SINGLE EQUATION LINEAR MODEL WITH CROSS-SECTIONAL DATA: CONTROL FUNCTIONS AND SPECIFICATION TESTING

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Control Function Approaches to Endogeneity
2. Correlated Random Coefficient Models
3. Testing for Endogeneity
4. Testing Overidentifying Restrictions
5. Labor Supply Application

1. CONTROL FUNCTION APPROACHES TO ENDOGENEITY

- Most models that are linear in parameters are estimated using two stage least squares (2SLS).
- An alternative, the control function (CF) approach, relies on the same kinds of identification conditions.
- Let y_1 be the response variable, y_2 the single endogenous explanatory variable (EEV), and $\mathbf{z}$ the $1 \times L$ vector of exogenous variables (with $z_1 = 1$):

$$y_1 = \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1, \tag{1}$$

where $\mathbf{z}_1$ is a $1 \times L_1$ strict subvector of $\mathbf{z}$.

- Consider the (weakest) exogeneity assumption

$$E(\mathbf{z}'u_1) = \mathbf{0}. \tag{2}$$

Reduced form for y_2:

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2,\ E(\mathbf{z}'v_2) = \mathbf{0} \tag{3}$$

where $\boldsymbol{\pi}_2$ is $L \times 1$. Write the linear projection of u_1 on v_2, in error form, as

$$u_1 = \rho_1 v_2 + e_1, \tag{4}$$

where $\rho_1 = E(v_2 u_1)/E(v_2^2)$ is the population regression coefficient. By construction, $E(v_2 e_1) = 0$ and $E(\mathbf{z}'e_1) = \mathbf{0}$.

• Plug (4) into (1):

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1, \tag{5}$$

where v_2 is an explanatory variable in the equation. The new error, e_1, is uncorrelated with y_2 as well as with v_2 and $\mathbf{z}$.

• Two-step procedure: (i) Regress y_{i2} on $\mathbf{z}_i$ and obtain the reduced form residuals, $\hat{v}_{i2}$; (ii) Regress

$$y_{i1} \text{ on } \mathbf{z}_{i1}, y_{i2}, \text{ and } \hat{v}_{i2}. \tag{6}$$

- Because we can write

$$y_{i1} = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \rho_1 \hat{v}_{i2} + e_{i1} + \rho_1 \mathbf{z}_i(\hat{\boldsymbol{\pi}}_2 - \boldsymbol{\pi}_2),$$

the error implicit in (6) is $e_{i1} + \rho_1 \mathbf{z}_i(\hat{\boldsymbol{\pi}}_2 - \boldsymbol{\pi}_2)$, which depends on the sampling error in $\hat{\boldsymbol{\pi}}_2$ unless $\rho_1 = 0$.

- Using results from Chapter 6 on two-step estimation, OLS estimators from (6) will be consistent for $\boldsymbol{\delta}_1, \alpha_1$, and ρ_1. Sometimes $\hat{v}_{i2} = y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$ is called a **generated regressor**.

• The OLS estimates from (6) are **control function** estimates.

• Using the Frisch-Waugh Theorem from OLS mechanics, the OLS estimates of $\boldsymbol{\delta}_1$ and α_1 from (6) can be shown to be *identical* to the 2SLS estimates starting from (1).

• Where does the CF estimator use the fact that $\mathbf{z}_i$ must contain at least one more element than $\mathbf{z}_{i1}$? Think of perfect collinearity in

$$y_{i1} = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \rho_1 \hat{v}_{i2} + error_i$$

• Now extend the model so that the EEV is in quadratic form:

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + u_1 \tag{7}$$

$$E(u_1|\mathbf{z}) = 0. \tag{8}$$

• Very difficult to get by without (8) once we include nonlinear functions in the model.

• Let z_2 be a non-binary scalar not also in $\mathbf{z}_1$. Under the (8) we can use, say nonlinear functions as IVs, say z_2^2 as an instrument for y_2^2. So the IVs would be $(\mathbf{z}_1, z_2, z_2^2)$ for $(\mathbf{z}_1, y_2, y_2^2)$.

