Do Race, Family Income, the One-child Situation in the Family, and Mother’s Education Influence Children’s School Choice Options?

Yanxi Zeng

MSSP 897 Applied Linear Modeling

Professor: Ezekiel Dixon-Román

December 16, 2020

Contents

1. **Introduction** **1**
2. **Method** **2**
   2.1 Data and sample 2
   2.2 Variables 3
   2.3 Rationale/description of the analysis 3
3. **Results** **4**
   3.1 Descriptive statistics and interpretation 4
   3.2 The logistic regression analysis 6
      3.2.1 Logistic model and model assumption evaluation 6
      3.2.2 Logistic regression results and interpretation 7
      3.2.3 Logistic model fit assessment 8
4. **Conclusion and implications** **9**
5. **Reference** **10**
6. **Appendix** **11**

1. Introduction

The research question of this study is: do race, family income, the one-child situation in the family, and mother's education influence children's school choice options? The children's school choice options discussed in this study also are the two most prevalent education options in America: the private school and public school.

Existing literature has determined that the child's race, parental education, family income, and the number of children in the family are significant social and family factors in determining children's school choice options. Firstly, the child's race as an important predictor of school choice, though the results are often inconclusive. For example, Coleman and Hoffer (1987) found that black and Hispanic students are under-represented in private schools when compared to white students. The second important factor of school choice can be parents' education. According to Ball (2003), parents with higher education more likely to value education and consider private school options with higher education quality for their children. Empirical evidence also consistently demonstrates a positive relationship between parents' education and the likelihood that they would send their children to a private school (Coleman et al., 1982; Lankford & Wyckoff, 1992). Next, family income is another demographic indicator that has been positively related to choosing schools. A higher family income, as an indicator of resources, can increase one's chances of affording an expensive private education. Coleman and Hoffer (1987) also found that the higher the family income, the higher the private school enrolment. Lastly, the number of children in the family is also likely to influence the education choices in the family: the number of children limits the amount of resources that can be spent on any one child (Reay & Lucey, 2000).

To sum up, based on the literature review, this study will identify variation by race, family income, the one-child situation in the family, and mother's completed education level in school choices. This study contributes to discussions about social and racial inequality in school choice options.

This report is divided into four parts to further examine the research question: introduction, methods, results, and conclusions. The syntax of R used for this research will also be provided as an appendix.

2. Method

2.1 Data and sample

The data/sample used for this study came from the Child Development Supplement to the Panel Study of Income Dynamics (PSID), a national probability sample (Sandra et al., 1997). This data source includes a nationally representative sample of families with young children, and the family wealth history data in PSID have few missing data (Yeuing & Conley, 2008). Therefore, this data source is appropriate to be used for this research question related to the children's school choice options and family backgrounds.

However, in the actual analysis process, 2534 observations are removed due to missingness or not belonging to the research scope (for example, children who attended kindergarten in 1997). Therefore, the sample size of this study is 1029.

2.2 Variables

- **Dependent variable**

This study's dependent variable is the children's school choice options (SCHOOL), which is a binary (or categorical) variable. This variable will be coded as 1 if the child attended a private school and coded as 0 if he/she attended a public school.

- **Independent variables**

There are four independent variables in this study: race (RACE), family income (faminc97), the one-child situation in the family (onlychild), and mother's completed education level (motheredu). Among them, only family income is a continuous variable, and the rest three are binary (or categorical) variables.

In detail, the race is coded as -0.5 if the child's race is black and coded as 0.5 if it is white; the one-child situation in the family is coded as 0.5 if the child is the only child in the family, -0.5 otherwise; mother's completed education level is coded as 0.5 if the mother attended college, -0.5 otherwise. In terms of the continuous variable, family income in 1997 has a broad range between \$0 and \$784610.59 (in 2002 constant dollars).

