EC508: Econometrics
Efficiency, Omitted Variable Bias

Jean-Jacques Forneron

Spring, 2023

Boston University

Efficiency of OLS with Gaussian Errors

- Suppose $u_i \sim \mathcal{N}(0, \sigma_u^2)$ iid, independent of X_i
- Conditional on $X_1, \ldots, X_n$:

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma_u^2)$$

- The joint distribution (Likelihood) of the data, conditional on X, is:

$$f(y|X,\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_u^2}} e^{-\frac{1}{2\sigma_u^2}(Y_i - [\beta_0 - \beta_1 X_i])^2}$$

- Log-likelihood $= \ln f(y|X,\beta)$:

$$\ell(\beta) = n \ln(\frac{1}{\sqrt{2\pi\sigma_u^2}}) - \sum_{i=1}^{n} \frac{1}{2\sigma_u^2}(Y_i - [\beta_0 - \beta_1 X_i])^2$$

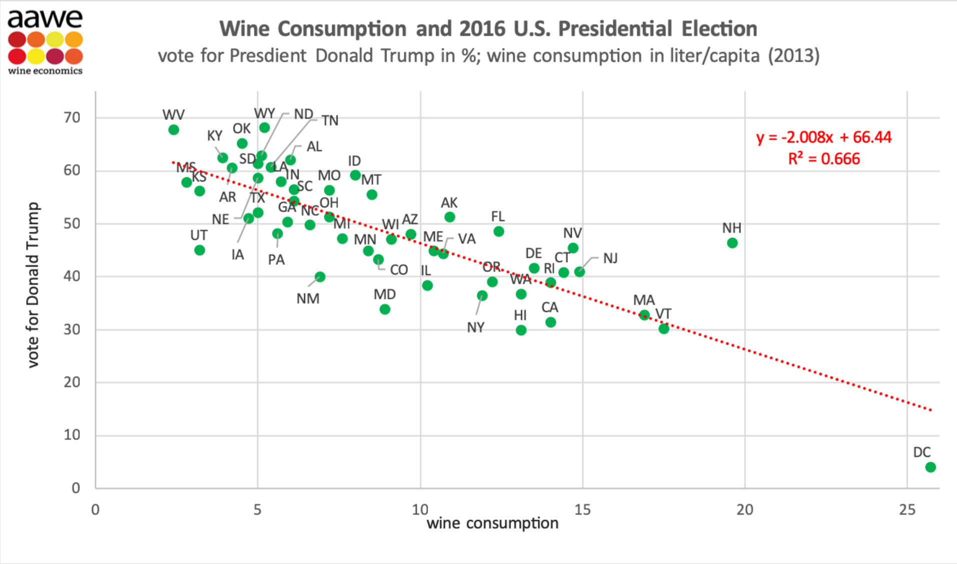
Efficiency of OLS with Gaussian Errors

- Log-likelihood $= \ln f(y|X,\beta)$:

$$\ell(\beta) = n\ln(\frac{1}{\sqrt{2\pi\sigma_u^2}}) - \sum_{i=1}^{n} \frac{1}{2\sigma_u^2}(Y_i - [\beta_0 - \beta_1 X_i])^2$$

- Maximize the (log-)Likelihood is equivalent to minimizing the Sum of Squared Residuals (SSR), i.e. the OLS objective
- This implies that $\hat{\beta}^{MLE} = \hat{\beta}^{OLS}$, MLE is efficient so OLS is efficient under Gaussian errors

In Vino Veritas: Correlation ≠ Causation



Wine Consumption and 2016 U.S. Presidential Election

vote for Presdient Donald Trump in %; wine consumption in liter/capita (2013)

Omitted Variable Bias

- The error u arises because of factors, or variables, that influence Y but are not included in the regression function. There are always **omitted variables**.
- Sometimes, the omission of those variables **can lead to bias** in the OLS estimator.

Omitted Variable Bias, cont'd

- The bias in the OLS estimator that occurs as a result of an omitted factor, or variable, is called **omitted variable** bias. For omitted variable bias to occur, the omitted variable "Z" must satisfy two conditions:
- The two conditions for omitted variable bias are
    1. Z is a determinant of Y (i.e. Z is part of u); and
    2. Z is correlated with the regressor X (i.e. $corr(Z, X) \neq 0$)
- **Both** conditions must hold for the omission of Z to result in omitted variable bias.

Omitted variable bias, cont'd

- In the test score example:
    1. English language ability (whether the student has English as a second language) plausibly affects standardized test scores: Z is a determinant of Y.
    2. Immigrant communities tend to be less affluent and thus have smaller school budgets and higher STR: Z is correlated with X.
- Accordingly, $\hat{\beta}_1$ is biased. What is the direction of this bias?
    - What does common sense suggest?
    - If common sense fails you, there is a formula...

