Regression.R

MRUNALINI K.
mrunalini0107@gmail.com

Linear Regression

```
Heart_data <- read.csv('HeartDiseaseTrainTest.csv')
head(Heart_data)
```

```
##   age    sex chest_pain_type resting_blood_pressure cholestoral
## 1  52   Male  Typical angina                    125         212
## 2  53   Male  Typical angina                    140         203
## 3  70   Male  Typical angina                    145         174
## 4  61   Male  Typical angina                    148         203
## 5  62 Female  Typical angina                    138         294
## 6  58 Female  Typical angina                    100         248
##      fasting_blood_sugar              rest_ecg Max_heart_rate
## 1   Lower than 120 mg/ml ST-T wave abnormality            168
## 2 Greater than 120 mg/ml                Normal            155
## 3   Lower than 120 mg/ml ST-T wave abnormality            125
## 4   Lower than 120 mg/ml ST-T wave abnormality            161
## 5 Greater than 120 mg/ml ST-T wave abnormality            106
## 6   Lower than 120 mg/ml                Normal            122
##   exercise_induced_angina oldpeak       slope vessels_colored_by_flourosopy
## 1                      No     1.0 Downsloping                           Two
## 2                     Yes     3.1   Upsloping                          Zero
## 3                     Yes     2.6   Upsloping                          Zero
## 4                      No     0.0 Downsloping                           One
## 5                      No     1.9        Flat                         Three
## 6                      No     1.0        Flat                          Zero
##         thalassemia target
## 1 Reversable Defect      0
## 2 Reversable Defect      0
## 3 Reversable Defect      0
## 4 Reversable Defect      0
## 5      Fixed Defect      0
## 6      Fixed Defect      1
```

```
colnames(Heart_data)
```

```
##  [1] "age"                          "sex"
##  [3] "chest_pain_type"              "resting_blood_pressure"
##  [5] "cholestoral"                  "fasting_blood_sugar"
##  [7] "rest_ecg"                     "Max_heart_rate"
##  [9] "exercise_induced_angina"      "oldpeak"
## [11] "slope"                        "vessels_colored_by_flourosopy"
## [13] "thalassemia"                  "target"
```

```
dim(Heart_data)
```

```
## [1] 1025   14
```

```r
str(Heart_data)
```

```
## 'data.frame':    1025 obs. of  14 variables:
##  $ age                          : int  52 53 70 61 62 58 58 55 46 54 ...
##  $ sex                          : chr  "Male" "Male" "Male" "Male" ...
##  $ chest_pain_type              : chr  "Typical angina" "Typical angina" "Typical angi
na" "Typical angina" ...
##  $ resting_blood_pressure       : int  125 140 145 148 138 100 114 160 120 122 ...
##  $ cholestoral                  : int  212 203 174 203 294 248 318 289 249 286 ...
##  $ fasting_blood_sugar          : chr  "Lower than 120 mg/ml" "Greater than 120 mg/ml"
"Lower than 120 mg/ml" "Lower than 120 mg/ml" ...
##  $ rest_ecg                     : chr  "ST-T wave abnormality" "Normal" "ST-T wave abn
ormality" "ST-T wave abnormality" ...
##  $ Max_heart_rate               : int  168 155 125 161 106 122 140 145 144 116 ...
##  $ exercise_induced_angina      : chr  "No" "Yes" "Yes" "No" ...
##  $ oldpeak                      : num  1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...
##  $ slope                        : chr  "Downsloping" "Upsloping" "Upsloping" "Downslop
ing" ...
##  $ vessels_colored_by_flourosopy: chr  "Two" "Zero" "Zero" "One" ...
##  $ thalassemia                  : chr  "Reversable Defect" "Reversable Defect" "Revers
able Defect" "Reversable Defect" ...
##  $ target                       : int  0 0 0 0 0 1 0 0 0 0 ...
```

```r
summary(Heart_data)
```

```
##       age            sex            chest_pain_type    resting_blood_pressure
##  Min.   :29.00   Length:1025        Length:1025        Min.   : 94.0
##  1st Qu.:48.00   Class :character   Class :character   1st Qu.:120.0
##  Median :56.00   Mode  :character   Mode  :character   Median :130.0
##  Mean   :54.43                                         Mean   :131.6
##  3rd Qu.:61.00                                         3rd Qu.:140.0
##  Max.   :77.00                                         Max.   :200.0
##   cholestoral  fasting_blood_sugar   rest_ecg         Max_heart_rate
##  Min.   :126   Length:1025         Length:1025        Min.   : 71.0
##  1st Qu.:211   Class :character    Class :character   1st Qu.:132.0
##  Median :240   Mode  :character    Mode  :character   Median :152.0
##  Mean   :246                                          Mean   :149.1
##  3rd Qu.:275                                          3rd Qu.:166.0
##  Max.   :564                                          Max.   :202.0
##  exercise_induced_angina    oldpeak         slope
##  Length:1025             Min.   :0.000   Length:1025
##  Class :character        1st Qu.:0.000   Class :character
##  Mode  :character        Median :0.800   Mode  :character
##                          Mean   :1.072
##                          3rd Qu.:1.800
##                          Max.   :6.200
##  vessels_colored_by_flourosopy thalassemia            target
##  Length:1025                   Length:1025        Min.   :0.0000
##  Class :character              Class :character   1st Qu.:0.0000
##  Mode  :character              Mode  :character   Median :1.0000
##                                                   Mean   :0.5132
##                                                   3rd Qu.:1.0000
##                                                   Max.   :1.0000
```

```r
# Using the lm() function to fit a simple linear regression model, with Max_heart_rate as
the response and age as the predictor.
```

```r
lm.fit <- lm(Max_heart_rate ~ age, data = Heart_data)
attach(Heart_data)
lm.fit <- lm(Max_heart_rate ~ age)
lm.fit
```

```
##
## Call:
## lm(formula = Max_heart_rate ~ age)
##
## Coefficients:
## (Intercept)          age
##    202.9793      -0.9895
```

```r
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Max_heart_rate ~ age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.680 -11.680   4.373  16.394  45.456
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 202.9793     4.0283   50.39   <2e-16 ***
## age          -0.9896     0.0730  -13.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.19 on 1023 degrees of freedom
## Multiple R-squared:  0.1523, Adjusted R-squared:  0.1514
## F-statistic: 183.8 on 1 and 1023 DF,  p-value: < 2.2e-16
```

```r
coef(lm.fit)
```

```
## (Intercept)         age
## 202.9792848  -0.9895469
```

```r
# To obtain a confidence interval for the coefficient estimates, we can use the confint()
command.
confint(lm.fit)
```

```
##                 2.5 %      97.5 %
## (Intercept) 195.074582 210.8839872
## age          -1.132789  -0.8463049
```

```r
# The predict() function can be used to produce confidence intervals and prediction inter
vals for the prediction of Max_heart_rate for a given value of age.
print("Confidence interval")
```

```
## [1] "Confidence interval"
```

```r
predict(lm.fit, data.frame(age = (c(5, 10, 15))), interval= "confidence")
```

```
##        fit      lwr      upr
## 1 198.0316 190.8324 205.2307
```

