HOMEWORK-8

Furkan Emir Akdemir

2024-12-17

1.Import and change name of the data

library(readr)

framingham <- read.csv("D:/HOMEWORK-8/framingham.csv")  
datasmt <- framingham

2.Load necessary packages

library(psych)

3.Exploratory Data Analysis

I used psych for more detailed descriptive statistics

describe(datasmt)

## vars n mean sd median trimmed mad min max range  
## male 1 4238 0.43 0.50 0.0 0.41 0.00 0.00 1.0 1.00  
## age 2 4238 49.58 8.57 49.0 49.31 10.38 32.00 70.0 38.00  
## education 3 4133 1.98 1.02 2.0 1.85 1.48 1.00 4.0 3.00  
## currentSmoker 4 4238 0.49 0.50 0.0 0.49 0.00 0.00 1.0 1.00  
## cigsPerDay 5 4209 9.00 11.92 0.0 6.89 0.00 0.00 70.0 70.00  
## BPMeds 6 4185 0.03 0.17 0.0 0.00 0.00 0.00 1.0 1.00  
## prevalentStroke 7 4238 0.01 0.08 0.0 0.00 0.00 0.00 1.0 1.00  
## prevalentHyp 8 4238 0.31 0.46 0.0 0.26 0.00 0.00 1.0 1.00  
## diabetes 9 4238 0.03 0.16 0.0 0.00 0.00 0.00 1.0 1.00  
## totChol 10 4188 236.72 44.59 234.0 234.67 43.00 107.00 696.0 589.00  
## sysBP 11 4238 132.35 22.04 128.0 130.10 19.27 83.50 295.0 211.50  
## diaBP 12 4238 82.89 11.91 82.0 82.20 11.12 48.00 142.5 94.50  
## BMI 13 4219 25.80 4.08 25.4 25.53 3.69 15.54 56.8 41.26  
## heartRate 14 4237 75.88 12.03 75.0 75.20 10.38 44.00 143.0 99.00  
## glucose 15 3850 81.97 23.96 78.0 78.98 11.86 40.00 394.0 354.00  
## TenYearCHD 16 4238 0.15 0.36 0.0 0.07 0.00 0.00 1.0 1.00  
## skew kurtosis se  
## male 0.29 -1.92 0.01  
## age 0.23 -0.99 0.13  
## education 0.69 -0.71 0.02  
## currentSmoker 0.02 -2.00 0.01  
## cigsPerDay 1.25 1.02 0.18  
## BPMeds 5.55 28.77 0.00  
## prevalentStroke 12.90 164.45 0.00  
## prevalentHyp 0.82 -1.33 0.01  
## diabetes 5.99 33.89 0.00  
## totChol 0.87 4.12 0.69  
## sysBP 1.14 2.15 0.34  
## diaBP 0.71 1.27 0.18  
## BMI 0.98 2.65 0.06  
## heartRate 0.64 0.90 0.18  
## glucose 6.21 58.56 0.39  
## TenYearCHD 1.94 1.76 0.01

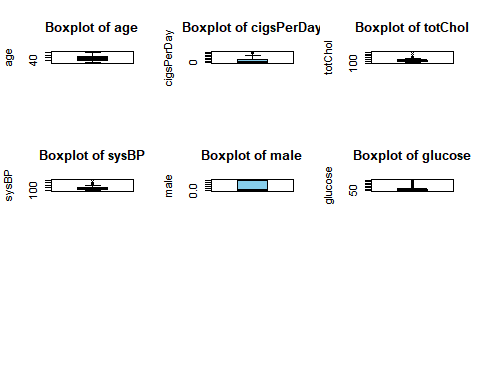
summary(datasmt)

