# SINGLE EQUATION LINEAR MODEL WITH CROSS-SECTIONAL DATA: IV

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

1. The Instrumental Variables Estimator in the Simple Model
2. IV Estimation of a General Equation
3. Two Stage Least Squares
4. Application: Endogeneity of Children in Labor Supply

## 1. THE INSTRUMENTAL VARIABLES ESTIMATOR IN THE SIMPLE MODEL

- Simple linear model in the population:

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

where $u$ is thought to be correlated with $x$ (which can have any kind of features – discrete, continuous, or hybrid). If $Cov(x,u) \neq 0$ ($x$ is "endogenous") then ordinary least squares (OLS) will be inconsistent for $\beta_1$.

- If we have a rich set of controls, we might be aboe to break the link between $u$ and $x$ – but that is an OLS solution.

- An instrumental variable, $z$, for $w$ has two properties:

$$Cov(z,u) = 0 \quad \text{(exogeneity)} \tag{2}$$

$$Cov(z,x) \neq 0 \quad \text{(relevance)} \tag{3}$$

A key difference is that we must take (2) on faith (or have other ways of checking it); we can always test the null that $z$ and $x$ are uncorrelated given a sample of data (and hope to reject the null).

- Under (2) and (3), we can show

$$\beta_1 = \frac{Cov(z,y)}{Cov(z,x)}, \tag{4}$$

which establishes that $\beta_1$ is identified. Replacing the population covariances with the sample covariances gives us the so-called instrumental variables estimator for the simple regression model.

$$\hat{\beta}_{1,IV} = \frac{\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})}.$$

- IV estimator is not unbiased when it is actually needed.

• Even if we restrict attention to consistency, it is not true that one should use a "slightly" endogenous instrument rather than OLS.

It is easy to show:

$$\text{plim}\ \hat{\beta}_{1,OLS} = \beta_1 + \frac{\sigma_u}{\sigma_x} \cdot Corr(x,u) \tag{5}$$

and

$$\text{plim}\ \hat{\beta}_{1,IV} = \beta_1 + \frac{\sigma_u}{\sigma_x} \cdot \frac{Corr(z,u)}{Corr(z,x)} \tag{6}$$

- If $Corr(z, x)$ is small – that is, $z$ is a “weak” instrument – then even a small correlation between $z$ and $u$ can produce a larger asymptotic bias than OLS. [In economics, very common to see IV estimates that are larger in magnitude than OLS estimates.] And weak instruments lead to large asymptotic standard errors, too.

Under a homoskedasticity assumption (to be made precise later),

$$Avar\sqrt{N}\,(\hat{\beta}_{1,IV} - \beta_1) = \frac{\sigma_u^2}{\sigma_x^2 \rho_{z,x}^2}. \tag{7}$$

When $\rho_{z,w}^2$ is small, the asymptotic variance can be very large. The formula for the OLS estimator omits $\rho_{z,w}^2$.

• Weak instruments also make the usual asymptotic approximations highly suspect, even with large sample sizes. [See Bound, Jaeger, Baker (1995); Staiger and Stock (1997).]

- EXAMPLE: Estimating the return to schooling via simple regression:

$$\log(wage) = \beta_0 + \beta_1 educ + u \tag{8}$$

Suggestions for $z$:

(i) mother's education

(ii) number of siblings

(iii) distance to the nearest college at age 16

(iv) last digit of Social Security Number

(v) $z$ binary, $z = 1$ if born in first quarter of year

## 2. IV ESTIMATION OF A GENERAL EQUATION

- Start again with the population model

$$y = \mathbf{x}\boldsymbol{\beta} + u, \tag{9}$$

where $\mathbf{x}$ is $1 \times K$, $\boldsymbol{\beta}$ is $K \times 1$, and in the vast majority of cases, $x_1 = 1$. This is the same *model* that we *estimated* by OLS.

