# PSTAT126 Project Step 2 11/12

Hanya Ansari, Carina Yuen, Daren Aguilera

2023-11-10

**Introduction:**

**Wine Quality Based on Physicochemical Tests from UCI Machine Learning Repository**

**No Missing Attribute Values**: 0

**Number of Instances:** red wine: 1599

**Number of Variables:** 12 total, 11 continuous, 1 discrete (fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH, sulphates, alcohol and 1 integer output variable: quality score between 0 and 10)

**Hypothesis**

**Null Hypothesis: all measured explanatory variables have no linear relationship to the wine quality score.**

$$H_0 : \beta_i = 0 \, \forall \, i \in f, v, c, r, ch, fs, ts, d, p, s, a$$

**Alternative Hypothesis: at least one measured explanatory variable has a linear relationship with the wine quality score.**

$$H_a : \beta_i \neq 0 \; for \, at \, least \, one \, i \in f, v, c, r, ch, fs, ts, d, p, s, a$$

**Partial F-test**

$$\beta_{\text{citric acid}} = \beta_{density} = 0$$

```
## Analysis of Variance Table
##
## Model 1: quality ~ volatile.acidity + fixed.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + pH + sulphates
## Model 2: quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1   1591 816.57
## 2   1587 666.41  4    150.16 89.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We see from the anova table that our F-value is 2.2e-16, therefore we reject our null hypothesis. (i.e. there is no significant difference in the features of citric acid and density.)

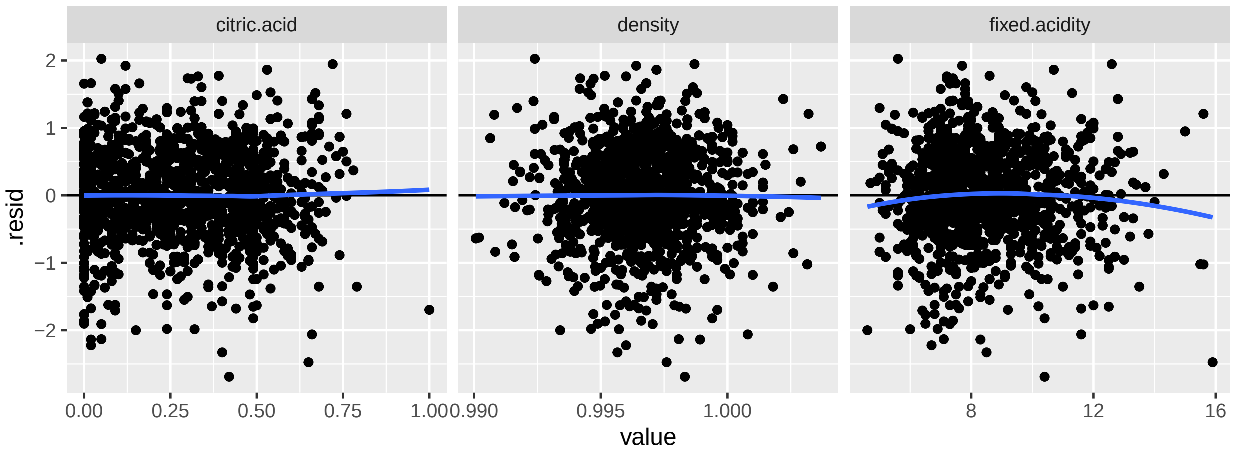

Figure 1: Panel of residual plots

**Assumptions for Linear Regression**

Alcohol vs Quality Score and Total Sulfur Dioxide Analysis: The residuals plot for alcohol vs quality score shows that it does not violate the assumption of linearity because the graph shows a pretty clear linear relationship.

Alcohol vs Residual Sugars Analysis: Residual Sugars The data shows a reverse fan shape,

Free Sulfur vs Quality Score Analysis: The residuals plot for free sulfur vs quality score analysis shows that there is not a clear linear relationship, it almost looks like a parabolic shape that's been rotated.

Residual Sugars Analysis: The data shows a reverse fan shape,

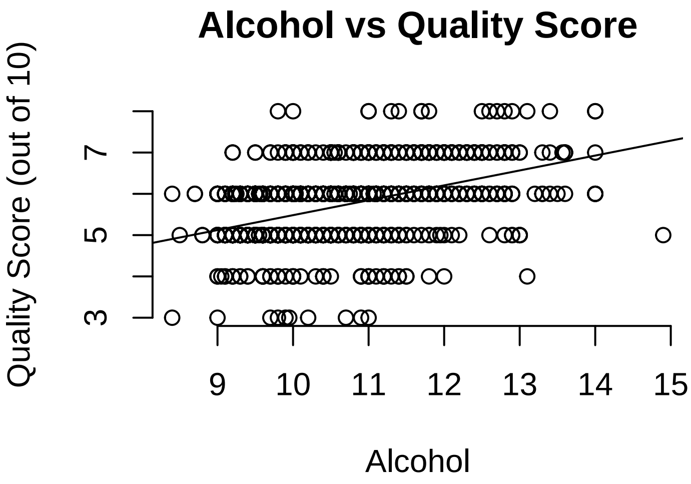

Figure 2: Checking Assumptions for linear regression

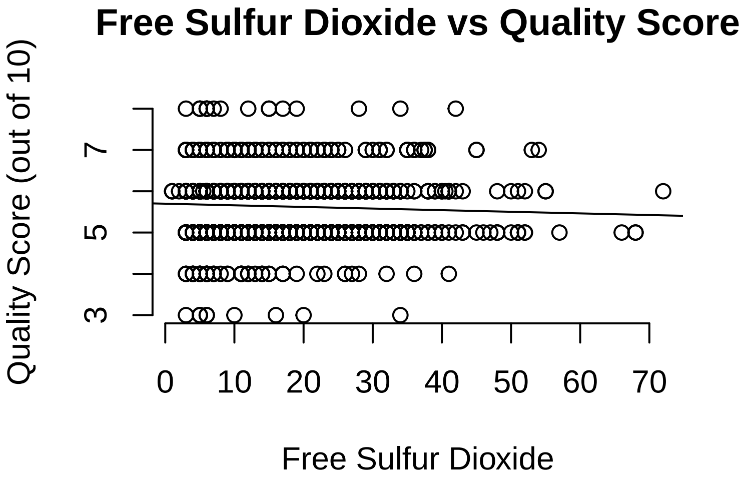

