

Econ 312: Introduction to Econometrics

Simple Linear Regression

Sang-Yeob Lee

March 23, 2010

Introduction to Multiple Regression

Outline

- 1 Omitted variable bias
- 2 Causality and regression analysis
- 3 Multiple regression and OLS
- 4 Measures of fit
- 5 Sampling Distribution of the OLS estimator

Omitted Variable Bias

The error ϵ arise because of factor that influence Y but are not included in the regression function; so, there are always omitted variables.

Sometimes, the omission of those variables can lead to bias in the OLS estimator.

Omitted variable bias,ctd

The bias in the OLS estimator that occurs as a result of an omitted factor is called omitted variable bias. For omitted variable bias to occur, the omitted factor “Z” must be:

- 1 A determinant of Y (i.e. Z is part of ϵ); and
- 2 Correlated with the regressor X (i.e. $\text{corr}(Z, X) \neq 0$)

Both conditions must hold for the omission of Z to result in omitted variable bias.

Omitted variable bias, ctd.

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, ctd.

A formula for omitted variable bias: recall the equation,

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

where $v_i = (X_i - \bar{X})\epsilon_i \approx (X_i - \mu_x)\epsilon_i$. Under Least Square Assumption 1,

$$E[(X_i - \mu_x)\epsilon_i] = \text{Cov}(X_i, \epsilon_i) = 0.$$

But what if $E[(X_i - \mu_x)\epsilon_i] = \text{Cov}(X_i, \epsilon_i) = \sigma_{x\epsilon} \neq 0$?

Omitted variable bias, ctd.

In general (that is, even if Assumption #1 is not true),

$$\begin{aligned}\hat{\beta}_1 - \beta_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})\epsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &\rightarrow^p \frac{\text{Cov}(X, \epsilon)}{\text{Var}(X)} = \frac{\sigma_{X\epsilon}}{\sigma_X^2} \\ &= \left(\frac{\sigma_\epsilon}{\sigma_X}\right) \times \left(\frac{\sigma_{X\epsilon}}{\sigma_X\sigma_\epsilon}\right) = \left(\frac{\sigma_\epsilon}{\sigma_X}\right)\rho_{X\epsilon},\end{aligned}$$

where $\rho_{X\epsilon} = \text{corr}(X, \epsilon)$. If assumption # 1 is valid, the $\rho_{X\epsilon} = 0$, but if not we have...

The omitted variable bias formula:

$$\hat{\beta}_1 \rightarrow^p \beta_1 + \left(\frac{\sigma_{\epsilon}}{\sigma_X}\right)\rho_{X\epsilon}$$

If an omitted factor Z is both:

- 1 a determinant of Y (that is, it is contained in ϵ); *and*
- 2 correlated with X ,

then $\rho_{X\epsilon} \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased (and is not consistent).

The math makes precise the idea that districts with few ESL students (1) do better on standardized tests and (2) have smaller classes (bigger budgets), so ignoring the ESL factor results in overstating the class size effect.

Digression on causality and regression analysis

What do we want to estimate?

- What is, precisely, a causal effect?
- The common-sense definition of causality isn't precise enough for our purposes.
- In this course, we define a causal effect as the effect that is measured in an ideal randomized controlled experiment.

Ideal Randomized Controlled Experiment

- Ideal: subjects all follow the treatment protocol (perfect compliance), no errors in reporting, etc.!
- Randomized: subjects from the population of interest are randomly assigned to a treatment or control group (so there are no confounding factors)
- Controlled: having a control group permits measuring the differential effect of the treatment
- Experiment: the treatment is assigned as part of the experiment: the subjects have no choice, so there is no reverse causality in which subjects choose the treatment they think will work best.

