lab_assignment_2

Ivan

11/14/2018

```
library(readr)
redwine <- read_delim("~/Desktop/MSIA/Courses/MSIA 400 Everything starts with data/Lab Assignment 1/Lab
```

```
## Parsed with column specification:
## cols(
##   QA = col_integer(),
##   FA = col_double(),
##   VA = col_double(),
##   CA = col_double(),
##   RS = col_double(),
##   CH = col_double(),
##   FS = col_double(),
##   SD = col_integer(),
##   DE = col_double(),
##   PH = col_double(),
##   SU = col_double(),
##   AL = col_double()
## )
```

```
## Warning: 2 parsing failures.
## row # A tibble: 2 x 5 col     row col   expected        actual file
```

```
#str(redwine)
#summary(redwine)
```

1

```
mean(redwine$RS,na.rm = T)
```

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

```
mean(redwine$SD,na.rm = T)
```

```
## [1] 46.25886
```

2

```
Sd <- redwine$SD[!is.na(redwine$SD)]
Fs <- redwine$FS[!is.na(redwine$SD)]
fit.sd <- lm(Sd~Fs)
coefficients(fit.sd)
```

```
## (Intercept)          Fs
##   12.963691    2.103455
```

3

```
test.sd <- data.frame(Fs = redwine$FS[is.na(redwine$SD)])
redwine$SD[is.na(redwine$SD)] <- predict(fit.sd,test.sd)
mean(redwine$SD)
```

```
## [1] 46.35588
```

4

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

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

5

```
redwine$RS <- avg.imp(redwine$RS,mean(redwine$RS[!is.na(redwine$RS)]))

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

```
##   (Intercept)            FA            VA            CA            RS
##  47.434916778   0.068546221  -1.097506221  -0.179577814   0.026132583
##            CH            FS            SD            DE            PH
##  -1.629534220   0.003606445  -0.002869843 -45.047229263   0.035821052
##            SU            AL
##   0.943987844   0.246735709
```

6

```
summary(winemodel)
```

```
##
## Call:
## lm(formula = QA ~ ., data = redwine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.78129 -0.36252 -0.05929  0.44558  1.98948
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.743e+01  1.782e+01   2.662 0.007847 **
## FA           6.855e-02  1.872e-02   3.663 0.000258 ***
```

```
## VA          -1.098e+00  1.212e-01  -9.052  < 2e-16 ***
## CA          -1.796e-01  1.473e-01  -1.219 0.223070
## RS           2.613e-02  1.420e-02   1.841 0.065872 .
## CH          -1.630e+00  4.096e-01  -3.978 7.25e-05 ***
## FS           3.606e-03  2.169e-03   1.663 0.096564 .
## SD          -2.870e-03  7.264e-04  -3.951 8.12e-05 ***
## DE          -4.505e+01  1.789e+01  -2.518 0.011891 *
## PH           3.582e-02  4.410e-02   0.812 0.416711
## SU           9.440e-01  1.135e-01   8.314  < 2e-16 ***
## AL           2.467e-01  2.266e-02  10.887  < 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:  0.3585, Adjusted R-squared:  0.354
## F-statistic: 80.61 on 11 and 1587 DF,  p-value: < 2.2e-16
```

PH has the largest p-value thus are least likely to be related to QA

7

```
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] #indices for kth part of data
}
Ind
}
Nrep <- 20
K <- 5
n = nrow(redwine)
SSE <- c()
for (j in 1:Nrep) {
  Ind<-CVInd(n,K)
  for (k in 1:K) {
    out <- lm(QA~.,data = redwine[-Ind[[k]],])
    yhat <- as.numeric(predict(out,redwine[Ind[[k]],]))
    SSE <- c(SSE,sum((redwine$QA[Ind[[k]]]-yhat)^2))
  }
}
mean(SSE)
```

```
## [1] 136.4407
```

8

```r
mean(redwine$PH)
```

```
## [1] 3.306202
```

```r
sd(redwine$PH)
```

```
## [1] 0.3924948
```

```r
PH.lb = mean(redwine$PH) - 3 * sd(redwine$PH)
PH.ub = mean(redwine$PH) + 3 * sd(redwine$PH)
redwine2 <- subset(redwine, redwine$PH < PH.ub & redwine$PH > PH.lb)
dim(redwine2)
```

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

```r
nrow(redwine)
```

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

19 observations are removed.

9

```r
fit = lm(QA ~ ., data = redwine2)
summary(fit)
```

```
##
## Call:
## lm(formula = QA ~ ., data = redwine2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.69007 -0.36398 -0.04395  0.45262  2.01461
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.078272  21.209017   0.900   0.3685
## FA            0.024396   0.026023   0.937   0.3487
## VA           -1.071707   0.122018  -8.783  < 2e-16 ***
## CA           -0.178537   0.148028  -1.206   0.2280
## RS            0.013102   0.014968   0.875   0.3815
## CH           -1.902735   0.420707  -4.523 6.56e-06 ***
## FS            0.004522   0.002193   2.062   0.0394 *
## SD           -0.003169   0.000740  -4.282 1.96e-05 ***
## DE          -14.999982  21.649949  -0.693   0.4885
## PH           -0.428552   0.192779  -2.223   0.0264 *
## SU            0.912227   0.114845   7.943 3.73e-15 ***
## AL            0.282625   0.026553  10.644  < 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:  0.363,  Adjusted R-squared:  0.3585
## F-statistic: 81.24 on 11 and 1568 DF,  p-value: < 2.2e-16
```

The new model is better because both F statistics and R square values increased. VA,CH,SD,SU,AL have the smallest p values so most likely to be related with QA.