Measuring Income Inequality of Opportunity

Focusing on Early Childhood Circumstances

Aman Desai
CUNY Graduate Center

PSID Annual User Conference 2024

12 September, 2024

Motivation

"The rise in inequality in the United States over the last three decades has reached the point that inequality in incomes is causing an unhealthy division in opportunities, and is a threat to our economic growth" (Alan Krueger, Center for American Progress, 12 January 2012)

Rigorous treatment to measurement of inequality of opportunity (IOp hereafter) is vital from policy perspective.

Contribution

- Categorization of circumstance and effort factors using the age of consent at 18 years.
- Accounting for the role of dynamic complementarity by constructing age-based circumstance sets in measuring the inequality of opportunity.
- Using supervised machine learning to construct counterfactual distribution of adult incomes based on circumstances.

Main Results

- About 32% of total inequality in an individual's adult income could be attributed to unequal circumstances faced in their childhood before or at age 5.
- This share rises to as high as 37% before the child reaches adulthood at age 18.

Inequality of Opportunity

- Seminal work by Roemer (1993). Success in adult life is considered to be influenced by
  - **Circumstance** : Beyond individual's control, hence for those the individual should not be held responsible and should be compensated for inequalities generated due to those.
  - **Effort** : Individual is in control of their effort and hence should be rewarded in the market economy.

Technology of Skill Formation

- Based on work by (Cunha and Heckman 2007; Cunha and Heckman 2009)
  - **Dynamic Complementarity** : Returns to investment in human capital at later stage in life is low if investment in early stage is low.

Inequality of Opportunity

Consider a population $\mathscr{N} = \{1, 2, \ldots, N\}$. Each individual in the population is characterized by a triple (y, C, e) where $C \in \Omega^c$, $e \in \Omega^e$, and $y = g(C, e)$, with $g : \Omega^c \times \Omega^e \Longrightarrow \mathbb{R}$.

- An individual in the population is identified by a *type* and a *tranch*.
- A *type* consists of individuals with the same circumstances beyond their control.
- A *tranch* consists of individuals with the same effort.
- According to Roemer, equality of opportunity is achieved when inequality generated due to differential circumstances is eliminated (between *types*),that is $F(y|C) = F(y)$.
- Inequality of opportunity is measured by the extent to which this principle is violated, that is $F(y|C) \neq F(y)$.

Existing Empirical Work

- Several empirical approaches in last twenty years. (Bourguignon, Ferreira, and Menéndez 2007; Pistolesi 2009; Ferreira and Gignoux 2011; Niehues and Peichl 2014; Hufe et al. 2017). The estimated shares of IOp in outcome inequality varies largely from 10% to as high as 70%.
- Usage of machine learning algorithms to model IOp (Brunori, Hufe, and Mahler 2023).
- Fixed set of circumstances where contingency on normative judgments is greatly amplified.
- Lower bound measures of IOp.

Technology of Skill Formation

Cunha and Heckman (2007) model technology for skill formation, conceptualized as a law of motion.

$$\omega_{i,t+1} = f(\omega_{i,t}, x_{i,t}, \omega_i^p, \epsilon_{i,t}) \tag{1}$$

- $f(.)$ is assumed to be twice continuously differentiable, increasing in all arguments, and concave in $x_{i,t}$.
- $x_{i,t}$ is the parental investment for the child i at age t.
- ω_i^p is parental human capital at time t.
- $\epsilon_{i,t}$ is an iid unobserved individual component.

Insight

Investment in period $t+k$ and investment in any prior years t are always complements as long as $\omega_{i,t+k}$ and $x_{i,t+k}$ are complements.

Idea

If a child can not consent before the age of 18, all the measurable data on the child including her achievements, before she turns 18, can be thought of beyond her control and hence should be considered circumstances.

Critical Stages in Childhood

To incorporate the idea of dynamic complementarity, age cutoffs are chosen based on critical stages in childhood.

