A Study of the Relationship Between Income and Opportunity Inequality Distribution Based on Big Data Analysis

Endong Zhu(✉)

School of China Economics and Management Academy, Central University of Finance and Economics, Beijing, China
18810788167@163.com

Abstract. Based on intergenerational equity theory, this paper empirically investigates the contribution of individual and family variables to income differences through the lens of "inequality of opportunity" and "inequality of effort" based on a big data perspective, using questionnaires primarily targeted at highly educated youth. The results show that individual income is less constrained by 'inequality of opportunity' factors, such as gender, family and parental education, and type of parental occupation. Party members, middle-class families, doctoral degrees and higher-paying industries were more likely to have higher incomes. In summary, inequality of opportunity has less of an impact on income in the current period than individuals.

Keywords: component · Inequality of Opportunity · Income Distribution · Intergenerational Equity

1 Introduction

1.1 Background of the Selected Topic and Significance of the Study

According to Roemer (2002), the factors affecting individual income can be categorized into objective environmental variables and subjective effort variables driven by subjective will [1]. In the decomposition of income differences, existing studies have defined the external environmental differences that are beyond the control of individuals, such as socioeconomic and family background, as "inequality of opportunity"; and the different levels of effort, such as the level of education and type of occupation, which are chosen later in life, as "inequality of effort" (Roemer, 2002; Marschall, 2002). (Roemer, 2002) [9]. When income inequality is determined by individual choice, policy authorities apply the "payoff principle" without any social intervention (Fleurbaey, 2009). In this paper, we explore the income differences of the population from the perspective of distinguishing between "inequality of opportunity" and "inequality of effort" [2], explore the degree of inequality in the income gap, and provide empirical suggestions for achieving income growth with equality of opportunity.

N. Radojević et al. (Eds.): ICAID 2022, AHIS 7, pp. 1219–1227, 2023.
https://doi.org/10.2991/978-94-6463-010-7_126

1.2 Combing the Literature on Inequality of Opportunity and Inequality of Efforts

To date, there has been a continuous innovation in the measurement and decomposition of inequality of opportunity, with Fleurbaey and Peragine (2009) distinguishing between ex ante (same environment) and ex post (same effort) approaches to measuring inequality of opportunity. In addition, Ferreira and Gignoux (2011) propose a parametric approach to measure inequality of opportunity. In an earlier study in China, Gong Feng (2010) proposed a measure of equality of opportunity in an intergovernmental fiscal transfer system; Wu Guixiao (2013) tested the effects of household registration, household economic status, years of parental education, and the number of children in the household on opportunities for further education from the perspective of inequality of educational opportunities for residents.

2 Research Design on Effort and the Impact of Inequality of Opportunity on Earnings

2.1 Sample Selection and Data Source

The data source of this study is the CEMA course paper questionnaire, which is mainly collected by posting the link and collecting the questionnaires online, and is aimed at the graduate students and social circle of the School of China Economics and Management Academy, Central University of Finance and Economics (CEMA), Class of 2020 [3]. The questionnaire was collected through an online link and surveyed from the postgraduate students and their social circle in the class of 2020 of the Central University of Finance and Economics (CUE) to investigate their personal and family wishes. Finally, the total number of questionnaire samples received was 214, among which 78 respondents were already working, while 136 respondents were enrolled in undergraduate and graduate schools and unemployed [4].

2.2 Model Variables

According to the measurement idea of Roemer (2002) and Bourguignon (2007), the education level is divided into less than bachelor's degree, bachelor's degree, master's degree and doctorate; the industry in which they work is divided into high return, medium return and low return according to Guo Congbin (2009) for different industry return indices in China.

2.3 Descriptive Statistics and Correlation Matrix

The correlation coefficient matrix shows that for the environmental variables, party membership, being born in a municipality directly under the central government, highly educated mother, father whose occupation is non-farm, and well-off family have significant income enhancing effects. The income-raising effect of middle-class families is significant at the 1% level and has an impact coefficient as high as 0.30, which is the most

prominent environmental variable, providing preliminary verification that inequality of opportunity leads to individual income differences [5].

