Regression Analysis on Heart Disease

Anjali Priya

2022-09-14

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
setwd("C:/personal files/data analytics/docs/MODULE 3/T3")
getwd()
```

```
## [1] "C:/personal files/data analytics/docs/MODULE 3/T3"
```

loading library

```
library(ggplot2)
library(tidyr)
library(readxl)
library(readr)
library(car)
```

```
## Loading required package: carData
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.6      v stringr 1.4.0
## v purrr   0.3.4      v forcats 0.5.1
## v dplyr   1.0.10
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(broom)
```

BRIEF

The dataset “heart.data” contains observations on the percentage of people cycling to work each day, the percentage of people smoking, and the percentage of people with heart diseases in a hypothetical sample of 498 towns. The rates of cycling to work range between 1 and 75%, rates of smoking between 0.5 and 30%, and rates of heart disease between 0.5% and 20.5%. The Surveyor want to check the relationship between cycling to work and heart diseases using above data

Q1. Read, call and view the data in r

```
cardiodata_1 = read_excel("C:/personal files/data analytics/docs/MODULE 3/T3/heart.data.xlsx"
)
View(cardiodata_1)
```

Q2. check the dimension and characteristics of the dataset.

```
dim(cardiodata_1)
```

```
## [1] 498   3
```

```
class(cardiodata_1)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

- *As seen from the given dataset are in tibble, dataframe class .*
- *Dimension of data set is [498 x 3] .*

Q3. check properties of the variables.

```
str(cardiodata_1)
```

```
## tibble [498 x 3] (S3: tbl_df/tbl/data.frame)
##  $ cycling       : num [1:498] 30.8 65.13 1.96 44.8 69.43 ...
##  $ smoking       : num [1:498] 10.9 2.22 17.59 2.8 15.97 ...
##  $ heart_diseases: num [1:498] 11.77 2.85 17.18 6.82 4.06 ...
```

- *As seen from the given data all the variables are in numeric form.*

Q4. Check the first and last few observations from the dataset

```
head(cardiodata_1 , 10)
```

```
## # A tibble: 10 x 3
##    cycling smoking heart_diseases
##      <dbl>   <dbl>          <dbl>
##  1   30.8    10.9           11.8
##  2   65.1     2.22           2.85
##  3    1.96   17.6           17.2
##  4   44.8     2.80           6.82
##  5   69.4    16.0            4.06
##  6   54.4    29.3            9.55
##  7   49.1     9.06           7.62
##  8    4.78   12.8           15.9
##  9   65.7    12.0            3.07
## 10   35.3    23.3           12.1
```

```
tail(cardiodata_1, 10)
```

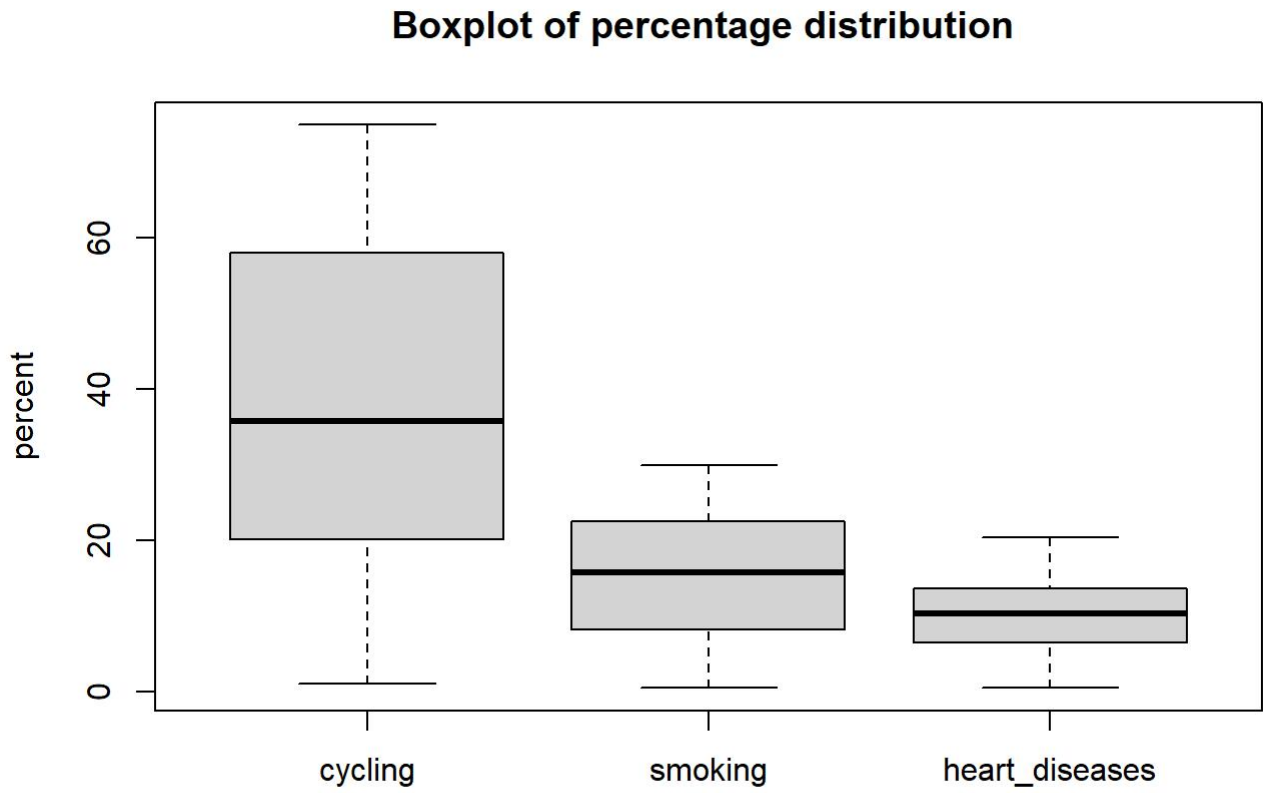
```
## # A tibble: 10 x 3
##    cycling smoking heart_diseases
##      <dbl>   <dbl>          <dbl>
##  1   41.6    0.533           5.93
##  2   70.2   16.3             4.71
##  3   50.8    6.03            5.59
##  4   68.9   10.5             3.11
##  5   21.6    7.60           12.4
##  6   47.7   27.6            11.3
##  7   45.1   21.4             9.62
##  8    8.28   6.42           13.5
##  9   42.3   20.7            10.1
## 10   30.8   23.6            11.8
```

