DS621_Group2_HW5_JC

Jagdish Chhabria

5/10/2021

Group 2 members: *Diego Correa, Jagdish Chhabria, Orli Khaimova, Richard Zheng, Stephen Haslett.*

Assignment 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: VARIABLE

Task 1: Data Exploration

Describe the size and the variables in the wine training data set.

```
## '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 ...
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
names(training_set)[1]<-"INDEX"
names(training_set)
```

```
##  [1] "INDEX"              "TARGET"             "FixedAcidity"
##  [4] "VolatileAcidity"    "CitricAcid"         "ResidualSugar"
##  [7] "Chlorides"          "FreeSulfurDioxide"  "TotalSulfurDioxide"
## [10] "Density"            "pH"                 "Sulphates"
## [13] "Alcohol"            "LabelAppeal"        "AcidIndex"
## [16] "STARS"
```

```
# Remove the index variable
training_set<-training_set%>%dplyr::select(-INDEX)
#%>%mutate(TARGET=as.factor(TARGET))
#training_set<-training_set%>%dplyr::mutate(TARGET=as.factor(TARGET))
```

```
# Check the structure of the training dataset.
str(training_set)
```

```
## 'data.frame':    12795 obs. of  15 variables:
##  $ 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 ...
```

```
describe(training_set)
```

```
##                    vars     n   mean     sd median trimmed    mad     min
## TARGET                1 12795   3.03   1.93   3.00    3.05   1.48    0.00
## FixedAcidity          2 12795   7.08   6.32   6.90    7.07   3.26  -18.10
## VolatileAcidity       3 12795   0.32   0.78   0.28    0.32   0.43   -2.79
## CitricAcid            4 12795   0.31   0.86   0.31    0.31   0.42   -3.24
## ResidualSugar         5 12179   5.42  33.75   3.90    5.58  15.72 -127.80
## Chlorides             6 12157   0.05   0.32   0.05    0.05   0.13   -1.17
## FreeSulfurDioxide     7 12148  30.85 148.71  30.00   30.93  56.34 -555.00
## TotalSulfurDioxide    8 12113 120.71 231.91 123.00  120.89 134.92 -823.00
## Density               9 12795   0.99   0.03   0.99    0.99   0.01    0.89
## pH                   10 12400   3.21   0.68   3.20    3.21   0.39    0.48
## Sulphates            11 11585   0.53   0.93   0.50    0.53   0.44   -3.13
```

	x
STARS	3359
Sulphates	1210
TotalSulfurDioxide	682
Alcohol	653
FreeSulfurDioxide	647
Chlorides	638
ResidualSugar	616
pH	395
TARGET	0
FixedAcidity	0
VolatileAcidity	0
CitricAcid	0
Density	0
LabelAppeal	0
AcidIndex	0

