HW 5: Count Regression on Wine Dataset

Team 1: Michael O’Donnell

November 19, 2020

1. Data Exploration

First, the wine data training set was explored for size and structure. The training set has 12795 observations and 16 variables. 15 of the variables are predictors, and the response variable is called TARGET. The response variable is count data, ranging from 0 to 8 with a median of 3 and a large number of 0’s.

The structure of the test set is identicle to the training set, but with only 3335 observations.

To explore the training data further, a few techniques were used: - first, the summary function exposed the means, medians, and quartiles of all variables - then, the str function showed the data type of each variable - next, the distribution of TARGET was explored with a histogram. The data was bimodal, with one peak at 0 and another around 3. Typical of count data, there are many 0’s in the distribution. - last, the correlation between predictors (or appropriate data type) and TARGET were visualized.

2. Data Preparation

The training and testing data were prepared for count regression. Specifically, they were prepared for a Generalized Linear Model with a poisson distribution.

So, the incomplete records were addressed first. In the training set, the following variables had incomplete records: - ResidualSugar - Chlorides - FreeSulfurDioxide - TotalSulfurDioxide - pH - Sulphates - Alcohol - STARS

Since the STARS variable had a strong correlation with TARGET, all rows with a missing STARS variable were removed from the training set and testing est.

For the rest of the variables, each missing variable was replaced by the mean of that variable in the training set. But, for the testing set, no rows were removed for these missing variables.

3. Build Models

First, a GLM with poisson distribution and log link function was built.

The distribution was poisson because the response variable was count data. The link function was log because the high number of 0’s in the response variable.

The model was first built with all variables, it had a residual deviance of 5836.9 on 9421 degrees of freedom.

4. Select Models

To select a model, many iterations and changes were made. Overall, all variables were removed that had a p-value > 0.05 and the distribution was changed from poisson to quasipoisson. The final model is named model2 in the appendix.

This final model was used to predict TARGET values for the test dataset. The results are shown in a histogram in the appendix.

Appendix

Import Libraries and Data

```
# load required packages
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(RCurl)
library(haven)
library(xtable)
```

Load data

```
# Loading the data
git_dir <- 'https://raw.githubusercontent.com/odonnell31/DATA621-HW5/main/data'
#class_data = read.csv(paste(git_dir, "/classification-output-data.csv", sep=""))
train_df = read.csv(paste(git_dir, "/wine-training-data.csv", sep=""))
test_df = read.csv(paste(git_dir, "/wine-evaluation-data.csv", sep = ""))
head(train_df, 2)
```

```
##   ï..INDEX TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar
## 1        1      3          3.2            1.16      -0.98          54.2
## 2        2      3          4.5            0.16      -0.81          26.1
##   Chlorides FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates
## 1    -0.567                NA                268 0.99280 3.33     -0.59
## 2    -0.425                15               -327 1.02792 3.38      0.70
##   Alcohol LabelAppeal AcidIndex STARS
## 1     9.9           0         8     2
## 2      NA          -1         7     3
```

Data Exploration & Preparation

See a summary of each column in the train_df set

```
# view a summary of all columns
summary(train_df)
```

```
##     ï..INDEX         TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##
##    CitricAcid      ResidualSugar        Chlorides       FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00
##  Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00
##                    NA's   :616        NA's   :638       NA's   :647
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
##  Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##  NA's   :682                         NA's   :395     NA's   :1210
##     Alcohol       LabelAppeal          AcidIndex          STARS
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##  NA's   :653                                          NA's   :3359
```