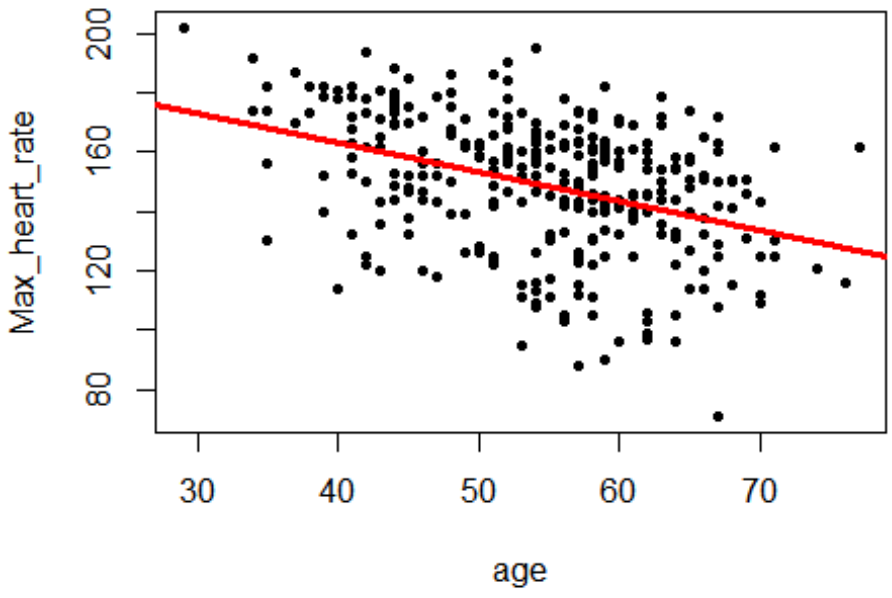
```
## 2 193.0838 186.5878 199.5798
## 3 188.1361 182.3400 193.9321
```

```
predict(lm.fit, data.frame(age = (c(5, 10, 15))), interval= "prediction")
```

