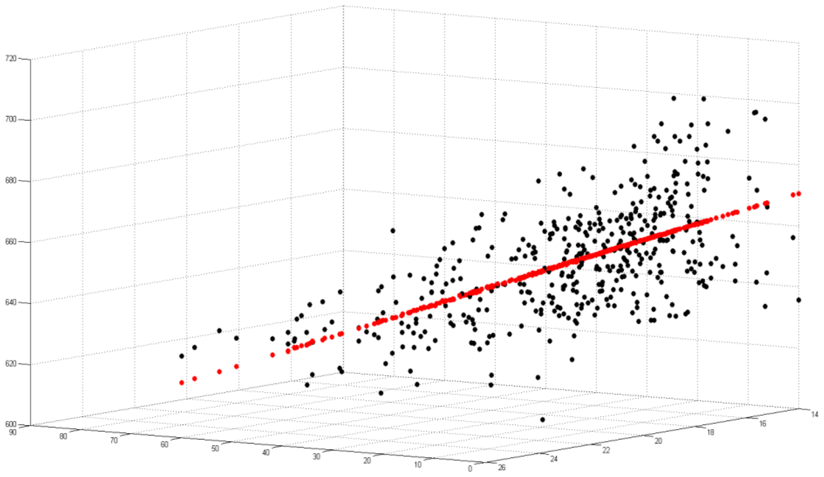
EC 3303: Econometrics I

Linear Regression with Multiple Regressors (Part 1)



Kelvin Seah

Outline

1. Omitted variable bias
2. Multiple regression and OLS

Initial Policy Question

Suppose new teachers are hired so *STR* falls by one student per class. How will this policy intervention (“treatment”) affect test scores?

- What did we find from our (simple) regression analysis?
- Negative relationship between test scores & class size (STR).
- But is this relationship ***causal***? Will smaller class sizes ***cause*** test scores to be higher?
- ***Correlation does not imply causation***!

- Reason to worry that the negative relationship is not causal.

- School districts with smaller classes (lower STR) tend also to be the wealthier ones. Hiring teachers costs money!

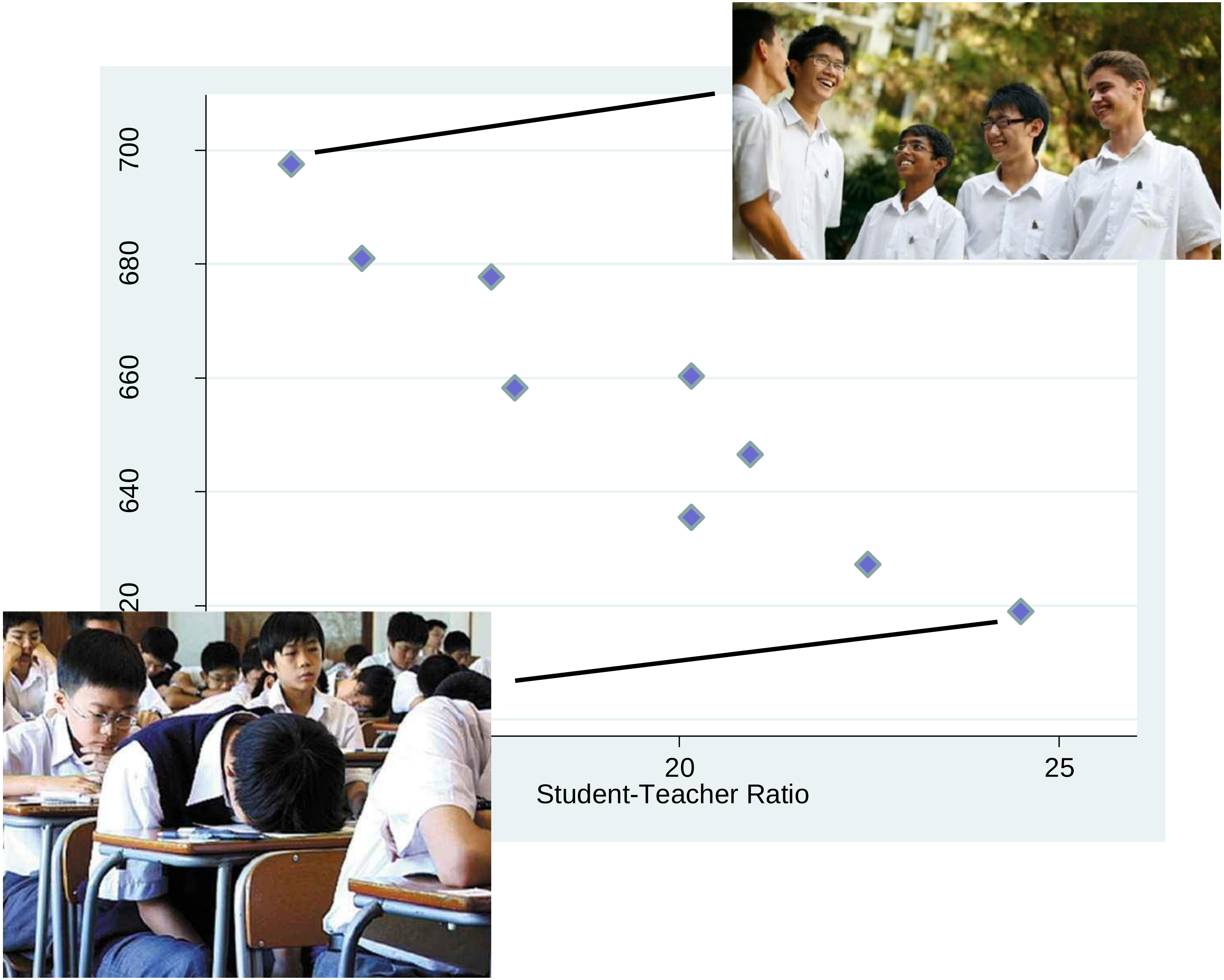
- …and students in wealthier school districts tend to come from more affluent families, have more resources, better quality teachers.

