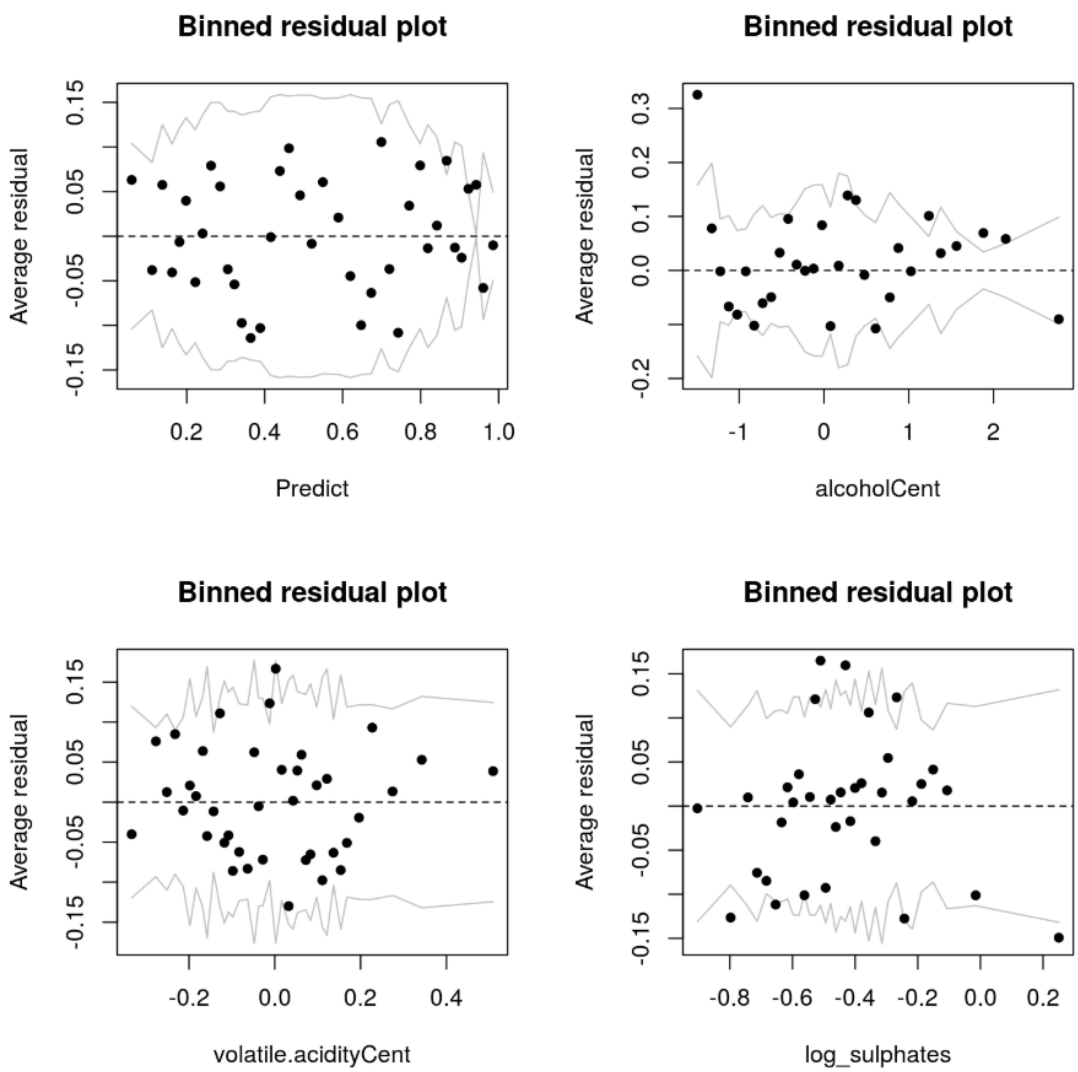
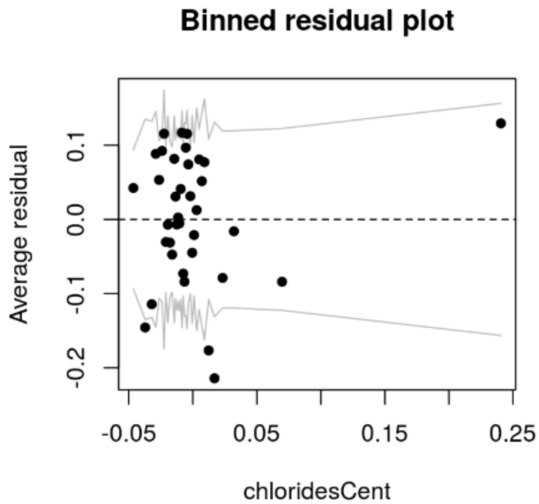
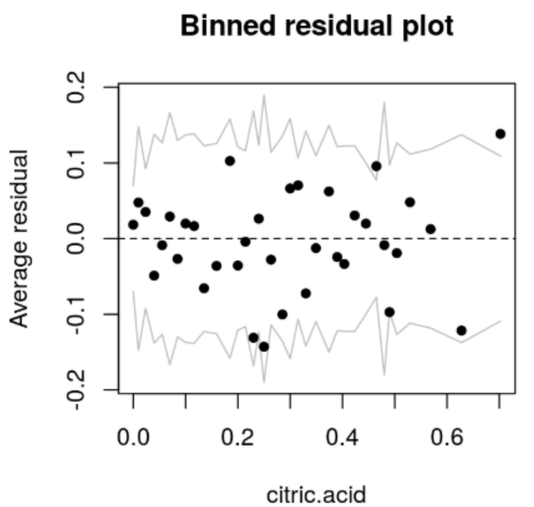
Assumptions

