BIGDATA CODE REPORT -STATS 4710/7710

Bruno, Drushya Ramesh and Fei Wang

## the whole idea, find a good model to explain or predict the relationship

## between the response variable and predictor variables.

**Step 1:** input data and use str() & summary()to see the detail of the data

data <- read.table("C:/Users/Dell/Downloads/fat.dat (2).txt")

column\_names <- c("Case Number","Brozek.equation","Siri.equation","Density",  
   
 "Age","Weight","Height","Adiposity.index","Fat.Free.Weight",  
   
 "Neck.cir","Chest.cir","Abdomen.cir","Hip.cir","Thigh.cir",  
   
 "Knee.cir","Ankle.cir","Extended.biceps.cir","Forearm.cir",  
   
 "Wrist.cir")

colnames(data) = column\_names

newdata <- data[c(-48, -76, -96, -182),]

str(newdata)

## 'data.frame': 248 obs. of 19 variables:  
## $ Case Number : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Brozek.equation : num 12.6 6.9 24.6 10.9 27.8 20.6 19 12.8 5.1 12 ...  
## $ Siri.equation : num 12.3 6.1 25.3 10.4 28.7 20.9 19.2 12.4 4.1 11.7 ...  
## $ Density : num 1.07 1.09 1.04 1.08 1.03 ...  
## $ Age : int 23 22 22 26 24 24 26 25 25 23 ...  
## $ Weight : num 154 173 154 185 184 ...  
## $ Height : num 67.8 72.2 66.2 72.2 71.2 ...  
## $ Adiposity.index : num 23.7 23.4 24.7 24.9 25.6 26.5 26.2 23.6 24.6 25.8 ...  
## $ Fat.Free.Weight : num 135 161 116 165 133 ...  
## $ Neck.cir : num 36.2 38.5 34 37.4 34.4 39 36.4 37.8 38.1 42.1 ...  
## $ Chest.cir : num 93.1 93.6 95.8 101.8 97.3 ...  
## $ Abdomen.cir : num 85.2 83 87.9 86.4 100 94.4 90.7 88.5 82.5 88.6 ...  
## $ Hip.cir : num 94.5 98.7 99.2 101.2 101.9 ...  
## $ Thigh.cir : num 59 58.7 59.6 60.1 63.2 66 58.4 60 62.9 63.1 ...  
## $ Knee.cir : num 37.3 37.3 38.9 37.3 42.2 42 38.3 39.4 38.3 41.7 ...  
## $ Ankle.cir : num 21.9 23.4 24 22.8 24 25.6 22.9 23.2 23.8 25 ...  
## $ Extended.biceps.cir: num 32 30.5 28.8 32.4 32.2 35.7 31.9 30.5 35.9 35.6 ...  
## $ Forearm.cir : num 27.4 28.9 25.2 29.4 27.7 30.6 27.8 29 31.1 30 ...  
## $ Wrist.cir : num 17.1 18.2 16.6 18.2 17.7 18.8 17.7 18.8 18.2 19.2 ...

**Step 2:** clear the data.

Obviously, the case number (column 1) is not a significant variable, keep it, but not use it as a predictor variable case 48, 76 and 96 have one digit in error; case 182, the body fat is zero, it means the raw data is wrong. these all can be defined as bad data, so delete them. case 42, fix the data value.

newdata $ Height[42] <- 69.5

newdata = newdata[,-c(1,3)]

**Step 3** This step where we created new variables to get the predictor variable.

newdata$hip\_thigh <- (newdata$Hip.cir+newdata$Thigh.cir)/2  
newdata$knee\_ankle <- (newdata$Knee.cir+newdata$Ankle.cir)/2  
newdata$hand <- (newdata$Extended.biceps.cir + newdata$Forearm.cir + newdata$Wrist.cir)/3

**Step 4** The step of creating new variables (hip\_thigh, knee\_ankle, hand) by combining existing measurements to simplify and enhance the correlation. These were then ploted and it made us ananlyase the differences.

library(ggplot2)

par(mfrow = c(4,1))

