Problem Set 6

EC 303: Empirical Economic Analysis

Due December 11, 2019 at 3:30pm.

1 Theoretical Problems

Problem 1.1: Regression by Hand.** Complete Exercise 12.13 from the textbook. This problem walks you through simple calculation of important regression aspects. You should know how to calculate all of these (plus the others covered) for exams.

- There is a typo in the book—the first line stating $\sum y_i = 346$ is correct, while the last line should say $\sum_i y_i^2 = 17,454$.

Problem 1.2: Properties of Simple Linear Regression. This problem points out important features of the OLS regression method. While these are introduced for the simple model, many of these properties hold for multiple regression.

a. Prove that the point $(\overline{x}, \overline{y})$ is on the line of a simple linear regression $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

b. Complete Exercise 12.65 from the textbook—that is, show that the test statistic for the hypothesis test $\mathcal{H}_0 : \beta_1 = 0$ for the simple OLS regression is equivalent to the test statistic for the hypothesis test of $\mathcal{H}_0 : \rho = 0$. What does this tell you?

   – *Hints:* Algebra will be easier if you use the facts that SST= S_{YY} and $r^2 = \frac{\text{SST}-\text{SSE}}{\text{SST}}$. Also don't forget that $S = \sqrt{\text{SSE}/n-2}$.

Problem 1.3: Confidence Intervals and Inference for OLS.** These two textbook problems use the same data set to build on each other.

i. First, complete Exercise 12.16 parts (c) through (e)

ii. Now, complete Exercise 12.33.

Problem 1.4: Leaving Out Regressors. This problem walks you through what happens if your regression model is missing an important variable, a problem called **omitted variable bias.** You're sure to see this again in one of your econometrics classes.

a. Suppose that we have a true regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \tag{1}$$

under which $\mathbb{E}[\epsilon | X_1, X_2] = 0$. Suppose, however, that we only estimate the model:

$$y = \beta_0 + \beta_1 x_1 + u, \tag{2}$$

where $u = \beta_2 x_2 + \epsilon$. Suppose that X_1 and X_2 have a correlation ρ. What regression assumption are we violating here?

b. Use the formula:

$$\hat{\beta}_1 = \frac{\sum (x_{1i} - \overline{x}_1)(y_i - \overline{y})}{\sum (x_{1i} - \overline{x}_1)(x_{1i} - \overline{x}_1)}$$

to show that $\hat{\beta}_1$ is no longer an unbiased estimator.

c. Under what conditions (for ρ) could $\hat{\beta}_1$ be unbiased, even if the variable is omitted from the regression? How does the size of the bias depend on ρ? Based on these answers, how is the omitted variable causing the observed bias?

d. Suppose that X_1 and X_2 are negatively correlated. What direction is the bias? Does this mean that your estimated regression *understates* or *overstates* the true association between X_1 and y?

Problem 1.5: Multiple Regression Interpretation.** Using the fringe data set, I estimated the following regression of 616 (note the sample size) people's annual earnings (in dollars) on their demographics and job characteristics. Standard errors are reported in square brackets beneath each coefficient:

$$\begin{aligned}\text{ann\_earn} = &-5040 + 90.1(\text{age}) + 512(\text{married}) + 426(\text{yrs\_educ}) + 3,980(\text{male}) + 1,270(\text{white}) \\ &\quad [2160] \quad [26] \quad [724] \quad [133] \quad [738] \quad [1060] \\ &- 804(\text{union}) + 760(\text{office}) + 2.25(\#\text{ vacation days}) + 8.38(\text{offers sick leave}) + 4.75(\text{offers health insurance}) \\ &\quad [766] \quad [800] \quad [1.74] \quad [3.37] \quad [1.31]\end{aligned}$$

Some variables (male, white, union, office sick leave, and health insurance) are called *dummy variables*—they take on a value of 1 if the indicated condition is met (if the earner is a man, if they are white, if they are part of a union, etc.) and 0 otherwise.

a. Interpret the coefficient on age. What might explain this feature?

b. How does the regression suggest earnings differ between (1) a 35-year old married minority man with 16 years of education and (2) a 55-year old single white woman with 18 years of education working in the same job at the same company?

c. One shortcut way to tell if a regression coefficient is significant at the 95% confidence level is if its estimated value is more than twice its standard error. Based on this shortcut, which variables are significantly associated with differences in earnings?

d. Based on the snippet of Stata output below, calculate and interpret the R^2 for this regression.

