MATH F432: APPLIED STATISTICAL METHODS

Project 3

Cirrhosis Analysis

Akhil Agnihotri (2016A4PS0353H)
Deeksha Kartik (2016A7PS0809H)
G. Jaya Aadityaa (2016A4PS0213H)
Skanda Vaidyanath (2016A7PS0236H)

Overview

- Dataset contained 11 continuous variables and 11 categorical variables (3 string and 8 numeric).
- This demanded a separate approach for analyzing both the types.
- A Multivariate ANOVA was run on the 11 continuous variables by status and gender to test the homogeneity of these variables for both the genders for a given status.
    - Model = status + gender + status * gender + intercept
- Kruskal-Wallis test was conducted for numeric categorical variables and it was concluded only the ‘edema’ gives a statistically significant difference in histolic stage.
- A chi-square goodness of fit test was conducted to show that females have a higher tendency of liver cirrhosis than males.
- Discriminant Analysis to find the status variable from the other columns.

Two way MANOVA

- The two-way multivariate analysis of variance (two-way MANOVA) is often considered as an extension of the two way ANOVA for situations where there is two or more dependent variables.
- The primary purpose of the two-way MANOVA is to understand if there is an interaction between the two independent variables (gender and status in our case) on two or more dependent variables. It generally has one primary aim: to understand whether the effect of one independent variable on the dependent variables (collectively) is dependent on the value of the other independent variable. This is called an "interaction effect".

Assumptions

1. The **two or more dependent variables** should be measured at the **interval** or **ratio level** (i.e., they are **continuous**).
2. The **two independent variables** should consist of **two or more categorical**, **independent groups**.
3. You should have **independence of observations**, which means that there is no relationship between the observations in each group or between the groups themselves.
4. You should have an **adequate sample size**.
5. There are **no univariate or multivariate outliers**.
6. There is **multivariate normality**.
7. There is a **linear relationship between each pair of dependent variables for all combinations of groups of your two independent variables**.
8. There is **homogeneity of variance-covariance matrices**.
9. There is **no multicollinearity**.

Wilks’ Lambda

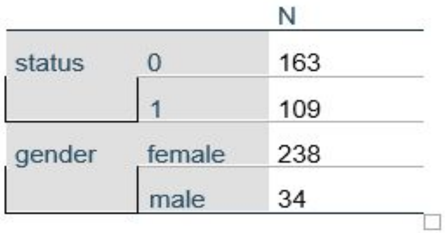
This is a test statistic used in MANOVA to test whether there are differences between the means of identified groups on a combination of dependent variables.

It has the same role as F-test in one-way analysis of variance.

It is a direct measure of the proportion of variance in the combination of dependent variables that is unaccounted for by the independent variable. If a large proportion of the variance is accounted for by the independent variable then it suggests that there is an effect from the grouping variable and that the groups have different mean values.

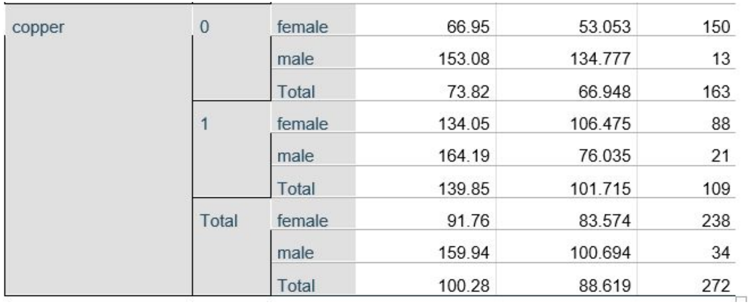
Between Subject Factors - Age and Gender

- For each independent variable, descriptive statistics are computed for N = 272 data points which had a mean age of 50.
- Effectively, a 2x2 matrix for each variable was constructed and mean and standard deviation were computed as in *figure 2.*



		N
status	0	163
	1	109
gender	female	238
	male	34

figure 1. Status and gender statistics



copper	0	female	66.95	53.053	150
		male	153.08	134.777	13
		Total	73.82	66.948	163
	1	female	134.05	106.475	88
		male	164.19	76.035	21
		Total	139.85	101.715	109
	Total	female	91.76	83.574	238
		male	159.94	100.694	34
		Total	100.28	88.619	272

figure 2. Variation of copper across status and gender

Between Subject Factors - Inferences

- A **healthy** human has a **low** level of **Bilirubin** and its usually slightly **higher** in men. However, we see that **women** with **Cirrhosis** have abnormally **high** levels of it, which is not so in **men**. Overall, women have a **higher range** of Bilirubin distribution as compared to men and the Bilirubin levels for men are **relatively uniform** across their **status**.
- **Clotting time** is slightly **higher** in the case of **diseased** humans, and this can be attributed to their **lower Platelet count.** Overall, males have a slightly **higher** clotting time than females.
- There is a significant **increase** in levels of **Alkaline Phosphate** and **Glutamic Oxaloacetic Transaminase** for those **afflicted** by the disease due to the **presence** of a **damaged liver.**

Kruskal-Wallis H test

The test is a non-parametric version of ANOVA and is used for test variables that are at least ordinal in nature. The ranks of the test variable are used for the test rather than the actual values of the test variable. We group the data by a categorical variable to get our blocks and then conduct the test on the separate blocks to check if the values of ranks are similar for each block.

Kruskal-Wallis H Test

The test variable was the histolic stage and the purpose of the test was to determine whether the different categories of gender, edema and drug had an effect on the histolic stage of the patient. However, we found that the test was statistically significant only for grouping by edema and found that ‘edema despite diuretic therapy’ has patients with a higher histolic stage than the other two categories of edema and as expected ‘no edema’ had patients with the least stage.

Chi-square test for goodness of fit

The test was conducted on a 2*2 contingency table to show that there is a difference in the proportion of afflicted (status column) in males and females. The test was conducted at 5% los and the Pearson Chi-square value was significant and showed that more females have liver cirrhosis than males.

Discriminant Analysis

- Log determinant values are fairly similar. Null hypothesis of equal population variances is not rejected as p value > alpha
- High eigenvalue of 0.68 signifies high variance. Wilk’s lambda is statistically significant.