Introduction

The following data is a modified form of a famous dataset from the UCI Machine Learning Data Repository. Researchers collected data related to various physical aspects of many patients. Their objective was to create a way of classifying whether their patients had heart disease or not without performing invasive procedures. To that end, they collected the following information from the patients.

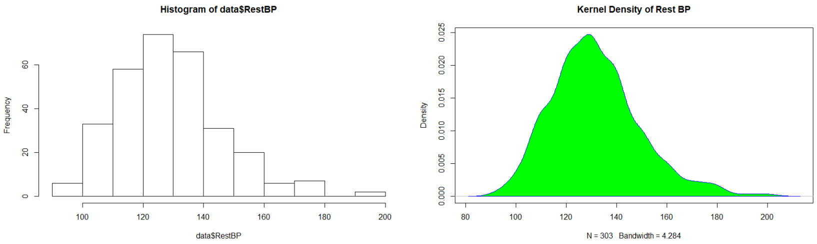
Data

The data file is available online as HeartAbridged.csv.

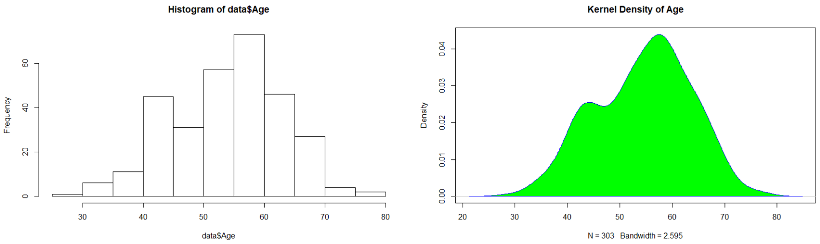
- **Age**: The person's age in years
- **RestBP**: The person's resting blood pressure (mm Hg on admission to the hospital)
- **Chol**: The person's cholesterol measurement in mg/dl
- **MaxHR**: The person's maximum heart rate achieved during controlled exercise
- **AHD**: Whether a person has heart disease. + Yes: The do have a heart disease. + No: They do not have a heart disease.

We have 139 entries in the dataset which says AHD=Yes (Heart Disease) and 164 entries in the dataset which says AHD=No (No Heart Disease). So, this is quite a balanced dataset.
For convenience during fitting a model, the AHD variable is converted into 0s and 1s, i.e., No Heart Disease and Heart Disease, respectively.

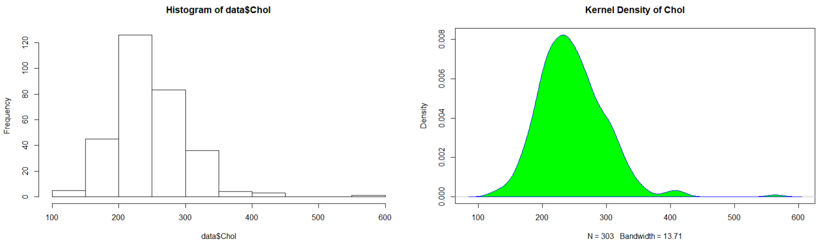
Graphing Distributions of Variables



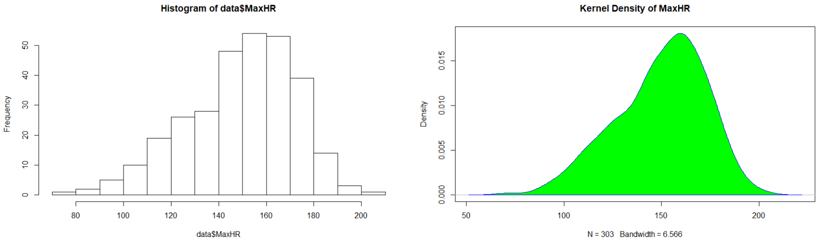
(Unimodal, somewhat Normal like distribution)



(Bimodal)