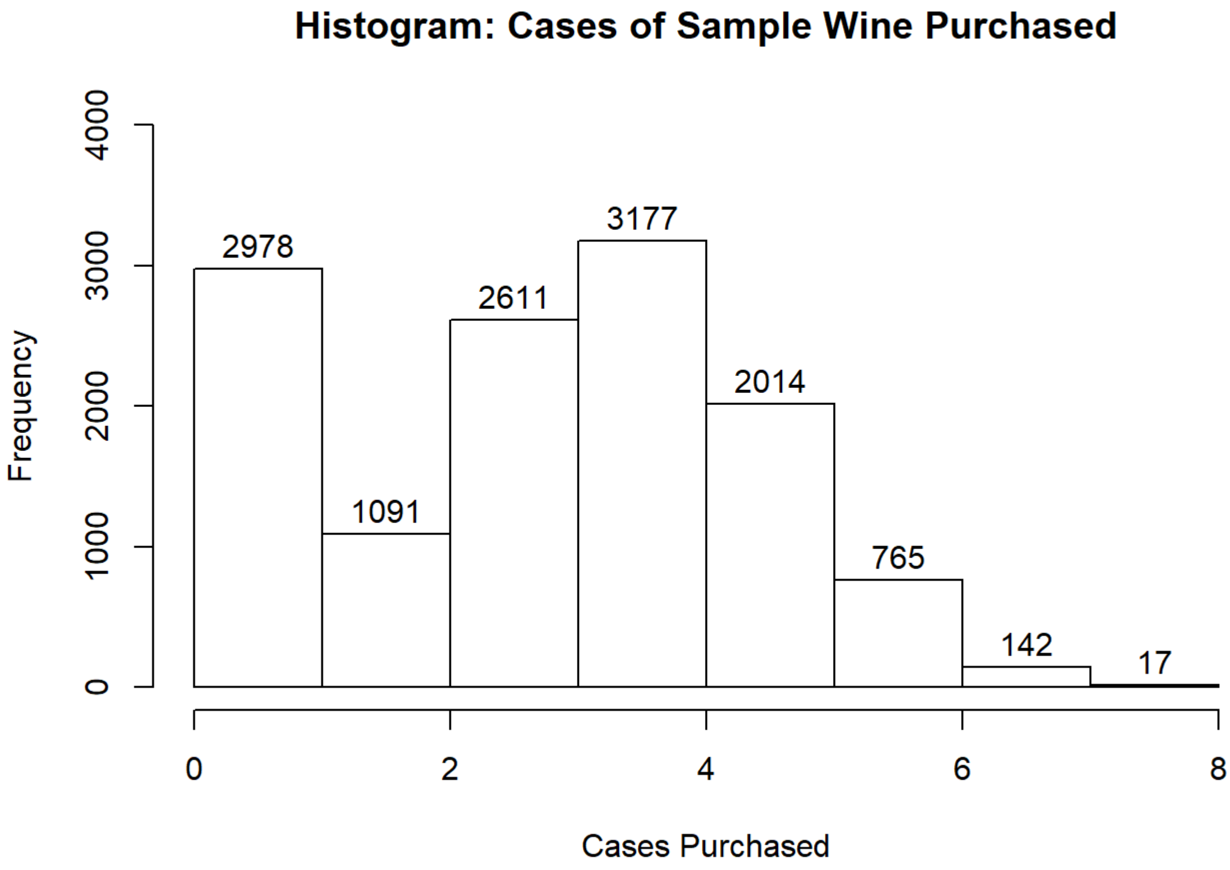
Look at the data type of each variable

```
# data type of predictors
str(train_df)
```

```
## 'data.frame':    12795 obs. of  16 variables:
##  $ ï..INDEX          : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET            : int  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       : int  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : int  2 3 3 1 2 NA NA 3 NA 4 ...
```