Source	SS	df	MS
Model	1.5878e+10	10	1.5878e+09
Residual	3.4081e+10	605	56331597.6
Total	4.9959e+10	615	81234289.6

e. Perform an F-test for this regression (with $\alpha = 0.01$). What do the results tell you?

2 Extra Credit Stata: Research Project

This exercise is extra credit, and is worth up to 5 points on your final exam (plus additional bonus points in parts (c) and (d)). Please send all solutions to this problem **directly to me** at alcobe@bu.edu. Note that the additional points available in parts (c) and (d) (the data viz contests) have an **earlier due date**—please plan accordingly.

Problem 2.1: Mini Research Project. To wrap up everything you've been exposed to in Stata this semester, your final extra credit opportunity is to perform a miniature research project. I've uploaded a data set to Blackboard called census.dta, that contains data from the U.S. census data between 2015 and 2017. It has many interesting variables that you can explore (unfortunately, only for the U.S. population, although there are interesting migration variables some of you might be interested in). The documentation for this data set (which includes a list of all variables and their descriptions) has also been uploaded to Blackboard.

a. **Identify a question.** Identify a research question for this project. This should involve the relationship between at least two variables, but could also incorporate variation over time, region, race, etc. Be as creative as you can! Describe your choice and why you think this is an interesting area of focus.

b. **Conduct a mini literature review.** Identify between 3 and 5 sources that address your chosen topic. These don't need to be scholarly—newspaper articles are fine. Write a 1-2 paragraph synthesizing these sources (it's okay if they've answered your question already!)

   – *Note:* The data set I've posted is rather large because it contains a lot of interesting variables. Once you've identified your question, your analysis will probably go faster if you drop variables that aren't related to your study.

c. **Bad data viz competition.** Make me a terrible data visualization that highlights some (real) aspect of the relationship between your variables. This could be any type of visualization (pie chart, scatter plot, map, etc.), but should be flawed in major ways—for example, it could have terrible design, be hard to read/interpret, or highlight some feature of the data that no one should doubt or care about. I will show these submissions in class, and the best bad visualization will receive **an extra 3 points** on the final exam.

d. **Good data viz competition.** Now repeat exercise (c), but with a better data visualization. This viz should have good design, use color appropriately, and highlight something really interesting about the relationship between your variables. Again, I will show these submissions in class, and the best visualization will receive **an extra 3 points** on the final exam.

   – **Note:** Submissions for parts (c) and (d) are due at 12:00PM on Wednesday, 12/11/2019 (so that I have time to organize them and pick winners).

e. **Summary Statistics**. Provide a table of summary statistics for your variables. That is, tell me any interesting things you think I should know about your data. This includes things like your sample size, the average value and standard deviation of your variables, and the correlation between them. It could also include things like the average value of a variable among different populations (would it be interesting for your story if the average value of one of your variables were different between low-income and high-income households? Men and women?) Try to think of interesting ways to subset and look at the data.

f. **Regression Analysis**. Provide two regressions—one that relates your two chosen variables directly, and one that includes additional controls (such as region, gender, household income, year, etc.). Interpret your regressions. Are your results significant (and if so, at what confidence?)

g. **Bootstrapping Confidence Intervals.** Adapt a bootstrapping procedure from earlier in the semester to construct a 95% confidence interval for your regression coefficient of interest in the model with many controls. I'm happy to help you adapt this code if you'd like. How do your results differ from your results in (f)?

h. **Interpretation.** Write a 1-2 paragraph interpretation of what you've found throughout this project. Can you interpret your results in a causal framework—that is, is one of your variables causing the change in the other one? If not, how might you make this causal? If you were to continue this project in a later course, what areas of focus have you identified as potentially interesting avenues for further research?

3 References

- **Census data:** Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas & Matthew Sobek. IPUMS USA: Version 9.0. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D010.V9.0.