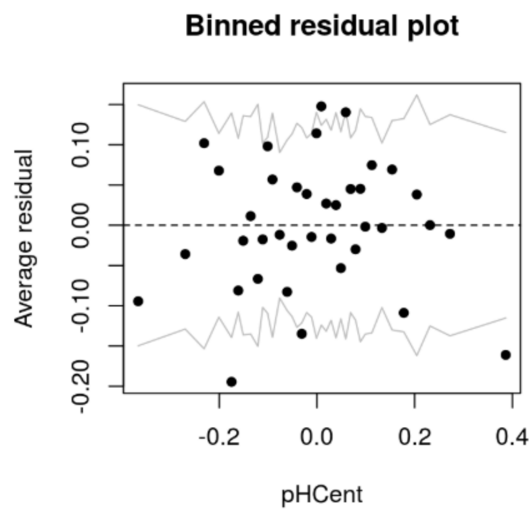
By: Bob Ding, Lynn Fan, Alice Jiang
December 15, 2018

Logistic Regression

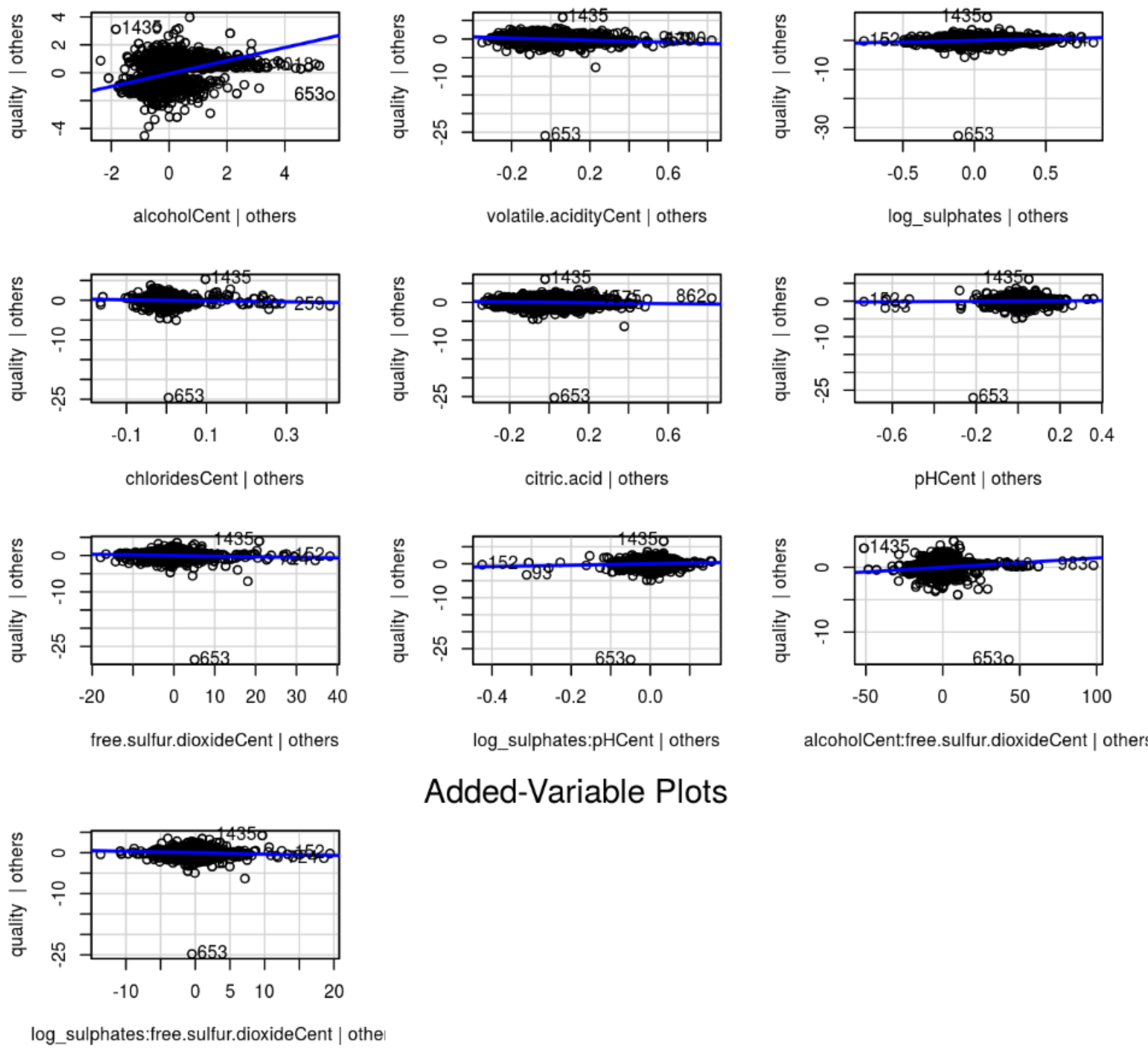




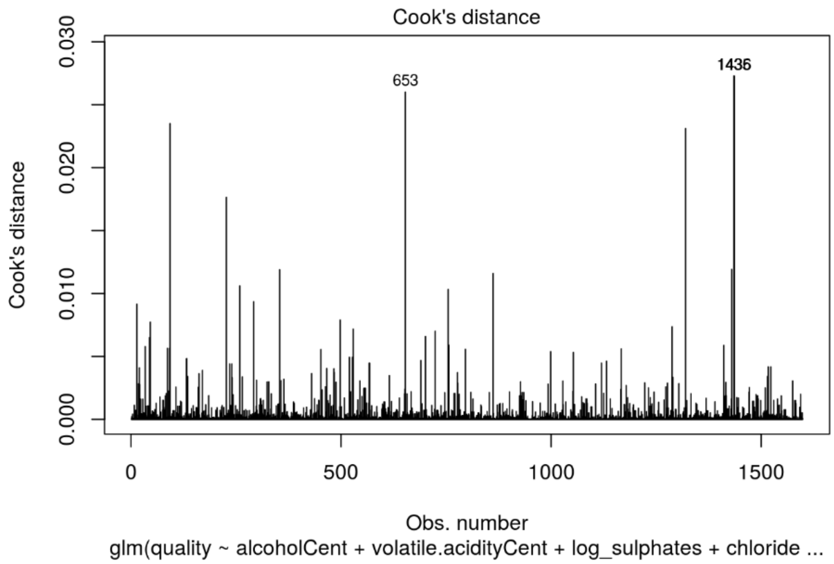




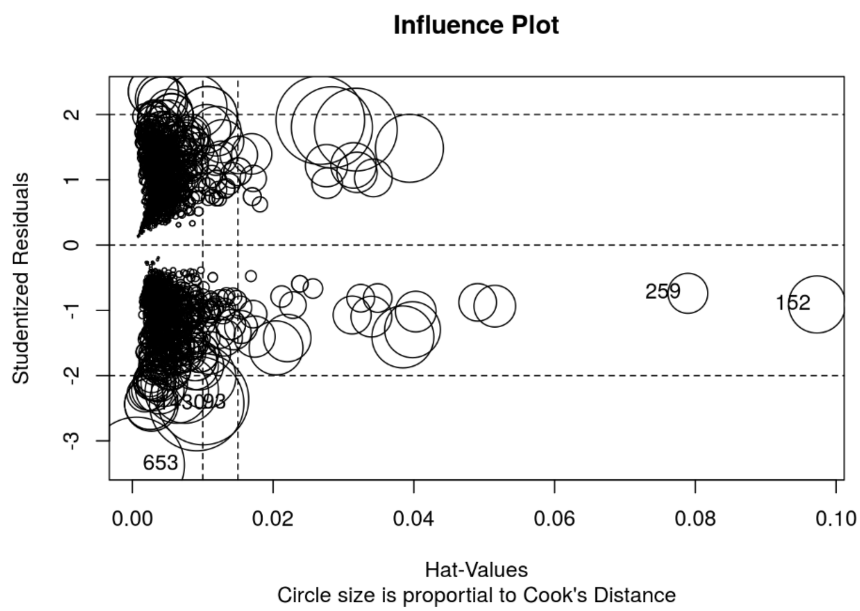
All of the binned residual plots show random pattern. The binned residual plot of alcoholCent seems to have an outlier on the top left and the binned residual plot of chloridesCent has an outlier on the far right. The outlier indicates that the average residual of the bin is different from that of other bins, and could be due to an outlier observation. We will investigate for influence points later. Overall, there is no major concerns of assumption violations.



Observations 653 and 1435 are consistently identified in AV Plots. They could be outliers and we will consider removing these observations if they are influential.

