Confounders & Omitted Variable Bias in Linear Regression

Erin Franke, Cheikh Fall, Vivian Powell

5/2/2022

Introduction

It is common knowledge that one of the core concepts of statistics is linear regression. Not only is it the highlight of Macalester's own Introduction to Statistics (STAT 155) course, but it has a vast range of applications to real life ranging from modeling fuel costs based on miles driven and car type to predicting ice cream sales based on temperature. But what are the underlying assumptions of this ever important modeling technique, and what happens if these assumptions do not hold? In this analysis, we take a closer look at one of the key assumptions of ordinary least squares, which is exogeneity, and the consequences of omitted variable bias if this assumption does not hold.

Background

Before covering omitted variable bias, we will introduce ordinary least squares and its assumptions. Ordinary least squares estimation, or OLS, estimates the parameters in a regression model by minimizing the sum of squared residuals, or errors (Frost, "Ordinary Least Squares"). Graphically, OLS can be thought of as the line that is closest to all points simultaneously (Addagatla). The linear equation produced by OLS for a regression with n cases and p predictors (where the subscript i denotes the case) takes the following form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

OLS is a desirable estimation technique because if certain assumptions hold, the OLS estimator is the best linear unbiased estimator (B.L.U.E). This means that it is the estimator that has the smallest variance among all linear unbiased estimators for the dependent variable (Moser). Yet in order to be the best linear unbiased estimator, the following five assumptions of the OLS model must hold (Wilms et al):

1. Linearity: all the parameters in the model are either a constant or a parameter multiplied by an independent variable (Frost, "7 Classical Assumptions").
2. No perfect multicollinearity: no explanatory variable is a perfect linear function of other explanatory variables (Frost, "7 Classical Assumptions").
3. Exogeneity: the independent variables are independent from the specified model's error term (Wilms et al).
4. Homoskedasticity and no autocorrelation: the variance of the errors should be consistent for all observations (Frost, "7 Classical Assumptions").
5. The expected value of the errors is zero

This paper will largely focus on the assumption **exogeneity** and the consequence of omitted variable bias if this assumption is not met. We will introduce this idea with an example and followed by a short series of proofs. Next we will discuss the consequences of omitted variable bias and finally conclude with steps to mitigate the bias.

Let's pretend that we are trying to model the risk of heart disease using exercise and age as explanatory variables. We understand from past studies that exercise has a negative causal relationship with heart disease - as you exercise more, you can generally expect your risk of heart disease to fall. On the other hand, age has a positive causal relationship with heart disease. The older you get, the more likely you are to develop heart disease. Using these two key explanatory variables, we define our true model as $y_i = \beta_0 + \beta_1(exercise_i) + \beta_2(age_i) + \epsilon_i$, where y_i represents heart disease risk and e_i represents the error term. But when conducting your experiment, you forgot to collect information on participant age and only tracked their activity levels. Having spent much time on the experiment, you decide to proceed anyway. Your model is therefore $y_i = \beta_0 + \beta_1(exercise_i) + e_i$, with e_i representing the error term. However, the error term, e_i, contains all variables that are not included in the specified model (Wilms et al). This means the error term essentially represents $\beta_2(age_i) + \epsilon_i$ from the true model. The exogeneity assumption for this OLS model requires that the explanatory variable (exercise) is independent of the model's error term. Is this true?

To address this question, we must ask ourselves if exercise levels and age are independent, meaning they have a covariance of 0. The answer is clear: it is common knowledge that generally, as people age, they tend to exercise less. Thus, the exogeneity assumption is violated, and age can be considered a confounding variable. **A confounding variable Z is defined as a variable that is causally associated with the outcome variable (meaning that the true value of the coefficient for Z is nonzero) and is also associated with the independent variable (this association does not need to be causal, and predictors cannot be causes of confounders).**

