Data 621 - HW5

Farhana Zahir, Vijaya Cherukuri, Scott Reed, Shovon Biswas, Habib Khan, Alain Kuiete Tchoupou

11/16/2020

Contents

Whining about Wine Sales

Overview

In this report we attempt to build a model for wine sales as would be predicted by a number of factors about the wine and its packaging. In the end a zero inflated poisson model seems to be the best model. Particularly important to sales variables seem to consist of Label Appeal, Acid Index, Stars and to a lesser extent Alcohol.

Data Exploration

Data was provided already split into a training dataset of 12,795 observations, and an evaluation dataset of 16,129 observations. There was one response variable, Number of Cases purchased, and 14 predictors.

Below is a short description of the variables of interest in the data set:

Variable Name	Description	Import
Index	Identification	Not used
Target	Number of cases purchased	Response variable
AcidIndex	Proprietary total acidity measure	
Alcohol	Alcohol content	
Chlorides	Chloride content	
Density	Density of Wine	
FixedAcidity	Fixed Acidity	
FreeSulfurDioxide	Sulfur Dioxide Content	
Label Appeal	Marketing Score indicating appeal of label	Expected positive
ResidualSugar	Residual Sugar	
STARS	Wine rating by experts	Positive
Sulphates	Sulfate content	
TotalSulfurDioxide	Total Sulfur Dioxide	
VolatileAcidity	Volatile Acid content	
pH	pH of wine	

On a first inspection, it was obvious that many variables had at least some missing data.

A sample from the training dataset is provided below:

Training Dataset

Table 2: Table continues below

TARGET	INDEX	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar
3	1	3.2	1.16	-0.98	54.2
3	2	4.5	0.16	-0.81	26.1
5	4	7.1	2.64	-0.88	14.8
3	5	5.7	0.385	0.04	18.8
4	6	8	0.33	-1.26	9.4
0	7	11.3	0.32	0.59	2.2

Table 3: Table continues below

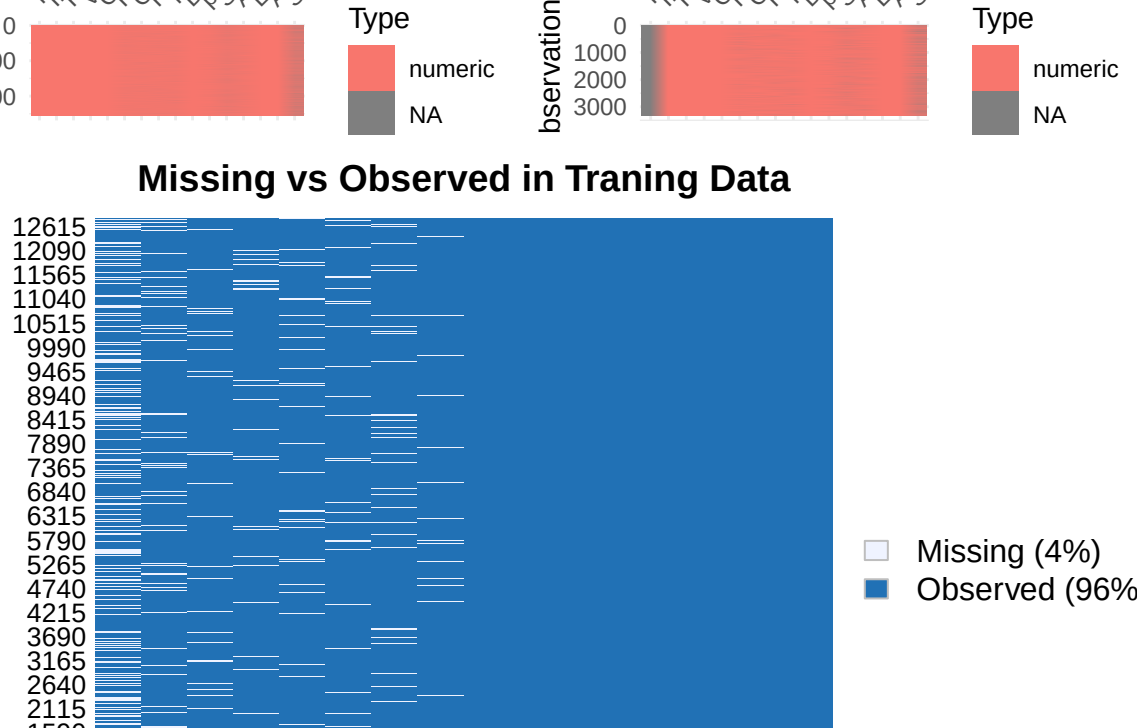
Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH
-0.567	NA	268	0.9928	3.33
-0.425	15	-327	1.028	3.38
0.037	214	142	0.9952	3.12
-0.425	22	115	0.9964	2.24
NA	-167	108	0.9946	3.12
0.556	-37	15	0.9994	3.2

Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
-0.59	9.9	0	8	2
0.7	NA	-1	7	3
0.48	22	-1	8	3
1.83	6.2	-1	6	1
1.77	13.7	0	9	2
1.29	15.4	0	11	NA

Check Data types and Missing values

The datasets both have missing values. There are 3 categorical variables (`LabelAppeal`,`AcidIndex`,`STARS`), 11 continuous variables and the target variable is categorical.


Missing Values and Data Type Check
Observations
Missing (4%)
Present (96%)
Missing (10.1%)
Present (89.9%)
Type
numeric
NA


Missing vs Observed in Traning Data
Missing (4%)
Observed (96%
STARS
Sulphates
TotalSulfurDioxide
Alcohol
FreeSulfurDioxide
Chlorides
ResidualSugar
pH
AcidIndex
LabelAppeal
Density
CitricAcid
VolatileAcidity
FixedAcidity
INDEX
TARGET

	Non_NAs	NAs	NA_Percent
TARGET	12795	0	0

INDEX 12795 0 0

FixedAcidity 12795 0 0

VolatileAcidity 12795 0 0

CitricAcid 12795 0 0

ResidualSugar 12179 616 0.04814

Chlorides 12157 638 0.04986

FreeSulfurDioxide 12148 647 0.05057

TotalSulfurDioxide 12113 682 0.0533

Density 12795 0 0

pH 12400 395 0.03087

Sulphates 11585 1210 0.09457

Alcohol 12142 653 0.05104

LabelAppeal 12795 0 0

AcidIndex 12795 0 0

STARS 9436 3359 0.2625

Data Statistics Summary

A binary logistic regression model is built using the `training set`, therefore the `training set` is used for the following data exploration.

The data types in the raw dataset are all 'doubles', however the counter `INDEX` and the response variable `target` are categorical.

The statistics of all variables are listed below:

Table 6: Table continues below

TARGET	FixedAcidity	VolatileAcidity	CitricAcid
Min. :0.000	Min. :-18.100	Min. :-2.7900	Min. :-3.2400
1st Qu.:2.000	1st Qu.: 5.200	1st Qu.: 0.1300	1st Qu.: 0.0300
Median :3.000	Median : 6.900	Median : 0.2800	Median : 0.3100
Mean :3.029	Mean : 7.076	Mean : 0.3241	Mean : 0.3084
3rd Qu.:4.000	3rd Qu.: 9.500	3rd Qu.: 0.6400	3rd Qu.: 0.5800
Max. :8.000	Max. : 34.400	Max. : 3.6800	Max. : 3.8600
NA	NA	NA	NA

Table 7: Table continues below

ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide
Min. :-127.800	Min. :-1.1710	Min. :-555.00	Min. :-823.0
1st Qu.: -2.000	1st Qu.:-0.0310	1st Qu.: 0.00	1st Qu.: 27.0
Median : 3.900	Median : 0.0460	Median : 30.00	Median : 123.0
Mean : 5.419	Mean : 0.0548	Mean : 30.85	Mean : 120.7
3rd Qu.: 15.900	3rd Qu.: 0.1530	3rd Qu.: 70.00	3rd Qu.: 208.0
Max. : 141.150	Max. : 1.3510	Max. : 623.00	Max. :1057.0
NA's :616	NA's :638	NA's :647	NA's :682

Table 8: Table continues below

Density	pH	Sulphates	Alcohol
Min. :0.8881	Min. :0.480	Min. :-3.1300	Min. :-4.70
1st Qu.:0.9877	1st Qu.:2.960	1st Qu.: 0.2800	1st Qu.: 9.00
Median :0.9945	Median :3.200	Median : 0.5000	Median :10.40
Mean :0.9942	Mean :3.208	Mean : 0.5271	Mean :10.49
3rd Qu.:1.0005	3rd Qu.:3.470	3rd Qu.: 0.8600	3rd Qu.:12.40
Max. :1.0992	Max. :6.130	Max. : 4.2400	Max. :26.50
NA	NA's :395	NA's :1210	NA's :653

LabelAppeal	AcidIndex	STARS
Min. :-2.000000	Min. : 4.000	Min. :1.000
1st Qu.:-1.000000	1st Qu.: 7.000	1st Qu.:1.000
Median : 0.000000	Median : 8.000	Median :2.000
Mean :-0.009066	Mean : 7.773	Mean :2.042
3rd Qu.: 1.000000	3rd Qu.: 8.000	3rd Qu.:3.000
Max. : 2.000000	Max. :17.000	Max. :4.000
NA	NA	NA's :3359

The statistics of TARGET Variable. TARGET: Number of Cases Purchased as Actual

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    StdD    Skew    Kurt
##    0.00    2.00    3.00    3.03    4.00    8.00    1.93   -0.33   -0.88
```

DATA EXPLORATION

Attributes

FixedAcidity: This variable tells us about the FixedAcidity of wine.

VolatileAcidity: This variable tells us about the Volatile Acidity content of Wine.

CitricAcid: This variable tells us about the Citric Acid Content of wine.

ResidualSugar: This variable tells us about the Residual Sugar of wine.

Chlorides: This variable tells us about the Chloride content of wine.

FreeSulfurDioxide : This variable tells us about the Sulfur Dioxide content of wine.

TotalSulfurDioxide : This variable tells us about the Total Sulfur Dioxide of Wine.

Density: This variable tells us about the Density of wine.

Sulphates: This variable tells us about the Sulphates content of wine.

Alcohol: This variable tells us about the Alcohol content.

LabelAppeal: Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design.

AcidIndex: Proprietary method of testing total acidity of wine by using a weighted average.

STARS: Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor. A high number of stars suggests high sales.

Outliers

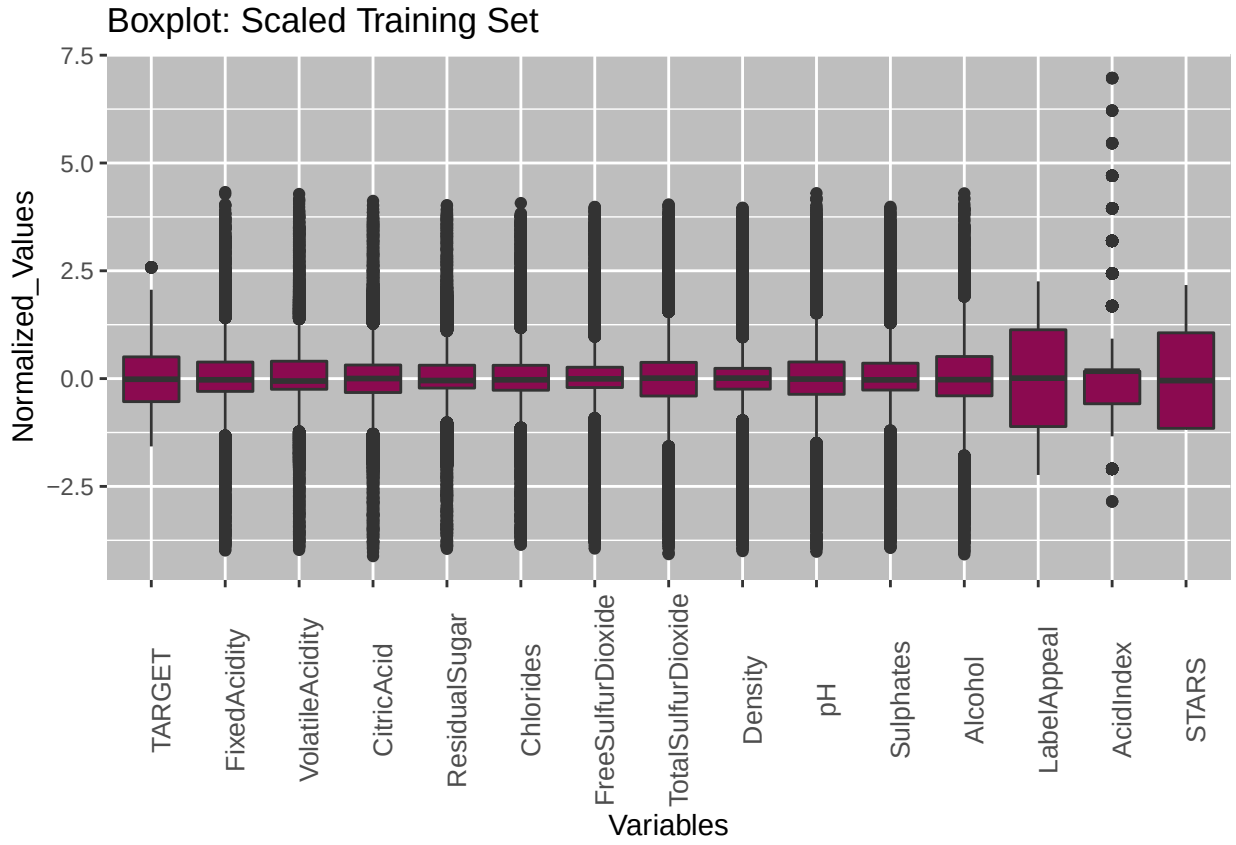
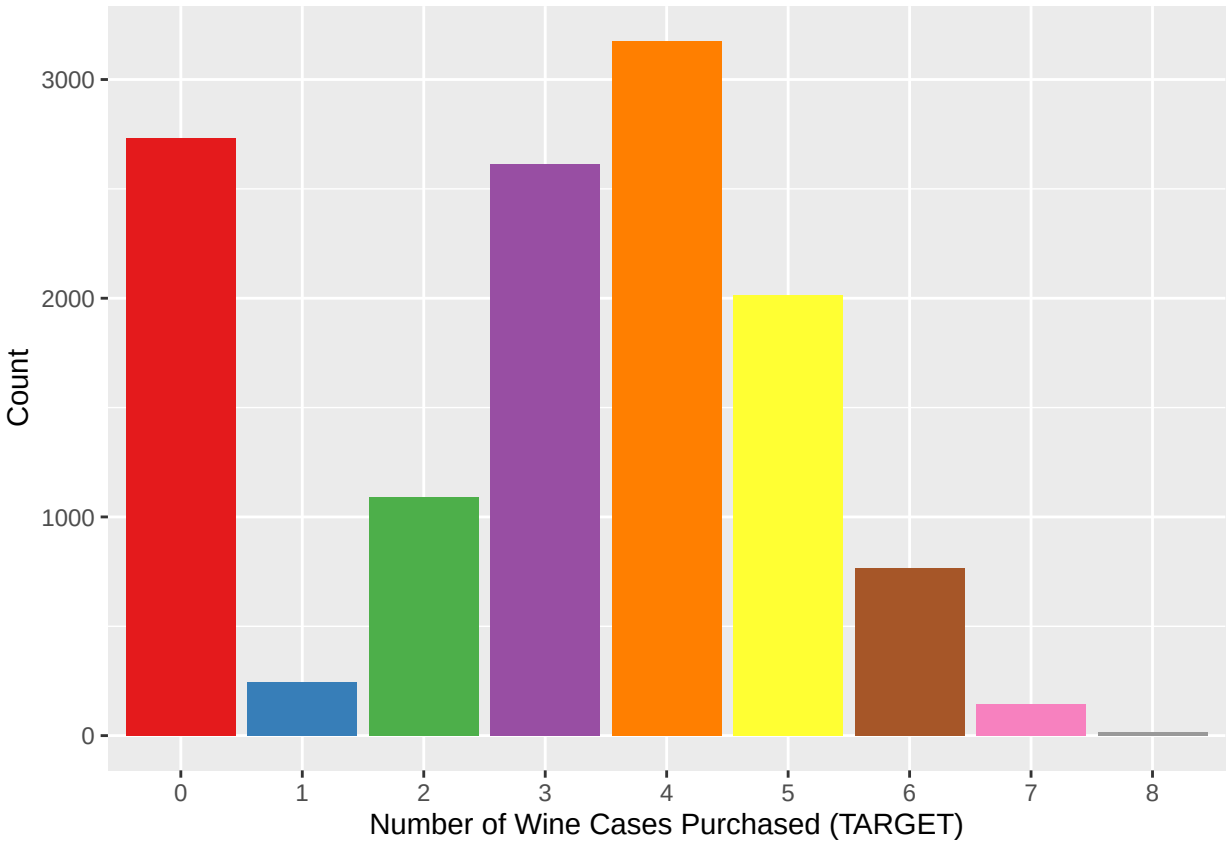


Figure 1: Boxplot: Scaled Training Set

The box plot shows that outliers exist in variables `FixedAcidity`, `VolatileAcidity`, `CitricAcid`, `ResidualSugar`, `Chlorides`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `Density`, `pH`, `Sulphates`, `Alcohol`, `LabelAppeal` and`AcidIndex`.

Univariate Analysis

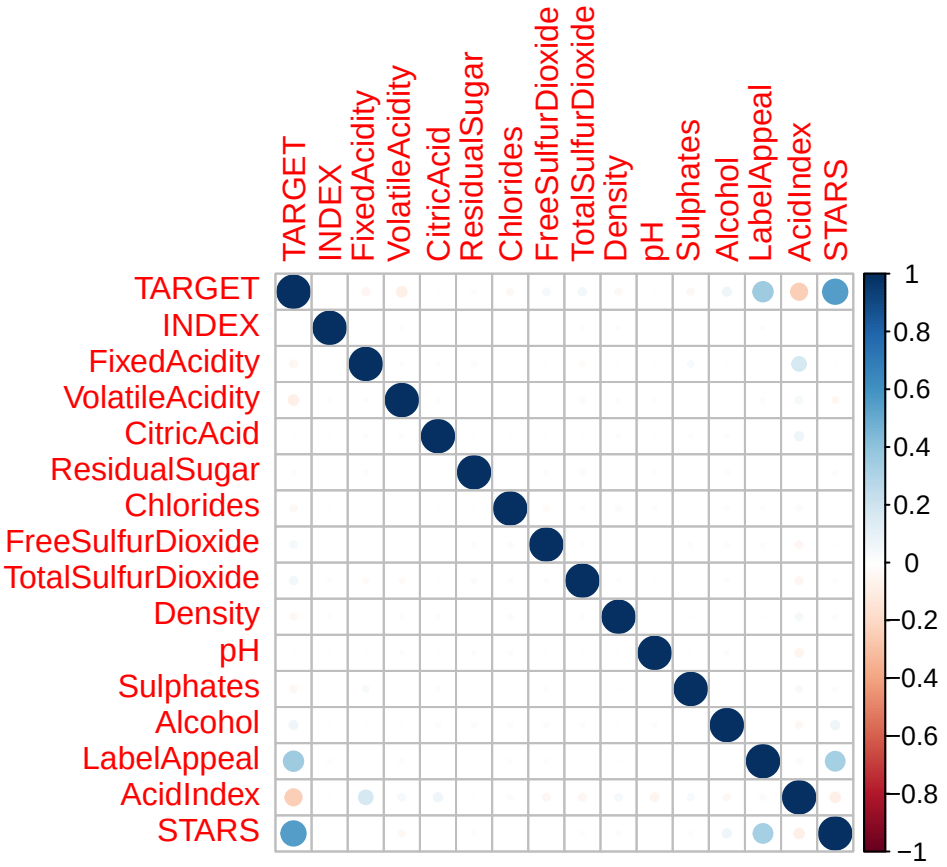
Response Variable



Upon examining the target variable we immediately see a normal distribution save for the large number of unsold cases.

Correlation Plot

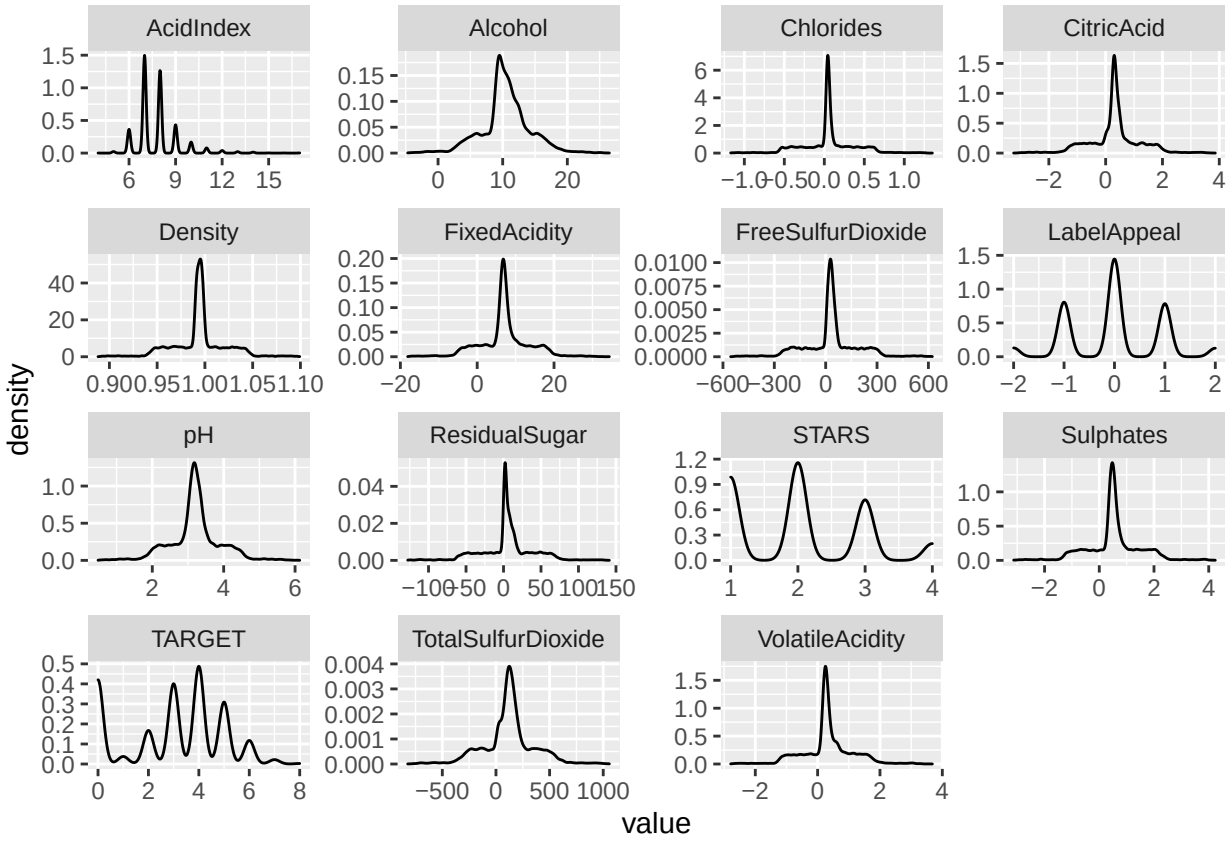
The correlation plot below shows how variables in the dataset are related to each other.



Here we see relatively little in strong correlations with almost all of the chemistry having minimal impact. Interestingly there is a some relation between star rating and label appeal. As one would expect there is a relationship with Acid index and fixed acidity.

Density Plot

Based on the below plots we can observe that `AcidIndex` is right skewed; `AcidIndex`, `STARS`, `LabelAppeal` and `TARGET` have multi-modal distribution (as expected because they are categorical). While most others seem to be normally distributed.



Summarized Data Dictionary

As a summary of the data exploration process, a data dictionary is presented below:

Variable	Missing_Value	Mean	Median	Max	Min	SD	Correlation_vs_Response
TARGET	No	NA	NA	NA	NA	NA	1.00
INDEX	No	NA	NA	NA	NA	NA	0.00
FixedAcidity	No	7.08	6.90	34.40	-18.10	6.32	-0.05
VolatileAcidity	No	0.32	0.28	3.68	-2.79	0.78	-0.09
CitricAcid	No	0.31	0.31	3.86	-3.24	0.86	0.01
ResidualSugar	No	NA	NA	NA	NA	NA	NA
Chlorides	No	NA	NA	NA	NA	NA	NA
FreeSulfurDioxide	No	NA	NA	NA	NA	NA	NA
TotalSulfurDioxide	No	NA	NA	NA	NA	NA	NA
Density	No	0.99	0.99	1.10	0.89	0.03	-0.04
pH	No	NA	NA	NA	NA	NA	NA
Sulphates	No	NA	NA	NA	NA	NA	NA
Alcohol	No	NA	NA	NA	NA	NA	NA
LabelAppeal	No	-0.01	0.00	2.00	-2.00	0.89	0.36
AcidIndex	No	7.77	8.00	17.00	4.00	1.32	-0.25
STARS	No	NA	NA	NA	NA	NA	NA

DATA PREPARATION

In the data preparation we will split data into training and test dataset.

MICE package (Multivariate Imputation by Chained Equations)implements a method to deal with missing data. The package creates multiple imputations (replacement values) for multivariate missing data. helps in inspecting, imputing, diagnose, analyze, pool the result, and generate simulated incomplete data

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
```

