Introduction

We will predict the reading scores of students from the United States of America on the 2009 PISA exam.

Goal

I would be running the cross validation across the entire Pisa2009 dataset. Using the feature selection and dummy variables we will be trying to find the best fit model. The model will include the first order, interaction terms and second order terms.

Cross Validation

Removing the variable X from Pisa2009 dataset. Creating a new dataset with name PisaS

```
PisaS <- Pisa2009[c(2:25)]
```

Separating the dataset d with 80:20 split.

```
> partition <- sample(2,nrow(PisaS),replace = TRUE, prob = c(0.80,0.20))
```

All the 1's are assigned to test.

```
> test <- PisaS[partition==1,]
```

All the 2's are assigned to train.

```
> train <- PisaS[partition==2,]
```

Building a linear model with readingScore vs. all the variables

```
> model <- lm(readingScore ~ ., data =PisaS)
```

Running prediction on model vs. test

```
> prediction <- predict(model,test)
```

Evaluating the actual based on Test dataset

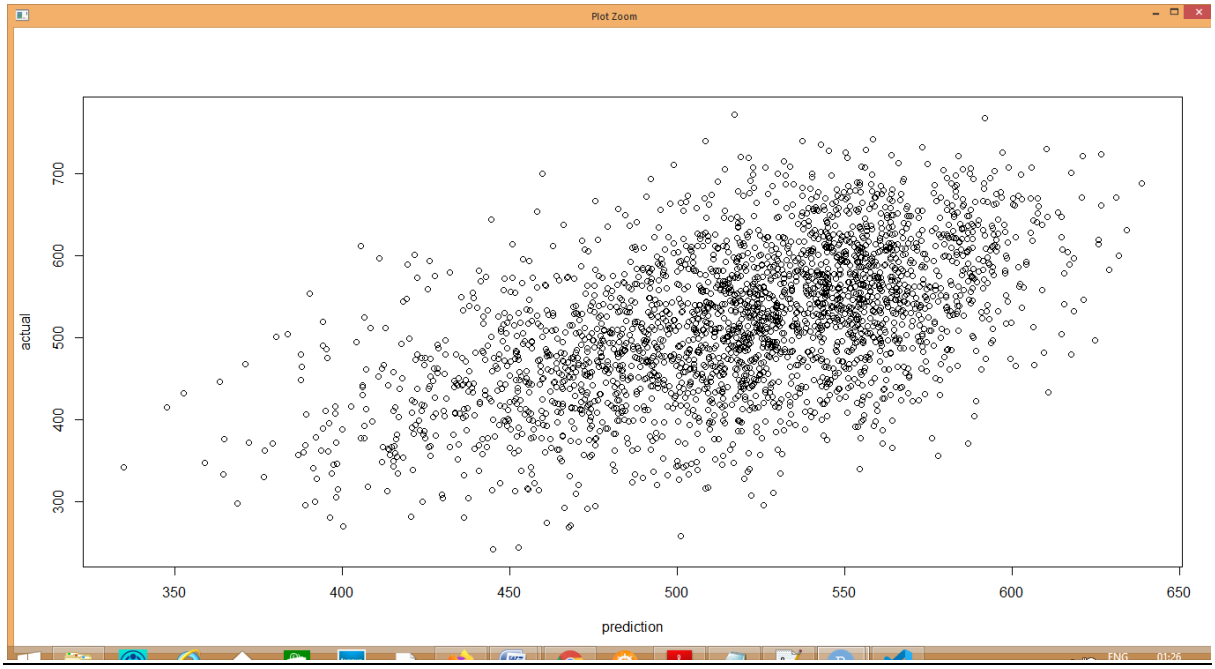
```
> actual =  test$readingScore
```

Running of correlation between prediction and model

```
> cor(prediction,actual)
[1] 0.539
```

