1. (Simple Regression)

A professor decides to run an experiment to measure the effect of time pressure on final exam scores. He gives each of the 400 students in his course the same final exam, but some students have 90 minutes to complete the exam while others have 120 minutes. Each student is randomly assigned one of the examination times based on the flip of a coin. Let \( Y_i \) denote the number of points scored on the exam by student \( i \) \((0 \leq Y_i \leq 100)\), let \( X_i \) denote the amount of time that the student has to complete the exam \((X_i = 90 \text{ or } 120)\), and consider the regression model \( Y_i = \beta_0 + \beta_1 X_i + u_i \).

(a) Explain what the term \( u_i \) represents. Why will different students have different values of \( u_i \)?

(b) Explain why \( E(u_i | X_i) = 0 \) holds for this regression model.

(c) Explain why \((X_i, Y_i)\) for \( i = 1, \ldots, n \) are i.i.d holds for this regression model.

(d) Explain why \( E(X^4), E(Y^4) < \infty \) holds for this regression model.

The estimated regression is

\[
\hat{Y}_i = 49 + 0.24 X_i
\]

(1)

(e) Construct a 95% confidence interval for \( \beta_1 \), the regression coefficient.

(f) Can the professor reject the hypothesis that exam time does not influence the test score at 5% significance level (two-sided test)?

Ans:

(a) The random term \( u_i \) represents factors other than time that influence the student’s performance on the exam including amount of time studying, aptitude for the material, and so forth. Some students will have studied more than average, other less; some students will have higher than average aptitude for the subject, others lower, and so forth.

(b) Because of random assignment, \( X_i \) is independent of \( u_i \). Since \( u_i \) represents deviations from average \( E(u_i) = 0 \). Because \( u_i \) and \( X_i \) are independent,

\[
E(u_i | X_i) = E(u_i) = 0.
\]
(c) \((X_i, Y_i)\) for \(i=1 \ldots n\) are i.i.d assumption is satisfied if this year’s class is typical of other classes, that is, students in this year’s class can be viewed as random draws from the population of students that enroll in the class.

(d) \(E(X^4), E(Y^4) < \infty\) assumption holds because \(0 \leq Y_i \leq 100\) and \(X_i\) can take on only two values (90 and 120).

(e) \(\beta_i \in (0.24 - 1.96*0.08, 0.24 + 1.96*0.08)\)

(f) \(H_0: \) No effects \((\beta_i = 0)\)

\(H_1: \) Some effects \((\beta_i \neq 0)\)

\[ \frac{\hat{\beta}_i - \beta_i}{S_{\beta_i}} = \frac{0.24 - 0}{0.08} = 3 > 1.96 \]

Reject the null hypothesis that exam time has no effects on exam score.

2. (OLS Estimator)

Given a linear regression model
\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]
\[ E(u_i | X_i) = 0 \]

where \(\{X_i, Y_i\}_{i=1}^n\) are i.i.d. distributed.

Suppose researcher A made a mistake—–he accidently replaces \(X\) by \(X + a\) (a is a constant), that is, \(\tilde{X}_i = a + X_i\), and \(\tilde{Y}_i = \tilde{\alpha}_0 + \tilde{\alpha}_1 \tilde{X}_i\)

(a) What is the least-squares estimator of \(\beta_1, \tilde{\alpha}_1\)?

(b) Is \(\tilde{\alpha}_1\) unbiased?

(c) Is \(\tilde{\alpha}_1\) consistent?