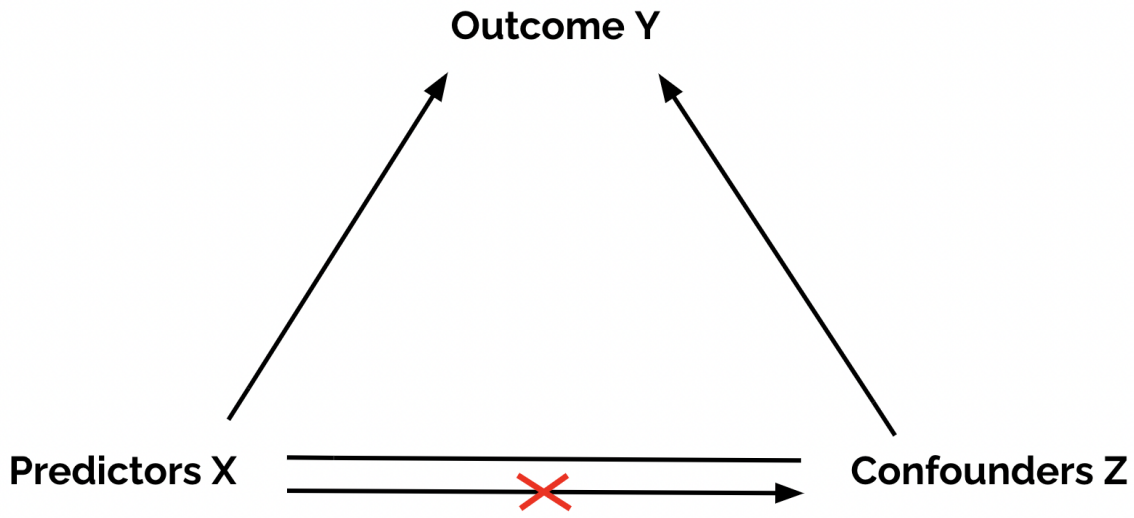

Figure 1: Visual representation of a confounding variable, where an arrow represents a causal relationship and a line represents a correlational relationship.

So why is it important that the assumption of exogeneity is violated by a confounding variable? Regardless of whether the assumption is true, when OLS is performed the estimated linear regression model forces the covariance between the independent variable (exercise) and estimated error term e_i to be zero. When in reality this covariance is *not* zero (as in this example) the resulting coefficient estimates are biased, meaning the arithmetic mean of the estimates do not yield the true value, and inconsistent, meaning that even with an increased sample size the estimates for the beta coefficients do not converge to their true value (Wilms et al). In other words, the model is wrong. To what extent the beta coefficients are biased is dependent on the strength of the correlation between the omitted and included variables.

To show this, let's return to our example using age, exercise, and heart disease risk. We ran a simulation using three different datasets - one with a strong correlation between age and exercise (r = -0.85), one with

medium correlation (r = -0.5), and one with a randomized relationship between age and exercise (r=0). In each case, we defined heart disease risk as $risk = 10 + age - 0.6(exercise)$ plus random noise. When fitting a linear model using each of the three datasets, we'd expect the coefficient on exercise to be -0.6 and the coefficient on age to be 1. We experimented with what happened to our coefficients when omitting age as a variable, thus fitting a model for risk as a function of exercise. In the highly correlated dataset, the β coefficient on exercise was -0.9. In other words, the coefficient on exercise was biased by being more negative than expected. In the dataset with medium correlation, the coefficient on exercise when omitting age was -0.8, which is closer to the true value of -0.6 than that of correlated dataset but still more negative than expected. Finally, in the dataset with no correlation between exercise and age - in other words, zero covariance - the coefficient on exercise was essentially exactly what was expected: -0.6. In summary, this simulation shows that as the covariance between the omitted and included variable strengthens, the bias on the included coefficient increases. Later in this report this concept will be proved mathematically.

Risk as a function of exercise (age omitted)	**Dataset**		
	No Correlation (R=0.0)	Medium Correlation (R=-0.5)	High Correlation (R=-0.85)
Expected beta coefficient	-0.6	-0.6	-0.6
Simulated beta coefficient	-0.6	-0.8	-0.9

Figure 2: Simulation Results

Results

We will now demonstrate omitted variable bias through a series of proofs.