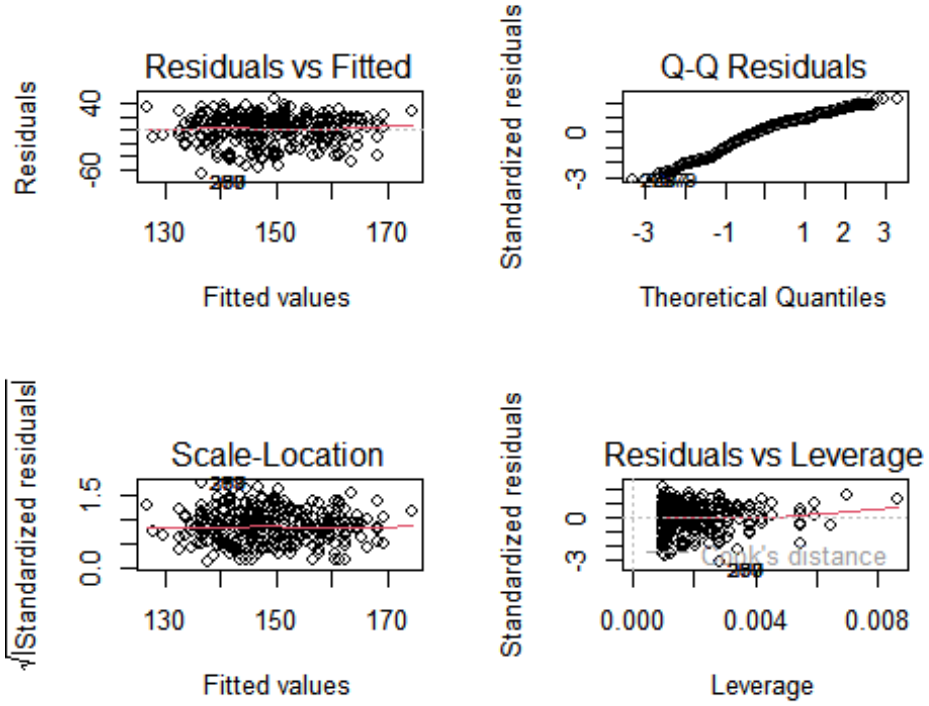
```
##        fit      lwr      upr
## 1 198.0316 155.8279 240.2352
## 2 193.0838 150.9945 235.1732
## 3 188.1361 146.1491 230.1231
```

```
# Note: The linear regression analysis reveals that age is negatively associated with Max
_heart_rate, as indicated by the coefficients of the fitted model. The intercept is estim
ated at 202.98, implying the expected Max_heart_rate at age zero, and the coefficient for
age is estimated at -0.99, suggesting a decrease in Max_heart_rate as age increases. The
95% confidence intervals for the coefficients provide a level of certainty about their tr
ue values. Predicted Max_heart_rate values at specific ages (5, 10, and 15) come with nar
rower confidence intervals, reflecting the precision of these predictions. However, the w
ider prediction intervals acknowledge the inherent variability and uncertainty in predict
ing individual observations. Overall, the analysis provides valuable insights into the ag
e-related trends in Max_heart_rate, offering both point estimates and intervals that capt
ure the uncertainty associated with the model and predictions.
```

```
# Plot Max_heart_rate and age along with the least squares regression line using the plot
() and abline() functions.
plot(age, Max_heart_rate , col = "black", pch = 20)
abline(lm.fit, lwd = 3, col = "red")
```




```
### Check For Homoscedasticity
par(mfrow = c(2,2))
plot(lm.fit)
```




```
# To plot the residuals against the fitted values.
plot(predict(lm.fit), residuals(lm.fit))
plot(predict(lm.fit), rstudent(lm.fit))

# On the basis of the residual plots, there is some evidence of non-linearity. Leverage s
tatistics can be computed for any number of predictors using the hatvalues() function
plot(hatvalues(lm.fit))

# The which.max() function identifies the index of the largest element of a vector
which.max(hatvalues(lm.fit))
```

