AMERICAN BAR ASSOCIATION
Section of
Litigation Trial Evidence

Treatment of Influential Observations and Outliers in Regression Analysis

By Narsid Golic, PhD – November 14, 2013

The validity of the results is frequently criticized by opposing experts. We break down why you shouldn't jump to conclusions.

The role of experts in the American jurisprudence system has been expanding over time and has become an increasingly valuable source of specialized technical knowledge that trial lawyers seek. Originally, courts allowed expert testimony only when facts were considered too complex for lay jurors to understand. The Federal Rules of Evidence have liberalized this rule, allowing greater use of experts. Rule 702, which governs testimony by expert witnesses, states:

> A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:
>
> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
> (b) the testimony is based on sufficient facts or data;
> (c) the testimony is the product of reliable principles and methods; and
> (d) the expert has reliably applied the principles and methods to the facts of the case.

One of the areas in which economists have specialized training is the application of statistical methods to the study of economic data and problems. One such method, regression analysis, has frequently been used by expert witnesses in judicial proceedings to explain the relationship (typically a linear relationship) among variables of interest. Regression analysis can provide the trier of fact with valuable insights into understanding what would have happened "but for" certain events. For example, regression analysis, in the form of an event study, is frequently used in securities litigation to predict the values of financial assets "but for" the defendant's alleged illegal actions. In labor and employment, regression analysis is used to test whether there is "statistically significant" evidence of discrimination in employment outcomes across a group of employees.

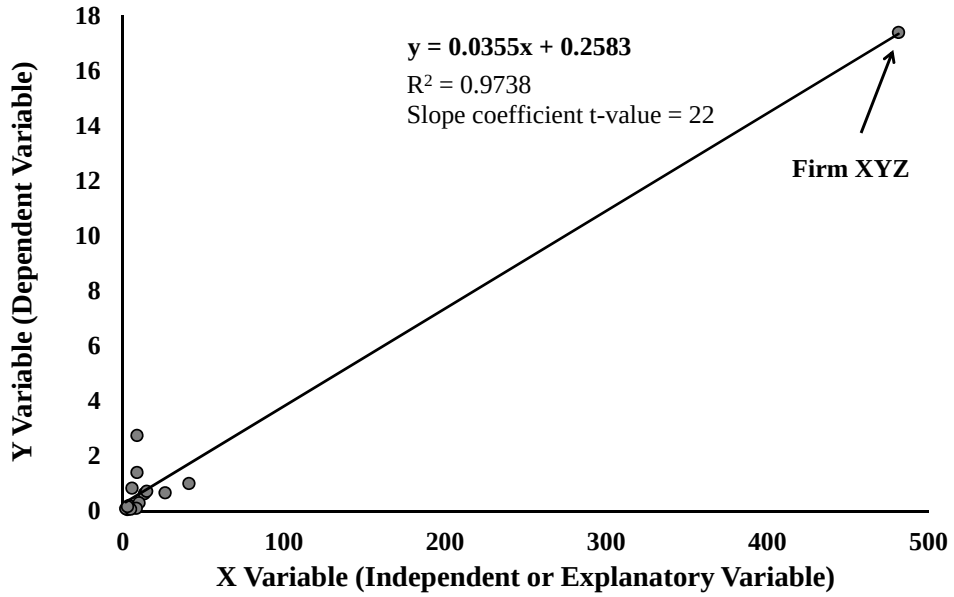
The validity of the results of regression analysis is frequently criticized by opposing experts on the basis of improper selection of the underlying data sample used for model estimation. An opposing expert may argue that the results of an expert's analysis are unreliable if they are based on sample data that include obvious "outliers." Using a case study example, this article suggests that, before criticizing the results of an opposing expert's regression analysis, it is essential to

1. implement proper statistical testing commonly used to detect "outliers" in the data and
2. make a statistical distinction between an "outlier" (an "outlier" in a dependent variable) and an "influential observation" (an "outlier" in an independent variable).