Plotting the graph of Prediction vs. Actual

```
> plot(prediction,actual)
```



Feature selection

Removing the variable X from Pisa2009 dataset. Creating a new dataset with name Pisa

```
Pisa <- Pisa2009[,c(-1)]
```

Dummy Variable

```
> Pisa2009$raceeth <- as.numeric(Pisa2009$raceeth)
```

In order to change the categorical variable into a numeric format, as.numeric function was applied on Pisa2009$raceeth

Building Regression Model (First and Multiple)

Building the initial regression model, plotting Pisa$readingScore against all the variables present in the model

```
m1 <- lm(Pisa$readingScore ~ ., data = Pisa)
summary(m1)

Call:
lm(formula = Pisa$readingScore ~ ., data = Pisa)

Residuals:
     Min       1Q   Median       3Q      Max
-247.456  -49.395   -0.066   49.931  255.372

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        95.172545  28.678974   3.319 0.000914 ***
grade              27.293518   2.535277  10.765  < 2e-16 ***
male              -12.566135   2.680703  -4.688 2.87e-06 ***
raceeth            11.605030   0.898844  12.911  < 2e-16 ***
preschool          -0.938827   2.993488  -0.314 0.753827
expectBachelors    53.827001   3.625772  14.846  < 2e-16 ***
motherHS            4.205431   5.122872   0.821 0.411754
motherBachelors    11.058074   3.328001   3.323 0.000901 ***
```

```
motherWork              -3.392260   2.993351  -1.133 0.257183
fatherHS                11.521901   4.681570   2.461 0.013900 *
fatherBachelors         19.598693   3.422947   5.726 1.12e-08 ***
fatherWork               4.024996   3.743967   1.075 0.282424
selfBornUS               0.382719   6.017062   0.064 0.949288
motherBornUS           -14.584984   5.595561  -2.607 0.009187 **
fatherBornUS            -2.516627   5.370133  -0.469 0.639361
englishAtHome           10.394287   5.810433   1.789 0.073720 .
computerForSchoolwork   21.960592   4.916157   4.467 8.19e-06 ***
read30MinsADay          33.734783   2.903936  11.617  < 2e-16 ***
minutesPerWeekEnglish    0.014602   0.009151   1.596 0.110673
studentsInEnglish       -0.041674   0.195134  -0.214 0.830897
schoolHasLibrary        -1.183307   7.680551  -0.154 0.877567
publicSchool           -20.134293   5.664973  -3.554 0.000384 ***
urban                   -3.033645   3.361369  -0.903 0.366854
schoolSize               0.006943   0.001833   3.787 0.000155 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.24 on 3380 degrees of freedom
Multiple R-squared:  0.2928,        Adjusted R-squared:  0.288
F-statistic: 60.84 on 23 and 3380 DF,  p-value: < 2.2e-16
```

The p value of overall model is ~ 0 and below the level of significance (a = 0.05). The **F test** value at **60.86** supports this assumption. Therefore we can conclude and reject the null hypothesis and accept the alternative hypothesis that at least one of the betas is not equal to zero.

The value of **R squared** is **0.2928**. This explains that 29% of variability on dependent variable can be explained by our model.
Considering the number of variables in our model, the value of **adjusted R squared** is **0.288**.

We have number of independent variables (highlighted in Yellow) having their p value above the level of significance. All the following variables can be excluded from the model.

```
preschool
motherHS
motherWork
fatherWork
selfBornUS
fatherBornUS
englishAtHome
minutesPerWeekEnglish
studentsInEnglish
schoolHasLibrary
urban
```

We can confirm the p value for all other variables is below the level of significance and the value betas for these variables are not equal to zero.

