Applied Econometrics: When Can an Omitted Variable Invalidate a Regression?

Hal J. Singer and Kevin W. Caves

In his *New York Times* best seller, *Thinking Fast and Slow*, Nobel Prize laureate Daniel Kahneman explained that the party in litigation proffering a regression model to a jury was bound to lose. The reason is that regression runs counter to the way we think: Our brains tend to overcompensate when we learn a random fact about a person or thing. Regression tells us that the best predictor is often the mean of the population;[1] hence, the popular phrase "regressing towards the mean." A prediction can often be improved by deviating from the mean, but only if certain conditions are satisfied.

One such condition is that the regression does not suffer from so-called omitted variable bias. The objective of this article is to illustrate this concept, which frequently comes up during antitrust litigation, for non-economists. We will answer the following questions: What is omitted variable bias? Why is it so important to so many econometric debates? When is it empirically relevant? And how do we deal with it in real-world antitrust litigation?

Omitted Variable Bias: Intuition

As illustrated in Figure 1, the essence of regression analysis is to use variation in *X* (the independent variable) to explain variation in *Y* (the dependent variable). We can see in Figure 1 that, when *X* is above its average, *Y* also tends to be above its average. Note that we need both forms of variation if we want our regression to make a meaningful prediction.

To illustrate, suppose that *Y* did not deviate at all from its average, such that all of the blue dots were clustered along the red line. In that case, there is literally nothing a regression can do to improve upon the average: Guessing the average will always predict *Y* perfectly. Knowledge of *X* would therefore be irrelevant to predicting *Y*. Conversely, suppose that there is variation in *Y*, but not in *X*. In that case, there would be no hope to use movements in *X* to predict *Y*, because X does not move at all. Our best guess would revert, once again, to the mean of *Y*.

Intuitively, omitted variable bias occurs when the independent variable (the *X*) that we have included in our model picks up the effect of some other variable that we have omitted from the model. The reason for the bias is that we are attributing effects to *X* that should be attributed to the omitted variable.

As illustrated in Figure 1, every regression model has one or more omitted variables, simply by virtue of the fact that the regression line does not always perfectly predict *Y*. (As seen below, these "errors" can be either positive or negative.) When an omitted variable is uncorrelated with *X*—for

■

***Hal J. Singer** is a Principal at Economists Inc. and a Senior Fellow at George Washington University's Institute for Public Policy; and **Kevin W. Caves** is a Vice President at Economists Inc.*

[1] Daniel Kahneman, Thinking Fast and Slow (2011).

