Independent Project

How Parenting Helps Children to Earn more:

The Effect of Concerted Cultivation on Children's Future Income

Introduction

The topic of my project is about the upward mobility and the certain practices of children's parents during their childhood, to be more specific, it is about the intergenerational transition of the family advantages and children's later success.

The certain question I would like to work on in this project is about the relationship between children's future income and their parents' parenting style.

According to the expert in this field, the sociologist Annette Lareau, middle-class family parents adopt "concerted cultivation"[1] to help their children get familiar with how institute works, and gain sufficient skills to guarantee success in these institutions. The practices in details includes arrange organized activities for their children, negotiate with them instead of hard punishment, work well with schools and teachers and be informative[2]. Therefore, in the light of her research findings, I pick out some variables that can partially represent these traits.

First of all, the major independent variables: 1) the organized activities they attend in hours every weekday; 2) their relationship with their parents; 3) their parents' attendance in school activities and how close they are with the institutes; 4) parents' company with their children and the activities they do together.

1 Lareau, A. (2011). Unequal Childhoods: Class, Race, and Family Life (2nd ed.). Berkeley: University of California Press.

2 Lareau, A. (2015). Cultural Knowledge and Social Inequality. American Sociological Review, 80(1), 1–27.

My dependent variable is children's income after high school.

My hypothesis is that even after controlling the effect of family income, the "good" parenting style still has an effect on children's future outcomes. The more time children devoted to organized activities, the closer parents are with their children, the more positive parents work with schools and teachers, the more likely their children would be having a well-paid job.

Description of Data Set and Variables

The dataset I adopt for this project is HSLS:09, namely High School Longitudinal Study of 2009. (https://nces.ed.gov/surveys/hsls09/index.asp). The data collectors are meant to represent all the ninth graders in 2009 around the United States and thus their samples are over 23,000 (and they got 21,444 observations who responded then) and students are from 944 schools. As it is a longitudinal study, the first round happened in 2009, and the follow-ups are in 2012 and 2016. They do not just ask questions about students' test scores in multiple disciplines but also about their lifestyles, their parents' backgrounds. Besides, they also collected data from respondents' math teachers, school counselors, and administrators.

The reasons why I choose this data are 1) they are longitudinal data, so they have both children's lifestyle records when they are still adolescents, and also their income and occupations information when they enter the job market. Therefore, most of the variables that I need are feasible in this dataset. 2) the data in HSLS:09 is relatively trustworthy --they sampled from a relatively complete sampling frame, and their response rate is relatively high (21444 students out of about 24000 samples), though the response rate does not always matter. Besides, they also provide weighted data based on the true distribution of high school 9th graders. Therefore,

although I cannot say for sure that the data are error-free, at least they tried to minimize the possible error.

The limitation of this dataset is that they do not include more information about students' lifestyle which is also part of concerted cultivation. For example, they do not have indicators showing how their parents communicate with them, or how often their parents will communicate with their teachers. Besides, mingling with friends or relatives is also an effective indicator to show the difference in childrearing. But unfortunately, they do not have this variable included in their study.

There are four categories of variables that I use in my model.

1) The **dependent variable** I use is the question in a recent follow-up they ask about the children's gross income throughout one year (in 2015, 2016 or 2017), depending on how the survey organizer gathered the information. This variable -- **incomecat** -- is shown in the categorical form:

1	No income
2	1,000 dollars or less
3	1,001 to 2,500 dollars
4	2,501 to 5,000 dollars
5	5,001 to 10,000 dollars
6	10,001 to 15,000 dollars
7	15,001 to 20,000 dollars
8	20,001 to 25,000 dollars
9	25,001 to 30,000 dollars
10	30,001 to 35,000 dollars
11	35,001 to 45,000 dollars
12	45,001 to 55,000 dollars
13	55,000 dollars and above

