EC 320: Introduction to Econometrics

Problem Set 4

Total: points
Due:

Learning Outcomes:

- Understanding hypothesis testing
- Understanding multiple regression models
- Understanding goodness of fit

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. Suppose that you run a regression of Y_i on X_i with 102 observations and obtain an estimate for the slope (*i.e.*, $\hat{\beta}_2$). Your estimate for the standard error of $\hat{\beta}_2$ is 1. You are considering two different hypothesis tests.

   The first is a **one-sided test**:

   $$\text{H:}_0 \;\; \beta_2 = 0, \quad \text{H:}_a \;\; \beta_2 > 0, \quad \alpha = 0.05.$$

   The second is a **two-sided test**:

   $$\text{H:}_0 \;\; \beta_2 = 0, \quad \text{H:}_a \;\; \beta_2 \neq 0, \quad \alpha = 0.05.$$

   (a) What values of $\hat{\beta}_2$ would lead you to reject the null hypothesis in the **one-sided** test?

   You reject H_0 if $t > t_{\text{crit}}$.

   Given $102 - 2$ degrees of freedom, you will find on your table that $t_{\text{crit}} = 1.66$.

   The t-statistic is $t = \dfrac{\hat{\beta}_2}{\text{SE}(\hat{\beta}_2)} = \dfrac{\hat{\beta}_2}{1}$.

   You will reject H_0 if $\hat{\beta}_2 > 1.66$.

   (b) What values of $\hat{\beta}_2$ would lead you to reject the null hypothesis in the **two-sided** test?

   You reject H_0 if $|t| > t_{\text{crit}}$.

   Given $102 - 2$ degrees of freedom, you will find on your table that $t_{\text{crit}} = 1.984$.

   The t-statistic is $t = \dfrac{\hat{\beta}_2}{\text{SE}(\hat{\beta}_2)} = \dfrac{\hat{\beta}_2}{1}$.

   You will reject H_0 if $\hat{\beta}_2 > 1.984$ or $\hat{\beta}_2 < -1.984$.

   (c) What values of $\hat{\beta}_2$ would lead you to reject the null hypothesis in the **one-sided** test, **but not** the **two-sided** test?

   You will reject H_0 in a one-sided test if $\hat{\beta}_2 > 1.66$.

   You will fail to reject H_0 in a two-sided test if $-1.984 \leq \hat{\beta}_2 \leq 1.984$.

   If $1.66 \leq \hat{\beta}_2 \leq 1.984$, then you will reject H_0 in the one-sided test, but not in the two-sided test.

(d) What values of $\hat{\beta}_2$ would lead you to reject the null hypothesis in the **two-sided** test, **but not** the **one-sided** test?

You will fail to reject H_0 in a one-sided test if $\hat{\beta}_2 \leq 1.66$.
You will reject H_0 in a two-sided test if $\hat{\beta}_2 > 1.984$ or $\hat{\beta}_2 < -1.984$.
If $\hat{\beta}_2 < -1.984$, then you will reject H_0 in the two-sided test, but not in the one-sided test.

2. Suppose that you are studying the effect of police officers on crime rates in several large American cities. When you estimate the model

$$\text{Crime}_i = \beta_1 + \beta_1 \text{Police}_i + u_i,$$

you obtain a positive slope estimate.

(a) What does your slope estimate imply about how the number of police officers affects crime?

The positive slope estimate ($\hat{\beta}_1$) suggests that crime increases as the number of police increase. In other words, "more police, more crime."

(b) Provide an example of an omitted variable that could explain why the slope estimate is positive.

The idea here is to think of a variable that—if omitted from a regression of crime on police—would bias the slope estimator upwards. There are many examples that you could give, but they need to produce positive omitted-variable bias. If the variable is called Omitted_i, then the bias from omitting Omitted_i is given by

$$\text{Bias} = \beta_{\text{Omitted}} \frac{\text{Cov}(\text{Police}_i, \text{Omitted}_i)}{\text{Var}(\text{Police}_i)}.$$

