Throw Me a Lifeline!
Regression Analysis Explained for Antitrust Practitioners

Elizabeth M. Bailey

Many years ago, I answered a phone call from an old friend who is now a lawyer. I knew he had one of his first matters headed toward litigation. He had no time for pleasantries and jumped right into the problem. He had just gotten off the phone with the economist he had retained.

“Progression!” he said to me, “My economist wants to run Progressions. What in the world is he talking about?”

“Did you say Progression?” I asked.

“Yes, Progression!”

“Is he also asking for data?” I probed.

“Yes, he says he needs a lot of data to do this Progression thing,” he sighed.

“The economist you hired said ‘Regression’ not ‘Progression’” I replied. “He wants to run Regressions.”

“OK, Regressions. But that does not change anything: What is a Regression and why is the economist I hired telling me he needs to run one?” he asked.

Similar anecdotes collected since that telephone call lead me to believe that the spirit of this conversation has played out hundreds of times across the antitrust bar.

Familiarity with regression analysis is important because it is a well-accepted scientific tool used in both merger and non-merger matters.[1] Conclusions drawn from regression analysis related to competitive dynamics are often discussed and relied upon by government antitrust agencies and in judicial opinions.[2] Regression analysis has been used to address a variety of data-driven issues that arise in competition matters, including defining product and geographic markets, identifying potential price effects, assessing commonality in class certification, and calculating damages. Given the breadth of settings in which regression analysis is used as technical evidence, it makes good sense to be familiar with its conceptual underpinnings and some of the common ways in which it is used in antitrust matters.

■
Elizabeth M. Bailey *is an economist and academic affiliate at NERA Economic Consulting in San Francisco where she handles mergers and acquisitions as well as other matters involving antitrust and competition issues. Kayvon Coffey provided excellent assistance in preparing this article.*

[1] For additional discussion of regression analysis, two non-technical references are Daniel L. Rubinfeld, *Quantitative Methods in Antitrust*, *in* Issues in Competition Law and Policy 723 (ABA Section of Antitrust Law 2008), and Daniel L. Rubinfeld, *Reference Guide on Multiple Regression*, *in* Reference Manual on Scientific Evidence 415 (Fed. Judicial Ctr., 3d ed. 2011).

[2] *See, e.g.*, FTC v. Swedish Match, 131 F. Supp. 2d 151 (D.D.C. 2000); FTC v. Whole Foods, Inc., 2007 U.S. Dist. LEXIS 61331 (D.D.C. Aug. 16, 2007); *see also* Jonathan Baker, *Econometric Analysis in* FTC v. Staples, 18 J. Pub. Pol’y & Mktg. 11 (1999).

What is Regression Analysis?

A regression is a statistical tool used to measure a precise relationship between two, and often more, factors or phenomena.[3]

If you drive to work in the morning, you know that the time it takes you to get to the office depends on how many other cars are also on the road. While it is useful to know it takes longer to drive to work when there are more cars on the road, in many situations, it is more useful to know how much longer it is going to take to get to the office. Is it five additional minutes or 30 additional minutes? A regression can be used to measure that relationship. If there are 100 more cars on the road because you leave the house an hour later than usual, a regression analysis calculates how many more minutes it is going to take, on average, to get into the office based on the increase in the number of cars.

Regression analysis starts by laying out a theoretical question that you want to measure. Following the drive to work example, suppose we are interested in knowing by how many minutes your commute time increases when the number of other cars on the road increases. Suppose we believe that a simple relationship describes how the time it takes to drive to work is related to the number of other cars on the road. The simple relationship is written as follows:

A regression is a statistical tool used to measure a precise relationship between two, and often more, factors or phenomena.

$$\text{Drive Time} = \alpha + \beta^{*}\text{Cars on the Road}$$

We call the formula used to describe this relationship the regression specification.[4] Translated into words, this regression specification tells us that we believe: (1) the time it takes to drive to work depends only on the number of other cars on the road, and (2) the time it takes to drive to work and the number of other cars on the road are related to each other linearly.

