109-2 EPM 7012 Statistical and Machine Learning: Assignment 1

Yi-Wen Hsiao(d08849010), Institute of Epidemiology and Preventive Medicine, NTU

March 31, 2021

Contents

Question 1

$$y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}\in\mathbb{R}^n\ ;\ \beta=\begin{bmatrix}w_1\\w_2\\\vdots\\w_p\end{bmatrix}\in\mathbb{R}_p\ ;\ x_n=\begin{bmatrix}x_1\\x_2\\\vdots\\x_p\end{bmatrix}\in\mathbb{R}_p\ ;\ X=\begin{bmatrix}x_{11}&\cdots&x_{1p}\\\vdots&\ddots&\vdots\\x_{n1}&\cdots&x_{np}\end{bmatrix}\in\mathbb{R}^{n\times p}$$

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}(y_i-\beta^T x_n)^2\right\}\ \text{subject to}\ \sum_{j=1}^{p}\beta_j^2\le s$$

$$\text{let}\quad E(\beta)=\sum_{i=1}^{n}(y_i-\beta^T x_n)^2+\lambda\beta^T\beta\overset{\Delta}{=}(y-X\beta)^T(y-X\beta)+\lambda\beta^T\beta$$

$$=(y-X\beta)^T(y-X\beta)+\lambda\beta^T\beta$$

$$=(y^Ty-y^TX\beta-(X\beta)^Ty+(X\beta)^TX\beta)+\lambda\beta^T\beta$$

$$a^Tb=b^Ta\qquad(AB)^T=B^TA^T$$

$$=(y^Ty-2(X\beta)^Ty+\beta^TX^TX\beta)+\lambda\beta^T\beta$$

$$\frac{\partial E(\beta)}{\partial\beta}=(-2X^Ty+2X^TX\beta)+2\lambda\beta\overset{\text{set}}{=}0\ \Rightarrow\ (-X^Ty+X^TX\beta)+\lambda\beta=0$$

$$X^TX\beta+\lambda\beta=X^Ty\ \Rightarrow\ \beta(X^TX+\lambda I)=X^ny\ \Rightarrow\ \beta=(X^TX+\lambda I)^{-1}X^Ty\quad\#$$

Question 2

let $\hat{\beta}$ is the point estimator of β

then error $= \hat{\beta} - \beta$ and the corresponding MSE is

$$MSE(\hat{\beta}) = E[(\hat{\beta} - \beta)^2] = E[\{(\hat{\beta} - E[\hat{\beta}]) + (E[\hat{\beta}] - \beta)\}^2]$$

$$= E[(\hat{\beta} - E[\hat{\beta}])^2] + (E[\hat{\beta}] - \beta)^2$$

$$\triangleq Var(\hat{\beta}) + [Bias(\hat{\beta})]^2$$

where $Bias(\hat{\beta}) = E[\hat{\beta}] - \beta$ ※

Question 3

3-0. Learning Objectives

- to understand the process of data analysis using R programming
  - Exploratory data analysis
  - Linear model
    - Linear regression
    - Regularized regression
- to learn how to interpret the analytic results

3-1. Data Description Due to the importance of the excellent wine quality evaluation for the marketing, this study used Wine Quality Data Set from UCI Machine Learning repository to model wine quality based on physicochemical tests, ensuring the quality of wine market. We further selected "red vinho verde wine samples" for downstream analysis. The basic descriptions of this dataset are as follows:

- Number of Instances: 1599
- Number of Attributes: 11 + output attribute
  - Input variables (based on physicochemical tests):
    - fixed acidity
    - volatile acidity
    - citric acid
    - residual sugar
    - chlorides
    - free sulfur dioxide
    - total sulfur dioxide
    - density
    - pH
    - sulphates
    - alcohol
  - Output variable (based on sensory data):
    - quality (score between 0 and 10)

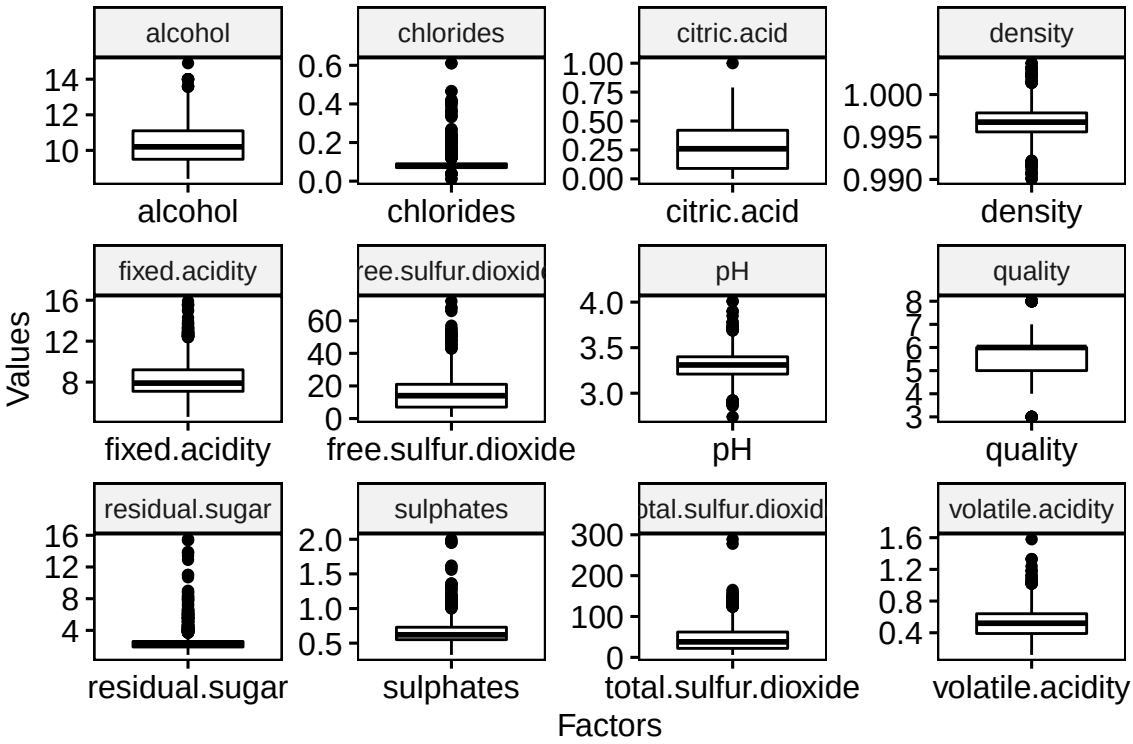
- Missing Attribute Values: None

3-2. Data Visualization In 'Data visualization' section, we provided a descriptive table and a boxplot for summary statistics, a heatmap and a scatter diagram for correlation analysis.

