Analytics for Competitive Advantage: Lab Exercise 2

Kristin Meier

November 1, 2016

Problem 1

Calculate averages of RS and SD by ignoring the missing values.

```
wine <- read.table(paste(filepath,"ex2_redwine.txt",sep=""),stringsAsFactors = F,header=T)

avg.RS <- mean(wine$RS,na.rm=T)
avg.SD <- mean(wine$SD,na.rm=T)
```

Ignoring missing values, the averages of RS and SD are 2.538 and 46.298, respectively.

Problem 2

Create vectors of SD.obs and FS.obs by omitting observations with missing values in SD. Build linear regression model to estimate SD.obs (response) using FS.obs (explanatory). Print coefficients.

```
SD.obs <- wine$SD[!is.na(wine$SD)]
FS.obs <- wine$FS[!is.na(wine$SD)]

fit.SD.FS <- lm(SD.obs~FS.obs)

coef.fit.SD.FS <- coefficients(fit.SD.FS)

coef.fit.SD.FS
```

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

The intercept is 13.186 and the coefficient for FS is 2.086. The model is $SD = 13.186 + 2.086FS$.

Problem 3

Create a vector (of length 17) of estimated SD values using the regression model in Problem 2 and FS values of the observations with missing SD values. Impute missing values of SD using the created vector. Print out the average of SD after the imputation.

```
# First get the vector of FS values where SD is missing
FS.obs2 <- wine$FS[is.na(wine$SD)]
# Fit the SD values using the linear regression model
fit.SD <- coef.fit.SD.FS[1] + coef.fit.SD.FS[2]*FS.obs2
# impute the missin values of SD in original data
```

```r
wine$SD[is.na(wine$SD)] <- fit.SD
# Find new average
avg.SD.noNA <- mean(wine$SD)

avg.SD.noNA
```

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

The new SD average with the imputed values is 46.302.

Problem 4

Impute missing values of RS using the average value imputation method from the lab. Print out the average of RS after the imputation.

```r
# replace RS missing values with the average
wine$RS[is.na(wine$RS)] <- avg.RS
avg.RS.noNA <- mean(wine$RS)

avg.RS.noNA
```

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

The new RS average with the imputed values is 2.538.

Problem 5

Build multiple linear regression model for the new dataset and save it as winemodel. Print out the coeficients of the regression model.

```r
winemodel <- lm(QA ~ FA + VA + CA + RS + CH + FS + SD + DE + PH + SU + AL, wine)
coef.winemodel <- round(coefficients(winemodel),3)
coef.winemodel
```

```
## (Intercept)          FA          VA          CA          RS          CH
##      47.203       0.068      -1.098      -0.179       0.026      -1.631
##          FS          SD          DE          PH          SU          AL
##       0.004      -0.003     -44.817       0.036       0.945       0.247
```

The model is:

$QA = 47.203 + 0.068FA + -1.098VA + -0.179CA + 0.026RS + -1.631CH + 0.004FS + -0.003SD + -44.817DE + 0.036PH + 0.945SU + 0.247AL$

Problem 6

Print out the summary of the model. Pick one attribute that is least likely to be related to QA based on p-values.

```r
summary(winemodel)
```

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

```r
# Find which has the max p-value
pvals.winemodel <- summary(winemodel)$coefficients[,4]
max.var <- rownames(summary(winemodel)$coefficients)[which(pvals.winemodel == max(pvals.winemodel))]
```

The attribute that is least likely to be related to QA based on p-values is PH with a p-value of 0.414.

