MDSC 206-Software lab in R

Mini Project- Heart Attack Prediction

Report.

reg.no - 20237

The report provides analysis of how Heart attack dataset is useful in prediction of heart attack possibility. The dataset has 14 variables which includes age, sex, chestpain levels,cholestrol ,resting ecg values etc. Using these variables we will try to build a model that predicts heart attack possibility.The variables in dataset are...

1) age

2) sex

3) chest pain type (4 values)

4) resting blood pressure

5) serum cholestoral in mg/dl

6)fasting blood sugar > 120 mg/dl

7) resting electrocardiographic results (values 0,1,2)

8) maximum heart rate achieved

9) exercise induced angina

10) oldpeak = ST depression induced by exercise relative to rest

11)the slope of the peak exercise ST segment

12) number of major vessels (0-3) colored by flourosopy

13) thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

14) target: 0= less chance of heart attack 1= more chance of heart attack

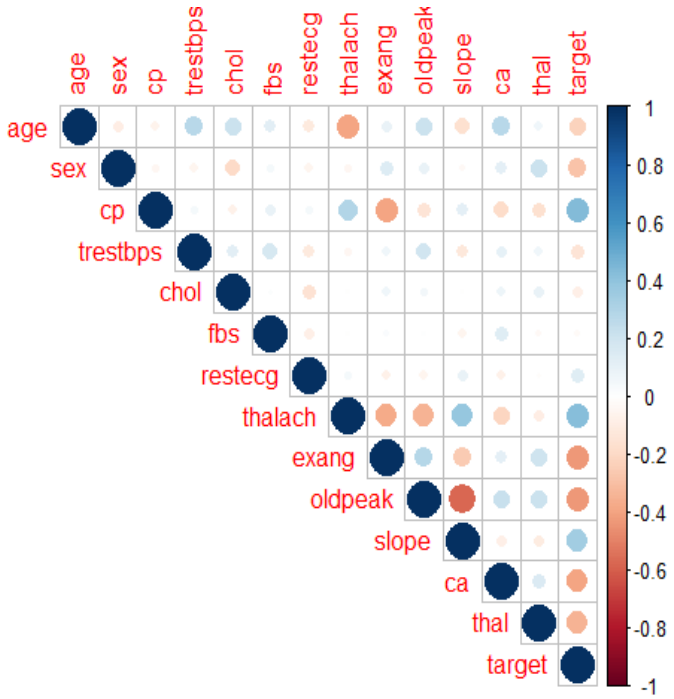
Basic Statistics.

First we found out some basic informations about the dataset. Like the dimension of the dataset which says there are 303 entries that is 303 observations for 14 different categories of information. Then we took out 6 point summary of each varialble which contains the mean,median, the min and max value and 1st and 3rd quartile values of each variable.

