ISBN: 978-605-307-788-6
www.rsepconferences.com

Decomposing the Effect of Childhood Environment on Earnings

Ivan Žilić[1]

Abstract

In this paper we contribute to the strain of literature on early childhood environment by analyzing the effect of different childhood conditions and environments on subsequent earnings. Using the Panel Study of Income Dynamics (PSID) we are able to link adult earnings with parental and family characteristics (for example, parental income, education and house value). Using family background variables we preform classification exercise via clustering algorithm and group individuals into two groups based on family background and childhood environment. Using these distinctive groups and decomposition methods we separate the total earnings gap on the explained part (composition effect: the effect of individual-level covariates) and the unexplained (structure effect: the return on the individual-level covariates). Results indicate that there are significant mean and quantile gaps between two groups and that the total gap is increasing in percentile difference. The explained effect is around 70 percent across the quantiles, which implies that 30 percent of differences in wages cannot be explained by individual characteristics but group membership, i.e. childhood environment. Furthermore, running a detailed decomposition, we conclude that single most important covariate is education as differences in education across the groups can explain approximately 50 percent of total difference and 80 percent of explained difference. These results, using a novel approach, corroborate conclusions regarding substantial influence of early childhood environment on earnings.

Keywords: childhood environment, adult earnings, PSID, decomposition methods.

JEL codes: I30, J31, J70.

1. Introduction and motivation

The importance of childhood environment on adult economic outcomes has captured a lot of attention in economic literature. The main message of this strain of literature can be summarized in a powerful statement by Conti and Heckman (2014): 'Children raised in disadvantaged environments start behind and usually stay behind throughout their lifetimes'.

As Heckman and Mosso (2014) state, at least 50 percent of variability of lifetime earning can be explained by personal attributes by age of 18, and any study that determines how conditions in childhood affect life outcome is indeed a study of family influence. These conclusions have been reached in numerous studies.

[1] The Institute of Economics, Zagreb, Email: izilic@eizg.hr

ISBN: 978-605-307-788-6

For example, the effect of childhood environment is analyzed through implemented programs such as Perry Preschool Program, Carolina Abecedarian Project, or Jamaican stimulation intervention (Heckman et al. (2013), Temple and Reynolds (2007), Gertler et al. (2013)). Heckman et al. (2013) show that Perry Preschool Program, which aimed at disadvantaged, low IQ African Americans aged from 3-4, had a profound positive effect on employment, education, earning, marriage, participation in healthy behavior and reduced participation in crime. Surprisingly, this intervention did not positively affect participants IQ, but it improved substantially what authors call Externalizing Behaviors (aggressive, antisocial and rule breaking behaviors). This example only emphasizes that personality traits (soft skills) are quite important for latter outcomes (Heckman and Kautz (2012)).

Gertler et al. (2013) show that early intervention programs such as aforementioned Perry Preschool Program, Carolina Abecedarian Project and Chicago Child-Parents Centers are much more effective than interventions than began during the school age. This emphasizes the importance of timing of intervention, the earlier the intervention, the higher the probability of remedying early disadvantage (Conti and Heckman (2014)).

Other branch of research has investigated the family and community effects through sibling correlations in various outcomes such as IQ, non-cognitive skills, years of schooling, and long run earnings, Bjorklund et al. (2009). Sibling correlations, a measure which shows what fraction of the variation of variable in interest can be attributed to factors that siblings share, show that at least one fifth of various outcomes originates from family and neighborhood influences (Bjorklund et al. (2009)).

The importance of family background can be seen in a fact that at age of 3 children from professional families can speak 50 percent more words than children from working class families and twice more than children from welfare families (Fernald et al. (2013)).

Aforementioned approaches show that family background and early childhood environment are profoundly important for adult labor market outcomes and one could invoke arguments that emphasize morality, social justice and fairness of early interventions. Heckman and Mosso (2014) claim that policies that redistribute resources toward disadvantaged children in early years are based on grounds of efficiency and not only on social sentiment.

In this paper, we contribute to this strain of literature by analyzing the effect of early childhood environment on individuals earning. In particular, using the Panel Study of Income Dynamics (PSID) we construct a data set where information about parent's characteristics (education, income, occupation, house value...) are constructed and linked with individual's variables. Doing so, we obtain a data set where we have individuals' earnings and covariates, but also variables that provide information regarding parent and family characteristics when individual was up to 15 years old. Using these family and parental variables, individuals are grouped (clustered) in distinctive groups; basic descriptive statistics shows that groups do differ in not only the values of the variables used for clustering but also in earnings. After clustering, decomposition method is used to separate the total earnings gap between two groups on the explained part (composition effect, covariate effect[2]) and the unexplained (structure effect).

To emphasize the distinction between the effect of covariates and the structure effect, it is useful to relate the problem to the literature on inequality of opportunity. For example, Roemer (1998) while discussing the inequality of opportunity, provides distinction between circumstances and efforts. Circumstances are a state for an individual, one cannot control them; and they are usually perceived as race, gender, age, socioeconomic and family background. Efforts on the other hand are in control of an individual, as for example education. In this framework, efforts are a function of circumstances; people with a more favorable set of circumstances can exhibit a higher level of efforts. Using the above mentioned classification, Bourguignon et al. (2007) and Pistolesi (2009) estimate the parts of inequality that can be attributed to circumstances and effort and conclude that inequality of opportunity presents between 20 and 43 percent of earnings inequality.

[2] Covariates are individual's characteristics, not the parent's. In fact, parents and family characteristics are used only in clustering algorithm.

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

In this paper in order to decompose the gap, generalization of Oaxaca-Blinder mean decomposition methods is used. This generalization enables the detail decomposition (contribution of each covariate to covariate and structure effect) beyond the mean. This generalization, developed by Firpo et al. (2009) uses rencetered influence function of desired statistic of outcome variable instead of outcome variable itself to decompose the gap for arbitrary quantile.

Decomposition results show that there are significant mean and quantile gap between two groups. This total gap is increasing in percentile difference (for example, 90th percentile difference between two group's earning is greater than mean difference). The explained effect is around 70 percent across the quantiles, which implies that 30 percent of differences in wages cannot be explained by individual characteristics but group membership of individuals. This result is consistent with aforementioned conclusions regarding substantial influence of early childhood environment in earnings.

Furthermore, running a detailed decomposition, we conclude that single most important covariate is education as differences in education across the groups can explain approximately 50 percent if total difference and 80 percent of explained difference. Race is explaining around 10 percent of total difference and 15 percent of explained difference. Gender is not playing significant role in gap explanation, which is intuitive as gender is not systematically assigned across groups. Decomposition was run on subsample of males, and subsample of white individuals. Most of the conclusions are valid also here, except the quantile difference across the groups has a steeper slope.