The variable on the left side, the time it takes to drive to work, is called the dependent variable because we are asking what this variable depends on. The variable on the right side, the number of other cars on the road, is called the explanatory variable[5] because we are using it to explain the dependent variable.

However, in most situations, relationships are more complex than just a simple two-variable relationship. For example, the time it takes to drive to work likely depends on more than just the number of other cars on the road. Other factors likely include weather conditions, whether there is road construction that is blocking or slowing lanes, whether a traffic accident has happened, and how many red stop lights you hit along the way. If we believe other factors also play a role in explaining the dependent variable, then it makes sense to include multiple explanatory variables in the regression analysis.[6]

To say we are going to run a regression means, in the context of our example, that we are going to use data on the time it takes to drive to work and the number of other cars on the road to obtain

[3] *See* William Greene, Econometric Analysis (3d ed. 1999) for an academic textbook reference on regression analysis and econometrics more generally. As with jelly beans, regression analysis comes in many flavors. Ordinary Least Squares (OLS) and Instrumental Variables (IV) are two of the more common regression techniques used for analyses in antitrust matters.

[4] This specification is for a simple linear regression. To generalize, such a regression specification is typically written as $Y = \alpha + \beta^{*}X$ where Y is the dependent variable, α is the intercept coefficient, β is the slope coefficient, and X is the explanatory variable. These variables and coefficients are described in detail in the text below.

[5] Sometimes the explanatory variables are referred to as independent variables. However, the use of the term independent variable implies that the variable is independent of (i.e., not related to) the other right side variables. The term explanatory variable is a more general term allowing for potential relationships among other variables in the regression specification.

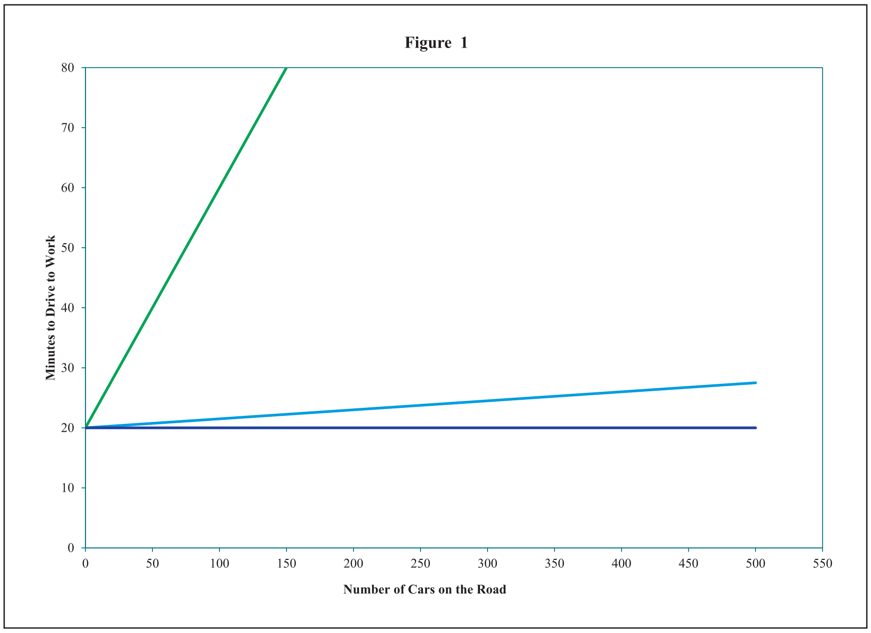
[6] We write the regression specification for a multiple regression analysis as $Y = \alpha + \beta_1{}^{*}X_1 + \beta_2{}^{*}X_2 + \beta_3{}^{*}X_3 + \ldots$. As discussed later in this article, failure to include all explanatory factors can lead to omitted variable bias.

an estimate of the values of α and β in our regression specification.[7] We call α and β the regression coefficients.[8]

