The data we analyzed is from the Early Childhood Longitudinal Survey conducted in 1998 by the Institute of Education Sciences. Our focus is on children in kindergarten and followed through the 8th grade. In this subset, there are 3600 participants, among which 1800 are male and 1800 are female, with their gender, race, highest education attainment by the father, math score range 0 to 300 for 7 periods from baseline age 4.5 to 6.6, and corresponding age are recorded. All analysis is conducted by using R 3.6.3.

We first add id for each participant and then convert their age from months to years. The summary statistics for gender, race, highest education attainment by the father, and math score are shown in **Table 1**, and here we see math scores in 7 subgroups according to the period. From the table, we can find most participants ‘father is with Associate’s or bachelor’s degree highest educational attainment, then with a high school diploma or GED, then Graduate or professional degree, and the least is doing not complete high school. Also, most children are White, non-Hispanic, then Black or African American, non-Hispanic, and then is Hispanic and the least is Asian. And there is no missing in education status and race. As for the math score variable, each subgroup has some missing except for period 1, resulting in a total of 5677 missing values. Since the father’s education may influence a child’s score, we will focus on the education status of the father. We study the missing pattern of math score variables in terms of gender and the highest educational attainment of the father. After studying the average number of missing in math score for each participant in different gender\*dad\_edu  groups, we find that the patterns of missing value are similar in both gender, with participants having Associate’s or Bachelor’s degree education attainments have most missing, then those with high school diploma or GED highest education attainments, followed by those with Graduate or professional degree, and finally those with did not complete high school. This pattern is reasonable since this is a similar sample size trend for each education attainments, the more the sample the more the missing data. Since there are many missing data in math score, we first delete all the t3 score because it has missing data more than 70%, then we delete row that has missing in any period math score. Then we change the long-form to wide form, combine all the period math score to one variable ’mathscore’.

| **Table 1. summary statistics for gender, race, age, highest education attainment and math score** | | | |
| --- | --- | --- | --- |
|  | **level** | **statistics** | **Overall (n=3600)** |
| **Gender** | male | N (% column) | 1800(50.0%) |
|  | female | N (% column) | 1800 (50.0%) |
| **Race** | ﻿White, non-Hispanic | N (% column) | 3027 (84.1%) |
|  | ﻿Black or African American, non-Hispanic | N (% column) | 238 (6.6%) |
|  | ﻿Hispanic | N (% column) | 170 (4.7%) |
|  | ﻿Asian | N (% column) | 165 (4.6%) |
| **Dad\_edu** | College or AD | N (% column) | 1723 (47.9%) |
|  | ﻿Graduate of Prof | N (% column) | ﻿570 (15.8%) |
|  | ﻿HS or GED | N (% column) | 1078 (29.9%) |
|  | ﻿less HS | N (% column) | ﻿229 (6.4%) |
| **Age** |  | Mean (SD) | 5.72(0.364) |
| **Math score** | Period 1 | Mean (SD) | ﻿28.6 (9.48) |
|  |  | Missing | 0 |
|  | Period 2 | Mean (SD) | ﻿39.2 (12.3) |
|  |  | Missing | ﻿35 (1.0%) |
|  | Period 3 | Mean (SD) | ﻿47.3 (14.5) |
|  |  | Missing | ﻿2563 (71.2%) |
|  | Period 4 | Mean (SD) | ﻿65.3 (18.2) |
|  |  | Missing | ﻿121 (3.4%) |
|  | Period 5 | Mean (SD) | ﻿105 (23.3) |
|  |  | Missing | ﻿522 (14.5%) |
|  | Period 6 | Mean (SD) | ﻿130 (21.5) |
|  |  | Missing | ﻿1086 (30.2%) |
|  | Period 7 | Mean (SD) | ﻿148 (18.6) |
|  |  | Missing | ﻿1350 (37.5%) |