This paper contributed to existing literature in twofold manner. Firstly, it investigates the importance of socioeconomic background using novel angle, i.e. using detailed quantile decompositions which gives new results and insights. This approach allows us to separate the effect of covariates and the effect of background, which is a novelty in this strain of literature. Secondly, in the decomposition methods usually the group variable is a priori given (for example, gender, race, and private/public sector). In this paper, groups that are used in decomposition are obtained via clustering algorithm. Clustering was implemented on variables that define early child hood environment thus giving the groups economic interpretation.

The paper is organized as follows: Section 2 explains how the data were obtained and which variables we use to cluster individuals; Section 3 explains the clustering method implemented and results the clustering procedure; Section 4 presents the RIF based decomposition methods, Section 5 brings estimation and presents the results while Section 6 concludes.

2 Data

In the next section we present in detail the data used. As every individual has to be linked with its parent (form age 1 to 15), detailed and rich data set is of paramount importance.

In particular, data set needs to be comprehensive enough to achieve two goals:

- Unconfoundedness, i.e. $D_g \perp \varepsilon \mid X$. Therefore the data set on individuals (children) has to be rich enough to claim that unobserved effect is randomly assigned across groups (which will be obtained later through clustering).
- Early childhood environment description, i.e. data set containing information about individual's parents needs to be rich enough to use this data to 'type' the family background.

For this purpose Panel Study of Income Dynamics (henceforth PSID) is used. PSID is a longitudinal data base that contains household data from 1968 (from 1968 to 1997 yearly data are available, from 1997 to 2011 biannual data is available). This data base has desired possibility of relating children and parents (using the Family Identification Mapping System-FIMS). For example, from 2011 PSID we can gather income data for a

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

43 year old individual and relate this information with data on his/her father in 1968 to 2011 (when individual was born), which makes PSID excellent data source for inter-generational research.

PSID is made out of PSID Family level and PSID Individual level data. Family level contains most of the data regarding household's income (in 2011 there are 5142 variables available), while individual level contains mostly variables that ease the identification of the individuals and basic information about them. First we present the data that are going to enter individual's wage equation. Individual data from 2011 PSID Family level are presented in Table 1.

Table 1: Individual data

Name of the variable	**PSID name (2011)**
Log hourly wage*	ER52237, ER46829, ER52175, ER46767
Age	ER47317
Gender	ER47318
Race	ER51904
Education	ER52405
Sector	ER47480
US state	ER47304
Tenure	ER47513, ER47515, ER47515
Experience	ER51955
Marriage status	ER47323

Source: PSID

Note that 'Log hourly wage' is in fact the log of 2009-2011 (two values) average. This is done to avoid idiosyncratic shocks in wages; we are trying to capture long run earnings component. For the variables 'Gender', 'Sector' and 'US state' dummy variables were created. Only individuals from 23-43 years old are in the sample, the upper bound (43) as we need to obtain information about their family when they were born, and PSID runs from 1968. Detailed description of the variables can be found on PSID website by searching their 2011 PSID name. In the sample only head of the families that worked more than 1040 hours last year are considered (to avoid people that are not working). Next we need to gather data that will define the environment to which a child was exposed to. To do so we can use two sources of data from PSID:

- The data about individual's family gathered from 2011 family file.
- The data about individual's family obtained from relating the individual with its parents.

Let us clarify the first group. In the family file, among other variables, there are information about persons parents (like education, were they poor...) and these data will be used as a part of data that will define the environment. The details are presented in Table 2.

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

Table 2: Family information form 2011 family file

Name of the variable	PSID name (2011)
Poor parents	ER51901
Live with parents	ER51902
Education father	ER51869
Occupation father	ER51874
Occupation mother	ER51884
Number of siblings	ER51887, ER51893

Source: PSID

Variable 'Poor parents' corresponds to individual's subjective impression whether (s)he was raised in a poor, average or well-off family. From this variable dummy were created. Variable 'Live with parents' gives information whether individual lived with both parents up to 16 years old. From variables 'Education father', 'Occupation father' and 'Occupation mother' dummy variables were created, while 'Number of siblings' was obtained by adding the number of sisters and brothers (+1).

Next group of data was obtained by connecting the information of the parents with the individual's information. The goal is to have information about the family where the individual was born and raised in (therefore when (s)he was 0-15 years old). To do so, a number of variables were taken from 1968-2003 (as individuals that are 43 in 2011 were born in 1968, and as individuals that are 23 in 2011 were 15 in 2003). The following table presents the detailed data:

Table 3: Family information obtained by connecting parents and children

Name of the variable	PSID name (2003)
House value*	ER21043
Hours worked father*	ER24080
Family income*	ER24100
Father income*	ER24116

Source: PSID

Note that * denotes that variables are available when individual was 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15 years old (skipping years comes from the fact that PSID from 1997 is released every two years). Therefore, in above table we have 52 variables. Also, it is worth mentioning how incomes and house values were transformed to get the comparable values (we cannot directly compare the individual's family income when they were for example 10, if one is 25 and other is 40, as there is 15 year mismatch). To do so, variable United States Department of Agriculture (USDA) needs standard (ER24139 for 2011), that captures how much food is needed for a specific household in a specific year, is used. All variables that are denominated in monetary unit are divided by this variable (household and year specific), so for example 'Family income' is ratio of how much did that family earn that year and how much they needed to buy the elementary food. After connecting individuals with their parents and cleaning the data we are left with 1803 individuals.

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

Next, the data from Table 2 and 3 (total 58 variables) are going to be used to cluster (type) individuals based on the family socio-economic background.

3. Clustering

As mentioned in the Data section, once we collect the variables defining the early childhood environment, we can proceed to cluster (categorize, type) individuals. Before clustering, we need to decide how to deal with missing values of environment variables (variables that define socio-economic background of a family where individual was raised in). Nominal variables, such as father's education ranging from 1-9 where each number corresponds to some level of completed education, are transformed into dummy variables and the missing values (non-response) were modeled through a dummy (as one cannot simply discard missing values- non response might not be random). Therefore, missing values are problem only in continuous (for example: hours of work, house value, family income) variables.

If we discard every individual that has at least one missing value in environment variables, we will lose 500 out of 1800 individuals. To mitigate this loss of observations, we will discard individuals for which the percentage of missing values is 25 or more.[3] For the rest of individuals (ones that have 75 percent of environment variable values), we will impute the missing data. In this way we preserve almost 200 individuals. Missing data is imputed using Iterative robust model-based imputation.[4] This is a regression based imputation technique where at each step of iteration one variable is used as a response variable and the remaining variables serve as the regressors; for the detailed version of the algorithm see Templ et al. (2011). After this pre-processing of the data, one can proceed to clustering. The two most used clustering families are: centroid based clustering and hierarchical clustering.

In centroid based clustering techniques (such as Kmeans or Kmeadians) the objects are assigned based on distance to central vector (mean or a median of cluster). As the objects are added to a cluster, the values in central vector change, and the algorithm stops once the values in central vector are not changing any more.

