DATA 621 Business Analytics and Data Mining
Homework #5 (Submitted by Group 1)

Calvin Wong, Sudhan Maharjan, Kevin Benson, Ravi Itwaru, Juanelle Marks

---

Contents

We have been given a dataset which we need to explore, analyze and model. It is a dataset containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high-end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

Data Exploration

After importing the dataset from the github, we came to a conclusion that some of the variables are not needed.

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 12795 obs. of  15 variables:
##  $ TARGET            : num  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num  -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
##  $ ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num  -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num  NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num  268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num  0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##  $ Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num  9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##  $ LabelAppeal       : num  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : num  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : num  2 3 3 1 2 NA NA 3 NA 4 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   INDEX = col_double(),
##   ..   TARGET = col_double(),
##   ..   FixedAcidity = col_double(),
##   ..   VolatileAcidity = col_double(),
```

After removing the INDEX column, the data set contains 15 numerical variables and 12,795 observations. Given that the NAs in the STARS variable are meaningful, we have changed those instances to zero to represent a very poor rating.

Data Dictionary

From the descriptions, we would expect that higher LabelAppeal and STARS values correspond with greater numbers of cases purchased. Variable names kind of indicate that some of them will be correlated with each other

- AcidIndex, CitricAcid, FixedAcidity & VolatileAcidity
- FreeSulfurDioxide & TotalSulfurDioxide
- FreeSulfurDioxide, Sulphates & TotalSulfurDioxide

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET	Number of Cases Purchased	None
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average	
Alcohol	Alcohol Content	
Chlorides	Chloride content of wine	
CitricAcid	Citric Acid Content	
Density	Density of Wine	
FixedAcidity	Fixed Acidity of Wine	
FreeSulfurDioxide	Sulfur Dioxide content of wine	
LabelAppeal	Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customes don't like the design.	Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales.
