Assignment 3

ECON 485: Causal Inference

Isaac Shon

October 13, 2023

Part I: Generating Data & IV Simulations

Consider the following data generating process:

$$log(Y_i) = \tau T_i + \beta_0 + \beta X_i^T + \gamma A_i^T + \epsilon_i \tag{1}$$

Where Y_i = Wages, $T_i = \mathbb{1}$[Person i finished high school], X_i = Observed covariates (e.g. parents' education), A_i = Unobserved covariates (e.g. ability) and ϵ_i = Error term. For simplicity, assume that X, A, and ϵ are all standard normal, i.e. distributed $\mathbb{N}(0, 1)$; and set $\tau = \beta_0 = \beta = \gamma = 1$. We also have three scholarships, Z_1, Z_2, and Z_3, which affect the probability of attendance in high school. Z_1 and Z_2 are randomly assigned with probability 0.5. Z_3 is almost exogenous - it has some randomness but some dependence on ability:

$$Z_3 = \mathbb{1}[\epsilon_3 + A_i > 0] \tag{2}$$

where $\epsilon_3 \sim \mathbb{N}(0, 3)$. T is positively correlated with both X and A, as well as with the instruments. The high school attendence equation is:

$$T_i = \mathbb{1}[5Z_{1i} + 0.01Z_{2i} + Z_{3i} + X_i + 10A_i + \epsilon_T > 0], \tag{3}$$

where $\epsilon_T \sim \mathbb{N}(0, 2)$.

Question 1: *Simulate 1000 observations of these data using Stata. Report a summary of the data.*

Table 1: Summary Table of Select Variables

	mean	sd	min	max
$log(Y_i)$	1.574375	2.007503	-4.416014	7.608584
T_i	.621	.4853809	0	1
X_i	-.0523292	.9954492	-3.123622	3.235672
A_i	.0409977	1.006152	-3.15944	3.385867
ϵ_i	-.0352937	1.041046	-3.569683	3.31665
Z_1	.51	.5001501	0	1
Z_2	.516	.499994	0	1
Z_3	.497	.5002412	0	1
ϵ_3	-.0814875	2.895713	-8.41365	7.965618
ϵ_T	-.069308	1.944285	-5.544947	5.840926
N	1000			

Question 2: *Generate a table of the parameter estimates. It should have 5 columns (OLS, IV instrumenting with Z_1, IV instrumenting with Z_2, IV instrumenting with Z_3, and IV instrumenting with Z_1 and Z_2). For the IV regressions, use the command ivreg2. You should always be using robust standard errors; nobody assumes homoskedasticity. Stata produces pre-formatted LaTeX tables using the outreg or esttab commands.*

	(1) OLS	(2) IV w/ Z_1	(3) IV w/ Z_2	(4) IV w/ Z_3	(5) IV w/ Z_1 and Z_2
T_i	2.645***	0.799	2.785	3.038***	0.799
	(0.079)	(0.611)	(11.073)	(0.360)	(0.611)
X_i	0.884***	0.983***	0.877	0.863***	0.983***
	(0.036)	(0.055)	(0.595)	(0.041)	(0.055)
Constant	-0.042	1.132***	-0.131	-0.292	1.132***
	(0.062)	(0.393)	(7.040)	(0.231)	(0.393)
N	1000	1000	1000	1000	1000

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Question 3: *Test statistically whether Z_1 is correlated with X, and report the results of that test. Use precise language to discuss whether this is a convincing suggestive test of the exclusion restriction when analyzing returns to education in a developing country.*

Table 2: Relationship Between X and Z_1

	(1) Dep. var = X
Z_1	0.039
	(0.063)
Constant	-0.072
	(0.045)
N	1000

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

There does not appear to be a statistically significant relationship between X and the scholarship Z_1. However, finding whether our scholarship Z_1 is correlated with X is not a convincing test of the exclusion restriction. The exclusion restriction condition states that an instrument Z_1 must have no **direct** effect on the outcome Y_i and that the **only effect** of the scholarship on wages is through whether or not a student was able to finish high school (T_i), and that $Cov(Z_1, u) = 0$. During our data generation process, scholarship Z_1 was randomly assigned independently of the observed covariates X and unobserved covariates A (i.e., independent of parents' education and inherent abilities, respectively). It also seems reasonable to suppose that the scholarship Z_1 may have an effect on T_i, and that the only way Z_1 affects Y_i would be solely through T_i. The exclusion restriction is untestable, but we can see if the instrument might at least be relevant if we see strong correlations between Z_1 and the endogenous regressor T.