To study the association between math score and age, we add the age according to the difference of the actual year on the baseline age for each participant. After that, we draw the boxplot (**Figure 1**) with age as x and math score as y. From the figure, we can find there seems a linear relationship between math score and age, with some outliers suggesting the further transformation of the outcome variable - math score. Therefore, we examine their univariate association by the first fitting simple linear regression model and then check possible modifiers by adding race, gender, or/and dad’s education and corresponding interactions to the model. We find the coefficient estimators for all of the interaction terms are significant within the model, which means the association between

![Chart, scatter chart

Description automatically generated]()

Figure1. Box plot for change in math score for increasing age

age and math scores vary by race, gender, and education attainments of the father. When has one of the race or gender or education in the model they are not significant but when add all three of them at the same time they are all significant, suggesting race, gender, and education of father are confounders, so we add race, gender and education attainments of the father in the model, for the univariate analysis we don’t include the interaction term for now. Next, we plot DFFITS (id) against id to check for outliers, there are many outliers. Next, we conduct the residual analysis and find the assumptions of linearity and homoscedasticity are not reasonable, accordingly, we tried with different remedy measures, including 1) fitting poison logistic model, 2) sqrt transformation on the outcome variable, 3) log transform on the outcome (requires first change math score=0 to 1, since the transformation is not reasonable when math score=0, 4) adding a higher-order term of age into the model to fit a nonlinear model. Comparing the residual analysis results of all different methods, we decided to choose the second method - sqrt transformation on the outcome variable. Therefore, our final model at this point is: where represents *Black or African American, non-Hispanic, Hispanic and Asian* in race relatively*,* , , represents *HS or GED*, *College or AD* and *Graduate or Prof* in highest education attainments of father categories relatively.

Considering not fixed individual effects, based on previous model, we fit a random effect model: , where And accordingly we get the models for within-sample group and out-of-sample group are as below:

1) To predict math score at for participant id=:

where is the estimated difference between the mean of group i and the grand mean (the overall intercept) ；

2) Given for a new participant :

.

And with above models, we can calculate the rates of math score increase as a function of age for different gender\*education of father groups, the results are as showed in **Table 2**. As for interpretation, take first row in the table as an example: For female with ‘less HS’ highest education attainment of father, the square root of math score increases 0.78 with one year increases in age, and the 95% confidence interval is (0.75, 0.81).

**Table 2. Estimator and 95% confidence interval for rates of math score with age**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Group** | | **Estimator** | **95% confidence interval** | |
| **Lower bound** | **Upper bound** |
| **Female** | less HS | 0.78 | 0.75 | 0.81 |
| HS or GED | 0.80 | 0.78 | 0.81 |
| College or AD | 0.80 | 0.79 | 0.81 |
| Graduate or Prof | 0.78 | 0.77 | 0.80 |
| **Male** | less HS | 0.79 | 0.74 | 0.84 |
| HS or GED | 0.81 | 0.76 | 0.85 |
| College or AD | 0.80 | 0.76 | 0.84 |
| Graduate or Prof | 0.79 | 0.75 | 0.83 |

Next, we calculate the math score increase as a function of age for different gender\*race, the results are as showed in **Table 3**. As for interpretation, take first row in the table as an example: For ‘White, non-Hispanic ‘female, the square root of math score increases 0.78 with one year increases in age, and the 95% confidence interval is (0.75, 0.81).

**Table 3. Estimator and 95% confidence interval for rates of math score with age**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Group** | | **Estimator** | **95% confidence interval** | |
| **Lower bound** | **Upper bound** |
| **Female** | White, non-Hispanic | 0.78 | 0.75 | 0.81 |
| Black or African American, non-Hispanic | 0.77 | 0.73 | 0.81 |
| Hispanic | 0.80 | 0.76 | 0.84 |
| Asian | 0.82 | 0.78 | 0.86 |
| **Male** | White, non-Hispanic | 0.79 | 0.74 | 0.84 |
| Black or African American, non-Hispanic | 0.78 | 0.74 | 0.82 |
| Hispanic | 0.81 | 0.75 | 0.86 |
| Asian | 0.83 | 0.77 | 0.88 |

From the table, we can find between female and male group, the math score increasing rate across each race and education is similar. The Asian seems doing best on math score and the children whose father has a high school diploma or Associate’s or bachelor’s degree may do better than others.