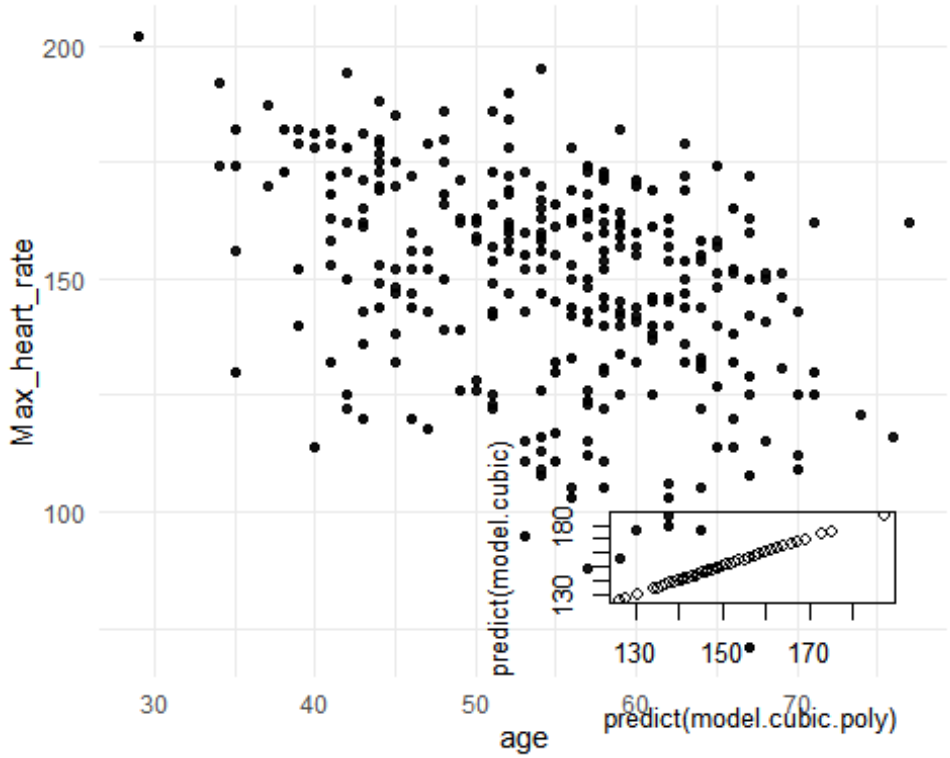
```
## 61
## 61
```

```
###

# Comment:
# The linear regression analysis suggests that there is a relationship between age and Ma
x_heart_rate. The regression model provides coefficient estimates along with confidence a
nd prediction intervals for Max_heart_rate based on age. The visualization of the regress
ion line on the scatter plot helps to understand the trend in the data. The residual plot
s indicate some evidence of non-linearity and potential outliers, which may need further
investigation. The leverage statistics highlight influential observations in the dataset.
Overall, this analysis provides valuable insights into the association between age and Ma
x_heart_rate in the context of heart disease data. Further refinement of the model or exp
loration of additional variables may be necessary for a more comprehensive understanding
of the relationship.
```

```r
# Polynomial Regression
#install.packages("MultiKink")
#install.packages("ggplot2")

library(MultiKink)

## Warning: package 'MultiKink' was built under R version 4.3.2

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.2

set.seed(1974)

# Using the poly() to predict Max_heart_rate with a forth-degree polynomial
heart.age.plot <- ggplot(Heart_data, aes(x = age, y = Max_heart_rate)) +
  geom_point(alpha=0.55, color="black") +
  theme_minimal()
heart.age.plot

model.cubic <- lm( Max_heart_rate ~ age + I(age^2) + I(age^3), Heart_data)
summary(model.cubic)

##
## Call:
## lm(formula = Max_heart_rate ~ age + I(age^2) + I(age^3), data = Heart_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.983 -12.098   3.776  16.162  46.472
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.402e+02  7.832e+01   4.344 1.54e-05 ***
## age         -8.410e+00  4.534e+00  -1.855   0.0639 .
## I(age^2)     1.296e-01  8.589e-02   1.509   0.1315
## I(age^3)    -7.337e-04  5.334e-04  -1.376   0.1692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.16 on 1021 degrees of freedom
## Multiple R-squared:  0.1564, Adjusted R-squared:  0.1539
## F-statistic: 63.08 on 3 and 1021 DF,  p-value: < 2.2e-16

model.cubic.poly <- lm( Max_heart_rate ~ poly(age,3), data = Heart_data)
summary(model.cubic.poly)

##
## Call:
## lm(formula = Max_heart_rate ~ poly(age, 3), data = Heart_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.983 -12.098   3.776  16.162  46.472
##
```

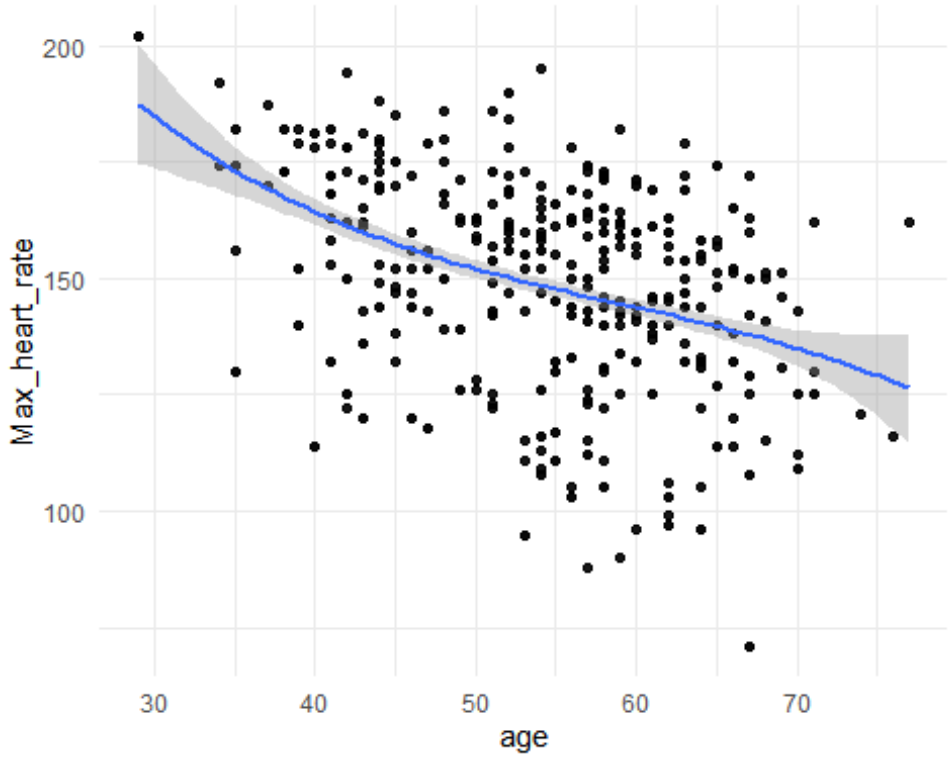
```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     149.114      0.661 225.597   <2e-16 ***
## poly(age, 3)1 -287.279     21.162 -13.575   <2e-16 ***
## poly(age, 3)2   37.056     21.162   1.751   0.0802 .
## poly(age, 3)3  -29.111     21.162  -1.376   0.1692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.16 on 1021 degrees of freedom
## Multiple R-squared:  0.1564, Adjusted R-squared:  0.1539
## F-statistic: 63.08 on 3 and 1021 DF,  p-value: < 2.2e-16
```

```
plot(predict(model.cubic.poly), predict(model.cubic))
```

