# Research Methods & Technical Writing

Lesson 9 - Week 9

Multivariate Analysis Techniques

Lecturer: Dr. Msagha J Mbogholi, PhD

# Flashback from Lesson 8 (1 of 2)

- A hypothesis (plural hypotheses) is a precise, testable statement of what the researcher(s) predict will be the outcome of the study. It is stated at the start of the study.
- Types of hypotheses include simple hypothesis, complex hypothesis, empirical hypothesis, statistical hypothesis, logical hypothesis, null hypothesis, and alternative hypothesis.
- Significance level: The significance level, also known as alpha or α, is a measure of the strength of the evidence that must be present in your sample before you will reject the null hypothesis and conclude that the effect is statistically significant.
- Decision rule or test of hypothesis: The decision rule is a statement that tells under what circumstances to reject the null hypothesis.

# Flashback from Lesson 8 (2 of 2)

- Type I error refers to the situation where we incorrectly reject $H_0$ when in fact it is true. This is also called a false positive result (as we incorrectly conclude that the research hypothesis is true when in fact it is not). When we run a test of hypothesis and decide not to reject $H_0$ (e.g., because the test statistic is below the critical value in an upper tailed test) then either we make a correct decision because the null hypothesis is true or we commit a Type II error.
- The 5 step approach to hypothesis testing as follows: **Step 1.** Set up hypotheses and determine level of significance, **Step 2.** Select the appropriate test statistic, **Step 3.** Set up decision rule, **Step 4.** Compute the test statistic, **Step 5.** Conclusion. (draw an inference -If the null hypothesis is rejected, then an exact significance level is computed to describe the likelihood of observing the sample data assuming that the null hypothesis is true. The exact level of significance is called the p-value and it will be less than the chosen level of significance if we reject $H_0$.)

# Content

- Introduction
- Classification
- Multivariate Techniques
- More about factor analysis

# Part 1

## Introduction

# Introduction

- So far in this course we have been analyzing data based on a single (or two) variables. You notice, for example, in the last lesson we were looking at data to determine whether to accept or reject hypothesis based on our hypothesis statement.
- This is usually the simplest form of analysis used. However, it is not always the case that we work with one set of variable. In today's world analysis is being done on multiple datasets with several variables.
- The idea is to find out how these variables are related to each other (if at all), and to make sense of the data so that we can understand what is happening (in real world situations), and consequently make informed decisions.
- More specifically multivariate techniques "are being applied in many fields such as economics, sociology, psychology, agriculture, anthropology, biology and medicine. These techniques are used in analyzing social, psychological, medical and economic data, specially when the variables concerning research studies of these fields are supposed to be correlated with each other and when rigorous probabilistic models cannot be appropriately used." (Kothari, 2004).

In order to understand the lesson better let us define a few terms in this part.

# Introduction

- The different types and categories of variables namely, categorical, ordinal, discrete, continuous, and so, have been described in previous lessons.
- The following definitions have been provided by (Sherpa, 2021):"
- Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable. There are many different ways people use univariate analysis. The most common univariate analysis is checking the central tendency (mean, median and mode), the range, the maximum and minimum values, and standard deviation of a variable.
- Bivariate analysis is where you are comparing two variables to study their relationships. These variables could be dependent or independent to each other. In Bivariate analysis is that there is always a Y-value for each X-value.
- Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables."

# Introduction

- Kothari (2004) further asserts that "multivariate techniques transform a mass of observations into a smaller number of composite scores in such a way that they may reflect as much information as possible contained in the raw data obtained concerning a research study. Thus, the main contribution of these techniques is in arranging a large amount of complex information involved in the real data into a simplified visible form."
- In this lesson we examine the classification of multivariate techniques, and describe different multivariate techniques (with examples where appropriate, for easier understanding).

# Part 2

## Classification

# 2.1 Definitions

- Before delving into the world of multivariate statistics it is important that we define a few terms that are used in the literature:
- Explanatory variable: An explanatory variable is a type of independent variable. It is what a researcher manipulates or observes changes in. In other words, an explanatory variable is the expected cause, and it explains the results. (Sirisilla, 2022). It is also known as a causal or predictor variable.
- Criterion variable: A **criterion variable** is simply another name for a *dependent variable* or a *response variable*. This is the variable that is being predicted in a statistical analysis. (Zach, 2019). Sirisilla(2022) adds, "…it is the one that changes the results. Furthermore, a response variable is the expected effect, and it responds to explanatory variables."
- For reminder purposes it is worth also mentioning the following types of variables (Kothari, 2004):
- *Observable variables and latent variables:* Explanatory variables described above are supposed to be observable directly in some situations, and if this is so, the same are termed as observable variables. However, there are some unobservable variables which may influence the criterion variables. We call such unobservable variables as latent variables.

