Lab 2: What Makes a Product Successful?

Ahmad Azizi, Jon Hand, Mickey Hua, Prashant Dhingra

Introduction

Making a better wine can lead to more awards, more brand recognition, and ultimately more sales. Creating a quality wine can be tricky though, as many different factors in the vinification process can influence the end product. The climate grapes are grown, the type of grapes that are used, and how long the grapes are fermented all contribute to the ultimate flavor and quality of a wine. Additionally, classifying a wine as tasting "good" is a difficult task, since the quality of taste is not something that can be measured with any instrument.

Despite these challenges, many different physicochemical properties that can be measured in a wine. We intend to use these objective chemical measurements to explain which chemicals influence the quality of the Portuguese "Vinho Verde" wine. To operationalize the concept of "quality" we will be using a variable from our dataset which is a quality rating given by wine experts in a blind taste test. These ratings are on a 0 (very bad) to 10 (very excellent) scale, based on sensory data made by wine experts with a median of at least 3 evaluations per wine. Additional chemical components contained in the wine are measured using physicochemical tests. Chemicals measured in the dataset include fixed acidity; volatile acidity; citric acid; residual sugar; chlorides; free sulfur dioxide; total sulfur dioxide; density; pH; sulphates; alcohol.

By performing regression analysis, we seek to answer the following research question:

> How does the percentage of alcohol in decently rated Portuguese wine affect its quality?

Data and Research Design

Our data source for analysis is the "Wine Quality" dataset from the UC Irvine Machine Learning Data Repository. We utilized two data sets of observations from this source, one which contains data on red wines and one on white wines, to arrive at 6,497 distinct observations. While combining the data sources, we also created a categorical variable called "wine type" which indicates if the observation is red or white wine.

Before we start our exploration, it is important to define the variables. "Table 1" shows all the variables of our data along with the first few rows of our data. From our limited understanding of chemistry, the variables are defined as follows:

- *alcohol* - the percent alcohol content of the wine
- *citric_acid* - weak organic acid that occurs naturally in citrus fruits and can add 'freshness' and flavor to wines
- *chlorides* - the amount of salt in the wine
- *density* - self explanatory
- *fixed_acidity* - acids involved with wine that are fixed (don't evaporate readily)

Table 1: Summary Table of Wine Dataset

fixed_acidity	volat_acidity	citric_acid	resid_sugar	chlorides	free_SO2	total_SO2	density	pH	sulphates	alcohol	quality	wine_type
7.5	0.50	0.36	6.1	0.071	17	102	0.9978	3.35	0.80	10.5	5	red
6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5	red
7.5	0.50	0.36	6.1	0.071	17	102	0.9978	3.35	0.80	10.5	5	red
7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5	red
8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.30	0.75	10.5	7	red
8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5	red

- *free_sulfur_dioxide* -free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulphate ion. It exhibits both germicidal and antioxidant properties
- *volatile_acidity* - the amount free of acetic acid in wine, which at high levels can lead to an unpleasant, vinegar taste
- *residual_sugar* - refers to any natural grape sugars that are leftover after fermentation stops. it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
- *total_sulfur_dioxide* - amount of free and bound forms of S02
- *pH* - from a winemaker's point of view, it is a way to measure ripeness in relation to acidity
- *sulphates* - a wine additive that can contribute to sulfur dioxide gas (S02) levels. It acts as an antimicrobial and antioxidant
- *quality* - output variable

Data Exploration

As mentioned in the 'description of data' section, we have 1875 data points across 13 columns. Since our dataset is large enough, we split our dataset into an exploration set and a testing set. The exploration set has 30% and the testing set has 70% of our data. We will use the exploration set to get an understanding of our data, build intuition and explore how the data is distributed. The testing set would be used to build our models.

Out of 6497 rows in the dataset, 6251 are clean rows. We used data 1875 for exploration and 4376 for testing/model.

Next, let's take a look at the summary of our data. As shown in Table 2 we can observe some properties as follows

- Residual sugar maximum value is 22.60 g/dmˆ3. So, there are no sweet wines in this data set
- There is a huge dispersion of Total SO2 from 6 to 290.
- Alcohol level varies from 8.4 to 14.05 %
- pH range 2.77 - 4.01
- quality minimum value is 5 and the maximum is 9 (refer to Limitations of Model section for an explanation)

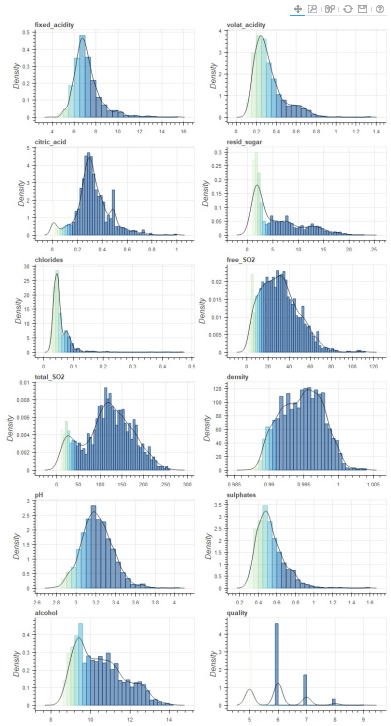
Next, we wanted to see how each variable is distributed, particularly our outcome variable. From figure 1, we can see observe the following.

- Most of our variables have a roughly normal distribution
- Most observations have values that are equal to greater than 5.
- Residual sugar and chlorides are skewed to the left.

Table 2: Summary Table of Wine Dataset

	fixed_acidity	volat_acidity	citric_acid	resid_sugar	chlorides	free_SO2	total_SO2	density	pH	sulphates	alcohol	quality	wine_type
	Min. : 3.900	Min. :0.0800	Min. :0.0000	Min. : 0.700	Min. :0.01400	Min. : 1.00	Min. : 6.0	Min. :0.9871	Min. :2.770	Min. :0.2500	Min. : 8.40	Min. :5.000	Length:1875
	1st Qu.: 6.400	1st Qu.:0.2200	1st Qu.:0.2500	1st Qu.: 1.900	1st Qu.:0.03800	1st Qu.: 18.00	1st Qu.: 80.0	1st Qu.:0.9924	1st Qu.:3.110	1st Qu.:0.4300	1st Qu.: 9.50	1st Qu.:5.000	Class :character
	Median : 7.000	Median :0.2900	Median :0.3100	Median : 3.500	Median :0.04700	Median : 30.00	Median :119.0	Median :0.9949	Median :3.200	Median :0.5000	Median :10.30	Median :6.000	Mode :character
	Mean : 7.204	Mean :0.3281	Mean :0.3223	Mean : 5.597	Mean :0.05556	Mean : 31.33	Mean :116.7	Mean :0.9947	Mean :3.213	Mean :0.5323	Mean :10.50	Mean :5.911	NA
	3rd Qu.: 7.700	3rd Qu.:0.3900	3rd Qu.:0.3900	3rd Qu.: 8.300	3rd Qu.:0.06300	3rd Qu.: 42.00	3rd Qu.:155.0	3rd Qu.:0.9970	3rd Qu.:3.310	3rd Qu.:0.6100	3rd Qu.:11.30	3rd Qu.:6.000	NA
	Max. :15.500	Max. :1.3300	Max. :1.0000	Max. :22.600	Max. :0.46400	Max. :112.00	Max. :260.0	Max. :1.0037	Max. :4.010	Max. :1.5900	Max. :14.05	Max. :9.000	NA

- Sulphates mainly lie between 0.25 and 0.7
- Most data has residual sugar less than 5.