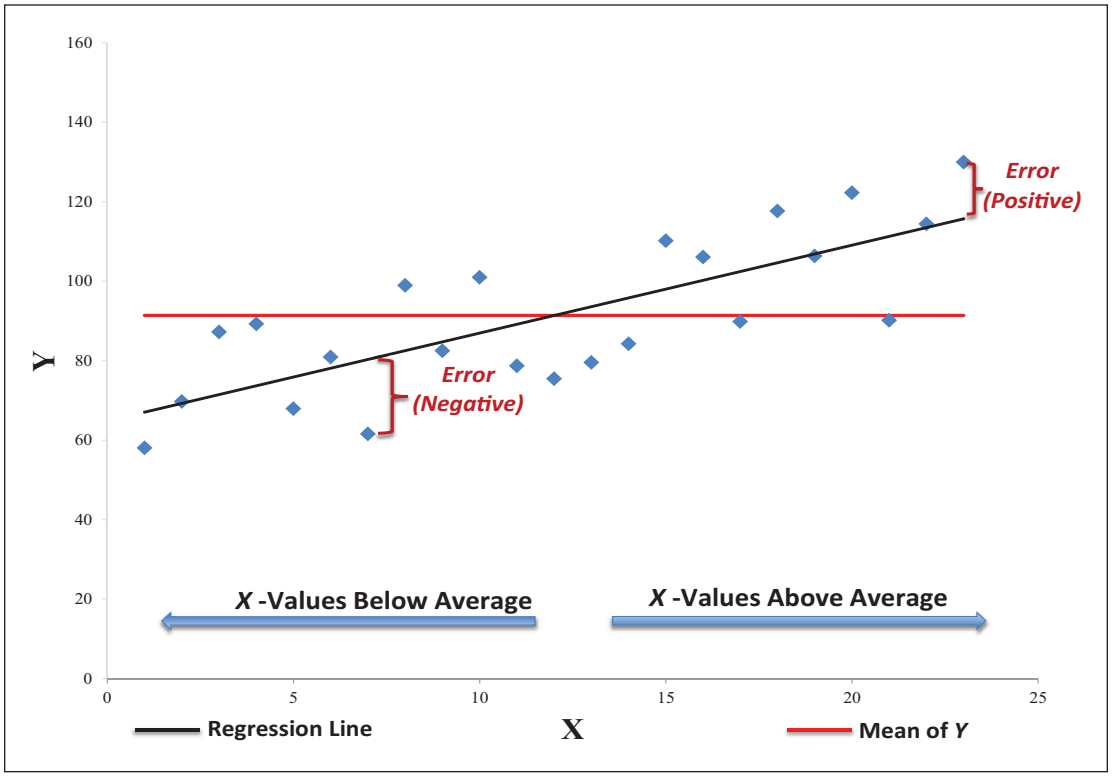

Figure 1

[O]mitted variable bias occurs when the independent variable (the X) that we have included in our model picks up the effect of some other variable that we have omitted from the model.

example, if the errors come from statistical white noise in the measurement of *Y*—then it does not present any problems. But if the omitted variable has an effect on the dependent variable (*Y*) *and* is correlated with the explanatory variable (*X*), the regression will mistakenly attribute the effects of the omitted variable to the explanatory variable, resulting in omitted variable bias.

Direction of Bias Caused by the Omitted Variable

To make matters more concrete, Figure 2 provides a stylized illustration of data points along a classic demand curve of the type covered in introductory microeconomics courses. Within this stylized example, one can see how a regression line can fit the data. The line that provides the best fit is the one that passes through the three data points. And we can see, simply by eyeballing the data, that the demand curve is indeed downward-sloping: When price goes up, the quantity demanded falls. For example, when the price is $4, the quantity demanded is 75 units; when the price is only $2, the quantity demanded increases to 125 units.

As empirical economists know all too well, real-world data are never this well behaved. Among other things, the example below assumes (naïvely) that the quantity demanded depends on the price, and *only* on the price.[2] But elementary economics teaches that the demand curve will tend to shift with changes in income (among other variables).

[2] The example also assumes (again naïvely) that price endogeneity can be ignored here. Although endogeneity (sometimes referred to as "simultaneity") is an important topic in empirical econometrics, it is outside the scope of this brief article.

