Assignment 3: Wine Sales Project

MSPA PREDICT 411-DL-SEC56

Darryl Buswell

1 Introduction

This document presents results of the third assignment for the Masters of Science in Predictive Analytics course: PREDICT 411. This assessment required the student to build predictive models which are able to predict the number of wine cases ordered by wine distribution companies. To achieve this, we built five predictive models, including a linear regression model and four generalized linear regression models. Each were specified using automated variable selection techniques. As a final step for this assessment, we present a SAS routine which is able to generate predictions of wine orders based on a withheld test set of data.

2 Data

The dataset contains 12,000 data records, with variables which characterize commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The target variable is the number of sample cases of wine that were purchased by wine distribution companies after they had sampled the wine. At a first pass, it seems the dataset has quite a large amount of scope. There are 14 variables tracking a number of attributes. The table below shows a list of variables included in the original dataset.

Table 2.1: Variable Descriptions

Original Variable	Renamed Variable	Description
AcidIndex	N_AcidIndex	Method of testing total acidity of wine
Alcohol	N_Alcohol	Alcohol Content
Chlorides	N_Chlorides	Chloride content of wine
CitricAcid	N_CitricAcid	Citric Acid Content
Density	N_Density	Density of Wine
FixedAcidity	N_FixedAcidity	Fixed Acidity of Wine
FreeSulfurDioxide	N_FreSulfDiox	Sulfur Dioxide content of wine
LabelAppeal	N_LabelAppeal	Marketing Score indicating the appeal of label
ResidualSugar	N_ResSugar	Residual Sugar of wine
STARS	N_STARS	Wine rating by a team of experts
Sulphates	N_Sulphates	Sulfate content of wine
TotalSulfurDioxide	N_TotSulfDiox	Total Sulfur Dioxide of Wine
VolatileAcidity	N_VolAcid	Volatile Acid content of wine
pH	N_pH	pH of wine

For this assessment, we have renamed each variable according to its format type. Since all variables are of numeric type, each have been renamed to include a 'N_' prefix.

3 Data Exploration

Prior to performing any model building, a number of data exploration routines were conducted. These routines allow us to gain an understanding of any potential limitations of the dataset including identifying variables which have missing observations, outlier observations, or those variables which may benefit from transformation.

3.1 Univariate Data Analysis

Summary statistics for each of the numeric variables is shown in the table below.

Table 3.1.1: Data Statistics

Variable	Minimum	Maximum	Mean	Std Dev	N Miss	N
N_FixedAcidity	-18.1	34.4	7.0757171	6.3176435	0	12795
N_VolAcid	-2.79	3.68	0.3241039	0.7840142	0	12795
N_CitricAcid	-3.24	3.86	0.3084127	0.8620798	0	12795
N_ResSugar	-127.8	141.15	5.4187331	33.749379	616	12179
N_Chlorides	-1.171	1.351	0.0548225	0.3184673	638	12157
N_FreSulfDiox	-555	623	30.8455713	148.7145577	647	12148
N_TotSulfDiox	-823	1057	120.7142326	231.9132105	682	12113
N_Density	0.88809	1.09924	0.9942027	0.0265376	0	12795
N_pH	0.48	6.13	3.2076282	0.6796871	395	12400
N_Sulphates	-3.13	4.24	0.5271118	0.9321293	1210	11585
N_Alcohol	-4.7	26.5	10.4892363	3.727819	653	12142
N_LabelAppeal	-2	2	-0.009066	0.8910892	0	12795
N_AcidIndex	4	17	7.7727237	1.3239264	0	12795
N_STARS	1	4	2.041755	0.90254	3359	9436

First, we can see that a number of variables suffer from missing observations and will therefore benefit from some form of imputation. In-fact only six variables were found to not include missing observations. Second, we note a number of inconsistencies in the data. For example, wine cannot have a negative alcohol content, indicating that the minimum value of -4.7 is erroneous (Robinson 2006). Similarly, the minimum residual sugar (-127.8 g/L) is implausible as wines typically have a sugar content greater than 1 g/L (Peynaud 1987). Third, we note that a comparison of the minimum/maximum versus standard deviation value suggests the existence of outliers for a number of variables.

We also looked to compare the mean and variance of the target variable in an attempt to check for the assumption of equality for the Poisson or Negative Binomial distribution. We found the target mean value (3.03) to be quite similar to its variance (3.71), which suggests that we would be in violation of the assumption of equal mean and variance for the Poisson distribution, but not in violation of the assumption that the variance be larger than the mean for the Negative Binomial distribution.

Visualization methods can also be used to gain a greater understanding of each variable. For this assessment, histogram and box plots were generated and reviewed for all numeric variables. We have selected a number of variables for further discussion below.

