AFRE 835: Introductory Econometrics

Chapter 15: Instrumental Variable Estimation and 2SLS

Spring 2017

Introduction

- A key issue in applied econometrics is the potential for one or more of the regressors being correlated with the unobservable factors captured by the error term u.
- We have already seen a number of tools for dealing with this problem, including
  - The use of proxy variables
  - The use of fixed effects or first differencing to control for time constant omitted variables.
- This chapter introduces another approach, the method of instrumental variables.
- We will focus on its application to a simple cross-section,

  ...though it can be applied as well to time series, pooled cross sections or panel data settings.

Outline

Omitted Variables in a Simple Regression Model

- Consider again the classic returns to education model where

$$ln(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + e \qquad (1)$$

- Without data on *abil* (or a reasonable proxy for it), we are left with the model

$$ln(wage) = \beta_0 + \beta_1 educ + u \qquad (2)$$

  where $u = \beta_2 abil + e$ is likely correlated with *educ*, leading to omitted variables bias.
- The key to the instrumental variables approach is to find a way to break the correlation between *educ* and u.
- What we need is a new piece of information - an instrumental variable - that is correlated with *educ* but not with u.

The Generic Problem

- Suppose that we have a simple regression model with

$$y = \beta_0 + \beta_1 x + u \tag{3}$$

  where $Cov(x, u) \neq 0$.
- An instrumental variable z needs to satisfies two conditions
  1. **Instrument relevance**: $Cov(z, x) \neq 0$
     - i.e., z must be linked positively or negatively to the endogenous regressor x.
     - Formally, in the population model

$$x = \pi_0 + \pi_1 z + v \tag{4}$$

       it must be the case that $\pi_1 \neq 0$.
  2. **Instrument exogeneity**: $Cov(z, u) = 0$
- While we cannot test the instrument exogeneity assumption, we can test instrument relevance using a sample of observations on (x, z).

Examples of Instruments

Wooldridge lists a number of examples

- In a model regressing wages on education,
  1. IQ would likely be a poor instrument, satisfying relevance, but violating exogeneity;
  2. Mothers or Father's education might be somewhat better, but one can still think of reasons why it might be correlated with unobservables like ability;
- In a model regressing final exam score on the number of days of classes skipped,
  1. Distance from school would likely satisfy the relevance assumption, but could still violate exogeneity by being correlated with household income, etc.
- In a model of the impact of military service on wages during the Vietnam War Era
  1. Draft numbers, randomly assigned to 18 year-olds, would serve as a useful instrument.

Using the Instrumental Variable

- In a simple regression model, it is clearly the case that

$$Cov(z, y) = \beta_1 Cov(z, x) + Cov(z, u) \tag{5}$$

- Given instrument exogeneity, the second term on the right-hand side of equation (5) is zero.
- *If* the instrumental variable also satisfies the relevance assumption, we can divide through by $Cov(z, x)$ to solve for β_1 as

$$\beta_1 = \frac{Cov(z, y)}{Cov(z, x)} \tag{6}$$

- Using the sample counterparts to these population moments yields the **instrumental variables (IV) estimator** for β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \tag{7}$$

- The intercept is estimated as $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

The Variance of the IV Slope Estimator

- Under the assumption of homoskedasticity (and the additional assumption that $E(u^2|z) = \sigma^2$), the asymptotic variance of the IV slope estimator in the simple regression model is given by

$$\frac{\sigma^2}{n\sigma_x^2 \rho_{x,z}^2} \tag{8}$$

  where σ_x^2 is the population variance of x and $\rho_{x,z}$ is the population covariance between x and z.
- Four factors influence the variance of the slope estimator:
  1. The residual variance σ^2;
  2. The sample size n;
  3. The variability of x (i.e., σ_x^2); and
  4. The strength of the connection between x and z (i.e., $\rho_{x,z}$).

Estimating the Asymptotic Standard Error for $\hat{\beta}_1$

- A consistent estimator for the standard error for $\hat{\beta}_1$ is given by the square root of

$$\frac{\hat{\sigma}^2}{SST_x \cdot R^2_{x,z}} \tag{9}$$

where $\hat{\sigma}^2$ is based on the IV residuals and $R^2_{x,z}$ is the R-squared from a regression of x on z (and a constant).

