1. Preface

In the upcoming some 40 summary papers I will demonstrate a comprehensive view of Applied Multivariate Statistical Modeling. First, I will start with a thorough introduction of AMSM. Then, I will explain the univariate descriptive statistics, sampling distribution, estimation, in addition to hypothesis testing. After that, I will do a comprehensive review of multivariate descriptive statistics, the normal distribution of it, and the inferential statistics. Having we accomplished that, it will be the time to discuss some various models: ANOVA, MANOVA, Multiple Linear Regression, and Multivariate Linear Regression. Furthermore, we will discuss, Principal Component analysis, Factor Analysis, and Cluster Analysis. At the end of this series of summaries, some intro to structural equation modeling (SEM), and correspondence analysis will be discussed. Prerequisite skills are, of which readers must have, basic knowledge of statistics and probability, in addition to some advanced knowledge of linear algebra. I have published summary papers in both disciplines, see the reference page.

2. Introduction

When I was a kid, I used to visualize everything in order to improve the speed of understanding. This probably helped me to understand, quite faster than other kids, rigid subjects such as mathematic and physics. Having I shared that, I will be using the exact same method throughout all upcoming summary papers. To this end, first let's try to anatomize the term, in which known as, Applied Multivariate Statistical Modeling. Well, what is “**Applied**”, “**Multivariate**”, “**Statistical**”, and “**Modeling**”, each separately?

- **Applied:** as you may probably know that science can be seen from two different windows: Pure science, and applied science. Pure science is an introvert attitude. It's depending on deduction from demonstrated truths, such as math or logic. In other words, it's used to develop information to explain phenomena in the natural world. it's linked to laws, theories and their developments (Borko, 1968). On the other hand, applied science is an extrovert attitude. in simplest form, it's the use of knowledge from the pure science to develop something for the benefit of mind. It is a discipline that is used to apply existing scientific knowledge to develop more practical applications, for example: technology or inventions (Borko, 1968).

Let me demonstrate these definitions in a better way. We all know something in basic statistic named “*normal distribution*”. Its probability density function formula is $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$. **This is the “pure science”.**

Now, the question is that how I can benefit from this knowledge in a real-life problem. Let's assume that I have a factory which produce a product named “**prod**”. And let's assume that I have three different machines “m1”, “m2”, and “m3”. The three machines produce the same product “**prod**”. Let's say that the management wants to figure out if there is or are different/ differences between the three machines in producing the final product. Now *the applied science makes sense to answer their question*. First, the management have identified three variables of the product “**prod**” to be examined. Let's say that “**prod**” have “**variable 1= the height in cm**”, “**variable 2= the thickness in mm**”, and finally “**variable 3= the weight in g**”. now let's say I will study each

variable separately for each machine using the knowledge of normal distribution (i.e. pure science). Then I could develop some figure as below:

Now let's assume that I'm studying *whether the distribution of the mean and the variance* for the first variable "**variable 1= the height in cm**" are the same in "**m1- the red color**", "**m2- the green color**", and "**m3- the blue color**" or there are any discrepancies. Same I can investigate for the second and third variables. Or probably in large scale, I can investigate the three variables at once. (i.e. simultaneously). **That's the applied science.**

- **Multivariate:** first we must understand a fundamental term, that is "variable". A variable is a symbol on whose value a function, polynomial, etc., depends. For example, the variables in the function $f(x,y)$ are x and y . We can categorize variables as "**fixed, or determined such as months, days, fess, etc.**" and "**random, or probabilistic such as any value for a variable**". See in the dataset below, assume we have " p " number of variables. As our previous examples we had "**variable 1= the height in cm**", "**variable 2= the thickness in mm**", and finally "**variable 3= the weight in g**". However, let's assume the list of variables is extended to number " p " of variables. Let's assume that I have started some observations on my production line, in other words I'm observing the bulk of product "**prod**". Thus, the dataset below suggests that, when I observed "**prod 1**", I found that **the value** of the first variable " $X_1 = x_{11}$ ", and for second variable " $X_2 = x_{12}$ ", all the way till last variable " $X_p = x_{1p}$ ". Same for "**prod 2, 3, ...**" till "**prod n**" where I have observed first variable " $X_1 = x_{n1}$ ", and for the second variable " $X_2 = x_{n2}$ ", all the way till the last variable " $X_p = x_{np}$ ". Items to mention:
 1. **Prod 1,2, ..., n**, they are **the cases** (i.e. instances) that I'm observing
 2. **V 1,2, 3,...,p**, they are **the variables** that I'm measuring on each case,
 3. x_{ij} where $i=1,2,..., n$ and $j=1,2,3,...,p$, this is the random value that can be assigned to any variable,
 4. for any "variable V , where $v=1,2, 3,...,p$ ", it can be any random value " x_{ij} " observed,
 5. the horizontal red circle we draw named "**one observation on multiple variables for a case**".
 6. the vertical red circle we draw named "**one observation on multiple cases for a variable**".