Q5. Create the summary

```
summary(cardiodata_1)
```

```
##     cycling          smoking        heart_diseases
##  Min.   : 1.119   Min.   : 0.5259   Min.   : 0.5519
##  1st Qu.:20.205   1st Qu.: 8.2798   1st Qu.: 6.5137
##  Median :35.824   Median :15.8146   Median :10.3853
##  Mean   :37.788   Mean   :15.4350   Mean   :10.1745
##  3rd Qu.:57.853   3rd Qu.:22.5689   3rd Qu.:13.7240
##  Max.   :74.907   Max.   :29.9467   Max.   :20.4535
```

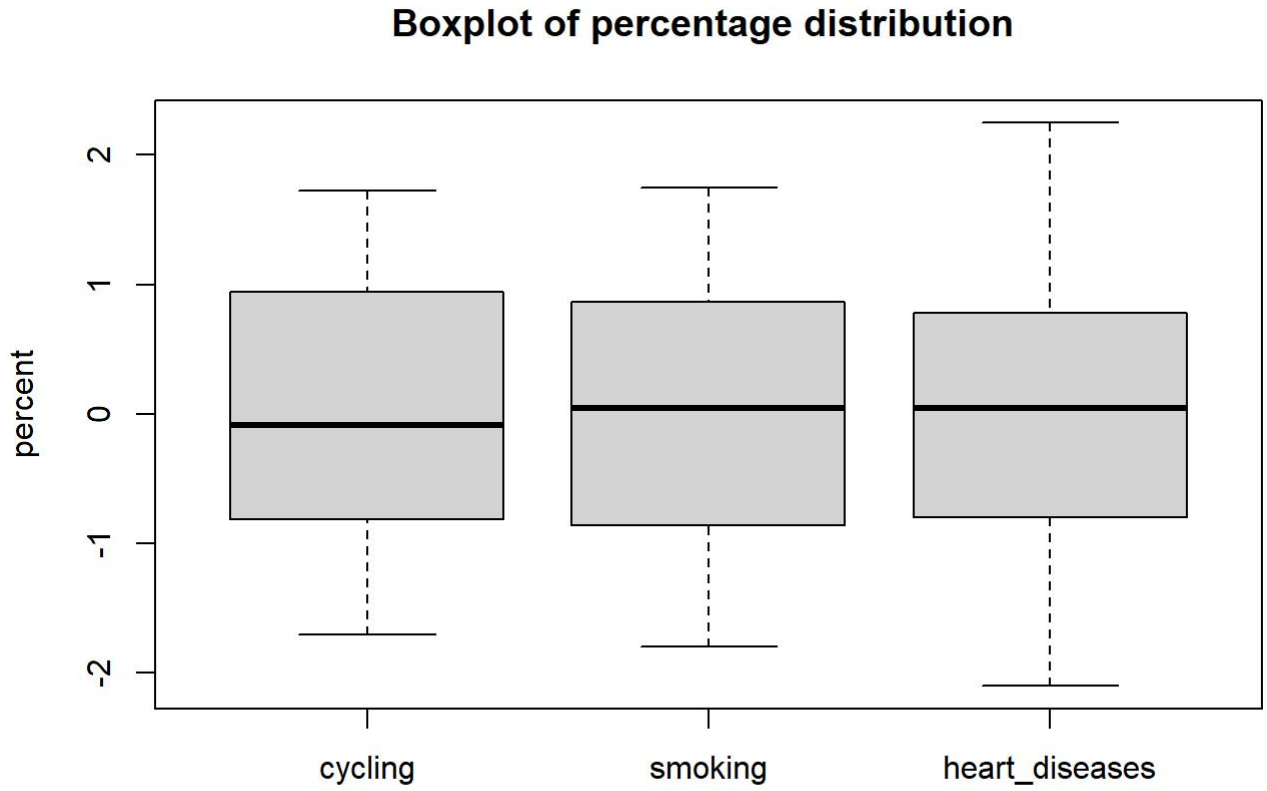
```
cardiodata_1 %>%
  boxplot( main = "Boxplot of percentage distribution " , ylab="percent")
```




```
lapply(cardiodata_1, boxplot.stats)
```

```
## $cycling
## $cycling$stats
## [1]  1.119154 20.197206 35.824459 57.978406 74.907111
##
## $cycling$n
## [1] 498
##
## $cycling$conf
## [1] 33.14949 38.49942
##
## $cycling$out
## numeric(0)
##
##
## $smoking
## $smoking$stats
## [1]  0.525850  8.278009 15.814614 22.585020 29.946743
##
## $smoking$n
## [1] 498
##
## $smoking$conf
## [1] 14.80166 16.82757
##
## $smoking$out
## numeric(0)
##
##
## $heart_diseases
## $heart_diseases$stats
## [1]  0.5518982  6.5126756 10.3852547 13.7241183 20.4534962
##
## $heart_diseases$n
## [1] 498
##
## $heart_diseases$conf
## [1]  9.874674 10.895836
##
## $heart_diseases$out
## numeric(0)
```

