Jalen Jackson
CPSC 375

Body Composition Analysis

Preparation

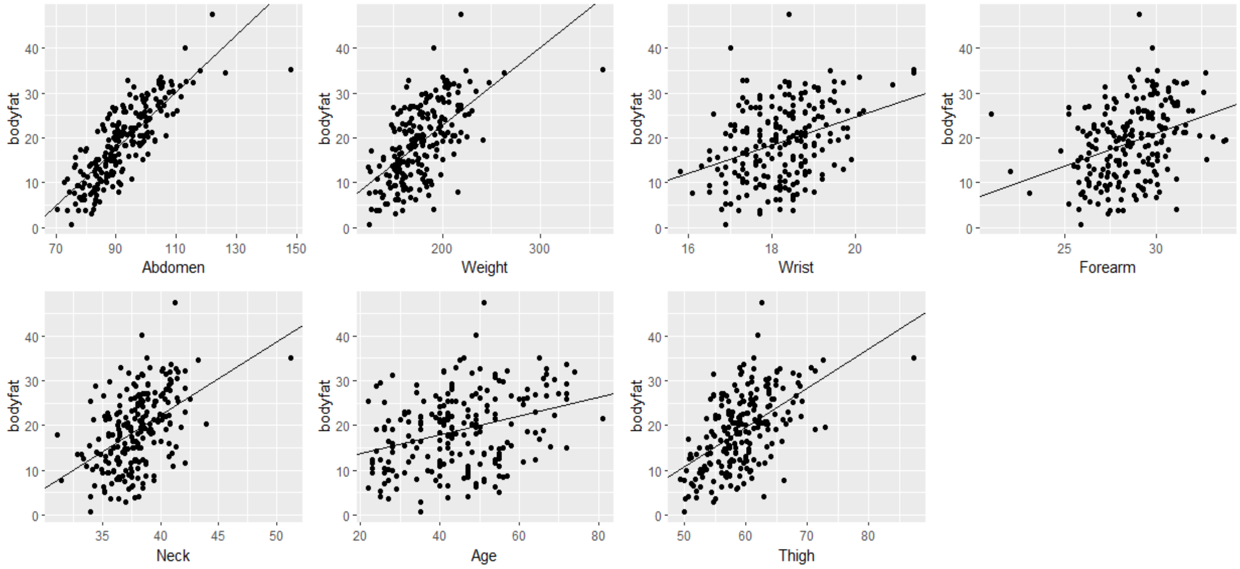
First, I read the data into R then I noticed that the density column must be removed according to the Project 1 Requirements. Also, I wanted to make sure that all of the observations' body fat percentages followed the Siri's Equation, $\textbf{Body Fat Percentage} = \frac{\textbf{495}}{\textbf{Density}} - \textbf{450},$ by filtering out the data that doesn't follow this equation. As a result, 36 observations were removed and the R^2 for the linear model that compares body fat and density increased from approximately 97% to 100%.

Picking Best Measurements to Calculate Body Fat

Next, I used the built-in **step** function in R to create linear models until the lowest Akaike information criterion (AIC) is found. The step function uses AIC to analyze the quality of a set of statistical linear regression models from body fat percentages and none of the variables to body fat percentages and all the variables. As the AIC gets lower, then the quality for its linear model gets higher. A variable gets added into the model if its calculated AIC reaches closer to the lowest AIC. If there are no variables that decrease the AIC, then we can say that this is the lowest AIC and the best model that neither over-fits or under-fits the dependent variable. In respect to the body composition data, I found that the lowest AIC is **631.94** and the best measurements to predict body fat according to this function is **Abdomen**, **Weight**, **Wrist**, **Forearm**, **Neck**, **Age**, and **Thigh**.

Plotting Best Measurements to Check Its Linearity

Then, I checked to make sure that the Abdomen, Weight, Wrist, Forearm, Neck, Age and Thigh measurements were linear to body fat by creating a linear model for each of the explanatory variables to body fat and plot it separately.



