Heart Disease UCI

Dhiraj Bankar

2020-Nov-14

**Heart Disease UCI Dataset analysis**

#### A. Short summary of data

## 'data.frame': 303 obs. of 14 variables:  
## $ age : int 63 37 41 56 57 57 56 44 52 57 ...  
## $ sex : int 1 1 0 1 0 1 0 1 1 1 ...  
## $ cp : int 3 2 1 1 0 0 1 1 2 2 ...  
## $ trestbps: int 145 130 130 120 120 140 140 120 172 150 ...  
## $ chol : int 233 250 204 236 354 192 294 263 199 168 ...  
## $ fbs : int 1 0 0 0 0 0 0 0 1 0 ...  
## $ restecg : int 0 1 0 1 1 1 0 1 1 1 ...  
## $ thalach : int 150 187 172 178 163 148 153 173 162 174 ...  
## $ exang : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ oldpeak : num 2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...  
## $ slope : int 0 0 2 2 2 1 1 2 2 2 ...  
## $ ca : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ thal : int 1 2 2 2 2 1 2 3 3 2 ...  
## $ target : int 1 1 1 1 1 1 1 1 1 1 ...

## age sex cp trestbps   
## Min. :29.00 Min. :0.0000 Min. :0.000 Min. : 94.0   
## 1st Qu.:47.50 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:120.0   
## Median :55.00 Median :1.0000 Median :1.000 Median :130.0   
## Mean :54.37 Mean :0.6832 Mean :0.967 Mean :131.6   
## 3rd Qu.:61.00 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:140.0   
## Max. :77.00 Max. :1.0000 Max. :3.000 Max. :200.0   
## chol fbs restecg thalach   
## Min. :126.0 Min. :0.0000 Min. :0.0000 Min. : 71.0   
## 1st Qu.:211.0 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:133.5   
## Median :240.0 Median :0.0000 Median :1.0000 Median :153.0   
## Mean :246.3 Mean :0.1485 Mean :0.5281 Mean :149.6   
## 3rd Qu.:274.5 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:166.0   
## Max. :564.0 Max. :1.0000 Max. :2.0000 Max. :202.0   
## exang oldpeak slope ca   
## Min. :0.0000 Min. :0.00 Min. :0.000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.00 1st Qu.:1.000 1st Qu.:0.0000   
## Median :0.0000 Median :0.80 Median :1.000 Median :0.0000   
## Mean :0.3267 Mean :1.04 Mean :1.399 Mean :0.7294   
## 3rd Qu.:1.0000 3rd Qu.:1.60 3rd Qu.:2.000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :6.20 Max. :2.000 Max. :4.0000   
## thal target   
## Min. :0.000 Min. :0.0000   
## 1st Qu.:2.000 1st Qu.:0.0000   
## Median :2.000 Median :1.0000   
## Mean :2.314 Mean :0.5446   
## 3rd Qu.:3.000 3rd Qu.:1.0000   
## Max. :3.000 Max. :1.0000

We have 303 observation with 14 variables:

* age: age
* sex: sex
* cp: chest pain type (4 values)
* trestbps: resting blood pressure
* chol: serum cholesterol in mg/dl
* fbs: fasting blood sugar > 120 mg/dl
* restecg: resting electrocardiograph results (values 0,1,2)
* thalach: maximum heart rate achieved
* exang: exercise induced angina
* oldpeak: ST depression induced by exercise relative to rest
* slope: the slope of the peak exercise ST segment
* ca: number of major vessels (0-3) colored by flourosopy
* thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
* target: Target Groups