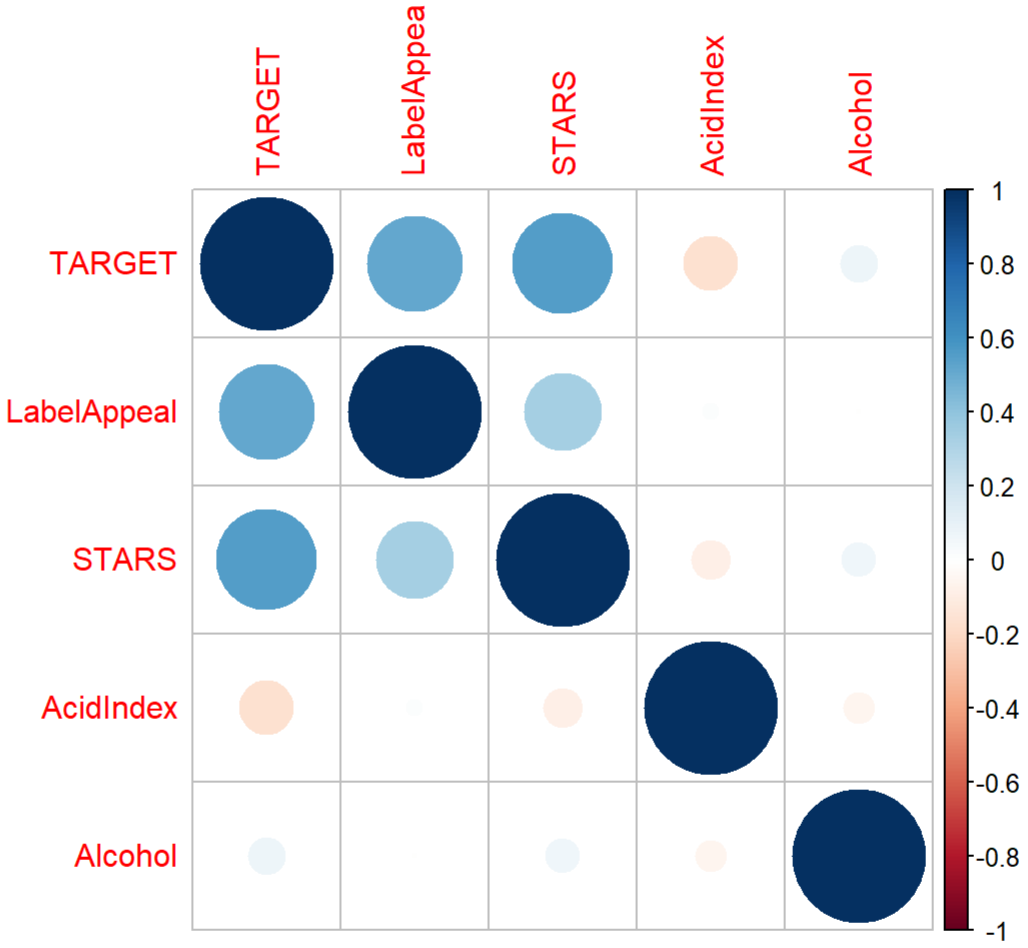
Look at the histogram of TARGET

```
h <- hist(train_df$TARGET, ylim = c(0,4000), breaks = 8,
          main="Histogram: Cases of Sample Wine Purchased",
          xlab = "Cases Purchased",
          ylab = "Frequency")
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))
```



Look at the correlation among each of the variables

```
cor_train = cor(train_df[c("TARGET", "LabelAppeal", "STARS",
                           "AcidIndex", "Alcohol")],  use = "na.or.complete")
corrplot(cor_train)
```



Check for varaibles with NA’s

```
has_NA = names(which(sapply(train_df, anyNA)))
has_NA
```

```
## [1] "ResidualSugar"      "Chlorides"          "FreeSulfurDioxide"
## [4] "TotalSulfurDioxide" "pH"                 "Sulphates"
## [7] "Alcohol"            "STARS"
```

Remove rows where STARS are missing

```
train_df <- train_df[complete.cases(train_df$STARS), ]
```

Remove incomplete rows for testing dataset

```
test_df <- test_df[complete.cases(test_df$STARS), ]
```

Replace NA’s with means for rest of variables

```
train_df$ResidualSugar[is.na(train_df$ResidualSugar)] <- mean(train_df$ResidualSugar, na.rm = T)
train_df$Chlorides[is.na(train_df$Chlorides)] <- mean(train_df$Chlorides, na.rm = T)
train_df$FreeSulfurDioxide[is.na(train_df$FreeSulfurDioxide)] <- mean(train_df$FreeSulfurDioxide, na.rm = T)
train_df$TotalSulfurDioxide[is.na(train_df$TotalSulfurDioxide)] <- mean(train_df$TotalSulfurDioxide, na.rm = T)
train_df$pH[is.na(train_df$pH)] <- mean(train_df$pH, na.rm = T)
train_df$Alcohol[is.na(train_df$Alcohol)] <- mean(train_df$Alcohol, na.rm = T)
train_df$Sulphates[is.na(train_df$Sulphates)] <- mean(train_df$Sulphates, na.rm = T)

# re-check for NA's
has_NA2 = names(which(sapply(train_df, anyNA)))
has_NA2
```

```
## character(0)
```