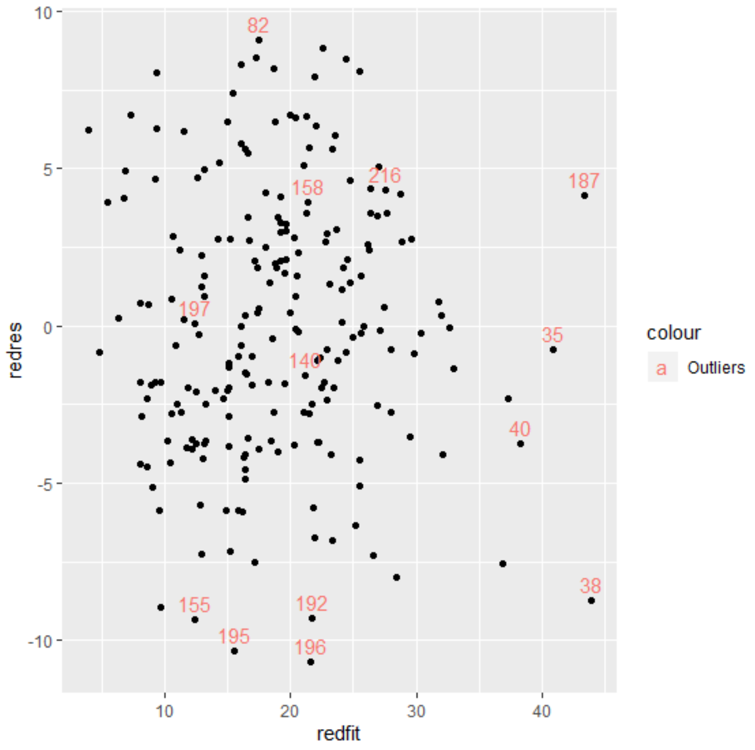
Significance & F-Test

Although Abdomen, Weight, Wrist, Forearm, Neck, Age and the Thigh measurements were the best variables according to the model that has the lowest AIC; these variables must be analyzed further for its statistical significance and the F-test so that I can obtain a simpler model. Out of all the 7 linear models created by removal or keeping a variable, 3 of the models removed a variable while 4 of them kept it. The best model that had the highest R^2 was the third model that **removed Hip**, **Neck**, and **Thigh** due to their P-values being greater than 5%, but **kept Abdomen**, **Weight**, **Wrist** & **Forearm** due to their P-values being less than 5%.

```
Model 1: bodyfat ~ Abdomen + Weight + Wrist + Forearm + Neck + Age
Model 2: bodyfat ~ Abdomen + Weight + Wrist + Forearm + Neck + Age + Thigh
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    209 3781.8
2    208 3740.0  1    41.853 2.3277 0.1286


Model 1: bodyfat ~ Abdomen + Weight + Wrist + Forearm + Neck
Model 2: bodyfat ~ Abdomen + Weight + Wrist + Forearm + Neck + Age
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    210 3832.5
2    209 3781.8  1    50.724 2.8032 0.09557 .


Model 1: bodyfat ~ Abdomen + Weight + Wrist + Forearm
Model 2: bodyfat ~ Abdomen + Weight + Wrist + Forearm + Neck
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    211 3878.7
2    210 3832.5  1    46.176 2.5302 0.1132


Model 1: bodyfat ~ Abdomen + Weight + Wrist
Model 2: bodyfat ~ Abdomen + Weight + Wrist + Forearm
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    212 4007.0
2    211 3878.7  1    128.24 6.9761 0.008879 **


Model 1: bodyfat ~ Abdomen + Weight + Forearm
Model 2: bodyfat ~ Abdomen + Weight + Wrist + Forearm
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    212 4085.4
2    211 3878.7  1    206.64 11.241 0.0009481 ***


Model 1: bodyfat ~ Abdomen + Wrist + Forearm
Model 2: bodyfat ~ Abdomen + Weight + Wrist + Forearm
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    212 4410.0
2    211 3878.7  1    531.32 28.904 2.009e-07 ***


Model 1: bodyfat ~ Weight + Wrist + Forearm
Model 2: bodyfat ~ Abdomen + Weight + Wrist + Forearm
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    212 9309.4
2    211 3878.7  1    5430.6 295.42 < 2.2e-16 ***
```

Removing Outliers

Observations like 38 where the weight is 363.15 pounds will skew our reduced model such that there will be highly inaccurate predictions for future observations and testing. We must remove the observations where they are outliers found due to high or low Abdomen, Weight, Wrist, and Forearm value. This is the graph between the residuals and fitted values for our reduced model and it displays the row the outlier is at and how much it is affecting our linear model based on its position in the graph.



Cross Validation

After outliers were removed, prediction tests can now be done to validate the linear model. First, I sampled 80% of the reduced data into a training set and 20% of it into a testing set into confirm our predictions. Then, I created another linear model that has the same formula as our base model but tested on the training set. I gathered the Abdomen, Weight, Wrist and Forearm values for the testing set and put it into its own data frame. Finally, I used the training model and test frame to predict a body fat percentage with a 95% prediction interval. Here are my results down below:

- Number of predictions within 95% prediction interval:
  40 out of 41 observations
- Number of predictions within ±5% prediction interval from actual body fat percentage:
  33 out of 41 observations

Therefore, we can conclude that

$$Body\ Fat\ Percentage = -36.6803 + 1.0156 * Abdomen - 0.1395 * Weight - 1.6907 * Wrist + 0.6207 * Forearm$$

is an accurate model for predicting body fat percentages.