Homework 4

Multiple Linear Regression

AUTHOR

Zeb Sorenson

Data and Description

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 (years), weight (pounds), height (inches), and three body circumference measurements (around the neck, chest, and abdominal (all in centimeters)) for 252 men. Each man's percentage of body fat was accurately estimated by an underwater weighing technique (the variable "brozek" is the percentage of body fat). The hope is to be able to use these data to create a model that will accurately predict body fat percentage using just the basic variables recorded, without having to use the tank submerging method.

The data can be found in the BodyFat data set on Canvas. Download "BodyFat.txt", and put it in the same folder as this R Markdown file.

0. Replace the text "< PUT YOUR NAME HERE >" (above next to "author:") with your full name.

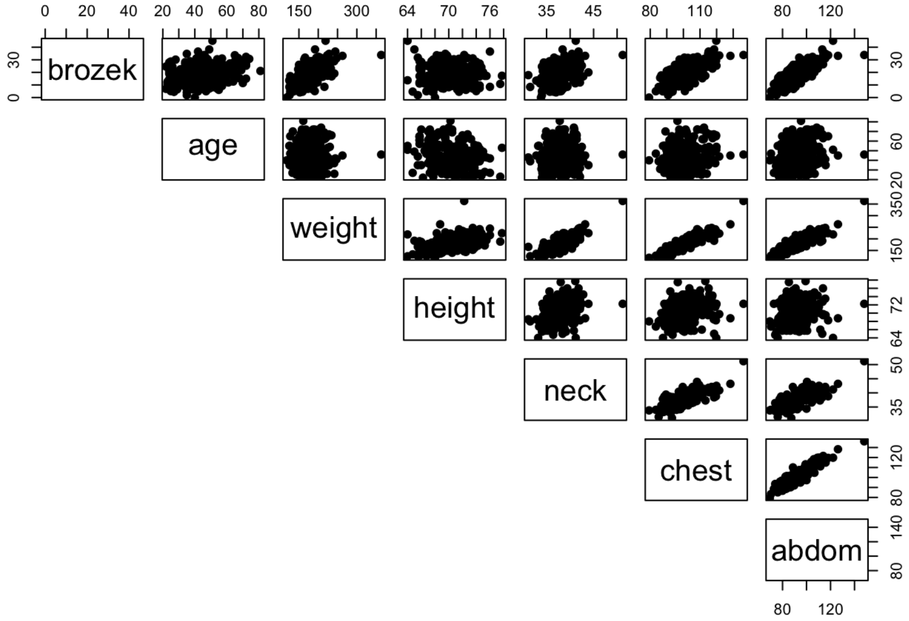
1. Read in the data set, and call the data frame "bodyfat". Print a summary of the data. **Remove the "row" column (which contains row numbers) from the data set.**

```r
orginalData <- read.csv("~/Desktop/Stat 330/BodyFat.txt", sep="")

bodyfat <- subset(orginalData, select = -row)

head(bodyfat)
```

```
  brozek age weight height neck chest abdom
1   12.6  23 154.25  67.75 36.2  93.1  85.2
2    6.9  22 173.25  72.25 38.5  93.6  83.0
3   24.6  22 154.00  66.25 34.0  95.8  87.9
4   10.9  26 184.75  72.25 37.4 101.8  86.4
5   27.8  24 184.25  71.25 34.4  97.3 100.0
6   20.6  24 210.25  74.75 39.0 104.5  94.4
```

2. Create and print a scatterplot matrix of the data.

```r
pairs(bodyfat, pch = 19, lower.panel = NULL) # this omits the duplicated half
```



```
#modified from in class 4
```

3. Based on the scatterplot matrix, briefly explain which variables you think will be "significant" for predicting brozek and which variables you think will *not* be helpful at predicting brozek. Explain how the scatterplot helped determine your answers.

Based on the scatter plot matrix, there are a number of variables that appear to possibly serve us well (before doing further investigation). We are looking for the variables that appear to have a strong positive linear correlation. The ones we can see from the plot are, chest, abdomen, and neck appears to be useful when pared with these previous three.