ResidualSugar	Residual Sugar of wine	
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor	A high number of stars suggests high sales
Sulphates	Sulfate conten of wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	

Statistical Summary

In the table below, we notice that we have missing values in about half the variables. Along with AcidIndex, LabelAppeal & STARS, the response variable TARGET is discrete, which makes this data set a good candidate for count regression. It seems unlikely that these are valid measurements of the variable, and we will address them in with our variable transformations.

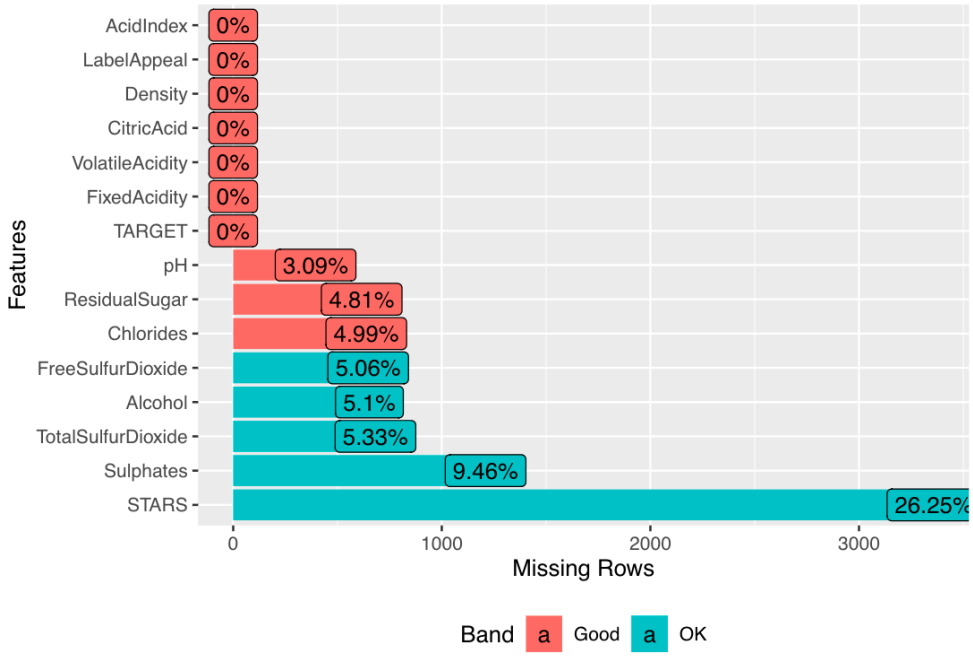
Table 1: Table continues below

	MEAN	MIN	MEDIAN	MAX	IQR	STD. DEV
TARGET	3.03	0	3	8	2	1.93
FixedAcidity	7.08	-18.1	6.9	34.4	4.3	6.32
VolatileAcidity	0.32	-2.79	0.28	3.68	0.51	0.78
CitricAcid	0.31	-3.24	0.31	3.86	0.55	0.86
ResidualSugar	5.42	-127.8	3.9	141.2	17.9	33.75
Chlorides	0.05	-1.17	0.05	1.35	0.18	0.32
FreeSulfurDioxide	30.85	-555	30	623	70	148.7
TotalSulfurDioxide	120.7	-823	123	1057	181	231.9
Density	0.99	0.89	0.99	1.1	0.01	0.03
pH	3.21	0.48	3.2	6.13	0.51	0.68
Sulphates	0.53	-3.13	0.5	4.24	0.58	0.93
Alcohol	10.49	-4.7	10.4	26.5	3.4	3.73
LabelAppeal	-0.01	-2	0	2	2	0.89
AcidIndex	7.77	4	8	17	1	1.32
STARS	2.04	1	2	4	2	0.9

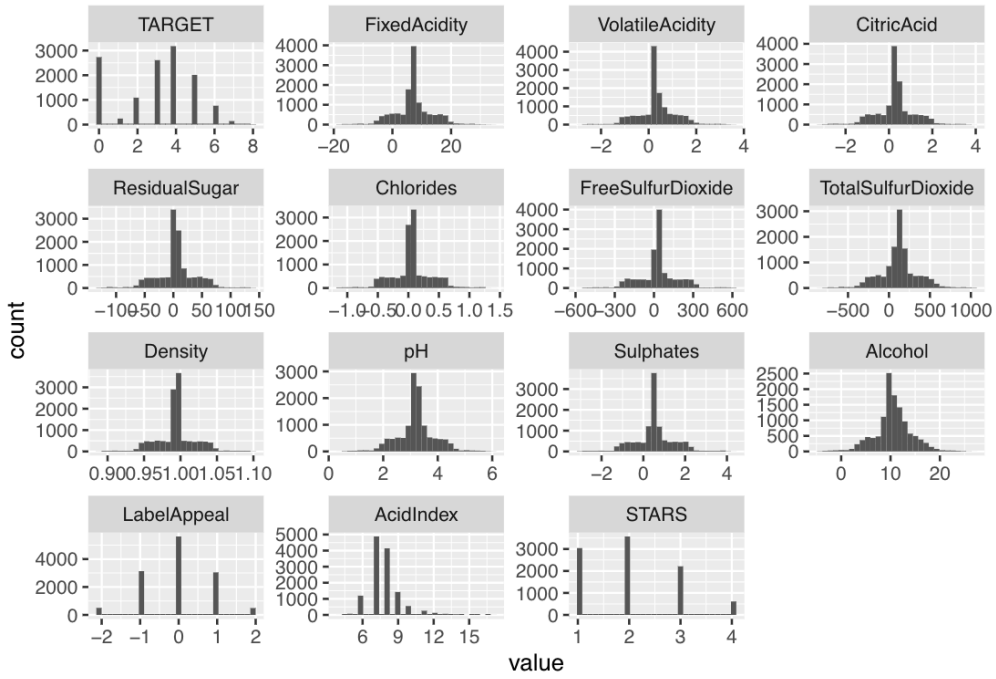
	SKEW	r_{TARGET}	NAs
TARGET	-0.33	1	0
FixedAcidity	-0.02	-0.01	0
VolatileAcidity	0.02	-0.08	0
CitricAcid	-0.05	0	0
ResidualSugar	-0.05	0	616
Chlorides	0.03	-0.03	638
FreeSulfurDioxide	0.01	0.02	647
TotalSulfurDioxide	-0.01	0.02	682
Density	-0.02	-0.05	0
pH	0.04	0	395
Sulphates	0.01	-0.02	1210
Alcohol	-0.03	0.07	653
LabelAppeal	0.01	0.5	0
AcidIndex	1.65	-0.17	0
STARS	0.45	0.55	3359

Visualizations

From the graph below, we can see that they are many variables that have a missing value.

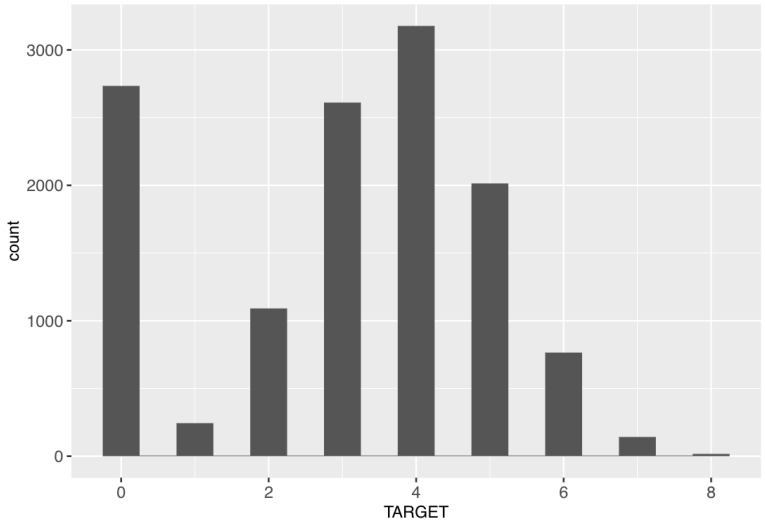


Let's present all the variables in a histogram.



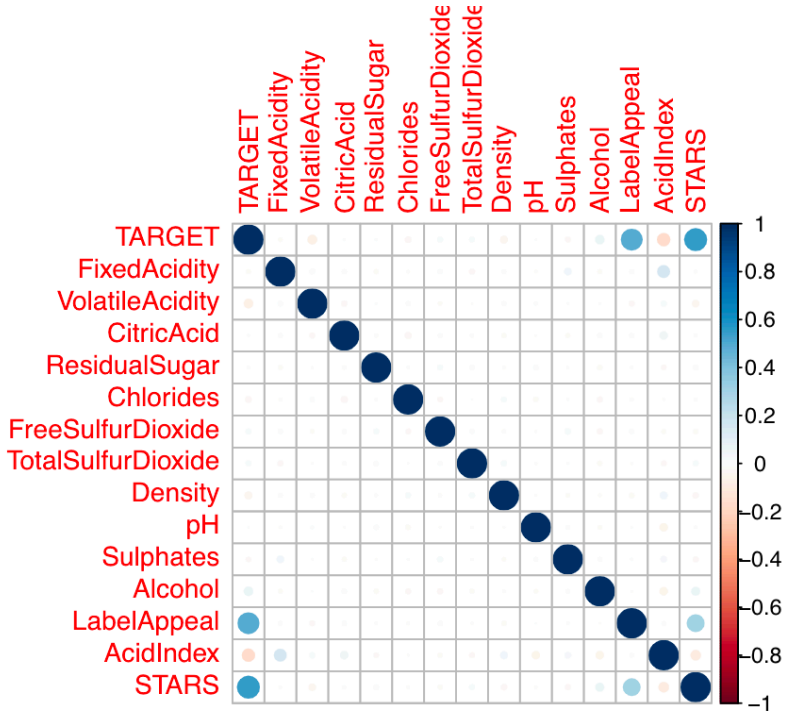
It seems majority of the variables have a normal distribution.

The target variable is a count variable, indicating the number of sample cases. The distribution below indicates that the distribution has a lot of ZERO values, which would indicate 'no sample purchased'. This appears to be a poisson distribution.

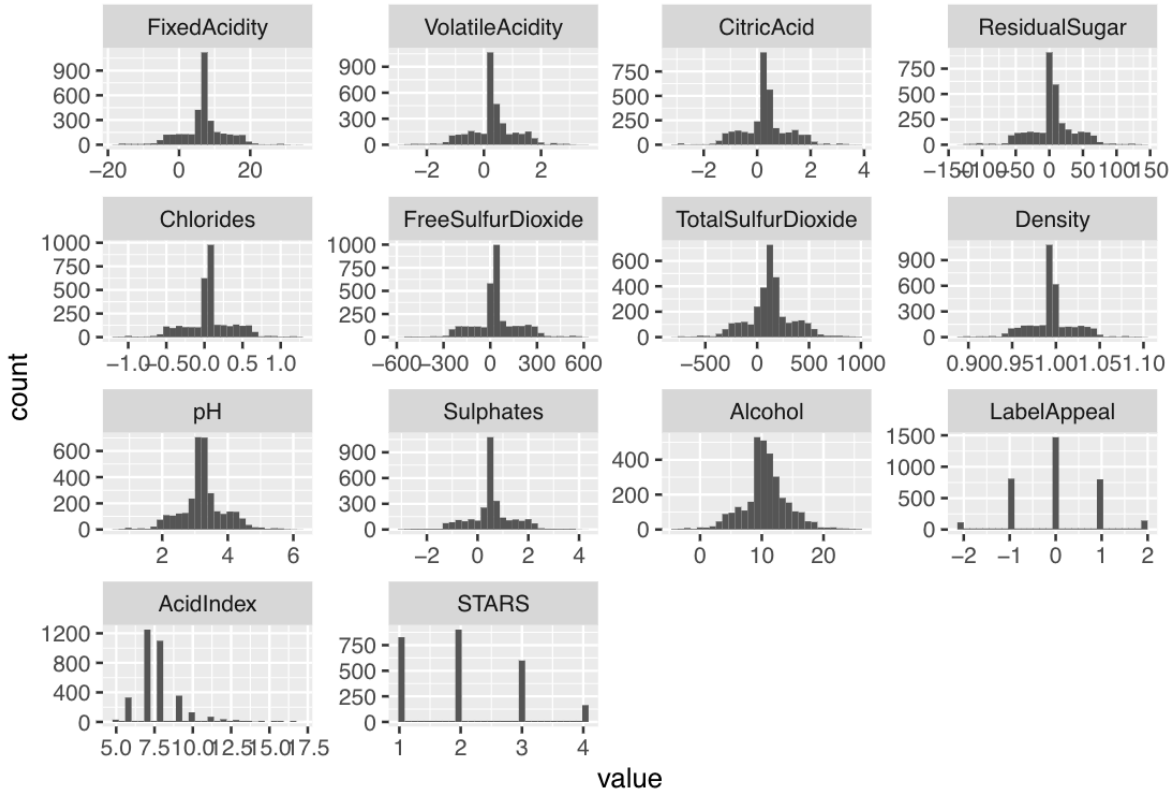


Correlations

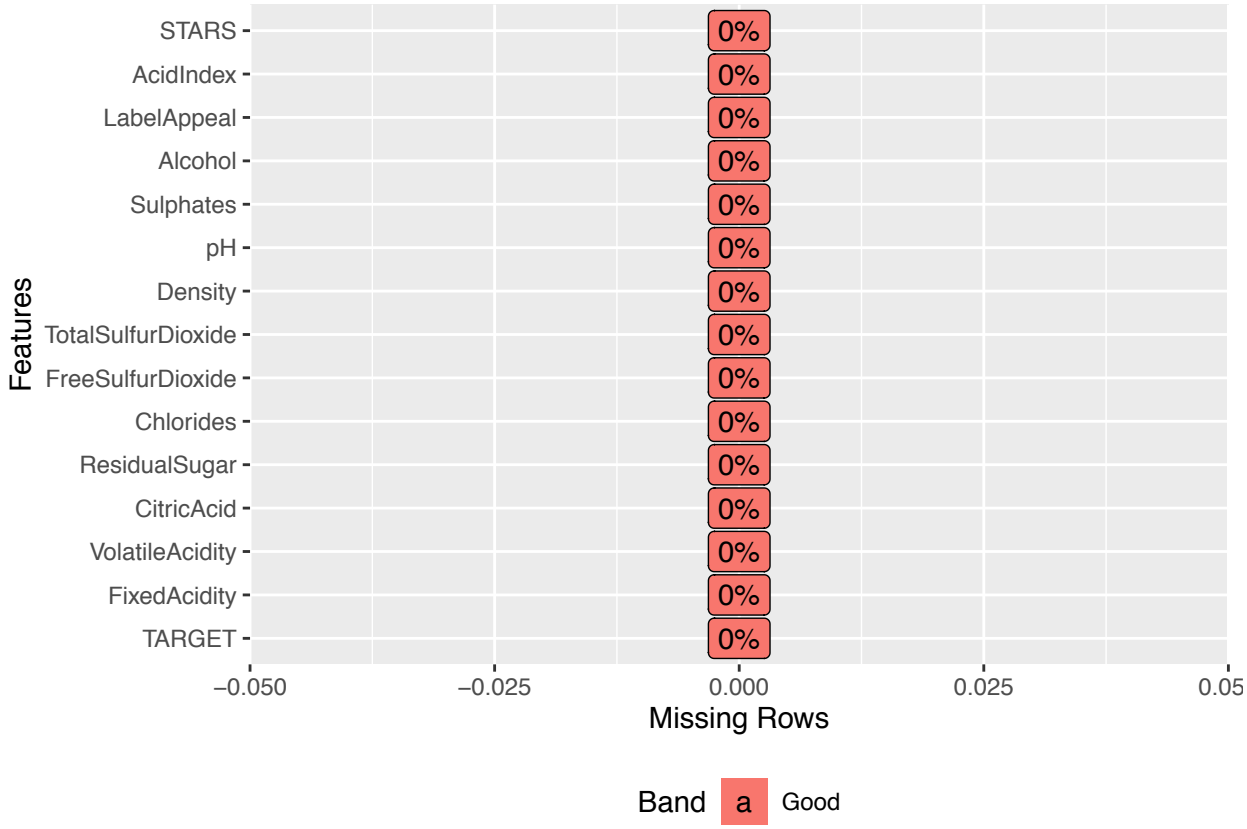
Few predictor variables predictors are correlated with the response variable. As we would expect from the previous boxplots, STARS and LabelAppeal have moderate positive correlations with TARGET, and AcidIndex has a slight negative correlation. STARS is only slightly correlated with LabelAppeal and AcidIndex.



