to the orthogonal space spanned by W. Let's denote $\widehat{v}$ and $\widehat{e}$ as the vectors of OLS residuals. The B matrix is given by:

$$\sum_j z_j^{\perp\prime}\widehat{v}_j\widehat{e}_j' z_j^{\perp}$$

For robust standard errors, $z_j^{\perp}$ is a $k_z \times 1$ vector, and $\widehat{v}_j$ and $\widehat{u}_j$ are scalars. For clustered standard errors, $z_j^{\perp}$ is a $k_z \times n_j$ matrix, and $\widehat{v}_j$ and $\widehat{u}_j$ are $n_j \times 1$ vectors, where n_j is the number of observations in cluster j.

The tests obtained here and by Chernozhukov and Hansen (2008) are closely related. They work with the following regression model:

$$y_i - Y_i\beta = Z_i\gamma + u_i \tag{7}$$

A simple t test, $\widehat{\gamma}/s_{\widehat{\gamma}}$, is the same as testing $H_0\colon \beta = \beta_0$, where $\widehat{\gamma}$ is the OLS estimator derived from (7) replacing β with β_0. Our AR test and the AR test of Chernozhukov and Hansen (2008) are identical. Our LM test, however, is only asymptotically equivalent to theirs; they are slightly different in small samples.[2]

4 The rivtest command

The software package accompanying this article contains a Stata command, `rivtest`, to implement the tests discussed above after using the `ivregress`, `ivreg2`, `ivprobit`, or `ivtobit` command.

2. A proof is available upon request.

4.1 Command description

For `ivregress` and `ivreg2`, `rivtest` supports limited-information maximum likelihood and two-stage least-squares models (the `liml` and `2sls` options of `ivregress`, respectively), as well as `vce(robust)` and `vce(cluster` *clustvar*`)` options for variance–covariance estimation. For `ivprobit` and `ivtobit`, `rivtest` supports all variance–covariance estimation options except the `vce(robust)` and `vce(cluster` *clustvar*`)` options. Weights are allowed as long as they are supported by the appropriate IV command.

`rivtest` calculates the minimum distance version of the AR test statistic. When the IV model contains more than one instrumental variable, `rivtest` also conducts the minimum distance versions of the CLR test, the LM test, the J overidentification test, and a combination of the LM multiplier and overidentification tests (LM-J). As a reference, `rivtest` also presents the Wald test.

The AR test is a joint test of the structural parameter and the overidentification restrictions. The AR statistic can be decomposed into the LM statistic, which tests only the structural parameter, and the J statistic, which tests only the overidentification restrictions. (This J statistic, evaluated at the null hypotheses, is different from the Hansen J statistic, which is evaluated at the parameter estimate.) The LM test loses power in some regions of the parameter space when the likelihood function has a local extrema or inflection. In the linear IV model with homoskedasticity, the CLR statistic combines the LM statistic and the J statistic in the most efficient way, thereby testing both the structural parameter and the overidentification restrictions simultaneously. The LM-J combination test is another approach for testing the hypotheses simultaneously. It is more efficient than the AR test and allows different weights to be put on the parameter and overidentification hypotheses. The CLR test is the most powerful test for the linear model under homoskedasticity (within a class of invariant similar tests), but this result has not been proven yet for other IV-type estimators, so we present all test results.

`rivtest` can also estimate confidence intervals based on the AR, CLR, LM, and LM-J tests. With `ivregress` there is a closed-form solution for these confidence intervals only when homoskedasticity is assumed. More generally, `rivtest` estimates confidence intervals through test inversion over a grid. The default grid is twice the size of the confidence interval based on the Wald test. As a reference, `rivtest` also presents the Wald confidence interval.

4.2 Syntax

The following is the command syntax for `rivtest`:

```
rivtest [ , null(#) lmwt(#) small ci grid(numlist) points(#)
   gridmult(#) usegrid retmat level(#) ]
```

4.3 Options

The options for `rivtest` relate to testing and confidence-interval estimation.

Testing options

`null(#)` specifies the null hypothesis for the coefficient on the endogenous variable in the IV model. The default is `null(0)`.

`lmwt(#)` is the weight put on the LM test statistic in the LM-J test. The default is `lmwt(0.8)`.

`small` specifies that small-sample adjustments be made when test statistics are calculated. The default is given by whatever small-sample adjustment option was chosen in the IV command.

Confidence-interval options

`ci` requests that confidence intervals be estimated. By default, these are not estimated because grid-based test inversion can be time intensive.