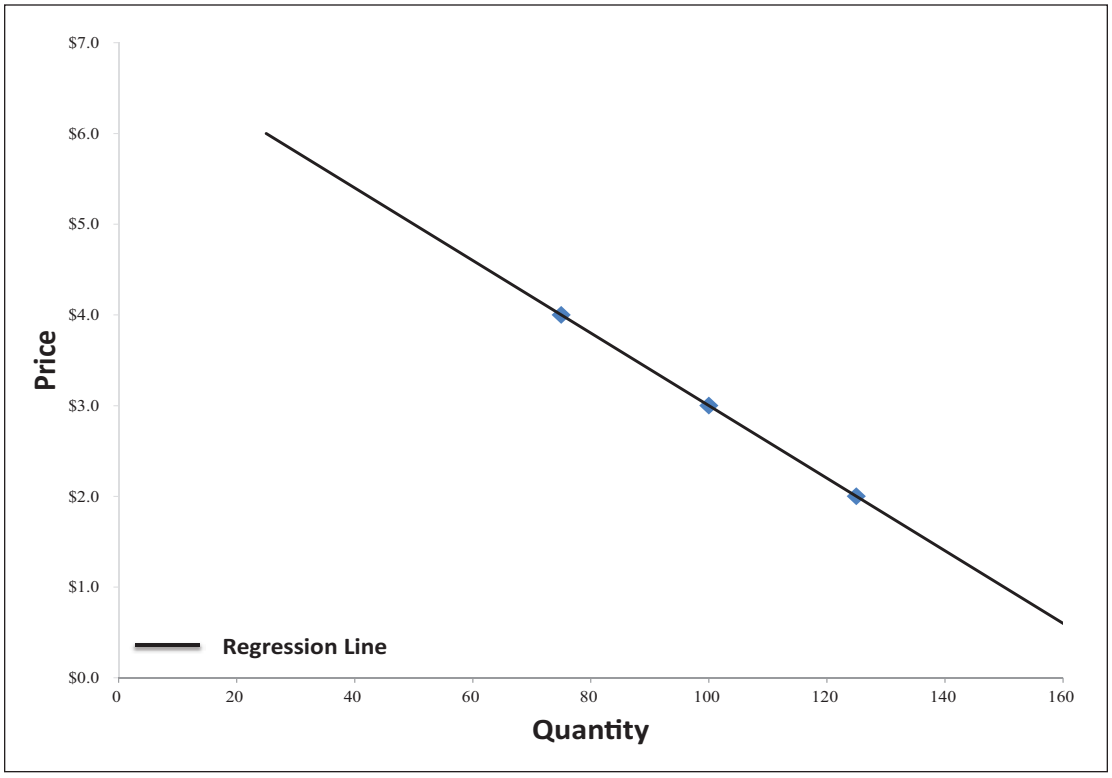

Figure 2

Let's examine a slightly more complex example, in which the data points are drawn from different areas, with varying income levels. Figure 3 displays the data:

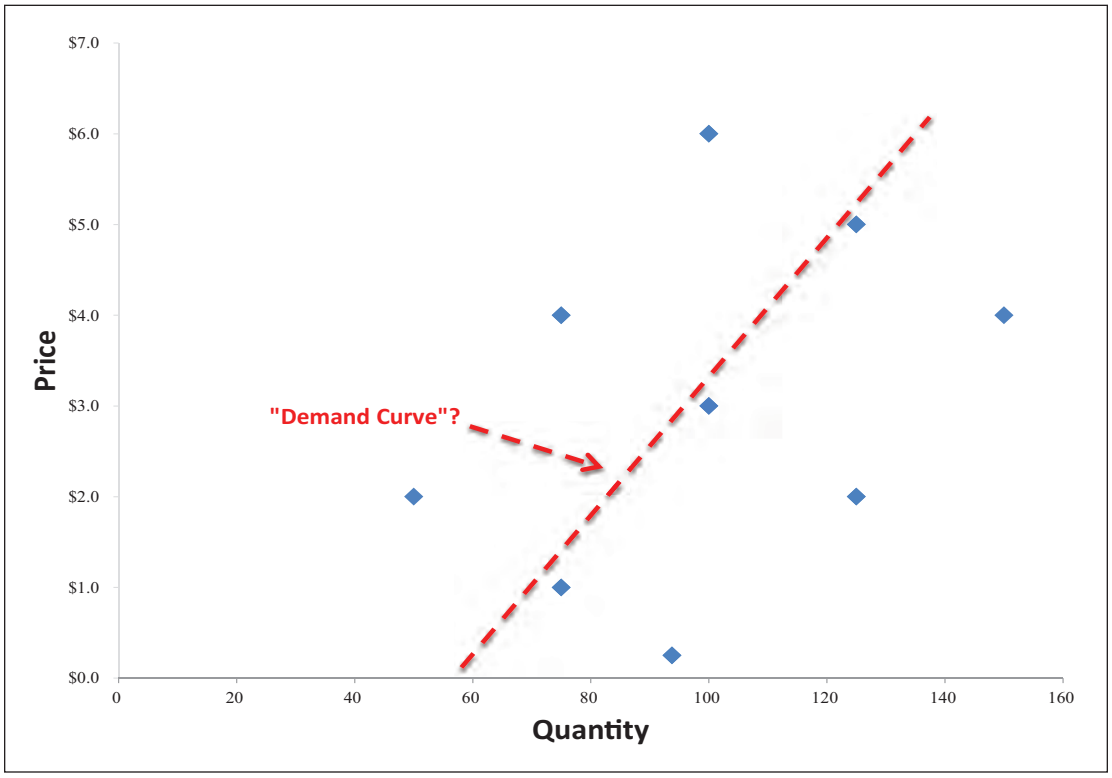

Figure 3

As seen above, we now have a cluster of data points without the downward-sloping relationship that was so readily discernible in Figure 2. If anything, a naïve interpretation of these data would seem to suggest that the quantity demanded *increases* when the price goes up. For example, the quantity demanded at a price of just $2 is only 50 units, but when the price increases to $5, the quantity demanded seemingly increases to 125 units. (This is actually possible in theory, in the case of so-called Giffen goods,[3] but these are so rare in practice that the odds that we've actually stumbled upon a Giffen good are vanishingly small).

Why do these consumers appear to demand more when the price is higher? The answer is that we have failed to control for a key variable—in this case, income. For any given price level, consumers in lower-income areas will demand less than consumers in higher-income areas. The consumers willing to purchase a quantity of 125 units at a price of $5 are wealthier than those willing to purchase a quantity of 50 units at the much lower price of $2. Figure 4 shows what happens when we include the omitted variable in our analysis, by allowing the demand curve to shift with income, and then fitting separate demand curves to the data points conditional on the income level:

What appeared to be an upward-sloping demand relationship was actually three separate demand curves . . .