2) One part of the independent variables include children's behaviors during their childhood, which are believed to be planned and arranged by their parents, also their reliance on their parents. These variables are collected when they were still 9th graders in 2009.

noActivity: this is a variable ask if students have any organized activities. 1 means no any activity, 0 means they have activities. (I will re-code them reversely.)

hrActivity: its question is "During a typical weekday during the school year how many hours do you spend participating in extracurricular activities such as sports teams, clubs, band, student government?" The answer is coded from range 1 to 6, 1 indicates less than one hour, 2 means one to two hours, 3 means two to three hours, 4 means three to four hours, 5 means four to five hours and 6 means five or more hours.

momTalkPrb; dadTalkPrb: these two variables indicates "Since the beginning of the last school year (2008-2009), which of the following people have you talked with about personal problems?". If teenagers chose mother or female guardian, then the variable **momTalkPrb** is 1, if they chose father or male guardian, then the variable **dadTalkPrb** will be 0, otherwise they will either or both be 0.

3) Another part of the independent variables are behaviors about parents and their relationship with children and schools. They are also collected in 2009.

p1hhTime, this variable comes from the question: "How much of the time does [your 9th-grader] live with you?", and the answers are from 1 to 5, 1 means all the time, 3 means half of the time and 5 means none of the time.

p1hwOften, this comes from "During this school year, about how many days in an average week do you or another adult in your household help [your 9th-grader] with homework?" 1 indicates never, 2 means less than once a week, 3 means one or two days a week, 4 means three or four days a week or 5 means five or more days a week.

p1ptoMtg, this is from "Since the beginning of this school year (2009-2010), have you or other adults in your household attended a meeting of the parent-teacher organization or association?" 1 means yes, 0 means no.

p1ptConfer, this is from the question "Since the beginning of this school year (2009-2010), have you or other adults in your household gone to a regularly scheduled **parent-teacher conference** with [your 9th-grader]'s teacher?" 1 means yes, 0 means no.

p1schevent, the question for this is "Since the beginning of this school year (2009-2010), have you or other adults in your household attended a school or class event such as a play, dance, sports event or science fair because of [your 9th-grader]?" 1 means yes, 0 means no.

p1fundRaise, "Since the beginning of this school year (2009-2010), have you or other adults in your household participated in fundraising for the school?" 1 means yes, 0 means no.

p1Volunteer, "Since the beginning of this school year (2009-2010), have you or other adults in your household served as a volunteer in [your 9th-grader]'s classroom or elsewhere in the school?"1 means yes, 0 means no.

4) The last part of the variables included in my model are control variables. They are the basic information about children's childhood and their families. The most major one is family income. I see this variable as a competing variable towards my independent

variables of interest, simply because some people view economic factors as the element that can explain everything. However, my trial in this project is to stress the effect of soft practices that matter. Other than that, I also include children's **sex**, **race**, **region**, both **their parents' educational level**, **employment status and race**. (I did not include their age in this model because they are all 9th graders in 2009.)

In terms of re-coding, the original dataset coded situations like "not applicable" or "non response" as negative numbers like -8 or -9. I coded all of these negative numbers as NA, and afterwards I can deal with NA to make the model results more reasonable.

I also recode the variable **p1hhtime** reversely, so that now 1 means little company time while 5 indicates all the time live together.

For **momRace**, **dadRace**, **momEmp**, **dadEmp**, **momEdu** and **dadEdu**, those who do not have female or male guardian in household are coded 0, I changed them to NA, dealing with them like missing data.

For **noactivity**, I re-coded it so that now when this variable is equal to 0, it means the student has no activity, when it is 1, meaning he or she has activity.

Descriptive Statistics

1. Control Variables (categorical variables)

Table 1 Control variable: sex

	sex	children
1	male	11973 (50.94%)
2	female	11524 (49.03%)
	NA’s	6 (0.00%)

There are almost half women half men in this dataset, though men are a little bit more than women.

