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.

$(\text{Final Grade})_i = \beta_1 + \beta_2(\text{Hours Studied})_i + u_i$.

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

Using the formulas, you should obtain $\hat{\beta}_1 \approx 19.28$ and $\hat{\beta}_2 \approx 8.11$. For full credit, you must have showed your work.

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

$\hat{\beta}_1$ tells us the predicted grade for a student who didn't study. Someone who does not study can expect to receive a final grade of 19.28%.

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

$\hat{\beta}_2$ tells us the grade boost from an additional hour of studying per week. An additional hour of studying per week increases a student's final grade by 8.11 percentage points, on average.

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

Using the formula, $R^2 \approx 0.77$, which implies that studying explains 77% of the variation in grades. For full credit, you must have showed your work.

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

The predicted final grade for someone who studies 5 hours per week is $19.28 + 8.11(5) = 59.83$.

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

The idea is to use the fitted relationship $(\text{Final }\hat{\text{Grade}})_i = 19.28 + 8.11(\text{Hours Studied})_i$ and then solve for $(\text{Hours Studied})_i$.

First, plug 100 into $(\text{Final }\hat{\text{Grade}})_i$: $100 = 19.28 + 8.11(\text{Hours Studied})_i$.

Next, subtract the intercept estimate from both sides: $80.72 = 8.11(\text{Hours Studied})_i$.

Finally, divide both sides by the slope estimate: $(\text{Hours Studied})_i \approx 9.95$.

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)

Start with the fitted regression line then replace $\hat{\beta}_1$ with the OLS intercept formula:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \\ &= \bar{Y} - \hat{\beta}_2 \bar{X} + \hat{\beta}_2 X_i \\ &= \bar{Y} + \hat{\beta}_2 (X_i - \bar{X}).\end{aligned}$$

Next, sum over i and divide by n:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n}\sum_{i=1}^{n} \bar{Y} + \frac{1}{n}\hat{\beta}_2 \sum_{i=1}^{n}(X_i - \bar{X}).$$

Notice that $\frac{1}{n}\sum_{i=1}^{n} \bar{Y} = \bar{Y}$. Then

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \bar{Y} + \frac{1}{n}\hat{\beta}_2 \sum_{i=1}^{n}(X_i - \bar{X}).$$

In class, we showed that $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$. Hence,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i &= \bar{Y} + \frac{1}{n}\hat{\beta}_2(0) \\ &= \bar{Y}.\end{aligned}$$

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)

A residual is defined as

$$\hat{u}_i = Y_i - \hat{Y}_i.$$

Define the sample mean of the residuals as

$$\bar{\hat{u}}_i = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i.$$

Plug in the definition of the residual to obtain

$$\begin{aligned}\bar{\hat{u}}_i &= \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i \\ &= \bar{Y} - \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i.\end{aligned}$$

In the previous exercise, you proved that the sample mean of $\hat{Y}_i$ is equal to the sample mean of Y_i. Thus,

$$\begin{aligned}\bar{\hat{u}}_i &= \bar{Y} - \bar{Y} \\ &= 0.\end{aligned}$$

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)

If $\bar{X} = 0$ and $\bar{Y} = 0$, then

$$\begin{aligned}\hat{\beta}_1 &= \bar{Y} - \hat{\beta}_2\bar{X} \\ &= (0) - \hat{\beta}_2(0) \\ &= 0\end{aligned}$$

and

$$\begin{aligned}
\hat{\beta}_2 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^{n}(X_i - [0])(Y_i - [0])}{\sum_{i=1}^{n}(X_i - [0])^2} \\
&= \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\hat{u}_i &= Y_i - \hat{Y}_i \\
&= Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i) \\
&= Y_i - ([0] + \hat{\beta}_2 X_i) \\
&= Y_i - \hat{\beta}_2 X_i.
\end{aligned}$$

The sample covariance of X_i and $\hat{u}_i$ is given by

$$S_{(X_i,\hat{u}_i)} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(\hat{u}_i - \bar{\hat{u}}).$$

The problem states that $\bar{X} = 0$ and you proved that $\bar{\hat{u}} = 0$, so

$$S_{(X_i,\hat{u}_i)} = \frac{1}{n-1}\sum_{i=1}^{n} X_i \hat{u}_i.$$

The denominator $n-1$ is nonzero, so we will restrict our attention to $\sum_{i=1}^{n} X_i \hat{u}_i$. Showing that this term is equal to zero completes the proof.