- The only difference between this variance and the corresponding OLS estimator's variance is the $R^2_{x,z}$ term, which will make the OLS variance smaller.

  ...so that *if* x is not endogenous, then the use of the IV estimator will unnecessarily increase the variance of the slope estimator.

  ...of course, this should not be surprising, since, if x is exogenous, then OLS is *BLUE*.

Example #1: The Returns to Education for Women (mroz.dta) - code

```
*****************************************************************************
*                                                                           *
*       First Stage Regression - Returns to Education for Women             *
*                                                                           *
*****************************************************************************;
reg     educ fatheduc, robust;
outreg  using "`TableA'", bdec(3) se tex title(IV Relevance)
        ctitle("", fatheduc) replace;
reg     educ motheduc, robust;
outreg  using "`TableA'", bdec(3) se tex title(IV Relevance)
        ctitle("", motheduc) merge;

*****************************************************************************
*                                                                           *
*       IV Regression                                                       *
*                                                                           *
*****************************************************************************;
reg     lwage educ, robust;
outreg  using "`TableB'", bdec(3) se tex title(OLS and IV Estimates)
        ctitle("", OLS) replace;
ivregress 2sls lwage (educ = fatheduc), robust;
outreg  using "`TableB'", bdec(3) se tex title(OLS and IV Estimates)
        ctitle("", feduc) merge;
ivregress 2sls lwage (educ = motheduc), robust;
outreg  using "`TableB'", bdec(3) se tex title(OLS and IV Estimates)
        ctitle("", meduc) merge;
```

Example #1: The Returns to Education for Women (mroz.dta)

Consider three possible instruments: *fatheduc* and *motheduc*

IV Relevance

	fatheduc	motheduc
fatheduc	0.282	
	(0.023)**	
motheduc		0.295
		(0.024)**
_cons	9.799	9.560
	(0.214)**	(0.242)**
R^2	0.20	0.19
N	753	753

* $p < 0.05$; ** $p < 0.01$

Example #1: The Returns to Education for Woman (cont'd)

OLS and IV Estimates

	OLS	IV-fatheduc	IV-motheduc
educ	0.109	0.059	0.039
	(0.013)**	(0.037)	(0.039)
_cons	-0.185	0.441	0.702
	(0.171)	(0.464)	(0.493)
R^2	0.12	0.09	0.07
N	428	428	428

* $p < 0.05$; ** $p < 0.01$

Notice what happens to the standard errors using the IV estimator. It's not clear that the education effect significantly different from OLS estimate.

Example #2: The Returns to Education for Men (WAGE2.dta)

Consider three possible instruments: *feduc*, *meduc*, and *sibs*

IV Relevance

	feduc	meduc	sibs
feduc	0.290		
	(0.021)**		
meduc		0.281	
		(0.024)**	
sibs			-0.228
			(0.028)**
_cons	10.650	10.575	14.139
	(0.225)**	(0.261)**	(0.116)**
R^2	0.18	0.13	0.06
N	741	857	935

Example #2: The Returns to Education for Men (cont'd)

OLS and IV Estimates

	OLS	IV-feduc	IV-meduc	IV-sibs
educ	0.060	0.097	0.111	0.122
	(0.006)**	(0.016)**	(0.017)**	(0.025)**
_cons	5.973	5.473	5.280	5.130
	(0.082)**	(0.217)**	(0.226)**	(0.330)**
R^2	0.10	0.05	0.03	.
N	935	741	857	935

* $p < 0.05$; ** $p < 0.01$

The direction of the change in $\hat{\beta}_1$ is inconsistent with the omitted variables bias story.

The Problem of Weak Instruments

- One consequence of using the IV estimator is that the standard errors can become large.
- This is particularly true if $\rho_{z,x}$ is small.
- A more serious problem arises if z is not truly exogenous.

