Final Project

Kat Moon, David Peng

1. Introduction

Nutrition and health are two of the most important aspects of our lives that we have to continuously take care of. However, we often neglect how to lead a healthy life either by having an unbalanced diet, fixating on extreme dieting, or not exercising enough. This becomes an even bigger issue when students leave home and enter college and they are no longer under parents' guidance. In response to this, we decided to investigate data on students' opinions and habits on food, eating, and health by analyzing the data "Food Choices: Food Choices and Preferences of College Students". The survey includes 126 responses from students at Mercyhurst University to more than 50 questions. In particular, we will explore how healthy students feel about themselves and how they perceive their weight, and analyze these by genders, actual weights, among other variables.

Acknowledgements: Thank you to Professor Reuning-Scherer, all of S&DS 230e's wonderful TF's and ULA's, all the students of Mercyhurst University who agreed to participate in the survey, and Kaggle user BoraPajo, who compiled and published the survey.

2. Data

List of Variables

Here are the variables we use throughout the document. Some are original, some recoded, and some new.

- `weight` : weight in pounds; continuous
- `GPA` : unrounded GPA; continuous
- `self_perception_weight` : self perception of weight; 1-slim, 5 - overweight; continuous
- `income` : income brackets; 1 - < $30,000, 5 - > $100,000; continuous
- `Gender` : gender; 'F' - Female, 'M' - Male; categorical (2)
- `employment` : employed (part or full) or unemployed; categorical (2)
- `exercise` : how often you exercise in a typical week; 1 - every day, 5 - never; continuous
- `parent_education` : composite score of parents education; 2 - both have less than high school, 10 - both have graduate degrees; continuous
  - made from similar variables `mother_education` and `father_education` - scale
- `Att_Variety` : composite score of likelihood of eating variety foods when available; with 9 - the most unlikely and 30 being the most likely - continuous
  - made from similar variables `thai_food, persian_food, greek_food, ethnic_food indian_food, italian_food` - scale
- `associate_food` : composite score of choosing the healthier food from two options; 4 - least healthy, 8 - most healthy; continuous
- `healthy_feeling` : agreement with the statement "I feel healthy!"; 1 - strongly agree, 10 - strongly disagree; continuous
- `eating_out` : frequency of eating out in a typical week; 1- never, 5- every day; continuous
- `calories_day` : importance of tracking calories per day; 1 - not knowing important, 4- very important; continuous
- `nutritional_check` : frequency of checking nutritional values; 1 - never, 5 - everything; continuous

- `cook` : frequency of cooking; 1 - every day, 5 - never; continuous
- `fruit_day` : likelihood of eating fruit in a regular day; 1 - very unlikely, 5 - very likely; continuous
- `veggies_day` : likelihood of eating veggies in a regular day; 1 - very unlikely, 5 - very likely; continuous

3. Data Cleaning

Get Raw Data

We start by reading the csv from our local directory and taking a look at the dimension. We won't print all the variables because there are 61.

```
## [1] 125  61
```

Choosing Variables

The amount of variables in the original dataset *is too damn high!* So we spent an hour going through variables one by one, sorting them into 1) definitely use, 2) maybe use, 3) probably don't use, and 4) don't use. We also somewhat labeled ones we thought we could visualize in histograms or use in plots, but this was in vain since we needed to actually graph data and attempt models in order to see what the connections actually were. We tried data from our "definitely use" pile first, then worked downwards.

One significant problem was that many variables were categorical or ordered but on a small scale. For ordered categorical ones, we grouped similar questions that could be combined into a score. For example, questions about how likely someone was to eat ____ food with a max of 5 became an `Att_Variety` composite score with a max of 30. We created three composite variables that we hoped would become more continuous and slightly more normally distributed.

Cleaning and Creating

We start with some clean up from characters to numbers. `GPA` and `weight` have some extra characters cleaned up before turning it into numbers. For `self_perception_weight` , the single nonanswer response to the scale question is turned into `NA` .

To help later, we'll also recode some variables. Income is recoded so there is one less bracket. Gender is changed to readable "F", "M". Employment is recoded to be binary: employed or unemployed.

We'll now create some new variables, including some composite scores. `Att_Variety` adds together several variables about willingness to eat variety foods, `parent_education` adds father's and mother's education scores, and `associate_food` adds healthy food choices between two options, coded in the same direction.

Finally, we make our data set `fc` that includes just the variables indicated in our list. There are now 17 variables. We'll also save an `fc_old` to use when comparing composite scores to their components.

```
## [1] 125  17
```

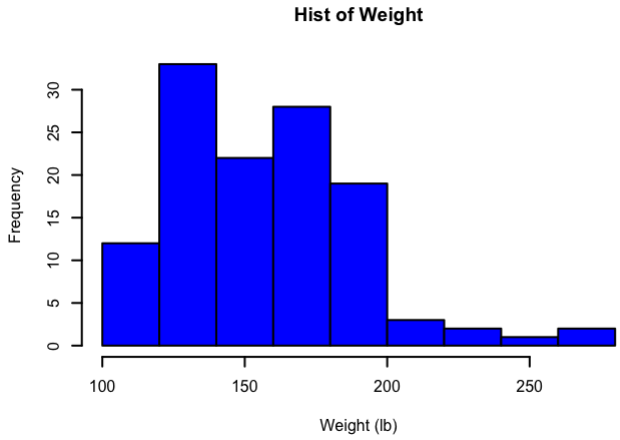
4. Data Exploration

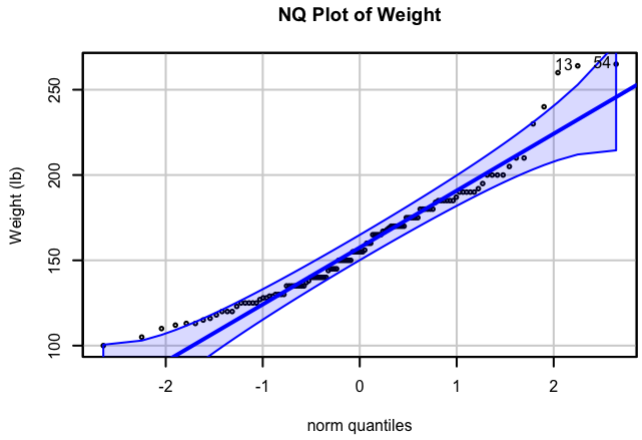
Hists of Continuous Variables

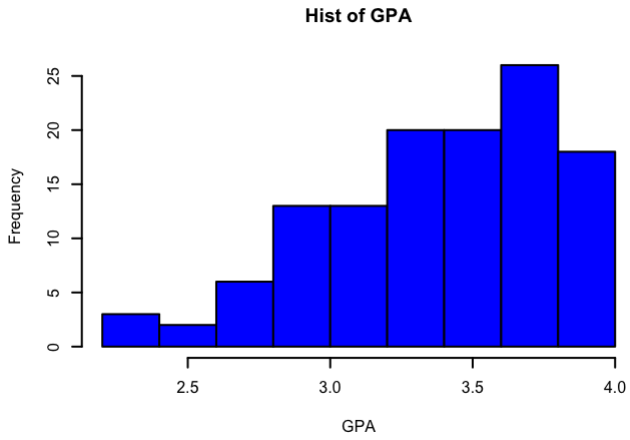
Let's visualize the variables we have! For our continuous. We'll check their distributions and NQ plots.

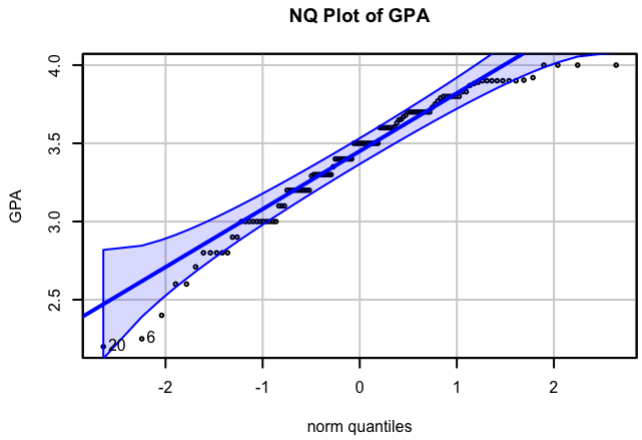
Note: we will treat scaled variables as continuous, and react later if it doesn't work well.

```
## [1] 54 13
```







