Take Home Final Project

Nikhil Agarwal

6/22/2021

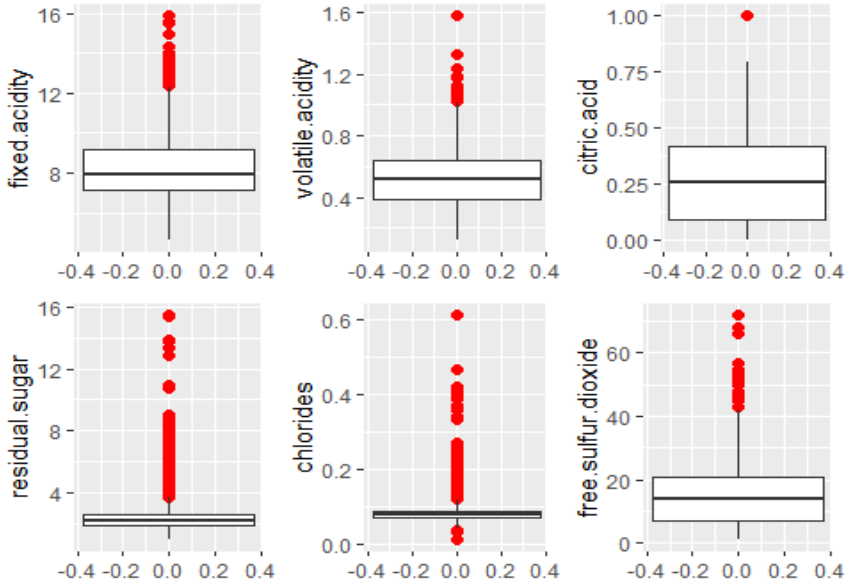
Contents

Part A. Background and data exploration

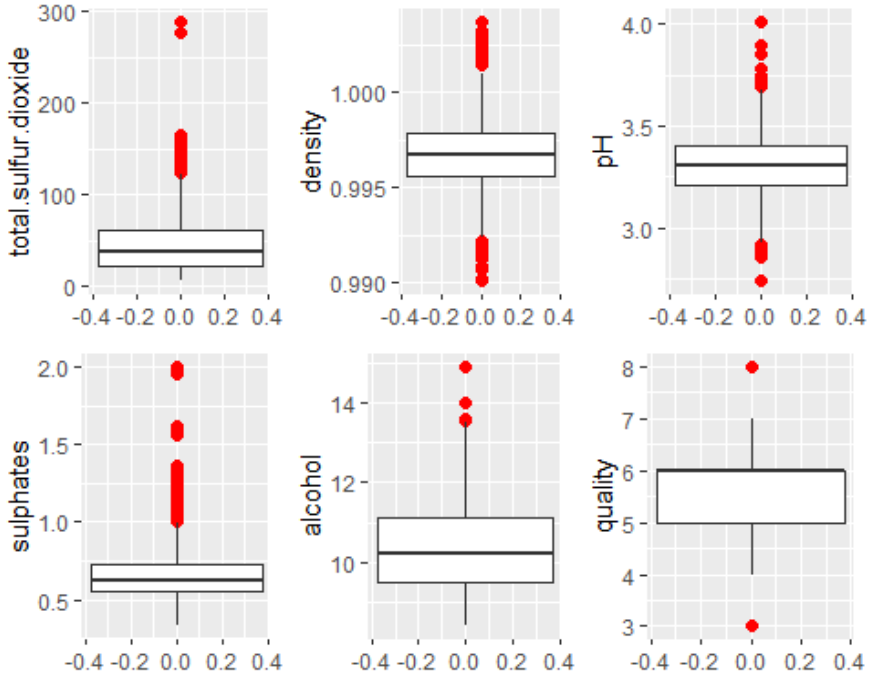
I have imported the red wine quality dataset and checked its structure and summary statistics. I have also plotted histogram to identify the outliers (highlighted by the red points). Please find the codes below

```
# Importing Dataset
wine <- read.csv("C:/UC/Stat Modelling/Final Project/winequality-red.csv",
                 sep = ";" )

# Check NA values and structure of the data: No NA values
sum(is.na(wine))

## [1] 0

str(wine)

## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.
5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.
065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3
.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57
0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...

summary(wine)

##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
```

```
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000

# Outlier Detection (please refer to the codes in the appendix section)
h1 <- grid.arrange(out1, out2, out3, out4, out5, out6, nrow = 2, ncol = 3)
```




```
h2 <- grid.arrange(out7, out8, out9, out10, out11, out12, nrow = 2, ncol = 3)
```


total.sulfur.dioxide
density
pH
sulphates
alcohol
quality

 # Part B: Visualization and initial models for a binary response

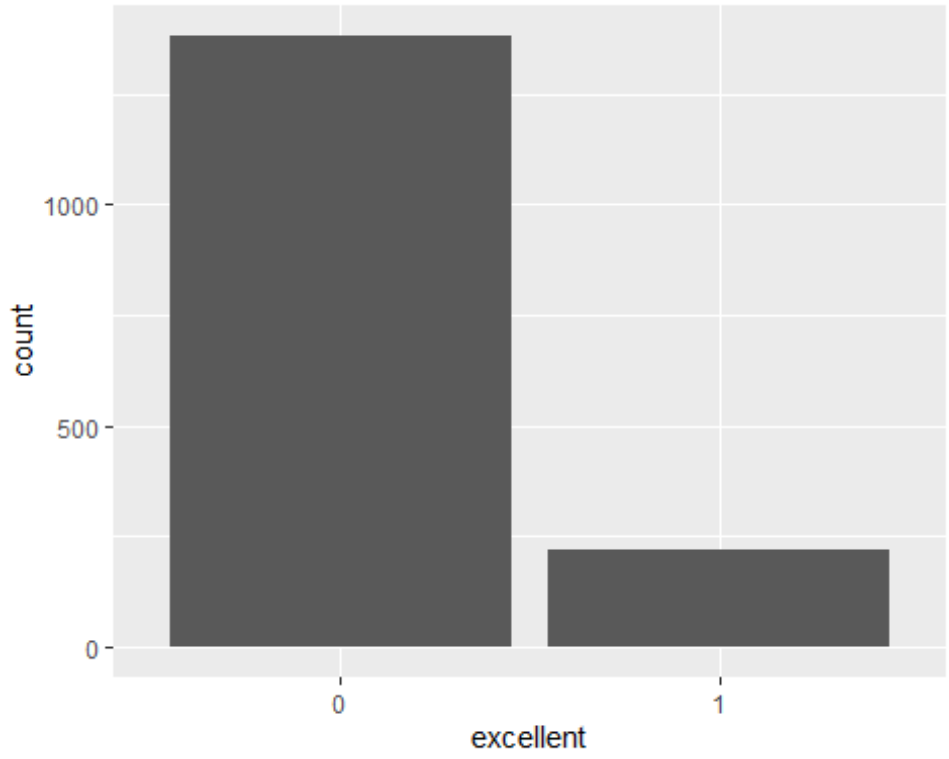
 # Q1 PartB: Create new variable excellent

```
wine$excellent <- ifelse(wine$quality >= 7, 1, 0)
wine$excellent <- as.factor(wine$excellent)
library(dplyr)
```

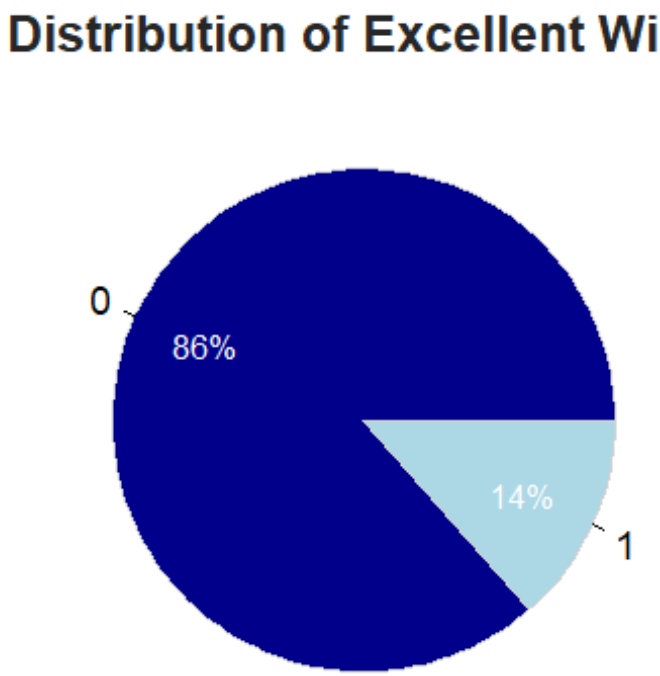
 # Q2 PartB: Proportion of the Excellent Red Wine

