From Baby Boomers to X Generation: The Evolution of Intergenerational Mobility in Italy

Francesco Bloise (Sapienza University of Rome), Teresa Barbieri (University of Turin), Michele Raitano (Sapienza University of Rome), and Francesca Subioli (Sapienza University of Rome)

This version: January 26, 2023

Long abstract plus main results, please do not circulate

Abstract

The economic literature on the evolution over time of intergenerational inequality – i.e., the association between parents' incomes and children's incomes – is limited because of the usual lack of proper longitudinal datasets tracking a large number of different generations (i.e., individuals born in different cohorts). To deal with the lack of parents' incomes information, the empirical literature has proposed to exploit the two-stage two sample least square (TSTSLS) technique to impute parents' incomes or to estimate intergenerational mobility regressing adult children's incomes on characteristics of their parents associated with their incomes (e.g., education, occupation). However, both methods might not be proper once one aims to estimate how intergenerational mobility has evolved along many cohorts (e.g., among children born in the different post World War II decades). On the one hand, repeated cross-sectional surveys – needed to carry out TSTSLS estimates – usually do not cover periods far back in time (e.g., before World War II). On the other hand, the capacity of the various parents' characteristics to proxy incomes and, more generally, to capture parental chances in transmitting economic advantages to offspring might have changed over time. This risk affects the reliability of estimates about the association between adult children's incomes and some parents' characteristics when one compares this association across many children's birth cohorts.

To overcome both issues, in this paper, we propose a new methodology based on machine learning techniques to build a distribution of parents' socioeconomic status when parents' incomes are missed. In detail, exploiting ML algorithms, we compute a data-driven distribution based on all background characteristics (e.g., education, occupation, family composition) that can predict adult children's earnings. This methodology thus allows us to summarize intergenerational mobility through a rank-rank regression of children's income ranks on the ranks of this surrogate distribution, which is based on children's predicted income rankings according to parents' characteristics. In addition, our methodology allows us to compare the extent to which parental characteristics explain children's incomes across cohorts.

To validate the capacity of our methodology to capture the association between children and parents' earnings, we use PSID data for the US and compare results about the rank-rank slopes obtained, regressing children's income percentiles with percentiles of the surrogate distribution, with rank-rank slopes obtained by using actual parents' incomes or parents' incomes predicted through the standard TSTSLS approach.

Once our methodology is validated, we use it to investigate the evolution over time of intergenerational mobility in Italy, a country characterized by low values of such mobility in international comparison. The administrative data we have at our disposal, tracking individual working lives from 1975 to 2018 and recording characteristics of the parents of these individuals, allow us to investigate how the association between children's labor earnings and parents' characteristics has evolved across many cohorts, namely from those born in 1945 up to those born in 1979.

Moreover, our dataset allows us to observe children's (males and females) earnings when they are aged from 30 to 40, thus allowing us to obtain estimates not affected by temporary income fluctuations and life-cycle bias.

We find that intergenerational mobility in Italy has followed an inversed U-shaped decreasing pattern over the second half of the XX century, even if a decrease in mobility is emerging for the youngest cohorts. When we estimate the rank-rank association, also adding some children's characteristics that might mediate the association between children's earnings and family background (e.g., education, occupation, region of birth), our measures of intergenerational mobility remain positive and statistically significant, and no changes in the time trend emerge. However, the inversed U-shaped trend disappears when we control for the number of positive children's earnings observations over the 11 years, thus suggesting that the influence of parental status on children's unemployment risk has changed over time.

Finally, observing children's earnings for an 11-year time window, we also estimate the association between measures of children's income growth and income volatility along this time window and our surrogate distribution of parents' characteristics, thus providing some insights about the link between intergenerational and intragenerational mobility. This issue has yet to be addressed in the current literature. Interestingly, we find more substantial income growth and lower volatility for workers from more advantaged backgrounds. At the same time, no clear-cut time trend in the association between measures of intragenerational mobility and parental status emerges.

In what follows, we include detailed slides of the paper where we briefly review the related literature, discuss the primary empirical limits to estimate intergenerational mobility along many cohorts, propose and explain our method to construct a surrogate distribution of parents through ML techniques and show main results about Italy. The validation test of our methodology using PSID data has been carried out but has still to be added to the paper.

