Ryan Sandfort & Carch Bott

The Storm Kings

4/17/2020

STAT 5100

Predicting Red Wine Quality Using Various Indicators

Introduction:

The global wine market was valued at over 302 billion dollars in 2017, and is expected to be worth over 400 billion by 2023. Wine comes in all different types and colors, such as whites, reds, roses, and sparkling. These wines all have very different tastes, tannin concentrations, and acidities. With all of the extreme variety within the world of wine, it can be very difficult to try to predict the quality of wine from just the type, year or fruit quality. This makes it extremely difficult for vineyards and wine connoisseurs alike to purchase new wines with confidence of them being of the quality they enjoy. The ability to predict the quality of wine using various indicators would allow for both people with personal and professional interest in the world of wine to produce or pick out high quality wines with less difficulty.

Data:

This dataset had 1600 observations, where red wine was analyzed using physicochemical tests and put on a relative scale for each variable. These variables include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The response variable, based on sensory data, was quality which ranged from 0-10. When these various predictor variables were placed against quality to determine various trends, there were a couple concerning points. Quality is only measured in whole numbers between 0-10 and is therefore a discrete set of data. This makes it difficult as there are steps that jump up instead of having a very apparent trend in some of this data. Another disconcerting factor is that several of these variables seem to have very influential points or very little correlation with quality, such as density (Figure 2). However, there are a couple variables such as alcohol that follows a very clear trend positively associated with quality (Figure 1). This indicates that after variable selection, there should still be a few variables that are strong in predicting the quality of wine. The heteroscedasticity of the data is also explained by the fact that the quality variable jumps from whole number to whole number. This would make it difficult for there to be even residuals as there are only a select number of points that observations can land on the Y-axis. 20% of the observations were randomly withheld and put into a test data set and will be used to test the accuracy of the final model.

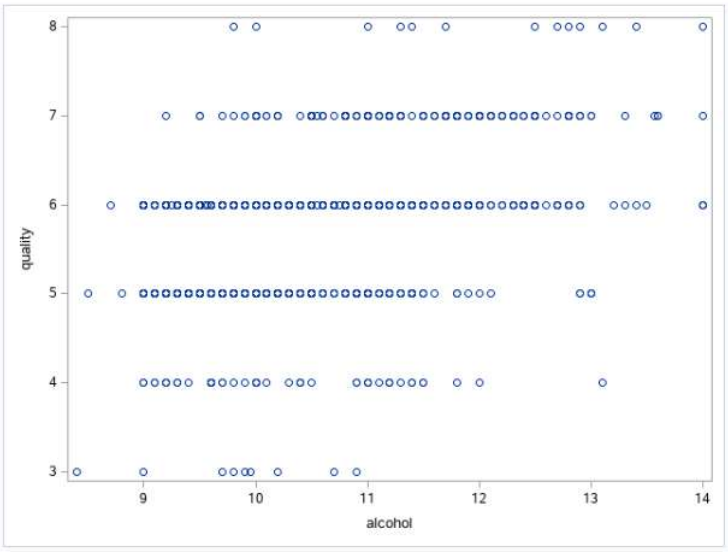


Figure 1: Scatterplot of Alcohol vs Quality Indicating Clear Positive Correlation

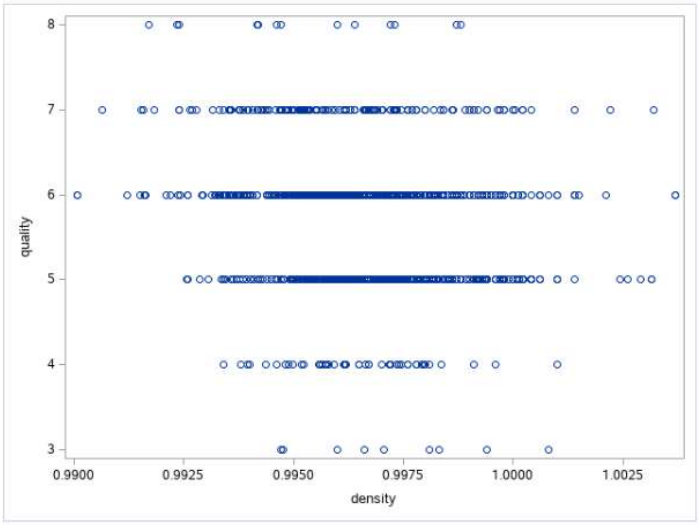
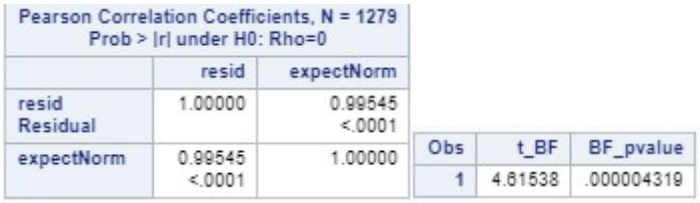


