Keiland Pullen
ID:1977120

DSC 423 | Spring 2021
Assignment 3

Honor statement:
"I have completed this work independently. The solutions given are entirely my own work"

Introduction

For this assignment, we are told that 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. The 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.

Data

The dataset contains 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). The dataset consists of 3,405 observations, each representing one student. The dataset is composed of 24 explanatory variables and 1 response variable. The response variable is labeled "readingScore" and is defined as the student's predicted reading score. The score is on a 1000 point scale, which means that 1000 is the highest score possible. The explanatory variables are 19 categorical or qualitative variables and 4 continuous or quantitative variables. There is also one column that is missing a label, the data in this column appears to be of continuous type and will be labeled as variable "X".

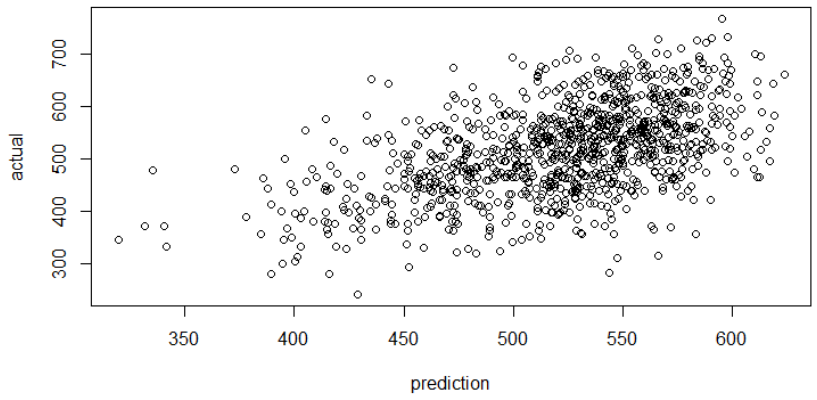
To prepare the data, several variables that are categorical data needed to be converted into factors. The race/ethnicity "raceeth" variable contains 7 levels with values "American Indian/Alaska Native", "Asian", "Black", "Hispanic", "Native Hawaiian/Other Pacific Islander", "White", and "More than one race". The variable "grade" contains 5 levels with values 8, 9, 10, 11, and 12. The remaining 17 categorical variable each have values of 0 or 1.

Training and Test Sets

To create training and test sets for N-fold validation, the following lines of code were executed. In this case, the training and test sets were created and then executed on the model after Dummy variables were created.

```
dataTrainTest <- sample(2, nrow(Pisa2009), replace=TRUE, prob=c(0.70, 0.30) )

piso2009Train <- Pisa2009[dataTrainTest == 1,]
piso2009Test  <- Pisa2009[dataTrainTest == 2,]

crossval <- lm(readingScore ~. - X -grade - raceeth - male, data=piso2009Train)
summary(crossval)

prediction <- predict(crossval,piso2009Test)
actual = piso2009Test$readingScore

cor(prediction,actual)  = 0.5230714
plot(prediction,actual)
```

The following is the plot for the Prediction vs Actual:



Dummy Variables

Due to nature of the categorical data in this dataset, dummy variables were required for the "Race" variable and the "Grade" variable. Fortunately, for 2 level variables, the R-studio application will perform the regression calculations on the factor for us. This was tested with the "Male" variable. The following code was used to create N-1 dummy variables for the N-levels of the 2 variables, "Race" and "Grade".

```
Pisa2009$grade9 <- (Pisa2009$grade == 9) * 1
Pisa2009$grade10 <- (Pisa2009$grade == 10) * 1
Pisa2009$grade11 <- (Pisa2009$grade == 11) * 1
Pisa2009$grade12 <- (Pisa2009$grade == 12) * 1

Pisa2009$raceethAsian <- (Pisa2009$raceeth == "Asian") * 1
Pisa2009$raceethBlack <- (Pisa2009$raceeth == "Black") * 1
Pisa2009$raceethHispanic1 <- (Pisa2009$raceeth == "Hispanic") * 1
Pisa2009$raceethMoreThan1 <- (Pisa2009$raceeth == "More than one race") * 1
Pisa2009$raceethNativeAmer1 <- (Pisa2009$raceeth == "Native Hawaiian/Other Pacific Islander") * 1
Pisa2009$raceethWhite1 <- (Pisa2009$raceeth == "White") * 1

Pisa2009$maleYes <- (Pisa2009$male == 1) * 1
```