- 2 years : A child starts to speak.
- 5 years : A child enters K-12 system.
- 14 years : A child enters high school.
- 18 years : A child becomes an adult and can consent.

Four datasets are constructed according to four age cutoffs.
i.e. $C^2 \subseteq C^5 \subseteq C^{14} \subseteq C^{18} \subseteq \Omega^c$

Data

Ideally, one would have an entire biography of the individual's childhood experiences.

Analytical Sample

- Database : Panel Study of Income Dynamics(Main Interview, FRM[1], FIMS[2]).
- Cohorts : 1978-1983 (restricted to SRC[3] sample).
- Number of Individuals : 639.
- Types of Factors : Demographic, Monetary/Market, Government/Community.
- Outcome Variable : Individual labor income at age 35 years[4].

The data in consideration is in wide format. Every observation reflects information on measurable factors for an individual over the first 18 years of their life.

[1] Family Relationship Matrix.

[2] Family Identification Mapping System.

[3] Data are restricted to the individuals in Survey Research Center sample to ensure representativeness of the population.

[4] Individual labor income excludes farm and unincorporated business income. All monetary variables including adult incomes are adjusted to 2018 dollars and individual longitudinal weights from 2012-2018 are used in the analyses.

Estimation

The data generating process:

$$y = h(C, e) = f(C) + u = E(y|C) + u \tag{2}$$

- $E(y|C)$ represents the variation in outcome due to observed circumstances.
- u is an iid residual term that captures the variation due to both unobserved circumstances and individual effort.
- Lower bound interpretation of IOp.

Estimation

Parametric Specification (Bourguignon, Ferreira, and Menéndez 2007; Ferreira and Gignoux 2011; Niehues and Peichl 2014)

$$ln(y_i) = \alpha_0 + \sum_{l=1}^{L} (\alpha_l C^s_{i,l}) + u_i \tag{3}$$

where y is the adult income, C is the collection of factors that are categorized as circumstance belonging to a finite set Ω^c, $s \in \{2, 5, 14, 18\}$ reflecting four different sets of circumstances based on chosen age cutoffs.

$$\hat{y}_i = exp\left[\alpha_0 + \sum_{l=1}^{L} (\hat{\alpha}_l C^s_{i,l})\right] \tag{4}$$

IGE

Estimation

The measurement of inequality of opportunity can be thought of as a two-step procedure: first, the actual distribution of y_i is transformed into a counterfactual distribution (obtain $\hat{y}_i$) that reflects only and fully the unfair inequality in y_i, while all the fair inequality is removed. In the second step, a measure of inequality[5] is applied to $\hat{y}_i$. I use mean logarithmic deviation as an inequality measure[6].

$$Absolute\ IOp = I(\hat{y}_{EA}) \tag{5}$$

where $I(\hat{y}_{EA})$ is the ex-ante measure of inequality of opportunity.

$$Relative\ IOp = \frac{I(\hat{y}_{EA})}{I(y)} \tag{6}$$

The value of relative IOp ranges from 0 to 1. If all income differences are solely due to circumstances, relative IOp will be 1.

[5] any standard measure of inequality that satisfies anonymity, the principle of transfers, population replication, and scale invariance.

[6] $MLD(x) = ln(\bar{x}) - \overline{ln(x)}$.

Estimation

I use supervised machine learning methods : Regression Trees, Random Forest.

Algorithm

I fit the models on training data, tune the hyper parameters on validation data, and then use the best model(with the lowest rmse) on the full data set. The algorithm runs as follows:

- Execute the random forest algorithm and use 5-fold cross validation for hyperparameter tuning. Select the models with hyperparameters that yield the lowest *rmse*. In each fold, the data is divided into $N_{train} = \frac{4}{5}N$ and $N_{validation} = \frac{1}{5}N$.
- Store the prediction functions $\hat{f}_{train}(\hat{\Omega}^c)$.
- Obtain final predictions using the full data $\hat{y} = \hat{f}_{train}(\hat{\Omega}^c_{fulldata})$.