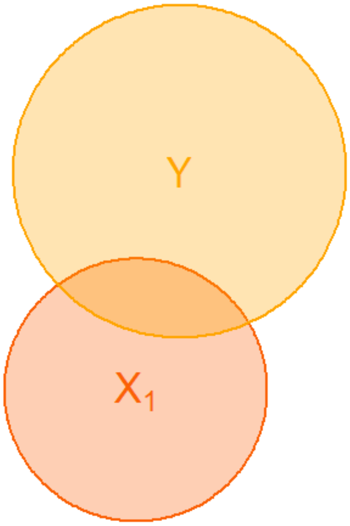
For Bias > 0, there are two types of omitted variables that you could describe: either 1) the omitted variable is positively correlated with crime (*i.e.*, $\beta_{\text{Omitted}} > 0$) and positively correlated with police or 2) it is negatively correlated with crime (*i.e.*, $\beta_{\text{Omitted}} < 0$) and negatively correlated with police. (If it is negatively correlated with one variable, but positively correlated with the other, then Bias < 0.)

One example of an omitted variable that could cause positive bias is the number of ex-convicts. For a variety of reasons, those who have been convicted of a crime are more likely to commit a new crime than those who have never been convicted, so cities with more ex-convicts are likely to exhibit higher crime rates. One might also expect officials in cities with more ex-convicts to recognize that fact and respond by hiring more police.

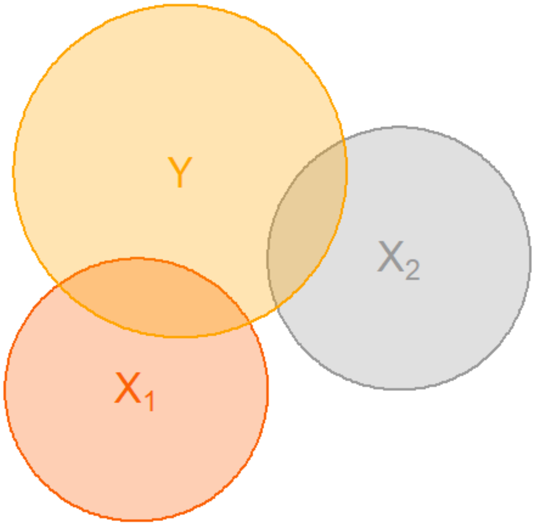
3. Suppose that your friend wrote a computer program that runs both simple and multiple linear regressions using OLS. Your friend asks you to test the software. After importing a dataset with 2000 observations and three cryptically named variables—Y, X_1, and X_2—you run two regressions. The first is a regression of Y on X_1, which gives you an intercept estimate of 10.4, a slope estimate of -3.8, and $R^2 = 0.215$. The second is a regression of Y on X_1 and X_2, which gives you an intercept estimate of 9.3, an X_1 slope estimate of -2.9, an X_2 slope estimate of -0.07, and $R^2 = 0.198$. You tell your friend that they must made a mistake somewhere in their code. Why?

The model with fewer variables had a higher R^2, which contradicts what we learned in class. If the model contains only one variable, then R^2 tells us the percentage of the variation in Y explained by X. If the model contains more than one variable, then R^2 tells us the percentage of the variation in Y explained by all of the variables in the model, collectively. All else being equal (*e.g.,* the dependent variable, sample size, *etc.*), adding explanatory variables to a regression generally increases R^2. In the edge case where an additional variable explains none of the variation in Y, then R^2 will not change. However, there is no case in which R^2 will decrease. You can illustrate this idea using Venn diagrams.

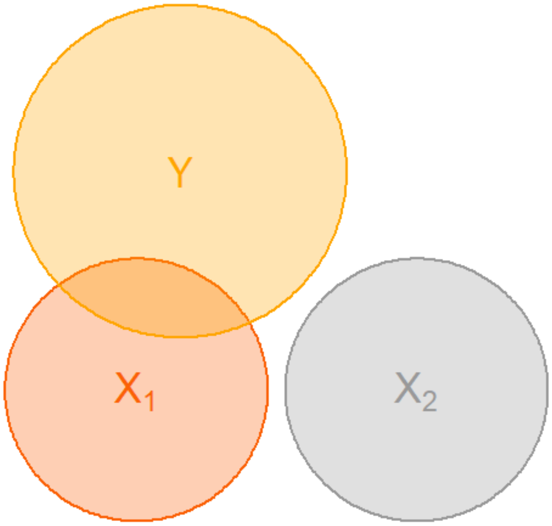
The orange circle describes the variation in Y, and the red circle describes the variation in X_1. In the simple linear regression, the overlap between the orange and red circles describes the variation in X_1 that explains Y.