Back to class size:

- Conceive an ideal randomized controlled experiment for measuring the effect on Test Score of reducing STR. . .
- How does our observational data differ from this ideal?
 - The treatment is not randomly assigned
 - Consider PctEL - percent English learners - in the district. It plausibly satisfies the two criteria for omitted variable bias: $Z = PctEL$ is:
 - ① a determinant of Y ; and
 - ② correlated with the regressor X .
 - The “control” and “treatment” groups differ in a systematic way - $corr(STR, PctEL) \neq 0$

Randomized controlled experiments:

- Randomization + control group means that any differences between the treatment and control groups are random - not systematically related to the treatment
- We can eliminate the difference in PctEL between the large (control) and small (treatment) groups by examining the effect of class size among districts with the same PctEL.
- If the only systematic difference between the large and small class size groups is in PctEL, then we are back to the randomized controlled experiment - within each PctEL group.
- This is one way to “control” for the effect of PctEL when estimating the effect of STR.

Return to omitted variable bias

Three ways to overcome omitted variable bias

- 1 Run a randomized controlled experiment in which treatment (STR) is randomly assigned: then PctEL is still a determinant of TestScore, but PctEL is uncorrelated with STR. (But this is unrealistic in practice.)
- 2 Adopt the “cross tabulation“ approach, with finer gradations of STR and PctEL - within each group, all classes have the same PctEL, so we control for PctEL (But soon we will run out of data, and what about other determinants like family income and parental education?)
- 3 Use a regression in which the omitted variable (PctEL) is no longer omitted: include PctEL as an additional regressor in a multiple regression.

The Population Multiple Regression Model

Consider the case of two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \quad i = 1, \dots, n$$

- Y is the dependent variable
- X_1, X_2 are the two independent variables (regressors)
- (Y_i, X_{1i}, X_{2i}) denote i th observation on Y, X_1 , and X_2 .
- β_0 = unknown population intercept
- β_1 = effect on Y of a change in X_1 , holding X_2 constant
- β_2 = effect on Y of a change in X_2 , holding X_1 constant
- ϵ_i = the regression error (omitted factors)

Interpretation of coefficients in multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \quad i = 1, \dots, n$$

Consider changing X_1 by ΔX_1 while holding X_2 constant:
Population regression line before the change:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Population regression line, after the change:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$$

- Before: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- After: $Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$
- Difference: $\Delta Y = \beta_1 \Delta X_1$
- so;
- $\beta_1 = \frac{\Delta Y}{\Delta X_1}$, holding X_2 constant
- $\beta_2 = \frac{\Delta Y}{\Delta X_2}$, holding X_1 constant
- β_0 = predicted value of Y when $X_1 = X_2 = 0$.

The OLS Estimator in Multiple Regression

With two regressors, the OLS estimator solves:

$$\min_{b_0, b_1, b_2} \sum_{i=1}^n [Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i})]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus
- This yields the OLS estimators of β_0, β_1 and β_2 .

Measures of Fit for Multiple Regression

$$\text{Actual} = \text{predicted} + \text{residual} : Y_i = \hat{Y}_i + \hat{\epsilon}_i$$

- SER = std. deviation of $\hat{\epsilon}_i$ (with d.f. correction)
- RMSE = std. deviation of $\hat{\epsilon}_i$ (without d.f. correction)
- R^2 = fraction of variance of Y explained by X
- \bar{R}^2 = “adjusted R^2 ” = R^2 with a degrees-of-freedom correction that adjusts for estimation uncertainty; $\bar{R}^2 < R^2$

SER and RMSE

As in regression with a single regressor, the SER and the RMSE are measures of the spread of the Y 's around the regression line:

$$SER = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^n \hat{\epsilon}_i^2}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2}$$

R^2 and \bar{R}^2

The R^2 is the fraction of the variance explained-same definition as in regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

Where $ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$, $SSR = \sum_{i=1}^n \hat{\epsilon}_i^2$, $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$.

- The R^2 always increases when you add another regressor (why?) - a bit of a problem for a measure of “fit”

R^2 and \bar{R}^2

The \bar{R}^2 (the "adjusted R^2 ") corrects this problem by "penalizing" you for including another regressor - the \bar{R}^2 does not necessarily increase when you add another regressor.

Adjusted R^2 : $\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right) \frac{SSR}{TSS}$

Note that $\bar{R}^2 < R^2$, however if n is large the two will be very close.