

Econometrics

Practical Session 9

Dummy Variables and Omitted Variables Bias



Ricardo Gouveia-Mendes
rgouveiamendes@ucp.pt

Spring 2025-26

Católica-Lisbon School of Business and Economics

Theoretical Wrap-up

Can We Use OLS With Categorical Regressors?

Can We Use OLS With Categorical Regressors?

- Yes, we can, using **binary variables** (a.k.a. indicator or **dummy variables**).
Example:

$$\text{wage} = \beta_0 + \beta_1 \text{male} + \beta_2 \text{female} + \beta_3 \text{north} + \beta_4 \text{center} + \beta_5 \text{south} + u$$

Can We Use OLS With Categorical Regressors?

- Yes, we can, using **binary variables** (a.k.a. indicator or **dummy variables**).

Example:

$$\text{wage} = \beta_0 + \beta_1 \text{male} + \beta_2 \text{female} + \beta_3 \text{north} + \beta_4 \text{center} + \beta_5 \text{south} + u$$

- **But can we really use OLS to estimate this model?**

Can We Use OLS With Categorical Regressors?

- Yes, we can, using **binary variables** (a.k.a. indicator or **dummy variables**).

Example:

$$\text{wage} = \beta_0 + \beta_1 \text{male} + \beta_2 \text{female} + \beta_3 \text{north} + \beta_4 \text{center} + \beta_5 \text{south} + u$$

- **But can we really use OLS to estimate this model?**
- **No!** Problem: **the dummy variable trap** → perfect multicollinearity:
 - male = 1 – female
 - north = 1 – center – south

Can We Use OLS With Categorical Regressors?

- Solution: use $j - 1$ dummy variables if you have j categories

$$\text{wage} = \beta_0 + \beta_1 \text{male} + \beta_2 \text{north} + \beta_3 \text{center} + u$$

- Notice that the meaning of the coefficients changes

Can We Use OLS With Categorical Regressors?

- Solution: use $j - 1$ dummy variables if you have j categories

$$\text{wage} = \beta_0 + \beta_1 \text{male} + \beta_2 \text{north} + \beta_3 \text{center} + u$$

- Notice that the meaning of the coefficients changes → average difference to the **reference category**
 - If male = 1 (man): $\widehat{\text{wage}} = \hat{\beta}_0 + \hat{\beta}_1$
 - If male = 0 (woman): $\widehat{\text{wage}} = \hat{\beta}_0$
 - $\hat{\beta}_0$ is the **average predicted wage** of a *woman* from the *south*
 - $\hat{\beta}_1$ is the **average predicted difference** in the wage *between a man and a woman* from the *south*

Can We Use OLS With Categorical Regressors?

- Dummy variables also allow us to create **interaction variables**. Example:

$$\text{wage} = \beta_0 + \beta_1 \text{male} + \beta_2 \text{educ} + \beta_3 \text{male} \times \text{educ} + u$$

- If male = 1 (man): $\widehat{\text{wage}} = \hat{\beta}_0 + \hat{\beta}_1 + (\hat{\beta}_2 + \hat{\beta}_3) \text{educ}$
- If male = 0 (woman): $\widehat{\text{wage}} = \hat{\beta}_0 + \hat{\beta}_2 \text{educ}$
- $\hat{\beta}_0$ is the **average predicted wage** of a *woman* without schooling
- $\hat{\beta}_1$ is the **average predicted difference** in the wage *between a man and a woman* without schooling

What If We Omit a Relevant Regressor?

- Suppose you are undecided about these two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, \quad \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$$

$$Y = \beta_0 + \beta_1 X_1 + v, \quad \tilde{\beta}_0, \tilde{\beta}_1$$

- **What happens if X_2 is relevant ($\beta_2 \neq 0$)?**

What If We Omit a Relevant Regressor?

- Suppose you are undecided about these two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, \quad \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$$

$$Y = \beta_0 + \beta_1 X_1 + v, \quad \tilde{\beta}_0, \tilde{\beta}_1$$

- **What happens if X_2 is relevant ($\beta_2 \neq 0$)?**

- If $\mathbb{E}[u|X_1, X_2] = 0 \rightarrow \hat{\beta}_1$ is unbiased

What If We Omit a Relevant Regressor?

