International School of Economics at TSU

Econometrics 2

Lasha Chochua

Problem Set 3

Instructions: You are encouraged to solve the problems before the recitation. Additionally, you are encouraged to work in groups. It is **not mandatory** to submit solutions unless stated otherwise. However, if you would like to share your solution, I would be happy to review it.

Problem 1: Data were collected from a random sample of 220 home sales from a community in 2013. Let *Price* denote the selling price (in \$1000s), *BDR* denote the number of bedrooms, *Bath* denote the number of bathrooms, *Hsize* denote the size of the house (in square feet), *Lsize* denote the lot size (in square feet), *Age* denote the age of the house (in years), and *Poor* denote a binary variable that is equal to 1 if the condition of the house is reported as "poor." An estimated regression yields

$$\widehat{Price} = 119.2 + 0.485BDR + 23.4Bath + 0.156Hsize + 0.002Lsize + 0.090Age - 48.8Poor$$
$$\bar{R}^2 = 0.72, \quad SER = 41.5.$$

a. Suppose a homeowner converts part of an existing family room in her house into a new bathroom. What is the expected increase in the value of the house?

b. Suppose a homeowner adds a new bathroom to her house, which increases the size of the house by 100 square feet. What is the expected increase in the value of the house?

c. What is the loss in value if a homeowner lets his house run down, so that its condition becomes "poor"?

d. Compute the R^2 for the regression.

Problem 2: A researcher plans to study the causal effect of police on crime, using data from a random sample of U.S. counties. He plans to regress the county's crime rate on the (per capita) size of the county's police force.

a. Explain why this regression is likely to suffer from omitted variable bias. Which variables would you add to the regression to control for important omitted variables?

b. Use your answer to (a) and the expression for omitted variable bias given in slides to determine whether the regression will likely over- or underestimate the effect of police on the crime rate. That is, do you think that $\hat{\beta}_1 > \beta_1$ or $\hat{\beta}_1 < \beta_1$?

Problem 3: Critique each of the following proposed research plans. Your critique should explain any problems with the proposed research and describe how the research plan might be improved. Include a discussion of any additional data that need to be collected and the appropriate statistical techniques for analyzing those data.

a. A researcher is interested in determining whether a large aerospace firm is guilty of sex bias in setting wages. To determine potential bias, the researcher collects data on salary and sex for all of the firm's engineers. The researcher then plans to conduct a difference-in-means test to determine whether the average salary for women is significantly less than the average salary for men.

b. A researcher is interested in determining whether time spent in prison has a permanent effect on a person's wage rate. He collects data on a random sample of people who have been out of prison for at least 15 years. He collects similar data on a random sample of people who have never served time in prison. The data set includes information on each person's current wage, education, age, ethnicity, sex, tenure (time in current job), occupation, and union status, as well as whether the person has ever been incarcerated. The researcher plans to estimate the effect of incarceration on wages by regressing wages on an indicator variable for incarceration, including in the regression the other potential determinants of wages (education, tenure, union status, and so on).

Problem 4:

Let (Y_i, X_{1i}, X_{2i}) satisfy the following assumptions:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots, n$$

where $\beta_1, \ldots, \beta_k$ are causal effects and:

1. $E(u_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$
2. $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$ are i.i.d.
3. All variables have finite fourth moments (no large outliers).
4. There is no perfect multicollinearity among the regressors.

Now, suppose that

$$\text{Var}(u_i \mid X_{1i}, X_{2i}) = 4, \quad \text{Var}(X_{1i}) = 6, \quad n = 400$$

A random sample of size $n = 400$ is drawn from the population.

a. Assume that X_1 and X_2 are uncorrelated. Compute the variance of $\hat{\beta}_1$.

b. Assume that $\text{corr}(X_1, X_2) = 0.5$. Compute the variance of $\hat{\beta}_1$.

c. Comment on the following statement:
"When X_1 and X_2 are correlated, the variance of $\hat{\beta}_1$ is larger than it would be if X_1 and X_2 were uncorrelated. Thus, if you are interested in β_1, it is best to leave X_2 out of the regression if it is correlated with X_1."

Problem 5: Consider the regression model

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

for $i = 1, \ldots, n$. (Notice that there is no constant term in the regression.)

a. Specify the least squares function that is minimized by OLS.

b. Compute the partial derivatives of the objective function with respect to b_1 and b_2.

c. Suppose that $\sum_{i=1}^n X_{1i}X_{2i} = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^n X_{1i}Y_i \big/ \sum_{i=1}^n X_{1i}^2$.

d. Suppose that $\sum_{i=1}^n X_{1i}X_{2i} \neq 0$. Derive an expression for $\hat{\beta}_1$ as a function of the data $(Y_i, X_{1i}, X_{2i}),\ i = 1, \ldots, n$.

e. Suppose that the model includes an intercept:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i.$$

Show that the least squares estimators satisfy:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

f. As in (e), suppose that the model contains an intercept. Also suppose that:

$$\sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 0$$

Show that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_{1i} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2}$$

How does this compare to the OLS estimator of β_1 from the regression that omits X_2?

Problem 6 Examine the following economic model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

a. Derive the formula for the sample least squares estimator for the parameters α and β.

b. In the regression of X on Y (the reverse of the above), what is the formula for the least squares estimator for the slope parameter on Y?

c. If the slope parameter for the reverse regression is δ, is the value of $\delta \times \beta = 1$?
Explain your reasoning.

d. Show that the geometric mean of δ and β is equal to the correlation coefficient.

Problem 7

In this exercise, you will investigate the relationship between earnings and height in R.
Use the *Earnings_and_Height.zip* file, and read the *Earnings_and_Height_description.pdf* file carefully.

a. What is the median value of height in the sample?

b. Answer the following questions:

i. Estimate average earnings for workers whose height is at most 67 inches.
ii. mEstimate average earnings for workers whose height is greater than 67 inches.
iii. On average, do taller workers earn more than shorter workers? How much more? What is a 95% confidence interval for the difference in average earnings?

c. Construct a scatterplot of annual earnings (*Earnings*) on height (*Height*).
Notice that the points on the plot fall along horizontal lines. (There are only 23 distinct values of *Earnings*.) Why? (*Hint: Carefully read the detailed data description.*)

d. Run a regression of *Earnings* on *Height*.

i. What is the estimated slope?
ii. Use the estimated regression to predict earnings for a worker who is 67 inches tall, for a worker who is 70 inches tall, and for a worker who is 65 inches tall.

e. Suppose height were measured in **centimeters** instead of inches.
Answer the following questions about the *Earnings* on *Height (in cm)* regression.

i. What is the estimated slope of the regression?
ii. What is the estimated intercept?
iii. What is the R^2?
iv. What is the standard error of the regression?

f. Run a regression of *Earnings* on *Height*, using data for **female** workers only.

i. What is the estimated slope?
ii. A randomly selected woman is 1 inch taller than the average woman in the sample. Would you predict her earnings to be higher or lower than the average earnings for women in the sample? By how much?

g. Repeat part **(f)** for **male** workers.

h. Do you think that height is uncorrelated with other factors that cause earning? That is, do you think that the regression error term, u_i, has a conditional mean of 0 given $Height(X_i)$?