Removing other variables with p value above the threshold

```
Pisa <- Pisa[,-c(4,6,8,11,12,14,15,18,19,20,22)]
```

Rebuilding the linear regression model using the remaining variables

```
> m1 <- lm(Pisa$readingScore ~ grade + male + raceeth  + expectBachelors  +
motherBachelors + fatherHS + fatherBachelors  + motherBornUS+ computerForSchoolwork +
read30MinsADay + publicSchool, data = Pisa)
> summary(m1)

Call:
lm(formula = Pisa$readingScore ~ ., data = Pisa)

Residuals:
     Min       1Q   Median       3Q      Max
-252.802  -49.624   -0.466   50.010  250.457

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            96.688857  26.729798   3.617 0.000302 ***
```

```
grade                  27.478764   2.526660  10.876  < 2e-16 ***
male                  -12.443507   2.671116  -4.659 3.31e-06 ***
raceeth                11.893819   0.867565  13.709  < 2e-16 ***
expectBachelors        53.891953   3.609619  14.930  < 2e-16 ***
motherBachelors        11.319782   3.281956   3.449 0.000569 ***
fatherHS               14.196926   4.231139   3.355 0.000801 ***
fatherBachelors        20.248542   3.401014   5.954 2.89e-09 ***
motherBornUS          -11.008256   3.838619  -2.868 0.004159 **
computerForSchoolwork  22.521164   4.860802   4.633 3.74e-06 ***
read30MinsADay         33.868301   2.893817  11.704  < 2e-16 ***
publicSchool          -17.756490   5.026672  -3.532 0.000417 ***
schoolSize              0.006116   0.001641   3.726 0.000198 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.23 on 3391 degrees of freedom
Multiple R-squared:  0.2907,        Adjusted R-squared:  0.2882
F-statistic: 115.8 on 12 and 3391 DF,  p-value: < 2.2e-16
```

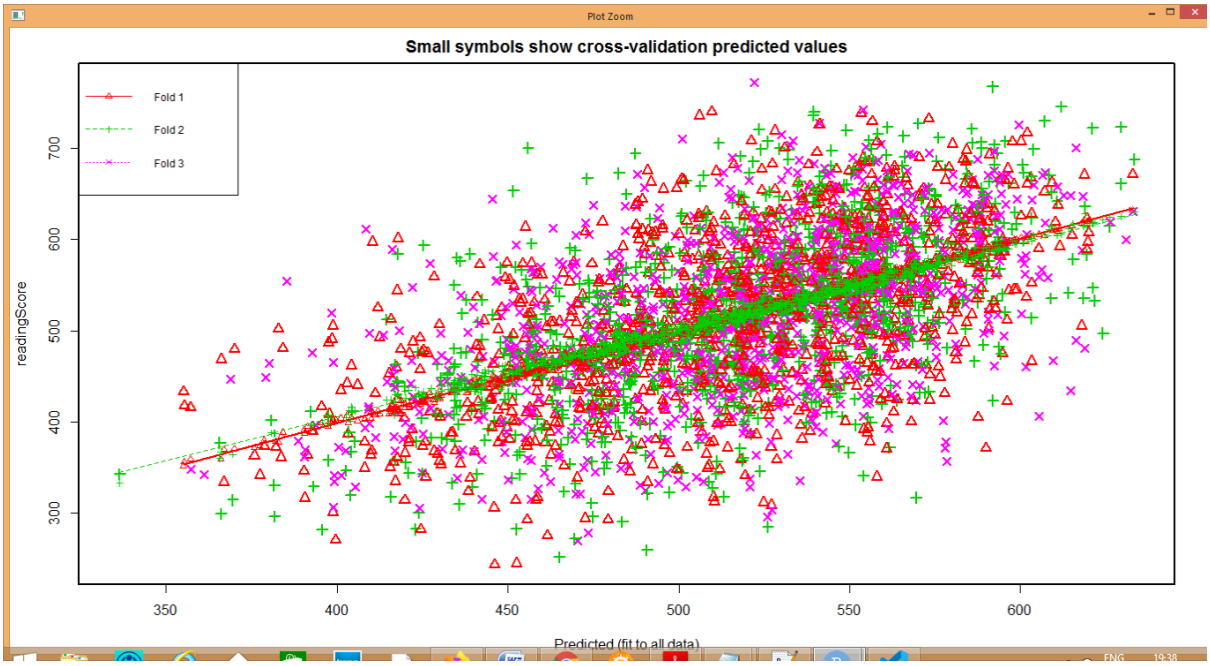
The p value of overall model is ~ 0 and below the level of significance (a = 0.05). The **F test** value at **75.23** supports this assumption. Therefore we can conclude and reject the null hypothesis and accept the alternative hypothesis that at least one of the betas is not equal to zero.