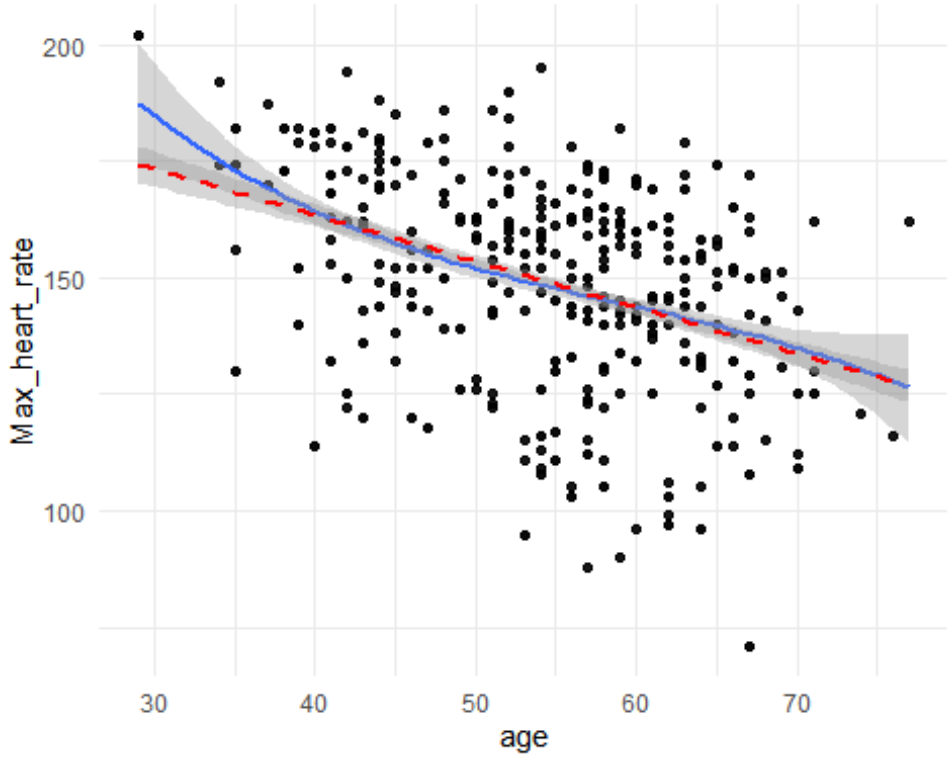



```
heart.age.plot + stat_smooth(method = "lm", formula = y ~ poly(x, 3, raw=T), size = 1)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
# Add a linear fit to the plot above
heart.age.plot + stat_smooth(method = "lm",  formula = y ~ poly(x, 3, raw=T), size = 1) +
  stat_smooth(method = "lm",  formula = y~poly(x,1,raw=T), lty = 2, col = "red" , size =
1)
```




```
###

# Comment:
# In this R script, polynomial regression is employed to model the relationship between a
```

ge and maximum heart rate using the Heart_data dataset. Two models are fitted: a cubic regression (model.cubic) and a polynomial regression of degree 3 (model.cubic.poly). Visualizations are created to showcase the scatter plot of the data and overlay cubic and linear fits. The cubic polynomial appears to capture non-linear trends in the data, suggesting its potential suitability for predicting maximum heart rate based on age. However, careful consideration should be given to the balance between model complexity and interpretability, as more complex models may be prone to overfitting. Further validation, such as cross-validation, is recommended to assess the models' performance on unseen data. Overall, the analysis provides insights into the complex relationship between age and maximum heart rate, highlighting the importance of considering non-linear terms in the regression modeling process.