- What does CF approach entail? We really need to impose much more on the reduced form; it is no longer just defined as a linear projection:

$$\begin{aligned} y_2 &= \mathbf{z}\boldsymbol{\pi}_2 + v_2 \\ E(v_2|\mathbf{z}) &= 0 \end{aligned}$$

which puts strong restrictions on $E(y_2|\mathbf{z})$.

- Further, *assume*

$$E(u_1|\mathbf{z}, y_2) = E(u_1|v_2) = \rho_1 v_2. \tag{9}$$

This has two parts. First, that $\mathbf{z}$ drops out of $E(u_1|\mathbf{z}, y_2)$. Independence of (u_1, v_2) and $\mathbf{z}$ is sufficient. Second, linearity of $E(u_1|v_2)$ is a real restriction.

- Under (9),

$$\begin{aligned} E(y_1|\mathbf{z}, y_2) &= E(y_1|\mathbf{z}, v_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + E(u_1|\mathbf{z}, v_2) \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + \rho_1 v_2. \end{aligned} \tag{10}$$

- A CF approach is immediate: OLS of

$$y_{i1} \text{ on } \mathbf{z}_{i1}, y_{i2},\ y_{i2}^2,\ \text{and } \hat{v}_{i2}. \tag{11}$$

- *Not* equivalent to a 2SLS estimate. If we use, say, IVs $(\mathbf{z}_{i1}, z_{i2}, z_{i2}^2)$ then the IV estimator is consistent under $E(u_1|\mathbf{z}) = 0$.
- CF accounts for endogeneity of y_2 and y_2^2 using a single control function, $\hat{v}_2$. CF is likely more efficient but definitely less robust.

2. CORRELATED RANDOM COEFFICIENT MODELS

- Modify the original equation as

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + a_1 y_2 + u_1, \tag{12}$$

where a_1, the “random coefficient” on y_2. Heckman and Vytlacil (1998) call (12) a **correlated random coefficient (CRC) model**. For emphasis,

$$y_{i1} = \eta_1 + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + a_{i1} y_{i2} + u_{i1} \tag{13}$$

- a_{i1} contains “ability” and “motivation”; y_{i2} is schooling. Return to schooling is individual-specific.

• In the population, write $a_1 = \alpha_1 + v_1$ where $\alpha_1 = E(a_1)$ is the object of interest: the **average partial effect** (**APE**). We can rewrite the equation as

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + v_1 y_2 + u_1 \tag{14}$$

$$\equiv \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + e_1. \tag{15}$$

where $e_1 = v_1 y_2 + u_1$. Generally, $E(e_1) = E(v_1 y_2) = Cov(v_1, y_2)$. Just having a nonzero unconditional mean is not much of a problem.

• The potential problem with applying instrumental variables is that the error term $e_1 = v_1 y_2 + u_1$ is not necessarily uncorrelated with the instruments $\mathbf{z}$, even with our maintained assumptions

$$E(u_1|z) = E(v_1|\mathbf{z}) = 0. \tag{16}$$

• We want to allow y_2 and v_1 to be correlated, $Cov(v_1, y_2) \equiv \tau_1 \neq 0$. A condition that still allows for any amount of *unconditional* correlation is

$$Cov(v_1, y_2|\mathbf{z}) = Cov(v_1, y_2), \tag{17}$$

and this is sufficient for 2SLS to consistently estimate $(\alpha_1, \boldsymbol{\delta}_1)$.

• Why is (17) sufficient? Because $E(v_1|\mathbf{z}) = 0$, $Cov(v_1, y_2|\mathbf{z}) = E(v_1 y_2|\mathbf{z})$. Therefore, if (17) holds, we can write

$$v_1 y_2 = \tau_1 + r_1 \tag{18}$$

$$E(r_1|\mathbf{z}) = 0. \tag{19}$$

So, the equation we estimate by usual 2SLS can be written as

$$y_1 = (\eta_1 + \tau_1) + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + (r_1 + u_1), \tag{20}$$

where by (16) and (19), $E(r_1 + u_1|\mathbf{z}) = 0$. Thus, the parameters in (20) are consistently estimated by 2SLS using IVs $\mathbf{z}$, which includes a constant.