The initial model was created with the following command:

```
model <- lm(readingScore ~. -X -grade - raceeth - male, data = Pisa2009)
summary(model)
```

Multicollinearity

There was an issue with executing the cor() command on the dataset due to the categorical variables. To address this issue, I had to Google the error message "X must be numeric". One of the results pages pointed to the Stack Overflow web-site which provided a code snippet which I used as a "model" to create the multicollinearity matrix:

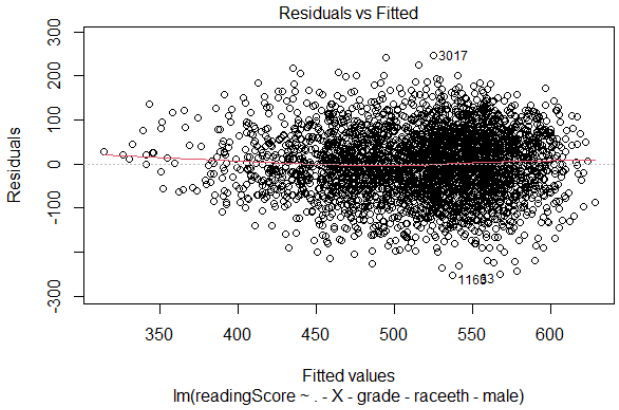
```
indx <- sapply(Pisa2009, is.factor)
Pisa2009[indx] <- lapply(Pisa2009[indx], function(x) as.numeric(as.character(x)))
cor(Pisa2009)
```

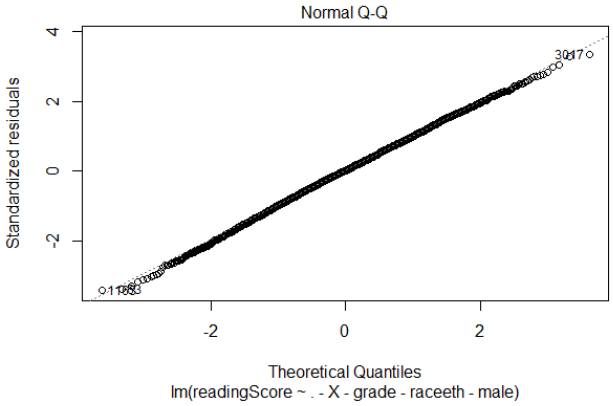
The matrix displayed a total of 36 variables. This includes the original variables including the dummy variables. The two variables with the highest percentage of collinearity are "FatherBornUS" and "MotherBornUS" at 78.4%. Both of these variables will be removed from the model. The following variables have questionable percentages for collinearity are:

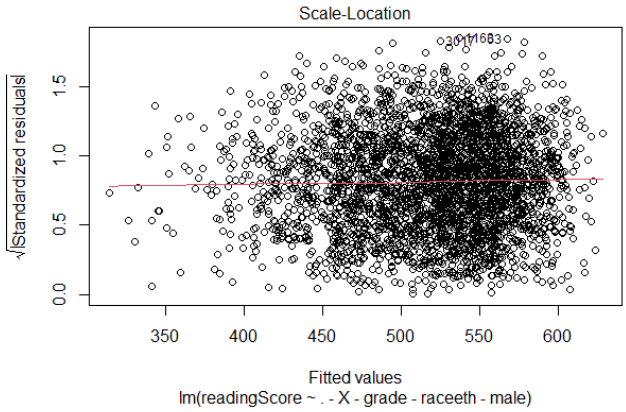
"EnglishAtHome" vs "MotherBornUS" - 66.5%
"EnglishAhHome" vs "FatherBornUS " – 64.1%
"RaceEthWhite" vs "RaceEthHispanic" – 63.9%

