Additional Work

Bob Ding, Lynn Fan, Alice Jiang

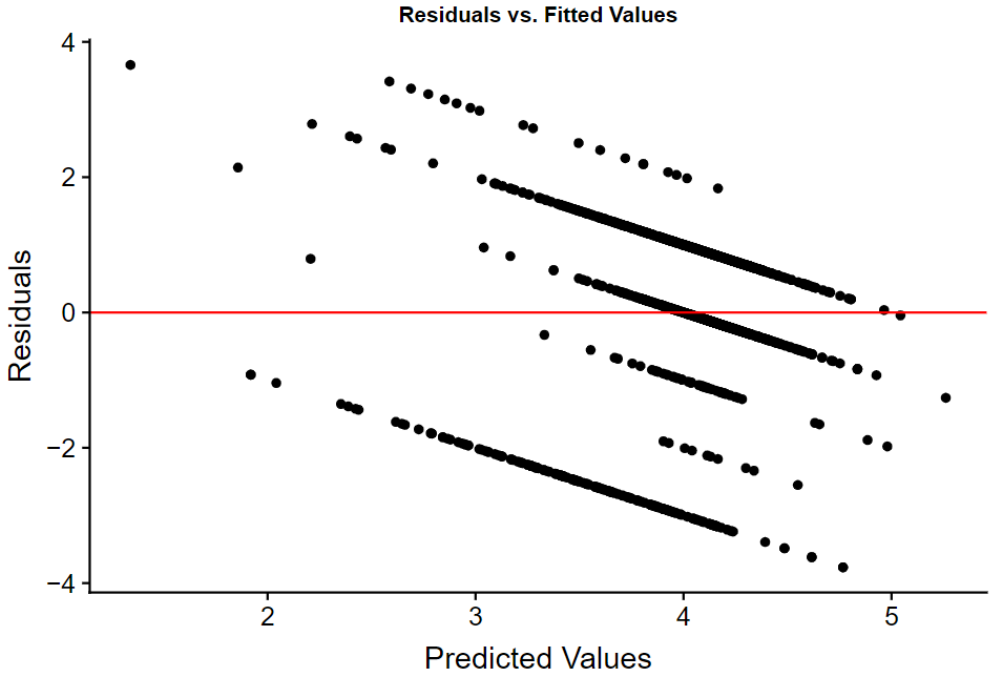
12/13/2018

Multiple Regression Analysis

Our final multiple reression model:
$quality = 4.2245 - 0.0512 * residual.sugarCent - 0.6498 * citric.acid + 0.0204 * free.sulfur.dioxideCent - 0.1836 * sulphatesCent + 57.4648 * densityCent - 0.1355 * alcoholCent - 0.0438 * citric.acid * free.sulfur.dioxideCent - 1.9995 * citric.acid * sulphatesCent + 0.0103 * free.sulfur.dioxideCent * alcoholCent - 1.3477 * sulphatesCent * alcoholCent$

