Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Problem Set 5
84 total points

Problem Set 5

Instructions: Answer the following questions in a Stata do-file. Submit your problem set as do-file and/or a PDF via email to `sean.p.corcoran@vanderbilt.edu`. Use your last name and problem set number as the filename. Working together is encouraged, but all submitted work should be that of the individual student.

Question 1. 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 random 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 a person's age at the time of the census and their quarter of birth. We can infer that any two children in the household with the same age and quarter of birth are twins. There are roughly 6,000 first births to mothers that are 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. **(50 points)**

(a) Read the *twins1sta.dta* dataset from Github:

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

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)**

(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)**

(c) Try using twins on first birth (*twin1st*) as an instrument for *second* in the main regression model of interest. That is, estimate the first-stage and reduced-form regression models, 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)**

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

(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)**

(f) You are concerned that twin births are not entirely random, and convey some information about the mother. Regress the following seven variables (individually) on *twin1st* and interpret your results: 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? **(5 points)**

(g) Now expand your 2SLS models in part (d) to include the covariates listed in (f). Interpret and compare your findings to the model without covariates. **(5 points)**

(h) 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 following part (g) and report your conclusion. **(4 points)**

(i) 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 models in part (g) and report your conclusion. **(4 points)**

(j) 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)**

(k) Finally, estimate the 2SLS model from part (g) but using the new set of three instruments created in (j). How does your result compare to that in part (g)? Compare both

the point estimate and standard error. Conduct a test of over-identifying restrictions. What is the degrees of freedom for this test, and what is the conclusion? **(6 points)**

Question 2. 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)**

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

(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)**

(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)**

(c) In our model of measurement error, we distinguished between the observed (noisy) measure x^*, the true measure x and the random noise e_0. 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)**

(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)**

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

(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)**

(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)**

Question 3. 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)**

Question 4. 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)**

(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)**

(c) Using the information in the table, what is the TOT? What is the ATE in the population? **(4 points)**