The problem with these methods is that initial center vector values are picked randomly and therefore repeating clustering will not yield the same cluster results. As this randomness in not suitable in this framework, we turn to hierarchical clustering methods (for survey of clustering techniques see Xu et al. (2005). Hierarchical clustering method that is used can be summarized in the following steps:

1. Each of n individuals forms a cluster (therefore there are $n = k$ clusters).
2. Compute $k \times k$ dissimilarity matrix, where the characteristic element is $d(i, j)$, i.e. distance between clusters i and j.
3. Based on a dissimilarity matrix, merge two closest cluster into one (therefore there are $k = n - 1$ clusters now).
4. Repeat steps 2 to 4 until there is only one cluster left.

Two objects that govern hierarchical clustering are distance measure and linkage function. As we are dealing with mixed data (categorical and continuous) the distance measure picked for clustering purpose is Gower distance[5], for details see Gower (1971). The linkage function is important as it defines how the distance is

[3] Percentage of missing values across variables varies from 9 to 16 percent (mean 14.3), so there is no obvious variable candidate to exclude from analysis.

[4] Command $irmi$ in R in VIM package. Note also that package $ameilaII$ was used which implements multiple imputation procedure.

[5] Default distance for mixed data in Stata.

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

measured between non singleton clusters. For example, suppose you want to compute the distance $d_{(ij)l}$, which measures how much are clusters C_{ij} (cluster with two sub-clusters, i and j, i.e. the ones that were joined in step 3 of the algorithm) and C_l close. Then, the Lance-Williams family of linkages functions is:

$$d_{(ij)l} = \alpha_i d_{il} + \alpha_j d_{jl} + \beta d_{ij} + \gamma \,|\, d_{il} - d_{jl} \,| \tag{1}$$

where, for example d_{ij} is the distance between cluster i and cluster j, and $\alpha_.$, β and γ are parameters that need to be set. Usual linkage functions, such as single, average and complete, can be represented as a specific case of linkage above (by changing the $\alpha_.$, β and γ parameters). For our purpose, we used a linkage that is used in a Ward algorithm[6], i.e.:

$$\alpha_i = \frac{n_i + n_l}{n_i + n_j + n_l}; \alpha_j = \frac{n_j + n_l}{n_i + n_j + n_l}; \beta = -\frac{n_l}{n_i + n_j + n_l}; \gamma = 0 \tag{2}$$

where n_i, n_j and n_l represent the sizes of corresponding clusters. This linkage is fairly standard and resembles average linkage. The reason why it is favored in our analysis is that it clusters individuals into similarly sized groups.[7]

After the algorithm has clustered all individuals in one group we can split the cluster into arbitrary number of groups based on a cluster dendogram. The intuition is to go backwards, i.e. see which clusters have been merged in the last iteration of the algorithm and split them into two clusters.

[6] Ward algorithm is executed if squared Euclidean distance measure is used with mentioned linkage function. However linkage function that is used in Ward algorithm can be used with other distances. To perform clustering functions *hclust* and *daisy* are used in R, variables are standardized before clustering.

[7] Other distance measures and linkage function were tried, they give similar results in terms of descriptive statistics of each group, but they do differ in size of groups.

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6
www.rsepconferences.com

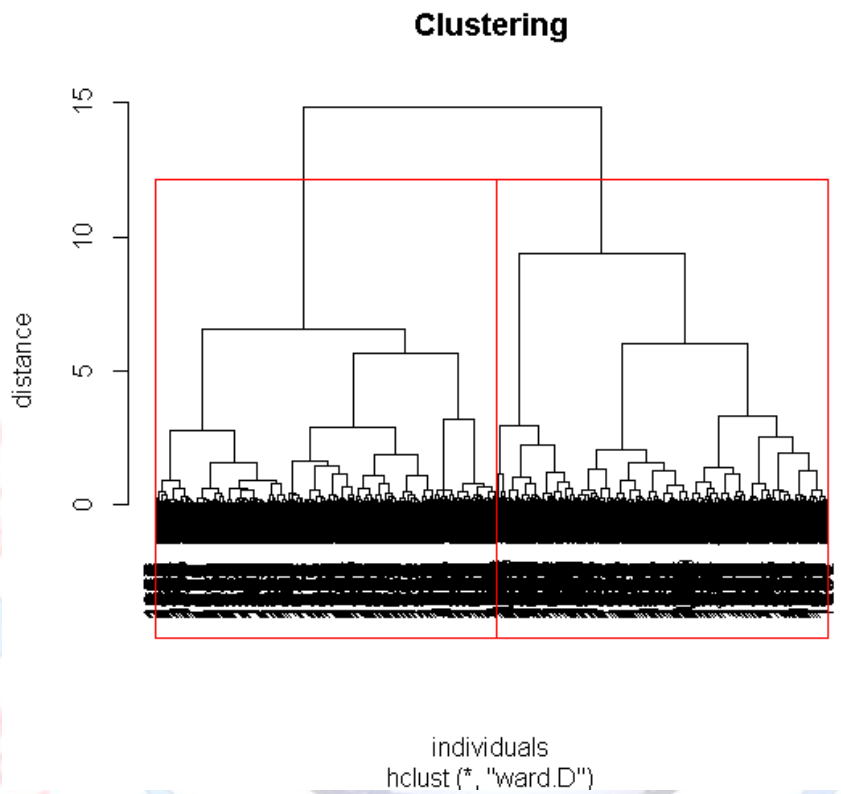

Figure 1: Clustering based on childhood environment

Source: PSID and author's calculations

Based on this clustering we have two distinct and disjoint groups. Next we present descriptive statistics of variables that were used to cluster the individuals.

Table 4: Chosen summary statistics of cluster 1

Variable	Obs.	Mean	Std. Dev.
Father's age when child was born	745	28.962	5.513
Parents were well off (dummy)	745	0.396	0.489
No. of siblings	745	3.004	1.516
Father had at least some college education (dummy)	745	0.685	0.465
House value at age 10	745	24.589	31.134
Family income ate age 10	745	13.992	17.45

Source: PSID and author's calculations

ISBN: 978-605-307-788-6

Table 5: Chosen summary statistics of cluster 2

Variable	Obs.	Mean	Std. Dev.
Father's age when child was born	770	27.168	6.546
Parents were well off (dummy)	770	0.287	0.453
No. of siblings	770	3.479	1.952
Father had at least some college education (dummy)	770	0.229	0.42
House value at age 10	770	9.378	11.224
Family income ate age 10	770	7.61	6.571

Source: PSID and author's calculations

First, we can see that two clusters are of similar sizes (745 and 770 individuals). Second, based on a selected descriptive statistics it is evident that individuals which are in cluster 1 come from the families where father was older when individual was born, father was more educated, the family was smaller, their house valued more, family income was higher and subjective perception of their parents wealth was more favorable. Therefore, cluster indeed separated individuals in way that they do differ in family socio-economic background. Therefore, we have disjoint groups for which we can attach economic interpretation and we can proceed to decomposition methods.[8] First, let us see if the wage distribution between two groups indeed differs.