The value of **R squared** is **0.2907**. This explains that 29% of variability on dependent variable can be explained by our model.
Considering the number of variables in our model, the value of **adjusted R squared** is **0.2882**.

We can confirm the p value for all other variables is below the level of significance and the value betas for these variables are not equal to zero.

n - Fold Cross Validation

```
out<-cv.lm(data=Pisa,form.lm = formula(readingScore ~ grade + male + raceeth  +
expectBachelors  + motherBachelors + fatherHS + fatherBachelors  + motherBornUS+
computerForSchoolwork + read30MinsADay + publicSchool),plotit="Observed",m=3)
```




```
Analysis of Variance Table

Response: readingScore
                       Df   Sum Sq Mean Sq F value  Pr(>F)
```

```
grade                    1  1335657 1335657  236.02 < 2e-16 ***
male                     1   278021  278021   49.13 2.9e-12 ***
raceeth                  1  1749886 1749886  309.22 < 2e-16 ***
expectBachelors          1  2480948 2480948  438.40 < 2e-16 ***
motherBachelors          1   512809  512809   90.62 < 2e-16 ***
fatherHS                 1   118264  118264   20.90 5.0e-06 ***
fatherBachelors          1   314171  314171   55.52 1.2e-13 ***
motherBornUS             1    86402   86402   15.27 9.5e-05 ***
computerForSchoolwork    1   111997  111997   19.79 8.9e-06 ***
read30MinsADay           1   761305  761305  134.53 < 2e-16 ***
publicSchool             1    36922   36922    6.52  0.0107 *
schoolSize               1    78575   78575   13.88  0.0002 ***
Residuals             3391 19189875    5659
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Overall (Sum over all 1135 folds)
    ms
319129
```

The average mean square error for 3 fold cross validation is 319129.

Correlation

The correlation function on Pisa2009 shows 1 independent variables (highlighted in Yellow) having high correlation between variables (above 0.5).

