Education Stats Analysis

Stella Li

12/8/2018

Abstract

This report analyzes education data collected by International Organizations, e.g. World Bank and OECD PISA test team. The focus of this analysis is on the effect of education status of parents on academic performances of students in 2012 PISA test. The linear regression and mixed effect models indicates that there is a possitive association between education backgroud of parents and test scores of students, and the association is interacting with the ESCS scores of the students.

Keywords: education, statistics, multilevel regression

1.Introduction

Background and Research Question

Understanding how well the education systems prepare their students and what factors impact the outcomes is essential for students, parents, teachers and governments.[1] This report attempts to explore the second question via analyzing results from international comparative census/survey and assessments. Specifically, I'm interested in the relationship between the education performance of students and the education status of parents, as well as factors such as studying strategy, access to education and social-economic backgrounds of students. The second fold of this study is to explore variances of education outcomes among countries over the world.

Data Source

The data used in this study is from two sources: World Bank EdStats All Indicator Query and OECD PISA database.

I downloaded the World Bank data from their account on Kaggle.com. The dataset "holds over 4,000 internationally comparable indicators that describe education access, progression, completion, literacy, teachers, population, and expenditures."[2] It actually contains data from several datasets maintained by other organizations, such as Barro-Lee Educational Attainment Data and UNESCO Institute for Statistics. I selected indicators in this dataset as predictors in my models.

The "dependent variable" is 2012 PISA test result. Programme for International Student Assessment (PISA) was created in 1997, as a measure of effectiveness of education systems across OECD (Organisation for Economic Co-operation and Development) countries "within a common internationally agreed framework".[1] PISA tests have a strong focus on the preparedness for the future, and include three subjects: reading, mathematics and science.[3] Via student-, teacher- and school- level questionnaires, PISA also collected information on socioeconomic background, studying strategy, attitudes, education access, etc. of students and schools. The student questionnaires data was downloaded from the PISA website.[4] I selected some of the features as my predictors.

2.Methods and Results

2.1 Exploratory Data Analysis

The EdStats data file contains information about 242 countries on 3,665 educational features from 1971 to 2017, as well as projections up to 2050. The 2012 PISA test data contains the full set of responses from 480,174 students from 65 countries and contains 634 variables. The data exploratory process therefore has four steps:
1) Deleted rows and columns with too many n/a, invalid or missing values; and used stratified sampling to get 10% students for each country;

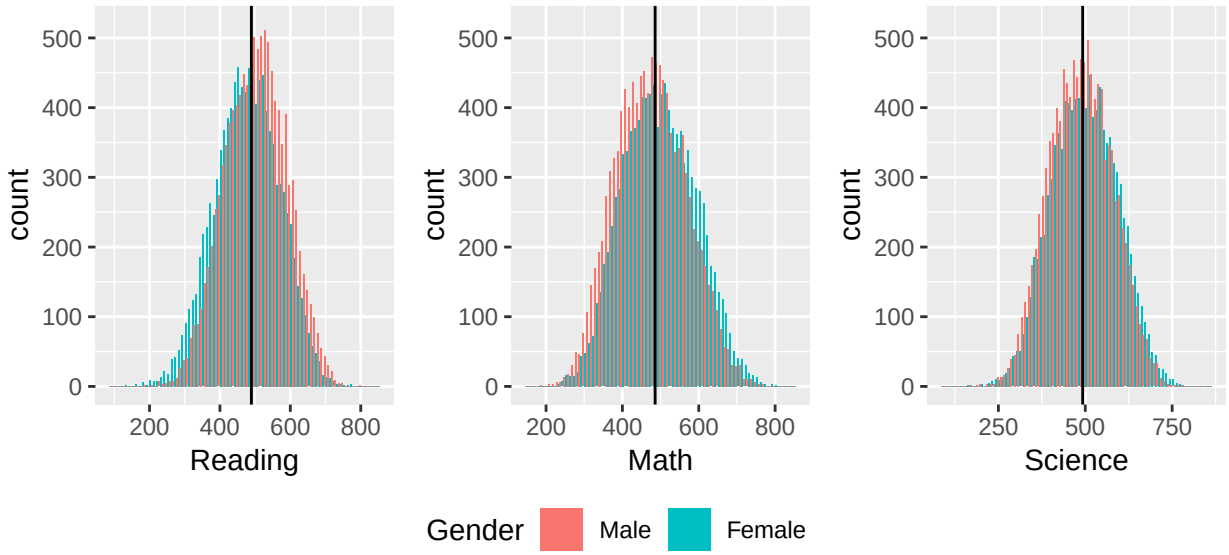


Figure 1: Distribution of PISA Score by Genders

2) Selected variables that are relative to the question of this study;
3) Did exploratory data analysis on variables of interest;
4) Finally, selected variables to include in the models.

2.1.1 PISA scores

First, I checked the distribution of PISA scores among different genders, countries and country income groups.

A. PISA scores vs gender

The Figure 1 shows the distribution of PISA scores of three subjects for two genders.
The black vertical line shows the mean score of all students for each subject respectively. The distributions are almost symmetric about their respective mean values, while there are slightly differences between the two genders in reading and math tests. More female students have higher-than-average reading scores while more male students have higher-than-average math scores.