Table 2 Control variables: Children's and Parents' Race

	Race	children	momRace	dadRace
1	Amer. Indian/Alaska Native, non-Hispanic	165(0.70%)		
2	Asian, non-Hispanic	1952(8.31%)	1241(5.28%)	1056(4.49%)
3	Black/African-American, non-Hispanic	2450(10.42%)	1355(5.77%)	960(4.08%)
4	Hispanic, no race specified	422(1.80%)	585(2.49%)	468(1.99%)
5	Hispanic, race specified	3375(14.36%)	1459(6.21%)	1146(4.88%)
6	More than one race, non-Hispanic	1941(8.26%)	514(2.19%)	352(1.50%)
7	Native Hawaiian/Pacific Islander, non-Hispanic	110(0.47%)		
8	White, non-Hispanic	12082(51.41%)	9512(40.47%)	8359(35.57%)
9	Other, non-Hispanic		147(0.63%)	107(0.46%)
	NA’s	1006(4.28%)	8690(36.97%)	11055(47.4%)

White people are always taking the largest proportion of the sample. There are more “more than one race” students than parents, meaning that some of these parents are married interracially. There are a lot of missing data in parents part, especially for father. So maybe I should learn how to reweight them later in order to get an unskewed results.

Table 3 Control Variables: Parents' Employment Status

	Employment Status	**momEmp**	**dadEmp**
1	Never worked for pay	733(3.12%)	180(0.77%)
2	Not currently working for pay, has in the past	4026(17.13%)	1806(7.68%)
3	Currently working for pay PT (<35 hrs/wk)	2698(11.48%)	513(2.18%)
4	Currently working for pay FT (>=35 hrs/wk)	8152(34.68%)	10549(44.88%)
	NA’s	7894(33.59%)	1006(44.48%)

The employment pattern based on gender is quite obvious. As more mother are doing non-paid job or quitted their job and doing part-time job now, more father are still the breadwinner and most of them are doing full-time job.

Table 4 Control Variables: Parents' Educational Level

	Highest Level of Education	**momEdu**	**dadEdu**
1	Less than high school	1284(5.46%)	1205(5.13%)
2	High school diploma or GED	6341(26.98%)	5413(23.03%)
3	Associate's degree	2519(10.72%)	1493(6.35%)
4	Bachelor's degree	3688(15.69%)	2911(12.39%)
5	Master's degree	1366(5.81%)	1211(5.20%)
7	Ph.D/M.D/Law/other high level prof degree	430(1.83%)	820(3.49%)
	NA’s	7875(33.51%)	10440(44.42%)

Men are more likely to have Ph.D or equivalent high level of professional degree, but women have more proportion of masters’ and bachelor’s degree holder. Meanwhile, they also have lower degree holders. These might because of the large missing data.

2. Dependent Variable (categorical variable)

Table 5 Dependent Variable: Children's Future Income

	Incomecat	Count (%)
1	no income	2125 (9.04%)
2	1,000 dollars or less	1465 (6.23%)
3	1,001 to 2,500 dollars	1612 (6.86%)
4	2,501 to 5,000 dollars	2488 (10.59%)
5	5,001 to 10,000 dollars	3335 (14.19%)
6	10,001 to 15,000 dollars	2314 (9.85%)
7	15,001 to 20,000 dollars	1508 (6.42%)
8	20,001 to 25,000 dollars	1058 (4.50%)
9	25,001 to 30,000 dollars	599 (2.55%)
10	30,001 to 35,000 dollars	315 (1.34%)
11	35,001 to 45,000 dollars	263 (1.12%)
12	45,001 to 55,000 dollars	143 (0.61%)
13	55,000 dollars and above	110 (0.47%)
	NA’s	6168 (26.24%)

Because by the time the income data was collected, those 9th graders are just about to graduate from university (if they attend ones), so this income data is skewed towards the lower end. Besides, there are many missing data, which makes the analysis more troublesome. All in all, this is not a perfect variable, but it is all we have so far. So I choose to bear with it and see what can be found from these data.