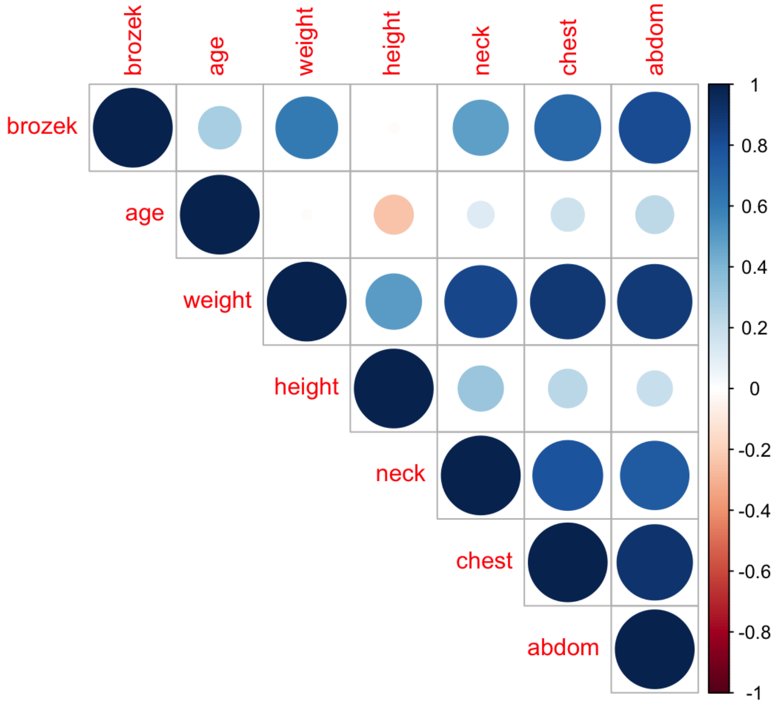
The ones that appear to not have a clear trend are height and age.

4. Create and print a correlation matrix (numeric or color- and shape-coded).

```
round(cor(bodyfat), 2)
```

```
       brozek   age weight height neck chest abdom
brozek   1.00  0.29   0.61  -0.02 0.49  0.70  0.81
age      0.29  1.00  -0.01  -0.24 0.11  0.17  0.23
weight   0.61 -0.01   1.00   0.49 0.83  0.89  0.89
height  -0.02 -0.24   0.49   1.00 0.33  0.24  0.20
neck     0.49  0.11   0.83   0.33 1.00  0.78  0.75
chest    0.70  0.17   0.89   0.24 0.78  1.00  0.92
abdom    0.81  0.23   0.89   0.20 0.75  0.92  1.00
```

```
corrplot(cor(bodyfat), type = "upper")
```



5. Based on the scatterplot matrix and the correlation matrix, are their any pairs of variables that you suspect will cause a problem in terms of multicollinearity? If so, which ones?

The one that jumps out immediately is age with height as it arrives in the negative correlation. Along with age with weight (This one is almost zero) and age with height, as both of these pairs are extremely close to one. The biggest concern would be age and weight as it is almost zero and therefore possibly pulling on the other variables in such a way that negativily affects our model.

6. Fit a multiple linear regression model to the data (no transformations). Print a summary of the results. Save the residuals to the `bodyfat` data frame.

```
bodyFat_lm <- lm(brozek ~ age+weight+height+neck+chest+abdom, data = bodyfat) #Add each one indi
summary(bodyFat_lm)
```

```
Call:
lm(formula = brozek ~ age + weight + height + neck + chest +
    abdom, data = bodyfat)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5811  -3.0358   0.0668   2.8879  10.2828

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.010e+01  1.413e+01  -1.423   0.1561
age          5.010e-03  2.540e-02   0.197   0.8438
weight      -8.733e-02  3.879e-02  -2.251   0.0253 *
height      -1.400e-01  1.520e-01  -0.921   0.3579
neck        -4.421e-01  2.019e-01  -2.189   0.0295 *
chest        4.844e-04  9.107e-02   0.005   0.9958
abdom        8.754e-01  7.924e-02  11.048   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.124 on 244 degrees of freedom
Multiple R-squared:  0.722, Adjusted R-squared:  0.7151
F-statistic: 105.6 on 6 and 244 DF,  p-value: < 2.2e-16
```

```
bodyfat$residuals <- bodyFat_lm$residuals
```

```
bodyfat$fits <- bodyFat_lm$fitted.values
```

7. Briefly comment on the “significance” of the variables: were you surprised by the results? Are there any variables that are significant that you think shouldn’t be? Are there any variables that are not significant that you think should be?

We have a large number of high P values in our summary. However, it appears that weight, neck and abdom appear to be linearly related to average brozek score for body fat. Weight and abdom make sense on a surface level analysis of a body fat score, however, I am curious to see if we do more analysis on our neck variable which intuitively does not jump out as important. I was also expecting height to have a much lower (significant) P value, so its 0.3579 is surprising to me age’s 0.8438 P value is also higher than something I would have predicted.