2.3 Rationale/description of the analysis

The logistics regression is utilized as the modeling technique in this study. This choice is mainly due to the binary outcomes of this research question: children attended a private school or public school. Because the general linear model cannot account for outcomes with binomial distributions which have non-linear relationships with other variables, the logistic modeling that focuses on analyzing binary outcomes is more appropriate for this study.

Then, the rationale of the analysis is as follows: (1) Conducting descriptive statistics of the sample. A brief interpretation of this descriptive statistics will also be given. (2) Constructing a logistic model for the research question and conducting the logistic regression for this model. The model assumptions and model fit will be evaluated; the regression results and relevant interpretation will also be provided.

3. Results

3.1 Descriptive statistics and interpretation

Table 1. Frequency Distribution of Children's School Choice Options

Children's School Choice Options	Number	Percentage
Private school	121	11.76%
Public school	908	88.24%

Table 2. Frequency Distribution of Race

Race	Number	Percentage
Black	331	32.17%
White	698	67.83%

Table 3. Frequency Distribution of Mother's Completed Education Level

Mother's Completed Education Level	Number	Percentage
Not attend college	797	77.45%
Attend college	232	22.55%

Table 4. Frequency Distribution of the One-child Situation in the Family

The One-child Situation in the Family	Number	Percentage
Only one child in the family	135	13.12%
More than one child in the family	894	86.88%

Table 5. Means, Median, Standard Deviations, and Observed Ranges for the Total Sample (N = 1029)

Variables	**Mean**	**Median**	**SD**	**Observed range**
Dependent variable				
Children's School Choice Options	0.12	0	0.32	1
Independent variables				
Family Income	71206.64	57164.49	59504.7	646743.3
Race	0.18	0.5	0.47	1
Mother’s Completed Education Level	-0.27	-0.5	0.42	1
The One-child Situation in the Family	-0.37	-0.5	0.34	1

After removing 2534 observations that are missing or not belonging to the research scope, the sample size of this study is 1029. Among them, 908 children attended a public school, accounting for nearly 88% of the sample, while only 121 children’s education option is the private school (Table 1). And there are 331 black children and 698 white children in the sample (Table 2). In terms of these children’s family background, only 13% of children (135) are the only child in their families in 1997, while 894 children came from a multi-child family; about only 23% of these children’s mothers once attended college (Table 3 and 4).

According to Table 5, the average family income in this sample is 71,206.64 dollars; however, the standard deviation of this variable is large (59,504.7), which means that the family income varies greatly for different participants. Especially, as the difference between its median (57,164.49 dollars) and mean is also large (14,042 dollars), we can further conclude that the family income data tends to be rightward skewed.

3.2 The logistic regression analysis

3.2.1 Logistic model and model assumption evaluation

Based on the research question, we constructed a logistic model with all variables as follows:

$$\log it(\pi_i) = \beta_0 + \beta_1 faminc97 + \beta_2 RACE + \beta_3 motheredu + \beta_4 onlychild$$

$$\text{(Or, } \log it[\pi(x)] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \text{)}$$

This model represents the children's school choice options (SCHOOL) is regressed on family income (faminc97), race (RACE), mother's completed education level (motheredu), and the one-child situation in the family (onlychild). In this model, logit(π_i)/logit[$\pi(x)$] represents the probability that children's education choice is the private school given X.

Besides, there are no assumption violations in this model. In detail, the outcome in this model is binary (attending a public school/private school), and the model is perfect multicollinearity. As Figure 1 shown, since the association between the family income variable and the log odds is not biased downward, the family income variable, as the only continuous independent variable in the model, also adheres to the assumption of linearity.

