Educational Progression of Two Generations of American Youths

A Probabilistic and Data-Driven Inference

Jimmy Kailun Cao

A research paper presented for the degree of
Master of Arts in Economics

Department of Economics
Concordia University
Canada

Acknowledgement

I thank Professor Prosper Dovonon for having accepted to be my supervisor and having given me an initial insight as to what I could work on. The adoption of the most important methodology – clustering – in the current paper is essentially derived by the initial idea that he had introduced to me. This paper was supposed to have two chapters, where the first chapter was a literature review comparing the performance of several estimators in panel data econometrics. It is regretful that some force majeure has made it impossible to finish in due course.

I also express my appreciation to Professor Christian Belzil, who has given his class an interesting project from which the current paper is extended. Lastly, I appreciate Professor Heejeong Kim for her support over semesters on the computational aspect. The book that she has lent me has been extremely useful and inspiring.

Abstract

This paper probes into education choices, in particular educational progression, of American adolescents in their early adulthood, notably by the age of 29. From 2 cohorts of National Longitudinal Survey of Youths (NLSY79 and NLSY97) carried out by the Bureau of Labour Statistics of the United States, we modelled their terminal educational attainment, which is categorised into 4 levels – 1. High School Dropout, 2. High School Graduate, 3. College Participation but not finished, and 4. 4-year College Graduate. This outcome variable is modelled with respect to a set of personal and familial demographics with a Multinomial-Logistic Type Conditionally-Binary Choice model design.

We began from assuming the existence of unobserved heterogeneity in the 2 samples. Therefore, our model allows it in the form of "intercept-terms" that account for unobserved factors within the samples. In addition, we only allow this unobserved heterogeneity to exist with a certain hidden group structure, which is also to be estimated by clustering, so that the curse of dimensionality is evaded. Due to the clustering procedure, our full model consists of a hyperparameter $\mathbf{G}$ standing for the number of clusters in the sample, which is to be determined. We calculate the model's prediction accuracy given a series of different values of G, and we found that $G = 3$ for cohort 1 and $G = 7$ for cohort 2 render the highest prediction accuracy.

Finally, it is found that the demographic variables considered have mixed effects on people's propensity to eventually achieve one of the four predefined levels. In particular, some are consistent with, while some are contradictory previous literatures. Among the significant variables, aptitude scores, parental educational background, gender agree with previous works and have diminished across cohorts. However, ethnicity surprisingly exhibits opposite signs from that of numerous literatures in both cohorts and all 4 levels.

Contents

1 Introduction

Schooling choice, as in the decision to progress into receiving higher levels of education, is one of the most important decisions in one's life. There are different reasons as to why some would choose to drop out from high school and why some would choose to enter college. The main goal of this research is to probe into the effects of some demographic features, such as natural aptitudes, family income and ethnicity, of American adolescents on their choices in education by the age of 29.

From existing research work, the connection between family income and educational attainment has been established and compared internationally. For instance, with evidence from the United Kingdom (see, for example, Rowntree, 1901; Glennerster, 1995; Blanden and Gregg, 2004), children coming from less prosperous family backgrounds tend to have worse educational attainment than their affluent counterparts. Based on British data, in addition, some evidence has been revealed that there exists an income effect independent from other demographic aspects of these British children (Gregg and Machin, 2000; and Hobcraft, 1998). Prais (1995) has shown that school leavers are less likely to progress to higher educational levels beyond the compulsory education. Data collected from the United States, on the other hand, suggest that the effects of identified family income on youngsters' terminal educational attainment from one research seem to largely differ from one another (for example Levy and Duncan, 2000; and Clark-Kaufman et al., 2003), but most research based on US data reveal positive effects of parental income on the number of years of education received (for example Taubman, 1989). With US

data, Brooks-Gunn et al. (1997) have also shown that connections exist between parental socioeconomic features, such as income and parents' education level, and children's developmental outcome, including educational achievements and attainment.

Apart from family income, ethnicity also plays a crucial role as a determinant in educational attainment. For instance, Orfield (1986) and Velez (1989) have illustrated that Hispanic students are more likely to drop out from school, and are less likely to attain higher level of education than Asians and Caucasians. In our studies, Asian and Caucasian (non-Hispanic white) are exactly merged as one category – non-black and non-Hispanic (hereafter the base category) – in the race variable, since the goal is to reveal the ethnic effects of African- and Hispanic-Americans.

Intriguingly, there are literatures mapping "parental perception" (or "parental beliefs") to their children's educational attainment, where ethnicity is often taken as a proxy of parents' beliefs towards upward mobility (Spera et al., 2009). This can be true especially when one considers the reality of a country like the US, where immigration is a norm to its population. For instance, 1st generation immigrants from East-Asia tend to believe that a college degree is a key, if not the key, to career success. Children brought up in this sort of family setting would naturally be more prone to enter and complete college. In this specific regard of parental perception, Borger et al. (1985) have established that it is a crucial factor driving parents' decision on their children's education. For black Americans, due to the history of slavery as well as the continuous voluntary immigration from Africa and the Caribbean, the situation is slightly more perplexed. Studies show that different identities within the "black" communities lead to different results in schooling and education (see Waters, 1994 and 1996). In particular, teenagers who self-identify as Black Americans see lower returns to their efforts, tend to perform worse in school and are therefore less likely to progress in education (Waters, 1999; and Vickerman, 1999).

Next, parental educational background, specifically mother's education in our study, is also an essential determinant in children's educational outcome established by, for

instance, Kane (1994); Keana and Wolpin (1997); Eckstein and Wolpin (1999); and Cameron and Heckman (1998, 2001). Bock and Moore (1984) have pointed out that mother's education has the largest effect on children's academic skills reflected by the ASVAB (Armed Services Vocational Aptitude Battery) tests, among all familial effects. They concluded from NLSY data that the positive impacts of mother's education are consistent with findings of previous literatures (such as Williams, 1979). As is the case for parental educational background, Belzil and Hansen (2003) have also reported that youngsters raised with both parents attain higher schooling. In addition, suggested by Maccoby and Jacklin (1974), females are generally superior in languages, resulting in more complex learning environments for children as mothers' educational attainment increases. In a nutshell, the more highly educated mothers' they are, the more they encourage their children academic excellence by fostering more complexed learning environments for their offspring.

Our results are consistent with these previous research work in the aspect of family income and parental educational background. Thanks to the design of the ASVAB tests aiming to reflect people's aptitudes in verbal, quantitative and practical skills, the ASVAB scores serve as a proxy measure of adolescents' innate aptitudes in this paper. We have shown that mother's education, both on its own and as a factor driving ASVAB, positively impacts one's propensity to progress in education. As for ethnic effects, contrary to Orfield et al. (2004), our estimation results indicate that the mean probability of high school graduation for African[1] (black)-Americans is respectively 0.12 and 0.1 lower than that of "non-Black and non-Hispanic" Americans. This is perhaps the most important finding in our paper.

[1]In the current literature review section, African-American is the conventional courteous manner to address Americans whose ancestors came from Africa, regardless of the reason why they had migrated to what it is today the United States of America. For our studies, however and hereafter, we use the term "African-American" only to refer to those who identify themselves as "Black Americans".

2 Methodology

In terms of methodology and technicality, an absolute majority of these aforementioned papers model the relationship between family income and educational attainment using the actual length of time they have spent in school, be it primary school, secondary school or university. That is to say, the models that they have estimated and drawn conclusions on are typically linear. However, quantifying educational choices by the actual length of education received does not lead to a straight forward interpretation, since translation is required to interpret the meanings behind the numbers. Another problem associated with linear models in this regard is that the length of schooling should be bounded above by, for instance, 22[2]. Linear models may also predict the years of schooling beyond this limit.

Therefore, in this paper, we transform this number into an ordinal categorical variable, with the 4 categories being the education levels, which serves as the outcome variable in our model, such that the model addresses the following questions: Does one proceed from one level to the next, and where does one arrive by the age of 29? The model perceives each possible transition from one level to the next as a binary choice of the individuals. Unlike Bradley and Taylor (2004) who have estimated a standard (simultaneous) multinomial choice model that explains British people's choices in different career possibilities, such as A-Level, vocational training, publicly funded training programme and permanent employment after the GCSE examination, our model imposes that people do not face all levels of educational simultaneously. That said, they face the choice of entering the next level if and only if they have completed the one before. More importantly, this is the one and only one choice that they face upon accomplishment of one level. Such a way of establishing our model is very intuitive: high school students cannot enter university before they graduate, and that university students cannot graduate if they drop out from college. This type of model specification empowers one to study people's choices, as well as what makes them opt for the options they have chosen. The

[2]The number 22 comes from 6 years of primary school, 6 years of secondary school, 4 years of undergraduate studies, 2 years of master's studies and 4 years of doctoral studies.

model is thus conditionally-binary multinomial.