• The original intercept, η_1, cannot be estimated.

- What would a control function approach look like? Write

$$\begin{aligned} y_2 &= \mathbf{z}\boldsymbol{\pi}_2 + v_2 && (21) \\ E(v_2|\mathbf{z}) &= 0. && (22) \end{aligned}$$

Add

$$E(u_1|\mathbf{z}, v_2) = \rho_1 v_2, \quad E(v_1|\mathbf{z}, v_2) = \xi_1 v_2. \tag{23}$$

Then

$$E(y_1|\mathbf{z},y_2) = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \xi_1 v_2 y_2 + \rho_1 v_2. \tag{24}$$

• Two-step method: (1) Regress y_2 on $\mathbf{z}$ to get $\hat{v}_2$ (residuals). (2) Run the OLS regression y_1 on $1, \mathbf{z}_1, y_2, \hat{v}_2 y_2, \hat{v}_2.$ Due to Garen (1984). Under the maintained assumptions, Garen's method consistently estimates $\boldsymbol{\delta}_1$ and α_1.

- Because the second step uses generated regressors, the standard errors should be adjusted for the estimation of $\boldsymbol{\pi}_2$ in the first stage.
- Garen relies on a linear model for $E(y_2|\mathbf{z})$. Further, Garen adds the assumptions that $E(u_1|v_2)$ and $E(v_1|v_2)$ are linear functions, something not needed by the IV approach.

3. TESTING FOR ENDOGENEITY

- In the general equation $y = \mathbf{x}\boldsymbol{\beta} + u$ with instruments $\mathbf{z}$, the **Durbin-Wu-Hausman** (**DWH**) test is based on the difference $\hat{\boldsymbol{\beta}}_{2SLS} - \hat{\boldsymbol{\beta}}_{OLS}$. If all elements of $\mathbf{x}$ are exogenous (and $\mathbf{z}$ is also exogenous – a maintained assumption), then 2SLS and OLS should differ only due to sampling error.
- Do not just blindly compute a test statistic. Are the differences in OLS and 2SLS practically important?

• The general approach suggested by Hausman (1978, *Econometrica*) maintains that one of the estimators is relatively (asymptotically) efficient under the null. In this case, under the null that $\mathbf{x}$ is exogenous (and $\mathbf{z}$, too), OLS is asymptotically efficient provided we add the homoskedasticity assumption

$$E(u^2\mathbf{w}'\mathbf{w}) = \sigma^2 E(\mathbf{w}'\mathbf{w})$$

where $\mathbf{w}$ is all nonredundant elements of $(\mathbf{x}, \mathbf{z})$.

• But it is important to know that the approach makes sense whenever both estimators are consistent under the null and at least on is inconsistent under the alternative.

• It makes no sense to make inference on $\boldsymbol{\beta}$ using, say, OLS robust to general heteroskedasticity and then assume homoskedasticity when obtaining a Hausman test. The traditional Hausman test that compares 2SLS and OLS does not have a limiting chi-square distribution when heteroskedasticity is present. Yet it has no systematic power for detecting heteroskedasticity.

• If in addition to $E(\mathbf{x}'u) = \mathbf{0}$, $E(\mathbf{z}'u) = \mathbf{0}$, the rank conditions for OLS and 2SLS, and the homoskedasticity assumption $E(u^2\mathbf{w}'\mathbf{w}) = \sigma^2 E(\mathbf{w}'\mathbf{w})$ (under the null), then

$$Avar[\sqrt{N}(\hat{\boldsymbol{\beta}}_{2SLS} - \hat{\boldsymbol{\beta}}_{OLS})] = \sigma^2[E(\mathbf{x}^{*\prime}\mathbf{x}^*)]^{-1} - \sigma^2[E(\mathbf{x}'\mathbf{x})]^{-1}, \tag{25}$$

which is simply the difference between the asymptotic variances.

