DS 740 Midterm Project Spring 2019

By Matt Allen

The dataset that was chosen for the Midterm project was the Australian Institute of Sport Data set. The quantitative response was chosen, which in this dataset is Body Fat. Body fat was chosen for the current study, but the qualitative factor response could also be an interesting investigation like finding what body measurements correlate to different sports for example. The Body Fat response is important, because it is difficult and expensive to measure body fat directly. Finding easy measurements that predict body fat is critical where body fat cannot be measured directly. The method developed in this analysis could be used by athletic trainers to predict an athlete's body fat based on easily measured quantities.

The model should be as simple as possible, but still provide enough predictors to give an accurate prediction of body fat. Having a simple model is critical to make measurements as easy as possible for the athletic trainer. A simple linear regression model could be used, so that the trainer can easily plug her measurements into a formula to calculate the athlete's body fat.

A histogram of the Body Fat response was created. Based on the histogram, there was visual evidence that the response was not normally distributed. This was further confirmed by a Shapiro test for normality. Based on the right skewness of body fat, the body fat was log transformed. Double cross validation was used to select and validate models using lasso and linear forward step model. A simple model was selected that just uses the athlete's sex, weight in kg and sum of skin folds. This is a good model, because it consists of quantities that are easily measured, and all predictors were significant at a 5% level.

In Figure 1, the diagnostic plots for the initial model are displayed. It can be seen that rows 200, 201, 202 in the data set are outliers. A good technique to emphasize these outliers less without throwing away data from an already small data set is to use the bisquare method. The regression was performed again using the bisquare method. The points in rows 200,201, and 202 were weighted down to 0.029, 0.005, and 0.00 respectively.

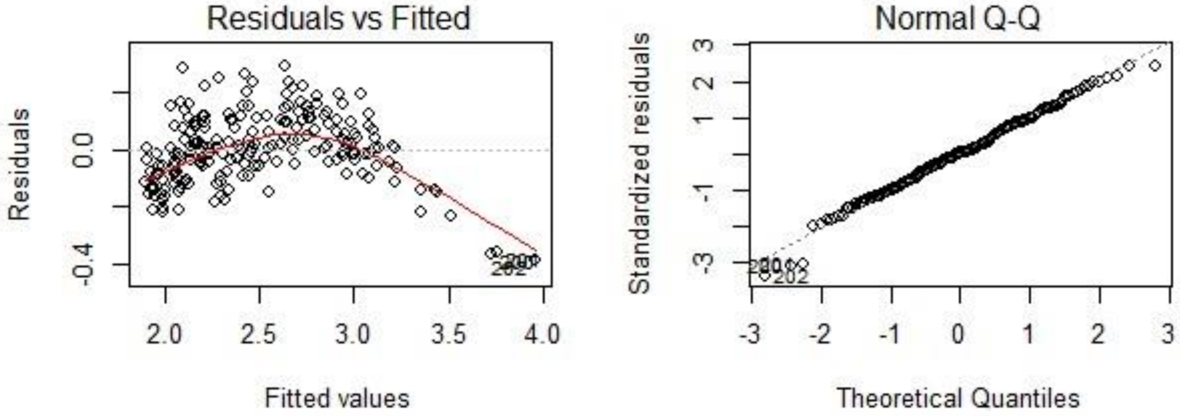

Figure 1. Diagnostic Plots of initial model.

The final model after using the biquare method is shown in Figure 2. Again, it has the benefit that the measurements are relatively easy and inexpensive to do for an athletic trainer. Sum of skin folds requires more expertise to perform correctly, but the tools to measure are inexpensive. Weight measurements simply require a scale.

$$log(BodyFat) = 1.403 + 0.329 * SexFemale + 0.002 * Wt + 0.011 * SSF$$

Figure 2. Body Fat Equation.