Inequality of Opportunity in Indian Society*

Arnaud Lefranc† and Tista Kundu‡

May 31, 2019

Abstract

Using data from the National Sample Survey we estimate inequality of opportunity for India, in consumption expenditure, wage earning and education, on the basis of caste, sex, region and parental backgrounds as our circumstances. We use the widely used methods of non-parametric and parametric analysis to find that even in 2011-12, more than one-fourth of total inequality in wage and education is due to unequal circumstances. But as compared to the other two outcomes, we find inequality of opportunity in consumption to be relatively low. We further provide the opportunity tree for India using the recently introduced method of the regression tree analysis and find parental backgrounds as the most important circumstances for all outcomes. The opportunity tree also reveals a hierarchical order among the circumstances that are most relevant for the underlying unequal opportunity in the country.

JEL Classification : D31, D63, I24

Keywords : Inequality of opportunity, caste, consumption expenditure, income distribution, education, parental background, multiple imputation, mean log deviation, regression tree, India

*We are grateful to Daniel Mahler and Geoffrey Teyssier for helpful comments and discussion.

†THEMA, Université de Cergy-Pontoise, Email: arnaud.lefranc@u-cergy.fr

‡Corresponding author. Department of Economics, ESSEC Business School, Paris & THEMA, Université de Cergy-Pontoise; Email: tista.kundu@essec.edu

1 Introduction

"*..The service to India means the service of the millions who suffer. It means the ending of poverty and ignorance and decease and inequality of opportunity.*" - Jawaharlal Nehru[1]

Seventy years have passed after this speech is made at the stroke of midnight on the very first day of independence of India. Over this span, India from an impoverished country, made her journey to one of the emerging global economy now. Especially since the late nineties, with a consistent high GDP growth rate of more than 7%, India has now become the sixth largest economy in the world. Much work has been accomplished with significant improvement in overall well-being of the country, but much enough, if not more, remains to be done or even addressed. Numerous studies have showed that the rapid growth of India has been accompanied by increasing inequality as well. However very few studies have yet been done to explore how much of the growing inequality is due to *inequality of opportunity*, that is how much of this high inequality is generated by factors that are purely fatalistic and therefore beyond any human control.

India followed an interventionist central planning for the first forty years after independence followed by 'neo-liberal' economic reforms at the beginning of 1990s. Since then, both the overall growth rate and inequality in India grew almost simultaneously, making it a very relevant and active area of research concerning India. A sharp increase in consumption inequality along with a slower pace of poverty reduction has almost become a distinct feature of the Indian economy, especially in the twenty-first century[2]. But for a very stratified society like India, while there are wealth of literature on analyzing the problem of inequality, linking it to social mobility, labor market discrimination, urbanization or poverty, only a handful of them analyze how much of this inequality is due to unequal opportunities arising from varying social and family backgrounds, for which no one can be held accounted for.

The present work aspires to quantify the degree of unequal opportunity in India by estimating how much of inequality in consumption, wage and education is due to differences in caste, sex, region, parental education and occupation. Traditionally inequality had been assessed following a welfarist approach, where inequality in the final outcome was the main focus of analysis. Unequal distribution of any desirable outcome (*e.g.* income, education, standard of living, health etc.) are of primary concern for assessing social welfare. However inequality can arise from an array of different factors, some of which are purely fatalistic to the individuals. This heterogeneity in the inequality generating factors had actually triggered a philosophical debate in the late twentieth century, criticizing the fact that the classical welfarist way of inequality analysis is an approach too consequentialist to take into account the multifaceted nature of the inequality generating process (Rawls 1971, Dworkin 1981*b*,*a*). The main point of the debate is that inequality arising from factors on which no individual has any control, like race, sex, ethnicity, religion, birthplace, parental and family background, should be of primary concern from an ethical standpoint and should therefore be considered as rather *unfair*. On the other hand inequality generated from

[1] Excerpt from *'Tryst with Destiny'* - a speech delivered on the first day of independence, 15th August 1947, by Jawaharlal Nehru, the first Prime Minister of independent India.

[2] *See* Deaton & Dreze (2002), Himanshu (2007), Dev & Ravi (2007), for example. For the recent updates on Indian inequality, *see* India inequality report by Himanshu (2018).

unregulated lifestyle, lack of perseverance, inadequate skill formation or poor managing ability, in other words, factors for which one can arguably be held responsible for, are not unethical and unfair in an egalitarian society.

This new approach of analyzing inequality by splitting it into fair and unfair part, brings about the question of individual responsibility in the domain of distributive justice and started to prioritize the analysis of inequality arising solely from the factors that are beyond subjective responsibility (Arneson 1989, Cohen 1989). Inspired by this philosophical debate on the responsibility sensitive egalitarian justice, Roemer (1993) formulates inequality of opportunity as that part of inequality that is generated by factors beyond any individual control. In the jargon of inequality of opportunity (*IOP*), all such factors that are outside the periphery of individual responsibility but are responsible for generating inequality, are called *circumstances*. On the other hand the inequality generating factors that the individual can presumably control, are called *efforts*. In this dichotomous standpoint of *effort* versus *circumstances*, inequality of opportunity is that (unfair) part of inequality that had been generated only by the *circumstance factors* (Roemer 1998).

Methodologically both non-parametric and parametric approaches serve the literature to estimate the measure of *IOP* in a society. The backbone structure of these methods attributes to Checchi & Peragine (2010) (for Italy) and Bourguignon, Ferreira & Menéndez (2007) (for Brazil), respectively for the non-parametric and the parametric estimates. Although the parametric estimates of *IOP* comes at the cost of a specific functional form assumption between the outcome and the circumstance variables, it is often recommended for studies with a broad range of circumstances. Whereas the use of non-parametric approach is more common for multi-country comparison studies that is limited to a comparable set of circumstances across the countries. So far in the literature there is no universal consensus to prioritize one approach over another. But in either set up to quantify the unfair part of inequality as a measure of *IOP*, majority of the literature use an index from the generalized entropy class of inequality indices, that of the index of *mean log deviation*. Using a slightly different non-parametric and parametric set up, Ferreira & Gignoux (2011) nevertheless showed that the estimates of *IOP* are significantly close regardless of the method adopted. This is the methodological set up that we will use for measuring the index of *IOP* in India[3].

Two of the major shortcomings of the above mentioned approaches are that they are based on a pre-specified number of circumstances and often uses all possible interactions of the chosen circumstances to estimate *IOP*, while in reality only some of the interactions may be meaningful. However there is no way that either of the non-parametric or the parametric set-up can point out the relevant interactions. Besides including all possible interactions also increases the total number of circumstance groups to compare, which may lead to an overestimated *IOP* as the number of observations per cell decreases. The problem is even more aggravated when some

[3] *See* Roemer & Trannoy (2013), Ramos & Van de Gaer (2012) for an extensive analysis on the major methodologies used in the literature. For some international estimates of IOP, *see* Brunori, Ferreira & Peragine (2013) (selected developed countries including some Nordic countries, selected Latin American, African, Middle-East and Asian countries), Ferreira & Gignoux (2011) (Latin American countries), Marrero & Rodrìguez (2011) (Unites States of America), Checchi, Peragine & Serlenga (2010) (European countries), Cogneau & Mesplè-Somps (2008) (African countries).

of the interactions are almost vacuous, leaving very few unusual observations in some cells. To address this problem, Brunori, Hufe & Mahler (2018) introduced a novel approach of analyzing *IOP* using the *regression tree* analysis that let the algorithm choose the most relevant circumstances in a statistically significant way from the submitted set of circumstances and generates a visually interpretative opportunity tree in the hierarchical order of circumstances. Therefore along with the non-parametric and parametric estimation of *IOP*, we adopt this approach for the present work as well to provide the opportunity structure for contemporary India.

India epitomizes a very hierarchical social structure historically, where the century old caste system is functional even in the twenty-first century. For such a stratified country there is almost no work analyzing unequal opportunity in India, with two notable exceptions. Using the National Sample Survey data, Asadullah & Yalonetzky (2012) analyzed educational opportunity in different states of India due to differences in sex, religion and caste. However for being a state-level study it is naturally focused more on the inter-state differences in terms of unequal opportunity in education rather than the national estimate. Besides due to the structure of the data base, their study can not take into account parental background as one of their circumstances which is repeatedly shown as one of the major driving factor behind unequal opportunities in a number of developed and developing countries. With a different survey Singh (2012) gives a national estimate of *IOP* in India for consumption and income, that includes father's educational and occupational background as two of the major circumstances. But due to the survey structure, the inclusion of parental background limits this study to Indian men only. Besides none of the above studies gives the recent picture of India, as the latest time frame in either work is *2004-05*. The scanty work on *IOP* in India leaves significant scope of further improvement. The aim of the present paper is to provide the latest estimates of *IOP* in India using both the non-parametric and parametric methodology, as well as to provide the opportunity structure for contemporary India adopting the recently introduced approach of the regression tree analysis.

In particular we choose three outcome variables to analyze, namely, consumption expenditure, wage earning and year of education, and analyze *IOP* for a set of five circumstance variables comprising of caste, sex, region, parental education and occupation. The present work contributes to the literature in several ways. First, using the most recent survey rounds of the National Sample Survey from *2004* to *2012*, our study gives a rather recent picture of unequal opportunity in India. We found that even by *2012*, more than one-fourth of the total wage and educational inequality is due to differences in the taken circumstances. This positions India as one of the high opportunity unequal countries in the global perspective. Second, due to the structure of the National Sample Survey it is difficult to incorporate parental information into the analysis, as the survey questionnaire have no direct provision of reporting this information. Instead parental attributes are only available for the co-resident households where parents are enumerated along with their offspring. This immediately raises the question of selection bias due to co-residence.

The present study overcome this problem by imputing information on parental background for the general sample by the widely used technique of multiple imputation (Rubin 1986). We thereby produce the estimates of *IOP* by taking into account the important circumstances of

parental backgrounds but without limiting the study to the co-resident households. In fact we found that ignoring parental backgrounds as circumstances results in considerable underestimation of *IOP*, as the loss in information due to omitting parental attributes can not be captured well by the other social circumstantial backgrounds considered, like caste, sex or region. Besides, in spite of the prevalent evidence of casteism in India, differences only on the basis of caste groups is found to be not enough to capture the differences in economic opportunity arising from other sources like family backgrounds. However the opportunity structure of India shows that while sex become rather relevant when parents have little or no experience of formal schooling, the forward caste premium is not limited to the lesser educated families only. Nevertheless, the historically destitute lower caste categories are most often the most disadvantageous people, especially if they are from the agricultural or relatively lower educated family backgrounds. This is the third contribution of our paper, that is to show how our circumstances are intertwined in generating unequal opportunity in the society.

The remaining of the paper is organized as follows. Section 2 sketch out the methodological framework of the non-parametric, parametric and the regression tree approach. Section 3 introduces our data and a clear clarification of all our variables, along with details on our sample selection criteria. Section 4 describes our results in different subsections. After discussing the main non-parametric and parametric measures of *IOP* in India, we give a brief account on the relative importance of caste and other social backgrounds, in comparison with the parental backgrounds. The opportunity structure for contemporary India is discussed next, separately for all of the different outcome variables. Section 5 concludes.

2 Theoretical and methodological background

In the analysis of inequality of opportunity, any social outcome is supposed to be generated by two broad classes of factors. Factors that are beyond individual responsibility or *circumstances (C)* and factors that are within individual control or *efforts* (e). Therefore, borrowing from Ferreira & Peragine (2015), the simplified outcome generating process can be written as -

$$y = f(C, e) \tag{1}$$

Such that the outcome to be analyzed, y, can be determined from a finite set of circumstances, C, and efforts, e. From the standpoint of responsibility sensitive egalitarianism any outcome inequality generated by C is ethically objectionable, whereas inequality arising from e can be considered legitimate[4].

