
Problem Set 6 *Solutions*

Question 1. Convince yourself that IV works using simulated data!

- (a) Create a simulated dataset using the syntax below. (Note this is the same syntax used in the in-class exercise). We are going to assume that w is an unobserved variable. It's used to create x and y here, but is unobserved to the researcher. **(9 points)**

```
clear
set obs 1000
gen z = rnormal()
gen w = rnormal()
gen x = -2*z + 2*w + rnormal()
gen y = 5*x + 10*w + rnormal()
```

Based on the above, briefly explain how z , x , and y are related in the population (in words). Using the `corr` command, how are they correlated in the data? **(3 points)**

Based on the data generating process above, y is determined by x and w (and a random error term). x is determined by the levels of z and w (and a random error term). In the output below, we see that all three variables are correlated in the data: x is strongly positively related to y and z is strongly negatively related to x . This is as expected given the syntax used to generate the data. Notice also that z is strongly negatively related to y . While z does not appear in the structural equation for y , it is strongly related to x , and x is in the structural equation for y .

```
. corr y x z
(obs=1,000)
```

	y	x	z
y	1.0000		
x	0.9485	1.0000	
z	-0.4265	-0.6496	1.0000

- (b) Estimate a simple regression of y on x . What is the estimated slope? Is this an unbiased estimate of the population slope? Why or why not? **(3 points)**

The estimated slope below is 7.3 with a 95% confidence interval of [7.1, 7.4]. Notice the confidence interval does not contain the known population slope of 5. This is because of the omitted unobserved variable w that is correlated with both x and y . w is positively related to x and y , so the estimated slope is biased upward.

```
. reg y x
```

Source	SS	df	MS	Number of obs	=	1,000
Model	521259.968	1	521259.968	F(1, 998)	=	8943.14
Residual	58169.448	998	58.28602	Prob > F	=	0.0000
				R-squared	=	0.8996
				Adj R-squared	=	0.8995
Total	579429.416	999	580.009426	Root MSE	=	7.6345

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	7.287046	.077056	94.57	0.000	7.135835	7.438256
_cons	-.1687383	.2414438	-0.70	0.485	-.642534	.3050574

- (c) Now calculate $\hat{\sigma}_{zy}/\hat{\sigma}_{zx}$ (the sample covariance between z and y divided by the sample covariance between z and x). How does this compare to the slope in part (b)? To the population slope on x ? **(3 points)**

See below. The estimate in this case is 5.04, very close to the true population slope of 5. It is clearly smaller than the biased OLS estimate of 7.3.

```
. corr y x z, cov
(obs=1,000)
```

	y	x	z
y	580.009		
x	71.604	9.82621	
z	-10.22	-2.02604	.989872

```
. matrix list r(C)
```

```
symmetric r(C)[3,3]
```

	y	x	z
y	580.00943		
x	71.604018	9.8262067	
z	-10.219962	-2.0260379	.98987202

```
. display el(r(C),3,1) / el(r(C),3,2)
5.0443094
```

Question 2. To obtain a consistent estimate of the causal effect of family size on female labor supply, some authors have suggested using twins on their first birth as an instrument for the number of children in the household. A twin birth is arguably unexpected, and by definition, the realization of a twin increases the number of children in the household relative to a singleton birth. The Stata dataset *twins1sta.dta* was created from the 1980 Public Use Micro Sample 5% Census data files, and includes women aged 21-40 with at least one child. The 1980 PUMS identifies household members' age at the time of the census and their quarter of birth. We infer that any two children in the household with the same age and quarter of birth are twins. In the data there are roughly 6,000 mothers of twins. While there are over 800,000 observations in the original data set, a random sample of 6,500 non-twin births has been retained, for a total of about 12,500 observations. **(45 points)**

- (a) What fraction of mothers in the sample worked in the previous year? What is the average weeks worked among women that worked? What is the median labor earnings for women who worked? **(3 points)**

See attached log. 60.4% of these mothers worked. Those who did work worked an average of 38.3 weeks with median earnings \$5,505 (this was 1979).

- (b) Construct an indicator variable *second* that equals 1 for women that have two or more children (and zero otherwise). What fraction of women had two or more children? Estimate a simple bivariate regression where *weeks* of work is regressed on *second*. Interpret the slope coefficient in words. Explain why this regression is likely to suffer from omitted variables bias, and speculate on the direction of the bias. **(5 points)**

See attached log. 85.5% of mothers had at least two children. The slope estimate of -6.8 tells us that women with 2 or more children worked 6.8 fewer weeks, on average, than those with 1 child. This regression likely suffers from omitted variables bias since the decision to have more children is endogenous. If women who expect to earn less in the labor market decide to stay home and raise more children, for example, this would produce the same negative association.

- (c) Try using twins on first birth (*twin1st*) as an instrument for *second* in the main regression model of interest: estimate the first-stage and reduced-form, then calculate the Wald estimate. (Again, *weeks* of work is the outcome of interest). Interpret the slope coefficients in both regressions, and compare the IV (Wald) estimate to the OLS. What is the R^2 from the regression of *second* on *twin1st*? **(5 points)**

See the first stage and reduced form regressions in the attached log. The Wald estimate is the reduced form (-0.99) divided by the first stage (0.275), or -3.6. This is nearly half the size of the OLS estimate in absolute value, which makes sense if we believe OLS overstates the effect of family size on

labor market participation (i.e., it reflects the influence of omitted variables associated with lower labor market participation).

