Homework Analysis 4: Body Fat: Kevin Toney

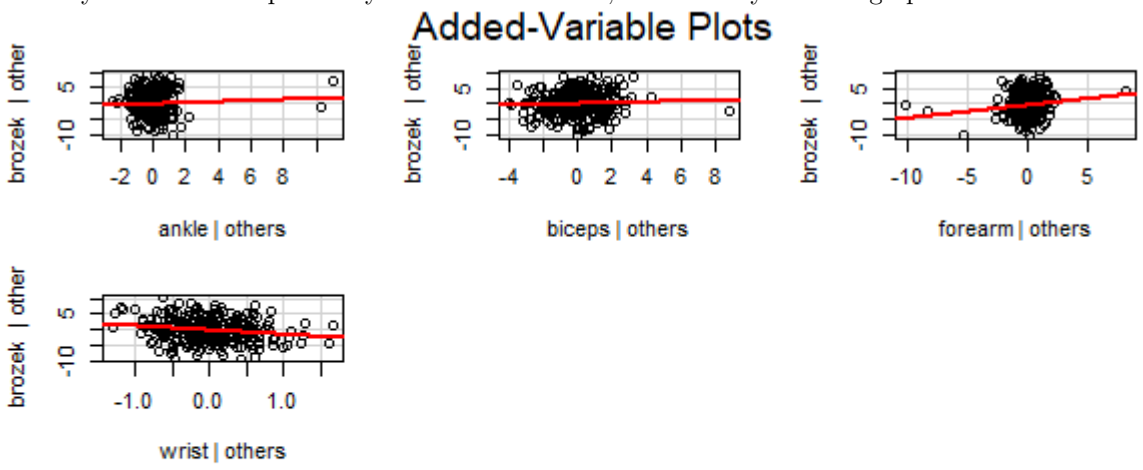
October 25, 2017

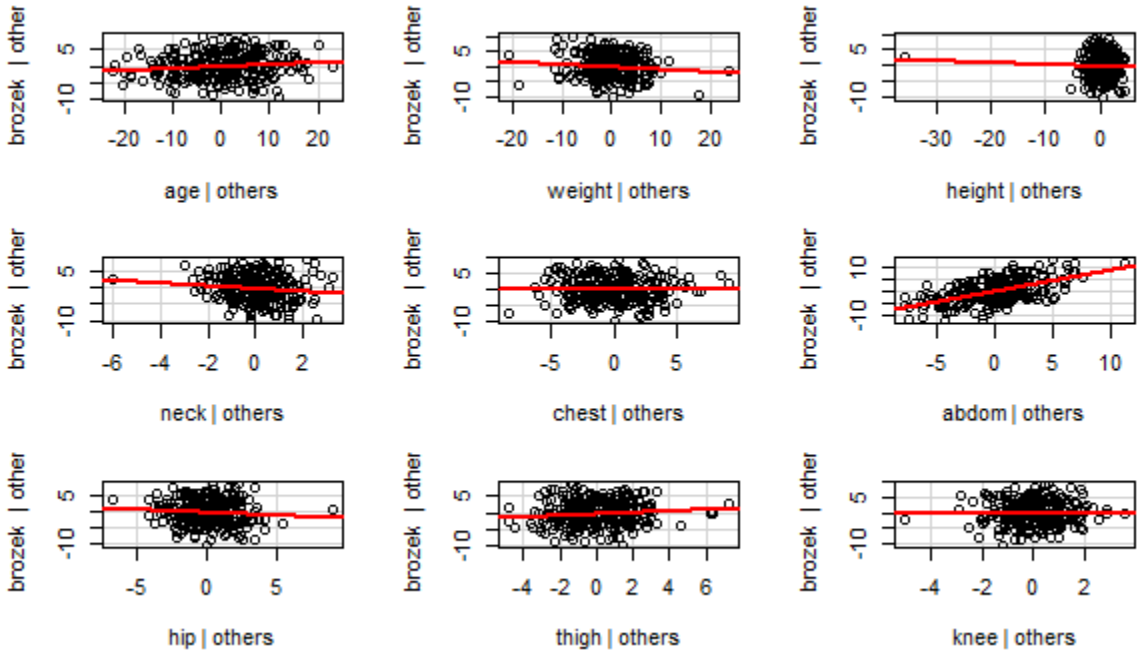
1 Introduction and Problem Background

The overarching purpose of this analysis is develop a simple method for estimating body fat. To help fullfil this purpose, researchers recorded each of the 252 subjects' age, weight (in pounds), height and body part circumferences (there were 10 different circumference measurements). This analysis will help predict a subject's body fat based these variables. In other words, statistical modeling, like linear regression, will enable us to input values, such as age, weight, height and circumferences, into linear equation and get the predicted body fat for the individual. This method will be simpler than using the under-water weighing technique.