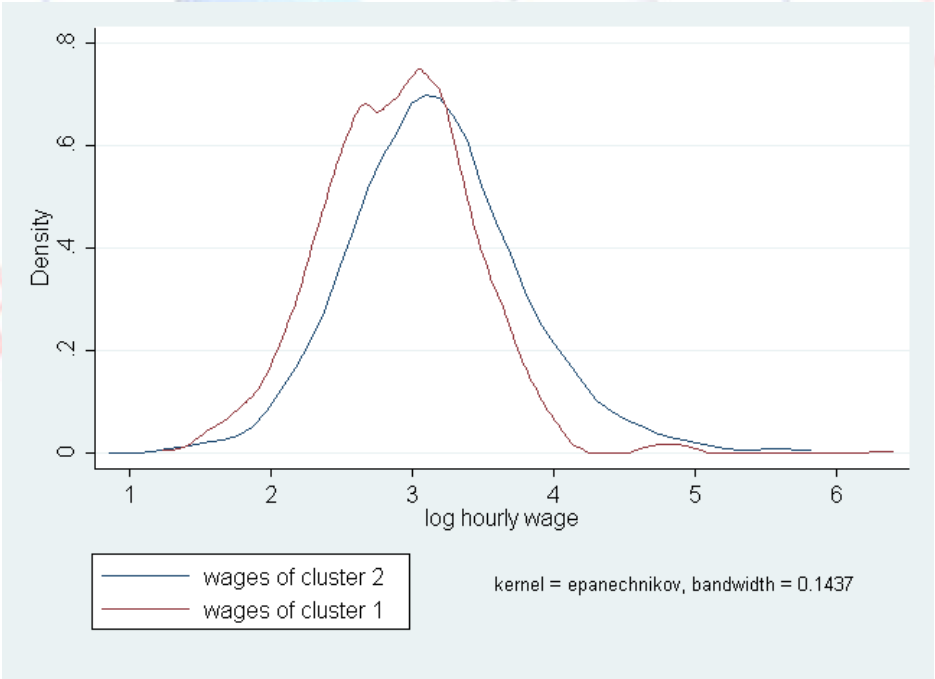

Figure 2: Density estimation of cluster 1 and 2 wages

Source: PSID and author's calculations

[8] With the same method we obtained also 3 and 4 clusters. It that case group 1 and 2 are further partitioned. The problem with the case with two or more groups is that we cannot order them. For example, one group is composed of individuals that have richer parents, while in other individuals have more educated parents, making the socioeconomic ordering of groups ambiguous.

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

From figure above it is evident that distribution of wages in cluster 1 is shifted to the right when compared to wages in cluster 2.

One part of this wage differential is generated by more favorable covariate distribution of individuals in cluster 1. For example, in group 1 the average of years of completed education is almost 15, while in groups it is just above 13.5. Also, group 1 is composed mostly of white individuals (95 percent) while in group 2 this percentage is lower (82 percent). The fundamental question of decomposition methods is to ask which portion of this wage differential comes from the group membership and which one from the different covariate distribution. In other words, do people from cluster 1 earn more (and how much) because they come from favorable socio-economic background or they simply have covariates that are more desirable on labor market?

Simple way to asses if the early childhood environment is indeed important is to run standard Mincerian equation[9] and add group membership dummy as a covariate[10].

Table 6: Mincerian equation including group membership variable

Variable	Coefficient	Std. Err.
Education	0.096	0.008
Age	0.014	0.003
Female	-0.086	0.038
White	0.167	0.037
Group 2 dummy	-0.072	0.030

Source: PSID and author's calculations

Therefore, from this simple model we can see that group membership indeed is a significant variable in explaining wage differences across individuals. In fact, the coefficient next to group 2 dummy (group that was exposed to less favorable early conditions) is comparable to coefficient next to female dummy. Indeed, even if control for standard covariates in wage equation, socio-economic background of an individual seems to play important role in labor outcomes.

4. Decomposition methods

Once the environmental variables are used to cluster data into two distinctive groups, decomposition methods can be applied. Decomposition methods were introduced in the seminal papers of Oaxaca (1973) and Blinder (1973). Research that followed dealt with generalization to detailed decomposition and decomposing beyond the mean; for example Juhn et al. (1993), DiNardo et al. (1996), Donald et al. (2000), Machado and Mata (2005), Fortin et al. (2011), Rothe (2010), and Chernozhukov et al. (2013).

[9] Regress the log hourly wage on education, age, tenure, squared tenure, experience, squared experience, US state dummies, sector dummies, race dummy, gender dummy and marriage dummy.

[10] Note that in this regression, as well as in all other, sample weights are used (ER33637 in 2011).

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

The basic idea of the decomposition is the following (Fortin et al. (2011)): we have two distinct groups that are labeled 2 for the 'not favorable' childhood environment and 1 for the 'favorable' childhood environment. Then $F_{Y_g|D_s}$, where $g,s=1,2$, is a distribution of the potential outcome Y_g for a worker that is in group s. If $g=s$ then the distribution is observed, while if $g \neq s$ the distribution is counterfactual. The distributional statistic of interest $\nu(F_{Y_g|D_s})$ can be mean or arbitrary quantile. The observed ν-difference in wages between the workers in two groups (the 'not favorable' childhood environment group and the 'favorable' childhood environment group) is then:

$$\Delta_O^\nu = \nu(F_{Y_1|D_1}) - \nu(F_{Y_2|D_2}) \tag{3}$$

The goal of decomposition methods is to decompose the ν-distributional statistic of the aggregate difference into:

$$\nu(F_{Y_1|D_1}) - \nu(F_{Y_2|D_2}) = [\nu(F_{Y_1|D_1}) - \nu(F_{Y_2^C:X=X|D_1})] + [\nu(F_{Y_2^C:X=X|D_1}) - \nu(F_{Y_2|D_2})] \tag{4}$$

where $\nu(F_{Y_2^C:X=X|D_1})$ is the counterfactual distributional statistic that represents what would an individual from group 2 earn if (s)he had the same distribution of covariates as an individual from group 1. Therefore, the first bracket gives us the wage structure effect, the ν-difference of wages between workers that differ only in group membership (distribution of the covariates is identical). The second bracket is the ν-difference that comes from the fact that workers from two groups have different covariate distributions. Therefore, the total difference can be decomposed into the structure and the covariate effect, i.e.:

$$\Delta_O^\nu = \Delta_S^\nu + \Delta_X^\nu \tag{5}$$

It is worthwhile to further explain what the structure difference represents. If the outcome variable for the individual i in group g is generated by: $Y_{gi} = m_g(X_i, \varepsilon_i)$, then the structure term Δ_S^ν is the difference in the function m_g between the two groups (difference in return on the covariates).