• Equation (25) is also the basis for showing 2SLS is asymptotically less efficient than OLS under OLS.1, OLS.2, OLS.3, and the corresponding 2SLS assumptions.

- One version of the DWH statistic uses the OLS estimate for σ^2:

$$(\hat{\boldsymbol{\beta}}_{2SLS} - \hat{\boldsymbol{\beta}}_{OLS})'[(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-}(\hat{\boldsymbol{\beta}}_{2SLS} - \hat{\boldsymbol{\beta}}_{OLS})/\hat{\sigma}^2_{OLS}, \tag{26}$$

where we must use a generalized inverse, except in the very unusual case that all elements of **x** are allowed to be endogenous under the alternative.

- The rank of $Avar[\sqrt{N}(\hat{\boldsymbol{\beta}}_{2SLS} - \hat{\boldsymbol{\beta}}_{OLS})]$ is equal to the number of elements of **x** allowed to be endogenous under the alternative. The singularity of the matrix in (26) makes computing the statistic cumbersome.

• Not surprising, the statistic in (26) is not robust to heteroskedasticity. A robust variance matrix estimator for $Avar[\sqrt{N}(\hat{\boldsymbol{\beta}}_{2SLS} - \hat{\boldsymbol{\beta}}_{OLS})]$ can be obtained, but not easily.

• With only a single suspected endogenous explanatory variable y_2, a Hausman t statistic can be used to determine whether y_2 is endogenous:

$$(\hat{\alpha}_{1,2SLS} - \hat{\alpha}_{1,OLS})/\{[se(\hat{\alpha}_{1,2SLS})]^2 - [se(\hat{\alpha}_{1,OLS})]^2\}^{1/2} \tag{27}$$

Under the null hypothesis, the t statistic has an asymptotically standard normal distribution.

• Unfortunately, there is no simple correction if one allows heteroskedasticity: the asymptotic variance of the difference is no longer the difference in asymptotic variances.

• A regression-based Hausm test uses the control function approach. Write

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \mathbf{y}_2\boldsymbol{\alpha}_1 + u_1, \tag{28}$$

where $\mathbf{z}_1$ is $1 \times L_1$, $\mathbf{y}_2$ is $1 \times G_1$, and the entire vector of all instruments is $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$, where $\mathbf{z}_2$ is $1 \times L_2$ with $L_2 \geq G_1$. The two-step procedure is

(i) Regress $\mathbf{y}_{i2}$ on $\mathbf{z}_i$ to obtain the $1 \times G_1$ reduced form residuals, $\hat{\mathbf{v}}_{i2}$ (one vector for each observation).

(ii) Run the regression

$$y_{i1} \text{ on } \mathbf{z}_{i1}, \mathbf{y}_{i2}, \hat{\mathbf{v}}_{i2} \tag{29}$$

and use a joint Wald test of $H_0 : \boldsymbol{\rho}_1 = \mathbf{0}$, where $\boldsymbol{\rho}_1$ is the vector of coefficients on $\hat{\mathbf{v}}_{i2}$. (This is often computed as an approximate F statistic by dividing the Wald statistic by G_1, the number of restrictions being tested.)

- The test need not be adjusted for the first-stage estimation (generated regressors, $\hat{\mathbf{v}}_{i2}$), and it is easily made robust to heteroskedasticity of unknown form.

• Sometimes we may want to test the null hypothesis that a subset of explanatory variables is exogenous while allowing another set of variables to be endogenous. Write an expanded model as

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \mathbf{y}_2\boldsymbol{\alpha}_1 + \mathbf{y}_3\boldsymbol{\gamma}_1 + u_1, \tag{30}$$

where $\boldsymbol{\alpha}_1$ is $G_1 \times 1$ and $\boldsymbol{\gamma}_1$ is $J_1 \times 1$. We allow $\mathbf{y}_2$ to be endogenous and test $H_0 : E(\mathbf{y}_3' u_1) = \mathbf{0}$. The relevant equation is now $y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \mathbf{y}_2\boldsymbol{\alpha}_1 + \mathbf{y}_3\boldsymbol{\gamma}_1 + \mathbf{v}_3\boldsymbol{\rho}_1 + e_1$, or, when we operationalize it,

$$y_{i1} = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \mathbf{y}_{i2}\boldsymbol{\alpha}_1 + \mathbf{y}_{i3}\boldsymbol{\gamma}_1 + \hat{\mathbf{v}}_{i3}\boldsymbol{\rho}_1 + error_i. \tag{31}$$

