BE303 - Biostatistics

ASSIGNMENT REPORT

submitted by

Aditya Sarkar
IIT Mandi
B19003

under the supervision of

Dr. Erwin Fuhrer

INDIAN INSTITUTE OF TECHNOLOGY, MANDI

May 2022

Abstract

In this programming assignment, we have improved our understanding of the statistical tests and their applications on real datasets. We have applied both parametric and non-parametric tests on the datasets.
This report is made in LaTeX

Contents

Chapter 1

Simulations

1.1 Brief description of the dataset

We have generated a population with 1,000,000 subjects using uniform distribution with the probability density function: $100 \leq x \leq 200 = 1$.

1.2 Conducting T-test

In this section, we will be discussing the T-test using the above data.

1.2.1 Experiment

We took a sample of size 10,000 and a reference value of 150. We will be testing the following Null and Alternate hypotheses is -

$$H_0 : \mu_s = \mu_p$$

$$H_1 : \mu_s \neq \mu_p$$

where μ_s is sample mean and μ_p is the population mean. Since we know that the population distribution is a uniform distribution with mean of 100+200/2 = 150, we took the reference mean as 150, which in turn makes our null hypothesis true.
We took an alpha threshold of 0.05.

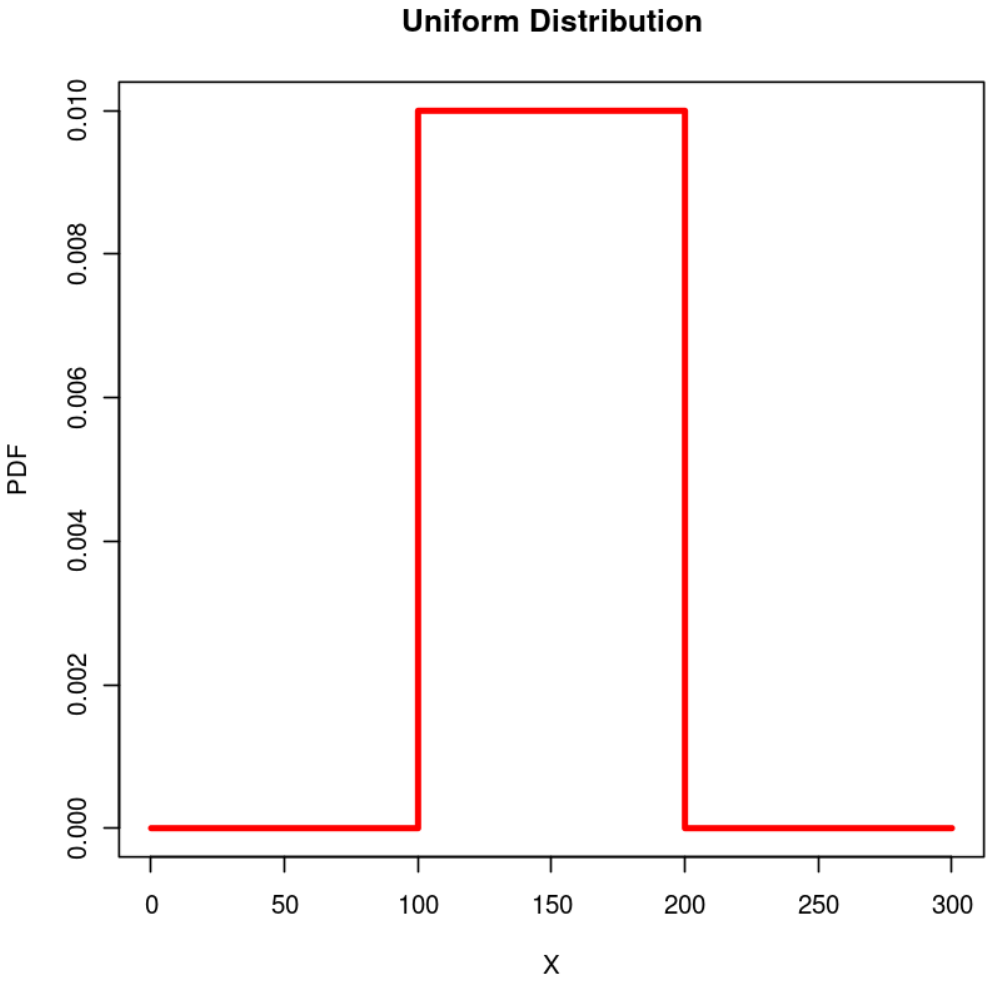

Fig. 1.1: Uniform Distribution

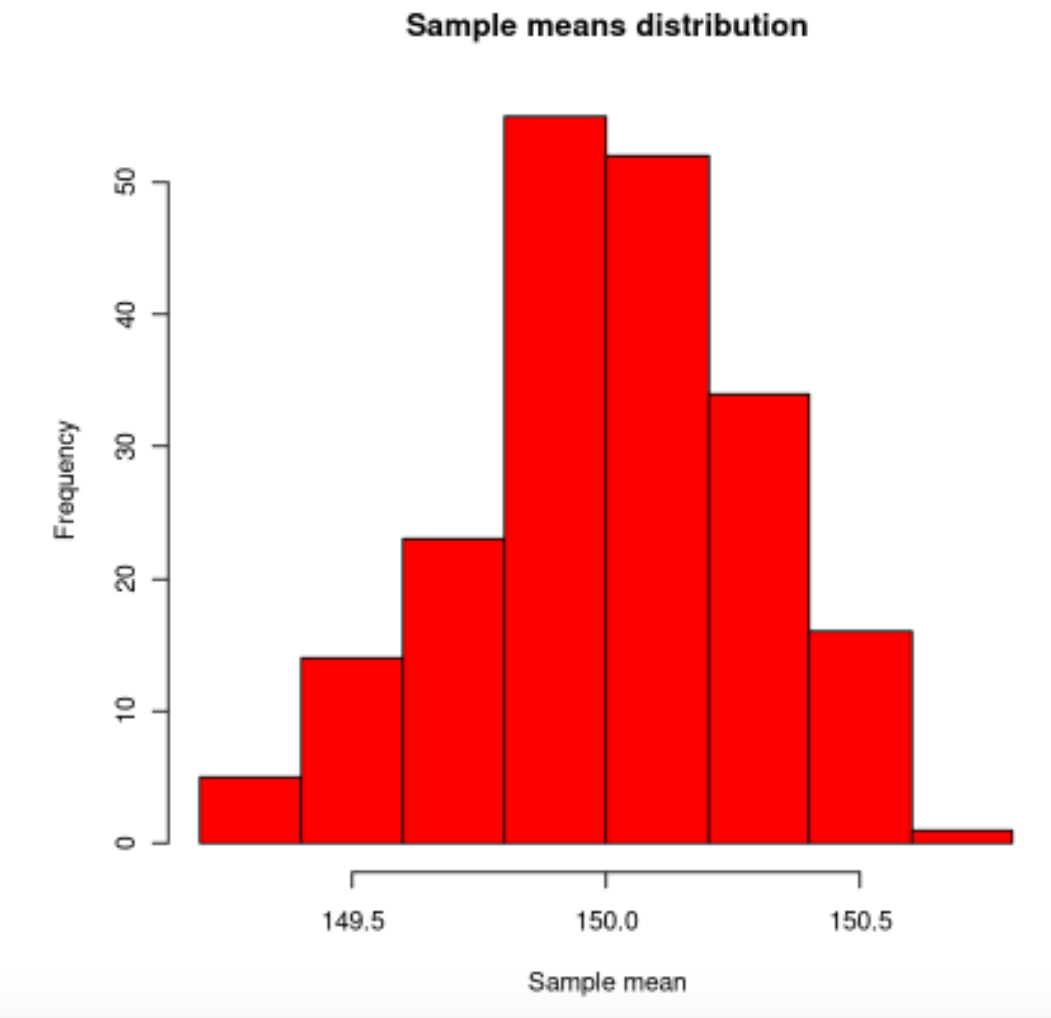

Fig. 1.2: Sample Distribution

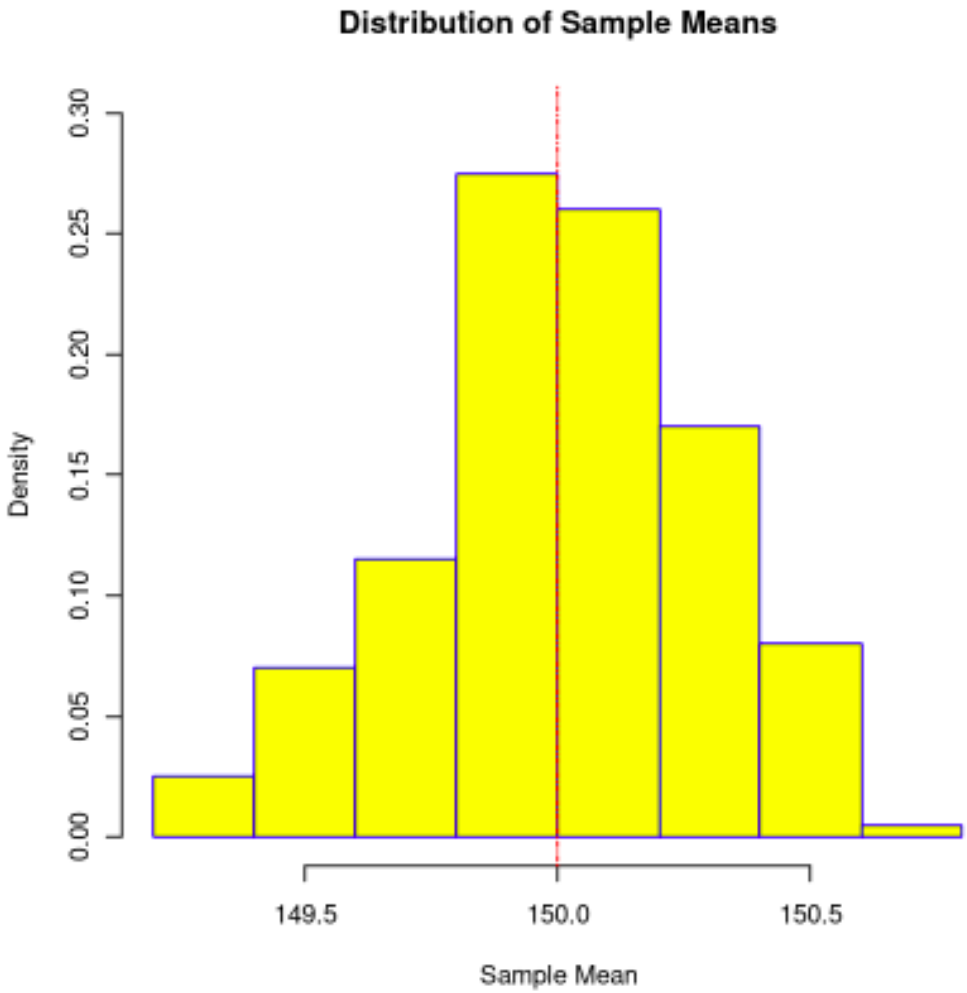

Fig. 1.3: Sample Distribution (with density)

1.2.2 Small Sample size

For small sample sizes, the type 1 error of T-test (0.055) is higher than that of large sample sizes (0.05). This can seen from the sample distributions as when the sample size is small, the distribution is skewed and not a Gaussian Distribution. This is proposed in the Central Limit theorem. Only when the sample size is large, it is Gaussian. Assumptions of T-test that are being violated are -

1. Distribution Assumption - It Requires normally distributed samples. It assumes the sample means have a t-distribution, which is not always the case, especially when the sample size is low.

1.2.3 Non-parametric test

Given the small sample size, the above assumptions are violated for T-test, as a result of which non-parametric tests are preferred. Parametric always assume a distribution for the samples. In the assignment, we implemented Rank Sum test, a non-parametric equivalent of T-test. We got the type-1 error for Rank Sum test as 0.035 while for T-test, it is 0.055. This was for sample size of 10. So it is high for T-test, which clearly shows that Rank Sum is better than T-test.

Chapter 2

Data Analysis - I

2.1 Brief description of the dataset

I have imported the population data from the excel sheet ”India Health Database”. To obtain population data for the years 2012 - 2017, I extrapolated the total population for each state and each year, with the data from 2011 and 2018 assuming a linear increase (or decrease).

2.2 Malaria Dataset

I took the dataset of Malaria and compute the mortality for corresponding years and state. The average mortality of all states was found out to be 1.60 x 10^{-6}, and the confidence interval was from 8.62 x 10^{-7} to 2.34 x 10^{-6}. Confidence intervals and mean mortality across all the states are present in the code outputs. We can observe that populations in almost all states increased except for Arunachal Pradesh, Manipur, Meghalaya, Mizoram, and Tamil Nadu, where the population decreased. A linear interpolation of the population between 2011 and 2018 shows the linearity in population growth or decline for each state. For getting the standard deviation, across the states, we considered the point where 68.25% of data lies. That is actually 1 standard deviation from the mean.

