Colin White
February 2021

HOMEWORK ANALYSIS #4 –

BODY FAT Measuring body fat is not simple. One method requires submerging the body underwater in a tank and measuring the increase in water level. A simpler method for estimating body fat would be preferred. In order to develop such a method, researchers recorded age, weight, height, and 10 body circumference measurements for 252 men. Each man's percentage of body fat was accurately estimated by an underwater weighing technique. The data can be found in the BodyFat dataset (the variable brozek is the percentage of body fat). For each of the following questions, assume that your audience are nurses with moderate statistical training. Use full sentences for answers that request commentary. Please attach your clearly commented code to the back of your answers as an appendix.

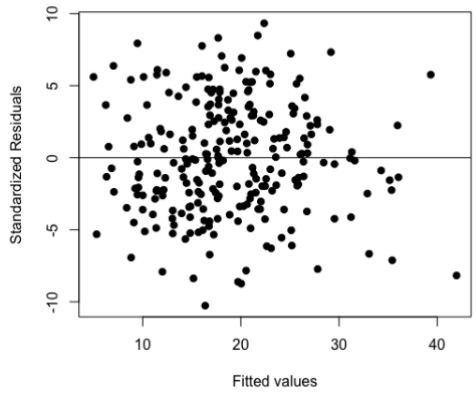
1. In your own words, summarize the overarching problem and any specific questions that need to be answered using the BodyFat data. Discuss how statistical modeling will be able to answer the posed questions.

Health care professionals need a cheaper and easier way to predict body fat than current methods. We are trying to see is measuring circumference of different body parts can predict overall body fat. If the circumference data passes the LINE assumptions, we can fit a MLR model to the data and use it to predict the body fat of individuals. This would work by taking a measure of one of more circumferences, plugging it into the MLR model, and the model predicting the body fat.

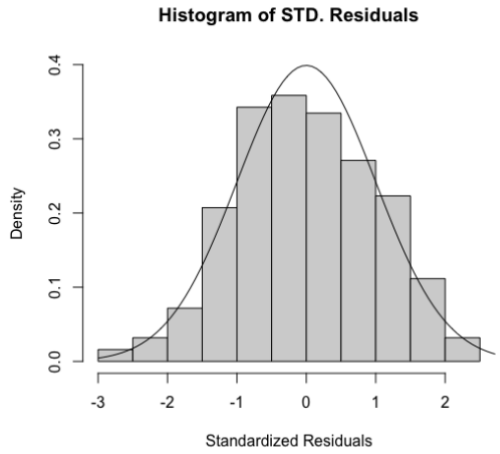
2. Explore the data using basic exploratory graphics and summary statistics. Comment on any potential relationships you see through this exploratory analysis.

For any data to be suitable to analyze with a MLR model, it needs to pass 4 assumptions. These assumptions are that there is a linear relationship between covariates (circumferences) and response (body fat), that errors are independent of one another and of the covariates (circumferences), that errors are normally distributed, and that there is equal variance for all ϵi.

The residuals vs. fitted values scatterplot bellow suggests that the data passes the Linear, Independent and Equal Variance Assumptions.



The histogram of standardized residuals below suggests that the data passes the normality assumption. I propose this is because of the relatively small number of observations, and if more observations were made that histogram would normalize.



The normal quantile-quantile (QQ) plot below also suggests that the data meets the normality assumption.

3. Write out (in mathematical form with greek letters) a MLR model that would help answer the questions you stated in #1. Provide an interpretation of the intercept and at least 1 slope coefficient included in your model.

$Y = ß_0 + ß_1*x_1 + ß_2*x_2 + ß_3*x_3 + ß_4*x_4 + ß_5*x_5 + ß_6*x_6 + ß_7*x_7 + ß_8*x_8 + ß_9*x_9 + ß_{10}*x_{10} + ß_{11}*x_{11} + ß_{12}*x_{12} + ß_{13}*x_{13} + ß_{14}*x_{14} + \epsilon_i$ where $\epsilon_i \sim {}^{iid}N(0, \sigma^2)$

Y = Predicted body fat
$ß_0$ = The predicted body fat on average if all the measurements were at zero. It is the Y axis intercept.

$ß_1$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_1 ((snowfall) in inches)
x_1 = Measured age

$ß_2$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_2
x_2 = Measured weight

$ß_3$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_3
x_3 = Measured height

$ß_4$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_4
x_4 = Measured neck circumference

$ß_5$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_5
x_5 = Measured chest circumference

$ß_6$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_6
x_6 = Measured abdomen circumference

$ß_7$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_7
x_7 = Measured hip circumference

$ß_8$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_8
x_8 = Measured thigh circumference

$ß_9$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_9
x_9 = Measured knee circumference

$ß_{10}$ = Holding all other x's constant this is the holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_{10}
x_{10} = Measured ankle circumference

$ß_{11}$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_{11}
x_{11} = Measured bicep circumference

$ß_{12}$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_{12}

x_{12} = Measured forearm circumference

$ß_{13}$ = Holding all other x's constant this is the amount the predicted body fat increase on average per unit increase x_{13}
x_{13} = Measured wrist circumference

ϵ = average of residuals, average distance to mean about line

This model will help us predict body fat based on circumference which is cheaper and easier to measure then using a tank of water.

