STAT 5214 Homework - 3

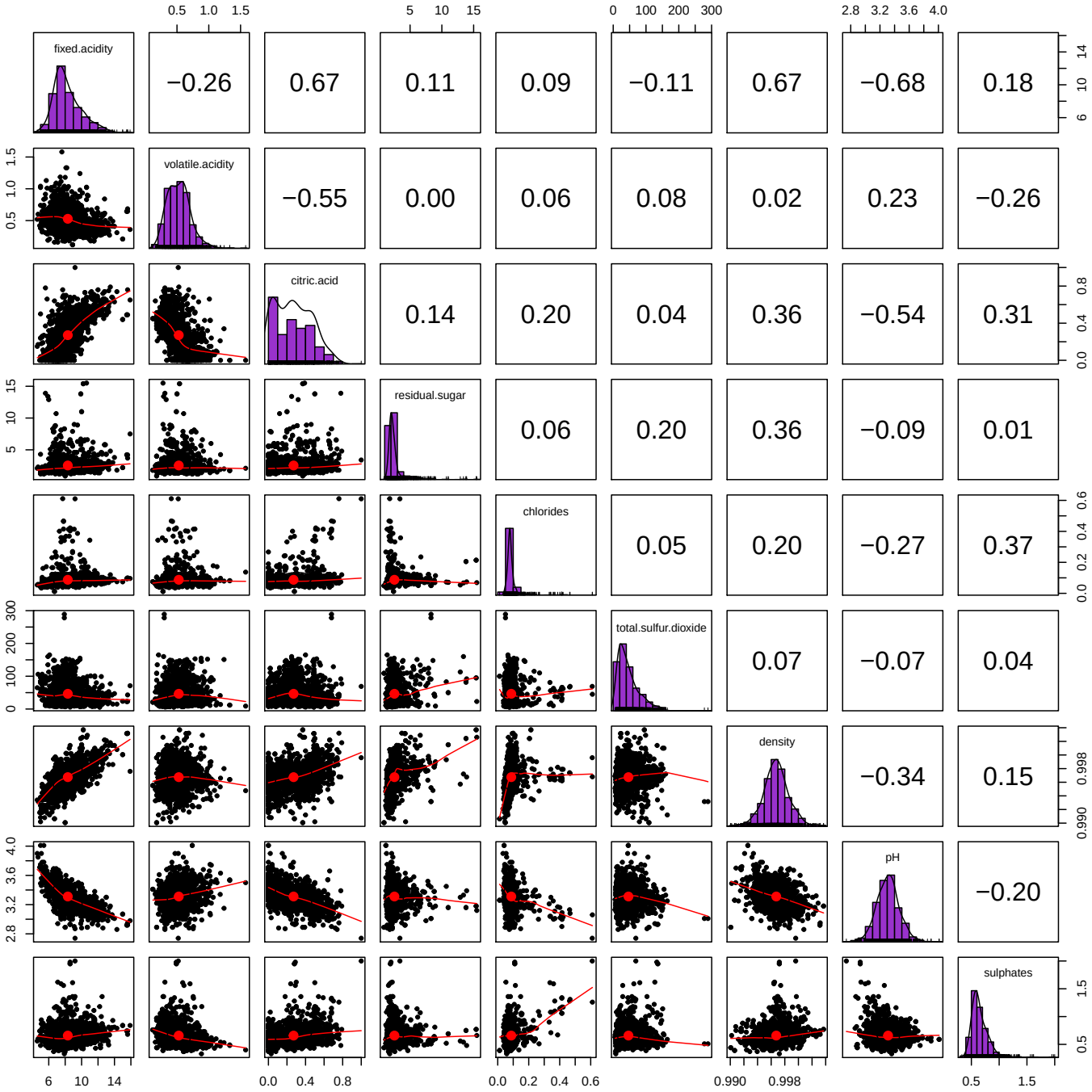
Bineyam Tafesse

June 13, 2021

1. [5 pts] Fit the winning model from Homework 1 and 2.

```
##
## Call:
## lm(formula = alcohol ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + +total.sulfur.dioxide + density +
##     pH + sulphates, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.06145 -0.39706 -0.03917  0.34928  2.44848
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.059e+02  1.302e+01  46.535  < 2e-16 ***
## fixed.acidity         5.300e-01  2.051e-02  25.846  < 2e-16 ***
## volatile.acidity      3.809e-01  1.128e-01   3.377 0.000749 ***
## citric.acid           8.548e-01  1.359e-01   6.289 4.12e-10 ***
## residual.sugar        2.827e-01  1.219e-02  23.198  < 2e-16 ***
## chlorides            -1.487e+00  3.949e-01  -3.766 0.000172 ***
## total.sulfur.dioxide -2.775e-03  5.123e-04  -5.416 7.02e-08 ***
## density              -6.160e+02  1.335e+01 -46.125  < 2e-16 ***
## pH                    3.739e+00  1.534e-01  24.369  < 2e-16 ***
## sulphates             1.242e+00  1.036e-01  11.984  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.614 on 1589 degrees of freedom
## Multiple R-squared:  0.6699, Adjusted R-squared:  0.668
## F-statistic: 358.2 on 9 and 1589 DF,  p-value: < 2.2e-16
```

2. [20 pts] Create a scatterplot matrix with correlations. Comment on what you observe.



