Measuring Income Inequality of Opportunity
Focusing on Early Childhood Circumstances

Aman Desai
The CUNY Graduate Center

08 October, 2024

Abstract

According to the egalitarian principle of justice, life success can be attributed to two main factors: circumstances beyond an individual's control and personal effort within an individual's control. Roemer's concept of equality of opportunity proposes that individuals should be compensated for inequalities resulting from unequal circumstances. Dynamic complementarity in skill formation suggests that early childhood skill gaps often persist into adulthood, leading to unequal outcomes. Using PSID data, I classify all measurable factors before the age of consent at 18 as circumstances. My findings show that 42.57% of inequality can be attributed to inequality of opportunity before an individual becomes an adult. Moreover, nearly 34% of total outcome inequality stems from circumstances faced by individual at or before age 5. These results highlight the importance of considering childhood circumstances—factors beyond individual control—when measuring inequality of opportunity rigorously.

1 Introduction

Some forms of inequality in society are unjust. However, deciding which types of inequality are fair can present an ethical dilemma. It is crucial to consider the mechanisms that enable individuals to succeed in life when addressing the issue of inequality. Starting with Rawls (1971), the concept of egalitarianism has shifted from focusing on welfare derived from the final outcome to paying attention to the processes that lead to the outcome. Economists have begun to incorporate the idea of fairness in rewarding individual responsibility while acknowledging the existence of unfair inequalities in their analyses.

One vital contribution by Roemer (1993) is the notion that success in life can be broadly determined by two elements: "circumstances," over which individuals have no control and for which they should not be held responsible, and "effort," which represents the factors within an individual's control. Equality of opportunity is achieved when the distribution of outcomes depends only on effort, not on circumstances. This formulation of equality of opportunity aligns with the concept of a "level playing field". The literature on redistributive preferences explains how individual views on redistributive policies correlate with beliefs about the impact of effort versus circumstances on outcomes (Alesina and Giuliano (2011)). Fong (2001) demonstrates that people are more accepting of inequality resulting from differential effort rather than unequal circumstances. From a behavioral perspective, Starmans, Sheskin, and Bloom (2017) uses laboratory studies, cross-cultural research, and experiments with infants and young children to show that humans naturally favor fair distribution over unequal distribution. When equality and fairness conflict, people prefer fair inequality to unfair equality.

The empirical literature measures the extent of inequality of opportunity (IOp here after) for various outcomes, including income, wages, and health, in many countries[1] . I add to the empirical literature on inequality of opportunity by measuring the income inequality due to unequal opportunities by creating age based circumstance sets using age of consent as a responsibility cutoff. Numerous empirical studies have estimated the extent of income inequality due to circumstances. In the case of the US, Pistolesi (2009) estimates the IOp between 20 and 43% of earnings inequality. In a study by Hufe et al. (2017), using NLSY79 data, the authors estimate IOp shares in income inequality from 27.1% to as high as 43.5% for the United States. The recent launch of the Global Estimates of Opportunity and Mobility (GEOM) database marks a

[1]For recent survey articles on both the theoretical and empirical literature, see Fleurbaey and Peragine (2013), Roemer and Trannoy (2016), Ferreira and Peragine (2015), or Ramos and Van de gaer (2016).

significant step toward understanding global inequality of opportunity. This public data repository currently includes estimates from 72 countries representing 67% of the world's population and aims to highlight how income inequality is influenced by circumstances beyond individual control, such as parental background and geographic location.

Lack of high-quality datasets that reflect all the circumstances faced by an individual leads to partial observability of circumstances which results in a downward bias in the estimates of the IOp. (Bourguignon, Ferreira, and Menéndez 2007; Ferreira and Gignoux 2011; Niehues and Peichl 2014). There is also an issue of arbitrary categorization of circumstances and effort variables. A factor considered by a researcher as a circumstance may not be categorized as such by others. Since the distinction between "effort" and "circumstances" is a value judgment, measuring the role of circumstances accurately in predicting adult outcomes is challenging. I take a rather radical but not unprecedented position (Hufe et al. (2017)). All measurable factors, behavioral or otherwise, before the age of consent are considered circumstances. The law determines when a child becomes an adult and is ready to stand on their own (for example, voting laws, drinking age, and other measures). Therefore, I propose using this societal value judgment to categorize variables as either "effort" or "circumstances." Theoretically, if I had all information about a child before the age of consent, I would categorize that information as circumstances. A child is unable to give consent before reaching the age of consent. Hence, the idea of equality of opportunity as explained by Roemer requires that the child should not be held responsible for factors affecting them before the age of consent, including their achievements.

Following this view, all the measurable factors before the age of consent, say before 18 years, could be categorized as circumstances. The inequality in outcomes generated via these circumstances could be considered "unfair" and hence should be addressed. Roemer (1993) proposes the individuals affected by adverse circumstances should warrant compensation. In this paper, I attempt to bring an insight, dynamic complementarity, from the literature on human capital development in childhood to contribute to the literature on inequality of opportunity[2]. While measuring IOp, we account for the idea of dynamic complementarity in skill formation. Skills gaps opened up early in childhood due to unequal circumstances tend to persist

[2]For survey, see Heckman and Mosso (2014).

in adult hood. Any policy to address these inequality using compensation later in life may prove inefficient if the skills gaps early in the childhood have not been addressed. It is important to measure inequality of opportunity rigorously using these early childhood circumstances to better inform policy decisions. We measure inequality of opportunity using circumstances a child faces at critical stages in her development before the age of consent.

Recent empirical studies in the literature have used machine learning algorithms to create counterfactual distributions of outcomes using circumstances as well as identification of circumstances. Using representative survey data from 31 European countries Brunori, Hufe, and Mahler (2023) show superiority of tree based models in creation of counterfactual distributions using data on circumstances. Machine learning algorithms such as decision trees as well their ensemble random forest also allow interaction among the circumstance factors. Use of machine learning algorithms allows to relax the assumption of fixed circumstance with linear specification and offer non linearity in relationship between circumstances and the outcome variable.

I follow this practice and utilize the random forest algorithm to calculate estimates of inequality of opportunity. To account for dynamic complementarity in skill formation, I create age based opportunity sets using circumstances before or at critical stages in childhood. Then I use random forest algorithm to create counterfactual distributions of adult income for these different circumstance sets and apply a scale invariant measure of inequality to obtain the estimates of IOp and subsequently its share in total income inequality. It is important to clarify that the nature of the problem being examined is not causal. The objective here is to examine the extent to which variations in adult incomes can be attributed to circumstances that are perceived as "unfair." This aspect classifies it as a prediction problem, which can be addressed using supervised machine learning.

Using Panel Study of Income Dynamics (PSID) data on both the Survey Research Center (SRC) sample and the full sample—which also includes the Survey of Economic Opportunity sample—I found that relative inequality of opportunity (IOp), that is the shares of income inequality in adult income (proxied as individual labor income at age 35) attributable to unequal circumstances faced by an individual before or at age 5, are about 34% in the full sample and around 29% in the SRC sample. When individual adult incomes are proxied by average age-adjusted incomes across four waves from 2012–2018, these estimates increase slightly

to 34.34% for the full sample and 30.64% for the SRC sample. These figures exceed those obtained using widely used fixed circumstances in the literature in OLS regression.

The paper is structured as follows: in the next section briefly covers theoretical framework explaining the concepts of inequality of opportunity and dynamic complementarity, section 3 covers data description, section 4 covers the measurement of IOp followed by results in section 4, and section 5 concludes.

2 Theoretical Framework

2.1 Inequality of Opportunity

Consider a population $\mathcal{N} = \{1, 2, ..., N\}$. 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 $g : \Omega^c \times \Omega^e \Longrightarrow R$. The outcome vector $y = (y_1, ..., y_N)$ represents the incomes of individuals, which depend on circumstances C and effort e. 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. If the population is divided into M mutually exclusive and exhaustive groups, called *types*, such that $\prod = \{\tau_1, \tau_2, ..., \tau_M\}$, then all individuals belonging to the same *type* τ_m share the same circumstances: $\tau_1 \cup \tau_2 \cup ... \cup \tau_M = \{1, ..., N\}$, $\tau_m \cap \tau_k = \emptyset, \forall m, k$ and $C_i = C_j, \forall i, j \mid i, j \in \tau_m, \forall m$..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 $F(y|C) \neq F(y)$.

Following the egalitarian project, Roemer (1993) argues for the *ex-post compensation* principle, which calls for compensation after the effort is realized. The *ex-post compensation* principle requires that individuals exerting the same degree of effort receive the same outcomes, regardless of their circumstances. Roemer proposes a model of optimal taxation where the social planner's objective function incorporates an aversion to inequality caused by circumstances beyond an individual's control. Effort is typically unobservable. Roemer offers a solution to identify the effort predicated on some assumptions.

1. The circumstances faced by the individuals are fully observed.
2. Outcome is monotonically increasing in effort. Higher effort implies higher outcome.

$$y^m(e_i) \geq y^m(e_j) \Leftrightarrow e_i^m \geq e_j^m, \forall m = 1, ...M, \forall\ e_i, e_j \in R$$

3. Effort, by definition, is orthogonal to circumstances.

Roemer (2002) argues that when comparing the efforts of different individuals, we should take into account their specific effort distributions based on their *types*, and individuals should not be held solely responsible for these differences. Indeed, Roemer distinguishes between "level of effort" and "degree of effort", with latter being ethically relevant which can be identified with the quantile of the *type*-specific effort distribution of the individual. We denote distribution of effort in *type* m with $G^m(e)$ and its quantiles with $\pi \in [0, 1]$. For example, consider two individuals, A and B, born into a wealthy family and a poor family respectively. If they exert the same level of effort, the degree of effort is higher for child B due to her less advantaged circumstances. Instead of directly comparing their effort levels, Roemer suggests comparing their ranks (quantiles) on individual *type*-specific effort distributions. Since, effort distributions are mostly unobservable, Roemer suggests to identify the degree of effort exerted by the individual with the quantile of their *type*-specific outcome distribution. i.e. $y^m(G^m(e)) = y^m(\pi)$.

Then the requirement for the same outcome due to same degree of effort exerted by the individuals is, $y^m(\pi) = y^k(\pi) \Leftrightarrow F^m(y) = F^k(y); \forall \pi \in [0, 1], \forall m; k = 1, ..., M.$

As explained, the implication of Roemer's adherence to the ex-post compensation principle is that society should compensate individuals for their unequal circumstances after individual effort is realized. This contrasts with the *ex-ante compensation* principle, where compensation is due before effort is realized by equalizing the opportunity sets available to everyone, regardless of their circumstances.

2.2 Importance of Skills in Early Childhood

Skills are multidimensional, covering cognition, personality, as well as mental and physical health. They reflect an individual's capacity to act. Recent studies show that both cognitive and non-cognitive skills have an impact on labor market outcomes[3].

[3]See Borghans et al. (2008) and Almlund et al. (2011) for comprehensive surveys.

Gaps in both cognitive and non-cognitive skills emerge early in childhood, across individuals and socio-economic groups. There is substantial evidence of early divergence in these skills even before schooling begins[4].These skill gaps correspond to gaps in family investment and the environment in which individuals are brought up. Hart and Risley (1995) showed that children from professional families speak 50% more words than children from working-class families, and twice as many as children from welfare families. There is a substantial literature, summarized in Cunha et al. (2006), Lareau (2011), Kalil (2015) showing that disadvantaged children have compromised early environments as measured on a variety of dimensions. Moreover, various skills and abilities are critical at different stages of the life cycle. Early life disadvantages have a lasting impact on a range of outcomes in adulthood. Cunha et al. (2006) provide a review of studies that examine the significance of early childhood environments on socioeconomic outcomes in adolescence and adulthood.

The empirical literature shows that investing in disadvantaged young children yields higher economic returns[5]. Early interventions have been shown to be more effective than targeted interventions later in life, as high-quality interventions during the early years promote the development of skills in disadvantaged young children that lead to greater economic returns in the future. Non-cognitive skills foster cognitive skills and are an important product of successful families and successful interventions in disadvantaged families.

2.3 Technology of Skill Formation

Both cognitive and non-cognitive skills, the technology used for their development, and parental investment, which includes their own skills, are crucial in determining the dynamics of family influence. Cunha and Heckman (2007) model technology for skill formation, where formulation of skills is conceptualized as a law of motion. Let $\omega_{i,t}$ denote the human capital of child i at age t, $\omega_{i,t+1}$ the human capital at age $t+1$. Let parental investment for the child i be $x_{i,t}$ at age t and parental human capital be ω_i^p. $\epsilon_{i,t}$ is an independently and identically distributed unobserved individual component.

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

[4]Cunha et al. (2006), and Cunha and Heckman (2007).

[5]for comprehensive survey of empirical literature on human development and social mobility see Heckman and Mosso (2014).

The equation 3 captures the idea of static complementarity between investment in human capital in period t and skills in t Becker and Tomes (1986). Children who are more intelligent, healthier, have better non cognitive skills acquire more capability from the same level of investment.

$f(.)$The function is assumed to be twice continuously differentiable, increasing in all arguments, and concave in $x_{i,t}$. Stock of skills $\omega_{i,t}$ and $\omega_{i,t+1}$ include both cognitive and non cognitive skills. The dimensions of $\omega_{i,t}$ and $f(.)$ are allowed to increase with the stage of the life cycle. The technology in the model is stage-specific and allows for critical and sensitive periods in the formation of capabilities and effectiveness of the investment. This formulation of technology has two implications.

