DATA621-HW5-SmoothOperators

Rob Hodde, Matt Farris, Jeffrey Burmood, Bin Lin

5/11/2017

Problem Description

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.

The 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. This variable is called TARGET.

Data Exploration

Data Exploration

There are numerous NAs in certain variables, and variables with negative values. Variables with negative values have nearly normal distributions so it is possible some previous data adjustments have been made. The variable data with negative values in stable, normal distributions will be used as-is. Below is a summary of variables by type, followed by their basic statistical summaries:

VAR	TYPE
TARGET	integer
FixedAcidity	double
VolatileAcidity	double
CitricAcid	double
ResidualSugar	double
Chlorides	double
FreeSulfurDioxide	double
TotalSulfurDioxide	double
Density	double
pH	double
Sulphates	double
Alcohol	double
LabelAppeal	integer
AcidIndex	integer
STARS	integer

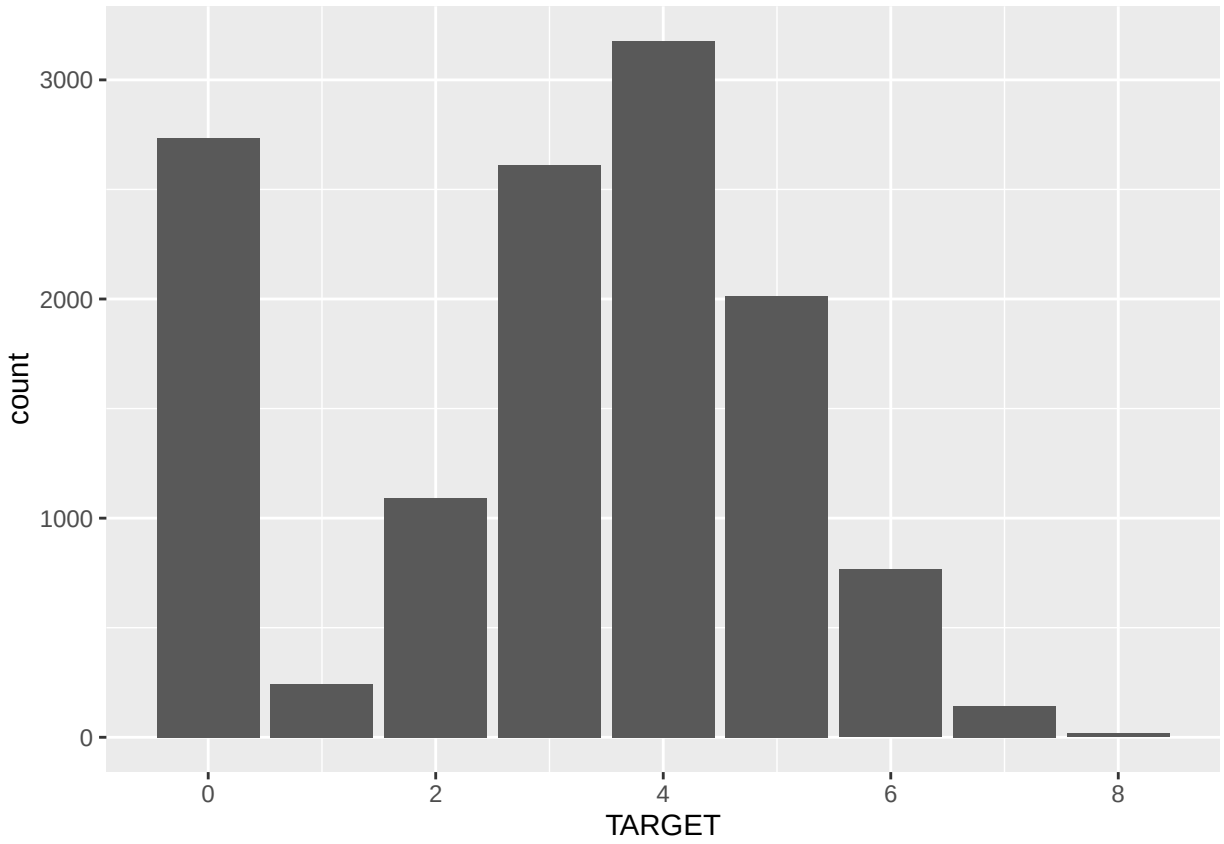
TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides
Min. :0.000	Min. :-18.100	Min. :-2.7900	Min. :-3.2400	Min. :-127.800	Min. :-1.1710
1st Qu.:2.000	1st Qu.: 5.200	1st Qu.: 0.1300	1st Qu.: 0.0300	1st Qu.: -2.000	1st Qu.:-0.0310
Median :3.000	Median : 6.900	Median : 0.2800	Median : 0.3100	Median : 3.900	Median : 0.0460
Mean :3.029	Mean : 7.076	Mean : 0.3241	Mean : 0.3084	Mean : 5.419	Mean : 0.0548
3rd Qu.:4.000	3rd Qu.: 9.500	3rd Qu.: 0.6400	3rd Qu.: 0.5800	3rd Qu.: 15.900	3rd Qu.: 0.1530
Max. :8.000	Max. : 34.400	Max. : 3.6800	Max. : 3.8600	Max. : 141.150	Max. : 1.3510
NA	NA	NA	NA	NA's :616	NA's :638

FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol
Min. :-555.00	Min. :-823.0	Min. :0.8881	Min. :0.480	Min. :-3.1300	Min. :-4.70
1st Qu.: 0.00	1st Qu.: 27.0	1st Qu.:0.9877	1st Qu.:2.960	1st Qu.: 0.2800	1st Qu.: 9.00
Median : 30.00	Median : 123.0	Median :0.9945	Median :3.200	Median : 0.5000	Median :10.40
Mean : 30.85	Mean : 120.7	Mean :0.9942	Mean :3.208	Mean : 0.5271	Mean :10.49
3rd Qu.: 70.00	3rd Qu.: 208.0	3rd Qu.:1.0005	3rd Qu.:3.470	3rd Qu.: 0.8600	3rd Qu.:12.40
Max. : 623.00	Max. :1057.0	Max. :1.0992	Max. :6.130	Max. : 4.2400	Max. :26.50
NA's :647	NA's :682	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

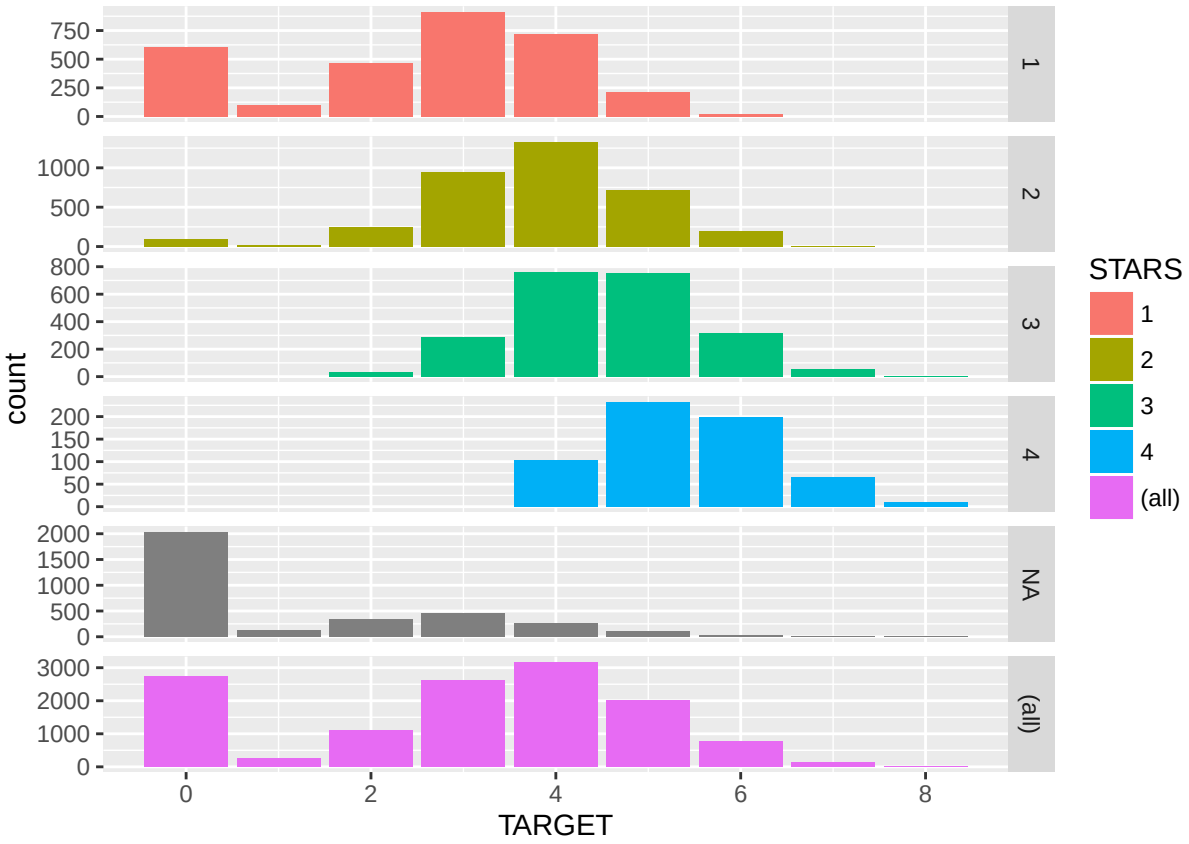
There are numerous NAs in certain variables, and variables with negative values. Variables with negative values have apparently normal distributions so it's possible some previous data adjustments have been made. The variable data with negative values in stable, normal distributions will be used as-is.

Below is a plot of the distribution of counts for the TARGET variable.