4. Fit your model in #3 to the BodyFat data and summarize the results by displaying estimated coefficients in a table (do NOT just provide a screen shot of the R output). Interpret at least 1 of the coefficients (not the intercept) in the context of the problem.
oefficients:

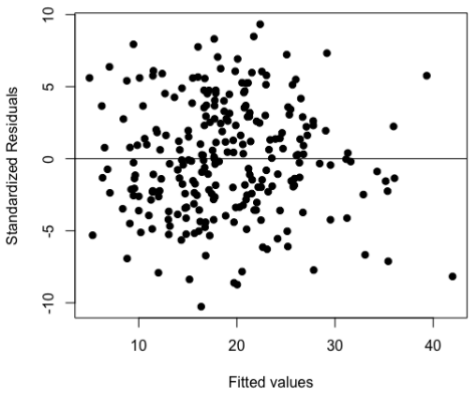
Body Fat = -15.29228 + age * 0.05678 + weight * -0.08031 + Height * -0.06461 + Neck * -0.43754 + Chest * -0.02360 + Abdom * 0.88543 + Hip * -0.19843 + Thigh * 0.23189 + Knee * -0.01166 + Ankle * 0.16355 + Biceps * 0.15280 + Forearm * 0.43050 + Wrist * -1.47657 + 3.996

Holding all other x's constant 0.05678 is the amount that predicted body fat increase on average per unit increase in age.

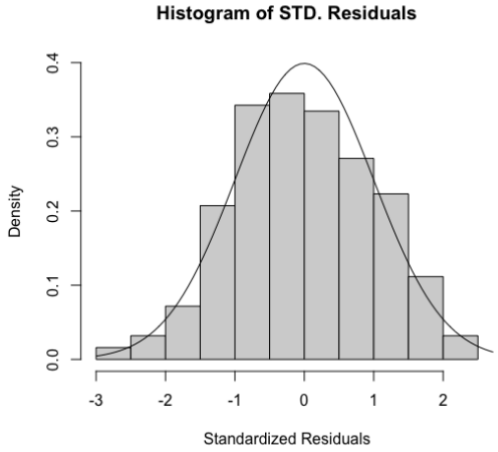
5. List your model assumptions, then justify them using appropriate graphics or summary statistics.

For any data to be suitable to analyze with a MLR model, it needs to pass 4 assumptions. These assumptions are that there is a linear relationship between covariates (circumferences) and response (body fat), that errors are independent of one another and of the covariates (circumferences), that errors are normally distributed, and that there is equal variance for all ϵi.

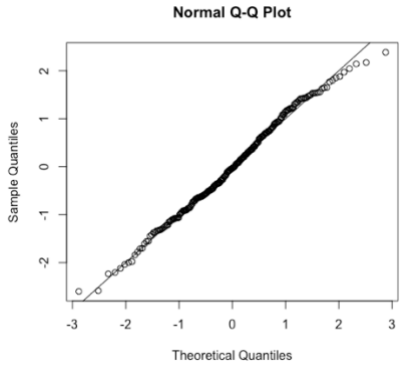
The residuals vs. fitted values scatterplot bellow suggests that the data passes the Linear, Independent and Equal Variance Assumptions.



The histogram of standardized residuals below suggests that the data passes the normality assumption. I propose this is because of the relatively small number of observations, and if more observations were made that histogram would normalize.



The normal quantile-quantile (QQ) plot below also suggests that the data meets the normality assumption.



6. Because prediction is so important, nurses want to know how accurate your predictions are. Carry out an appropriate study that will evaluate the ability of your model to perform predictions. Display appropriate numerical summaries and interpret these summaries on the level of your target audience. Draw conclusions about how "good" your predictions are relative to the original spread of the response variable. Separately, please discuss model fit by reporting and interpreting R2.

For the MLR model above the R-squared value is 0.7464. This means the 74.64% of the variation of body fat is explained by the circumferences.
Cross validation:
mean(bias) = -0.02202383
mean(rpmse) = 3.931307
mean(coverage) = 0.950775
mean(width) = 16.26007
range= (0.0, 45.1)
Standard Deviation = 7.726733

The RPMSE spans hardly any of the range and it is smaller than the standard deviation indicating that the predictive accuracy of the model is “good”.

