Unit 3 Project

Nelson, Jon

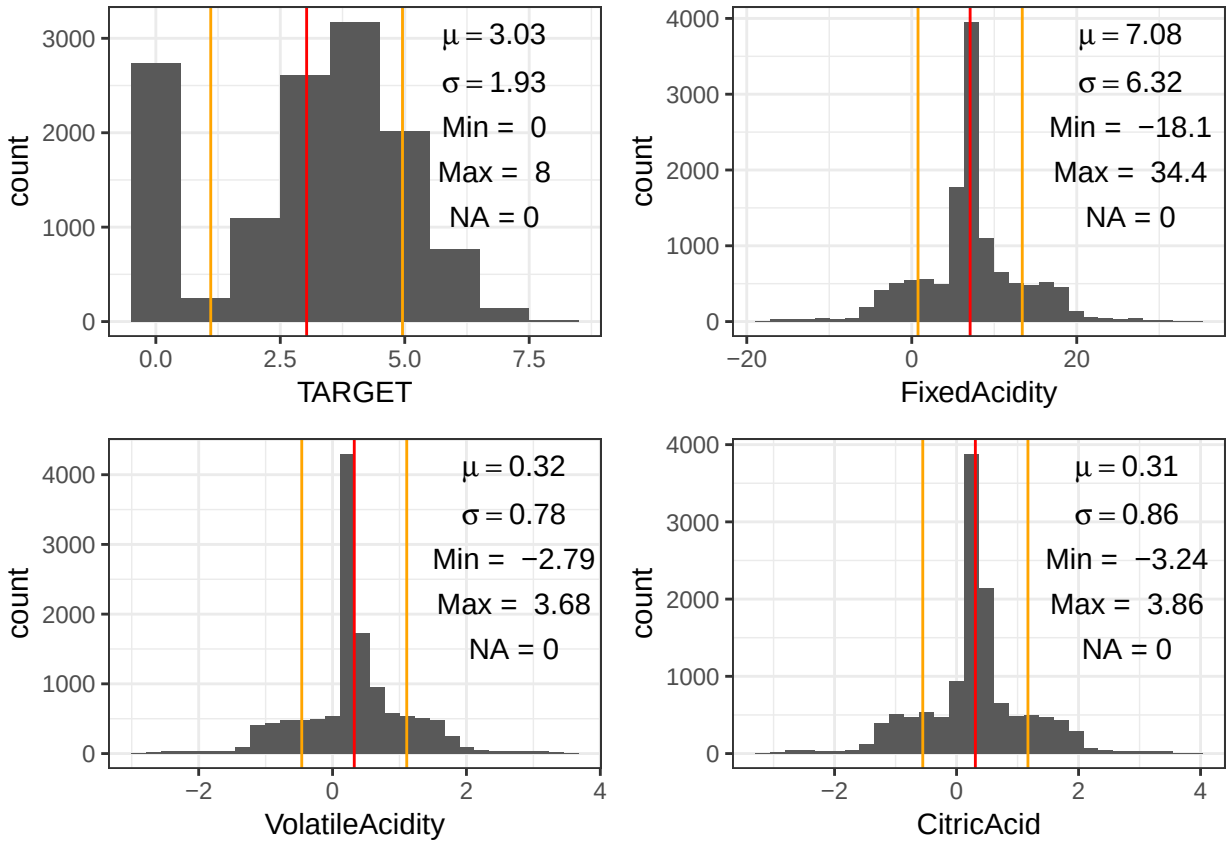
August 7, 2018

Part I: Take a sip, swish it around in your mouth, spit it out

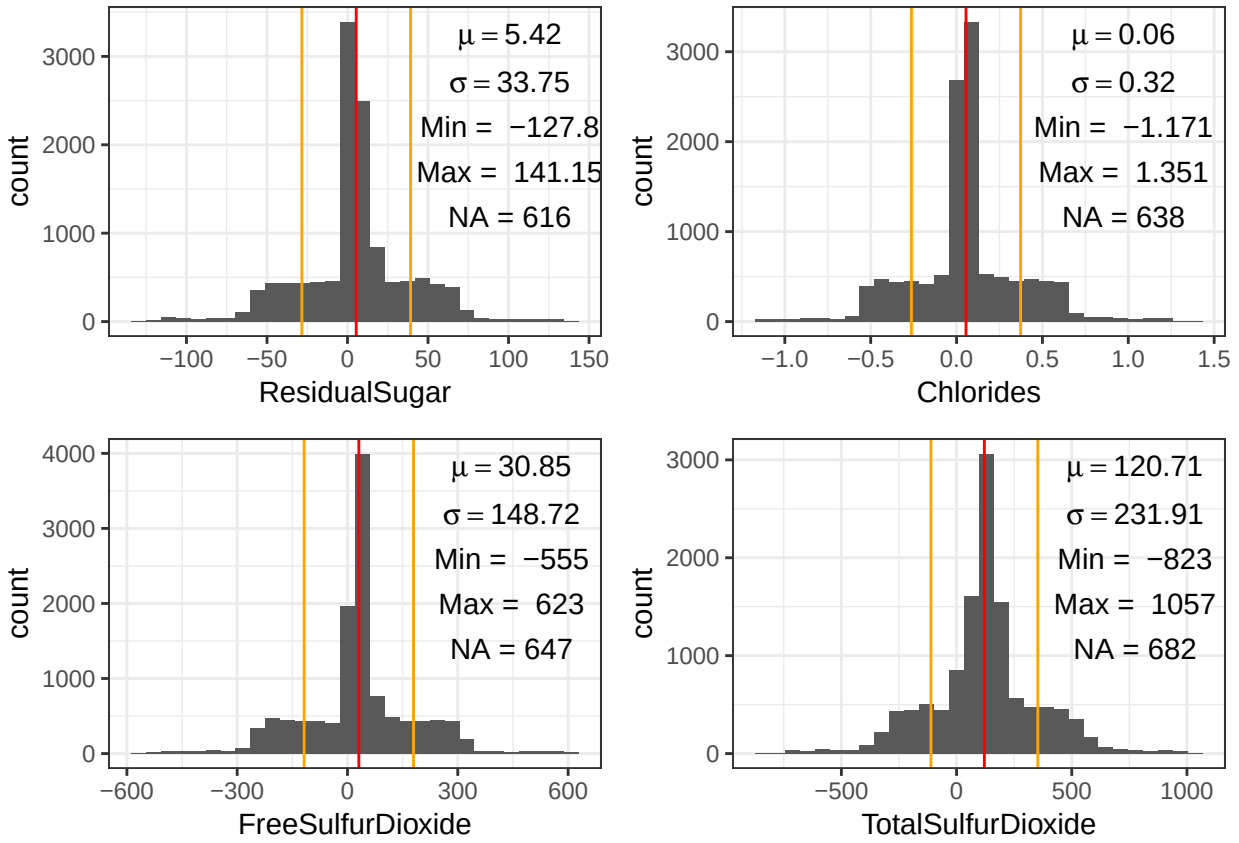
Since almost all of the variables are one form of measurement or another, the quantity of negative values in the distributions is a little concerning. There aren't many negative measurements out there, so to avoid confusion, *all* continuous variables will be Z transformed to make regression coefficients interpretable in terms of relationship to the mean. Furthermore, there are a variety of different units at play here, so standardizing in terms of Z score will aid in interpreting coefficients for relative impact. Z transformed variables will be denoted with a ".z" suffix.

On to the variables themselves:

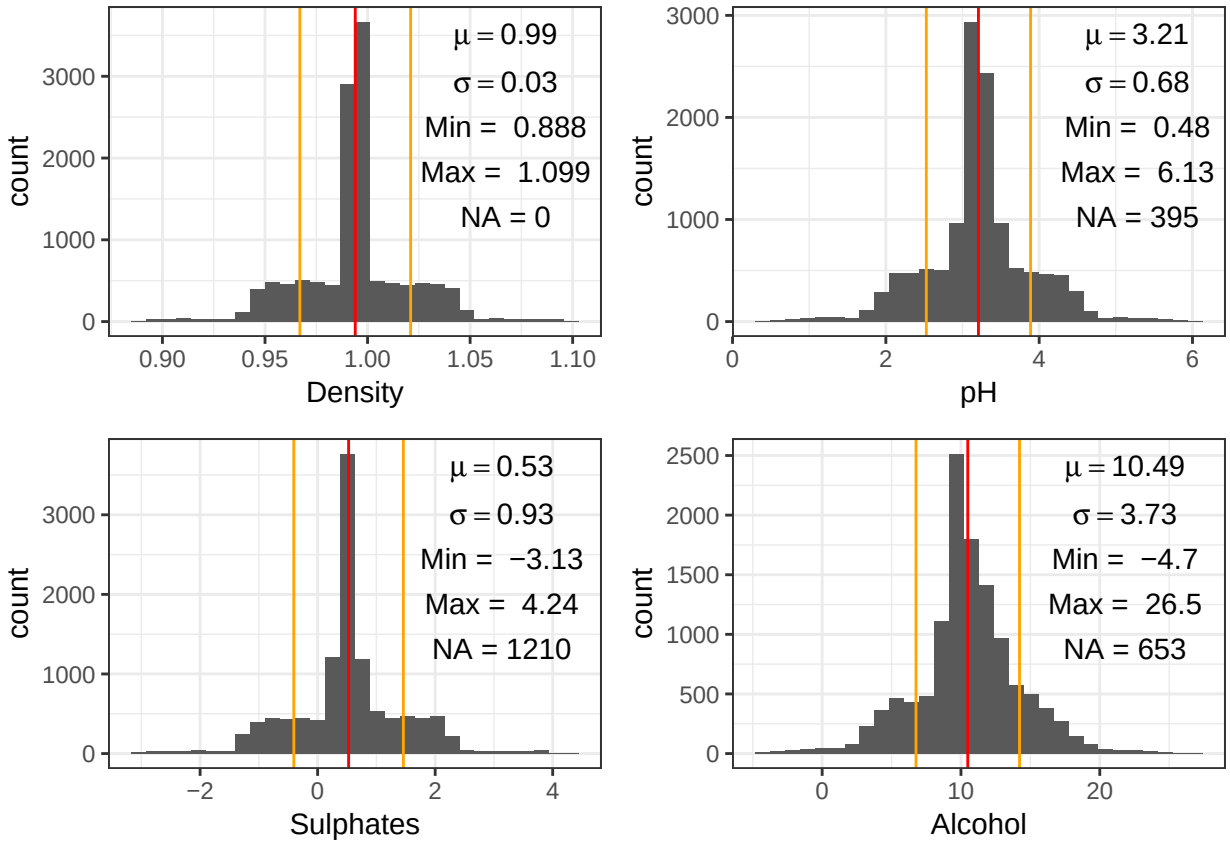
- TARGET appears to be Poisson distributed with a large spike at zero, suggesting a Zero Inflated Poisson model will be the most appropriate course of action in making useful predictions with this data. OLS regression is unlikely to perform well as it will, in all likelihood, either fail to capture the large number of zero values, or will allow negative values. There are 2734 zero values (21.37%), and the remaining values have a mean (μ) of 3.852 and variance (σ^2) of 1.548. Since Poisson distributions are characterized by a λ parameter, which captures (and equalizes) both mean and variance, and they are a special case of the binomial distribution where the variance is at a minimum (i.e. equal to the mean), negative biniomial should not provide any benefit over standard Poisson regression. This will be validated in part III.
- FixedAcidity, VolatileAcidity, and CitricAcid are all roughly normally distributed, though with a a large spike at the mean and a relatively flat dispersion to two standard deviations, with a number of outliers.

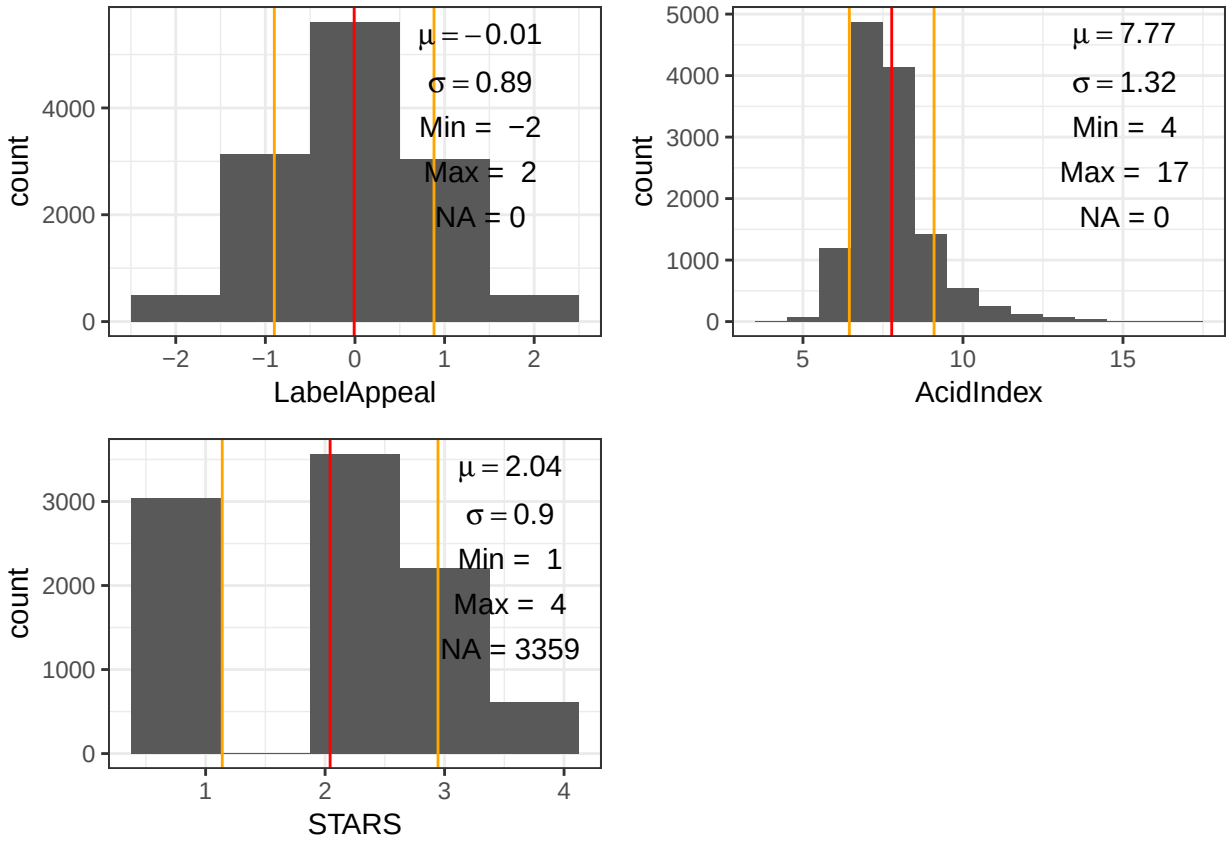


ResidualSugar, Chlorides, FreeSulfurDioxide, and TotalSulfurDioxide all share similarly shaped distributions, however each series has between 616 and 682 missing values. These records will have the mean value assigned to them.



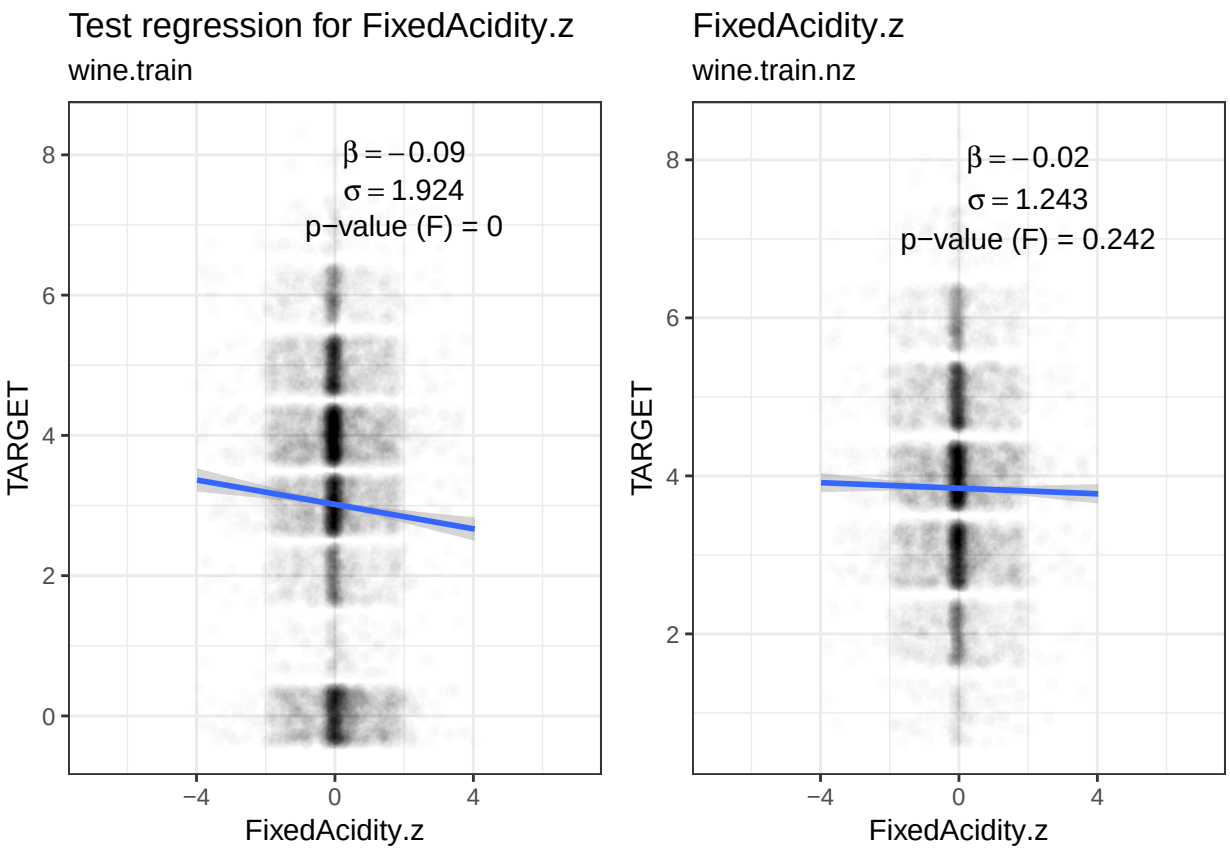
The same pattern continues to appear in Density, pH, Sulphates, and Alcohol. pH, Sulphates, and Alcohol all have missing values. The mean will be assigned to these records. Note: The pH variable contains records below a pH of 2, which is the same acidity as lemon juice. Indeed, there are even a few records under a pH of 1, which is sulphuric acid. This analyst does not recommend drinking these wines, should they exist. In any case, in the real world, wine appears to range from a pH of 3 with sweet wines up to "4+", which we'll call 4.5 (https://winefolly.com/review/understanding-acidity-in-wine/). Values outside this range will be replaced with the mean.