# 2.1 Definitions

- *Discrete variable and continuous variable:* Discrete variable is that variable which when measured may take only the integer value whereas continuous variable is one which, when measured, can assume any real value (even in decimal points).
- Dummy variable: A **dummy** variable is a binary variable that takes a value of 0 or 1. One adds such variables to a regression model to represent factors which are of a binary nature i.e. they are either observed or not observed. (Date, 2022)

# 2.2 Classification

- Multivariate techniques are classified as being one of two methods: dependence and interdependence methods. Dependence methods are those where one variable depends on another; interdependence methods are applied where variables do not depend on one another.
- Other questions that help in classification is whether the number of variables exhibiting dependency, and whether the data is metric or not (quantitative or qualitative).
- Sheth (as cited in Kothari, 2004) developed a chart that shows the criteria to use to determine which multivariate technique to use depending on the nature of the data and dependence/interdependence of the variables. This is captured in fig 1.

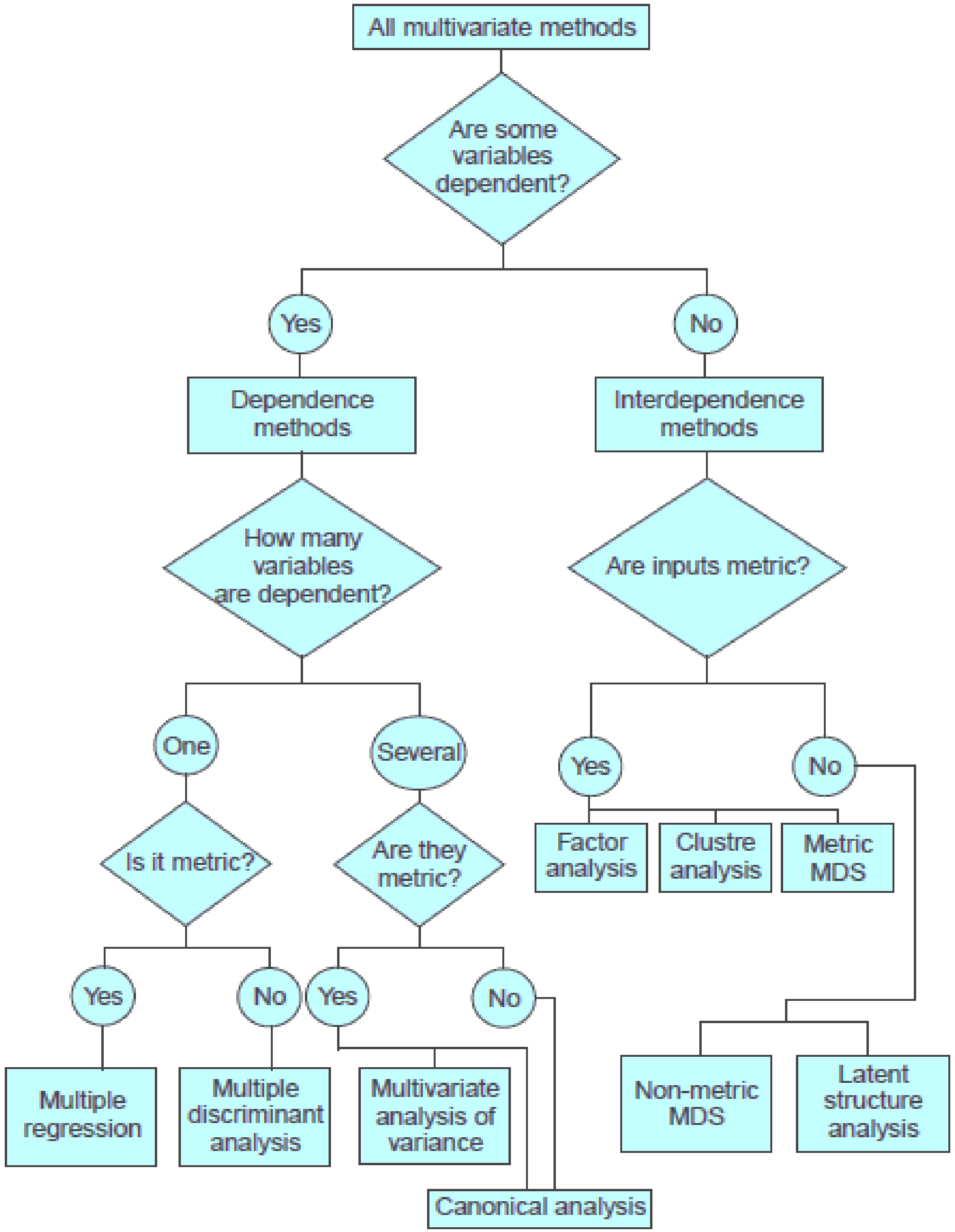




Fig 1. Multivariate technique criteria (Sheth, as cited in Kothari, 2004)