Figure 2: Scatterplot of Density vs Quality Indicating no Correlation

Assumptions:

The assumptions of linear regression have been partially met by the original data in this set. The data is proven to be normally distributed, with both an even looking histogram and a normality-indicating Pearson Correlation Coefficient. The issue is that the residuals seem to be heteroscedastic. The p-value for the Brown-Forsythe test on the untransformed data is under .05, which indicates that the data's residuals are not homoscedastic. This was expected, as the response variable is discrete, which makes it very difficult for homoscedasticity to be met. This means that linear regression would be appropriate for this model even with a very low p-value for the Brown-Forsythe Test**.** After going through remedial measures and altering the data with a log transformation, the Brown-Forsythe Test's p-value increased up to barely over .05, which is our set threshold to accept the null hypothesis that the data is homoscedastic. This indicates that the log transformation helped to minimize the heteroscedasticity caused by the discrete response variable. The log transformation also maintains an even histogram and a normality-indicating Pearson Correlation Coefficient. This indicates that the data after being transformed by the log transformation is much better suited for linear regression.



Pearson Correlation Coefficients, N = 1279
Prob > |r| under H0: Rho=0

	resid	expectNorm
resid Residual	1.00000	0.99545 <.0001
expectNorm	0.99545 <.0001	1.00000

Obs	t_BF	BF_pvalue
1	4.61538	.000004319

Figure 3: Pearson Correlation Coefficients and Brown-Forsythe Test for Untransformed Data

Pearson Correlation Coefficients, N = 1279
Prob > |r| under H0: Rho=0

	resid	expectNorm
resid Residual	1.00000	0.98305 <.0001
expectNorm	0.98305 <.0001	1.00000

Obs	t_BF	BF_pvalue
1	1.90021	0.057630

Figure 4: Pearson Correlation Coefficients and Brown-Forsythe Test for Log Transformation of Quality

Remedial Measures:

Because the study was performed with each variable on its own relative physicochemical scale and quality defined strictly from values 0-10, there were no outliers in the dataset that warranted removing or other action to be taken. A CooksD Plot was used to find influential points in the model. The plot indicated one key influential point, point 119 (Figure 5). This observation seems to be an influential point due to its very high volatile acidity and citric acid. The volatile acidity is rather high but in normal ranges, but the citric acid is a value of 1, which is the highest value citric acid can have. Due to our lack of knowledge about whether or not this value is a mistype, we attempted some remedial measures to fix it to no avail. We attempted to use the log of citric acid, but that made the variable statistically insignificant. This led to us keeping the influential point within the model. This is the only value that A Box-Cox analysis on several of the tentative predictor variables, such as pH shown in Figure 6, indicated that a log transformation would be most efficient at making our model best suited for linear regression. Several other transformations were attempted as well, such as square root and cubed root transformations, but none were as effective at increasing the Brown-Forsythe Test's p-value as the log transformation. This indicated that the best model that met assumptions of linear regression was the log transformation of quality in our dataset. We continued onto variable selection to find a tentative model using the log transformed data.

