Modeling Assignment 4

Taoran Zhang

Exploratory data analysis and data prep

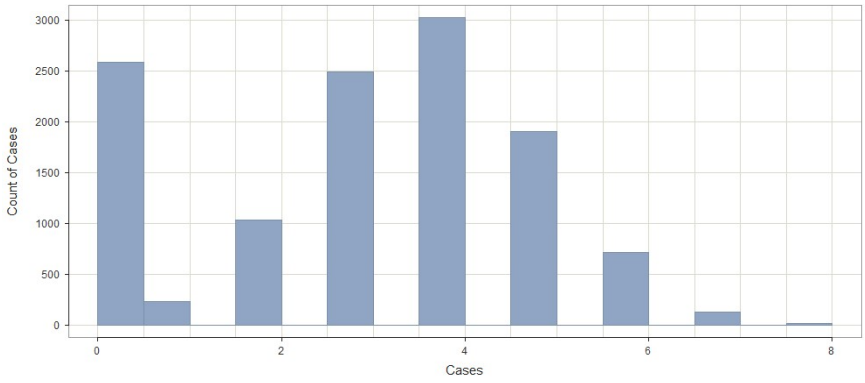
The wine study data set contains 17 columns, and the response variable of choice in this study is the number of cases of wine sold (CASES). The Index and Purchase variables are dropped from the dataset as they are not effective explanatory variables. The number of missing values in the rest 15 columns were summarized here:

Vairable	No. NA
Chlorides	638
FreeSulfurDioxide	647
TotalSulfurDioxide	682
ResidualSugar	616
Sulphates	1210
Alcohol	653
pH	395
STARS	3359

There is a total of 12795 rows, and 3359 of them have STARS variable missing, accounting for 26.5% of whole sample. I decide to drop the STARS variable as too many values are missing.

I also dropped samples with 2 or more missing variables. A total of 660 rows were removed. The rest of the NA values were imputed with median of each corresponding column. The resulted table contains 12135 rows and 14 variables, with Cases as the response variable.

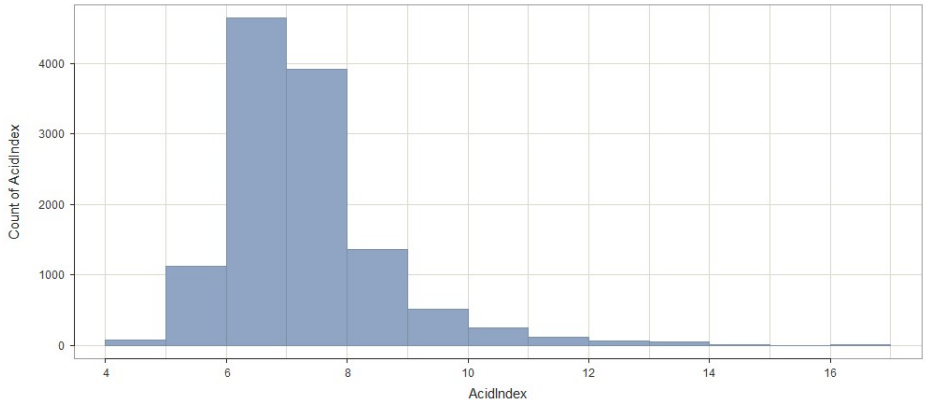
Histograms were plotted for each variables and here are some representative graphs:



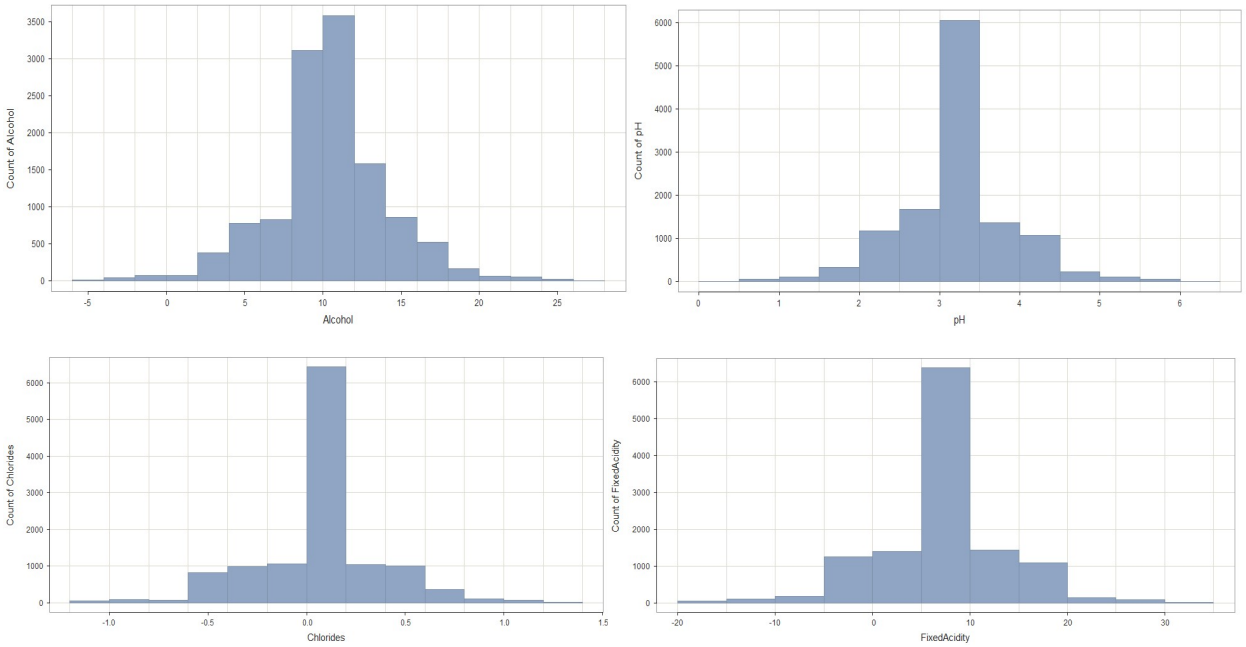
The response variable Cases has a distribution resembling Poisson Distribution, with mean 3.028 close to the variance 3.693. There are significant number of 0 cases, so I will also explore zero inflated models.

Here are some more statistic on the Cases variable:

Min. :0.000	1st Qu.:2.000	Median :3.000	Mean :3.028	3rd Qu.:4.000	Max. :8.000



The AcidIndex distribution has a long right tail, with a few extreme value above 16. Most of the samples are normally distributed.



The Alcohol, pH, Chlorides and most other variables show close to normal distribution.

Modeling

Three different models were built and evaluated. The dataset was split randomly to training data and validation data. There are 8529 rows in training data and 3606 in the validation data, approximating a 70%:30% split.

OLS Model

The first explored model is a linear regression model. The StepAIC function was used for variable selection. The upper model is the Full Model containing all 13 predictor variables in the variable pool, and the lower model as the Intercept Model. The model containing single variable ‘Chlorides’ is used to initialize the stepwise model selection.

The selected OLS model contains 11 explanatory variables with coefficients summarized here:

	Estimate	Pr(>|t|)
(Intercept)	7.31275762	2E-16
Chlorides	-0.19694714	0.00093
LabelAppeal	0.76958337	2E-16
AcidIndex	-0.36229725	2E-16
VolatileAcidity	-0.17121272	3.36E-13
Alcohol	0.02413318	2.24235E-06
TotalSulfurDioxide	0.00030001	0.000221
FreeSulfurDioxide	0.00041997	0.001036
Sulphates	-0.05962638	0.004126
Density	-1.53348256	0.028271
pH	-0.0546552	0.047036
CitricAcid	0.03907205	0.07127

It has an R square 0.201 and AIC =33453. The VIF value of all variables are close to 1.0, indicating no sign of strong collinearity.

This model was used to make prediction for the validation data set and the resulted mean absolute residual is 1.34. The mean absolute residual for training dataset is 1.342. That indicates no sign of over-fitting. However, the R squared value is 0.2, so the overall goodness of fit is not ideal.

Poisson Distribution Model

A Poisson distribution general liner model was fitted using all 13 explanatory variables.

