An Econometrics Primer for Lawyers

ROY J. EPSTEIN

ECONOMETRICS IS WIDELY USED in modern antitrust analysis to help assess the empirical support for alternative theories of competitive effects. Typical examples include measurement of the overcharge due to an alleged price-fixing conspiracy and determination of the extent of common impact in class certification. A basic understanding of econometrics and its pitfalls can provide a lawyer with vital advantages in many key tasks: deposing an expert, understanding and presenting one's own witness, and drafting persuasive briefs and motions. This article explains the essential principles of regression analysis, the fundamental econometric procedure, from a practical and non-technical point of view.[1]

Lawyers already have the necessary skills to analyze antitrust theories using documentary and testimonial evidence. At bottom, an econometric model is a different method to test theories about economic effects by using a particular set of facts (data) and assumptions (economic principles). Fortunately, the ability to understand an econometric model does not require advanced mathematics. What a lawyer needs is guidance in identifying the important features in a model that drive the results. These are generally down-to-earth and even intuitive issues. With a modest investment in learning some basic terminology and concepts, most litigators will be able to take on econometric evidence with greater confidence.

In practice, regression analysis entails a considerable amount of judgment by the econometrician. Tests of "statistical significance," while important, are only a small part of assessing the overall reliability of the conclusions. Econometric results can depend greatly on other factors, such as the selection of the variables in the model, the quality of the data used, and methods used to calibrate the model to the data. A lawyer's goal should be to learn how to investigate all of these elements of an econometric model.

Econometrics 101

The most common type of econometric model to test hypotheses about causation and economic effects is called "ordinary least squares multiple regression," or simply "regression" when the context is clear. This technique relates an outcome, the "dependent" variable, to one or more hypothesized causal factors, referred to as "explanatory" or "independent" variables. It is conceptually similar to plotting the data points and drawing the straight line that comes closest to connecting them. The result is basically a formula that quantifies the relationship between each explanatory variable and the outcome in terms of regression "coefficients."

An example is a regression to predict bushels per acre of wheat (dependent variable) based on annual inches of rainfall and tons of fertilizer per acre (explanatory variables) for a sample of farms. Suppose the result is

$$\text{Bushels} = 0.5 \times \text{Rainfall} + 12 \times \text{Fertilizer} + 4.$$

The regression coefficients are 0.5 for rainfall and 12 for fertilizer. The model indicates that each additional inch of rain causes yields to increase by half a bushel, while each additional ton of fertilizer causes yields to increase by 12 bushels. In addition, the regression computes a term called the "intercept" or "constant" that summarizes the effect of all other influences on the dependent variable that are not explicitly represented in the model. In this example, the intercept has the value 4 and can be interpreted as the expected yield if rainfall and fertilizer are both zero.

The coefficients, the explanatory variables, and the intercept together predict the dependent variable. For example, if 20 inches of rain are expected for the season and the plan is to apply 3 tons of fertilizer, the predicted yield is 50 bushels per acre (0.5 times 20 plus 12 times 3 plus 4).

Regression can model a surprising variety of potential causal factors. In addition to numerical data such as inches of rain, it can analyze the relationship between the dependent variable and qualitative factors that have a purely yes/no or on/off nature. Qualitative variables have many uses in antitrust to control for such factors as location (e.g., East Coast), supplier (e.g., firm A), customer, product type, and so forth. They can also represent the period when an alleged conspiracy was "on," as opposed to a pre- or post-conspiracy "off" period. Because each of these variables implicitly separates the data into two distinct categories (e.g., conspiracy vs. non-conspiracy), they are referred to as "categorical" or "binary" variables or simply as "dummy" variables.