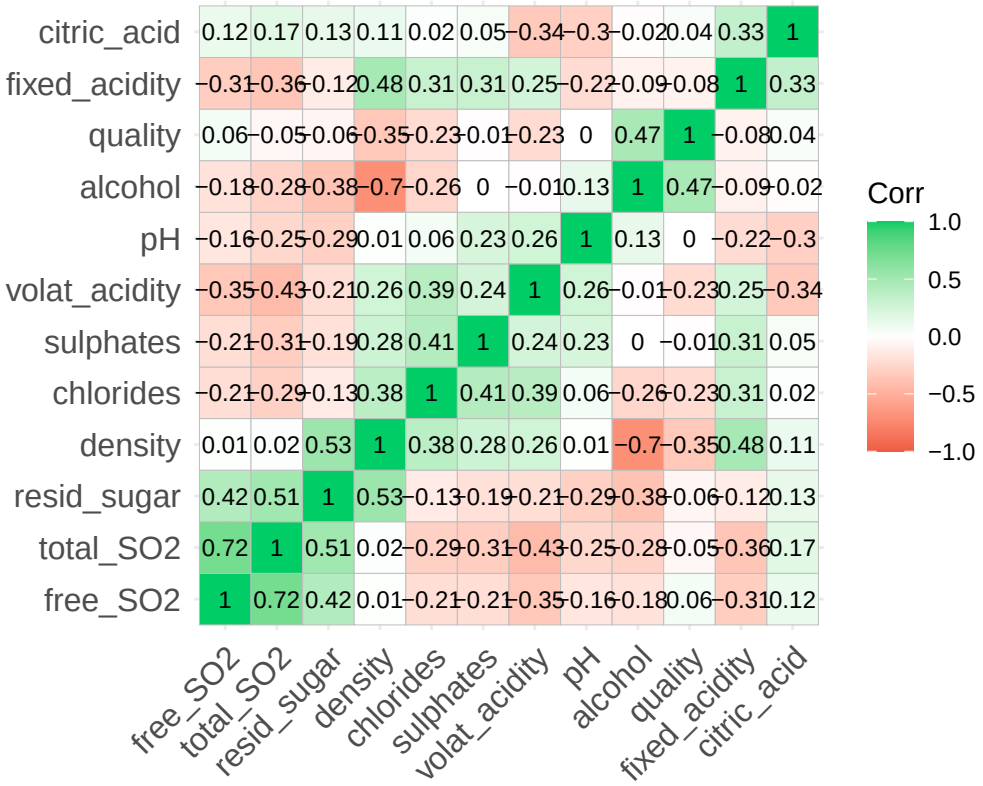


It would be also helpful to understand the correlations of our variables with each other. Figure 2 shows a correlation plot and we can observe the following.

- Quality has the highest correlation with alcohol (0.47)
- Alcohol has a very high correlation with density (-0.70)
- Free SO2 has a very high correlation with total SO2(0.72)

Since quality is the outcome variable, we wanted to see how each variable is correlated with quality with exception of wine type. From figure 3 and (optional table 3), we can see that alcohol and density have the highest correlation with quality; density has a negative correlation while alcohol has a high correlation. This information would be important when we are building our model.

Figure 2: Correlation of Variables



Now we draw detail plot to show relationship of quality (outcome variable y) with individual variables e.g. alcohol, sulphates, pH, density, total_SO2.

Figure 3: Box plot to show quality with each variable

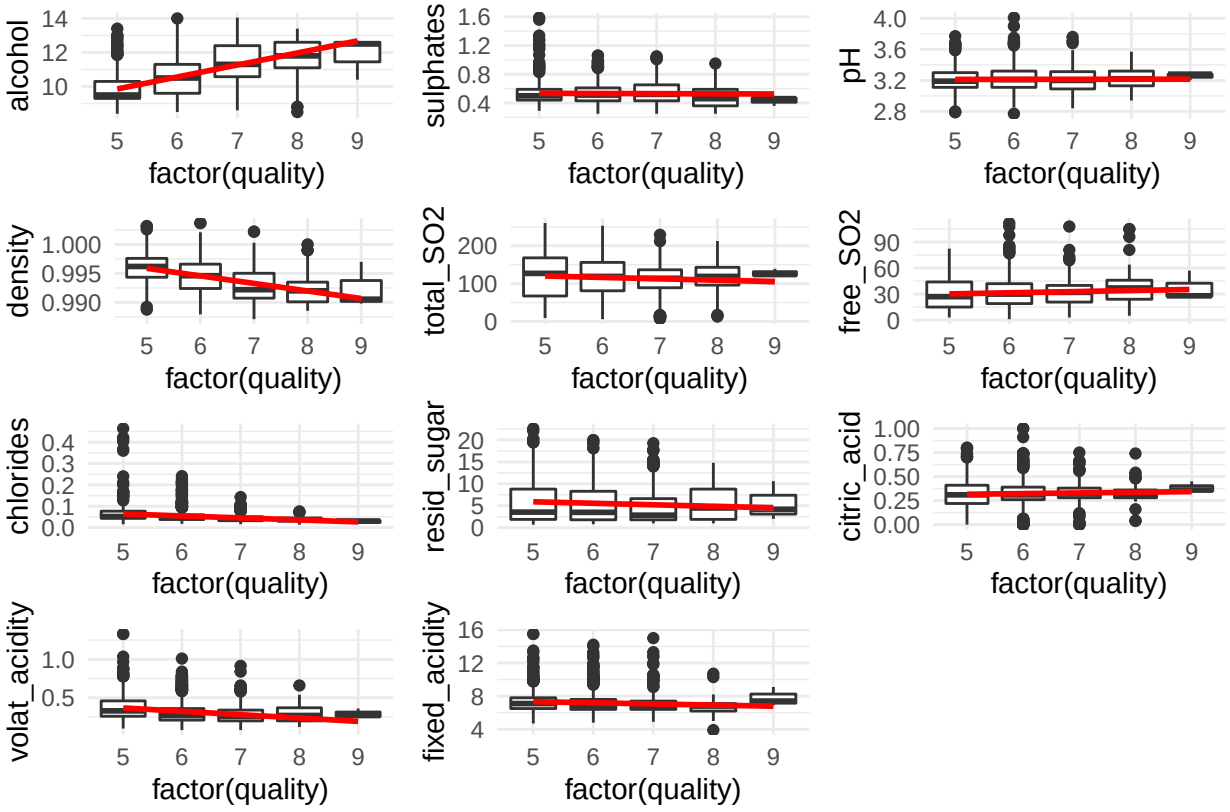


Figure 3 above shows how individual variable relates to Quality e.g. Quality increases with an increase in Alcohol. Quality increase with a decrease in density.

Model building

Model 1

$$Quality = \beta_0 + \beta_1 \cdot alcohol$$

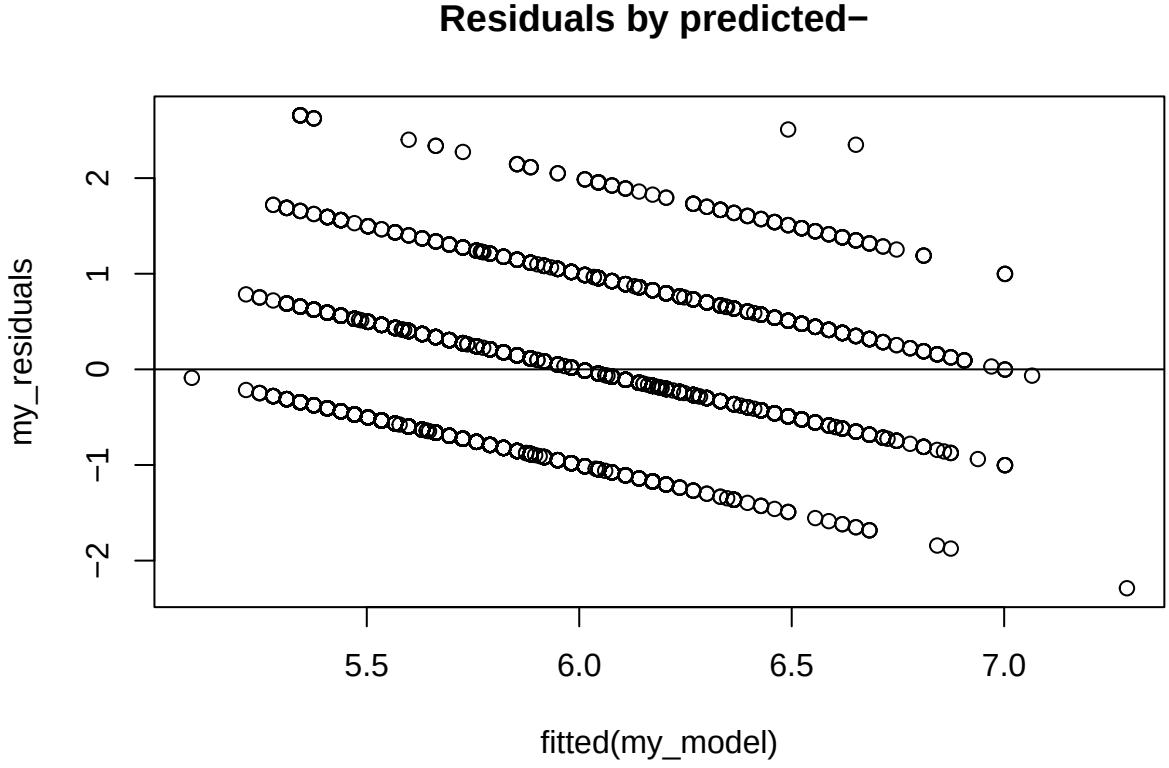
β_0 quality rating the variable gets with no additional input. β_1 would be the alcohol content coefficient.