Case Study: Expert A's Proposed Regression Analysis

The data used and the case presented below were created for illustrative purposes only. Assume that Expert A has been asked by legal counsel to investigate whether a certain executive has been harmed by his company's executive compensation policy relative to predetermined industry standards. Expert A suggests that the level of executive incentive compensation is directly related to a company's performance (earnings). Expert A hypothesizes that the variation in the executive's compensation package is related in large part to the variations in the performance of the company that this executive manages. The ability to show a strong correlation between a company's performance and incentive pay across companies in the industry enables Expert A to determine the level that an executive should have been paid given the level of his or her company's performance and to determine whether this level is significantly different than what the executive actually did receive. Expert A collects data on incentive pay (Y variables) and companies' performances (X variables) across a number of companies in the industry and quantifies the linear relationship between them using the regression analysis presented in figure 1 below.

Figure 1: Data Sample Used In Regression By Expert A



The regression in figure 1 demonstrates the existence of a "statistically significant" relationship between two variables of interest, as measured by the slope coefficient t-value of 22. The t-value (or statistic) is the estimated value of the regression coefficient divided by its standard error and is used in determining whether the estimated regression coefficient meets the test of statistical significance. The statistical significance of the coefficient of an explanatory variable means that there is reasonable evidence that this explanatory variable will have an effect on a dependent variable. A t-statistic with an absolute value of 2.58, 1.96, or 1.65 or greater denotes statistical significance at the 1 percent, 5 percent, or 10 percent confidence level, respectively. It is conventional to use a 5 percent confidence level, but 10 percent or 1 percent are also commonly used. See Chris Brooks, *Introductory Econometrics for Finance* 72 (Cambridge Univ. Press

2002). A 5 percent confidence level is usually accepted as a threshold value by courts. In this illustrative case, the independent variables X (companies' performances) are statistically significant in explaining the dependent variables Y (incentive pays) at the 1 percent confidence level.

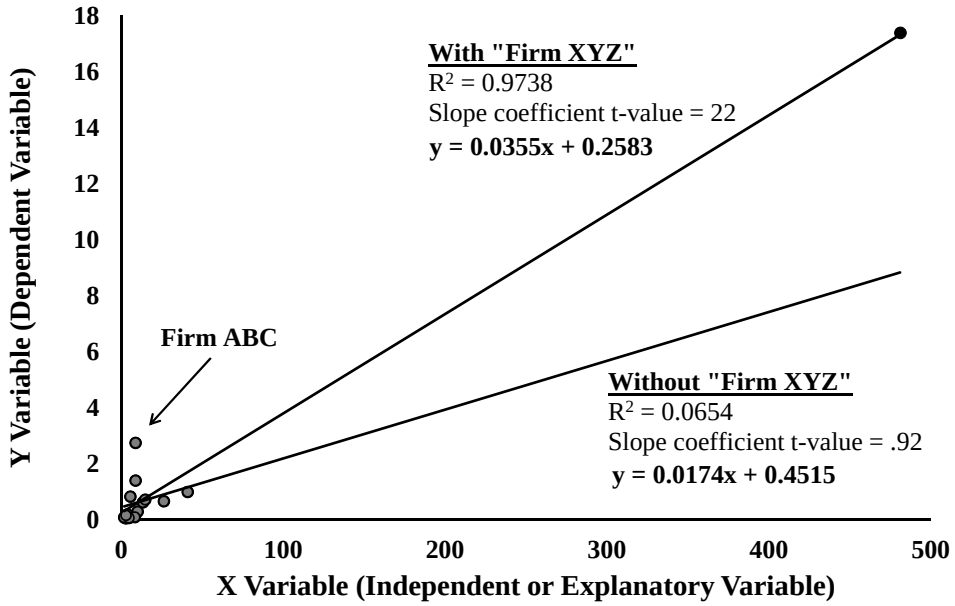
The regression in figure 1 also shows that the linear model explains 97 percent of the variation in incentive pays in the selected sample of companies as measured by R^2 of 0.9739. The R^2 statistic, also called the coefficient of determination, measures the proportion of variability in a dependent variable that is accounted for by a statistical model or the movement in independent (explanatory) variables. In this illustration, the R^2 statistic provides a statistical measure of "goodness of fit," or how well the regression line approximates the real data points. An R^2 of 1 indicates that the regression line perfectly reflects the underlying data. In such a case, one can predict the value of a dependent variable with 100 percent certainty, given the values of independent variables used in a regression model.