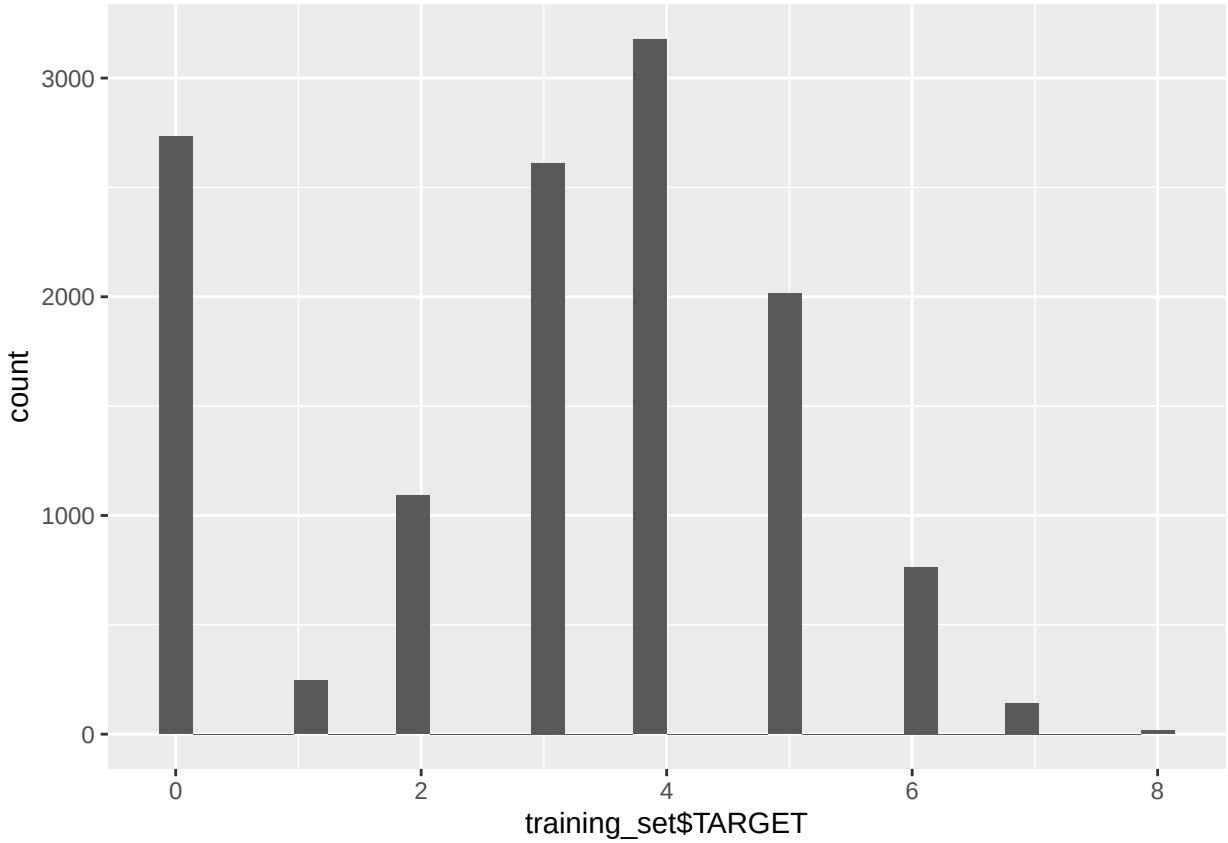
```
## Alcohol              12 12142  10.49   3.73  10.40   10.50   2.37   -4.70
## LabelAppeal          13 12795  -0.01   0.89   0.00   -0.01   1.48   -2.00
## AcidIndex            14 12795   7.77   1.32   8.00    7.64   1.48    4.00
## STARS                15  9436   2.04   0.90   2.00    1.97   1.48    1.00
##                        max   range  skew kurtosis   se
## TARGET                8.00    8.00 -0.33    -0.88 0.02
## FixedAcidity         34.40   52.50 -0.02     1.67 0.06
## VolatileAcidity       3.68    6.47  0.02     1.83 0.01
## CitricAcid            3.86    7.10 -0.05     1.84 0.01
## ResidualSugar       141.15  268.95 -0.05     1.88 0.31
## Chlorides             1.35    2.52  0.03     1.79 0.00
## FreeSulfurDioxide   623.00 1178.00  0.01     1.84 1.35
## TotalSulfurDioxide 1057.00 1880.00 -0.01     1.67 2.11
## Density               1.10    0.21 -0.02     1.90 0.00
## pH                    6.13    5.65  0.04     1.65 0.01
## Sulphates             4.24    7.37  0.01     1.75 0.01
## Alcohol              26.50   31.20 -0.03     1.54 0.03
## LabelAppeal           2.00    4.00  0.01    -0.26 0.01
## AcidIndex            17.00   13.00  1.65     5.19 0.01
## STARS                 4.00    3.00  0.45    -0.69 0.01
```

Missing Values

```
# Check for missing values
sapply(training_set, function(x) sum(is.na(x))) %>% sort(decreasing = TRUE) %>% kable() %>% kable_stylin
```

```
ggplot(training_set,aes(x=training_set$TARGET))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Check count and proportion of 0 values for the TARGET variable
training_set%>%filter(TARGET==0)%>%summarise(n=n())%>%mutate(freq=round(n/nrow(training_set),4))
```

```
##      n   freq
## 1 2734 0.2137
```

