This section discusses multiple linear regression models, and focuses in particular on omitted variable bias and multicollinearity.

1 Introduction

Why would we turn to a multiple regression analysis?

Unlike the single linear regression model, the multiple linear regression model allows us to explicitly control for other variables that could be affecting our dependent variable Y. By adding more variables into our regression function, we can explain more of the variation in Y and better predict the outcomes we are interested in. We can thus introduce more general functional form relationships.

The following equation illustrates these benefits explicitly:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \epsilon \tag{1}$$

In equation 1, we want to know the effect of education on wages. Here, we explicitly control for experience. Compared to the single linear regression model, we have effectively taken *experience* out of the error term and put it explicitly in the equation. In a simple regression analysis, we would have had to assume that experience is uncorrelated with education, an unrealistic assumption.

2 Omitted Variable Bias

Recall the zero conditional mean assumption of the linear regression model, which says that $\mathbb{E}[\epsilon|X] = 0$. This is the key assumption needed to get an unbiased estimate of β_1. If this assumption does not hold, then we cannot expect our estimate $\hat{\beta}_1$ to be close to the true value β_1 on average. One way this assumption fails is by omitting important variables from the multiple linear regression model, leading to **omitted variable bias (OVB)**. That is, the bias in our estimate that comes from failing to include key variables in the model → $\mathbb{E}[\hat{\beta}_1] \neq \beta_1$. The motivation of multiple regression is therefore to take this key variable out of the error term and include it in our estimation.

2.1 Bias

As a reminder, an estimator $\hat{\theta}_1$ is unbiased for β_1 when:

$$E[\hat{\theta}_1] = \beta_1$$

If an estimator of β_1 is unbiased for β_1, then the estimator doesn't miss the mark on average: it's neither too high, nor too low. We can view omission of a set of relevant variables as equivalent to imposing an incorrect restriction on the population model. Such an omission may be the consequence of an erroneous exclusion of a variable for which data are available or of exclusion of a variable that is not directly observed. In other words, the bias of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated.

We begin with a case where the true population model has two explanatory variables and an error term:

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

We are interested in the **causal** effect of X_1 on Y. Think of Y as hourly wage (or log of hourly wage), X_1 = education, and X_2 = a measure of innate ability. To get an unbiased estimator of β_1, we need to regress Y on X_1 and X_2. However, we don't have information on ability (X_2), and so we estimate the model by *excluding* X_2. We estimate

$$Y = \theta_0 + \theta_1 X_1 + \nu$$

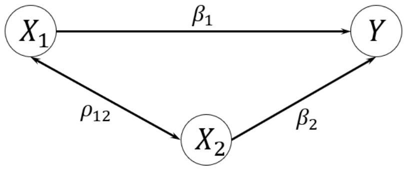
where $\nu = \beta_2 X_2 + \epsilon$. We perform the simple regression of Y on X_1 and obtain the equation:

$$\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X_1$$

To see why there is a bias, recall that with the 3 assumptions in Handout 4 satisfied, $\hat{\theta}_1$ is a consistent estimator for θ_1 as follows, and plug in the true population model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, into our formula for θ_1 and simplify:

$$\begin{aligned}
\theta_1 &= \frac{Cov\left(Y, X_1\right)}{Var\left(X_1\right)} \\
&= \frac{Cov\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, X_1\right)}{Var\left(X_1\right)} \\
&= \beta_1 \frac{Cov\left(X_1, X_1\right)}{Var\left(X_1\right)} + \beta_2 \frac{Cov\left(X_2, X_1\right)}{Var\left(X_1\right)} + \frac{Cov\left(X_1, \epsilon\right)}{Var\left(X_1\right)} \\
&= \beta_1 + \beta_2 \frac{Cov\left(X_2, X_1\right)}{\sqrt{Var\left(X_1\right) Var\left(X_2\right)}} \sqrt{\frac{Var\left(X_2\right)}{Var\left(X_1\right)}} \\
&= \beta_1 + \beta_2 \rho_{12} \sqrt{\frac{Var\left(X_2\right)}{Var\left(X_1\right)}}
\end{aligned}$$

When $\mathbb{E}\left[\hat{\theta}_1\right] \neq \beta_1$ then we say $\hat{\theta}_1$ is biased. What this means is that on average, our regression estimate is going to miss the true population parameter by $\beta_2 \rho_{12} \sqrt{\frac{Var(X_2)}{Var(X_1)}}$. Note: ρ_{12} is the correlation coefficient between X_2 on X_1.



