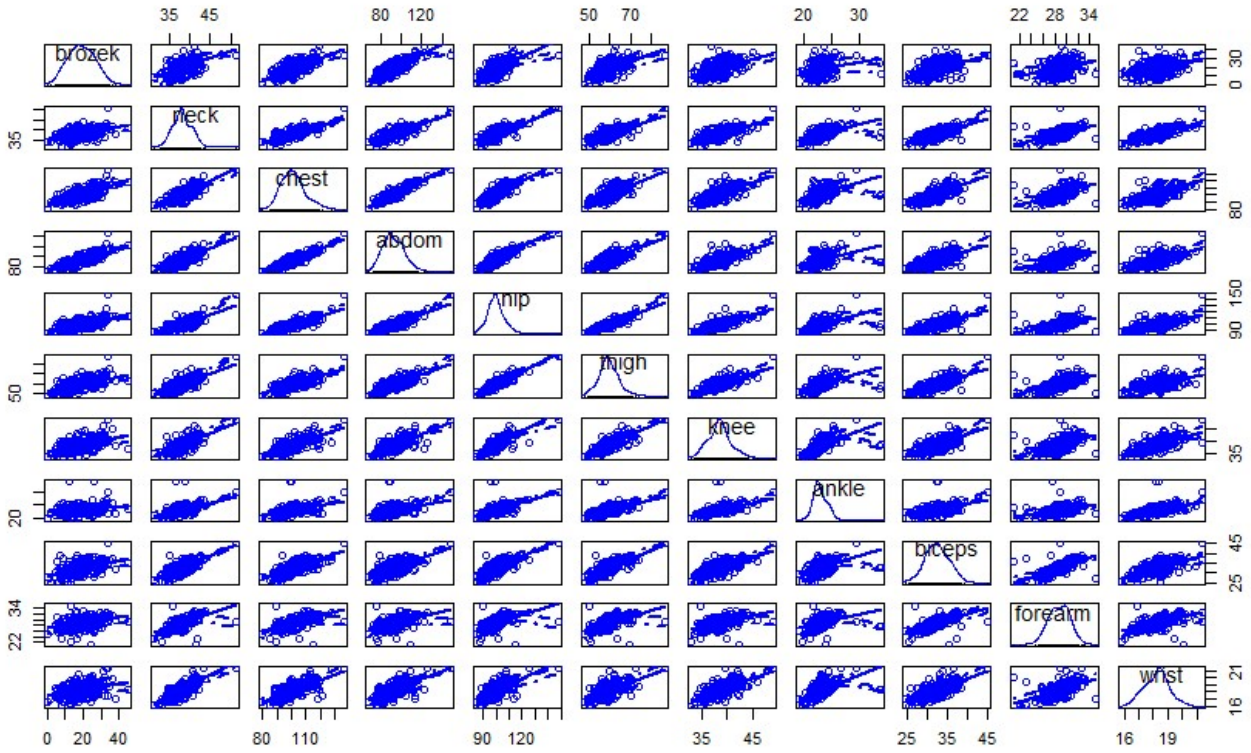
Finding the Best Model to Predict Body Fat Content

By Kian Khaffajian

This statistical report aims to find the "best" model for predicting body fat content using 10 body measurements from 202 men. The independent variables in question are the neck, chest, abdomen (abdom), hip, thigh, knee, ankle, biceps, forearm and wrist, and of course our dependant variable is body fat content (brozek). We will begin by conducting some exploratory research to highlight anything interesting about the data set. Next, we begin the model selection process, which comprises of finding and testing different models. Finally, we validate these models by checking for skewness and looking at residual plots. To assess the predictive capabilities of the model, we use some testing data, where the same measurements for 50 more men were taken. The aim is to see how our model compares with the full model.

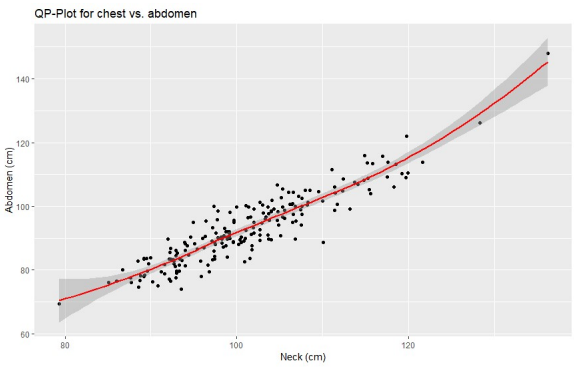
To being, let's look at the scatterplot matrix below. The correlation between body fat content and all the other body parts is positive with its relationship with abdomen being the strongest and ankle being the weakest. Moreover, we can see that variability of the data comparing body fat with covariates tends to increase as we move along from left to right. The relationship between all the covariates are positive, with most having a relatively low amount of variability. The histograms from brozek to ankle all have a slight positive skew, bicep is symmetrical and the remaining two have a slight negative skew. This can be reinforced by looking at the relationship between each variable's mean and median (e.g. positive skew if (mean>median) from the summary of the data file, located beneath the scatterplot matrix.