Figure 3.1.1 Histogram: Density

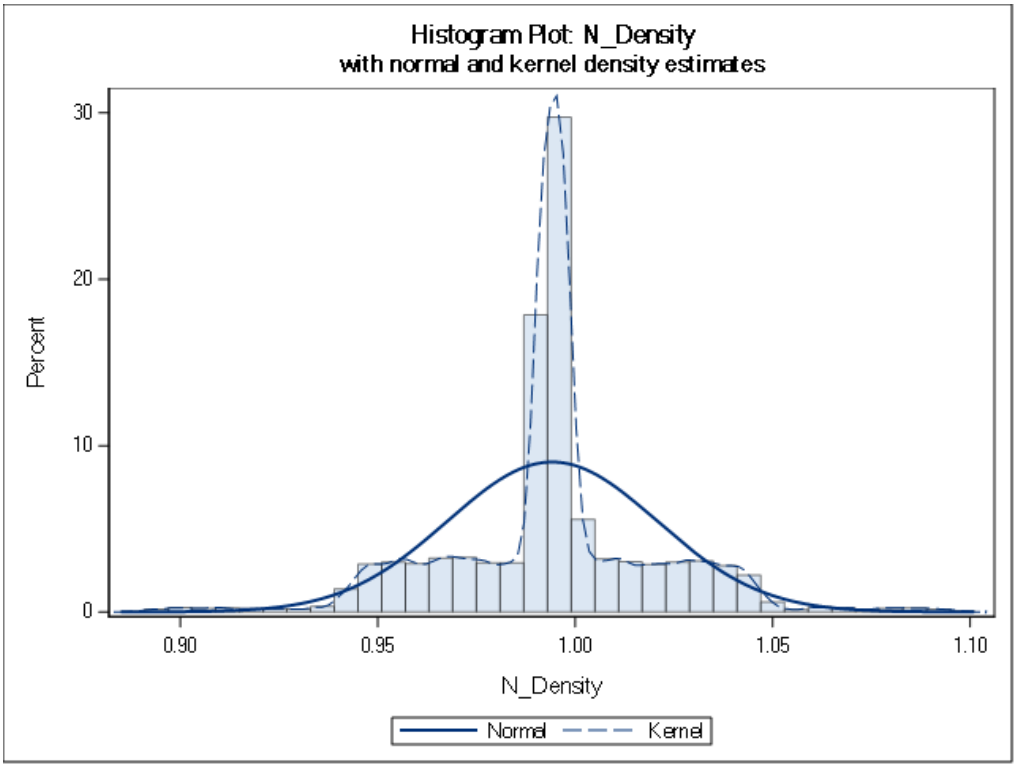


Figure 3.1.2 Box Plot: Density

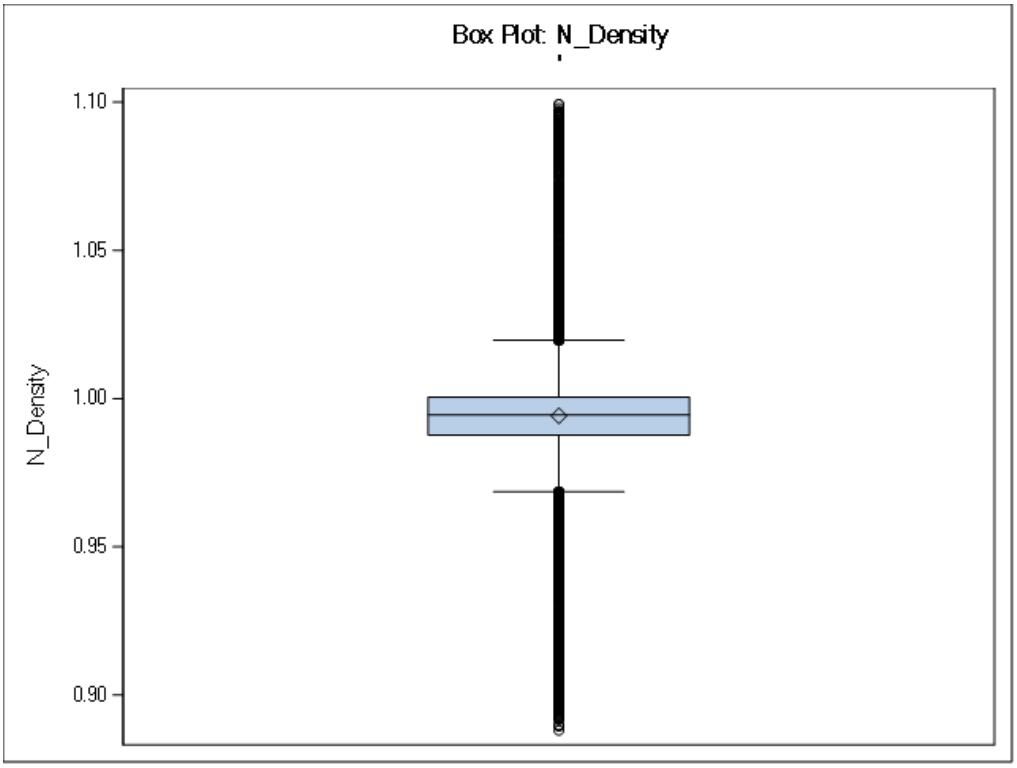


Figure 3.1.3 Histogram: Acid Index

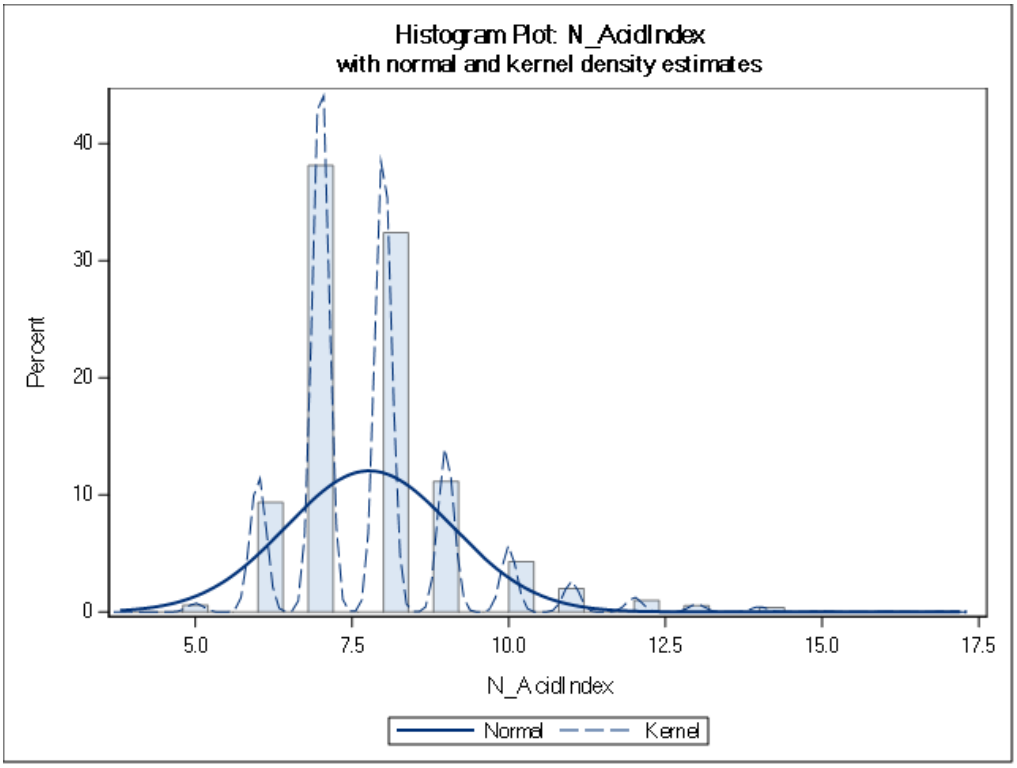
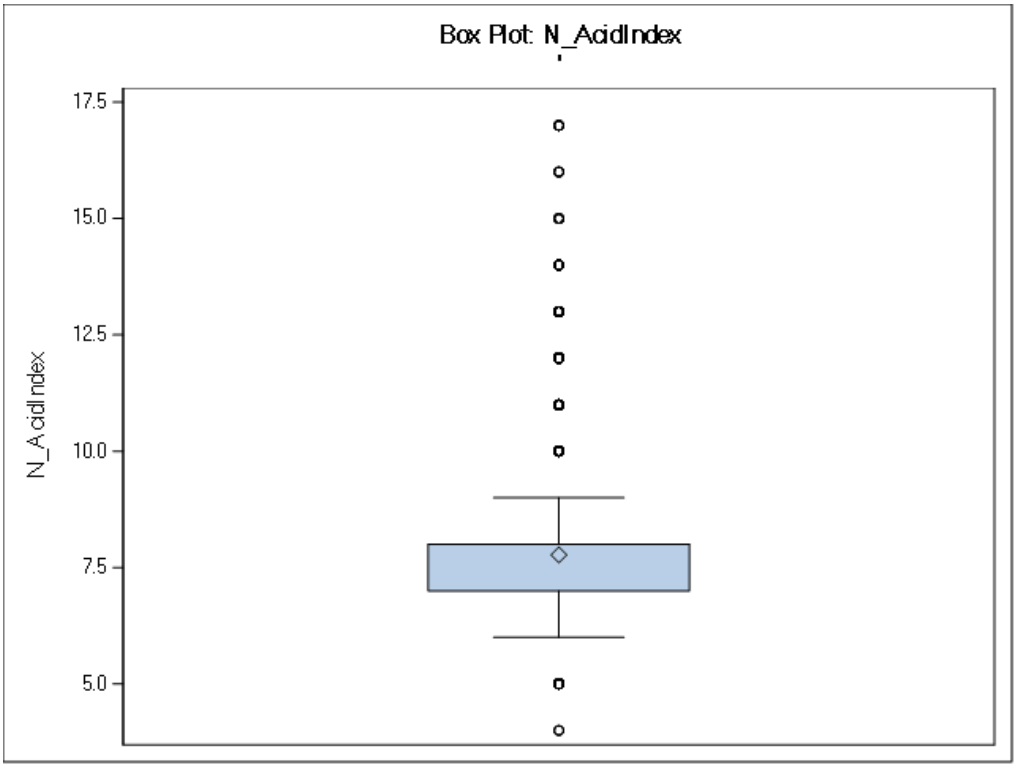


Figure 3.1.4 Box Plot: Acid Index



Many variables share the same characteristics as density (N_Density). That is, a tight collection of observations around the mean and fat tails, suggesting a high amount of kurtosis. Chlorides (N_Chlorides) in particular seems to have the highest amount of kurtosis of each variable. This quality makes outlier identification difficult, as can be seen by the box plot for density above.

There are also some examples of variables which are non-continuous in nature. Acid index (N_AcidIndex) and the appeal label (N_LabelAppeal) for example have their observations binned over certain values. We note that this is also the case for the target variable, which does seem to be normally distributed, however its non-continuous nature and amount of zero observations will cause issues when fitting standard Ordinary Least Square (OLS) based models.

3.2 Bivariate Data Analysis

Since we intend on building a prediction model for the number of purchased cases of wine, we have an interest in those variables which have explanatory power over this variable. As such, we first check the Pearson correlation coefficient in relation between each numeric variable and our target.

The table below summarizes the correlation coefficients between our target and each numeric variable.

Table 3.2.1: Correlations for Purchased Cases vs. Numeric Data

Variable	Correlation
N_FixedAcidity	-0.04901
N_VolAcid	-0.08879
N_CitricAcid	0.00868
N_ResSugar	0.01649
N_Chlorides	-0.03826
N_FreSulfDiox	0.04382
N_TotSulfDiox	0.05148
N_Density	-0.03552
N_pH	-0.00944
N_Sulphates	-0.03885
N_Alcohol	0.06206
N_LabelAppeal	0.3565
N_AcidIndex	-0.24605
N_STARS	0.55879

None of the numeric variables are reported to have a strong positive or negative correlation coefficient with the response variable, with the greatest absolute correlation being reported by wine rating (N_STARS) and appeal label (N_LabelAppeal) at 0.56 and 0.36 respectively. However, it is well established that wine taste is determined by the level of sugar. alcohol, acids and tannins in the beverage (Robinson 2006). This may indicate that a diverse range of wines appealed to the wine tasters and/or their preferences may have been influenced by factors other than taste.

We can also use scatter plots with a Locally Estimated Scatter Plot Smoother (LOESS) overlay to further explore the relationship between the target and each predictor variable. Scatter plots for a collection of variables against the response variable were generated and reviewed with two of these plots selected for further discussion below.

