Graphical user interface, text

Description automatically generated with medium confidence

Logo

Description automatically generated

**Abhilash Dikshit**

Course: ALY 6010

Instructor: Mohsen Soltanifar

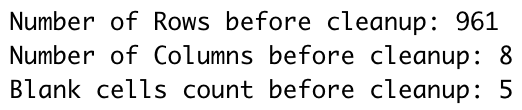
Week 5: R Practice

Date: 2022/12/11

**Introduction**

Predictive modelling techniques such as regression analysis is used to determine the relationship between a dataset’s dependent (goal) and independent variables. It is widely used when the dependent and independent variables are linked in a linear or non-linear fashion, and the target variable has a set of continuous values.

Initially, the data shared by the professor is imported to R studio and data cleanup was performed.

 Text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Part 1: Produce and export at least one correlation table or correlation chart. A correlation chart is diagnostic and should not be larger than 5 variables for reporting purposes. Why is this? Provide several sentences describing the key analytical findings.

#### **Shapiro-Wilk normality test**

|  |  |
| --- | --- |
|  | Shapiro Test on Hormone Therapy Data to check whether the considered data is normally distributed data or not.  Shapiro Test on White Blood cells' Data to check whether the considered data is normally distributed data or not. |

|  |  |
| --- | --- |
| Analysis using Quantile-Quantile Plots for Hormone Therapy and White Blood Cells | |
|  |  |
|  |  |

Both Hormone Therapy and White Blood Cells are not normally distributed.

Next, we created a matrix for Hormone > 50 and Hormone < 50 as shown below.

|  |  |  |
| --- | --- | --- |
| Text  Description automatically generated  Correlation matrix for Hormone\_more | Text  Description automatically generated  Correlation matrix for Hormone\_less | |
| Visualizing the correlation matrix for Hormone > 50 and Hormone < 50 | | | |
|  | |  | |
|  | | Checking which variables are normally distributed and which are not using Quantile-Quantile Plots for Hormone >50 and Hormone<50. | |

Part 2: Produce and export at least one regression table. You may pick your own outcome and predictor variables. How does regression analysis differ from correlation analysis? Provide several sentences discussing the key results.

**Simple Linear Regression:**

Two subsets were created using 3 variables Hormone\_more w.r.t Age and Smoking habit and Hormone\_less w.r.t Age and Smoking habit and stargazer library is installed for further representation.

Text

Description automatically generated with medium confidence

|  |  |
| --- | --- |
| Boxplot for Univariate Regression table | |
|  |  |

|  |  |
| --- | --- |
| Simple Regression Table for Hormone >50 w.r.t Age and smoking habit | Simple Regression Table for Hormone <50 w.r.t Age and smoking habit |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
| Text  Description automatically generated |  |

**Multiple Line Regression:**

Hormone therapy and White blood cells was considered along with Age for Multiple Regression.

**Text, table

Description automatically generated**

**Text

Description automatically generated**

We can determine the distribution and correlation of the variables we'll be utilising for Multiple Regression using the plot() method. The figure below demonstrates this as follows:

|  |  |
| --- | --- |
|  |  |
|  |  |

**Reference**:

1. Bevans ([2022, November 11](http://127.0.0.1:55283/rmd_output/4/#ref-R-Career));Datanovia ([2019, December 26](http://127.0.0.1:55283/rmd_output/4/#ref-R-Action));Linear Regression Example in r Using Lm() Function ([n.d.](http://127.0.0.1:55283/rmd_output/4/#ref-R-Cran));RPubs - How Do i Get p-Values and Critical Values from r? ([2017, March 1](http://127.0.0.1:55283/rmd_output/4/#ref-R-Material1));Investopedia ([2022, August 31](http://127.0.0.1:55283/rmd_output/4/#ref-R-Material2))
2. Bevans, R. 2022, November 11. Hypothesis Testing | a Step-by-Step Guide with Easy Examples. <https://www.scribbr.com/statistics/hypothesis-testing/>.
3. Datanovia. 2019, December 26. How to Do a t-Test in r: Calculation and Reporting. <https://www.datanovia.com/en/lessons/how-to-do-a-t-test-in-r-calculation-and-reporting/>.
4. Investopedia. 2022, August 31. What Is a Confidence Interval and How Do You Calculate It?<https://www.investopedia.com/terms/c/confidenceinterval.asp>.
5. Linear Regression Example in r Using Lm() Function. n.d. <https://www.learnbymarketing.com/tutorials/linear-regression-in-r/>.
6. RPubs - How Do i Get p-Values and Critical Values from r? 2017, March 1. <https://rpubs.com/mdlama/spring2017-lab6supp1>.