Also, as the covariate effect is a sum of effects of particular covariates: $\Delta_X^\nu = \sum_{k=1}^{K} \Delta_{Xk}^\nu$, one can also estimate each Δ_{Xk}^ν (effect of particular covariate in the covariate term). The same also can be done for the structure effect term (but summing up does not necessarily hold) when one obtains the difference in the parameters associated with covariate k between m_g.

The basic mean Oaxaca Blinder decomposition can estimated easily (following Nopo(2008)). Due to the Law of iterated expectations and consequently the fact that mean regression has unconditional interpretation, the outcome variables (log hourly wage) of two groups can be written as $\overline{Y}^1 = \hat{\beta}^1 \overline{X}^1$ and $\overline{Y}^2 = \hat{\beta}^2 \overline{X}^2$. Where

 DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

$\overline{X}^g$ represents the the average of characteristics X of group g. If we add and substract the counterfactual $\hat{\beta}^1\overline{X}^2$ (how much would a person coming from favorable economic background earn if (s)he had characteristics of a person from not favorable environment) to the difference of above equations we obtain:

$$\overline{Y}^1 - \overline{Y}^2 = (\hat{\beta}^1\overline{X}^1 - \hat{\beta}^1\overline{X}^2) + (\hat{\beta}^1\overline{X}^2 - \hat{\beta}^2\overline{X}^2) \tag{6}$$

or

$$\overline{Y}^1 - \overline{Y}^2 = \hat{\beta}^1(\overline{X}^1 - \overline{X}^2) + (\hat{\beta}^1 - \hat{\beta}^2)\overline{X}^2 \tag{7}$$

The first part on the right hand side represents the composition (covariate, explained) effect, while the second part represents structure (unexplained) effect.[11] Also, this simple procedure can be extended to obtain contribution of each particular covariate to composition and structure effect.

To generalize this procedure beyond the mean is not a simple task. The main problem arises from the fact that quantile regression, unlike mean regression, does not have unconditional interpretation.

For example if $E[Y \mid X] = \beta X$ then $E(Y) = E[E[Y \mid X] \mid X] = E(X)\beta$. But there is no such rule for the quantiles, therefore if for the arbitrary $\tau \in (0,1)$ $Q_\tau(X) = X\beta_\tau$ but $Q_\tau \neq E[Q_\tau(X) \mid X] = E(X)\beta_\tau$. This enables straightforward generalization of Oaxaca-Blinder decompositions to arbitrary quantiles. In order to obtain decomposition that that is valid for quantiles, counterfactual unconditional density of outcome of interest must be applied, like in Chernozhukov et al. (2013). In order to circumvent the problem related to quantile regression interpretation and at the same time avoid cumbersome application of methods from Chernozhukov et al. (2013), RIF regression from Firpo et al. (2009) (quantile regression with unconditional interpretation) will be used. This allows us to obtain detailed quantile decomposition.

The basic intuition of the method is to find a way to apply Law of iterated expectations. As Firpo et al. (2009) show, this can be achieved introducing influence functions. Influence function measures how arbitrary statistic changes when one observation of the sample changes its value (small perturbation of the distribution of the outcome variable). For the quantile Q_τ, where $\tau \in (0,1)$ the influence function is:

$$IF(Y;Q_\tau) = \frac{\tau - \mathbf{1}[Y \leq \tau]}{f_Y(Q_\tau)} \tag{8}$$

where $1[\cdot]$ denotes indicator function and $f_Y(Q_\tau)$ represents density of outcome variable at a quantile Q_τ. Note that $E[IF(Y;Q_\tau)] = 0$ (this holds for any statistic). Recentered influence function is sum of statistic of interest and its influence function. In our case, the recententered influence function is:

[11] Notice that counterfactual exercise was reversed with respect to the explanation of decomposition methods above. In fact, in the covariate effect one could evaluate differences covariates at $\hat{\beta}^1$, $\hat{\beta}^2$ or any combination of them. In our estimation $\hat{\beta}^{pooled}$ will be used.

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

$$RIF(Y;Q_\tau) = Q_\tau + \frac{\tau - \mathbf{1}[Y \le \tau]}{f_Y(Q_\tau)} \tag{9}$$

As the expectation of a $IF(Y;Q_\tau)$ is zero $E[RIF(Y;Q_\tau)] = Q_\tau$. Firpo et al. (2009) take the conditional expectation of to be linear,[12] i.e. $E[RIF(Y;Q_\tau) \mid X] = X\beta_\tau^{RIF}$. Therefore, due to the law of iterated expectations:

$$E[RIF(Y;Q_\tau)] = E[X]\beta_\tau^{RIF} \tag{10}$$

Hence, using RIF straightforward generalization of Oaxaca Blinder beyond mean can be obtained. As Firpo et al. (2009) state, if $Y = \beta X + \varepsilon$ and $\varepsilon \perp X$, unconditional quantile partial effect is equal to the structural (true) quantile parameter β_τ. This method can be easily implemented in Stata, using command 'rifreg' recentered influence function of outcome variable for quantile of interest can be estimated; these estimates are then used in Oaxaca Blinder decomposition instead of outcome variable to decompose the gap for the desired quantile.

5. Estimation and Results

Next we turn to estimation and results. The Oaxaca Blinder and RIF regression decomposition between two groups will be estimated using standard Mincerian equation where dependent variable is log hour wage and explanatory variables are: education, age, tenure, tenure squared, work experience, work experience squared, dummy for females, dummy for white, dummy for married, 49 dummies for US states and 23 dummies for employment sector.

Following table compares Oaxaca Blinder decomposition with RIF decompositions for 10th, 25th, 50th, 75th and 90th percentile.

Table 7: OB and RIF aggregate decompositions

	(1) **Oaxaca Blinder**	**(2)** **RIF_10**	**(3)** **RIF_25**	**(4)** **RIF_50**	**(5)** **RIF_75**	**(6)** **RIF_90**
Total difference	0.272	0.207	0.222	0.267	0.304	0.433
t statistics	(7.19)	(3.82)	(4.98)	(5.85)	(5.73)	(5.25)
Decomposition						
Explained	0.189	0.144	0.149	0.187	0.230	0.303
t statistics	(6.67)	(4.22)	(4.95)	(5.87)	(5.97)	(5.35)
Unexplained	0.0834	0.0626	0.0732	0.0795	0.0743	0.130
t statistics	(2.55)	(1.08)	(1.74)	(1.95)	(1.55)	(1.68)

Source: PSID and author's calculations

[12] Firpo et al. (2009) estimate conditional expectation nonparametrically, the results in their application are quite similar.