Adding X_2 to the regression provides an additional region of overlap: the overlap between the orange and gray circles describes the variation in X_2 that explains Y. The important thing to notice is that—compared to the simple linear regression—more of the variation in Y is explained by the X variables. In other words, R^2 has increased.



However, if X_2 is irrelevant in the sense that it explains none of the variation in Y, then the multiple regression model will not explain any more of the variation in Y than the simple regression model. In this case R^2 does not change by adding X_2.



There is no way for me to add a variable that "un-explains" Y. I can't erase any of the existing overlap between X_1 and Y by adding X_2. In other words, R^2 cannot decrease.

4. A useful application of multiple regression analysis is *Hedonic modeling.* Hedonic models seek to explain the price of a good—such as a house—in terms of its attributes (*e.g.,* number of bedrooms, square footage, or distance from the nearest toxic waste dump). Consider the following Hedonic model of home sale prices:

$$\text{Price}_i = \beta_0 + \beta_1(\text{Square footage})_i + \beta_2\text{Bathrooms}_i + \beta_3\text{Bedrooms}_i + u_i.$$

Using data from 37 home sales, you estimate the model and obtain $\hat{\beta}_0 = 90000$, $\hat{\beta}_1 = 1100$, $\hat{\beta}_2 = 16000$, $\hat{\beta}_3 = 35000$, $\text{SE}(\hat{\beta}_1) = 650$.

(a) Interpret each coefficient.

$\hat{\beta}_0$: The average sale price of an empty lot (*i.e.*, a home with zero square footage, zero bedrooms, and zero bathrooms) is \$90,000.

$\hat{\beta}_1$: Holding the number of bedrooms and the number of bathrooms constant, each additional square foot of living space increases a home's sale price by \$1,100, on average.

$\hat{\beta}_2$: Holding the number of bedrooms and square footage constant, each additional bathroom increases a home's sale price by \$16,000, on average.

$\hat{\beta}_3$: Holding the number of bathrooms and square footage constant, each additional bedroom increases a home's sale price by \$35,000, on average.

(b) What is the model's forecasted sale price for a 2500-square-foot home with 3 bedrooms and 2.5 bathrooms?

The forecasted price is \$2,985,000:

$$\begin{aligned}\hat{\text{Price}}_i &= 90000 + 1100 \cdot (2500) + 16000 \cdot (2.5) + 35000 \cdot (3) \\ &= 2985000.\end{aligned}$$

(c) In a remodeling frenzy, a homeowner adds an additional bedroom and an additional bathroom by splitting up existing rooms. What is the forecasted change in the price of her home?

The forecasted change in price is \$51,000:

$$\begin{aligned}\Delta \hat{\text{Price}}_i &= \hat{\beta}_2 \cdot \Delta \text{Bathrooms}_i + \hat{\beta}_3 \cdot \Delta \text{Bedrooms}_i \\ &= 16000 \cdot (1) + 35000 \cdot (1) \\ &= 51000.\end{aligned}$$

Alternatively, you could have started with the house described in part (b) and added 1 bedroom and 1 bathroom. Then you could have taken the difference in forecasted prices between the remodeled home and the original home to obtain the expected change in sale price. This alternative approach yields the same forecasted change in price demonstrated above.

(d) A homeowner adds a 450-square-foot bedroom and a 75-square-foot bathroom by extending the footprint of his home into an area that used to be a driveway. What is the forecasted change in the price of his home?

The forecasted change in price is \$628,500:

$$\begin{aligned}\Delta \hat{\text{Price}}_i &= \hat{\beta}_1 \cdot \Delta (\text{Square footage})_i + \hat{\beta}_2 \cdot \Delta \text{Bathrooms}_i + \hat{\beta}_3 \cdot \Delta \text{Bedrooms}_i \\ &= 1100 \cdot (525) + 16000 \cdot (1) + 35000 \cdot (1) \\ &= 628500.\end{aligned}$$

Alternatively, you could have started with the house described in part (b) and added 525 square feet, 1 bedroom, and 1 bathroom. Then you could have taken the difference in forecasted prices between the remodeled home and the original home to obtain the expected change in sale price. This alternative approach yields the same forecasted change in price demonstrated above.