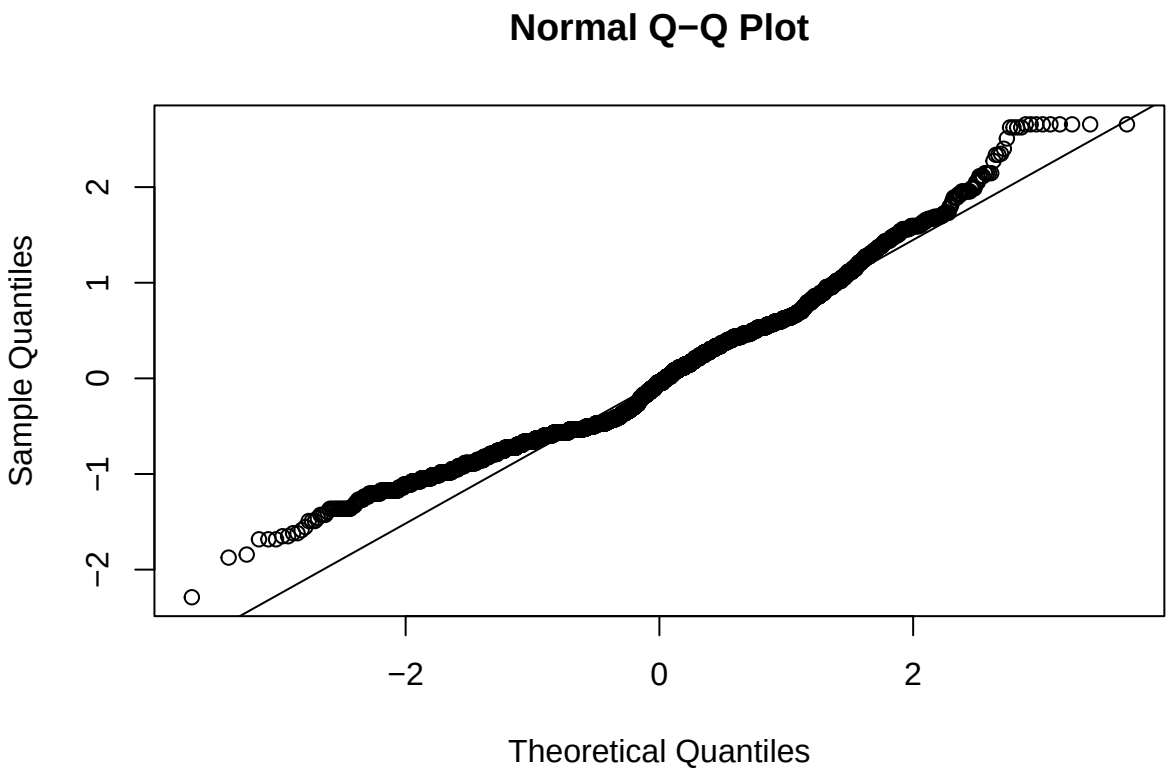
For our first model, we only included the alcohol,

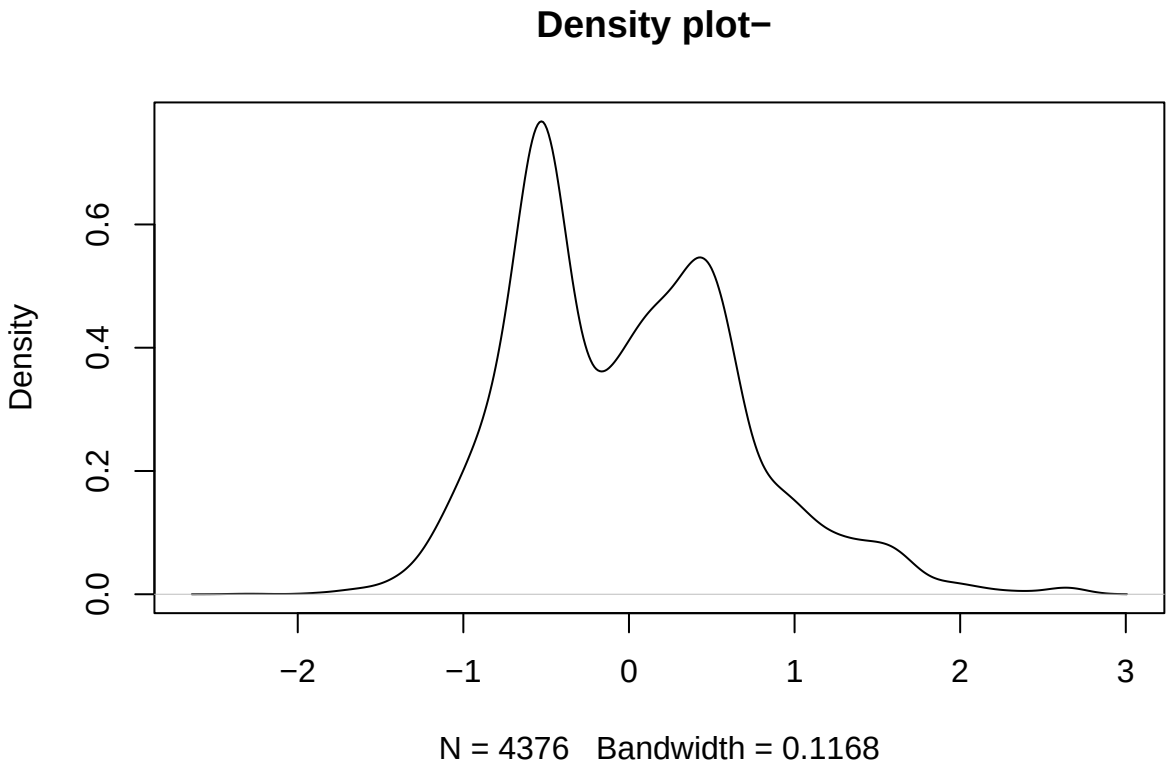
Null Hypothesis : Alcohol content does not impact quality. Alternative Hypothesis : Alcohol content impacts quality.

```
##
## Call:
## lm(formula = quality ~ alcohol, data = test_data_wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.28892 -0.53484 -0.04512  0.46516  2.65652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.536950   0.092817   27.33   <2e-16 ***
```

```
## alcohol     0.318924   0.008777   36.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6944 on 4374 degrees of freedom
## Multiple R-squared:  0.2319, Adjusted R-squared:  0.2317
## F-statistic:  1320 on 1 and 4374 DF,  p-value: < 2.2e-16
##
## [1] "MSE:-0.482007910937604"
```




Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−2
0
2



After running our linear regression, we can see that the p-value is significant and we can reject the null hypothesis that alcohol level does not effect quality. For each alcohol percent increase, we get roughly a 0.32 increase in quality rating. The distribution of residuals seems "ok" and could potentially pass as symmetrical.
The standard error seems to be very low which is great for our model. In other words, the increase in alcohol quality by an increase in alcohol percent varies by 0.008777, which seems good. Adjusted R-squared is about 0.2317. This means that roughly 23% variance found in the outcome variable(quality) can be explained by the predictable variable(alcohol).

Let's see if we can improve our model by introducing additional variables.