B. PISA score vs country

I then checked the distribution of PISA test scores among all the countries.
In Figure 2, the colorful lines shows the score distributions for each country and the dashed line shows the score distribution of all students. It is clear that there are many differences in the distributions, therefore, it is necessary to construct a multilevel model to analyze data.

C. PISA scores vs. Income and Region

a) The OECD categorized all the countries into four income groups. So first I checked the distribution of average PISA scores of each country among different income groups.

In Figure 3, the black vertical lines shows the average of all country means.

As it shows in Figure, most of the mean scores of countries in "High income: OECD" group are higher than average and most of the mean scores of countries in "Upper middle income" group. However, being in higher income group does not necessarily associated with having higher scores. It's also interesting that the histogram is bi-modal, meaning that the countries in these groups may have very different features.

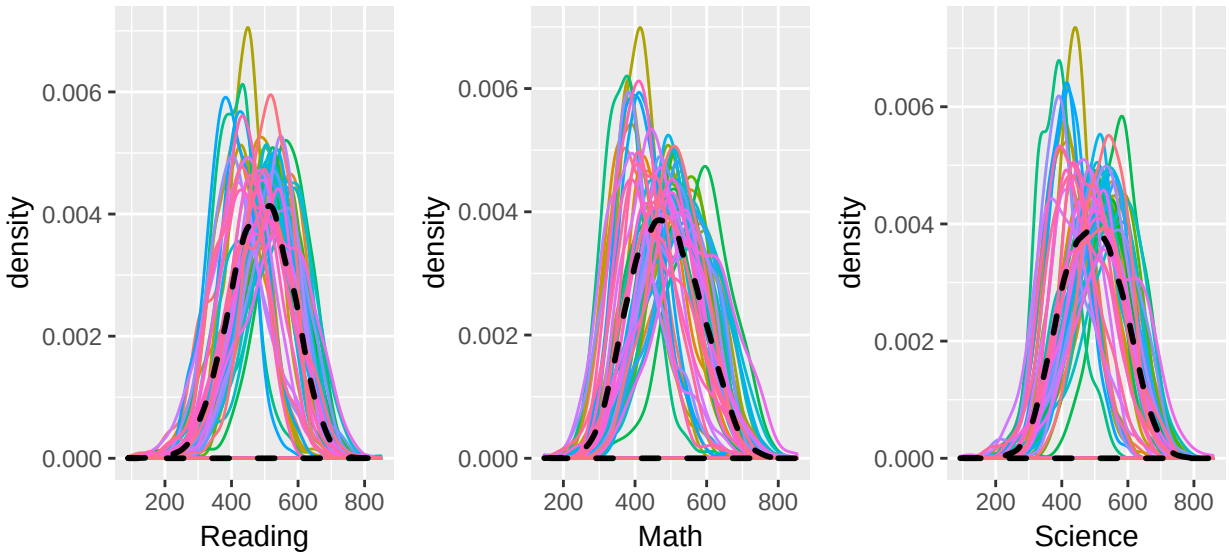


Figure 2: Distribution of PISA Scores by Contries

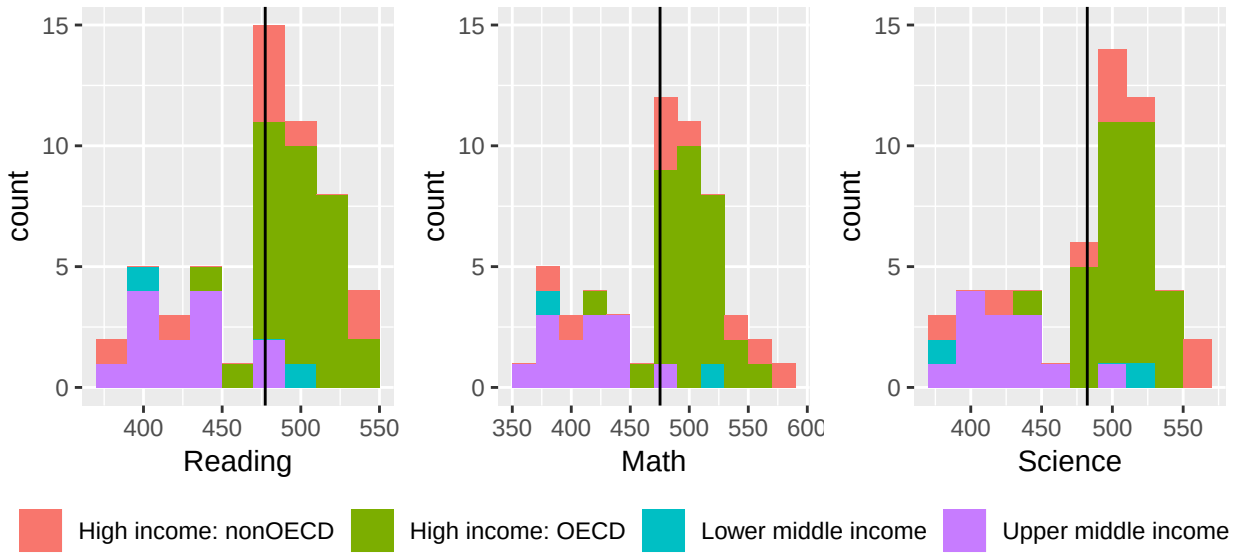
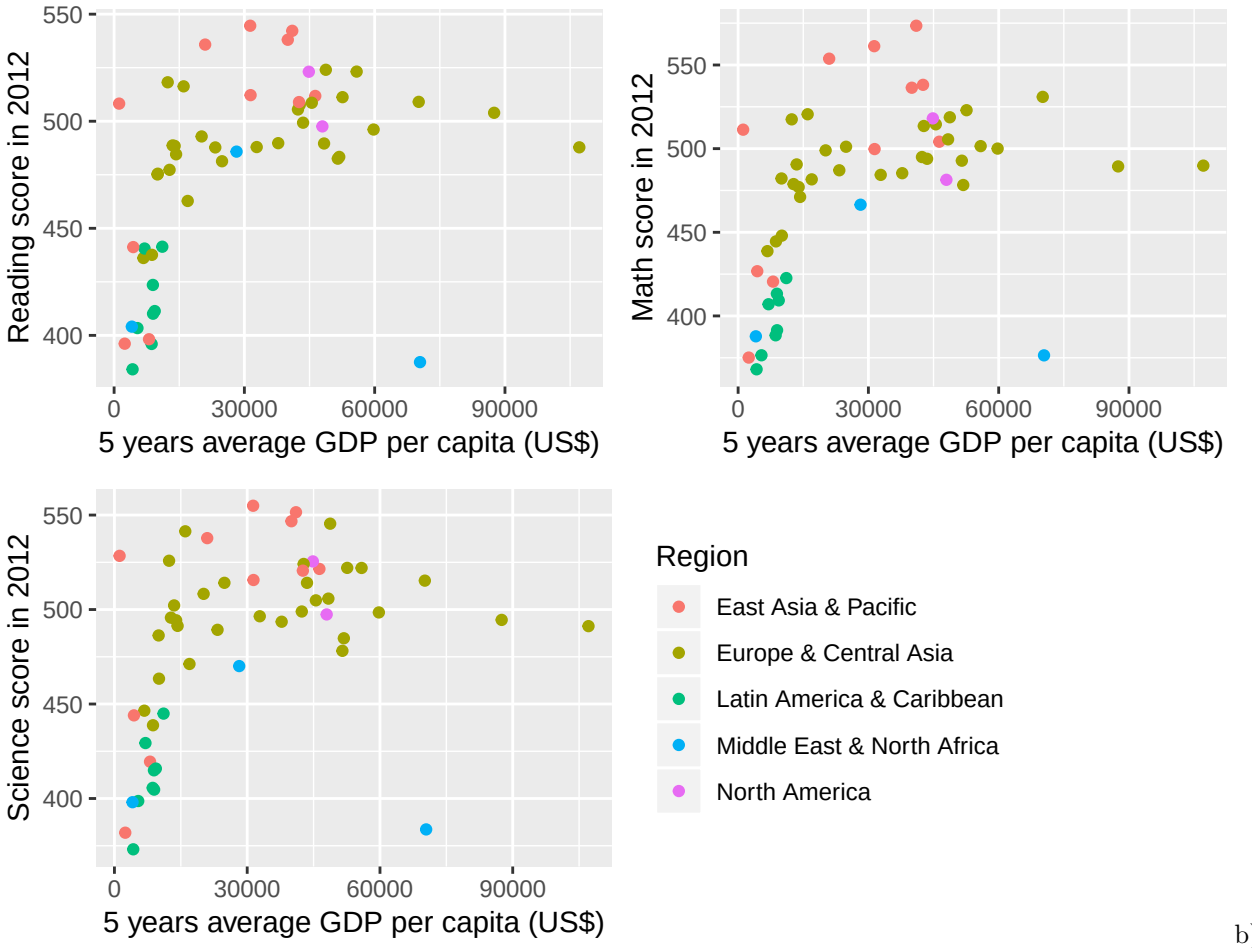


Figure 3: Distribution of Average PISA Score



b) I further explored the relationship between average PISA test scores for each country and the 5-year average GDP per capita.

In Figure **??**, different colors indicating different regions. For countries in the "Europe & Central Asia" region, though the 5-year average GDP per capita change a lot, the range of mean scores of these countries is relatively small than the ranges of other regions.

From the above analyses, it is clear that it is reasonable and necessary to use multilevel models to analyze the data.

D. PISA Scores vs Parent Education Background

As I'm very interested in the association between parent education background and the students performance, I checked the relationships on both student level and country level. On student level, Figure 4 shows that the overall trend of relationship between parent education and students reading scores are similar for different countries, however, the intercepts vary a lot. On the other hand, on country level, countries in different income groups have different trends.

2.1.2. Country Level Predictors

Initially I selected 17 features as country level predictors. I focused on information about two time periods: 2010-2012 and 1995-1997. As I'm very interested in learning about the parents of the students who took the PISA test in 2012, I selected the indicators about 20-29 years old population in 1995-1997 and 35-44 years old population in 2010-2012, assuming that the data can represent the characteristics of parents of those 2012 PISA test takers.