$\frac{\partial \omega_{i,t+1}}{\partial \omega_{i,t}} > 0$, that is, when higher stocks of skills in one period create higher stocks of skills in the next period. $\frac{\partial^2 \omega_{i,t+1}}{\partial \omega_{i,t} \partial x_{i,t}} > 0$, that is when stocks of skills acquired by period t, $\omega_{i,t}$, make investment in period $t+1$, $x_{i,t}$ more productive. For the case of skill vectors, this includes own and cross effects. These generate dynamic complementarity between investment in period t and in period k where $k > t$. Higher investment in period t increases $\omega_{i,t+1}$ as $f(.)$ is increasing in $x_{i,t}$. This in turn raises $\omega_{i,k}$ because technology is increasing in $\omega_{i,m}$ for any m between t and k. This in turn leads to $\frac{\partial f}{\partial x_{i,k}} > 0$, since $\omega_{i,k}$ and $x_{i,k}$ are complements. It follows that

$$\frac{\partial^2 \omega_{i,t+k+1}}{\partial x_{i,t} \partial x_{i,t+k}} > 0, \quad \forall k \geq 1. \tag{2}$$

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. These properties help explain why early investment in disadvantaged children can yield high productivity, which is both fair and economically efficient. Conversely, the return on investment tends to be lower at later stages for disadvantaged children, due to their lower skill base and hence reduced complementarity effect. While this may seem fair, it's less economically efficient.

The concept of dynamic complementarity implies that early differences in skill investments can lead to enduring disparities in adult outcomes. I argue that a child encounters situations beyond their control before reaching the age of consent. By applying the principle of dynamic complementarity, I highlight the unequal opportunities arising from unequal circumstances in early life stages when measuring inequality of opportunity. By identifying specific age milestones in childhood, I can analyze the extent to which inequality

in adult incomes can be attributed to circumstances before or during these stages. In the United States, for instance, children typically start speaking at age 2, attend kindergarten at age 5, begin high school at age 14, and transition into adulthood at age 18. By focusing on these significant stages of development, I can more accurately measure the opportunity gaps resulting from circumstances preceding these critical childhood stages.

3 Data Description

The data used in this study comes from the Michigan Panel Study of Income Dynamics (PSID). PSID is the longest-running longitudinal survey of the population in the United States, beginning in 1968 with a coverage of 4,800 households. It ran annually until 1997, and has since been conducted biennially. The genealogical design of PSID data allows to link individuals of interest to their parents and grandparents.

3.1 Analytical Sample

PSID was originally created to study poverty. As a result, it disproportionately sampled individuals from poor households, which are included in the SEO (Survey of Economic Opportunity) sample. The analytical sample comprises two subsets. The first subset is restricted to individuals from the SRC (Survey Research Center) sample, ensuring a representative sample of the US population. The second subset, or full sample, includes individuals from both the SRC and SEO samples. To account for the inclusion of the SEO sample in the full sample, I apply appropriate weights provided by the PSID during the analyses. This approach yields a larger sample size compared to limiting the analysis solely to the SRC sample.

The individuals of interest are the heads of the family and their spouses/partners who were present during the interviews conducted in the 2013-2019 waves. Since any measurable data on a child before the age of consent is considered circumstances, the goal is to use the data before the age of consent, which in the American context for certain rights and privileges I take to be 18, for an individual to predict their outcome in the labor market. I look at the individuals (the heads of the family and their spouses/partners present during the interview in the 2013-2019 waves) who were born in 1978-1983 and follow them on various characteristics including of the families they grew up in during first 18 years of their lives. I use the Family Relationship

Matrix (FRM) to identify the family of the individual. I also use the Family Identification Mapping System (FIMS) provided by PSID to link these individuals to their parents in their first 18 years of life.

The sample consists of data on family heads, their spouses, and in some cases, other family members. The head of the family can be a father, a step-father, a grandfather, or in some cases, a single mother. Therefore, some children at some point in time may or may not have their parents as the heads of their family. If the family head is a parent, then using FIMS, we ensure those parents are the individual's biological parents.

Data on all six birth cohorts includes around 60 variables on 1022 individuals in the full sample and 639 individuals in the SRC sample across 18 years. Although I do have longitudinal data, I do not make use of the panel nature of it. Instead I create wide datasets according to pre-determined age cut-offs where each row reflects a biography of an individual across their first 18 years. Since the data only includes variables before a child's consent age, all these factors should be considered circumstance variables.

As mentioned earlier, various skills and abilities are critical at different stages of the life cycle. Dynamic complementarity suggests that gaps in skills attained at different critical and sensitive periods of childhood tend to persist in adulthood and lead to unequal outcomes. Moreover, dynamic complementarity and self productivity together suggest that lack of investments in skills at early stages lead to low returns to human capital investment in later stages of life. The Panel Study of Income Dynamics (PSID) includes data on measurable factors at various stages of childhood. This makes it suitable for accounting for critical stages before the age of consent, allowing for the creation of age-based circumstance sets.

To illustrate how factors are included in this study, imagine the circumstance sets as displayed in figure 2 that affect individuals during different stages of childhood based on their age. The age cutoffs (or responsibility cutoffs) for this study align with crucial stages of childhood development. Each circle represents a set of circumstances relevant to a specific age group. These sets coincide with significant milestones in a child's growth. For instance, in the United States, children typically start speaking at age 2, attend kindergarten at age 5, enter high school at age 14, and transition into adulthood at age 18. Hence, the smallest circle represents all the circumstances the child faces before or at age 2. The largest circle represents all the circumstances the child faces before transitioning into adulthood.

By focusing on these pivotal developmental stages, I can more accurately assess the opportunity gaps

resulting from conditions that occur before these critical points. Larger circles indicate a greater number of circumstances included in that set. As a child matures into an adult, the number of circumstances they encounter increases, which aligns with the evolving dimensions of factors in skill formation technology. Some factors may appear in all sets, such as the child's demographic characteristics like sex and race, family income and wealth, and specific characteristics of the family environment over the first 18 years of the child's life.[6]

The main outcome of interest in this study is individual's permanent income. To proxy for an individual's permanent income, I use two measures. First, I use their labor income at age 35. This means that for individuals born in 1978, their labor income is measured in 2013. Similarly, for individuals from other birth cohorts, such as those born in 1983, their incomes are measured in 2018. For simplicity, the labor income at age 35 excludes farm and unincorporated business income. Additionally, I omit individuals with income below $500 from the analyses. The decision to measure labor income at age 35 provides a large enough sample size while ensuring it serves as a reasonable proxy for an individual's permanent income.

Of course, labor income measured at a single point in time is susceptible to measurement error and can lead to attenuation bias (Solon 1992; Nybom and Stuhler 2017). To minimize this attenuation bias, I proxy the individual's permanent income using their labor income in adulthood, averaged over four years from 2013-2019. For individuals with missing income data in any wave, I calculate their average income using only the available years. For example, if an individual has income data for 2012, 2014, and 2018, but missing in 2016, I compute their average income using three years (2012, 2014, 2018). For the cohorts under consideration (born in 1978-1983), average incomes are measured at different ages. For example, in 2015, an individual born in 1978 is on average 37 years old, while someone born in 1982 is on average 33 years old.

The outcome variables are measured in logarithmic terms. Finally, I use urban CPI series (CUUR0000SA0) from Bureau of Labor Statistics for consumer price index to convert all monetary variables in 2018 dollars.As explained previously, even though I have panel data, I do not utilize its longitudinal nature. Hence, I use individual cross-sectional weights from the survey waves in which individuals were identified. For example, for individuals born in 1978, their cross-sectional weights were pulled from the 2013 survey wave. For those born in 1979 and 1980, their cross-sectional weights were taken from the 2015 survey wave. This same approach

[6]The complete list of variables is provided in the appendix.

applies to individuals in the remaining cohorts. I perform analyses using both outcome variables explained above for both the full sample and the SRC sample.

4 Measuring Inequality of Opportunity

Various measures have been proposed based on differing normative views[7]. I use a widely adopted ex-ante utilitarian measure of inequality of opportunity (Van de Gaer 1993; Checchi and Peragine 2010.). The idea is to construct a counterfactual distribution of outcomes obtained by removing inequality within types(circumstances) from the original outcome distribution. Measuring the inequality of opportunity involves two steps. First, I form a counterfactual distribution of outcomes based on individual types, or circumstances. Then, I apply a standard measure of inequality that satisfies anonymity, the principle of transfers, population replication, and scale invariance[8], to the counterfactual distribution conditional solely on circumstances. The data generating process can be written as follows:

$$y = h(C, e) = f(C) + u = E(y|C) + u \tag{3}$$

Here, $E(y|C)$ represents the variation in outcome due to observed circumstances. The residual term u, which is independently and identically distributed, captures the variation due to both unobserved circumstances and individual effort. Therefore, with this specification, one can only interpret inequality of opportunity as a lower bound. I use what is referred to as parametric specification[9] in the literature for estimation of lower bounds of IOp (Bourguignon, Ferreira, and Menéndez 2007; Ferreira and Gignoux 2011; Niehues and Peichl 2014).

$$ln(y_i) = \alpha_0 + \sum_{l=1}^{L} (\alpha_l C_{i,l}^s) + u_i \tag{4}$$

where y is the adult income, C is the collection of factors that are categorized as circumstance belonging

[7] see Ramos and Van de gaer (2016) for extensive survey

[8] See Cowell (2016) for more information

[9] There is also a non-parametric approach used by Checchi and Peragine (2010). Briefly, partitions in the sample are created based on mutually exclusive *types* based on realization of observed circumstances. Then $E(y|C)$ is obtained from average incomes within *types*.

to a finite set Ω^c. $s \in \{2, 5, 14, 18\}$ reflecting four different sets of circumstances based on chosen age cutoffs.

$$\hat{y}_i = exp\left[\alpha_0 + \sum_{l=1}^{L}(\hat{\alpha}_l C_{i,l}^s)\right] \tag{5}$$

If any measurable data in a child's life before the age of consent is considered part of the circumstance set Ω^c, then the data that ideally includes a biography of a child across the first 18 years will form a circumstance set for us. However, in reality, I do not observe a full set of circumstances, that is $C^s \subseteq \Omega^c$, and hence as I stated earlier, I obtain lower bound estimates of inequality of opportunity. The PSID offers an extensive list of factors across the first 18 years that make up the circumstance set. To account for dynamic complementarity, I construct four circumstance sets. I use age based circumstances to create opportunity structures.

These circumstance sets are created using the age cut offs based on critical stages of development in childhood.

$$C^2 \subseteq C^5 \subseteq C^{14} \subseteq C^{18} \subseteq \Omega^c \tag{6}$$

This formulation allows us to expand the circumstance set with age to account for additional circumstances a child faces at different stages of childhood before she becomes an adult. For instance $C^2 \subseteq \Omega^c$ includes the circumstances the child faces prior to or at age 2. $C^{18} \subseteq \Omega^c$ will make use of full set of circumstances in the data that include factors across first 18 years of the child's life. Similar interpretation holds for all other circumstance sets.

Obtaining $E(y|C)$ is a prediction problem, where the relationship between circumstances and outcome is unknown *a priori*. Various methods for estimating $E(y|C)$ have been proposed in existing literature. The lack of interaction terms and higher order polynomials in equation 5 can be addressed by directly integrating them into the equation.

Economists are increasingly using machine learning techniques to solve prediction problems (Kleinberg et al. 2015). However, addressing these "prediction policy problems" requires more than just simple regression techniques. These are designed to generate unbiased estimates of coefficients, not to minimize prediction error. Brunori, Hufe, and Mahler (2023) demonstrate the superiority of tree-based supervised machine learning

models over existing estimation methods. They use conditional inference trees and forests to generate $E(y|C)$ predictions. These models outperform the standard OLS and latent class models (Donni, Rodríguez, and Dias 2015), especially when the potential number of types exceeds the available degrees of freedom. I adopt their estimation strategy, but unlike Brunori, Hufe, and Mahler (2023), I use random forests for our predictions. Random forests are an interactive function class and hence allow for non-linearity among the "types", that is circumstances. After obtaining the adult income predictions based on circumstances, I apply an inequality measure, mean logarithmic deviation to the predicted income distribution to calculate absolute inequality of opportunity.

$$Absolute\ IOp = I(\hat{y}_{EA}) \tag{7}$$

where $I(\hat{y}_{EA})$ is the ex-ante measure of inequality of opportunity. I also report relative inequality of opportunity as a ratio of inequality in predicted income distribution to inequality in adult income. It can be interpreted as the share of inequality in adult income that is attributed to inequality of opportunity.

$$Relative\ IOp = \frac{I(\hat{y}_{EA})}{I(y)} \tag{8}$$

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

4.1 Regression Trees

An algorithm used in the literature to identify *types* (circumstances) is a regression tree. Similar to a linear regression function, a regression tree also predicts an outcome value for each feature vector. The prediction function takes the form of a tree that splits the feature space into two at every node. At each node, a single variable (individual's biological sex) determines whether the left (if female) or right (if male) child node is considered next. When a terminal node, or "leaf," is reached, a prediction is returned. Trees are thus a highly interactive function class. and allow to create "*types*", that is, circumstances from the data.