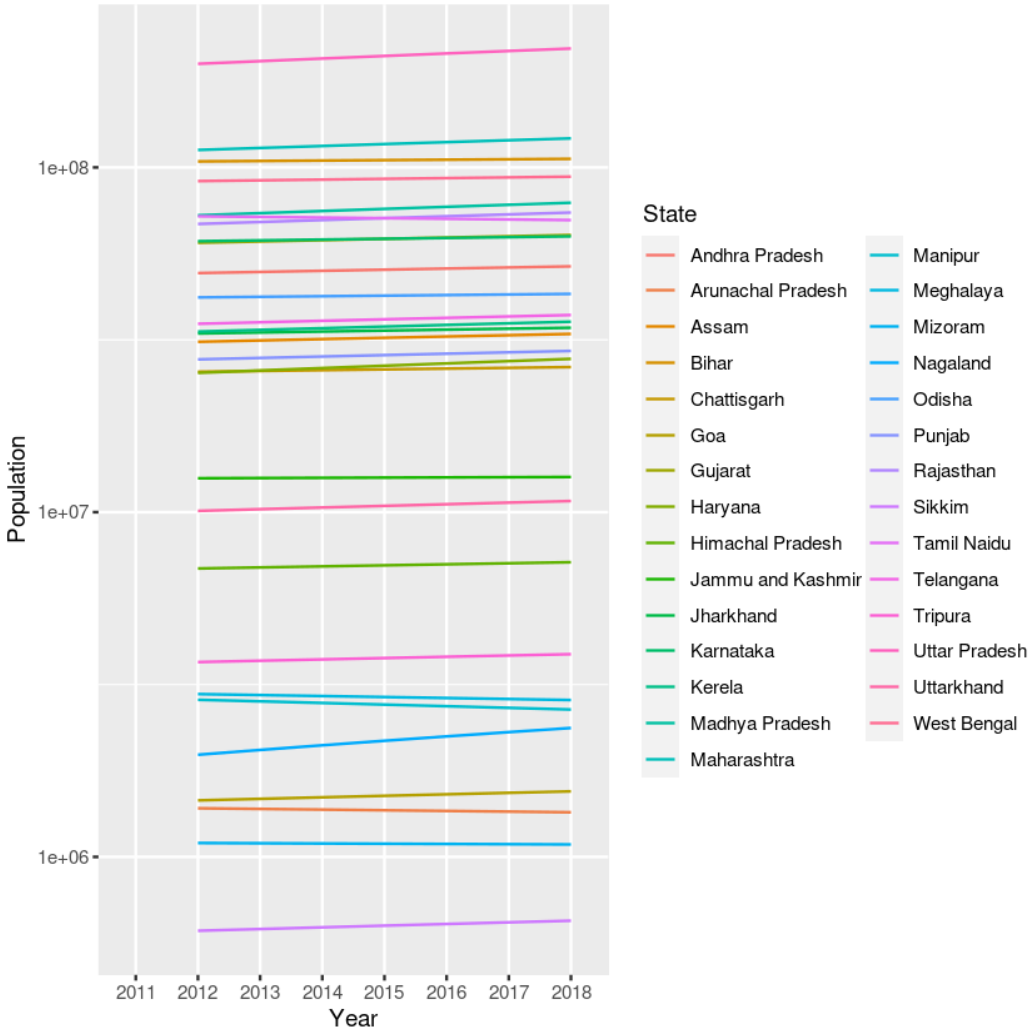

Fig. 2.1: Quick plot for Population

2.3 Difference in Mortality

We have less number of samples in cases and deaths. So parametric tests will not work here. The years can be considered as 1 factor and the groups will be 2013, 2014, 2015, 2016, 2017 and 2018. So the test will be non-parametric equivalent of One way Anova. Hence we applied a Kruskal Wallis test, a non-parametric test. Non-parametric is preferred here because it does not assume any distribution of the mean of the samples. Null Hypothesis will be that the mean ranks of the groups (all the years in this case) are the same. Alternative will be that they are not same. We found the p-value as 0.169, which shows that the null hypothesis is accepted over the alternative hypothesis. That is, there is no significant difference in the

mortality of the different years.

2.4 Correlation with urbanisation level

We calculated the correlation between the two columns using Pearson correlation coefficient for all the years that are from 2013 to 2018. The table is found in the code. Here the positive correlation means the variables are directly proportional to each other or if one increases (/decreases), the other one also increases (/decreases). Opposite of this is negative correlation. Magnitude of correlation shows the strength of the correlation or relationship between two variables. For the correlations, we can see that the mortality is weakly correlated with urbanisation level. Hence urbanisation does not bring any mortality. However, most of them are positive, which is concerning as increase in urbanisation is having an impact on mortality, though with a weak relation.

```
Correlation b/w urbanisation vs mortality for year  2013 : 0.0126469612185968
Correlation b/w urbanisation vs mortality for year  2014 : 0.0441934249494912
Correlation b/w urbanisation vs mortality for year  2015 : 0.020293975457508
Correlation b/w urbanisation vs mortality for year  2016 : -0.0568859890028357
Correlation b/w urbanisation vs mortality for year  2017 : 0.0466640845307434
Correlation b/w urbanisation vs mortality for year  2018 : -0.186801100889142
```