8. Briefly comment on the sign (+/-) of the coefficients for the variables. Are their any variables where the sign is the opposite of what you expected?

I was extremely concerned about the negative weight coefficient of -0.08733 until further discussion with the TA’s and Dr. Sandholtz. Looking at this one, holding all else constant, when weight goes up by 1 pound, the average body fat percentage decreases by 0.08733. Which doesn’t make sense at first, but when we hold all else constant, this is almost certainly muscle being added, which will increase a person’s weight. The other negative coefficients make sense, height may be a bit strange but the value is quite small.

The positive coefficients also do not seem concerning. It makes sense that our average chest would increase (muscle again) I would think that average abdomen would decrease but I am sure this can be explained by the type of people being sampled (overweight vs skinny)

Overall, nothing too out of the ordinary with these coefficients.

9. Mathematically write out the *fitted* multiple linear regression model for this data set using the coefficients you found above (do not use betas). Do not use “X” and “Y” in your model - use variable names that are fairly descriptive.

Without doing any type of transformations, our fitted model looks like this.

$$\widehat{\text{BodyFat}}_i = -20.1 + 0.00501 \cdot (\text{Age}_i) - 0.08733 \cdot (\text{Weight}_i) - 0.14 \cdot (\text{Height}_i) - 0.4421 \cdot (\text{Neck}_i) + 0.0004844 \cdot (\text{Chest}_i) + 0.8$$

10. *Assuming* the model assumptions are all met, how would you interpret the coefficient for Weight?

Holding all else constant, when weight goes up by 1 pound, the average body fat percentage decreases by .087

11. Briefly explain what it means to “hold all else constant” when interpreting a coefficient.

It often doesn’t make literal sense when we make our interpretations. It’s almost inconceivable that body fat percentage and weight could change while variables like chest, abdomen and age would not change.

However, we can take the approach of looking at persons that more or less meet the given criteria. For example, among people around the same age that have the same height, neck, chest and abdomen, for every increase of weight of 1 pound, their body fat percentage will decrease by 0.087

12. Briefly explain what the F-test indicates, as given in the model output from question 6.

This gives us an overall test of significance of whether or not at least one of the predictor variables is is significantly linearly correlated with the response (Brozek/body fat percentage).

We get a P value of less than 2.2e-16 meaning that at least one of the variables is linearly associated with the average body fat percentage.

13. Briefly interpret the *adjusted* R-squared, as reported in the model output from question 6.

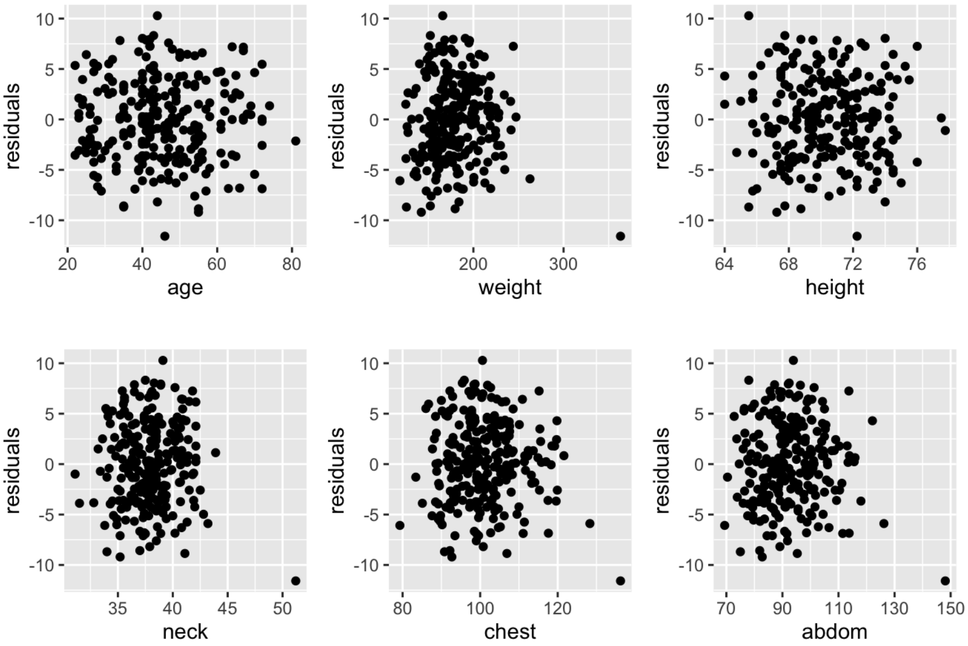
About 71% of the total variation in the brozek body fat percentage score is explained by the predictors in the model, after accounting for the number of variables in the model (This was taken direly from the slides).