`grid(`*numlist*`)` specifies the grid points over which to calculate the confidence sets. The default grid is centered around the point estimate with a width equal to twice the Wald confidence interval. That is, if $\widehat{\beta}$ is the estimated coefficient on the endogenous variable, $\widehat{\sigma}_\beta$ is its estimated standard error, and $1-\alpha$ is the confidence level, then the default endpoints of the interval over which confidence sets will be calculated are $\widehat{\beta} \pm 2z_{\alpha/2}\widehat{\sigma}_\beta$. With weak instruments, this is often too small of a grid to estimate the confidence intervals. `grid(`*numlist*`)` may not be used with the other two grid options: `points(#)` and `gridmult(#)`. If one of the other options is used, only input from `grid(`*numlist*`)` will be used to construct the grid.

`points(#)` specifies the number of equally spaced values over which to calculate the confidence sets. The default is `points(100)`. Increasing the number of grid points will increase the time required to estimate the confidence intervals, but a greater number of grid points will improve precision.

`gridmult(#)` is another way of specifying a grid to calculate confidence sets. This option specifies that the grid be # times the size of the Wald confidence interval. The default is `gridmult(2)`.

`usegrid` forces grid-based test inversion for confidence-interval estimation under the homoskedastic linear IV model. The default is to use the analytic solution. Under the other models, grid-based estimation is the only method.

`retmat` returns a matrix of test results over the confidence-interval search grid. This matrix can be large if the number of grid points is large, but it can be useful for graphing confidence sets.

`level(#)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`. Because the LM-J test has no p-value function, we report whether the test is rejected. Changing `level(#)` also changes the level of significance used to determine this result: [100-`level(#)`]%.

4.4 Saved results

`rivtest` saves the following in `r()`:

Scalars

`r(null)`	null hypothesis
`r(clr_p)`	CLR test p-value
`r(clr_stat)`	CLR test statistic
`r(ar_p)`	AR test p-value
`r(ar_chi2)`	AR test statistic
`r(lm_p)`	LM test p-value
`r(lm_chi2)`	LM test statistic
`r(j_p)`	J test p-value
`r(j_chi2)`	J test statistic
`r(lmj_r)`	LM-J test rejection indicator
`r(rk)`	rk statistic
`r(wald_p)`	Wald test p-value
`r(wald_chi2)`	Wald test statistic
`r(points)`	number of points in grid used to estimate confidence sets

Macros

`r(clr_cset)`	confidence set based on CLR test
`r(ar_cset)`	confidence set based on AR test
`r(lm_cset)`	confidence set based on LM test
`r(lmj_cset)`	confidence set based on LM-J test
`r(wald_cset)`	confidence set based on Wald test
`r(inexog)`	list of instruments included in the second-stage equation
`r(exexog)`	list of instruments excluded from the second-stage equation
`r(endo)`	endogenous variable
`r(grid)`	range of grid used to estimate confidence sets

Matrices

`r(citable)`	table with test statistics, p-values, and rejection indicators for every grid point over which hypothesis was tested

5 Examples: Married female labor market participation

We demonstrate the use of the `rivtest` command in a set of applications with the data from Mroz (1987), available from the Stata web site at http://www.stata.com/data/jwooldridge/eacsap/mroz.dta. These examples are related to married female labor supply and illustrate the differences between robust and non-robust inference when instruments are potentially weak.

5.1 Example 1: Two-stage least squares with unknown heteroskedasticity

In this example, we fit a two-stage least-squares model with Stata's `ivregress` command using the robust variance–covariance estimation option to account for arbitrary

heteroskedasticity. We regress working hours (`hours`), on log wages (`lwage`), other household income in logs (`nwifeinc`), years of education (`educ`), number of children less than 6 years old (`kidslt6`), and the number of children at least 6 years old (`kidsge6`). As instruments for the wage, we use labor market experience (`exper`) and its square (`expersq`), and father's and mother's years of education (`fatheduc` and `motheduc`). We consider the subsample of women who are participating in the labor market and have strictly positive wages.

```
. use http://www.stata.com/data/jwooldridge/eacsap/mroz.dta
. ivregress 2sls hours nwifeinc educ age kidslt6 kidsge6 (lwage = exper expersq
> fatheduc motheduc) if inlf==1 , first vce(robust)
First-stage regressions
```

```
                                        Number of obs   =        428
                                        F(   9,    418) =      10.78
                                        Prob > F        =     0.0000
                                        R-squared       =     0.1710
                                        Adj R-squared   =     0.1532
                                        Root MSE        =     0.6655
```

