Data 621: Assignment 5

*Wine Data*

*Ritesh Lohiya*

*July 12, 2018*

## *Overview*

**In this homework assignment, you will explore, analyze and model a data set 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. Your objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:**

**Data Exploration:**

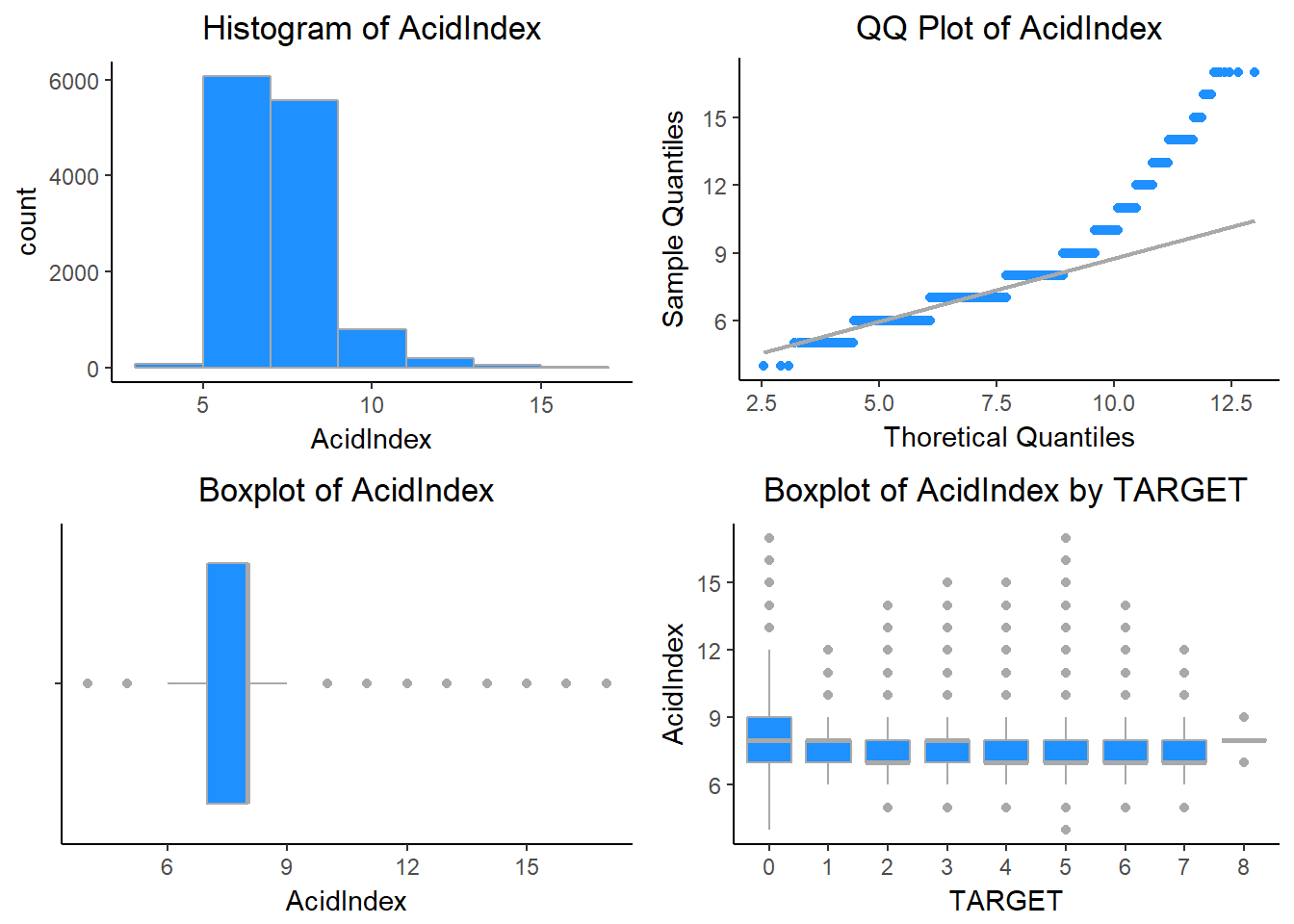
The training data set includes 12795 observations, with 16 variables: 14 predictors, 1 response variables, and one record identifier. Below is a brief description of the included variables:

First we will remove the INDEX column.

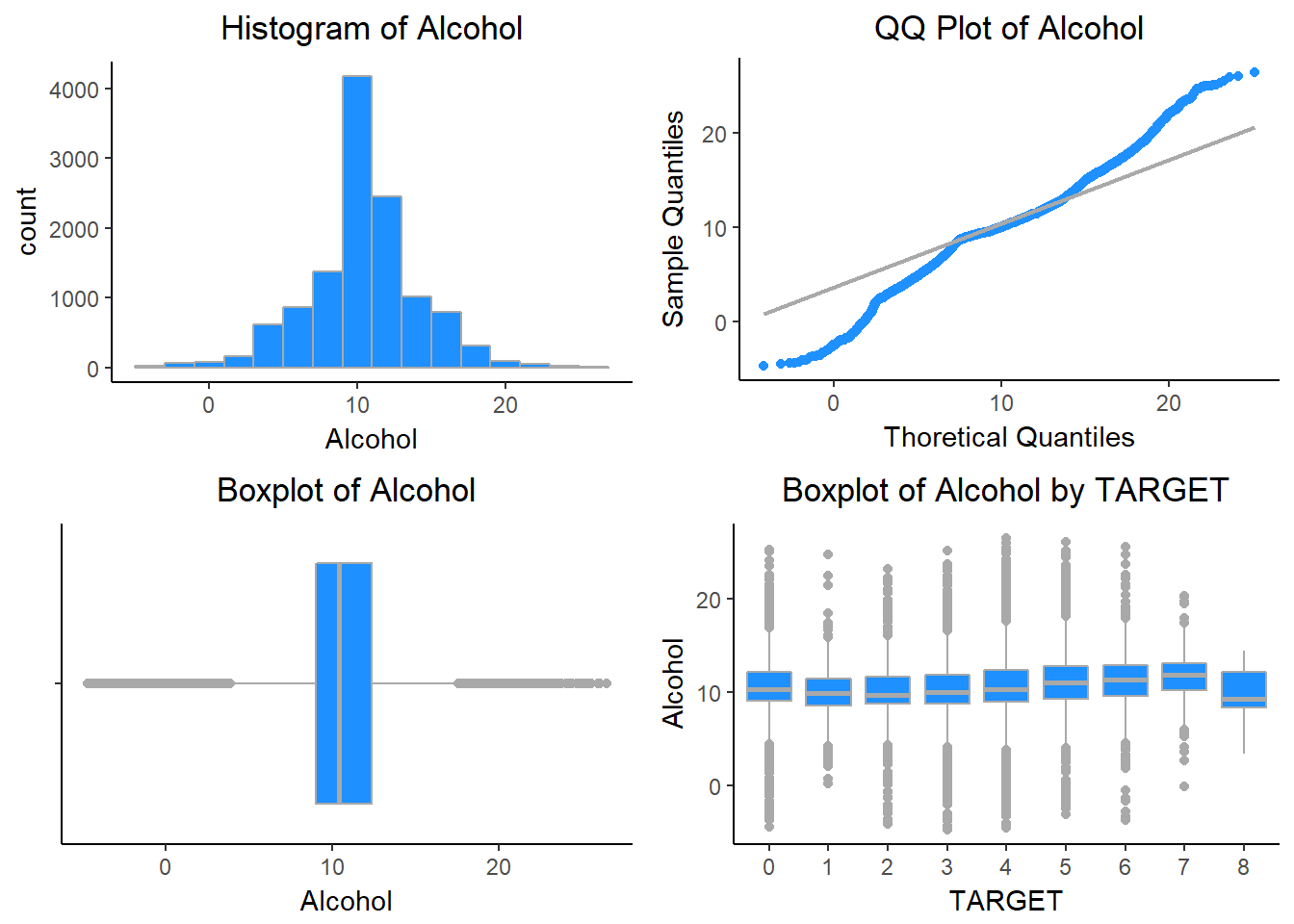
Visual Exploration:

Let's dig into our available variables.

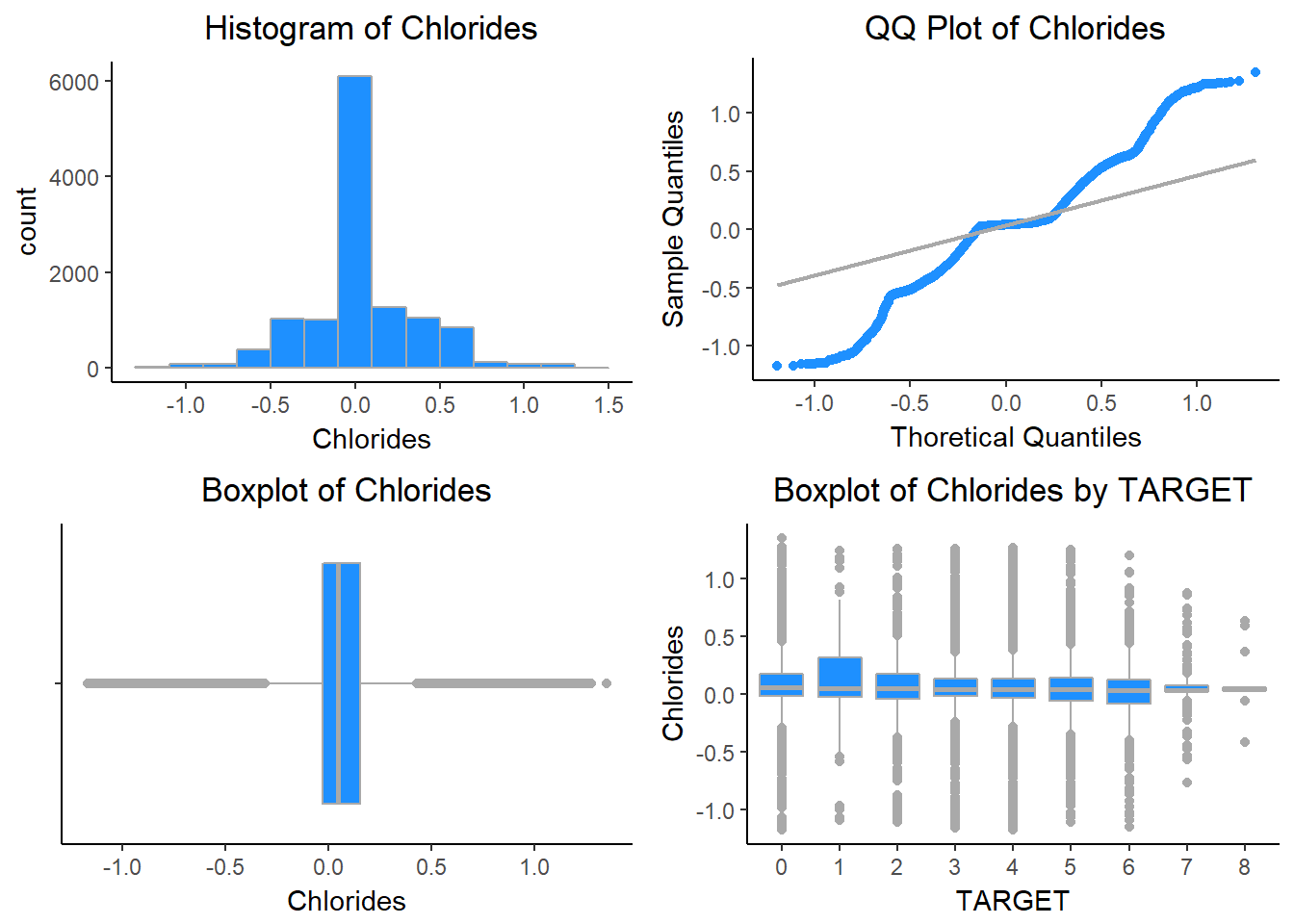
AcidIndex - Proprietary method of testing total acidity of wine by using a weighted average. From the plot below looks like is slightly right skewed. Also correlation with the target is low.



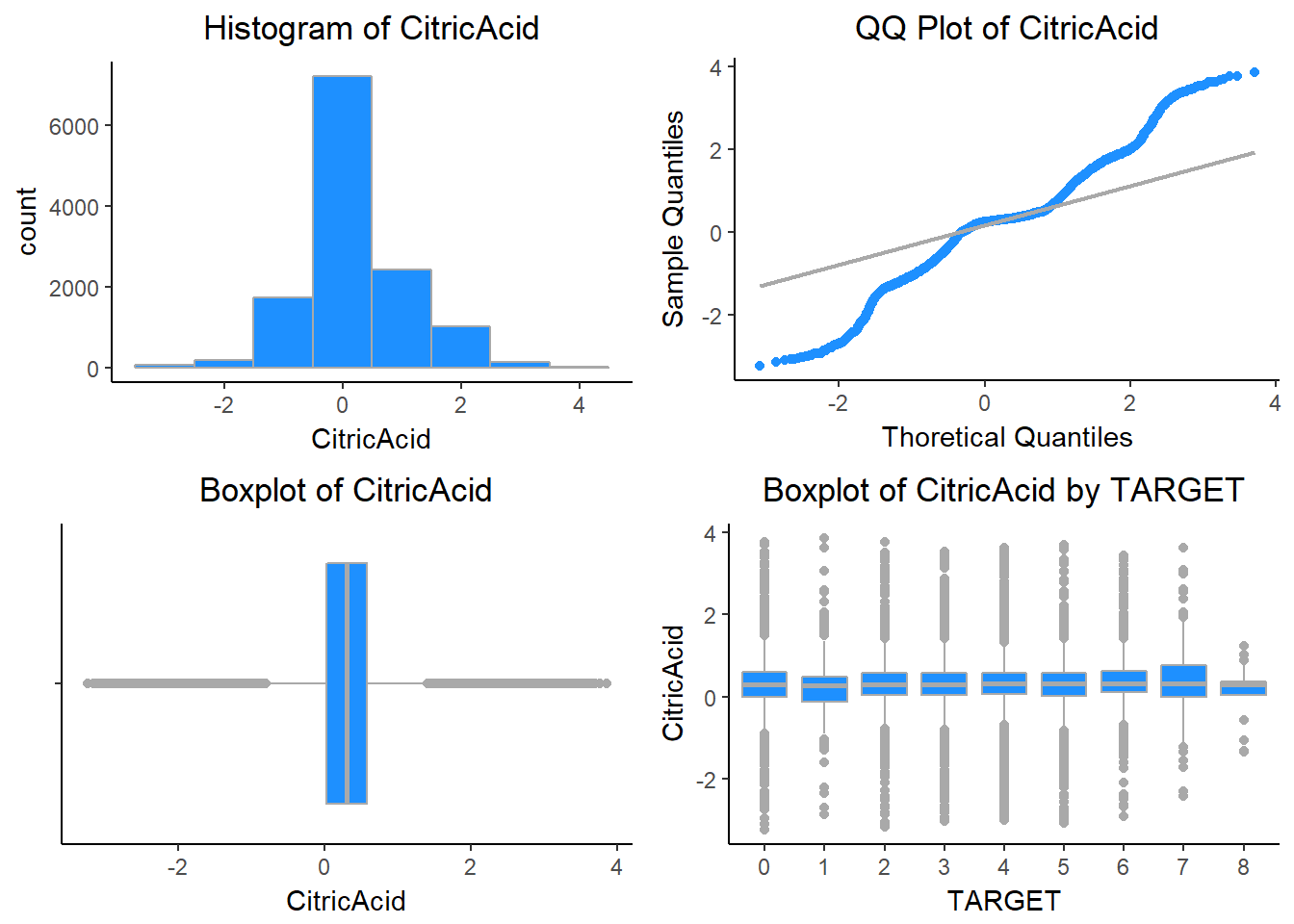
Alcohol - This variable tells us about the Alcohol content. The variable Alcohol is normally distributed.



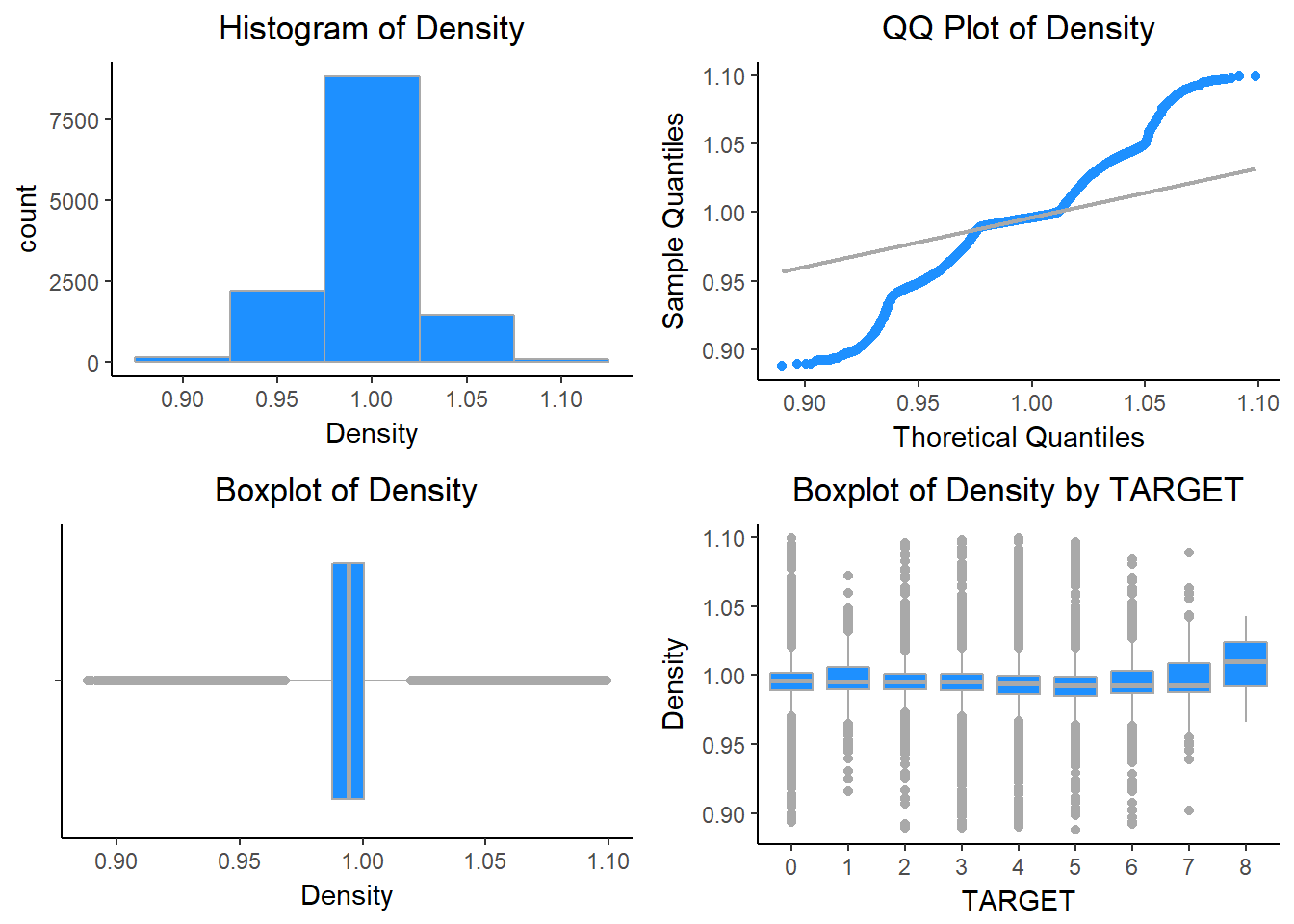
Chlorides - This variable tells us about the Chloride content of wine. The variable Chlorides is normally distributed.



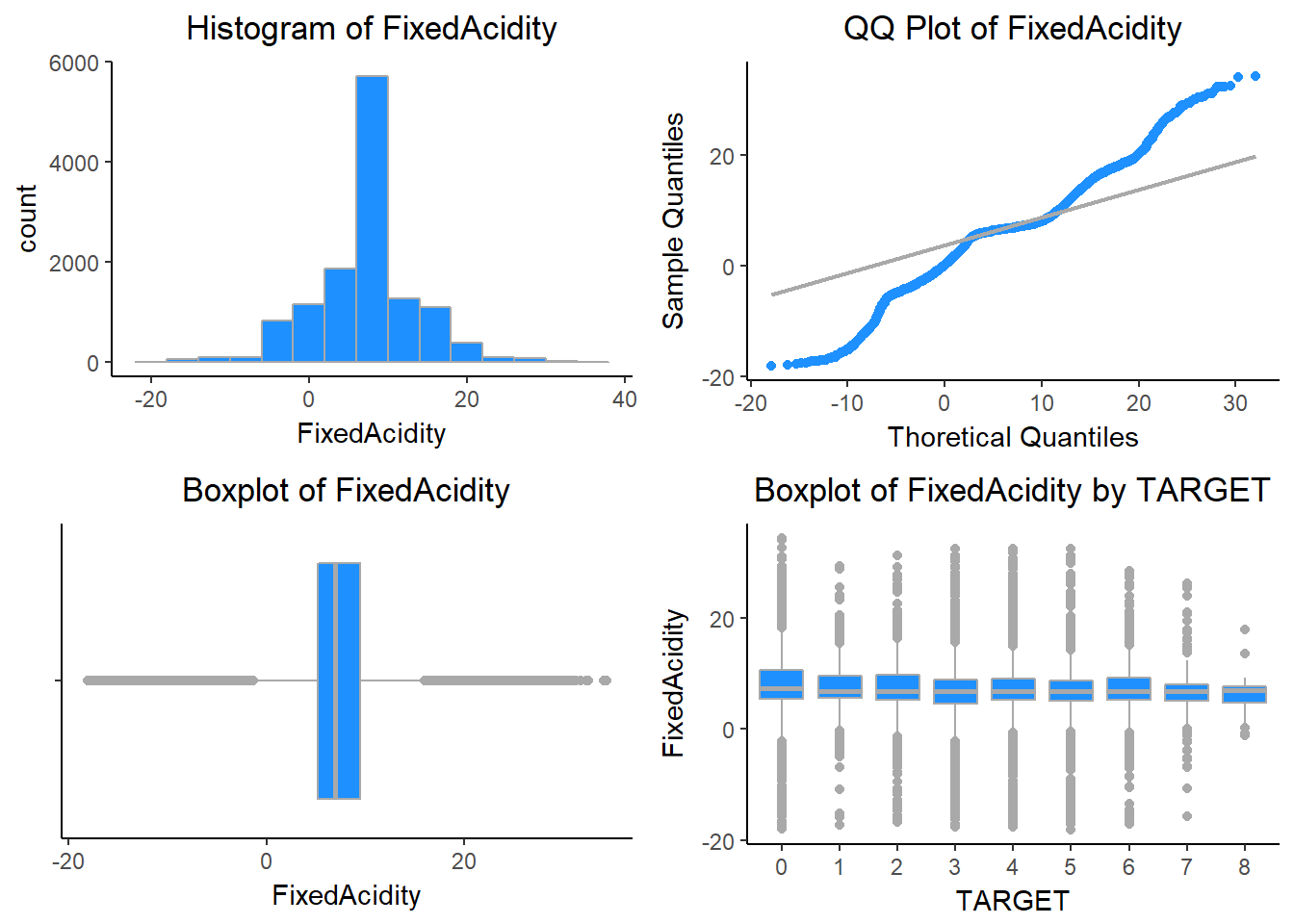
CitricAcid - This variable tells us about the Citric Acid Content of wine. This variable is also normally distributed.



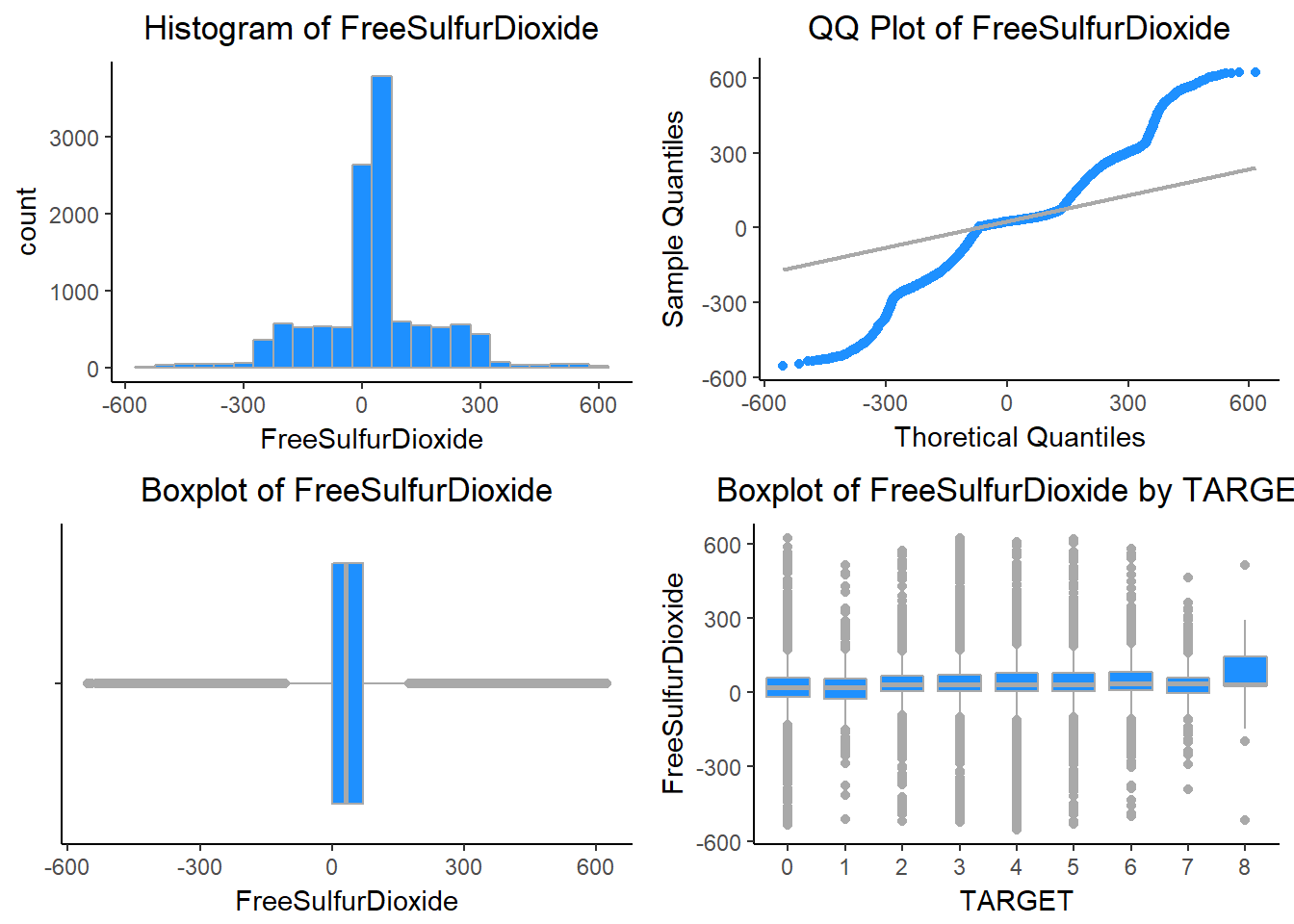
Density - This variable tells us about the Density of wine. Density is also normally distributed.