Keywords: Intergenerational mobility, Italy, machine learning

JEL No. J24, J31, J62

Outline

Data requirements and suboptimal data

- Global evidence on intergenerational mobility is often based on low-quality data
- In many cases the available observations do not permit to establish a direct longitudinal parent-child link
- This limitation is of particular relevance for developing countries and for historical analyses of mobility in societies at various stages of economic development
- Amongst the existing estimates of intergenerational mobility, a significant number (virtually all of those for lower-income countries) are obtained through the TSTSLS methodology introduced by Björklund & Jäntti (1997)
- Since background information (e.g., parental education, occupation, area of birth, etc.) is more likely to be available in cross-sectional datasets compared to parental income, the TSTSLS methodology allowed the estimation of intergenerational income mobility for a significantly larger number of countries and historical periods
- Recent evidence shows that ML can improve the reliability of intergenerational mobility estimates based on suboptimal data (Bloise et al., 2021a)

Motivation

- Italy shows **less intergenerational income mobility** than most other countries with similar degrees of development (Björklund & Jäntti, 2009; Blanden, 2013; Corak, 2013), as concerns both the intergenerational elasticity and the Rank-Rank correlation (Barbieri et al., 2020, Bloise & Raitano, 2020)
- No studies have estimated **trends in economic mobility** in Italy
- The capability of parents to pass on economic advantages to their children might have changed over time because of the significant economic and social transformation that occurred over the last century in Italy

Contribution

- By exploiting the longitudinal AD-SILC dataset, we focus on the Italian case and on two underinvestigated issues in the literature on intergenerational mobility:
  1. **the evolution of intergenerational mobility across cohorts** - i.e., the association between parental background and adult children earnings – across 31 moving workers birth cohorts in Italy from 1945 to 1979
  2. **the link between parental background and intragenerational mobility** (i.e., earnings mobility patterns from age 30 to age 40)
- To overcome the lack of information on parental income across different cohorts, we rely on **machine learning (ML) techniques** to better proxy parental background
- By proposing and developing a data-driven ranking of parental background we contribute to the literature on intergenerational mobility in two ways:
  1. Using ML we are not forced to rely on arbitrary indexes of parental background or on a single proxy, whose role may have changed over time (e.g., parental education, parental occupation)
  2. ML allows to build a reliable background distribution for most of countries with unavailable longitudinal data on parents-children pairs from different birth cohorts

Previous evidence

- Because of data limitations, very few papers have investigated trends of intergenerational mobility over time:
  1. For the US a mixed evidence has emerged: if some studies do not detect an increase in intergenerational income mobility over the second half of the 20th century (Hertz, 2007; Lee & Solon 2009; Chetty et al 2014), others show that intergenerational mobility declined sharply (Davis & Mazumder 2017)
  2. Income mobility declined in UK for cohorts between 1950s and 1970s (Blanden et al.,2007; Nicoletti & Ermisch, 2008)
  3. Income mobility appears to have increased in Nordic Countries (Bratberg et al. 2007; Björklund et al. 2009)
  4. Rising intergenerational income persistence in China from the 1970–1980 birth cohort to the 1981–1988 birth cohort (Fan et al., 2021)
  5. Bukowski et al. (2022) find no relevant changes in social mobility in Hungary by exploiting rare surnames as a measure of social status
- No studies about the evolution of intergenerational income mobility in Italy both considering incomes or other parents' characteristics as background variables

Building a background distribution on suboptitmal data

- How to rank children according to their background when information on parental income is unobserved? Can we use a single parental feature?
- Raitano and Vona (2015) and Bloise et al. (2021b) build a distribution of parental background using information on parents' characteristics in a hierarchical order
- Basing on the socioeconomic literature (e.g. Granovetter, 1995), they rank individuals by deciding "ex-ante" which parental characteristics are more relevant in influencing children economic outcomes when adult
- They first take father and mother occupation, respectively, as a good proxy for the influence of the family on children's outcomes as it encompasses unobservable aspects of human capital, socio-economic status and family networks.
- Previous studies take, in a hierarchical order, all other parents' characteristics provided by EU-SILC: father and mother education, country of birth of fathers and mothers, the presence of both parents in the household, the number of siblings and the number of income recipients in the household

Issues in the previous approach

- The previous approach exploited to build a background distribution might have two basic issues:

1. Within a selected birth cohort: it is difficult to decide ex-ante which background variable (and which categories of a given background variables) are more relevant in predicting economic success of a children when adult
2. The influence of a specific background category (i.e., having a tertiary graduated mother) on children's income level is likely to change over time due to an interplay of both composition and price effects: the structure of the population in terms of educational level change over time
3. This is why, we use ML to select background categories that are relevant for economic success of children in a specific birth cohort and we build the parental background distribution accordingly

