The impacts of the external shocks between 2001-2003 on intergenerational mobility: An application of estimating permanent income with machine learning

A thesis submitted by

Yangyuntao Xu

in partial fulfillment of the requirements for the degree of

Master of Science

in

Economics

Tufts University

May 2019

Adviser: Prof. Jeffrey Zabel

Abstract

This paper finds that machine learning can overcome the life-cycle bias in the process of estimating permanent income, where the permanent income is defined as at least 5-years' average income between 30 and 40. In order to measure the impacts of external shocks, the Bush Tax cuts and economic recession, on intergenerational mobility, dividend income is chosen as identification strategy. This quasi-experiment shows that economic recession brought about a heterogeneous decrease in mobility by 35.7%, and the Bush Tax Cut at 2003 brought about a heterogeneous decrease in mobility by 16.9%. Though the effect of the tax cut is not significant, it might be caused by an inaccurate estimated error in the regression. In addition, this paper finds that any instrumental variables used in the estimation are not supposed to be used in the second stage regression.

Table of Contents

List of Figures:

List of tables:

The impacts of the external shocks between 2001-2003 on intergenerational mobility: An application of estimating permanent income with machine learning

1. Introduction

It is always interesting to predict individuals' future decisions basing on current information. For example, economists believe in the permanent income hypothesis, individual will have a smooth current consumption basing on future information to maximize his utility. However, the permanent income is hard to estimate, people might have different prediction on his permanent income basing on the change of current information. A similar problem exists in the estimation of intergenerational mobility, and the concept, intergenerational mobility, measures how much personal success depends on individual ability rather than family background. A typical way of estimating intergenerational mobility is to calculate intergenerational income elasticity (IGE) using the income of fathers and sons in their 30s-40s. Age matters in the current income, for example, the income in one's early career has a higher chance be lower than the later stage. So, only panel survey can be used for research on mobility, and this bias is referred as life-cycle bias as Haider and Solon (2006) stated.

Chetty et. al. (2014) state that the Unites States is hailed as the "land of opportunity", and their findings suggest that the U.S. is a collection of different societies, including those with high or low mobility. In fact, in nearly every country, wealth and social capital accumulation in one generation affects the conditions of the next generation. Notably, it is widely believed that social mobility in the Northern European countries is higher than the rest of the world. Pekkarinen et. al. (2009) argue that it is a result of a high tax rate and systematic social benefits. Tax policies in these countries have served as a redistribution function across generations, which maintains high intergenerational mobility. The U.S. has performed several tax cuts in recent years, so an analysis of the Bush Tax Cuts from 2001 and 2003 and their impact on intergenerational mobility is a

worthwhile undertaking. Since there was also an economic recession during 2001, both events are the external shocks included in this analysis.

If people were affected by the events of 2001 or 2003, then the effects on their permanent income can only be observed when they turn 40. However, if the identification of the impact is based on whether they were 18 in 2001 or 2003, the results can only be observed in 2023 at the earliest. The choice of age 18 is a result of considering whether their higher education decisions were affected by the increase of their father's income. If more samples are needed, it would be better to have panel data ranged from 2020 to 2030. Therefore, current panel datasets cannot cover the range of the data without eliminating the life-cycle bias. This paper tries to use recent data from the Panel Study of Income and Dynamics (PSID), but the number of father-child pairs is limited due to the time limitation. As for the Current Population Survey (CPS), though it has enough observations, but it can't overcome the life-cycle bias. In addition, household surveys are rarely used in the IGE estimation because the father and the child need to be in the same household to be recorded in a household survey, and the income can't be tracked across different waves. Though the life-cycle bias can be solved by machine learning, the co-living bias is hard to eliminate. The co-living bias also indicates that the child might have a stronger correlation with the father. However, there is no research showing whether the bias will increase or decrease the IGE. These typical biases are reasons why recent events have not been analyzed by researchers. If panel survey data is used for this analysis, then the life-cycle bias and co-living bias will affect the estimation results.

Basing on these considerations, this paper uses machine learning to eliminate the life-cycle bias in the CPS. With the development of artificial intelligence and data size, it would be interesting to test what machine learning can do in economic research. A common

application of machine learning in economics research is prediction, so if permanent income can be obtained from some observed current data, then household surveys can be used to estimate the IGE without being subject to the life-cycle bias. This paper tries to use recent data from the PSID to build a model which can transfer current income to permanent income. Resultantly, the estimation model was transferred into the CPS. Though the CPS has a co-living bias, the bias can be reduced if a difference-in-difference method is used. The co-living bias cannot be avoided for the CPS, so using an estimated permanent income would be the best option for this paper. In addition, the co-living bias might not be that relevant in this research if the effect of the tax cut is uniformly distributed between the co-living and non-co-living families.

The second part of this paper is a policy analysis of the Bush tax cuts and the economic recession. It is widely accepted that the tax cut in 2001 focused on labor income, while the tax cut in 2003 focused on capital income. This paper finds a significant impact of the both tax cuts using a single dummy of time. As for a difference-in-difference model, this paper uses 'whether a person has a dividend' as the identification strategy of event effect. Though the dividend is closely related to wealth accumulation, the experiment is not a perfect random experiment. It is hard to find a perfect random experiment in economic research, but a feature that can identify treatment and control groups is good enough for a quasi-experiment. In addition, having a dividend in a certain year does not mean that the same individual has a dividend in another year. It is possible to consider the dividend as a random choice influenced by risk preference. The results might not indicate causality, but they do present a heterogeneous effect on different social groups. If the findings of this paper are reliable, then the capital income tax cut decision likely underestimated its impact on social mobility.

This paper assumes that external shocks will change a father's income, leading to a change in higher education investment which will affect intergenerational mobility. When a father makes his decision on his child's higher education, a sudden increase or decrease of his income might lead to a different result if he has a budget constraint. Especially, families with insufficient higher education investment can then increase their investment. In this way, if higher education is critical for intergenerational mobility, tax cuts would have an impact on intergeneration mobility, when there a sudden change cause by external shocks. Overall, the tax cuts had a heterogeneous decrease on mobility by 16.9%, not statistically significant, and the economic recession had a heterogeneous decrease on mobility by 35.7%, statistically significant. There was an overall increase in mobility of 35%. Clearly, the treatment groups affected by these events experienced a relatively smaller increase in mobility, while the control group enjoyed a larger increase in mobility.

In short, there are three contributions this paper provides. First, this paper shows that machine learning can provide estimates of unobserved variables, such as permanent income. Second, if the estimated variable is used as an instrument in machine learning for estimation, then these instruments are not supposed to be used for a second time in an econometric regression: the estimation result is highly related to the instrument, so any estimated coefficients would be lower. Lastly, this paper finds that the economic recession caused a heterogeneous decrease in mobility by 35.7%, and the Bush Tax Cut of 2003 brought about a heterogeneous decrease of mobility by 16.9%. The heterogeneous decrease is between the non-random treatment and control group in the event.

2. Literature review

2.1 Estimation of permanent income and life-cycle bias

There are various discussions on the estimation of the relationship between current income and life-time income. Haider and Solon (2006) wrote a classic paper on this relationship. They estimate it by equation (2.1.1):

$$y_{it} = \gamma_t * y_i + \alpha_i \ldots (2.1.1)$$

y_{it} is the log of current income for individual i in year t and y_i is the log of life-time income for individual i. If $\gamma_t = 1$. This is called the textbook errors-in-variables problem. If y_{it} proxies to for y_i, then the slope coefficient equals to true coefficient times a factor:

$$\frac{\mathrm{var}(y_i)}{\mathrm{Var}(y_i) + \mathrm{Var}(\alpha_i t)}$$

If $\gamma_i \neq 1$, then y_{it} needs to be recalculated. Because y_{it} no longer proxies to for y_i. First suppose a worker's income in year t is determined by the following, where γ_t is the growth rate of earnings across the population. This equation means the income in year t is cumulated differently according to individual i.

$$y_{it} = \alpha_i + \gamma_i * t \ldots (2.1.2)$$

Then the permanent income in logs will be:

$$V_i = \sum_{s=0}^{\infty} e^{(\alpha_i + \beta_i s)(1+r)^{-s}} \cong e^{\alpha_i((1+r)/(r+\beta_i))} \ldots (2.1.3)$$

Hence, now V_i is different from y_i, because y_i is proxied by y_{it} and V_i is a function of t, r and varies according to individuals.

$$Log(V_i) \cong \alpha_i + r - logr + \beta_i/r \ldots (2.1.4)$$

So, the coefficient of the permanent income to current income is:

$$\lambda_t = \frac{cov(logV_i, y_{it})}{Var(logV_i)} = \frac{\sigma_\alpha^2 + t\sigma_\gamma^2/r}{\sigma_\alpha^2 + \sigma_\gamma^2/r^2} \quad \dots (2.1.5)$$

Using the model here, they found that the relationship between current income and life-time income was not stable. It is under 1 before the age of 40, and above 1 for individuals in their 40s-50s, and finally, it decreased below 1 after 50s. People would adjust their consumption and income behavior because they can observe the change of the permanent income. When this paper talks about perfect correlation, it is referring the coefficient equals to 1.