# Part 3

Multivariate techniques

# 2.1 Simple regression analysis

- Before describing multiple regression (which is the one we use in multivariate analysis), let us introduce regression analysis from a simple perspective (before you learnt to run, you learnt to crawl, then walk first).
- What is regression analysis in the first place? It is simply a statistical method used to show the relationship between two or more variables.
- Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables). (Beers, 2022).
- We normally speak of simple regression or multiple regression.
- Simple regression is also known as ordinary least squares (OLS), is a form of linear regression where we establish a linear relationship between two variables based on what is known as a line of best fit. Non-linear regression does exist too, but is outside the scope of this course.
- Let us demonstrate simple regression using an easy to understand example.

# 2.1 Simple regression analysis

- This example has been shared by (Zach, 2019):
- We may fit a simple linear regression model to a dataset using *weight* to predict the value for *height* for a group of people. In this case, our criterion (or dependent) variable is *height* since that's the value we're interested in predicting.
- If we plotted the values for height and weight on a scatterplot, the criterion variable *height* would be on the y-axis, as seen in fig 2.
- Based on the graph we can draw a line of fit and predict height for the different weights.
- Mathematically regression is shown in the form of $Y = a + bX$, where Y is the dependent variable (height in our above example), and X is the independent (predictor, in our example, weight) variable. a and b are known as the regression coefficients.

