Cardio-R.R

Student: Nikola Dmitrašinović br.Index: 2018430094 2020-03-28

#tražimo radni folder  
getwd()

## [1] "D:/R"

#učitavamo dataset koji smo sredili u Pythonu  
cardio <- read.csv("cardio\_clean.csv", TRUE, sep = ",")  
#Proveravamo dataset  
class(cardio)

## [1] "data.frame"

head(cardio)

## id age gender height weight ap\_hi ap\_lo cholesterol gluc smoke alco active  
## 1 0 50 2 168 62 110 80 1 1 0 0 1  
## 2 1 55 1 156 85 140 90 3 1 0 0 1  
## 3 2 51 1 165 64 130 70 3 1 0 0 0  
## 4 3 48 2 169 82 150 100 1 1 0 0 1  
## 5 8 60 1 151 67 120 80 2 2 0 0 0  
## 6 9 60 1 157 93 130 80 3 1 0 0 1  
## cardio  
## 1 0  
## 2 1  
## 3 1  
## 4 1  
## 5 0  
## 6 0

str(cardio)

## 'data.frame': 62502 obs. of 13 variables:  
## $ id : int 0 1 2 3 8 9 12 13 15 16 ...  
## $ age : num 50 55 51 48 60 60 61 48 61 51 ...  
## $ gender : int 2 1 1 2 1 1 2 1 1 2 ...  
## $ height : int 168 156 165 169 151 157 178 158 169 173 ...  
## $ weight : num 62 85 64 82 67 93 95 71 80 60 ...  
## $ ap\_hi : int 110 140 130 150 120 130 130 110 120 120 ...  
## $ ap\_lo : int 80 90 70 100 80 80 90 70 80 80 ...  
## $ cholesterol: int 1 3 3 1 2 3 3 1 1 1 ...  
## $ gluc : int 1 1 1 1 2 1 3 1 1 1 ...  
## $ smoke : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ alco : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ active : int 1 1 0 1 0 1 1 1 1 1 ...  
## $ cardio : int 0 1 1 1 0 0 1 0 0 0 ...

#Simple Linear Regression  
#Pravimo linerni model  
SimpleLM = lm(ap\_hi ~ ap\_lo, data = cardio)  
#Prikaz linearnog modela  
SimpleLM

##   
## Call:  
## lm(formula = ap\_hi ~ ap\_lo, data = cardio)  
##   
## Coefficients:  
## (Intercept) ap\_lo   
## 19.062 1.314

#kontrola rezultata poređenjem dobijenog ML i stvarnog  
#Y = m \* X + B  
#Y=api\_high m(slope)=1.314 x=ap\_lo b(intercept)=19.062  
ap\_high\_1 = 1.34\*cardio$ap\_lo[2] + 19.062  
#donji pritisak  
cardio$ap\_lo[2]

## [1] 90

#gornji pritisak  
cardio$ap\_hi[2]

## [1] 140

#dobijena vrednost za gornji pritisak primenom modela SLM  
ap\_high\_1

## [1] 139.662

#grafički prikaz vrednosti gornjeg i donjeg pritiska  
plot(ap\_hi~ap\_lo, data = cardio)  
#dodavanje linije regresije  
abline(SimpleLM, col="red")  
#Prikaz za intercept i slope  
SimpleLM$coefficients

## (Intercept) ap\_lo   
## 19.061509 1.314075

#Regresijom izračunavamo gornji pritisak na osnovu modela  
predicted\_ap\_hi <- predict(SimpleLM)  
head(predicted\_ap\_hi)

## 1 2 3 4 5 6   
## 124.1875 137.3283 111.0468 150.4690 124.1875 124.1875

#Računamo odstupanje modela u odnosu na realne vrednosti  
cardio\_residuals <- cardio$ap\_hi - predicted\_ap\_hi  
head(cardio\_residuals)

## 1 2 3 4 5 6   
## -14.1875332 2.6717137 18.9532199 -0.4690394 -4.1875332 5.8124668

#Računamo srednju vrednost odstupanja  
mean\_cardio\_residuals <- mean(cardio\_residuals)  
mean\_cardio\_residuals

## [1] 5.801155e-11

#Prosek odstupanja na kvadrat  
rmse\_cardio <- sqrt(mean\_cardio\_residuals)  
rmse\_cardio

## [1] 7.616531e-06

#Koren iz proseka odstupanja na kvadrat  
summary(SimpleLM)$r.squared

## [1] 0.4979123

#Multiple Linear Regression  
MultipleLM = lm(ap\_hi~ap\_lo+age+gender+height+weight+cholesterol+gluc+smoke+alco+active+cardio, data = cardio)  
summary(MultipleLM)

