Programme for International Student Assesment (PISA) Data Set Evaluation

We are trying to predict the reading scores of students from the of students of students from the United States of from the United States of America based on the 2009 PISA exam. Before we have our final model, there are many steps to take in order to ensure we have the most accurate, robust model. In the report I will be using cross validation, checking for dummy variables, checking for multicollinearity, performing feature selection, checking for interaction and second order terms, transforming variables as needed, and providing an evaluation of the final model.

Cross Validation

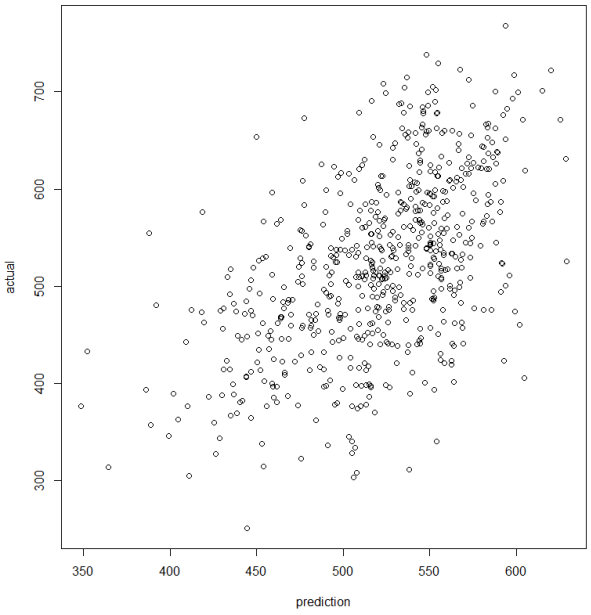
We will first be performing cross validation which involves creating a training and a testing set. Using R, it is important to know that we building our model off of the training data (split 80% training and 20% test), but evaluating the model based on the test data. By using this, we are able to see the correlation of our prediction v. actual model below:

Model 1

```
> cor(prediction,actual)
[1] 0.5404587
```

This tells us that there is 54% correlation between our prediction of the reading scores and the actual reading scores. We will be working further with our model, improve our predictions using the methods listed in the first paragraph. Below in model 1, we are able to see a plot of this correlation.

Model 2



As we can see above, there is a somewhat strong, positive correlation with some outliers they may need further looking near the first half of the x axis.

Dummy Variables

Race/Ethnicity contains seven different types, so we will need six dummy variables (k-1). Dummy variables are used for categorical data. We are able to do this using the below code which displays the below models (how to do the code was found in scholarly journal that will be cited at the end of the paper. I wanted to see other ways besides professor Gemmel's In Module 7):

Model 2

```
Pisa2009$raceeth.f<-factor(Pisa2009$raceeth)
is.factor(Pisa2009$raceeth.f)
(a<-contrasts(Pisa2009$raceeth.f))
contrasts(Pisa2009$raceeth.f)<-contr.treatment(7)
summary(lm(readingScore~raceeth.f,data = Pisa2009))
```

Model 3

```
> (a<-contrasts(Pisa2009$raceeth.f))
                                  2 3 4 5 6 7
American Indian/Alaska Native     0 0 0 0 0 0
Asian                             1 0 0 0 0 0
Black                             0 1 0 0 0 0
Hispanic                          0 0 1 0 0 0
More than one race                0 0 0 1 0 0
Native Hawaiian/Other Pacific Islander 0 0 0 0 1 0
White                             0 0 0 0 0 1
```

Model 4

```
lm(formula = readingScore ~ raceeth.f, data = Pisa2009)

Residuals:
     Min       1Q   Median       3Q      Max
-292.602  -58.887    2.518   59.276  255.320

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   439.29      15.20  28.895  < 2e-16 ***
raceeth.f2    109.13      16.80   6.495 9.50e-11 ***
raceeth.f3     25.45      15.95   1.596   0.1107
raceeth.f4     45.83      15.54   2.950   0.0032 **
raceeth.f5     72.61      17.11   4.242 2.27e-05 ***
raceeth.f6     90.04      22.51   4.000 6.47e-05 ***
raceeth.f7     97.79      15.32   6.385 1.95e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84.65 on 3397 degrees of freedom
Multiple R-squared:  0.1004,    Adjusted R-squared:  0.09878
F-statistic: 63.16 on 6 and 3397 DF,  p-value: < 2.2e-16
```

Model 3 is a matrix that shows the values given to the categorical variables. The "intercept" in Model 4 is American Indian/Alaska Native and the remaining are the dummy variables in the order of Model 3. As shown in Model 4, there is high p-value for "raceeth.f3" (Black), but since the others are low, we will keep all of the dummy variables in this model.