However, there are many missing values in the dataset, as some census/surveys were conducted in some certain countries or in some certain years. Therefore, after further selection, correlation analysis and processing the raw data, I got 54 countries with 5 predictors: region, income group, average 5 years GDP per capita, government expenditure on education as % of GDP (%), and percentage of population age 40-44 completed secondary schooling (in 2010). (Summary of these features is in Appendix Table)

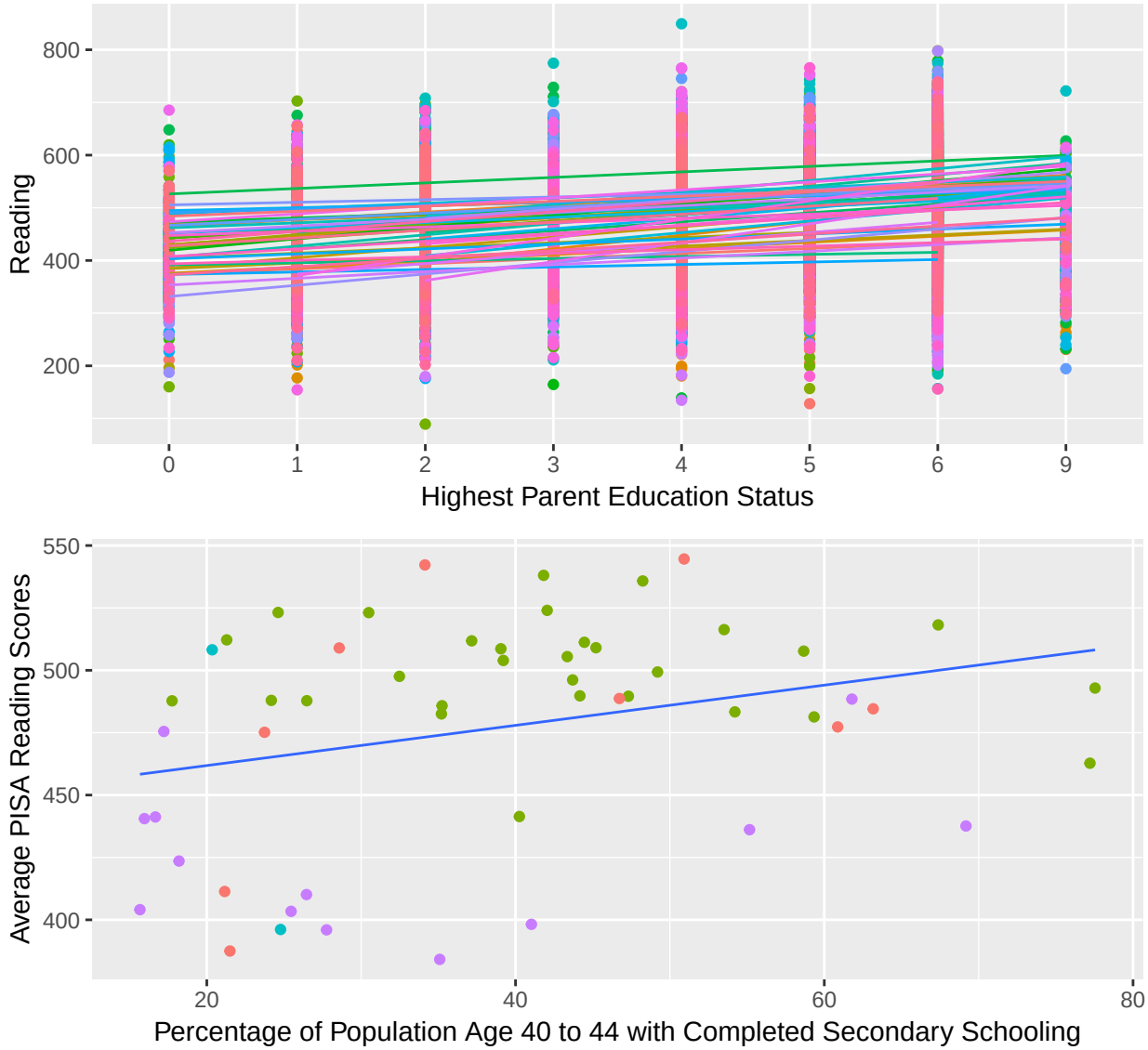


Figure 4: PISA Scores vs Parents Education

- `Country Code`: Three-letter country code for each country; treated as factors.
- `Region`: The region of a country. There are 5 regions.
- `Income Group`: The income group categories defined by OECD. There are four levels.
- `ave5yrGDP`: The average GDP per capita of 2007 to 2011 in current U.S. dollars.
- `ParentSchool6`: Percentage of population age 40-44 with completed secondary schooling in 2010.
- `Expediture2`: The average government expenditures on education as % of GDP (%) in 2010-2012.

2.1.3. Student Level Predictors

I extracted 17 features and 3 subject scores for each student, and then deleted the students with too many n/a, invalid or missing values. Then students were grouped by countries and 10% of students were randomly selected. Then I joined the student data table with country data table, and finally got information on 22,388 students (from 54 countries) with 9 predictors. Those predictors are about gender, education attitude and education access (out of school study time and total learning time for a certain subject), socioeconomic background, and education status of parents.

- `Country Code`: Three-letter country code for each country; treated as factors.
- `male`: Gender, male is coded as 1 and female is coded as 0.
- `ESCS`: Index of economic, social and cultural status.
- `parentEd`: Highest parent education status, from 1 to 9; treated as continuous variables.
- `sameLanguage`: Whether the test language is the same as the language student use at home.
- `OutHours`: Out-of-school study time per week in hours.
- `LTime<subject>`: Hours per week learning for each subject.

I also checked the correlations among the numeric predictors for both country- and student- level predictors (with sampled students) and also learned about the distribution of each predictors. Some of the predictors need to be log transformed to get close-to-symmetric distributions.

