Kyle Salitrik (kps168)
STAT 462
Homework 4 Report

Applied Analysis

The following reduction was performed on the data in order to remove influential points, outliers and errors.

- Samples below 2% and above 40% body fat were excluded. The minimum was obtained through the chart on the (ACE) website and maximum was set to limit outliers.
- The minimum height accepted was cut to be over 50 inches to remove outliers.
- The maximum weight was also clipped to be below 300 pounds. Although the outlier may have been legitimate, it could have also been an error.
- Observation number 41 was removed as an influential point.

Part A: Backward Elimination

Beginning with the full model below, at every step the predictor with the highest p-value was removed until all predictors were below the threshold of $\alpha = 0.15$.

$$\begin{aligned} y =& \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Weight} + \beta_3 \text{Height} + \beta_4 \text{NeckC} + \beta_5 \text{ChestC} + \\ & \beta_6 \text{HipC} + \beta_7 \text{ThighC} + \beta_8 \text{KneeC} + \beta_9 \text{AnkleC} + \beta_{10} \text{BicepsC} + \\ & \beta_{11} \text{ForearmC} + \beta_{12} \text{WristC} + \beta_{13} \text{Over45} + \epsilon \end{aligned}$$

Through this process, the following predictors were removed in the order presented:

- Knee Circumference
- Thigh Circumference
- Weight
- Ankle Circumference
- Biceps Circumference
- Chest Circumference
- Forearm Circumference
- Hip Circumference

The final model obtained from backwards elimination is:

$$y = \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Height} + \beta_3 \text{NeckC} + \beta_4 \text{WristC} + \epsilon$$

Part B: Forward Selection

Starting with zero predictors, 13 models were generated (each with one of the 13 predictors) and the p-values of the new predictors were saved off. The predictor with the lowest p-value was added to the real model and the process was repeated for the 12 remaining predictors. This continued until any additional predictors had a p-value of greater than 0.15.

The following presents the order in which predictors were added:

- Abdomen Circumference
- Weight

- Wrist Circumference
- Height

The model below was then fit based on this selection:

$$y = \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Weight} + \beta_3 \text{WristC} + \beta_4 \text{Height} + \epsilon$$

Part C: RSS, Mallow's C, AIC, and BIC Selection

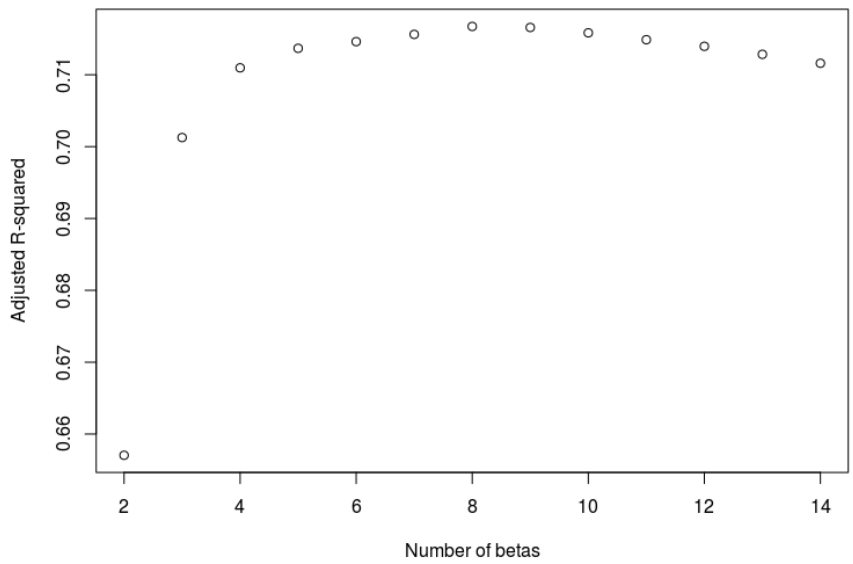
The "leaps" R package was used to exhaustively search for the best models at a set number of predictors. These results are displayed below (with the intercept term removed from the output) and were used to determine the predictors present in the corresponding models.

	AbdomenC	Weight	Height	NeckC	ChestC	HipC	ThighC	KneeC	AnkleC	BicepsC	ForearmC	WristC	Over45
1	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
4	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
5	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
6	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
7	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
8	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE
9	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
11	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
12	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
13	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Listing 1: regsubsets Function Output

Beginning with the method of model selection using adjusted R-squared values, the plot below shows that the highest value occurs at 8 betas, or 7 predictors.