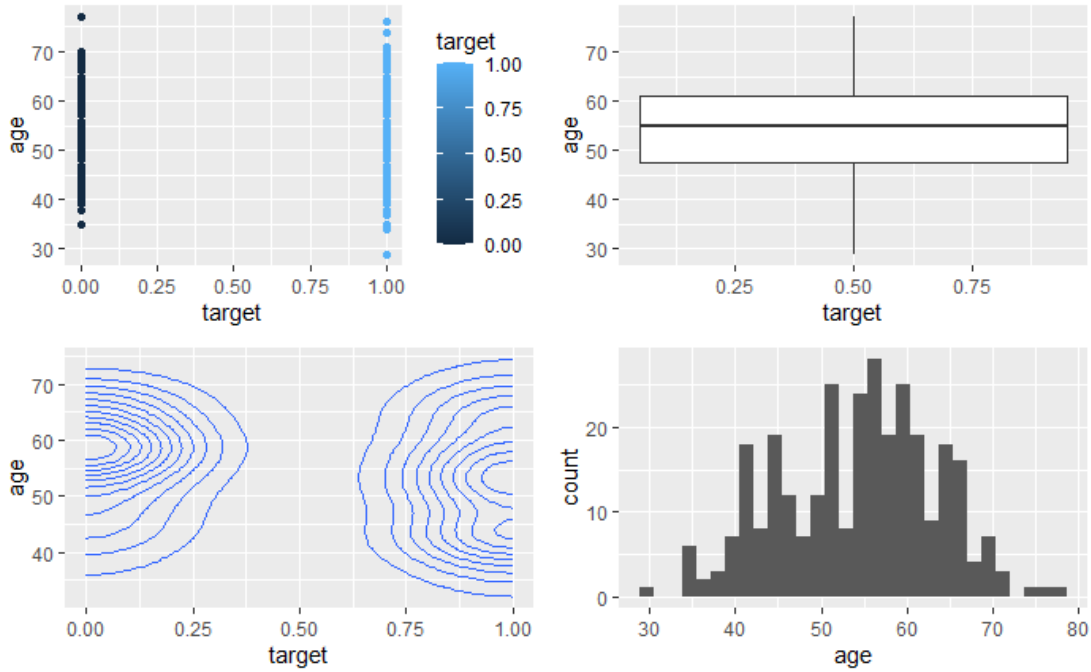
Exploratory Data Analysis

At first we will see the distributions of each variables.

Then we try to find correlation between variables by looking at correlation plot.



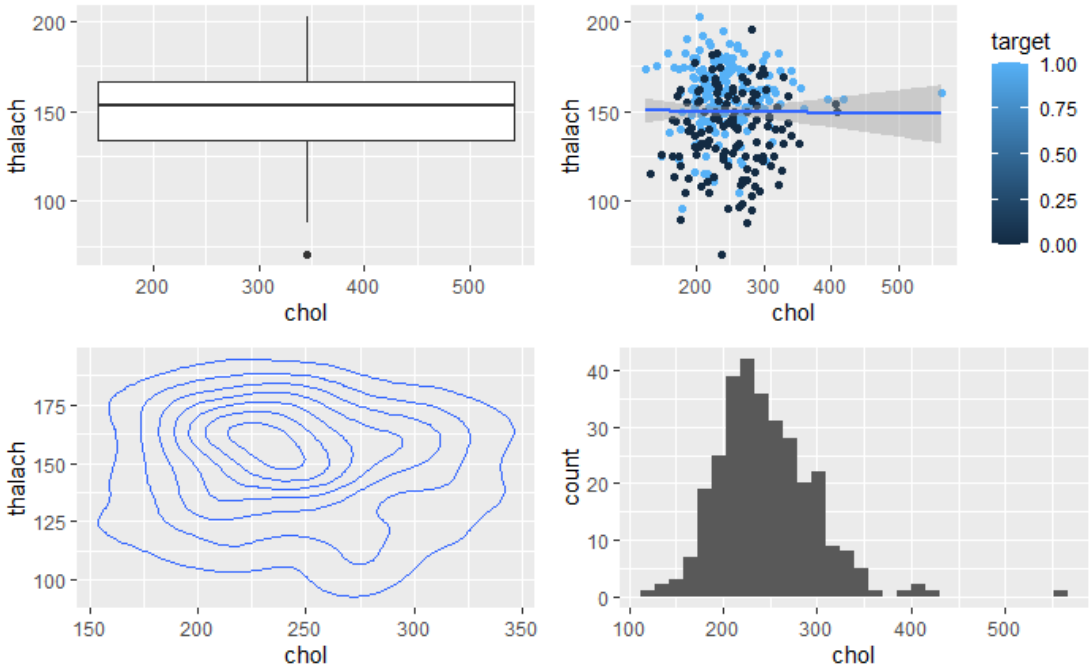
The darker and wider circles represent strong relation between coefficients i.e the variables. With the help of this plot we can see that many variables like cp, thalach,slope... are somewhat strongly correlated to target variable.Using this information we carry our further analysis. So now we try to see how is relation of age with target using different plots.