- Suppose you are undecided about these two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, \quad \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$$

$$Y = \beta_0 + \beta_1 X_1 + v, \quad \tilde{\beta}_0, \tilde{\beta}_1$$

- **What happens if X_2 is relevant ($\beta_2 \neq 0$)?**
 - If $\mathbb{E}[u|X_1, X_2] = 0 \rightarrow \hat{\beta}_1$ is unbiased
 - If $\mathbb{E}[v|X_1] = 0 \Rightarrow \text{cov}(X_1, X_2) = 0 \rightarrow \tilde{\beta}_1$ is also unbiased

What If We Omit a Relevant Regressor?

- Suppose you are undecided about these two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, \quad \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$$

$$Y = \beta_0 + \beta_1 X_1 + v, \quad \tilde{\beta}_0, \tilde{\beta}_1$$

- **What happens if X_2 is relevant ($\beta_2 \neq 0$)?**

- If $\mathbb{E}[u|X_1, X_2] = 0 \rightarrow \hat{\beta}_1$ is unbiased
- If $\mathbb{E}[v|X_1] = 0 \Rightarrow \text{cov}(X_1, X_2) = 0 \rightarrow \tilde{\beta}_1$ is also unbiased
- If $\text{cov}(X_1, X_2) \neq 0 \Rightarrow \mathbb{E}[v|X_1] \neq 0 \rightarrow \tilde{\beta}_1$ **is biased**

$$\tilde{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{v, X_1} \cdot \sigma_v / \sigma_{X_1}$$

Exercises

Stock & Watson – Exercise 6.4

Consider the following table of estimated regressions, where:

- *AHE* = average hourly earnings
- *College* = binary variable (1 if college, 0 if high school)
- *Female* = binary variable (1 if female, 0 if male)
- *Age* = age (in years)
- *Northeast* = binary variable (1 if Region = Northeast, 0 otherwise)
- *Midwest* = binary variable (1 if Region = Midwest, 0 otherwise)
- *South* = binary variable (1 if Region = South, 0 otherwise)
- *West* = binary variable (1 if Region = West, 0 otherwise)

The data set consists of information on 7178 full-time, full-year workers, in 2015 from the Current Population Survey. The highest educational achievement for

Stock & Watson – Exercise 6.4

each worker was either a high school diploma or a bachelor's degree. The workers' ages ranged from 25 to 34 years. The data set also contains information on the region of the country where the person lived, marital status, and number of children.

Stock & Watson – Exercise 6.4

Dependent variable: average hourly earnings (*AHE*).

Regressor	(1)	(2)	(3)
College (X_1)	10.47	10.44	10.42
Female (X_2)	-4.69	-4.56	-4.57
Age (X_3)		0.61	0.61
Northeast (X_4)			0.74
Midwest (X_5)			-1.54
South (X_6)			-0.44
Intercept	18.15	0.11	0.33

Stock & Watson – Exercise 6.4

Summary Statistics

SER	12.15	12.03	12.01
R^2	0.165	0.182	0.185
\bar{R}^2			
n	7178	7178	7178

Stock & Watson – Exercise 6.4

Using the regression results in column (3):

$$\widehat{AHE} = 0.33 + 10.42 \text{ College} - 4.57 \text{ Female} + 0.61 \text{ Age} \\ + 0.74 \text{ Northeast} - 1.54 \text{ Midwest} - 0.44 \text{ South}$$

a) Do there appear to be important regional differences?

Stock & Watson – Exercise 6.4

Using the regression results in column (3):

$$\widehat{AHE} = 0.33 + 10.42 \text{ College} - 4.57 \text{ Female} + 0.61 \text{ Age} \\ + 0.74 \text{ Northeast} - 1.54 \text{ Midwest} - 0.44 \text{ South}$$

a) Do there appear to be important regional differences?

Stock & Watson – Exercise 6.4

Using the regression results in column (3):

$$\widehat{AHE} = 0.33 + 10.42 \text{ College} - 4.57 \text{ Female} + 0.61 \text{ Age} \\ + 0.74 \text{ Northeast} - 1.54 \text{ Midwest} - 0.44 \text{ South}$$

a) Do there appear to be important regional differences?