Figure 1: Adjusted R^2 Plot

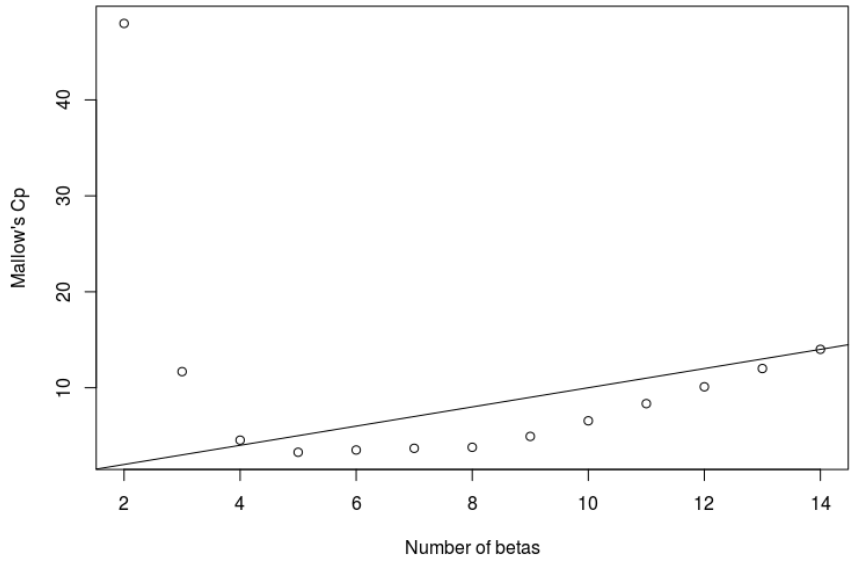


Using the regsubsets output and this information, we can obtain the following model:

$$y = \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Height} + \beta_3 \text{NeckC} + \beta_4 \text{ChestC} + \beta_5 \text{HipC} + \beta_6 \text{ForearmC} + \beta_7 \text{WristC} + \epsilon$$

The second method of model selection explored in this section is using the Mallow's C values for a given number of predictors to determine the best model. When looking at the differences between the C values and the number of predictors, the closest fit is at C_3. This indicates that 4 betas (or 3 predictors) is the best model based on this method. The following plot contains the Mallow's C values for up to 13 predictors.

Figure 2: Mallow's C_p Plot

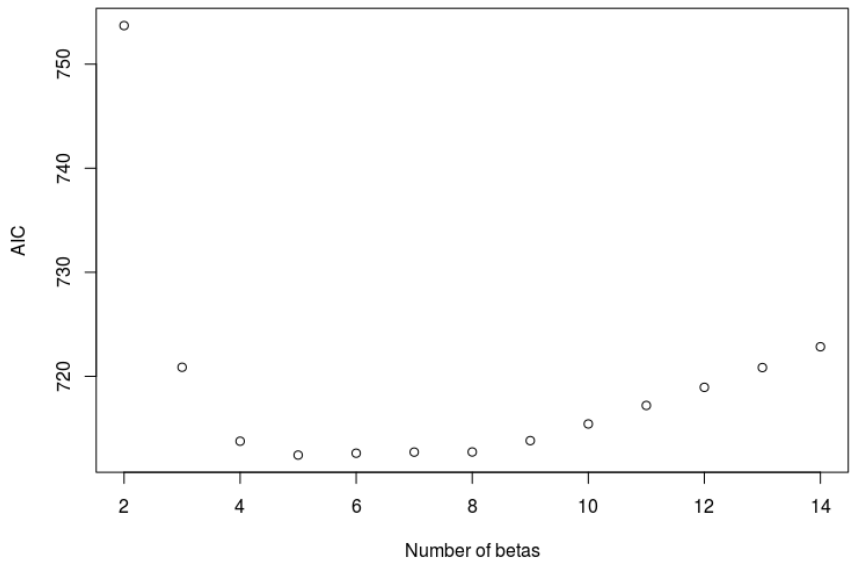


Once again, using the regsubsets information the linear model below is obtained:

$$y = \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Weight} + \beta_3 \text{WristC} + \epsilon$$

For the third method, AIC, the minimum value indicates which model is the best fit according to the metric. The plot of values and linear model can be seen below, with $p = 4$. Predictors in the model were obtained from the regsubsets function.