Any analysis of *IOP* therefore begins with the clear classification of the circumstance and the effort variables. However there are no fixed list of circumstance or effort variables to be taken into account, as they are subject to data availability and are rather determined in the social or political space that varies between different societies (Roemer & Trannoy 2013). Nevertheless as common to any empirical exercise, estimates of *IOP* crucially depends on the data structure and partial observabiltiy of circumstance or effort factors severely limit the study. Data availability

[4]Lefranc et al. (2009) introduced a third factor, that of *luck*, in the study of *IOP*, which we did not consider in the present work.

on effort factors in particular, are even more limited for a large number of surveys. However *IOP* is the amount of inequality generated by circumstances only and *efforts*, the so called legitimate source of inequality, can itself be determined by the existing social circumstances. Hence effort variables themselves are often assumed to be a function of circumstances, so that the outcome generating process in *equation* (1) can actually be reformulated as a reduced form equation, $y = g(C)$, where outcome is a function of circumstances only (Ferreira & Gignoux 2011). Of course higher the number of circumstances taken into account, more realistic is the measure of *IOP*. But with addition of new circumstances into the analysis *IOP* will always increase as long as the added circumstances are not orthogonal to the outcome in concern. Since it is impossible for any survey to provide a complete exhaustive list of circumstances, Ferreira & Gignoux (2011) therefore advice to interpret any resulting estimates of *IOP* as the lower bound of the true *IOP* in the society.

Unlike the traditional inequality approach, social welfare in the responsibility sensitive domain is not judged on the basis of total inequality in the outcome variable, $I\{y\}$. Rather *IOP* is the measure of only that part of outcome inequality that is generated by the circumstance factors, C, exclusively. So the main methodological challenge for quantifying *IOP* is to quarantine this unfair part of outcome inequality. This is usually done in the literature by constructing suitable counterfactual distributions, y^{CF}, such that by construction, y^{CF} is able capture the variability in the outcome arising uniquely from the differences in the circumstance variables, C. The measure of *absolute IOP* in the society can then be measured by the inequality in the counterfactual distribution, $I\{y^{CF}\}$. However since *IOP* is estimated as that part of total inequality which is unfair and morally objectionable, it is a common practice in the literature to provide the estimates of *relative IOP* as the share of unfair inequality in the total outcome inequality by $I\{y^{CF}\}/I\{y\}$. The construction of the counterfactual distributions and hence the measurement of *IOP*, varies with the non-parametric or the parametric statistical model of analysis as discussed below.

2.1 Non-parametric approach

The non-parametric method for the present analysis have been adopted from the work of Ferreira & Gignoux (2011). Consider a finite population set, $i \in \{1, ..., N\}$, characterized by $\{y_i, C_i\}$, standing for outcome and circumstances respectively. Assume that the vector C_i consists of J elements and each of the element can take x_j number of values or categories. Usually groups formed by all possible interactions of the circumstances are called *types*. In this framework, the population under study can thus be partitioned into a maximum of $\bar{K} = \prod_{j=1}^{J} x_j$, exhaustive and mutually exclusive *types*.

From the viewpoint of *IOP* any inequality *between types* is 'unfair'. To isolate this unfair inequality each of the k types are represented by a 'smoothed distribution' of their respective mean outcomes. Thus every individual in a *type*, $i \in \{k, k = 1, \ldots, \bar{K}\}$, are assumed to be characterized by the *type-mean* outcome, μ^k, for each $k = 1, \ldots, \bar{K}$. Therefore the counterfactual distribution to quarantine the inequality generated exclusively from the differences in *types*, is represented by, $y^{CF} = \{\mu_1, \ldots, \mu_{\bar{K}}\}$. The absolute and relative measure of *IOP* can then be

estimated as[5] -

$$\theta_a^{NP} = I(\{\mu^k\}) \tag{2}$$

$$\theta_r^{NP} = \frac{I(\{\mu^k\})}{I(\{y_i\})} \tag{3}$$

Where $I(\{x\})$ denotes inequality in the distribution of x. Following the extant literature, $I(\cdot)$ is measured by the index of Mean Log Deviation (MLD)[6].

2.2 Parametric approach

The parametric approach in the present work has been adopted from Ferreira & Gignoux (2011) as well, which also essentially estimates *IOP* by the mean outcome conditional on *types* by the OLS estimates, but differs from the non-parametric set up in its construction of the counterfactual distribution to isolate the ethically unfair part of inequality.

The parametric set up usually assumes a log-linear relationship between the outcome and the circumstance/effort variables. So the income generating process can be written as -

$$\ln y_i = \alpha C_i + \beta e_i + u_i \tag{4}$$

However, as mentioned before, the effort factors can fairly be assumed as a function of circumstances as below -

$$e_i = \gamma C_i + v_i \tag{5}$$

with u_i and v_i being the random errors.

Hence, from the structural model (4) and (5), the *reduced form* income generating process can be summarized as -

$$\begin{aligned} \ln y_i &= \alpha C_i + \beta(\gamma C_i + v_i) + u_i \\ &= (\alpha + \beta\gamma)C_i + (\beta v_i + u_i) \\ &= \Psi C_i + \varepsilon_i \end{aligned} \tag{6}$$

From the OLS estimates of *equation* (6), $\hat{\Psi}$,*IOP* is then measured in comparison to a hypothesized distribution, $\{\tilde{y_i}\}$, that eliminates any differences in individual circumstances, as -

$$\tilde{y_i} = \exp[\bar{C_i}\hat{\Psi} + \hat{\varepsilon_i}] \tag{7}$$

where, $\bar{C_i}$ is the mean of circumstance variables across the population. Thus *equation* (7) eliminates the differences in circumstances by replacing them with their mean values and the associated inequality, $I(\{\tilde{y_i}\})$, is therefore segregated as fair, by construction. The measure of absolute *IOP* can then eventually be estimated from the counterfactual distribution, $y^{CF} = (\{y_i\} - \{\tilde{y_i}\})$, that isolates the outcome variations generated from the differences in individual

[5] NP stands for Non-parametric, r for relative measure and a for absolute measure

[6] MLD(x)=$\frac{1}{N}\sum_1^N \ln \frac{\bar{x}}{x}$

circumstances only. So the relative share of *IOP* in the total inequality is given by[7] -

$$\theta_r^P = \frac{I(\{y_i\}) - I(\{\tilde{y_i}\})}{I(\{y_i\})} \tag{8}$$

Similar to the non-parametric approach, we use the same index of *MLD* for the parametric estimates of *IOP* as well.

2.3 Regression tree approach

Circumstances by definition are all possible factors that are beyond individual responsibility and it is physically impossible for any data set to capture all such factors under a single or multiple survey. Research on *IOP* is therefore always restricted to a subset of the total set of circumstances. But as long as the omitted circumstances have non-trivial effect in predicting the outcome variable, addition of each such circumstance will increase the estimate of *IOP* by virtue of finer partitioning of the population. Clearly higher the circumstances taken into account, more realistic is the estimate of *IOP*. However addition of new circumstances also comes at a cost. Not only that this finer sample partitioning leaves fewer observations for each *type*, but some *types* may have very unusual observations due to this unrestricted partitioning. This may contaminate the associated measure of *IOP*. The regression tree analysis coined in the literature by Brunori et al. (2018), makes an attempt to allay this issue in the fashion of machine learning.

Once again assume that for individual i, the circumstance vector, C_i consists of J elements, $C_i \in \{C_i^1, \ldots, C_i^J\}$, each of which can take x_j number of values, where $j \in \{1, \ldots, J\}$. Unrestricted partitioning will then divide the population into, $\bar{K} = \prod_{j=1}^{J} x_j$, number of *types*, considering all possible interactions among the circumstances. However for a large number of C_i and/or x_j variables, observations in all or some of the cells in $\bar{K}$ may get too crunched to allow the researcher to use all available *types*, especially when sample size is relatively less. Besides in case of unrealistic vacuous interactions, some cells may suffer from no observations at all. Since there is no way to point out the relevant interactions in either of the non-parametric or the parametric modeling, the conventional resort is either to regroup the circumstances in broader categories (less x_j) or sacrificing some of the circumstances (less C_i) or both. In the regression tree approach instead, the researcher submits the full set of available circumstances, C_i, to the program and let the algorithm choose the relevant partitioning of the sample under study in a non-arbitrary way, by *recursive binary splitting* to be precise.

The recursive binary splitting is a type of permutation test, because it rearranges the labels on the observed data set multiple times and computes test statistic (and p-value) for each of this rearrangement. It starts by dividing the full sample into two distinct groups based on one circumstance factor and then continue the same for each split, potentially based on another circumstance, into more subgroups and so on. The criteria for the selection of splitting circumstances depends on the type of regression tree used. Brunori et al. (2018) uses the conditional inference tree algorithm to determine the splitting criteria as follows.

The algorithm runs in two stages as -

[7] *Where* P and r in the superscript and subscript stand for parametric and relative measure respectively.

- **Stage - I**: Selecting the initial splitting circumstance
  - It starts with the simultaneous testing of the J partial hypothesis, $H_0^{C^j} : D(Y|C^j) = D(Y)$ for $j \in \{1, \ldots, J\}$. Notice, this precisely is the testing of the existence of *IOP*, to see if any circumstances have any effect on the outcome.
  - Adjusted p-values, $p_{adj}^{C^j}$, are then computed with the standard adjustment for multiple hypothesis testing[8] and identifies the circumstance, C^*, with the highest degree of association, that is, the circumstance with the minimum p-value, $C^* = \{C^j : \text{argmin} \quad p_{adj}^{C^j}\}$[9].
  - The algorithm stops if the p-value associated to C^* is greater than some pre-specified significance level, α[10]. Hence, if $p_{adj}^{C^*} > \alpha$, the null of equality of opportunity for the society, can not be rejected at $\alpha\%$ level of significance. Otherwise, the circumstance, C^*, is selected as the initial splitting variable.
- **Stage - II**: Growing the opportunity tree
  - Once C^* is selected, it is split by the binary split criterion to grow the tree. For each possible binary partition, s, involving C^*, the entire sample can be split into two distinct parts as, $Y_s = \{Y_i : C_i^* < x_j\}$ and $Y_{-s} = \{Y_i : C_i^* \geqslant x_j\}$.
  - For each binary split, s, the goodness of split is tested by testing the discrepancy between Y_s and Y_{-s}[11]. The split, s^*, with the maximum discrepancy, that is with the minimum p-value, is then selected as the optimum binary split point, based on which the sample is now partitioned into two sub-samples, constructing the initial two branch of the opportunity tree.
  - The entire algorithm is then repeated for each branch separately, to construct the full opportunity tree.

3 Data, variables and sample selection

3.1 Data

For the present analysis of inequality of opportunity in India we have taken data from the National Sample Survey (*NSS*). This is the biggest nationally representative micro level database for India, collected by the National Sample Survey Organization (*NSSO*), India. Among the many national level surveys conducted by *NSSO* we have taken the *Employment Unemployment Survey* in particular. This survey is conducted for a year in every five years, covering the whole country except some remote inaccessible area[12]. For focusing on the recent scenario in India,

[8] The adjustment is the Bonferroni correction, $p_{adj}^{C^j} = 1 - (1 - p^{C^j})$ (Brunori et al. 2018, p. 8).

[9] To test the association between the outcome variable and the covariates, the linear statistics form, along with its mean and variance, is provided in Hothorn, Hornik & Zeileis (2006), where from the relevant test statistic and p-value can be formulated.

[10] Like Brunori et al. (2018) we also choose $\alpha = 0.01$

[11] This is tested by the two sample test statistics, provided in Hothorn et al. (2006). The entire algorithm can be executed by an R package, developed by the same authors.

[12] So conflict areas of Ladakh & Kargil districts of Jammu & Kashmir, some remote interior villages of Nagaland, few unreachable areas of Andaman & Nicobar Islands and those villages recorded as uninhabited by respective population census, are kept out of these surveys.

we have taken the latest two employment-unemployment survey rounds of *NSS*, covering years *2004-05* and *2011-12*[13].