Descriptive Statistics

Characteristic	N = 639
Individual labor income at age 35 (in log)	10.80 (10.32, 11.20)
Total family income at age 1 (in log)	11.09 (10.62, 11.39)
Sex	
Male	311 (49%)
Female	328 (51%)
Race	
White	556 (87%)
Black	70 (11%)
AIAE	6 (0.9%)
Other	7 (1.1%)
Occupation of the head at age 1	
Inap	58 (9.1%)
Professional, Technical, and Kindred Workers	152 (24%)
Managers and Administrators, except Farm	63 (9.9%)
Sales Workers	22 (3.4%)
Clerical and Kindred Workers	29 (4.5%)
Craftsman and Kindred Workers	152 (24%)
Operatives, except Transport	79 (12%)
Transport Equipment Operatives	24 (3.8%)
Laborers, except Farm	21 (3.3%)
Farmers and Farm Managers	11 (1.7%)
Farm Laborers and Farm Foremen	2 (0.3%)
Service Workers, except Private Household	26 (4.1%)
Years of education of the head at age 1	12.44 (12.00, 15.00)
Years of education of the spouse at age 1	12.0 (12.0, 14.2)

[1] Median (Q1, Q3); n (%)

Note:

Age referes to the age of the invidual when she was a child.

Head refers to the head of the family the child grew up in during the childhood.

Spouse refers to the spouse of the family head.

Figure 1: Descriptive Statistics for Selected Variables

Table 1: Absolute IOp Estimates for Different Circumstance Sets

	Income Inequality	Absolute IOp	N
Fixed Set of Circumstances			
Baseline	0.337	0.069	639
Age-based Circumstances			
2	0.337	0.089	639
5	0.337	0.107	639
14	0.337	0.117	639
18	0.337	0.124	639

- Fixed circumstances include individual's sex, race as well as the occupation of the family head, total family income, education of the head and the spouse (all measured when child's age is 1).

Results

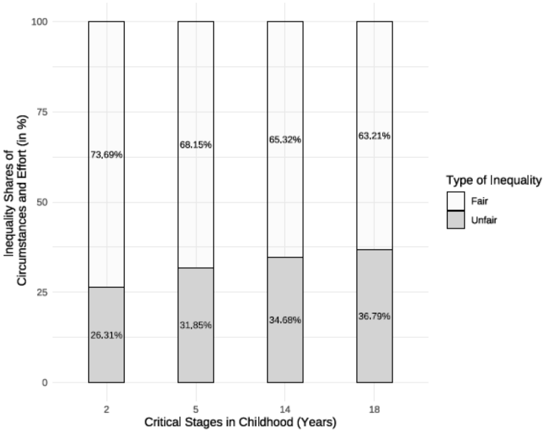

Figure 2: Share of IOp in Total Income Inequality (MLD)

▸ Gini ▸ Average

Results

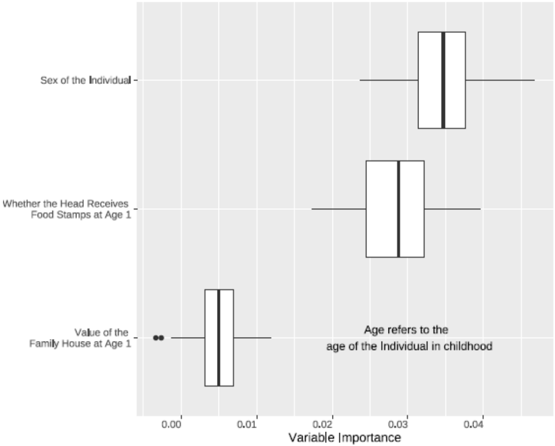

Figure 3: Variable Importance Plot of Circumstances at Age 5

Results

Intergenerational Income Elasticity

- In recent times, policy discussions have shifted from inequality of outcome to inequality of opportunity, informed by intergenerational mobility(Corak 2013; Chetty et al. 2014).
- IGE is measured as a coefficient in a Galtonian regression of a child's income on parental income.

$$ln(y_{child}) = \alpha + \beta_{IGE}\ ln(y_{parent}) + u \tag{7}$$

- Recent evidence suggest that the timing of parental income measured may be as or more important than a single measure of parental income (Carneiro et al. 2021).