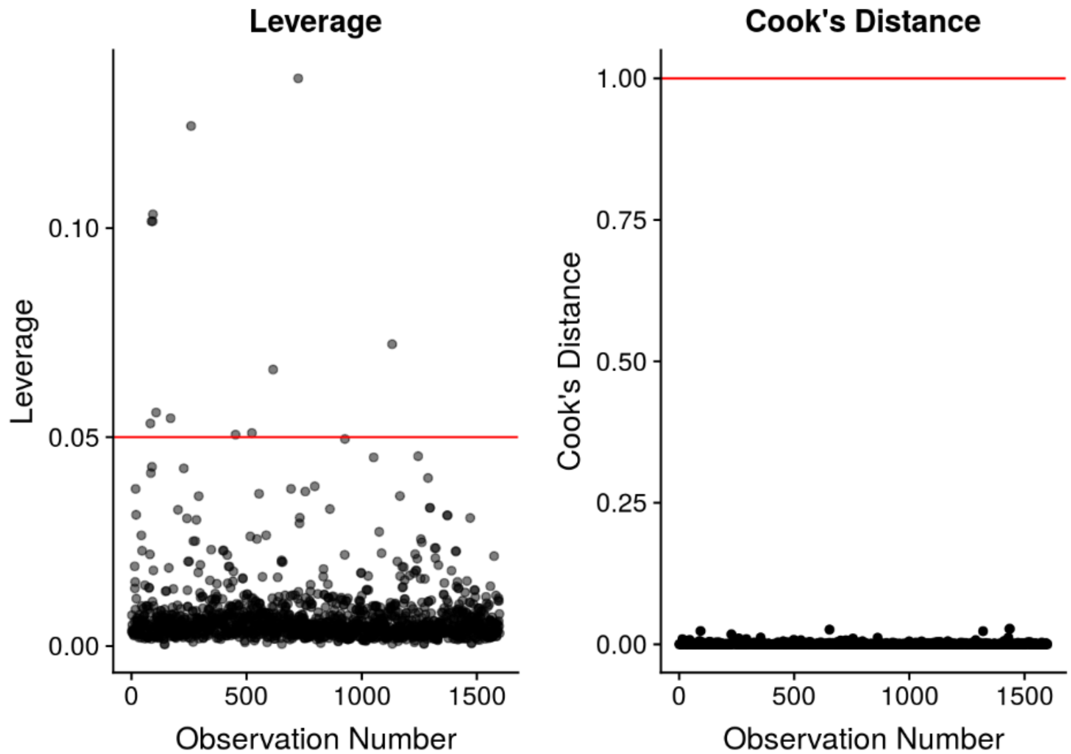


We can see that observations 653 and 1435 have noticeably higher Cook's Distance than other observations. However, their Cook's Distance is still much smaller than the threshold of 1.




```
##         StudRes          Hat       CookD
## 93   -2.4287943 0.0091359820 0.019405024
## 152  -0.9101741 0.0972935691 0.007088194
## 259  -0.7393382 0.0789403851 0.003462105
## 653  -3.3710586 0.0005899936 0.019939563
## 1430 -2.4435909 0.0027063910 0.006228563
```

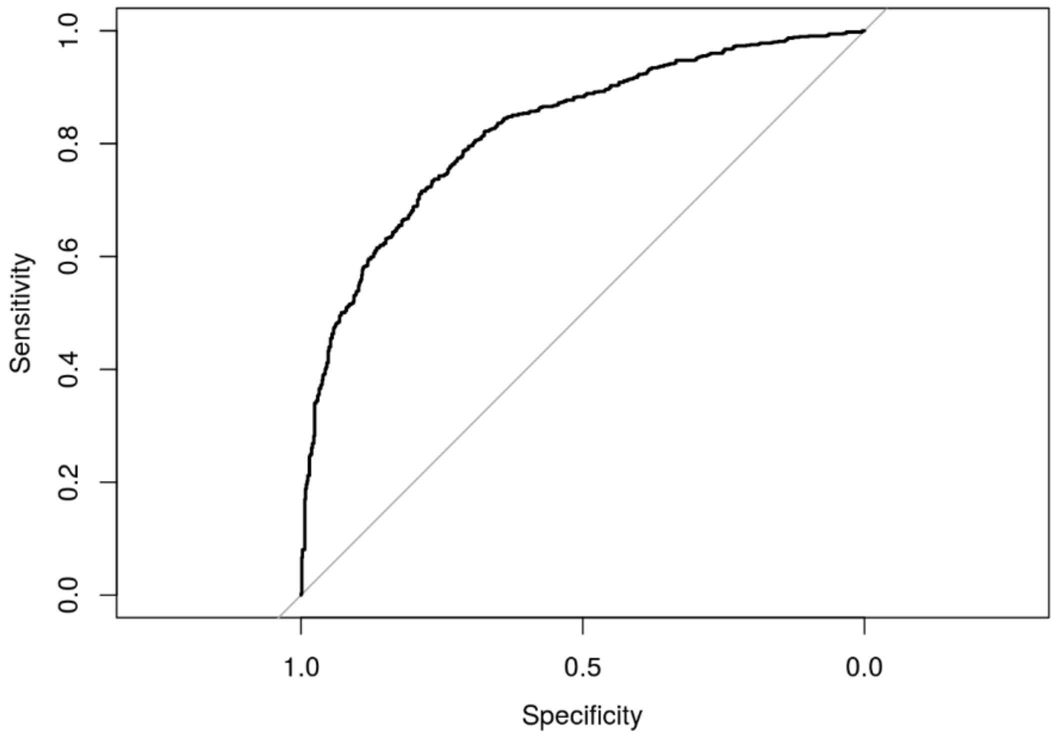
We can see that some observations have high studentized residual value with magnitude greater than 2, but they have very low hat values and acceptable circle size. Observations 259 and 152 have very high leverage (hat values) but are within reasonable studentized residual value and circle size. Observation 653 has high magnitude of studentized residual, but very low leverage and reasonable circle size. Overall, there is no outstanding influence point in the data set, based on considerations of leverage and cook's distance.