# 2.1 Simple regression analysis

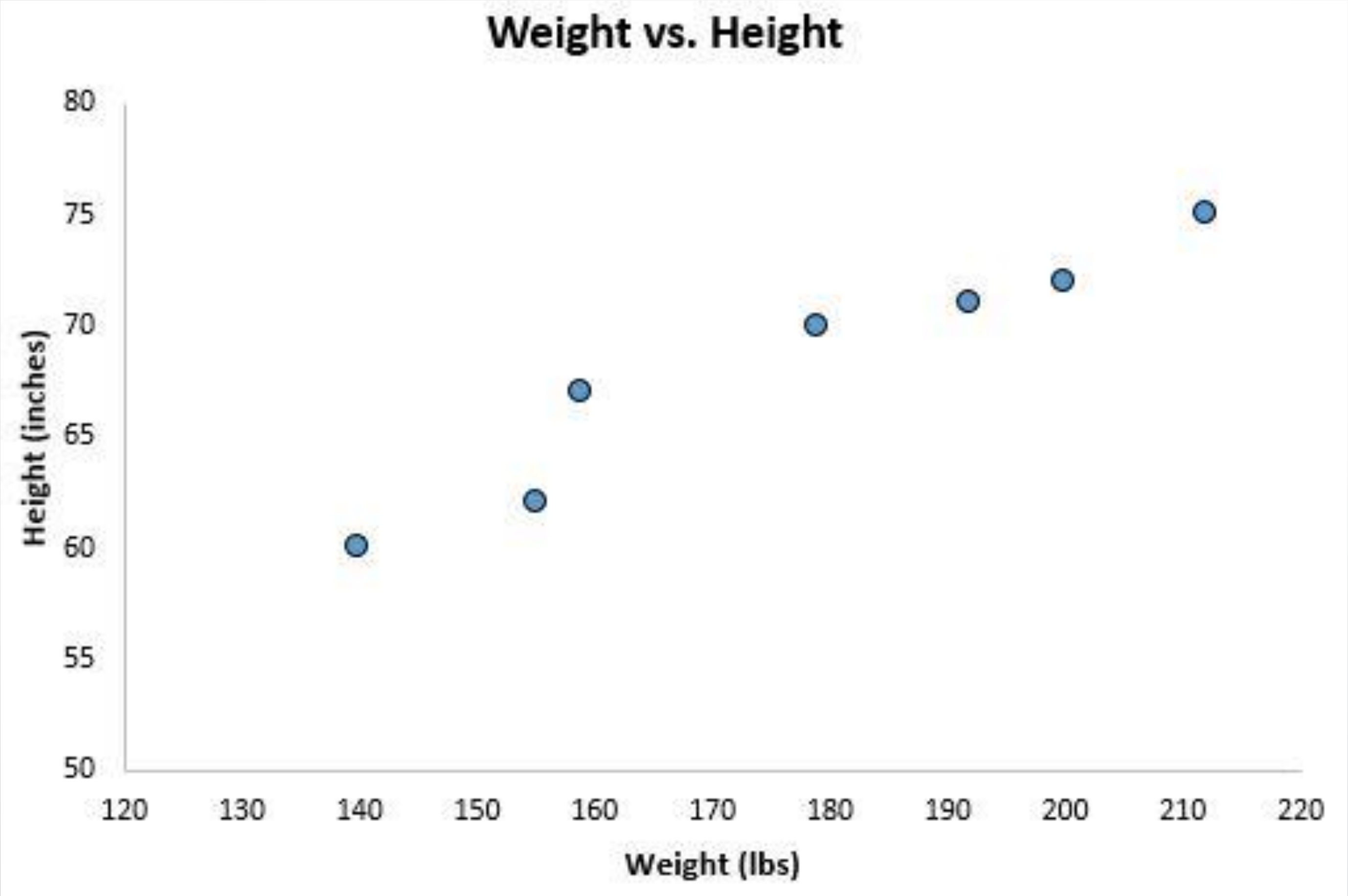


Fig 2. Regression analysis of height vs weight (Zach, 2019)

# 2.1 Simple regression

- To find the value of the regression coefficients we use the formula:
- $a = n(\sum xy) - (\sum x)(\sum y) / n(\sum x^2) - (\sum x)^2$

- $b = (\sum y)(\sum x^2) - (\sum x)(\sum xy) / n(\sum x^2) - (\sum x)^2$ where n is the number of data points in the set; for example from fig 2, n = 7.
- Therefore, given a set of data points you only need to substitute the values into the formula and you will be able to determine how your line of best fit appears. In today's digital age a scientific calculator will make your work that much easier.
- If you're interested in knowing how strongly the variables are related then you will use the correlation coefficient: **Correlation coefficients** are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. **Pearson's correlation** (also called Pearson's *R*) is a **correlation coefficient** commonly used in linear regression. (Glen, 2022). (please visit this reference for more information, as we can't cover it here in detail due to time and space constraints)

# 2.2 Multiple regression

- Consider the following example:
- A retail business needs to predict sales figures for the next month (or the dependent variable). It is difficult to know, since there are so many variables surrounding that number (the independent variables)—the weather, a new model release, what your competitors do, or the maintenance work going on to the pavement outside. (Tibco Cloud, 2023)
- This is the kind of scenario that you will face in most industry problems; a case where there is more than one independent variable. How do we go about solving such problems?
- This is where multiple (linear) regression comes in. It is used to solve equations of the form:
- $Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$, where Y is the dependent variable and X is the independent variable.

# 2.2 Multiple regression

- Five main assumptions underlying multiple regression models must be satisfied: (1) linearity, (2) homoscedasticity (homogeneity of variances in the different groups), (3) independence of errors, (4) normality, and (5) independence of independent variables. (CFA Institute, 2023)
- Let us demonstrate step by step how to solve a problem where we have 2 independent variables; we are keen to solve for the regression coefficients in this example, so we can obtain the best line of fit.
- This example is obtained from (Zach, 2019). The dataset is shown in table 1.

# 2.2 Multiple regression

Table 1. Dataset for multiple linear regression (Zach, 2019)

| y | $X_1$ | $X_2$ |
|---|---|---|
| 140 | 60 | 22 |
| 155 | 62 | 25 |
| 159 | 67 | 24 |
| 179 | 70 | 20 |
| 192 | 71 | 15 |
| 200 | 72 | 14 |
| 212 | 75 | 14 |
| 215 | 78 | 11 |

# 2.2 Multiple regression

- **Step 1: Calculate $X_1^2$, $X_2^2$, $X_1y$, $X_2y$ and $X_1X_2$.**
- These are captured in table 2

Table 2. Step 1 calculations (Zach, 2019)

| | y | $X_1$ | $X_2$ |
|---|---|---|---|
| | 140 | 60 | 22 |
| | 155 | 62 | 25 |
| | 159 | 67 | 24 |
| | 179 | 70 | 20 |
| | 192 | 71 | 15 |
| | 200 | 72 | 14 |
| | 212 | 75 | 14 |
| | 215 | 78 | 11 |
| **Mean** | 181.5 | 69.375 | 18.125 |
| **Sum** | 1452 | 555 | 145 |

| | $X_1^2$ | $X_2^2$ | $X_1y$ | $X_2y$ | $X_1X_2$ |
|---|---|---|---|---|---|
| | 3600 | 484 | 8400 | 3080 | 1320 |
| | 3844 | 625 | 9610 | 3875 | 1550 |
| | 4489 | 576 | 10653 | 3816 | 1608 |
| | 4900 | 400 | 12530 | 3580 | 1400 |
| | 5041 | 225 | 13632 | 2880 | 1065 |
| | 5184 | 196 | 14400 | 2800 | 1008 |
| | 5625 | 196 | 15900 | 2968 | 1050 |
| | 6084 | 121 | 16770 | 2365 | 858 |
| **Sum** | 38767 | 2823 | 101895 | 25364 | 9859 |

