Patrick Keener
SID: 1385832
DSC 423 – Assignment 2

Assignment 3

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

Overview

The Programme for International Student Assessment (PISA) is given every 3 years to 15-year-old students from around the world to evaluate their math, reading, and science skills, facilitating performance comparisons between countries. The goal of this analysis is to predict reading scores from the US on the 2009 PISA exam. Information contained in the dataset cover demographics of the students and schools. Each row represents one student taking the exam.

The analysis should contain documentation describing:

- Variable transformations
- Feature selection
- Appropriate dummy variable creation
- Interaction and Second order term analysis
- Multicollinearity testing and analysis
- N-fold cross validation
- Evaluation of final model

Feature Engineering

The PISA dataset contains 3404 observations on 24[1] variables covering demographics of the students and their schools, as well as the student's reading score. All observations were complete.

Feature Generation

After familiarization with the available features, other variables that could be derived from existing variables were considered.

Four separate binary variables were included that indicated whether mother or father had obtained an HS or Bachelor's degree. These were combined in various ways such has "mother's highest level of education", and second-order derivations such as "highest level of education obtained by any parent".

Additional feature subjects include number of parents born in the US, the number of parents that work, and the number of years of schooling completed (combining pre-school and HS level).

Typically, these variables had superior explanatory power than the variables that they were created from, as well as reduced collinearity, with the exception of the number of years in school which provided no additional explanatory power over separate variables.

[1] The data set came with an index column that was removed.

Patrick Keener
SID: 1385832
DSC 423 – Assignment 2

Data-Typing

Each variable was examined to determine whether it should be treated as a continuous or categorical variable. With addition of the newly generated features, 34 were examined in total. Of these 34 variables, 4 were treated as numerical (2 numeric and 2 integers) and 30 were categorical (factors). Of particular interest, the dependent variable (readingScore) is numeric, and grade was converted to a factor. While a binary variable is not strictly required to be set as a factor to perform adequately in a regression, the designation allows more elegant use of R's linear model functionality and thus they were converted.

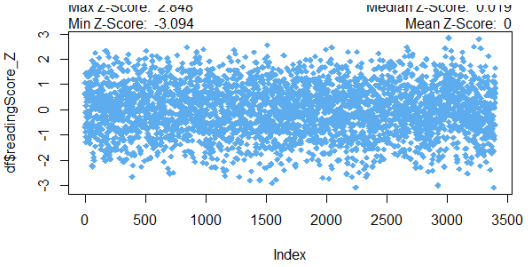
Table 1 Variable Types Post-Conversion

Variable	Type	Variable	Type	Variable	Type
readingScore	numeric	selfBornUS	factor	schoolSize	integer
grade	factor	motherBornUS	factor	mothersEducation	factor
male	factor	fatherBornUS	factor	fathersEducation	factor
raceeth	factor	englishAtHome	factor	familyEducation	factor
preschool	factor	computerForSchoolwork	factor	numParentBachelors	factor
expectBachelors	factor	read30MinsADay	factor	parentBachelors	factor
motherHS	factor	minutesPerWeekEnglish	integer	parentHS	factor
motherBachelors	factor	studentsInEnglish	integer	parentHighestEdu	factor
motherWork	factor	schoolHasLibrary	factor	numParentsBornUS	factor
fatherHS	factor	publicSchool	factor	numParentsWork	factor
fatherBachelors	factor	urban	factor	yrsInSchool	factor
fatherWork	factor				

Outlier Analysis

Understanding outliers is an important part of the analysis process. Outliers can distort distributions and cause poor model performance. Analysis of the numerical variables shows that the dependent variable (Figure 1) does not have any outliers. Analysis of the other numeric variables found a significant number of 4-15 standard deviation observations which implies normalization is necessary.