There are 1,382 red wines which are not good and 217 red wines which are excellent. Please find the distribution and pie chart below:

```
w <- wine$excellent
summary(w)

##    0    1
## 1382  217

ggplot(wine, aes(x = excellent)) + geom_bar()
```



```
pie2 <- data.frame(excellent = wine$excellent)
library("lessR")
PieChart(excellent, hole = 0, values = "%", data = wine, fill = c("darkblue",
"lightblue"), main = "Distribution of Excellent Wine")
```

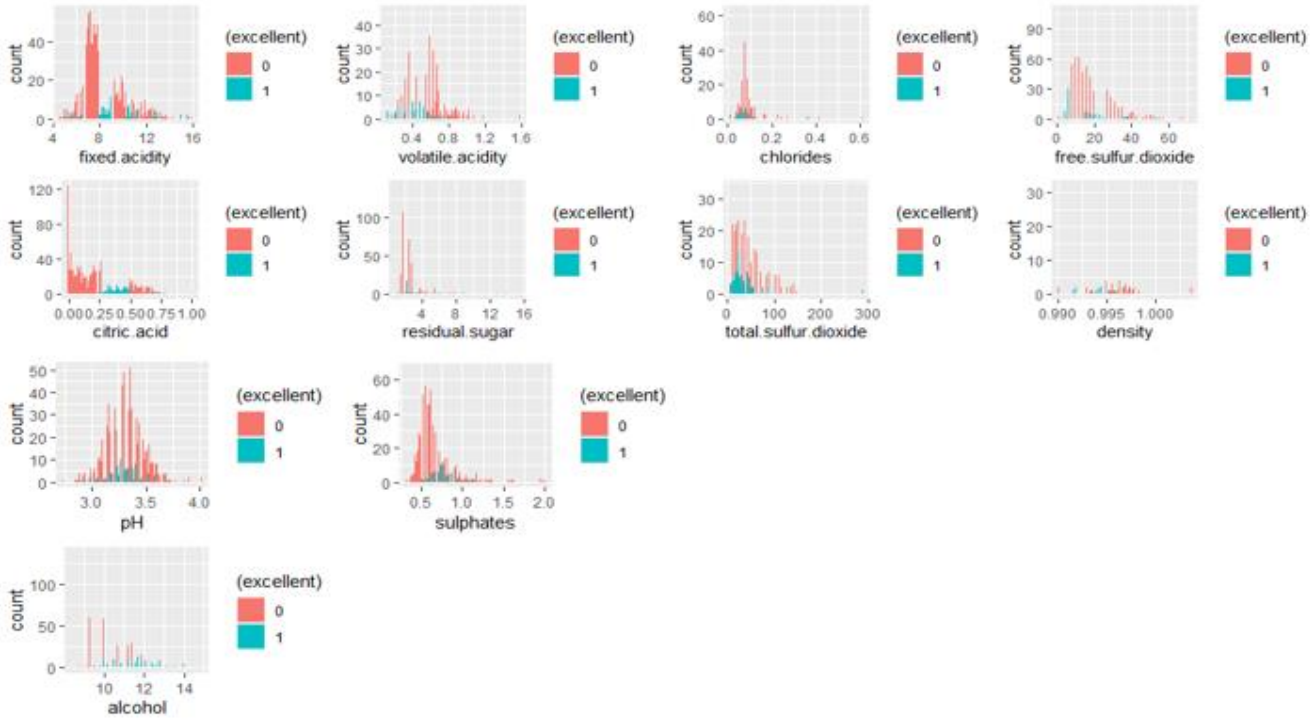


 # Q3 PartB: Visualization to Inspect Association between binary response and explanatory variables

Please refer to the codes in the appendix:

I observe that the alcohol has a slightly direct positive correlation with the quality response. Also Volatile acidity has a slightly negative association with the response

	quality
fixed.acidity	0.12405165
volatile.acidity	-0.39055778
citric.acid	0.22637251
residual.sugar	0.01373164
chlorides	-0.12890656
free.sulfur.dioxide	-0.05065606
total.sulfur.dioxide	-0.18510029
density	-0.17491923
pH	-0.05773139
sulphates	0.25139708
alcohol	0.47616632
quality	1



 # Q4 Part B Fit a linear model on excellent

There is collinearity between some pairs, such as citric acid and fixed acidity, pH and fixed acidity, and fr ee sulfur dioxide and total sulfur dioxide.

Upon fixing the linear regression model, I found that variables fixed.acidity, volatile.acidity, residual.sug ar, chlorides, total.sulfur.dioxide, density, sulphates, and alcohol were significant determined by their p-values (< 0.05). Please check the tables below. I have also used step function to find the best fit model, p lease find the results below

	Estimate	Pr(>|t|)	Association with Quality
fixed.acidity	0.0340033	0.00462	Positive
volatile.acidity	-0.1782816	0.001468	Negative
citric.acid	0.0868324	0.201774	Positive
residual.sugar	0.0251971	0.000286	Positive
chlorides	-0.6556214	0.00073	Negative
total.sulfur.dioxide	-0.0006658	0.048159	Negative
density	-36.677326	0.000251	Negative
sulphates	0.3515431	3.89E-11	Positive
alcohol	0.0761776	6.111E-10	Positive

Correlation Table (1/2)

	fixed.acidity vo	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide
fixed.acidity	1.00	-0.26	0.67	0.11	0.09	-0.15
volatile.acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01
citric.acid	0.67	-0.55	1.00	0.14	0.20	-0.06
residual.sugar	0.11	0.00	0.14	1.00	0.06	0.19
chlorides	0.09	0.06	0.20	0.06	1.00	0.01
free.sulfur.dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00
total.sulfur.dioxide	-0.11	0.08	0.04	0.20	0.05	0.67
density	0.67	0.02	0.36	0.36	0.20	-0.02
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07
quality	0.12	-0.39	0.23	0.01	-0.13	-0.05

Correlation Table (2/2)

	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
fixed.acidity	-0.11	0.67	-0.68	0.18	-0.06	0.12
volatile.acidity	0.08	0.02	0.23	-0.26	-0.20	-0.39
citric.acid	0.04	0.36	-0.54	0.31	0.11	0.23
residual.sugar	0.20	0.36	-0.09	0.01	0.04	0.01
chlorides	0.05	0.20	-0.27	0.37	-0.22	-0.13