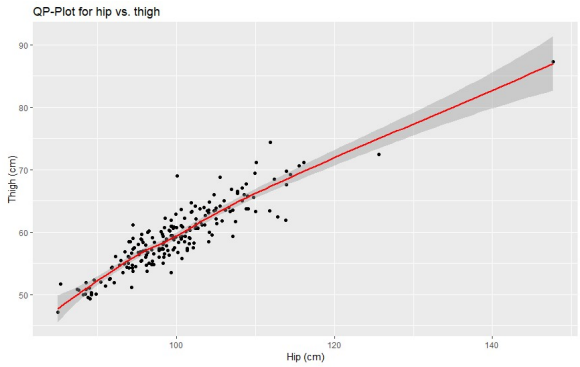
```
Train <- read.csv("C:/Users/kian/Desktop/R/SMM Cw/Train.txt",sep="")
summary(Train)
```

```
##      brozek           neck           chest           abdom       
##  Min.   : 0.00   Min.   :31.50   Min.   : 79.30   Min.   : 69.40  
##  1st Qu.:13.40   1st Qu.:36.50   1st Qu.: 94.45   1st Qu.: 84.45  
##  Median :19.00   Median :38.00   Median :100.30   Median : 91.05  
##  Mean   :19.09   Mean   :38.19   Mean   :101.18   Mean   : 92.97  
##  3rd Qu.:24.60   3rd Qu.:39.98   3rd Qu.:106.15   3rd Qu.: 99.80  
##  Max.   :45.10   Max.   :51.20   Max.   :136.20   Max.   :148.10  
##       hip             thigh            knee            ankle       
##  Min.   : 85.00   Min.   :47.20   Min.   :33.00   Min.   :19.10  
##  1st Qu.: 95.50   1st Qu.:56.15   1st Qu.:36.92   1st Qu.:22.00  
##  Median : 99.35   Median :58.95   Median :38.40   Median :22.75  
##  Mean   :100.02   Mean   :59.40   Mean   :38.56   Mean   :23.07  
##  3rd Qu.:103.67   3rd Qu.:62.45   3rd Qu.:39.80   3rd Qu.:23.88  
##  Max.   :147.70   Max.   :87.30   Max.   :49.10   Max.   :33.90  
##      biceps          forearm          wrist      
##  Min.   :24.80   Min.   :21.00   Min.   :15.80  
##  1st Qu.:30.23   1st Qu.:27.40   1st Qu.:17.62  
##  Median :32.40   Median :28.75   Median :18.30  
##  Mean   :32.40   Mean   :28.69   Mean   :18.26  
##  3rd Qu.:34.48   3rd Qu.:30.10   3rd Qu.:18.80  
##  Max.   :45.00   Max.   :34.90   Max.   :21.40
```

It is, however, important to look at some of these plots in more detail. The two QP-Plots below were chosen as they have a very high positive correlation with a low amount of variability, which suggests that one variable from each graph can be discarded from the model as they both are assumed to affect body fat in similar ways. We also notice some potential outliers or high leverage points, which will become important later in the model selection process.