The values of α and β describe what the linear regression line looks like.[9] The coefficient α, the intercept,[10] tells us what the value of the dependent variable would be if the explanatory variable were equal to zero (i.e., where the line crosses the Y-axis). In our example, the coefficient α tells us, in words, how long it would take to drive to work if there were no other cars on the road.[11]

The coefficient β, the slope, measures how the time it takes to drive to work is related to the number of other cars on the road. Specifically, the coefficient β tells us precisely by how much the time it takes to drive to work changes as the number of cars on the road changes. For example, suppose the time it takes to drive to work is measured in minutes. If we were to find β equals 0.05, then when the number of cars on the road increases by 100, the time it takes to drive to work would increase by five minutes (=0.05*100 cars). Similarly, if the number of cars on the road were to decline by 100, the time it takes to drive to work would decrease by five minutes (=0.05*-100 cars).

The coefficient β tells us whether the line describing the relationship between number of other cars on the road and the time it takes to drive to work is flat or steep. As shown in Figure 1, the blue line shows a relatively flat regression line (β close to zero) while the green line shows a relatively steep regression line (β relatively far from zero).



[7] Technically, regression analysis finds the estimates of α and β that provide the best fit to the data. In the context of a one explanatory variable regression, of all the possible lines we could draw through the pairs of data points, we want to find the line that best fits the data. One criterion that is frequently used to determine the best fit is to minimize the sum of squared errors. A regression using this criterion is referred to as ordinary least squares.

[8] In the context of ordinary least squares regression analysis, another term for regression coefficient is regression parameter.

[9] In the context of a two-variable regression, the regression line represents the two-dimensional straight line that best fits the scatterplot of data points. In a multiple regression model with three variables (i.e., two explanatory variables), the regression line that best fits the data is a three-dimensional plane.

[10] The intercept is sometimes referred to as the constant term because its value does not depend on the values taken on by the explanatory variables (i.e., it is constant regardless of the value of the explanatory variables).

[11] If Cars on the Road is equal to zero, then: Drive Time = α + β*Cars on the Road = α + β*0 = α + 0 = α.

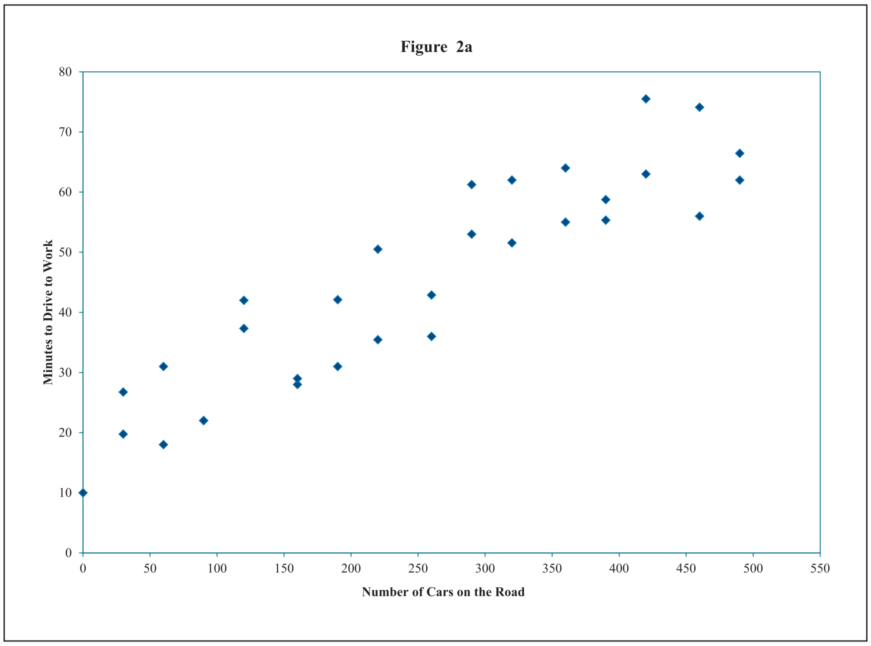
What if β equals zero? If we believe we correctly specified our regression and that the data we are using to estimate the regression line are measured appropriately, then we interpret β equal to zero as telling us that the number of cars on the road does not explain the time it takes to drive to work (see the pink line in Figure 1). That is, the time it takes to drive to work does not change no matter how many other cars are—or are not—on the road. If β equals zero, we conclude that these two variables are not related to one another.