2.2 Models and Results

2.2.1. Student level preditors (complete pooling)

In this model, I constructed models for three subjects separately, only taking student level indicators (do not include "country" as a variable). The models can be presented as:

$score_i = \alpha + \beta x_i + \epsilon_i,$
$\epsilon_i \sim N(0,\ \sigma_y^2)$

I noticed that the coefficients of `parentEd` are negative in models for three subjects, which is counter-intuitive. Correlation table shows that there are strong correlation between `parentEd` and `ESCS`. It is reasonable that the socioeconomic status of a student is highly associated with the education background of their parents. Since I'm interesting the impact of parents on students' outcomes, I decide to keep `parentEd` and also check its interaction with `ESCS`.
The coefficients of the variable `OutHours` for science and math are not significant, and the effect sizes are small comparing with those of learning time for each subject, but there is no strong evidence that it should be excluded from the models, so I decide to keep it.

Table 1: Regression Models

	Dependent variable:		
	Reading	Math	Science
	(1)	(2)	(3)
male	-30.87^{***} (1.09)	16.21^{***} (1.14)	5.93^{***} (1.11)
ESCS	42.24^{***} (0.77)	45.28^{***} (0.80)	43.10^{***} (0.78)
parentEd	-7.50^{***} (0.51)	-8.23^{***} (0.53)	-7.89^{***} (0.52)
sameLanguage	14.49^{***} (1.44)	2.25 (1.50)	10.84^{***} (1.46)
log(OutHours + 1)	3.06^{***} (0.69)	1.01 (0.72)	0.31 (0.70)
log(LTimeReading + 1)	8.39^{***} (1.76)		
log(LTimeMath + 1)		21.24^{***} (1.83)	
log(LTimeScience + 1)			24.98^{***} (1.20)
Constant	515.05^{***} (4.20)	487.11^{***} (4.30)	488.81^{***} (3.52)
R^2	0.22	0.21	0.21
Adjusted R^2	0.22	0.21	0.21
Residual Std. Error (df = 22381)	81.16	84.49	82.27
F Statistic (df = 6; 22381)	$1{,}027.32^{***}$	966.16^{***}	988.74^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 2: Regression Model Comparison

	Dependent variable:					
	Reading		Math		Science	
	(1)	(2)	(3)	(4)	(5)	(6)
ESCS	42.24^{***}		45.28^{***}		43.10^{***}	
	(0.77)		(0.80)		(0.78)	
parentEd	-7.50^{***}	5.15^{***}	-8.23^{***}	5.48^{***}	-7.89^{***}	5.16^{***}
	(0.51)	(0.39)	(0.53)	(0.40)	(0.52)	(0.39)
ESCS:parentEd		6.75^{***}		7.13^{***}		6.80^{***}
		(0.15)		(0.16)		(0.15)
log(OutHours + 1)	3.06^{***}	2.73^{***}	1.01	0.70	0.31	0.04
	(0.69)	(0.70)	(0.72)	(0.73)	(0.70)	(0.71)
log(LTimeReading + 1)	8.39^{***}	11.09^{***}				
	(1.76)	(1.80)				
log(LTimeMath + 1)			21.24^{***}	21.26^{***}		
			(1.83)	(1.88)		
log(LTimeScience + 1)						
R^2	0.22	0.18	0.21	0.17	0.21	0.17
Adjusted R^2	0.22	0.18	0.21	0.17	0.21	0.17
Residual Std. Error (df = 22381)	81.16	82.78	84.49	86.42	82.27	84.06
F Statistic (df = 6; 22381)	$1{,}027.32^{***}$	842.67^{***}	966.16^{***}	758.58^{***}	988.74^{***}	790.25^{***}
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 2 shows the changes in the coefficients of `parentEd`, `ESCS`, `ESCS:parentEd`, and some regression summary information. When I replace `ESCS` with the interaction term, the F Statistics drop significantly. However, among all models, the R^2 statistics are around 0.2, not changing too much. So there are more factors that can explain the variance among performance of students. Take a look at models estimating Math scores as an example. In the left figure in Figure 5, the pink dots represent the revised model while the black dots represent the original model. The two models are very similar to each other, and are both not good at estimating actual score. The range of fitted values from both models are smaller than the real data.

I also checked residual plots for all models. The middle and the right figures in Figure 5 are residual plot shows

