Measuring Inequality of Opportunity - A Machine Learning Approach Using the PSID

Aman Desai

The Graduate Center, CUNY

03 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.
- Accounting for the role of dynamic complementarity in measuring the inequality of opportunity.
- Using supervised machine learning to construct counterfactual distribution of adult incomes based on circumstances.

Main Results

- Share of inequality due to circumstances beyond a child's control account for around **25 percent** of income inequality before the child turns 3.
- This share increases as high as **33.5** percent before the child becomes an adult.
- Since childhood skill gaps due to differing circumstances persist into adulthood, investing in human capital early in life is more efficient than compensating for a lack of opportunities in adulthood.

Related Literature

Inequality of Opportunity

- Seminal work by @roemer_pragmatic_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.
  - **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_technology_2007; @cunha_economics_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

Idea

Let each and every individual be fully characterized by the triple (y, C, e), and

$$C \in \Omega$$

$$e \in \Theta$$

$$y = g(C, e)$$

$$g : \Omega \times \Theta \Longrightarrow R$$

- Let all elements of the vector C, as well as e, be discrete.
- Let $y_{ij} = g(C_i, e_j)$.
- Let a *type* consists of all individuals with identical circumstances.
- Let a *tranch* consists of all individuals with identical effort levels.
- Let there be n types and m tranches.
- The population then can be represented by the $n \times m$ matrix $[Y_{ij}]$.

Inequality of Opportunity

Roemerian Algorithm

- Each column represents a tranch.
- Each row represents a type.
- Outcome differences due to factors beyond and individual's responsibility (circumstances) are unfair and should be compensated.
- Eliminate inequality across types **after** effort is realized, by eliminating inequality among people exerting the same degree of effort. (i.e. eliminate inequality within tranches)

Existing empirical work

- Several empirical approaches in last twenty years. [@bourguignon_inequality_2007; @pistolesi_inequality_2009; @ferreira_measurement_2011; @niehues_upper_2014; @hufe_inequality_2017]. The estimated shares of IOp in outcome inequality varies largely from 10 percent to as high as 60 percent.
- Usage of machine learning algorithms to model IOp. @brunori_roots_2023
- Contingency on normative judgments is greatly amplified: Not only can different indices be used to measure inequality on $[Y_{ij}]$, but the matrix itself can be constructed in different ways.
- Lower bound measures of IOp.

Technology of Skill Formation

@cunha_technology_2007 model technology for skill formation, where the vector θ_t evolves according to a law of motion affected by investments broadly defined as actions taken to promote learning, and parental skills (environmental variables).

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

- $f^{(t)}$ is assumed to be twice continuously differentiable, increasing in all arguments, and concave in I_t.
- I_t is human capital investment at time t.
- $\theta_{P,t}$ is parental skills at time t.
- The dimension of θ_t and $f^{(t)}$ is likely to increase with the stage of the life cycle.

Technology of Skill Formation

$$\omega_{i,t+1} = f(\omega_{i,t}, x_{i,t}, \omega_i^p, \epsilon_{i,t}) \tag{2}$$

The equation captures two ideas:

- investments in skills do not fully depreciate within a given period.
- stocks of skills can act in unison. For instance, as @cunha_formulating_2008 and @cunha_estimating_2010 point out, a higher level of non-cognitive skills promotes a higher level of cognitive skills.

Technology of Skill Formation

Complementarity

$\frac{\partial^2 \theta_{t+1}}{\partial \theta_t \partial I_t} > 0$ i.e. when stocks of skills acquired by period $t-1$, θ_t, make investment in period t , I_t more productive. Such complementarity explains why returns to educational investments are higher at later stages of the child's life cycle for more able children (those with higher θ_t).

Self-productivity

$\frac{\partial \theta_{t+1}}{\partial \theta_t} > 0$. i.e. when higher stocks of skills in one period create higher stocks of skills in the next period. For the case of skill vectors, this includes own and cross effects.

The concept of dynamic complementarity suggests that early gaps in circumstances beyond a child's control during critical stages of childhood can lead to persistent disparities in adult outcomes.

Ex-post Compensation

vs

Ex-ante Investment

Data

Proposal: Use age of consent as a boundary to separate circumstances and effort and look at circumstances at different stages in childhood to construct counterfactual distributions.

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.

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 of the individual when the individual is 35 years old.