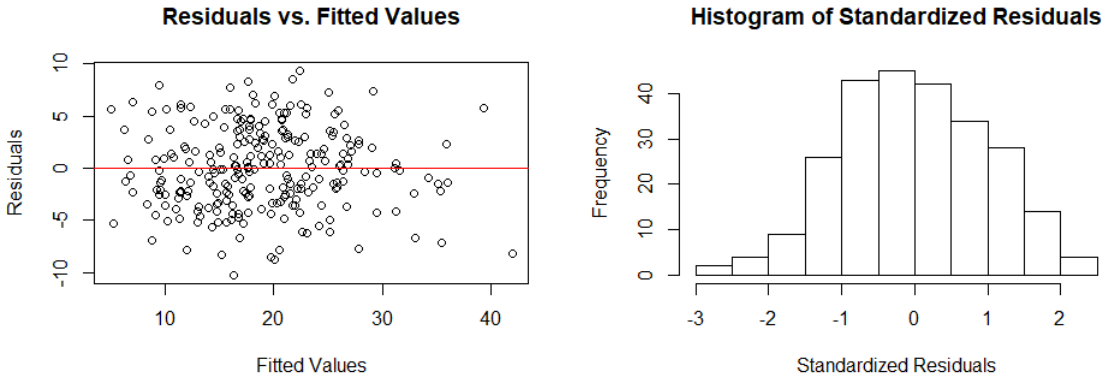
2 Exploratory Data Analysis

Multiple linear regression is suitable to analyze the body fat measurements from the data. The relationships of the body fat and the explanatory variables are linear, as shown by the two graphs below.





Each graph shows the percentage of body fat vs. age, weight, height or circumference. There is a linear relationship in each plot. As a matter of fact, the correlation scores between the body fat and the explanatory variables are all positive. Therefore, we expect that the percentage of body fat to increase as the other variables increase.



Also, four assumptions, which must be justified in order to use multiple linear regression are met. First, I already showed that the relationships between the body fat and explantory variables are linear. Second, the residuals (shown by the histogram above) of body fat follow a normal distribution. Third, the body fat has a constant variance throughout all of the other variables, as per the residuals vs. fitted values plot. Finally, the body fat measurements are independent from each other. These assumptions will be explained more in depth later.

3 Mathematical Model

A justifiable multiple linear regression model, with centering the explanatory variables (age, weight, height and circumferences) around their means is the following:

$y[i] \sim N(\beta_0 + \beta_{age} * (age[i] - \bar{age}[i]) + \beta_{weight} * (weight[i] - \bar{weight}) + \beta_{height} * (height[i] - \bar{height}) + \beta_{neck} * (neck[i] - \bar{neck}) + \beta_{chest} * (chest[i] - \bar{chest}) + \beta_{abdom} * (abdom[i] - \bar{abdom}) + \beta_{hip} * (hip[i] - \bar{hip}) +$

$\beta_{thigh} * (thigh[i] - \bar{thigh}) + \beta_{knee} * (knee[i] - \bar{knee}) + \beta_{ankle} * (ankle[i] - \bar{ankle}) + \beta_{biceps} * (biceps[i] - \bar{biceps}) + \beta_{forearm} * (forearm[i] - \bar{forearm}) + \beta_{wrist} * (wrist[i] - \bar{wrist}), \sigma^2)$

This model will be able to estimate and predict the body fat of a person becuase we will be able to input the age, weight, height and circumference measurements of a person, into the linear equation and get a predicted value. In order to use the multiple linear regression model, we assumed the relationship between the body fat and each of the centered explanatory variables is linear. We also assumed the residuals of the body fat follows a normal distribution. We assumed the variance of the body fat is equal throughout the data, and we assumed the data was collected independently.

