EFFECT OF PHYSICOCHEMICAL PROPERTY ON WINE QUALITY

WEIQI WENG

COLLEGE OF COMPUTER AND INFORMATION SCIENCE

NORTHEASTERN UNIVERSITY

INTRODUCTION

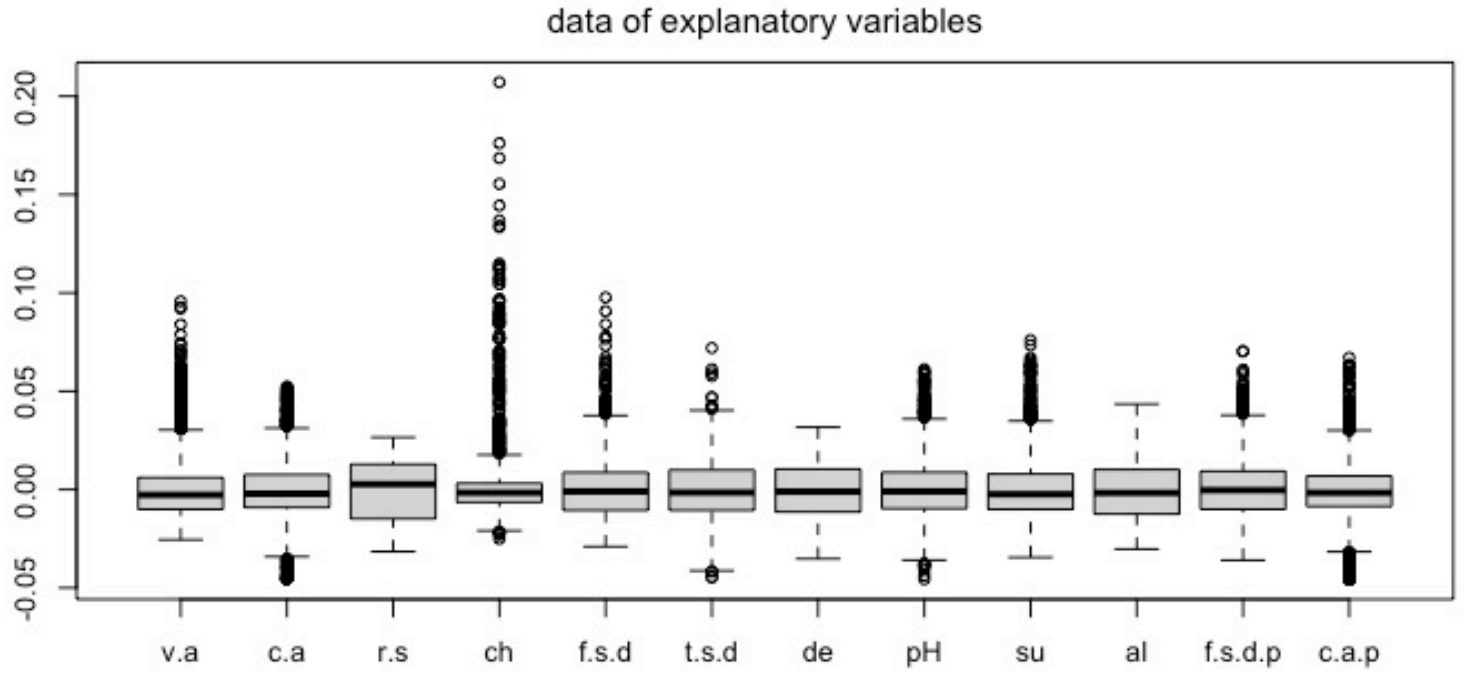
- What kind of wine has better quality?
- Can we predict the quality based on physicochemical property?
- How does physicochemical property correlate to quality?