Figure 1 Dependent Variable Normalized



Variable Normalization

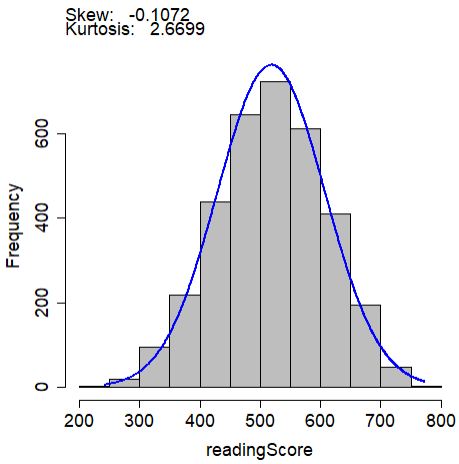
Linear Regressions assumes that the underlying data is normally distributed. As part of the data preparation, each numeric variable is assessed for normality and transformed if necessary.

Patrick Keener
SID: 1385832
DSC 423 – Assignment 2

There are 4 numeric variables to assess include the Dependent variable: readingScore, minutesPerWeekEnglish, studentsInEnglish, and schoolSize.

Reading score is perhaps the most important variable to get right, so a normality analysis is conducted despite showing normality in its z-score distribution. Figure 2 is a histogram of readingScore and shows a skew of .11 and kurtosis of 2.67, which is very well distributed. Therefore, the dependent variable does not need any type of normalization.

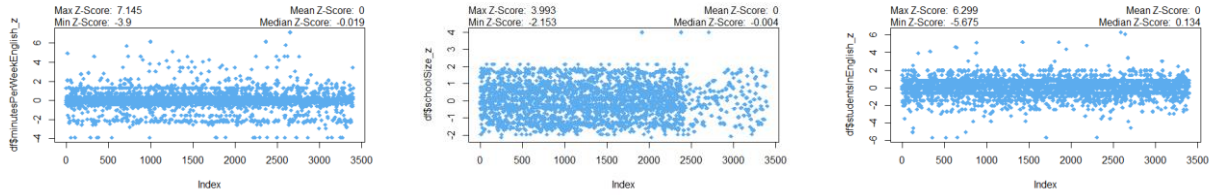
Figure 2 Dependent Variable Distribution



The other 3 numeric variables were assessed for normality and each has only marginal improvements, the largest being in schoolSize which was improved by taking its square root. The variable studentsInEnglish was insignificant, and the other two were significant but explained less than 1% of the variance on their own.

After transforming the variables, the distributions were re-assessed and still found to be abnormal. Four outliers were identified for schoolSize (center chart, below), however the impact on p-value and r-squared was meaningless, therefore to preserve data in other variables these outliers were not removed. Overall, analysis of the numeric variables did not appear promising.

Figure 3 Numeric Variables Normalized



Additional analysis was conducted on these variables to determine their p-value and R^2, results can be found in table Table 2, below.

Table 2 Analysis of Numeric Variables

Variable	p-value	R^2
studentsInEnglish	.28	0
minutesPerWeekEnglish	.0086	0
schoolSize	.0011	.028

Feature Selection and Model Building

Once the dataset is prepared the features may be selected for inclusion in the model. This process includes analyzing each variable individually (univariate analysis) then together (multivariate analysis). Models are evaluated first on the f-tests, then each variable is assessed for significance with t-tests. If the model passes these tests the R^2 is evaluated to gauge the utility of the model. If the R^2 is reasonable then a VIF test is performed, and finally validation is conducted to evaluate the robustness of the model.