# 2.2 Multiple regression

- **Step 2: Calculate Regression Sums.**
- Next, make the following regression sum calculations:
- $\Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2 / n = 38{,}767 - (555)^2 / 8 =$ **263.875**
- $\Sigma x_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2 / n = 2{,}823 - (145)^2 / 8 =$ **194.875**
- $\Sigma x_1 y = \Sigma X_1 y - (\Sigma X_1 \Sigma y) / n = 101{,}895 - (555*1{,}452) / 8 =$ **1,162.5**
- $\Sigma x_2 y = \Sigma X_2 y - (\Sigma X_2 \Sigma y) / n = 25{,}364 - (145*1{,}452) / 8 =$ **-953.5**
- $\Sigma x_1 x_2 = \Sigma X_1 X_2 - (\Sigma X_1 \Sigma X_2) / n = 9{,}859 - (555*145) / 8 =$ **-200.375**
- This is captured in table 3.

# 2.2 Multiple regression

Table 3. Regression sums (Zach, 2019)

| | y | $X_1$ | $X_2$ |
|---|---|---|---|
| | 140 | 60 | 22 |
| | 155 | 62 | 25 |
| | 159 | 67 | 24 |
| | 179 | 70 | 20 |
| | 192 | 71 | 15 |
| | 200 | 72 | 14 |
| | 212 | 75 | 14 |
| | 215 | 78 | 11 |
| **Mean** | 181.5 | 69.375 | 18.125 |
| **Sum** | 1452 | 555 | 145 |

| | $X_1^2$ | $X_2^2$ | $X_1y$ | $X_2y$ | $X_1X_2$ |
|---|---|---|---|---|---|
| | 3600 | 484 | 8400 | 3080 | 1320 |
| | 3844 | 625 | 9610 | 3875 | 1550 |
| | 4489 | 576 | 10653 | 3816 | 1608 |
| | 4900 | 400 | 12530 | 3580 | 1400 |
| | 5041 | 225 | 13632 | 2880 | 1065 |
| | 5184 | 196 | 14400 | 2800 | 1008 |
| | 5625 | 196 | 15900 | 2968 | 1050 |
| | 6084 | 121 | 16770 | 2365 | 858 |
| **Sum** | 38767 | 2823 | 101895 | 25364 | 9859 |
| **Reg Sums** | 263.875 | 194.875 | 1162.5 | -953.5 | -200.375 |

# 2.2 Multiple regression

- **Step 3: Calculate $b_0$, $b_1$, and $b_2$.**
- The formula to calculate $b_1$ is: $[(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$
- Thus, $\mathbf{b_1} = [(194.875)(1162.5) - (-200.375)(-953.5)] / [(263.875)(194.875) - (-200.375)^2] =$ **3.148**
- The formula to calculate $b_2$ is: $[(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$
- Thus, $\mathbf{b_2} = [(263.875)(-953.5) - (-200.375)(1152.5)] / [(263.875)(194.875) - (-200.375)^2] =$ **-1.656**
- The formula to calculate $b_0$ is: $y - b_1 X_1 - b_2 X_2$
- Thus, $\mathbf{b_0} = 181.5 - 3.148(69.375) - (-1.656)(18.125) =$ **-6.867**

# 2.2 Multiple regression

- **Step 5: Place $b_0$, $b_1$, and $b_2$ in the estimated linear regression equation.**
- The estimated linear regression equation is: $\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2$
- In our example, it is $\mathbf{\hat{y} = -6.867 + 3.148x_1 - 1.656x_2}$
- **How to Interpret a Multiple Linear Regression Equation**
- Here is how to interpret this estimated linear regression equation: $\hat{y} = -6.867 + 3.148x_1 - 1.656x_2$
- $\mathbf{b_0 = -6.867}$. When both predictor variables are equal to zero, the mean value for y is -6.867.
- $\mathbf{b_1 = 3.148}$. A one unit increase in $x_1$ is associated with a 3.148 unit increase in y, on average, assuming $x_2$ is held constant.
- $\mathbf{b_2 = -1.656}$. A one unit increase in $x_2$ is associated with a 1.656 unit decrease in y, on average, assuming $x_1$ is held constant.