motherBachelors and fatherBachelors

```
> cor(Pisa)
                          grade         male      raceeth expectBachelors
grade                  1.00000000 -0.088509655 -0.023883350     0.115848353
male                  -0.08850965  1.000000000  0.020436746    -0.092327173
raceeth               -0.02388335  0.020436746  1.000000000     0.033879941
expectBachelors        0.11584835 -0.092327173  0.033879941     1.000000000
motherBachelors        0.03535829  0.052540996  0.158999759     0.177168825
fatherHS               0.05552169  0.028284741  0.229714049     0.160542581
fatherBachelors        0.05796257  0.058504910  0.170621567     0.220152567
motherBornUS          -0.07373164  0.000600294  0.497586309    -0.001410829
computerForSchoolwork  0.08356420 -0.017935135  0.086565810     0.153391509
read30MinsADay         0.04119317 -0.200024132 -0.008331321     0.113815504
publicSchool          -0.04858833 -0.088921910 -0.048847234    -0.109910940
schoolSize             0.06804436 -0.002999718 -0.197084800     0.038534475
readingScore           0.22219025 -0.120639795  0.247034181     0.343325644
                      motherBachelors    fatherHS fatherBachelors motherBornUS
grade                     0.035358291  0.05552169      0.05796257 -0.073731639
male                      0.052540996  0.02828474      0.05850491  0.000600294
raceeth                   0.158999759  0.22971405      0.17062157  0.497586309
expectBachelors           0.177168825  0.16054258      0.22015257 -0.001410829
motherBachelors           1.000000000  0.20296915      0.55020342  0.133454617
fatherHS                  0.202969145  1.00000000      0.27239147  0.316446910
fatherBachelors           0.550203420  0.27239147      1.00000000  0.070311623
motherBornUS              0.133454617  0.31644691      0.07031162  1.000000000
computerForSchoolwork     0.137948974  0.16505635      0.16002405 -0.002305598
read30MinsADay            0.029851359  0.03886818      0.04837057  0.014734711
publicSchool             -0.186334800 -0.08390245     -0.19195167  0.016657713
schoolSize               -0.003737008 -0.08071990      0.02060398 -0.244063789
readingScore              0.228639885  0.19503890      0.27895304  0.073225129
                      computerForSchoolwork read30MinsADay publicSchool
grade                          0.083564197    0.041193166  -0.04858833
male                          -0.017935135   -0.200024132  -0.08892191
raceeth                        0.086565810   -0.008331321  -0.04884723
expectBachelors                0.153391509    0.113815504  -0.10991094
motherBachelors                0.137948974    0.029851359  -0.18633480
fatherHS                       0.165056348    0.038868183  -0.08390245
fatherBachelors                0.160024047    0.048370566  -0.19195167
motherBornUS                  -0.002305598    0.014734711   0.01665771
computerForSchoolwork          1.000000000   -0.019575314  -0.07161030
read30MinsADay                -0.019575314    1.000000000   0.01037715
publicSchool                  -0.071610301    0.010377147   1.00000000
schoolSize                     0.066657923   -0.015735763   0.25831668
readingScore                   0.178640095    0.224203070  -0.11865063
                      schoolSize readingScore
grade                 0.068044358   0.22219025
male                 -0.002999718  -0.12063980
raceeth              -0.197084800   0.24703418
```

```
expectBachelors         0.038534475   0.34332564
motherBachelors        -0.003737008   0.22863989
fatherHS               -0.080719896   0.19503890
fatherBachelors         0.020603982   0.27895304
motherBornUS           -0.244063789   0.07322513
computerForSchoolwork   0.066657923   0.17864009
read30MinsADay         -0.015735763   0.22420307
publicSchool            0.258316680  -0.11865063
schoolSize              1.000000000   0.03022833
readingScore            0.030228332   1.00000000
```

VIF

```
> vif(m1)
                 Pisa2009$grade                   Pisa2009$male
                       1.045977                       1.080319
              Pisa2009$raceeth              Pisa2009$preschool
                       1.490615                       1.072133
      Pisa2009$expectBachelors               Pisa2009$motherHS
                       1.128145                       1.550637
      Pisa2009$motherBachelors             Pisa2009$motherWork
                       1.523636                       1.059743
             Pisa2009$fatherHS        Pisa2009$fatherBachelors
                       1.517669                       1.583438
          Pisa2009$fatherWork             Pisa2009$selfBornUS
                       1.043980                       1.415781
        Pisa2009$motherBornUS           Pisa2009$fatherBornUS
                       3.187589                       2.958987
       Pisa2009$englishAtHome Pisa2009$computerForSchoolwork
                       2.195858                       1.104264
      Pisa2009$read30MinsADay Pisa2009$minutesPerWeekEnglish
                       1.064788                       1.009932
    Pisa2009$studentsInEnglish       Pisa2009$schoolHasLibrary
                       1.111020                       1.040894
         Pisa2009$publicSchool                  Pisa2009$urban
                       1.480455                       1.565741
           Pisa2009$schoolSize
                       1.478538
```

None of the variables have a VIF value greater than 10. We can confirm there is no multicollinearity amongst the independent variables.