Overall, there is no significant influence point.

names	x
alcoholCent	1.248964
volatile.acidityCent	1.546294
log_sulphates	1.544255
chloridesCent	1.377441
citric.acid	2.095763
pHCent	4.868276
free.sulfur.dioxideCent	5.463815
log_sulphates:pHCent	4.529185
alcoholCent:free.sulfur.dioxideCent	1.185423
log_sulphates:free.sulfur.dioxideCent	5.468602

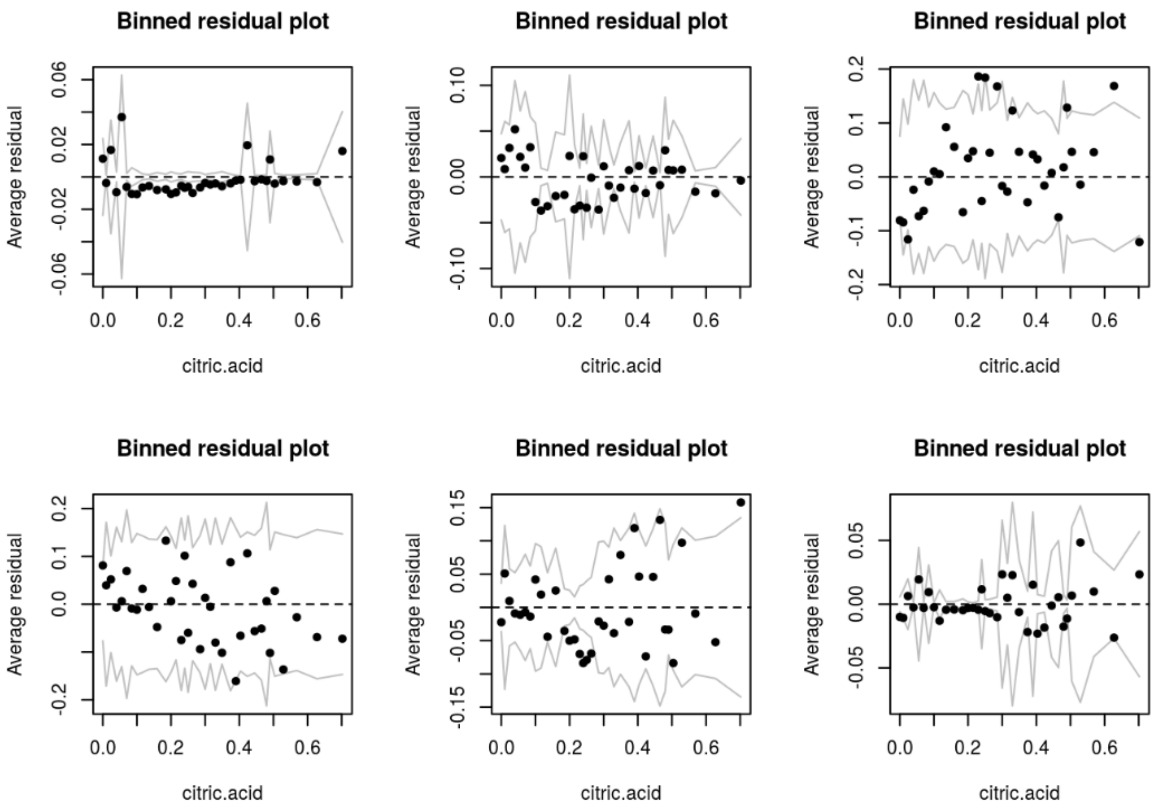
VIF values are small, so there is no major concerns of multicollinearity.