For example, suppose a price-fixing conspiracy is alleged from 2001 to 2004 and data are available for the years 1990 to 2007. The model is intended to test whether prices were elevated during the conspiracy period in order to measure damages. The dependent variable is price, measured as $ per

Roy J. Epstein consults as a damages expert with a specialty in econometrics and is also adjunct professor of finance at Boston College Carroll School of Management. He is the author of the book, A History of Econometrics. *He thanks the editors, Andrew Goodrich, Judy Roberts, Danny Shoag, and especially Fred Knapp for helpful comments.*

pound. The dummy variable would be constructed as the value 1 ("on") for the four conspiracy years, and 0 ("off") for the other time periods. This would be used as an explanatory variable in addition to any other supply and demand factors in the model. The estimated coefficient on the dummy could be interpreted as the average difference in prices when the conspiracy was on.[2] If the estimated coefficient had the value 0.11, for example, this could be evidence of an average overcharge equal to $0.11 per pound.[3]

A vital question about any regression analysis is the reliability of the results. Each estimated regression coefficient is subject to uncertainty for a variety of reasons stemming from the underlying model and data. The true coefficient may therefore differ from the value estimated by the regression. A fundamental question in this context is whether the true coefficient might actually be zero, regardless of the regression result. If the true coefficient were zero, the associated explanatory variable would have no effect on the outcome.

The starting point for assessing reliability is determining if the estimated regression coefficients meet the test of "statistical significance." Statistical significance means the true coefficient is unlikely to be zero—i.e., there is reasonable evidence of an effect—even if the regression estimate is imprecise. Statistical significance is thus essential for evaluating econometric evidence.

The idea of statistical significance originated with games of chance. Suppose a coin is flipped three times and comes up heads each time. There would be no basis to conclude the coin is crooked because this outcome is also reasonably likely for a fair coin. This corresponds to an estimated regression coefficient that is not statistically significant—the result could be due simply to chance. But getting 10 heads in a row is a different story, which would justify the opposite conclusion for the coin. Similarly, a statistically significant coefficient can be interpreted as strong evidence of an effect: according to the model, it is too improbable for the effect to be zero.

Statistical significance is conventionally tested using a standard calculation called a "95 percent confidence interval."[4] This interval, included in the regression output, determines a range of values above and below the estimated regression coefficient. If the range includes zero, this indicates a non-trivial possibility that the true effect of the explanatory variable is zero and the coefficient is not considered to be statistically significant. On the other hand, a coefficient is statistically significant when the interval does not include zero—the estimation results reject a conclusion of no effect. When the 95 percent confidence interval indicates statistical significance, an equivalent statement is that the result is significant at the "5% level." Both terminologies are in common use.

In the wheat yield example, suppose the coefficient of 12 on fertilizer is statistically significant but the coefficient of 0.5 on rainfall is not—i.e., the rainfall confidence interval includes zero. The results would support the inference that fertilizer, but not rainfall, is a causal factor for the yield. Even though one might think rainfall should play a role, the regression model (meaning the variables chosen and the data used) does not provide sufficient evidence to support the conclusion that there is an effect.

A different test of statistical significance called an "F test" (also built into modern regression software) is required in a particular situation involving models with two or more explanatory variables.[5] The F test is used to examine the joint statistical significance of two or more regression coefficients when each of those coefficients individually fails to achieve statistical significance using confidence intervals. If the F test indicates joint significance, there likely is an effect from one or more explanatory variables in the group but the data do not identify them more specifically (a condition known as "collinear" data). A negative test result, however, would indicate no joint significance and would finally allow the inference of no effect from any of the nominally insignificant variables.

The wheat yield regression can again illustrate. Suppose the coefficients for rainfall and fertilizer were not significant according to their separate 95 percent confidence intervals. A joint effect could still exist if, for example, the data came from a poorly designed experiment in which the farms that used the most fertilizer also received the most rainfall. Those farms could enjoy higher yields as a result but the regression may not accurately sort out the increase corresponding to one factor versus the other. The result could be wide confidence intervals for the individual coefficients that both include zero even though there is a reliable effect on yield from the two explanatory variables considered jointly.

Tests of statistical significance help guard against spurious regression results. The usual practice is to retain the explanatory variables with statistically significant coefficients but to rerun the regression to exclude the variables with insignificant estimated coefficients.

Significance tests are not infallible, however. Problems such as poor quality or quantity of data may prevent the regression from showing statistical significance for explanatory variables thought to be relevant (a "false negative"). More data in the crop yield example might show a significant effect for rainfall after all. On the other hand, a confidence interval can indicate statistical significance when the true coefficient is zero (a "false positive").[6] A false positive can arise by chance, like a fair coin that actually does come up heads 10 times in a row. The risk of a false negative or positive is a fact of life in any empirical analysis and cannot be eliminated.

