PISA

Saeid Abolfazli

May 2, 2016

Credit:

** The dataset and below description is provided by MITx:** 15.071x The Analytics Edge Team @ EDX.

Description

The **Programme for International Student Assessment (PISA)** is a test given every three years to 15-year-old students from around the world to evaluate their performance in mathematics, reading, and science. This test provides a quantitative way to compare the performance of students from different parts of the world. In this homework assignment, we will predict the reading scores of students from the United States of America on the 2009 PISA exam.

The datasets *pisa2009train.csv* and *pisa2009test.csv* contain information about the demographics and schools for American students taking the exam, derived from 2009 PISA Public-Use Data Files distributed by the United States National Center for Education Statistics (NCES). While the datasets are not supposed to contain identifying information about students taking the test, by using the data you are bound by the NCES data use agreement, which prohibits any attempt to determine the identity of any student in the datasets.

Each row in the datasets pisa2009train.csv and pisa2009test.csv represents one student taking the exam. The datasets have the following variables:

- **grade:** The grade in school of the student (most 15-year-olds in America are in 10th grade)
- **male:** Whether the student is male (1/0)
- **raceeth:** The race/ethnicity composite of the student
- **preschool:** Whether the student attended preschool (1/0)
- **expectBachelors:** Whether the student expects to obtain a bachelor's degree (1/0)
- **motherHS:** Whether the student's mother completed high school (1/0)
- **motherBachelors:** Whether the student's mother obtained a bachelor's degree (1/0)
- **motherWork:** Whether the student's mother has part-time or full-time work (1/0)
- **fatherHS:** Whether the student's father completed high school (1/0)
- **fatherBachelors:** Whether the student's father obtained a bachelor's degree (1/0)
- **fatherWork:** Whether the student's father has part-time or full-time work (1/0)
- **selfBornUS:** Whether the student was born in the United States of America (1/0)
- **motherBornUS:** Whether the student's mother was born in the United States of America (1/0)
- **fatherBornUS:** Whether the student's father was born in the United States of America (1/0)
- **englishAtHome:** Whether the student speaks English at home (1/0)

- **computerForSchoolwork:** Whether the student has access to a computer for schoolwork (1/0)
- **read30MinsADay:** Whether the student reads for pleasure for 30 minutes/day (1/0)
- **minutesPerWeekEnglish:** The number of minutes per week the student spend in English class
- **studentsInEnglish:** The number of students in this student's English class at school
- **schoolHasLibrary:** Whether this student's school has a library (1/0)
- **publicSchool:** Whether this student attends a public school (1/0)
- **urban:** Whether this student's school is in an urban area (1/0)
- **schoolSize:** The number of students in this student's school
- **readingScore:** The student's reading score, on a 1000-point scale

Problem 1.1 - Dataset size

Load the training and testing sets using the read.csv() function, and save them as variables with the names pisaTrain and pisaTest.

```
PISA_train <- read.csv("data/pisa2009train.csv")
PISA_test <-  read.csv("data/pisa2009test.csv")
str(PISA_train)
```

```
## '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 Native",..: 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 502 899 ...
##  $ readingScore         : num  476 575 555 458 614 ...
```

How many students are there in the training set? **3663**

Problem 1.2 - Summarizing the dataset

Using tapply() on pisaTrain, what is the average reading test score of males? **483.5325**

```
tapply(PISA_train$readingScore,PISA_train$male ,mean)
```

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

Of females? **512.9406**

Problem 1.3 - Locating missing values

Which variables are missing data in at least one observation in the training set? Select all that apply.

```
summary(PISA_train)
```

```
##      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
```

```r
colnames(PISA_train)[colSums(is.na(PISA_train)) == 0]
```

```
## [1] "grade"        "male"         "publicSchool" "urban"
## [5] "readingScore"
```

1. grade
2. male
3. raceeth
4. preschool
5. expectBachelors
6. motherHS
7. motherBachelors
8. motherWork
9. fatherHS
10. fatherBachelors
11. fatherWork
12. selfBornUS
13. motherBornUS
14. fatherBornUS
15. englishAtHome
16. computerForSchoolwork
17. read30MinsADay

18. minutesPerWeekEnglish

19. studentsInEnglish

20. schoolHasLibrary

21. publicSchool

22. urban

23. schoolSize

24. readingScore

Answer: **3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,23**

Problem 1.4 - Removing missing values

Linear regression discards observations with missing data, so we will remove all such observations from the training and testing sets. Later in the course, we will learn about imputation, which deals with missing data by filling in missing values with plausible information.