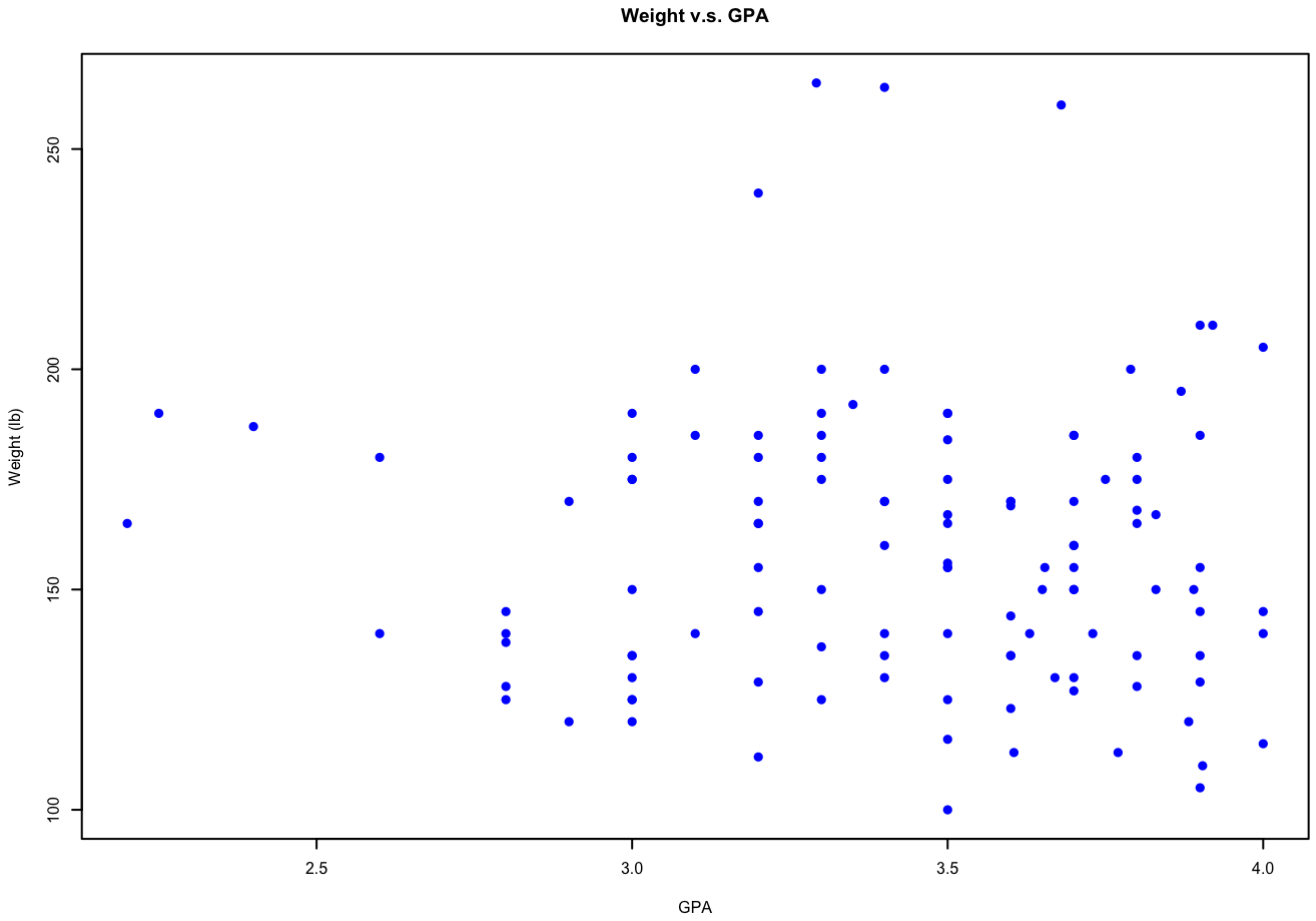



```
## [1] 20  6
```

Weight seems slightly right skewed; GPA slightly left skewed. Their NQ plots are all approximately linear, although the tails don't quite fit.

Scatterplot

Why not try graphing aginst each other the two variables that are most continuous from this dataset: `weight` and `GPA` ?

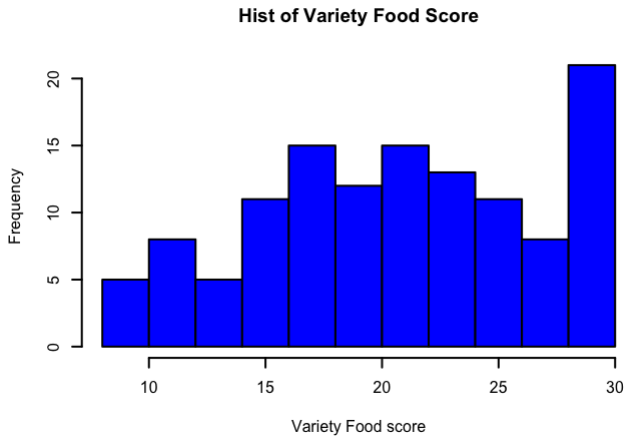


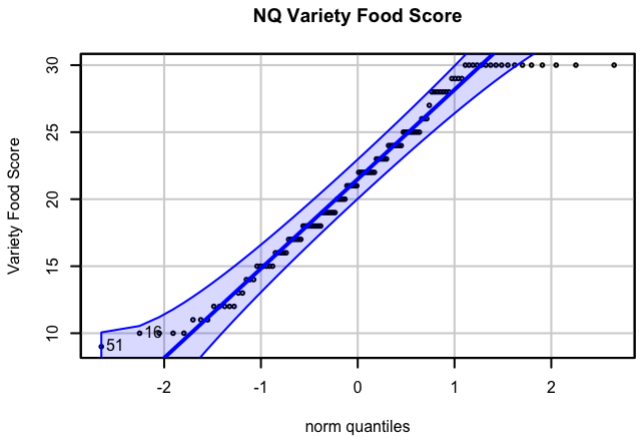
Unfortunately, the correlation is a weak value of -0.04, and the p-value of 0.66 is larger than 0.05, so the correlation is not statistically significantly different from 0.

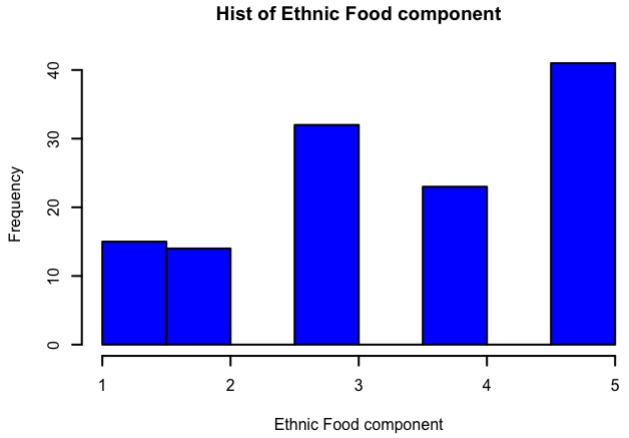
Hists of Composite vs Component Scores

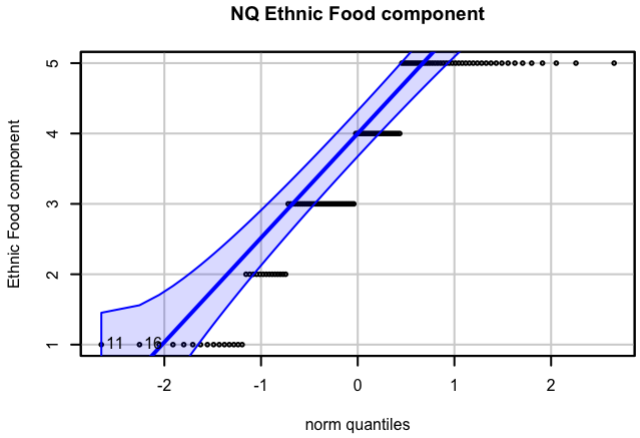
Let's also see what composite scores look like compared to some of their component questions used to create them.

```
## [1] 51 16
```









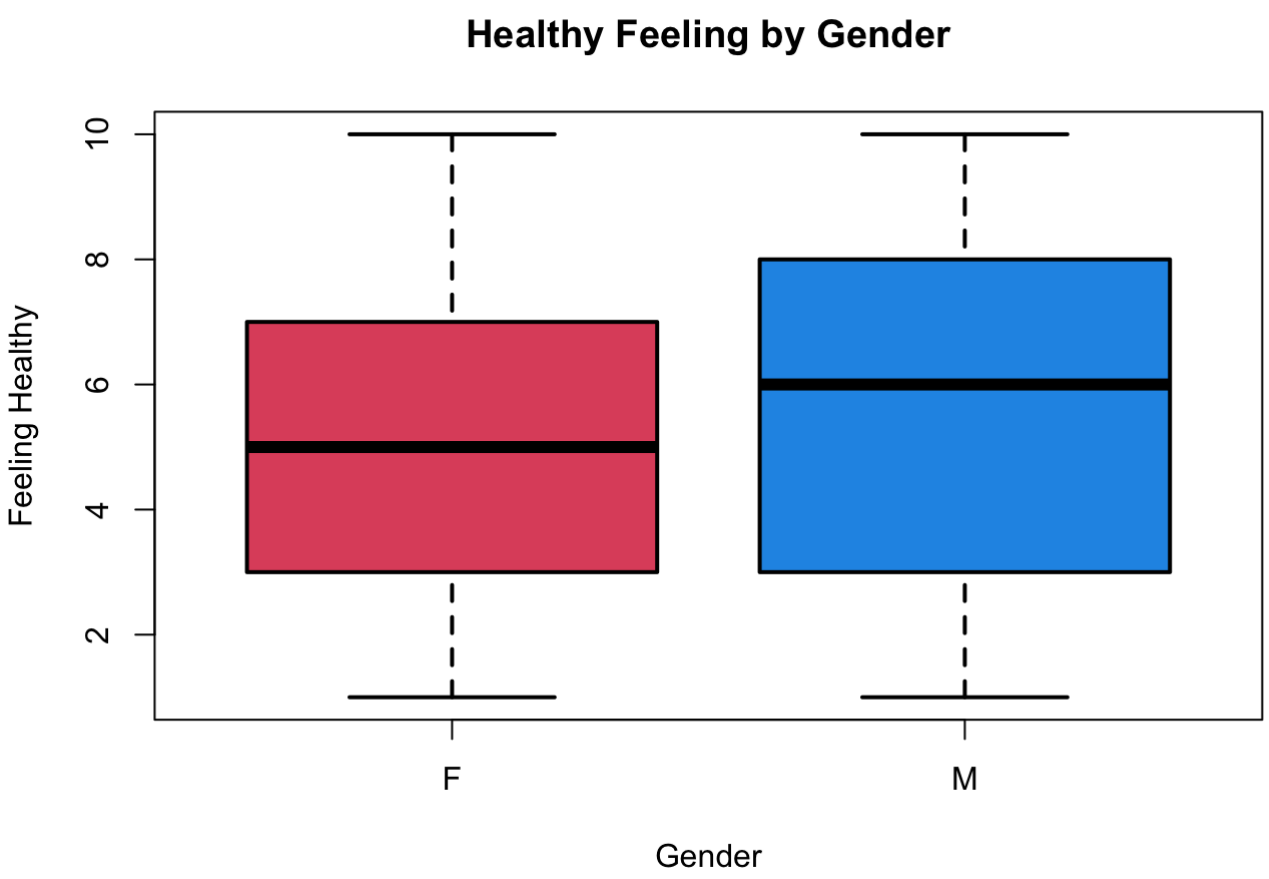

```
## [1] 11 16
```

For the variety foods score, the plot seems more symmetrical and approximately normally distributed, save for the people who rated every component question 5, and the NQ plot is more linear than the example component that was used to create the score.

