Outline

1. Omitted variable bias
2. Multiple regression and OLS
3. Measures of fit
4. 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.

Omitted variable bias, ctd.

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

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*) ≠ 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…

The omitted variable bias formula:

$$\boxed{\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}}$$

- 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 $\rho_{Xu} \neq 0$ 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*?

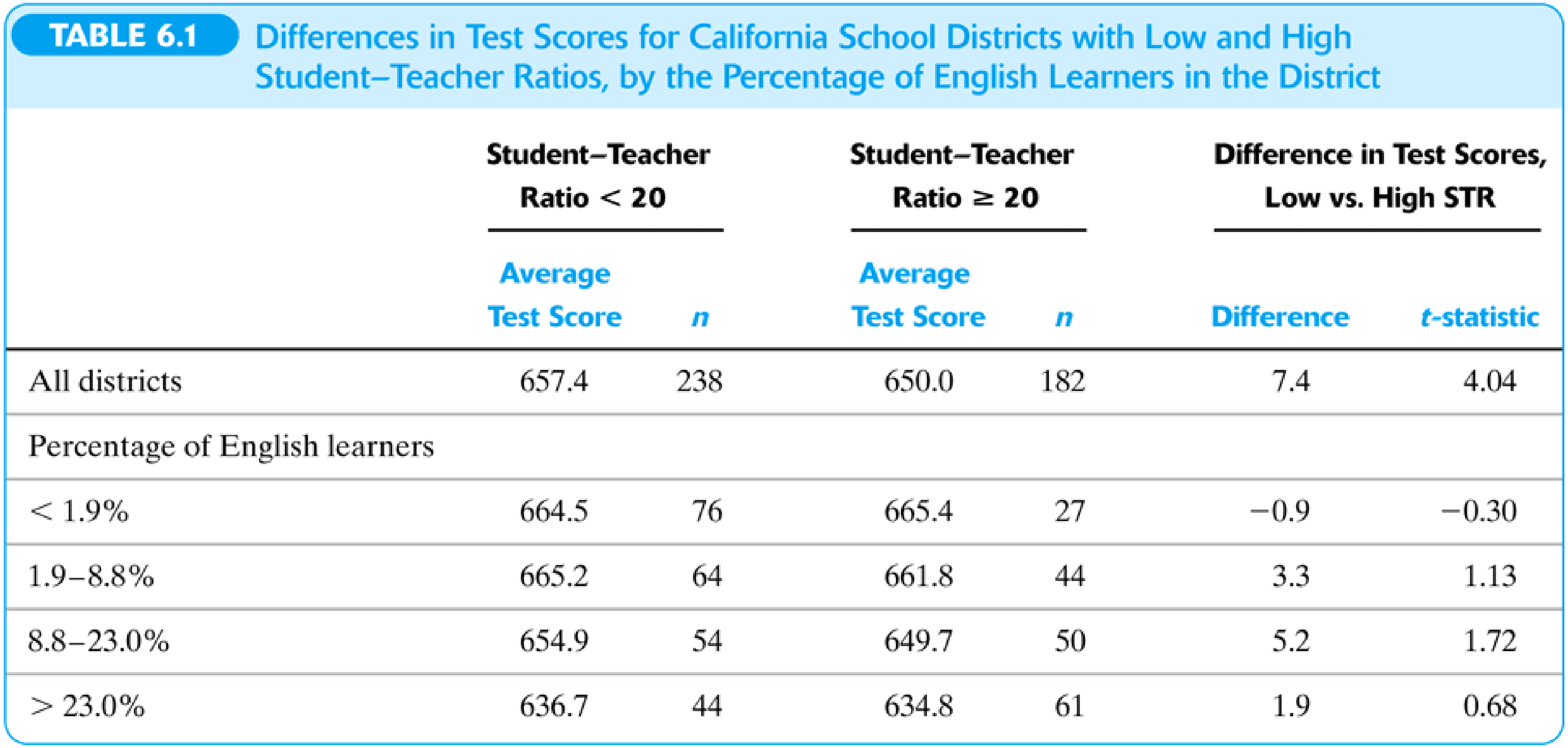

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	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?

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.*

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 OV 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.