The first stage slope coefficient tells us that mothers with twins on their first birth were 27.5 percentage points more likely to (ultimately) have 2 or more children than mothers who did not have twins. The reduced form slope coefficient tells us that mothers with first birth twins worked about 1 week less, on average, than mothers who did not have twins. The first stage slope coefficient (0.275) is not equal to 1.0 since many women who did *not* have twins went on to have 2 or more children. The R^2 from the first stage is 0.15.

- (d) Repeat part (c) but use `ivregress` 2SLS and compare your results. Estimate the model a second time but allow for heteroskedasticity by using the robust standard errors. Does this change your inference about the slope coefficient β ? (4 points)

See attached log. The coefficient of -3.6 on *second* is identical to the Wald estimate in part (c). The heteroskedasticity-robust standard errors are virtually the same as the traditional standard errors, leading to the same inference.

- (e) Carefully state the assumptions required for interpreting $\hat{\beta}_{IV}$ in this case as an estimate of the causal effect of having two or more children on mothers' labor supply. (4 points)

The assumptions required for causal inference are: (1) instrument relevance: non-zero covariance between the instrument and explanatory variable ($\text{Cov}(Z, X) \neq 0$), and (2) the independence/exclusion restriction: no covariance between the instrument and error term in the structural equation ($\text{Cov}(Z, u) = 0$). In this application, there must be a significant association between having twins on the first birth and the propensity to have two or more children; the first stage regression provides strong evidence for this. Independence means the instrument (twins on first birth) is uncorrelated with other factors in the error term of the weeks worked equation. This seems unlikely, if some women are systematically more likely to have twins on their first birth (e.g., women who use IVF).

- (f) You are concerned that twin births are not entirely random, and convey some information about the mother. Add the following seven covariates to your regressions: mother's education, age at first birth, current age, married, white, Black, other race. (You will need to create dummy variables for the last three in this list). Which of these have statistically significant relationships with *twin1st*? Are they meaningful in size? How does this affect your results vis-a-vis the model without covariates? (5 points)

See attached log. With the covariates, the coefficient of -3.84 on *second* is

similar to the model without covariates. There is a slight difference since the covariates are correlated with the twins instrument.

- (g) You remain concerned that the covariates do not fully account for correlation between the instrument and the error term, which could lead to inconsistency. This remaining correlation would be especially problematic if the instruments were weak. Conduct a weak instruments test after part (f) and report your conclusion. (4 points)

The first stage F statistic is very large (see log). Inconsistency could be a problem in the presence of weak instruments, but this does not appear to be a concern here.

- (h) OLS would be preferable if in fact family size (as represented here by *second*) were exogenous. Explain why. Conduct a test for endogeneity following the model in part (f) and report your conclusion. (4 points)

See attached log. The null hypothesis in the Durbin-Wu-Hausman test is that the explanatory variable of interest (*second*) is exogenous. The large test statistic and small p-value leads us to reject this hypothesis, suggesting that IV is appropriate.

- (i) Create three new dummy variables that indicate whether the mother's age at first birth was before age 20, between ages 20 and 24 (inclusive), or above age 24. Call these *age1st1*, *age1st2*, and *age1st3*. Next, create variables called *twin1st1*, *twin1st2*, and *twin1st3* that are interactions between the *age1st* variables and *twin1st*. Estimate a first stage regression that includes all of the covariates in (f), the three new *age1st* dummy variables and the three interactions. (Leave out the original *agefst*). Explain why the interaction terms can be considered instruments, and why they (might) improve upon the original single instrument *twin1st*.

Use an F-test to test two different hypotheses. First, test whether the coefficients on all three instruments are the same. Then, test whether the coefficients on all three instruments are zero. (Use the `test` command after `regress`). (5 points)

See attached log. For comparison, the original first stage had a coefficient on *twin1st* of 0.285. The new first stage includes the new “age at first birth” dummies (with one category necessarily omitted) and the new instruments: interactions between the age at first birth dummies and twins on first birth. First, notice that women who are older at their first birth are less likely to have second children. Second, notice that the effect of having twins on having 2+ children is larger for older women. This makes sense if the counterfactual (older women who don't have twins on their first birth) are less likely to have 2+ children. Both F tests reject the null hypothesis. So there is strong evidence that the effect of twins differs by age at first

birth, and strong evidence that the instruments jointly explain variation in *second*.

- (j) Finally, estimate the 2SLS model from part (f) but using the new set of three instruments created in (i). How does your result compare to that in part (f)? Compare both the point estimate and standard error. Conduct a test of over-identifying restrictions. What is the null hypothesis for this test, and what is the conclusion? **(6 points)**

The first stage and 2SLS estimates are reported below. The 2SLS coefficient estimate for *second* is -3.37 with a standard error of 1.36. This is very similar to the results in part (g). The overid test is also shown. There are 2 degrees of freedom, the total number of additional restrictions. (Three instruments minus one endogenous explanatory variable). We cannot reject the null hypothesis that the model is appropriately specified.

Question 3. This problem will examine the role of measurement error using the dataset *cps87.dta* on Github. These data are a subsample of working men from the Current Population Survey of 1987. **(16 points)**

- (a) First create a variable that is the natural log of weekly earnings (*lnweekly*) and regress this on the individual's years of education (*years_educ*). What is the estimated slope coefficient and standard error? **(2 points)**

See log. The estimated slope coefficient on *years_educ* is 0.074, with a standard error of 0.0012. The interpretation is a predicted 7.4% increase in weekly earnings with every additional year of education.

