Joseph Griffin

DATA 512

Assignment 1

3/21/2021

First of all, multivariate data analysis is important to study because it can be applied to almost every industry and facet of life. Furthermore, it is important to study because there are almost always multiple variables at play in a given situation in the real world. When discussing this topic in his office hours, I really liked what Dr. Al-khassaweneh said: “The goal is to capture as much of reality as possible.”

Getting into the basics of multivariate data analysis, the variate refers to a linear combination of variables with empirically determined weights. Weights are a crucial factor as certain elements obviously have greater influence on the dependent variable than other elements. Weights should be calculated so as to produce an equation that resembles reality. Next, the data can be categorized by a couple different elements. It can be first categorized as either nonmetric/qualitative or metric/quantitative. This can be further divided into subcategories. Within qualitative, we can have nominal or ordinal scales, and within quantitative, there are interval or ratio scales.

One topic to always consider is measurement error. There are two main characteristics of measurement error: validity and reliability. Validity refers to how accurately a measurement represents what it should, and reliability refers to how true the measurement is. Another way to think of reliability is whether or not you continue to get the same results if you repeat the measurement multiple times.

In statistics, an important distinction to make has to do with hypotheses. If our data does not support the hypothesis, we say we reject the hypothesis. However, when the data does in fact support the hypothesis, we then say that we fail to reject the hypothesis -- not ‘accept’ the hypothesis. We say fail to reject rather than accept because it is impossible to have data on the entire population. Thus, the data in front of us supports the hypothesis, but perhaps the results would be slightly different if there was a greater sample size or if it was somehow possible to analyze the entire population.

Next, we must consider the types of scenarios that can occur in terms of rejecting or failing to reject the null hypothesis. Type I (alpha) and type II (beta) are the two types of possible errors. A type I error refers to rejecting the null hypothesis when it is actually true, and a type II error refers to failing to reject the null hypothesis when it is actually false. On the other hand, there are two possibilities for non-errors: failing to reject the null hypothesis when it is true in reality (1-alpha), and rejecting the null hypothesis when it is false in reality (1-beta, also known as power). In reality, the null hypothesis can only be true or false, and our prediction can either be correct or not. This is why for each possibility, alpha and beta must add up to 1. The total probability of us being right or wrong for each has to be 100%, or 1. The power (1-beta) is determined by three factors: effect size, alpha, and sample size. All three of these factors have direct relationships with power -- meaning when these factors increase, so does the power. However, it is good to note that the effect size cannot be directly modified.

There are a few other things that are important when considering power levels in practice. In general, researchers should achieve a power level of at least 0.8 at the desired significance level. This can be difficult with small significance levels. The best way to increase the power level is through larger sample sizes, but lowering the alpha can also help.

Also, as a side note: thank you Dr. Al-khassaweneh for the guilty / not guilty metaphor, that was really helpful!

There are a number of techniques that can be used to perform multivariate statistical analysis. There are two main types of techniques: dependence and independence. Dependence considers how one (dependent) variable is impacted by other (independent) variables, while interdependence does not identify specific independent or dependent variables, and considers all variables simultaneously. Dependence techniques can also be subdivided into methods for one or multiple dependent variables. And to further break it down, there are certain methods for each depending on whether the data is metric or nonmetric.

Some examples of dependence techniques include multiple regression, multiple discriminant analysis, logit/logistic regression, multivariate analysis of variance and covariance, conjoint analysis, and canonical correlation. Interdependence techniques include principal components and common factor analysis and cluster analysis.

When deciding which multivariate techniques to use, the first question to ask is whether the data has an dependent or interdependent relationship. If just one variable is being examined, then dependence is the way to go. If you have multiple variables but they are dependent variables, then dependence is still the way to go. If you are looking at the relationships between multiple variables, then interdependence would be the right option.

In general, there are some good things to keep in mind when performing multivariate analysis. It is important to decide on both practical and statistical significance, remember the significant impact of sample size, be familiar with your data, use the right model, consider any errors, and don’t forget to validate the results.