These rounds on average, survey 110000 households enumerating about 0.4 to 0.6 million individuals. India is predominantly rural even to date with a rural-urban ratio around 70 : 30 on average. Initially we have to drop about 1000 observations per round to clean for valid age, sex, sector, caste specification, marital status and some other criterion, depending on different rounds. *NSS* provides details on several household and individual characteristics. Some of the major household provisions include household size, religion, caste and consumption expenditure, whereas age, sex, education, occupation and many other demographic characteristics are recorded for each member of the household. However not everybody reported as 'employed' do have information on their income, rather wage earning is selectively reported in the *NSS* data only for the regular and the casual wage earners who are not self-employed. Another possible drawback in the structure of *NSS* data base is that it does not report information on parental background directly for every individual. Rather this crucial information is only available for households where the offspring is enumerated along with his/her parents.

3.2 Definition of variables

3.2.1 Circumstance variables

For the present analysis, we have chosen a set of five circumstance factors, that of caste, sex, region of residence, parental education and father's occupation. We can label the first three of them as social backgrounds, while parental education and occupation constitute parental background. With all possible interactions of these five circumstance variables, we have a total of 324 *types*.

Caste system in India is a century old hierarchical social structure based on occupation. However the historical occupational perspective in its way became hereditary over time and children always inherit the caste of their father that is unchangeable for lifetime. There are thousands of castes in the country, which are regrouped in fewer caste categories by the constitution of India for the purpose of caste based affirmative policy or reservations. We consider three caste categories in our analysis. The lower caste category consists of the *Scheduled Castes* and the *Scheduled Tribes* caste categories together (*SC/ST*). They are the most historically disadvantageous caste groups in India and are designated the reservation status since 1950. Around mid-eighties, the socially and economically backward castes among the non-*SC/ST*s are further categorized as the *Other Backward Classes* (*OBC*) who are entitled to certain reservation quotas in higher education and Government jobs since the beginning of nineties. Indian nationals do not belong to any of the above mentioned caste categories are formally called as the *General* category individuals and are excluded from any caste based affirmative policies by rule. *OBC*s can be thought of as the middle level caste category who are usually little more advantageous than the historically disadvantageous caste categories of *SC/ST*, but have lesser economic advantage as compared to the forward *General* caste category.

[13]This means we have taken *Schedule 10* survey of *NSS*, for rounds 61 (2004-05) and 68 (2011-12). Details of these database are in the NSSO data appendix Kundu (2019).

Considering the bulk of literature on gender discrimination in India, we take two categories of sex, male and female, as our next social circumstance. To consider region as one of our circumstances, we have to take region of residence, although the ideal circumstance factor would be the birth region. Due to unavailability of information on birth place, we have to consider the present residing region as a proxy for birth region, which is not a far fetched assumption given the low rate of inter-state and inter-district migration in India as per the recent migration survey report of NSSO (2008). To further minimize migration related contamination, we take six broad regional categories for our analysis as - North, East, Central, North-East, South and West[14].

Our next batch of circumstances consists of parental background that includes two kind of parental attributes, that of parental education and occupation. By combining father's and mother's education, we take three categories of parental education as - (i) both parents have no formal schooling (ii) at least one parent has primary or below primary schooling (that means the other parent, either have the same level of schooling or less) and (iii) at least one parent has above primary schooling. It is worth a mention, that 'no formal schooling' is not equivalent to illiterate parents, as they may have exposed to other informal adult literacy programs, but have never experienced formal schooling. Due to considerably low information on mother's occupation, we took three categories of father's occupation as a proxy of parental occupation. The occupational categories are taken as father's employment in - (i) white collar job (ii) blue collar job and (iii) agricultural occupation. White collar job category includes all sorts of professional, executive and managerial jobs. Whereas, sales and service workers falls in the domain of blue collar workers. Agricultural job includes horticulture, fishing and hunter-gatherers as well.

3.2.2 Outcome variables

The analysis of *IOP* on India is executed for three different outcome variables - consumption, wage and education. All the three outcome variables are considered as continuous variables. While first two of them is measured in monetary units (Indian Rupee, INR), education is measured as the years of school/college education.

Consumption is considered as the monthly per capita consumption expenditure (*MPCE*). This is the total monthly expenditure on certain durable and non-durable goods incurred by the household over the last thirty days prior to the date of the survey. This data is therefore reported at the household level, which we divide by the respective household size to get the individual level values. The list of goods, expenditure on which is to be reported is a selection of goods that has been considered as the most important ones by the respective survey. Borrowing from Hnatkovska, Lahiri & Paul (2012), we use the real *MPCE* as our outcome variable, upon dividing *MPCE* by the state level absolute poverty lines[15].

[14]Statewise composition: Jammu & Kashmir, Himachal Pradesh, Punjab, Haryana and Uttarakhand - *North*; Bihar, Jharkhand, Orissa, West Bengal - *East*; Uttar Pradesh, Rajasthan, Madhya Pradesh, Chattisgarh - *Central*; Sikkim, Arunachal Pradesh, Assam, Nagaland, Meghalaya, Manipur, Mizoram, Tripura - *North-East*; Karnataka, Andhra Pradesh, Tamilnadu, Pondichery, Kerala, Lakshadeep - *South* and Gujrat, Daman & Diu, Dadra & Nagar Haveli, Maharashtra, Goa - *West*.

[15]We use poverty lines, that can account for the differences in standard of living across the states of India. Besides, the measure of absolute poverty line is provided by the Planning Commission of India using data collected by the same survey, that of the National Sample Survey, the one we use for the present analysis. Another commonly used deflator is the consumer price index, which we did not use, as it was measured on the basis of a different survey

Our next outcome variable is the wage earning, which is reported only for the regular and casual wage earners. Therefore the wage data is not available for a large chunk of self-employed individuals who constitute nearly 40% of the working adults in India. Unlike *MPCE*, wage is reported as the weekly wage received or receivable for multiple activities, by each regular/casual earning members of the household over the last week prior to survey. The main reason for reporting wage as an weekly input is that many of the Government or non-Government public work programs in India are transitory in nature that employ a huge number of rural casual laborers. However we consider wage corresponding to the major activity that had been pursued for the maximum number of days over the reference week. In case of equal number of days spent on more than one activity, we prioritize those having valid wage entry and occupation information. In particular we consider the daily real wage earning as our outcome variable by dividing total weekly wage by the number of days engaged in that major activity. Similar to *MPCE*, the corresponding real wages are generated upon division by the state level absolute poverty lines.

Our third outcome variable is education. However our data base provides information on education in different categorical codes, that is recorded for each individual as their highest educational achievement at the time of the survey. We converted the education codes in suitable number of school/college years, based on the standardized norm in the country and use years of education as our outcome variable in concern. We assign 1 year of education to the lowest category of 'without formal schooling'. As 'without formal schooling' incorporates literacy through other informal medium, we reserve one year of education for this category as a cognitive margin of low education. Further, 2 to 4 years of education is assigned to educational categories corresponding to primary or below-primary level of schooling. For education above the formal primary schooling 8 to 16 years of education are assigned, that covers a broader range of reported educational codes from below secondary level to graduate level college education. Since our data has the further provision of some additional technical education (like certain under-graduate or graduate levels diploma/certificate course), we update the years of education accordingly, for those who have reported to have some technical education[16].

3.3 Sample selection

As mentioned before, *NSS* does not provide information on parental attributes for every individual, making this data limited to the co-resident households that consists of both offspring and parents as the respondents. Provided the instrumental role of parental backgrounds in the analysis of unequal opportunities for a number of countries, the study on India will remain incomplete had we not consider that. Therefore given the data structure, the biggest challenge in the sample selection procedure is how to best incorporate the valuable information of parental backgrounds in our analysis of *IOP* in India.

and prior to 2011, the combined rural and urban price indices are not provided (instead, consumer price index used to comprise of multiple series like, urban non-manual labor, agricultural labor, rural labor and industrial workers).

[16]We draw upon the work of Hnatkovska et al. (2012) while updating the year of education for technical education. The mapping of education codes to years of education is provided in the NSSO data appendix (Kundu 2019).

Studies for which parental information may be important, like the analysis of inter-generational mobility or inequality of opportunity, when using the *NSS* data base, usually deal with this issue either by restricting their analysis to the co-resident households (*e.g.* Hnatkovska, Lahiri & Paul (2013) for inter-generational mobility analysis) or by sacrificing the parental background data (*e.g.* Asadullah & Yalonetzky (2012) for educational opportunity analysis). As mentioned before we already ruled out the second option considering the importance of parental attributes in *IOP*. However to analyze *IOP* we want our sample to be restricted to working adults who have reportedly finished their education. Provided that, the other option to include parental attributes is to limit our analysis to households with adult inter-generational co-residence. Although adult parent-child co-residence is not an uncommon social pattern in India, it may raise the issue of selectivity bias. So to provide estimates of *IOP* in India with a nationally representative sample, we impute the parental attributes for our sample using the technique of *multiple imputation*. Our sample therefore consists of working adults who are aged between 18 to 45 years, are not currently enrolled in any educational institution, are from male-headed households (who also are the only head of the household) and have valid information on education and occupation, both for themselves and for their parents[17]. However for estimating *IOP* in wage, we further restrict our sample to those who additionally provide valid data on wage.

The theory of multiple imputation was introduced by Rubin (1976, 1986) for dealing with the problem of missing data due to non-response in large survey data sets. Although mostly popular in the statistical and medical research, the use of multiple imputation to handle missing values is expanding in economics as well, especially in the survey data based econometric analysis[18]. In particular, Teyssier (2017) showed the efficacy of multiple imputation for imputing parental information for a data set on Brazil, for which this information is also available without the co-residence issue. We want to impute two parental attributes in particular, that of parental education and father's occupation, both of which are considered as categorical variables in our estimation of *IOP*.

We first form our sample as per the sample selection criteria mentioned above, except the criteria related to parental attributes. We can now think of this sample as the union of two exhaustive and mutually exclusive parts - the 'response' and the 'non-response' part. While the 'response' part have valid information on parental background, this crucial information is missing for the other part. The exercise of multiple imputation is to use information from the 'response' part to *impute* values for the 'non-response' part, using all possible auxiliary information provided by the data set that are non-missing for both of the 'response' and the 'non-response' part. In our case the 'response' part consists of the co-resident data points for which parental background is observed[19]. Table 8 in Appendix A reports the summary statistics

[17]We exclude multi-headed and female headed households in India, as they are rare and subject to special constraints. Over 90% of heads are male and 99% households are single-headed-household.

[18]For application of multiple imputation technique in poverty and inequality analysis, *see* Alon (2009), Jong-Sung & Khagram (2005), for example. Whereas, Salehi-Isfahani, Hassine & Assaad (2014), Teyssier (2017), provide estimates of IOP using multiply imputed circumstances.

[19]In particular we consider our 'response' part to constitute of samples who are living with their parents, with father as the household head. However a co-resident household may consist of other members with information on parents as well. Two cases in particular are excluded. First we did not take grandchildren of the household head for simplicity. Secondly, households where the adult working child share the headship and is living with one of his/her parents should also be taken into account, but could not be, because in this case *NSSO* reports father/mother/father-in-law/mother-in-law by a single code, making it impossible to extract information on bio-

of the 'response' and the 'non-response' sub-samples. It shows that co-residence does not seem to make a marked difference in terms of caste, occupation and rural-urban composition. But notice that the samples of the 'response' part, as expected, are relatively younger. Hence is the justification of taking relatively younger adults (18-45 years) for our analysis, so as to keep parity between the 'response' and the 'non-response' part.

The two parental variables in concern, that of the parental education and father's occupation are then estimated for the 'response' part by a suitable imputation model (an ordered logistic regression, in our case), using a broad range of predictors including households, individuals and some survey related characteristic variables that are strictly non-missing for both the 'response' and the 'non-response' part[20]. Parental attributes for the 'non-response' part is then imputed from simulated draws of the posterior distribution of these estimates. However as the name suggests, the imputation of the missing values is done for a multiple number of times generating multiple number of 'completed' data-sets, where none of the attributes are missing any longer. We adopt the sequential regression multiple imputation algorithm of Raghunathan, Lepkowski, Van Hoewyk & Solenberger (2001) and use 20 imputations in particular. Both the non-parametric and parametric measures of IOP are then analyzed separately over each of the 'completed' data-set and combined by Rubin's rule (Rubin 1986) to give the final measures of *IOP*.