Questions 14-20 involve using diagnostics to determine if the linear regression assumptions are met. For each assumption, (1) perform

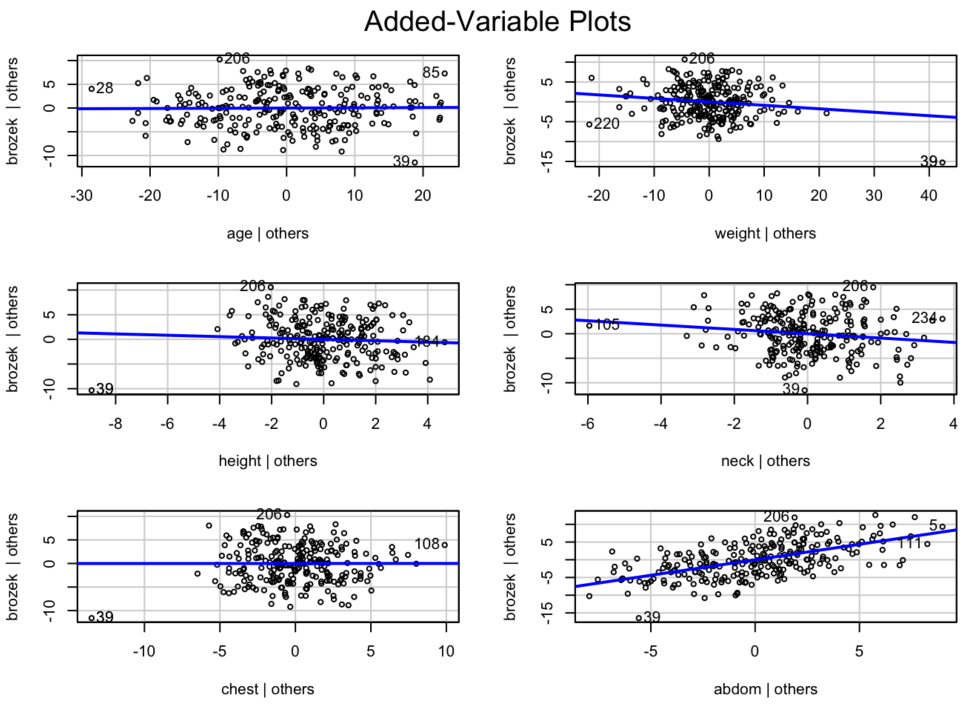
appropriate diagnostics to determine if the assumption is violated, and (2) explain whether or not you think the assumption is violated and why you think that.

14. The X's vs Y are linear (use the residual vs. predictor plots, partial regression plots, and the residual vs. fitted values plot).