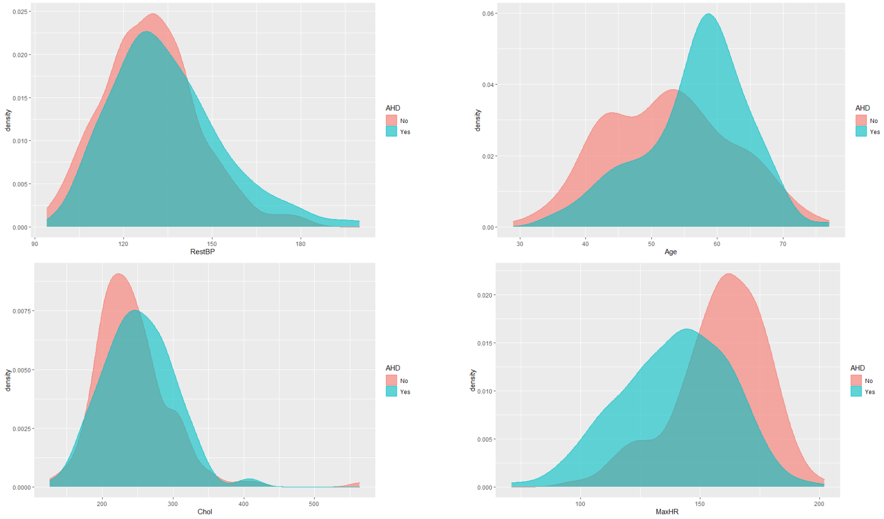


(Unimodal, Normal like distribution if right tail is ignored)



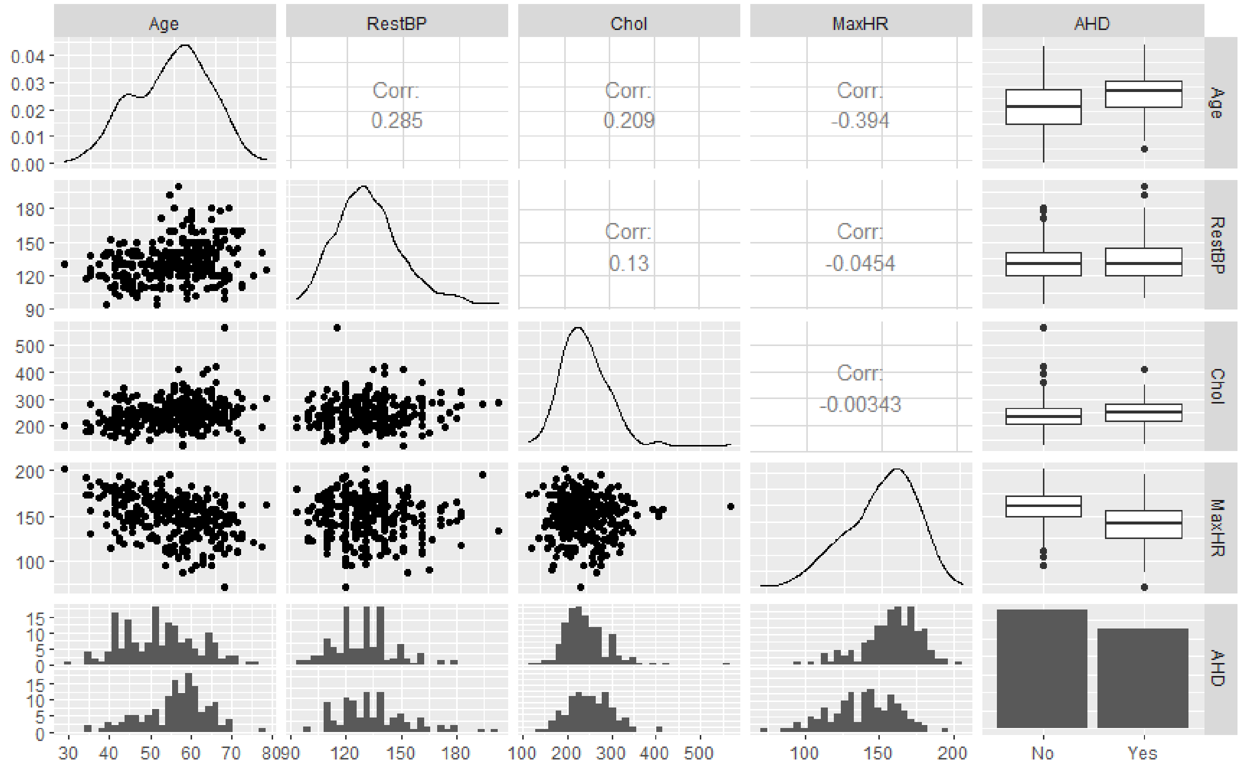
(Unimodal, slightly left tailed)

Graphing Distributions of Variables grouped by AHD



RestBP and Chol have pretty much similar density for heart disease and no heart disease. MaxHR has higher density for cases with no heart disease whereas Age has higher density for cases with heart disease.