DATA BACKGROUND

- Feature pool: numerical data
- fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol
- Response variable: quality
- Significance level 0.05

DATA PREPROCESSING: SKEWNESS AND OUTLIERS

- Logarithmic transformation on residual sugar
- Box plot



DATA PREPROCESSING: OUTLIERS

- Preceding Research

Component	Range (mg/L)	Threshold	Component	Range (mg/L)	Threshold
tartaric acid	1000 – 4000		Acetic acidity	600 – 900	
Malic acid	0 – 8000		Volatile acidity	0 – 1100 (legal)	(0, 1100]
Citric acid	0 – 500	(0, 500]	Sulfur dioxide	0 – 350 (legal)	(0, 350]
Succinic acid	500 – 2000		Fixed acidity	1500 – 14000	(0, 14000]

(Doug Nierman, 2004)

Density > 1000 mg/L removed
Residual sugar 65.8 removed
Legal limit for chloride concentration varies
pH level 3.0 – 3.4

UCDAVIS UNIVERSITY OF CALIFORNIA

Waterhouse Lab

DATA PREPROCESSING: RELATIVE INDICATOR

- Fixed acidity contains citric acidity.
- $r = 0.659$
- Total sulfur dioxide contains free sulfur dioxide.
- $r = 0.614$

$$citirc\ acid\ proportion = \frac{citric\ acid}{fixed\ acidity}$$

$$free\ sulfur\ dioxide\ proportion\ = \frac{free\ sulfur\ dioxide}{total\ sulfur\ dioxide}$$

FEATURE SELECTION: EXHAUSTIVE SUBSET SELECTION

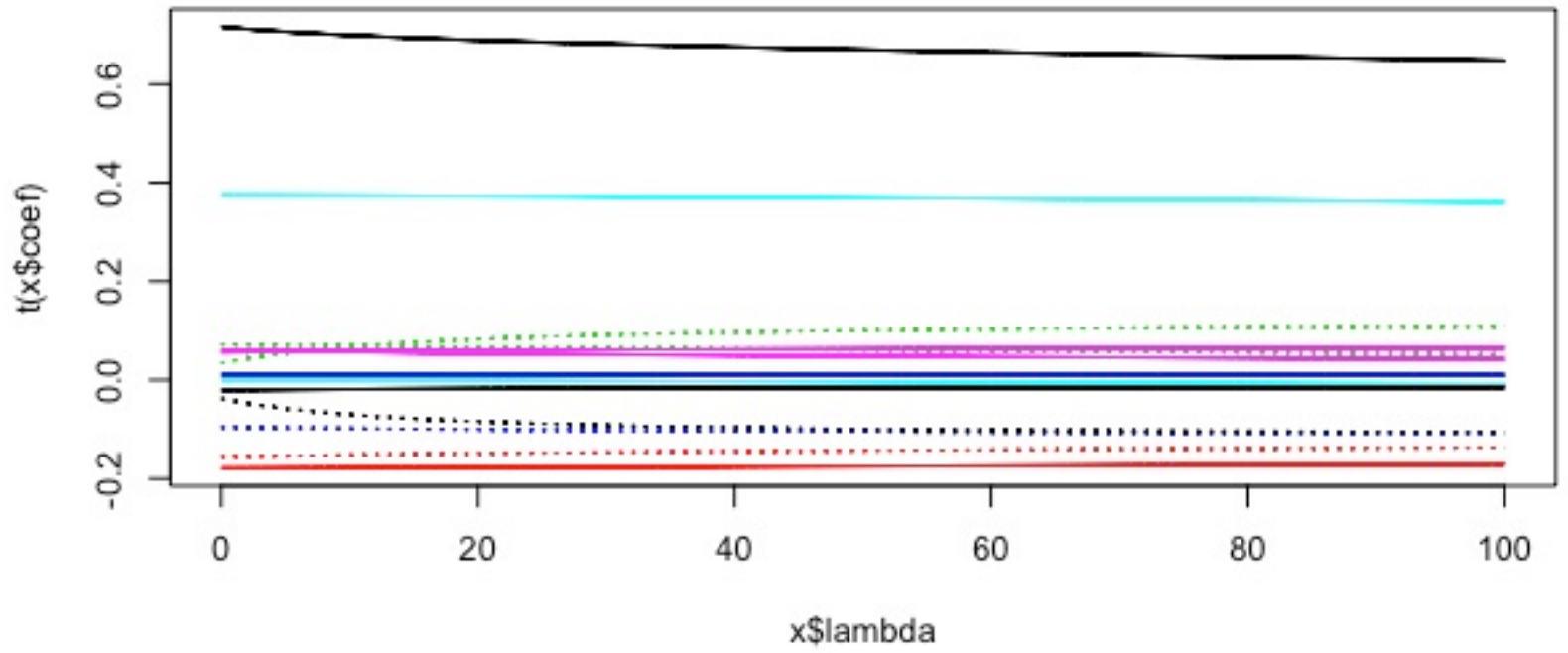
- Criterion: $BIC = -2\ln(L) + k\ln(n)$

Variable	VIF	coefficient
Fixed Acidity	2.238	0.846
Volatile Acidity	1.093	-1.798
Residual Sugar	6.501	-0.193
Density	17.697	-3.705
pH	1.824	0.322
Alcohol	7.389	0.376
Free Sulfur Dioxide Proportion	3.302	0.905