The data in consideration are 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

Methodology

I follow [@ferreira_measurement_2011; @niehues_upper_2014] who use a parametric specification to estimate IOp.

$$ln(y_i) = \alpha_0 + \sum_{p=1}^{\tilde{P}} (\alpha_p C_i^p) + u_i$$

The right hand side of the specification hence involves only circumstance variables and can be used to predict the adult income. The estimate of IOp in this reduced form measures overall effect of circumstances on adult income.

Methodology

I construct a counterfactual distribution Φ by replacing adult incomes by their predictions. In this counterfactual distribution, all the individuals with the same circumstances have the same income. Although, I only observe a subset of circumstances $\tilde{\Omega} \subseteq \Omega$ of size $\tilde{P}$

$$\hat{y}_i = exp\left[\sum_{p=1}^{\tilde{P}} (\hat{\alpha}_p C_i^p)\right]$$

In this counterfactual distribution, all the individuals with the same circumstances have the same income.

Methodology

The measurement of inequality of opportunity can be thought of as a two-step procedure: first, the actual distribution y_i is transformed into a counterfactual distribution $\hat{y}_i$ that reflects only and fully the unfair inequality in y_{ij}, while all the fair inequality is removed. In the second step, a measure of inequality is applied to $\hat{y}_{ij}$.

$$IOp = \frac{MLD[\hat{y}_i]}{MLD[y_i]}$$

where MLD is mean logarithmic deviation

If all the income differences are due to circumstances, the IOp measure would be a unity.

Methodology

- Limitations of the OLS
  - Focus on obtaining unbiased estimates precisely by minimizing the in sample error. That leads to higher out of sample errors (over-fitting).
  - Not suitable for high dimensional data due to loss of many degrees of freedom.
- Why Machine Learning?
  - Machine learning techniques were developed specifically to maximize prediction performance by providing an empirical way to make the bias-variance trade-off. @hastie_elements_2009.
  - Supervised machine learning algorithms handle high dimensional data pretty well.

Tree Based Algorithm

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

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

where, $|T|$ is the number of terminal nodes of the tree, R_j is the region corresponding to j^{th} terminal node, α is complexity parameter that defines the cost-complexity measure of a given tree,and $\hat{y}_{Rj}$ the predicted value of the outcome variable in the region R_j, which is the mean value of the observations in the training data in that region.

Methodology

Random Forest

Although regression trees have simple interpretations and are easy to understand, they have low bias but high variance. The goal here is to reduce the variance and one solution would be to use a method called bagging (bootstrap aggregating). The idea here is to create B bootstrap samples of training data and fit a regression tree for each dataset so that we have B regression tree predictions. Finally all these B set of predictions are averaged to reduce the variance.

Random forest algorithm uses this technique with a small tweak that sees a random subset of predictors is used for each bootstrap sample and hence reduces the correlation among the regression trees.

Cross Validation

- Divide sample in K folds
- Choose some value of the tuning parameter, λ or α
- For each fold k = 1,..,K
  - Train model leaving out fold k
  - Generate predictions in fold k
  - Compute MSE for fold k : $MSE_k = \frac{1}{n_k} \sum_{i \in k} (y_i - \hat{y}_i)^2$
- Compute overall MSE corresponding to the current choice of λ : $MSE(\lambda) = \frac{1}{K} \sum_{k=1}^{K} MSE_k$

