Zeeshan Latifi

Northwestern University Winter 2018

Predict 411

Wine sales project

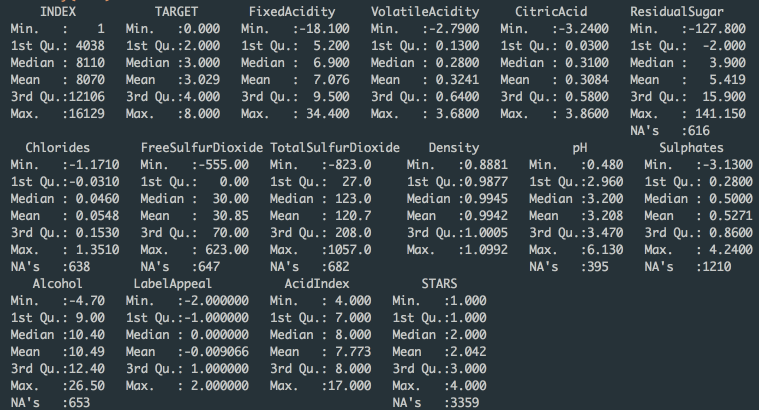
**Introduction:**

The purpose of this analysis is to determine whether we can predict the number of wine sample cases that were purchased by a wine distribution company after sampling the wine by using the wine characteristics as the predictor variables. The data set contains information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The target variable is what we will be predicting for as mentioned above. These cases are used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. If we can accurately predict the number of cases, then that manufacturer will be able to adjust their wine offerings to maximize sales.

**Data Exploration:**

The data contains 16 variables with 12,795 records of wine data. Figure 1 below highlights the summary of the data for each variable. As you can see there are several variables that contain missing values. For these variables with “NA’s” we’ll impute values using the simple average method.

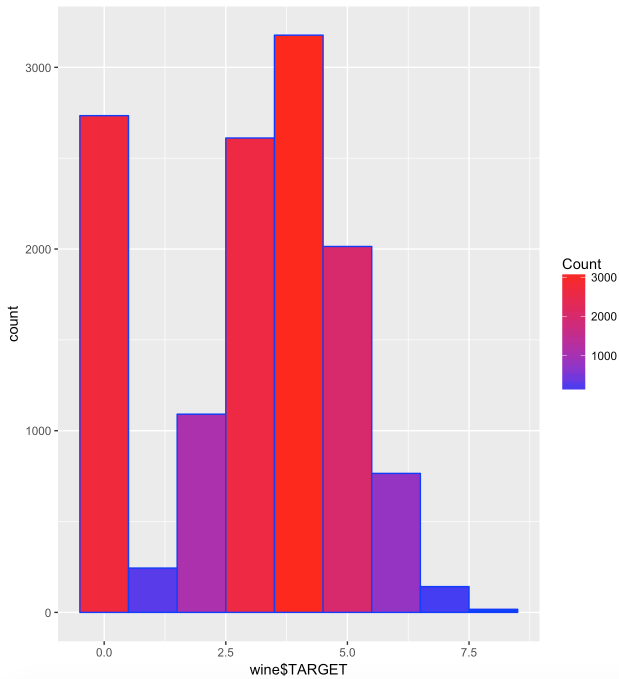
Figure 1



Notice how the Target variable minimum and maximum are 0 and 8 respectively with the mean being just over 3. We would expect these results when evaluating our model with our test data set. About 25% of the data for STARS is missing, we’ll need to impute these values before using that variable in our model.

First let’s begin by looking at our Target variable. Figure 2 below outlines the Target variable count. You can see that there is a large proportion of wine orders that are 0. Other than this, the rest of the data seems to be fairly distributed with a slight right skew.

Figure 2



Let’s now look at some of the predictor variables. Figure 3 below shows the distribution for both Acid Index and well as the Alcohol content in the wine. Acid Index seems to be skewed right, with many upper outliers. Alcohol content levels between 7 and 12 seem to most prevalent; however, we can say the distribution is fairly normal.

