# What Practitioners Need to Know . . .

by Mark Kritzman

## . . . About Regressions

How can we predict uncertain outcomes? We could study the relations between the uncertain variable to be predicted and some known variable. Suppose, for example, that we had to predict the change in profits for the airline industry. We might expect to find a relation between GNP growth in the current period and airline profits in the subsequent period, because economic growth usually foreshadows business travel as well as personal travel. We can quantify this relation through a technique known as regression analysis.

Regression analysis can be traced to Sir Francis Galton (1822–1911), an English scientist and anthropologist who was interested in determining whether or not a son's height corresponded to his father's height. To answer this question, Galton measured a sample of fathers and computed their average height. He then measured their sons and computed their average height. He found that fathers of above-average height had sons whose heights tended to exceed the average. Galton termed this phenomenon "regression toward the mean."

### Simple Linear Regression

To measure the relation between a single independent variable (GNP growth, in our earlier example) and a dependent variable (subsequent change in airline profits), we can begin by gathering some data on each variable—for example, actual GNP growth in each quarter of a given sample period and the change in the airline industry's profit over each subsequent quarter. We can then plot the intersects of these observations. The result is a scatter diagram such as the one shown in Figure A.

The horizontal axis represents a quarter's GNP growth and the vertical axis represents the percentage change in profits for the airline industry in the subsequent quarter. The plotted points in the figure indicate the actual percentage change in airline profits associated with a given level of GNP growth. They suggest a positive relation; that is, as GNP increases so do airline profits. The straight line sloping upward from left to right measures this relation.

This straight line is called the *regression line*. It has been fitted to the data in such a way that the sum of the squared differences of the observed airline profits from the values along the line is minimized. The values along the regression line corresponding to the vertical axis represent the predicted change in airline profits given the corresponding prior quarter's GNP growth along the horizontal axis. The difference between a value predicted by the regression line and the actual change in airline profits is the error, or the residual.

Given a particular value for GNP growth, we can predict airline profits in the subsequent quarter by multiplying the GNP growth value by the slope of the regression line and adding to this value the intercept of the line with the vertical axis. The equation is:

$$\hat{Y}_i = \alpha + \beta \cdot X_i.$$

Here $\hat{Y}_i$ equals the predicted percentage change in airline profits, $\alpha$ equals the intercept of the regression line with the vertical axis, $\beta$ equals the slope of the regression line and $X_i$ equals the prior quarter's growth in GNP.

We can write the equation for the *actual* percentage change in airline profits, given our observation of the prior quarter's GNP growth, by adding the error to the prediction equation:

$$Y_i = \alpha + \beta \cdot X_i + e_i.$$

Here $Y_i$ equals the actual percentage change in airline profits and $e_i$ equals the error associated with the predicted value.

Positive errors indicate that the regression equation underestimated the dependent variable (airline profits) for a particular value of the independent variable (GNP growth), while negative errors indicate that the regression equation overestimated the dependent variable. Figure B illustrates these notions.

### Analysis of Variance

To determine whether or not our regression equation is a good predictor of the dependent variable, we can

**Figure A** Scatter Diagram

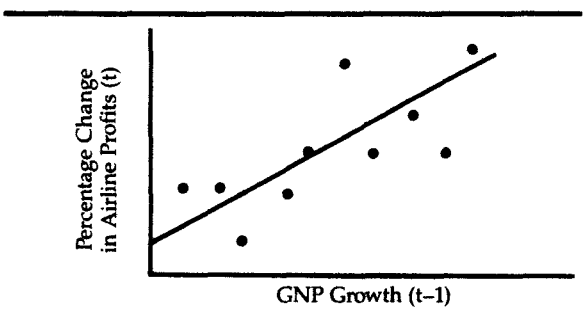
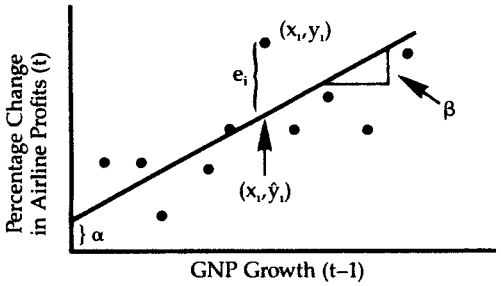