Type the following commands into your R console to remove observations with any missing value from pisaTrain and pisaTest:

```
PISA_train <-  na.omit(PISA_train)

PISA_test = na.omit(PISA_test)
```

How many observations are now in the training set? **2414**

How many observations are now in the testing set? **990**

Problem 2.1 - Factor variables

Factor variables are variables that take on a discrete set of values, like the "Region" variable in the WHO dataset from the second lecture of Unit 1. This is an unordered factor because there isn't any natural ordering between the levels. An ordered factor has a natural ordering between the levels (an example would be the classifications "large," "medium," and "small").

Which of the following variables is an unordered factor with at least 3 levels? (Select all that apply.)

1. grade
2. male
3. raceeth

Which of the following variables is an ordered factor with at least 3 levels? (Select all that apply.)

1. grade
2. male
3. raceeth

Problem 2.2 - Unordered factors in regression models

```
str(PISA_train$grade)
```

```
##  int [1:2414] 11 10 10 10 10 10 10 10 11 9 ...
```

```
is.factor(PISA_train$grade)
```

```
## [1] FALSE
```

```
str(PISA_train$male)
```

```
##  int [1:2414] 1 0 1 0 1 0 0 0 1 1 ...
```

```
is.factor(PISA_train$male)
```

```
## [1] FALSE
```

```
str(PISA_train$raceeth)
```

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

```
is.factor(PISA_train$raceeth)
```

```
## [1] TRUE
```

```
is.ordered(PISA_train$grade)
```

```
## [1] FALSE
```

To include unordered factors in a linear regression model, we define one level as the "reference level" and add a binary variable for each of the remaining levels. In this way, a factor with n levels is replaced by n-1 binary variables. The reference level is typically selected to be the most frequently occurring level in the dataset.

As an example, consider the unordered factor variable "color", with levels "red", "green", and "blue". If "green" were the reference level, then we would add binary variables "colorred" and "colorblue" to a linear regression problem. All red examples would have colorred=1 and colorblue=0. All blue examples would have colorred=0 and colorblue=1. All green examples would have colorred=0 and colorblue=0.

Now, consider the variable "raceeth" in our problem, which has levels "American Indian/Alaska Native", "Asian", "Black", "Hispanic", "More than one race", "Native Hawaiian/Other Pacific Islander", and "White". Because it is the most common in our population, we will select White as the reference level.

Which binary variables will be included in the regression model? (Select all that apply.)

```
table(PISA_train$raceeth)
```

```
##
##          American Indian/Alaska Native
##                                     20
##                                  Asian
##                                     95
##                                  Black
##                                    228
##                               Hispanic
##                                    500
##                     More than one race
##                                     81
## Native Hawaiian/Other Pacific Islander
##                                     20
##                                  White
##                                   1470
```

raceethAmerican Indian/Alaska Native raceethAsian raceethBlack raceethHispanic raceethMore than one race raceethNative Hawaiian/Other Pacific Islander raceethWhite

Problem 2.3 - Example unordered factors

Consider again adding our unordered factor race to the regression model with reference level “White”.

For a student who is Asian, which binary variables would be set to 0? All remaining variables will be set to 1. (Select all that apply.)

raceethAmerican Indian/Alaska Native raceethAsian raceethBlack raceethHispanic raceethMore than one race raceethNative Hawaiian/Other Pacific Islander

Answer: **All except raceethAsian**

For a student who is white, which binary variables would be set to 0? All remaining variables will be set to 1. (Select all that apply.)

raceethAmerican Indian/Alaska Native raceethAsian raceethBlack raceethHispanic raceethMore than one race raceethNative Hawaiian/Other Pacific Islander

Answer: **ALL**

Problem 3.1 - Building a model

Because the race variable takes on text values, it was loaded as a factor variable when we read in the dataset with read.csv() – you can see this when you run str(pisaTrain) or str(pisaTest). However, by default R selects the first level alphabetically (“American Indian/Alaska Native”) as the reference level of our factor instead of the most common level (“White”). Set the reference level of the factor by typing the following two lines in your R console:

pisaTrain$raceeth = relevel(pisaTrain$raceeth, “White”)

pisaTest$raceeth = relevel(pisaTest$raceeth, “White”)

```
PISA_train$raceeth <- relevel(PISA_train$raceeth,"White")
PISA_test$raceeth <- relevel(PISA_test$raceeth,"White")
```

