Altman Weil, Inc.
The leader in legal consulting.

Legal Managers' Progression with Regression:
A Lawyer's Gentle Introduction to Data Insights from Linear Regression

By Rees W. Morrison

A well-known statistical tool called **regression** enables organizations to craft better strategies based on their data. With regression law firms and law departments can learn how closely some important fact, such as likely write-offs, billable hours, or effective billing rates, is associated with other numbers related to that fact, such as client revenue, years with the firm, or total timekeepers, respectively. Moreover, regression also makes predictions regarding the fact when new data arrives.

The goal of this article is to introduce leaders of lawyers to regression. Many, many kinds of regression have been developed, but this article will focus on **linear regression**. In that basic method, the **regression equation** we discuss later has constant and justifiable numbers multiplied by other numbers, rather than powers of numbers or other transformations. Furthermore, regression also works with non-numeric information, such as gender, publicly-traded or privately-held status, or cities where offices are located, but that is beyond our scope.

How might regression help you? What are some decisions where managers might apply regression? A general counsel might investigate whether and how the size of law firms retained is associated with average effective billing rate. Or she might predict whether and how more matters assigned to a firm is associated with lower effective billing rates. Law firm partners might look at several pieces of information about associates and use regression to estimate the likelihood of a particular associate making partner. Or the head of marketing might learn to what degree the number of participants in a survey influences how many times the report is downloaded. Countless examples of regression can illuminate law firm and law department operations and strengthen management decisions.

A few terms of art will prove useful:

1. The instances of data to be studied (e.g., a group of law departments or a set of associates) is referred to as **observations**.

2. Each observation has related numbers, which analysts call **variables**.

3. The question we're studying becomes the **response variable**. It will be influenced by the other numbers/variables to greater or lesser degrees.

4. Those other numbers are **predictors**, because by the methods of regression they can predict the estimated value of the response variable.

So, let's practice those terms on an example. The amount of an invoice written off by clients might be a response variable and what a firm knows about each invoice might make up the predictor variables: total fees charged on the invoice, total disbursements, the number of timekeepers, revenue of the client company, level of the person who is reviewing the bill, the number of years the firm has represented the company, and others. Once regression has digested all the data about the observations' response and predictor variables, regression will explain their linkages and importance and will be able to estimate the likely write-off amount of the next invoice for which you provide data.

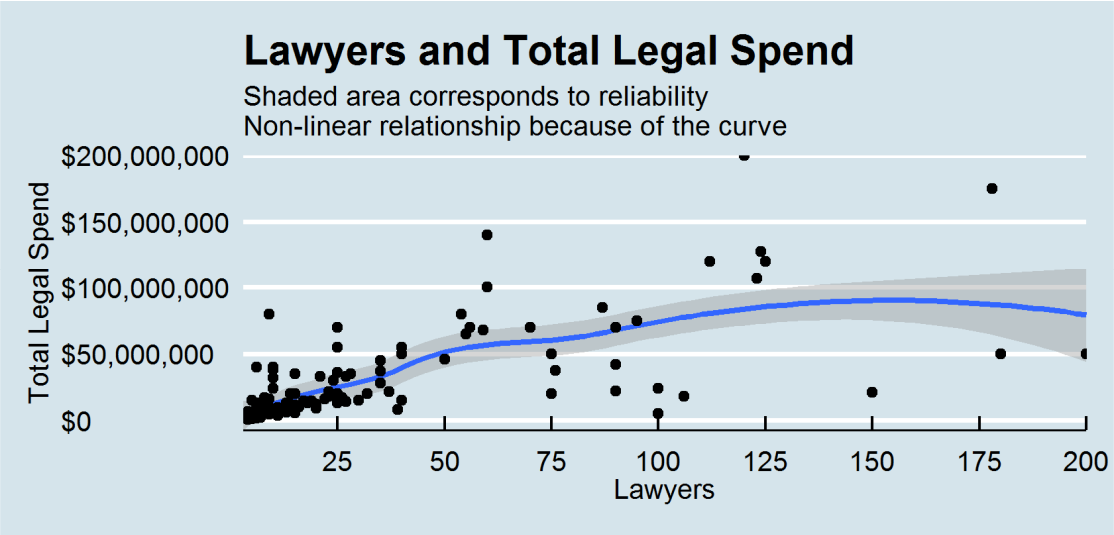
A word on software. Linear regression can be done with Excel, but for more sophisticated and easier applications analysts often use free software such as the R or Python programming languages, or they use licensed software such as Tableau, Matlab or SPSS. Whatever the software, the analyst must bring the data set into the software and tell it the response variable and which variables are predictors. The software will create a model almost instantaneously and in doing so will generate a range of information about it. Importantly, the model has **coefficients** that create the regression equation.