From the above, we can see that about 21% of the records have a count = 0. Given that more than a fifth of the target variable values are 0, this could be considered as a "zero-inflated" dataset.

```
# Fit the Poisson model for the count variable, with all predictors included
summary(model1 <- glm(TARGET~., family="poisson", data=training_set))
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = "poisson", data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2158  -0.2734   0.0616   0.3732   1.6830
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.593e+00  2.506e-01   6.359 2.03e-10 ***
## FixedAcidity         3.293e-04  1.053e-03   0.313  0.75447
## VolatileAcidity     -2.560e-02  8.353e-03  -3.065  0.00218 **
```

```
## CitricAcid         -7.259e-04  7.575e-03  -0.096  0.92365
## ResidualSugar      -6.141e-05  1.941e-04  -0.316  0.75165
## Chlorides          -3.007e-02  2.056e-02  -1.463  0.14346
## FreeSulfurDioxide   6.734e-05  4.404e-05   1.529  0.12620
## TotalSulfurDioxide  2.081e-05  2.855e-05   0.729  0.46618
## Density            -3.725e-01  2.462e-01  -1.513  0.13026
## pH                 -4.661e-03  9.598e-03  -0.486  0.62722
## Sulphates          -5.164e-03  7.051e-03  -0.732  0.46398
## Alcohol             3.948e-03  1.771e-03   2.229  0.02579 *
## LabelAppeal         1.771e-01  7.954e-03  22.271  < 2e-16 ***
## AcidIndex          -4.870e-02  5.903e-03  -8.251  < 2e-16 ***
## STARS               1.871e-01  7.487e-03  24.993  < 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: 5844.1  on 6435  degrees of freedom
## Residual deviance: 4009.1  on 6421  degrees of freedom
##   (6359 observations deleted due to missingness)
## AIC: 23172
##
## Number of Fisher Scoring iterations: 5
```

Based on the p-values above, it looks like the following predictors have a significant impact on the number of wine cases ordered: VolatileAcidity, Alcohol, LabelAppeal, AcidIndex, STARS. All the co-efficients are very small though.

```
# Check for dispersion with this model
dispersiontest(model1,trafo=1)
```

```
##
##  Overdispersion test
##
## data:  model1
## z = -44.739, p-value = 1
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##     alpha
## -0.573328
```

From the above results, it seems that there is underdispersion in the data, since $c < 0$.

```
# Fit the Negative Binomial model for the count variable, with all predictors included
summary(model2<-glm.nb(TARGET~.,data=training_set))
```

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = training_set, init.theta = 140198.134,
##     link = log)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.2157  -0.2733   0.0616   0.3732   1.6830
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.593e+00  2.506e-01   6.359 2.03e-10 ***
## FixedAcidity         3.293e-04  1.053e-03   0.313  0.75446
## VolatileAcidity     -2.560e-02  8.353e-03  -3.065  0.00218 **
## CitricAcid          -7.259e-04  7.575e-03  -0.096  0.92365
## ResidualSugar       -6.141e-05  1.941e-04  -0.316  0.75166
## Chlorides           -3.007e-02  2.056e-02  -1.463  0.14347
## FreeSulfurDioxide    6.734e-05  4.404e-05   1.529  0.12621
## TotalSulfurDioxide   2.081e-05  2.855e-05   0.729  0.46618
## Density             -3.725e-01  2.462e-01  -1.513  0.13026
## pH                  -4.661e-03  9.598e-03  -0.486  0.62722
## Sulphates           -5.164e-03  7.052e-03  -0.732  0.46398
## Alcohol              3.948e-03  1.771e-03   2.229  0.02579 *
## LabelAppeal          1.771e-01  7.954e-03  22.271  < 2e-16 ***
## AcidIndex           -4.870e-02  5.903e-03  -8.251  < 2e-16 ***
## STARS                1.871e-01  7.487e-03  24.992  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(140198.1) family taken to be 1)
##
##     Null deviance: 5843.9  on 6435  degrees of freedom
## Residual deviance: 4009.1  on 6421  degrees of freedom
##   (6359 observations deleted due to missingness)
## AIC: 23174
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  140198
##           Std. Err.:  234984
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -23141.85
```

```
# Check for overdispersion with this model
odTest(model2)
```

```
## Likelihood ratio test of H0: Poisson, as restricted NB model:
## n.b., the distribution of the test-statistic under H0 is non-standard
## e.g., see help(odTest) for details/references
##
## Critical value of test statistic at the alpha= 0.05 level: 2.7055
## Chi-Square Test Statistic =  -0.1055 p-value = 0.5
```

Based on the above test statistic value, we fail to reject the Null Hypothesis which states that the Poisson model is better suited for this dataset. So we stick with the Poisson model instead of the Negative Binomial model.