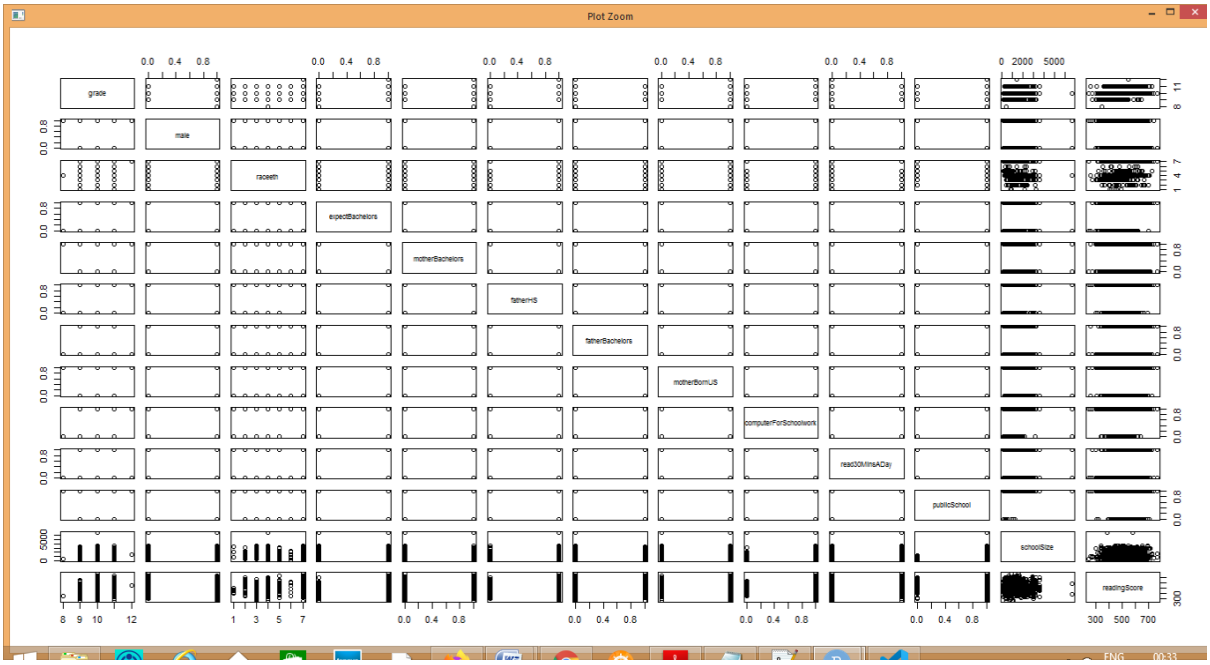
Second Order Terms

Sampling 1000 records from the dataset and plotting the graph.

```
PisaSample <- Pisa[sample(1:nrow(d),1000,replace = FALSE),]

plot(PisaSample)
```

Following is the plot of 1000 samples of Pisa (modified dataset).



None of the above plots show a curvature. Therefore we can confirm there are no second order terms.

