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. This was done for both lambda.min and lambda.1se. For each interation, the data is 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) 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 were calculated and observations that exceeded ±3.5 standard deviations were removed from the dataset before running a new iteration. A total of 14 and 15 residual outliers were found of which 3 seemed to be highly influential for the original and PCA scores, respectively.

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.

2. 19 predictors

   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. Also, time variables were removed because the date frames did not have consistent time frames.

3. 6 predictors

   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.

2. 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.

3. Less 7 residual outliers‘

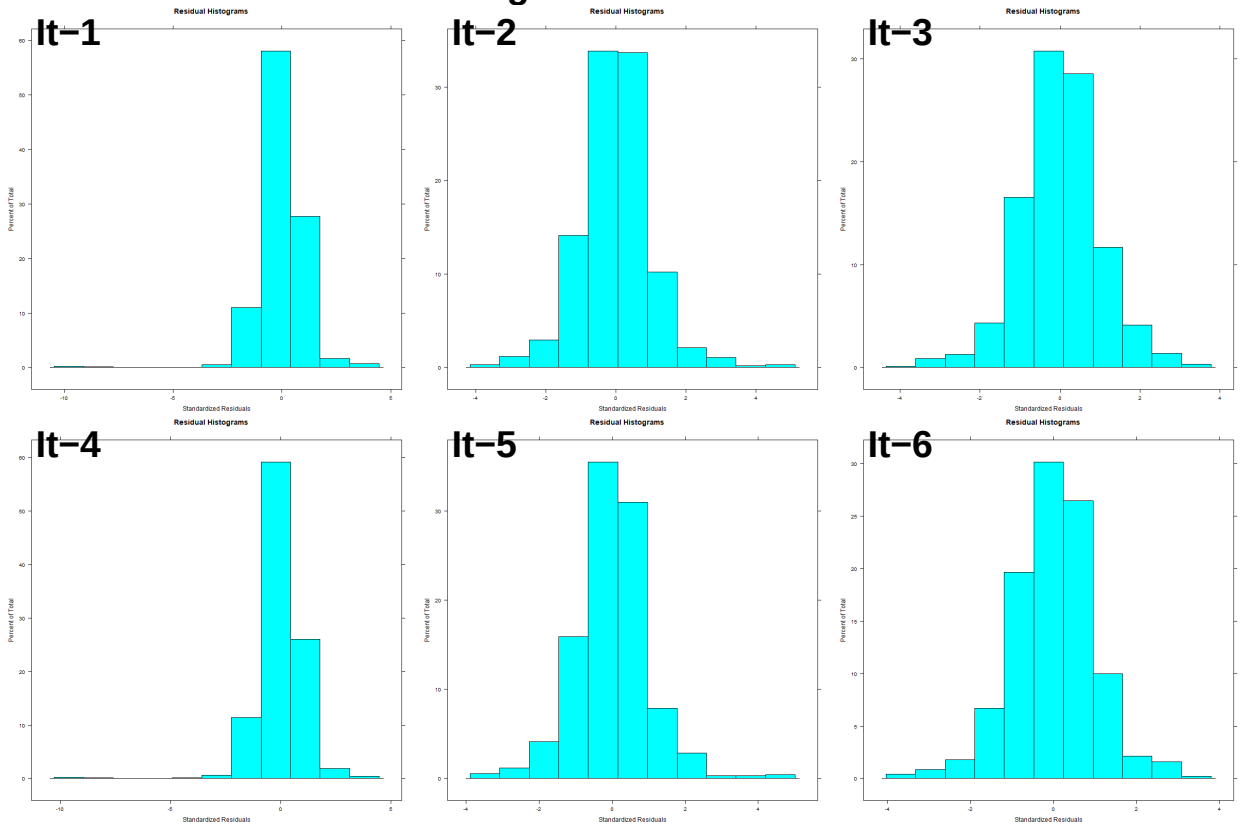
Total number of observations is 941 Overall, The best model was also produced with elasticNet using Lambda 1SE at alpha = 0.05.

Residual Analysis

Residual analysis did not yield good results for the 1st or 4th iteration. Very few elements displayed low degrees of heteroscedasticity. There are very obvious outlier that is apparant in iterations 1,2,4 and 5 with some standardized residuals exceeding -8. 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

