EARLY CHILDHOOD LONGITUDINAL STUDY

Aniket Shah

*International School of Engineering*

March 20th, 2016

Purpose

The data set for this is a collection of 7 surveys of the 1998-1999 Kindergarten over several years to analyze the short-term and long-term impact of environmental stressors on cognitive development and self-efficacy. The purpose is to observe how students with certain attributes perform in the U.S schooling system; identify key variables that might be affecting their cognitive abilities, and self – efficacy.

The U.S government spent $154 billion dollars on education in the fiscal year of 2015, which is the third highest dollar-per-pupil investment in the world. Considering this, one would expect that the performance of U.S students would be at par with the best in international assessments. However, that’s not the case, they rank 35th in terms of performance. Therefore, it is important for the administrators to understand the environmental stressors causing the dip in performance. Only with proper analysis, can they spend the funds efficiently to bridge this gap.

In this study, cognitive abilities are being measured by standard tests in Math, Reading, and Science. Self – efficacy variables are also being observed, such as how angry is the child when having trouble learning, and how much do they enjoy/like Math and Reading. Out of the 7 surveys, two were done in kindergarten, two in Grade 1, and the rest were done through to fifth grade. A data-scientific analysis done to determine significant attributes that impact the reading, math, & science scores across the waves for children in poverty and children not in poverty.

State-of-the-art

The Early Childhood Longitudinal Survey (ECLS) is one of the first nationally representative studies of early childhood development and educational experiences. It provides detailed information on children's health, early care and early school experiences.1 Understanding cognitive abilities, and emotional actions of children is a difficult task. Some attributes such as test scores are objective, but questions that the parents, and teachers answer are subjective. Tests designed to asses student abilities in Math, Reading, and Science are also questionable as there is no perfect recipe for testing all students; how they understand the topic and express their abilities can vary widely.

The ECLS-K tries to gain insight into how student environment affects their experience in school. It was designed to address a vast array of research issues like: 1) schooling and performance, 2) status and transitions, and 3) the interaction of school, family, and community.

For school performance, ECLSK analyzes the interaction between children's backgrounds and their performance in different learning settings. Data was collected on curriculum, instructional practices, resources, school climate, and background characteristics of teachers and administrators in order to examine school performance over time.1

Children learn at varying pace, and they have differing levels of preparation as they enter kindergarten. ECLS-K study addresses this by measuring children's skills and knowledge at several intervals from kindergarten through eighth grade. ECLS-K also provides critical information on the roles that parents and families play in supporting their children's education. ECLS-K also focused on the resources of the family, the home environment, and the community, which can have a profound impact on children's success in school. This data was used in studies such as, *Big Math for Little Kids: The Effectiveness of a Preschool and Kindergarten Mathematics Curriculum*, and *A Multilevel Analysis of Asian Immigrant Children's Reading Achievement in the Early Years*.

Method

To decide on an overall method for analysis, one needs to understand the properties of the data. ELCS-K involves a combination of categorical, and numerical attributes. Machine learning techniques that purely deal with numerical attributes can only be used if categorical attributes are converted into dummy variables, but if one does go that route, distance measure will need to be well thought out. To gain insight into this data, descriptive analytics is the focus, not predictive. Is the dataset dense or sparse? Well, initially dataset contains 529 attributes, and 3,143 which is sparse, but when the dataset is broken down into the respective 7 waves, the datasets become dense as row to column ratio increases. This means regression techniques such as linear regression (StepAIC) can give reliable results. However, high number of attributes, especially categorical, means regression output will have low explicability, and hence, to improve ease of understanding, feature selection techniques must be used.

Feature selection can be done using PCA, SVD or Random Forest. For SVD, categorical variables need to converted into dummy variables, making the data sparse, and the technique unreliable. PCA can be used, but it will return components which are combinations of the original attributes. There will be no explicability left if these components are used for regression. Random Forest is ideal as it can handle a mixture of attributes, automatically bins the numerical attributes. It ranks the attributes based on the Gini Index. A high number means the attribute provides great information gain, meaning it helps a great deal in explaining the target variable. Random Forest also increases the probability of attaining the best global solution. It chooses attributes at random, orders them with respect to information gain, and runs a decision tree over them; it does this multiple times to provide the best estimate. After the attributes have been narrowed down, analysis will depend on the type of the target variable. For cognitive ability attributes (numerical), various regression techniques such as linear, and lasso can be used. For self-efficacy (categorical), a decision tree can be used to identify rules. The resulting outcomes are visualized to attain greater understanding of the underlying data.