The correlation coefficients among the main explanatory variables in the model are low and none of the absolute values exceed 0.8, which do not affect the reliability of the regression results. Meanwhile, variables with correlation coefficients below 0.3 in absolute value account for more than 70% of the variables, and they can be regarded as extremely weakly correlated or uncorrelated [6]. For the highly correlated explanatory variables, the model will further explore whether there is multicollinearity (Multicollinearity) among the variables through the equation inflation factor VIF.

2.4 Preliminary Model Setting

The data used in this paper are cross-sectional data from questionnaires, and environmental and income variables are selected to study inequality of opportunity and inequality of effort in the income determination equation, and benchmark regressions are conducted and White heteroskedasticity tests are done (graphs omitted).

$$\begin{aligned} Ln\_income_i = {} & \alpha + \beta_1 Gender_i + \beta_2 Age_i + \beta_3 Party_i + \beta_4 Married_i \\ & + \beta_{5\sim7} Urban\_area_i + \beta_{8\sim9} Father\_edu_i + \beta_{10\sim11} Mother\_edu_i \\ & + \beta_{12} Father\_party_i + \beta_{13} Mother\_party_i + \beta_{14} Father\_industry_i \\ & + \beta_{15\sim16} Family\_income_i + \beta_{17\sim19} Education_i + \beta_{20\sim21} Industry_i \\ & + \beta_{22\sim23} Property_i + \beta_{24\sim26} Workinghours_i + \varepsilon_i \end{aligned} \tag{1}$$

The variance inflation factor (VIF) measures the degree of multicollinearity of the variables. The test results showed that the VIF of Age and Age_square variables were significantly outliers (above the threshold). With reference to previous literature practices, Age_square variables were considered for exclusion.

3 Empirical Test Results and Analysis

3.1 Model Setting

Based on the results of multicollinearity test and variable screening, this paper uses model (1) to test the effects of inequality of opportunity and inequality of effort on personal income, using multiple explanatory variables for OLS regression, in which the total number of environmental variables is 16 and the number of effort variables is 10, and obtains the reported results shown in column (1).

$$\begin{aligned} \text{Ln\_incomei} = {} & \alpha + \beta 1\text{Genderi} + \beta 2\text{Agei} + \beta 3\text{Partyi} + \beta 4\text{Marriedi} \\ & + \beta 5 \sim 7\text{Urban\_areai} + \beta 8 \sim 9\text{Father\_edui} + \beta 10 \sim 11\text{Mother\_edui} \\ & + \beta 12\text{Father\_partyi} + \beta 13\text{Mother\_partyi} + \beta 14\text{Father\_industryi} \\ & + \beta 15 \sim 16\text{Family\_incomei} + \beta 17 \sim 19\text{Educationi} + \beta 20 \sim 21\text{Industryi} \\ & + \beta 22 \sim 23\text{Propertyi} + \beta 24 \sim 26\text{Workinghoursi} + \varepsilon i \end{aligned}$$

Based on this, "stepwise regression" is selected as the control for the benchmark OLS regression [7], and backward elimination is used to reduce the overfitting problem of the model. The regression analysis model is established, and the parameters are estimated as shown in column (1).

Table 1. BASELINE OLS REGRESSION AND STEPWISE

Dependent Variable	OLS Returns	Gradual return
	(1)	(2)
Gender	−0.026	
	(−0.20)	
Age	−0.006	
	(−0.38)	
Party	0.423***	0.454***
	(2.89)	(3.98)
Married	−0.327	−0.410***
	(−1.47)	(−2.71)
Urban_area1	−0.285	
	(−1.26)	
Urban_area2	−0.449**	−0.268*
	(−2.36)	(−1.87)
Urban_area3	−0.207	−0.190*
	(−1.43)	(−1.71)
Father_edu1	0.119	
	(0.64)	
Father_edu2	0.025	
	(0.08)	
Mother_edu1	−0.119	
	(−0.68)	
Mother_edu2	0.071	
	(0.19)	
Father_party	0.137	
	(0.88)	
Mother_party	−0.258	
	(−1.14)	
Father_industry	−0.107	
	(−0.53)	

(*continued*)

Table 1. (*continued*)