```r
# Fit a zero-inflation poisson model with all predictors included
summary(model3<-zeroinfl(TARGET~.,data=training_set,dist="poisson"))
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ ., data = training_set, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.30606 -0.29065  0.03564  0.34185  3.92873
##
## Count model coefficients (poisson with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.426e+00  2.563e-01   5.563 2.66e-08 ***
## FixedAcidity         1.428e-04  1.078e-03   0.132 0.894616
## VolatileAcidity     -1.190e-02  8.507e-03  -1.399 0.161826
## CitricAcid          -1.957e-03  7.721e-03  -0.253 0.799891
## ResidualSugar       -1.680e-04  1.963e-04  -0.856 0.392041
## Chlorides           -2.617e-02  2.096e-02  -1.249 0.211808
## FreeSulfurDioxide    2.461e-05  4.456e-05   0.552 0.580819
## TotalSulfurDioxide -1.950e-05  2.832e-05  -0.689 0.491086
## Density             -2.982e-01  2.520e-01  -1.183 0.236631
## pH                   6.278e-03  9.795e-03   0.641 0.521553
## Sulphates            1.210e-03  7.181e-03   0.169 0.866189
## Alcohol              6.279e-03  1.804e-03   3.481 0.000499 ***
## LabelAppeal          2.107e-01  8.161e-03  25.813  < 2e-16 ***
## AcidIndex           -1.673e-02  6.203e-03  -2.697 0.006987 **
## STARS                1.128e-01  7.913e-03  14.258  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -6.3822133  2.8198499  -2.263  0.02362 *
## FixedAcidity       -0.0103170  0.0118808  -0.868  0.38519
## VolatileAcidity     0.2660691  0.0926319   2.872  0.00407 **
## CitricAcid          0.0001215  0.0841838   0.001  0.99885
## ResidualSugar      -0.0037930  0.0021582  -1.757  0.07884 .
## Chlorides           0.1573073  0.2433982   0.646  0.51809
## FreeSulfurDioxide  -0.0010635  0.0005144  -2.068  0.03867 *
## TotalSulfurDioxide -0.0010277  0.0003173  -3.239  0.00120 **
## Density             3.2382780  2.7460326   1.179  0.23830
## pH                  0.2470412  0.1102338   2.241  0.02502 *
## Sulphates           0.1858210  0.0800655   2.321  0.02029 *
## Alcohol             0.0451117  0.0201085   2.243  0.02487 *
## LabelAppeal         0.7542524  0.0929042   8.119 4.72e-16 ***
## AcidIndex           0.5101960  0.0514554   9.915  < 2e-16 ***
## STARS              -3.8028969  0.3683015 -10.325  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 41
## Log-likelihood: -1.113e+04 on 30 Df
```

```r
# Fit the Poisson model for the count variable, with selected predictors only
summary(model4<-glm(TARGET~VolatileAcidity+Alcohol+LabelAppeal+AcidIndex+STARS, family="poisson", data=
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Alcohol + LabelAppeal +
##     AcidIndex + STARS, family = "poisson", data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2222  -0.2698   0.0641   0.3712   1.6551
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.214470   0.044713  27.162  < 2e-16 ***
## VolatileAcidity -0.024054   0.007083  -3.396 0.000683 ***
## Alcohol           0.004689   0.001486   3.156 0.001601 **
## LabelAppeal       0.181146   0.006713  26.984  < 2e-16 ***
## AcidIndex        -0.049533   0.004962  -9.982  < 2e-16 ***
## STARS             0.184905   0.006308  29.311  < 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: 8176.2  on 8962  degrees of freedom
## Residual deviance: 5564.4  on 8957  degrees of freedom
##   (3832 observations deleted due to missingness)
## AIC: 32245
##
## Number of Fisher Scoring iterations: 5
```