Figure 3

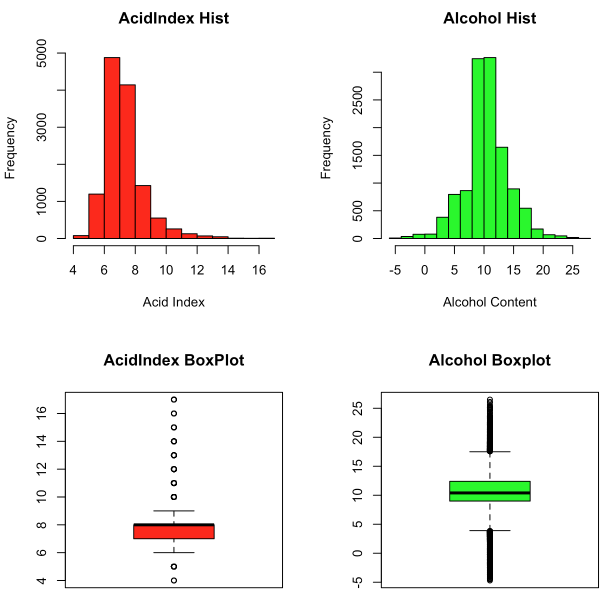
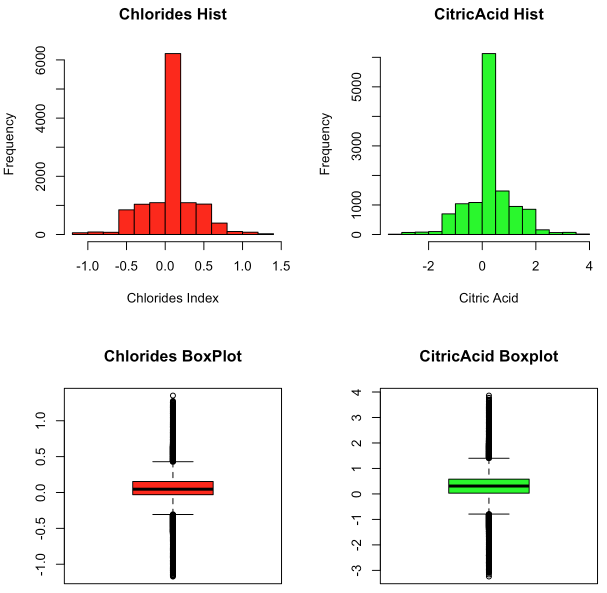


Figure 4 below outlines the histogram and boxplot for Chlorides and Citric Acid. Here there is a little different story. We can see that there is a large number of wines that fall within the a very low amount of Chloride as well as Citric Acid. Because of this, the boxplots for each are showing many outliers. The same can be said for Density and Fixed Acidity.

Figure 4



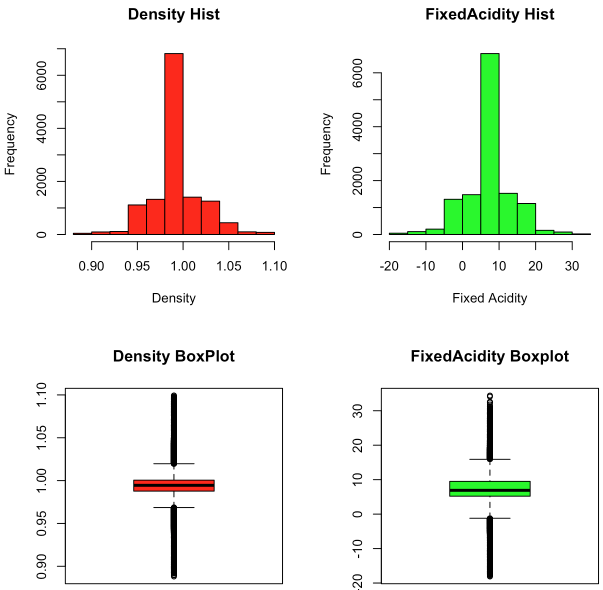


Figure 5 below showcases a little different story. We wanted to see the distribution of the Label Appeal and number of Stars for each wine (if the data exists). We can see that most of the data that we have is a Star rating of 2 for the wines. The majority of wines showcase a Label Appeal of 0.

