ECON 7310 Elements of Econometrics
Week 4: Linear Regression with Multiple Regressors

David Du[1]

[1] University of Queensland

Draft

Outline

- Omitted variable bias
- Causality and regression analysis
- Multiple regression and OLS
- Measures of fit
- Sampling distribution of the OLS estimator

Omitted Variable Bias (SW Section 6.1)

- 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.
- 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 the following two conditions:
  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, i.e., OLS estimators are biased and inconsistent.

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, i.e., Z is part of u.
2. Immigrant communities tend to be less affluent and thus have smaller school budgets and higher STR: Z is correlated with X.

Accordingly, $\widehat{\beta}_1$ is biased. What is the direction of this bias?

In the test score example:

Suppose that the **true** model is given as

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Z_i + e_i$$

where Z_i is the proportion of ESL students in district i. Now, let's assume that the following stories are reasonable;

- $STR \uparrow \Rightarrow TestScore \downarrow$. That is, $\beta_1 < 0$
- $Z_i \uparrow \Rightarrow$ English skill $\downarrow \Rightarrow TestScore_i \downarrow$. So, Z_i is part of u_i. Indeed $\beta_2 < 0$
- $Z_i \uparrow \Rightarrow$ Educ Budget $\downarrow \Rightarrow STR_i \uparrow$. So, $corr(Z_i, STR_i) > 0$

Hence, if the equation without Z_i is estimated,

$$TestScore_i = \beta_0 + \beta_1 STR_i + \underbrace{u_i}_{=\beta_2 Z_i + e_i},$$

the effect of Z_i on *TestScore* will be partially absorbed into the effect of *STR* on *TestScore*.
That is, OLS estimate for β_1 will overestimate the effect of *STR* on *TestScore*.

What does the sample say about this?

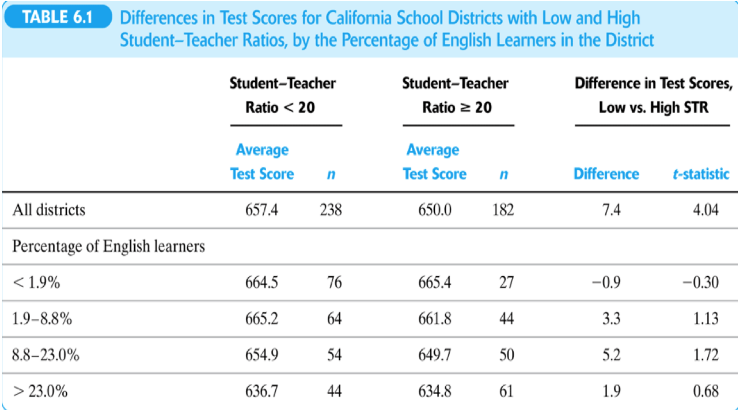


TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	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

- This example (test score/STR/fraction English Learners) 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, $\widehat{\beta}_1$ 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 scatter plot as a simple summary of the data to which we attach no substantive meaning.
   - This can be useful at times, but isn't interesting intellectually and isn't what this course is about.
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 can be done by regression methods without considering causal effects.
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 decided to cut class size by 2 students. What would be the effect on test scores? This is a causal question (what is the causal effect of *STR* on test scores?).
   - Except when we discuss forecasting, the aim of this course is the estimation of causal effects using regression methods.

What is a causal effect?

- “Causality” is a complex concept! In this course, we take a practical approach to defining causality:
- **A causal effect is defined to be the effect measured in an 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 (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

Imagine an ideal randomized controlled experiment for measuring the effect on Test Score of reducing STR.

- In that experiment, students would be randomly assigned to classes, which would have different sizes.
- Because they are randomly assigned, all student characteristics (and thus u_i) would be distributed independently of STR_i.
- Thus, $E(u_i|STR_i) = 0$, that is, LSA #1 holds in a randomized controlled experiment.

How does our observational data differ from this ideal?

- The treatment is often 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.
- Thus, the “control” and “treatment” groups differ in a systematic way, so $corr(STR, PctEL) \neq 0$.
- This means that $E(u_i|STR_i) \neq 0$ because $PctEL$ is included in u and LSA #1 is violated.