- No such thing as an OLS "model" or IV "model." OLS and IV are different *estimation* methods that can be applied to the same model. They are consistent under different assumptions.

- Let $\mathbf{z} = (z_1, z_2, \ldots, z_L)$ be a $1 \times L$ vector, where $z_1 = 1$ almost always. Further, $\mathbf{z}$ contains all exogenous elements of $\mathbf{x}$. But, if one or more elements of $\mathbf{x}$ is correlated with $u$, $\mathbf{z}$ must contain some outside variables.
- $\mathbf{z}$ is exogenous;

$$E(\mathbf{z}'u) = \mathbf{0}. \tag{10}$$

How does this assumption – these moment or orthogonality conditions – allow us to identify $\boldsymbol{\beta}$?

- Suppose $L = K$. For example, $\mathbf{x} = (1, x_2, \ldots, x_{K-1}, x_K)$ and $\mathbf{z} = (1, x_2, \ldots, x_{K-1}, z_1)$, so that only $x_K$ is (possibly endogenous) and $z_1$ as an IV for $x_K$. Using (2.1) and (2.2),

$$\begin{aligned} E(\mathbf{z}'y) &= E(\mathbf{z}'\mathbf{x})\boldsymbol{\beta} + E(\mathbf{z}'u) \\ &= E(\mathbf{z}'\mathbf{x})\boldsymbol{\beta} \text{ by (10).} \end{aligned} \tag{11}$$

If we assume the *rank condition*

$$rank\, E(\mathbf{z}'\mathbf{x}) = K \tag{12}$$

then

$$\boldsymbol{\beta} = [E(\mathbf{z}'\mathbf{x})]^{-1}E(\mathbf{z}'y), \tag{13}$$

which extends the moment condition for OLS (which is the special case $\mathbf{z} = \mathbf{x}$).

- Give a random sample,

$$\hat{\boldsymbol{\beta}}_{IV} = \left( N^{-1} \sum_{i=1}^{N} \mathbf{z}_i' \mathbf{x}_i \right)^{-1} \left( N^{-1} \sum_{i=1}^{N} \mathbf{z}_i' y_i \right) \tag{14}$$

$$= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y}, \tag{15}$$

and $\text{plim}(\hat{\boldsymbol{\beta}}_{IV}) = \boldsymbol{\beta}$ under (10) and (12).

• How can we test the rank condition? Difficult in general, but with a single endogenous explanatory variable, easy. Write the *reduced form* of $x_K$ as

$$x_K = \delta_1 + \delta_2 x_2 + \ldots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_k \tag{16}$$

where, by *definition*,

$$E(r_K) = 0,\ Cov(x_j, r_K) = 0,\ j = 2, \ldots, K-1,\ Cov(z_1, r_K) = 0. \tag{17}$$

- In other words, the linear projection of $x_k$ on $(1, x_2, \ldots, x_{K-1}, z_1)$ is

$$L(x_K|1, x_2, \ldots, x_{K-1}, z_1) = \delta_1 + \delta_2 x_2 + \ldots + \delta_{K-1} x_{K-1} + \theta_1 z_1. \tag{18}$$

- Can show: the rank condition (12) holds if and only if

$$\theta_1 \neq 0. \tag{19}$$

- OLS consistently estimates the parameters of a linear projection (not necessarily unbiased).
- Need to reject

$$H_0 : \theta_1 = 0 \tag{20}$$

in favor of (19) convincingly. Heteroskedasticity-robust inference can be used.

• Do not care about the $\delta_j$ in (18), but $x_2, \ldots, x_{K-1}$ must be partialled out. ($z_1$ could be correlated with $x_K$, but it must be *partially* correlated.)

• $x_K$ can be discrete, continuous, some mixture. Regardless of the nature of $x_K$, the linear projection is well-defined. The IV estimator is consistent under $E(\mathbf{z}'u) = \mathbf{0}$ if

$$L(x_K|1, x_2, \ldots, x_{K-1}, z_1) \neq L(x_K|1, x_2, \ldots, x_{K-1}).$$

This last condition is just another way to say that, in a linear sense, $z_1$ helps to predict $x_K$ controlling for the other exogenous variables.