Figure 5

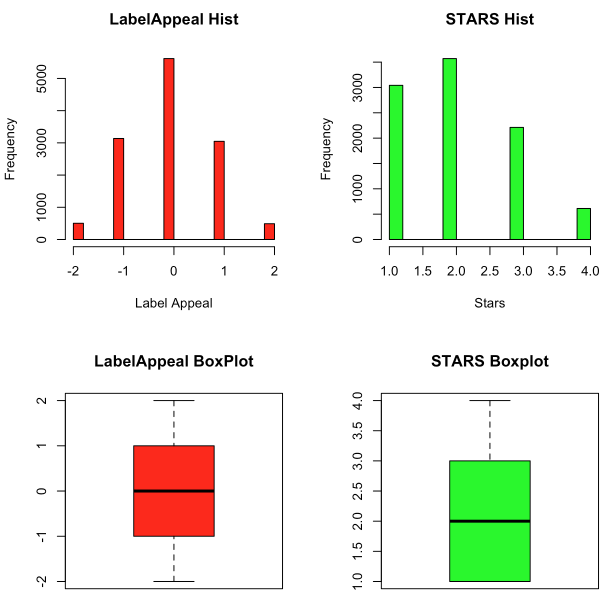
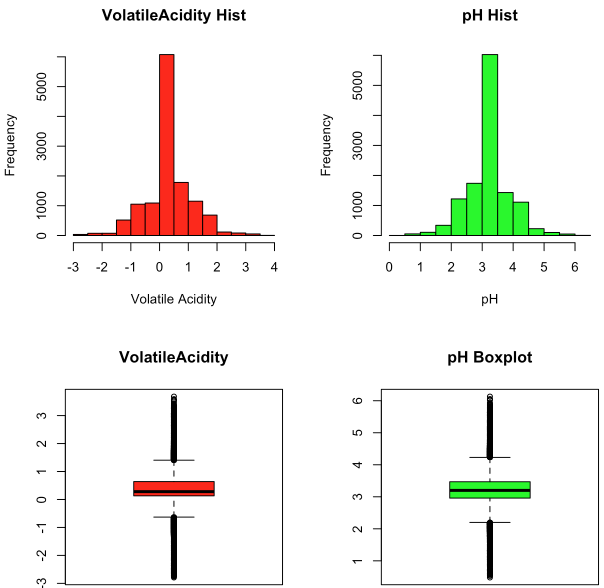


Figure 6 below looks very similar to the variables above in Figures 3 and 4. Volatile Acidity and pH levels seems to be predominately in the middle of the pack. Therefore, the boxplots show many outliers on either tail end. There may be a correlation to these variables. We’ll dive deeper in the next section.