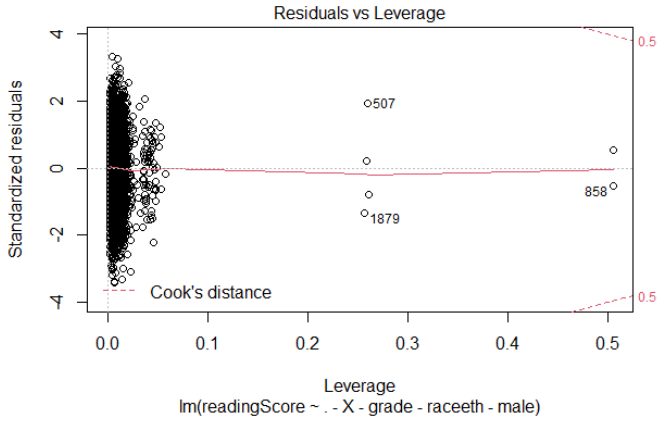
Variance Inflation Factor and Residual plots were performed on the initial model and produced the following results:

```
> vif(model)
      preschool  expectBachelors        motherHS  motherBachelors      motherWork        fatherHS
       1.079266        1.132415        1.563185        1.528714        1.064363        1.556160
  fatherBachelors      fatherWork      selfBornUS    motherBornUS    fatherBornUS   englishAtHome
       1.596956        1.049627        1.441143        3.375888        3.052399        2.285400
computerForSchoolwork read30MinsADay minutesPerWeekEnglish  studentsInEnglish  schoolHasLibrary      publicSchool
       1.111614        1.067278        1.011349        1.120310        1.043295        1.487218
        urban      schoolSize          grade9         grade10         grade11         grade12
       1.576362        1.492232      122.516806      346.895181      277.936055        3.027201
     raceethAsian    raceethBlack raceethHispanic1 raceethMoreThan1 raceethNativeAmer1  raceethWhite1
       5.819105       10.101315       19.793089        4.651023        1.894085       27.070906
         maleYes
       1.084635
```









Feature Selection

For this step, the technique that I will use for feature selection is “Step-wise”. This will be done by pruning the least significant features identified by their highest p-values, one by one, until the Adjusted R-Square value of the model is consistent.

Due to the total number of initial variables, I created a first-order model with the following lines of code:

```
model1 <- lm(readingScore ~. - X  - grade - raceeth - male + raceethAsian + raceethBlack +
raceethHispanic1 + raceethMoreThan1 + raceethNativeAmer1 + raceethWhite1
          + grade9 + grade10 + grade11 + grade12 + maleYes , data = Pisa2009)

summary(model1)
```

The summary of the first-order model is:

```
Call:
lm(formula = readingScore ~ . - X - grade - raceeth - male +
    raceethAsian + raceethBlack + raceethHispanic1 + raceethMoreThan1 +
    raceethNativeAmer1 + raceethWhite1 + grade9 + grade10 + grade11 +
    grade12 + maleYes, data = Pisa2009)

Residuals:
     Min       1Q   Median       3Q      Max
-252.698  -48.479    0.481   49.936  247.243

Coefficients:
                          Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)             299.488737  55.703871    5.376  8.11e-08  ***
preschool1               -2.008505   2.956941   -0.679  0.497026
expectBachelors1         53.227613   3.576399   14.883  < 2e-16   ***
motherHS1                4.375418    5.063943    0.864  0.387631
motherBachelors1         11.151077   3.281944    3.398  0.000687  ***
motherWork1              -2.268512   2.953436   -0.768  0.442486
fatherHS1                6.891077    4.667189    1.476  0.139905
fatherBachelors1         17.604801   3.384319    5.202  2.09e-07  ***
fatherWork1              3.033776    3.695971    0.821  0.411799
selfBornUS1              0.796308    5.976750    0.133  0.894016
motherBornUS1            -8.337533   5.669328   -1.471  0.141482
fatherBornUS1            2.556788    5.369816    0.476  0.634005
englishAtHome1           10.905428   5.835964    1.869  0.061757  .
computerForSchoolwork1   19.807206   4.856144    4.079  4.63e-05  ***
read30MinsADay1          32.736380   2.862328   11.437  < 2e-16   ***
minutesPerWeekEnglish    0.011938    0.009016    1.324  0.185567
studentsInEnglish        -0.182103   0.192915   -0.944  0.345260
schoolHasLibrary1        -1.019002   7.570382   -0.135  0.892933
publicSchool1            -18.794210  5.590012   -3.362  0.000782  ***
urban1                   -1.563926   3.320545   -0.471  0.637682
schoolSize               0.007573    0.001813    4.177  3.03e-05  ***
grade9                   40.285804   52.722943   0.764  0.444859
grade10                  90.414303   52.582799   1.719  0.085621  .
grade11                  104.977136  52.650018   1.994  0.046247  *
grade12                  153.124134  64.477284   2.375  0.017612  *
raceethAsian             59.289097   15.422137   3.844  0.000123  ***
raceethBlack             -3.245780   14.086553  -0.230  0.817782
raceethHispanic1         28.478156   13.967874   2.039  0.041545  *
raceethMoreThan1         42.834427   15.091587   2.838  0.004563  **
raceethNativeAmer1       52.643342   20.069600   2.623  0.008754  **
raceethWhite1            62.865269   13.555355   4.638  3.66e-06  ***
maleYes                  -12.629264  2.644476   -4.776  1.87e-06  ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74.07 on 3372 degrees of freedom
Multiple R-squared:  0.3161,    Adjusted R-squared:  0.3099
F-statistic: 50.29 on 31 and 3372 DF,  p-value: < 2.2e-16
```