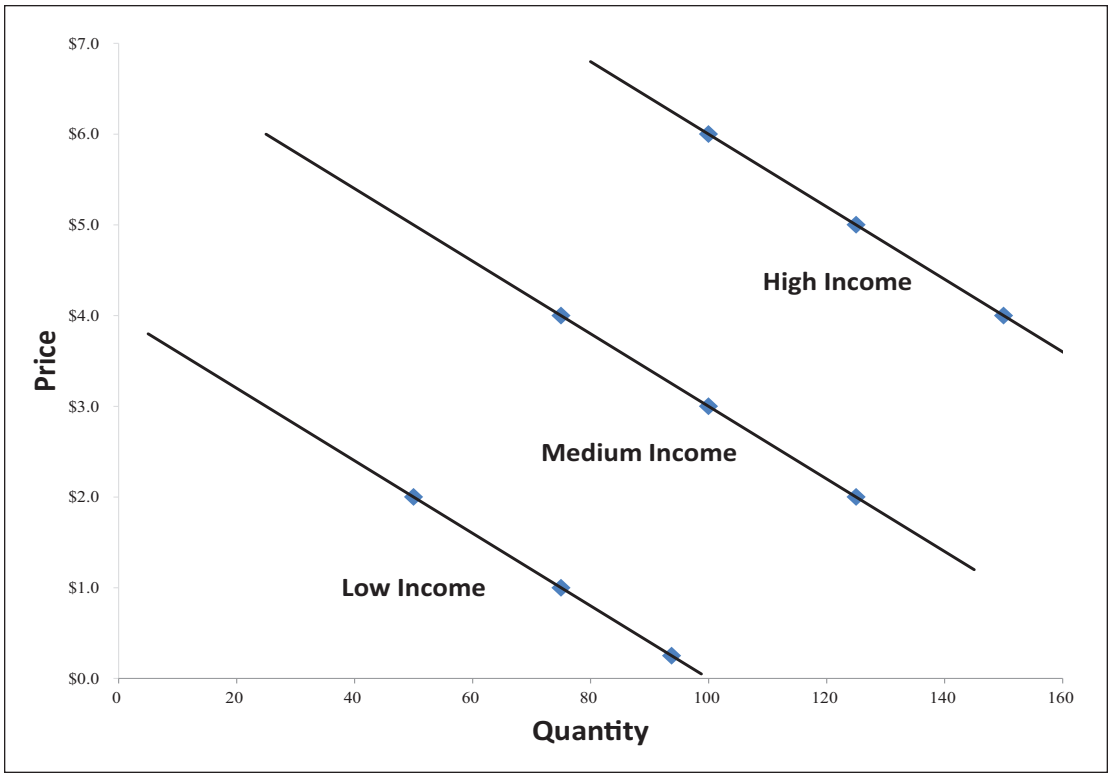

Figure 4

As seen above, once we control for income, the paradox is resolved: What appeared to be an upward-sloping demand relationship was actually three separate demand curves, corresponding to three groups of consumers with three distinct income levels.

[3] Named after Scottish economist Sir Robert Giffen, a "Giffen good" is a good for which quantity demanded increases as its price increases, rather than falls. Giffen noted that as the price of a food staple like bread rises, the poor can no longer afford to supplement their diet with better foods and must consume more of the staple food.

As the prior example illustrates, omitting a key variable can be detrimental to a regression model, making it impossible to accurately tease out even the most fundamental economic relationships in the data and resulting in biased econometric estimates. The reason that omitting income is so problematic in this example is that income is correlated with price *and* has a positive effect on quantity demanded: Consumers in higher-income areas tend to pay higher prices, and to purchase larger quantities, than do consumers in lower-income areas. If one fails to control for income, it is easy to mistake this as evidence that higher prices cause consumers to purchase more than they would otherwise, when in fact it is a change in income that leads to higher prices and higher quantity demanded. In this example, omitted variable bias leads us to conclude incorrectly that the demand curve is upward-sloping.

On the other hand, omitting income would not create omitted variable bias if income had no effect on demand or if income were uncorrelated with price (or both). If income were uncorrelated with price, then there would be no danger of attributing changes in quantity demanded caused by changes in income to changes in price. If income had no effect on demand, then movements in income would not cause movements in quantity demanded at all.

The direction of the bias introduced by an omitted variable depends on the sign of the correlation between the omitted variable and the independent variable, as well as the sign of the effect of the omitted variable on the dependent variable. Table 1 summarizes the possible biases introduced by an omitted variable. Whenever the omitted variable is positively correlated with the independent variable and has a positive effect on the dependent variable, the direction of the bias is positive: In the example above, income was positively correlated with price and had a positive effect on quantity demanded, so the direction of the bias was positive (as seen in the first row of Table 1). The positive bias was so severe that the relationship between price and quantity demanded, which should have been negative, was actually pushed into positive territory, implying (counterfactually) that the demand curve was upward-sloping. This positive bias derives from incorrectly attributing the effects of an increase in income to an increase in the price level.

Table 1

Effect of the Omitted Variable on Y	Correlation Between Omitted Variable and X	Direction of Bias
Positive	Positive	Positive
Positive	Negative	Negative
Negative	Positive	Negative
Negative	Negative	Positive

When the correlations change, so does the direction of the bias. To understand why, suppose next that income has a *negative* effect on the quantity demanded (as might be the case for inexpensive domestic cars), but that income is still positively correlated with price (as would be the case if higher-income purchasers of domestic cars were more likely to pay sticker price). In contrast to the prior example, higher incomes are now associated with lower quantities (because fewer domestic cars are sold in higher-income areas) but also with higher prices (because domestic cars tend to sell for higher prices in higher-income areas). As seen in Table 1, this means that the direction of the bias is negative.