Figure 3.2.1 Scatter: Purchased Cases vs. Density

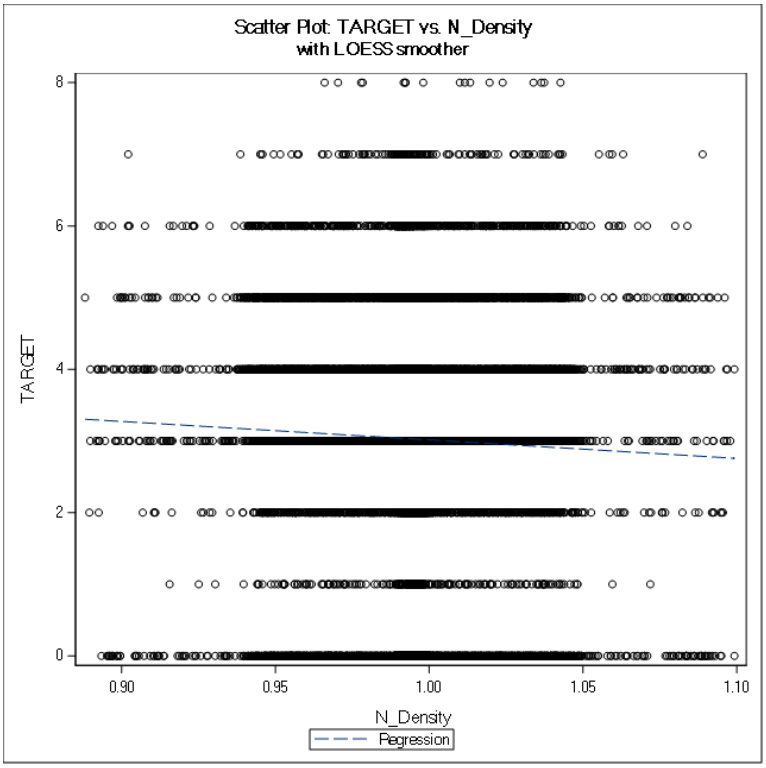
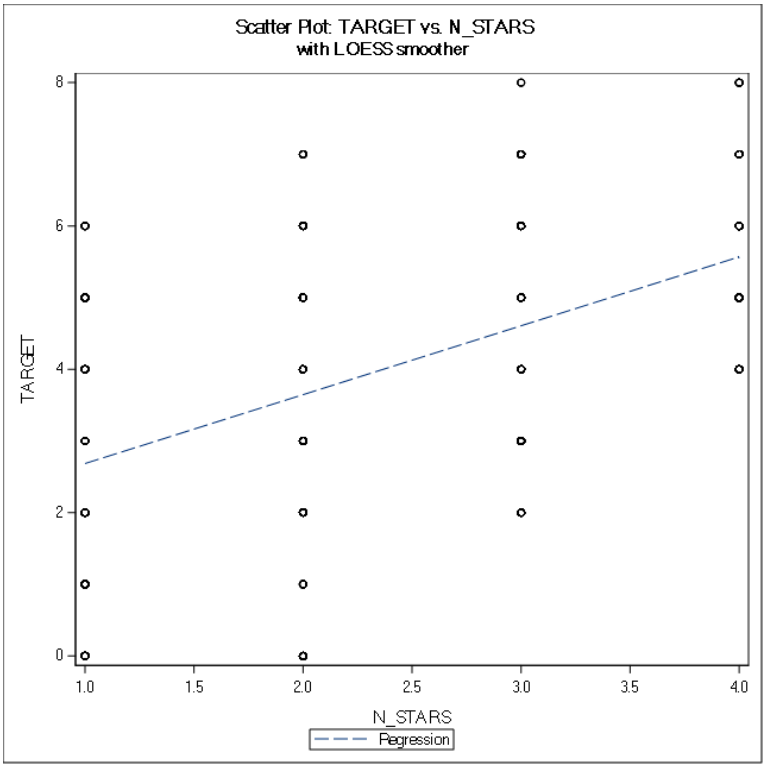


Figure 3.2.2 Scatter: Purchased Cases vs. Stars



We can immediately see the non-continuous nature of the target variable in the scatter plots above. But perhaps more concerning is the lack of relationship between each variable and the target. In fact, all variables other than wine rating (N_STARS) and appeal label (N_LabelAppeal) share a similar scatter plot against the target as density (N_Density) above.

4 Data Preparation

The data preparation routine for this assessment follows a three step process. This includes 1) trimming variables to account for outliers, 2) imputing variables to account for missing values, and finally, 3) performing a log transformation of all existing and newly created variables. Note that during this process, new dummy variables are created in order to reflect any identified outlier or missing observations.

4.1 Data Outliers

From the univariate analysis above, we have identified that the majority of variables have a high amount of kurtosis and what may be considered outlier observations. A review of the percentiles for each variable confirms this, with a rather large gap between the min, max and the 1st and 99th percentile for each variable. For example, levels of 'total sulfur dioxide' span -823.0mg/L to 1,057.0mg/L. Not only is it implausible for total sulfur dioxide levels in wine to be negative, the legal limit for sulfur dioxide in wine in the United States is 350mg/L (N. Jackowetz 2011). It would therefore be probable that values in excess of this are erroneous. A summary of percentiles for total sulfur dioxide and each of the other variables can be found in the table below.

Table 4.1: Quantiles Summary

Variable	Min	0.01	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.99	Max
N_FixedAcidity	-18.1	-10.9	-3.6	-1.2	5.2	6.9	9.5	15.6	17.8	24.4	34.4
N_ResSugar	-2.8	-1.9	-1.0	-0.7	0.1	0.3	0.6	1.4	1.6	2.6	3.7
N_CitricAcid	-3.2	-2.2	-1.2	-0.8	0.0	0.3	0.6	1.4	1.8	2.7	3.9
N_ResSugar	-127.8	-91.0	-52.7	-39.7	-2.0	3.9	15.9	49.8	62.7	99.2	141.2
N_Chlorides	-1.2	-0.9	-0.5	-0.4	0.0	0.0	0.2	0.5	0.6	1.0	1.4
N_FreSulfDiox	-555.0	-388.0	-224.0	-171.0	0.0	30.0	70.0	230.0	284.0	469.0	623.0
N_TotSulfDiox	-823.0	-531.0	-273.0	-185.0	27.0	123.0	208.0	422.0	514.0	767.0	1057.0
N_Density	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.1
N_pH	0.5	1.3	2.1	2.3	3.0	3.2	3.5	4.1	4.4	5.1	6.1
N_Sulphates	-3.1	-2.1	-1.1	-0.7	0.3	0.5	0.9	1.8	2.1	3.2	4.2
N_Alcohol	-4.7	0.1	4.1	5.7	9.0	10.4	12.4	15.2	16.7	20.3	26.5
N_LabelAppeal	-2.0	-2.0	-1.0	-1.0	-1.0	0.0	1.0	1.0	1.0	2.0	2.0
N_AcidIndex	4.0	6.0	6.0	7.0	7.0	8.0	8.0	9.0	10.0	13.0	17.0
N_STARS	1.0	1.0	1.0	1.0	1.0	2.0	3.0	3.0	4.0	4.0	4.0

For this assessment, we have elected to generate trimmed copies of each numeric variable by their 1st/99th, 5th/95th and 10th/90th percentiles. Trimmed variables include the suffix '_T99', '_T95' and '_T90' respectively. Based on the quantile summary above, we note that the majority of variables which are trimmed by their 1st/99th percentiles will still retain many observations which may be classed as outliers. A new set of dummy variables are also created in order to capture those variables which were identified as having outlier observations according to the percentile threshold discussed above. These dummy variables include the suffix '_OF'.

4.2 Missing Data

Following introducing copies of trimmed numeric variables, we then look towards imputing values for missing observations. For this assessment, we elected to impute values according to the variable's median value.

By recalculating the median value for each variable after trimming, we are able to avoid imputing skewed values. Note that in order to simplify the SAS logic used for this assessment, all variables will include the suffix '_IME', however only those variables shown to have missing observations in the previous sections have actually received imputation. A new set of dummy variables are also created in order to capture those variables which were identified as having missing observations. These variables include the suffix '_MF'.

4.3 Data Transformation

We perform a natural logarithm transformation of each of the numeric variables. Variables which have been transformed include the suffix '_LN'. Such a transformation will help penalize extreme values and may provide an improved fit within subsequent regression models. This transformation is performed for all of the newly created variables discussed above.

4.4 Dummy Variables

Finally, we create a number of dummy variables based on bins of wine rating (N_STARS) and appeal label (N_LabelAppeal). The dummy variables created for this assessment are detailed in the table below.

Table 4.2: Dummy Variable Summary

Dummy	Criteria
N_STARS_0	(0.0 <= N_STARS < 0.5);
N_STARS_1	(0.5 <= N_STARS < 1.5);
N_STARS_2	(1.5 <= N_STARS < 2.5);
N_STARS_3	(2.5 <= N_STARS < 3.5);
N_STARS_4	(3.5 <= N_STARS <= 4.0);
N_STARS_GTE2	(1.5 <= N_STARS <= 4.0);
N_STARS_GTE3	(2.5 <= N_STARS <= 4.0);
N_LabelAppeal_1	(-2.0 <= N_LabelAppeal < -1.5);
N_LabelAppeal_2	(-1.5 <= N_LabelAppeal < -0.5);
N_LabelAppeal_3	(-0.5 <= N_LabelAppeal < 0.5);
N_LabelAppeal_4	(0.5 <= N_LabelAppeal < 1.5);
N_LabelAppeal_5	(1.5 <= N_LabelAppeal <= 2.0);
N_LA_GTE3	(-0.5 <= N_LabelAppeal <= 2.0);
N_LA_GTE4	(0.5 <= N_LabelAppeal <= 2.0);

Note that the newly created dummy variables include an appropriate suffix to reflect its criteria.

5 Model Development

For this section, we build five prediction models. This includes a linear regression model and four generalized linear regression models of the following forms; Poisson, Negative Binomial, Zero Inflated Poisson and Zero Inflated Negative Binomial. For each model, we use a stepwise selection technique with a SLENTRY and SLSTAY value of 0.15 in order to determine which variables are included in each specification.

5.1 Model 1: Linear Regression

Parameter estimates for the linear regression model (Model_LinR_S) are shown below.