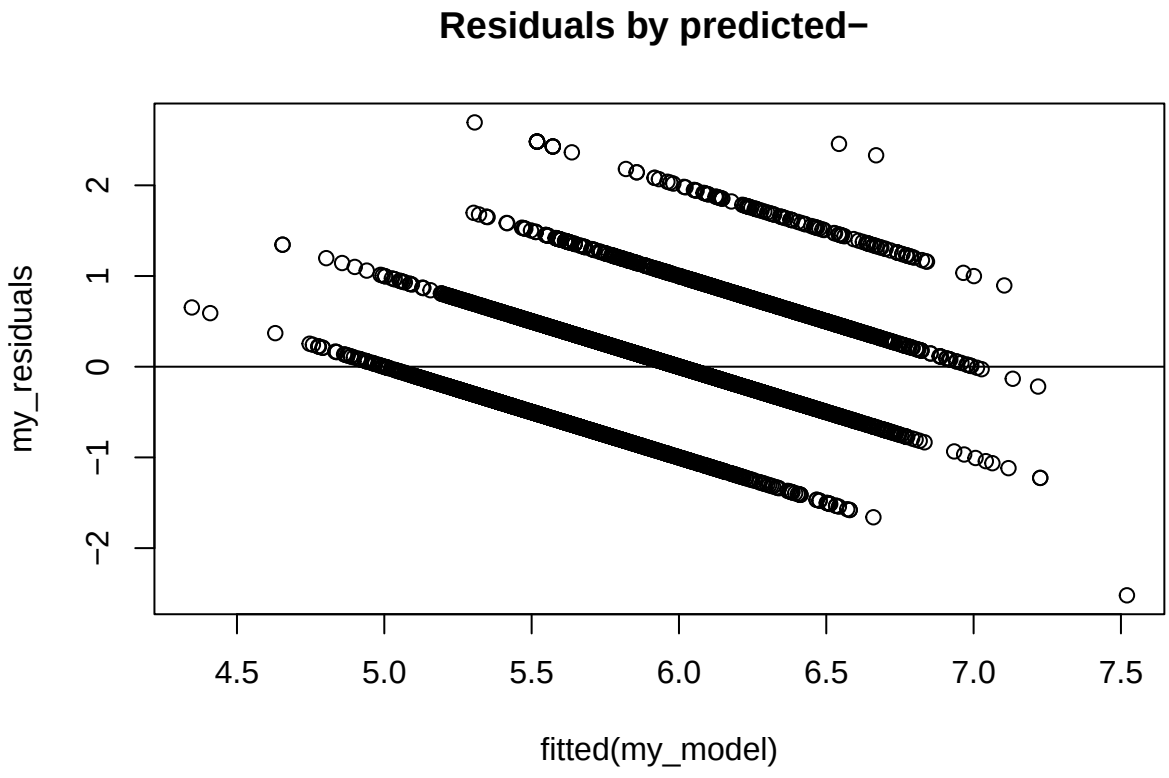
Model 2

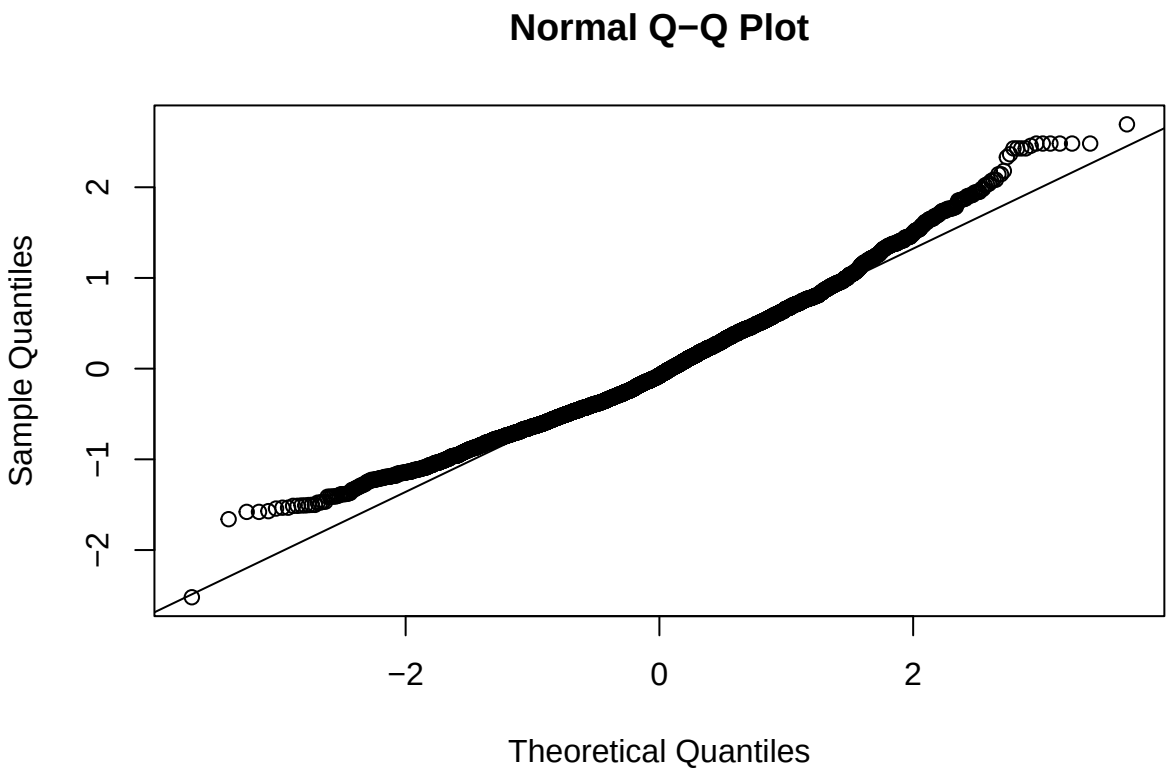
$Quality = \beta_0 + \beta_1 \cdot alcohol + \beta_2 \cdot wine.type + \beta_3 \cdot fixed.acidity + \beta_4 \cdot volatile.acidity + \beta_5 \cdot citric.acid + \beta_6 \cdot residual_sugar + \beta_7 \cdot chlorid$

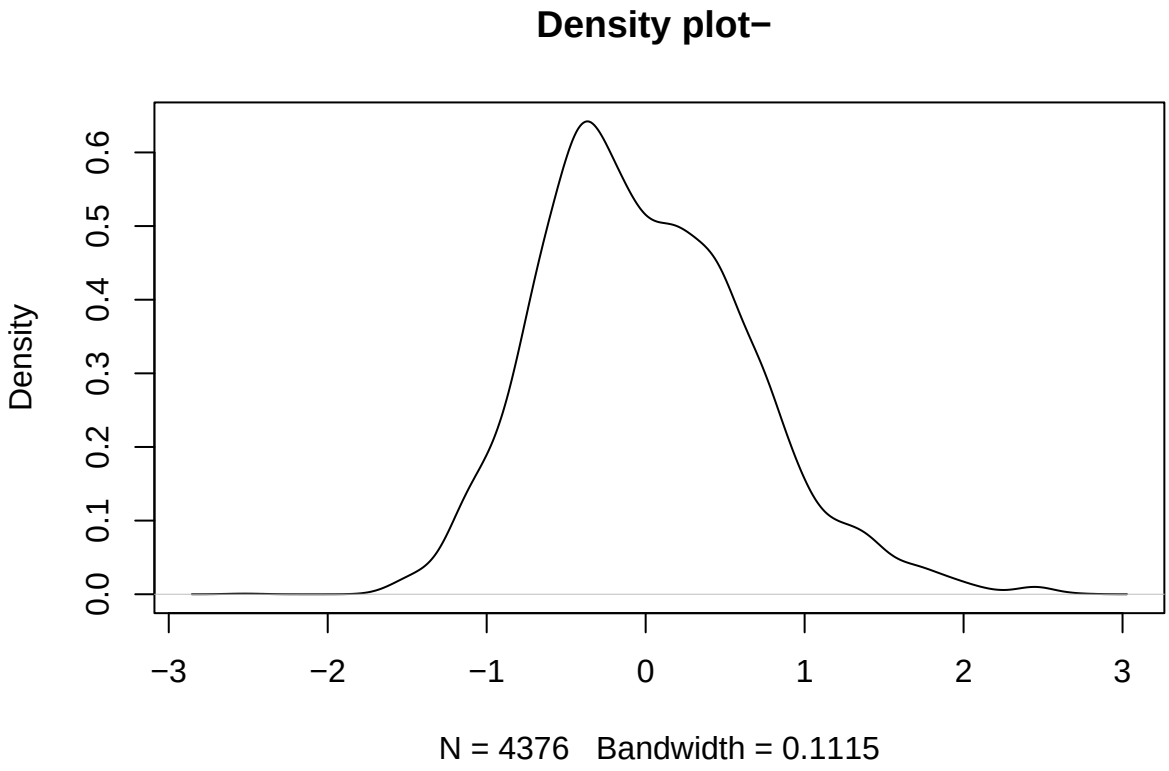
For model 2, we decided to include all variables, except density and free_SO2. We excluded density because density has a very high correlation with alcohol and we excluded free_SO2 because free_SO2 has a very high correlation with total_SO2. In both cases, we wanted to avoid co-linearity.

```
##
## Call:
## lm(formula = quality ~ alcohol + wine_type + fixed_acidity +
##     volat_acidity + citric_acid + resid_sugar + chlorides + total_SO2 +
##     pH + sulphates, data = test_data_wine)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.52040 -0.47019 -0.07748  0.43423  2.69335
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.7363979  0.3372925   5.148 2.75e-07 ***
## alcohol         0.3193962  0.0101605  31.435  < 2e-16 ***
## wine_typewhite  0.0296254  0.0530817   0.558 0.576798
## fixed_acidity   0.0131259  0.0114127   1.150 0.250163
## volat_acidity  -1.2038075  0.0921194 -13.068  < 2e-16 ***
## citric_acid    -0.0488960  0.0875092  -0.559 0.576359
## resid_sugar     0.0205011  0.0025487   8.044 1.11e-15 ***
## chlorides      -0.8157430  0.3682560  -2.215 0.026801 *
## total_SO2      -0.0010005  0.0002899  -3.451 0.000564 ***
## pH              0.2564570  0.0785394   3.265 0.001102 **
## sulphates       0.6104648  0.0811947   7.519 6.68e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6633 on 4365 degrees of freedom
## Multiple R-squared:  0.3007, Adjusted R-squared:  0.2991
## F-statistic: 187.7 on 10 and 4365 DF,  p-value: < 2.2e-16
##
## [1] "MSE:-0.438797738557331"
```