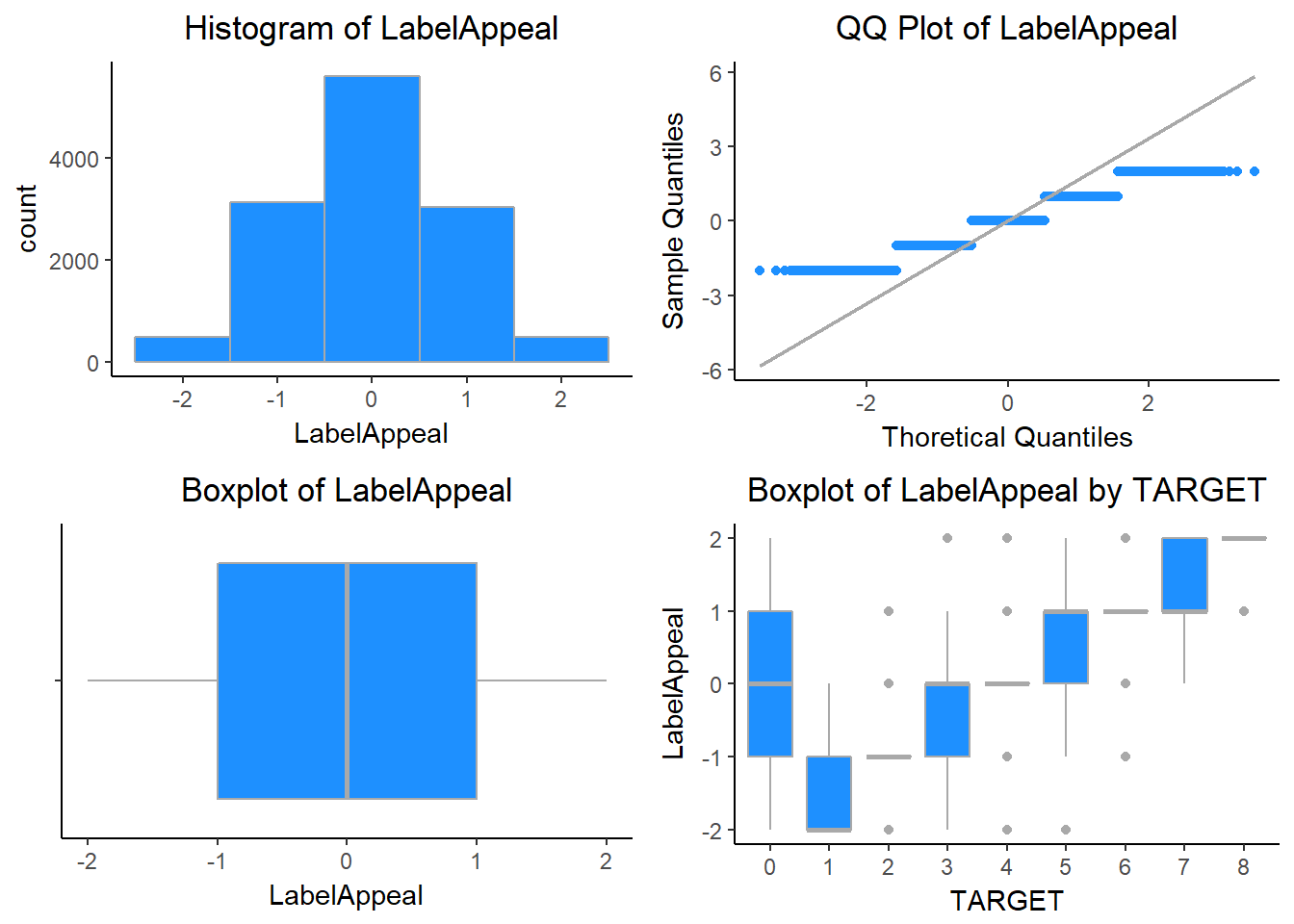
FixedAcidity - This variable tells us about the FixedAcidity of wine. Its also normally distributed.



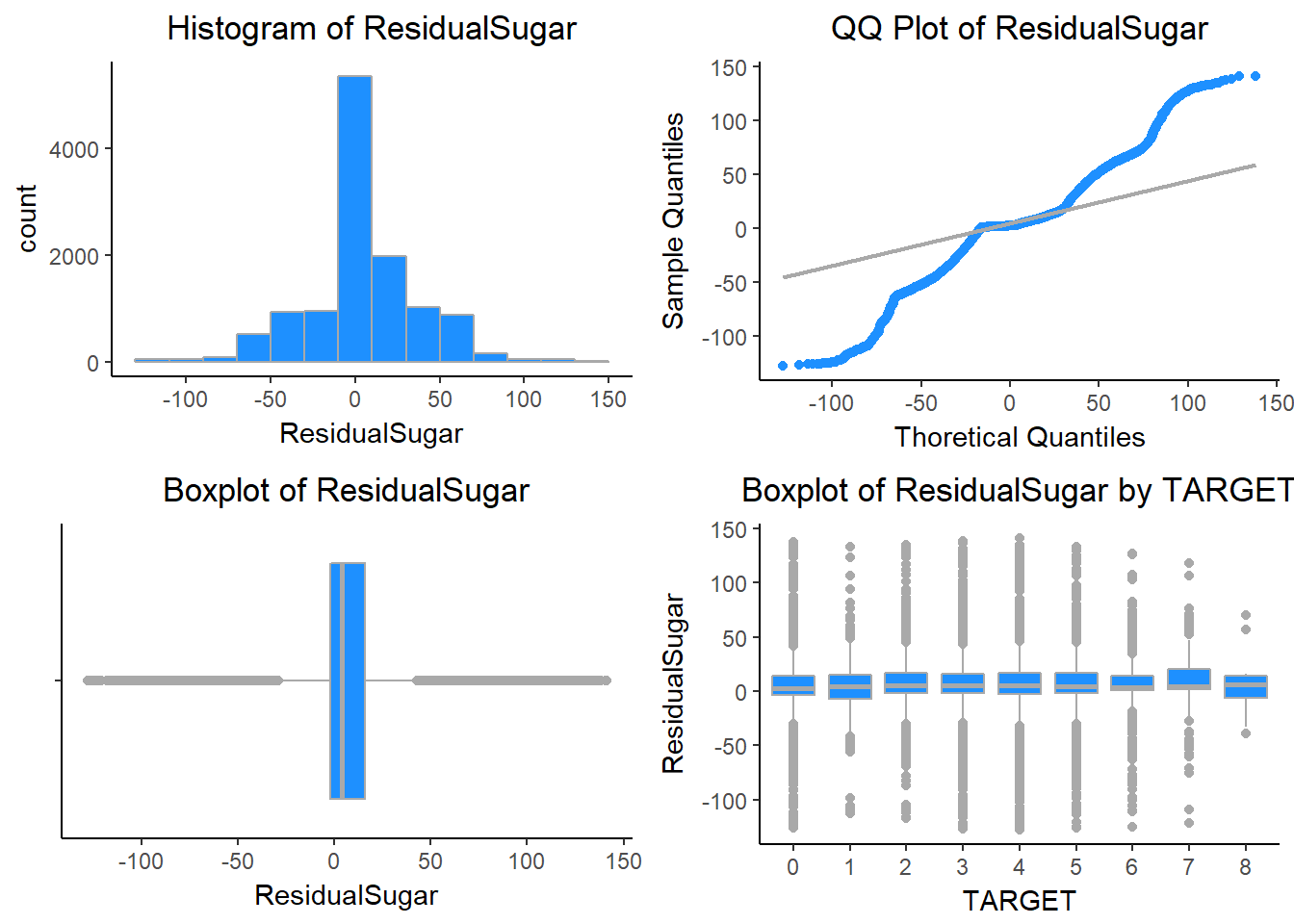
FreeSulfurDioxide - This variable tells us about the Sulfur Dioxide content of wine. It is slightly right skewed.



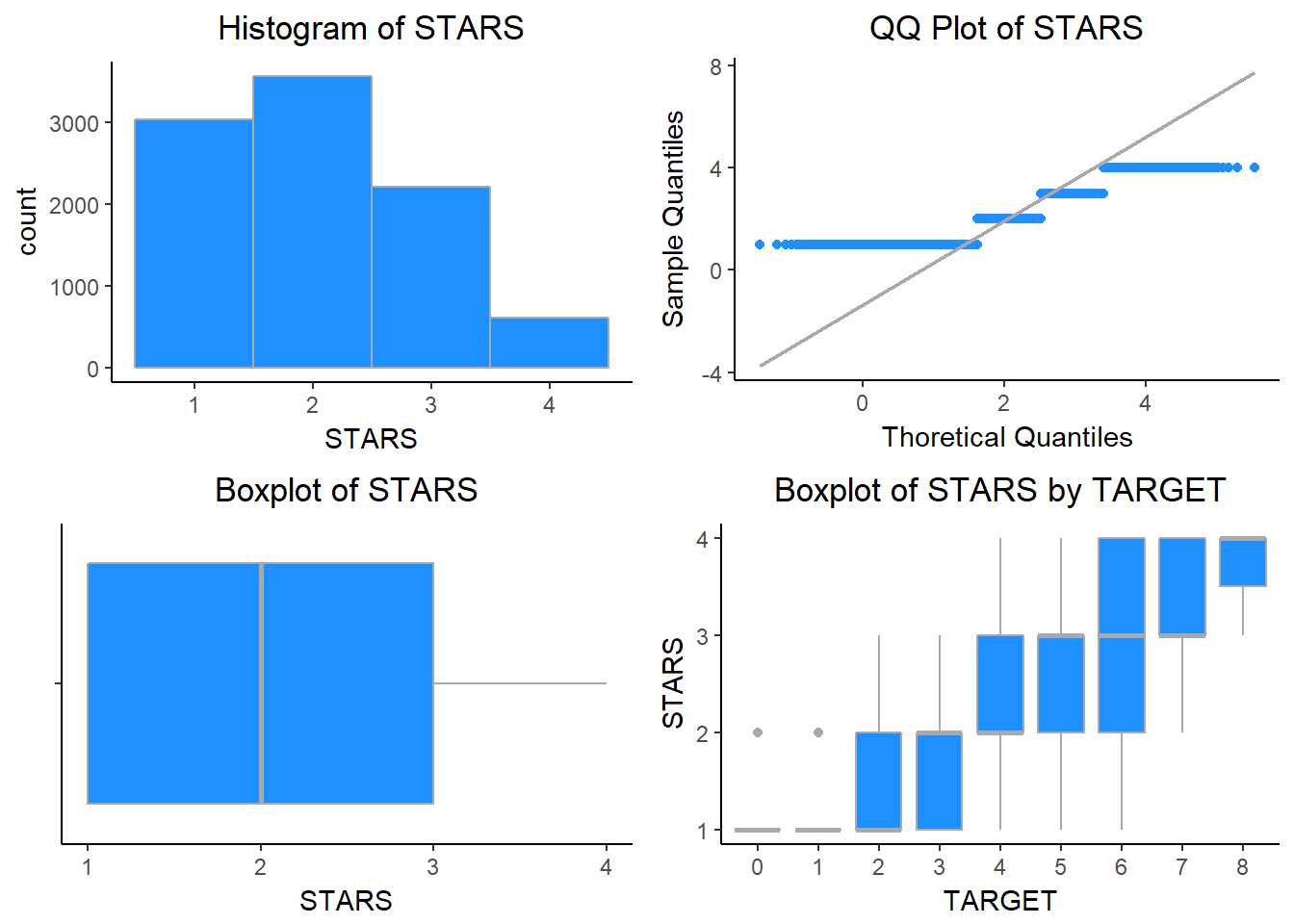
LabelAppeal - Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customers don’t like the design. Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales.



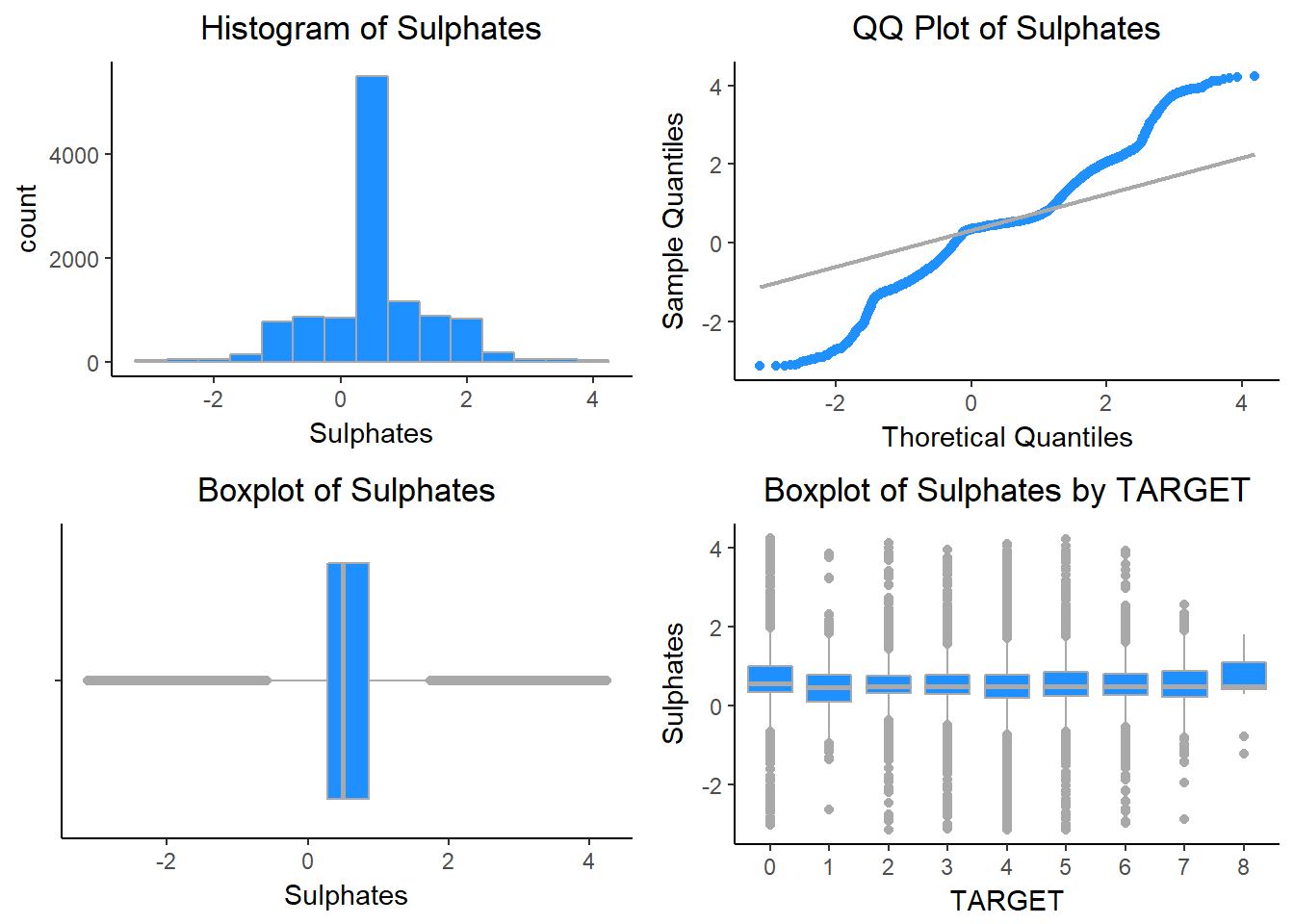
ResidualSugar - This variable tells us about the ResidualSugar of wine. ResidualSugar is also normally distributed.



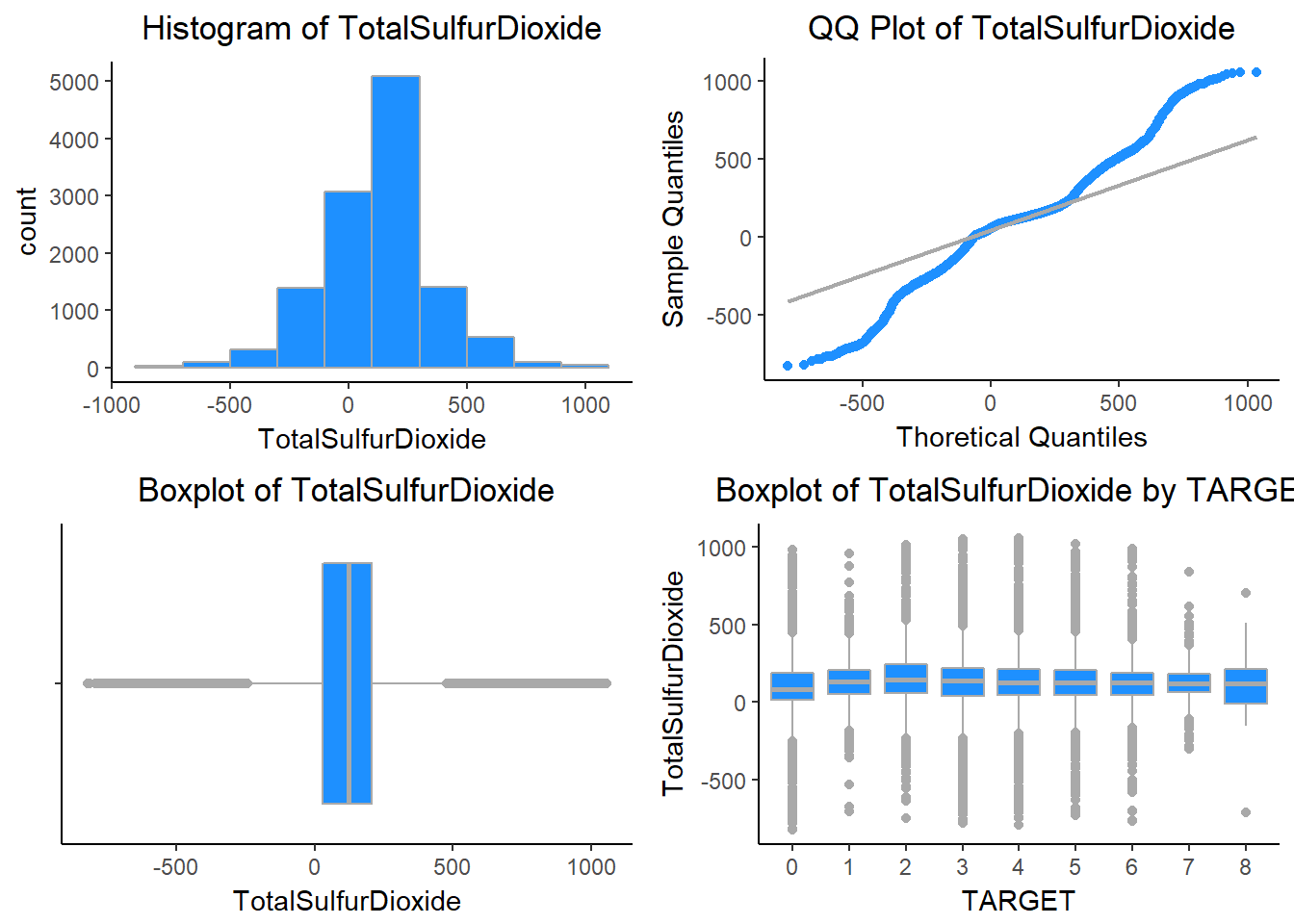
STARS - Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor. A high number of stars suggests high sales.



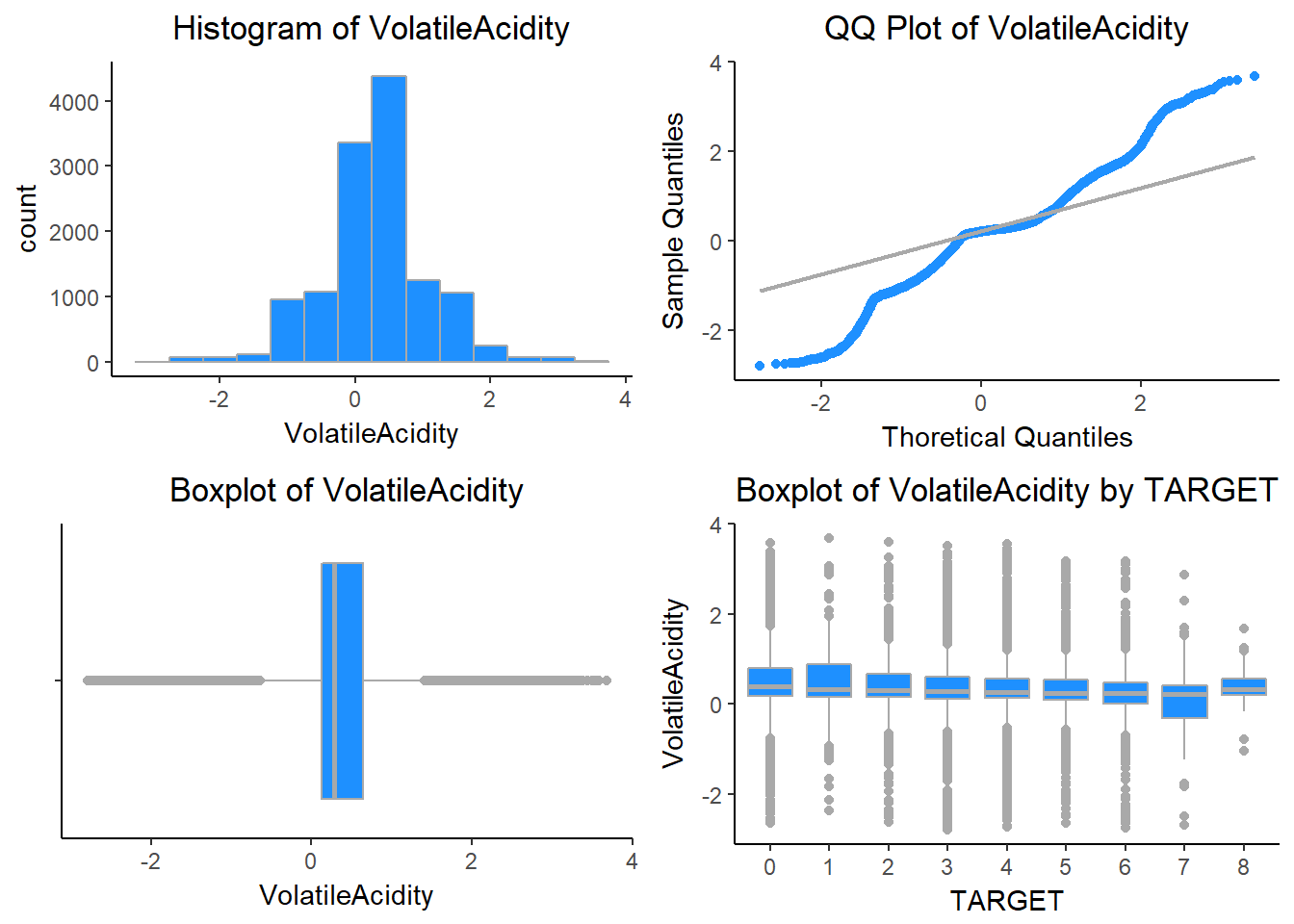
Sulphates - This variable tells us about the Sulphates content of wine.



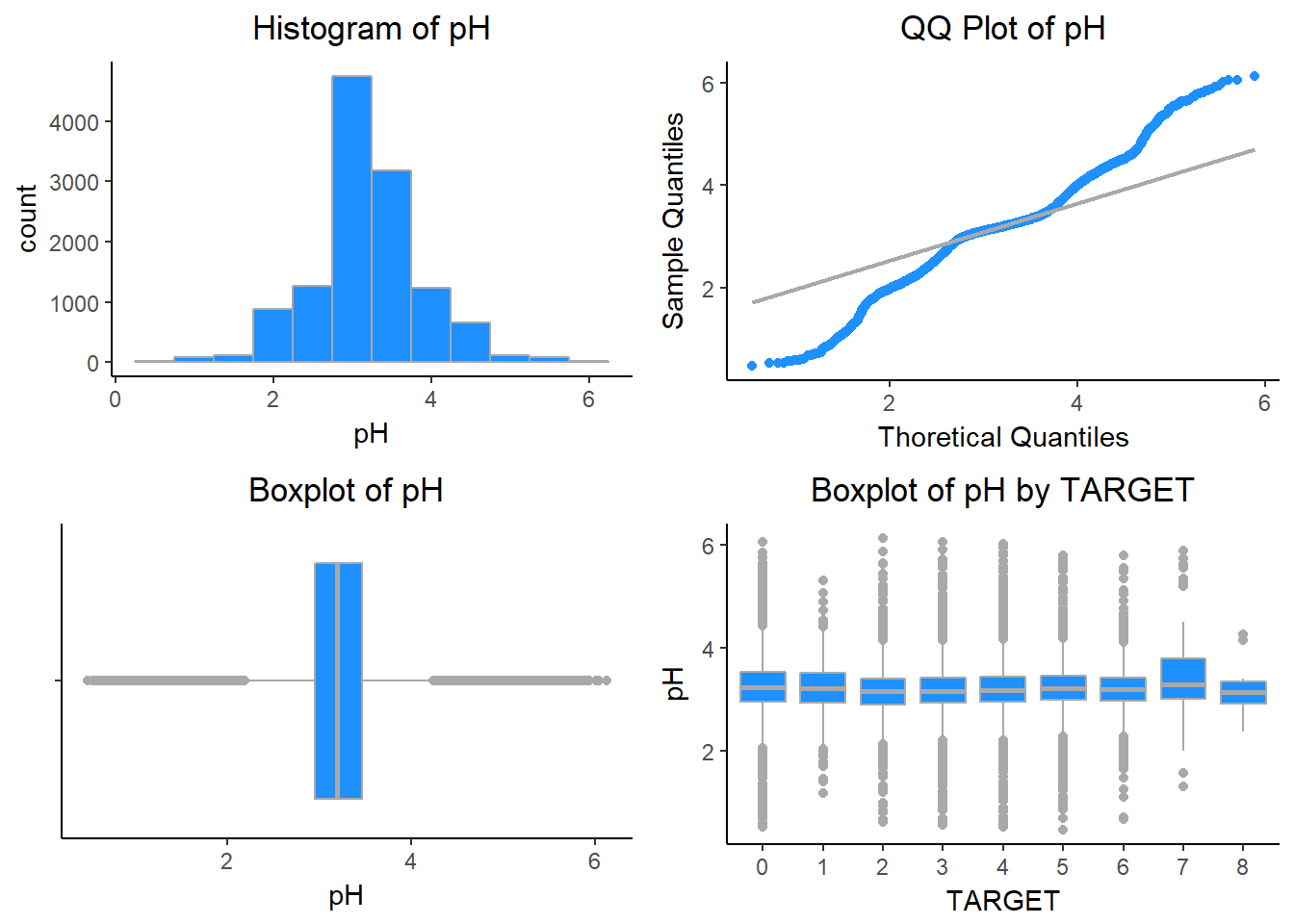
TotalSulfurDioxide - This variable tells us about the Total Sulfur Dioxide of Wine.



VolatileAcidity - This variable tells us about the VolatileAcidity content of Wine.

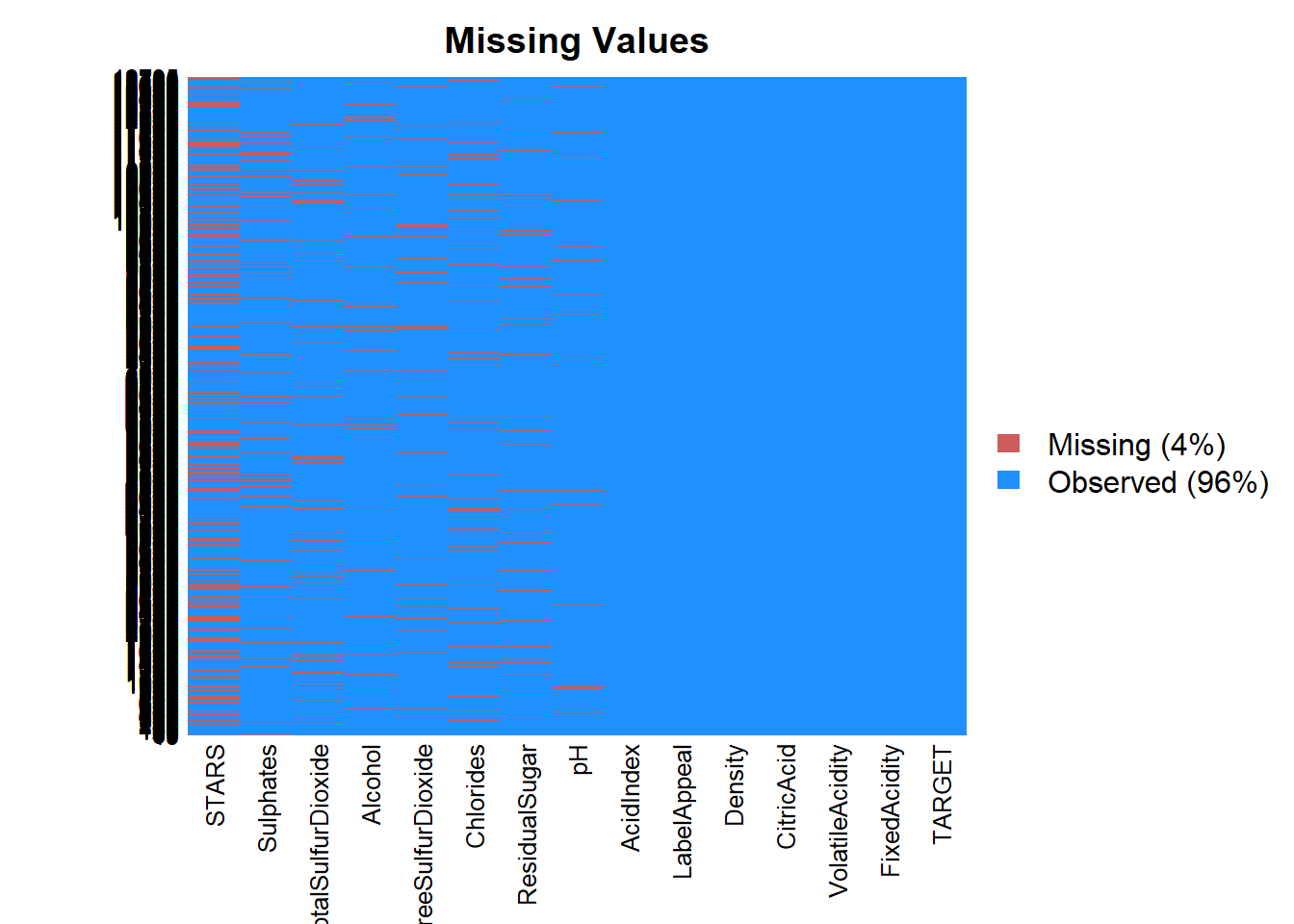


pH - This variable tells us about the pH of Wine.