Omitted variable bias, cont'd

- A formula for omitted variable bias: recall the equation

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n) u_i}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \frac{\frac{1}{n}\sum_{i=1}^n v_i}{\frac{n-1}{n} s_X^2}$$

- where $v_i = (X_i - \bar{X}_n) u_i \simeq (X_i - \mu_X) u_i$. Under Least Squares Assumption #1:

$$\mathbb{E}([X_i - \mu_X] u_i) = cov(X_i, u_i) = 0.$$

- But what if $\mathbb{E}([X_i - \mu_X] u_i) = cov(X_i, u_i) = \sigma_{X,u} \neq 0$?

Omitted variable bias, cont'd

- Under LSA #2 and #3 (that is, even if LSA #1 is not true)

$$
\begin{aligned}
\hat{\beta}_1 - \beta_1 &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} \\
&\xrightarrow{p} \frac{\sigma_{X,u}}{\sigma_X^2} \\
&= \frac{\sigma_u}{\sigma_X}\frac{\sigma_{X,u}}{\sigma_X \sigma_u} = \frac{\sigma_u}{\sigma_X}\rho_{X,u}
\end{aligned}
$$

- where $\rho_{X,u} = corr(X, u)$. If assumption #1 is correct, then $\rho_{X,u} = 0$, but if not we have...

The omitted variable bias formula

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\sigma_u}{\sigma_X}\rho_{X,u}$$

- If an omitted variable Z is both:
    1. a determinant of Y (that is, it is contained in u); and
    2. correlated with X
- then $\rho_{X,u} \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased and is not consistent.
- For example, districts with few ESL students (1) do better on standardized tests and (2) have smaller classes (bigger budgets), so ignoring the effect of having many ESL students factor would result in overstating the class size effect. Is this is actually going on in the CA data?

The omitted variable bias, illustration

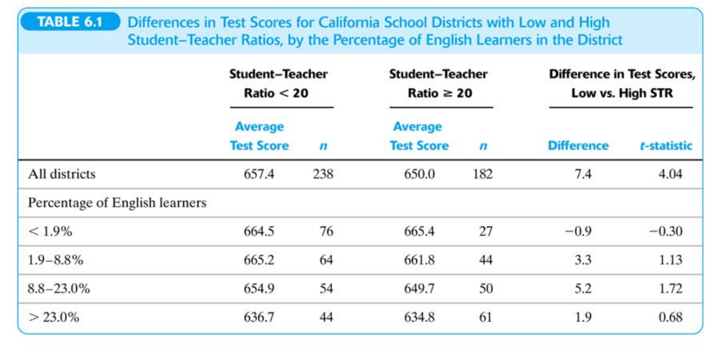


TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

- Districts with fewer English Learners have higher test scores
- Districts with lower percent EL (PctEL) have smaller classes
- Among districts with comparable PctEL, the effect of class size is small (recall overall “test score gap” = 7.4)

Causality and regression analysis

- The test score/STR/fraction English Learners example shows that, if an omitted variable satisfies the two conditions for omitted variable bias, then the OLS estimator in the regression omitting that variable is biased and inconsistent. So, even if n is large, $\hat{\beta}_1$ will not be close to β_1.
- This raises a deeper question: how do we define β_1? That is, what precisely do we want to estimate when we run a regression?

What precisely do we want to estimate when we run a regression?

- There are (at least) three possible answers to this question:
    1. We want to estimate the slope of a line through a scatterplot as a simple summary of the data to which we attach no substantive meaning.
    2. We want to make forecasts, or predictions, of the value of Y for an entity not in the data set, for which we know the value of X.
    3. We want to estimate the causal effect on Y of a change in X.

What precisely do we want to estimate when we run a regression?

- 1. We want to estimate the slope of a line through a scatterplot as a simple summary of the data to which we attach no substantive meaning
- This can be useful at times, but isn't very interesting intellectually and isn't what this course is about.

What precisely do we want to estimate when we run a regression?

- 2. We want to make forecasts, or predictions, of the value of Y for an entity not in the data set, for which we know the value of X
- Forecasting is an important job for economists, and excellent forecasts are possible using regression methods without needing to know causal effects. We will return to forecasting later in the course.

What precisely do we want to estimate when we run a regression?

- 3. We want to estimate the causal effect on Y of a change in X
- This is why we are interested in the class size effect. Suppose the school board decided to cut class size by 2 students per class. What would be the effect on test scores? This is a causal question (what is the causal effect on test scores of STR?) so we need to estimate this causal effect. Except when we discuss forecasting, the aim of this course is the estimation of causal effects using regression methods.