Least Squares estimator with no confounding variable

Assume that the true relationship between Y and X is denoted by the equation $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Using least squares estimation, we have derived that the estimator for β_1 is:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x} y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, \text{ moving } \bar{x} \text{ within the sum} \\
&= \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x} y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^n x_i y_i - \bar{x} y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, \text{ combining the contents of the sum} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i)}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}
\end{aligned}$$

Assume that the error term ϵ_i has a mean $\mu = 0$. To prove that $\hat{\beta}_1$ is unbiased for β_1, we must show that $E(\hat{\beta}_1) = \beta_1$, since Bias$(\beta_1) = E(\hat{\beta}_1) - \beta_1$.

$$
\begin{aligned}
E(\hat{\beta}_1) &= E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right] \\
&= \frac{1}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} E\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i)\right], \text{ because we can treat x as fixed and known} \\
&= \frac{\sum_{i=1}^{n} E\left[(x_i - \bar{x})(y_i)\right]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \text{ by linearity of expectation} \\
&= \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})E(y_i)\right]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})E(\beta_0 + \beta_1 x_i + \epsilon_i)\right]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \text{ since } y_i = \beta_0 + \beta_1 x_i + \epsilon_i \\
&= \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})(\beta_0 + \beta_1 x_i + E(\epsilon_i))\right]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})(\beta_0 + \beta_1 x_i)\right]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \text{ since } E(\epsilon_i) = 0 \text{ by assumption} \\
&= \frac{\sum_{i=1}^{n}\left[\beta_0 x_i + \beta_1 x_i^2 - \beta_0\bar{x} - \beta_1\bar{x}x_i\right]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\beta_0\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n} x_i^2 - \beta_0\sum_{i=1}^{n}\bar{x} - \beta_1\bar{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\beta_0\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n} x_i^2 - \beta_0\sum_{i=1}^{n}\bar{x} - \beta_1\bar{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\beta_0 n\bar{x} - \beta_0 n\bar{x} + \beta_1\sum_{i=1}^{n} x_i^2 - \beta_1\bar{x}n\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \text{ since } \sum_{i=1}^{n} x_i = n\bar{x} \\
&= \frac{\beta_1\sum_{i=1}^{n} x_i^2 - \beta_1 n\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\beta_1(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \beta_1, \text{ as desired}
\end{aligned}
$$

Least squares estimator with a confounding variable

Now, let us assume that there is a second predictor Z of Y such that the true model is $y = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$. Z is also associated with X such that $Cov(X, Z) \neq 0$. Therefore, as established above, Z is a confounding variable. However, without the knowledge of a potential confounding variable, we fit a model with only X, not accounting for Z.

In this case, our least squares estimator for β_1, $\hat{\beta}_1$, is the same as before. Namely,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Following the same steps as in the case without confounding, an expression for the expected value of $\hat{\beta}_1$ is:

$$E(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})E(y_i)\right]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Let us now derive expression for $E(\hat{\beta}_1)$ in terms of X and Z:

$$
\begin{aligned}
E(\hat{\beta}_1) &= \frac{\sum_{i=1}^n [(x_i - \bar{x})E(y_i)]}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^n [(x_i - \bar{x})E(\beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i)]}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^n [(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \beta_2 z_i + E(\epsilon_i))]}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^n [(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \beta_2 z_i)]}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, \text{ since } E(\epsilon_i) = 0 \text{ by assumption} \\
&= \frac{\sum_{i=1}^n \left[\beta_0 x_i + \beta_1 x_i^2 + \beta_2 x_i z_i - \beta_0 \bar{x} - \beta_1 \bar{x} x_i - \beta_2 \bar{x} z_i\right]}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 + \beta_2 \sum_{i=1}^n x_i z_i - \beta_0 \sum_{i=1}^n \bar{x} - \beta_1 \bar{x} \sum_{i=1}^n x_i - \beta_2 \bar{x} \sum_{i=1}^n z_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 - \beta_0 \sum_{i=1}^n \bar{x} - \beta_1 \bar{x} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} + \frac{\beta_2 \sum_{i=1}^n x_i z_i - \beta_2 \bar{x} \sum_{i=1}^n z_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\beta_0(\sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x}) + \beta_1(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i)}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} + \frac{\beta_2(\sum_{i=1}^n x_i z_i - \bar{x} \sum_{i=1}^n z_i)}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\beta_0(n\bar{x} - n\bar{x}) + \beta_1(\sum_{i=1}^n x_i^2 - n\bar{x}^2)}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} + \frac{\beta_2(\sum_{i=1}^n x_i z_i - \bar{x} \sum_{i=1}^n z_i)}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, \text{ since } \sum_{i=1}^n x_i = n\bar{x} \\
&= \frac{\beta_1(\sum_{i=1}^n x_i^2 - n\bar{x}^2)}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} + \frac{\beta_2(\sum_{i=1}^n x_i z_i - n\bar{x}\bar{z})}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, \text{ since } \sum_{i=1}^n z_i = n\bar{z} \\
&= \beta_1 + \beta_2 \frac{\sum_{i=1}^n x_i z_i - n\bar{x}\bar{z}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}
\end{aligned}
$$