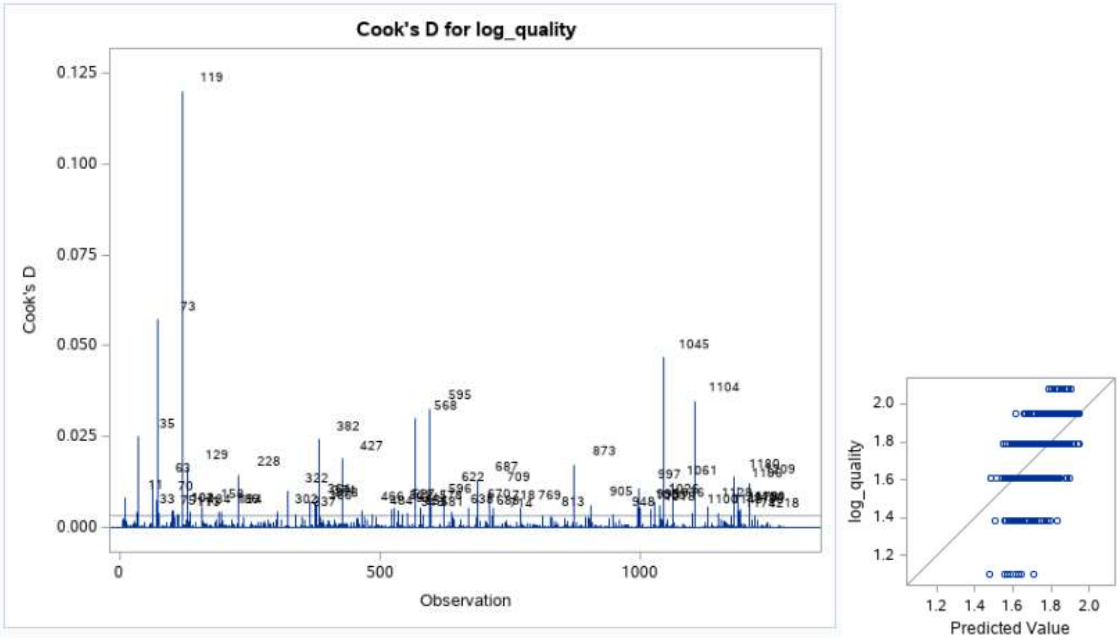


Figure 5: CooksD Plot on Log Quality After Variable Selection

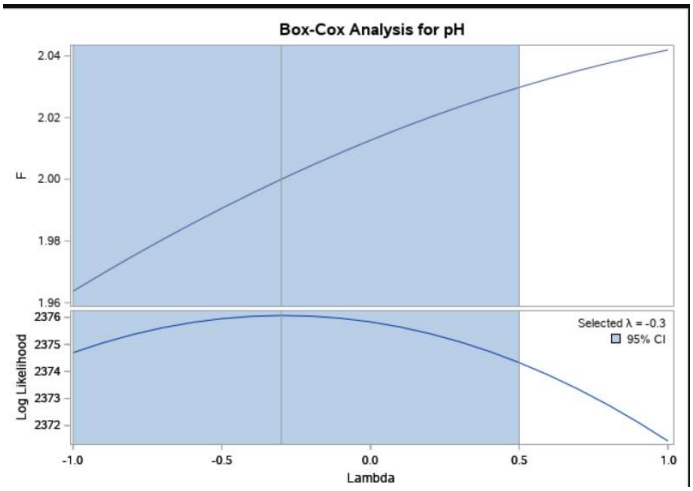


Figure 6: Box-Cox Analysis for pH on Quality

Variable Selection:

With over 10 variables in our initial dataset, variable selection was a necessary part of the model creation process. We decided to use a combination of Adjusted R-Squared, Mallows C(p), AIC and SBC techniques to determine the best model. Out of the 11 initial predictor variables, the variable selection determined that the best model was one that contained only 7. The retained variables were Alcohol, Chlorides, Citric Acid, pH, Sulfates, Total Sulfur Dioxide, and Volatile Acidity. We chose this model for a variety of reasons, primarily that it had less variables than other models with very similar Adjusted R-Squares. This model had an Adjusted R-Squared of .3207. Compared to the model with the highest Adjusted R-Squared of .3212, and a total of 10 variables, we determined that the former model would be much more effective. In the end, we lost a miniscule amount of predictive power in return for more interpretive ability in the form of fewer variables.