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
nwifeinc	.0057445	.0027375	2.10	0.036	.0003636	.0111255
educ	.1127654	.0154679	7.29	0.000	.0823609	.1431699
age	-.0053092	.0063134	-0.84	0.401	-.0177191	.0071007
kidslt6	-.066367	.103709	-0.64	0.523	-.2702231	.137489
kidsge6	-.0192837	.0292029	-0.66	0.509	-.0766866	.0381191
exper	.0404503	.0151505	2.67	0.008	.0106697	.0702309
expersq	-.0007512	.0004056	-1.85	0.065	-.0015485	.000046
fatheduc	-.0061784	.0106541	-0.58	0.562	-.0271208	.0147639
motheduc	-.016405	.0119691	-1.37	0.171	-.039932	.0071221
_cons	-.2273025	.3343392	-0.68	0.497	-.8844983	.4298933

```
Instrumental variables (2SLS) regression            Number of obs =     428
                                                    Wald chi2(6)  =   18.22
                                                    Prob > chi2   =  0.0057
                                                    R-squared     =       .
                                                    Root MSE      =  1143.2
```

hours	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwage	1265.326	473.6747	2.67	0.008	336.9408	2193.711
nwifeinc	-8.353995	4.57849	-1.82	0.068	-17.32767	.6196797
educ	-148.2865	54.38669	-2.73	0.006	-254.8824	-41.69053
age	-10.23769	9.299097	-1.10	0.271	-28.46358	7.98821
kidslt6	-234.3907	181.9979	-1.29	0.198	-591.1001	122.3187
kidsge6	-59.62672	49.24854	-1.21	0.226	-156.1521	36.89865
_cons	2375.395	535.4835	4.44	0.000	1325.867	3424.923

```
Instrumented:  lwage
Instruments:   nwifeinc educ age kidslt6 kidsge6 exper
               expersq fatheduc motheduc
```

```
. rivtest, ci grid(-1000(10)8000)
Estimating confidence sets over grid points
----+--- 1 ----+--- 2 ----+--- 3 ----+--- 4 ----+--- 5
..................................................   50
..................................................  100
..................................................  150
..................................................  200
..................................................  250
..................................................  300
..................................................  350
..................................................  400
..................................................  450
..................................................  500
..................................................  550
..................................................  600
..................................................  650
..................................................  700
..................................................  750
..................................................  800
..................................................  850
..................................................  900
.
Weak instrument robust tests and confidence sets for linear IV with robust VCE
H0: beta[hours:lwage] = 0
```

Test	Statistic	p-value	95% Confidence Set
CLR	stat(.) = 27.27	Prob > stat = 0.0000	[810, 5330]
AR	chi2(4) = 32.61	Prob > chi2 = 0.0000	[770, 6930]
LM	chi2(1) = 21.22	Prob > chi2 = 0.0000	
	[-830, -670] U [790, 5460]		
J	chi2(3) = 11.39	Prob > chi2 = 0.0098	
LM-J	H0 rejected at 5% level		[760, 5940]
Wald	chi2(1) = 7.14	Prob > chi2 = 0.0076	[336.941, 2193.71]

```
Note: Wald test not robust to weak instruments. Confidence sets estimated for
901 points in [-1000,8000].
```

The confidence intervals derived from weak-instrument robust tests are wider than the Wald confidence interval, indicating that instruments are not strong and that point estimates are biased. The negative values of the LM confidence set are discarded in the LM-J confidence interval, indicating the spurious behavior of the LM test in that part of the parameter space. The above result suggests a positive effect of wages on the labor supply, but `rivtest` is unable to predict the magnitude of the effect.

5.2 Example 2: Endogenous probit

Next we fit a model of labor force participation for the married women in the sample. The binary variable `inlf` equals one if the woman is in the labor market and zero otherwise. The endogenous explanatory variable is nonwife household income, which is instrumented by husband's hours of work (`hushrs`), father's education, mother's education, and the county-level unemployment rate (`unem`). As exogenous variables,

we include education, years of labor market experience, experience squared, number of children less than 6 years old, number of children at least 6 years old, and a dummy for whether the individual lives in a metropolitan area (`city`).

```
. ivprobit inlf educ exper expersq kidslt6 kidsge6 city (nwifeinc = hushrs
> fatheduc motheduc unem), twostep first
Checking reduced-form model...
First-stage regression
```

Source	SS	df	MS		
Model	18057.3855	10	1805.73855		
Residual	83739.7301	742	112.856779		
Total	101797.116	752	135.368505		

Number of obs	=	753			
F(10, 742)	=	16.00			
Prob > F	=	0.0000			
R-squared	=	0.1774			
Adj R-squared	=	0.1663			
Root MSE	=	10.623			

nwifeinc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hushrs	.0029782	.0006719	4.43	0.000	.0016591	.0042972
fatheduc	.1760206	.1385697	1.27	0.204	-.0960147	.4480558
motheduc	-.1395621	.1458037	-0.96	0.339	-.425799	.1466749
unem	.1652976	.1283373	1.29	0.198	-.0866498	.417245
educ	1.218966	.2011015	6.06	0.000	.8241703	1.613762
exper	-.3562876	.1406571	-2.53	0.012	-.632421	-.0801543
expersq	.0031554	.0045229	0.70	0.486	-.0057239	.0120346
kidslt6	-.3788863	.7624489	-0.50	0.619	-1.8757	1.117928
kidsge6	-.1729039	.3105805	-0.56	0.578	-.782625	.4368172
city	4.949449	.8419922	5.88	0.000	3.296478	6.602419
_cons	-2.916913	2.883583	-1.01	0.312	-8.577865	2.744039