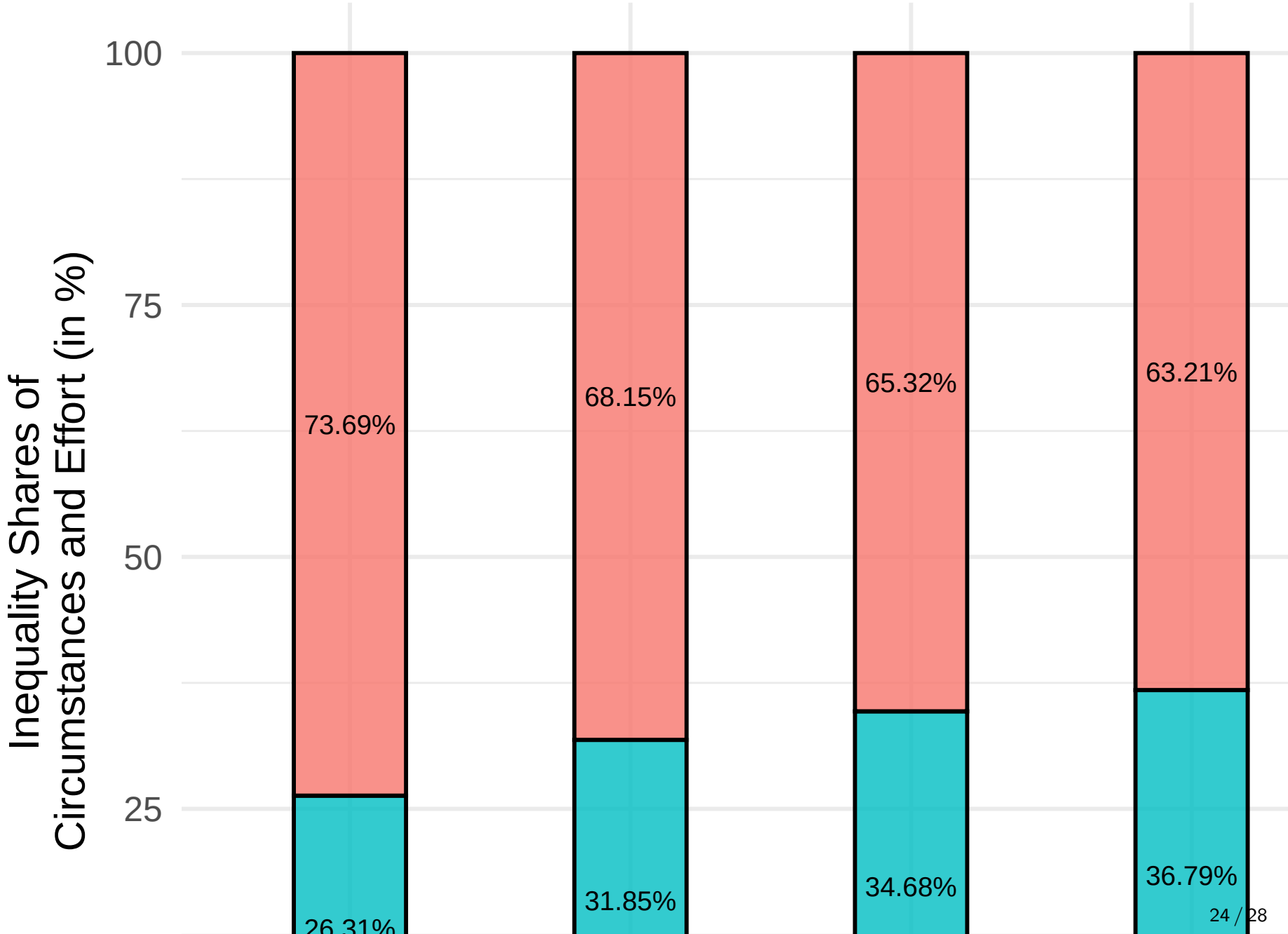
Methodology

I start by splitting the sample into a training set with $i_{train} \in \{1, ..., N_{train}\}$ and a test set with $i_{test} \in \{1, ..., N_{test}\}$. Stratified on adult income, $N_{train} = \frac{4}{5}N$,$N_{validation} = \frac{1}{4}N_{train}$ $N_{test} = \frac{1}{5}N$. I fit the models on training data, tune the hyper parameters on validation data, and evaluate their performance using test data.