**Figure B** Regression Model




start by performing an analysis of variance. This involves dividing the variation in the dependent variable (change in airline profits) into two parts—that explained by variation in the independent variable (prior quarter's GNP growth) and that attributable to error.

In order to proceed, we must first calculate three values—the total sum of the squares, the sum of the squares due to regression and the sum of the squares due to error. The total sum of the squares is calculated as the sum of the squared differences between the observed values for the dependent variable and the average of those observations. The sum of the squares due to regression is calculated as the sum of the squared differences between the predicted values for the dependent variable and the average of the observed values for the dependent variable. Finally, the sum of the squares due to error is calculated as the sum of the squared differences between the observed values for the dependent variable and the predicted values for the dependent variable.

The ratio of the sum of the squares due to regression to the total sum of the squares equals the fraction of variation in the dependent variable that can be explained by variation in the independent variable. It is referred to as R-squared ($R^2$), or the coefficient of determination. It ranges in value from 0 to 1. A high value for R-squared indicates a strong relation between the dependent and independent variables, whereas a low value for R-squared indicates a weak relation.[1]

The square root of R-squared is called the correlation coefficient. It measures the strength of the association between the dependent and independent variables. In the case of an inverse relation—that is, where the slope of the regression line is negative—we must adjust the sign of the correlation coefficient to accord with the slope of the regression line. The correlation coefficient ranges in value from $-1$ to $+1$.

## Residual Analysis

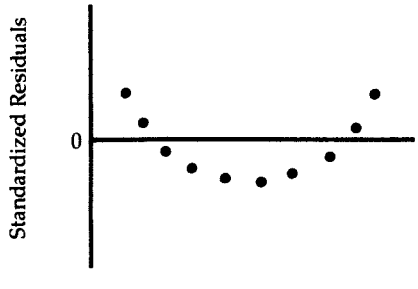
R-squared is only a first approximation of the validity of the relation between the dependent and independent variables. Its validity rests on several assumptions: (1) the independent variable (GNP growth in the example) must be measured without error; (2) the relation between the dependent and independent variables must be linear (as indicated by the regression line); (3) the errors, or residuals, must have constant variance (that is, or they must not increase or decrease with the level of the independent variable); (4) the residuals must be independent of each other; and (5) the residuals must be normally distributed. Unless these assumptions are true, the measured relation between the dependent and independent variables, even if it has a high R-squared, may be spurious.

The importance of the first assumption is self-evident and should not require elaboration. The importance of some of the remaining assumptions may require some elaboration. In order to analyze the residuals, it is convenient to standardize each residual by dividing it by the standard error.[2] We can then plot the residuals to determine whether or not the above assumptions are satisfied.

Figure C shows a plot of standardized residuals. These seem to trace a convex curve. The errors associated with low values of the independent variable are positive; but they become increasingly negative with higher levels of the independent variable and then become positive again as the independent variable increases still more. In this case, it is apparent that the relation between the dependent and independent variables violates the assumption of linearity. The dependent variable increases with the independent variable but at a decreasing rate. That is to say, the independent variable has less and less effect on the dependent variable.

