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.

Solution

a. \$23,400 (recall that *Price* is measured in \$1000s).

b. In this case $\Delta BDR = 1$ and $\Delta Hsize = 100$. The resulting expected change in price is:

$$23.4 + 0.156 \times 100 = 39.0 \text{ thousand dollars or \$39,000.}$$

c. The loss is $48,800.

d. From the text $\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2)$, so:

$$R^2 = 1 - \frac{n-k-1}{n-1}(1-\bar{R}^2), \quad \text{thus } R^2 = 0.727.$$

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 $\hat{\beta}_1 < \beta_1$?

Solution

a. This regression is likely to suffer from omitted variable bias because there are factors that influence both the crime rate and the size of the police force, which are not included in the regression. For example:

- **Population density**: More densely populated counties may have both higher crime rates and larger police forces.
- **Socioeconomic status**: Counties with lower income or higher poverty may have higher crime and might also employ more police.
- **Historical crime levels**: A county with historically high crime may have increased its police force in response.

If these variables are omitted, the estimated coefficient on police force size will capture not only the causal effect but also the influence of these confounders.

To mitigate this bias, the regression should include control variables such as:

- Population or population density
- Median income or poverty rate
- Unemployment rate

- Urban vs. rural indicator
- Prior crime rates

b. The omitted variable bias formula is:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu}\left(\frac{\sigma_u}{\sigma_X}\right)$$

Where:

- X = police force size (per capita)
- u = omitted variables (e.g., crime severity, poverty, etc.)

In many cases, counties with more crime are more likely to increase police size, meaning $\rho_{Xu} > 0$. Also, if σ_u is substantial, then the bias term is positive.

So the estimated effect $\hat{\beta}_1$ will be **too large in magnitude** (less negative or even positive), leading to:

$$\hat{\beta}_1 > \beta_1$$

Thus, the regression **overestimates** the (negative) causal effect of police presence on crime.

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).

Solution

a. The proposed research in assessing the presence of sex bias in setting wages is too limited. There might be some potentially important determinants of salaries: type of engineer, amount of work experience of the employee, and education level. The sex with the lower wages could reflect the type of engineer, the amount of work experience of the employee, or the education level of the employee.

The research plan could be improved with the collection of additional data as indicated above; an appropriate statistical technique for analyzing the data would then be a multiple regression in which the dependent variable is wages and the independent variables would include a binary variable for sex, indicator variables for type of engineer, work experience and education level (highest grade level completed, for example). The potential importance of the suggested omitted variables makes a "difference in means" test inappropriate for assessing the presence of sex bias in setting wages.

So,

- The proposed difference-in-means test does not account for potential confounding variables such as education, experience, tenure, job role, and performance.
- There may be **systematic differences** in the types of roles or seniority levels held by men and women, which the test would ignore.
- As a result, the difference in average wages might **reflect structural role differences**, not direct sex bias.

Improvement:

- Instead of a simple comparison of means, the researcher should run a **multiple regression** with salary as the dependent variable and sex as one of the independent variables, **controlling for other factors** such as:
  - Years of experience
  - Education
  - Seniority/position level
  - Department or role type
- This approach allows for a **conditional comparison**, isolating the effect of sex on salary.

b. The description suggests that the research goes a long way towards controlling for potential omitted variable bias. Yet, there still may be problems. Omitted from the analysis are characteristics associated with behavior that led to incarceration (excessive drug or alcohol

use, gang activity, and so forth) that might be correlated with future earnings. Ideally, data on these variables should be included in the analysis as additional control variables.

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."

Solution

a.
Using the formula:

$$\text{Var}(\hat{\beta}_1) = \frac{1}{n} \left(\frac{1}{1 - \rho^2_{X_1, X_2}} \right) \cdot \frac{\text{Var}(u_i \mid X_{1i}, X_{2i})}{\text{Var}(X_{1i})}$$

Given:

- $\rho_{X_1, X_2} = 0$

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

We substitute:

$$\text{Var}(\hat{\beta}_1) = \frac{1}{400} \cdot \left(\frac{1}{1-0^2}\right) \cdot \frac{4}{6} = \frac{4}{2400} = \frac{1}{600} \approx 0.00167$$

b.
Now assume $\rho_{X_1,X_2} = 0.5$.

Then:

$$\text{Var}(\hat{\beta}_1) = \frac{1}{400} \cdot \left(\frac{1}{1-0.5^2}\right) \cdot \frac{4}{6} = \frac{1}{400} \cdot \frac{1}{0.75} \cdot \frac{4}{6} = \frac{4}{1800} = \frac{2}{900} \approx 0.00222$$

c.
The statement is **partially correct**.