• Regressing $x_k$ on $1, x_2, \ldots, x_{K-1}, z_1$ using the data is often called the **first-stage regression**. It should be done to establish sufficient partial correlation between $x_K$ and $z_1$.

• A reduced form also exists for $y$, and can be written

$$y = \pi_1 + \pi_2 x_2 + \ldots + \pi_{K-1} x_K + \pi_K z_1 + v$$

The $\pi_j$ can be consistently estimated by OLS.

• If $x_K$ is participation in a program and $z_1$ is eligibility, the reduced form for $y$ estimates the effect of eligibility on $y$. The "structural" equation, $y = \beta_1 + \beta_2 x_2 + \ldots + \beta_K x_K + u$, estimated by IV, attempts to get at the causal effect of the program itself.

## 3. TWO STAGE LEAST SQUARES

- In some cases, we have more instruments than we need. For example, if we can use mother's education as an IV, why not father's education, too?
- Again write

$$y = \mathbf{x}\boldsymbol{\beta} + u \tag{21}$$

$$E(\mathbf{z}'u) = \mathbf{0} \tag{22}$$

where $L = \dim(\mathbf{z}) \geq \dim(\mathbf{x}) = K$.

- When $L > K$, have more than one IV estimator. We say the model (21) is (potentially) **overidentified**.
- When $L = K$ and the rank condition holds, the model is **just identified**.
- Suppose $z_1$ and $z_2$ are IVs for $x_K$. Which should we use? Under a homoskedasticity assumption, the best IV for $x_K$ is the linear combination of *all* exogenous variables defined by the linear projection.

• In general, the best vector of IVs for $\mathbf{x}$ is the vector of linear projections of each element of $\mathbf{x}$ on $\mathbf{z}$. We can write

$$\underset{1\times K}{\mathbf{x}} = \underset{1\times L}{\mathbf{z}}\ \underset{L\times K}{\boldsymbol{\Pi}} + \underset{1\times K}{\mathbf{r}}$$

where $\boldsymbol{\Pi}$ is the $L \times K$ matrix

$$\boldsymbol{\Pi} = [\underset{L\times L}{E(\mathbf{z}'\mathbf{z})}]^{-1}[\underset{L\times K}{E(\mathbf{z}'\mathbf{x})}]$$

and

$$E(\mathbf{z}'\mathbf{r}) = \mathbf{0}.$$

• For each $x_j$ we can write

$$x_j = \mathbf{z}\boldsymbol{\pi}_j + r_j \equiv x_j^* + r_j$$

where $\boldsymbol{\pi}_j$ $(L \times 1)$ is the $j^{th}$ column of $\boldsymbol{\Pi}$.

• For any $x_j \in \mathbf{z}$, $x_j^* = x_j$, so exogenous variables act as their own instruments.

• In the general case, use

$$\mathbf{x}^* = \mathbf{z}\boldsymbol{\Pi}$$

as the $1 \times K$ vector of instruments for $\mathbf{x}$.

- Because $\mathbf{z}$ is exogenous, so is $\mathbf{x}^*$:

$$E(\mathbf{x}^{*\prime}u) = \mathbf{0}.$$

- The rank condition becomes

$$rank\, E(\mathbf{x}^{*\prime}\mathbf{x}) = K.$$

But

$$E(\mathbf{x}^{*\prime}\mathbf{x}) = \mathbf{\Pi}' E(\mathbf{z}'\mathbf{x}) = E(\mathbf{x}'\mathbf{z})[E(\mathbf{z}'\mathbf{z})]^{-1}E(\mathbf{z}'\mathbf{x}).$$

• Formally, here are the first two assumptions for 2SLS, stated in the population. (We will assume access to a random sample).