Problem 7

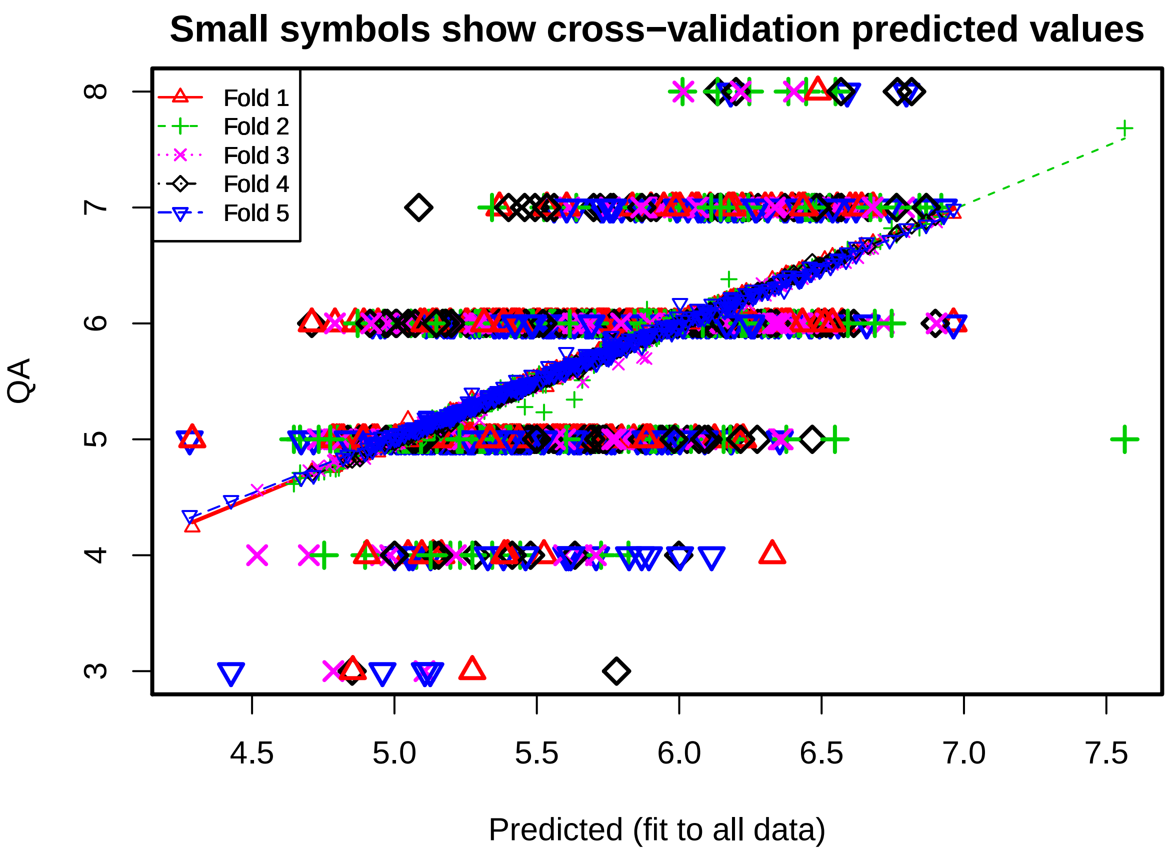
Perform 5-fold cross validation for the model you just built. Print out the average error rate.

```r
library(DAAG)
```

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

```r
wine.validation <- CVlm(data=wine, m=5,
                   form.lm=formula(QA ~ FA + VA + CA + RS + CH + FS + SD + DE + PH + SU + AL),
                   printit=F, plotit=T)
```

```
## Warning in CVlm(data = wine, m = 5, form.lm = formula(QA ~ FA + VA + CA + :
##
##  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
```




```
# average error rate
# (actual-pred)^2
avg.err <- round(mean((wine.validation$QA - wine.validation$cvpred)^2),3)
avg.err.check <- round(attr(wine.validation,"ms"),3)
```

The average error rate is 0.426.

Problem 8

Mr. Klabjan is informed that the attribute picked in Problem 6 actually contains outliers. Calculate the average and standard deviation of the selected attribute. Create a new data set after removing observations that is outside of the range [avg-3sd; avg+3sd] and name the data set as redwine2. Print out the dimension of redwine2 to know how many observations are removed.

```
# PH was chosen above
avg.PH <- mean(wine$PH)
```

```r
sd.PH <- sd(wine$PH)
PH.range.min <- avg.PH - 3*sd.PH
PH.range.max <- avg.PH + 3*sd.PH

redwine2 <- wine[(wine$PH > PH.range.min & wine$PH < PH.range.max),]
dim(redwine2)
```

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

The range for PH is 2.129 to 4.484. After removing observations outside of this range, the new dataset, redwine2, has dimenstions 1580 x 12.

Problem 9

Build regression model winemodel2 using the new data set from Problem 8 and print out the summary. Compare this model with the model obtained in Problem 6 and decide which one is better. Pick 5 attributes that is most likely to be related to QA based on p-values.

```r
winemodel2 <- lm(QA ~ FA + VA + CA + RS + CH + FS + SD + DE + PH + SU + AL, redwine2)
coef.winemodel2 <- round(coefficients(winemodel2),3)
summary(winemodel2)
```

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

```r
#
# Find which have the min 5 p-values
pvals.winemodel2 <- summary(winemodel2)$coefficients[,4]
min5 <- sort(pvals.winemodel2)[1:5]
min.val <- rownames(summary(winemodel2)$coefficients)[which(pvals.winemodel2 %in% min5)]

min.val
```

```
## [1] "VA" "CH" "SD" "SU" "AL"
```

The new model is:

$QA = 19.036 + 0.025FA + -1.072VA + -0.178CA + 0.013RS + -1.903CH + 0.004FS + -0.003SD + -14.974DE + -0.425PH + 0.913SU + 0.283AL$

Compared to the original model with an R^2 of 0.358, this one has an R^2 value of 0.363, which means more of the variance in QA is explained by the new model. Also, the F-statistic of the new model is 81.211, which is larger and more significant than that of the original model (80.6).

The 5 attributes most likely to be related to QA are the 5 with the lowest p-values, VA, CH, SD, SU, AL.