  ... In this case, the IV estimator can be seriously biased.
- For the IV estimator,

$$plim\hat{\beta}_{1,IV} = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x} \tag{10}$$

... whereas

$$plim\hat{\beta}_{1,OLS} = \beta_1 + Corr(x,u) \cdot \frac{\sigma_u}{\sigma_x} \tag{11}$$

The Problem of Weak Instruments (cont'd)

- Combining these results, the relative inconsistency of IV versus OLS is given by

$$\frac{plim\left[\hat{\beta}_{1,IV} - \beta_1\right]}{plim\left[\hat{\beta}_{1,OLS} - \beta_1\right]} = \frac{Corr(z,u)}{Corr(x,u)} \cdot \frac{1}{Corr(z,x)} \tag{12}$$

- Even if $Corr(z,u) < Corr(x,u)$, a weak instrument (a small $Corr(z,x)$) can lead to greater inconsistency using IV than using OLS.
- It is not enough that the instrument be statistically significant in the relevance test.
- Staiger and Stock (1997) suggest rules of thumb for avoiding weak instruments.

Example: Birthweight and Smoking

OLS and IV Estimates

RHS variable:	Packs	Bwght (OLS)	Bwght (IV-cigprice)
cigprice	0.00028		
	(0.00078)		
packs		-0.09	2.99
		(0.02)**	(8.98)
_cons	0.06743	4.77	4.45
	(0.10254)	(0.01)**	(0.94)**
R^2	0.00	0.02	.
N	1,388	1,388	1,388

IV in the Multiple Regression Model

- Extending the IV estimator to the multiple regression model setting is straightforward.
- Wooldridge introduces some slightly different notation, with the regression model:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \tag{13}$$

  where y_2 denotes our (potentially) endogenous variable and z_1 denotes an additional exogenous variable (with $E(u_1|z_1) = 0$).
- Equation (13) is often referred to as a *structural equation* for y_1.
- An example in modeling of the returns to education would be to specify $y_1 = ln(wages)$, $y_2 = educ$ and $z_1 = exper$.

Instrument Assumptions

- Our instrument, denoted by z_2, must still satisfy two conditions:
  1. Instrument exogeneity: $Cov(u_1, z_2) = 0$.
  2. Instrument relevance:
     ... This condition is bit more complex in this setting, requiring $\pi_2 \neq 0$ in the **reduced form model**
     $$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2. \tag{14}$$
     where, by construction, $E(v_2) = 0$ and $Cov(z_j, v_2) = 0 \ \ (j = 1, 2)$.
     ... Essentially, we require that z_2 contribute something (over and above z_1) in *explaining* the variation in y_2.
- Generalizing the model further to include additional exogenous variables to our structural model for y_1 is straightforward.

  ... It boils down to treating z_1 as a vector rather than a scalar above, including in the reduced form model in equation (14)

Deriving the IV Estimator

- The IV estimator makes use of the following three assumptions:
  1. $E(u) = 0$;
  2. $Cov(z_1, u) = 0$;
  3. $Cov(z_2, u) = 0$.
- The sample counterpart to these assumptions imply the underlying **instrumental variable (IV) estimator**:
  1. $\sum_{i=1}^{n}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} + \hat{\beta}_2 z_{i1}) = 0$;
  2. $\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} + \hat{\beta}_2 z_{i1}) = 0$;
  3. $\sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} + \hat{\beta}_2 z_{i1}) = 0$.
- The IV estimator in matrix form is $\hat{\beta} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}_1$ where
  $$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{y}_2 & \boldsymbol{z}_1 \end{bmatrix} \tag{15}$$

Wooldridge Example 15.1 (CARD.dta) - IV code

```
. ivregress 2sls lwage exper expersq black smsa south (educ = nearc4), robust;

Instrumental variables (2SLS) regression          Number of obs   =      3,010
                                                  Wald chi2(6)    =     792.07
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.2252
                                                  Root MSE        =     .39058

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1322888   .0485213     2.73   0.006     .0371888    .2273889
       exper |    .107498   .0211129     5.09   0.000     .0661175    .1488785
     expersq |  -.0022841   .0003463    -6.59   0.000    -.0029629   -.0016053
       black |  -.1308019   .0514513    -2.54   0.011    -.2316445   -.0299592
        smsa |   .1313237   .0297684     4.41   0.000     .0729787    .1896686
       south |  -.1049005   .0228997    -4.58   0.000    -.1497831   -.0600179
       _cons |   3.752781   .8167498     4.59   0.000     2.151981    5.353582
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq black smsa south nearc4
```

Uses college proximity as a instrument for *educ*.

