READING TEST SCORES

Anuranjan

25 September 2016

Load the training and testing sets

```
pisaTrain=read.csv("pisa2009train.csv")
pisaTest=read.csv("pisa2009test.csv")
summary(pisaTrain)
```

```
##      grade            male                      raceeth    
##  Min.   : 8.00   Min.   :0.0000   White             :2015  
##  1st Qu.:10.00   1st Qu.:0.0000   Hispanic          : 834  
##  Median :10.00   Median :1.0000   Black             : 444  
##  Mean   :10.09   Mean   :0.5111   Asian             : 143  
##  3rd Qu.:10.00   3rd Qu.:1.0000   More than one race: 124  
##  Max.   :12.00   Max.   :1.0000   (Other)           :  68  
##                                   NA's              :  35  
##    preschool      expectBachelors     motherHS    motherBachelors
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:1.00   1st Qu.:0.0000  
##  Median :1.0000   Median :1.0000   Median :1.00   Median :0.0000  
##  Mean   :0.7228   Mean   :0.7859   Mean   :0.88   Mean   :0.3481  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.00   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00   Max.   :1.0000  
##  NA's   :56       NA's   :62       NA's   :97     NA's   :397     
##    motherWork        fatherHS      fatherBachelors    fatherWork    
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.0000  
##  Median :1.0000   Median :1.0000   Median :0.0000   Median :1.0000  
##  Mean   :0.7345   Mean   :0.8593   Mean   :0.3319   Mean   :0.8531  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##  NA's   :93       NA's   :245      NA's   :569      NA's   :233     
##    selfBornUS      motherBornUS     fatherBornUS    englishAtHome   
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:1.0000  
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000  
##  Mean   :0.9313   Mean   :0.7725   Mean   :0.7668   Mean   :0.8717  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##  NA's   :69       NA's   :71       NA's   :113      NA's   :71      
##  computerForSchoolwork read30MinsADay   minutesPerWeekEnglish
##  Min.   :0.0000        Min.   :0.0000   Min.   :   0.0       
##  1st Qu.:1.0000        1st Qu.:0.0000   1st Qu.: 225.0       
##  Median :1.0000        Median :0.0000   Median : 250.0       
##  Mean   :0.8994        Mean   :0.2899   Mean   : 266.2       
##  3rd Qu.:1.0000        3rd Qu.:1.0000   3rd Qu.: 300.0       
##  Max.   :1.0000        Max.   :1.0000   Max.   :2400.0       
##  NA's   :65            NA's   :34       NA's   :186          
##  studentsInEnglish schoolHasLibrary  publicSchool        urban       
##  Min.   : 1.0      Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:20.0      1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000  
##  Median :25.0      Median :1.0000   Median :1.0000   Median :0.0000  
##  Mean   :24.5      Mean   :0.9676   Mean   :0.9339   Mean   :0.3849  
##  3rd Qu.:30.0      3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :75.0      Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##  NA's   :249       NA's   :143                                        
##    schoolSize    readingScore
##  Min.   : 100   Min.   :168.6
##  1st Qu.: 712   1st Qu.:431.7
##  Median :1212   Median :499.7
##  Mean   :1369   Mean   :497.9
##  3rd Qu.:1900   3rd Qu.:566.2
##  Max.   :6694   Max.   :746.0
##  NA's   :162
```

```
str(pisaTrain)
```

```
## 'data.frame':    3663 obs. of  24 variables:
##  $ grade                : int  11 11 9 10 10 10 10 10 9 10 ...
##  $ male                 : int  1 1 1 0 1 1 0 0 0 1 ...
##  $ raceeth              : Factor w/ 7 levels "American Indian/Alaska Nativ
e",..: NA 7 7 3 4 3 2 7 7 5 ...
##  $ preschool            : int  NA 0 1 1 1 1 0 1 1 1 ...
##  $ expectBachelors      : int  0 0 1 1 0 1 1 1 0 1 ...
##  $ motherHS             : int  NA 1 1 0 1 NA 1 1 1 1 ...
##  $ motherBachelors      : int  NA 1 1 0 0 NA 0 0 NA 1 ...
##  $ motherWork           : int  1 1 1 1 1 1 1 0 1 1 ...
##  $ fatherHS             : int  NA 1 1 1 1 1 NA 1 0 0 ...
##  $ fatherBachelors      : int  NA 0 NA 0 0 0 NA 0 NA 0 ...
##  $ fatherWork           : int  1 1 1 1 0 1 NA 1 1 1 ...
##  $ selfBornUS           : int  1 1 1 1 1 1 0 1 1 1 ...
##  $ motherBornUS         : int  0 1 1 1 1 1 1 1 1 1 ...
##  $ fatherBornUS         : int  0 1 1 1 0 1 NA 1 1 1 ...
##  $ englishAtHome        : int  0 1 1 1 1 1 1 1 1 1 ...
##  $ computerForSchoolwork: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ read30MinsADay       : int  0 1 0 1 1 0 0 1 0 0 ...
##  $ minutesPerWeekEnglish: int  225 450 250 200 250 300 250 300 378 294 ...
##  $ studentsInEnglish    : int  NA 25 28 23 35 20 28 30 20 24 ...
##  $ schoolHasLibrary     : int  1 1 1 1 1 1 1 1 0 1 ...
##  $ publicSchool         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ urban                : int  1 0 0 1 1 0 1 0 1 0 ...
##  $ schoolSize           : int  673 1173 1233 2640 1095 227 2080 1913 502 89
9 ...
##  $ readingScore         : num  476 575 555 458 614 ...
```