From here we may define a **variate** as a generalization of the concept of a **random variable** that is defined without reference to a particular type of probabilistic experiment. It is defined as the **set of all random variables** that obey a given probabilistic law. It is common practice to denote a variate with a capital letter (most commonly **X**). The set of all values that **X** can take is then called **the range**, denoted R_X (Forbes, Evans, Hastings, & Peacock, 2011). Specific elements in the range of **X** are called quantiles and denoted **x**, and the probability that a variate **X** assumes the value **x** is denoted $P(X=x)$. A function having a single variable (i.e. X_1) is said to be **univariate**, one having two variables (e.g. X_1 and X_2) is said to be **bivariate**, and one having two or more variables (e.g. $X_1, X_2, \dots, X_p$) is said to be **multivariate**. In other words, multivariate means a linear combination of variables as well as multiple observations.

	X_1	X_2	X_3	..	X_p
prod 1	x_{11}	x_{12}	x_{13}	..	x_{1p}
prod 2	x_{21}	x_{22}	x_{23}	..	x_{2p}
..	..	..	..	..	..
prod j	x_{j1}	x_{j2}	x_{j3}	..	x_{jp}
..	..	..	..	..	..
prod n	x_{n1}	x_{n2}	x_{n3}	..	x_{np}

- **Statistical:** it is a kind of discipline or a field of study dealing with collection, organization, analysis, interpretation, and presentation of data.
- **Modeling:** first let's try to get the meaning of **model** as a term. a model is a representation of an idea, an object or even a process or a system that is used to describe and explain phenomena that cannot be experienced directly. Models have a variety of uses – from providing a way of explaining complex data to presenting as a hypothesis. There may be more than one model proposed by scientists to explain or predict what might happen in particular circumstances. Model mimic reality and explain regularity. It can be a law or theory.

Modeling is the process of building the model. In other words, Scientists start with a small amount of data and build up a better and better representation of the phenomena they are explaining or using for prediction as time goes on. In some situations, models are developed by scientists to try and predict things. The best example is climate change. Humans don't know what effect they are having on the planet, but we do know a lot about carbon cycles, water cycles and weather. Using this information and an understanding of how these cycles interact, scientists are trying to figure out what might happen. You may see at the end this formula:

Data = Pattern + Error. It's clear here that the pattern is the regularity that I'm trying to capture from a set of data. Of course, plus the margin of error that might occur while doing so.

3. Why Applied Multivariate Statistical Modeling

Well, answering this question needs a real-life example. In the reference page you will find a reference to a dataset which I have uploaded. The dataset is a fictitious data set illustrating a manufacturing

product which I named “Prod”. Let’s assume that I manufacture a product name “prod” and this product in nature has got certain attribute to describe its functionality and behavior. **The attributes or in other words the variables that describe “prod” are variable one “the diameter”, variable two “the depth”, variable three “the length”, variable four “the weight”, variable five “the curve”, and variable six “the resistance.** Now to keep the discussion as simple as possible let’s say that each of these attribute/variables has a kind of measurement.

