## 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)

A screenshot of a computer

Description automatically generated

* 1. Interpret the coefficient on educ.

When moving from 0 years of education to 1 year of education we predict a $0.719 decrease in wages.

* 1. Interpret the coefficient on the square of the variable educ.

For every additional increase in years of education, we predict that the slope increases by $0.105.

* 1. Describe the predicted relationship between education and wages in this model.

Taken together, the predicted relationship between education and wages decreases at a decreasing rate prior to increasing at an increasing rate. This is due to the fact that the quadratic and linear differences are fractional and of opposite sign.

Therefore, we continue to expect an increase in education to lead to an increase in wages. However, this relationship is primarily seen once a particular threshold of education is achieved.

Could also take an example of moving between different points:

0 - > 1 year of education

11 - > 12 years of education

15 - > 16 years of education

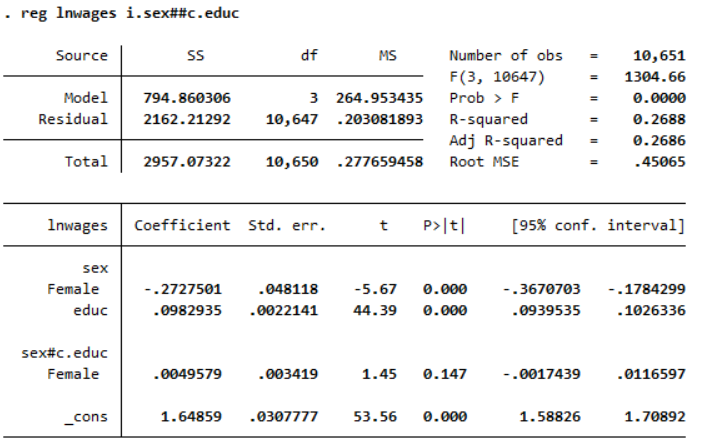
1. Create a new variable representing the natural log of wages with the following code.

gen lnwages=ln(wages)

Why might we make this transformation? (3 points)

We make this transformation in order to reduce the data dispersion in the wages variable. Since wages are often bunched at different levels and can vary quite widely, taking the natural log transformation of wages would reduce this variance and allow us to interpret differences in wages in percent terms rather than raw dollar increases, which may be meaningless in context.

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)



For every one year increase in education, we predict a 9.83% increase in wages.

Since the interactive term (educ\*sex) is not significant at the 95% significance level [t-stat = 1.45 < 1.96; p-value > 0.05], education does not have a differential effect for wages for women versus those of men.

1. 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

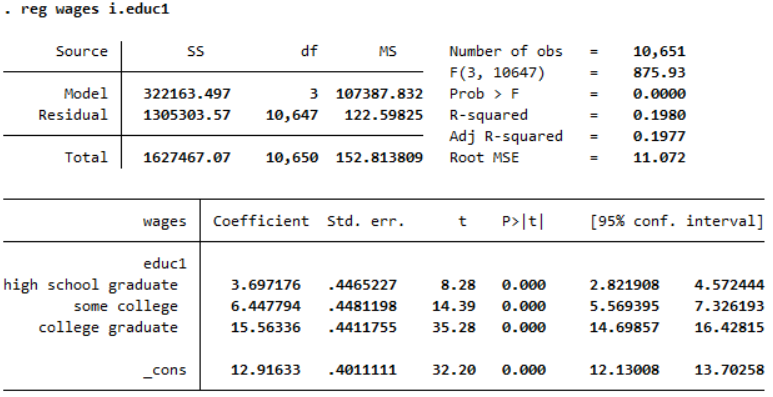
replace educ1=3 if educ>12 & educ<16

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”.

We predict that any individual with some years of college completed will have $6.45 higher wages compared to those who did not complete high school.

1. Are bachelor’s degree holders predicted to earn more than someone with some college? If so, how much more?

test 3.educ1 == 4.educ1

p-value < 0.000

Using the code above, we know that the difference in wages between those with some college and those with a college degree is different. We can also analyze both this difference and the magnitude of such from the above regression output where the 95% confidence intervals for wages do not overlap and have a minimum magnitude of 14.699 – 7.326 = 7.373.

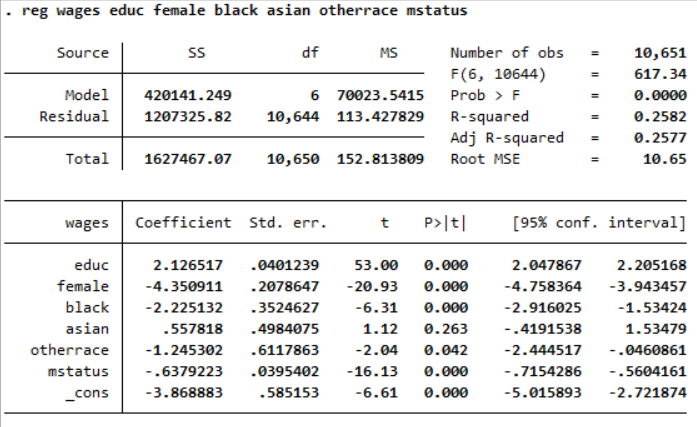
1. What is the predicted wage for someone who did not complete high school?

