Understanding
Multivariate Regression Analysis

The goal in any data analysis is to extract from raw information, an accurate estimation or prediction of a desired outcome or the strength of a relationship between select variables or the answer to a specific problem you're trying to solve.

Before any testing or estimation work, careful data cleansing is essential to review for errors, followed by data summarization. One of the most important and common questions asked is: Is there a statistical relationship between a response variable (Y) and explanatory variables (Xi). To answer this question, we can employ regression analysis, of which there are various types. It can be used to predict the response variable for any arbitrary set of explanatory variables.

Expressed another way, these methods allow us to assess the impact of multiple variables on the response variable we'd like to predict.

Inferential statistics are used to answer questions about the data, to test hypotheses, to generate a measure of effect such as a ratio of rates or risks, to describe associations (correlations) or to model relationships (regression) within the data and other functions.

Usually estimates are measures of associations or the magnitude of effects. Confounding (surprising or confusing) measurement errors and random errors mean that our estimates are unlikely to equal the true ones. In the estimation process, random error is not avoidable.

What is Multivariate Regression

Multivariate regression is a method used to measure the degree to which more than one independent variable (the predictors) and more than one dependent variable (responses), are linearly* related. The method is broadly used to predict the behavior of the response variables associated with changes in the predictor variables, once a desired degree of relationship has been established.

Multivariate regression is a particularly important and powerful form of data analysis and is more accurate when applied to the real world. In the world of business, specifically, situations are rarely ever influenced by a single factor. Usually, there are many factors working in concert to create a result or outcome. When data is collected on certain sets of conditions, this type of data analysis allows you to predict data in related conditions.

*Linear suggests that the relationship between dependent and independent variables can be expressed in a straight line.

Examples of Multivariate Regression

Multivariate regression is one of the simplest Machine Learning algorithms and falls under the class of Supervised Learning Algorithms. Some of the problems it can solve include:

Example 1

A doctor has collected data on cholesterol, blood pressure and weight. She also collected data on the eating habits of the subjects (eg. how many ounces of red meat, fish, dairy products and chocolate was consumed per week). She wants to investigate the relationship between the three health measures and the eating habits.

Example 2
A realtor wants to predict housing prices by using various factors such as the neighborhood, size of house, number of rooms, age of the house, attached facilities, distance to nearest train station, distance to nearest shopping area, etc.

Example 3
Can a supermarket owner maintain the right stock of water, ice cream, frozen foods, canned foods and meat as a function of temperature, tornado chance and gas prices during the tornado season?

In this example, several obvious assumptions can be drawn:

- If it is too hot, ice cream sales increase;
- If a tornado hits, sales of water and canned foods will increase while ice cream, frozen foods and meat will decrease; and
- If gas prices increase, prices on all goods will increase.

A mathematical model based on multivariate regression analysis will address this and other more complicated questions.

Simple Regression vs Multivariate Regression. What is the Difference?

Simple Regression

Also known as Univariate Regression, is a model with only one predictor and one response variable.

This is the simplest form of data analysis where the data being analyzed contains only one variable.

For example, a model that can predict car prices based on engine size. So one dependent variable (car price) is predicted using one independent variable.

And what if we feed the model more inputs such as horsepower, fuel type, number of doors, length, width, height etc? Will it improve the accuracy?

If we enhance the model by feeding it more input data, ie. more independent variables, we've now entered the world of multivariate regression.

Multivariate Regression

This model has more than one predictor and more than one response.

The interpretation of a multivariate model provides the impact of each independent variable on the dependent variable (the target).

How Does Multivariate Analysis (MVA) Work?

MVA attempts to identify the factors that effect a specified dependent variable, ie. the variable we'd like to predict. It may use patterns of association between factors (variables) to suggest pseudo-causal models.

MVA enables us to elaborate the measured bi-variate association (between X and Y variables) by considering other variables. Thus MVA essentially does two things.

First: it acts as a check on the assumed relationship between X and Y as revealed by the simple bivariate correlation. Multivariate analysis is used to establish the non-spurious (ie. genuine) relationship between X and Y by computing correlations and controlling for other variables that might explain away the observed relationship. If the relationship between X and Y persists when other variables are considered, then the correlation is regarded as 'non-spurious'. In other words, it is not a correlation that can be explained away by a pre-existing variable. For example, a relationship between income and education might be explained away by age.

Second: MVA acts to elaborate the association between X and Y by specifying other interrelated variables. Thus Y (the variable we want to predict) is seen to be associated with not simply a single X variable but by a combination of variously weighted Xs (multiple variables).

MVA also considers not only the relationship of independent (X) to dependent (Y) variables but also the interrelationship between independent variables. It estimates how much variance in a dependent variable can be attributed to variance in a combined set of independent variables.