Dependent Variable	OLS Returns	Gradual return
	(1)	(2)
Family_income1	0.166	
	(1.16)	
Family_income2	0.794***	0.683***
	(3.03)	(3.58)
Education_Bachelor	0.121	
	(0.46)	
Education_Master	0.199	
	(0.63)	
Education_Docter	0.989*	0.957**
	(1.71)	(2.12)
Industry_1	−0.055	
	(−0.30)	
Industry_2	0.436***	0.464***
	(2.71)	(4.66)
Property_G	−0.433***	−0.395***
	(−2.90)	(−3.75)
Property_C	−0.184	
	(−1.24)	
Workinghours_1	−0.337	
	(−1.17)	
Workinghours_2	−0.257	
	(−0.87)	
Workinghours_3	−0.135	
	(−0.40)	
Constant	2.640***	2.195***
	(4.53)	(23.39)
Observations	76	76
R-squared	0.595	0.507

3.2 Analysis of Baseline OLS Regression Results

- Environmental Variables

According to the OLS regression results, the gender factor is not significant in that women have higher annual income, contrary to the established literature that "male labor force

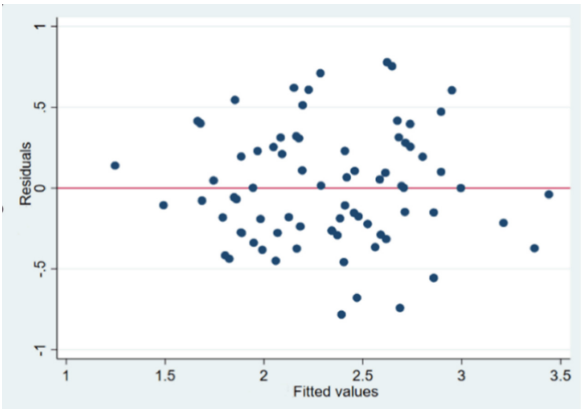

Fig. 1. OLS residual diagram

tends to have higher wages", and there is less gender discrimination among highly educated youth. Also, Age has a negligible effect on wages [8]. Party membership increases annual wages by 42.3% at the 1% confidence level compared to mass or communist membership. The effect of marital status (Married) on income is not significant.

- Effort Variables

In terms of personal effort, according to the regression estimated coefficients, the earnings of those with a master's degree are higher than those of workers with less than a bachelor's degree. Employed workers with a PhD have their earnings increased by 98.9% compared to those employed with less than a bachelor's degree at the 10% level, reflecting the advantage of education for personal development, a finding consistent with established research (Zhou Xing, 2015).

3.3 Heteroskedasticity Test

Since the samples collected from the questionnaire survey are cross-sectional data, there is a higher possibility of heteroskedasticity problem in the model, i.e., heterogeneity in the fluctuation of annual personal income due to the size of the explanatory variables. To address this problem, we first conduct a preliminary analysis by scatter plots of the OLS regression residuals to determine whether there is a trend pattern; and then conduct a White heteroskedasticity test to indicate whether it is consistent with the basic assumption of the random error term, so as to ensure that the regression parameter estimates have good statistical properties (Gauss-Markov property and accurate t-test) (Fig. 1).

The residual scatter plot shows that the variance of the random error term is not regular and can be considered as homoscedasticity. The White heteroskedasticity test was further done and the 2 statistic was obtained, so the original hypothesis was accepted that there was no heteroskedasticity in the model.

Table 2. LASSO REGRESSION

Dependent Variable	Post-lasso regularization	PDS selection and all-variable regression
	(1)	(2)
Party	0.234*	0.244**
	(1.87)	(2.10)
Father_edu2		0.071
		(0.46)
Family_income2	0.595***	0.593***
	(2.75)	(2.97)
Industry_1		−0.280*
		(−1.68)
Industry_2	0.283*	0.282**
	(1.92)	(2.08)
Constant		2.084***
		(18.82)

3.4 Analysis of Stepwise Regression Results

Due to the small number of significant variables in the OLS regression, this paper uses the backward method of stepwise regression (Backward elimination) to compare different models. According to the stepwise regression results in column (2) of Table 1, there are eight variables in the model with strong to weak explanatory power for individual income, in order of "high-yielding industry (Industry-2), party member (Party), state-owned unit (Property-G), middle-class family (Family-income2), married (Married) [9], PhD (Education_Docter), provincial capital city (Urban-area2), and ordinary prefecture-level city (Urban-area3)", and the above variables are significant at least at the 10% level.