Table 5.1.1: Linear Regression Parameter Estimates

Variable	DF	Est.	S.E.	t Value	$Pr >	t
Intercept	1	1.41942	1.79746	0.79	0.4297	0
N_Alcohol_OF	1	0.09899	0.04565	2.17	0.0301	1.02754
N_STARS_1	1	1.66956	0.17525	9.53	<.0001	42.05406
N_STARS_GTE2	1	2.38672	0.03189	74.84	<.0001	1.92077
N_LabelAppeal_5	1	0.13799	0.06724	2.05	0.0402	1.25799
N_AcidIndex_IME	1	-0.12299	0.0215	-5.72	<.0001	6.12241
N_AcidIndex_T99_IME	1	-0.58704	0.09362	-6.27	<.0001	96.43776
N_Alcohol_IME	1	0.00733	0.00364	2.02	0.0439	1.31901
N_Alcohol_T90_IME	1	0.1616	0.04896	3.3	0.001	57.08027
N_Chlorides_IME	1	-0.12143	0.03711	-3.27	0.0011	1.00284
N_Density_T90_IME	1	-2.22961	0.92926	-2.4	0.0164	1.00357
N_FreSulfDiox_T99_IME	1	0.00030222	0.0000904	3.34	0.0008	1.00525
N_LabelAppeal_IME	1	0.45554	0.01507	30.22	<.0001	1.36263
N_STARS_IME	1	1.14507	0.32996	3.47	0.0005	494.36929
N_Sulphates_IME	1	-0.03011	0.01299	-2.32	0.0205	1.00297
N_TotSulfDiox_IME	1	0.00021852	0.00005114	4.27	<.0001	1.00597
N_VolAcid_IME	1	-0.09513	0.01473	-6.46	<.0001	1.00719
N_pH_T90_IME	1	-0.13897	0.03582	-3.88	0.0001	1.00728
N_AcidIndex_T99_IME_LN	1	4.60162	0.849	5.42	<.0001	92.23336
N_Alcohol_T90_IME_LN	1	-1.51042	0.54083	-2.79	0.0052	56.7597
N_CitricAcid_T90_IME_LN	1	0.11136	0.03953	2.82	0.0049	1.00515
N_STARS_IME_LN	1	-2.04254	1.2314	-1.66	0.0972	749.03152

For Model_LinR_S, the majority of coefficient estimates have significant p-values at the 95% level, allowing us to reject the null hypothesis and conclude that each have non-zero coefficients. The only exception is the coefficient estimate for N_STARS_IME_LN. We also note that the VIF values for many coefficient estimates suggest that multicollinearity may be an issue.

Goodness-of-fit information for Model_LinR_S is shown below.

Table 5.1.2: Linear Regression Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	25846	1230.75735	726.75	<.0001
Error	12773	21631	1.69352		
Corrected Total	12794	47477			

The model has reported a large F-value suggesting that the observations and regression differ from the grand mean. Likewise, the F-value has a highly significant p-value under the null hypothesis that there is no linear relationship between the predictor and response variable.

Model performance statistics for Model_LinR_S are shown below.

Table 5.1.3: Linear Regression Performance Metrics

Measure	Statistic	Measure	Statistic
MSE	1.6906	R-Square	0.5444
MAE	1.01846	Adj R-Sq	0.5436
Root MSE	1.30135	C(p)	22

Measure	Statistic	Measure	Statistic
Dependent Mean	3.02907	AIC	6762.4748
Coeff Var	42.96203	BIC	6764.5506

The R-square value above suggests that Model_LinR_S explains approximately 54% of the variability in the target using each of the included predictor variables. The adjusted R-squared value indicates a similar level of explanatory power.

5.2 Model 2: GLM: Poisson

Parameter estimates for the Poisson based generalized linear model (Model_Poi_S) are shown below.

Table 5.2.1: GLM: Poisson Parameter Estimates

Parameter	Step	DF	Est.	S.E.	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2487	0.4518	0.3	0.582
N_STARS_GTE2		1	1.0737	0.0183	3454.82	<.0001
N_AcidIndex_IME	4	1	1.2052	0.5481	4.83	0.0279
N_AcidIndex_IME	5	1	1.0712	0.4517	5.62	0.0177
N_AcidIndex_IME	6	1	1.1071	0.4478	6.11	0.0134
N_AcidIndex_IME	7	1	1.0711	0.4476	5.73	0.0167
N_AcidIndex_IME	8	1	1.0392	0.4476	5.39	0.0202
N_AcidIndex_IME	9	1	0.9271	0.4478	4.29	0.0384
N_AcidIndex_IME	10	1	0.7725	0.4485	2.97	0.085
N_AcidIndex_IME	11	1	0.4052	0.451	0.81	0.369
N_AcidIndex_IME	12	1	0.3936	0.4551	0.75	0.3871
N_AcidIndex_IME	13	1	0.5515	0.4572	1.45	0.2278
N_AcidIndex_IME	14	1	0.4552	0.4663	0.95	0.3289
N_AcidIndex_IME	15	1	0.8889	0.5126	3.01	0.0829
N_AcidIndex_IME	16	1	0.2454	0.6327	0.15	0.6981
N_Alcohol_T90_IME		1	0.0108	0.0029	14.29	0.0002
N_Chlorides_IME		1	-0.0383	0.0165	5.4	0.0201
N_LabelAppeal_IME	-2	1	-0.6994	0.0424	271.49	<.0001
N_LabelAppeal_IME	-1	1	-0.4574	0.025	334.89	<.0001
N_LabelAppeal_IME	0	1	-0.2679	0.0229	137.24	<.0001
N_LabelAppeal_IME	1	1	-0.1348	0.0232	33.83	<.0001
N_STARS_IME	1	1	0.5163	0.028	339.6	<.0001
N_STARS_IME	2	1	-0.2394	0.0199	144.57	<.0001
N_STARS_IME	3	1	-0.1221	0.0202	36.48	<.0001
N_TotSulfDiox_IME		1	0.0001	0	10.1	0.0015
N_VolAcid_IME		1	-0.0291	0.0065	19.86	<.0001
N_pH_T90_IME		1	-0.043	0.0159	7.3	0.0069
N_CitricAcid_T90_IME		1	0.0261	0.0127	4.22	0.0399
N_FreSulfDiox_IME_LN		1	0.0034	0.0013	6.25	0.0124

An assessment of coefficients can be achieved by taking the $100 \cdot (exp(\beta) - 1)$ of its estimate. For example, the coefficient estimate for N_Alcohol_T90_IME suggests that a one unit increase in alcohol translates to a 1.1% increase in the number of wine cases purchased. Likewise, the coefficient estimates with a step value indicates a percentage change in the target based on that value of the predictor. For example, the coefficient estimate for N_LabelAppeal_IME with a step of -2 suggests a 50% decrease in the amount of cases purchased when appeal is equal to -2.

With this in mind, we are critical of a number of coefficient estimates for the above model. For instance, we would not expect such a large magnitude of change in the number of cases purchased from a one unit increase in many of the included predictors. We do note that the polarity of coefficient for the appeal variables seems appropriate. However, it is more difficult to assess the polarity of coefficient estimate for the chemical properties, since we are effectively assessing perceived quality. For example, our research suggests that taste is a balance of acidity, tannins, sugar and alcohol, with relative changes in these quantities influencing the sweetness, sourness and bitterness (Robinson 2006). In contrast, estimates relating to N_VolAcid_IME were intuitive. The model output indicated that as the volatile acidity of wine increases, the volume of wine purchased tends to decrease which would be due to the increasingly unpalatable levels of acetic acid in the wine (Neeley 2015).

The results, along with model performance criteria are shown below.

Table 5.2.3: GLM: Poisson Performance Metrics

Criterion	DF	Value	Value/DF
Deviance	13000	13522.5677	1.0593
Scaled Deviance	13000	13522.5677	1.0593
Pearson Chi-Square	13000	11177.4783	0.8756
Scaled Pearson X2	13000	11177.4783	0.8756
Log Likelihood		8864.8766	
Full Log Likelihood		-22732.2947	
AIC (smaller is better)		45522.5894	
AICC (smaller is better)		45522.7257	
BIC (smaller is better)		45738.8369	

We note the AIC and BIC values above, as these will be used as a comparison against other specifications.

5.3 Model 3: GLM: Negative Binomial

Parameter estimates for the Negative Binomial based generalized linear model (Model_NB_S) are shown below.

Table 5.3.1: GLM: Negative Binomial Parameter Estimates

Parameter	Step	DF	Est.	S.E.	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2487	0.4518	0.3	0.582
N_STARS_GTE2		1	1.0737	0.0183	3454.82	<.0001
N_AcidIndex_IME	4	1	1.2052	0.5481	4.83	0.0279
N_AcidIndex_IME	5	1	1.0712	0.4517	5.62	0.0177
N_AcidIndex_IME	6	1	1.1071	0.4478	6.11	0.0134
N_AcidIndex_IME	7	1	1.0711	0.4476	5.73	0.0167
N_AcidIndex_IME	8	1	1.0392	0.4476	5.39	0.0202
N_AcidIndex_IME	9	1	0.9271	0.4478	4.29	0.0384
N_AcidIndex_IME	10	1	0.7725	0.4485	2.97	0.085
N_AcidIndex_IME	11	1	0.4052	0.451	0.81	0.369
N_AcidIndex_IME	12	1	0.3936	0.4551	0.75	0.3871
N_AcidIndex_IME	13	1	0.5515	0.4572	1.45	0.2278
N_AcidIndex_IME	14	1	0.4552	0.4663	0.95	0.3289
N_AcidIndex_IME	15	1	0.8889	0.5126	3.01	0.0829
N_AcidIndex_IME	16	1	0.2454	0.6327	0.15	0.6981
N_Alcohol_T90_IME		1	0.0108	0.0029	14.29	0.0002
N_Chlorides_IME		1	-0.0383	0.0165	5.4	0.0201

Parameter	Step	DF	Est.	S.E.	Wald Chi-Square	Pr > ChiSq
N_LabelAppeal_IME	-2	1	-0.6994	0.0424	271.49	<.0001
N_LabelAppeal_IME	-1	1	-0.4574	0.025	334.89	<.0001
N_LabelAppeal_IME	0	1	-0.2679	0.0229	137.24	<.0001
N_LabelAppeal_IME	1	1	-0.1348	0.0232	33.83	<.0001
N_STARS_IME	1	1	0.5163	0.028	339.6	<.0001
N_STARS_IME	2	1	-0.2394	0.0199	144.57	<.0001
N_STARS_IME	3	1	-0.1221	0.0202	36.48	<.0001
N_TotSulfDiox_IME		1	0.0001	0	10.1	0.0015
N_VolAcid_IME		1	-0.0291	0.0065	19.86	<.0001
N_pH_T90_IME		1	-0.043	0.0159	7.3	0.0069
N_CitricAcid_T90_IME		1	0.0261	0.0127	4.22	0.0399
N_FreSulfDiox_IME_LN		1	0.0034	0.0013	6.25	0.0124

We see that the automated variable selection technique has mirrored the same specification as Model_Poi_S. This is perhaps not surprising due to the target variance being close to equal its mean.