Unobserved heterogeneity plays a key role in our model because one never observes all characteristics from data. We assume that the unobserved factors are permanent, which essentially represent non-time varying unobservable characteristics of the individuals such as their taste of schooling, motivation and aspiration. Instead of allowing individual-specific heterogeneity, only group-specific heterogeneity is permitted in the model. As a result, the number of groups (G), as a hyperparameter, is required before estimating the model. In contrast to some previous work that postulate the number of types of individuals or apply nonparametric method to "integrate out" the unobserved heterogeneity (such as Belzil and Leonardi (2009), Belzil and Hansen (2002), and Meghir and Rivkin (2010)), we have adopted 2 different data-driven and probabilistic approaches: **Gap Statistic** and **Grid Search Method**, and offer a comparison of the two methods in terms of statistical properties and explanatory power. First, Gap Statistic is adopted to estimate G. Tibshirani et al. (2001) has shown that the Gap Statistic outperforms other heuristics that serve the same goal proposed by, for instance, Kaufman and Rousseeuw (1990); Calinski and Harabasz (1974); Krzanowski and Lai (1988); and Hartigan (1975). We let the data express their group structures if they exist. Indeed, this method has its drawbacks that it might work incorrectly in some cases (see Yan and Ye, 2007; Yin et al., 2008; Dudoit and Fridlyand, 2002; and Mohajer et al., 2010), but it appears from our studies that our application of Gap Statistic yields a reasonable terminal model performance. Next, given a series of different values of G, we obtain estimate the parameters, obtain the bootstrap p-values of the estimates, screen out redundant variables, and then examine the model performance in terms of prediction accuracy. Herein, the niche of our paper lies in the fact that, to our best knowledge, no existing work has been carried out with this method.

Having obtained G, the model with G intercept terms is estimated with Broyden's algorithm (Broyden, 1962), and marginal effects are calculated from their analytical expressions. Since the individuals' group memberships are estimated whenever $G > 1$, the standard maximum likelihood theory may be spurious, and thus an empirical bootstrap is

conducted to obtain the standard errors and p-values. Finally, we probe into the effects of ASVAB scores, family income, gender, race, family structure, and mother's education on youngsters' schooling progression, as well as whether or not these effects have significantly changed across the two generations.

3 The Model

In this section, all equations and methodologies are applicable to both cohorts. Due to the nature that one does not face all 4 options simultaneously, the problem cannot and shall not be modelled using any conventional multinomial choice models such as the multinomial and ordered logit model. However, we will see that our model is of logistic type, and each logistic probability stands for the probability of transitioning from one level to the next, conditioning on that they have accomplished the previous level. Unobserved heterogeneity of individuals are also taken into accounts. As mentioned in section 2, group-specific heterogeneity, instead of individual-specific heterogeneity is permitted in our model, since the individual-specific heterogeneity would make estimation infeasible.

3.1 A Natural but Non-parsimonious Model

To begin with, we define the state space $\Omega = \{1, 2, 3, 4\}$ such that

$$p_{jk} := Pr(\text{progress to state } k \mid \text{state } j \text{ had been attained})$$

for $j \in \Omega$ and $k = j + 1 \leq \max \Omega = 4$. $p_0 := 1 - p_{12}$ is the probability of stopping at the initial state, which every individual must have at least gone through. For instance, for an individual who has stopped studying since graduating from high school, his probability expression would be $p_{12} \cdot (1 - p_{23})$. For any individual i, given any current j, we are interested in the following latent utility model driving their decisions to progress to state

$k = j + 1$, which is parametrised as

$$U_{ijk} = \alpha_i + X_i^{'}\beta_{jk} + \eta_{ijk},$$

where β_{jk} is the vector of (slope) parameters for the independent variables X_i specific to this progression; α_i is an individual-specific fixed-effects (intercept) parameter that captures any unobserved heterogeneity of each individual; and η_{ijk} is a standard logistic error independently and identically distributed for all i, j, k. It is assumed that unobserved heterogeneity enters the model uniquely through the intercept term. Furthermore, the unobserved heterogeneity shall be considered as an innate endowment of ability to proceed to higher levels of education that does not vary over time. Jump transitions, such as progressing from level 2 to 4 without going through 3, are strictly forbidden. This utility model evaluates to the following conditional probability:

$$p_{ijk} = \frac{exp(\alpha_i + X_i^{'}\beta_{jk})}{1 + exp(\alpha_i + X_i^{'}\beta_{jk})},$$

whereas if one did not progress, his conditional likelihood would become

$$1 - p_{ijk} = \frac{1}{1 + exp(\alpha_i + X_i^{'}\beta_{jk})}.$$

It follows that individuals' likelihood is given by

$$L_i = (1 - p_{i12})^{D_{i1}}(p_{i12} \cdot (1 - p_{i23}))^{D_{i2}}(p_{i12} \cdot p_{i23} \cdot (1 - p_{i34}))^{D_{i3}}(p_{i12} \cdot p_{i23} \cdot p_{i34})^{D_{i4}}, \quad (1)$$

and hence,

$$L = \prod_{i=1}^{n} L_i \quad (2)$$

is the sample joint-likelihood, where n is the sample size. Estimation of this model is non-parsimonious because there are n intercepts (α_i) and K slope parameters, summing to $n + K$ parameters, while there are only n observations.

3.2 A Feasible Modification to the Non-parsimonious Model

As a consequence, we thus require that the assumption that individuals have their unique unobserved ability factors be tightened. That is, we ought to impose that observations are latently heterogeneous in groups. It shares merely the same idea as the grouped fixed-effect (GFE) in say, Bester and Hansen (2016) and Bonhomme and Manresa (2015), except that our model is now purely cross-sectional. Clustering analysis is to be carried out to estimate the most reasonable number of groups G that exists in the samples, in order to reduce the number of model parameters and form the feasible model. Let C_g denote the g^{th} cluster. Given $i \in C_g$, it is

$$U_{ijk}^g = \alpha_g + X_i^{'}\beta_{jk} + \eta_{ijk}$$

and

$$p_{ijk}^g = \frac{exp(\alpha_g + X_i^{'}\beta_{jk})}{1 + exp(\alpha_g + X_i^{'}\beta_{jk})}.$$

Hence, the individuals' likelihood function becomes

$$L_i^g = (1 - p_{i12}^g)^{D_{i1}}(p_{i12}^g \cdot (1 - p_{i23}^g))^{D_{i2}}(p_{i12}^g \cdot p_{i23}^g \cdot (1 - p_{i34}^g))^{D_{i3}}(p_{i12}^g \cdot p_{i23}^g \cdot p_{i34}^g)^{D_{i4}} \quad (3)$$

and the joint log-likelihood is

$$l(\theta) = \sum_{g=1}^{G} \sum_{i \in C_g} ln(L_i^g) \quad (4)$$

where $\theta = (\beta_{12}^{'}, \beta_{23}^{'}, \beta_{34}^{'}, \alpha_1, ..., \alpha_G)^{'} \in \mathbb{R}^{48+G}$.

θ is estimated by the Maximum Likelihood (ML) method, i.e. $\hat{\theta}_{MLE} = argmax_{\theta}\, l(\theta)$. In our case, since estimation of group membership of the observations (i.e. clustering) is required to estimate this version of the model, there is no guarantee that the asymptotic distribution of $\hat{\theta}_{MLE}$ would be the one under the standard maximum likelihood theory (see Wilks, 1962). As a consequence, an empirical bootstrap re-sampling scheme will be carried out to obtain the empirical distribution of the ML estimator. All test-statistics and associated hypothesis testing procedures in the latter sections will be performed based on the bootstrap whenever membership estimation is required.

4 The Data

In this paper, analyses are carried out on the 1979 and 1997 cohort of the National Longitudinal Survey of Youth (NLSY79 and NLSY97). NLSY79 is a nationally representative sample of 12686 Americans who were between 14 and 21 year-old as of 1^{st} January 1979. NLSY97, on the other hand, covers approximately 9000 American who were 12 to 16 year-old as of 31^{st} December, 1996. Re-interviews were carried out every subsequent year until, respectively, 1994 (U.S. Department of Defense, 1982; Bock and Moore, 1986).