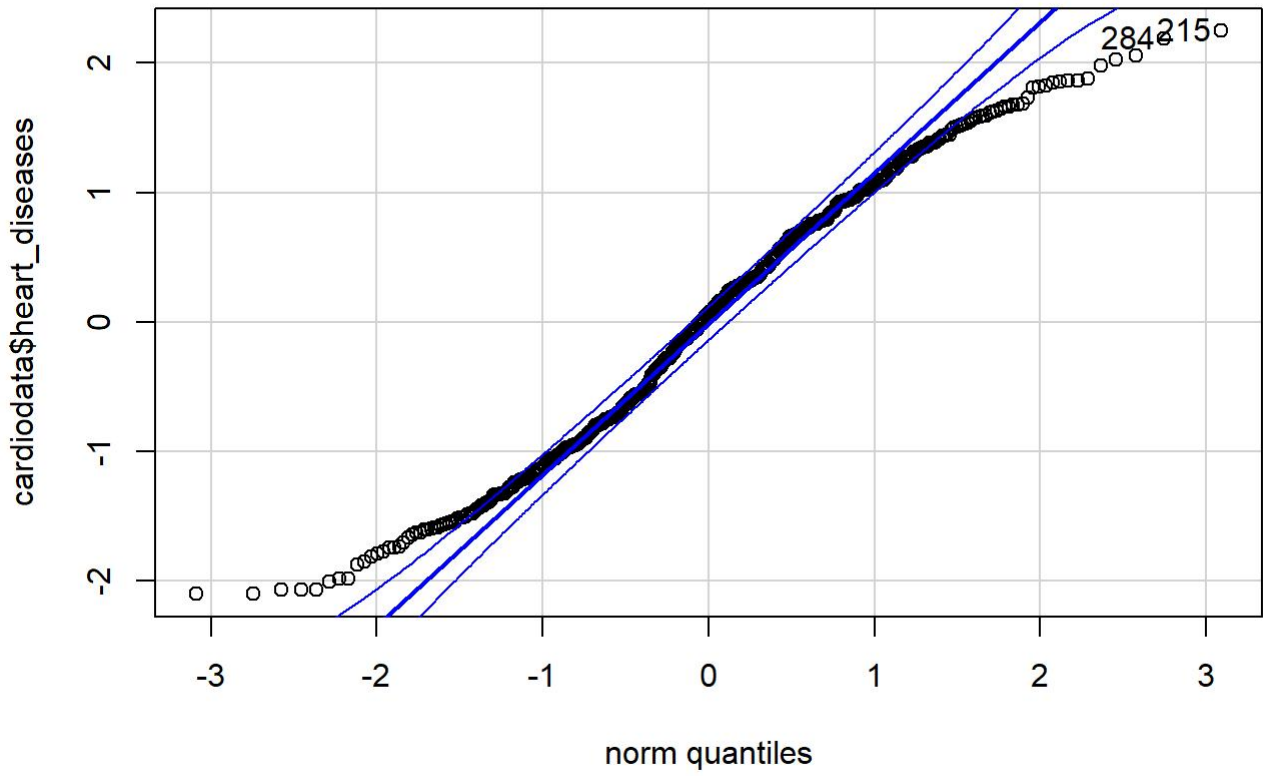
```
## standarizing data
cardiodata=as.data.frame( scale(cardiodata_1))
cardiodata %>%
  boxplot( main = "Boxplot of percentage distribution " , ylab="percent")
```



- *data under each category is normally distributed around median*
- *data under cycling category has largest spread and higher median for 498 towns*

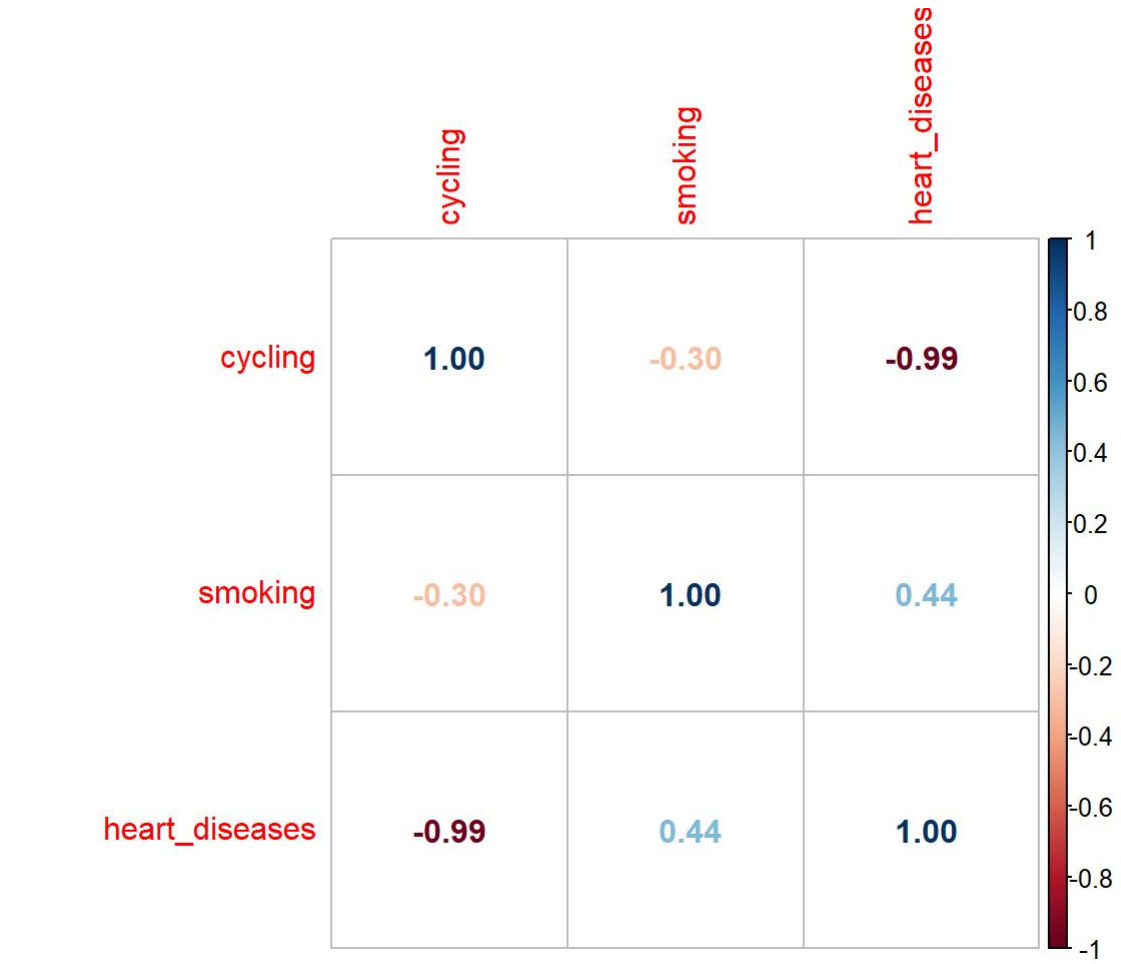
Q6. Check normality of dependent variable and linearity between variables

```
qqPlot(cardiodata$heart_diseases)
```

