Regression Analysis

Harshal Shivaji Holam
2102451

TITLE: Heart Disease Analysis

DOMAIN:

The chosen domain is Health (Heary disease).

OBJECTIVE:

To predict the risk of heart disease

METHOD/TECHNIQUE

The method used is Multiple Linear regression Model. Linear regression follows the linear mathematical model for determining the value of one dependent variable from value of one/more given independent variable(s).

Formula and Calculation of Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

DATASET

The dataset (https://cdn.scribbr.com/wp-content/uploads//2020/02/heart.data_.zip) contains observations on the percentage of people biking to work each day, the percentage of people smoking, and the percentage of people with heart disease in an sample of 500 towns.In this data, the percentage of people who smoke and the percentage of people who ride a bicycle are independent of each other. The percentage of people who smoke is dependent on both percentages of people who smoke and the percentage of people who do bicycling.

Analysis through multiple linear regression

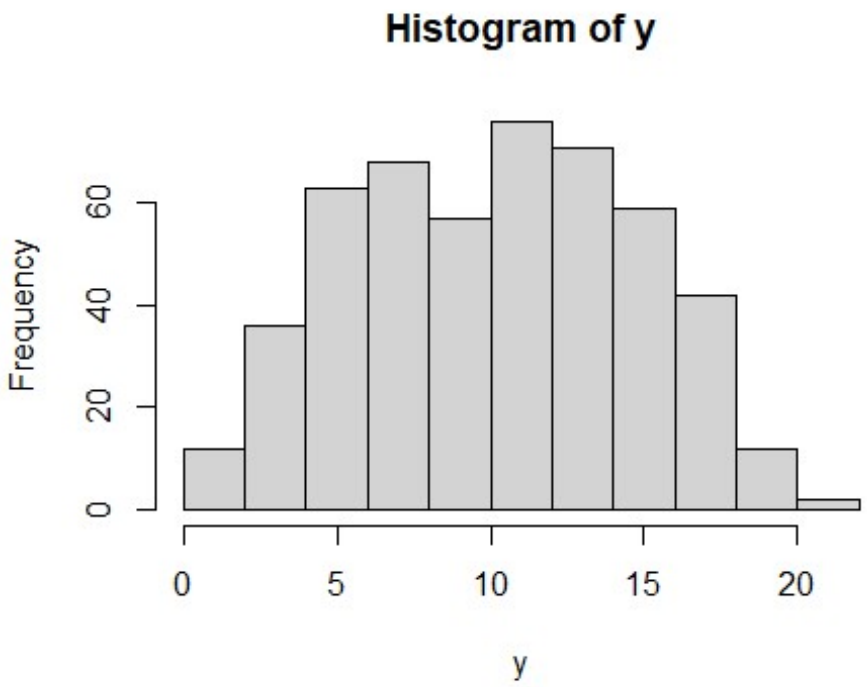
```
library(readxl)

## Warning: package 'readxl' was built under R version 4.1.2

heart <- read_excel("C:/Users/LENOVO/Downloads/regression project/heart.xlsx")

## New names:
## * `` -> ...1

View(heart)

#assign variable
```

```
y=heart$heart.disease
x1=heart$biking
x2=heart$smoking

#regression model
Reg = lm(y~x1+x2)
Reg

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##     14.9847      -0.2001       0.1783

#regression equation is given by
#y=14.9847 - 0.2001*x1 + 0.1783*x2
```

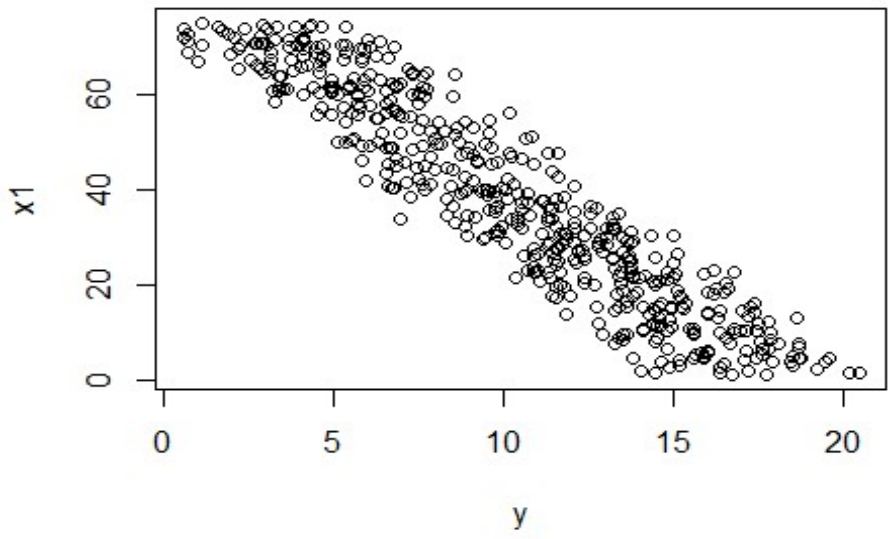
- Assumptions: The assumptions for multiple regression analysis were checked and all the assumptions were satisfied.
  1.Residuals vs fitted graph is used to check the linear relationship assumptions.
  2.Normal Q-Q plot is used to determine whether the residuals are normally distributed.

