## 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 .

For every one point increase in ACT score, we predict a 0.102 point increase in GPA.

1. The standard error of is 0.036. Please test the null hypothesis that is equal to zero.

is statistically significantly different from zero. We therefore reject the null hypothesis that is equal to zero.

1. 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.”

This recommendation does not make sense as the association between GPA and ACT has not been proven to be causal utilizing this regression. We are not appropriately accounting for potential *omitted variables* that may *bias* our estimates of the causal relationship between GPA and ACT scores. For example, it may be that students who score highly on the ACT have greater resources at home that also benefit their GPA.

**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)

is not statistically significant at the 95% confidence level because the p-value is greater than 0.05 (equivalent to the absolute value of the t-stat being less than 1.96).

1. 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)
2. 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)

The estimate of is statistically insignificant, so the findings are inconclusive. That said, if the finding is “real,” one could argue that the voucher has a meaningfully large (and negative) effect on test scores. Ideally, one would compare the 10.03 point gap in scores to the overall variation in test scores (the standard deviation). As another benchmark, one could note that that the 10.03 point different amounts to (10.03/72.82) 13.8% of the mean for public school students.

1. 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)

Given the randomized design of the instituted program, if this were a significant effect, it would be interpreted as causal. (There is no reason to be concerned about omitted variables bias in a randomized controlled trial). As noted, however, the findings are inconclusive due to the lack of statistical significance.

1. Evaluate the . What does this measure of fit tell us about the calculated estimate? (3 points)

The can be interpreted as 6.7% of the variation in test scores is explained by the sector of school attended (public or private). This tells us that the difference described accounts for a small portion of the overall variation among student test scores. This does not explain the causal effect.

**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?



No, the distribution of wages given education is **NOT** homoscedastic. The variation in wages is different for different years of education. For example, there is less variation in wages for individuals with ten years of education and more variation in wages for individuals with fourteen years of education.

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?

**reg wages educ**

Source | SS df MS Number of obs = 10,651

-------------+---------------------------------- F(1, 10649) = 2659.78

Model | 325251.353 1 325251.353 Prob > F = 0.0000

Residual | 1302215.71 10,649 122.285258 R-squared = 0.1999

-------------+---------------------------------- Adj R-squared = 0.1998

Total | 1627467.07 10,650 152.813809 Root MSE = 11.058

------------------------------------------------------------------------------

wages | Coefficient Std. err. t P>|t| [95% conf. interval]

-------------+----------------------------------------------------------------

educ | 2.130196 .0413045 51.57 0.000 2.049231 2.21116

\_cons | -8.206163 .579967 -14.15 0.000 -9.343007 -7.069319

------------------------------------------------------------------------------

**reg wages educ, robust**

Linear regression Number of obs = 10,651

F(1, 10649) = 2104.34

Prob > F = 0.0000

R-squared = 0.1999

Root MSE = 11.058

------------------------------------------------------------------------------

| Robust

wages | Coefficient std. err. t P>|t| [95% conf. interval]

-------------+----------------------------------------------------------------

educ | 2.130196 .0464367 45.87 0.000 2.039171 2.221221

\_cons | -8.206163 .615163 -13.34 0.000 -9.411998 -7.000329

------------------------------------------------------------------------------

On average, one additional year of education is associated with a $2.13 increase in hourly wages. This relationship is statistically significant (p<0.001).

The standard errors of the regression coefficients are slightly larger in the second regression than the first. Relatedly, the t-statistics are slightly smaller in the second regression compared to the first.

The second regression (reg wages educ, robust) is more appropriate in this situation since it does not assume homoscedasticity.

1. 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

Linear regression Number of obs = 10,651

F(5, 10645) = 526.42

Prob > F = 0.0000

R-squared = 0.2400

Root MSE = 10.779

------------------------------------------------------------------------------

| Robust

wages | Coefficient std. err. t P>|t| [95% conf. interval]

-------------+----------------------------------------------------------------

educ | 2.163807 .04592 47.12 0.000 2.073795 2.253819

female | -4.484593 .2092242 -21.43 0.000 -4.894712 -4.074475

black | -2.980114 .3069107 -9.71 0.000 -3.581717 -2.378512

asian | .5834352 .556834 1.05 0.295 -.5080634 1.674934

otherrace | -1.711251 .6110332 -2.80 0.005 -2.90899 -.5135116

\_cons | -6.241799 .6054466 -10.31 0.000 -7.428587 -5.05501

------------------------------------------------------------------------------

On average, one additional year of education is associated with a $2.16 increase in hourly wages, holding constant gender and race. This relationship is statistically significant (p<0.001). The coefficient is slightly larger than in part B.

1. How do we interpret the coefficient on the indicator for “female?” What about the coefficient associated with the variable “black?”

On average, females earn $4.48 per hour less than males, holding constant years of education and race. This relationship is statistically significant (p<0.001).

On average, people who are Black earn $2.98 per hour less than people who are white, holding constant years of education and gender. This relationship is statistically significant (p<0.001).