Multiple Regression

```
str(Heart_data)
```

```
## 'data.frame':    1025 obs. of  14 variables:
##  $ age                       : int  52 53 70 61 62 58 58 55 46 54 ...
##  $ sex                       : chr  "Male" "Male" "Male" "Male" ...
##  $ chest_pain_type           : chr  "Typical angina" "Typical angina" "Typical angi
na" "Typical angina" ...
##  $ resting_blood_pressure    : int  125 140 145 148 138 100 114 160 120 122 ...
##  $ cholestoral               : int  212 203 174 203 294 248 318 289 249 286 ...
##  $ fasting_blood_sugar       : chr  "Lower than 120 mg/ml" "Greater than 120 mg/ml"
"Lower than 120 mg/ml" "Lower than 120 mg/ml" ...
##  $ rest_ecg                  : chr  "ST-T wave abnormality" "Normal" "ST-T wave abn
ormality" "ST-T wave abnormality" ...
##  $ Max_heart_rate            : int  168 155 125 161 106 122 140 145 144 116 ...
##  $ exercise_induced_angina   : chr  "No" "Yes" "Yes" "No" ...
##  $ oldpeak                   : num  1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...
##  $ slope                     : chr  "Downsloping" "Upsloping" "Upsloping" "Downslop
ing" ...
##  $ vessels_colored_by_flourosopy: chr  "Two" "Zero" "Zero" "One" ...
##  $ thalassemia               : chr  "Reversable Defect" "Reversable Defect" "Revers
able Defect" "Reversable Defect" ...
##  $ target                    : int  0 0 0 0 0 1 0 0 0 0 ...
```

```
# Using the lm() function to fit a Multiple Regression model, with Max_heart_rate as the
response and age, cholesterol, resting_blood_pressure as the predictor
lm.fit <- lm(Max_heart_rate ~ age + cholestoral + resting_blood_pressure, data = Heart_da
ta)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Max_heart_rate ~ age + cholestoral + resting_blood_pressure,
##     data = Heart_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.386 -12.349   4.958  15.391  41.152
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           188.92376    6.04532  31.251   <2e-16 ***
## age                    -1.07055    0.07699 -13.904   <2e-16 ***
## cholestoral             0.02784    0.01314   2.119   0.0344 *
## resting_blood_pressure  0.08827    0.03922   2.250   0.0246 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.11 on 1021 degrees of freedom
## Multiple R-squared:  0.1607, Adjusted R-squared:  0.1583
## F-statistic: 65.18 on 3 and 1021 DF,  p-value: < 2.2e-16

dim(Heart_data)

## [1] 1025   14

lm.fit <- lm(Max_heart_rate ~ ., data = Heart_data)
summary(lm.fit)

##
## Call:
## lm(formula = Max_heart_rate ~ ., data = Heart_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.542 -10.394   1.353  12.598  40.709
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          177.98489    9.61364  18.514  < 2e-16
## age                                   -0.79446    0.07036 -11.291  < 2e-16
## sexMale                                1.00727    1.40642   0.716  0.47404
## chest_pain_typeAtypical angina        -2.75198    2.52216  -1.091  0.27548
## chest_pain_typeNon-anginal pain       -4.53010    2.33607  -1.939  0.05276
## chest_pain_typeTypical angina         -9.61497    2.32566  -4.134 3.86e-05
## resting_blood_pressure                 0.10595    0.03447   3.074  0.00217
## cholestoral                            0.03635    0.01136   3.201  0.00141
## fasting_blood_sugarLower than 120 mg/ml -1.01860   1.62654  -0.626  0.53130
## rest_ecgNormal                         4.58068    4.74030   0.966  0.33411
## rest_ecgST-T wave abnormality          3.41834    4.77985   0.715  0.47468
## exercise_induced_anginaYes            -7.85625    1.40151  -5.606 2.68e-08
## oldpeak                               -1.41964    0.64582  -2.198  0.02816
## slopeFlat                            -11.47690    1.37291  -8.360  < 2e-16
## slopeUpsloping                        -6.39916    2.64043  -2.424  0.01555
## vessels_colored_by_flourosopyOne      -1.17053    4.47034  -0.262  0.79350
## vessels_colored_by_flourosopyThree    -3.93404    4.90165  -0.803  0.42240
## vessels_colored_by_flourosopyTwo       8.24134    4.70887   1.750  0.08039
## vessels_colored_by_flourosopyZero      1.84570    4.27258   0.432  0.66584
## thalassemiaNo                         -7.94837    6.74146  -1.179  0.23867
## thalassemiaNormal                     -5.57686    2.51914  -2.214  0.02707
## thalassemiaReversable Defect           0.49294    1.42809   0.345  0.73004
## target                                 5.30758    1.67321   3.172  0.00156
##
## (Intercept)                          ***
## age                                  ***
## sexMale
## chest_pain_typeAtypical angina
```

```
## chest_pain_typeNon-anginal pain         .
## chest_pain_typeTypical angina           ***
## resting_blood_pressure                  **
## cholestoral                             **
## fasting_blood_sugarLower than 120 mg/ml
## rest_ecgNormal
## rest_ecgST-T wave abnormality
## exercise_induced_anginaYes              ***
## oldpeak                                 *
## slopeFlat                               ***
## slopeUpsloping                          *
## vessels_colored_by_flourosopyOne
## vessels_colored_by_flourosopyThree
## vessels_colored_by_flourosopyTwo        .
## vessels_colored_by_flourosopyZero
## thalassemiaNo
## thalassemiaNormal                       *
## thalassemiaReversable Defect
## target                                  **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.39 on 1002 degrees of freedom
## Multiple R-squared:  0.4411, Adjusted R-squared:  0.4288
## F-statistic: 35.95 on 22 and 1002 DF,  p-value: < 2.2e-16
```

```
summary(lm(Max_heart_rate ~ age * cholestoral, data = Heart_data))
```