Question 4: *Now, re-generate the data with 100,000 observations. Report the summary statistics.*

Table 3: Summary Table of Select Variables

	mean	sd	min	max
$log(Y_i)$	1.605847	1.989658	-6.679304	9.485506
T_i	.61168	.4873704	0	1
X_i	-.0063222	.9993476	-4.312932	4.056283
A_i	-.0018753	1.001005	-4.812684	4.475568
ϵ_i	.0023645	.9995656	-4.473359	4.013322
Z_1	.50168	.4999997	0	1
Z_2	.50198	.4999986	0	1
Z_3	.50073	.500002	0	1
ϵ_3	.0108799	2.996293	-14.25454	12.38698
ϵ_t	-.0074057	2.007894	-8.219564	9.36622
N	100000			

Question 5a: *Now, we will construct the Wald estimator. In constructing this Wald estimator, we have the choice of conditioning on X or not. Is this important here? Why or why not? Your answer probably will have two parts, and will involve the words "consistency" and "efficiency."*

In population, the Wald estimator is given by:

$$\lambda = \frac{\rho}{\phi} = \frac{\mathbb{E}[log(Y_i)|Z_i = 1] - \mathbb{E}[log(Y_i)|Z_i = 0]}{\mathbb{E}[T_i|Z_i = 1] - \mathbb{E}[T_i|Z_i = 0]} \tag{4}$$

The Wald estimator is given by taking the difference in wage outcomes between the groups that were intended vs unintended to receive the scholarship (ρ), and dividing it by the difference in the high school completion status of those groups (ϕ). Because our treatment variable T is endogenous, basic OLS estimates of the effect of T on Y are biased/inconsistent because T is correlated with both observed and unobserved traits X and A (as well as our instruments Z_1, Z_2, Z_3).

If we condition on X when constructing our Wald estimator, we notice significant changes in the coefficient on T, so in terms of consistency the Wald estimator is still unbiased when we include the exogenous variable X. However, we find that X will reduces the standard error on the coefficient on T, thus when we condition on X when constructing our Wald estimator we increase its efficiency.

Question 5b: *There are four expectations in the Wald estimator. Construct and report the sample averages of each of those four expectations.*

Table 4: Subsample Means Conditional on Z_1

$Z_1 =$	0	1
$log(Y_i)$	1.51122	1.69984
T_i	0.52003	0.70272
N	100000	

Question 5c: *Construct the Wald estimator using those subsample means from 5b.*

$$\widehat{\beta}_{IV} = \frac{\overline{y}_1 - \overline{y}_0}{\overline{T}_1 - \overline{T}_0} \tag{5}$$

$$\widehat{\beta}_{IV} = \frac{1.699836 - 1.511224}{0.7027189 - 0.5200273} = 1.0324089$$

Question 5d: *Estimate the analogous Wald estimator directly using ivreg2. By "analogous," I mean that the coefficient estimate should be identical.*

Table 5: IV w/ Z_1

	(1)
T_i	1.03241***
	(0.05960)
Constant	0.97434***
	(0.03684)
N	100000

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Part II: Replication of Angrist, J.D. & Lavy, V. (1999)

Use the dataset provided called "final4.dta". This comes from: Angrist, J. D., & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. The Quarterly Journal of Economics, 114(2), 533–575.