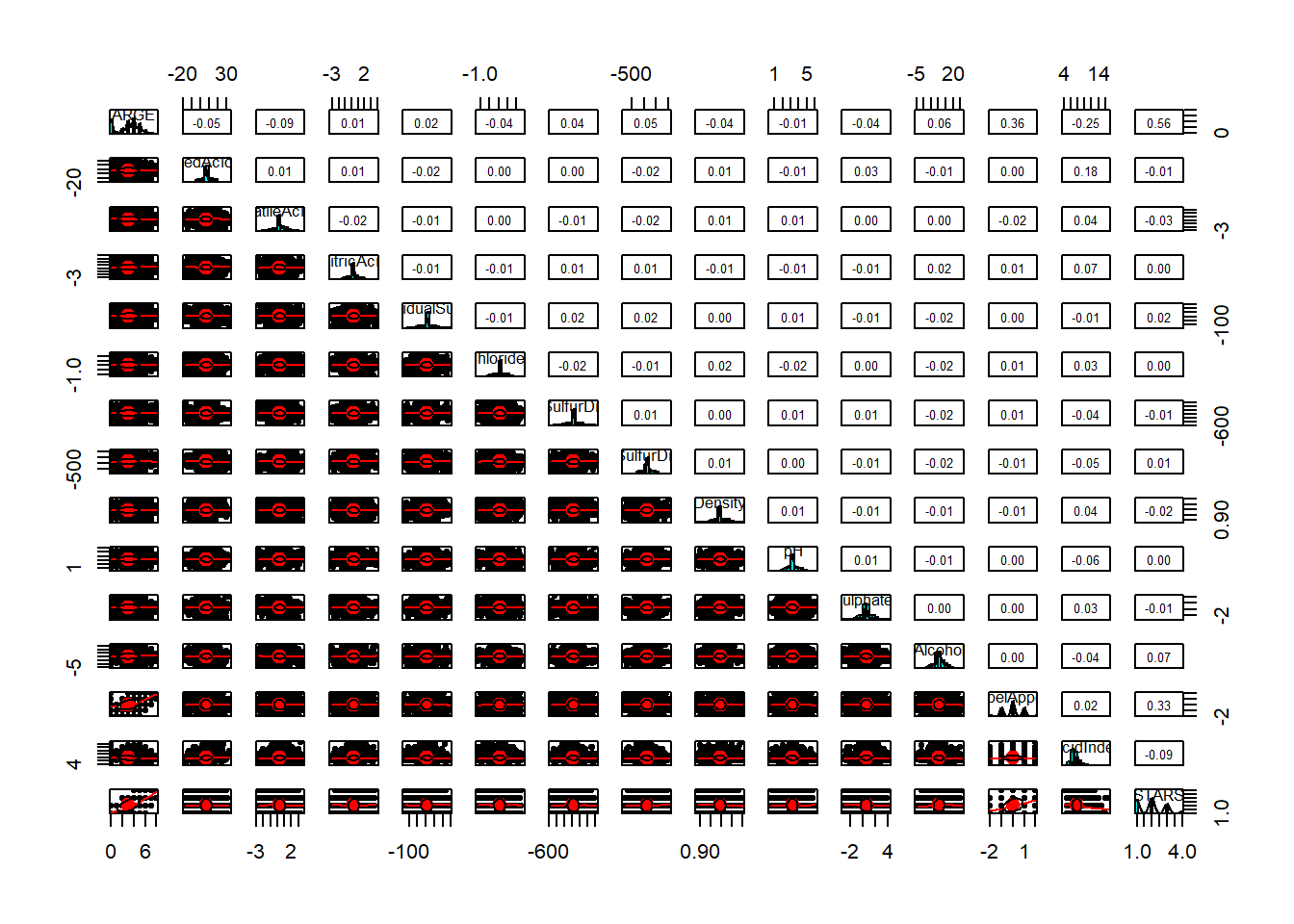
### Missing Values

Now we will see the missing values in the dataset. For this i have used Amelia package. The below table shows a summary of the NA values in the data. Only STARS had an NA frequency higher than 10%, so this was a concern.  There needs to be taken care while we do data preparation.



### Correlations:

Finding correlations: The correlation plot below shows how variables in the dataset are related to each other. Looking at the plot, we don’t see much correlations.



Now we will see the TARGET Variable.

TARGET - Number of Cases Purchased

## 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 2.12

# Data Preparation:

# Let’s first split the data into training and test. We split the data into 80:20.

set.seed(999)

sampl = sample.split(wine\_train$TARGET, SplitRatio = .80)

wine\_train1 <- subset(wine\_train, sampl == TRUE)

wine\_test1 <- subset(wine\_train, sampl == FALSE)

# Also used the mice package to impute missing values. There is very low correlation between AcidIndex and TARGET, so I applied log transformation on AcidIndex.

wine\_train2$AcidIndex <- log(wine\_train2$AcidIndex)

wine\_test2$AcidIndex <- log(wine\_test2$AcidIndex)

# Rest of the data looks good and don’t think we need any transformation.

**Build Models:**

We will build a variety of models using both the imputed and non-imputed data.

1. Poisson model without imputations.

model1 = glm(TARGET ~ ., data=wine\_train1, family=poisson)

summary(model1)

##

## Call:

## glm(formula = TARGET ~ ., family = poisson, data = wine\_train1)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.2128 -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 8.90e-09 \*\*\*

## FixedAcidity 6.705e-04 1.177e-03 0.570 0.56901

## 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 0.93232

## Chlorides -3.764e-02 2.314e-02 -1.627 0.10377

## FreeSulfurDioxide 5.671e-05 4.892e-05 1.159 0.24630

## TotalSulfurDioxide 2.230e-05 3.177e-05 0.702 0.48274

## Density -4.025e-01 2.749e-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 19.530 < 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 poisson family taken to be 1)

##

## Null deviance: 4720.5 on 5143 degrees of freedom

## Residual deviance: 3242.8 on 5129 degrees of freedom

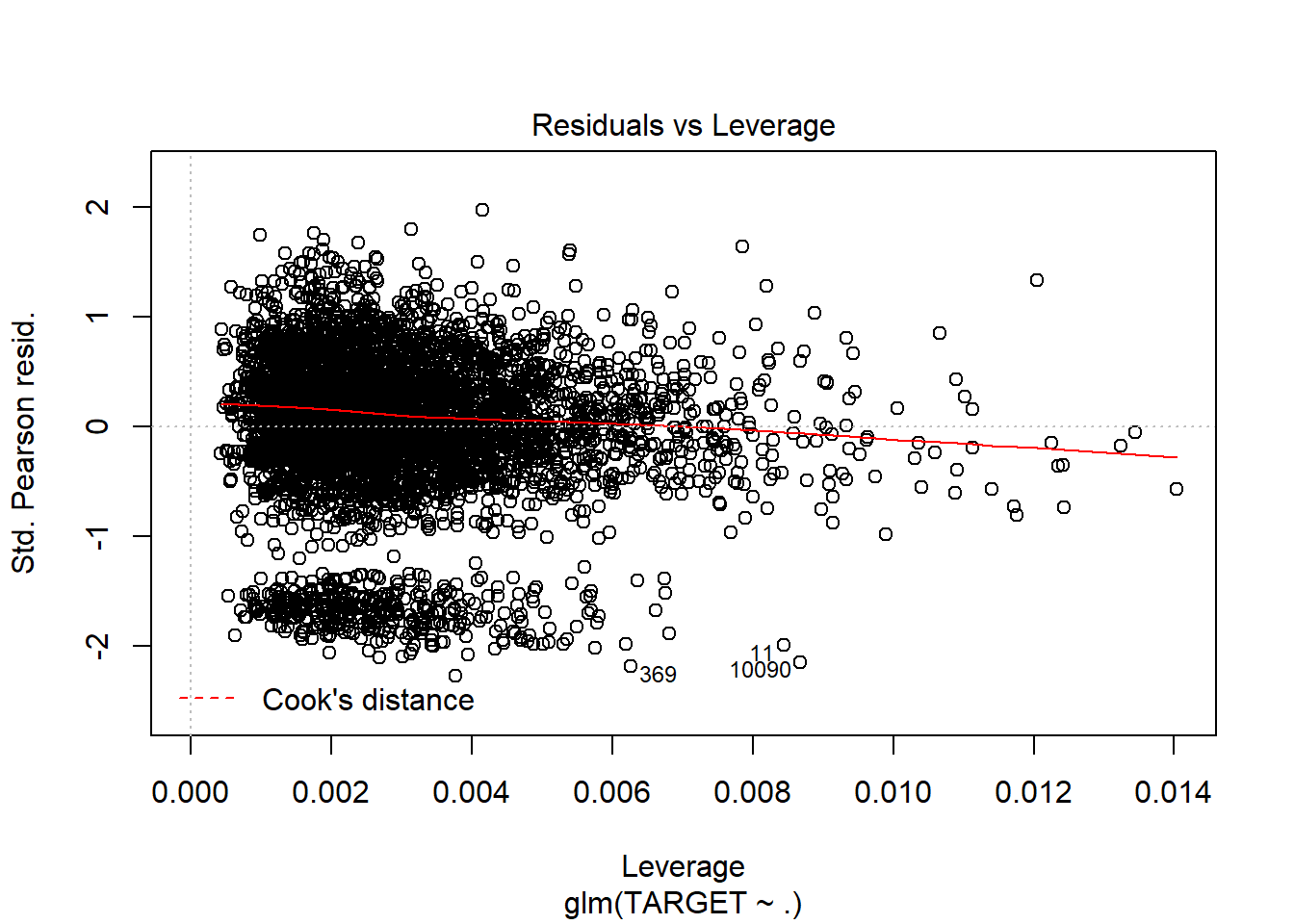
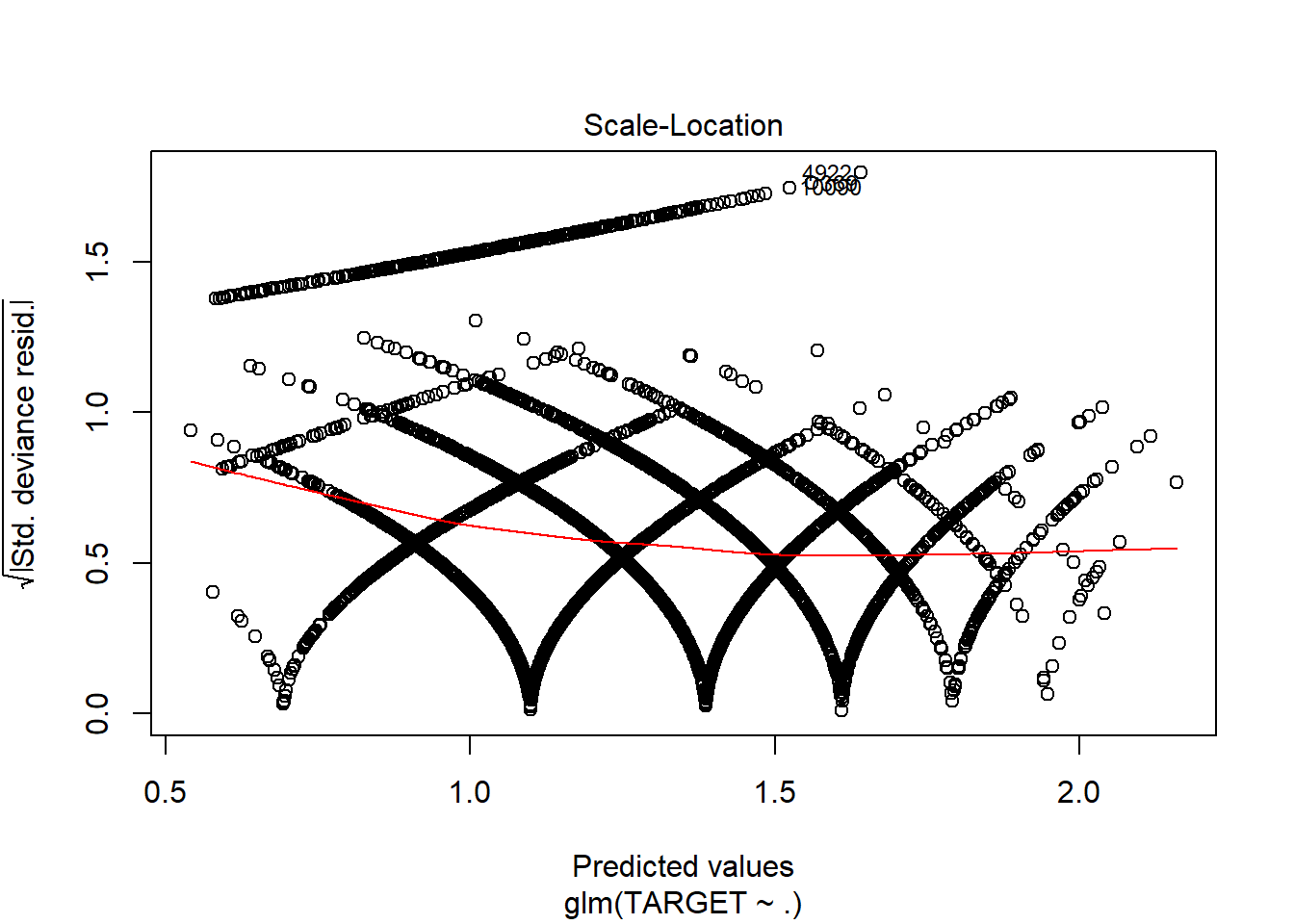
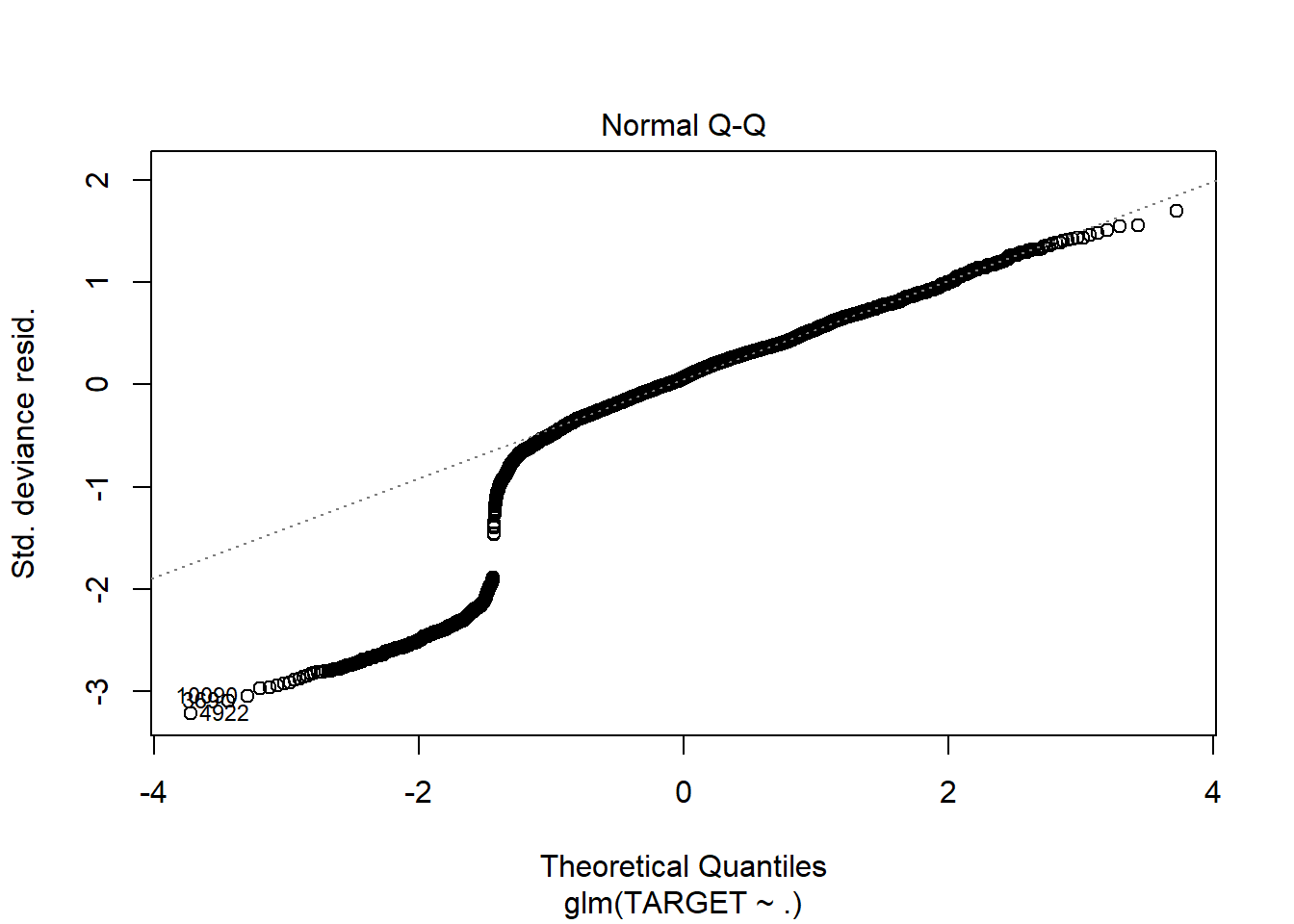
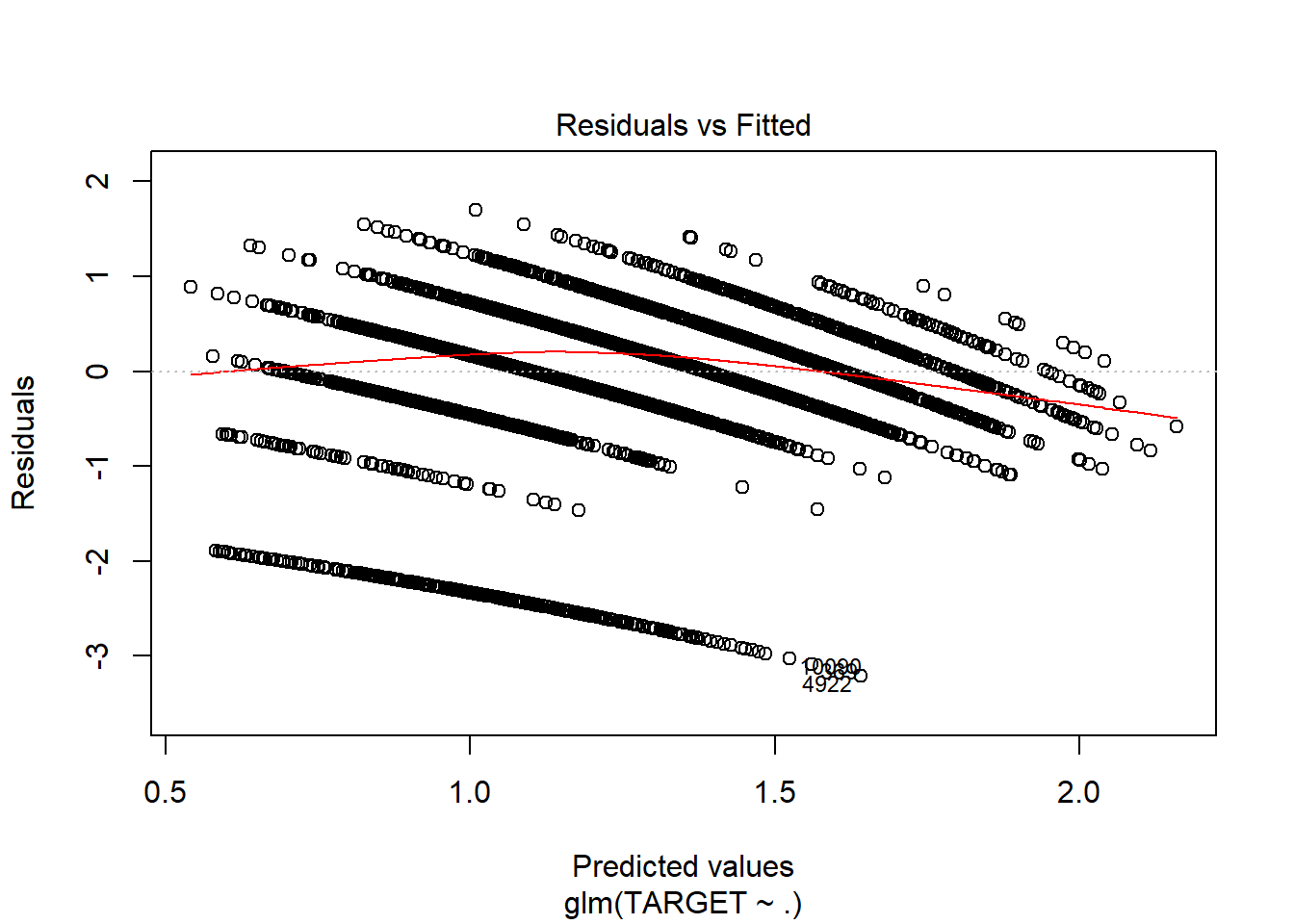
## (5093 observations deleted due to missingness)

## AIC: 18545

##

## Number of Fisher Scoring iterations: 5

plot(model1)

****

1. Poisson model without imputations and only significant variables.

model2 = glm(TARGET ~ .-FixedAcidity-CitricAcid-ResidualSugar-Chlorides-FreeSulfurDioxide-TotalSulfurDioxide-Density-pH-Sulphates-Alcohol, data=wine\_train1, family=poisson)

summary(model2)

##

## Call:

## glm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -

## Chlorides - FreeSulfurDioxide - TotalSulfurDioxide - Density -

## pH - Sulphates - Alcohol, family = poisson, data = wine\_train1)

##

## 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) 1.251442 0.054724 22.868 < 2e-16 \*\*\*

## VolatileAcidity -0.027581 0.009278 -2.973 0.00295 \*\*

## LabelAppeal 0.173177 0.008853 19.562 < 2e-16 \*\*\*

## AcidIndex -0.050616 0.006553 -7.724 1.13e-14 \*\*\*

## STARS 0.194208 0.008292 23.421 < 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: 4720.5 on 5143 degrees of freedom