##   
## Call:  
## lm(formula = ap\_hi ~ ap\_lo + age + gender + height + weight +   
## cholesterol + gluc + smoke + alco + active + cardio, data = cardio)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -55.397 -6.304 -0.741 3.136 65.302   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 27.546783 1.089052 25.294 < 2e-16 \*\*\*  
## ap\_lo 1.136428 0.005419 209.705 < 2e-16 \*\*\*  
## age 0.120841 0.005928 20.384 < 2e-16 \*\*\*  
## gender 0.657543 0.098252 6.692 2.21e-11 \*\*\*  
## height -0.068793 0.006236 -11.031 < 2e-16 \*\*\*  
## weight 0.081628 0.003410 23.938 < 2e-16 \*\*\*  
## cholesterol 0.761540 0.065641 11.602 < 2e-16 \*\*\*  
## gluc -0.010938 0.075698 -0.144 0.88511   
## smoke 0.416571 0.152383 2.734 0.00626 \*\*   
## alco 0.216632 0.183954 1.178 0.23894   
## active 0.344360 0.096716 3.561 0.00037 \*\*\*  
## cardio 5.546877 0.084508 65.637 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.576 on 62490 degrees of freedom  
## Multiple R-squared: 0.551, Adjusted R-squared: 0.5509   
## F-statistic: 6971 on 11 and 62490 DF, p-value: < 2.2e-16

MultipleLM$coefficients

## (Intercept) ap\_lo age gender height weight   
## 27.54678343 1.13642818 0.12084052 0.65754261 -0.06879334 0.08162764   
## cholesterol gluc smoke alco active cardio   
## 0.76153999 -0.01093762 0.41657068 0.21663189 0.34436043 5.54687720

#kontrola rezultata poređenjem dobijenog ML i stvarnog  
ap\_high\_2 = cardio$ap\_lo[2]\*1.13642818 + 0.65754261\*0.12084052 + cardio$gender[2]\*0.65754261 + cardio$height[2]\*-0.06879334 + cardio$weight[2]\*0.08162764 + cardio$cholesterol[2]\*0.76153999 + cardio$gluc[2]\*-0.01093762 + cardio$smoke[2]\*0.41657068 + cardio$alco[2]\*0.21663189 + cardio$active[2]\*0.34436043 + cardio$cardio[2]\*5.54687720 + 27.54678  
ap\_high\_1

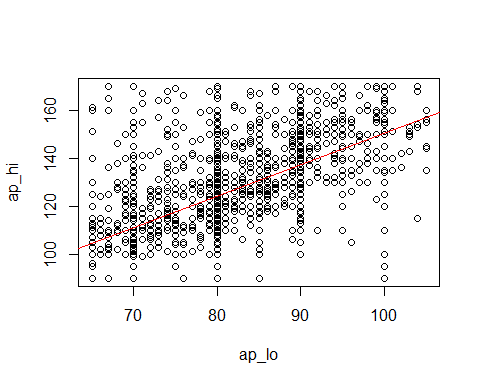
## [1] 139.662

#dobijena vrednost za gornji pritisak primenom modela MLM  
ap\_high\_2

## [1] 134.9338

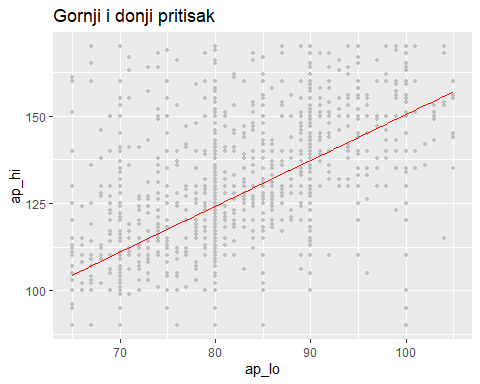
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.3



ggplot(cardio, aes(x=ap\_lo, y=ap\_hi)) +   
 ggtitle("Gornji i donji pritisak") +  
 geom\_point(color="gray", size=1) +  
 stat\_smooth(method = "lm", color="red", se=FALSE, fullrange = TRUE, size=0.5)

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



#Regresijom izračunavamo gornji pritisak na osnovu modela  
predicted\_ap\_hi\_2 <- predict(MultipleLM)  
head(predicted\_ap\_hi\_2)

## 1 2 3 4 5 6   
## 120.4167 141.5006 115.6110 150.0143 122.9515 125.7779

#Računamo odstupanje modela u odnosu na realne vrednosti  
multiple\_cardio\_residuals <- cardio$ap\_hi - predicted\_ap\_hi\_2  
head(multiple\_cardio\_residuals)

## 1 2 3 4 5 6   
## -10.41674486 -1.50059976 14.38900702 -0.01426425 -2.95147439 4.22212887

#Računamo srednju vrednost odstupanja  
multiple\_mean\_cardio\_residuals <- mean(multiple\_cardio\_residuals)  
multiple\_mean\_cardio\_residuals

## [1] 5.716842e-11

#Prosek odstupanja na kvadrat  
multiple\_rmse\_cardio <- sqrt(multiple\_mean\_cardio\_residuals)  
multiple\_rmse\_cardio

## [1] 7.56098e-06

#Koren iz proseka odstupanja na kvadrat  
summary(MultipleLM)$r.squared

## [1] 0.5509897