Figure 5 illustrates these relationships. In contrast to our prior example, higher incomes are now associated with higher prices and lower quantities demanded, instead of higher prices and higher quantities demanded. As seen below, when we control for income, we again trace out three

separate demand curves corresponding to three different income levels, each of which is relatively steep. Had we failed to control for income, we would have traced out a demand curve that is too flat, as illustrated by the dashed red line. A flat demand curve means that the quantity demanded is highly sensitive to movements in price, while a steep demand curve means that the quantity demanded does not change much with the price. In this example, omitted variable bias did not change the sign of the relationship—we still find that quantity is negatively related to price. But the bias makes the relationship too negative: Quantity demanded appears overly sensitive to price when we fail to control for income.

Quantity demanded appears overly sensitive to price when we fail to control for income.

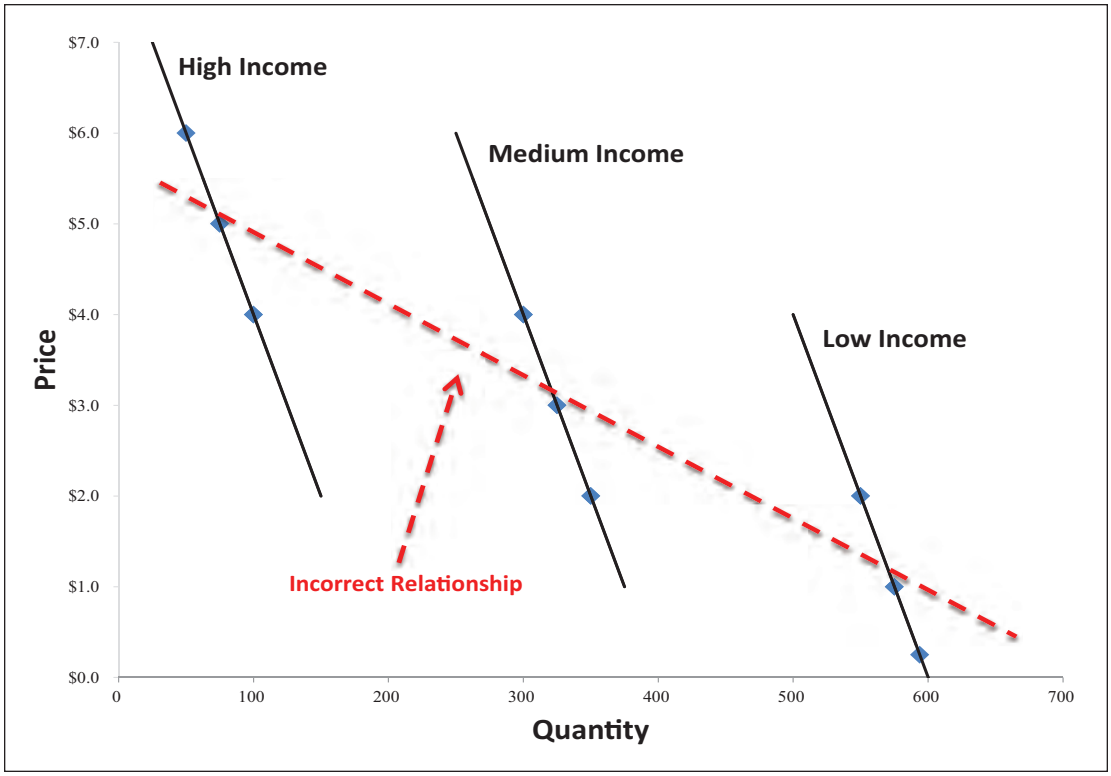

Figure 5

For example, consider what happens to price and quantity demanded when we move from a lower-income area to a higher-income area. Because higher incomes are positively related to prices, prices will increase. And because quantity demanded and income have a negative relationship, the quantity will decrease. If we fail to control for income, we will mistakenly fully attribute the decline in quantity demanded to the increase in price (a movement along the demand curve), which in reality is partially caused by the fact that higher-income consumers are less interested in purchasing domestic cars than lower-income consumers (a shift in the demand curve).