Figure 6

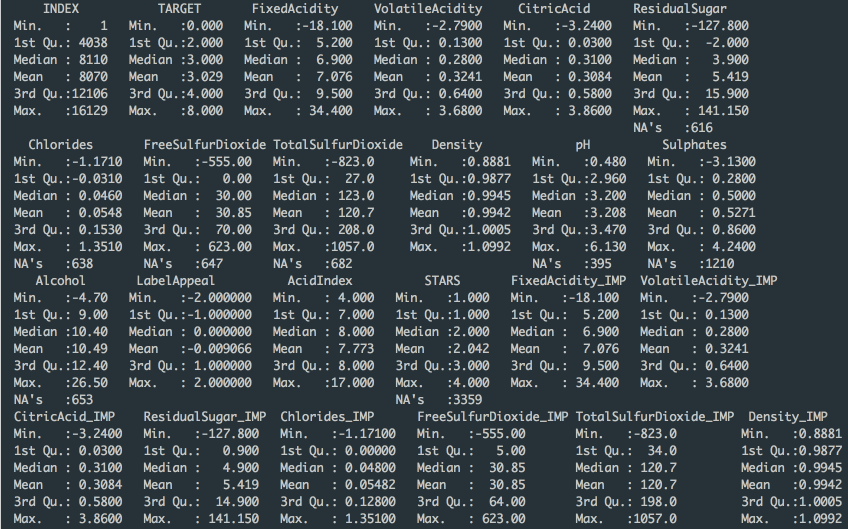


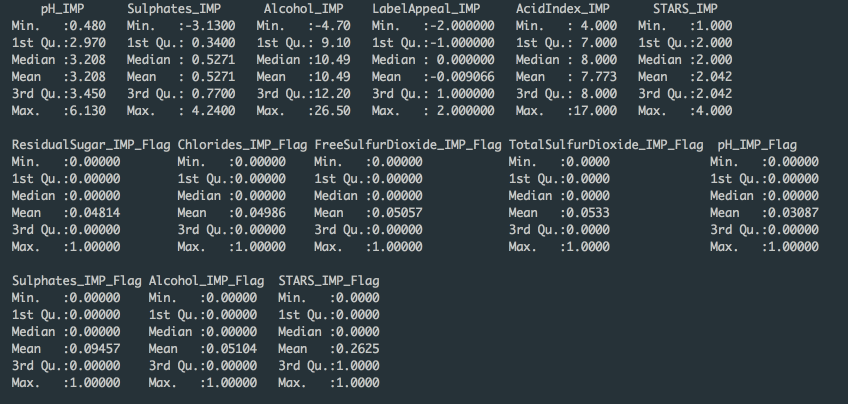
**Data Preparation:**

To being our data preparation, we’ll impute the values needed for our missing variables. To do this, we’ll use the simple average method and just replace our missing values with NA’s.

To start, we need to impute the missing values in our data. In order to do this, we’ve replaced our missing values with the mean of the actual values. The newly imputed values are given a new name with the suffix: IMP. It should be mentioned that every transformation, categorization, or imputation were also replicated on our test data set. This is to ensure proper predictability when deploying our model.

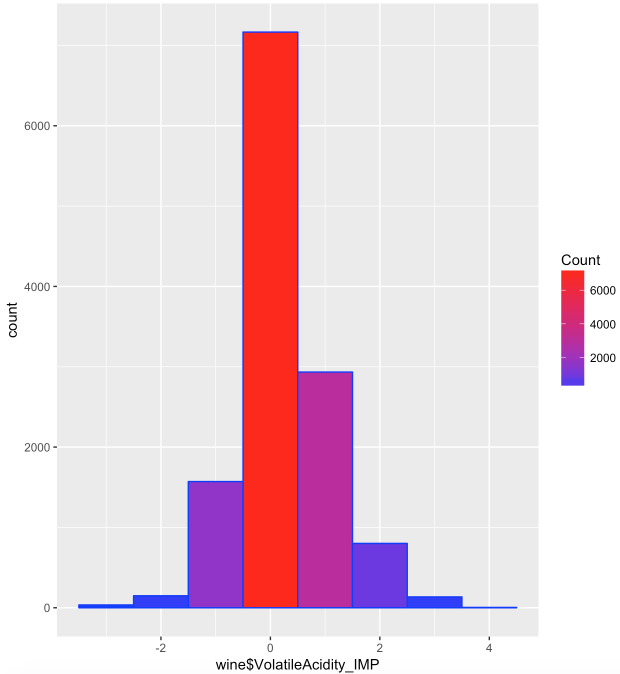
Figure 7





Next, we also created some flags within our data set to ensure accuracy and detectability. We flagged our variables that were imputed with a 1, 0 if actual value. We also broke out the type of wine, red or white, by looking at the volatile acidity. Figure 8 below has the breakdown of wine of red versus white. White wine is represented by the blue color.