Figure 3: Checking Assumptions for linear regression

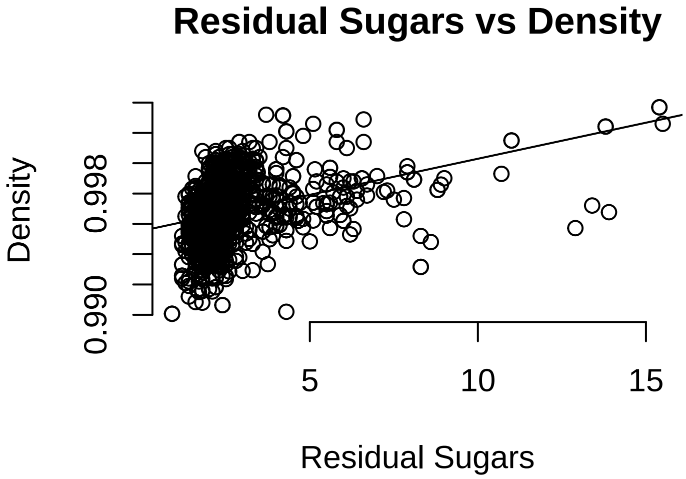

Figure 4: Checking Assumptions for linear regression

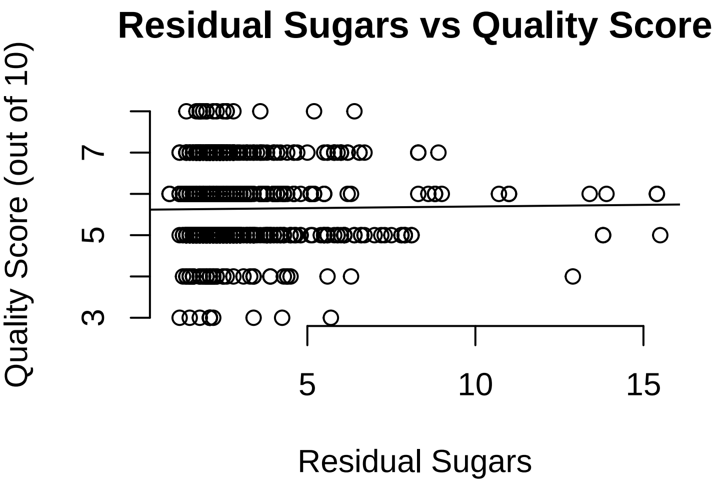

Figure 5: Checking Assumptions for linear regression

**ANOVA Testing:**

We created four models: fit_alc, fit_quality, fit2_quality, and fit3_quality that have polynomials of order 1, 2, and 3, respectively. We applied the poly() function to wine quality variables with lower p-values, as we thought these would be the most influential on the response variable of quality score. The variables we chose to apply poly() to were the following: x_v, x_ch, x_fs x_ts, x_p x_s, x_a (measuring volatile acidity, chloride content, free sulfur dioxide, total sulfur dioxide, phosphates, sulphates, and alcohol content).

Overall Analysis of fitalc (Alcohol Content and Density): The sum of squares value for the alcohol content variable x_a is 236.3, which measures the variation between the group mean and the overall mean. The F value is 521.6 which is positive because a larger F value means that it is more likely alcohol content is causing the variation in density. Given that the p value (<2e-16) is low, it seems like alcohol content has an impact on the quality score.

Two-Way ANOVA for Alcohol Content and Density (continuous variables case): x_a measures alcohol percentage. The percentage of alcohol content in the wine directly affects the wine's density, so it makes sense that are consistent with one another. The F value was 521.6, which indicates that it is more likely that alcohol content is causing variation in density. In addition, the p value is also very low <2e-16, further supporting this.

ANOVA for Alcohol Content and Quality Data (discrete variable with integer score 1-10): The F value was 468.3, which indicates that it is likely alcohol content causes changes in Final Quality Score of each wine. The p value is also very low <2e-16.

Residual Plot Analysis: The fitted vs residual plots for fit2 vs fit3 (poly order 2 vs 3) are quite similar, indicating that the change in the polynomial order did not change the model for data quality score much.

R^2 values: “R squared value: fit alc” [1] 0.6680158 [1] “R squared value: fit 1” [1] 0.3561195 [1] “R squared value: fit 2” [1] 0.3761969 [1] “R squared value: fit 3” [1] 0.3912501