Pairwise Scatter Plots for all variables



Out of the 4 variables: Age, MaxHR, Chol, RestBP, there seems to be a positive correlation of 0.285 between Age and RestBP. But the magnitude of correlation for Age and MaxHR is higher. There is a relatively strong negative correlation of -0.394 between Age and MaxHR compared to most other variable pairs. This can be seen in the graphs below. As the age increases, the MaxHR is seen to reduce significantly.

99% confidence for the difference in the means of those with heart disease and those without heart disease

For RestBP and Age, Mean for Heart Disease is more than mean for no Heart Disease.

For Chol and MaxHR, Mean for No Heart Disease is more than mean for Heart Disease.

Hypothesis test for each variable to see if the mean of those with heart disease differs from the mean of those without heart disease

On performing the t-test, we get the following results:

Variable	t value	p value	95% CI	Mean in Group with Heart Disease	Mean in Group with No Heart Disease
RestBP	-2.6152	0.009409	[-9.321,-1.314]	134.57	129.25

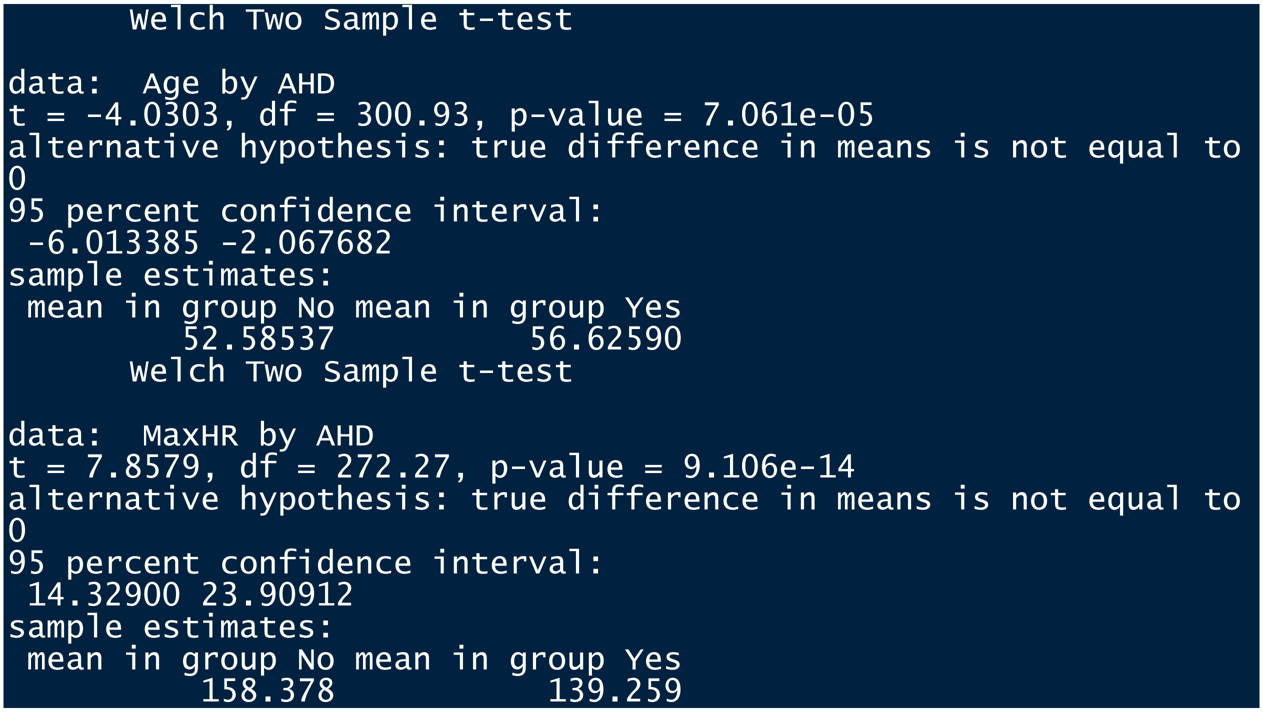
Age	-4.0303	7.061*10^-5	[-6.013,-2.067]	56.63	52.59
MaxHR	7.8579	9.106*10^-14	[14.329,23.909]	139.26	158.38
Chol	-1.4924	0.1366	[-20.484,2.815]	251.48	242.64

Mean of heart patients with RestBP, Age, Chol is much higher to have the disease compared to heart patients with MaxHR, which has a statistically significance of lesser chance of having a heart disease.