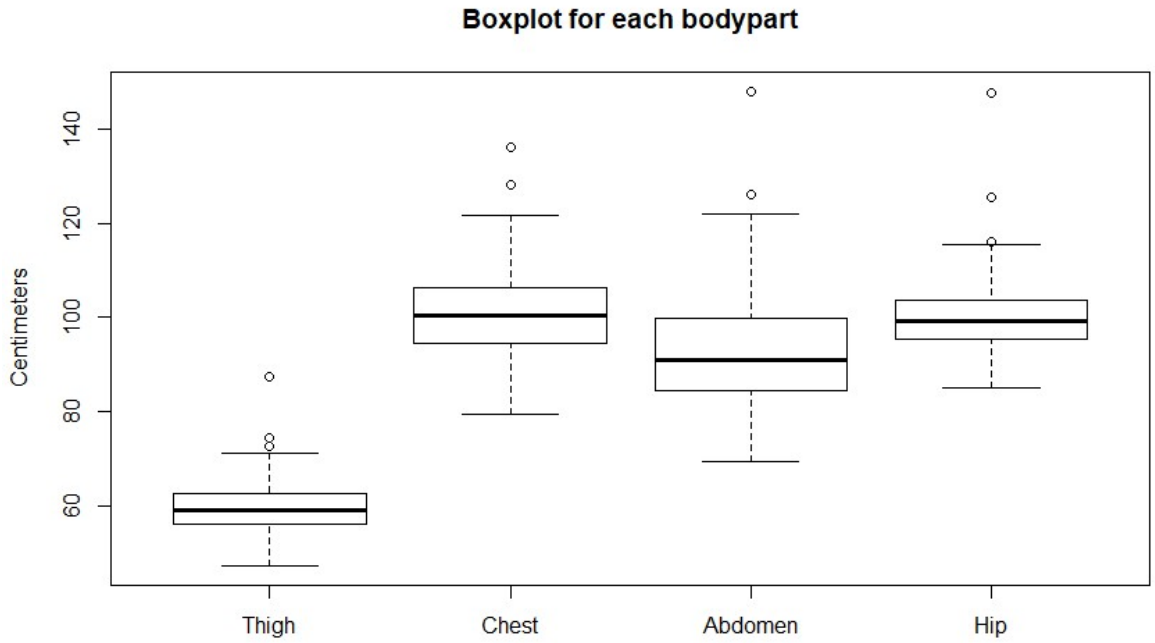
Correlation = 0.9162027

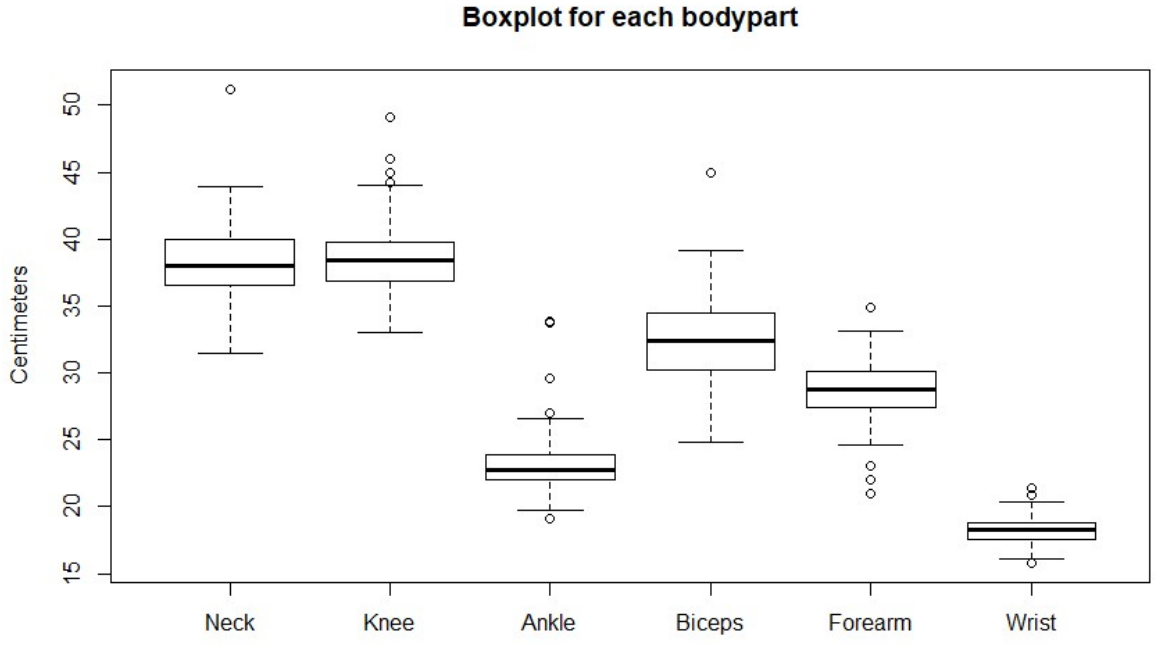


Correlation = 0 .9056402

Next, we can look at the boxplots for each variable. I have grouped the data in two sets to make it easier to read. Again, we can observe the skewness for the data and see that median line lies virtually in the centre of all the boxplots; one could argue that these variables are relatively symmetrically skewed and so won't require any transformations. This applies less to ankle, as the skew is noticeably

larger. Furthermore, we notice a fair number of outliers particularly in the ankle set, suggesting that the dataset may need altering.





We can now begin the model selection process. The general method requires the implementation of automated model selection functions in R on the null model and more a complex model. These are the best subset regression (BSR), which tries every model in the hierarchy (all 2^{10} models) and the stepwise method, which looks at adding or removing one variable on each step via the AIC criterion.

Then, depending on the models outputted, an ANOVA test is carried out to help determine the superior model.

Let's begin by fitting the null model and the full model. The best subset regression function is used on the full model, and we determine the appropriate variables to use in our model by finding the row corresponding to the number or parameters p such that Mallow's $C_p \approx p$. In this case $p = 5$, which corresponds to the fourth row (because the intercept is not considered in the method) and thus, we use the variables neck, abdom, hip and wrist, as show in the R output below.