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
```

'AcidIndex' and 'TARGET' have low correlation between them. We will apply a log transformation to it even if it doesn't seem likely to provide a large model improvement.

BUILD MODELS

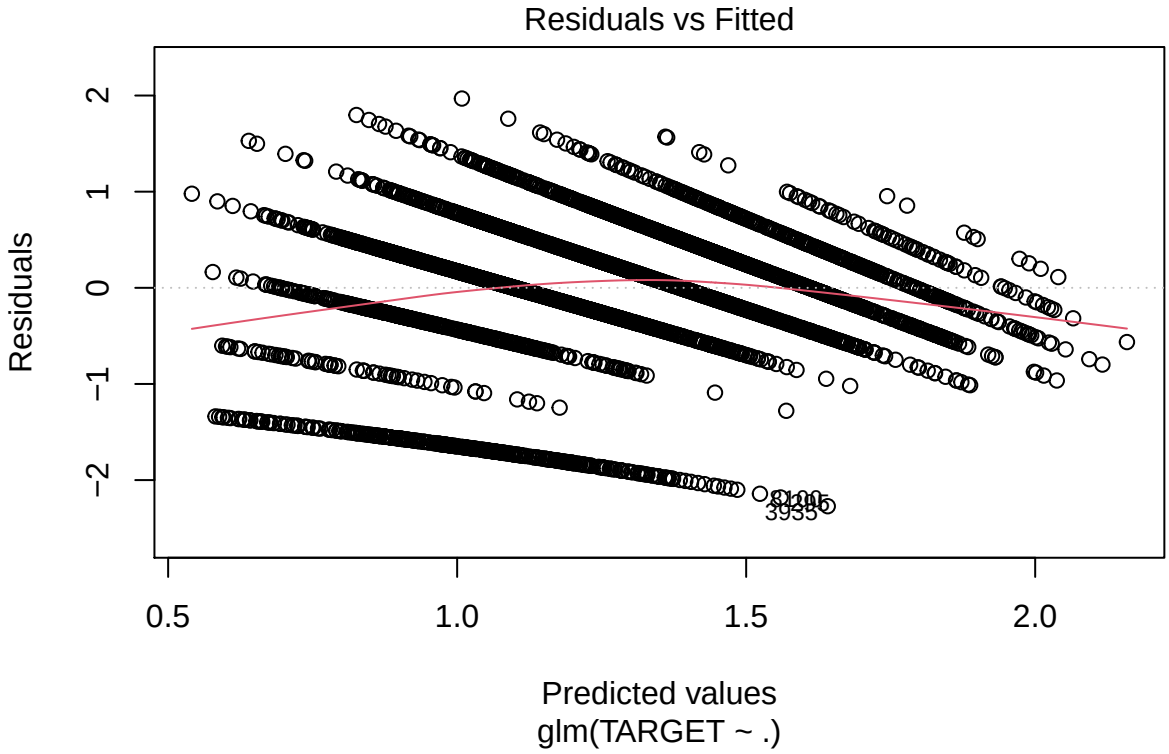
Poisson Model

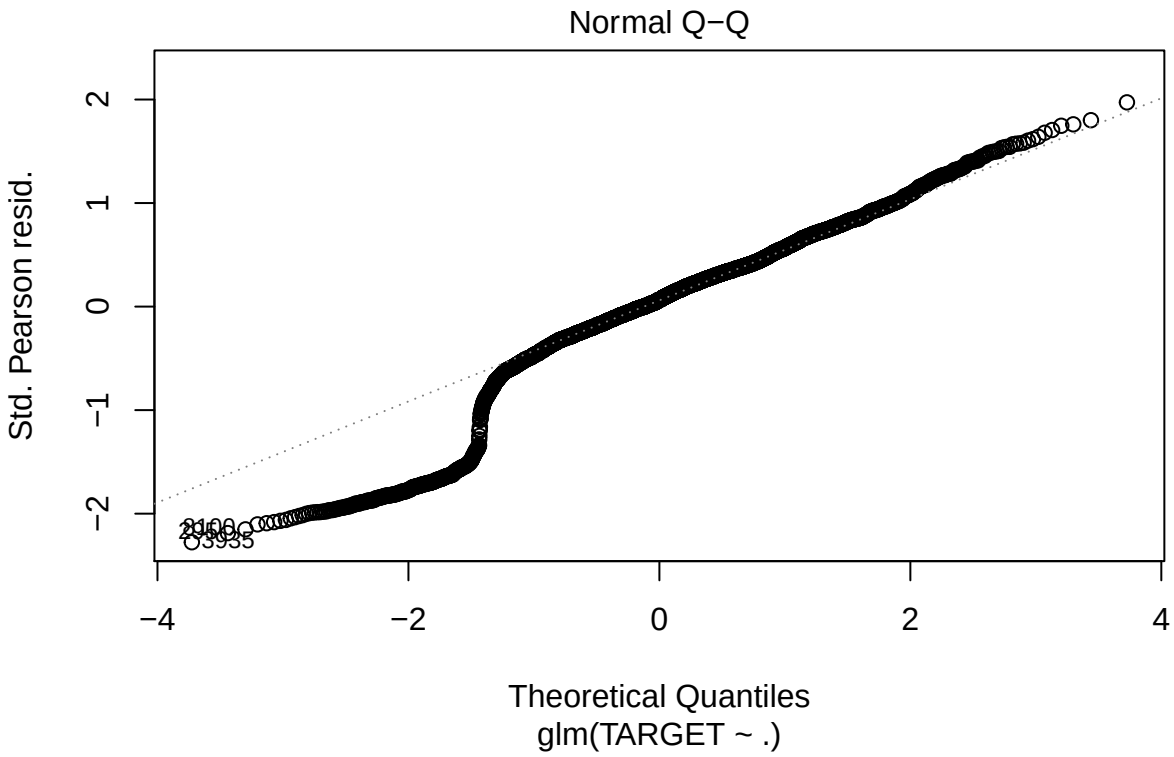
Model 1: Poisson Model without imputations

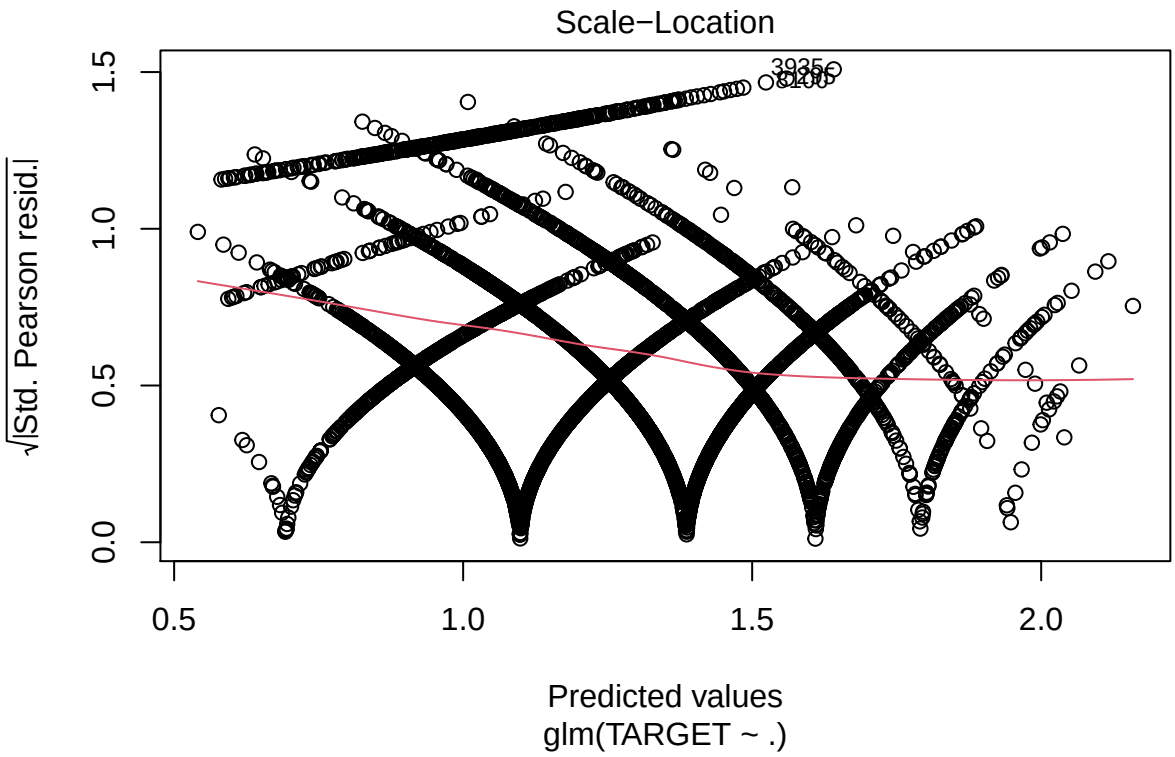
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.608	0.2796	5.75	8.902e-09
FixedAcidity	0.0006705	0.001177	0.5695	0.569
VolatileAcidity	-0.0275	0.009283	-2.963	0.00305
CitricAcid	-0.003835	0.008519	-0.4502	0.6526
ResidualSugar	1.828e-05	0.0002152	0.08493	0.9323
Chlorides	-0.03764	0.02314	-1.627	0.1038
FreeSulfurDioxide	5.671e-05	4.892e-05	1.159	0.2463
TotalSulfurDioxide	2.23e-05	3.177e-05	0.7019	0.4827
Density	-0.4025	0.2749	-1.464	0.1433
pH	0.0002307	0.01085	0.02127	0.983
Sulphates	-0.005984	0.007973	-0.7505	0.4529
Alcohol	0.003262	0.002004	1.628	0.1036
LabelAppeal	0.173	0.008858	19.53	6.135e-85
AcidIndex	-0.04967	0.006666	-7.451	9.281e-14
STARS	0.1929	0.008328	23.16	1.146e-118

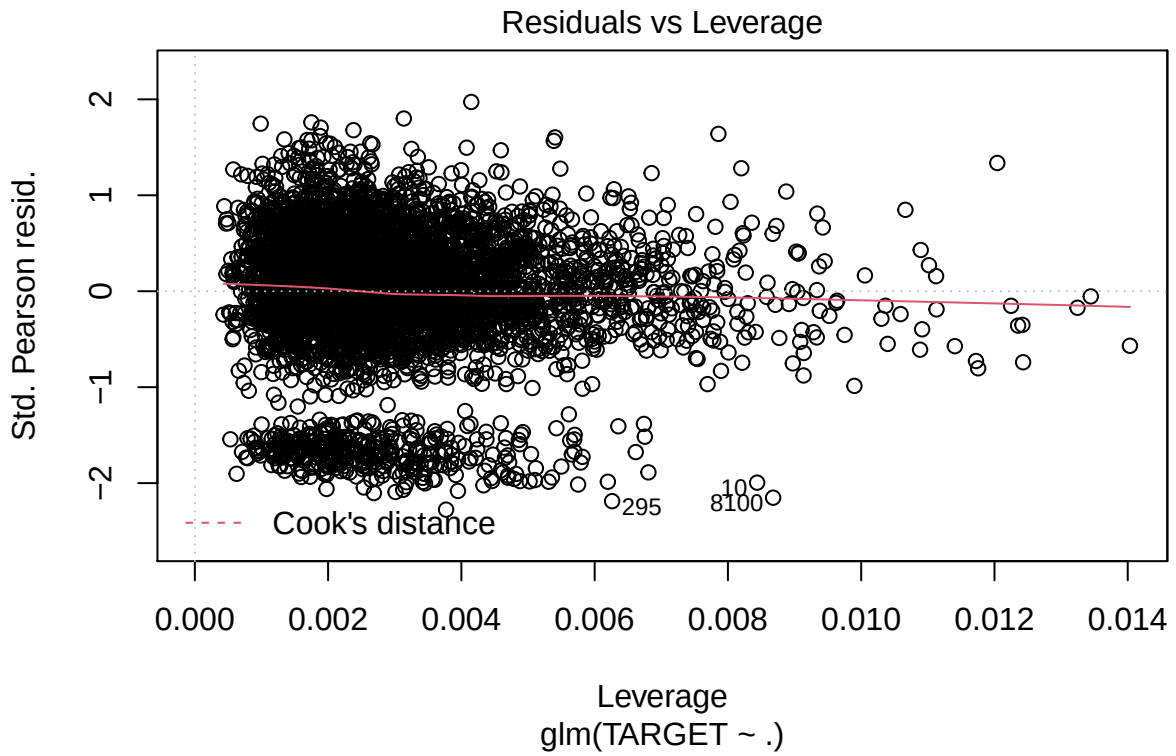
(Dispersion parameter for poisson family taken to be 1)

Null deviance:	4721 on 5143 degrees of freedom
Residual deviance:	3243 on 5129 degrees of freedom




Normal Q−Q
Std. Pearson resid.
−2
−1
0
1
2
−4
−2
0
2
4
Theoretical Quantiles
glm(TARGET ~ .)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
Predicted values
glm(TARGET ~ .)

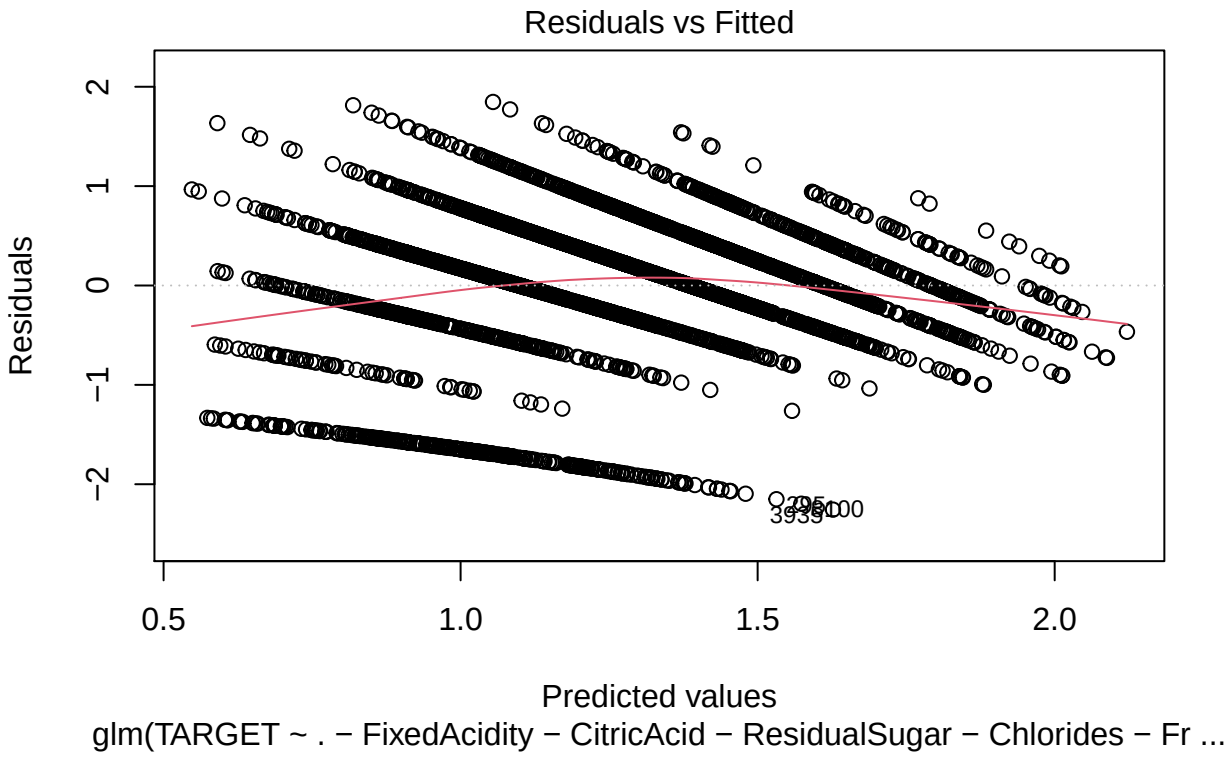


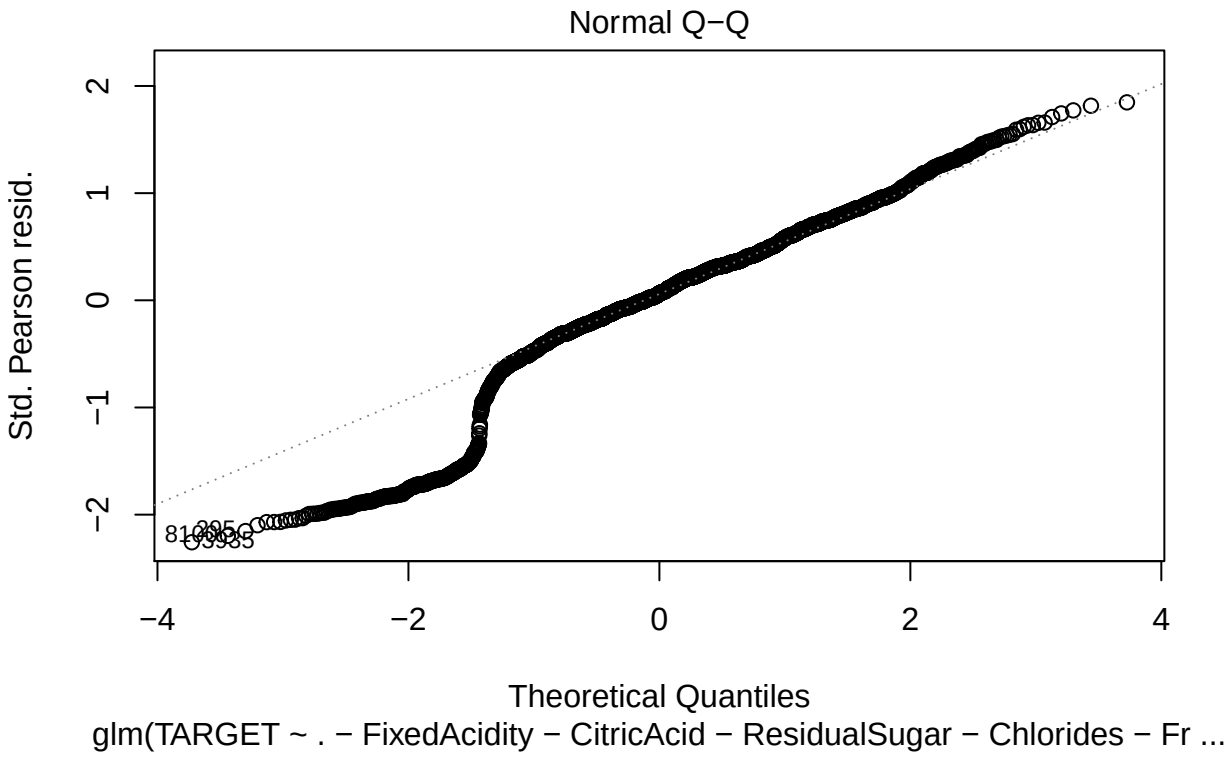
Stars and Label appeal seem to have a strong positive impact on sales, which is to be expected (and will be seen in all models). On the acid front the acidindex value seems a better predictor than the acid components. ### Model 2: Poisson Model without imputations and only significant variables

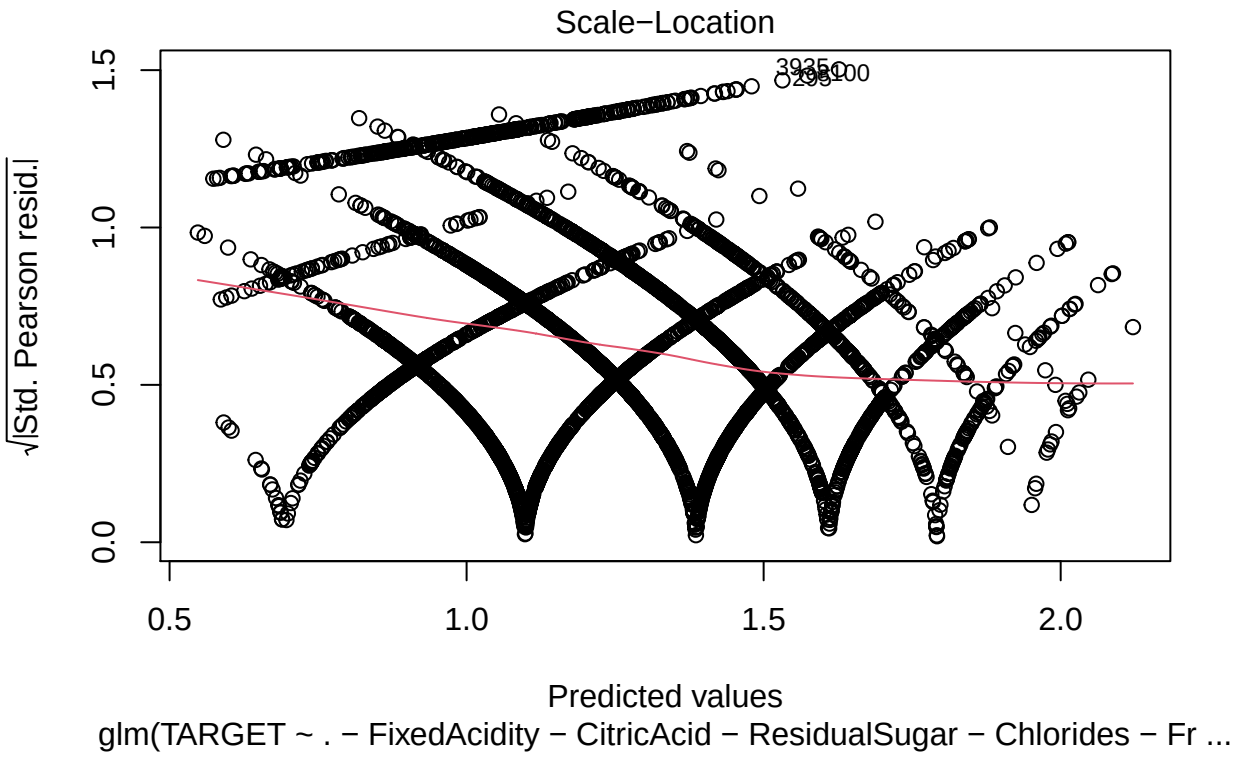
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.251	0.05472	22.87	9.626e-116
VolatileAcidity	-0.02758	0.009278	-2.973	0.002953
LabelAppeal	0.1732	0.008853	19.56	3.247e-85
AcidIndex	-0.05062	0.006553	-7.724	1.129e-14
STARS	0.1942	0.008292	23.42	2.601e-121

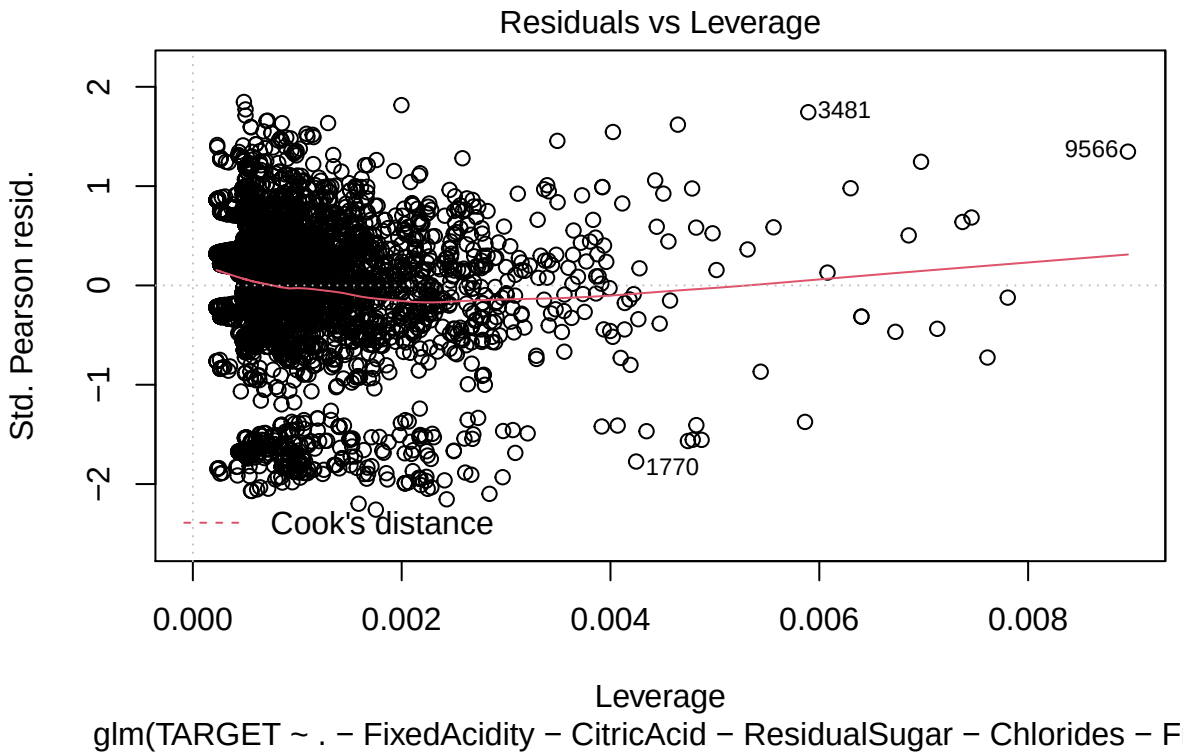
(Dispersion parameter for poisson family taken to be 1)

Null deviance:	4721 on 5143 degrees of freedom
Residual deviance:	3253 on 5139 degrees of freedom


Residuals vs Fitted
Residuals
Predicted values
glm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Chlorides − Fr ...
2
1
0
−1
−2
0.5
1.0
1.5
2.0


Normal Q−Q
Std. Pearson resid.
2
1
0
−1
−2
−4
−2
0
2
4
Theoretical Quantiles
glm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Chlorides − Fr ...


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
Predicted values
glm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Chlorides − Fr ...



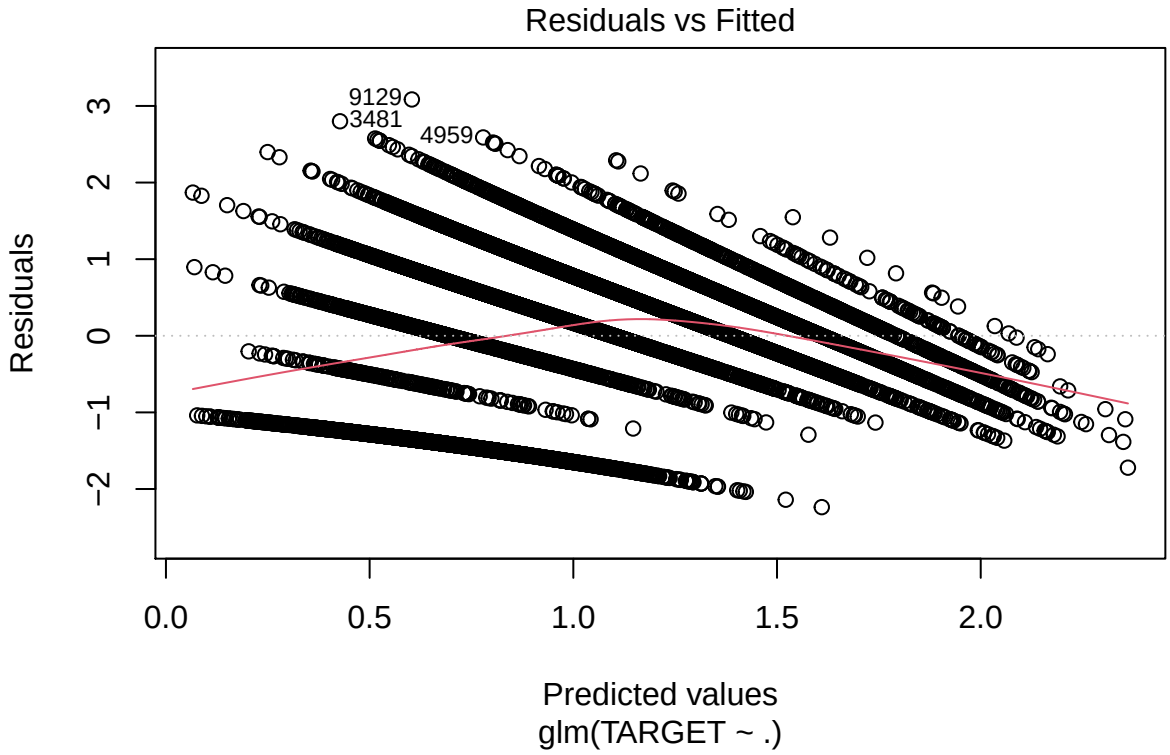
Reducing the variables in the model doesn't have a major impact on the model, and will probably improve its performance in actual fact.

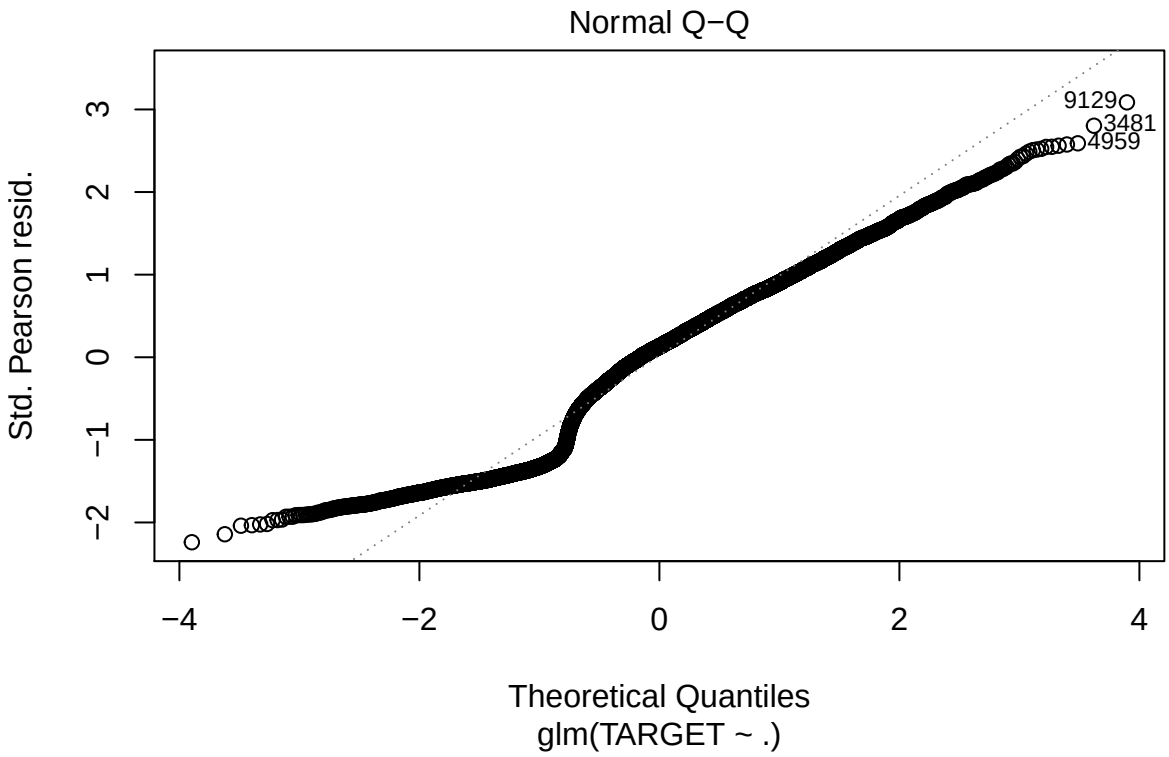
Model 3: Poisson Model with imputations

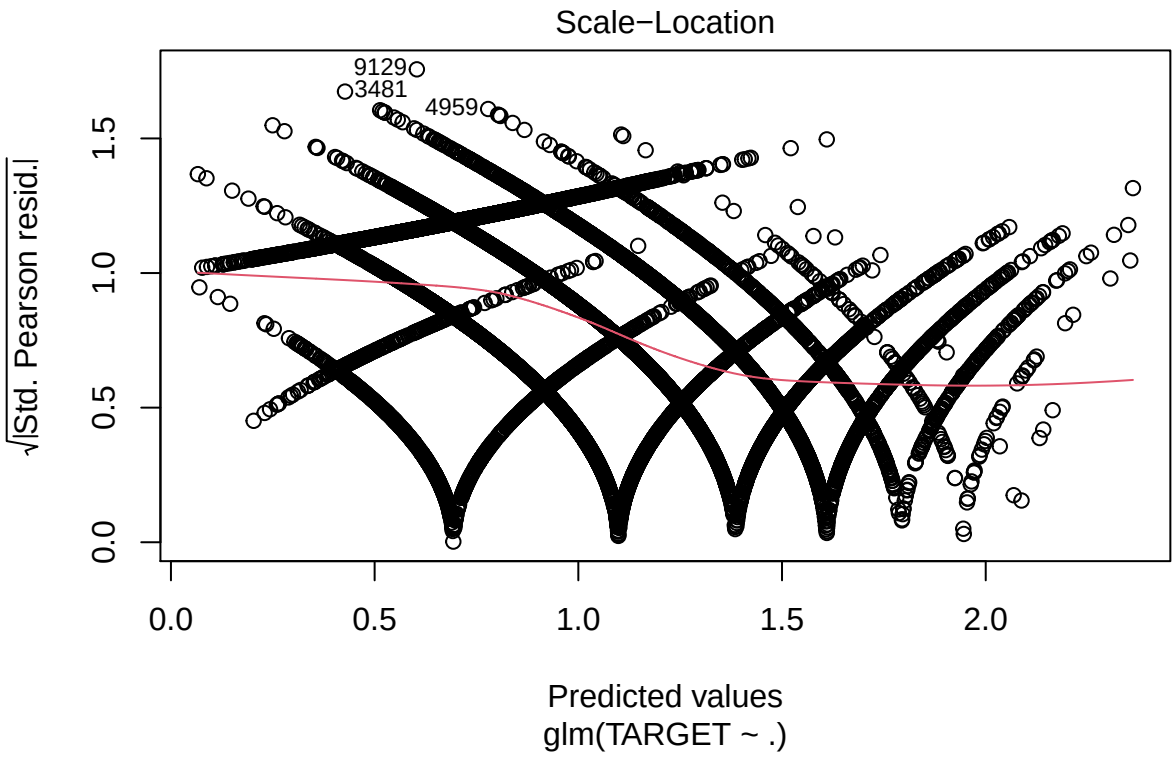
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.337	0.2281	10.24	1.281e-24
FixedAcidity	0.000225	0.000919	0.2448	0.8066
VolatileAcidity	-0.04313	0.007286	-5.919	3.232e-09
CitricAcid	0.008534	0.006573	1.298	0.1942
ResidualSugar	0.0001271	0.0001675	0.7587	0.448
Chlorides	-0.06572	0.0179	-3.673	0.0002401
FreeSulfurDioxide	0.0001336	3.804e-05	3.512	0.0004439