**Assumption 2SLS.1** (Exogenous Instruments): $E(\mathbf{z}'u) = \mathbf{0}$.

**Assumption 2SLS.2** (Rank Condition): (a) *rank* $E(\mathbf{z}'\mathbf{z}) = L$; (b) *rank* $E(\mathbf{z}'\mathbf{x}) = K$.

• Part (a) rules out perfect collinearity among the exogenous variables (which means we cannot use linear combinations of exogenous variables as additional instruments). Part (b) is the practically important restriction, and requires $L \geq K$.

- Deriving 2SLS: With

$$\boldsymbol{\beta} = [E(\mathbf{x}^{*\prime}\mathbf{x})]^{-1}E(\mathbf{x}^{*\prime}y),$$

need to worry about unknown $\boldsymbol{\Pi}$ because $\mathbf{x}_i^* = \mathbf{z}_i\boldsymbol{\Pi}$. Two-step estimation:

(1) Run the regression $\mathbf{x}_i$ on $\mathbf{z}_i$, $i = 1,\ldots,N$ to obtain $\hat{\boldsymbol{\Pi}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. Obtain the vector fitted values,

$$\hat{\mathbf{x}}_i = \mathbf{z}_i\hat{\boldsymbol{\Pi}},\ i = 1,\ldots,N.$$

(This is the same as regressing each element of $\mathbf{x}_i$ not in $\mathbf{z}_i$ on $\mathbf{z}_i$, and obtaining the fitted values. Any element of $\mathbf{x}_i$ in $\mathbf{z}_i$ is used as its own fitted value.)

(2) Use $\hat{\mathbf{x}}_i$ as the vector of IVs for $\mathbf{x}_i$:

$$\hat{\boldsymbol{\beta}}_{IV} = \left( N^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i' \mathbf{x}_i \right)^{-1} \left( N^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i' y_i \right).$$

- We can write this differently. Because

$$\mathbf{x}_i = \hat{\mathbf{x}}_i + \hat{\mathbf{r}}_i$$

$$\sum_{i=1}^{N} \hat{\mathbf{x}}_i' \hat{\mathbf{r}}_i = \mathbf{0} \ \text{(by OLS FOCs).}$$

So

$$\sum_{i=1}^{N} \hat{\mathbf{x}}_i' \mathbf{x}_i = \sum_{i=1}^{N} \hat{\mathbf{x}}_i' \hat{\mathbf{x}}_i$$

and then the IV estimator can be written as a two stage least squares estimator:

$$\hat{\boldsymbol{\beta}}_{2SLS} = \left( N^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i' \hat{\mathbf{x}}_i \right)^{-1} \left( N^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i' y_i \right).$$

- The first-stage regression is $\mathbf{x}_i$ on $\mathbf{z}_i$ to get the fitted values, $\hat{\mathbf{x}}_i$. The second-stage regression is $y_i$ on $\hat{\mathbf{x}}_i$.
- The two-stage least squares algorithm is not really the best way to think about the estimator. For one, standard errors from second-stage regression are not correct. Also, we will see that using the fitted values as IVs is not the same as 2SLS approach for some panel data applications. Finally, the two-step approach can lead to abuse for simple nonlinear models.

- Using full data matrices and some algebra, we can write

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{2SLS} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{Y} \\
&= [(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Y}) \\
&= \boldsymbol{\beta} + [(\mathbf{X}'\mathbf{Z}/N)(\mathbf{Z}'\mathbf{Z}/N)^{-1}(\mathbf{Z}'\mathbf{X}/N)]^{-1}(\mathbf{X}'\mathbf{Z}/N)(\mathbf{Z}'\mathbf{Z}/N)^{-1}(\mathbf{Z}'\mathbf{U}/N)
\end{aligned}$$

where the last expression can be used to show consistency by applying the WLLN to each term, along with the rank condition and $E(\mathbf{z}'u) = \mathbf{0}$

Key Result: Under 2SLS.1 and 2SLS.2, $\hat{\boldsymbol{\beta}}_{2SLS}$ on a random sample is consistent for $\boldsymbol{\beta}$.

