WQD.2 - Multiple Regression

Sample R code for Multiple Regression

Linear regression is fitted to the Training data.

Model I: All predictors in the model

Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)	VIF
(Intercept)	166.40	38.60	4.31	0.00	
fixed.acidity	0.13	0.04	3.38	0.00	3.15
volatile.acidity	-1.85	0.22	-8.27	0.00	1.12
citric.acid	0.05	0.18	0.30	0.76	1.12
residual.sugar	0.08	0.01	5.59	0.00	19.06
chlorides	-3.63	2.05	-1.77	0.08	1.59
free.sulfur.dioxide	0.00	0.00	2.88	0.00	1.87
total.sulfur.dioxide	0.00	0.00	0.40	0.69	2.52
density	-167.20	39.12	-4.27	0.00	47.87
pH	0.85	0.18	4.68	0.00	2.41
sulphates	0.81	0.18	4.64	0.00	1.14
alcohol	0.16	0.05	3.24	0.00	13.03

For extremely high VIF density was removed from the model. There are other predictors with high VIF, but they were not removed at this step.

Model II: After removal of density VIFs improved

Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)	VIF
(Intercept)	1.47	0.56	2.61	0.01	
fixed.acidity	0.00	0.02	0.14	0.89	1.28
volatile.acidity	-1.91	0.22	-8.54	0.00	1.12
citric.acid	-0.01	0.18	-0.03	0.97	1.11
residual.sugar	0.02	0.00	5.29	0.00	1.49
chlorides	-5.47	2.01	-2.72	0.01	1.52
free.sulfur.dioxide	0.01	0.00	3.63	0.00	1.82
total.sulfur.dioxide	0.00	0.00	-0.79	0.43	2.34
pH	0.31	0.13	2.38	0.02	1.27
sulphates	0.60	0.17	3.58	0.00	1.05
alcohol	0.34	0.02	18.32	0.00	1.98

Not all predictors are significant. A forward selection method is employed to build a working model. The sample R output follows:

```
Start:  AIC=-680.65
quality ~ 1

                      Df Sum of Sq    RSS      AIC
+ alcohol              1   283.176 1173.8 -1118.89
+ chlorides            1   130.970 1326.0  -870.52
+ total.sulfur.dioxide 1    48.673 1408.3  -747.86
+ residual.sugar       1    25.581 1431.4  -714.73
+ volatile.acidity     1    21.608 1435.3  -709.09
+ pH                   1    13.037 1443.9  -696.96
+ fixed.acidity        1     4.089 1452.9  -684.38
<none>                             1457.0  -680.65
+ citric.acid          1     1.238 1455.7  -680.38
+ sulphates            1     1.002 1456.0  -680.05
+ free.sulfur.dioxide  1     0.240 1456.7  -678.99

Step:  AIC=-1118.89
quality ~ alcohol

                      Df Sum of Sq    RSS     AIC
+ volatile.acidity     1    40.585 1133.2 -1188.6
+ free.sulfur.dioxide  1    17.564 1156.2 -1147.6
+ residual.sugar       1    11.564 1162.2 -1137.0
+ sulphates            1     5.907 1167.9 -1127.2
+ chlorides            1     5.772 1168.0 -1126.9
+ pH                   1     2.837 1171.0 -1121.8
+ total.sulfur.dioxide 1     1.706 1172.1 -1119.8
+ citric.acid          1     1.421 1172.4 -1119.3
<none>                             1173.8 -1118.9
+ fixed.acidity        1     0.244 1173.5 -1117.3

Step:  AIC=-1188.56
quality ~ alcohol + volatile.acidity

                      Df Sum of Sq    RSS     AIC
+ residual.sugar       1   18.8659 1114.3 -1220.8
+ free.sulfur.dioxide  1   16.5614 1116.6 -1216.5
+ sulphates            1    5.7863 1127.4 -1197.0
+ total.sulfur.dioxide 1    5.6644 1127.5 -1196.8
+ chlorides            1    4.1606 1129.0 -1194.1
+ pH                   1    2.3397 1130.9 -1190.8
<none>                             1133.2 -1188.6
+ fixed.acidity        1    0.8165 1132.4 -1188.0
+ citric.acid          1    0.0908 1133.1 -1186.7

Step:  AIC=-1220.76
quality ~ alcohol + volatile.acidity + residual.sugar

                      Df Sum of Sq    RSS     AIC
+ free.sulfur.dioxide  1    9.2408 1105.1 -1235.7
+ sulphates            1    7.9372 1106.4 -1233.3
+ pH                   1    4.9199 1109.4 -1227.8
+ chlorides            1    3.8495 1110.5 -1225.8
+ total.sulfur.dioxide 1    2.0808 1112.2 -1222.6
<none>                             1114.3 -1220.8
+ fixed.acidity        1    1.0437 1113.3 -1220.7
+ citric.acid          1    0.0002 1114.3 -1218.8

Step:  AIC=-1235.72
quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide

                      Df Sum of Sq    RSS     AIC
+ sulphates            1    7.4161 1097.7 -1247.4
+ pH                   1    4.2500 1100.8 -1241.6
+ chlorides            1    4.0537 1101.0 -1241.2
<none>                             1105.1 -1235.7
+ fixed.acidity        1    0.6325 1104.5 -1234.9
+ total.sulfur.dioxide 1    0.1885 1104.9 -1234.1
+ citric.acid          1    0.0467 1105.0 -1233.8

Step:  AIC=-1247.44
quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
    sulphates

                      Df Sum of Sq    RSS     AIC
+ chlorides            1    4.3656 1093.3 -1253.6
+ pH                   1    3.1849 1094.5 -1251.4
<none>                             1097.7 -1247.4
+ total.sulfur.dioxide 1    0.6983 1097.0 -1246.7
+ fixed.acidity        1    0.6129 1097.1 -1246.6
+ citric.acid          1    0.1387 1097.5 -1245.7
```

```
Step:  AIC=-1253.56
quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
    sulphates + chlorides

                      Df Sum of Sq    RSS     AIC
+ pH                   1    3.3650 1090.0 -1257.8
<none>                             1093.3 -1253.6
+ fixed.acidity        1    0.4827 1092.8 -1252.5
+ total.sulfur.dioxide 1    0.2152 1093.1 -1252.0
+ citric.acid          1    0.0848 1093.2 -1251.7

Step:  AIC=-1257.84
quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide + sulphates
+ chlorides + pH

                      Df Sum of Sq    RSS     AIC
<none>                             1090.0 -1257.8
+ total.sulfur.dioxide 1   0.33456 1089.6 -1256.5
+ citric.acid          1   0.00436 1089.9 -1255.8
+ fixed.acidity        1   0.00167 1089.9 -1255.8
```