Two-step probit with endogenous regressors

Number of obs	=	753
Wald chi2(7)	=	136.69
Prob > chi2	=	0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nwifeinc	-.0631912	.0292417	-2.16	0.031	-.1205038	-.0058785
educ	.2148807	.0473224	4.54	0.000	.1221304	.307631
exper	.1067194	.0225831	4.73	0.000	.0624574	.1509813
expersq	-.0022201	.0006423	-3.46	0.001	-.003479	-.0009611
kidslt6	-.5794973	.1113274	-5.21	0.000	-.797695	-.3612996
kidsge6	.1284411	.0429235	2.99	0.003	.0443126	.2125696
city	.1421479	.1805589	0.79	0.431	-.2117411	.4960368
_cons	-2.038166	.3551659	-5.74	0.000	-2.734279	-1.342054

```
Instrumented:  nwifeinc
Instruments:   educ exper expersq kidslt6 kidsge6 city
               hushrs fatheduc motheduc unem
```

Wald test of exogeneity: chi2(1) = 3.05 Prob > chi2 = 0.0808

```
. rivtest, ci grid(-.2(.001).6)
Estimating confidence sets over grid points
----+--- 1 ----+--- 2 ----+--- 3 ----+--- 4 ----+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
..................................................   550
..................................................   600
..................................................   650
..................................................   700
..................................................   750
..................................................   800
.
Weak instrument robust tests and confidence sets for IV probit
H0: beta[inlf:nwifeinc] = 0
```

Test	Statistic	p-value	95% Confidence Set
CLR	stat(.) = 5.82	Prob > stat = 0.0249	[-.172, -.01]
AR	chi2(4) = 9.50	Prob > chi2 = 0.0498	[-.197, -.001]
LM	chi2(1) = 4.75	Prob > chi2 = 0.0293	
	[-.177, -.008] U [.17, .534]		
J	chi2(3) = 4.75	Prob > chi2 = 0.1913	
LM-J	H0 rejected at 5% level		[-.186, -.005]
Wald	chi2(1) = 4.67	Prob > chi2 = 0.0307	[-.120504,-.005879]

```
Note: Wald test not robust to weak instruments. Confidence sets estimated for
801 points in [-.2,.6].
```

In the endogenous probit model, the `rivtest` command uses the normalization of Newey's minimum chi-squared estimator, $\sigma_\varepsilon = 1$ in (6), which is different from the default normalization used in maximum likelihood estimation, $\sigma_u = 1$ in (1) (see [R] **ivprobit** for further explanation). Therefore, the confidence intervals produced by `rivtest` and the maximum likelihood version of `ivprobit` are not comparable.

In this example, although one instrument, husband's hours of work, has a first-stage t statistic greater than 4, the confidence intervals produced from the weak-instrument tests are significantly larger than the nonrobust Wald confidence interval; for example, the LM-J confidence interval is 50% larger than the Wald confidence interval. Thus the presence of only one strong instrument in the first stage among other weaker ones does not imply that classical inference is correct.

5.3 Example 3: Endogenous tobit

In the following example, we fit an endogenous tobit model with Stata's `ivtobit` command. We regress hours of work, including the many observations in which the woman does not supply labor, on the same regressors as in the previous example.

```
. ivtobit hours educ exper expersq kidslt6 kidsge6 city (nwifeinc = hushrs
> fatheduc motheduc unem), ll(0) first nolog
Tobit model with endogenous regressors           Number of obs   =        753
                                                 Wald chi2(7)    =     173.12
Log likelihood = -6686.3386                      Prob > chi2     =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hours						
nwifeinc	-71.02316	33.59912	-2.11	0.035	-136.8762	-5.170087
educ	183.002	51.47165	3.56	0.000	82.11939	283.8846
exper	121.0376	23.65995	5.12	0.000	74.66493	167.4102
expersq	-2.478807	.623252	-3.98	0.000	-3.700358	-1.257255
kidslt6	-639.99	116.5606	-5.49	0.000	-868.4446	-411.5353
kidsge6	74.23684	41.79029	1.78	0.076	-7.670611	156.1443
city	187.9859	194.1849	0.97	0.333	-192.6095	568.5814
_cons	-1436.843	351.9196	-4.08	0.000	-2126.593	-747.0931
nwifeinc						
educ	1.284978	.198927	6.46	0.000	.8950883	1.674868
exper	-.368858	.1399175	-2.64	0.008	-.6430913	-.0946248
expersq	.0033886	.0044934	0.75	0.451	-.0054183	.0121955
kidslt6	-.3558916	.75725	-0.47	0.638	-1.840074	1.128291
kidsge6	-.1665826	.308437	-0.54	0.589	-.771108	.4379429
city	4.833468	.8349314	5.79	0.000	3.197033	6.469904
hushrs	.0027375	.0007263	3.77	0.000	.001314	.0041611
fatheduc	.1481241	.1277639	1.16	0.246	-.1022886	.3985368
motheduc	-.2084148	.1309959	-1.59	0.112	-.465162	.0483325
unem	.2506685	.1163957	2.15	0.031	.022537	.4787999
_cons	-2.883293	2.871029	-1.00	0.315	-8.510407	2.743821
/alpha	57.91175	34.02567	1.70	0.089	-8.777325	124.6008
/lns	7.062261	.0372561	189.56	0.000	6.989241	7.135282
/lnv	2.356454	.02581	91.30	0.000	2.305867	2.40704
s	1167.081	43.48089			1084.897	1255.491
v	10.55346	.2723849			10.03287	11.10106