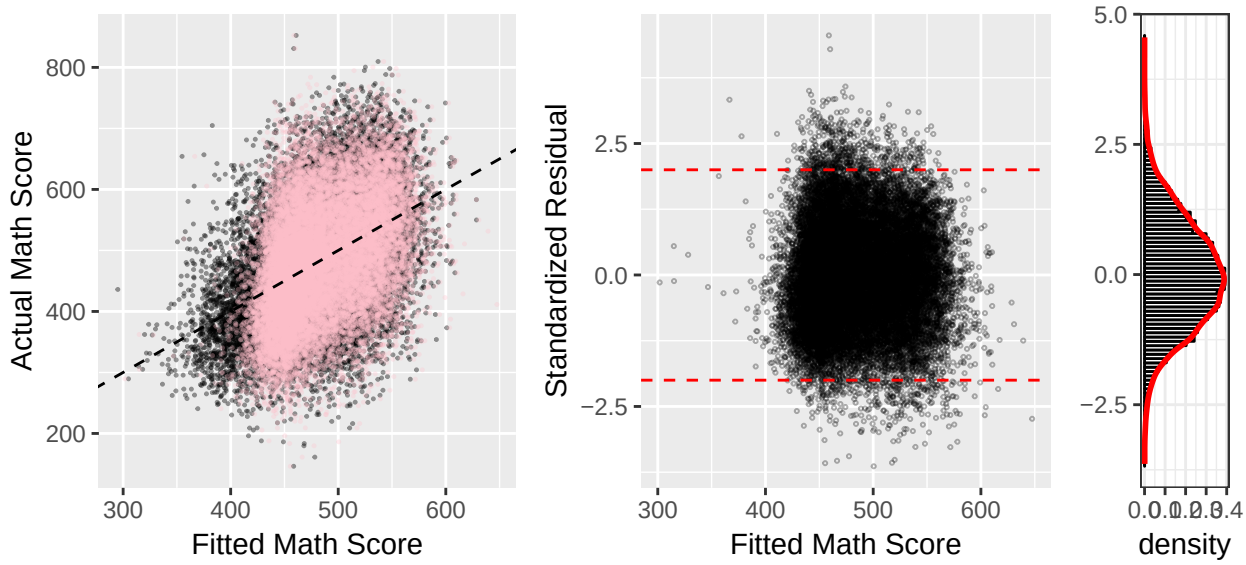


Figure 5: Residual Plot

the residual plot of the revised student level model for math. The residuals are almost symmetrically distributed and there is no trend. However, there are many data points outside of the $\pm 2\sigma$ range. The other residual plots follow a similar pattern to this one.

2.2.2. Add country level predictors (no pooling)

Based on the findings from exploratory data analysis, I add country level predictor to the model. First starting with country name as a variable and force it without an intercept. Then the models can be represented as:
$score_i = \alpha_{j[i]} + \beta x_i + \epsilon_i$,
$\epsilon_i \sim N(0,\ \sigma_y^2)$

Table 3: Regression Models with 'Country'

	Dependent variable:		
	Reading	Math	Science
	(1)	(2)	(3)
male	−30.83*** (1.03)	15.32*** (1.03)	5.55*** (1.01)
parentEd	3.65*** (0.37)	3.97*** (0.37)	3.43*** (0.37)
sameLanguage	26.21*** (1.60)	18.83*** (1.61)	24.51*** (1.57)
log(OutHours + 1)	6.63*** (0.69)	4.66*** (0.69)	4.12*** (0.68)
log(LTimeReading + 1)	1.96 (1.94)		
log(LTimeMath + 1)		25.93*** (1.88)	
log(LTimeScience + 1)			31.67*** (1.20)
ESCS:parentEd	5.72*** (0.15)	5.89*** (0.15)	5.49*** (0.14)
R^2	0.98	0.98	0.98
Adjusted R^2	0.98	0.98	0.98
Residual Std. Error (df = 22328)	75.78	76.17	74.51
F Statistic (df = 60; 22328)	15,734.05***	15,375.62***	16,510.08***
Note:			*p<0.1; **p<0.05; ***p<0.01

In Table 3, I omitted the printing of coefficients for all the countries as the list is too long. All the p-values of the coefficients for `Country Code` are less than 0.05. Though the signs of other predictors do not change comparing

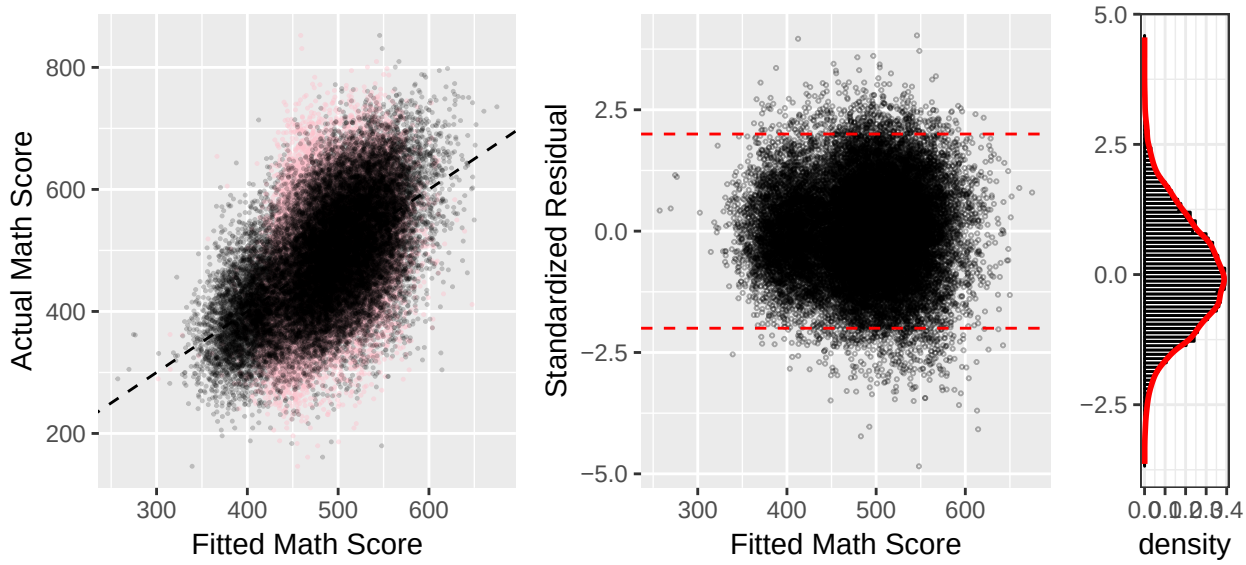


Figure 6: Residual Plot

with previous models, the scale of impact changed a lot. In these models for three subjects, the coefficients of `log(OutHours)` now are statistically significant.