- So the negative estimated relationship between test scores and class size could actually be reflecting the influences of these other factors instead.

- These other factors ("omitted variables") could mean that the OLS analysis done so far could be misleading.

- To address the problem, need to isolate the effect on test score of changing the class size, while holding these other factors constant (*Ceteris Paribus*).

- ***Multiple regression*** allows us to study how changes in one variable affects another while holding other factors constant.


700
680
660
640
20
20
25
Student-Teacher Ratio

Omitted Variable Bias

$$TestScore_i = \beta_0 + \beta_1 STR_i + u_i$$

- By focusing only on STR, have ignored some potentially important determinants of $Testscore$ by collecting them in u.

- These omitted factors include school characteristics (teacher quality, resources), student characteristics (family background),…

- Consider an omitted student characteristic: prevalence of English learners in the school district.

Background

- In U.S., immigrant communities tend to be less affluent and so attend poorer schools with higher *STR*.

- Immigrant students (English Language Learners/ELL) also tend to perform worse academically than native students.

Background

- Since districts with larger classes also tend to have a higher percentage of ELL, an OLS regression of *TestScore* on *STR* would erroneously find a large negative estimated coefficient, even if the true causal effect of reducing class size may be very small or even zero.

- In other words, OLS estimator $\hat{\beta}_1$ will be biased.

$$E(\hat{\beta}_1) \neq \beta_1$$

Omitted Variable Bias

- Factors not included in the regressor are in the error u. So, there are **always omitted variables.**

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Sometimes, **omitted variables** can lead to **bias** in the OLS estimator.

$$Y_i = \beta_0 + \beta_1 X_i + [\beta_2 Z_i + error_i]$$

- When?

- If the omitted factor "Z" satisfies 2 conditions:
    - it is a determinant of the dependent variable "Y"
    - it is correlated with the included regressor X (*i.e.*, $corr(Z, X) \neq 0$),

then the omission will cause a bias in the OLS estimator, called ***omitted variable bias.***

$$Y_i = \beta_0 + \beta_1 X_i + [\beta_2 Z_i + error_i]$$

violates **LSA#1:** $\boldsymbol{E}[\boldsymbol{u_i}|\boldsymbol{X_i}] \ = \ \mathbf{0}$

Omitted Variable Bias & LSA#1

- Omission of "Z" means that LSA#1 is violated.

Why?