Average reading test score of males and females

Male - 1 Female - 0

```
tapply(pisaTrain$readingScore,pisaTrain$male,mean,na.rm=TRUE)
```

```
##        0        1 
## 512.9406 483.5325
```

Variables that are missing data in at least one observation in the training set

```
colnames(pisaTrain)[colSums(is.na(pisaTrain))>0]
```

```
##  [1] "raceeth"                "preschool"            
##  [3] "expectBachelors"        "motherHS"             
##  [5] "motherBachelors"        "motherWork"           
##  [7] "fatherHS"               "fatherBachelors"      
##  [9] "fatherWork"             "selfBornUS"           
## [11] "motherBornUS"           "fatherBornUS"         
## [13] "englishAtHome"          "computerForSchoolwork"
## [15] "read30MinsADay"         "minutesPerWeekEnglish"
## [17] "studentsInEnglish"      "schoolHasLibrary"     
## [19] "schoolSize"
```

Linear regression discards observations with missing data, so we will remove all such observations from the training and testing sets.

```
pisaTrain=na.omit(pisaTrain)
pisaTest=na.omit(pisaTest)
colnames(pisaTrain)[colSums(is.na(pisaTrain))>0]
```

```
## character(0)
```

```
colnames(pisaTest)[colSums(is.na(pisaTest))>0]
```

```
## character(0)
```

For variable raceeth, by default, R selects first level alphabetically as shown by code.

```
str(pisaTrain$raceeth)
```

```
##  Factor w/ 7 levels "American Indian/Alaska Native",..: 7 3 4 7 5 4 7 4 7
7 ...
```

We can reset the reference level using following code.

```
pisaTrain$raceeth = relevel(pisaTrain$raceeth, "White")
pisaTest$raceeth = relevel(pisaTest$raceeth, "White")
str(pisaTrain$raceeth)
```

```
##  Factor w/ 7 levels "White","American Indian/Alaska Native",..: 1 4 5 1 6 5
1 5 1 1 ...
```

Creating Linear Model.