However, the current analysis is with limitations. The intraclass correlation (ICC) of random effects models is ICC=0.723, it’s relatively high but that may because we delete all the missing data in math score variables. We first delete all the t3 score because it has missing data more than 70%, then we delete row that has missing in any period math score. We delete all the missing because a large percent of missing in math score variables may bias the result, since different participants may lack value for math score at different ages, then a different number of values for each age are used to conduct the model. However, just based on the non-missing data may also lead to biased and unreliable results. After deleting, the sample size became quite small, also causing a negative influence on the analysis, it cannot represent the actual trend in the original dataset.

Besides, the interpretation based on current analysis models is complex, due to square-root transformed dependent variable (math score) in our models. With this transformation, we can get simple explanation for each main coefficient as the change in square root of math score for per unit change in corresponding independent variable. However, this is not straightforward for us to learn about the exact rate of change in math score. And if we want to use another approach by squaring both sides to get the derivative of math score over main independent variables, then we will have many coefficients to consider, which means it is not quite easy to calculate the final result.

We use random intercept model that’s because the intercepts are allowed to vary across different child groups, meaning that they are different in different child. This shows that the student-specific effect, i.e. the deviation in math score just for that student being who they are, can be seen as an additional source of variance. The student effects are random, and specifically are normally distributed with mean of zero and some estimated standard deviation. In other words, conceptually the only difference between random intercept model and a standard regression is the student effect, which on average is no effect, but typically varies from student to student by some amount that is on average.

**Appendix 1. – Analysis Code**

## PART I. Exploratory Data Analysis------------------------------------------------------------------

## install required packadges

library(lme4)

library(table1)

library(tidyr)

## import the dataset

dat <- get(load('/Users/mengjiawei/Desktop/2020fall/525/Midterm/Midterm2020\_Data.RData'))

## add id to the data

dat$id <- 1:nrow(dat)

## 1. summary statistics for gender, race, dad\_edu, age and mathscore------------------------------------------------

## use table 1 function

label(dat$t1) <- "Period 1 score"

label(dat$t2) <- "Period 2 score"

label(dat$t3) <- "Period 3 score"

label(dat$t4) <- "Period 4 score"

label(dat$t5) <- "Period 5 score"

label(dat$t6) <- "Period 6 score"

label(dat$t7) <- "Period 7 score"

##convert age to years

dat$age<- round(dat$age\*0.08333333,1)

dat$race<- factor(dat$race, levels = c(1, 2, 3, 5), labels = c("White, non-Hispanic","Black or African American, non-Hispanic", "Hispanic","Asian"))

dat$gender <- factor(dat$gender, levels = c(1, 2), labels = c("male", "female"))

# descriptive table

table1(~ gender + race + dad\_edu + age + t1 + t2 + t3 + t4 + t5 + t6 + t7, data=dat)

## 2. checking missing pattern------------------------------------------------------------------------------

# 2.1 missing value for each column

library(mice)

md.pattern(dat, plot = FALSE, rotate.names = FALSE)

colMeans(is.na(dat))

# race gender dad\_edu age t1 t2 t3 t4 t5

#0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.009722222 0.711944444 0.033611111 0.145000000

#t6 t7 id

#0.301666667 0.375000000 0.000000000

# t3 has NA more than 71%, maybe need to drop that variable

# 2.2 study pattern focusing on score variable, group by gender\*education

# write a loop to calculate the number of missing value in score for each id

#change the wide form to long form

dat1 = gather (dat, key=sample, value = mathscore, c(t1,t2,t3,t4,t5,t6,t7))

dat1

dat1$id

dat1$period = rep (c(1,2,3,4,5,6,7), 3600)

dat1 <- dat1[order(dat1$id, dat1$sample, dat1$period),]

dat1

num <- length(unique(dat$id))

nmv <- vector()

for (i in 1:num) {

nmv[i] = nrow(dat1[dat1$id==unique(dat1$id)[i] & is.na(dat1$mathscore),])

