Assignment 2

Spencer Moon

10/28/2017

Data was loading using the following code:

```
library(tidyverse)

redwine <- read.table('redwine.txt', header = TRUE, sep = "\t", na.strings = 'NA')
```

Problem 1

The averages of RS and SD without the missing values are as follows:

```
rs_avg <- mean(redwine$RS, na.rm = TRUE)
sd_avg <- mean(redwine$SD, na.rm = TRUE)

# Print averages
paste('RS average:', round(rs_avg, digits = 2))
```

```
## [1] "RS average: 2.54"
```

```
paste('SD average:', round(sd_avg, digits = 2))
```

```
## [1] "SD average: 46.3"
```

Problem 2

The coefficients of a linear model between SD and FS are as follows:

```
# Creat SD and FS vectors
SD.obs <- redwine$SD[is.na(redwine$SD) == FALSE]
FS.obs <- redwine$FS[is.na(redwine$SD) == FALSE]

# Build simple regression and print coefficients
SDFS_fit <- lm(SD.obs ~ FS.obs)
coefficients(SDFS_fit)
```

```
## (Intercept)      FS.obs
##   13.185505    2.086077
```

Problem 3

Missing values of SD were imputed using the following code:

```
# Create vector of estimated SD
FS_fill <- data.frame(FS.obs = redwine$FS[is.na(redwine$SD) == TRUE])
SD_est <- data.frame(predict(SDFS_fit, FS_fill))
```

```
# Create function for imputing missing SD values
estimp <- function(field, est)
{
  missing   <- is.na(field)
  n.missing <- sum(missing)
  field.obs <- field[!missing]
  imputed   <- field
  for (i in 1:n.missing)
  {
    imputed[missing][i] = est[i,]
  }
  return(imputed)
}

# Use function and print new RS average
redwine$SD <- estimp(redwine$SD, SD_est)
paste('SD new average:', round(mean(redwine$SD), digits = 2))
```

```
## [1] "SD new average: 46.3"
```

Problem 4

Missing values of RS were imputed using the following code:

```
# Create function for imputing missing RS values
avgimp <- function(field, avg)
{
  missing   <- is.na(field)
  n.missing <- sum(missing)
  field.obs <- field[!missing]
  imputed   <- field
  imputed[missing] <- avg
  return(imputed)
}

# Use function and print new RS average
redwine$RS <- avgimp(redwine$RS, rs_avg)
paste('RS new average:', round(mean(redwine$RS), digits = 2))
```

```
## [1] "RS new average: 2.54"
```

Problem 5

Below is the multiple linear regression model:

```
winemodel <- lm(QA ~ ., redwine)
coefficients(winemodel)
```

```
##   (Intercept)            FA            VA            CA            RS
##  47.202815335   0.068406796  -1.097686420  -0.178949797   0.025926958
##            CH            FS            SD            DE            PH
##  -1.631290466   0.003530106  -0.002854970 -44.816652166   0.035996993
```

```
##            SU            AL
##   0.944871182   0.247046550
```

Problem 6

Below is the summary of the model:

```
summary(winemodel)
```

```
##
## Call:
## lm(formula = QA ~ ., data = redwine)
##
## 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
```

Based on the summary shown above, the PH variable is least likely to be related to QA as it has the highest p-value compared to other variables.

