MSiA 400 Lab 2 Harish Chockalingam

```
#Problem 1
redwine<-read.table('redwine.txt',header=T)
mean_RS<-mean(redwine$RS,na.rm=T)
mean_SD<-mean(redwine$SD,na.rm=T)
mean_RS
```

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

```
mean_SD
```

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

```
#The RS and SD average after removing NAs is 2.53 and 46.29 respectively
```

```
#Problem 2
M<-cbind(redwine$FS,redwine$SD)
M<-na.omit(M)
FS.obs<-M[,1]
SD.obs<-M[,2]
ABC<-lm(SD.obs~FS.obs)
coef<-coefficients(ABC)
coef
```

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

```
#There are 17 missing SD values. After removing the missing SD and respective
#FD values, fitting yields a
#intercept of 13.18 and coefficient of 2.08
```

```
#Problem 3
SD<-redwine$SD
missingSD <- is.na(SD)
FS_17<-redwine$FS[missingSD]

SD_predict<-coef[1]+coef[2]*FS_17
redwine$SD[missingSD]<-SD_predict
mean(redwine$SD)
```

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

```
#The mean for SD after imputation is 46.30182, not a huge change
```

```
#Problem 4
avg.imp <- function (a, avg){
        missing <- is.na(a)
        imputed <- a
        imputed[missing] <- avg
        return (imputed)
        }

RS_ave<-mean(na.omit(redwine$RS))
RS_imp<-avg.imp(redwine$RS,RS_ave)
redwine$RS<-RS_imp
mean(RS_imp)
```

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

```
#The average value for RS is 2.537952
```

```
#Problem 5
winemodel<-lm(redwine$QA~redwine$FA+redwine$VA+redwine$CA+redwine$RS+redwine$CH+
                redwine$FS+redwine$SD+redwine$DE+redwine$PH+redwine$SU+redwine$AL)
coefficients(winemodel)
```

```
##   (Intercept)    redwine$FA    redwine$VA    redwine$CA    redwine$RS 
##  47.202815335   0.068406796  -1.097686420  -0.178949797   0.025926958 
##    redwine$CH    redwine$FS    redwine$SD    redwine$DE    redwine$PH 
##  -1.631290466   0.003530106  -0.002854970 -44.816652166   0.035996993 
##    redwine$SU    redwine$AL 
##   0.944871182   0.247046550
```

```
#The coefficients:
#Intercep: 47.202 FA:0.0684 VA:-1.097 CA:-0.179 RS:0.026  CH:-1.631
#FS:0.0035 SD:-0.0028  DE:-44.817  PH:0.036  SU: 0.944 AL:0.247
```

```
#Problem 6
summary(winemodel)
```

```
## 
## Call:
## lm(formula = redwine$QA ~ redwine$FA + redwine$VA + redwine$CA + 
##     redwine$RS + redwine$CH + redwine$FS + redwine$SD + redwine$DE + 
##     redwine$PH + redwine$SU + redwine$AL)
## 
## 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 ** 
## redwine$FA   6.841e-02  1.872e-02   3.654 0.000267 ***
## redwine$VA  -1.098e+00  1.213e-01  -9.053  < 2e-16 ***
## redwine$CA  -1.789e-01  1.474e-01  -1.214 0.224954    
## redwine$RS   2.593e-02  1.419e-02   1.827 0.067944 .  
## redwine$CH  -1.631e+00  4.097e-01  -3.982 7.14e-05 ***
## redwine$FS   3.530e-03  2.159e-03   1.635 0.102262    
## redwine$SD  -2.855e-03  7.248e-04  -3.939 8.54e-05 ***
## redwine$DE  -4.482e+01  1.789e+01  -2.505 0.012329 *  
## redwine$PH   3.600e-02  4.409e-02   0.816 0.414413    
## redwine$SU   9.449e-01  1.136e-01   8.321  < 2e-16 ***
## redwine$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
```

```
#From the summary R^2 is 0.3584, and based on a significance level of 0.05
#the PH attribute is least likely related to QA
#as it has a high p-value of 0.4144
```

```
#Problem 7
CVInd  <-  function(n,K){    #n  is  sample  size;  K  is  number  of  parts;
  #returns  K-length  list  of indices for each part
  m<-floor(n/K)  #approximate size of each part
  r<-n-m*K
  I<-sample(n,n)  #random reordering of the indices
  Ind<-list()  #will be list of indices for all K parts
  length(Ind)<-K
  for (k in 1:K) {
    if (k <= r) kpart <- ((m+1)*(k-1)+1):((m+1)*k)
      else kpart<-((m+1)*r+m*(k-r-1)+1):((m+1)*r+m*(k-r))
    Ind[[k]] <- I[kpart]  #indices for kth part of data
    }
    Ind }

Nrep<-20 #number of replicates of CV
K<-5  #K-fold CV on each replicate
n=nrow(redwine)
y<-redwine$QA
SSE<-matrix(0,Nrep,1)
for (j in 1:Nrep) {
  Ind<-CVInd(n,K)
  yhat11<-y
  for (k in 1:K) {
    out11<-lm(QA~.,redwine[-Ind[[k]],])
    yhat11[Ind[[k]]]<-as.numeric(predict(out11,redwine[Ind[[k]],]))
    } #end of k loop
  SSE[j]=c(sum((y-yhat11)^2))
  } #end of j loop
SSE
```

```
##           [,1]
##  [1,] 682.8796
##  [2,] 676.9501
##  [3,] 688.1601
##  [4,] 680.7316
##  [5,] 678.6632
##  [6,] 684.9521
##  [7,] 688.2641
##  [8,] 682.6897
##  [9,] 681.5459
## [10,] 682.6100
## [11,] 680.7892
## [12,] 683.7520
## [13,] 684.2307
## [14,] 684.7442
## [15,] 680.9374
## [16,] 684.5377
## [17,] 683.3969
## [18,] 690.2042
## [19,] 682.1277
## [20,] 689.2895
```

```
apply(SSE,2,mean)
```

```
## [1] 683.5728
```

```
#The average error rate after 20 replications is 683.4685. The MSE (683.4685/(1599-11-1))=0.431
```

```
#Problem 8
PH_omit<-na.omit(redwine$PH)
PH_mean<-mean(PH_omit)
PH_std<-sd(na.omit(redwine$PH))
PH_lb<-PH_mean-3*PH_std
PH_ub<-PH_mean+3*PH_std

redwine2<-subset(redwine,redwine$PH<PH_ub & redwine$PH>PH_lb)
dim(redwine2)
```