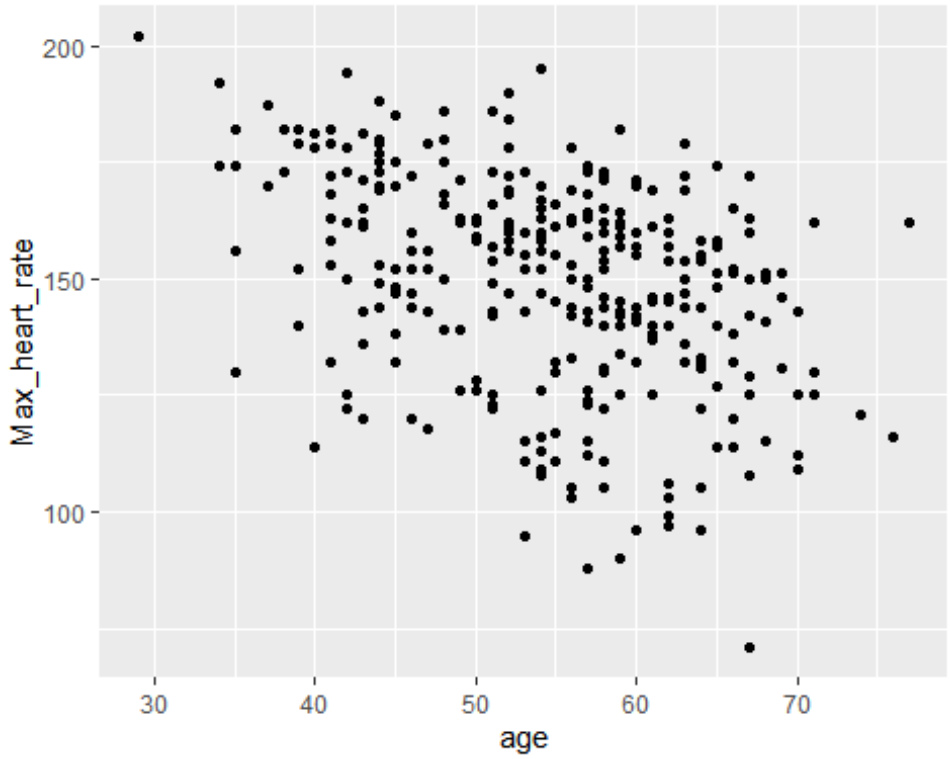
```
##
## Call:
## lm(formula = Max_heart_rate ~ age * cholestoral, data = Heart_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.681 -12.481   4.354  16.111  45.202
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     252.235835  19.340374  13.042  < 2e-16 ***
## age              -1.993092   0.340499  -5.853 6.48e-09 ***
## cholestoral      -0.203130   0.081242  -2.500  0.01256 *
## age:cholestoral   0.004101   0.001411   2.907  0.00372 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.07 on 1021 degrees of freedom
## Multiple R-squared:  0.1635, Adjusted R-squared:  0.1611
## F-statistic: 66.52 on 3 and 1021 DF,  p-value: < 2.2e-16
```

```
cor(Heart_data$age, Heart_data$cholestoral)
```

```
## [1] 0.2198225
```

```
plotting.data <- expand.grid(age = seq(min(Heart_data$age),
                max(Heart_data$age), length.out=30),
                  cholestoral=c(min(Heart_data$cholestoral), mean(Heart_data$cholestoral)
,
```

```
                    max(Heart_data$cholestoral),
                    resting_blood_pressure = c(min(Heart_data$resting_blood_pressure),
                      mean(Heart_data$resting_blood_pressure),
                        max(Heart_data$resting_blood_pressure))))

View(plotting.data)

heart.plot <- ggplot(Heart_data, aes(x = age, y=Max_heart_rate)) + geom_point()
heart.plot
```




```
###
```

Comment:

```
# The provided R code conducts a thorough analysis of the "Heart_data" dataset, employing
multiple regression models to investigate the relationship between predictor variables wh
ich are age, cholesterol, and resting_blood_pressure and the response variable, Max_heart
_rate. The code includes fitting models with and without interaction terms, assessing cor
relations between age and cholesterol, and preparing data for visualization. While the co
de generates a scatter plot to visualize the relationship between age and Max_heart_rate,
a specific conclusion cannot be drawn without the actual results and interpretations of t
he regression analyses. A comprehensive conclusion would involve examining coefficients,
significance levels, and model fit statistics to make informed statements about the impac
t of the predictor variables on the response variable in the context of the dataset.
```

Natural Cubic Spline

```
#install.packages("gam")

library(splines)
library(MultiKink) #for the data
library(ggplot2)   #for the plots
```

```r
set.seed(1974)     #fix the random generator seed

attach(Heart_data)
```