Model performance statistics for Model_NB_S are shown below.

Table 5.3.2: GLM: Negative Binomial Performance Metrics

Criterion	DF	Value	Value/DF
Deviance	13000	13522.5677	1.0593
Scaled Deviance	13000	13522.5677	1.0593
Pearson Chi-Square	13000	11177.4694	0.8756
Scaled Pearson X2	13000	11177.4694	0.8756
Log Likelihood		8864.8766	
Full Log Likelihood		-22732.2947	
AIC (smaller is better)		45524.5894	
AICC (smaller is better)		45524.7351	
BIC (smaller is better)		45748.2937	

We see that the AIC and BIC for Model_NB_S is slightly higher (worse) than that of Model_Poi_S. However, there is no difference between the ratio of deviance to degree of freedom. We may have expected a Negative Binomial based model to perform better than a Poisson based model, as it allows for some overdispersion, however in this case, it seems that overdispersion is not an issue.

5.4 Model 4: GLM: Zero Inflated Poisson

Parameter estimates for the Zero Inflated Poisson generalized linear model (Model_ZPoi_S) are shown below.

Table 5.4.1: GLM: Poisson Parameter Estimates

Parameter	Step	DF	Est.	S.E.	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.4688	0.0677	470.47	<.0001
N_STARS_GTE2		1	0.4752	0.0285	277.32	<.0001
N_AcidIndex_T95_IME		1	-0.0249	0.006	17.46	<.0001
N_Alcohol_IME		1	0.0041	0.0017	6.21	0.0127
N_Alcohol_T90_IME		1	0.0095	0.0034	8.04	0.0046
N_LabelAppeal_IME	-2	1	-1.0236	0.0458	499.52	<.0001
N_LabelAppeal_IME	-1	1	-0.6331	0.026	591.12	<.0001

Parameter	Step	DF	Est.	S.E.	Wald Chi-Square	Pr > ChiSq
N_LabelAppeal_IME	0	1	-0.3518	0.0233	227.54	<.0001
N_LabelAppeal_IME	1	1	-0.1637	0.0235	48.45	<.0001
N_STARS_IME	1	1	0.1579	0.0351	20.23	<.0001
N_STARS_IME	2	1	-0.1822	0.0199	83.38	<.0001
N_STARS_IME	3	1	-0.099	0.0202	24.03	<.0001
N_VolAcid_IME		1	-0.018	0.0068	7.02	0.008

Table 5.4.2: GLM: Zero Inflated Poisson Parameter Estimates

Parameter		DF	Est.	S.E.	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.394	756.2719	0	0.9954
N_STARS_IME	1	1	18.2319	674.7371	0	0.9784
N_STARS_IME	2	1	18.5193	674.7371	0	0.9781
N_STARS_IME	3	1	-7.4969	796.0038	0	0.9925
N_LabelAppeal_IME	-2	1	-2.8423	0.5858	23.54	<.0001
N_LabelAppeal_IME	-1	1	-1.4548	0.1697	73.47	<.0001
N_LabelAppeal_IME	0	1	-0.8331	0.1493	31.16	<.0001
N_LabelAppeal_IME	1	1	-0.4174	0.1518	7.56	0.006
N_AcidIndex_IME	4	1	-13.7932	341.5851	0	0.9678
N_AcidIndex_IME	5	1	-15.3206	341.5814	0	0.9642
N_AcidIndex_IME	6	1	-15.0953	341.581	0	0.9648
N_AcidIndex_IME	7	1	-15.032	341.581	0	0.9649
N_AcidIndex_IME	8	1	-14.703	341.581	0	0.9657
N_AcidIndex_IME	9	1	-14.0103	341.581	0	0.9673
N_AcidIndex_IME	10	1	-13.331	341.581	0	0.9689
N_AcidIndex_IME	11	1	-12.3995	341.581	0	0.971
N_AcidIndex_IME	12	1	-12.072	341.5811	0	0.9718
N_AcidIndex_IME	13	1	-12.4809	341.5811	0	0.9709
N_AcidIndex_IME	14	1	-12.1375	341.5812	0	0.9717
N_AcidIndex_IME	15	1	-13.3732	341.582	0	0.9688
N_AcidIndex_IME	16	1	0.3679	434.0462	0	0.9993

We again note similarity in the final predictors included for Model_ZPoi_S and the previously discussed models. However, the zero inflated parameter estimates suggest that the above model has aided in treating the high amount of zero observations within the label appeal variable. The key difference between this and earlier models is the production of two tables - the Poisson Count Model and Logit Model for Predicting Excess Zeros. (Digital Research and Education 2016) notes that this approach assumes that the excess zeros in the model are produced by a separate process and therefore need to be modelled independently.

Similar to other models tested above, the estimates for N_STARS_IME are not intuitive. The data dictionary defines wines with higher STARS ratings (e.g. 3-4 stars) as being of a higher quality than wines with lower STARS ratings (e.g. 1-2 stars). However, the parameter estimates show that higher STARS ratings do not necessarily correspond with more wine purchases.

Model performance statistics for Model_ZPoi_S are shown below.

Table 5.4.3: GLM: Zero Inflated Poisson Performance Metrics

Criterion	DF	Value	Value/DF
Deviance		43577.0023	
Scaled Deviance		43577.0023	

Criterion	DF	Value	Value/DF
Pearson Chi-Square	13000	6704.5151	0.5254
Scaled Pearson X2	13000	6704.5151	0.5254
Log Likelihood		9808.6701	
Full Log Likelihood		-21788.5011	
AIC (smaller is better)		43645.0023	
AICC (smaller is better)		43645.1888	
BIC (smaller is better)		43898.5338	

We see an improvement in both the AIC and BIC for Model_ZPoi_S is lower (better) than that of Model_Poi_S.

5.5 Model 5: GLM: Zero Inflated Negative Binomial

Parameter estimates for the Zero Inflated Poisson generalized linear model (Model_ZNB_S) are shown below.

Table 5.5.1: GLM: Negative Binomial Parameter Estimates

Parameter	Step	DF	Est.	S.E.	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.4629	0.0681	461.64	<.0001
N_STARS_GTE2		1	0.4813	0.0289	278.01	<.0001
N_AcidIndex_T95_IME		1	-0.0249	0.006	17.25	<.0001
N_Alcohol_IME		1	0.0042	0.0017	6.25	0.0124
N_Alcohol_T90_IME		1	0.0094	0.0034	7.79	0.0053
N_LabelAppeal_IME	-2	1	-1.0086	0.0456	490.13	<.0001
N_LabelAppeal_IME	-1	1	-0.6308	0.0262	579.38	<.0001
N_LabelAppeal_IME	0	1	-0.3507	0.0235	222.94	<.0001
N_LabelAppeal_IME	1	1	-0.1633	0.0237	47.57	<.0001
N_STARS_IME	1	1	0.1625	0.0354	21.08	<.0001
N_STARS_IME	2	1	-0.1828	0.02	83.12	<.0001
N_STARS_IME	3	1	-0.0994	0.0203	23.94	<.0001
N_VolAcid_IME		1	-0.0185	0.0068	7.4	0.0065

The Negative Binomial Parameter Estimates technique has produced similar results to Model_ZPoi_S and the previously discussed models.