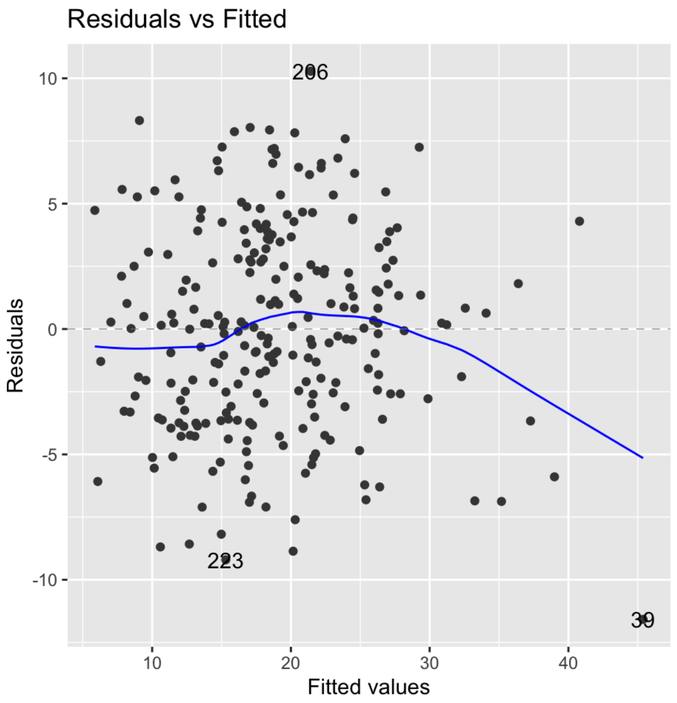
```
# residual vs. predictor plots

resid_vs_age <- ggplot(data = bodyfat) +
  geom_point(mapping = aes(x = age, y = residuals)) +
  theme(aspect.ratio = 1)

resid_vs_weight <- ggplot(data = bodyfat) +
  geom_point(mapping = aes(x = weight, y = residuals)) +
  theme(aspect.ratio = 1)

resid_vs_height <- ggplot(data = bodyfat) +
  geom_point(mapping = aes(x = height, y = residuals)) +
  theme(aspect.ratio = 1)

resid_vs_neck<- ggplot(data = bodyfat) +
  geom_point(mapping = aes(x = neck, y = residuals)) +
  theme(aspect.ratio = 1)

resid_vs_chest <- ggplot(data = bodyfat) +
  geom_point(mapping = aes(x = chest, y = residuals)) +
  theme(aspect.ratio = 1)

resid_vs_abdom <- ggplot(data = bodyfat) +
  geom_point(mapping = aes(x = abdom, y = residuals)) +
  theme(aspect.ratio = 1)

# put plots in 2 rows & 3 columns using the patchwork package
(resid_vs_age | resid_vs_weight | resid_vs_height) /
  (resid_vs_neck | resid_vs_chest | resid_vs_abdom)
```



```
# partial regression plots

avPlots(bodyFat_lm)
```



```
# residual vs fitted values

autoplot(bodyFat_lm, which = 1, ncol = 1, nrow = 1) +
  theme(aspect.ratio = 1)
```



Linearity is looking great!

With our residual vs predictor plots, we do not see any striking trend for any of the variables. There are no curves at any of our x values.

For our partial regression plots, we have mostly flat lines. We do some some curves in neck and abdom, but nothing too concerning. For the most part, straight lines.

In our Residuals vs. Fitted, again, we are mostly flat around zero. We do see some curving towards the larger x values, but it appears that this is being caused by a single point. Which we will investigate further into our analysis. Because of this appearing to be caused by one point, I am not too concerned with this curve towards the end.

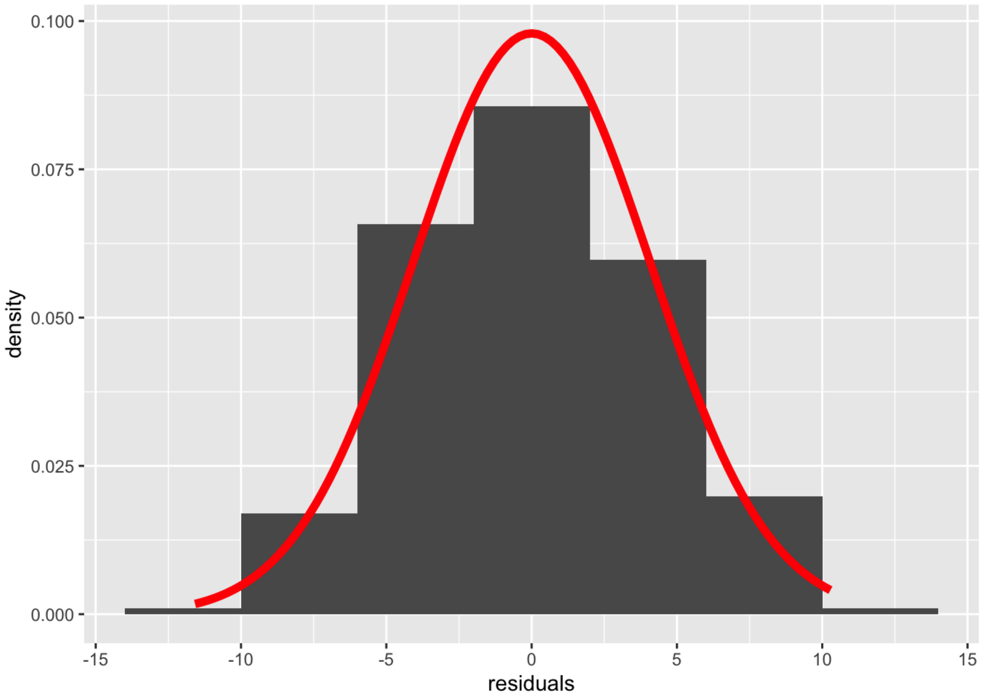
Considering these 3 graphs, it is safe to assume linearity is met.