There is also another way of thinking about this problem. People can form expectations for their life-time income based on their current income, because people can only observe their current income. Though people might have different ways to calculate a life-time income, using machine learning as the best estimation might be a useful to reveal an individual's preferences and inter-temporal decisions. This assumption can be realized by machine learning and the equation 2.1.6. Though this equation is not the final estimation method for the machine learning approach, it can provide a basic understanding over the mechanism.

$$y_i = \alpha_t y_{it} + \beta_i X + \mu_{it} \qquad \dots (2.1.6)$$

2.2 Current researches on the IGE in the US

Solon (1999) states the importance of intergenerational mobility, and notes that two societies with similar income inequality can have different mobility. He also summarized researchers' findings concerning IGE using different measurements of the permanent income and father-child pairs using the PSID. Minicozzi (1997) uses the log of the 2-year average of annual earnings with an age range for the son of 28-29. As for the father, it is

defined as the log of present discounted value of the father's lifetime earnings, and the regression result is 0.42. If the income is the parent's income, the estimated IGE is 0.53. Mulligan (1997) uses log of multi-year (up to 5-year) average of annual earnings with the son's age range between 23-37. The definition of the father's income is also the log of multi-year (up to 5-year) average of annual earnings, and the estimation result is 0.32. Counch and Dunn (1997) use log of multi-year (up to 6-year) average of annual earnings as income for both father and son, and they have an estimation result of 0.13. Lillard and Reville (1996) use a 3-year average of log annual earnings for both father and son, and they have an estimation result of 0.34. For daughters, Minicozzi (1997) uses the log of 2-year average of annual earnings for both daughters. As for the father, it is defined as log of present discounted value of father's lifetime earnings. He finds an IGE of 0.41 for his definition. Shea (1997) uses a longer range of average income and finds an IGE of 0.56. Their findings show that the females have lower mobility compared with males. This might be a result of social inequality because males have greater freedom to choose the work they prefer.

It can therefore be concluded that if the income is defined as multi-year average, the longer average years is used, a higher IGE will get. In recent years, Chetty et al. (2017) estimates the IGE across the U.S. is around 0.5 using tax information, and this conclusion is similar for this paper's finding from PSID, where the IGE is estimated at 0.49. For this paper, the estimated IGE with PSID and an at least 5-years' average income, the IGE for overall is 0.49, for father-son in is 0.27, for father-daughter is 0.63. This result should be close to others' findings mentioned above, and the gender gap is increasing over time. Resultantly, the data cleaning process and the definition of the permanent income with the PSID are valid, and more estimation details will be mentioned in the following sections.

2.3 A difference-in-difference model for the policy analysis

Firstly, Pekkarinen et al. (2009) report that comprehensive school reform in Finland could reduce IGE from 0.30 to 0.23. Their paper use a DID approach to analyze this reform, and following their approach, this paper forms the estimation model in the following equations:

$$\ln\left({y_{ijt}}^{Child}\right) = \alpha_t + \beta_{jt} * \ln\left({y_{ijt}}^{Father}\right) + \varepsilon \ldots (2.3.1)$$

According to definition, β_{jt} is the coefficient of IGE and j stands for whether affected by treatment and t stands for the different control and experiment cohorts. To examine the effect of the treatment by $cohort_t$, the effect of being different group by $area_j$, and compose an interaction term to measure the difference-in-difference (DID) effect as listed below.

$$\beta_{jt} = \beta_0 + \beta_1 * area_j + \beta_2 * cohort_t + \beta_3 * area_j * cohort_t \ldots (2.3.2)$$

In their case, the reform took place in regions of Finland at different times. Some districts started earlier than others, so the treatment here is extending the length of comprehensive education in primary school. The data comes from the Finnish Longitudinal Census by Statistics Finland. It was conducted every fifth year between 1970 and 2000. They randomly selected 10% of the 6.3 million individuals in Finland as their research sample. They define child as son and calculate his income by his log taxable earnings in 2000; the father's income is calculated as the average of log taxable earnings in 1970, 1975, 1980, 1985, and 1990, and then converted into 2000 prices. The treatment is whether affected by the cohort, specifically, 1 for being younger than the start year in one area; the area is a dummy for different areas.

After combining these two equations here and adding the main effects of the treatment, cohort, and interaction status, a new equation is obtained as below. Here y_{ijt} stands for

the income of observation (i), the difference in treatment by j and the difference in cohorts by t. Then add main effects to equation. They stand for the direct impact of the dummy, treatment and cohort, to children's income.

$$\ln\left(y_{ijt}^{Child}\right) = \alpha_t + \beta_0 * \ln\left({y_{ijt}}^{Father}\right) + \beta_1 * area_j * \ln\left({y_{ijt}}^{Father}\right) + \beta_2 * cohort_t$$
$$* \ln\left({y_{ijt}}^{Father}\right) + \beta_3 * cohort_t * area_j * \ln({y_{ijt}}^{Father})$$
$$+ \beta_4 cohort_t * area_j + \beta_5 area_j + \beta_6 * cohort_t + \beta_7 * X$$
$$+ \varepsilon_{ijt} \ \ldots \ (2.3.3)$$

According to the introduction to the DID method, β_3 stands for the estimated DID effect, and this coefficient is also the interaction term of the other two dummy variables. β_2 is the effect of income inequality and β_3 is the DID effect.

2.4 Selected machine learning models and applications

Athey (2018) provides a relatively narrow definition of machine learning: *"machine learning is a field that develops algorithms designed to be applied to data sets, with the main areas of focus being predication (regression), classification, and clustering or grouping tasks."* Athey argues machine learning is an intermediate step in empirical works in economics.

Chen and Wang (2018) demonstrate the difference between machine learning and OLS regression in their working paper. In the following equations, some derived errors and biases of a RIDGE regression and an OLS regression is listed. The key difference between a RIDGE regression and an OLS regression is the penalty term, $\lambda \sum_{j=1}^{m} {\beta_j}^2$, for model complexity. After applying this term, the solution for the target equation is completed, and in most cases, a numerical solution would be a better answer.

$$\widehat{\boldsymbol{\beta}}_{Ridge} = argmin \sum_{i=1}^{n} (y_i - \sum_j^m \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{m} {\beta_j}^2 = (X^T X + \lambda I)^{-1} X^T Y \ldots (2.4.1)$$

Here, the coefficient is $\widehat{\boldsymbol{\beta}}_{Ridge} = \frac{\boldsymbol{\beta}}{1+\lambda}$ so the bias and variance are the below.

$$Bias_{Ridge} = \frac{\boldsymbol{\beta}}{1+\lambda} - \boldsymbol{\beta} \;\; Variance_{Ridge} = \frac{\sigma^2}{(1+\lambda)^2} \;\; \text{...(2.4.2)}$$

As for a typical OLS regression, the bias and variance are easy to obtain, as below.

$$\hat{\beta}_{OLS} = argmin \sum_{i=1}^{n} (y_i - \sum_j^m \beta_j x_{ij})^2 = (X^T X)^{-1} X^T Y \;\; \text{...(2.4.3)}$$

$$Bias_{OLS} = 0 \;\; Variance_{OLS} = \sigma^2 \;\; \text{...(2.4.4)}$$

After comparing the result from the equations 2.4.2 and 2.4.4, the variance from the OLS regression is constantly larger than then variance for the RIDGE regression.

As for the error in the equation 2.4.6, the signal of this equation depends on λ. After calculating its first order condition and putting it back into the equation, the equation becomes 2.4.7, where all values are positive. Then, we can always find a λ , letting the estimation power of the RIDGE regression work better than the OLS. Theobald (1974) argues that even in the condition of $\lambda < 2\sigma^2(\beta^T\beta)^{-1}$, the error from the RIDGE regression is smaller than the error from the OLS regression.

$$Err_{OLS} - Err_{RIDGE} = 0 + \sigma^2 - (\frac{\lambda\beta}{1+\lambda})^2 - \frac{\sigma^2}{(1+\lambda)^2} \;\; \text{...(2.4.6)}$$

$$Err_{OLS} - Err_{RIDGE} = \frac{2(\beta^T\beta)(\sigma^T\sigma) + (\sigma^T\sigma)^2}{\beta^T\beta + \sigma^T\sigma} \;\; \text{...(2.4.7)}$$

Chen and Wang (2018) have written a summary on current findings in the application of machine learning in social science. They conclude that there are three major areas of application: data generation, prediction, and causal inference. In addition, the source of data can be natural language, images, and tables. Athey (2018) argues that it would be impossible to use machine learning for causal inference due to the limitations of machine learning itself. Because the estimated coefficient is biased, so even if it not different from 0, the number doesn't reflect any economic intuition.