Fig. 2.2: Correlations

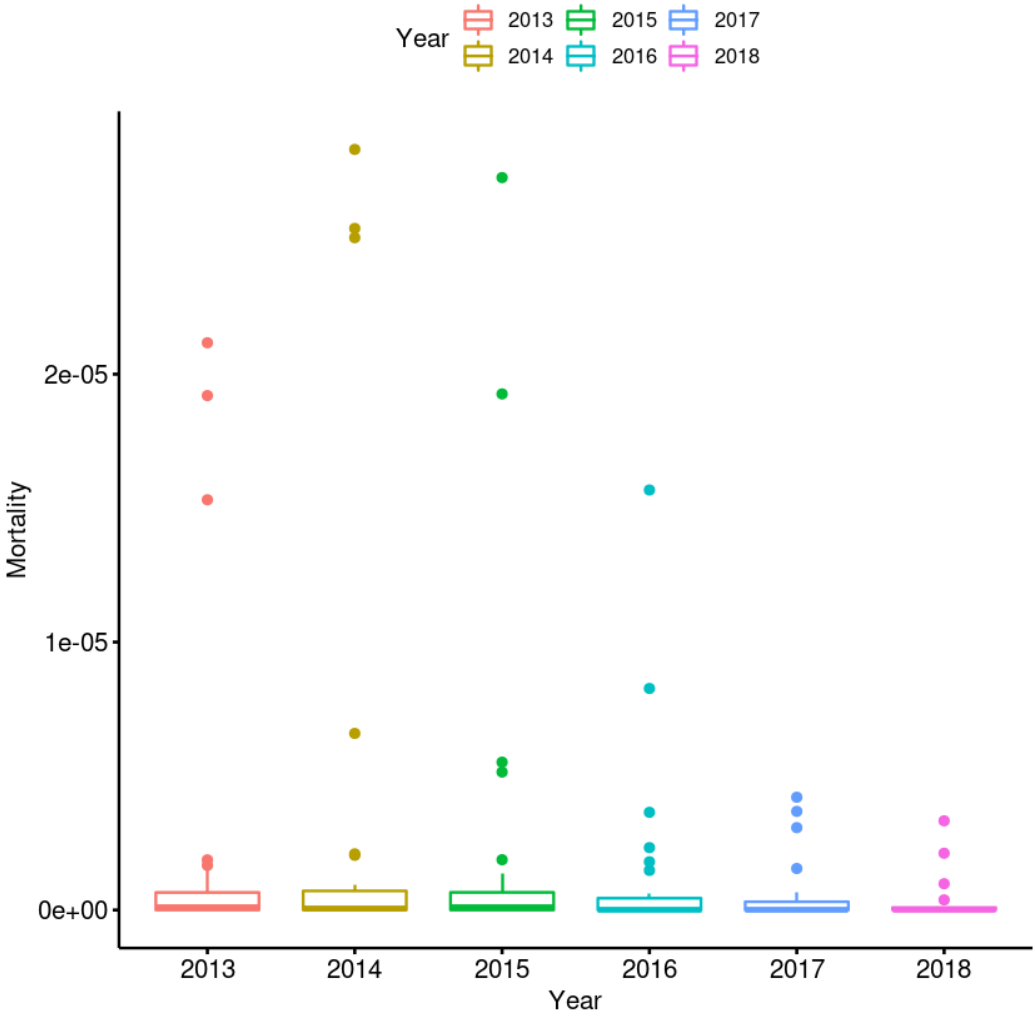

Fig. 2.3: Boxplots for different years

Chapter 3

Data Analysis - II

3.1 Brief description of the dataset

In the attached excel sheet "Glucose BP levels" is a dataset of glucose levels and the systolic blood pressure. The research question that we will be investigating here is if there is a significant relation between glucose levels and high blood pressure.

3.2 Relation between glucose levels and high blood pressure

We will be using linear regression to analyse the linear relation between the two variables. We have this equation of the line y = $b_0 + b_1 x_1$. So we can consider y as Glucode level and x as blood pressure. Using R, we fit this line on the data and we got b_0 = 1.915 and b_1 = -898.94. The plot showing the same is given below.

We can also predict the further values of x considering this linear relation exists. So the regression equation is y = mx+b. From the given data, we have predicted the value of m and the value of b, so on giving the value of x, we can get the value of y. It may not be the exact value and hence there is a slight error between the predicted value and the actual value, which one can capture using mean squared error. However, values predicted beyond