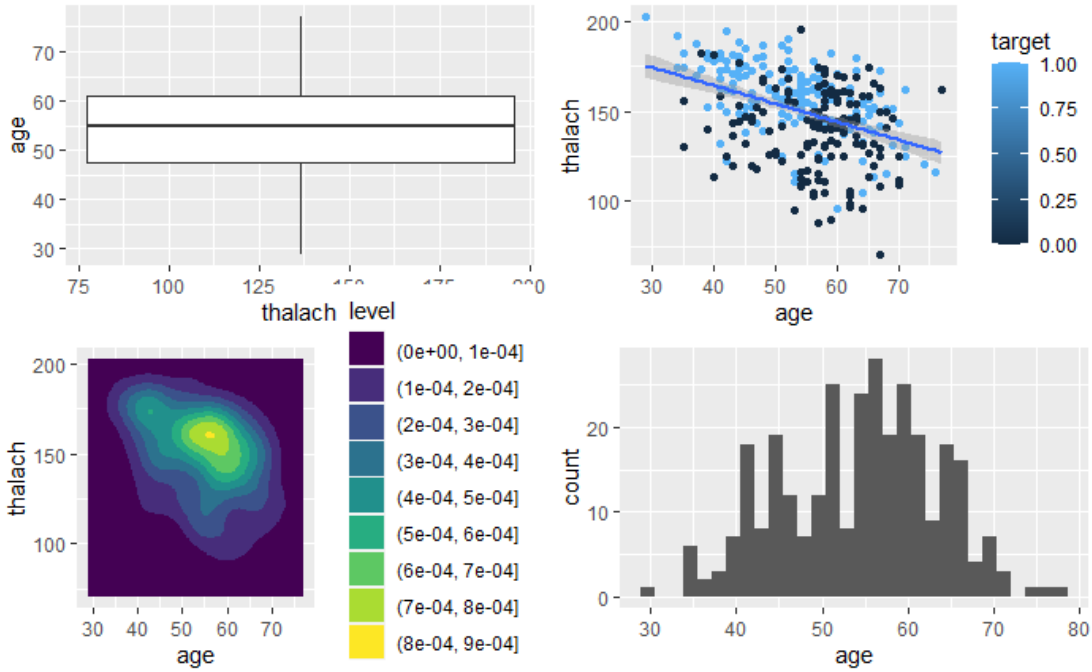
Now for our linear model we need a continuous variable, we will consider thalach i.e. the maximum heart rate so using this we can compare the target heart rate and estimated heart

rate of a person for this we see the correlation plot and plot some variables relation to thalach variable.

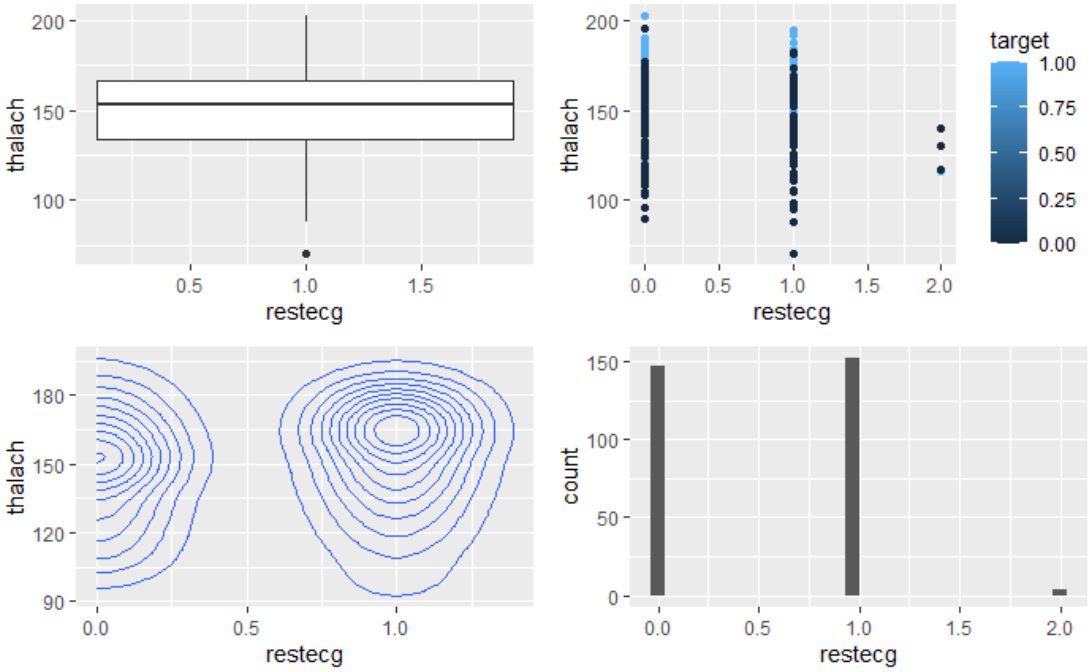
First we see plot of cholestrol against max heart rate.