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
9	0.3214	0.3262	8.0387	-5392.0188	-5340.48048	alcohol chlorides citric acid density free sulfur dioxide pH sulphates total sulfur dioxide volatile acidity
8	0.3212	0.3255	7.3241	-5392.7219	-5346.33737	alcohol chlorides citric acid free sulfur dioxide pH sulphates total sulfur dioxide volatile acidity
9	0.3210	0.3258	8.8163	-5391.2341	-5339.69572	fixed acidity alcohol chlorides citric acid free sulfur dioxide pH sulphates total sulfur dioxide volatile acidity
9	0.3210	0.3257	8.8489	-5391.2012	-5339.66290	alcohol chlorides citric acid free sulfur dioxide pH residual sugar sulphates total sulfur dioxide volatile acidity
10	0.3209	0.3262	10.0111	-5390.0467	-5333.35449	alcohol chlorides citric acid density free sulfur dioxide pH residual sugar sulphates total sulfur dioxide volatile acidity
10	0.3209	0.3262	10.0111	-5390.0467	-5333.35448	fixed acidity alcohol chlorides citric acid density free sulfur dioxide pH sulphates total sulfur dioxide volatile acidity
8	0.3207	0.3250	8.3104	-5391.7277	-5345.34318	alcohol chlorides citric acid density pH sulphates total sulfur dioxide volatile acidity
7	0.3207	0.3244	7.3320	-5392.6987	-5351.46803	alcohol chlorides free sulfur dioxide pH sulphates total sulfur dioxide volatile acidity

Figure 7: Variable Selection Table (Chosen Model with Model Number 7)

Interaction Terms:

Testing for interaction terms between several of the variables in combinations that were still interpretable for the model yielded only one significant interaction between alcohol and pH level as they affect wine quality. After adding the interaction, indicated as temp1 in Figure 8, the Adjusted R-Squared was boosted up slightly, however the interaction variable was not statistically significant as shown by the p-value of .051 (Figure 8). It also made pH a statistically insignificant variable. We decided to remove the interaction term for this reason, and continue on with the initial model that was decided after variable selection.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.32457	0.63969	0.51	0.6120	0
alcohol	1	0.16809	0.05949	2.83	0.0048	340.68330
chlorides	1	-0.32603	0.08143	-4.00	<.0001	1.38387
citric acid	1	-0.04566	0.02506	-1.82	0.0687	2.12052
pH	1	0.26959	0.19025	1.42	0.1567	77.57273
sulphates	1	0.16160	0.02284	7.08	<.0001	1.32266
total sulfur dioxide	1	-0.00032878	0.00010804	-3.04	0.0024	1.08740
volatile acidity	1	-0.19978	0.02351	-8.50	<.0001	1.57913
temp1	1	-0.03464	0.01772	-1.95	0.0508	486.87571

*Figure 8: Parameter Estimates of the Model with Interaction Term Temp1 (Alcohol*pH) added*

Multicollinearity and Ridge Regression:

Collinearity diagnostics showed signs of potentially problematic collinearity between predictor variables due to some combinations of large proportion of variance values and large condition index values. To attempt to reduce and mitigate the present multicollinearity, the data will be standardized and ridge regression will be performed on the tentative reduced model. Ridge regression introduces a slight bias to our dataset so that values with less variance, meaning data that we are more confident in, is given slight preference over data with less variance. This way we make more sure data more valuable in exchange for no longer having unbiased analysis.