Normal Q−Q Plot
Sample Quantiles
−2
−1
0
1
2
−2
0
2
Theoretical Quantiles



After running our linear regression, we can see that the p-value is significant for all variables, except wine type, fixed acidity, and citric acid. In other words, wine_type, citric and acid, and fixed acidity fail to reject the null hypothesis that these three variables do not affect quality.

The distribution of residuals has improved and could potentially pass as symmetrical. The standard errors(SDE), as expected, vary for each variable. SDE seems to have increased for alcohol, while the coefficient has increased a bit for alcohol, which is good. Volatile acidity seems to have the highest coefficient and affects the outcome variable the most. Adjusted R-squared is about 0.0.2991. This means that roughly 30% variance found in the outcome variable(quality) can be explained by the predictable variables. This could be attributed to the introduction of more variables.

Tweaking our model a bit will make it better.

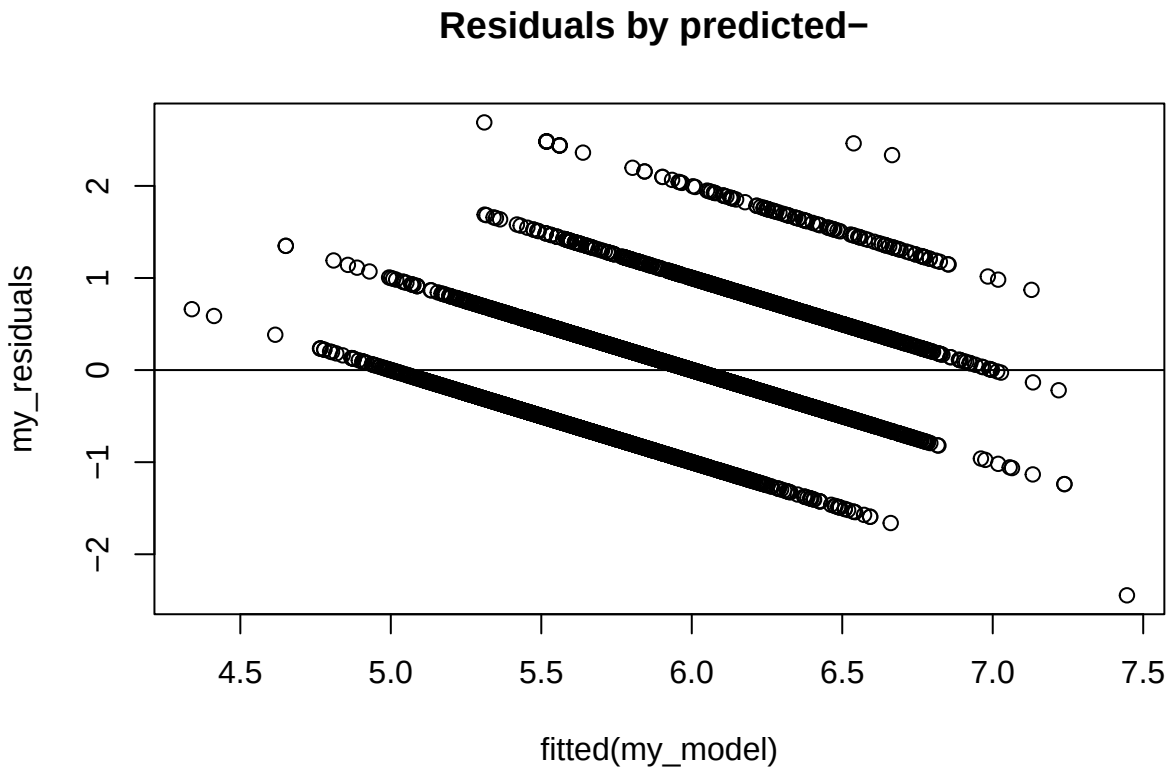
Model 3

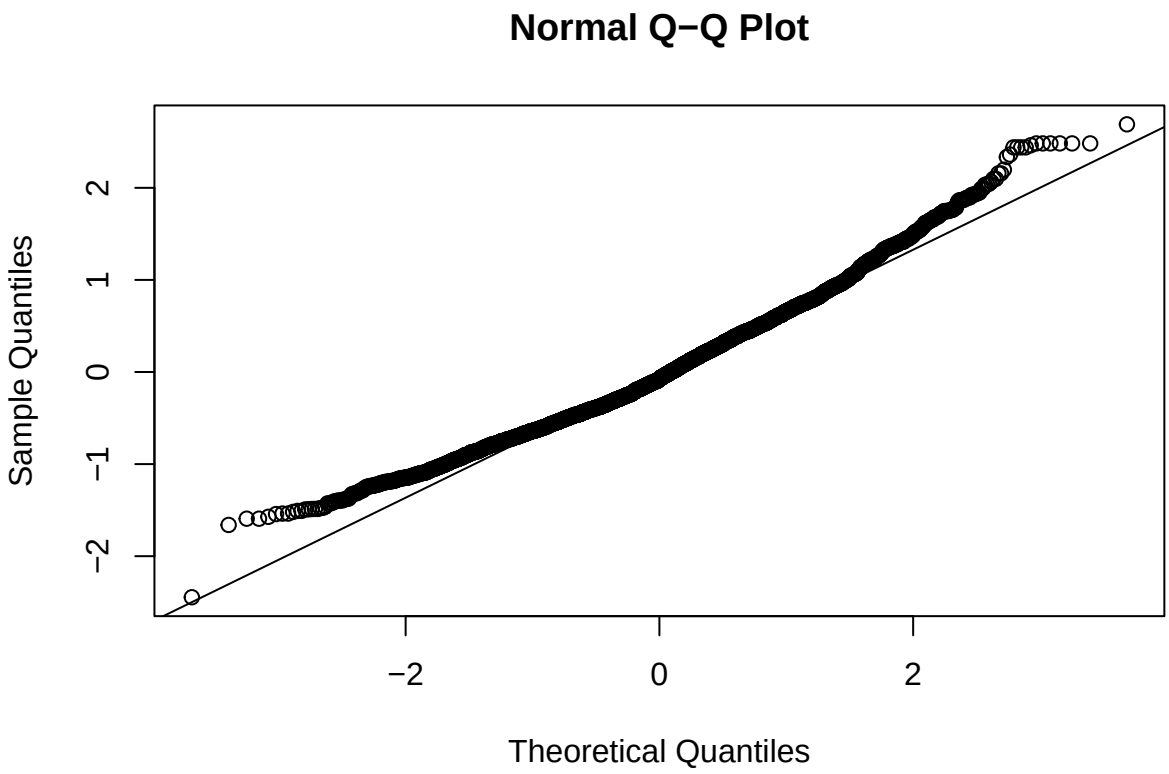
$Quality = \beta_0 + \beta_1 \cdot alcohol + \beta_2 \cdot wine.type + \beta_3 \cdot volatile.acidity + \beta_4 \cdot residual_sugar + \beta_5 \cdot chlorides + \beta_6 \cdot total.SO2 + \beta_7 \cdot pH + \beta_8 \cdot sulp$