We can always use the expression above to sign the bias. Just recall that the sign of β_2 is obtained from thinking about how the dependent variable Y is correlated with the omitted variable ($cov(Y, X_2)$) and the sign of ρ is obtained from thinking about how your independent variable of interest X_1 is correlated with the omitted variable X_2 ($cov(X_1, X_2)$).

2.2 Intuition

Suppose we want to understand how car theft rates (per capita) are affected by changes in financing for job training programs (per capita). Presumably, giving job training programs more resources will lower the rate of car thefts in a given area by increasing the economic return to legal employment. Let us imagine that the population model of car thefts looks like this, with an index of gang presence in a given district as an additional explanatory variable.

$$cartheft = \beta_0 + \beta_1 jobtrainingfinance + \beta_2 gangs + \epsilon \tag{2}$$

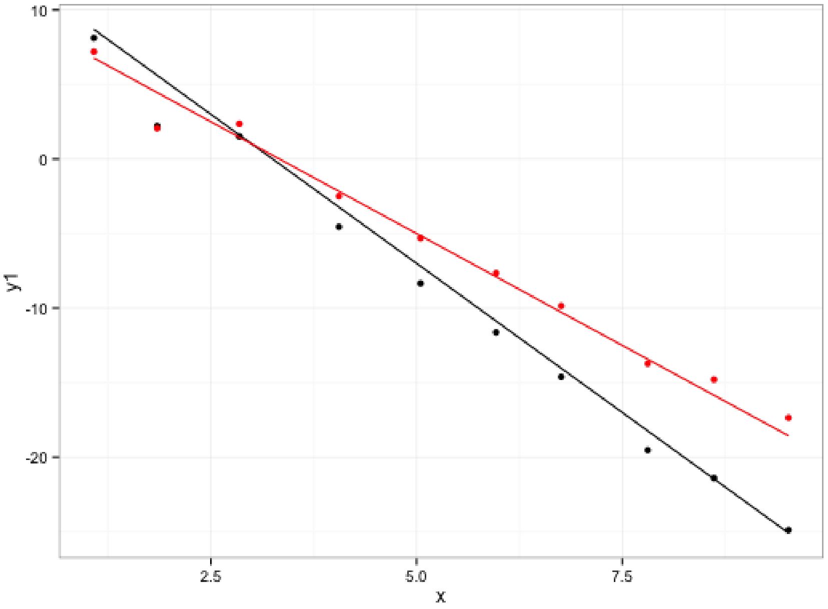
However, let's pretend that we didn't think to collect data on prevalence of gangs in each district, so that the model we have in mind is:

$$cartheft = \hat{\theta}_0 + \hat{\theta}_1 jobtrainingfinance + \nu \tag{3}$$

What happens if we omit a variable like *gangs* above? We can sign the bias we expect by determining the signs of two correlations:

1. The correlation between our omitted variable *gang* and our dependent variable *cartheft*. Here, it seems reasonable to think that $Cov(cartheft, gangs) > 0$ because the presence of gangs is generally associated with higher car theft rates.
2. The correlation between the omitted variable *gang* and our independent variable *jobtrainingfinance*. Here, $Cov(jobtrainingfinance, gangs) > 0$ if we assume that areas with higher gang activity get more financing for job training programs.

Let us consider the graph of results below. The red line represents our regression line for equation (3). For this case, we find that financing for job training programs reduces car thefts by -2. We might be tempted to say that a \$1000 increase in financing for job programs leads to a decrease in the number of car thefts by 2. Is this statement correct?