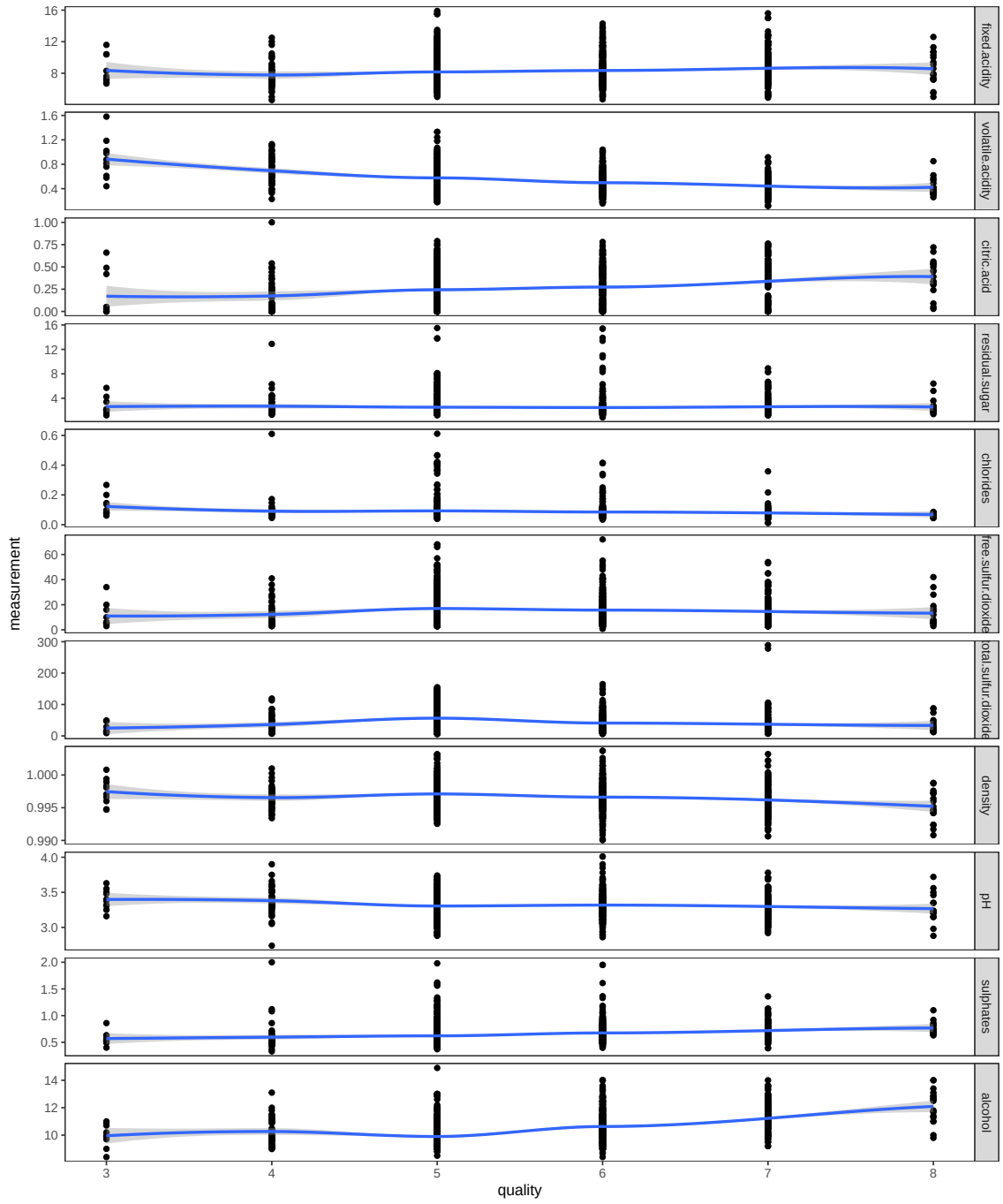
- **Summary statistics**

```
##                        Min. 1st Qu.   Median        Mean   3rd Qu.      Max.
## fixed.acidity        4.60000  7.1000  7.90000  8.31963727  9.200000  15.90000
## volatile.acidity     0.12000  0.3900  0.52000  0.52782051  0.640000   1.58000
## citric.acid          0.00000  0.0900  0.26000  0.27097561  0.420000   1.00000
## residual.sugar       0.90000  1.9000  2.20000  2.53880550  2.600000  15.50000
## chlorides            0.01200  0.0700  0.07900  0.08746654  0.090000   0.61100
## free.sulfur.dioxide  1.00000  7.0000 14.00000 15.87492183 21.000000  72.00000
## total.sulfur.dioxide 6.00000 22.0000 38.00000 46.46779237 62.000000 289.00000
## density              0.99007  0.9956  0.99675  0.99674668  0.997835   1.00369
## pH                   2.74000  3.2100  3.31000  3.31111320  3.400000   4.01000
## sulphates            0.33000  0.5500  0.62000  0.65814884  0.730000   2.00000
## alcohol              8.40000  9.5000 10.20000 10.42298311 11.100000  14.90000
## quality              3.00000  5.0000  6.00000  5.63602251  6.000000   8.00000
```

