Israel Nolazco
Concepts of Statistics 2
SU20-Data-51200-002
July 12, 2020

Multivariate Data Analysis

As a Data Science student, we are constantly reminded that a Data Scientist is nothing more than a Mathematician with strong computational knowledge, a Computational Scientist with strong Mathematical knowledge, and an individual with acute business insight within their designated field of work. However, a simpler definition would be, someone who can do two things at the same time. Because Data Scientist is solving problems whilst analyzing data; so, what is the secret? The answer, multivariate analysis. Multivariate analysis is any statistical technique that analysis multiple measurements on individuals/object under investigation. Thus, “doing two things at the same time” parallel.

Before, learning about multivariate methods, I often approach problems with a narrow mindset. I was asked a question, I pull the data, do some math to solve the problem, turn in my answer, and boom done. I would never look back and think twice about the work I submitted. Even after learning about Statistics, Statistical Programming, and Data Mining, I would treat the problem as a regular math problem. Find the answer and move on. I sure learned about what TYPE I & TYPE II errors were. I learned about the cluster analysis, I learned about quantitative vs non-quantitative data; I knew they were all important, but never bother to see the bigger picture. Maybe, it was the years off of academic work that made me lazy. Because in the real world solving/finding “X” does not mean the problem is fixed, in fact, it only means there is more to come.

In this single and first chapter, I feel like I finally come to realize how all of the knowledge I compile from the very beginning of my academic career comes together. Additionally, this chapter narrows it down to a six-step procedure.

1. Establish Practical Significance as well as statistical significance.

This step is the most crucial because it verifies the need for applying and interpreting multivariate analysis. While analyzing a problem, the very first step is to recognize what type of problem is it? What are we trying to understand? Will analyzing the problem bring forward new insight? Does the data analysis need special mathematical analysis, or will a simple algebraic equation suffice?

2. Recognize that sample size affects all results

After recognizing the need for multivariate analysis, the following question should be the sample size. This is where you got to think like a Statistician because you do not want to have a too-small sample size that cannot be generalize. Or, a sample size too big, where the analytical research gets clouded the sheer volume of the data. Additionally, a Statistician must recognize the sample size also affects the TYPE I and TYPE II errors.

3. Know your data

This step can be a bit confusing. Because, in theory there are two types of data, metric and nonmetric. However, within both types of data, there are sublevels and the book best explain it as the following. Additionally, both metric and non-metric can be further analyzed by inner changing their scalar value. i.e. True/False with 1/0.

Nonmetric

Nominal – size of number is not related to the amount of the characteristic being measured

Ordinal – larger numbers indicate more (or less) of the characteristic measured, but not how much more (or less).

Metric

Interval – contains ordinal properties, and in addition, there are equal differences between scale points.

Ratio – contains interval scale properties, and in addition, there is a natural zero point.

4. Strive for model parsimony

This step is, to look for the simplest way to explain a model with the smallest number of parameters.

5. Look at your errors

As explained in step 2. The sample size affects all results, it is unavoidable to run and create errors. However, this does not mean we cannot be prepared for such eventuality. The best way to approach this step is by recognizing the Validity of the research conducted. Or, the degree to which measure accurately represents what it is supposed to. And, the Reliability of the research conducted. Or, the degree to which the observed variable measures the "true" value (how close are we to the truth?).

6. Validate your results

After conducting a multivariate analysis, it is crucial to analyze the results and not only gain knowledge from the research sample, but rather from the representative population the sample was derived from. In other words, do not over generalize/under generalize results.

Multivariate analysis is a wonderful approach to drive meaningful interpretations to complex data, but in order to have a successful analysis one must stop thinking like mathematician and start thinking like a scientist. Additionally, while these steps may seem linear, they are rather continuous. This means the steps may repeat in a cycle until the desire conclusion is approach. Or, will instead reveal hidden meaning behind the data which in turn will reset the steps once more.