- (b) Now create a “random noise” variable drawn from the standard normal distribution: `gen v=rnormal(0,1)`. Add this random noise to the years of education variable to create an education variable measured with classical measurement error (call it *years_educ2*). What are the means and standard deviations of *years_educ*, *years_educ2*, and *v*? **(2 points)**

See log. The mean of the original years of education variable is 13.16. The mean of the new (noisy) education variable is 13.17, only slightly higher. In expectation, the new variable should have the same mean, but my mean for *v* turned out to be a little higher than 0. By construction, the standard deviation of *v* is close to 1. The standard deviation of the original education variable is 2.80 years, while the standard deviation of the new variable is 2.96 years. Note the increase is not 1; that is, adding a random variable *v* with a standard deviation of 1 does not increase the standard deviation by 1. Why? Let years of education be *x*. If *x* and *v* are uncorrelated, we know that $\text{Var}(x + v) = \text{Var}(x) + \text{Var}(v)$. However, it is not the case that $\text{SD}(x + v) = \text{SD}(x) + \text{SD}(v)$.

- (c) In our model of measurement error, we distinguished between the observed (noisy) measure *x**, the true measure *x* and the random noise *e₀*. Here, those variables are *years_educ2*, *years_educ*, and *v*. Regress log weekly earnings on *years_educ2* rather than *years_educ*. What is the estimated slope coefficient and standard error, and how does it compare to part (a)? Does this change make sense to you? Explain. **(2 points)**

See log. The estimated slope coefficient on the noisy measure of education is 0.066, with a standard error of 0.0011. That the slope coefficient is smaller in absolute value than the one in part (1) is expected, since classical measurement error in the explanatory variable will attenuate the slope estimate (that is, bias it toward zero).

- (d) Calculate the “reliability ratio” (or attenuation factor) below. How does it compare to the ratio of slope coefficients in (c) and (a)? **(2 points)**

$$\frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}$$

See log. The attenuation factor is 0.888, which is approximately the ratio of the slopes in (c) and (a): $0.0659/0.0741 = 0.889$.

- (e) Repeat parts (b)-(d) but with a “noisier” v term: `gen v2=normal(0,2)`. How does this change the estimated slope coefficient, standard error, and reliability ratio when regressing log weekly earnings on the mis-measured education variable? (4 points)

See log. The estimated coefficient is now 0.048, with a standard error of 0.001. The slope estimate is attenuated further toward zero. Accordingly, the reliability ratio is smaller, at 0.657.

- (f) Finally, create a mis-measured version of log weekly earnings: `gen y2=lnweekly+v`. Regress this on the (correct) measure of education, *years_educ*. How do the slope coefficient and standard error compare with earlier results? (4 points)

See attached log. The slope coefficient of 0.070 is now close to the original OLS estimate of 0.074, and the standard error (0.0028) is higher than the original (0.0012). This is expected since classical measurement error in the dependent variable does not bias the OLS estimator, but does make it less precise.

Question 4. A researcher has collected data on alcohol consumption for 50 students each from 100 different colleges. The outcome of interest (y_i) is the number of drinks consumed in the past 30 days. The researchers have developed an index (x_i) that represents the strictness of a college's alcohol use policy with higher values meaning a more strict policy. The authors are interested in the following model:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

The researchers are concerned about measurement error in y_i . In particular, they believe that students at schools with stricter alcohol policies may be less likely to report actual drinking because they are not supposed to drink. In this case, let y_i be actual consumption and y_i^* be reported consumption: $y_i^* = y_i + e_i$. We will assume that $E(u_i) = 0$ and that $Cov(x_i, u_i) = 0$, but the measurement error is systematic such that $Cov(e_i, x_i) < 0$. In this case, with this form of measurement error, will the OLS estimate generated from a regression of y_i^* on x_i still be unbiased and consistent? If not, is the estimate biased upward or downward? Explain. (6 points)

Since we are forced to use the mismeasured y_i^* , the regression we are estimating is:

$$y_i^* = \beta_0 + \beta_1 x_i + \underbrace{u_i + e_i}_{v_i}$$

Using the OVB formula, in large samples we know that the OLS estimator $\hat{\beta}_1$ converges in probability to:

$$\text{plim } \hat{\beta}_1 = \beta_1 + \frac{Cov(x_i, v_i)}{Var(x_i)}$$

or:

$$\text{plim } \hat{\beta}_1 = \beta_1 + \frac{Cov(x_i, e_i)}{Var(x_i)}$$

since we assume the covariance between x and u is zero. If we believe there is a negative covariance between x (strictness of the alcohol policy) and e (measurement error in y), then the second term is negative. If we also believe that $\beta_1 < 0$ —the true relationship between strictness of alcohol policies and drinking is negative—then our estimated β_1 will be “too negative”.

Put another way, we are regressing reported alcohol consumption on the strictness of a college's alcohol use policy. If this relationship works as hypothesized, then $\beta_1 < 0$. That is, stricter alcohol policies reduce alcohol consumption. However, we believe that students in stricter environments are also more likely to under-report alcohol consumption. If this is the case, the relationship between alcohol consumption and the strictness of a college's alcohol use policy will be overstated. It will appear that the policies are more effective than they are.