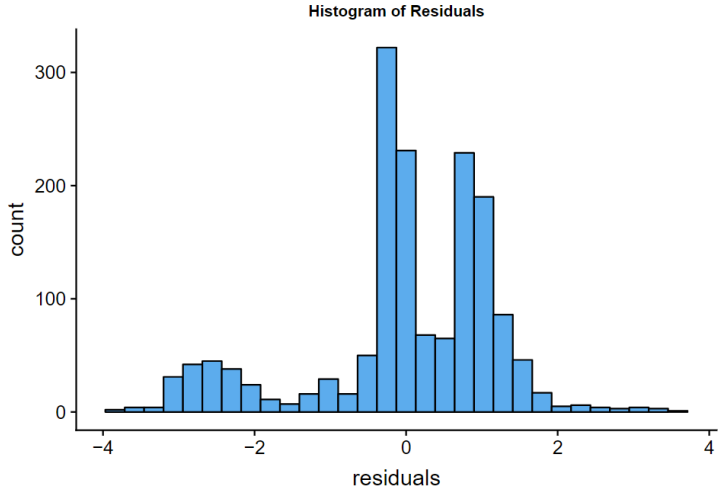
Assumptions

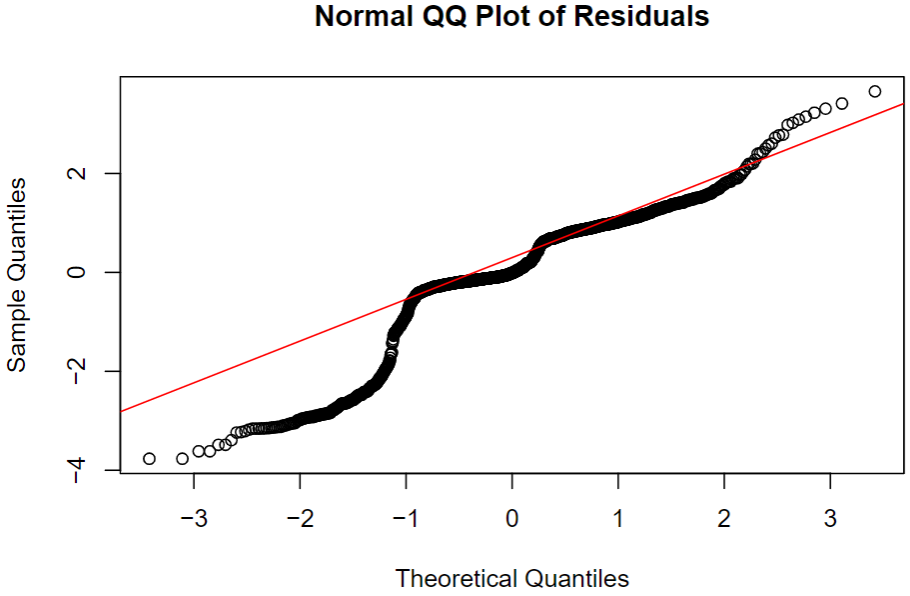
Residual Plots



This residual plot is a result of having categorical response variable.

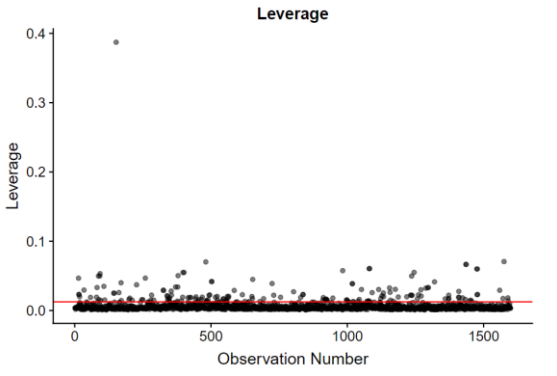
Normality



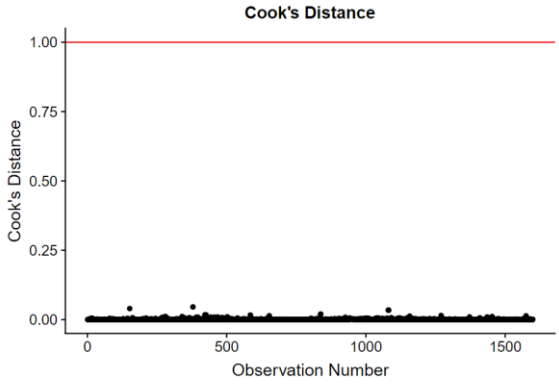


The distribution of residuals appear to be bimodal. The QQ plot also suggests the same conclusion, since we can see a very prominent deviation from the diagonal normal line on the left side. Overall, the Normaltiy Assumption seems to be violated. This could be due to the fact that we are dealing with a categorical response varaible quality.