However the exercise of multiple imputation does *not* mean to 'create' the missing values in a deterministic fashion, but rather to capture the additional features of the 'response' part to use it in the final analysis. Therefore two of the important criteria for a successful imputation are, that the imputation model should provide good estimates of the missing parental attributes from a bunch of non-missing variables and that the relation between them remain the same for the 'non-response' part as well. While the former can be tested by the imputation model diagnostics, given the data-set the latter can at best be reasonably assumed (Marchenko & Eddings 2011). In particular the second criteria of a good imputation requires that the probability of the missing data does not depend on any unobservable factor and hence can be imputed successfully from the imputation model (Allison 2000). Our imputation exercise and eventually the measures of *IOP* also bank on this assumption, which is the so called assumption of 'missing at random' (*MAR*)[21]. Summary statistics of our *sample*, as well as our sub-sample for the wage analysis (wage sample), is given in Table 1.

logical parents. However these two cases together do not exclude more than 10% of the sample, as far as adults are considered.

[20]This includes some household characteristics like household size, caste, sector (rural/urban), religion, consumption expenditure and offspring's' characteristics like their age, relation to head, marital status, region of residence, sex, occupation, education, along with some other survey-specific attributes. Further details of our imputation model are provided in Appendix A.

[21]Since we can never actually test whether the missing-ness depend on some unobservable factor not provided by the data-set, we have to assume MAR. However, since adult inter-generational co-residence is the rather prevalent social pattern for most part of India, it is quite reasonable to assume that parental attributes does not depend on some hidden unobservable factors beyond the provision of the survey. Another assumption that of 'missing completely at random' (MCAR) is also mentioned in the literature, which assumes that the probability of missing-ness is random. This is rarely the case for any survey data and so for NSS, because co-residence is clearly more probable for younger males and less for females (for female migration due to marriage). However a number of literature suggests that the assumption of MAR is good enough for a reasonable imputation (Rubin 1976, Little 1988, Allison 2000, Raghunathan et al. 2001). Appendix A provides further details of our imputation algorithm and diagnostics.

	age	hhsize	%rural	%married	%noschool	%agri	%wage	N
Working sample								
2004-05	32.11	5.5	0.76	.82	.36	.53	.41	127002
[61]	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)		
2011-12	32.74	5.0	0.72	.82	.24	.45	.48	90574
[68]	(0.05)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)		
Wage sample								
2004-05	31.80	5.0	.69	.80	.37	.42	1.0	48127
[61]	(0.05)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)		
2011-12	32.24	4.7	.65	.79	.24	.33	1.0	41619
[68]	(0.07)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)		

Table 1: Work sample summary statistics [a]

[a]standard errors are in parentheses and round numbers are in squared brackets. 'age' and 'hhsize' reports the mean age and household size of our sample. %rural, %married, %noschool, %agri and %wage reports the share of rural sample, married individuals, samples without any formal schooling, samples engaged in agricultural jobs and samples who further have the information on wage data, respectively. The last column (N) reports the respective sample size.

Table 1 reports the mean age, household size (hhsize), share of rural sector, share of married samples, share of individuals without any formal schooling (noschool) and share of population engaged in agriculture (agri) in our working sample along with the respective sample size. First of all, similar to the general picture of the whole country, our sample is predominantly rural with a substantial population in agricultural occupations. However even in *2012*, nearly one-fourth of our sample have no experience of formal schooling ever. The last but one column reports the percentage share of our working sample to have information on wage data. It shows that more than half of our working sample do not have information on wage data, which explains the massive reduction of sample size for our wage sample. The lower panel of Table 1 shows that the regular and casual wage earners are usually less rural and less agricultural.

circumstances →	Caste	Sex	Region	Parental education	Father's occupation
share of →	SC/ST	male	north	no schooling	agriculture
Working sample					
2004-05	29.1%	77.2%	6.6%	55.0%	62.8%
2011-12	29.7%	82.1%	6.9%	46.9%	54.3%
Wage sample					
2004-05	32.0%	80.0%	7.9%	56.5%	56.7%
2011-12	34.1%	83.9%	7.8%	46.4%	48.1%

Table 2: Circumstance specific summary statistics[a]

[a]Each column shows the percentage share of our samples who are - SC/ST, males, residents of Northern region, have both parents without any formal schooling and have agricultural fathers, respectively.

Table 2 gives the circumstance specific composition for each of our five circumstance variables (caste, sex, region, parental education, father's occupation). The samples across the rounds looks comparable with identical proportion of circumstances, especially in terms of social background circumstances (caste, sex, region). Due to low female labor force participation, both of our working and wage sample are rather male dominated with even higher proportion of males in

the wage sample[22]. Besides as per with the national population, Northern India has relatively less number of samples[23]. Although our wage sample has relatively more lower caste individuals, caste composition for either of our sample is close to the national proportion. For both of the survey rounds, nearly 30% of our sample are from the destitute caste groups of *SC/ST* which is similar to the caste proportions in the country as a whole. About 46-56% of both of our working and wage sample have neither parents with any formal schooling experience. Besides most of the samples are from agricultural households where fathers are engaged in agro-based occupations. However similar to their own education and occupation as provided in Table 1, parental education seems better for the latest round as well, along with a lesser share of agricultural fathers.

4 Results and discussion

4.1 Measures of IOP in India

To quantify the degree of unequal opportunity in Indian society for consumption, wage and education, we adopt both the non-parametric and the parametric approaches for at least two good reasons. First, it will serve as a robustness check to our measures of *IOP*. With the same set of circumstances, the amount of unfair inequality should not have much variation under the non-parametric and the parametric set up. Second, most of the international measures of *IOP* have used either or both of these methods. Estimating *IOP* for India under both the approaches will therefore be helpful for international comparisons. Following the extant literature, both inequality and *IOP* are always measured by the index of mean log deviation. Besides both the non-parametric and the parametric measures of *IOP* are based on all possible interaction of the full set of circumstances, *viz.* caste, sex, region, parental education and occupation, leaving us a total of 324 *types* to compare[24].

Table 3 reports the *relative IOP* as well as the measure of total inequality, for MPCE (consumption), wage and education. The first row reports the amount of total inequality in each of the outcome variables separately. Inequality is highest for education that lies between 0.39 to 0.46 over the time frame of *2004-12*. Whereas over the same time span inequality in consumption and wage is close by and hovers around 0.24 on average. Notice that other than education, inequality in all other outcome variables are actually increasing over the eight years time span considered here. Particularly consumption inequality for our sample shows a rather sharp increase for the latest survey year which is at par with the recent trend in Indian economy. A number of literature documents the increasing consumption and earning inequality in India since the country switched from a centrally interventionist policy to a rather neo-liberal open-market policy regime around early nineties[25].

[22]In the chosen age group (18-45 yrs.), about 30% females are currently employed, while, on average, 65% are reported as not in labor force for attending domestic duties.

[23]Notice that many parts of the Northern and North-Eastern India are more likely to be out of the *NSS* sample coverage for having relatively more conflict zones and remote areas.

[24]The 324 *types* correspond to the interaction of - caste(3)×sex(2)×region(6)×parental education(3)×father's occupation(3), where number of categories for each circumstances are in parentheses.

[25]*See* for example, Deaton & Dreze (2002), Himanshu (2018).

	MPCE		Wage		Education	
	2004-05	2011-12	2004-05	2011-12	2004-05	2011-12
Inequality	0.19681	0.28527	0.22519	0.25101	0.46248	0.39063
Measures of relative IOP						
Non-parametric	0.11051	0.11172	0.32896	0.39310	0.31968	0.26707
Parametric	0.10378	0.10661	0.31461	0.37747	0.30591	0.24803

Table 3: Measures of Inequality of opportunity in India[a]

[a]All IOP measures are the relative measures of IOP and therefore reports the percentage share of IOP in the total inequality upon multiplied by 100. So the non-parametric estimation of IOP in education for 2011-12 tells that 26.7% of educational inequality is due to unequal circumstances in that survey year.

The last two rows of Table 3 reports the non-parametric and the parametric measures of relative *IOP* respectively, using all possible interaction of the chosen circumstances. So the non-parametric *IOP* for education says that 26.7% of the high educational inequality is due to differences in the chosen set of circumstances during the survey year of 2011-12 and therefore strictly unfair from an ethical perspective. Similar to Ferreira & Gignoux (2011), we found the non-parametric measures for each outcome to be always little higher than the corresponding parametric measures of *IOP*. However for all the respective outcome variables, the measures of *IOP* are close-by under both of the statistical set-ups (non-parametric and parametric), indicating that our results are actually robust to the method adopted.

Among the three outcome variables considered, Table 3 shows that the share of ethically unfair inequality is relatively low for *MPCE*. About 11% of consumption inequality is due to unequal opportunities arising from the differences in the chosen circumstances. The degree of consumption *IOP* in India is still a bit higher than most of the developed countries and in fact positions India closer to the Sub-Saharan African countries (Cogneau & Mesplè-Somps 2008). The same can not be said for wage and education though. Over the time span of *2004-12*, about 33-39% of wage inequality in India is conditioned by unequal social and parental backgrounds. At least in terms of wage *IOP* with a comparable set of circumstances, India seems worse than Brazil that has found to be as one of the most opportunity unequal country in Latin America (Ferreira & Gignoux 2011). Education on the other hand, in spite of having a much higher level of inequality than wage, shows a comparable degree of *IOP* that on average accounts for about 29% of total inequality. However even by *2012*, more than one-fourth of educational inequality and more than one-third of earning inequality in India is due to unequal opportunities, arising from differences in circumstances that are beyond any subjective control.

Although Consumption and wage are often analyzed side by side in many of the development studies as two comparable source of standard of living, this is not the case for the present analysis. This is because *NSS* data does not report these two variables in a comparable format and we can point out at least three major sources of variation in the reporting of the consumption and the wage data in our data base. First of all, *MPCE* is a household level data reported as the total expenditure of the household and is therefore unable to capture any intra-household differences. Wage on the other hand is likely to be rather varying in nature, as it is reported not only for every regular/casual earning members of the household but also for multiple number of activities. Second, *MPCE* is recorded for a larger recall period of a month. Whereas due to

the transitory nature of many casual wage earning jobs, wage is reported for the reference week prior to the date of the survey. Together a shorter recall period along with a finer reporting unit makes the wage data to be more variant and responsive to changes in the individual circumstance factors. Finally, wage and consumption are estimated for different samples and the same set of circumstances may have a differentiated effect for different samples. In particular a large body of self-employed individuals are excluded exclusively from the wage analysis.

4.2 Effect of caste in comparison with parental background

India is one of the very few countries where the century old caste system is well embedded even to date. The origin of the caste system was found in the ancient Hindu text, where the society was divided in hierarchical occupational structure. Upper castes are supposed to be engaged in occupations that are more pure in nature like worshiping deities or serving the country as soldiers. Whereas the major occupation of the lower caste categories is to serve the upper caste 'masters'. Caste in its way became hereditary and is identified at birth that is not convertible for lifetime. Although that makes caste a classic circumstance factor in the context of *IOP*, it is certainly not the only source of hierarchy in the Indian society and may have its effect through many channels. The purpose of the present section is not to explore these different channels, rather to show the relative importance of caste as a circumstance factor as compared to parental background and other social backgrounds, in the context of estimating *IOP* for India.