The R^2 of these three models are significant larger than the previous models. It doesn't mean the models are "nearly perfect". In these models, I force the intercept to be 0. In other words, I set the "expected value" of the outcome to be 0. Since the equation of R^2 is

$$R_0^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

by setting intercepts to be 0, the new R^2 is now

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i)^2}$$

Though SSE and SST will both increase for new R^2, but SST tends to change more. Therefore, $\frac{SSE}{SST}$ decreases, and the R^2 increases a lot. We still need to look at fitted data from the new model.

The left figure in Figure 6 shows the comparison between the complete pooling model with and the new model. The pink points represent the previous model and the black points represent the new model. It looks like the new model improves estimation.

Taking a look at residual plots in Figure 6. The plots also look better than before: the data points that outside of the $\pm 2\sigma$ range are fewer.

2.2.3. Multi-level models (partial pooling)

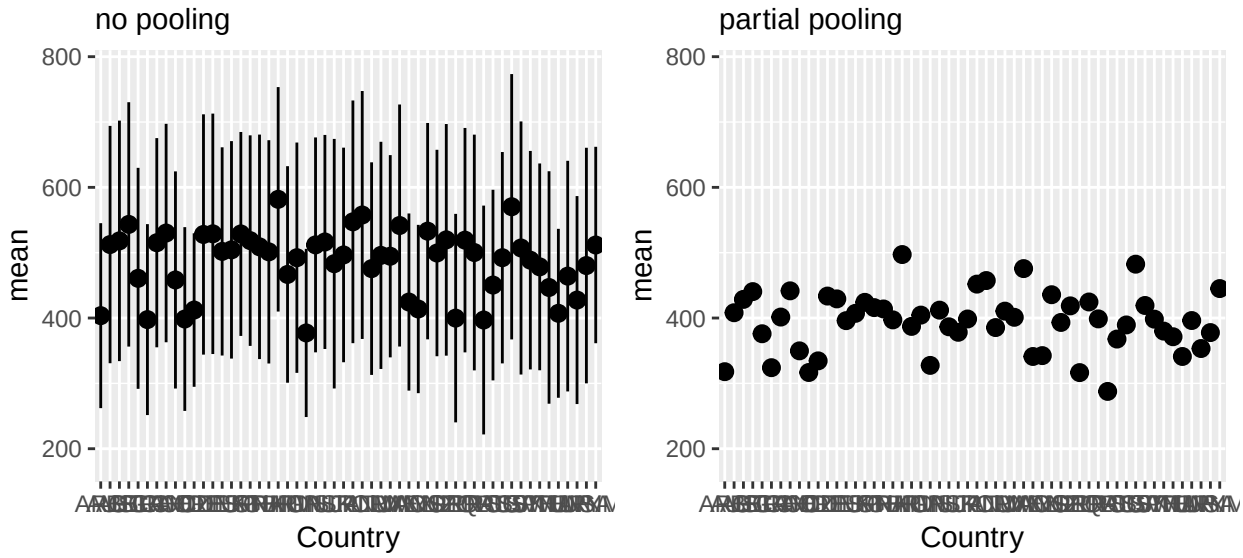
I fit several models in this part, and use Math score as an example.
A. Partial pooling, varying intercept according to `Country Code`.

$score_i = \alpha_{j[i]} + \beta x_i + \epsilon_i$,
$\epsilon_i \sim N(0,\ \sigma_y^2)$,
$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$

The below Table **??** shows the comparison between coefficients of the variables (except for those of `Country Code`) in the last model and fixed effects of varying intercept partial pooling model. The two models are very similar.

Table 4

	Dependent variable:	
	Math	
	Partial Pooling	No Pooling
	(1)	(2)
male	15.33*** (1.03)	15.32*** (1.03)
parentEd	3.97*** (0.37)	3.97*** (0.37)
sameLanguage	18.77*** (1.61)	18.83*** (1.61)
log(OutHours + 1)	4.63*** (0.69)	4.66*** (0.69)
log(LTimeMath + 1)	25.90*** (1.88)	25.93*** (1.88)
ESCS:parentEd	5.90*** (0.15)	5.89*** (0.15)
Constant		
Adjusted R^2		0.98
Akaike Inf. Crit.	257,802.70	
Residual Std. Error		76.17 (df = 22328)
F Statistic		15,375.62*** (df = 60; 22328)
Note:		*p<0.1; **p<0.05; ***p<0.01



The effect of varying intercept is to shrink the variance, as showed in Figure **??**

B. Add a country level predictor: 5-year average GDP

The Table **??** shows the result of three models for math scores.
* only introducing `log(ave5yrGDP)` variable to the varying intercept model
* introducing `log(ave5yrGDP)` variable and the interaction term between `Region` and `log(ave5yrGDP)`
* introducing `log(ave5yrGDP)`, and allow the slope of `log(ave5yrGDP)` to vary by `Region`