When Does Omitted Variable Bias Matter, and What Should Be Done About It?

It is easy to claim, in the abstract, that a regression has failed to account for some unspecified factor. This is because there will always be at least some movements in *Y* that *X* cannot fully account for. Omitted variable bias is therefore most effective as a methodological critique when one can (1) identify a plausible candidate for the omitted variable; (2) predict the direction of the bias based on its expected correlation with *X* and *Y*; and (3) (ideally) demonstrate this effect empirically by controlling for the omitted variable, and showing that the results change substantially.

Consider an example of a horizontal price-fixing conspiracy in which the defendants allegedly entered into an agreement as of a certain date. Suppose that the plaintiffs present a regression indicating that prices increased by 30 percent on average after the start date of the alleged conspiracy, relative to beforehand. An economist for the defense might argue that the plaintiffs' regression model suffers from omitted variable bias because the plaintiffs' economist neglected to control for changes in the defendants' costs that took place around the time of the alleged conspiracy.

Note that the direction of the bias is important here, and depends critically on whether and how the omitted variable (cost) is correlated with the challenged conduct: If costs are *positively* correlated with the conduct, then the direction of the bias is positive (as seen in the first row of Table 1), implying that the plaintiffs' model has overstated the effect of the conspiracy on prices. This would imply that the plaintiffs' regression was mistakenly attributing an observed price increase to the conspiracy, when in fact some or all of the increase was driven by higher costs. But if costs are *negatively* correlated with the conduct, then the direction of the bias is negative (as seen in the second row of Table 1), implying that the plaintiffs' model has *understated* the effect of the conspiracy on prices. (That is, but for falling costs, the conspiracy would have driven prices still higher). Finally, if costs are uncorrelated with the conduct, then omitting them from the regression model does not bias the plaintiffs' regression model.

The ideal solution would be to obtain cost data from the defendants, so that costs can be directly controlled for in the regression model. If the plaintiffs' regression still detects a positive and significant effect of the conspiracy on prices, the defense can no longer plausibly argue that the plaintiffs' estimate of the effect of the conduct is biased (unless some other omitted variable is identified). But if the plaintiffs' regression no longer shows a significant effect of the conspiracy, then the plaintiffs cannot plausibly prove liability or claim damages based on their regression model. On the other hand, it is possible that suitable cost data are unavailable. (For example, perhaps costs are aggregated in the defendants' records, with no specific information available for the products at issue.) In this case, other forms of record evidence (e.g., testimony from input suppliers) could be used to investigate the likely direction of the correlation between the omitted variable and the conduct, and thus, the likely direction of the bias.

Claims of omitted variable bias were raised by the defense in *In re High-Tech Employee Antitrust Litigation*.[4] In that case, the plaintiffs alleged that top executives at some of Silicon Valley's most prominent companies, including Apple, Google, Intel, and Adobe, conspired to restrict the recruiting and hiring of high-tech workers as a mechanism for suppressing compensation. To quantify this effect, the plaintiffs' economist used an econometric model in which the dependent variable was real annual employee compensation, and the independent variable was a measure of the challenged conduct, calculated as the proportion of months within a given year during which a given employer was subject to one or more of the anti-solicitation agreements challenged by the plaintiffs.[5] The results of the regression indicated that the compensation paid to class members was negatively related to the challenged conduct.

---

[4] No. 11-CV-2509 LHK (N.D. Cal.).

[5] The regression controlled for a variety of additional factors, including employee age, gender, years at the company, employer revenue, and the number of new hires. *See* Kevin Caves & Hal Singer, *Analyzing High-Tech Employee: The Dos and Don'ts of Proving (and Disproving) Classwide Impact in Wage Suppression Cases*, ANTITRUST SOURCE (Feb. 2015), http://www.americanbar.org/content/dam/aba/publishing/antitrust_source/feb15_caves_2_11f.pdf.