- (Randomization + control group) $\Rightarrow$ 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. (This solution to Omitted Variable bias is rarely feasible.)
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 you 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 (SW Section 6.2)

- Consider the case of two regressors:

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

- Y is the dependent variable (or LHS variable)
- X_1, X_2 are the two independent variables (regressors, RHS variables)
- (Y_i, X_{1i}, X_{2i}) denote the 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
- u_i = the regression error (omitted factors)

Interpretation of coefficients in multiple regression

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

- 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 **before** the change:

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

- Difference: $\Delta Y = \beta_1 \Delta X_1$. So,

$$\begin{aligned} \beta_1 &= \Delta Y / \Delta X_1 \text{ holding } X_2 \text{ constant,} \\ \beta_2 &= \Delta Y / \Delta X_2 \text{ holding } X_1 \text{ constant,} \\ \beta_0 &= \text{predicted value of } Y \text{ when } X_1 = X_2 = 0. \end{aligned}$$

The OLS Estimator in Multiple Regression (SW Section 6.3)

- 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 can be solved using calculus
- This yields the OLS estimators of $(\beta_0, \beta_1, \beta_2)$.

Example: the California test score data

- Regression of TestScore against STR:

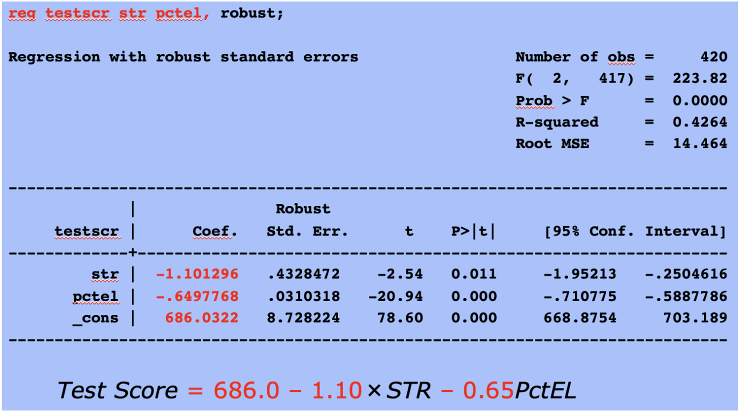
$$\mathit{TestScore} = 698.9 - 2.28 \times \mathit{STR}$$

- Now include percent English Learners in the district (PctEL):

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

- What happens to the coefficient on *STR*?

Multiple regression in STATA



```
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
------------------------------------------------------------------------------
```

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

Measures of Fit for Multiple Regression (SW Section 6.4)

- Actual = predicted + residual: $Y_i = \widehat{Y}_i + \widehat{u}_i$
- SER = standard deviation of $\widehat{u}_i$ (with d.f. correction)
- RMSE = standard deviation of $\widehat{u}_i$ (without d.f. correction)
- R^2 = fraction of variance of Y explained by X
- $\overline{R}^2$ = "adjusted R^2" = R^2 with a degrees-of-freedom correction that adjusts for estimation uncertainty; $\overline{R}^2 < R^2$

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:

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2}$$

R^2 and adjusted 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 $ESS = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$, $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, $SSR = \sum_{i=1}^{n} \widehat{u}_i^2$

- The R^2 always increases when you add another regressor – a bit of a problem for a measure of “fit”
- The $\overline{R}^2$ (the “adjusted R^2”) corrects this problem by “penalizing” you for including another regressor – the $\overline{R}^2$ does not necessarily increase when you add another regressor.

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

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

Measures of fit (continued)

Test score example:

1. $\textit{TestScore} = 698.9 - 2.28 \times \textit{STR}$ with $R^2 = 0.05$ and $\textit{SER} = 18.6$
2. $\textit{TestScore} = 686.0 - 1.10 \times \textit{STR} - 0.65 \times \textit{PctEL}$
   with $R^2 = 0.426$, $\overline{R}^2 = 0.424$, and $\textit{SER} = 14.5$

- Including *PctEL* substantially improves the goodness of fit.
  - *SER* reduces (unit of *SER* = unit of *TestScore*)
  - R^2 substantially increases.
  - Note:$R^2 \approx \overline{R}^2$ because n is large.
- Question: how to choose a variable – should we maximize $\overline{R}^2$?
  Chapter 7 will discuss about how to choose a variable for a regression analysis.

The Least Squares Assumptions for Multiple Regression (SW Section 6.5)

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

1. The conditional distribution of u given X has mean zero, that is, $E(u_i | X_{1i} = x_1, \ldots, X_{ki} = x_k) = 0$.
2. $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, are i.i.d.
3. Large outliers are unlikely: $X_1, \ldots, X_k$, and Y have four moments: $E(X_{1i}^4) < \infty, \ldots, E(X_{ki}^4) < \infty, E(Y_i^4) < \infty$
4. There is no perfect multicollinearity.

