





This book is in **Open Review**. We want your feedback to make the book better for you and other students. You may annotate some text by selecting it with the cursor and then click the  on the pop-up menu. You can also see the annotations of others: click the  in the upper right hand corner of the page



6.1 Omitted Variable Bias

The previous analysis of the relationship between test score and class size discussed in Chapters 4 and 5 has a major flaw: we ignored other determinants of the dependent variable (test score) that correlate with the regressor (class size). Remember that influences on the dependent variable which are not captured by the model are collected in the error term, which we so far assumed to be uncorrelated with the regressor. However, this assumption is violated if we exclude determinants of the dependent variable which vary with the regressor. This might induce an estimation bias, i.e., the mean of the OLS estimator's sampling distribution is no longer equals the true mean. In our example we therefore wrongly estimate the causal effect on test scores of a unit change in the student-teacher ratio, on average. This issue is called *omitted variable bias* (OVB) and is summarized by Key Concept 6.1.

1

Omitted Variable Bias in Regression with a Single Regressor

Key Concept 6.1

Omitted variable bias is the bias in the OLS estimator that arises when the regressor, X , is *correlated* with an omitted variable. For omitted variable bias to occur, two conditions must be fulfilled:

1. X is correlated with the omitted variable.
2. The omitted variable is a determinant of the dependent variable Y .

Together, 1. and 2. result in a violation of the first OLS assumption $E(u_i|X_i) = 0$. Formally, the resulting bias can be expressed as

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}. \quad (6.1)$$

See Appendix 6.1 of the book for a detailed derivation. (6.1) states that OVB is a problem that cannot be solved by increasing the number of observations used to estimate β_1 , as $\hat{\beta}_1$ is inconsistent: OVB prevents the estimator from converging in probability to the true parameter value. Strength and direction of the bias are determined by ρ_{Xu} , the correlation between the error term and the regressor.

In the example of test score and class size, it is easy to come up with variables that may cause such a bias, if omitted from the model. As mentioned in the book, a highly relevant variable could be the percentage of English learners in the school district: it is plausible that the ability to speak, read and write English is an important factor for successful learning. Therefore, students that are still learning English are likely to perform worse in tests than native speakers. Also, it is conceivable that the share of English learning students is bigger in school districts where class sizes are relatively large: think of poor urban districts where a lot of immigrants live.

Let us think about a possible bias induced by omitting the share of English learning students ($PctEL$) in view of (6.1). When the estimated regression model does not include $PctEL$ as a regressor although the true data generating process (DGP) is

$$TestScore = \beta_0 + \beta_1 \times STR + \beta_2 \times PctEL \quad (6.2)^1$$

where STR and $PctEL$ are correlated, we have

$$\rho_{STR, PctEL} \neq 0.$$

Let us investigate this using R. After defining our variables we may compute the correlation between STR and $PctEL$ as well as the correlation between STR and $TestScore$.

```
3 load the AER package
library(AER)

3 load the data set
data(CASchools)

3 define variables
CASchools$STR <- CASchools$students/CASchools$teachers
CASchools$score <- (CASchools$read + CASchools$math)/2

3 compute correlations
cor(CASchools$STR, CASchools$score)
```

```
## [1] -0.2263627
```

```
cor(CASchools$STR, CASchools$english)
```

```
## [1] 0.1876424
```

The fact that $\hat{\rho}_{STR, Testscore} = -0.2264$ is cause for concern that omitting *PctEL* leads to a negatively biased estimate $\hat{\beta}_1$ since this indicates that $\rho_{Xu} < 0$. As a consequence we expect $\hat{\beta}_1$, the coefficient on *STR*, to be too large in absolute value. Put differently, the OLS estimate of $\hat{\beta}_1$ suggests that small classes improve test scores, but that the effect of small classes is overestimated as it captures the effect of having fewer English learners, too.

What happens to the magnitude of $\hat{\beta}_1$ if we add the variable *PctEL* to the regression, that is, if we estimate the model

$$TestScore = \beta_0 + \beta_1 \times STR + \beta_2 \times PctEL + u$$

instead? And what do we expect about the sign of $\hat{\beta}_2$, the estimated coefficient on *PctEL*?

Following the reasoning above we should still end up with a negative but larger coefficient estimate $\hat{\beta}_1$ than before and a negative estimate $\hat{\beta}_2$.

Let us estimate both regression models and compare. Performing a multiple regression in R is straightforward. One can simply add additional variables to the right hand side of the `formula` argument of the function `lm()` by using their names and the `+` operator.

```
3 estimate both regression models
```

```
mod <- lm(score ~ STR, data = CASchools)
```

```
mult.mod <- lm(score ~ STR + english, data = CASchools)
```

```
3 print the results to the console
```

```
mod
```

```
##  
## Call:  
## lm(formula = score ~ STR, data = CASchools)  
##  
## Coefficients:
```

```
## (Intercept)      STR  
##      698.93      -2.28
```

```
mult.mod
```

```
##  
## Call:  
## lm(formula = score ~ STR + english, data = CASchools)  
##  
## Coefficients:  
## (Intercept)      STR      english  
##    686.0322    -1.1013    -0.6498
```

We find the outcomes to be consistent with our expectations.

The following section discusses some theory on multiple regression models.