Harris School of Public Policy

Program Evaluation

Jaskirat Kaur

TA session 1
Oct 22, 2023

Learning Goals

1. Interpret OLS estimates (ceteris paribus)
2. Identify the direction of bias in case of OVB (know the table)
3. Name the Gauss-Markov assumptions and state what they imply about OLS
4. Understand goodness of fit - difference between R^2 and adjusted R^2

Goal 1: Interpreting OLS Estimates

- Population Model:
  $COL\_GPA = B_0 + B_1 HS\_GPA + B_2 ALT + u$
- Estimated Model:
  $COL\_GPA_i = 1.29 + 0.453\text{HS\_GPA}_i + 0.0094\text{ACT}_i$
- Interpretations:
  1. Interpret the intercept coefficient.
  2. Interpret the coefficient for HS_GPA.
  3. Interpret the coefficient for ACT.
  4. Expected college GPA with 3.3 high school GPA and ACT score of 34:

$$COL\_GPA = 1.29 + 0.453(3.3) + 0.0094(34) = 3.1045$$

.

Language to Use

- A one unit increase in X is statistically associated with a B unit increase in Y.
- For every one unit increase in X, the conditional mean of Y changes by B (holding all other X equal).

Goal 2: Identifying the Direction of OVB

- Model: $COL\_GPA = B_0 + B_1 US\_GPA + B_2 ACT + u$
- Omitting ACT: $COL\_GPA - B_0 + B_1 US\_GPA$
- Questions:
  1. Where does the omitted factor go?
  2. What OLS assumption about a relevant regressor is broken when omitted?
  3. What information do we need to identify the direction of the bias?
  4. When we say that a coefficient is biased, what does that mean mathematically?
  5. Based on question 3, what is the direction of the bias?

General Case for the Simple Case

- One regressor, one omitted (X_1, X_2):

	$Cov(X1, X2) > 0$	$Cov(X1, X2) < 0$
$B2 > 0$	$(+)$ bias	(-) bias
$B2 < 0$	(-) bias	$(+)$ bias

- Overestimate: $(+)$ bias
- Underestimate: (-) bias
- Bias $= E(\hat{B}_1) - B_1 = \frac{Cov(X_1, X_2)}{Var(X_1)}$

Goal 3: Gauss-Markov Assumptions and What They Imply

1. The population model is linear in parameters:

$$y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k + u$$

2. We have a random sample of size n $(X_{i1}, X_{i2}, \ldots, X_{ik}, y_i), i = 1, 2, \ldots, n$ from the population model.
   The sample model is:

$$y_i = B_0 + B_1X_{i1} + B_2X_{i2} + \ldots + B_kX_{ik} + u_i$$

3. There is variation in each X_i and no exact linear relationships among them.
4. $E(u|X_1, X_2, \ldots, X_k) = 0$
5. $Var(u) = \sigma^2$ (homoscedasticity)

Goal 4: Understanding Measures for Goodness of Fit

- Some Definitions:
  - $SST = \sum(y_i - \bar{y})^2 \quad (n-1)$
  - $SSE = \sum(y_i - \hat{y}_i)^2 \quad (n-k-1)$
  - $SSR = \sum(\hat{y}_i - \bar{y})^2 \quad k$
- $SST = SSE + SSR$
- $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

  What happens to the R squared of a model when we add a regressor?
- Adjusted $R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(\frac{SSR}{SST}\right)$

  What happens to the Adjusted R squared of a model when we add a regressor?

Graphical Depiction