  3.Scale-location(or spread-location) is homogeneity of variance(homoscedasticity) determination
  4.Residuals vs Leverage is for identifying influential points, which are points that might impact regression results when included or excluded from the analysis

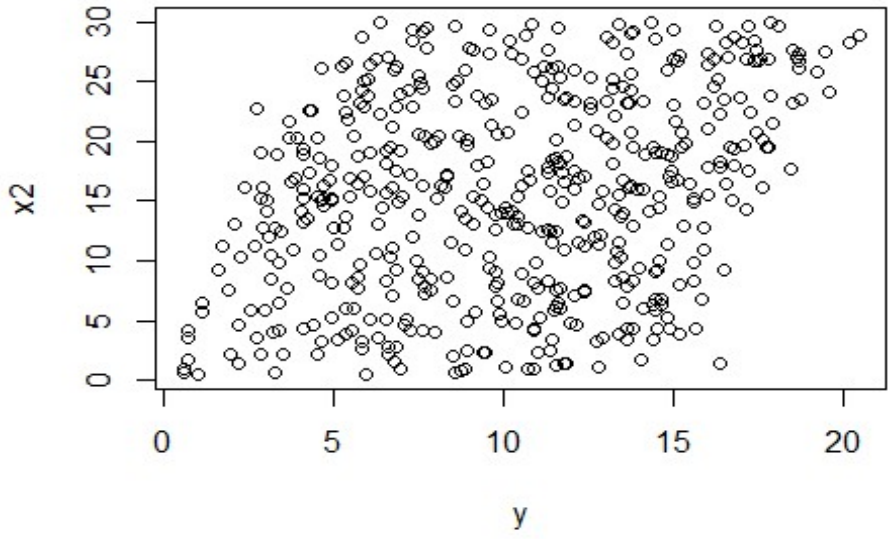
```
#check the assumption
#1.independece of the observation  (no autocorrelation)
cor(x1,x2)

## [1] 0.01513618

#The correlation between biking and smoking is small (0.015 is only a 1.5% correlation),
so we can include both parameters in our model.

#2.Normality of y
hist(y)
```



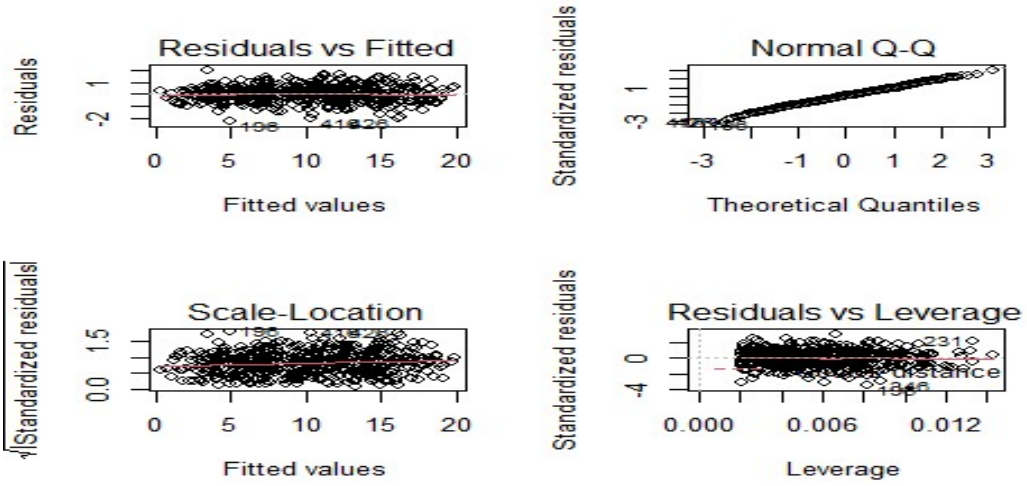

```
#here dependent variable is normally distributed

#3.Linearity
#first check for y and x1
plot(y,x1)
```




```
#here y and x1 are negatively correlated hence there is linear relationship between x1 an
d y
#noe for y and x2
plot(y,x2)
```




```
#the relationship between smoking and heart disease is a bit less clear, it still appears
linear.

#4.error follow normal distribution
par(mfrow=c(2,2))
plot(Reg)
```




```
#from first plot we can conclude that model is adequate
#fron second plot we can conclude that data is normally distributed.


#anova table and test the significance of the regression
a=anova(Reg)
a

## Analysis of Variance Table
##
```

```
## Response: y
##             Df Sum Sq Mean Sq F value    Pr(>F)
## x1           1 9090.6  9090.6 21251.7 < 2.2e-16 ***
## x2           1 1086.0  1086.0  2538.8 < 2.2e-16 ***
## Residuals 495  211.7     0.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test for significance of regression.

Hypothesis:

For $\beta_{i,}$

H0: $\beta_i = 0$ V/s

H1: $\beta_i \neq 0$ for at least one i where i=1, 2, 3… k

Where ,k=No. of regressors in the model

By P-value criteria,

Reject Ho, when p value is less than LOS (LOS=0.05)

Hence,

P value is= 2.2e-16 < 0.05

Hence we reject Ho

i.e . there is at least one regressor variable is significant in the model.