```
## [1] "Full and Nested Models"
```

```
## [1] "R squared value: fit alc, fit 1, fit 2, fit 3"
```

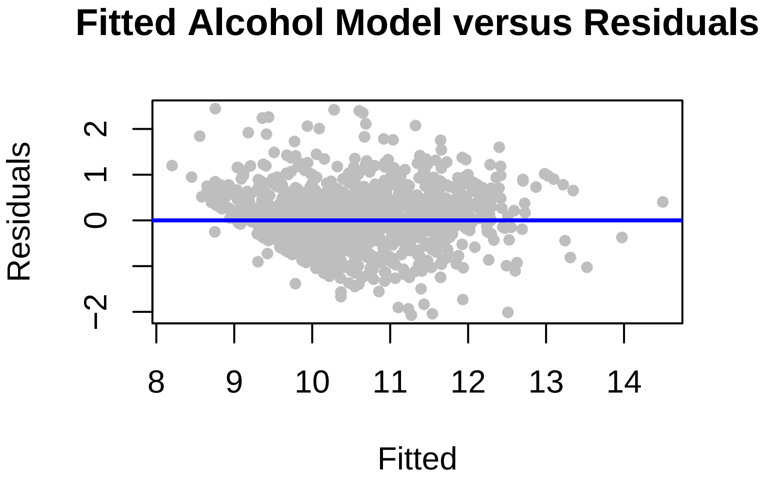

Figure 6: Full and nested models

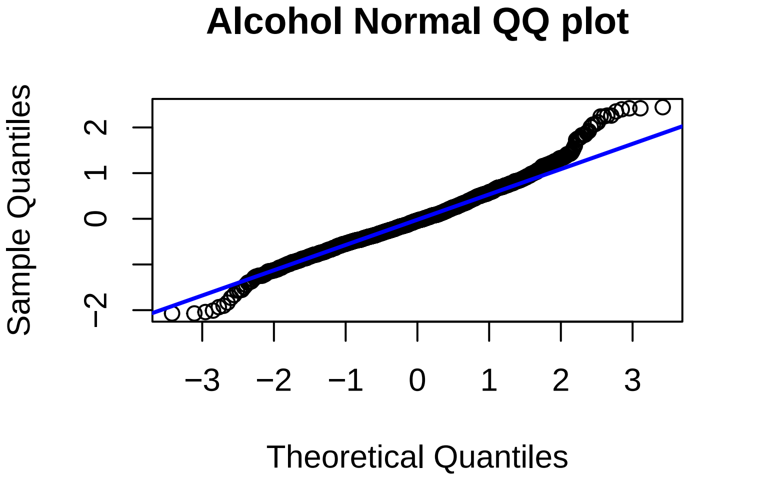

Figure 7: Full and nested models

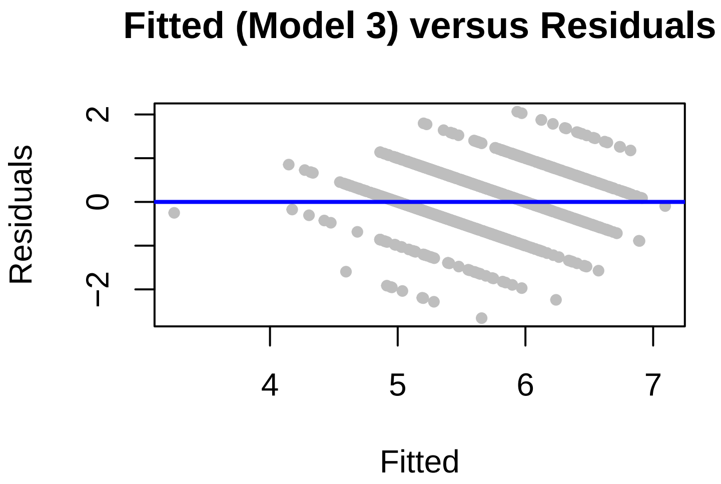