KEY TAKEAWAYS

- Being in the Northeast increases AHE by about \$0.74 relative to the West
- Being in the Midwest decreases AHE by about \$1.54 relative to the West
- Being in the South decreases AHE by about \$0.44 relative to the West
- So yes, there are important regional differences in average hourly earnings

Stock & Watson – Exercise 6.4

b) Why is the regressor *West* omitted from the regression? What would happen if it were included?

Stock & Watson – Exercise 6.4

b) Why is the regressor *West* omitted from the regression? What would happen if it were included?

b) Why is the regressor *West* omitted from the regression? What would happen if it were included?

KEY TAKEAWAYS

- The regressor was omitted to avoid the *dummy variable trap*
- If it was included it would result in a problem of perfect multicollinearity: the intercept could be written as a perfect linear function of the four regional regressors

Stock & Watson – Exercise 6.4

c) Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

Stock & Watson – Exercise 6.4

c) Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

Stock & Watson – Exercise 6.4

c) Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

KEY TAKEAWAYS

- Juanita and Jennifer share all characteristics but the region where they live
- Then:

$$\Delta_{JJ} = \hat{\beta}_5 - \hat{\beta}_6 = -1.54 - (-0.44) = -1.1$$

- Jennifer is expected to earn -\$1.1 than Juanita

Stock & Watson – Exercise 6.6

A researcher plans to study the causal effect of a strong legal system on the number of scandals in a country, using data from a random sample of countries in Asia. The researcher plans to regress the number of scandals on how strong a legal system is in the countries (an indicator variable taking the value 1 or 0, based on expert opinion).

Stock & Watson – Exercise 6.6

a) Do you think this regression suffers from omitted variable bias? Explain why. Which variables would you add to the regression?

Stock & Watson – Exercise 6.6

a) Do you think this regression suffers from omitted variable bias? Explain why. Which variables would you add to the regression?

Stock & Watson – Exercise 6.6

a) Do you think this regression suffers from omitted variable bias? Explain why. Which variables would you add to the regression?

KEY TAKEAWAYS

- The regression suffers from omitted variable bias
- Other factors that may affect the number of scandals: the country's economic conditions, political stability, or cultural norms
- Examples of variables to include: GDP per capita, the level of education, the level of technological development, the number of police officers per capita...

Stock & Watson – Exercise 6.6

b) Using the expression for omitted variable bias:

$$\hat{\beta}_j \xrightarrow{p} \beta_j + \rho_{X_j,u} \cdot \frac{\sigma_u}{\sigma_X},$$

assess whether the regression will likely over- or underestimate the effect of a strong legal system on the number of scandals in a country. That is, do you think that $\hat{\beta}_1 > \beta_1$ or $\hat{\beta}_1 < \beta_1$?

Stock & Watson – Exercise 6.6

b) Using the expression for omitted variable bias:

$$\hat{\beta}_j \xrightarrow{p} \beta_j + \rho_{X_j,u} \cdot \frac{\sigma_u}{\sigma_X},$$

assess whether the regression will likely over- or underestimate the effect of a strong legal system on the number of scandals in a country. That is, do you think that $\hat{\beta}_1 > \beta_1$ or $\hat{\beta}_1 < \beta_1$?

Stock & Watson – Exercise 6.6

b) Using the expression for omitted variable bias:

$$\hat{\beta}_j \xrightarrow{p} \beta_j + \rho_{x_j, u} \cdot \frac{\sigma_u}{\sigma_x},$$

assess whether the regression will likely over- or underestimate the effect of a strong legal system on the number of scandals in a country. That is, do you think that $\hat{\beta}_1 > \beta_1$ or $\hat{\beta}_1 < \beta_1$?

KEY TAKEAWAYS

- Suppose that the number of scandals is negatively related to the number of police officers per capita (omitted relevant variable)
- But that variable also contributes to the strength of the legal system ($\rho_{xu} > 0$)
- In that case, we expect $\hat{\beta}_1 > \beta_1$ (overestimation)