There are several papers that use machine learning in their research. Concerning data generation for example, Bleakley and Ferrie (2016) studied whether the accumulation of

wealth would increase education investment for the next generation. They had a problem with some father and child pairs being hard to group, such as when females change their last name after marriage. The solution they utilized was to use machine learning to predict family relationships. Iaria et al. (2018) tested whether World War I impacted the exchange of global academic research. They argued that the similarity between paper titles reflected academic collaboration and cooperation and used latent semantic analysis to test the similarity of 40,000 paper titles from different sides of the war. They found a significant drop in the similarity of titles after 1914.

Regarding prediction, Blumenstock et al. (2015) worked to investigate the allocation of wealth in developing countries. However, the quality of official data in these countries is low, so they used mobile phone metadata including recent calls. The authors first collected 856 individuals and their economics status, then set up a model to predict economics status. This model could predict the distribution of wealth with a high prediction power. Engstrom et al. (2017) have done similar work with satellite images to predict wealth conditions in developing countries. One possible advantage of these methods is that they can observe a data size that expand the boundary of data sizes and sources in economics research. However, though these kinds of data address a wider range than survey data, they might also be biased by the way they are obtained. For example, for phone metadata, people need to have a phone as a minimum requirement to be collected into the sample, and in developing countries this can rule out large groups of people. As for the satellite images, it could be affected by different weather conditions. For example, higher levels of traffic might be observed in a sunny day after long rainy seasons, so the observed clustering of cars does not imply a higher volume of car ownership.

In fact, the casual inference of machine learning is combined with current econometric tools, such as instrument variables and difference-in-difference analysis. However, due to the nature of machine learning, when it chooses to reduce variance with a biased estimation the estimation coefficient can never be used for casual inference. The estimation results can be a better way to pre-process raw data. For example, in the first stage of IV estimation, the goal is to use exogenous variables to predict endogenous variables, which is when machine learning can provide a better estimation than OLS regression. Some papers have used LASSO and RIDGE to build the first stage of estimation (Belloni, e al. 2012; Carrasco, 2012; Hansen and Kozbur, 2014), while Hartford et al. 2016 have used neural networks to conduct this estimation.

3. Estimate permanent income using machine learning;

3.1 Definition of permanent income;

The definition of permanent income is one of the key questions to answer in the estimation of IGE. Theoretically, Permanent income is defined as the average yearly income for the total life-time income. However, as Haider and Solon (2006) report that the incomes of individuals in their 30-40 are perfectly correlated to permanent income. Therefore, using the average income in these years will be an estimation of the permanent income. Researchers have used several different ways to define this variable, for example, they use annul income, at least 3-year average, and results are reported in the previous review. So, practically, to compare the findings in this paper to others, it is defined as at least five years' average income between 30 and 40. Specifically, to compare income across years, all income is converted into 1968 dollars, all observations need to have at least 15 recorded years in the PSID, 2% interest rate is used for the accumulation

process and all records with a recorded age below 18 are dropped. The second concept is the definition for the father-child pairs. This definition has a major impact on the estimation of the IGE: typically, it is defined as father-son pairs, and often only with the first-born son. However, in order to obtain the largest possible sample, here it is defined as the father and all his children, including daughters and sons.

The key data source for this process is the PSID. This is one of the largest panel data sets in the U.S., and it began in 1968 with over 18,000 individuals in 5,000 families in the U.S. Information in the PSID contains variables such as income, health, marriage, childbearing, and numerous other topics. This dataset is conducted by the University of Michigan and the data is widely used by researchers, policy analysts, and teachers around the globe. In addition, the application for the Current Population Survey (CPS) provides primary statistics on the U.S. labor force. As the oldest, largest, and most well-recognized dataset in the United States, it is sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS). It provides monthly information about 60,000 households on many aspects, including work, earnings, and education. For the analysis of the impact from external shocks on mobility and inequality, the CPS 2003-2013 is used for analysis. It also includes the information of dividend income and homeownership, which is used in the model setting. It was downloaded from the IPUMS with the selected variables that this paper is concerned with.

3.2 The data cleaning process and results in the PSID;

Before model training, obtaining variables that are available across all waves was of great importance. The PSID does provide a cross-year variable table, and it provides information on variable names. There is a table 1 from the PSID which summarizes all variables that can be tracked across years from 1968 to 2015. Unfortunately, the number of trackable

variables is far less than the total variables. This unexpected fact might help this model to be generalized into other datasets because machine learning needs as much information as possible to perform a better prediction on an outcome. The list of obtained variables is in table 1, and the final variables used for training is in the table 2. For the variables in the table 2, they are the variable that co-exist in both the PSID and the CPS. The only variable that needed recoding was education, and it is recorded into year of education with the same definition as with the PSID.

Summary table 1 shows there are 443 observations in the PSID sample. Only the pairs of head and child are kept here: those of father-head, father-brother, or sister and child-grandchild do not remain in the sample. The observations are recorded with a new identification number for the first time they have a new family relationship code. Since the purpose of testing the IGE in the PSID is to compare the results with findings from previous research. Such a choice is one way to start the comparation for the data structure.

Here the summary statistics show that the average yearly income of the child is lower than the father. This limitation might be caused by the data source, because the sample are selected to the paired father-children. As it is impossible to trace everyone over their whole lifetime, it is normal that only the early stage of the child is recorded in this dataset, with future information to be recorded in other waves. In addition, the standard deviation of the father's income is lower than the children'. This is also another signal that there is still an increase in the range of income. As for the observed mean age of the child, it is around 22, with a highest value of 42. In data preparation, the range of age was kept between 18-65, and all observations need to have income. Therefore, it is understandable that the first recorded age for a child is around 22. However, 8 observations are first

recorded above 22. These might be abnormal observations in the database, but as the size is small, they were dropped, so the final dataset contains 443 observations in total. The father's age range is also expected: normally they became the new head or answer the questionnaire at age 24, on average, and are recorded as household head around age 42. The education is recorded as 0, if it is not recorded in that year. These are replaced by the first observed education level in the dataset. The working status includes working now, temporarily laid off, unemployed, retired, disabled, student, and keeping house. This variable is not included in machine learning, so it is not recoded.

In the PSID, the first approach provides a higher IGE at 0.49 and 0.27 for males. These results are listed in the regression tables 1 and 2. This average provides a similar result to the findings of others. In addition, the five-year estimation provides a higher IGE than the three-year average. This result is also consistent with other findings (Lillard and Reville, 1996). Therefore, this approach was adopted for this paper. Regression table 1 shows the results. Lastly, when the income is accumulating, an interest rate shall be considered. In this paper, it is set at 2%. Though the mean income of the children is lower than the father's, the standard deviation is larger. As it is stated in other sections, the selected pairs might be biased on many conditions, so the children have a higher income than the parents. The distribution of income might indicates that the diversity of the next generation is higher than pervious expectation.

3.3 The process of machine learning estimation;

A typical example of machine learning is predicting the housing selling prices. In this case, there is information about the size of the house, rooms, location, the construction year, and other information that might be important. This information serves as independent variables, and housing price is the dependent variable that needs prediction. If there are

enough cases with sale price, the price for a new house can be estimated. Similarly, if there are enough cases about one's permanent income in different years, then estimating one's expected permanent income based on one year's information would be possible. Therefore, the data structure is the information about one's current status and the estimating variable is the c permanent income. In this way, the machine learning task can be defined. For the datasets such as the CPS, it is not a panel, meaning it is typically considered impossible to conduct IGE analysis. These data structures can therefore be expanded into the scope of research.

These parameters can be considered as a choice of the penalty on the model complexity. The goal of the selection process it to maximize one of the selected conditions. If it is a classification problem, then accuracy rate would be the first choice. As for the prediction of permanent income, using the R^2 is one possible choice. Then the selected variable will be fitted into a model with highest R^2, which would be used for the estimation in the CPS. The process of choosing parameters for the best prediction is as follows: first, split the data into training and testing sets. Prediction from the trained model can then be tested with the current dataset to see whether the model has made a good prediction. Second, choose a parameter in a machine learning model to maximize the R^2 in the test data set and then minimize the gap between the testing and training data sets. This condition can reduce the possibility of obtaining an overfitted model. In the figure 5, the three machine learning models used are random forest regression, Ada-boost regression, and gradient-boost regression to maximize prediction power. The y-axis is the R^2, the dotted lines are the results from training sets, and the line with crosses are the results from the testing sets; the higher the lines with crosses, the better the model is. All of them are simple machine learning models, but they provide enough prediction power for this paper's

target. Since the aim of this paper is economic application, there might be more useful machine learning models with higher estimation power.

First, consider the random forest regression. The explanations are following some documents online, Turi machine learning guide. It is one of the most effective additive models for predictive analytics. It combines decisions from a sequence of base models and can be written in the following equation:

$$\mathrm{g(x)} = f_1(x) + f_2(x) + f_3(x) + \cdots (3.3.1)$$

where the final model g is the sum of simple base models fi. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called model ensembling. In random forests, all the base models are constructed independently by using different subsamples of the data.