Here is another look at the TARGET variable, stratified by the number of Stars rating given for each wine.

```
ggplot(wine, aes(TARGET, fill = STARS)) + geom_bar(stat = "count") + facet_grid(STARS ~
    ., margins = TRUE, scales = "free")
```

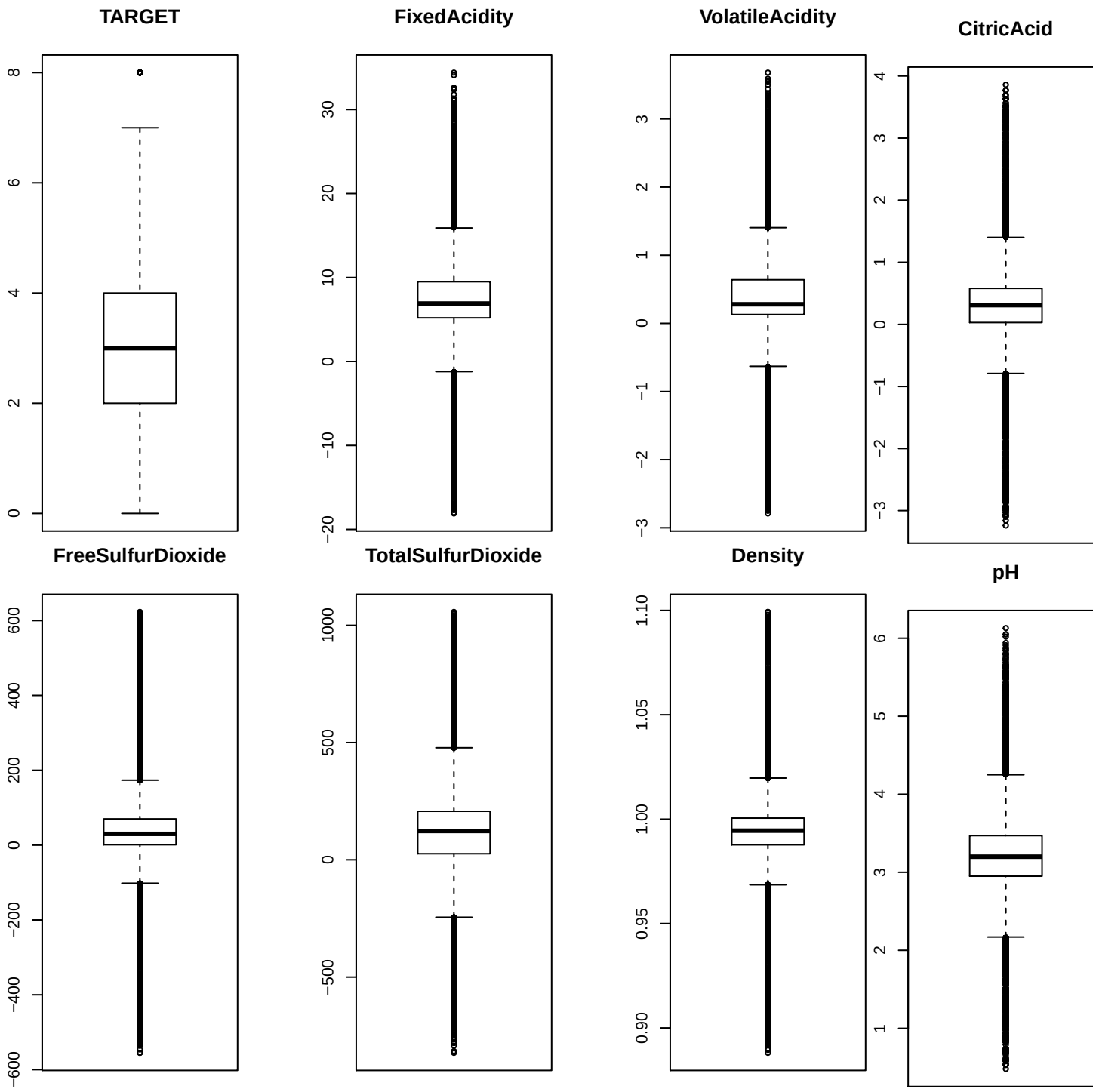


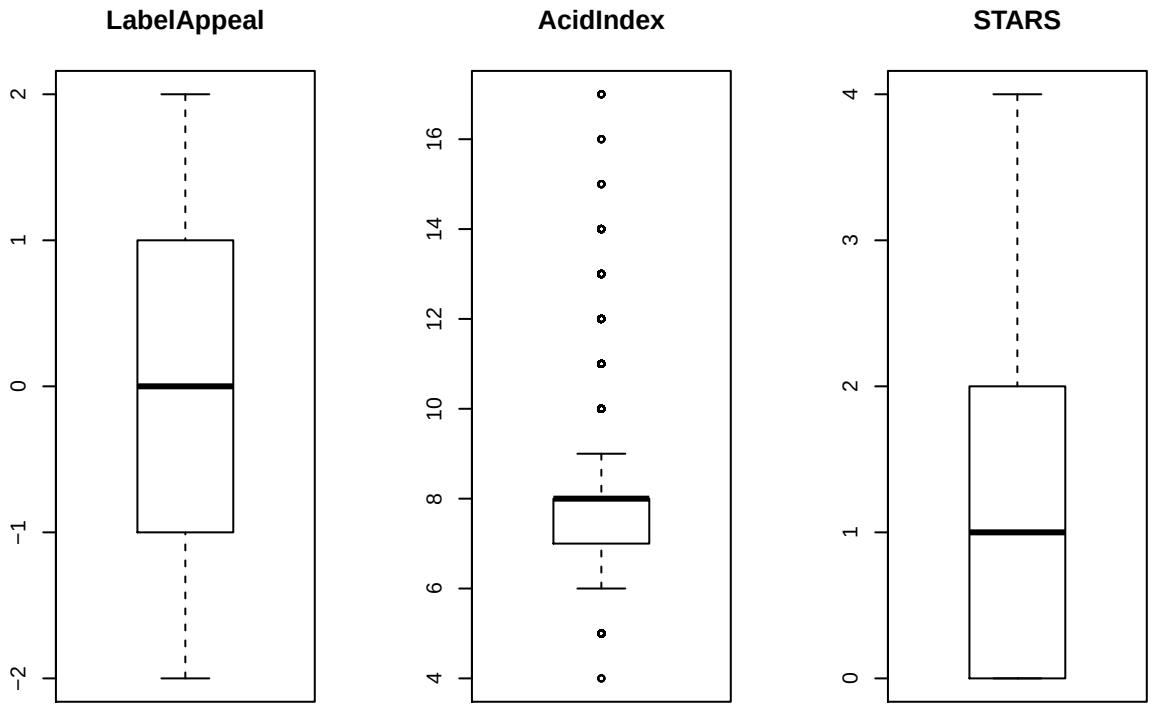
Data Preparation

Data Preparation

We will cleanse the data by removing the index column, using the MICE package to replace NA's with meaningful values, and setting the unrated wines (no stars) to zero stars, so they can be analyzed quantitatively.

Below are boxplots of the independent variables, which illustrate the normality of the data.


TARGET
FixedAcidity
VolatileAcidity
CitricAcid
FreeSulfurDioxide
TotalSulfurDioxide
Density
pH

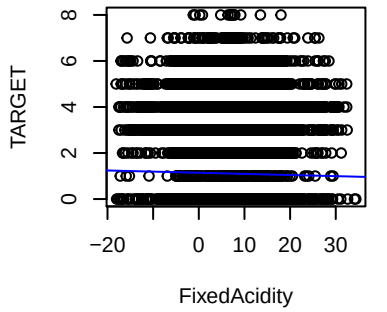


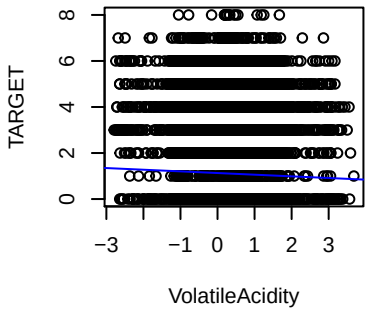
Build Models

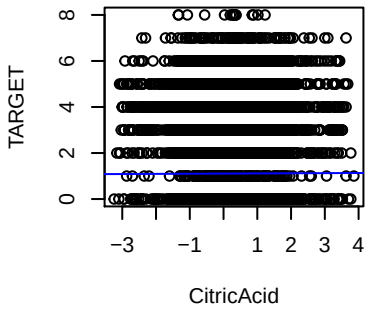
Build Models

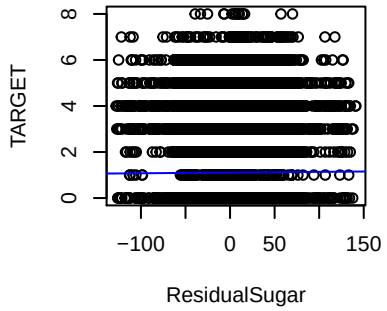
Regular Poisson

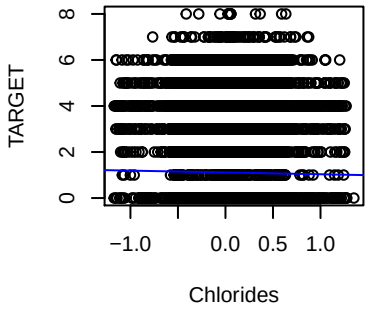
To take a deeper look at the data, first we create a model for each variable individually - to get a sense of how each variable interacts with the outcome on its own, as a means to inform us how we might use groups of variables to build the best models.

