Session 5: ANOVA, MANOVA, and ANCOVA

ANLY 608

Amalesh Sharma

What is multivariate statistical analysis?

- **Univariate analysis:** analysis of only one variable. For example, number of footfalls in the mall, incidence of a disease, breakdown of a machine.
- **Bivariate analysis:** two variables are at work. For example, rice production and amount of rainfall, incentives and employee attrition. Remember the relationship between X and Y in a linear regression model.
- **Multivariate analysis:** allows a host of variables.
  - For example, imagine the breakdown of a machine or frequency of malfunction. It could be dependent on certain factors, which are easily acceptable, and others, which are not easily acceptable.
  - Factors like incidence, age distribution, time and pressure on the machine can be accounted for more easily when compared to mechanical faults, wrong loads, repeated running, etc.
  - Note that this analysis includes techniques to solving problems where more than one dependent variable is analyzed simultaneously with other variables (this is critical, such as simultaneous equations modeling, seemingly unrelated regression (SUR), corrected models, etc.)
- Defining MVSA
  - An approach that allows multiple measurements on each unit (individual, machine), and where relationships among multiple variables and their structure are critical.
- Merits and Demerits
  - Conclusions drawn are more accurate (because of multiple variables), realistic, and nearer to real-life contexts.
  - Required complex calculation to reach to satisfactory conclusions. Time consuming and process intensive (in data collection).

Selection of Appropriate Multivariate Technique

- Depends on two criteria
  - Is there a clear demarcation of DV and IVs? If yes, how many variables are treated as dependent in a single equation (i.e., if one DV then a linear regression, if two DVs, SUR, etc.).
  - How are variables measured?
- Depending on first criteria, multivariate analysis can be classified into two methods: *dependence methods* and *interdependence methods*.
  - **Dependence methods:** Multiple regression analysis, MANOVA, Canonical Correlation Analysis, Multiple Discriminant Analysis, etc.
  - **Interdependence methods:** Factor Analysis, Cluster analysis, Multidimensional Scaling, Correspondence analysis, etc.

Objectives of MVA

1. Predict relationships between variables. Also used for hypotheses testing
2. Investigation of dependence among variables: The nature of the relationships among variables is of interest. Are all the variables mutually independent or are one or more variables dependent on the others?
3. Grouping of variables
4. Patterns of data, to make clear comparisons, to discard unwanted information and to study multiple factors at once (Data reduction or structural simplification)

MVSA-Techniques

- **Analysis of variance (ANOVA) and Multivariate Analysis of variance (MANOVA):** In ANOVA, differences among various group means on a single-response variable are studied. In MANOVA, the number of response variables (DVs) is increased to two or more.
- **Regression Analysis:** For example, a model where dependent variable is Y, and independent variables are X1, X2, X3, etc. We have already discussed this technique in detail. We will revisit this topic through learning how to code some of these models in R. If you forgot the details of a regression analysis, please revisit our materials from Fall 2023.
- **Principal component analysis (PCA):** Suppose we have multiple variables representing similar things (i.e., prices or nature of sales from various channels). All the information we have might not be beneficial because of the correlation. This method allows us to interpret data (i.e., price data or sales data) in its simplest form by introducing new uncorrelated variables (i.e., one or two variables indicating price).
- **Factor analysis:** This method is similar to PCA (instead of principal component, we find factored score) used to handle big data (in terms of tons of variables telling same/similar things) into small, interpretable forms. It is a way to condense the data in many variables into just a few variables, also known as "dimension reduction" method. In practice, it groups variables based on high correlations. This approach includes PCA and common factor analysis.
- **Discriminant analysis:** An approach that allows us to classify various observations in two or more distinct set of categories. The idea here is to "discriminate" two or more groups based on certain attributes (variables).

MVSA-Techniques (green highlighted ones will be discussed in the class)

- **Cluster analysis:** This approach allows us to find similarities across observations, and then identify various groups. Note that there is no prior information about the group or cluster membership for any of the objects. We first partition the set of data into groups based on data similarity and then assign the labels to the groups. Unlike classification, is that it is adaptable to changes and helps single out useful features that distinguish different groups.
- **Conjoint Analysis: Hope you have learned in a previous class**. Known as trade-off analysis, is useful for identifying how people like or dislike different attributes of a product or service. You can use this analysis to find the ideal combination of attributes, such as features, benefits and colors.
- **Canonical Correlation Analysis:** An approach to investigate linear relations between two sets of variables. It is the multivariate extension of correlation analysis. It is mainly used for two purposes: data reduction and data interpretation. The motivation behind this approach is similar to PCA.
- **Multidimensional Scaling:** A technique that creates a map displaying the relative positions of several objects, given only a table of the distances between them. The map may consist of one, two, three, or even more dimensions. The program calculates either the metric or the non-metric solution. The table of distances is known as the proximity matrix. It arises either directly from experiments or indirectly as a correlation matrix.
- **Correspondence analysis:** A method for visualizing the rows and columns of a table of non-negative data as points in a map, with a specific spatial interpretation.