For our purpose, the longitudinal dimension of the samples is not considered since it focuses on the effects of non-time-varying attributes to educational progression, such as a set of aptitude test scores (the Armed Services Vocational Aptitude Battery, a.k.a. ASVAB), gender, family income (income), mother's education (med), race, whether or not they are from nuclear families (nuclear) and whether or not they live in urban areas (urban). The raw samples that we utilise are subsamples of non-immigrant youngsters' who have no siblings and have participated in the ASVAB tests due to the invitation of the U.S. Department of Defense. For ease of reading and consistency, the sample from 1979 cohort is named **D1**; and that of 1997 is named **D2** hereafter.

4.1 ASVAB – Proxies of Innate Aptitude/Intelligence

The motivation of considering ASVAB scores is two-folded. Firstly, it is the intelligence test with the most valid responses, among other tests such as SAT and ACT since NLSY79 and NLSY97 subjects were invited to complete the test with renumeration for updating the database of the U.S. Department of Defense and Military Services in regards of the aptitudes of newly recruited personnel in the U.S. armed forces (U.S. Department of Defense, 1980). Secondly, the ASVAB itself is designed for reflecting examinees' ability in the realms covered by its subtests, herein it could serve as a reliable proxy measuring one's intelligence in multiple aspects, and is capable to represent youth aptitudes nation-wise.

In the era of sampling the 2 cohorts, ASVAB had 9 components (subtests), which are General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Coding Speed (CS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), Numerical Operations (NO), Electronic Information (EI), Auto Information (AI), Shop Information (SI) and Assembling Objects (AO). Due to the major revision in ASVAB grading scale by the time that the 2nd cohort took the test, all ASVAB scores are standardised in order to conduct a fair test. To achieve a more robust estimation, all continuous variables have been standardised. On the other hand, in the 1979 cohort, AI and SI were administered in one component labelled AS, and they were only separated by the time when the 1997 cohort took them. AO has also been added to the test framework so that it is taken by only almost the entire 1997 cohort. Since there is no single manner to compare the effects of these problematic variables, AS, AI, SI and AO are dropped, and, as a result, GS, AR, WK, PC, NO, CS, MK, MC and EI are retained in both samples. For the purpose of our studies, emphasis is put on 2 major domains of intelligence, namely a). verbal skills and b). quantitative skills.

4.1.1 Verbal Skills

Of the 9 components, WK and PC are considered as verbal tests. WK measures the skills in identifying the correct words presented in context and to select the best synonyms of a given word; and PC measures one's ability to read and comprehend passages.

4.1.2 Quantitative Skills

Although AR and MK are considered quantitative tests, it is crucial to distinguish the difference between what they respectively measure. Despite their high correlation in both cohorts (0.8216 and 0.8032 respectively), they denote completely different notions. On one hand, AR gauges one's natural aptitude to resolve arithmetic word problems that does not require sophisticated learning processes and time investment; and on the other hand, the MK examines one's understanding in mathematics up to the high school level, which requires substantial training efforts and time.

4.2 Other Personal Demographics

4.2.1 Ethnicity

For ethnicity, due to a new category (Mixed, Non-Hispanic) added to the NLSY97, this category has been re-merged into "Non-Black / Non-Hispanic" as in NLSY79, yielding 3 resultant categories that require 2 dummy variables, namely "black", "hisp". We would like to clarify that "black" means that the interviewee self-recognises as "African-American", instead of any other possible scenarios where, for instance, one is of Caribbean descent. For detailed classification of "Black-American" self-identities, please refer to Rong and Brown (2001).

4.2.2 Family Income

Family income in our samples are annual, and are denominated in 10,000 real dollars with the base year being the first year of sampling for each cohort. Having taken into accounts of inflation, this variable is standardised in order to conduct a fair test across cohorts. In D1, this variable contains 2 missing values out of $n_1 = 2151$ observations. For this reason, observations with missing values in this variable are deleted.

4.2.3 Mother's Education

This variable captures the cumulative number of years of education that the subjects' mothers have received by the end of the last survey year of each NLSY cohort. It is treated as a continuous variable throughout our studies. In spite of the absence of missing values, it contains 8 zero-valued cases in D1, which are also deleted. We rule out the possibility that one's mother has never received any education.

4.2.4 Family Structure (Nuclearity)

By definition, Nuclear Family, as known as Conjugal Family definition, is a family group consisting of only two parents and their child or children (Encyclopædia Britannica, 2011). In our data, the indicator variable *nuclear* equals to 1 if a respondent comes from such type of family. 78.8% and 58.46% of respondents are reported to be from nuclear families respectively in cohort 1 and 2.

4.3 ACGRD – the Outcome Variable

The outcome variable of interest is the maximum years of education attained by 29 years old. This piece of information is contained in the set of ACGRD (indexed from 1 to 13) variables, which records the cumulative years of education received as of the time of the re-interviews (1^{st} May of each survey year) since they were 16 year-old. This is, in turn, the only longitudinal component of the effective sample we use. That said,

$ACGRD_1$ measures the number of years of education attained at the age of 16, $ACGRD_2$ measures that at the age of 17, and so on. It is recorded until the interviewees turned 29.

Table 1: Missing rate in ACGRD on D1 (%)

ACGRD	$1-6$	7	8	9	10	11	12	13
Total	0	2.978	5.212	7.678	21.685	46.812	70.451	92.834
Marginal	NA	2.978	2.234	2.466	14.007	25.128	23.639	22.383

Table 2: Missing rate in ACGRD on D2 (%)

ACGRD	$1-6$	7	8	9	10	11	12	13
Total	0	1.180	2.246	3.350	4.416	31.671	55.995	80.548
Marginal	NA	1.180	1.066	1.104	1.066	27.255	24.324	24.553

In principle, one would want to use the ACGRD variable with the highest index (13 in our samples) as it is the supposed to contain the most comprehensive information of their educational attainment. The higher the index we refer to, the more completely we would be able to extract information. However, some preliminary assessments indicate that the amount of missing values in ACGRD can be enormous. The marginal rate of missing data is the lowest when we refer to $ACGRD_9$ in D1, the highest grade attained at the 9th year of surveying the interviewees, and to $ACGRD_{10}$ in D2 (see table 1 and 2).

Since the purpose of this paper is to study the educational choices in terms of progression, we are interested in modelling the probability that one progresses from one level of academic attainment to the next. Given by the values of $ACGRD_9$ in D1 and $ACGRD_{10}$ in D2, 4 states of achievements have been defined and are labelled as follows:

State	Definition	Criteria
1	High school drop-out	$ACGRD \leq 11$
2	High school graduates	$ACGRD = 12$
3	College participants	$13 \leq ACGRD \leq 15$
4	4-year college graduates	$ACGRD \geq 16$

And since the outcome variable of interest is people's educational attainment by the age of 29, contingencies where people suspend and return to school for any reason are not

taken into accounts. The details on the outcome variable are displayed in table 3, from which it seems that the distribution of the 4 states is moderately even for both cohorts; and that the effective sample sizes turn out to be $n_1 = 1976$ and $n_2 = 2511$.

Table 3: State distribution in both cohorts (%)

Cohort/ State	1	2	3	4
1	15.38462	39.32186	23.02632	22.26721
2	19.11589	22.54082	27.16049	31.18280

5 Estimation

5.1 Step 1: Cluster Analysis

5.1.1 Objectives

Clustering is carried out based on all 17 attributes of each subject, including the outcome variable. The motivation of clustering is to group similar observations, such that homogeneity within groups can be reasonably assumed. Due to the presence of categorical variables, the conventional K-means algorithm shall not be adopted since it creates of artificial centroids that are "means" but not necessarily actual data points (see the actual algorithm in MacQueen, 1967; Lloyd, 1982 and Forgy, 1965). This is non-sensical when there are non-continuous variables. As a result, we adopt the K-medoid (a.k.a. Partitioning Around Medoid (PAM)) algorithm (Kaufman and Rousseeuw, 1987), a more robust clustering approach which allows for non-continuity in data and more generalised distance metrics when it searches for the optimal partition. It also preserves actual data points as the centroid (exemplar) of clusters (Kaufman and Rousseeuw, 1990). Contrary to K-means that uses Euclidean distance, our cluster analysis is implemented by transforming data into dissimilarity matrices in Gower distance (Gower, 1971; Kaufman and Rousseeuw, 1990).

