628 Module2

2024-10-03

library(MASS)  
library(car)  
library(caret)  
data <- read.csv("BodyFat.csv")

# Using IQR to find outliers

find\_iqr\_outliers <- function(x) {  
 Q1 <- quantile(x, 0.25, na.rm = TRUE)  
 Q3 <- quantile(x, 0.75, na.rm = TRUE)  
 IQR\_value <- Q3 - Q1  
 lower\_bound <- Q1 - 1.5 \* IQR\_value  
 upper\_bound <- Q3 + 1.5 \* IQR\_value  
 x < lower\_bound | x > upper\_bound  
}  
  
data\_no\_id <- data[ , !(names(data) %in% c("IDNO"))]  
iqr\_outliers\_flags <- as.data.frame(lapply(data\_no\_id, find\_iqr\_outliers))  
outliers\_original <- data[apply(iqr\_outliers\_flags, 1, any), ]  
outliers\_columns <- apply(iqr\_outliers\_flags[apply(iqr\_outliers\_flags, 1, any), ], 1, function(x) {  
 colnames(data\_no\_id)[x]  
})  
outliers\_original$Outlier\_Variables <- sapply(outliers\_columns, paste, collapse = ", ")  
print(outliers\_original)

## IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN HIP  
## 31 31 12.3 1.0716 32 182.00 73.75 23.6 38.7 100.5 88.7 99.8  
## 35 35 31.1 1.0263 41 247.25 73.50 32.2 42.1 117.0 115.6 116.1  
## 39 39 33.8 1.0202 46 363.15 72.25 48.9 51.2 136.2 148.1 147.7  
## 41 41 33.1 1.0217 45 262.75 68.75 39.1 43.2 128.3 126.2 125.6  
## 42 42 31.7 1.0250 44 205.00 29.50 29.9 36.6 106.0 104.3 115.5  
## 45 45 8.4 1.0814 39 125.25 68.00 19.1 31.5 85.1 76.0 88.2  
## 86 86 25.8 1.0386 67 167.00 67.50 26.0 36.5 98.9 89.7 96.2  
## 106 106 17.8 1.0578 43 165.50 68.50 24.8 31.1 93.1 87.3 96.6  
## 152 152 19.1 1.0542 26 241.75 74.50 30.7 41.8 108.3 102.9 114.4  
## 159 159 12.8 1.0704 30 136.50 68.75 20.3 35.9 88.7 76.6 89.8  
## 169 169 34.7 1.0180 35 228.25 69.50 33.3 40.4 114.9 115.9 111.9  
## 175 175 24.6 1.0414 36 226.75 71.75 31.0 41.5 115.3 108.8 114.4  
## 192 192 36.5 1.0140 42 244.25 76.00 29.8 41.8 115.2 113.7 112.4  
## 206 206 16.6 1.0610 44 208.75 73.00 27.6 41.9 105.6 96.3 102.0  
## 216 216 45.1 0.9950 51 219.00 64.00 37.6 41.2 119.8 122.1 112.8  
## 226 226 12.8 1.0703 55 126.50 66.75 20.0 33.4 88.8 78.2 87.5  
## 242 242 33.6 1.0207 65 224.50 68.25 33.9 38.8 119.6 118.0 114.3  
## 244 244 31.4 1.0256 67 227.75 72.75 30.3 41.3 115.8 113.4 109.8  
## 252 252 30.7 1.0271 74 207.50 70.00 29.8 40.8 112.4 108.5 107.1  
## THIGH KNEE ANKLE BICEPS FOREARM WRIST  
## 31 57.5 38.7 33.9 32.5 27.7 18.4  
## 35 71.2 43.3 26.3 37.3 31.7 19.7  
## 39 87.3 49.1 29.6 45.0 29.0 21.4  
## 41 72.5 39.6 26.6 36.4 32.7 21.4  
## 42 70.6 42.5 23.7 33.6 28.7 17.4  
## 45 50.0 34.7 21.0 26.1 23.1 16.1  
## 86 54.7 37.8 33.7 32.4 27.7 18.2  
## 106 54.7 39.0 24.8 31.0 29.4 18.8  
## 152 72.9 43.5 25.1 38.5 33.8 19.6  
## 159 50.1 34.8 21.8 27.0 34.9 16.9  
## 169 74.4 40.6 24.0 36.1 31.8 18.8  
## 175 69.2 42.4 24.0 35.4 21.0 20.1  
## 192 68.5 45.0 25.5 37.1 31.2 19.9  
## 206 63.3 39.8 24.1 37.3 23.1 19.4  
## 216 62.5 36.9 23.6 34.7 29.1 18.4  
## 226 50.8 33.0 19.7 25.3 22.0 15.8  
## 242 61.3 42.1 23.4 34.9 30.1 19.4  
## 244 65.6 46.0 25.4 35.3 29.8 19.5  
## 252 59.3 42.2 24.6 33.7 30.0 20.9  
## Outlier\_Variables  
## 31 ANKLE  
## 35 HIP  
## 39 WEIGHT, ADIPOSITY, NECK, CHEST, ABDOMEN, HIP, THIGH, KNEE, ANKLE, BICEPS, WRIST  
## 41 WEIGHT, ADIPOSITY, CHEST, ABDOMEN, HIP, THIGH, WRIST  
## 42 HEIGHT  
## 45 NECK, FOREARM  
## 86 ANKLE  
## 106 NECK  
## 152 THIGH  
## 159 FOREARM  
## 169 THIGH  
## 175 FOREARM  
## 192 KNEE  
## 206 FOREARM  
## 216 BODYFAT, DENSITY, ADIPOSITY, ABDOMEN  
## 226 FOREARM, WRIST  
## 242 ADIPOSITY  
## 244 KNEE  
## 252 WRIST