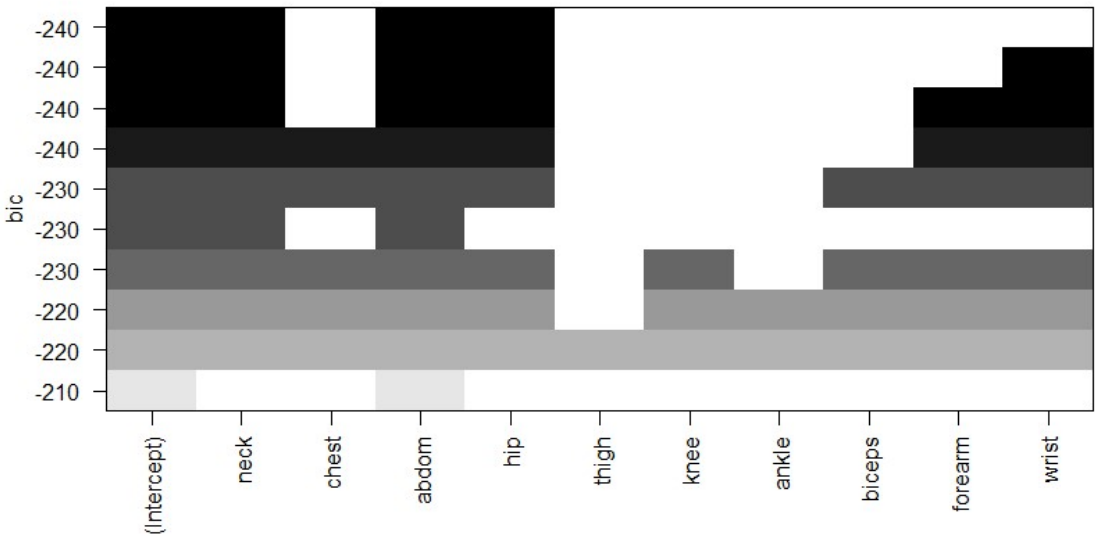
```
fit1 <- lm(brozek~., data=Train)
fit0N <- lm(brozek~1, data=Train)

#First test
BSR <- regsubsets(brozek~., data=Train,, nbest=1, nvmax=10)
summary.out <- summary(BSR)
summary.out
```

```
## Subset selection object
## Call: regsubsets.formula(brozek ~ ., data = Train, , nbest = 1, nvmax =
10)
## 10 Variables  (and intercept)
##         Forced in Forced out
## neck        FALSE      FALSE
## chest       FALSE      FALSE
## abdom       FALSE      FALSE
## hip         FALSE      FALSE
## thigh       FALSE      FALSE
## knee        FALSE      FALSE
## ankle       FALSE      FALSE
## biceps      FALSE      FALSE
## forearm     FALSE      FALSE
## wrist       FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: exhaustive
##          neck chest abdom hip thigh knee ankle biceps forearm wrist
## 1  ( 1 )  " "  " "   "*"   " " " "  " "  " "   " "    " "     " "
## 2  ( 1 )  "*"  " "   "*"   " " " "  " "  " "   " "    " "     " "
## 3  ( 1 )  "*"  " "   "*"   "*" " "  " "  " "   " "    " "     " "
## 4  ( 1 )  "*"  " "   "*"   "*" " "  " "  " "   " "    " "     "*"
## 5  ( 1 )  "*"  " "   "*"   "*" " "  " "  " "   " "    "*"     "*"
## 6  ( 1 )  "*"  "*"   "*"   "*" " "  " "  " "   " "    "*"     "*"
## 7  ( 1 )  "*"  "*"   "*"   "*" " "  " "  " "   "*"    "*"     "*"
## 8  ( 1 )  "*"  "*"   "*"   "*" " "  "*"  " "   "*"    "*"     "*"
## 9  ( 1 )  "*"  "*"   "*"   "*" " "  "*"  "*"   "*"    "*"     "*"
## 10  ( 1 ) "*"  "*"   "*"   "*" "*"  "*"  "*"   "*"    "*"     "*"
```

```
summary.out$cp
```

```
##  [1] 52.964732 24.461290  8.574988  6.153231  3.324409  4.391811  5.542
002
##  [8]  7.215071  9.017266 11.000000
```



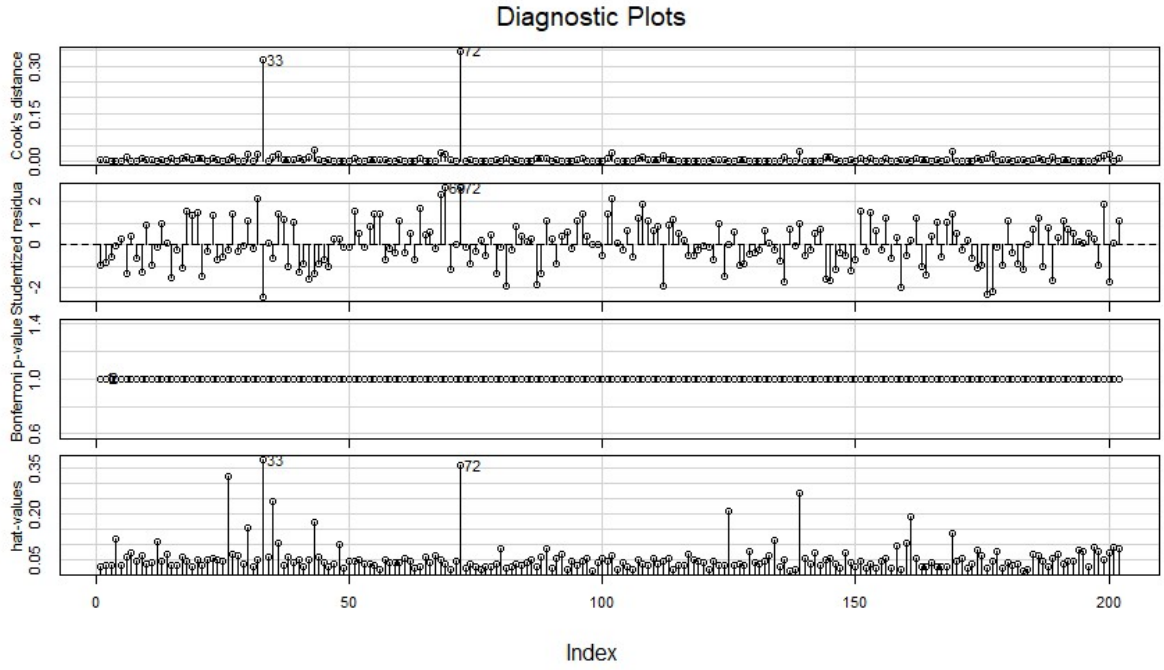
The graph above considers the Bayesian information criterion (BIC) and theory says that we can find a good model by minimising the BIC. This graph suggests that we use the four models with a BIC of -240 but we will consider the with one that matches the previously found model, and so we have

