Problem Set 3

ECON 480 — Fall 2021

Due by Thursday October 7

Theory and Concepts

1. In your own words, describe what exogeneity and endogeneity mean, and how they are related to bias in our regression. What things can we learn about the bias if we know X is endogenous?

2. In your own words, describe what R^2 means. How do we calculate it, what does it tell us, and how do we interpret it?

3. In your own words, describe what the standard error of the regression (SER) means. How do we calculate it, what does it tell us, and how do we interpret it?

4. In your own words, describe what homoskedasticity and heteroskedasticity mean: both in ordinary English, and in terms of the graph of the OLS regression line.

5. In your own words, describe what the variation in $\hat{\beta_1}$ (either variance or standard error) means, or is measuring. What three things determine the variation, and in what way?

6. In your own words, describe what a p-value means, and how it is used to establish statistical significance.

7. A researcher is interested in examining the impact of illegal music downloads on commercial music sales. The author collects data on commercial sales of the top 500 singles from 2017 (Y) and the number of downloads from a web site that allows 'file sharing' (X). The author estimates the following model

$$\text{music sales}_i = \beta_0 + \beta_1 \text{illegal downloads}_i + u_i$$

The author finds a large, positive, and statistically significant estimate of $\widehat{\beta}_1$. The author concludes these results demonstrate that illegal downloads actually *boost* music sales. Is this an unbiased estimate of the impact of illegal music on sales? Why or why not? Do you expect the estimate to overstate or understate the true relationship between illegal downloads and sales?

8. A pharmaceutical company is interested in estimating the impact of a new drug on cholesterol levels. They enroll 200 people in a clinical trial. People are randomly assigned the treatment group or into the control group. Half of the people are given the new drug and half the people are given a sugar pill with no active ingredient. To examine the impact of dosage on reductions in cholesterol levels, the authors of the study regress the following model:

$$\text{cholesterol level}_i = \beta_0 + \beta_1 \text{dosage level}_i + u_i$$

For people in the control group, dosage $\text{level}_i = 0$ and for people in the treatment group, dosage level_i measures milligrams of the active ingredient. In this case, the authors find a large, negative, statistically significant estimate of $\widehat{\beta}_1$. Is this an unbiased estimate of the impact of dosage on change in cholesterol level? Why or why not? Do you expect the estimate to overstate or understate the true relationship between dosage and cholesterol level?

Theory Problems

For the following questions, please *show all work* and explain answers as necessary. You may lose points if you only write the correct answer. You may use `R` to *verify* your answers, but you are expected to reach the answers in this section "manually."

9. A researcher wants to estimate the relationship between average weekly earnings (AWE, measured in dollars) and Age (measured in years) using a simple OLS model. Using a random sample of college-educated full-time workers aged 25-65 yields the following:

$$\widehat{AWE} = 696.70 + 9.60\,Age$$

a. Interpret what $\hat{\beta_0}$ means in this context.

b. Interpret what $\hat{\beta_1}$ means in this context.

c. The $R^2 = 0.023$ for this regression. What are the units of the R^2, and what does this mean?

d. The SER, $\hat{\sigma_u} = 624.1$ for this regression. What are the units of the SER in this context, and what does it mean? Is the SER large in the context of this regression?

e. Suppose Maria is 20 years old. What is her predicted $\widehat{AWE}$?

f. Suppose the data shows her *actual* AWE is $430. What is her residual? Is this a relatively good or a bad prediction?[1]

g. What does the error term, $\hat{u}_i$ represent in this case? What might individuals have different values of u_i?

h. Do you think that *Age* is exogenous? Why or why not? Would we expect $\hat{\beta}_1$ to be too *large* or too *small*?

[1] Hint: compare your answer here to your answer in Part D.

10. Suppose a researcher is interested in estimating a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

In a sample of 48 observations, she generates the following descriptive statistics:

- $\bar{X} = 30$
- $\bar{Y} = 63$
- $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 6900$
- $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 29000$
- $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = 13800$
- $\sum_{i=1}^{n}\hat{u}^2 = 1656$

a. What is the OLS estimate of $\hat{\beta}_1$?

b. What is the OLS estimate of $\hat{\beta}_0$?

c. Suppose the OLS estimate of $\hat{\beta}_1$ has a standard error of 0.072. Could we probably reject a null hypothesis of $H_0 : \beta_1 = 0$ at the 95% level?

d. Calculate the R^2 for this model. How much variation in Y is explained by our model?

e. How large is the average residual?