Second, consider the Ada-boost regression. An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases. Lastly, consider the gradient-boost regression. It builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function. Since this paper is focusing more on the application, technical details are not included here. Similarly, the smaller the gap between the training and test sets, the better the model is. The final choice for this paper is the gradient-boost regression with n estimators at 100. This model provides an R^2 at 0.65, which provides considerable prediction power for estimating permanent income.

However, in the model training process, the estimated testing R^2 is around 0.65 for the CPS. If all across year variables in the PSID can be used, the R^2 can reach 0.7, so the result from a smaller variable space is lower than expected. if the R^2 can reach 0.7, the estimation results might be more reliable. For other empirical research, Blumenstock and Cadamuro (2015) also used a machine learning model with an estimation of R^2 at 0.68 when they used metadata from phones to predict wealth allocation in several African countries. In light of this, the R^2 at 0.65 for this paper should be good enough to make predictions.

After the whole process of estimation, there are something findings. First, setting the models for different age and gender cannot increase the estimation power of the machine learning model. Second, the income from the children aged under 18 does not equal a full-time job's income. These observations would make the estimated IGE non-linear, so they were dropped before estimation. Third, from this model selection process, having a clear definition of the estimation target is the first step in applying machine learning. The ambiguous definition of permanent income brings considerable challenges to the process. After adjustments to the training set and target definition and setting permanent income to be at least five-observed-year average between 30 and 40, the final model provides robust estimation results.

3.4 Estimating IGE in the CPS with the predicted permanent income;

When the trained model is applied to the CPS dataset, the remaining variables are current income, year, race, age, cpi, education, state, and gender. Information on state and education is transformed into dummy variables, hence, there are 64 remaining variables. It would be better if the PSID could contain more co-existing variables than it does currently. This limitation might decrease the credibility of this paper. Under this condition,

the testing result still provided a R^2 at 0.65. In addition,the father-child pairs include both sons and daughters, and one father might be paired to several offspring. For the CPS, the age range of the children is between 16-21 in 2001, and the father must be younger than 65 in the recorded year in the CPS. In addition, Chetty et. al. (2014) show there is a non-linear relationship in the IGE for families with different incomes. Therefore, the top and bottom 10% paired samples in term of fathers are dropped. When the PSID is used for calculation of the IGE, a similar selection process is adopted.

Though this is not expected, and there is an increase in standard deviation, there remains a significant number of children who earn a higher wage than their fathers. It is reasonable to argue that the distribution feature of the PSID is successfully transferred into the CPS. In addition, a decrease in the estimated IGE with the increase of father's income is also observed in the PSID. This might be a result of the co-living bias, and it is surprising to find that this bias will decrease the IGE in the data. One possible explanation is that homeownership might be the pre-requirement for co-living (the father can afford a house while the child cannot do the same). This would lead to a decrease in the IGE.

Before using machine learning, the estimated IGEs in the CPS are mostly insignificant and negative. These estimations are clearly biased due to various reasons: after adopting the machine learning estimation, the results become positive and significant. Summary table 2 shows the distribution of each variable. In the CPS with estimated permanent income, the father's income is still higher than the children's. The age range of the offspring is between 18-33, while the age range of the fathers is between 37 and 64. The estimated father's income is higher than the child's income on average, which is consistent with the summary statistics in the PSID. The impact of life-cycle variation in income still exists. If there is a better training source than the PSID, a better result can be estimated.

Lastly, it would be interesting to demonstrate the distribution of the IGE in the CPS for different ages of the children, the age in the recorded years. Figure 1 shows the IGE distribution by ages with control variables. For some ages in the graph, the estimated IGE is below 0 and the average is around 0.10. While in the figure 2, most of the IGE are above 0, it only becomes unstable around 60. In the final model, the age range of the children is between 18 and 33 to keep the age in 2001 between 13-21. Therefore, the distribution of the IGE for children aged between 18-40 without controls in the figure 3 is more stable. The results show a stable IGE around 0.30. In figure 4, there is also another figure showing 1% samples of total samples for the relationship between the predicted permanent income and current income in standardized form. There is also a linear fitted line in the graph showing a linear relationship between the predicted permanent income and current income. This is not the machine learning model used for prediction but demonstrates their relationship in a linear model.

4. How to use variables estimated by machine learning in regressions.

4.1 A comparation between instrument variable estimation and machine learning estimation

Another important discussion concerns the dependent variables. If the machine learning estimation process is considered as a similar approach to using instrument variables, then the V_{ijt} are all used for the first stage estimation in the equation 4.1.1, and those variables are not used for the second stage estimation. Since education, age, gender, and location have been used as instruments in the first stage, including them again in the second stage would reduce the estimated IGE.

$$\widehat{\ln}\left(y_{ijt}{}^{Father}\right) = \alpha_t + \beta_1 * V_{ijt} + \beta_2 * X_{ijt} \dots (4.1.1)$$

$$\ln\left(y_{ijt}^{Child}\right) = \alpha_t + \beta_1 * \widehat{\ln}\left(y_{ijt}{}^{Father}\right) + \beta_2 * X_{ijt} + \varepsilon_{ijt} \ \ldots \ (4.1.2)$$

However, there are still some differences between an IV estimation and a machine learning approach. First, if it is considered as an IV estimation, father's information is used to estimate father's income, but machine learning also uses children's information to estimated children's income, where the dependent and independent variables are used in the estimation process. (such as equation 4.1.3.)

$$\widehat{\ln}\left(y_{ijt}\right) = \alpha_t + \beta_1 * V_{ijt} + \beta_2 * X_{ijt} + \varepsilon_{ijt} \ \ldots \ (4.1.3)$$

$$\widehat{\ln}\left(y_{ijt}^{Child}\right) = \alpha_t + \beta_1 * \widehat{\ln}\left(y_{ijt}{}^{Father}\right) + \beta_2 * X_{ijt} + \varepsilon_{ijt} \ \ldots \ (4.1.4)$$

When there is a regression of children's income on children's information, the coefficients of these control variables will be significant and large if it is equation 4.1.4. As a result, the coefficients of other variables will be small. Since the dependent variable is predicted basing on these IVs, a high correlation and statistically significant result will not be a surprise. However, if other new variables are still significant with these IVs, then these findings are something that can explain some of the unexplained variations in the prediction. In addition, if the tax cuts have a small effect on the IGE, then without control variables, a simple regression might lead to an insignificant result. However, this does not mean that there is no statistically significant result. If the machine learning error is different from an OLS regression error, then testing whether the distribution of the error still follows a t-distribution would be an interesting undertaking. However, this question is beyond the purpose of this paper, which will present results without the variables used to predict permanent income.

As a short summary, this paper proposes a conjecture: for a regression with a predicted value from machine learning, the regressions that do not include the instruments used to obtain the predicted value provide more accurate estimates. In addition, if the results

are not significant, it would be useful to check whether there is a significant result with the used instruments.

5. The estimated impacts of external shocks;

5.1 Identification strategies

The Bush government carried out two different tax cuts: in 2001 and 2003. Auten, Carroll, and Gee (2008) and Yagan (2015) both share a similar understanding of the nature of the tax cuts, which is that the 2001 tax cut focused on labor income, while the 2003 tax cut focused on capital income. In addition, Burke and McCouch (2007) state that there is a tax cut on estate tax in 2001, and all policy was extended to be permanent by the later tax reforms. However, there was also an economic recession in 2001. During this period, the burst of the internet bubble led to a considerable drop in the stock market, but the recession did not expand to other industries. Though there was a drop in the stock asset price, real estate and other assets were affected less significantly. Additional information is therefore needed to identify which groups were affected by the external shocks. The additional grouping strategy used here is the dividend income. In the best case, if a father is identified with dividend income in 2001, then his decision on education investment was affected by the recession rather than the tax cuts. However, if he is identified in 2003, then he is affected by the tax cuts. In addition, in order to reduce the impact of other historical events, a small gap of comparation is used, so the treatment age is between 16-18 and the control group age is from 13-15, and 19-21. In this way, the time dummy might capture the external shocks better.

This paper assumes that the external shocks will provide a sudden change on family income. Ioannides (2019) states that tax reform is also another possible factor that will affect the IGE. If the family is facing a budget constraint on education investment, then a

sudden increase in income could provide the children in that family with a chance to attend college. Therefore, the children who cannot afford tuition before the tax cut can afford education afterwards. However, this would affect both higher education and primary education, and it is reasonable to believe the cost of high school is close to 0 in the US. The change of income would affect their higher education decision more significantly. Another strong assumption for this paper is that higher education plays a crucial part in intergenerational mobility. The current high school enrollment rate is high, and college enrollment is relatively lower. With the growth of information technology, a college degree matters much more than a high school diploma, and it is harder to find a well-paid job with only a high school diploma. Therefore, arguing that the impact on higher education matters more in intergenerational mobility is more reasonable.