TotalSulfurDioxide	9.235e-05	2.46e-05	3.754	0.0001737
Density	-0.3404	0.2144	-1.588	0.1124
pH	-0.01962	0.008417	-2.331	0.01974
Sulphates	-0.01569	0.006157	-2.549	0.01081
Alcohol	0.002951	0.001554	1.898	0.05763
LabelAppeal	0.1409	0.006798	20.72	2.1e-95
AcidIndex	-0.7709	0.03998	-19.28	7.981e-83
STARS	0.3407	0.00627	54.34	0

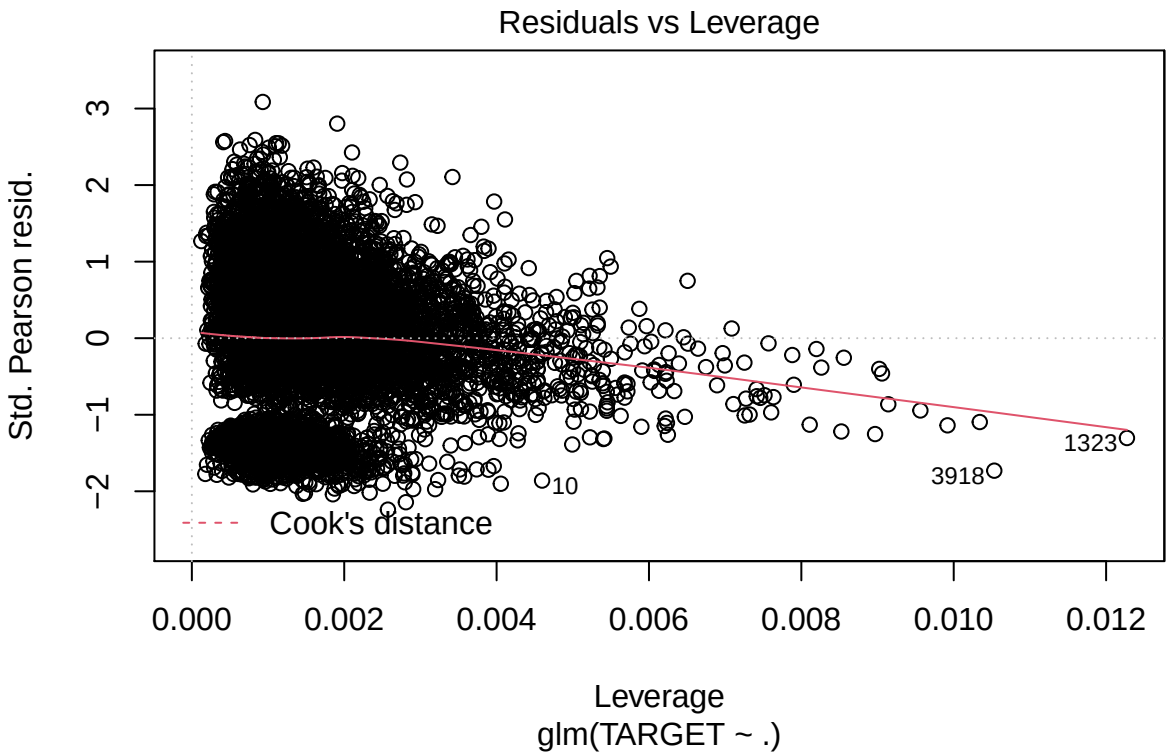
(Dispersion parameter for poisson family taken to be 1)

Null deviance:	18291 on 10236 degrees of freedom
Residual deviance:	12829 on 10222 degrees of freedom




Normal Q−Q
Std. Pearson resid.
−2
−1
0
1
2
3
9129
3481
4959
−4
−2
0
2
4
Theoretical Quantiles
glm(TARGET ~ .)


Scale−Location
9129
3481
4959
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
Predicted values
glm(TARGET ~ .)



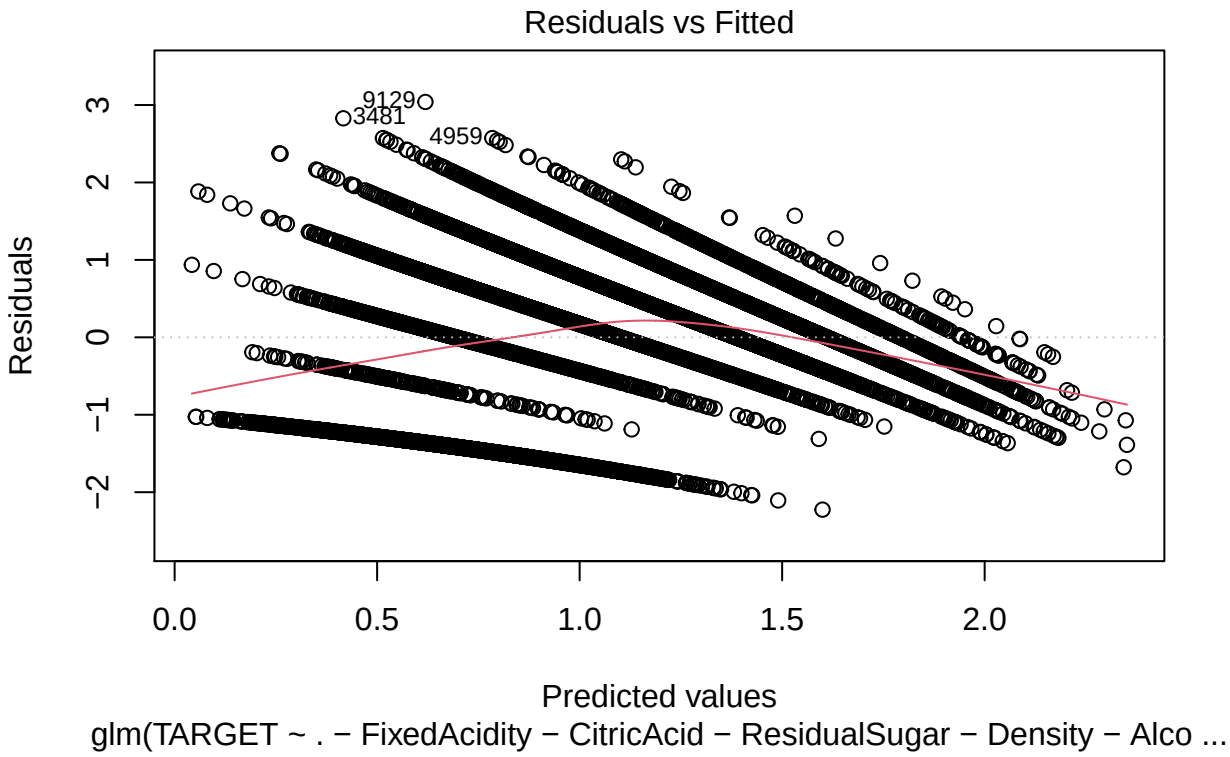
With imputations a number of variables become significant, including acidity, and Sulfur/sulphate counts. This makes some intuitive sense as those would affect the taste of the wine. Interestingly, the coefficient is positive on the sulfur dioxide. One would rarely expect sulfur dioxide (the smell of burnt matches) to be an improvement for wine.

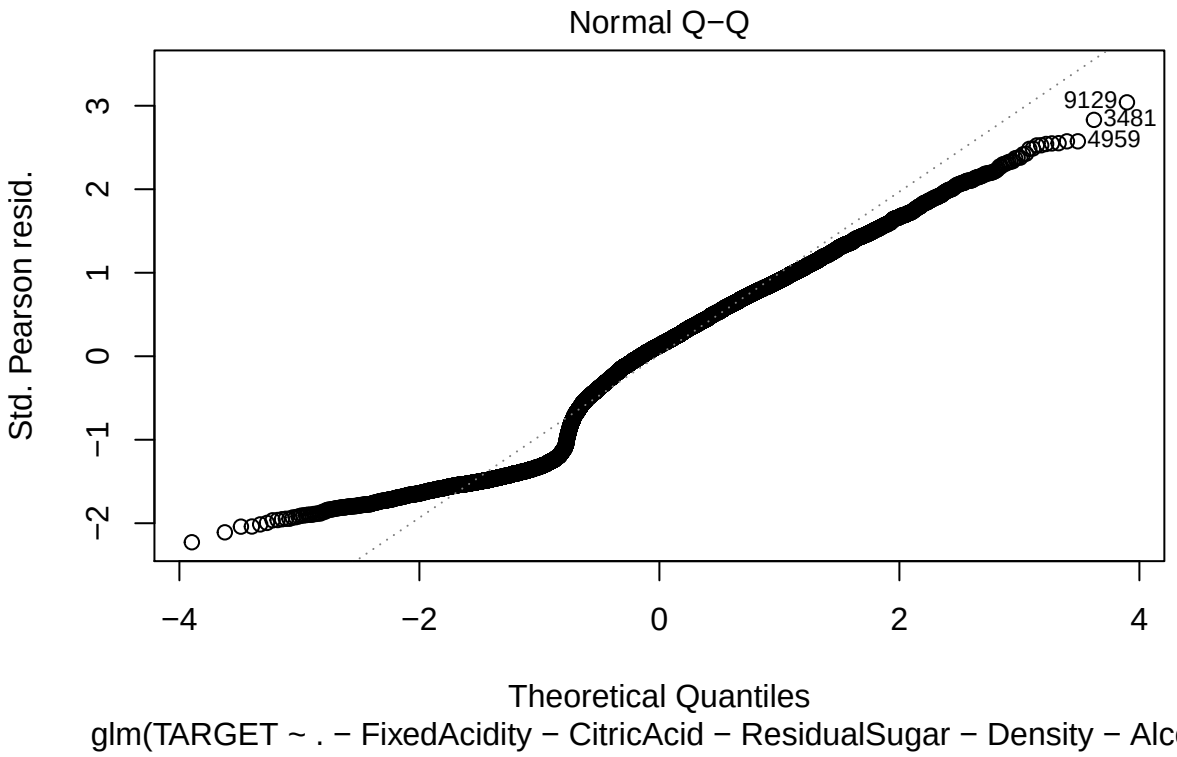
Model 4: Poisson Model with imputations and only significant variables

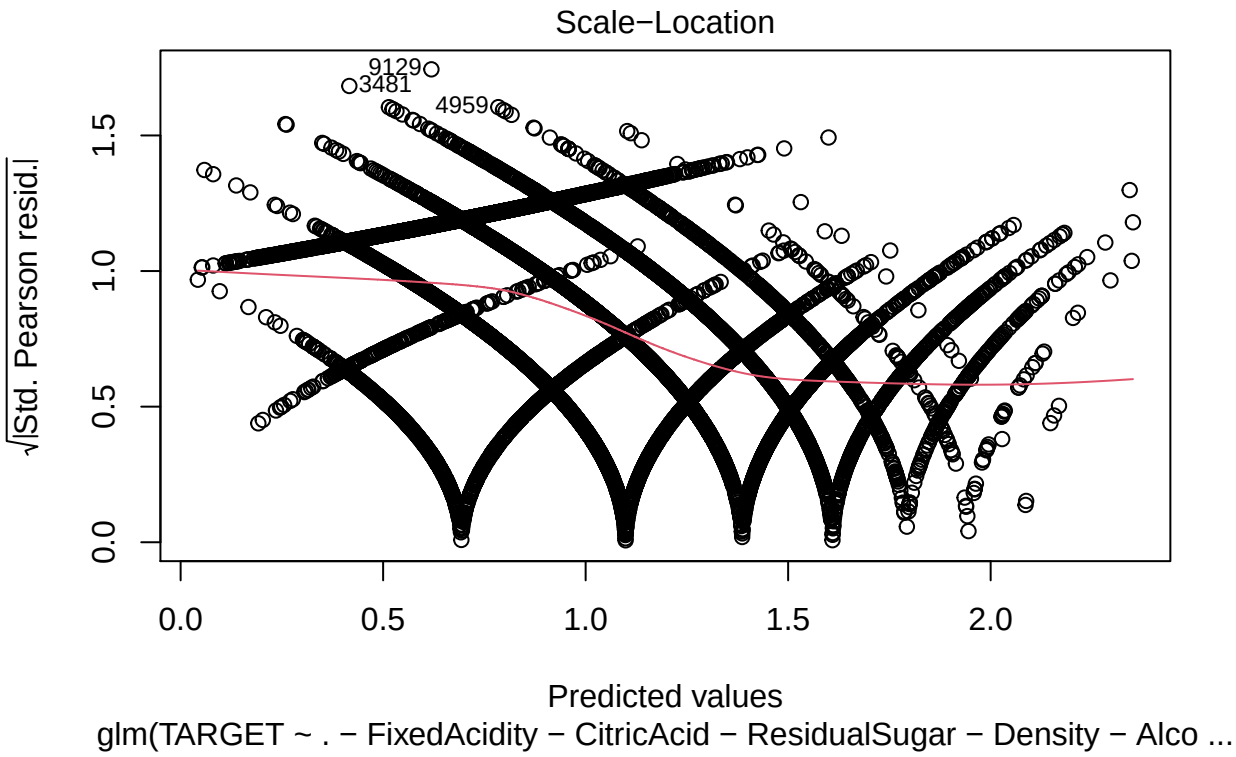
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.038	0.0884	23.05	1.395e-117
VolatileAcidity	-0.04348	0.007284	-5.969	2.39e-09
Chlorides	-0.06725	0.01789	-3.76	0.0001702
FreeSulfurDioxide	0.0001316	3.801e-05	3.461	0.0005373
TotalSulfurDioxide	9.15e-05	2.458e-05	3.723	0.0001968
pH	-0.01991	0.008415	-2.366	0.018
Sulphates	-0.01563	0.006153	-2.54	0.01109
LabelAppeal	0.1409	0.006798	20.73	1.987e-95
AcidIndex	-0.7729	0.03936	-19.64	7.554e-86
STARS	0.3417	0.006255	54.63	0

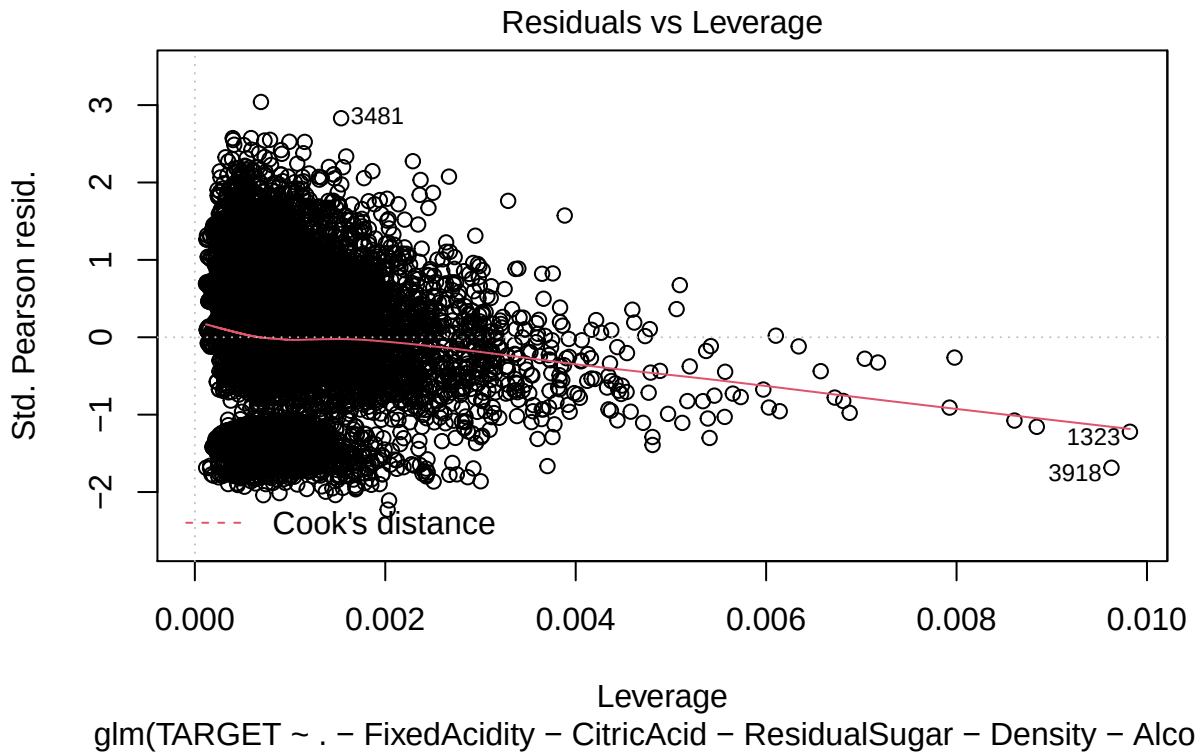
(Dispersion parameter for poisson family taken to be 1)

Null deviance:	18291 on 10236 degrees of freedom
Residual deviance:	12837 on 10227 degrees of freedom


Residuals vs Fitted
9129
3481
4959
Residuals
3
2
1
0
−1
−2
0.0
0.5
1.0
1.5
2.0
Predicted values
glm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − Alco ...


Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − Alco ...
9129
3481
4959
−2
−1
0
1
2
3
−4
−2
0
2
4


Scale−Location
9129
3481
4959
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
Predicted values
glm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − Alco ...



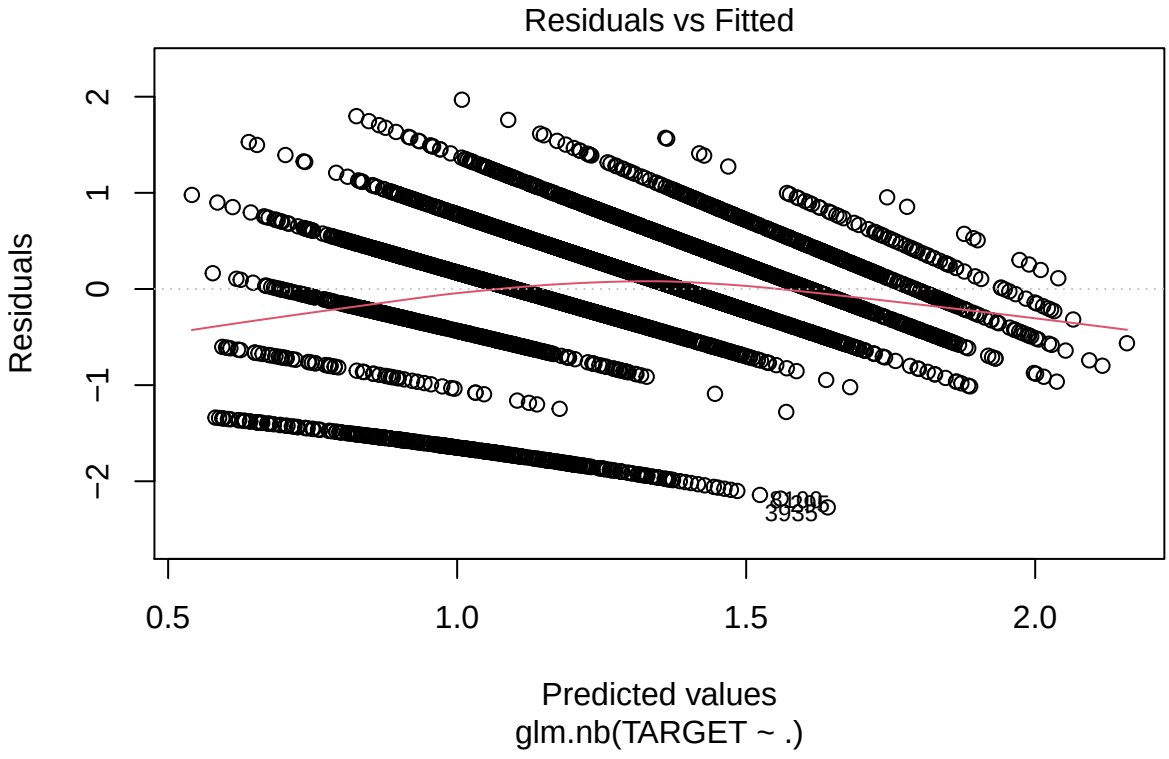
Using just the significant variables we do not get much of an improvement in AIC, but the model seems to lose little.

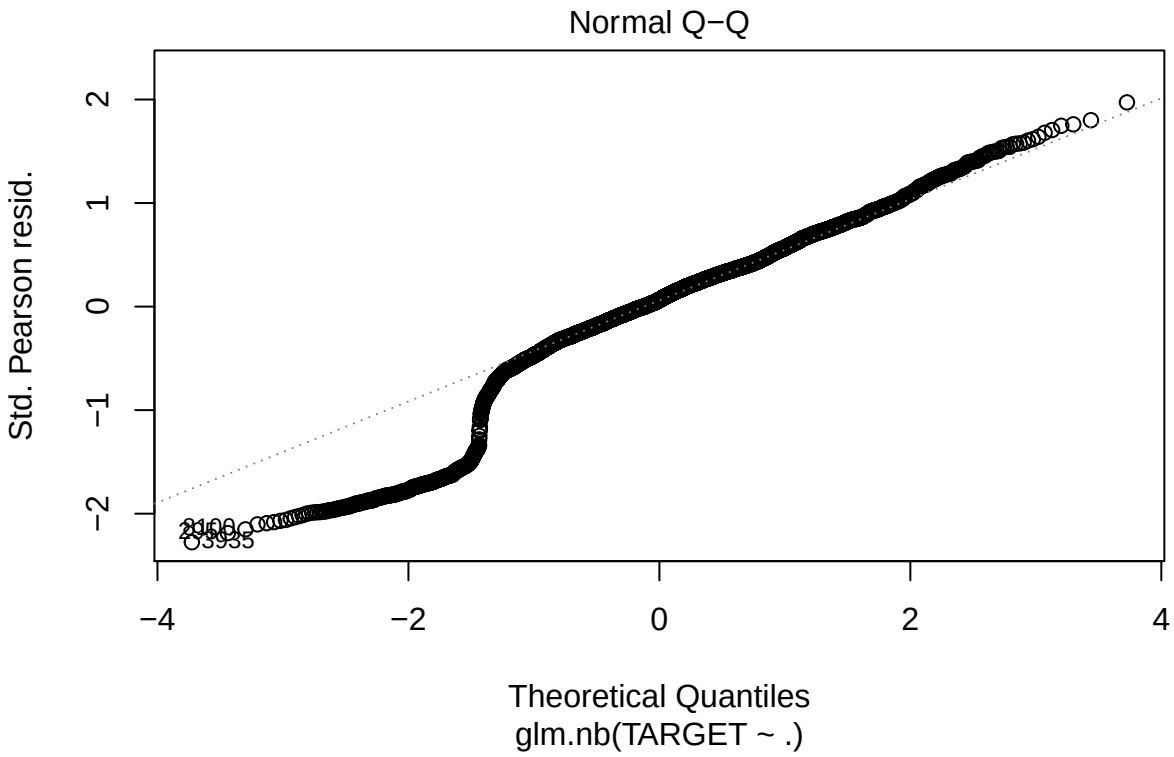
Negative Binomial Model

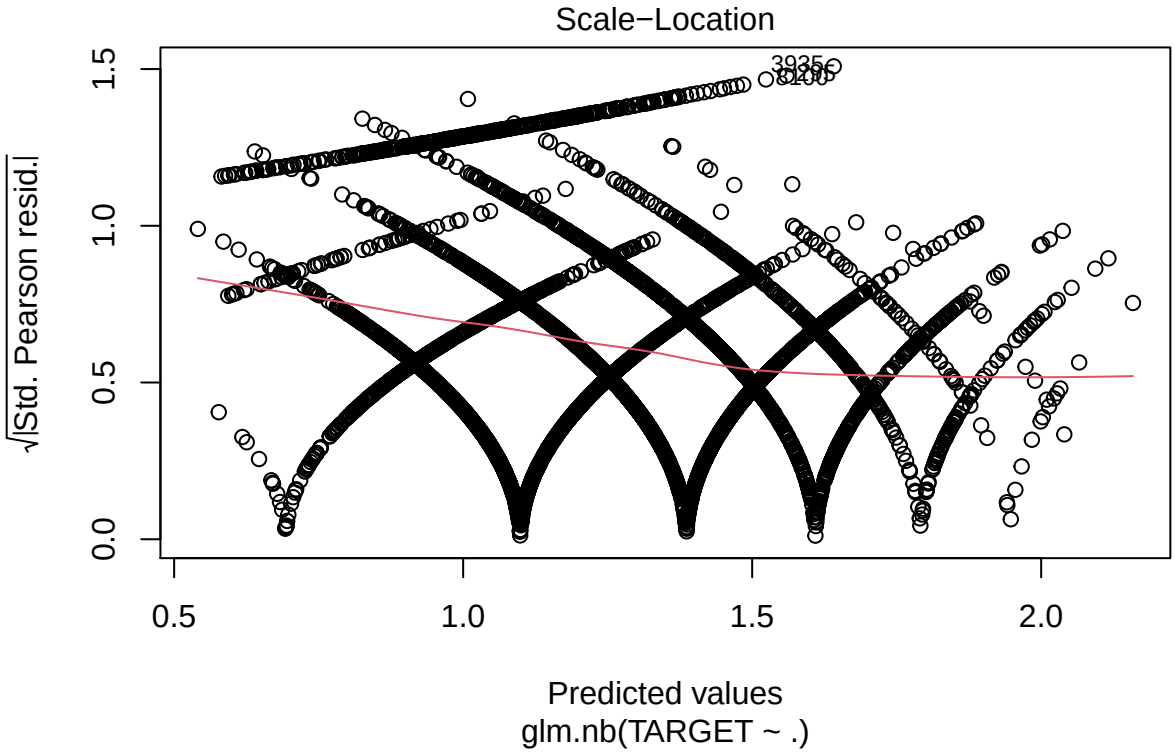
Model 5 : Negative Binomial Model without imputations

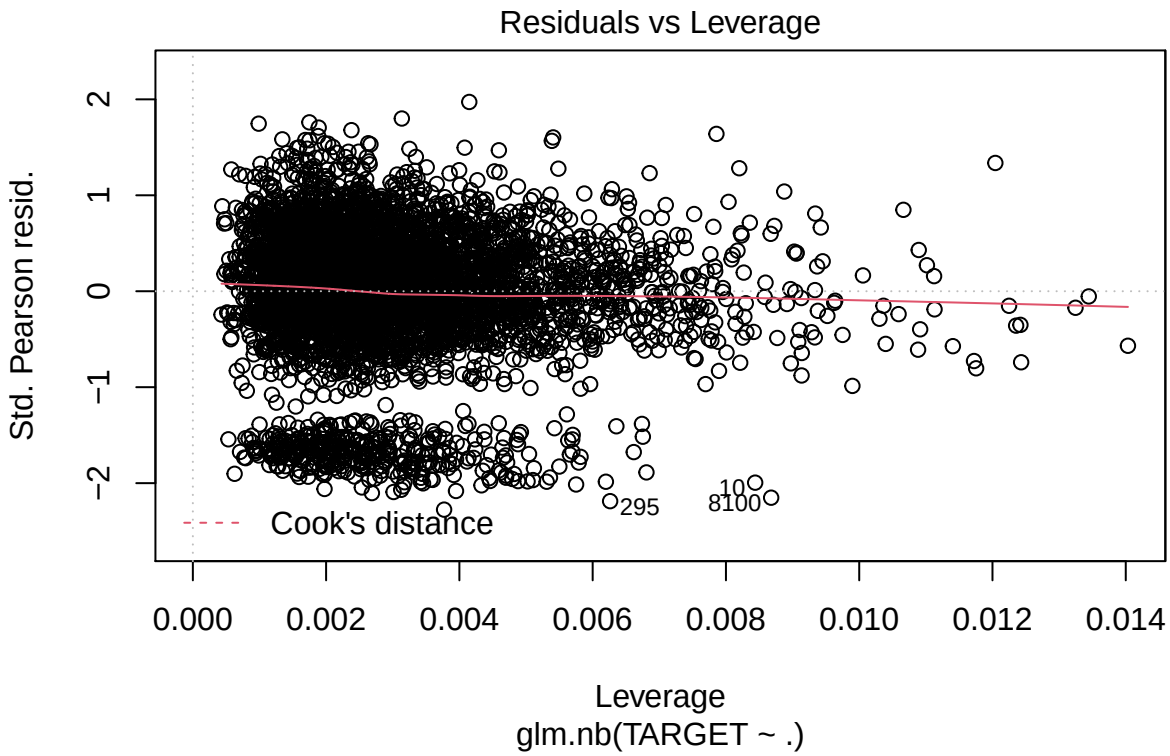
```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = wine_train1, init.theta = 138898.9107,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2127  -0.2757   0.0647   0.3766   1.6981
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.608e+00  2.796e-01   5.750 8.91e-09 ***
## FixedAcidity         6.705e-04  1.177e-03   0.570  0.56900
## VolatileAcidity     -2.750e-02  9.283e-03  -2.963  0.00305 **
## CitricAcid          -3.835e-03  8.519e-03  -0.450  0.65259
## ResidualSugar        1.828e-05  2.152e-04   0.085  0.93231
## Chlorides           -3.764e-02  2.314e-02  -1.627  0.10378
## FreeSulfurDioxide    5.671e-05  4.892e-05   1.159  0.24630
## TotalSulfurDioxide   2.230e-05  3.177e-05   0.702  0.48275
```