```
Instrumented:  nwifeinc
Instruments:   educ exper expersq kidslt6 kidsge6 city
               hushrs fatheduc motheduc unem
Wald test of exogeneity (/alpha = 0): chi2(1) =     2.90  Prob > chi2 = 0.0888
  Obs. summary:        325  left-censored observations at hours<=0
                       428     uncensored observations
                         0 right-censored observations
```

```
. rivtest, ci points(500) gridmult(14)
Estimating confidence sets over grid points
----+--- 1 ----+--- 2 ----+--- 3 ----+--- 4 ----+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
Weak instrument robust tests and confidence sets for IV Tobit
H0: beta[hours:nwifeinc] = 0
```

Test	Statistic	p-value	95% Confidence Set
CLR	stat(.) = 5.35	Prob > stat = 0.0315	[-176.335, -10.053]
AR	chi2(4) = 11.53	Prob > chi2 = 0.0212	[-154.164,-17.4433]
LM	chi2(1) = 3.73	Prob > chi2 = 0.0535	
	[-202.201, 1.03251] U [122.973, 813.968]		
J	chi2(3) = 7.81	Prob > chi2 = 0.0502	
LM-J	H0 not rejected at 5% level		[-216.982, 4.72767]
Wald	chi2(1) = 4.47	Prob > chi2 = 0.0345	[-136.876,-5.17009]