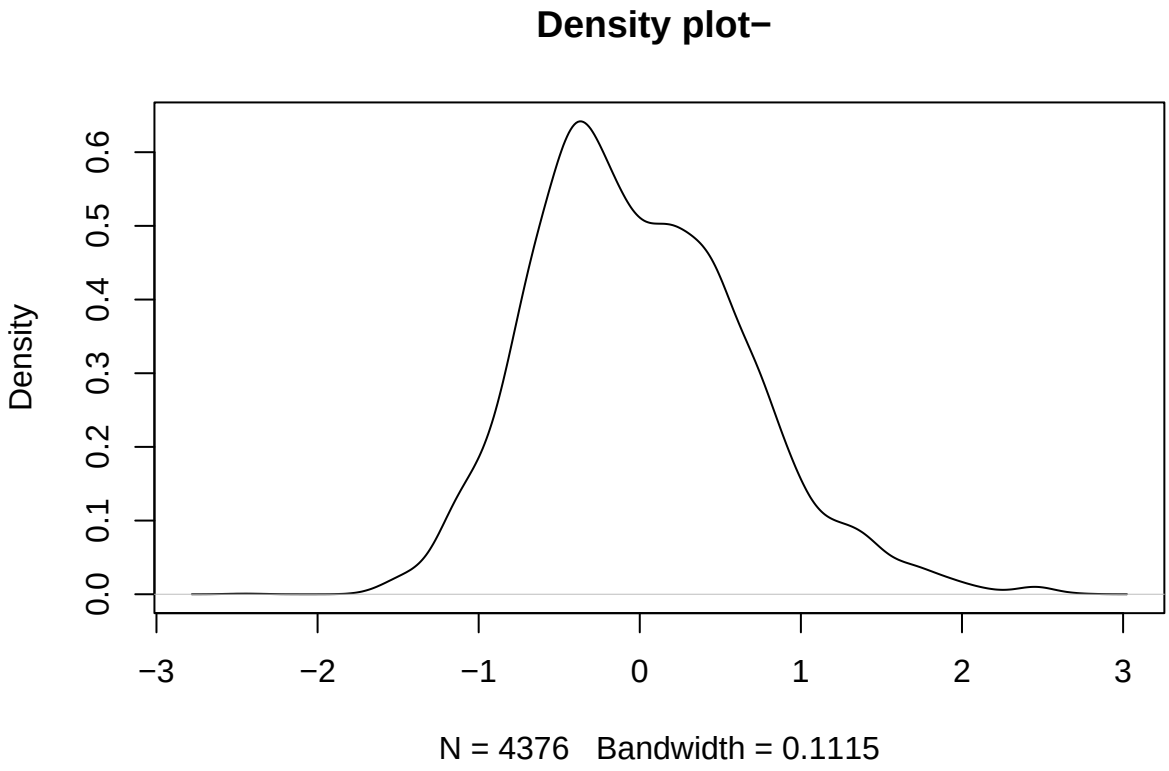
For model 3, we decided to remove variables that failed to reject the null, except wine type. In other words, we removed fixed acidity and citric acid.

```
##
## Call:
## lm(formula = quality ~ alcohol + wine_type + volat_acidity +
```

```
##     resid_sugar + chlorides + total_SO2 + pH + sulphates, data = test_data_wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44594 -0.47253 -0.07695  0.43563  2.68944
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.9637376  0.2558351   7.676 2.02e-14 ***
## alcohol         0.3184149  0.0100630  31.642  < 2e-16 ***
## wine_typewhite  0.0046395  0.0479251   0.097  0.92288
## volat_acidity  -1.1995962  0.0861738 -13.921  < 2e-16 ***
## resid_sugar     0.0203650  0.0025415   8.013 1.42e-15 ***
## chlorides      -0.8631939  0.3617531  -2.386  0.01707 *
## total_SO2      -0.0010145  0.0002879  -3.523  0.00043 ***
## pH              0.2196809  0.0681596   3.223  0.00128 **
## sulphates       0.6156907  0.0805893   7.640 2.66e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6632 on 4367 degrees of freedom
## Multiple R-squared:  0.3005, Adjusted R-squared:  0.2992
## F-statistic: 234.5 on 8 and 4367 DF,  p-value: < 2.2e-16
##
## [1] "MSE:-0.438932139429592"
```




Normal Q−Q Plot
Sample Quantiles
−2
−1
0
1
2
−2
0
2
Theoretical Quantiles



> After running our linear regression, we can see that the p-value is significant for all variables, except wine type again.

> The distribution of residuals has gotten a bit worse, but it could still pass as symmetrical The standard errors(SDE) remain roughly the same as model 2 Adjusted R-squared is about 0.0.2992. This has also stayed roughly the same as the model.

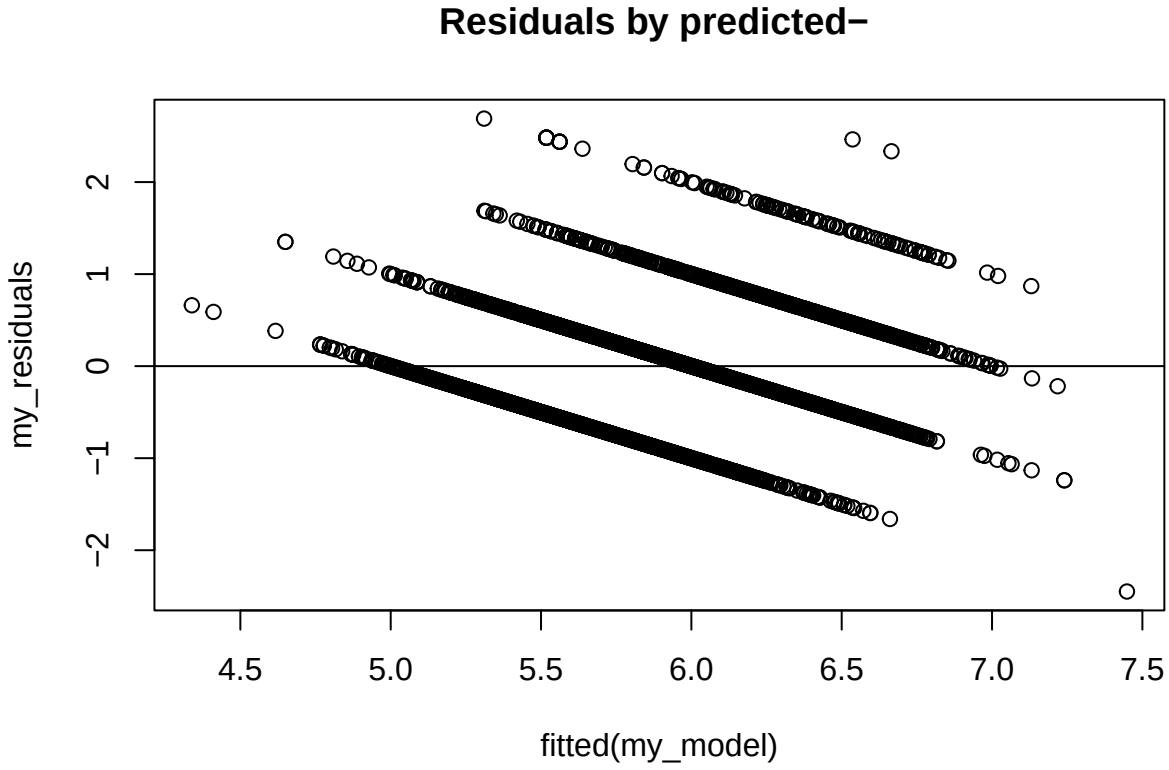
For our last model, we will remove the wine type as well because it fails to reject the null hypothesis.