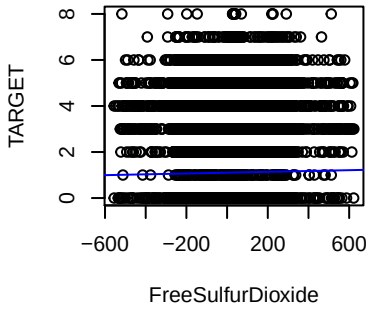

TARGET
0 2 4 6 8
−20 0 10 20 30
FixedAcidity

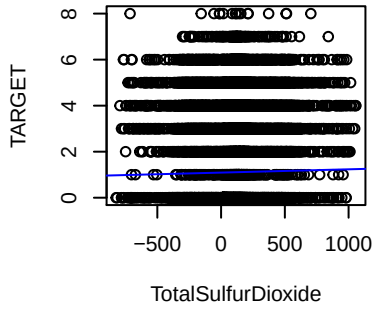

TARGET
0 2 4 6 8
−3 −1 0 1 2 3
VolatileAcidity

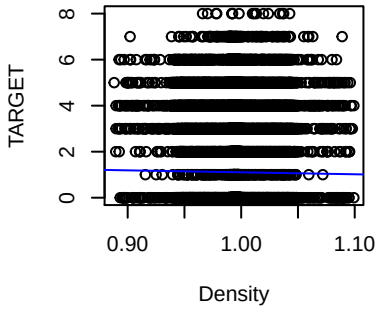

TARGET
0 2 4 6 8
−3 −1 1 2 3 4
CitricAcid

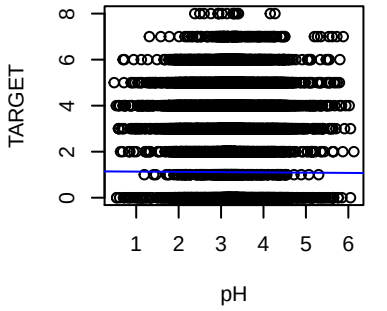

TARGET
0 2 4 6 8
−100 0 50 150
ResidualSugar

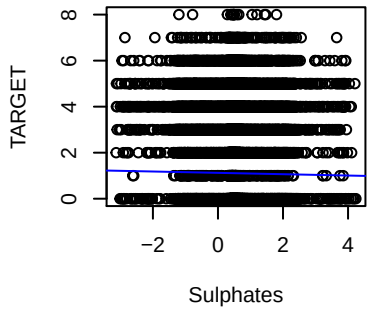

TARGET
0 2 4 6 8
−1.0 0.0 0.5 1.0
Chlorides

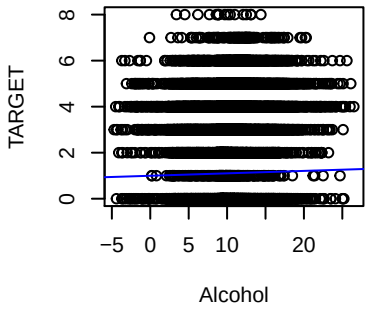

TARGET
0 2 4 6 8
−600 −200 200 600
FreeSulfurDioxide

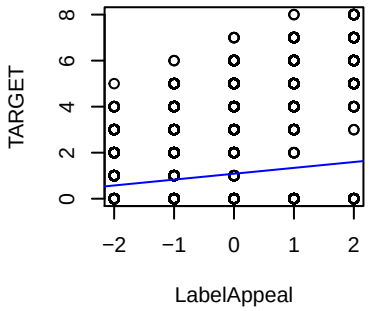

TARGET
0 2 4 6 8
−500 0 500 1000
TotalSulfurDioxide

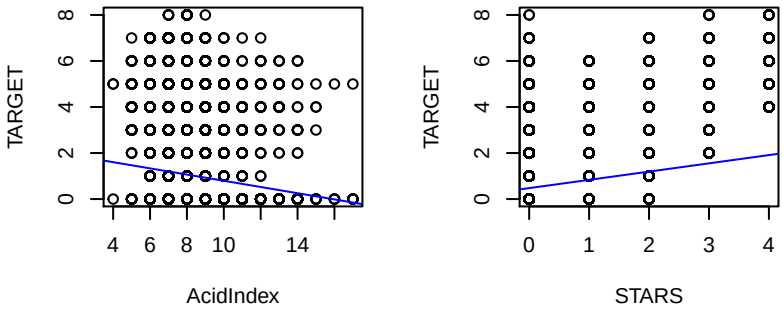

TARGET
0 2 4 6 8
0.90 1.00 1.10
Density


TARGET
0 2 4 6 8
1 2 3 4 5 6
pH


TARGET
0 2 4 6 8
−2 0 2 4
Sulphates


TARGET
0 2 4 6 8
−5 0 5 10 20
Alcohol


TARGET
0 2 4 6 8
−2 −1 0 1 2
LabelAppeal



By looking at these models we suspect there may be two forces at work. The first we will call Perception. The two Perception variables are Stars and Label Appeal. Based on the high coefficients and high significance, Perception seems to impact the outcome much more than anything else. The second force we will call Chemistry. All the other variables could belong to this group. The pattern we see here is that the best outcome (highest number of cases purchased) tends to occur when the Chemistry variables are close to the mean.

Next we will create a generalized linear model, Poisson family, that combines all the variables:

```
##
## Call:
## glm(formula = as.formula(paste(colnames(wine)[1], "~", paste(colnames(wine)[-1],
##     collapse = "+"), sep = "")), family = poisson(), data = wine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9694  -0.7225   0.0673   0.5787   3.2259
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.521e+00  1.954e-01   7.788 6.81e-15 ***
## FixedAcidity        -3.047e-04  8.206e-04  -0.371 0.710378
## VolatileAcidity     -3.336e-02  6.516e-03  -5.120 3.06e-07 ***
## CitricAcid           7.720e-03  5.893e-03   1.310 0.190241
## ResidualSugar        2.742e-05  1.509e-04   0.182 0.855824
## Chlorides           -4.251e-02  1.597e-02  -2.661 0.007789 **
## FreeSulfurDioxide    1.175e-04  3.427e-05   3.430 0.000604 ***
## TotalSulfurDioxide   8.710e-05  2.215e-05   3.933 8.38e-05 ***
## Density             -2.797e-01  1.920e-01  -1.457 0.145029
## pH                  -1.582e-02  7.519e-03  -2.104 0.035413 *
## Sulphates           -1.435e-02  5.498e-03  -2.609 0.009077 **
## Alcohol              2.427e-03  1.373e-03   1.768 0.077074 .
## LabelAppeal          1.332e-01  6.063e-03  21.971  < 2e-16 ***
## AcidIndex           -8.693e-02  4.549e-03 -19.109  < 2e-16 ***
## STARS                3.111e-01  4.531e-03  68.658  < 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: 14724  on 12780  degrees of freedom
## AIC: 46696
##
## Number of Fisher Scoring iterations: 5
```

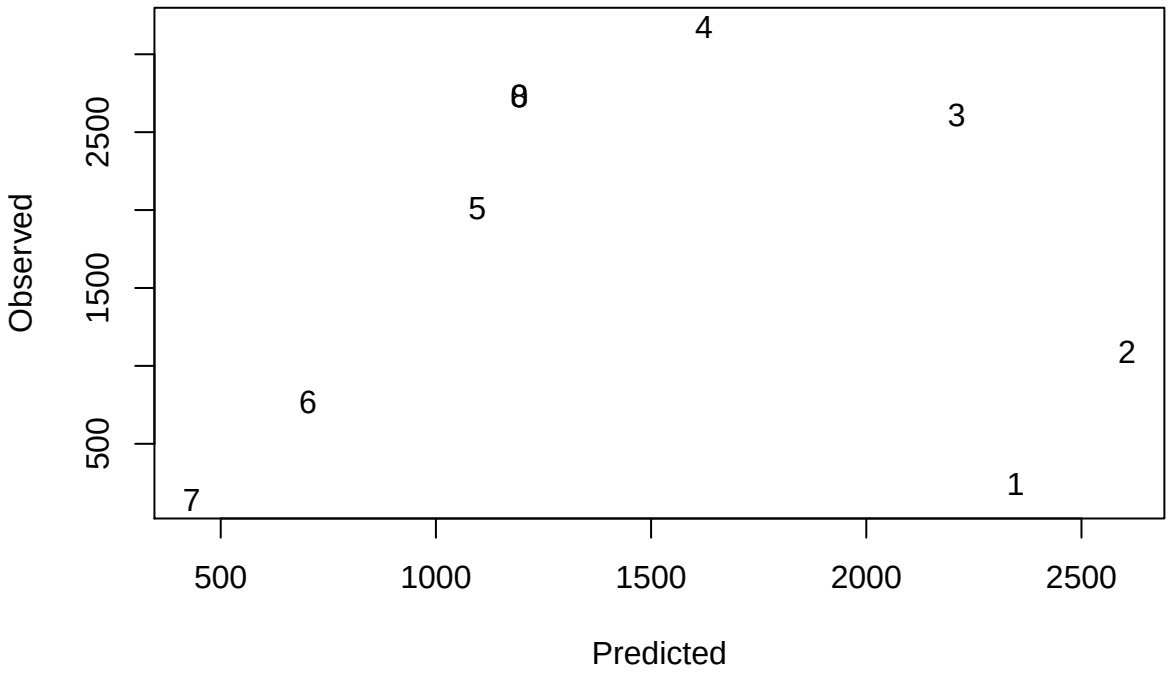
Here we see that the Perception variables have an outsize impact on the outcome.

Let's create a Poisson model using only the two Perception variables:

```
##
## Call:
## glm(formula = TARGET ~ STARS + LabelAppeal, family = poisson(),
##     data = wine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8852  -0.7533   0.0842   0.6161   3.2856
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.516912   0.010057   51.40   <2e-16 ***
## STARS       0.329083   0.004437   74.16   <2e-16 ***
## LabelAppeal 0.125476   0.006042   20.77   <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: 15221  on 12792  degrees of freedom
## AIC: 47169
##
## Number of Fisher Scoring iterations: 5
```

Zero-inflated Poisson Model

We next explore the seemingly high number of zero cases in the TARGET count as seen in the previous histogram. We can easily see if the number of zeros observed is in line with the number of zeros predicted by the Poisson model alone.



The number of observed zero cases and the predicted zero cases do not match up well so we'll move to look at the influence of the zero counts on the model by separating out the modeling of zero counts and the modeling of the non-zero counts.