Comments:

- None of the correlation coefficients between pairs of explanatory variables exceed the 0.8 or 0.9
- This is a good indication that we probably do NOT have a multicollinearity problem

3. [25 pts] Use tolerance/VIF's to assess whether there are near linear relationships of three or more variables. Comment on what you observe.

Variables	Tolerance	VIF
fixed.acidity	0.1851076	5.402262
volatile.acidity	0.5786811	1.728068
citric.acid	0.3365335	2.971472
residual.sugar	0.7990415	1.251499
chlorides	0.6831588	1.463788
total.sulfur.dioxide	0.8308890	1.203530
density	0.3714034	2.692490
pH	0.4205328	2.377936
sulphates	0.7650299	1.307138

Comments:

- None of the values of VIF are above 10
- This is a good indication that we probably do NOT have a multicollinearity problem

4. [25 pts] Use condition indices to evaluate whether there is multicollinearity in this model.

Table 1: Table continues below

Eigenvalue	Condition Index	intercept	fixed acidity	volatile acidity	citric acid	residual sugar
8.815	1	0	0	0.001	0.001	0.002
0.387	4.77	0	0	0.025	0.186	0.002
0.302	5.399	0	0	0.016	0.029	0.032
0.211	6.462	0	0	0	0	0.51
0.174	7.124	0	0.002	0.002	0.015	0.299
0.062	11.89	0	0.002	0.36	0.172	0.01
0.029	17.41	0	0.005	0.527	0.089	0.005
0.018	22.18	0	0.324	0.048	0.488	0
0.001	123.1	0.001	0.219	0.004	0	0.001
0	3600	0.999	0.446	0.016	0.022	0.139

chlorides	total sulfur dioxide	density	pH	sulphates
0.002	0.003	0	0	0.001
0.005	0.105	0	0	0.001
0.025	0.596	0	0	0.002
0.21	0.147	0	0	0.003
0.46	0.009	0	0	0.001
0.011	0	0	0	0.358
0.152	0.006	0	0.003	0.626
0.008	0.049	0	0.005	0.007
0.069	0.047	0	0.906	0
0.058	0.039	1	0.085	0.002

Comments:

- There are two condition indices that can be labeled as large (>30) which are 123.1 and 3600.
- The condition index 123.051 does NOT have a variable with variance proportion larger than 50%

- The condition index 3600 does NOT have a variable with variance proportion larger than 50% except the intercept
- It appears we do NOT have a multicollinearity problem

5. [25 pts] Based on all the information you have (homework 1 - homework 3), what can you conclude about this model?

Comments:

- In HM 1 we used a Stepwise Selection and Forward Selection methods to reach to the winning model and also examine R squared, Adj. R squared, AIC, RSS and Sum Sq.
- We've also performed tests to determine whether alcohol is dependent on at least one of the predictors and to determine whether the acohol content is associated with chlorides. In both these tests we reject the null hypothesis with p- values 2.2e-16 and 0.00.
- In HM 2, we did a thorough diagnostic check such as Normal probability plot and Residuals vs. Predicted (using both Regular and Studentized Residuals) and concluded that the model show no evidence of a problem.
- We also performed Residuals vs. Regressor in the model with both Regular and Studentized Residuals and concluded that we might have have a problem of non-constant variance.
- We further investigated for outliers, leverage and influential points on the winning model by calculating and plotting for leverage points (for each $4p/n$ *and* 2p/n) Cook's D, DFBETAS and DFFITS. We then re-fitted the model after excluding the influential records and compared the values of MSE, R Squared, Adj. R Squared, F statistic and regression coefficient estimates with Winning model and concluded that there is no significance change from the original winning model, and also there was no change in sign & in significance when excluding the outliers.
- In HM 3 we performed tests to see if we have multicollinearity problem by using scatter plot matrix with correlation, VIF tolerance and condition indices and we concluded that we do NOT have any multicollinearity problem.
- based on all above listed assessments, we can conclude that the winning model is robost and best model to determine for alcohol content (y).