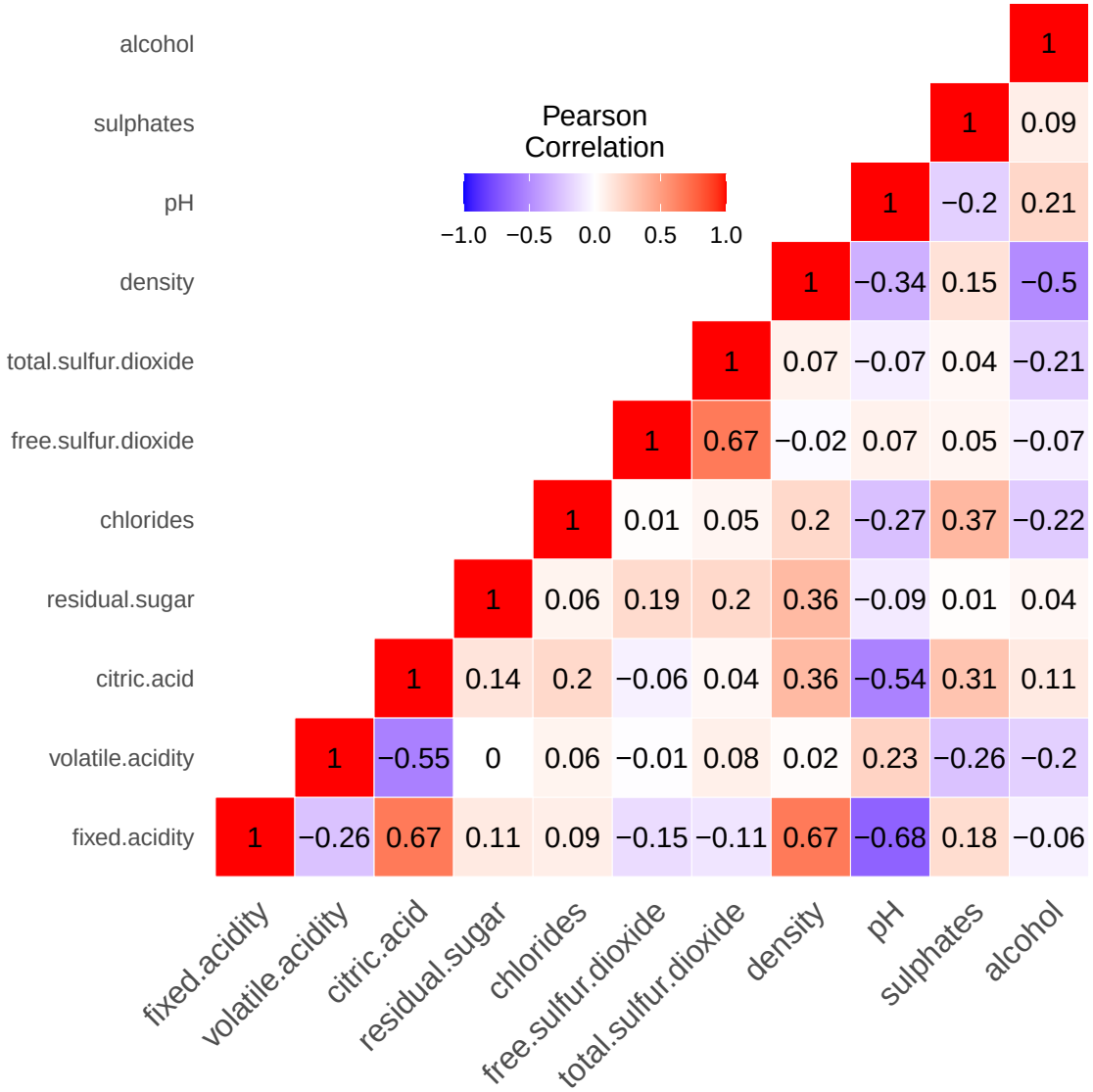
- **Boxplot**



- **Correlation analysis**
  - **Scatter diagram**: 11 variables against 1 outcome



– **Heatmap**: correlation relationship among 11 variables



- **Brief summary**
  - The quality of red wine was not uniformly distributed, ranged from 5 to 7.
  - The ***wine quality*** was strongly positively correlated to ***density*** and ***residual sugar***. On the contrary, a strong negative correlation between ***wind quality*** and ***alcohol*** was observed.

3-3. Linear Regression

Here, we used linear regression model using Forward, Backward, and Bi-direction methods to predict the **quality** of the red wine.

- **Training and Test sets : 80/20 split**

```
set.seed(1234)
data <- read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequ
    sep = ";")
n = nrow(data)
train_index = sample(n, ceiling(n * 0.8), replace = F)
```

```
train_set = data[train_index, ]
dim(train_set)
```

```
## [1] 1280   12
```

```
test_set = data[-train_index, ]
dim(test_set)
```

```
## [1] 319  12
```

- **By forward selection method**

```
model_full = lm(quality ~ ., data = train_set)
model_int = lm(quality ~ -., data = train_set)

scopeformula = formula(model_full)
scopeformula
```

```
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
```

```
fwd_sel = step(object = model_int, scope = scopeformula,
    direction = "forward")
```

```
## Start:  AIC=-555.46
## quality ~ -(fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol)
##
##                        Df Sum of Sq    RSS     AIC
## + alcohol               1   193.648 634.43 -894.42
## + volatile.acidity      1   134.091 693.98 -779.58
## + sulphates             1    56.199 771.88 -643.41
## + citric.acid           1    36.284 791.79 -610.81
## + total.sulfur.dioxide  1    25.901 802.17 -594.13
## + density               1    25.415 802.66 -593.36
## + chlorides             1    15.068 813.01 -576.96
## + fixed.acidity         1     9.695 818.38 -568.53
## <none>                              828.07 -555.46
## + free.sulfur.dioxide   1     1.129 826.95 -555.20
## + pH                    1     1.086 826.99 -555.14
## + residual.sugar        1     0.737 827.34 -554.60
##
## Step:  AIC=-894.42
## quality ~ alcohol
##
##                        Df Sum of Sq    RSS      AIC
## + volatile.acidity      1    83.005 551.42 -1071.91
## + sulphates             1    37.159 597.27  -969.68
## + citric.acid           1    20.493 613.93  -934.45
## + pH                    1    16.113 618.31  -925.35
## + fixed.acidity         1    15.849 618.58  -924.81
## + total.sulfur.dioxide  1     5.806 628.62  -904.19
## + density               1     4.898 629.53  -902.34
```

```
## <none>                              634.43  -894.42
## + chlorides             1     0.802 633.62  -894.04
## + free.sulfur.dioxide   1     0.026 634.40  -892.48
## + residual.sugar        1     0.015 634.41  -892.45
##
## Step:  AIC=-1071.91
## quality ~ alcohol + volatile.acidity
##
##                         Df Sum of Sq    RSS     AIC
## + sulphates              1   13.8174 537.60 -1102.4
## + total.sulfur.dioxide   1    4.6578 546.76 -1080.8
## + pH                     1    1.9323 549.49 -1074.4
## + fixed.acidity          1    1.8819 549.54 -1074.3
## + density                1    1.1616 550.26 -1072.6
## <none>                              551.42 -1071.9
## + chlorides              1    0.6595 550.76 -1071.4
## + citric.acid            1    0.3565 551.06 -1070.7
## + free.sulfur.dioxide    1    0.0727 551.35 -1070.1
## + residual.sugar         1    0.0109 551.41 -1069.9
##
## Step:  AIC=-1102.39
## quality ~ alcohol + volatile.acidity + sulphates
##
##                         Df Sum of Sq    RSS     AIC
## + chlorides              1    7.0971 530.51 -1117.4
## + total.sulfur.dioxide   1    5.7967 531.81 -1114.3
## + citric.acid            1    1.7499 535.85 -1104.6
## + fixed.acidity          1    0.9108 536.69 -1102.6
## <none>                              537.60 -1102.4
## + pH                     1    0.7890 536.81 -1102.3
## + free.sulfur.dioxide    1    0.2059 537.40 -1100.9
## + density                1    0.0951 537.51 -1100.6
## + residual.sugar         1    0.0148 537.59 -1100.4
##
## Step:  AIC=-1117.4
## quality ~ alcohol + volatile.acidity + sulphates + chlorides
##
##                         Df Sum of Sq    RSS     AIC
## + total.sulfur.dioxide   1    6.2460 524.26 -1130.6
## + pH                     1    2.1582 528.35 -1120.6
## + fixed.acidity          1    1.1231 529.38 -1118.1
## <none>                              530.51 -1117.4
## + citric.acid            1    0.4755 530.03 -1116.5
## + free.sulfur.dioxide    1    0.3254 530.18 -1116.2
## + density                1    0.1365 530.37 -1115.7
## + residual.sugar         1    0.1062 530.40 -1115.7
##
## Step:  AIC=-1130.56
## quality ~ alcohol + volatile.acidity + sulphates + chlorides +
##     total.sulfur.dioxide
##
##                        Df Sum of Sq    RSS     AIC
## + pH                    1   2.50884 521.75 -1134.7
## + free.sulfur.dioxide   1   2.21355 522.05 -1134.0
```

```
## + residual.sugar       1   0.88756 523.37 -1130.7
## <none>                             524.26 -1130.6
## + fixed.acidity        1   0.53263 523.73 -1129.9
## + citric.acid          1   0.23386 524.03 -1129.1
## + density              1   0.04860 524.21 -1128.7
##
## Step:  AIC=-1134.7
## quality ~ alcohol + volatile.acidity + sulphates + chlorides +
##     total.sulfur.dioxide + pH
##
##                       Df Sum of Sq    RSS     AIC
## + free.sulfur.dioxide  1    3.1993 518.55 -1140.6
## + citric.acid          1    1.9334 519.82 -1137.5
## <none>                             521.75 -1134.7
## + residual.sugar       1    0.6602 521.09 -1134.3
## + fixed.acidity        1    0.1836 521.57 -1133.2
## + density              1    0.0268 521.72 -1132.8
##
## Step:  AIC=-1140.58
## quality ~ alcohol + volatile.acidity + sulphates + chlorides +
##     total.sulfur.dioxide + pH + free.sulfur.dioxide
##
##                   Df Sum of Sq    RSS     AIC
## + citric.acid      1   1.32032 517.23 -1141.8
## <none>                         518.55 -1140.6
## + residual.sugar   1   0.42218 518.13 -1139.6
## + fixed.acidity    1   0.15703 518.40 -1139.0
## + density          1   0.01724 518.53 -1138.6
##
## Step:  AIC=-1141.84
## quality ~ alcohol + volatile.acidity + sulphates + chlorides +
##     total.sulfur.dioxide + pH + free.sulfur.dioxide + citric.acid
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         517.23 -1141.8
## + residual.sugar   1   0.68381 516.55 -1141.5
## + density          1   0.20777 517.02 -1140.3
## + fixed.acidity    1   0.08162 517.15 -1140.0
```

```
summary(fwd_sel)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     chlorides + total.sulfur.dioxide + pH + free.sulfur.dioxide +
##     citric.acid, data = train_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.69517 -0.36210 -0.03285  0.43262  1.95746
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.529310   0.498052   9.094  < 2e-16 ***
## alcohol               0.294996   0.018707  15.769  < 2e-16 ***
```

```
## volatile.acidity     -1.194813   0.127280  -9.387  < 2e-16 ***
## sulphates             0.909408   0.125303   7.258 6.83e-13 ***
## chlorides            -1.913482   0.438430  -4.364 1.38e-05 ***
## total.sulfur.dioxide -0.003461   0.000762  -4.541 6.12e-06 ***
## pH                   -0.493422   0.143952  -3.428 0.000628 ***
## free.sulfur.dioxide   0.005923   0.002350   2.521 0.011824 *
## citric.acid          -0.240574   0.133561  -1.801 0.071904 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6379 on 1271 degrees of freedom
## Multiple R-squared:  0.3754, Adjusted R-squared:  0.3714
## F-statistic: 95.48 on 8 and 1271 DF,  p-value: < 2.2e-16
```

```
FwdSelection_AIC = AIC(fwd_sel)
FwdSelection_AIC
```

```
## [1] 2492.644
```