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

There are several conclusions that are apparent from the above table. First, there is difference in means (and in the chosen quantiles) of log hour wages between the two groups. Also the difference is higher as we go towards higher percentiles of wage distribution. For example, the 90th percentile difference between two groups is 0.433 log points while the mean difference is 0.267. Existence of this difference enables us to proceed with the decomposition.

Second, 70 percent of mean difference can be explained with differences in the covariates (explained part), while the rest is left unexplained. Therefore, for the mean gap, 30 percent of difference comes from the family environment a person has been exposed when young.

This is true for the quantile gaps as well. In order to facilitate the reading of results, following graphs are presented

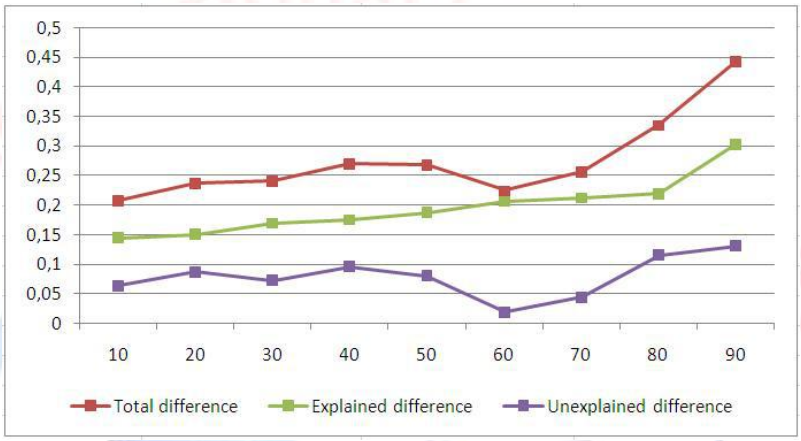

Figure 3: Total, explained and unexplained difference

Source: PSID and author's calculations

As already stated, it is clear that the difference in the two hourly wages is increasing as we considered higher percentile gap. The explained difference is increasing across all the quantiles, sharply rising after 80th percentile. This is true for unexplained difference as well, if the decline after the median is disregarded.

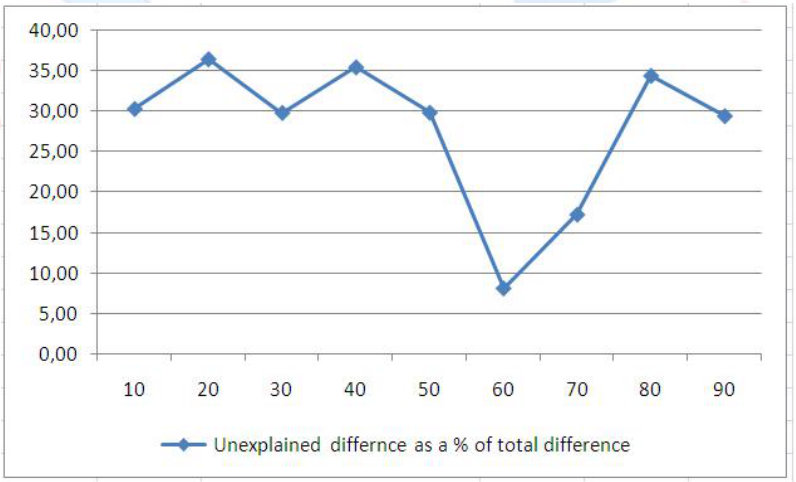

Figure 4: Unexplained difference as % of total difference

Source: PSID and author's calculations

 DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

Figure 4 shows that although unexplained difference is increasing, its share in total difference is stable around 30 percent (again, subtracting from obvious gap after the median). Therefore, almost one third of gap across the distribution of two wages is due to the family background.

As already noted in the explanation of the method, a detailed decomposition is available as well. Following figure graphs the contribution of chosen covariates in the explained effect (for solution to problem of categorical variables contribution see Yun (2005)). This is informative as it gives insight on sources of variation of wages between two groups. Three chosen covariates are education, gender and race. Differences in levels of education can explain roughly 50 percent of the total difference. Hence, half of the gap between individuals from good favorable environment and not so favorable environment comes from the more finished education of the first group. This can be seen though descriptive statistics of education across groups. In group with favorable background the average of finished years of education is 14.9 (minimum 8 and maximum of 17), while in unfavorable family background group average years of finished years of education is 13.5 (minimum 0 and maximum of 17). Also, differences in education accounts for around 75 percent of explained differences. Therefore, as expected, education is single most important covariate in Mincerian wage equation and indeed a means to promote egalitarian policies.

Race explains slightly below 10 percent of total difference, and around 15 percent of explained differences. This can be seen also in racial composition of the groups, group with favorable family background has 95 white individuals, while the other has 82 percent of whites. Therefore, racial composition explained a portion of differences between groups.

Gender seem not to explain much of the total or explained differences. This comes from the fact that gender is random across two groups (both groups have roughly one quarter of females). This results is also intuitive as there should not be any systematical difference between genders of children across different socioeconomic groups.