Interaction Terms

Following the trial and error of multiple interaction terms, we can conclude the following terms are the best fit.

```
m2 <- lm(Pisa$readingScore ~ grade + male + raceeth  + expectBachelors  +
motherBachelors + fatherHS + fatherBachelors  + motherBornUS+ computerForSchoolwork +
read30MinsADay + publicSchool + grade*publicSchool + publicSchool*schoolSize
+raceeth*motherBornUS + motherBachelors*fatherHS , data = Pisa)
> summary(m2)

Call:
lm(formula = Pisa$readingScore ~ . + grade * publicSchool + publicSchool *
    schoolSize + raceeth * motherBornUS + motherBachelors * fatherHS,
    data = Pisa)

Residuals:
     Min       1Q   Median       3Q      Max
-251.334  -49.057    0.648   50.247  248.602

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             366.20819   87.88292   4.167 3.16e-05 ***
grade                     3.24452    8.51236   0.381  0.70311
male                    -12.79353    2.66159  -4.807 1.60e-06 ***
raceeth                   2.52041    1.94062   1.299  0.19411
expectBachelors          53.52213    3.58486  14.930  < 2e-16 ***
motherBachelors         -21.82354   11.52574  -1.893  0.05838 .
fatherHS                 10.21404    4.44540   2.298  0.02164 *
fatherBachelors          17.65069    3.41146   5.174 2.43e-07 ***
motherBornUS            -61.47343   10.32040  -5.956 2.84e-09 ***
computerForSchoolwork    22.71560    4.84762   4.686 2.90e-06 ***
read30MinsADay           33.66398    2.87654  11.703  < 2e-16 ***
publicSchool           -259.12719   91.12432  -2.844  0.00449 **
schoolSize                0.03819    0.01401   2.726  0.00644 **
grade:publicSchool       25.63423    8.89489   2.882  0.00398 **
publicSchool:schoolSize  -0.03225    0.01412  -2.284  0.02243 *
raceeth:motherBornUS     11.44344    2.16554   5.284 1.34e-07 ***
motherBachelors:fatherHS 35.52363   11.94912   2.973  0.00297 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 74.7 on 3387 degrees of freedom
Multiple R-squared:  0.3014,        Adjusted R-squared:  0.2981
F-statistic: 91.33 on 16 and 3387 DF,  p-value: < 2.2e-16
```

We have included the following four interaction terms (highlighted in Yellow) in the model.

```
grade:publicSchool
publicSchool:schoolSize
raceeth:motherBornUS
motherBachelors:fatherHS
```

The p value of overall model is ~ 0 and below the level of significance (a = 0.05). The **F test** value at **91.33** supports this assumption. Therefore we can conclude and reject the null hypothesis and accept the alternative hypothesis that at least one of the betas is not equal to zero.

The value of **R squared** is **0.3014**. This explains that 30% of variability on dependent variable can be explained by our model.
Considering the number of variables in our model, the value of **adjusted R squared** is **0.2981**.
There is a slight improvement in the value of adjusted R squared compared to the initial model.

We have some independent variables (highlighted in Blue) with p value above the level of significance (a = 0.05).

