Homework on Instrument Variable—Chapter 15

1. Consider the simple regression model $y_i = \beta_0 + \beta_1 x_i + u_i$ and let z be a binary instrumental variable for x. Use (15.10) in the book to show that the IV estimator $\hat{\beta}_1$ can be written as

   $\hat{\beta}_1 = \dfrac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$ where $\bar{y}_1$ and $\bar{x}_1$ are the sample averages of y_i and x_i over the part of the sample where z=1 and $\bar{y}_0$ and $\bar{x}_0$ are the sample averages of y_i and x_i over the part of the sample where z=0.

   This estimator, known as the grouping estimator, was first suggested by Wald (1940). In the next problem and in the empirical part of the problem set below, we will refer to this Wald estimator.

Step 1: Rewrite the numerator in the formula for $\hat{\beta}_1$ dropping the $\bar{z}$

Remember, this is allowed because $\sum_{i=1}^{n}(z_i - \bar{z})\,(x_i - \bar{x}) = \sum_{i=1}^{n} z_i (x_i - \bar{x})$ and similarly when we replace x with y. (If you need to verify that statement, crank through the algebra to show it.)

$$\sum_{i=1}^{n} z_i (y_i - \bar{y}) = \sum_{i=1}^{n} z_i y_i - \left(\sum_{i=1}^{n} z_i \right) \bar{y} = n_1 \bar{y}_1 - n_1 \bar{y},$$

where $n_1 = \sum_{i=1}^{n} z_i$ is the number of observations with $z_i = 1$, and we have used the fact that $\left(\sum_{i=1}^{n} z_i y_i \right) / n_1 = \bar{y}_1$, the average of the y_i over the i with $z_i = 1$. So far, we have shown that the numerator in $\hat{\beta}_1$ is $n_1(\bar{y}_1 - \bar{y})$.

Step 2: Write $\bar{y}$ as a weighted average of the averages over the two subgroups:

$$\bar{y} = (n_0/n)\,\bar{y}_0 + (n_1/n)\,\bar{y}_1,$$

where $n_0 = n - n_1$. Therefore,

$$\bar{y}_1 - \bar{y} = [(n - n_1)/n]\;\bar{y}_1 - (n_0/n)\;\bar{y}_0 = (n_0/n)\,(\bar{y}_1 - \bar{y}_0).$$

Therefore, the numerator of $\hat{\beta}_1$ can be written as

$$(n_0 n_1/n)(\bar{y}_1 - \bar{y}_0).$$

Step 3: By simply replacing y with x, the denominator in $\hat{\beta}_1$ *can be expressed as* $(n_0 n_1/n)(\bar{x}_1 - \bar{x}_0)$. *When we take the ratio of these, the terms involving* n_0, n_1, *and n, cancel, leaving*

$$\hat{\beta}_1 = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0).$$

2. Take the model $y_i = \beta_0 + \beta_1 x_i + u_i$ and data, (y_i, x_i, z_i) i= 1,..., *n* where *i* denotes entities, *y* is the dependent variable, and *x* is an explanatory variable for each entity and *z* is an instrument that takes on the value of either 0 or 1 (a dummy variable). Assume that both *x* and *y* are continuous. Note that the 2SLS estimator will be the Wald Estimator discussed above.

The following is some data to make this more concrete.

Sample

y	x	z
20	3	0
20		0
30	3	0
	6	0
50	3	0
40	4	0
65	2	0
70		0
45	8	0
30	9	0
	8	1
75	9	1
60	8	1
60		1
55	7	1
	8	1
90	7	1
85	9	1
75	4	1
90	7	1

Note: In the table, I blacked out some of the values of the data, but these were included the regressions that follow. The idea is that you cannot calculate $\hat{\beta}_1$ using a computer package (or by hand doing averages).

Given the information provided below, what is $\hat{\beta}_1$**?** (Note—not all of the following information may be relevant.)