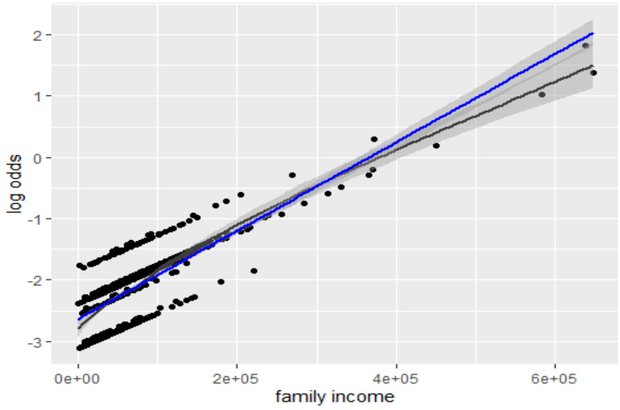


Figure 1. The Association Between the Family Income Variable and the Log Odds

3.2.2 Logistic regression results and interpretation

The logistic regression results of this model are shown in Table 6.

Table 6. The Logistic Regression Results

	Logit coefficients	Odds ratio
(Intercept)	-2.452*** (0.203)	0.086
Family Income	5.748e-06** (1.391e-06)	1.0000057
Race	0.731** (0.260)	2.076
Mother's Completed Education Level	0.057 (0.232)	1.059
The One-child Situation in the Family	0.547* (0.267)	1.728
N	1029	/

*p < 0.05, **p < 0.01, ***p < 0.001

According to Table 6, the p-value of family income, race, and the one-child situation in the family variables are all less than 0.05, indicating these three independent variables are statistically significant. On the other hand, the p-value of mother's completed education level variable is greater than 0.05, which means that this variable is not statistically significant, and changes in this predictor (mother's completed education level) are not associated with the odds or odds ratio changes in children's school choice options.

The odds ratio for family income can express the percent change in the odds for each dollar increase in the family income. However, the family income variable's effect on the likelihood that children attended a private school cannot be represented explicitly due to the improper measurement scale. After enlarging its measurement scale, we can conclude that when 1,000 dollars increase in family income, there will be a roughly 0.57% increase in the odds that children attended a private school rather than a public school.

The odds ratio for race is 2.08, which means that the predicted odds of attending a private school are 2.08 times greater for white children than black children. Lastly, the odds ratio for the one-child situation in the family is 1.73, indicating that the predicted odds that children's education choice is the private school are 1.73 times greater for children who is the only child in their families than children from a multi-children family.

3.2.3 Logistic model fit assessment

Table 7. Model Fit Statistics

	The Null Model Without Predictors	**The Model with Predictors**
AIC	747.1922	716.1904
BIC	752.1286	740.8721
neg2LL	745.1922	706.1904

As Table 7 shown, the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Negative 2 Log Likelihood (-2LL) are utilized to assess the model fit. Since the values of AIC, BIC, and -2LL corresponding to the model with predictors are all smaller than those corresponding to the intercept-only model, it can be concluded that the model with predictor variables we constructed is a better fit.

4. Conclusion and implications

This study's research question is to examine whether race, family income, the one-child situation in the family, and mother's completed education level have an influence on children's school choice options. The logistic modeling is utilized in this study to answer this research question.

After analyzing the regression results of the logistic model, we can conclude that race, family income, and the one-child situation in the family, as social and racial factors, can influence children's school choice options. However, mother's completed education level variable has no such influence on school options: changes in mother's completed education level are not associated with the odds or odds ratio changes in children's school choice options.

Meanwhile, as a possible limitation of this study, only the mother's completed education level but not the more convictive parental education level is being added into the logistic model as an independent variable.

5. Reference

Ball, S. J. (2003). *Class strategies and the education market: The middle classes and social advantage*. London: Routledge Farmer.

Coleman, J., and T. Hoffer. (1987). *Public and private high schools.* New York: Basic Books.

Coleman, J., T. Hoffer, and S. Kilgore. (1982). *High school achievement: Public, Catholic, and private schools compared*. New York: Basic Books.