Staying with our concepts of Perception and Chemistry, we will look treating the high number of zero counts using the Perception variables of STARS and LabelAppeal, and the non-zero counts will use all other variables as the Chemistry variables.

```
##
## Call:
## zeroinfl(formula = TARGET ~ . - (STARS + LabelAppeal) | STARS +
##     LabelAppeal, data = wine, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -1.96050 -0.49226  0.04247  0.52347  4.77891
##
## Count model coefficients (poisson with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.860e+00  2.012e-01   9.247  < 2e-16 ***
## FixedAcidity         2.115e-05  8.349e-04   0.025  0.97979
## VolatileAcidity     -2.360e-02  6.679e-03  -3.534  0.00041 ***
## CitricAcid           4.169e-03  6.025e-03   0.692  0.48901
## ResidualSugar       -7.931e-07  1.536e-04  -0.005  0.99588
## Chlorides           -2.122e-02  1.630e-02  -1.302  0.19285
## FreeSulfurDioxide    3.839e-05  3.452e-05   1.112  0.26617
## TotalSulfurDioxide -1.696e-05  2.190e-05  -0.774  0.43872
## Density             -4.156e-01  1.977e-01  -2.102  0.03557 *
## pH                   7.972e-03  7.691e-03   1.037  0.29995
## Sulphates           -1.972e-03  5.638e-03  -0.350  0.72650
## Alcohol              8.918e-03  1.384e-03   6.444 1.16e-10 ***
## AcidIndex           -2.992e-02  5.041e-03  -5.936 2.93e-09 ***
##
```

```
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.57361    0.03747   15.31   <2e-16 ***
## STARS       -2.28644    0.05259  -43.48   <2e-16 ***
## LabelAppeal  0.55884    0.03628   15.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 22
## Log-likelihood: -2.196e+04 on 16 Df
```

```
## [1] 3110.048
```

```
## [1] "Chi-Square Test =  0.496382636947722"
```

Given the large p-value from the chi-square test, we conclude our model approach for Chemsitry vs Perception is valid.

After analyzing the p-values for the Chemistry portion of the zero-inflated model, there are only 4 statistically significant variables: VolatileAcidity, Density, Alcohol, and AcidIndex. We'll re-reun the zero-inflated Poisson model with just these variables in the Poisson portion.

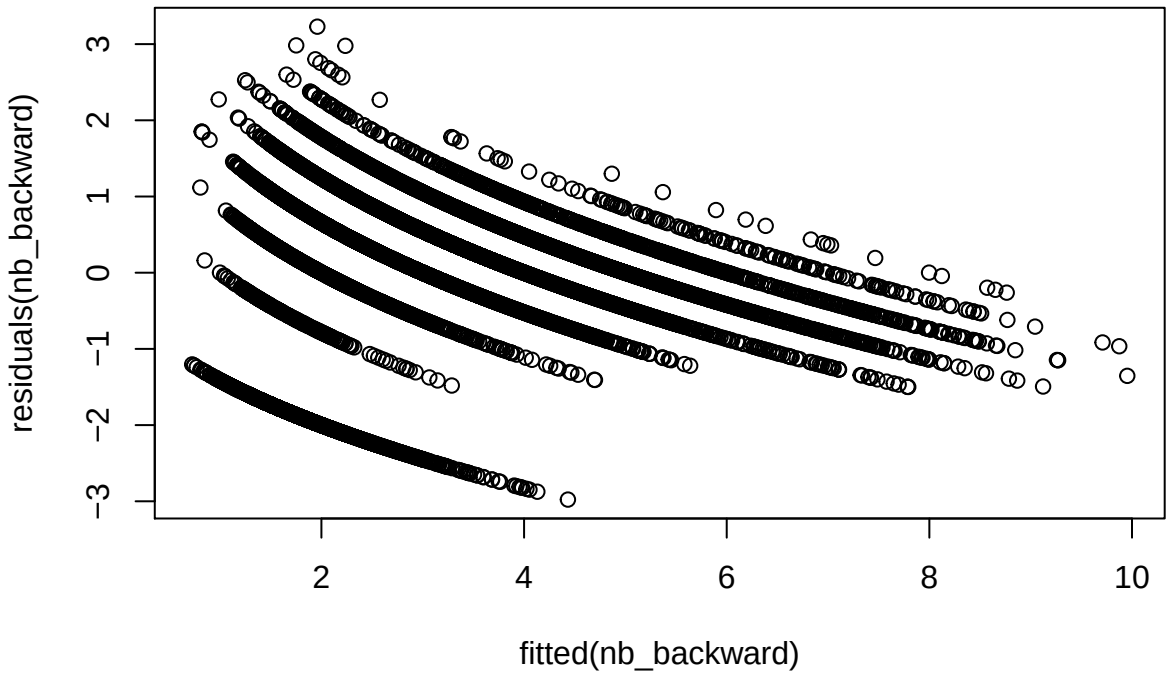
```
##
## Call:
## zeroinfl(formula = TARGET ~ (VolatileAcidity + Density + Alcohol +
##     AcidIndex) - (STARS + LabelAppeal) | STARS + LabelAppeal, data = wine,
##     dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -1.95756 -0.49178  0.04267  0.52820  4.79578
##
## Count model coefficients (poisson with log link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.891862   0.199179   9.498  < 2e-16 ***
## VolatileAcidity -0.023753   0.006677  -3.557 0.000375 ***
## Density         -0.421259   0.197592  -2.132 0.033010 *
## Alcohol          0.008953   0.001383   6.475 9.46e-11 ***
## AcidIndex       -0.030184   0.004977  -6.065 1.32e-09 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.57366    0.03747   15.31   <2e-16 ***
## STARS       -2.28638    0.05257  -43.49   <2e-16 ***
## LabelAppeal  0.55907    0.03628   15.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 13
## Log-likelihood: -2.196e+04 on 8 Df
```

We have reduced the degrees-of-freedom from 16 down to 8 which is as far as we'll go with the zero-inflated Poisson model.

Regular Negative Binomial Model

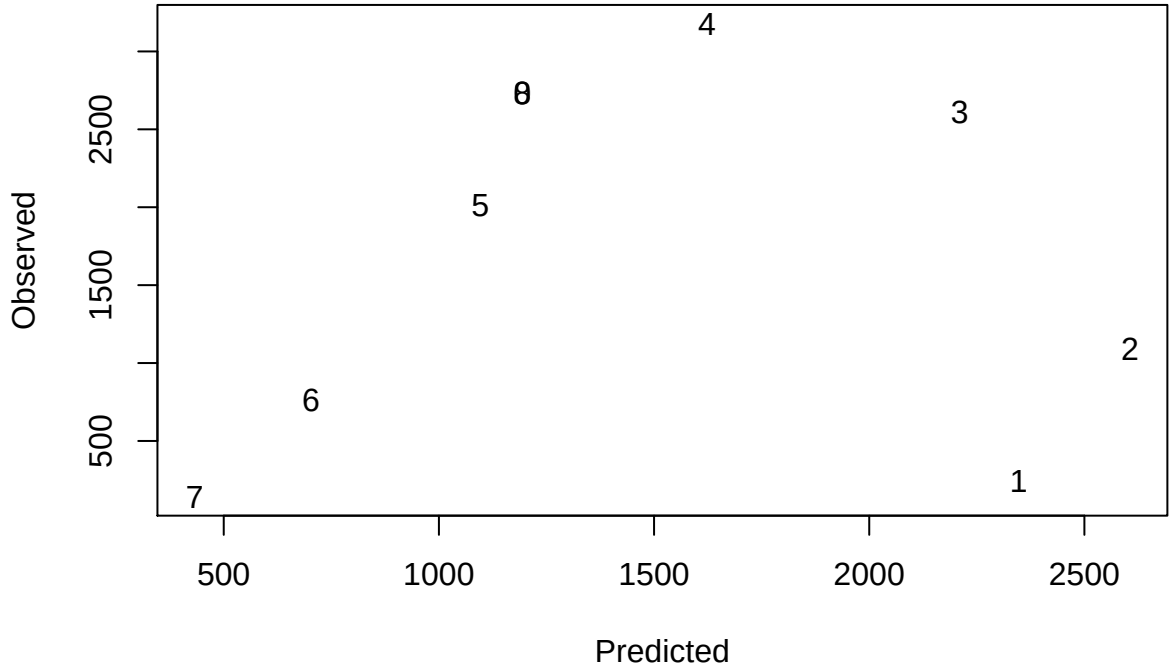
For regular negative binomial model, we start with all the dependent variables, and perform a backward stepwise algorithm. Initially, we have 14 dependent variables; using this process we reduce to 10 variables. The AIC is 46692

```
##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS, data = wine, init.theta = 48985.97639,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9775  -0.7247   0.0696   0.5808   3.2287
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.524e+00  1.953e-01   7.802 6.12e-15 ***
## VolatileAcidity     -3.356e-02  6.515e-03  -5.150 2.60e-07 ***
## Chlorides           -4.271e-02  1.597e-02  -2.674 0.007506 **
## FreeSulfurDioxide    1.177e-04  3.426e-05   3.436 0.000591 ***
## TotalSulfurDioxide   8.720e-05  2.214e-05   3.939 8.19e-05 ***
## Density             -2.834e-01  1.919e-01  -1.477 0.139772
## pH                  -1.575e-02  7.518e-03  -2.095 0.036211 *
## Sulphates           -1.450e-02  5.496e-03  -2.639 0.008323 **
## Alcohol              2.470e-03  1.372e-03   1.801 0.071775 .
## LabelAppeal          1.333e-01  6.063e-03  21.979  < 2e-16 ***
## AcidIndex           -8.682e-02  4.493e-03 -19.324  < 2e-16 ***
## STARS                3.112e-01  4.530e-03  68.688  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48985.98) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 14725  on 12783  degrees of freedom
## AIC: 46694
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48986
##           Std. Err.:  50730
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -46667.84
```



Zero-inflated Negative Binomial Regession Model

We'll continue our exploration of the seemingly high number of zero cases in the TARGET count as seen in the previous histogram. In this case, we'll see if the number of zeros observed is in line with the number of zeros predicted by the negative binomial model alone.



The number of observed zero cases and the predicted zero cases do not match up well so we'll move to look at the influence of the zero counts on the model by separating out the modeling of zero counts and the modeling of the non-zero counts.