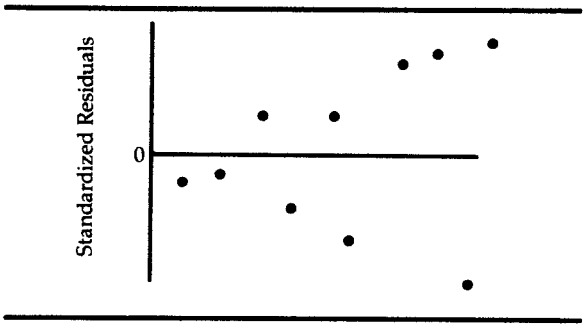
This pattern is characteristic of the relation between the level of advertising expenditures and sales, for example. Suppose a company distributes a product in several regions, and it varies the level of advertising expenditures across these regions to measure advertising's effect. The company will likely observe higher sales in a region where it advertises a little than in a region where it does not advertise at all. And as advertising increases from region to region, corresponding sales should also increase. At some level of

**Figure C** Non-Linearity




1. Footnotes appear at the end of article.

**Figure D** Heteroscedasticity




sales, however, a region will start to become saturated with the product; additional advertising expenditures will have less and less impact on sales.

The obvious problem with using a linear model when the independent variable has a diminishing effect on the dependent variable is that it will overestimate the dependent variable at high levels of the independent variable. In many instances, we can correct this problem by transforming the values of the independent variable into their reciprocals and then performing a linear regression of the dependent variable on these reciprocals.

Figure D illustrates a case in which the absolute values of the standardized residuals *increase* as the values for the independent variable increase. In this case, the errors involved in predicting the dependent variable will grow larger and larger, the higher the value of the independent variable. Our predictions are subject to larger and larger errors. This problem is known as heteroscedasticity. It can often be ameliorated by transforming the independent variables into their logarithmic values.

Figure E shows a plot in which all the standardized residuals are positive with the exception of a single very large negative residual. This large negative residual is called an outlier, and it usually indicates a specious observation or an event that is not likely to recur. If we had included GNP growth in the last quarter of 1990 as one of the observations used to predict airline profitability, for example, we would have grossly overestimated airline profits in the first quarter of 1991; both business and personal air travel dropped precipitously in early 1991 because of the threat of terrorism stemming from the Gulf War. In this case, we would simply eliminate the outlying observation and rerun the regression with the remaining data.

In all these examples, the residuals are in violation of the independence assumption. That is, the plotted points in Figures C, D and E form patterns, rather than random distributions. This suggests that the residuals are not independent of one another but are correlated with one another, or autocorrelated.

Without examining the residuals explicitly, we can test for first-order autocorrelation (correlation between successive residuals) by calculating a Durbin-Watson statistic. The Durbin-Watson statistic is approximately equal to $2(1 - R)$, where R equals the correlation coefficient measuring the association between successive residuals. As the Durbin-Watson statistic approaches 2, we should become more confident that the residuals are independent of each other (at least successively). Depending on the number of variables and number of observations, we can determine our level of confidence specifically.

With economic and financial data, it is often useful to transform the data into percentage changes, or first differences. This reduces autocorrelation.

## Multiple Linear Regression

We have so far focused on simple linear regressions—that is, regressions between a dependent variable and a single independent variable. In many instances, variation in a dependent variable can be explained by variation in several independent variables. Returning to our example of airline profits, we may wish to include changes in energy prices as a second independent variable, given the relatively high operating leverage associated with the airline industry.

We can express this multiple regression equation as follows:

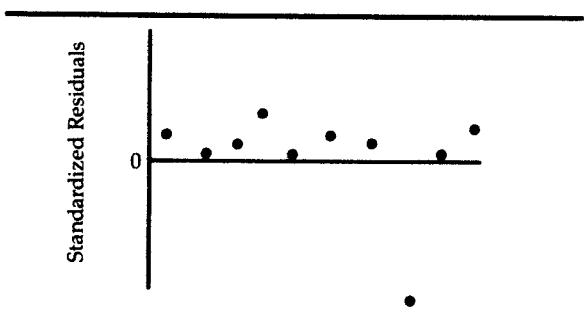
$$\hat{Y}_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2}.$$

Here $X_{i1}$ and $X_{i2}$ equal the two independent variables (GNP growth and changes in energy prices) and $\beta_1$ and $\beta_2$ equal their coefficients.