- Because $\mathbf{y}_2$ is allowed to be endogenous under H_0, we cannot estimate (31) by OLS in order to test $\mathrm{H}_0 : \boldsymbol{\rho}_1 = \mathbf{0}$. Instead, we apply 2SLS to (31) with instruments $(\mathbf{z}_i, \mathbf{y}_{i3}, \hat{\mathbf{v}}_{i3})$; remember, $(\mathbf{y}_3, \mathbf{v}_3)$ are exogenous in the augmented equation. In effect, we still instrument for $\mathbf{y}_{i2}$ but $\mathbf{y}_{i3}$ and $\hat{\mathbf{v}}_{i3}$ act as their own instruments.
- The usual Wald statistic for 2SLS (possibly implemented as an F-type statistic) for testing $\mathrm{H}_0 : \boldsymbol{\rho}_1 = \mathbf{0}$ is asymptotically valid under H_0. As usual, it may be prudent to allow heteroskedasticity of unknown form under H_0, which is easily done in many software packages.

Question: What would a test for the null of y_2 exogenous look like for the CRC model? Remember, under

$$\begin{aligned} y_2 &= \mathbf{z}\boldsymbol{\pi}_2 + v_2 \\ E(v_2|\mathbf{z}) &= 0. \end{aligned}$$

$$E(u_1|\mathbf{z}, v_2) = \rho_1 v_2,\;\; E(v_1|\mathbf{z}, v_2) = \xi_1 v_2$$

we derived

$$E(y_1|\mathbf{z}, v_2) = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \xi_1 v_2 y_2 + \rho_1 v_2.$$

Solution: First, regress y_{i2} on $\mathbf{z}_i$ and get the OLS residuals, $\hat{v}_{i2}$. Then, test $H_0 : \xi_1 = 0, \rho_1 = 0$ using OLS on

$$y_{i1} = \eta_1 + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \xi_1 \hat{v}_{i2} y_{i2} + \rho_1 \hat{v}_{i2} + error_i$$

• Under the null hypothesis, the generated regressors problem does not matter asymptotically. Can use a heteroskedasticity-robust Wald test.

4. TESTING OVERIDENTIFYING RESTRICTIONS

• If we have more instruments than we need we can, in a (weak) sense, test whether some of them are exogenous. Write the equation as

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \mathbf{y}_2\boldsymbol{\alpha}_1 + u_1 \tag{32}$$

where $\mathbf{z}_1$ is $1 \times L_1$ and $\mathbf{y}_2$ is $1 \times G_1$. The entire vector of instruments is $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$, where $\mathbf{z}_2$ is $1 \times L_2$. the equation is overidentified if $L_2 > G_1$.

• The 2SLS estimator uses $L_1 + G_1$ moment conditions, so $L_2 - G_1$ overidentifying restrictions can be tested.

- A traditional version of the Hausman test, under the 2SLS homoskedasticity assumption, directly compares the 2SLS estimator using all instruments to a just identified IV estimator. Turns out not to matter which just identified IV estimator we use.
- In the case of, say, a scalar y_2 and two elements in $\mathbf{z}_2 = (z_{21}, z_{22})$, can directly compare the two IV estimators using each IV in turn (but neither is relatively efficient, so computation is not straightforward).

EXAMPLE: $y_2 = educ$ and $\mathbf{z}_2 = (motheduc, fatheduc)$. Problem is the test will have weak power if the two IV estimators are biased in a similar way (likely in this example).

• In other words, a failure to reject should not make us too confident. A rejection indicates that one or both IVs fail the exogeneity requirement; we do not know which one or whether it is both.