1. 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

Linear regression Number of obs = 10,651

F(2, 10648) = 1243.55

Prob > F = 0.0000

R-squared = 0.2343

Root MSE = 10.818

------------------------------------------------------------------------------

| Robust

wages | Coefficient std. err. t P>|t| [95% conf. interval]

-------------+----------------------------------------------------------------

educ | 2.190271 .0460244 47.59 0.000 2.100055 2.280487

female | -4.609988 .209746 -21.98 0.000 -5.021129 -4.198846

male | 0 (omitted)

\_cons | -6.865002 .6051123 -11.35 0.000 -8.051135 -5.678869

------------------------------------------------------------------------------

No, it is not possible to run a regression model with both a control for female and male.

In this data set, every individual is classified as either female or male (i.e. they have a correlation of exactly -1). This makes it impossible to interpret the female and male indicators together. The interpretation for the female indicator would be the average difference in hourly wage between females (female=1) and males (female=0), holding all other controls in the model constant. It doesn’t make sense to talk about a difference between females and males holding male constant. A change from female=1 to female=0 necessarily implies a change in males from male=0 to male=1.

This is also called perfect multicollinearity. Female and male are perfectly correlated. Therefore, we estimate their respective relationships with wages utilizing one variable, where the constant can be interpreted as the predicted earnings for men (with zero years of education).

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.

Our null hypothesis is that there is no relationship between wages and education, race, or gender. Our alternative hypothesis is that there is a relationship between wages and at least one of these factors.

In order to test this hypothesis, I used a global F test. Based on the regression output, the F-statistic is equal to 1243.55. The p-value is less than 0.0001. Thus, there is sufficient evidence to reject the null hypothesis that there is no relationship between wages and education, race, and gender at the 0.05 significance level.

1. 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.

I think the coefficient on education **is a biased estimate of the causal effect** of education on wages. There are a number of other factors that are associated with both years of education and wages which may be captured in this estimate.

For example, individuals from families with higher income often have an easier time accessing post-secondary education and thus have more years of education than those from lower-income families. However, individuals from higher-income families may also have an easier time finding jobs with higher wages due to their families social and professional networks. In this case, we may be over-estimating the impact of education on wages because we are not accounting for the effects of socioeconomic status on both years of education and wages. (Another omitted variable could be rural versus urban location, to the extent earnings and education both tend to be lower in rural areas).

**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?

On average, males spent 1.26 fewer hours on homework per week than females.

If we were to take repeated samples from the same population, we would expect the difference we would calculate between males and females to vary by around 0.14 hours per week on average.

The difference in time spent on homework between males and females is statistically significant (p<0.01).

1. 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?

Effect size: difference/sd = 1.26/5.7 = 0.22 standard deviations

Percent difference from the mean: difference/mean = 1.26/5.7 = 0.22 -> 22% decrease for the average student

I’d say that the gender gap in homework time is practically significant. A difference of 1.26 hours represents an effect size of 0.22 standard deviation units or a decrease of 22% in time spend on homework for the average student. Over the course of a full year (~36 weeks), this weekly difference adds up to around 45 fewer hours.

A good deal of learning can occur during this amount of time so this may have practical implications for academic outcomes by gender. In contrast, time spent on homework is also time not spent doing something else. In this way, girls may be missing out on using their out-of-school time in other potentially beneficial ways.

1. 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)).

In model 3 the authors added controls for father’s education level, mother’s education level, and household income.

1. 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?

On average, students whose fathers have a college degree spend 1.49 more hours per week studying than students whose fathers don’t have a high school degree, holding constant students’ gender, mothers’ education level, and household income.

On average, students with a household income of $50k-75k spend 0.34 more hours per week studying than students with a household income of less than $25k, holding constant gender, fathers’ education level, and mothers’ education level.

Gender, Father-Some college, Father-College degree, Mother-College degree, HH Income - $100k-$200k, and HH Income >$200k are all statistically significant at the 0.05 level.

1. 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)?

No, the approach in column 3 wouldn’t change the authors’ overall conclusions about the gender gap. The estimated gap between males and female is very similar in models 1 and 3. The estimates didn’t change when adding controls because gender is generally uncorrelated with parental education and household income.

1. 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?

Model 4 includes controls for students’ academic achievement as measured by 9th grade gpa and their ELA and math assessment score quartiles.

The estimated gender gap is smaller in model 4 than in the previous models. However, it is still meaningfully large and statistically significant. I don’t think model changes the authors’ overall findings that there is a difference in time spent on homework between males and females. This model additionally shows that differences by gender persist even when accounting for gender differences in academic achievement.

Gender, Father-College Degree, Household Income >$200, 9th grade GPA, Bottom Quartile for ELA, Bottom Quartile for Math, and Top Quartile for Math are all statistically significant at the 0.05% level.

1. Provide an interpretation of the adjusted R-squared in column (5).

Gender, parental education level, house-hold income, academic achievement, other “baseline” and activity controls can explain around 14% of the variation in time spent on homework after adjusting for the number of variables in the model.