free.sulfur.dioxide	0.67	-0.02	0.07	0.05	-0.07	-0.05
total.sulfur.dioxide	1.00	0.07	-0.07	0.04	-0.21	-0.19
density	0.07	1.00	-0.34	0.15	-0.50	-0.17
pH	-0.07	-0.34	1.00	-0.20	0.21	-0.06
sulphates	0.04	0.15	-0.20	1.00	0.09	0.25
alcohol	-0.21	-0.50	0.21	0.09	1.00	0.48
quality	-0.19	-0.17	-0.06	0.25	0.48	1.00

```
# fitting linear models
wine_lm <- wine
wine_lm$excellent <- as.numeric(wine_lm$excellent)
wine_lm <- wine_lm[-12]

lm_full <- lm(excellent ~ . , data = wine_lm)
summary(lm_full)

##
## Call:
## lm(formula = excellent ~ ., data = wine_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92396 -0.17446 -0.04220  0.05006  0.99838
##
## Coefficients:
##                         Estimate  Std. Error t value        Pr(>|t|)
## (Intercept)           36.4319904   9.7918285   3.721        0.000206
## fixed.acidity          0.0340033   0.0119881   2.836        0.004620
## volatile.acidity      -0.1782816   0.0559484  -3.187        0.001468
## citric.acid            0.0868324   0.0679949   1.277        0.201774
## residual.sugar         0.0251971   0.0069309   3.635        0.000286
## chlorides             -0.6556214   0.1937076  -3.385        0.000730
## free.sulfur.dioxide   -0.0005526   0.0010031  -0.551        0.581817
## total.sulfur.dioxide  -0.0006658   0.0003367  -1.977        0.048159
## density              -36.6773260   9.9944256  -3.670        0.000251
## pH                     0.0172498   0.0885174   0.195        0.845516
## sulphates              0.3515431   0.0528236   6.655 0.0000000000389
## alcohol                0.0761776   0.0122353   6.226 0.0000000006111
##
## Residual standard error: 0.2994 on 1587 degrees of freedom
## Multiple R-squared:  0.2416, Adjusted R-squared:  0.2363
## F-statistic: 45.96 on 11 and 1587 DF,  p-value: < 0.00000000000000022

lm_zero <- lm(excellent ~ 1, data = wine_lm)

lm_step <- step(lm_zero, scope = list(lower = lm_zero, upper = lm_full),
                direction = "both", k = log(nrow(wine_lm)))
```

```
## Start:  AIC=-3419.41
## excellent ~ 1
##
##                         Df Sum of Sq    RSS     AIC
## + alcohol                1   31.1157 156.44 -3702.1
## + volatile.acidity       1   13.7446 173.81 -3533.7
## + citric.acid            1    8.6466 178.90 -3487.5
## + sulphates              1    7.4635 180.09 -3477.0
## + density                1    4.2458 183.31 -3448.6
## + total.sulfur.dioxide   1    3.6507 183.90 -3443.5
## + fixed.acidity          1    2.7035 184.85 -3435.3
## + chlorides              1    1.7759 185.78 -3427.2
## + free.sulfur.dioxide    1    0.9655 186.59 -3420.3
## <none>                               187.55 -3419.4
## + pH                     1    0.6154 186.94 -3417.3
## + residual.sugar         1    0.4281 187.12 -3415.7
##
## Step:  AIC=-3702.11
## excellent ~ alcohol
##
##                         Df Sum of Sq    RSS     AIC
## + volatile.acidity       1    6.9349 149.50 -3767.2
## + citric.acid            1    5.4833 150.95 -3751.8
## + sulphates              1    4.9266 151.51 -3745.9
## + fixed.acidity          1    3.9681 152.47 -3735.8
## + pH                     1    3.8956 152.54 -3735.1
## <none>                               156.44 -3702.1
## + density                1    0.6635 155.77 -3701.5
## + total.sulfur.dioxide   1    0.6087 155.83 -3701.0
## + free.sulfur.dioxide    1    0.3562 156.08 -3698.4
## + residual.sugar         1    0.1764 156.26 -3696.5
## + chlorides              1    0.0103 156.43 -3694.8
## - alcohol                1   31.1157 187.55 -3419.4
##
## Step:  AIC=-3767.23
## excellent ~ alcohol + volatile.acidity
##
##                         Df Sum of Sq    RSS     AIC
## + sulphates              1    2.6129 146.89 -3788.1
## + fixed.acidity          1    1.7401 147.76 -3778.6
## + pH                     1    1.6073 147.89 -3777.1
## + citric.acid            1    1.1700 148.33 -3772.4
## <none>                               149.50 -3767.2
## + total.sulfur.dioxide   1    0.4690 149.03 -3764.9
## + free.sulfur.dioxide    1    0.4399 149.06 -3764.6
## + density                1    0.3299 149.17 -3763.4
## + residual.sugar         1    0.2008 149.30 -3762.0
## + chlorides              1    0.0031 149.50 -3759.9
## - volatile.acidity       1    6.9349 156.44 -3702.1
## - alcohol                1   24.3060 173.81 -3533.7
```

```
##
## Step:  AIC=-3788.05
## excellent ~ alcohol + volatile.acidity + sulphates
##
##                        Df Sum of Sq    RSS     AIC
## + fixed.acidity         1    1.2486 145.64 -3794.3
## + pH                    1    1.0402 145.85 -3792.0
## <none>                               146.89 -3788.1
## + chlorides             1    0.6600 146.23 -3787.9
## + total.sulfur.dioxide  1    0.6541 146.23 -3787.8
## + citric.acid           1    0.5767 146.31 -3787.0
## + free.sulfur.dioxide   1    0.5650 146.32 -3786.8
## + residual.sugar        1    0.1945 146.69 -3782.8
## + density               1    0.0577 146.83 -3781.3
## - sulphates             1    2.6129 149.50 -3767.2
## - volatile.acidity      1    4.6212 151.51 -3745.9
## - alcohol               1   23.5784 170.47 -3557.4
##
## Step:  AIC=-3794.32
## excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity
##
##                        Df Sum of Sq    RSS     AIC
## + density               1    0.7453 144.89 -3795.2
## + chlorides             1    0.7347 144.90 -3795.0
## <none>                               145.64 -3794.3
## + total.sulfur.dioxide  1    0.4408 145.20 -3791.8
## + free.sulfur.dioxide   1    0.3118 145.33 -3790.4
## + pH                    1    0.1474 145.49 -3788.6
## - fixed.acidity         1    1.2486 146.89 -3788.1
## + residual.sugar        1    0.0916 145.55 -3788.0
## + citric.acid           1    0.0013 145.64 -3787.0
## - sulphates             1    2.1214 147.76 -3778.6
## - volatile.acidity      1    3.3115 148.95 -3765.8
## - alcohol               1   24.5820 170.22 -3552.3
##
## Step:  AIC=-3795.15
## excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity +
##     density
##
##                        Df Sum of Sq    RSS     AIC
## + residual.sugar        1    0.7918 144.10 -3796.5
## + chlorides             1    0.7467 144.15 -3796.0
## <none>                               144.89 -3795.2
## - density               1    0.7453 145.64 -3794.3
## + total.sulfur.dioxide  1    0.3622 144.53 -3791.8
## + free.sulfur.dioxide   1    0.2417 144.65 -3790.4
## + citric.acid           1    0.0042 144.89 -3787.8
## + pH                    1    0.0004 144.89 -3787.8
## - fixed.acidity         1    1.9363 146.83 -3781.3
## - sulphates             1    2.5032 147.40 -3775.1
```

```
## - volatile.acidity      1    2.5863 147.48 -3774.2
## - alcohol               1   11.9413 156.84 -3675.9
##
## Step:  AIC=-3796.54
## excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity +
##     density + residual.sugar
##
##                        Df Sum of Sq    RSS     AIC
## + chlorides             1    0.8870 143.22 -3799.0
## + total.sulfur.dioxide  1    0.6948 143.41 -3796.9
## <none>                              144.10 -3796.5
## - residual.sugar        1    0.7918 144.89 -3795.2
## + free.sulfur.dioxide   1    0.4807 143.62 -3794.5
## + pH                    1    0.1366 143.97 -3790.7
## + citric.acid           1    0.0024 144.10 -3789.2
## - density               1    1.4456 145.55 -3788.0
## - volatile.acidity      1    2.3570 146.46 -3778.0
## - fixed.acidity         1    2.5889 146.69 -3775.4
## - sulphates             1    2.8107 146.91 -3773.0
## - alcohol               1    8.0121 152.11 -3717.4
##
## Step:  AIC=-3799.03
## excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity +
##     density + residual.sugar + chlorides
##
##                        Df Sum of Sq    RSS     AIC
## + total.sulfur.dioxide  1    0.7759 142.44 -3800.3
## <none>                              143.22 -3799.0
## + free.sulfur.dioxide   1    0.5457 142.67 -3797.8
## - chlorides             1    0.8870 144.10 -3796.5
## - residual.sugar        1    0.9321 144.15 -3796.0
## + citric.acid           1    0.0424 143.17 -3792.1
## + pH                    1    0.0219 143.19 -3791.9
## - density               1    1.5593 144.77 -3789.1
## - volatile.acidity      1    1.9168 145.13 -3785.2
## - fixed.acidity         1    2.7610 145.98 -3775.9
## - sulphates             1    3.6730 146.89 -3765.9
## - alcohol               1    6.5255 149.74 -3735.2
##
## Step:  AIC=-3800.34
## excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity +
##     density + residual.sugar + chlorides + total.sulfur.dioxide
##
##                        Df Sum of Sq    RSS     AIC
## <none>                              142.44 -3800.3
## - total.sulfur.dioxide  1    0.7759 143.22 -3799.0
## - chlorides             1    0.9681 143.41 -3796.9
## + citric.acid           1    0.1708 142.27 -3794.9
## + free.sulfur.dioxide   1    0.0510 142.39 -3793.5
## - residual.sugar        1    1.3159 143.75 -3793.0
```

```
## + pH                    1    0.0001 142.44 -3793.0
## - density               1    1.6974 144.14 -3788.8
## - volatile.acidity      1    1.8413 144.28 -3787.2
## - fixed.acidity         1    2.5822 145.02 -3779.0
## - sulphates             1    4.0327 146.47 -3763.1
## - alcohol               1    5.2833 147.72 -3749.5
```

```
lm_forward <- step(lm_zero, scope = list(lower = lm_zero, upper = lm_full),
                   direction = "forward", k = log(nrow(wine_lm)))
```

```
## Start:  AIC=-3419.41
## excellent ~ 1
##
##                        Df Sum of Sq    RSS     AIC
## + alcohol               1   31.1157 156.44 -3702.1
## + volatile.acidity      1   13.7446 173.81 -3533.7
## + citric.acid           1    8.6466 178.90 -3487.5
## + sulphates             1    7.4635 180.09 -3477.0
## + density               1    4.2458 183.31 -3448.6
## + total.sulfur.dioxide  1    3.6507 183.90 -3443.5
## + fixed.acidity         1    2.7035 184.85 -3435.3
## + chlorides             1    1.7759 185.78 -3427.2
## + free.sulfur.dioxide   1    0.9655 186.59 -3420.3
## <none>                             187.55 -3419.4
## + pH                    1    0.6154 186.94 -3417.3
## + residual.sugar        1    0.4281 187.12 -3415.7
##
## Step:  AIC=-3702.11
## excellent ~ alcohol
##
##                        Df Sum of Sq    RSS     AIC
## + volatile.acidity      1    6.9349 149.50 -3767.2
## + citric.acid           1    5.4833 150.95 -3751.8
## + sulphates             1    4.9266 151.51 -3745.9
## + fixed.acidity         1    3.9681 152.47 -3735.8
## + pH                    1    3.8956 152.54 -3735.1
## <none>                             156.44 -3702.1
## + density               1    0.6635 155.77 -3701.5
## + total.sulfur.dioxide  1    0.6087 155.83 -3701.0
## + free.sulfur.dioxide   1    0.3562 156.08 -3698.4
## + residual.sugar        1    0.1764 156.26 -3696.5
## + chlorides             1    0.0103 156.43 -3694.8
##
## Step:  AIC=-3767.23
## excellent ~ alcohol + volatile.acidity
##
##                        Df Sum of Sq    RSS     AIC
## + sulphates             1   2.61288 146.89 -3788.1
## + fixed.acidity         1   1.74012 147.76 -3778.6
## + pH                    1   1.60731 147.89 -3777.1
```

```
## + citric.acid           1   1.16996 148.33 -3772.4
## <none>                              149.50 -3767.2
## + total.sulfur.dioxide  1   0.46898 149.03 -3764.9
## + free.sulfur.dioxide   1   0.43985 149.06 -3764.6
## + density               1   0.32990 149.17 -3763.4
## + residual.sugar        1   0.20079 149.30 -3762.0
## + chlorides             1   0.00313 149.50 -3759.9
##
## Step:  AIC=-3788.05
## excellent ~ alcohol + volatile.acidity + sulphates
##
##                         Df Sum of Sq    RSS     AIC
## + fixed.acidity          1   1.24864 145.64 -3794.3
## + pH                     1   1.04016 145.85 -3792.0
## <none>                              146.89 -3788.1
## + chlorides              1   0.66001 146.23 -3787.9
## + total.sulfur.dioxide  1   0.65413 146.23 -3787.8
## + citric.acid           1   0.57672 146.31 -3787.0
## + free.sulfur.dioxide   1   0.56495 146.32 -3786.8
## + residual.sugar        1   0.19449 146.69 -3782.8
## + density               1   0.05766 146.83 -3781.3
##
## Step:  AIC=-3794.32
## excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity
##
##                         Df Sum of Sq    RSS     AIC
## + density                1   0.74532 144.89 -3795.2
## + chlorides              1   0.73469 144.90 -3795.0
## <none>                              145.64 -3794.3
## + total.sulfur.dioxide  1   0.44080 145.20 -3791.8
## + free.sulfur.dioxide   1   0.31178 145.33 -3790.4
## + pH                     1   0.14741 145.49 -3788.6
## + residual.sugar        1   0.09160 145.55 -3788.0
## + citric.acid           1   0.00130 145.64 -3787.0
##
## Step:  AIC=-3795.15
## excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity +
##     density
##
##                         Df Sum of Sq    RSS     AIC
## + residual.sugar        1   0.79184 144.10 -3796.5
## + chlorides              1   0.74672 144.15 -3796.0
## <none>                              144.89 -3795.2
## + total.sulfur.dioxide  1   0.36217 144.53 -3791.8
## + free.sulfur.dioxide   1   0.24169 144.65 -3790.4
## + citric.acid           1   0.00422 144.89 -3787.8
## + pH                     1   0.00039 144.89 -3787.8
##
## Step:  AIC=-3796.54
## excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity +
```

```
##     density + residual.sugar
##
##                         Df Sum of Sq    RSS     AIC
## + chlorides              1   0.88701 143.22 -3799.0
## + total.sulfur.dioxide   1   0.69482 143.41 -3796.9
## <none>                               144.10 -3796.5
## + free.sulfur.dioxide    1   0.48073 143.62 -3794.5
## + pH                     1   0.13657 143.97 -3790.7
## + citric.acid            1   0.00237 144.10 -3789.2
##
## Step:  AIC=-3799.03
## excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity +
##     density + residual.sugar + chlorides
##
##                         Df Sum of Sq    RSS     AIC
## + total.sulfur.dioxide   1   0.77591 142.44 -3800.3
## <none>                               143.22 -3799.0
## + free.sulfur.dioxide    1   0.54568 142.67 -3797.8
## + citric.acid            1   0.04244 143.17 -3792.1
## + pH                     1   0.02190 143.19 -3791.9
##
## Step:  AIC=-3800.34
## excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity +
##     density + residual.sugar + chlorides + total.sulfur.dioxide
##
##                        Df Sum of Sq    RSS     AIC
## <none>                              142.44 -3800.3
## + citric.acid           1  0.170764 142.27 -3794.9
## + free.sulfur.dioxide   1  0.051036 142.39 -3793.5
## + pH                    1  0.000144 142.44 -3793.0
```