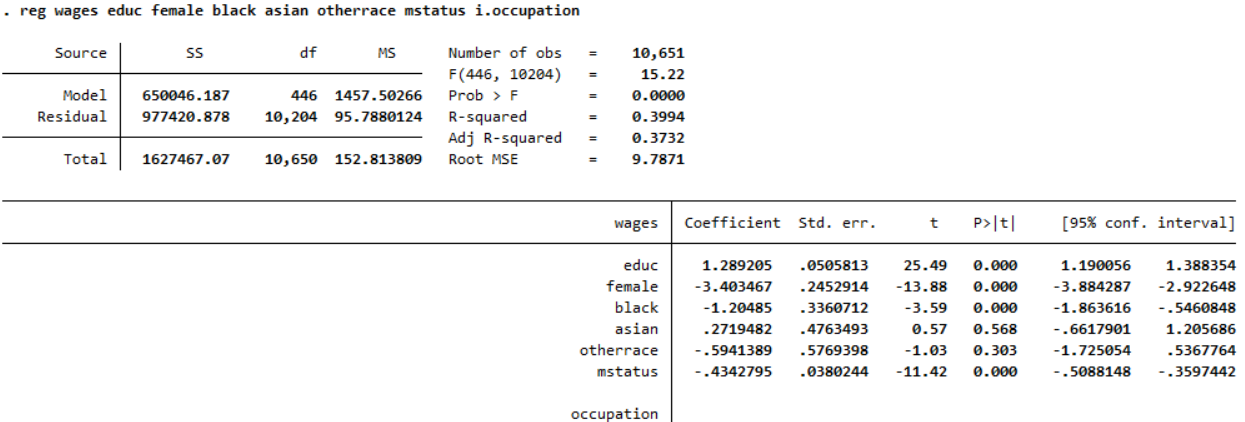
The predicted wages for someone who did not complete high school are $12.92.

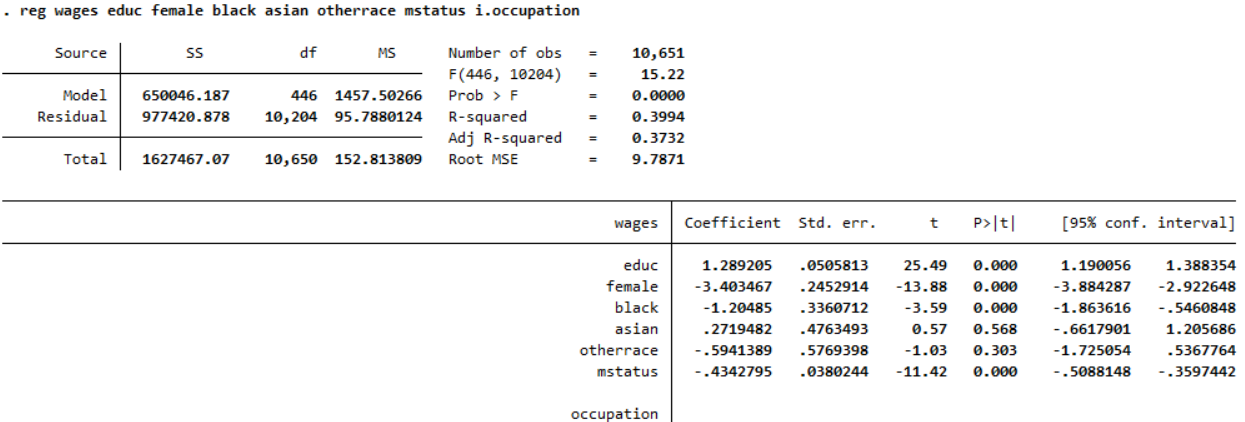
1. How much of the variation in wages is explained by the level of education?

Given the adjusted of 0.1977, 19.77% of the variation in wages is explained by these four levels of education.

1. 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)







The coefficient on educ decreases from 2.127 to 1.289. This means that in the first model we expect for every one year increase in education, wages are likely to increase by $2.127; whereas in the second model we expect for every one year increase in education, wages are likely to increase by $1.289. Measures from both models are statistically significant.

The coefficient on black increases from -2.225 to -1.205. This means that in the first model we expect that wages are likely $2.225 lower for black people than white people; whereas in the second model we expect that wages are likely $1.205 lower for black people than white people. Measures from both models are statistically significant.