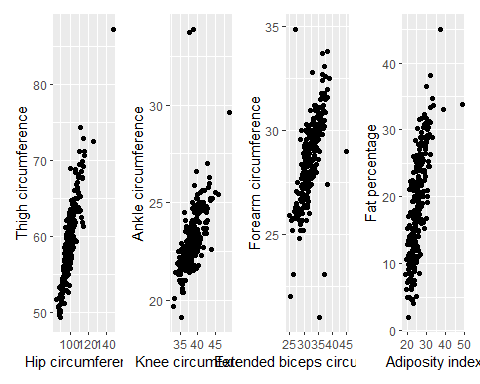
p1 = ggplot(newdata) + aes(x= Hip.cir, y = Thigh.cir) + geom\_point() + labs(x = "Hip circumference", y = "Thigh circumference")

p2 = ggplot(newdata) + aes(x= Knee.cir, y = Ankle.cir) + geom\_point() + labs(x = "Knee circumference", y = "Ankle circumference")

p3 = ggplot(newdata) + aes(x= Extended.biceps.cir, y = Forearm.cir) + geom\_point() + labs(x = "Extended biceps circumference", y = "Forearm circumference")

p4 = ggplot(newdata) + aes(y=Brozek.equation , x = Adiposity.index) + geom\_point() + labs(y = "Fat percentage", x = "Adiposity index")

p1 + p2 + p3 + p4 + plot\_layout(ncol = 4)



**Step 5** Check for missing values

if (any(is.na(newdata))) {  
 # If missing values exist, remove them  
 newdata <- na.omit(newdata)  
}

**Step 6** Check for any empty dataset

# Check if the dataset is empty  
if (nrow(newdata) == 0 || ncol(newdata) == 0) {  
 stop("The dataset is empty.")  
}

**Step 6**

The steps involve preparing data for training and evaluating a predictive model. First, the dataset is split into training (train\_data) and testing (test\_data) sets using a 70-30 ratio.

library(caret)  
set.seed(1)   
train\_index <- createDataPartition(newdata$Brozek.equation, p = 0.7, list = FALSE)  
train\_data <- newdata[train\_index, ]  
test\_data <- newdata[-train\_index, ]

**Step 7**

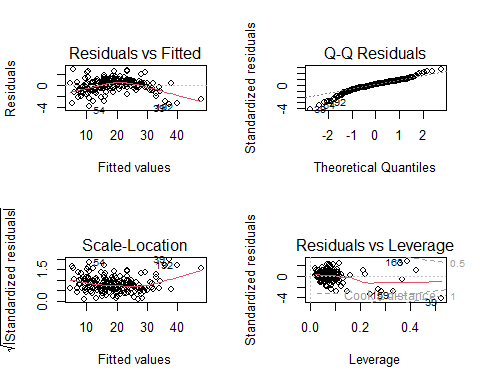
This step involves fitting a linear regression model lm to the training data to predict fat percentage based on several predictor variables including age, weight, height, body circumferences, fatfree weight and neck circumference. Then summarized and ploted to check for the linearity, independence of variables, outliers and to check if variables were constant or not.

# Fit the model  
fit <- lm(Brozek.equation ~ Age + Weight + Height + Adiposity.index + Fat.Free.Weight + Neck.cir +  
 Abdomen.cir + Hip.cir + Thigh.cir + Knee.cir + Ankle.cir + Extended.biceps.cir + Forearm.cir +  
 Wrist.cir , data = train\_data)

summary(fit)

##   
## Call:  
## lm(formula = Brozek.equation ~ Age + Weight + Height + Adiposity.index +   
## Fat.Free.Weight + Neck.cir + Abdomen.cir + Hip.cir + Thigh.cir +   
## Knee.cir + Ankle.cir + Extended.biceps.cir + Forearm.cir +   
## Wrist.cir, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8113 -0.5059 0.2003 0.7342 2.8342   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -64.286330 12.400429 -5.184 6.43e-07 \*\*\*  
## Age 0.012952 0.011745 1.103 0.271772   
## Weight 0.176770 0.034485 5.126 8.40e-07 \*\*\*  
## Height 0.954487 0.172023 5.549 1.16e-07 \*\*\*  
## Adiposity.index 0.825715 0.242038 3.412 0.000817 \*\*\*  
## Fat.Free.Weight -0.511657 0.013523 -37.836 < 2e-16 \*\*\*  
## Neck.cir 0.044949 0.080345 0.559 0.576628   
## Abdomen.cir 0.181352 0.036604 4.954 1.82e-06 \*\*\*  
## Hip.cir -0.087143 0.048822 -1.785 0.076160 .   
## Thigh.cir 0.198590 0.054880 3.619 0.000396 \*\*\*  
## Knee.cir -0.029052 0.088725 -0.327 0.743764   
## Ankle.cir 0.116931 0.071544 1.634 0.104128   
## Extended.biceps.cir 0.112626 0.060912 1.849 0.066290 .   
## Forearm.cir 0.341568 0.080728 4.231 3.89e-05 \*\*\*  
## Wrist.cir -0.004984 0.204746 -0.024 0.980609   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.264 on 161 degrees of freedom  
## Multiple R-squared: 0.9755, Adjusted R-squared: 0.9734   
## F-statistic: 457.7 on 14 and 161 DF, p-value: < 2.2e-16