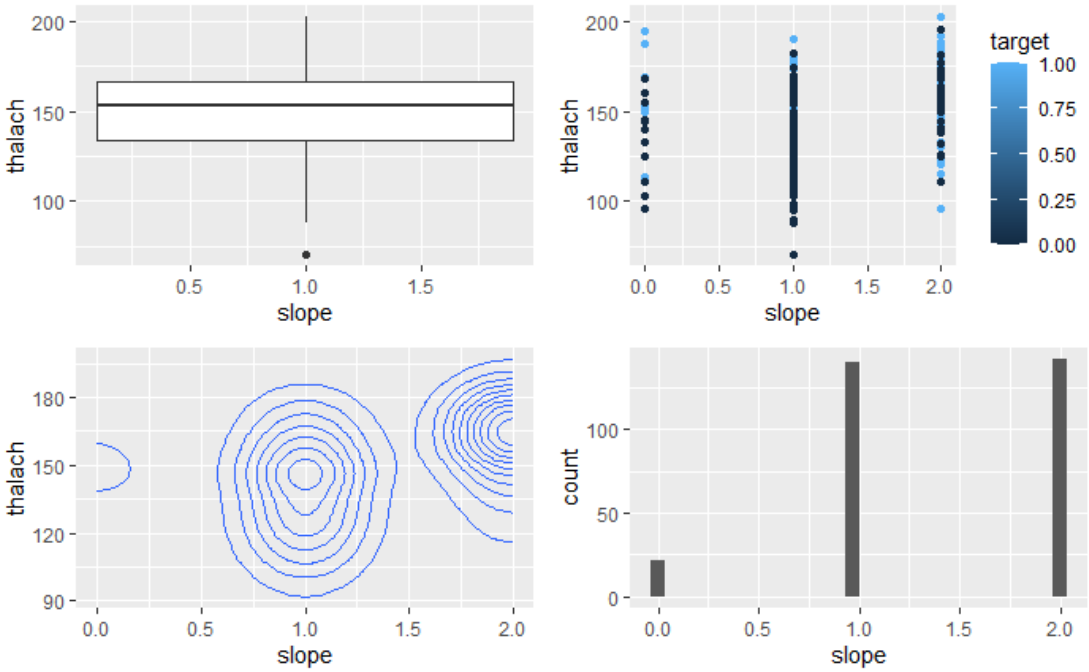
Now we see some plots of age against max heart rate(thalach).



Restecg vs Thalach



slope vs thalach.