## male age education currentSmoker   
## Min. :0.0000 Min. :32.00 Min. :1.000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:42.00 1st Qu.:1.000 1st Qu.:0.0000   
## Median :0.0000 Median :49.00 Median :2.000 Median :0.0000   
## Mean :0.4292 Mean :49.58 Mean :1.979 Mean :0.4941   
## 3rd Qu.:1.0000 3rd Qu.:56.00 3rd Qu.:3.000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :70.00 Max. :4.000 Max. :1.0000   
## NA's :105   
## cigsPerDay BPMeds prevalentStroke prevalentHyp   
## Min. : 0.000 Min. :0.00000 Min. :0.000000 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.:0.000000 1st Qu.:0.0000   
## Median : 0.000 Median :0.00000 Median :0.000000 Median :0.0000   
## Mean : 9.003 Mean :0.02963 Mean :0.005899 Mean :0.3105   
## 3rd Qu.:20.000 3rd Qu.:0.00000 3rd Qu.:0.000000 3rd Qu.:1.0000   
## Max. :70.000 Max. :1.00000 Max. :1.000000 Max. :1.0000   
## NA's :29 NA's :53   
## diabetes totChol sysBP diaBP   
## Min. :0.00000 Min. :107.0 Min. : 83.5 Min. : 48.00   
## 1st Qu.:0.00000 1st Qu.:206.0 1st Qu.:117.0 1st Qu.: 75.00   
## Median :0.00000 Median :234.0 Median :128.0 Median : 82.00   
## Mean :0.02572 Mean :236.7 Mean :132.4 Mean : 82.89   
## 3rd Qu.:0.00000 3rd Qu.:263.0 3rd Qu.:144.0 3rd Qu.: 89.88   
## Max. :1.00000 Max. :696.0 Max. :295.0 Max. :142.50   
## NA's :50   
## BMI heartRate glucose TenYearCHD   
## Min. :15.54 Min. : 44.00 Min. : 40.00 Min. :0.000   
## 1st Qu.:23.07 1st Qu.: 68.00 1st Qu.: 71.00 1st Qu.:0.000   
## Median :25.40 Median : 75.00 Median : 78.00 Median :0.000   
## Mean :25.80 Mean : 75.88 Mean : 81.97 Mean :0.152   
## 3rd Qu.:28.04 3rd Qu.: 83.00 3rd Qu.: 87.00 3rd Qu.:0.000   
## Max. :56.80 Max. :143.00 Max. :394.00 Max. :1.000   
## NA's :19 NA's :1 NA's :388

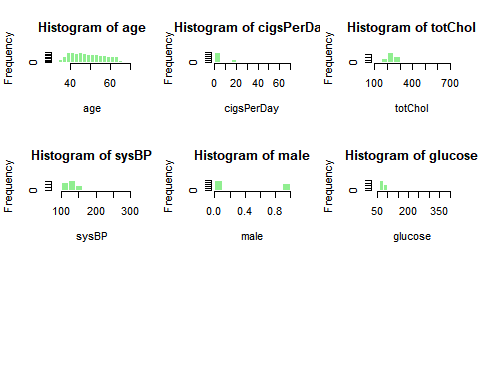
4.Visualization the data to identify outliers and distributions.

Boxplots identify outliers.

# Boxplots for selected numeric variables  
  
numeric\_columns <- c("age", "cigsPerDay", "totChol", "sysBP", "male", "glucose")  
  
#Loop through the numeric columns and create boxplots  
  
par(mfrow = c(3, 3)) # # Layout: 3x3 grid for multiple plots  
  
for (col in numeric\_columns) {  
 boxplot(datasmt[[col]],   
 main = paste("Boxplot of", col),  
 col = "skyblue",  
 ylab = col) }

 Histograms identify distributions.

# Histograms for numeric columns  
  
par(mfrow = c(3, 3)) # Reset to 3x3 layout  
  
for (col in numeric\_columns) {  
 hist(datasmt[[col]],   
 main = paste("Histogram of", col),   
 xlab = col,  
 col = "lightgreen",  
 border = "white")  
}

 #Comments about boxplots and histograms

-cigsPerDay has extreme outliers (heavy smokers). -totChol shows a few high cholesterol values. -sysBP have notable outliers, indicating hypertension cases. -glucose has extreme outliers, suggesting potential diabetes cases -male and age has no outliers. -age appears well-distributed , others is not well-distributed.

5.To decide which model is best

#Multi Linear Regression   
datasmt <- na.omit(datasmt) #Remove missing values  
lm\_model <- lm(TenYearCHD ~ age + male + cigsPerDay + totChol + sysBP + glucose, data = datasmt)  
summary(lm\_model) #summary model

##   
## Call:  
## lm(formula = TenYearCHD ~ age + male + cigsPerDay + totChol +   
## sysBP + glucose, data = datasmt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7356 -0.1925 -0.1086 -0.0098 1.1063   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.7193393 0.0482875 -14.897 < 2e-16 \*\*\*  
## age 0.0075048 0.0007450 10.073 < 2e-16 \*\*\*  
## male 0.0559223 0.0121059 4.619 3.98e-06 \*\*\*  
## cigsPerDay 0.0023978 0.0005129 4.675 3.05e-06 \*\*\*  
## totChol 0.0001131 0.0001346 0.840 0.401   
## sysBP 0.0024422 0.0002814 8.677 < 2e-16 \*\*\*  
## glucose 0.0012617 0.0002393 5.272 1.42e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3416 on 3649 degrees of freedom  
## Multiple R-squared: 0.098, Adjusted R-squared: 0.09651   
## F-statistic: 66.07 on 6 and 3649 DF, p-value: < 2.2e-16