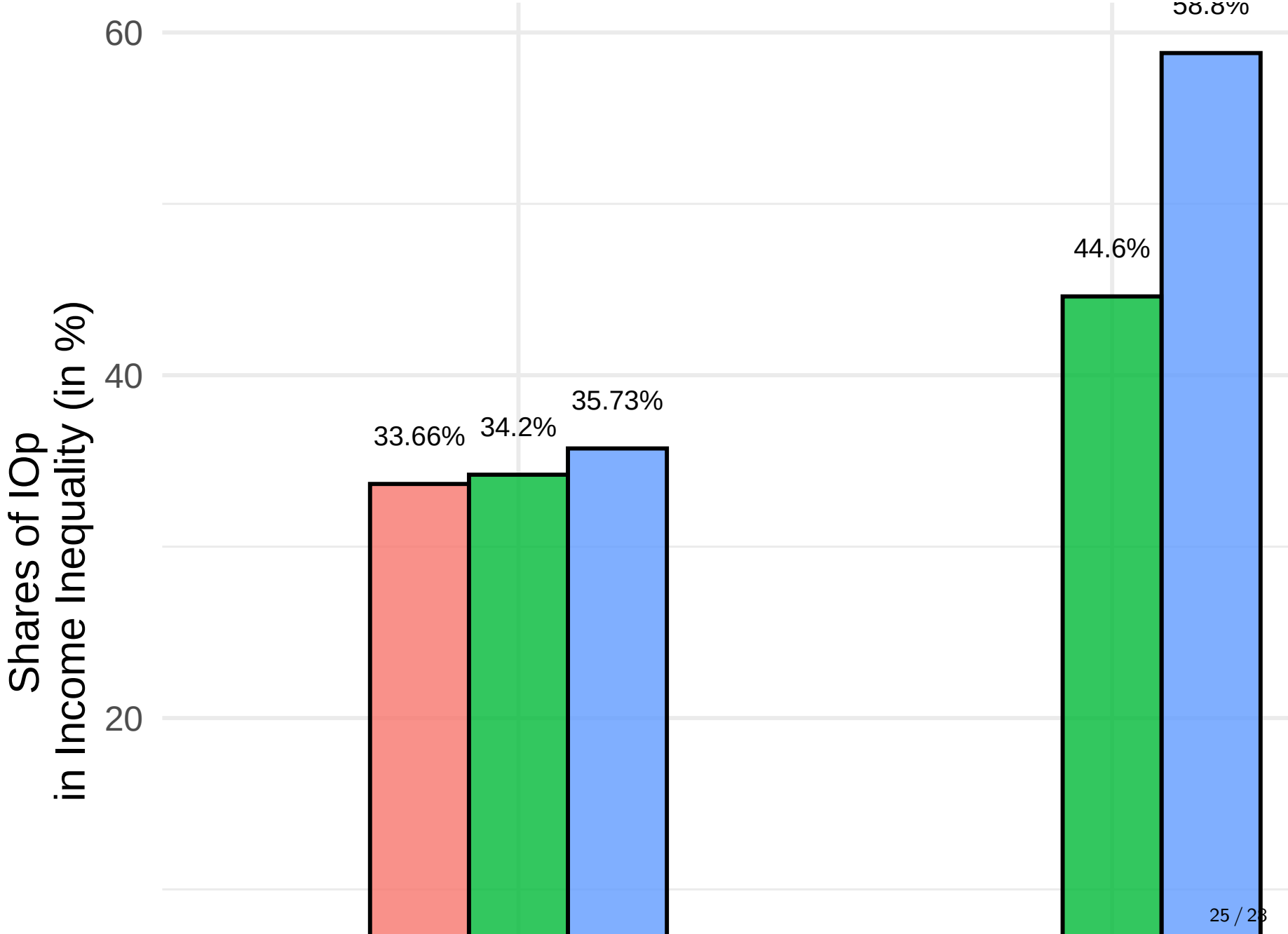
Procedure

- Run the random forest algorithm on the training data.
- Perform 10-fold cross validation (repeated twice) for hyperparameter tuning and chose the models with the hyperparameters that ensure the lowest MSE.
- Store the prediction functions $\hat{f}_{train}(\hat{\Omega})$.
- Obtain final predictions using the test data $\hat{y} = \hat{f}_{train}(\hat{\Omega}_{test})$.
- Calculate the out-of-sample error $MSE^{test} = \frac{1}{N_{test}} \sum_{i_{test}} (y_{i_{test}} - \hat{y}_{i_{test}})^2$.

Results


Inequality Shares of
Circumstances and Effort (in %)
100
75
50
25
73.69%
26.31%
68.15%
31.85%
65.32%
34.68%
63.21%
36.79%

Results


Shares of IOp
in Income Inequality (in %)
60
40
20
33.66%
34.2%
35.73%
44.6%
58.8%

Results


Sex of the Individual
Whether the Head Receives Food Stamps at Age 1
Occupation of the Head at Age 5
Value of the Family House at Age 1
Value of the Family House at Age 4
Whether the Head Receives Food Stamps at Age 2
Age ref
age of the indiv

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 the achieving success in the adulthood.
- By the time the child turns 3, the circumstances account for majority of inequality of opportunity, with IOp measure being 25.4 percent.
- Early intervention to equalize IOp could be more efficient than ex-post compensation advocated by Roemer (1993).

Next Steps: Exploration of differences in IOp across gender, geography

References I