```
## Density            -4.025e-01  2.750e-01  -1.464  0.14326
## pH                  2.307e-04  1.085e-02   0.021  0.98303
## Sulphates          -5.984e-03  7.973e-03  -0.751  0.45293
## Alcohol             3.262e-03  2.004e-03   1.628  0.10360
## LabelAppeal         1.730e-01  8.858e-03  19.529  < 2e-16 ***
## AcidIndex          -4.967e-02  6.666e-03  -7.451 9.28e-14 ***
## STARS               1.929e-01  8.328e-03  23.160  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(138898.9) family taken to be 1)
##
##     Null deviance: 4720.4  on 5143  degrees of freedom
## Residual deviance: 3242.7  on 5129  degrees of freedom
##   (5093 observations deleted due to missingness)
## AIC: 18547
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  138899
##           Std. Err.:  259921
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -18515.07
```




Normal Q−Q
Std. Pearson resid.
−2
−1
0
1
2
−4
−2
0
2
4
Theoretical Quantiles
glm.nb(TARGET ~ .)

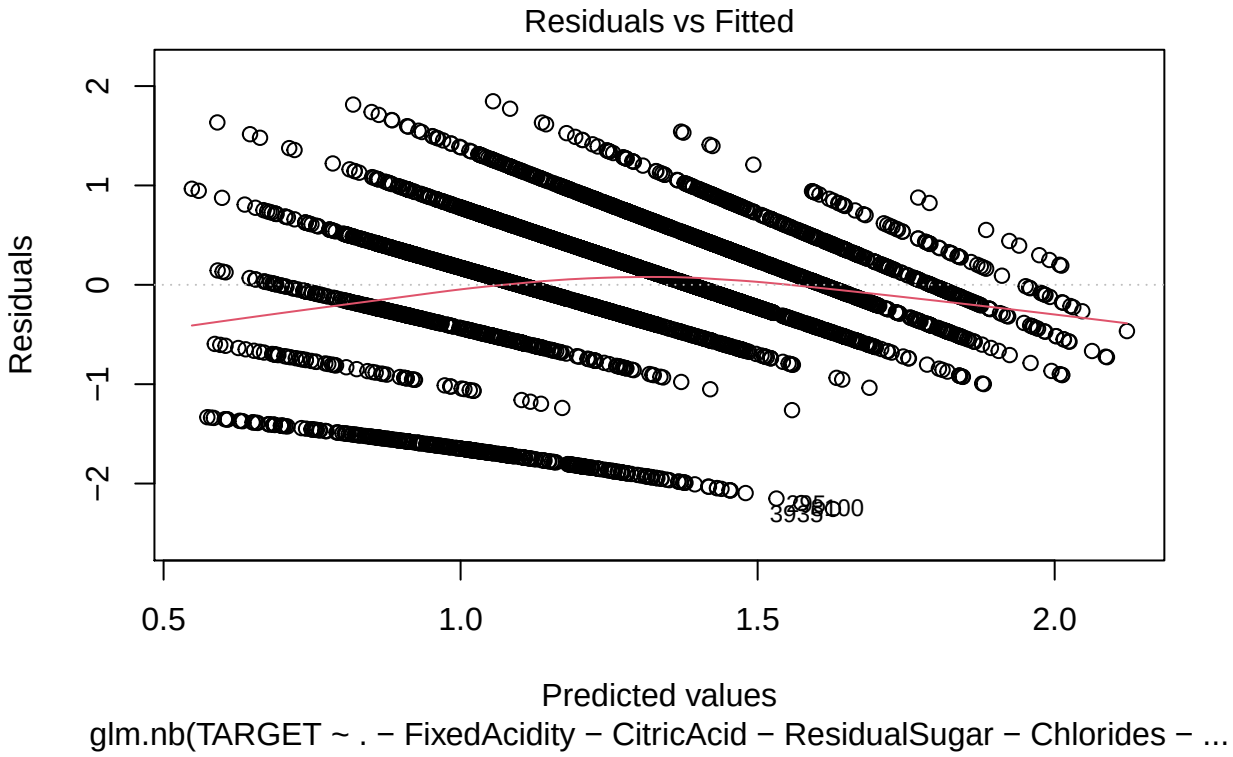

Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
Predicted values
glm.nb(TARGET ~ .)

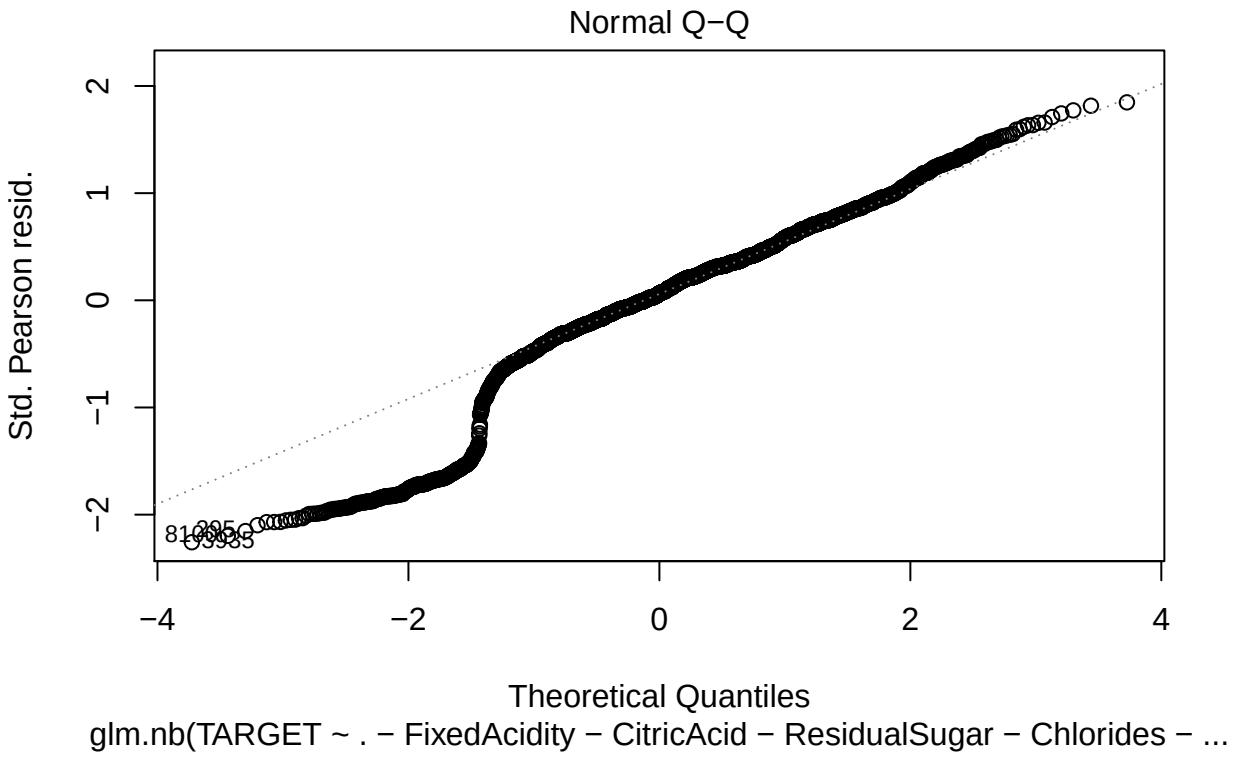


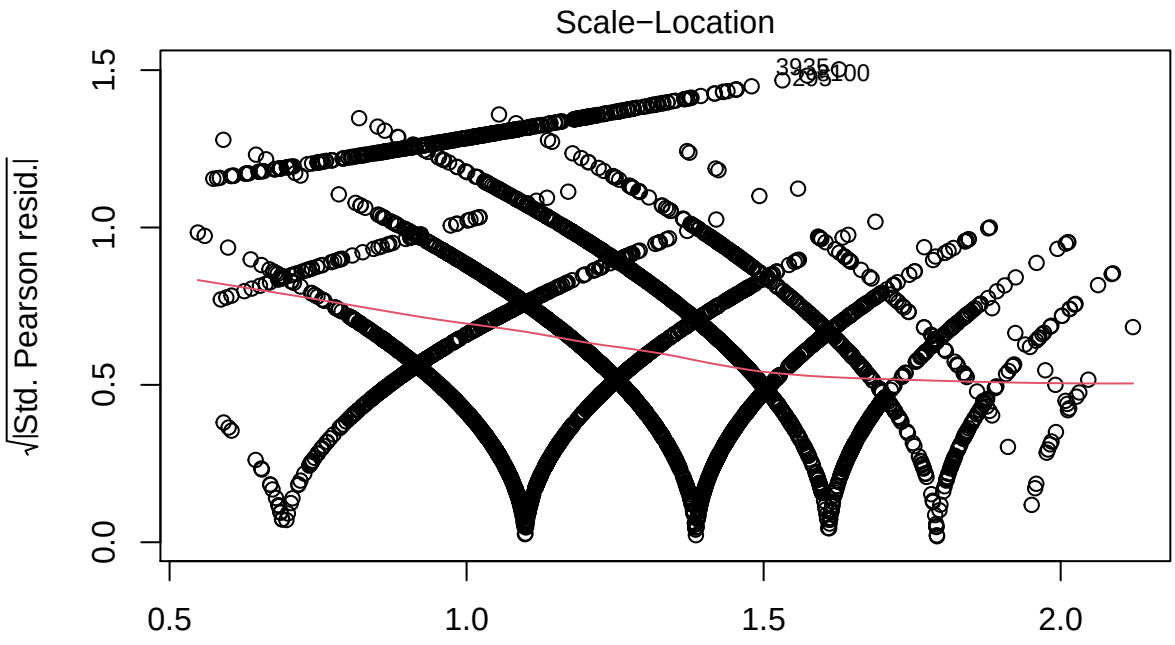
AIC is similar to Poisson without imputations, as are some of the coefficients. ### Model 6 : Negative Binomial Model without imputations and only significant variables

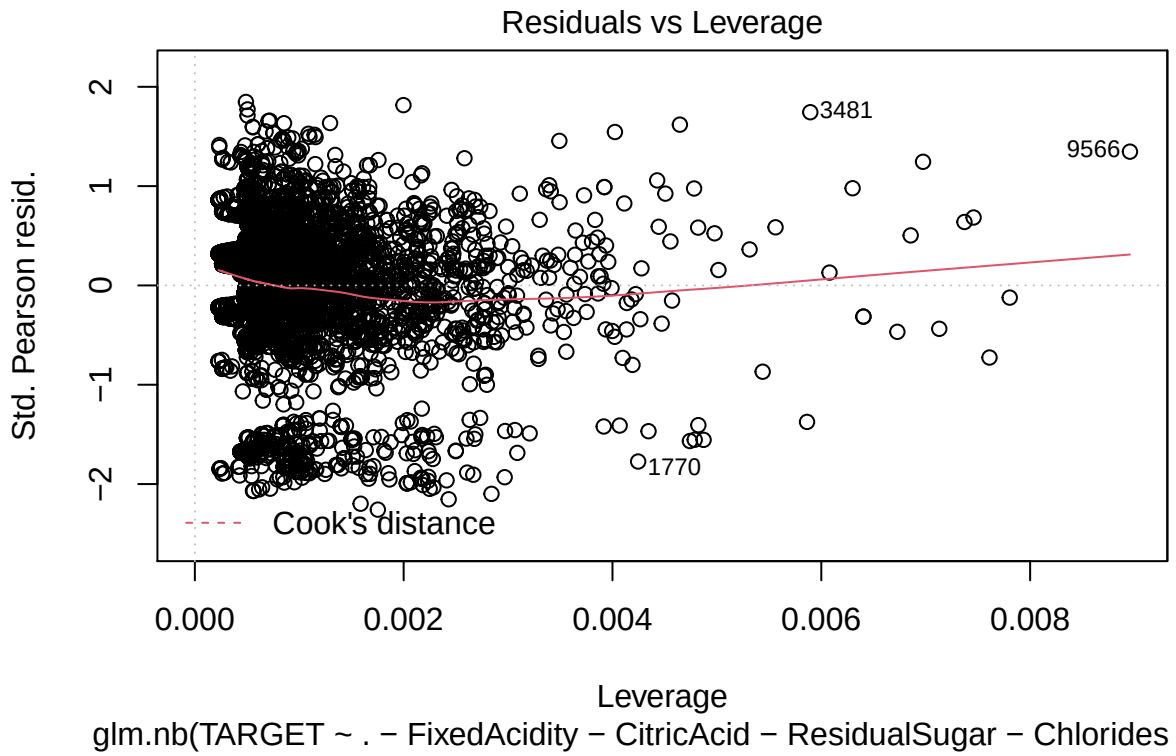
```
##
## Call:
## glm.nb(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##     Chlorides - FreeSulfurDioxide - TotalSulfurDioxide - Density -
##     pH - Sulphates - Alcohol, data = wine_train1, init.theta = 138402.5261,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1898  -0.2777   0.0622   0.3764   1.6086
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.251443   0.054725  22.868  < 2e-16 ***
## VolatileAcidity -0.027581   0.009279  -2.973  0.00295 **
## LabelAppeal      0.173177   0.008853  19.562  < 2e-16 ***
## AcidIndex       -0.050616   0.006553  -7.724 1.13e-14 ***
## STARS            0.194209   0.008292  23.421  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(138402.5) family taken to be 1)
##
##     Null deviance: 4720.4  on 5143  degrees of freedom
```

