**Analysis of Predictors for Body Fat Percentage**

**Executive Summary**

**Shawn Chapler**

**Background**

Body fat percentage is a factor known to affect athletic performance. Monitoring changes in body fat percentage is often used as one method of predicting and optimizing athletic performance in various sports. Distance running and cycling are examples where lower body fat levels can have a significant positive impact on overall performance while sports requiring more “power” such as power lifters and baseball players tend to perform better with higher body fat levels. Athletes, trainers, scientists, and coaches expend time analyzing optimal body fat levels and its correlation with performance. This analysis aims to identify those factors that provide the best estimate of body fat percentage.

**Study**

This study is performed on data collected at the Australian Institute of Sport (AIS), on 102 male and 100 female athletes. The data was analyzed to derive a statistical model that optimized predictive capabilities of the independent variables and body fat. The model selected predicts body fat percentage given the available variables and identifies those that are most reliable in estimating body fat. Additional details on the study are found in the following appendix.

The data contains the following variables which were assessed in their ability to predict percent body fat. The variables considered were: sex, height, weight, lean body mass (LBM), red blood cell count, white blood cell count, hematocrit, hemoglobin, plasma ferritin concentration, body mass index (BMI), sum of skin folds (SSF), and one of ten sports the athlete participated in. It should be noted that while sport participation may correlate with body fat percentage, this in and of itself lacks causation. As such, body fat was not considered as a predictor variable in this study.

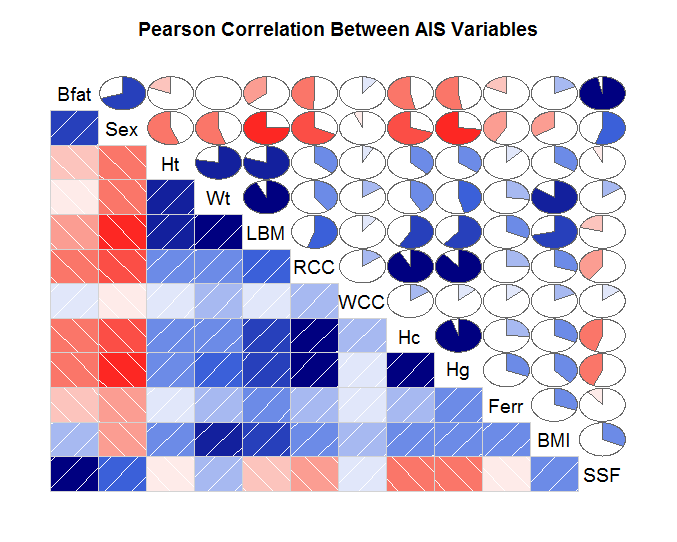
The data was analyzed for normality, variance, and collinearity to ensure the appropriate models were selected for consideration. The below graph reflects relationships amongst pairs of predictor variables as well as body fat. The graph shows very strong relationships between SSF and sex with body fat as evidenced by the dark blue pie charts in the graph. Additionally, there is evidence of collinearity amongst the predictor variables such height, weight, LBM, and BMI. This is not surprising as weight is used in the LBM and BMI respective calculations. 

Figure 1: Correlation between Pairs of Variables

From this analysis, two types of models were selected: multiple linear regression and random forests. Multiple linear regression as graphical plots indicated a linear relationship was likely to exist with certain predictors in the data set. Random forests modeling was selected for its ability to better estimate new data points when certain predictors, such as SSF or sex, have a stronger relationship with body fat than others predictors such as hemoglobin.

Four multiple linear regression models were selected each with different combinations of predictor variables shown to have a relationship with body fat percentage. Two random forest models were used one considering all eleven variables at each decision point and one considering six.

**Findings**

Based upon the best fitted model, the most important variables are the sum of skin folds and sex in estimating body fat percentage. This suggests sum of skin folds test is a superior method to estimate body fat relative to other available metrics such as lean body mass, body mass index, height, or weight.

**Appendix**

**1.0 Data Study**

Various statistical methods were deployed on the dataset to identify relationships between body fat and the available predictor variables. Several models were considered.

The first step involved reviewing the data for completeness, sample size for variables with factor levels, normality, variance, outliers, and significance of each variable’s relationship with body fat.

**1.1 Variable Selection: Removal of Sport**

In the overall analysis of variables, it was observed the sport variable contained ten levels ranging in sample sizes of 4 to 37. While certain sports displayed a correlation with body fat, the relationship was not particularly strong. It must be noted that while a correlation may exist in that many athletes participating in a particular sport may have similar body fat measures, sport in of itself is not a factor that can be reliably used to estimate body fat percentage in an individual.

**1.2 Normality/Variance/Outliers**

The predictor variables were individually analyzed through use of plotting for normality and equal variance. Most variables had outliers with SSF showing the most significant skew. A log transformation was done on SSF to create a more normal distribution.