The larger concern is that significance tests are not sufficient to establish reliability of a regression. Each side in litigation can be counted on to report favorable regression analyses. When the plaintiff's model implies large and statistically significant damages and the defendant's model indicates small and statistically insignificant damages, it is not because one side is using defective statistical software. Instead, the conflicting results arise from differences in the modeling assumptions made by the econometricians. The main assumptions

involve the choice of variables, the data, and the estimation procedure used to obtain the coefficients.

A full assessment of econometric reliability requires investigating the ability of a claimed result to survive reasonable changes in the modeling approach. If an estimated coefficient is consistently statistically significant—or not—in a range of alternative models that sufficiently test the key assumptions, one can have much greater faith that a valid result has been obtained. Probing these alternatives is where a lawyer needs to focus.

Choice of Variables

The first step in analyzing reliability is to review the list of explanatory variables to see what potential causal factors are accounted for in the analysis. A model that omits a relevant explanatory variable can result in estimated coefficients that are wildly off the mark. That, in turn, can distort tests of statistical significance for the variables that are included. This problem is known as "omitted variables bias."

In the textbook world, the form of a model is known in advance and a single regression is sufficient to complete the analysis. But a practicing econometrician uses a large cutting room floor. Because there is seldom consensus on the most appropriate specification of the model, the norm is to experiment with different combinations of variables, data, and estimation procedures. Typically an econometrician reports results for only a fraction of the different models tested. Cherry picking favorable results is one source of omitted variables bias.

Assessing the importance of omitted variables bias depends on economic theory as well as the availability of additional explanatory variables that could be relevant to the issue under consideration. An estimated coefficient may not have a plausible algebraic sign or magnitude. In common terms, the results may strain credibility. For example, a model that implies that higher input costs are related to lower prices is not economically sensible and could be biased due to omitted variables. If additional variables (i) have statistically significant and economically plausible coefficients and (ii) change a previous conclusion (e.g., regarding the statistical significance of an estimated overcharge), potential omitted variables bias should be a genuine concern.

Returning to the wheat yield example, suppose the data come from farms in North Dakota and Texas. The model should probably also include temperature as an additional explanatory variable. It could well turn out that with this revision all three variables—rainfall, fertilizer, and temperature—would be found to have statistically significant coefficients with plausible magnitudes. This result would indicate the initial model was unreliable due to the omitted temperature variable.

Conversely, analyses are sometimes presented with variables that are not statistically significant. The model in this case is said to contain "extraneous" variables. Generally, it is considered to be poor econometric practice to include such variables in a model. At a minimum, it would be important to know if re-estimating the model without the extraneous variables affects the conclusions.

The possibility of omitted variables bias and the presence of extraneous variables do not in themselves indicate that a model generates inaccurate results. Criticisms should be put to the test by suitably revising the model at issue to see if the conclusions are sufficiently robust to withstand such changes.

A case involving allegations of price fixing illustrates a number of these principles. The Justice Department had sued the Ivy League, MIT, and several other schools, alleging an illegal restraint of trade involving financial aid awards.[7] Among other defenses, MIT believed the challenged aid policies were purely redistributive and had no effect on average tuition charged.[8] The author developed an econometric model to test these propositions as a member of the economic expert team assisting MIT.

The model used the average tuition charged by each school in a large national sample as the dependent variable.[9] The explanatory variables included a dummy variable (1 if defendant school, otherwise 0). The regression coefficient on this variable measured the average difference in tuitions between the defendants and the other schools.

But the model also included many other explanatory variables to control for different characteristics of the schools. For example, a dummy variable (1 if a state college or university, otherwise 0) was included on the hypothesis that a publicly supported school will charge a lower average price than a comparable private university. Other variables to represent school quality were included to control for the possibility that higher quality schools charge higher average tuitions. Econometrically, leaving out school characteristics could overstate the estimated coefficient for the defendant dummy variable due to omitted variables bias.

The simple comparison based on the defendant/nondefendant categories did show higher prices. But MIT argued that the challenged conduct had no statistically significant effects after accounting for the other factors represented in the model. Moreover, according to MIT, similar results were obtained in virtually all of the alternative models estimated, which reinforced the overall reliability of the analysis.