Multicollinearity

Before going further, we display what our model looks like at its current state. This we way we can check the f test, t test, and the adjusted R-squared. This is shown below in Model 5:

Model 5

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          378.662981  18.998202  19.932  < 2e-16 ***
male                 -15.119220   2.687626  -5.625 2.00e-08 ***
raceeth.f2            60.050964  15.645886   3.838 0.000126 ***
raceeth.f3            -5.243507  14.366468  -0.365 0.715147
raceeth.f4            26.479657  14.220641   1.862 0.062682 .
raceeth.f5            41.436651  15.392327   2.692 0.007137 **
raceeth.f6            53.728960  20.454687   2.627 0.008660 **
raceeth.f7            62.814078  13.829361   4.542 5.76e-06 ***
preschool             -1.125057   3.008742  -0.374 0.708480
expectBachelors       56.805241   3.633769  15.633  < 2e-16 ***
motherHS               3.618382   5.165591   0.700 0.483677
motherBachelors       11.064621   3.346652   3.306 0.000956 ***
motherWork            -1.394464   3.010339  -0.463 0.643233
fatherHS              10.076380   4.723582   2.133 0.032980 *
fatherBachelors       17.959631   3.452365   5.202 2.09e-07 ***
fatherWork             3.931537   3.766085   1.044 0.296591
selfBornUS             2.568901   6.056676   0.424 0.671488
motherBornUS          -8.833656   5.096202  -1.733 0.083119 .
englishAtHome         12.077792   5.851679   2.064 0.039095 *
computerForSchoolwork 23.232637   4.944363   4.699 2.72e-06 ***
read30MinsADay        33.431804   2.918570  11.455  < 2e-16 ***
minutesPerWeekEnglish  0.016249   0.009192   1.768 0.077189 .
studentsInEnglish     -0.101253   0.196507  -0.515 0.606402
schoolHasLibrary      -2.783196   7.714895  -0.361 0.718304
publicSchool         -20.062051   5.696467  -3.522 0.000434 ***
urban                 -0.537920   3.383580  -0.159 0.873694
schoolSize             0.008123   0.001846   4.400 1.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.59 on 3377 degrees of freedom
Multiple R-squared:  0.2868,    Adjusted R-squared:  0.2813
F-statistic: 52.24 on 26 and 3377 DF,  p-value: < 2.2e-16
```

Our f test shows that at least one of the betas is not 0. We can now reject the null hypothesis, and accept the alternative. The adjusted R-squared is quite low at .2813. This means that 28.13% of variability in our dependent is explained by our independent variables. If we look at the t-tests, there are 12 independent variables that are over .05 (do not pass): raceeth.f3,raceeth.f4, preschool,mothersHS,motherWork,fatherWork,SelfBornUS,motherBornUS,minutesperweekEnglish,StudentsInEnglish,schoolhaslibrary,and urban. We will leave in all raceeth variables since the others pass the t-test. The other variables will be considered to be taken out.

To further look into the accuracy of our model, we will check for multicollinearity in our independent variables. This will ensure we do not have rounding errors, incorrect beta estimates, wrong positive or negative values, or t-tests giving back incorrect information. For this, we will first run a test to check in for a variable inflation factor greater than 10. We will be using all of the independent variables un the model.

Model 6

```
> vif(modl1)
                            GVIF Df GVIF^(1/(2*Df))
male                    1.075857  1        1.037235
raceeth.f               2.482316  6        1.078710
preschool               1.073065  1        1.035888
expectBachelors         1.122642  1        1.059548
motherHS                1.562016  1        1.249806
motherBachelors         1.526504  1        1.235518
motherWork              1.061887  1        1.030479
fatherHS                1.530733  1        1.237228
fatherBachelors         1.595867  1        1.263276
fatherWork              1.046577  1        1.023023
selfBornUS              1.421210  1        1.192145
motherBornUS            2.619576  1        1.618510
englishAtHome           2.206535  1        1.485441
computerForSchoolwork 1.106635  1        1.051967
read30MinsADay          1.065594  1        1.032276
minutesPerWeekEnglish 1.009489  1        1.004733
studentsInEnglish       1.116286  1        1.056544
schoolHasLibrary        1.040506  1        1.020052
publicSchool            1.483109  1        1.217830
urban                   1.571821  1        1.253723
schoolSize              1.485374  1        1.218759
```

As shown above, since all of the variance inflation factors are below 10, we can state there is little multicollinearity between the between the independent variables.

Model Adjustment

In our next step we will remove the independent variables that do not pass the t test in model 5(besides the 'raceeth' variables). This will make sure that we are only working with variables p-test that are within our bounds (p<.05).

Model 7