Using ML to rank individuals according to parental background

- We first assume an unknown data generating process for log income of children i from birth cohort b:

$$y_{i,b} = f_b\left(back_{i,b}\right) + \varepsilon_{i,b} \tag{1}$$

- Here f_b is some fixed but unknown function of $back_1, back_2..., back_p$, and $\epsilon_{i,b}$ is the error term
- As f_b describes an unknown data generating process, we should not assume that the relationship between log $y_{i,b}$ and background characteristics must be linear. We could also have polynomials, pairwise or high-order interactions between regressors
- Machine learning algorithms are thus exploited to select and order those background categories included in the vector $back_{i,b}$ maximizing the 'out-of-sample' capability of the function f_b to predict economic success of children when adult
- We thus select the model which maximize the 'out-of-sample', rather than the 'in-sample' MSE, in order to minimize overfitting and multicollinearity issues

The variance-bias trade-off

- In the statistical learning framework, the Mean Squared error (MSE) can be decomposed as follows:

$$MSE = \mathbb{E}\{f(x_i) - \mathbb{E}[f(x_i)]\}^2 + \{\hat{f}(x_i) - \mathbb{E}[f(x_i)]\}^2 + \sigma^2 \quad (2)$$

- The first term is the **variance** of the model; the second term is the **square bias**; the last is the noise term, which cannot be reduced
- The bias is a measure of the distance between the expected value of the prediction and the unknown function which captures the true relationship between the outcome variable and predictors. It derives from the fact that we generally approximate a complex date generating process using a simple function
- The variance is the expected variability of a model prediction around its expected value. it captures the model sensitivity to different samples
- Very complex models will tend to have low bias and large variance (**overfitting**). On the other hand, simple models are characterized by high bias and low variance (**underfitting**).

How we tune ML algorithms: the K-fold cross-validation

- This approach involves randomly dividing the set of observations into k equally-sized groups, or folds.
- The first fold is treated as a validation set, and the model is estimated on the remaining K-1 folds.
- The mean squared error, MSE_1, is then computed on the observations in the held-out fold.
- This procedure is repeated K times. Each time, a different group of observations is treated as a validation set.
- This process results in k estimates of the prediction error, $MSE_1, MSE_2, ..., MSE_k$
- The K-fold CV estimate is computed by averaging these K values

Assessing the predictive accuracy of our models

- To assess the relative prediction ability of our models we follow Mullainathan and Spiess (2017) and proceed in four steps:

1. We randomly split the sample into two subsets: the training set (70% of the observations) and the test set (30%).
2. Our methods need regularization/complexity choice: we select the most appropriate set of parameters by 5-fold cross-validation.
3. We then use the tuning parameters that produce the smallest cross-validation MSE to estimate the prediction model and we store the prediction function
4. We then estimate the prediction error "out-of-sample" on the test sample for non-tuned learners stored in step 3

Estimating intergenerational mobility on suboptimal data (1)

- Intergenerational mobility can be estimated using the rank correlation (Rank-Rank slope), that is the association between the relative position of parents and children in their respective generation
- Absent parental earnings, we obtain a predicted distribution of earnings using ML and information recalled by children about their parents
- In other terms, parental background is measured as the capability of background characteristics to predict children's earnings and, thus, their rank along a parental background distribution
- To predict the background rank, we train and calibrate different potential algorithms. Specifically, among all possible state of the art Learners, we train:
  1. LASSO, Elastic Net for dimension reduction
  2. Random Forest and XgBoost for dimension reduction and to take into account possible high-order non-linearities among regressors

Estimating intergenerational mobility on suboptimal data (2)

- In the second step, we estimate the intergenerational correlation for cohort b using the following equation:

$$Rank(y_{i,b}) = \alpha + \rho_b Rank(\hat{y}_{i,b}) + u_{i,b} \quad (3)$$

- $Rank(y_{i,b})$ is the earnings percentile of a child from cohort b
- $Rank(\hat{y}_{i,b})$ is the percentile on the predicted distribution of children's earnings calculated by exploiting a high-dimensional vector of background characteristics
- To minimize the prediction error, we use the full set of background characteristics of fathers and mothers provided by the 2005 wave of IT-SILC