- u_i in the regression model with a single regressor represents all other factors, other than X_i, that are determinants of Y_i.
- If one of those factors is correlated with X_i, then u_i (which contains this factor) is correlated with X_i.
- Since u_i and X_i are correlated, $cov(u_i, X_i) \neq 0$ & $E(u_i|X_i = x) \neq 0$

Consequently:

$$E\left(\hat{\beta}_1\right) \neq \beta_1$$

bias does not disappear, even in large samples, so the OLS estimator is inconsistent.

Class size example

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

Consider *Z*: Percentage of English Language Learners:

1. Is Z a determinant of *Y*?

language ability affects standardized test scores so a larger percentage of ELL will affect districtwide test scores.

2. Is Z correlated with *X?*

In US, immigrant communities tend to be less affluent and thus attend schools with higher *STR*.

Thus, the OLS estimator is biased. What is the direction of this bias?

Omitted Variable Bias: An E.g.

- Does classical music make you smart?

$$TestScore_i = \beta_0 + \beta_1 ClassicalMusic_i + u_i$$

- Rauscher et al. (1993): listening to Mozart can raise your intelligence

$$TestScore_i = \beta_0 + \beta_1 ClassicalMusic_i + [\beta_2 Z_i + error_i]$$

Here, the Z_is could be student's socioeconomic background, innate ability, …

Omitted Variable Bias: An E.g.

- Do more educated people earn a higher salary?

$$Wage_i = \beta_0 + \beta_1 Education_i + u_i$$

```
. reg lwage educ, robust

Linear regression                               Number of obs =     935
                                                F(  1,   933) =   96.89
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0974
                                                Root MSE      =  .40032


                          Robust
       lwage      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

        educ   .0598392   .0060791     9.84   0.000      .047909    .0717694
       _cons   5.973063   .0822718    72.60   0.000     5.811603    6.134522
```

$$Wage_i = \beta_0 + \beta_1 Education_i + [\beta_2 Z_i + error_i]$$

What could the Z_is be here?

$$Wage_i = \beta_0 + \beta_1 Education_i + [\beta_2 Z_i + error_i]$$

What could the Z_is be here?

a) Luck
b) Innate ability
c) Motivation level
d) B & C
e) All of the above

Omitted Variable Bias

- What is the magnitude and direction of the omitted variable bias?
- Let the correlation between X_i and u_i be $Corr(X_i, u_i) = \rho_{Xu}$. Suppose LSA#2 & #3 hold, then

As $n \to \infty$,

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$$

- Because $\hat{\beta}_1$ does not converge in probability to the true value β_1, $\hat{\beta}_1$ is biased and inconsistent.
- $\rho_{Xu} \frac{\sigma_u}{\sigma_X}$ is the bias in $\hat{\beta}_1$, that persists even in large samples.

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$$

- size of the bias depends on ρ_{Xu}; the larger $|\rho_{Xu}|$, the larger the bias.
- direction of the bias in $\hat{\beta}_1$ depends on whether X and u are positively or negatively correlated:
  - If X and u are positively correlated: $\hat{\beta}_1$ will have a positive bias (i.e. $E(\hat{\beta}_1) > \beta_1$).
  - If X and u are negatively correlated: $\hat{\beta}_1$ will have a negative bias (i.e. $E(\hat{\beta}_1) < \beta_1$).
- What is the direction of the bias in the class size e.g.?

$$TestScore_i = \beta_0 + \beta_1 STR_i + [\beta_2 Z_i + error_i]$$

- omitted factor – Z – percentage of ELL has a negative effect on test scores.
  - So Z enters u_i with a negative sign.
- Meanwhile, the percentage of ELL is positively correlated with STR.
  - hence, STR would be negatively correlated with the error term u.
  - so $\rho_{Xu} < 0$ and $E(\hat{\beta}_1) < \beta_1$
  - in the simple regression of test scores on STR, $\hat{\beta}_1$ would be biased toward a negative number.
  - so one reason why the estimated slope (-2.28) suggests small classes improve test scores may be that districts with small classes have fewer ELL.

Population Multiple Regression Model

Multiple regression helps ***reduce the omitted variable bias*** by ***controlling for other related factors***.