For more details, let us consider what some of the parameters mean.

The average body fat of all the participants is β_0 if all the explanatory variables are at the mean.

Holding all the other explanatory variables constant, the average body fat for all the individuals increases by β_{weight} if the weight of the individual increases by one pound in the positive direction from the mean weight.

4 Results

The estimates of the β parameters, which show the relationship the body fat has with the centered explanatory variables is shown by this table.

parameters	estimates
b0	18.889
b_age	0.057
b_weight	-0.08
b_height	-0.065
b_neck	-0.438
b_chest	-0.024
b_abdom	0.885
b_hip	-0.198
b_thigh	0.232
b_knee	-0.012
b_ankle	0.164
b_biceps	0.153
b_forearm	0.43
b_wrist	-1.477

To interpret this table, let us consider the estimates of three parameters.

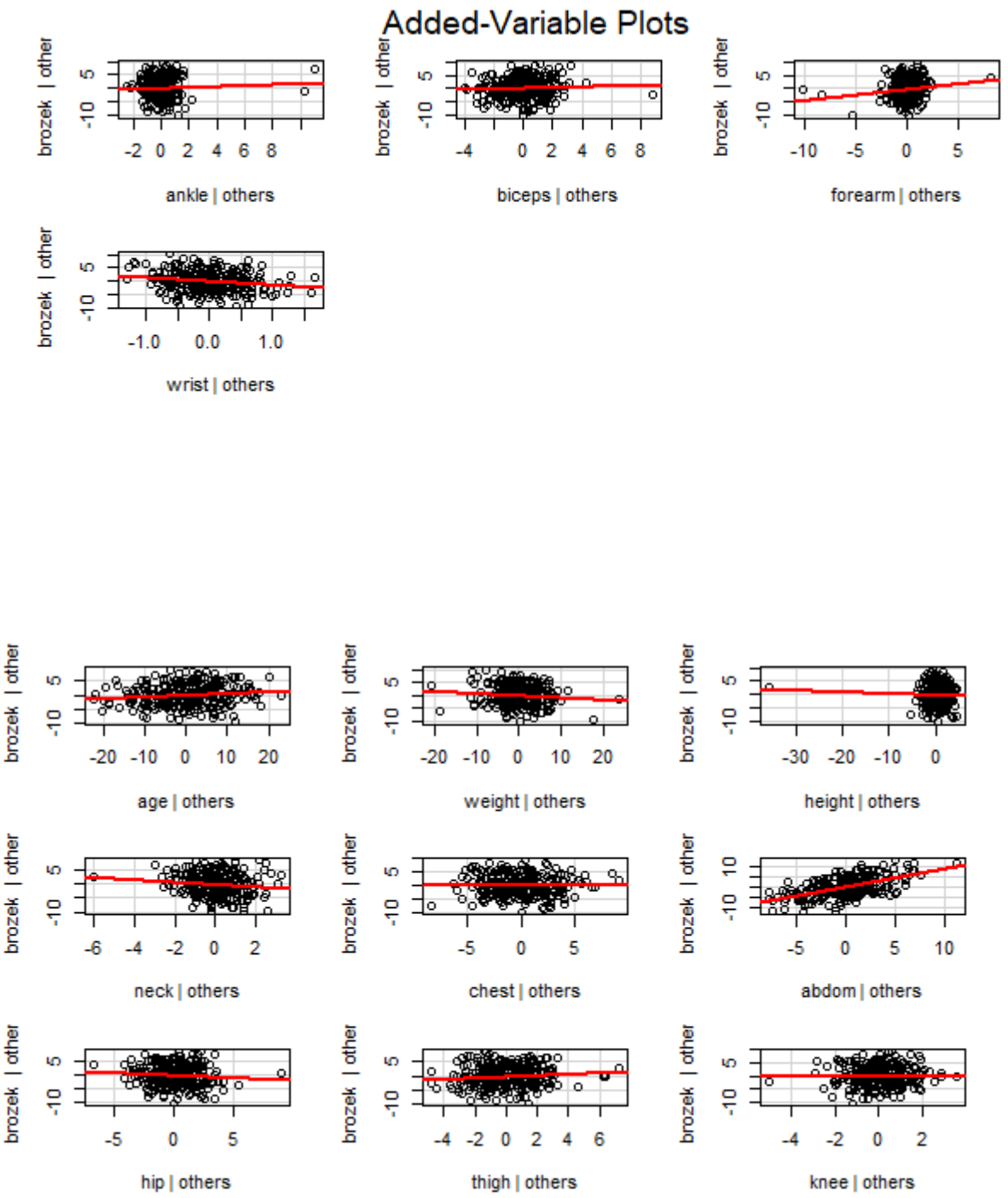
β_0 = 18.889 = the average body fat of all the participants is 18.889 if all the explanatory variables, are at the mean.