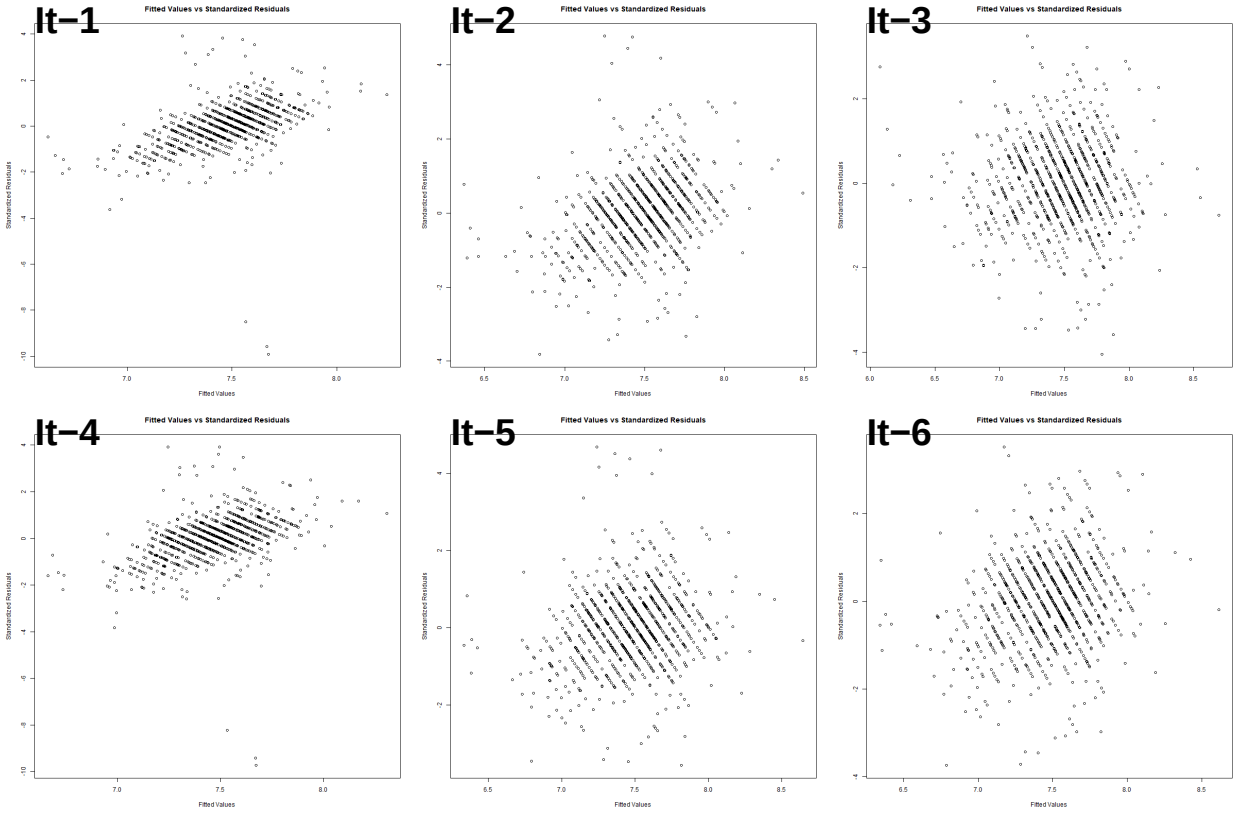
```
plot_grid(
  ggdraw() + draw_label(
    "Standardized Residual Histograms",
    fontface = 'bold',x = 0,hjust = 0),

  plot_grid(
iteration1[[1]]$Regularization$lambda.1se$Best[[1]]$`Standardized Residuals Histgram`,
iteration2[[1]]$Regularization$lambda.1se$Best[[1]]$`Standardized Residuals Histgram`,
iteration3[[1]]$Regularization$lambda.min$Best[[1]]$`Standardized Residuals Histgram`,

iteration4[[1]]$Regularization$lambda.1se$Best[[1]]$`Standardized Residuals Histgram`,
iteration5[[1]]$Regularization$lambda.1se$Best[[1]]$`Standardized Residuals Histgram`,
iteration6[[1]]$Regularization$lambda.1se$Best[[1]]$`Standardized Residuals Histgram`,

    align = "h",nrow = 2,ncol = 3,labels = c("It-1","It-2","It-3","It-4","It-5","It-6")
    ),

  align = "h", ncol = 1,rel_heights = c(.05,1)
)
```

Standardized Residual Histograms




```
plot_grid(
  ggdraw() + draw_label(
    "Fitted Values vs Standardized Residuals",
    fontface = 'bold',x = 0,hjust = 0),

  plot_grid(
iteration1[[1]]$Regularization$lambda.1se$Best[[1]]$`Fitted Values vs Standardized Residuals`,
iteration2[[1]]$Regularization$lambda.1se$Best[[1]]$`Fitted Values vs Standardized Residuals`,
iteration3[[1]]$Regularization$lambda.min$Best[[1]]$`Fitted Values vs Standardized Residuals`,

iteration4[[1]]$Regularization$lambda.1se$Best[[1]]$`Fitted Values vs Standardized Residuals`,
iteration5[[1]]$Regularization$lambda.1se$Best[[1]]$`Fitted Values vs Standardized Residuals`,
iteration6[[1]]$Regularization$lambda.1se$Best[[1]]$`Fitted Values vs Standardized Residuals`,

    align = "h",nrow = 2,ncol = 3,labels = c("It-1","It-2","It-3","It-4","It-5","It-6")
    ),

  align = "h", ncol = 1,rel_heights = c(.05,1)
)
```