Therefore my final linear model is **excellent ~ alcohol + volatile.acidity + sulphates + fixed.acidity + density + residual.sugar + chlorides + total.sulfur.dioxide**

 # Q5 Part B Fitting logistic models

```
# dropping quality
wine_glm <- wine[-12]
str(wine_glm)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.
5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.
065 0.073 0.071 ...
```

```
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3
.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57
0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ excellent           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 2 1 .
..
```

```
log_full <- glm(excellent ~ ., family = "binomial", data = wine_glm)
summary(log_full)
```

```
##
## Call:
## glm(formula = excellent ~ ., family = "binomial", data = wine_glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9878  -0.4351  -0.2207  -0.1222   2.9869
##
## Coefficients:
##                         Estimate  Std. Error z value         Pr(>|z|)
## (Intercept)           242.762519  108.053640   2.247         0.024660
## fixed.acidity           0.274953    0.125278   2.195         0.028183
## volatile.acidity       -2.581002    0.784285  -3.291         0.000999
## citric.acid             0.567794    0.838510   0.677         0.498313
## residual.sugar          0.239464    0.073733   3.248         0.001163
## chlorides              -8.816365    3.364760  -2.620         0.008788
## free.sulfur.dioxide     0.010821    0.012235   0.884         0.376469
## total.sulfur.dioxide   -0.016531    0.004894  -3.378         0.000731
## density              -257.797579  110.398918  -2.335         0.019536
## pH                      0.224185    0.998361   0.225         0.822327
## sulphates               3.749879    0.541583   6.924 0.00000000000439
## alcohol                 0.753339    0.131609   5.724 0.00000001040036
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  870.86  on 1587  degrees of freedom
## AIC: 894.86
##
## Number of Fisher Scoring iterations: 6
```

Comparing the Full linear regression model with the logistic regression model

	Linear Model		Logistic Model	
	Estimate	**Pr(>|z|)**	**Estimate**	**Pr(>|t|)**
fixed.acidity	0.274953	0.028183	0.0340033	0.00462
volatile.acidity	-2.581002	0.000999	-0.1782816	0.001468
citric.acid	0.567794	0.498313	0.0868324	0.201774
residual.sugar	0.239464	0.001163	0.0251971	0.000286
chlorides	-8.816365	0.008788	-0.6556214	0.00073
free.sulfur.dioxide	0.010821	0.376469	-0.0005526	0.581817
total.sulfur.dioxide	-0.016531	0.000731	-0.0006658	0.048159
density	-257.797579	0.019536	-36.677326	0.000251
pH	0.224185	0.822327	0.0172498	0.845516
sulphates	3.749879	4.39E-12	0.3515431	3.89E-11
alcohol	0.753339	1.04004E-08	0.0761776	6.111E-10

I observed that only the beta coefficients have changed and significance level remained the same

 # Part C

#Q1 Trim Logisitc Model

Free sulphur dioxide and total sulphur dioxide are collinear therefore we have to drop one of the variable in the full model. I will drop free sulphur dioxide, since it was not significant in the above steps.

I deployed stepwise selection using AIC and BIC criterion along with Chi-Squared test for finding the significant variables. A summary of the final models obtained is presented below:

	Final Model	No of Variables	AIC	BIC
Step AIC	excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar + density	8	890.08	938.47
Step BIC	excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity	6	897.17	934.81
Chi Square (using drop1 function)	excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar + density	8	890.08	938.47

Step AIC function has same number of significant variables as coming out from chi squared test. Since my Step AIC has maximum number of variables, lowest AIC and comparable BIC with step BIC method. I will go with the Step AIC method for variable selection.

Final Model:

excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar + density

```
wine_glm <- wine_glm[-6]
log_full2 <- glm(excellent ~ ., data = wine_glm, family = "binomial")
log_zero <- glm(excellent ~ 1, family = "binomial", data = wine_glm)
log_step_BIC <- step(log_zero, scope = list(lower = log_zero, upper = log_ful
l2),
                direction = "both", k = log(nrow(wine_glm)))
```

```
log_step_AIC <- step(log_zero, scope = list(lower = log_zero, upper = log_ful
l2),
                     direction = "forward", k = 2)

AIC(log_step_AIC)

## [1] 890.076

BIC(log_step_BIC)

## [1] 934.8134

BIC(log_step_AIC)

## [1] 938.4702

# Chi Square test
log_chi <- drop1(log_full2 ,test = "Chi")
```

 # Q2 Quantify and explain the influence of alcohol, residual sugar, and pH

I have alcohol and residual.sugar in the final model. Quantifying these variables using the final logistic model obtained above:

	Estimate	Std. Error	Pr(>|t|)	Odds Ratio
alcohol	0.782253	0.112017	2.88E-12	2.186392663
residual.sugar	0.232843	0.070088	0.000893	1.262183301

It means odds of response variable (excellent) will increase by 1.18 (118%) if alcohol content is increased by 1 vol% unit.