## Residual deviance: 3253.1 on 5139 degrees of freedom

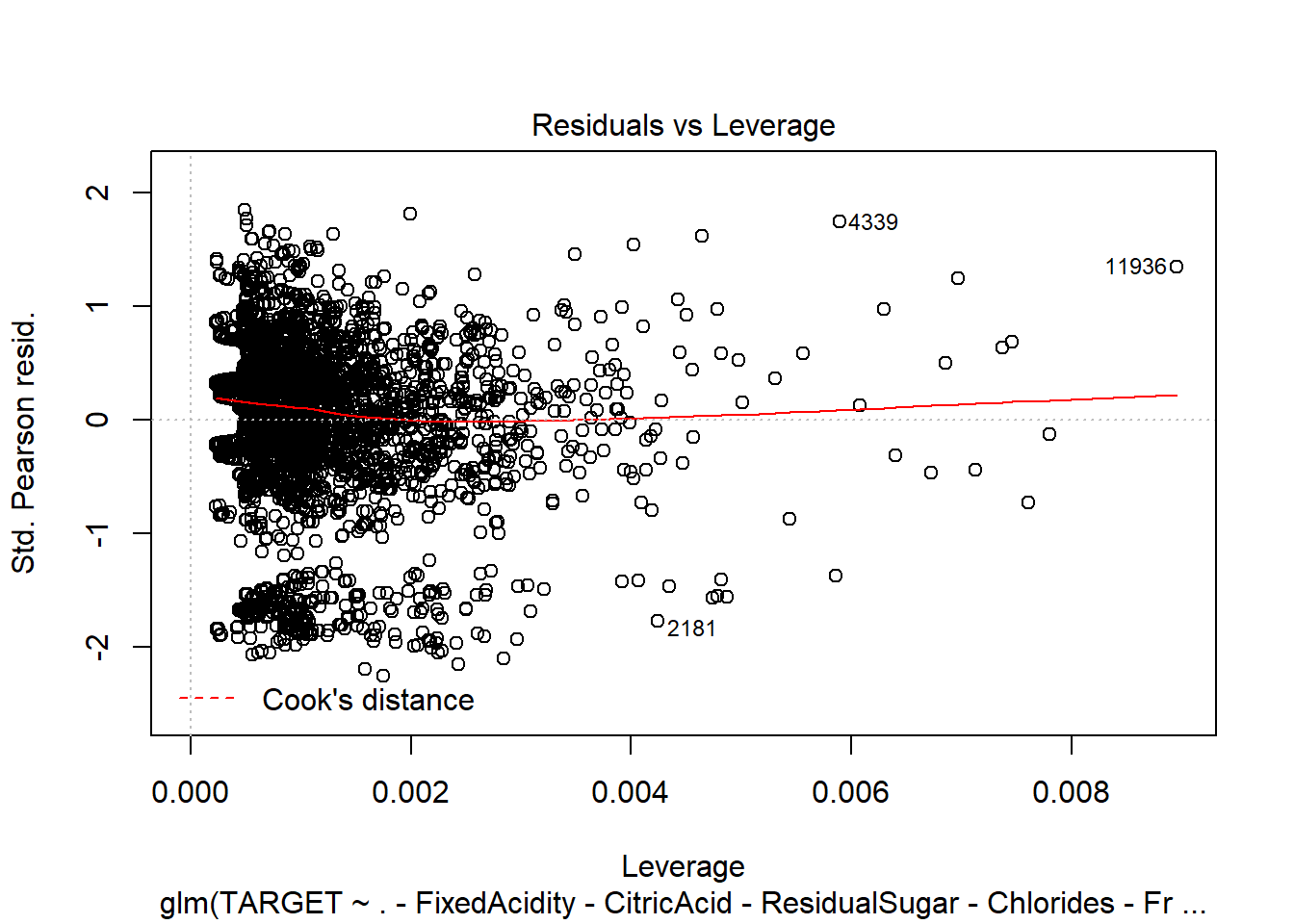
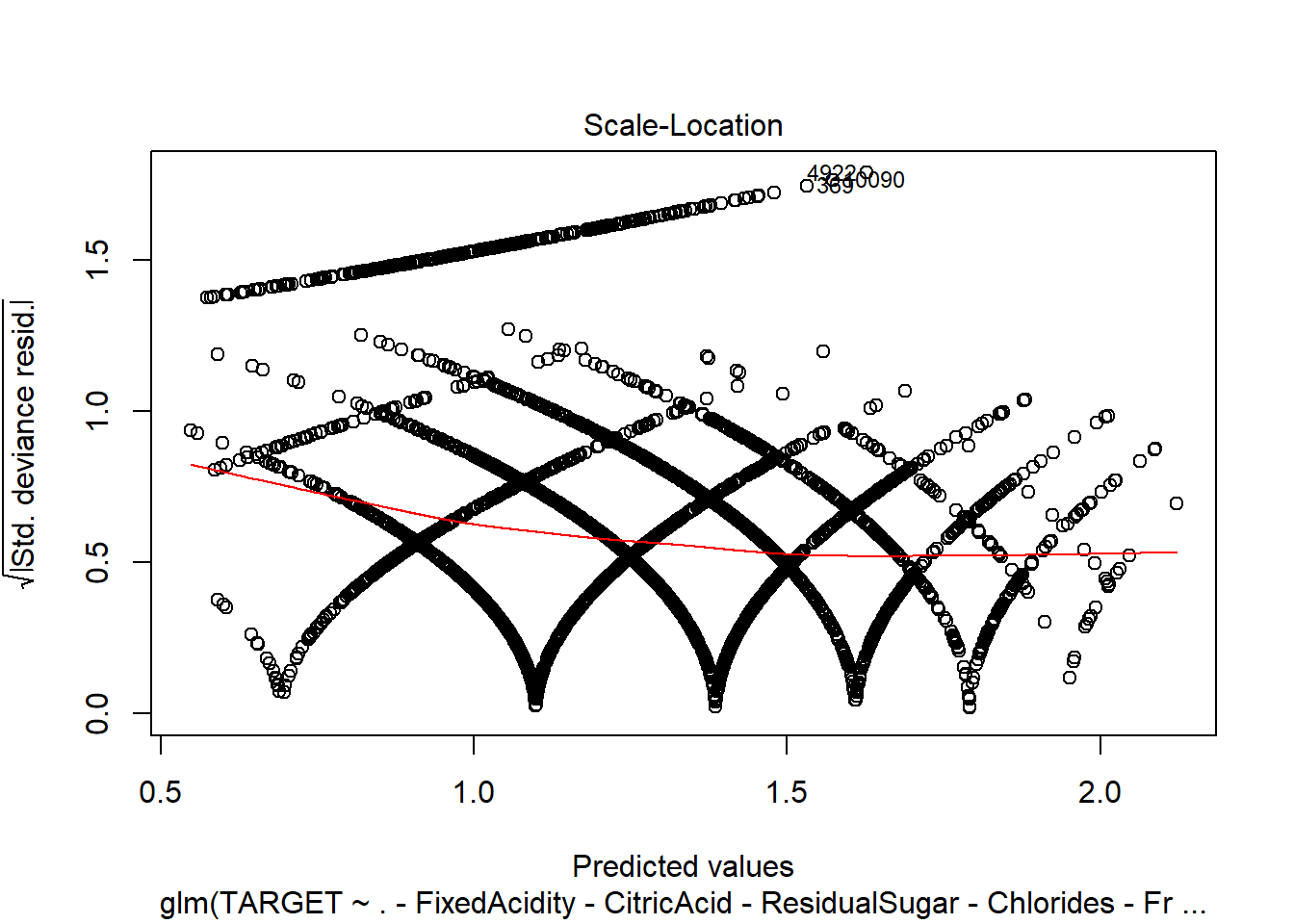
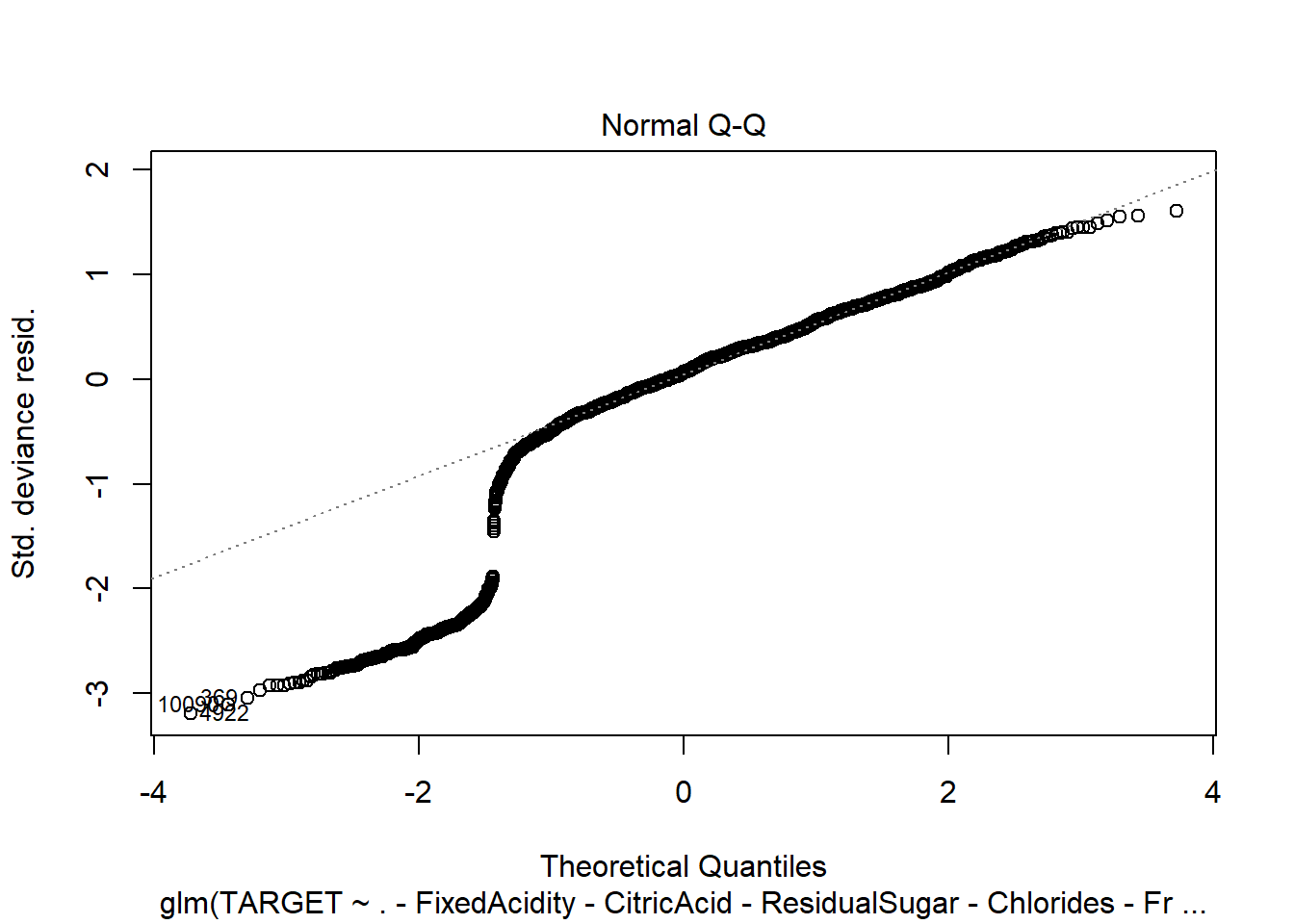
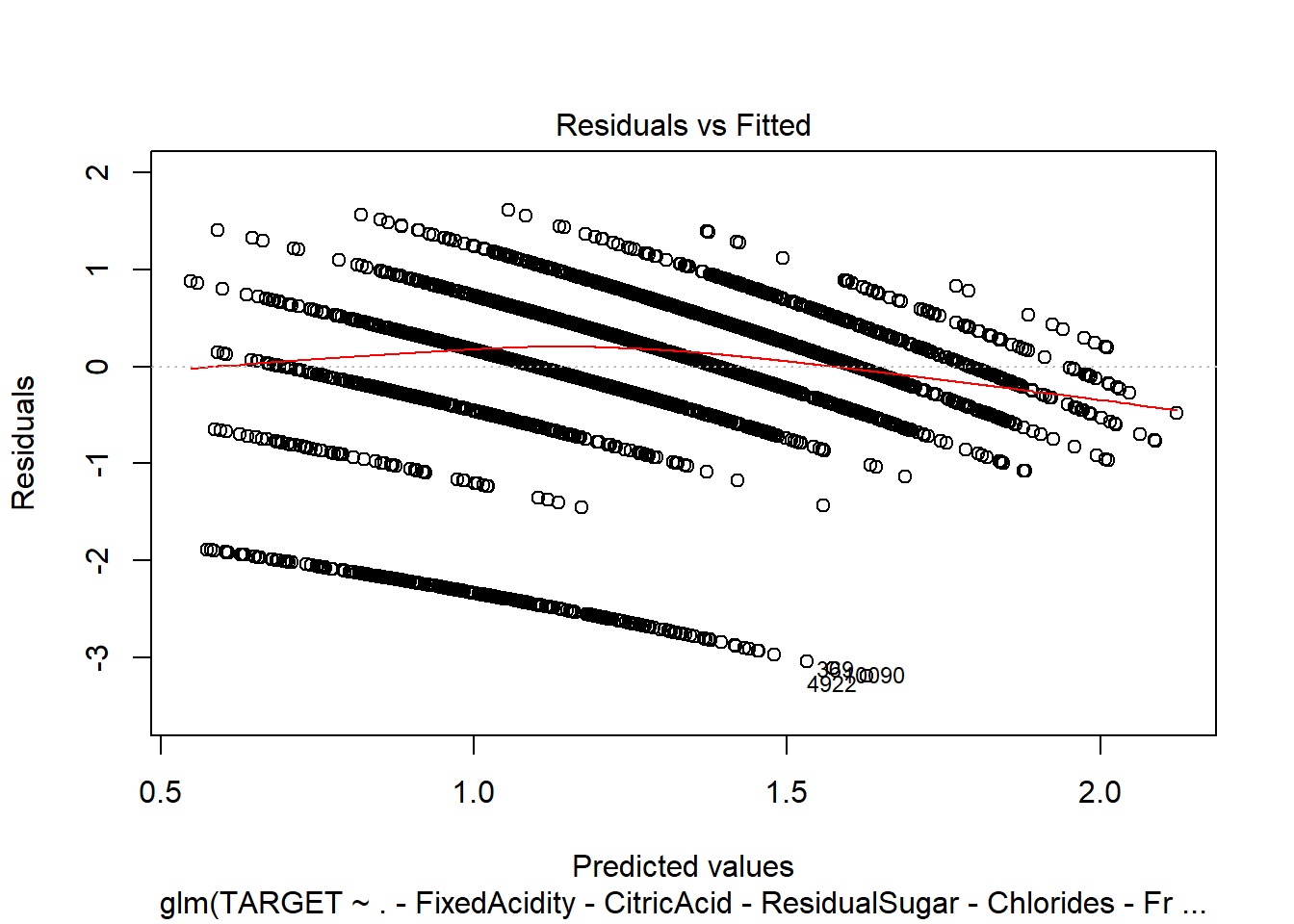
## (5093 observations deleted due to missingness)

## AIC: 18535

##

## Number of Fisher Scoring iterations: 5

plot(model2)

****

1. Poisson model with Imputation.

model3 = glm(TARGET ~ ., data=wine\_train2, family=poisson)

summary(model3)

##

## Call:

## glm(formula = TARGET ~ ., family = poisson, data = wine\_train2)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.1516 -0.6809 0.1304 0.6390 2.4033

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 2.382e+00 2.277e-01 10.463 < 2e-16 \*\*\*

## FixedAcidity -1.332e-04 9.197e-04 -0.145 0.88487

## VolatileAcidity -4.351e-02 7.275e-03 -5.982 2.21e-09 \*\*\*

## CitricAcid 8.883e-03 6.576e-03 1.351 0.17679

## ResidualSugar 1.508e-04 1.675e-04 0.900 0.36797

## Chlorides -6.506e-02 1.791e-02 -3.633 0.00028 \*\*\*

## FreeSulfurDioxide 1.143e-04 3.804e-05 3.005 0.00266 \*\*

## TotalSulfurDioxide 8.709e-05 2.446e-05 3.560 0.00037 \*\*\*

## Density -4.047e-01 2.141e-01 -1.890 0.05876 .

## pH -1.788e-02 8.407e-03 -2.126 0.03347 \*

## Sulphates -1.327e-02 6.163e-03 -2.153 0.03129 \*

## Alcohol 2.690e-03 1.546e-03 1.740 0.08187 .

## LabelAppeal 1.432e-01 6.783e-03 21.107 < 2e-16 \*\*\*

## AcidIndex -7.622e-01 4.005e-02 -19.029 < 2e-16 \*\*\*

## STARS 3.401e-01 6.252e-03 54.395 < 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: 18291 on 10236 degrees of freedom

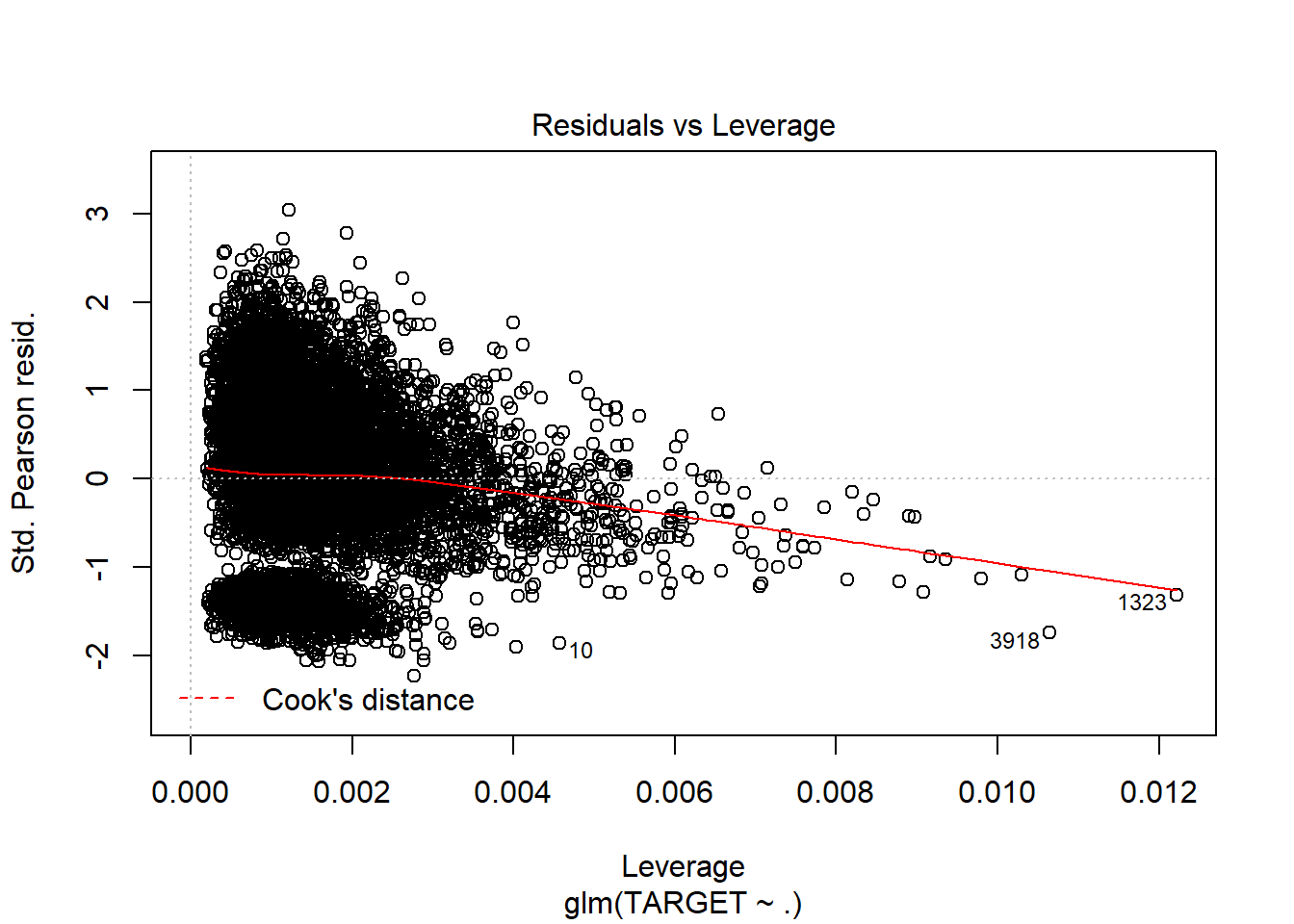
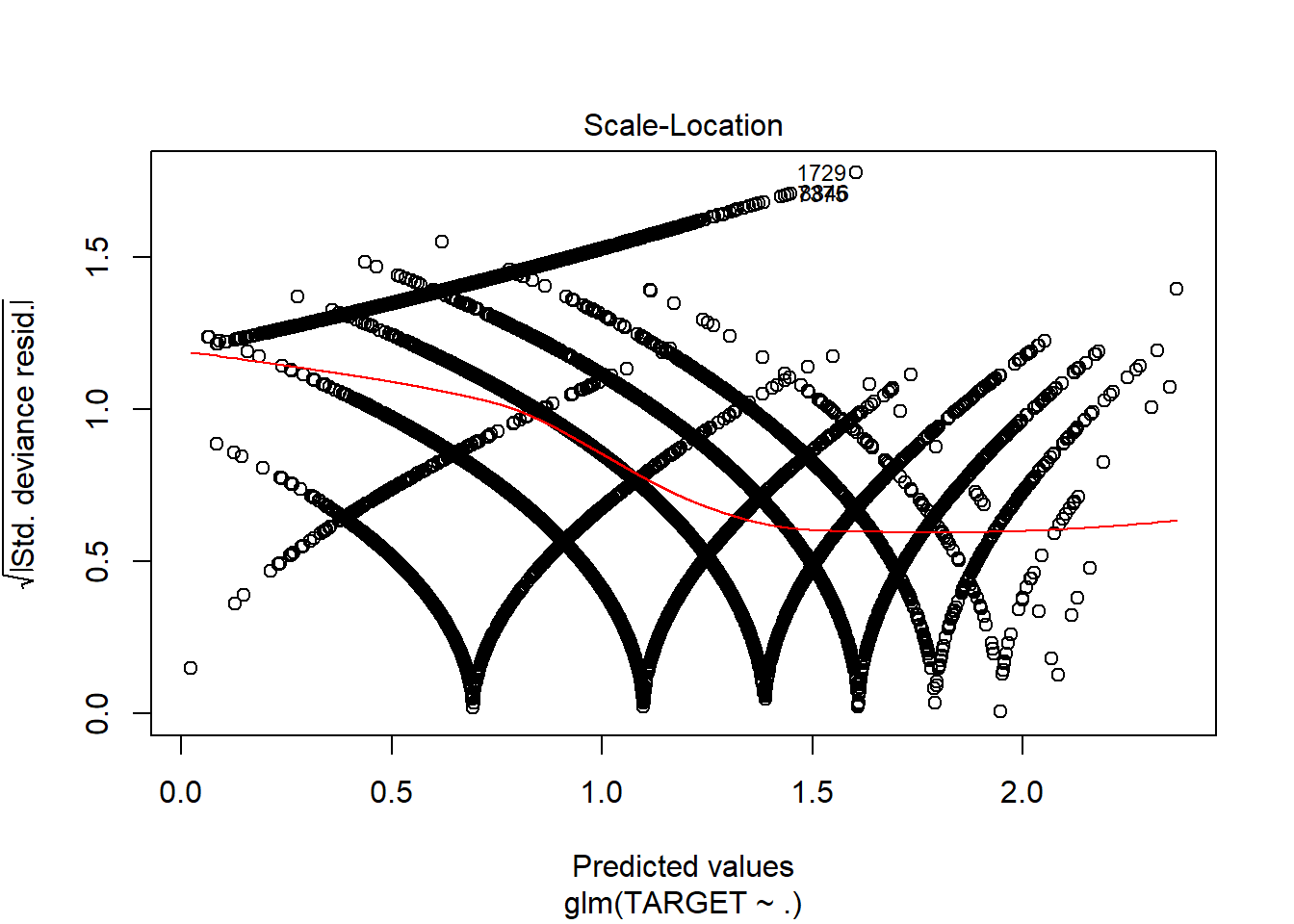
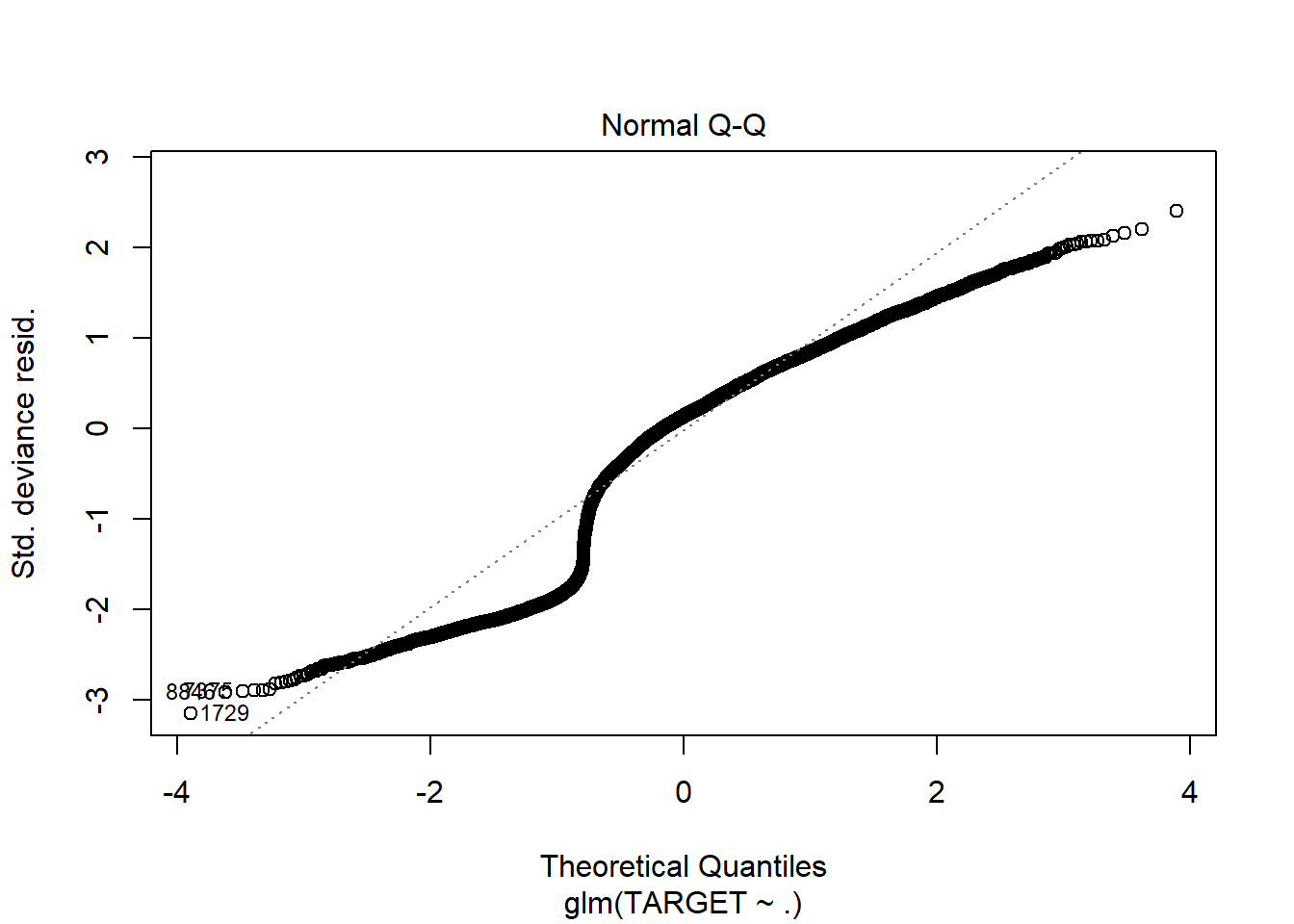
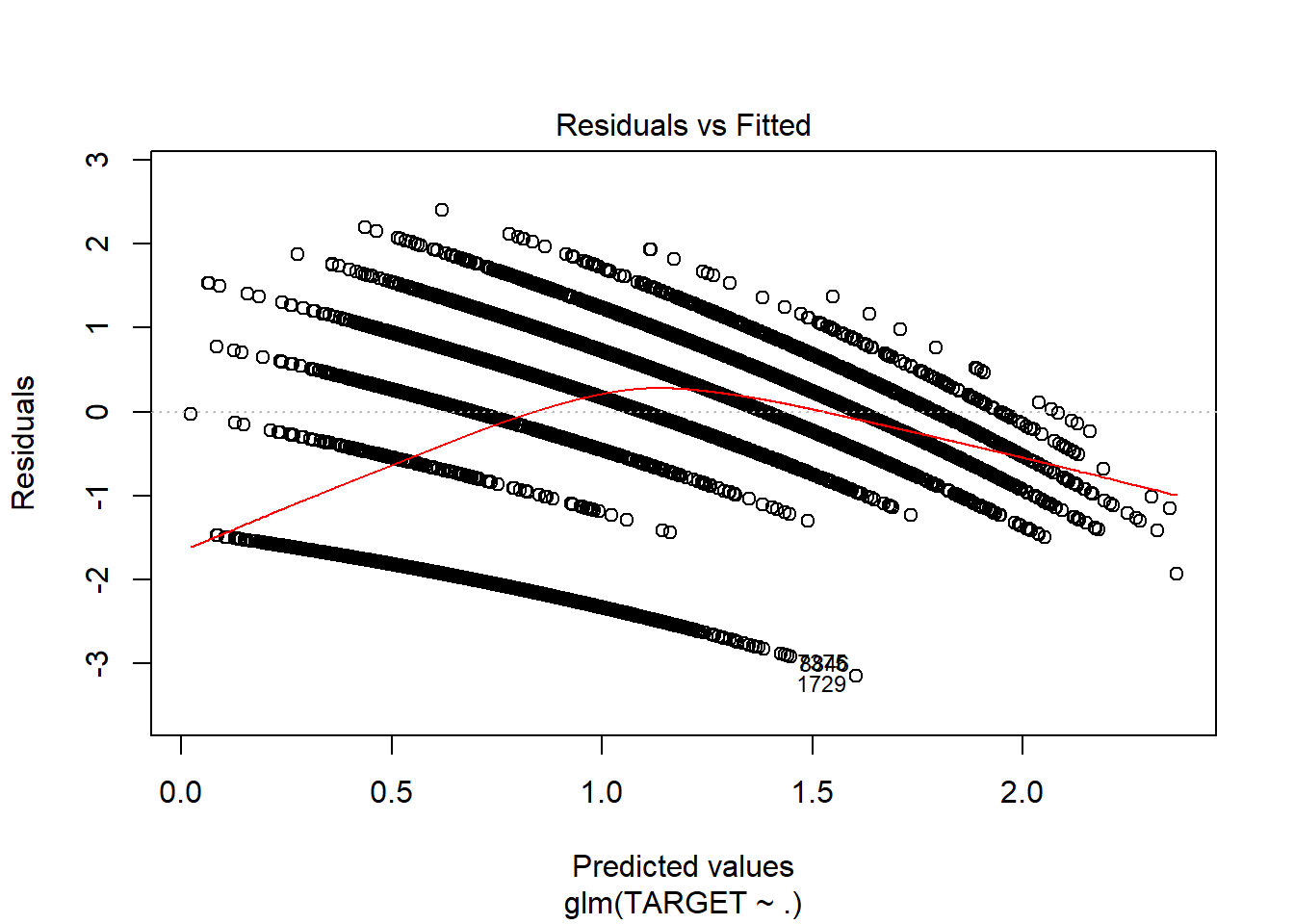
## Residual deviance: 12830 on 10222 degrees of freedom

## AIC: 38418

##

## Number of Fisher Scoring iterations: 5

plot(model3)

****

1. Poisson model with imputations and only significant variables.

model4 = glm(TARGET ~ .-FixedAcidity-CitricAcid-ResidualSugar-Density-Alcohol, data=wine\_train2, family=poisson)

summary(model4)

##

## Call:

## glm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -

## Density - Alcohol, family = poisson, data = wine\_train2)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.1405 -0.6852 0.1288 0.6412 2.4039

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 2.019e+00 8.848e-02 22.820 < 2e-16 \*\*\*

## VolatileAcidity -4.388e-02 7.273e-03 -6.033 1.61e-09 \*\*\*

## Chlorides -6.711e-02 1.790e-02 -3.750 0.000177 \*\*\*

## FreeSulfurDioxide 1.119e-04 3.802e-05 2.943 0.003256 \*\*

## TotalSulfurDioxide 8.560e-05 2.442e-05 3.505 0.000457 \*\*\*

## pH -1.818e-02 8.404e-03 -2.164 0.030488 \*

## Sulphates -1.327e-02 6.157e-03 -2.155 0.031143 \*

## LabelAppeal 1.433e-01 6.783e-03 21.120 < 2e-16 \*\*\*

## AcidIndex -7.665e-01 3.941e-02 -19.448 < 2e-16 \*\*\*

## STARS 3.410e-01 6.237e-03 54.673 < 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: 18291 on 10236 degrees of freedom