μ = 0.99
σ = 0.03
Min = 0.888
Max = 1.099
NA = 0
count
Density
μ = 3.21
σ = 0.68
Min = 0.48
Max = 6.13
NA = 395
pH
μ = 0.53
σ = 0.93
Min = −3.13
Max = 4.24
NA = 1210
Sulphates
μ = 10.49
σ = 3.73
Min = −4.7
Max = 26.5
NA = 653
Alcohol

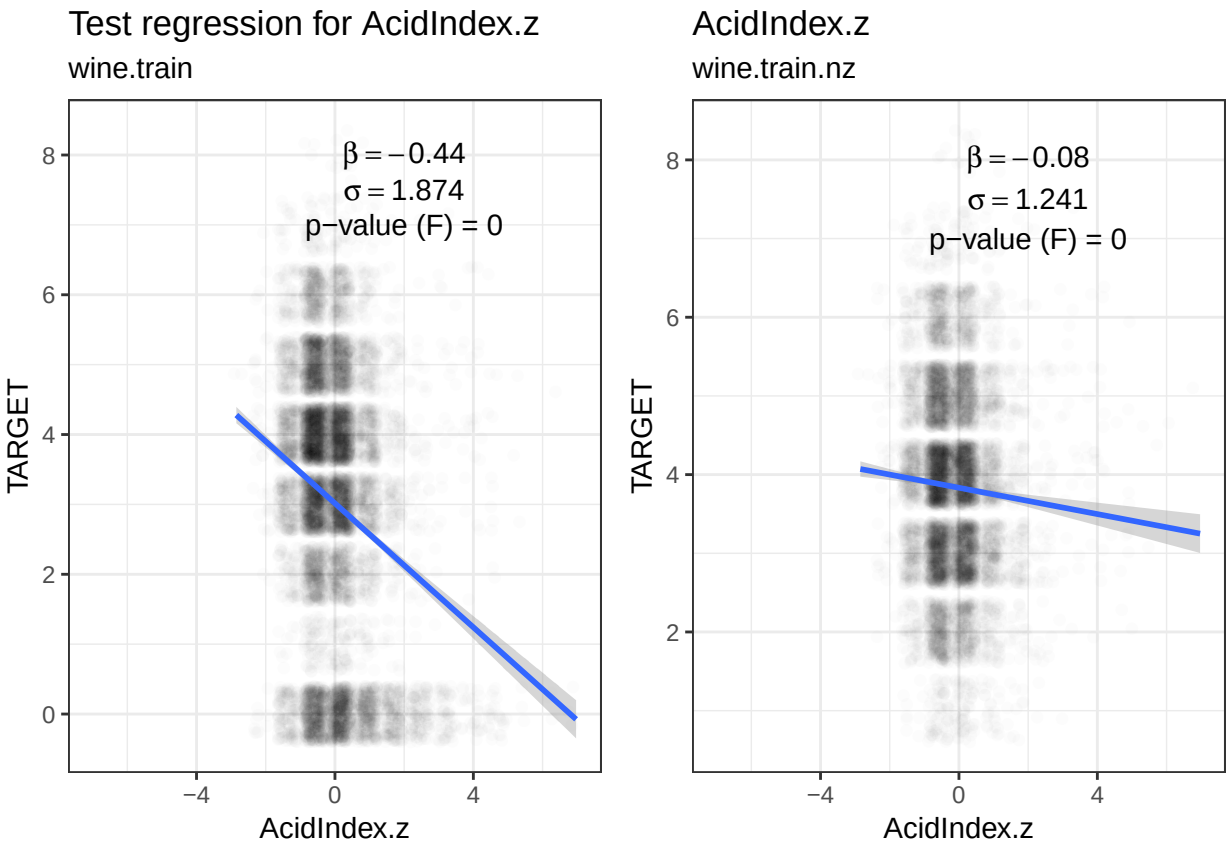


Since a significant number of these variables are measurements about of chemical properties of the wine, I surmised that there is likely to be an optimal range for each of the variables, and that as values moved outside that optimal range, it would have a deleterious effect on the demand for that wine. Theoretically, this would result in an "upside down U" shaped distribution. To test this, for each of the measurement variables, I plotted the TARGET versus the variable as well as the absolute value of the variable side by side. I added a least squares regression line to each plot to identify if there was a meaningful relationship between the two. While I expected there would be some cases where the variable would result in a flat-ish regression line for the unmodified (after Z transform) variable and an upward or downward slope for the absolute value, there were no cases like this. Note that this part of the analysis, since it is involving some basic statistical testing, is being performed *after* the data has been split into training and validation sets. No variables saw a statistically significant improvement in fit by being transformed by either absolute value or squaring.

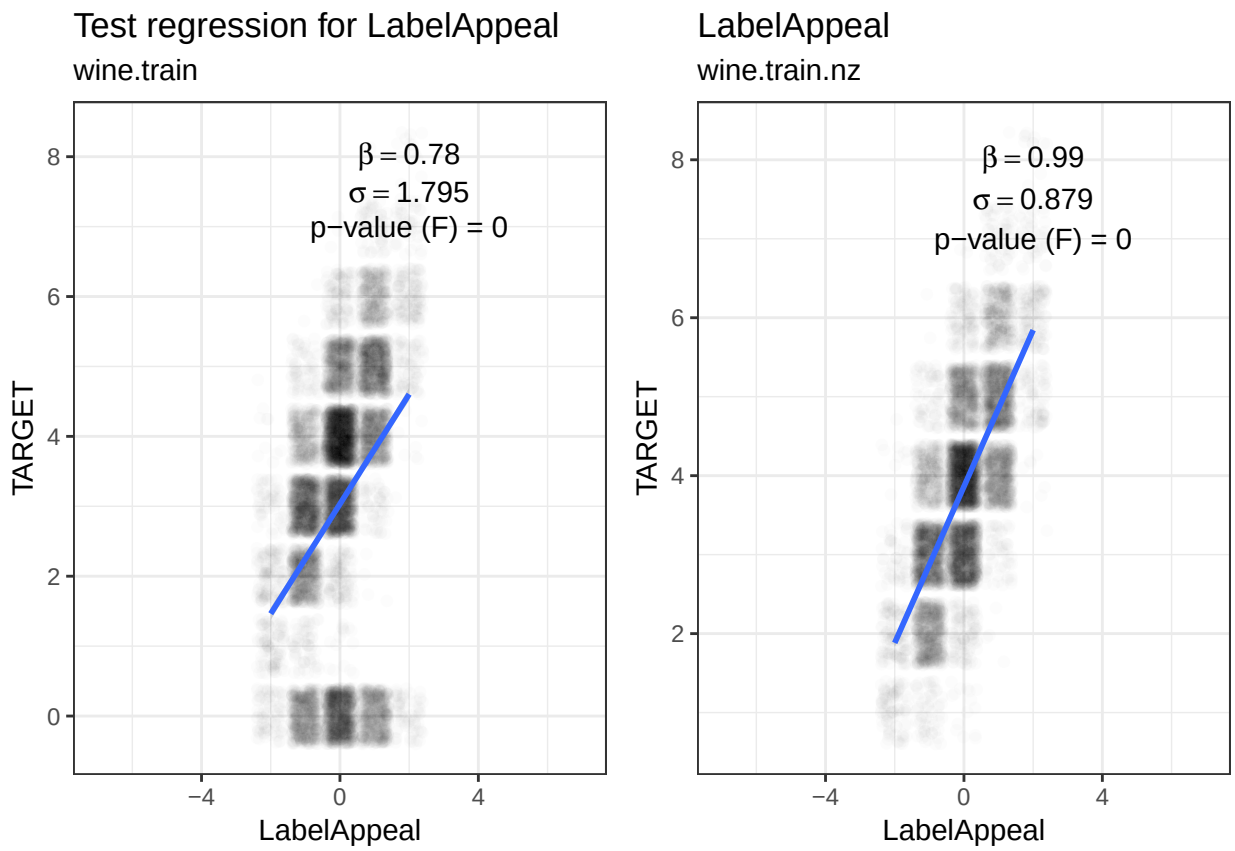
A more advanced EDA process to identify potential relationships between the predictor variables and the TARGET variable identified some potential hypothesis to validate during the modeling process. In many cases, the obvious relationship between a predictor variable and the response variable evaporates when zero values are removed from the data set. This suggests we need to think of this problem in two pieces: identifying what drives the zeroes, and running the normal regression on everything else.