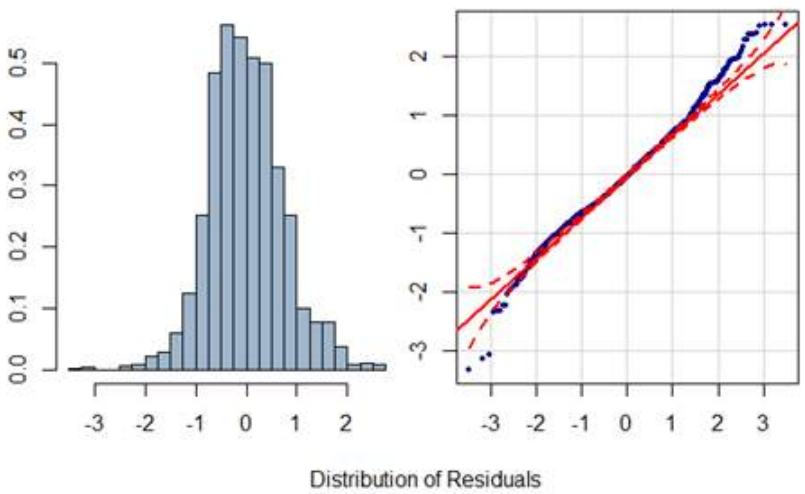
Model III: Working model

Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)
(Intercept)	1.49	0.46	3.27	0.00
alcohol	0.35	0.02	19.20	0.00
volatile.acidity	-1.95	0.22	-9.03	0.00
residual.sugar	0.02	0.00	5.24	0.00
free.sulfur.dioxide	0.005	0.001	3.95	0.00
sulphates	0.59	0.17	3.51	0.00
chlorides	-5.74	1.97	-2.91	0.00
pH	0.30	0.12	2.50	0.01

Sample R output:

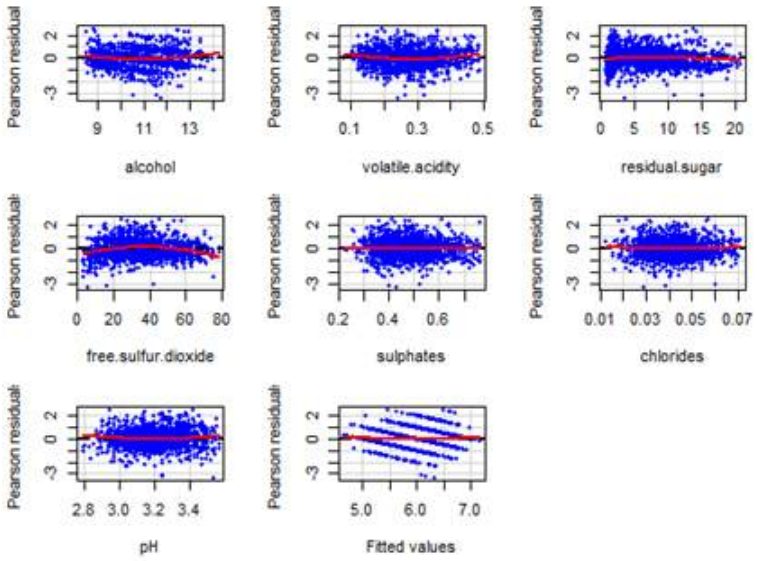
```
Residual standard error: 0.7329 on 2029 degrees of freedom
Multiple R-squared:  0.2519,   Adjusted R-squared:  0.2493
F-statistic:  97.6 on 7 and 2029 DF,  p-value: < 2.2e-16
```

Note that multiple R^2 is 25%. Regression diagnostics are examined for possible improvement of the model.