Figure 8: Full and nested models

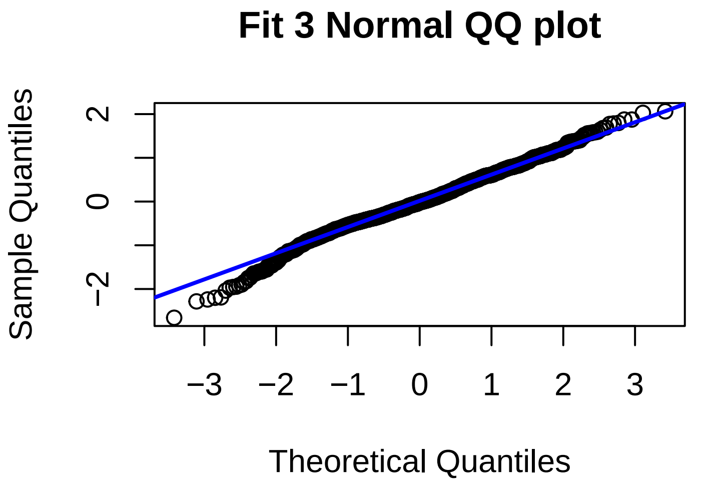

Figure 9: Full and nested models

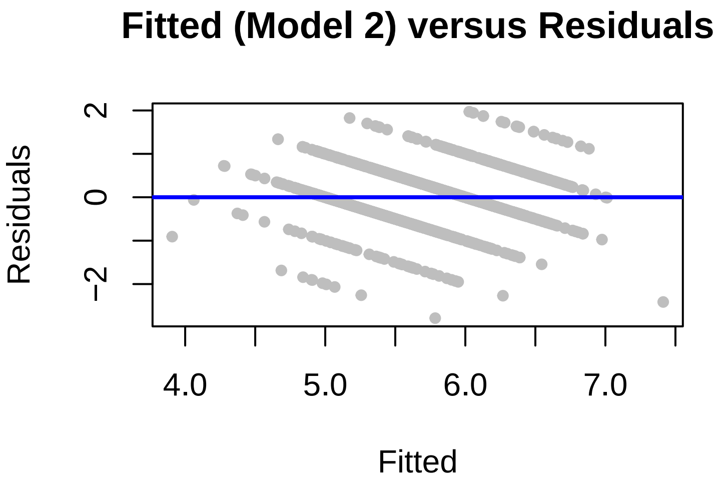

Figure 10: Full and nested models

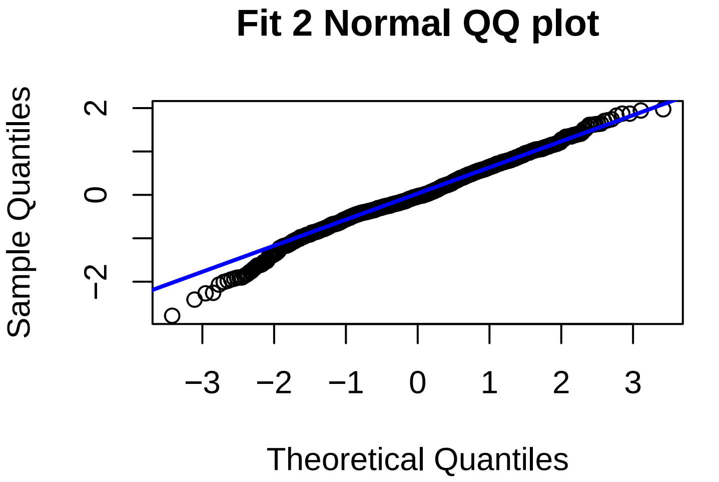

Figure 11: Full and nested models