```
lm(formula = readingScore ~ male + raceeth.f + expectBachelors +
    motherBachelors + fatherHS + fatherBachelors + englishAtHome +
    computerForSchoolwork + read30MinsADay + publicSchool + schoolSize,
    data = Pisa2009)

Residuals:
     Min       1Q   Median       3Q      Max
-256.944  -49.745    2.724   51.477  264.963

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         378.610499  15.837138  23.906  < 2e-16 ***
male                -14.945582   2.678330  -5.580 2.59e-08 ***
raceeth.f2           64.481504  15.293499   4.216 2.55e-05 ***
raceeth.f3           -5.565268  14.337773  -0.388 0.697926
raceeth.f4           28.742772  14.078178   2.042 0.041263 *
raceeth.f5           42.192246  15.353550   2.748 0.006027 **
raceeth.f6           57.505756  20.233939   2.842 0.004509 **
raceeth.f7           62.691881  13.792518   4.545 5.68e-06 ***
expectBachelors      56.919345   3.617913  15.733  < 2e-16 ***
motherBachelors      10.955029   3.300909   3.319 0.000914 ***
fatherHS             10.696276   4.361710   2.452 0.014244 *
fatherBachelors      18.628499   3.424245   5.440 5.70e-08 ***
englishAtHome         8.612783   4.923874   1.749 0.080349 .
computerForSchoolwork  24.014201   4.884002   4.917 9.21e-07 ***
read30MinsADay       33.580514   2.908269  11.547  < 2e-16 ***
publicSchool        -19.859581   5.057256  -3.927 8.77e-05 ***
schoolSize            0.008001   0.001658   4.827 1.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75.57 on 3387 degrees of freedom
Multiple R-squared:  0.2851,    Adjusted R-squared:  0.2817
F-statistic: 84.42 on 16 and 3387 DF,  p-value: < 2.2e-16
```

Per the above model, the adjusted r-squared is still quite low, and we will work to increase it in our next steps. The “englishAtHome” t test increased to .08, we will leave it in for now and see what happened when we create second order and interaction terms.

Feature Selection and Second Order Terms

Now we will perform feature selection with our second order terms. This will narrow down our model to the recommended number of variables we should put in our model. We will start with backwards step regression. Both of our backward and forward selection provided the recommended variables as their final model. This is shown below, in Model 8:

step3<-stepAIC(modl3,direction="backward")

step4<-stepAIC(modl4,direction="forward",scope=list(upper=modl3,lower=modl4))

(modl4 represents a blank model)

Model 8

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.760e+02  1.588e+01  23.684  < 2e-16 ***
Pisa2009$maleSQ               -1.505e+01  2.677e+00  -5.621 2.05e-08 ***
Pisa2009$expectBachelorsSQ     5.642e+01  3.624e+00  15.569  < 2e-16 ***
Pisa2009$motherBachelorsSQ     1.094e+01  3.299e+00   3.316 0.000921 ***
Pisa2009$fatherHSSQ            1.099e+01  4.362e+00   2.519 0.011812 *
Pisa2009$fatherBachelorsSQ     1.817e+01  3.429e+00   5.300 1.24e-07 ***
Pisa2009$englishAtHomeSQ       8.962e+00  4.924e+00   1.820 0.068838 .
Pisa2009$computerForSchoolworkSQ 2.348e+01  4.888e+00   4.803 1.63e-06 ***
Pisa2009$read30MinsADaySQ      3.371e+01  2.907e+00  11.594  < 2e-16 ***
Pisa2009$publicSchoolSQ       -2.139e+01  5.107e+00  -4.189 2.88e-05 ***
Pisa2009$schoolSizeSQ         -2.336e-06  1.109e-06  -2.106 0.035235 *
raceeth.f2                     6.316e+01  1.530e+01   4.129 3.74e-05 ***
raceeth.f3                    -6.991e+00  1.435e+01  -0.487 0.626070
raceeth.f4                     2.819e+01  1.407e+01   2.003 0.045226 *
raceeth.f5                     4.094e+01  1.536e+01   2.666 0.007724 **
raceeth.f6                     5.656e+01  2.023e+01   2.796 0.005204 **
raceeth.f7                     6.187e+01  1.379e+01   4.486 7.49e-06 ***
schoolSize                     1.636e-02  4.299e-03   3.805 0.000144 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.53 on 3386 degrees of freedom
Multiple R-squared:  0.286,     Adjusted R-squared:  0.2825
F-statistic:  79.8 on 17 and 3386 DF,  p-value: < 2.2e-16
```

Compared to model 7, the increase is minimal with using second order terms. We will keep model 7, and then look for interaction terms to increase the adjusted R squared.

