WEEK 1 PAPER

Concepts of Statistics 2 – DATA-51200 | Spring 2 2020

Christina Morgenstern

My Chapter 1 summary of Multivariate Data Analysis by Joseph F. Hair Jr, William C. Black, Barry J. Babin and Rolphe E. Anderson, 7th edition, Pearson.

We are living in the era of "big data" with vast amounts of data being constantly generated and from which knowledge is drawn and decisions are inferred. Multivariate statistical techniques are statistical techniques that allow for analyzation of multiple variables simultaneously and are applied in this area in order to generate new insights into data. "Old statistics" has been able to deal with single variables in univariate analysis or with two variables in bivariate analysis, multivariate statistics extends the "old statistics toolbox" and makes it applicable to more than two variables. To make multivariate analysis work, the variables must be randomly distributed and in such a way related to each other that no individual interpretation thereof is meaningful.

The focal point of interest in multivariate analysis is the **variate**, a combination of a set of variables with empirically determined weights. In applying multivariate techniques, the goal is to identify and measure variation within a set of variables or between a dependent variable and one or more independent variables.

As a researcher, one has to get acquainted with and understand the data at hand in depth. One property of the data that needs to be understood is the type of data which can be measured in metric or nonmetric scales. Whereas metric variables differ in the amount or degree of a particular attribute, nonmetric variables describe differences in discrete properties. Nonmetric data can be further divided into categorical (or nominal) and ordinal types with the latter one being able to be ranked and the former one providing numbers with no quantitative meaning. Metric data can further be divided into interval and ratio types. The difference between the two being that the latter uses an absolute zero point whereas the former using an arbitrary zero point. In general, mathematical operations cannot be applied to nonmetric data whereas we can use mathematical operations to interpret metric data.

When measuring the value of variables in our analysis, we can be certain that these come with some **measurement error** which can be due to missing data, imprecise measurements or the incorrect information obtained from respondents. This is "noise" that gets added to the observed measured variables and needs to be taken into account. **Validity** and **reliability** are two important characteristics that must be addressed in this context. The former represents the degree to which a measure accurately represents the true value and the latter describes the degree to which the observed variable measures the true value.

In multivariate measurements we are dealing with summated scales which means that several variables which are representing different facets of one concept serve as indicators.

With the exception of cluster analysis and perceptual mapping, all multivariate techniques are based on **statistical inference**, the process of drawing conclusions about populations from noisy data. The noise can be considered as statistical error resulting from a sample (thus also termed **sampling error**) leading to differences between the obtained value and the true value. Generally, we are dealing with two main types of statistical error: Type I error (also alpha) states the probability of rejecting the null hypothesis when it is actually true and thus is an indicator of false positives. Type II error (also beta) states the probability of not rejecting the null hypothesis when it is actually false representing false negatives. Ideally, one would like to minimize both error rates however in reality it is a trade-off between the two and decreasing one type of error results in increasing the other type of error.

An extension to the Type II error is the notion of statistical power (1-beta) which states the probability that the test correctly rejects a false null hypothesis. In order to estimate the power of a statistical test, one needs to have information about three further pieces: 1. Effect size, a quantified measure of a population result, 2. Significance (alpha), the level of significance used in the test, which is often set to 0.05 and 3. The sample size, representing the observations in a sample. Since we cannot control effect size because it comes directly from the data (e.g. variance, mean, etc.) we can adjust alpha and sample size. To obtain a higher power in our statistical test we can thus increase alpha and sample size. However, we have to be cautious as we are increasing the number of false positives.

When choosing a specific multivariate technique to analyze your data, a few questions aid in the decision process such as the variables being independent or dependent as well as the number of dependent variables and the measurement scales of the variables. Broadly, **dependence techniques** are applied if a dependent variable is being predicted by other independent variables. **Interdependence techniques** are used when variables cannot be classified as either dependent or independent which makes an analysis of all variables simultaneously necessary.

Various multivariate methods have been described. The following lists and very briefly describes the most common techniques.

I. Dependence multivariate techniques

1. Multiple Regression: dependent and independent variables are metric and predict one single metric value
2. Multiple Discriminant Analysis: used to classify instances into one or more nonmetric categories based several, metric independent variables
3. Logistic Regression: predicts nonmetric categorical variable based on several, metric independent variables.
4. Multivariate Analysis of Variance (MANOVA) and Covariance (MANCOVA): categorical variables are used to predict a metric variable.
5. Cojoint Analysis: applied in product research where a product is broken down into component attributes for analysis of decision making and prediction of future decisions.
6. Canonical Analysis: determines the relationships between groups of metric variables in a data set.
7. Structural Equation Modelling and Confirmatory Factor Analysis: for the analysis of structural relationships

II. Interdependence multivariate techniques

1. Principal Component and Common Factor Analysis: data reduction techniques, with PCA being a linear combination of variables and CFA a measurement model of latent variable.
2. Cluster Analysis: classifies objects into similar groups (clusters)
3. Multidimensional Scaling / Perceptual Mapping: aids in identifying key dimensions in underlying customer evaluations of e.g. products. Perceptual maps are used to visualize relative positioning of all product.
4. Correspondence Analysis: summarizes a nonmetric data set in two-dimensional graphical form.

Although all these techniques work in different ways, for validation of the results the following strategies can be employed:

1. Split your sample (in training and test sets)
2. Gather a separate sample for validation
3. Employ a bootstrapping technique i.e. split your sample in min samples, perform the analyses and assess the weights

In general, the following steps are recommended in pursuing multivariate statistical analysis:

1. Define the research problem in conceptual terms including the objectives, the multivariate technique, the fundamental relationships.
2. Design an analysis plan representing sample size, variables, estimation metrics etc.
3. Evaluate your assumptions
4. Estimate the multivariate model and assess overall model fit
5. Interpret the variate
6. Validate your multivariate model

References:

[1] Multivariate Data Analysis by Joseph F. Hair Jr, William C. Black, Barry J. Babin and Rolphe E. Anderson, Pearson, 7th edition, 2010