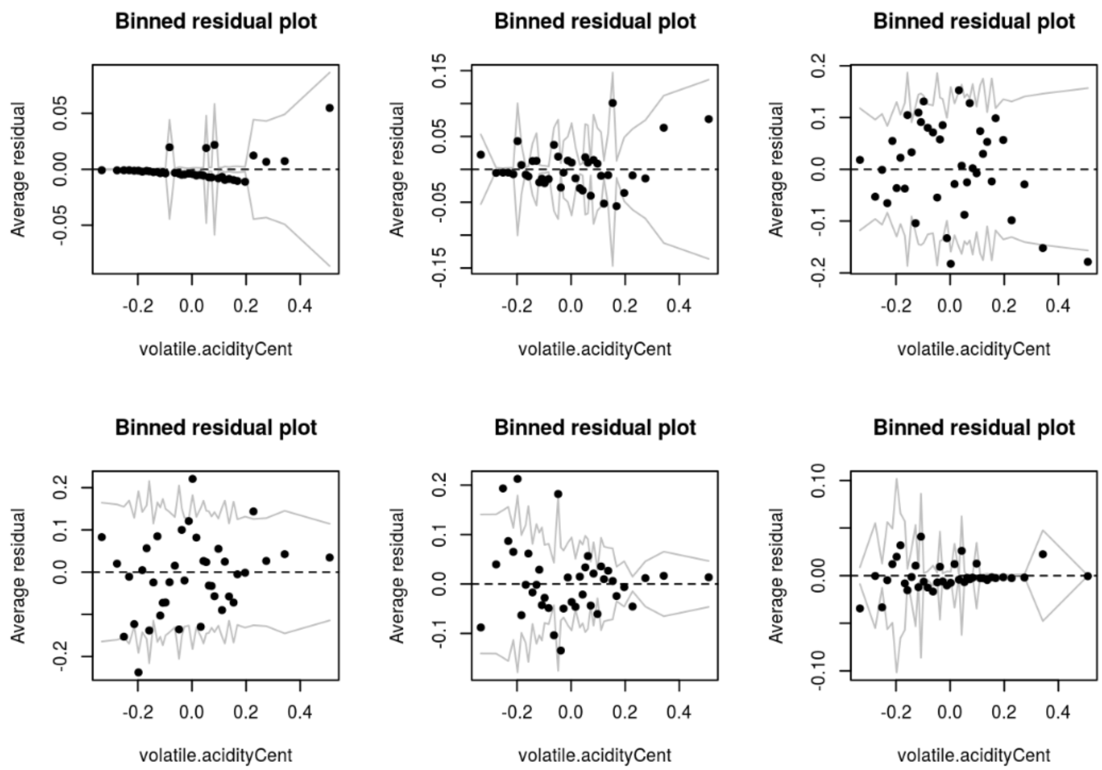

```
## Area under the curve: 0.824
```

From the ROC curve and AUC calculation, we can see the curve is fairly close to the top left corner (area under the curve is close to 1). This shows that the logistic model is able to distinguish between good and not good quality, so this is a pretty good model.