Table 5.5.2: GLM: Zero Inflated Negative Binomial Parameter Estimates

Parameter		DF	Est.	S.E.	Wald Chi-Square	Pr > ChiSq
Intercept		1	-6.8329	1.354	25.47	<.0001
N_STARS_IME	1	1	3.2252	0.3943	66.9	<.0001
N_STARS_IME	2	1	3.521	0.3912	81.02	<.0001
N_STARS_IME	3	1	-0.8452	0.5111	2.73	0.0982
N_LabelAppeal_IME	-2	1	-2.3416	0.4016	34	<.0001
N_LabelAppeal_IME	-1	1	-1.4016	0.1678	69.76	<.0001
N_LabelAppeal_IME	0	1	-0.7878	0.1476	28.49	<.0001
N_LabelAppeal_IME	1	1	-0.3817	0.1499	6.49	0.0109
N_AcidIndex_IME	4	1	3.8158	2.0135	3.59	0.0581
N_AcidIndex_IME	5	1	2.2842	1.3622	2.81	0.0936
N_AcidIndex_IME	6	1	2.3033	1.2944	3.17	0.0752
N_AcidIndex_IME	7	1	2.3539	1.2895	3.33	0.0679

Parameter		DF	Est.	S.E.	Wald Chi-Square	Pr > ChiSq
N_AcidIndex_IME	8	1	2.6837	1.2897	4.33	0.0375
N_AcidIndex_IME	9	1	3.3782	1.2912	6.85	0.0089
N_AcidIndex_IME	10	1	4.0638	1.2939	9.86	0.0017
N_AcidIndex_IME	11	1	4.989	1.299	14.75	0.0001
N_AcidIndex_IME	12	1	5.2714	1.3105	16.18	<.0001
N_AcidIndex_IME	13	1	4.9162	1.3232	13.8	0.0002
N_AcidIndex_IME	14	1	5.2288	1.3404	15.22	<.0001
N_AcidIndex_IME	15	0	4.2238	0	.	.
N_AcidIndex_IME	16	1	4.3348	1.7154	6.39	0.0115

Model performance statistics for Model_ZNB_S are shown below.

Table 5.5.3: GLM: Zero Inflated Negative Binomial Performance Metrics

Criterion	DF	Value	Value/DF
Deviance		43695.6413	
Scaled Deviance		43695.6413	
Pearson Chi-Square	13000	6695.4366	0.5247
Scaled Pearson X2	13000	6695.4366	0.5247
Log Likelihood		-21847.8206	
Full Log Likelihood		-21847.8206	
AIC (smaller is better)		43765.6413	
AICC (smaller is better)		43765.8388	
BIC (smaller is better)		44026.6296	

6 Model Selection

In order to select the optimum model, we have collated and assessed the AIC, AICC and BIC of each. A summary of performance metrics over each model is shown below.

Table 6.1: Model Performance Metric Summary

Model	AIC	AICC	BIC
Model_Poi_S	45522.5894	45522.7257	45738.8369
Model_NB_S	45524.5894	45524.7351	45748.2937
Model_ZPoi_S	43645.0023	43645.1888	43898.5338
Model_ZNB_S	43765.6413	43765.8388	44026.6296

Based on the results above, this assessment concludes that Model_ZPoi_S is the superior model. Its performance metrics were among the most favorable, and maintained a ratio of deviance to degree of freedom close to one.

7 Model Deployment Code

Please see Appendix A for the final deployment code.

8 Conclusion

Five predictive models were built, each specified using automated variable selection techniques, and we predicted the number of wine orders based on a withheld test set of data using a SAS routine. To assess the performance of each model we compared the AIC, AICC and BIC values. The Zero Inflated Poisson model produced the most favorable performance metrics, with the lowest AIC, AICC and BIC metrics of all the models tested. In addition, the Zero Inflated Poisson model maintained a ratio of deviance to degree of freedom close to one.