```
#Summary of the model
s=summary(Reg)
s

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1789 -0.4463  0.0362  0.4422  1.9331
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.984658   0.080137  186.99   <2e-16 ***
## x1          -0.200133   0.001366 -146.53   <2e-16 ***
## x2           0.178334   0.003539   50.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.654 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
```

for testing the hypothesis $H_0: \beta_0=0$, $H_0: \beta_1=0$, $H_0: \beta_2=0$.

Hypothesis:

For β_0,

H0: $\beta_0 = 0$ V/s H1: $\beta_0 \neq 0$

Test Procedure

We reject H0, if $|$ **t Cal** $| >$ **t** $_{n-2,\alpha/2}$ o.w. we accept it

Here $|$ t Cal $| = 186.99 > t_{n-2,\alpha/2} = 1.64792$

hence we rejct H0.

P-value criteria

If p-value < α ,

we reject H0,

Here, p-value = 2e-16 < 0.05,

Hence we reject H0.

X0 contributes significantly to the model

Hypothesis:

For β_1,

H0: $\beta_1 = 0$ V/s H1: $\beta_1 \neq 0$

Test Procedure

We reject H0, if $|$ t Cal $| > t_{n-2,\alpha/2}$ o.w. we accept it

Here $|$ t Cal $|$

$= 146.53 > t_{n-2,\alpha/2} = 1.64792$

hence we rejct H0.

P-value criteria

If p-value < α ,

we reject H0

Here, p-value = 2e-16 < 0.05,

Hence we reject H0.

X1 contributes significantly to the model

For β_2,

H0: $\beta_2 = 0$ V/s H1: $\beta_2 \neq 0$

Test Procedure

We reject H0, if $|$ t Cal $| > t_{n-2,\alpha/2}$ o.w. we accept it

Here $|$ t Cal $|$

$= 50.39 > t_{n-2,\alpha/2} = 1.64792$

hence we rejct H0.

P-value criteria

If p-value < α ,

we reject H0

Here, p-value = 2e-16 < 0.05,

Hence we reject H0.

X2 contributes significantly to the model

```
#Multiple r sq and adjusted R-sq
M_R_2 = s$r.squared
M_R_2

## [1] 0.9796175
```

```
#Interpretation: Coefficient of determination of 97.96% shows that 97.96% of the variatio
n in Y is explained by regressors X1, X2.
#here we have two regressor x1 and x2
adj_R2 = s$adj.r.squared
adj_R2

## [1] 0.9795351

sigma_2 = a$`Mean Sq`[3] #sigma_2
sigma_2

## [1] 0.4277581

confint(Reg)

##                   2.5 %     97.5 %
## (Intercept) 14.8272075 15.1421084
## x1          -0.2028166 -0.1974495
## x2           0.1713800  0.1852878
```

#for β_1 the confidence interval is [**-0.2028166 , -0.1974495**]

>#hence CI does not include 0 which indicate that regressor x1 is significant.

#for B2 the confidence interval is **0.1713800 , 0.1852878**]

```
>#hence CI does not include 0 which indicate that regressor x2 is significant
```

#checking the multicollinearity

```
library(car)

## Warning: package 'car' was built under R version 4.1.2

## Loading required package: carData

vif(Reg)

##       x1       x2
## 1.000229 1.000229

#here we can conclude that Vif value are less than 5 or 10 hence multicollinearity is abs
ent
```

Conclusion

Thus, from the data and after applying Multiple linear regression it is clear that the risk of heart disease increases with smoking and decreases with any form of physical exercise.

Final model is given by

```
Heart disease =14.9847 - 0.2001*biking + 0.1783*smoking
```