Our intuition is No. Suppose we actually collect data on gangs and run the regression (2). We then plot the relationship with the black line. We find that the slope is actually -4: \$1000 increase in financing for job training programs leads to a decrease in the number of car thefts by 4.

What is the direction of the bias in regression (3)? We can think about the difference between our biased estimate and our unbiased estimate. Our biased estimate $\hat{\theta}_1$ is -2 and our unbiased estimate $\hat{\beta}_1$ is -4. So, the bias will exactly be: $(-2-(-4)=+2$. Because this bias is positive, we say that we have an **upward bias** here. This upward bias simply means that the coefficient we estimate when we exclude gangs from the regression should be higher than the one that we get when we include gangs in the regression. We will thus have an **upward bias** in our overall estimate of the effect of financing for job training programs on car theft rates (if the biased estimate is -2 and the unbiased estimate is -4).

While this example can help us build intuition on the bias, in practice, we may never be able to collect variables on the omitted variable or even know which variables are omitted. So, we may never know the exact value of the bias. The good news is that we don't need the exact value of the bias and we can just refer to the correlations to sign it.[1]

In general, we can refer to the below table to know the sign of the bias based on the covariance between the outcome variable and the excluded independent variable and the covariance between the included independent variable and the excluded independent variable:

	$Cov(X, X_{ov}) > 0$	$Cov(X, X_{ov}) < 0$
$Cov(Y, X_{ov}) > 0$	Upward bias	Downward bias
$Cov(Y, X_{ov}) < 0$	Downward bias	Upward bias

2.3 Examples

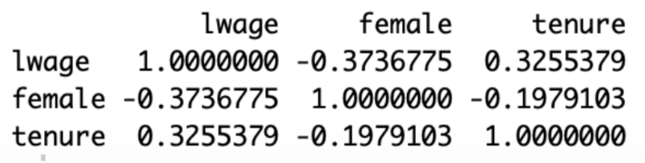
Example 1

In this subsection, we use some wage data to demonstrate the consequences of omitted variable bias and show how the OVB formula works. Let's pretend that we have a sample of 500 people, which constitutes our whole population of interest, so that when we run our regressions, we are actually revealing the true parameters instead of just estimates. We're interested in the relationship between wages and gender, and our "omitted" variable will be tenure (how long the person has been at their job). Suppose our population model is:

$$\log(wage)_i = \beta_0 + \beta_1 female_i + \beta_2 tenure_i + \epsilon_i \tag{4}$$

First let's look at the correlations between our variables and see if we can predict how omitting tenure will bias $\hat{\beta}_1$:

[1] In the above example, we actually didn't have to do the calculation to sign the bias. We knew that the bias will be positive based on our discussion on the sign of the two correlations.



```
            lwage     female     tenure
lwage   1.0000000 -0.3736775  0.3255379
female -0.3736775  1.0000000 -0.1979103
tenure  0.3255379 -0.1979103  1.0000000
```

So we have

- $Cov(wage, tenure)$ or $Cov(Y, omitted\ variable) > 0$
- $Cov(female, tenure)$ or $Cov(X, omitted\ variable) < 0$

If we instead ran the following regression:

$$\log(wage)_i = \theta_0 + \theta_1 female_i + \nu_i \tag{5}$$

Using the correlation table above, we would expect: $E(\hat{\theta}_1) < \beta_1$. How would we explain this in words? We can note that females have lower tenure, and lower tenure leads to lower wages - so if we fail to control for tenure then it will look like women have much lower wages than men because of their gender, when in reality women have lower tenure as well, and that contributes to lower wages - not just their gender.[2] Note here that we didn't need to run any regression to sign the bias.

Example 2

Consider the following population regression equation of housing prices in a county on pollution levels in the county and a dummy for whether the county is rural or urban:

$$price = \beta_0 + \beta_1 pollution + \beta_2 rural + u$$

Imagine a scenario where your friend fails to include a dummy for rural, and estimates the model as:

$$\widehat{price} = \hat{\theta}_0 + \hat{\theta}_1 pollution \quad (1)$$

And gets $\hat{\theta}_1 = -2$.