Let’s assume that one day I have decided to monitor the output of a bulk of “Prod” product. So, let’s say that in a random selection I have selected a sample of “**308**” pieces of “**prod**”. But the question is that, what is the purpose? Well, let’s assume that I have received a complain that “**prod**” resistance attribute/variable is vary from one to other pieces in a large scale. Let’s say that according to “prod” specification the resistance attribute/variable should be within a certain range. To this end, I have decided to investigate some issues:

1. How much it varies the value of resistance attribute/variable form one piece to another as well as the curve variable?
2. Are other attributes/variables affect the resistance attribute/variable and the curve variable?
3. If yes, which of these other attributes is most likely to be influential?
4. And finally, will I be able to design a model where can predict the resistance value of “prod” as well as the curve value, by knowing the values of other attributes?

Know answering these questions is the core of what so called “AMSM”. You see that I might be able to investigate one variable at a time for example “the depth”, or the diameter. Or maybe two of them at a time. However, if my aim is to investigate them all and try to understand the behavior of one of these attributes based on the values from other attributes, then the multivariate analysis would be the solution.

4. Essentials

Variables and data types are essential knowledge that you have to understand before we navigate further. In this section I will do brief the definition of both knowledge.

- **Type of Variables**, we can view variables as two types:
 1. **Discrete**, that can only take on a finite number of values. All qualitative variables are discrete. Some quantitative variables are discrete, such as performance rated as 1,2,3,4, or 5, or temperature rounded to the nearest degree.
 2. **Continues**, that can take on an infinite number of possible values. Let’s say an example the height of a person or the time measurement of something.

Please note, you don’t have to stuck here and try to find if this variable is discrete or continues. Just try to grasp the concept and as time pass, you will be able to find our naturally.

- **Types of Data**, I believe it’s an essential skill to type your data and to understand each type in a better way.

1. **Nominal**, in a very simple way to say, the nominal data is just a name. if you are well, consider it as odorless, and tasteless. It's a label, for example (eye color, gender, race, etc.). **(NO ARITHMATIC OPERATIONS, no +, -, *, or /)**
 2. **Ordinal**, here it's as well a label, however this label has got an order. For example, survey is asking your rank of a service (bad, not bad, good, very good). See here are labels/names however choosing one of them has got an order and that affect the result. That's a scale. **(NO ARITHMATIC OPERATIONS, no +, -, *, or /)**.
 3. **Interval**, it's same as ordinal somehow. However, in interval measurement the distance between attributes does have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values is interpretable. Because of this, it makes sense to compute an average of an interval variable, where it doesn't make sense to do so for ordinal scales. In addition, that zero doesn't hold any value. But note that in interval measurement ratios don't make any sense - 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large). **(ALL ARITHMATIC OPERATIONS "+, -, *" are allowed but no DIVISION "/)**.
 4. **Ratio**, in ratio measurement there is always an absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. In applied social research most "count" variables are ratio, for example, the number of clients in past six months. Why? Because you can have zero clients and because it is meaningful to say that "...we had twice as many clients in the past six months as we did in the previous six months.". **(ALL ARITHMATIC OPERATIONS are possible to use "+, -, *, and /")**.
- **Source of Data**, there are may sources of data, however we may consider three important sources of data as:
 1. **Primary**, from the source where it's generated.
 2. **Secondary**, stored in a repository or collected by someone else,
 3. **Tertiary**, common knowledge or general reference for example Wikipedia,

5. Statistical Approach

From modeling bullet in the introduction section above, we have written a distinguished formula **"Data = Pattern + Error"**. The question is what an approach I can follow to extract **the pattern from a data-set**. Here's a circle approach could help to address such an issue. **First**, Identify the practical problem. **Second**, overturn it into a statistical problem. **Third**, introduce the statistical solution, and finally, **fourth**, apply it as a practical solution.

Note: every problem has its own nature. If you want to become good at multivariate analysis, then first you **MUST** identify the practical problem. Then sort out the statistical problem as we will do in the following section. Noteworthy, there are many techniques in multivariate statistical modeling (e.g. classification, regression, clustering, etc.), BUT we will not consider any technique at this step.

Having you sorted out the statistical problem, it's the time where you can work on assuming the relationship between the variables that you have identified. In other words, you will find out the best technique to capture the variability between the variables, and what cause this variation. That will help you to manifest the way you understand this discipline. That lead to the final design of the model of which you are building. This model will be nothing but the MINING MACHINE that will extract the pattern. Let's now discuss each step in thorough.

If you find this note is not clear enough, please keep calm and patient and follow till the end. I can assure you a positive outcome.