Appendix A: Final Deployment Code

```
* Set variables / global macros;

%LET key = INDEX;
%LET response = TARGET;
%LET varname = name;

%LET data = wine;
%LET contents = &data._contents;



* Load the dataset;

libname mydata '/sscc/home/d/dgb2583/411/' access = readonly;

DATA &data.;
    *SET mydata.wine;
    SET mydata.wine_test;
RUN; QUIT;

PROC CONTENTS DATA = &data. OUT = &contents.;
RUN; QUIT;

*PROC PRINT DATA = &contents. (OBS=20);
*RUN; QUIT;

PROC MEANS DATA = &data. MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;



* Data rename;

%MACRO rename_num(varname);
    DATA &data_def.;
        SET &data_def. (RENAME = (&varname. = N_&varname.));
    RUN; QUIT;
%MEND;

TITLE1 '';
TITLE2 '';

DATA &data._name;
    SET &data.
        (RENAME = (TotalSulfurDioxide   = TotSulfDiox
                   FreeSulfurDioxide    = FreSulfDiox
                   ResidualSugar        = ResSugar
                   VolatileAcidity      = VolAcid));
RUN; QUIT;

PROC CONTENTS DATA = &data._name OUT = &contents._name;
```

```
RUN; QUIT;

DATA &contents._name;
    SET &contents._name;
        IF name = "&key." then DELETE;
        IF name = "&response." then DELETE;
RUN; QUIT;

%LET data_def = &data._name;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._name NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%rename_num('||name||')');
    END;
RUN; QUIT;

PROC MEANS DATA = &data._name MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;

PROC CONTENTS DATA = &data._name OUT = &contents._name;
RUN; QUIT;



* Data preparation;

%MACRO means(varname);
    PROC means DATA = &data_def. noprint;
    OUTPUT OUT = &varname. (DROP = _freq_ _type_)
        nmiss(&varname.)    = &varname._nmiss
        n(&varname.)        = &varname._n
        mean(&varname.)     = &varname._mean
        median(&varname.)   = &varname._median
        mode(&varname.)     = &varname._mode
        std(&varname.)      = &varname._std
        skew(&varname.)     = &varname._skew
        P1(&varname.)       = &varname._P1
        P5(&varname.)       = &varname._P5
        P10(&varname.)      = &varname._P10
        P25(&varname.)      = &varname._P25
        P50(&varname.)      = &varname._P50
        P75(&varname.)      = &varname._P75
        P90(&varname.)      = &varname._P90
        P95(&varname.)      = &varname._P95
        P99(&varname.)      = &varname._P99
        min(&varname.)      = &varname._min
        max(&varname.)      = &varname._max
        qrange(&varname.)   = &varname._qrange
        ;
    RUN; QUIT;
%MEND;
```

```
%MACRO transpose(varname);
    PROC transpose DATA = &varname. OUT = &varname._t;
        var _numeric_;
    RUN; QUIT;
%MEND;

%MACRO symputx_num(varname);
    DATA _null_;
        SET &varname._t;
            CALL symputx(_name_, strip(col1), 'g');
    RUN; QUIT;
%MEND;

%MACRO outlier(varname);
    DATA &data_def.;
        SET &data_def.;
            *IF (&varname. < &&&varname._P10) OR (&varname. > &&&varname._P90) THEN
            *   &varname._OF = 1.0; *ELSE &varname._OF = 0.0;

            *IF (&varname. < &&&varname._P5) OR (&varname. > &&&varname._P95) THEN
            *   &varname._OF = 1.0; *ELSE &varname._OF = 0.0;

            IF (&varname. < &&&varname._P1) OR (&varname. > &&&varname._P99) THEN
                &varname._OF = 1.0; ELSE &varname._OF = 0.0;
    RUN; QUIT;
%MEND;

%MACRO trim(varname);
    DATA &data_def.;
        SET &data_def.;
            &varname._T90 = &varname.;
            *&varname._T90 = max(min(&varname.,&&&varname._P90),&&&varname._P10);
            IF (&varname._T90 < &&&varname._P10) OR (&varname._T90 > &&&varname._P90) THEN
                &varname._T90 = '.';

            &varname._T95 = &varname.;
            *&varname._T95 = max(min(&varname.,&&&varname._P95),&&&varname._P5);
            IF (&varname._T95 < &&&varname._P5) OR (&varname._T95 > &&&varname._P95) THEN
                &varname._T95 = '.';

            &varname._T99 = &varname.;
            *&varname._T99 = max(min(&varname.,&&&varname._P99),&&&varname._P1);
            IF (&varname._T99 < &&&varname._P1) OR (&varname._T99 > &&&varname._P99) THEN
                &varname._T99 = '.';
    RUN; QUIT;
%MEND;

%MACRO missing(varname);
    DATA &data_def.;
        SET &data_def.;
            IF missing(&varname.) THEN
                &varname._MF = 1.0; ELSE &varname._MF = 0.0;
    RUN; QUIT;
```

```
%MEND;

%MACRO impute(varname);
    DATA &data_def.;
        SET &data_def.;
            *&varname._IMU = &varname.;
            *IF missing(&varname._IMU) THEN
            *   &varname._IMU = &&&varname._mean;

            *&varname._IMO = &varname.;
            *IF missing(&varname._IMO) THEN
            *   &varname._IMO = &&&varname._mode;

            &varname._IME = &varname.;
            IF missing(&varname._IME) THEN
                &varname._IME = &&&varname._median;
    RUN; QUIT;
%MEND;

%MACRO transform(varname);
    DATA &data_def.;
        SET &data_def.;
            &varname._LN = sign(&varname.) * log(abs(&varname.)+1);
            *&varname._SQ = (&varname.*&varname.);
            *&varname._RT = sqrt(&varname.);
    RUN; QUIT;
%MEND;

%MACRO drop(varname);
    DATA &data_def.;
        SET &data_def.;
            DROP &varname.;
    RUN; QUIT;
%MEND;

TITLE1 '';
TITLE2 '';

* Adhoc changes;

DATA &data._clean;
    SET &data._name;
RUN; QUIT;

* Create new dataset of flags for continuous variables;

DATA &data._flag;
    SET &data._clean;
RUN; QUIT;

PROC CONTENTS DATA = &data._flag OUT = &contents._flag;
RUN; QUIT;
```

```
DATA &contents._flag;
    SET &contents._flag;
        IF name = "&key." then DELETE;
        IF name = "&response." then DELETE;
RUN; QUIT;

%LET data_def = &data._flag;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._flag NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%means('||name||')');
                CALL EXECUTE('%transpose('||name||')');
                CALL EXECUTE('%symputx_num('||name||')');
    END;
RUN; QUIT;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._flag NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%missing('||name||')');
                CALL EXECUTE('%outlier('||name||')');
    END;
RUN; QUIT;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._name NOBS = NUM;
            CALL EXECUTE('%drop('||name||')');
    END;
RUN; QUIT;

DATA &data._flag;
    MERGE &data._flag &data.(KEEP = &key.);
RUN; QUIT;

PROC MEANS DATA = &data._flag MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;

PROC CONTENTS DATA = &data._flag OUT = &contents._flag;
RUN; QUIT;

* Create dummy variables;

DATA &data._dum;
    SET &data._clean;
RUN; QUIT;

PROC CONTENTS DATA = &data._dum OUT = &contents._dum;
RUN; QUIT;
```

```
DATA &contents._dum;
    SET &contents._dum;
        IF name = "&key." then DELETE;
        IF name = "&response." then DELETE;
RUN; QUIT;

DATA &data._dum;
    SET &data._dum;
        N_STARS_0       =   (0.0 <= N_STARS < 0.5);
        N_STARS_1       =   (0.5 <= N_STARS < 1.5);
        N_STARS_2       =   (1.5 <= N_STARS < 2.5);
        N_STARS_3       =   (2.5 <= N_STARS < 3.5);
        N_STARS_4       =   (3.5 <= N_STARS <= 4.0);
        N_STARS_GTE2    =   (1.5 <= N_STARS <= 4.0);
        N_STARS_GTE3    =   (2.5 <= N_STARS <= 4.0);

        N_LabelAppeal_1     =   (-2.0 <= N_LabelAppeal < -1.5);
        N_LabelAppeal_2     =   (-1.5 <= N_LabelAppeal < -0.5);
        N_LabelAppeal_3     =   (-0.5 <= N_LabelAppeal < 0.5);
        N_LabelAppeal_4     =   (0.5 <= N_LabelAppeal < 1.5);
        N_LabelAppeal_5     =   (1.5 <= N_LabelAppeal <= 2.0);
        N_LabelAppeal_GTE3  =   (-0.5 <= N_LabelAppeal <= 2.0);
        N_LabelAppeal_GTE4  =   (0.5 <= N_LabelAppeal <= 2.0);
RUN; QUIT;

%LET data_def = &data._dum;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._name NOBS = NUM;
            CALL EXECUTE('%drop('||name||')');
    END;
RUN; QUIT;

DATA &data._dum;
    MERGE &data._dum &data.(KEEP = &key.);
RUN; QUIT;

PROC MEANS DATA = &data._dum MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;

PROC CONTENTS DATA = &data._dum OUT = &contents._dum;
RUN; QUIT;

* Add trimmed series to original dataset;

DATA &data._trim;
    SET &data._clean;
RUN; QUIT;

PROC CONTENTS DATA = &data._trim OUT = &contents._trim;
RUN; QUIT;
```

```
DATA &contents._trim;
    SET &contents._trim;
        IF name = "&key." then DELETE;
        IF name = "&response." then DELETE;
RUN; QUIT;

%LET data_def = &data._trim;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._trim NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%means('||name||')');
                CALL EXECUTE('%transpose('||name||')');
                CALL EXECUTE('%symputx_num('||name||')');
    END;
RUN; QUIT;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._trim NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%trim('||name||')');
    END;
RUN; QUIT;

* Impute all continuous series in original dataset;

DATA &data._imp;
    SET &data._trim;
RUN; QUIT;

PROC CONTENTS DATA = &data._imp OUT = &contents._imp;
RUN; QUIT;

DATA &contents._imp;
    SET &contents._imp;
        IF name = "&key." then DELETE;
        IF name = "&response." then DELETE;
RUN; QUIT;

%LET data_def = &data._imp;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._imp NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%means('||name||')');
                CALL EXECUTE('%transpose('||name||')');
                CALL EXECUTE('%symputx_num('||name||')');
    END;
RUN; QUIT;
```

```
DATA _null_;
    DO i = 1 to NUM;
        SET &contents._imp NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%impute('||name||')');
    END;
RUN; QUIT;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._imp NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%drop('||name||')');
    END;
RUN; QUIT;

* Transform all continuous series in original dataset;