- **By reverse selection method**

```
model_full = lm(quality ~ ., data = train_set)
scopeformula = formula(model_full)

back_sel = step(object = model_full, scope = scopeformula,
    direction = "backward")
```

```
## Start:  AIC=-1137.7
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## - density               1     0.018 516.50 -1139.7
## - fixed.acidity         1     0.059 516.54 -1139.5
## - residual.sugar        1     0.538 517.02 -1138.4
## <none>                              516.48 -1137.7
## - citric.acid           1     1.381 517.86 -1136.3
## - pH                    1     1.792 518.27 -1135.3
## - free.sulfur.dioxide   1     2.170 518.65 -1134.3
## - chlorides             1     6.897 523.38 -1122.7
## - total.sulfur.dioxide  1     7.707 524.19 -1120.7
## - sulphates             1    20.402 536.88 -1090.1
## - volatile.acidity      1    34.501 550.98 -1056.9
## - alcohol               1    39.418 555.90 -1045.6
##
## Step:  AIC=-1139.65
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS      AIC
## - fixed.acidity         1     0.048 516.55 -1141.53
## - residual.sugar        1     0.650 517.15 -1140.04
## <none>                              516.50 -1139.65
```

```
## - citric.acid           1     1.383 517.88 -1138.23
## - free.sulfur.dioxide   1     2.213 518.71 -1136.18
## - pH                    1     2.936 519.44 -1134.40
## - chlorides             1     7.022 523.52 -1124.37
## - total.sulfur.dioxide  1     7.776 524.28 -1122.52
## - sulphates             1    21.529 538.03 -1089.38
## - volatile.acidity      1    35.173 551.67 -1057.33
## - alcohol               1    97.449 613.95  -920.42
##
## Step:  AIC=-1141.53
## quality ~ volatile.acidity + citric.acid + residual.sugar + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates +
##     alcohol
##
##                         Df Sum of Sq    RSS     AIC
## - residual.sugar         1     0.684 517.23 -1141.8
## <none>                               516.55 -1141.5
## - citric.acid            1     1.582 518.13 -1139.6
## - free.sulfur.dioxide    1     2.263 518.81 -1137.9
## - pH                     1     4.697 521.25 -1132.0
## - chlorides              1     7.929 524.48 -1124.0
## - total.sulfur.dioxide   1     8.765 525.31 -1122.0
## - sulphates              1    21.871 538.42 -1090.5
## - volatile.acidity       1    36.419 552.97 -1056.3
## - alcohol                1    98.563 615.11  -920.0
##
## Step:  AIC=-1141.84
## quality ~ volatile.acidity + citric.acid + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
##
##                         Df Sum of Sq    RSS      AIC
## <none>                               517.23 -1141.84
## - citric.acid            1     1.320 518.55 -1140.58
## - free.sulfur.dioxide    1     2.586 519.82 -1137.45
## - pH                     1     4.781 522.01 -1132.06
## - chlorides              1     7.752 524.98 -1124.80
## - total.sulfur.dioxide   1     8.393 525.63 -1123.23
## - sulphates              1    21.436 538.67 -1091.86
## - volatile.acidity       1    35.861 553.09 -1058.03
## - alcohol                1   101.195 618.43  -915.12
```

```
summary(back_sel)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + citric.acid + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates +
##     alcohol, data = train_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.69517 -0.36210 -0.03285  0.43262  1.95746
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           4.529310   0.498052   9.094  < 2e-16 ***
## volatile.acidity     -1.194813   0.127280  -9.387  < 2e-16 ***
## citric.acid          -0.240574   0.133561  -1.801 0.071904 .
## chlorides            -1.913482   0.438430  -4.364 1.38e-05 ***
## free.sulfur.dioxide   0.005923   0.002350   2.521 0.011824 *
## total.sulfur.dioxide -0.003461   0.000762  -4.541 6.12e-06 ***
## pH                   -0.493422   0.143952  -3.428 0.000628 ***
## sulphates             0.909408   0.125303   7.258 6.83e-13 ***
## alcohol               0.294996   0.018707  15.769  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6379 on 1271 degrees of freedom
## Multiple R-squared:  0.3754, Adjusted R-squared:  0.3714
## F-statistic: 95.48 on 8 and 1271 DF,  p-value: < 2.2e-16
```

```
BackSelection_AIC = AIC(back_sel)
BackSelection_AIC
```

```
## [1] 2492.644
```