All data analysis, and neither regression nor the software you use is an exception, involves cleaning the data. When cleaning your data, you need to be careful about **missing data**. Most collections of data have holes. You don't have the law school graduation year for this associate or the number of matters worked on last year for that associate. When you include those associate's information in your regression model, the software may drop all of the associate's data totally even though only one piece is missing. You don't want that to happen because then you have also lost the remaining, valid data of the associate. It is also important to scrutinize **outliers**, which are very unusual and overly influential data points. Mostly you don't want mistakes in data collection or data entry to warp your regression

model. Several kinds of calculations or graphics can help you spot and deal with missing data and outliers.

With your shiny data, you still need to pass a few tests, because linear regression assumes certain things.

Linear regression assumes that the numbers you have collected are in, said in normal phrasing, decent condition. For example, you need to check that when your predictors estimate the response, the accuracy of those predictions displays no noticeable patterns at the extremes. You also do not want any predictor variables to be highly **correlated** with another predictor variable because you can't disentangle their respective contributions to the response variable. If either of those assumptions are not satisfied, you need to use a more sophisticated kind of regression or transform your data in some way.



Software can vet the assumptions. A variety of calculations will show whether your data satisfies the assumptions of linear regression and you can also create helpful graphs. Here[1] is an example of a scatter plot of the predictor variable of number of lawyers along the bottom against total legal spend on the vertical axis. The plot clearly shows that the relationship between the number of lawyers in our law departments and total legal spending is not a straight line. For this reason, we did not include the number of lawyers in our regression.[2]

[1] Data sample from a subset of participants in the Altman Weil 2018 Chief Legal Officer Survey.
[2] Many techniques exist to salvage the lawyer data and include it in the regression model, but those methods are more advanced than fits this article.

Once you have scrubbed your data and confirmed it's appropriate for regression, you are ready to roll.

To roll into regression we will use data collected from the most recent Altman Weil *Chief Legal Officer Survey*.[3] Of the hundreds of participating departments, we chose a subset of responses from the survey that met our criteria for completeness across a range of variables. Each response includes these key variables:

- Average total legal spending (internal and external legal costs) over the past two years
- Number of lawyers
- Organization revenue

and, most importantly for this article, variables that might influence total legal spending:

- Number of lawyers based in the department's largest office location
- Number of lawyers based outside of the United States
- Number of specialist lawyers (as compared to generalists who handle multiple practice areas).[4]

Our final tweak was to drop the very large and very small departments. With this set of data, in a spreadsheet with rows for each law department and columns for each piece of information about it, we are set for venturing into the pleasures of multiple regression.

We created a **regression model** of total legal spend as the response variable – the one we want to understand better and make predictions about – and three predictors: the number of lawyers, the number of lawyers in the largest office, and the revenue of the company divided by one million, e.g., a $2.5 billion company would have revenue of $2,500.

The linear regression model tells us the regression equation for the model. It starts with what is called the **intercept** but I have called the "baseline" below. The equation then has in it three numbers (one coefficient for each predictor variable: 162,085; 751,128 and 381) that are multiplied by the corresponding new data. The resulting sum of four numbers is the estimated total legal spend. Here are the equation's values to multiply by its corresponding variable:

[3] The Altman Weil 2018 Chief Legal Officer Survey is available to download at www.altmanweil.com/CLO2018.

[4] We also asked about the number of direct reports to the General Counsel, the company's industry, and the average years out of law school of the department's lawyers, but found that for various reasons those figures were not useful for the regression. To create the final data set we also dropped 18 departments that had less than three lawyers and three departments that had more than 400 lawyers.