Based on the above results, Expert A concludes that incentive pay is related primarily to variations in a company's performance and that the predicted incentive pay of a hypothetical executive is not significantly different from the actual pay to suggest an disparate impact on the executive.

Case Study: Expert B's Criticism of Expert A's Regression Results

Expert B suggests that visual inspection of the data plot used in Expert A's regression model shows that Firm XYZ is a single obvious "outlier" because observations for all other firms are clustered in the lower left corner of the graph. Expert B opines that relying on the results of a regression that does not account for (or includes) outliers is a fundamental mistake, especially when it can be found that the results of the regression are dominated by one or more outliers in the sample. For this reason, Expert B suggests modeling the linear relationship between incentive pay and companies' performances using a data sample that excludes the outlier (observation for the Firm XYZ). The results of this regression are presented in figure 2.

Figure 2: Regression Result with and without Observation for Firm XYZ



Expert B opines that removing the observation for Firm XYZ completely changes the regression results, as also illustrated in figure 2. The new regression shows no statistically significant relationship between incentive pay and a company's performance (the slope coefficient t-value of 0.92 is lower than the critical value of 1.96 at the 5 percent confidence level), and it suggests that only 6.5 percent of the variation in incentive pay can be explained or attributed to variations in companies' performances. For this reason, Expert B opines that Expert A's conclusions are fundamentally flawed and unsupported because they are based on predictions that rely on a flawed regression model.

Case Study: Which Expert Is Right?

Before criticizing Expert A's regression analysis on the basis of identifying an outlier by visual inspection of the data plot, the appropriate steps Expert B should have taken would be

1. to implement proper statistical testing that is commonly used to detect outliers in data and
2. to make a statistical distinction between an outlier and an influential observation.

Testing for outliers. An outlier in linear regression is defined as an observation with a large "residual." See Peter Kennedy, *A Guide to Econometrics* 373 (MIT Press 5th ed. 2003). In other words, an outlier is an observation whose dependent variable (Y variable) value is unusual given values of the independent or explanatory variables (X variables). "Studentized residuals" (i.e., the residual divided by its standard error) have been commonly used in statistics to detect outliers in data. See Dennis R. Cook, "Detection of Influential Observations in Linear Regression," 19 *Technometrics* 15–18 (Feb. 1977). The idea behind using studentized residuals is to identify observations whose residual (difference between actual value of Y or dependent variable and model predicted value of Y) is statistically different from the residuals of other observations. As a rule of thumb,

an outlier can be considered any observation with a studentized residual above 1.96 (one should be even more concerned about observations with studentized residuals above 2.5 or even 3). Calculation of studentized residuals for each observation used in Expert A's regression shows no statistical evidence that Firm XYZ is an outlier because its studentized residual is 0.95 and, therefore, much lower than the critical value of 1.96.

Now, consider the observation for another company, Firm ABC, shown in figure 2. There is strong evidence that Firm ABC is an outlier because its studentized residual is 5.66. Note that visual observation of data plots does show Firm ABC having an unusually high value of Y for a given level of X relative to other observations.