Table 4 reports the non-parametric relative measures of *IOP* with different set of circumstances. The first row gives the non-parametric *IOP* with the full set of circumstances and is therefore the same as the non-parametric measures in Table 3. From the second row onward we provide the associated estimates of *IOP* after omitting one or more of the circumstances from our analysis. Measures corresponding to the second row reports the index of non-parametric relative *IOP* after caste is omitted from our set of circumstances. Similarly the third row estimates *IOP* without taking any parental attributes (parental education and father's occupation) as our circumstances and the last row reports the same when all circumstances other than caste are omitted from the analysis. However unless the omitted circumstances are completely orthogonal to the outcome in concern, *IOP* will always increase with addition of new circumstances. It is the reason why Ferreira & Gignoux (2011) suggested to interpret the resulting *IOP* estimates as a lower bound of the true *IOP* in the society because no study can ever take into account the complete exhaustive set of circumstances. Therefore as expected, *IOP* mostly decreases as we move down in Table 4 from more to lesser number of circumstances.

	Measures of relative *IOP* (non-parametric)					
	MPCE		Wage		Education	
Taken circumstances	2004-05	2011-12	2004-05	2011-12	2004-05	2011-12
caste+sex+region+parental backgrounds	0.111	0.112	0.329	0.393	0.320	0.267
sex+region+parental backgrounds	0.069	0.099	0.313	0.363	0.310	0.248
caste+sex+region	0.086	0.047	0.131	0.161	0.123	0.102
caste	0.049	0.014	0.030	0.079	0.029	0.045

Table 4: Effect of omitted circumstances in the measure of IOP[a]

[a]'Parental background' is abbreviated to indicate circumstances related to parents and therefore includes parental education and father's occupation.

Notice that as compared to the first row with full set of circumstances, *IOP* decreases both for the second and the third row of Table 4, but it is the latter for which the fall in the value of *IOP* is larger for most cases. Even after omitting caste, earning and educational *IOP* in India are still mostly over 30% and consumption *IOP* for the latest round (2011-12) also decreases marginally. On the other hand after omitting parental backgrounds from the analysis, only about 10-16% of the total inequality is deemed unfair for the presence of *IOP* in wage and education. So *IOP* more than doubled for most of the outcomes when parental background is considered as additional circumstances along with the social backgrounds (caste, sex, region), whereas it decreases marginally when only caste is omitted from the analysis. This implies that the omitted effect of caste can be captured to a large extent by the other social and parental attributes considered. The only exception is the outcome of *MPCE* for 2004-05, where the omitted caste effect is higher than that of parental backgrounds. But even after controlling for caste, sex and region, differences in parental background have non-trivial additional effect in generating unequal opportunities for all the outcome variables. Hence is the necessity of multiple imputation of information on parental backgrounds, as the social attributes alone are not sufficient to take into account the discriminatory effect of parental backgrounds.

With caste as the only circumstance variable, *IOP* in India is no more than 8% for any outcome which is even less than some of the developed countries. However a comparison in this regard is not really appropriate as most of the international studies on quantifying *IOP* involves at least one circumstance regarding parental information. Nevertheless the low estimates of *IOP* for the last row of Table 4 does not indicate that caste has no role to play in generating unequal opportunities in the Indian society, rather it is indicative of the fact that caste alone can not capture well the differences in other circumstances especially that of parental backgrounds. The most historically disadvantageous caste category of *SC/ST*s are indeed found to be clearly dominated by the relatively advantageous upper caste categories even in the twenty-first century and the caste premium enjoyed by the forward caste group is actually increasing over time as far as earning opportunity is concerned (Kundu 2019, chapter-2). So the practice of casteism surely adds an extra deep rooted level of hierarchy even in the social fabric of twenty-first century India, but taking caste as the only responsible factor for quantifying *IOP* may be too coarse to account for the underlying unequal opportunity in the society.

4.3 Opportunity tree for contemporary India

Either of the non-parametric or the parametric approach uses a fixed model specification for analyzing *IOP*, where all the circumstances are given equal importance while estimating the resulting measures of *IOP* in India. However it is possible that caste may matter more in some part of the country with certain family backgrounds or educational opportunity is always less with lesser educated parents but even more when father is an agricultural worker. Neither of the non-parametric or the parametric measures have an answer to this question in the context of *IOP*. So to investigate the intertwining of our circumstances we adopt the regression tree approach that has been recently introduced in the literature by Brunori et al. (2018).

Because of our data structure we have to impute the information on parental backgrounds throughout our analysis. Although we computed the non-parametric and parametric estimates

on multiply imputed data set for more precision, it is difficult to perform the same for the regression tree analysis as far as the drawing of opportunity tree is concerned. Since each imputed data set may generate slightly different opportunity trees depending on the imputed values of parental education and occupation, the interpretation of the multiple opportunity trees for a single outcome variable becomes rather complicated. We therefore pick a randomly chosen imputed data set and draw the opportunity tree for that single imputed data-set for the survey year of *2011-12*, separately for each of our outcome variables.

All the opportunity trees are drawn on the basis of the same set of circumstances as they are considered for the non-parametric and parametric analysis. So the opportunity tree for all outcome variables are therefore drawn on the basis of - (i) three categories of caste - General [Gen], Other Backward Classes [OBC] and Scheduled Castes/Scheduled Tribes [SCST] (ii) two categories of sex - male [M] and female [F] (iii) six categories of region - North [N], East [E], Central [C], North-East [NE], South [S], West [W] (iv) three categories of parental education - none of the parents have any formal schooling [No], at least one have below primary schooling (considered as medium education) [Med] and at least one of them have above primary schooling (considered as high education) [High] (v) three categories of father's occupation - white collar [WC], blue collar [BC] and agriculture [Agr], where abbreviations in the square brackets are used to label the corresponding categories in the opportunity trees (Figures 1, 2, 3).

We submit this full set of circumstances to the program and let the algorithm choose the most relevant ones to draw out the opportunity tree, where the initial node represents the most important circumstance for the respective outcome. Unlike the non-parametric and parametric approaches, *types* in the regression tree are not all possible combination of the circumstances, rather each terminal node of the tree now correspond to a different *type* and is represented by the mean outcome of that *type*. *IOP* is then measured as the inequality between these *type*-mean outcomes. The major difference with the non-parametric and parametric analysis is that the regression tree traces out the most important interactions among the circumstances in a statistically significant way and estimates *IOP* only on the basis of those limited number of interactions which are chosen by the program as the most relevant ones. The opportunity tree is therefore able to produce an estimate of *IOP* that escapes the possible risk of over-fitting arising from unregulated number of interactions. Indeed during *2011-12*, Table 5 shows that *IOP* in consumption is less than 7% when it is estimated using the regression tree algorithm. For the same year, unequal opportunity in wage and education as estimated by the regression tree, are still about 32% and 23% of their corresponding total inequality, respectively. But for all the outcomes, *IOP* estimated by the regression tree are considerably lower than their corresponding non-parametric and parametric estimates[26].

The opportunity trees for MPCE, wage and education are presented in Figures 1, 2 and 3, respectively. First of all notice that except for MPCE, parental education has turned out to be the most important circumstance factor for all other outcomes, as denoted by the initial nodes of Figures 2 and 3. Whereas for *MPCE*, the most crucial circumstance is the occupational category

[26]Notice that although we draw the respective opportunity trees on the basis of a randomly chosen single imputed data-set, the same is not done for quantifying *IOP* under the regression tree approach. Similar to the non-parametric and parametric analysis, *IOP* is measured in the regression tree analysis using all the 20 imputed data-sets and by the index of mean log deviation.

	Measures of relative *IOP*		
	Regression tree	Parametric	Non-parametric
MPCE	0.068	0.107	0.112
Wage	0.318	0.377	0.393
Education	0.225	0.248	0.267

Table 5: Different estimations of IOP (2011-12)[a]

[a]All IOP estimates are measured by the index of mean log deviation on multiply imputed data-sets.

of fathers and the average monthly consumption expenditure is always higher when fathers are engaged in non-agricultural jobs (Figure 1). Not only for consumption, having an agricultural family background is always less advantageous for all outcomes, whenever it matters,. Another common feature across most of the outcome variables is that the effect of some discriminatory social attributes like sex and caste, seem to be rather conditioned by differences in parental backgrounds. Figures 2 and 3 for example reflect that females have always less earning and educational opportunity than males, but even more so if parents have no or very little experience of formal schooling. Except for North-East India, women have less earning opportunity than men as well, if parents have no or medium level of formal schooling (Figure 2).

Similar to sex, the role of caste in the circumstance hierarchy also comes after parental attributes, but the forward caste premium is not limited to individuals with lesser educated parents only. In fact for educational opportunity, casteism has turned out to be rather relevant when parents are relatively more educated (Figure 3). However the forward *General* caste category has always better educational opportunity than the relatively disadvantageous caste groups of *OBC* and *SC/ST*, and even more so if their fathers are also engaged in white collar occupations. Similar to education, Figure 2 shows that the forward *General* caste categories also have an even better earning opportunity for most part of the country, when parents are comparatively more educated and fathers are in non-agricultural professions.

The geographical habitat have a distinguishing effect for all of our outcome variables, but its order of relevance varies across the outcomes. As compared to *MPCE* and wage, the region of residence seems to be relatively less important for generating unequal opportunity in education. While region of residence (zone) is the second most important circumstance variable after parental backgrounds for consumption and income (Figures 1 and 2), it becomes relevant for education at a later stage (Figure 3). Residents of East and Central regions however seem to have lesser opportunity on average, in both consumption and education[27].

Opportunity tree for wage earning is however limited to the casual or regular wage earners who are necessarily non-self employed and the wage tree structure may well be very different for India with the inclusion of the self-employed workers. Nevertheless as far as regular/casual earning opportunity is concerned, the circumstance of region divides the country in two parts. Although father's occupation is selected as the next important circumstance after region for the whole country, the historically destitute caste categories of *SC/ST* almost always have a better earning opportunity than the relatively upper castes in the North-Eastern region, irrespective of

[27]On a separate note, East and Central India are also found as two of the worst performing regions in terms of educational opportunity for children as well (*see* (Kundu 2019, Chapter 3)).

their father's occupational background. For the rest of the country however, earning opportunity is always better for the forward *General* caste individuals, as the average wage of them is always higher than that of the lower castes. This seemingly counter-intuitive caste dynamics is rather a region specific feature of the North-Eastern part of the country, that embodies not more than 5% of the national population, but is often called the tribal hub of India for having a much higher concentration of the marginalized lower castes of *SC/ST*s (particularly *ST*s).