Smaller p value for Age and MaxHR indicates strong evidence to reject Heart Disease hypothesis and relatively larger p value for RestBP and Chol indicates weak evidence to reject Heart Disease hypothesis.

<u>Hypothesis Testing of Age and MaxHR</u>

Let us perform a Welch Two Sample T-test.



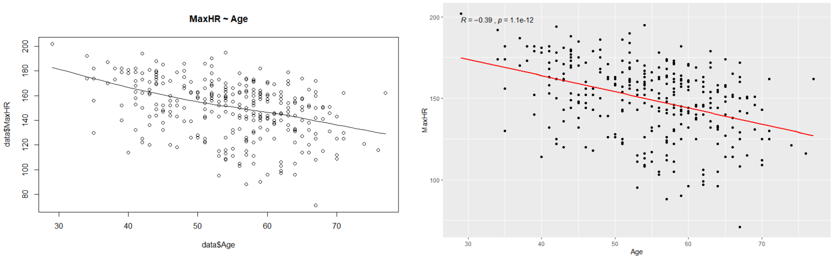
```
      Welch Two Sample t-test

data:  Age by AHD
t = -4.0303, df = 300.93, p-value = 7.061e-05
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -6.013385 -2.067682
sample estimates:
 mean in group No mean in group Yes
         52.58537          56.62590
      Welch Two Sample t-test

data:  MaxHR by AHD
t = 7.8579, df = 272.27, p-value = 9.106e-14
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 14.32900 23.90912
sample estimates:
 mean in group No mean in group Yes
          158.378           139.259
```

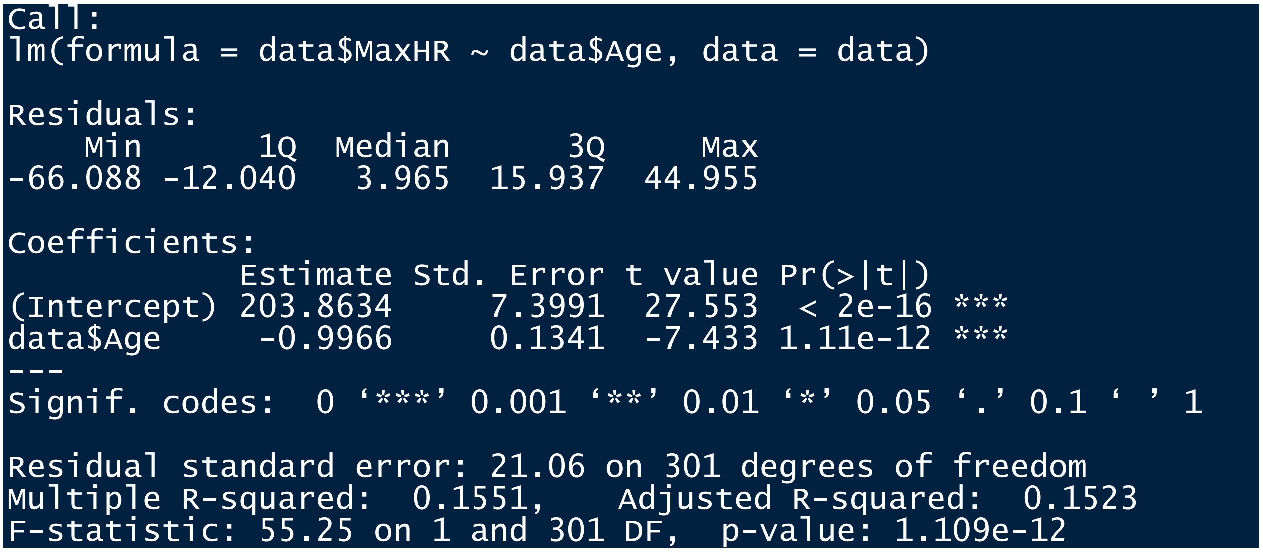
Both the variables are statistically significant because of very low p values. MaxHR might be a more decisive variable because of a lower p value compared to Age. This makes sense because heart disease can be attributed better towards a maximum heart rate achieved compared to the age of an individual. Nowadays, we see people in their early 30s and 40s also suffering from a heart disease. So, MaxHR might be a more crucial factor.

