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.

Data Exploration

Data Exploration

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
4 :3177	Min. :-18.100	Min. :-2.7900	Min. :-3.2400	Min. :-127.800	Min. :-1.1710
0 :2734	1st Qu.: 5.200	1st Qu.: 0.1300	1st Qu.: 0.0300	1st Qu.: -2.000	1st Qu.:-0.0310
3 :2611	Median : 6.900	Median : 0.2800	Median : 0.3100	Median : 3.900	Median : 0.0460
5 :2014	Mean : 7.076	Mean : 0.3241	Mean : 0.3084	Mean : 5.419	Mean : 0.0548
2 :1091	3rd Qu.: 9.500	3rd Qu.: 0.6400	3rd Qu.: 0.5800	3rd Qu.: 15.900	3rd Qu.: 0.1530
6 : 765	Max. : 34.400	Max. : 3.6800	Max. : 3.8600	Max. : 141.150	Max. : 1.3510
(Other): 403	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

```
#ggplot(wine, aes(TARGET, fill = STARS)) + geom_histogram(binwidth = 1, stat = "count") + facet_grid(ST
#    ., margins = TRUE, scales = "free")
```

EDIT TO BE REMOVED: In the below examples, I took Rob's runs and extrapolated some Negative Binomial Regression to see if it greatly differed. From What I saw, it didn't, though I didnt compare the chi-sqaure test yet.

```
##
## Call:
## glm(formula = as.formula(paste(colnames(train)[1], "~", paste(colnames(train)[-1],
##     collapse = "+"), sep = "")), family = poisson, data = train)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9945  -0.7089   0.0591   0.5775   3.2498
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.717e+00  2.664e-01   6.448 1.14e-10 ***
## FixedAcidity        -8.333e-05  1.121e-03  -0.074 0.940734
## VolatileAcidity     -3.463e-02  8.921e-03  -3.882 0.000104 ***
## CitricAcid           1.884e-03  8.051e-03   0.234 0.814987
## ResidualSugar       -8.061e-05  2.051e-04  -0.393 0.694357
## Chlorides           -5.673e-02  2.164e-02  -2.621 0.008773 **
## FreeSulfurDioxide    1.270e-04  4.673e-05   2.718 0.006562 **
## TotalSulfurDioxide   8.448e-05  3.017e-05   2.800 0.005108 **
## Density             -4.789e-01  2.620e-01  -1.828 0.067541 .
## pH                  -1.571e-02  1.022e-02  -1.536 0.124484
## Sulphates           -9.854e-03  7.499e-03  -1.314 0.188818
## Alcohol              1.396e-03  1.895e-03   0.737 0.461356
## LabelAppeal          1.340e-01  8.299e-03  16.143  < 2e-16 ***
## AcidIndex           -8.550e-02  6.199e-03 -13.794  < 2e-16 ***
## STARS                3.128e-01  6.230e-03  50.213  < 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: 12143.9  on 6893  degrees of freedom
## Residual deviance:  7882.5  on 6879  degrees of freedom
##   (3342 observations deleted due to missingness)
## AIC: 25165
##
## Number of Fisher Scoring iterations: 5
```

```
pm2 <- glm(TARGET ~ STARS + LabelAppeal,data = train,family = poisson)
summary(pm2)
```

```
##
## Call:
## glm(formula = TARGET ~ STARS + LabelAppeal, family = poisson,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8887  -0.7644   0.0787   0.6151   3.2902
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.514677   0.011250   45.75   <2e-16 ***
## STARS       0.331690   0.004969   66.75   <2e-16 ***
## LabelAppeal 0.125219   0.006773   18.49   <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: 18225  on 10235  degrees of freedom
## Residual deviance: 12107  on 10233  degrees of freedom
## AIC: 37677
##
## Number of Fisher Scoring iterations: 5
```

```
pm3 <- glm.nb(TARGET ~ STARS + LabelAppeal,data = train)
```

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

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

```
summary(pm3)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ STARS + LabelAppeal, data = train,
##     init.theta = 48547.83031, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8887  -0.7644   0.0787   0.6151   3.2901
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.514670   0.011250   45.75   <2e-16 ***
## STARS       0.331694   0.004969   66.75   <2e-16 ***
## LabelAppeal 0.125219   0.006773   18.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48547.83) family taken to be 1)
##
##     Null deviance: 18224  on 10235  degrees of freedom
## Residual deviance: 12106  on 10233  degrees of freedom
## AIC: 37679
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48548
##           Std. Err.:  56938
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -37671.1
```

```
pm4 <- glm.nb(TARGET ~ ., data = train)
```

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

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

```
summary(pm4)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = train, init.theta = 49150.47669,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9944  -0.7089   0.0591   0.5774   3.2497
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.717e+00  2.664e-01   6.447 1.14e-10 ***
## FixedAcidity        -8.331e-05  1.121e-03  -0.074 0.940748
## VolatileAcidity     -3.463e-02  8.922e-03  -3.882 0.000104 ***
## CitricAcid           1.884e-03  8.051e-03   0.234 0.814966
## ResidualSugar       -8.060e-05  2.051e-04  -0.393 0.694380
## Chlorides           -5.673e-02  2.165e-02  -2.621 0.008774 **
## FreeSulfurDioxide    1.270e-04  4.673e-05   2.718 0.006563 **
## TotalSulfurDioxide   8.448e-05  3.017e-05   2.800 0.005108 **
## Density             -4.789e-01  2.620e-01  -1.828 0.067548 .
## pH                  -1.571e-02  1.022e-02  -1.536 0.124488
## Sulphates           -9.855e-03  7.499e-03  -1.314 0.188810
## Alcohol              1.396e-03  1.895e-03   0.737 0.461392
## LabelAppeal          1.340e-01  8.300e-03  16.143  < 2e-16 ***
## AcidIndex           -8.551e-02  6.199e-03 -13.794  < 2e-16 ***
## STARS                3.129e-01  6.231e-03  50.212  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(49150.48) family taken to be 1)
##
##     Null deviance: 12143.4  on 6893  degrees of freedom
## Residual deviance:  7882.2  on 6879  degrees of freedom
##   (3342 observations deleted due to missingness)
## AIC: 25167
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  49150
##           Std. Err.:  69155
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -25135.19
```

```r
pchisq(2 * (logLik(pm) - logLik(pm4)), df = 1, lower.tail = FALSE)
```

```
## 'log Lik.' 0.6988478 (df=15)
```