Fitted Values vs Standardized Residuals



Influential Outliers

```
df1rc
```

```
## # A tibble: 8 x 22
##   cupperPts species harvestYr process aroma flavor aftertaste acidity  body
##       <dbl> <fct>       <dbl> <fct>   <dbl>  <dbl>      <dbl>   <dbl> <dbl>
## 1      6    Arabica      2013 Washed~  6.5    6.33       6.5     7.5   7.33
## 2      5.42 Arabica      2017 Washed~  7.17   7.58       7.5     7.5   8
## 3      5.25 Arabica      2015 Natura~  7.83   7.75       7.67    7.83  7.83
## 4      5.17 Arabica      2015 Other    7.75   7.83       7.58    7.75  7.92
## 5      8.5  Arabica      2018 Natura~  7.5    7.75       7.58    7.5   7.58
## 6      8.5  Arabica      2013 Washed~  7.67   7.75       7.67    7.58  7.58
## 7      8.25 Arabica      2016 Natura~  7.58   7          6.92    6.92  7.67
## 8      8.42 Arabica      2012 Natura~  7.83   7.58       7.25    7.5   7.5
## # ... with 13 more variables: balance <dbl>, uniformity <dbl>, cleanCup <dbl>,
## #   sweetness <dbl>, moisture <dbl>, oneDefect <dbl>, quakers <dbl>,
## #   color <fct>, twoDefect <dbl>, Bag_Weight <dbl>, avgAltitude <dbl>,
## #   res <dbl>, fitted <dbl>
```

```
df2rc
```

```
## # A tibble: 6 x 22
##   cupperPts species harvestYr process aroma flavor aftertaste acidity  body
##       <dbl> <fct>       <dbl> <fct>   <dbl>  <dbl>      <dbl>   <dbl> <dbl>
## 1      6.17 Arabica      2013 Washed~  7.08   6.92       6.33    7     6.92
## 2      8.33 Arabica      2014 Washed~  7.67   7.67       7.5     7.5   7.67
## 3      8.25 Arabica      2013 Washed~  7.75   7.42       7.33    7.5   7.58
## 4      8.08 Arabica      2013 Washed~  7.5    7.17       7.17    7.17  7.5
## 5      8.17 Arabica      2014 Natura~  7.25   7.33       7.25    7.5   7.42
## 6      8    Arabica      2014 Washed~  7      7.08       7.33    7.5   7.5
## # ... with 13 more variables: balance <dbl>, uniformity <dbl>, cleanCup <dbl>,
## #   sweetness <dbl>, moisture <dbl>, oneDefect <dbl>, quakers <dbl>,
## #   color <fct>, twoDefect <dbl>, Bag_Weight <dbl>, avgAltitude <dbl>,
## #   res2 <dbl>, fitted2 <dbl>
```

Results

Overall, The best model was produced with elasticNet using lambda Min with an alpha of 0.3 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 were small, but it does suggest that the model has at least a low degree of overfitting to idiosyncrasis in the data.

```
round(iteration3[[1]]$Regularization$lambda.min$Results$`Best Model RMSE Results`,2)
```

```
##   alpha RMSETest R2Test RMSETrain R2Train
## 8  0.35     0.16   0.79      0.15    0.81
```

```
round(iteration6[[1]]$Regularization$lambda.1se$Results$`Best Model RMSE Results`,2)
```

```
##   alpha RMSETest R2Test RMSETrain R2Train
## 5   0.2     0.17   0.79      0.16    0.78
```

Lasso produced the most parsimonious model with only 3 predictors. Just as Net at 0.50, the beta weights included are not trivially zero. Aftertaste, Flavor, and Balance appear to be some of the most important predictors. Ridge does not give very interpretable results because all the variables have beta weights that are trivially zero. ElasticNet at 0.50 also produced a sparse model. Compared to that of alpha 0.05, the beta weights are not trivially zero. 5 distinct predictors standout - Aroma, Flavor, Aftertaste, Acidity, Body, and Balance"

Linear models overall all three linear models produced terrible, overfit results. This is mainly due to the many levels in the categorical variables that can make the interpretations very difficult. The linear model with all the variables expectedly produced the worst model. Due to the many factor levels it is very difficult to differentiate between important and unimportant predictors". The backwards model produced a majority of the betas to be statistically significant, however, it did so by removing some of the categorical country and regions. It has an F statistic less than alpha suggesting that at least one of the betas is not equal to zero. The adjusted R^2 is 74.65%. Respectable, but it does not help interpreting the relationships. All the residuals appear to have a multiplicative process, that is, as actual values increase, so do the residuals. Additionally, all the plots display a high degree of heteroscadisticity. There may be a need to transform the variables or standardize the variables in order to correct the residual patterns. The QQplot suggests that the residuals are not normal. The majority of the points lie above and below the QQ line. The deviations are most apparant at the tails. The histogram looks normal but with extreme tails on both ends. It depicts outliers exceeeding 5 SDs. Residuals against the response appear to have a curved relationship. Suggesting the need to transform the response variable. All the residuals have a high degree of homoscedasticity with a multiplicative residual pattern.