The bias is also much smaller than the standard deviation confirming that the predictive accuracy of the model is good. Because the coverage is so much smaller than the range, this also indicates the model is good.

7. Nurses wish to make a prediction of percentage of body fat for the following person: age= 50, weight= 203, height= 67, neck= 40.2, chest=114.8, abdom=108.1, , hip=102.5, thigh=61.3, knee= 41.1, ankle= 24.7, biceps= 34.1, forearm= 31, wrist= 18.3. Describe how you would use your fitted model in #4 to

I would use my fitted model by inputting the measurements and interpreting the fitted value output.

With the given measurements the MLR model predicts the percentage of body fat is 31.30359% with lower prediction interval of 23.19192% and an upper prediction interval of 39.41525%. We are 95% confident that with the given measurements the body fat percentage between 23.19192% and 39.41525%.

Appendix A: Source code

```
#This code is derived from examples in class

library(MASS)     ## stdres
library(car)      ## added-variable plots
library(normtest) ## jb.norm.test
library(lmtest)   ## bptest


## Read in Supervisor Data
setwd("~/Desktop/1A School/1A Winter 2021/STAT330/HW4")
super = read.table("fileDownload.txt",header = TRUE)

head(super)
names(super)
n = nrow(super)

## Explore the Data

plot(super,cex = 0.5,pch =20)
pairs(super,cex = 0.8,pch =20)


pairs(super[,c(1,2:4)],cex = 0.8,pch =20)
pairs(super[,c(1,5:7)],cex = 0.8,pch =20)

round(cov(super),2)
round(cor(super),2)

cor(super)[1,]

## Fit a MLR Model

super.lm = lm(brozek ~ age + weight + height + neck +
              chest + abdom + hip + thigh + knee + ankle +
              biceps + forearm + wrist, data=super)
super.lm = lm(brozek ~ .,data=super)

summary(super.lm)

## Check model assumptions
```

```
## Added variable plot to check linearity assumption
###### For the pth variable
#### y-axis - plot residuals of a regression model y ~ all X but x_p
## the part of the response that isn't explained by another covariate
#### x-axis - plot residuals of a regression model x_p ~ all X but x_p
## the part of the pth covariante that isn't explained by another covariate


##each of these should be linear
avPlots(super.lm,pch = 20,cex = 0.8)

plot(super.lm$fitted.values,super.lm$residuals,pch=19, ylab="Standardized Residuals",
xlab="Fitted values")
abline(a=0,b=0)

## equal variance

plot(super.lm$fitted.values,super.lm$residuals,pch=19)
abline(a=0,b=0)

plot(super$neck,super.lm$residuals,pch=19)
abline(a=0,b=0,col = "red")

bptest(super.lm)

## Normality

hist(stdres(super.lm),freq = FALSE, xlab="Standardized Residuals", ylab="Density", ylim=c(0,.4),
main="Histogram of STD. Residuals")
curve(dnorm,from = -3,to = 4,add = TRUE)

qqnorm(stdres(super.lm))
abline(0,1)


jb.norm.test(stdres(super.lm),nrepl = 1e5)
ks.test(stdres(super.lm),"pnorm")


## Predict for a new supervisor where
##Complaints=100,Privileges=0,Learn=100,Raises=100,Critical=100,Advance=0

predict.lm(super.lm,newdata=data.frame(age= 50, weight= 203,
```

```
                          height= 67, neck= 40.2,
                          chest=114.8, abdom=108.1,
                          hip=102.5, thigh=61.3,
                          knee= 41.1, ankle= 24.7,
                          biceps= 34.1, forearm= 31,
                          wrist= 18.3),
                         interval = "prediction",level = 0.95)

apply(super,2,range)

## Perform a a series of cross-validation studies
n.test = 4
n.cv = 1e4

bias = numeric(n.cv)
rpmse = numeric(n.cv)
coverage = numeric(n.cv)
width = numeric(n.cv)


for(i in 1:n.cv){
  test.obs = sample(1:nrow(super),n.test)
  super.test = super[test.obs,]
  super.train = super[-test.obs,]
  train.lm = lm(brozek~.,data=super.train)
  preds = predict.lm(train.lm,newdata=super.test,interval = "prediction",
                level = 0.95)
  bias[i] = mean(preds[,1]-super.test$brozek)
  rpmse[i] = sqrt(mean((preds[,1]-super.test$brozek)^2))
  width[i] = mean(preds[,3]-preds[,2])
  coverage[i] = mean(preds[,2] < super.test$brozek & preds[,3] > super.test$brozek)

}



mean(bias)
mean(rpmse)
mean(coverage)
mean(width)


range(super$brozek)
sd(super$brozek)
```