Therefore, potential cut off points will be the college admission years. Then, as for the age of the children in 2001, it would be 18 for the 2001 tax cuts or 16 for the 2003 tax cuts. Using another variable to identify the sub-groups experiencing the heterogeneous effect of tax cuts will improve the estimation results. Since dividend income is chosen as the identification variable, and it has a direct connection to the stock market bust in 2001 and the tax cut in 2003. It can be used to measure the heterogeneous effect of the policy analysis among the affected and non-affected groups. After applying this DID method, the heterogeneous effect can eliminate the co-living bias in the CPS.

The tax cut is an exogenous variable in this setting, but the dividend income is not. Having dividend income indicates a higher income and higher mobility, therefore being in the sub-group or not is not totally random. Though some researchers report that risk preferences are irrelevant to wealth accumulation, wealth accumulation might affect risk preferences in the reverse. Resultantly, it is rich families that tend to be in possession of

risky assets, such as stocks. In contrast, if the 2003 tax cut is considered only as a capital income cut, then dividends can identify the group that is not affected by the cut. As for the 2001 recession, the impact was limited to the stock market. Therefore, a dividend in both periods is also a way to identify different social groups affected by the event. The events do not affect citizens randomly, however, so this is only a quasi-experiment. The results are heterogeneous effects on different social groups. In addition, other variables, such as father's education level, father's race, father's immigration status, and father's income level have been considered to separate the policy effect, but there are no statistically significant results for those under these separations. Therefore, using whether have dividend income is the best way to identify the external shocks.

5.2 Econometric model

Here is the final regression model used for estimating the events impact on the IGE. The equation is modified from equation 2.9 without adding dummy variables such as child's education years and the year of observation. Hence, equations 5.2.1 and 5.2.2 are the final regression for the empirical results. However, the regressions with these dummies are also included in the regression for comparation.

$$\ln\left(y_{ijt}^{Child}\right) = \alpha_t + \beta_0 * \ln\left({y_{ijt}}^{Father}\right) + \beta_1 * cohort_t * \ln\left({y_{ijt}}^{Father}\right) + \beta_2 * cohort_t + \varepsilon_{ijt} \quad \dots (5.2.1)$$

$$\ln\left(y_{ijt}^{Child}\right) = \alpha_t + \beta_0 * \ln\left({y_{ijt}}^{Father}\right) + \beta_1 * treatment_j * \ln\left({y_{ijt}}^{Father}\right) + \beta_2 * cohort_t * \ln\left({y_{ijt}}^{Father}\right) + \beta_3 * cohort_t * treatment_j * \ln({y_{ijt}}^{Father}) + \beta_4 * treatment_j + \beta_5 * cohort_t + \beta_6 * cohort_t * treatment_j + \varepsilon_{ijt} \quad \dots (5.2.2)$$

In this equation, i stands for the paired father and child, t stands for the identification of before the event year and after event year, and j stands for the identification of the tax cut in 2001. Though the differences can be considered as natural experiments, the selection criteria are somewhat related to tax reform. Therefore, arguing the regressions estimated heterogeneous effect of different social groups will give a more accurate description to this paper's findings.

In order to decrease the co-living bias, only children who are aged above or equal to 18 and fathers who are aged below 65 are kept in the final sample. For children under 22, the age for college graduation, keeping or dropping them will not affect regression results. This paper assumes that machine learning estimation can predict one's permanent income, even if the income earned as a part-time job can also be used for estimation. In addition, Chetty et. al. (2017) find that the IGE is non-linear between different families, so the top 10% and bottom 10% of father's income data are dropped. Therefore, there are total 34,815 paired father-children in the dataset.

5.3 Estimation results

Regression table 3 uses the model 5.2.1 where cut-off ages are used as a single dummy. The results are quite consistent across different cut-off ages. Models 1 and 2 set 18 as the cut-off age, so the 2001 recession is identified. Models 3 and 4 set 16 as the cut-off age, with the 2003 tax cut being identified. Models 5 and 6 set 16-18 as the cut-off age, and both tax cuts are identified. Models 2, 4, and 6 show the results when the year and education dummies are included. In these models, the estimated IGE is much lower than in other three models, and the time effect is stronger. As previously discussed, they are not supposed to be included in this model, though they are here to demonstrate the different results. Overall, the estimated increase of the mobility is around 35.9%. Though

a small age gap is chosen here to measure the effect from the tax cuts more precisely, it is still hard to conclude that all these changes were caused by the tax cuts. Therefore, a more detailed analysis with the DID model is required.

Regression table 4 presents the estimation results from equation 5.2.2 with different cut-off ages. In regression table 4, it identifies the heterogeneous effects of the tax cut and recession by the difference-in-difference measurement. This effect is significant in model 1, but not in the models 2, 3, and 4. The difference in these models is the cut-off age to identify the tax cut effect and the treatment is having a dividend. As discussed in the previous section, if a father is identified with dividend in 2001, then he is affected by the recession, while if in 2003, he is affected by the Bush tax cut. Model 1 provides a significant estimation result for all key variables, which means the recession brings a heterogeneous effect on decreasing intergenerational mobility for the family with dividend by 35.9%, and an increase for the overall population at 42%. In addition, if a child enters university during the recession, his income will increase by 1.34 in the log term if his father also has dividend, then his income will increase by another 0.66 in log term.

It is possible that people who enter the job market after a recession will benefit from an economic boom. As a result, their income is higher than those who went through the recession. At the same time, families with dividend income tend to have diversified investments, so when there is a recession, their investment and consumption behavior might be affected less, while families without dividend income might face a decrease in labor income or even the chance of losing their job. In this case, their higher education investment for their children would be affected. The regression here might demonstrate an economic intuition.

However, model 4 shows there is no statistically significant relationship between the 2003 tax cuts on capital and the IGE. The coefficient of the heterogeneous effect is at 0.049, a 16.4% decrease for the family with dividend. As discussed, the test results for this kind of regression might be unreliable. Though the bootstrap and other measurements will not change the result, using homeownership as a cluster would make the DID coefficient a significant indicator. Due to the purpose of this paper, there is no evidence related to the error calculation that can be given. It is possible that the distribution of different groups is not random, but highly related to the wealth level. Therefore, a heterogeneous distribution of the IGE is anticipated. With this correction, the effect should also be taken into consideration.

A similar story is behind these results: when there is a capital tax cut in society, though overall investment on education might increase (bringing about a lower IGE), the families with dividend income might have a better chance to use their influence to help their children gain an opportunity to attend higher education. Since these kinds of connections and theories are harder to observe, the effect of the tax cut is only half of the impact of the recession.

Therefore, it is possible that the Bush tax cuts have little impact on the IGE, and the significant effect captured in the single-dummy model is the effect of the recession. If the machine learning changes the distribution of the variations, and the result in regression table 2, then model 3 is accurate. If the family asset preference is independent to intergenerational mobility, then the capital income here is a good indicator for natural-experiment, and the heterogeneous effect here is the difference-in-difference effect. There is some evidence supporting this assumption: Parada-Contzen (2017) has a working

paper stating that risk preference has nothing to do with wealth accumulation. I would like to argue here that the heterogeneous effect is the true effect of the treatments.

Robustness tests were completed for this paper. First, using race, education, homeownership, income, and immigration status as identification strategies did not produce any significant results. Second, setting a different cut-off age did not produce significant results. Lastly, different error measures are listed in regression table 5, where only the clustering on the homeownership for the 2003 tax brought significant results.

6. Conclusion

Overall, this paper finds that machine learning can overcome the life-cycle bias in the process of estimating permanent income, and some interesting results on tax cuts, economic recession, and intergenerational mobility. Though the co-living bias cannot be solved by machine learning, it is identified in the estimation results. Since a difference-in-difference comparison is used to identify reform effect, the estimation results are robust. The permanent income is defined as at least 5-years' average income between 30 and 40. There are three contributions this paper provides. First, this paper shows that machine learning can provide estimations of unobserved variables, such as permanent income. Second, if the estimated variable is used for regression, any instrumental variables used in the estimation are not supposed to be used in the main regression. Because the estimation result is highly related to these variables, any estimated coefficients would therefore be lower. Lastly, this paper finds that economic recession brought about a heterogeneous decrease in mobility by 35.7%, and the Bush Tax Cut at 2003 brought about a heterogeneous decrease in mobility by 16.9%. The heterogeneous decrease is between the non-random treatment and control group in the event. Though the effect

of the tax cut is not significant, it might be caused by an inaccurate estimated error in the regression.

It is possible that rich families have already invested enough resources on their children's higher education, a tax cut on the top 20% might not be able to change the education decision. While an economic recession brought more impact on the middle class' education decision, and this brings a more significant impact. In both cases, the treatment groups have a less increase in mobility, while the control group have a large increase. Therefore, both shocks won't affect these family with enough education investment, while these middle-class families are affected more.

