Multivariate Regression, Part 1

In This Issue

1. Learn how regression analysis finds the unique best-fitting line through a set of data points
2. Learn how to run and interpret a sample multiple regression using Risk Simulator

"What is multivariate regression?"

Theory

It is assumed that the user is knowledgeable about the fundamentals of regression analysis. The general bivariate linear regression equation takes the form of

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where β_0 is the intercept, β_1 is the slope, and ε is the error term. It is bivariate as there are only two variables, a *Y,* or dependent variable, and an *X,* or independent variable, where *X* is also known as the regressor (sometimes a bivariate regression is also known as a univariate regression as there is only a single independent variable *X*). The dependent variable is so named because it *depends* on the independent variable; for example, sales revenue depends on the amount of marketing costs expended on a product's advertising and promotion, making the dependent variable "sales" and the independent variable "marketing costs." An example of a bivariate regression is seen as simply inserting the best-fitting line through a set of data points in a two-dimensional plane, as seen on the left in Figure 1. In other cases, a multivariate regression can be performed, where there are multiple, or *k* number of, independent *X* variables or regressors, where the general regression equation will now take the form of

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 ... + \beta_k X_k + \varepsilon$$

In this case, the best-fitting line will be within a *k + 1* dimensional plane.

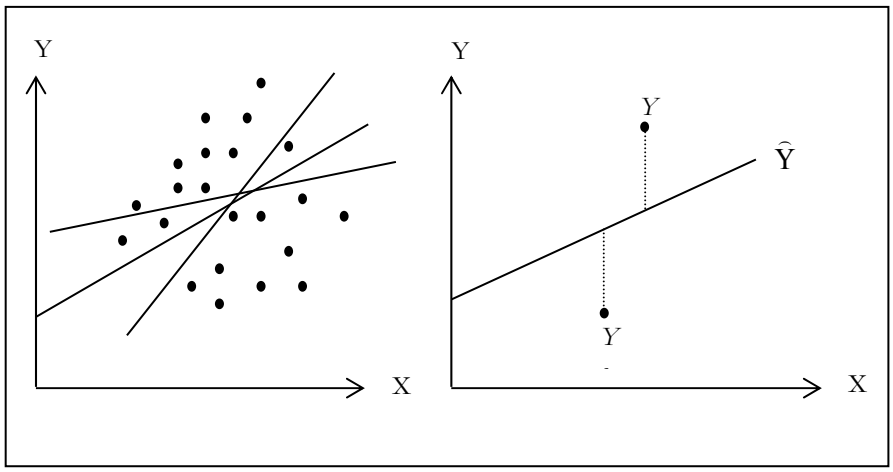

Figure 1. Bivariate Regression

However, fitting a line through a set of data points in a scatter plot as in Figure 1 may result in numerous possible lines. The best-fitting line is defined as the single unique line that minimizes the total vertical errors, that is, the sum of the absolute distances between the actual data points (Y_i) and the estimated line (**Error! Objects cannot be created from editing field codes.**), as shown on the right of Figure 1. To find the best-fitting unique line that minimizes the errors, a more sophisticated approach is applied, using regression analysis. Regression analysis, therefore, finds the unique best-fitting line by requiring that the total errors be minimized, or by calculating

$$Min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Contact Us

Real Options Valuation, Inc.

4101F Dublin Blvd., Ste. 425, Dublin, California 94568 U.S.A.