Figure 3: AIC Plot

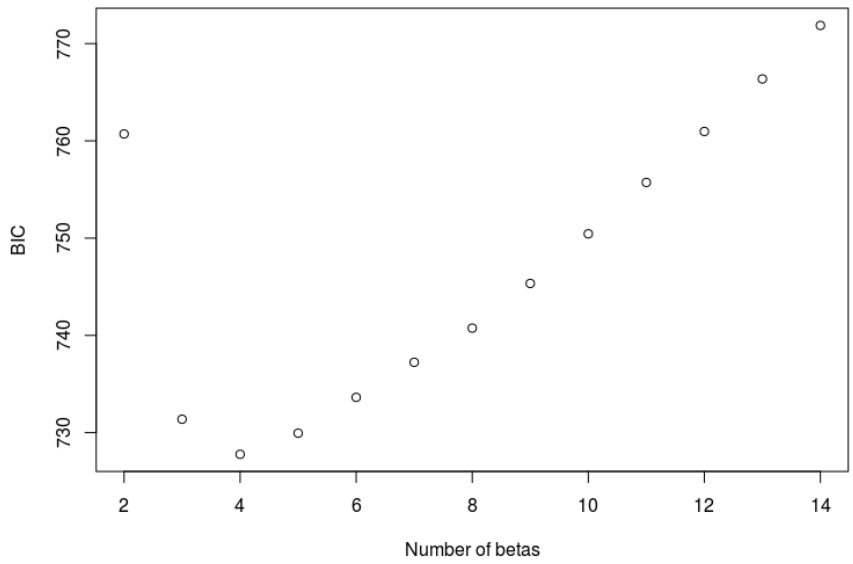


It is worth noting that the best model for $p = 4$ is identical to the model obtained by backward elimination.

$$y = \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Weight} + \beta_3 \text{NeckC} + \beta_4 \text{WristC} + \epsilon$$

Finally, the execution and interpretation is similar to that of the AIC method in that the minimum value of the results corresponding to a given number of predictors indicates the number of predictors for the best fit. The number of predictors obtained here is the same as the Mallow's C values, and is therefore the same model.

Figure 4: BIC Plot



$$y = \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Weight} + \beta_3 \text{WristC} + \epsilon$$

Part D: Model Evaluation and Refinement

VIF

At this point four unique models exist that are shown below.

$$\begin{aligned}
y_{BWD} &= \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Height} + \beta_3 \text{NeckC} + \beta_4 \text{WristC} + \epsilon \\
y_{FWD} &= \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Weight} + \beta_3 \text{WristC} + \beta_4 \text{Height} + \epsilon \\
y_{RSS} &= \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Height} + \beta_3 \text{NeckC} + \beta_4 \text{ChestC} + \beta_5 \text{HipC} + \\
&\quad \beta_6 \text{ForearmC} + \beta_7 \text{WristC} + \epsilon \\
y_{BIC} &= \beta_0 + \beta_1 \text{AbdomenC} + \beta_2 \text{Weight} + \beta_3 \text{WristC} + \epsilon
\end{aligned}$$

The first step taken in the model evaluation was to examine the variance inflation factors (VIF) for each unique model to determine whether or not colinearity issues are present.

The forward elimination model (y_{FWD})contains two predictors with high VIF values: AbdomenC and Weight. In this model Weight has a VIF value of ≈ 9.78; with a value near 10 it indicates issues with colinearity and was removed.

In the model chosen by Adjusted R^2, AbdomenC has a VIF value of ≈ 7.6, which is significantly higher than all of the other VIF values. For this reason it was removed from the model.

The model chosen from Mallow's C and BIC contains two predictors with moderately high VIF values. For this model, AbdomenC is ≈ 4.03 and Weight is ≈ 5.28. Logically, it makes sense that there would be colinearity between these two predictors, but further investigation is necessary.

After adjusting the models and re-evaluating the VIF values, the models chosen by Adjusted R^2 and forward elimination both have VIF values under 4 for all predictors.

Diagnostic Information

Examining the summaries for each model, the model chosen from the Adjusted R^2 method has a significantly lower adjusted R^2 value than any of the other models after removing the weight as a predictor (≈ 0.54 vs ≈ 0.71). The other three models have very close adjusted R^2 values. This information is displayed below.