- **By bidirectional(both) selection method**

```
model_full = lm(quality ~ ., data = train_set)
scopeformula = formula(model_full)

both_sel = step(object = model_full, scope = scopeformula,
    direction = "both")
```

```
## Start:  AIC=-1137.7
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## - density               1     0.018 516.50 -1139.7
## - fixed.acidity         1     0.059 516.54 -1139.5
## - residual.sugar        1     0.538 517.02 -1138.4
## <none>                              516.48 -1137.7
## - citric.acid           1     1.381 517.86 -1136.3
## - pH                    1     1.792 518.27 -1135.3
## - free.sulfur.dioxide   1     2.170 518.65 -1134.3
## - chlorides             1     6.897 523.38 -1122.7
## - total.sulfur.dioxide  1     7.707 524.19 -1120.7
## - sulphates             1    20.402 536.88 -1090.1
## - volatile.acidity      1    34.501 550.98 -1056.9
## - alcohol               1    39.418 555.90 -1045.6
##
## Step:  AIC=-1139.65
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS      AIC
## - fixed.acidity         1     0.048 516.55 -1141.53
```

```
## - residual.sugar        1     0.650 517.15 -1140.04
## <none>                              516.50 -1139.65
## - citric.acid           1     1.383 517.88 -1138.23
## + density               1     0.018 516.48 -1137.70
## - free.sulfur.dioxide   1     2.213 518.71 -1136.18
## - pH                    1     2.936 519.44 -1134.40
## - chlorides             1     7.022 523.52 -1124.37
## - total.sulfur.dioxide  1     7.776 524.28 -1122.52
## - sulphates             1    21.529 538.03 -1089.38
## - volatile.acidity      1    35.173 551.67 -1057.33
## - alcohol               1    97.449 613.95  -920.42
##
## Step:  AIC=-1141.53
## quality ~ volatile.acidity + citric.acid + residual.sugar + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates +
##     alcohol
##
##                         Df Sum of Sq    RSS     AIC
## - residual.sugar         1     0.684 517.23 -1141.8
## <none>                              516.55 -1141.5
## + fixed.acidity          1     0.048 516.50 -1139.7
## - citric.acid            1     1.582 518.13 -1139.6
## + density                1     0.008 516.54 -1139.5
## - free.sulfur.dioxide    1     2.263 518.81 -1137.9
## - pH                     1     4.697 521.25 -1132.0
## - chlorides              1     7.929 524.48 -1124.0
## - total.sulfur.dioxide   1     8.765 525.31 -1122.0
## - sulphates              1    21.871 538.42 -1090.5
## - volatile.acidity       1    36.419 552.97 -1056.3
## - alcohol                1    98.563 615.11  -920.0
##
## Step:  AIC=-1141.84
## quality ~ volatile.acidity + citric.acid + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
##
##                         Df Sum of Sq    RSS      AIC
## <none>                              517.23 -1141.84
## + residual.sugar         1     0.684 516.55 -1141.53
## - citric.acid            1     1.320 518.55 -1140.58
## + density                1     0.208 517.02 -1140.35
## + fixed.acidity          1     0.082 517.15 -1140.04
## - free.sulfur.dioxide    1     2.586 519.82 -1137.45
## - pH                     1     4.781 522.01 -1132.06
## - chlorides              1     7.752 524.98 -1124.80
## - total.sulfur.dioxide   1     8.393 525.63 -1123.23
## - sulphates              1    21.436 538.67 -1091.86
## - volatile.acidity       1    35.861 553.09 -1058.03
## - alcohol                1   101.195 618.43  -915.12
```

```
summary(both_sel)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + citric.acid + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates +
```

```
##     alcohol, data = train_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.69517 -0.36210 -0.03285  0.43262  1.95746
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.529310   0.498052   9.094  < 2e-16 ***
## volatile.acidity     -1.194813   0.127280  -9.387  < 2e-16 ***
## citric.acid          -0.240574   0.133561  -1.801 0.071904 .
## chlorides            -1.913482   0.438430  -4.364 1.38e-05 ***
## free.sulfur.dioxide   0.005923   0.002350   2.521 0.011824 *
## total.sulfur.dioxide -0.003461   0.000762  -4.541 6.12e-06 ***
## pH                   -0.493422   0.143952  -3.428 0.000628 ***
## sulphates             0.909408   0.125303   7.258 6.83e-13 ***
## alcohol               0.294996   0.018707  15.769  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6379 on 1271 degrees of freedom
## Multiple R-squared:  0.3754, Adjusted R-squared:  0.3714
## F-statistic: 95.48 on 8 and 1271 DF,  p-value: < 2.2e-16
```

```
BidirSelection_AIC = AIC(both_sel)
BidirSelection_AIC
```

```
## [1] 2492.644
```

- **Comparison of three models using AIC**

```
AIC_compare = data.frame(FwdSelection = FwdSelection_AIC,
    BackSelection = BackSelection_AIC, BidirSelection = BidirSelection_AIC)
rownames(AIC_compare) = c("AIC")
AIC_compare
```

```
##     FwdSelection BackSelection BidirSelection
## AIC     2492.644      2492.644       2492.644
```

- **Prediction results**

```
##    Step_forward Step_backward Step_both
## 1      5.045331      5.045331  5.045331
## 2      5.136983      5.136983  5.136983
## 3      5.211085      5.211085  5.211085
## 5      5.045331      5.045331  5.045331
## 7      5.113628      5.113628  5.113628
## 12     5.629172      5.629172  5.629172
```

- **Mean Squared Error (MSE) calculation**

```
##           Model       MSE
## 1  Step_forward 0.4736429
```

```
## 2 Step_backward 0.4736429
## 3     Step_both 0.4736429
```

- **Feature selection** Because AIC scores and MSEs of all 3 models are same, these models are suitable for this dataset. Final variables selected from these models based on their p-value: ***alcohol***, ***volatile.acidity***, ***sulphates***, ***total.sulfur.dioxide***, ***chlorides***, ***pH*** and ***free.sulfur.dioxide***.
- **Final prediction model** Then, we trained the linear model again using these selected features.

```
selected_features <- c("alcohol", "volatile.acidity",
    "sulphates", "total.sulfur.dioxide", "chlorides",
    "pH", "free.sulfur.dioxide", "quality")
train_set_final <- train_set[, selected_features]
model_final = lm(quality ~ ., data = train_set_final)
summary(model_final)
```

```
##
## Call:
## lm(formula = quality ~ ., data = train_set_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73240 -0.35855 -0.03705  0.45022  1.95217
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.1074477  0.4399351   9.336  < 2e-16 ***
## alcohol                0.2883442  0.0183551  15.709  < 2e-16 ***
## volatile.acidity      -1.0816035  0.1107768  -9.764  < 2e-16 ***
## sulphates              0.8961539  0.1251968   7.158 1.38e-12 ***
## total.sulfur.dioxide -0.0036630  0.0007544  -4.856 1.35e-06 ***
## chlorides             -2.0512615  0.4320859  -4.747 2.29e-06 ***
## pH                    -0.3766086  0.1286299  -2.928  0.00347 **
## free.sulfur.dioxide    0.0065220  0.0023281   2.801  0.00516 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6385 on 1272 degrees of freedom
## Multiple R-squared:  0.3738, Adjusted R-squared:  0.3703
## F-statistic: 108.5 on 7 and 1272 DF,  p-value: < 2.2e-16
```