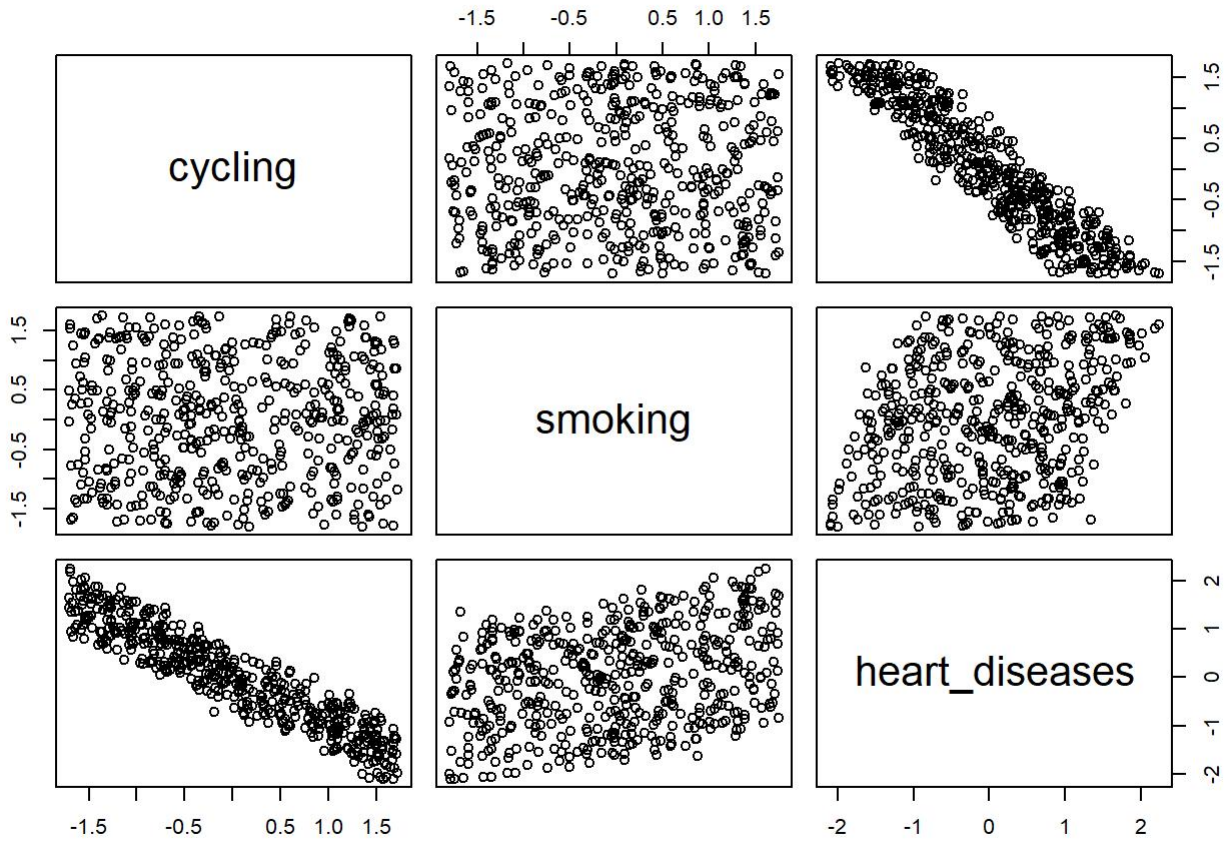



```
## [1] 215 284
```

```
correlationmatrix = cor(cardiodata)
corrplot(cor(correlationmatrix),method ="number")
```




```
plot(cardiodata)
```



- Seeing the qqplot we can say that the data are approximately normally distributed
- There is high correlation negative correlation between cycling-heart disease , weak positive correlation between smoking and heart disease and no correlation between cycling and smoking*
- *No outliers in dataset

> Q7. To check if there is a linear relationship between “cycling to work”, “smoking”, and “heart disease” in our hypothetical survey of 498 towns. Create and run a regression model using “heart.data” dataset and also create summary of the regression model

```
model_1=lm(heart_diseases~ cycling + smoking ,cardiodata )
summary(model_1)
```

```
##
## Call:
## lm(formula = heart_diseases ~ cycling + smoking, data = cardiodata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47658 -0.09762  0.00792  0.09671  0.42283
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.119e-17  6.410e-03    0.00        1
## cycling     -9.403e-01  6.418e-03 -146.53   <2e-16 ***
## smoking      3.234e-01  6.418e-03   50.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1431 on 495 degrees of freedom
## Multiple R-squared:  0.9796, Adjusted R-squared:  0.9795
## F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

```
model_2=lm(heart_diseases ~ cycling , cardiodata)
summary(model_2)
```

```
##
## Call:
## lm(formula = heart_diseases ~ cycling, data = cardiodata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88111 -0.26369 -0.00088  0.25179  0.79674
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.015e-17  1.585e-02    0.00        1
## cycling     -9.355e-01  1.587e-02  -58.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3538 on 496 degrees of freedom
## Multiple R-squared:  0.8751, Adjusted R-squared:  0.8748
## F-statistic:  3474 on 1 and 496 DF,  p-value: < 2.2e-16
```

- As seen from the summary stats of linear regression its evident that including cycling and smoking both give the better fit model by looking at the Adjusted R sqrd value which is 97.6% and all the parameters are highly significant.

Q8. Store the output of the regression model and print coefficients.

```
model_1
```

```
##
## Call:
## lm(formula = heart_diseases ~ cycling + smoking, data = cardiodata)
##
## Coefficients:
## (Intercept)      cycling      smoking
##  -6.119e-17   -9.403e-01    3.234e-01
```

```
coff_m1=model_1$coefficients
```

Q9. Print the Estimate, Std. Error, t-value, and p-value for the independent variables (i.e., cycling and smoking)

```
summary(model_1)$coefficients
```

```
##                  Estimate  Std. Error       t value      Pr(>|t|)
## (Intercept) -6.119484e-17 0.006410474 -9.546071e-15  1.000000e+00
## cycling     -9.403500e-01 0.006417655 -1.465255e+02  0.000000e+00
## smoking      3.233643e-01 0.006417655  5.038667e+01 5.192235e-197
```

Q10. Print residual standard error, r-squared, adjusted r-squared, f-statistic, p-value from the output

