Regularization

# Regularization

When modeling for cupper points, six iterations of model building were created. Three of the interations used the original date set and the other three used the calculated PCA scores. Each iteration of model building follows a systematic approach. First, OLS models were created including forward and backward stepwise regression for comparisons.Then, ridge,lasso and elasticNet with varying levels of alphas were also created and done for both lambda.min and lambda.1se. For each interation, the data was seperated into training and test sets (80/20 split) in order to calculate the root mean squared error (RMSE). At each iteration, feature selection was performed by way of lasso, elasticNet and p-value (OLS), and AIC stepwise selection methods. The most common variables between each feature selection methods were then used to create the next interation of models and repeat the process. In addition to feature selection, the standardized residuals for the best model in each iteration was calculated and any observations that exceeded standard deviations were removed from the dataset before running a new iteration.

## Original Dataset Iterations

1. 26 predictors

* species | country | region | bagCount | weight | harvestYr | gradeDate | variety | process | aroma | flavor | aftertaste | acidity | body | balance | uniformity | cleanCup | sweetness | moisture | oneDefect | quakers | color | twoDefect | expirDate | certBody | avgAltitude

Total number of observations is 956. Overall, The best model was produced with elasticNet using Lambda 1SE at alpha = 0.2.

1. 19 predictors, Less 8 residual outliers

* species | harvestYr | process | aroma | flavor | aftertaste | acidity | body | balance | uniformity | cleanCup | sweetness | moisture | oneDefect | quakers | color | twoDefect | Bag\_Weight | avgAltitude

Total number of observations is 948 Overall, The best model was produced with elasticNet using Lambda 1SE at alpha = 0.05. Many of the variables removed had too many levels without enough obervations within each level. Time variables were removed because the date frames did not have consistent time frames. In addition, 3 transformations were performed. Moisture received a log transformation, avgAltitude was reduced by a factor of 1000, and 2000 was subtracted from harvest year (that way, the data would only reflect the decade it was harvest, 2014 = 14)

1. 6 predictors, Less 6 residual outliers

* flavor | aftertaste | acidity | body | balance | avgAltitude

Total number of observations is 942 Overall, The best model was produced with elasticNet using Lambda Min at alpha = 0.35.

## PCA Scores Iterations

1. All Observations

Total number of observations is 956. Overall, The best model was produced with elasticNet using Lambda 1SE at alpha = 0.2.

1. Less 8 residual outliers

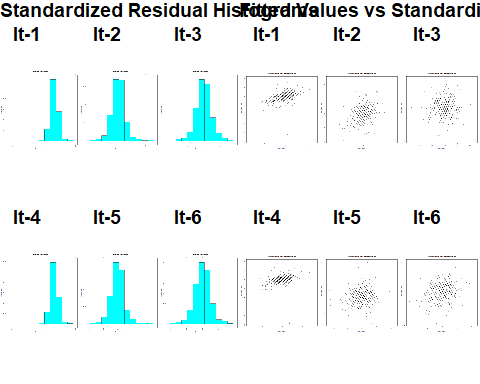
Total number of observations is 948 Overall, The best model was produced with elasticNet using Lambda 1SE at alpha = 0.05.

1. Less 7 residual outliers`

Total number of observations is 941 Overall, The best model was also produced with elasticNet using Lambda 1SE at alpha = 0.05.

## Residual Analysis

Residual analysis did not yield good results for the 1st or 4th iteration. The standardized residuals in the 1st and 4th interation are highly skewed. This can can also be noted by the high degree of heteroscedasticity when looking at the fitted values against the residuals.



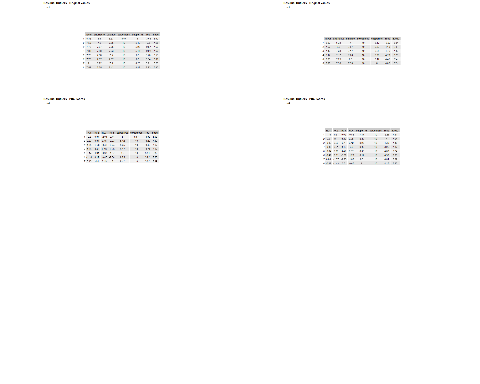
There are three obvious outliers in the bottom left hand corner that have standardized residuals exceeding -9. This is a very drastic deviation and may be strongly influencing the effects of the model. This may be an indication of some variables that are giving the model too much weight. For the final models, the residuals against the predictors and fitted residuals were homoscedastic.

### Influential Outliers

From all the iterations, most of the residual outliers tended to be on the outer edges. As most of the observations below have absolute residual values between 3.5 and 4. The influential outliers can be argued to be in It-1. Where 3 of the residuals had absolute scores above 9.

The same be seen using the PCA scores. In fact the same 3 observations were removed. This is an indication of how the model is unable to represent those few points or something else may be going on.

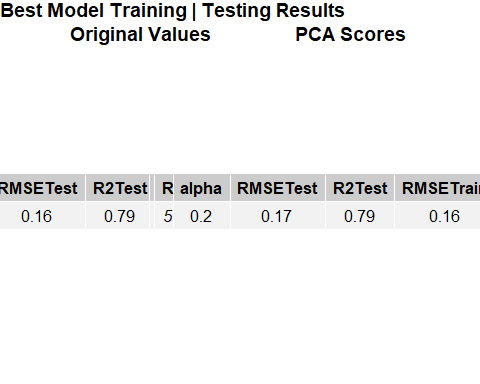
## png   
## 2



The most immediate pattern that can be discerned is that all the of observations had very high scores in sweetness, while moderatly high values in everything else. It seems to suggest that those observations, although moderatly strong scores in other categories, the downfall was the overpowering sweetness that contributed to the lower actual scores. The model doesn’t have enouch observations to account for this type of potential pattern to recognize the need for a penalty for being too sweet.

## Results

Overall, The best model was produced with elasticNet using lambda Min with an alpha of 0.35 with the original values and lambda at 1 SE with an alpha of 0.2. This produced the smalled RMSE error when comparing the training and test sets. The difference between the training and test RMSE are very small, but it does suggest that the model has, at most, a low degree of overfitting due to idiosyncrasis in the data. Both models have a strong goodness of fit capturing between 78% and 81% of the variance in cupper points.



The beta coefficients suggest that the most important variables that can explain the variability in cupper points are **flavor, aftertaste, acidity, body and balance and altitude** More importantly, it suggests that flavor, aftertaste and balance are the strongest contributing indicators by beta weight that may influence the final score in cupper point ratings. Another important note is the beta coefficient of Altitude. Although below it is depicted as 0 it is in fact non-zero, just a very small beta weight. Without the variable included in the model the RMSE nearly doubles. For that reason, it remains in the model due to its significance and the reduction in RMSE.

## Warning: Graphs cannot be horizontally aligned unless the axis parameter is set.  
## Placing graphs unaligned.