```
## Residual deviance: 3253.0  on 5139  degrees of freedom
##   (5093 observations deleted due to missingness)
## AIC: 18537
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  138403
##           Std. Err.:  258834
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -18525.37
```




Normal Q−Q
Std. Pearson resid.
2
1
0
−1
−2
−4
−2
0
2
4
Theoretical Quantiles
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Chlorides − ...

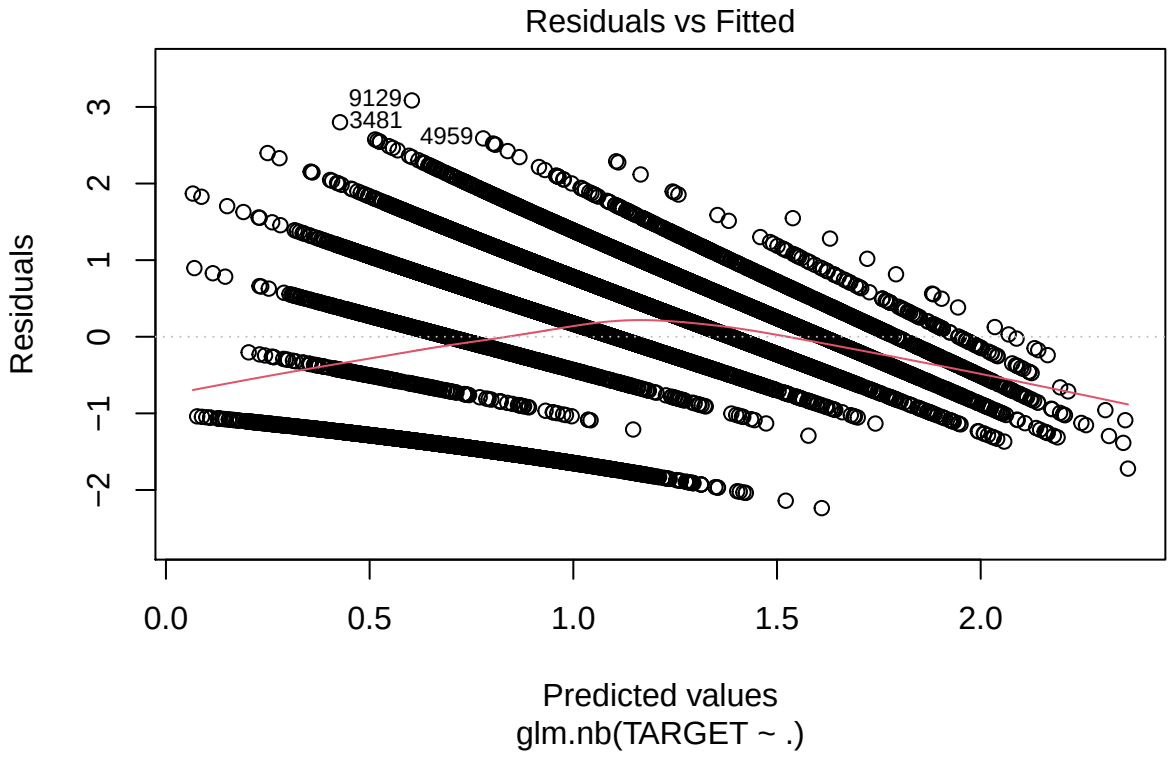

Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
Predicted values
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Chlorides − ...

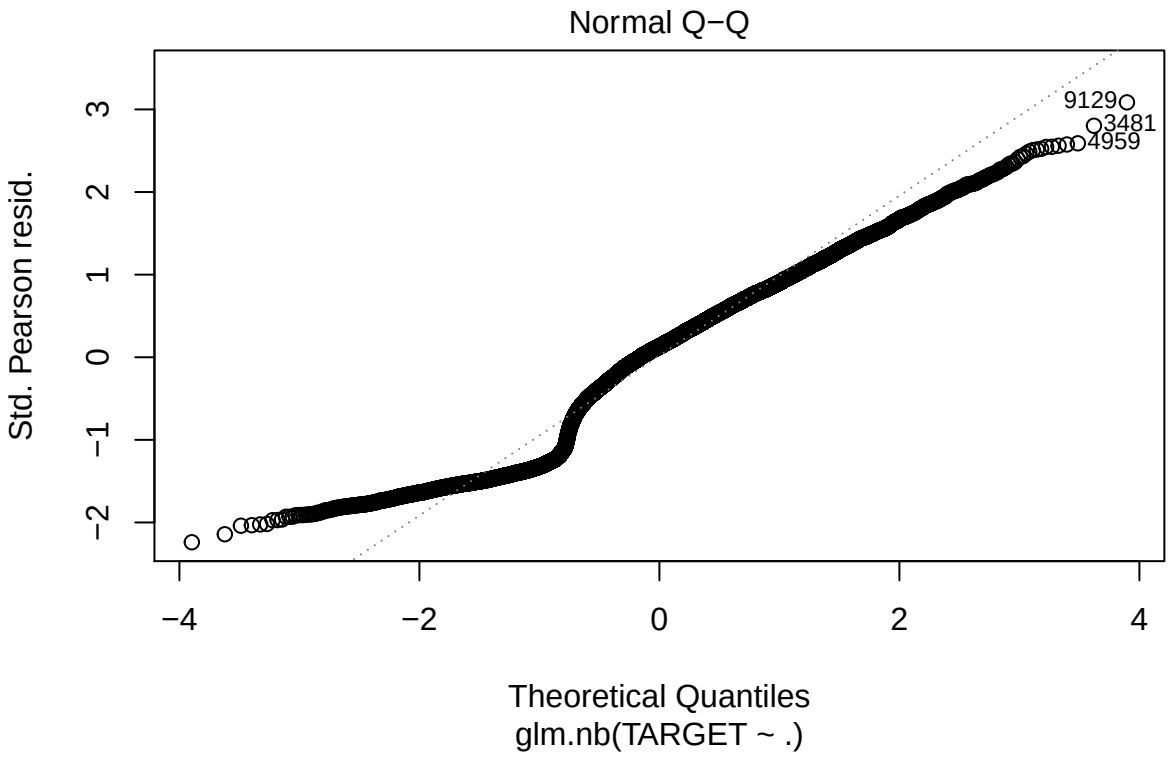


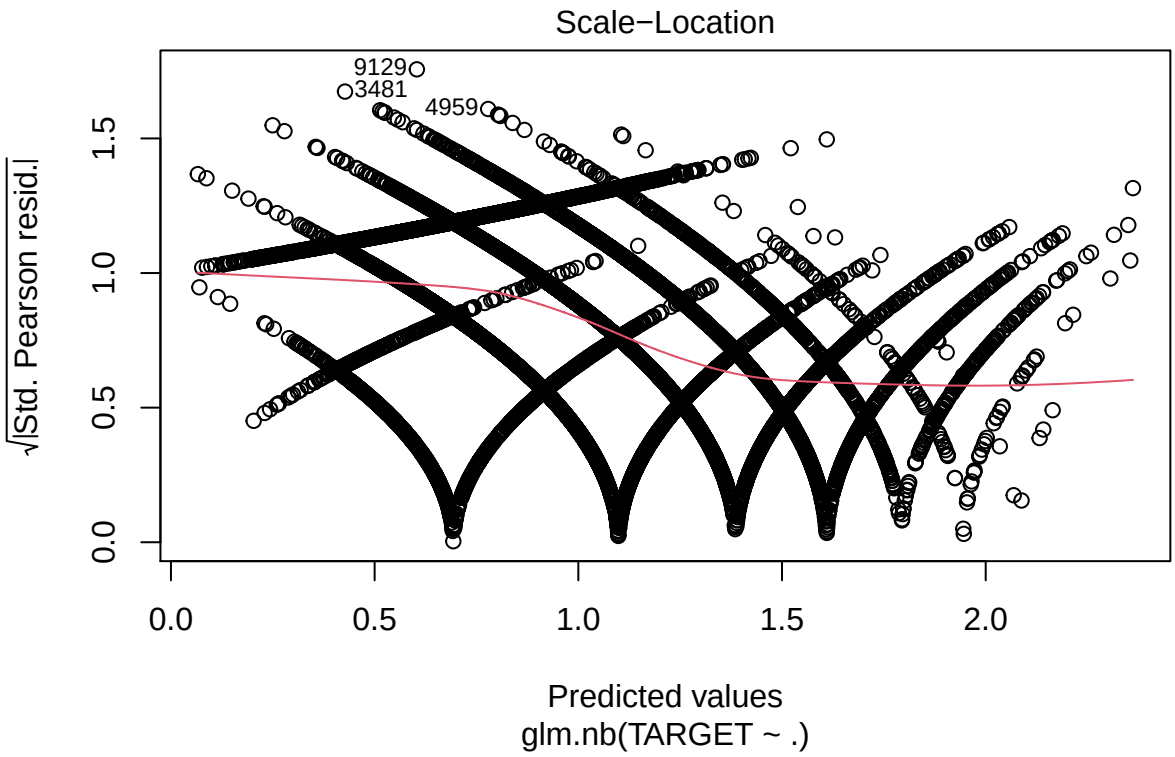
Removing non significant variables doesn't seem to affect the model's accuracy much. ### Model 7 : Negative Binomial Model with imputations

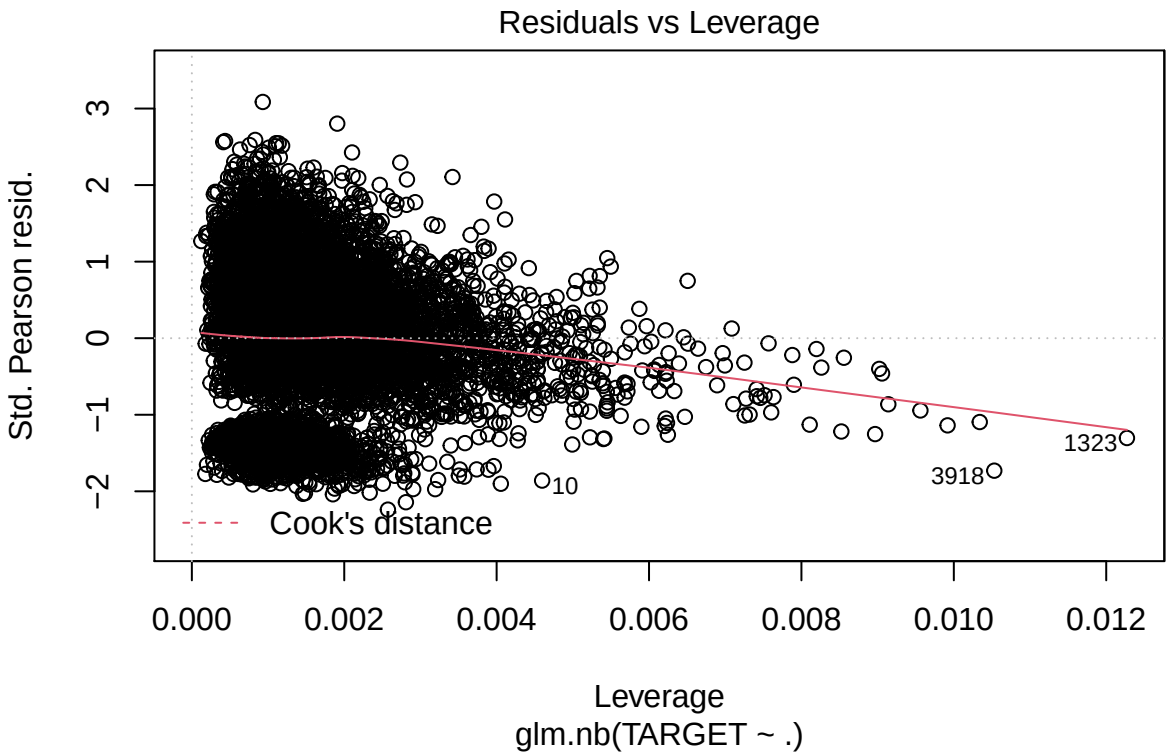
```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = wine_train2, init.theta = 49078.50992,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1629  -0.6739   0.1305   0.6337   2.4320
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.337e+00  2.281e-01  10.242  < 2e-16 ***
## FixedAcidity         2.250e-04  9.190e-04   0.245 0.806608
## VolatileAcidity     -4.313e-02  7.286e-03  -5.919 3.23e-09 ***
## CitricAcid           8.534e-03  6.573e-03   1.298 0.194177
## ResidualSugar        1.271e-04  1.675e-04   0.759 0.448021
## Chlorides           -6.573e-02  1.790e-02  -3.673 0.000240 ***
## FreeSulfurDioxide    1.336e-04  3.804e-05   3.512 0.000444 ***
## TotalSulfurDioxide   9.235e-05  2.460e-05   3.754 0.000174 ***
## Density             -3.404e-01  2.144e-01  -1.588 0.112389
## pH                  -1.962e-02  8.418e-03  -2.331 0.019745 *
## Sulphates           -1.569e-02  6.157e-03  -2.549 0.010806 *
## Alcohol              2.951e-03  1.554e-03   1.898 0.057642 .
## LabelAppeal          1.409e-01  6.798e-03  20.723  < 2e-16 ***
```

