Graphical user interface, text

Description automatically generated with medium confidence

Logo

Description automatically generated

**Abhilash Dikshit**

Course: ALY 6010

Instructor: Mohsen Soltanifar

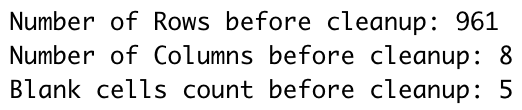
Week 6: R Practice

Date: 2022/12/19

**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: Using an appropriate variable, create dummy variables to subset your dataset. Then rerun your regression line for your dependent variable. How many subsets did you create? How many lines are there? Create a scatterplot with multiple regression lines. How does this impact your understanding of the impact of the categorical variable on the regression?

Visualized the data for Treatments using scatter plot:

Chart

Description automatically generated

#### Installed the package “fastDummies” for creating dummy variable for Treatment.

Table

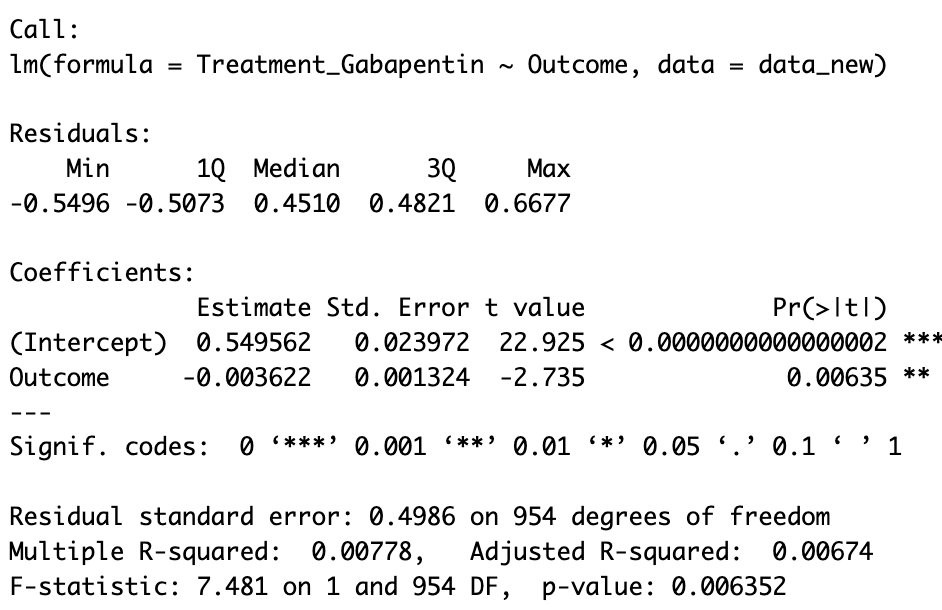
Description automatically generated

One category variable, called treatment, appears in our dataset. Therefore, I made a dummy variable for the treatment variable. After making the dummy variable, we now have two more variables: treatment Gabapentin and treatment Placebo.

Currently, two subsets have been formed after this dummy variable.

The single line regression for the dependent variable must now be run.

We first perform regression for the treatment (gabapentin) and the result.



We can see that their residuals' minimum and maximum values are probably close to the fitted line, and the same is true for 1Q and 3Q. So, we can conclude that this is distributed symmetrically along the line. We have intercept value is 0.55 and slope is -0.003.

Equation:

*Treatment\_Gabapentin = 0.55 – 0.003 \* outcome*

Here treatment\_Gabapentin is dependent variable.

Chart, line chart

Description automatically generated

Dummy variable: Treatment\_Placebo

Text

Description automatically generated

For this is also symmetrical distributed around fitted line because residuals are at similar distance from fitted line. Here we have intercept 0.45 and slope is 0.003, so this has positive regression.

Equation:

*Treatment\_Placebo = 0.45 + 0.003 \* outcome*

Treatment\_Placebo is Dependent Variable.

Chart, line chart

Description automatically generated

Single regression line as certain results between only one X and one Y variable.

Above two regression line graphs demonstrate the trend of a dummy variable with outcome. In the first Graph, dummy variable trend is flat with a decline as outcome value increase. In the second Graph, for second dummy variable trend is flat with a substantial tilt as outcome increase.

**Multiple Linear Regression:**

For Multiple Regression, Let’s see if there is any linear relationship between Outcome, white Blood Cells, Treatment\_Placebo and Treatment\_Gabapentin.

Text

Description automatically generated

From the residuals, we can see that there is large difference between distance from min and max value from fitted line so we cannot say that it is significantly distributed around the line.

The standard errors for these regression coefficients are very small, and the t-statistics are large. The p-values reflect these small errors and t-statistics. Here we can see that NA values for Treatment\_Gabapentin in our summary so from this we can say that Treatment\_Gabapentin is linearly related to the other variables.

Equation:

*Outcome = 13.80 – 0.22 \* WhiteBloodCells + 2.09 \* Treatment\_Placebo + Treatment\_Gabapentin*

Chart, scatter chart

Description automatically generated

In this Scatter Plot, we can see that there are two regression line where red line indicates for Gabapentin and blue is for Placebo for outcome vs White Blood Cells.

This above graph explains the slightly negative trend, as the regression line is slightly tiled in the downward direction.

Which helps us to understand that the White Blood Cells has some role in outcome with Treatment\_Gabapantin and Treatment\_Placebo.

Part 2: Using the appropriate subset data from step 1, create separate regression lines for each subset. How do these regression lines differ from the regression lines in step 1? How does this method of looking at the data impact your understanding of the data?

We will take previous Multiple regression for this question

**Text

Description automatically generated with medium confidence**

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

Here we can see through residuals that min and max have not equal distance from fitted line so we can say that it is not symmetrically Distributed with line.

Equation:

*Outcome = 16.037 - 0.2469 \* White Blood Cells + Treatment\_Placebo*

*Chart, scatter chart

Description automatically generated*

This is scatter plot with regression line with slightly flat with tiny decline as wbc increase. With this we can say that wbc has some relationship with outcome with Treatment\_Placebo

Now we can take another regression for different treatment

We will check regression between outcome, White Blood Cells, and Treatment\_Gabapentin.

**Text, letter

Description automatically generated**

From the residuals, we can see that there is less difference between distance from min and max value but not similar to each one of them from fitted line so we can say that it is less significantly distributed around the line.

Equation:

*Outcome = 13.6764 – 0.2054 \* WhiteBloodCells + Treatment\_Gabapentin*

**Chart, scatter chart

Description automatically generated**

This is scatter plot with regression line for Treatment\_Gabapentin with slightly flat with tiny decline as White Blood Cells increase.

This identifies as White Blood Cells has relationship with outcome variable with Gabapentin treatment.

There is negligible difference between multiple linear regression which we mentioned earlier in Part 1 with These two regressions which we discussed in Part 2.

**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>.