Patrick Keener
SID: 1385832
DSC 423 – Assignment 2

Univariate Analysis

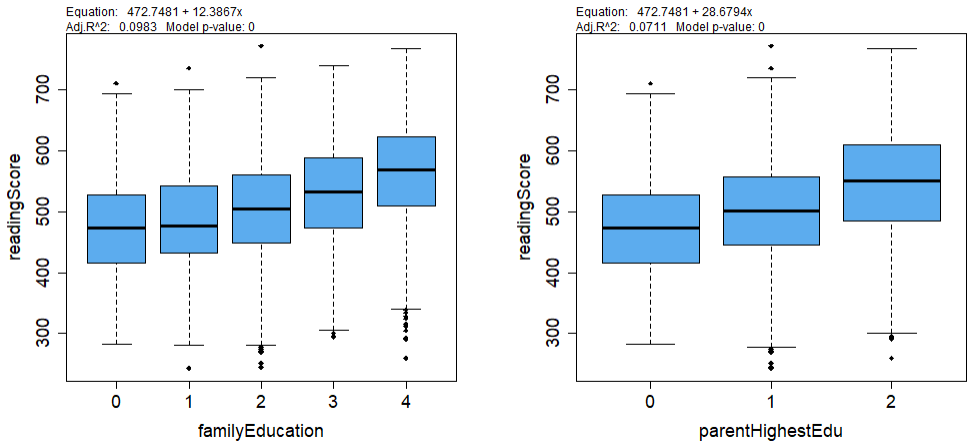
A visual inspection of the outputs is helpful in identifying candidate variables. Analysis of numeric variables was conducted during the transformation phase; this section pertains only to categorical variables.

Each variable was plotted against the dependent variable and evaluated for significance and explainability. 29 box plots were generated, and only variables with a p-value below .05 were included in the analysis. The excluded variables were "urban" which had a p-value of .445 and R^2 of 0, and schoolHasLibrary, which had a p-value of .936 and R^2 of 0.

These 27 variables were filtered to remove those with an R^2 < .01 which removed another 8 variables. The remaining 21 variables were sorted into logical groupings that were unlikely to overlap, and 1 variable from each group was selected. For example, public schools typically have larger class sizes, so both class size and public/private school variables would not be selected. Additionally, any variables that were used in the construction of a variable would naturally be included in the same group, so if family level of education was selected, mother's or father's level of academic achievement would not be included.

The larges surprise during this level of analysis was when analyzing the effectiveness of "family education" against "parent highest education". The first variable gives 1 point per level achieved, so if mother graduated from high school and college this would be worth 1 point each, and if father graduated from high school the score would be 3, which is in contrast to highest education, which simply checks the highest level of education of any parent. This was surprising because many studies seem to use the latter, whereas the former is more predictive (.0711 R^2 vs .0983 R^2, respectively; both have p-values of 0).

Figure 4 Comparison of Family Education to Parent Highest Education



Categorical variables do not lend themselves to other forms of univariate analysis, so the following 10 variables were selected for further analysis:

Patrick Keener
SID: 1385832
DSC 423 – Assignment 2

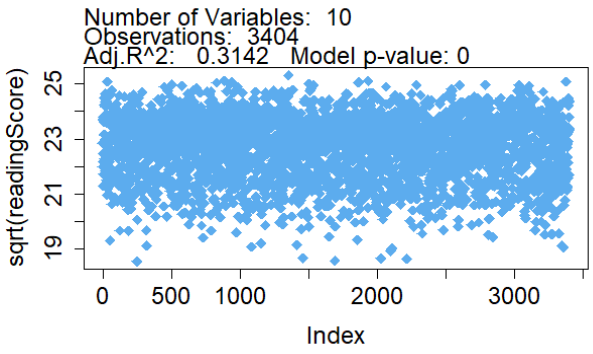
Variable	p-value	R^2
Male	.0143	0
minutesPerWeekEnglish[2]	.0086	0
FamilyEducation	.0983	0
Grade	.0668	0
raceEth	.0988	0
englishAtHome	.016	0
publicSchool	.0138	0
computerForSchoolwork	.0316	0
Read30MinsADay	.05	0
expectBachelors	1176	0

Multivariate Analysis

Multivariate analysis would typically begin with a correlation matrix; however, this is not practical given the large number of categorical variables. Instead, forward and backward StepAIC is used to set a baseline, the results of which can be found in the appendix.