Question 6a: *Replicate columns 7-12 of Table 2.*

Table 6: OLS Estimates for 1991

	(7)	(8)	(9)	(10)	(11)	(12)
Class Size	0.135***	-0.054**	-0.042	0.211***	0.050*	0.003
	(0.030)	(0.023)	(0.027)	(0.033)	(0.030)	(0.035)
Percent disadvantaged		-0.339***	-0.341***		-0.288***	-0.281***
		(0.013)	(0.014)		(0.014)	(0.015)
Enrollment			-0.004			0.015***
			(0.004)			(0.005)
Root MSE	7.941	6.641	6.642	8.670	7.838	7.827
R^2	0.012	0.309	0.309	0.024	0.202	0.205
N	2055.000	2055.000	2055.000	2055.000	2055.000	2055.000

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Question 6b: *Interpret the coefficient on class size in columns 7 and 9. Why do these change?*

As can be seen in the table above, in column 7 the coefficient on class size is +0.135, which is statistically significant at the $\alpha = 0.01$ level. This differs from the results from the original article due to differences in the sample size. Here, we see that as the class size grows by one student, average reading comprehension scores appear to increase by 0.135. However, as we fit the control variables of enrollment and percent disadvantaged into our regression model, we find that in column 9 the relationship between class size and average verbal scores is negative (coefficient of -0.042) and is not a statistically significant variable at all. This is due to omitted variable bias, where the effect of our excluded variables of enrollment and percent disadvantaged on the average verbal scores were instead captured in the class size variable, which biased our specification in column 7.

Question 6c: *Interpret the coefficient on class size in columns 10 and 12. Why do these change?*

As can be seen in the table above, in column 10 the coefficient on class size is +0.211, which is statistically significant at the $\alpha = 0.01$ level. This differs from the results from the original article due to differences in the sample size. Here, we see that as the class size grows by one student, average mathematics scores appear to increase by 0.211. However, similar to columns 7-9, as we fit the control variables of enrollment and percent disadvantaged into our regression model, we find that in column 12 the relationship between class size and average math scores is closer to 0 (coefficient of 0.003) and is not a statistically significant variable at all. This is again due to omitted variable bias, where the effect of our excluded variables of enrollment and percent disadvantaged on the average math scores were instead captured in the class size variable, which biased our specification in column 10.

Question 6d: *Follow the "partialling out" approach to obtain the estimate of β_1 (coefficient on class size) in column 12. Present this set of regressions to show that with this process you arrive at the same exact point estimate for β_1 as you did in part (c). You will need to run the specification for column 12 first, then estimate the auxiliary regression with the following condition: "if e(sample)" to keep your sample the same (this is just a weird thing about this particular dataset). So the regression syntax for the auxiliary regression will look like this: regress ... if e(sample).*

Table 7: OLS and Partialled-Out Estimates for 1991

	(12)	Partialled-Out
Class Size	0.003	
	(0.035)	
classize_hat		0.003
		(0.037)
Percent disadvantaged	-0.281***	
	(0.015)	
Enrollment	0.015***	
	(0.005)	
Constant	71.480***	68.856***
	(1.002)	(0.194)
N	2055	2055

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Question 6e: *Present two scatterplots: one of the raw data (scatter math score and class size, with a fitted line from the regression in column 10) and with a scatterplot + predicted line depicting the point estimate from part (d).*

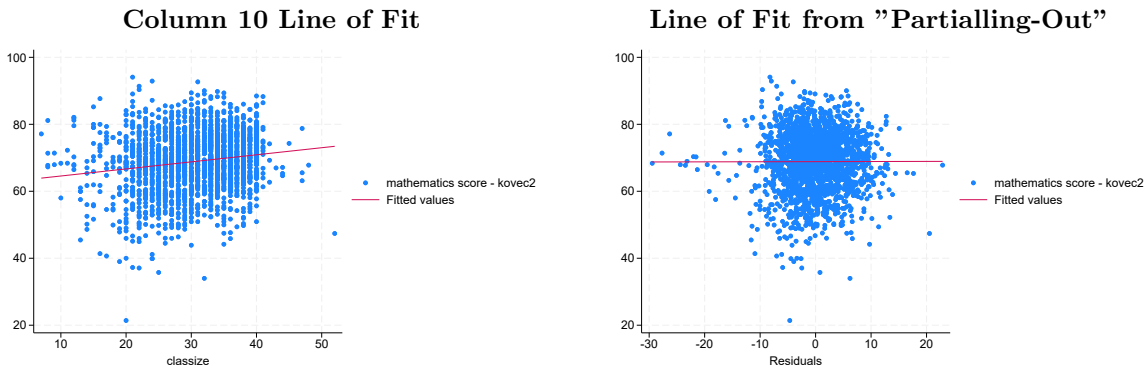

Figure 1: Scatterplots for 6e