# 2.3 Multiple discriminant analysis

- Discriminant analysis differs from regression analysis in that "discriminant analysis is a technique that is used by the researcher to analyze the research data when the criterion or the dependent variable is categorical and the predictor or the independent variable is interval in nature." (Lani, 2009).
- Lani (2009) further adds: "The objective of discriminant analysis is to develop discriminant functions that are nothing but the linear combination of independent variables that will discriminate between the categories of the dependent variable in a perfect manner. It enables the researcher to examine whether significant differences exist among the groups, in terms of the predictor variables. It also evaluates the accuracy of the classification."
- The major distinction to the types of discriminant analysis is that for a two group, it is possible to derive only one discriminant function. On the other hand, in the case of multiple discriminant analysis, more than one discriminant function can be computed. (Lani, 2009).
- In many ways, discriminant analysis parallels multiple regression analysis. The main difference between these two techniques is that regression analysis deals with a continuous dependent variable, while discriminant analysis must have a discrete dependent variable. (Ogbogo, 2019)

# 2.3 Multiple discriminant analysis

- Let us look at an example, from (Berman, n.d.):
- The SAT is an aptitude test taken by high school juniors and seniors. College administrators use the SAT along with high school grade point average (GPA) to predict academic success in college.
- Table 4 shows the SAT score and high school GPA for ten students accepted to Acme College. And it shows whether each student ultimately graduated from college.
- For this exercise, using data from the table, we are going to complete the following task:
- Define a discriminant function that classifies incoming students as graduates or non-graduates, based on their SAT score and high school GPA.

Table 4. Scores (Berman, n.d.)

| Graduate | SAT | GPA |
|---|---|---|
| Yes | 1300 | 2.7 |
| Yes | 1260 | 3.7 |
| Yes | 1220 | 2.9 |
| Yes | 1180 | 2.5 |
| Yes | 1060 | 3.9 |
| No | 1140 | 2.1 |
| No | 1100 | 3.5 |
| No | 1020 | 3.3 |
| No | 980 | 2.3 |
| No | 940 | 3.1 |

# 2.3 Multiple discriminant analysis

- Dummy Variable Recoding
- Look at the data table above. The dependent variable (Graduate) is a categorical variable that takes the values "Yes" or "No". To use that variable in regression analysis, we need to make it a quantitative variable.
- We can make Graduate a quantitative variable through dummy variable recoding. That is, we can express the categorical variable Graduate as a dummy variable (Y), like so:
- Y = 1 for students that graduate.
- Y = 0 for students that do not graduate.
- Now, we replace the categorical variable Graduate with the quantitative variable Y in our data table. We set the value of Y equal to 1 for students who graduated; 0, for students who did not graduate.
- Table 5 shows the changes

Table 5. Changed values (Berman, n.d.)

| Y | SAT | GPA |
|---|---|---|
| 1 | 1300 | 2.7 |
| 1 | 1260 | 3.7 |
| 1 | 1220 | 2.9 |
| 1 | 1180 | 2.5 |
| 1 | 1060 | 3.9 |
| 0 | 1140 | 2.1 |
| 0 | 1100 | 3.5 |
| 0 | 1020 | 3.3 |
| 0 | 980 | 2.3 |
| 0 | 940 | 3.1 |

# 2.3 Multiple discriminant analysis

- We input data from the above table into our statistical software to conduct a standard regression analysis. Outputs from the analysis include a regression coefficients (*we will only be concerned with the discriminant function)
- Discriminant Function
- The first task in our analysis is to define a linear, least-squares regression equation to predict academic performance, based on SAT and GPA. That equation will be our discriminant function. Since we have two independent variables, the equation takes the following form:
- $\hat{y} = b_0 + b_1 SAT + b_2 GPA$
- In this equation, $\hat{y}$ is the *predicted* academic performance (i.e., whether the student graduates or not). The independent variables are SAT and GPA. The regression coefficients are $b_0$, $b_1$, and $b_2$. On the right side of the equation, the only unknowns are the regression coefficients; so to specify the equation, we need to assign values to the coefficients.