Model 1

MULTICOLLINEARITY

- Ridge Regression



Estimator	Value
HKB	9.942
L-W	14.432
GCV	8.6

Model 2

MULTICOLLINEARITY

- Drop density

$r(density, residual\ sugar) = 0.758$

$r(density, alcohol) = -0.808$

$r(residual\ sugar, alcohol) = -0.172$

- 5-fold Cross Validation

Drop density in Model 1, estimate parameters, run CV on validation set $MSE = 0.817$ ★

Set ridge parameter 8.6 in Model 2, estimate parameters, run CV on validation set $MSE = 0.837$

PARAMETER ESTIMATION

Coefficient	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Fixed Acidity	0.844	0.008	102.156	$< 2^{-16}$
Volatile Acidity	-1.786	0.063	-28.198	$< 2^{-16}$
Residual Sugar	-0.194	0.007	-26.494	$< 2^{-16}$
pH	0.318	0.045	7.024	$2.81 \cdot 10^{-12}$
Alcohol	0.376	0.005	69.518	$< 2^{-16}$
Free Sulfur Dioxide Proportion	1.097	0.066	16.612	$< 2^{-16}$

Model 3

$R^2 = 0.8295, R_a^2 = 0.8987$

T TEST FOR COEFFICIENTS

Tests for β_k

null hypothesis H_0:

$$\beta_k = 0,$$

the predictor corresponding to β_k is not associated with Quality.
alternative hypothesis H_a:

$$\beta_k \neq 0,$$

the predictor corresponding to β_k is associated with Quality.
test statistic:

$$t^* = \frac{b_k}{s(b_k)}$$

Require $t_{0.975}(2364 - 7) = 1.961$. In table 2, we see $|t^*|$ of each predictor is larger than 1.961, along with an extremely small $p - value$, so we conclude H_a, $\beta_k \neq 0$ for each predictor. This means the selected predictors are associated with Quality.

F TEST FOR REGRESSION RELATION

F Test for Regression Relation

null hypothesis H_0:

$$\beta_1 = \beta_2 = \cdots = \beta_6 = 0,$$

there is no general regression relation between the selected predictors and Quality.
alternative hypothesis H_a:

$$\text{not all } \beta_k \text{ equal to } 0,$$

there is general regression relation between the selected predictors and Quality.
test statistic:

$$F^* = \frac{MSR}{MSE}$$

Require $F_{0.95}(6, 2357) = 2.102$. We use the test statistic $F^* = 3271 > 2.102$ to conclude H_a, that is the existence of a regression relation.

F TEST FOR LACK OF FIT

F Test for Lack of Fit

null hypothesis H_0:

$$E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6,$$

the current model is complex enough to explain Quality.
alternative hypothesis H_a:

$$E(Y) \neq \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6,$$

the current model is not complex enough to explain Quality.
test statistic:

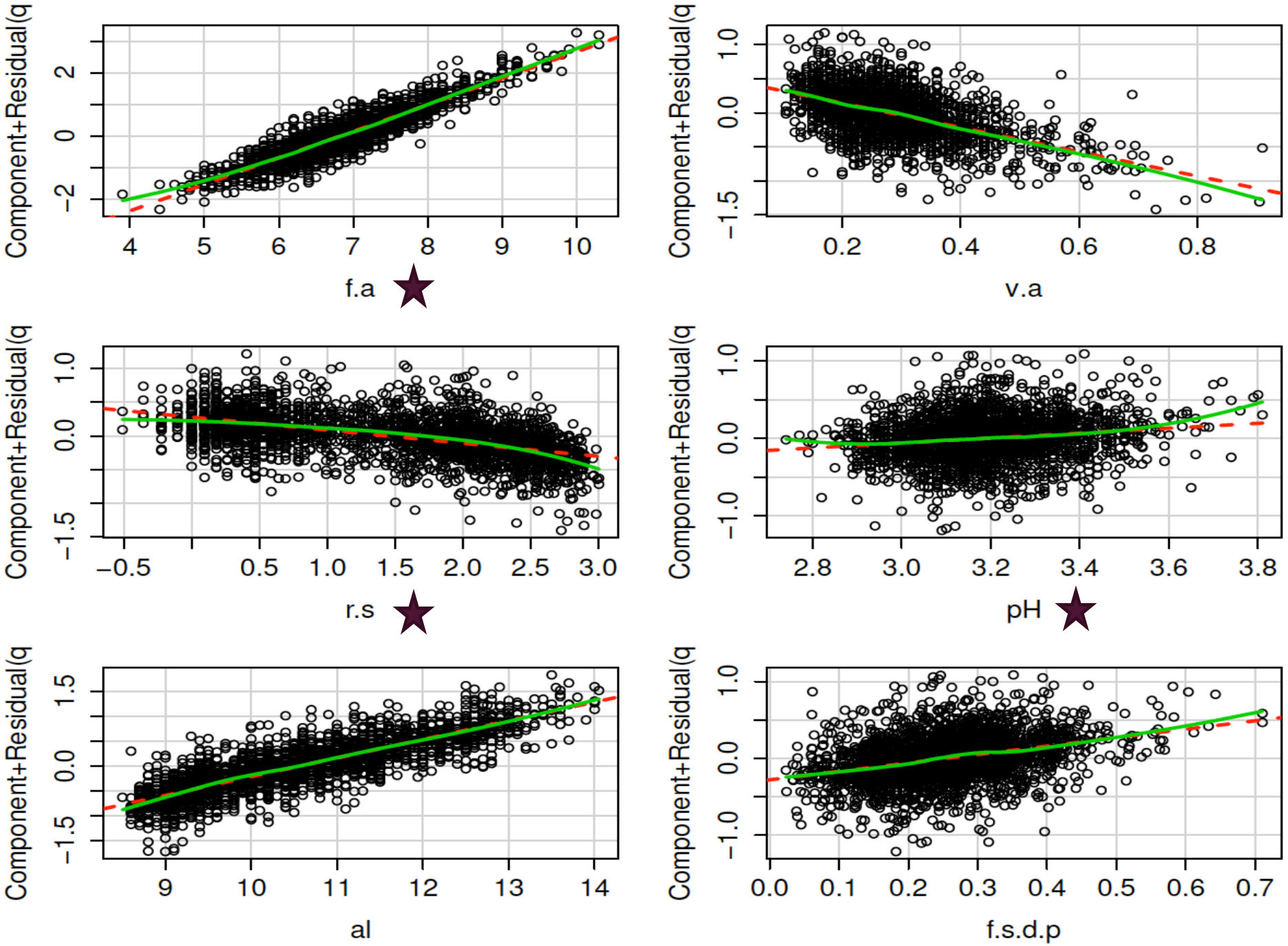
$$F^* = \frac{SSLF}{c-p} / \frac{SSPE}{n-c}$$

Here $n = 2364, p = 7, c = 1972$. The test statistic is $F^* = (203.20 - 29.80)/(1972 - 7)/0.076 = 1.161$. Requiring $F_{0.95}(1965, 392) = 1.141$, we have $F^* > F$ and claim that the current model is not complex enough to explain Quality.

CURVATURE AND INTERACTION INVESTIGATION

- Partial residual plots

$$residual + \widehat{\beta}_i X_i \sim X_i$$


Component+Residual(q
f.a
v.a
r.s
pH
al
f.s.d.p

CURVATURE AND INTERACTION INVESTIGATION

- Acids impart the sourness or tartness that is a fundamental feature in wine taste (Doug Nierman, 2004)
- Residual sugar affects taste.
- Fixed Acidity and Volatile Acidity contributes to smoothness and fragrance. In addition, they make long-term preservation of wine reasonable since it takes time for full fermentation whose by-product is acidity, especially Fixed Acidity.