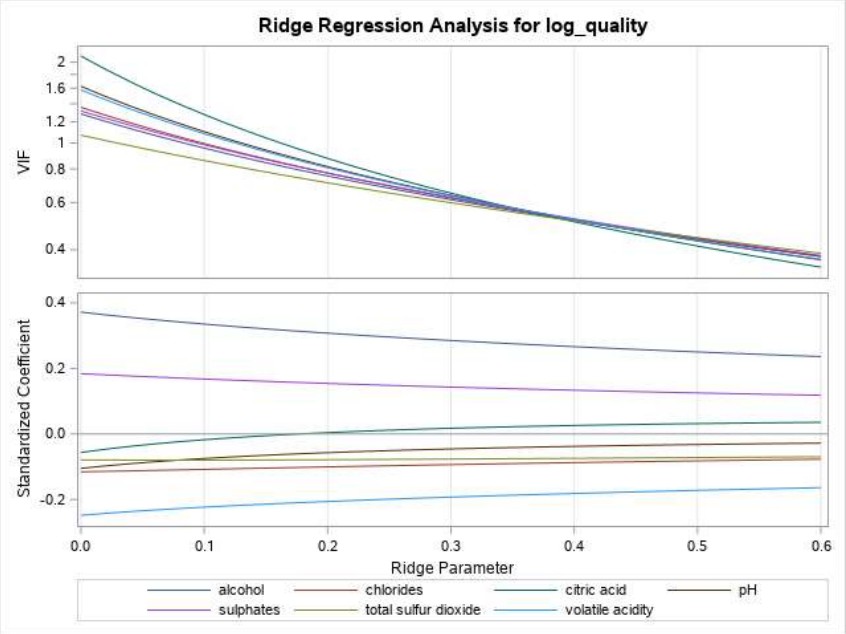


Figure 9: Ridge Regression Analysis Testing for Ridge Parameters on Log-Quality

Testing a wide range of ridge parameters (see Figure 9) gave an optimal ridge parameter value of 0.4. This value is optimal because it reduces the different variables in their variance inflation and both puts a higher value on less variance. This value was chosen over a higher ridge parameter because we want to choose the least powerful bias that will still solve the issue of large variance inflation values. Ridge regression of the tentative model of the log-transformed data gave the model:

Log_quality = 2.03544 + 0.03727*alcohol - 0.26271*chlorides + 0.01938*citric acid - 0.03547*pH + 0.115*sulphates - 0.00033*total sulfur dioxide + 0.01367*volatile acidity

This model seems to have fair predicting power on wine quality, as observed from Figure 10 comparing observed values and predicted values of wine quality having a roughly linear positive relationship. This model has an R-Squared value of 0.2683 and an Adjusted R-Squared value of 0.2677. This shows there is a lot of variability in the data, and that the sacrifice of reducing multicollinearity by introducing bias resulted in a model with less predictive power than the ordinary least squares regression model.

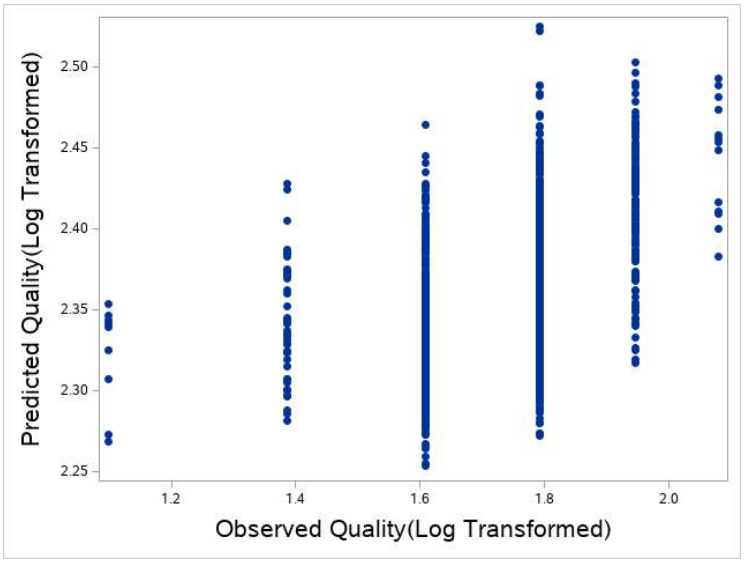


Figure 10: Observed Quality Compared to Predicted Quality

Model Assumptions For Final Model:

The final model of log-transformed quality using Ordinary Least Squares Regression is:

Log_Quality = 1.5614 + 0.052*alcohol - 0.3474*chlorides - 0.0421*citric acid

- 0.0984*pH + 0.1585*sulphates - 0.000356*total sulfur dioxide - 0.2015*volatile acidity

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.56135	0.09461	16.50	<.0001	0
alcohol	1	0.05200	0.00366	14.23	<.0001	1.28319
chlorides	1	-0.34736	0.08078	-4.30	<.0001	1.35905
citric acid	1	-0.04208	0.02502	-1.68	0.0929	2.10920
pH	1	-0.09840	0.02758	-3.57	0.0004	1.62675
sulphates	1	0.15851	0.02281	6.95	<.0001	1.31631
total sulfur dioxide	1	-0.00035590	0.00010727	-3.32	0.0009	1.06947
volatile acidity	1	-0.20151	0.02352	-8.57	<.0001	1.57690