}

nmv

# create a subset containing the id, gender, edu, # of NA in math score for each patient

ref = dat1[!duplicated(dat1$id),1:3]

ref$num\_NA = nmv

ref

# get summary statistics for missing value for different groups according to gender\*edu

library(dplyr)

sta = ref %>%

filter(!is.na(dad\_edu)) %>%

group\_by(gender,dad\_edu) %>%

summarise(N=length(num\_NA),n=length(num\_NA[num\_NA!=0]), mu = mean(num\_NA[num\_NA!=0]))

sta

#drop t3 because of many NA

dat3<-subset(dat,select = -t3)

colMeans(is.na(dat3))

#Still many NA in mathscore, delete the row if they have NA in t1-t7

dat4<-na.omit(dat3, cols=c("t1", "t2","t4","t5","t6","t7"))

dim(dat4)

#change the long form to wide form

library (tidyr)

dat5 = gather (dat4, key=sample, value = mathscore, c(t1,t2,t4,t5,t6,t7))

dat5$period = rep (c(1,2,4,5,6,7), 2129)

dat5 <- dat5[order(dat5$id, dat5$sample, dat5$period),]

#add different age according to actual year to baseline

for (f in 1:max(dat5$id)){

dat5$age[dat5$id == f][2]<-dat5$age[dat5$id == f][1] + 0.5

dat5$age[dat5$id == f][3]<-dat5$age[dat5$id == f][2] + 1

dat5$age[dat5$id == f][4]<-dat5$age[dat5$id == f][3] + 2

dat5$age[dat5$id == f][5]<-dat5$age[dat5$id == f][4] + 2

dat5$age[dat5$id == f][6]<-dat5$age[dat5$id == f][5] + 3

}

#plot to see trend for male and female mathscore vs age

library (ggplot2)

p <- ggplot(dat5, aes(x = age, y = mathscore, group = id, color = factor(id)))

p + geom\_line() + facet\_grid(. ~ gender, labeller = labeller(gender=c('1'="Male", '2'="Female")))+

guides(colour=FALSE)

## 3. check trend of math score with agefit model --------------------------------------------------------------

# 3.1 three visualize plots

plot(mathscore~age,data=dat5)

boxplot(mathscore~age,data=dat5,ylab = "mathscore",col="red") ## outliers

hist(dat5$mathscore)

# 3.2 univariate association between age and mathscore

## fit SLR model

dat5$dad\_edu <- relevel(factor(dat5$dad\_edu),ref = "less HS") # set reference group

dat5$race <- relevel(factor(dat5$race),ref = "White, non-Hispanic") # set reference group

dat5[dat5$gender==2,]$gender=0

model1 = lm(mathscore~age,data=dat5)

summary(model1)

## conduct residual analysis

library(car)

qqnorm(model1$residuals)

qqline(model1$residuals)

## check modifiers and confounders

#1) for interaction and confounder

#########gender

M1 <- lm(mathscore ~ age, data=dat5)

M2 <- lm(mathscore ~ age + gender + age\*gender, data=dat5)

summary(M1)

summary(M2) #gender has interaction

#confounder

M3<- lm(mathscore ~ age + gender, data=dat5)

((coef(M1)[2]-coef(M3)[2])/coef(M1)[2])\*100 #gender not confounder

#########race

M4 <- lm(mathscore ~ age + factor(race) + age\*factor(race), data=dat5)

summary(M4) #race has interaction

#confounder

M5 <- lm(mathscore ~ age + factor(race), data=dat5)