5.1.2 Gower Distance

Proposed by Gower (1971) and generalised by Kaufman and Rousseeuw (1990), Gower distance is a measure of dissimilarity between data points. The niche of this distance metric lies in the existence of mixed-type variables in data. Concerning the data types, they are distinguished by their nature. The terminal dissimilarity between the i^{th} and j^{th} data point is given by

$$d(i,j) := \frac{\sum_k \delta_{ijk} \cdot d_{ijk} \cdot w_k}{\sum_k \delta_{ijk} \cdot w_k},$$

where w_k is a weight assigned to the k^{th} variable (equals to 1 by default); d_{ijk} stands for the distance between the i^{th} and j^{th} data point for the k^{th} variable; and $\delta_{ijk} = 1_{(x_{ik}=x_{jk}=0||(x_{ik}\neq x_{jk}))}$. As for d_{ijk},

1. For binary variables, such as *male* and *urban*, $d_{ijk} = 1_{(x_{ik}=x_{jk})}$;
2. for factors, such as race, $d_{ijk} = 1_{(x_{ik}=x_{jk})}$;
3. for numeric (continuous) variables, $d_{ijk} := \frac{|x_{ik}-x_{jk}|}{R_k}$, where R_k is the range in the k^{th} variable; and
4. for ordinal categorical variables, $d_{ijk} := \frac{|x_{ik}-1|}{max(x_{ik}-1)}$, where x_{ik} would be the factor level of the i^{th} observation.

For our samples, all ASVAB scores, *med* and *income* are considered numeric variables, whereas *male*, *nuclear*, *urban*, *black* and *hisp* are considered binary.

5.2 Step 2: Maximum Likelihood Estimation (MLE)

Owing to the quasi-logistic model specification, the Broyden's method (Broyden, 1965), a quasi-Newton method, has been proven to be numerically stable and fast-convergent. It is therefore adopted to estimate our feasible model. The exact algorithm is given in the Appendix 10.1. The initial guess sets all parameters to 0. The algorithm iterates until convergence, and each iteration has a time complexity of $O((48+G)^2)$.

At this step, given $G = g$, the algorithm is as follows:

1. estimate our model in expression (4) with MLE;
2. carry out an empirical bootstrap to obtain the standard errors (s.e.) and the p-values of estimates;
3. screen out redundant variables according to the p-values; and
4. make prediction and record its accuracy

The bootstrap is necessary for any $G > 1$ since, as mentioned in section 2.2, group membership estimation is required for this case, and it may lead to noise in the ML estimation procedure, and the standard asymptotic theory may not necessarily hold. For all bootstrapping procedures in this paper, the bootstrap size is $B = 1000$. It gives us an idea of the empirical distribution of the ML estimator.

6 Model Selection

The clustering problem remains to estimate the number of groups G that exists in the sample. This is what is meant by model selection - selecting the right value of G. To determine this number, we focus on 2 criteria:

1. the Gap statistic (Tibshirani et al., 2001); and
2. Grid Search method for G that renders the maximum prediction accuracy.

6.1 Method 1: Gap Statistic

Proposed by Tibshirani et al. (2001), Gap statistic formalises heuristics like the Elbow Method and Average Silhouette Score; and has been illustrated to perform well in terms of robustness (Tibshirani et al., 2001; Yan and Ye, 2007). We realise that the quality and performance of cluster analyses is sensitive to the scale of any dimension in the data. Non-uniform scales across variables could lead to inappropriate conclusions as

variables with larger scale could dominate how clusters are defined, hence all continuous variables, including med, are standardised. By observing the behaviours of clusters over $G = 1, 2, \ldots, 20$ and setting $B = 500$, a sufficiently large Monte Carlo sample (of size B) has been collected from a Uniform distribution over a "box aligned with the first principal component (PC) of the data" for each G. This particular data-specific distribution is often referred to as the null or reference distribution when computing Gap statistic, whose importance had been illustrated by Gordon (1996). Nevertheless, starting from $G = 1$, one additional cluster is formed at a time to calculate

$$Gap_B := E_B[\log(W_B^*)] - \log(W_g),$$

where $W_g := \sum_{r=1}^{g} \frac{D_r}{2n_r}$; $D_r := \sum_{i,j \in C_g} d_{ij}$; d_{ij} is the Gower distance; and $E_B[\cdot]$ is the estimated sample first-moment with respect to the null distribution over the B simulated observations. Suggested by Tibshirani et al., the optimal G is given by

$$\hat{G} := min[G : Gap_B(G) \geq Gap_B(G+1) - s_{g+1}],$$

where s_{g+1} is the Monte Carlo standard deviation of $E_B[\log(W_B^*)]$ corrected for simulation errors.

The vertical blue dotted line (figure 1) is the optimal number of groups in the samples suggested by this method. It follows that **Case 1**:

$$G_1 = G_2 = 1,$$

which is somehow trivial since it suggests that there is no obvious heterogeneity that exists in clusters in both cohorts.

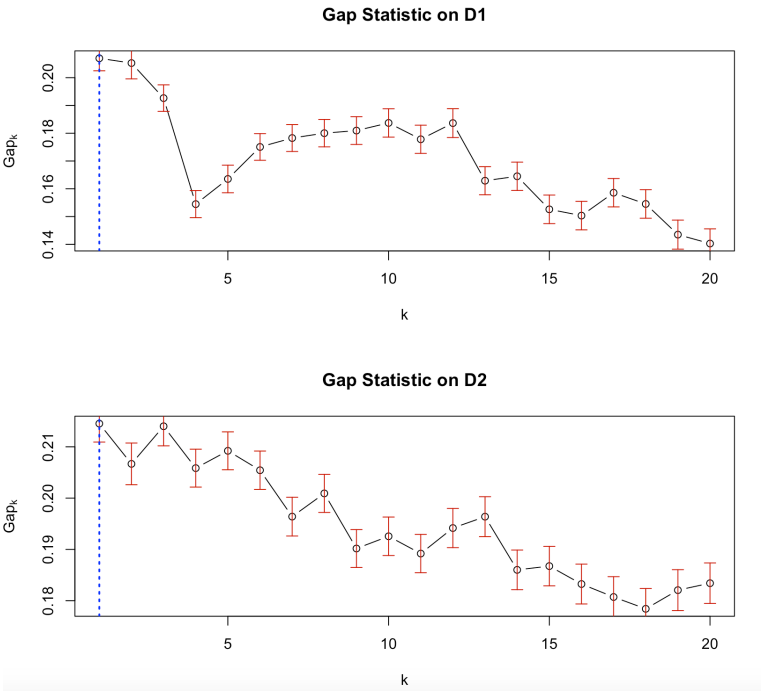

Figure 1: Gap statistic plot on D1 and D2

6.2 Method 2: Grid Search Method

As mentioned in the previous subsection, the Gap statistic has provided us some data-driven insight as to how clusters or group structures should appear. Next, we are off to look up the value of G that renders the maximum prediction accuracy on both cohorts. Notice that this perspective is reasonable because the goal of this project is not to build a universally applicable machine learning model that is to be applied to predict one's educational outcome prior to a certain age. Instead, it aims to develop an explanatory model to explain the phenomena observed from the data. Therefore, it suffices to select the model with the highest prediction accuracy within cohorts. Here, we essentially reiterate the algorithm presented in section 5.2 over $G \in \{1, 2, ..., 20\}$.