data <- data[!(data$IDNO %in% c(39, 41, 216)), ]

Delete 39, 41, 216 because they are outliers in many variables.

# Using adiposity to adjust height for 42

data$HEIGHT <-round(data$HEIGHT \* 2.54 , 2)  
weight\_kg <- data$WEIGHT[42] \* 0.453592   
height\_m <- sqrt(weight\_kg / data$ADIPOSITY[42])   
height\_cm <- round(height\_m \* 100, 2)   
data$HEIGHT[42] <- height\_cm  
data[data$IDNO == 42, ]

## IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN HIP  
## 42 42 31.7 1.025 44 205 74.93 29.9 36.6 106 104.3 115.5  
## THIGH KNEE ANKLE BICEPS FOREARM WRIST  
## 42 70.6 42.5 23.7 33.6 28.7 17.4

#Preprocessed data  
df <- data[, !names(data) %in% c("IDNO", "DENSITY")]  
write.csv(df, "data.csv", row.names = FALSE)

# Stepwise Regression

d <- read.csv("data.csv")  
dim(d)

## [1] 249 15

full\_model <- lm(BODYFAT ~ ., data = d)  
model1 <- stepAIC(full\_model, direction = "both", trace = FALSE)  
summary(model1)

##   
## Call:  
## lm(formula = BODYFAT ~ AGE + WEIGHT + NECK + ABDOMEN + THIGH +   
## FOREARM + WRIST, data = d)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.7112 -2.7272 -0.2259 2.6502 9.3716   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -29.33016 8.35770 -3.509 0.000536 \*\*\*  
## AGE 0.06812 0.02859 2.383 0.017959 \*   
## WEIGHT -0.08891 0.03246 -2.739 0.006616 \*\*   
## NECK -0.32208 0.20465 -1.574 0.116847   
## ABDOMEN 0.81452 0.06668 12.215 < 2e-16 \*\*\*  
## THIGH 0.23141 0.10694 2.164 0.031455 \*   
## FOREARM 0.36390 0.18005 2.021 0.044373 \*   
## WRIST -1.43551 0.47741 -3.007 0.002919 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.921 on 241 degrees of freedom  
## Multiple R-squared: 0.7345, Adjusted R-squared: 0.7268   
## F-statistic: 95.27 on 7 and 241 DF, p-value: < 2.2e-16

adjusted\_r\_squared1 <- round(summary(model1)$adj.r.squared,3)