((coef(M1)[2]-coef(M5)[2])/coef(M1)[2])\*100 #race not confounder

#########dad\_edu

M6 <- lm(mathscore ~ age + factor(dad\_edu) + age\*factor(dad\_edu), data=dat5)

summary(M6) #edu has interaction

#confounder

M7 <- lm(mathscore ~ age + factor(dad\_edu) , data=dat5)

((coef(M1)[2]-coef(M7)[2])/coef(M1)[2])\*100 #edu not confounder

#########gender, race and dad\_edu

M8 <- lm(mathscore ~ age + gender + factor(race) + factor(dad\_edu), data=dat5)

summary(M8) #if put them in the model together, the gender, race and edu are all confounder for the mathscore so add all of them in the model

#final model for univariate analysis

model.int = lm(mathscore ~ age + gender + factor(race) + factor(dad\_edu) , data=dat5)

summary(model.int)

#this is the primary final model

## conduct residual analysis

library(car)

qqnorm(model.int$residuals)

qqline(model.int$residuals)

## need transformation

#outlier plot

par(mfrow=c(2,2))

plot(model.int)

#dffits

IDs<-dat5$id

n <- nrow(dat5)

p <- 6

dat5$fits <- dffits(model.int)

sub\_fit <- dat5[abs(dat5$fits) > 2\*sqrt(p/n),]

plot(IDs, dat5$fits, xlab="ID", ylab="DFFITS")

points(sub\_fit$id,sub\_fit$fits,col = "red")

# 3.3 remedy measures

#1) poisson regression

model.fn = glm(mathscore ~ age + gender + factor(race) + factor(dad\_edu),data=dat5,family = poisson)

summary(model.fn)

qqnorm(model.fn$residuals)

qqline(model.fn$residuals)

## not good

par(mfrow=c(2,2))

plot(model.fn)

#2) sqrt transformation on math score

model.sq = lm(sqrt(mathscore) ~ age + gender + factor(race) + factor(dad\_edu), data=dat5) ## due to 0 mathscore

summary(model.sq)

qqnorm(model.sq$residuals)

qqline(model.sq$residuals)

par(mfrow=c(2,2))

plot(model.sq)

#3) log transformation on math score

dat6 = na.omit(dat5)

dat6[dat6$mathscore==0,"mathscore"]=1

model.log = lm(log(mathscore) ~ age + gender + factor(race) + factor(dad\_edu), data=dat6)

summary(model.log)

qqnorm(model.log$residuals)

qqline(model.log$residuals)

## not good

#4) add higher order term of age

model.ho = lm(mathscore ~ age + age^2 + gender + factor(race) + factor(dad\_edu), data=dat5) ## due to 0 mathscore

summary(model.ho)

qqnorm(model.ho$residuals)

qqline(model.ho$residuals)

## not good

## PART II. MODELING-------------------------------------------------------------------------------------------

##1) rates in math score as a function of age

## fit random effect model, with three interaction terms, group by id

library(lmerTest)

fit\_ran <- lmer((sqrt(mathscore))~(1|id)+age + gender + factor(race) + factor(dad\_edu) + age\*gender + age\*factor(race) + age\*factor(dad\_edu), data=dat5)

summary(fit\_ran)

vcov(fit\_ran)

#id (Intercept) 0.4244 0.6515

#Residual 1.1099 1.0535

## calculate 95% ci for each change rate in 8 gender\*education group

#1. female dad\_edu1

est1 = 7.839e-01

ci.f1 = 7.839e-01+c(-1,1)\*1.96\*1.530e-02

#2. female dad\_edu2

est2 = 7.839e-01 + 1.334e-02

ci.f2 = est2+c(-1,1)\*1.96\*sqrt(1.530e-02^2+2\*(-2.210905e-04)+ 1.592e-02^2)

#3. female dad\_edu3

est3 = 7.839e-01 + 1.304e-02

ci.f3 = est3+c(-1,1)\*1.96\*sqrt(1.530e-02^2+2\*(-2.217271e-04)+1.547e-02^2)