Interaction Terms

After testing multiple terms for significant interaction, we were unable to find any that significantly increased the adjusted R-squared. One the terms we tested for interaction is motherBachelors, and fatherBachelors. This seemed intuitive since parent with the same education level may expect a certain reading score from their child. The below model shows the minimal change in adjusted R-squared by using this variable.

Model 9

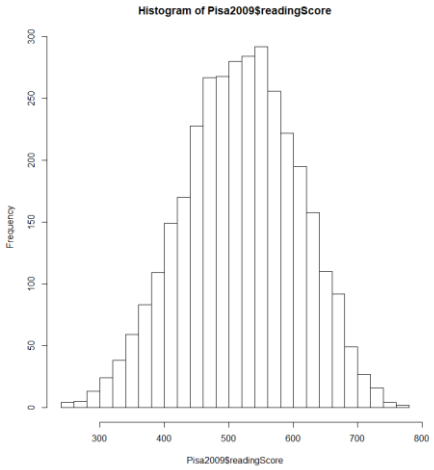
```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 377.981391  15.825104  23.885  < 2e-16 ***
male                        -14.936130   2.675992  -5.582 2.57e-08 ***
raceeth.f2                   64.466000  15.280135   4.219 2.52e-05 ***
raceeth.f3                   -5.698277  14.325333  -0.398 0.690820
raceeth.f4                   28.873637  14.065963   2.053 0.040175 *
raceeth.f5                   42.195901  15.340133   2.751 0.005979 **
raceeth.f6                   56.426887  20.220413   2.791 0.005291 **
raceeth.f7                   62.621233  13.780491   4.544 5.71e-06 ***
expectBachelors              57.099039   3.615396  15.793  < 2e-16 ***
motherBachelors               3.473244   4.354042   0.798 0.425097
fatherHS                     11.149077   4.361293   2.556 0.010620 *
fatherBachelors              10.238273   4.676231   2.189 0.028634 *
englishAtHome                 9.151159   4.923822   1.859 0.063178 .
computerForSchoolwork        24.320807   4.881125   4.983 6.59e-07 ***
read30MinsADay               33.492673   2.905920  11.526  < 2e-16 ***
publicSchool                -19.155231   5.059919  -3.786 0.000156 ***
schoolSize                    0.008008   0.001656   4.835 1.39e-06 ***
motherBachelors:fatherBachelors  17.303705   6.574439   2.632 0.008528 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75.5 on 3386 degrees of freedom
Multiple R-squared:  0.2866,    Adjusted R-squared:  0.283
F-statistic:    80 on 17 and 3386 DF,  p-value: < 2.2e-16
```

Like all of the other interactions tested, this interaction minimally increased adjusted R-squared. Since the impact is not that significant, we will take out the interaction term. It is important to only leave in terms that are significant, because if you are at a firm it can more expensive, the more variables you add. It is also important to notice that the t-test of motherBachelors significantly increased due to this interaction term, it was below .01 before and now it is at .425. That is another reason to leave out this interaction term.

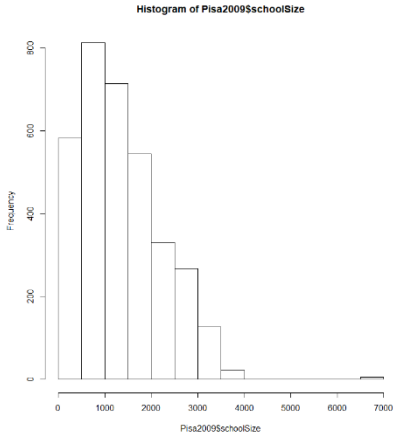
Transforming Variables

It is important to check if you now want to transform any variables. We can look at the distribution of our numerical variables to see if they need transformation. If we are working off of model 7, that leaves us with checking the distribution of readingScore (our dependent variable) and schoolSize (an independent variable). Below are the histogram for both:

Model 10

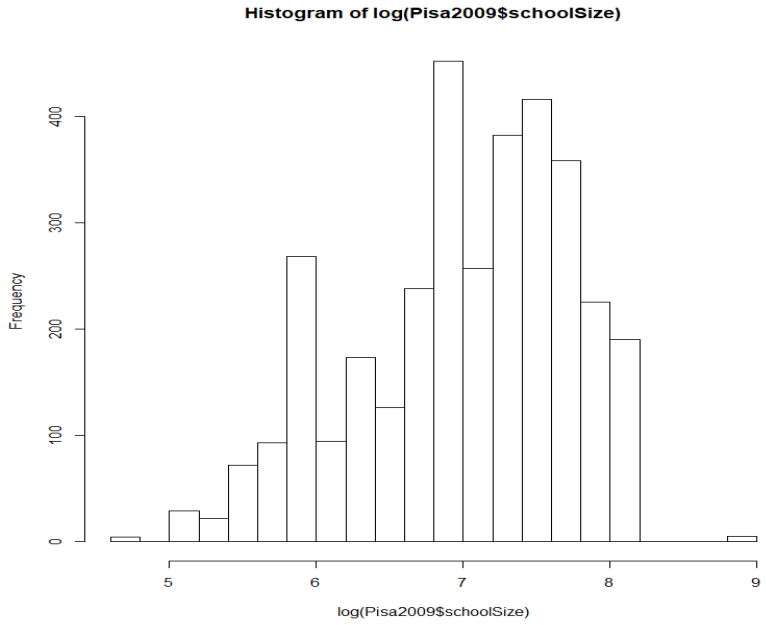


Model 11



As shown above, model 10 is the distribution of the independent variable, readingScore. This is normally distributed, so it is not necessary to log this variable. Model 11 is the distribution of the school size, which has a right a skew. Below we will show that the distribution gets closer to normal after we apply log:

Model 12



As shown above, the distribution has gone back closer to normal. The log adjusts the distribution because it takes out data points that may be extreme. We can now adjust this in our model to see what happens to the adjusted R-squared in our model below:

Model 13

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          320.095     19.337  16.553  < 2e-16 ***
male                 -15.201      2.678  -5.677 1.49e-08 ***
raceeth.f2            62.918     15.297   4.113 4.00e-05 ***
raceeth.f3            -7.717     14.347  -0.538 0.590678
raceeth.f4            27.808     14.075   1.976 0.048265 *
raceeth.f5            40.443     15.354   2.634 0.008477 **
raceeth.f6            56.049     20.230   2.771 0.005625 **
raceeth.f7            61.434     13.789   4.455 8.65e-06 ***
expectBachelors       56.527      3.619  15.620  < 2e-16 ***
motherBachelors       10.929      3.298   3.313 0.000931 ***
fatherHS              11.113      4.359   2.549 0.010834 *
fatherBachelors       18.218      3.426   5.317 1.12e-07 ***
englishAtHome          8.937      4.923   1.816 0.069533 .
computerForSchoolwork 23.403      4.888   4.788 1.76e-06 ***
read30MinsADay        33.679      2.907  11.587  < 2e-16 ***
publicSchool         -21.076      5.084  -4.146 3.47e-05 ***
log(schoolSize)       10.347      1.955   5.294 1.28e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.52 on 3387 degrees of freedom
Multiple R-squared:  0.2861,    Adjusted R-squared:  0.2827
F-statistic: 84.83 on 16 and 3387 DF,  p-value: < 2.2e-16
```

As shown above, we see the log of schoolSize has increased the adjusted R-squared from Model 7. Although the increase is not significant, we will leave in the transformed variable because it has replaced the previous variable.

Evaluating Final Model

After training / testing the data, creating appropriate dummy variables, checking for multicollinearity, performing feature selection, check for interaction/second order terms, and transforming school size our final model will be Model 13. We can see that the f-test states that the p-value is <2.2e16. We can interpret this to reject the null hypothesis, at least one of the betas is not equal to zero. The individual t-tests are all <.05 , besides English at home (which was originally <.05) and raceeth.f3. We will leave in English at home because by taking it out, it will decrease the adjusted R-squared. We will also leave in raceeth.f3 because the rest of the dummy variables pass the test. We decided to not include second order terms or interaction terms because they did not add significant value to the adjusted R-squared. Our final model's (Model 13) adjusted R-squared is .2827, which we can interpret as 28.27% of our independent variable is explained by our model.

Citations

1. “HOME.” IDRE Stats, https://stats.idre.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-models/.