Data Pre-Processing

The first step is to clean the dataset, and make it suitable for machine learning models. The following processing steps were done over the entire dataset (all 7 waves):

* Import the data into R, and omit attributes that have type “WEIGHT”, or have values “SUPPRESSED”
* Attributes that have greater than 30% NA’s are removed
* Split the data into smaller datasets, specific to waves 1 to 7
  + The attributes that start with
    - WK belongs to - Waves 1&2
    - W1 : Waves 3&4
    - W3 : Wave 5
    - W5: Wave 6
    - W7 : Wave7
  + After these were sorted, the remaining attributes were assigned according to the rule: second position of the attribute name corresponds to the wave number. Sorting was done with the grep function

Once the waves were separated into their own datasets, further pre-processing was required before Random Forest could be applied. These steps were applied to all the waves:

* Changed "NOT APPLICABLE","REFUSED","DON'T KNOW", and "NOT ASCERTAINED“ to NA
* Given certain assumptions, substituted -1 with zero for some numerical attributes like P1EARLY, P1ADLTL2 etc.
* Numerical attributes that had a positive range, negative numbers represented instances that added no additional information about the data; replaced them with NA
* Removed columns where more than 30% of the values are NA, and rows where more than 20% of the values are NA
* If two attributes had a correlation value > 0.85, I removed one of the attributes. For example, P1HSDAYS and P1HSHRS were highly correlated, removed P1HSDAYS
* Performed imputation on the remaining data (knnImputation), standardized numerical attributes, and split the wave into two; students above poverty, and students below poverty.

As stated above, for some numerical attributes, negative values were changed to NA barring some exceptions. This is explained in greater detail below.

Many numeric variables have negative values that are not part of the continuous range, instead they symbolize something about the survey data. -7 (REFUSED), -8(DON’T KNOW), -9 (NOT ASCERTAINED). These were replaced with NA for a couple of reasons, one they add nothing to the data, and two, data would be distorted during imputation. After replacing with NA, I can remove the attributes that have more than 30% NA values. -1 represented NOT APPLICABLE, but the attribute information did not provide any reasons as to why they were not applicable. Hence, for attributes that had -1 occurring with high frequency, certain assumptions had to be made. For example, in Wave 1:

P1HSDAYS

Description: How many days in the week did child go to HEAD START program. Values ranged from 1 to 7, but many were -1 (NOT APPLICABLE)

Assumption: Not Applicable represented children that did not attend the program at all /were not a part of it. Replaced -1 with 0.

P1HSHRS

Description: How many hours in the week did child go to HEAD START program. Values ranged from 1 to 60, but many were -1 (NOT APPLICABLE)

Assumption: Not Applicable represented children that did not attend the program at all /were not a part of it. Replaced -1 with 0.

P1HAGEYR

Description: How old was child in years when he/she joined HEAD START program. Values ranged from 2 to 5, but many were -1 (NOT APPLICABLE)

Assumption: Not Applicable represented children that did not attend the program at all /were not a part of it. Replaced -1 with NA as 0 or any other value does not make sense with respect to this attribute

P1HAGEMO

Description: How old was child in months when he/she joined HEAD START program. Values ranged from 0 to 11, but many were -1 (NOT APPLICABLE)

Assumption: Not Applicable represented children that did not attend the program at all /were not a part of it. Replaced -1 with NA as 0 or any other value doesn’t make sense in the context.

P1EARLY

Description: How many days or weeks was the child born early? Values ranged from 1 to 31, but many were -1 (NOT APPLICABLE)

Assumption: Not Applicable represented children that were not premature. Replaced -1 with NA as 0 in 0 days early.

P1ADLTL2

Description: {Besides {CHILD}'s biological {mother/father/parents}, how/How} many adults, 18 years or older at the time, once lived with {CHILD} for at least four months, but no longer do? Values range between 1 and 20. -1 is frequent

Assumption: Not Applicable (-1) represented children that had only parents stay with them for a long duration. Replaced -1 with 0 as in no other adults other than the parents.

