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

**Assignment 2**

**Submission Guidelines:**

* **Please submit an electronic copy of your group’s solutions as a PDF file by 4:00 pm, Monday, February 19.**
* **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. A regression estimation by hand.** (20 points – 5 each)

You have (admittedly) a very small data set with 8 observations from students in a particular college. Each individual observation includes the ACT test score in high school and the first semester GPA in college. You are interested in estimating the relationship between ACT scores and GPA. Here is the data and the scatter graph:

|  |  |  |
| --- | --- | --- |
| Student | ACT | GPA |
| 1 | 21 | 2.8 |
| 2 | 24 | 3.4 |
| 3 | 26 | 3 |
| 4 | 27 | 3.5 |
| 5 | 29 | 3.6 |
| 6 | 25 | 3 |
| 7 | 25 | 2.7 |
| 8 | 30 | 3.7 |

Based on the graph, you assume a linear model:

(1)

1. Please estimate and by hand by calculating the least squares slope and intercept. Show all your work. (Note: you can show your manual calculation in Excel if that helps).
2. Please interpret the coefficient .
3. The standard error of is 0.036. Please test the null hypothesis that is equal zero.
4. The association between college GPA and ACT scores might lead some policymakers to put more emphasis on preparation for the ACT. Does this recommendation make sense? Why or why not? In your answer, include the concept of “omitted variables bias.”

**Question 2. Regression interpretation and inference** (24 points)

Suppose that 1,000 low-income students applied for a publicly-funded voucher to attend private school. 500 were randomly selected to receive the voucher. At the end of the first year, these two groups’ reading test scores were compared using the following regression:

where *Private School* = 1 for students who received the voucher and enrolled in private school (and zero otherwise).

1. Write out the formal hypothesis to test whether differences in test scores for students in private schools versus public schools are statistically significant. (3 points)

The researchers obtained the following results from this regression:

1. Calculate the test-statistic and *p*-value for the estimated . Is statistically significant at the 95% confidence level? Show your work. (5 points)
2. What is the mean test score for students in private school? What is the mean test score for those in public school? Show your calculation. (5 points)
3. Is the estimate of practically significant? In other words, does the voucher to attend private school have a meaningfully “large” or “small” influence on test scores for these students? Explain your rationale. (4 points)
4. Should the result of this regression be interpreted as the *causal effect* of the voucher to attend private school on reading test scores? Why or why not? (4 points)
5. Evaluate the . What does this measure of fit tell us about the calculated estimate? (3 points)

**Question 3: Stata exercise** (28 points – 4 points each)

Use the CPS dataset from Assignment 1, Question 2 to assess the relationship between wages and education using multiple regression.

1. Plot the scatter plot and fitted line showing the relationship between wages and years of education using the following code:

generate wages=earnings\_pw/hours\_pw

twoway (scatter wages educ) (lfit wages educ)

Is the distribution of wages given education *homoscedastic*? Why or why not?

1. Fit a simple regression of wages on education. Use this code below to run the regression twice. (1) reg wages educ (2) reg wages educ, robust. Interpret the estimated coefficient on educ. What differences do you see in the regression output from these two approaches? Based on your response to part (a), which approach should be used for this analysis?
2. Next add controls for race and gender using the code below. (The first step is to create separate indicators for each race/ethnic group). Interpret the coefficient on education. How does it differ, if at all, from the one in part (b)?

Use this code:

gen race1="white" if race==1

replace race1="black" if race==2

replace race1="asian" if race==4

replace race1="other" if race==3

replace race1="other" if race>4

tab race1, gen(race\_)

rename race\_1 asian

rename race\_2 black

rename race\_3 otherrace

rename race\_4 white

gen female=0 if sex==1

replace female=1 if sex==2

reg wages educ female black asian otherrace, robust

1. How do we interpret the coefficient on the indicator for “female?” What about the coefficient associated with the variable “black?”
2. Now fit a regression model for wages that includes education, an indicator for female and an indicator for male as explanatory variables. Is it possible to run such model (with all these controls)? Please explain why or why not.

Use this code:

gen male=0 if sex==2

replace male=1 if sex==1

reg wages educ female male, robust

1. Using the model in Part (c) (wages against education, race, and gender) test if the point estimates for education, gender, and race are *jointly significant*. Please discuss the test that you are using.
2. Suppose you are interested in the *causal effect* of education on wages. Explain whether you believe the coefficient on education estimated in part (b) is biased or not, and any concerns you have about omitted variables.

**Question 4: Research paper on the gender gap in homework time** (28 points – 4 each)

In a 2015 paper, [Gershenson and Holt](https://journals.sagepub.com/doi/epub/10.3102/0013189X15616123) used data from the American Time Use Survey and the Educational Longitudinal Study of 2002 to estimate differences by gender in time spent doing homework during high school. They used additional data from this survey to try to “explain” the gender gap in homework time they observed (e.g., by household income, time spent in out-of-school activities, differences in coursework, or parental involvement).

The main regression the authors used was:

where *Ti* is time spent studying outside of school, *Malei*=1 for students identifying as male, *SESi* represents a set of indicator variables for parents’ socio-economic status (income and educational attainment), and *Xi* represents a set of other control variables. Note: these authors use Greek letters other than β’s to represent slope coefficients—this is ok, and very common!

Use Table 3 excerpted below to answer the following questions. *Ti* is measured here as hours per week. (Note: the table notes say that the standard errors—in parentheses—are clustered, and the regressions are weighted using sampling weights. These details are not important for this question).

A table of text with numbers

Description automatically generated with medium confidence

1. Provide a written interpretation of the coefficient on “Male” in column (1), along with its standard error. Is the gender gap in homework time statistically significant?
2. In the sample as a whole, students spent an average of 5.7 hours per week studying outside of school, with a standard deviation of 5.7 hours. Would you say the gender gap in homework time is *practically significant*? Why or why not?
3. Explain in words how the regression in column (3) differs from the one in column (1). (Don’t interpret the results yet, just explain what the authors are doing differently in column (3)).
4. Provide a written interpretation of the coefficients “College degree” (R’s father) and “HH income $50k-75k” in column (3). Which of the explanatory variables in column (3) are statistically significant at the 0.05 level or below?
5. Did the approach in column (3) change the authors’ conclusions about the gender gap in homework time? Explain how you know. Why do you think the coefficient on Male changed (or not) as much as it did between columns (1) and (3)?
6. Explain in words how the regression in column (4) differs from the one in column (3). Did this approach change the authors’ conclusions about the gender gap in homework time? What variables in column (4) are significant predictors of weekly homework time?
7. Provide an interpretation of the adjusted R-squared in column (5).