The results show that the high-yield industry for compared to the low-yield industry income increase of 46.4%, party membership makes the income higher by 45.4%, while state-owned enterprises and institutions compared to private/private enterprises decreased by 39.5%, doctoral degree makes the labor force income higher by 95.7%, middle-class families children's income is 68.3% higher than low-income families, children born in municipalities directly under the central government income instead lower than county or rural areas, and the mechanism of action of the above studied variables is the same as OLS regression (Table 2).

Lasso regressions were used to obtain 2 screening variables (Selected controls) among the remaining study variables, namely father's higher education and medium-earnings industry [11]. The medium-earnings industry significantly reduces personal income at the 10% level, which may be explained by the wage increase in the low earnings industry due to the transformation and upgrading of the manufacturing industry and the increase in the share of services.

3.5 Lasso Regression

Limited by the data capacity of the questionnaire, 26 independent variables are too many relative to 76 observations, i.e., $p > n$, which can easily lead to overfitting as there is no unique solution for the OLS estimated parameters.

Belloni et al. (2014) proposed that control variables can be selected by applying the Lasso twice, i.e., PDS (post-double-selection) method. In contrast to the OLS method, Lasso adds a penalty term (also known as a "regularization function") to the optimization problem and controls the penalty value by adjusting the parameters to find the one that minimizes the out-of-sample mean squared error MSE.

The results of the OLS and stepwise regressions are combined, and "Party membership (Party), middle class or higher household income (Family-income2), and high yield industry (Industry-2)" are used as endogenous variables to explore the selection of the remaining 23 variables through Lasso regression. The results are as follows.

4 Conclusion

The regression results find that a labor force with party membership, from a middle-class or higher family, with the highest education, and in a high-yielding industry will earn more generously, a finding consistent with the established literature in the field of income determination. The remaining explanatory variables in the model are not as expected due to the questionnaire respondent group.

The characteristic of "highly educated youth" makes the regression equation results highly specific, for example, in terms of educational attainment, only a PhD has a significant contribution to higher earnings. If the sample selection bias is resolved, the all-age, all-education labor force data, individual education, parents' education level, parents' occupation type, family residence, workplace, and work hours may have significant effects on earnings.

Of course, the sample data obtained in this paper fully demonstrate how the earnings of today's highly educated youth are determined: less constrained by "inequality of opportunity" factors such as gender, household and parental education, and parental occupation type, and more focused on the individual's entry industry, unit type, and education level through The "inequality of effort" provides significant life-changing opportunities.

References

1. Gong F, Li Z, Lei X et al (2017) The impact of effort on inequality of opportunity measurement and comparison. Econ Res 03(382):78–92
2. Jing L (2019) Does intergenerational mobility of education level affect income disparity - based on the perspective of inequality of opportunity. Econ Res Ref 000(003):56–64
3. Wu G (2013) Inequality of educational opportunities among urban and rural residents in China and its evolution (1978–2008). China Soc Sci
4. Zhou X, Zhang P et al (2015) Intergenerational occupational mobility and income mobility: an empirical study from urban and rural households in China. Econometrics (Quarterly) 01(v.14; No.55):355–376

5. Arawatari R, Ono T (2013) Inequality, mobility and redistributive politic. J Econ Theor
6. Asadullah MN, Yalonetzky G (2012) Inequality of educational opportunity in India: changes over time and across states. World Dev 40(6):1151–1163
7. Ferreira F, Gignoux J (2011) The measurement of inequality of opportunity: theory and an application to Latin America. Rev Income Wealth 57
8. Fleurbaey M, Peragine V (2013) Ex ante versus ex post equality of opportunity. Economica 80
9. Marrero GA, Rodríguez JG (2013) Inequality of opportunity and growth. J Dev Econ 104:107–122
10. Roemer JE (2002) Equality of opportunity: a progress report. Soc Choice Welf 19(2):455–471
11. Jacobs IS, Bean CP (1963) Fine particles, thin films and exchange anisotropy. In: Rado GT, Suhl H (eds) Magnetism, vol III. Academic, New York, pp 271–350