MVA can be used in parametric as well as non-parametric tests.

- **Parametric tests:** Makes certain assumptions regarding the distribution of data, i.e., within a fixed parameter. Based on interval/ratio scale, outliers are absent, data is uniform and have equal variance and have large sample size.
- **Non-parametric tests:** Do not make assumptions with respect to distributions. On the contrary, the distribution of data is assumed to be free of distribution. Based on ordinal/nominal scale, outliers might present, data is non-uniform and variance is unequal, and sample is small.

ANOVA (Analysis of Variance)

- **Situation 1:** Joy is the manager of an agriculture firm, where the company produces ‘avocados.’ The company acquired a new piece of land where they plan to produce avocados, but the company is not sure if the land is perfect/better for avocados. Joy was asked to find if the production of avocados in the newly acquired land is significantly better than their current cultivation ground (treatment 1), and their base cultivation ground (control). The avocados are picked up every 15 days and the observations last for 6 months. This means the company picks avocados from the new ground (treatment 2), current cultivation ground (treatment 1), and base ground (control) at the interval of 15 days. This also means that Joy has 12 datapoints from each piece of land to understand if the newly acquired land produces significantly higher number of avocados. Joy finds that the average number of avocados across three places are different. However, he does not know how to identify if the production in the new land is different statistically and that the difference is not due to anything random. Here, ANOVA can be helpful.
- **Situation 2:** Jenny finds that customers in her Home Warranty Business are churning. She asks a question, “Do our customers churn more in summer, fall, winter or spring? Here, the independent variable (IV) is “season”. In an ANOVA, our IVs are organized in categorical groups. For example, here we have 4 groups (seasons) for analysis.

What is ANOVA?

- Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures that is mainly utilized to analyze the existing differences among the values of means.
- **One way ANOVA** test is used to compare the variance in the group means within a sample considering only one independent variable or factor. It compares three or more than three categorical groups to establish whether there is a difference between them. Within each group, we should have three or more observations, and then the means are compared.
- **Assumptions:** DV is continuous, each sample is believed to be taken from a normally distributed population. Each sample is drawn independent of others. No heterogeneity in the variance (i.e., the variance among various groups is same).
- **<u>How to test hypothesis using One-way ANOVA?</u>**
  - H0: There is no difference between the groups and equality between means (churn is the same in different seasons).
  - H1: There is a difference between the means and groups (churn is different weights in different seasons).

Two-way ANOVA

- Examines the effect of two factors on a DV. For example, if Jenny is interested in finding if “churn is more in different seasons and if customer satisfaction (low/high) drives it”, then we will need two-way ANOVA as we have now ‘seasons’ and ‘satisfaction’. It will also be important to consider the levels of a variable (i.e., for satisfaction, we have two levels, but one can have three levels, four levels, etc.).
  - Assumptions
    - Churn (DV) should be continuous
    - IVs should be categorical (seasons, and satisfaction).
    - Each sample is believed to be taken from a normally distributed population. Each sample is drawn independent of others. No heterogeneity in the variance (i.e., the variance among various groups is same).
- **How to test hypothesis using Two-way ANOVA?**
- Group 1 hypotheses
  - H0: The means of all season groups are equal.
  - H1: The mean of at least one season group is different.
- Group 2 hypotheses
  - H0: The means of the satisfaction groups are equal.
  - H1: The means of the satisfaction groups are different.
- Group 3 hypotheses
  - H0: There is no interaction between the seasons and satisfaction.
  - H1: There is interaction between the seasons and satisfaction.

Key Difference between One-way and Two-way ANOVA

- One-way tests quality between three or more means. Two-way tests the interrelationship of two independent variables on a dependent variable.
- One-way involves one IV. Two-way has two IVs.
- The IV in one-way should have three or more categorical groups. Two-way compares multiple groups of two factors.