```
## Warning: Removed 2055 rows containing non-finite values (stat_bin).
```



After removing the missing values, we got the above output for the variables.



###Models

3. BUILD MODELS (25 Points) Using the training data set, build at least two different poisson regression models, at least two different negative binomial regression models, and at least two multiple linear regression models, using different variables (or the same variables with different transformations). Sometimes poisson and negative binomial regression models give the same results. If that is the case, comment on that. Consider changing the input variables if that occurs so that you get different models. Although not covered in class, you may also want to consider building zero-inflated poisson and negative binomial regression models. You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? In this case, about the only thing you can comment on is the number of stars and the wine label appeal. However, you might comment on the coefficient and magnitude of variables and how they are similar or different from model to model. For example, you might say "pH seems to have a major positive impact in my poisson regression model, but a negative effect in my multiple linear regression model". Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

##Model 1 - Our kitchen sink regression. This model basically has all the predictor variables.

```
#model1
mod1 <- glm(TARGET ~. , data = train)
(mod1sum <- summary(mod1))
```

```
##
## Call:
## glm(formula = TARGET ~ ., data = train)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8614  -0.7404   0.3705   1.1230   4.7256
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.479e+00  5.542e-01   9.886  < 2e-16 ***
## FixedAcidity        -1.902e-03  2.935e-03  -0.648 0.516948
## VolatileAcidity     -1.682e-01  2.600e-02  -6.471 1.01e-10 ***
## CitricAcid           5.384e-02  2.382e-02   2.260 0.023826 *
## ResidualSugar       -1.020e-04  5.927e-04  -0.172 0.863305
## Chlorides           -1.372e-01  6.182e-02  -2.219 0.026507 *
## FreeSulfurDioxide    2.645e-04  1.370e-04   1.931 0.053477 .
## TotalSulfurDioxide   3.547e-04  9.106e-05   3.895 9.86e-05 ***
## Density             -1.344e+00  5.441e-01  -2.471 0.013500 *
## pH                  -6.038e-02  2.157e-02  -2.800 0.005123 **
## Sulphates           -7.610e-02  2.308e-02  -3.297 0.000979 ***
## Alcohol              1.917e-02  3.981e-03   4.816 1.48e-06 ***
## LabelAppeal          5.940e-01  1.691e-02  35.122  < 2e-16 ***
## AcidIndex           -3.291e-01  1.118e-02 -29.438  < 2e-16 ***
## STARS                7.507e-01  1.950e-02  38.488  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.659694)
##
##     Null deviance: 47477  on 12794  degrees of freedom
## Residual deviance: 33991  on 12780  degrees of freedom
## AIC: 48844
##
## Number of Fisher Scoring iterations: 2
```

##Model2- A poisson regression model

```
#model2
mod2 <- glm(TARGET ~., family=poisson, data=train)
(mod2sum <- summary(mod2))
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2816  -0.5113   0.1984   0.6365   2.7547
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.055e+00  1.962e-01  10.477  < 2e-16 ***
## FixedAcidity        -7.872e-04  1.046e-03  -0.752 0.451908
## VolatileAcidity     -5.870e-02  9.416e-03  -6.234 4.55e-10 ***
## CitricAcid           1.750e-02  8.292e-03   2.110 0.034840 *
## ResidualSugar        2.906e-05  2.087e-04   0.139 0.889221
## Chlorides           -4.213e-02  2.196e-02  -1.919 0.055005 .
## FreeSulfurDioxide    9.024e-05  4.811e-05   1.876 0.060681 .
```

```
## TotalSulfurDioxide  1.183e-04  3.173e-05   3.728 0.000193 ***
## Density            -4.523e-01  1.922e-01  -2.353 0.018600 *
## pH                 -2.262e-02  7.626e-03  -2.967 0.003010 **
## Sulphates          -2.583e-02  8.267e-03  -3.125 0.001780 **
## Alcohol             5.547e-03  1.410e-03   3.935 8.33e-05 ***
## LabelAppeal         1.963e-01  6.020e-03  32.612  < 2e-16 ***
## AcidIndex          -1.232e-01  4.454e-03 -27.673  < 2e-16 ***
## STARS               2.212e-01  6.466e-03  34.203  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 18475  on 12780  degrees of freedom
## AIC: 50447
##
## Number of Fisher Scoring iterations: 5
```

##Model3 - Stepwise regression model of model: mod2