5. Basic Test

T-test and Bootstrap for Healthy Feeling

We start with a boxplot to see differences between feelings of healthiness between Females and Males.

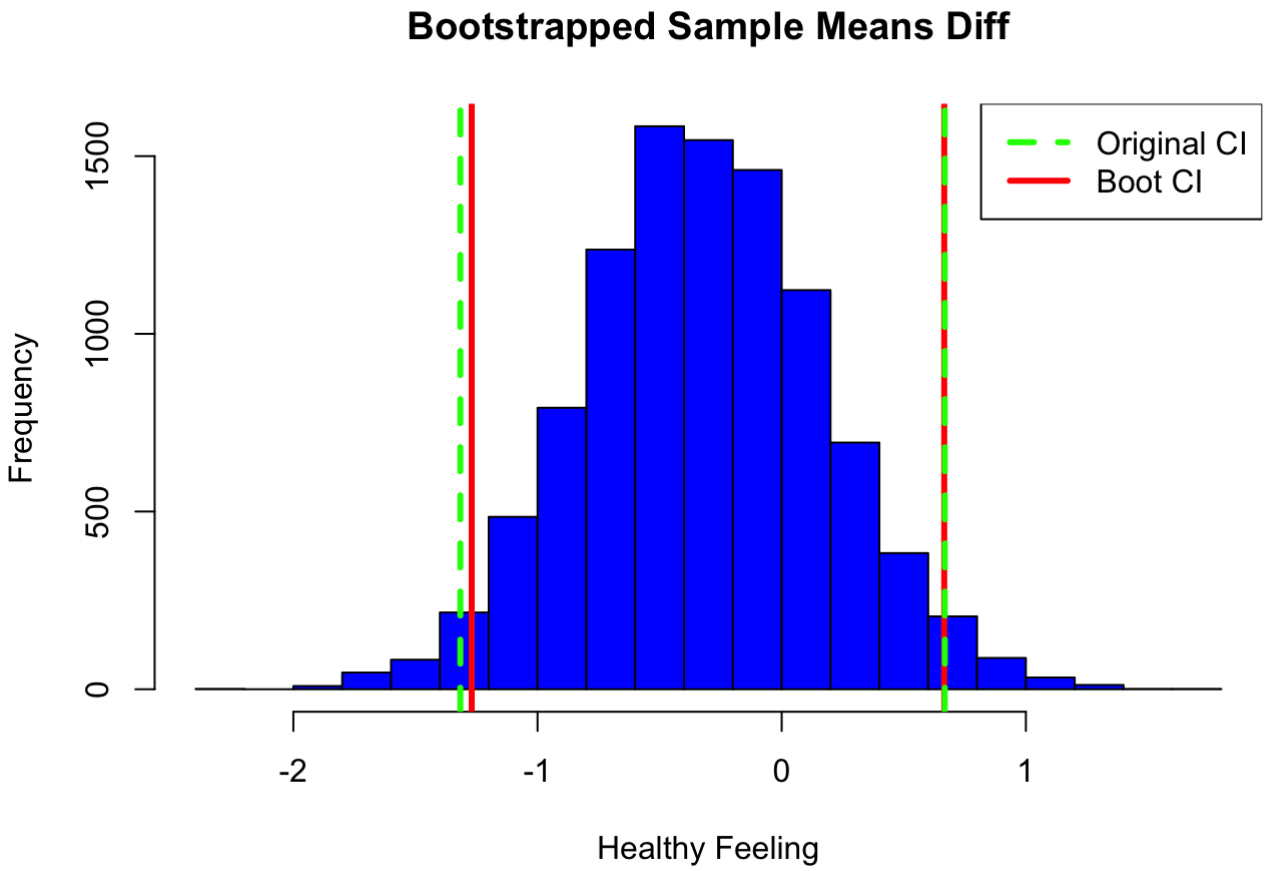


According to the boxplot, males visually seem to have a slightly higher median and a larger interquartile range of feeling healthy than females. To test if the difference is actually statistically significant, we ran a two sample T-test across gender groups.

```
##
##  Welch Two Sample t-test
##
## data:  healthy_feeling by Gender
## t = -0.64946, df = 85.702, p-value = 0.5178
## alternative hypothesis: true difference in means between group F and group M is not e
qual to 0
## 95 percent confidence interval:
##  -1.316249  0.668021
## sample estimates:
## mean in group F mean in group M
##        5.328947        5.653061
```

We fail to reject the null hypothesis, that is, there is no evidence of a statistically significant difference between two groups, because the p-value 0.5178 is greater than the significance level of 0.05. 95% confidence interval (-1.3, 0.7) also contains 0, further supporting the conclusion.

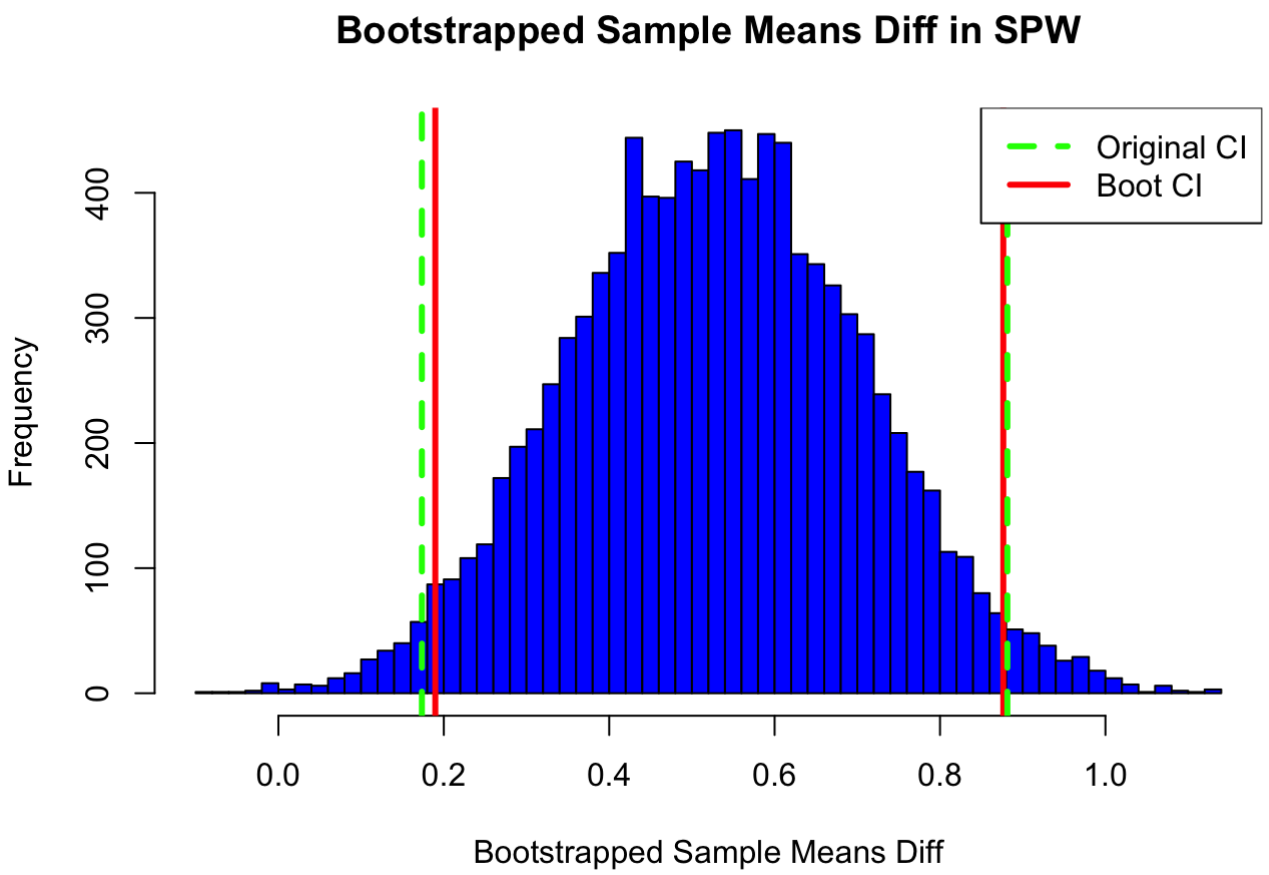
The Central Limit Theorem states that any sample with size greater than 30 can be considered to have a normal distribution of sample means. Though our sample size is 126 and thus satisfies the condition, we would still like to perform bootstrap and compare the confidence interval of bootstrapped mean difference to the original confidence interval calculated above.



Rounding to the first decimal place, both bootstrapped and original confidence intervals are (-1.3, 0.7), although the bootstrapped confidence interval to be slightly slimmer than the original interval in the plot. Since both include a difference in means of 0, we conclude that there is no statistically significant difference between the mean healthy feeling reported by Females and Males.