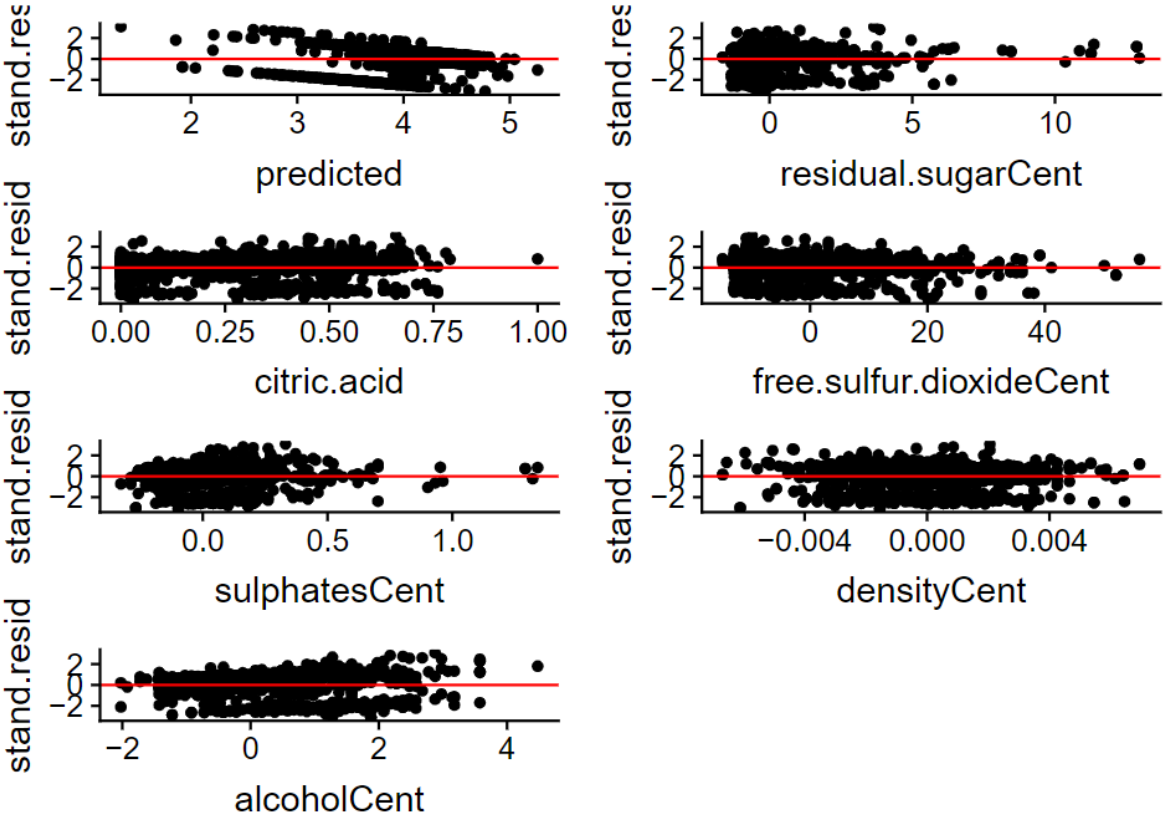
Influence Point



We can see that there is one point with a significantly high leverage around 0.4, comparing with other observations. This could be an outlier and might be an influence point.



The Cook's Distance for all observations are far below the threshold of 1.



The standardized residuals show some points with magnitude greater than 2, but overall, in combination with our observation from Cook's Distance and just one data point with high leverage away from other points, we can conclude there isn't any obvious influential points in this model.

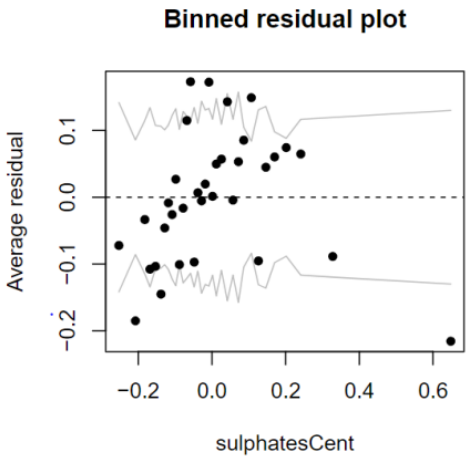
names	x
residual.sugarCent	1.402748
citric.acid	1.425081
free.sulfur.dioxideCent	3.169724
sulphatesCent	3.922514
densityCent	2.274231
alcoholCent	1.825881
citric.acid:free.sulfur.dioxideCent	3.249458
citric.acid:sulphatesCent	3.832337
free.sulfur.dioxideCent:alcoholCent	1.097922
sulphatesCent:alcoholCent	1.187895

Overall, the final model does not have any major multicollinearity concerns – all VIF values are fairly low and below the threshold of 10. However, multiple regression model does not seem to be the best fitting for our data set. Given we have a categorical response variable, we will try different types of logistic regression models.

Logistic Regression with sulphatesCent

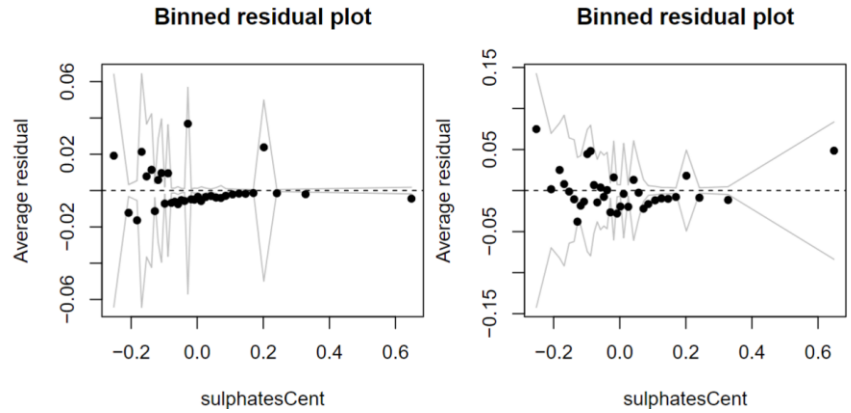
term	estimate	std.error	statistic	p.value
(Intercept)	0.4951634	0.1207867	4.099486	0.0000414
alcoholCent	0.9510158	0.0709582	13.402476	0.0000000
volatile.acidityCent	-3.5486411	0.4515594	-7.858636	0.0000000
sulphatesCent	2.7030432	0.4463930	6.055299	0.0000000
chloridesCent	-3.7700684	1.4668058	-2.570257	0.0101623
free.sulfur.dioxideCent	-0.0132536	0.0058580	-2.262476	0.0236680
citric.acid	-0.8755285	0.3918921	-2.234106	0.0254761