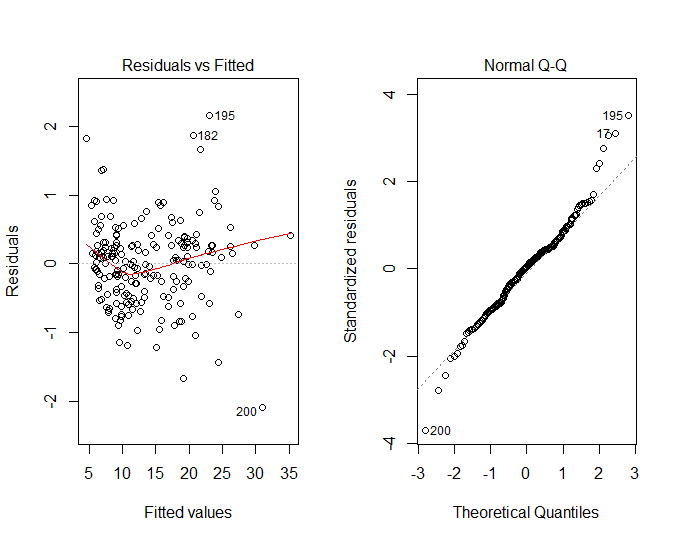
The Normal QQ charts reflect a general linear relationship, but there is skew at both ends, which appears to be the result of outliers in the dataset as evidenced in the boxplots. Residuals are also generally normally distributed with the exception of outliers and potential heteroskedacity as evidenced by the wider variance of the residuals versus fitted toward the right of the plot.

Figure 2: Normal QQ Plot and Fitted vs Residuals

**1.3 Variable Importance/Multicollinearity**

From this analysis, the following variables were found to have strong relationships with body fat: sex and SSF as shown in the chart below. Additionally, there is high correlation found between weight, height, LBM, and BMI. Blood composition variables as a class have a weak correlation with body fat composition, but high correlation with each other as evidenced in Figure 1.

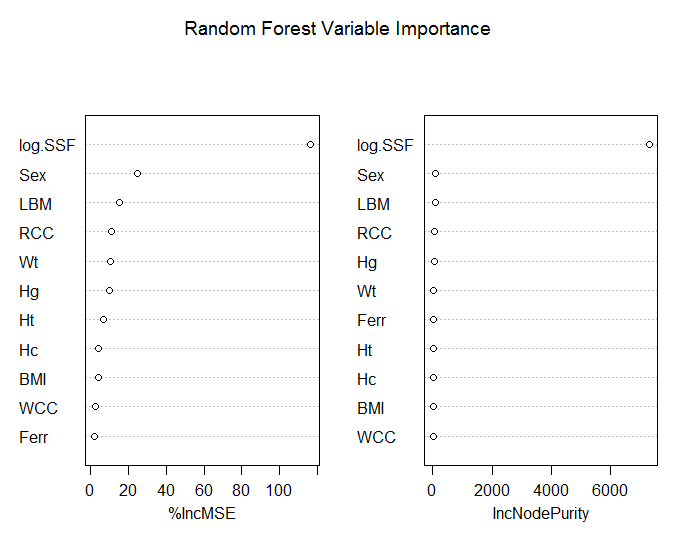


Figure 3: Variable Importance Random Forest

**2.0 Model Selection Methodology**

Two primary models were selected to predict body fat.

1. Multiple linear regression was selected given the general linear relationship between body fat and the stronger predictor variables.
2. The random decision forest model was selected given there are strong correlations between predictor variables (height, weight, BMI, LBM) and several predictors are more informative about the response variable than the others such as SSF and sex.

While body fat also does not have a normal distribution, the use of the box cox test indicated a lambda of 1 when SSF was not included in the set and a lambda of .71 when it was included. As such, no transformation was made pending variable selection through formalized variable selection processes.

**2.1 Regression Models**

Regsubsets and stepwise regression in both directions were used to select the variable combinations that resulted in the lowest error values. Figure 4 reflects the lowest MSE with five variables, however models with three to eleven predictors produced MSE’s within one standard deviation. Both the optimal model with five variables and the simplest model with three variables were selected in addition to the full model of all available predictors.

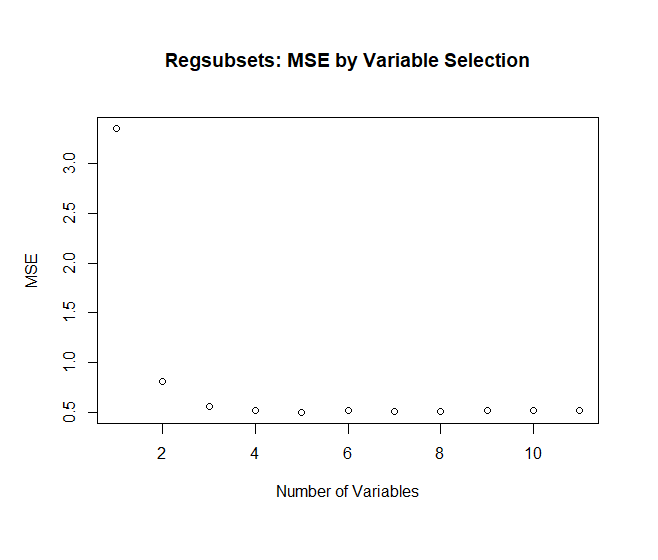


Figure 5 Model MSE through Regsubsets

Model 1: All variables which maximizes R2 and Adjusted R2

Sex+Ht+Wt+LBM+RCC+WCC+Hc+Hg+Ferr+BMI+log.SSF

Model 2:

Several methods were employed to identify predictors to be included for the multiple regression models. Stepwise regression and regsubsets produced a singular model that minimizes AIC, BIC, and Mallows CP.

Sex+Ht+Wt+LBM+BMI+log.SSF

Model 3:

This model was selected through the regsubsets selection process as it was the model with the minimum number of predictor variables within one standard deviation of the mean square error (MSE) of the optimal model. Analysis of variance indicated the model was not significantly different than full model.

Sex+Ht+Wt

Model 4:

This model was selected after performing Breusch-Pagan tests for heteroskedacity which is known to exist amongst the predictor variables. This is a model where variances are equal.

Sex+Wt+BMI+log.SSF

**2.2 Random Forest Models**

Model 5:

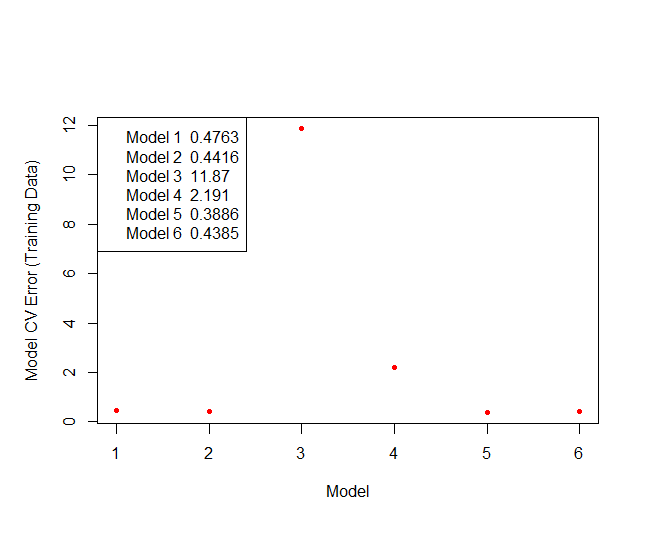
Random forest with eleven predictors considered at each branch (mtry=11) where the importance of predictors is considered.

Model 6:

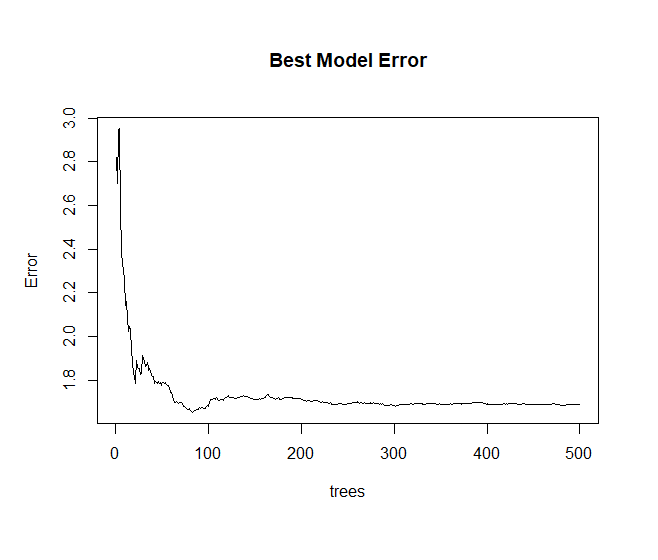
Random forest with six predictors considered where the importance of predictors is considered. Six was selected after applying the tuning analysis which shows the out of bag error dropping significantly until six. Thus six was selected over the model default of p/3.

**3.0 Double Cross Validation for Model Selection and Assessment**

Double cross validation was applied on all six models. The best model was selected based upon the model that minimized the tenfold CV error rate. In general, the random forest models performed better than the regression models as shown in Figure 6.

Figure 6 Cross Validation Error Comparison on Training Data

The final fit of the random forest model with 11 predictor variables at each split against the full AIS dataset resulted in a CV value of 0.3126825 and an R2 value of .941471. Figure 7 shows the decreasing and flattening of the error rate with the number of trees added.

Figure 7: Random Forest Best Model Error Rate

**4.0 Conclusion**

The full random forest model is superior in predicting body fat composition relative to the other models given the available variables. The variables sex and sum of skin folds test have stronger predictive value in estimating body fat percentage.