Yes, $\text{Var}(\hat{\beta}_1)$ increases with correlation between X_1 and X_2 due to the factor:

$$\frac{1}{1-\rho^2_{X_1,X_2}}$$

However, omitting X_2 from the regression may introduce **omitted variable bias** if X_2 is also related to Y.

So even though variance decreases without X_2, the estimator $\hat{\beta}_1$ may become **biased**. That bias typically outweighs the benefit of smaller variance.

Conclusion: If X_2 is a relevant predictor of Y, it should be included, even if it increases the variance of $\hat{\beta}_1$.

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?

Solution

a.

$$\sum (Y_i - b_1 X_{1i} - b_2 X_{2i})^2$$

b.

$$\frac{\partial}{\partial b_1} \sum (Y_i - b_1 X_{1i} - b_2 X_{2i})^2 = -2 \sum X_{1i}(Y_i - b_1 X_{1i} - b_2 X_{2i})$$

$$\frac{\partial}{\partial b_2} \sum (Y_i - b_1 X_{1i} - b_2 X_{2i})^2 = -2 \sum X_{2i}(Y_i - b_1 X_{1i} - b_2 X_{2i})$$

c. From (b), $\hat{\beta}_1$ satisfies

$$\sum X_{1i}(Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) = 0$$

or

$$\hat{\beta}_1 = \frac{\sum X_{1i}Y_i - \hat{\beta}_2 \sum X_{1i}X_{2i}}{\sum X_{1i}^2}$$

d. Following analysis as in (c):

$$\hat{\beta}_2 = \frac{\sum X_{2i}Y_i - \hat{\beta}_1 \sum X_{1i}X_{2i}}{\sum X_{2i}^2}$$

Substituting this into the expression for $\hat{\beta}_1$ in (c):

$$\hat{\beta}_1 = \frac{\sum X_{1i}Y_i - \frac{\sum X_{2i}Y_i - \hat{\beta}_1 \sum X_{1i}X_{2i}}{\sum X_{2i}^2} \sum X_{1i}X_{2i}}{\sum X_{1i}^2}$$

Solving for $\hat{\beta}_1$ yields:

$$\hat{\beta}_1 = \frac{\sum X_{2i}^2 \sum X_{1i}Y_i - \sum X_{1i}X_{2i} \sum X_{2i}Y_i}{\sum X_{2i}^2 \sum X_{1i}^2 - (\sum X_{1i}X_{2i})^2}$$

e. The least squares objective function is:

$$\sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2$$

The partial derivative with respect to b_0 is:

$$\frac{\partial}{\partial b_0} \sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2 = -2 \sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})$$

Setting this to zero and solving for $\hat{\beta}_0$ yields:

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

f. Substituting $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$ into the least squares objective function yields:

$$\sum (Y_i - \hat{\beta}_0 - b_1 X_{1i} - b_2 X_{2i})^2 = \sum \left((Y_i - \bar{Y}) - b_1 (X_{1i} - \bar{X}_1) - b_2 (X_{2i} - \bar{X}_2) \right)^2$$

This is identical to the least squares objective function in part (a), except that all variables have been replaced with deviations from sample means. The result then follows as in (c).

Notice that the estimator for β_1 is identical to the OLS estimator from the regression of Y onto X_1, omitting X_2. Said differently, when

$$\sum (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 0,$$

the estimated coefficient on X_1 in the OLS regression of Y onto both X_1 and X_2 is the same as the estimated coefficient in the OLS regression of Y onto X_1.

Problem 5 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.

Solution

a. We use the least squares approach to estimate α and β:

- **Slope ($\hat{\beta}$):**

$$\hat{\beta} = \frac{s_{XY}}{s_X^2}$$

Where:

- $s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ is the sample covariance,
- $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance of X.
- Intercept ($\hat{\alpha}$):

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

b. If we reverse the regression, we estimate:

$$X_i = \gamma + \delta Y_i + \eta_i$$

Then the slope estimator is:

$$\hat{\delta} = \frac{s_{XY}}{s_Y^2}$$

c. In general:

$$\delta \times \beta = \frac{s_{XY}}{s_Y^2} \cdot \frac{s_{XY}}{s_X^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2} = r^2$$

So the product $\delta \cdot \beta$ equals the **square of the correlation coefficient**, not 1 (unless $r = \pm 1$).

d. We compute:

$$\sqrt{\delta \cdot \beta} = \sqrt{\frac{s_{XY}}{s_Y^2} \cdot \frac{s_{XY}}{s_X^2}} = \sqrt{\frac{s_{XY}^2}{s_X^2 s_Y^2}} = \frac{s_{XY}}{s_X s_Y} = r$$

Therefore, the geometric mean of δ and β equals the correlation coefficient r.