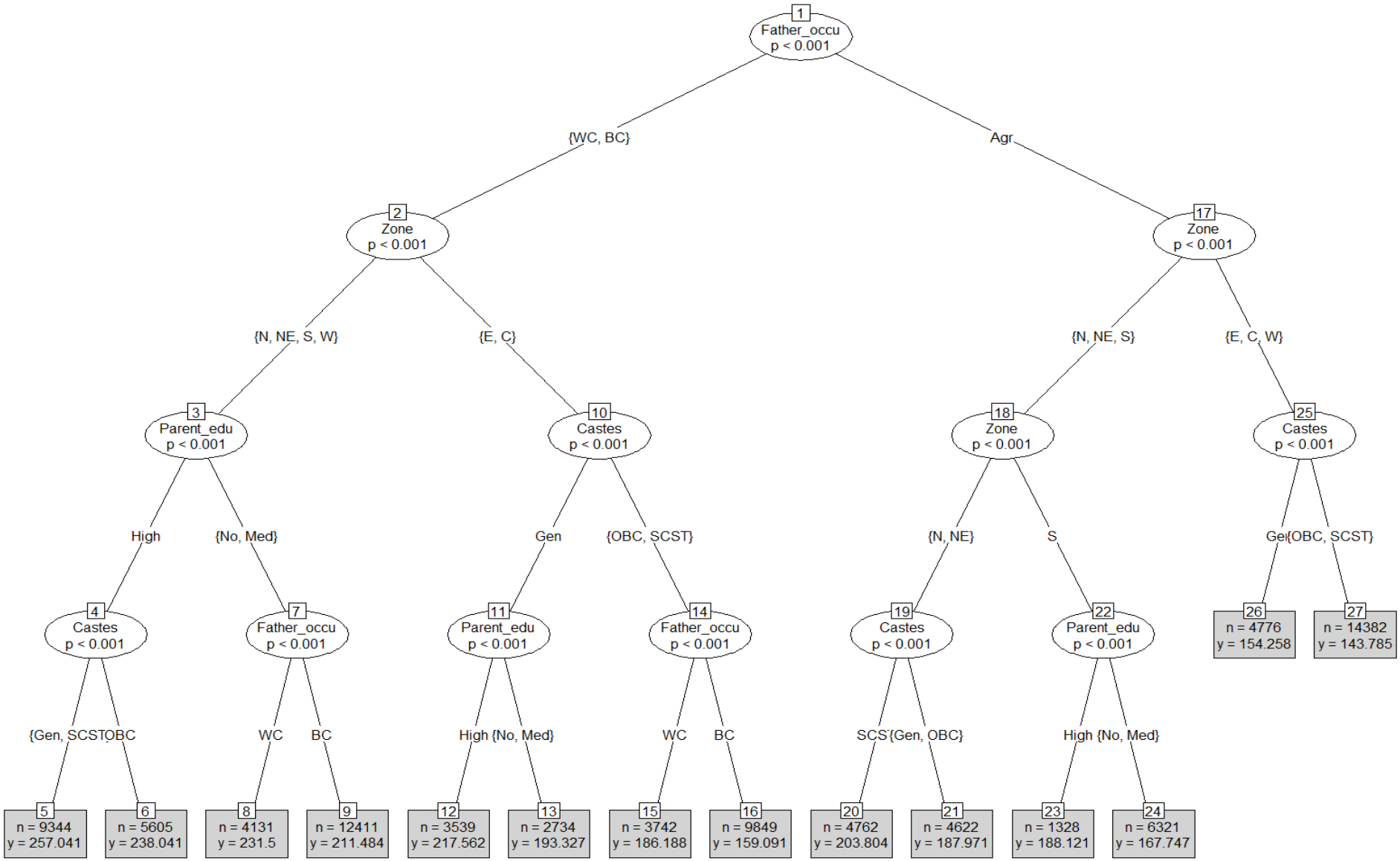

Figure 1: MPCE (*2011-12*)[a]

[a]'n' and 'y' denote the sample size and the mean MPCE in INR (Indian Rupee), respectively, for the corresponding terminal node. Parent_edu, Father_occu and zone represent the circumstances of parental education, father's occupation and region of residence, respectively.

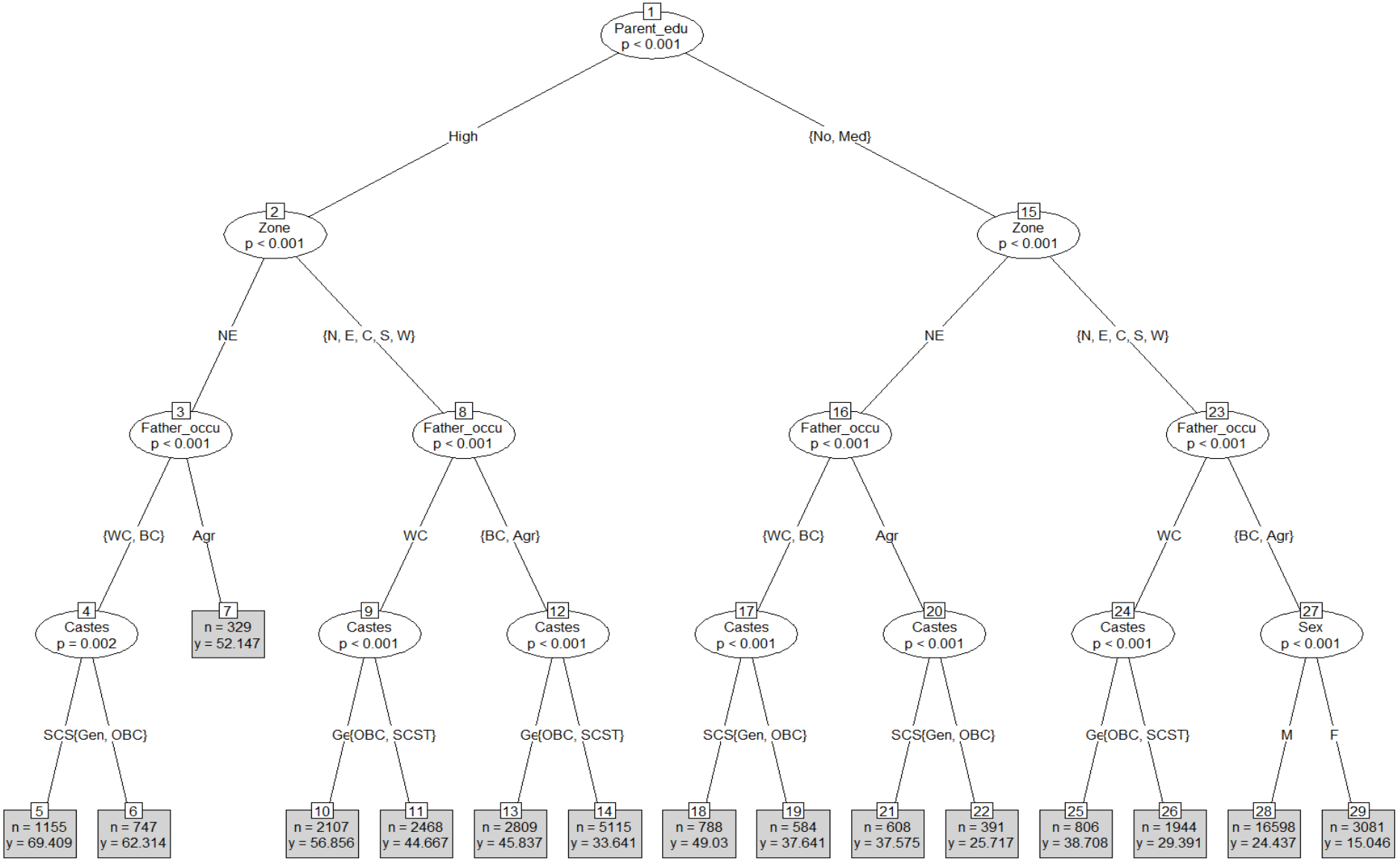

Figure 2: Wage (*2011-12*)[a]

[a]'n' and 'y' denote the sample size and the mean (daily) wage in INR (Indian Rupee), respectively, for the corresponding terminal node. Parent_edu, Father_occu and zone represent the circumstances of parental education, father's occupation and region of residence, respectively.

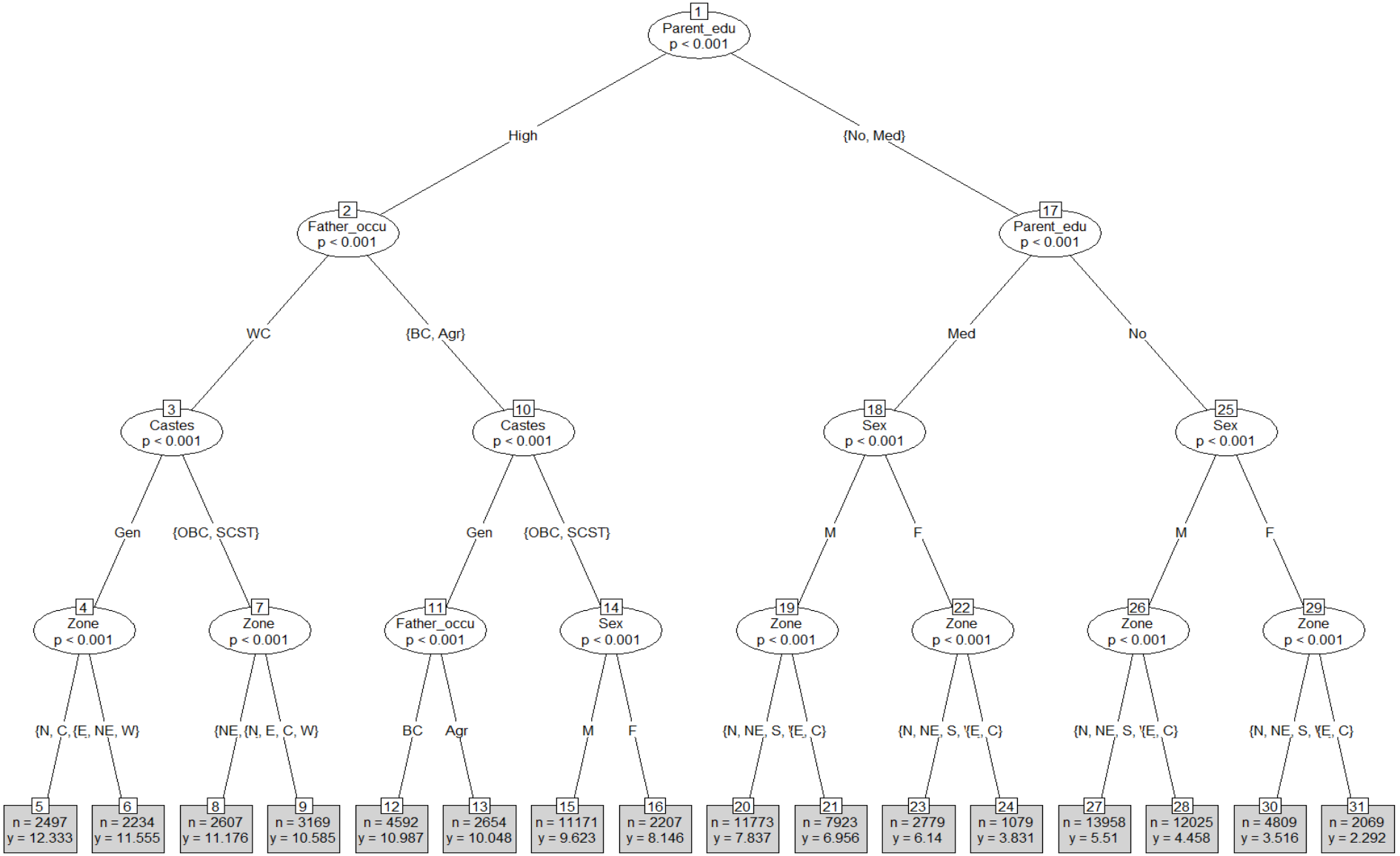

Figure 3: Education (*2011-12*)[a]

[a]'n' and 'y' denote the sample size and the mean year of education, respectively, for the corresponding terminal node. Parent_edu, Father_occu and zone represent the circumstances of parental education, father's occupation and region of residence, respectively.

5 Concluding remarks

In this paper we estimate the amount of *IOP* for India in consumption expenditure, wage earning and education, using the last two survey years of the *NSS* data base. In our assessment of *IOP* in the Indian society, we consider a set of five circumstance factors comprising of caste, sex, region, parental education and father's occupation. Using the most widely used methodologies in estimating *IOP*, we found that 27-32% of educational inequality in due to unequal opportunity. Whereas during the time frame of *2004-12* earning opportunity in India is around 32-39%, which is higher than some of the most opportunity unequal countries in Latin America. However due to the selective reporting of wage data in *NSS*, our wage analysis is limited to the non-self-employed regular or casual workers of the country and excludes a substantial portion of self-employed working adults. On the other hand, both of the non-parametric and parametric methods estimate that the share of unfair inequality in consumption (MPCE) is around 11%. But consumption for being reported as the total monthly household consumption expenditure, may not be well responsive to changes in the individual circumstances and thereby has a chance to be underestimated.

Due to the structure of *NSS* information on parental attributes is provided for 'co-resident' households only where the offspring are enumerated along with their parents. The other social circumstances like caste, sex and region on the other hand, are non-missing for the entire sample. However we found the degree of *IOP* to be substantially underestimated if parental backgrounds are omitted from the set of circumstances, whereas this is not the case when caste is omitted. In fact *IOP* in India is estimated even lower than some of the developed countries while taking the social circumstances alone (caste, sex, region). In addition we also found that in spite of numerous evidence on caste discrimination in the Indian society, taking caste as the only circumstance factor is not enough as far as quantifying *IOP* is concerned. The hierarchical division of caste is therefore not able to capture well the differences in other omitted circumstances, especially that of parental backgrounds.