build poisson GLM model

```
model1 <- glm(formula = TARGET ~ FixedAcidity + LabelAppeal +
                VolatileAcidity + CitricAcid + ResidualSugar +
                Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
                Density + pH + Sulphates + Alcohol + STARS +
                AcidIndex,
              family = poisson(link = "log"),
              data = train_df)

summary(model1)
```

```
## 
## Call:
## glm(formula = TARGET ~ FixedAcidity + LabelAppeal + VolatileAcidity + 
##     CitricAcid + ResidualSugar + Chlorides + FreeSulfurDioxide + 
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol + 
##     STARS + AcidIndex, family = poisson(link = "log"), data = train_df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.2318  -0.2696   0.0649   0.3722   1.6947  
## 
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         1.486e+00  2.063e-01   7.204 5.85e-13 ***
## FixedAcidity        2.811e-04  8.653e-04   0.325 0.745277    
## LabelAppeal         1.819e-01  6.543e-03  27.805  < 2e-16 ***
## VolatileAcidity    -2.329e-02  6.899e-03  -3.376 0.000736 ***
## CitricAcid          2.984e-03  6.221e-03   0.480 0.631447    
## ResidualSugar       1.622e-08  1.637e-04   0.000 0.999921    
## Chlorides          -2.621e-02  1.741e-02  -1.505 0.132306    
## FreeSulfurDioxide   6.427e-05  3.698e-05   1.738 0.082245 .  
## TotalSulfurDioxide  2.333e-05  2.410e-05   0.968 0.332837    
## Density            -2.748e-01  2.023e-01  -1.358 0.174383    
## pH                 -2.759e-03  8.084e-03  -0.341 0.732864    
## Sulphates          -5.605e-03  6.087e-03  -0.921 0.357139    
## Alcohol             4.711e-03  1.486e-03   3.170 0.001526 ** 
## STARS               1.836e-01  6.135e-03  29.927  < 2e-16 ***
## AcidIndex          -4.832e-02  4.902e-03  -9.858  < 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: 8597.2  on 9435  degrees of freedom
## Residual deviance: 5836.9  on 9421  degrees of freedom
## AIC: 33958
## 
## Number of Fisher Scoring iterations: 5
```

