msia 400 lab 2 Finn Qiao

Lab 2 for MSIA 400

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_knit$set(root.dir = "/Users/finn/MSIA/msia400-finn")
options(warn=-1)
```

```
wine_df <- read.table('redwine.txt', header = TRUE)
dim(wine_df)
```

```
## [1] 1599   12
```

```
head(wine_df)
```

```
##   QA   FA   VA   CA  RS    CH FS SD     DE   PH   SU  AL
## 1  5  7.4 0.70 0.00 1.9 0.076 11 34 0.9978 3.51 0.56 9.4
## 2  5  7.8 0.88 0.00 2.6 0.098 25 67 0.9968 3.20 0.68 9.8
## 3  5  7.8 0.76 0.04 2.3 0.092 15 54 0.9970 3.26 0.65 9.8
## 4  6 11.2 0.28 0.56 1.9 0.075 17 60 0.9980 3.16 0.58 9.8
## 5  5  7.4 0.70 0.00 1.9 0.076 11 34 0.9978 3.51 0.56 9.4
## 6  5  7.4 0.66 0.00 1.8 0.075 13 40 0.9978 3.51 0.56 9.4
```

Problem 1

```
mean(wine_df$RS, na.rm = TRUE)
```

```
## [1] 2.537952
```

```
mean(wine_df$SD, na.rm = TRUE)
```

```
## [1] 46.29836
```

Problem 2

```
SD <- wine_df$SD[!is.na(wine_df$SD)]
FS <- wine_df$FS[!is.na(wine_df$SD)]
lmfit <- lm(SD ~ FS)
summary(lmfit)
```

```
##
## Call:
## lm(formula = SD ~ FS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.489 -13.530  -7.155   7.252 197.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.18551    1.11502   11.82   <2e-16 ***
```

```
## FS           2.08608    0.05867   35.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.39 on 1580 degrees of freedom
## Multiple R-squared:  0.4445, Adjusted R-squared:  0.4441
## F-statistic:  1264 on 1 and 1580 DF,  p-value: < 2.2e-16
```

Problem 3

```
missing_SD <- subset(wine_df, complete.cases(wine_df$SD) == FALSE)
missing_SD$SD <- predict(lmfit, missing_SD)
```

```
missing_SD_rows <- as.numeric(rownames(missing_SD))
wine_df_imputed <- wine_df
wine_df_imputed[missing_SD_rows,] <- missing_SD
mean(wine_df_imputed$SD)
```

```
## [1] 46.30182
```

After imputation of SD, the average value is 46.3.

Problem 4

```
avg.imp <- function(a,avg) {
  missing <- is.na(a)
  imputed <- a
  imputed[missing] <- avg
  return (imputed)
}

wine_df_imputed$RS <- avg.imp(wine_df$RS, mean(wine_df$RS, na.rm = TRUE))
mean(wine_df_imputed$RS)
```

```
## [1] 2.537952
```

After imputation of RS, the average is 2.54.

Problem 5

```
winemodel <- lm(QA ~ FA + VA + CA + RS + CH + FS + SD + DE + PH + SU + AL, wine_df_imputed)
summary(winemodel)
```

```
##
## Call:
## lm(formula = QA ~ FA + VA + CA + RS + CH + FS + SD + DE + PH +
##     SU + AL, data = wine_df_imputed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.78010 -0.36249 -0.06331  0.44595  1.98828
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.720e+01  1.782e+01   2.649 0.008151 **
## FA           6.841e-02  1.872e-02   3.654 0.000267 ***
## VA          -1.098e+00  1.213e-01  -9.053  < 2e-16 ***
## CA          -1.789e-01  1.474e-01  -1.214 0.224954
## RS           2.593e-02  1.419e-02   1.827 0.067944 .
## CH          -1.631e+00  4.097e-01  -3.982 7.14e-05 ***
## FS           3.530e-03  2.159e-03   1.635 0.102262
## SD          -2.855e-03  7.248e-04  -3.939 8.54e-05 ***
## DE          -4.482e+01  1.789e+01  -2.505 0.012329 *
## PH           3.600e-02  4.409e-02   0.816 0.414413
## SU           9.449e-01  1.136e-01   8.321  < 2e-16 ***
## AL           2.470e-01  2.265e-02  10.906  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6491 on 1587 degrees of freedom
## Multiple R-squared:  0.3584, Adjusted R-squared:  0.354
## F-statistic:  80.6 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Problem 6

Based on the coefficients above, the greatest p value belongs to PH. It has a p value of 0.414480 and is highly statistically insignificant.