# Simple Linear Regression

variables <- colnames(d)  
variables <- variables[variables != "BODYFAT"]  
  
max\_adjusted\_r\_squared <- 0   
best\_variable <- ""  
for (var in variables) {  
 formula <- as.formula(paste("BODYFAT ~", var))  
 model <- lm(formula, data = d)  
 adjusted\_r\_squared <- summary(model)$adj.r.squared  
 if (adjusted\_r\_squared > max\_adjusted\_r\_squared) {  
 max\_adjusted\_r\_squared <- adjusted\_r\_squared  
 best\_variable <- var  
 }  
}  
  
best\_formula <- as.formula(paste("BODYFAT ~", best\_variable))  
model2 <- lm(best\_formula, data = d)  
adjusted\_r\_squared2 <- round(max\_adjusted\_r\_squared, 3)

# Two variable linear regression

max\_adjusted\_r\_squared <- 0  
best\_variable\_pair <- c("","")  
  
  
for (i in 1:(length(variables) - 1)) {  
 for (j in (i + 1):length(variables)) {  
 var1 <- variables[i]  
 var2 <- variables[j]  
 formula <- as.formula(paste("BODYFAT ~", var1, "+", var2))  
 model <- lm(formula, data = d)  
  
 adjusted\_r\_squared <- summary(model)$adj.r.squared  
  
 if (adjusted\_r\_squared > max\_adjusted\_r\_squared) {  
 max\_adjusted\_r\_squared <- adjusted\_r\_squared  
 best\_variable\_pair <- c(var1, var2)  
 }  
 }  
}  
  
best\_formula <- as.formula(paste("BODYFAT ~", best\_variable\_pair[1], "+", best\_variable\_pair[2]))  
model3 <- lm(best\_formula, data = d)  
adjusted\_r\_squared3 <- round(max\_adjusted\_r\_squared, 3)

# Bivariate product term

max\_adjusted\_r\_squared <- 0  
best\_variable\_pair <- c("","")  
  
for (i in 1:(length(variables) - 1)) {  
 for (j in (i + 1):length(variables)) {  
 var1 <- variables[i]  
 var2 <- variables[j]  
 d$interaction\_term <- d[[var1]] \* d[[var2]]  
 formula <- as.formula("BODYFAT ~ interaction\_term")  
 model <- lm(formula, data = d)  
 adjusted\_r\_squared <- summary(model)$adj.r.squared  
 if (adjusted\_r\_squared > max\_adjusted\_r\_squared) {  
 max\_adjusted\_r\_squared <- adjusted\_r\_squared  
 best\_variable\_pair <- c(var1, var2)  
 }  
 }  
}  
  
  
d$best\_interaction\_term <- d[[best\_variable\_pair[1]]] \* d[[best\_variable\_pair[2]]]  
best\_formula <- as.formula("BODYFAT ~ best\_interaction\_term")  
model4 <- lm(best\_formula, data = d)  
adjusted\_r\_squared4 <- round(max\_adjusted\_r\_squared, 3)

Notice that all the best model includes the variable ABDOMEN, so try its higher-order terms.