#### B. Find the correleation between all veriables

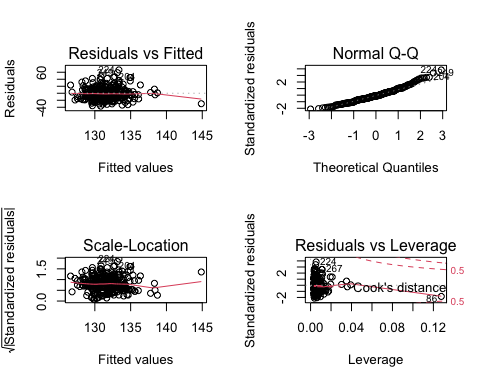
## age sex cp trestbps chol  
## age 1.00000000 -0.09844660 -0.06865302 0.27935091 0.213677957  
## sex -0.09844660 1.00000000 -0.04935288 -0.05676882 -0.197912174  
## cp -0.06865302 -0.04935288 1.00000000 0.04760776 -0.076904391  
## trestbps 0.27935091 -0.05676882 0.04760776 1.00000000 0.123174207  
## chol 0.21367796 -0.19791217 -0.07690439 0.12317421 1.000000000  
## fbs 0.12130765 0.04503179 0.09444403 0.17753054 0.013293602  
## restecg -0.11621090 -0.05819627 0.04442059 -0.11410279 -0.151040078  
## thalach -0.39852194 -0.04401991 0.29576212 -0.04669773 -0.009939839  
## exang 0.09680083 0.14166381 -0.39428027 0.06761612 0.067022783  
## oldpeak 0.21001257 0.09609288 -0.14923016 0.19321647 0.053951920  
## slope -0.16881424 -0.03071057 0.11971659 -0.12147458 -0.004037770  
## ca 0.27632624 0.11826141 -0.18105303 0.10138899 0.070510925  
## thal 0.06800138 0.21004110 -0.16173557 0.06220989 0.098802993  
## target -0.22543872 -0.28093658 0.43379826 -0.14493113 -0.085239105  
## fbs restecg thalach exang oldpeak  
## age 0.121307648 -0.11621090 -0.398521938 0.09680083 0.210012567  
## sex 0.045031789 -0.05819627 -0.044019908 0.14166381 0.096092877  
## cp 0.094444035 0.04442059 0.295762125 -0.39428027 -0.149230158  
## trestbps 0.177530542 -0.11410279 -0.046697728 0.06761612 0.193216472  
## chol 0.013293602 -0.15104008 -0.009939839 0.06702278 0.053951920  
## fbs 1.000000000 -0.08418905 -0.008567107 0.02566515 0.005747223  
## restecg -0.084189054 1.00000000 0.044123444 -0.07073286 -0.058770226  
## thalach -0.008567107 0.04412344 1.000000000 -0.37881209 -0.344186948  
## exang 0.025665147 -0.07073286 -0.378812094 1.00000000 0.288222808  
## oldpeak 0.005747223 -0.05877023 -0.344186948 0.28822281 1.000000000  
## slope -0.059894178 0.09304482 0.386784410 -0.25774837 -0.577536817  
## ca 0.137979327 -0.07204243 -0.213176928 0.11573938 0.222682322  
## thal -0.032019339 -0.01198140 -0.096439132 0.20675379 0.210244126  
## target -0.028045760 0.13722950 0.421740934 -0.43675708 -0.430696002  
## slope ca thal target  
## age -0.16881424 0.27632624 0.06800138 -0.22543872  
## sex -0.03071057 0.11826141 0.21004110 -0.28093658  
## cp 0.11971659 -0.18105303 -0.16173557 0.43379826  
## trestbps -0.12147458 0.10138899 0.06220989 -0.14493113  
## chol -0.00403777 0.07051093 0.09880299 -0.08523911  
## fbs -0.05989418 0.13797933 -0.03201934 -0.02804576  
## restecg 0.09304482 -0.07204243 -0.01198140 0.13722950  
## thalach 0.38678441 -0.21317693 -0.09643913 0.42174093  
## exang -0.25774837 0.11573938 0.20675379 -0.43675708  
## oldpeak -0.57753682 0.22268232 0.21024413 -0.43069600  
## slope 1.00000000 -0.08015521 -0.10476379 0.34587708  
## ca -0.08015521 1.00000000 0.15183213 -0.39172399  
## thal -0.10476379 0.15183213 1.00000000 -0.34402927  
## target 0.34587708 -0.39172399 -0.34402927 1.00000000