Staying with our concepts of Perception and Chemistry, we will look at treating the high number of zero

counts using the Perception variables of STARS and LabelAppeal, and the non-zero counts will use all other variables as the Chemistry variables.

```
##
## Call:
## zeroinfl(formula = TARGET ~ . - (STARS + LabelAppeal) | (STARS +
##     LabelAppeal), data = wine, dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -1.96051 -0.49226  0.04247  0.52346  4.77883
##
## Count model coefficients (negbin with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.861e+00  2.012e-01   9.247  < 2e-16 ***
## FixedAcidity         2.122e-05  8.349e-04   0.025  0.97972
## VolatileAcidity     -2.360e-02  6.679e-03  -3.534  0.00041 ***
## CitricAcid           4.169e-03  6.025e-03   0.692  0.48903
## ResidualSugar       -8.129e-07  1.536e-04  -0.005  0.99578
## Chlorides           -2.122e-02  1.630e-02  -1.302  0.19288
## FreeSulfurDioxide    3.838e-05  3.452e-05   1.112  0.26623
## TotalSulfurDioxide -1.695e-05  2.190e-05  -0.774  0.43894
## Density             -4.157e-01  1.977e-01  -2.102  0.03551 *
## pH                   7.972e-03  7.691e-03   1.037  0.29996
## Sulphates           -1.972e-03  5.638e-03  -0.350  0.72652
## Alcohol              8.918e-03  1.384e-03   6.444 1.16e-10 ***
## AcidIndex           -2.992e-02  5.041e-03  -5.936 2.93e-09 ***
## Log(theta)           1.725e+01  8.998e+00   1.917  0.05519 .
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.57361    0.03747   15.31   <2e-16 ***
## STARS       -2.28642    0.05259  -43.48   <2e-16 ***
## LabelAppeal  0.55882    0.03628   15.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 31088483.5665
## Number of iterations in BFGS optimization: 60
## Log-likelihood: -2.196e+04 on 17 Df
```

```
## [1] 3110.049
```

```
## [1] "Chi-Square Test =  0.496381668017019"
```

Given the large p-value from the chi-square test, we conclude our model approach for Chemistry versus Perception is valid.

After analyzing the p-values for the Chemistry portion of the zero-inflated model, there are only four statistically significant variables: VolatileAcidity, Density, Alcohol, and AcidIndex. We'll re-reun the zero-inflated Poisson model with just these variables in the negative binomial portion.

```
##
```

```
## Call:
## zeroinfl(formula = TARGET ~ (VolatileAcidity + Density + Alcohol +
##     AcidIndex) - (STARS + LabelAppeal) | STARS + LabelAppeal, data = wine,
##     dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -1.95756 -0.49179  0.04266  0.52822  4.79556
##
## Count model coefficients (negbin with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.891982   0.199179   9.499  < 2e-16 ***
## VolatileAcidity -0.023758   0.006677  -3.558 0.000374 ***
## Density         -0.421433   0.197592  -2.133 0.032937 *
## Alcohol          0.008954   0.001383   6.476 9.40e-11 ***
## AcidIndex       -0.030179   0.004977  -6.064 1.33e-09 ***
## Log(theta)      15.269348   8.542154   1.788 0.073852 .
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.57362    0.03747   15.31   <2e-16 ***
## STARS       -2.28626    0.05256  -43.50   <2e-16 ***
## LabelAppeal  0.55904    0.03628   15.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 4279504.8666
## Number of iterations in BFGS optimization: 39
## Log-likelihood: -2.196e+04 on 9 Df
```

We have reduced the degrees-of-freedom from 17 down to 9 which is as far as we'll go with the zero-inflated negative binomial model.

Select Models

Linear Regression Models:

Lastly we are going to look at a regular linear model, as a comparison to the analysis shown above. Again we are going to compare Chemistry vs. Perception.

```
lin.mod.perc <- lm(TARGET ~ . - STARS - LabelAppeal,data = wine)
summary(lin.mod.perc)
```

```
##
## Call:
## lm(formula = TARGET ~ . - STARS - LabelAppeal, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5958 -1.3093  0.2833  1.3240  5.5732
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.409e+00  6.255e-01  11.845  < 2e-16 ***
## FixedAcidity        -1.183e-03  2.637e-03  -0.449  0.65361
## VolatileAcidity     -1.876e-01  2.093e-02  -8.963  < 2e-16 ***
## CitricAcid           4.570e-02  1.906e-02   2.398  0.01649 *
## ResidualSugar        7.015e-04  4.855e-04   1.445  0.14850
## Chlorides           -1.833e-01  5.125e-02  -3.576  0.00035 ***
## FreeSulfurDioxide    4.168e-04  1.104e-04   3.777  0.00016 ***
## TotalSulfurDioxide   3.253e-04  7.078e-05   4.596 4.36e-06 ***
## Density             -1.743e+00  6.181e-01  -2.820  0.00481 **
## pH                  -6.855e-02  2.417e-02  -2.837  0.00456 **
## Sulphates           -6.982e-02  1.766e-02  -3.954 7.74e-05 ***
## Alcohol              2.782e-02  4.401e-03   6.323 2.66e-10 ***
## AcidIndex           -3.438e-01  1.269e-02 -27.094  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.853 on 12782 degrees of freedom
## Multiple R-squared:  0.07601,    Adjusted R-squared:  0.07514
## F-statistic: 87.63 on 12 and 12782 DF,  p-value: < 2.2e-16
```

We can see that not all the chemistry variables are significant, using backward stepwise elimination, so we eliminate the insignificant independent variables.

```
lin.mod.back <- step(lin.mod.perc)
```

```
## Start:  AIC=15791.22
## TARGET ~ (FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS) -
##     STARS - LabelAppeal
##
##                       Df Sum of Sq   RSS   AIC
## - FixedAcidity         1      0.69 43869 15789
## <none>                              43868 15791
## - ResidualSugar        1      7.17 43876 15791
## - CitricAcid           1     19.74 43888 15795
## - Density              1     27.29 43896 15797
## - pH                   1     27.62 43896 15797
## - Chlorides            1     43.90 43912 15802
## - FreeSulfurDioxide    1     48.95 43917 15804
## - Sulphates            1     53.64 43922 15805
## - TotalSulfurDioxide   1     72.48 43941 15810
## - Alcohol              1    137.20 44006 15829
## - VolatileAcidity      1    275.73 44144 15869
## - AcidIndex            1   2519.47 46388 16504
##
## Step:  AIC=15789.42
## TARGET ~ VolatileAcidity + CitricAcid + ResidualSugar + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + AcidIndex
##
##                       Df Sum of Sq   RSS   AIC
```

```
## <none>                           43869 15789
## - ResidualSugar       1      7.24 43876 15790
## - CitricAcid          1     19.72 43889 15793
## - Density             1     27.29 43896 15795
## - pH                  1     27.64 43897 15796
## - Chlorides           1     43.87 43913 15800
## - FreeSulfurDioxide   1     48.79 43918 15802
## - Sulphates           1     54.07 43923 15803
## - TotalSulfurDioxide  1     72.63 43942 15809
## - Alcohol             1    137.19 44006 15827
## - VolatileAcidity     1    275.85 44145 15868
## - AcidIndex           1   2614.75 46484 16528
```

```
lin.mod.back <- lm(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
    TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
    AcidIndex, data = wine)
summary(lin.mod.back)
```

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     AcidIndex, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6826 -1.3032  0.2837  1.3299  5.6247
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.430e+00  6.255e-01  11.878  < 2e-16 ***
## VolatileAcidity    -1.888e-01  2.092e-02  -9.022  < 2e-16 ***
## Chlorides          -1.854e-01  5.125e-02  -3.617 0.000299 ***
## FreeSulfurDioxide   4.208e-04  1.103e-04   3.813 0.000138 ***
## TotalSulfurDioxide  3.288e-04  7.077e-05   4.646 3.42e-06 ***
## Density            -1.763e+00  6.181e-01  -2.852 0.004355 **
## pH                 -6.842e-02  2.417e-02  -2.831 0.004649 **
## Sulphates          -7.089e-02  1.765e-02  -4.016 5.96e-05 ***
## Alcohol             2.793e-02  4.400e-03   6.347 2.26e-10 ***
## AcidIndex          -3.429e-01  1.246e-02 -27.513  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.853 on 12785 degrees of freedom
## Multiple R-squared:  0.07543,    Adjusted R-squared:  0.07478
## F-statistic: 115.9 on 9 and 12785 DF,  p-value: < 2.2e-16
```

Comparing this to just the perception data we can see the following:

```
lin.mod.app <- lm(TARGET ~ STARS + LabelAppeal,data = wine)
summary(lin.mod.app)
```