```
Note: Wald test not robust to weak instruments. Confidence sets estimated for
500 points in [-992.966, 850.92].
```

After the `rivtest` command, we have requested two `rivtest` options related to confidence estimation: `points(500)` and `gridmult(14)`, which specify that confidence set estimation should be performed on a grid of 500 points over a width of 14 times the Wald confidence interval (centered around the IV point estimate).[3]

Here we obtain similar results to the ones in the endogenous probit example. While the estimated confidence sets are generally consistent with a negative effect of nonwife income on labor supply, the estimated confidence sets from the weak-instrument tests are wider than the Wald confidence interval.

6 Monte Carlo simulations

To show the performance of the tests, we perform Monte Carlo simulations of the `rivtest` command with linear IV, IV probit, and IV tobit. We show simulations from small ($N = 200$) samples, but results were qualitatively similar with larger samples. We performed simulations with both weak ($\pi_z = 0.1$) and nonweak ($\pi_z = 1$) instruments. The coefficient β is 0.5 and the excluded instruments are drawn from independent standard normal distributions and are the same for all simulations. Finally, we experimented

3. Calculation of the test statistics is almost instantaneous, but grid-based confidence-interval estimation takes time (increasing linearly with the number of grid points). In the IV tobit example, the command required about 2 seconds for 100 grid points and 8 seconds for 500 points.

with three levels of correlation between the error terms in the two equations (ρ): 0.1 for low levels of simultaneity, 0.5 for moderate simultaneity, and 0.8 for a high degree of simultaneity. For each Monte Carlo experiment, we generated 5,000 simulations and computed the rejection probability under the true null hypothesis. All simulations were performed in Stata with the built-in regression commands with our `rivtest` command.[4]

Table 1 shows the results of Monte Carlo simulations for the linear IV model under homoskedasticity and arbitrary heteroskedasticity. Panel A shows the results when the errors are homoskedastic. Here we see that the Wald test does not have the correct size when the instrument is weak for all different degrees of simultaneity. For example, with a highly correlated interequation error ($\rho = 0.8$), the Wald test incorrectly rejected the true parameter in 44.94% of the simulations.

Panel B shows the results when the errors are arbitrarily heteroskedastic.[5] The performance of the Wald test with weak instruments ($\pi = 0.1$) is similar to the previous case: it overrejects the null hypothesis when the errors in the two equations are moderately or highly correlated ($\rho = 0.5$ or $\rho = 0.8$), and underrejects the null hypothesis when the simultaneity is low ($\rho = 0.1$). For the case of strong instruments ($\pi = 1$), the tests have similar nominal sizes.

4. The Monte Carlo simulations include five instruments excluded from the second-stage equation, but only one of the instruments has a nonzero coefficient in the first stage. In the tables, we refer to this coefficient as π. Also, two control variables entered the model, including a vector of ones. The error terms were drawn from a bivariate standard normal distribution with correlation coefficient ρ.
5. We generated this heteroskedasticity by multiplying homoskedastic errors by an independently drawn uniform random variable between zero and two—separately for each equation error.

Table 1. Size (in percent) for testing $H_0 : \beta = 0.5$ at the 5% significance level in the linear IV model under homoskedasticity and arbitrary heteroskedasticity

Models Simulation parameters			Test size Rejection rate for tests (percent)					
A. 2SLS *with homoskedasticity*								
N	π	ρ	CLR	AR	LM	J	LM-J	Wald
200	0.1	0.8	5.34	5.40	5.34	5.30	5.62	44.94
			(0.32)	(0.32)	(0.32)	(0.32)	(0.33)	(0.70)
200	0.1	0.5	5.22	5.08	5.42	5.48	5.38	13.28
			(0.31)	(0.31)	(0.32)	(0.32)	(0.32)	(0.48)
200	0.1	0.1	5.84	5.52	6.00	5.02	5.56	0.90
			(0.33)	(0.32)	(0.34)	(0.31)	(0.32)	(0.13)
200	1	0.8	5.06	5.38	5.08	5.40	5.28	5.68
			(0.31)	(0.32)	(0.31)	(0.32)	(0.32)	(0.33)
200	1	0.5	4.64	5.34	4.68	5.36	4.94	4.96
			(0.30)	(0.32)	(0.30)	(0.32)	(0.31)	(0.31)
200	1	0.1	5.32	5.52	5.34	5.10	5.46	5.10
			(0.32)	(0.32)	(0.32)	(0.31)	(0.32)	(0.31)
B. 2SLS *with arbitrary heteroskedasticity*								
N	π	ρ	CLR	AR	LM	J	LM-J	Wald
200	0.1	0.8	6.34	6.68	6.08	6.42	6.16	36.66
			(0.34)	(0.35)	(0.34)	(0.35)	(0.34)	(0.68)
200	0.1	0.5	6.60	6.72	6.18	6.58	6.22	11.60
			(0.35)	(0.35)	(0.34)	(0.35)	(0.34)	(0.45)
200	0.1	0.1	6.80	6.46	6.30	6.44	6.56	0.84
			(0.36)	(0.35)	(0.34)	(0.35)	(0.35)	(0.13)
200	1	0.8	6.26	6.84	6.22	5.92	6.76	6.20
			(0.34)	(0.36)	(0.34)	(0.33)	(0.36)	(0.34)
200	1	0.5	5.70	6.46	5.72	6.36	6.42	5.38