Sample Summary Statistics:

$\bar{y} = 54.25$ $\bar{x} = 6.1$ $\bar{Z} = 0.5$

stdev(y)= 22.37 stdev(x)=2.31 stdev(z)=0.51

Regression #1Dependent Variable: X
Method: Least Squares
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Constant	4.900000	0.636832	7.694332	0.0000
Z	2.400000	0.900617	2.664840	0.0158

R-squared 0.282908 Mean dependent var 6.100000
Adjusted R-squared 0.243069 S.D. dependent var 2.314713
S.E. of regression 2.013841 Akaike info criterion 4.332604
Sum squared resid 73.00000 Schwarz criterion 4.432177
Log likelihood -41.32604 F-statistic 7.101370
Durbin-Watson stat 1.514521 Prob(F-statistic) 0.015786

Regression #2 Dependent Variable: Y
Method: Least Squares
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Constant	36.48330	14.13098	2.581795	0.0188
X	2.912574	2.172712	1.340524	0.1967

R-squared 0.090772 Mean dependent var 54.25000
Adjusted R-squared 0.040259 S.D. dependent var 22.37686
S.E. of regression 21.92180 Akaike info criterion 9.107479
Sum squared resid 8650.172 Schwarz criterion 9.207052
Log likelihood -89.07479 F-statistic 1.797005
Durbin-Watson stat 0.836087 Prob(F-statistic) 0.196750

Regression #3 Dependent Variable: Y
Method: Least Squares
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Constant	42.00000	6.015027	6.982512	0.0000
Z	24.50000	8.506533	2.880139	0.0100

R-squared 0.315464 Mean dependent var 54.25000
Adjusted R-squared 0.277435 S.D. dependent var 22.37686
S.E. of regression 19.02119 Akaike info criterion 8.823623
Sum squared resid 6512.500 Schwarz criterion 8.923197
Log likelihood -86.23623 F-statistic 8.295202
Durbin-Watson stat 1.181612 Prob(F-statistic) 0.009963

Answer:

Following the previous problem, $\hat{\beta}_1 = \dfrac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0} = \dfrac{24.5}{24} = 10.208$

3. (From 15.7) The following is a simple model to measure the effect of a school choice program on standardized test performance (see Rouse[1998])

   $score = \beta_0 + \beta_1 choice + \beta_2 faminc + u$

   Where *score* is the score on a statewide test, *choice* is a binary variable indicating whether a student attended a choice school in the last year, and *faminc* is family income. The IV for *choice* is *grant*, the dollar amount granted by the government to students to use for tuition at choice schools. The grant amount differed by family income level, which is why we control for *faminc* in the equation.

   (a) Even with *faminc* in the equation, why might choice be correlation with *u*?

Even at a given income level, some students are more motivated and more able than others, and their families are more supportive (say, in terms of providing transportation) and enthusiastic about education. Therefore, there is likely to be a self-selection problem: students that would do better anyway are also more likely to attend a choice school.

   (b) If within each income class, the grant amounts were assigned randomly, is *grant* uncorrelated with *u*?

Assuming we have the functional form for faminc correct, the answer is yes. Since u_1 does not contain income, random assignment of grants within income class means that grant designation is not correlated with unobservables such as student ability, motivation, and family support.

   (c) What other condition needs to be satisfied for *grant* to be a good instrument for *choice*?

Grant needs to be correlated with choice: it seems plausible here that larger grants make it more likely that families will send their child to a choice school.

   (d) Write the reduced form equation for *choice* (that is, choice as a function of all exogenous variables). What is needed for *grant* to be partially correlated with *choice*?

The reduced form is

$$choice = \pi_0 + \pi_1 faminc + \pi_2 grant + v_2,$$

and we need $\pi_2 \neq 0$. In other words, after accounting for income, the grant amount must have some affect on choice. This seems reasonable, provided the grant amounts differ within each income class.

   (e) Write the reduced form equation for *score* (that is, score as a function of all exogenous variables). Explain why this equation is useful. How do you interpret the coefficient on *grant*?

The reduced form for score is just a linear function of the exogenous variables

$$score = \alpha_0 + \alpha_1 faminc + \alpha_2 grant + v_1.$$

This equation allows us to directly estimate the effect of increasing the grant amount on the test score, holding family income fixed. From a policy perspective this is itself of some interest.