Scatterplot of Age and MaxHR with regression line on it



The response variable is Age and the predictor variable is MaxHR. It makes sense to give the age of a person and predict the MaxHR. General studies also involve identifying heart rate of patients within a certain age period. The scatterplot with the regression line does not seem to fit all the points very well on the line because of many points farther away from the regressor line. The error rate is very high. Since the R value of -0.39 is not very high, this could not be a good prediction.

Creating the linear regression model between the two variables



```
Call:
lm(formula = data$MaxHR ~ data$Age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-66.088 -12.040   3.965  15.937  44.955 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 203.8634     7.3991  27.553  < 2e-16 ***
data$Age     -0.9966     0.1341  -7.433 1.11e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.06 on 301 degrees of freedom
Multiple R-squared:  0.1551,	Adjusted R-squared:  0.1523 
F-statistic: 55.25 on 1 and 301 DF,  p-value: 1.109e-12
```

Now that we have built the linear model, we also have established the relationship between the predictor and response in the form of a mathematical formula for MaxHR as a function for Age. For the above output, we can notice the ‘Coefficients’ part having two components:

Intercept: 203.8634, *Age*: -0.996. These are also called the beta coefficients.

In other words,
MaxHR = Intercept + (β ∗ Age) => MaxHR = 203.8634 - (0.9966∗Age)

The linear model is statistically significant because of a lower p value than 0.05. *Pr(>|t|)* is low, the coefficients are significant (significantly different from zero).

Since the hypothesis test for MaxHR gave a mean value between a heart disease or not within [139, 158] and from the results below using predict(),

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  127.1   143.1   148.1   149.6   156.0   175.0 
```

We can safely say that MaxHR is a good predictor variable compared to other variables.

99% Confidence interval for the slope is:

```
                 0.5 %     99.5 %
(Intercept) 184.683104 223.043676
data$Age     -1.344226  -0.649058
```

Quantiles of the age variable:

```
20% 40% 60% 80% 
 45  53  58  62 
```

Predictions of MaxHR with the above quantiles:

```
      fit              lwr              upr       
 Min.   :127.1   Min.   :120.7   Min.   :133.5  
 1st Qu.:143.1   1st Qu.:140.1   1st Qu.:146.0  
 Median :148.1   Median :145.6   Median :150.5  
 Mean   :149.6   Mean   :146.4   Mean   :152.9  
 3rd Qu.:156.0   3rd Qu.:153.1   3rd Qu.:158.9  
 Max.   :175.0   Max.   :167.8   Max.   :182.1  
```

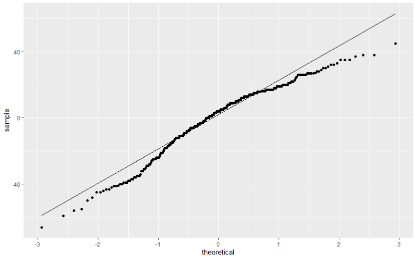
The MaxHR confidence interval associated with Ages is (120.7, 182.1). This means that, according to our model, a person with a certain age of has, on average, a MaxHR ranging between 120.7 and 182.1.

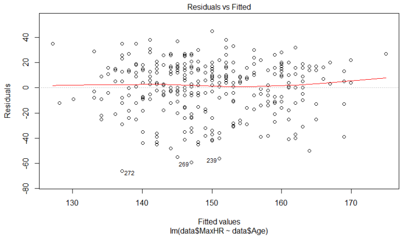
Checking Regression Assumptions

A QQ-Plot is drawn to assess Normality. Normality does not hold true completely most of the points do not fit well on the line. We observe many outliers.

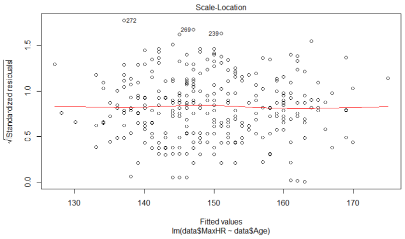
sum(data$residuals - (data$MaxHR - data$fitted)) = `1.256772e-13`

Residuals have a consistent mean of zero and constant variability



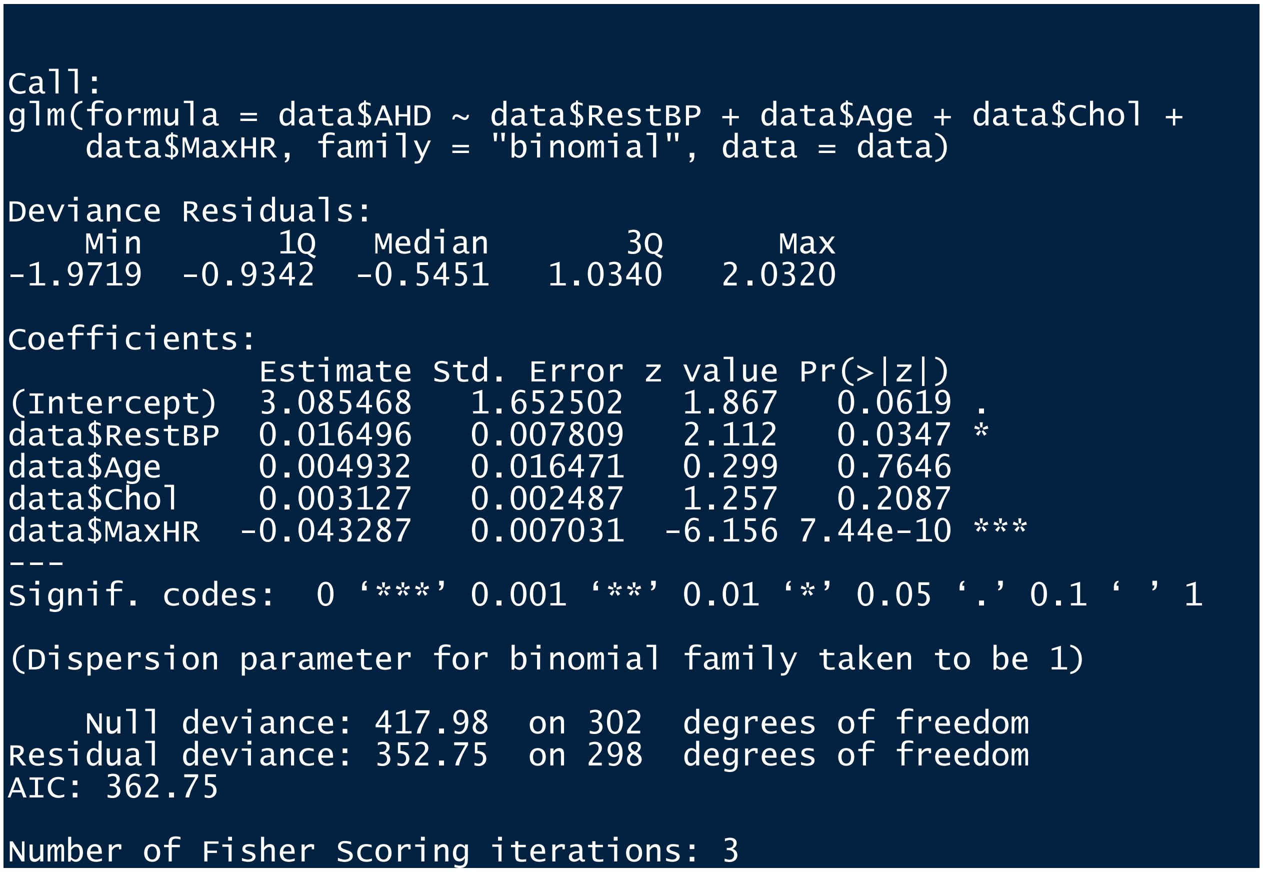


The residual plot shows no fitted pattern. The red line is horizontal at zero.



There is no pattern in the residuals plot. So, there is a linear relationship between the response and predictor variables.

Logistic Regression Model



```
Call:
glm(formula = data$AHD ~ data$RestBP + data$Age + data$Chol +
    data$MaxHR, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9719  -0.9342  -0.5451   1.0340   2.0320

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.085468   1.652502   1.867   0.0619 .
data$RestBP  0.016496   0.007809   2.112   0.0347 *
data$Age     0.004932   0.016471   0.299   0.7646
data$Chol    0.003127   0.002487   1.257   0.2087
data$MaxHR  -0.043287   0.007031  -6.156 7.44e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.98  on 302  degrees of freedom
Residual deviance: 352.75  on 298  degrees of freedom
AIC: 362.75