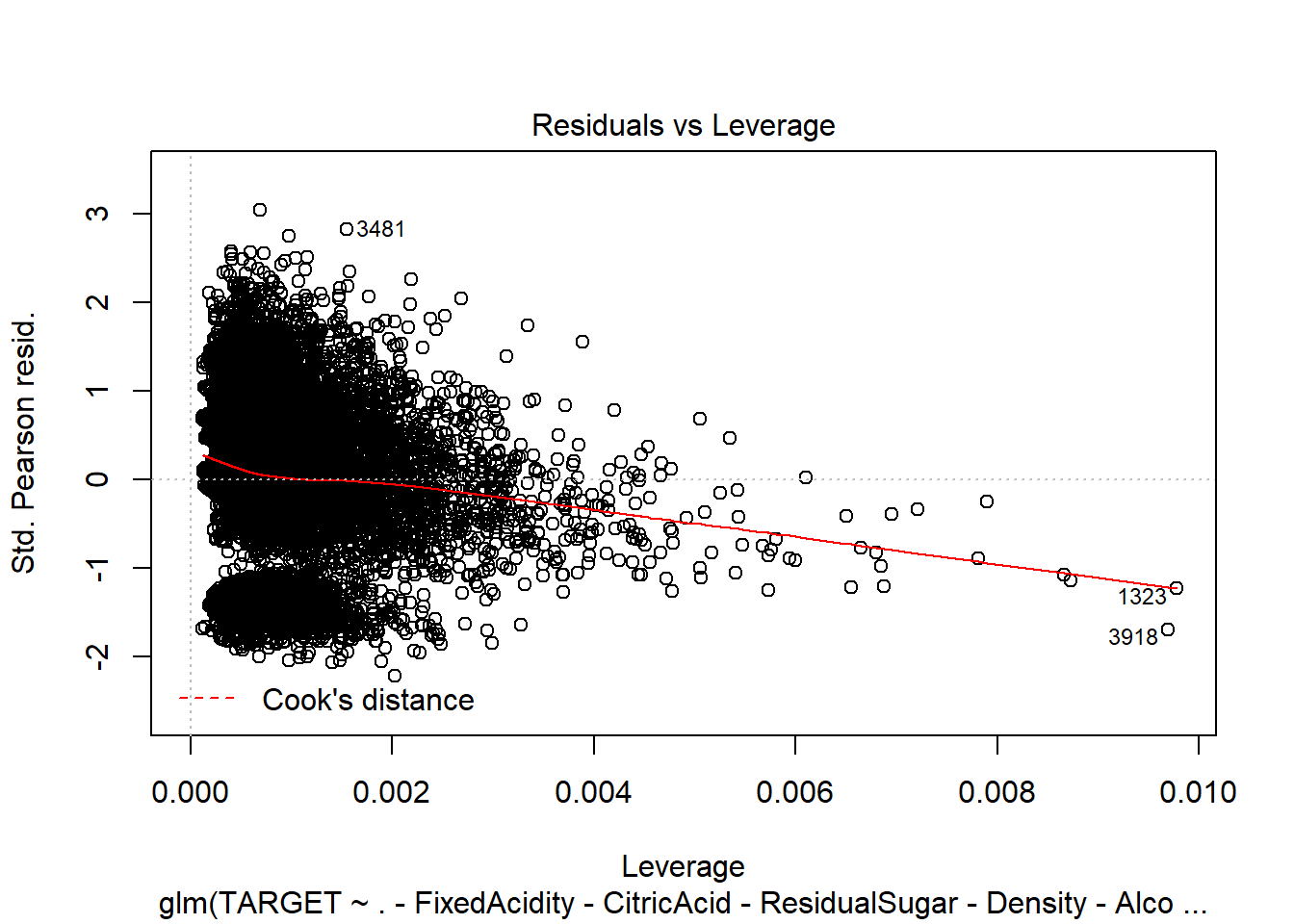
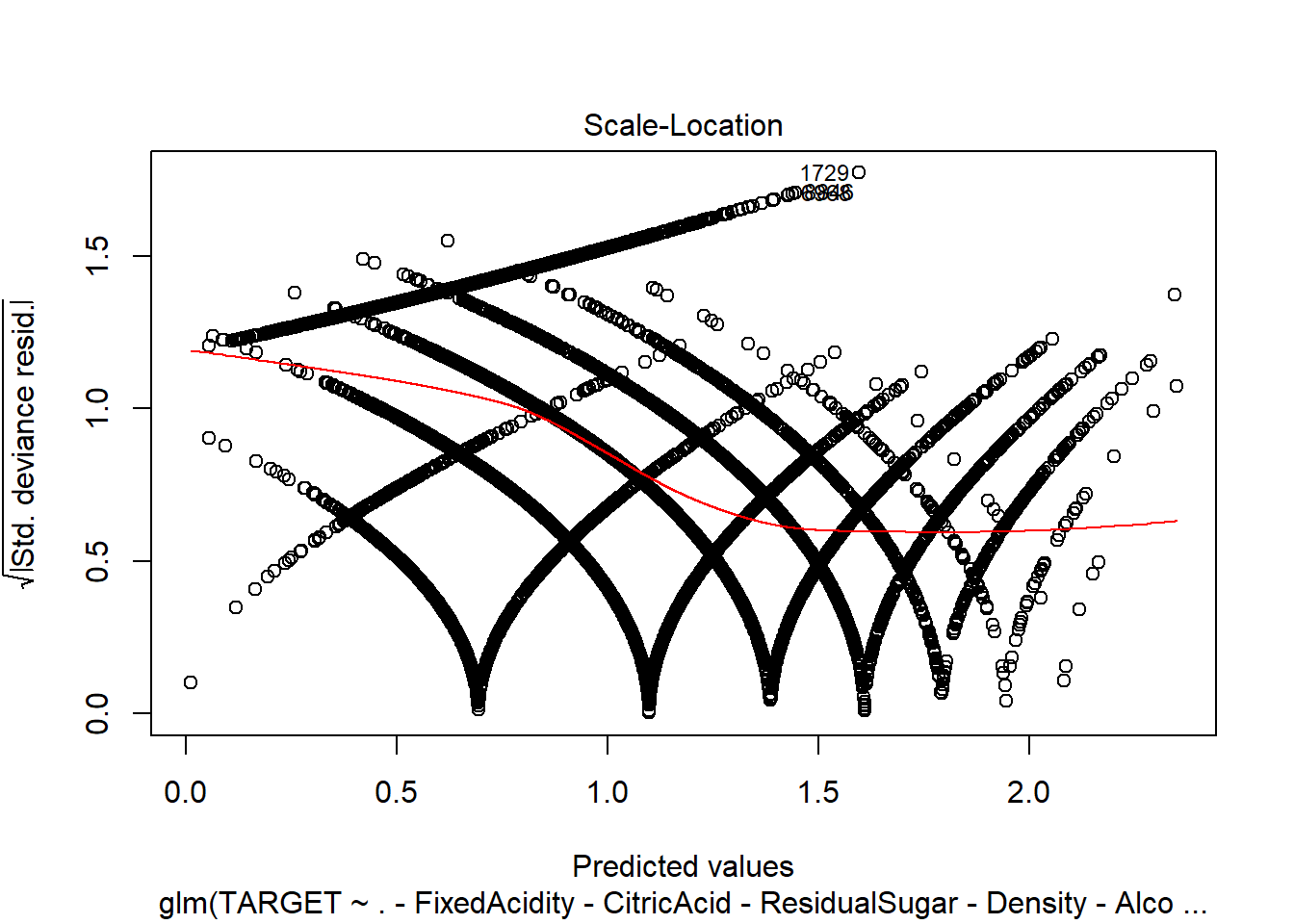
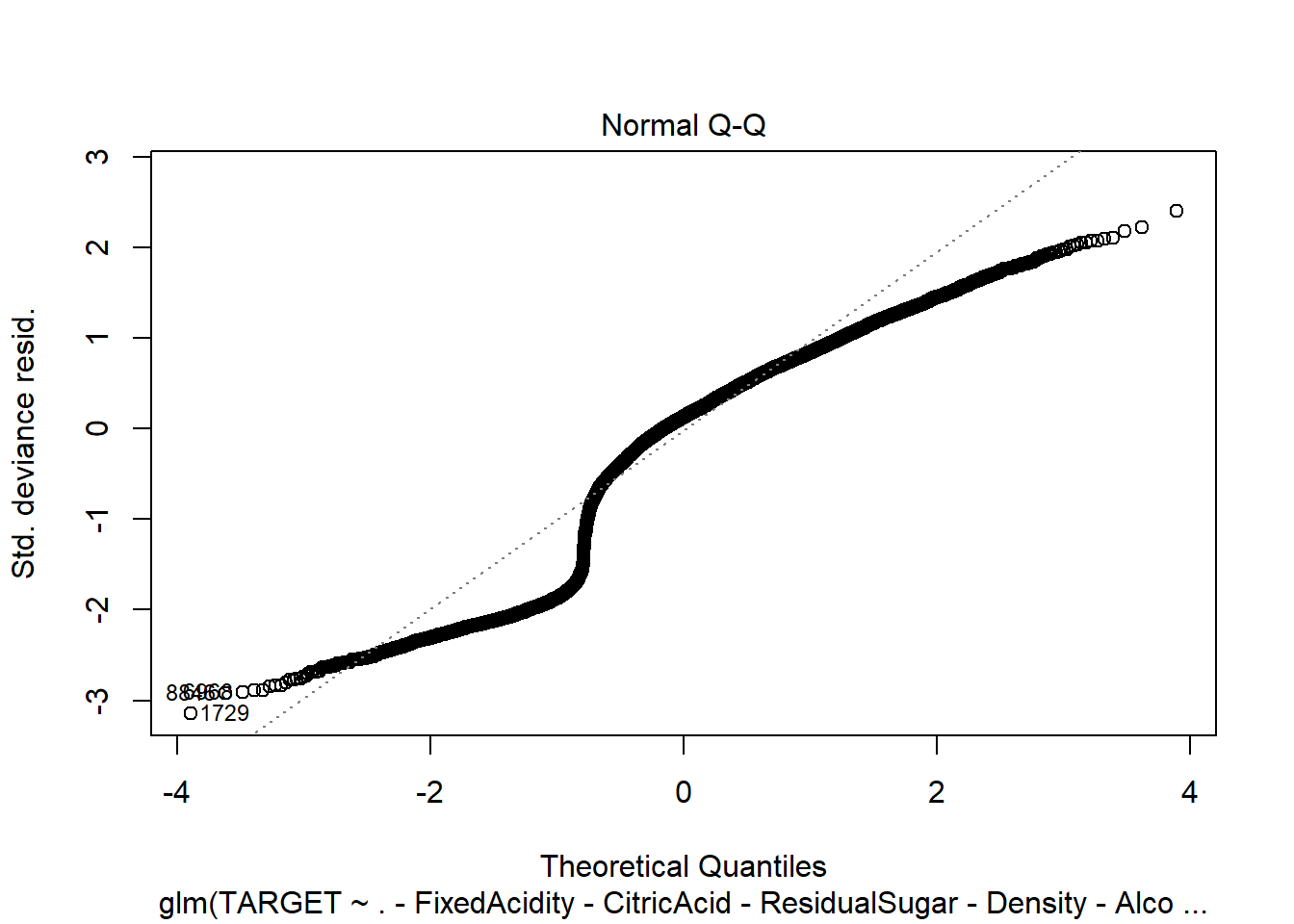
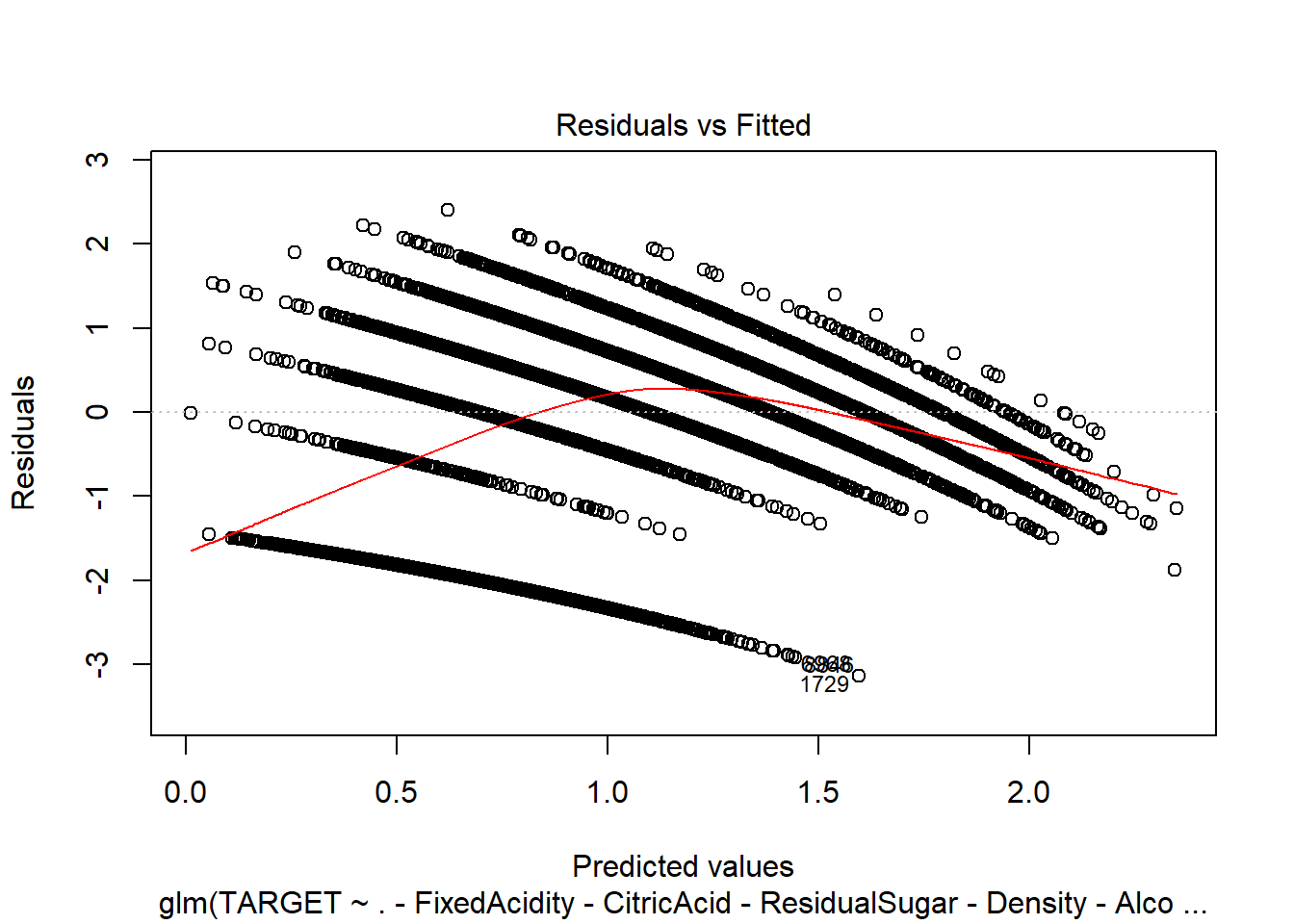
## Residual deviance: 12839 on 10227 degrees of freedom

## AIC: 38417

##

## Number of Fisher Scoring iterations: 5

plot(model4)

****

1. Negative Binomial without imputations:

model5 <- glm.nb(TARGET ~ ., data = wine\_train1)

## 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 = 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

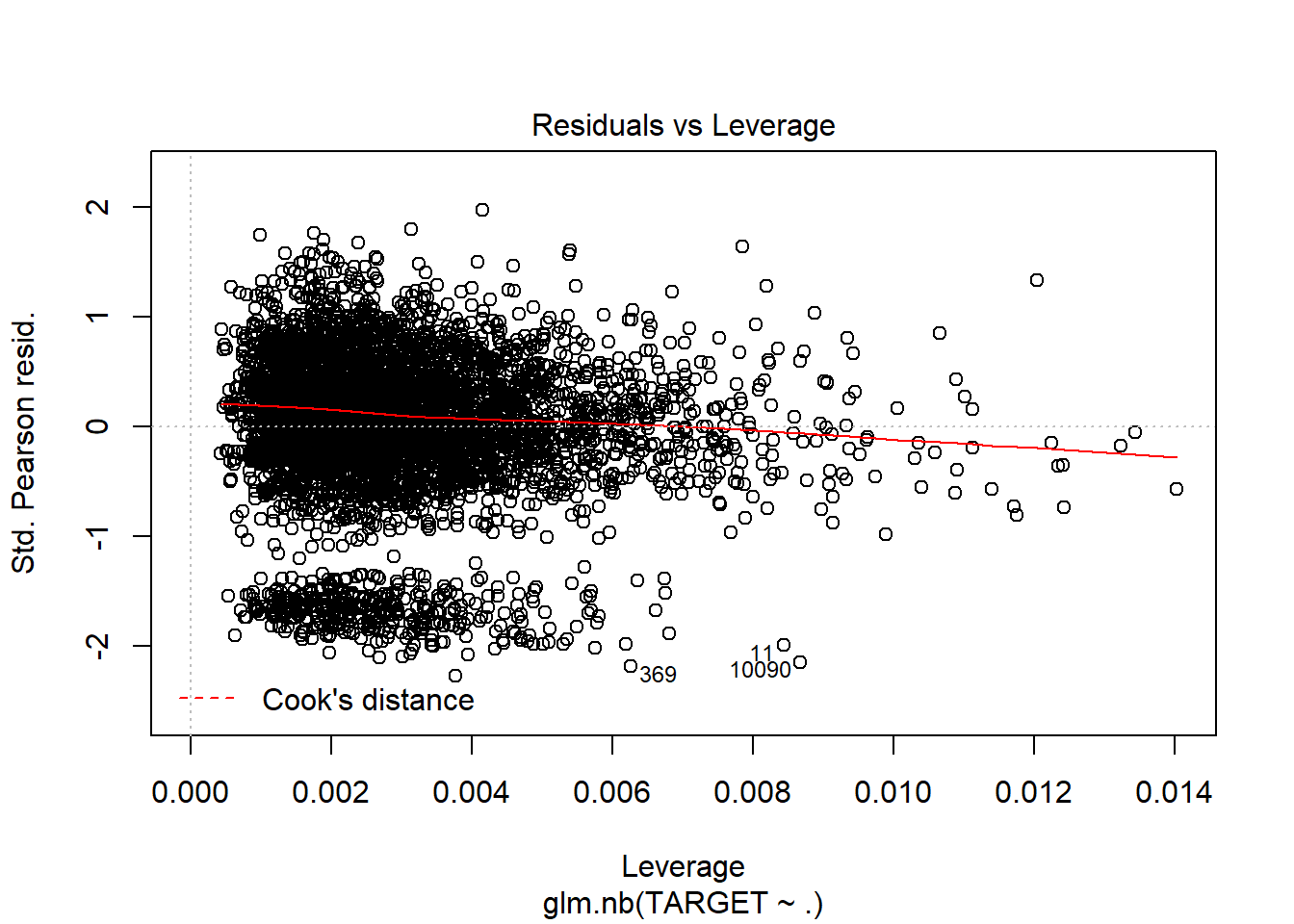
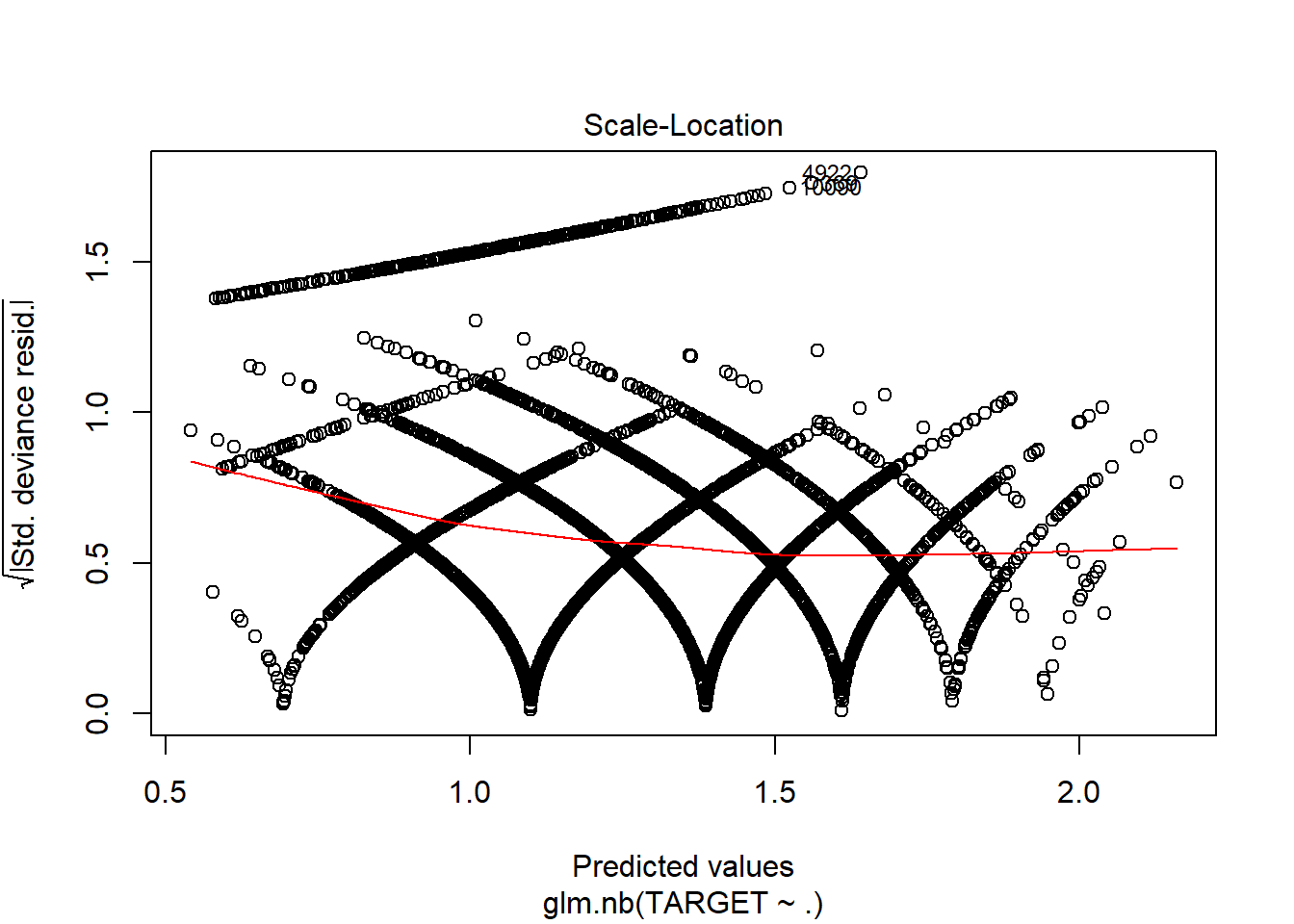
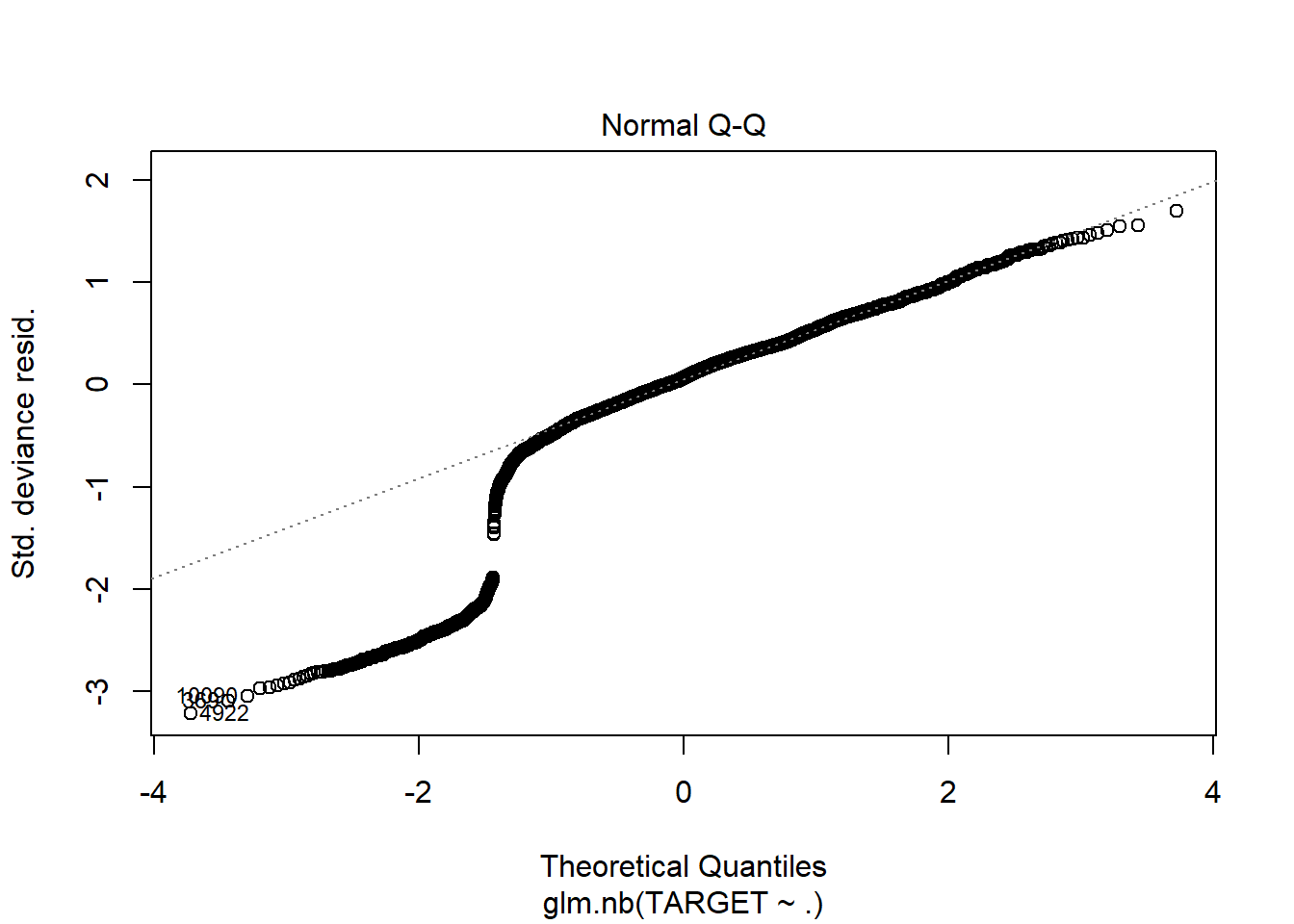
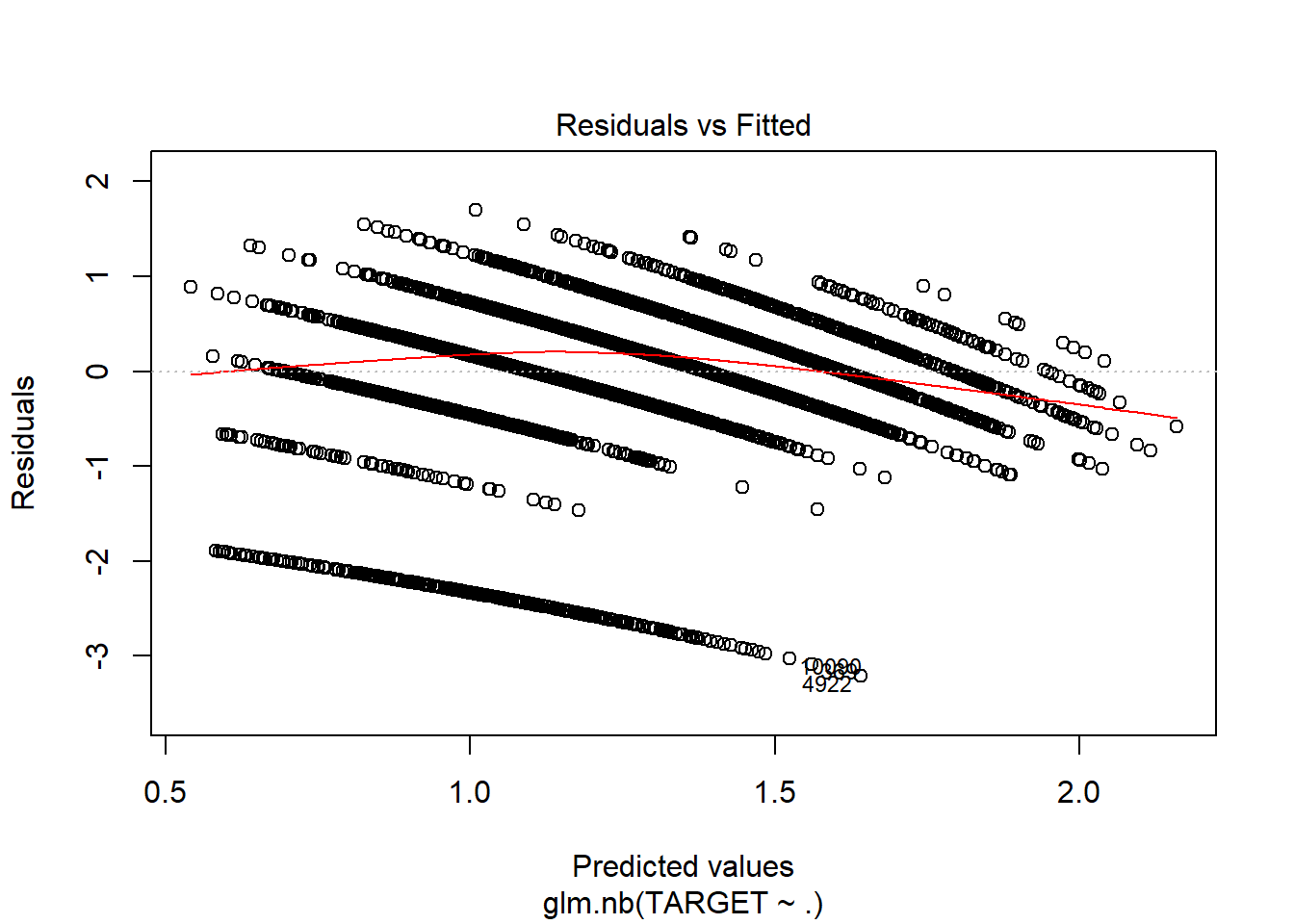
## Std. Err.: 259921