▸ IOp

Results

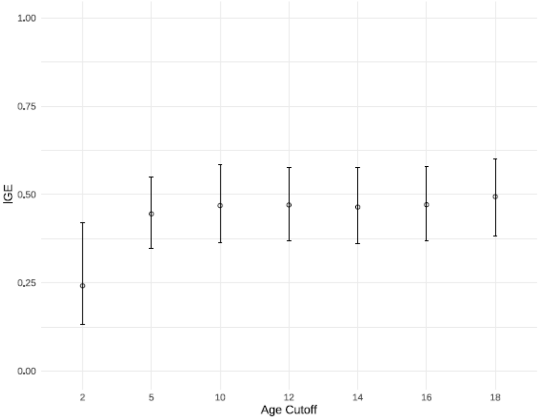

Figure 4: IGE Estimates

Results

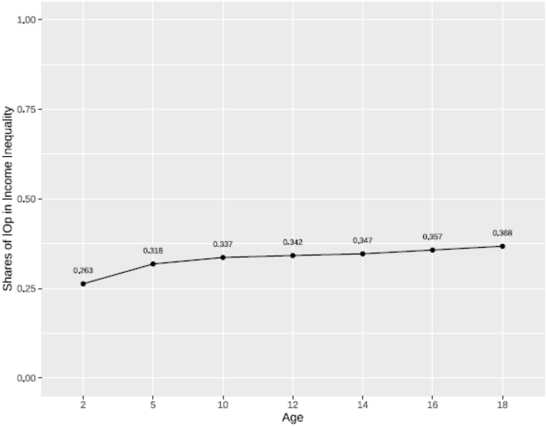

Figure 5: Relative IOp for Different Age Cutoffs

Results

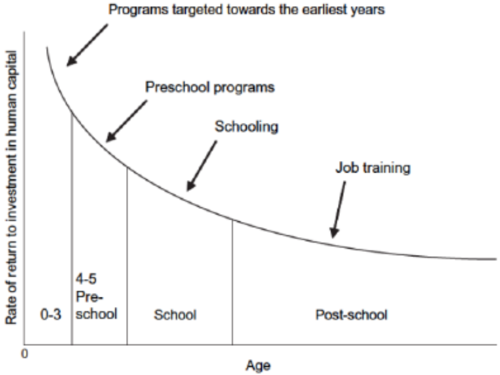

Figure 6: Source : Heckman Equation

Conclusion

- The inequality of opportunity is evaluated via role of the circumstances in the childhood.
- Lower bounds of IOp, as one might argue about the persistent effects of childhood circumstances in achieving success in the adulthood.
- Inequality stemming from unequal circumstances account for 32% of total income inequality.
- Childhood skill gaps resulting from unequal circumstances often persist into adulthood. Therefore, including unequal childhood circumstances in the measurement of inequality of opportunity (IOp) is valuable.

References I

Bourguignon, François, Francisco H. G. Ferreira, and Marta Menéndez. 2007. "Inequality of Opportunity in Brazil." *Review of Income and Wealth* 53 (4): 585–618. https://doi.org/10.1111/j.1475-4991.2007.00247.x.

Brunori, Paolo, Paul Hufe, and Daniel Mahler. 2023. "The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees and Forests*." *The Scandinavian Journal of Economics* 125 (4): 900–932. https://doi.org/10.1111/sjoe.12530.

Carneiro, Pedro, Italo López García, Kjell G. Salvanes, and Emma Tominey. 2021. "Intergenerational Mobility and the Timing of Parental Income." *Journal of Political Economy*, March. https://doi.org/10.1086/712443.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *Quarterly Journal of Economics* 129 (4): 1553–623.