After performing step-wise selection, the following command-line created the model that follows:

```
model1 <- lm(readingScore ~. - X  - grade - raceeth - urban - motherBornUS - fatherBornUS - selfBornUS
        - motherHS - fatherHS - preschool - minutesPerWeekEnglish – studentsInEnglish
        - schoolHasLibrary - fatherWork - motherWork- male + raceethAsian + raceethBlack
        + raceethHispanic1 + raceethMoreThan1  + raceethNativeAmer1 + raceethWhite1 + grade9
        + grade10 + grade11 + grade12 + maleYes , data = Pisa2009)

summary(model1)
```

```
Call:
lm(formula = readingScore ~ . - X - grade - raceeth - urban -
    motherBornUS - fatherBornUS - selfBornUS - motherHS - fatherHS -
    preschool - minutesPerWeekEnglish - studentsInEnglish - schoolHasLibrary -
    fatherWork - motherWork - male - raceethAsian - raceethBlack -
    raceethHispanic1 - raceethMoreThan1 - raceethNativeAmer1 -
    raceethWhite1 - grade9 - grade10 - grade11 - grade12 + maleYes,
    data = Pisa2009)

Residuals:
    Min      1Q  Median      3Q     Max
-262.079  -52.056    1.669   54.536  259.522

Coefficients:
                          Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)               416.536764  8.437070   49.370   < 2e-16 ***
expectBachelors1          57.745741   3.742382   15.430    < 2e-16 ***
motherBachelors1          12.318560   3.419747    3.602   0.000320 ***
fatherBachelors1          27.361257   3.484587    7.852   5.44e-15 ***
englishAtHome1            22.648522   4.264381    5.311   1.16e-07 ***
computerForSchoolwork1    31.578293   5.024140    6.285   3.69e-10 ***
read30MinsADay1           33.820912   3.021410   11.194   < 2e-16 ***
publicSchool1             -19.951813  5.218231   -3.823   0.000134 ***
schoolSize                0.004939    0.001684    2.932   0.003387 **
maleYes                   -13.968289  2.779922   -5.025   5.30e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 78.6 on 3394 degrees of freedom
Multiple R-squared:  0.225,          Adjusted R-squared:  0.223
F-statistic: 109.5 on 9 and 3394 DF,  p-value: < 2.2e-16
```

Interaction and Second Order Terms

For this data, Interaction or Second-Order terms did not display any significant change with the Adjusted-R-Square.

Evaluation of Final Model

For our the Final Model, the F-tests looks good as the null hypothesis was rejected and the alternative was accepted. The T-tests look good and its p-values tell us to reject the null hypothesis and accept the alternative that at least one Beta is equal to zero. What genuinely has me confused is the value of the Adjusted R-Squared. That value is 22%, which tells us that this model is NOT a good fit for the data.