Therefore, the prediction can be obtained from the following equation:

$$\hat{y} = 4.1 + 0.288 * x_{alcohol} - 1.082 * x_{volatile.acidity} + 0.896 * x_{sulphates}$$

$$-0.004 * x_{total.sulfur.dioxide} - 2.0512 * x_{chlorides}$$

$$-0.377 * x_{pH} + 0.007 * x_{free.sulfur.dioxide}$$

3-4. Regularized Regression

Here, we used three regularized regression models including Ridge, Lasso and Elastic-Net, to predict the **quality** of the red wine.

- **Ridge Regression**

```
ridge_model <- glmnet(x = as.matrix(train_set[, -12]),
    y = as.matrix(train_set$quality), alpha = 0)  #alpha =0 for ridge
summary(ridge_model)
```

```
##           Length Class     Mode
## a0         100   -none-    numeric
## beta      1100   dgCMatrix S4
## df         100   -none-    numeric
## dim          2   -none-    numeric
## lambda     100   -none-    numeric
## dev.ratio  100   -none-    numeric
## nulldev      1   -none-    numeric
## npasses      1   -none-    numeric
## jerr         1   -none-    numeric
## offset       1   -none-    logical
## call         4   -none-    call
## nobs         1   -none-    numeric
```

- **Lasso Regression**

```
lasso_model <- glmnet(x = as.matrix(train_set[, -12]),
    y = as.matrix(train_set$quality), alpha = 1)  #alpha =1 for lasso
summary(lasso_model)
```

```
##           Length Class     Mode
## a0         70    -none-    numeric
## beta      770    dgCMatrix S4
## df         70    -none-    numeric
## dim         2    -none-    numeric
## lambda     70    -none-    numeric
## dev.ratio  70    -none-    numeric
## nulldev     1    -none-    numeric
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
## offset      1    -none-    logical
## call        4    -none-    call
## nobs        1    -none-    numeric
```

- **Elastic-Net Regression**

```
elastic_model <- glmnet(x = as.matrix(train_set[, -12]),
    y = as.matrix(train_set$quality), alpha = 0.5)  #alpha =0.5 (range: 0-1) for elastic
summary(elastic_model)
```

```
##           Length Class     Mode
## a0         71    -none-    numeric
## beta      781    dgCMatrix S4
## df         71    -none-    numeric
## dim         2    -none-    numeric
## lambda     71    -none-    numeric
## dev.ratio  71    -none-    numeric
## nulldev     1    -none-    numeric
```

```
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
## offset      1    -none-    logical
## call        4    -none-    call
## nobs        1    -none-    numeric
```

- **Prediction results**

```
##       Ridge    Lasso  Elastic
## 1  5.079855 5.215961 5.171407
## 2  5.128504 5.171887 5.109886
## 3  5.209950 5.275295 5.231433
## 5  5.079855 5.215961 5.171407
## 7  5.135643 5.286102 5.206312
## 12 5.699006 5.711564 5.715147
```

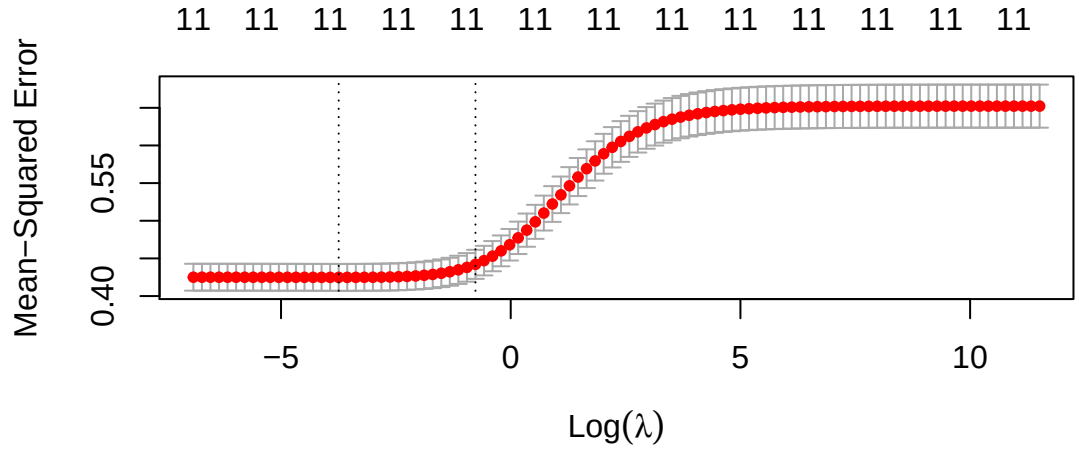
- **Mean Squared Error (MSE) calculation**

```
##     Model       MSE
## 1   Ridge 0.4681614
## 3 Elastic 0.4818979
## 2   Lasso 0.4980975
```

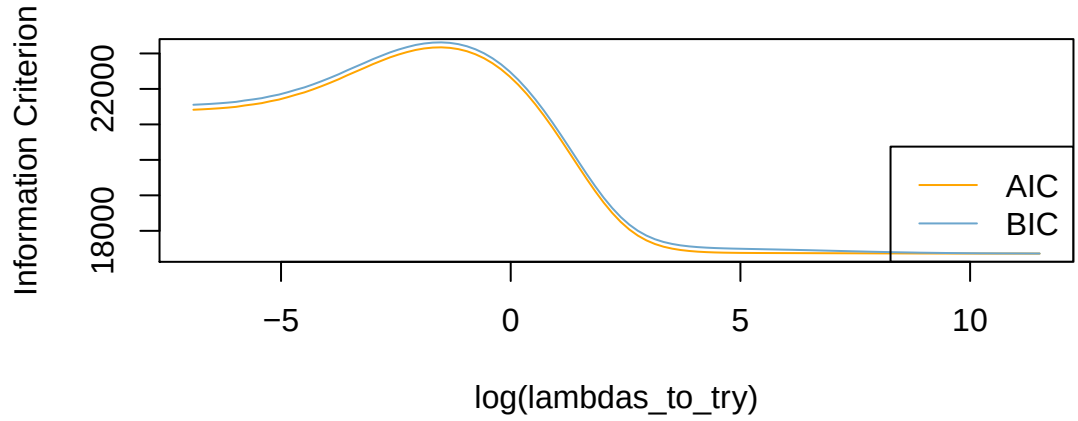
MSE is the one of indicators for evaluating model performance. Lower MSE value reveals that better performance of the model with selected variables.

According the MSE calculation, Ridge regression model had the best prediction performance (MSE = 0.468), compared with other two models.

