MSiA 400 Lab Assignment 1 Solution

Oct 13, 2014

- Due: 11:59pm Oct 20, 2014
- This is an open book assignment.
- Please submit one report file that includes : short answer, related code and print for each problem if necessary.

Cortez *et al.* (2009) model wine quality based on physicochemical tests. The data includes 11 input (explanatory or independent) variables and 1 output (response or dependent) variable for regression analysis. The 11 explanatory variables include:

> fixed acidity(FA), volatile acidity(VA), citric acid(CA), residual sugar(RS),
> chlorides(CH), free sulfur dioxide(FS), total sulfur dioxide(SD), density(DE),
> pH(PH), sulphates(SU), and alcohol(AL).

The output variable is quality(QA) (score between 0 and 10).

Please find the attached data file *redwine.txt*. Unlike the original data set of Cortez *et al.* (2009), *redwine.txt* contains missing values in attributes SD and RS. The summary of the data set is following.

Name of the data set	redwine
Number of columns	12 (11 explanatory variables and 1 response variable)
Number of observations	1599
Number of missing values	39 (22 in RS and 17 in SD)

Answer the following questions.

Problem 1

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

Solution

```
> mean(redwine$RS, na.rm=T);
[1] 2.537952
> mean(redwine$SD, na.rm=T);
[1] 46.29836
```

Problem 2

After correlation analysis, Mr. Park 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).

Solution

```
> missing.SD = is.na(redwine$SD)
> SD.obs = redwine$SD[!missing.SD]
> FS.obs = redwine$FS[!missing.SD]
> reg.obs = lm(SD.obs~FS.obs);
> coefficients(feg.obs);
(Intercept)      FS.obs
  13.185505    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.

Solution

```
> beta0 = 13.186
> beta1 = 2.086
> FS.imp = redwine$FS[missing.SD]
> SD.imp = beta0+FS.imp*beta1;
> redwine$SD[missing.SD]=SD.imp;
> mean(redwine$SD);
[1] 46.30181
```

Problem 4

Mr. Park 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.

Solution

```
> RS.avg = mean(redwine$RS, na.rm=T);
> missing.RS = is.na(redwine$RS);
> redwine$RS[missing.RS]=RS.avg;
> mean(redwine$RS)
[1] 2.537952
```

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$QA∼redwine$FA+...+redwine$AL)

Solution

```
> 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    redwine$CH
  47.202781     0.068407      -1.097686     -0.178950     0.025927      -1.631291
redwine$FS    redwine$SD    redwine$DE    redwine$PH    redwine$SU    redwine$AL
0.003530      -0.002855     -44.816617    0.035997      0.944871      0.247047
```

Problem 6

Printout the summary of the model. Pick one attribute that is least likely to be related to QA based on p-values.

Solution
Pick PH since the p-value is the largest.

```
> 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
```

Problem 7

Mr. Park 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.

Solution

```
> PH.avg = mean(redwine$PH);
> PH.sd = sd(redwine$PH);
> PH.lb = PH.avg - 3*PH.sd;
> PH.ub = PH.avg + 3*PH.sd;
> redwine2 = subset(redwine, PH<PH.ub & PH>PH.lb);
> dim(redwine2);
[1] 1580   12
```

Problem 8

Build regression model winemodel2 using the new data set from Problem 7 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.

Solution
Winemodel2 is better since it has lower RSS and greater r^2.
Pick VA,CH,SD,SU,AL

```
> 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)  19.036121  21.211610   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 -14.973604  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
```

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J.(2009) Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, **47**, 547–553.