3. Independent Variables (continuous variables)

Table 6 Independent Variables

	Variable name	Mean	Standard deviation	NA's	Observations	Min	Max
dv	incomecat	4.85	2.57	6168	17335	1	13
id	noActivity	0.87	0.33	132	23371	0	1
id	hrActivity	2.48	1.45	1897	20598	1	6
id	momTalkPrb	0.59	0.49	2958	20545	0	1
id	dadTalkPrb	0.38	0.49	4281	19222	0	1
pid	p1hhTime	4.87	0.51	7731	15772	1	5
pid	p1hwOften	2.47	1.10	7792	15711	1	5
pid	p1ptoMtg	0.38	0.49	8011	15492	0	1
pid	p1ptConfer	0.57	0.50	8023	15480	0	1
pid	p1SchEvent	0.69	0.46	7986	15517	0	1
pid	p1FundRaise	0.53	0.50	7990	15513	0	1
pid	p1Volunteer	0.31	0.46	7984	15519	0	1

There are a lot of missing when parents answering their questionnaires (variables start with p1). As I mentioned above, this is not a perfect dataset, so we should still question it when get the results. As we can see from above table, most of the students have attended activities, and spend relatively long hours on it. They tend to talk with their mother instead of father. Most of the parents who response the questionnaire are always living with their children, and they from time to time help them with their homework. They might attend PTO less and do volunteer work less, but about half of them who answered these questions keep in touch with teachers often, the raise fund to schools and usually attend school events.

Initial Models

Table 7 Initial Models

	Model 1	Model 2	Model 3
noActivity	-.005 (.082)	-.108 (.081)	.162+ (.083)
hrActivity	.032+ (.018)	.043* (.018)	.030+ (.018)
momTalkPrb	-.259*** (.057)	-.257*** (.057)	-.158** (.058)
dadTalkPrb	.064 (.057)	.142* (.057)	.016 (.059)
p1ptConfer	-.057 (.052)	-.049 (.051)	-.008 (.051)
p1hwOften	-.016 (.023)	-.023 (.023)	-.013 (.023)
p1fundRaise	-.071 (.057)	-.005 (.056)	-.100+ (.057)
p1ptoMtg	-.236*** (.054)	-.230*** (.053)	-.126* (.054)
p1hhTime	-.200*** (.051)	-.176*** (.050)	-.157** (.052)
p1Volunteer	-.267*** (.059)	-.182*** (.059)	-.158** (.059)
p1schEvent	-.024 (.062)	-.023 (.062)	.013 (.062)
famIncome		-.109*** (.008)	-.048*** (.011)
sex			-.528*** (.049)
Control Variables			omit
Constant	6.177*** (0.265)	6.296*** (0.264)	5.884*** (.427)
Observations	10,520	10,520	10,034
R_2	.013	.029	.086
Adjusted R_2	.012	.028	.079
Residual Std. Error	2.471 (df = 10508)	2.451 (df = 10507)	2.371 (df = 9960)
F Statistic	12.249*** (df = 11; 10508)	26.037*** (df = 12; 10507)	12.819*** (df = 73; 9960)
Notes:	+ p *P < .05; **P < .01; ***P < .001		

See table 7, in model 1, I ran a simple multinomial linear regression with only my independent variables of interest, without any control variables. In model 2, I add my major control variable, also my competing hypothesis, family income (famIncome) into the model.

Afterwards, I added all the control variables I mentioned above into the model, getting results showing in model 3.

As we can see in model 3 results, even after controlling the effects of family income, sex, race and parents' employment status and so on, the effect of organized extracurriculum activities is still positive and statistically significant. For example, the coefficient for variable **noActivity** is 0.162 and it is significant at the 10% level, meaning that when students attend these activities, their income in the future will be 0.162 score (in the range from 1 to 13) higher than their counterparts who did not attend these activities.