```
#model3
mod3 <- step(mod2, direction = "backward")
```

```
## Start:  AIC=50446.57
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - ResidualSugar        1    18475 50445
## - FixedAcidity         1    18475 50445
## <none>                      18474 50447
## - FreeSulfurDioxide    1    18478 50448
## - Chlorides            1    18478 50448
## - CitricAcid           1    18479 50449
## - Density              1    18480 50450
## - pH                   1    18483 50453
## - Sulphates            1    18484 50454
## - TotalSulfurDioxide   1    18488 50458
## - Alcohol              1    18490 50460
## - VolatileAcidity      1    18514 50484
## - AcidIndex            1    19294 51264
## - LabelAppeal          1    19539 51509
## - STARS                1    19624 51594
##
## Step:  AIC=50444.59
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - FixedAcidity         1    18475 50443
## <none>                      18475 50445
## - FreeSulfurDioxide    1    18478 50446
```

```
## - Chlorides           1    18478 50446
## - CitricAcid          1    18479 50447
## - Density             1    18480 50448
## - pH                  1    18483 50451
## - Sulphates           1    18484 50452
## - TotalSulfurDioxide  1    18488 50456
## - Alcohol             1    18490 50458
## - VolatileAcidity     1    18514 50482
## - AcidIndex           1    19294 51262
## - LabelAppeal         1    19539 51507
## - STARS               1    19624 51592
##
## Step:  AIC=50443.15
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## <none>                     18475 50443
## - FreeSulfurDioxide   1    18479 50445
## - Chlorides           1    18479 50445
## - CitricAcid          1    18480 50446
## - Density             1    18481 50447
## - pH                  1    18484 50450
## - Sulphates           1    18485 50451
## - TotalSulfurDioxide  1    18489 50455
## - Alcohol             1    18491 50457
## - VolatileAcidity     1    18515 50481
## - AcidIndex           1    19326 51292
## - LabelAppeal         1    19540 51506
## - STARS               1    19624 51590
```

```
(mod3sum <- summary(mod3))
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS, family = poisson,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2796  -0.5132   0.1982   0.6344   2.7527
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.053e+00  1.961e-01  10.471  < 2e-16 ***
## VolatileAcidity     -5.875e-02  9.416e-03  -6.239  4.4e-10 ***
## CitricAcid           1.757e-02  8.291e-03   2.120 0.034023 *
## Chlorides           -4.208e-02  2.196e-02  -1.916 0.055316 .
## FreeSulfurDioxide    9.035e-05  4.810e-05   1.878 0.060357 .
## TotalSulfurDioxide   1.184e-04  3.173e-05   3.730 0.000191 ***
## Density             -4.512e-01  1.922e-01  -2.348 0.018876 *
## pH                  -2.269e-02  7.625e-03  -2.975 0.002927 **
```

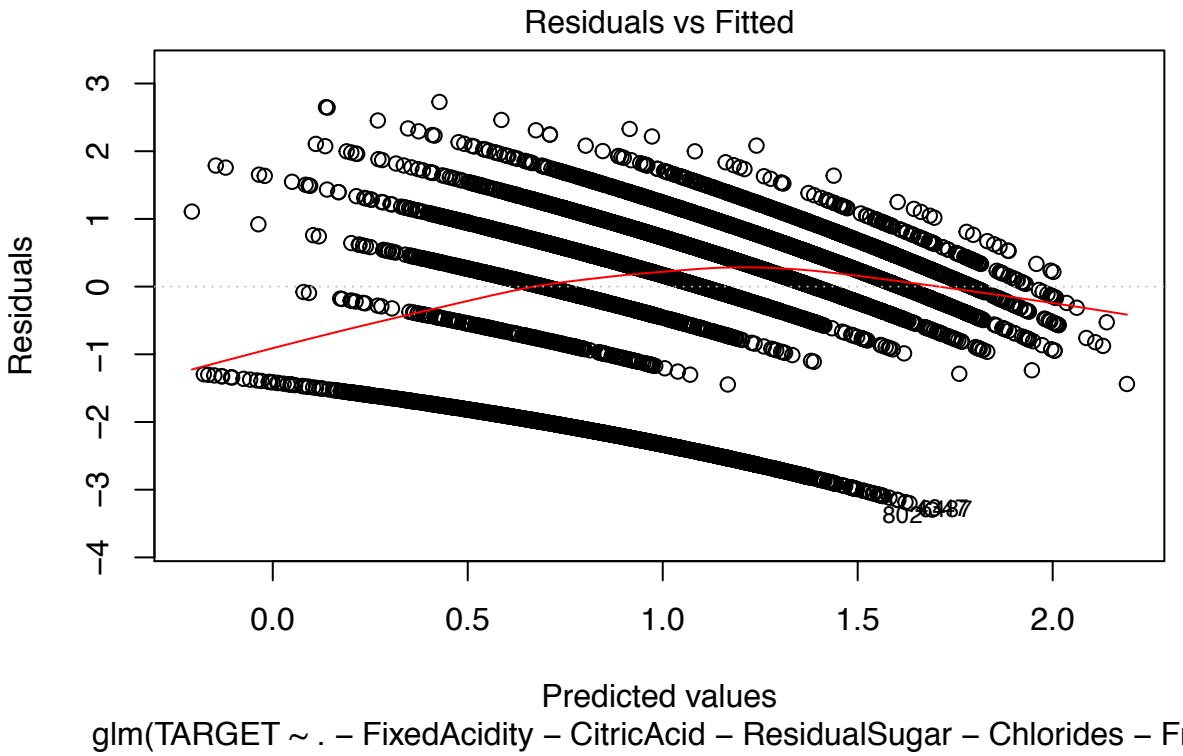
```
## Sulphates          -2.591e-02  8.266e-03  -3.135 0.001720 **
## Alcohol             5.549e-03  1.410e-03   3.936  8.3e-05 ***
## LabelAppeal         1.964e-01  6.020e-03  32.617  < 2e-16 ***
## AcidIndex          -1.238e-01  4.400e-03 -28.127  < 2e-16 ***
## STARS               2.212e-01  6.466e-03  34.205  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 18475  on 12782  degrees of freedom
## AIC: 50443
##
## Number of Fisher Scoring iterations: 5
```

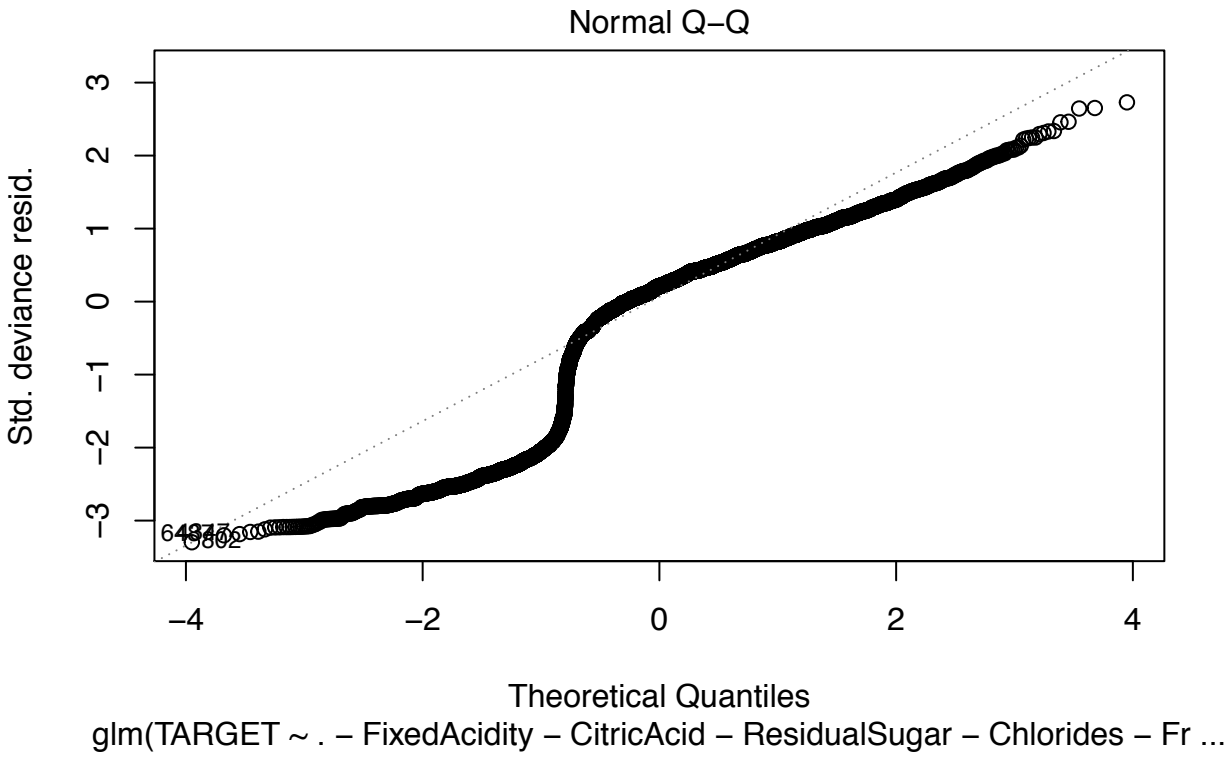
##Model 4- Poisson model with only significant variables

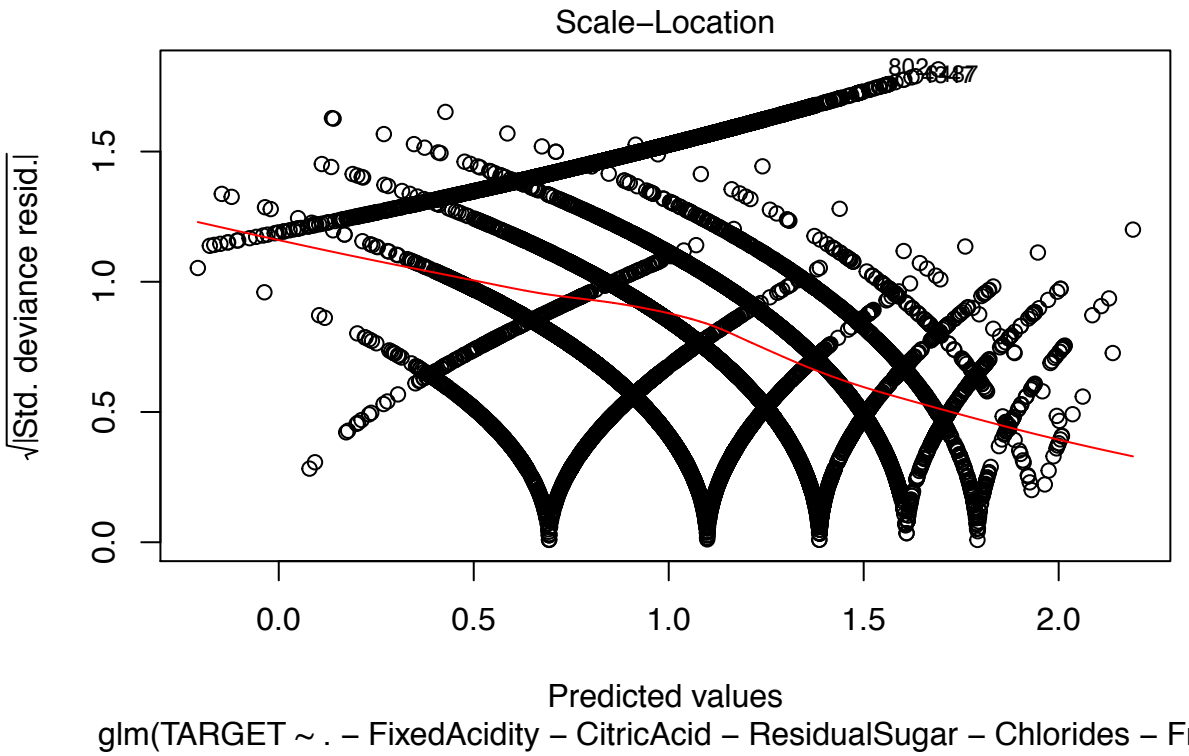
```
#model4
model4 = glm(TARGET ~  .-FixedAcidity-CitricAcid-ResidualSugar-Chlorides-FreeSulfurDioxide-TotalSulf
summary(model4)
```

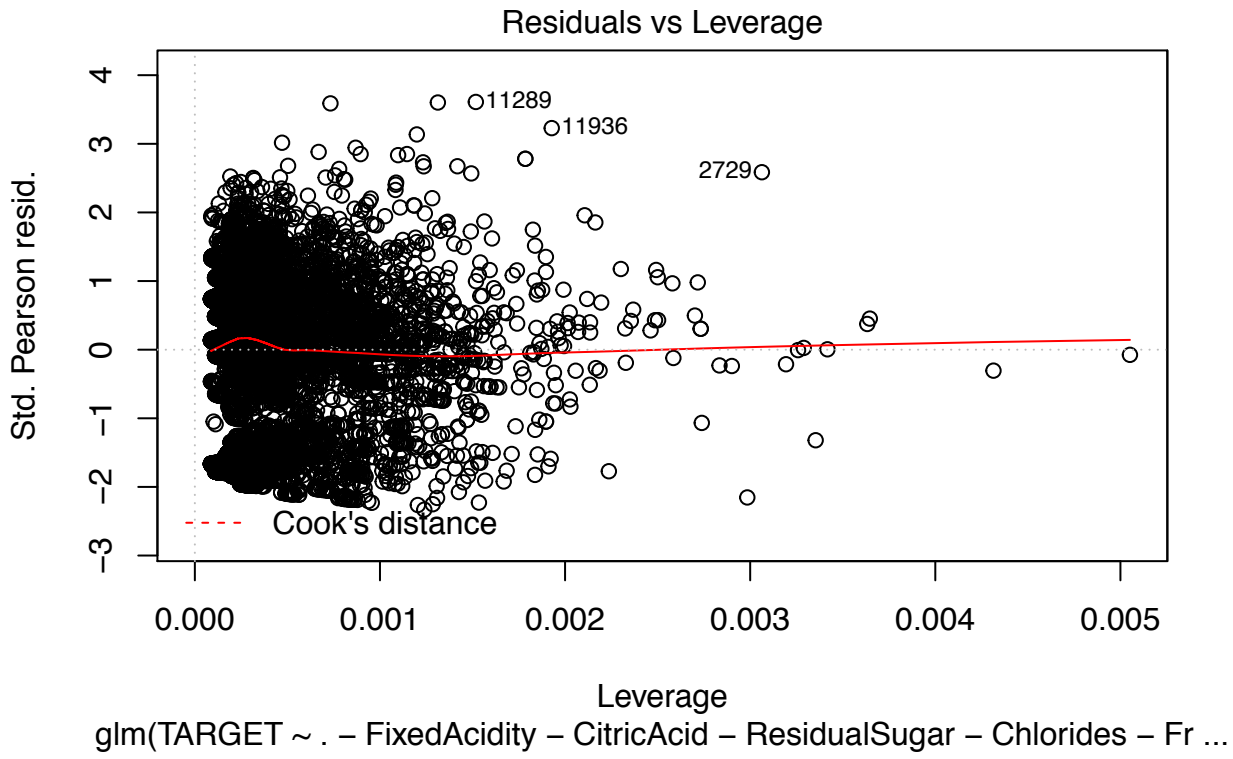
```
##
## Call:
## glm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##     Chlorides - FreeSulfurDioxide - TotalSulfurDioxide - Density -
##     pH - Sulphates - Alcohol, family = poisson, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2943  -0.5088   0.2112   0.6394   2.7273
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.610986   0.037820   42.60  < 2e-16 ***
## VolatileAcidity -0.060226   0.009411   -6.40 1.56e-10 ***
## LabelAppeal      0.196203   0.006017   32.61  < 2e-16 ***
## AcidIndex       -0.124891   0.004373  -28.56  < 2e-16 ***
## STARS            0.223126   0.006451   34.59  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 18541  on 12790  degrees of freedom
## AIC: 50493
##
## Number of Fisher Scoring iterations: 5
```