15. The residuals are independent (I will answer this one for you, no need to modify response. No points for this question).

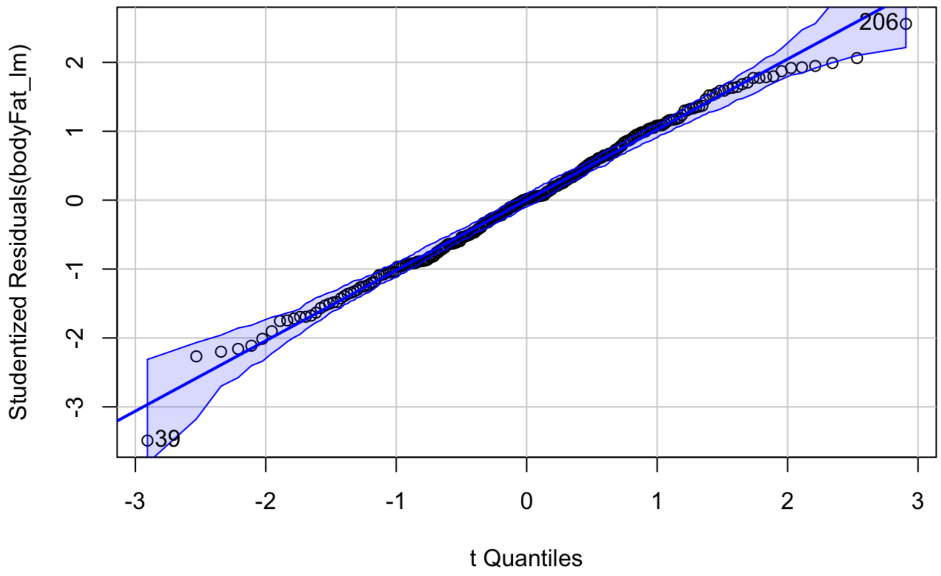
Since we do not know if the 252 men were randomly sampled from a population, we do not know if the residuals are independent or not. We will assume that they are independent for this analysis.

16. The residuals are normally distributed (use a histogram, qq-plot, and Shapiro-Wilk test)

```
# Diagnostic 1 Histogram


ggplot(data = bodyfat) +
  geom_histogram(aes(x = residuals, y = after_stat(density)),
                 binwidth = 4) +
  stat_function(fun = dnorm, color = "red", linewidth = 2,
                args = list(mean = mean(bodyfat$residuals),
                            sd = sd(bodyfat$residuals)))
```



```
# Diagnostic 2 qq plot


qqPlot(bodyFat_lm)
```



```
[1]  39 206
```

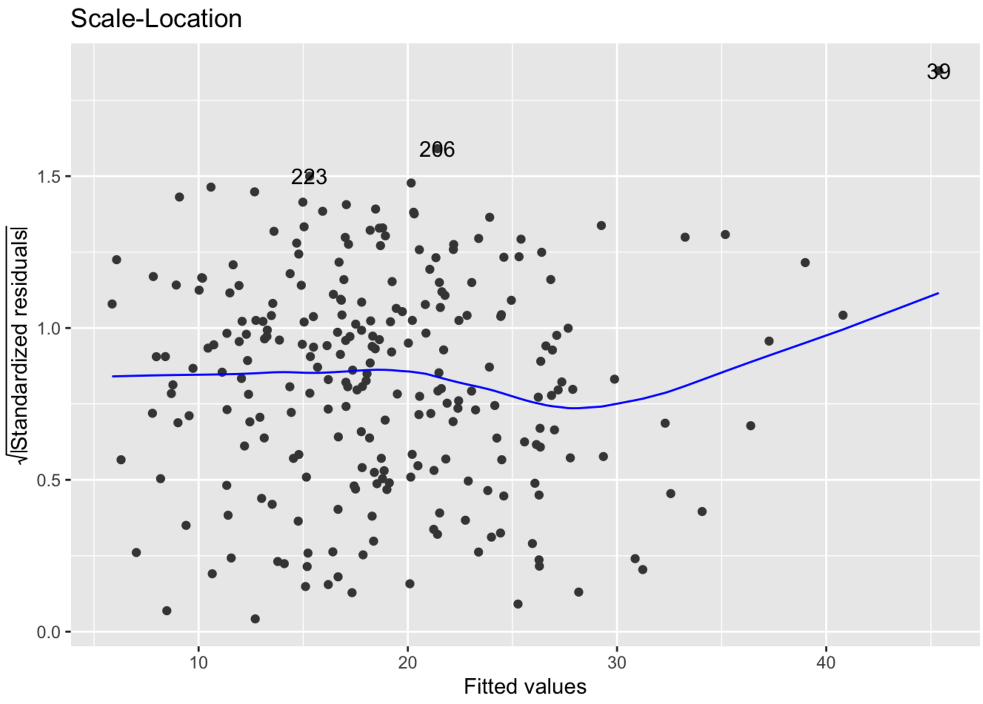
```
# Diagnostic 3 shapiro Wilk


shapiro.test(bodyfat$residuals)
```

```

	Shapiro-Wilk normality test

data:  bodyfat$residuals
W = 0.99416, p-value = 0.4432
```