	Relevance	OLS	IV(nearc4)
RHS Variable:	educ	lwage	lwage
nearc4	0.337		
	(0.083)**		
educ		0.074	0.132
		(0.004)**	(0.049)**
exper	-0.410	0.084	0.107
	(0.034)**	(0.007)**	(0.021)**
expersq	0.001	-0.002	-0.002
	(0.002)	(0.000)**	(0.000)**
black	-1.006	-0.190	-0.131
	(0.090)**	(0.017)**	(0.051)*
smsa	0.404	0.161	0.131
	(0.085)**	(0.015)**	(0.030)**
south	-0.291	-0.125	-0.105
	(0.079)**	(0.015)**	(0.023)**
_cons	16.659	4.734	3.753
	(0.176)**	(0.070)**	(0.817)**
R^2	0.47	0.29	0.23

Multiple Instruments

- There are going to be settings in which we have more than one instrumental variable for an endogenous variable.
- In the modeling the returns to education, we have already considered four instruments for education:
  1. Mother's education
  2. Father's education
  3. Number of siblings
  4. Proximity to a four year college
- The question is how best to use these competing instruments.
- Using them individually leaves us with competing sets of parameter estimates.
- Instead, we want to consider how to use them in combination.

  ...A linear regression provides a natural approach for doing so.

Multiple Instruments with a Single Endogenous Regressor

- Consider again our structural model:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \tag{16}$$

  where now we have, say, two instrumental variables z_2 and z_3.
- The assumptions that our instruments do not appear in (16) *and* are uncorrelated with u_1 are referred as **exclusion restrictions**.
- The reduced form expression for our endogenous variable becomes:

$$\begin{aligned} y_2 &= \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2 \\ &= y_2^* + v_2 \end{aligned} \tag{17}$$

  where $E(v_2) = 0$ and $Cov(z_j, v_2) = 0 \quad j = 1, 2, 3$.
- Effectively, we have segments y_2 into two parts
  1. y_2^* which is uncorrelated with u_1
  2. v_2 which is potentially correlated with u_1.

Forming the Instrument for y_2

- We then form the instrument for y_2 using OLS, yielding
$$\hat{y}^2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3 \tag{18}$$
Note: It is key that z_1 is included in (18).
- Our instruments, z_2 and z_3, must still satisfy two conditions:
  1. Instrument exogeneity: $Cov(u_1, z_j) = 0 \quad j = 2, 3$.
  2. Instrument relevance:
     In the reduced form model
     $$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2. \tag{19}$$
     we need either $\pi_2 \neq 0$ or $\pi_3 \neq 0$.
     i.e., similar to the single instrument case, we require that z_2 and z_3 *together* contribute something (over and above z_1) in *explaining* the variation in y_2.
- Again, generalizing the model further to include additional exogenous variables to our structural model for y_1 is straightforward.
  ...It boils down to treating z_1 as a vector rather than a scalar above, including in the reduced form model in (19)

The Two Stage Least Squares (2SLS) Estimator

- The IV estimator in this case solves the following set of equations:
  1. $\sum_{i=1}^{n}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} + \hat{\beta}_2 z_{i1}) = 0;$
  2. $\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} + \hat{\beta}_2 z_{i1}) = 0;$
  3. $\sum_{i=1}^{n} \hat{y}_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} + \hat{\beta}_2 z_{i1}) = 0.$
- This is also known as the **Two Stage Least Squares (2SLS) Estimator** since it can be obtained using the two stages:
  1. Regressing y_2 on all of the exogenous variables z_1, z_2 and z_3, and forming $\hat{y}_2$.
  2. Regressing y_1 on z_1 and $\hat{y}_2$ (and a constant).

  Important note: the standard errors from this second stage regression *will not* be correct, but must be adjusted to reflect the use of an estimator for y_2^*.

  Stata does this correction for you.