Data and sample selection

- We exploit the AD-SILC dataset, built matching the 2005 IT-SILC wave with information from Italian National Social Security Institute Archives (INPS), where individuals are followed from the labour market entry up to 2019
- We use the IT-SILC 2005 cross section where detailed information on parents' characteristics are recorded
- Male and female workers are grouped in 31 moving birth cohorts: 1945-1949, 1946-1950, 1947-1951, 1948-1952 and so on until the 31st cohort 1975-1979
- We observe workers real gross annual earnings from employment and self-employment from age 30 to age 40 to minimize life cycle bias and depurate from temporary income fluctuations (Haider & Solon, 2005; Nybom & Stuhler, 2016)
- In baseline estimates we average earnings also considering years with zero income

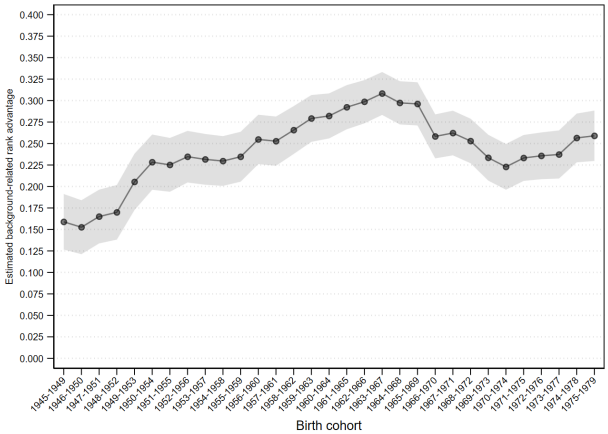
An excerpt of our sample

Cohort	No. Observations	Historical period in which earnings are observed
1945-1949	2,206	1975-1989
1950-1954	2,186	1980-1994
1955-1959	2,729	1985-1999
1960-1964	3,196	1990-2004
1965-1969	3,455	1995-2009
1970-1974	3,209	2000-2014
1975-1979	2,725	2005-2019

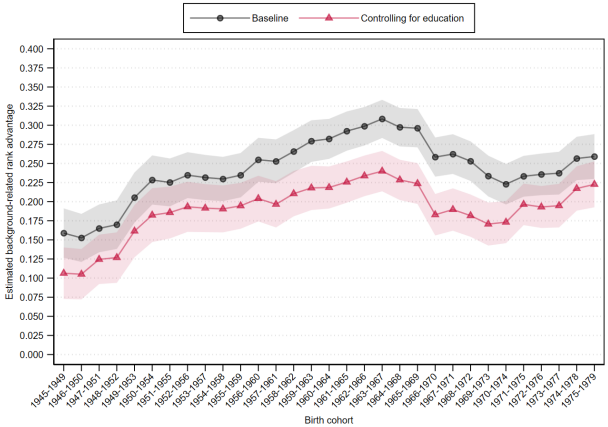
Background categories

Background variable definition	Background variable name	Nr. of binary variables
Family composition	Back1	4
Number of siblings	Back2	7
Highest ISCED level attained by the father	Back3	6
Highest ISCED level attained by the mother	Back4	6
Activity status of the father	Back5	5
Occupation of the father	Back6	11
Activity status of the mother	Back7	5
Occupation of the mother	Back8	10
Financial problems in the household when young teenager	Back9	5
Total	9	59

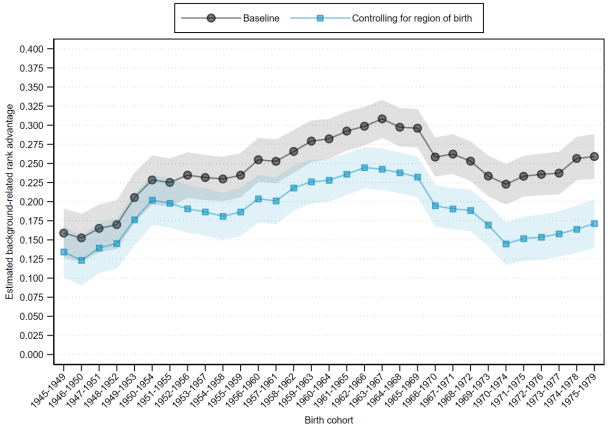
Preliminary results



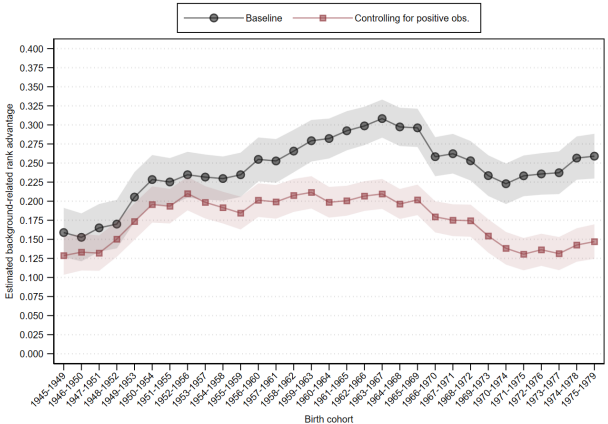
Preliminary results: controlling for education