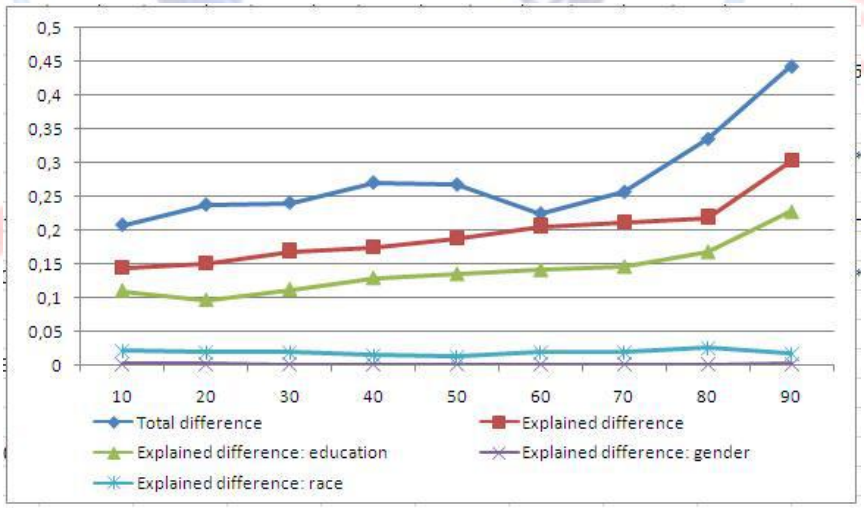

Figure 5: Detail decomposition

Source: PSID and author's calculations

 DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

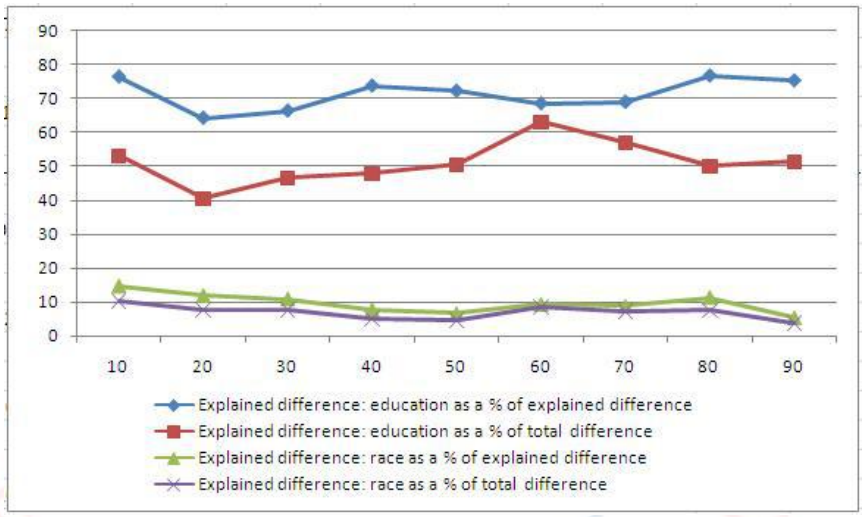

Figure 6: Detail decomposition %

Source: PSID and author's calculations

It is worthwhile decomposing the gap only for the subpopulation of males (enough observations; for subpopulation of females, as there are not many observations, results are not very robust). If we compare the gap for the males to gap with the whole sample the difference for males is around 80 percent of the difference of the whole sample up to the median difference. After the median this percentage rises culminating at 90th percentile difference where the gap for males is higher than the gap for the whole sample. Therefore, percentile difference for the males follows a steeper profile than the one for the whole sample. The rest of the conclusions are similar to those ones obtained from the whole sample. Around one third of the gap cannot be explained with covariate differences, differences in education explain half of the total wage gap, and two thirds of the explained gap.

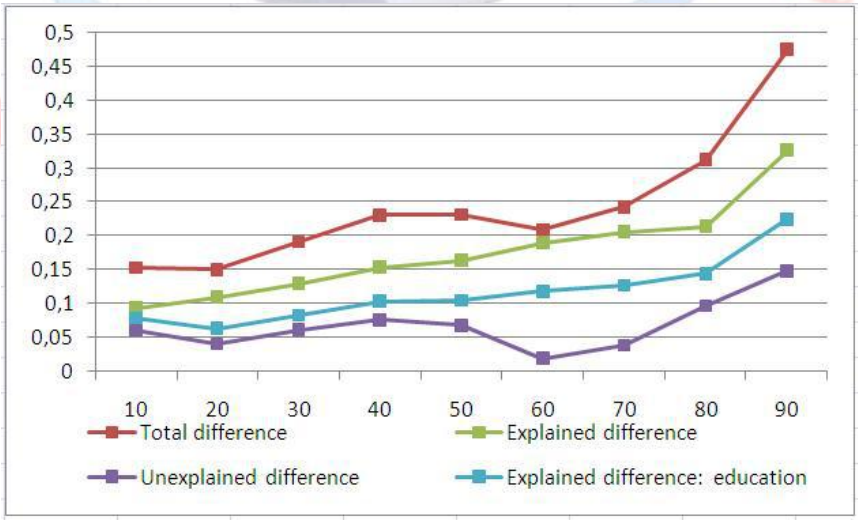

Figure 7: Detail decomposition males only

Source: PSID and author's calculations

 DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

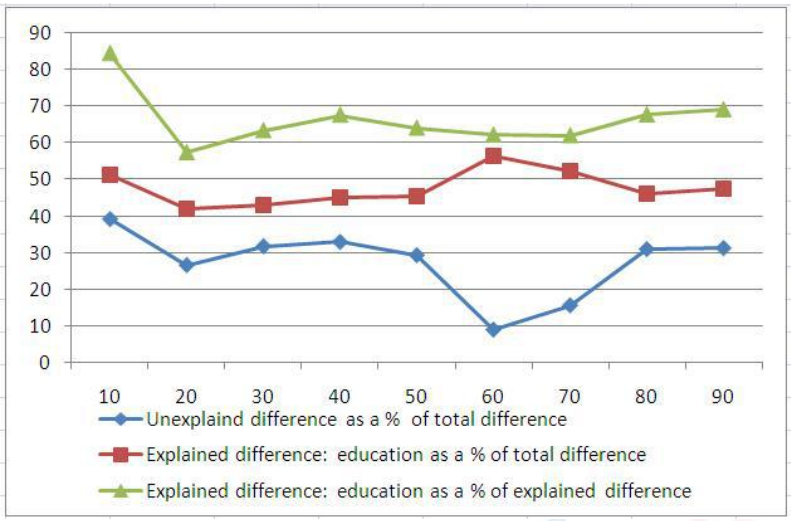

Figure 8: Detail decomposition males only %

Source: PSID and author's calculations

The same can be done for the subpopulation of white individuals. The total gap for whites only shows same patterns as the total gap obtained with the whole sample. They do slightly differ in the detail decomposition; education seem to have profound effect on the explained and total gap. Differences in education across groups can explain close to 60 percent of total differences and around 80 percent of the explained differences.