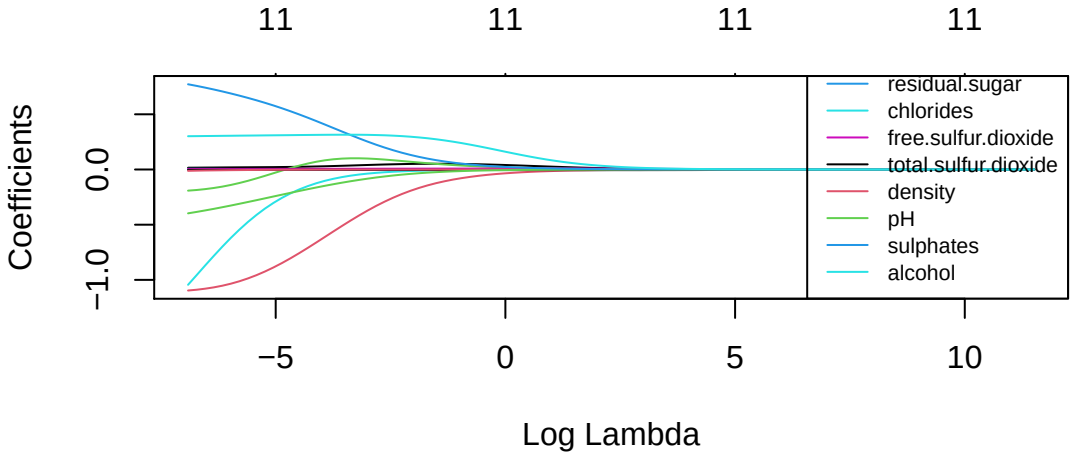
- **Parameter turning** Here, we tested the parameters, such as lambda and that are suitable for ridge regression and lasso regression.
  - *Ridge regression*
    - Perform 10-fold cross-validation to select lambda



    - Plot information criteria against tried values of lambdas



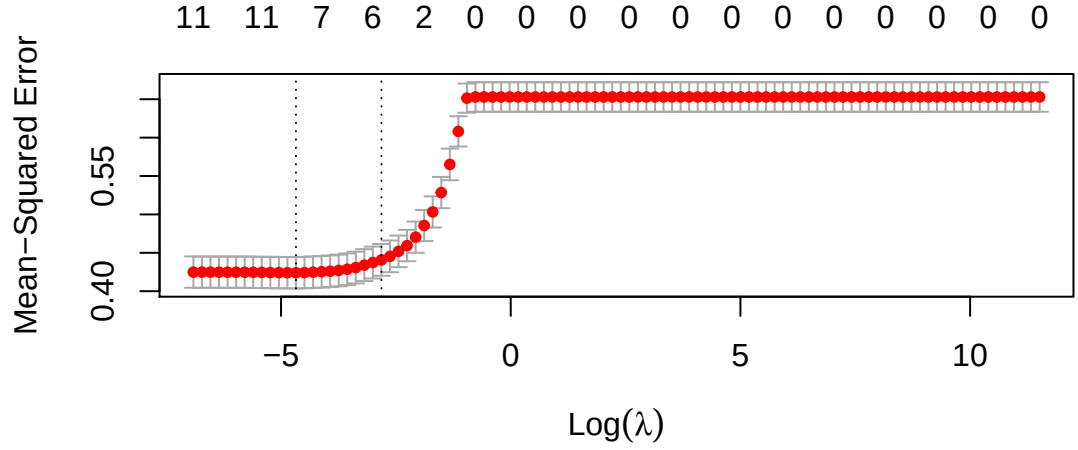
∗ See how increasing lambda shrinks the coefficients



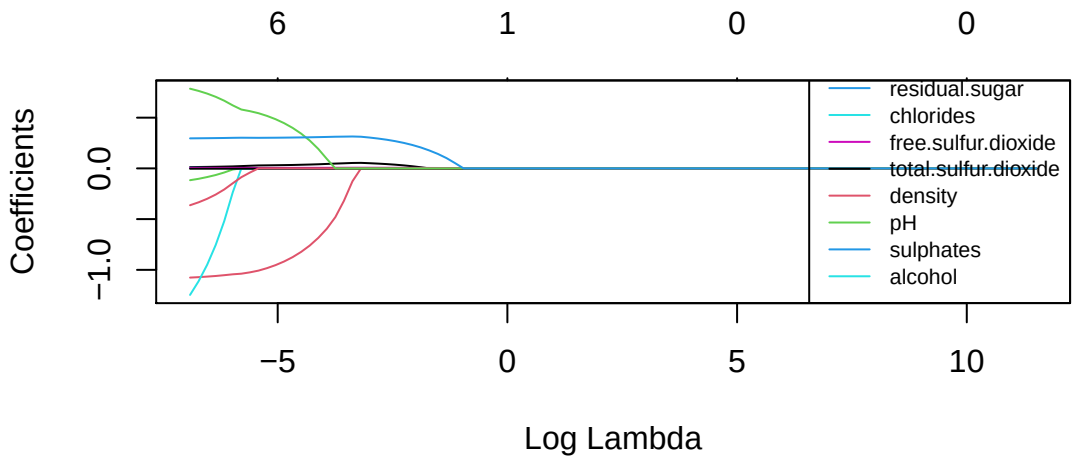
Each line shows coefficients for one variables under different lambdas. The higher the lambda, the more the coefficients are shrinked towards zero.

– *Lasso regression*

∗ Perform 10-fold cross-validation to select lambda



∗ See how increasing lambda shrinks the coefficients



Each line shows coefficients for one variables under different lambdas. The higher the lambda, the more the coefficients are shrinked towards zero.

- Model comparisons

```
##                          R-squared
## ridge cross-validated 0.3604093
## ridge AIC             0.3318570
## ridge BIC             0.3318570
## lasso cross_validated 0.3584408
```

- **Brief summary**
  - According to the shrinkage plots of Lasso and ridge regression, the former can set some coefficients to zero, thus performing variable selection, while the latter cannot.
  - The MSE plot of Lasso suggested that 6-10 features influence the response, which is consistent with the results obtained from linear regression.
  - Our results suggest that the performance of ridge is slightly better than Lasso. Since 7 out of total 11 features were selected by linear and lasso regression, ridge tends to do well if a large number of significant parameters while Lasso works well when only a few predictors actually impact the response.

3-5. Discussion

There are three main issues that need to be further discussed.

- Converting ordinal outcome to binary outcome and then applying logistic regression may be the suitable way to tackle the issue of the unbalanced outcome (i.e. wine quality).
- In 'Parameter turning' section, we only applied whole dataset to test turning process. Using hold-on samples subset from training data to turn the parameters may be the most proper approach.
- Ridge regression assumes the predictors are standardized and the response is centered; hence, future work should take this step into consideration to minimize the error of computing precision and maximize the model performance.