This is also true for the variable **hrActivity**, which means how many hours students spent on organized activities. When students have 1 score higher in this variable (ranging from 1 to 6), their future income will increase 0.03, and this coefficient is significant at the 10% level, meaning there is vaguely a positive relationship between hours students devoted to organized activities and their future income.

However, all the other hypothesis of mine seems no hold in this model. Some of the variables indicating parents emphasis on education do not have a relationship with their children's later income, like how close children and their dad are (**dadTalkPrb**), parents helping children with their homework (**p1hwOften**), and parents' close relationship with schools and teachers (**p1ptConfer** and **p1schEvent**). What is worse is that some of the variables have a negative relationship with children's future income, indicating that parents' cooperation with institutions undermines the future achievement of their children's. For example, the more children prefer to talk with their mom about their trouble (**momTalkPrb**), or the more often parents participate in parent-teacher organization (**p1ptoMtg**), or the larger numbers of hours parents spend with their children (**p1hhTime**), or the more parents volunteer in class or raise

funds to school (**p1Volunteer** and **p1fundRaise**), the less their children will earn in the future. This is against what sociologists found in the fieldwork. I have to hold this conclusion to see if there is any mistake with the data or if I miss anything to control.

As for the competing hypothesis (that the richer one's family is, they can get a better outcome than their poor counterpart), the coefficient of family income (famIncome) indicates the opposite. The coefficient is -0.048, and is significant at the 1% level, meaning that I can trust there exists a negative relationship between family income and children's future income. This is also weird, although this out of the scope of this project, but this negative relationship is against what other researchers have found, and against almost all social scientists' notion -- that there is social class stratification. Again, I have to find out some flaw in the modeling and data cleaning.

This model is not sufficient not just because the two bizarre findings I mentioned above, I also want to inspect how family income interacting with children's activity attendance, as Lareau found, middle-class families are more likely to adopt this kind of cultivation. Therefore, I have to add the interaction term into my model.

Besides, the VIF test shows that the multicollinearity in this model are super high, especially for parents' race and children's race. So in the final model, I will remove parents' race from the control variables.

Table 8 Multicollinearity test

	GVIF	Df	GVIFDf))
hrActivity	1.141	1	1.068
momTalkPrb	1.516	1	1.231
dadTalkPrb	1.551	1	1.246
noActivity	1.153	1	1.074
p1ptConfer	1.136	1	1.066
p1hwOften	1.047	1	1.023
p1fundRaise	1.387	1	1.178
p1ptoMtg	1.223	1	1.106
p1hhTime	2.210	1	1.487
p1Volunteer	1.398	1	1.183
p1schEvent	1.373	1	1.172
famIncome	1.717	1	1.310
momEdu	1.789	1	1.337
as.factor(race)	141.002	7	1.424
sex	1.095	1	1.046
as.factor(locale)	1.120	3	1.019
momEmp	1.079	1	1.039
as.factor(momOcc2)	2.362	22	1.020
dadEmp	1.098	1	1.048
as.factor(dadOcc2)	2.189	22	1.018
as.factor(parPattern)	2.438	3	1.160
as.factor(momRace)	273.021	6	1.596
as.factor(dadRace)	173.729	6	1.537