T-test and Bootstrap for Self Perceived Weight

Let us consider another continuous variable `self_perception_weight`. We suspected that females might have harsher standards about their weights compared to males. We will perform a two-sample t-test and a bootstrap to get the CI's for the mean difference in self perceived weight between Females and Males.



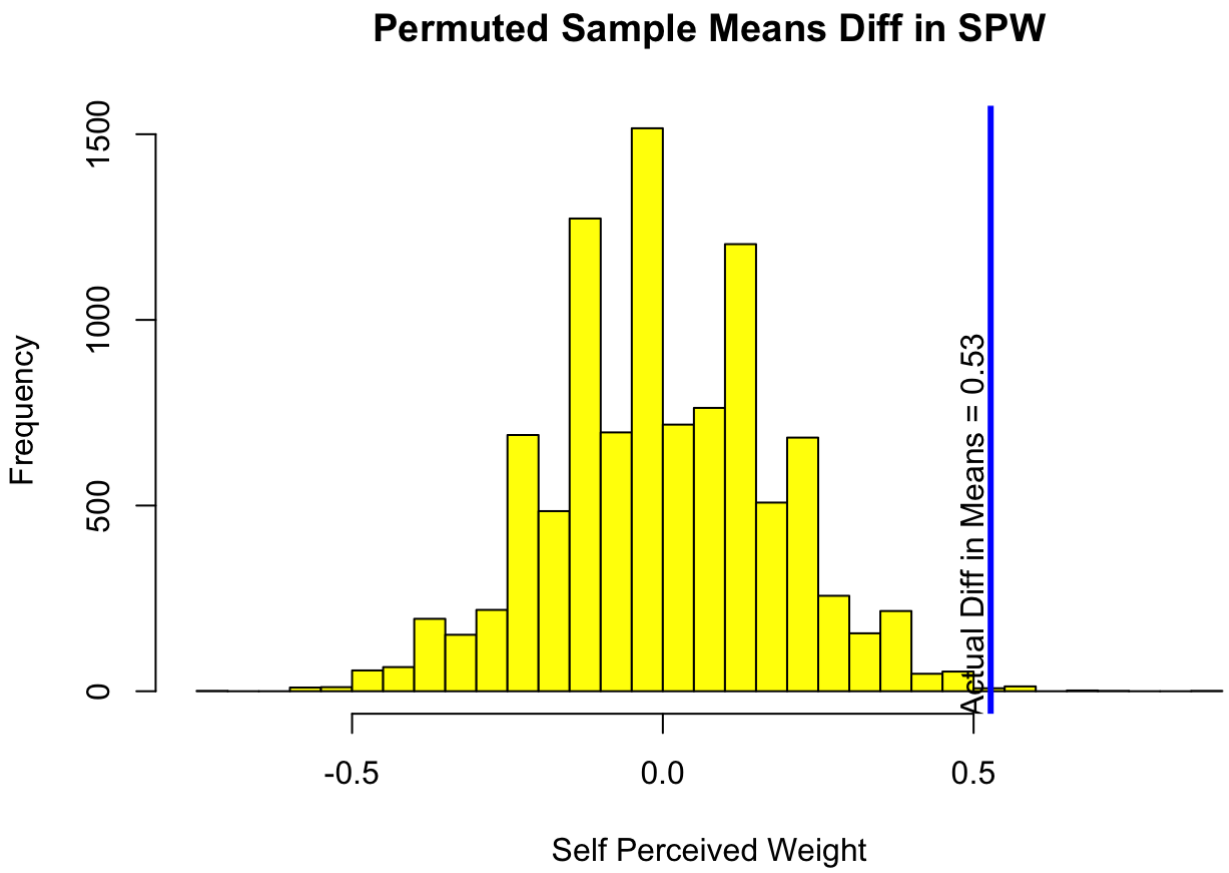
Rounding to the first decimal place, the CI's have the same range and bound again of (0.2, 0.9). Since both do not include a difference in means of 0, we conclude that there is a statistically significant difference between the mean healthy feeling reported by Females and Males.

Permutation Test

Since we know that the means in self-perception of weights (SPW) are statistically different in males and females, now we want to test if females have statistically higher SPW compared to men. To test our hypothesis, we will do a permutation test for two samples. Below are the null and one-sided alternative hypotheses.

$$H_0 : \mu_F - \mu_M = 0$$

$$H_a : \mu_F - \mu_M > 0$$



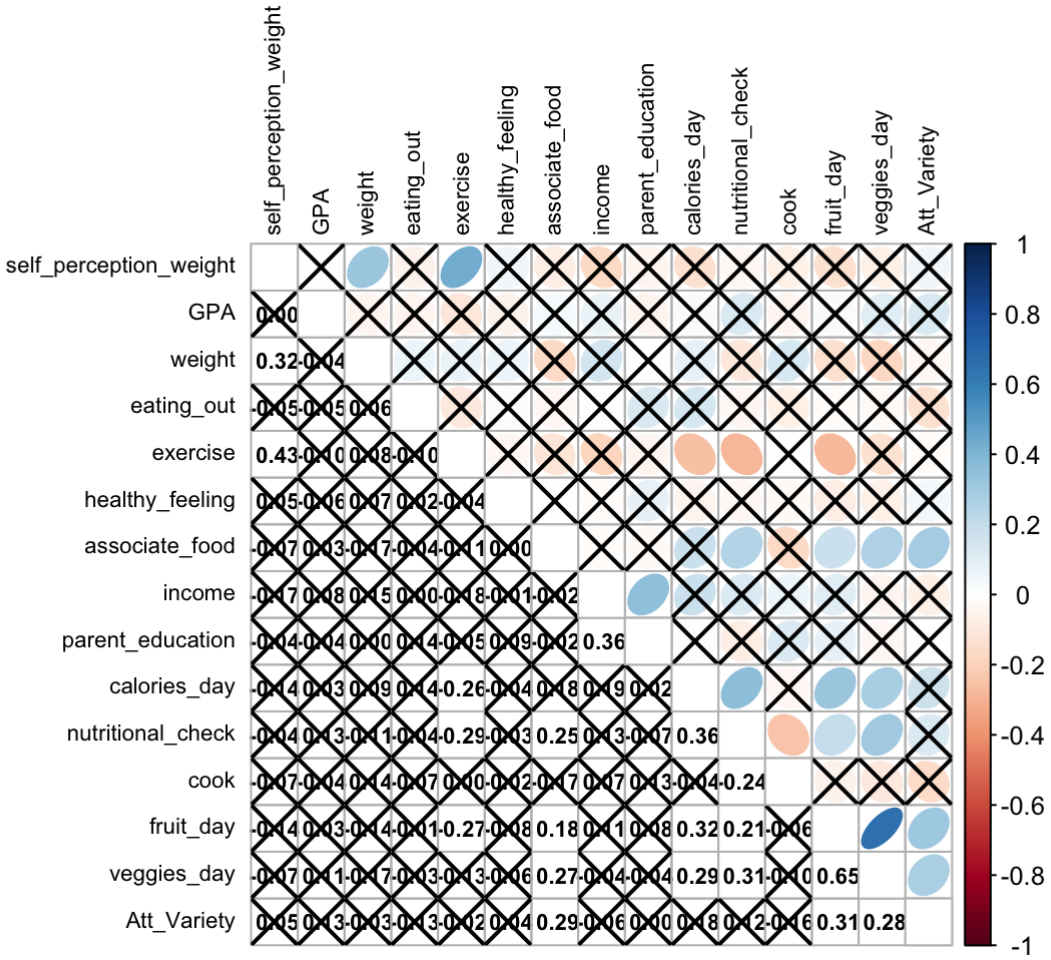

```
## [1] 0.0025
```

Since the p-value of 0.002 is less than the significance level of 0.05, we reject the null hypothesis and accept our alternative one-sided hypothesis: the mean SPW is higher for females than for males.

6. Multiple Regression

Next, we will use multiple regression to find the model for predicting self-perception of weights. We first examine the correlations among variables, possible issues of multicollinearity, and need for transforming the variables. Then we will use best subsets regression, and R-squared, Adjusted R-squared, BIC, and Cp Statistics to determine the final model containing statistically significant predictors.

We will use the following continuous variables: `GPA`, `weight`, `eating_out`, `exercise`, `healthy_feeling`, `associate_food`, `income`, `parent_education`, `calories_day`, `nutritional_check`, `cook`, `fruit_day`, `veggies_day`, and `Att_variety`.