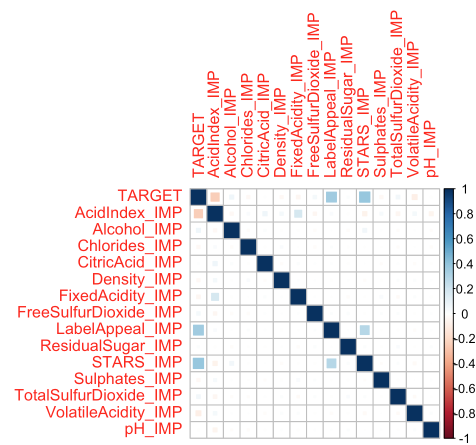
Figure 8



This indicator was created so that when our model is complete, we can see the breakdown of what cases of wine were ordered by the type of wine. As shown the majority of the wine that falls within a 0 range for Volatile Acidity happens to be red wine.

Additionally, we also built a correlation plot in Figure 9. You can see that there is a positive correlation between our Target and the number of Stars a wine gets as well as the Label Appeal. This makes sense logically, the better the rating, the more you’ll sell. There is also somewhat of a negative correlation between the Acid Index and the Target variable as shown.

Figure 9



**Build Linear Regression Models:**

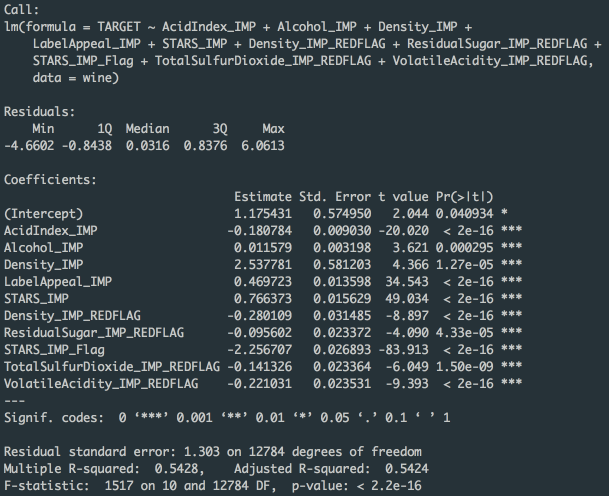
From our cleansed data, we now begin building an optimal model for predicting our response variable. In order to optimize our search for the most accurate model, we will use an R function (regsubsets) to score the variables that will have the most impact on generating the best model. Based on the automated variable selection, the best variables are as follows:

* AcidIndex\_IMP
* Alcohol\_IMP
* Density\_IMP
* LabelAppeal\_IMP
* STARS\_IMP
* Density\_IMP\_REDFLAG
* ResidualSugar\_IMP\_REDFLAG
* STARS\_IMP\_Flag
* TotalSulfurDioxide\_IMP\_REDFLAG
* VolatileAcidity\_IMP\_REDFLAG

Figures 10-15 below indicate the summary output for each of our models. The following models were used for this analysis (in order of occurrence):

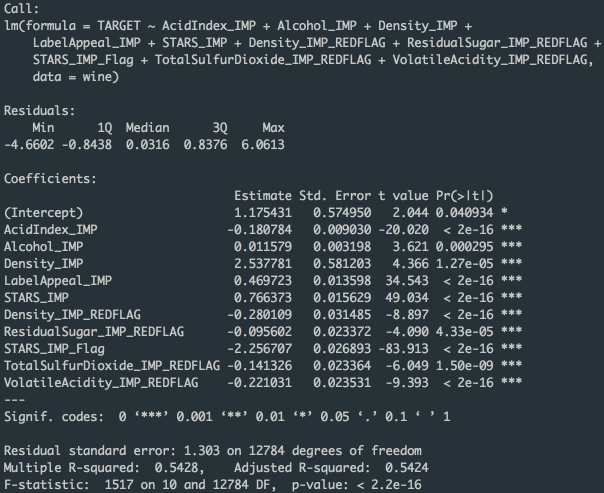
1. Regular Linear Regression
2. Regular Linear Regression using Stepwise Variable Selection
3. Poisson
4. Negative Binomial Distribution
5. Zero Inflated Poisson
6. Zero Inflated Negative Binomial Regression

Figure 10 (1)