A regression tree algorithm makes predictions by stratifying the feature space through a process called

recursive binary splitting. This top-down, greedy approach starts at the top of the tree and splits the predictor space into two new branches further down the tree at each split. During each step of the tree-building process, the best split is created at that step, rather than looking ahead and selecting a split that will lead to a better tree in a future step. The goal is to minimize the loss function

$$\sum_{j=1}^{|T|} \sum_{i:x_i \in C_j} (y_i - \hat{y}_{C_j})^2 + \alpha|T| \tag{9}$$

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.

The algorithms works as follows :

1. To grow a large tree on the training data using recursive binary splitting, continue splitting until each terminal node has fewer than a specified minimum number of observations.

2. To obtain a sequence of best subtrees as a function of α, apply cost complexity pruning[10] to the large tree.

3. To tune the cost complexity hyper parameter α, use the k-fold cross validation or bootstrap resampling to obtain validation set results as function of α. Then, pick the value of α that minimizes the root mean squared error (rmse).

4. For the chosen value of α obtain the subtree fitted in step 2.

4.2 Random Forest Construction

Although regression trees are easy to interpret and understand, they have low bias but high variance, making them prone to overfitting. To reduce overfitting, I use a tree ensemble algorithm called Random Forest. The

[10]A strategy here is to grow a very large tree and then prune it back to a smaller simple subtrees that can perform better on test data. In broad terms, without the cost complexity parameter the algorithm provides the biggest tree as it only reduces the first term of the loss function. As α increases, the price to be paid for a tree with many terminal nodes increases and hence the loss function above minimizes for a small enough sub tree.

idea is to create B bootstrap samples of training data and fit a regression tree for each dataset, resulting in B regression tree predictions. Finally, these B sets of predictions are averaged to reduce the variance.

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

$$\widehat{y} = F(C) = \frac{1}{K}\sum_{k=1}^{K} h_k(C) \tag{10}$$

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. $h_k(C)$ denotes predictions from each tree. Averaging predictions from K trees reduces the overall variance.

4.3 Procedure

The data is in wide format with 396 circumstance variables as features in the model. Number of individuals in the data is 639 for the SRC sample and 1022 for the full sample that also includes individuals from the SEO sample. Each row reflects an incomplete biography of an individual across the first 18 years of their life. To train the model effectively, I use as much data as possible. As such, I do not create a separate test data set. Instead, I employ the full dataset with a cross-validation process to minimize overfitting and tune the hyper parameters. The goal is to calculate the shares of inequality of income opportunity as shown in equation 8. To that extent, I obtain predictions using the final model on the whole data to obtain absolute and relative IOp estimates for all age cutoffs. 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 $N_{validation} = \frac{1}{5}N$. This 5-fold cross-validation process helps minimize overfitting. To increase efficiency, I repeat this cross-validation process twice[11].

[11] Cross validation process and results of hyperparameter tunning are provided in the appendix.

- Store the prediction functions $\hat{f}_{train}(\hat{\Omega}^c)$.

- Obtain final predictions using the full data $\hat{y}_{EA} = \hat{f}_{train}(\hat{\Omega}^c_{fulldata})$.

This procedure is repeated for all circumstance sets in consideration based on cut-offs at age 2, 5, 14, and 18 for both the full sample as well as the SRC sample using individual labor income at age 35 and average age adjusted income across 2013-2019 waves as proxy for permanent income.

5 Results

5.1 Descriptive Statistics

Table 1 displays unweighted summary statistics for the outcome variable and selected variables that compose the circumstance sets. These represent the fixed set of circumstances frequently used in the literature to estimate the IOp. I provide descriptive statistics for both the Survey Research Center (SRC) sample and the full sample, which also includes the Survey of Economic Opportunity (SEO) sample. The SRC sample is representative of the US population, while the SEO sample includes a disproportionate number of poor households. Throughout the analysis, I use appropriate PSID survey weights to ensure the samples are representative of the population. Table 1 displays unweighted summary statistics for the outcome variable and selected variables that compose the circumstance sets. These represent the fixed set of circumstances frequently used in the literature to estimate the IOp. I provide descriptive statistics for both the Survey Research Center (SRC) sample and the full sample, which also includes the Survey of Economic Opportunity (SEO) sample. The SRC sample is representative of the US population, while the SEO sample includes a disproportionate number of poor households. Throughout the analysis, I use appropriate PSID survey weights to ensure the samples are representative of the population. Note that in table 1 "age" refers to the individual's age when they were a child. "head" refers to the head of the family in which the child grew up during childhood. "spouse" refers to the spouse of the family head.

Table 1: Descriptive Statistics of Selected Variables

Characteristic	Full Sample N = 1,022	SRC Sample N = 639
Individual labor income at age 35 (in log)	10.65 (10.04, 11.08)	10.80 (10.32, 11.20)
Total family income at age 1 (in log)	10.89 (10.34, 11.29)	11.09 (10.58, 11.40)
Sex		
Male	474 (46%)	311 (49%)
Female	548 (54%)	328 (51%)
Race		
White	559 (55%)	554 (87%)
Black	446 (44%)	72 (11%)
AIAE	8 (0.8%)	6 (0.9%)
Other	9 (0.9%)	7 (1.1%)
Occupation of the head at age 1		
Inap	178 (17%)	60 (9.4%)
Professional, Technical, and Kindred Workers	168 (16%)	157 (25%)
Managers and Administrators, except Farm	72 (7.0%)	62 (9.7%)
Sales Workers	22 (2.2%)	20 (3.1%)
Clerical and Kindred Workers	51 (5.0%)	28 (4.4%)
Craftsman and Kindred Workers	219 (21%)	151 (24%)
Operatives, except Transport	128 (13%)	72 (11%)
Transport Equipment Operatives	45 (4.4%)	23 (3.6%)
Laborers, except Farm	41 (4.0%)	24 (3.8%)
Farmers and Farm Managers	13 (1.3%)	12 (1.9%)
Farm Laborers and Farm Foremen	5 (0.5%)	2 (0.3%)
Service Workers, except Private Household	79 (7.7%)	28 (4.4%)
Private Household Workers	1 (<0.1%)	
Years of education of the head at age 1	12.00 (11.00, 14.00)	12.06 (12.00, 15.00)
Years of education of the spouse at age 1	12.0 (9.9, 13.1)	12.0 (12.0, 14.0)

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

5.2 IOp Estimates Across Critical Stages in Childhood

Table 2 presents estimates of total income inequality in adult labor income and absolute inequality of opportunity (IOp) for both the full sample and the Survey Research Center (SRC) sample. As mentioned in the data section, adult labor income is represented by individual labor income at age 35, as well as age-adjusted incomes averaged over survey waves from 2012–2018. The table reports total income inequality measured by the mean logarithmic devaition (MLD) of the income distribution of labor income[12]. For labor income measured at age 35, total income inequality is 0.368 for the full sample and 0.337 for the SRC sample. For an age-adjusted distribution of average incomes across four years, total income inequality is 0.327 for the full sample and 0.308 for the SRC sample.

[12] $MLD(x) = ln(\bar{x}) - \overline{ln(x)}$. MLD of 0 reflects everyone has the same income,i.e. perfect equality.

Table 2: Absolute IOp Estimates for Different Circumstance Sets

	Full Sample (N = 1022)		SRC sample (N = 639)	
	Income Inequality	Absolute IOp	Income Inequality	Absolute IOp
Outcome : Labor Income at age 35				
Baseline	0.368	0.083	0.337	0.065
Age cutoff at 2 years	0.368	0.104	0.337	0.091
Age cutoff at 5 years	0.368	0.125	0.337	0.099
Age cutoff at 14 years	0.368	0.141	0.337	0.113
Age cutoff at 18 years	0.368	0.156	0.337	0.128
Outcome : Age-adjusted Labor Income				
Baseline	0.327	0.078	0.308	0.063
Age cutoff at 2 years	0.327	0.106	0.308	0.091
Age cutoff at 5 years	0.327	0.112	0.308	0.094
Age cutoff at 14 years	0.327	0.128	0.308	0.111
Age cutoff at 18 years	0.327	0.140	0.308	0.125

I report the absolute IOp estimates for a baseline model which uses an OLS specification using fixed set of circumstances. Fixed set of circumstances includes individual's race, gender, occupation of the head of the family in which the child grew up as well as completed years of education of the head and their spouse in the family when child was 1 year old[13]. For the full sample, absolute IOp is 0.083 when adult income is proxied by individual labor income at age 35 and 0.078 when the adult income is calculated as an average income of incomes across waves from 2013-2019. In case of SRC sample, the absolute IOp is 0.065 for the baseline specification when individual's labor income at age 35 is considered. Averaging incomes across aforementioned waves does not affect the absolute IOp measure with estimate being 0.062.

In addition to the absolute IOp estimates for the baseline circumstances, estimates for the age based circumstance sets are reported, too. It is easy to see these estimates are greater than those obtained using fixed circumstance set. These absolute estimates also increase with age as expected given the expansion of circumstance sets with age is allowed. In case of adult incomes proxied by labor income at age 35 the absolute IOp estimates range from 0.104 to 0.156 for age cut off for circumstances sets at age 2 and 18 respectively for the full sample. For the SRC sample the estimates are smaller. Considering average incomes across four waves as a measure of adult incomes, the absolute IOp estimates are lower compared to the case above, as they range from 0.078 to 0.140 for age cutoffs at 2 years and 18 years for for the full sample. For the SRC sample, these estimates are 0.063 to 0.125 for the same age cutoffs.

[13]This includes any child from newborn to someone who hasn't turned 2 by the time of the next PSID survey.

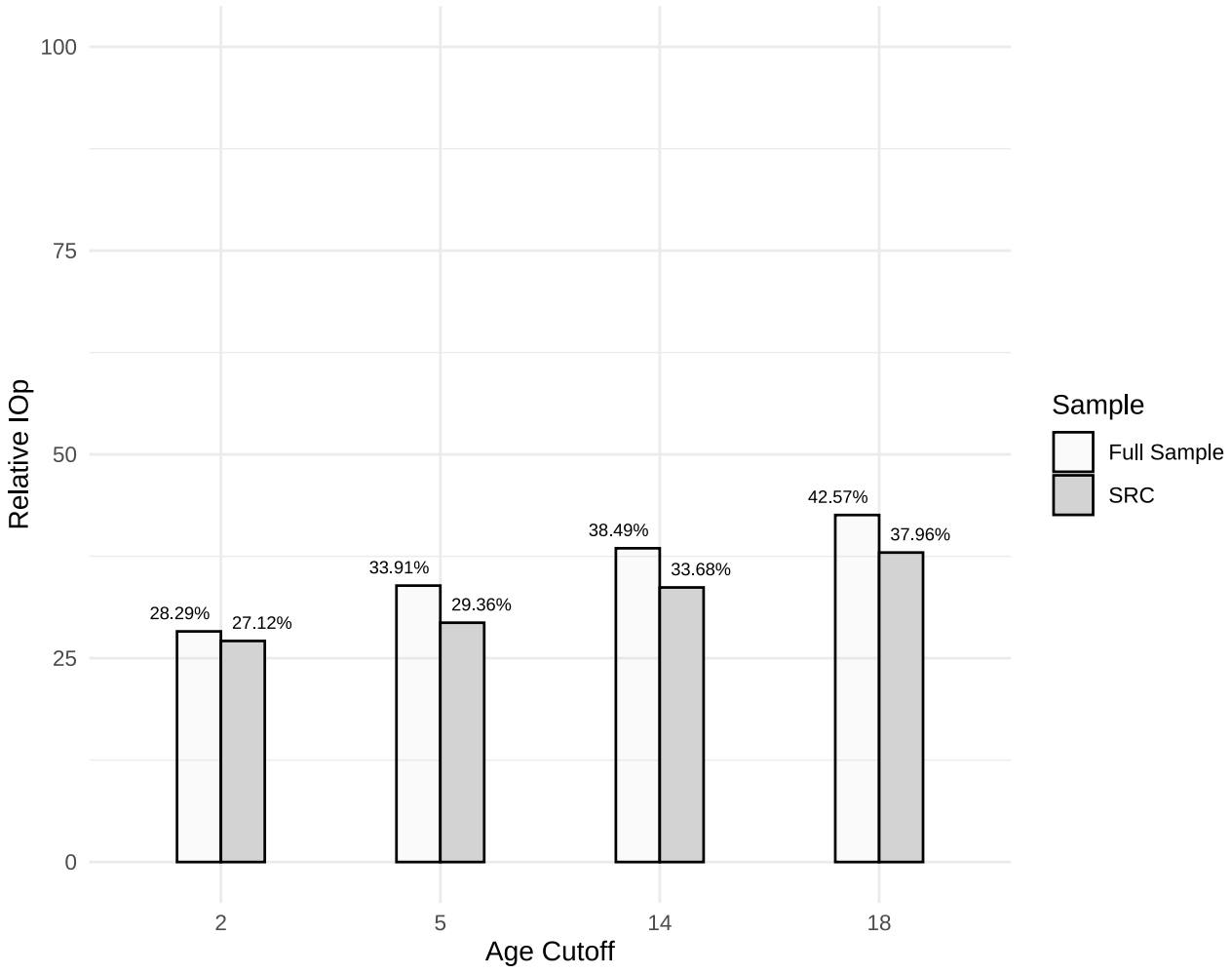

Figure 1: Relative IOp Estimates Across Age Cutoffs (Using Individual Labor Income at Age 35)

In Figure 1, a graphical representation of relative IOp, the share of inequality of opportunity in total income inequality for all age-based circumstance sets are reported. These shares are calculated using labor income at age 35 as an outcome variable to proxy individual income in adulthood. About 34% of income inequality in adult income at age 35 can be ascribed to unequal circumstances faced by an individual before or at age 5 while performing the analysis on the full sample. For the SRC sample, the relative IOp is 29.36%. As expected, the role of circumstances, and consequently the share of inequality of opportunity in income inequality, increases with age as the circumstance sets expand. This share of inequality increases up to 42.57% considering all circumstances encountered before or at the age of consent at 18 for the full sample and 37.96% for the SRC sample. These estimates of relative IOp are greater than those obtained from OLS regressions

using a fixed set of circumstances. For example, the relative IOp estimates are 22.5% for the full sample and 19.2% for the SRC sample when adult income is proxied by labor income at age 35.