As for the estimation process of the permanent income with machine learning, the details are listed below. First, the PSID dataset was used to build the relationship between current information and permanent income. The PSID is a longitudinal survey for individuals across years. If there are variables that are trackable over years, then after calculating permanent income these variables can be used to predict permanent income. In the process, one individual must have at least 5 recorded years in the data, and the average of income with an interest rate at 2% between 30 to 40 is used for permanent income. Second, when the dataset is built, machine learning was used to find the optimal coefficients that provide the highest R2 in the model. Third, the same variables were used in the CPS dataset to predict permanent income there. Lastly, this new income was used to test the policy effect of the Bush tax cut. In addition, there is information about homeownership and dividend in the supplements of the CPS, and this information is also included in the dataset.

After data processing, the result shows an estimated IGE in the PSID of 0.49 and 0.39 for father-son pairs. Some papers report similar IGE using the PSID and the same

measurement of the permanent income. While a pre-processed IGE in the CPS is 0.06, after the permanent income is estimated by the machine learning model the IGE is 0.24. These results are expected. The IGE from the CPS before processing is low, so it should be affected by the life-cycle bias. Meanwhile, the processed IGE is lower than the IGE from the PSID, and it should be affected by the co-living bias. After comparing the summary statistics, father's income is higher than the child's income in both datasets. Though this was not expected, there is an increase in standard deviation, and there are still significant numbers of children who have a higher wage than their fathers.

There are some shortcomings to this paper. First, if the tax information could be used in the analysis, then the outcome would be more stable. The training process could use more recent data and a wider range of variables to measure permanent income, and the empirical analysis could avoid the co-living bias. Second, it would be better to find some theoretical research on the different effects of labor tax and capital tax on intergenerational mobility. In addition, more results related to risk-preference and wealth accumulation should be included. Lastly, it would be better to try to find how machine learning estimations changed the distribution of the predicted error.

With these findings, it is reasonable to believe that machine learning can be used in economic research in various ways, and especially for predicting unobserved variables. This paper also opens a discussion on how to regress the estimation results from machine learning and error calculation. The theoretical discussion of the econometric model is beyond the scope of this paper, so that is not included here. In addition, if there is any theoretical model can demonstrate the difference of tax cuts and economic recession on the IGE, it can support this paper's finding. As for policy makers, they often pay little attention to changes of social mobility; after all, the effects cannot be measured 30 years

later. With the implementation of machine learning, the impact can be measured in a shorter range.

In short, this paper shows that machine learning can be used in economic research, especially for predicting unobserved variables, though this estimation process also brings about more questions that need answering. This method expands current information so that it can be used to estimate future status, and this idea might be adopted in many other research projects concerned with human decisions. Lastly, this paper argues that a tax cut on capital income and recession would decrease social mobility. Hence, any policy related to tax cuts should consider both income inequality and social mobility.

Appendix:

Figure 1: The estimation of the IGE in the CPS by ages (less than 60) with control variables

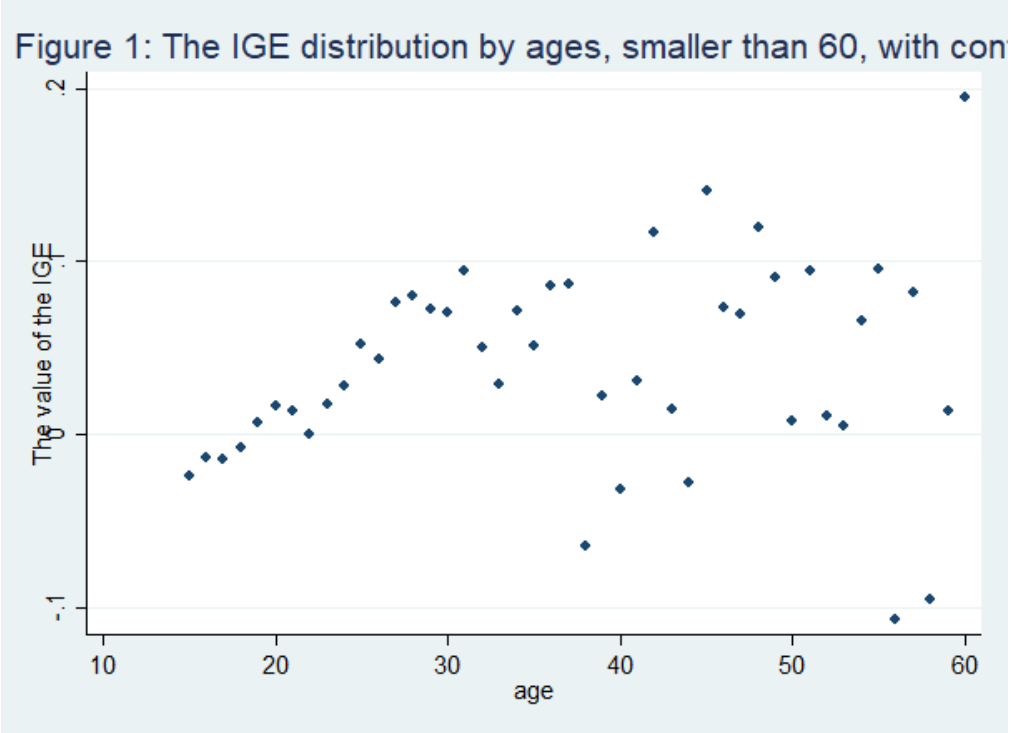


Figure 2: The estimation of the IGE in the CPS by ages with control variables.

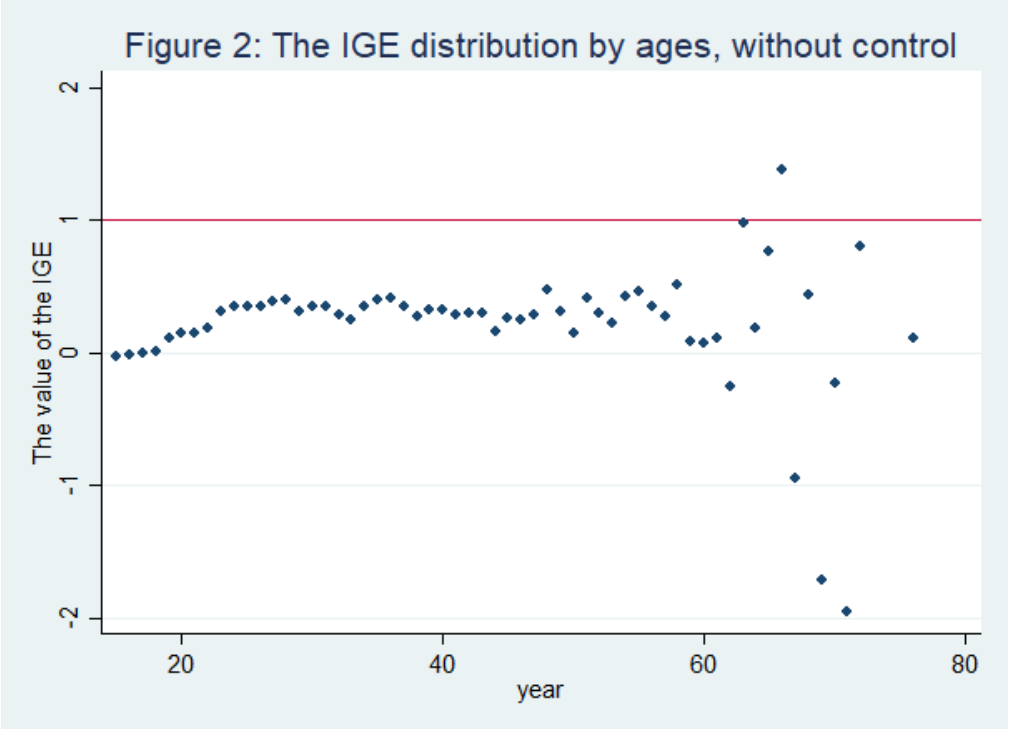


Figure 3: The estimation of the IGE in the CPS with an age range from 18-40 without controls