β_{weight} = -0.08 = holding all the other centered explanatory variables constant, the average body fat for all the individuals increases by -0.08 if the weight of the individual increases by one pound away from the mean weight.

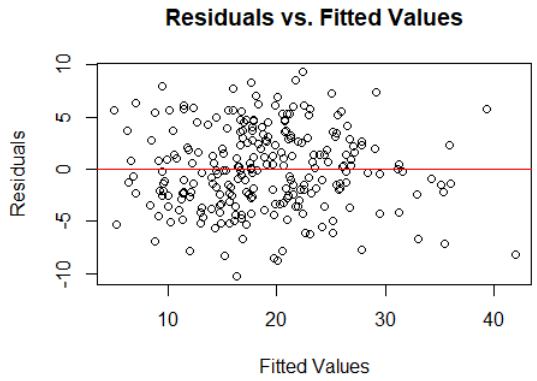
β_{biceps} = 0.153 = holding all the other centered explanatory variables constant, the average body fat for the all the individuals increases by 0.153 if the circumference of the individual's biceps increases by 1 centimeter from the mean bicep circumference.

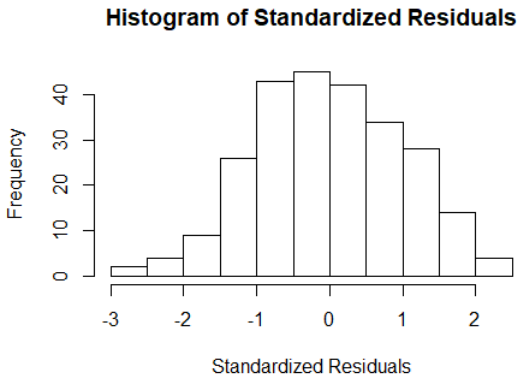
5 Justifying Model Assumptions

The plots below will help justify the assumptions needed in order to use multiple linear regression and answer your questions.



These plots show the relationships of the body fat with the other characteristics of the individuals are linear. The correlation scores for body fat compared to the other variables are all positive.





The residuals vs. fitted values plot doesn't follow a noticeable pattern like a wave or a clump. Therefore, the body fat measurements are independent of each other. The design of the study supports this conclusion. The body fat measurements were on different people. Thus, the measurements couldn't affect each other. They must be independent.

The residuals vs. fitted values shows a constant variance, or a constant difference of the data points from the mean. I performed a BP test to see if there was any evidence the variance wasn't equal in parts of the data. The BP test gave a p-value of 0.1209. There is not enough evidence to suggest the variance isn't constant. Our assumption is met.

Finally, the histogram of the standardized residuals shows a normal distribution. The KS test strengthens this conclusion. The test returned a p-value of 0.822, which is much larger than an alpha value of 0.05. According to the rules of hypothesis testing, we cannot prove the residuals do not follow a normal distribution. Therefore, all of our four assumptions are met.

The R^2 score of our multiple linear regression model is 0.7464. Therefore, 74.64 percent of the variation in the percentage of body fat is explained by the explanatory variables, if they are centered around the mean.

6 Making a Prediction

I used the multiple linear regression model, centered around the explanatory variables' means, to predict the 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. I inputted these values into the linear equation given in section three and got the predicted body fat percentage of the person, which is 65.485. Moreover, I found an interval called the prediction interval. I am 95 percent confident the percentage of body fat for this individual is between 32.88 percent and 98.08 percent.

7 Predictive Accuracy

In order to estimate the accuracy of my predictions, I ran multiple regression on the same model and made predictions 250 times. On average, my predictions overestimated body fat percentages compared to the added variable regression lines. On average, the predictions were 3.866 percent points away from the mean values. Compared to the range of body fat percentenges in the test subjects, which is 45 percentage points, the average prediction error of 3.866 is minimal.