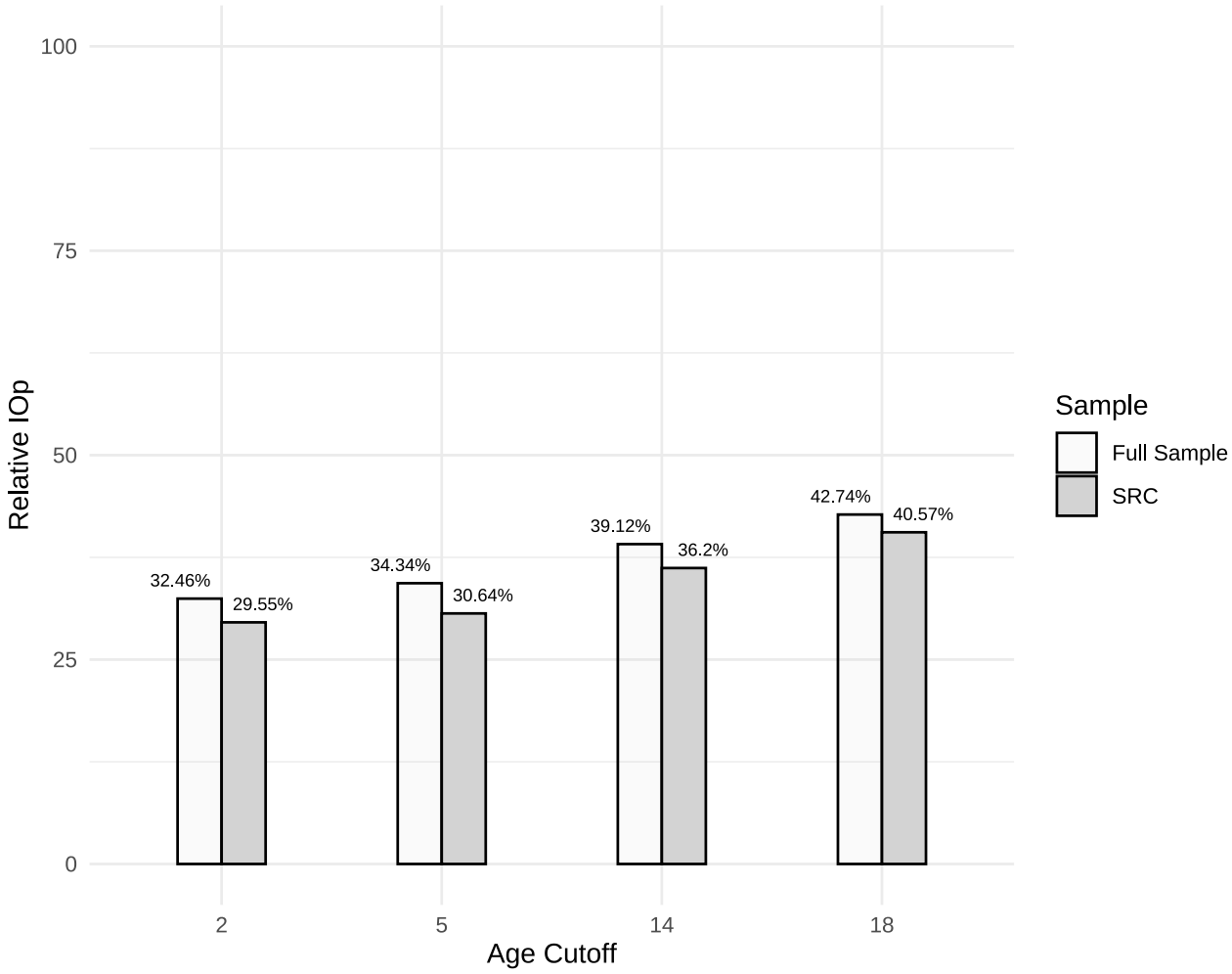

Figure 2: Relative IOp Estimates Across Age Cutoffs (Using Averaged Age-adjusted Incomes Across 2013-2019 Waves)

As outlined in the data section, I also conduct the analysis using an average of individual incomes from 2012–2018 waves to represent adult income as an outcome variable. Figure 2 shows 34.34% of income inequality in adult income can be attributed to unequal circumstances faced by individuals before or at age 5 when analyzing the full sample. For the SRC sample, the relative IOp is approximately 30.64%. In this analysis, since incomes are averaged at different ages for individuals, I also include their birth year in the equation 5. An individual cannot choose their birth year, so it enters the equation 5 as a circumstance and

helps account for average incomes measured at different ages in adulthood. The share of inequality increases up to 42.74% considering all circumstances encountered before or at the age of consent at 18 for the full sample and 40.57% for the SRC sample. Once again, these estimates of relative IOp are greater than those obtained from OLS regressions using a fixed set of circumstances. For example, the relative IOp estimates are 23.9% for the full sample and 20.6% for the SRC sample when adult income is proxied by labor incomes averaged over four survey waves from 2013-2019.

By definition, the share of IOp in income inequality that stems from unequal circumstances is deemed to be unfair inequality. The residual inequality could, therefore, be regarded as fair inequality. However, it's crucial to highlight that these IOp measures are in fact lower bound estimates. Despite the Panel Study of Income Dynamics (PSID) facilitating a more comprehensive account of circumstance factors than other datasets, we still lack a complete overview of an individual's life course across their first 18 years. Consequently, the unfair portion of inequality might be underestimated.

Figure 3 illustrates the profile of relative IOp—the proportion of inequality of opportunity in total income inequality—across all age cut-offs used in this study. These profiles are created for both the full and SRC samples, using two income measures as outcome variables to calculate relative IOp. The shares of IOp consistently increase with age across all cases. As expected, the highest IOp share stems from considering all measurable data on a child before the age of consent at 18.

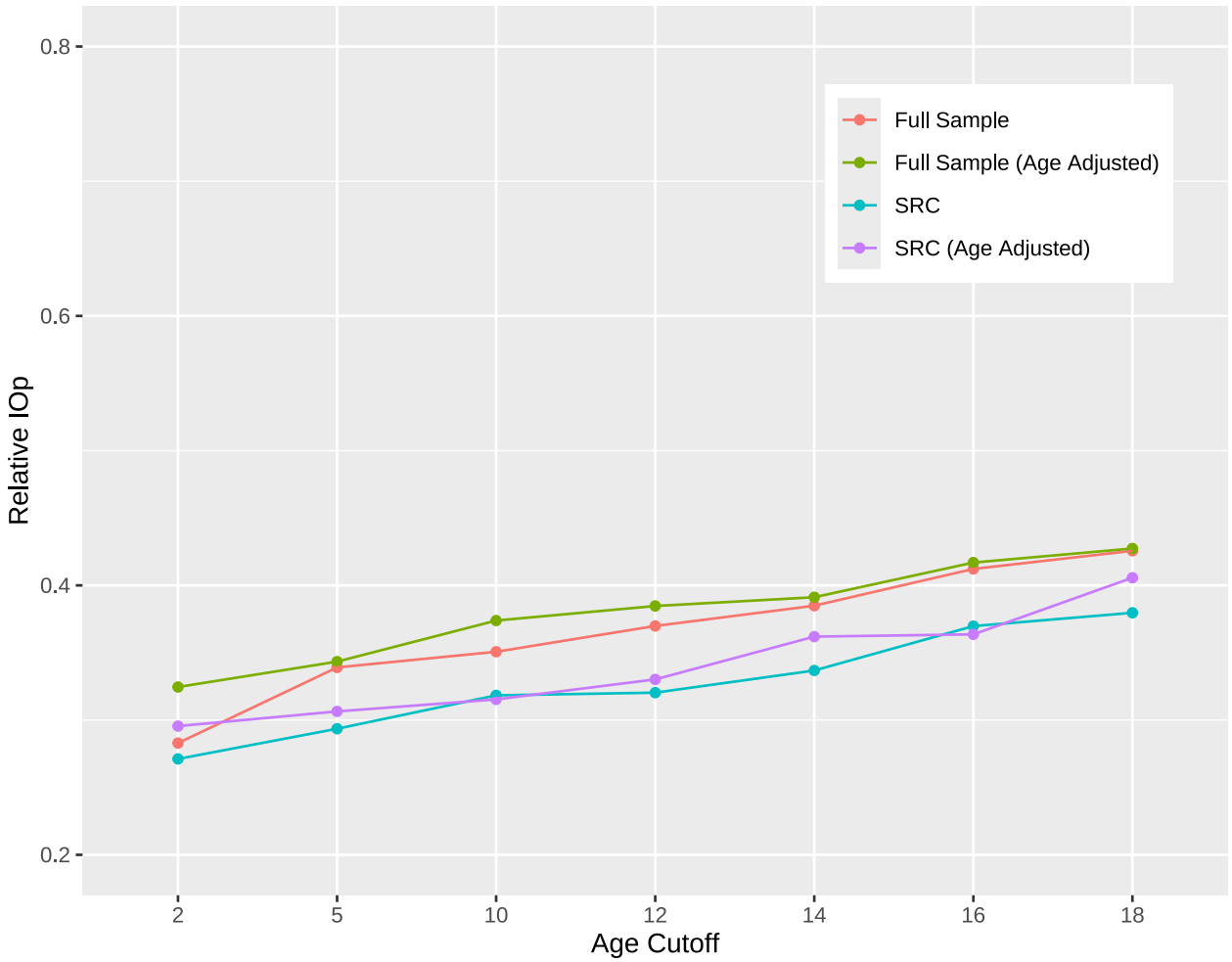

Figure 3: Relative IOp Profiles Across All Age Cutoffs

The results of this study are comparable to those found in other studies that estimate the inequality of opportunity in the US context. For instance, Pistolesi (2009) estimated the IOp to be between 20 and 43 % of earnings inequality in the US. I also compare my findings with those of Hufe et al. (2017), a study which comes the closest to the analysis performed in this paper. They estimated the IOp using circumstances at birth, at age 12, and age 16 in the US. While they used adult income measured in different years, they also reported results with average adult income for 2008-2012 using NLSY79 data. They report higher proportion of income inequality due to circumstances for age cutoffs at age 12 and 16. I compare these estimates with obtained in this study. Using labor income at 35, I obtained relative IOp 37% and 41.2% for age cut offs at 12 and 16 respectively for the full sample as well as 32% and 37% for the same cutoff using the SRC sample.

When average age adjusted incomes are used as a proxy for income in adulthood, for the full sample relative IOp estimates are 38.5% and 41.7% for age cutoffs at 12 and 16 respectively. For the SRC sample, these estimates are 33% and 36.4%.

It is important to note that I do not account for ability variables such as IQ or other test scores explicitly. One implication of dynamic complementarity is that early investments in cultivating non-cognitive skills can promote cognitive skills. A lack of early investments in disadvantaged children may lead to a lower stock of skills in subsequent years. Since children do not have control over their circumstances, these missed opportunities early in life may lead to lower stocks of skills in the future. Consequently, the inequalities generated due to these factors in outcomes should be accounted for in the measurement of IOp.

Hufe et al. (2017) argue that ability is a circumstance and categorize it as such, which may have led to estimates of relative IOp as high as 58.8 % for age cutoff at 16 years. However, I present results from circumstance set created at an age cutoff of 10, as it is well documented that IQ is rank stable after age 10 (Mackintosh 2011). For the full sample using average age adjusted incomes, 37.4%of IOp could be attributable to the circumstances faced by the child before or at age 10. This estimate is 31.5% for the analysis performed on the SRC sample. In addition to that, I permit the set of circumstances to expand with age, consistent with the formulation of the skill formation technology. Therefore, unlike Hufe et al. (2017), circumstances may reappear as the set enlarges. A set of circumstances that includes data on a child up to age 14 will be a superset of a set that contains data on a child up to age 5. The biggest set will be the set of circumstances including data on the child up to the age of consent at age 18.

5.3 Variable Importance

Displaying results from a random forest algorithm using a single tree is challenging due to the construction of multiple trees during the model fitting process. Instead, I can explore feature importance plots to understand the "importance" of different variables in constructing the trees and predicting adult income (Breiman 2001; Fisher, Rudin, and Dominici 2019).