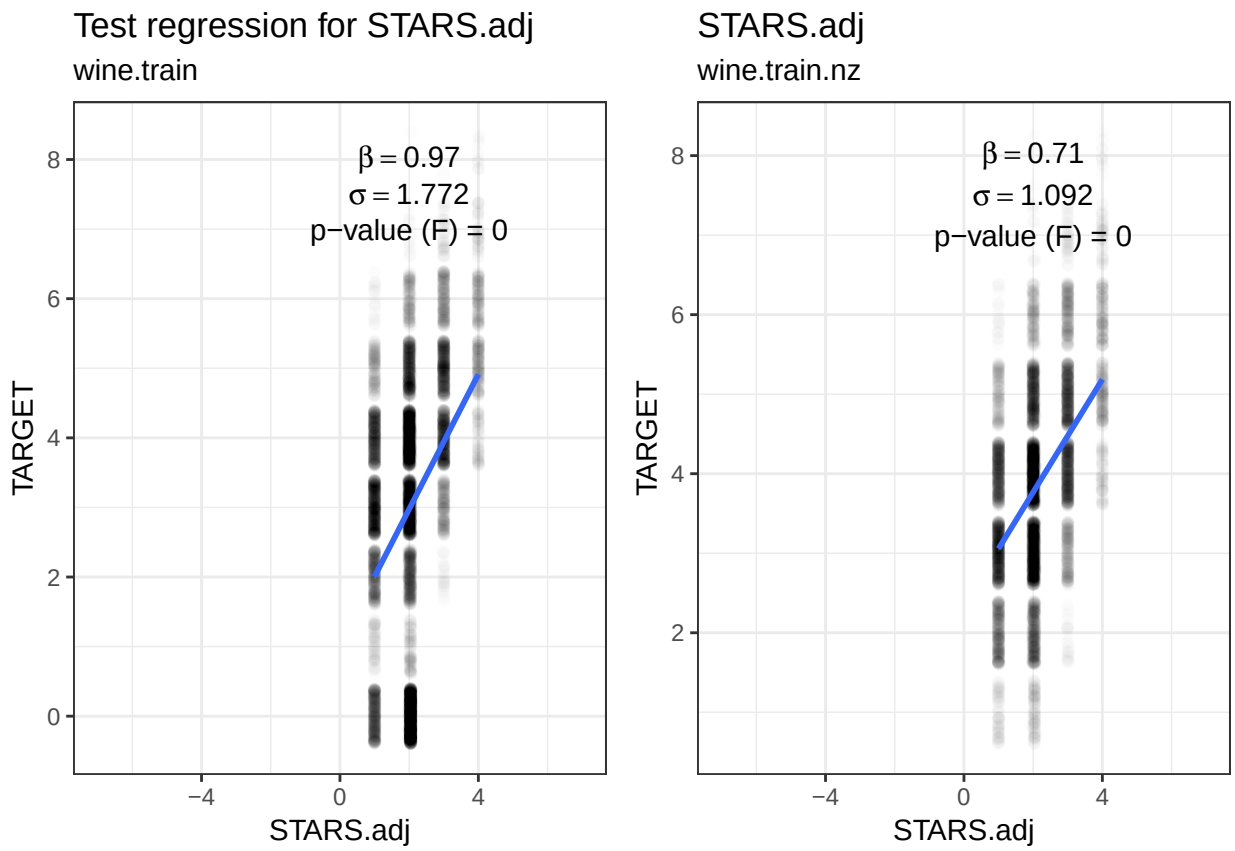
For example, with zeroes included in the data set, AcidIndex seems to be negatively correlated with wine sales, however removing the zero values seems to weaken the relationship by about 80%. This suggests that AcidIndex.z may be a useful predictor for identifying the zero values in the classification step of the zero inflated Poisson regression portion of this analysis.



LabelAppeal shows an obviously strong relationship, especially so after the zero values have been removed. We can also easily see how much the zero values in the series bring down the regression line, away from the main grouping of the data. Since the observations with a zero target value appear to be roughly normally distributed with respect to the LabelAppeal, we shouldn't expect this predictor to be particularly strong for the zero values. The difference in slope between the two regression lines is likely almost entirely driven by the squared distance between the majority of the observations without a zero target value, and the zero target values.



Finally, STARS seems to show a similar relationship as LabelAppeal, particularly after zero values are removed. Predictably, wines with a high star count seem to never fall in the zero sales category, so we should expect this variable to be useful in the categorization step in part IV.



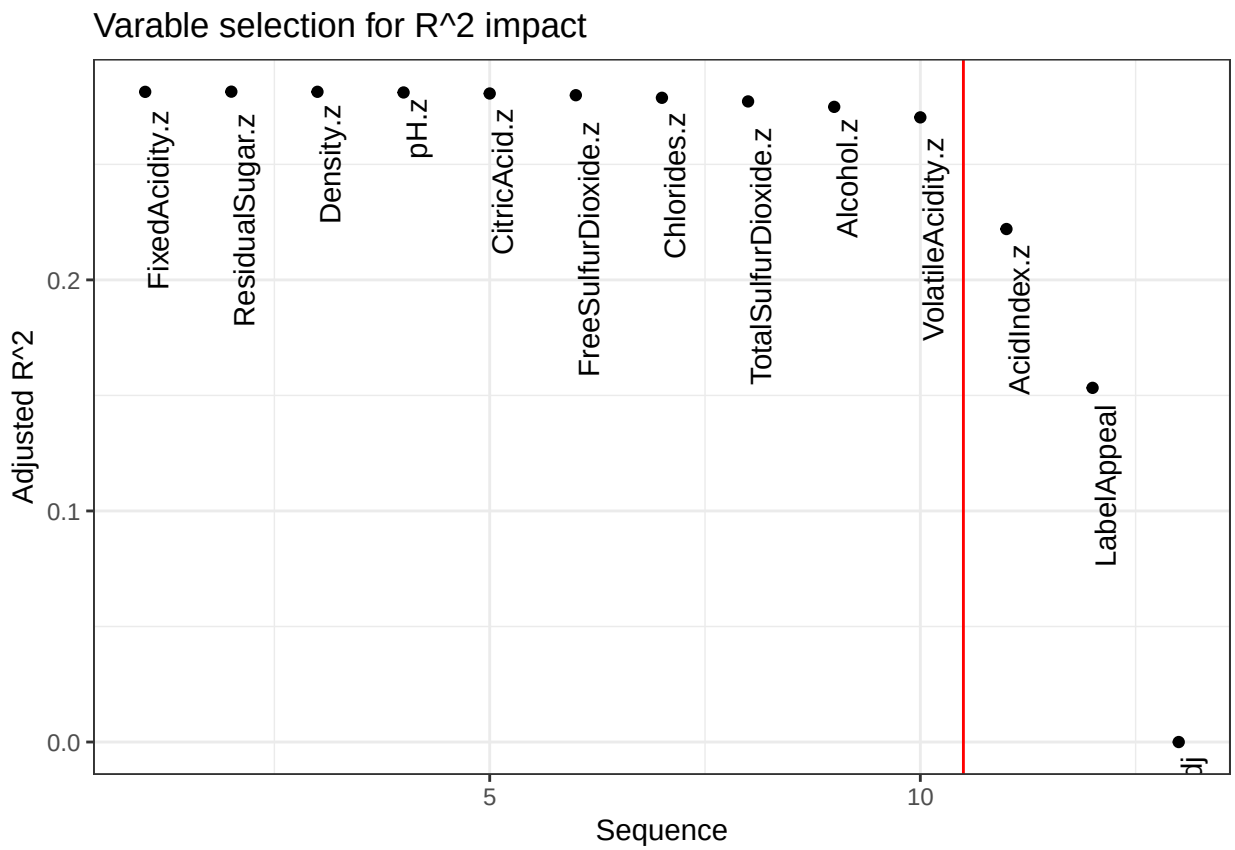
Given the sigma results of each of these "high signal" models when zero values are included, setting the expectation that we should not expect to get an especially low mean absolute error seems justified for this data. The 3-4 best variables all had a standard error (σ) of a little over 1 when zeros were excluded, and nearly 2 with it included.

The full data was split randomly into 70% training data and 30% validation data (for part VI.)

Part II: Steak, greens, Pinot Noir.

Basic.

To test the hypotheses about predictor variables from the EDA portion of the analysis, the automatic variable selection based on adjusted R^2 impact will be applied again.



The three variables that had almost all of the impact on the TARGET variable were AcidIndex, LabelAppeal, and STARS. All three were projected to be useful predictors for the number of cases of wine sold.