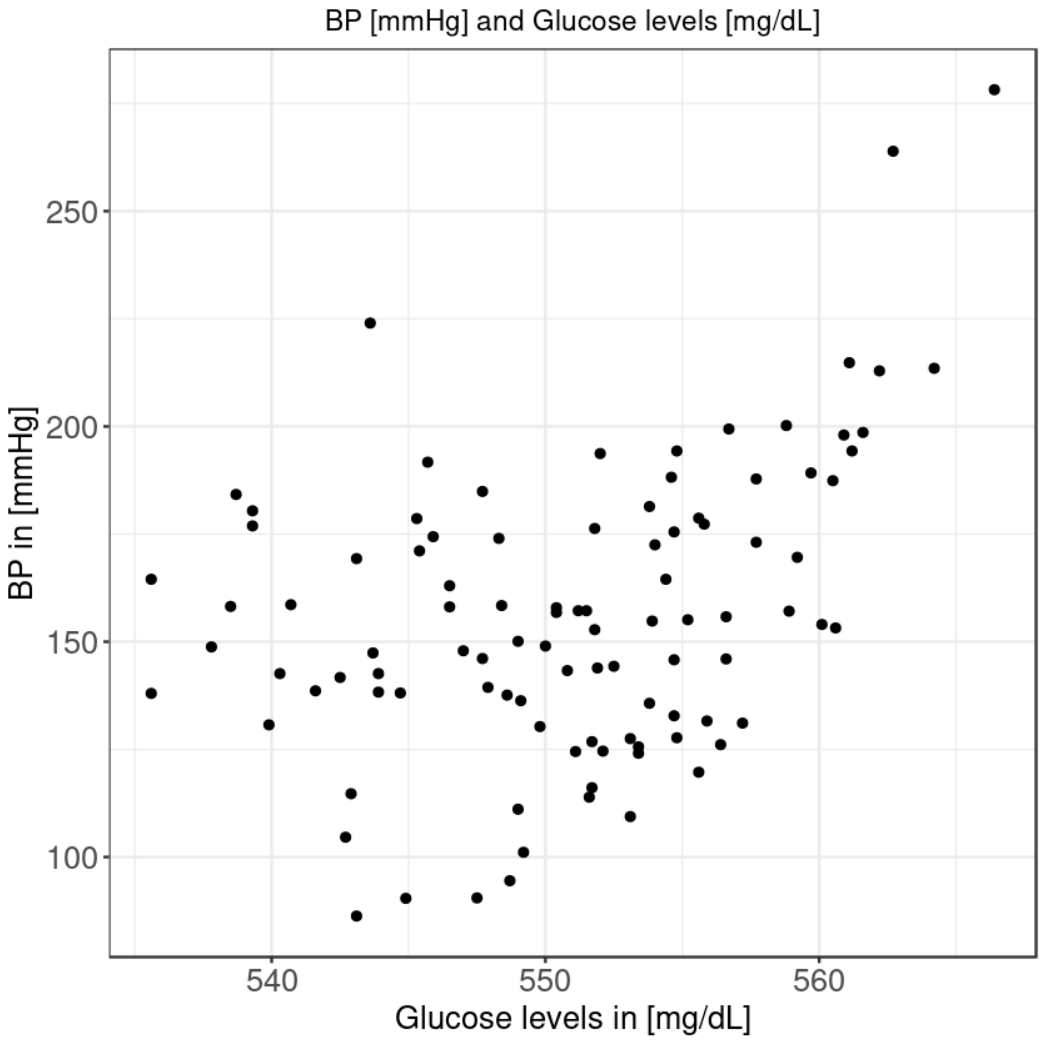

Fig. 3.1: Scatter Plot

a certain threshold make no sense, as the error will be too large.

3.3 Confidence intervals

To compute the the confidence interval for the regressors, we need to compute the standard errors (SE) first. Then we computed the standard errors for both regressors - that are the slope and y-intercept. For the slope, it was around 0.4598 and for the intercept, it was around 253.36. We then computed the confidence intervals. The graph is as shown below.