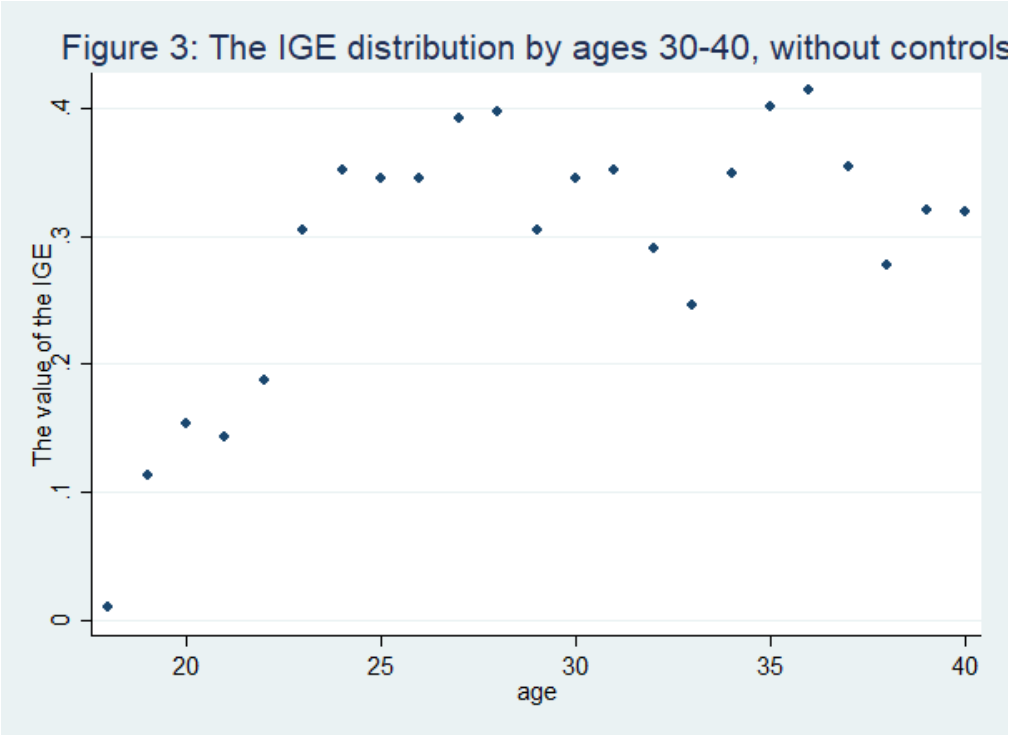


Figure 4: Predicted income and actual income in the CPS, 1% of total samples. Both incomes are standardized

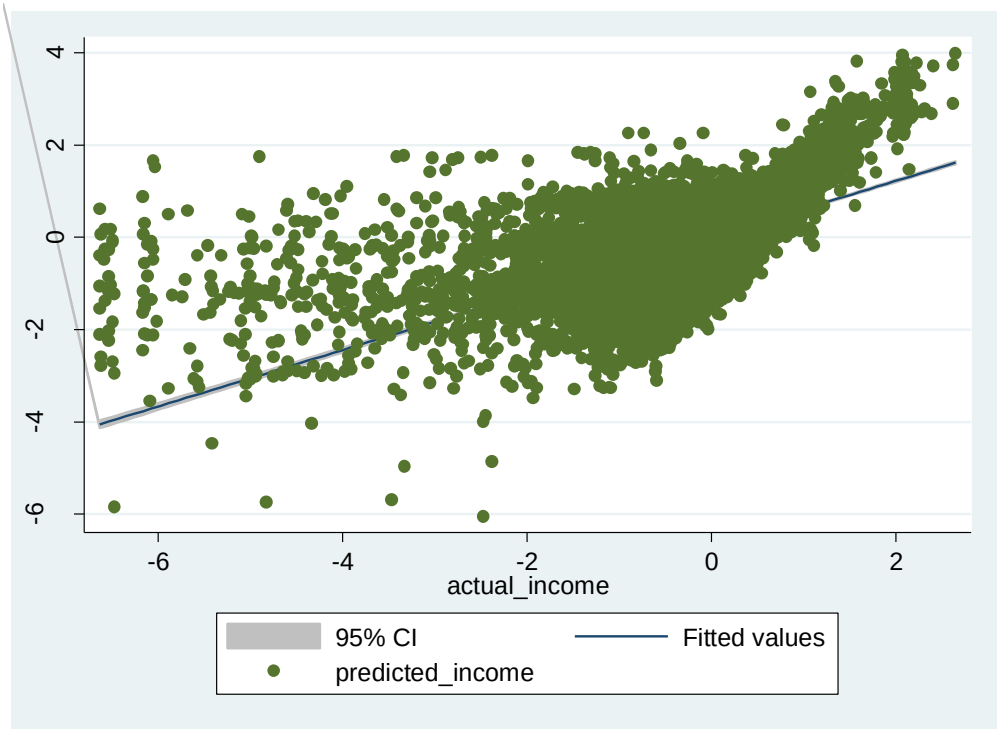
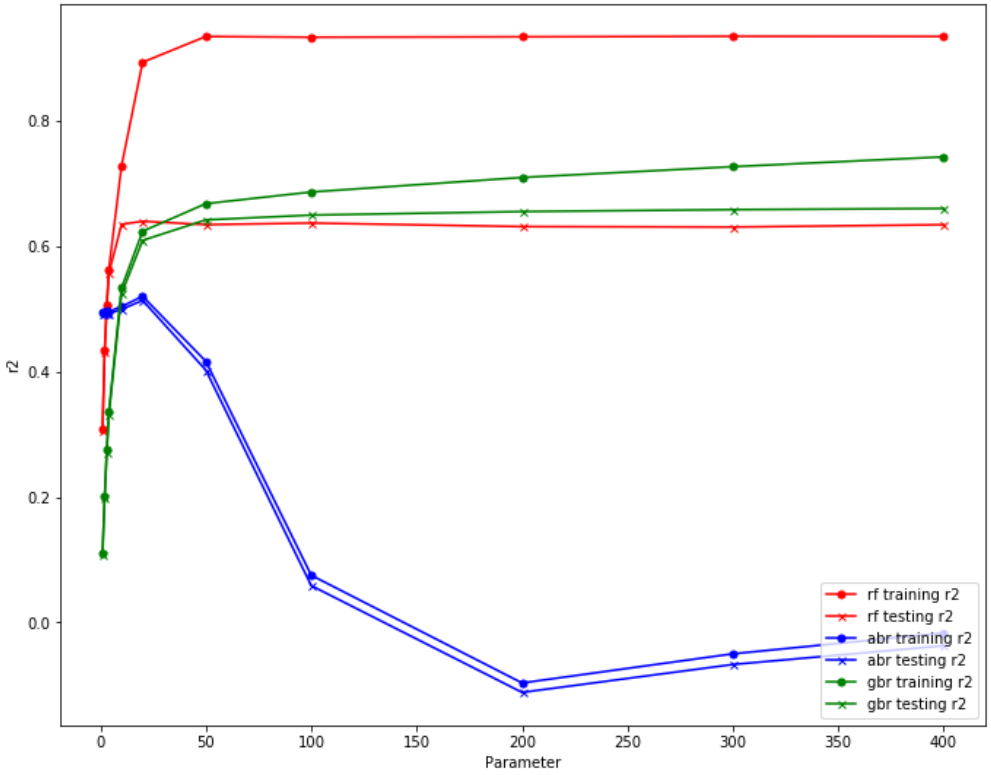


Figure 5: The testing and training R^2 for different machine learning models.
Note: Parameters are used to find a model setting that provide with lowest error summation.



Regression table 1: The regression results for the PSID without controls. (Permanent income is defined as the average income of at least 5 observed years in the PSID)

	(1)	(2)	(3)	(4)	(5)
	All sample	Top 10%	Bottom 10%	Son	Daughter
VARIABLES	lnc	lnc	lnc	lnc	lnc
Father's log income	0.6143	0.2120	1.4030	0.3859	0.7628
	(0.097)***	(0.523)	(1.180)	(0.104)***	(0.148)***
Constant	2.9312	6.7827	-3.5333	5.3623	1.1882
	(0.868)***	(5.060)	(9.247)	(0.932)***	(1.315)
Observations	443	60	32	231	212
R-squared	0.083	0.003	0.045	0.057	0.113

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Regression table 2: The regression results for the PSID with controls. (Permanent income is defined as the average income of at least 5 observed years in the PSID)

	(1)	(2)	(3)	(4)	(5)
	All sample	Top 10%	Bottom 10%	Son	Daughter
VARIABLES	lnc	lnc	lnc	lnc	lnc
Father's log income	0.4940	0.0283	1.2983	0.2743	0.6386
	(0.104)***	(0.509)	(1.229)	(0.111)**	(0.158)***
Child's age	-0.1627	-0.1982	0.0276	-0.1795	-0.1153
	(0.057)***	(0.143)	(0.240)	(0.061)***	(0.085)
Child's edu	0.1045	0.2223	-0.0787	0.1168	0.1253
	(0.035)***	(0.082)***	(0.174)	(0.035)***	(0.058)**
Child's Observed year	0.0009	0.0665	-0.0767	-0.0040	0.0216
	(0.015)	(0.039)*	(0.086)	(0.016)	(0.023)
Constant	5.6495	4.0094	3.8036	8.5502	1.1531
	(1.470)***	(5.347)	(12.171)	(1.492)***	(2.351)
Observations	443	60	32	231	212
R-squared	0.114	0.137	0.089	0.128	0.134

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1: Cross year index in the PSID