Problem 7

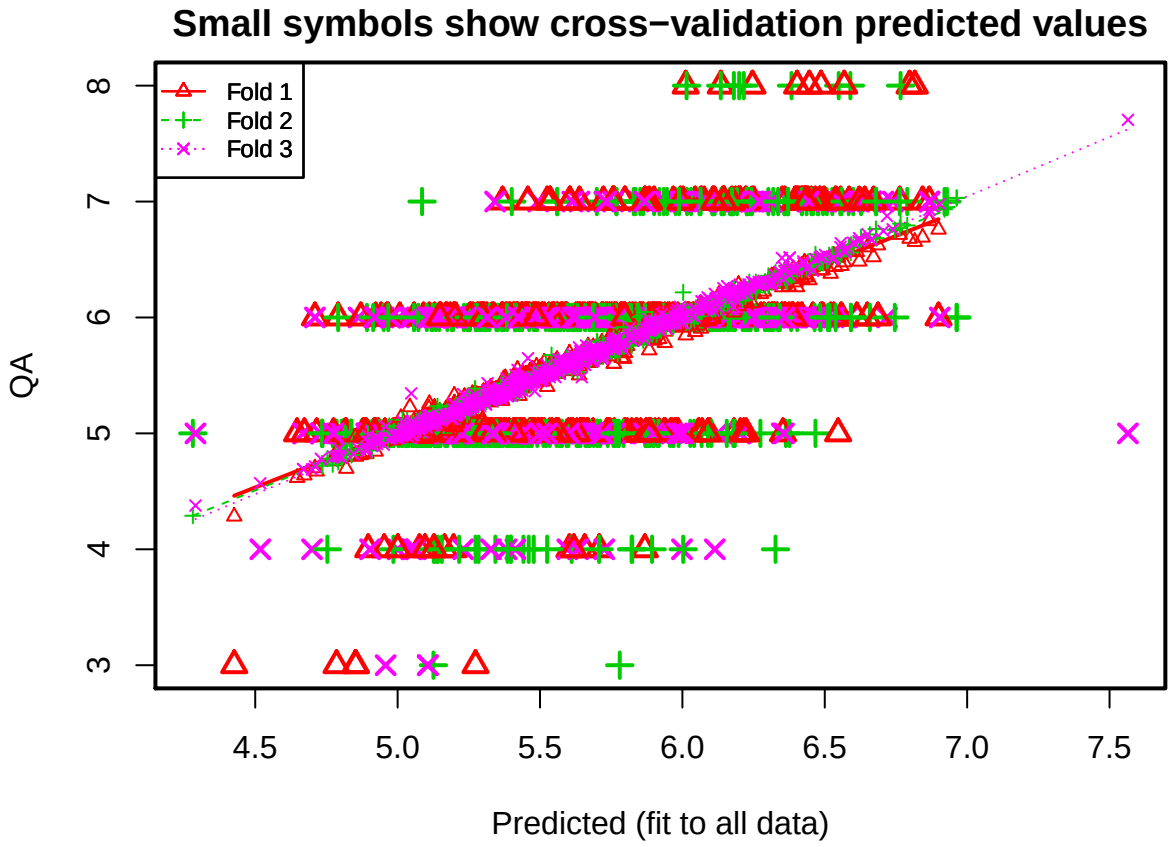
```
library(DAAG)
```

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

```
validation <- CVlm(data = redwine, m = 3, form.lm = winemodel, printit = FALSE)
```

```
## Warning in CVlm(data = redwine, m = 3, form.lm = winemodel, printit = FALSE):
##
##  As there is >1 explanatory variable, cross-validation
```

```
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```



Error is 0.43.

Problem 8

The average and standard deviation are shown below:

```
ph_avg <- mean(redwine$PH)
ph_std <- sd(redwine$PH)

paste('PH average:', round(ph_avg, digits = 2))
```

```
## [1] "PH average: 3.31"
```

```
paste('PH standard deviation:', round(ph_std, digits = 2))
```

```
## [1] "PH standard deviation: 0.39"
```

Below are the dimensions of *redwine* and *redwine2* as well as the number of rows removed:

```
redwine2 <- subset(redwine , PH < ph_avg + 3*ph_std & PH > ph_avg - 3*ph_std)
dim(redwine)
```

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

```r
dim(redwine2)
```

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

```r
paste("Number of rows removed:", dim(redwine)[1] - dim(redwine2)[1])
```

```
## [1] "Number of rows removed: 19"
```

Problem 9

Below is the new model:

```r
winemodel2 <- lm(QA ~ ., redwine2)
summary(winemodel2)
```

```
##
## Call:
## lm(formula = QA ~ ., data = redwine2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68933 -0.36336 -0.04368  0.45221  2.01272
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.036170  21.211609   0.897   0.3696
## FA            0.024613   0.026019   0.946   0.3443
## VA           -1.072147   0.122031  -8.786  < 2e-16 ***
## CA           -0.178017   0.148120  -1.202   0.2296
## RS            0.012955   0.014968   0.866   0.3869
## CH           -1.902552   0.420766  -4.522 6.60e-06 ***
## FS            0.004421   0.002182   2.026   0.0429 *
## SD           -0.003145   0.000738  -4.261 2.16e-05 ***
## DE          -14.973653  21.652465  -0.692   0.4893
## PH           -0.424704   0.192653  -2.205   0.0276 *
## SU            0.913456   0.114860   7.953 3.46e-15 ***
## AL            0.282744   0.026553  10.648  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6475 on 1568 degrees of freedom
## Multiple R-squared:  0.3629, Adjusted R-squared:  0.3585
## F-statistic: 81.21 on 11 and 1568 DF,  p-value: < 2.2e-16
```

The new model is worse. Even though we have a very small increase in R squared, we have less significant coefficients. Based on having the smallest p-values, variables VA, CH, SD, SU, and AL are most likely to be related to QA.