## Warning while fitting theta: iteration limit reached

##

## 2 x log-likelihood: -18515.07

plot(model5)

****

1. Negative Binomial without imputations and only significant variables:

model6 <- glm.nb(TARGET ~ .-FixedAcidity-CitricAcid-ResidualSugar-Chlorides-FreeSulfurDioxide-TotalSulfurDioxide-Density-pH-Sulphates-Alcohol, data = wine\_train1)

## 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 = 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

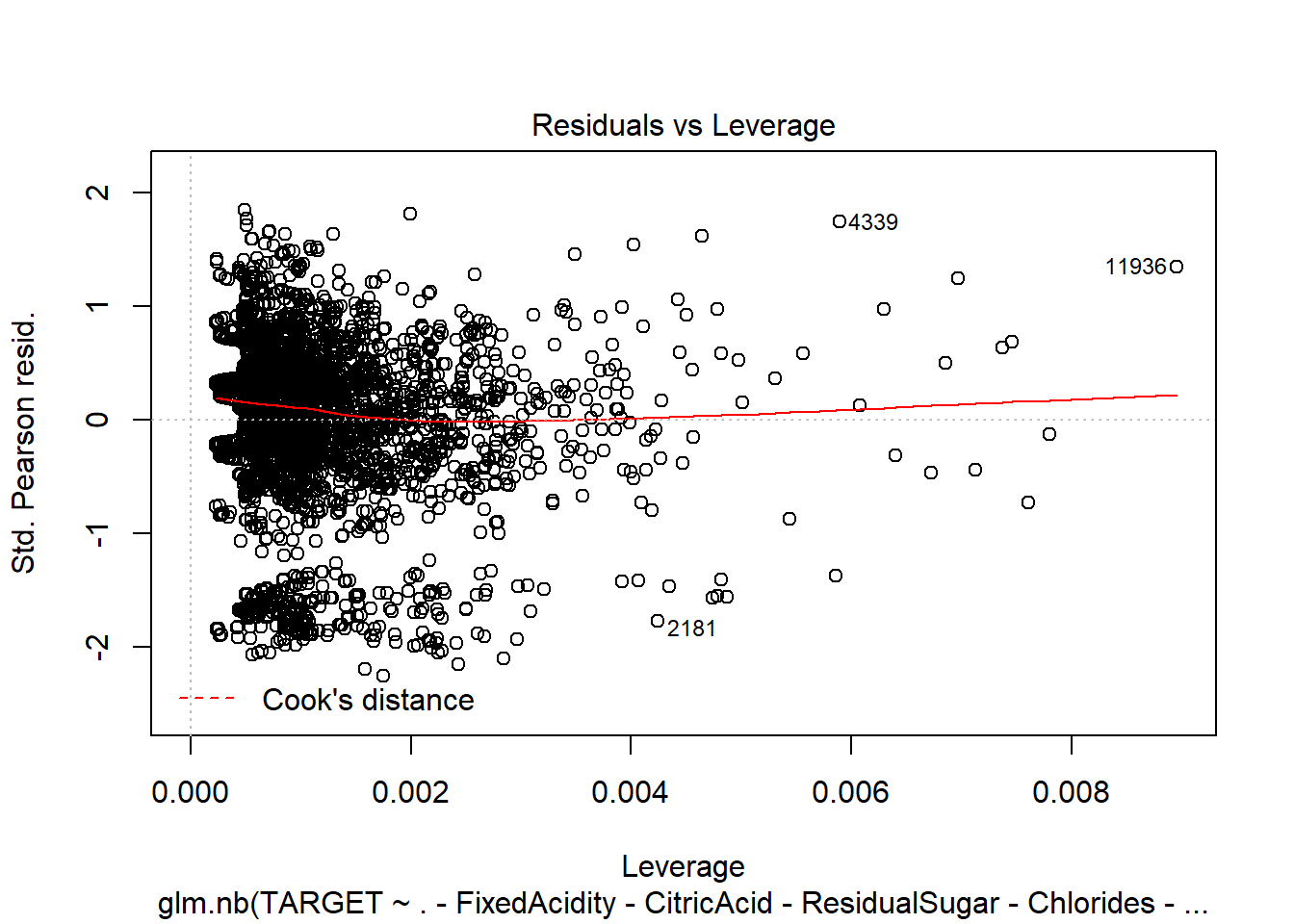
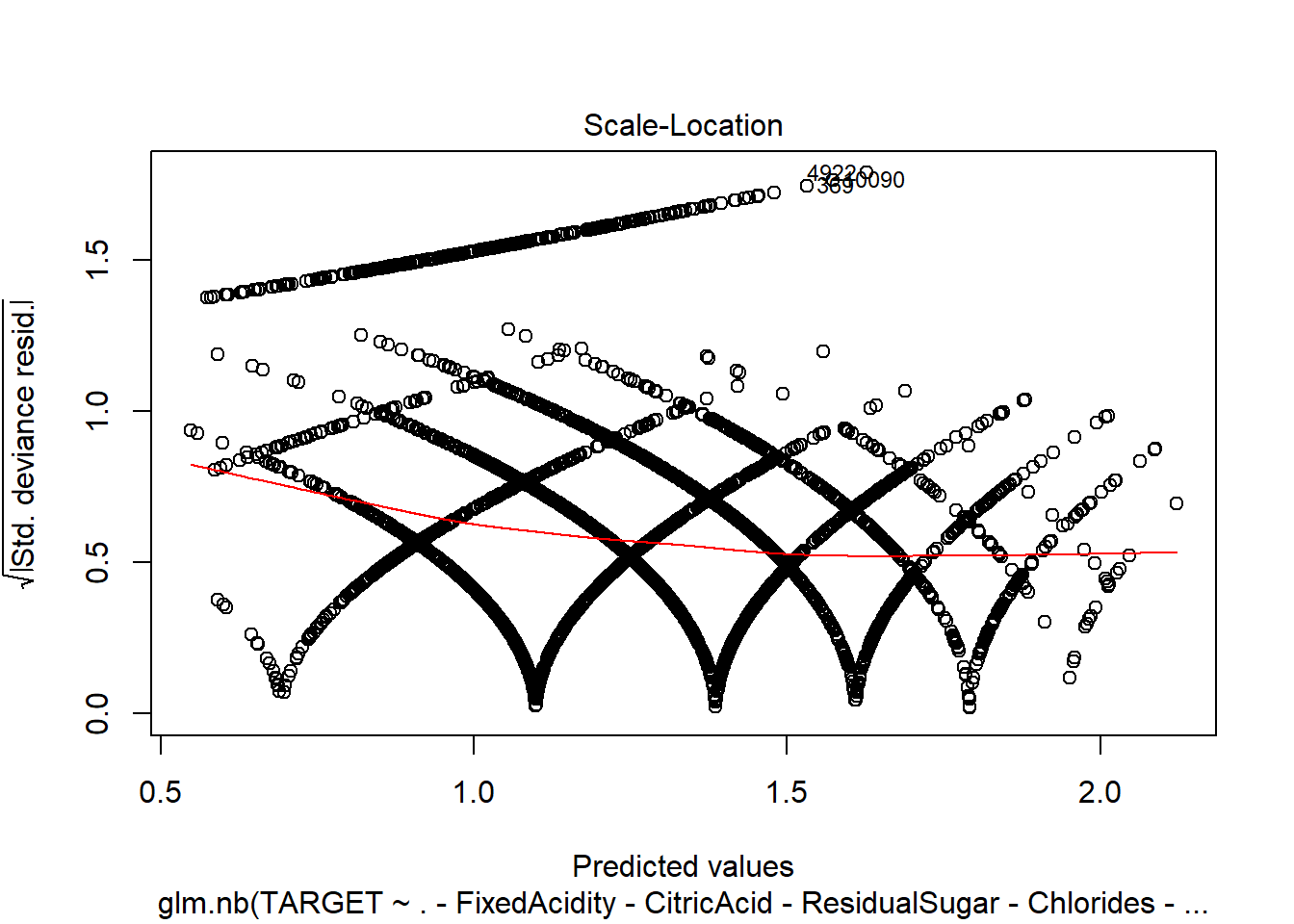
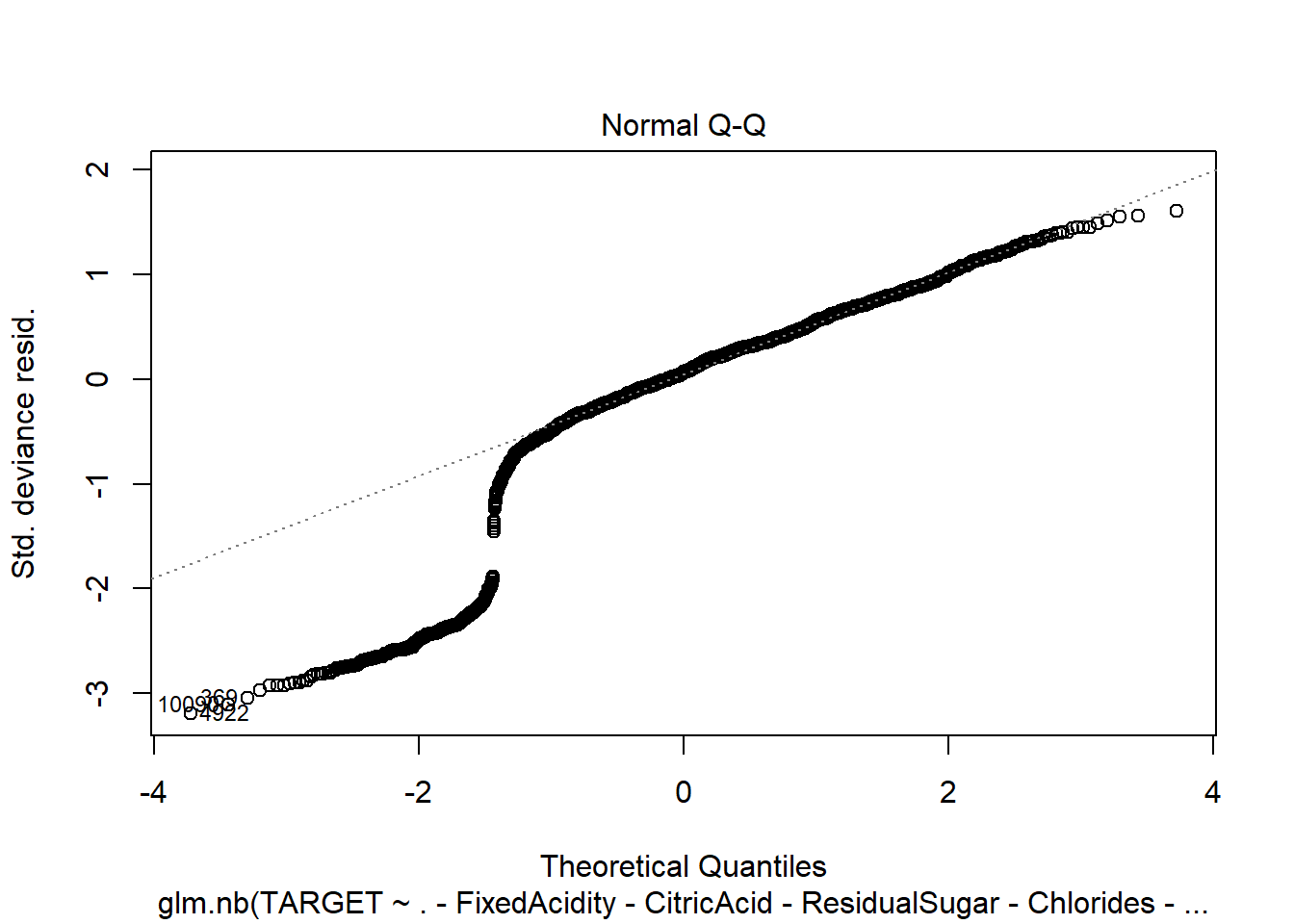
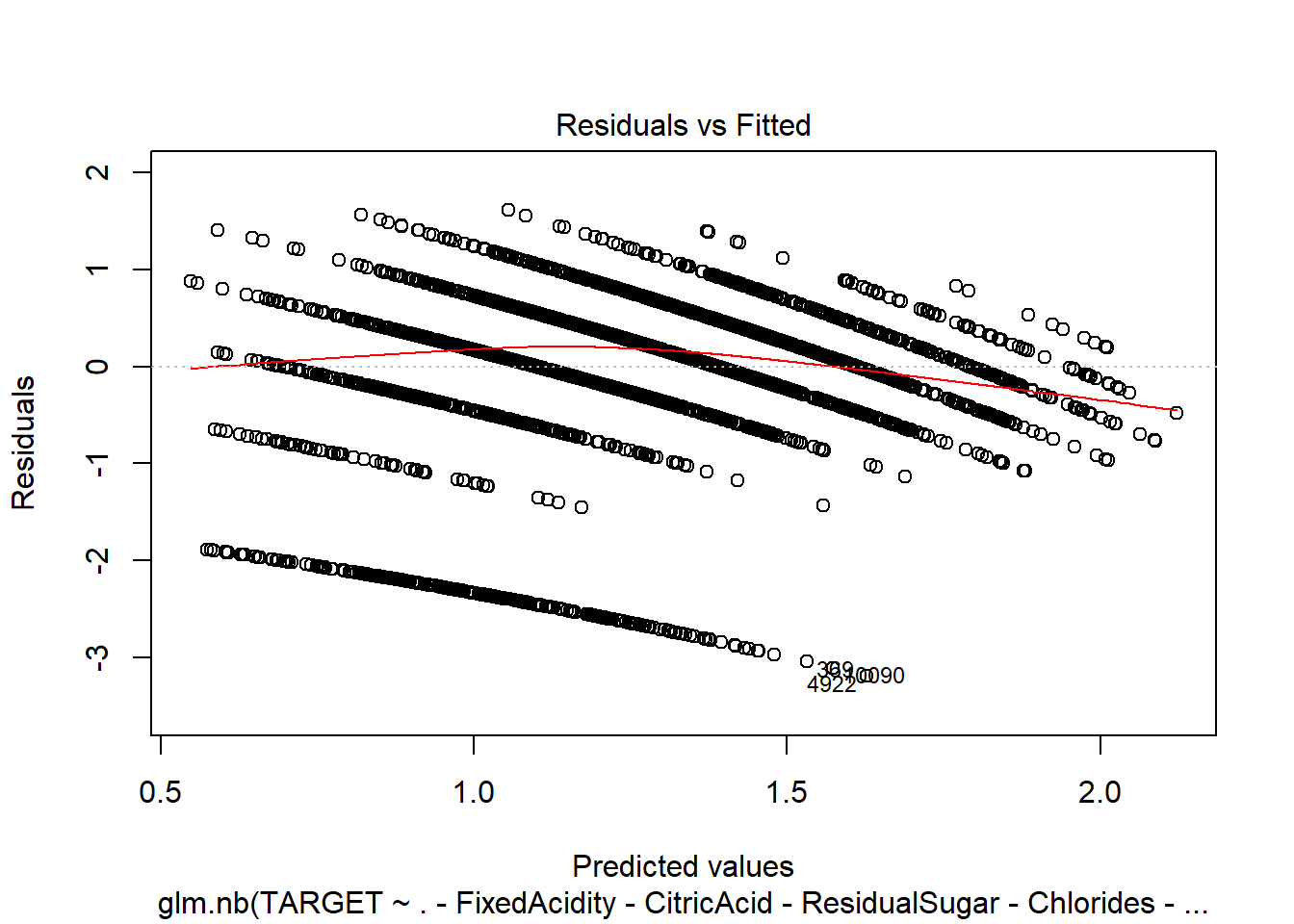
## Std. Err.: 258834

## Warning while fitting theta: iteration limit reached

##

## 2 x log-likelihood: -18525.37

plot(model6)

****

1. Negative Binomial with imputations:

model7 <- glm.nb(TARGET ~ ., data = wine\_train2)

## 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(model7)

##

## Call:

## glm.nb(formula = TARGET ~ ., data = wine\_train2, init.theta = 48897.24324,

## link = log)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.1515 -0.6808 0.1304 0.6390 2.4032

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 2.382e+00 2.277e-01 10.463 < 2e-16 \*\*\*

## FixedAcidity -1.332e-04 9.197e-04 -0.145 0.884879

## VolatileAcidity -4.351e-02 7.275e-03 -5.981 2.21e-09 \*\*\*

## CitricAcid 8.883e-03 6.577e-03 1.351 0.176804

## ResidualSugar 1.508e-04 1.675e-04 0.900 0.367960

## Chlorides -6.506e-02 1.791e-02 -3.633 0.000280 \*\*\*

## FreeSulfurDioxide 1.143e-04 3.804e-05 3.005 0.002657 \*\*

## TotalSulfurDioxide 8.709e-05 2.446e-05 3.560 0.000371 \*\*\*

## Density -4.047e-01 2.141e-01 -1.890 0.058762 .

## pH -1.788e-02 8.407e-03 -2.126 0.033466 \*

## Sulphates -1.327e-02 6.164e-03 -2.153 0.031286 \*

## Alcohol 2.690e-03 1.546e-03 1.740 0.081887 .

## LabelAppeal 1.432e-01 6.783e-03 21.106 < 2e-16 \*\*\*

## AcidIndex -7.622e-01 4.005e-02 -19.029 < 2e-16 \*\*\*

## STARS 3.401e-01 6.252e-03 54.393 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for Negative Binomial(48897.24) family taken to be 1)

##

## Null deviance: 18290 on 10236 degrees of freedom

## Residual deviance: 12830 on 10222 degrees of freedom

## AIC: 38420

##

## Number of Fisher Scoring iterations: 1

##

##

## Theta: 48897

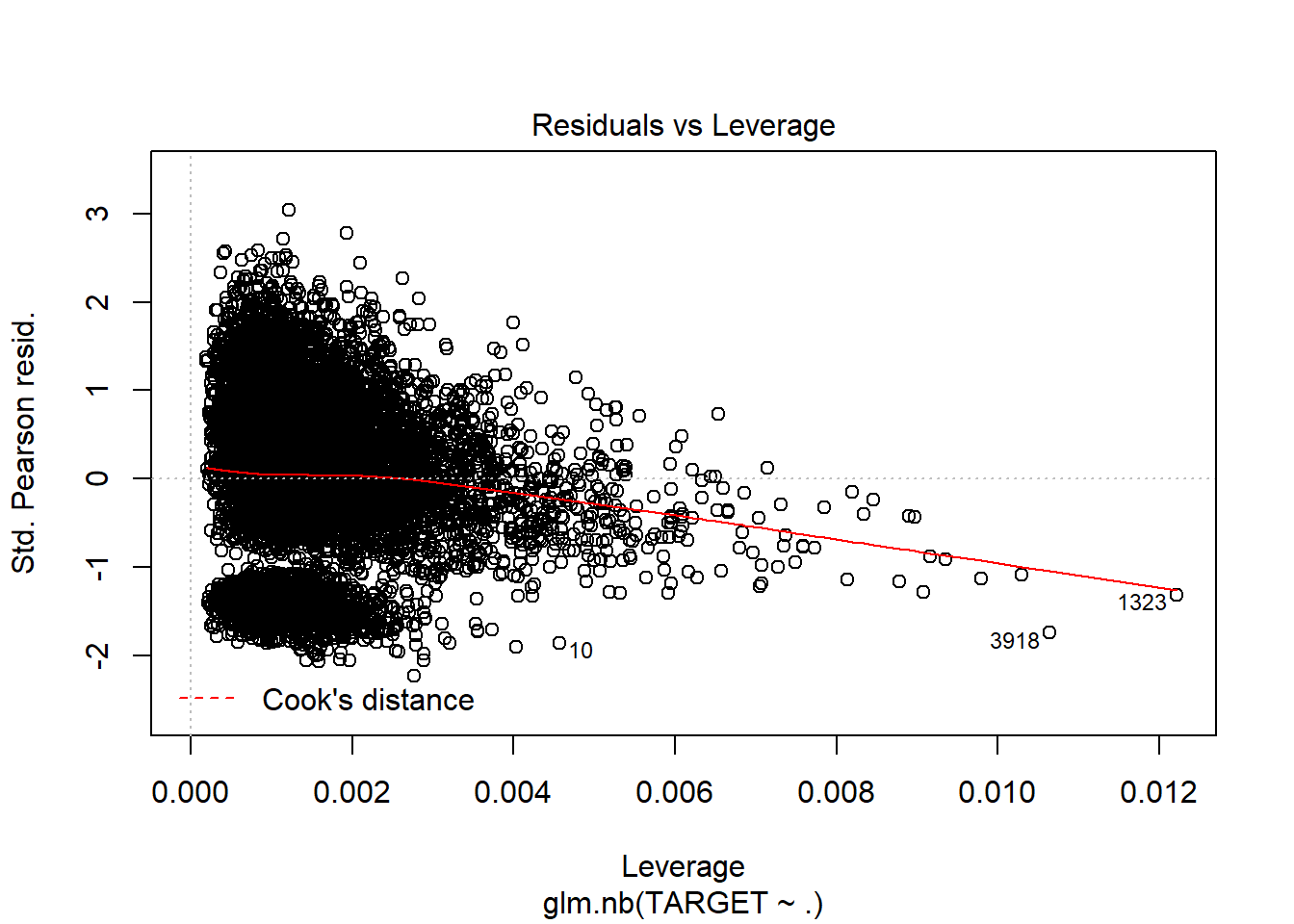
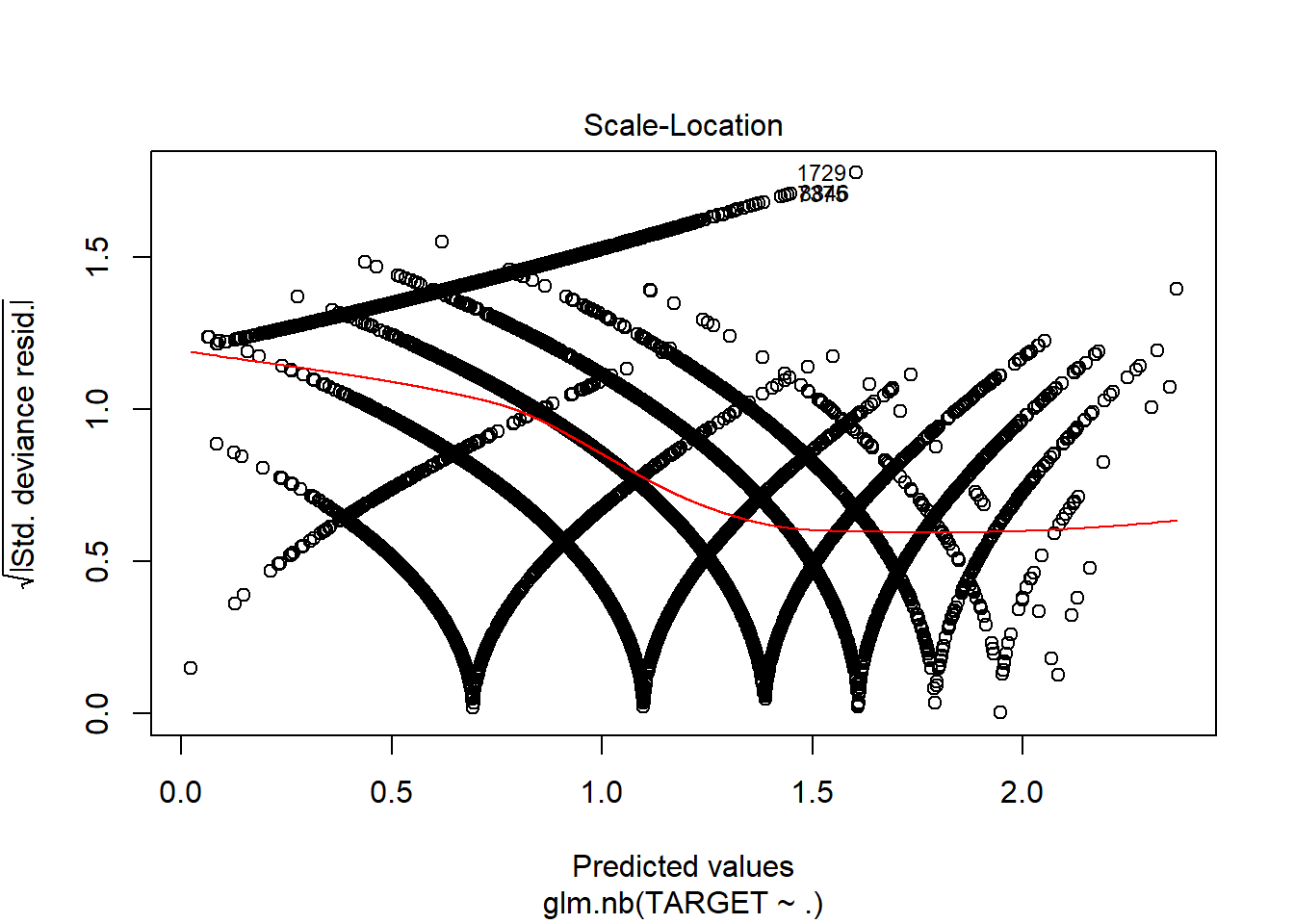
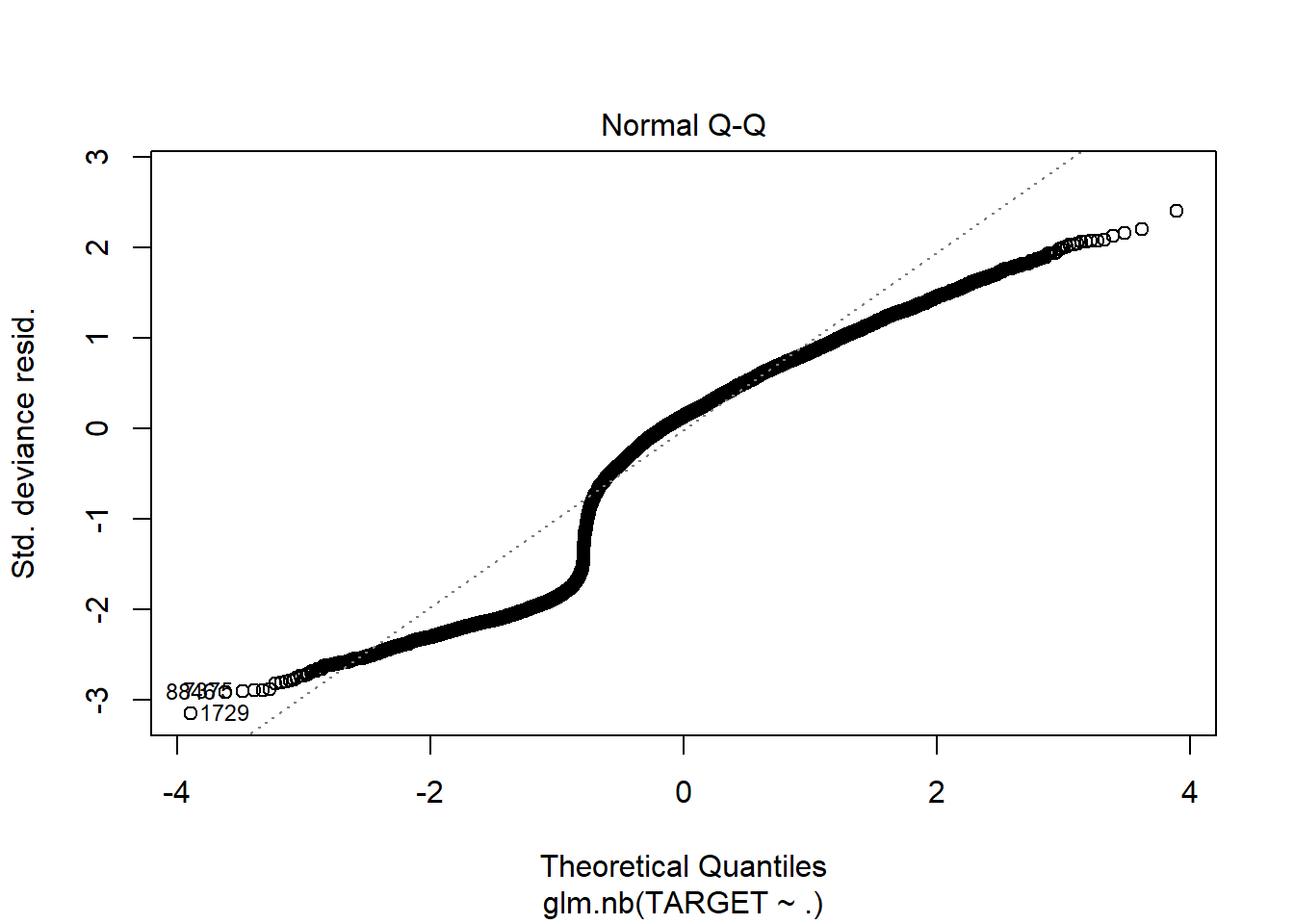
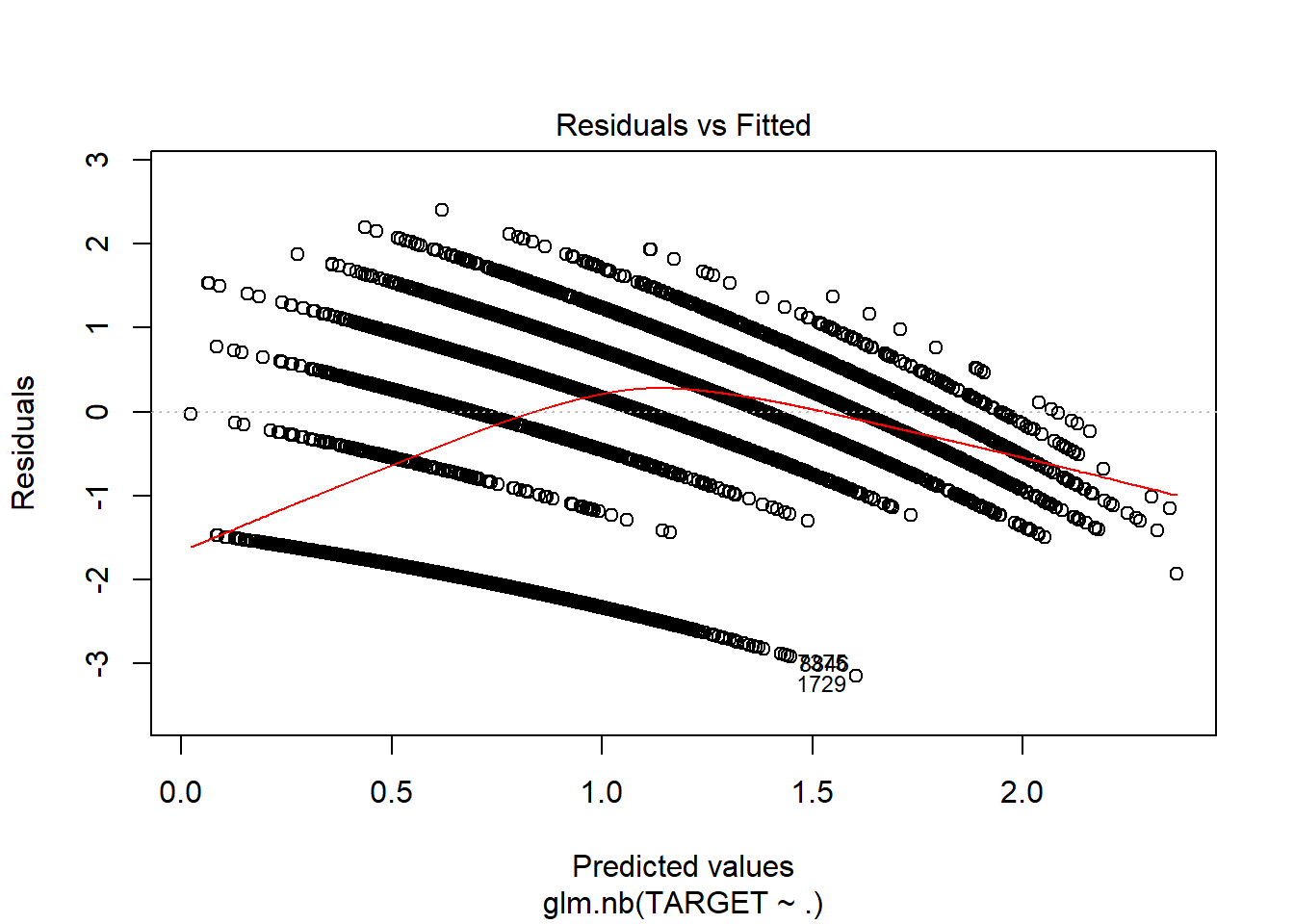
## Std. Err.: 63448