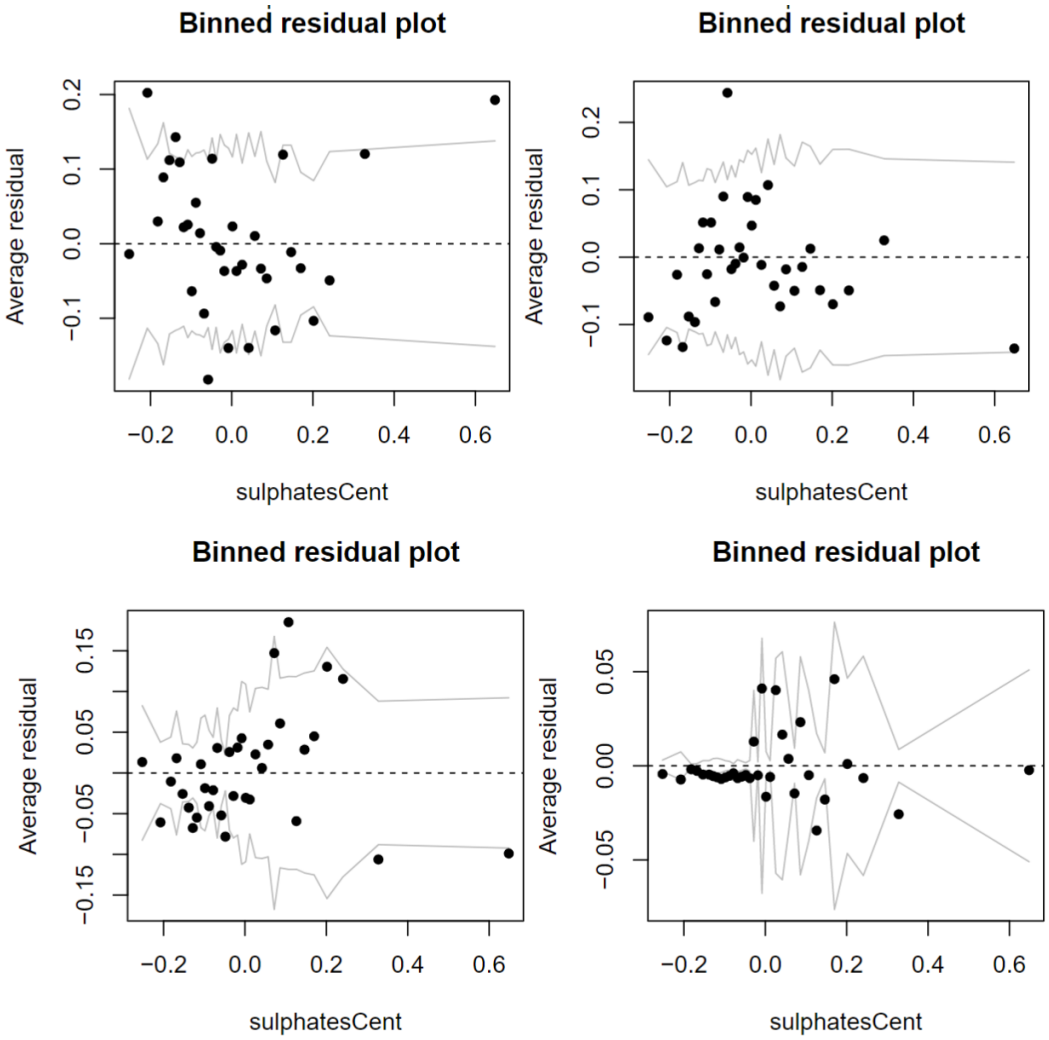
All the variables have p-values much smaller than 0.05, so there is significant evidence that they are important predictors of the log-odds (and therefore odds) of wine quality (good vs. not good). Based on the model, it appears that the percent alcohol content of the wine is the strongest predictor of wine quality. alcoholCent has the largest test statistic magnitude of 13.402476. The positive test statistic value also shows that as the % alcohol content of the wine increases, the logs-odds of good vs. not good wine quality increases. The amount of acetic acid in wine (volatile.acidityCent) and the amount of wine additive (sulphatesCent) are also strong predictors.



The binned residual plot for sulphatesCent appears to show some linear trend. We will try log transformation.

Ordinal Logistic Regression with SulphatesCent





Some of the binned residual plots suggest potential linear pattern. There also appears to be outliers. To improve the model, we will log transform sulphatesCent.