Preliminary results: controlling for region of birth



Preliminary results: controlling for the number of non-zero observations between 30 and 40 years old



Linking parental background and earnings dynamics

- Up to now we have measured children outcome as the average income during a relevant phase of the career – age 30-40
- Now we want to extend the concept of outcome looking at children income path
- Does a good parental background affect the probability of having a faster and less volatile career progression?
- Now we test whether a better parental background is associated with better performance for children in terms of income dynamics
- We consider a "better" – more desirable – income path one that is growing faster but less fluctuating:

  "[. . .] it seems likely that individuals tend to prefer a constant income stream, or one which is growing steadily, to one which continually fluctuates" Shorrocks (1978a, p. 392)

How to Measure intragenerational mobility

- We adopt two indicators to measure the characteristics of individual income paths:
  - ≫ **Individual income growth**:*computed as the percentage change from starting income and ending income*:

$$IncomeGrowth_{i.b} = \frac{\frac{1}{3}\sum_{a=38}^{40} y_{i,b,a}}{\frac{1}{3}\sum_{a=30}^{32} y_{i,b,a}} - 1$$

  - ≫ **Income volatility** *as the individual standard deviation of the percentage change in income from one year to the next*:

$$Volatility_{i,b} = \sqrt{\frac{1}{11}\sum_{a=30}^{40}\left[\left(\frac{y_{i,b,a}}{y_{i,b,a-1}} - 1\right) - \frac{1}{11}\sum_{a=30}^{40}\left(\frac{y_{i,b,a}}{y_{i,b,a-1}} - 1\right)\right]^2}$$

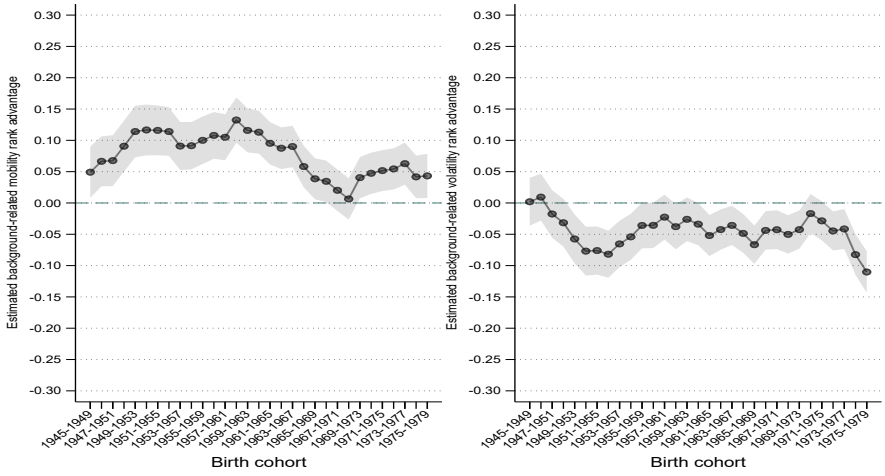
Estimating the association between parental background and intragenerational mobility

- We build for both income growth and volatility a within-cohort ranking
- Using our measure of parental background, we regress separately the rank of children in terms of mobility and volatility on the rank in terms of parental background:

$$Rank(Growth_{i,b}) = \alpha + \delta_b Rank(\hat{y}_{i,b}) + \omega_{i,b} \tag{4}$$

$$Rank(Volatility_{i,b}) = \alpha + \gamma_b Rank(\hat{y}_{i,b}) + \lambda_{i,b} \tag{5}$$

Preliminary results: the link between parental background and intragenerational mobility



Conclusions

- In the second half of the last century, the intergenerational association between children's earnings and parental background followed an inverted U-shaped pattern
- Education and occupation have little explanatory power
- When we control for the number of positive observations we can observe a significant reduction in intergenerational persistence: those coming from more disadvantaged backgrounds are more likely to have unstable careers
- For those coming for more advantaged background we observe faster and less volatile careers