```
##
```

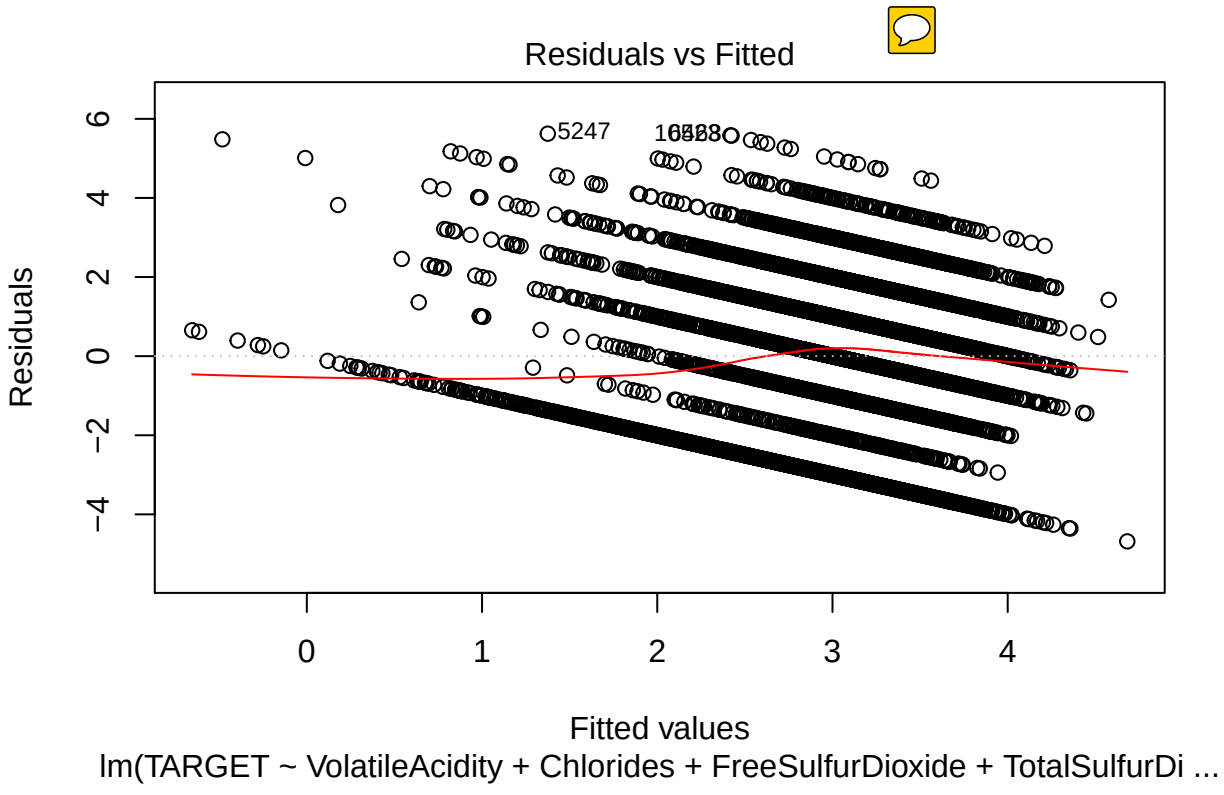
```
## Call:
## lm(formula = TARGET ~ STARS + LabelAppeal, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3568 -1.0721  0.0184  0.9279  6.1139
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.47910    0.01988   74.41   <2e-16 ***
## STARS        1.03182    0.01050   98.30   <2e-16 ***
## LabelAppeal  0.40701    0.01398   29.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.359 on 12792 degrees of freedom
## Multiple R-squared:  0.5027, Adjusted R-squared:  0.5026
## F-statistic:  6466 on 2 and 12792 DF,  p-value: < 2.2e-16
```

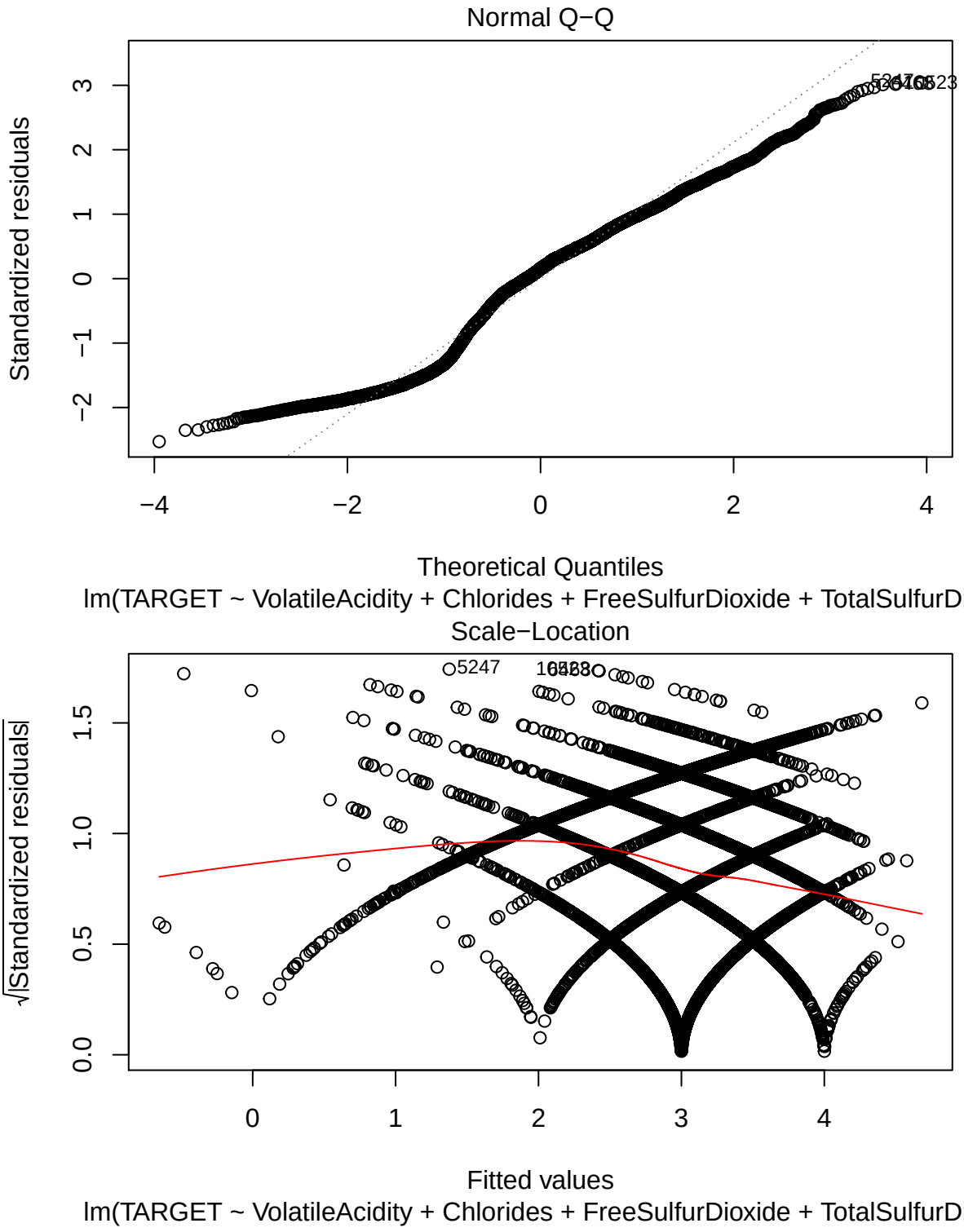
From the the models, we can clearly see that the perception data was a much more appropriate fit. This can be seen through the R-Squared value, which shows that the perception model explains roughly 50% of the variance in the model. This is a pretty "good-fit." Using the chemistry data, we also see a significant model, however, the fit is much worse, with practically none of the variance explained.

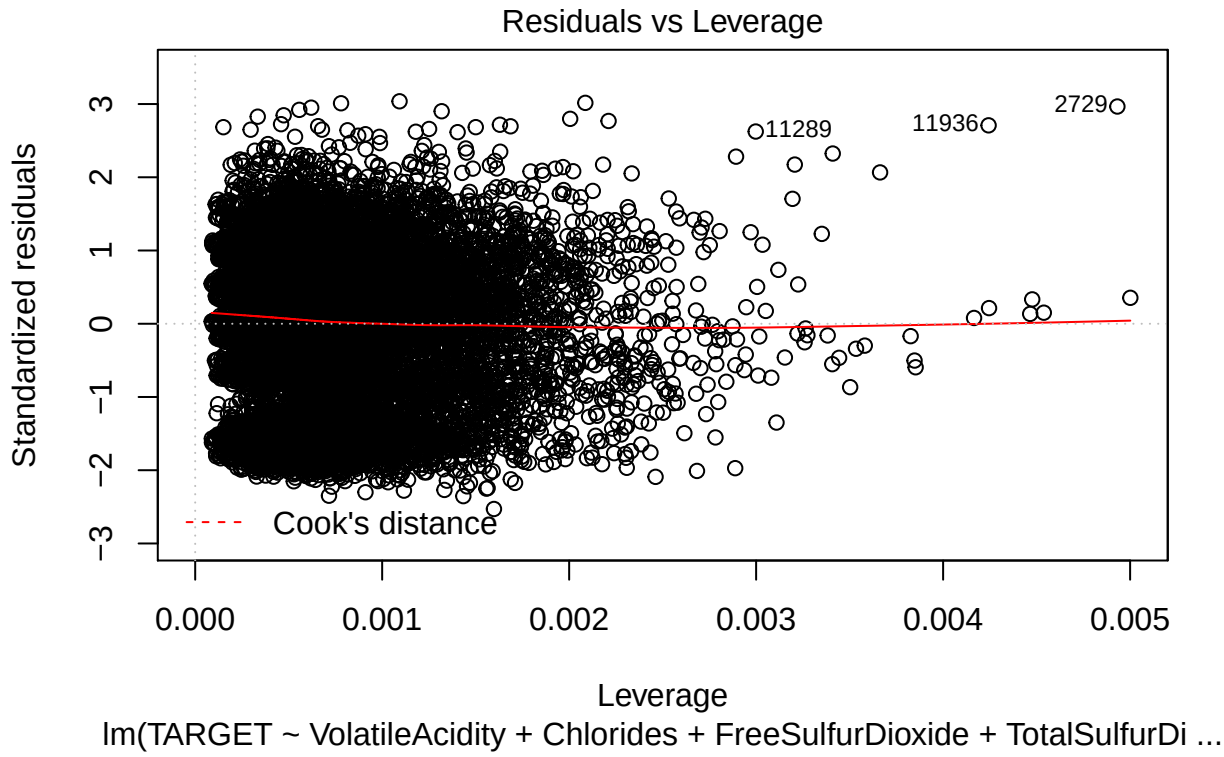
Checking the residuals we can see the following:

Chemistry model:

```
plot(lin.mod.back)
```

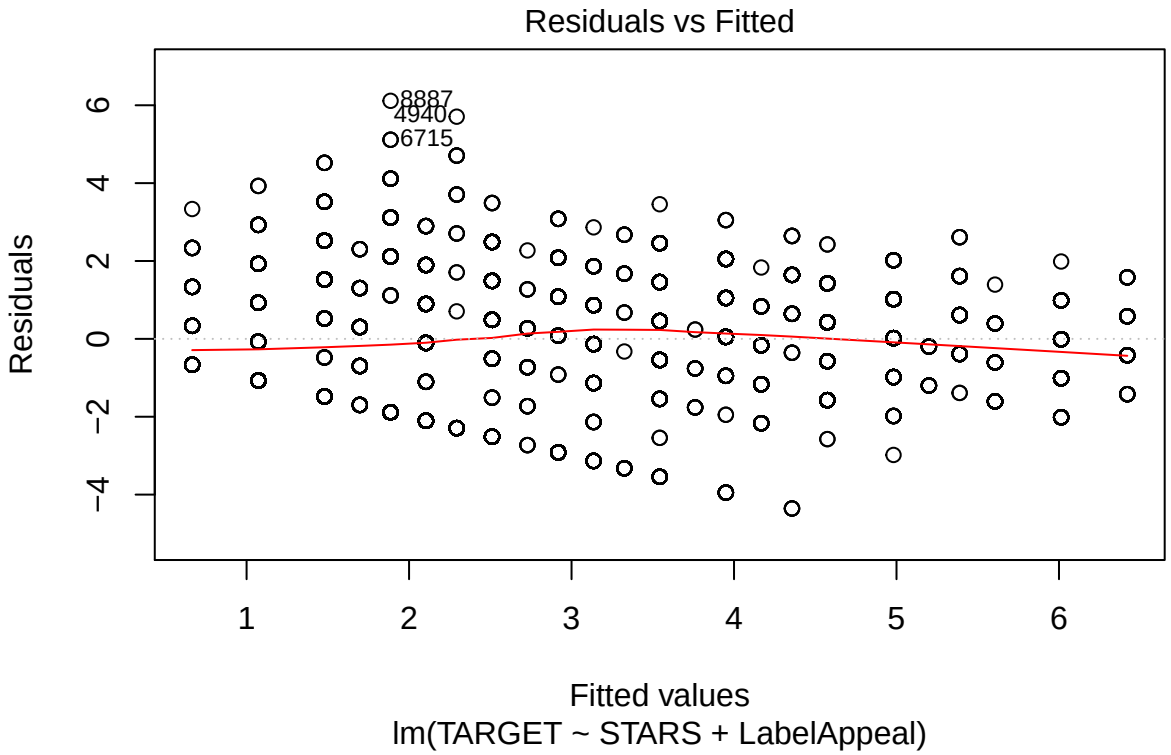


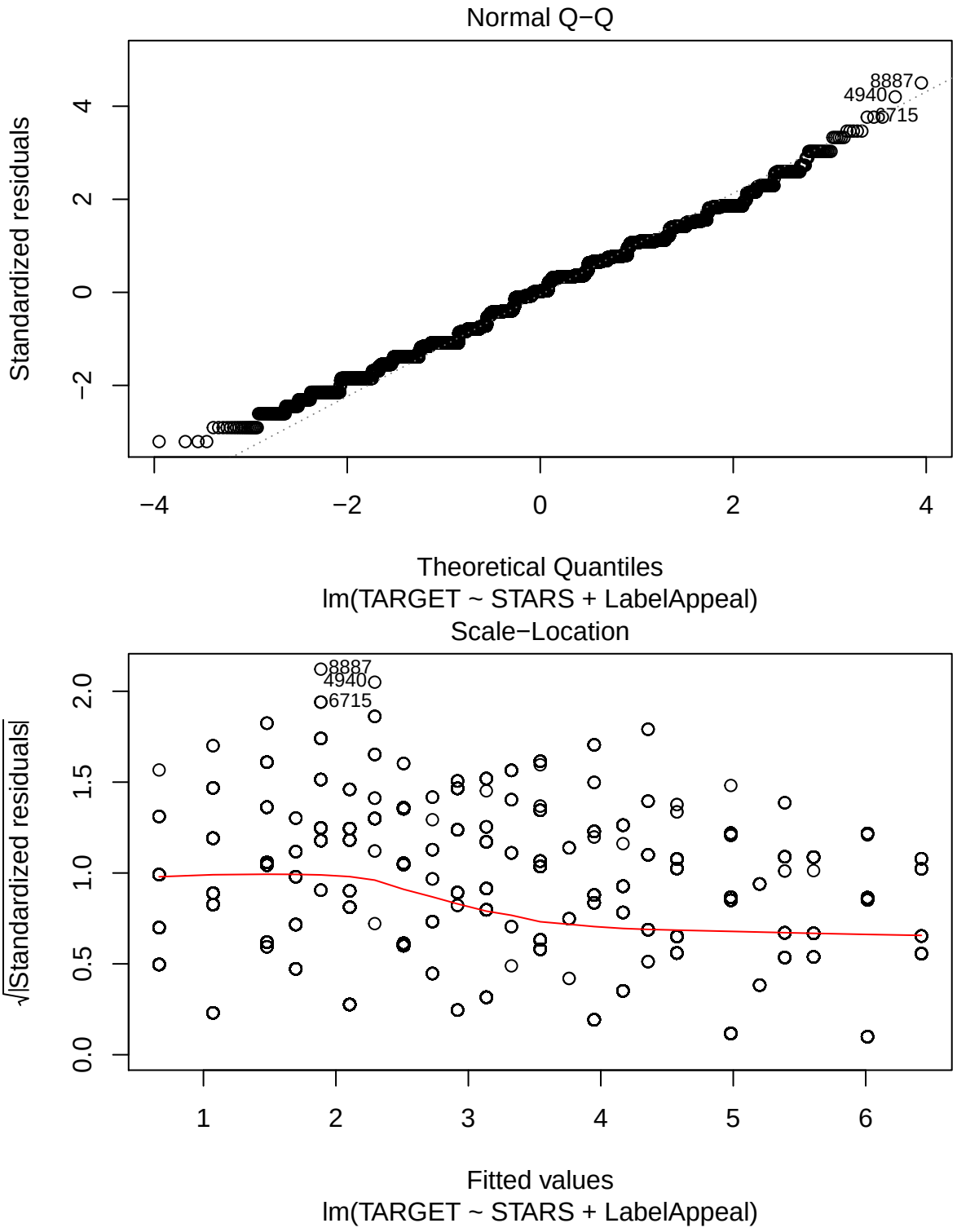

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide + TotalSulfurDi ...
Scale−Location
√|Standardized residuals|
Fitted values
lm(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide + TotalSulfurDi ...

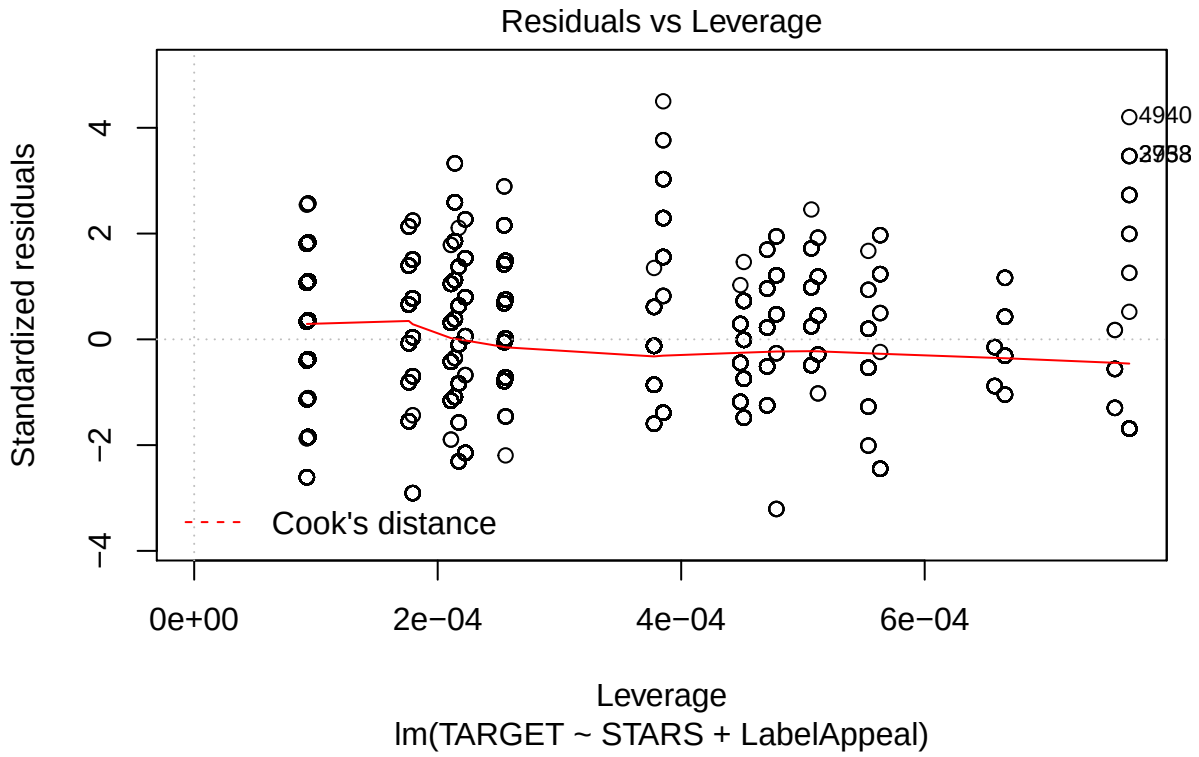


Perception Model:

```
plot(lin.mod.app)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TARGET ~ STARS + LabelAppeal)
8887
4940
6715
Scale−Location
√|Standardized residuals|
Fitted values
lm(TARGET ~ STARS + LabelAppeal)



From the plots above, we can clearly see that the first chemistry model shows clear patterns in the residuals, which indicates that linear modeling is not at all a good choice for these particular variables. However, just using the perception variables we see a much better picture, with a more random residual distribution with no clear patterns that would suggest another choice in models.

Select Models

To compare all our regular models first, we build a dataframe which contains all the performance parameters of the models. Out of the four regular models, it is clear that the regular linear model with focus on perception is the best model. It has the lowest AIC and BIC. The Log-Likelihood is also the highest among the four.

```
##                       Models      AIC      BIC    LogLik Deviance
## 1            Regular Poisson 47168.98 47191.35 -23581.49 15220.96
## 2 Regular Negative Binomial 46693.84 46790.78 -23333.92 14725.03
## 3    Regular Linear Science 52105.86 52187.88 -26041.93 43895.82
## 4 Regular Linear Perception 44157.01 44186.84 -22074.51 23609.92
##   df.residual
## 1       12792
## 2       12783
## 3       12785
## 4       12792
```

When we compare the two zero-inflated models against each other, the following code tells us that the performance differences between two zero inflated models (in terms of LogLik) is not statistically significant since the p value is 0.9595, which is much higher than the significance level 0.05. Their Log Likelihood are both -21962, we have to compare some other performance parameters such as AIC. zero inflated poisson model has slightly smaller AIC (43939.9) compare to the other one(43941.9)

```
#Compare two zero-inflated models
lmtest::lrtest(zn.simplified, zp.simplified)
```

```
## Likelihood ratio test
##
## Model 1: TARGET ~ (VolatileAcidity + Density + Alcohol + AcidIndex) -
##     (STARS + LabelAppeal) | STARS + LabelAppeal
## Model 2: TARGET ~ (VolatileAcidity + Density + Alcohol + AcidIndex) -
##     (STARS + LabelAppeal) | STARS + LabelAppeal
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   9 -21963
## 2   8 -21963 -1 0.0049     0.9439
```

```
AIC(zn.simplified)
```