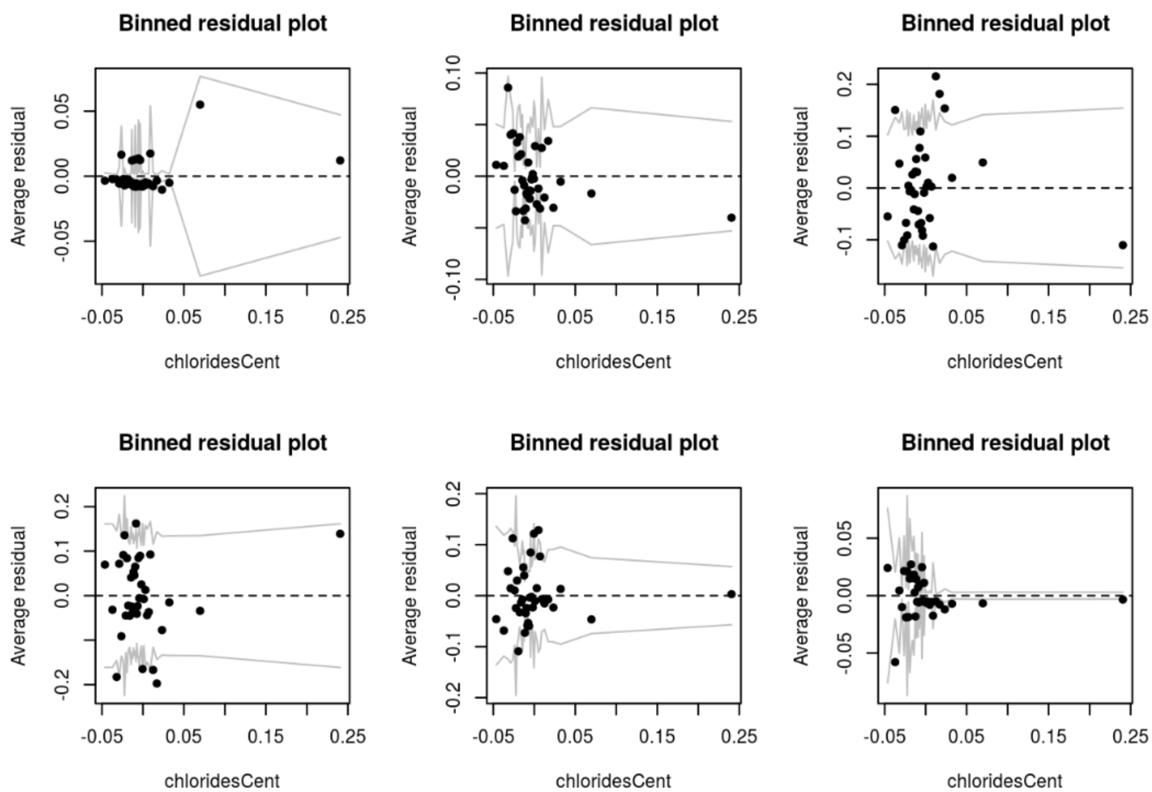
Ordinal Logistic Regression



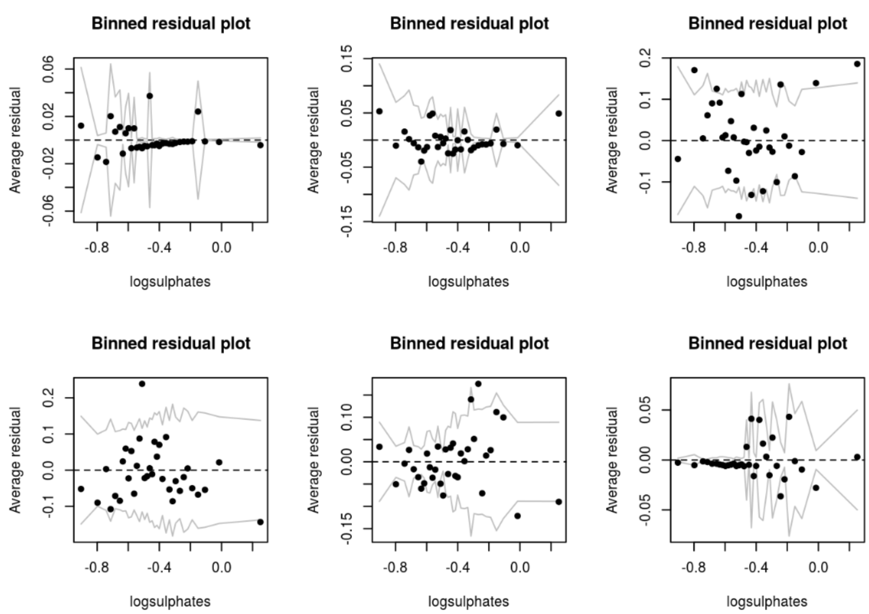
The binned residual plots of citric.acid at each quality level show random pattern and raises no obvious concerns.



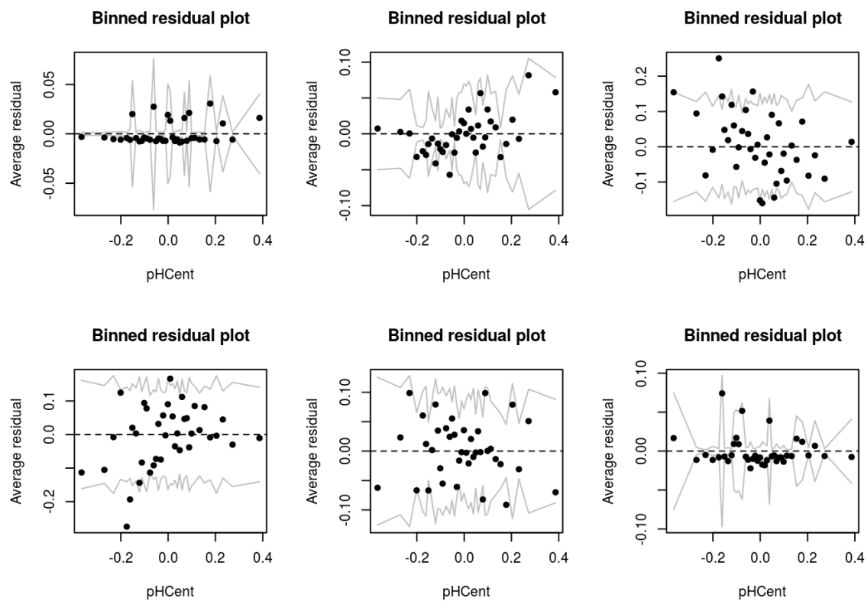
The binned residual plots of volatile.acidityCent also satisfies assumptions fairly well.