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

```
#dimensions of redwine2 is 1580 x 12. The imputed redwine dataset had 1599
#values thus there were 19 outliers
```

```
#Problem 9
winemodel2<-lm(redwine2$QA~redwine2$FA+redwine2$VA+redwine2$CA+redwine2$RS+
                 redwine2$CH+redwine2$FS+redwine2$SD+redwine2$DE+redwine2$PH+redwine2$SU+redwine2$AL)
summary(winemodel2)
```

```
##
## Call:
## lm(formula = redwine2$QA ~ redwine2$FA + redwine2$VA + redwine2$CA +
##     redwine2$RS + redwine2$CH + redwine2$FS + redwine2$SD + redwine2$DE +
##     redwine2$PH + redwine2$SU + redwine2$AL)
##
## 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
## redwine2$FA   0.024613   0.026019   0.946   0.3443
## redwine2$VA  -1.072147   0.122031  -8.786  < 2e-16 ***
## redwine2$CA  -0.178017   0.148120  -1.202   0.2296
## redwine2$RS   0.012955   0.014968   0.866   0.3869
## redwine2$CH  -1.902552   0.420766  -4.522 6.60e-06 ***
## redwine2$FS   0.004421   0.002182   2.026   0.0429 *
## redwine2$SD  -0.003145   0.000738  -4.261 2.16e-05 ***
## redwine2$DE -14.973653  21.652465  -0.692   0.4893
## redwine2$PH  -0.424704   0.192653  -2.205   0.0276 *
## redwine2$SU   0.913456   0.114860   7.953 3.46e-15 ***
## redwine2$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
```

```
#Compared to problem 6 the R^2 went up from 0.3584 to 0.3629 (not significant increase).
#After removing the outliers the signifance of coefficients has changed, but
#both models still have 4 non-significant coefficients. Thus, winemodel2 is a slightly
#better model to predict QA.
#The five attributes likely related to QA are VA, CH, SD, SU, AL they have
#p-values close to zero (signifcance level used 0.05)
```