Question 5. You are conducting a randomized experiment of an intervention designed to improve graduation rates among a vulnerable student population. Assume 50% of your study sample is offered the intervention and 50% is not. In your population, assume that 60% of individuals are “compliers,” 30% are “always takers,” and 10% are “never-takers.” (There are no defiers). These three groups have mean potential outcomes as shown in the table below. **(12 points)**

Table 1: Mean potential outcomes (graduation rates)			
	Compliers	Always-takers	Never-takers
$D_i = 1$	0.62	0.85	0.55
$D_i = 0$	0.55	0.70	0.50
Treatment effect	0.07	0.15	0.05

- (a) Calculate the intent-to-treat (ITT) effect of the intervention. **(4 points)**

Let Z_i indicate treatment assignment. The ITT is $E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)$.

Presuming random assignment worked, the $Z_i = 1$ group will consist of compliers, always-takers, and never-takers in their same proportion as in the population (60%, 30%, and 10%). The average graduation rate among this group would be: $(0.62 * 0.60) + (0.85 * 0.30) + (0.50 * 0.10) = 0.677$. Note the 0.62, 0.85, and 0.50 correspond to the *actual* treatment status (D_i) observed in these groups when $Z_i = 1$.

Similarly, the $Z_i = 0$ group will consist of compliers, always-takers, and never-takers in the same proportions as above. The average graduation rate among this group would be: $(0.55 * 0.60) + (0.70 * 0.30) + (0.50 * 0.10) = 0.635$.

Putting these two together, the ITT is $0.677 - 0.635 = 0.042$.

- (b) Calculate the first stage, and show that the IV (Wald) estimate equals the treatment effect for the compliers. (In other words, it is a LATE for the compliers). **(4 points)**

The first stage is $E(D_i|Z_i = 1) - E(D_i|Z_i = 0)$. In the $Z_i = 1$ group, 90% receive the intervention (everyone but the never-takers), so this is the first term. In the $Z_i = 0$ group, 30% receive the intervention (the always-takers), so this is the second term. The first stage is therefore: $0.90 - 0.30 = 0.60$.

The Wald estimate is the ITT/first stage, or $0.042/0.6 = 0.07$. This is the same as the treatment effect for the compliers shown in the table. Why is this the case? Notice that the graduation rates for the always-takers and

never-takers cancel out in the ITT (they are the same value, on average, in the $Z_i = 1$ and $Z_i = 0$ group.) Compliers only represent 60% of the ITT, however. (The ITT equals some value for the compliers and zero for the other two groups). Dividing by 0.6 gives you the treatment effect specific to the compliers.

- (c) Using the information in the table, what is the ATT? What is the ATE in the population? Are these different from the LATE? (4 points)

The ATT would be the average treatment effect for those treated. In this example, among those with $Z_i = 1$, the treated include the compliers and always-takers. Among those with $Z_i = 0$, the treated group includes the always-takers. Suppose the population were of size 100. The treated would include 30 compliers (50×0.6) and 30 always-takers ($50 \times 0.3 + 50 \times 0.3$). In other words, the treated would be an even split of compliers and always-takers. (That's not always the case, it just worked out that way here, since the complier group is twice the size of the always-takers group, and half of these are treated). Generally, the ATT would be a weighted average of the treatment effects for these two treated groups: $(1/2) \times 0.07 + (1/2) \times 0.15 = 0.11$. This is larger than the LATE of 0.7, which is intuitive because always-takers see a larger treatment effect than compliers. Factoring them in yields a larger ATT.

The ATE would be the average treatment effect in the *population*. This would be a weighted average of treatment effects across the three groups: $(0.60 \times 0.07) + (0.30 \times 0.15) + (0.10 \times 0.05) = 0.092$

```
.
. // *****
. // LPO-8852 Problem set 6 solutions
. // Last updated: November 8, 2023
. // *****
.
```

```
. // *****
. // Question 1
. // *****
.
```

```
. clear
```

```
. set seed 2001
```

```
. set obs 1000
```

```
number of observations (_N) was 0, now 1,000
```

```
. gen z = rnormal()
```

```
. gen w = rnormal()
```

```
. gen x = -2*z + 2*w + rnormal()
```

```
. gen y = 5*x + 10*w + rnormal()
```

```
.
. corr y x z
(obs=1,000)
```

		y	x	z
-----	+	-----	-----	-----
y		1.0000		
x		0.9485	1.0000	
z		-0.4265	-0.6496	1.0000

```
. reg y x
```

Source	SS	df	MS	Number of obs	=	1,000
-----	-----	-----	-----	F(1, 998)	=	8943.14
Model	521259.968	1	521259.968	Prob > F	=	0.0000
Residual	58169.448	998	58.28602	R-squared	=	0.8996
-----	-----	-----	-----	Adj R-squared	=	0.8995
Total	579429.416	999	580.009426	Root MSE	=	7.6345

	y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----	-----	-----	-----	-----	-----	-----
x		7.287046	.077056	94.57	0.000	7.135835 7.438256
_cons		-.1687383	.2414438	-0.70	0.485	-.642534 .3050574

```
.
. corr y x z, cov
(obs=1,000)
```

		y	x	z
-----	+	-----	-----	-----
y		580.009		
x		71.604	9.82621	
z		-10.22	-2.02604	.989872

```

. matrix list r(C)

symmetric r(C) [3,3]
      y      x      z
y  580.00943
x  71.604018  9.8262067
z -10.219962 -2.0260379  .98987202

. display e1(r(C),3,1) / e1(r(C),3,2)
5.0443094

.
.
. // *****
. // Question 2
. // *****
. // ****
. // (a)
. // ****
.
. clear

. estimates drop _all

. use https://github.com/spcorcor18/LPO-8852/raw/main/data/twins1sta.dta

.
. sum worked

      Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
      worked |   12,500      .60456      .4889646          0          1

. sum weeks if worked==1

      Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
      weeks |    7,557   38.30899   16.53096          1          52

. sum lincome if worked==1,det

      moms labor income, 1979
-----+-----
Percentiles      Smallest
1%              0          0
5%              45          0
10%             415          0      Obs          7,557
25%             2005          0      Sum of Wgt.    7,557

50%             5505
75%             9645      Largest
90%            14005      58515
95%            17005      60005      Variance      3.23e+07
99%            23005      70005      Skewness       1.727431
                        75000      Kurtosis       11.62867

. nmissing

.
. // ****
. // (b)
. // ****
. tabulate kids,miss