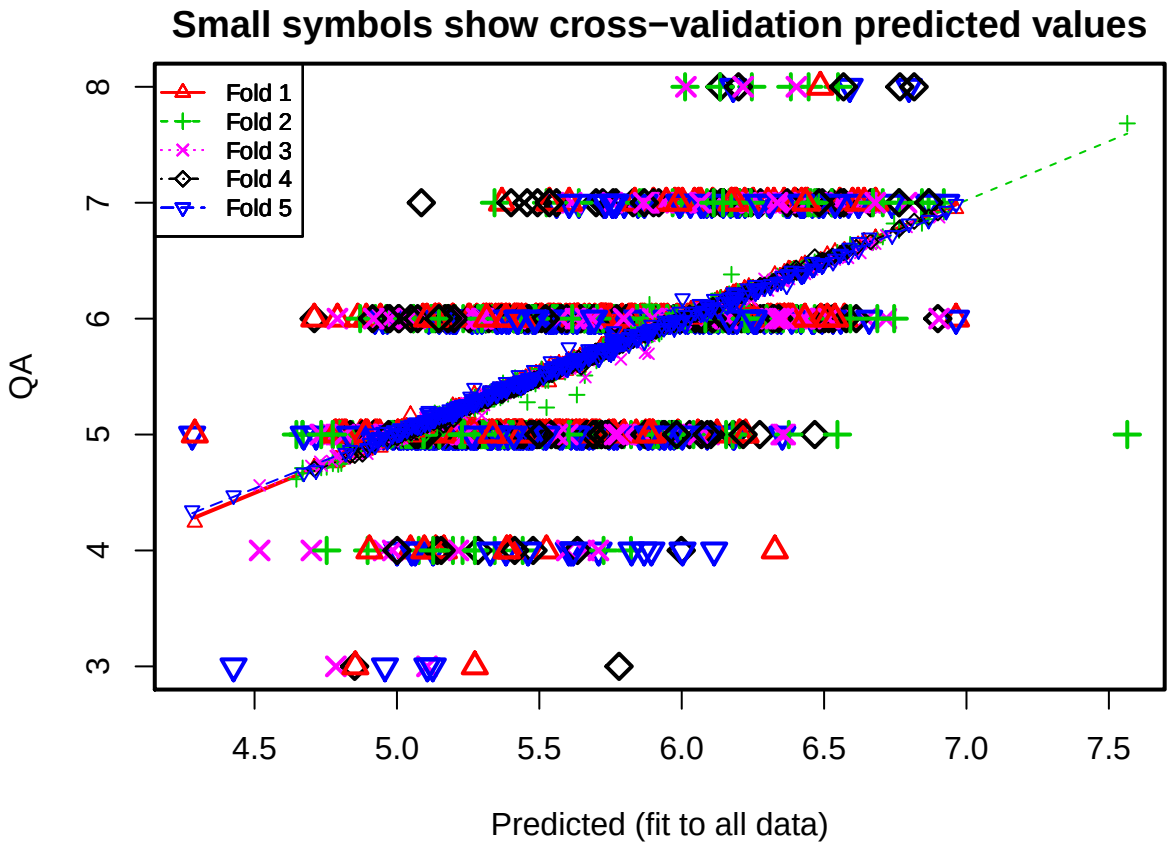
Problem 7

```
library('DAAG')
```

```
## Loading required package: lattice
```

The DAAG package was imported to run the cv.lm fit for cross validation. 5 folds was passed in as the parameter.

```
cv.lm(wine_df_imputed, winemodel, m=5)
```



Mean square is 0.426.

Problem 8

```
ph_mean <- mean(wine_df_imputed$PH)
ph_sd <- sd(wine_df_imputed$PH)
```

```
ph_lower_bound <- ph_mean - 3 * ph_sd
ph_upper_bound <- ph_mean + 3 * ph_sd
```

```
redwine2 <- wine_df_imputed[with(wine_df_imputed, !(PH < ph_lower_bound | PH > ph_upper_bound)),]
dim(redwine2)
```

```
## [1] 1580   12
```

19 rows were dropped.

Problem 9

```
winemodel <- lm(QA ~ FA + VA + CA + RS + CH + FS + SD + DE + PH + SU + AL, redwine2)
summary(winemodel)
```

```
##
## Call:
## lm(formula = QA ~ FA + VA + CA + RS + CH + FS + SD + DE + PH +
##     SU + AL, data = redwine2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6893 -0.3634 -0.0437  0.4522  2.0127
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.90e+01   2.12e+01    0.90    0.370
## FA           2.46e-02   2.60e-02    0.95    0.344
## VA          -1.07e+00   1.22e-01   -8.79  < 2e-16 ***
## CA          -1.78e-01   1.48e-01   -1.20    0.230
## RS           1.30e-02   1.50e-02    0.87    0.387
## CH          -1.90e+00   4.21e-01   -4.52  6.6e-06 ***
## FS           4.42e-03   2.18e-03    2.03    0.043 *
## SD          -3.14e-03   7.38e-04   -4.26  2.2e-05 ***
## DE          -1.50e+01   2.17e+01   -0.69    0.489
## PH          -4.25e-01   1.93e-01   -2.20    0.028 *
## SU           9.13e-01   1.15e-01    7.95  3.5e-15 ***
## AL           2.83e-01   2.66e-02   10.65  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1568 degrees of freedom
## Multiple R-squared:  0.363,  Adjusted R-squared:  0.358
## F-statistic: 81.2 on 11 and 1568 DF,  p-value: <2e-16
```

The five most significant attributes seem to be VA, CH, SD, SU, and AL.

The previous model with the outliers included seemed to have more significant predictors. The R^2 and F statistic are slightly higher for this new model which suggests a slightly better fit. This later model has the same most significant variables and has less variables in the model which suggests a more stable and better fit.