Linear Models

We choose our dependent variable to be thalach

Residuals:

Min 1Q Median 3Q Max

```
-58.999 -10.398   1.847  11.663  48.252

Coefficients:    Estimate Std. Error t value Pr(>|t|)

(Intercept) 152.46400   12.35466  12.341  < 2e-16 ***

age         -0.85317    0.12923  -6.602 1.94e-10 ***

sex          1.16007    2.49406   0.465 0.642187

cp           2.15525    1.19332   1.806 0.071943 .

trestbps     0.12492    0.06450   1.937 0.053752 .

chol         0.03724    0.02160   1.724 0.085696 .

fbs          1.80824    3.06602   0.590 0.555807

restecg     -1.39223    2.05629  -0.677 0.498912

exang       -9.48144    2.61776  -3.622 0.000345 ***

oldpeak     -0.89911    1.19081  -0.755 0.450837

slope        7.55891    2.14669   3.521 0.000499 ***

ca          -0.41182    1.16383  -0.354 0.723707

thal         2.14087    1.86376   1.149 0.251634

target       7.97976    2.98741   2.671 0.007988 **

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 18.21 on 289 degrees of freedom

Multiple R-squared:  0.3952,        Adjusted R-squared:  0.368

F-statistic: 14.53 on 13 and 289 DF,  p-value: < 2.2e-16
```

We see that the Adjusted R squared value is 0.368 which is not good. We observe that the summary shows some significant variables like age, exang,slope, target. Now we try focus on these significant variables and comeup with a better model.

Our final model looks like.

model2 <- lm(thalach~age+cp+trestbps+chol+exang+slope+target,data=data)

```
Call:

lm(formula = thalach ~ age + cp + trestbps + chol + exang + slope +

    target, data = data)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-59.288 -10.140   1.905  11.671  44.759

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 155.87051   10.64180  14.647  < 2e-16 ***
age          -0.86973    0.12444  -6.989 1.85e-11 ***
cp            2.28549    1.17251   1.949  0.05221 .
trestbps      0.12711    0.06288   2.021  0.04415 *
chol          0.03843    0.02072   1.854  0.06467 .
exang        -9.23031    2.57770  -3.581  0.00040 ***
slope         8.32783    1.83600   4.536 8.36e-06 ***
target        7.37733    2.59998   2.837  0.00486 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.12 on 295 degrees of freedom
Multiple R-squared:  0.3886,        Adjusted R-squared:  0.3741
F-statistic: 26.79 on 7 and 295 DF,  p-value: < 2.2e-16
```

The summary of this Linear model shows that R squared value is 0.3741 which is also not good but atleast better than previous ones. Now we check for normality of this model using Shapiro Wilks test. And the test shows that p-value is less than 0.05 which means we reject our null hypothesis that the model is Normal.

So we normalise the model using boxcox. After normalising the data we get the Adjusted R-squared value to 0.3883 which is the best so far.So we can say that the max heart rate can be predicted using the heart attack dataset but the results are weak they can be better with more data or better variables.

ANOVA

ONE WAY ANOVA

For one way ANOVA we consider the variables restbps i.e. resting blood pressure and electro cardio graph results

First we check the basic assumptions of ANOVA

Variable type: ANOVA requires a mix of one continuous quantitative dependent variable and one qualitative independent variable (with at least 2 levels which will determine the groups to compare).

Independence: the data, collected from a representative and randomly selected portion of the total population, should be independent.