• Again, regression-based tests are convenient. Under homoskedasticity, 2SLS.3, obtain NR_u^2 (generally, the uncentered *R*-squared, but almost always the usual *R*-squared) from

$$\hat{u}_{i1} \text{ on } \mathbf{z}_i, \tag{33}$$

where $\hat{u}_{i1}$ are the 2SLS residuals and $\mathbf{z}$ is the vector of all exogenous variables.

- The motivation for (33) is the sample moment conditions

$$N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\hat{u}_{i1} \approx \mathbf{0} \tag{34}$$

under the null. But we also know $K_1 = L_1 + G_1$ exact moment conditions hold in the sample,

$$N^{-1}\sum_{i=1}^{N}(\mathbf{z}_i\hat{\mathbf{\Pi}}_1)'\hat{u}_{i1} = \mathbf{0}, \tag{35}$$

where $\hat{\mathbf{\Pi}}_1$ is the $L \times K_1$ matrix from $\mathbf{x}_1$ on $\mathbf{z}$, so there are not as many degrees-of-freedom as (34) seems to suggest.

- Under the null hypothesis

$$E(\mathbf{z}'u) = \mathbf{0} \tag{36}$$

$$E(u^2\mathbf{z}'\mathbf{z}) = \sigma^2 E(\mathbf{z}'\mathbf{z}) \tag{37}$$

it can be shown

$$NR_u^2 \overset{a}{\sim} \chi^2_{L_2-G_1}. \tag{38}$$

- Easy to compute, but not robust to heteroskedasticity.
- The test has the wrong asymptotic size if (37) fails, but the test has no systematic power for detecting failure of (37).

• A heteroskedasticity-robust form requires a little more work. Separate the instrumental variables into two groups. Let $\mathbf{z}_2$ be the $1 \times L_2$ vector of exogenous variables excluded from (32) and write $\mathbf{z}_2 = (\mathbf{g}_2, \mathbf{h}_2)$, where $\mathbf{g}_2$ is $1 \times G_1$ – the same dimension as $\mathbf{y}_2$ – and $\mathbf{h}_2$ is $1 \times Q_1$ – the number of overidentifying restrictions.

• Provided $\mathbf{h}_2$ has Q_1 elements it matters not how it is chosen.

• Now, we need the 2SLS residuals, $\hat{u}_1$, as before, but we also need the fitted values $\hat{\mathbf{y}}_2$ from the first-stage regression.

• We partial out $\hat{\mathbf{y}}_2$ from each element of $\mathbf{h}_2$. So, run a multivariate regression of $\mathbf{h}_2$ on $\hat{\mathbf{y}}_2$ and obtain the residuals, $\hat{\mathbf{r}}_2$ (so Q_1 residuals for each observation).

• Run the regression

$$\hat{u}_1 \text{ on } \hat{\mathbf{r}}_2$$

(without a constant) and compute a heteroskedasticity-robust Wald test that all coefficients on $\hat{\mathbf{r}}_2$ are zero.

