Everything Starts with Data Lab Exercise #2

Johnny Chiu

11/3/2017

```
redwine <- read.table('_data/redwine.txt', header = TRUE)
```

Problem 1

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

Answer

The averages of RS and SD by ignoring the missing values are shown in the table below.

```
rs_mean = mean(redwine$RS, na.rm=TRUE)
sd_mean = mean(redwine$SD, na.rm=TRUE)

table_1 <- data.frame(Feature=c('RS','SD'), 'mean ignoring NA'=c(rs_mean,
sd_mean))
kable(table_1)
```

Feature	mean.ignoring.NA
RS	2.537952
SD	46.298356

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).

Answer

The coefficients of the regression model are shown in the table below.

```
df_2 = redwine %>% select(SD, FS) %>% filter(!is.na(SD))

model_2 = lm(SD~FS, data=df_2)
```

```
model_2_coefficients = summary(model_2)$coefficients[1:2]

model_2_coefficients_df = data.frame(summary(model_2)$coefficients)

kable(model_2_coefficients_df %>% select(Estimate))
```

	Estimate
(Intercept)	13.185505
FS	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.

Answer

```
redwine_imputed = redwine

df_3 = redwine_imputed %>% select(SD, FS) %>% filter(is.na(SD))

missing_sd = is.na(redwine_imputed$SD)
redwine_imputed$SD[missing_sd] =
predict(model_2,newdata=data.frame('FS'=df_3$FS))
```

The average of SD after the imputation is 46.3018197.

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.

Answer

```
missing_rs = is.na(redwine_imputed$RS)
redwine_imputed$RS[missing_rs] = mean(redwine_imputed$RS, na.rm = TRUE)
```

The average of SD after the imputation is 2.5379518.

Problem 5

We have imputed all missing values in the data set. Build multiple linear regression model for the new data set and save it as winemodel. Print out the coefficients of the regression model.

Hint 1: built multiple linear regression by winemodel = lm($redwine\$QA \sim redwine\$FA + \ldots + redwine\$AL$*)*

Answer

The coefficients of the regression model are shown in the table below.

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

winemodel_coefficients_df = data.frame(summary(winemodel)$coefficients)
kable(winemodel_coefficients_df %>% select(Estimate))
```

	Estimate
(Intercept)	47.2028153
FA	0.0684068
VA	-1.0976864
CA	-0.1789498
RS	0.0259270
CH	-1.6312905
FS	0.0035301
SD	-0.0028550
DE	-44.8166522
PH	0.0359970
SU	0.9448712
AL	0.2470466

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.

Answer

According to the summary table below, the attribute that is least likely to be related to QA is *PH*, with the highest p-value of 0.414413.

```
summary(winemodel)

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

Problem 7

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

Answer

The average error rate for *winemodel* performing 5-fold cross validation is shown as below, which is 0.6552711.

```
cvFit(winemodel, data = redwine_imputed, y = redwine_imputed$QA, K = 5, seed
= 1)

## 5-fold CV results:
##        CV
## 0.6552711
```

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 observa- tions that is outside of the range [mean − 3 std, mean + 3 std] and name the data set as redwine2. Print out the dimension of redwine2 to know how many observations are removed.

Answer

```
ph.std = sd(redwine_imputed$PH,na.rm = TRUE)
ph.mean = mean(redwine_imputed$PH,na.rm = TRUE)

ph.ub = ph.mean + 3*ph.std
ph.lb = ph.mean - 3*ph.std
```

```
redwine2 <- redwine_imputed %>% filter(PH < ph.ub & PH > ph.lb)

dim(redwine2)

## [1] 1580   12
```

The dimension of redwine2 is 1580. There are 19 observations being 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.

Answer

According to the overall Adjusted R-squared, winemodel2 is slightly better than winemodel. 5 attributes that is most likely to be related to QA are *VA*,*CH*,*SD*,*SU*,and *AL*.

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