I show variable importance plots based on permutation. The idea is to calculate the increase in model's prediction error after permuting a feature. A feature is "important" if shuffling its values increases the model

error as it implies that the model relied on the feature for prediction. If prediction error of the model does not change by much while shuffling the feature values, the feature is considered unimportant in predicting the outcome, adult income.

Let $x_1, x_2, ..., x_j$ be the features of interest and let $rmse_{base}$ be the baseline performance metric for the trained model. The permutation-based importance scores can be computed as follows:

1. For $i = 1, 2, ..., j$:
    1. Permute the values of feature x_i in the training data.
    2. Recompute the performance metric on the permuted data $rmse_{perm}$.
    3. Record the difference from baseline using $vi(x_i) = \frac{rmse_{perm}}{rmse_{base}}$.
2. Return the vi scores $vi(x_1), vi(x_2), ..., vi(x_j)$.

Figure 4 displays the results of the process explained above for the set that includes all the circumstances up to age 5. In the appendix, I show variable importance plots for all the age cut-offs. This permutation approach introduces randomness into the procedure. I repeat the procedure for 100 simulations to obtain mean importance scores and their corresponding standard deviations. Each point reflects the importance score of one of the features from the circumstance set. The bars reflect one standard deviation above and below the mean score. I show top 3 factors from the circumstance set $C_5^s \subseteq \Omega^c$ that were important in predicting the outcome variable, adult income proxied by average age adjusted labor incomes across four PSID survey waves. I show variable importance scores for both full sample and the SRC sample. Age in the labels on the X-axis refers to the age of the child.

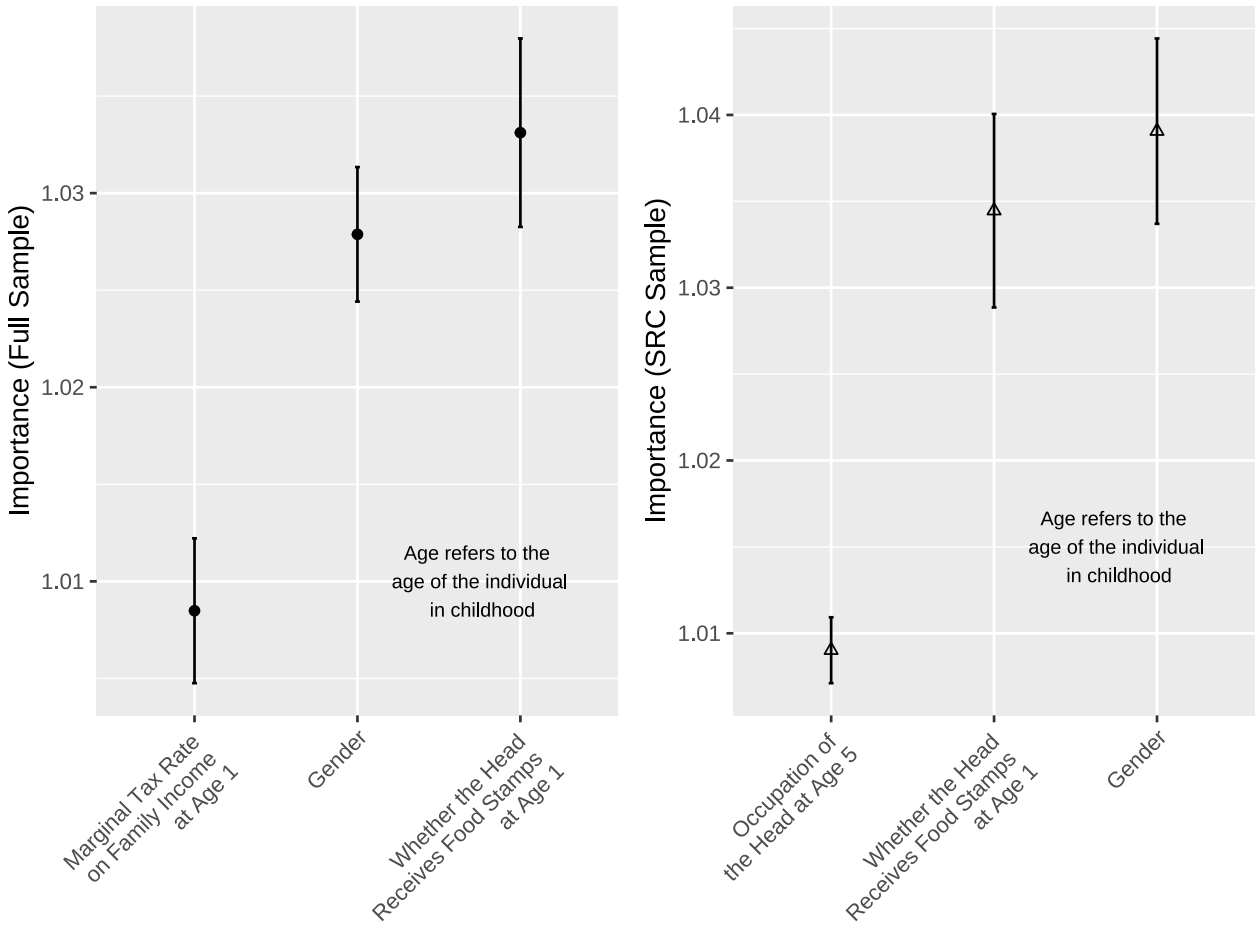

Figure 4: Variable Importance Plots for Circumstances up to Age 5

The left panel shows the important predictors in the analysis using the full sample, while the right panel displays those obtained using the SRC sample. Gender appears to be a significantly important circumstance factor in predicting an individual's adult income. This is evident as gender is the most important predictor in the SRC sample analysis and the second most important in the full sample. For the full sample, the most crucial predictor of adult income is whether the family head received food stamps when the child was 1 year old. The marginal tax rate on family income when the child was 1 year old is also a key predictor of adult income and, consequently, a contributor to inequality of opportunity using the circumstance set that includes all measurable data on a child up to age 5. In the SRC sample, the family head's occupation when the child was 5 years old is also an important predictor of adult income, along with the family head's access to food stamps when the child was one year old. It is important to note that these variable importance scores do not

indicate a causal relationship. However, they can provide an idea of how different circumstances measured at various stages in life can influence the prediction of adult income and contribute to inequality of opportunity. For example, recent evidence suggests that the timing of food stamps can have long term implications(Bond et al. 2022).

5.4 Intergenerational Income Elasticity

Policy discussions have shifted from inequality of outcome to inequality of opportunity, informed by intergenerational mobility(Corak 2013; Chetty et al. 2014). The literature on intergenerational income mobility offers variety of measures, the most popular being Intergenerational Income Elasticity (IGE). IGE is measured as a coefficient in a Galtonian regression of a child's income on parental income^[for a comprehensive discussion on full set of measures, see Deutscher and Mazumder (2023). Indeed, the IGE measure is a special case of IOp estimated using equation 6, where parental income is the sole circumstance variable(Brunori, Ferreira, and Salas-Rojo n.d. for a theoretical framework on inherited inequalities.). Usually, in the measurement of IGE, the child's income and the parent's income are averaged over multiple years to address the attenuation bias (Solon 1992; Mazumder 2005).

$$ln(y_{child}) = \alpha + \beta_{IGE}\, ln(y_{parent}) + u \tag{11}$$

Evidence suggest that the timing of parental income measured may be as or more important than a single measure of parental income (P. Carneiro et al. (2021)). I show the IGE estimates to compare them with the IOp estimates obtained in the study, considering the parental income averaged over years before and at critical stages of childhood. For parental incomes, I use family incomes from the first 18 years of the child's life. I proxy a child's permanent income using their labor income in adulthood, averaged over four years from 2013-2019. For individuals with missing income data in any wave, I calculate their average income using only the available years. For example, if an individual has income data for 2012, 2014, and 2018, but missing in 2016, I compute their average income using three years (2012, 2014, 2018). For the cohorts under consideration (born in 1978-1983), average incomes are measured at different ages. For example, in 2015, an individual born in 1978 is on average 37 years old, while someone born in 1982 is on average 33 years old. To

account for these age differences when measuring income, I include both age and age-squared terms in the equation 11. This approach follows standard practice in the intergenerational income mobility literature.

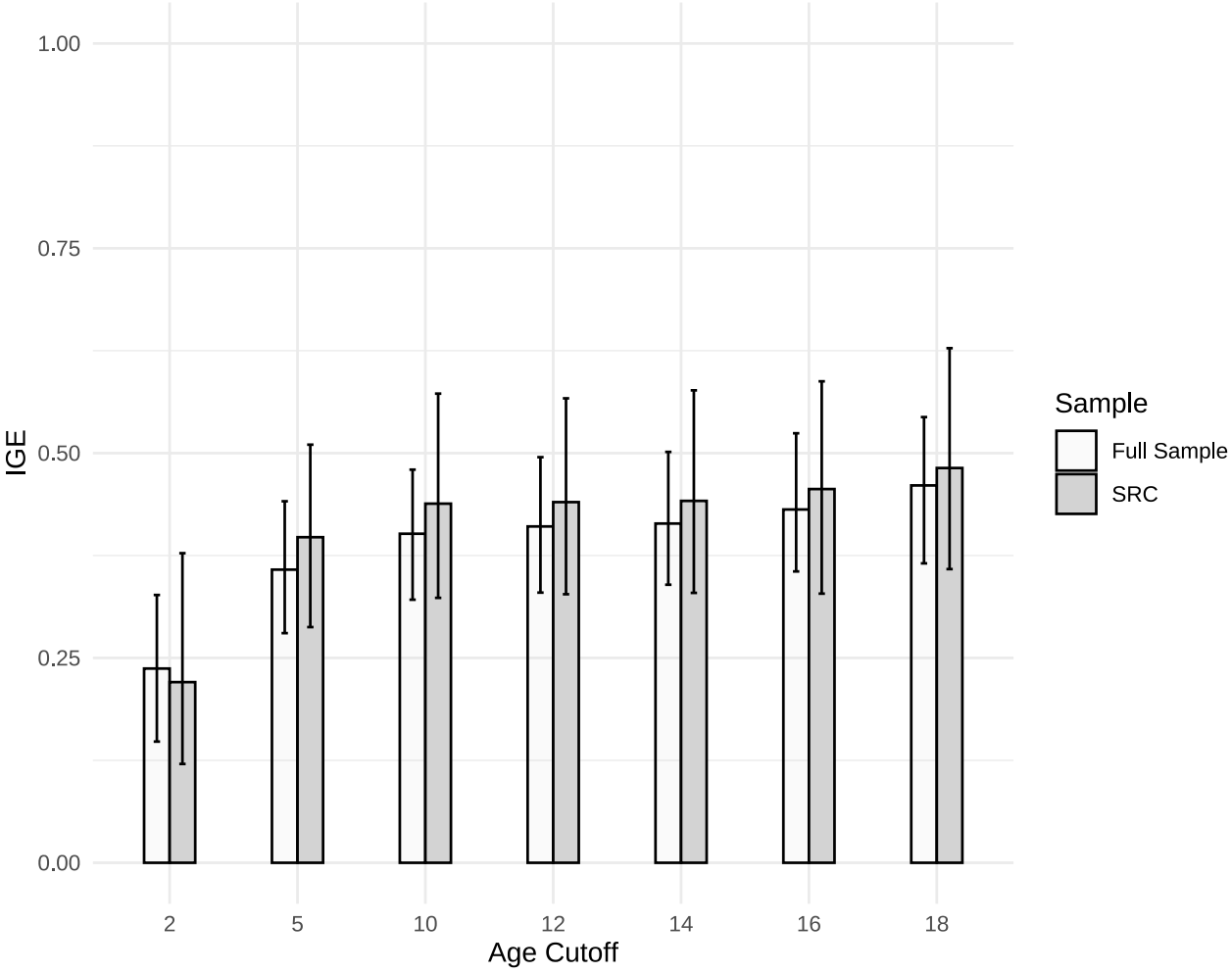

Figure 5: IGE Estimates Based on Age Cutoffs

Figure 5 displays the coefficients obtained from the IGE equation 11 , accounting for age differences at which the incomes of children are measured, with family incomes averaged over the years before and at the age cutoffs considered in the study. Once again, I report IGE coefficients using both the full sample as well as the SRC sample. For instance, IGE is 0.237 when computed using the equation with family income averaged over the first two years of the child's life for the full sample. For the SRC sample, this estimate is 0.22. Similarly, using family income averaged over the first 18 years of the child's life, I compute IGE at around 0.461 for the full sample and 0.482 for the SRC sample. We can see the IGE measures estimated in

the full and SRC samples using age cutoff at 5 is 0.358 and 0.397 respectively. The estimates do increase with age but at a decreasing rate. These measures are not causal but are comparable to IOp measures obtained previously and follow the same pattern directionally. To obtain the intervals, I use 95% bootstrap confidence intervals with 500 iterations.