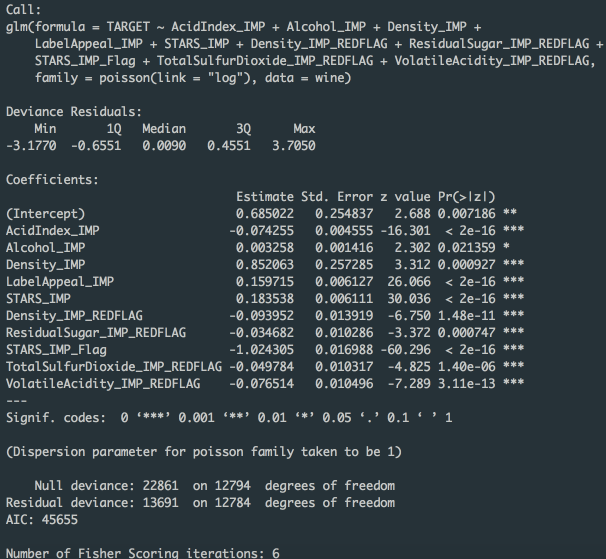
As expected, the regular linear regression is not going to be optimal here as we have zero inflated variables. However, for a good threshold, it’s always good to see what the model output is. Here we can see that Density and STARS have a significant impact on the model. STARS seems to be counterintuitive. Having more stars should be a positive impact on the cases of wine being sold.

Figure 11 (2)



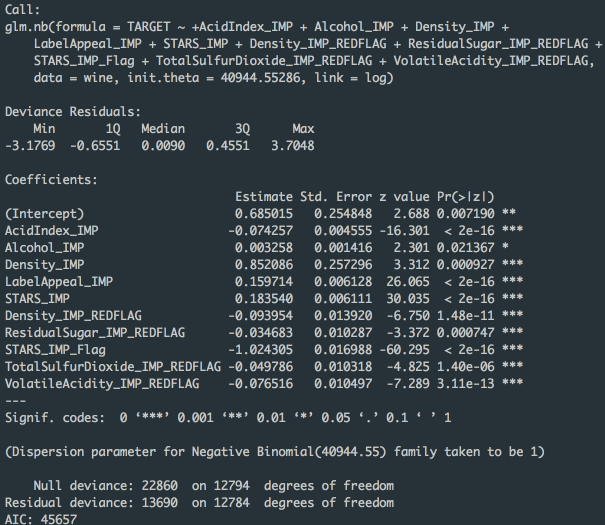
For our regular linear regression with a stepwise variable selection (Figure 11), we have very similar results to the initial regression model as shown in Figure 10. The adjusted R2 is the same for both models at 0.5424.

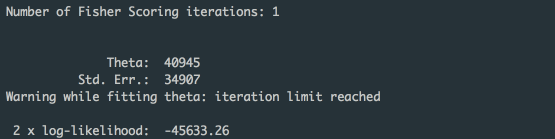
Figure 12 (3)



For our Poisson regression model, we can see that the STARS variable is also having a positive impact, although not at the magnitude at our previous models. Of the remaining variables, no other variable stands out as a significant coefficient of magnitude.

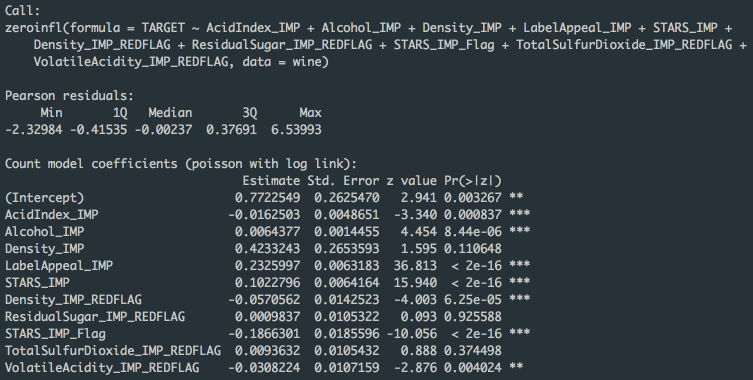
Figure 13 (4)