admin@realoptionsvaluation.com
www.realoptionsvaluation.com
www.rovusa.com

where only one unique line minimizes this sum of squared errors. The errors (vertical distances between the actual data and the predicted line) are squared to avoid the negative errors from canceling out the positive errors. Solving this minimization problem with respect to the slope and intercept requires calculating first derivatives and setting them equal to zero:

$$\frac{d}{d\beta_0}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=0 \text{ and } \frac{d}{d\beta_1}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=0$$

which yields the bivariate regression's least squares equations:

$$\beta_1=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\frac{\sum_{i=1}^{n}X_iY_i-\frac{\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{n}}{\sum_{i=1}^{n}X_i^2-\frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}}$$

$$\beta_0=\bar{Y}-\beta_1\bar{X}$$

For multivariate regression, the analogy is expanded to account for multiple independent variables, where

$$Y_i=\beta_1+\beta_2X_{2,i}+\beta_3X_{3,i}+\varepsilon_i$$

and the estimated slopes can be calculated by:

$$\hat{\beta}_2=\frac{\sum Y_iX_{2,i}\sum X_{3,i}^2-\sum Y_iX_{3,i}\sum X_{2,i}X_{3,i}}{\sum X_{2,i}^2\sum X_{3,i}^2-\left(\sum X_{2,i}X_{3,i}\right)^2}$$

$$\hat{\beta}_3=\frac{\sum Y_iX_{3,i}\sum X_{2,i}^2-\sum Y_iX_{2,i}\sum X_{2,i}X_{3,i}}{\sum X_{2,i}^2\sum X_{3,i}^2-\left(\sum X_{2,i}X_{3,i}\right)^2}$$

In running multivariate regressions, great care must be taken to set up and interpret the results. For instance, a good understanding of econometric modeling is required (e.g., identifying regression pitfalls such as structural breaks, multicollinearity, heteroskedasticity, autocorrelation, specification tests, nonlinearities, and so forth) before a proper model can be constructed.

Procedure

- Start Excel and type in or open your existing dataset (the illustration in Figure 2 uses the file *Multiple Regression* in the examples folder).
- Check to make sure that the data are arranged in columns and select the data including the variable headings, and click on *Risk Simulator | Forecasting | Multiple Regression*.
- Select the dependent variable and check the relevant options (lags, stepwise regression, nonlinear regression, and so forth) and click *OK* (Figure 2).