```
RSE = summary(model_1)$sigma
R_sqrd=summary(model_1)$r.squared
AdjR_sqrd=summary(model_1)$adj.r.squared
F_value=summary(model_1)$fstatistic
```

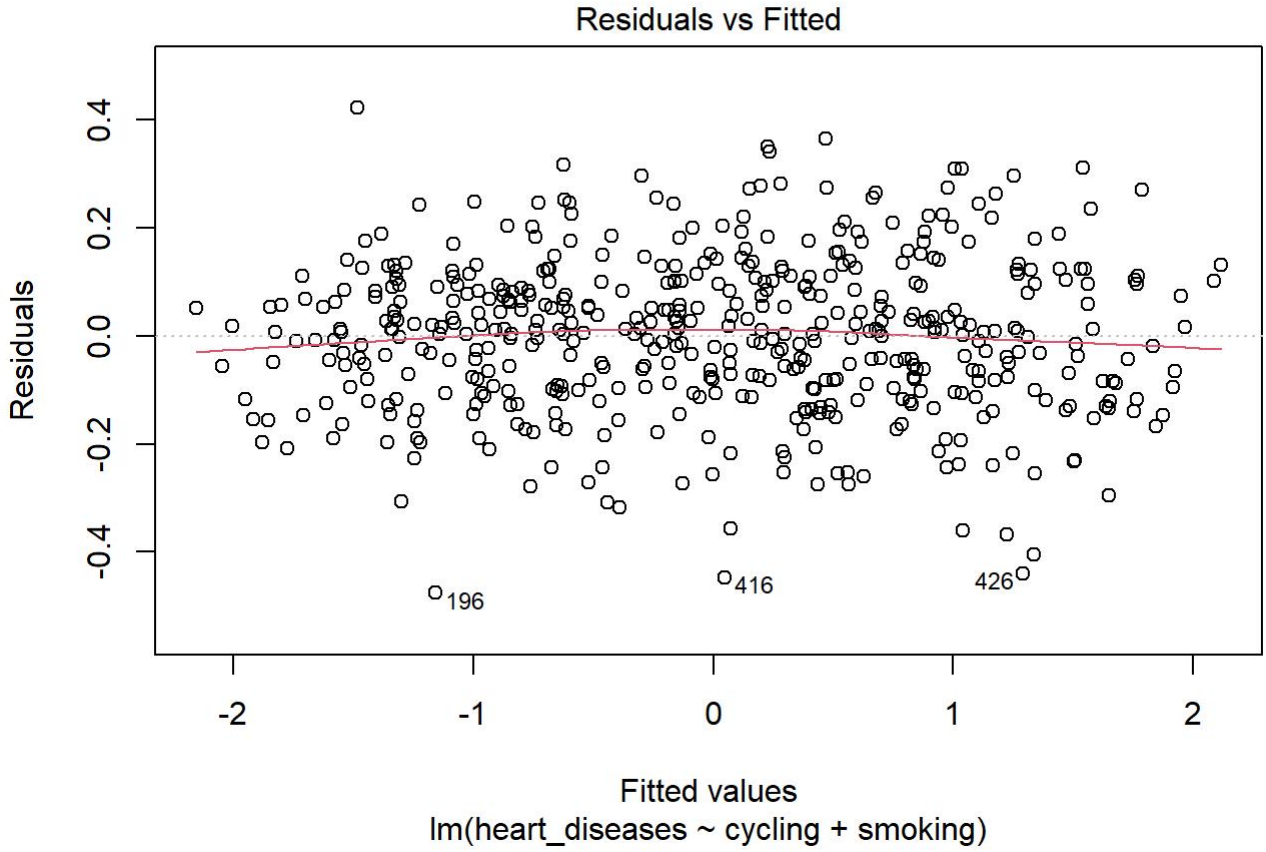
Q11. Compute the confidence intervals

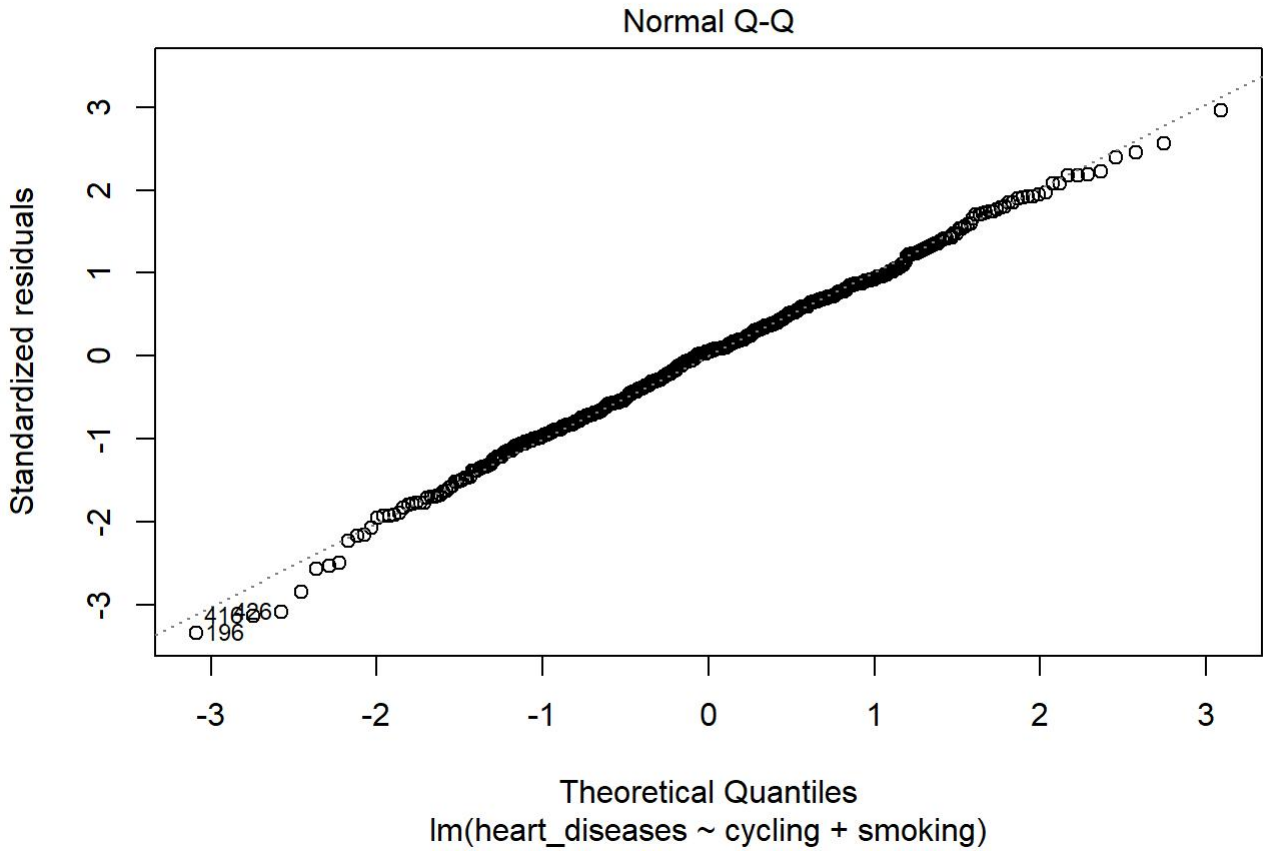
```
conf_interval=predict(model_1 , cardiodata , interval ="confidence")
head(conf_interval,10)
```

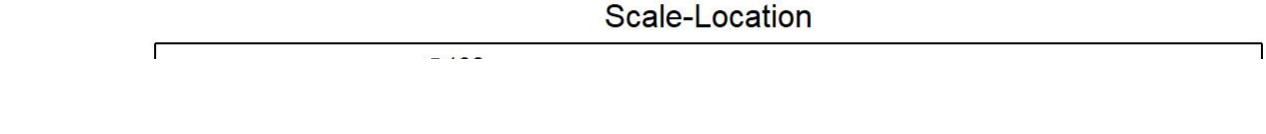
```
##           fit        lwr        upr
## 1   0.1288328  0.1139247  0.1437409
## 2  -1.7123317 -1.7411415 -1.6835219
## 3   1.6523906  1.6276168  1.6771644
## 4  -0.7996914 -0.8230828 -0.7763000
## 5  -1.3639947 -1.3864395 -1.3415499
## 6  -0.1852071 -0.2115583 -0.1588559
## 7  -0.7418803 -0.7591522 -0.7246083
## 8   1.3433182  1.3199248  1.3667116
## 9  -1.3575020 -1.3788956 -1.3361084
## 10  0.4167086  0.3992823  0.4341350
```

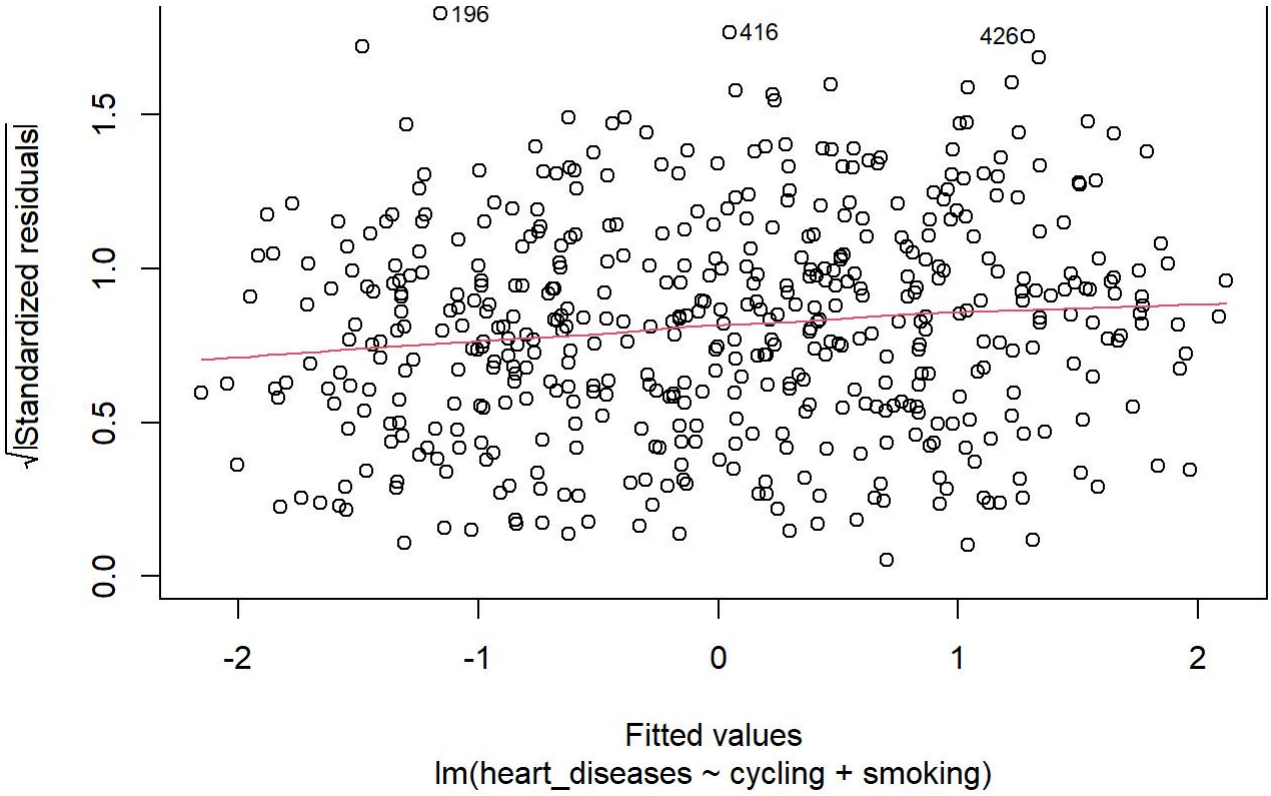
Q12. Create the diagnostic plots and discuss them in brief.

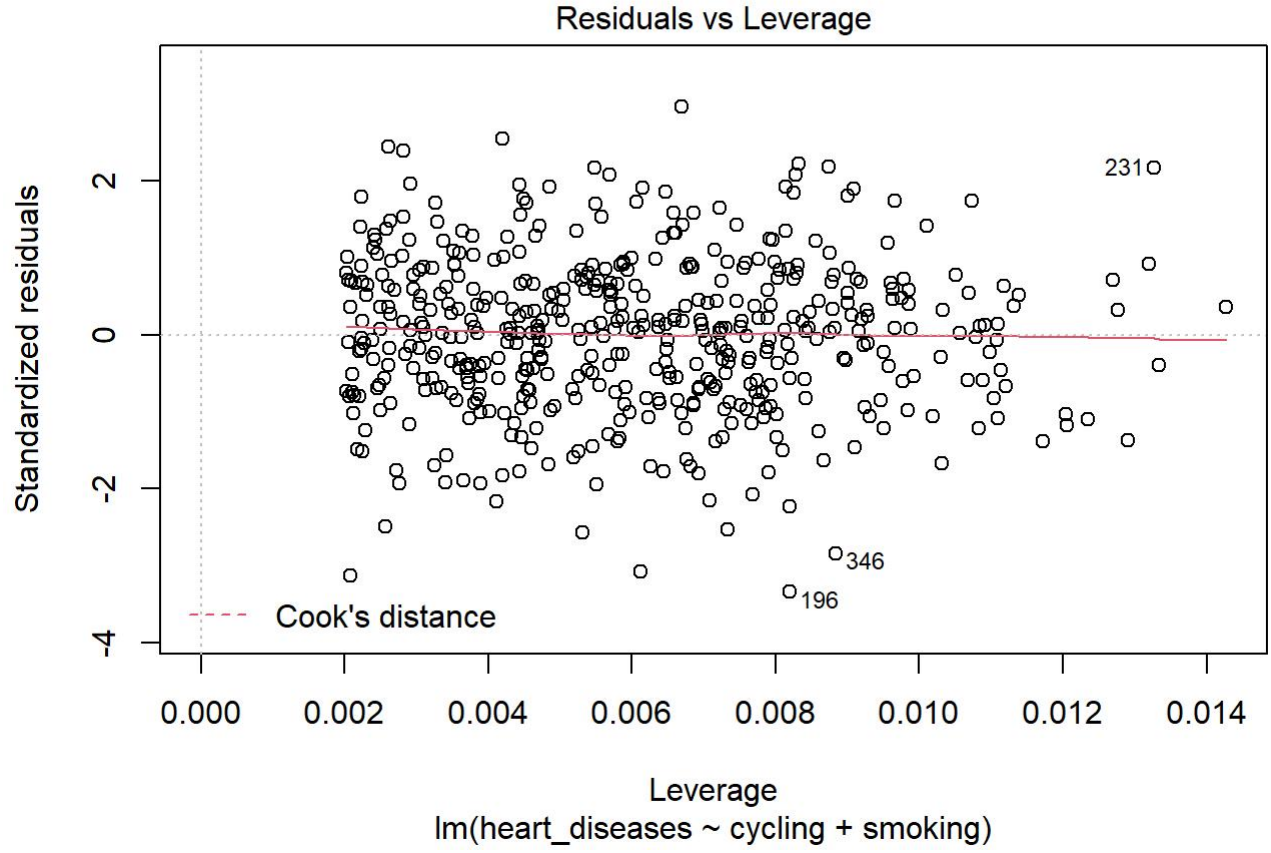
```
plot(model_1)
```


Residuals vs Fitted
Residuals
0.4
0.2
0.0
-0.2
-0.4
196
416
426
-2
-1
0
1
2
Fitted values
lm(heart_diseases ~ cycling + smoking)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
196
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(heart_diseases ~ cycling + smoking)


Scale-Location






