LabAssignment2

Kejin Qian

11/12/2018

import redwine.txt

```
redwine = read.table('redwine.txt', header=T);
```

problem 1

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

```
avgRS <- mean(redwine$RS, na.rm = T)
print(paste('the average of RS by ignoring the missing values is', avgRS))
```

```
## [1] "the average of RS by ignoring the missing values is 2.53795180722892"
```

```
avgSD <- mean(redwine$SD, na.rm = T)
print(paste('the average of SD by ignoring the missing values is', avgSD))
```

```
## [1] "the average of SD by ignoring the missing values is 46.2983565107459"
```

Answer:

- The average of RS by ignoring the missing values is 2.53795180722892
- The average of SD by ignoring the missing values is 46.2983565107459

---

problem 2

After correlation analysis, Mr. Klabjan observed that there exists a significant correlation between SD and FS. Create vectors of SD.obs and FS.obs by omitting observations with missing values in SD. Build (simple) linear regression model to estimate SD.obs using FS.obs. That is, SD.obs is used as response variable and FS.obs is used as explanatory variable for the regression analysis. Print out the coefficients of the regression model. Hint: If you save the output from lm function to ABC, then the coefficients of the regression model can be obtained by coefficients(ABC).

```
SDFS <- cbind(redwine$SD, redwine$FS)
SDFS.omit <- na.omit(SDFS)
colnames(SDFS.omit) <- c('SD','FS')
#vector of SD.obs by omitting missing values
SD_omit <- SDFS.omit[,1]
#vector of FS.obs by omitting missing values in SD
FS_omit <- SDFS.omit[,2]
model1 <- lm(SD_omit ~ FS_omit)
summary(model1)
```

```
##
## Call:
```

```
## lm(formula = SD_omit ~ FS_omit)
##
## 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_omit      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
```

```
coefficients(model1)
```

```
## (Intercept)     FS_omit
##   13.185505    2.086077
```

From the above regression model summary output and coefficients output, we get intercept = 13.185505 and slope = 2.086077.

---

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.

```
FS_omit = redwine$FS[is.na(redwine$SD)]
pred <- predict(model1, data.frame(FS_omit))
#Create a vector of estimated SD values using model1 and FS values of the observations with missing SD
pred_vector <- as.vector(pred)
pred_vector
```

```
##  [1] 44.47667 38.21843 36.13236 38.21843 97.67164 15.27158 27.78805
##  [8] 86.19821 44.47667 88.28429 61.16528 38.21843 29.87412 27.78805
## [15] 44.47667 50.73490 23.61589
```

```
redwine$SD[is.na(redwine$SD)] <- pred_vector
avgSDnew <- mean(redwine$SD)
print(paste('the average of SD is', avgSDnew))
```

```
## [1] "the average of SD is 46.3018196746507"
```

Answer: The created vector of estimated SD values using the regression model in Problem 2 and FS values of the observations with missing SD values is printed in the output above.

The average of SD after the imputation is 46.3018196746507.

---

problem 4

Mr. Klabjan decided RS is not significantly correlated to other attributes. Impute missing values of RS using the average value imputation method from the lab. Print out the average of RS after the imputation.

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

```
redwine$RS = avg.imp(redwine$RS, avgRS)
avgRSnew <- mean(redwine$RS)
print(paste('the average of RS is', avgRSnew))
```

```
## [1] "the average of RS is 2.53795180722892"
```

Answer:

The average of RS after the imputation is 2.53795180722892.

---

problem 5

We have imputed all missing values in the data set. Build multiple linear regression model for the new dataset and save it as winemodel. Print out the coefficients of the regression model. Hint 1 : built multiple linear regression by winemodel=lm(redwine$QA * redwine$FA+...+redwine$AL)

```
winemodel <- lm(QA ~ ., data = 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
```

Answer: *(Intercept) 47.202815335

*FA 0.068406796

*VA -1.097686420

*CA -0.178949797

*RS 0.025926958

*CH -1.631290466

*FS 0.003530106

*SD -0.002854970

*DE -44.816652166

*PH 0.035996993

*SU 0.944871182

*AL 0.247046550

---

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.

```
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
```

PH has the highest p-value(0.414413) among all the predictors. So PH is least likely to be related to QA based on p-values.

---

problem 7

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

```
CVInd <- function(n,K) {
  m<-floor(n/K)
  r<- n-m*K
  I<-sample(n,n)
  Ind<-list()
  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]}
  Ind
  }
```

```
# Repeat the 5-fold cross validation 20 times and take the average of SSE
Nrep <- 20
K <- 5
n <- nrow(redwine)
y <- redwine$QA
SSE <- matrix(0,Nrep,1)
```

```
for (j in 1:Nrep) {
  Ind <- CVInd(n,5)
  yhat <- y
  for (k in 1:K) {
    out <- lm(QA~.,redwine[-Ind[[k]],])
    yhat[as.vector(Ind[[k]])] <- predict(out,redwine[Ind[[k]],-1])
  }
  SSE[j,]=sum((y-yhat)^2)
}
mean(SSE)
```

```
## [1] 683.5239
```

```
print(paste("the average error rate (SSE) is", mean(SSE)))
```

```
## [1] "the average error rate (SSE) is 683.523908256945"
```

The average error rate (SSE) is printed above.

---

problem 8

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

```
PHmean <- mean(redwine$PH)
PHstd <- sd(redwine$PH)
PHupper <- PHmean + 3 * PHstd
PHlower <- PHmean - 3 * PHstd
redwine2 <-subset(redwine, PH<PHupper & PH>PHlower)
dim(redwine)
```

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

```
dim(redwine2)
```

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

Before removing the outliers, we have 1599 observations in redwine dataset. After removing observations that is outside of the range $[\mu+3\sigma;\ \mu+3\sigma]$, our dataset redwine2 has 1580 features. So 19 features were removed.

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.

```
winemodel2 <- lm(QA ~ ., data = redwine2)
coefficients(winemodel2)
```

```
##   (Intercept)            FA            VA            CA            RS
##  19.036169888   0.024613355  -1.072147125  -0.178017195   0.012955228
##            CH            FS            SD            DE            PH
##  -1.902551975   0.004421387  -0.003144516 -14.973653485  -0.424704272
##            SU            AL
##   0.913456157   0.282744390
```

```
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
```

1. In problem 6, the model we fit has $R^2_{adj} = 0.354$. After removing outliers from PH, we get a higher $R^2_{adj} = 0.3585$ from the new fitted model.
2. The models from problem 6 and problem 9 both have 7 predictors which have statistically significant coefficients at significance level of 0.05.

3. In problem 6, the model we fit has overall F-statistic = 80.6 (p-value < 2.2e-16) while in problem 9, the overall F-statistic is 81.21 (p-value < 2.2e-16) which is slightly higher.

So based on the above three comparisons, I think the model from Problem 9 is better than the model from Problem 6 because it has a higher R^2_{adj} and more significant overall F-statistic.

Pick 5 attributes that is most likely to be related to QA based on p-values:

1. VA (p-value < 2e-16)
2. AL (p-value < 2e-16)
3. SU (p-value < 3.46e-15)
4. CH (p-value < 6.60e-06)
5. SD (p-value < 2.16e-05)