However, we also can show that:

$$
\begin{aligned}
\frac{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} &= \frac{\sum_{i=1}^n (z_i x_i - z_i \bar{x} - \bar{z} x_i + \bar{z}\bar{x})}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^n z_i x_i - \bar{x} \sum_{i=1}^n z_i - \bar{z} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{z}\bar{x}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^n z_i x_i - n\bar{x}\bar{z} - n\bar{z}\bar{x} + n\bar{z}\bar{x}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, \text{ since } \sum_{i=1}^n x_i = n\bar{x} \text{ and } \sum_{i=1}^n z_i = n\bar{z} \\
&= \frac{\sum_{i=1}^n z_i x_i - n\bar{x}\bar{z}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}
\end{aligned}
$$

Thus, we can rewrite the expected value of $\hat{\beta}_1$ as:

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

It is important to note that the bias of $\hat{\beta}_1$ (the difference between its expected value and the true value of the parameter) is $\beta_2 \frac{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}{\sum_{i=1}^n x_i^2 - nx^2}$. This can be rewritten as $\beta_2 \frac{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$, which is equivalent to $\beta_2 \frac{\widehat{cov(x,z)}}{\widehat{var(x)}}$, where $\widehat{cov(x,z)}$ and $\widehat{var(x)}$ represent the sample covariance of X and Z and the sample variance of X, respectively. This indicates that as the sample covariance of X and Z increases and the coefficient

of the true effect of Z on Y increases, the bias of $\hat{\beta}_1$ will increase too. Therefore, the strength of omitted variable bias is contingent upon how much X (the independent variable) and Z (the omitted variable) vary together within the sample and upon the strength of the true relationship between Z and Y (the outcome variable). Note that if Z and X have no association in the data (the sample covariance is zero), then the estimator will be unbiased. Similarly, if $\beta_2 = 0$ and there is no true causal relationship between Z and Y, the estimator will also be unbiased. This indicates that an omitted variable must be a confounding variable in order to cause omitted variable bias.

Least Squares estimator in Matrix Form

Let us now estimate the following model:

$$E[y|x_1, x_2, ..., x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p.$$

Consider the vector of outcomes $\mathbf{y}$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

the vector of covariates $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

and the matrix of covariates (sometimes referred to as the "design matrix") $\mathbf{X}$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix}.$$

Then, we can write our linear regression model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $E[\boldsymbol{\epsilon}] = \mathbf{0}$.

Using matrix notation, we can formulate the least squares problem as follows:

$$\text{argmin}_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Let us find the value of $\boldsymbol{\beta}$ that minimizes the sum of squared residuals $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.