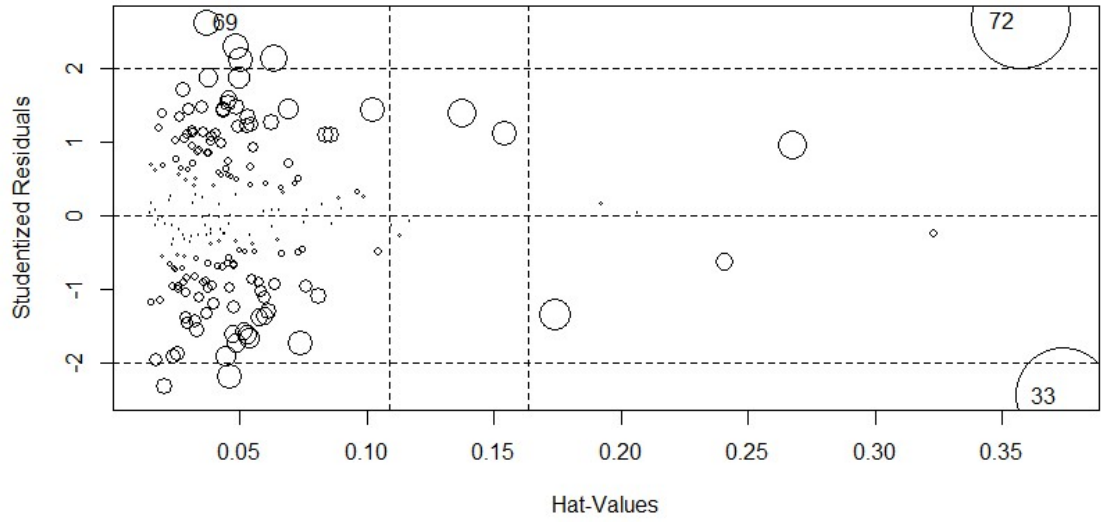
$$Brozek = a + b * neck + c * abdom + d * hip + e * wrist + \varepsilon.$$

The next step is to perform backward stepwise regression on the full model and the forward regression on the null model. These methods are generally used when there is a large amount of data and can doesn't consider all models unlike the BSR, making it slightly unreliable. In any case, we find that the best combination of variables to use are neck, abdom, hip, forearm and wrist – this contradicts the aforementioned model. An ANOVA test suggests that there is evidence that latter model is the superior one. This also matches ones of the models shown in the BIC graph. We now have,

$$Brozek = a + b * neck + c * abdom + d * hip + e * forearm + f * wrist + \varepsilon.$$

As mentioned in the exploratory research, there are some potential outliers and high leverage points that need to be spotted. Below is a pair of diagnostic plots that help determine these alleged outliers. We can define and outlier as a point with a Cook's distance greater than 0.5 or studentized residuals *t,* greater than 2 or less than -2. The diagnostic plots below suggest that we consider men 33,69 and 72. This is largely due to 33 having the largest abdomen and consequently neck, as well as thigh and hip – this is the outlier denoted in the QP-plots. In addition to this all three have *t,* such that $|t|>2$, which meets our criteria for them to be considered outliers.





It is worth repeating the analysis removing these rows from the data set. Let's repeat the testing process, first removing man 33 only. The series of tests leaves us with the simpler model,

$$Brozek = a + b * neck + c * abdom + d * hip + e * wrist + \varepsilon.$$

The deviance has decreased from approximately 3158 to 3110, suggesting that this is a better model.

Removing all three outliers from the data and repeating this process once more, we find a more complex model,

$$Brozek = a + b * neck + c * abdom + d * hip + e * ankle + f * forearm + g * wrist + \varepsilon,$$

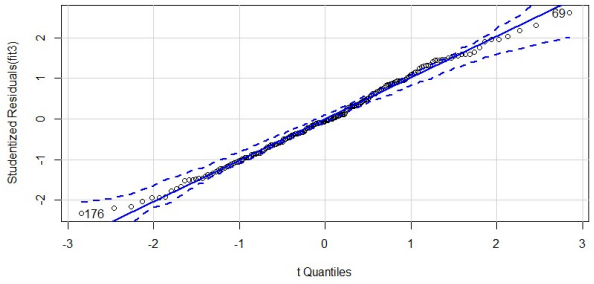
with a deviance of approximately 2864. This substantial decrease in the deviance leads us to believe that this is a better model, however, this sizeable change may be since we have added two more variables back in the model. This leads us to believe simpler of these two models is a better choice. In addition, removing just 69 or 72 either increases the deviance or reduces it slightly, which is not good for our model selection.