Can Visual Inspection of a Chart Be Used Instead of Regression Analysis?

There are situations in which a simple visual inspection of a scatterplot (sometimes colloquially referred to as the eyeball test) can be extremely informative. However, there are at least two reasons why a simple visual inspection of a scatterplot is usually not sufficient to measure a relationship between two or more factors or phenomena.

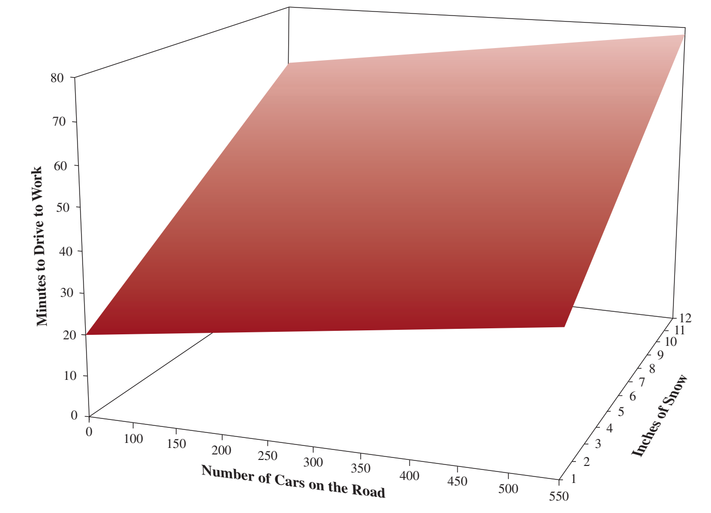
Quote . . . . .?

First, while a scatterplot chart of two variables may be able to suggest visually that the two variables appear to move together, a visual inspection cannot precisely measure that relationship. As shown in Figure 2a, through visual inspection alone, it is unlikely that we would be able to determine whether the precise relationship between the drive time to work and number of cars on the road is that every 100 additional cars on the road leads to a 10-minute increase in drive time or whether every 100 additional cars on the road leads to a 12-minute increase in drive time, and so on. In situations in which we are interested in measuring the precise relationship between two or more variables, visual inspection is generally not sufficient.



Second, even if a scatterplot chart could be used to precisely measure a relationship between two variables, in most situations, more than one factor explains the movement in another factor. When multiple factors are related, as is likely in the case of drive time to work, it is difficult to impossible to use a visual display of the data to measure those relationships. A standard two-dimensional scatterplot can display the visual relationship between number of other cars on the road and drive time to work (see Figure 2a), and a three dimensional display may be possible—though often cumbersome—to view the relationship between number of other cars on the road, weather conditions, and drive time to work (see Figure 2b). However, a visual display of the rela-

Figure 2b



tionship between more than three factors is difficult to impossible because four or more dimensions would need to be condensed into two (or three) dimensions to be readily visualized. Examples include the addition of road construction blocking traffic lanes, severity of a traffic accident, and so on.

What Are Some of the Ways Antitrust Practitioners Use Regression Analysis?

Regression analysis is used to provide data-driven evidence for a wide variety of questions that arise in merger and non-merger antitrust matters. While far from an exhaustive list of all of the settings in which regression analysis has been used, common settings include assessing the scope of the relevant geographic or product market, estimating the magnitude of elasticities of demand, predicting potential price effects from a merger, and estimating overcharges in matters involving damage claims.