Variable Number	Variable Name
1	Income (in 1967 dollars)
2	Yearly rent paid
3	Household head's wife's age
4	Household head's father's education
5	Whether household head is a veteran or not
6	Whether household head's parents grew up in poor?
7	Household head's education
8	Household head's employment status
9	The house is owned or rented
10	Self's marriage
12	Self's age
13	Self's house work time
14	Self's education
15	Self's intention to move
16	Household head's age
17	Interview year
18	Household head's race
19	Household's composition change
20	Self's current state of residence

Table 2: Cross year index in the PSID and CPS

Variable Number	Variable Name
1	Income (in 1967 dollars)
10	Self's marriage
12	Self's age
13	Self's house work time
14	Self's education
17	Interview year
18	Household head's race
20	Self's current state of residence

Summary table 1: For the selected sample from the PSID

Variable	Obs	Mean	Std. Dev.	Min	Max
Father's first recorded year	443	68.13	0.88	68	80
Father's first recorded age	443	31.73	3.82	18	44
Father's marriage status	443	0.99	0.08	0	1
Father's education year	443	11.92	2.82	4	17
Father's working status	443	1.04	0.34	1	5
Father's permanent income	443	8407.48	4266.72	1532.53	34950.50
Child's observed year	443	79.88	3.72	70	92
Child's observed age	443	18.49	0.99	18	22
Child's marriage status	436	0.00	0.10	0	2
Child's education year	443	12.33	1.56	5	17
Child's working status	443	1.21	0.84	1	7
Child's gender	443	1.48	0.50	1	2
Child's permanent income	443	6564.55	5450.22	13.11	48594.02

Summary table 2: For the selected sample from the CPS

Variable	Obs	Mean	SD	Min	Max
Father's age	34,815	51.20	5.65	37	64
Father's education year	34,815	13.85	2.43	0	22
Father's survey year	34,815	106.71	2.79	103	113
Father's dividend	34,815	380.48	2450.52	0	100000
Father's income	34,815	8627.22	2169.85	5186.69	13979.68
Homeownership	34,815	1.10	0.33	1	3
Child's age	34,815	21.97	3.13	18	33
Child's gender	34,815	1.45	0.50	1	2
Child's education year	34,815	1.44	1.17	1	7
Child's race	34,815	13.34	1.73	0	22
Child's dividend	34,815	33.19	598.90	0	38224
Child's income	34,815	6228.71	2823.55	-228.61	47776.50
Father's income in ln form	34,815	9.03	0.25	8.55	9.55
Child's income in ln form	34,814	8.65	0.42	5.94	10.77

Regression table 3: The regression results of the single dummy model, cut-off age is listed at the top. For model 2, 4 and 6, year dummies and education dummies are not included in this table, and some of them are significant.

	(1)	(2)	(3)	(4)	(5)	(6)
Cut-off age	18	18	16	16	16-18	16-18
VARIABLES	lnc	lnc	lnc	lnc	lnc	lnc
Father's log income	0.3353	0.0856	0.2936	0.0693	0.3353	0.0856
	(0.021)***	(0.013)***	(0.014)***	(0.009)***	(0.021)***	(0.013)***
Inc*Time	-0.1294	-0.0403	-0.1035	-0.0275	-0.1294	-0.0403
	(0.023)***	(0.014)***	(0.018)***	(0.011)**	(0.023)***	(0.014)***
Time	1.2313	0.3730	0.9610	0.2574	1.2313	0.3730
	(0.207)***	(0.126)***	(0.167)***	(0.101)**	(0.207)***	(0.126)***
Child_age2		0.0024		0.0024		0.0024
		(0.000)***		(0.000)***		(0.000)***
Child's age		-0.1544		-0.1522		-0.1544
		(0.006)***		(0.006)***		(0.006)***
Child's gender		-0.5523		-0.5524		-0.5523
		(0.003)***		(0.003)***		(0.003)***
Constant	5.5682	10.1842	5.9796	10.3082	5.5682	10.1842
	(0.186)***	(0.141)***	(0.127)***	(0.118)***	(0.186)***	(0.141)***
Observations	34,814	34,814	34,814	34,814	34,814	34,814
R-squared	0.023	0.643	0.020	0.643	0.023	0.643

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Regression table 4: The regression results of the DID model, cut-off age is listed at the top. For model 2 and 4, year dummies and education dummies are not included in this table, and some of them are significant.

	(1)	(2)	(3)	(4)
Cut-off age	18	18	16	16
VARIABLES	lnc	lnc	lnc	lnc
Father's log income	0.3482	0.0979	0.2956	0.0708
	(0.024)***	(0.015)***	(0.016)***	(0.010)***
Inc*Time	-0.1419	-0.0498	-0.1012	-0.0220
	(0.026)***	(0.016)***	(0.021)***	(0.013)*
Inc*Treat	-0.1940	-0.0584	-0.1245	-0.0145
	(0.056)***	(0.034)*	(0.037)***	(0.022)
Inc*Ti*Tr	0.1246	0.0403	0.0485	-0.0200
	(0.062)**	(0.037)	(0.049)	(0.029)
Treat	1.8192	0.5298	1.1857	0.1356
	(0.512)***	(0.310)*	(0.340)***	(0.205)
Time	1.3460	0.4575	0.9450	0.2084
	(0.238)***	(0.145)***	(0.192)***	(0.117)*
Ti*Tr	-1.1559	-0.3622	-0.4682	0.1794
	(0.566)**	(0.342)	(0.443)	(0.268)
cage2		0.0024		0.0024
		(0.000)***		(0.000)***
Child's age		-0.1543		-0.1522
		(0.006)***		(0.006)***
Child's gender		-0.5523		-0.5524
		(0.003)***		(0.003)***
Constant	5.4441	10.0721	5.9527	10.2939
	(0.213)***	(0.154)***	(0.147)***	(0.126)***
Observations	34,814	34,814	34,814	34,814
R-squared	0.025	0.644	0.022	0.643

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Bibliography

Athey, S. (2018). The impact of machine learning on economics. In *the Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Auten, G., Carroll, R., & Gee, G. (2008). The 2001 and 2003 tax rate reductions: An overview and estimate of the taxable income response. *National Tax Journal*, 345-364.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369-2429.

Bleakley, H., & Ferrie, J. (2016). Shocking behavior: Random wealth in antebellum Georgia and human capital across generations. *The quarterly journal of economics*, *131*(3), 1455-1495.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073-1076.

Burke, K. C., & McCouch, G. M. (2007). Turning Slogans into Tax Policy. *Va. Tax Rev.*, *27*, 747.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, *170*(2), 383-398.

Chen, S., Wang, X.Y. (2018). The application of machine learning in social science: review and expectation. Fudan University

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1553-1623.

Couch, K. A., & Dunn, T. A. (1997). Intergenerational correlations in labor market status: A comparison of the United States and Germany. *Journal of Human Resources*, 210-232.

Engstrom, R., Hersh, J., & Newhouse, D. (2017). Poverty from space: using high-resolution satellite imagery for estimating economic well-being.

Hansen, S., McMahon, M., & Prat, A. (2017). Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801-870.

Haider, S., & Solon, G. (2006). Life-cycle variation in the association between current and lifetime earnings. *American Economic Review*, *96*(4), 1308-1320.

Hungerford, T. L. (2011). Changes in the distribution of income among tax filers between 1996 and 2006: The role of labor income, capital income, and tax policy.

Lillard, L. A., & Reville, R. T. (1997). Intergenerational mobility in earnings and occupational status. *Unpublished manuscript, Rand, Santa Monica, CA*.

Minicozzi, A. L. (1999). Nonparametric analysis of intergenerational income mobility.

Mulligan, C. B. (1997). *Parental priorities and economic inequality*. University of Chicago Press.

Iaria, A., Schwarz, C., & Waldinger, F. (2018). Frontier knowledge and scientific production: Evidence from the collapse of international science. *The Quarterly Journal of Economics*, *133*(2), 927-991.

Paravisini, D., Rappoport, V., & Ravina, E. (2016). Risk aversion and wealth: Evidence from person-to-person lending portfolios. *Management Science*, *63*(2), 279-297.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Pekkarinen, T., Uusitalo, R., & Kerr, S. (2009). School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. *Journal of Public Economics*, *93*(7-8), 965-973.

Shea, J. (2000). Does parents' money matter?. *Journal of public Economics*, *77*(2), 155-184.

Solon, G. (1999). Intergenerational mobility in the labor market. In *Handbook of labor economics* (Vol. 3, pp. 1761-1800). Elsevier.

Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(1), 103-106.

Turi Machine Learning Platform User Guide. (n.d.). Retrieved April 8, 2019, from https://turi.com/learn/userguide/supervised-learning/random_forest_regression.html

Yagan, D. (2015). Capital tax reform and the real economy: The effects of the 2003 dividend tax cut. *American Economic Review*, *105*(12), 3531-63.