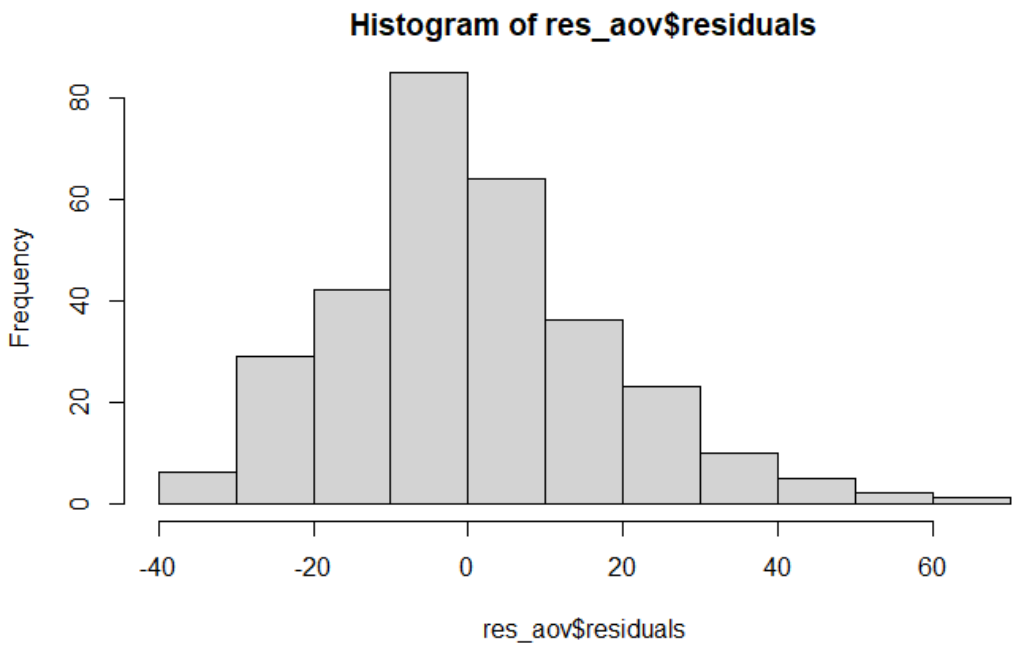
Normality: Residuals2 should follow approximately a normal distribution.

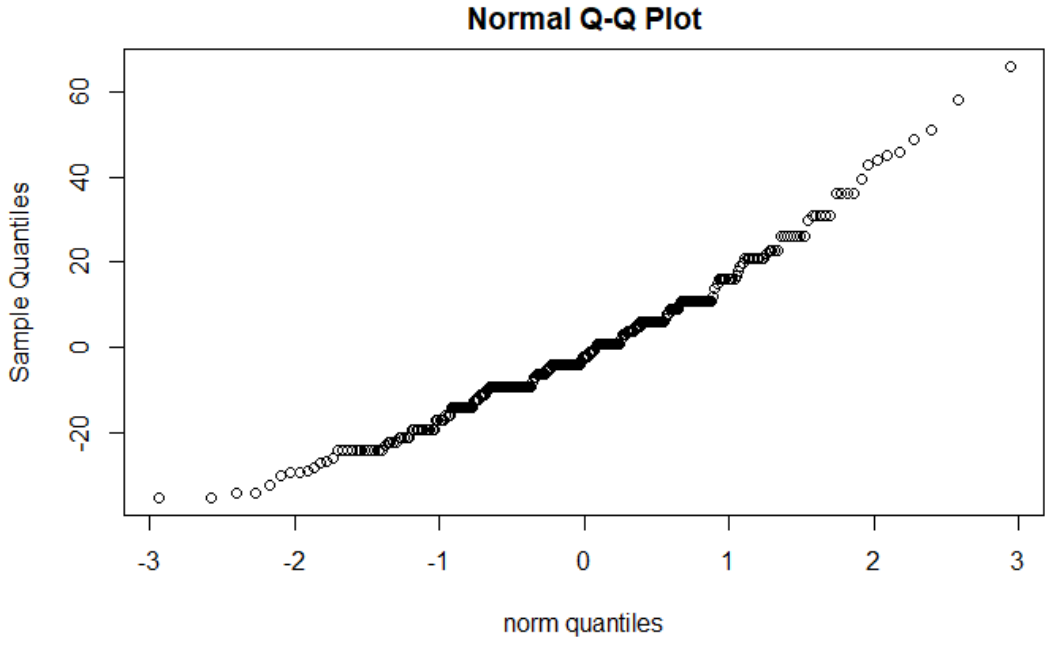
Equality of variances: the variances of the different groups should be equal in the populations (an assumption called homogeneity of the variances, or even sometimes referred as homoscedasticity, as opposed to heteroscedasticity if variances are different across groups).