Sandra, H., Pamela, E., Jean, D. & Jonathan, F. (1997). The Child Development Supplement to the Panel Study of Income Dynamics 1998 User Guide. Retrieve from https://psidonline.isr.umich.edu/CDS/cdsi_userGD.pdf

Reay, D., & Lucey, H. (2000). Children, school choice, and social differences. Educational Studies, 26(1), 83–100.

Yeung, W., & Conley, D. (2008). Black-White Achievement Gap and Family Wealth. Child Development, 79(2), 303-324. Retrieved from http://www.jstor.org/stable/27563485

6. Appendix

Import dataset and library.

```
good <- read.csv('good.csv',header = TRUE, sep = ',')
library(ggplot2)
```

Set required variables

```
# Set the RACE variable
good$RACE <- ifelse(good$CHRACE == 9, NA,
                  ifelse(good$CHRACE == 1, .5,
                         ifelse(good$CHRACE == 2, -.5, NA)))

# Set the SCHOOL variable
good$SCHOOL <- ifelse(good$Q1G10 == 0, NA,
                    ifelse(good$Q1G10 == 1, 0,
                           ifelse(good$Q1G10 == 2, 1, NA)))

# Set the onlychild variable
good$onlychild <- ifelse(good$childnum == 1, .5,
                       ifelse(good$childnum > 1, -.5, NA))


# Set the motheredu variable
good$motheredu <- ifelse(good$ER11743 == 0, NA,
                       ifelse(good$ER11743 == 98, NA,
                              ifelse(good$ER11743 == 99, NA,
                                     ifelse(good$ER11743 <= 5, -.5,
                                            ifelse(good$ER11743 >5, .5,
 NA)))))
```

Set the schoolset dataset including required variables

```
schoolset <- na.omit(good[,c("RACE","faminc97","motheredu","onlychil
d","SCHOOL")])
```

Descriptive statistics

```
nrow(good)

## [1] 3563

nrow(schoolset)

## [1] 1029
```

```
# Therefore, the sample size is 1029 (3563 - 2534)
summary(schoolset)
```

```
##       RACE             faminc97        motheredu         onlychild

##  Min.   :-0.5000   Min.   :     0   Min.   :-0.5000   Min.   :-0.500
0
##  1st Qu.:-0.5000   1st Qu.: 40351   1st Qu.:-0.5000   1st Qu.:-0.50
00
##  Median : 0.5000   Median : 57165   Median :-0.5000   Median :-0.50
00
##  Mean   : 0.1783   Mean   : 71207   Mean   :-0.2745   Mean   :-0.36
88
##  3rd Qu.: 0.5000   3rd Qu.: 86945   3rd Qu.:-0.5000   3rd Qu.:-0.50
00
##  Max.   : 0.5000   Max.   :646743   Max.   : 0.5000   Max.   : 0.500
0
##      SCHOOL
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1176
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

```
sapply(schoolset, sd, na.rm=TRUE)
```

```
##         RACE     faminc97    motheredu    onlychild       SCHOOL
## 4.673449e-01 5.950466e+04 4.180892e-01 3.377780e-01 3.222785e-01
```

```
table(schoolset$SCHOOL)
```

```
##
##   0   1
## 908 121
```

```
table(schoolset$RACE)
```

```
##
## -0.5  0.5
##  331  698
```

```
table(schoolset$motheredu)
```

```
##
## -0.5  0.5
##  797  232
```

```r
table(schoolset$onlychild)
```

```
##
## -0.5  0.5
##  894  135
```

Conduct the logistic regression

```r
lm1<-glm(SCHOOL~RACE + faminc97 + motheredu + onlychild, family=binom
ial(link='logit'), data=schoolset)
exp(cbind(OR = coef(lm1), confint(lm1)))
```