You tell your friend their mistake and they go back to collect data on rural/urban status. They then estimates the model as:

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 pollution + \hat{\beta}_2 rural \quad (2)$$

And gets $\hat{\beta}_1 = -5$, $\hat{\beta}_2 = -3$.

First, if we are given the estimates and asked to sign the bias, we can compare the two and see that $\hat{\theta}_1$ from the biased model (1) is "not as low as it should be" compared to the unbiased model (2). So we would infer that there is upward bias.

Second, if we are asked to infer $cov(rural, pollution)$, we can do this in the following way:

[2] As a side note, the reason women have lower tenure could be due to their gender itself. We put this valid consideration aside for the purpose of this example in order to simplify the model.

- We know that we have an upward bias by comparing the biased estimate $\hat{\theta}_1$ and the unbiased estimate $\hat{\beta}_1$.
- We can argue that $cov(prices, rural) < 0$ since price of houses are lower in rural areas.
- Therefore we infer that $cov(pollution, rural) < 0$.

  So when we include the variable *rural*, our estimate becomes more negative (i.e they become lower). Given that rural areas have lower housing prices, in order to get a more negative coefficient when we include the variable *rural*, it must be the case that rural areas also have lower pollution levels. This way, when we omit rural, we increase the effect of pollution on prices.

Third, suppose we did not know the values of the estimates but had to figure out the sign of the bias. We need two pieces of information to figure this out:

- $cov(prices, rural)$: We can argue that $cov(prices, rural) < 0$ since price of houses are lower in rural areas.
- $cov(pollution, rural)$: We can argue that $cov(pollution, rural) < 0$ since pollution is lower in rural areas.

Therefore we infer an upward bias from the formula

$$\mathbb{E}[\hat{\theta}_1] = \underbrace{\beta_1}_{-} + \underbrace{\beta_2}_{-}\underbrace{\rho_{12}}_{-}\sqrt{\frac{Var(X_{rural})}{(Var(pollution))}} > \beta_1$$

We observe both low pollution levels and low housing prices in rural counties. So we won't see a strong (negative) effect of pollution on prices for these counties. This can mislead us into concluding that the relationship between pollution and prices is not as negative as we may anticipate, when we exclude the variable *rural*. So the absence of a strong negative relationship between pollution and house price when we exclude the variable *rural* is because some of the low pollution counties are also rural, and rural areas have lower housing prices.

2.4 Exercises

1. Consider the following regressions:

$$\widehat{\ln(wage)} = 1.19 + 0.101 educ + 0.011 exp$$
$$\widehat{\ln(wage)} = 1.06 + 0.117 educ + 0.011 exp - 0.25 female$$

   (a) How does the coefficient on education change when we add in a dummy variable for being female. Is the bias upward or downward?

(b) Is female positively or negatively correlated with education?

2. In this exercise, we look at traffic fatalities and primary seatbelt laws. Using data for 49 US states, we can examine how primary seatbelt laws (an officer can pull you over just for not wearing your seatbelt) impact annual traffic fatalities. We have data on whether or not the state had a primary seatbelt law in place, and the total population of the state. In 2000, just 35% of the 49 states had primary seatbelt laws (the rest had what's called a secondary seatbelt law). Suppose we run the following regression:

$$\widehat{fatalities} = \hat{\theta}_0 + \hat{\theta}_1 pop + \hat{\theta}_2 primary \tag{6}$$

Here, $primary$ is a variable that is equal to 1 if the state has a primary seatbelt law and 0 otherwise. pop is the population of the state and $fatalities$ is the number of fatalities in the state.

We get

$$\widehat{fatalities} = 156.002 + 0.1232 pop + 17.258 primary \tag{7}$$

(a) What do the results tell us about the relationship between seatbelt laws and fatalities?