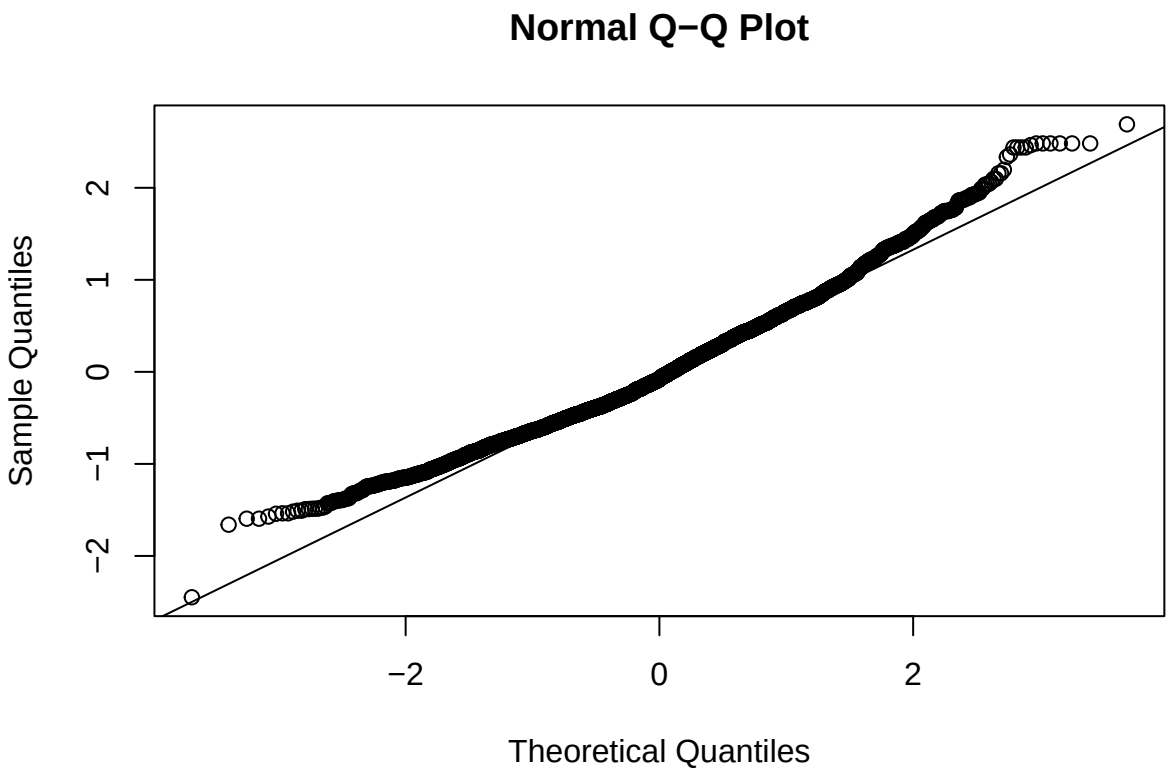
Model 4

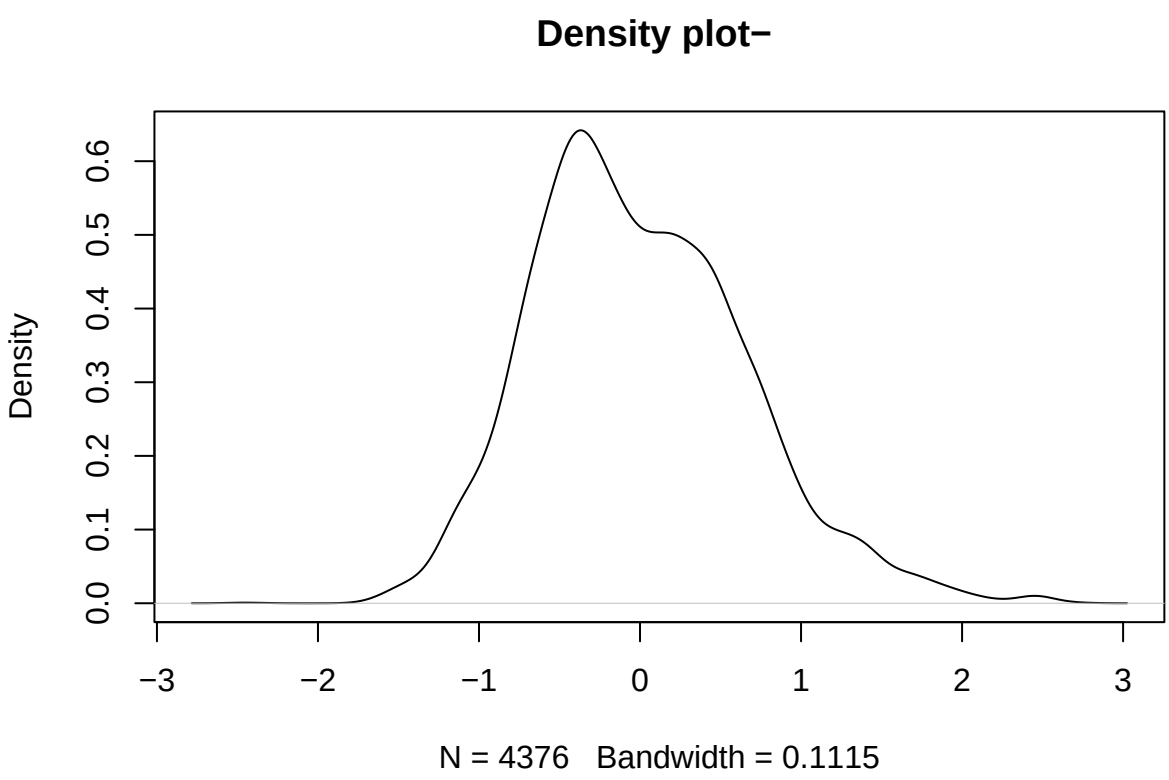
$Quality = \beta_0 + \beta_1 \cdot alcohol + \beta_2 \cdot volatile.acidity + \beta_3 \cdot residual_sugar + \beta_4 \cdot chlorides + \beta_5 \cdot total.SO2 + \beta_6 \cdot pH + \beta_7 \cdot sulphates$

```
##
## Call:
## lm(formula = quality ~ alcohol + volat_acidity + resid_sugar +
##     chlorides + total_SO2 + pH + sulphates, data = test_data_wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44870 -0.47295 -0.07662  0.43507  2.68937
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.9695772  0.2485937   7.923 2.92e-15 ***
```

```
## alcohol        0.3186398  0.0097900  32.547  < 2e-16 ***
## volat_acidity -1.2039104  0.0737469 -16.325  < 2e-16 ***
## resid_sugar    0.0203765  0.0025384   8.027 1.27e-15 ***
## chlorides     -0.8711414  0.3522746  -2.473  0.01344 *
## total_SO2     -0.0009979  0.0002311  -4.317 1.62e-05 ***
## pH             0.2186307  0.0672832   3.249  0.00117 **
## sulphates      0.6130148  0.0756918   8.099 7.13e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6631 on 4368 degrees of freedom
## Multiple R-squared:  0.3005, Adjusted R-squared:  0.2994
## F-statistic: 268.1 on 7 and 4368 DF,  p-value: < 2.2e-16
##
## [1] "MSE:-0.438933081367215"
```




Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−2
0
2



After running our linear regression, we can see that the p-value is significant for all variables, except wine type again.

The distribution of residuals remains roughly the same as model 3, but it could still pass as symmetrical The standard errors(SDE) remain roughly the same as model 3. Volatile acidity and chlorides have the highest coefficients. In other words, the amount of acetic acid in wine, which at a high level can lead to an unpleasant, vinegary taste which corresponds to volatile acidity has the highest effect on the wine. Furthermore, the level of salt in wine, which corresponds to chlorides Adjusted R-squared is about 0.2992. This has also stayed roughly the same as the previous model.