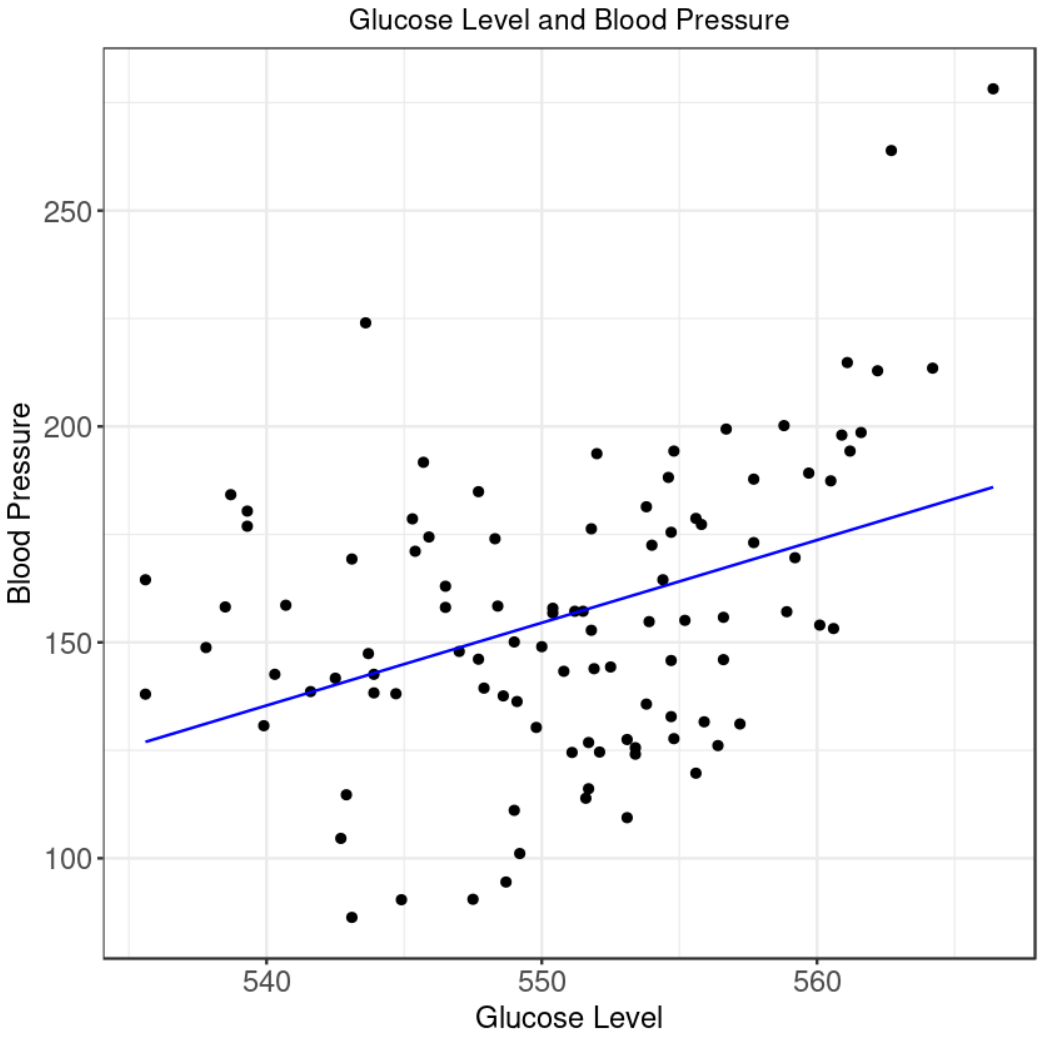

Fig. 3.2: Linear fit

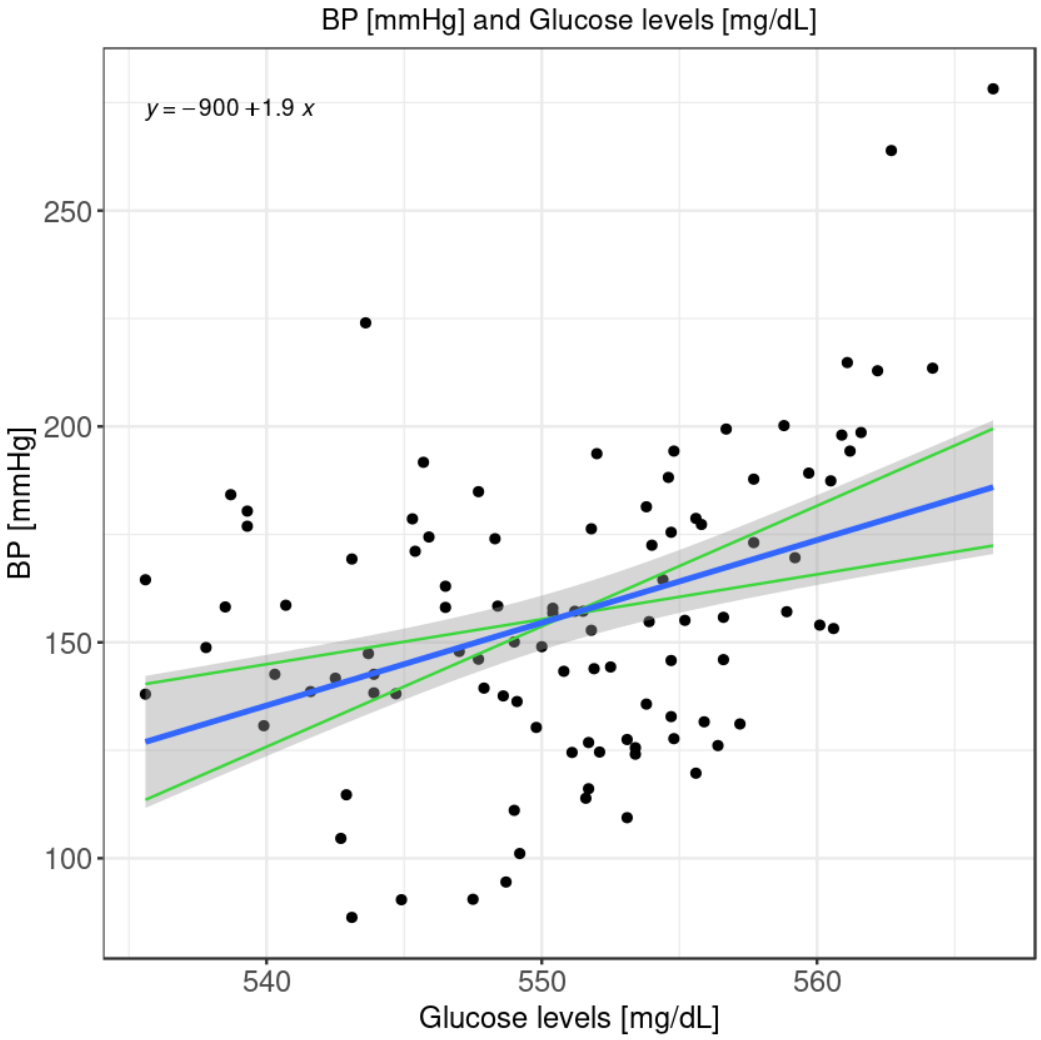

Fig. 3.3: Confidence Intervals

3.4 Residuals of model

The plot for the residuals is given below. We can clearly see that it is a biased and heteroscedastic residual. It is biased because average value of any thin vertical strip is not zero. It is heteroscedastic because the spread of the residuals is not equal in any thin strip. This is also evident from the plot below.