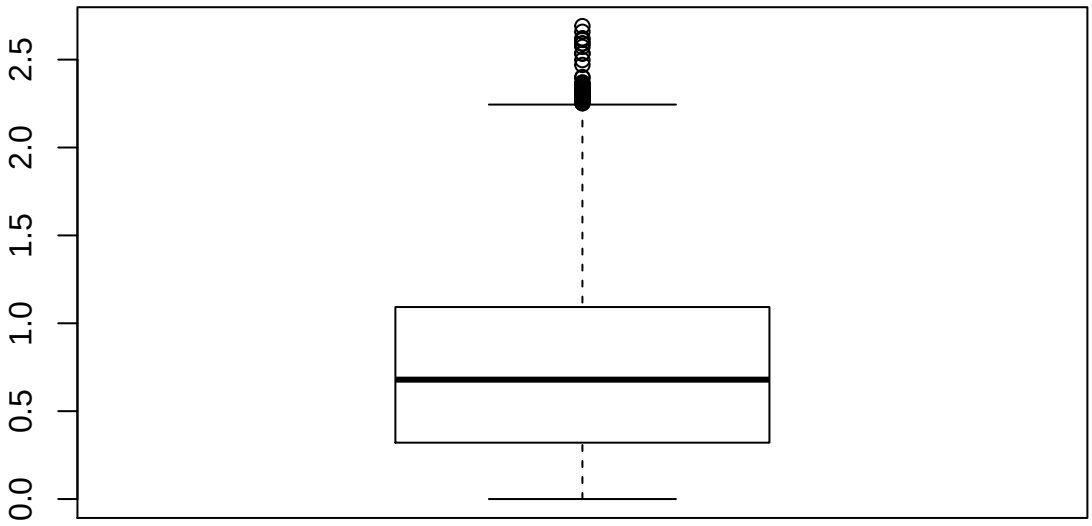
```
## [1] 43943.39
```

```
AIC(zp.simplified)
```

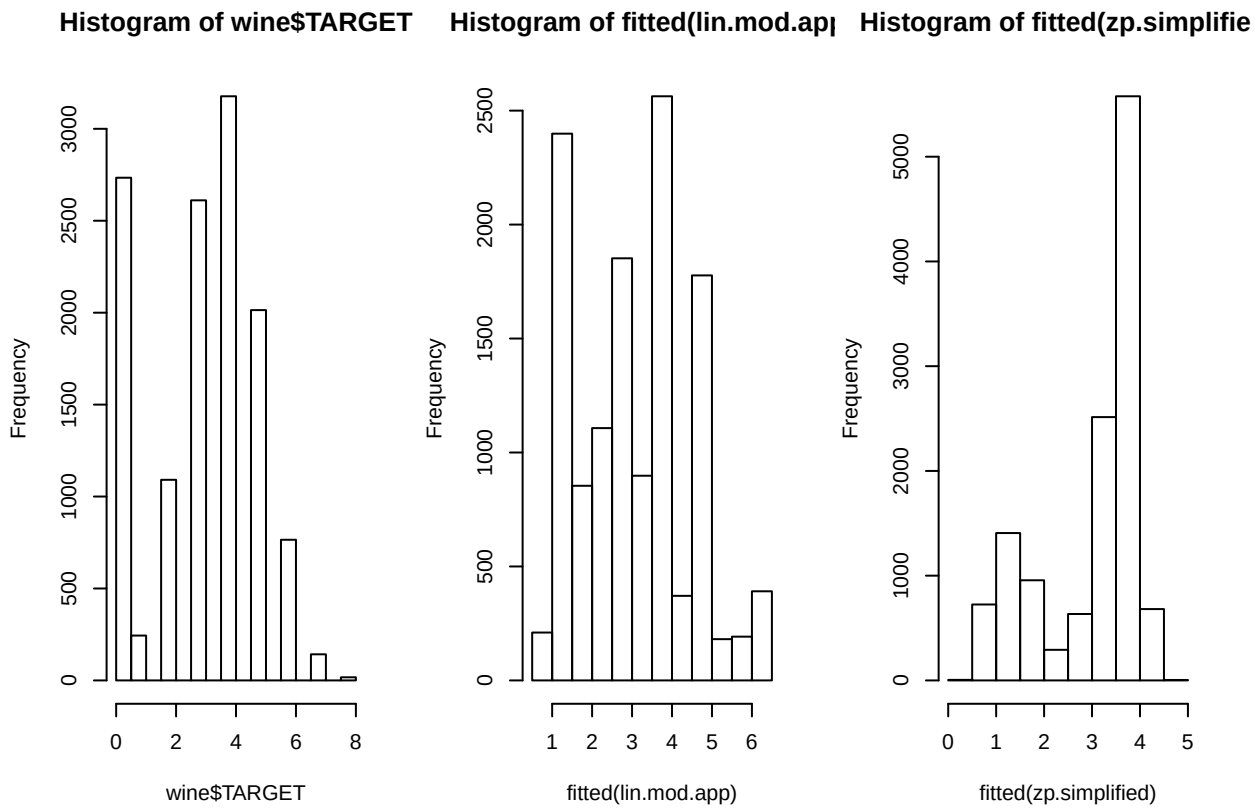
```
## [1] 43941.39
```

By comparing the regular linear model with focus on perception to the zero inflated poisson model, the histogram shows that zero inflated model takes good care of those structural zeros, which are not really zero but more like out of scope. Both models generate predictions that peak at 4 cases of wine, which correspond to the actual observation. Another thing we notice is that the predictions made by two models differ quite significantly. It is recognized according to the boxplot of the absolute differences between two models' residuals. However, based on AIC and Log Likelihood, zero inflated poisson model is still the winner here.

```
boxplot(abs(resid(lin.mod.app) - resid(zp.simplified)))
```




```
par(mfcol=c(1,3))
hist(wine$TARGET)
hist(fitted(lin.mod.app))
hist(fitted(zp.simplified))
```




```
AIC(lin.mod.app)
```

```
## [1] 44157.01
```

```
AIC(zp.simplified)
```

```
## [1] 43941.39
```

```
logLik(lin.mod.app)
```

```
## 'log Lik.' -22074.51 (df=4)
```

```
logLik(zp.simplified)
```

```
## 'log Lik.' -21962.69 (df=8)
```

```
#The following code is just the data preparation step for the evaluation dataset, before we apply our m
wine_eval <- wine_eval[2:length(wine_eval)]
wine_eval$STARS[is.na(wine_eval$STARS)] <- 0
wine_eval <- mice(wine_eval, m = 3, print=F)
wine_eval <- complete(wine_eval,1)
```

Our final predicted results.

```r
wine_eval$TARGET <- predict(zp.simplified, newdata = wine_eval, type = "response")
head(wine_eval, 20)
```

```
##       TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar
## 1  2.0663548          5.4          -0.860       0.27         -10.7
## 2  4.0936927         12.4           0.385      -0.76         -19.7
## 3  2.9409406          7.2           1.750       0.17         -33.0
## 4  3.4675904          6.2           0.100       1.80           1.0
## 5  1.1900288         11.4           0.210       0.28           1.2
## 6  3.8607438         17.6           0.040      -1.15           1.4
## 7  3.2218685         15.5           0.530      -0.53           4.6
## 8  0.9109515         15.9           1.190       1.14          31.9
## 9  1.0786103         11.6           0.320       0.55         -50.9
## 10 1.4487108          3.8           0.220       0.31          -7.7
## 11 3.1040063          6.8           1.680       0.44         -13.3
## 12 0.9006947          9.0          -0.210       0.04          51.4
## 13 3.5645712         24.6           0.030      -1.20           1.3
## 14 1.2948329         13.0           0.210       0.32          -3.2
## 15 1.9086669         17.9          -0.420      -0.91           7.1
## 16 3.4885346         10.0           0.200       1.27          30.9
## 17 3.4106196          7.4           0.290       0.50           8.5
## 18 0.7296859         11.7           1.180      -0.94         -62.0
## 19 3.7438359          9.7           0.410      -1.00           4.7
## 20 4.1823390         -5.2          -0.980      -0.08           6.4
##    Chlorides FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates
## 1      0.092                23                398 0.98527 5.02      0.64
## 2      1.169               -37                 68 0.99048 3.37      1.09
## 3      0.065                 9                 76 1.04641 4.61      0.68
## 4     -0.179               104                 89 0.98877 3.20      2.11
## 5      0.038                70                 53 1.02899 2.54     -0.07
## 6      0.535              -250                140 0.95028 3.06     -0.02
## 7      1.263                10                 17 1.00020 3.07      0.75
## 8     -0.299               115                381 1.03416 2.99      0.31
## 9      0.076                35                 83 1.00020 3.32      2.18
## 10     0.039                40                129 0.90610 4.72     -0.64
## 11     0.046                47                583 1.00833 3.36      1.64
## 12     0.237              -213               -527 0.99516 3.16      0.70
## 13     0.035               241                297 0.99232 2.22      0.50
## 14    -0.263               111                141 0.95918 3.20      0.38
## 15     0.045              -177                169 0.95307 3.17     -1.12
## 16     0.050                19                152 0.99400 3.36      0.42
## 17    -0.480               178                647 0.97275 3.45      0.50
## 18     0.675                 7               -393 0.99974 3.96      0.69
## 19    -0.235                24                113 0.99772 3.44      0.53
## 20     0.046               180                166 0.99400 3.30      2.18
##    Alcohol LabelAppeal AcidIndex STARS
## 1    12.30          -1         6     0
## 2    16.00           0         6     2
## 3     8.55           0         8     1
## 4    12.30          -1         8     1
## 5     4.80           0        10     0
## 6    11.40           1         8     4
## 7     8.50           0        12     3
```

```
## 8    11.40           1         7     0
## 9    -0.50           0        12     0
## 10   10.90           0         7     0
## 11   12.60           0         8     1
## 12   14.70           1        10     0
## 13    9.80           0         9     2
## 14    4.20           0         8     0
## 15   13.20          -1         9     0
## 16   13.80          -1        11     2
## 17   10.20          -1         8     1
## 18    5.20           1        13     0
## 19    9.80           0         7     2
## 20    9.90           1         5     3
```

Smooth Operators - All Done!