(b) Think of another variable or factor that you think affects traffic fatalities:

(c) Is this factor positively or negatively correlated with $fatalities$?

(d) Is this factor positively or negatively correlated with $primary$?

(e) Does omitting this factor lead to an upward or a downward bias of the true effect of the primary seatbelt law on fatalities[3]?

[3] This true effect would correspond to β_2 in the following multiple regression model: $fatalities = \beta_0 + \beta_1 pop + \beta_2 primary + \beta_3 other + \epsilon$, where the variable $other$ is the other variable omitted from the regression.

2.5 Irrelevant Variables

What happens if we include a variable that doesn't have any effect on Y (holding other variables constant) in our linear regression model? In other words, what if we include in our sample regression function a variable whose true population coefficient is zero? For example, we estimate:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

Even though the population regression function is $E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. The addition of the third variable X_3 will have no effect and will not bias $\hat{\beta}_1$ and $\hat{\beta}_2$.

3 No Perfect Multicollinearity

3.1 Definition

Recall that one of the assumptions of the linear regression model from previous handouts and sections is the following: The matrix $[\mathbf{X}_i \mathbf{X}_i^T]$ is invertible. One of the implications of this assumption is that two variables should not be **perfectly multicollinear**. By definition, two variables are said to be perfectly multicollinear if one variable is a linear combination of the other variable (where a linear combination of X_1 would be any expression of the form $aX_1 + b$). When do we run up against multicollinearity?

The simplest case of perfect multicollinearity is when one variable is a constant multiple of another. This might happen if a researcher inadvertently decides to include income measured in dollars as well as income measured in thousands of dollars.

Perfect multicollinearity also arises when one variable can be expressed as an exact linear combination of two or more of the other independent variables. For example, we may want to estimate the effect of campaign spending on voting outcomes.

Therefore, we run the following regression:

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 totexpend + \epsilon$$

where $voteA$ is the percentage of vote for Candidate A, $expendA$ are campaign expenditures for Candidate A, $expendB$ are campaign expenditures for Candidate B, and $totexpend$ are total expenditures on both campaigns. Here $totexpend = expendA + expendB$, which means we run into perfect multicollinearity.

3.2 Intuition

Why is multicollinearity a problem? So what if we have two variables that tell us the exact same information? Let's think about this in the "holding variables constant framework" that we introduced when interpreting parameters estimates in the multiple linear regression analysis.

If we include two variables X_1 and X_2 that are perfectly multicollinear (which means they vary in the exact same way) and we hold one of these variables constant (i.e. we hold X_2 fixed and don't allow it to move), then I can't by definition allow my variable of interest X_1 to increase, which is what we want to do to determine its effect on Y! So in the previous example: β_1 is supposed to measure the effect of increasing expenditures on Campaign A by one dollar, holding total spending and Campaign B's expenditures fixed. But this doesn't make sense, because if $totalexpend$ and $expendB$ are held fixed, then we can't increase $expendA$.

3.3 Exercise

The following equation represents the effects of tax revenue mix on subsequent employment growth for the *population* of counties in the United States:

$$growth = \beta_0 + \beta_1 share_P + \beta_2 share_I + \beta_3 share_S + other factors + \epsilon$$

where $growth$ is the percentage change in employment from 1980 to 1990, $share_P$ is the share of property taxes in total tax revenues, $share_I$ is the share of income taxes in total revenues, and $share_S$ is the share of sales tax in total tax revenues. All of these variables are measured in 1980. The omitted share, $share_F$, includes fees and miscellaneous taxes. By definition, the four shares add up to one. Other factors would include expenditures on education, infrastructure, and so on (all measured in 1980).

1. Why must we omit one of the tax share variables from the equation?

2. Give a careful interpretation of β_1.

3.4 Caveat

The 'no perfect multicollinerarity' assumption doesn't rule out correlation between variables, just perfect multicollinearity. Indeed, if we didn't allow for any correlation between variables, it wouldn't make much sense to perform multiple regression analysis.