First, take the derivative with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \left(\mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \right) \\ &= -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \end{aligned}$$

Then, set this equal to zero and solve for $\boldsymbol{\beta}$:

$$\begin{aligned} -2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &\stackrel{set}{=} 0 \\ 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &= 2\mathbf{X}^\top\mathbf{y} \\ \mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^\top\mathbf{y} \\ (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \end{aligned}$$

Bias Property of the Ordinary Least Squares Estimator without Confounder: Matrix Form

In this case, we assume we have fit the correct model and accounted for all potential confounders in the matrix $\mathbf{X}$ .To prove that $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$, we must show that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}] \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E[\mathbf{y}], \text{ since } \mathbf{X} \text{ is known} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}], \text{ since } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \text{ by assumption} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{E}[\epsilon]), \text{ since } \mathbf{X} \text{ is known} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}, \text{ since } E[\boldsymbol{\epsilon}] = 0 \text{ by assumption} \\ &= \boldsymbol{\beta}, \text{ as desired} \end{aligned}$$

Bias Property of the Ordinary Least Squares Estimator with Confounder: Matrix Form

Let us now assume that the true model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}$$

with one confounding variable $\mathbf{Z}$ such that:

$$\mathbf{Z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix},$$

However, without knowledge of the confounding variable, we fit the model as we did in the case with no confounding. Thus, the least squares estimator is still:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

In this case, let us find $E(\hat{\boldsymbol{\beta}})$.

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E[\mathbf{y}], \text{ since } \mathbf{X} \text{ is known} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}], \text{ since } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon} \text{ by assumption} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{E}[\epsilon]), \text{ since } \mathbf{X} \text{ and } \mathbf{Z} \text{ are known} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta}), \text{ since } E[\boldsymbol{\epsilon}] = 0 \text{ by assumption} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}\boldsymbol{\delta} \\
&= \boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}\boldsymbol{\delta}
\end{aligned}
$$

Thus, this derivation suggests that $E(\hat{\boldsymbol{\beta}})$ is biased if $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}\boldsymbol{\delta} \neq 0$. Similar to the derivation with one independent variable, we can divide this bias term into two parts. $\boldsymbol{\delta}$ represents the relationship between $\mathbf{Z}$ and $\mathbf{Y}$, accounting for $\mathbf{X}$. Thus, if there is no association between $\mathbf{Z}$ and $\mathbf{Y}$ ($\boldsymbol{\delta} = \mathbf{0}$), $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$. Therefore, the stronger the association between the outcome and the confounder, the larger the bias on our estimate $\hat{\boldsymbol{\beta}}$ if the confounder is omitted.

The second part of this bias term, $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}$, can be interpreted as the relationship between the confounder $\mathbf{Z}$ and the vector of covariates $\mathbf{X}$. Similar to the case where there is only one independent variable we accounted for, this result also shows that the stronger the relationship between the confounder and other independent variables present in the model, the larger the bias on $\hat{\boldsymbol{\beta}}$ if the confounder is omitted.

In sum, the matrix form of the omitted variable bias derivations serves as an extension of our derivation with only one covariate. With matrices, we are able to account for multiple covariates and we find very similar consequences of not including a confounding variable in the regression.

Discussion

Omitted variable bias carries significant consequences for studies that use linear regression. Under the assumption of exogeneity, ordinary least-squares estimators are unbiased and consistent, and conclusions are drawn from linear regressions based on that assumption. However, issues arise when exogeneity is broken by the omission of a confounder. Depending on the strength and signs of the correlation between the explanatory variable, outcome variable, and omitted variable, a regression that is subject to omitted variable bias can mask, heighten, or change the sign of the true relationship between two variables (Wilms et al). Using biased estimates is dangerous because it can lead to assumptions of causation that could be untrue. For example, let's consider a case where the omission of a confounding variable could drastically inflate the coefficient for an included predictor far beyond the true value of its relationship with the outcome variable. Seeing such a strong relationship (when in fact the relationship is weak) could lead to conclusions about the data that are plainly incorrect and could cause harm. Conversely, seeing a small coefficient when the true parameter is much larger could lead studies to miss important relationships among variables that prevent necessary actions from being taken. Omitted variable bias also affects the validity of hypothesis testing and p-values, which rely on the estimate of the beta-coefficient being unbiased (Wilms et al). Conducting hypothesis tests with a biased estimate can invalidate results and lead researchers to mistakenly label a relationship as statistically significant or insignificant.