1. **Practical Problem:** let's have the example in the "why AMSM" section reviewed again. Assume a company produce a product "**Prod**" in a large quantity. And as we stated in "why AMSM" section the manager is interested to find out answers for following questions:
 - How much it varies the value of resistance attribute/variable form one piece to another as well as the curve variable?
 - Are other attributes/variables affect the resistance attribute/variable and the curve variable?
 - If yes, which of these other attributes is most likely to be influential?
 - And finally, will I be able to design a model where can predict the resistance value of "prod" as well as the curve value, by knowing the values of other attributes?

The process of writing the problem in wordy context is the definition of practical problem.

2. **Statistical Problem:** having we identified the practical problem in words, we can now convert these words into statistical format.
 - **First identify variables of interest:** we have identified 6 variables as **variable one "the diameter"**, **variable two "the depth"**, **variable three "the length"**, **variable four "the weight"**, **variable five "the curve"**, and **variable six "the resistance"**.
 - Identify responses variables: $Y_1 = \text{resistance}$, $Y_2 = \text{curve}$
 - Identify explanatory variables: $X_1 = \text{diameter}$. $X_2 = \text{depth}$. $X_3 = \text{length}$, $X_4 = \text{weight}$
 - Find our dependent relationship: $Y = f(x)$.
3. **Statistical Solution**, following the same example, we can now assume a technique from statistical domain.
 - **Variability in** $Y_1 = \text{resistance}$, $Y_2 = \text{curve}$
 - **Caused by** $X_1 = \text{diameter}$. $X_2 = \text{depth}$. $X_3 = \text{length}$, $X_4 = \text{weight}$
 - There maybe a "**Linear relationship**".

That's the point, see here I have considered a technique. In other words, I have assumed a relationship. This assumption fits this problem. Probably for other problems, we may assume other techniques. Following section, I will brief some of these available techniques and their functionality.

4. **Practical Solution**, that's the last step. As I have assumed the technique, then I will build the model and test its accuracy. Please note, a model is a model if and only if (iff) it solves a problem and

predict something. Otherwise, it's just a descriptive chart. See the purpose of multivariate modeling below.

6. The Purpose of Multivariate Modeling

Three purposes for multivariate modeling. Note, if one of these steps are missing, your final work **IS NOT A MODEL**.

- Description, to describe the problem and the details that caused that problem
- Explanation, to explain the relationship between the involved variables (i.e. dependence or independence relationship).
- Prediction, to predict future outcome.

7. Multivariate models

We may consider two categories of multivariate models:

- Dependence Models
 1. MANOVA
 2. Multiple and Multivariate Regression
 3. Path Model
 4. Structural Equation Modeling
 5. Logistic Regression
 6. Discriminant Analysis
- Interdependence Models
 1. Principal Component Analysis
 2. Factor Analysis
 3. Cluster Analysis

Upcoming Summary papers, I will keep working on the same dataset, and discuss various practical problems where I can use as much as possible these models.

Closing

In this summary paper, I have introduced a thorough introduction about Applied Multivariate Statistical Modeling. First, I have explained each term in "AMSM", separately, to help readers make sense of the domain of the study. After that, I discussed why should we care to learn "AMSM". Following by essential knowledge about this domain such as type of variable, and data. After that I introduced a statistical approach to extract pattern from data. Following the three purpose of multivariate modeling. And finally, the common multivariate models. Stay tuned till next summary paper.

- Borko, H. (1968). Information science: What is it? *American Documentation*, 19(1), 3–5.
<https://doi.org/10.1002/asi.5090190103>
- Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions*. Wiley. Retrieved from <https://www.wiley.com/en-us/Statistical+Distributions%2C+4th+Edition-p-9780470390634>
- Maiti, J. (1999). An investigation of multivariate statistical models to evaluate mine safety performance. Unpublished Ph. D. thesis, Department of Mining Engineering, Indian Institute of Technology, Kharagpur.
- Dahman, M. R. (2018, October 11). Applied Multivariate Statistical Modeling. Retrieved from osf.io/3x76a
- Dahman, M. R. (2016). AoDS Solution. Retrieved from <https://bizmodule.blogspot.com/>