Now, build a linear regression model (call it lmScore) using the training set to predict readingScore using all the remaining variables.

```
lmScore <- lm(readingScore ~ grade + male+ raceeth+ preschool +expectBachelors+ motherHS +motherBachelor

lmScore <- lm(readingScore ~ ., data=PISA_train)
```

It would be time-consuming to type all the variables, but R provides the shorthand notation "readingScore ~ ." to mean "predict readingScore using all the other variables in the data frame." The period is used to replace listing out all of the independent variables. As an example, if your dependent variable is called "Y", your independent variables are called "X1", "X2", and "X3", and your training data set is called "Train", instead of the regular notation:

LinReg = lm(Y ~ X1 + X2 + X3, data = Train)

You would use the following command to build your model:

LinReg = lm(Y ~ ., data = Train)

What is the Multiple R-squared value of lmScore on the training set? **0.3251**

```
summary(lmScore)
```

```
##
## Call:
## lm(formula = readingScore ~ ., data = PISA_train)
##
## 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 this R-squared is lower than the ones for the models we saw in the lectures and recitation. 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).

Problem 3.2 - Computing the root-mean squared error of the model

What is the training-set root-mean squared error (RMSE) of lmScore?

```
sqrt(1/nrow(PISA_train)*sum(lmScore$residuals^2))
```

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

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

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

Problem 3.3 - Comparing predictions for similar students

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?

```
(29.542707 * 11) - (29.542707 * 9)
```

```
## [1] 59.08541
```

-59.09 -29.54 0 29.54 59.09 The difference cannot be determined without more information about the two students

Problem 3.4 - Interpreting model coefficients

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

1. Predicted average reading score of an Asian student

2. Difference between the average reading score of an Asian student and the average reading score of a white student
3. Difference between the average reading score of an Asian student and the average reading score of all the students in the dataset
4. Predicted difference in the reading score between an Asian student and a white student who is otherwise identical

Answer: **4**

The only difference between an Asian student and white student 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.

Problem 3.5 - Identifying variables lacking statistical significance

Based on the significance codes, which variables are candidates for removal from the model? Select all that apply. (We'll assume that the factor variable raceeth should only be removed if none of its levels are significant.)

grade male raceeth preschool expectBachelors motherHS motherBachelors motherWork fatherHS fatherBachelors fatherWork selfBornUS motherBornUS fatherBornUS englishAtHome computerForSchoolwork read30MinsADay minutesPerWeekEnglish studentsInEnglish schoolHasLibrary publicSchool urban schoolSize

Amswer: Look at the last coluumns of the coefficients table and select those without mar or dot.

Problem 4.1 - Predicting on unseen data

Using the "predict" function and supplying the "newdata" argument, use the lmScore model to predict the reading scores of students in pisaTest. Call this vector of predictions "predTest". Do not change the variables in the model (for example, do not remove variables that we found were not significant in the previous part of this problem). Use the summary function to describe the test set predictions.

What is the range between the maximum and minimum predicted reading score on the test set?

```
Predic <- predict(lmScore,newdata = PISA_test)
summary(Predic)
```

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

```
min(Predic) - max(Predic)
```

```
## [1] -284.4683
```

Problem 4.2 - Test set SSE and RMSE

What is the sum of squared errors (SSE) of lmScore on the testing set?

```
SSE <- sum((Predic - PISA_test$readingScore)^2)
SSE
```

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

What is the root-mean squared error (RMSE) of lmScore on the testing set?

```
RMSE <- sqrt(mean((Predic - PISA_test$readingScore)^2))
RMSE
```

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

Problem 4.3 - Baseline prediction and test-set SSE

```
SST <- sum((PISA_test$readingScore - mean(PISA_train$readingScore))^2)
SST
```

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

What is the predicted test score used in the baseline model? Remember to compute this value using the training set and not the test set.

Answer: **517.9628873**

What is the sum of squared errors of the baseline model on the testing set? HINT:** We call the sum of squared errors for the baseline model the total sum of squares (SST).

```
SST
```

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

Answer : 7802354

Problem 4.4 - Test-set R-squared

What is the test-set R-squared value of lmScore?

```
1-SSE/SST
```

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

Answer : **0.2614944**