**Exam 1**

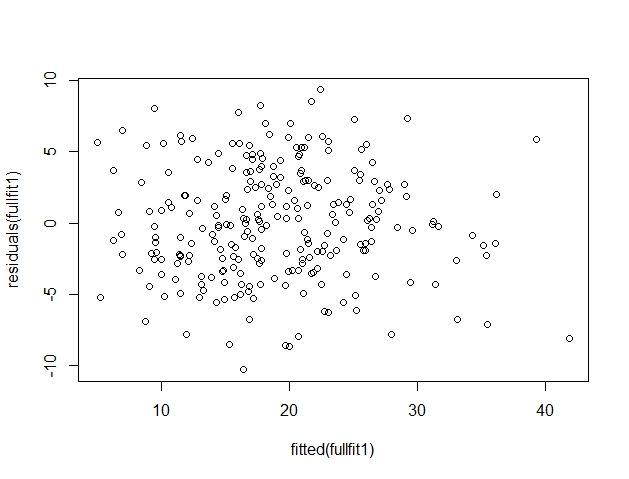
Luyao Zhang (NetID: lzhang94)

**Regression Model Selection Problem**

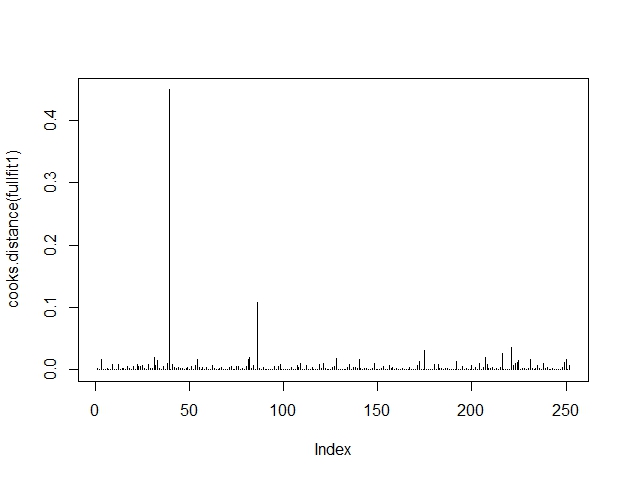
First I corrected the error in the data by changing the height for Case 42 from 29.5 to 69.5. Then, I first checked the diagnostics plots for the data in order to detect potential outliers that may need to be deleted from the data set prior to any further analysis.

To do the diagnostics, a linear regression model was fitted using all predictors (except for densities and siri). The residuals vs. fitted values plot was obtained (Figure 1), and is shown below. Judging from the spread of the data points, the residuals generally seem to spread evenly around the 0 line, and there doesn’t seem to be any case that stands out from the basic random pattern of the residuals, which indicating the possibility of no outliers.

However, to figure out more accurately if there are any outliers, I obtained the Cook’s Distance for all the cases in the data set, and the plot is shown as Figure 2. Most of the Cook’s Ds look fine, except a case where the Cook’s D is much bigger than all the other cases (almost as large as 0.5), and turns out this is Case 39. Therefore, Case 39 is removed as an outlier, which leaves 251 cases in the data set.



**Figure 1. Residual vs. Fitted values**

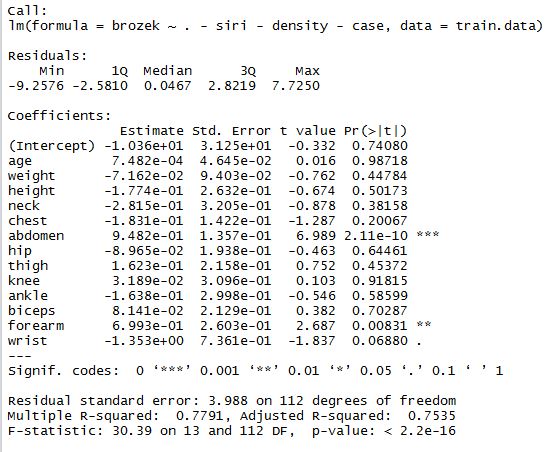


**Figure 2. Cook’s Distance**

Then the data set was split evenly into a training data set and a test data set. The training data set has 126 cases while the test data set has 125 cases, because the number of cases in the whole data (i.e., 251) set is an odd number.