```
heart_fitted=augment(model_1)$.fitted
```

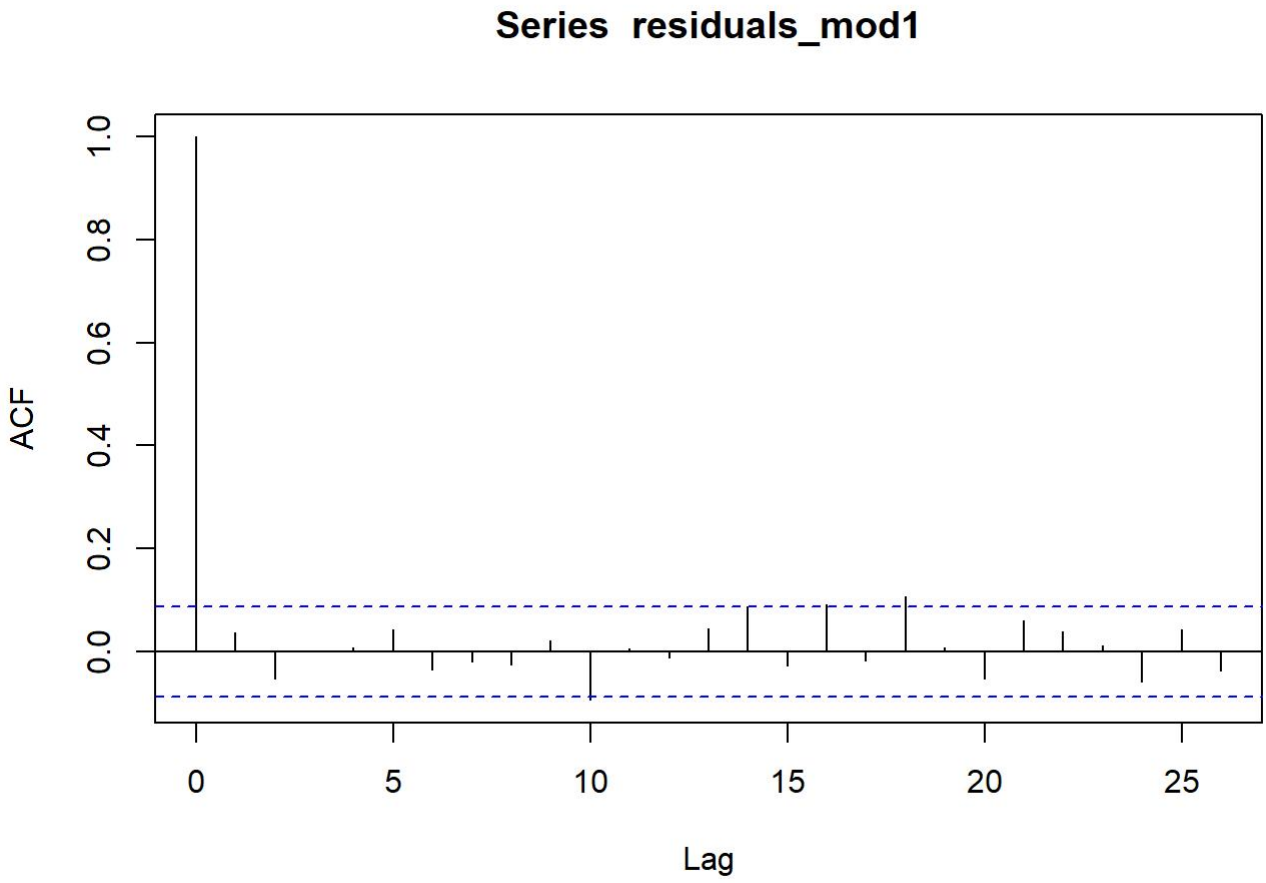
- Residual vs Fitted value has uniform spread and the red lines is deviates a little at ends but overall its more or less is perfect straight across plot .Hence we can declare that the residual follow linear pattern.
- We can see that in QQplot the points fall roughly along the straight line , Hence points are normally distributed

- Residual vs Fitted value has uniform spread and the red lines is almost straight across plot .Hence we can declare that the assumption of equal variance in not violated hence the model ids homoscedastic.
- In Residual vs Leverage we can see that observation 196 i s on cooks line but doesnot cross it .This means there is no potential influencers.

Q13. Check autocorrelation and heteroscedasticity using appropriate statistical test.

- **Test for Autocorrelation with the ACF Plot**