Final Models

Table 9 Final Models

	Model 3	Model 4
noActivity	.162+ (.083)	.385** (.149)
hrActivity	.030+ (.018)	.031+ (.018)
momTalkPrb	-.158** (.058)	-.159** (.058)
dadTalkPrb	.016 (.059)	.010 (.059)
p1ptConfer	-.008 (.051)	-.009 (.051)
p1hwOften	-.013 (.023)	-.013 (.023)
p1fundRaise	-.100+ (.057)	-.099+ (.057)
p1ptoMtg	-.126* (.054)	-.125* (.052)
p1hhTime	-.157** (.052)	-.156** (.052)
p1Volunteer	-.158** (.059)	-.156** (.059)
p1schEvent	.013 (.062)	.009 (.062)
famIncome	-.048 (.011)	.011 (.034)
sex	-.528*** (.049)	-.525*** (.049)
Control Variables	omit	omit
noActivity : famIncome		-.062+ (.034)
Constant	5.884*** (.427)	5.684*** (.441)
Observations	10,034	10,034
R^2	.086	.086
Adjusted R^2	.079	.079
Residual Std. Error	2.371 (df = 9960)	2.371 (df = 9959)
F Statistic	12.819*** (df = 73; 9960)	12.693*** (df = 74; 9959)
Notes:		*P < .05; **P < .01; ***P < .001

This model is better is because the interaction term in the model 4 is significant and meaningful. It also increase the R-squared a little bit, although it is not substantive. I also run the

partial F-test to see if there is any statistic difference if I include the interaction term in my model, and the test result is significant at 10% level, indicating a minor difference between this model and the last one.

The relationship between activity attendance and attending hours still hold in this model. This can partially prove my hypothesis that this parenting practice -- letting children attend organized activities -- can help improve their income.

However, the result of this interaction is also against the common sense, as the coefficient for the interaction term between **noActivity** and **famIncome** is -0.062, indicating that if one child attend organized activity, the effect of his or her family income on their future income will be further decreased by 0.062. This is totally the opposite of what I assumed. By saying middle-class parents are more likely to adopt the concerted cultivation and middle-class children are more likely to succeed, I mean there is a synergy between the family income and the effect of activities, and therefore I was expecting the positive coefficients for both family income and their interaction term.

One minor good news is that the negative relationship between family income and children's future income is gone, the coefficient for famIncome is not statistically significant in the model 4, indicating that we do not trust there is a relationship between them.

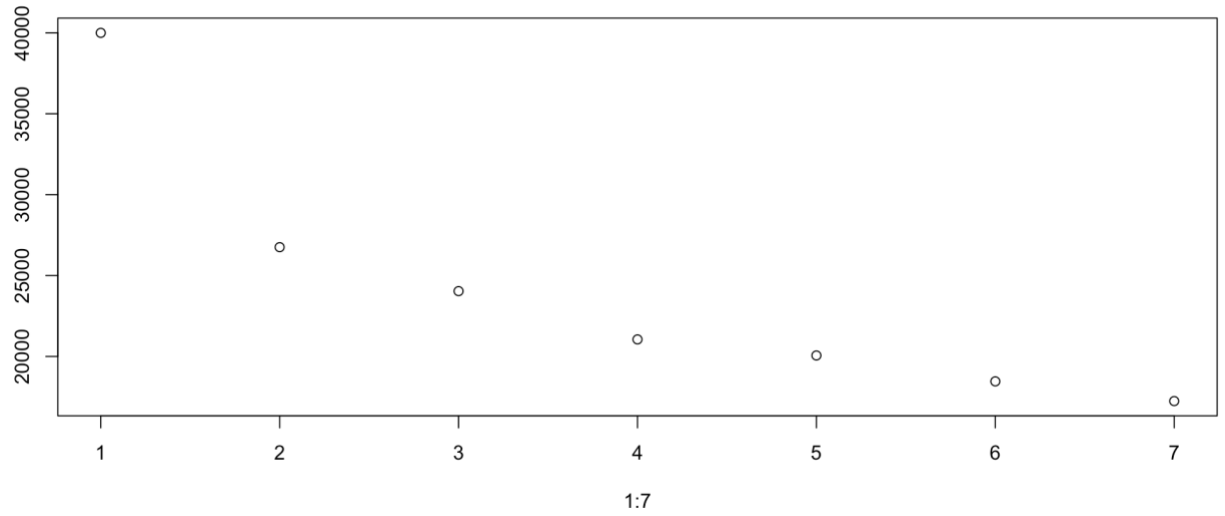
The VIF for other variables are less severe after I remove parents' race variables.