The fallacy is that the variance depends on explanatory variable. Thus the variance is not equal. It is a big problem for statistical tools that assume same variance such as Regression analysis and ANOVA. Hence for this kind of situation, we need very advanced statistical tools.

The model is a good fit wrt the R2 values, but as we have seen that many target values lie outside the 99% confidence interval. This shows that the model does not predict with much of a accuracy. We can also observe this in the above plots.

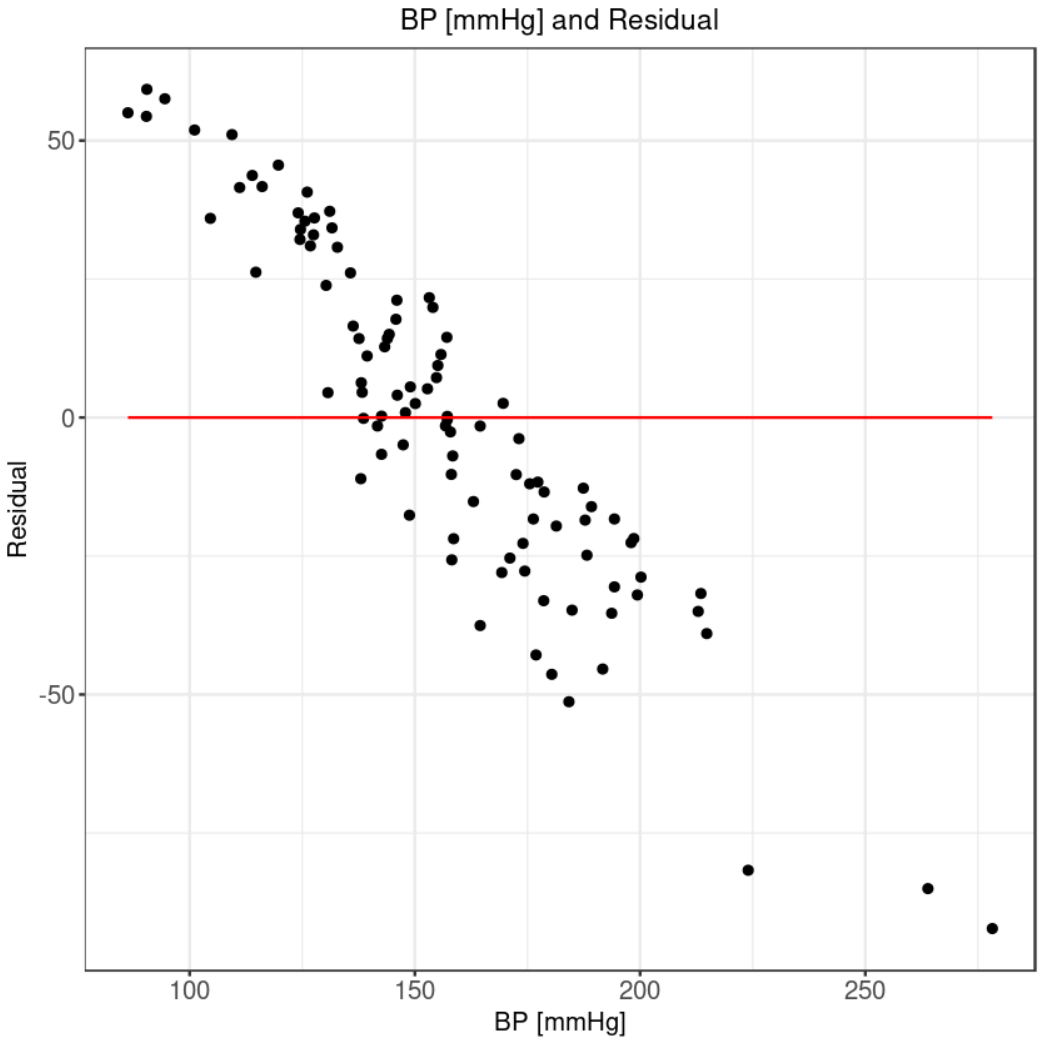

Fig. 3.4: Residuals plot

References