Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Due November 21, 2025
74 total points

Problem Set 8

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.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. In an article published in the *American Economic Journal: Economic Policy*, Briana Ballis and Katelyn Heath (2021) used two research designs—difference-in-differences and instrumental variables—to estimate the long-run impact of providing special education services in school. Use this article (accessible here) to answer the following questions. The full paper is worth reading, but you will find these sections most useful: Introduction, IIB (sample), IIIA (for variable definitions), IIIB (IV design), and IVB (IV results). **(28 points)**

(a) What is the "treatment" variable that the authors are ultimately interested in estimating the causal effect of, and what is their instrumental variable? **(4 points)**

(b) What are the two key assumptions for causal identification in their instrumental variables approach? Be specific to this study's context. **(4 points)**

(c) What are some ways in which the exclusion restriction might be violated in this context? This assumption is impossible to test directly. Briefly, what empirical evidence do the authors bring to bear on this assumption? **(5 points)**

(d) Carefully interpret the coefficient on "treatment" in column (1) of Table 4. Does the instrument pass the weak instrument test, and how do you know? **(5 points)**

NOTE: there seems to be an error in the notes to Table 4. These are not DiD estimates, but IV estimates, as the title indicates. The main DiD estimates are in Table 3.

(e) Carefully interpret the coefficients under "high school completion" in columns (2)-(4) of Table 4 (Reduced form, OLS, and IV) and comment on how and why the OLS and IV coefficients differ. **(5 points)**

(f) Briefly explain why the treatment effects estimated in Table 4 should be considered LATEs. To whom should the effects generalize? **(5 points)**

Question 2. In this question you will estimate the effects of fertility (number of children) on women's labor supply. The data are a sample of married women aged 21-35 with two or more children from the 1980 Census. The dataset is accessible from Github **(30 points)**:

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

(a) Estimate the simple regression of weeks worked (*weeksm1*) on the variable *morekids* which equals 1 if the woman had more than two children and 0 otherwise. Interpret the slope coefficient and assess its statistical and practical significance. What proportion of women had more than two children? **(5 points)**

(b) Explain why the regression in part (a) is most likely inappropriate for estimating the causal effect of fertility on labor supply. **(4 points)**

(c) Another variable in the dataset, *samesex*, is equal to 1 if the woman's first two children were of the same sex and 0 otherwise. Are women whose first two children are of the same sex more likely to have a third child? Is the effect large? Statistically significant? **(4 points)**

(d) Make a case for *samesex* to be used as an instrumental variable for *morekids*, with attention to the two identifying assumptions for IV. **(4 points)**

(e) Use two stage least squares (2SLS) to estimate the same regression in part (a) but with *samesex* as an instrument for *morekids*. Interpret the results, and contrast with those in part (a). What does the difference between these two results suggest to you about the direction of bias, if any, in part (a)? **(4 points)**

(f) Is *samesex* a weak instrument? How do you know? **(2 points)**

(g) OLS would be preferable if in fact the original model did not suffer from OVB. Briefly explain why. Conduct a test for endogeneity following th model in (e) and report your conclusion. **(4 points)**

(h) Finally, re-estimate the model in part (e) but with the other covariates found in the dataset: *agem1, black, hipan,* and *othrace*. Do these affect the results? **(3 points)**

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)**

```
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)**