5. LABOR SUPPLY APPLICATION

```
. use C:\mitbook1_2e\statafiles\labsup.dta

. * data are for black or Hispanic females

. des hours nonmomi kids educ age black hispan samesex

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
hours           byte   %8.0g                  hours of work per week, mom
nonmomi         float  %9.0g                  ’non-mom’ income, $1000s
kids            byte   %8.0g                  number of kids
educ            byte   %8.0g                  mom’s years of education
age             byte   %8.0g                  age of mom
black           byte   %8.0g                  =1 of black
hispan          byte   %8.0g                  =1 if hispanic
samesex         byte   %8.0g                  first two kids are of same sex
```

```
. sum hours nonmomi kids educ age black hispan

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       hours |     31857    21.22011    19.49892          0         99
     nonmomi |     31857     31.7618    20.41241  -39.93675    157.438
        kids |     31857    2.752237    .9771916          2         12
        educ |     31857    11.00534    3.305196          0         20
         age |     31857    29.74175    3.613745         21         35
-------------+--------------------------------------------------------
       black |     31857    .4129705    .4923753          0          1
      hispan |     31857     .593182    .4912481          0          1
```

```
. * First use OLS to estimate the effects of children on hours worked:

. reg hours kids nonmomi educ age agesq black hispan, robust

Linear regression                                      Number of obs =   31857
                                                       F(  7, 31849) =  377.87
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0727
                                                       Root MSE      =  18.779

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -2.325836   .1155164   -20.13   0.000    -2.552253   -2.099419
     nonmomi |  -.0578328   .0053515   -10.81   0.000     -.068322   -.0473436
        educ |   .5860083   .0374881    15.63   0.000     .5125302    .6594865
         age |   2.048793   .4483823     4.57   0.000     1.169946    2.927639
       agesq |  -.0277198   .0076957    -3.60   0.000    -.0428036    -.012636
       black |   1.058285    1.35088     0.78   0.433    -1.589492    3.706063
      hispan |  -5.114147    1.35152    -3.78   0.000    -7.763179   -2.465116
       _cons |  -10.44695   6.588891    -1.59   0.113    -23.36143    2.467528
------------------------------------------------------------------------------
```

```
. * Now use samesex and multi2nd as IVs for kids.

. * Estimate the reduced form:

. reg kids samesex multi2nd nonmomi educ age agesq black hispan, robust

Linear regression                                      Number of obs =   31857
                                                       F(  8, 31848) =  410.77
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1244
                                                       Root MSE      =  .91452

------------------------------------------------------------------------------
             |               Robust
        kids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |     .07044   .0102481     6.87   0.000     .0503533    .0905267
    multi2nd |   .7632484   .0546856    13.96   0.000     .6560626    .8704342
     nonmomi |  -.0027879   .0002562   -10.88   0.000    -.0032901   -.0022858
        educ |  -.0853114   .0020267   -42.09   0.000    -.0892838   -.0813391
         age |   .0563395    .020282     2.78   0.005      .016586    .0960929
       agesq |   .0000436   .0003551     0.12   0.902    -.0006524    .0007396
       black |   .0105681   .0645589     0.16   0.870    -.1159698    .1371059
      hispan |  -.0420447   .0646128    -0.65   0.515    -.1686882    .0845988
       _cons |   2.043467   .2924263     6.99   0.000       1.4703    2.616634
------------------------------------------------------------------------------
```

```
. test samesex multi2nd

 ( 1)  samesex = 0
 ( 2)  multi2nd = 0

       F(  2, 31848) =  117.38
            Prob > F =    0.0000

. * Clearly the two IV candidates are partially correlated with kids,
. * both in the direction (positive) that we expect.

. * Get the reduced form residuals.

. predict v2h, resid
```

```
. * Test the null that kids is exogenous in the hours equation:

. reg hours kids nonmomi educ age agesq black hispan v2h, robust

Linear regression                                      Number of obs =   31857
                                                       F(  8, 31848) =  330.79
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0727
                                                       Root MSE      =  18.779

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -2.986165   1.284302    -2.33   0.020    -5.503447   -.4688828
     nonmomi |  -.0596653   .0064263    -9.28   0.000     -.072261   -.0470696
        educ |   .5296332   .1154311     4.59   0.000     .3033839    .7558825
         age |    2.08815   .4545537     4.59   0.000     1.197208    2.979093
       agesq |  -.0277261   .0076958    -3.60   0.000    -.0428101   -.0126422
       black |   1.067778   1.350595     0.79   0.429     -1.57944    3.714995
      hispan |  -5.140945   1.352129    -3.80   0.000    -7.791169   -2.490721
         v2h |    .665256   1.290263     0.52   0.606     -1.86371    3.194222