CURVATURE AND INTERACTION INVESTIGATION

Coefficient	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Fixed Acidity	0.371	0.2833	1.310	0.190
Volatile Acidity	-1.837	0.063	-29.249	$< 2^{-16}$
Residual Sugar	-0.398	0.201	-1.993	0.046
pH	-1.501	1.638	-0.917	0.359
alcohol	0.358	0.006	63.671	$< 2^{-16}$
free sulfur dioxide proportion	1.145	0.065	17.646	$< 2^{-16}$
Fixed Acidity^2	0.028	0.007	3.782	0.000
pH^2	0.236	0.210	1.121	0.263
Residual Sugar^2	-0.088	0.010	-8.903	$< 2^{-16}$
Fixed Acidity $\times$ pH	0.017	0.064	0.263	0.793
Fixed Acidity $\times$ Residual Sugar	0.020	0.009	2.177	0.030
Residual Sugar $\times$ pH	0.097	0.051	1.905	0.057

CURVATURE AND INTERACTION INVESTIGATION

- Observations

Coefficient of pH 0.318 → -1.501

Std. Error of pH 0.045 → 1.638

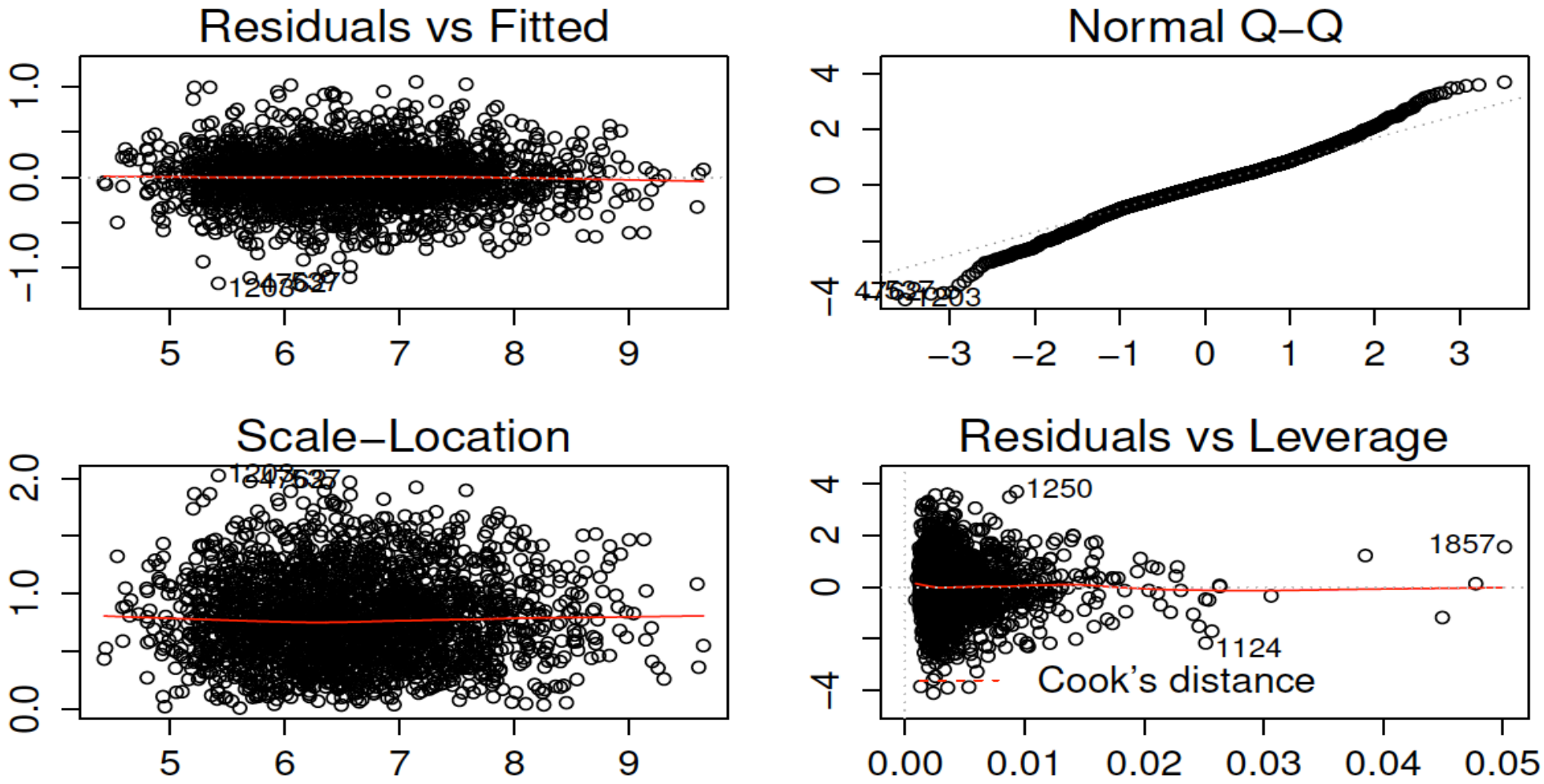
Failing T test for coefficients

- Yeasts can grow in a pH range of 4 to 4.5 and moulds can grow from pH 2 to 8.5, but favour an acid pH (Mountney and Gould, 1988).
- Stable effect of pH level on fermentation