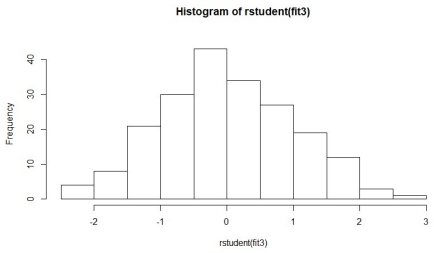
We can plot influence index and influence plots once more to determine any possible outliers. The graphs below show 5 more outliers in the data set. From this we can conclude that the model with the adjusted data values is not as useful as we once thought, since discarding said data gave rise to more anomalies. Moreover, removing more variables can make the data too bias, resulting in an unreliable model.

Finally, we can conclude from the model selection process, that the model found from the first round of tests is the "best" model. That is,

$$Brozek = a + b * neck + c * abdom + d * hip + e * forearm + f * wrist + \varepsilon.$$

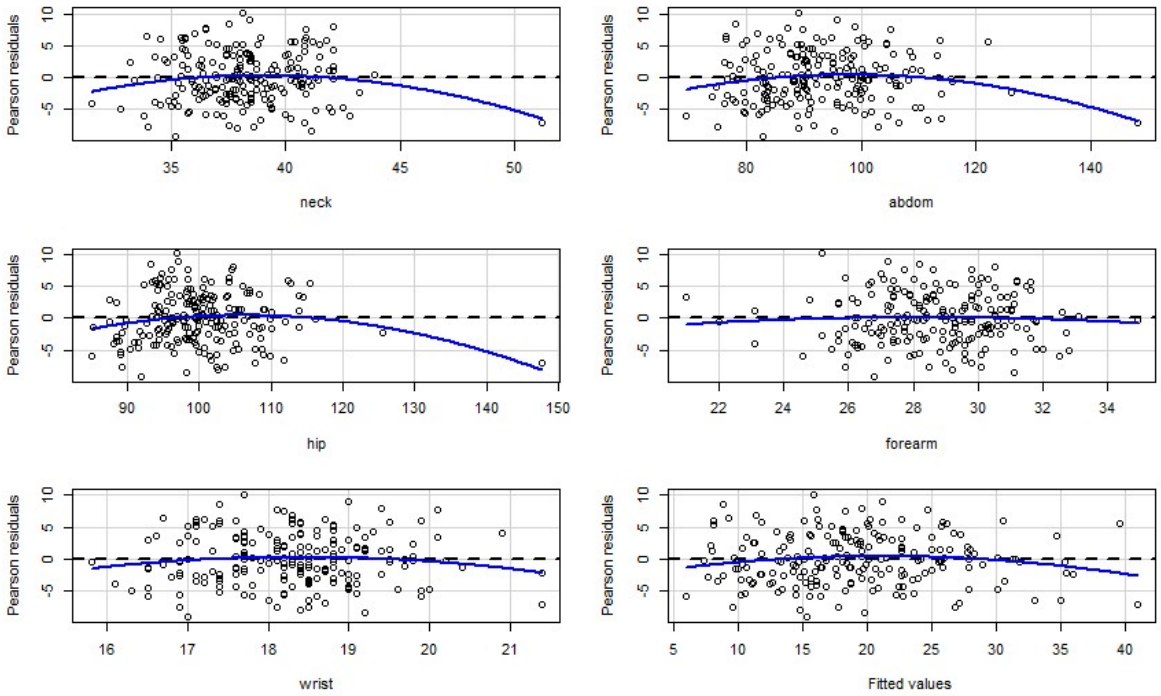
Our next job is to validate this model. We can begin by looking at the QQ-plot and the histogram of residuals for the model.