Note: what MVA can do is reveal and elaborate associations between measured variables, which is not the same as identifying causes. A cause involves a constant conjunction between X (or a combination of Xs) and Y, such that whenever X (or Xs) occur Y results. In principle this requires a perfect correlation (R=1) between the identified Xs and Y. MVA, of course, is used to indicate causal factors when this condition is not obtained.

MVA cannot reveal the existence or nature of any causal links in the sense of proving them or providing the basis of causal laws at a theoretical level. MVA is a pragmatic device that suggests macro-sociological causal factors and points to possible causal combinations.

What are the Issues Related to Multivariate Regression Analysis

- Insufficient sample size can reflect inaccurate associations; a small sample size results in the model not being trustworthy
- Accuracy of the performance data has a big impact on the results of the analysis
- Measurement error in which participants provide careless or non-genuine responses (note however that the presence of measurement errors in behavioral research is the rule rather than the exception, and that the reliability of many measures used in the behavioral sciences are, at best, moderate)
- Heteroscedasticity can lead to distortion of findings and weaken the analysis
- Having too many input variables in the model may cause overfitting and an inflated variance
- Having too few input variables in the model may cause underfitting and poor explanation of the data

What is Predictive Modeling

Predictive modeling is a statistical data mining approach that builds a prediction function from observed data, such as data that might be collected from an incumbent sales team.

The function is then used to estimate (predict) the value of a dependent variable using new data.

Can Algorithms Beat People

The findings in various of fields covering decades of research and hundreds of studies, are very clear. People, including ‘lower skilled’ people or ‘beginners’ armed with smart algorithms, consistently and uniformly beat or tie experts in hundreds of skills previously believed to require decades of expertise.

Below is a quote from one of the most widely cited studies in this area.

> *On average, mechanical-prediction techniques were about 10% more accurate than clinical predictions. Superiority for mechanical-prediction techniques was consistent, regardless of the judgment task, type of judges, judges’ amounts of experience, or the types of data being combined. Clinical predictions performed relatively less well when predictors included clinical interview data. These data indicate that mechanical predictions of human behaviors are equal or superior to clinical prediction methods for a wide range of circumstances. (Source: Practical and Theoretical Implications of 85 Years of Research Findings. Hunter and Schmidt, 1998)*

Research suggest that humans are still very good at collecting data, but we tend to get less good (relative to machines) when we have to combine large volumes of data and assign weights. This is where machines generally outperform human.

Many people and organizations resist this finding. However, it is very robust and validated across decades now, in dozens of fields.

A study conduced by the National Bureau of Economic Research demonstrated that:

> *People want to believe they have good instincts, but when it comes to hiring, they can't best a computer. Hiring managers select worse job candidates than the ones recommended by an algorithm, new research from the National Bureau of Economic Research finds. Looking across 15 companies and more than 300,000 hires in low-skill service sector jobs, such as data entry and call center work, NBER researchers compared the tenure of employees who had been hired based on the algorithmic recommendations of a job test with that of people who'd been picked by a human. The test asked a variety of questions about technical skills, personality, cognitive skills, and fit for the job. The applicant's answers were run through an algorithm, which then spat out a recommendation: Green for high-potential candidates, yellow for moderate potential, and red for the lowest-rated. (Source: The Validity and Utility of Selection Methods in Personnel Psychology).*

Further Reading

Anderson, T.W. (1958). *An Introduction to Multivariate Analysis, New York*: Wiley ISBN 0471026409; 2e (1984) ISBN 0471889873; 3e (2003) ISBN 0471360910

Grimm, G.; P. Yarnold (1995). *Reading and Understanding Multivariate Statistics*. ISBN: 978-1-55798-273-5

Spicer, J. (2004). *Making Sense of Multivariate Data Analysis*. ISBN: 9781412904018

Tabachnick, B. G.; Fidell, L. S. *Using Multivariate Statistics*. ISBN: 9780134790541

Krzanowski, W. (2000). *Principles of Multivariate Analysis*. ISBN: 9780198507086

Sen, Pranab Kumar; Anderson, T. W.; Arnold, S. F.; Eaton, M. L.; Giri, N. C.; Gnanadesikan, R.; Kendall, M. G.; Kshirsagar, A. M.; et al. (June 1986). *"Review: Contemporary Textbooks on Multivariate Statistical Analysis: A Panoramic Appraisal and Critique"*. Journal of the American Statistical Association. 81 (394): 560–564. doi:10.2307/2289251. ISSN 0162-1459. JSTOR 2289251.(Pages 560–561)

Schervish, Mark J. (November 1987). *"A Review of Multivariate Analysis"*. Statistical Science. 2 (4): 396–413. doi:10.1214/ss/1177013111. ISSN 0883-4237. JSTOR 2245530.

Raykov, T. (2008). *An Introduction to Applied Multivariate Analysis*. ISBN-139780805863758

Staines, H.; Dugard, P.; Todman, J. (2014). *Approaching Multivariate Analysis.*