Total Legal Spend (TLS) = $5,382,185 as the baseline
+ $162,085 times the number of lawyers in the department
+ $751,128 times the number of lawyers in the largest office
+ $381 times the revenue of the company divided by one million.

As an example, taking a 10-lawyer department that offices 8 of them in its largest location and a $2 billion company (2,000 millions), the model estimates its total legal spend. We plug those numbers into the model's equations, with its coefficients:

TLS = $5,382,185 + $162,085 x 10 [lawyers] + $751,128 x 8 [lawyers in the largest office] + $381 x 2000 [revenue divided by one million]. Thus, TLS would be predicted to be $13,774,059.

This equation tells us the estimated spending figure based on any of the three predictor variables if the other two variables are held constant. It isolates the effect of each predictor when the others are not factors in the prediction. This is a powerful tool!

The model also tells us whether each predictor variable influences the response variable to such a degree that we can rely on that influence. Imagine if you were able to find lots of new data sets like the one you have and you ran the same regression over and over, less than 5% of the time would you come up with a different conclusion. This is what it means to find that a predictor variable is **statistically significant**.

With our model, surprisingly, the number of lawyers in a law department is not a significant predictor of total legal spend. Both Largest Office, very much, and Revenue, to a degree, however are statistically significant in their association with legal spend. Each additional lawyer added to the largest office increases total legal spend by approximately a quarter of a million dollars. Assuming the data is correct and the regression model was done right, why might this counter-intuitive result be?

1. It may be that as the in-house lawyers of a U.S. company cluster more in one location, the company is less likely to have significant numbers of foreign lawyers – who are paid much less and who also reduce the demand for expensive outside counsel.

2. Possibly the co-location of in-house lawyers at one site tends to be in an expensive metropolitan area, with higher costs all around.

3. Consider that the additional lawyer probably costs the company more than $350,000 when all compensation, benefits and overhead are accounted for (half the estimated

increase in TLS) and the additional lawyer may surface more legal issues, thus triggering the remaining additional spending on outside counsel.

4. Possibly this finding suggests that locating lawyers near business units makes more budget sense.

5. It may be that the in-house lawyers in the largest locations tend to be more senior, and therefore more costly – including administrative support.

The model also tells us the portion of the estimated total legal spend that is explained by the predictors in the model. That piece of useful information is called **Adjusted R-squared**. Interesting in its own right, Adjusted R-squared helps when you compare different models for the same data set, such as when you try different combinations of predictor variables. Our model touts an Adjusted R-squared of 55 percent, which is modest, and warns us that other predictors that we don't have available account for a sizeable portion of the estimate. Industry category would likely claim some portion of the unexplained portion but the industry data we collected was too varied across the sample to be a useful predictor.

We have seen how a linear regression model can tell us how much a predictor influences the response variable, whether that influence is statistically significant, and how much our model explains. But wait, there's more!

The example above shows how linear regression models enable predictions. Let's illustrate that capability with a hypothetical law firm that wants to predict an associate's annual billable hours based on the number of partners that associate worked for during the year. The observations of the data set would be the firm's associates. For each associate the number of hours he or she billed during the most recent year would be the response variable and the number of partners who assigned him or her work would be the predictor variable. Linear regression would generate an equation and, assuming partners worked for is statistically significant, the firm could predict billable hours for any associate based on the number of partners who gave them work.

With this particular illustration, the value of regression as a prediction tool may be low, but as a tool to understand the relationship between partner-worked-for numbers and billable hours, it could be insightful for partners.

Regression, therefore, lets you weave the straw of your data into golden insights. It can tell you how influential predictors are; it can tell you how much of your important fact (the response variable) is

explained by those drivers; and, it can let you predict a new instance if you can provide the appropriate information on drivers. Regression offers an exciting tool that will deepen your understanding of your data.

Rees W. Morrison, a principal at Altman Weil, Inc., helps law firm and law department leaders make better decisions through data analytics. Contact Mr. Morrison at (973) 568-9110 or rwmorrison@altmanweil.com.

If any reader would like to talk about their data and how they might benefit from linear regression, the author is offering a free, no-charge hour of telephone discussion. Email me and let's figure out what will help you learn from regression! Likewise, for more on regression please visit the author's blog, JurisDatoris.com.