Similar to the extant literature, both of our non-parametric and parametric measures of *IOP* are based on all possible interactions of the five taken circumstances, while in reality some of them may be more relevant than the others. To explore the intertwining of our circumstances we further provide the opportunity structure for India using the recently introduced approach of the regression tree analysis. We found parental education to be the most important circumstance for wage and education, whereas it is the occupational category of father that seems the most important source of unequal opportunity in consumption. Irrespective of the outcomes, individuals from agricultural family backgrounds are always worse off. In addition, the opportunity structure of India reveals the interesting way caste and sex are connected to parental education. Sex seems rather relevant when parents have lesser experience of formal schooling. Especially for earning and educational opportunity, females of poorly educated parents have significantly less opportunity than males. The forward caste premium on the other hand, is also prominent for higher educated families as well. The opportunity tree also brings forth the special case of the tribal part of India, the North-Eastern region, where the most historically disadvantageous caste categories (*SC/ST*) are actually found to be better off than the upper castes, at least in terms of regular/casual wage earning, which is never the case for the rest of the country.

Appendices

A Multiple imputation

A.1 The algorithm of multiple imputation of chained equation

To impute parental education and father's occupation, we adopt a multivariate imputation approach, in particular, the sequential regression multiple imputation algorithm of Raghunathan et al. (2001). Two other popularly used method for multivariate imputation are multivariate normal imputation and sequential univariate monotone imputation. We could not use the former, as that is applicable in case of continuous imputed variables with multivariate normal distribution. The latter is a rather quick method, but is only applicable if the pattern of the missing data is monotone, which means, if missing values for one variable is completely nested in that of the other. For either rounds, father's occupation is missing for about 20% of the co-resident data points, because they are recorded against currently employed fathers. Whereas, parental education is almost non-missing or have very few missing values for all years. But, it is only one round, the missing pattern among these two variables are monotone. Hence we have to use the iterative multivariate imputation process.

The multivariate imputation algorithm draws the imputed values through a series of univariate regressions, or equivalently, through a series of chained equations and hence, is also called the multiple imputation of chained equations (*MICE*). The underlying imputation model specification takes all the variables as predictors except the one to be imputed. First, the variables to be imputed are ordered from the least to the highest (in terms of missing values) and then start imputing the variable for which missing information is minimum, using predictors without any missing value. The next ordered variable (with second least number of missing values) is then imputed using the non-missing predictors, as well as the imputed value of the first variable. The process continues till the variable with highest number of missing value is imputed. Further, each imputation consists of multiple cycles or iterations to get more stable set of imputed values, based on which, the final vector of imputed values are drawn for the entire working sample. The algorithm is detailed in Raghunathan et al. (2001)[28]. For two imputed variables, the regression sequence is described as below.

Let X_1 and X_2 be the variables to be imputed with the fully specified vector of variables denoted by Z and let X_1 be the variable with the least number of missing values (which in our case, is parental education for all rounds). In the first cycle, X_1 is regressed on Z (*i.e.* $X_1 \rightarrow Z$) and the missing values in X_1 are imputed by simulated draws from the posterior distribution of X_1. Then X_2 is regressed on Z along with the imputed values of X_1, (*i.e.* $X_2 \rightarrow X_1^m, Z$) and imputed values of X_2 are drawn similarly. In the cycles thereafter, each of X_1 and X_2 are regressed on the fully specified variables along with the previously imputed variables. Thus, in the second cycle, the prediction sequence is $(X_1 \rightarrow X_2^m, Z)$, $(X_2 \rightarrow X_1^m, Z)$ and so on. The cycles are continued (often upto 10 to 20 iterations) to converge to a set a stable imputed values $\{X_1^1, X_2^1\}$, that constitutes the first imputed data set. The entire process with the same number of iterations are then repeated M times, to produce M copies of the imputed data sets, with

[28] Also see Royston et al. (2011), Azur et al. (2011).

imputed variables $\{(X_1^1, X_2^1), \ldots, (X_1^M, X_2^M)\}$. The non-parametric and parametric measures of *IOP* are then estimated for each of these M imputed data sets and the final estimate of *IOP* is then estimated as the average of all the imputed data sets [Rubin's rule (Rubin 1986)].

Notice that, after the first cycle, all the missing values are imputed. If the missing pattern is monotone, that is, if X_2 is missing only if X_1 is missing, there is no need of further iteration. Only cycle one is repeated M times to produce multiple copies of the imputed data set. In that case the prediction sequence is like - $(X_1 \rightarrow Z)$; $(X_2 \rightarrow X_1^m, Z)$. Since X_2 is only missing when X_1 is missing, this sequence is enough to draw sensible imputed values for both the variables (Raghunathan et al. 2001). When missing pattern is arbitrary, iteration is needed so as to get a stable set of imputed values, that is repeatedly predicted by old and newly imputed values. Accordingly, we performed no iterations for year *2011-12*, for which missing pattern is detected as monotone. For the other round, we choose 20 iterations for each imputation.

A.2 Imputation model and diagnostics

The variables to be imputed in our case, are - parental education and father's occupation, where the former is generated by combining father's and mother's education[29]. To reduce imputational rigor, we consider to impute the combined parental education, instead of imputing each of the father's and mother's education (much in the spirit of 'transform then impute' (Von Hippel 2009)). We estimate an ordered logistic regression as our imputation model, to estimate parental background with a broad range of covariates, that are not missing for the entire work sample. Following the literature (Rubin 1986, Little 1988, Schafer 1999), we include three broad set of covariates - (*i*) the *analysis model variables* (caste, sex, zone along with their all possible interactions), (*ii*) the *auxiliary variables* (household size, consumption expenditure, sector, religion, along with children age, age squared, education, occupation, sex, marital status, relation to head) and (*iii*) the *survey specific variables* (sub round, second stage stratum, first stage units[30]). Following Teyssier (2017), who have used MI for the same purpose of imputing parental background for Brazil, we include the sample weight as a predictor as well (along with the normal use of sample weights in the logit model). In addition, children wage and its interaction with age is also considered for the wage sample imputation. The imputation model does not have any claim of causality, but it should fit the data well. With highly significant model chi-square statistics for all rounds, Table 6 does not indicate that our chosen imputation model is a poor fit for any of the imputed variables.

Across the rounds, 66-71% of our working sample have missing information on parental background that we needed to impute. Multiple imputation is a simulation based algorithm and hence, the power and precision of the multiply imputed values are likely to increase with the number of imputations, especially when missing data proportion is large. So far in the literature, there is no unequivocal rule to choose an *optimum* number of imputations. However, even with a high fraction of missing information, a number of literature often recommends that a modest number of imputation is good enough to generate statistically sound imputed values (Rubin

[29] In case of single-parent household, that constitute about 8% of the co-resident sample, parental education is the education of the single parent.

[30] Definition of these variables are provided in the *NSSO* data appendix (Kundu 2019).

Year	Likelihood Ratio Chi-square				Pseudo R^2	
	Parental education	[p-value]	Father's occupation	[p-value]	Parental education	Father's occupation
Work sample						
2004-05	6439.4	[0.000]	6608.6	[0.000]	0.184	0.396
2011-12	2978.2	[0.000]	4118.6	[0.000]	0.181	0.418
Wage sample						
2004-05	2646.0	[0.000]	2281.2	[0.000]	0.221	0.391
2011-12	1632.9	[0.000]	1779.4	[0.000]	0.215	0.388

Table 6: Imputation model check[a]

[a]We report McFadden R^2 in particular.

1986, Schafer 1999)[31]. As shown by Rubin (1986), the relative efficiency of an infinite number of imputations subject to a finite one, is $(1+\gamma/m)^{-1/2}$, where γ and m are the fraction of missing information and the number of imputations, respectively[32]. In case of 70% missing information ($\gamma = 0.7$), the relative large sample efficiency is already 0.96 with 10 imputations, that increases to 0.98 for 20 imputations. Since in case of large degrees of freedom, each additional imputation adds little to the efficiency of the estimated parameter (Schafer & Olsen 1998), we choose to do 20 imputations for each of our imputed variables (for imputing around 68% missing data for a sample size of 0.1 million, on average). Further, in case of arbitrary missing pattern, each imputation is generated from a simulated draw of 20 iterations.

However, "a naive imputation is worse than doing nothing" (Little 1988, p 288). We have a total of 20 imputed data-set. For a randomly chosen imputation, Table 7 reports the distribution of the imputed variables in the observed data-set ('response'), the imputed data-set ('non-response') and the completed data-set ('response'+'non-response'), for both of our final working sample and the wage sub-sample. At a glance, father's occupation seem to have been imputed better, for it has similar distribution across all the data-sets. Whereas, more parents are pointed as having no formal education for the imputed data-set. But that does not mean a faulty imputation of parental education, and in fact, the difference in its distribution is indicative of a rather sensible imputation. The non-co-resident sample, who are, on average, 10 years older than the co-resident ones, are supposed to have older parents. Provided the substantial educational improvement over time for all generations, as is reflected by Table 8 and 9, older parents are more likely to be deprived of formal education, exactly as they are imputed. On the other hand, Table 8 also shows that occupational composition of the samples does not seem to be markedly different due to co-incidence. Provided low occupation mobility in India, this is likely to be true for parents as well[33]. Besides, as a robustness check, we found that the pattern of the distributions of the imputed values are similar for many other imputed data sets as well.

[31]Besides, in case of a complex imputation model with large number of variables and sample size, even a single imputation takes hours to complete, and so more, if it is iterative. The computational effort associated with the higher number of imputations in these cases, are often too prohibitively high to make little sense to increase the number of imputations for a marginal increase in efficiency (Allison 2003, Von Hippel 2005, Azur et al. 2011).

[32]Missing information, strictly speaking, is not the same as the number of missing data points. With high correlation between the missing variables and the observed covariates, γ is actually lesser than the percentage of missing values (Graham et al. 2007). However, they are the same in the simplest setting.

[33]Also note from Table 9, that in *2011-12*, 56% of co-resident sample have their fathers working in agricultural sector, while 45% of them are in agricultural job themselves (Table 8).

	2004-05			*2011-12*		
	obs.	imp.	comp.	obs.	imp.	comp.
Work sample imputation diagnostics						
Parental education						
No_schooling	0.390	0.524	0.478	0.305	0.379	0.354
Below primary	0.286	0.246	0.260	0.280	0.263	0.269
Above primary	0.324	0.230	0.262	0.415	0.358	0.378
Father's occupation						
White collar	0.119	0.105	0.109	0.198	0.194	0.195
Blue collar	0.322	0.359	0.348	0.353	0.408	0.393
Agricultural	0.559	0.536	0.542	0.449	0.398	0.412
Wage sample imputation diagnostics						
Parental education						
No_schooling	0.412	0.495	0.471	0.326	0.379	0.363
Below primary	0.278	0.238	0.250	0.270	0.247	0.254
Above primary	0.311	0.266	0.279	0.404	0.374	0.383
Father's occupation						
White collar	0.115	0.097	0.101	0.181	0.124	0.138
Blue collar	0.405	0.443	0.435	0.473	0.489	0.485
Agricultural	0.480	0.460	0.464	0.346	0.387	0.377

Table 7: Imputation diagnostics[a]

[a]Where 'obs.', 'imp' and 'comp.' stand for *observed*, *imputed* and *completed* data set, respectively. For reporting the imputed and the completed data set, we choose one imputation at random (among 20 imputations).