- For inference, it is useful to show

$$N^{1/2}(\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}) = \left( N^{-1} \sum_{i=1}^{N} \mathbf{x}_i^{*\prime} \mathbf{x}_i^{*} \right)^{-1} \left( N^{-1/2} \sum_{i=1}^{N} \mathbf{x}_i^{*\prime} u_i \right) + o_p(1)$$

where the $\mathbf{x}_i^* = \mathbf{z}_i \boldsymbol{\Pi}$ are the linear projections.

- It follows that

$$N^{1/2}(\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}) \xrightarrow{d} Normal(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})$$

$$\mathbf{A} = E(\mathbf{x}_i^{*\prime} \mathbf{x}_i^{*})$$

$$\mathbf{B} = E(u_i^2 \mathbf{x}_i^{*\prime} \mathbf{x}_i^{*}).$$

**Assumption 2SLS.3** (Homoskedasticity):

$$E(u^2\mathbf{z}'\mathbf{z}) = E(u^2)E(\mathbf{z}'\mathbf{z}) \equiv \sigma^2 E(\mathbf{z}'\mathbf{z}).$$

Key Result: Under 2SLS.1, 2SLS.2, and 2SLS.3,

$$N^{1/2}(\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}) \xrightarrow{d} Normal(\mathbf{0}, \sigma^2\mathbf{A}^{-1})$$

- 2SLS residuals:

$$\hat{u}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{2SLS}$$

(That is *not* $\hat{\mathbf{x}}_i$ in the definition of $\hat{u}_i$. That is, these are not the OLS residuals from the second stage.)

- Consistent (not unbiased) estimators of $\sigma^2$ and **A**:

$$\hat{\sigma}^2 = (N-K)^{-1} \sum_{i=1}^{N} \hat{u}_i^2 \xrightarrow{p} \sigma^2$$

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i' \hat{\mathbf{x}}_i$$

$$\widehat{Avar}(\hat{\boldsymbol{\beta}}_{2SLS}) = \hat{\sigma}^2 \hat{\mathbf{A}}^{-1}/N = \hat{\sigma}^2 (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1}$$

- Heteroskedasticity-Robust Inference:

$$\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N = \frac{N}{(N-K)}\left(\sum_{i=1}^{N}\hat{\mathbf{x}}_i'\hat{\mathbf{x}}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2\hat{\mathbf{x}}_i'\hat{\mathbf{x}}_i\right)\left(\sum_{i=1}^{N}\hat{\mathbf{x}}_i'\hat{\mathbf{x}}_i\right)^{-1}$$

which is not the same as using the heteroskedasticity-robust inference in the second stage regression $y_i$ on $\hat{\mathbf{x}}_i$.

- Efficiency: 2SLS has the smallest asymptotic variance among all IV estimators using linear functions of $\mathbf{z}_i$ as instruments under 2SLS.1, 2SLS.2, and 2SLS.3. (Only meaningful when $L > K$.)

• Potential pitfalls with 2SLS:

(1) A “little” endogeneity of one or more instruments can lead to large inconsistency if the instruments are weak, that is, only slightly partially correlated with the endogenous explanatory variables (EEVs).

(2) The standard errors of 2SLS can be large. Suppose $x_K$ is the only EEV.

$$Avar(\hat{\beta}_K) \approx \frac{\sigma^2}{\widehat{SST}_K(1 - \hat{R}_K^2)}$$

where the denominator statistics are from the regression

$$\hat{x}_{iK} \text{ on } 1, x_{i2}, \ldots, x_{i,K-1}.$$

- Do not get a perfect fit because of extra elements of $\mathbf{z}_i$. Still, $\widehat{SST}_K < SST_K$, and $\widehat{SST}_K$ is often much smaller.
- Plus, $\hat{R}_K^2$ can be close to unity with a poor instrument.