```
plot(model4)
```


Residuals vs Fitted
Residuals
Predicted values
glm(TARGET ~ . – FixedAcidity – CitricAcid – ResidualSugar – Chlorides – Fr ...
3
2
1
0
−1
−2
−3
−4
0.0
0.5
1.0
1.5
2.0


Normal Q–Q
Std. deviance resid.
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
Theoretical Quantiles
glm(TARGET ~ . – FixedAcidity – CitricAcid – ResidualSugar – Chlorides – Fr ...


Scale−Location
√|Std. deviance 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 − Chlorides − Fr ...



###Model 5- Negative binomial

```
#model5
model5 <- glm.nb(TARGET ~ ., data = train)
```

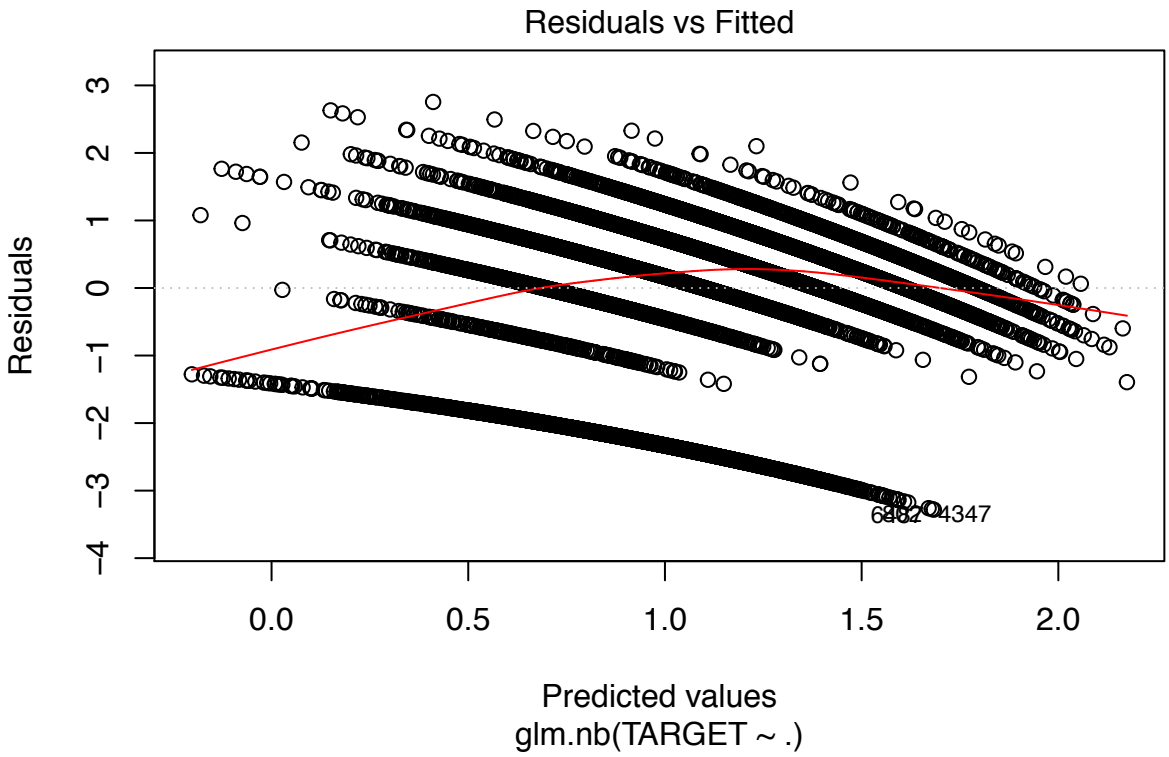
```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached
```

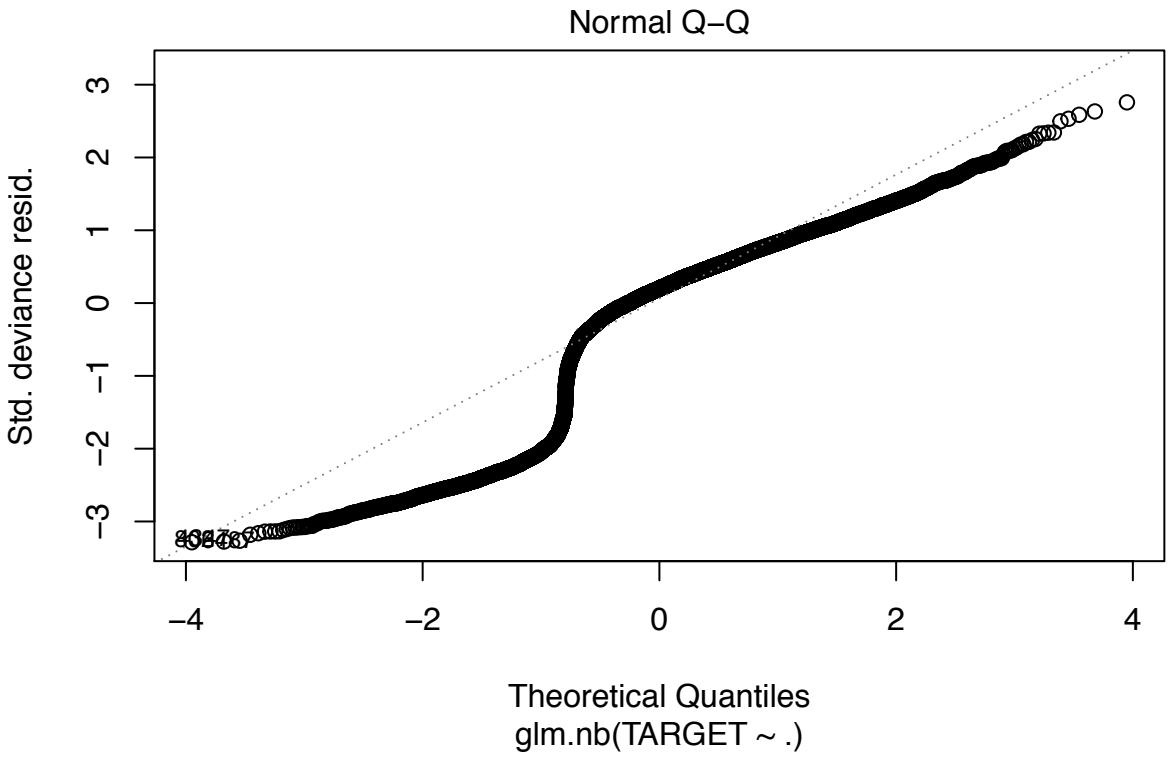
```
summary(model5)
```

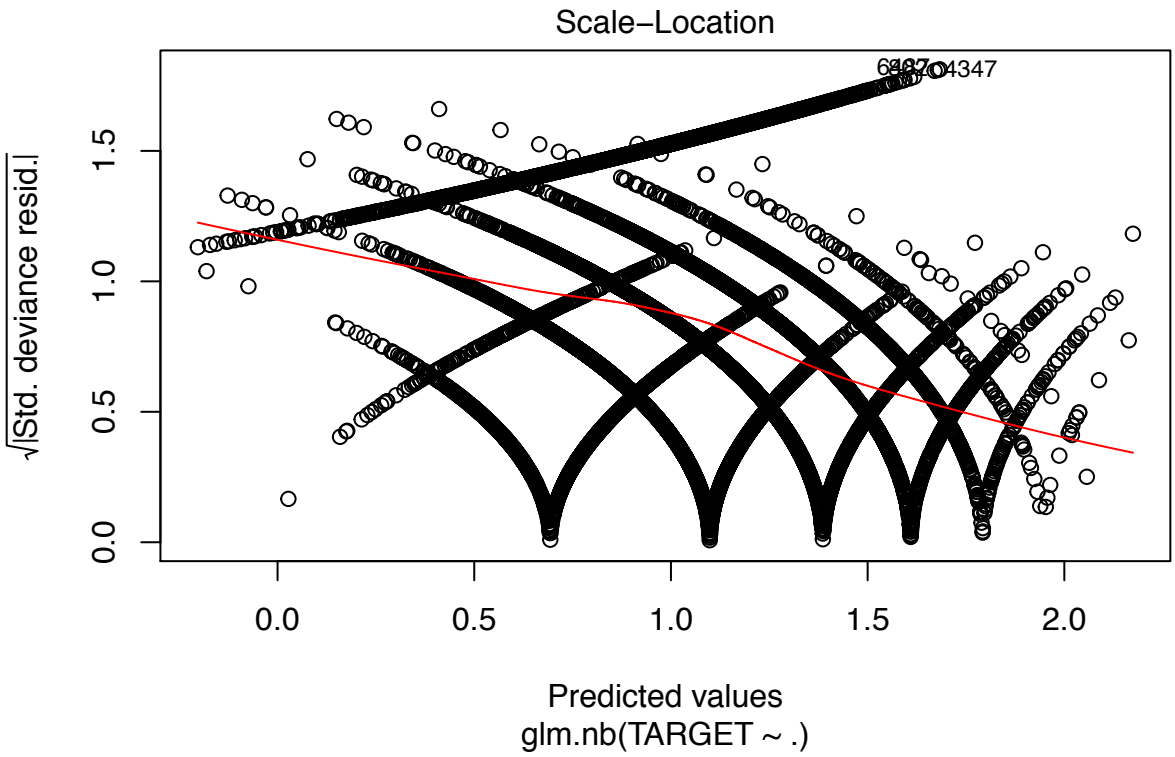
```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = train, init.theta = 39158.3886,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2815  -0.5113   0.1984   0.6365   2.7546
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.055e+00  1.962e-01  10.476  < 2e-16 ***
## FixedAcidity        -7.872e-04  1.047e-03  -0.752 0.451910
## VolatileAcidity     -5.870e-02  9.416e-03  -6.234 4.55e-10 ***
## CitricAcid           1.750e-02  8.293e-03   2.110 0.034849 *
```

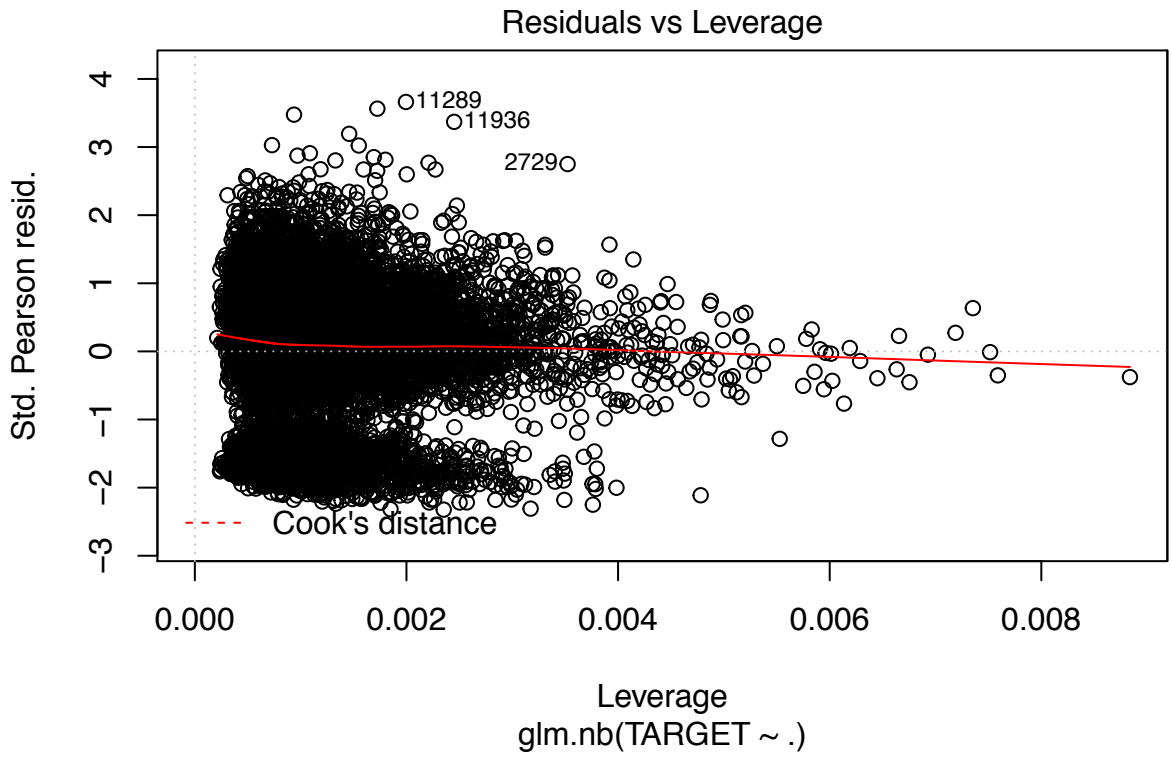
```
## ResidualSugar       2.907e-05  2.087e-04   0.139 0.889222
## Chlorides          -4.213e-02  2.196e-02  -1.919 0.055014 .
## FreeSulfurDioxide   9.024e-05  4.811e-05   1.876 0.060690 .
## TotalSulfurDioxide  1.183e-04  3.173e-05   3.728 0.000193 ***
## Density            -4.523e-01  1.922e-01  -2.353 0.018602 *
## pH                 -2.262e-02  7.626e-03  -2.967 0.003010 **
## Sulphates          -2.583e-02  8.268e-03  -3.125 0.001780 **
## Alcohol             5.547e-03  1.410e-03   3.935 8.33e-05 ***
## LabelAppeal         1.963e-01  6.021e-03  32.611  < 2e-16 ***
## AcidIndex          -1.232e-01  4.454e-03 -27.673  < 2e-16 ***
## STARS               2.212e-01  6.466e-03  34.201  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(39158.39) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 18474  on 12780  degrees of freedom
## AIC: 50449
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  39158
##           Std. Err.:  59656
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -50416.69
```

```
plot(model5)
```


Residuals vs Fitted
Residuals
3
2
1
0
−1
−2
−3
−4
0.0
0.5
1.0
1.5
2.0
4347
Predicted values
glm.nb(TARGET ~ .)


Normal Q−Q
Std. deviance resid.
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
Theoretical Quantiles
glm.nb(TARGET ~ .)


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



0.1 Model 6- Negative binomial with significant variables

