ECN 480/PUB 580
Assignment #5
Due: Thursday, March 31, 2022 by end of day

Directions: Answer each question electronically in a MS Word or .pdf file. Compile your answers into a single computer file, and then upload it to Canvas under "Assignment #5." Contact me if you have any questions.

1. Download the data set entitled "MEAP93.dta" from Canvas that is posted along with assignment #5. This contains data on the percentage of students in a school district being proficient on the math portion of the MEAP test along with some additional variables regarding the school district. The MEAP test is a test given to all high school students to see if they are proficient in various subjects. Suppose you suspect that math proficiency is related to poverty in the district and how much is spent on the schools in the district. The variables of interest are:

- math10: the percent of students passing the math section of the MEAP test.
- lnchprg: the percentage of students receiving free or reduced school lunch, as lower income students receive free or reduced lunch.
- expend: spending per student in the school district
- staff: number of staff per 1,000 students
- salary: average teacher salary in the school district
- benefits: average teacher benefits in the school district
- droprate: the dropout rate in the school district
- gradrate: the graduation rate in the school district

Suppose you think that math10 = f(lnchprg, expend, salary, benefits, staff droprate, gradrate)

a. Estimate the multiple regression in Stata for this. Are there signs of multicollinearity? Why or why not? Refer to Lecture #15 for the signs regarding multicollinearity. **(3 points)**

b. Calculate the variance inflation factors to see if multicollinearity is present. You can do this by typing `estat vif` after you estimate the regression. Which variables have a problematic variance inflation factor? Refer to Lecture #10 on how to interpret the variance inflation factor. **(3 points)**

c. Suppose you think a fix for the multicollinearity is to drop the staff and salary variables and then re-estimate the model. Does this fix the multicollinearity? **(2 points)**

d. Which variables are statistically significant for explaining the percentage of students who are proficient in math? Use a one or two-tailed test as you deem appropriate. **(3 points)**

e. Suppose you have a school district that has 25% of students receiving free or reduced lunch, spends $5,550 per student, pays teachers a benefit package worth $6,500 dollars, has a 4% dropout rate and a 90% graduation rate. What percentage of students will be proficient in math? **(2 points)**

2. Load in the data set entitled "smoke.dta" which is posted along with assignment #5. This is a data set containing the number of cigarettes a smoker smokes per day, the price of a pack of cigarettes in a state, along with some other variables. Suppose you think the number of cigarettes a smoker smokes a function of the smoker's education, age, square of age, the income of the smoker, and whether restaurants in the state restrict smoking. The variables of interest are:

- cigs: the number of cigarettes smoked per day
- educ: the education level of the smoker
- age: the age of the smoker
- agesq: the square of the age of the smoker
- lincome: the natural logarithm of the smoker's income
- restaurant: dummy variable = 1 if restaurants restrict smoking in that state.

You think that: cigs = f(educ, age, agesq, lincome, restaurant)

a. Estimate this regression in Stata. If you get an additional year of education, how many more (or less) cigarettes do you smoke per day as a result? **(3 points).**

b. Which $\hat{\beta}$ are statistically different from zero? Compare the test-statistic for each $\hat{\beta}$ to the two-tailed critical value for a 5% level of significance. **(3 points)**

c. Test for heteroscedasticity using the Breusch-Pagan test using the `hettest, rhs iid` test we described in Lecture #17. Is there evidence of heteroscedasticity at a 5% level of significance? At a 10% level of significance? How do you know? **(3 points).**

d. Estimate the same regression as in part a., but have Stata calculate the heteroscedasticity-robust standard errors for you by adding `,robust` like we talked about in Lecture #16. Does anything change in terms of statistical significance? Compare these results to your results in part b. **(3 points)**

e. Suppose you think the heteroscedasticity is a function of the smoker's age, since there is such a large variance in the age of smokers in the data set. Estimate a weighted least squares regression by adding `[aweight = age]` to the end of the regression command like we talked about in Lecture #18. Do NOT include `,robust` Does anything change in terms of statistical significance? Compare your results to those in d. and in a. **(3 points)**

f. We have estimated three different regressions to explain the number of cigarettes a smoker smokes per day: a regular regression, a regression with robust standard errors, and weighted least squares. Are there some common results across the three that suggest what determines the number of cigarettes a smoker smokes per day? **(2 points)**