## LPO 7870 Research Design and Data Analysis II, 2024

**Assignment 3**

**Submission Guidelines:**

* **Please submit an electronic copy of your group’s solutions as a PDF file by 4:00 pm, Monday, March 4.**
* **All submissions via Brightspace in the “Assignments” section —only one submission per group.**
* Please include the Stata output (you can copy and paste the main Stata outcomes).
* Pro-tip: if you use Courier New font, Stata output will line up as it does in the output window.

**Question 1. Multiple regression with nonlinear terms.** (27 points)

Use the CPS dataset (cps\_2008.dta) utilized in previous assignments.

1. Regress wages on educ and include the square of the variable educ in the model. (You can use Stata’s factor variable notation, i.e., regress wages c.educ##c.educ) (5 points)
   1. Interpret the coefficient on educ.
   2. Interpret the coefficient on the square of the variable educ.
   3. Describe the predicted relationship between education and wages in this model.
2. Create a new variable representing the natural log of wages with the following code.

Gen lnwages=ln(wages)

Why might we make this transformation? (3 points)

1. Regress lnwages on educ and add an indicator for sex to the model, interacting sex and educ (i.sex##c.educ). How do you interpret the coefficient on educ? Does education have a differential effect on wages for men and women? (4 points)
2. Create a variable for educational attainment called educ1 with four mutually exclusive categories: less than high school, high school graduate, completed some college or an associate’s degree, college graduate (4-year degree or more). Now run a model of wages on educ1. (Note we are going back to the original wages variable). Use the code below. (10 points)

gen educ1=1 if educ<12

replace educ1=2 if educ==12

replace educ1=3 if educ>12 & educ<16

replace educ1=4 if educ>=16

label def reduced\_educ 1 "less than high school" 2 "high school graduate" 3 "some college" 4 "college graduate"

label val educ1 reduced\_educ

reg wages i.educ1

1. Interpret the coefficient on “some college”.
2. Are bachelor’s degree holders predicted to earn more than someone with some college? If so, how much more?
3. What is the predicted wage for someone who did not complete high school?
4. How much of the variation in wages is explained by the level of education?
5. Now run a model of wages on educ with indicators for race, gender, and marital status. (Note we are going back to the original educ variable). Then add the indicator variable occupation to the model (i.occupation). How do the coefficients on educ and black change when controlling for occupation? Is this evidence of omitted variable bias? Explain. (3 points)
6. Is multicollinearity a problem in the model from part (e)? Why or why not? (2 points)

**Question 2: Statistical power** (28 points – 4 each)

Part 1

Data are gathered on 16 students. Half of the students were randomly assigned to a new tutoring program while the other half continued their usual schooling experience. An impact evaluation found that test scores for the tutored students were 10 points higher, on average, than those for the other students, with a *p*-value for the *t-*test of 0.3. The mean test score overall was 100, with a standard deviation of 20.

1. Is the study’s finding *statistically significant*? Explain why or why not.
2. Is the study’s finding *practically significant*? Explain how you determined this.
3. Would you conclude that the tutoring program is effective based on these results? Ineffective? Briefly explain.
4. What is the chief limitation of this research design? Carefully explain.

Part 2

1. A researcher is interested in whether students who attend charter schools perform better academically than those who attend regular public schools. She plans to randomly assign 30 new students to attend charter schools (or not), and then test these students at the end of the year. She anticipates a 5 point advantage for charter school students, on an exam that typically has a standard deviation of 6 points. Her statistician tells her that her study design has an estimated power of 0.5962. Explain in words what this quantity is telling us. Is this a “good” level of power? Explain why or why not.
2. The same statistician from part (e) tells the researcher that the *minimum detectable effect size* for her study is 6.36. Explain in words what this quantity means.
3. How might this researcher improve the power of her study, if she wished to do so? Explain the intuition behind your answer.

**Question 3: Research paper on a textbook intervention in Kenya** (25 points – 5 each)

One of the earliest randomized controlled trials in development economics was a textbook intervention in Kenya in 1995 (See Glewwe, P., Kremer, M., & Moulin, S. 2009. “Many children left behind? Textbooks and test scores in Kenya.” *American Economic Journal: Applied Economics*, 1(1), 112–135.) The intervention was straightforward: 100 schools were allocated randomly to receive textbooks or not. (The actual intervention was a little more complicated than this, but for the purpose of this question you can assume a group of schools that received textbooks, and a group that did not). The theory of the intervention was that students were not learning because they did not have access to textbooks.

The original Table 4 of the paper—reproduced below—shows the effect of the program. The authors estimated a regression like the following:

is a standardized test score (mean 0, standard deviation 1) for student *i*, in school *j*, at time *t*. The variable is an indicator for whether the school *j* received textbooks for its students ( or not (). and are other controls. is the error term. Column 1 of Table 4 presents the effects of one year of treatment, and column two, the effects of two years of treatment, captured by the coefficient .



Use Table 4 to answer the following questions.

1. Please interpret the effect of provision of textbooks in both years. Discuss the statistical significance of both coefficients. Present all your calculations.
2. Does provide the causal effect of providing textbooks on learning? Briefly justify your answers, with reference to omitted variable bias.

In a further analysis, the authors wanted to assess if the textbooks were more beneficial for students with higher achievement at the beginning of the intervention. The data that the authors had was test scores before the intervention (“pretest score”, ). For their regression, they created an indicator variable *Pretest* = 1 for those with scores above the average and =0 for those with scores below the average. They then estimated a different regression:

The results of this regression are shown in Table 8, reproduced below:



1. Please interpret the coefficients for Textbook school () in columns 1 and 2. Are they statistically significant for year 1? For year 2?
2. Please interpret the coefficients for Pretest score () in columns 1 and 2. Are they statistically significant for year 1? Year 2?
3. Please interpret the coefficients of the interaction between Textbook school and Pretest score () in columns 1 and 2. Are they statistically significant for year 1? Year 2?