1. **The full model:**

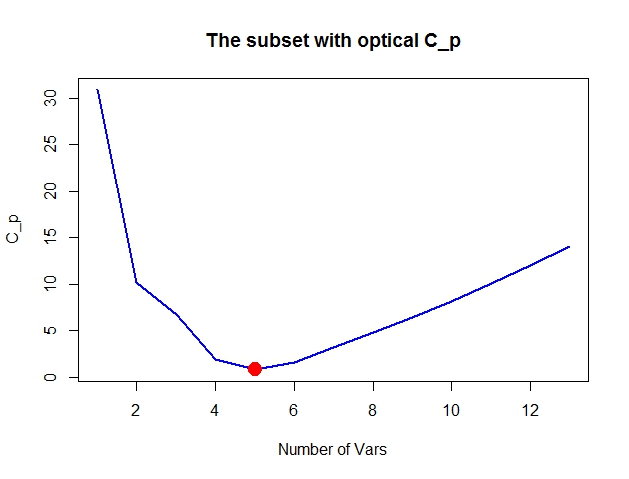
First of all, a linear model is fitted using the training data set with all predictors except for densities, siri, and cases, and the summary of the model can be found below (Figure 3). It looks like “abdomen” and “forearm” are the only predictors that are statistically significant at 0.05 levels in predicting “brozek”.



**Figure 3. Summary of the linear model using the training data**

1. **Best subset selection based on the optimal** 

Next, I tried to find the best subset of predictors for the dependent variable based on. The results support a 5-predictor model, with 5 predictors being respectively “height”, “chest”, “abdomen”, “forearm”, and “wrist”. Note that both predictors that were found significant using the linear regression model have also been included here.

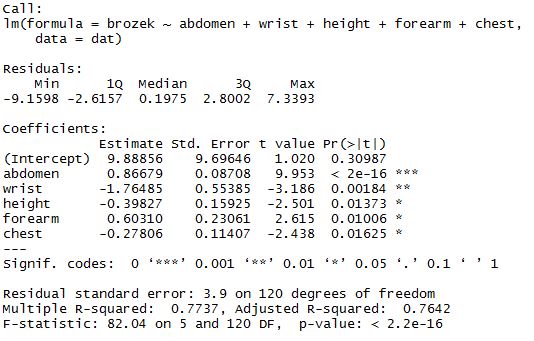


**Figure 4. Best subset selection based on the optimal **

1. **Forward Selection based on AIC**

The forward selection method was used to select the best subset of predictors for the model to predict the dependent variable. The criteria used was the AIC (or the F values <3). This means each time the model is fitted, the predictor with the smallest AIC will be included in the model, and the selection won’t stop until the all F-values of the remaining predictors are below 3. The final model selected using forward selection method is shown below (Figure 5).

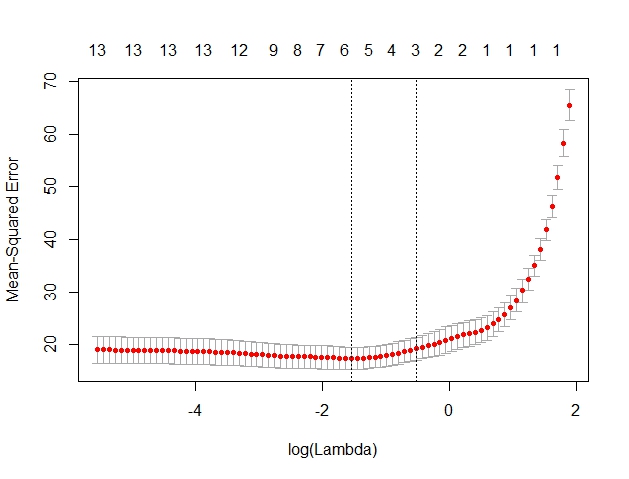
The final model obtained here is exactly the same as what was obtained from the best subset with .



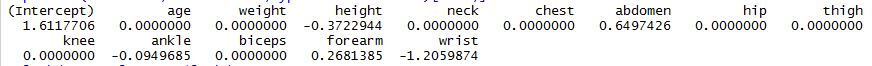
**Figure 5. Final model using forward selection**

1. **Lasso with a well-chosen lambda.**

Lasso was fitted, and the MSE of using different sizes of predictors can be found below (Figure 6). The minimum of lambda is 0.2139. Using the minimum of lambda, a 5-predictor model is selected, and the predictors are “height”, “abdomen”, “ankle”, “forearm”, and “wrist”, and their coefficients can be found in Figure 7. This model selected is partially different from the model selected with optimal  or the forward selection method. The predictor that got left out with Lasso is “chest”, and the predictor that was included in here but not by the previous two models is “ankle”.



**Figure 6. MSE vs. Log (Lambda)**



**Figure 7. Coefficients of the 5-predictor model**

1. To compare the 4 models above, MSEs were computed, and since Model 2 () and Model 3 (forward selection) are the same, only one MSE was computed for the two. According to R, the MSE for the full linear model with all the predictors except for densities, siri, and case has yielded MSE equal to 17.33. The model with the best subset selected with and the forward selection method have obtained MSE equal to 17.44. Lasso has the largest MSE, which is equal to 16.72.
2. According to the results above, in terms of MSE, Lasso gave the best model, as it has the smallest MSE, which is equivalent to having the largest adjusted R^2 (i.e., the proportion of variation in the dependent variable explained by the predictors is the largest under Lasso, compared with the other 3 models).