Figure 11: Final OLS Model using Log Transformation of Quality

Due to present multicollinearity, the interpretation using this model should be regarded with hesitation and the ridge regression model listed earlier would perhaps be a better model when wanting interpretation of individual variables because of ridge regression's resistance to multicollinearity. This OLS model has a Brown-Forsythe test of constant variance p-value of 0.11634. Because of this p-values nonsignificance we conclude the data meets our assumptions of homoscedasticity. The Pearson Correlation Coefficient was highly significant with a p-value less than .0001 indicates that our assumptions of normality are met.

Accuracy:

To compare the accuracy of both our final OLS model, ridge regression model, and a model using just the intercept with no predictor variables, the 20% of randomly withheld observations will be used to test the prediction power and find the Mean Square Error of each of these models. The MSE for our final OLS model was 0.0117099 as seen in Figure 12.

The MEANS Procedure

Analysis Variable : MSE				
N	Mean	Std Dev	Minimum	Maximum
320	0.0117099	0.0233604	1.2058131E-8	0.2814834

Figure 12: MSE of Final OLS Model

The MEANS Procedure

Analysis Variable : MSE				
N	Mean	Std Dev	Minimum	Maximum
320	0.0136631	0.0295655	4.1145911E-8	0.4026622

Figure 13: MSE of Final Ridge Regression Model

The MEANS Procedure

Analysis Variable : MSE				
N	Mean	Std Dev	Minimum	Maximum
320	0.0196477	0.0319079	0.0058201	0.3805135

Figure 14: MSE of Model with Only the Intercept

Slightly larger was the MSE for our ridge regression model was 0.0136631 (Figure 13). Lastly, a model of just the intercept with no predictor variables was 0.0196477 (Figure 14). These values tell us that both models do a better job of predicting wine quality on new data than simply using a constant, and that our OLS model does a better job of predicting on new data than our ridge regression model does.

Interpretation and Conclusion:

We conclude from both models that density, fixed acidity, free sulfur dioxide, and residual sugar does not have a significant effect on the observed quality of red wine. This information can be used by both professional and hobbyist wine tasters to pay little heed to these details if they are advertised on how they impact the quality of red wine. Though our ridge regression model had slightly less predicting power than the other final model, it's coefficients on variables will be used because they reduce multicollinearity and will give more understandable interpretation. Alcohol, citric acid, sulphates, and volatile acidity all increase the quality of red wine as their coefficients increase. Looking specifically at the alcohol coefficient of 0.03727, we expect to see the quality of red wine increase by 0.03727 for every point higher in alcohol concentration while holding other variables constant. This means that generally a more alcoholic red wine will taste better than a less alcoholic counterpart, holding other aspects of the

wine constant. This is wonderful news for industrial wine producers and home wine brewers alike and can help focus their brewing efforts by increasing these attributes during the fermentation process to create a finer red wine. Likewise chlorides, pH, and total sulfur dioxide generally decreased wine quality as their numbers increased. Chlorides, for example, will generally decrease the quality of red wine by 0.26271 for every point higher the wine's chloride concentration is.

Ridge Estimates for Variable Coefficients, with ridge parameter c = 0.4

Obs	_TYPE_	_RMSE_	alcohol	chlorides	citric acid	pH	sulphates	total sulfur dioxide	volatile acidity
1	PARMS	0.12110	0.05200	-0.34736	-0.04208	-0.09840	0.15851	-0.00036	-0.20151
2	SEB	0.12110	0.00366	0.08078	0.02502	0.02758	0.02281	0.00011	0.02352
3	RIDGEVIF	.	0.51403	0.51727	0.50871	0.51786	0.52177	0.51223	0.51768
4	RIDGE	0.12288	0.03727	-0.26271	0.01938	-0.03547	0.11500	-0.00033	-0.14746
5	RIDGESEB	0.12288	0.00235	0.05057	0.01247	0.01579	0.01457	0.00008	0.01367

Figure 15: Parameter Estimates for Final Model using Ridge Regression of Log of Quality

This is also incredibly helpful information for wine brewers and tasters as excess chlorides used in filtering materials after fermentation could rather surprisingly lead to a less tasty red wine. This already massive and growing industry of wine would benefit greatly knowing what aspects of red wine improve, deteriorate, or are negligible to quality.