Alternatively, one can test for the existence of an outlier by using a modified form of the original regression model that, in addition to the X variables (companies' performances), would also have an "observation-specific dummy" (an observation that takes a value of 1 if observation was for the Firm XYZ; otherwise, zero) as an explanatory variable. See Kennedy, *supra*, , at 379. The statistical significance of the coefficient on this "dummy variable" will indicate whether this particular observation (observation for Firm XYZ) is an outlier. In other words, the "coefficient value" on the "dummy variable" for Firm XYZ would explain by how much the difference between actual and predicted dependent variable Y (incentive pay) for Firm XYZ is different from the rest of the companies. If the difference is statistically significant, then the difference cannot be attributed to chance and the observation is an outlier. This approach also shows no statistically significant evidence that Firm XYZ is an outlier (t_{DUMMY}=0.95), but it does find statistically significant evidence that Firm ABC is an outlier (t_{DUMMY}=5.67).

If an argument is made for the exclusion of an outlier from the regression analysis, then only Firm ABC should be a possible candidate for exclusion, suggesting that the opinions offered by Expert B are incorrect. Figure 3 shows the results of Expert A's regression analysis, with and without observations for Firm XYZ, using a data sample that excludes observations for Firm ABC.

Figure 3: Expert A's Regression Result, without observation for Firm ABC (Results with and without Observation for Firm XYZ)

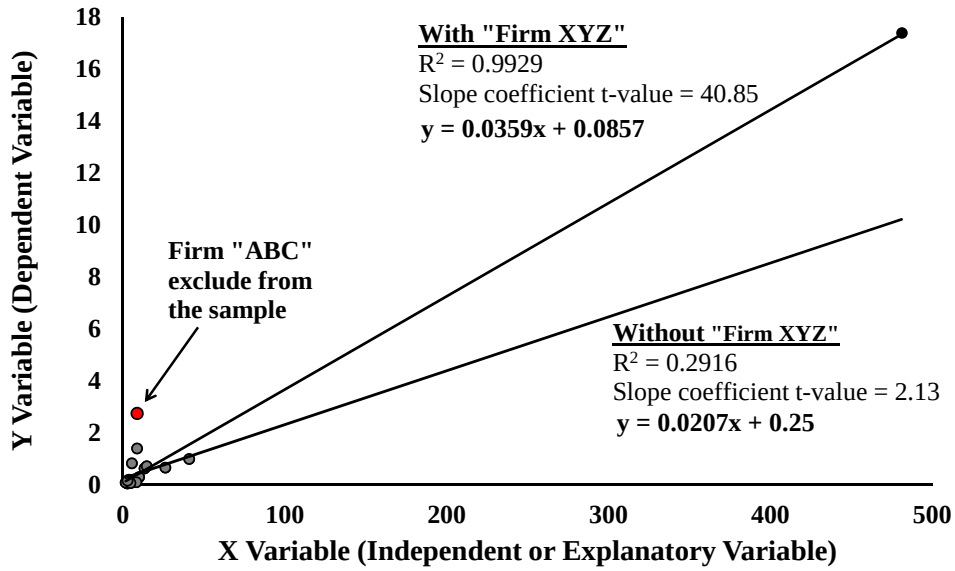


Figure 3 illustrates that the regression without Firm ABC shows a slight improvement in the "fit" relative to Expert A's original regression. There is a slight increase in the explanatory power of the model, from 97 percent to 99 percent, and there is a significant increase in statistical significance of the slope coefficient (the t-value increases from 22 to 40.85).

The distinction between outliers and influential observations. When talking about outliers, it is important to distinguish between an outlier and an "influential observation." The type of outlier discussed above, an outlier in the dependent or response Y variable, typically signals model failure. An influential observation, on the other hand, is an observation with an unusual value of an independent or explanatory X variable. A good example would be a graph of the dependent variable plotted against a single explanatory variable, with a group of observations clustered in a small area, and a single observation with a markedly different value of the explanatory variable in the same pattern of data displayed in Expert A's regression. See Peter Kennedy, *A Guide to Econometrics* 373–74 (MIT Press, 5th ed. 2003). Influential observations that are not outliers improve the precision of the regression coefficients.