**Classification Problem**

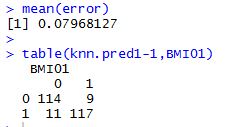
According to MSE obtained from the first section, LASSO supports the most informative model (the model has the smallest MSE), which has “height”, “abdomen”, “ankle”, “forearm”, and “wrist” as predictors. Therefore, the classification cross-validation analysis will be based on this model.

First, a new variable BMI was created using the formula given in the question, and was combined with the original data set. Then a binary variable BMI01 was created, where 1 stands for when BMI > 25 and 0 stands for when BMI is smaller or equal to 25.

1. **K-nearest neighbors**

K = 10 was selected because normally K is between 3 and 10, and according to previous HW assignments, this value of K works the best compared with smaller values such as 1 and 5, or larger values such as 50, and 100. The test error rate for classifying is 7.97%, which means the percentage of correct prediction is 92.03%.

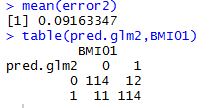
The confusion table is shown below (Figure 8). Based on the table, this classifier did similarly good jobs predicting both negative (not overweight) and positive (overweight) results. The sensitivity (true positive) of the classifiers is 117/ (117+9) = 92.86%, while the specificity (true negative) is 91.2%. In this sense, when the person is overweight (i.e., BMI > 25), the model works slightly better predicting than when the person is not overweight.



**Figure 8. Test error rate and the confusion table for KNN**

1. **Logistic regression**

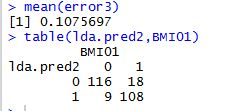
A logistic regression model was fitted and the cross-validation was carried out. The test error rate obtained was 9.16%, which means that the percentage of correct prediction is 90.84%. The confusion table (Figure 9) showed that the model performed almost equally well in terms of predicting positive and negative results. The sensitivity of the model is 114/ (12+114) = 90.48% and the specificity is 114/ (114+11) = 91.2%.



**Figure 9. Test error rate and the confusion table for logistic regression**

1. **LDA**

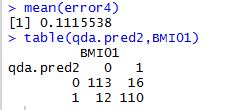
LDA and cross-validation were carried out, and the results are as shown in Figure 10. The test error rate was 10.76%, which means that the percentage of correct prediction is 89.24%. The confusion table showed that the model performed much better when predicting negative results, with the specificity being 116/ (116+9) = 92.8%. When predicting positive results, however, the model did a much poorer job, with sensitivity being 108/ (18+108) = 85.71%.



**Figure 10. Test error rate and the confusion table for LDA**

1. **QDA**

QDA and cross-validation were conducted, and the results are shown in Figure 11. The test error rate is 11.16, which means the model predicts correctly 88.84% of the time. The model performed better when predicting negative results than positive. The sensitivity of the classifier is 113/ (12+113) = 90.4% and the specificity is 110/ (110+16) = 87.30%.



1. **Comparison & Discussion**

Considering the overall test error rate, as well as the sensitivity and specificity of the model, K-nearest neighbors with K=10 did the best job in terms of classification.

Logistic regression method has the second best test error rate, and performed almost equally well predicting negative and positive results (sensitivity is the same as specificity). Logistic has the same specificity as KNN, but a lower sensitivity.

LDA actually has the best specificity among all 4 methods. However, LDA is the least sensitive among the 4 methods, and also has overall error rate that is higher than KNN and logistic regression. Therefore, if we are concerned about detecting overweight people or classification accuracy as a whole, LDA is not a good choice, but if we only care about correctly detecting those who are not overweight, LDA should be considered.

QDA, with the highest error rate among the 4 methods, as well as unimpressive sensitivity nor specificity, seems to be worst method among the four.

To conclude, according to results obtained from the analysis above, KNN with K=10 is the best classifier, because it has the lowest overall error rate and the highest sensitivity, which is usually what we are the most interested in. LDA may be a good choice if our focus is only on correctly classifying participants under the BMI < 25 category. Logistic regression is the second best method in terms of overall error rate. QDA is the least ideal method.

According to the four confusion tables, it looks like sensitivity is a bigger problem than specificity, because different methods have bigger differences in sensitivity than in specificity. However, sensitivity, which is the ability to classify correctly a person under the overweight category, is usually what we care about the most in real life.

In terms of the choice of K, it may be ideal to perform a cross-validation. However, based on previous HW assignment, I believe the having K equal to 10 is appropriate.