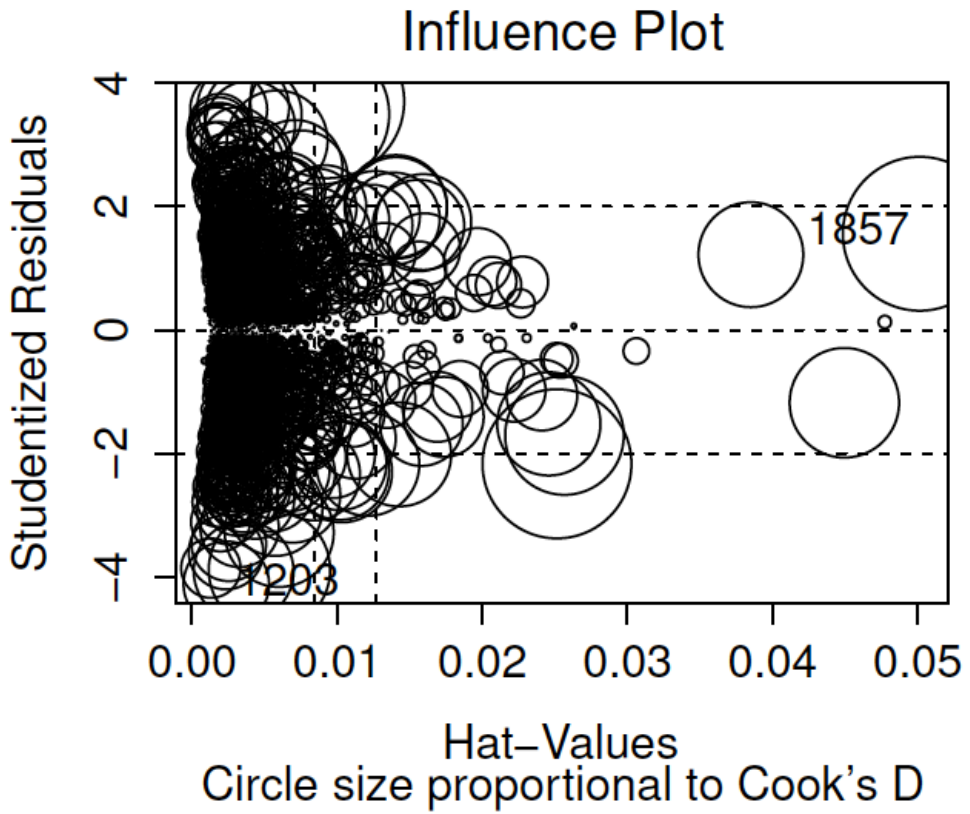
CURVATURE AND INTERACTION INVESTIGATION

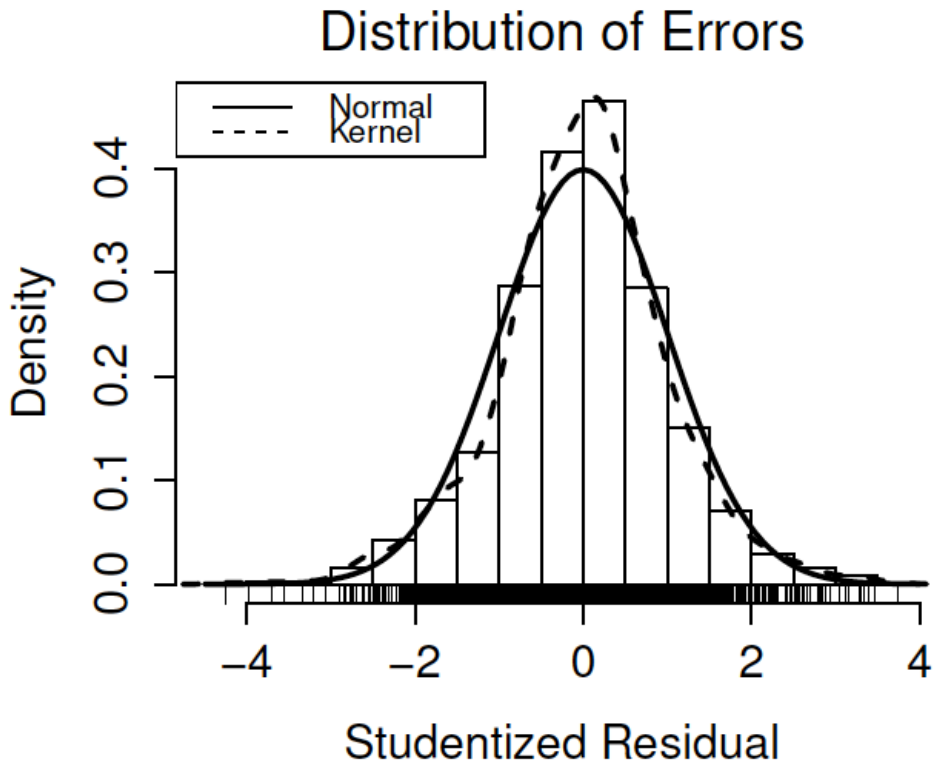
Variable	Coefficient estimate	Std. Error	T statistic	P(>|t|)
Fixed Acidity	0.378	0.093	4.063	$5*10^{-5}$
Volatile Acidity	-1.880	0.063	-29.697	$<2^{-16}$
Residual Sugar	-0.033	0.061	-0.550	0.583
pH	0.249	0.044	5.686	$1.46*10^{-8}$
Alcohol	0.360	0.005	66.174	$<2^{-16}$
Free Sulfur Dioxide Proportion	1.146	0.063	18.071	$<2^{-16}$
Fixed Acidity^2	0.032	0.007	4.930	0.000
Fixed Acidity*Residual Sugar	0.011	0.008	1.363	0.173
Residual Sugar^2	-0.087	0.010	-8.890	$<2^{-16}$