5.5 The Heckman Equation and IOp

The shares of income inequality in adult income, proxied as individual labor income at age 35, attributable to unequal circumstances faced by an individual before or at age 5, are about 34% for the full sample and 29% for the SRC sample. When individual adult incomes are proxied by average age-adjusted incomes across four waves from 2012–2018, these estimates increase slightly to 34.34% for the full sample and 30.64% for the SRC sample., which is substantial especially before a child even enters kindergarten. The purpose of this study is to bring together the measurement of IOp with the idea of dynamic complementarity. Figure 6 shows the Heckman Equation that stemmed from the well documented literature in human capital and child development. There is enough evidence of dynamic complementarity in human capital investments across the childhood. Skills, or lack there of in early life tend to stay persistent in the adulthood. Return on investments made later in life to compensate for these persistent gaps are not higher if enough investment to close the early skill gaps are made.

The literature on human capital and child development provides ample evidence of how early circumstances can predict adult outcomes. P. M. Carneiro and Heckman (2003) demonstrate that significant differences in children's skills, depending on their family backgrounds, emerge at an early age and persist over time. These skill differences impact success in the labor market and other life aspects. Cunha, Heckman, and Navarro-Lozano (2004) report that approximately 60% of the residual variance in log wages can be attributed to skills developed by late adolescence. Neal and Johnson (1996) link a significant portion of the black-white wage gap for men to cognitive skill disparities identified years before these individuals enter the job market. Furthermore, Heckman, Stixrud, and Urzua (2006) highlight that both cognitive and non-cognitive skills directly influence not only labor market outcomes but also a wide range of life experiences. These include the likelihood of unemployment, welfare usage, teenage pregnancy, criminal activity participation, and drug use.

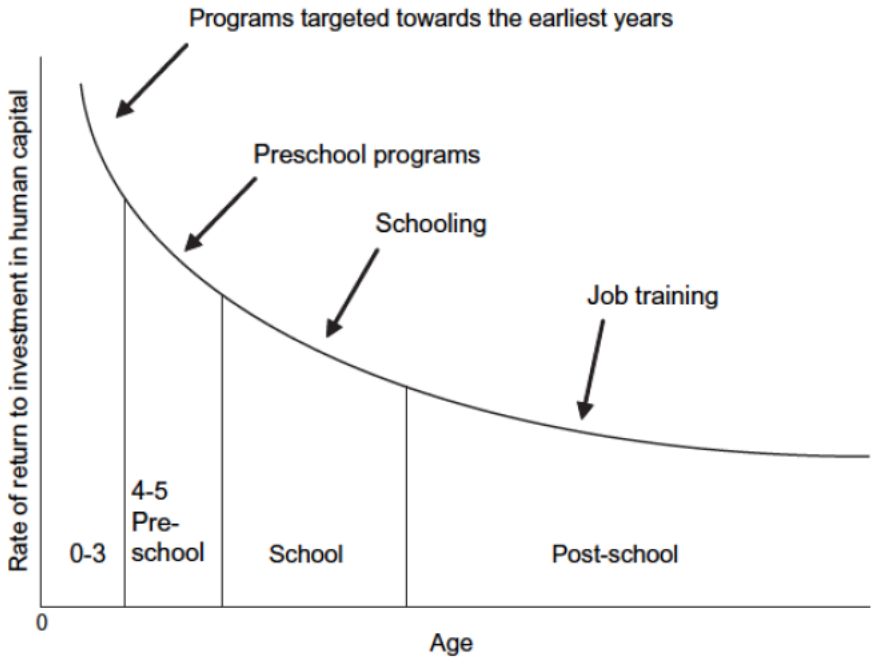

Figure 6: Source: Heckman Equation

I analyze the graph in Figure 3 in relation to the Heckman Equation to emphasize the significance of inequality generated by early childhood circumstances when measuring IOp. Figure 3 demonstrates that the majority of relative IOp stems from unequal circumstances faced by individuals up to age 5. This evidence, combined with the potential persistence of skill gaps later in life due to dynamic complementarity, warrants the consideration of early childhood opportunities in measuring inequality of opportunity, rather than relying solely on a fixed set of circumstances.

6 Conclusion

In this study, I measure income inequality of opportunity using age-based circumstance sets while accounting for the dynamic complementarity in human capital investments across the first 18 years of individuals' lives.

Using Panel Study of Income Dynamics (PSID) data on both the Survey Research Center (SRC) sample and the full sample—which also includes the Survey of Economic Opportunity sample—I found that relative inequality of opportunity (IOp) is significant. The shares of income inequality in adult income, proxied as individual labor income at age 35, attributable to unequal circumstances faced by an individual before or at

age 5, are about 34% for the full sample and 29% for the SRC sample. When individual adult incomes are proxied by average age-adjusted incomes across four waves from 2012–2018, these estimates increase slightly to 34.34% for the full sample and 30.64% for the SRC sample. These figures are higher than those obtained using a fixed set of circumstances in OLS regression. Focusing on circumstances faced by individuals before or at age 5, I used random forest—a supervised machine learning algorithm—to identify gender, access to food stamps, and housing as "important" circumstances in predicting adult incomes. This identification was based on a variable importance plots. Accounting for dynamic complementarity, the relative IOp estimates increase with the expansion of circumstances as age increases, but at a decreasing rate. Unequal circumstances up to the age of consent account for about 37% to 43% of the inequality in adult incomes.I also calculated intergenerational income elasticity (IGE), another measure of inequality of opportunity. Using average family incomes from a child's first five years as a proxy for parental income, I found IGE coefficients of 0.358 for the full sample and 0.397 for the SRC sample. The pattern of IGE estimates across age cutoffs mirrors the trend of relative IOp measured in this study.

These findings have limitations. First, due to partial observability of circumstances, I only obtain lower bounds of absolute and relative IOp. Consequently, the role of unequal circumstances in total inequality in adult income is underestimated. Second, the boundary between circumstances and effort factors at age 18 could be seen as arbitrary, given the persistent effects of circumstances even after this age. I acknowledge this, though the goal here is to focus on early childhood circumstances and account for dynamic complementarity in human capital investments during children's formative years. It is difficult to argue that a 5-year-old has any control over their circumstances. Finally, I recognize that the estimation of inequality of opportunity using machine learning algorithms is only as good as the next best algorithm. There's always room for improvement through better algorithms and hyperparameters.

Rigorously measuring unfair income inequality is crucial if equalizing opportunities is a public policy goal. This study identifies "unfair" inequality by incorporating early childhood circumstances into the measurement of inequality of opportunity (IOp). This approach to measuring IOp could inform the development of public policies aimed at equalizing opportunities in early childhood. Additionally, Roemer (1993) proposes ex-post compensation for individuals who experience outcome inequalities due to their unequal circumstances.

Measuring IOp by accounting for unequal childhood circumstances will better inform policies based on the principle of compensation. While this research focuses on childhood circumstances in measuring inequality of opportunity (IOp), it would be valuable to examine cross-country differences in the share of unequal opportunities using age-based circumstance sets. This approach could complement the current practice of using fixed sets of circumstances. Furthermore, this analysis could be extended to other outcomes such as health and education, as well as exploring geographical heterogeneity in IOp estimates.

7 References

Alesina, Alberto, and Paola Giuliano. 2011. "Chapter 4 - Preferences for Redistribution." In *Handbook of Social Economics*, edited by Jess Benhabib, Alberto Bisin, and Matthew O. Jackson, 1:93–131. North-Holland. https://doi.org/10.1016/B978-0-444-53187-2.00004-8.

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz. 2011. "Personality Psychology and Economics." In *Handbook of the Economics of Education*, 4:1–181. Elsevier. https://doi.org/10.1016/B978-0-444-53444-6.00001-8.

Becker, Gary S., and Nigel Tomes. 1986. "Human Capital and the Rise and Fall of Families." *Journal of Labor Economics* 4 (3): S1–39. https://www.jstor.org/stable/2534952.

Bond, Timothy N., Jillian B. Carr, Analisa Packham, and Jonathan Smith. 2022. "Hungry for Success? SNAP Timing, High-Stakes Exam Performance, and College Attendance." *American Economic Journal: Economic Policy* 14 (4): 51–79. https://doi.org/10.1257/pol.20210026.

Borghans, Lex, Angela Duckworth, James Heckman, and Bas Weel. 2008. "The Economics and Psychology of Personal Traits." *The Journal of Human Resources* 43 (February). https://doi.org/10.1353/jhr.2008.0017.

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.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Brunori, Paolo, Francisco H. G. Ferreira, and Pedro Salas-Rojo. n.d. "Inherited Inequality: A General Framework and an Application to South Africa." *Working Papers*. Accessed July 30, 2024. https://ideas.repec.org//p/inq/inqwps/ecineq2023-658.html.

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 Manuel, and James J. Heckman. 2003. "Human Capital Policy." {SSRN} {Scholarly} {Paper}. Rochester, NY. https://doi.org/10.2139/ssrn.434544.

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.

Checchi, Daniele, and Vito Peragine. 2010. "Inequality of Opportunity in Italy." *J Econ Inequal* 8 (4): 429–50. https://doi.org/10.1007/s10888-009-9118-3.

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.

Cowell, Frank. 2016. "Inequality and Poverty Measures." *Oxford Handbooks Online*, January. https://www.academia.edu/69017672/Inequality_and_Poverty_Measures.

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.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov. 2006. "Chapter 12 Interpreting the Evidence on Life Cycle Skill Formation." In *Handbook of the Economics of Education*, 1:697–812. Elsevier. https://doi.org/10.1016/S1574-0692(06)01012-9.

Cunha, Flavio, James J. Heckman, and Salvador Navarro-Lozano. 2004. "Separating Uncertainty from Heterogeneity in Life Cycle Earnings." {SSRN} {Scholarly} {Paper}. Rochester, NY. https://doi.org/10.2139/ssrn.643622.

Deutscher, Nathan, and Bhashkar Mazumder. 2023. "Measuring Intergenerational Income Mobility: A Synthesis of Approaches." *Journal of Economic Literature* 61 (3): 988–1036. https://doi.org/10.1257/jel.20211413.

Donni, Paolo Li, Juan Gabriel Rodríguez, and Pedro Rosa Dias. 2015. "Empirical Definition of Social Types in the Analysis of Inequality of Opportunity: A Latent Classes Approach." *Social Choice and Welfare* 44 (3): 673–701. https://www.jstor.org/stable/43662611.

Eskelson, Bianca, Hailemariam Temesgen, Valerie Lemay, Tara Barrett, Nicholas Crookston, and Andrew Hudak. 2009. "The Roles of Nearest Neighbor Methods in Imputing Missing Data in Forest Inventory

and Monitoring Databases." *United States Department of Agriculture, Forest Service / University of Nebraska-Lincoln: Faculty Publications*, January. https://digitalcommons.unl.edu/usdafsfacpub/217.

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.

Ferreira, Francisco H. G., and Vito Peragine. 2015. "Equality of Opportunity: Theory and Evidence." {SSRN} {Scholarly} {Paper}. Rochester, NY. https://papers.ssrn.com/abstract=2584375.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2019. "All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *J Mach Learn Res* 20: 177. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8323609/.

Fleurbaey, Marc, and Vito Peragine. 2013. "Ex Ante Versus Ex Post Equality of Opportunity." *Economica* 80 (317): 118–30. https://doi.org/10.1111/j.1468-0335.2012.00941.x.

Fong, Christina. 2001. "Social Preferences, Self-Interest, and the Demand for Redistribution." *Journal of Public Economics* 82 (2): 225–46. https://doi.org/10.1016/S0047-2727(00)00141-9.

Gower, J. C. 1971. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics* 27 (4): 857–71. https://doi.org/10.2307/2528823.

Hart, Betty, and Todd R. Risley. 1995. *Meaningful Differences in the Everyday Experience of Young American Children*. Brookes Publishing Company, Inc.

Heckman, James J., and Stefano Mosso. 2014. "The Economics of Human Development and Social Mobility." *Annu. Rev. Econ.* 6 (1): 689–733. https://doi.org/10.1146/annurev-economics-080213-040753.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." {SSRN} {Scholarly} {Paper}. Rochester, NY. https://papers.ssrn.com/abstract=881240.

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.

Kalil, Ariel. 2015. "Inequality Begins at Home: The Role of Parenting in the Diverging Destinies of Rich and

Poor Children." In *Families in an Era of Increasing Inequality: Diverging Destinies*, 63–82. National Symposium on Family Issues. Cham, Switzerland: Springer International Publishing/Springer Nature. https://doi.org/10.1007/978-3-319-08308-7_5.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95. https://doi.org/10.1257/aer.p20151023.

Lareau, Annette. 2011. "Unequal Childhoods: Class, Race, and Family Life." In *Unequal Childhoods*. University of California Press. https://doi.org/10.1525/9780520949904.

Mackintosh, Nicholas. 2011. *IQ and Human Intelligence*. Second Edition, Second Edition. Oxford, New York: Oxford University Press.

Mazumder, Bhashkar. 2005. "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data." *The Review of Economics and Statistics* 87 (2): 235–55. https://doi.org/10.1162/0034653053970249.

Neal, Derek A., and William R. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy* 104 (5): 869–95. https://doi.org/10.1086/262045.

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.

Nybom, Martin, and Jan Stuhler. 2017. "Biases in Standard Measures of Intergenerational Income Dependence." *Journal of Human Resources* 52 (3): 800–825.

Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas. 2012. "How Many Trees in a Random Forest?" In *Machine Learning and Data Mining in Pattern Recognition*, edited by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, et al., 7376:154–68. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31537-4_13.

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.