Corak, Miles. 2013. "Income Inequality, Equality of Opportunity, and Intergenerational Mobility." *Journal of Economic Perspectives* 27 (3): 79–102.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97 (2): 31–47. https://doi.org/10.1257/aer.97.2.31.

References II

Cunha, Flavio, and James J. Heckman. 2009. "The Economics and Psychology of Inequality and Human Development." *J Eur Econ Assoc* 7 (2): 320–64. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2832600/.

Ferreira, Francisco H. G., and Jérémie Gignoux. 2011. "The Measurement of Inequality of Opportunity: Theory and an Application to Latin America." *Review of Income and Wealth* 57 (4): 622–57. https://doi.org/10.1111/j.1475-4991.2011.00467.x.

Hufe, Paul, Andreas Peichl, John Roemer, and Martin Ungerer. 2017. "Inequality of Income Acquisition: The Role of Childhood Circumstances." *Soc Choice Welf* 49 (3-4): 499–544. https://doi.org/10.1007/s00355-017-1044-x.

Niehues, Judith, and Andreas Peichl. 2014. "Upper Bounds of Inequality of Opportunity: Theory and Evidence for Germany and the US." *Social Choice and Welfare* 43 (1): 73–99. https://www.jstor.org/stable/43662521.

Pistolesi, Nicolas. 2009. "Inequality of Opportunity in the Land of Opportunities, 1968–2001." *J Econ Inequal* 7 (4): 411–33. https://doi.org/10.1007/s10888-008-9099-7.

Roemer, John E. 1993. "A Pragmatic Theory of Responsibility for the Egalitarian Planner." *Philosophy & Public Affairs* 22 (2): 146–66. https://www.jstor.org/stable/2265444.

IOp shares in Total Inequality using MLD

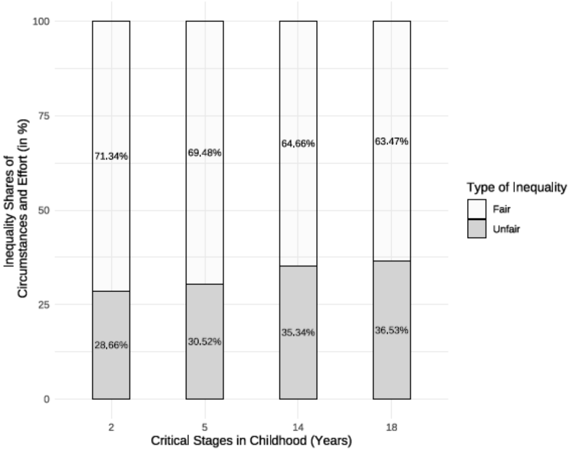

Figure 7: Share of IOp in Total Income Inequality (Adult incomes are averaged across 2012-2018 waves)

Main

IOp shares in Total Inequality using Gini

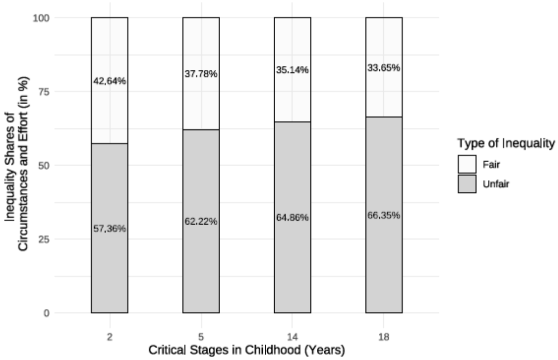

Figure 8: Share of IOp in Total Income Inequality

Main

IOp shares in Total Inequality using Gini

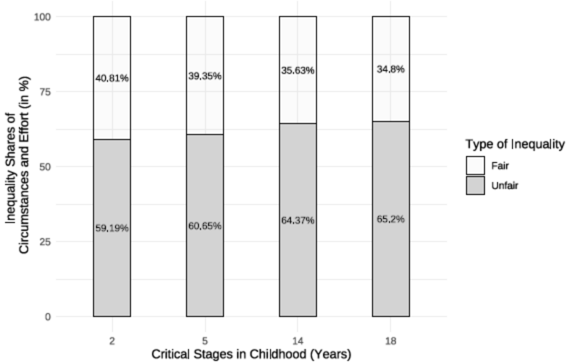

Figure 9: Share of IOp in Total Income Inequality (Adult incomes are averaged across 2012-2018 waves)