Performing a regression on these three variables results in a model of the form:

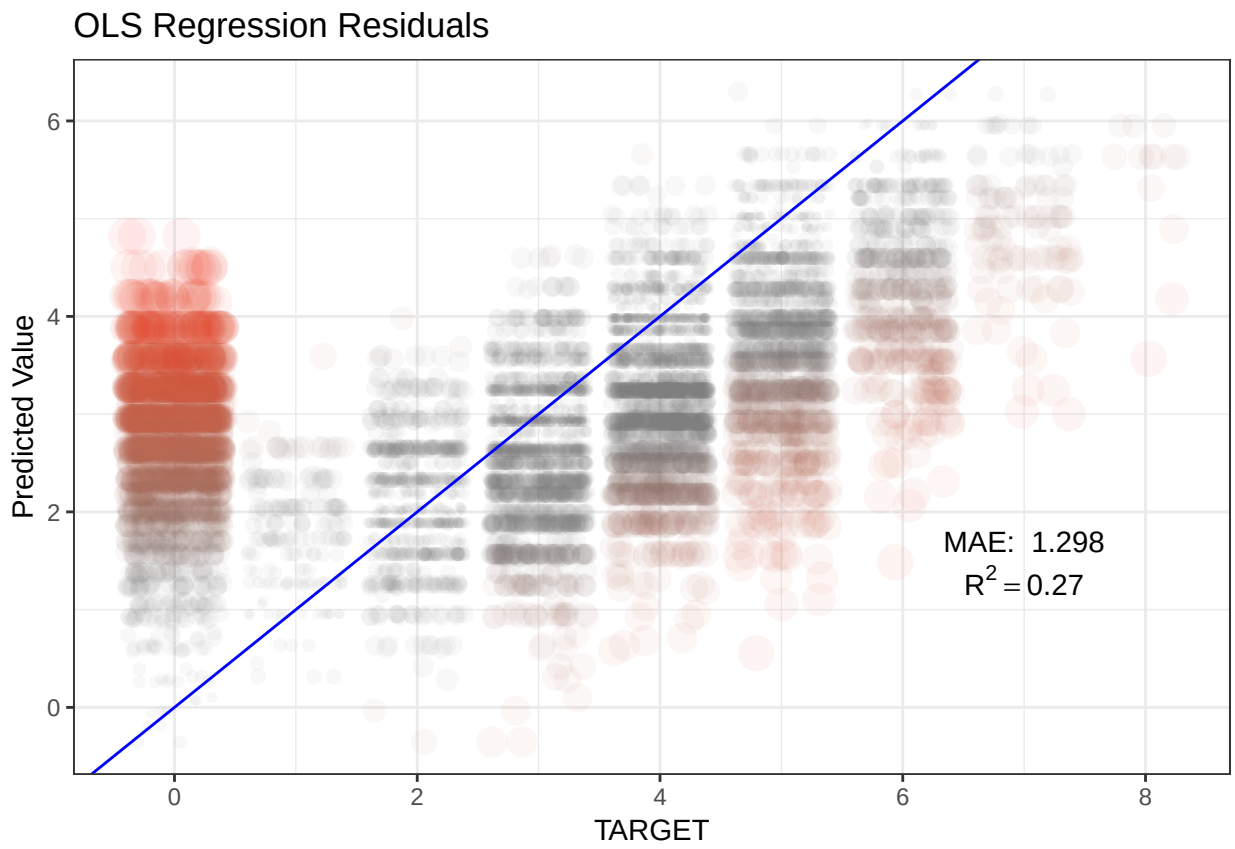
$\hat{y} = 1.5122 + \text{-}0.4238 * \text{AcidIndex.z} + 0.6134 * \text{LabelAppeal} + 0.7417 * \text{STARS.adj}$

These coefficients suggest that, all other variables held constant, a single point increase in AcidIndex corresponds linearly to a decrease of -0.4238 in the TARGET value, whereas LabelAppeal and STARS increase by 0.6134 and 0.7417 respectively.

Simply put, the biggest impact on the number of cases of wine sold is the number of stars it has received, while subjective label appeal has about a 20% smaller impact on sales as compared to the rating (stars). The major driver pushing down demand for a particular wine is the AcidIndex. Since it is unclear what is actually being measured with respect to that variable, all we can reliably say is that higher values drive down sales at a somewhat smaller magnitude than the marginal impact of label appeal.

Residual diagnostics reveal a model that makes not unreasonable predictions, given the nuances of the data, however the errors are distributed heavily at the zero target value prediction. Residual plots begin turning red after they have exceeded the mean absolute error of the residual for the model in order to visually identify where in the distribution of data we are seeing observations that are adversely impacting MAE. Ideally, we would like to see the residuals grouped as close to the blue line as possible.

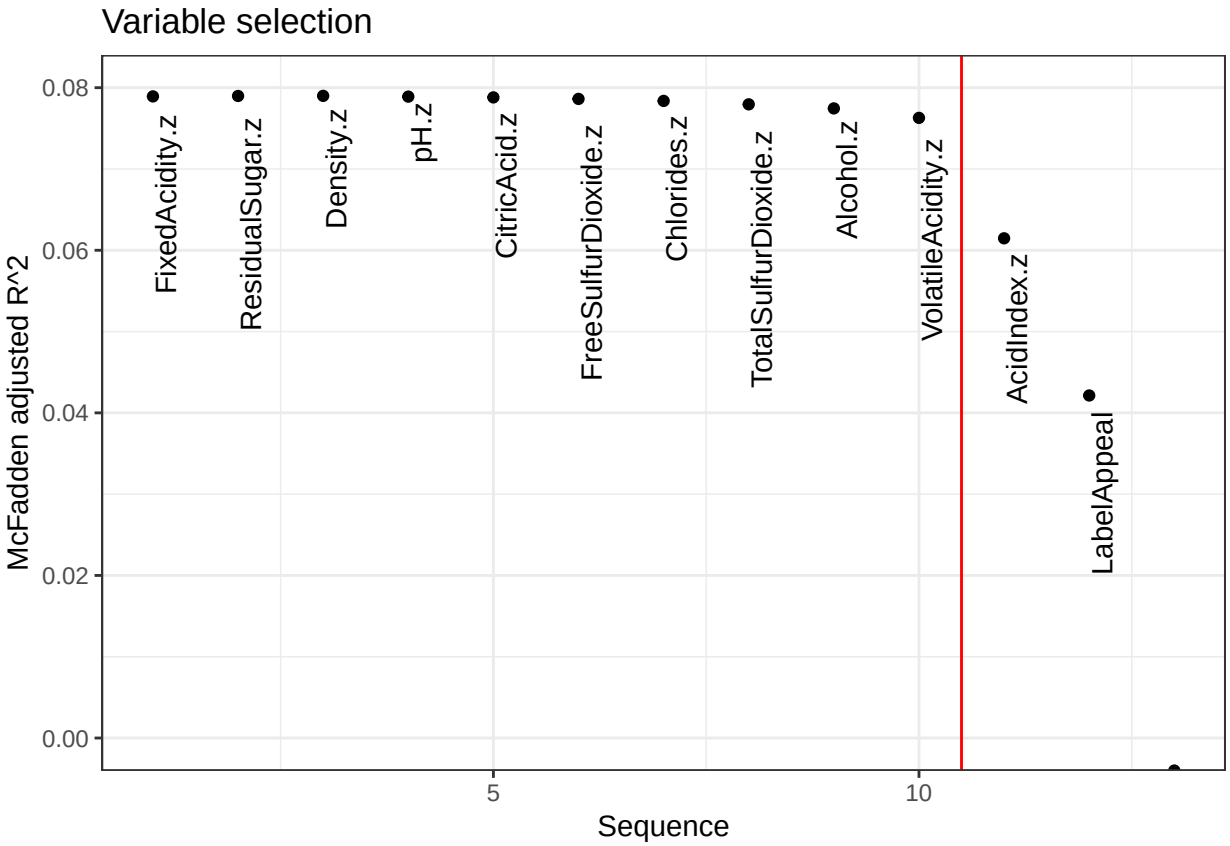
This behavior is not entirely unexpected given the results of the test regression for LabelAppeal, where as the independent variable increases, the regression line will tend to underpredict the response variable.



Part III: Give a man a fish... if he's smart he'll poach it in Chardonnay

...and serve with lemon and capers.

Once again, variable selection logic was adapted to use McFadden adjusted R^2 impact as the criteria for selecting predictors. Once again, AcidIndex, LabelAppeal, and STARS all survived the cut for inclusion in the model.