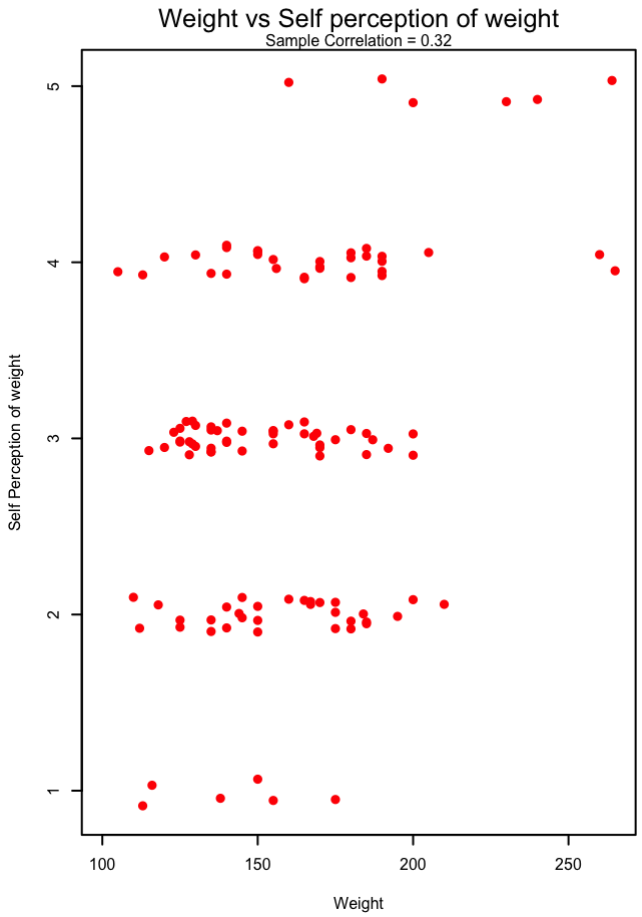


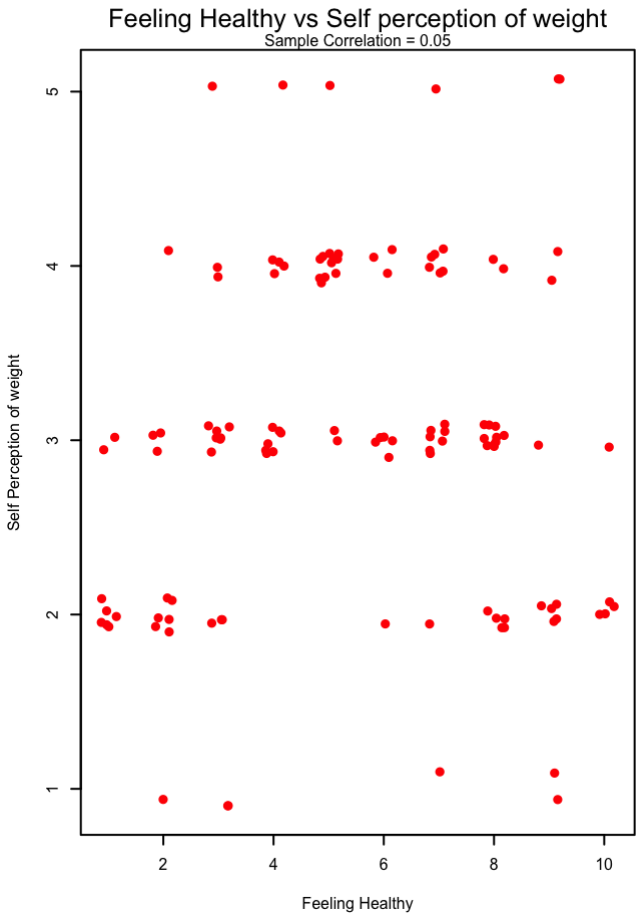
The plot above shows pairwise correlation values across variables with an X mark over correlations that are statistically non-significant at a significance level of 0.05. The variable `self_perception_weight` has relatively strong positive correlations to `weight` (0.32) and `exercise` (0.43). This is not surprising because the heavier a student weighs, the the more overweight the student may think of him/herself. Similarly, if the student exercises more often (recall that 1 in `exercise` is "exercise every day"), they are more likely to consider themselves to be fit or slim (1 in `self_perception_weight` is "slim").

Overall, however, there are some strong and significant correlations among the variables, such as `fruit_day` & `veggies_day` or `nutritional_check` & `exercise`, indicating an issue of multicollinearity.

Since we are unsure if there are other interesting nonlinear relationships across variables, we made a plot using `pairsJDRS`.

First, we found that there might be a few influential points in the variable `weight` that increase the slope and strengthens the correlation between `weight` and `self_perception_weight`. In the jittered scatterplot below, we can see that there are datapoints on the top right corner that might skew the correlation.





Next, we found that there is a nonlinear, quadratic-looking relationship between `healthy_feeling` and `self_perception_weight`, as shown below in the jittered plot.

Finally, we will perform best subsets regression and the Bayesian Information Criteria (BIC) to determine the best model predict `self_perception_weight`.

According to the BIC, the best model contains 2 variables: `weight` and `exercise`. So if we fit the model, we get the following:

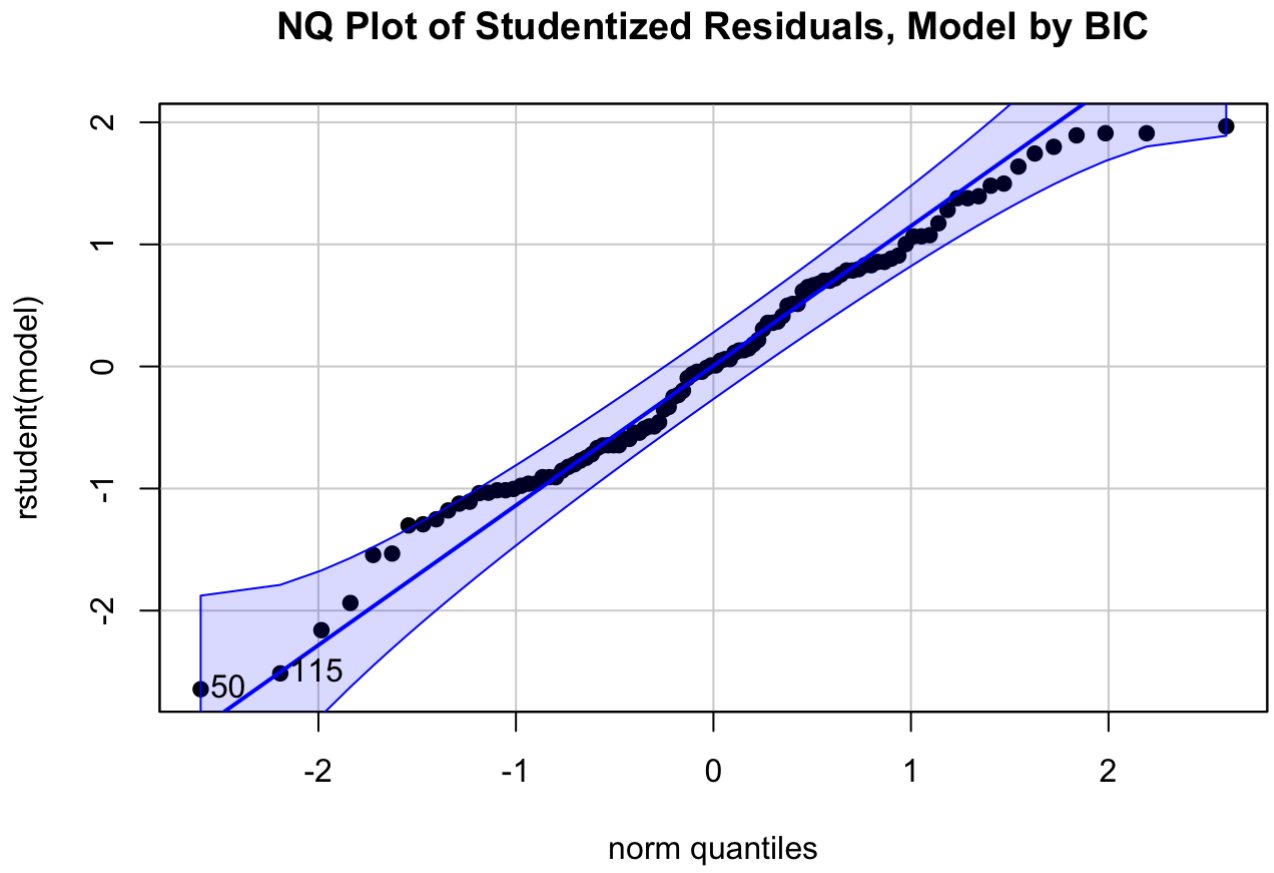
```
##
## Call:
## lm(formula = self_perception_weight ~ ., data = fc2bic1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1211 -0.6223  0.0071  0.6403  1.5950
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 0.663067   0.474896   1.396    0.16564
## weight      0.008549   0.002757   3.101    0.00249 **
## exercise    0.587880   0.120623   4.874 0.00000399 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8313 on 103 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.2574, Adjusted R-squared:  0.243
## F-statistic: 17.85 on 2 and 103 DF,  p-value: 0.0000002201
```