```
#model6
model6 <- glm.nb(TARGET ~ .-FixedAcidity-CitricAcid-ResidualSugar-Chlorides-FreeSulfurDioxide-TotalS
```

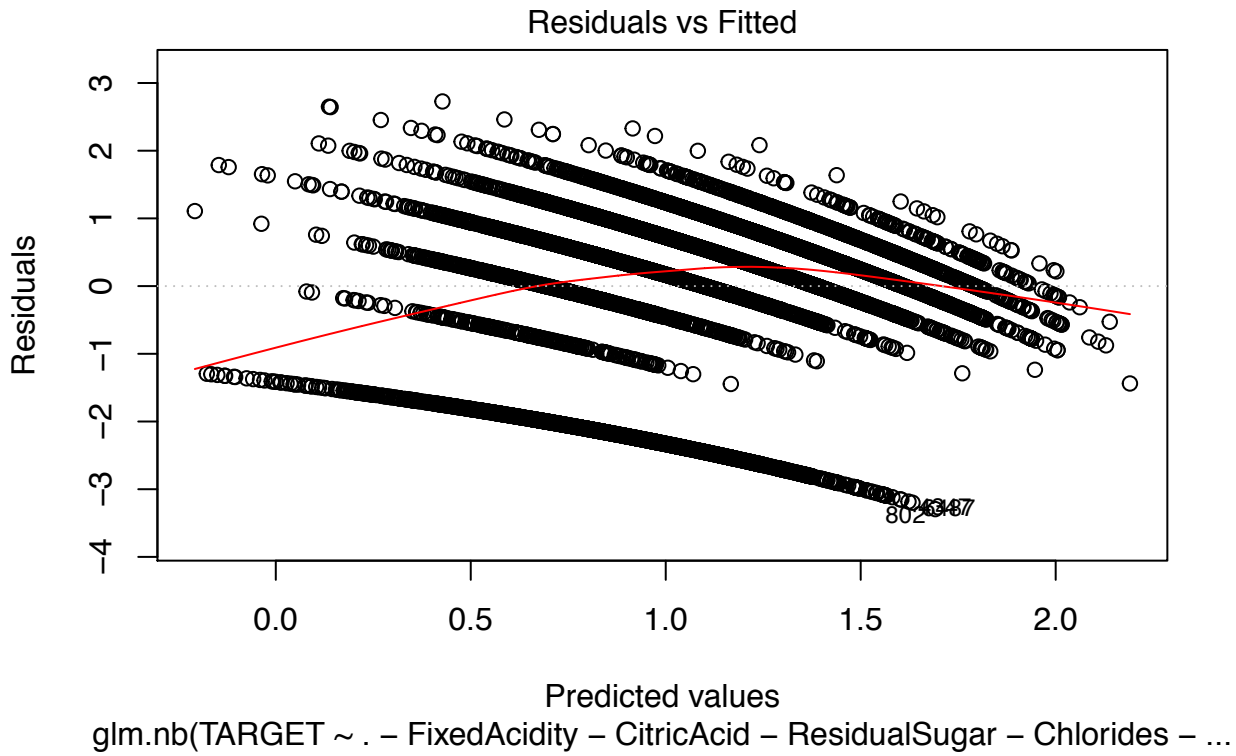
```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached
```

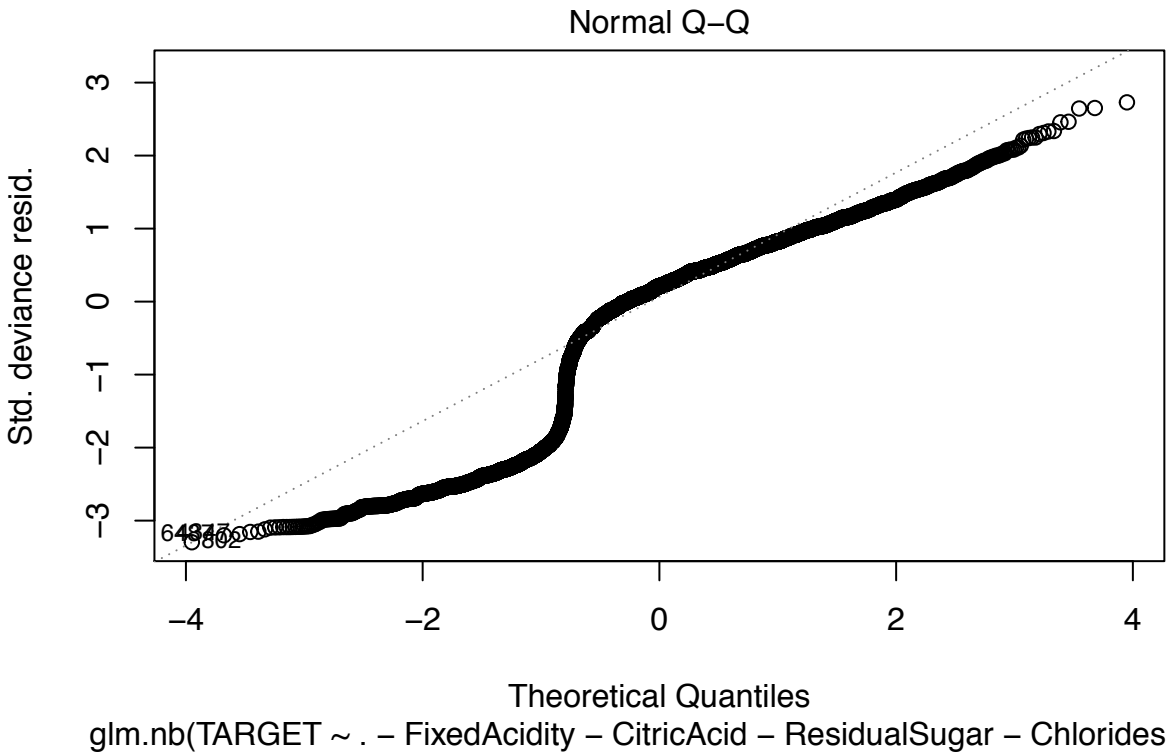
```
summary(model6)
```

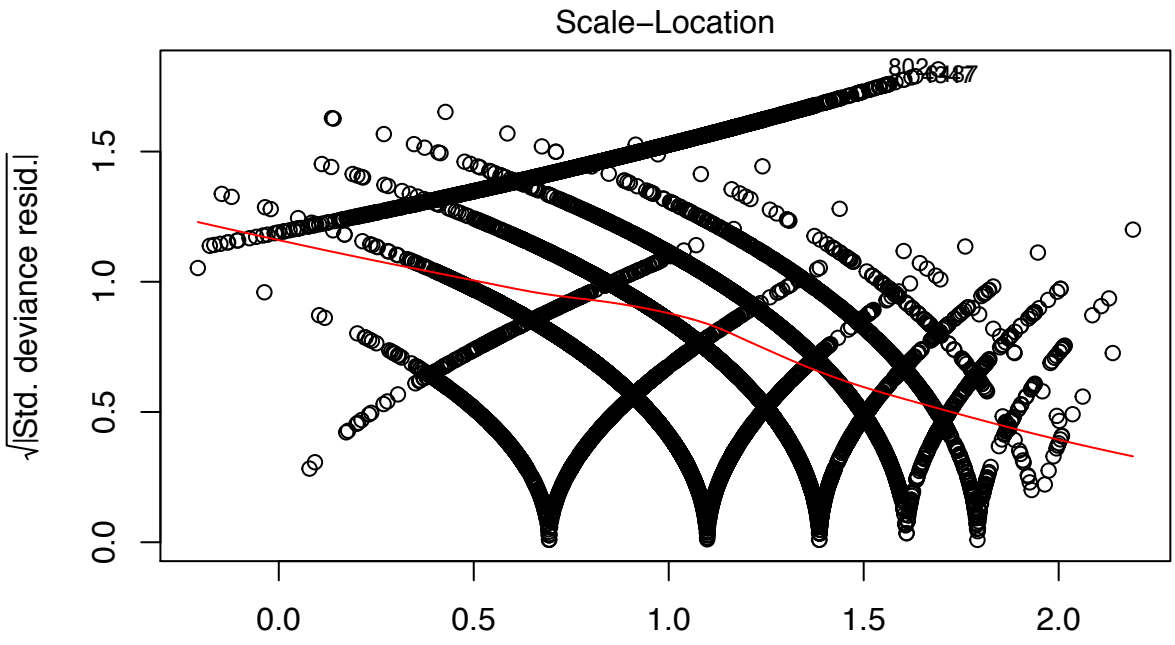
```
##
## Call:
## glm.nb(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##     Chlorides - FreeSulfurDioxide - TotalSulfurDioxide - Density -
##     pH - Sulphates - Alcohol, data = train, init.theta = 38796.83639,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2942  -0.5087   0.2112   0.6394   2.7273
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.611002   0.037822   42.59  < 2e-16 ***
```

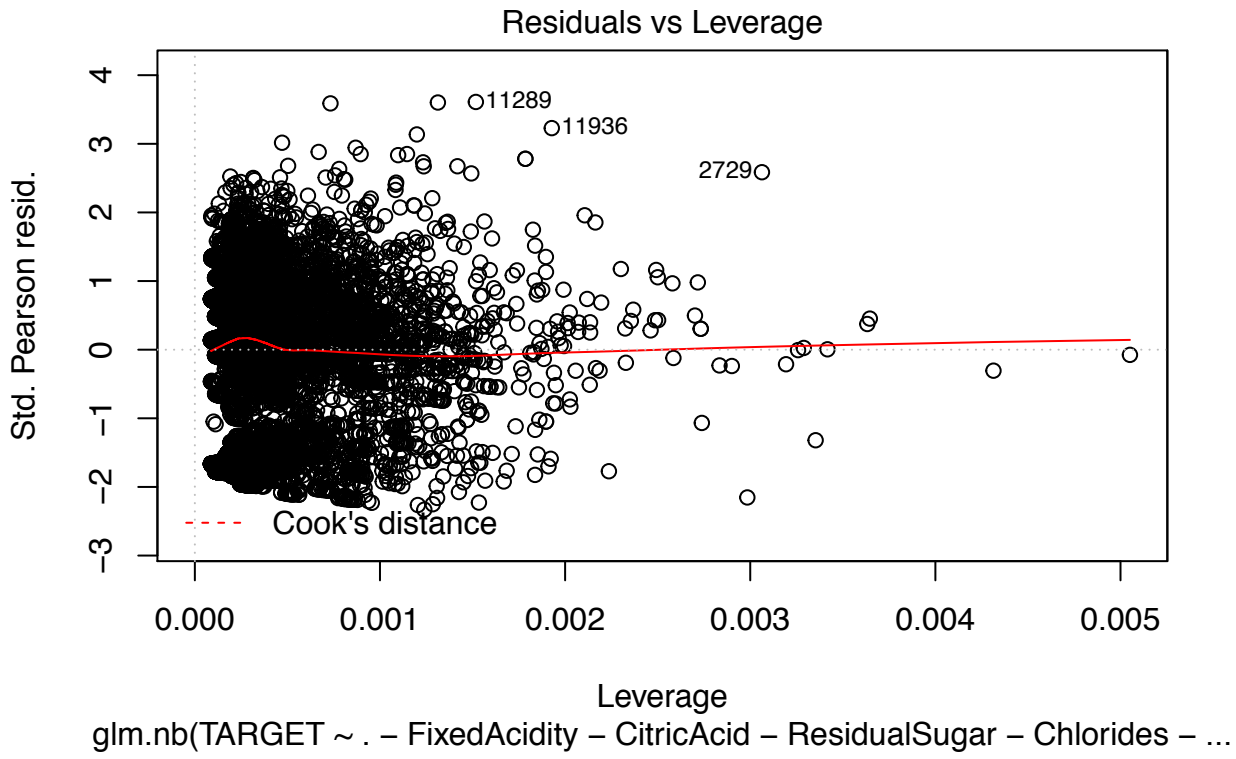
```
## VolatileAcidity -0.060227   0.009411   -6.40 1.56e-10 ***
## LabelAppeal      0.196204   0.006017   32.61  < 2e-16 ***
## AcidIndex       -0.124893   0.004374  -28.56  < 2e-16 ***
## STARS            0.223125   0.006451   34.59  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(38796.84) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 18540  on 12790  degrees of freedom
## AIC: 50495
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  38797
##           Std. Err.:  59786
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -50482.94
```