DIAGNOSTICS PLOTS

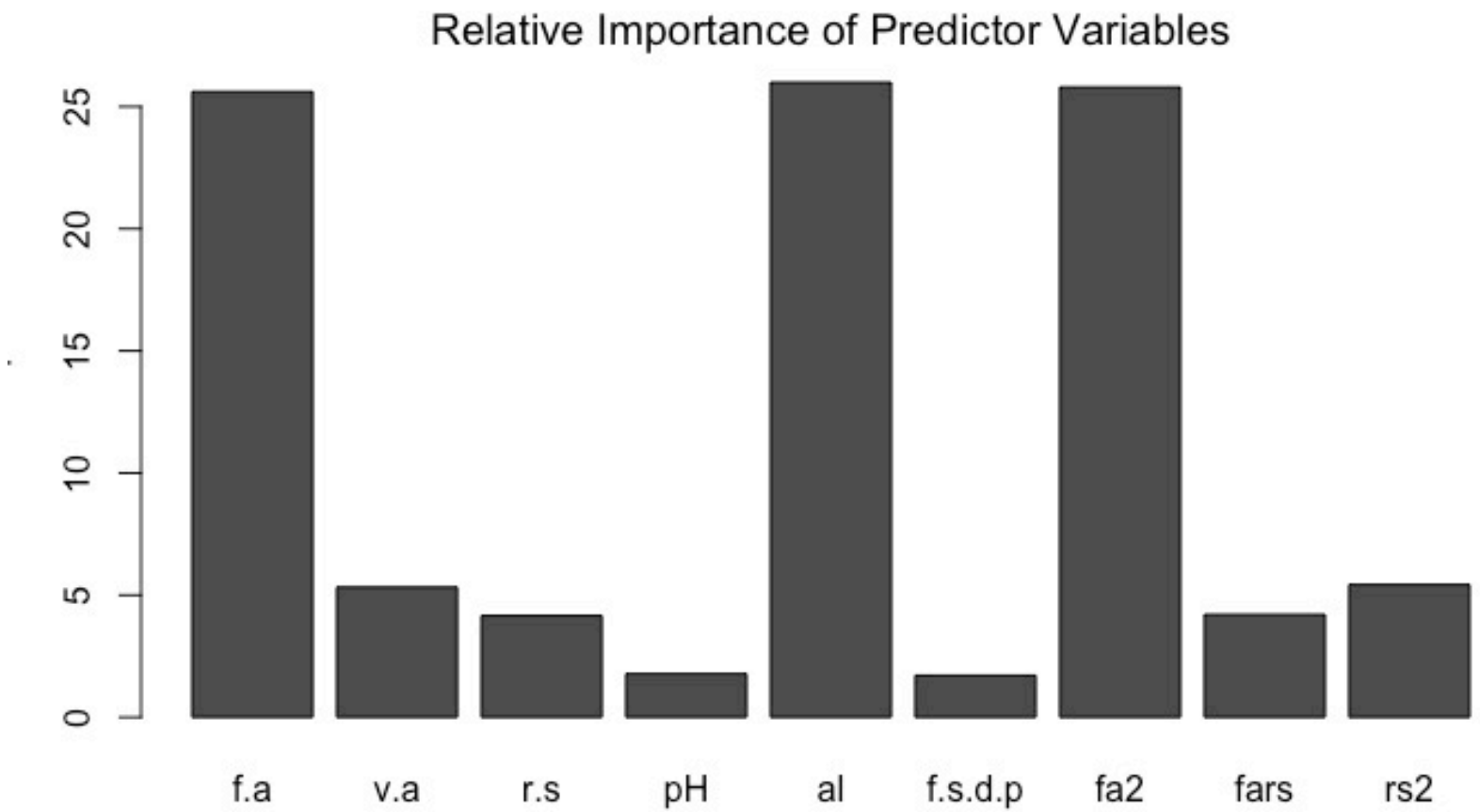


INFLUENCE PLOT





MODEL DISCUSSION



MODEL DISCUSSION

X_1	Fixed Acidity	X_2	Volatile Acidity	X_3	Residual Sugar
X_4	pH	X_5	Alcohol	X_6	Free Sulfur Dioxide Proportion

$$Y_i = 0.378X_{i1} - 1.880X_{i2} - 0.033X_{i3} + 0.249X_{i4} + 0.360X_{i5} + 1.146X_{i6} + 0.032X_{i1}^2 + 0.011X_{i1}X_{i3} - 0.087X_{i3}^2 + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

$$\frac{\partial Y}{\partial X_i}, i = 2,4,5,6$$

$$\frac{\partial Y}{\partial X_i}, i = 1,3 \qquad \frac{\partial Y}{\partial X_1} = 0.378 + 0.064X_1 + 0.011X_3 > 0\ ?$$

MODEL DISCUSSION

Hypothesis Testing 1

Define μ_1 as the mean of X_1, μ_3 as the mean of $(-0.011X_3 - 0.378)/0.064$. Now we care about the mean of population so Central Limit Theorem comes to help.
null hypothesis H_0:

$$\mu_1 - \mu_3 = 0$$

alternative hypothesis H_a:

$$\mu_1 - \mu_3 > 0$$

test statistic:

$$T = \frac{\bar{X}_1 - \bar{X}_3}{\sqrt{s_1^2/n_1 + s_3^2/n_3}} = 1063.038$$

degree of freedom:

$$df = \frac{(s_1^2/n_1 + s_3^2/n_3)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_3^2/n_3)^2/(n_3 - 1)} \approx 5063$$

Require $t_{0.95}(5063) = 1.645$. Since $T > t_{0.95}(5063)$, we conclude H_a.

CONCLUSION

- Fixed acidity, pH, alcohol, free sulfur dioxide proportion
- Volatile acidity and residual sugar
- Fixed acidity and residual sugar

$$MSPE = \frac{\sum_{i=1}^{n^*}(Y_i - \widehat{Y}_i)^2}{n^*} = 0.85$$