For example, whether a potential restraint, such as a joint venture or a refusal to deal, gives rise to competitive harm often turns on the scope of the relevant geographic or product market. Regression analysis can be used to inform the proper relevant market. For example, a central question may be whether the relevant geographic market is limited to Region A (e.g., the United States) or is properly expanded to include Region B (e.g., Asia). A multiple regression can be used to identify the magnitude—and speed—of an import response from Region B into Region A as a result of a relative price change between the two regions, controlling for other factors. Data-driven evidence on the magnitude of the increase in imports in response to a relative price increase informs the scope of the relevant geographic market. In general, the larger the response and the quicker the response, the more likely Region B is properly included in the relevant geographic market around Region A.

Likewise, regression analysis is frequently used to estimate a price elasticity of demand, which also can be used to inform the scope of the relevant market. The price elasticity of demand measures how responsive quantity demanded is to changes in price, everything else held equal. A multiple regression can be used to estimate the price elasticity of demand, with quantity as the dependent variable, price as an explanatory factor, and additional explanatory variables included to control for the other factors that also would be expected to affect quantity demanded, such as the price of a potential substitute product as well as demographic factors.

Additionally, the key analytical exercise in evaluating a proposed merger between two firms is to predict the potential price effect resulting from the merger. If the products sold by the two firms are particularly close or unique in the eyes of consumers, then the merger may be anticompetitive. On the other hand, if rival firms provide strong competitive constraints on the two firms because consumers view those products as equivalent substitutes, then the merger is unlikely to be anticompetitive. Multiple regression analysis can assess the extent to which consumers consider other firms' products close competitive alternatives to those offered by the parties.

One way to do this is to estimate by how much one of the merging party's (Firm A) store-level sales were affected by entry from the other party to the proposed merger (Firm B) compared to entry by third-party rival firms (Firms W, X, Y, and Z). If the results of the regression analysis indicate that the magnitude of the hit to sales from entry by each of these rival firms is similar, then the results are likely to support the argument that the parties' products are not particularly unique or special relative to products available from rivals in the marketplace. On the other hand, if the magnitude of the hit to sales from entry by the other party to the proposed merger is large and the magnitude of the hit to sales from entry by third-party rivals is substantially smaller, then the evidence from the regression analysis may be more consistent with rivals providing only a limited competitive constraint.

Sound regression analysis does not take a spaghetti-on-the-wall approach, in which factors are included through a trial-and-error method . . .

Last, regression analysis is also frequently used in matters involving potential damages. Regression analysis can be used to estimate the magnitude of the damages, if any, by comparing prices before, during, and/or after an alleged event, such as a price-fixing agreement, took place. To determine by how much prices increased as the result of an alleged agreement, multiple regression analysis, with price as the dependent variable, is often used. The regression specification would typically include an "on"/"off" indicator variable for the period during which the agreement was alleged to be in effect ("on"), as well as additional explanatory variables to control for the other supply and demand factors on which price would be expected to depend, such as input costs and macroeconomic conditions.

How Can Regression Analysis Mislead?

While regression analysis is a powerful scientific tool to use to measure the relationship between two or more variables, regression analysis can also create confusion. Regression analysis can be misused and the coefficient estimates can be misinterpreted, rendering results and inferences unreliable.

Although regression analysis does not need to be as flawless as science performed in a controlled laboratory setting in order to be of value, it is important to know where to look for the flaws and the blemishes. The larger the flaws and the blemishes, the less reliable the results of the regression analysis will be to the fact finder. While regression analysis can mislead for many reasons, the most common in my experience fall into three categories.

Untethered from Common Sense. Sound regression analysis does not take a spaghetti-on-the-wall approach, in which factors are included through a trial-and-error method in order to choose the relationship that appears to work out most favorably. Sound use of regression analysis relates the specific facts of the matter at hand to the choice of regression specification. Regression analysis should not be undertaken in a vacuum, untethered to the non-statistical information also available in the matter. The choice of regression specification and the factors that are expected to matter should be grounded in—and consistent with—available documentary evidence, deposition testimony from relevant business personnel, and/or sensible economic theory.