```

# of kids ever born to mom	Freq.	Percent	Cum.
1	1,808	14.46	14.46
2	5,958	47.66	62.13
3	3,248	25.98	88.11
4	1,054	8.43	96.54
5	318	2.54	99.09
6	75	0.60	99.69
7	24	0.19	99.88
8	11	0.09	99.97
9	3	0.02	99.99
10	1	0.01	100.00
Total	12,500	100.00	

. gen byte second=kids>=2

. tabulate second

second	Freq.	Percent	Cum.
0	1,808	14.46	14.46
1	10,692	85.54	100.00
Total	12,500	100.00	

. _eststo ols: reg weeks second

Source	SS	df	MS	Number of obs	=	12,500
Model	71801.5838	1	71801.5838	F(1, 12498)	=	140.68
Residual	6378669.1	12,498	510.375188	Prob > F	=	0.0000
Total	6450470.68	12,499	516.078941	R-squared	=	0.0111
				Adj R-squared	=	0.0111
				Root MSE	=	22.591

weeks	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
second	-6.813862	.5744749	-11.86	0.000	-7.939921	-5.687803
_cons	28.98838	.531307	54.56	0.000	27.94694	30.02983

.
. // ****
. // (c)
. // ****
. // Wald estimate
. reg weeks twinlst

Source	SS	df	MS	Number of obs	=	12,500
Model	3054.30028	1	3054.30028	F(1, 12498)	=	5.92
Residual	6447416.38	12,498	515.875851	Prob > F	=	0.0150
Total	6450470.68	12,499	516.078941	R-squared	=	0.0005
				Adj R-squared	=	0.0004
				Root MSE	=	22.713

weeks	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
twinlst	-.990038	.4068821	-2.43	0.015	-1.78759	-.1924865
_cons	23.62865	.279916	84.41	0.000	23.07997	24.17732

```
. scalar rf=_b[twinlst]
```

```
. reg second twinlst
```

Source	SS	df	MS	Number of obs	=	12,500
Model	234.976907	1	234.976907	F(1, 12498)	=	2239.20
Residual	1311.51397	12,498	.104937908	Prob > F	=	0.0000
				R-squared	=	0.1519
				Adj R-squared	=	0.1519
Total	1546.49088	12,499	.123729169	Root MSE	=	.32394

second	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
twinlst	.2746051	.0058031	47.32	0.000	.2632301 .2859801
_cons	.7253949	.0039923	181.70	0.000	.7175694 .7332204

```
. scalar fs=_b[twinlst]
```

```
. display rf/fs
```

```
-3.6053155
```

```
.  
. // ****  
. // (d)  
. // ****  
. // 2SLS  
. _eststo ivl: ivregress 2sls weeks (second=twinlst)
```

Instrumental variables (2SLS) regression	Number of obs	=	12,500
	Wald chi2(1)	=	5.97
	Prob > chi2	=	0.0145
	R-squared	=	0.0087
	Root MSE	=	22.618

weeks	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
second	-3.605315	1.475498	-2.44	0.015	-6.497239 -.7133917
_cons	26.24392	1.278193	20.53	0.000	23.73871 28.74913

```
Instrumented: second  
Instruments: twinlst
```

```
. _eststo ivlr: ivregress 2sls weeks (second=twinlst), robust
```

Instrumental variables (2SLS) regression	Number of obs	=	12,500
	Wald chi2(1)	=	5.96
	Prob > chi2	=	0.0146
	R-squared	=	0.0087
	Root MSE	=	22.618

weeks	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
second	-3.605315	1.476209	-2.44	0.015	-6.498632 -.7119987
_cons	26.24392	1.276994	20.55	0.000	23.74106 28.74679

```
Instrumented: second  
Instruments: twinlst
```

```

. // ****
. // (f)
. // ****
. gen white=race==1

. gen black=race==2

. gen other=race==3

.
. // 2SLS with covariates
. _eststo iv2: ivregress 2sls weeks educm agefst agem married black other (secon
> d=twin1st)

```

```

Instrumental variables (2SLS) regression      Number of obs   =    12,500
                                              Wald chi2(7)     =    799.03
                                              Prob > chi2      =    0.0000
                                              R-squared        =    0.0713
                                              Root MSE        =    21.892

```

weeks	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
second	-3.840711	1.388089	-2.77	0.006	-6.561314	-1.120107
educm	1.338171	.0850866	15.73	0.000	1.171404	1.504938
agefst	-1.00932	.0702044	-14.38	0.000	-1.146918	-.8717218
agem	.893219	.052759	16.93	0.000	.7898133	.9966247
married	-6.005684	.5624385	-10.68	0.000	-7.108044	-4.903325
black	2.761305	.6253911	4.42	0.000	1.535561	3.987049
other	2.651669	1.174782	2.26	0.024	.3491376	4.9542
_cons	8.371989	1.810752	4.62	0.000	4.822981	11.921

```

Instrumented:  second
Instruments:  educm agefst agem married black other twin1st