Ramos, Xavier, and Dirk Van de gaer. 2016. "Approaches to Inequality of Opportunity: Principles, Measures

and Evidence." *Journal of Economic Surveys* 30 (5): 855–83. https://doi.org/10.1111/joes.12121.

Rawls, John. 1971. *A Theory of Justice: Original Edition*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9z6v.

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.

———. 2002. "Equality of Opportunity: A Progress Report." *Social Choice and Welfare* 19 (2): 455–71. https://www.jstor.org/stable/41106460.

Roemer, John E., and Alain Trannoy. 2016. "Equality of Opportunity: Theory and Measurement." *Journal of Economic Literature* 54 (4): 1288–1332. https://doi.org/10.1257/jel.20151206.

Solon, Gary. 1992. "Intergenerational Income Mobility in the United States." *The American Economic Review* 82 (3): 393–408.

Starmans, Christina, Mark Sheskin, and Paul Bloom. 2017. "Why People Prefer Unequal Societies." *Nat Hum Behav* 1 (4): 1–7. https://doi.org/10.1038/s41562-017-0082.

Tutz, Gerhard, and Shahla Ramzan. 2015. "Improved Methods for the Imputation of Missing Data by Nearest Neighbor Methods." *Computational Statistics & Data Analysis* 90 (C): 84–99. https://ideas.repec.org//a/eee/csdana/v90y2015icp84-99.html.

Van de Gaer, Dirk. 1993. *Equality of Opportunity and Investment in Human Capital*. Reeks van de Faculteit Der Economische En Toegepaste Economische Wetenschappen, Katholieke Universiteit Te Leuven.

Appendix

7.1 Family Relationship Matrix

The Family Relationship Matrix File (FRM), spanning from 1968 to 2019, is designed to consolidate all existing relationship data gathered during these years as part of the Panel Study of Income Dynamics (PSID). This file outlines the relationship between each individual in a PSID Family Unit (FU) and all other members within the same FU for each wave from 1968 to 2019. Each record set for a specific individual shows their relationship to all other FU members during that wave. Only individuals who resided in the FU at the time of the interview are included in the file for each wave.I use this FRM file to identify the family heads of the individuals of interest during their first 18 years of life. As the individuals were born between 1978-1983, the interview waves are limited to the period from 1978-2001. This matrix provides information on whether the family head was the individual's parent before they reached the age of 18. It's also possible for the family head to be a non-parent. Moreover, the family head can change over the course of an individual's first 18 years.

7.2 Family Identification Mapping System

FIMS relies on information stored in the Parent Identification File (PID), a cumulative file compiled from several PSID data sources. The PID summarizes information about parent-child relationships collected from various sources since the 1983 wave of the Panel Study of Income Dynamics (PSID). This file includes identifier variables that link children to their birth and adoptive parents, and it also indicates the source of the information. I use FIMS to identify whether the parent-head of a child is a mother or a father.

7.3 Handling Missing Data

The analysis uses wide data, where each observation is a joint distribution of all measurable data on individuals. All of these are considered circumstances over which individuals have no control. The sample suffers from missing data issues.

When the training sample is moderate in size, one effective method to impute missing data is the K

nearest neighbors algorithm (Eskelson et al. 2009; Tutz and Ramzan 2015). I use this algorithm to impute missing data in my sample.

The procedure finds a sample with one or more missing values and then identifies the K most similar samples in the complete training data with no missing values. Similarity of samples is defined by a distance metric. After computing this distance metric, the nearest K samples to the sample with the missing value are identified, and the mean value is calculated. This mean value is then used to replace the missing value in the sample.

Usually, Euclidean distance is used as a sample similarity metric when all the features are numeric. In this study, I have 396 features (for a full dataset) with missing values for both numeric and categorical features. Instead of using Euclidean distance, I use a good alternative called Gower's distance (Gower (1971)). This distance metric uses separate measures for both numeric and categorical features. For a categorical feature, the distance between two samples is 0 if the samples have the same value and 1 otherwise. For a numeric feature, the sample distance between two observations is defined as

$$d(x_i, x_j) = 1 - \frac{|x_i - x_j|}{R_x} \tag{12}$$

where R_x is the range of the feature for which the missing values are being imputed using KNN. This measure is computed for each feature, and the average distance is used as an overall distance. Once the K neighbors are found, their mean values are used to impute the missing data. For categorical features, the mode is used, while an average or a median can be used for numeric data. I use the average in my analyses. I explain the feature engineering steps in next section.

7.4 Data Preprocessing

While converting long data to wide data based on an individual's age in childhood, some columns have all values missing. I start by removing these columns. Next, I exclude features where more than 50% of the values are missing from my analysis.

I then run the KNN algorithm to impute the missing values in the rest of the data for all quantitative and qualitative features using Gower's distance to measure the distance between neighbors. A rule of thumb

is to use $k = \sqrt{n}$, where k is the number of neighbors and n is the number of observations in the sample. I settled on k = 31 while using the KNN algorithm for missing data imputation.

For the remaining features, I remove numerical features with near-zero variance. Finally, to address multicollinearity, I remove numeric features with a correlation greater than 0.8 with other features.

7.5 Tuned Hyperparameters

Table 3: Tuned Hyperparameters

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

Table 3 in the appendix lists the hyperparameter values obtained through a 5-fold cross-validation process for each circumstance set, based on their respective age cut-offs. I utilize three essential hyperparameters for building a random forest model.

- *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.

To reduce the complexity and run time of the code, I only tune *min_n* parameter using 5-fold cross validation. The number of trees are chosen following the standard practice in the literature of machine

learning. I keep the number of trees arbitrarily high and do not tune that parameter Oshiro, Perez, and Baranauskas (2012). Following the same practice, I do not tune *mtry* hyperparameter. I choose high enough number for this hyperparameter instead of tuning it to reduce the run time as well as the model complexity.

Hyperparameter *min_n* indirectly controls the tree depth. Higher the tree depth, lesser the model complexity. Each value of *min_n* in the table is obtained through a 5-fold cross-validation process, repeated twice. The model with the lowest root mean square error (rmse) - as indicated by the hyperparameters in the table - is selected. This model is then fitted on the entire dataset to generate a counterfactual distribution of predictions, based on the factors in the respective circumstance sets.

7.6 IOp Estimates for All Age Cutoffs

Table 4: IOp Estimates for All Age Cutoffs

	Full Sample (N = 1022)			SRC sample (N = 639)		
	Income Inequality	Absolute IOp	Relative IOp	Income Inequality	Absolute IOp	Relative IOp
Outcome : Labor Income at age 35						
Cutoff at age 2	0.368	0.104	0.283	0.337	0.091	0.271
Cutoff at age 5	0.368	0.125	0.339	0.337	0.099	0.294
Cutoff at age 10	0.368	0.129	0.351	0.337	0.107	0.318
Cutoff at age 12	0.368	0.136	0.370	0.337	0.108	0.320
Cutoff at age 14	0.368	0.141	0.385	0.337	0.113	0.337
Cutoff at age 16	0.368	0.152	0.412	0.337	0.125	0.370
Cutoff at age 18	0.368	0.156	0.426	0.337	0.128	0.380
Outcome : Age-adjusted Labor Income						
Cutoff at age 2	0.327	0.106	0.325	0.308	0.091	0.296
Cutoff at age 5	0.327	0.112	0.343	0.308	0.094	0.306
Cutoff at age 10	0.327	0.122	0.374	0.308	0.097	0.315
Cutoff at age 12	0.327	0.126	0.385	0.308	0.102	0.330
Cutoff at age 14	0.327	0.128	0.391	0.308	0.111	0.362
Cutoff at age 16	0.327	0.136	0.417	0.308	0.112	0.364
Cutoff at age 18	0.327	0.140	0.427	0.308	0.125	0.406

7.7 Relative IOp Estimates Using Gini

In my main study, I use mean logarithmic deviation (MLD). Any standard inequality measure that satisfies anonymity, the principle of transfers, population replication, and scale invariance could be used. Here, I present the IOp estimates along with their contributions to total inequality, as measured by the Gini coefficient.

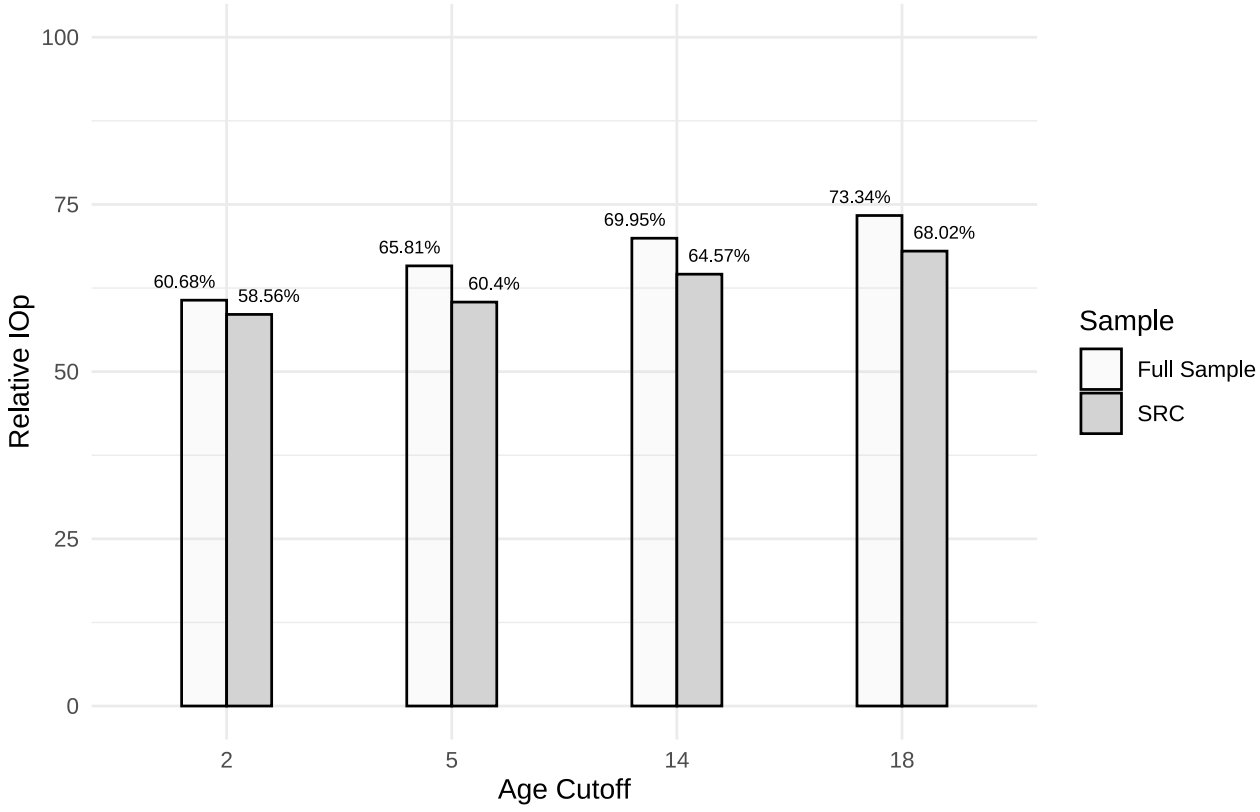

Figure 7: Relative IOp Estimates Across Age Cutoffs (Using Individual Labor Income at Age 35)

Figure 7 shows the shares of IOp using the Gini coefficient as the inequality measure. Despite the shares being higher, the upward trend until the age of consent at 18 aligns with what is observed when using MLD as the inequality measure. Most of the income inequality, approximately 60%, attributed to the inequality of opportunity, stems from circumstances at or before the age of 2.