```
## AcidIndex          -7.709e-01  3.999e-02 -19.279  < 2e-16 ***
## STARS               3.407e-01  6.270e-03  54.335  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(49078.51) family taken to be 1)
##
##     Null deviance: 18290  on 10236  degrees of freedom
## Residual deviance: 12828  on 10222  degrees of freedom
## AIC: 38419
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  49079
##           Std. Err.:  63619
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -38387.04
```




Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm.nb(TARGET ~ .)
9129
3481
4959
−2
−1
0
1
2
3
−4
−2
0
2
4

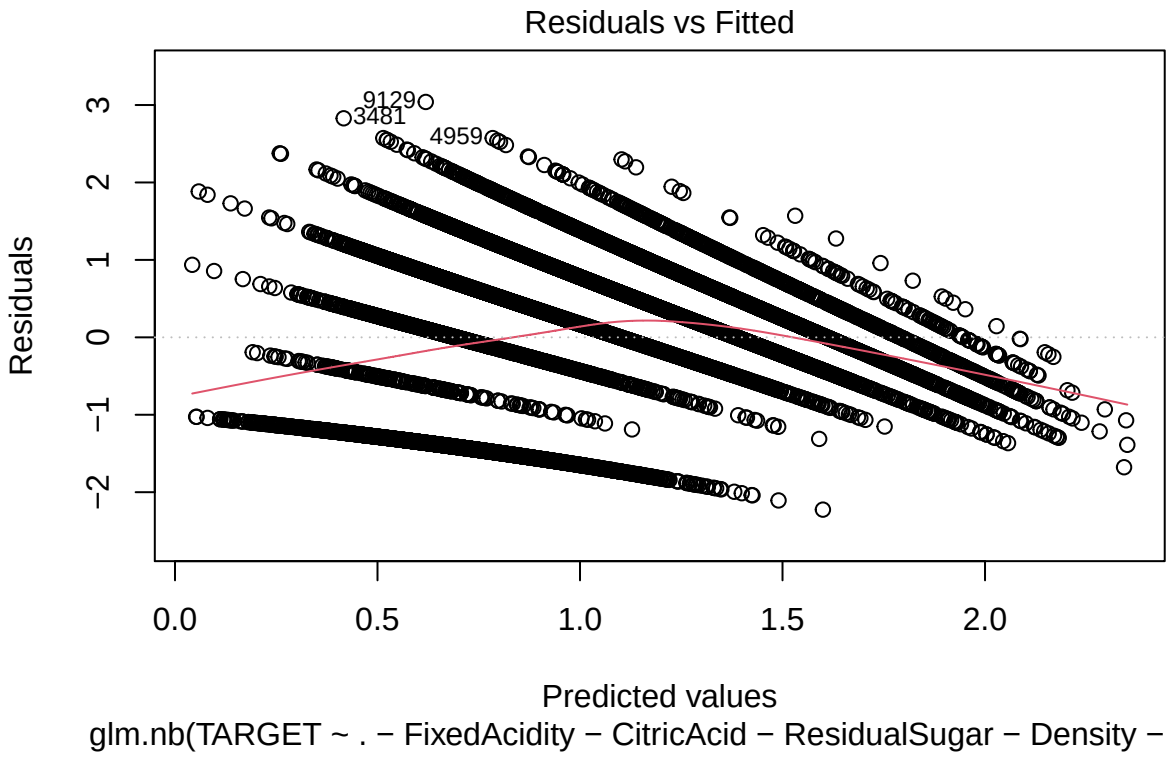

Scale−Location
9129
3481
4959
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
Predicted values
glm.nb(TARGET ~ .)



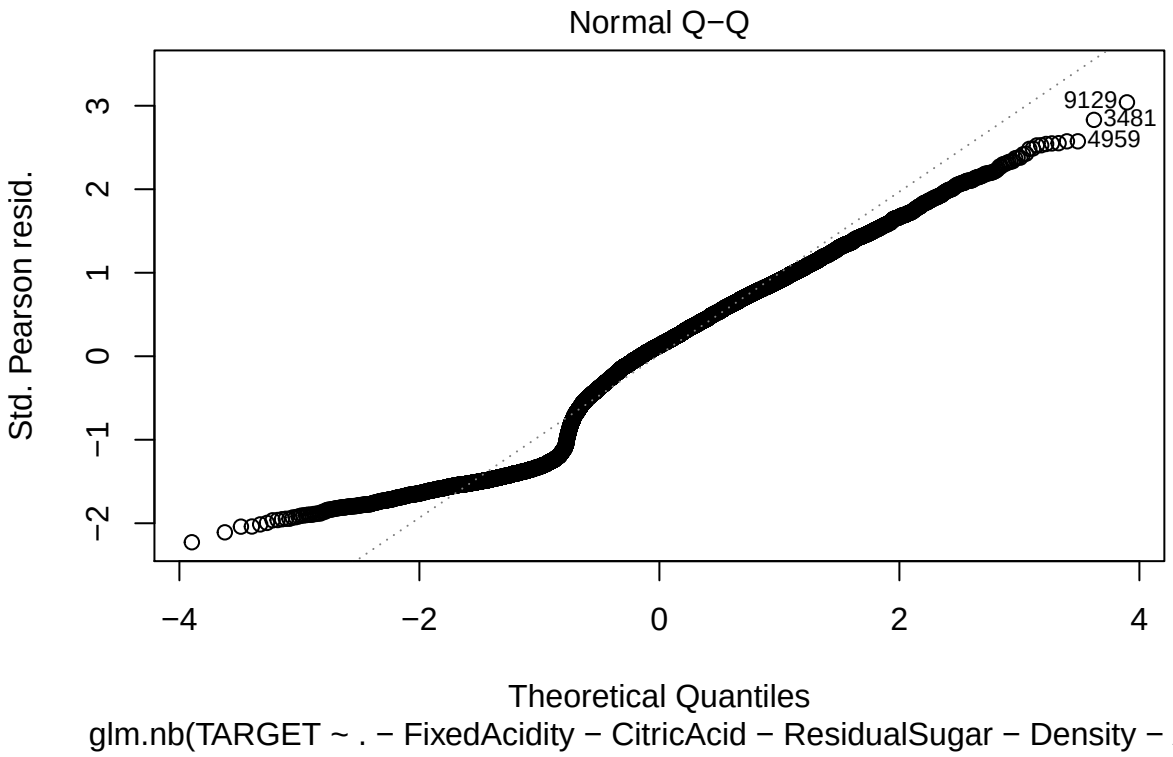
With imputations we find surprisingly similar results to a the Poisson model with imputations. ### Model 8 : Negative Binomial Model with imputations and only significant variables

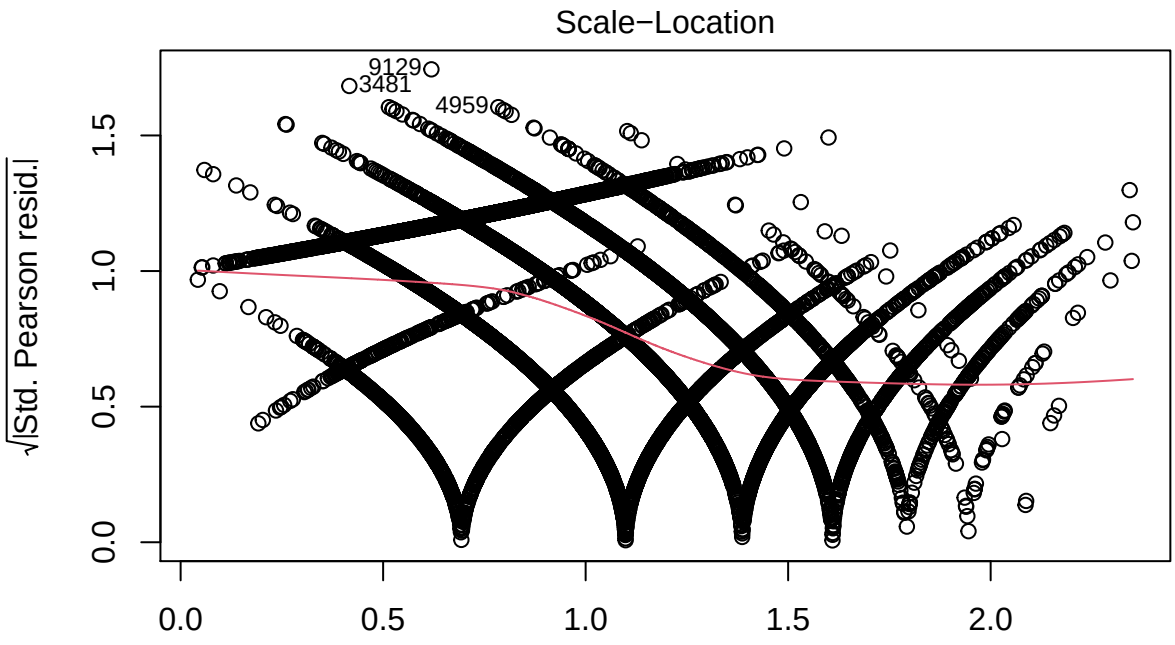
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.038	0.0884	23.05	1.412e-117
VolatileAcidity	-0.04348	0.007284	-5.969	2.391e-09
Chlorides	-0.06726	0.01789	-3.76	0.0001702
FreeSulfurDioxide	0.0001316	3.801e-05	3.461	0.0005374
TotalSulfurDioxide	9.15e-05	2.458e-05	3.723	0.0001969
pH	-0.01991	0.008415	-2.366	0.018
Sulphates	-0.01563	0.006153	-2.54	0.01109
LabelAppeal	0.1409	0.006798	20.73	2.021e-95
AcidIndex	-0.773	0.03936	-19.64	7.593e-86
STARS	0.3417	0.006255	54.63	0

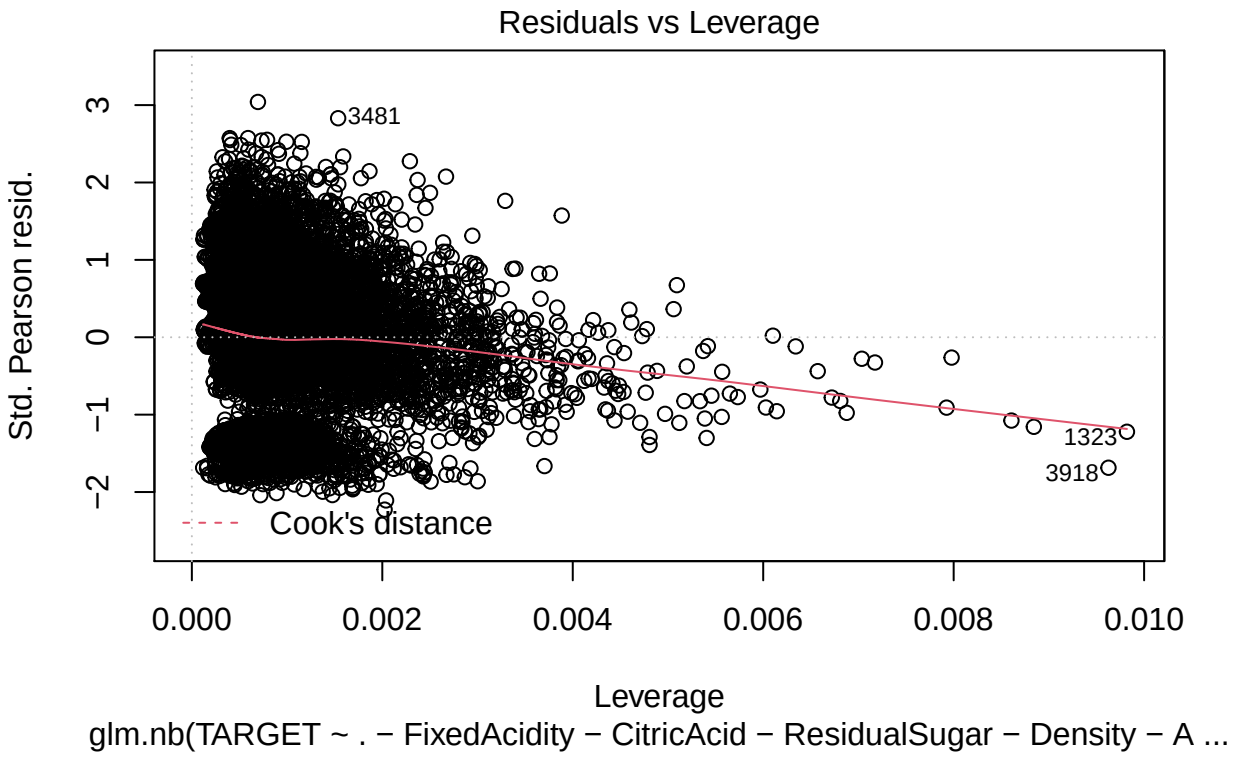
Quitting from lines 352-355 (Data621_Hw5.Rmd) Error in data.frame(Observations = length(x*residuals*), '*ResidualStd.Erro* *x*sigma, : arguments imply differing number of rows: 1, 0 Calls: ... pandoc.table -> cat -> pandoc.table.return -> data.frame In addition: There were 18 warnings (use warnings() to see


Residuals vs Fitted
9129
3481
4959
Residuals
3
2
1
0
−1
−2
0.0
0.5
1.0
1.5
2.0
Predicted values
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − A ...

them)


Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − A ...
9129
3481
4959
−2
−1
0
1
2
3
−4
−2
0
2
4


Scale−Location
9129
3481
4959
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
Predicted values
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − A ...



Linear Model

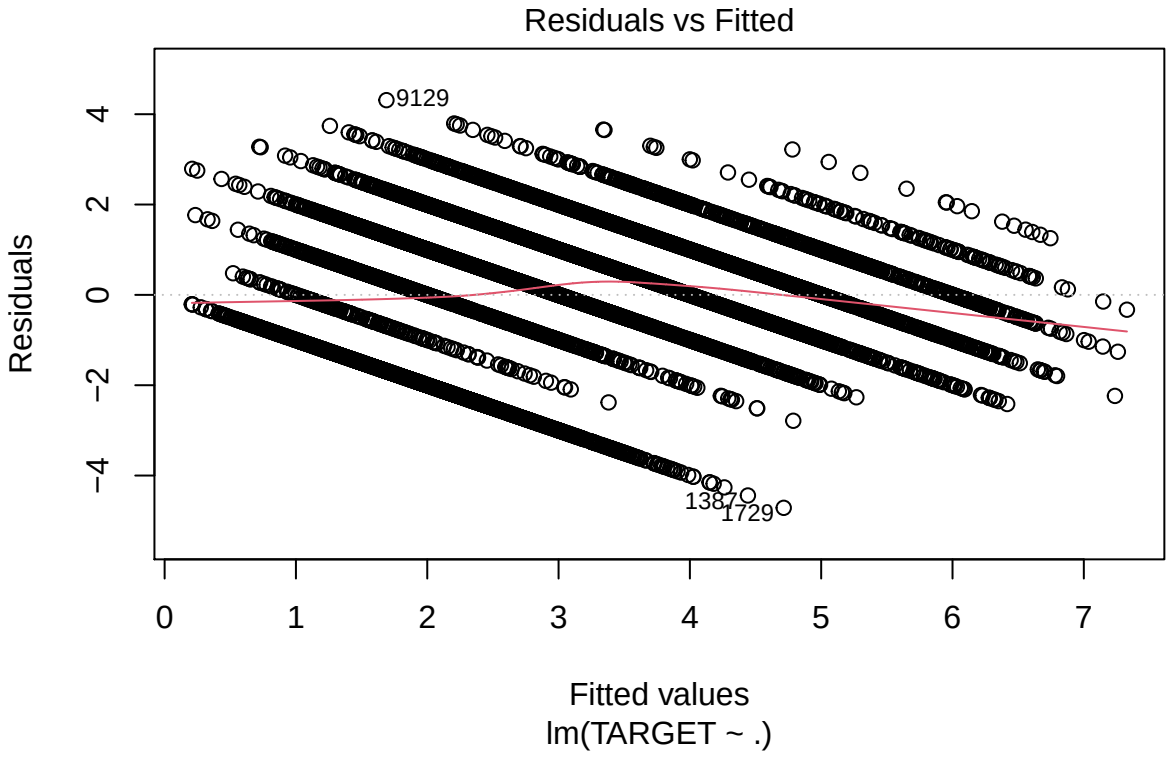
Model 9 : Linear Model with imputations

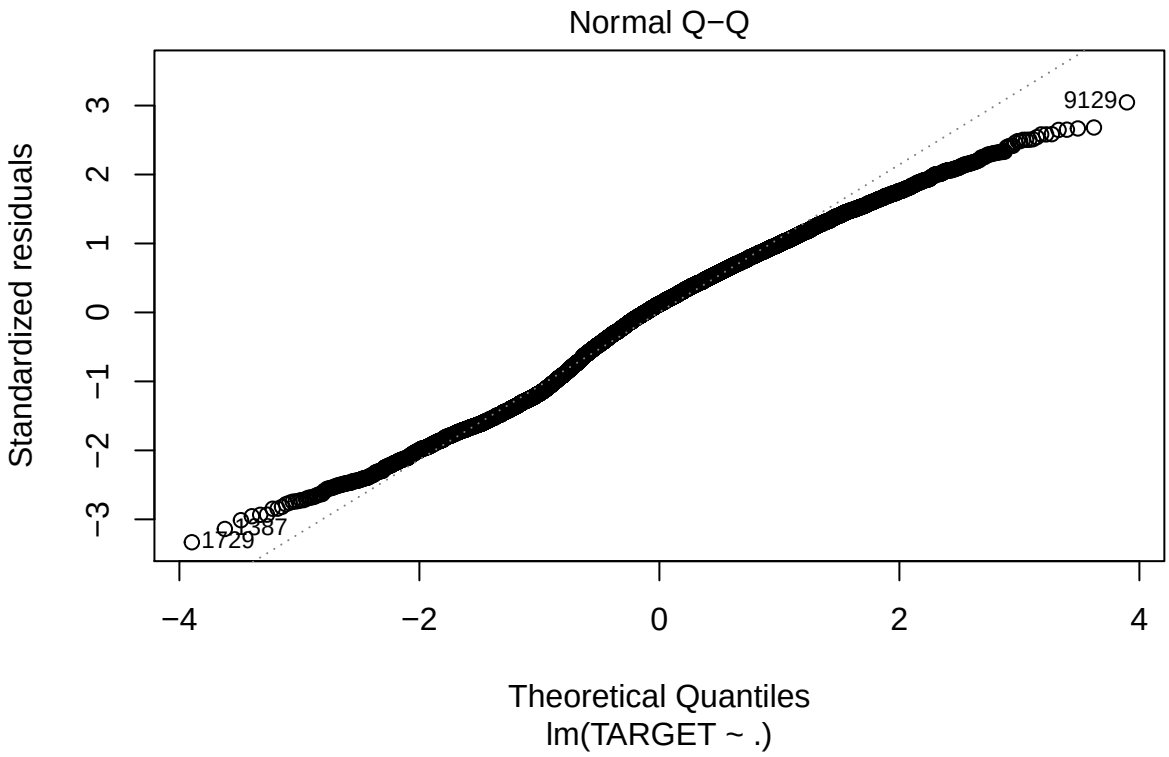
Use imputed training data on Linear regression model

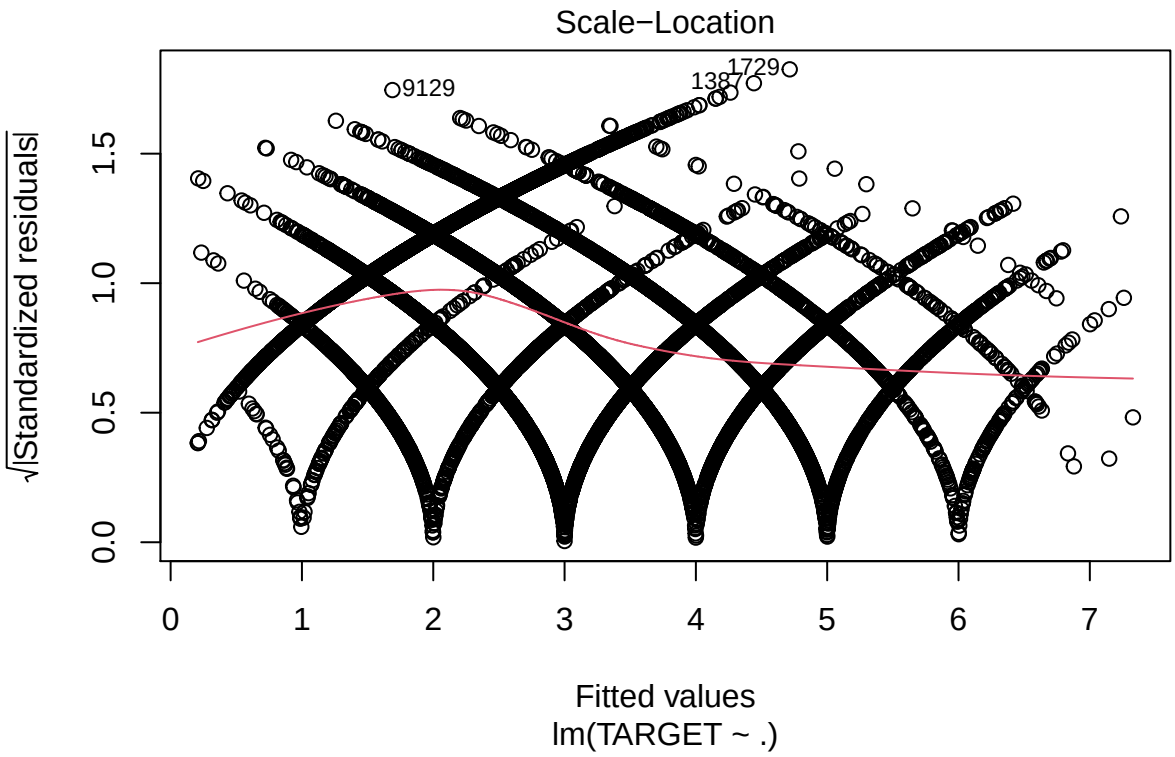
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.968	0.5567	10.72	1.149e-26
FixedAcidity	0.001297	0.002253	0.5756	0.5649
VolatileAcidity	-0.1269	0.01791	-7.085	1.477e-12
CitricAcid	0.02625	0.01629	1.611	0.1071
ResidualSugar	0.0004231	0.0004132	1.024	0.3059
Chlorides	-0.2023	0.04391	-4.606	4.149e-06
FreeSulfurDioxide	0.0003635	9.387e-05	3.873	0.0001082
TotalSulfurDioxide	0.0002432	6.023e-05	4.038	5.425e-05
Density	-0.8659	0.526	-1.646	0.09974
pH	-0.0473	0.02072	-2.283	0.02242
Sulphates	-0.04212	0.01512	-2.786	0.005346
Alcohol	0.01251	0.003807	3.285	0.001025
LabelAppeal	0.4311	0.01646	26.19	2.115e-146
AcidIndex	-2.068	0.09237	-22.39	1.875e-108
STARS	1.167	0.01671	69.8	0

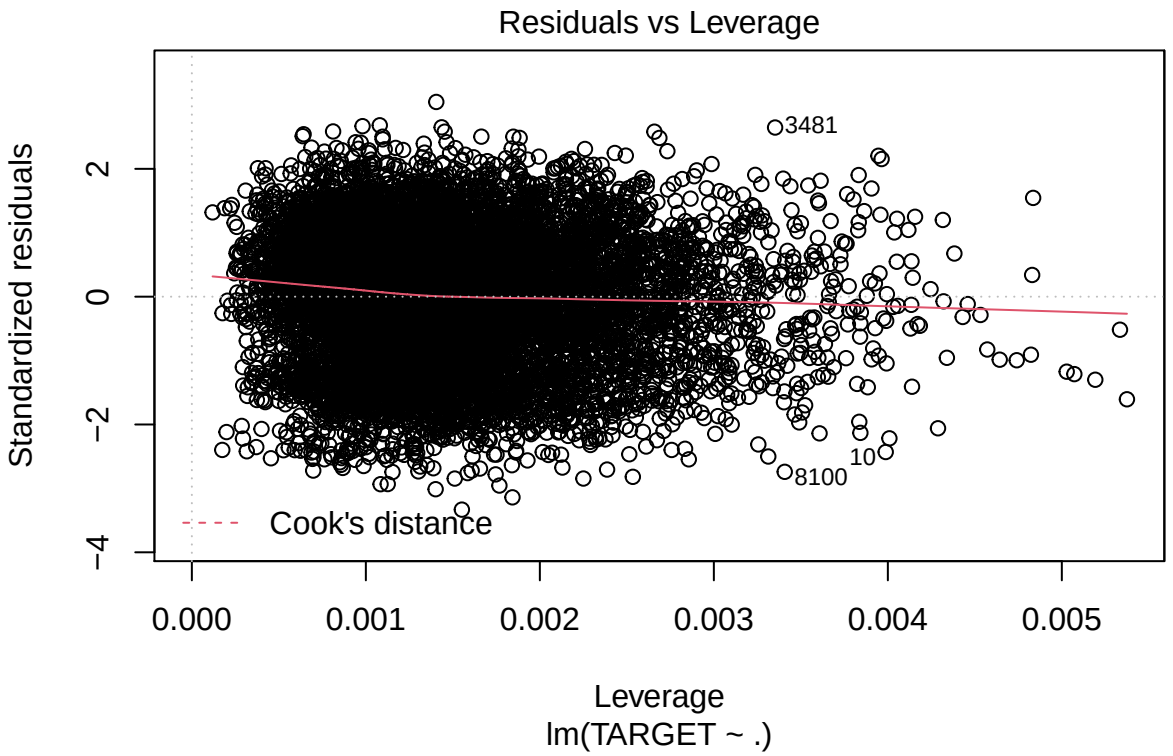
Table 20: Fitting linear model: TARGET ~ .

Observations	Residual Std. Error	R^2	Adjusted R^2
10237	1.416	0.4605	0.4598




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TARGET ~ .)
9129
1729
−4
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
9129
13857
17729
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
1
2
3
4
5
6
7
Fitted values
lm(TARGET ~ .)



A plain linear regression model is a surprisingly decent performer when coupled with imputations.

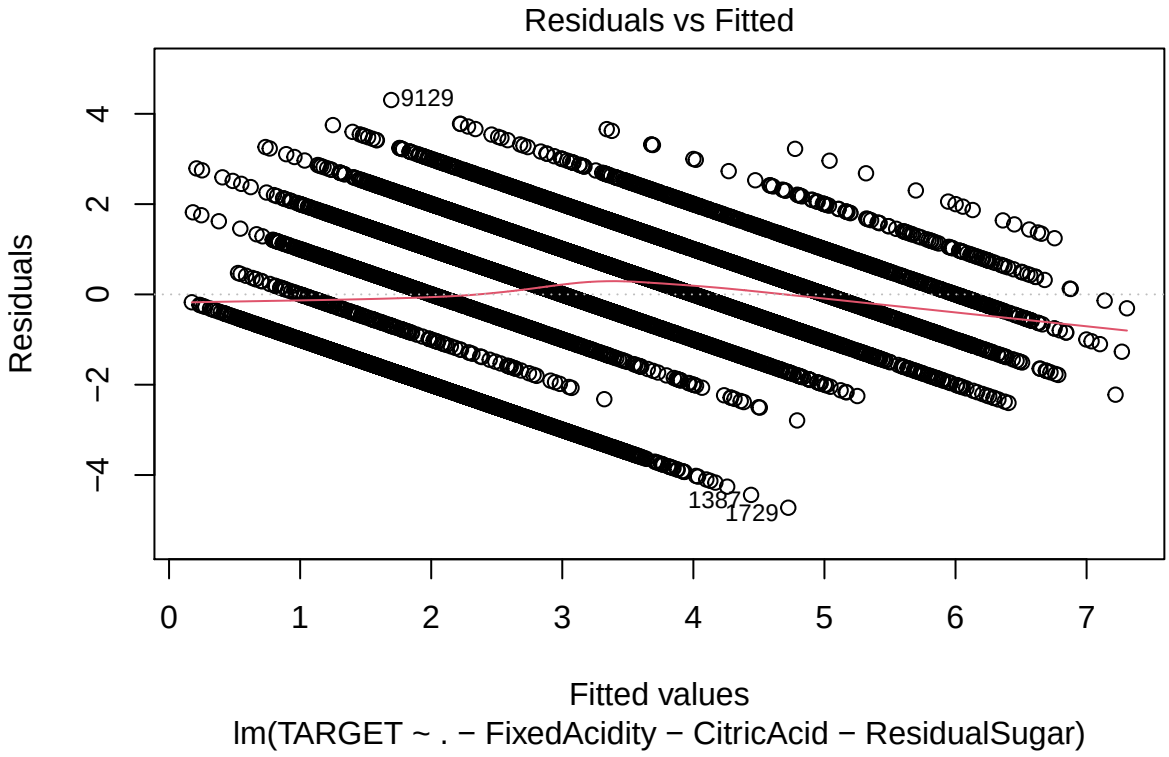
Model 10 : Linear Model with imputations and only significant variables.

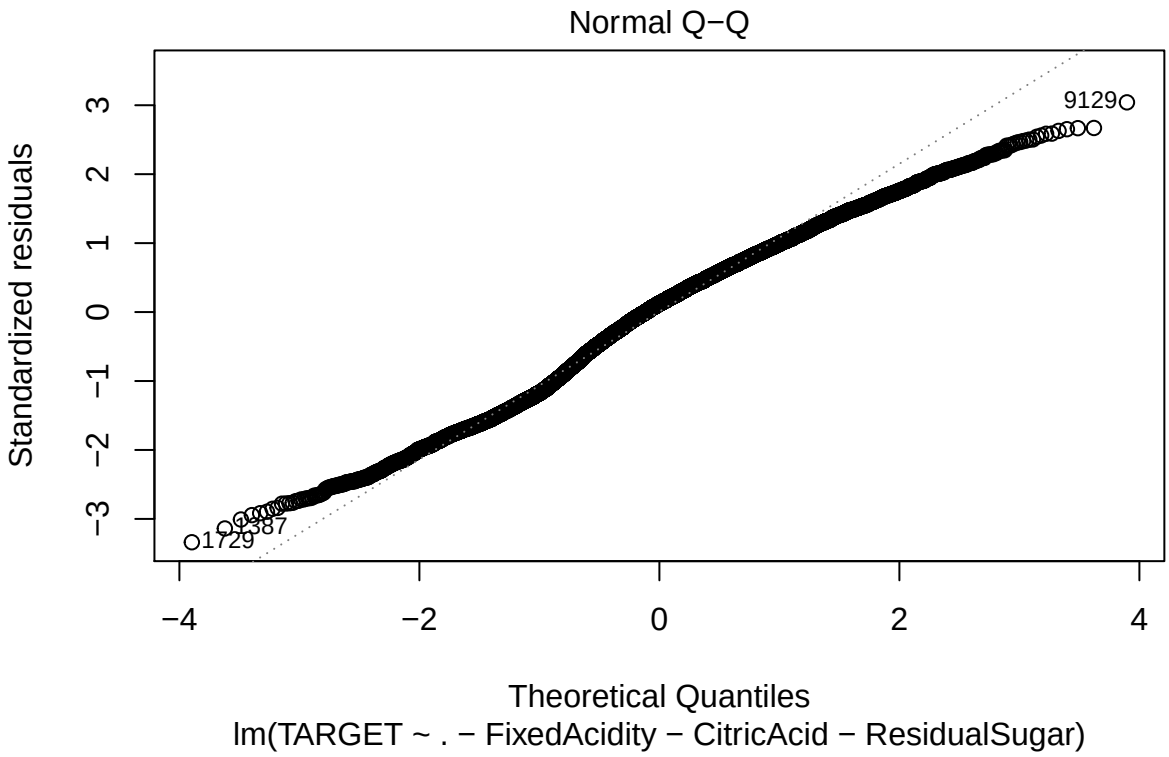
We got `FixedAcidity`, `CitricAcid` and `ResidualSugar` as significant variables and use same variables on Linear regression model with imputed training data.

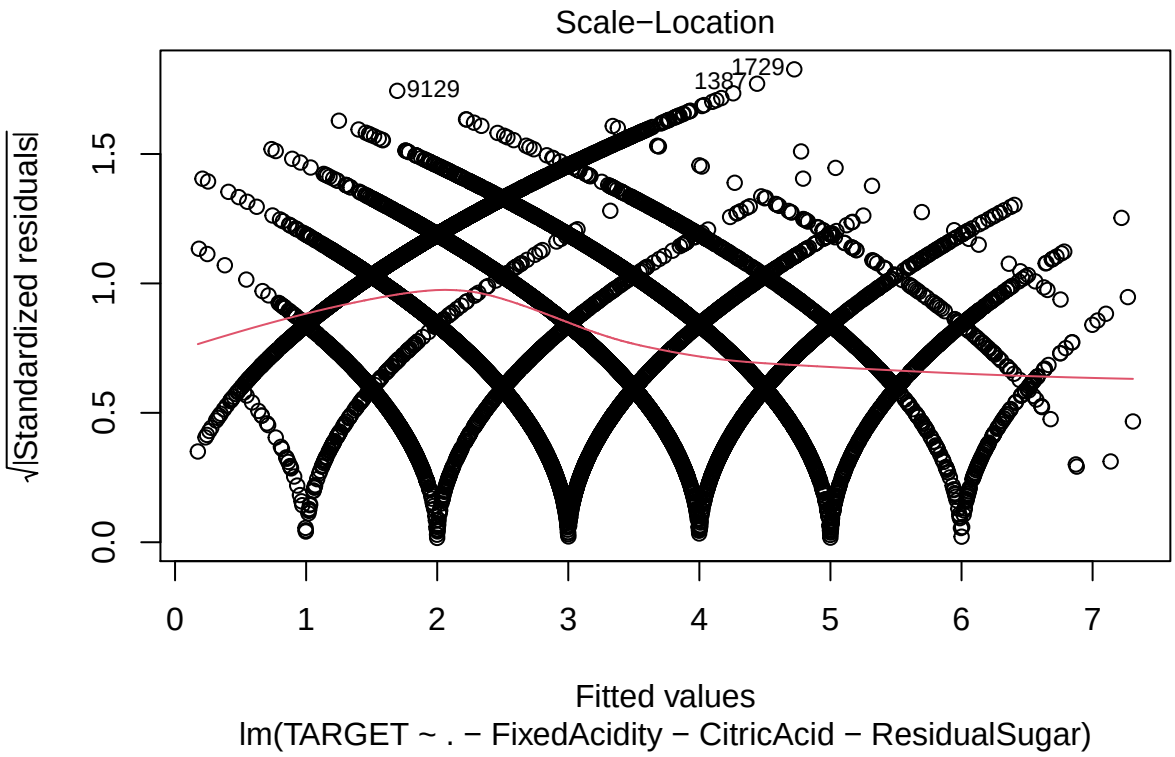
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.951	0.5566	10.69	1.541e-26
VolatileAcidity	-0.1278	0.01791	-7.138	1.015e-12
Chlorides	-0.2032	0.04391	-4.627	3.748e-06
FreeSulfurDioxide	0.000366	9.386e-05	3.899	9.711e-05
TotalSulfurDioxide	0.0002454	6.021e-05	4.075	4.635e-05
Density	-0.8712	0.526	-1.656	0.09768
pH	-0.04728	0.02071	-2.282	0.02249
Sulphates	-0.04236	0.01511	-2.803	0.005066
Alcohol	0.01249	0.003807	3.281	0.001036
LabelAppeal	0.431	0.01646	26.19	2.4e-146
AcidIndex	-2.048	0.09074	-22.57	3.931e-110
STARS	1.167	0.01671	69.83	0

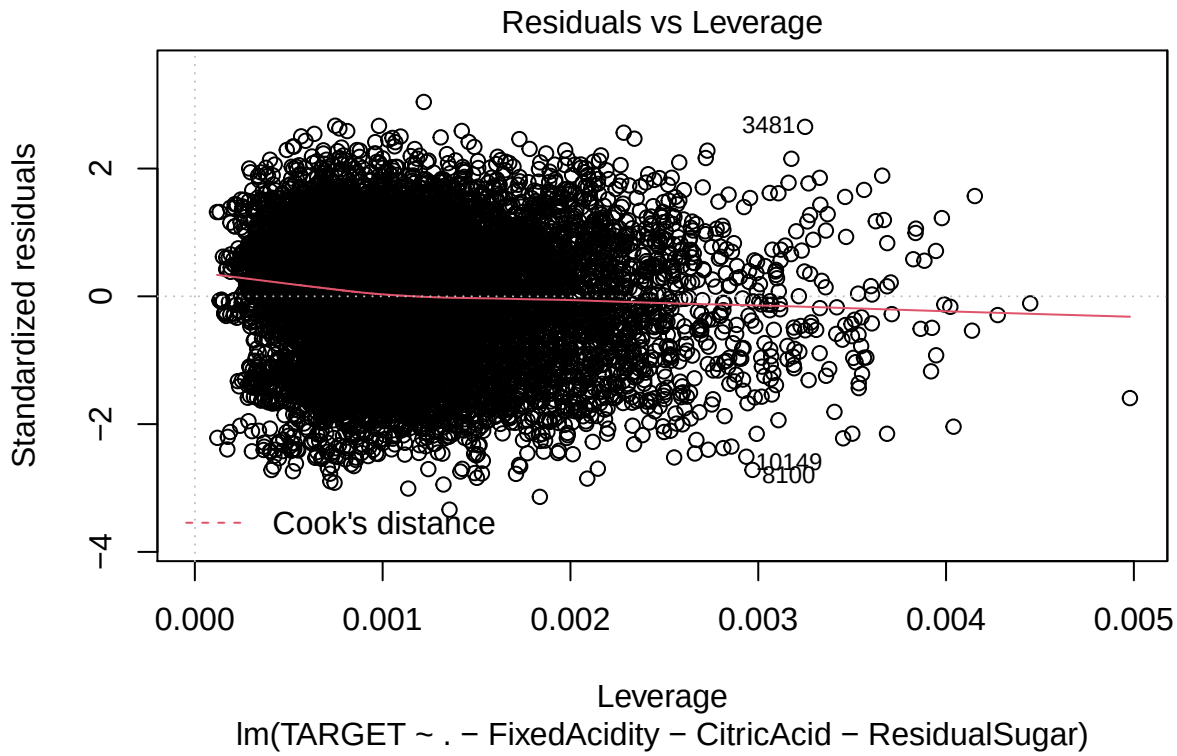
Table 22: Fitting linear model: TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar

Observations	Residual Std. Error	R^2	Adjusted R^2
10237	1.416	0.4603	0.4598




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar)
9129
1729
−4
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
9129
1387
1729
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
1
2
3
4
5
6
7
Fitted values
lm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar)



Again, removing less significant variables has little impact on the model, and is recommended to reduce overfitting.

Ordinal Logistic Regression

Since Ordinal logistic regression uses ordered factors we might find this as one of the top model based on our use cases.

Call: polr(formula = TARGET ~ ., data = polrDF, Hess = TRUE)

Table 23: Coeficients

	Value	Std. Error	t value
FixedAcidity	0.002182	0.002905	0.751
VolatileAcidity	-0.1556	0.02328	-6.685
CitricAcid	0.02897	0.02112	1.372
ResidualSugar	0.0003196	0.000532	0.6008
Chlorides	-0.2627	0.05667	-4.637
FreeSulfurDioxide	0.0004607	0.0001216	3.788
TotalSulfurDioxide	0.0002716	7.83e-05	3.469
Density	-1.298	0.1491	-8.707
pH	-0.03141	0.02681	-1.172
Sulphates	-0.03391	0.01967	-1.724
Alcohol	0.02691	0.004897	5.495
LabelAppeal	0.8256	0.02377	34.73
AcidIndex	-2.665	0.1251	-21.3

	Value	Std. Error	t value
STARS	1.468	0.02567	57.21

Table 24: Intercepts

	Value	Std. Error	t value
0|1	-5.921	0.1357	-43.64
1|2	-5.784	0.1355	-42.67
2|3	-5.181	0.1351	-38.35
3|4	-3.813	0.135	-28.26
4|5	-1.966	0.1372	-14.33
5|6	0.003405	0.1437	0.0237
6|7	2.207	0.1675	13.18
7|8	4.548	0.3034	14.99

Residual Deviance: 30016.23

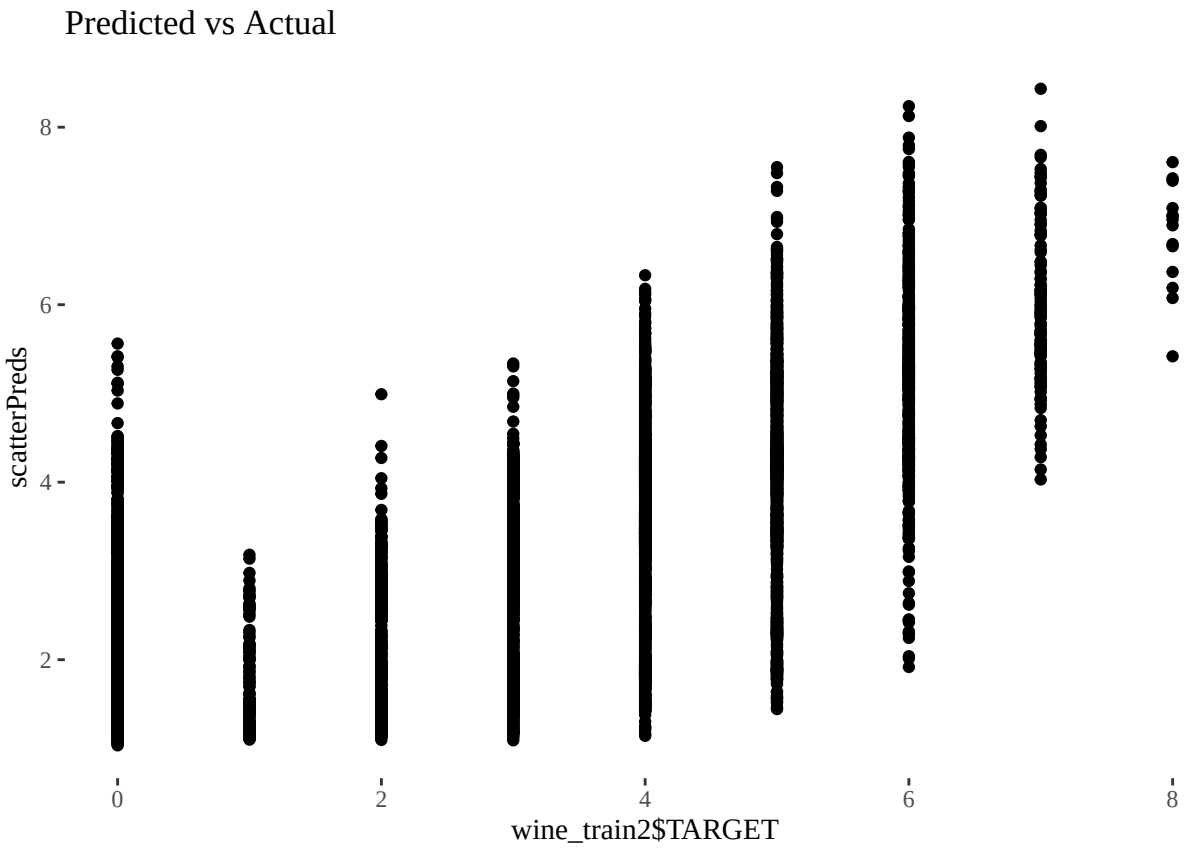
AIC: 30060.23

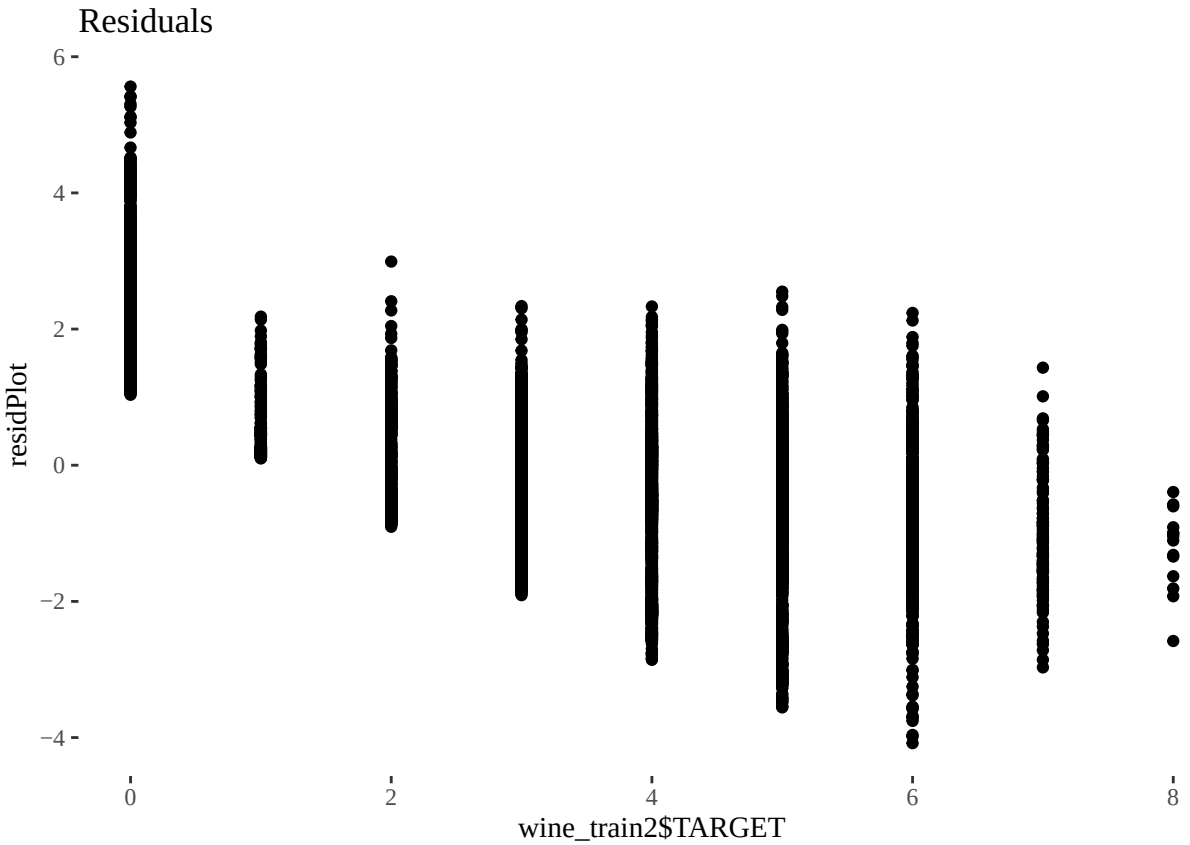
Zero inflation

Zero-inflated poisson regression is used to model count data that has an excess of zero counts. Further, theory suggests that the excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently. In Data exploration we saw many zero values, considering this we might get this as one of our best model.