DATA &data._trans;
    SET &data._imp;
RUN; QUIT;

PROC CONTENTS DATA = &data._trans OUT = &contents._trans;
RUN; QUIT;

DATA &contents._trans;
    SET &contents._trans;
        IF name = "&key." then DELETE;
        IF name = "&response." then DELETE;
RUN; QUIT;

%LET data_def = &data._trans;

DATA _null_;
    DO i = 1 to NUM;
        SET &contents._trans NOBS = NUM;
            WHERE type = 1;
                CALL EXECUTE('%transform('||name||')');
    END;
RUN; QUIT;

PROC MEANS DATA = &data._trans MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;

PROC CONTENTS DATA = &data._trans OUT = &contents._trans;
RUN; QUIT;

* Merge Datasets;

DATA &data._merged;
    MERGE &data._flag &data._dum &data._trans;
    *DROP where TYPE _CHARACTER_;
    &response._FLAG = (&response. > 0);
```

```
    &response._AMT  = (&response. - 1);
    IF &response._FLAG = 0 then &response._AMT = .;
RUN; QUIT;

PROC CONTENTS DATA = &data._merged OUT = &contents._merged;
RUN; QUIT;

PROC MEANS DATA = &data._merged MIN P5 P50 P90 P95 P99 MAX MEAN STDDEV NMISS N;
RUN; QUIT;



* Testing;

DATA &data._scored (KEEP = INDEX P_:);
    SET &data._merged;

    p_target_reg = 1.41942 +
    (N_Alcohol_OF   *   0.09899)   +
    (N_STARS_1   *   1.66956)   +
    (N_STARS_GTE2   *   2.38672)   +
    (N_LabelAppeal_5   *   0.13799)   +
    (N_AcidIndex_IME   *   -0.12299)   +
    (N_AcidIndex_T99_IME   *   -0.58704)   +
    (N_Alcohol_IME   *   0.00733)   +
    (N_Alcohol_T90_IME   *   0.1616)   +
    (N_Chlorides_IME   *   -0.12143)   +
    (N_Density_T90_IME   *   -2.22961)   +
    (N_FreSulfDiox_T99_IME   *   0.00030222)   +
    (N_LabelAppeal_IME   *   0.45554)   +
    (N_STARS_IME   *   1.14507)   +
    (N_Sulphates_IME   *   -0.03011)   +
    (N_TotSulfDiox_IME   *   0.00021852)   +
    (N_VolAcid_IME   *   -0.09513)   +
    (N_pH_T90_IME   *   -0.13897)   +
    (N_AcidIndex_T99_IME_LN   *   4.60162)   +
    (N_Alcohol_T90_IME_LN   *   -1.51042)   +
    (N_CitricAcid_T90_IME_LN   *   0.11136)   +
    (N_STARS_IME_LN   *   -2.04254);

    p_target_reg = ROUND(p_target_reg, 1);


    p_target_poi = -0.2487 +
    (N_STARS_GTE2   *   1.0737)   +
    ((N_AcidIndex_IME  in (4)) *   1.2052)   +
    ((N_AcidIndex_IME  in (5)) *   1.0712)   +
    ((N_AcidIndex_IME  in (6)) *   1.1071)   +
    ((N_AcidIndex_IME  in (7)) *   1.0711)   +
    ((N_AcidIndex_IME  in (8)) *   1.0392)   +
    ((N_AcidIndex_IME  in (9)) *   0.9271)   +
    ((N_AcidIndex_IME  in (10)) *   0.7725)   +
    ((N_AcidIndex_IME  in (11)) *   0.4052)   +
```

```
    ((N_AcidIndex_IME  in (12)) *   0.3936)   +
    ((N_AcidIndex_IME  in (13)) *   0.5515)   +
    ((N_AcidIndex_IME  in (14)) *   0.4552)   +
    ((N_AcidIndex_IME  in (15)) *   0.8889)   +
    ((N_AcidIndex_IME  in (16)) *   0.2454)   +
    ((N_AcidIndex_IME  in (17)) *   0)   +
    (N_Alcohol_T90_IME   *   0.0108)   +
    (N_Chlorides_IME   *   -0.0383)   +
    ((N_LabelAppeal_IME  in (-2)) *   -0.6994)   +
    ((N_LabelAppeal_IME  in (-1)) *   -0.4574)   +
    ((N_LabelAppeal_IME  in (0)) *   -0.2679)   +
    ((N_LabelAppeal_IME  in (1)) *   -0.1348)   +
    ((N_LabelAppeal_IME  in (2)) *   0)   +
    ((N_STARS_IME  in (1)) *   0.5163)   +
    ((N_STARS_IME  in (2)) *   -0.2394)   +
    ((N_STARS_IME  in (3)) *   -0.1221)   +
    ((N_STARS_IME  in (4)) *   0)   +
    (N_TotSulfDiox_IME   *   0.0001)   +
    (N_VolAcid_IME   *   -0.0291)   +
    (N_pH_T90_IME   *   -0.043)   +
    (N_CitricAcid_T90_IME   *   0.0261)   +
    (N_FreSulfDiox_IME_LN   *   0.0034);

    p_target_poi = EXP(p_target_poi);
    p_target_poi = ROUND(p_target_poi, 1);


    p_target_nb = -0.2487 +
    (N_STARS_GTE2   *   1.0737)   +
    ((N_AcidIndex_IME  in (4)) *   1.2052)   +
    ((N_AcidIndex_IME  in (5)) *   1.0712)   +
    ((N_AcidIndex_IME  in (6)) *   1.1071)   +
    ((N_AcidIndex_IME  in (7)) *   1.0711)   +
    ((N_AcidIndex_IME  in (8)) *   1.0392)   +
    ((N_AcidIndex_IME  in (9)) *   0.9271)   +
    ((N_AcidIndex_IME  in (10)) *   0.7725)   +
    ((N_AcidIndex_IME  in (11)) *   0.4052)   +
    ((N_AcidIndex_IME  in (12)) *   0.3936)   +
    ((N_AcidIndex_IME  in (13)) *   0.5515)   +
    ((N_AcidIndex_IME  in (14)) *   0.4552)   +
    ((N_AcidIndex_IME  in (15)) *   0.8889)   +
    ((N_AcidIndex_IME  in (16)) *   0.2454)   +
    ((N_AcidIndex_IME  in (17)) *   0)   +
    (N_Alcohol_T90_IME   *   0.0108)   +
    (N_Chlorides_IME   *   -0.0383)   +
    ((N_LabelAppeal_IME  in (-2)) *   -0.6994)   +
    ((N_LabelAppeal_IME  in (-1)) *   -0.4574)   +
    ((N_LabelAppeal_IME  in (0)) *   -0.2679)   +
    ((N_LabelAppeal_IME  in (1)) *   -0.1348)   +
    ((N_LabelAppeal_IME  in (2)) *   0)   +
    ((N_STARS_IME  in (1)) *   0.5163)   +
    ((N_STARS_IME  in (2)) *   -0.2394)   +
    ((N_STARS_IME  in (3)) *   -0.1221)   +
```

```
((N_STARS_IME  in (4)) *   0)   +
(N_TotSulfDiox_IME   *   0.0001)   +
(N_VolAcid_IME   *   -0.0291)   +
(N_pH_T90_IME   *   -0.043)   +
(N_CitricAcid_T90_IME   *   0.0261)   +
(N_FreSulfDiox_IME_LN   *   0.0034);

p_target_nb = EXP(p_target_nb);
p_target_nb = ROUND(p_target_nb, 1);


p_target_zip_all = 1.4688 +
(N_STARS_GTE2   *   0.4752)   +
(N_AcidIndex_T95_IME   *   -0.0249)   +
(N_Alcohol_IME   *   0.0041)   +
(N_Alcohol_T90_IME   *   0.0095)   +
((N_LabelAppeal_IME  in (-2)) *   -1.0236)   +
((N_LabelAppeal_IME  in (-1)) *   -0.6331)   +
((N_LabelAppeal_IME  in (0)) *   -0.3518)   +
((N_LabelAppeal_IME  in (1)) *   -0.1637)   +
((N_LabelAppeal_IME  in (2)) *   0)   +
((N_STARS_IME  in (1)) *   0.1579)   +
((N_STARS_IME  in (2)) *   -0.1822)   +
((N_STARS_IME  in (3)) *   -0.099)   +
((N_STARS_IME  in (4)) *   0)   +
(N_VolAcid_IME   *   -0.018);

p_target_zip_zero = -4.394 +
((N_STARS_IME  in (1)) *   18.2319)   +
((N_STARS_IME  in (2)) *   18.5193)   +
((N_STARS_IME  in (3)) *   -7.4969)   +
((N_STARS_IME  in (4)) *   0)   +
((N_LabelAppeal_IME  in (-2)) *   -2.8423)   +
((N_LabelAppeal_IME  in (-1)) *   -1.4548)   +
((N_LabelAppeal_IME  in (0)) *   -0.8331)   +
((N_LabelAppeal_IME  in (1)) *   -0.4174)   +
((N_LabelAppeal_IME  in (2)) *   0)   +
((N_AcidIndex_IME  in (4)) *   -13.7932)   +
((N_AcidIndex_IME  in (5)) *   -15.3206)   +
((N_AcidIndex_IME  in (6)) *   -15.0953)   +
((N_AcidIndex_IME  in (7)) *   -15.032)   +
((N_AcidIndex_IME  in (8)) *   -14.703)   +
((N_AcidIndex_IME  in (9)) *   -14.0103)   +
((N_AcidIndex_IME  in (10)) *   -13.331)   +
((N_AcidIndex_IME  in (11)) *   -12.3995)   +
((N_AcidIndex_IME  in (12)) *   -12.072)   +
((N_AcidIndex_IME  in (13)) *   -12.4809)   +
((N_AcidIndex_IME  in (14)) *   -12.1375)   +
((N_AcidIndex_IME  in (15)) *   -13.3732)   +
((N_AcidIndex_IME  in (16)) *   0.3679)   +
((N_AcidIndex_IME  in (17)) *   0);

p_target_zip_all = EXP(p_target_zip_all);
```

```
p_target_zip_zero = EXP(p_target_zip_zero) / (1 + EXP(p_target_zip_zero));
p_target_zip = p_target_zip_all * (1 - p_target_zip_zero);
p_target_zip = ROUND(p_target_zip, 1);
DROP p_target_zip_all p_target_zip_zero;


p_target_zinb_all = 1.4629 +
(N_STARS_GTE2   *   0.4813)   +
(N_AcidIndex_T95_IME   *   -0.0249)   +
(N_Alcohol_IME   *   0.0042)   +
(N_Alcohol_T90_IME   *   0.0094)   +
((N_LabelAppeal_IME  in (-2)) *   -1.0086)   +
((N_LabelAppeal_IME  in (-1)) *   -0.6308)   +
((N_LabelAppeal_IME  in (0)) *   -0.3507)   +
((N_LabelAppeal_IME  in (1)) *   -0.1633)   +
((N_LabelAppeal_IME  in (2)) *   0)   +
((N_STARS_IME  in (1)) *   0.1625)   +
((N_STARS_IME  in (2)) *   -0.1828)   +
((N_STARS_IME  in (3)) *   -0.0994)   +
((N_STARS_IME  in (4)) *   0)   +
(N_VolAcid_IME   *   -0.0185);

p_target_zinb_zero = -6.8329 +
((N_STARS_IME  in (1)) *   3.2252)   +
((N_STARS_IME  in (2)) *   3.521)   +
((N_STARS_IME  in (3)) *   -0.8452)   +
((N_STARS_IME  in (4)) *   0)   +
((N_LabelAppeal_IME  in (-2)) *   -2.3416)   +
((N_LabelAppeal_IME  in (-1)) *   -1.4016)   +
((N_LabelAppeal_IME  in (0)) *   -0.7878)   +
((N_LabelAppeal_IME  in (1)) *   -0.3817)   +
((N_LabelAppeal_IME  in (2)) *   0)   +
((N_AcidIndex_IME  in (4)) *   3.8158)   +
((N_AcidIndex_IME  in (5)) *   2.2842)   +
((N_AcidIndex_IME  in (6)) *   2.3033)   +
((N_AcidIndex_IME  in (7)) *   2.3539)   +
((N_AcidIndex_IME  in (8)) *   2.6837)   +
((N_AcidIndex_IME  in (9)) *   3.3782)   +
((N_AcidIndex_IME  in (10)) *   4.0638)   +
((N_AcidIndex_IME  in (11)) *   4.989)   +
((N_AcidIndex_IME  in (12)) *   5.2714)   +
((N_AcidIndex_IME  in (13)) *   4.9162)   +
((N_AcidIndex_IME  in (14)) *   5.2288)   +
((N_AcidIndex_IME  in (15)) *   4.2238)   +
((N_AcidIndex_IME  in (16)) *   4.3348)   +
((N_AcidIndex_IME  in (17)) *   0);

p_target_zinb_all = EXP(p_target_zinb_all);
p_target_zinb_zero = EXP(p_target_zinb_zero) / (1 + EXP(p_target_zinb_zero));
p_target_zinb = p_target_zinb_all * (1 - p_target_zinb_zero);
p_target_zinb = ROUND(p_target_zinb, 1);
DROP p_target_zinb_all p_target_zinb_zero;
```

```
RUN; QUIT;

PROC PRINT DATA = &data._scored (OBS = 20);
RUN; QUIT;

PROC EXPORT DATA = &data._scored
    OUTFILE = '/sscc/home/d/dgb2583/411/out.csv'
    DBMS = csv
    REPLACE;
RUN; QUIT;

DATA '/sscc/home/d/dgb2583/411/out';
    SET &data._scored;
RUN; QUIT;
```

References

Digital Research, Institute for, and Education. 2016. "SAS Data Analysis Examples Zero-Inflated Poisson Regression." http://www.ats.ucla.edu/stat/sas/dae/zipreg.htm.

N. Jackowetz, E. Li & R. Mira de Orduna. 2011. "Sulfur Dioxide Content of Wines: The Role of Winemaking and Carbonyl Compounds." Cornell University. https://grapesandwine.cals.cornell.edu/sites/grapesandwine.cals.cornell.edu/files/shared/documents/Research-Focus-2011-3.pdf.

Neeley, E. 2015. "Volatile Acidity." http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity.

Peynaud, E. 1987. "The Taste of Wine: The Art and Science of Wine Appreciation." The Wine Appreciation Guild.

Robinson, J. 2006. "The Oxford Companion to Wine." Oxford University Press.