Problem Set 3 (Multivariate regression 1)

ECON 441, 2019 Spring

Please solve all problems below, and do not forget to submit the Stata outputs somehow (copy here, print it separately, or upload it as an attachment to the submission on Canvas). This is NOT a group exercise, so everyone needs to write up their own solutions. WARNING: the exercises are modified slightly to better fit your needs!

1 Wooldridge, Chapter 3, Exercise 1 (20 pts)

Using the data in GPA2 on 4,137 college students, the following equation was estimated by OLS:

$$\widehat{colgpa} = 1.340 - 0.0205\ hsperc + 0.00145\ sat.$$

In addition, we got that $n = 4,137$ and $R^2 = 0.294$. Here *colgpa* is measured on a four-point scale, *hsperc* is the percentile in the high school graduating class (defined so that, for example, $hsperc = 5$ means the top 5% of the class), and *sat* is the combined math and verbal score of the student achievement test.

1. Why does it make sense for the coefficient on *hsperc* to be negative?

2. What is the predicted college GPA, if $hsperc = 30$ and $sat = 1,020$ (make sure to use notation from above)?

3. Suppose that two high school graduates, A and B graduated in the same percentile from high school, but student A's SAT was 100 points higher (about one standard deviation in the sample). What is the predicted difference in college GPA for these two students? Is the difference large?

4. Holding *hsperc* fixed, what difference in SAT score leads to a predicted college GPA difference of 0.6 (one-half grade point)? Comment on your answer (large small, positive, negative the SAT difference).

2 Wooldridge, Chapter 3, Exercise 4 (20 pts)

The median starting salary (*salary*, hourly, dollars) for new law school graduates[1] is determined by the following model

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + U,$$

where LSAT is the median LSAT score for the graduating class, GPA is the median college GPA for the class, *libvol* is the number of volumes in the law school library, *cost* is the annual cost of attending the school and *rank* is a law school ranking (with $rank = 1$ being the best school).

1. Explain why we expect $\beta_5 \leq 0$?

2. What signs do you expect for the other slope parameters? Justify your answers.

3. Write down ALL the assumptions of the corresponding linear model (you should repeat the equation above). Do you think exogeneity holds? (Justify your answer by saying what intuition you are using.)

[1] You have many law school classes, and the median salary for each class.

4. What estimator would you use to estimate this model?

5. Using your data, the estimated equation is

   $$\widehat{\log(salary)} = 7.56 + 0.0049 LSAT + 0.251 GPA + 0.092 \log(libvol) + 0.035 \log(cost) - 0.0032 rank$$

   with $n = 136$ and $R^2 = 0.842$. Depending on you answer in the previous point, interpret the coefficient on GPA, $\log(cost)$, $\log(libvol)$ and $rank$.

6. Would you say it is better to attend a higher ranked law school? You have two schools: law school A is ranked 3 places higher than law school B, but law school A costs 10% more than law school B for 5 years. Given your results, how much is the median salary difference between the two schools?

3 Wooldridge, Chapter 3, Exercise C8 (30 pts)

Use the DISCRIM data set to answer this question. These are ZIP-code level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast food restaurants charge higher prices in areas with higher concentration of African-American inhabitants.[2]

1. Find the average values of *prpblck* and *income*, along with their standard deviations. What are these variables (read it from Stata, including units of measurements)? Do you think they are correlated? To check this in Stata, either run the regression of *income* on *prpblck*, or calculate the correlation with the 'correlate varname1 varname2' command. Give the result (sign, strength).

2. Estimate the following simple linear model:

$$psoda = \beta_0 + \beta_1 prpblck + U$$

   Interpret the slope coefficient. (Think about exogeneity first.)

3. Consider the model on the price of soda in fast-food restaurants (*psoda*) in terms of the proportion of African-Americans and low income.

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + U.$$

   Estimate this model by OLS, and report your results here as well (equation, sample size, measure of fit). Interpret the coefficient on *prpblck* and *income*. Do you think they are large, economically speaking? What happened with the discrimination effect? Why?

[2] I am going to use the variable names of the book in the following exercise.

4. Your friend says that the model with constant price-elasticity would be more appropriate than the linear model. You agree. Write down the new model (just the equation suffices). (Note that *prpblck* is already a percentage, so it should enter as a RHS variable without any modification.)

5. Estimate the new model, report the results. Interpret the coefficient on *prpblck* now - you can assume exogeneity if you would like to.

6. Now add the *prppov* variable to the regression. What happens to $\hat{\beta}_{prpblck}$?

4 Wooldridge, Chapter 3, C10 (30 pts)

Use data in HTV to answer this question. The data set includes information on wages, education, parents' education and several other variables for 1,230 working men in 1991.

1. What is the range (=maximum and minimum) of the *educ* variable in the sample. (Also write down the definition/unit of the variable, please.) Do the men or their parents' have on average higher education?

2. Optional: Use the 'tabulate variablename' command to see what percentage of men completed twelfth grade, but no higher grade.

3. Estimate the regression model

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + U.$$

Report your results. How much sample variation of education explained by the parents' education? (English sentence, please.) Interpret the coefficient on *motheduc*. (Comment on exogeneity briefly.)

4. There is a measure of cognitive ability in the data set, the variable *abil*. Add this variable to the regression, report your estimation results. How does the measure of fit change after adding *abil* to the regression?

5. Optional: Check if ability is indeed a factor that would endogeneity in the regression model at part 3. (Hint: use correlate command). Also, use the 'scatter varname1 varname2' command to check the scatter plot of *educ* and *abil*. Is the relationship linear?

6. You decided that the relationship between *educ* and *abil* may not be linear. Your intuition is that as we increase the ability measure, the expected level of education increases more and more ('increasing returns'). What can you do to fix this modeling problem, given your intuition? Write down the new model (the equation is enough). Run the regression, report your results.

7. (You decided to include a quadratic term for *abil*, I will call it *abilsq*.) Interpret the sign for the coefficient on *abilsq*. It is what you expected? Do you prefer the model with the quadratic terms? Why?

8. Give the unitary effect of the ability score: What is the average change of education level if the ability measure is increased from 1.8 to 2.8? What if it increased from -1.2 to -0.2?

9. OPTIONAL: What is the effect of increasing it by 2 points (roughly 1 standard deviation) from 2 to 4? What is the marginal effect at the average? What is the average marginal effect? Do they coincide?