B Additional tables and figures

	age	hhsize	%male	%rural	%SC/ST	%married	%noschool	%agri	%wage	N
Working sample (total)										
2004-05	32.1	5.5	0.77	0.76	0.29	0.82	0.36	0.53	0.41	127002
2011-12	32.8	5.0	0.82	0.72	0.30	0.82	0.24	0.45	0.48	90574
Non-response part (non-co-resident)										
2004-05	35.1	4.7	0.71	0.76	0.30	0.96	0.45	0.54	0.43	83201
2011-12	35.5	4.4	0.77	0.71	0.31	0.96	0.31	0.46	0.49	59592
Response part (co-resident)										
2004-05	25.7	7.0	0.92	0.76	0.26	0.50	0.18	0.49	0.37	43801
2011-12	26.5	6.5	0.93	0.72	0.26	0.50	0.10	0.42	0.45	30982

Table 8: Summary statistics: working sample, response part and non-response part[a]

[a]Response part correspond to the co-resident sample for which parental information is provided in the data-set, whereas the non-response part are the non-co-resident samples for which parental backgrounds are needed to be imputed. Working sample is the union of the response and the non-response part. 'age' and 'hhsize' reports the mean age and household size of the respective sample. %male, %rural, %SC/ST, %married, %noschool, %agri and %wage reports the share of males, rural inhabitants, SC/STs, married individuals, samples without any formal schooling, samples engaged in agricultural jobs and samples who further have the information on wage data, respectively. The last column (N) reports the respective sample size.

	age *father*	age *mother*	%noschool *father*	%noschool *mother*	%noschool *both*	edu_year *child*	edu_year *father*	edu_year *mother*	%dom_duty *mother*	%agri *father*
Co-resident parents										
2004-05	54.0	48.8	0.47	0.75	0.45	6.6	3.8	2.1	0.60	0.63
[61]	(0.07)	(0.06)	(0.00)	(0.00)	(0.00)	(0.03)	(0.03)	(0.02)	(0.00)	(0.00)
2011-12	54.5	49.5	0.42	0.68	0.40	7.7	4.2	2.5	0.72	0.56
[68]	(0.09)	(0.09)	(0.01)	(0.01)	(0.01)	(0.05)	(0.05)	(0.03)	(0.01)	(0.01)

Table 9: Co-resident sample summary of parents[a]

[a]Standard errors are in parentheses and rounds in squared brackets. In particular, 'noschool father/mother' indicates fathers/mothers who are deprived of any formal schooling, whereas 'noschool both' means none of the parents have any formal schooling. 'edu_yr' abbreviates as the year of education. '%dom_duty mother' denotes the share of mothers who have reported not to be in the labor market for attending domestic duties and '%agri_father' are the share of fathers engaged in agriculture related jobs.

	MPCE		Wage		Education	
	2004-05	2011-12	2004-05	2011-12	2004-05	2011-12
Ref: General						
OBC	-0.157***	-0.058***	-0.037***	-0.136***	-0.082***	-0.192***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
SC/ST	-0.288***	-0.063***	-0.074***	-0.146***	-0.189***	-0.278***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Ref: Primary plus						
Primary or below	-0.066***	-0.067***	-0.345***	-0.274***	-0.437***	-0.385***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
No schooling	-0.109***	-0.102***	-0.498***	-0.409***	-1.029***	-0.933***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Ref: White collar						
Blue collar	-0.068***	-0.075***	0.066**	-0.094***	-0.129***	-0.105***
	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.00)
Agricultural	-0.008	-0.226***	-0.001	-0.303***	-0.291***	-0.208***
	(0.53)	(0.00)	(0.29)	(0.00)	(0.00)	(0.00)
Ref: North						
East	-0.293***	-0.359***	-0.338***	-0.229***	-0.283***	-0.247***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Central	-0.214***	-0.403***	-0.313***	-0.267***	-0.188***	-0.205***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
North-East	-0.082***	-0.179***	-0.031	-0.022	0.050*	0.027
	(0.00)	(0.00)	(0.20)	(0.39)	(0.03)	(0.37)
South	-0.088***	-0.196***	-0.197***	-0.055**	-0.069***	-0.009
	(0.00)	(0.00)	(0.00)	(0.02)	(0.00)	(0.62)
West	-0.161***	-0.281***	-0.330***	-0.299***	0.032	0.015
	(0.00)	(0.00)	(0.00)	(0.00)	(0.10)	(0.58)
Ref: Male						
Female	0.039***	0.010	-0.403***	-0.316***	-0.576***	-0.429***
	(0.00)	(0.38)	(0.00)	(0.00)	(0.00)	(0.40)
Intercept	5.36***	5.38***	3.48***	3.90***	2.36***	2.45***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Table 10: Reduced form OLS: for MPCE, Wage and Education[a]

[a]Standard errors are in parenthesis. (***, **, *) correspond to 1%, 5% and 10% level of significance, respectively.

References

Allison, P. D. (2000), 'Multiple imputation for missing data: A cautionary tale', *Sociological methods & research* **28**(3), 301–309.

Allison, P. D. (2003), 'Missing data techniques for structural equation modeling.', *Journal of abnormal psychology* **112**(4), 545.

Alon, S. (2009), 'The evolution of class inequality in higher education: Competition, exclusion, and adaptation', *American Sociological Review* **74**(5), 731–755.

Arneson, R. (1989), 'Equality of opportunity and welfare ', *Philosophical Studies* **56**, 77–93.

Asadullah, M. N. & Yalonetzky, G. (2012), 'Inequality of educational opportunity in India: Changes over time and across states', *World Development* **40**(6), 1151–1163.

Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. (2011), 'Multiple imputation by chained equations: what is it and how does it work?', *International journal of methods in psychiatric research* **20**(1), 40–49.

Bourguignon, F., Ferreira, F. H. & Menéndez, M. (2007), 'Inequality of opportunity in Brazil ', *Review of income and wealth* **53**(4), 585–618.

Brunori, P., Ferreira, F. & Peragine, V. (2013), *Inequality of Opportunity, Income Inequality, and Economic Mobility: Some International Comparisons*, Paus E. (eds) Getting Development Right; Palgrave Macmillan, New York.

Brunori, P., Hufe, P. & Mahler, D. G. (2018), 'The roots of inequality: Estimating inequality of opportunity from regression trees.'.

Checchi, D. & Peragine, V. (2010), 'Inequality of opportunity in Italy ', *Journal of economic inequality* **8**, 429–450.

Checchi, D., Peragine, V. & Serlenga, L. (2010), 'Fair and unfair income inequalities in Europe', *IZA discussion paper No. 5025* .

Cogneau, D. & Mesplè-Somps, S. (2008), 'Inequality of opportunity for income in five countries of Africa', *John Bishop, Buhong Zheng (ed.); Inequality and Opportunity: Papers from the Second ECINEQ Society Meeting, Emerald Group Publishing Limited* **16**, 99–128.

Cohen, G. A. (1989), 'On the currency of egalitarian justice ', *Ethics* **99**, 906–944.

Deaton, A. & Dreze, J. (2002), 'Poverty and Inequality in India: A Re-Examination ', *Economic and Political Weekly* **37**(36), 3729–3748.

Dev, S. M. & Ravi, C. (2007), 'Poverty and Inequality: All-India and States, 1983-2005', *Economic and Political Weekly* **42**(6), 509–521.

Dworkin, R. (1981*a*), 'What is equality? Part 1: Equality of resources ', *Philosophy & public affairs* **10**, 283–345.

Dworkin, R. (1981*b*), 'What is equality? Part 1: Equality of welfare ', *Philosophy & public affairs* **10**, 185–246.

Ferreira, F. H. & Gignoux, J. (2011), 'The measurement of inequality of opportunity: theory and an application to Latin America ', *The review of income and wealth* **57**(4).

Ferreira, F. H. & Peragine, V. (2015), 'Equality of Opportunity: Theory and evidence ', *Policy research working paper* (WPS 7217, Washington, D.C: World Bank Group).

Graham, J. W., Olchowski, A. E. & Gilreath, T. D. (2007), 'How many imputations are really needed? some practical clarifications of multiple imputation theory', *Prevention science* **8**(3), 206–213.

Himanshu (2007), 'Recent Trends in Poverty and Inequality: Some Preliminary Results', *Economic and Political Weekly* **42**(6), 497–508.

Himanshu (2018), 'Widening Gaps: India Inequality Report 2018', *Oxfam India* .

Hnatkovska, V., Lahiri, A. & Paul, S. B. (2012), 'Caste and labor mobility ', *Applied economics* **4**(2).

Hnatkovska, V., Lahiri, A. & Paul, S. B. (2013), 'Breaking the Caste Barrier: Intergenerational Mobility in India', *Journal of human resources* **48**(2), 435–473.

Hothorn, T., Hornik, K. & Zeileis, A. (2006), 'Unbiased recursive partitioning: A conditional inference framework', *Journal of Computational and Graphical statistics* **15**(3), 651–674.

Jong-Sung, Y. & Khagram, S. (2005), 'A comparative study of inequality and corruption', *American sociological review* **70**(1), 136–157.

Kundu, T. (2019), 'Inequality of opportunity: the case of india', *PhD Thesis, Université de Cergy-Pontoise (laboratoire THEMA) & ESSEC Business School, Paris* .

Lefranc, A., Pistolesi, N. & Trannoy, A. (2009), 'Equality of opportunity and luck: definitions and testable conditions, with an application to income in France (1979-2000)', *Journal of public economics* **93**, 1189–1207.

Little, R. J. (1988), 'Missing-data adjustments in large surveys', *Journal of Business & Economic Statistics* **6**(3), 287–296.

Marchenko, Y. V. & Eddings, W. (2011), 'A note on how to perform multiple-imputation diagnostics in stata', *College Station, TX: StataCorp* .

Marrero, G. A. & Rodrìguez, J. G. (2011), 'Inequality of opportunity in the United States: trends and decomposition', *Research on Economic Inequality* **19**, 217–216.

NSSO (2008), 'NSS Report No. 533: Migration in India: July, 2007-June, 2008', *National Sample Survey Organization, Ministry Of Statistics and Program Implementation (MOSPI), Govt. of India* .

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. & Solenberger, P. (2001), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', *Survey methodology* **27**(1), 85–96.

Ramos, X. & Van de Gaer, D. (2012), 'Empirical approaches to inequality of opportunity: Principles, measures and evidence ', *IZA discussion paper no. 6672* .

Rawls, J. (1971), *A theory of justice*, Cambridge: Harvard University Press.

Roemer, J. (1993), 'A Pragmatic Theory of Responsibility for the Egalitarian Planner', *Philosophy & Public Affairs* **22**, 146–166.

Roemer, J. (1998), *Equality of Opportunity*, Harvard University Press, Cambridge, MA.

Roemer, J. E. & Trannoy, A. (2013), 'Equality of Opportunity', *Cowles foundation discussion paper no. 1921* .

Royston, P., White, I. R. et al. (2011), 'Multiple imputation by chained equations (mice): implementation in stata', *J Stat Softw* **45**(4), 1–20.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

Rubin, D. B. (1986), 'Basic Ideas of Multiple Imputation for Nonresponse', *Survey Methodology, Statistics Canada* **12**(1), 37–47.

Salehi-Isfahani, D., Hassine, N. B. & Assaad, R. (2014), 'Equality of opportunity in educational achievement in the middle east and north africa', *The Journal of Economic Inequality* **12**(4), 489–515.

Schafer, J. L. (1999), 'Multiple imputation: a primer', *Statistical methods in medical research* **8**(1), 3–15.

Schafer, J. L. & Olsen, M. K. (1998), 'Multiple imputation for multivariate missing-data problems: A data analyst's perspective', *Multivariate behavioral research* **33**(4), 545–571.

Singh, A. (2012), 'Inequality of opportunity in earnings and consumption expenditure: The case of Indian men', *The review of income and wealth* **58**(1), 79–106.

Teyssier, G. (2017), 'Inequality of opportunity: New measurement methodology and impact on growth', *Seventh ECINEQ Meeting, New-York City (mimeo)* .

Von Hippel, P. T. (2005), 'Teacher's corner: How many imputations are needed? a comment on hershberger and fisher (2003)', *Structural Equation Modeling* **12**(2), 334–335.

Von Hippel, P. T. (2009), '8. how to impute interactions, squares, and other transformed variables', *Sociological methodology* **39**(1), 265–291.