The resulted coefficients are:

	Estimate	Pr(>|z|)
(Intercept)	2.65127528	2E-16
FixedAcidity	-0.00009866	0.921241
VolatileAcidity	-0.05747563	2.67E-13
CitricAcid	0.01354596	0.06198
ResidualSugar	0.0001458	0.439567
Chlorides	-0.06587943	0.001032
FreeSulfurDioxide	0.00013284	0.001972
TotalSulfurDioxide	0.00010343	0.000162
Density	-0.53301999	0.023434
pH	-0.02054664	0.027093
Sulphates	-0.02028456	0.00372
Alcohol	0.0074481	1.40933E-05
LabelAppeal	0.25651454	2E-16
AcidIndex	-0.13662837	2E-16

AIC = 34322 for this model, slightly more than that of the OLS model. The mean absolute residual based on the training set is 0.489, improving from the OLS model. However, when predictions were made on validation dataset using this model, the result mean absolute residual is 2.417 – higher than that of OLS model and significantly increased comparing to the training data set. That indicates this Poisson Distribution Model overfits the training data.

Zero Inflated Poisson Distribution Model

A ZIP model is fitted using the training dataset. And here are the resulted coefficients:

Poisson Model		
	Estimate	Pr(>|z|)
(Intercept)	1.351301	2E-16
VolatileAcidity	-0.0149	0.0343
Alcohol	0.009272	3.51E-10
LabelAppeal	0.282632	2E-16
AcidIndex	-0.020365	5.19845E-05
Zero-inflation model coefficients		
(Intercept)	-6.1010632	2E-16
VolatileAcidity	0.2509351	5.6E-13
CitricAcid	-0.084158	0.007458
Chlorides	0.311452	0.000278
FreeSulfurDioxide	-0.0006704	0.000277
TotalSulfurDioxide	-0.0008572	3.48E-13
pH	0.2040195	3.90008E-07
Sulphates	0.120401	6.04103E-05
LabelAppeal	0.1595718	3.40481E-07
AcidIndex	0.4884067	2E-16

The Zero-Inflation model contains 9 explanatory variables, and the Poisson Model contains 4 explanatory variables.

The coefficients should be interpreted by two parts:

Logistic: When other variables are held constant, the logit of Cases=0 increase by 0.251 when Volatile Acidity variable increase by 1 unit. The odds of Cases=0 increase by exp(b1)-1=28.52% when Volatile Acidity increase by 1 unit. The same calculation applied to other 8 variables in the logistic model. The intercept is not particularly interpretable as 0 is out of the range of several explanatory variables.

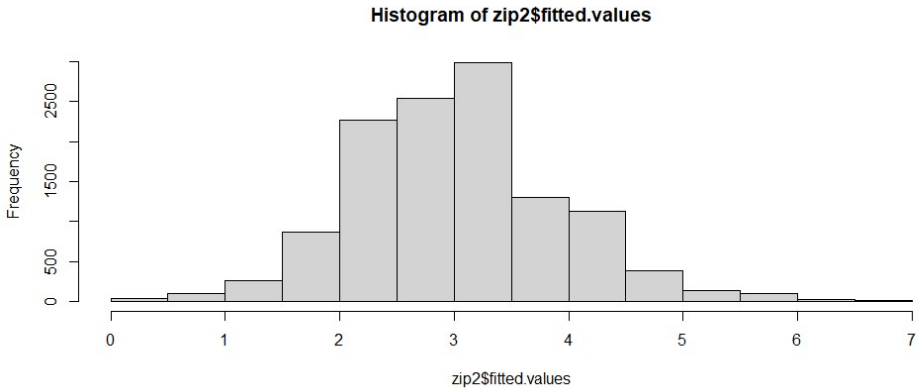
Poisson: Using Label Appeal variable as an example, when other variables are held constant, the predicted number of cases increase by exp (0.283) = 1.327 cases per unit increase of Label Appeal. While many of the variables do not have obvious logical relationship to number of cases sold, the positive coefficient makes sense for the Label Appeal variable – if a label is more appealing, a wine is more likely to sell better. Same interpretation applied for the other 3 variables and the intercept is not particularly interpretable as 0 is out of the range of several explanatory variables.