# 2.3 Multiple discriminant analysis

- To assign values to regression coefficients, we consult the regression coefficients table produced by Excel (*we skip the calculation as it is done in an excel spreadsheet*):
- Here, we see that the regression intercept ($b_0$) is -3.8392, the regression coefficient for SAT ($b_1$) is 0.003233, and the regression coefficient for GPA ($b_2$) is 0.23955. So the least-squares regression equation is:
- $\hat{y} = -3.8392 + 0.003233 * SAT + 0.23955 * GPA$
- This is the discriminant function that we can use to classify incoming students as likely graduates or non-graduates.
- Thus we see how we can compute a discriminant function using the same technique as with regression.

# 2.4 Multivariate analysis of variance (MANOVA)

- **Multivariate analysis of variance (MANOVA)** is an extension of the univariate analysis of variance (ANOVA). In an ANOVA, we examine for statistical differences on one continuous dependent variable by an independent grouping variable. The MANOVA extends this analysis by taking into account multiple continuous dependent variables, and bundles them together into a weighted linear combination or composite variable. The MANOVA will compare whether or not the newly created combination differs by the different groups, or levels, of the independent variable. In this way, the MANOVA essentially tests whether or not the independent grouping variable simultaneously explains a statistically significant amount of variance in the dependent variable. (Statistics Solutions, n.d.)
- Questions that can be answered by MANOVA include (Statistics Solutions, n.d.):
- Do the various school assessments vary by grade level?
- Do the rates of graduation among certain state universities differ by degree type?

Which diseases are better treated, if at all, by either X drug or Y drug?

# 2.5 Canonical correlation analysis

- "Canonical Correlation Analysis is one way to find associations between two data sets. Like the Correlation Coefficient, CCA measures the relationship between variables. Where Canonical Correlation Analysis differs is that it is specifically used to find the relationships between *two sets* of variables. For example, an educational researcher may want to find the association between different measures of scholastic ability and success in school.
- It's appropriate to use CCA in the same situations as you might use multiple regression analysis, but when you have multiple intercorrelated outcome variables. CCA is not recommended for small data sets.
- The purpose of Canonical Correlation Analysis is to explain the variability within and between sets through identification of several sets of canonical variates. Canonical variates are new variables formed by making a linear combination of two of more variables from the data sets. When running CCA, you choose weights that maximize the correlation between these sets of variates." (Glen, n.d.)

# 2.6 Factor analysis

- "Factor Analysis (FA) is an exploratory technique applied to a set of outcome variables that seeks to find the underlying factors (or subsets of variables) from which the observed variables were generated. For example, an individual's response to the questions on an exam is influenced by underlying variables such as intelligence, years in school, age, emotional state on the day of the test, amount of practice taking tests, and so on. The answers to the questions are the observed or outcome variables. The underlying, influential variables are the factors." (NCSS Statistical Software, n.d.)
- Important methods of factor analysis are:
  - the centroid method;
  - the principal components method;
  - the maximum likelihood method.
- We describe these briefly.

# 2.6.1 Centroid method

- The sequence of steps involve in centroid method are:
- Compute a matrix of correlation;
- Work out the sum of coefficients in each column of the correlation matrix;
- Obtain the sum (T) of columns;
- Divide the sum of each column by the square root of T.

# 2.6.2 Principal component method

- Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
- Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data points much easier and faster for machine learning algorithms without extraneous variables to process.
- So, to sum up, the idea of PCA is simple — **reduce the number of variables of a data set, while preserving as much information as possible**. (Jaadi, 2019)
- The mathematics of *factor analysis* and *principal component analysis* (PCA) are different. Factor analysis explicitly assumes the existence of latent factors underlying the observed data. PCA instead seeks to identify variables that are composites of the observed variables. Although the techniques can get different results, they are similar to the point where the leading software used for conducting factor analysis (SPSS Statistics) uses PCA as its default algorithm. (Block, 2018). An easy to follow worked example is also provided here.

# 2.6.3 Maximum likelihood method

- Maximum likelihood estimation, is a technique used in statistics to estimate the parameters of an assumed **probability distribution** based on specific observed data. This is accomplished by optimizing a likelihood function in such a way that, according to the statistical model that is being assumed, the observed data has the highest probability. (Team, 2022)

# Summary (1 of 2)

- Multivariate techniques transform a mass of observations into a smaller number of composite scores in such a way that they may reflect as much information as possible contained in the raw data obtained concerning a research study. Thus, the main contribution of these techniques is in arranging a large amount of complex information involved in the real data into a simplified visible form.
- An explanatory variable is a type of independent variable. It is what a researcher manipulates or observes changes in. In other words, an explanatory variable is the expected cause, and it explains the results. It is also known as a causal or predictor variable.
- A criterion variable is simply another name for a *dependent variable* or a *response variable*. This is the variable that is being predicted in a statistical analysis.
- Five main assumptions underlying multiple regression models must be satisfied: (1) linearity, (2) homoscedasticity (homogeneity of variances in the different groups), (3) independence of errors, (4) normality, and (5) independence of independent variables.

