Lab Assignment 2

Joe Gilbert

10/23/2017

```
options(digits=4)
```

```
redwine <- read.delim("redwine.txt")
```

Problem 1

The mean of RS is given by:

```
rsmean <- mean(redwine$RS,na.rm=TRUE)
rsmean
```

```
## [1] 2.538
```

The mean of SD is given by:

```
sdmean <- mean(redwine$SD,na.rm=TRUE)
sdmean
```

```
## [1] 46.3
```

Problem 2

The coffecients of the regression model of SD on FS are given by:

```
SDmissing <- is.na(redwine$SD)
SD.obs <- redwine$SD[-SDmissing]
FS.obs <- redwine$FS[-SDmissing]
SDmodel <- lm(SD.obs ~ FS.obs)
coefficients(SDmodel)
```

```
## (Intercept)      FS.obs
##      13.188       2.086
```

Problem 3

```
estSD <- redwine$FS[SDmissing]*coefficients(SDmodel)[2]
estSD <- estSD + coefficients(SDmodel)[1]
estSD
```

```
##  [1] 44.48 38.22 36.13 38.22 97.67 15.27 27.79 86.20 44.48 88.28 61.17
## [12] 38.22 29.88 27.79 44.48 50.74 23.62
```

The new mean of SD after imputing values is given by:

```
redwine[SDmissing,"SD"] <- estSD
sdmean.replaced <- mean(redwine$SD)
sdmean.replaced
```

```
## [1] 46.3
```

Problem 4

The new mean of RS after imputing values is given by:

```
RSmissing <- is.na(redwine$RS)
redwine[RSmissing,"RS"] <- rsmean
rsmean.replaced <- mean(redwine$RS)
rsmean.replaced
```

```
## [1] 2.538
```

It does not change since we input mean for all missing values.

Problem 5

The coefficents of the model are given by:

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

```
## (Intercept)          FA          VA          CA          RS          CH
##   47.202825    0.068407   -1.097686   -0.178949    0.025927   -1.631291
##          FS          SD          DE          PH          SU          AL
##    0.003530   -0.002855  -44.816662    0.035997    0.944871    0.247047
```

Problem 6

```
summary(winemodel)
```

```
##
## Call:
## lm(formula = QA ~ ., data = redwine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7801 -0.3625 -0.0633  0.4460  1.9883
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.72e+01   1.78e+01    2.65  0.00815 **
## FA           6.84e-02   1.87e-02    3.65  0.00027 ***
## VA          -1.10e+00   1.21e-01   -9.05  < 2e-16 ***
## CA          -1.79e-01   1.47e-01   -1.21  0.22496
## RS           2.59e-02   1.42e-02    1.83  0.06794 .
## CH          -1.63e+00   4.10e-01   -3.98  7.1e-05 ***
```

```
## FS           3.53e-03   2.16e-03    1.63  0.10226
## SD          -2.85e-03   7.25e-04   -3.94  8.5e-05 ***
## DE          -4.48e+01   1.79e+01   -2.51  0.01233 *
## PH           3.60e-02   4.41e-02    0.82  0.41441
## SU           9.45e-01   1.14e-01    8.32  < 2e-16 ***
## AL           2.47e-01   2.27e-02   10.91  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.649 on 1587 degrees of freedom
## Multiple R-squared:  0.358,  Adjusted R-squared:  0.354
## F-statistic: 80.6 on 11 and 1587 DF,  p-value: <2e-16
```

Based on P-values, the attribute PH is least likely to be related to QA.

Problem 7

```
library("boot")
```

The average error rate is given by:

```
winemodelglm <- glm(QA~.,data=redwine)
validation <- cv.glm(redwine,winemodelglm,K=5)
validation$delta[1]
```

```
## [1] 0.4271
```

Problem 8

```
phmean <- mean(redwine$PH)
phstdev <- sd(redwine$PH)
ph.ub <- phmean + 3*phstdev
ph.lb <- phmean - 3*phstdev
redwine2 <- subset(redwine,PH>ph.lb & PH<ph.ub)
nrow(redwine2)
```

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

The original redwine dataset had 1599 observations, so 19 outliers were removed.

Problem 9

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

```
##
## Call:
## lm(formula = QA ~ ., data = redwine2)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -2.6893 -0.3634 -0.0437  0.4522  2.0127
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.90e+01   2.12e+01    0.90    0.370
## FA           2.46e-02   2.60e-02    0.95    0.344
## VA          -1.07e+00   1.22e-01   -8.79  < 2e-16 ***
## CA          -1.78e-01   1.48e-01   -1.20    0.230
## RS           1.30e-02   1.50e-02    0.87    0.387
## CH          -1.90e+00   4.21e-01   -4.52  6.6e-06 ***
## FS           4.42e-03   2.18e-03    2.03    0.043 *
## SD          -3.14e-03   7.38e-04   -4.26  2.2e-05 ***
## DE          -1.50e+01   2.17e+01   -0.69    0.489
## PH          -4.25e-01   1.93e-01   -2.20    0.028 *
## SU           9.13e-01   1.15e-01    7.95  3.5e-15 ***
## AL           2.83e-01   2.66e-02   10.65  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1568 degrees of freedom
## Multiple R-squared:  0.363,  Adjusted R-squared:  0.358
## F-statistic: 81.2 on 11 and 1568 DF,  p-value: <2e-16
```

The second model has slightly higher R-squared, so it is better. Based on p-values, 5 attributes that are most likely to be related to QA are VA, CH, SD, SU, and AL.