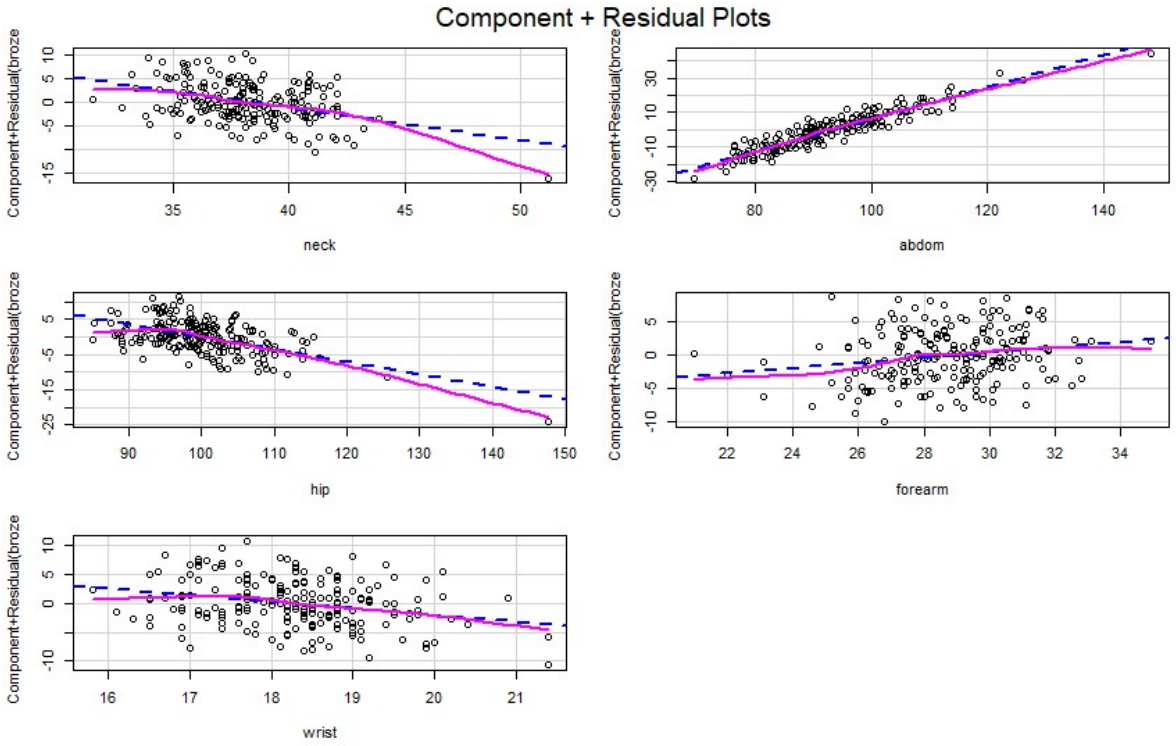




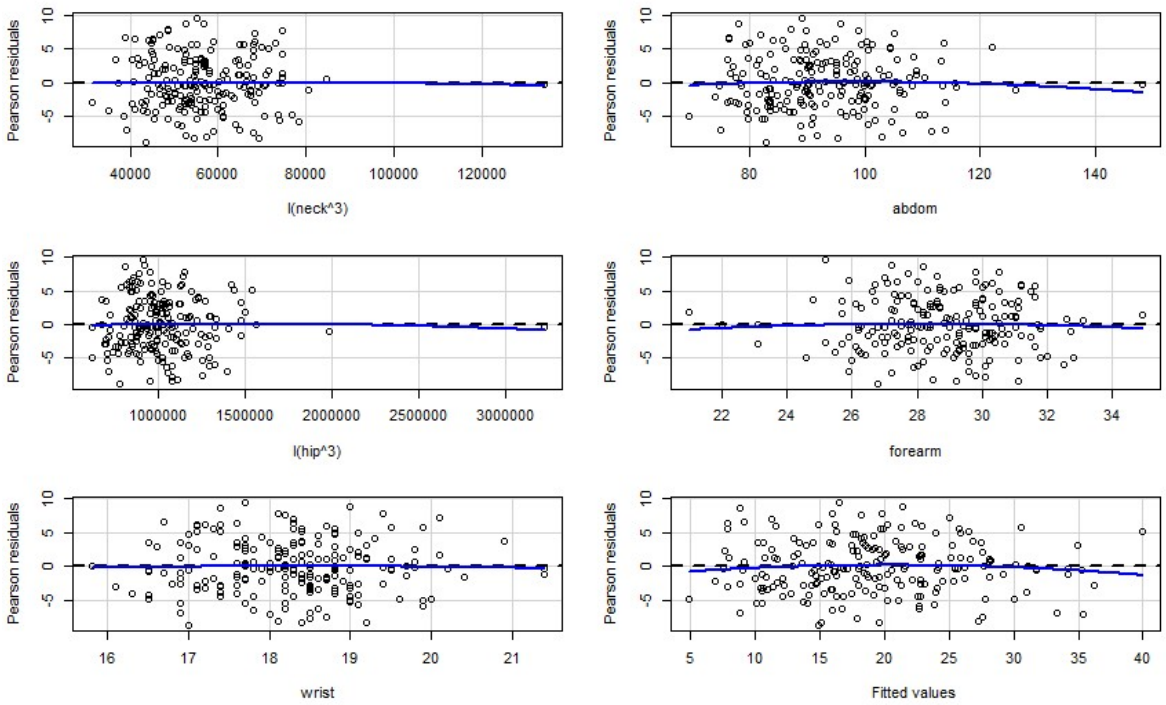
The QQ-plot shows that the model follows the normal distribution, however the histogram of residuals shows a slight positive skew (as did the original histogram for brozek in the scatterplot matrix). This doesn't seem like enough of a skew that requires transformation on the dependant variable, so we can move on.