Also, I calculated prediction intervals. Prediction intervals are the range, in which we are 95 percent confident, the body fat percentages of one person will be. These prediction intervals contained the actual body fat measurements 94.8 percent of the time. The average difference between the upper values of the prediction intervals and the lower values was 16.227 percentage points. In other words, the average prediction interval range was 16.227 percentage points.

8 Appendix

```
rm(list=ls())

#########################
#Homework 4: Body Fat
#########################
library(MASS)
library(lmtest)
library(car)

bodyfat <- read.table("C:/Users/kevin/Desktop/Fall 2017/STAT 330/Homework/Homework4/bodyfat

pairs(bodyfat)
cor(bodyfat)

multlin <- lm(brozek ~ ., data=bodyfat)
summary(multlin)

resids <- multlin$residuals
std.res <- stdres(multlin)

#check for normality
hist(std.res, xlab="Standardized Residuals",
     main="Histogram of Standardized Residuals")
ks.test(std.res, "pnorm")
#the residuals follow a normal distribution.

bptest(multlin)
plot(multlin$fitted.values, resids, xlab = "Fitted Values",
     ylab="Residuals",
     main = "Residuals vs. Fitted Values")
abline(0,0, col="red")
#The variance is equal throughout the data.

avPlots(multlin)
#The relationships between the body fat and the other variables are linear.

#The residuals in the residuals vs. fitted values plot do not
#follow any normal patterns, such as a wave or a clump.
#The plot shows the body fat measurements are independent of each other.
#Also, the body fat measurements came from different subjects.
#Therefore, the body fat measurements do not affect each other.

num.pred <- ncol(bodyfat)-1
predictors <- bodyfat[,-1]
for(i in 1:num.pred){

  predictors[,i] <- predictors[,i]-mean(predictors[,i])

}
brozek <- bodyfat$brozek
bodyfat.cent <- cbind(brozek, predictors)
```

```
multlin.cent <- lm(brozek ~ ., data=bodyfat.cent)
summary(multlin.cent)

resids.cent <- multlin.cent$residuals
std.res.cent <- stdres(multlin.cent)

#check for normality
hist(std.res.cent, xlab="Standardized Residuals",
     main="Histogram of Standardized Residuals")
ks.test(std.res.cent, "pnorm")
#the residuals follow a normal distribution.

bptest(multlin.cent)
plot(multlin.cent$fitted.values, resids.cent, xlab = "Fitted Values",
     ylab="Residuals",
     main = "Residuals vs. Fitted Values")
abline(0,0, col="red")
#The variance is equal throughout the data.

avPlots(multlin.cent)


predict.lm(multlin.cent, newdata = data.frame(brozek=1, 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")


n.cv <- 250
bias <- rep(NA, n.cv)
rpmse <- rep(NA, n.cv)
pred.int.width <- rep(NA, n.cv)
coverage <- rep(NA, n.cv)

for(i in 1:n.cv) {
  ## Step 1: split data into test and training sets
  adv.test <- sample(1:nrow(bodyfat), 3)
  test.data <- bodyfat[adv.test,]
  train.data <- bodyfat[-adv.test,]

  ## Step 2: Fit model to training data
  my.model <- lm(brozek ~ ., data = train.data)
  #if I am using a predict.lm statement, I need to fit the model
  #using the column names of the data frame, not variables
  #I created before.
  ## Step 3: predict for test data
  test.preds <- predict.lm(my.model, newdata = test.data)
  ## Step 4: calculate the bias and RPMSE
  bias[i] <- mean((test.preds - test.data$brozek))
  rpmse[i] <- sqrt(mean((test.preds - test.data$brozek)^2))

  pred.int <- predict.lm(my.model, newdata= test.data, interval = "prediction", level=0.95)
  #do my prediction intervals contain the five data points from the test data
  coverage[i] <- pred.int[,2] < test.data$brozek & test.data$brozek < pred.int[,3]
```

```
  pred.int.width[i] <- mean(pred.int[,3] - pred.int[,2])
  #get the average of the differences between the upper and the lower
  #that number equals the prediction interval width.
}

mean(coverage)
mean(pred.int.width)
mean(bias)
mean(rpmse)
```