## Warning while fitting theta: iteration limit reached

##

## 2 x log-likelihood: -38388.3

plot(model7)

****

1. Negative Binomial with imputations and only significant variables:

model8 <- glm.nb(TARGET ~ .-FixedAcidity-CitricAcid-ResidualSugar-Density-Alcohol, data = wine\_train2)

## 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(model8)

##

## Call:

## glm.nb(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -

## Density - Alcohol, data = wine\_train2, init.theta = 48805.90033,

## link = log)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.1405 -0.6852 0.1288 0.6412 2.4038

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 2.019e+00 8.849e-02 22.820 < 2e-16 \*\*\*

## VolatileAcidity -4.388e-02 7.273e-03 -6.033 1.61e-09 \*\*\*

## Chlorides -6.711e-02 1.790e-02 -3.750 0.000177 \*\*\*

## FreeSulfurDioxide 1.119e-04 3.802e-05 2.942 0.003257 \*\*

## TotalSulfurDioxide 8.561e-05 2.443e-05 3.505 0.000457 \*\*\*

## pH -1.818e-02 8.404e-03 -2.164 0.030489 \*

## Sulphates -1.327e-02 6.157e-03 -2.155 0.031144 \*

## LabelAppeal 1.433e-01 6.783e-03 21.119 < 2e-16 \*\*\*

## AcidIndex -7.665e-01 3.941e-02 -19.447 < 2e-16 \*\*\*

## STARS 3.410e-01 6.237e-03 54.671 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for Negative Binomial(48805.9) family taken to be 1)

##

## Null deviance: 18290 on 10236 degrees of freedom

## Residual deviance: 12839 on 10227 degrees of freedom

## AIC: 38420

##

## Number of Fisher Scoring iterations: 1

##

##

## Theta: 48806

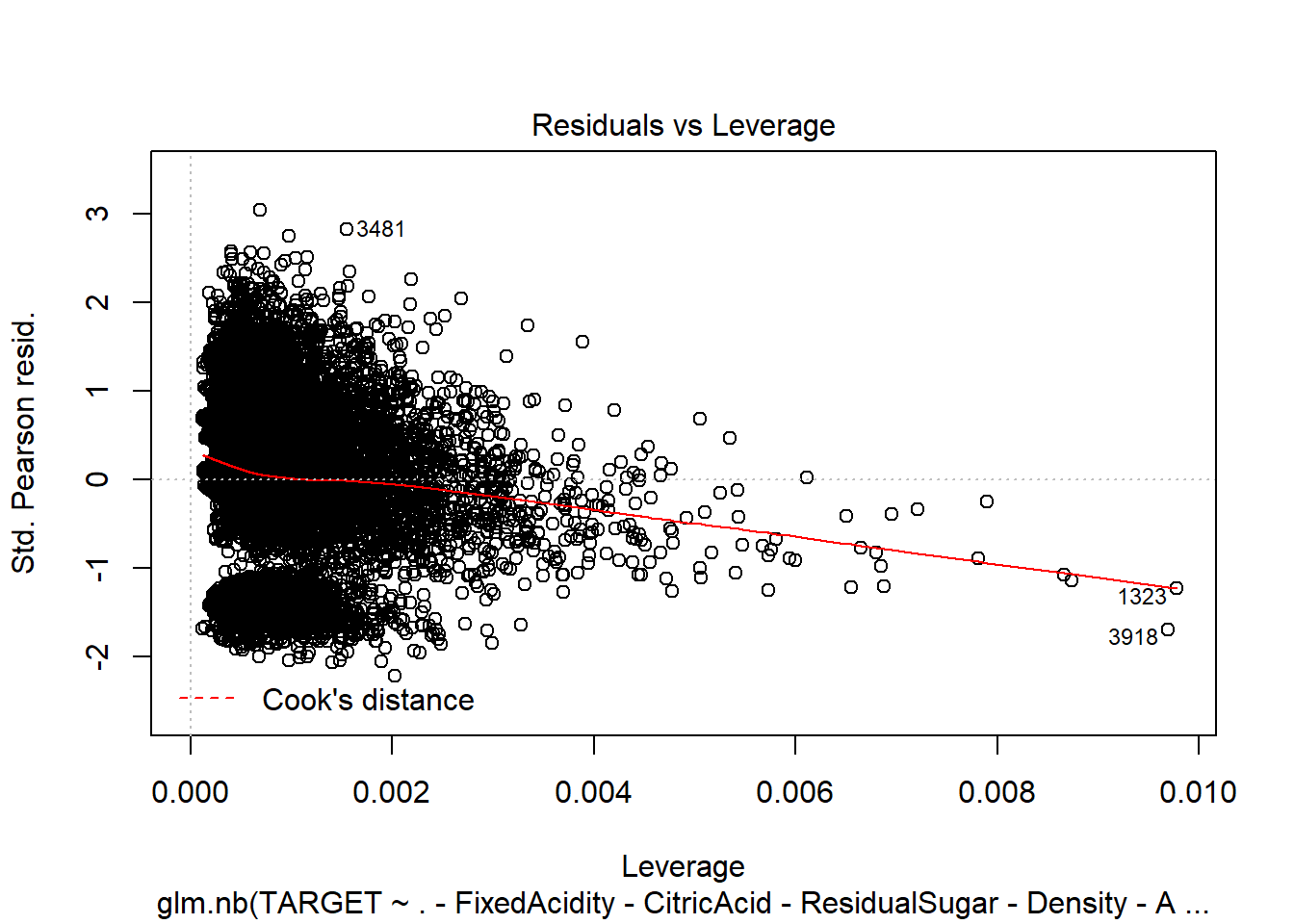
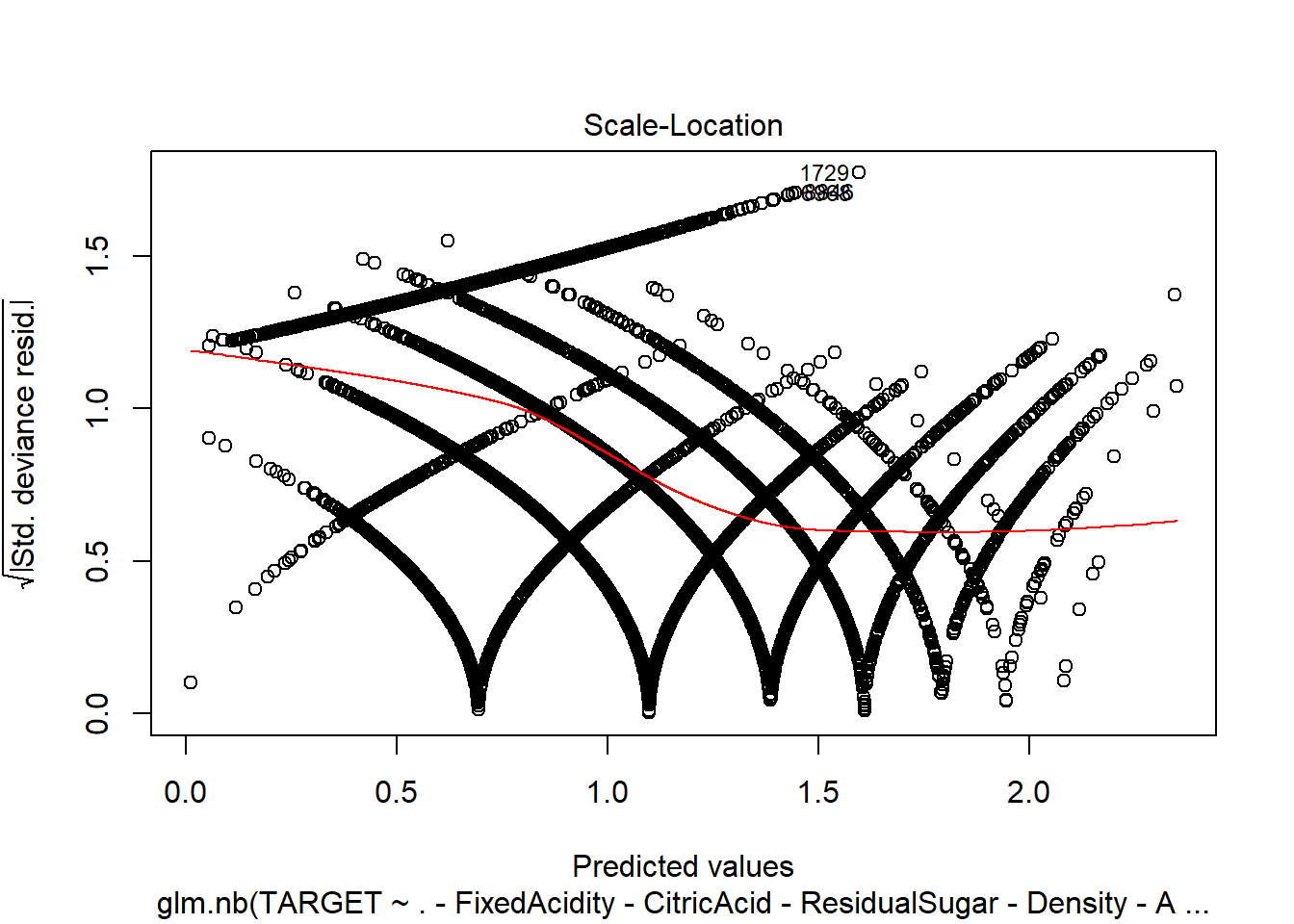
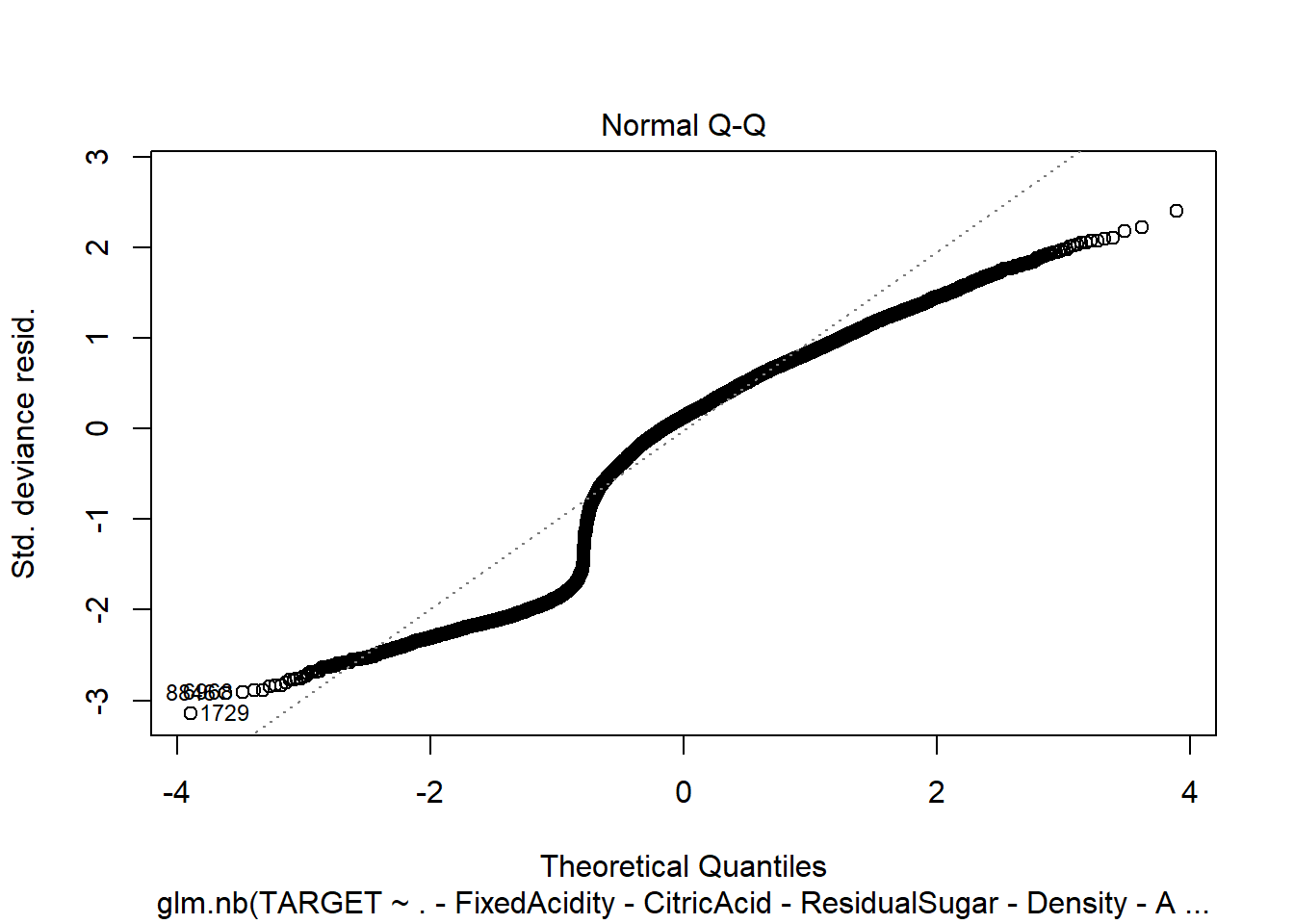
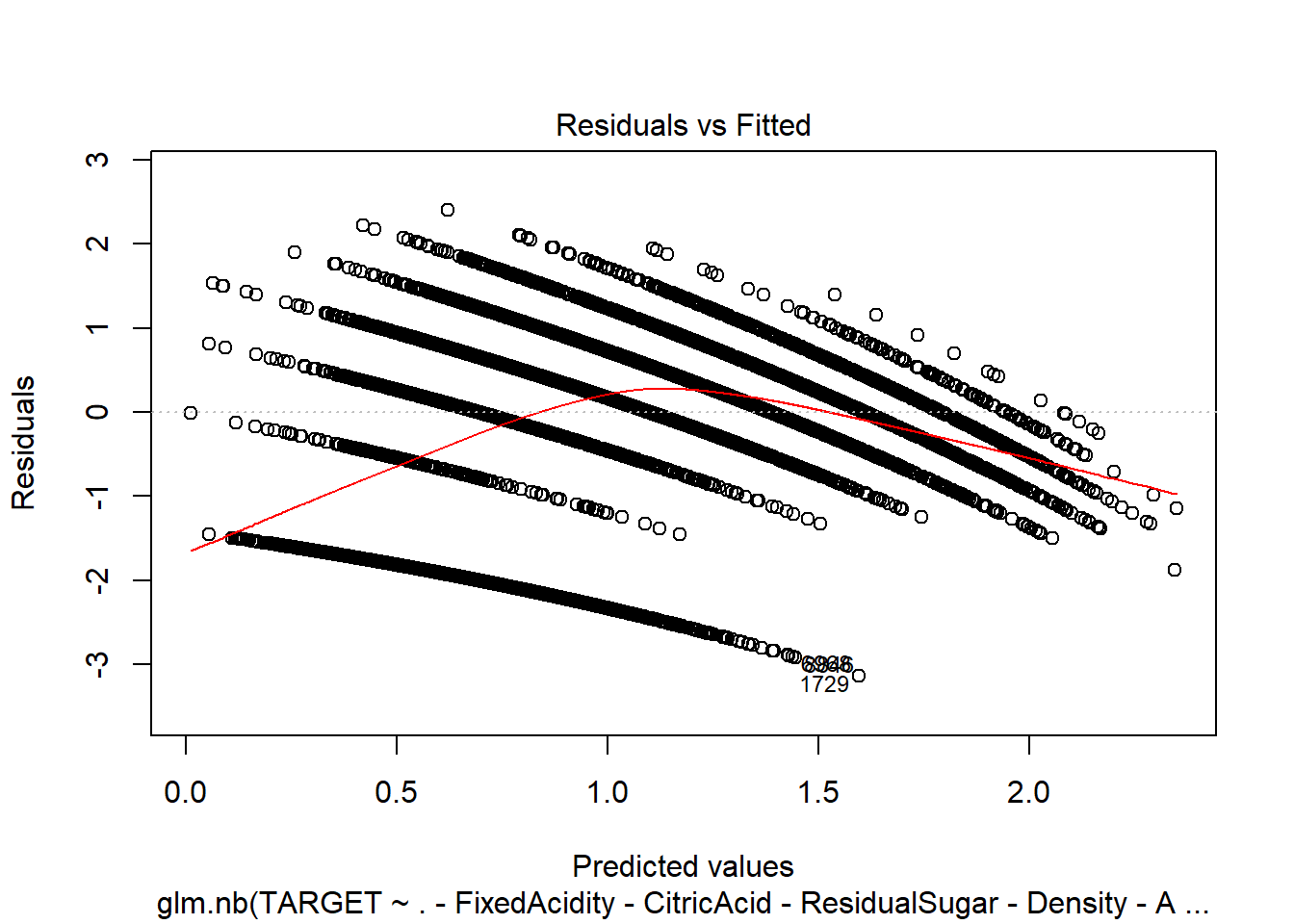
## Std. Err.: 63368

## Warning while fitting theta: iteration limit reached

##

## 2 x log-likelihood: -38397.65

plot(model8)

****

1. Linear Model with imputations.

model9 <- lm(TARGET ~ ., data = wine\_train2)

summary(model9)

##

## Call:

## lm(formula = TARGET ~ ., data = wine\_train2)

##

## Residuals:

## Min 1Q Median 3Q Max

## -4.6944 -1.0191 0.1692 1.0335 4.2502

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 6.149e+00 5.564e-01 11.052 < 2e-16 \*\*\*