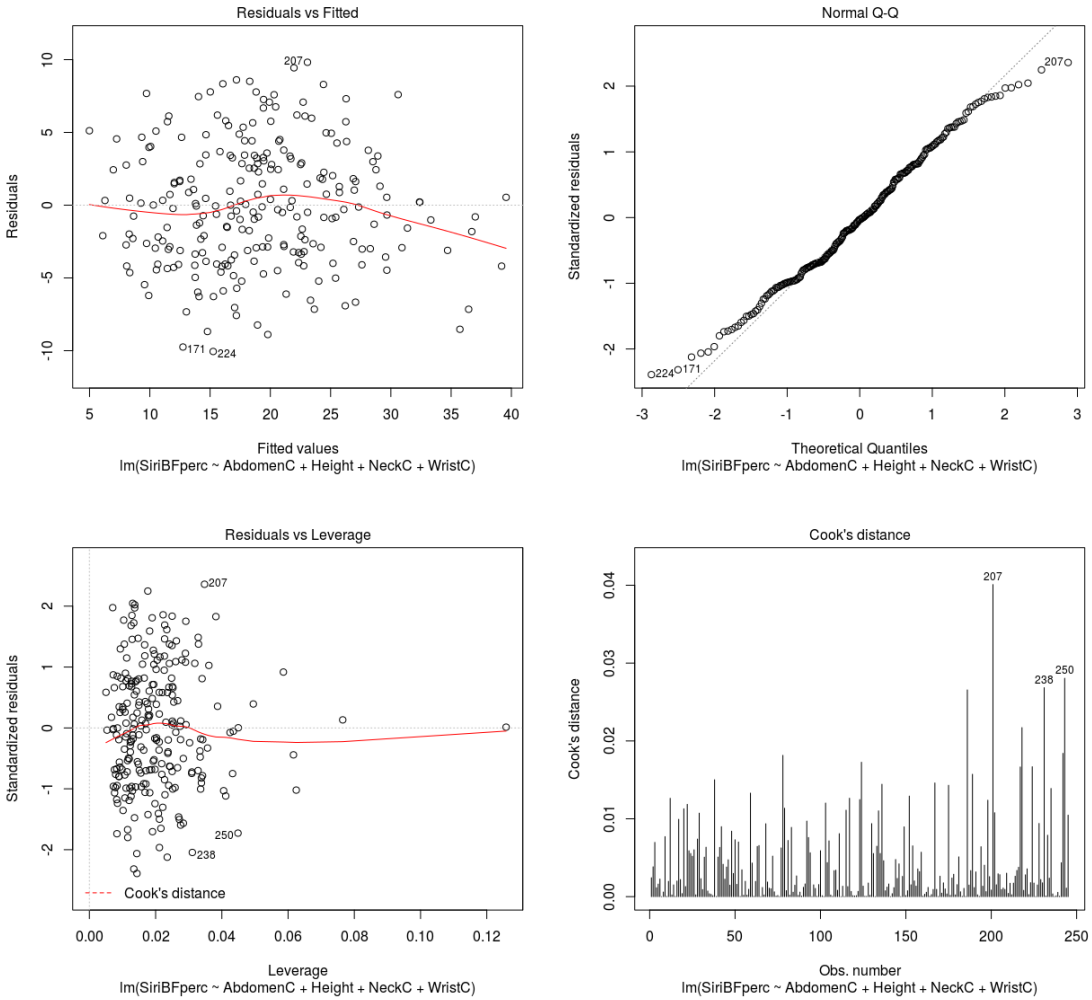
```
   lm_rss     lm_fwd     lm_bic     lm_bwd
0.5354380 0.7101028 0.7109660 0.7136836
```