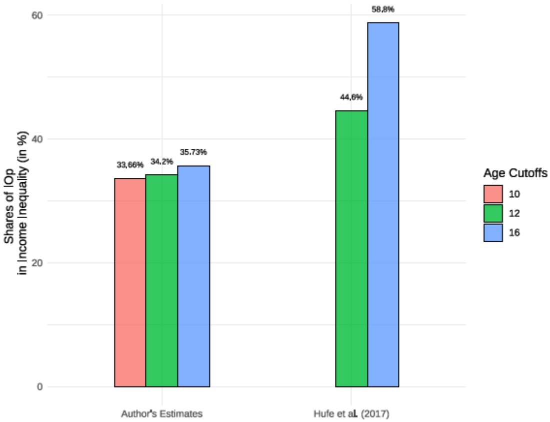

Figure 10: Relative IOp Estimates for Different Age Cutoffs

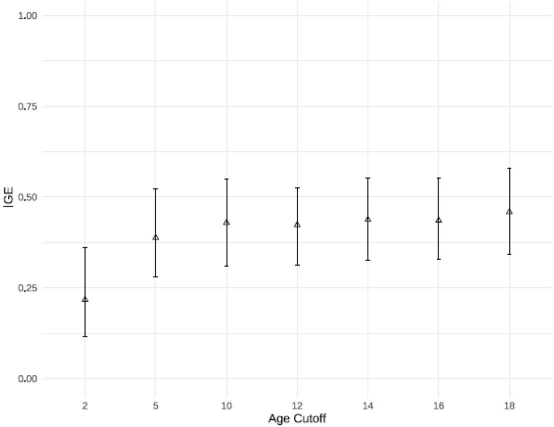

Figure 11: IGE Estimates Adjusted by Gender and Race

Regression Trees

A regression tree algorithm makes predictions by stratifying the feature space through a process called *recursive binary splitting*. The goal is to minimize the loss function

$$\sum_{j=1}^{|T|} \sum_{i:x_i \in C_j} (y_i - \hat{y}_{C_j})^2 + \alpha |T| \tag{8}$$

where, $|T|$ is the number of terminal nodes of the tree, C_j is the region corresponding to j^{th} terminal node, and $\hat{y}_{C_j}$ the predicted value of the outcome variable in the region C_j, which the mean value of the observations in the training data in that region.
α, the hyper parameter controls a trade-off between the subtree's complexity and its fit to the training data.

Random Forest

The process of tree construction is similar to a single decision tree, with some modifications. In each iteration, a tree is constructed using a random subsample. The number of features in these subsamples is determined through hyperparameter tuning. Random sampling in each iteration ensures less correlation among the regression trees constructed. The prediction function in my case becomes

$$\hat{y} = F(C) = \frac{1}{K} \sum_{k=1}^{K} h_k(C) \tag{9}$$

C stands for circumstances, which are a subset of the full set of circumstances in consideration. C is chosen randomly before constructing each tree. K is the total number of trees.

Tuned Hyperparmeters

Responsibility Cutoffs	n_trees	mtry	min_n
2	500	30	13
5	500	25	46
10	500	25	95
12	500	25	111
14	500	30	255
16	500	25	182
18	500	20	84

Figure 12: Tuned Hyperparamters

- *mtry*: An integer representing the number of predictors that will be randomly selected at each split during the tree model creation.
- *n_trees*: An integer representing the number of trees in the ensemble.
- *min_n*: An integer representing the minimum number of data points a node must contain before it can be split further.