The Data

Econometric results can be sensitive to at least three kinds of problems with the underlying data: "outlier" observations that may not be relevant; errors in constructing the variables; and erroneous assumptions by the econometrician. Scrutinizing the data is a humble but important part of quality control in a model.

An outlier is a value for a variable that is far from the bulk of the data. In the wheat yield model described above, suppose the rainfall observations were all in the range of 15 to 40 inches with the exception of a single observation of 100 inches. Perhaps the 100 inch observation reflects an error or, if accurate, a set of growing conditions that are not relevant and

comparable to the rest of the data. Even a single extreme value can strongly influence a least squares regression. It would be important to test whether the estimated coefficients and confidence intervals are sensitive to excluding the outlier observation.[10]

As an initial screen to identify outliers, it is a good practice to include a table in the analysis that shows the average, median, standard deviation, and minimum and maximum values for each variable used in the regression. An outlier can also be detected visually by plotting the values for an explanatory variable against those for the dependent variable.

The increased utilization of micro-level information in antitrust econometrics, such as invoice databases and survey data, also poses new challenges for reliable modeling. A great deal of care is necessary, especially when modeling transaction prices. Rebates, which may be recorded only periodically and in separate accounting records, need to be credited back to the underlying gross prices. Sales to customers outside the relevant geographic market should be excluded. Sales must also be limited to the relevant product market, which may exclude specialty items, off-quality goods, and samples. It can also be important to identify the terms of sale, such as purchases under long-term contracts and purchases made under "winner take all" competitive bidding, to determine comparability of the price data.

Aggregating invoice-level data up to broader average prices to infer price trends can also introduce problems of interpretation. Suppose there is a basic product A with a price of $1 and sales of 100 units and a premium product B with a price of $2 and sales of 100 units. The average price is $1.50 (total revenue of $300 divided by 200 total units). In the next period, suppose the prices of A and B remain the same but sales of A fall to 80 units and sales of B rise to 120 units. The average price would rise to $1.60 (total revenue of $320 divided by 200 units). A regression showing a "statistically significant" increase in average prices between the two periods may be an artifact of changes in sales mix, with no implication for competitive conduct.

When using transactions data for different firms in the same model, it is common to include dummy variables to control for firm-specific effects that cannot be identified more precisely. For example, firm A may consistently charge higher prices because it offers a better warranty. But when firm-specific effects are found, it is also important to check that the estimated regression coefficients for the other explanatory variables are the same across the individual firms. More generally, econometric tests of "structural change" can be used to test the validity of "pooling" potentially non-comparable data and variables in a single regression.

The definition of the damages period dummy can have a major effect in an econometric model. In a recently settled conspiracy case, the plaintiffs' model for price per pound defined the damages period as the 72 months from January 1995 through December 2000 and included a corresponding dummy variable (1 for the damages period, otherwise 0) as well as other variables. The estimated regression coefficient for the damages dummy was positive and statistically significant, which the plaintiffs interpreted as proof of the per pound overcharge.

Yet other facts in the case indicated uncertainty about the precise timing of the alleged conspiracy. The defendants analyzed two different time periods with a comparable model to address the timing issue. The first one extended the damages period for another 3 months, until March 2001. The second one considered damages that commenced in January 1997 instead of 1995. In both of the defendants' alternative scenarios, the regression coefficient associated with the estimated overcharge not only failed to achieve statistical significance, but even became negative, which according to the defendants showed no antitrust impact.

Constructing Valid Significance Tests

Significance tests can be unreliable when improper data are used or relevant variables are omitted from a model. They can still be seriously misleading even without these problems. In brief, there are two principal checks on a regression that signal when the model has not been formulated correctly and may be generating invalid significance tests. Modern regression software is equipped to perform these checks and make appropriate corrections to the model.

First is a check for "serial correlation" or "autocorrelation." This problem is particularly relevant for models that use time-series data, i.e., where the observations come from different time periods. Statistically, serial correlation usually causes a regression to overstate the precision of the coefficient estimates, which biases confidence intervals to be too narrow. This bias can lead to false conclusions about statistical significance.