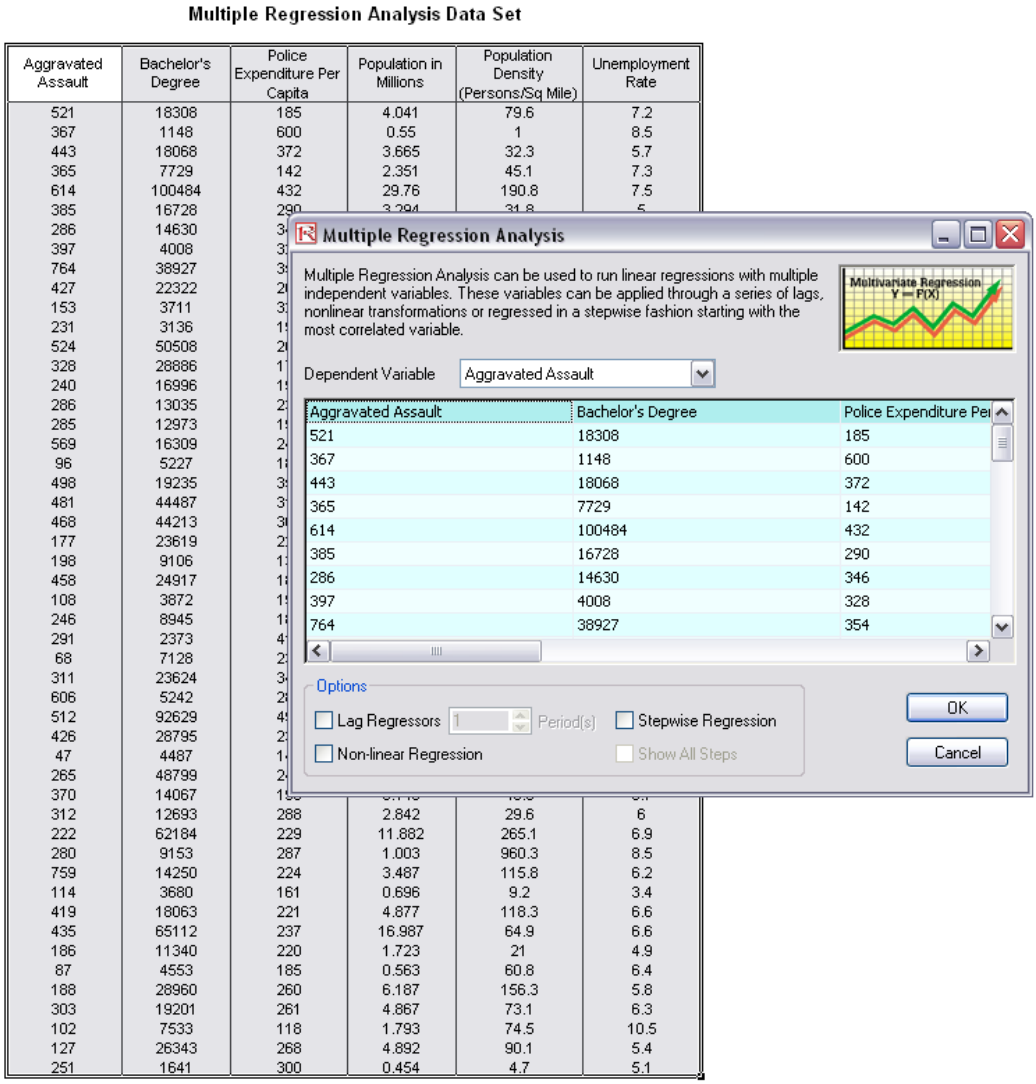

Figure 2. Running a Multivariate Regression

Results Interpretation

Figure 3 (on the next page) illustrates a sample multivariate regression result report generated. The report comes complete with all the regression results, analysis of variance results, fitted chart, and hypothesis test results.

In "**Multivariate Regression, Part 2**," you will learn about a powerful automated approach to regression analysis known as "stepwise regression" and about how goodness-of-fit statistics provide a glimpse into the accuracy and reliability of the estimated regression model.

Regression Analysis Report

Regression Statistics

R-Squared (Coefficient of Determination)	0.8146
Adjusted R-Squared	0.7776
Multiple R (Multiple Correlation Coefficient)	0.9026
Standard Error of the Estimates (SEy)	0.5725
nObservations	7

The R-Squared or Coefficient of Determination indicates that 0.81 of the variation in the dependent variable can be explained and accounted for by the independent variables in this regression analysis. However, in a multiple regression, the Adjusted R-Squared takes into account the existence of additional independent variables or regressors and adjusts this R-Squared value to a more accurate view of the regression's explanatory power. Hence, only 0.78 of the variation in the dependent variable can be explained by the regressors.

The Multiple Correlation Coefficient (Multiple R) measures the correlation between the actual dependent variable (Y) and the estimated or fitted (Y) based on the regression equation. This is also the square root of the Coefficient of Determination (R-Squared).

The Standard Error of the Estimates (SE_Y) describes the dispersion of data points above and below the regression line or plane. This value is used as part of the calculation to obtain the confidence interval of the estimates later.

Regression Results

	Intercept	Ad Size
Coefficients	4.3643	0.0845
Standard Error	0.5826	0.0180
t-Statistic	7.4911	4.6877
p-Value	0.0007	**0.0054**
Lower 5%	2.8667	0.0382
Upper 95%	5.8619	0.1309

Degrees of Freedom		Hypothesis Test	
Degrees of Freedom for Regression	1	Critical t-Statistic (99% confidence with df of 5)	4.0321
Degrees of Freedom for Residual	5	Critical t-Statistic (95% confidence with df of 5)	2.5706
Total Degrees of Freedom	6	Critical t-Statistic (90% confidence with df of 5)	2.0150

The Coefficients provide the estimated regression intercept and slopes. For instance, the coefficients are estimates of the true; population b values in the following regression equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$. The Standard Error measures how accurate the predicted Coefficients are, and the t-Statistics are the ratios of each predicted Coefficient to its Standard Error.