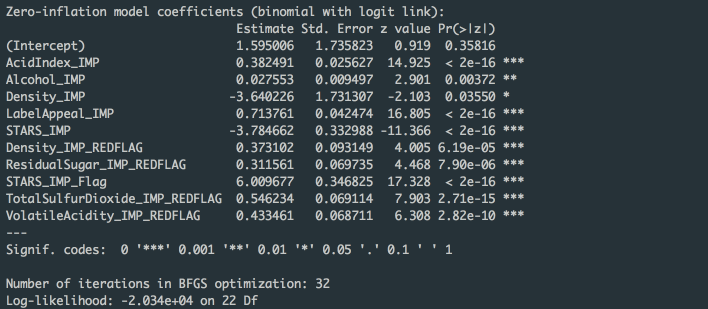




For our negative binomial distribution, we can see that there is again a positive coefficient for STARS. This makes logical sense, however, the STARS\_IMP\_FLAG is negative here.

Figure 14 (5)





Based on our data, we’re predicting this model to work best. The Zero Inflated Poisson model, accounts for the zero inflated variables. We can also see the count model above to have a positive coefficient for STARS. The data for many of our variables are zero inflated. Therefore, this model would be best for using that as a reliable predictor.

Figure 15 (6)

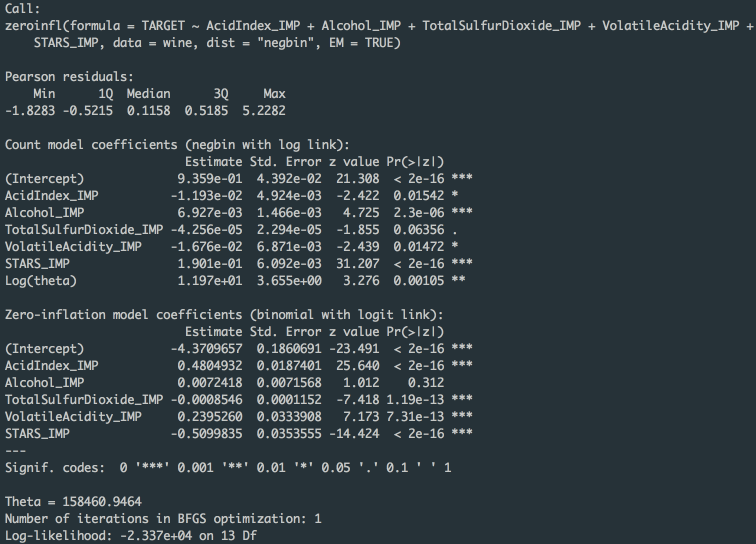


Figure 15 above outlines the last of our models. Here we’ve accounted for the zero inflated variables, but now we have negative binomial regression. For the count models, no variable stands out as significant for our regression, indicating a more conservative mode. However, when accounting for zero inflation, the intercept is of high magnitude.

**Select Best Model and Stand Alone Scoring:**

To asses our model, we will use the AIC and prediction values. The table below outlines the model metrics.

Figure 16

|  |  |
| --- | --- |
| **Model** | **AIC** |
| **Model 1** | 43098.72 |
| **Model 2** | 43098.72 |
| **Model 3** | 45654.86 |
| **Model 4** | 45657.26 |
| **Model 5** | 40730.06 |
| **Model 6** | 46761.49 |

Based on the values above in Figure 16, we can see the best model fit is when using Zero Inflated Poisson (Model 5). Just as we predicted, this was the best model. This is because when looking at the data, we had several zero inflated variables. In addition, when looking at the target variable, we noticed the variance is greater than the mean. The variance is 3.71 and the mean 3.03. This means that using a Poisson regression model would work best. When looking back at the values from the original wine data set, the min was 0, the max was 8, and the mean was just a little over 3. When building our target prediction using our test data set, the values came very close to the actual values from the training data set. We can then reasonably infer that the Zero Inflation Poisson Regression model is the best fit model to predict the cases of wine that will be sold.