While we have gone through examples of running models with and without a particular variable, often in real life situations that choice is not an option. The missing confounding variable sometimes is simply not attainable. In such cases, what are steps that we can take to reduce omitted variable bias?

Whenever possible, one of the best ways to minimize omitted variable bias is through experimental design (Wilms et al). If participants are assigned randomly into experimental and control groups with sufficiently large sample sizes, aside from the treatment variable the variables are distributed equally to both groups and any omitted variable is missing from both groups. Therefore, the groups only differ with respect to treatment and observed differences can be attributed to this treatment (Wilms et al). However, in reality a large majority of research questions cannot be executed as an experiment. Thus, another solution is needed.

One solution is a *proxy variable*. A proxy variable is defined as an observed variable that is related to but not identical to an unobserved explanatory variable in multiple regression analysis (Van Kammen). For example, a proxy variable for quality of life might be per-capita GDP, and one for true body fat percentage could be BMI (Frost, "Proxy Variables: The Good Twin of Confounding Variables"). To understand a little more about the math behind a proxy variable, let's think of an example of predicting wages using education (x_1), labor experience (x_2), and ability (x_3^*). Ability is essentially unmeasurable and thus we will use IQ (x_3) as a proxy variable for it. If x_3 is a proxy for ability (x_3^*), the relationship between them might look like $x_3^* = \delta_0+\delta_3 x_3+v_3$, where error term v_3 represents the non-identical relationship between the proxy and unobserved variable (Van Kammen). δ_3 represents the relationship between IQ and ability, and thus if δ_3 is 0 we understand that IQ is a poor proxy for ability. We can substitute our equation for x_3^* into our linear regression equation for wages, which is $y = \beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\epsilon$, and get $y = \beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3(\delta_0+\delta_3 x_3+v_3)+\epsilon$. Rearranging, we end up with $y = (\beta_0+\beta_3\delta_0)+\beta_1 x_1+\beta_2 x_2+\beta_3\delta_3 x_3+(\beta_3 v_3+\epsilon)$. In order to have consistent estimates of β_1 and β_2, the expected value of the new error term $(\beta_3 v_3+\epsilon)$ must be 0, which is satisfied if $E(v_3|x_1,x_2,x_3) = E(\epsilon|x_1,x_2,x_3) = 0$. In other words, the relationship between ability and IQ does not depend on education and labor experience, allowing us to estimate β_1 and β_2 without bias. The coefficient on our proxy variable (IQ) itself will actually not be unbiased or consistent, but generally the goal is for consistent and unbiased estimates of β_1 and β_2 and overall bias reduction (Van Kammen). Thus, using a proxy variable is a common and relatively trusted solution to omitted variable bias.

Works Cited

Addagatla, Arun. "Ordinary Least Squares Regression." Medium, Geek Culture, 20 Apr. 2021, https://medium.com/geekculture/ordinary-least-squares-regression-41f40400a58d./

Frost, Jim. "7 Classical Assumptions of Ordinary Least Squares (OLS) Linear Regression." Statistics By Jim, 8 Sept. 2021,https://statisticsbyjim.com/regression/ols-linear- regression-assumptions/.

Frost, Jim. "Ordinary Least Squares." Statistics By Jim, 5 May 2017,https://statisticsbyjim.com /glossary/ordinary-least-squares/.

Frost, Jim. "Proxy Variables: The Good Twin of Confounding Variables" Statistics By Jim, https://statisticsbyjim.com/regression/proxy-variables.

Moser, Barry Kurt. "Linear Unbiased Estimator." Linear Unbiased Estimator - an Overview | ScienceDirect Topics,https://www.sciencedirect.com/topics/mathematics/linear -unbiased-estimator.

Van Kammen, Ben. More on Specification and Data Issues.https://web.ics.purdue.edu/~bvankam m/Files/360%20Notes/08%20-%20Specification%20and%20Data%20Issues.pdf.

Wilms, R., et al. "Omitted Variable Bias: A Threat to Estimating Causal Relationships." Methods in Psychology, Elsevier, 10 Sept. 2021,https://www.sciencedirect.com/science /article/pii/S2590260121000321.