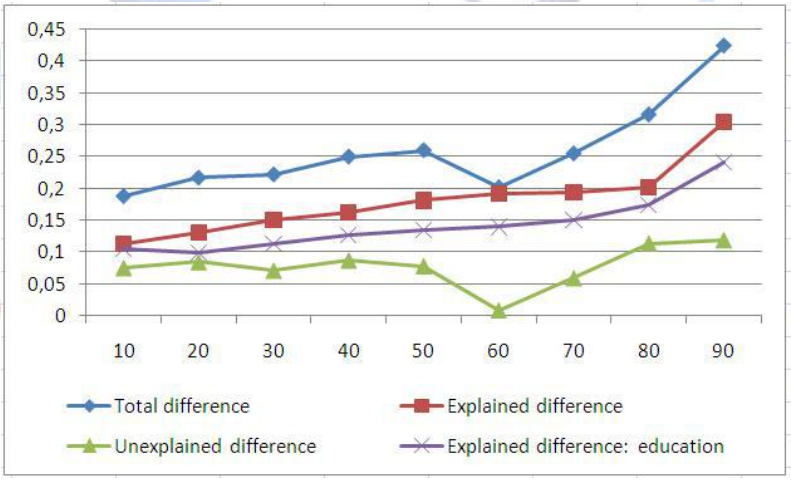

Figure 9: Detail decomposition white only

Source: PSID and author's calculations

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

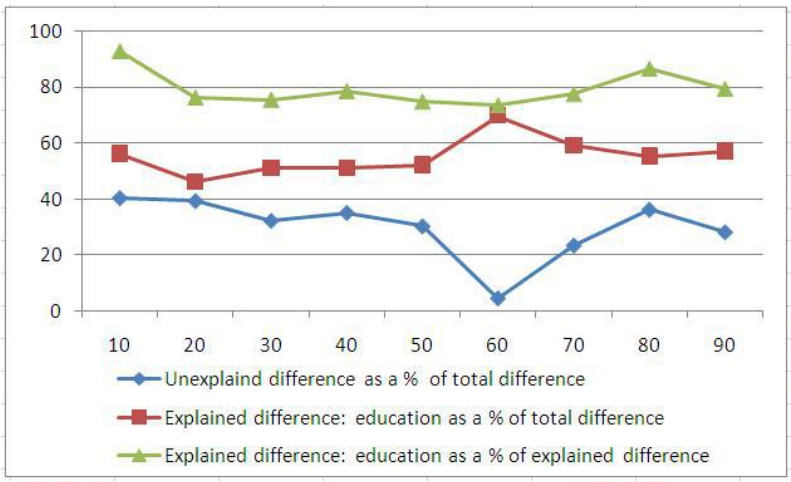

Figure 10: Detail decomposition white only %

Source: PSID and author's calculations

6. Conclusion

The fact that the socioeconomic background of the family individual was born in is important for the whole array of adult outcomes is well established result as well as continuous subject of research. This paper contributes to the literature of family influences by decomposing the effect of childhood environment of earnings. Differences from the wages between individuals that differ in socioeconomic background can come from: more favorable distribution of covariates and unexplained effect that goes beyond the covariates. In order to estimate this covariate and structure effect Panel Study of Income Dynamics (PSID) was used to construct a data set where information about parents characteristics linked with children's variables. Doing so, we obtain a data set where we have individual's earnings and covariates, but also variables that provide information regarding parent and family characteristics when individual was up to 15 years old. As the data on parents is multivariate and we need to split the individuals in two groups, individuals are clustered on parent's characteristics. Using hierarchical clustering algorithm, individuals were grouped into distinctive groups that indeed differ in socioeconomic background. For example, one group is made out the individuals that have richer, more educated and older parents compared to individuals from other group. Using this grouping and applying RIF based decompositions, we obtained the results that are in line with other papers in this field, but add insight as total differences are separated on effect of covariates and unexplained effect. The explained effect is around 70 percent across the quantiles, which implies that 30 percent of differences in wages cannot be explained by individual characteristics but group membership of individuals. Furthermore, running a detailed decomposition, we conclude that single most important covariate is education as differences in education across the groups can explain approximately 50 percent of total difference and 80 percent of explained difference. Therefore, 50 percent differences in wages between individuals that differ in socio-economic background can be explained with education, while one third of the differences are explained by family background

DOI: 10.19275/RSEPCONFERENCES110

ISBN: 978-605-307-788-6

References

Björklund, A., M. Jäntti M. & M. J. Lindquist (2009). "Family background and income during the rise of the welfare state: brother correlations in income for Swedish men born 1932–1968 ". *Journal of Public Economics*, 930 (5):0 671–680.

Blinder, A. S. (1973). "Wage discrimination: reduced form and structural estimates ". *Journal of Human resources*, pages 436–455.

Bourguignon, F., F. H. G. Ferreira & M. Menéndez (2007). "Inequality of opportunity in Brazil ". *Review of Income and Wealth*, 530 (4):0 585–618.

Chernozhukov, V., I. Fernández-Val & B. Melly (2013). "Inference on counterfactual distributions ". *Econometrica*, 810 (6):0 2205–2268.

Conti, G. & J. Heckman (2014). *Economics of Child Well-Being*. Springer.

DiNardo, J., N. M. Fortin & T. Lemieux (1996). "Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach ". *Econometrica*, 640 (5):0 1001–1044.

Donald, S. G., D. A. Green & H. J. Paarsch (2000). "Differences in wage distributions between Canada and the United States: An application of a flexible estimator of distribution functions in the presence of covariates ". *The Review of Economic Studies*, 670 (4):0 609–633.

Fernald, A., V. A. Marchman & A. Weisleder (2013). "SES differences in language processing skill and vocabulary are evident at 18 months". *Developmental Science*, 160 (2):0 234–248.

Firpo S., N. M. Fortin & T. Lemieux (2009). "Unconditional quantile regressions". *Econometrica*, 770 (3):0 953–973.

Fortin N., T. Lemieux, & S. Firpo (2011). "Decomposition methods in economics". *Handbook of labor economics*, 4:0 1–102.

Gertler P., J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. M. Chang & S. Grantham-McGregor (2013). "Labor market returns to early childhood stimulation: a 20-year followup to an experimental intervention in Jamaica". Technical report, National Bureau of Economic Research.

Gower G. C. (1971). "A general coefficient of similarity and some of its properties". *Biometrics*, pages 857–871, 1971.

Heckman, J. & T. Kautz (2012). "Hard evidence on soft skills". *Labour economics*, 190 (4):0 451–464.

ISBN: 978-605-307-788-6

Heckman, J. & S. Mosso (2014). "The economics of human development and social mobility". Technical report, National Bureau of Economic Research.

Heckman, J., R. Pinto & P. Savelyev (2013). "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes". *American Economic Review*, 1030 (6):0 2052–86.

Juhn, C., K. M. Murphy & B. Pierce (1993). "Wage inequality and the rise in returns to skill". *Journal of Political Economy*, p 410–442.

Machado, J. & J. Mata (2005). "Counterfactual decomposition of changes in wage distributions using quantile regression". *Journal of Applied Econometrics*, 200 (4):0 445–465.

Ñopo, H. (2008). "Matching as a tool to decompose wage gaps". *The Review of Economics and Statistics*, 900 (2):0 290–299.

Oaxaca, R. (1973). "Male-female wage differentials in urban labor markets". *International Economic Review*, pages 693–709.

Pistolesi, N. (2009). "Inequality of opportunity in the land of opportunities, 1968–2001". *The Journal of Economic Inequality*, 70 (4):0 411–433.

Roemer, J. E. (1998). *Equality of opportunity*. Cambridge University Press.

Rothe, C. (2010). "Nonparametric estimation of distributional policy effects". *Journal of Econometrics*, 1550 (1):0 56–70.

Temple, J. A. & A. J. Reynolds (2007). "Benefits and costs of investments in preschool education: Evidence from the child–parent centers and related programs". *Economics of Education Review*, 260 (1):0 126–144.

Templ, M., A. Kowarik & P. Filzmoser (2011). "Iterative stepwise regression imputation using standard and robust methods". *Computational Statistics & Data Analysis*, 550 (10):0 2793–2806.

Xu, R. & D. Wunsch, (2005). "Survey of clustering algorithms". *Neural Networks, IEEE Transactions on Neural Networks*, 160 (3):0 645–678, 2005.

Yun, M. S. (2005). "A simple solution to the identification problem in detailed wage decompositions". *Economic Inquiry*, 430 (4):0 766–772.