As a simplified example, suppose a pollster surveys 1000 people for their political views by interviewing 500 couples. The statistical ideal is to choose the different observations at random. If, however, couples tend to share political views (i.e., their views are correlated), the sample would not provide as much information as 1000 people chosen at random, and the poll would have lower accuracy. The standard confidence interval would assume the pollster is using 1000 random observations. But because of the correlation, that confidence interval would be biased. The correct interval would be wider. Similarly, the regression procedure would require an adjustment when serial correlation is present in order to generate appropriate confidence intervals that reflect the true amount of information in the data.

Second is a check for "heteroscedasticity." In this situation, certain data points are statistically more informative than others, which also leads a least squares regression analysis to produce inaccurate confidence intervals. The necessary adjustment entails giving the more informative observations more weight in the regression analysis.

Panel data models, which combine time-series and cross-section data, can be subject to combined effects of serial cor-

relation and heteroscedasticity. Ordinary least squares can also be adjusted to correct this more complex situation.

In a recently settled case, the plaintiff's expert presented a model that pointed to damages in excess of $300 million. The defendant's expert argued that the model suffered from undiagnosed and massive serial correlation. When the defendant's expert re-estimated the model with a correction for serial correlation, the regression coefficient for the overcharge was not statistically significant, supporting the defendant's argument that a more appropriate conclusion was zero damages.

In other situations, a heteroscedasticity correction can materially alter the conclusions from a model. An estimated regression coefficient may switch from being statistically significant to not significant, and vice versa.

A model that fails to check for problems due to serial correlation and heteroscedasticity is seriously deficient and does not pass a threshold test for reliability of confidence intervals and significance tests. This scrutiny should be part of every regression analysis.

Concluding Observations

Regression analysis is useful in antitrust to study alleged anticompetitive conduct—in terms of liability, impact, and damages—and to assess issues relevant to class certification. It measures the relationship between different hypothesized causal factors and economic outcomes, such as prices, and tests each relationship for statistical significance.

Econometrics is intended to help make the most reliable empirical inferences with the available data. But interpreting statistical evidence has inherent uncertainties. In this regard, the significance tests for a particular regression are seldom sufficient to establish reliability of the conclusions. What counts more is the sensitivity of the results to reasonable changes in the assumptions underlying a model, since favorable results may have been cherry-picked. The battles are fought over the variables included, the data and time periods used, the plausibility of the estimated coefficients, and whether the model has problems with serial correlation or heteroscedasticity.

The ability to unpack an econometric model is the key for effective use of much of the empirical evidence in a modern antitrust case. Lawyers can excel in this realm and find it a most satisfying professional challenge. ■

---

[1] For a more technical discussion of regression analysis, see ABA Section of Antitrust Law, Econometrics (2005).

[2] Econometrically, this use of a dummy variable implements a "before and after" methodology for measuring damages.

[3] Similarly, in a model to analyze class certification, dummy variables might identify different purchasers. This could help assess whether there was impact on different members of a proposed class, whether the impact was due to the same causal factors, and whether the impact was similar across purchasers.

[4] While other confidence intervals can be calculated, the most common practice in social science is to use the 95 percent interval as the threshold for significance. *See* Federal Judicial Center, Reference Manual on Scientific Evidence 124, 194 (2000).

[5] The F test is named for Sir Ronald Fisher, a pioneering statistician in the early 20th century.

[6] A false positive is the regression equivalent of the "correlation does not prove causation" principle. As discussed further in this article, statistical significance is no guarantee that a result is sensible in terms of economics or other facts in a case.

[7] United States v. Brown Univ., 805 F. Supp. 288, 289 (E.D. Pa. 1992), *rev'd*, 5 F.3d 658 (3d Cir. 1993).

[8] *Id.* at 296–97.

[9] The econometric model and the theory of the case are discussed in detail in Roy J. Epstein, Dennis Carlton & Gustavo Bamberger, *Antitrust and Higher Education: Was There a Conspiracy to Restrict Financial Aid?* 26 RAND J. Econ. 131–47 (1995).

[10] According to Professor Hovenkamp, *Conwood v. U.S. Tobacco* is a case in which the damages analysis hinged on whether a single observation was included as a valid data point (implying $1 billion in damages) or excluded as an outlier (implying zero damages). Herbert Hovenkamp, The Antitrust Enterprise: Principle and Execution 81 (2005).