Replace $\hat{u}_i$ with $Y_i - \hat{\beta}_2 X_i$ and simplify:

$$\begin{aligned}
\sum_{i=1}^{n} X_i \hat{u}_i &= \sum_{i=1}^{n} X_i [Y_i - \hat{\beta}_2 X_i] \\
&= \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_2 \sum_{i=1}^{n} X_i^2.
\end{aligned}$$

Replace $\hat{\beta}_2$ with the OLS slope formula:

$$\begin{aligned}
\sum_{i=1}^{n} X_i \hat{u}_i &= \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 \\
&= \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \sum_{i=1}^{n} X_i^2 \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i Y_i \\
&= 0.
\end{aligned}$$

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)

Using the transformed data, you estimate $\tilde{Y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 X_i + \tilde{u}_i$ to obtain the fitted model $\hat{\tilde{Y}}_i = \hat{\tilde{\beta}}_1 + \hat{\tilde{\beta}}_2 X_i$. From here, we will find an expression for $\hat{\tilde{Y}}_i$ in terms of $\hat{Y}_i$. We will then plug this into the formula for $\tilde{R}^2$ to complete the proof.

First, notice that

$$\begin{aligned}
\bar{\tilde{Y}} &= \frac{1}{n} \sum_{i=1}^{n} \tilde{Y}_i = \frac{1}{n} \sum_{i=1}^{n} 2Y_i \\
&= 2 \frac{1}{n} \sum_{i=1}^{n} Y_i \\
&= 2\bar{Y}.
\end{aligned}$$

This will prove useful in the next step.

Second, relate $\hat{\tilde{\beta}}_2$ to $\hat{\beta}_2$:

$$\begin{aligned}
\hat{\tilde{\beta}}_2 &= \frac{\sum_{i=1}^{n} (\tilde{Y}_i - \bar{\tilde{Y}})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (2Y_i - 2\bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= 2 \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= 2\hat{\beta}_2.
\end{aligned}$$

Third, relate $\hat{\tilde{\beta}}_1$ to $\hat{\beta}_1$:

$$\begin{aligned}\hat{\tilde{\beta}}_1 &= \bar{\tilde{Y}} - \hat{\tilde{\beta}}_2\bar{X} = 2\bar{Y} - 2\hat{\beta}_2\bar{X} \\ &= 2(\bar{Y} - \hat{\beta}_2\bar{X}) \\ &= 2\hat{\beta}_1.\end{aligned}$$

Next, use the expressions for $\hat{\tilde{\beta}}_1$ and $\hat{\tilde{\beta}}_2$ to express $\hat{\tilde{Y}}_i$ in terms of $\hat{Y}_i$.

$$\begin{aligned}\hat{\tilde{Y}}_i &= \hat{\tilde{\beta}}_1 + \hat{\tilde{\beta}}_2 X_i = 2\hat{\beta}_1 + 2\hat{\beta}_2 X_i \\ &= 2(\hat{\beta}_1 + \hat{\beta}_2 X_i) \\ &= 2\hat{Y}_i.\end{aligned}$$

Finally, use the expression for $\hat{\tilde{Y}}_i$ in the definition of $\tilde{R}^2$:

$$\begin{aligned}\tilde{R}^2 &= \frac{\sum_{i=1}^n(\hat{\tilde{Y}}_i - \bar{\tilde{Y}})^2}{\sum_{i=1}^n(\tilde{Y}_i - \bar{\tilde{Y}})^2} = \frac{\sum_{i=1}^n(2\hat{Y}_i - 2\bar{Y})^2}{\sum_{i=1}^n(2Y_i - 2\bar{Y})^2} \\ &= \frac{\sum_{i=1}^n[2(\hat{Y}_i - \bar{Y})]^2}{\sum_{i=1}^n[2(Y_i - \bar{Y})]^2} = \frac{4\sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2}{4\sum_{i=1}^n(Y_i - \bar{Y})^2} \\ &= \frac{\sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n(Y_i - \bar{Y})^2} \\ &= R^2.\end{aligned}$$

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. (2 points each)

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.

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.