Session Information

```
## - Session info ---------------------------------------------------------------
```

```
##  setting  value
##  version  R version 4.0.3 (2020-10-10)
##  os       macOS Big Sur 10.16
##  system   x86_64, darwin17.0
##  ui       X11
##  language (EN)
##  collate  zh_TW.UTF-8
##  ctype    zh_TW.UTF-8
##  tz       Asia/Taipei
##  date     2021-03-31
##
## - Packages -------------------------------------------------------------------
##  package     * version date       lib source
##  abind         1.4-5   2016-07-21 [1] CRAN (R 4.0.2)
##  assertthat    0.2.1   2019-03-21 [1] CRAN (R 4.0.2)
##  backports     1.2.1   2020-12-09 [1] CRAN (R 4.0.2)
##  broom         0.7.3   2020-12-16 [1] CRAN (R 4.0.2)
##  car           3.0-10  2020-09-29 [1] CRAN (R 4.0.2)
##  carData       3.0-4   2020-05-22 [1] CRAN (R 4.0.2)
##  cellranger    1.1.0   2016-07-27 [1] CRAN (R 4.0.2)
##  cli           2.2.0   2020-11-20 [1] CRAN (R 4.0.2)
##  codetools     0.2-16  2018-12-24 [1] CRAN (R 4.0.3)
##  colorspace    2.0-0   2020-11-11 [1] CRAN (R 4.0.2)
##  crayon        1.3.4   2017-09-16 [1] CRAN (R 4.0.2)
##  curl          4.3     2019-12-02 [1] CRAN (R 4.0.1)
##  data.table    1.13.6  2020-12-30 [1] CRAN (R 4.0.2)
##  DBI           1.1.1   2021-01-15 [1] CRAN (R 4.0.2)
##  digest        0.6.27  2020-10-24 [1] CRAN (R 4.0.2)
##  dplyr       * 1.0.3   2021-01-15 [1] CRAN (R 4.0.2)
##  ellipsis      0.3.1   2020-05-15 [1] CRAN (R 4.0.2)
##  evaluate      0.14    2019-05-28 [1] CRAN (R 4.0.1)
##  fansi         0.4.2   2021-01-15 [1] CRAN (R 4.0.2)
##  farver        2.0.3   2020-01-16 [1] CRAN (R 4.0.2)
##  forcats       0.5.0   2020-03-01 [1] CRAN (R 4.0.2)
##  foreach       1.5.1   2020-10-15 [1] CRAN (R 4.0.2)
##  foreign       0.8-80  2020-05-24 [1] CRAN (R 4.0.3)
##  formatR       1.7     2019-06-11 [1] CRAN (R 4.0.2)
##  generics      0.1.0   2020-10-31 [1] CRAN (R 4.0.2)
##  ggplot2     * 3.3.3   2020-12-30 [1] CRAN (R 4.0.2)
##  ggpubr      * 0.4.0   2020-06-27 [1] CRAN (R 4.0.2)
##  ggsignif      0.6.0   2019-08-08 [1] CRAN (R 4.0.2)
##  glmnet      * 4.1-1   2021-02-21 [1] CRAN (R 4.0.2)
##  glue          1.4.2   2020-08-27 [1] CRAN (R 4.0.2)
##  gtable        0.3.0   2019-03-25 [1] CRAN (R 4.0.2)
##  haven         2.3.1   2020-06-01 [1] CRAN (R 4.0.2)
##  hms           1.0.0   2021-01-13 [1] CRAN (R 4.0.2)
##  htmltools     0.5.1.1 2021-01-22 [1] CRAN (R 4.0.2)
##  iterators     1.0.13  2020-10-15 [1] CRAN (R 4.0.2)
##  knitr       * 1.31    2021-01-27 [1] CRAN (R 4.0.2)
##  labeling      0.4.2   2020-10-20 [1] CRAN (R 4.0.2)
##  lattice       0.20-41 2020-04-02 [1] CRAN (R 4.0.3)
##  lifecycle     0.2.0   2020-03-06 [1] CRAN (R 4.0.2)
##  magrittr      2.0.1   2020-11-17 [1] CRAN (R 4.0.2)
##  Matrix      * 1.2-18  2019-11-27 [1] CRAN (R 4.0.3)
```

```
##  mgcv          1.8-33  2020-08-27 [1] CRAN (R 4.0.3)
##  mnormt        2.0.2   2020-09-01 [1] CRAN (R 4.0.2)
##  munsell       0.5.0   2018-06-12 [1] CRAN (R 4.0.2)
##  nlme          3.1-149 2020-08-23 [1] CRAN (R 4.0.3)
##  openxlsx      4.2.3   2020-10-27 [1] CRAN (R 4.0.2)
##  pillar        1.4.7   2020-11-20 [1] CRAN (R 4.0.2)
##  pkgconfig     2.0.3   2019-09-22 [1] CRAN (R 4.0.2)
##  plyr          1.8.6   2020-03-03 [1] CRAN (R 4.0.2)
##  psych       * 2.1.3   2021-03-27 [1] CRAN (R 4.0.3)
##  purrr         0.3.4   2020-04-17 [1] CRAN (R 4.0.2)
##  R6            2.5.0   2020-10-28 [1] CRAN (R 4.0.2)
##  Rcpp          1.0.6   2021-01-15 [1] CRAN (R 4.0.2)
##  readxl        1.3.1   2019-03-13 [1] CRAN (R 4.0.2)
##  reshape2    * 1.4.4   2020-04-09 [1] CRAN (R 4.0.2)
##  rio           0.5.16  2018-11-26 [1] CRAN (R 4.0.2)
##  rlang         0.4.10  2020-12-30 [1] CRAN (R 4.0.2)
##  rmarkdown     2.7     2021-02-19 [1] CRAN (R 4.0.2)
##  rstatix       0.6.0   2020-06-18 [1] CRAN (R 4.0.2)
##  scales        1.1.1   2020-05-11 [1] CRAN (R 4.0.2)
##  sessioninfo   1.1.1   2018-11-05 [1] CRAN (R 4.0.2)
##  shape         1.4.5   2020-09-13 [1] CRAN (R 4.0.2)
##  stringi       1.5.3   2020-09-09 [1] CRAN (R 4.0.2)
##  stringr       1.4.0   2019-02-10 [1] CRAN (R 4.0.2)
##  survival      3.2-7   2020-09-28 [1] CRAN (R 4.0.3)
##  tibble        3.0.5   2021-01-15 [1] CRAN (R 4.0.2)
##  tidyr       * 1.1.2   2020-08-27 [1] CRAN (R 4.0.2)
##  tidyselect    1.1.0   2020-05-11 [1] CRAN (R 4.0.2)
##  tmvnsim       1.0-2   2016-12-15 [1] CRAN (R 4.0.2)
##  vctrs         0.3.6   2020-12-17 [1] CRAN (R 4.0.2)
##  withr         2.4.0   2021-01-16 [1] CRAN (R 4.0.2)
##  xfun          0.20    2021-01-06 [1] CRAN (R 4.0.2)
##  yaml          2.2.1   2020-02-01 [1] CRAN (R 4.0.2)
##  zip           2.1.1   2020-08-27 [1] CRAN (R 4.0.2)
##
## [1] /Library/Frameworks/R.framework/Versions/4.0/Resources/library
```