```
##
## Call:
## zeroinfl(formula = TARGET ~ . | STARS, data = wine_train2, dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.09180 -0.49650  0.07134  0.48208  2.08565
##
## Count model coefficients (negbin with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.824e+00  2.385e-01   7.648 2.04e-14 ***
## FixedAcidity         4.523e-04  9.473e-04   0.477 0.633017
## VolatileAcidity     -1.936e-02  7.560e-03  -2.561 0.010429 *
## CitricAcid           2.116e-03  6.718e-03   0.315 0.752830
## ResidualSugar       -6.323e-05  1.722e-04  -0.367 0.713427
## Chlorides           -3.245e-02  1.851e-02  -1.754 0.079503 .
## FreeSulfurDioxide    4.932e-05  3.854e-05   1.280 0.200613
## TotalSulfurDioxide   4.760e-06  2.460e-05   0.194 0.846535
## Density             -2.990e-01  2.224e-01  -1.344 0.178840
## pH                  -1.910e-03  8.749e-03  -0.218 0.827223
## Sulphates           -4.586e-03  6.391e-03  -0.718 0.473029
## Alcohol              5.950e-03  1.591e-03   3.741 0.000184 ***
## LabelAppeal          2.240e-01  7.112e-03  31.491  < 2e-16 ***
## AcidIndex           -2.682e-01  4.492e-02  -5.971 2.36e-09 ***
```