Example #1: MROZ (female returns to education)

OLS and IV Estimates

	OLS	IV Stage 1	IV Stage 2
educ	0.1075		0.0614
	(0.0132)**		(0.0332)
exper	0.0416	0.085	0.0442
	(0.0153)**	(0.026)**	(0.0155)**
expersq	-0.0008	-0.002	-0.0009
	(0.0004)	(0.001)*	(0.0004)*
fatheduc		0.185	
		(0.024)**	
motheduc		0.186	
		(0.026)**	
_cons	-0.5220	8.367	0.0481
	(0.2017)**	(0.280)**	(0.4278)
N	428	753	428

Miscellaneous 2SLS Issues/Topics

- The 2SLS estimator is more susceptible to the problem multicollinearity because
  - $\hat{y}_2$ will have less variability than y_2.
  - Corr$(\hat{y}_2, z_1)$ will generally be much higher than Corr(y_2, z_1).
- One can have multiple endogenous regressors, but this requires *at least* one instrument for each regressor.
- It is important not to rely on 2SLS R-squared or SSR's reported for most regression packages.

  ...Instead one should use Stata's *test* command, which computes the correct F-stats.
- 2SLS can be adapted to
  - allow for heteroskedasticity (See Wooldridge 15.6)
  - time series and panel data settings (See Wooldridge Sections 15.7 and 15.8, respectively)

IV Solution to the Errors-in-Variables Problem

- Recall in the classic error-in-variables (CEV) setting, the observed (mis-measured) variable $x_1 = x_1^* + e_1$ is correlated with the composite error term $\tilde{u} = u - \beta_1 e_1$, since

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u \\ &= \beta_0 + \beta_1(x_1 - e_1) + \beta_2 x_2 + u \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \tilde{u} \end{aligned} \tag{20}$$

- To the extent that there is an instrument for our mis-measured variable, and the instrument is not correlated with the measurement error, one can reduce to attenuation bias due to measurement error.

Testing for Endogeneity - The Control Function Approach

- We've already seen that 2SLS is less efficient than OLS when our regressors are exogenous.
- It makes sense, then, to test for exogeneity.

  ...If we fail to reject exogeneity, then we can return to using OLS.
- The test uses what's known as a **control function** approach.
- Suppose we have a single endogenous variable, two exogenous variables and two instruments; i.e., a structural equation for y_1 given by

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1 \tag{21}$$

  and a reduced form equation for y_2 given by;

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \tag{22}$$

  with $Cov(z_j, u_1) = 0 \ \ \forall j$.

Testing for Endogeneity - The Control Function Approach

- The only way in which y_2 is correlated with u_1 is if v_2 is correlated with u_1.
- We can segment u_1 into two components:
  - The part that is correlated with v_2
  - The part that is uncorrelated with v_2
- We do this by writing u_1 as $\delta_1 v_2 + e_1$, where $Corr(e_1, v_2) = 0$ and $E(e_1) = 0$.
- If $\delta_1 = 0$, then $u_1 = e_1$ and $Corr(y_2, u_1) = 0$.
- While we do not have v_2, we can use the estimated residuals from the reduced form regression for y_2 to obtain $\hat{v}_2$ and estimate

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + error. \tag{23}$$

- Testing $H_0 : \delta_1 = 0$ is our test for exogeneity.

A Few More Notes on the Endogeneity Test

- Even before testing for endogeneity, one should compare the OLS and 2SLS estimates.
- If they are substantially different, that suggest potential endogeneity problem

  ...or other problems with the underlying assumptions.
- The estimates of the β_j's in equation (23) will be identical to those obtained via 2SLS
- The 2SLS and control function approaches both break the endogeneity, but using different tacks:
  - 2SLS subdivides y_2
  - The control function approach subdivides u_1.

Testing Overidentifying Restrictions

- If our model is just identified, then we cannot test the suitability of the instrument.
- However, with multiple instruments for a single endogenous variable, we can test the implied *overidentifying restrictions*.

  ... The basic idea is that we can obtain separate estimates of the parameter on our endogenous variables for each instrument.

  ... If all of our instruments are valid, these estimates should be the same (asymptotically).
- Three steps are involved in the test
  1. Estimate the structural equation for y_1 and construct $\hat{u}_1$.
  2. Regress $\hat{u}_1$ on *all exogenous* variables, obtaining R_1^2.
  3. Under the null hypothesis that all of the IV's are uncorrelated with u_1, $nR_1^2 \overset{a}{\sim} \chi_q^2$ where q is the number of instrumental variables minus the number of endogenous variables.
- Note: Passing this test is no guarantee that the instruments are fine.