Distribution of Residuals

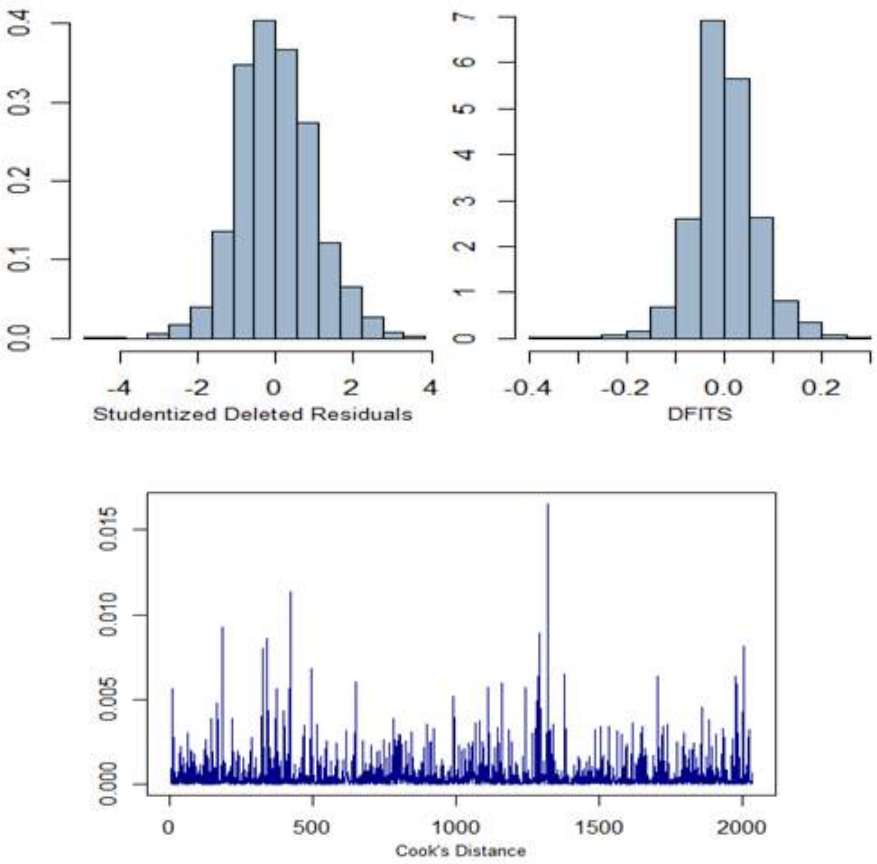
Residuals have an approximately symmetric distribution but there seems to be outliers at both ends. Partial residual plots are given below. Note the pattern in the fitted value plot. Since the response actually takes only integer values but has been assumed to be continuous, such pattern arises.



Outliers and leverage points are identified through the following:

- Studentized deleted residuals (a point is outlier if residual is outside of [-3, 3] limits
- DFITS (a point is outlier if residual is outside of [-1, 1] limits
- Cook's distance

All three plots are given below. Note that no point is identified as outlier with DFITS value.



Only 26 points are identified as outliers according to the above criteria. A final model is fit after eliminating these points and a slight improvement in the R^2 value is noted.

Model IV: Final model

Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)
(Intercept)	1.41	0.43	3.25	0.00
alcohol	0.35	0.02	20.42	0.00
volatile.acidity	-1.99	0.20	-9.72	0.00
residual.sugar	0.02	0.00	5.58	0.00
free.sulfur.dioxide	0.004	0.001	3.21	0.00
sulphates	0.56	0.16	3.57	0.00
chlorides	-5.79	1.87	-3.10	0.00
pH	0.34	0.11	2.94	0.00

Sample R output:

```
Residual standard error: 0.6884 on 2003 degrees of freedom
Multiple R-squared:  0.2809,   Adjusted R-squared:  0.2784
F-statistic: 111.8 on 7 and 2003 DF,  p-value: < 2.2e-16
```

Application of this model on test data gives sum of square of differences between the actual response and predicted response to be 1196.205 whereas sum of square of deviations of actual response is 1554.754. Ratio of these two may be taken as the ratio of Error sum of squares and total sum of squares. Hence a measure similar to that of R^2 may be computed as 1 – 1196.205/1554.754 = 0.2306.

Sample R code for Final Model

Source URL: https://onlinecourses.science.psu.edu/stat857/node/225