```
## STARS               1.227e-01  6.997e-03  17.531  < 2e-16 ***
## Log(theta)          1.860e+01  2.758e+00   6.742 1.56e-11 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.31678    0.11092   20.89   <2e-16 ***
## STARS       -2.66899    0.09689  -27.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 119421399.3444
## Number of iterations in BFGS optimization: 22
## Log-likelihood: -1.725e+04 on 18 Df
```





Interestingly Sulfur Dioxides and Sulphates are not significant in this model, while Alcohol is.

SELECT MODELS

Compare Models based on MSE/AIC

	MSE	AIC
Model1	6.929787	18544.98
Model2	6.926722	18535.29
Model3	6.849005	38416.87
Model4	6.849920	38415.39
Model5	6.929788	18547.07
Model6	6.926723	18537.37
Model7	6.849001	38419.04
Model8	6.849916	38417.56
Model9	2.002207	NA
Model10	2.002977	NA
Model11	NA	30060.23
Model12	1.988813	NA

Similar MSE values are observed for Poisson and negative binomial models.

We can compare our zero inflated model using a Vuong test to a normal Poisson model.

```
vuong(model4,model12)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                   -34.62149 model2 > model1 < 2.22e-16
## AIC-corrected         -34.49684 model2 > model1 < 2.22e-16
## BIC-corrected         -34.04599 model2 > model1 < 2.22e-16
```

Model2, or our Zero Inflated model, would seem to be better than our non inflated model.

Compare Models by Loss

Use test data and check the output

In order to validate we will use squared loss and squared difference to select model (MSE) from predicting on selected training datasets. Smaller numbers would indicate a truer fit.

```
## # A tibble: 12 x 1
##    `Loss:`
##      <dbl>
##  1    5.47
##  2    5.46
##  3    6.83
##  4    6.83
##  5    5.47
##  6    5.46
##  7    6.83
##  8    6.83
##  9    2.03
## 10    2.03
## 11    3.68
## 12    2.00
```

Based on above results these are our observation

-> Linear model performed well. -> Poisson regression model and Negative binomial model did not performed as expected. -> We expected Ordinal logistic regression to be a better model but it did not performed well.

At this point we are concentrated more on square loss which tells us the accuracy of our model

Zero Poisson Inflation seems to be the most accurate model with least loss score, and had good results from a Vuong test.

If we consider all the factors like least loss, good MSE and AIC score we found 'Zero Inflated Poisson' as our best one.

Prediction on Evaluation Data

Here we use MICE just like how we used earlier for imputing and log transformation for AcidIndex.

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
```

Display the Predicted values

– Column specification ————————————————————————————- cols(TARGET = col_double(), FixedAcidity = col_double(), VolatileAcidity = col_double(), CitricAcid = col_double(), ResidualSugar = col_double(), Chlorides = col_double(), FreeSulfurDioxide = col_double(), TotalSulfur-Dioxide = col_double(), Density = col_double(), pH = col_double(), Sulphates = col_double(), Alcohol = col_double(), LabelAppeal = col_double(), AcidIndex = col_double(), STARS = col_double())

Table 25: Table continues below

TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar
3.015	5.4	-0.86	0.27	-10.7
3.622	12.4	0.385	-0.76	-19.7
1.738	7.2	1.75	0.17	-33
1.507	6.2	0.1	1.8	1
1.677	11.4	0.21	0.28	1.2
5.664	17.6	0.04	-1.15	1.4

Table 26: Table continues below

Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH
0.092	23	398	0.9853	5.02
1.169	-37	68	0.9905	3.37
0.065	9	76	1.046	4.61
-0.179	104	89	0.9888	3.2
0.038	70	53	1.029	2.54
0.535	-250	140	0.9503	3.06

Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
0.64	12.3	-1	1.792	2
1.09	16	0	1.792	2
0.68	8.55	0	2.079	1
2.11	12.3	-1	2.079	1
-0.07	4.8	0	2.303	1
-0.02	11.4	1	2.079	4

For TARGET: Number of Cases Purchased as Predicted

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    StdD    Skew    Kurt
##    1.03    1.91    3.30    3.24    4.17    8.27    1.38    0.51   -0.32
```

Predicted Evaluation data

https://github.com/vijay564/Data621/blob/main/Evaluation_Full_Data.csv

Appendix

https://github.com/vijay564/Data621/blob/main/Data621_Hw5.Rmd