```
##                    OR      2.5 %    97.5 %
## (Intercept) 0.0861164 0.05703591 0.1263898
## RACE        2.0762686 1.26909062 3.5366944
## faminc97    1.0000057 1.00000312 1.0000086
## motheredu   1.0584963 0.66374975 1.6518187
## onlychild   1.7278226 1.00285812 2.8737336
```

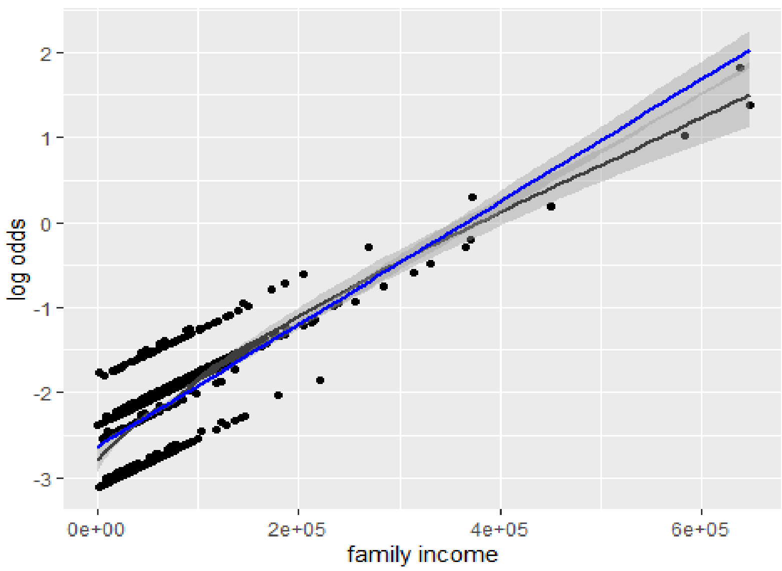
```r
summary(lm1)
```

```
##
## Call:
## glm(formula = SCHOOL ~ RACE + faminc97 + motheredu + onlychild,
##     family = binomial(link = "logit"), data = schoolset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7937  -0.5226  -0.4667  -0.3366   2.4903
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.452e+00  2.025e-01 -12.106  < 2e-16 ***
## RACE         7.306e-01  2.602e-01   2.807   0.0050 **
## faminc97     5.748e-06  1.391e-06   4.133 3.58e-05 ***
## motheredu    5.685e-02  2.319e-01   0.245   0.8064
## onlychild    5.469e-01  2.674e-01   2.045   0.0408 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 745.19  on 1028  degrees of freedom
## Residual deviance: 706.19  on 1024  degrees of freedom
## AIC: 716.19
```

```
##
## Number of Fisher Scoring iterations: 5
```

Diagnose the model assumptions

```
# diagnose the multicollinearity
library(car)

## Loading required package: carData

vif(lm1)

##      RACE  faminc97 motheredu onlychild
##  1.069082  1.076852  1.067369  1.021380

#As VIFs for these variables are less than 10, we can conclude that t
his model is perfect multicollinearity

# check the assumption of linearity
linearity <- glm(SCHOOL ~ . , family=binomial(link='logit'),
                 data=schoolset)
logodds <- predict(linearity)
plotlin <- with(schoolset, data.frame(faminc97 = faminc97,
                                      logit = logodds))
ggplot(plotlin, aes(x = faminc97, y = logit))+
  geom_point()+
  labs(x = "family income", y = "log odds") +
  geom_smooth(method = "loess", col = "#3e3e3e")+
  geom_smooth(method = "lm", col = "blue")

## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



```
# As the plot shown, the faminc97 variable adheres to the assumption
of linearity.
```

Assess the model fit

```
# Construct  the null model without predictors
lm1null<-glm(SCHOOL~1, family=binomial(link='logit'), data= schoolse
t)

# -2LL
lm1$null.deviance

## [1] 745.1922

lm1$deviance

## [1] 706.1904

# AIC
lm1null$aic

## [1] 747.1922

lm1$aic

## [1] 716.1904

# Schwarz Criterion
BIC(lm1null)

## [1] 752.1286

BIC(lm1)

## [1] 740.8721
```