## age trestbps  
## age 1.00 0.28  
## trestbps 0.28 1.00  
##   
## n= 303   
##   
##   
## P  
## age trestbps  
## age 0   
## trestbps 0

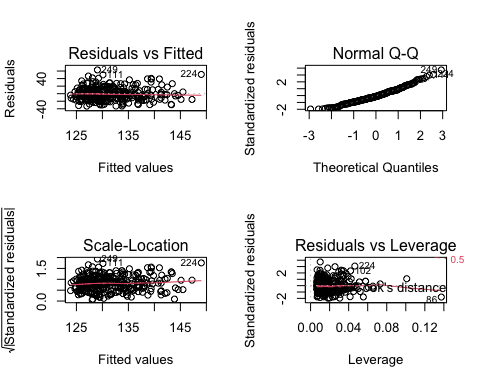
##   
## Pearson's product-moment correlation  
##   
## data: heartDF$age and heartDF$trestbps  
## t = 5.0475, df = 301, p-value = 7.762e-07  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1720897 0.3800657  
## sample estimates:  
## cor   
## 0.2793509

From correlation result below are my observations on heart disease data \* The trestbps has 27% affecting to the heart. \* cp has 27% affecting to the heart. \* Cholesterol has 21% affecting to the heart. \* oldpeakhas 21% affecting to the heart. \* fbs has 12% affecting to the heart. \* exang has 9% affecting to the heart. \* thal has 6% affecting to the heart. \* All other variables has a -ve correlation values(Means they are not impacting or not related to heart disease). We are going to ignore it for further analysis \* For Further analysis lets use only age, trestbps , chol, fbs, exang, oldpeak, ca, thal for main analysis

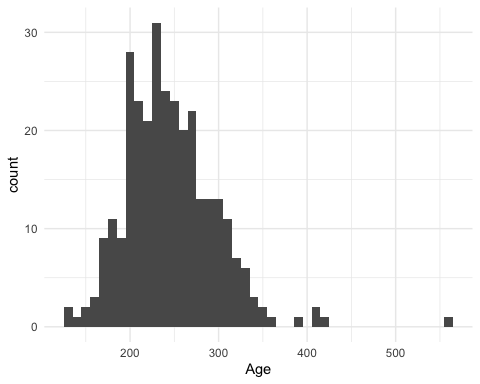
##   
## Call:  
## lm(formula = trestbps ~ chol, data = heartDF)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36.821 -11.259 -1.946 9.513 66.637   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 121.35976 4.87052 24.917 <2e-16 \*\*\*  
## chol 0.04168 0.01935 2.153 0.0321 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.43 on 301 degrees of freedom  
## Multiple R-squared: 0.01517, Adjusted R-squared: 0.0119   
## F-statistic: 4.637 on 1 and 301 DF, p-value: 0.03208



##   
## Call:  
## lm(formula = trestbps ~ chol + oldpeak + exang + fbs + ca, data = heartDF)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -33.003 -11.418 -2.039 10.695 62.957   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 118.08407 4.79559 24.623 < 2e-16 \*\*\*  
## chol 0.03690 0.01890 1.952 0.05187 .   
## oldpeak 2.71054 0.89579 3.026 0.00270 \*\*   
## exang 0.02688 2.17444 0.012 0.99015   
## fbs 8.41367 2.76588 3.042 0.00256 \*\*   
## ca 0.51587 0.99098 0.521 0.60306   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 16.95 on 297 degrees of freedom  
## Multiple R-squared: 0.08154, Adjusted R-squared: 0.06608   
## F-statistic: 5.273 on 5 and 297 DF, p-value: 0.0001172