Here is a summary of some key metrics comparing these 3 models:

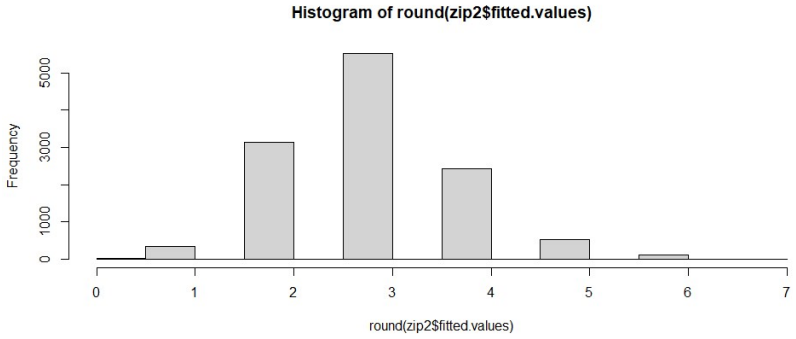
	Zip	poisson	OLS
AIC	43313	34322	33453
BIC	43424	34421	33545
Mean Abs Residual (train)	1.327	0.489	1.34
Mean Abs Residual (test)	1.325	2.417	1.342

Although the zip model has higher AIC and BIC values (likely due to the combination of 2 models), it yields significantly lower mean absolute residual on the test dataset. There is no indication of overfit either.

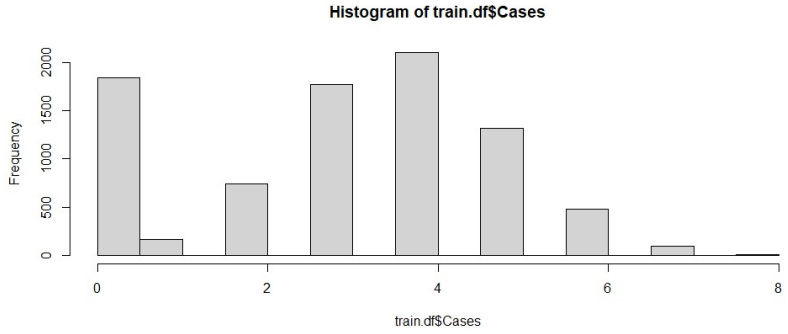
Here is the histogram of zip model predicted Cases values (training set)



I also rounded the fitted values to the nearest integers:

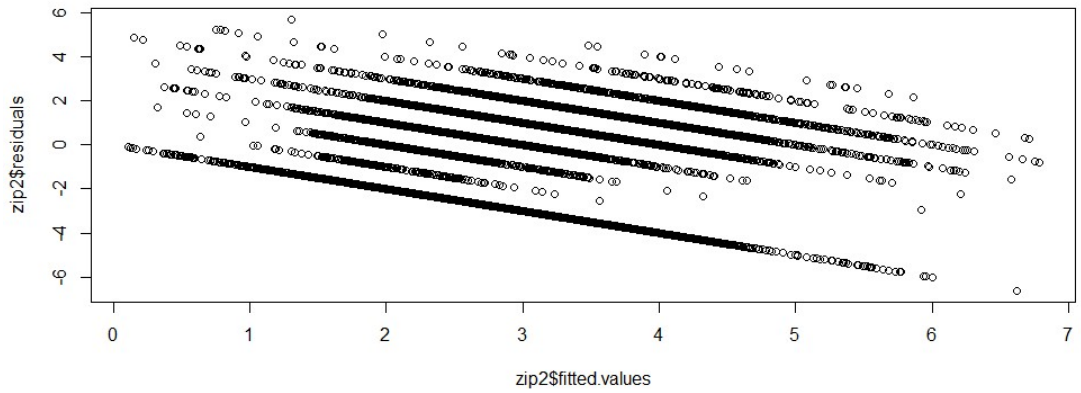


And here is the histogram of actual Cases distribution:



The zip model is able to capture the distribution of sample data for non-zero cases. The logistic part of this model is still not ideal but better than the other two models in making Cases=0 predictions. Given these comparison results, the final model is the zip model.

Here is a residual vs fitted value scatter plot:



There are linear patterns in the residual plot and the model tend of over-estimate the lower fitted values, and under-estimate the higher fitted values.

Conclusions

The fitted zip model indicates that a few variables are most crucial for the prediction of number of cases a wine can be sold. Customers tend to prefer lower volatile acidity and lower acid index, and Label Appeal and alcohol level are positively correlated to number of cases sold. Meanwhile, a number of variables affect whether a wine will be sold at all (if cases=0), for example volatile acidity, acid index and chloride are positively correlated to the odds of selling zero cases. From a data analysis perspective, I would suggest producing wines with lowers volatile acidity, acid index and choloride levels, and boosting Label Appeal as much as possible to promote the Sale performance.