```
## [1] 0.6680158
```

```
## [1] 0.3561195
```

```
## [1] 0.3761969
```

```
## [1] 0.3912501
```

```
## [1] "Density and Alcohol Content ANOVA"
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## density        1  446.8   446.8   521.6 <2e-16 ***
## Residuals   1597 1368.0     0.9
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## [1] "Quality Score ANOVA"
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## x_a            1  236.3   236.3   468.3 <2e-16 ***
## Residuals   1597  805.9     0.5
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

### Confidence Interval and Prediction Interval

**Variables of Interest** fit_alc: The mean alcohol content value (volume %) is about 10.42. Our confidence interval output suggests the 95% confidence interval is between 9.55 and 9.69 (one of randomly chosen intervals) when only taking into account variables of interest. Our prediction interval output suggests that the 95% prediction interval for a particular value ranges from about 8.41 to 10.82 when also only taking into account variables of interest. Here we see that the prediction interval is larger than the confidence interval as we expect. It seems that given the other wine parameters, we can predict pretty well the mean alcohol value.

Variables of Interest and fit3quality: The mean quality score is about 5.06/10. Our output suggests the confidence interval is between 4.97 and 5.14 with 95% confidence when only taking into account variables of interest. Our output suggests that the prediction interval for a particular value ranges from about 3.8 to 6.29 with 95% confidence when also only taking into account variables of interest. Here we see that the prediction interval is larger than the confidence interval as we expect.

### Conclusion

From our ANOVA testing of alcohol content and final quality score, the F value was 468.3, which indicates that it is likely alcohol content causes changes in Final Quality Score of each wine.

Our two-way ANOVA testing of alcohol content and density indicated consistency amongst both continuous variables. The F value was 521.6 and the low p value was (<2e-16) indicating that alcohol content causes variation in density.

The fitted vs residual plots for fit2 vs fit3 (poly order 2 vs 3) are quite similar, with fit3 being the better model. The Rˆ2 values are 0.6680158 for fit alc, 0.3561195 for fit1, 0.3761969 for fit 2, and 0.3912501 for fit 3.

Our confidence interval for the variables of interest and fit_alc indicated that the 95% confidence interval was between 9.55 and 9.69 when only considering the variables of interest and their impact. In relation to the variables of interest and fit3quality, our output suggested that the confidence interval was between 4.97 and 5.14 with 95% confidence. Our output suggests that the prediction interval, as a result, ranges from about 3.8 to 6.29 with 95% confidence when only taking into account variables of interest.

Analysis for nested model (without citric acid and density terms): We see from the anova table that our F-value is 2.2e-16, therefore we reject our null hypothesis. (i.e. there is no significant difference in the features of citric acid and density.)