EC 320: Introduction to Econometrics

Problem Set 3

Total: 50 points
Due: Tuesday January 11th, 2020 at 1 pm

Learning Outcomes:

- Understanding regression analysis
- Understanding how to do basic proofs
- Understanding omitted variable bias

Checklist Before Handing In:

- Did you answer all questions?
- Did you answer all parts for each question?
- Were your answers too vague? If so, make them more precise to make sure they really answer the question being asked.

Instructions: You are encouraged to work with other students in the class, but you must provide original responses. To receive full credit, justify your answers and list your collaborators. For full credit on the computational exercises, include your code and output in addition to your answers. You will turn in digital copies of your responses on Canvas. Please note the list of acceptable file types on the submission page.

Name:

Collaborator 1:

Collaborator 2:

Collaborator 3:

Analytical Questions

1. Consider the data on hours studied per week and final grades in EC 320 from a sample of four students: (2 points each)

Student	Hours Studied	Final Grade
Hinata	7	72%
Michelle	10	97%
Jerry	4	43%
Jorge	6	84%

   (a) Suppose that you want to use the data to learn about the effect of studying on grades. Write down a simple linear regression model tailored to your objective.

   (b) Calculate the parameter estimates for the intercept ($\hat{\beta}_1$) and slope ($\hat{\beta}_2$) using the OLS formulas from class.

   (c) Interpret $\hat{\beta}_1$. What does this parameter estimate tell us?

   (d) Interpret $\hat{\beta}_2$. What does this parameter estimate tell us?

   (e) Calculate the coefficient of determination (R^2) for the regression you estimated. What does it tell us about the relationship between studying and grades?

   (f) What is the predicted final grade for a student who studies 5 hours per week?

   (g) Based on the regression you estimated, how many hours would a student have to study to expect a score of 100%?

2. Consider the predicted values of Y_i from a simple linear regression of Y_i on X_i, which we refer to as $\hat{Y}_i$. Prove that the sample mean of $\hat{Y}_i$ is equal to the sample mean of Y_i (i.e. $\bar{Y}$). (3 points)

3. Consider the residuals ($\hat{u}_i$) from a simple linear regression. Prove that the sample mean of $\hat{u}_i$ is equal to zero. (3 points)

4. Assume that $\bar{Y}$ and $\bar{X}$ are equal to zero. Prove that the sample covariance of X_i and $\hat{u}_i$ is equal to zero. (3 points)

5. Suppose that you run a regression of Y_i on X_i and obtain parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. Then, using the same data, you decide to run a regression of $\tilde{Y}_i$ on X_i, where $\tilde{Y}_i = 2Y_i$. Prove that both regressions have the same R^2. (3 points)

Computational Questions

For this portion of the problem set, you will use the `fertility.csv` excel file in the Problem Set 3 folder on Canvas. The file contains a random sample of data on the fertility and education of women interviewed in the Demographic and Health Survey in India. To complete this assignment, you will need to load the `tidyverse` and `stargazer` package.

The Demographic and Health Survey interviews women and asks them information about their fertility and health behavior. The variables in the data set are described below.

Variable Name	Description
`id`	Unique id of each woman interviewed
`educ`	Years of schooling completed by the woman (respondent)
`peduc`	Years of schooling completed by the partner
`kids`	Number of children
`agem`	Age of marriage for the woman

Economic theory suggests that education can decrease fertility. One hypothesized mechanism is that education improves job prospects, which increases the opportunity cost of having children. Another hypothesized mechanism—particularly relevant in developing countries—is that education increases awareness of effective contraceptive methods. We will delve into the effect of parental education on fertility behavior in India. (2 points each)

1. Does marrying later in life reduce fertility, as measured by the number of children?

   (a) Write down a simple linear regression model that describes the effect of age at marriage on fertility. Use variable names that relate to the research question.

   (b) Using the `lm()` function, estimate the simple linear regression model that you specified above.

   (c) Make a regression table using `stargazer`. Make sure you specify `type = "html"`. To reduce clutter, pass `keep.stat = c("rsq", "n")` as an argument to `stargazer()`.

   (d) Visualize your regression results by making a scatter plot with a fitted regression line. **Hint:** Use `stat_smooth(method = "lm", se = FALSE)` in your `ggplot` code.

   (e) Interpret the intercept estimate. Is this reasonable?

   (f) Interpret the slope estimate.

   (g) Is the slope coefficient statistically significant?

   (h) Does the slope estimate warrant causal interpretation? In other words, do you think that the slope describes the causal effect of education on fertility? Why or why not? If not, identify a potential source of selection bias or omitted-variable bias.

2. Does education reduce fertility, as measured by the number of children?

   (a) Write down two simple linear regression models that describe the effect of parent's education on fertility. Use variable names that relate to the research question.

   (b) Run two separate simple linear regressions that allow you to estimate the effect of mother's education and father's education on fertility. Use one independent variable per regression. Summarize the results in a regression table.

   (c) Identify and interpret the R^2 from the regression of fertility on father's education.

   (d) Of the three independent variables you considered, which one is the best predictor of fertility? Justify your answer.