# ABDOMEN^2

d$ABDOMEN\_squared <- d$ABDOMEN^2  
model5 <- lm(BODYFAT ~ ABDOMEN\_squared, data = d)  
adjusted\_r\_squared5 <- round(summary(model5)$adj.r.squared, 3)

# ABDOMEN^2 + other 1-order

variables <- variables[!variables %in% c("BODYFAT", "ABDOMEN", "ABDOMEN\_squared")]  
  
max\_adjusted\_r\_squared <- 0  
best\_variable <- ""  
  
for (var in variables) {  
 formula <- as.formula(paste("BODYFAT ~ ABDOMEN\_squared +", var))  
 model <- lm(formula, data = d)  
 adjusted\_r\_squared <- summary(model)$adj.r.squared  
 if (adjusted\_r\_squared > max\_adjusted\_r\_squared) {  
 max\_adjusted\_r\_squared <- adjusted\_r\_squared  
 best\_variable <- var  
 }  
}  
  
best\_formula <- as.formula(paste("BODYFAT ~ ABDOMEN\_squared +", best\_variable))  
model6 <- lm(best\_formula, data = d)  
adjusted\_r\_squared6 <- round(max\_adjusted\_r\_squared, 3)

# ABDOMEN^2 +ABDOMEN +other variable 1-order

max\_adjusted\_r\_squared <- 0  
best\_variable <- ""  
for (var in variables) {  
   
 formula <- as.formula(paste("BODYFAT ~ ABDOMEN + ABDOMEN\_squared +", var))  
 model <- lm(formula, data = d)  
 adjusted\_r\_squared <- summary(model)$adj.r.squared  
 if (adjusted\_r\_squared > max\_adjusted\_r\_squared) {  
 max\_adjusted\_r\_squared <- adjusted\_r\_squared  
 best\_variable <- var  
 }  
}  
  
best\_formula <- as.formula(paste("BODYFAT ~ ABDOMEN + ABDOMEN\_squared +", best\_variable))  
model7 <- lm(best\_formula, data = d)  
adjusted\_r\_squared7 <- round(max\_adjusted\_r\_squared, 3)

# 10-fold CV

train\_control <- trainControl(method = "cv", number = 10)  
mse\_results <- c()  
  
for (i in 1:7) {  
 model <- get(paste0("model", i))  
 cv\_model <- train(  
 formula(model),  
 data = d,   
 method = "lm",   
 trControl = train\_control,   
 metric = "RMSE"   
 )  
 mse\_results <- c(mse\_results, mean((cv\_model$results$RMSE)^2))  
}

model\_names <- c(  
 "BODYFAT ~ AGE + WEIGHT + NECK + ABDOMEN + THIGH + FOREARM + WRIST",   
 "BODYFAT ~ ABDOMEN ",   
 "BODYFAT ~ WEIGHT + ABDOMEN",   
 "BODYFAT ~ ABDOMEN \* WEIGHT",   
 "BODYFAT ~ ABDOMEN\_squared",   
 "BODYFAT ~ ABDOMEN\_squared + WEIGHT",   
 "BODYFAT ~ ABDOMEN + ABDOMEN\_squared + WEIGHT"   
)  
mse\_results <- round(mse\_results, 3)  
  
results\_df <- data.frame(  
 model = model\_names,  
 adjusted\_R\_squared = c(adjusted\_r\_squared1, adjusted\_r\_squared2, adjusted\_r\_squared3, adjusted\_r\_squared4, adjusted\_r\_squared5, adjusted\_r\_squared6, adjusted\_r\_squared7),  
 MSE = mse\_results  
)  
  
print(results\_df)

## model  
## 1 BODYFAT ~ AGE + WEIGHT + NECK + ABDOMEN + THIGH + FOREARM + WRIST  
## 2 BODYFAT ~ ABDOMEN   
## 3 BODYFAT ~ WEIGHT + ABDOMEN  
## 4 BODYFAT ~ ABDOMEN \* WEIGHT  
## 5 BODYFAT ~ ABDOMEN\_squared  
## 6 BODYFAT ~ ABDOMEN\_squared + WEIGHT  
## 7 BODYFAT ~ ABDOMEN + ABDOMEN\_squared + WEIGHT  
## adjusted\_R\_squared MSE  
## 1 0.727 15.416  
## 2 0.666 18.568  
## 3 0.708 16.483  
## 4 0.615 21.400  
## 5 0.659 19.057  
## 6 0.696 17.063  
## 7 0.709 16.582