```
plot(model6)
```




Normal Q–Q
Std. deviance resid.
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
Theoretical Quantiles
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Chlorides − ...


Scale−Location
√|Std. deviance 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 − Chlorides − ...



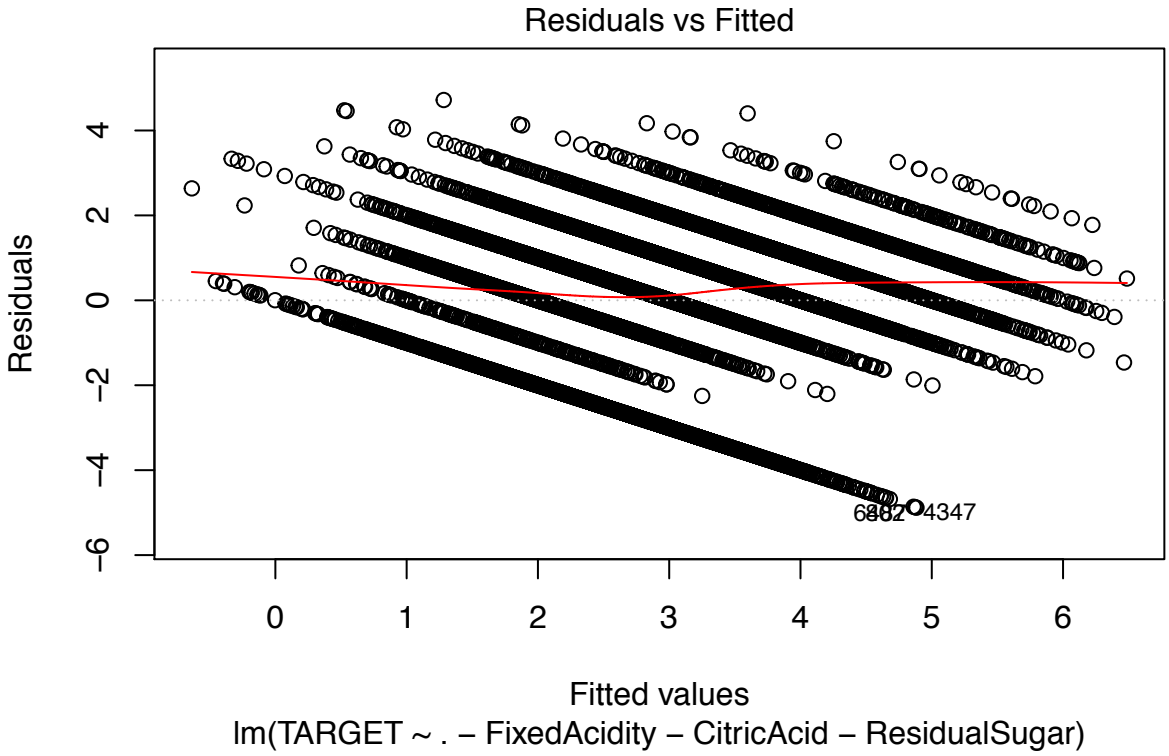
##Model 7 -Multiple linear regression

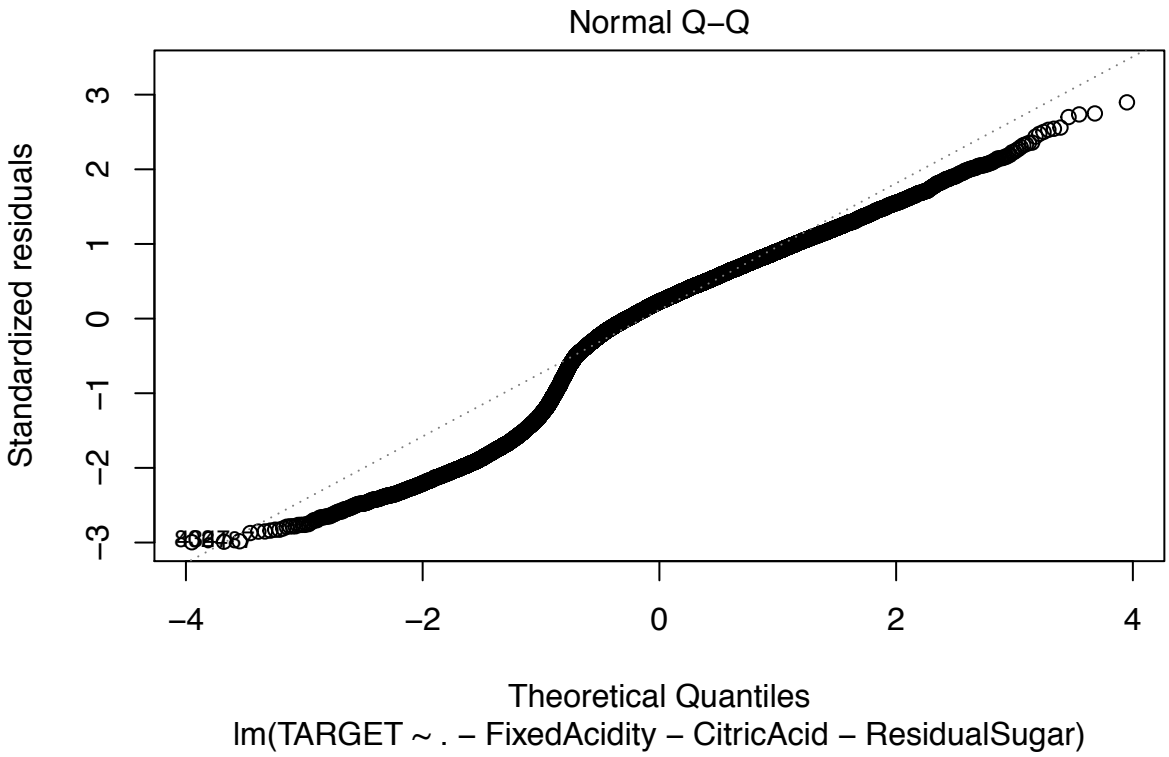
```
#model7
model7 <- lm(TARGET ~ .-FixedAcidity-CitricAcid-ResidualSugar, data = train)
summary(model7)
```

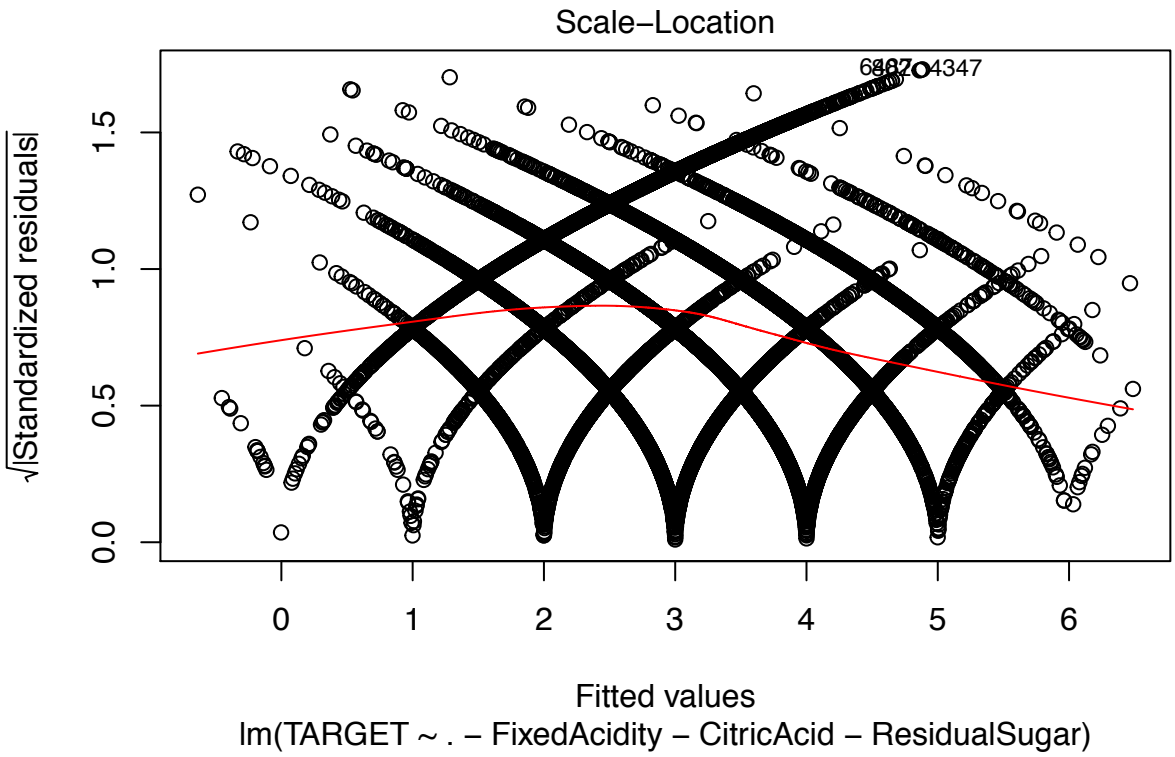
```
##
## Call:
## lm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8835 -0.7409  0.3701  1.1238  4.7173
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.513e+00  5.536e-01   9.958  < 2e-16 ***
## VolatileAcidity     -1.685e-01  2.600e-02  -6.481 9.47e-11 ***
## Chlorides           -1.377e-01  6.183e-02  -2.228  0.02592 *
## FreeSulfurDioxide    2.659e-04  1.370e-04   1.941  0.05224 .
## TotalSulfurDioxide   3.569e-04  9.106e-05   3.920 8.90e-05 ***
## Density             -1.357e+00  5.441e-01  -2.494  0.01265 *
## pH                  -6.022e-02  2.157e-02  -2.792  0.00524 **
## Sulphates           -7.542e-02  2.308e-02  -3.268  0.00108 **
## Alcohol              1.914e-02  3.981e-03   4.808 1.54e-06 ***
## LabelAppeal          5.946e-01  1.691e-02  35.163  < 2e-16 ***
```

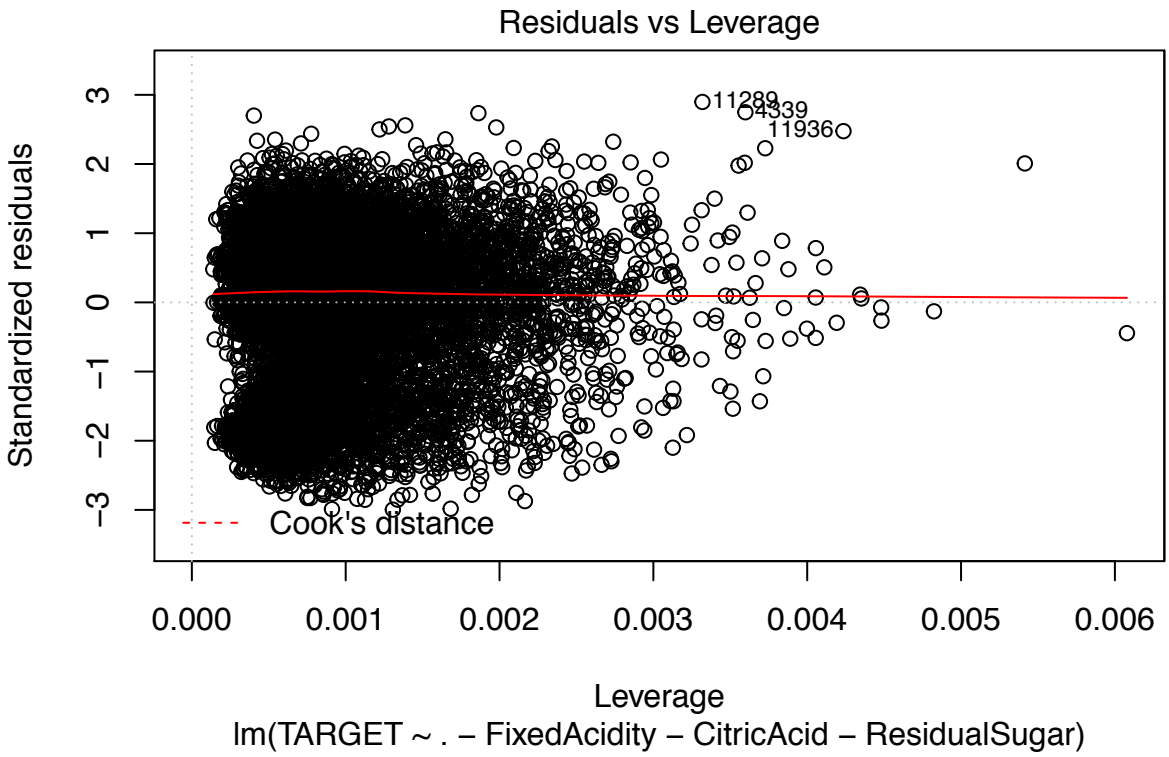
```
## AcidIndex          -3.294e-01  1.099e-02 -29.972  < 2e-16 ***
## STARS               7.507e-01  1.951e-02  38.484  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.631 on 12783 degrees of freedom
## Multiple R-squared:  0.2837, Adjusted R-squared:  0.2831
## F-statistic: 460.4 on 11 and 12783 DF,  p-value: < 2.2e-16
```

```
plot(model7)
```




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


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



##Model 8- Zero inflated poisson regression model