Figure 2 reports the accuracy over the first 20 natural numbered G, where the blue dotted lines indicate the value of G at which it attains its maximum. Hence, we take **Case 2**:

$$G_1 = 2 \text{ and } G_2 = 7.$$

It is, however, worth noting that adding extra number of intercept terms does not signifi-

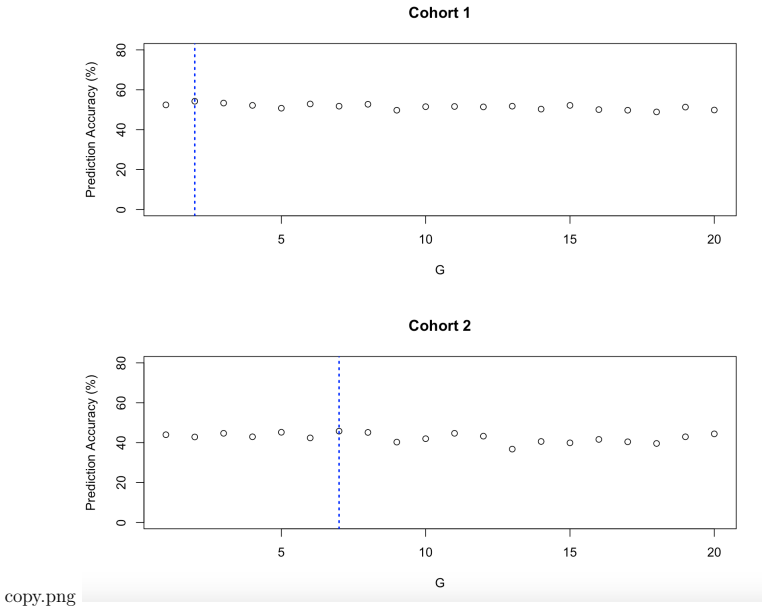

Figure 2: Prediction accuracy over G

cantly improve accuracy, as the curve is quite flat. Table 6 and 7 illustrate some summary statistics of the 3 and 7 clusters respectively on D1 and D2.

Table 4: Some summary statistics of clusters in D1

Cluster	Size	Avg. Dissimilarity	Diameter	Separation	Avg. Width
1	465	0.1568	0.5679	0.0402	0.1739
2	666	0.1225	0.4486	0.0242	0.3444
3	845	0.1382	0.5901	0.0242	0.1715

Table 5: Some summary statistics of clusters in D2

Cluster	Size	Avg. Dissimilarity	Diameter	Separation	Avg. Width
1	444	0.0847	0.3855	0.0299	0.3426
2	375	0.0954	0.4064	0.0301	0.2309
3	406	0.1115	0.4479	0.0245	0.1760
4	220	0.1221	0.4843	0.0245	0.1639
5	386	0.1097	0.4413	0.0290	0.1215
6	383	0.0924	0.3286	0.0290	0.2669
7	297	0.0995	0.3938	0.0334	0.1821

6.3 Statistical Goodness-of-fit

Parameter estimates and model statistics are illustrated in table 13 – 16 in appendix 10.3. In these 4 tables, each parameter estimate occupies 3 rows, each of which respec-

tively represents the level of progression that the logistic probability expression stands for. That is, the 1^{st} row maps to p_{12}; 2^{nd} row maps to p_{23}; and 3^{rd} row maps to p_{34}. Throughout the entire paper, all statistical hypotheses are tested by the bootstrap method at the size $\alpha = 0.05$.

Notice that for both cohorts, the (bootstrap) Likelihood Ratio test is implemented by also re-estimating the models with only the appropriate number of intercepts depending on the case. That is, each LR statistic comes from the hypothesis test:

$$H_0 : \beta_{jk} = \mathbf{0} \; \forall j, \, k$$

$$H_1 : H_0 \text{ is false.}$$

6.3.1 Cohort 1

Table 6: Model Statistics on D1

Case	G	LR	p-value	McFadden's R^2
1	1	1769.389	0	0.3044
2	2	1503.488	0	0.2722

Extracted from table 13 and 15, table 6 illustrates some basic model statistics on cohort 1. From the perspective of log-likelihood improvement, case 1 outperforms case 2.

6.3.2 Cohort 2

Table 7: Model Statistics on D2

Case	G	LR	p-value	McFadden's R^2
1	1	1462.266	0	0.2028
2	7	1029.261	0	0.1534

Extracted from table 14 and 16, table 7 illustrates the basic model statistics on cohort 2. From the perspective of log-likelihood improvement, case 1 also outperforms case 2 on cohort 2.

6.3.3 Case-by-case Comparison

From the 2 sections above, it follows that, for both cohorts, both LR and pseudo-R^2 are uniformly higher in case 1 than in case 2. We would like to now examine the statistical significance of adding, respectively, 2 and 6 intercept terms to the model in case 1. In other words, for each cohort, we test

$$H_0 : G = 1$$

$$H_1 : G = G_j,$$

to decide if there is significant improvement in the model's log-likelihood by adding more intercepts terms.

Table 8: Bootstrap LR test for case-by-case comparisons

Cohort	G_j	LR	Bootstrap p-value
1	2	24.5024	0.019
2	7	66.9291	0.005

where G_j is the number of groups for cohort j in case 2. Table 8 indicates that H_0 is rejected for both cohorts. This means that configuring additional intercept terms indeed improves statistical goodness-of-fit. It follows that case 1 fits better to cohort 1, and case 2 is more preferable to cohort 2.

6.4 Model Accuracy

6.4.1 Marginal (Overall) Accuracy

Apart from model statistics, we shall also evaluate the intra-cohort performance of our models in the 2 cases in terms of the overall prediction accuracy, which are in turn tabulated in table 9 in the form of a marginal table[3]. All variables in the models are

[3]Marginal table is a notion in statistics that refers to a table that marginalises all cases in 1 dimension of categorical data to reduce that dimension in a K-way table. Having been marginalised, a (K-1)-way table is rendered. For all notions concerning categorical data analysis in statistics, see Agresti (2014).

retained if and only if they are individually significant at size $\alpha = 0.05$. Only statistically significant variables are retained for the purpose of prediction.

Table 9: Marginal model accuracy (%)

Cohort/Case	1	2	Differential (Gain)
1	52.4292	54.2510	1.8218
2	43.9665	45.7587	1.7922

Revealed by the marginal table (table 9), the model assuming there is no unobserved heterogeneity (case 1) performs slightly worse overall for both cohorts. This is, somehow, contradictory to the conclusion we have drawn from section 6.1. Adding group structure and the appropriate number of intercept terms to the model, the gains in marginal prediction accuracy is almost equal in both cohorts.

6.4.2 Conditional Accuracy

Acknowledging the Simpson's paradox (Simpson, 1951), the conditional tables (in the form of confusion matrices) are in table 10 and 11. We break down the prediction accuracy by conditioning it on the actual states to which individuals belong.

Table 10: Conditional model accuracy on D1 (Row %)

(a) Case 1

Actu/Pred	1	2	3	4
1	47.37	43.09	8.55	0.99
2	16.99	59.85	16.60	6.56
3	7.47	31.87	37.14	23.52
4	1.59	10.68	25.68	62.05

(b) Case 2

Actu/Pred	1	2	3	4
1	32.89	60.20	5.26	1.64
2	10.42	70.40	9.78	9.40
3	5.71	38.90	23.96	31.43
4	0.45	14.09	13.64	71.82

Cohort 1 When we visualise the intra-cohort case-by-case comparison, the contrast between the 2 cases on cohort 1 becomes more apparent. The principal diagonal of the tables are the correct prediction rates given each state as individuals' final educational attainment. Case 1 offers a less extreme accuracy distribution of prediction accuracy across all 4 states.

Table 11: Conditional model accuracy on D2 (Row %)

(a) Case 1

Actu/Pred	1	2	3	4
1	48.75	26.25	14.17	10.83
2	25.44	27.39	24.38	22.79
3	12.46	17.89	26.69	42.96
4	4.21	4.98	22.73	68.07

(b) Case 2

Actu/Pred	1	2	3	4
1	65.42	17.92	4.58	12.08
2	41.70	23.14	6.18	28.98
3	21.26	16.72	7.92	54.11
4	7.79	5.62	3.58	83.01

Cohort 2 This time, Case 1 yields a more even distribution on prediction accuracy. This is particularly obvious with an accuracy of 7.92% on state 3 in case 2.

7 Discussions

Goodness-of-fit Section 6 demonstrated that following the suggestion of the Gap statistic approach would lead to a model with better performance in terms of statistical goodness-of-fit (see table 6 and 7). This alone does not supply a direct comparison between the 2 cases in the cohorts, although the log-likelihood improvement and McFadden's R^2 are uniformly higher in case 1 on both cohorts. Table 8 subsequently suggests the contrast that at 95% level of confidence, we have sufficient evidence that case 1 is not true and should be rejected.

Prediction Accuracy From the perspective of in-sample explanatory power, the Grid search method has revealed similar results. On both cohorts, the prediction accuracy in case 2 is higher that in case 1, but to a rather limited extent. Note that the goal of the current paper is not building a generally-applicable supervised learning model that classifies individuals' attributes and features to the 4 states of educational attainment by the age of 29. Instead, it is to explain and compare the effects of personal attributes on adolescents' schooling attainment between 2 generations of Americans, by building a model that well fits our within-cohort data.