The random effect of the varying slope is:

```
##                          (Intercept) log(ave5yrGDP)
## East AsiaPacific          -84.000130      14.019239
## EuropeCentral Asia        -37.678139       6.288310
## Latin AmericaCaribbean     68.228133     -11.386964
## Middle EastNorth Africa    60.849519     -10.155507
## North America              -7.399384       1.234923
```

Table 5

	Dependent variable:		
	Math		
	(1)	(2)	(3)
male	15.35***	15.34***	15.34***
	(1.03)	(1.03)	(1.03)
parentEd	3.96***	3.96***	3.95***
	(0.37)	(0.37)	(0.37)
sameLanguage	18.83***	18.92***	18.95***
	(1.60)	(1.60)	(1.60)
log(OutHours + 1)	4.68***	4.66***	4.67***
	(0.69)	(0.69)	(0.69)
log(LTimeMath + 1)	25.87***	26.07***	26.03***
	(1.88)	(1.88)	(1.88)
log(ave5yrGDP)	20.22***	17.19***	7.77
	(5.69)	(4.86)	(6.97)
ESCS:parentEd	5.89***	5.89***	5.89***
	(0.15)	(0.15)	(0.15)
log(ave5yrGDP):RegionEuropeCentral Asia		−3.08***	
		(1.08)	
log(ave5yrGDP):RegionLatin AmericaCaribbean		−9.41***	
		(1.53)	
log(ave5yrGDP):RegionMiddle EastNorth Africa		−10.11***	
		(1.93)	
log(ave5yrGDP):RegionNorth America		−5.29**	
		(2.16)	
Constant	194.08***	261.38***	299.04***
	(56.87)	(46.77)	(55.59)
Akaike Inf. Crit.	257,787.90	257,745.80	257,767.10

Note: *p<0.1; **p<0.05; ***p<0.01

3.Discussion

3.1.Indications:

1. The model I fit all indicates that there are positive association between parent education status and students' performance on PISA test, when controlling for other variables. It also interact with the ESCS status of students. In our first groups of model, controlling for other variables, for students who all have 1 as ESCS score, those whose parents have 1 higher level of education status are expected to have around 12 more points on average, in the three PISA tests. While for students who all have 2 as ESCS score, those whose parents have 1 higher level of education status are expected to have around 18 more points on average.
2. On country level, in different regions, the relationships between national income and students' PISA scores are different. For example, in varying slope model, when controlling for student level variables, the associations between math scores and national income are positive. However, the association is stronger for students from countries in Europe & Central Asia than students from countries in Middle East & North Africa.
3. Education acheivement is a complicated issue. In this study, I just tested several elements that I am interested in. Although the data supports that there are certain associations between the predictors and the PISA test scores, there is no easy conclusion about causal relationship or what strategy will definite be effective to improve education acheivment. Though I have been being interested in PISA test for a long time, this is my first attempt to analyze the data myself, and it has been a great learning opportunity for me to understand the complexity of international level school-based test.

3.2.Limitations:

1. In first step of collecting data and cleaning data, I was first surprised by how international organizations are open to share data. Then quickly, I got frustrated by the quality of data. There are many missing or invalid data in both database. Therefore I have to make a lot of assumptions when cleaning and processing data. For example, when cleaning data for country level indicators, I use average over five years time span instead of a certain year . Because no year that all country have data entry, I have to assume that the three-year-average can be a good approximate of the real data. I'm very curious about how people leverage these two database, especially how PISA team work on the data to generate thorough analysis.
2. The PISA Student Level data were collected via questionnaires. Therefore, the self-reported data, e.g. learning time, parents education status can be different than real status.

Appendix

References

[1] OECD (2013), PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264190511-en.
[2] Cresswell, J., U. Schwantner and C. Waters (2015), A Review of International Large-Scale Assessments in Education: Assessing Component Skills and Collecting Contextual Data, PISA, The World Bank, Washington, D.C./OECD Publishing, Paris, https://doi.org/10.1787/9789264248373-en.
[3] Education Statistics From World Bank Open Data https://www.kaggle.com/theworldbank/education-statistics/home
[4] Student Questionnaire data file http://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm

Tables and Figures

1. Summary of country level predictors
2. Summary of student level predictors

Table 6: Country Level Predictors

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Reading	54	477.4	44.3	384.2	441.3	508.9	544.6
Math	54	475.2	51.9	368.1	440.2	509.9	573.5
Science	54	482.3	48.9	373.1	445.3	519.3	554.9
ave5yrGDP	54	29,827.6	23,516.3	1,162.4	9,515.7	45,351.8	107,119.1
ParentSchool6	54	39.3	16.5	15.7	25.0	49.0	77.5
Expenditure2	54	5.0	1.2	2.8	4.2	5.6	8.2

Table 7: Student Level Predictors

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
male	22,388	0.5	0.5	0	0	1	1
ESCS	22,388	−0.2	1.1	−4.7	−0.9	0.7	2.8
LTimeReading	22,388	3.7	1.6	0	2.9	4.2	30
LTimeMath	22,388	3.8	1.6	0.0	3.0	4.2	25.5
LTimeScience	22,388	3.5	2.1	0	2	4.5	30
Math	22,388	485.6	94.8	146.2	415.6	553.0	852.4
Reading	22,388	489.4	91.6	89.3	426.0	554.0	849.4
Science	22,388	492.6	92.5	92.4	425.6	558.5	857.8