Regression vs ANOVA vs t-test

- Regression is used to make estimates or predictions for the DV with the help of single or multiple IVs. ANOVA is used to find a common mean between variables of different groups.
- In the case of regression, the error term is one, but in the case of ANOVA, the number of the error term is more than one.
- Regression is applied to independent variables or fixed variables. ANOVA is applied to variables which are random in nature.
- The t-test is the statistical test used to examine whether there is any difference between the means of two groups. The test assumption includes the normal distribution of the samples or groups.
  - Both compares means and have similar assumptions.
  - ANOVA can accommodate a huge population count. In t-test, sample population should be less than 30.
  - Test statistical value is F-test in ANOVA and t-test in t-test
  - In ANOVA, variance between the data set, group, or sample, along with the variance inside the data set, group, or sample, is calculated. In t-test, mean differences, the standard deviation, and the number of data values are calculated.

ANOVA IN PRACTICE

MANOVA

- A multivariate (MORE THAN ONE DV) version of the ANOVA model. It is a procedure for comparing the multivariate sample means, when we have two or more DVs.
- The one-way MANOVA is mainly utilized to determine whether there are any differences between the independent groups on more than one continuous dependent variable.
- Flybig inc. launched a new product in last December across four regions. The company observed the penetration of the new product in 4 regions, along with the (%) profits from the customers for the company. The company wants to understand if regions are associated with various levels of penetration of the new product and % profit. In this, we want to test if one region is significantly differently associated with penetration and % profit.
  - In the example, you will see differences will be based on profit and new product penetration (i.e., two DVs). MANOVA determines if the DVs get significantly affected by changes in the IVs. The MANOVA uses the covariance-variance between the existing variables to test for the difference between vectors of means.

MANOVA vs ANOVA

A. ANOVA mainly checks the differences between the means of two samples/populations while MANOVA checks for the differences between multiple sample/populations. It means that ANOVA analyses the difference between 2 or higher numbers of groups in their means based on the single DV, whereas MANOVA analyses the difference between multiple groups in the means based on multiple DVs.

B. MANOVA uses covariance-variance relationship of considering more than one DV. ANOVA allows only one DV, but the numbers of IVs may vary as per the data set.

C. ANOVA concerns about two variables, while MANOVA concerns the differences in multiple variables simultaneously.

D. ANOVA is parametric in nature, MANOVA is non-parametric in nature

Additional information on MANOVA

- For MANOVA, the data should be such that the observations per group (in our case, A, B,C,D) should be more than the number of DVs. We have two DVs and more than two observations per group.
- **Hypotheses:** Because MANOVA uses more than one dependent variable, the null and the alternative hypotheses are slightly changed:
  - H0: Group mean vectors are the same for all groups or they don’t differ significantly.
  - H1: At least one of the groups mean vectors is different from the rest.
- How is MANOVA model performance analyzed?
  - MANOVA in R uses Pillai’s Trace test for the calculations, which is then converted to an F-statistic when we want to check the significance of the group mean differences. Other tests, such as Wilk’s Lambda, Roy’s Largest Root, or Hotelling-Lawley’s test can also be used. However, Pillai’s Trace test is the most powerful one.
- Assumptions:
  a. *Multivariate normality* – Each combination of IV or DVs should have a multivariate normal distribution. Use Shapiro-Wilk’s test to verify.
  b. *Linearity* – DVs should have a linear relationship with each group of the independent variable.
  c. *No multicollinearity* – The DVs should have very high correlations.
  d. *No outliers* – Analysis will perform the best if no outliers in the dependent variables.

ANCOVA

- An approach to test the effect of categorical variables (and their interactions on a continuous dependent variable, after controlling for a continuous independent variable, which should covary with the dependent variable.
- For example, if a company wants to know the effect of market type (develop vs emerging) and product type (green product vs not green product) on sales, after controlling for advertising spend (which in practice should covary with the sales), ANCOVA is the right approach.
- **Dataset:** Session 2-SAS-ANCOVA-DRUG TEST. Informs us about the three drugs (A,B, and C), which could influence recovery score of patients. Such recovery score covaries with vitality score. Our aim is to identify if treatment (A,B,C) influence recovery, and if yes, which one is more influential.

MANOVA IN PRACTICE

tusind tak
ngiyabonga
dziękuję
谢谢
dakujem vám
ありがとう
merci
suksema
danke
thank
you
baie dankie
धन्यवाद
molte grazie
gracias
takk
obrigada
dank u
teşekkür edirem
mahalo
obrigado
gràcies
teşekkür ederim
شكرا
tänan
tack så mycket