SAS Appendix

```
/*This line will have to be changed for it to read in the data*/
FILENAME REFFILE '/home/u42653432/winequality-red.csv';

PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=WORK.IMPORT1;
        GETNAMES=YES;
RUN;
proc contents data=work.import1; run;

/* Separate Into Training and Test Sets.
Only Fit Models to the Training Set. The variable
"Selected" separates training (0) from test (1) */
proc surveyselect data=import1 seed=420 out=dataset
    rate=0.2 outall; /* Withold 20% for validation */
run;

data train; set dataset;
if Selected = 0;
run;

data test; set dataset;
if Selected = 1;
run;
```

```
217
218  /* Rough model of Quality using all predictor variables*/
219  proc reg data=train;
220          model quality = 'fixed acidity'n alcohol chlorides 'citric acid'n density
221          'free sulfur dioxide'n pH 'residual sugar'n sulphates 'total sulfur dioxide'n
222          'volatile acidity'n / vif collin;
223  output out=out1 r=resid p=pred;
224  run;
225  %resid_num_diag(dataset=out1, datavar=resid,
226   label='Residual', predvar=pred,
227   predlabel='Predicted Value');
228  run;
229
230  /*Scatterplot of rough model*/
231  proc sgplot data=train;
232    scatter x=ID y=quality;
233  run;
234
235  /*Log-transformed rough model*/
236  data train; set train;
237  log_quality=log(quality);
238  proc reg data=train;
239          model log_quality = 'fixed acidity'n alcohol chlorides 'citric acid'n density
240          'free sulfur dioxide'n pH 'residual sugar'n sulphates 'total sulfur dioxide'n
241          'volatile acidity'n;
```

```
output out=out2 r=resid p=pred;
run;
%resid_num_diag(dataset=out2, datavar=resid,
 label='Residual', predvar=pred,
 predlabel='Predicted Value');
run;

/*Cube root-transformed rough model*/
data train; set train;
cubeRT_quality=(quality)**(1/3);
proc reg data=train;
        model cubeRT_quality = 'fixed acidity'n alcohol chlorides 'citric acid'n density
        'free sulfur dioxide'n pH 'residual sugar'n sulphates 'total sulfur dioxide'n
        'volatile acidity'n;
output out=out3 r=resid p=pred;
run;
%resid_num_diag(dataset=out3, datavar=resid,
 label='Residual', predvar=pred,
 predlabel='Predicted Value');
run;

/*Looking at Criterion for Variable Selection - log*/
proc reg data=train;
        model log_quality = 'fixed acidity'n alcohol chlorides 'citric acid'n density
        'free sulfur dioxide'n pH 'residual sugar'n sulphates 'total sulfur dioxide'n
```

```
267             'volatile acidity'n
268      / selection=AdjRSq Cp AIC SBC;
269      title1 'Compare Selection';
270     run;
271
272     /*Model of Selected Variables w/ log transformation*/
273     proc reg data=train;
274             model log_quality =  alcohol chlorides 'citric acid'n pH sulphates
275             'total sulfur dioxide'n 'volatile acidity'n / vif collin;
276     output out=out4 r=resid p=pred;
277     store regModel;
278     run;
279     %resid_num_diag(dataset=out4, datavar=resid,
280      label='Residual', predvar=pred,
281      predlabel='Predicted Value');
282     run;
283
284     data train; set train;
285     temp1 = alcohol*pH;
286     run;
287     proc reg data=train;
288      model log_quality = temp1 alcohol pH / vif; run;
289
290      proc reg data=train;
291      model log_quality = temp1 chlorides 'citric acid'n sulphates
```

```
292             'total sulfur dioxide'n 'volatile acidity'n
293     / vif;   title1 'Check for interaction';
294     run;
295
296     /*Looking at Ridge Regression*/
297    proc reg data=train ridge=0 to 0.6 by .02