- allows us to study the effect on Y of changing one variable (say X_1) while holding other variables (say X_2) constant

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

- Y : *dependent variable*
- X_1, X_2 : *independent variables*
- (Y_i, X_{1i}, X_{2i}) : value of Y, X_1, and X_2 for the i^{th} observation
- β_0 : unknown population intercept
- β_1 : effect on Y of a unit change in X_1, holding X_2 constant
- β_2 : effect on Y of a unit change in X_2, holding X_1 constant
- u_i : population error

- multiple regression allows isolating the effect on test scores (Y) of the class size (X_1) by holding constant the percentage of ELL in the district (X_2)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ = population regression line / function
    - summarizes the relationship between Y and the regressors (X_1 & X_2) that holds ***on average*** in the population.
- u_i = population error
    - observations do not fall exactly on the population regression line because many other factors influence the dependent variable. The influence of these other factors is contained in u_i .

Interpretating Coefficients

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Consider changing X_1 by ΔX_1 , while holding X_2 constant:

- Population regression line, ***before*** the change:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (1)$$

- Population regression line, ***after*** the change:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 \qquad (2)$$

(2) – (1)

$$\Delta Y = \beta_1 \Delta X_1 \qquad (3)$$

$\beta_1 = \dfrac{\Delta Y}{\Delta X_1}$, **holding X_2 constant**

$\beta_2 = \dfrac{\Delta Y}{\Delta X_2}$, **holding X_1 constant**

β_0 = predicted value of Y when $X_1 = X_2 = 0$.

General Case

- In practice, there might be multiple factors omitted from the single-regressor model.

  - E.g. ignoring students' economic background etc. might result in omitted variable bias, just as ignoring the percentage of English learners does.

- More generally, can have a multiple regression model with k regressors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

OLS Estimator in Multiple Regression

- Use OLS to estimate the unknown population intercept and slopes (β_0, β_1, β_2,…, β_k).

- Let b_0, b_1,…, b_k be some estimators of β_0, β_1,.., β_k . Regression line based on these estimators is $b_0 + b_1X_1 + \cdots + b_kX_k$.

- value of Y_i predicted using this line is $b_0 + b_1X_{1i} + \cdots + b_kX_{ki}$ for all $i = 1, \ldots n$

- the mistake in predicting the i^{th} observation is

$$Y_i - (b_0 + b_1X_{1i} + \cdots + b_kX_{ki})$$

- ***sum*** of the ***squared prediction mistakes*** over all n observations is

$$\sum_{i=1}^{n}[Y_i - (b_0 + b_1X_{1i} + \cdots + b_kX_{ki})]^2 \quad (1)$$

- ***sum*** of the ***squared prediction mistakes*** over all n observations is:

$$\sum_{i=1}^{n}[Y_i - (b_0 + b_1 X_{1i} + \cdots + b_k X_{ki})]^2 \qquad (1)$$

- estimators that minimize (1) are the OLS estimators of β_0, β_1,.., β_k.

- minimization problem is solved using calculus.

OLS Terminology in Multiple Regression

- Same as before:

- OLS regression line / function:

$$\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$$

- Predicted value of Y_i given the X_i's, based on the OLS regression line:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$$

- residual for the i^{th} observation:

$$\hat{u}_i = Y_i - \hat{Y}_i$$

Application: Test Scores & Class Size

- earlier regression of test scores on STR yielded:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

- estimated relationship might be misleading because it may be picking up not only the effect of class size but also the effect of other omitted factors.

- recall: districts with larger classes tend to have a greater percentage of ELL.

- let's control for the percentage of ELL in the district by including it as a regressor.

```
reg testscr str pctel, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  223.82
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4264
                                                      Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
-------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$$

More on this printout later…

$$\widehat{TestScore} = 698.9 - 2.28STR \quad (2)$$

$$\widehat{TestScore} = 686.0 - 1.10STR - 0.65PctEL \quad (3)$$

- estimated effect on test scores of a change in the STR in the second regression is only half as large as the one in the first.

- difference occurs because the coefficient on STR in the multiple regression is the effect of a one unit change in STR, holding constant (controlling for) $PctEL$, whereas in the single-regressor regression, $PctEL$ is not held constant.

- because districts with a high percentage of ELL tend to have both low test scores and high STR, omitting $PctEL$ from the regression will result in a larger estimated increase in test score from a unit decrease in the STR. However, this estimate not only reflects the effect of a decrease in the STR but also the (omitted) effect of having fewer ELL in the district.

- if adding another regressor changes the estimated coefficient on the variable of interest, this is indicative of omitted variable bias in the original regression.