```

```

. _eststo iv2r: ivregress 2sls weeks educm agefst agem married black other (seco
> nd=twin1st), robust

```

```

Instrumental variables (2SLS) regression      Number of obs   =    12,500
                                              Wald chi2(7)     =    871.98
                                              Prob > chi2      =    0.0000
                                              R-squared        =    0.0713
                                              Root MSE        =    21.892

```

weeks	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
second	-3.840711	1.388178	-2.77	0.006	-6.56149	-1.119931
educm	1.338171	.0824623	16.23	0.000	1.176548	1.499794
agefst	-1.00932	.0703404	-14.35	0.000	-1.147185	-.8714552
agem	.893219	.0521858	17.12	0.000	.7909367	.9955014
married	-6.005684	.5608533	-10.71	0.000	-7.104937	-4.906432
black	2.761305	.6359378	4.34	0.000	1.51489	4.007721
other	2.651669	1.189649	2.23	0.026	.3199998	4.983338
_cons	8.371989	1.77135	4.73	0.000	4.900208	11.84377

```

Instrumented:  second
Instruments:  educm agefst agem married black other twin1st

```

```

. // ****

```



```
. // (g)
. // ****
. // F-test for weak instruments
. quietly ivregress 2sls weeks educm agefst agem married black other (second=twi
> nlst)
```

```
. estat firststage
```

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	F(1,12492)	Prob > F
second	0.2354	0.2350	0.1738	2627.73	0.0000

Minimum eigenvalue statistic = 2627.73

Critical Values # of endogenous regressors: 1
Ho: Instruments are weak # of excluded instruments: 1

	5%	10%	20%	30%
2SLS relative bias		(not available)		
2SLS Size of nominal 5% Wald test	16.38	8.96	6.66	5.53
LIML Size of nominal 5% Wald test	16.38	8.96	6.66	5.53

```
. quietly ivregress 2sls weeks educm agefst agem married black other (second=twi
> nlst), robust
```

```
. estat firststage
```

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	Robust F(1,12492)	Prob > F
second	0.2354	0.2350	0.1738	2779.11	0.0000

```
. // ****
. // (h)
. // ****
. // Endogeneity test
. quietly ivregress 2sls weeks educm agefst agem married black other (second=twi
> nlst)
```

```
. estat endog
```

Tests of endogeneity
Ho: variables are exogenous

Durbin (score) chi2(1) = 18.5511 (p = 0.0000)
Wu-Hausman F(1,12491) = 18.5653 (p = 0.0000)

```
. quietly ivregress 2sls weeks educm agefst agem married black other (second=twin1st), robust
```

```
. estat endog
```

```
Tests of endogeneity
Ho: variables are exogenous
```

```
Robust score chi2(1)      = 18.5198   (p = 0.0000)
Robust regression F(1,12491) = 18.5472   (p = 0.0000)
```

```
. // ****
. // (i)
. // ****
```

```
. gen agefst1=(agefst<20)
```

```
. gen agefst2=(agefst>=20 & agefst<=24)
```

```
. gen agefst3=(agefst>24)
```

```
. gen twin1st1=(agefst1*twin1st)
```

```
. gen twin1st2=(agefst2*twin1st)
```

```
. gen twin1st3=(agefst3*twin1st)
```

```
. reg second twin1st1 twin1st2 twin1st3 educm agefst2 agefst3 agem married black
> other
```

Source	SS	df	MS	Number of obs	=	12,500
Model	364.042163	10	36.4042163	F(10, 12489)	=	384.50
Residual	1182.44872	12,489	.094679215	Prob > F	=	0.0000
				R-squared	=	0.2354
				Adj R-squared	=	0.2348
Total	1546.49088	12,499	.123729169	Root MSE	=	.3077

second	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
twin1st1	.2272634	.010031	22.66	0.000	.2076012	.2469256
twin1st2	.2617009	.007974	32.82	0.000	.2460705	.2773312
twin1st3	.4141127	.0121261	34.15	0.000	.3903436	.4378817
educm	-.0040774	.0011842	-3.44	0.001	-.0063986	-.0017563
agefst2	-.0997083	.0088162	-11.31	0.000	-.1169895	-.0824271
agefst3	-.2954771	.0119809	-24.66	0.000	-.3189615	-.2719926
agem	.0179598	.0006236	28.80	0.000	.0167375	.0191822
married	.0939062	.0077174	12.17	0.000	.0787789	.1090336
black	-.0268822	.0088008	-3.05	0.002	-.0441332	-.0096311
other	.0011631	.0165179	0.07	0.944	-.0312145	.0335407
_cons	.2470463	.0234947	10.51	0.000	.2009929	.2930996

```
. test twin1st1=twin1st2=twin1st3
```

```
( 1) twin1st1 - twin1st2 = 0
( 2) twin1st1 - twin1st3 = 0

F( 2, 12489) = 77.48
Prob > F = 0.0000
```

```
. test twin1st1 twin1st2 twin1st3
```

```
( 1) twin1st1 = 0
( 2) twin1st2 = 0
( 3) twin1st3 = 0
```

```
F( 3, 12489) = 916.69
Prob > F = 0.0000
```

```
. // ****
. // (j)
. // ****
```

```
. _eststo iv3: ivregress 2sls weeks educm agefst2 agefst3 agem married black oth
> er ///
> (second=twin1st1 twin1st2 twin1st3), first
```