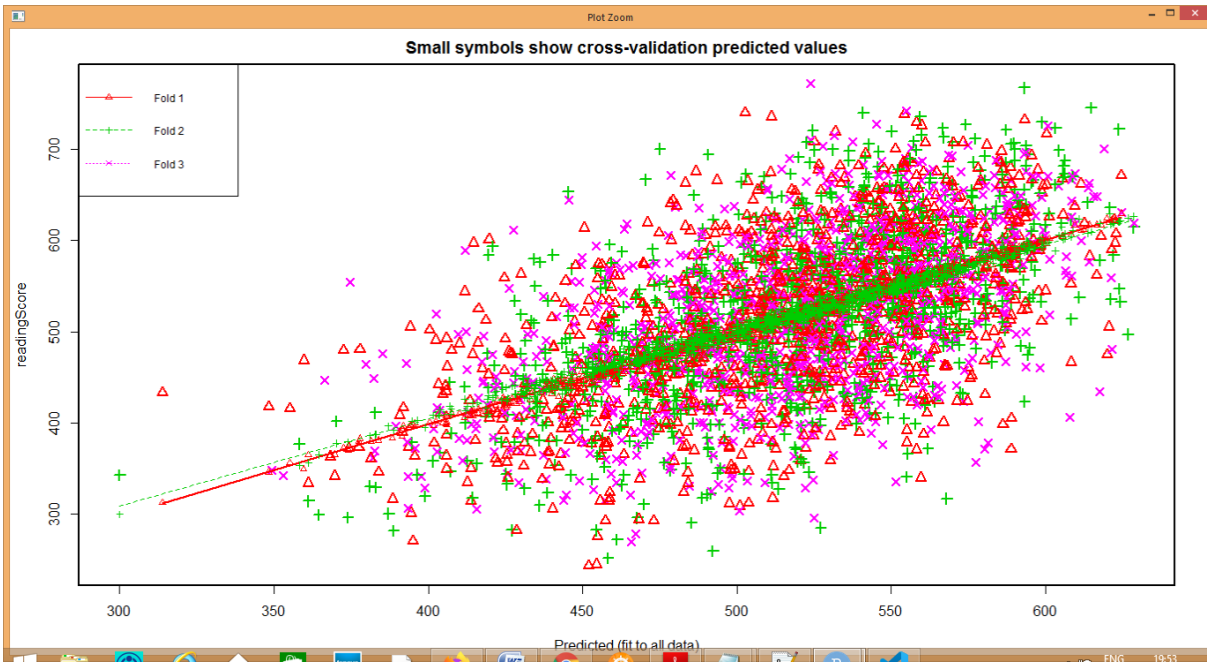
```
grade
raceeth
motherBachelors
```

Since the interaction terms were included in the model, we will include these variables in the model.

We can confirm the p value for all other variables is below the level of significance (a = 0.05) and therefore we can conclude the betas for these variables are not equal to zero.

n - Fold Cross Validation

```
out<-cv.lm(data=Pisa,form.lm = formula(readingScore ~ grade + male + raceeth  +
expectBachelors  + motherBachelors + fatherHS + fatherBachelors  + motherBornUS+
computerForSchoolwork + read30MinsADay + publicSchool + grade*publicSchool +
publicSchool*schoolSize +raceeth*motherBornUS + motherBachelors*fatherHS
),plotit="Observed",m=3)
```



```
Analysis of Variance Table

Response: readingScore
                          Df   Sum Sq Mean Sq F value  Pr(>F)
grade                      1  1335657 1335657  239.35 < 2e-16 ***
male                       1   278021  278021   49.82 2.0e-12 ***
raceeth                    1  1749886 1749886  313.59 < 2e-16 ***
expectBachelors            1  2480948 2480948  444.60 < 2e-16 ***
motherBachelors            1   512809  512809   91.90 < 2e-16 ***
fatherHS                   1   118264  118264   21.19 4.3e-06 ***
fatherBachelors            1   314171  314171   56.30 7.9e-14 ***
motherBornUS               1    86402   86402   15.48 8.5e-05 ***
computerForSchoolwork      1   111997  111997   20.07 7.7e-06 ***
read30MinsADay             1   761305  761305  136.43 < 2e-16 ***
publicSchool               1    36922   36922    6.62 0.01015 *
schoolSize                 1    78575   78575   14.08 0.00018 ***
grade:publicSchool         1    45505   45505    8.15 0.00432 **
publicSchool:schoolSize    1    30850   30850    5.53 0.01877 *
raceeth:motherBornUS       1   163922  163922   29.38 6.4e-08 ***
motherBachelors:fatherHS   1    49319   49319    8.84 0.00297 **
Residuals               3387 18900280    5580
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Overall (Sum over all 1135 folds)
    ms
313293
```

The average mean square error for 3 fold cross validation is 313293.

Summary

Final Model:

```
lm(Pisa$readingScore ~ grade + male + raceeth  + expectBachelors  + motherBachelors +
fatherHS + fatherBachelors  + motherBornUS+ computerForSchoolwork + read30MinsADay +
publicSchool + grade*publicSchool + publicSchool*schoolSize +raceeth*motherBornUS +
motherBachelors*fatherHS , data = Pisa)
```

Our final model has 13 independent variables and 4 interaction variables. The p value of overall model is ~ 0 and below the level of significance (a = 0.05). The **F test** value at **91.33** supports this assumption.

The value of **R squared** is **0.3014**. This explains that 30% of variability on dependent variable can be explained by our model.

Considering the number of variables in our model, the value of **adjusted R squared** is **0.2981**.

We can confirm the p value for all other variables is below the level of significance (a = 0.05) and therefore we can conclude the betas for these variables are not equal to zero.