R Questions

Answer the following questions using R. When necessary, please write answers in the same document (knitted Rmd to html or pdf, typed .doc(x), or handwritten) as your answers to the above questions. Be sure to include (email or print an .R file, or show in your knitted markdown) your code and the outputs of your code with the rest of your answers.

11. Download the MLBattend dataset. This data contains data on attendance at major league baseball games for all 32 MLB teams from the 1970s-2000. We want to answer the following question:

    "How big is home-field advantage in baseball? Does a team with higher attendance at home games over their season have score more runs over their season?"

a. Clean up the data a bit by mutate()-ing a variable to measure home attendance in millions. This will make it easier to interpret your regression later on.
b. Get the correlation between Runs Scored and Home Attendance.
c. Plot a scatterplot of Runs Scored (y) on Home Attendance (x). Add a regression line.
d. We want to estimate a regression of Runs Scored on Home Attendance:

$$\widehat{\text{runs scored}}_i = \beta_0 + \beta_1 \text{home attendance}_i$$

Run this regression in R. What are $\hat{\beta}_0$ and $\hat{\beta}_1$ for this model? Interpret them in the context of our question. [Hint: make sure to save your regression model as an object, and get a summary() of it. This object will be needed later.]

e. Write out the estimated regression equation.
f. Make a regression table of the output (using the huxtable package).
g. Check the goodness of fit statistics. What is the R^2 and the SER of this model? Interpret them both in the context of our question.
h. Now let's start running some diagnostics of the regression. Make a histogram of the residuals. Do they look roughly normal? [Hint: you will need to use the broom package's augment() command on your saved regression object to add containing the residuals (.resid), and save this as a new object - to be your data source for the plot in this question and the next question.]
i. Make a residual plot.
j. Test the regression for heteroskedasticity. Are the errors homoskedastic or heteroskedastic? [Hint: use the lmtest package's bptest() command on your saved regression object.] Run another regression using robust standard errors. [Hint: use the estimatr package's lm_robust() command and save the output like the following:

```
reg_robust <-lm_robust(y ~ x, data = the_data, # change y, x, and data names to yours
                              se_type = "stata") # we'll use this method to calculate
```

Now make another regression output table with huxtable, with one column using regular standard errors (just use your original saved regression object) and another using robust standard errors (use this new saved object).

k. Test the data for outliers. If there are any, identify which team(s) and season(s) are outliers. [Hint: use the car package's outlierTest() command on your saved regression object.]
l. Look back at your regression results. What is the marginal effect of home attendance on runs scored? Is this statistically significant? Why or why not?
m. Now we'll try out the infer package to understand the p-value for our observed slope in our regression model. First, save the (value of) our sample $\hat{\beta}_1$ from your regression in Part D as an object, I suggest:

```
our_slope = 123 # replace "123" with whatever number you found for the slope in part D
```

Then, install and load the infer package, and then run the following simulation:

```
# save our simulations as an object (I called it "sims")
sims <- data %>% # "data" here is whatever you named your dataframe!
  specify(y ~ x) %>% # replacing y and x with your variable names
  hypothesize(null = "independence") %>% # H_0 is that slope is 0, x and y are independent
  generate(reps = 1000,
           type = "permute") %>% # make 1000 samples assuming H_0 is true
  calculate(stat = "slope") # estimate slope in each sample

# look at it
sims

# calculate p value
sims %>%
  get_p_value(obs_stat = our_slope,
              direction = "both") # a two-sided H_a: slope =/= 0
```

Compare to the p-value in your original regression output in previous parts of this question.

n. Make a histogram of the simulated slopes, and plot our sample slope on that histogram, shading the p-value. [You can pipe `sims` into `visualize(obs_stat = our_slope)`, or use `ggplot2` to plot a histogram in the normal way, using `sims` as the data source and add a `geom_vline(xintercept = our_slope)` to show our finding on the distribution.]