We can consider this a representation of omitted variable bias if either statistical or practical significance changes. Since statistical significance does not change for either variable among both models, we examine the practical significance of the difference in wages between the two models. Both magnitudes change by approximately $1.1 per hour when controlling for occupational differences. This is approximately $528 per year for a full time job. This could be considered of practical significance. We can therefore conclude that without the control of occupation, our model may have suffered from omitted variable bias.

1. Is multicollinearity a problem in the model from part (e)? Why or why not? (2 points)

Multicollinearity could be considered a problem due to the fact that the occupations are set up in sectors that may overlap in wage structures. To solve this, we can reduce the number of occupational markers by grouping occupations into tighter wage bundles.

On the other hand, given that the entirety of occupations are not represented in this list, we have no issues of perfect multicollinearity. By including the variable as a factor variable in the regression output, it automatically removes one of the outputs and sets “Chief Executive Officers” as our baseline.

**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.

No, the results are not statistically significant at the 90%, 95%, or 99% significance levels since the p-value of 0.3 > 0.1.

1. Is the study’s finding *practically significant*? Explain how you determined this.

Although not statistically significant, the effect found is large enough to be practically significant. A difference of 10 points when divided by the standard deviation of scores is an effect size of 0.5. This is a very large effect compared to existing studies and interventions in the field of education.

1. Would you conclude that the tutoring program is effective based on these results? Ineffective? Briefly explain.

I would not be confident in saying that the program was effective given these results without qualifying my statement. Since the results are not statistically significant, there is a larger chance that they occurred due to random chance than we are typically comfortable with for establishing program effects. However, the evidence is promising so further attempts to scale up the program and evaluation may be warranted.

1. What is the chief limitation of this research design? Carefully explain.

This study is underpowered.

The sample size of the study is very small. This means that there is large sampling variability, i.e., the estimated effect could vary widely from sample to sample. This makes it more likely that we could have drawn a sample with a very large estimated effect, even if the program actually has no effect. Given this variability, we cannot rule out that large estimated effects we actually observed due to random chance. Thus, the study does not have the power to accurately detect even quite large true program effects.

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.

This means that if attending charter schools does in fact provide students with a 5-point advantage, a study with this design would be able to correctly reject the null hypothesis of no effect for 59.62% of samples.

No, this is not a good level of power. It is quite likely (~40% of samples) that we could happen to draw a sample that incorrectly fails to reject the hypothesis of no effect when there is in fact an effect. It is likely not worth the investment to run a study where the chances to accurately identify an effect are this low.

1. 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.

This is the smallest actual effect of charter school attendance that a study with this design will be able to reliability identify, i.e., correctly reject the null hypothesis of no effect for a larger percent of potential samples.

1. How might this researcher improve the power of her study, if she wished to do so? Explain the intuition behind your answer.

To improve the power of her study should could increase the number of students. This will decrease the standard error (sampling variation) of the estimated effect, allowing her to more precisely estimate the true effect and correctly reject the null for a higher percent of potential samples.

**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.

In both years, on average students in schools that received textbooks scored ~0.02 standard deviation units higher on standardized tests than students in the schools that didn’t receive textbooks. This difference is not statistically significant in either year since the 95% confidence intervals for the estimated effects include zero.

95% CI Year 1: 0.023 +/- 1.96\*0.087 -> (-0.148, 0.194)

95% CI Year 2: 0.020 +/- 1.96\*0.104 -> (-0.184, 0.224)

1. Does provide the causal effect of providing textbooks on learning? Briefly justify your answers, with reference to omitted variable bias.

Yes, does provide the causal effect of providing textbooks. Since schools were randomly assigned to treatment, on average across multiple samples whether or not you receive a textbook will be uncorrelated with any other school or student characteristics. Thus, there is no 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?

In year 1, on average students who scored below average on pretests in schools that received textbooks **scored 0.060 points higher** than students who scored below average on pretests in school that did not receive textbooks, holding constant pretest score. This difference was not statistically significant.

In year 2, on average students who scored below average on pretests in schools that received textbooks **scored 0.016 points lower** than students who scored below average on pretests in school that did not receive textbooks, holding constant pretest score. This difference was not statistically significant.

1. Please interpret the coefficients for Pretest score () in columns 1 and 2. Are they statistically significant for year 1? Year 2?

In year 1, on average for a one-point increase in students’ pretest score there is a predicted increase of 0.430-point increase in their posttest score, holding constant school treatment assignment. This difference is statistically significant.

In year 2, on average for a one-point increase in students’ pretest score there is a predicted increase of 0.342-point increase in their posttest score, holding constant school treatment assignment. This difference is statistically significant.

1. 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?

In year 1, on average the effect of receiving textbooks on posttest scores was 0.057 points larger for students who scored above averaged on the pretest than for students who scored below average on the pretest. This difference is statistically significant.

In year 2, on average the effect of receiving textbooks on posttest scores was 0.061 points larger for students who scored above averaged on the pretest than for students who scored below average on the pretest. This difference is statistically significant.