			(0.33)	(0.35)	(0.33)	(0.35)	(0.35)	(0.32)
200	1	0.1	6.06	6.32	6.02	6.28	6.12	5.08
			(0.34)	(0.34)	(0.34)	(0.34)	(0.34)	(0.31)

Note: Simulation standard errors are in parentheses.

In table 2, we present the result from some Monte Carlo simulations for the linear IV model when the errors have intracluster dependence.[6] We experimented with different combinations of overall sample sizes (N), number of clusters (G), and resulting cluster sizes (M_g). In general, asymptotics related to cluster–robust variance–covariance estimation apply only to the case when the cluster sample sizes are small and the num-

6. Within clusters, errors were drawn from a multivariate normal distribution with a nondiagonal covariance matrix. The off-diagonal blocks are multiplied by the cross-equation correlation coefficient. Across clusters, the errors are independent.

ber of clusters goes to infinity. In our simulations, we find that this is true for the weak-instrument robust tests as well.

Table 2. Size (in percent) for testing $H_0 : \beta = 0.5$ at the 5% significance level in the linear IV model with intracluster-dependent errors

Models Simulation parameters					Test size Rejection rate for tests (percent)					
N	G	M_g	π	ρ	CLR	AR	LM	J	LM-J	Wald
400	100	4	0.1	0.8	6.44	6.52	5.96	6.20	6.28	1.12
					(0.35)	(0.35)	(0.33)	(0.34)	(0.34)	(0.15)
400	100	4	0.1	0.5	6.88	7.08	6.34	6.20	6.66	1.10
					(0.36)	(0.36)	(0.34)	(0.34)	(0.35)	(0.14)
400	100	4	0.1	0.1	6.82	7.20	6.42	6.46	6.54	0.98
					(0.36)	(0.37)	(0.35)	(0.35)	(0.35)	(0.14)
400	100	4	1	0.8	6.30	7.16	6.22	6.76	6.76	4.78
					(0.34)	(0.36)	(0.34)	(0.36)	(0.36)	(0.30)
400	100	4	1	0.5	5.98	7.16	5.96	6.84	6.22	4.86
					(0.34)	(0.36)	(0.33)	(0.36)	(0.34)	(0.30)
400	100	4	1	0.1	6.26	7.08	6.22	6.48	6.90	4.94
					(0.34)	(0.36)	(0.34)	(0.35)	(0.36)	(0.31)
500	50	10	0.1	0.8	8.46	8.74	7.68	7.18	8.64	1.50
					(0.39)	(0.40)	(0.38)	(0.37)	(0.40)	(0.17)
500	50	10	0.1	0.5	7.88	8.04	7.26	6.94	7.94	1.12
					(0.38)	(0.38)	(0.37)	(0.36)	(0.38)	(0.15)
500	50	10	0.1	0.1	8.56	8.72	7.66	7.46	8.62	1.38
					(0.40)	(0.40)	(0.38)	(0.37)	(0.40)	(0.14)
500	50	10	1	0.8	6.90	8.50	6.92	7.90	7.74	4.62
					(0.36)	(0.39)	(0.36)	(0.38)	(0.38)	(0.30)
500	50	10	1	0.5	6.98	8.40	6.98	7.38	7.98	5.05
					(0.36)	(0.39)	(0.36)	(0.37)	(0.38)	(0.31)
500	50	10	1	0.1	7.82	8.94	7.86	7.66	8.70	4.98
					(0.38)	(0.40)	(0.38)	(0.38)	(0.40)	(0.31)

Note: Simulation standard errors are in parentheses.

In the first six simulations, with 400 observations split into 100 clusters, the weak-instrument robust tests slightly overreject the null hypothesis, having a nominal size between 5% and 8%. This holds with weak or nonweak instruments. The Wald test, however, has a less predictable pattern; it consistently underrejects when instruments are weak but has the correct size when instruments are not weak. In the second six simulations, with 500 observations split into 50 clusters (an example consistent with many applications that use cross-sectional data from U.S. states), the weak-instrument robust tests also overreject, but their performance is still closer to the correct size than the Wald tests when instruments are weak.

We also conducted simulations with larger and smaller numbers of clusters and different numbers of observations within cluster. We found that the number of clusters is the most important element in determining the rejection probability of the tests. The overrejection decreases as the number of clusters increases.[7] We recommend bootstrapping the test to find appropriate critical values when the number of clusters is small (less than 50). A discussion of techniques that work well in the single equation linear model can be found in Cameron, Gelbach, and Miller (2008).

In table 3, we present the results from Monte Carlo simulations for the endogenous probit and tobit models (panels A and B, respectively). To avoid having to rescale the maximum likelihood test in the endogenous probit model, we let the population parameter, β, equal zero.[8]

7. Results are available upon request.
8. When $\beta = 0$, we have $\beta/\sigma_u = \beta/\sigma_\varepsilon = 0$ for positive values of σ_u and σ_ε.

Table 3. Size (in percent) for testing $H_0 : \beta = 0$ at the 5% significance level in the endogenous probit model and $H_0 : \beta = 0.5$ at the 5% significance level in the endogenous tobit model

Models Simulation parameters			Test size Rejection rate for tests (percent)					
A. *IV probit* ($\beta = 0$)								
N	π	ρ	CLR	AR	LM	J	LM-J	Wald
200	0.1	0.8	3.58	3.52	4.59	4.07	4.01	32.95
			(0.26)	(0.26)	(0.30)	(0.28)	(0.28)	(0.67)
200	0.1	0.5	3.99	3.93	5.03	4.49	4.77	41.94
			(0.28)	(0.28)	(0.31)	(0.29)	(0.30)	(0.70)
200	0.1	0.1	4.90	4.70	5.24	4.68	4.90	45.17
			(0.31)	(0.30)	(0.32)	(0.30)	(0.31)	(0.70)
200	1	0.8	3.94	3.88	3.96	4.72	3.82	5.12
			(0.28)	(0.27)	(0.28)	(0.30)	(0.27)	(0.31)