refine the poisson GLM model

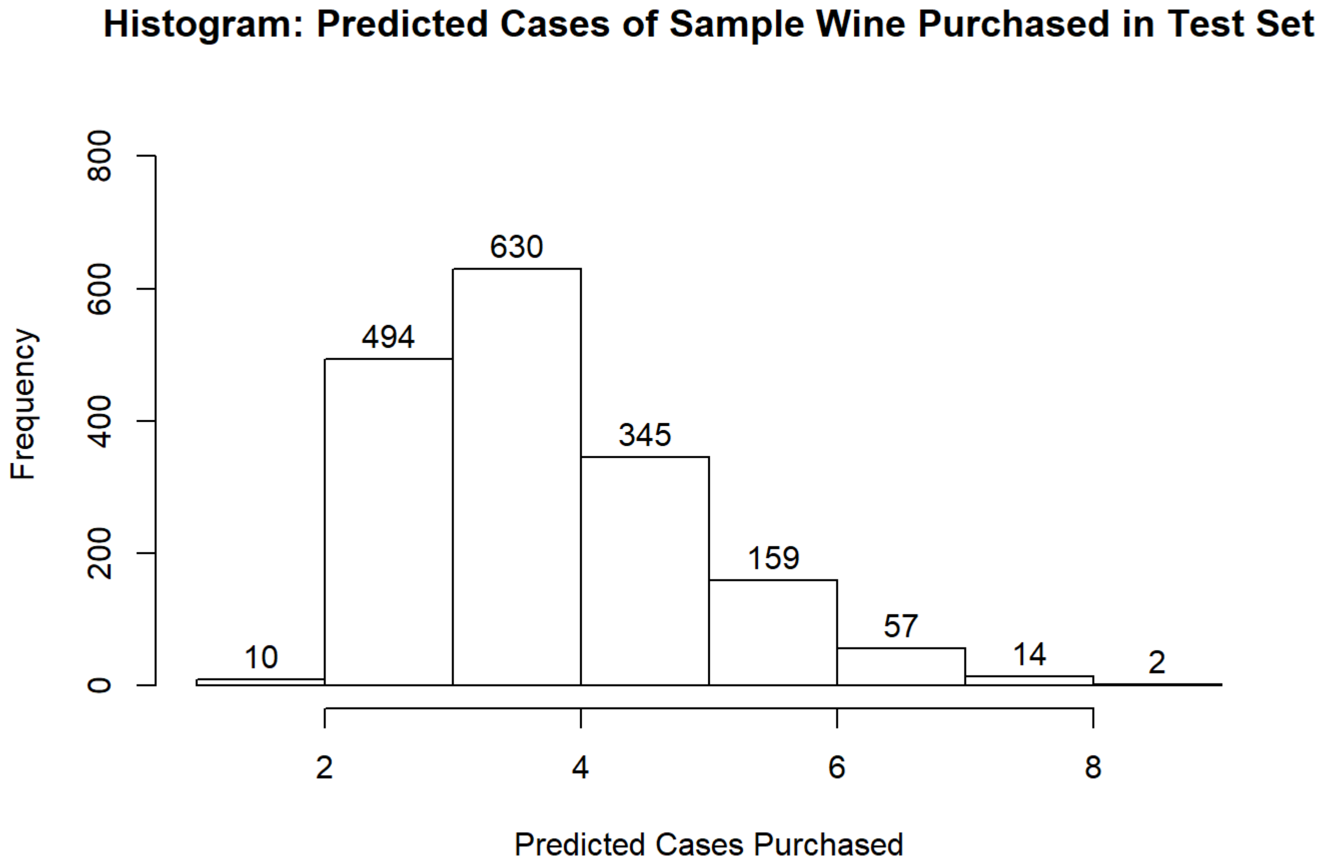
```
model2 <- glm(formula = TARGET ~ LabelAppeal + 
                VolatileAcidity + Alcohol + STARS + 
                AcidIndex,
              family = quasipoisson(link = "log"),
              data = train_df)

summary(model2)
```

```
##
## Call:
## glm(formula = TARGET ~ LabelAppeal + VolatileAcidity + Alcohol +
##     STARS + AcidIndex, family = quasipoisson(link = "log"), data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2253  -0.2704   0.0637   0.3703   1.6541
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.2096310  0.0282991  42.745  < 2e-16 ***
## LabelAppeal      0.1819200  0.0042391  42.915  < 2e-16 ***
## VolatileAcidity -0.0236753  0.0044698  -5.297 1.21e-07 ***
## Alcohol          0.0047205  0.0009625   4.905 9.52e-07 ***
## STARS            0.1837144  0.0039736  46.234  < 2e-16 ***
## AcidIndex       -0.0485056  0.0031302 -15.496  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.4201478)
##
##     Null deviance: 8597.2  on 9435  degrees of freedom
## Residual deviance: 5846.3  on 9430  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
predict <- predict(model1, newdata=test_df, type = 'response')
test_df$TARGET <- predict

h <- hist(test_df$TARGET, ylim = c(0,850), breaks = 8,
          main="Histogram: Predicted Cases of Sample Wine Purchased in Test Set",
          xlab = "Predicted Cases Purchased",
          ylab = "Frequency")
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))
```


Histogram: Predicted Cases of Sample Wine Purchased in Test Set
Frequency
0
200
400
600
800
10
494
630
345
159
57
14
2
2
4
6
8
Predicted Cases Purchased