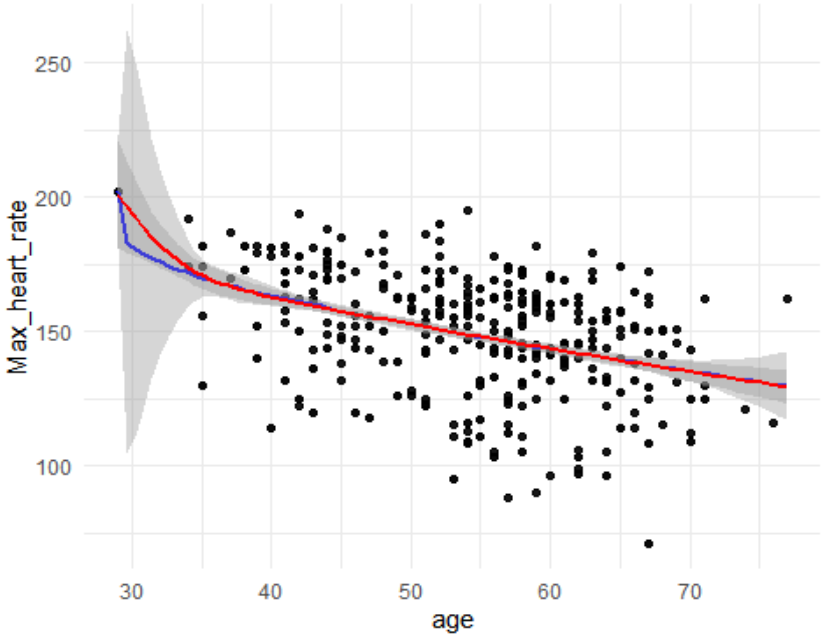
```
## The following objects are masked from Heart_data (pos = 6):
##
##     age, chest_pain_type, cholestoral, exercise_induced_angina,
##     fasting_blood_sugar, Max_heart_rate, oldpeak, rest_ecg,
##     resting_blood_pressure, sex, slope, target, thalassemia,
##     vessels_colored_by_flourosopy
```

```r
#linear model with the natural cubic splines function
cub.splines.bs <- lm(Max_heart_rate ~ bs(age, knots = c(5,10,20,30,40)), data=Heart_data)
summary(cub.splines.bs)
```

```
##
## Call:
## lm(formula = Max_heart_rate ~ bs(age, knots = c(5, 10, 20, 30,
##     40)), data = Heart_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -66.66 -12.53   3.84  16.12  46.11
##
## Coefficients: (3 not defined because of singularities)
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            129.596      6.454  20.080  < 2e-16 ***
## bs(age, knots = c(5, 10, 20, 30, 40))1       NA         NA      NA       NA
## bs(age, knots = c(5, 10, 20, 30, 40))2       NA         NA      NA       NA
## bs(age, knots = c(5, 10, 20, 30, 40))3   72.404     12.389   5.844 6.85e-09 ***
## bs(age, knots = c(5, 10, 20, 30, 40))4   53.428     45.894   1.164 0.244634
## bs(age, knots = c(5, 10, 20, 30, 40))5   42.253     11.007   3.839 0.000131 ***
## bs(age, knots = c(5, 10, 20, 30, 40))6   22.043      6.580   3.350 0.000838 ***
## bs(age, knots = c(5, 10, 20, 30, 40))7    9.913     12.841   0.772 0.440303
## bs(age, knots = c(5, 10, 20, 30, 40))8       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.15 on 1019 degrees of freedom
## Multiple R-squared:  0.1589, Adjusted R-squared:  0.1548
## F-statistic: 38.51 on 5 and 1019 DF,  p-value: < 2.2e-16
```

```r
#simple scatter
Heart.age.plot <- ggplot(Heart_data, aes(x = age, y = Max_heart_rate)) + geom_point(alpha
=0.55, color="black") +  theme_minimal()

Heart.age.plot + stat_smooth(method = "lm", formula = y~bs(x,knots = c(5,10,20,30,40)), l
ty = 1, col = "blue") +
  stat_smooth(method = "lm", formula = y~ns(x,knots = c(5,10,20,30,40)), lty = 1, col = "
red")
```

```
## Warning in predict.lm(model, newdata = data_frame0(x = xseq), se.fit = se, :
## prediction from rank-deficient fit; attr(*, "non-estim") has doubtful cases
```



Comment:

In this analysis, a linear model incorporating natural cubic splines was applied to the relationship between age and maximum heart rate using the 'gam' package in R. The knots were strategically placed at ages 5, 10, 20, 30, and 40 to capture potential non-linearities in the data. The resulting model, as summarized, provides insights into the intricate relationship between age and maximum heart rate. The scatter plot illustrates the raw data, while the blue and red curves represent the fitted models using natural cubic splines and alternative smoothing methods, respectively. The use of natural cubic splines allows for flexibility in capturing the underlying patterns in the data, revealing potential non-linear trends. Further interpretation of the model coefficients and examination of the plotted curves can guide a comprehensive understanding of how age influences maximum heart rate in the dataset.

Conclusion:

In this R script, a comprehensive analysis of the "Heart_data" dataset is conducted, focusing on regression modeling techniques. The script begins with a linear regression analysis, fitting a model to predict maximum heart rate (Max_heart_rate) based on age. The interpretation of coefficients, confidence intervals, and prediction intervals is demonstrated. Residual plots are examined for potential non-linearity, and leverage statistics are computed to identify influential observations. Subsequently, polynomial regression is applied to capture non-linear trends in the relationship between age and Max_heart_rate, showcasing cubic and linear fits. The importance of balancing model complexity and interpretability is emphasized. The script then extends to multiple regression, incorporating additional predictors such as cholesterol and resting blood pressure. Model summaries and significance tests provide insights into the impact of these variables on Max_heart_rate. Finally, natural cubic splines are introduced as a flexible modeling approach, allowing for non-linear relationships. While the script provides valuable insights, it underscores the necessity of thorough interpretation and validation, such as cross-validation, to ensure the reliability of the regression models.