Similarly odds of response variable (excellent) will increase by 0.26 (26%) if residual.sugar content is increased by 1 gm/dm^3.

```
log_final <- glm(excellent ~ alcohol + volatile.acidity + sulphates + total.s
ulfur.dioxide +
                   chlorides + fixed.acidity + residual.sugar + density,
                 data = wine_glm, family = "binomial")
```

 # Q3 Calculate Confusion Matrix, specificity, and sensitivity

Using the cut-off probability of 0.5, I got the following confusion matrix, along with specificity and sensitivity values

	Predicted Probability	
Excellent	No	Yes
0	1336	46
1	145	72

Specificity = 1336 / (1336 + 46) = 0.9667 or 96.69%;

Sensitivity = 72 / (72 + 145) = 0.2350 or 23.50%

Since increasing the cut-off probability increases the specificity and decreases the sensitivity. Setting the cut-off probability depends upon the business case. However, in this case, I feel it is necessary that our model does not predict that wine is good and in reality it is bad (FP). It is okay if wine is predicted bad, even if it is good for the customers. Therefore cut-off probability should increase by some margin (upto 0.6)

```
logpredprob <- predict(log_final,type ="response")
log_predprob_cut <- ifelse(logpredprob < 0.5, "no", "yes")
cut_dataframe <- data.frame(wine_glm, logpredprob, log_predprob_cut)
tab1 <- xtabs(~ excellent + log_predprob_cut, cut_dataframe)
tab1
```

```
##          log_predprob_cut
## excellent   no  yes
##         0 1336   46
##         1  145   72
```

```
specificity1 <- tab1[1,1]/(tab1[1,1] + tab1[1,2])
sensitivity1 <- tab1[2,2]/(tab1[2,1] + tab1[2,2])
specificity1
```

```
## [1] 0.9667149
```

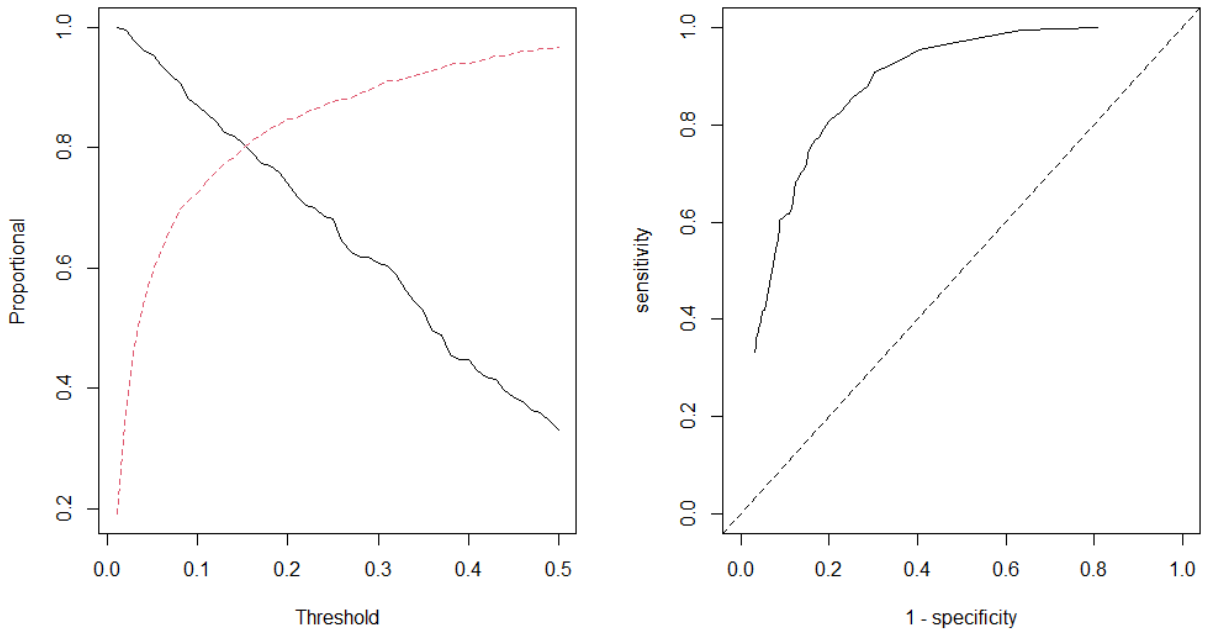
```
sensitivity1
```

```
## [1] 0.3317972
```

```
#We can increase the cut-off probability to reduce FP
```

#Q4 ROC curve and AUC

ROC curve and corresponding AUC values are presented below:



AUC: 0.882

```
thresh <- seq(0.01, 0.5, 0.01)
sensitivity <- specificity <- rep(NA, length(thresh))
for(j in seq(along = thresh)) {
  pp <- ifelse(cut_dataframe$logpredprob < thresh[j], "no", "yes")
  xx <- xtabs(~ excellent + pp, cut_dataframe)
  specificity[j] <- xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity[j] <- xx[2,2]/(xx[2,1]+xx[2,2])
  }

par(mfrow = c(1,2))
matplot(thresh, cbind(sensitivity,specificity),type="l",
        xlab ="Threshold", ylab ="Proportional", lty =1:2)
plot(1 - specificity, sensitivity, type="l", xlim = c(0,1), ylim = c(0,1)); a
bline(0,1, lty=2)

library(pROC)
auc(cut_dataframe$excellent, logpredprob)

## Area under the curve: 0.882
```

#Q5 Prediction

Using predict (), I get the chances of the 1st bottle is excellent. Using type = "link" and taking inverse of logit finction (ilogit), we can calculate the 95% CIs. My predicted outcomes for the 1st and 268th bottle:

- **1st Bottle:**
  - **Chances : 0.00768** ; **95% CI Values : (0.004481932, 0.013130163)**
- **268th Bottle:**
  - **Chances: 0.6286813 ; 95% CI Values : (0.4853097, 0.7524842)**

```
# Prediction for the 1st and 268th bottle
pred_1st_bottle <- predict(log_final, wine_glm[1,], type = "link", se.fit = T
RUE)
library(faraway)
p1 <- ilogit(pred_1st_bottle$fit)
p1_ci <- ilogit(c(pred_1st_bottle$fit - 1.96 * pred_1st_bottle$se.fit,
                  pred_1st_bottle$fit + 1.96 * pred_1st_bottle$se.fit))
p1

##           1
## 0.007680065

p1_ci

##           1           1
## 0.004481932 0.013130163

pred_268_bottle <- predict(log_final, wine_glm[268,], type = "link", se.fit =
TRUE)
p2 <- ilogit(pred_268_bottle$fit)
p2_ci <- ilogit(c(pred_268_bottle$fit - 1.96 * pred_268_bottle$se.fit,
                  pred_268_bottle$fit + 1.96 * pred_268_bottle$se.fit))
p2

##       268
## 0.6286813

p2_ci

##       268       268
## 0.4853097 0.7524842
```

Part D. Link functions and Dispersion Parameters

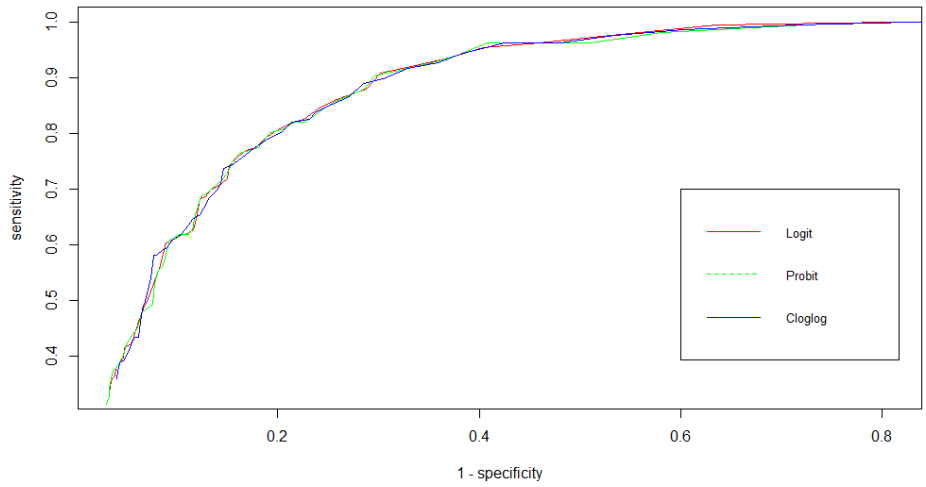
#Q1 Fit models with other link functions and compare