The regression also enabled the plaintiffs' economist to quantify this effect. In particular, the regression yielded a prediction of how much class member compensation would have increased, on average, if each employer had participated in anti-solicitation agreements for zero months (equivalent to eliminating the challenged conduct). This yielded an estimate of aggregate damages—calculated as the difference between what class members were actually paid and the amount that they would have been paid absent the allegedly anticompetitive agreements among employers—of approximately $3 billion.[6]

The defendants' economists argued in the abstract that the plaintiffs' regression model might suffer from omitted variable bias, which "arises when some of the same unmeasured common factors drive both the independent and dependent variables."[7] By invoking omitted variable bias, the defense was asserting that the plaintiffs' measure of the challenged conduct was correlated with some other variable, which the plaintiffs had omitted from their model, and that it was this omitted variable that was actually causing lower compensation to be paid to class members. As the court observed in its class certification order, the defense had failed to specify what the omitted variable might be, or to explain why excluding it from the model would have biased the plaintiffs' regression in the matter claimed by defendants.[8]

Controlling for the omitted variable could have reduced the plaintiffs' damages estimates, perhaps to zero.

Both omissions are important: A plausible omitted variable is, first and foremost, something that affects the dependent variable. In this context, defendants' experts would have had to offer up some factor that would be expected to have a significant effect on class member compensation, yet was not already controlled for in plaintiffs' regression model. Second, one would have to be able to plausibly claim that the omitted variable had the correct correlation with the challenged conduct. If the omitted variable were uncorrelated with the conduct, or if the correlation had the wrong sign, then the critique falls apart.

To illustrate, suppose the defense had posited an omitted variable that had a *positive* effect on class member compensation. In this case, the defense would have had to argue that the omitted variable was *negatively* correlated with the challenged conduct—so that, as the challenged conduct took effect, the omitted variable would decline, causing class member compensation to fall. If that were the case, then the plaintiffs' measure of the challenged conduct would have been picking up the effect of the omitted variable, implying that the plaintiffs' damages estimate was biased upward. Controlling for the omitted variable could have reduced the plaintiffs' damages estimates, perhaps to zero. Had the defense in *High Tech Employee* succeeded in demonstrating this empirically, it is not at all clear how the plaintiffs could have prevailed. On the other hand, if the correlation between the conduct and omitted variable were *positive*, then the omitted variable would simply have made the plaintiffs' damages estimates conservative, understating the effect of the conduct on class member compensation.

Conversely, if the defense posited an omitted variable that had a negative effect on class member compensation, the defense also would have had to argue that the omitted variable was positively correlated with the conduct—so that, as the challenged conduct took effect, the omitted variable would increase, causing class member compensation to fall. If the correlation

---

[6] The court agreed that the analysis "was capable of showing that Defendants' total expenditures on compensation [were] less than they would have been in the absence of anti-solicitation agreements and thus capable of showing classwide damages." *See* Order Granting Plaintiffs' Supplemental Motion for Class Certification at 56, *In re* High-Tech Employee Antitrust Litig., No. 11-CV-02509 (N.D. Cal. Oct. 24, 2013).

[7] *Id.* at 73.

[8] *Id.* at 74.

between the conduct and omitted variable was *negative*, this would again imply that the plaintiffs' regression model had underestimated the true effect of the challenged conduct.

Conclusion

We have explored omitted variable bias, a fundamental regression concept that frequently arises in antitrust litigation. Every regression has omitted some variables. The relevant question is whether the omission generates bias that significantly compromises the reliability of the regression model. For this to happen, it must be the case that the model has failed to control for a variable that affects *Y* and is correlated with *X*. In litigation, the direction of the bias is often highly relevant (e.g., are damages estimated conservatively, or are they over-stated?). The direction of the bias can be predicted based on how the omitted variable is correlated with the independent variable. Better yet, it can be known with certainty by controlling for the omitted variable. ●