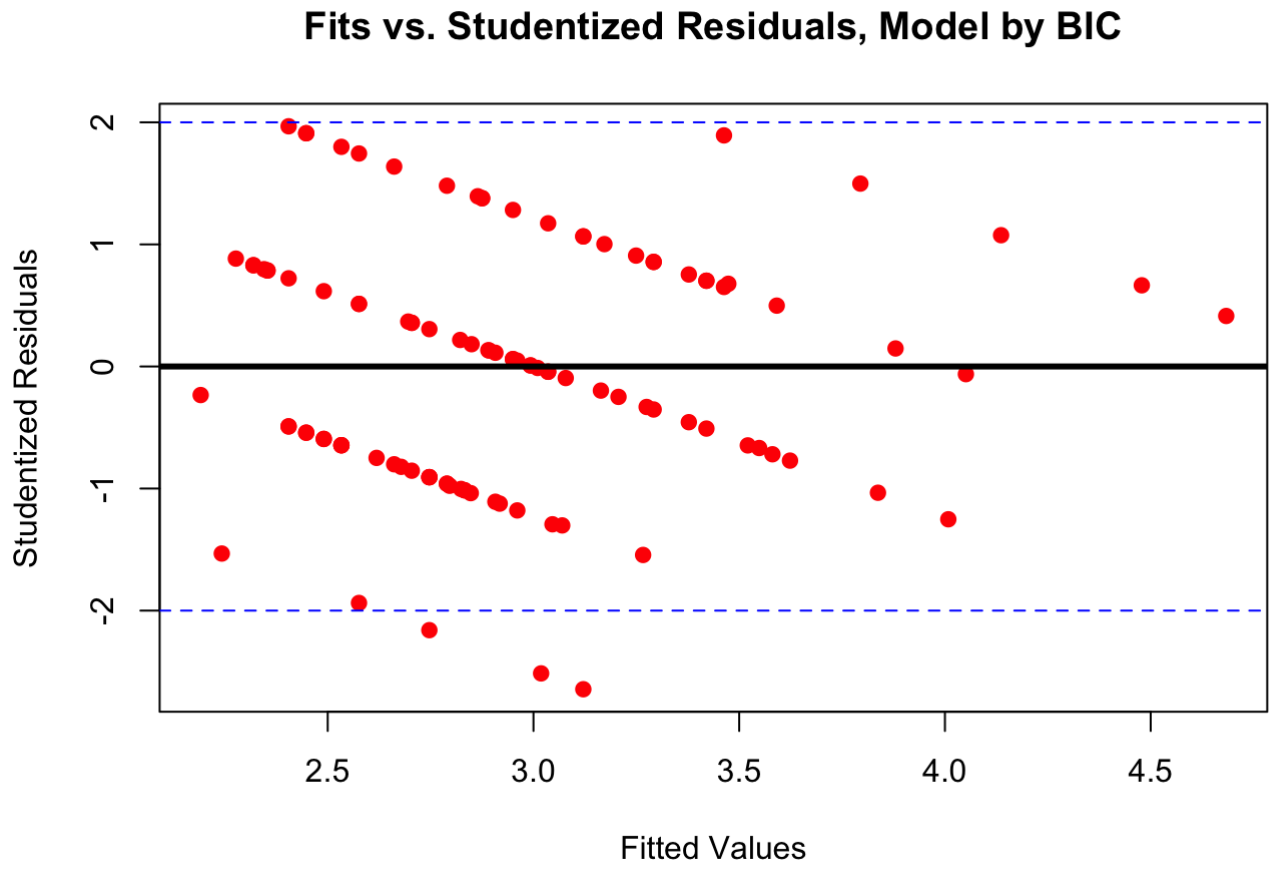
Weight and Exercise are significant predictors of the self perception of weight since both p-values are below the significance level of 0.05. The coefficient for weight is positive, which means as the weight increases, the self-perceived weight also increases. Similarly, the coefficients for exercise is positive, which means the more frequently a person exercises (recall that 1 is coded as “every day”), the less the self-perceived weight is. The R-squared value of 0.2574 indicates that approximately 26 percent of the variability is explained by this model.

There were other models suggested by the R-squared, adjusted R-squared values and Cp statistic. R-squared model recommended us to include all variables in the model as it has the highest R-squared value. However, since the R-squared increases with the number of predictors, it is not a relevant or useful factor in determining statistically significant predictors in our case. The model from adjusted R-squared values suggested us to include 4 variables. However, when we fit them into a linear model, two of the variables, `eating_out` and `income`, were statistically insignificant. Finally, both the BIC and Cp statistic suggested two variables, `weight` and `exercise`, and because both were statistically significant, we determined this to be our final model.

There were 19 observations deleted due to missingness in the BIC model. This is 1 less in comparison to the model suggested by the adjusted R value, and 23 less than the model from R squared value.

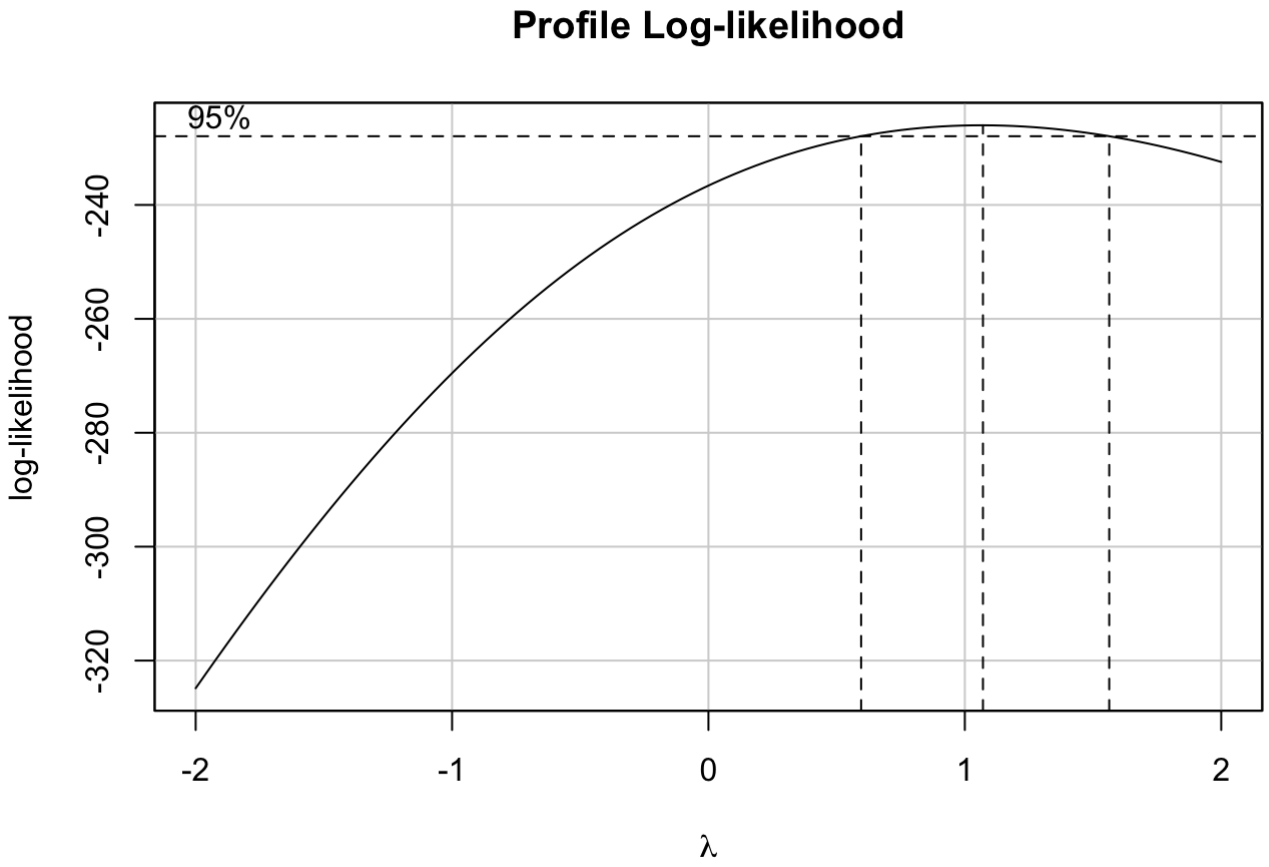
Now that we have a model, we can graph residual plots.


NQ Plot of Studentized Residuals, Model by BIC
rstudent(model)
norm quantiles
50
115
-2
-1
0
1
2


Fits vs. Studentized Residuals, Model by BIC
Studentized Residuals
Fitted Values
2.5
3.0
3.5
4.0
4.5
-2
-1
0
1
2

The plots indicate that we have met the assumptions about our regression models because the data points in the normal quantile plot are approximately in a straight line, and there are no outstanding outliers in the residual plot. The residual plot also does not fan out, make a curve, or show unwanted trends, indicating little evidence of heteroskedasticity. The plot has five slashes, however, because the response values in `self_perception_weight` are discrete numbers from 1-5 (though we may regard it as continuous).

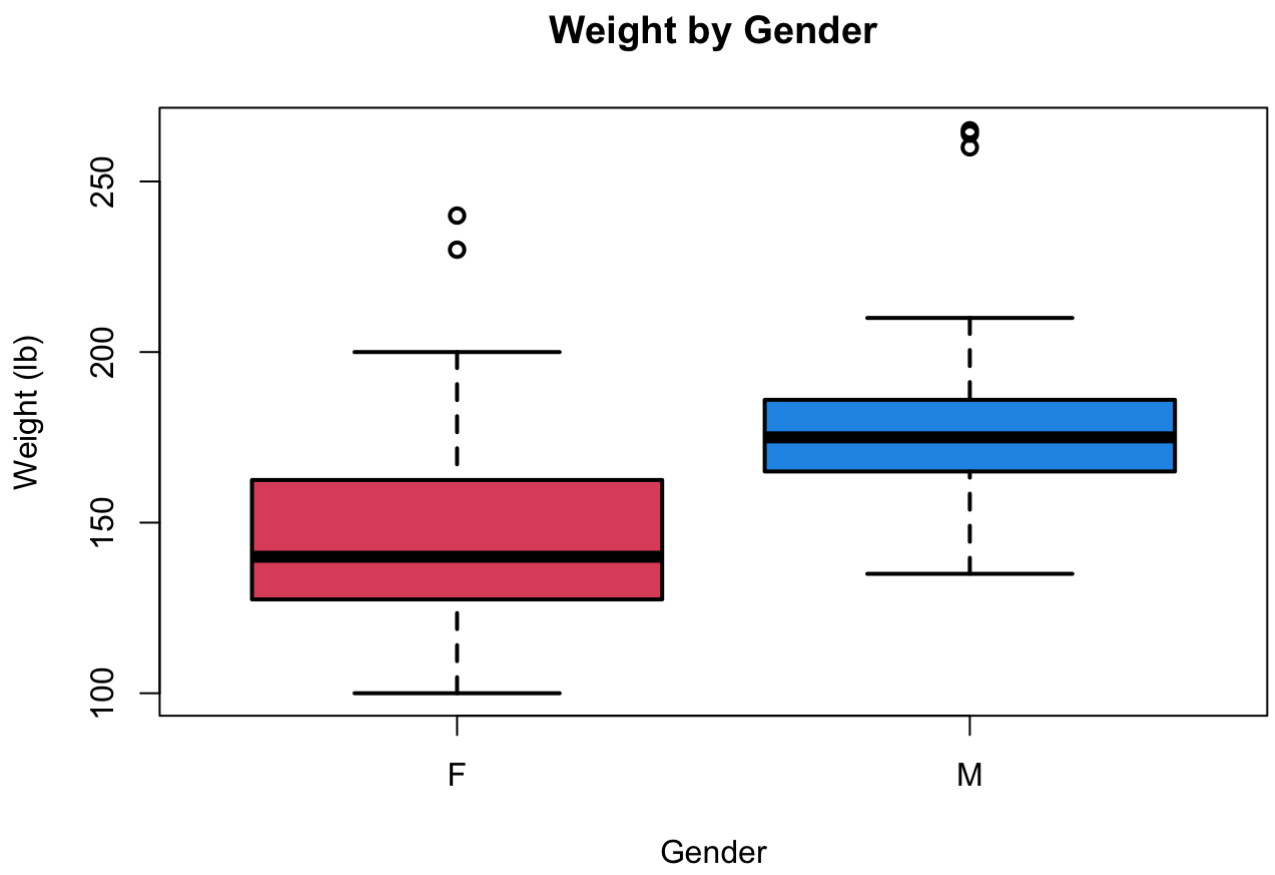
Though we do not see clear signs of non-normal distribution of residuals or heteroskedasticity, we can ensure this by performing Box-Cox procedure.