#### C. What is most common age when heart disease gets started?

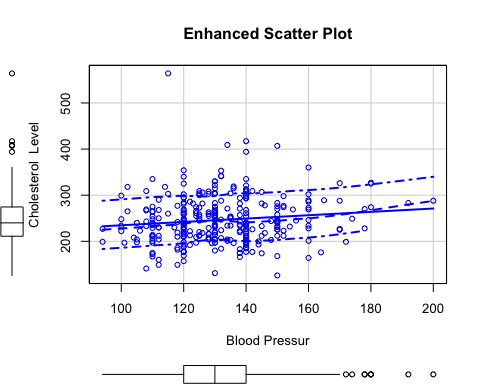
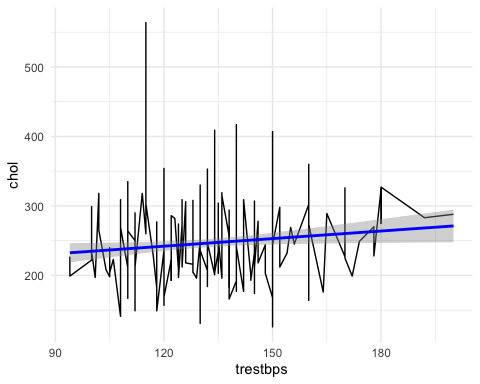


* From the above histogram I would say heart disease are getting started from thirties of the age

#### D. How much is the effect of cholesterol on heart?

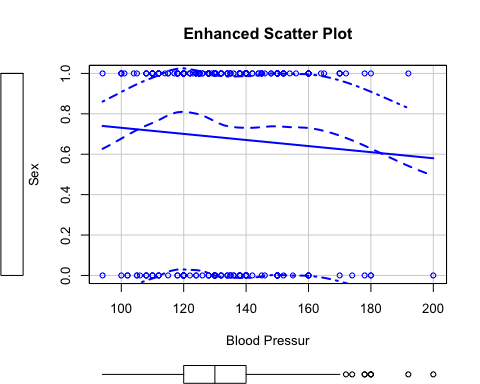
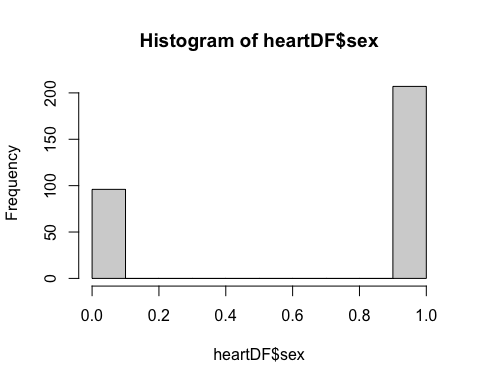
From the correlation found that the Cholesterol has 21% affecting to the heart. Initially before starting the analysis I thought that the Cholesterol is the most important factor which cause the heart disease. But blood pressure and stress has 27% of effect on heart disease.

## `geom\_smooth()` using formula 'y ~ x'



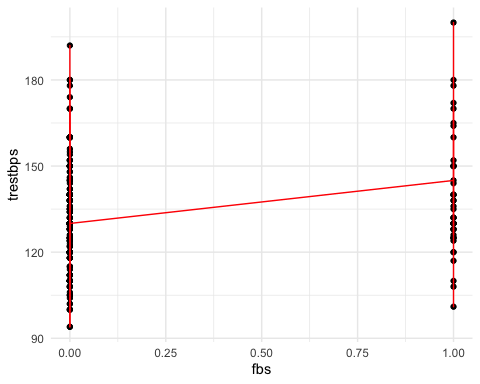
* The major factor of heart disease is blood pressure and Cholesterol. If you observe above plot, the increase in Cholesterol increases the blood pressure level.

#### E. Who is most affected by heart disease (Male or Female)?



* Mail has more heart disease that women.

#### F. Is fasting blood sugar has a relation with heart disease?



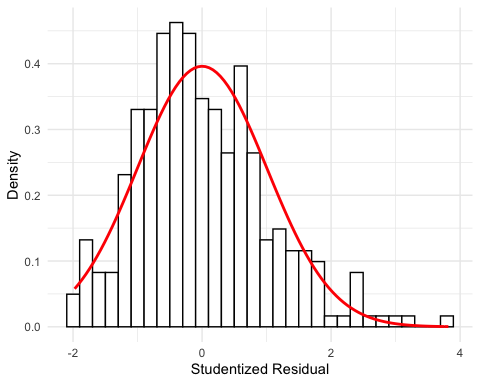
* With fasting the blood pressure increases so the patient with heart disease should not do long fasting.