## 4. APPLICATION: ENDOGENEITY OF CHILDREN IN LABOR SUPPLY

Data are a subset from Angrist and Evans (AER, 1998).

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

. 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
```

```
. tab kids

  number of |
       kids |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |     16,215       50.90       50.90
          3 |     10,014       31.43       82.33
          4 |      3,736       11.73       94.06
          5 |      1,374        4.31       98.37
          6 |        323        1.01       99.39
          7 |        134        0.42       99.81
          8 |         47        0.15       99.96
          9 |          6        0.02       99.97
         10 |          4        0.01       99.99
         11 |          2        0.01       99.99
         12 |          2        0.01      100.00
------------+-----------------------------------
      Total |     31,857      100.00

. tab samesex

  first two |
kids are of |
   same sex |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     15,840       49.72       49.72
          1 |     16,017       50.28      100.00
------------+-----------------------------------
      Total |     31,857      100.00
```

```
. * First use OLS:

. 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
------------------------------------------------------------------------------

. * Each child beyond the first two reduces estimated hours by about 2.3 hours,
. * other things fixed.
```

```
. * But what if kids is endogenous?
. * Assume "samesex" is exogenous to the labor supply equation.
. * Is samesex partially correlated with kids?
. * Estimate the reduced form for kids (first-stage regression):

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

Linear regression                                      Number of obs =   31857
                                                       F(  7, 31849) =  437.80
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1191
                                                       Root MSE      =  .91724

------------------------------------------------------------------------------
             |               Robust
        kids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |   .0703744   .0102783     6.85   0.000      .0502285    .0905202
     nonmomi |  -.0027871    .000257   -10.85   0.000     -.0032907   -.0022834
        educ |  -.0853676   .0020296   -42.06   0.000     -.0893457   -.0813895
         age |   .0589312   .0203278     2.90   0.004       .019088    .0987744
       agesq |   1.98e-06   .0003559     0.01   0.996     -.0006956    .0006995
       black |   .0128681   .0644422     0.20   0.842      -.113441    .1391772
      hispan |  -.0424722   .0644997    -0.66   0.510     -.1688941    .0839498
       _cons |   2.010258   .2930274     6.86   0.000      1.435913    2.584603
------------------------------------------------------------------------------

. * Yes: Having the first two children the same gender means the expected
. * number of children is estimated to be .07 higher.
```

```
. * Now compute the IV (2SLS) estimates:

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

Instrumental variables (2SLS) regression               Number of obs =   31857
                                                       F(  7, 31849) =  304.81
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0583
                                                       Root MSE      =  18.924

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -4.878903   3.013547    -1.62   0.105    -10.78557    1.027766
     nonmomi |  -.0649179   .0099359    -6.53   0.000    -.0843926   -.0454432
        educ |    .368042   .2595992     1.42   0.156    -.1407823    .8768664
         age |   2.200964   .4845126     4.54   0.000       1.2513    3.150627
       agesq |  -.0277443    .007744    -3.58   0.000     -.042923   -.0125657
       black |   1.094986   1.376742     0.80   0.426    -1.603482    3.793454
      hispan |  -5.217758   1.381364    -3.78   0.000    -7.925284   -2.510232
       _cons |  -5.253976   9.037541    -0.58   0.561     -22.9679    12.45995
------------------------------------------------------------------------------
Instrumented:  kids
Instruments:   nonmomi educ age agesq black hispan samesex
------------------------------------------------------------------------------

. * Much bigger effect using IV, but only marginally statistically significant.
```

```
. corr kids samesex
(obs=31857)

             |     kids  samesex
-------------+------------------
        kids |   1.0000
     samesex |   0.0358   1.0000

. * The partial correlation is even smaller. It's not surprising the IV estimate
. * is much less precise than OLS.
```