The Box-Cox procedure returns a lambda value of 1.07, but since this value is verly close to 1, which is also included in the interval, we may conclude that no transformation is necessary.

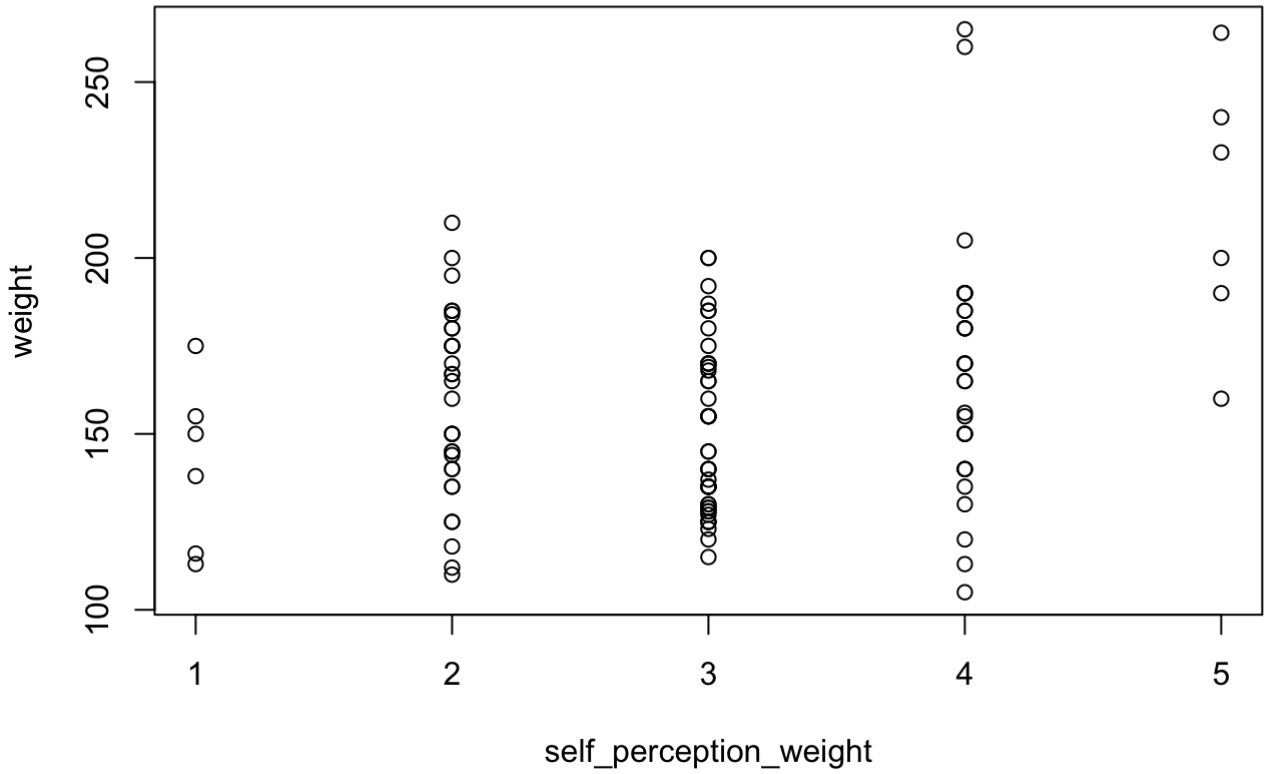
7. ANCOVA

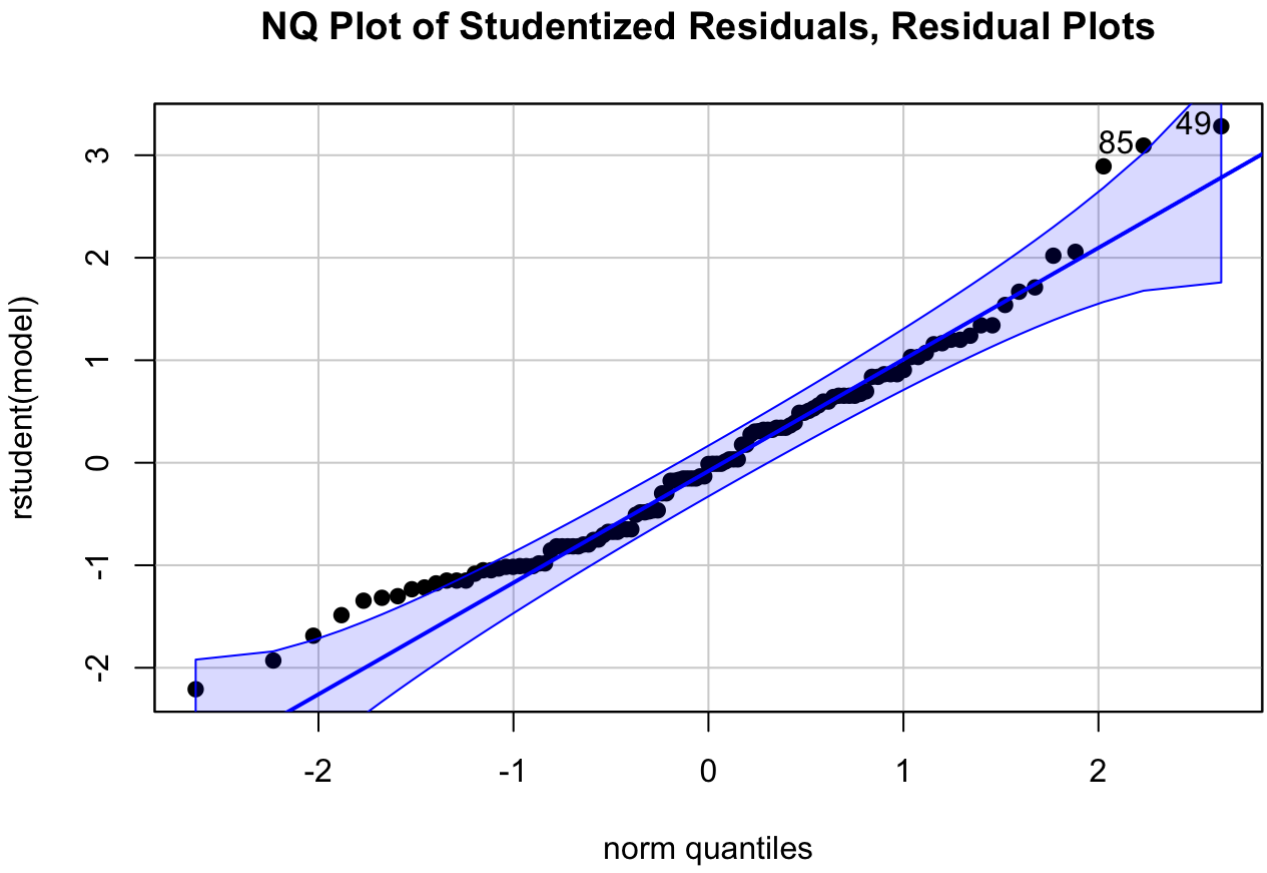
As we now know from the linear model above, the self perception of weight is a significant predictor of weight. Notice from the boxplot below that females and males seem to have different weights. We will investigate how that relationship affected by Gender by performing ANCOVA.