Listing 2: Adjusted R^2 Values

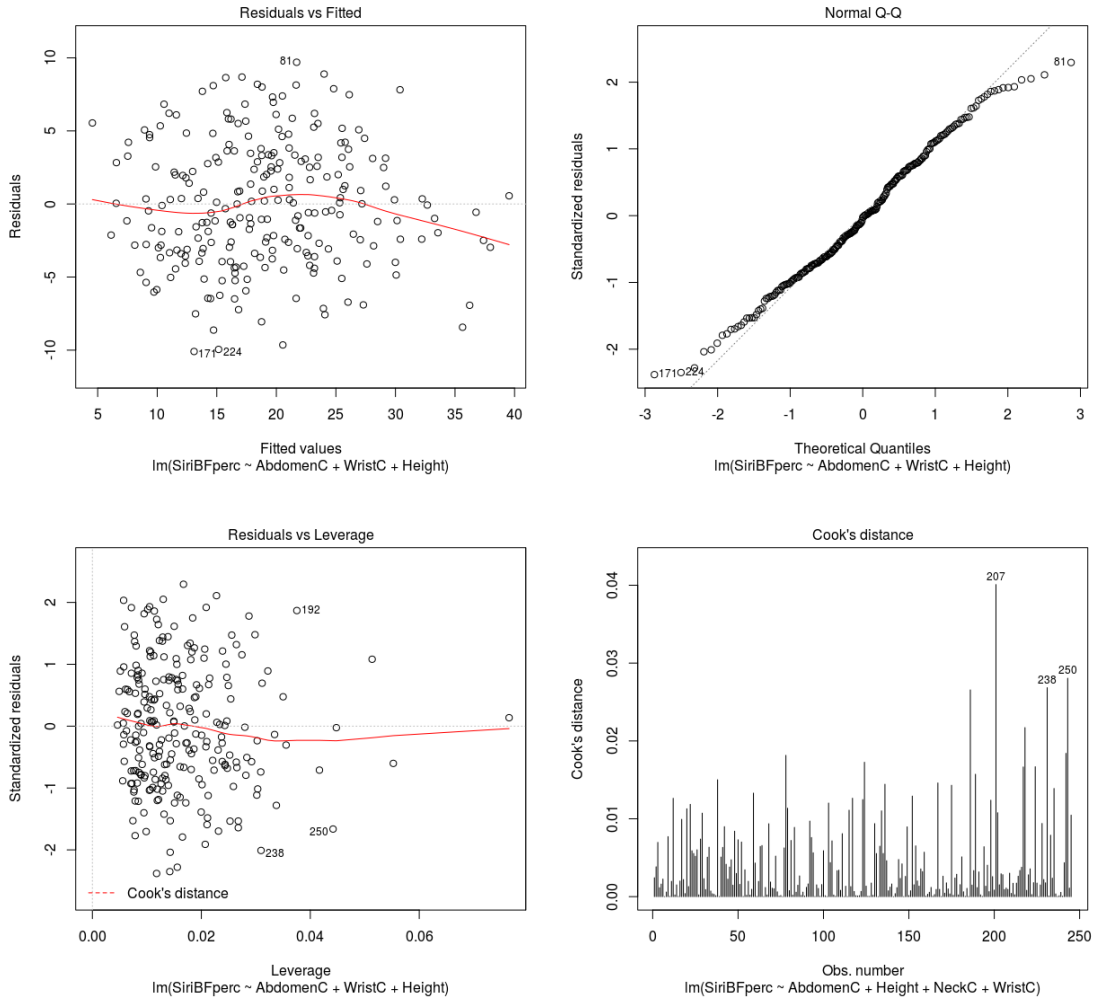
With the similarities between the remaining models combined with the potential colinearity issues in the BIC model, the decision was made to remove this model from consideration. If the model had a significantly higher (or even the highest) adjusted R^2 value out of the remaining models, further investigation would be considered. However, the forward and backward elimination models have VIF values less than 3 for all of their predictors.

Below are two sets of plots containing (from top right to bottom left): Residuals vs Fitted values, the Q-Q plot of the residuals, Residuals vs. Leverage, and the Cook's distance. The first set of plots is for the backward elimination model and the second set is for the forward selection model.

For the backward elimination model, there do not appear to be any major trends that occur in the residuals vs fitted values. Towards the higher fitted BFP values, there appears to be a reduction in variance. The Q-Q plot shows that the left and right tails are slightly light, but not extreme. Likewise, the standardized residual plot does not show any points with extreme values (absolute values larger than 3). The cooks distance plot shows that there are a few points with higher distances relative to the rest of the observations.



The diagnostic plots for the forward selection model show nearly identical information as the backward elimination plots.



Prediction Intervals

Due to the high similarities in the diagnostics of the two models, both models will be used to generate prediction intervals for body fat percentage. Either model may be appropriate depending on the intended usage. To compute a prediction interval about the means, the following values were obtained as the mean for every quantitative predictor used in the model:

```
AbdomenC    Height    NeckC   WristC
92.16245  70.33367  37.9302  18.2102
```

Listing 3: Mean Predictor Values

The following prediction intervals were obtained using the given mean values:

```
             fit       lwr       upr
lm_bwd  19.03347  10.67003  27.39691
lm_fwd  19.03347  10.61807  27.44887
```

Listing 4: Mean Predictor Values

Final Models

After refining the models, the following are the final model estimates

$$\hat{y}_{BWD} = 10.64971 + 0.81416\text{AbdomenC} - 0.42159\text{Height} - 0.40964\text{NeckC} - 1.17854\text{WristC}$$

$$\hat{y}_{FWD} = 8.8642 + 0.7719\text{AbdomenC} - 1.6304\text{WristC} - 0.4448\text{Height}$$

As we can see in both models a larger abdomen circumference tends to show an increase in body fat percentage, while height and wrist circumference appear to indicate a lower body fat percentage. For the backward selection model, larger neck circumference also appears to indicate a lower body fat percentage.