The t-Statistic is used in hypothesis testing, where we set the null hypothesis (Ho) such that the real mean of the Coefficient = 0, and the alternate hypothesis (Ha) such that the real mean of the Coefficient is not equal to 0. A t-test is is performed and the calculated t-Statistic is compared to the critical values at the relevant Degrees of Freedom for Residual. The t-test is very important as it calculates if each of the coeffients is statististically significant in the presence of the other regressors. This means that the t-test statistically verifies whether a regressor or independent variable should remain in the regression or it should be dropped.

The Coefficient is statistically significant if its calculated t-Statistic exceeds the Critical t-Statistic at the relevant degrees of freedom (df). The three main confidence levels used to test for significance are 90%, 95% and 99%. If a Coefficient's t-Statistic exceeds the Critical level, it is considered statistically significant. Alternatively, the p-Value calculates each t-Statistic's probability of occurrence, which means that the smaller the p-Value, the more significant the Coefficient. The usual significant levels for the p-Value are 0.01, 0.05, and 0.10, corresponding to the 99%, 95%, and 99% confidence levels.

The Coefficients with their p-Values highlighted in blue indicate that they are statistically significant at the 90% confidence or 0.10 alpha level, while those highlighted in red indicate that they are not statistically significant at any other alpha levels.

Analysis of Variance

	Sums of Squares	Mean of Squares	F-Statistic	p-Value	Hypothesis Test	
Regression	7.2014	7.2014	21.9747	**0.0054**	Critical F-statistic (99% confidence with df of 4 and 3)	16.2582
Residual	1.6386	0.3277			Critical F-statistic (95% confidence with df of 4 and 3)	6.6079
Total	8.8400				Critical F-statistic (90% confidence with df of 4 and 3)	4.0604

The Analysis of Variance (ANOVA) table provides an F-test of the regression model's overall statistical significance. Instead of looking at individual regressors as in the t-test, the F-test looks at all the estimated Coefficients' statistical properties. The F-statistic is calculated as the ratio of the Regression's Mean of Squares to the Residual's Mean of Squares. The numerator measures how much of the regression is explained, while the denominator measures how much is unexplained. Hence, the larger the F-statistic, the more significant the model. The corresponding p-Value is calculated to test the null hypothesis (Ho) where all the Coefficients are simultaneously equal to zero, versus the alternate hypothesis (Ha) that they are all simultaneously different from zero, indicating a significant overall regression model. If the p-Value is smaller than the 0.01, 0.05, or 0.10 alpha significance, then the regression is significant. The same approach can be applied to the F-statistic by comparing the calculated F-statistic with the critical F values at various significance levels.

Forecasting

Period	Actual (Y)	Forecast (F)	Error (E)
1	5.9	5.3786	0.5214
2	5.6	5.8857	(0.2857)
3	5.5	6.3929	(0.8929)
4	7.2	6.9000	0.3000
5	8	7.4071	0.5929
6	7.7	7.9143	(0.2143)
7	8.4	8.4214	(0.0214)

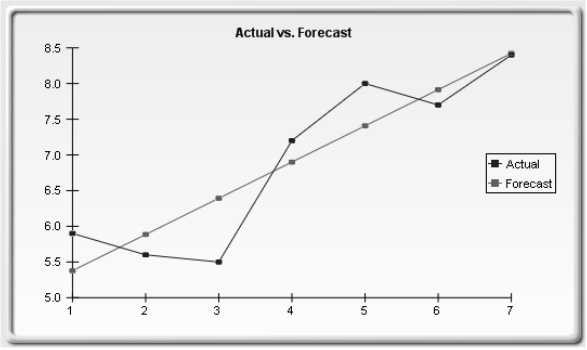

Figure 3. Multivariate Regression Results

TO BE CONCLUDED IN "Mulitvariate Regression, Part 2"