7.8 Relative IOp Estimates Using Gini

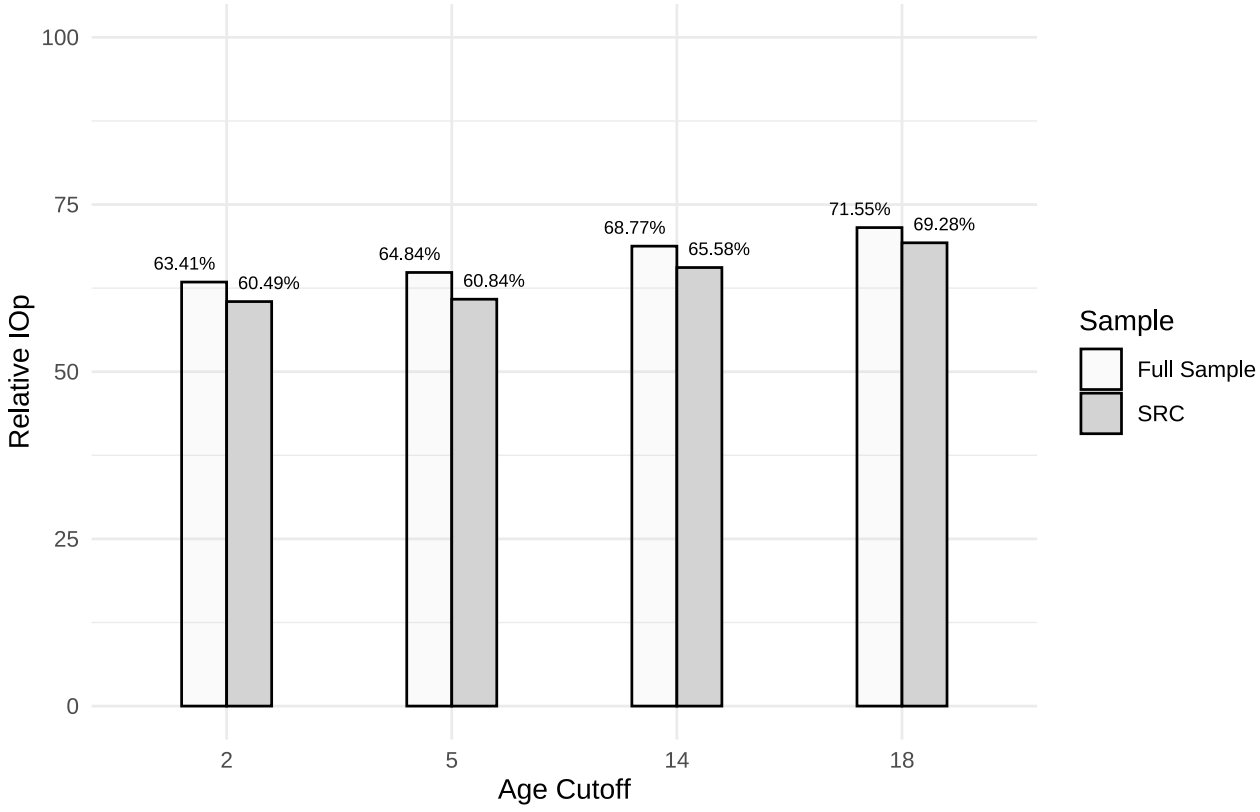

Figure 8: Relative IOp Estimates Across Age Cutoffs (Using Averaged Age-adjusted Incomes Across 2013-2019 Waves)

7.9 Variable Importance Plots for Different Age Cut-offs

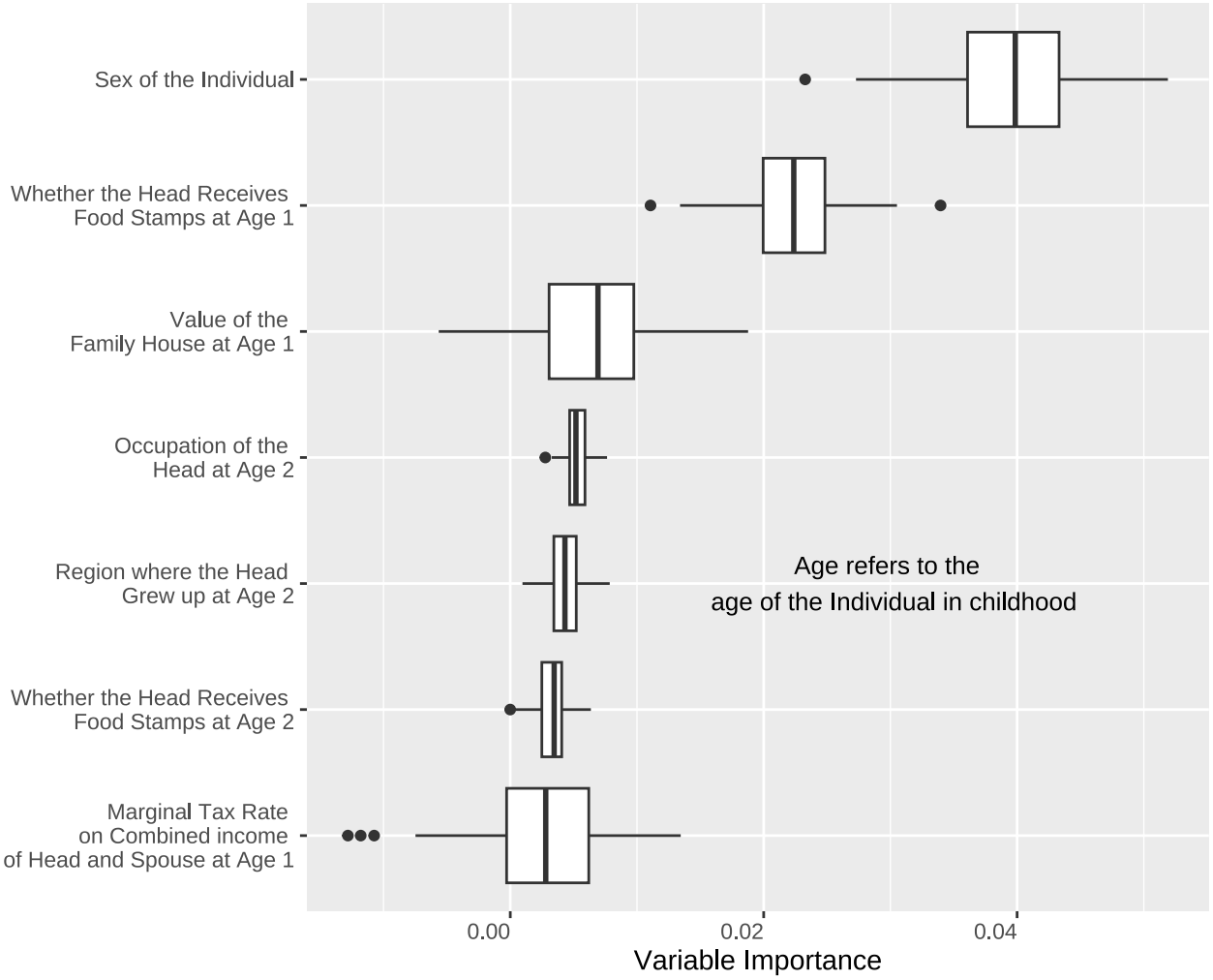

Figure 9: Variable Importance Plot of Circumstances at Age 2

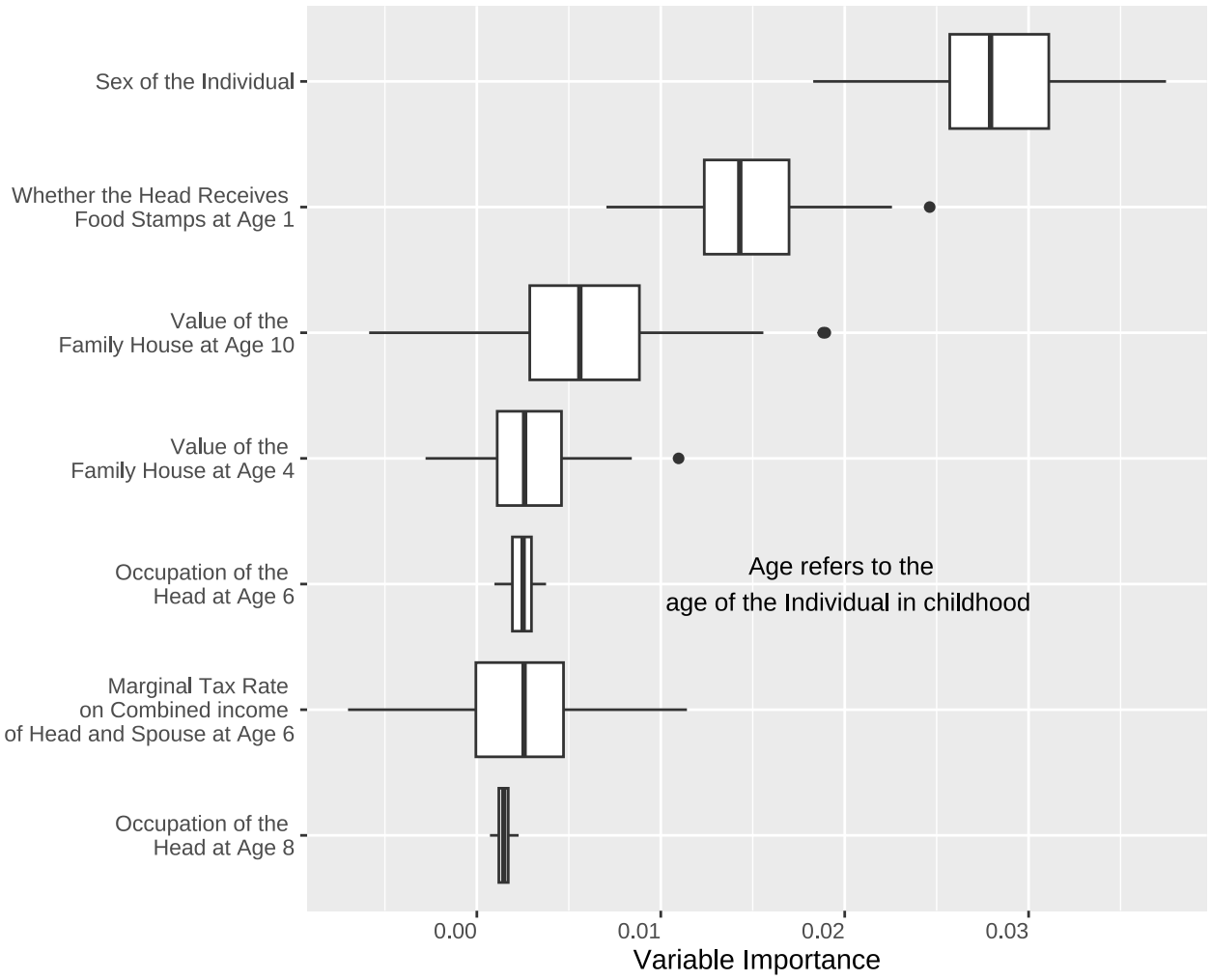

Figure 10: Variable Importance Plot of Circumstances at Age 14

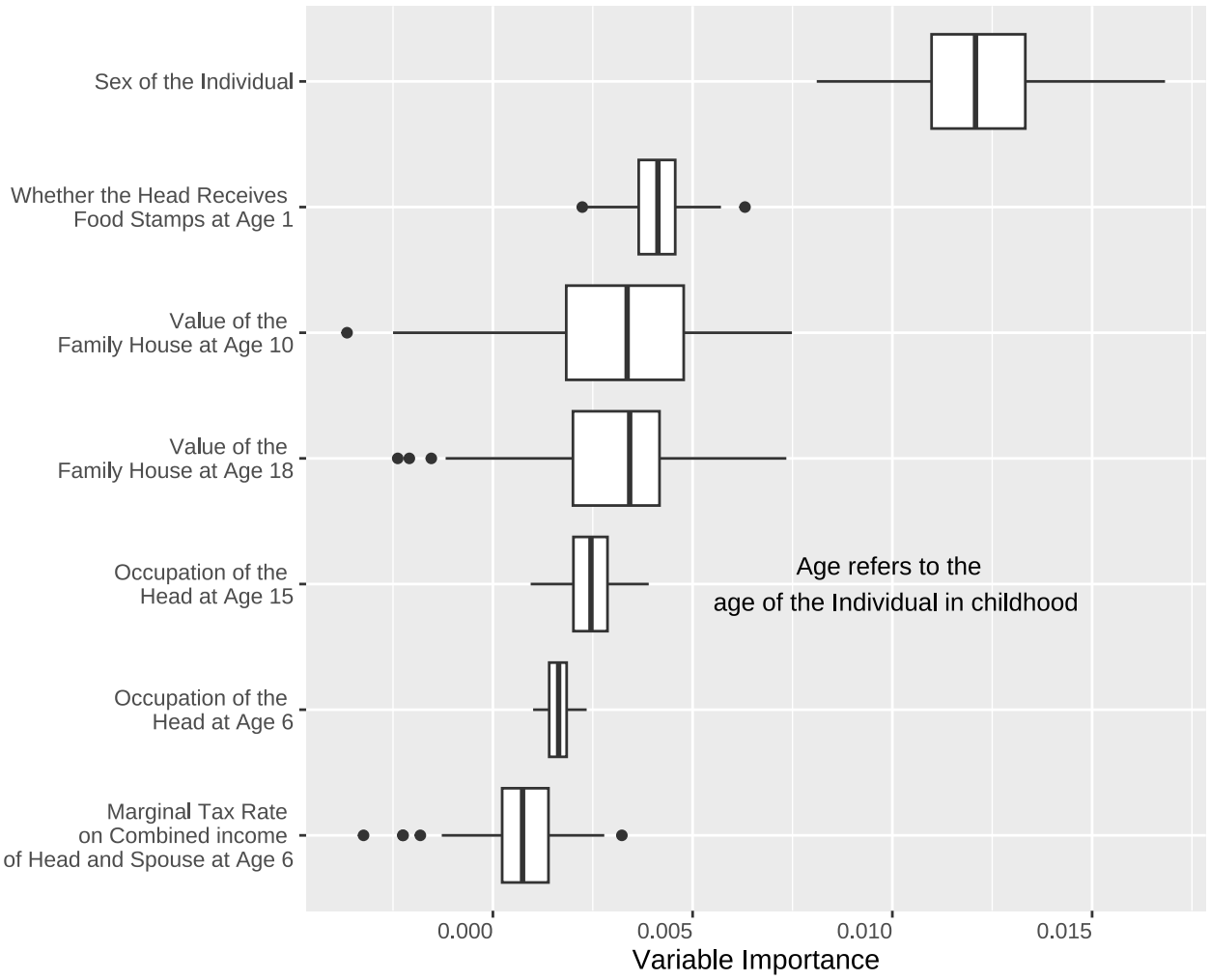

Figure 11: Variable Importance Plot of Circumstances at Age 18