Correlation Is Not Causality. Although regression analysis establishes a correlation between two variables, regression analysis does not typically establish causality between those two variables. Economic theory and common sense, not linear regression analysis, tell us whether the first factor causes the second factor, whether the second factor causes the first factor, or whether the causality runs in both directions. In fact, it is even possible for two factors to appear related to one another but have no meaningful relationship to one another or any causal relationship whatsoever.[12]

For example, imagine a research study that found that people who sleep a large number of hours of the day are also people who are more likely to be sick. That is, there is a positive correlation between hours spent sleeping and being sick. Now, while there may well be a relationship between the two variables, it would take a large leap of the imagination to conclude that sleeping more hours is what causes people to be sick. It is much more plausible that the causality runs the other direction: being sick is what causes those same people to sleep more hours of the day.

In antitrust matters, relationships often have causality running in both directions. For example, it can be informative to use regression analysis to estimate an elasticity of demand. However, in doing so, it is important to take into account that the relationship between price and quantity is bidirectional. In particular, through the demand-side relationship, quantity demanded is usually negatively related to price (because a lower price increases consumer demand for the product). At the same time, through the supply-side relationship, price is often positively related to the quantity supplied (because, after some point, a greater volume supplied increases production costs and thus the price, of the product).

A Duck Is Not Always a Duck (Multi-Collinearity, Omitted Variables, and Measurement Error). In regression analysis, the coefficient on each of the explanatory variables, β, is interpreted as providing the precise measure of the relationship between that explanatory variable and the dependent variable. For example, if the explanatory variable is labeled Other Cars on the Road, then the coefficient on this variable is interpreted as the change in the dependent variable when the number of other cars on the road changes. This interpretation is correct if the regression is correctly specified and the data used to run the regression are available and appropriately measured.

However, there are instances in which, just because a factor is identified as representing the number of other cars on the road (a duck, so to speak), that coefficient need not be correctly interpreted as the contribution attributed to the Other Cars on the Road factor. The potential for multicollinearity, omitted variables, and measurement error are three reasons why a coefficient may not be correctly interpreted as the factor appears to be labeled.

Multicollinearity arises when two, or more, explanatory factors are highly related to each other.[13] For example, the listing price of a house and the ultimate sale price of a house are likely to be highly collinear. A higher listing price will go hand in hand with a higher sales price. Similarly, crude oil prices and gasoline prices at the pump are likely to be highly collinear. It is difficult to identify the true coefficient estimates for variables that are highly collinear with one another because the separate effect from one explanatory factor compared to the other explanatory fac-

[12] "Spurious correlation" is the term used when two variables appear related but in fact have no meaningful relationship to one another. In addition to simple coincidence, another reason two variables may appear related to one another but in fact have no causal relationship to each other is because a common third factor is driving both variables to move in a similar way. For example, suppose we observe rising housing prices in San Francisco and rising housing prices in Washington, DC. While the two series of data would be positively related to each other, it is unlikely that one is causing the other. Rather, a third factor, such as common changing preferences toward city-living and/or job growth common to both cities, may be the underlying factor driving the increase in both price series.

[13] In the extreme, two explanatory factors can be perfectly collinear. As an example, if a variable takes on one of two values, e.g., On/Off, then including an explanatory variable for On and an explanatory variable for Off in a regression specification would result in perfect collinearity because the data that comprise these two variables would be the exact opposite of one other. That is, knowing the values taken on by the On variable for each observation would allow the values taken on by the Off variable to be known perfectly.

tor cannot be disentangled. Factors that are collinear are not typically estimated with precision. The estimated coefficients may have the wrong sign (e.g., negative instead of positive), unreasonable magnitude, and/or appear to have no relationship with the dependent variable.[14]

Multicollinearity can be a confounding factor in antitrust matters. As an example, consider a proposed merger between two brick-and-mortar retailers. In evaluating the likely competitive effect of the proposed merger, it is often useful to understand how, if at all, the prices for the parties' products depend on the presence or absence of other nearby brick-and-mortar retailers that sell similar products. A common regression specification is to consider prices for particular stock-keeping-units (SKU) from one of the merging parties as the dependent variable and a series of explanatory variables reflecting the presence (or absence) of other potentially price-constraining brick-and-mortar retailers in geographic proximity (e.g., located in the same shopping center as the merging party, located within a one-mile distance band of the merging party, or located within a five-mile distance band of the merging party).