#Logit Model  
logit\_model <- glm(TenYearCHD ~ age + male + cigsPerDay + totChol + sysBP + glucose, data = datasmt, family = binomial(link="logit"))  
summary(logit\_model)

##   
## Call:  
## glm(formula = TenYearCHD ~ age + male + cigsPerDay + totChol +   
## sysBP + glucose, family = binomial(link = "logit"), data = datasmt)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.129843 0.475530 -19.199 < 2e-16 \*\*\*  
## age 0.065896 0.006426 10.254 < 2e-16 \*\*\*  
## male 0.561446 0.106845 5.255 1.48e-07 \*\*\*  
## cigsPerDay 0.019226 0.004176 4.604 4.14e-06 \*\*\*  
## totChol 0.002272 0.001123 2.024 0.043 \*   
## sysBP 0.017534 0.002149 8.159 3.38e-16 \*\*\*  
## glucose 0.007280 0.001677 4.342 1.41e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3120.5 on 3655 degrees of freedom  
## Residual deviance: 2762.5 on 3649 degrees of freedom  
## AIC: 2776.5  
##   
## Number of Fisher Scoring iterations: 5

#Probit Model  
probit\_model <- glm(TenYearCHD ~ age + male + cigsPerDay + totChol + sysBP + glucose, data = datasmt, family = binomial(link= "probit"))  
summary(probit\_model)

##   
## Call:  
## glm(formula = TenYearCHD ~ age + male + cigsPerDay + totChol +   
## sysBP + glucose, family = binomial(link = "probit"), data = datasmt)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.0644162 0.2517734 -20.115 < 2e-16 \*\*\*  
## age 0.0361972 0.0035092 10.315 < 2e-16 \*\*\*  
## male 0.2897210 0.0580668 4.989 6.06e-07 \*\*\*  
## cigsPerDay 0.0110349 0.0023319 4.732 2.22e-06 \*\*\*  
## totChol 0.0011302 0.0006242 1.811 0.0702 .   
## sysBP 0.0097450 0.0012246 7.958 1.75e-15 \*\*\*  
## glucose 0.0041004 0.0009808 4.181 2.90e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3120.5 on 3655 degrees of freedom  
## Residual deviance: 2763.8 on 3649 degrees of freedom  
## AIC: 2777.8  
##   
## Number of Fisher Scoring iterations: 5

Compare Models

Linear Regression: The r-squared value is 0.098 and the r-squared model explains %9.8, this value is very low, so we can say that the model is not suitable.

Logit Model and Probit Model: AIC value of the logit model = 2776.5, AIC value of the probit model = 2777.8, the AIC values of both models are quite close, but since the lower AIC value is more compatible with the model, we say that the logit model is more compatible.

As a result, the best fitting model is the logit model.

5.Comment about model and variables

Logit Model “Pr(>|z|)” shows the p-value for each coefficient. Since all independent variables have p-values < 0.05, they are all statistically significant. This indicates that there is strong evidence that these variables affect the risk of developing coronary heart disease within 10 years.

age:The coefficient is 0.065896. This means that each unit increase in age (for example, 1 year) increases the log-odds ratio of the probability of developing coronary heart disease within 10 years by 0.065896 units.

male:The coefficient is 0.561446. Men have a log-odds ratio of 0.561446 units greater chance of developing coronary heart disease within 10 years than women.

cigsPerDay:The coefficient is 0.019226. It means that each cigarette smoked per day increases the log-odds ratio of the probability of developing coronary heart disease within 10 years by 0.019226 units.

totChol: The coefficient is 0.002272. It means that each unit increase in total cholesterol increases the log-odds ratio of the probability of developing coronary heart disease within 10 years by 0.002272 units.

sysBP:The coefficient is 0.017534. It means that each unit increase in systolic blood pressure increases the log-odds ratio of the probability of developing coronary heart disease within 10 years by 0.017534 units.

Glucose:The coefficient is 0.007280. It means that each unit increase in blood glucose level increases the log-odds ratio of the probability of developing coronary heart disease within 10 years by 0.007280 units.

This logistic regression model examines the factors that affect the risk of developing coronary heart disease over 10 years. The model shows that age, gender, number of cigarettes per day, total cholesterol, systolic blood pressure, and blood sugar levels significantly affect this risk.

#Real life application

I think that even though a life-threatening event like a heart attack may not have much of an effect on gender, it does have a significant effect on age and other factors.Because daily cigarette consumption has important effects on organs such as the heart and lungs, at the same time, high glucose levels, cholesterol and blood pressure levels are important values for the human body and health and should be checked regularly.