       _cons |  -9.103833   7.093029    -1.28   0.199    -23.00644    4.798776
------------------------------------------------------------------------------

. * The test statistic is only about .52, so there is little evidence that kids
. * is endogenous.
```

```
. * Now compute the 2SLS estimates:

. ivreg hours nonmomi educ age agesq black hispan (kids = samesex multi2nd),
        robust

Instrumental variables (2SLS) regression               Number of obs =   31857
                                                       F(  7, 31849) =  310.81
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0717
                                                       Root MSE      =  18.789

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -2.986165    1.28219    -2.33   0.020    -5.499307    -.473022
     nonmomi |  -.0596653   .0064235    -9.29   0.000    -.0722555   -.0470751
        educ |   .5296332   .1152961     4.59   0.000     .3036484     .755618
         age |    2.08815   .4545798     4.59   0.000     1.197156    2.979144
       agesq |  -.0277261   .0076979    -3.60   0.000    -.0428143    -.012638
       black |   1.067778   1.355563     0.79   0.431    -1.589178    3.724733
      hispan |  -5.140945   1.357096    -3.79   0.000    -7.800906   -2.480985
       _cons |  -9.103834   7.092956    -1.28   0.199     -23.0063    4.798632
------------------------------------------------------------------------------
Instrumented:  kids
Instruments:   nonmomi educ age agesq black hispan samesex multi2nd
------------------------------------------------------------------------------

. * Note that these are the same as the CF estimates.
```

```
. predict u1h, resid

. * Test the single overidentifying restriction using nonrobust test:

. reg u1h samesex multi2nd nonmomi educ age agesq black hispan

      Source |       SS       df       MS              Number of obs =   31857
-------------+------------------------------           F(  8, 31848) =    0.06
       Model |  176.258976     8   22.032372           Prob > F      =  0.9999
    Residual |  11242898.1 31848  353.017398           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0002
       Total |  11243074.3 31856  352.934277           Root MSE      =  18.789

------------------------------------------------------------------------------
         u1h |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |  -.1331695   .2105507    -0.63   0.527    -.5458569    .2795179
    multi2nd |    .357619   1.136161     0.31   0.753    -1.869301    2.584539
     nonmomi |   .0000221   .0053906     0.00   0.997    -.0105436    .0105879
        educ |   .0000136   .0353226     0.00   1.000      -.06922    .0692472
         age |   .0000577   .4481451     0.00   1.000    -.8783239    .8784393
       agesq |  -2.46e-06   .0077015    -0.00   1.000    -.0150978    .0150929
       black |   .0017749     1.3505     0.00   0.999    -2.645257    2.648807
      hispan |   .0037765   1.352616     0.00   0.998    -2.647404    2.654957
       _cons |   .0605262     6.5755     0.01   0.993    -12.82771    12.94876
------------------------------------------------------------------------------
```

```
. * R-squared is zero to four decimal places, but N is large.

. di e(N)*e(r2)
.49942587

. di chi2tail(1,.499)
.47993984

. * So the p-value is about .48, showing little evidence against the
. * overidentifying restriction
```

```
. * Now compute the heteroskedasticity-robust test.

. qui reg kids samesex multi2nd nonmomi educ age agesq black hispan

. predict kidsh
(option xb assumed; fitted values)

. qui reg samesex kidsh nonmomi educ age agesq black hispan

. predict r21h, resid

. qui reg multi2nd kidsh nonmomi educ age agesq black hispan

. predict r22h, resid

. reg u1h r21h, nocons robust

Linear regression                                      Number of obs =   31857
                                                       F(  1, 31856) =    0.51
                                                       Prob > F      =  0.4767
                                                       R-squared     =  0.0000
                                                       Root MSE      =  18.786

------------------------------------------------------------------------------
             |               Robust
         u1h |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        r21h |   -.166174   .2335323    -0.71   0.477    -.6239062    .2915583
------------------------------------------------------------------------------
```

```
. reg u1h r22h, nocons robust

Linear regression                                      Number of obs =   31857
                                                       F(  1, 31856) =    0.51
                                                       Prob > F      =  0.4767
                                                       R-squared     =  0.0000
                                                       Root MSE      =  18.786

------------------------------------------------------------------------------
             |               Robust
         u1h |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        r22h |   1.800574   2.530425     0.71   0.477    -3.159156    6.760305
------------------------------------------------------------------------------

. * Get the same answer since only the absolute value of the t matters.
. * Equivalently, use the F statistic reported in the upper right-hand
. * corner.
```