(e) Conduct two-sided tests of the hypothesis that square footage has no effect on sale price at the 10, 5, and 1 percent levels.

Here, I'm telling you to test the null hypothesis

$$\text{H}_0 : \beta_1 = 0$$

against the alternative hypothesis

$$\text{H}_a : \beta_1 \neq 0$$

at $\alpha = 0.1$, $\alpha = 0.05$, and $\alpha = 0.01$.

For each of these tests, you will use the same t statistic:

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_1^0}{\text{SE}(\hat{\beta}_1)} \\ &= \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} \\ &= \frac{1100}{650} \\ &\approx 1.69. \end{aligned}$$

You also need to find three different critical values of t (*i.e.*, t_{crit}). Each of these will have the same degrees of freedom:

$$\text{dof} = n - k - 1 = 37 - 3 - 1 = 33.$$

Your table does not have a row for 33 degrees of freedom, so "play it safe" by picking t_{crit} with 30 degrees of freedom. Doing this ensures that the Type I error rate of your tests does not exceed the chosen significance levels (the αs).

For $\alpha = 0.1$, $t_{\text{crit}} = 1.697$.

For $\alpha = 0.05$, $t_{\text{crit}} = 2.042$.

For $\alpha = 0.01$, $t_{\text{crit}} = 2.75$.

You will reject the null hypothesis if $|t| > t_{\text{crit}}$.

For the 10 percent test, $|t| = 1.69 > t_{\text{crit}} = 1.697$ is false. This implies that you fail to reject the null hypothesis at the 10 percent level.

The other two t_{crit}s are greater than 1.697, which implies that you will also fail to reject the null hypothesis at the 5 percent and 1 percent levels.

(f) Construct a 95 percent confidence interval for β_1.

The 95 percent confidence interval for β_1 is outlined by

$$\hat{\beta}_1 \pm t_{\text{crit}} \cdot \text{SE}(\hat{\beta}_1).$$

The appropriate t_{crit} is 2.042 (based on 30 degrees of freedom and $\alpha = 0.05$ for a two-sided test).

Then the confidence interval for β_1 is outlined by

$$1100 \pm 2.042 \cdot 650.$$

Then the 95 percent confidence interval for β_1 is

$$-227 < \beta_1 < 2,427.$$

Computational Problems

For this portion of the problem set, you will use the `apple.csv` file in the `Problem Set 4` folder on Canvas. The file contains data from an experimental survey. The survey presented participants with **randomly determined** prices for "eco-labeled" apples and regular apples and then asked how many eco-labeled and regular apples they would buy at those prices. For reference, eco-labeling helps consumers identify sustainably-produced (or "green") products and helps firms command higher prices for their products. You will estimate the demand for eco-labeled and regular apples by running regressions of apple quantity on prices. The fact that the prices were randomly assigned means that the exogeneity assumption holds—so long as both prices are included in the model. To complete this assignment, you will need to load the `tidyverse`, `stargazer`, and `broom`. Use `Problem_Set_04_template.Rmd` as a template.

Variable Name	**Description**
`reglbs`	Pounds of regular apples demanded
`ecolbs`	Pounds of eco-labeled apples demanded
`regprc`	Price of regular apples (per pound)
`ecoprc`	Price of eco-labeled price (per pound)

1. Run a regression of `reglbs` on `regprc`. Interpret the slope coefficient. Is the sign of the slope consistent with what you know about demand curves?

2. Run a regression of `ecolbs` on `ecoprc`. Interpret the intercept coefficient.

3. Run a regression of `reglbs` on `regprc` and `ecoprc`. How does the estimated coefficient on `regprc` change? What does this tell you about the correlation between `regprc` and `ecoprc`? Justify your answer and then use `R` to verify.

4. Run a regression of `ecolbs` on `regprc` and `ecoprc`. Identify and interpret R^2.

5. Summarize your regression results in a table. Which two of the four regressions have the highest R^2? Why?

6. Construct a 99 percent confidence interval for the `ecoprc` coefficient from the regression described in exercise 4.