Number of Fisher Scoring iterations: 3
```

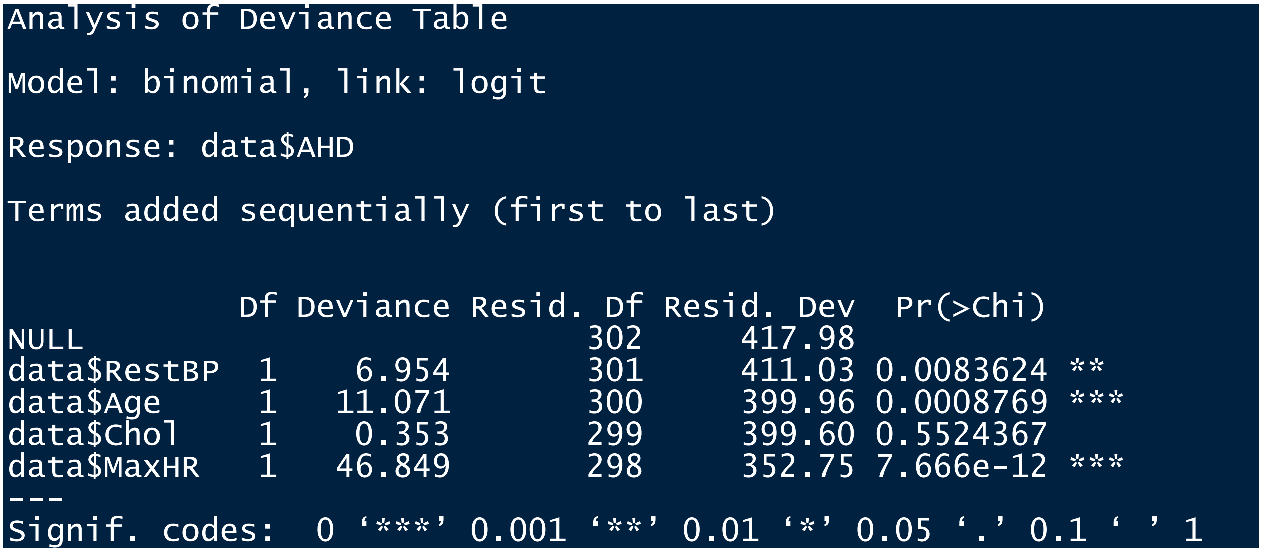
For every one-unit change in RestBP, the log odds of AHD (versus No AHD) increases by 0.016

For every one-unit change in Age, the log odds of AHD (versus No AHD) increases by 0.004

For every one-unit change in Chol, the log odds of AHD (versus No AHD) increases by 0.0.003

For every one-unit change in MaxHR, the log odds of AHD (versus No AHD) decreases by 0.0.043

MaxHR is the most statistically significant variable compared to the other three variables.



```
Analysis of Deviance Table

Model: binomial, link: logit

Response: data$AHD

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                          302     417.98              
data$RestBP  1    6.954       301     411.03 0.0083624 ** 
data$Age     1   11.071       300     399.96 0.0008769 ***
data$Chol    1    0.353       299     399.60 0.5524367    
data$MaxHR   1   46.849       298     352.75 7.666e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better. Analyzing the table we can see the drop in deviance when adding each variable one at a time.

We would like to see how the model is doing when predicting *y* on a new set of data. By setting the parameter type=’response’, R will output probabilities in the form of P(y=1|X). Our decision boundary will be 0.5. If P(y=1|X) > 0.5 then y = 1 otherwise y=0.

We get an accuracy of about 71% which is not that great. Linear Regression could be a better model to fit.

Conclusions, Limitations and Future Work

As we have more data in our hand, the accuracy of our model and the insights that we gain would obviously be better. Having said that, we could probably focus on other parameters such as obesity, diabetes, smoking habits, family genetics, inactivity, etc. Identifying the impact of these parameters could enhance the current work. The problem is it would be hard to gather such information from patients because they might be hesitant or wrong information. The cost of gathering data would go high. Identifying a specific gender where the cases are high would be important. It is important to understand the preexisting conditions, allergies, and other medical conditions of patients. There is huge investment to cut down this disease and active research is prevailing. Scientists are trying to predict the disease before a patient tries to show symptoms. All these factors could greatly influence in a better model. Using wearable technology or cameras would show the actions of patients and this can be an extraordinarily useful indication. Having all these parameters, we might have to build a more complex model to do our prediction.