```r
# Fit a zero-inflation poisson model with selected predictors only

summary(model5<-zeroinfl(TARGET~VolatileAcidity+Alcohol+LabelAppeal+AcidIndex+STARS,data=training_set,di
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ VolatileAcidity + Alcohol + LabelAppeal +
##     AcidIndex + STARS, data = training_set, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.31606 -0.29156  0.04495  0.34676  5.45180
##
## Count model coefficients (poisson with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.152739   0.046283  24.906  < 2e-16 ***
## VolatileAcidity -0.010646   0.007214  -1.476 0.139985
## Alcohol           0.006699   0.001509   4.440 8.99e-06 ***
## LabelAppeal       0.212277   0.006872  30.891  < 2e-16 ***
## AcidIndex        -0.018016   0.005221  -3.451 0.000559 ***
## STARS             0.112437   0.006660  16.881  < 2e-16 ***
```

```
##
## Zero-inflation model coefficients (binomial with logit link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.34945    0.52589  -4.468 7.91e-06 ***
## VolatileAcidity  0.29586    0.07811   3.788 0.000152 ***
## Alcohol          0.03994    0.01601   2.495 0.012596 *
## LabelAppeal      0.71368    0.07866   9.073  < 2e-16 ***
## AcidIndex        0.50622    0.04266  11.866  < 2e-16 ***
## STARS           -3.83863    0.34936 -10.988  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 20
## Log-likelihood: -1.555e+04 on 12 Df
```

```
# Perform a vuong test to compare model 4 and model 5
vuong(model4, model5)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                  -14.67892 model2 > model1 < 2.22e-16
## AIC-corrected        -14.52437 model2 > model1 < 2.22e-16
## BIC-corrected        -13.97564 model2 > model1 < 2.22e-16
```

The Vuong test compares the zero-inflated model (model 5) with the ordinary Poisson regression model (model 4). In this case, we can see that our test statistic is significant, indicating that the zero-inflated model is superior to the standard Poisson model.

```
# Predict the target count for the evaluation dataset
evaluation_set$predicted_target<-predict(model5,type = 'response',newdata = evaluation_set)
```

```
str(evaluation_set)
```

```
## 'data.frame':    3335 obs. of  17 variables:
##  $ IN                : int  3 9 10 18 21 30 31 37 39 47 ...
##  $ TARGET            : logi  NA NA NA NA NA NA ...
##  $ FixedAcidity      : num  5.4 12.4 7.2 6.2 11.4 17.6 15.5 15.9 11.6 3.8 ...
##  $ VolatileAcidity   : num  -0.86 0.385 1.75 0.1 0.21 0.04 0.53 1.19 0.32 0.22 ...
##  $ CitricAcid        : num  0.27 -0.76 0.17 1.8 0.28 -1.15 -0.53 1.14 0.55 0.31 ...
##  $ ResidualSugar     : num  -10.7 -19.7 -33 1 1.2 1.4 4.6 31.9 -50.9 -7.7 ...
##  $ Chlorides         : num  0.092 1.169 0.065 -0.179 0.038 ...
##  $ FreeSulfurDioxide : num  23 -37 9 104 70 -250 10 115 35 40 ...
##  $ TotalSulfurDioxide: num  398 68 76 89 53 140 17 381 83 129 ...
##  $ Density           : num  0.985 0.99 1.046 0.989 1.029 ...
##  $ pH                : num  5.02 3.37 4.61 3.2 2.54 3.06 3.07 2.99 3.32 4.72 ...
##  $ Sulphates         : num  0.64 1.09 0.68 2.11 -0.07 -0.02 0.75 0.31 2.18 -0.64 ...
##  $ Alcohol           : num  12.3 16 8.55 12.3 4.8 11.4 8.5 11.4 -0.5 10.9 ...
##  $ LabelAppeal       : int  -1 0 0 -1 0 1 0 1 0 0 ...
##  $ AcidIndex         : int  6 6 8 8 10 8 12 7 12 7 ...
##  $ STARS             : int  NA 2 1 1 NA 4 3 NA NA NA ...
##  $ predicted_target  : num  NA 3.94 2.49 2.45 NA ...
```