Model 4 seems to be the best out of the 4 models.

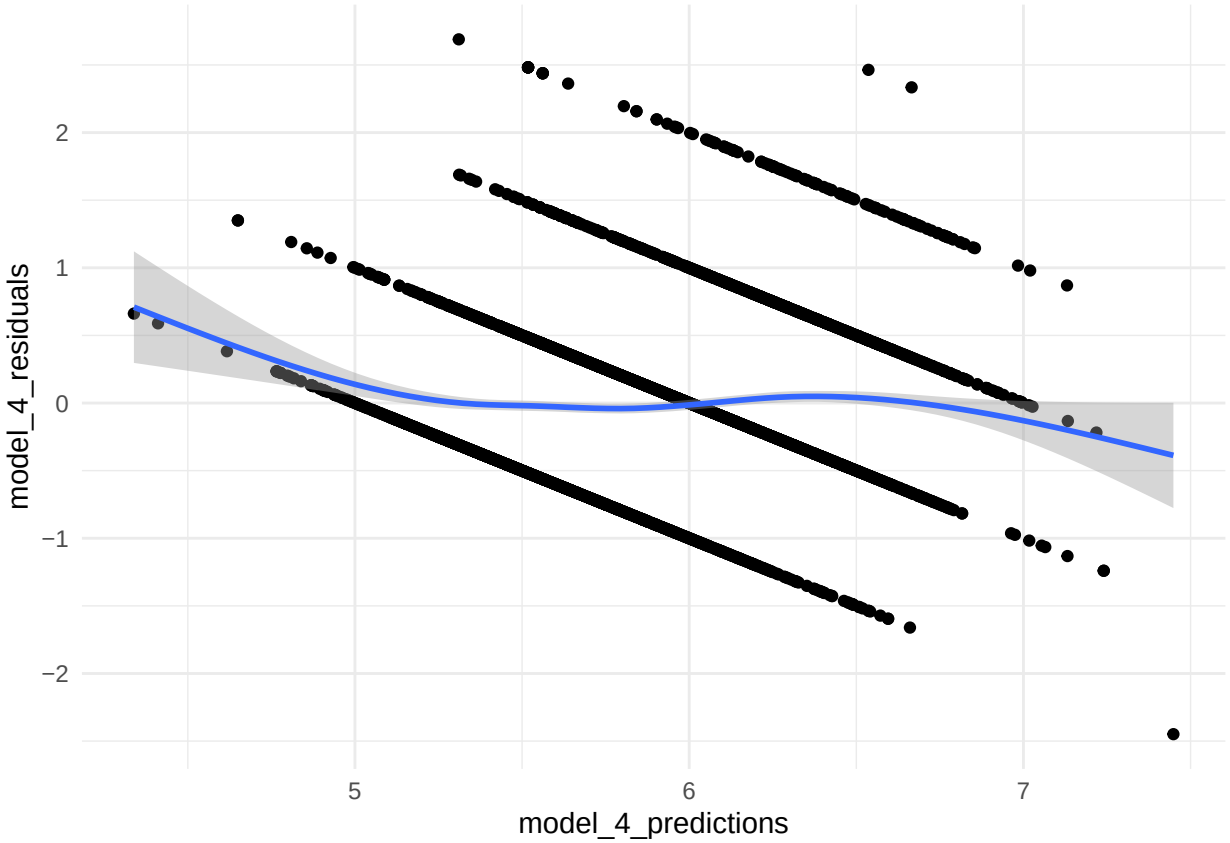
Let's study the mean of squared errors(MSE) between all 4 models to gain better insight. We can see that model 1 had the best MSE, however, the difference between the MSE is roughly the same. Figure (number) shows us model 4 prediction vs residuals. Our model is roughly linear.

```
## [1] 0.4820079
```

```
## [1] 0.4387977
```

```
## [1] 0.4389321
```

```
## [1] 0.4389331
```



Results

Wine quality refers to the factors that go into producing a wine, as well as the indicators or characteristics that tell you if the wine is of high quality.

When you know what influences and signifies wine quality, you'll be in a better position to make better wine products. The above exercise demonstrate how alcohol impacts wine quality.

Table (number) shows the summary of all 4 models next to each other.

```
##
## ================================================
##                         Dependent variable:
##                 ---------------------------------
##                               quality
##                   (1)       (2)       (3)       (4)
## ------------------------------------------------
## alcohol         0.319*** 0.319***  0.318***  0.319***
##                 (0.009)   (0.011)   (0.011)   (0.010)
##
## volat_acidity           -1.204*** -1.200*** -1.204***
##                          (0.091)   (0.084)   (0.070)
##
## resid_sugar             0.021***  0.020***  0.020***
##                          (0.003)   (0.003)   (0.003)
```

```
##
## chlorides                -0.816**  -0.863*** -0.871***
##                           (0.325)   (0.315)   (0.314)
##
## total_SO2                -0.001*** -0.001*** -0.001***
##                          (0.0003)  (0.0003)  (0.0002)
##
## pH                       0.256***  0.220***  0.219***
##                           (0.082)   (0.072)   (0.071)
##
## sulphates                0.610***  0.616***  0.613***
##                           (0.083)   (0.083)   (0.078)
##
## wine_typewhite            0.030     0.005
##                           (0.052)   (0.047)
##
## fixed_acidity             0.013
##                           (0.012)
##
## citric_acid              -0.049
##                           (0.083)
##
## Constant       2.537*** 1.736***  1.964***  1.970***
##                (0.097)   (0.354)   (0.272)   (0.263)
##
## ------------------------------------------------------
## Observations    4,376     4,376     4,376     4,376
## R2              0.232     0.301     0.300     0.300
## Adjusted R2     0.232     0.299     0.299     0.299
## ======================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

Model Limitations

Statistical Limitations

A basic limitation to our model is that we are basing the "success" of our product on an ordinal variable(quality rating). Since the quality rating is an ordinal variable in nature, the difference between ratings might not be consistent. As a result, the estimate for mean change in quality with a unit change in our inputs might not be as accurate as possible. We kept the ratings as the metric for "success" because measuring taste is subjective in nature and there was no better alternative.

Since our data has over 6,000 different wines, we can follow the large sample model assumptions. The first assumption requires that the samples are independent and identically distributed. The exact selection process for the taste test was not specified. It is possible that only decently tasting wine was selected to not subject the tasters to really disgusting tasting wine. This assumption has limitations as the wine quality only ranges from 3 to 9 with a majority of the wine ranging between 5 and 6. This means that there were no terrible wines (0-2) and most were rated at the middle of the range or higher. Because of this, our model holds no guarantees about the entire population of wines (not informative about really bad wines). Instead, it is informative of higher-rated wines (ones that are considered decent to great). To account for this, we decided to only use wines rated 5 and higher. By reframing our research question to only focus on highly rated wines, we can still model for alcohol levels in wines and still have our focus applicable to the quality of wine that we want to emulate. The other assumption is that the best linear predictor must exist. The

distribution of average rating for decent wines looks reasonably well behaved. The Central Limit Theorem should work and will satisfy the assumptions necessary for the large-sample model to produce consistent estimates.

Structural Limitations

One important omitted variable is how long the wine matured. The longer the wine matures, the smoother the wine, leading to a better taste and over higher rating. As the wine age increases, more phenolic compounds link, and the wine's primary flavor can fade. This can lead to wines with a heavier alcohol content having a less pronounced alcoholic flavor. Because the omitted variable (age of the wine) has a positive effect on the quality of the wine, but also a negative effect on the alcohol variable, it pushes the estimator (beta_alcohol) toward zero. Since the effect is toward zero, omitting the age of the wine does not call any of the results into question.

Conclusion

Improving the quality of a wine is a massive topic. Countless books, essays, and papers exist on the subject. The above exercise is built the assumption that wine quality is easier to detect than define. Second assumption is that quality is explainable and can be improved by knowing factors that makes it more enjoyable. In a future study, we could use sales data as a replacement metric for "success."

Our study focused on alcohol and the role it played in improving the quality of highly rated Portuguese wines. We found that wine experts preferred a higher alcohol content in their wines as the coefficient for alcohol was positive. We also found several other factors that contributed to the quality rating of the wines. We hope our findings will allow the company to improve the production success of their wine product.