Using other link functions: Probit and Cloglog: I noticed the following observations for the three different link functions:

- **Significance:** All link functions showed had equal significance (i.e. $p < 0.05$)
- **Size and Sign of Beta Estimates:** All the three link functions had same sign for the Beta estimates. However, size of Beta estimates was different in all the link functions. Beta estimates for Cloglog were closer to Logit

	Beta Estimate Logit	Beta Estimate Probit	Beta Estimate Cloglog	Beta Sign Logit	Beta Sign Probit	Beta Sign Cloglog
Alcohol	0.7823	0.4297	0.5715	+	+	+
volatile.acidity	-2.9128	-1.5667	-2.4380	-	-	-
Sulphates	3.6987	2.0394	2.8042	+	+	+
total.sulfur.dioxide	-0.0136	-0.0075	-0.0114	-	-	-
Chlorides	-8.4408	-4.4028	-6.6640	-	-	-
fixed.acidity	0.2812	0.1442	0.2133	+	+	+
residual.sugar	0.2328	0.1237	0.2175	+	+	+
Density	-240.9465	-128.2683	-172.5074	-	-	-

	AIC	BIC	Residual Deviance DoF = Degree of Freedom	AUC	Specificity (0.5 cut-off)	Sensitivity (0.5 cut-off)
Logit	890.08	938.47	872.08 on 1590 DoF	0.882	0.967	0.235
Probit	886.86	935.25	868 on 1590 DoF	0.882	0.969	0.313
Cloglog	952.91	1001.37	934.91 on 1590 DoF	0.8812	0.959	0.359



AUC of the Probit is the highest. Moreover, its AIC, BIC, and residual deviance are least as compared to Logit and Cloglog model. Therefore I will suggest Probit model to the wine expert

```
probit_final <- glm(excellent ~ alcohol + volatile.acidity + sulphates +
      total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +
      density, family = binomial(link = probit), data = wine_glm)

clog_final <- glm(excellent ~ alcohol + volatile.acidity + sulphates +
                      total.sulfur.dioxide + chlorides + fixed.acidity + resi
dual.sugar +
                      density, family = binomial(link = cloglog), data = wine
_glm)
summary(log_final)

##
## Call:
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +
##     density, family = "binomial", data = wine_glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0158  -0.4314  -0.2220  -0.1255   2.9883
##
## Coefficients:
##                          Estimate  Std. Error z value         Pr(>|z|)
## (Intercept)           226.750350   91.628690   2.475         0.013336
## alcohol                 0.782253    0.112017   6.983 0.00000000000288
## volatile.acidity       -2.912834    0.646698  -4.504 0.00000666350911
```

```
## sulphates               3.698736    0.528654   6.997 0.00000000000262
## total.sulfur.dioxide   -0.013600    0.003447  -3.946 0.00007948205551
## chlorides              -8.440804    3.258784  -2.590         0.009593
## fixed.acidity           0.281169    0.080289   3.502         0.000462
## residual.sugar          0.232843    0.070088   3.322         0.000893
## density              -240.946491   92.021524  -2.618         0.008835
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  872.08  on 1590  degrees of freedom
## AIC: 890.08
##
## Number of Fisher Scoring iterations: 6
```

```
summary(probit_final)
```

```
##
## Call:
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +
##     density, family = binomial(link = probit), data = wine_glm)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.07806  -0.45118  -0.20474  -0.08787   3.12995
##
## Coefficients:
##                         Estimate  Std. Error z value         Pr(>|z|)
## (Intercept)           120.515485   50.679679   2.378          0.01741
## alcohol                 0.429729    0.061785   6.955 0.00000000000352
## volatile.acidity       -1.566669    0.345886  -4.529 0.00000591396162
## sulphates               2.039440    0.297365   6.858 0.00000000000696
## total.sulfur.dioxide   -0.007457    0.001861  -4.006 0.00006173828928
## chlorides              -4.402780    1.643923  -2.678          0.00740
## fixed.acidity           0.144192    0.043882   3.286          0.00102
## residual.sugar          0.123710    0.039051   3.168          0.00154
## density              -128.268310   50.865476  -2.522          0.01168
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  868.86  on 1590  degrees of freedom
## AIC: 886.86
##
## Number of Fisher Scoring iterations: 7
```

```
summary(clog_final)
```

```
##
## Call:
```

```
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +
##     density, family = binomial(link = cloglog), data = wine_glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.1486  -0.4995  -0.2983  -0.1888   2.7585
##
## Coefficients:
##                         Estimate  Std. Error z value         Pr(>|z|)
## (Intercept)           161.765500   76.998997   2.101         0.035652
## alcohol                 0.571494    0.090962   6.283 0.00000000033255
## volatile.acidity       -2.438035    0.507949  -4.800 0.00000158851933
## sulphates               2.804176    0.410807   6.826 0.00000000000873
## total.sulfur.dioxide   -0.011363    0.003156  -3.600         0.000318
## chlorides              -6.664031    2.858720  -2.331         0.019747
## fixed.acidity           0.213270    0.064002   3.332         0.000861
## residual.sugar          0.217469    0.073083   2.976         0.002924
## density              -172.507419   77.147742  -2.236         0.025347
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  934.91  on 1590  degrees of freedom
## AIC: 952.91
##
## Number of Fisher Scoring iterations: 25

AIC(log_final)

## [1] 890.076

AIC(probit_final)

## [1] 886.8559

AIC(clog_final)

## [1] 952.9129

BIC(log_final)

## [1] 938.4702

BIC(probit_final)

## [1] 935.2501

BIC(clog_final)

## [1] 1001.307
```

```r
probitpredprob <- predict(probit_final,type ="response")
probit_predprob_cut <- ifelse(probitpredprob < 0.5, "no", "yes")
cut_dataframe <- data.frame(wine_glm, probitpredprob, probit_predprob_cut)
tab2 <- xtabs(~ excellent + probit_predprob_cut, cut_dataframe)
tab2
```

```
##           probit_predprob_cut
## excellent   no  yes
##         0 1340   42
##         1  149   68
```

```r
specificity2 <- tab2[1,1]/(tab2[1,1] + tab2[1,2])
sensitivity2 <- tab2[2,2]/(tab2[2,1] + tab2[2,2])
specificity2
```

```
## [1] 0.9696093
```

```r
sensitivity2
```

```
## [1] 0.3133641
```

```r
thresh <- seq(0.01, 0.5, 0.01)
sensitivity_probit <- specificity_probit <- rep(NA, length(thresh))
for(j in seq(along = thresh)) {
  pp_probit <- ifelse(cut_dataframe$probitpredprob < thresh[j], "no", "yes")
  xx_probit <- xtabs(~ excellent + pp_probit, cut_dataframe)
  specificity_probit[j] <- xx_probit[1,1]/(xx_probit[1,1]+xx_probit[1,2])
  sensitivity_probit[j] <- xx_probit[2,2]/(xx_probit[2,1]+xx_probit[2,2])
}

clogpredprob <- predict(clog_final,type ="response")
clog_predprob_cut <- ifelse(clogpredprob < 0.5, "no", "yes")
cut_dataframe <- data.frame(wine_glm, clogpredprob, clog_predprob_cut)
tab3 <- xtabs(~ excellent + clog_predprob_cut, cut_dataframe)
tab3
```

```
##           clog_predprob_cut
## excellent   no  yes
##         0 1326   56
##         1  139   78
```