On the surface, although case 2 outperforms case 1 in overall accuracy, this superiority is subtle. By breaking it down into conditional prediction accuracy (illustrated in table 10 and 11), we have also established that case 1 yields a more even prediction accuracy across states. As a result, a trade-off comes up: would we exchange a more even conditional explanatory power for a rather subtle additional marginal accuracy? Having taken consideration of all the criteria presented in this paper, we therefore have chosen the latter.

Technicality Since Gower distance is restricted in $[0, 1]$ per se, it may lead to a lower sensitivity for the Gap statistic algorithm than if the usual Euclidean distance was used instead. Low sensitivity could have led to problematic results when using clustering anal-

ysis to exploit the unobservable group structures in the samples. It might be the reason why it suggests $G = 1$ on both samples. We can nonetheless observe from figure 1 that there are close runner-ups in the magnitude of Gap statistic to $G = 1$ for both cohorts ($G_1 = 2$ and $G_2 = 3$). However, these potential issues could (or would) have been resolved by parallelly selecting G by grid search.

As the current section unfolds our findings, we refer all estimation results to table 17 – 18 in appendix 10.3. Note that numbers presented in this section are averages, and that analyses assume that other factors be constant.

7.1 Effects of Innate Aptitudes

7.1.1 Verbal Skills

We observe that the effects of word knowledge (WK) reduced the probability that one drops out from high school by as much as 0.0288, and increased that of participating in college by 0.0563 on cohort 1. However, its effects seems to have faded out on cohort 2, since it no longer affects the probability of dropping out from high school and that its effects on college participation has dropped to merely 0.0337.

Regarding paragraph comprehension (PC), its marginal effect on the high school drop-out probability has changed from -0.0496 to -0.0431. This number has increased in the ordinal sense but it signals a slight diminution in magnitude. That is, its strength of reducing the probability of dropping out from high school has slightly decreased. In contrast, its effects on college graduation has increased from 0.0344 to 0.0405.

Hence, it is apparent that effects of verbal skills on high school graduation and college entrance have gone down over the 2 generations, whereas it somehow becomes a stronger factor driving one's propensity of college graduation.

7.1.2 Quantitative Skills

Arithmetic reasoning (AR) surges and becomes a determinant that increases the probability of graduating from college (by 0.034) on cohort 2, as it was believed to be completely irrelevant in determining one's educational attainment on cohort 1.

Mathematics knowledge (MK), contrarily, serves consistently as an influential factor on high school drop-out, college entrance and college graduation. On the probability of high school drop-out, its effects have subtly changed from -0.0525 to -0.0565, meaning that it drops more significantly on cohort 2 than on cohort 1 given the same increment in the MK scores. When considering the probability of not entering college after high school graduation, the figure changes from -0.0995 to -0.0307. It follows that even if it is still less likely for people with superior high school mathematics background not to progress into college, the effects of knowledge in mathematics on remaining in state 2 has decreased. Lastly, its marginal effect on college graduation has also diminished by almost 50% from 0.1295 to 0.0611.

Herein, it is believed that quantitative skills generally played a more crucial role governing their educational attainment on cohort 1 than on cohort 2.

7.2 Effects of Ethnicity

Since our ethnicity attribute has $K = 3$ categories, all comparisons are against the base category, which is "non-black and non-Hispanic (others, a.k.a. Caucasian or Asian)".

For self-identified "black" Americans, their probability of dropping out of high school is shown to be respectively 0.1205 and 0.1034 lower than their "non-black and non-Hispanic" counterparts. It is also found that "black" Americans have lower probability, by 0.0884 and 0.0465 respectively, to end up only with a high school diploma without entering college than "non-black and non-Hispanic" Americans. Lastly, being "black" appears to tremendously improve the tendency to enter college, as its marginal effects on

college entry probability are as large as 0.2211 and 0.1168. Based on this observation, we also suspect that the effects of being "black" have in fact decreased over the 2 generations. This contradicts previous research by, for instance, Waters (1999), Vickerman (1999) and Orfield et al. (2004).

As for the Hispanic community, Hispanic Americans in cohort 2 have exhibited a probability of high school drop-out 0.0659 lower than those of the base category. In cohort 1, their ethnicity had empowered them to be more likely to enter college after high school graduation than their "non-black and non-Hispanic" counterparts, whilst their ethnicity does not influence the same decisions in cohort 2. This result with respect to the Hispanic group is also contradictory to what Arias (1989), Orfield (1986) and Velez (1989) have discovered.

7.3 Effects of Family Income

Consistently, family income appears to alter only the probability of being in state 2 and 4, notably graduating from high school but not proceed to college and graduating from college, for both cohorts. Over the 2 cohorts, it is observed that a unit increase in family income leads to a 0.0305 and 0.0294 reduction in the probability of ending up with only a high school diploma; and we visualise an increase of 0.0411 and 0.0326 in that of graduating from college eventually. We therefore establish that there are significant marginal effects of family income in educational attainment (such as in Taubman, 1989; Levy and Duncan, 2000), but it has fallen over the 2 generations.

7.4 Effects of Mother's Education

The marginal effects of *med* on the high school drop-out probability becomes significant and negative (-0.0397) on cohort 2. In other words, per unit increase in *med*, the probability of dropping out high school reduces by 0.0397 on average. Next, the negativity in state 2 on both cohorts suggest that the higher educated the mothers are, the

less likely their children would stop their education after high school, even if its marginal effect has fallen over time. Probing into college entry and graduation, the marginal effects of *med* are reported to be positive, and that it has increased over time for the probability of college graduation.

Our results regarding mother's education generally coheres with with previous research findings of, say, Bock and More (1984) and Williams (1979). The higher educated the mothers are, the more likely their children would enter college and thus graduate.

7.5 Effects of Family Structure

Our results partly support what has been found in previous literatures such as Belzil and Hansen (2003); and Cameron and Heckman (1998, 2001), as the mean marginal effects of *nuclear* are mostly statistically positive. Yet, they could be negative in some scenarios.

It is reported that adolescents raised in nuclear families are less likely to drop out from high school, and are more likely to either stop schooling or enter college after high school, ceteris paribus, than those who are not. Its marginal effect on the high school dropout probability is steady at about -0.163 on both cohorts, proving that family structure, in particular nuclearity, significantly reduces the tendency of high school incompletion. A switch of sign in the marginal effects of *nuclear* on college graduation is also observed. For cohort 1, holding other factors constant, the probability that a person from a nuclear family graduates from college is 0.0447 **less** than someone from a non-nuclear family; while for cohort 2, this probability is expected to be 0.0723 higher. This negativity in the marginal effects of *nuclear* on college graduation in cohort 1 is essentially the part that disagrees with the aforementioned literatures. Meanwhile, the disagreement is rectified in cohort 2.

7.6 Effects of Gender

The gender effect has generally become more prominent over the generations when determining one's tendency to enter college and hence graduate. Over cohorts, we observe that males have become more likely to participate in college without graduating by 29 years old, but less likely to graduate and obtain a college degree than females. Males in cohort 2 have, on average, a probability 0.0682 higher than females to discontinue college studies, and a probability 0.0965 lower than females to finish and graduate from college. In cohort 1, however, gender was believed not to influence the 2 respective probabilities. Gender also plays a stable role deciding the likelihood that one remains in state 2. In particular, its marginal effects remains at about 0.062 in both cohorts, implying that males have larger tendency to stop schooling after high school than females, ceteris paribus.

8 Conclusion

We have reached the discussions above by estimating a multinomial-type conditionally binary choice model, in which unobserved heterogeneity is permitted in groups. The data, however, are later suggested to be homogeneous by Gap statistic. Although this reduces our model to one that assumes homogeneity, we have shown that this version of the model is superior to the one allowing heterogeneity. Hence, it is considered optimal and is retained for our studies.

The main reason why we have adopted cluster analysis is that we would like that unobserved heterogeneity be discretised, in the sense that "similar" individuals should behave sufficiently homogeneously. Instead of simply postulating the number of "types" of individuals in the samples, we first allow unobservable heterogeneous factors, such as the taste of schooling, to enter the model as an intercept term; and we thus let the data reflect how many such "types" (or "groups") of individuals there are by applying the Gap statistic. To the author's best knowledge, there has not been existing literatures estimat-

ing similar types of models with these data-scientific and data-driven approaches. It is thus the paper's niche to model people's educational attainment with the aforementioned methodologies, in order to supply different insights as to how observations are classified by "types".