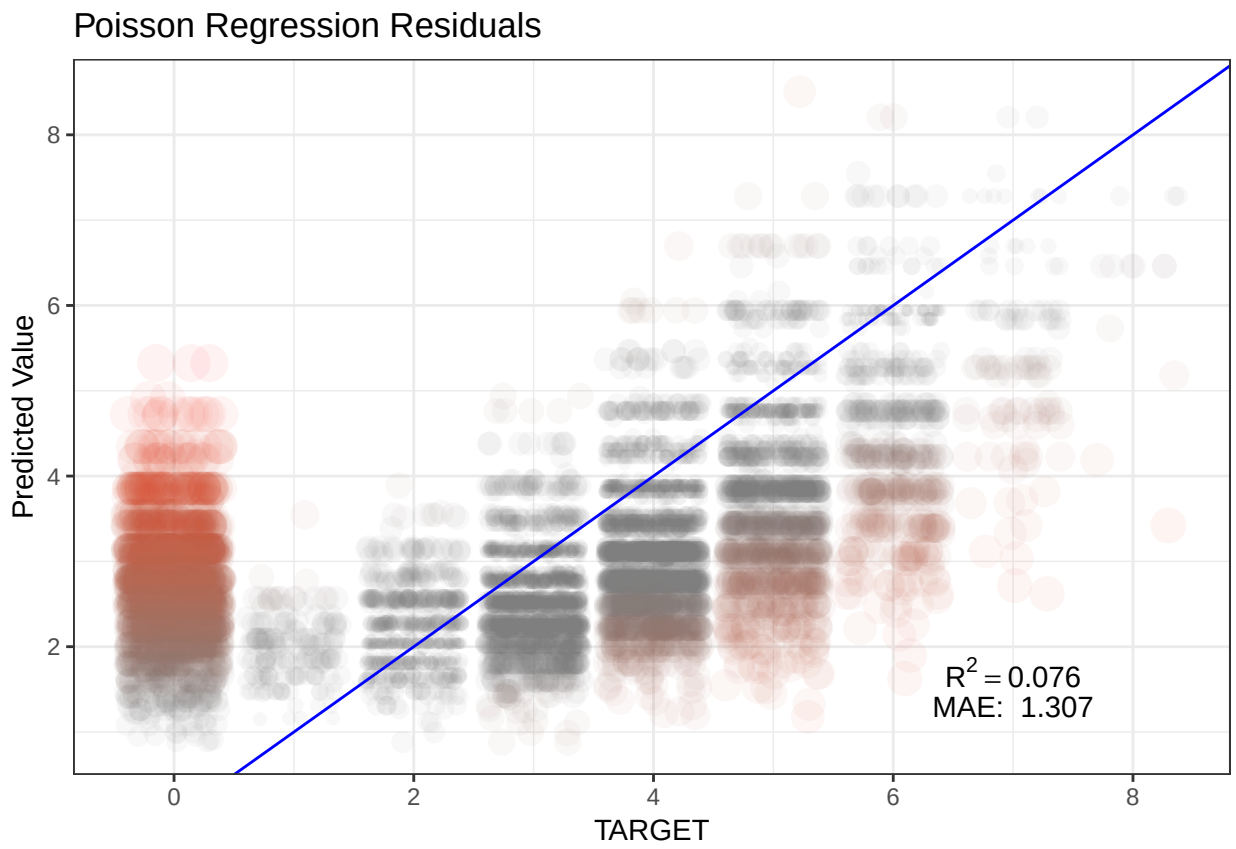


The resulting model is of the form:

$\hat{y}$ = 0.6026 + -0.1583 * AcidIndex.z + 0.2034 * LabelAppeal + 0.2209 * STARS.adj

The coefficients in this model suggest that all things held constant, a single point increase in AcidIndex corresponds with a 14.6% decrease in the TARGET value. Conversely, a single point increase in LabelAppeal and STARS corresponds with increases of 22.6% and 125% in the TARGET value respectively.

In this case, the magnitudes of the coefficients are smaller than the ordinary least squares model generated, but they are proportionally almost identical. Interestingly, this model performed slightly worse in terms of mean absolute error than the previous model. Looking at the residuals, it appears the errors at the zero target value are reduced somewhat, however there is much larger diserpsion as the target value increases



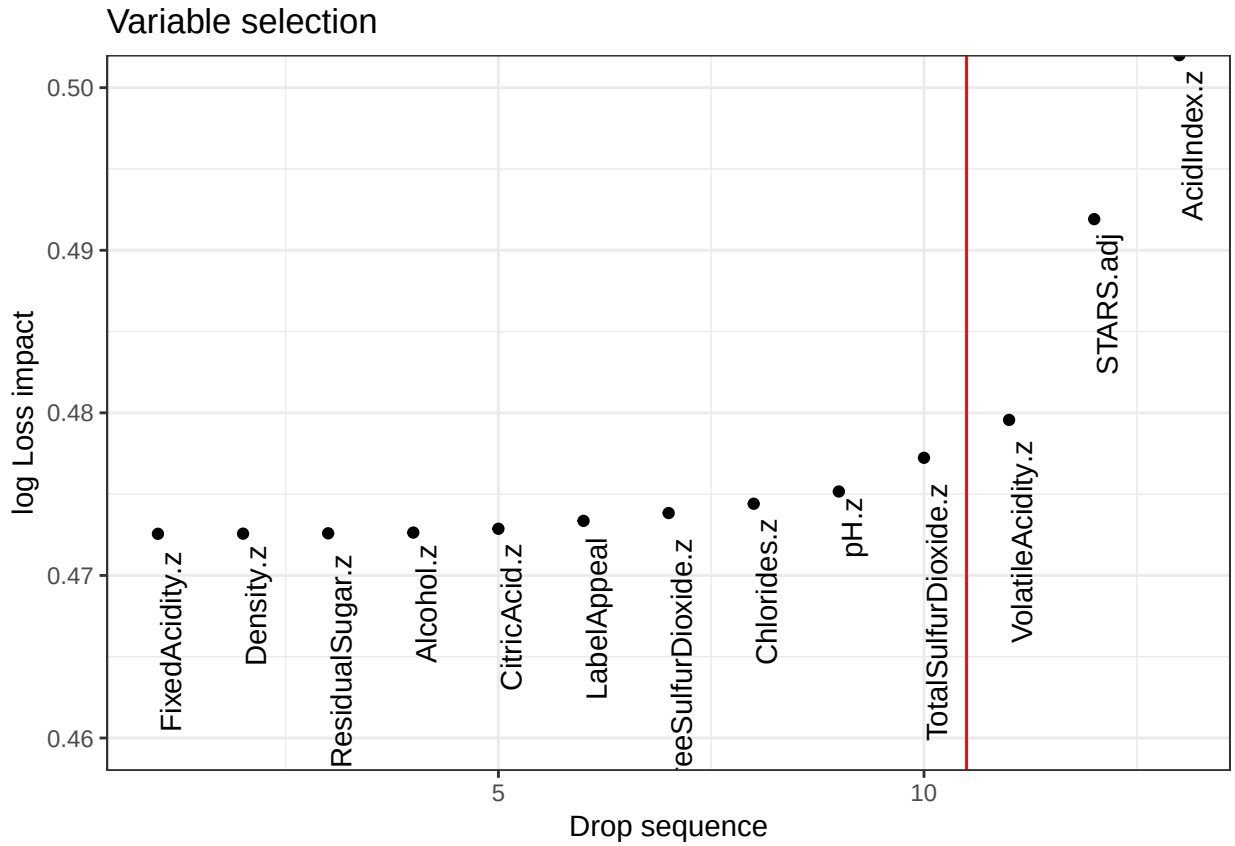
The same predictors were passed to a negative binomial model, which was described by the following equation:

$\hat{y} = 0.6026 + \text{-}0.1583 * \text{AcidIndex.z} + 0.2034 * \text{LabelAppeal} + 0.2209 * \text{STARS.adj}$

This is identical to the Poisson model, which is to be expected, since Poisson is a special case of negative biomial where the mean is equal to the variance. Negative binomial introduces an extra dispersion parameter θ to provide for greater variance in the model. Since in this data, the variance was smaller than the mean, there will be no benefit to negative binomial.

Part IV: *Zi*nfandel

When I ran the log loss variable selecton for the zero TARGET value classification, I was initially surprised to see LabelAppeal perform poorly relative to the other high signal variables, then I went back to my EDA section to review and discovered that it made perfect sense, as I observed that there appeared to be no relationship between TARGET zero and LabelAppeal. Conversely, there are two variables (AcidIndex and VolatileAcidity) relating to the degree of acidity in the wine that seem to be effective predictors of zero sales.



The resulting logistic regression model to identify zero values is of the form:

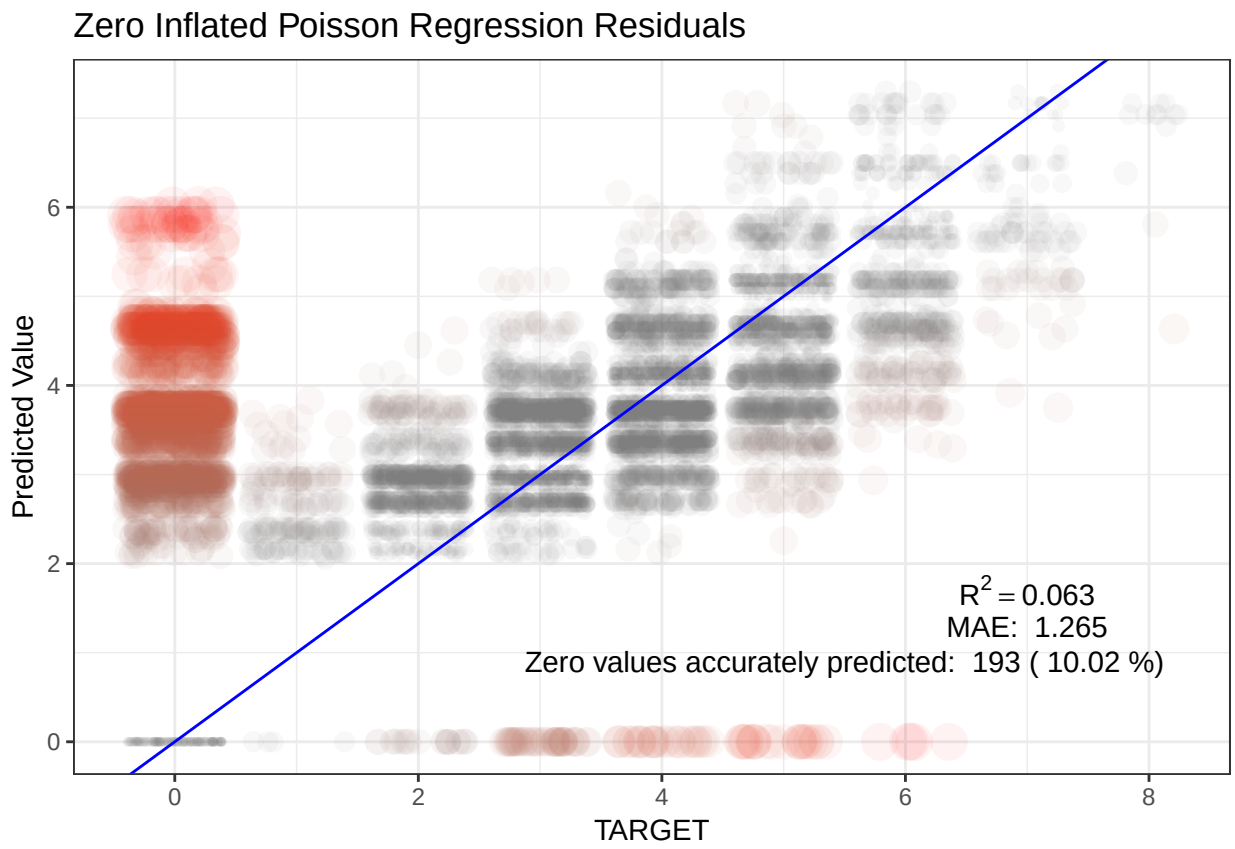
$\hat{y}$ = -0.3092 + 0.1773 * VolatileAcidity.z + -0.5457 * STARS.adj + 0.5468 * AcidIndex.z

These coefficients suggest a reduction of one star in the rating of the wine has a similar impact as a one standard deviation increase in the AcidIndex of the wine, and about three times the impact of a one standard deviation increase in VolatileAcidity. For example, the baseline chance of a wine with three stars having 0 TARGET value is 12.5%, and falls to 7.6% with an additional star. Conversely, a 1 standard devation increase from the mean in AcidIndex would push the probability to 19.8%.

The Poisson regression model for the nonzero values is described by the equation:

$\hat{y}$ = 1.1115 + -0.0238 * AcidIndex.z + 0.226 * LabelAppeal + 0.098 * STARS.adj

Of the models tested thus far, this model clearly performs the best. This model fails to predict any TARGET values below three, save for the zero values predicted by the logisitc regression. This is likely because the two models share two of their predictors, so the two models are highly correlated (negatively, the signs swap on the coefficients for AcidIndex and STARS). On the bright side, this model manages to acccurately predict 193 zero values correctly. The coefficients in this model imply that each unit increase in LabelAppeal is responsible for an 25.4% increase in the TARGET value.



Part V: Portmant(abl)eau

One thing that is fairly obvious about this particular data set is that but for 4 of the predictor variables, the variables are largely useless for predicting the TARGET variable. To illustrate this, we'll test a zero inflated Poisson model with all of the *other* remaining nine variables, to see if we can get anything out of them.

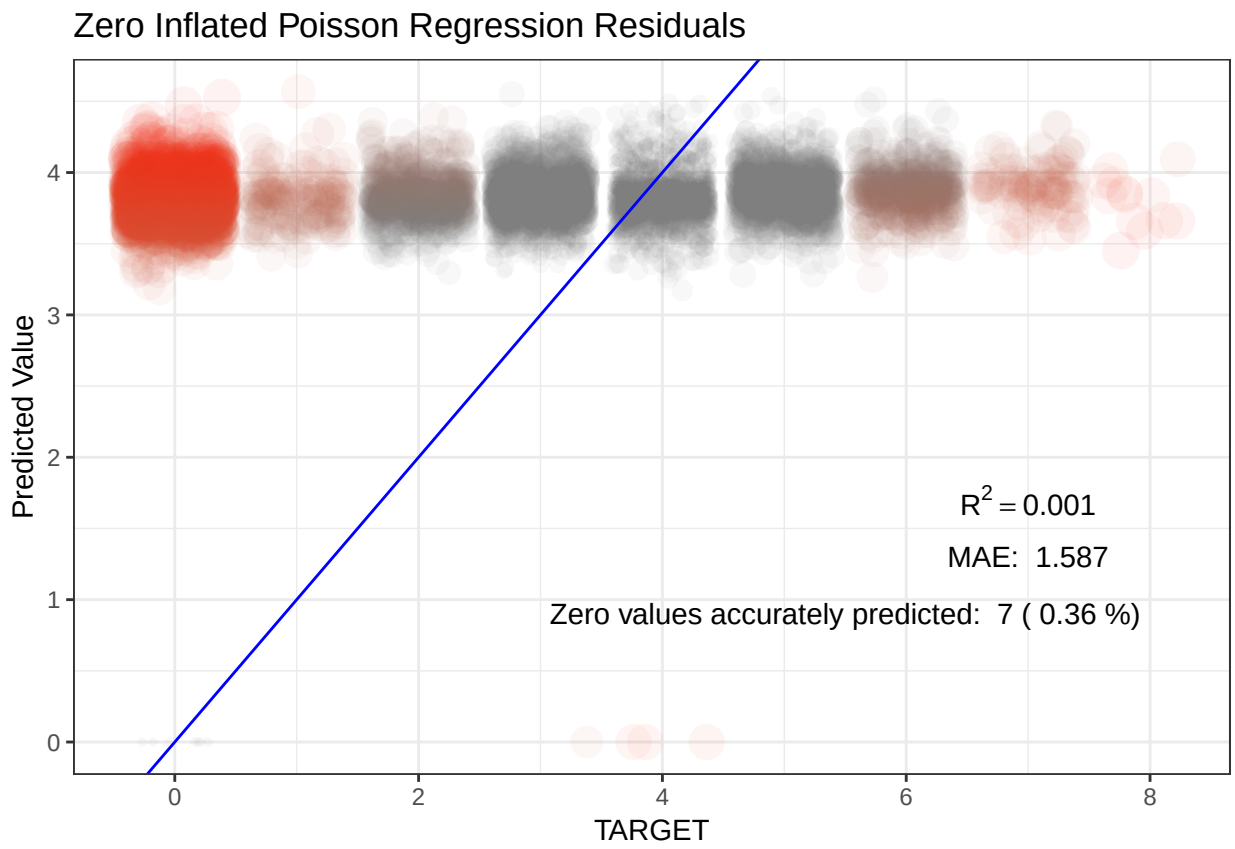
An initial review of a model containing all predictors except those in the ZIP model described in part IV revealed that there remained a strong link between LabelAppeal and sales

The zero classifier model performs so poorly that with a probability threshold of .5, it failed to generate any positive results. Lowering it to .4 resulted in 8 accurate zero predictions. The remaining model actually performs poorly in terms of mean absolute error, at 1.587 versus 1.265 from the three-predictor ZIP model.

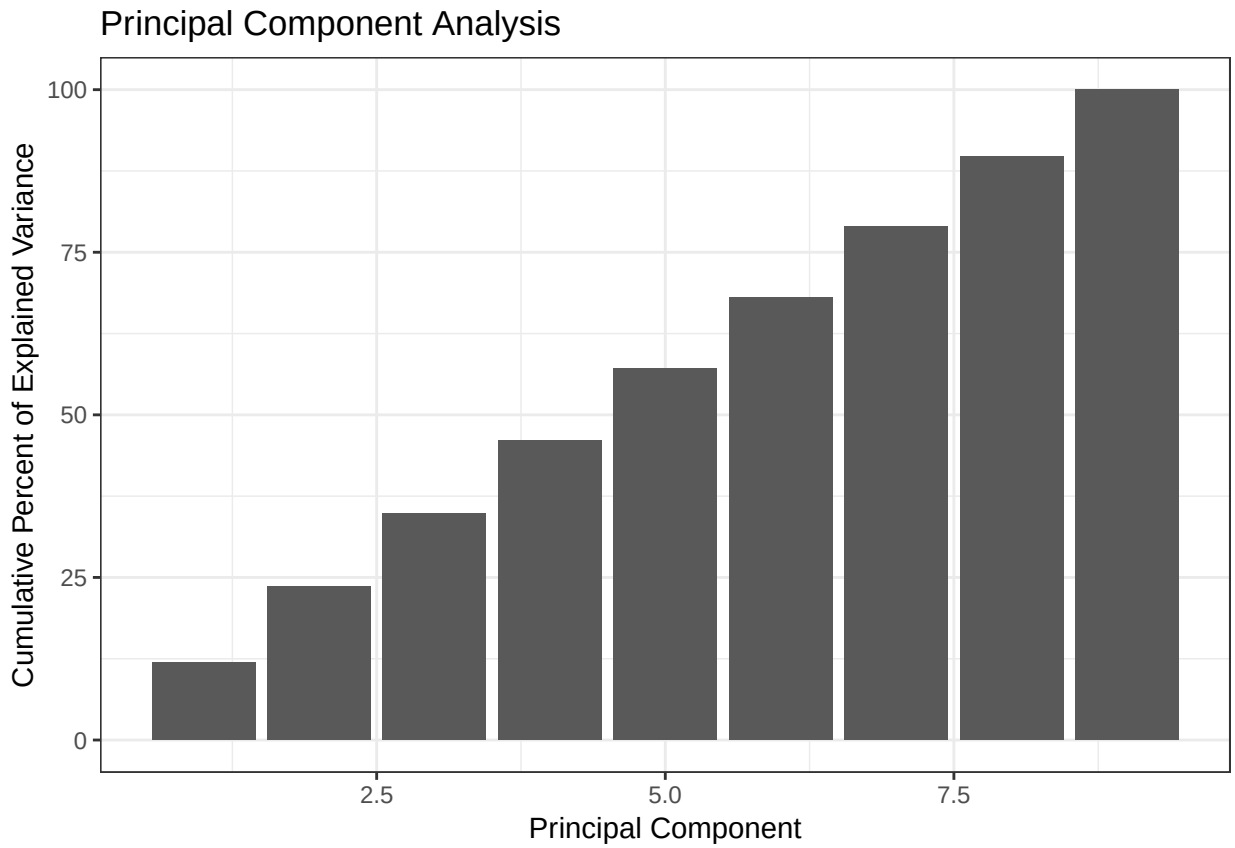
The Poisson model is of the form:

$\hat{y} = 1.3444 + \text{-}0.0037 * \text{FixedAcidity.z} + 0.0025 * \text{ResidualSugar.z} + \text{-}0.009 * \text{Density.z} + 0.0099 * \text{pH.z} + 0.004 * \text{CitricAcid.z} + 0.0053 * \text{FreeSulfurDioxide.z} + \text{-}0.0043 * \text{Chlorides.z} + \text{-}0.0069 * \text{TotalSulfurDioxide.z} + 0.034 * \text{Alcohol.z}$

Based on the results of the residual plot below, we can infer that this model is for all intents no improvement over an intercept-only (mean) model, as just using the mean results in a mean absolute error of 1.5976. It makes a couple accurate zero value predictions, but it trades the benefit of that for a number of inacurrate ones as well. It doesn't appear there is much signal here.

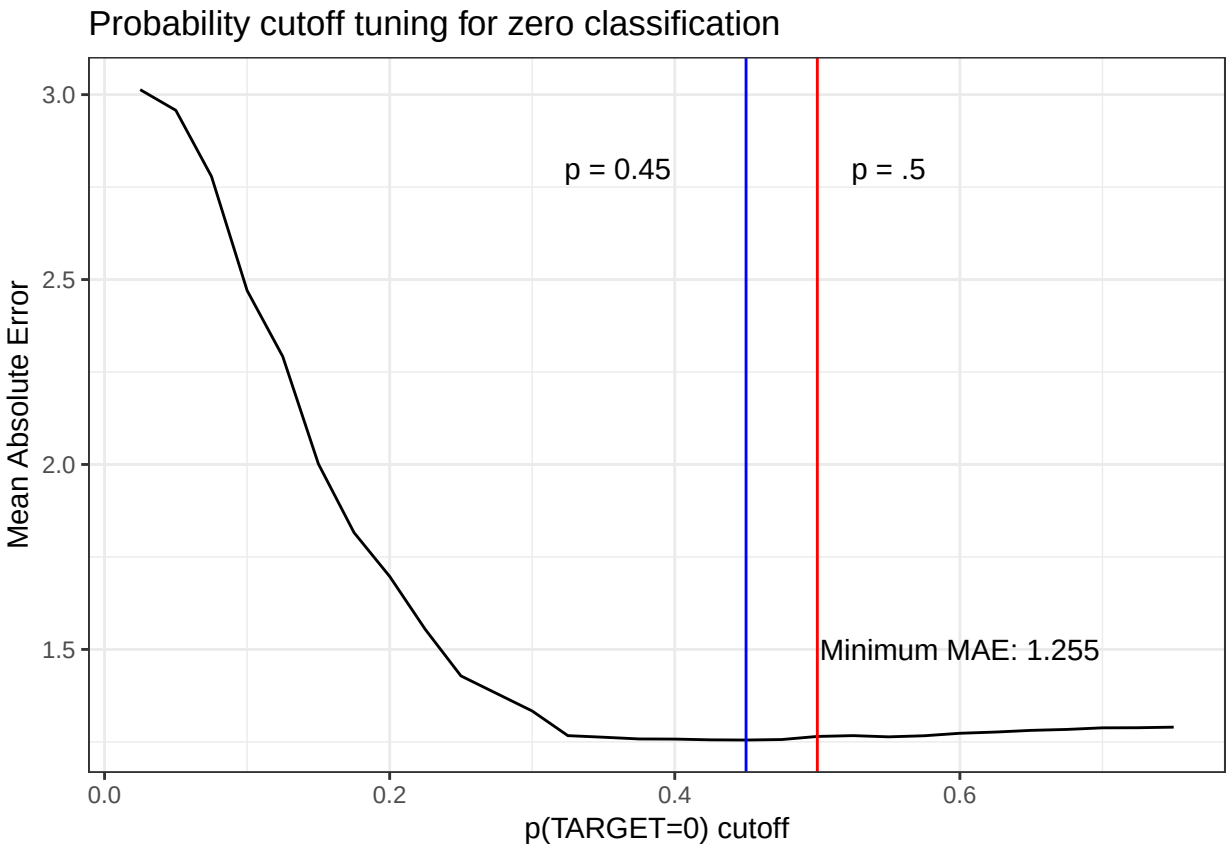


One great way to test that is to reduce dimensionality using PCA. By decorrelating the data and breaking it into the principal components of it's variance, we may be able to create one or more composite predictors to aid in our predictive model. Performing PCA on the remaining 9 variables not included in the Poisson regression reveals no underlying patterns. Each variable contributes almost exactly 11% of the variance, suggesting they are already completely uncorrelated. In fact, one might even think this is completely artificial garbage data, but that seems perhaps conspiratorial. We get enough conspiracy theory in this country as it is these days. If there were underlying correlations in this data, we'd expect to see the cumulative explained variance to start higher than $\frac{1}{k}$ where k is the number of components, and for the incremental increase of each component to fall asymptotically toward zero. Since this is essentially 11% for each component, we can infer that the data is totally uncorrelated, and digging deeper here is unlikely to yield anything. Since we have gained nothing from the PCA decorrelation, we'll stick to the four variables identified and utilized in the zero inflated Poisson model discussed in part IV.



So that leaves us with the remaining four variables already established: AcidIndex, VolatileAcidity, LabelAppeal, and STARS, and those have already been fully modeled in part IV. About all we can do, short of training a deep neural net or a random forest on this data, is to tune the cutoff point for deciding when to classify an observation as a zero, since setting it at greater than .5 resulted in relatively few zero predictions compared to the training data. On the training data, testing each probability cutoff for impact on model mean absolute error reveals that predictive accuracy on training data reaches an optimum value at $p = 0.45$.

The residuals of this model are going to look virtually indistiguishable from the model from part IV, but this obviously performs slightly better on training data. We'll see if that persists on validation data. It could just be overfit, though it has persisted through a couple different random seed settings, which is like a garbage-tier version of cross validation.



Part VI: Grappa or brandy?

¸.·´¯·.·´¯·.·¸¸.·´¯‘·.¸><(((º>

I re-ran the entire workbook with a half dozen or so different random seeds, and the benefit of reducing the probability threshold for identifying a zero target value seemed to persist in all but one of the iterations. We'll call it a provisional win. Note that for OLS, all $\hat{y}$ values less than or equal to 0.5 were treated as zero (ie. rounding down). Interestingly, the standard Poisson regression performed the worst, since it is unable to make any sort of zero predictions (which comprised a large number of the TARGET observations). The zero inflated Poisson performed the best, with the probability threshold tuned version making an almost imperceptable improvement over the typical version.

Note that in the performance comparison below, normal metrics for models like AIC, BIC, or deviance are not included, since the zero inflated Poisson results are not apples-to-apples comparable (the Poisson component does not include any zero TARGET values, thus it is fundamentally different training data, and the zero classifier logistic regression has nothing to compare it to)

Model	MAE	Zero accuracy
Intercept-only (mean)	1.5976	0
OLS	1.3084	0.0124
Poisson	1.3205	0
Zero inflated Poisson	1.2377	0.1943
Zero Inflated Poisson (alt. p threshold)	1.234	0.2698

Ultimately, the best model is the zero inflated Poisson regression, though with the adjusted cutoff for zero classification. Overall, this modeling work I think was largely succcessful, as we were able to reduce the MAE from an intercept-only model by 22.8%.

Residual diagnostics for the zero classifier and no-zero data Poisson models only suggest areas of major deviance ($\triangle D_i >= 4$, Hoffman) occuring where the model confidently predicted a zero, but the ground truth was different (see plot below). We would likely need more variables that contain actual signal to make any improvements in our modeling of this data.