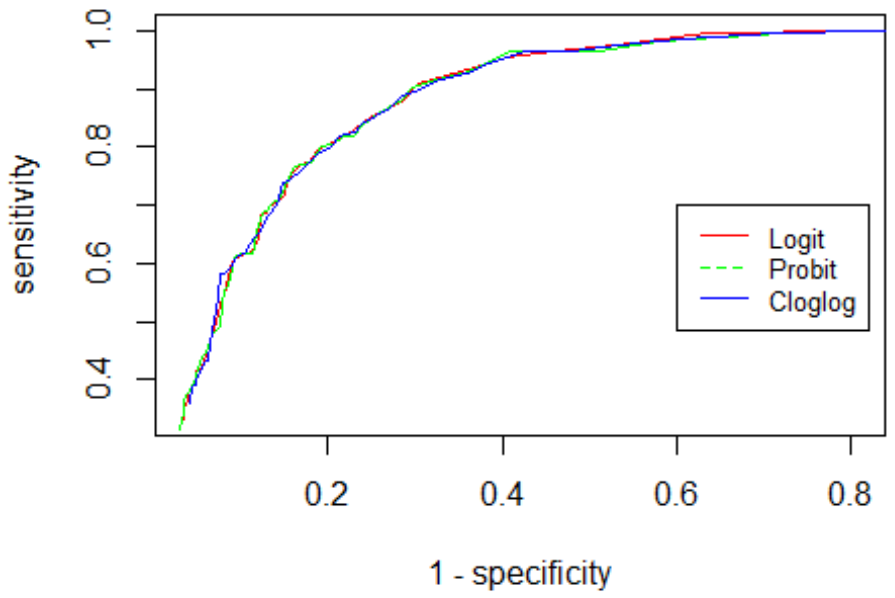
```r
specificity3 <- tab3[1,1]/(tab3[1,1] + tab3[1,2])
sensitivity3 <- tab3[2,2]/(tab3[2,1] + tab3[2,2])
specificity3
```

```
## [1] 0.959479
```

```r
sensitivity3
```

```
## [1] 0.359447
```

```r
thresh <- seq(0.01, 0.5, 0.01)
sensitivity_clog <- specificity_clog <- rep(NA, length(thresh))
for(j in seq(along = thresh)) {
  pp_clog <- ifelse(cut_dataframe$clogpredprob < thresh[j], "no", "yes")
  xx_clog <- xtabs(~ excellent + pp_clog, cut_dataframe)
  specificity_clog[j] <- xx_clog[1,1]/(xx_clog[1,1]+xx_clog[1,2])
  sensitivity_clog[j] <- xx_clog[2,2]/(xx_clog[2,1]+xx_clog[2,2])
}

# Combined ROCs
par(mfrow = c(1,1))

plot(1 - specificity, sensitivity, type = 'l', col = 'red');
lines(1 - specificity_probit, sensitivity_probit, col = 'green')
lines(1 - specificity_clog, sensitivity_clog, col = 'blue')
legend(0.6, 0.7, legend = c("Logit", "Probit", "Cloglog"),
       col = c("red", "green", "blue"), lty = 1:2, cex = 0.8)
```




```r
# AUC of the curves
auc(cut_dataframe$excellent, logpredprob)

## Area under the curve: 0.882

auc(cut_dataframe$excellent, probitpredprob)

## Area under the curve: 0.882
```

```
auc(cut_dataframe$excellent, clogpredprob)

## Area under the curve: 0.8812
```

#Q2 Dispersion Parameter

Calculating the dispersion parameter by using number of rows of wine dataset and 11 columns. sigma_squared = sum(residuals(final model, type = "pearson")^2)/(nrow(wine_glm) - 11)

We get a dispersion parameter of 0.886. Therefore our model had under-dispersion.

If we have lower dispersion then the variance of the estimated betas decreases and hence we have more certainty in determining the response.

Comparing the models where dispersion = 0.86 with dispersion = 1

	Dispersion = 0.8816				Dispersion = 1			
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	120.5155	47.5870	2.5330	0.0113240	120.5155	50.6797	2.3780	0.0174100
alcohol	0.4297	0.0580	7.4070	0.0000000	0.4297	0.0618	6.9550	0.0000000
volatile.acidity	-1.5667	0.3248	-4.8240	0.0000014	-1.5667	0.3459	-4.5290	0.0000059
sulphates	2.0394	0.2792	7.3040	0.0000000	2.0394	0.2974	6.8580	0.0000000
total.sulfur.dioxide	-0.0075	0.0017	-4.2660	0.0000199	-0.0075	0.0019	-4.0060	0.0000617
chlorides	-4.4028	1.5436	-2.8520	0.0043410	-4.4028	1.6439	-2.6780	0.0074000
fixed.acidity	0.1442	0.0412	3.4990	0.0004660	0.1442	0.0439	3.2860	0.0010200
residual.sugar	0.1237	0.0367	3.3740	0.0007410	0.1237	0.0391	3.1680	0.0015400
density	-128.2683	47.7614	-2.6860	0.0072400	-128.2683	50.8655	-2.5220	0.0116800

We see that the resgression coefficients are same; however, standard errors for each covariate is lesser in the case where dispersion = 0.8816.

We also observe that the p-values have been reduced in the case where dispersion = 0.8816.

Thus we can say that our original model with dispersion of 0.8816 is better then the model where dispersion parameter = 1.

```
sigma.squared <- sum(residuals(probit_final, type = "pearson")^2)/(nrow(wine_
glm) - 11)
summary(probit_final, dispersion = sigma.squared)
```

```
##
## Call:
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +
##     density, family = binomial(link = probit), data = wine_glm)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.07806  -0.45118  -0.20474  -0.08787   3.12995
##
## Coefficients:
##                         Estimate  Std. Error z value          Pr(>|z|)
## (Intercept)           120.515485   47.721612   2.525          0.011557
## alcohol                 0.429729    0.058178   7.386 0.000000000000151
## volatile.acidity       -1.566669    0.325697  -4.810 0.000001507765345
## sulphates               2.039440    0.280008   7.284 0.000000000000325
## total.sulfur.dioxide   -0.007457    0.001753  -4.254 0.000020962334639
## chlorides              -4.402780    1.547970  -2.844          0.004452
## fixed.acidity           0.144192    0.041321   3.490          0.000484
## residual.sugar          0.123710    0.036771   3.364          0.000767
## density              -128.268310   47.896564  -2.678          0.007406
##
## (Dispersion parameter for binomial family taken to be 0.886671)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  868.86  on 1590  degrees of freedom
## AIC: 886.86
##
## Number of Fisher Scoring iterations: 7
```

```
summary(probit_final)
```

```
##
## Call:
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +
##     density, family = binomial(link = probit), data = wine_glm)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.07806  -0.45118  -0.20474  -0.08787   3.12995
##
## Coefficients:
##                         Estimate  Std. Error z value         Pr(>|z|)
## (Intercept)           120.515485   50.679679   2.378          0.01741
## alcohol                 0.429729    0.061785   6.955 0.00000000000352
## volatile.acidity       -1.566669    0.345886  -4.529 0.00000591396162
## sulphates               2.039440    0.297365   6.858 0.00000000000696
## total.sulfur.dioxide   -0.007457    0.001861  -4.006 0.00006173828928
## chlorides              -4.402780    1.643923  -2.678          0.00740
```

```
## fixed.acidity           0.144192    0.043882   3.286          0.00102
## residual.sugar          0.123710    0.039051   3.168          0.00154
## density              -128.268310   50.865476  -2.522          0.01168
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  868.86  on 1590  degrees of freedom
## AIC: 886.86
##
## Number of Fisher Scoring iterations: 7
```

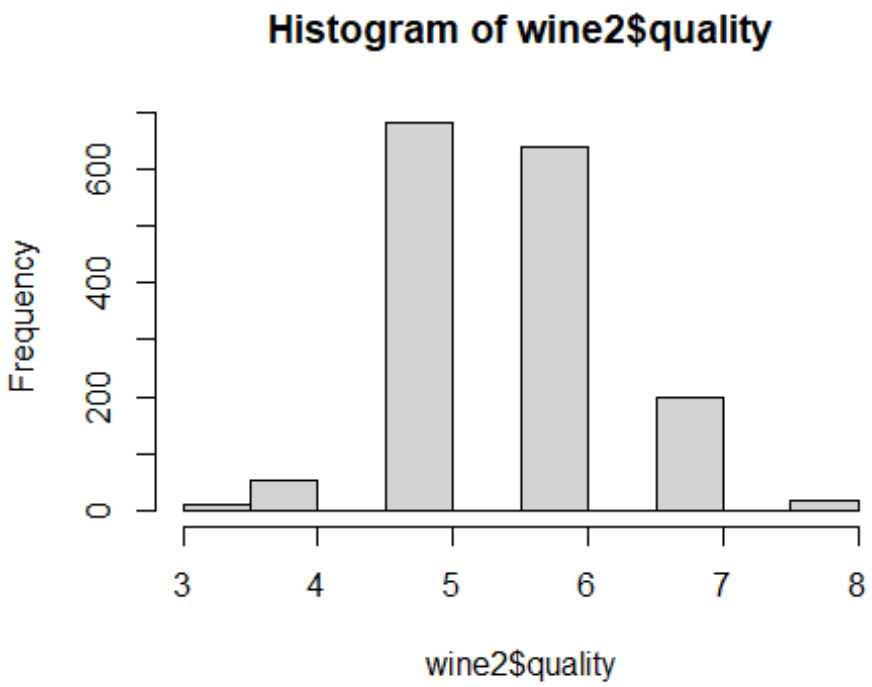
 # Part E

#Q1 Histogram and levels of Ordinal Variable

I have plotted histogram of the quality variable and frequency of different categorical levels in the quality variable are as follows:

Level	3	4	5	6	7	8
Frequency	10	53	681	638	199	18

```
wine2 <- read.csv("C:/UC/Stat Modelling/Final Project/winequality-red.csv",
                  sep = ";" )
hist(wine2$quality)
```




```
quality <- as.factor(wine2$quality)
summary(quality)

##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

#Q2 Multinomial Distribution

I Should use the logistic distribution and extend to multinomial distribution to model the quality variable . Quality variable is the multinomial variable and there is an order (ordinal variable) as ranking of the win e matters.

#Q3 Correlation

Rank	Variable	Kendall's Tau correlation with Quality Variable
1	alcohol	0.380
2	sulphates	0.299
3	citric.acid	0.167
4	fixed.acidity	0.088
5	residual.sugar	0.026
6	pH	-0.034
7	free.sulfur.dioxide	-0.046

8	density	-0.137
9	chlorides	-0.149
10	total.sulfur.dioxide	-0.157
11	volatile.acidity	-0.301

I see that alcohol had the highest positive correlation with the quality. Volatile.acidity had the most negative correlation with quality

```
library(VGAM)
k <- cor(wine2, method="kendall")
k[3]

## [1] 0.4842712

krank <- k[order(-k$quality),]
krank <- krank[12]
```

#Q4 Subset Selection

I am picking alcohol, volatile.acidity, sulphates, total.sulfur.dioxide, chlorides, fixed.acidity, residual.sugar, and density as the predictors from steps A- D and step 3 (Part E) for regressing quality variable.

Q5 Multinomial Model Fit

I have fitted a proportional odds model using the predictors from the above step for the response variable quality. I have chosen a multinomial model as it gives the flexibility for modeling ordinal variables.

I can interpret this model considering factor level – 8 as the baseline and the exponentiated coefficients can deliver the idea of change in odds of a particular rating w.r.t the level of 8. For instance, upon increasing alcohol by 1 unit, the (odds for rating 3, 4, 5, 6, and 7 w.r.t 8) will decrease by ~ (1 – 0.4707 = ~0.529) or 52.9%.

Please find the summary of the fitted model below:

```
multi.model <- vglm(quality ~ alcohol + volatile.acidity + sulphates +
                      total.sulfur.dioxide + chlorides + fixed.acidity + resi
dual.sugar +
                      density, data = wine2, family = cumulative(parallel = T
RUE))
summary(multi.model)

##
## Call:
## vglm(formula = quality ~ alcohol + volatile.acidity + sulphates +
```

```
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +
##     density, family = cumulative(parallel = TRUE), data = wine2)
##
## Coefficients:
##                          Estimate  Std. Error z value             Pr(>|z|)
## (Intercept):1        -130.226497   54.763195  -2.378              0.01741
## (Intercept):2        -128.315800   54.763832      NA                   NA
## (Intercept):3        -124.637287   54.767155  -2.276              0.02286
## (Intercept):4        -121.780469   54.758191  -2.224              0.02615
## (Intercept):5        -118.763010   54.752824  -2.169              0.03008
## alcohol                -0.753438    0.071205 -10.581 < 0.0000000000000002
## volatile.acidity        3.083311    0.327365   9.419 < 0.0000000000000002
## sulphates              -2.954007    0.352861  -8.372 < 0.0000000000000002
## total.sulfur.dioxide    0.008462    0.001673   5.059          0.000000421
## chlorides               5.362256    1.236606   4.336          0.000014492
## fixed.acidity          -0.154000    0.047697  -3.229              0.00124
## residual.sugar         -0.115281    0.043659  -2.641              0.00828
## density               133.643013   54.916816      NA                   NA
##
## Number of linear predictors:  5
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3]), logitlink(P[Y<=4]), logitlink(P[Y<=5])
##
## Residual deviance: 3084.313 on 7982 degrees of freedom
##
## Log-likelihood: -1542.157 on 7982 degrees of freedom
##
## Number of Fisher scoring iterations: 6
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2', 'density'
##
##
## Exponentiated coefficients:
##              alcohol     volatile.acidity           sulphates
##        4.707452e-01         2.183056e+01         5.213042e-02
## total.sulfur.dioxide            chlorides        fixed.acidity
##        1.008498e+00         2.132055e+02         8.572723e-01
##       residual.sugar              density
##        8.911154e-01         1.097547e+58
```

Q6 Comparison of Multinomial model with Logistic model

	Multinomial Model		Logit Model	
	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)
(Intercept):1	-130.226497	0.01741	226.75035	0.013336
(Intercept):2	-128.3158	NA		
(Intercept):3	-124.637287	0.02286		
(Intercept):4	-121.780469	0.02615		
(Intercept):5	-118.76301	0.03008		
alcohol	-0.753438	2E-16	0.782253	2.88E-12
volatile.acidity	3.083311	2E-16	-2.912834	6.66351E-06
sulphates	-2.954007	2E-16	3.698736	2.62E-12
total.sulfur.dioxide	0.008462	0.000000421	-0.0136	7.94821E-05
chlorides	5.362256	0.000014492	-8.440804	0.009593
fixed.acidity	-0.154	0.00124	0.281169	0.000462
residual.sugar	-0.115281	0.00828	0.232843	0.000893
density	133.643013	NA	-240.946491	0.008835

I observe that all the beta coefficients have changed their signs as proportional odds model is w.r.t the ordinal level of '8'. I also observed that the significance remained the same

Q7 Prediction

Yes, I can use this model to get the binary response excellent or not excellent. Since my model follows a probability distribution function where order was important, I can use them to classify between excellent (>7) vs non excellent (<7).

I plan to use the probabilities of 3, 4, 5, and 6 and sum them. If the sum is greater then 0.5 then it is 'not excellent' and 'excellent' otherwise.

Prediction for the 1st bottle: Not Excellent

Prediction for the 268st bottle: Excellent

Therefore I conclude that the results match with the logistic model I created earlier. Please find the R codes below:

```
# Part E 7.
first_bottle_multi <- wine2[1,]
class(first_bottle_multi)

## [1] "data.frame"
```

```r
library(VGAM)
first_bottle_multi <- predict(multi.model, newdata = wine2[1,], "response")
ifelse(sum(first_bottle_multi[,1:4]) > 0.5,"Not Excellent","Excellent")
```

```
## [1] "Not Excellent"
```

```r
twosixtyeight_bottle_multi <- predict(multi.model, wine2[268,],type ='respons
e')
ifelse(sum(twosixtyeight_bottle_multi[,1:4]) > 0.5,"Not Excellent","Excellent
")
```

```
## [1] "Excellent"
```

Appendix

Outlier detection:

```
library(gridExtra)
out1 <- ggplot(wine, aes(y= fixed.acidity)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)

out2 <- ggplot(wine, aes(y= volatile.acidity)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)

out3 <- ggplot(wine, aes(y= citric.acid)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)

out4 <- ggplot(wine, aes(y= residual.sugar)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)

out5 <- ggplot(wine, aes(y= chlorides)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)

out6 <- ggplot(wine, aes(y = free.sulfur.dioxide)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)

out7 <- ggplot(wine, aes(y= total.sulfur.dioxide)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)

out8 <- ggplot(wine, aes(y= density)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)
```

```
out9 <- ggplot(wine, aes(y= pH)) +
  geom_boxplot(outlier.colour="red",outlier.size=2)

out10 <- ggplot(wine, aes(y= sulphates)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)

out11 <- ggplot(wine, aes(y= alcohol)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)

out12 <- ggplot(wine, aes(y= quality)) +
 geom_boxplot(outlier.colour="red",outlier.size=2)

h1 <- grid.arrange(out1, out2, out3, out4, out5, out6, nrow = 2, ncol = 3)
h2 <- grid.arrange(out7, out8, out9, out10, out11, out12, nrow = 2, ncol = 3)
```

Part B Q3

Plotting visualizations for the association of response with other variables

```
g1 <- ggplot(wine, aes(x = fixed.acidity, fill = (excellent))) +
  geom_bar(position="dodge")

g2 <- ggplot(wine, aes(x = volatile.acidity, fill = (excellent))) +
  geom_bar(position="dodge")

g3 <- ggplot(wine, aes(x = citric.acid, fill = (excellent))) +
  geom_bar(position="dodge")

g4 <- ggplot(wine, aes(x = residual.sugar, fill = (excellent))) +
  geom_bar(position="dodge")

g5 <- ggplot(wine, aes(x = chlorides, fill = (excellent))) +
  geom_bar(position="dodge")
```

```r
g6 <- ggplot(wine, aes(x = free.sulfur.dioxide, fill = (excellent))) +
  geom_bar(position="dodge")

g7 <- ggplot(wine, aes(x = total.sulfur.dioxide, fill = (excellent))) +
  geom_bar(position="dodge")

g8 <- ggplot(wine, aes(x = density, fill = (excellent))) +
  geom_bar(position="dodge")

g9 <- ggplot(wine, aes(x = pH, fill = (excellent))) +
  geom_bar(position="dodge")

g10 <- ggplot(wine, aes(x = sulphates, fill = (excellent))) +
  geom_bar(position="dodge")

g11 <- ggplot(wine, aes(x = alcohol, fill = (excellent))) +
  geom_bar(position="dodge")

grid.arrange(g1, g2, g3, g4, g5, g6, ncol=3, nrow = 2)
```

```r
grid.arrange(g7, g8, g9, g10, g11, ncol=3, nrow = 2)
```