200	1	0.5	4.68	4.88	4.66	4.90	4.38	5.68
			(0.30)	(0.30)	(0.30)	(0.31)	(0.29)	(0.33)
200	1	0.1	5.24	5.10	5.26	5.32	5.16	6.18
			(0.32)	(0.31)	(0.32)	(0.32)	(0.31)	(0.34)
B. *IV tobit*								
N	π	ρ	CLR	AR	LM	J	LM-J	Wald
200	0.1	0.8	5.18	5.38	5.24	5.16	5.06	18.10
			(0.31)	(0.32)	(0.32)	(0.31)	(0.31)	(0.54)
200	0.1	0.5	5.34	5.50	5.16	5.44	5.24	7.20
			(0.32)	(0.32)	(0.31)	(0.32)	(0.32)	(0.37)
200	0.1	0.1	6.28	5.86	6.02	5.36	6.10	0.74
			(0.34)	(0.33)	(0.34)	(0.32)	(0.34)	(0.12)
200	1	0.8	5.12	5.22	5.10	5.40	5.22	5.14
			(0.31)	(0.31)	(0.31)	(0.32)	(0.31)	(0.31)
200	1	0.5	5.30	5.66	5.24	5.26	5.44	5.20
			(0.32)	(0.33)	(0.32)	(0.32)	(0.32)	(0.31)
200	1	0.1	5.16	5.84	5.26	5.72	5.26	5.04
			(0.31)	(0.33)	(0.32)	(0.33)	(0.32)	(0.31)

Note: Simulation standard errors are in parentheses.

(*Continued on next page*)

With any value of the simultaneity parameter, we find that the Wald test performs poorly when the instruments are weak ($\pi = 0.1$) in both the endogenous probit and tobit models. Surprisingly, the rejection probability for the Wald test in the endogenous probit model with weak instruments is above 30% independent of the degree of simultaneity, which contrasts with patterns observed in the linear IV and endogenous tobit models.[9] Regardless of the strength or weakness of the instruments, our tests are estimated to have rejection rates between 3.5% and 6.3%, close to the correct size of 5%.

7 Acknowlegments

We thank Mark Schaffer, Alan Barreca, Tom Palmer, and an anonymous referee for helpful comments, and the Tulane Research Enhancement Fund and the Committee on Research Summer Fellowship for funding.

8 References

Anderson, T. W., 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.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock. 2007. Performance of conditional Wald tests in IV regression with weak instruments. *Journal of Econometrics* 139: 116–132.

Baum, C. F., M. E. Schaffer, and S. Stillman. 2007. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal* 7: 465–506.

Cameron, A. C., J. B. Gelbach, and D. L. Miller. 2008. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics* 90: 414–427.

Chernozhukov, V., and C. Hansen. 2008. The reduced form: A simple approach to inference with weak instruments. *Economics Letters* 100: 68–71.

Dufour, J.-M. 2003. Identification, weak instruments and statistical inference in econometrics. Working Paper 2003s-49, Scientific Series, CIRANO. http://www.cirano.qc.ca/pdf/publication/2003s-49.pdf.

9. With weak instruments, the endogenous probit models have a difficult time converging. For example, for the weak-instrument ($\delta = 0.1$) simulations shown in table 3, 32 simulations did not converge under high simultaneity ($\rho = 0.8$), 10 simulations did not converge under moderate simultaneity ($\rho = 0.5$), and 1 simulation did not converge under low simultaneity ($\rho = 0.1$). In our experiments, the rejection rates for the weak-instrument robust tests were not affected by the proportion of simulations that did not converge. The Wald test, however, will reject more frequently under non-convergence. In table 3, we exclude test results for simulations that did not achieve convergence in the IV model, so the Wald rejection rates for IV probit with weak instruments should be considered lower bounds.

Hansen, L. P., J. Heaton, and A. Yaron. 1996. Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* 14: 262–280.

Kleibergen, F. 2007. Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. *Journal of Econometrics* 139: 181–216.

Magnusson, L. M. 2008a. Inference in limited dependent variable models robust to weak identification. Working Paper 0801, Department of Economics, Tulane University. http://ideas.repec.org/p/tul/wpaper/0801.html.

———. 2008b. Tests in censored models when the structural parameters are not identified. Working Paper 0802, Department of Economics, Tulane University. http://ideas.repec.org/p/tul/wpaper/0802.html.

Mikusheva, A. 2005. Robust confidence sets in the presence of weak instruments. Working Paper No. 07-27, Department of Economics, Massachusetts Institute of Technology. http://ssrn.com/abstract=1021366.

Mikusheva, A., and B. P. Poi. 2006. Tests and confidence sets with correct size when instruments are potentially weak. *Stata Journal* 6: 335–347.

Moreira, M. J. 2003. A conditional likelihood ratio test for structural models. *Econometrica* 71: 1027–1048.

Moreira, M. J., and B. P. Poi. 2003. Implementing tests with correct size in the simultaneous equations model. *Stata Journal* 3: 57–70.

Mroz, T. A. 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* 55: 765–799.

Stock, J. H., J. H. Wright, and M. Yogo. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20: 518–529.

About the authors

Keith Finlay and Leandro M. Magnusson are assistant professors in the Department of Economics at Tulane University.