The stepAIC revealed that the variable schoolSize_sqrt, which had been ruled out in the first step, may in fact have explanatory power worth exploring. After several iterations, the following model made it to the top of testing:

$$\sqrt{readingScore} = read30MinsADay + male + raceeth + computerForSchoolwork + familyEducation + expectBachelors + grade + publicSchool + \sqrt{minutesPerWeekEnglish} + \sqrt{schoolSize} + familyEducation * expectBachelors + grade * publicSchool$$



VIF analysis was performed on each model (forward/backward selection and the final model). None of the VIF scores exceeded 4 with the exception of those involved in interaction or second-order terms. The VIF scores for the final model can be found in the appendix.

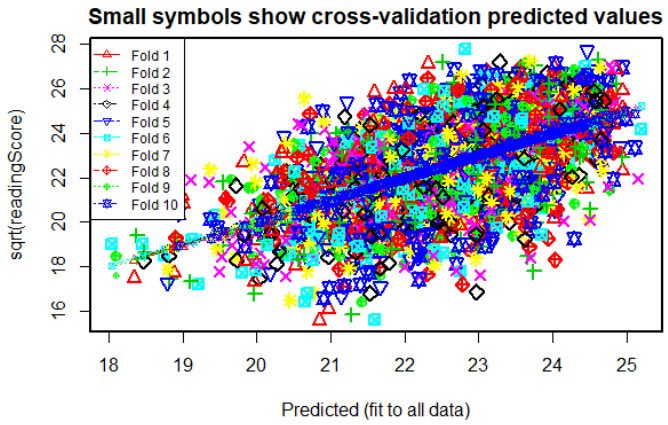
[2] MPW English was included even though its R^2 was below the threshold so a continuous variable would be available for testing. MPW English and englishAtHome are similar and would not be considered for inclusion in the same model.

Patrick Keener
SID: 1385832
DSC 423 – Assignment 2

Validation

The final step is to ensure the model is robust through different data sets, and not overfitted. To accomplish this, the library Caret was used in order to find the performance metrics: Adj. R^2, RMSE, and MAE. Additionally, the library DAAG was used to generate a plot as well as provide a second analysis. DAAG agreed with Caret's evaluation.

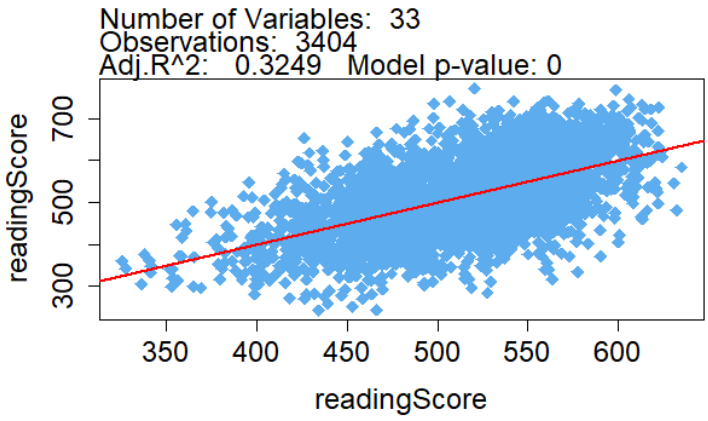
Fold	1	2	3	4	5	6	7	8	9	10
Adj.R^2	0.321	0.313	0.354	0.256	0.295	0.32	0.346	0.31	0.293	0.391
RMSE	1.65	1.65	1.61	1.68	1.69	1.66	1.68	1.61	1.65	1.56
MAE	1.29	1.3	1.27	1.33	1.34	1.32	1.33	1.28	1.31	1.21



Final Model Evaluation

Overall, the model had an averaged RMSE of 1.64, Adj.R^2 of .32, and MAE of 1.3. Furthermore, a VIF test was conducted (found in the relevant appendix) and found that the only variables exceeding the threshold of 10 were those that were expected: interaction terms and dummy variables. The F-Statistic of the final model is 59.5 with a p-value of 0, and the t-tests showed all variables at significant levels: 2nd order variables and 1st order variables without corresponding 2nd order variables had p-values below .05.