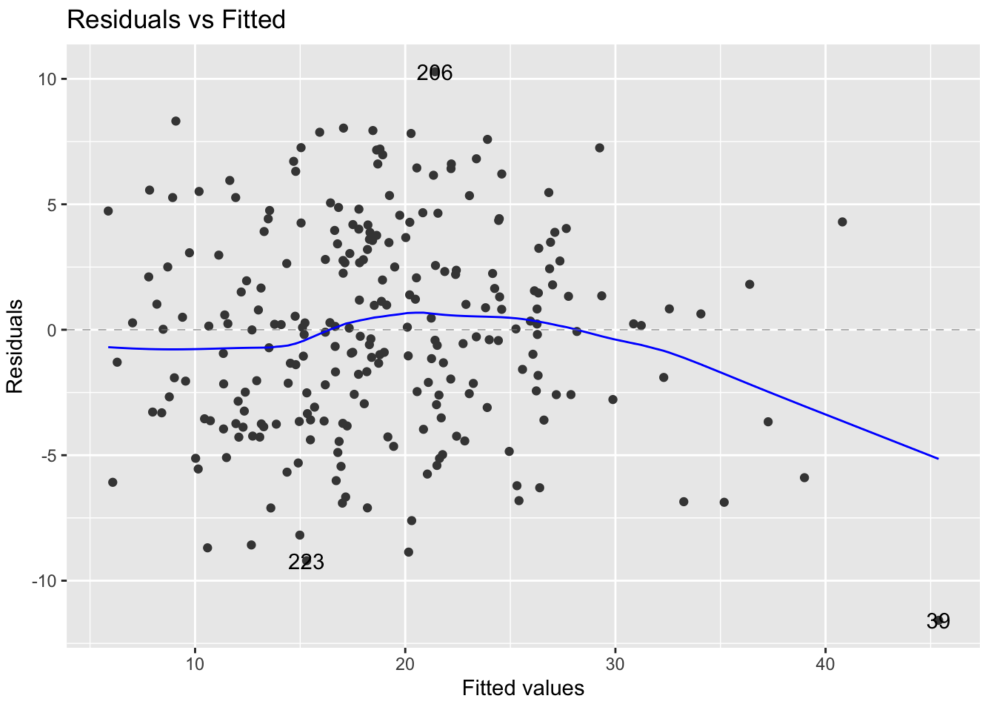
Normally distributed residuals are also looking great! With the the histogram and qq plot as we can see a nice normal curve and our value stay within the boundaries in the qq plot. Finally, our Shaprio Wilk test gives us a large P value, further confirming that the residuals are indeed normally distributed and that this assumption is met.

17. The residuals have equal/constant variance across all values of X (check Scale-location plot and residuals vs. fitted values plot)

```
#Scale location
autoplot(bodyFat_lm, which = 3, nrow = 1, ncol = 1)
```



```
# residuals vs fitted values

autoplot(bodyFat_lm, which = 1, nrow = 1, ncol = 1)
```