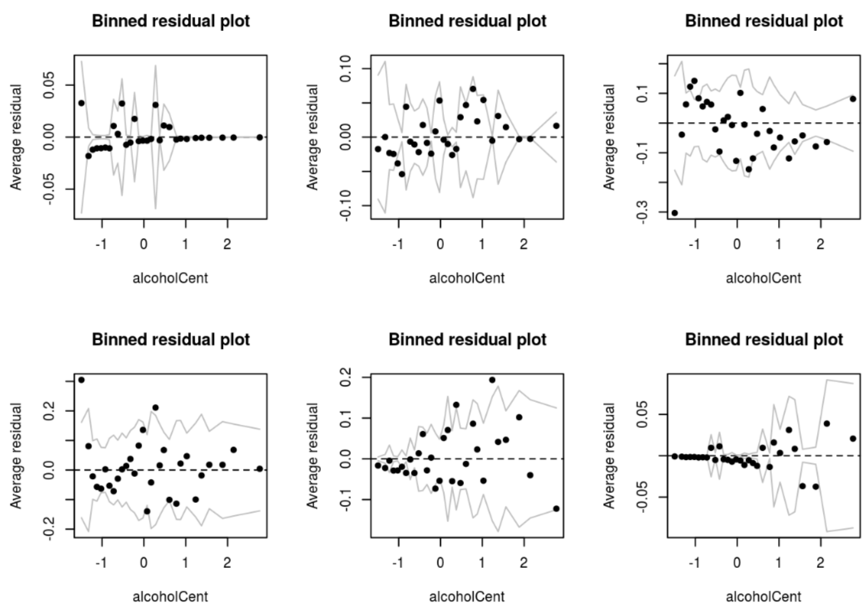
The binned residual plots of chloridesCent has an obvious outlier on the right for all six quality levels. Other than that, the plots contain random patterns and nothing alarming.



The log transformation of logsulphates has slightly improved its binned residual plots. But in the binned residual plot of quality at or below level 7 or otherwise still show potential linear trend.



Some of the binned residual plots with pHCent sugget potential linear trend, but the non-randomness is not very obvious, thus not too concerning.



The binned residual plots of alcoholCent for wines with quality at or below 6 appear to have a nonrandom pattern. The plot for wines with quality at or below 7 indicates strong fanning patterns.

names	x
volatile.acidityCent	6.310810
citric.acid	1.314155
chloridesCent	6.902109
logsulphates	1.574871
pHCent	1.484354
alcoholCent	1.618167

VIF measures is still below 10 for all variables, so no concerning multicollinearity in the model.

Confusion Matrix

pred.comp	3	4	5	6	7	8
4	1	NA	NA	NA	NA	NA
5	8	41	513	219	11	NA
6	1	12	164	389	139	11
7	NA	NA	3	30	49	7
8	NA	NA	1	NA	NA	NA

```r
# misclassification rate
(10+53+168+249+150+18)/1599
```

```
## [1] 0.4052533
```

The misclassification rate of the model is around 40.5%. We can see that the model does not predict any quality at level 3, most likely because there are only 10 such observations in the data set. The model predicts quality level 5 and 6 pretty well. We also remember from the exploratory data analysis, most of the observations are around quality level 5 and 6. Overall, the model is pretty good.

```
## # A tibble: 11 x 5
##    term                 estimate std.error statistic   p.value
##    <chr>                   <dbl>     <dbl>     <dbl>     <dbl>
##  1 volatile.acidityCent   -3.47     0.379      -9.16 5.13e- 20
##  2 citric.acid            -0.910    0.377      -2.41 1.59e-  2
##  3 chloridesCent          -5.61     1.29       -4.36 1.30e-  5
##  4 logsulphates            2.50     0.267       9.35 8.42e- 21
##  5 pHCent                 -1.68     0.416      -4.03 5.52e-  5
##  6 alcoholCent             0.943    0.0581     16.2  2.28e- 59
##  7 3|4                    -7.31     0.378     -19.4  1.94e- 83
##  8 4|5                    -5.35     0.235     -22.8  6.92e-115
##  9 5|6                    -1.65     0.178      -9.32 1.18e- 20
## 10 6|7                     1.19     0.179       6.67 2.55e- 11
## 11 7|8                     4.23     0.289      14.6  2.47e- 48
```

The p-values of all the variables are extremely small, so they are all significant predictors of the log-odds of the wine falling in or below quality j. We should beware to not extrapolate beyond quality j=3,4,…,7.