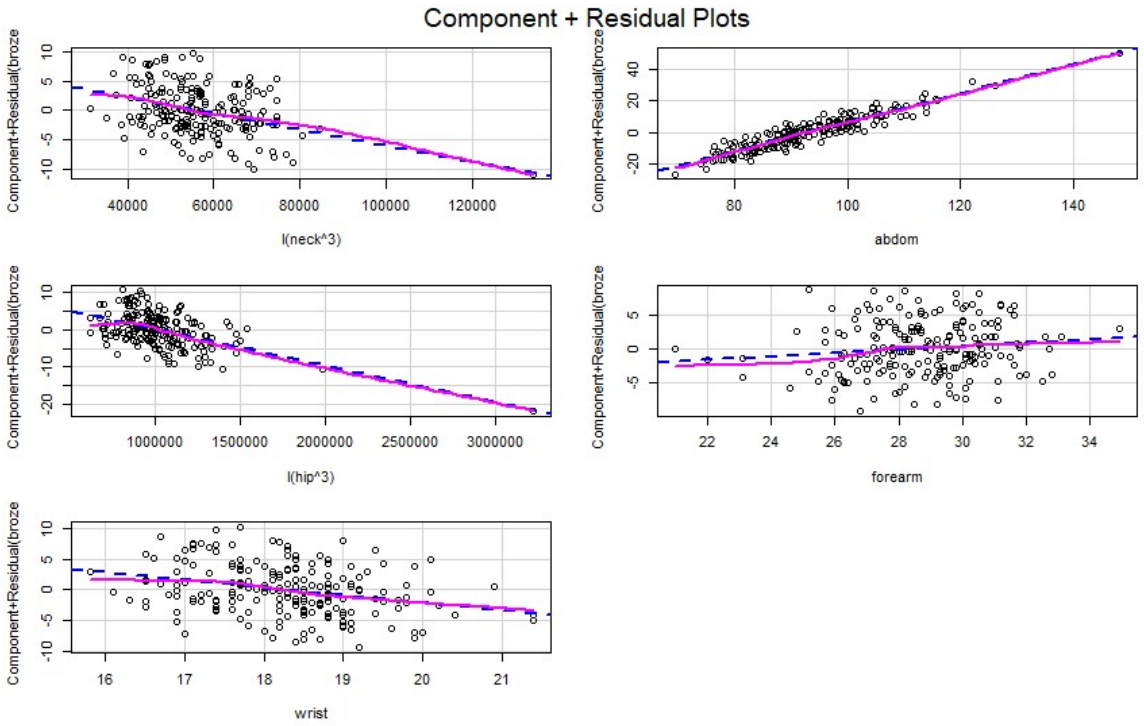
We can now check the residual plots for this model. The two sets of graphs on the next page show some non-linearity, particularly in how the response depends upon neck and hip. We could try transforming these down the ladders of powers, but the relationships appear to be monotone, in which case we will need to fit a quadratic model. We can try transforming up the ladders of power, as suggested by Mosteller and Turkey's bulging rule.


Pearson residuals
neck
abdom
hip
forearm
wrist
Fitted values


Component + Residual Plots
Component+Residual(broze
neck
abdom
hip
forearm
wrist

After several trials, the most suitable transformation was found to be the cubing of neck and hip, which gave a near perfect fit.



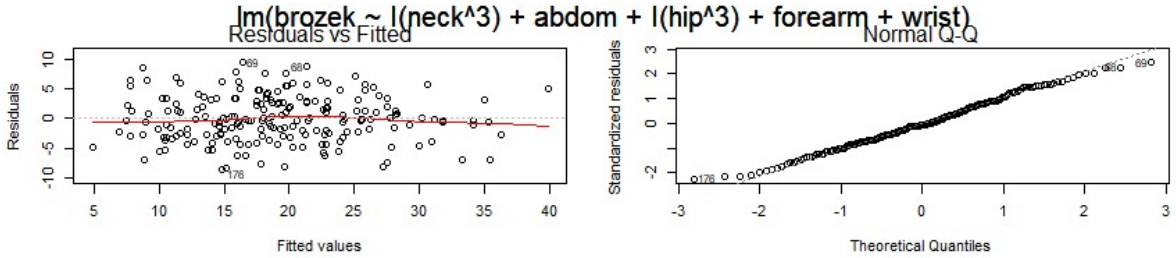


The leftmost graph below shows us that the residuals "bounce" randomly around the horizontal line where the residuals equal 0. We can infer from this graph that the model satisfies the assumption the relationship is linear. Finally, the QQ-Pot demonstrates that model follows a normal distribution, which satisfies the assumption that the residuals are independent and identically distributed and follow a normal distribution. The coefficients of each covariate is determined using the coef() function in R and thus we have,

$$Brozek = -33.6 - 0.000134 * neck^3 + 0.920 * abdom - 0.00000971 * hip^3 + 0.248 * forearm - 1.23 * wrist + \varepsilon,$$

$$where\ \varepsilon\ is\ the\ error$$

As our final model.



We are now ready to test our model. The idea is to use our model to predict the body fat content of these men and see how that compares with the full model. The code below shows us fitting the full model and our "best" model and calculates the mean squared error (MSE) between them. Fitting the simpler model reduces the MSE a significant amount, indicating that the errors have been reduced. We conclude that our model denoted above is the "best" model in predicting body fat.

```
fit1b <- lm(brozek~., data = Train)
fit2b <- lm(brozek ~ I(neck^3)+abdom+I(hip^3)+forearm+wrist, data = Train)

TestResponses=select(Test, brozek)$brozek

predictions1 <- predict(fit1b, newdata=select(Test, -brozek))
predictions2 <- predict(fit2b, newdata=select(Test, -brozek))

MSE1 <- mean((predictions1 - TestResponses)^2)
MSE2 <- mean((predictions2 - TestResponses)^2)

MSE1

## [1] 18.299

MSE2

## [1] 17.45787
```