Essentially, our results are partly consistent with numerous literatures in this domain, notably in the aspect of innate aptitudes, family income and parental educational background. Yet, some of them contradict previous findings, especially in the ethnic effects. While it is generally believed that black- and Hispanic-Americans youths tend to drop out from high school than their Caucasian and Asian counterparts, the reverse is reported by this paper. From the perspective of family structure, while it had been established that children raised in nuclear families tend to achieve higher in education, our findings in cohort 1 partially stand against it. Nevertheless, for the other demographic factors that were considered, despite the consistency with previous research in terms of their effects on educational attainment, some of them have diminished over the two generations of Americans. For instance, verbal and quantitative skills seem to have diminishing marginal effects; family structure has a generally increasing marginal effects in 3 of the 4 educational outcomes in the model; family income and ethnicity both exhibit diminishing marginal effects.

9 Future Work

9.1 Model Specifications

As mentioned before, the econometric objectives of the current paper require that the model be adequate to fit and predict observations sufficiently well within each sample. Indeed, the current paper adopts a relatively simplistic model which tremendously differs from the ones in previous literatures. Therefore, establishing more perplexed model specifications, for instance introducing structures, might further enhance the goodness-of-fit.

9.2 Clustering

The adoption of PAM algorithm has made G, the number of clusters, a hyperparameter of our main model. With other clustering algorithms, particularly Affinity Propagation (AP) (Frey and Dueck, 2007) that do not require this hyperparameter, there could be different insights as to how unobserved group structures exist in the samples. Nevertheless, AP does not form clusters with solely geometry (distance metrics). Even if Gap Statistic already outperforms other heuristics (Kaufman and Rousseeuw, 1990; Calinski and Harabasz, 1974; Krzanowski and Lai, 1988; Hartigan, 1975) in determining this hyperparameter (Tibshirani et al., 2001), it would be intriguing to apply AP to explain and establish clusters in the samples.

References

1. Agresti, A. (2014) "*Categorical Data Analysis*" (Hoboken: Wiley)

2. Bellman, R. (1961) "*Adaptive Control Process*" (Princeton University Press)

3. Belzil, C. and Hansen, J. (2001) "Estimating the Intergenerational Educational Correlation from a Dynamic Programming Model" *Centre interuniversitaire de recherche en analyse des organisations* (CIRANO) 2001s-20

4. Belzil, C. and Hansen, J. (2002) "Unobserved Ability and the Return to Schooling" *IZA Institute of Labor Economics Discussion Paper* No. 508

5. Belzil, C. and Hansen, J. (2003) "Structural Estimates of the Intergenerational Education Correlation" *Journal of Applied Econometrics* 18(6): 679-696

6. Belzil, C. and Leonardi, M. (2009) "Risk Aversion and Schooling Decisions" *HAL* 00411099

7. Bester, C. A. and Hansen, C. B. (2016) "Grouped Effects Estimators in Fixed Effects Models" *Journal of Econometrics* 190(7): 197–208

8. Blanden, J. and Gregg, P. (2004) "Family Income and Educational Attainment: A Review of Approaches and Evidence for Britain" *Oxford Review of Economic Policy* 20(2): 245–263

9. Bock, R. D. and Moore G. J. (1984) "*Profile of American Youth: Demographic Influences on ASVAB Test Performance*" (Washington DC: National Opinion Research Center, Office of the Assistance Secretary of Defense (Manpower, Installations and Logistics))

10. Bonhomme, S., Lamadon, T. and Manresa, E. (2017) "Discretizing Unobserved Heterogeneity" *University of Chicago, Becker Friedman Institute for Economics Working Paper* No. 2019–16

11. Bonhomme, S., Manresa, E. (2015) "Grouped Patterns of Heterogeneity in Panel

Data" *Econometrica* 83(3): 1147–1184

12. Borger, J. B., Lo, C., Oh, S. and Walberg, H. J. (1985) "Effective Schools: A Quantitative Synthesis of Constructs" *The Journal of Classroom Interaction* 20(2): 12–17

13. Bradley, S. and Taylor, J. (2004) "Ethnicity, Education Attainment and the Transition from School" *The Manchester School* 72(3): 317–346

14. Frey, B. J. and Dueck, D. (2007) "Clustering by passing messages between data points" *Science* 315(5814): 972–976

15. Brooks-Gunn, J., Duncan, G., and Aber, J. L. (Eds.) (1997) "*Neighborhood poverty II: Policy implications for studying neighborhoods*" (New York: Russell Sage)

16. Broyden, C. G. (1965) "A Class of Methods for Solving Nonlinear Simultaneous Equations" *Mathematics of Computation* 19(92): 577–593

17. Calinski, T. and Harabasz, J. (1974) "A Dendrite Method for Cluster Analysis" *Communications in Statistics – Theory and Methods* 3(1): 1–27

18. Cameron, S. and Heckman, J. (1998) "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males" *Journal of Political Economy* 106(2): 262-333

19. Cameron, S. and Heckman, J. (2001) "The Dynamics of Educational Attainments for Black, Hispanic and White Males" *Journal of Political Economy* 109(3): 455-499

20. Clark-Kauffman, E., Duncan, G. J., and Morris, P. (2003) "How Welfare Policies Affect Child and Adolescent Achievement" *American Economic Review* 93(2): 299–303

21. Colby, E. and Bair, E. (2013) "Cross-Validation for Nonlinear Mixed Effects Models" *Journal of Pharmacokinetics and Biopharmaceutics* 40(2): 243-252

22. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm" *Journal of the Royal Statistical Society*

Series B 39(1): 1–38

23. Dudoit, S. and Fridlyand, J. (2002) "A prediction-based resampling method for estimating the number of clusters in a dataset" *Genome Biology* 2002 3(7): 0036.1–0036.21

24. Eckstein, Z. and Wolpin, K. (1999) "Why Youth Drop Out of High School: The Impact of Preferences, Opportunities and Abilities" *Econometrica* 67(6): 1295-1339

25. Gordon, A. (1996) "Null Models in Cluster Validation" *From Data to Knowledge* (eds W. Gaul and P. Pfeifer), 32–44 (New York: Springer)

26. Gower, J. C. (1971) "A General Coefficient of Similarity And Some of Its Properties" *Biometrics* 27: 857–874

27. Glennerster, H. (1995) "*British Social Policy since 1945*" (Oxford: Blackwell)

28. Gregg, P. and Machin, S. (2000) "Child Development and Success or Failure in the Youth Labor Market" *Youth Employment and Joblessness in Advanced Countries* University of Chicago Press, 247–288

29. Hartigan, J. A. (1975) "*Clustering algorithms*" (New York: John Wiley & Sons)

30. Hobcrafe, J. (1998) "Intergenerational and Life-Course Transmission of Social Exclusion: Influences and Childhood Poverty, Family Disruption and Contact with the Police" *LSE STICIED Research Paper* No. CASE015

31. Jiménez, J. D. and Salas-Velasco. M. (2000) "Modelling Educational Choices. A Binomial Logit Model Applied to the Demand for Higher Education" *Higher Education* 40: 293–311

32. Kane, T. (1994) "College Entry by Blacks Since 1970: The Role of College Costs, Family Background, and the Returns to Education" *Journal of Political Economy* 102(5): 878-911

33. Kaufman, L. and Rousseeuw, P. J. (1990) "*Finding Groups in Data: An Introduction to Cluster Analysis*" (Hoboken, New Jersey: John Wiley & Sons)

34. Kaufman, L. and Rousseeuw, P. J. (1987) "*Clustering by Means of Medoids, in*

Statistical Data Analysis Based on the L1-Norm and Related Methods" (North-Holland, Dodge, Y.), 405–416

35. Keane, M. P. and Wolpin K. (1997) "The Career Decisions of Young Men" *Journal of Political Economy* 105(3): 473-522

36. Krazanowski, W. J. and Lai, Y. T. (1988) "A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering" *Biometrics* 44(1): 23–34

37. Levy, M. and Duncan, G. (2000) "Using Siblings Samples to Assess the Effect of Childhood Family Income on Completed Schooling" *JCPR Working Papers* 168, Northwestern University/University of Chicago Joint Center for Poverty Research