The Least Squares Assumptions for Multiple Regression (SW Section 6.5)

Assumption #1: the conditional mean of u given the included X's is zero.

$$E(u_i | X_{1i} = x_1, \ldots, X_{ki} = x_k) = 0$$

- This has the same interpretation as in regression with a single regressor.
- This condition fails when there exists an omitted variable, i.e.,
  1. belongs in the equation (so is in u) **and**
  2. is correlated with an included X
- then this condition fails and there is Omitted Variable bias.
- The best solution, if possible, is to include the omitted variable in the regression.
- A second, related solution is to include a variable that controls for the omitted variable (discussed in Ch. 7)

The Least Squares Assumptions for Multiple Regression (SW Section 6.5)

Assumption #2: $(X_{1i}, \ldots, X_{ki}, Y_i)$**,** $i = 1, \ldots, n$**, are i.i.d.**

- This is satisfied automatically if the data are collected by simple random sampling.

Assumption #3: large outliers are rare (finite fourth moments)

- This is the same assumption as we had before for a single regressor. As in the case of a single regressor, OLS can be sensitive to large outliers, so you need to check your data (scatter plots!) to make sure there are no crazy values (typos or coding errors).

Assumption #4: There is no perfect multicollinearity

- Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors. Stata automatically drops out the problematic variables.

The Sampling Distribution of the OLS Estimator (SW Section 6.6)

Under the four Least Squares Assumptions,

- $E[\widehat{\beta}_j] = \beta_j$ for $j = 0, 1, \ldots, k$, i.e., OLS estimators are unbiased
- $V(\widehat{\beta}_j)$ is inversely proportional to n
- For n large,
$$\frac{\widehat{\beta}_j - \beta_j}{SE(\widehat{\beta}_j)} \overset{approx}{\sim} \mathcal{N}(0,1)$$
- Conceptually, there is nothing new here! The way we test a simple hypothesis such as $H_0 : \beta_j = \beta_j^0$ is the same. When $\alpha = 0.05$, Reject H_0
  1. if $|\frac{\widehat{\beta}_j - \beta_j}{SE(\widehat{\beta}_j)}| > 1.96$
  2. if p-value is smaller than 0.05
  3. if β_j^0 is outside the 95% confidence interval, $\widehat{\beta}_j \pm 1.96SE(\widehat{\beta}_j)$

Multicollinearity, Perfect and Imperfect (SW Section 6.7)

- **Perfect multicollinearity** is when one of the regressors is an exact linear function of the other regressors.
- Some more examples of perfect multicollinearity
  1. Include the same variable twice, i.e., $X_1 = X_2$.
  2. Regress *TestScore* on a constant, D, and B, where D is dummy for $STR \leq 20$ and B is dummy for $STR > 20$. So, $B = 1 - D$.
- 2 above is an example of ‘dummy variable trap’. More explicitly, suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive
- That is, there are multiple categories and every observation falls in one and only one category. If you include all these dummy variables and a constant, you will have perfect multicollinearity. (Why?)
- Solutions: (1) omit one of the groups **or** (2) omit the intercept. The interpretation of the coefficients is different between (1) and (2)!!

Perfect multicollinearity (continued)

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- If you have perfect multicollinearity, your statistical software will let you know – either by crashing or returning an error message or by “dropping” one of the variables arbitrarily
- The solution to perfect multicollinearity is to modify your list of regressors so that you no longer have perfect multicollinearity.

Imperfect multicollinearity

- Imperfect and perfect multicollinearity are quite different despite the similarity of the names.
- **Imperfect multicollinearity** occurs when two or more regressors are highly correlated.
- Why the term “multicollinearity”? If two regressors are highly correlated, then their scatterplot will pretty much look like a straight line – they are “co-linear” – but unless the correlation is exactly ± 1, that collinearity is imperfect.

Imperfect multicollinearity, ctd.

- Imperfect multicollinearity implies that one or more of the regression coefficients will be imprecisely estimated (large standard errors).
- The idea: the coefficient on X_1 is the effect of X_1 holding X_2 constant; but if X_1 and X_2 are highly correlated, there is very little variation in X_1 once X_2 is held constant.
- So the data don't contain much information about what happens when X_1 changes but X_2 doesn't. If so, the variance of the OLS estimator of the coefficient on X_1 will be large.
- Example: X_1 is dummy for a woman and X_2 is dummy for a lipstick user.
- Having high standard errors is a natural result: when X_1 and X_2 are highly correlated, it is hard to disentangle the effect of X_1 on Y from the effect of X_2 on Y. So, the estimates naturally have a lot of uncertainty.