## FixedAcidity -1.428e-04 2.255e-03 -0.063 0.94952

## VolatileAcidity -1.265e-01 1.792e-02 -7.056 1.82e-12 \*\*\*

## CitricAcid 2.771e-02 1.630e-02 1.699 0.08927 .

## ResidualSugar 4.479e-04 4.138e-04 1.083 0.27904

## Chlorides -1.956e-01 4.398e-02 -4.448 8.77e-06 \*\*\*

## FreeSulfurDioxide 2.930e-04 9.398e-05 3.117 0.00183 \*\*

## TotalSulfurDioxide 2.365e-04 6.006e-05 3.938 8.28e-05 \*\*\*

## Density -1.099e+00 5.263e-01 -2.088 0.03678 \*

## pH -4.064e-02 2.071e-02 -1.962 0.04978 \*

## Sulphates -3.621e-02 1.519e-02 -2.384 0.01713 \*

## Alcohol 1.131e-02 3.782e-03 2.991 0.00279 \*\*

## LabelAppeal 4.379e-01 1.644e-02 26.633 < 2e-16 \*\*\*

## AcidIndex -2.041e+00 9.250e-02 -22.067 < 2e-16 \*\*\*

## STARS 1.162e+00 1.665e-02 69.754 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

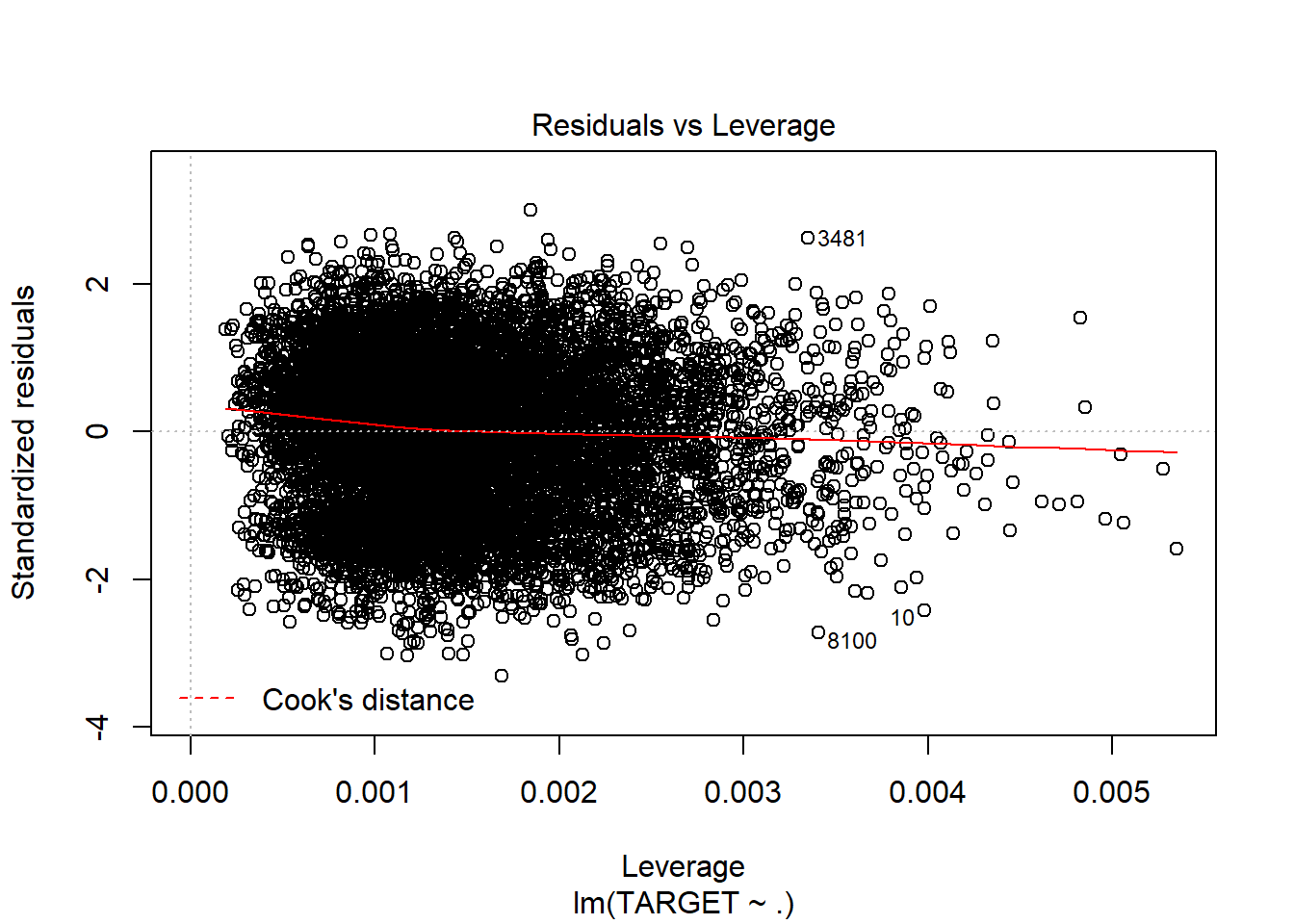
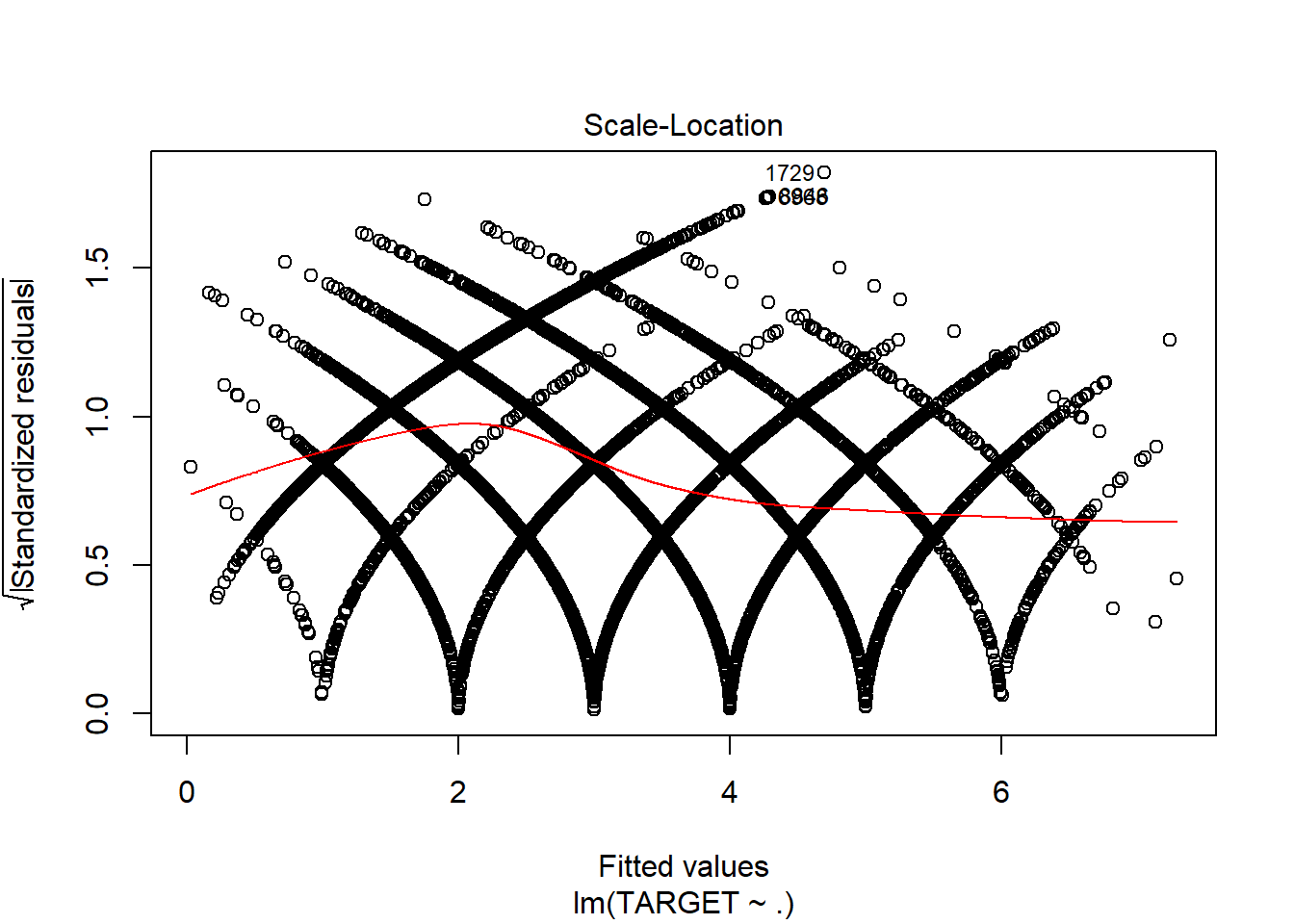
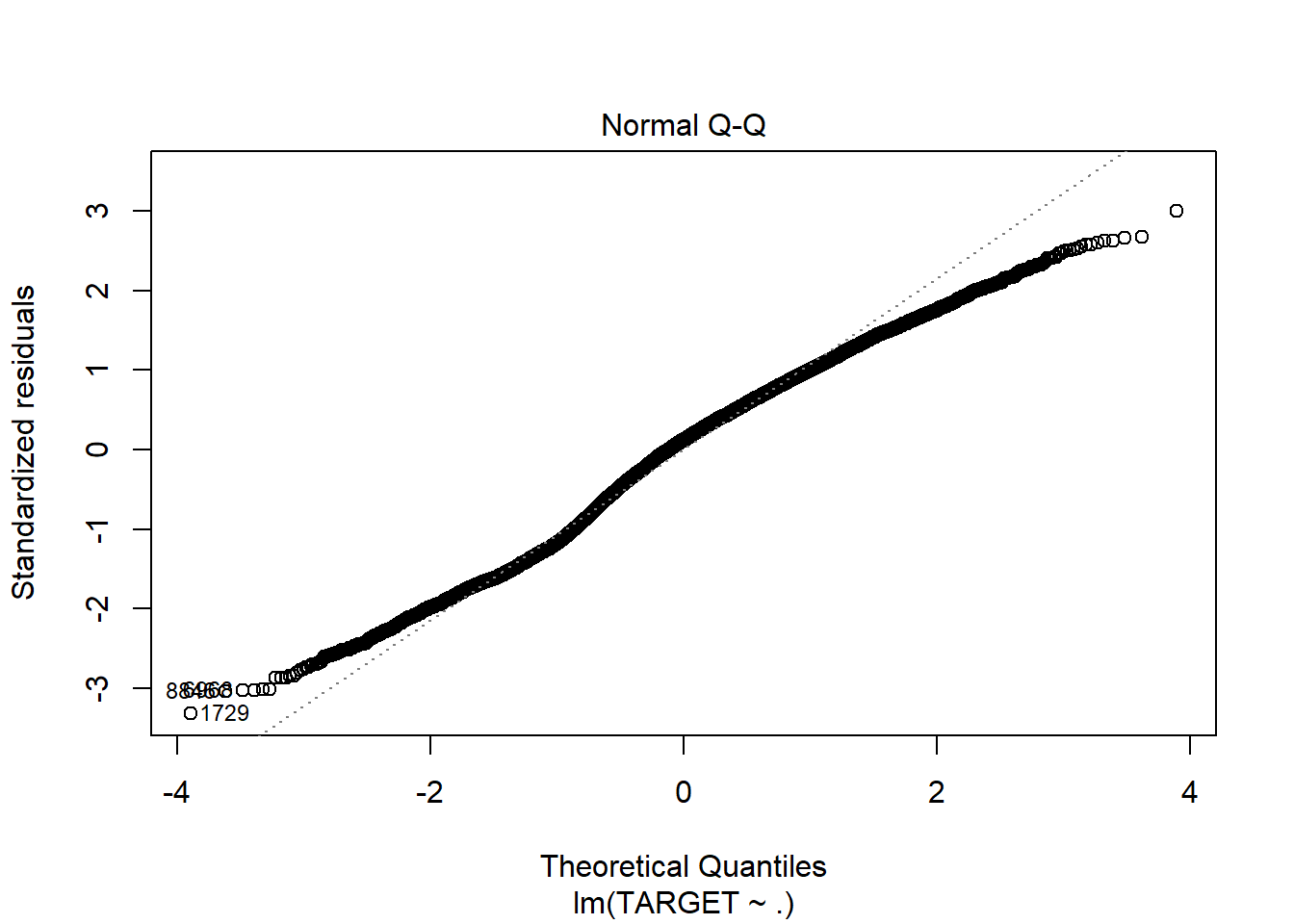
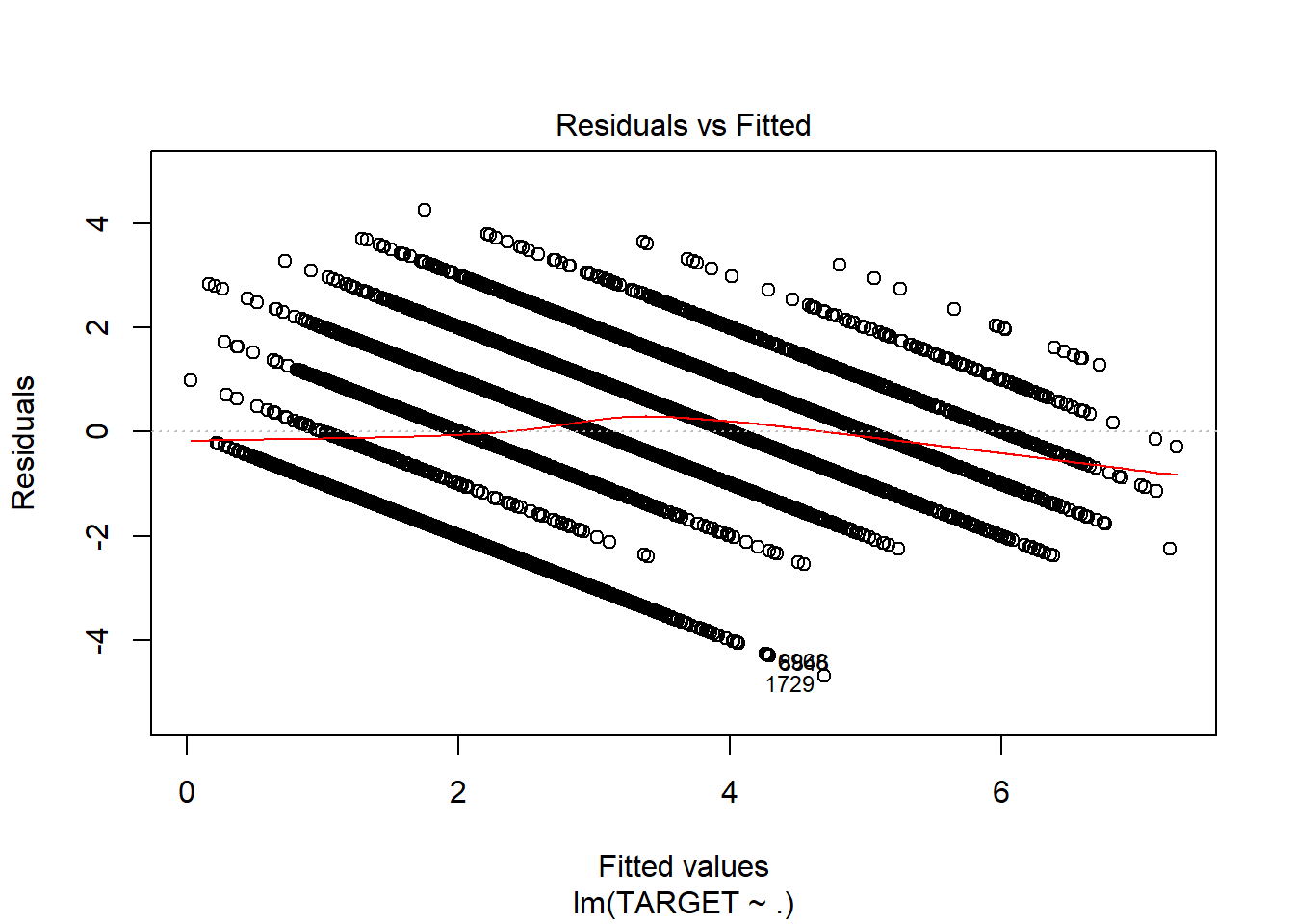
##

## Residual standard error: 1.417 on 10222 degrees of freedom

## Multiple R-squared: 0.4598, Adjusted R-squared: 0.4591

## F-statistic: 621.5 on 14 and 10222 DF, p-value: < 2.2e-16

plot(model9)

****

1. Linear Model with imputations and only significant variables.

model10 <- lm(TARGET ~ .-FixedAcidity-CitricAcid-ResidualSugar, data = wine\_train2)

summary(model10)

##

## Call:

## lm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar,

## data = wine\_train2)

##

## Residuals:

## Min 1Q Median 3Q Max

## -4.7075 -1.0195 0.1718 1.0343 4.2907

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 6.139e+00 5.563e-01 11.036 < 2e-16 \*\*\*

## VolatileAcidity -1.273e-01 1.792e-02 -7.104 1.30e-12 \*\*\*

## Chlorides -1.970e-01 4.397e-02 -4.479 7.57e-06 \*\*\*

## FreeSulfurDioxide 2.939e-04 9.397e-05 3.128 0.00177 \*\*

## TotalSulfurDioxide 2.389e-04 6.003e-05 3.980 6.94e-05 \*\*\*

## Density -1.101e+00 5.263e-01 -2.093 0.03638 \*

## pH -4.059e-02 2.071e-02 -1.960 0.05008 .

## Sulphates -3.699e-02 1.517e-02 -2.437 0.01481 \*

## Alcohol 1.136e-02 3.781e-03 3.005 0.00266 \*\*

## LabelAppeal 4.379e-01 1.644e-02 26.632 < 2e-16 \*\*\*

## AcidIndex -2.031e+00 9.085e-02 -22.351 < 2e-16 \*\*\*

## STARS 1.162e+00 1.665e-02 69.794 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

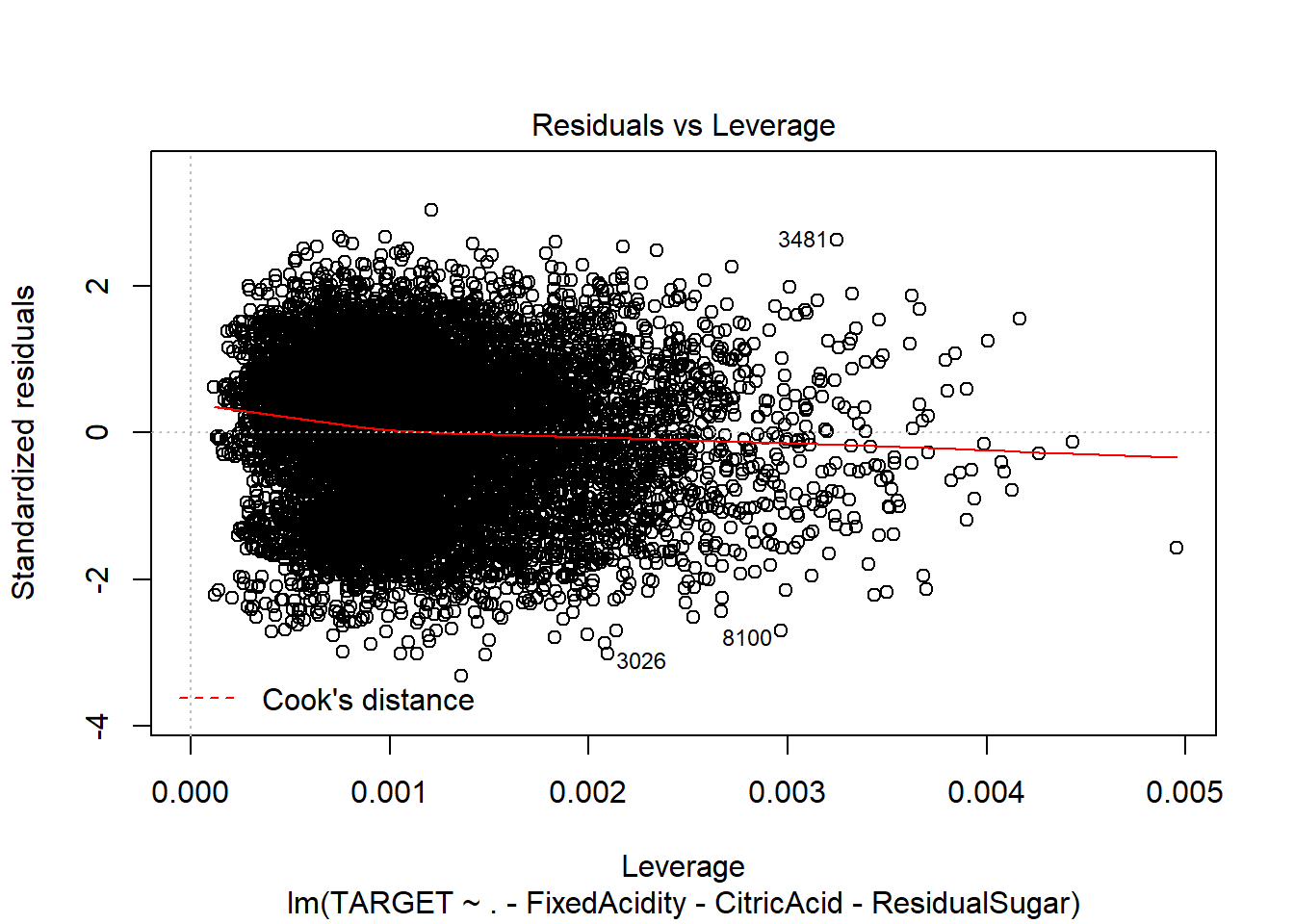
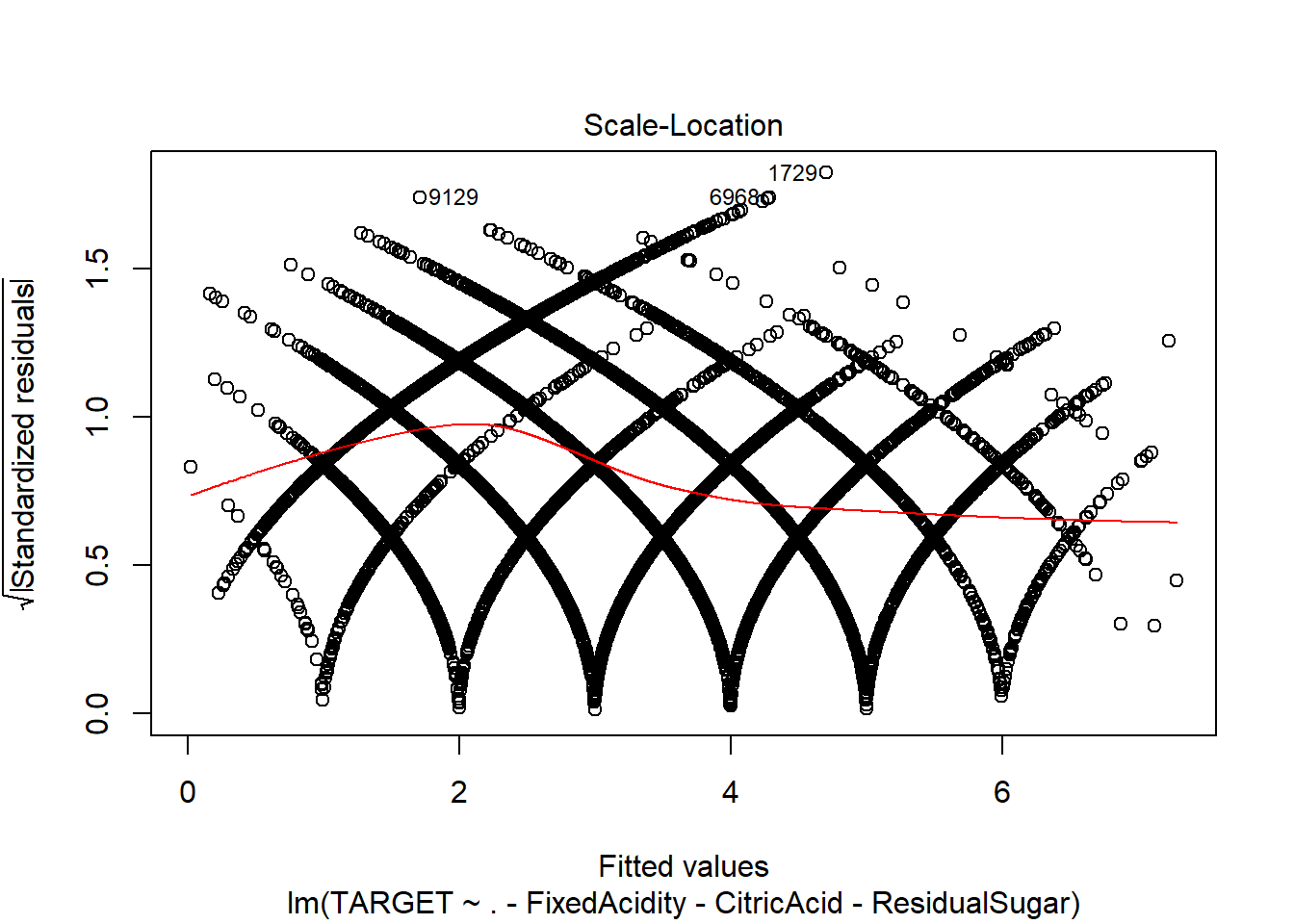
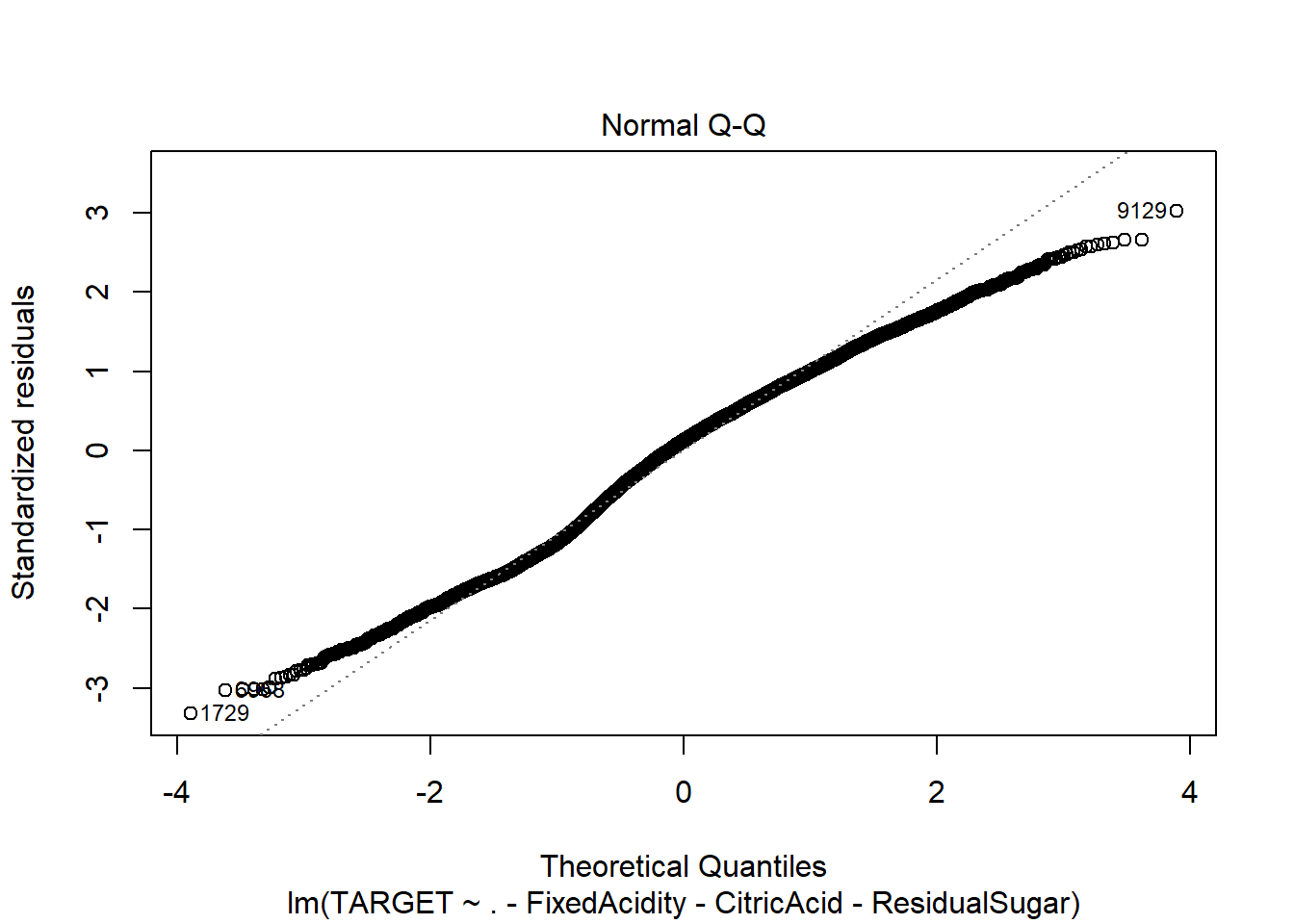
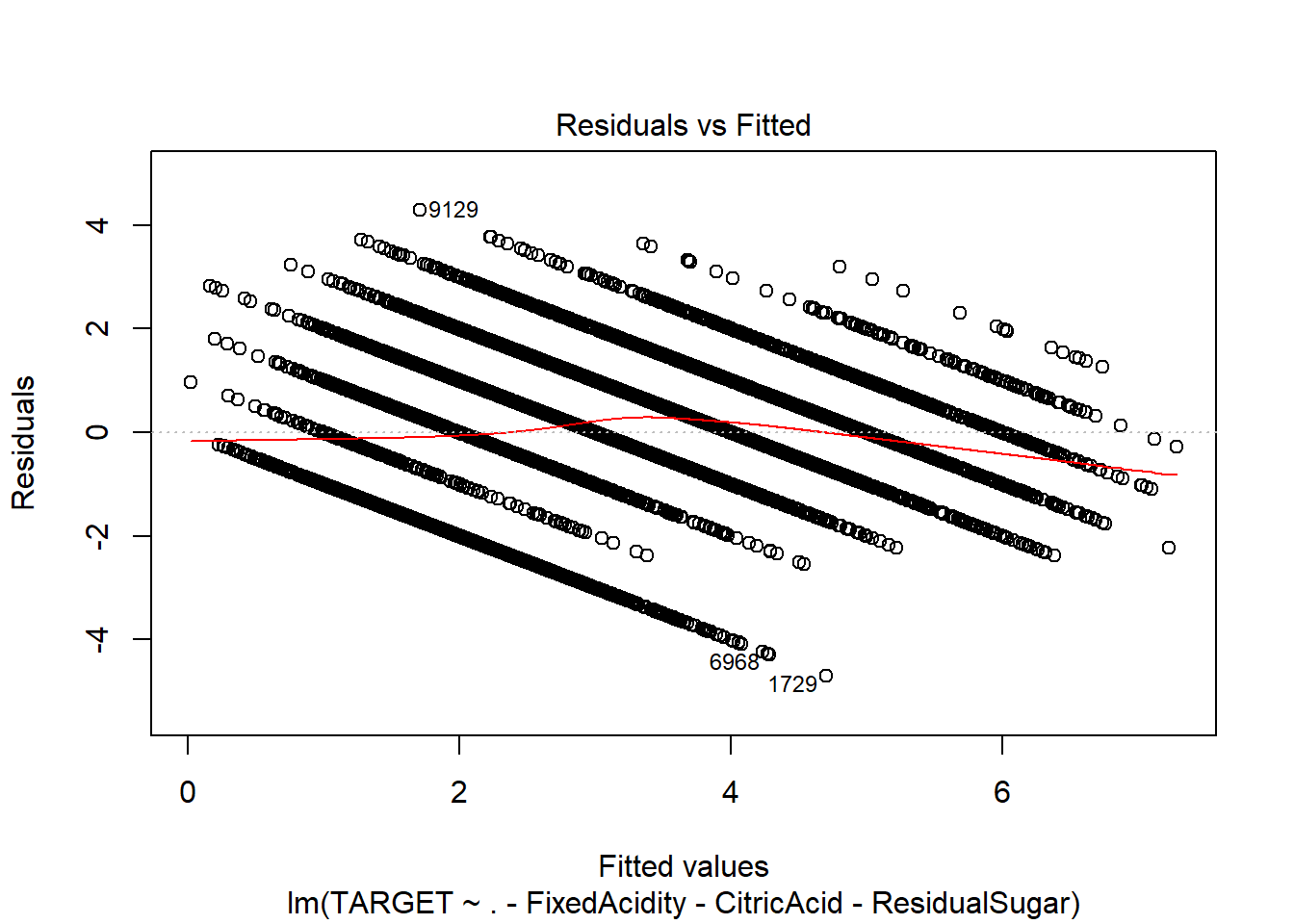
##

## Residual standard error: 1.417 on 10225 degrees of freedom

## Multiple R-squared: 0.4596, Adjusted R-squared: 0.459

## F-statistic: 790.6 on 11 and 10225 DF, p-value: < 2.2e-16

plot(model10)

****

Now let’s see the output of the Models using test data:

We will use the squared loss to validate the model.

modelValidation <- **function**(mod, test){

preds = predict(mod, test)

diffMat = as.numeric(preds) - as.numeric(test$TARGET)

diffMat = diffMat^2

loss <- mean(diffMat)

**return**(loss)

}

Poisson model with imputations.

modelValidation(model3, wine\_test2)

## [1] 6.852209

Poisson model with imputations and only significant variables.

modelValidation(model4, wine\_test2)

## [1] 6.854547

Negative Binomial with imputations:.

modelValidation(model7, wine\_test2)

## [1] 6.852205

Negative Binomial with imputations and only significant variables.

modelValidation(model8, wine\_test2)

## [1] 6.854543

Linear Model with imputations.

modelValidation(model9, wine\_test2)

## [1] 2.029061

Linear Model with imputations and only significant variables.

modelValidation(model10, wine\_test2)

## [1] 2.030002

**Model Selection:**

From the above models, i would like to go with Model10 - Linear Model with imputations and only significant variables as it uses less variables and is parsimonious. Also the R2 looks fine. The squared loss is also fine.

# Prediction:

We will use the same method to impute and use log transformation for AcidIndex.

wine\_eval <- read.csv("https://raw.githubusercontent.com/Riteshlohiya/Data621-Assignment-5/master/wine\_evaluation\_data.csv", stringsAsFactors = FALSE)

do\_factors <- **function**(wine\_instance){

wine\_instance <- within(wine\_instance, {

LabelAppeal <- factor(LabelAppeal)

AcidIndex <- factor(AcidIndex)

STARS <- factor(STARS)

})

**return** (wine\_instance)

}

summary(wine\_eval)

wine\_eval <- mice(wine\_eval, m=1, maxit = 5, seed = 42)

wine\_eval <- complete(wine\_eval)

wine\_eval <- as.data.frame(wine\_eval)

wine\_eval$AcidIndex <- log(wine\_eval$AcidIndex)

wine\_eval$TARGET1 <- predict(model10, newdata=wine\_eval)

write.csv(wine\_eval,"Evaluation\_Full\_Data.csv", row.names=FALSE)

The output is present at the below given location.

Link for output: <https://github.com/Riteshlohiya/Data621-Assignment-5/blob/master/Evaluation_Full_Data.csv>