If two of the brick-and-mortar retailers used as explanatory factors are typically co-located together in the same shopping center, then the physical location of these two retailers will be highly collinear, making it hard to disentangle the separate effect from each of the two co-located retailers on the party's SKU prices. The difficulty here comes from the fact that the same signal (the co-location) is being used to tease out two potentially different effects (the two rival retailers). In this case, identifying which of the two co-located rival retailers is a more important factor in constraining the party's SKU prices may only be possible with the addition of non-statistical evidence from business documents and/or testimony from business personnel.

Omission of one or more relevant explanatory variables in a regression specification can happen for several reasons. The regression may be incorrectly specified in the sense that an explanatory factor that is expected to help explain the dependent variable is excluded from the regression specification. Alternatively, while the regression may be correctly specified, data may not be available for one or more factors or phenomena that are expected to matter. If there are omitted variables, the effects of the omitted variables are typically soaked up by the explanatory variables that are included. As a result, the magnitude of the estimated coefficients for the included explanatory variables may be too big or too small compared to their "true" values because some of the effect of the omitted variable is attributed (incorrectly) to the included explanatory variables.

For example, in transactions in which products are sold through a bid process, regression analysis is often used to measure the relationship between explanatory factors, such as the identities of rival bidders, and the price bid by a particular firm. While sometimes difficult to measure, a firm's backlog (that is, the firm's accumulation of unfinished products in the queue to be supplied to customers) can be one factor that contributes to the explanation of a firm's bid price on future products to be supplied. In some situations, the greater the firm's current backlog, the higher the firm's bid price for sale and delivery of products in the future. If documents or testimony suggests that backlog may be a factor that explains bid prices, then the failure to include backlog as an explanatory factor may result in the coefficients for included variables of interest, such as the identity of rival bidders, being too big or too small compared to their true contribution.

Measurement error occurs when the data used for one or more factors in the regression are imprecisely measured relative to how that factor is described in the regression specification. For

[14] With multicollinearity, the coefficients for the two (or more) explanatory factors may not be individually statistically different from zero, but taken together, have joint statistical significance.

example, it may be difficult to accurately measure the number of other cars on the road because we observe the count of cars passing through only one particular intersection rather than along the entire route of the commute. Similarly, it may be difficult to accurately measure road conditions. For example, we may observe only inches of snow accumulated, not the more relevant factor of how well-plowed the roads remain during the snowstorm. In the presence of measurement error, the estimated coefficients may be too big or too small compared to their true values had the data been measured without error.

Conclusion

Regression analysis is well-accepted scientific evidence relied on in merger and non-merger antitrust matters. Combined with other non-statistical evidence, such as documents and testimony, technical evidence from regression analysis can be a powerful tool for measuring the relationship between two, or more, factors or phenomena.

Regression analysis, however, should not be a substitute for careful and critical thinking about the regression specification, the data being used, and the results being generated. Regression analysis should not replace common-sense thinking that is tethered to the specific facts at hand. If garbage goes in regression analysis, whether in the form of a poor regression specification or poor data availability, then garbage will come out and create confusion. This will render the inferences from the regression analysis output unreliable. Regression analysis does not need to be flawless in order to be of value, but it is important to know where to look for the blemishes in order to evaluate the reliability of the results. The scientific and technical nature of the evidence generated from regression analysis requires lawyers, economists, expert witnesses, and fact finders to be careful consumers of regression analysis. ●