This assumption can be tested graphically (boxplot or dotplot), or more formally via the Levene's test (`leveneTest(variable ~ group)` from the `{car}` package) or Bartlett's test, among others.

First we check visually and then use the levenTest...

```
res_aov <- aov(trestbps ~ as.factor(restecg),data=data)
```





visually things look normal. We check again using LeveneTest.

we get that

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	2	0.7079	0.4935

300

So we see that the p-value is higher than 0.05 that means we do not reject our null hypothesis of assuming that variances are equal

Applying ONE WAY ANOVA:

We get the p-value = 0.6037 which means we accept that means of the target data(0-low possibility 1-high possibility) have same mean for resting blood pressure in the dataset

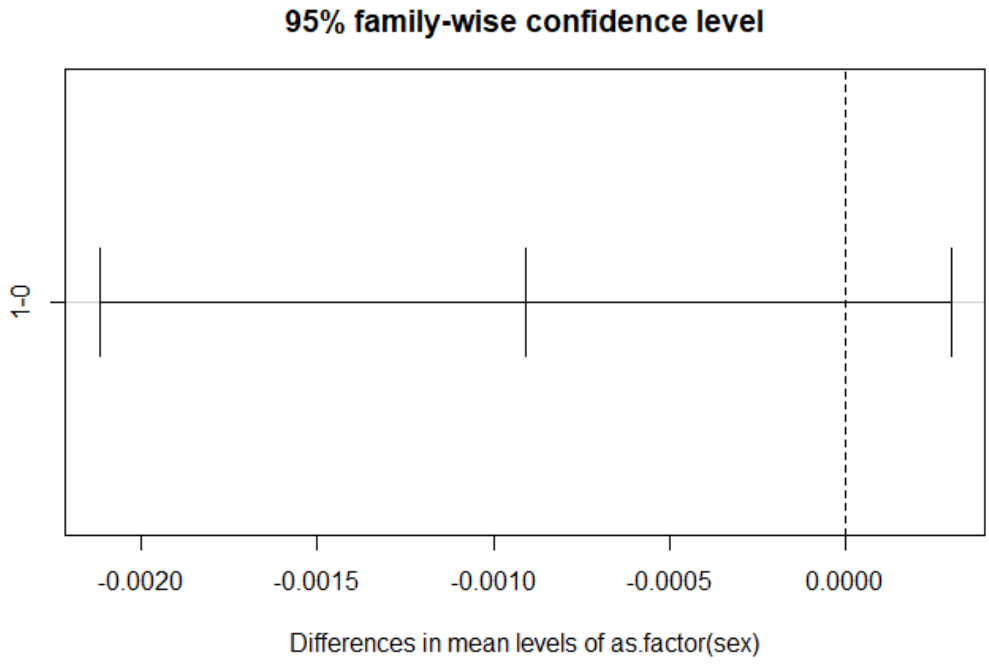
TWO WAY ANOVA

For Two way ANOVA we select the variables resting blood pressure , target and gender .

Using Shapiro wilks test we see that the model for two way ANOVA is not normal. So we apply boxcox to it to normalise it.

After normalising the model we check with levene test and get the p-value 0.435

Also after applying TukeyHSD test we get the plot where mean value lies below 0 that means the that means mean target-values(0-low possibility, 1-high possibility) and gender(0-female, 1-male) is same w.r.t. resting blood pressure.



Logistic Regression

Now we apply logistic regression to the data in order to predict the heart attack possibility.

First we divide i.e. split the dataset into training data and test data some 250 rows for training and remaining 53 for test data.

After creating the model with oldpeak , ca, cp, exang and thal variables against target we get accuracy of 0.836

And we get Sensitivity as 0.9270, Specificity as 0.7257 , Balanced Accuracy as 0.8263 which is pretty balanced.