Data Analysis

Once the data was ready for modelling, a t-test was applied on the various cognitive attributes to determine if the scores were significantly different between children who were above the poverty line vs. ones who were below poverty line. A sample output for Wave 1 Math scores is shown in Figure 1. As can be seen, they are significantly different with a 95% confidence interval, Math scores of above poverty children are on average 7 points better.

data: eclsw1\_APvrty$C1R4MSCL and eclsw1\_BPvrty$C1R4MSCL

t = 20.895, df = 1033.1, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

6.654390 8.033772

sample estimates:

mean of x mean of y

28.68214 21.33806

**Figure 1**

Similarly, the scores were significantly different for all waves. A Normal Q-Q plot was also observed before performing the t-test to ensure the assumption of normal distribution is correct.

Next step was applying the Random Forest model for all the target attributes, and selecting the top 20 attributes for each mini dataset. The attributes that were not part of the top 20 were dropped before applying linear regression, and StepAIC. A sample code for identifying the top 20 attributes is shown in Figure 2.

library(randomForest)

set.seed(1234)

fit.forest\_eclsw1 <- randomForest(C1R4MSCL ~ . , data= eclsw1A\_MSCL,

na.action=na.roughfix,

importance=TRUE, ntree = 150)

# following function determines variable importance

importance(fit.forest\_eclsw1, type=2)

#Arrange in decreasing order

order\_rf <- sort(importance(fit.forest\_eclsw1, type=2)[,1], decreasing = TRUE)

# Top 20 Attributes

feature\_stringw1 <- paste(names(order\_rf)[1:20])

#Data with top 20 attributes that will be used further for modeling

eclsw1M\_impAttri <- eclsw1A\_MSCL[, feature\_stringw1]

#Adding back the target attribute

eclsw1A\_MSCL\_RF <- cbind(eclsw1M\_impAttri, C1R4MSCL = eclsw1A\_MSCL$C1R4MSCL)

**Figure 2 – Wave 1 Math Scores Random Forest (Above Poverty)**

The selected attributes, along with the target variable, are then fed into stepAIC using a linear regression model. The direction for stepAIC is “both”. Sample code shown in Figure 3.

**Figure 3 – Wave 1 Math Scores stepAIC (Above Poverty)**

#Performing stepAic on the top attributes selected through Random Forest

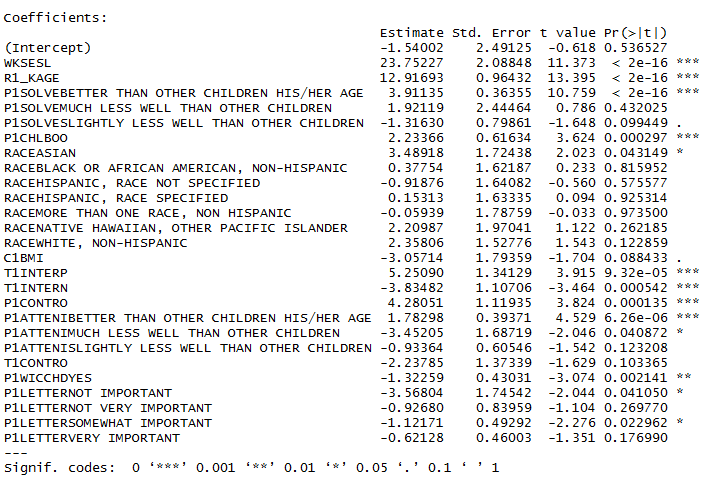
library(MASS)

stepw1\_MSCL <- stepAIC(lm(C1R4MSCL ~ ., data=eclsw1A\_MSCL\_RF), direction="both")

summary(stepw1\_MSCL)

The output of this code was a list of attributes, with their estimated regression coefficients, and significance. This analysis was done across all the 14 waves to determine short-term, and long term environmental stressors that affect cognitive abilities.