I then, also tried to use k-means to cluster parents styles. So that I can compare what type of parents can help their children earn more.

I categorize those variables regarding parents' behaviors towards their children and schools and teachers.

I ran the k-means and find out the proper number of cluster at first. (to make it work, I have to remove all of those who have missing data in any one of these variables.)

The total withinss changing with the numbers of the cluster is show below. In the light of this plot, I decided to cluster these observations into 4 groups based on parents' related variables (p1ptConfer, p1ptoMtg, p1hwOften, p1fundRaise, p1hhTime, p1Volunteer, p1schEvent).



The mean results based on the 4 clusters is shown below.

	P1PTCONFER	P1PTOMTG	P1HWOFTEN	P1FUNDRAISE	P1HHTIME	P1VOLUNTEER	P1SCHEVENT
1	0.833	0.703	2.652	0.884	1.076	0.669	0.953
2	0.487	0.279	1.428	0.468	1.124	0.207	0.676
3	0.728	0.483	4.345	0.601	1.116	0.355	0.746
4	0.297	0.130	2.874	0.180	1.196	0.027	0.350

As we can see, the first group are parents who devote a lot of time and money in school and classes, although not so many time with their children. Parents in the second group care less about the institute. The third group help their children do a lot of homework, and care about the school events too. The last group care the least about the school affairs but do accompany their children a lot.

I used these groups as independent variables and run the multinomial regression again. And below is the model result after controlling the basic control variables I used previously.

	INCOMECAT
hrActivity	.050** (.019)
famIncome	-.049*** (.011)
as.factor(parenttype)2	.164* (.064)
as.factor(parenttype)3	.077 (.086)
as.factor(parenttype)4	.280*** (.080)
sex	-.568*** (.052)
Control Variables	Omit
Constant	3.568***
	(.497)
Observations	8,240
R_2	.103
Adjusted R_2	.094
Residual Std. Error	2.330 (df = 8160)
F Statistic	11.832*** (df = 79; 8160)
Notes:	*P < .05; **P < .01; ***P < .001

As we can see from the coefficients for these types of parents, type 2 and type 4 have statistically significant and positive coefficients, meaning that compared with the reference type – type 1, who devote a lot of time and money in school and teachers, those who care more about their children (having higher score in helping children with their homework and living with their longer than others) are more likely to have children earning higher in the future.

In sum, we can roughly conclude that parents who care more about their children, spend more time with their children (doing homework and live with them) can help their children have more income than those parents who devote a lot time to maintain a great relationship with schools and teachers.

Conclusion

From this project, I learnt how to choose over variables. This can be simple when I just imagine the process and read others' papers, but when I tried to choose them from over

thousands of variables from one survey dataset, it can be grueling. Therefore, I found the importance of theories, with the guide of previous works, I can be less likely to be lost in the ocean of variables.

My hypothesis is only partly supported by my model, as there is a strong pattern of how attending organized activities help children earn more in the future. Also, accompanying children can help, if we consider the result from the k-means model. However, other hypotheses regarding parents' close relationship with schools and teachers did not indicate children's later achievement, some of them even have an inverse relationship against my hypothesis.

I also find variables indicating the "counter-school culture" inside one's family, like do you think high school is a waste of time, etc. I wish I could employ them in my research, but they are missing too many observations to get a good model, therefore I gave them up for now.

I still want to keep on working at using PCA to utilize more variables and get more clear type of parents (like informative ones, counter school culture ones, and rich ones etc.) and see how these types of parents affect their children's future income.

The inspiration that this project provides to my thesis is the positive relationship between the activity attendance and the future income, they are relatively strong even after controlling various variables.

The limitations of my project are obvious, and I have mentioned them a lot of times. Part of the results are against the general findings in this area of study. Besides, there are too many missing data in the dataset, I just removed them to run the model, but I am suspicious about the strength of my inference after I simply removing them.