It seems reasonable to expect that as fuel prices rise, profit margins in the airline industry will fall and vice versa. This would mean a negative relation between airline profits and energy prices. Thus $\beta_2$ would be a negative value. But an increase in economic activity could increase demand for energy and contribute to a rise in energy prices. Thus the two independent variables, GNP growth and changes in energy prices, may not be independent of each other. This problem is known as multicolinearity.

Suppose we run two simple linear regressions using two independent variables. If the variables are

**Figure E** Outlier



independent of each other, then the sum of the R-squares from the two regressions will equal the R-squared from a multiple linear regression combining the two variables. To the extent that the independent variables are correlated with each other, however, the R-squared from the multiple regression will be less than the sum from the two simple regressions.

When the independent variables in a multiple regression are colinear, we must take care in interpreting their coefficients. The coefficients $\beta_1$ and $\beta_2$ in the above equation represent the marginal sensitivity of a change in airline profits to a one-unit change in GNP growth and to a one-unit change in energy prices in the prior quarter. If $\beta_1$ equals 0.7 per cent and $\beta_2$ equals $-0.15$ per cent, for example, we would expect airline profits to increase by 0.7 per cent if GNP grew 1 per cent in the prior quarter and energy prices remained constant. If energy prices increased by 1 per cent in the prior quarter and GNP remained constant, we would expect airline profits to decrease by 0.15 per cent. To the extent there is multicolinearity between the independent variables, these responses would not equal the sensitivity of airline profits to the same independent variables as measured by simple linear regressions.

Regression analysis is a powerful tool for the financial analyst. But, as we have attempted to demonstrate, the summary statistics from regression analysis can be misleading.

## Footnotes

1. As part of their output, most regression packages include measures of statistical significance such as an F-value and a t-statistic. The F-value is computed as the ratio of the sum of the squares due to regression (adjusted by the degrees of freedom) to the sum of the squares due to error (also adjusted by the degrees of freedom). Its significance depends on the number of variables and observations. The t-statistic measures the significance of the coefficients of the independent variables. It is computed as the ratio of the coefficient to the standard error of the coefficient. The F-test and the t-test are the same for simple linear regressions, but not necessarily for multiple linear regressions.
2. The standard error measures the dispersion of the residuals around the regression line. It is calculated as the square root of the average squared differences of the observed values from the values predicted by the regression line. To estimate the average of the squared differences, we divide the sum of the squared differences by the number of observations less one.

---

**From the Board** *concluded from page 8.*

toward shared risk-reward pension models are more persuasive explanations for the persistence of 60/40 pension fund investment policies.

## Congress And the FASB

Bodie is on the mark when he points to the central roles Congress and the Financial Accounting Standards Board (FASB) have played in promoting fuzzy thinking about pension finance and investments. The ABO now plays a central role in Congress' funding rules and in the FASB's financial disclosure rules. No doubt these developments have given this liability measure a great deal more respectability than it deserves.

Though he didn't, Bodie could also have fingered the enshrinement of pension assets in ERISA as trust assets through the "exclusive benefit rule." Both this rule and the ABO concept pull pension thinking away from legitimate pension models (e.g., the integrated finance, pure defined benefit model or some form of a fair shared risk-reward model). Instead, they push pension fiduciaries to contemplate perverse, unstable pension models such as the one described by Bodie (i.e., unshared risk-shared reward arrangements, which lead to stakeholder gaming).

## Three Important Lessons

In the end, there are three important lessons in this tale of the two investment policies and their origins. These lessons apply as much to public-sector pension funds as they do to corporate funds.

1. There can be no clearly defined investment policies without clearly defined pension deals.
2. Perverse legal and accounting pension rules promote perverse pension deals.
3. Perverse pension deals in turn promote perverse investment policies.