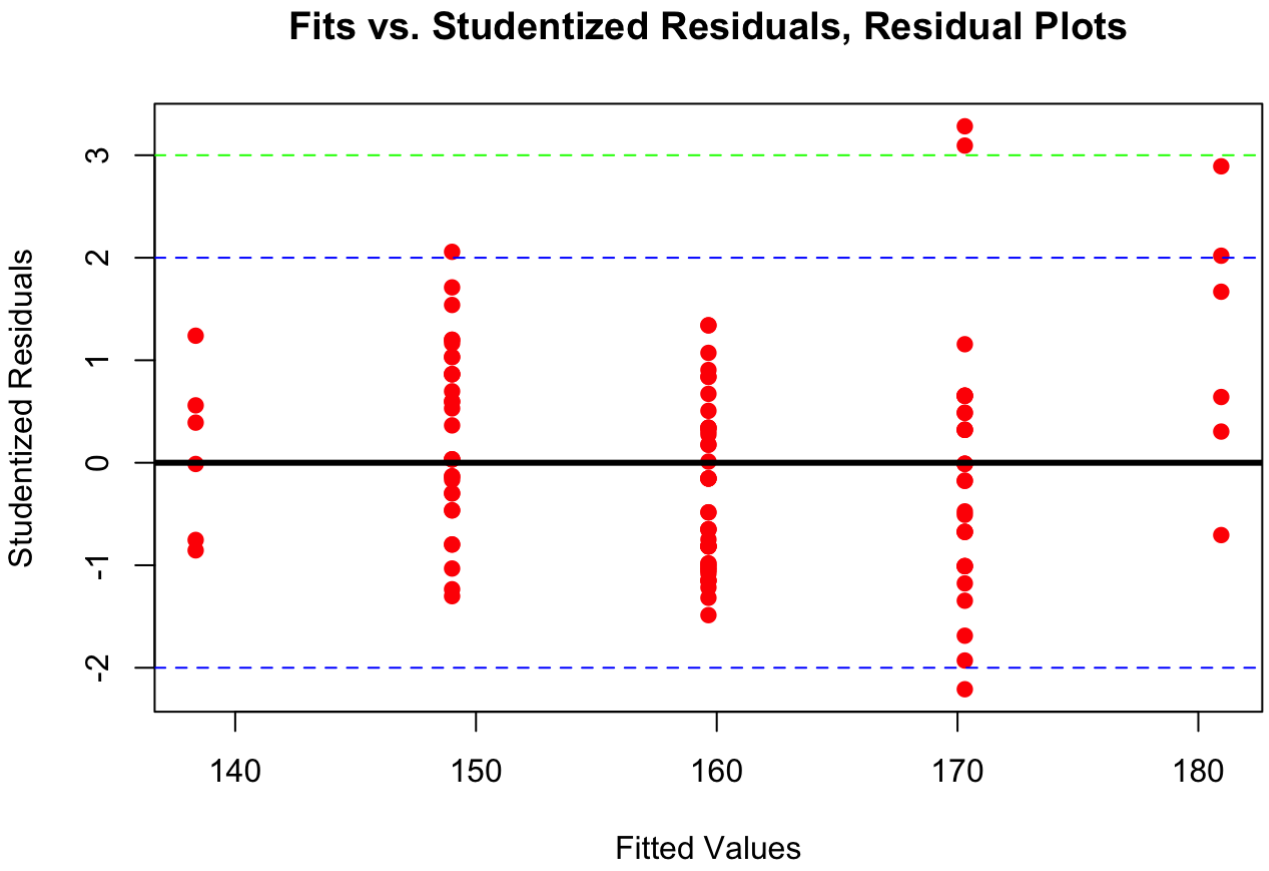


First,

we will try fitting a basic linear model.


weight
100
150
200
250
1
2
3
4
5
self_perception_weight


NQ Plot of Studentized Residuals, Residual Plots
rstudent(model)
-2
-1
0
1
2
3
85
49
-2
-1
0
1
2
norm quantiles

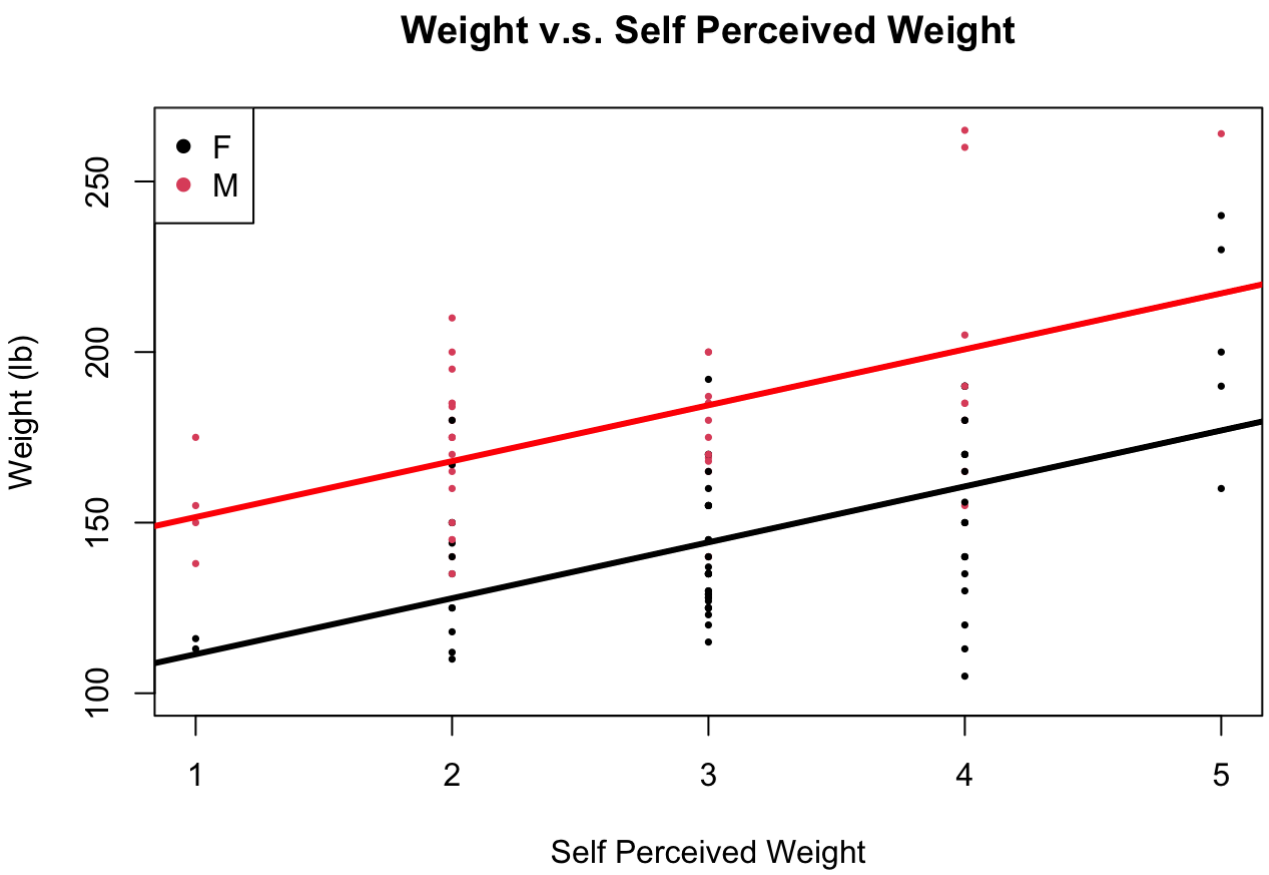


The NQ plot shows a line that is almost but not quite linear, and the plot of fits v.s. residuals seems to have some heteroskedascitiy and a few outstanding outliers. However, this is probably due to the discrete character of variables and the slightly right-skewed distribution of `weight`. Hence, we will continue to perform ANCOVA in the raw scale.

Now, we'll try adding in the effect of `Gender`.

```
## Anova Table (Type III tests)
##
## Response: fc2$weight
##                             Sum Sq  Df F value                Pr(>F)
## (Intercept)                  77433   1 136.565 < 0.00000000000000022 ***
## fc2$self_perception_weight   26964   1  47.555   0.00000000031538834 ***
## fc2$Gender                   41104   1  72.493   0.00000000000007724 ***
## Residuals                    64639 114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                (Intercept) fc2$self_perception_weight
##                   95.03246                   16.40547
##                fc2$GenderM
##                   40.18419
```



The R-squared is 0.45, which means that 45% of the variability is explained by the model. Furthermore, both the predictors `gender` and `self_perception_weight` are significant as their p-values are less than 0.05. We can see that males have higher intercept than females by 47.6 pounds, which makes sense because males generally weigh more than females. We also notice that the `self_perception_weight` by `Gender` increases as the `weight` increases.

Perhaps the slopes are different for Females and Males, so we should fit more than one slope based on Gender by adding an interaction term between `self_perception_weight` and `Gender`.

```
## Anova Table (Type III tests)
##
## Response: fc$weight
##                                   Sum Sq  Df F value              Pr(>F)
## (Intercept)                        47346   1 82.7697 0.00000000000000373 ***
## fc$self_perception_weight          15430   1 26.9751 0.00000092137373041 ***
## fc$Gender                           4147   1  7.2497             0.00817 **
## fc$self_perception_weight:fc$Gender    1   1  0.0014             0.96993
## Residuals                          64638 113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                          (Intercept)           fc$self_perception_weight
##                              95.2801                             16.3283
##                         fc$GenderM fc$self_perception_weight:fc$GenderM
##                              39.6574                              0.1827
```

The model looks almost exactly the same as the previous ANCOVA model. This is because the interaction between Gender and self perception of weight is not significant (the p-value is 0.96 > 0.05) and the coefficient for the interaction term is only 0.183 compared to the slope of females, which is 16.3. Hence the regression lines still appear to be parallel. The R-squared value does not change either.

8. Discussion

We investigated data on students' opinions and habits on food, eating, and health by analyzing the data "Food Choices: Food Choices and Preferences of College Students". After making and exploring a useable dataset, we found no difference in healthy feeling but a difference in self perceived weight between females and males using T-tests, bootstraps, and a permutation test. Our multiple regression model found that the more frequent the exercise, the less the self perceived weight, and the higher the actual weight, the more the self perceived weight. Lastly, we got that there is no significant interaction between gender and self perception of weight when predicting for weight from ANCOVA models.