par(mfrow = c(2,2))  
plot(fit)



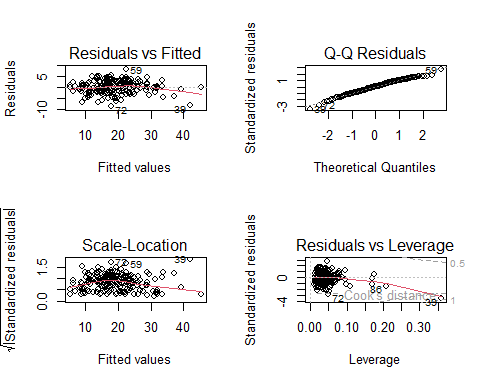
**Step 8** After fitting the initial linear regression model (fit) with a set of predictor variables, a second regression model (fit2) is constructed using a refined subset of predictors (Age, Adiposity.index, Fat.Free.Weight, Neck.cir, hip\_thigh, knee\_ankle, hand) on the train\_data. Then plotted the same.

fit2 <- lm(Brozek.equation ~ Age + Adiposity.index + Fat.Free.Weight + Neck.cir   
 + hip\_thigh + knee\_ankle + hand, data = train\_data)

summary(fit2)

##   
## Call:  
## lm(formula = Brozek.equation ~ Age + Adiposity.index + Fat.Free.Weight +   
## Neck.cir + hip\_thigh + knee\_ankle + hand, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.5448 -1.9960 0.2446 2.1597 7.9377   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -67.89947 6.56032 -10.350 < 2e-16 \*\*\*  
## Age 0.05154 0.02152 2.394 0.017749 \*   
## Adiposity.index 0.76365 0.16117 4.738 4.57e-06 \*\*\*  
## Fat.Free.Weight -0.40311 0.02300 -17.528 < 2e-16 \*\*\*  
## Neck.cir 0.52020 0.18125 2.870 0.004632 \*\*   
## hip\_thigh 0.64853 0.11409 5.684 5.69e-08 \*\*\*  
## knee\_ankle 0.95229 0.21437 4.442 1.61e-05 \*\*\*  
## hand 0.84875 0.24353 3.485 0.000627 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.021 on 168 degrees of freedom  
## Multiple R-squared: 0.8539, Adjusted R-squared: 0.8478   
## F-statistic: 140.2 on 7 and 168 DF, p-value: < 2.2e-16

par(mfrow = c(2,2))  
  
plot(fit2)

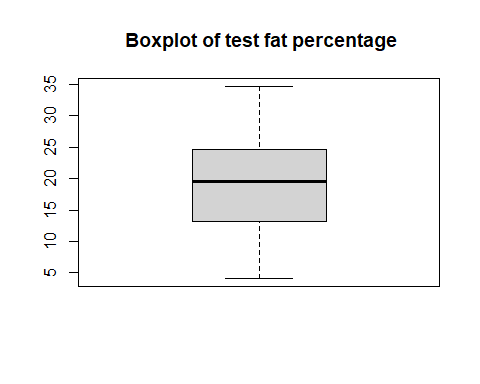


**Step 9** These steps are essential for evaluating the refined linear regression model (fit2) on unseen test data. First, we use fit2 to predict Brozek.equation (fat percentage) values for the test dataset. The mean squared error was analyzed and then box plot was plotted to visualize the data.

predicted\_fat <- predict(fit2, newdata = test\_data)  
  
mean((predicted\_fat-test\_data$Brozek.equation)^2)

## [1] 8.340314

boxplot(test\_data$Brozek.equation, main = " Boxplot of test fat percentage")



**Conclusion:**

These steps collectively assess the effectiveness and reliability of the model in predicting fat percentages based on the selected predictor variables.