#4. female dad\_edu4

est4 = 7.839e-01 - 7.546e-04

ci.f4 = est4+c(-1,1)\*1.96\*sqrt(1.530e-02^2+2\*(-2.225971e-04)+1.669e-02^2)

#5. male dad\_edu1

est5 = 7.839e-01 + 7.901e-03

ci.f5 = est5+c(-1,1)\*1.96\*sqrt(1.530e-02^2+2\*1.557137e-04+6.211e-03^2)

#6. male dad\_edu2

est6 = est5+1.334e-02

ci.f6 = est6+c(-1,1)\*1.96\*sqrt(1.530e-02^2+1.592e-02^2+6.211e-03^2+

2\*1.557137e-04+

2\*(-2.210905e-04)+

2\*1.579553e-05)

#7. male dad\_du3

est7 = est5+1.304e-02

ci.f7 = est7+c(-1,1)\*1.96\*sqrt(1.530e-02^2+1.547e-02^2+6.211e-03^2+

2\*1.557137e-04+

2\*(-2.217271e-04)+

2\*1.749730e-05)

#8. male dad\_edu4

est8 = est5-7.546e-04

ci.f8 = est8+c(-1,1)\*1.96\*sqrt(1.530e-02^2+1.669e-02^2++6.211e-03^2+

2\*1.557137e-04+

2\*(-2.225971e-04) +

2\*1.625832e-05 )

est = c(est1,est2,est3,est4,est5,est6,est7,est8)

ci = c(ci.f1,ci.f2,ci.f3,ci.f4,ci.f5,ci.f6,ci.f7,ci.f8)

round(est,2)

round(ci,2)

## intraclass correlation

rou <- 1.1099/(1.1099+0.4244) #0.7234

rou

## calculate 95% ci for each change rate in 8 race\*education group

#1. female race1

est11 = 7.839e-01

ci.f11 = 7.839e-01+c(-1,1)\*1.96\*1.530e-02

#2. female race2

est22 = 7.839e-01 -1.390e-02

ci.f22 = est22+c(-1,1)\*1.96\*sqrt(1.530e-02^2+2\*(-3.522905e-05)+ 1.745e-02^2)

#3. female race3

est33 = 7.839e-01 + 1.415e-02

ci.f33 = est33+c(-1,1)\*1.96\*sqrt(1.530e-02^2+2\*(-4.093082e-05)+1.522e-02^2)

#4. female race5

est44 = 7.839e-01 + 3.453e-02

ci.f44 = est44+c(-1,1)\*1.96\*sqrt(1.530e-02^2+2\*(-2.201436e-05)+1.544e-02^2)

#5. male race1

est55 = 7.839e-01 + 7.901e-03

ci.f55 = est55+c(-1,1)\*1.96\*sqrt(1.530e-02^2+2\*1.557137e-04+6.211e-03^2)

#6. male race2

est66 = est55-1.390e-02

ci.f66 = est66+c(-1,1)\*1.96\*sqrt(1.530e-02^2+1.745e-02^2+6.211e-03^2+

2\*1.557137e-04+

2\*(-2.210905e-04)+

2\*1.048046e-06)

#7. male race3

est77 = est55+1.415e-02

ci.f77 = est77+c(-1,1)\*1.96\*sqrt(1.530e-02^2+1.522e-02^2+6.211e-03^2+

2\*1.557137e-04+

2\*(-3.522905e-05)+

2\*3.967480e-06)

#8. male race5

est88 = est55 + 3.453e-02

ci.f88 = est88+c(-1,1)\*1.96\*sqrt(1.530e-02^2+1.544e-02^2++6.211e-03^2+

2\*1.557137e-04+

2\*(-2.201436e-05) +

2\*1.790134e-05)

estr = c(est11,est22,est33,est44,est55,est66,est77,est88)

cir = c(ci.f11,ci.f22,ci.f33,ci.f44,ci.f55,ci.f66,ci.f77,ci.f88)

round(estr,2)

round(cir,2)