We can use dot(.) if we want to use all the independent variables in our model.

```
lmScore=lm(readingScore ~ . , data=pisaTrain)
summary(lmScore)
```

```
## 
## Call:
## lm(formula = readingScore ~ ., data = pisaTrain)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -247.44  -48.86    1.86   49.77  217.18 
## 
## Coefficients:
##                                                 Estimate Std. Error
## (Intercept)                                   143.766333  33.841226
## grade                                          29.542707   2.937399
## male                                          -14.521653   3.155926
## raceethAmerican Indian/Alaska Native          -67.277327  16.786935
## raceethAsian                                   -4.110325   9.220071
## raceethBlack                                  -67.012347   5.460883
## raceethHispanic                               -38.975486   5.177743
## raceethMore than one race                     -16.922522   8.496268
## raceethNative Hawaiian/Other Pacific Islander  -5.101601  17.005696
## preschool                                      -4.463670   3.486055
## expectBachelors                                55.267080   4.293893
## motherHS                                        6.058774   6.091423
## motherBachelors                                12.638068   3.861457
## motherWork                                     -2.809101   3.521827
## fatherHS                                        4.018214   5.579269
## fatherBachelors                                16.929755   3.995253
## fatherWork                                      5.842798   4.395978
## selfBornUS                                     -3.806278   7.323718
## motherBornUS                                   -8.798153   6.587621
## fatherBornUS                                    4.306994   6.263875
## englishAtHome                                   8.035685   6.859492
## computerForSchoolwork                          22.500232   5.702562
## read30MinsADay                                 34.871924   3.408447
## minutesPerWeekEnglish                           0.012788   0.010712
## studentsInEnglish                              -0.286631   0.227819
## schoolHasLibrary                               12.215085   9.264884
## publicSchool                                  -16.857475   6.725614
## urban                                          -0.110132   3.962724
## schoolSize                                      0.006540   0.002197
##                                               t value Pr(>|t|)    
## (Intercept)                                     4.248 2.24e-05 ***
## grade                                          10.057  < 2e-16 ***
## male                                           -4.601 4.42e-06 ***
## raceethAmerican Indian/Alaska Native           -4.008 6.32e-05 ***
## raceethAsian                                   -0.446  0.65578    
## raceethBlack                                  -12.271  < 2e-16 ***
## raceethHispanic                                -7.528 7.29e-14 ***
## raceethMore than one race                      -1.992  0.04651 *  
## raceethNative Hawaiian/Other Pacific Islander  -0.300  0.76421    
## preschool                                      -1.280  0.20052    
## expectBachelors                                12.871  < 2e-16 ***
## motherHS                                        0.995  0.32001    
## motherBachelors                                 3.273  0.00108 ** 
## motherWork                                     -0.798  0.42517    
## fatherHS                                        0.720  0.47147    
## fatherBachelors                                 4.237 2.35e-05 ***
## fatherWork                                      1.329  0.18393    
## selfBornUS                                     -0.520  0.60331    
## motherBornUS                                   -1.336  0.18182    
## fatherBornUS                                    0.688  0.49178    
## englishAtHome                                   1.171  0.24153    
## computerForSchoolwork                           3.946 8.19e-05 ***
## read30MinsADay                                 10.231  < 2e-16 ***
## minutesPerWeekEnglish                           1.194  0.23264    
## studentsInEnglish                              -1.258  0.20846    
## schoolHasLibrary                                1.318  0.18749    
## publicSchool                                   -2.506  0.01226 *  
## urban                                          -0.028  0.97783    
## schoolSize                                      2.977  0.00294 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 73.81 on 2385 degrees of freedom
## Multiple R-squared:  0.3251, Adjusted R-squared:  0.3172 
## F-statistic: 41.04 on 28 and 2385 DF,  p-value: < 2.2e-16
```

Note that the R-squared is lower than the usuals. This does not necessarily imply that the model is of poor quality. More often than not, it simply means that the prediction problem at hand (predicting a student's test score based on demographic and school-related variables) is more difficult than other prediction problems (like predicting a team's number of wins from their runs scored and allowed, or predicting the quality of wine from weather conditions).

Calculating Root Mean Square Error.

```
SSE = sum(lmScore$residuals^2)
RMSE = sqrt(SSE/nrow(pisaTrain))
RMSE
```

```
## [1] 73.36555
```

A alternative way of getting this answer would be with the following command:

```
sqrt(mean(lmScore$residuals^2))
```

```
## [1] 73.36555
```

Question : Consider two students A and B. They have all variable values the same, except that student A is in grade 11 and student B is in grade 9. What is the predicted reading score of student A minus the predicted reading score of student B?

```
2*lmScore$coefficients["grade"]
```

```
##    grade
## 59.08541
```

Question: What is the meaning of the coefficient associated with variable raceethAsian?

```
lmScore$coefficients["raceethAsian"]
```

```
## raceethAsian
##    -4.110325
```

The only difference between an Asian student and white student(set as reference level) with otherwise identical variables is that the former has raceethAsian=1 and the latter has raceethAsian=0. The predicted reading score for these two students will differ by the coefficient on the variable raceethAsian.

Predicting on Test data

```
predTest=predict(lmScore,newdata = pisaTest)
summary(predTest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   353.2   482.0   524.0   516.7   555.7   637.7
```

```
SSE_test = sum((predTest - pisaTest$readingScore)^2)
RMSE_test = sqrt(SSE_test/nrow(pisaTest))
SSE_test
```

```
## [1] 5762082
```

```
RMSE_test
```

```
## [1] 76.29079
```

```
baseline = mean(pisaTrain$readingScore)
SST_Test = sum((baseline - pisaTest$readingScore)^2)
SST_Test
```

```
## [1] 7802354
```

```
R2_Test = 1 - SSE_test/SST_Test
R2_Test
```

```
## [1] 0.2614944
```