298     outvif outest=ridgests
299     plots(only)=ridge(VIFaxis=log);
300     model log_quality =  alcohol chlorides 'citric acid'n pH sulphates
301             'total sulfur dioxide'n 'volatile acidity'n / vif;
302     title1 'Ridge Regression';
303     title2 '(find adequate ridge parameter)';
304    run;
305
306    proc reg data=train outest=ridgenew outseb ridge=0.4
307     outvif noprint;
308     model log_quality =  alcohol chlorides 'citric acid'n pH sulphates
309             'total sulfur dioxide'n 'volatile acidity'n;
310     title1 'Log Quality Ridge Regression (c=.4)';
311    run;
312    proc print data=ridgenew;
313     var _type_ _rmse_ alcohol chlorides 'citric acid'n pH sulphates
314             'total sulfur dioxide'n 'volatile acidity'n;
315     title1 'Ridge Estimates for Variable Coefficients,';
316     title2 'with ridge parameter c = 0.4';
```

```
317 run;
318
319 proc means data=train mean;
320  var log_quality alcohol chlorides 'citric acid'n pH sulphates
321          'total sulfur dioxide'n 'volatile acidity'n;
322  title1 'Summary Statistics';
323 run;
324 /*Getting the interecepts from the above coefficients*/
325 data temp;
326 b0 = 1.71547 + 10.3994*0.03727 - 0.08848*0.26271 + 0.26901*0.01938
327 - 3.31249*0.03547 + 0.65983*0.115 - 46.1293980*0.00033 + 0.5303*0.01367;
328 proc print data=temp;
329 var b0;
330 title1 'Ridge Regression Intercept';
331 run;
332
333 /*Ridge Regression Model on Log-Transformed Data */
334 data train; set train;
335 pred_log_quality = 2.03544 + 0.03727*alcohol - 0.26271*chlorides +
336 0.01938*'citric acid'n - 0.03547*pH + 0.115*sulphates -
337 0.00033*'total sulfur dioxide'n + 0.01367*'volatile acidity'n;
338 /*The below plot should show a sort of y=x graph, which is seems
339 to do a pretty good job at*/
340  proc sgplot data=train;
341  scatter x=log_quality y=pred_log_quality / markerattrs=(symbol=CIRCLEFILLED);
```

```
342     xaxis label='Observed Quality(Log Transformed)' labelattrs=(size=15pt);
343     yaxis label='Predicted Quality(Log Transformed)' labelattrs=(size=15pt);
344     title1;
345    run;
346
347    proc reg data=train;
348              model log_quality =  pred_log_quality / vif collin;
349    output out=out5 r=resid p=pred;
350    store regModel2;
351    run;
352
353    /* Calculate MSPR */
354    data test; set test;
355    log_quality = log(quality);
356    pred_log_quality = 2.03544 + 0.03727*alcohol - 0.26271*chlorides +
357    0.01938*'citric acid'n - 0.03547*pH + 0.115*sulphates -
358    0.00033*'total sulfur dioxide'n + 0.01367*'volatile acidity'n;
359    proc plm restore=regModel;
360     score data=test out=newTest predicted;
361     run;
362
363     proc reg data=train;
364              model log_quality =;
365              store intercept;
366              proc plm restore=intercept;
```

```
 score data=test out=newTest1 predicted;
 run;

 data newTest1; set newTest1;
MSE = (log_quality - Predicted)**2;
run;
proc means data = newTest1;
var MSE;
run;

data newTest; set newTest;
MSE = (log_quality - Predicted)**2;
run;
proc means data = newTest;
var MSE;
run;

/*Ridge Regression MSPR */
proc plm restore=regModel2;
 score data=test out=newTest2 predicted;
 run;
data newTest2; set newTest2;
MSE = (log_quality - Predicted)**2;
run;
proc means data = newTest2;
```

```
392     var MSE;
393     run;
394
395     /*Box Cox Analysis*/
396     proc transreg data= train;
397       model boxcox(pH sulphates / lambda=-1 to 1 by 0.1)
398       = identity(quality);  title1 'Box-Cox Transformation pH and sulphates on quality';
399       run;
400
```