Figure 5 Final Model



Patrick Keener
SID: 1385832
DSC 423 – Assignment 2

Appendices

Final Model

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.64 on 3375 degrees of freedom
Multiple R-squared:  0.33,    Adjusted R-squared:  0.325
F-statistic: 59.5 on 28 and 3375 DF,  p-value: <2e-16
```

```
lm(formula = sqrt(readingScore) ~ read30MinsADay + male + raceeth +
    computerForSchoolwork + minutesPerWeekEnglish_sqrt + schoolSize_sqrt +
    familyEducation * expectBachelors + grade * publicSchool,
    data = df2)

Residuals:
   Min     1Q Median     3Q    Max
-6.100 -1.022  0.088  1.118  4.979
```

VIF Test

	Variables	Tolerance	VIF
1	read30MinsADay1	0.9409	1.06
2	male1	0.9256	1.08
3	raceethAsian	0.1836	5.45
4	raceethBlack	0.0991	10.09
5	raceethHispanic	0.0517	19.32
6	raceethMore than one race	0.2156	4.64
7	raceethNative Hawaiian/Other Pacific Islander	0.5367	1.86
8	raceethWhite	0.0371	26.98
9	computerForSchoolwork1	0.9065	1.10
10	minutesPerWeekEnglish_sqrt	0.9840	1.02
11	schoolSize_sqrt	0.8254	1.21
12	familyEducation1	0.1457	6.87
13	familyEducation2	0.0682	14.67
14	familyEducation3	0.0705	14.19
15	familyEducation4	0.0373	26.84
16	expectBachelors1	0.1016	9.84
17	grade9	0.0000	Inf
18	grade10	0.0000	Inf
19	grade11	0.0000	Inf
20	grade12	0.0000	Inf
21	publicSchool1	0.0000	Inf
22	familyEducation1:expectBachelors1	0.1423	7.03
23	familyEducation2:expectBachelors1	0.0549	18.23
24	familyEducation3:expectBachelors1	0.0613	16.30
25	familyEducation4:expectBachelors1	0.0332	30.12
26	grade9:publicSchool1	0.0000	Inf
27	grade10:publicSchool1	0.0000	Inf
28	grade11:publicSchool1	0.0000	Inf
29	grade12:publicSchool1	0.0000	Inf

Step AIC Backward Results

```
lm(formula = readingScore ~ minutesPerWeekEnglish_sqrt + schoolSize_sqrt +
    grade + male + raceeth + expectBachelors + motherBachelors +
    fatherHS + fatherBachelors + englishAtHome + computerForSchoolwork +
    read30MinsADay + publicSchool + parentBachelors, data = df2)


Residual standard error: 74.1 on 3384 degrees of freedom
Multiple R-squared:  0.3131,   Adjusted R-squared:  0.3093
F-statistic:  81.2 on 19 and 3384 DF,  p-value: < 2.2e-16
```

Patrick Keener
SID: 1385832
DSC 423 – Assignment 2

Step AIC Forward Results

```
lm(formula = readingScore ~ expectBachelors + raceeth + read30MinsADay +
    grade + familyEducation + computerForSchoolwork + male +
    schoolSize_sqrt + publicSchool + minutesPerWeekEnglish_sqrt +
    parentHighestEdu + numParentBachelors + englishAtHome, data = df2)


Residual standard error: 74.08 on 3385 degrees of freedom
Multiple R-squared:  0.3133,   Adjusted R-squared:  0.3097
F-statistic: 85.81 on 18 and 3385 DF,  p-value: < 2.2e-16
```