First-stage regressions

```
-----
```

```
Number of obs      =    12,500
F( 10, 12489)      =    384.50
Prob > F            =    0.0000
R-squared           =    0.2354
Adj R-squared       =    0.2348
Root MSE           =    0.3077
```

second	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educm	-.0040774	.0011842	-3.44	0.001	-.0063986	-.0017563
agefst2	-.0997083	.0088162	-11.31	0.000	-.1169895	-.0824271
agefst3	-.2954771	.0119809	-24.66	0.000	-.3189615	-.2719926
agem	.0179598	.0006236	28.80	0.000	.0167375	.0191822
married	.0939062	.0077174	12.17	0.000	.0787789	.1090336
black	-.0268822	.0088008	-3.05	0.002	-.0441332	-.0096311
other	.0011631	.0165179	0.07	0.944	-.0312145	.0335407
twin1st1	.2272634	.010031	22.66	0.000	.2076012	.2469256
twin1st2	.2617009	.007974	32.82	0.000	.2460705	.2773312
twin1st3	.4141127	.0121261	34.15	0.000	.3903436	.4378817
_cons	.2470463	.0234947	10.51	0.000	.2009929	.2930996

Instrumental variables (2SLS) regression

```
Number of obs      =    12,500
Wald chi2(8)       =    759.80
Prob > chi2         =    0.0000
R-squared           =    0.0671
Root MSE           =    21.941
```

weeks	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
second	-3.370982	1.359771	-2.48	0.013	-6.036083	-.7058805
educm	1.26522	.0847011	14.94	0.000	1.099209	1.431231
agefst2	-3.65137	.494139	-7.39	0.000	-4.619865	-2.682876
agefst3	-8.971728	.6795455	-13.20	0.000	-10.30361	-7.639843
agem	.8275766	.0510731	16.20	0.000	.7274751	.927678
married	-6.164801	.5626361	-10.96	0.000	-7.267548	-5.062055
black	2.976924	.626407	4.75	0.000	1.749188	4.204659
other	2.477429	1.177288	2.10	0.035	.1699871	4.78487
_cons	-7.189732	1.711086	-4.20	0.000	-10.5434	-3.836065

```
Instrumented: second
```

```
Instruments: educm agefst2 agefst3 agem married black other twin1st1
twin1st2 twin1st3
```

```
. estat overid
```

Tests of overidentifying restrictions:

Sargan (score) $\chi^2(2) = 4.32266$ (p = 0.1152)
 Basmann $\chi^2(2) = 4.32035$ (p = 0.1153)

```
. estat first
```

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	F(3,12489)	Prob > F
second	0.2354	0.2348	0.1805	916.693	0.0000

Minimum eigenvalue statistic = 916.693

Critical Values # of endogenous regressors: 1
 Ho: Instruments are weak # of excluded instruments: 3

	5%	10%	20%	30%
2SLS relative bias	13.91	9.08	6.46	5.39
2SLS Size of nominal 5% Wald test	10%	15%	20%	25%
LIML Size of nominal 5% Wald test	22.30	12.83	9.54	7.80
	6.46	4.36	3.69	3.32

```
. estimates table ols iv*, b(%4.3f) se(%4.3f)
```

Variable	ols	iv1	iv1r	iv2	iv2r	iv3
second	-6.814	-3.605	-3.605	-3.841	-3.841	-3.371
	0.574	1.475	1.476	1.388	1.388	1.360
educm				1.338	1.338	1.265
				0.085	0.082	0.085
agefst				-1.009	-1.009	
				0.070	0.070	
agem				0.893	0.893	0.828
				0.053	0.052	0.051
married				-6.006	-6.006	-6.165
				0.562	0.561	0.563
black				2.761	2.761	2.977
				0.625	0.636	0.626
other				2.652	2.652	2.477
				1.175	1.190	1.177
agefst2						-3.651
						0.494
agefst3						-8.972
						0.680
_cons	28.988	26.244	26.244	8.372	8.372	-7.190
	0.531	1.278	1.277	1.811	1.771	1.711

legend: b/se

```
.  
.  
. // *****
```

```

. // Question 3
. // *****
. // ****
. // (a)
. // ****
. clear

. estimates drop _all

. use https://github.com/spcorcor18/LPO-8852/raw/main/data/cps87.dta

.
. gen lnweekly = ln(weekly_earn)

. _eststo parta: reg lnweekly years_educ

-----+-----
Source |      SS      df      MS      Number of obs   =    19,906
-----+-----
Model |    854.28055      1    854.28055   F(1, 19904)      =    3877.62
Residual |  4385.05814   19,904   .220310397   Prob > F          =    0.0000
-----+-----
Total |  5239.33869   19,905   .263217216   R-squared         =    0.1631
                                           Adj R-squared     =    0.1630
                                           Root MSE         =    .46937

-----+-----
lnweekly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
years_educ |   .0741141   .0011902    62.27   0.000   .0717813   .076447
_cons |   5.091872   .0160138   317.97   0.000   5.060484   5.123261
-----+-----

.
. // ****
. // (b)
. // ****
. // random noise drawn from N(0,1)
. gen v=rnormal(0,1)

. gen years_educ2 = years_educ + v

. sum years_educ years_educ2 v

-----+-----
Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
years_educ |   19,906   13.16126   2.795234         0         18
years_educ2 |   19,906   13.15335   2.966171  -1.699685   21.84595
v |   19,906   -.0079049   .9994347  -4.19409    4.329064

.
. // ****
. // (c)
. // ****
. // regress lnweekly on noisy educ
. _eststo partc: reg lnweekly years_educ2

-----+-----
Source |      SS      df      MS      Number of obs   =    19,906
-----+-----
Model |    762.986377      1    762.986377   F(1, 19904)      =    3392.60
Residual |  4476.35231   19,904   .224897122   Prob > F          =    0.0000
-----+-----
Total |  5239.33869   19,905   .263217216   R-squared         =    0.1456
                                           Adj R-squared     =    0.1456
                                           Root MSE         =    .47423

-----+-----
lnweekly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
years_educ2 |   .0660056   .0011332    58.25   0.000   .0637844   .0682269
_cons |   5.199112   .0152799   340.26   0.000   5.169162   5.229062
-----+-----