**Figure 4 – Wave 1 stepAIC output (Math-Above Poverty)**



Outputs, similar to the one shown in Figure 4, were compared across all the waves to determine common long-term attributes that affected cognitive abilities. The two numerical attributes were Age, and Socio Economic Status. Age always had a positive regression coefficient, meaning older kids in a particular class did better at Math, Reading, and Science tests than their peers. Regression Coefficients for Socio Economic Status (SES) were all positive as well, meaning better economic status of the family meant their kids would do well in school. These were true for both above, and below poverty.

A long term categorical attribute that appeared across all waves was Race. Race has 8 levels, and when linear regression was done, it gave coefficients for 7 levels, and considered *Race:Native Americans* as the base category. Therefore, the coefficient estimates can be seen as the change in cognitive abilities between Native American children and the corresponding race. A positive regression coefficient means the children of that particular race, perform better on average than Native American students.

**Figure 5 – Wave 1 Race (Math-Above Poverty)**

Figure 5 shows how all races perform in comparison to Native American students on Math tests, considering they are above poverty. In the short term (Waves 1 & 2), there is not much difference, but in the long run, all races tend to outperform Native American children. This positive difference in scores is significant when it comes to Asian, and White students. The trend is similar for Reading scores (Figure 6).

**Figure 6 – Wave 1 Race (Reading-Above Poverty)**

What about for below poverty students, does race matter then? As Figure 7 demonstrates, when children come from families who are poor, race does not seem to matter! Yes, the coefficients do differ, and tend to increase for later waves, but they are insignificant in a 95% confidence interval; Race did not even make into the stepAIC model for wave 7. This is true for both Math and Reading scores.

**Figure 7 – Wave 1 Race (Math-Below Poverty)**

Few of the attributes were common among a few waves (2-4), but not in all, these were considered short term stressors.

* BMI – Higher BMI resulted in lower scores, but not significantly
* Internalizing Problem Behaviours – Scores reduced as this attribute increased
* Self-Control – Better self-control meant higher scores
* Parental Education – Children of parents with better education (Masters, Doctorate) performed better on tests

Self-Efficacy Attributes

To analyze categorical self-efficacy characteristics, a decision-tree algorithm was used. For the one numerical attribute SADLON (How sad/lonely the students felt?) measured in Waves 1,2 and 4, stepAIC with a linear model is used. Students that were more social, and had more self-control were less likely to be sad, whereas students with high impulsive behavior were more likely to be sad (Figure 8).

**Figure 8 – Sad/Lonely (Above & Below Poverty)**

The categorical attributes were only measured in Wave 7, output of a decision tree algorithm is used to assess what environmental stressors might be impacting them. Figure 9 shows percentage of attribute usage in the rules derived by the algorithm. Higher the usage percentage, better the target attribute is defined by it. After analyzing specific rules and their lifts, it is observed that how angry the student felt when having trouble learning depended commonly on how worried they were about doing well (C7WRYWEL), how ashamed they felt on making mistakes (C7ASHAME), and the education of their parents.

Attribute usage:

94.47% C7WRYWEL

67.51% S7FLCH\_I

56.49% C7ASHAME

41.68% W8PARED

**Figure 9 – Attribute % for C7ANGRY (Above Poverty)**

Similar analysis was done for enjoy/like Math, and Reading. Whether the students enjoyed Math or Reading depended highly on their competence of the subject. For reading, race also mattered, White children were more likely to love reading than Black or Native American students.

Conclusion

ECLS-K was an ambitious project in data collection, it required the co-operation of children, teachers, and parents over several years! It was done though to aid several research questions such as what environmental stressors affect the cognitive abilities of children as they pass through the American school system. From the analysis performed in this report, it is clear that economic status has a big impact on test scores, and for children above poverty, Native American children perform poorly in comparison to other races. For children below poverty, Race differences are not as pronounced because the negative impact of their economic status overwhelms the influence of Race. Further exploration and analysis can be done using a different technique such as Lasso Regression (done in code, but has been commented out due to lack of visual comparison), and correlations between categorical variables can be looked at. Based on the observations, the U.S government needs to efficiently invest on Native American Students, and on schools in poor neighborhoods in order to improve student performance.

Appendix A –Code

Code is given in R files attached with this report. ECLS.R needs to be run before any specific Wave R file.