# Summary (2 of 2)

- Discriminant analysis differs from regression analysis in that "discriminant analysis is a technique that is used by the researcher to analyze the research data when the criterion or the dependent variable is categorical and the predictor or the independent variable is interval in nature
- In an ANOVA, we examine for statistical differences on one continuous dependent variable by an independent grouping variable. The MANOVA extends this analysis by taking into account multiple continuous dependent variables, and bundles them together into a weighted linear combination or composite variable
- The purpose of Canonical Correlation Analysis is to explain the variability within and between sets through identification of several sets of canonical variates.
- Factor Analysis (FA) is an exploratory technique applied to a set of outcome variables that seeks to find the underlying factors (or subsets of variables) from which the observed variables were generated. Common methods of FA are: Centroid method, principal component method, and maximum likelihood estimation.

# References


- Beers, B. (2022, August 18). *Regression definition*. Investopedia. https://www.investopedia.com/terms/r/regression.asp#:~:text=A%20regression%20is%20a%20statistical
- Berman, H. B. (n.d.). *Discriminant Analysis*. Stattrek.com. Retrieved May 15, 2023, from https://stattrek.com/multiple-regression/discriminant-analysis
- Block, T. (2018, August 12). *Factor Analysis and Principal Component Analysis: A Simple Explanation*. Displayr. https://www.displayr.com/factor-analysis-and-principal-component-analysis-a-simple-explanation/
- CFA Institute. (2023). *Basics of Multiple Regression and Underlying Assumptions*. Www.cfainstitute.org; CFA Institute. https://www.cfainstitute.org/en/membership/professional-development/refresher-readings/multiple-regression#:~:text=Five%20main%20assumptions%20underlying%20multiple
- Date, S. (2022, July 26). *What Are Dummy Variables And How To Use Them In A Regression Model*. Time Series Analysis, Regression, and Forecasting. https://timeseriesreasoning.com/contents/dummy-variables-in-a-regression-model/#:~:text=A%20dummy%20variable%20is%20a

# References


- Glen, S(n.d). "Principal Component Analysis (PCA), Regression & Parafac" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/principal-component-analysis/
- Jaadi, Z. (2019, September 4). *A Step by Step Explanation of Principal Component Analysis*. Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis
- Kothari, C. R. (2004). *Research methodology : methods & techniques* (2nd ed.). New Age International (P) Ltd., Publishers, Cop.
- Lani, J. (2009, October 28). *Discriminant Analysis*. Statistics Solutions. https://www.statisticssolutions.com/discriminant-analysis/#:~:text=The%20major%20distinction%20to%20the
- NCSS Statistical Software. (n.d.). *Multivariate Analysis | Factor Analysis | PCA | MANOVA | NCSS*. Www.ncss.com; NCSS Statistical Software. Retrieved May 15, 2023, from https://www.ncss.com/software/ncss/multivariate-analysis-in-ncss/#FA

# References


- Ogbogo, S. (2019). Discriminant Analysis: An Analysis of Its Predictship Function. *Journal of Education and Practice* , *Vol.10*(5). https://doi.org/10.7176/JEP/10-5-04
- Sherpa, D. (2021, March 16). *Univariate, Bivariate and Multivariate Analysis*. Analytics Vidhya. https://medium.com/analytics-vidhya/univariate-bivariate-and-multivariate-analysis-8b4fc3d8202c
- Sirisilla, S. (2022, October 17). *Explanatory Vs Response Variables | Definitions & Examples*. Enago Academy. https://www.enago.com/academy/explanatory-response-variable-statistics/
- Statistics Solutions. (n.d.). *MANOVA*. Statistics Solutions. Retrieved May 15, 2023, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/manova/
- Team, W. (2022, November 1). *Factor Analysis*. WallStreetMojo. https://www.wallstreetmojo.com/factor-analysis/#types

# References


- Tibco Cloud. (2023). *What is Regression Analysis?* TIBCO Software; Cloud Software Group, Inc. https://www.tibco.com/reference-center/what-is-regression-analysis#:~:text=Regression%20analysis%20is%20a%20statistical
- Zach. (2019, May 24). *What is a Criterion Variable? (Explanation + Examples)*. Statology. https://www.statology.org/criterion-variable/#:~:text=A%20criterion%20variable%20is%20simply
- Zach. (2020, November 18). *Multiple Linear Regression by Hand (Step-by-Step)*. Statology. https://www.statology.org/multiple-linear-regression-by-hand/