Implementation of IV and Two-Stage Least Squares
14.32, Spring 2007

Review of IV

In lecture we saw the schooling model

$$y_i = \alpha + \rho S_i + \gamma A_i + \varepsilon_i$$

where y_i is log income of individual i, S_i is schooling, and A_i is unmeasured ability. Since ability is unobservable, the OLS regression becomes

$$y_i = \alpha + \rho S_i + u_i$$

where $u_i = \gamma A_i + \varepsilon_i$. From the omitted variables bias formula, the coefficient we estimate has $p\lim(\hat{\rho}) = \rho + \gamma\beta_{AS}$, where β_{AS} is the coefficient of the bivarite regression of ability on schooling.

One way to deal with this problem is to put in enough covariates so that schooling is no longer correlated with the unobserved error term. This works in some cases, but often you can't find the appropriate covariates, e.g., how can you measure ability? Some people have used test scores, but these results are not 100% convincing.

Instrumental variables estimation gets around the omitted variables bias problem by finding another variable, called an instrument, which affects the variable of interest but nothing else in the regression. More formally, you find some instrument z where $cov(z_i, u_i) = 0$. Then, you can back out the parameter of interest using covariances:

$$\begin{aligned} cov(y_i, z_i) &= cov(\alpha + \rho S_i + u_i, z_i) = \rho cov(S_i, z_i) \\ \Rightarrow \frac{cov(y_i, z_i)}{cov(S_i, z_i)} &= \frac{\beta_{yz}}{\beta_{Sz}} = \rho \end{aligned}$$

From the law of large numbers, the sample analog will give you a consistent estimate, i.e.,

$$p\lim \frac{\widehat{cov(y_i, z_i)}}{\widehat{cov(S_i, z_i)}} = p\lim \frac{\widehat{\beta_{yz}}}{\widehat{\beta_{Sz}}} = \rho$$

The important assumptions for the instrument z are $cov(z_i, S_i) \neq 0$ and $cov(z_i, u_i) = 0$. In words we can state these two assumptions as:

1. The insturment z is correlated with the endogenous variable S (there is a first stage).

2. The instrument z only affects y through the variable S (exclusion restriction).

Two-Stage Least Squares

Practically, IV is usually implemented through a process called two-stage least squares (2SLS). This is actually a simple type of simultaneous equations problem. Let's set up the relationships between y, S, z and u as

$$\begin{aligned} y_i &= \alpha + \rho S_i + u_i \\ S_i &= \kappa + \delta z_i + \nu_i \end{aligned}$$

To run 2SLS:

1. Regress S on z

2. Using the estimated parameters, estimate $\hat{S}$

3. Regress y on $\hat{S}$

4. The estimated $\hat{\rho}$ is your IV estimator.

It turns out that through this process the estimated $\hat{\rho}$ is numerically equivalent to the IV estimator developed above, $\frac{\widehat{cov(y,z)}}{\widehat{cov(z,S)}}$. Here's the proof (supressing some hats and $i's$ to make it easier to read):

$$
\begin{aligned}
\hat{\rho}_{2SLS} &= \frac{cov(y,\hat{S})}{var(\hat{S})} \\
&= \frac{cov(y, \frac{cov(S,z)}{var(z)} z + \hat{\kappa})}{var(\frac{cov(S,z)}{var(z)} z + \hat{\kappa})} \\
&= \frac{\frac{cov(S,z)}{var(z)} cov(y,z)}{\left(\frac{cov(S,z)}{var(z)}\right)^2 var(z)} \\
&= \frac{cov(y,z)}{cov(S,z)} = \hat{\rho}_{IV}
\end{aligned}
$$

Computing 2SLS Estimates

2SLS makes computing IV estimators easy. Here's an example with SAS. As we have seen in class, there are a number of instruments one can use to estimate the wage-schooling equation

$$y_i = \alpha + \rho S_i + u_i$$

An alternative (and clever) instrument for schooling was developed by Card (1995). His insight was that people who live closer to colleges have a lower cost of attending school, and therefore higher education levels. For his instrument he constructed an indicator for whether the individual lived near a four-year college. The quality of this instrument has been the subject of debate, and you can make your own judgement as to its own validity.
To compute 2SLS, we first estimate the first-stage regression:

$$S_i = \kappa + \delta * nearc4_i + v_i$$

Then take the fitted values and estimate

$$y_i = \alpha + \rho \hat{S}_i + \zeta_i$$

Note that the standard errors from the second OLS estimation need to be corrected for 2SLS. The basic idea of this is that once we have a consistent estimate of $\hat{\rho}$, we use S, not $\hat{S}$, to compute the residuals. SAS does this automatically with PROC SYSLIN.

```
/*2SLS DEMONSTRATION*/


data one;
  set 'card';


/*OLS REGRESSION*/
proc reg data=one;
  model lwage=educ;
  title "Simple OLS";

/*FIRST STAGE OF 2SLS REGRESSION*/
proc reg data=one;
  model educ=nearc4;
  output out=one
  p=educhat;
  title "First Stage" ;

/*SECOND STAGE*/
proc reg data=one;
  model lwage=educhat;
  title "Second Stage";

/*DOING IT ALL AT ONCE USING PROC SYSLIN*/
proc syslin data=one 2sls;
  endogenous educ;
  instruments nearc4;
  model lwage=educ;
  title "2SLS with PROC SYSLIN";
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: lwage

Number of Observations Read	3010
Number of Observations Used	3010

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	58.51536	58.51536	329.54	<.0001
Error	3008	534.12627	0.17757		
Corrected Total	3009	592.64163			

Root MSE	0.42139	R-Square	0.0987
Dependent Mean	6.26183	Adj R-Sq	0.0984
Coeff Var	6.72948		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.57088	0.03883	143.47	<.0001
educ	1	0.05209	0.00287	18.15	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: educ

Number of Observations Read	3010
Number of Observations Used	3010

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	448.60420	448.60420	63.91	<.0001
Error	3008	21113	7.01911		
Corrected Total	3009	21562			

Root MSE	2.64936	R-Square	0.0208
Dependent Mean	13.26346	Adj R-Sq	0.0205
Coeff Var	19.97488		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.69801	0.08564	148.27	<.0001
nearc4	1	0.82902	0.10370	7.99	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: lwage

Number of Observations Read	3010
Number of Observations Used	3010

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	15.86603	15.86603	82.74	<.0001
Error	3008	576.77560	0.19175		
Corrected Total	3009	592.64163			

Root MSE	0.43789	R-Square	0.0268
Dependent Mean	6.26183	Adj R-Sq	0.0264
Coeff Var	6.99299		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3.76747	0.27433	13.73	<.0001
educhat	Predicted Value of educ	1	0.18806	0.02067	9.10	<.0001

The SYSLIN Procedure
Two-Stage Least Squares Estimation

Model	lwage
Dependent Variable	lwage

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	15.86603	15.86603	51.17	<.0001
Error	3008	932.7531	0.310091		
Corrected Total	3009	592.6416			

Root MSE	0.55686	R-Square	0.01673
Dependent Mean	6.26183	Adj R-Sq	0.01640
Coeff Var	8.89289		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.767472	0.348862	10.80	<.0001
educ	1	0.188063	0.026291	7.15	<.0001