#### G. How much help or improvement we can achieve with the help of exercise and diet?

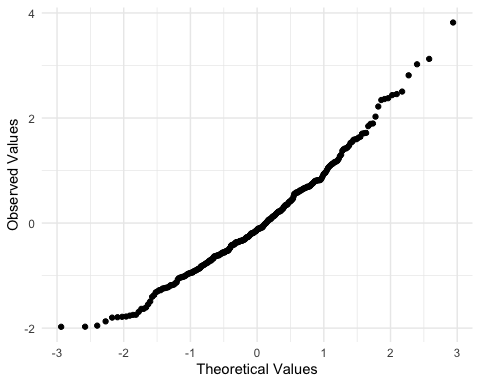
* I couldn’t find the exact answer of this question.

#### H. Will create a multi linear model find the appropriate results.

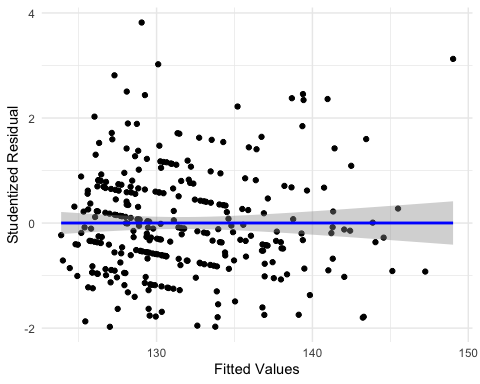
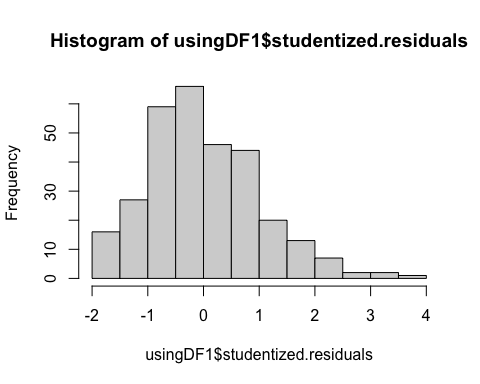
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

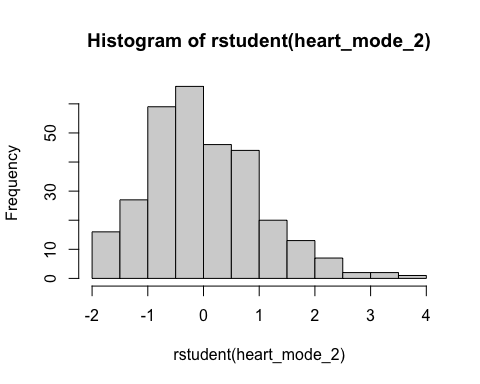


## Warning: `stat` is deprecated



## `geom\_smooth()` using formula 'y ~ x'



* Overall regression model is unbiased. We could summarize saying that the model appears, in most senses, to be both accurate for the sample and generalizable to the population.

#### I. How much is confidence level of the model?

## 2.5 % 97.5 %  
## (Intercept) 1.086464e+02 127.52170683  
## chol -3.002895e-04 0.07410088  
## oldpeak 9.476387e-01 4.47343480  
## exang -4.252385e+00 4.30613667  
## fbs 2.970462e+00 13.85687880  
## ca -1.434361e+00 2.46609683

* What I observed from the outcome of confident interval. A good model will have a small confidence interval, indicating that the value of b is of the b-values tells us about the direction of the relationship between the predictors and outcome. If you observe outcome of my model is very small, means this model is good.

## Add Citations

* Discovering Statistics Using R(Field, Miles, and Field 2012)

# References

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using R*. SAGE Publications. <https://books.google.com/books?id=wd2K2zC3swIC>.