```
residuals_mod1 =model_1$residuals
acf(residuals_mod1 , type ="correlation")
```



After the lag-0 correlation, the subsequent correlations drop quickly to zero and stay (mostly) between the limits of the significance level (dashed blue lines). Therefore, we can conclude that the residuals of this model meet the assumption of no autocorrelation.

- **Durbin-Watson Test to Check Autocorrelation**
  - H0 : First order autocorrelation do not exist
  - H1 : First order autocorrelation exist

```
#Durbin-Watson Test to Check Autocorrelation

library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
lmtest::dwtest(model_1)
```

```
##
##  Durbin-Watson test
##
## data:  model_1
## DW = 1.9174, p-value = 0.1773
## alternative hypothesis: true autocorrelation is greater than 0
```

- *As the DW value is 1.91 which lies in the range of 1.5 to 2.5 hence there is no autocorrelation between the residuals*

Testing Heteroskedacity

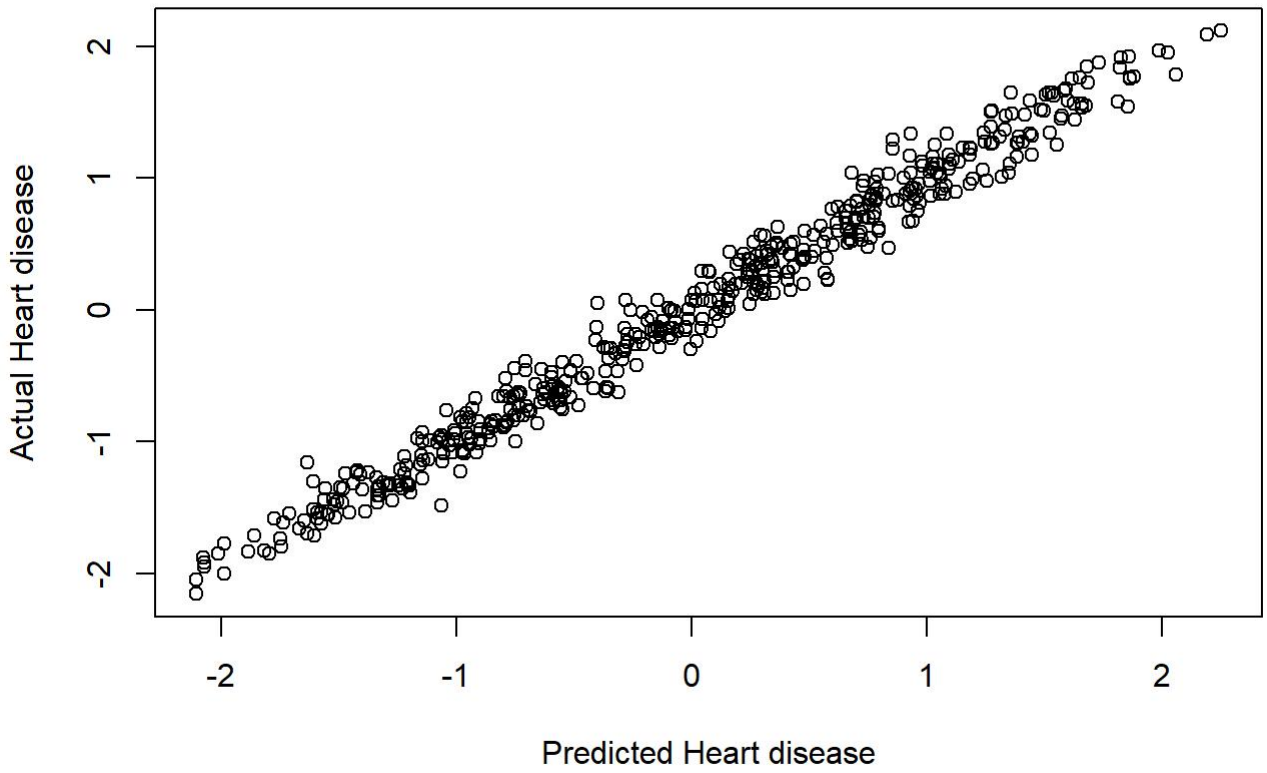
- **Perform the Breusch–Pagan Test to Check Heteroscedasticity**
  - H0 : Residuals are distributed with equal variance(i.e homoscedastic)
  - H1 : Residuals are distributes with non equal variance(i.e heteroscedastic)

```
lmtest::bgtest(model_1)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  model_1
## LM test = 0.63776, df = 1, p-value = 0.4245
```

- *p value is greater than .05 hence we fail to reject the null confirming that model is homoscedastic*

```
plot(cardiodata$heart_diseases ,heart_fitted , xlab= "Predicted Heart disease" , ylab="Actual
Heart disease")
```




```
## unscaling data for accuracy check and rmse of model
targetmean = mean(cardiodata_1$heart_diseases)
targetsd = sd(cardiodata_1$heart_diseases)
unscaledtest.obs = round(cardiodata$heart_diseases *targetsd + targetmean , 0)
unscaledtest.pred = round (heart_fitted *targetsd + targetmean , 0)
#####


Accuracy=mean(unscaledtest.obs ==unscaledtest.pred )
Accuracy
```

```
## [1] 0.4959839
```

```
rmse= sqrt(mean((unscaledtest.obs-unscaledtest.pred)^2))
rmse
```

```
## [1] 0.8127992
```