38. Liu, Y., Cai, J., Yin, J. and Fu, A. W. (2008) "Clustering Text Data Streams" *Journal of Computer Science and Technology* 23(1): 112–128

39. McCullagh, P., and Nelder, J. A. (1989) "*Generalized Linear Models*" (London: Chapman and Hall)

40. Meghir, C. and Rivkin, S. G. (2010) "Econometric Methods for Research in Education" *NBER Working Paper* No. 16003

41. Mohajer, M., Englmeier, K. and Schmind, V. J. (2010) "A Comparison of Gap Statistic Definitions with and without Logarithm Function" *Technical Report Number* 096, 2010, University of Munich

42. Maccoby, E. E. & Jacklin, C. N. (1974) "Myth, reality and shades of gray: What we know and don't know about sex differences" *Psychology Today* 8(7): 109-112

43. Orfield, G. (1986) "Hispanic Education: Challenges, Research, and Policies" *The American Journal of Education* 95(1): 1-25

44. Orfield, G., Losen, D., Wald, J. and Swanson, C. B. (2004) "*Losing Our Future: How Minority Youth are Being Left Behind by the Graduation Rate Crisis*" (Cambridge, MA: The Civil Rights Project at Harvard University) Contributors: Advocates for Children of New York, The Civil Society Institute

45. Pelleg, D. and Moore, A. W. (2000) "X-means: Extending K-means with Efficient Estimation of the Number of Clusters" *Proceedings of the Seventeenth International Conference on Machine Learning* (ICML 2000)

46. Prais, S. J. (1995) "*Productivity, Education and Training: Facts and Policies in International Perspective*" Cambridge University Press

47. Rong, X. L. and Frank, B. (2001) "The Effects of Immigrant Generation and Ethnicity on Educational Attainment among Young African and Caribbean Blacks in the United States" *Harvard Educational Review* 71(3): 536-565

48. Rowntree, S. B. (1901) "Poverty: A Study of Town Life" *International Journal of Ethics* 13(1): 129-130

49. Rubin, D. B. (1976) "Inference and Missing Data" *Biometrika* 63(3): 581-592

50. Simpson, E. H. (1951) "The Interpretation of Interaction in Contingency Tables" *Journal of the Royal Statistical Society* Series B 13(2): 238–241

51. Spera, C., Wentzel, K. R., and Matto, H. C. (2008) "Parental Aspirations for Their Children's Educational Attainment: Relations to Ethnicity, Parental Education, Children's Academic Performance, and Parental Perceptions of School Climate" *Journal of Youth and Adolescence* 38(8): 1140–1152

52. Taubman, P. (1989) "Role of Parental Income in Educational Attainment" *The American Economic Review* 79(2): 57–61

53. The Editors of Encyclopædia Britannica (2015) "Nuclear family" In *Encyclopædia Brittanic* (Chicago, IL: Encyclopædia Britannica, Inc.) Retrieved March 11, 2020, from https://www.britannica.com/topic/nuclear-family

54. Tibshirani, R., Walther, G. and Hastie, T. (2001) "Estimating the number of clusters in a data set via the gap statistic" *Journal of the Royal Statistics Society* Series B 63(2): 411–423

55. U.S. Department of Defense (1982) "*Profile of American Youth: 1980 Nationwide*

Administration of the Armed Services Vocational Aptitude Battery" (Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics))

56. Velez, W. (1989) "High School Attrition Among Hispanic and Non-Hispanic White Youths" *Sociology of Education* 62(2): 119-133

57. Vickerman, M. (1999) "*Crosscurrents: West Indian Immigrants and Race*" (New York: Oxford University Press)

58. Waters, M. C. (1994) "Ethnic and Racial Identities of Second-Generation Black Immigrants in New York City" *International Migration Review* 28(4): 795

59. Waters, M. C. (1996) "The intersection of gender, race and ethnicity in identity development of Caribbean American teens" In Leadbeater, B. and Way, N. (Eds.), Urban girls: Resisting stereotypes, creating identities 65–81 (New York: New York University Press)

60. Waters, M. C. (1999) "Black Identities: West Indian Immigrant Dreams and American Realities" In Grusky, D. B. (Eds.), *Social Stratification: Class, Race, and Gender in Sociological Perspective* (New York: Russell Sage Foundation) 608–610

61. Wilks, S. (1938) "The large-sample distribution of the likelihood ratio for testing composite hypotheses" *The Annals of Mathematical Statistics 9*: 60–62

62. Williams, J. T. (1979) "Uncertainty, and the Accumulation of Human Capital over the Life Cycle." *Journal of Business* 52(4): 521–524

63. Yan, M. and Ye, K. (2007) "Determining the number of clusters using the weighted gap statistic" *Biometrics* 63(4): 1031–1037

10 Appendix

10.1 Broyden's Algorithm

Step 1: Initialise θ_0, obtain $l(\theta_0)$, $s(\theta_0)$, $H(\theta_0)$ and $H^{-1}(\theta_0)$. Put $t = 1$, $\delta \to 0^+$.

Step 2: Obtain $\theta_1 \leftarrow \theta_0 - H^{-1}(\theta_0)s(\theta_0)$. Hence, obtain $s(\theta_1)$.

Step 3: While $2 \leq t \leq maxiter$, for some $maxiter$,

$$\delta_t := \theta_t - \theta_{t-1}$$

$$\epsilon_t := s(\theta_t) - s(\theta_{t-1}).$$

If $|\delta_t|_F < \delta$ or $|s(\delta_t)|_F < \delta$, terminate and exit to ***Step 4***. Else,

$$H^{-1}(\theta_{t-1}) \leftarrow H^{-1}(\theta_{t-1}) + \frac{(\delta_t - H^{-1}(\theta_{t-1})\epsilon_t) \cdot \delta_t' \cdot H^{-1}(\theta_{t-1})}{\delta_t' H^{-1}(\theta_{t-1})\epsilon_t}$$

$$\theta_{t-1} \leftarrow \theta_t$$

$$\theta_t \leftarrow \theta_t - H^{-1}(\theta_t)s(\theta_t)$$

$$s(\theta_{t-1}) \leftarrow s(\theta_t)$$

$$t \leftarrow t + 1$$

and repeat ***Step 3***.

Step 4: Return θ_t, $|s(\theta_t)|_F$, $l(\theta_t)$ and $H(\theta_t)$, where $|\cdot|_F$ is the Forbenius norm.

10.2 Empirical Bootstrap

Put B large. Let $u := g(X)$ be any statistic computed from the data
$X = (x_1, x_2, ..., x_n)'$; $F^*(X)$ be the Empirical Distribution Function (EDF) of X; and n be the sample size.

Step 1: For each $b \in \{1, 2, ..., B\}$, re-sample

$$X_b^* := (x_1^*, x_2^*, ..., x_n^*)_b'$$

which has the same size n as the original sample; and obtain

$$u_b^* = g(X_b^*),$$

the same statistic computed based on the re-sample. This is equivalent to re-sampling X_b^* from $F^*(X)$ with replacement, as if $F^*(X)$ was the true distribution of X.

Step 2: $D^* := (u_1^*, u_2^*, ..., u_B^*)'$ is then the bootstrap distribution of u.

In particular, letting

$$u = argmax_\theta l(\theta)$$

yields the bootstrap distribution of $\hat{\theta}_{MLE}$. Taking the sample covariance matrix of D^* and taking square-root of its principal diagonal yields the bootstrap standard errors of the ML estimates. As for the bootstrap C.I., we let

$$u = \hat{\theta} - \theta$$

and obtain its D^*. That is, $D^* = (u_1^*, u_2^*, ..., u_B^*)'$, where $u_b^* = \hat{\theta}_b^* - \hat{\theta}_{MLE}$. Next, locate the $\frac{\alpha}{2}^{th}$ and $(1 - \frac{\alpha}{2})^{th}$ percentile of D^*, and the $100(1 - \alpha)\%$ bootstrap C.I. is given by

$$[\hat{\theta}_{MLE} - D_{1-\frac{\alpha}{2}}^*, \hat{\theta}_{MLE} - D_{\frac{\alpha}{2}}^*],$$

where $D_{1-\frac{\alpha}{2}}^*$ and $D_{\frac{\alpha}{2}}^*$ are respectively the $(1 - \frac{\alpha}{2})^{th}$ and $\frac{\alpha}{2}^{th}$ percentile of D^*. In this paper, $\alpha = 0.05$, $B = 1000$.

10.3 Tables of Estimation Results