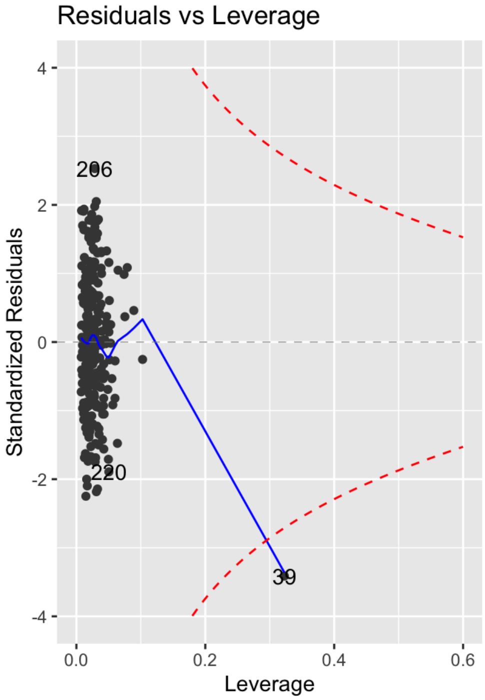
Here we have a consistent spread in both of plots without any blaring patterns. Our line mostly stays straight except for towards the end, which appears again to be caused by a single point which will be further investigated.

Constant variance is satisfied.

18. Check for no influential points using Cook's distance. Also, in your response, reference the plots you created in previous questions.

```
# Cook's Distance

cd_cont_pos <- function(leverage, level, model) {sqrt(level*length(coef(model))*(1-leverage)/leve
cd_cont_neg <- function(leverage, level, model) {-cd_cont_pos(leverage, level, model)}
```

```
cd_threshold <- 0.5
autoplot(bodyFat_lm, which = 5) +
  stat_function(fun = cd_cont_pos,
                args = list(level = cd_threshold, model = bodyFat_lm),
                xlim = c(0, 0.6), lty = 2, colour = "red") +
  stat_function(fun = cd_cont_neg,
                args = list(level = cd_threshold, model = bodyFat_lm),
                xlim = c(0, 0.6), lty = 2, colour = "red") +
  scale_y_continuous(limits = c(-4, 4))
```



Here we see the confirmation of what we have been suspecting for some time now. There is indeed a point (39) which violates cook's distance which will need to be addressed (I'm guessing in the next homework). We saw this point in question 17, as it pulled on both of our graphs causing the most curvature. As well as the residual vs fitted values in question 14. Again, causing the curve we saw in our graph.

This is definitely an issue we will want to look into.

19. Check for extreme multicollinearity. For this (tacit) model assumption, compute the variance inflation factors (VIFs) and compare the VIFs to your response in question 5. Is there agreement? Is this assumption met (recall: the rule of thumb is that each VIF should be less than 10 and the average of the VIFs should be close to 1)?

```
new_bodyFat_lm <-lm(brozek ~ age + height+ neck + chest + abdom, data = bodyfat)

#VIFS

fat_vifs <- vif(bodyFat_lm)
fat_vifs
```

```
      age    weight    height      neck     chest     abdom
 1.510958 19.128777  2.306983  3.545106  8.604128 10.684888
```

```
max(fat_vifs)
```

```
[1] 19.12878
```

```
mean(fat_vifs)
```

```
[1] 7.63014
```

```
#What are the rules of thumb for the VIFS again?
```

It appears that I was correct in being concerned with weight, as it has the largest VIF of 19.13. The rest look to meet our standards, except for abdom which is slightly above 10.

Unfortunately our average is much higher than 1. Sitting at 7.63, this is something I'm sure we will look into for our next analysis.

Note: your next homework assigment will use this same data set, and you will be asked to fix the assumptions that were broken.

20. Briefly summarize what you learned, personally, from this analysis about the statistics, model fitting process, etc.

Again, this class is doing a great job of teaching us the importance of investigating our data before jumping into analysis which I believe is crucial to real world work and any type of situation that involves making decisions based on numbers. It is interesting to see what we have learned so far this semester and apply it to models that have multiple variables. It opens the door to a much wider range of the type of regression analysis we should be able to do in the future, which I think is quite cool.

21. Briefly summarize what you learned from this analysis *to a non-statistician*. Write a few sentences about (1) the purpose of this data set and analysis and (2) what you learned about this data set from your analysis. Write your response as if you were addressing a business manager (avoid using statistics jargon) and just provide the main take-aways.

Measuring body fat is a complicated, time consuming task. We are hoping to create a method that can take measurements such as a person's height, weight, age, and other measurements of the neck, chest and abdomen that can use these measurements to predict a person's body fat percentage.

This will help save time and money as we should be able to avoid traditional methods of estimating body fat.

After running an analysis on the data, it is promising that we can create such a method to predict body fat percentage with these data points. However, there were a few minor issues that we would like to investigate before confidently using the method But for now, it is safe to say that we have a rough method that can make a good estimate that we are going to keep working on.