Ans

(a)

Minimizing \(\sum_{i=1}^{n}(Y_i - \tilde{\alpha}_0 + \tilde{\alpha}_1 \tilde{X}_i)^2\) yields \(\tilde{\alpha}_1\). Following the two first order conditions, we have,

\[ \tilde{\alpha}_1 = \frac{\sum_{i=1}^{n}(\tilde{X}_i - \tilde{X})(Y_i - \tilde{Y})}{\sum_{i=1}^{n}(\tilde{X}_i - \tilde{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \]

(b)

Since this estimator is identical to OLS estimator, it is an unbiased estimator.

\[ E(\tilde{\alpha}_1) = \beta_1 \]
Since this estimator is identical to OLS estimator, it is also a consistent estimator.
\[ \hat{\alpha}_i \rightarrow \beta_i \]

3. (Multiple Regression)

Using data on course evaluations and professor characteristics for 463 courses at the University of Texas at Austin, this question investigates how course evaluations are related to the professor’s beauty. Course evaluations are on a scale of 1 (very unsatisfactory) to 5 (excellent). Beauty is a measure of instructor physical appearance by a panel of six students, averaged across the six panelists, shifted to have mean zero. Table 1 shows regression results for various specifications.

| Table 1: The Effect of Beauty on Course Evaluation\(^1\) |
|---------------------------------|——-|——-|——-|——-|——-|
|                                | (1) | (2) | (3) | (4) | (5) |
| Beauty                         | 0.133 | 0.133 | 0.149 | 0.143 | 0.138 |
|                                | [0.022]** | [0.032]** | [0.032]** | [0.031]** | [0.037]** |
| Female                         | -0.198 | -0.215 | -0.211 |
|                                | [0.050]** | [0.053]** | [0.053]** |
| Minority                       | -0.073 | -0.046 |
|                                | [0.082] | [0.085] |
| Non-native English speaker     | -0.305 | -0.321 |
|                                | [0.102]** | [0.102]** |
| Professor’s age                | 0.036 |
|                                | [0.024] |
| age squared/100                | -0.039 |
|                                | [0.025] |
|                                | [0.025]** | [0.025]** | [0.033]** | [0.553]** | [0.563]** |
| Observations                   | 463 | 463 | 463 | 463 | 463 |
| R-squared                      | 0.04 | 0.04 | 0.07 | 0.1 | 0.1 |

\(^1\) Robust standard errors in brackets, * significant at 5%, ** significant at 1%.

(a) Model (1) regresses course evaluation on beauty, assuming homoskedastic random errors. Based on estimates of Model (1), can we reject the null hypothesis that beauty does not affect course evaluation at 1% level?

(b) Model (2) re-estimate course evaluation on beauty, but adds the “robust” option at the end of “regress” command. Why is the standard deviation of beauty in model
(2) larger than that in model (1)? Please explain carefully.

(c) Given the change in the coefficients of beauty between models (2) and (3), is beauty and female positively or negatively correlated? Please explain carefully.

Ans:

\( H_0 : \) No effects \( (\beta_{\text{beauty}} = 0) \)

\( H_1 : \) Some effects \( (\beta_{\text{beauty}} \neq 0) \)

\[
\frac{\hat{\beta}_{\text{beauty}} - \beta}{SE(\hat{\beta}_{\text{beauty}})} = \frac{0.133 - 0}{0.22} = 6 > 2.58
\]

Reject the null hypothesis that beauty has no effects on course evaluation.

(b) \( SE(\hat{\beta}_{\text{beauty}}) \) is smaller in Model (1) because the error term is assumed to be homoskedastic. Adding “robust” option at the end of regress command allows heteroskedastic random errors, resulting in a larger \( SE(\hat{\beta}_{\text{beauty}}) \) in Model (2).

(c) Recall the omitted variable formula. If the true model is the following,

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i, \quad \text{Cov}(X_i, u_i) = 0, \]

and the estimated model is

\[ Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i, \quad \epsilon_i \equiv \beta_2 Z_i + u_i \]

Then the omitted variable bias can be described as

\[
\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\text{Cov}(X_i, \epsilon_i)}{\text{Var}(X_i)} = \beta_1 + \beta_2 \frac{\text{Cov}(X_i, Z_i)}{\text{Var}(X_i)}
\]

Since the coefficient of \( \beta_1 \) increases in model (2) when beauty is included as a regressor, and \( \beta_2 < 0, \text{Cov}(\text{beauty}; \text{female}) > 0 \). Beauty and female are positively correlated.