```
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
#model7
model8 <-zeroinfl(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + TotalSulfur
summary(model8)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity +
##     TotalSulfurDioxide + Chlorides + Density, data = train, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.8669 -0.3556  0.1678  0.5060  4.8585
##
## Count model coefficients (poisson with log link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.573e+00  2.037e-01   7.719 1.18e-14 ***
## STARS               1.036e-01  6.431e-03  16.106  < 2e-16 ***
```

```
## LabelAppeal         2.510e-01  6.382e-03  39.335  < 2e-16 ***
## AcidIndex          -1.894e-02  4.929e-03  -3.843 0.000122 ***
## VolatileAcidity    -1.226e-02  9.758e-03  -1.256 0.209114
## TotalSulfurDioxide -3.809e-05  3.253e-05  -1.171 0.241601
## Chlorides          -1.744e-02  2.300e-02  -0.758 0.448247
## Density            -3.440e-01  2.024e-01  -1.700 0.089190 .
##
## Zero-inflation model coefficients (binomial with logit link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.0098192  1.0254467  -4.885 1.03e-06 ***
## STARS              -0.6846710  0.0381101 -17.966  < 2e-16 ***
## LabelAppeal         0.3240544  0.0331183   9.785  < 2e-16 ***
## AcidIndex           0.4836262  0.0193138  25.040  < 2e-16 ***
## VolatileAcidity     0.2566653  0.0454236   5.650 1.60e-08 ***
## TotalSulfurDioxide -0.0010364  0.0001878  -5.518 3.43e-08 ***
## Chlorides           0.1451603  0.1142885   1.270    0.204
## Density             0.9575049  1.0205714   0.938    0.348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 22
## Log-likelihood: -2.258e+04 on 16 Df
```

4. SELECT MODELS (25 Points) Decide on the criteria for selecting the best count regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models. For the count regression model, will you use a metric such as AIC, average squared error, etc.? Be sure to explain how you can make inferences from the model, and discuss other relevant model output. If you like the multiple linear regression model the best, please say why. However, you must select a count regression model for model deployment. Using the training data set, evaluate the performance of the count regression model. Make predictions using the evaluation data set.

```
summary(eva)
```

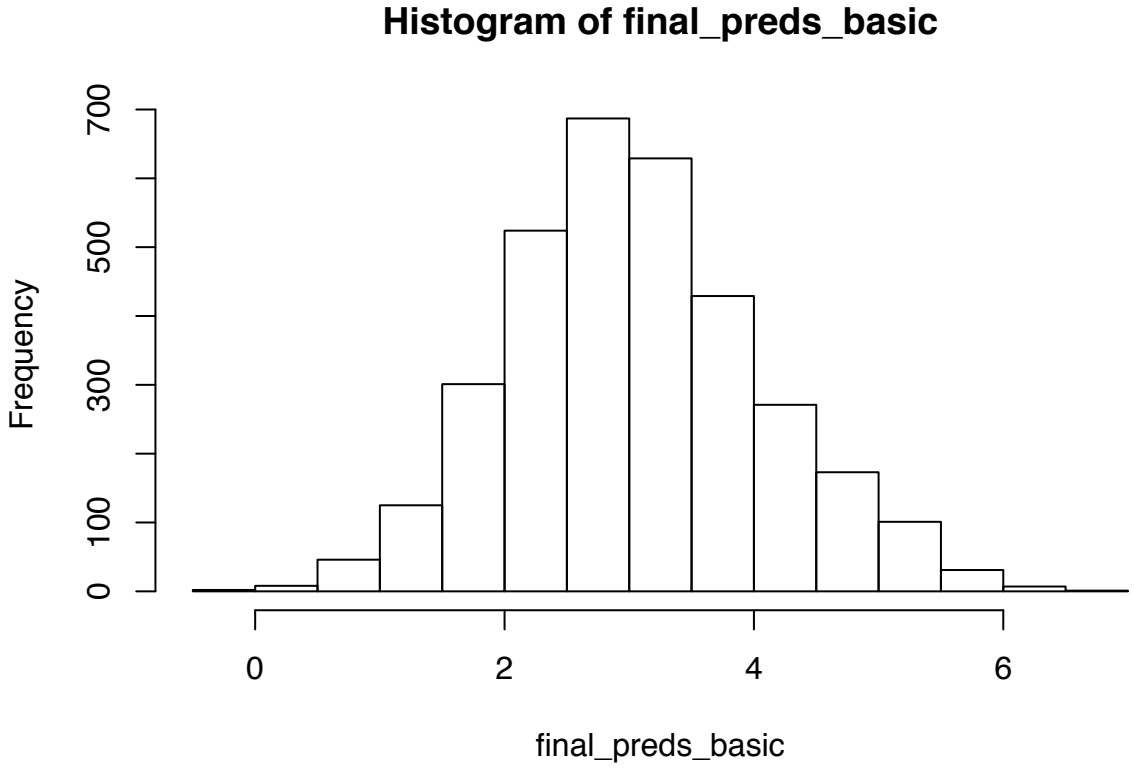
```
##   FixedAcidity    VolatileAcidity    CitricAcid     ResidualSugar
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   :  0.10
##  1st Qu.: 5.700   1st Qu.:0.2500   1st Qu.:0.2900   1st Qu.:  4.00
##  Median : 7.000   Median :0.4200   Median :0.4400   Median : 14.90
##  Mean   : 7.967   Mean   :0.6542   Mean   :0.6969   Mean   : 23.79
##  3rd Qu.: 9.400   3rd Qu.:0.9300   3rd Qu.:1.0000   3rd Qu.: 36.95
##  Max.   :33.500   Max.   :3.6100   Max.   :3.7600   Max.   :145.40
##    Chlorides      FreeSulfurDioxide TotalSulfurDioxide    Density
##  Min.   :0.0000   Min.   :  0.0     Min.   :   0.0     Min.   :0.8898
##  1st Qu.:0.0430   1st Qu.: 28.0     1st Qu.: 100.0     1st Qu.:0.9883
##  Median :0.0870   Median : 60.0     Median : 161.0     Median :0.9946
##  Mean   :0.2121   Mean   :107.2     Mean   : 201.5     Mean   :0.9947
##  3rd Qu.:0.3590   3rd Qu.:167.5     3rd Qu.: 252.0     3rd Qu.:1.0005
##  Max.   :1.2630   Max.   :617.0     Max.   :1004.0     Max.   :1.0998
##        pH          Sulphates         Alcohol       LabelAppeal
##  Min.   :0.600   Min.   :0.0000   Min.   :-4.20   Min.   :-2.00000
##  1st Qu.:2.990   1st Qu.:0.4500   1st Qu.: 9.10   1st Qu.:-1.00000
##  Median :3.200   Median :0.6300   Median :10.60   Median : 0.00000
##  Mean   :3.229   Mean   :0.8486   Mean   :10.61   Mean   : 0.01349
##  3rd Qu.:3.460   3rd Qu.:1.0000   3rd Qu.:12.40   3rd Qu.: 1.00000
##  Max.   :6.210   Max.   :4.1800   Max.   :25.60   Max.   : 2.00000
```

```
##    AcidIndex          STARS
##  Min.   : 5.000   Min.   :1.00
##  1st Qu.: 7.000   1st Qu.:2.00
##  Median : 8.000   Median :2.00
##  Mean   : 7.748   Mean   :2.03
##  3rd Qu.: 8.000   3rd Qu.:2.00
##  Max.   :17.000   Max.   :4.00
```

Yes, any analyst would definitely select a models that best suites the dataset even though it has a slightly worse performance. The model doesn't mean that you have to only show the good things about the data. You should be able to show what the dataset is trying to tell you. If you have the good anlaysis of a dataset, you can predict the best model and give the best fitted solution for that model.

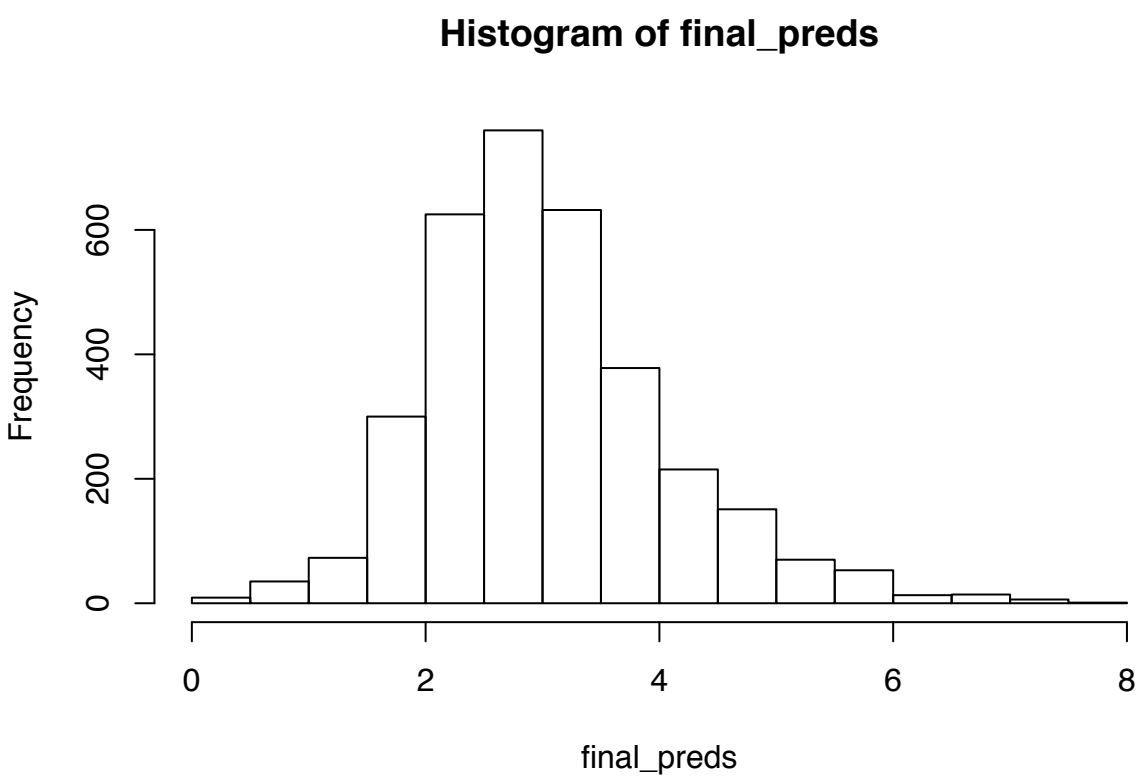
Model 1 has the lowest AIC = 48844.

```
#histogram for prediction basic model 1
final_preds_basic <- predict(mod1, eva)
finaldf <- cbind(TARGET_FLAG=final_preds_basic)
hist(final_preds_basic)
```



Compared with the training dataset, Model8(zero inflated poisson model) would make a good sense. Zero inflated model is used to model count data which has excess zero values.

```
#histogram for predictions (zero inflated poisson)
final_preds <- predict(model8, eva)
finaldf <- cbind(TARGET_FLAG=final_preds)
hist(final_preds)
```



Both model 1 and model 8 gives similar output but the only issue with Model 8 is not all the zero values are a non-values. Zero value doesn't mean it is a null value. It has a value which is equivalent to 0.