```

```

. // ****
. // (d)
. // ****
. // reliability ratio
. sum years_educ

Variable | Obs Mean Std. Dev. Min Max
-----+-----
years_educ | 19,906 13.16126 2.795234 0 18

. local varx=r(Var)

. sum v

Variable | Obs Mean Std. Dev. Min Max
-----+-----
v | 19,906 -.0079049 .9994347 -4.19409 4.329064

. local varv=r(Var)

. display `varx'/(`varx' + `varv')
.88664924

. reg lnweekly years_educ

Source | SS df MS Number of obs = 19,906
-----+----- F(1, 19904) = 3877.62
Model | 854.28055 1 854.28055 Prob > F = 0.0000
Residual | 4385.05814 19,904 .220310397 R-squared = 0.1631
-----+----- Adj R-squared = 0.1630
Total | 5239.33869 19,905 .263217216 Root MSE = .46937

lnweekly | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-----+-----
years_educ | .0741141 .0011902 62.27 0.000 .0717813 .076447
_cons | 5.091872 .0160138 317.97 0.000 5.060484 5.123261

. display _b[years_educ]*(`varx'/(`varx' + `varv'))
.06571324

. // ****
. // (e)
. // ****
. // "noisier" term drawn from N(0,2)
. gen v2=rnormal(0,2)

. gen years_educ3 = years_educ + v2

. sum years_educ years_educ3 v2

Variable | Obs Mean Std. Dev. Min Max
-----+-----
years_educ | 19,906 13.16126 2.795234 0 18
years_educ3 | 19,906 13.16669 3.440259 -4.437566 24.63555
v2 | 19,906 .0054279 2.012806 -8.229694 7.52805

. _eststo parte: reg lnweekly years_educ3

Source | SS df MS Number of obs = 19,906
-----+----- F(1, 19904) = 2535.07
Model | 591.917515 1 591.917515 Prob > F = 0.0000
Residual | 4647.42117 19,904 .233491819 R-squared = 0.1130
-----+----- Adj R-squared = 0.1129
Total | 5239.33869 19,905 .263217216 Root MSE = .48322

```

lnweekly	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
years_educ3	.0501255	.0009956	50.35	0.000	.0481741	.0520768
_cons	5.407321	.0135481	399.12	0.000	5.380766	5.433877

```
.
. // reliability ratio
. sum years_educ
```

Variable	Obs	Mean	Std. Dev.	Min	Max
years_educ	19,906	13.16126	2.795234	0	18

```
. local varx=r(Var)
```

```
. sum v2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
v2	19,906	.0054279	2.012806	-8.229694	7.52805

```
. local varv2=r(Var)
```

```
. display `varx'/(`varx' + `varv2')
.65853478
```

```
.
. reg lnweekly years_educ
```

Source	SS	df	MS	Number of obs	=	19,906
Model	854.28055	1	854.28055	F(1, 19904)	=	3877.62
Residual	4385.05814	19,904	.220310397	Prob > F	=	0.0000
Total	5239.33869	19,905	.263217216	R-squared	=	0.1631
				Adj R-squared	=	0.1630
				Root MSE	=	.46937

lnweekly	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
years_educ	.0741141	.0011902	62.27	0.000	.0717813	.076447
_cons	5.091872	.0160138	317.97	0.000	5.060484	5.123261

```
. display _b[years_educ]*(`varx'/(`varx' + `varv2'))
.04880674
```

```
.
. // ****
. // (f)
. // ****
. // mis-measured dependent variable
. gen y2=lnweekly + v
```

```
. _eststo partf: reg y2 years_educ
```

Source	SS	df	MS	Number of obs	=	19,906
Model	833.708809	1	833.708809	F(1, 19904)	=	681.38
Residual	24353.8461	19,904	1.22356542	Prob > F	=	0.0000
Total	25187.5549	19,905	1.26538834	R-squared	=	0.0331
				Adj R-squared	=	0.0331
				Root MSE	=	1.1061

y2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
years_educ	.0732163	.0028049	26.10	0.000	.0677185	.0787141
_cons	5.095784	.0377391	135.03	0.000	5.021812	5.169755

```
. estimates table part*, b(%4.3f) se(%4.3f)
```

Variable	parta	partc	parte	partf
years_educ	0.074 0.001			0.073 0.003
years_educ2		0.066 0.001		
years_educ3			0.050 0.001	
_cons	5.092 0.016	5.199 0.015	5.407 0.014	5.096 0.038

legend: b/se

```
.
. capture log close
```