Detection of influential observations is typically done by comparing regression coefficient estimates calculated using the entire data set to regression estimates calculated using the entire data set less one observation. Any observation that, when eliminated, causes the regression estimates to change markedly is identified as an influential observation. *Id.* at 379. Expert B's opinion that the removal of Firm XYZ from the regression completely changes the regression results is a good indication that Firm XYZ is an influential observation. There are two statistics that are commonly use to identify influential observations: DFITS and Cook's D. *Id.*

DFITS measures the (normalized) change in the regression estimate of the *i*th value of the dependent variable resulting from omitting the *i*th observation when calculating the regression coefficient estimate. A rough rule is that observations with $DFITS_i$ greater than $2 * \sqrt{p/n}$ should be investigated, where p is the number of independent variables and n is the number of observations. See D.A. Belsley, E. Kuh & R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (John Wiley 1980). In our illustrative case, this translates to $2*\sqrt{1/15}$ or 0.51. This test identifies two observations that can be considered "influential": Firm ABC (DFITS=1.58) and Firm XYZ (DFITS=11.54).

Alternatively, Cook's Distance (Cook's D) can also be used to detect influential observations. Cook's D is based on the sum of squared differences between estimated Y values using all observations and estimated Y values eliminating the *i*th observation. This value is then normalized by dividing it by the estimated variance of the error term (the mean square error of the regression model) times the number of explanatory variables. The conventional threshold value for detection of influential observations is 4/n. See K.A. Bollen & R. Jackman, "Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases," *in Modern Methods of Data Analysis* 257–91 (J. Fox & J. Scott Long eds., Sage 1990). Cook's D again identifies Firm ABC (Cook's D=.37) and Firm XYZ (Cook's D=67.08) as influential observations.

Although removal of Firm ABC from the regression may be warranted given that this observation has been identified as an outlier in addition to being influential (bad influential observation), the same cannot be said for Firm XYZ. Kennedy concludes that removing influential observations from a model would be a major mistake. He posits that influential observations are the most valuable observations in a data set because they may reflect unusual facts that could lead to an improvement in the model's specification. For example, if energy prices do not change significantly over time, when they do change, the new observations would be very useful in providing an estimate of future prices. *See* Kennedy, *supra*, at 374. Given that there is no evidence suggesting that this data point is a result of measurement error or other irregularity, there is no justification for removing this variable from the regression. Expert B is therefore wrong in opining that Expert A's conclusions are fundamentally flawed and unsupported because they are based on predictions produced by a flawed regression model. Expert A's regression is correct in including Firm XYZ observations in the model estimation. Figure 3 shows that, even if Firm XYZ is removed from the regression in addition to removing Firm ABC, the resulting coefficient on the X variable still shows a statistically significant positive relationship between incentive pay and company performance.

Conclusion

Although regression analysis has found widespread acceptance throughout the judicial system, the results of regression analysis have not gone unchallenged by opposing experts. An opposing expert may challenge the results of a regression analysis on the basis of the appropriateness of inclusion or exclusion of certain explanatory variables in or from the regression model. Or the opposing expert may argue that the results of a regression analysis are unreliable because they are based on sample data that include obvious outliers. Focusing on this latter criticism, this article illustrates that one should not jump to quick conclusions and criticize the results of an opposing expert's regression analysis based on the detection of outliers through visual inspection

of data plots only. To support an opinion for exclusion of an outlier from the sample used for regression estimation, one is required to implement proper statistical testing commonly used to detect outliers in data and make a statistical distinction between an outlier and an influential observation.

Keywords: litigation, trial evidence, regression analysis, statistical significance, outlier, influential observation

Narsid Golic, PhD, is a senior economist at Compass Lexecon in Chicago, Illinois.