Sources of Income Inequality in China: Individual's Effort or Circumstances?

Dongjie Wu, Prasada Rao, Kam Ki Tang, and Pravin Trivedi (University Of Queensland, Australia)

Paper prepared for the 34th IARIW General Conference

Dresden, Germany, August 21-27, 2016

Session 2B: Economic Change and Insecurity

Time: Monday, August 22, 2016 [Afternoon]

Sources of Income Inequality in China: Individual Effort or Circumstances?

Dongjie Wu, Prasada Rao, Kam Ki Tang and Pravin Trivedi
School of Economics
The University of Queensland

August 16, 2016

Paper prepared for presentation in Session 2B:Economic Change and Insecurity at the IARIW 34th General Conference.

Contents

List of Figures

List of Tables

Abstract

Our study examines the extent to which the dynamics of income inequality in contemporary China is driven by individual's effort and circumstances at the national and regional levels. The framework we use is inequality of opportunity. Under this framework, the factors contributing to income inequality are categorized into either circumstances or effort. No existing Chinese datasets, however, has sufficient information for examining the contribution of inequality of opportunity to the variation of inequality over time. To circumvent this issue, we make use of information in cross-regional variation in development and inequality. Our data come from the China Family Panel Study, which contains 33,600 individual observations for years 2010 and 2012. Our empirical analysis allows zero-income observations using a hurdle model, parameterizes and estimates the heteroskedasticity —the indirect effect of circumstances— through maximum likelihood estimation (MLE), and implements the Shapley decomposition to identify the contribution of each circumstance and effort to income inequality respectively. We find that at the national level, inequality of opportunity accounts for around 31% in 2010 and 43% in 2012, a higher figure than U.S. and most Latin American countries. At the regional level, as we move from low-income regions to high-income ones, inequality of effort decreases significantly while the level of inequality of opportunity increases slightly, with a net effect of small total income inequality.

Keywords: Inequality of Opportunity, China, Development.

1 Introduction

The real per capita income in China grew at an impressive rate in the last two decades, but so was income inequality. Gini coefficient, an indicator of income inequality, rose from under 0.3 before 1980 to 0.55 in 2012 (Xie and Zhou, 2014); it is higher than US (0.41) (The World Bank, 2016a) but similar to some Latin American countries such as Brazil (0.53) and Colombia (0.54). The increase in inequality was not due to a fall in the income levels of the poor [1], but due to more rapid income growth of the rich (Li et al., 2013). This finding raises questions about the change of income distribution in China and its sources. In particular, what is the main source of the divergence in income growth between the poor and the rich?

The public is also aware of high income inequality in China. The International Social Survey Program(ISSP) (2009) surveyed perceptions of economic inequality in 38 countries in 2009. Most Chinese respondents tended to agree with the statement: "Income differences in China are too high" and the conceding rate is on par with other 37 countries. More importantly, Chinese respondents have the lowest "feeling of procedural justice" (Larsen, 2016) among all respondents in the ISSP survey[2]. Most respondents strongly believed that socio-political connections and parents' socio-economic backgrounds were important for getting ahead in society.

Many researchers have studied income inequality and its determinants in China. They found that income inequality rose with regional disparities (John Knight, 1993, Wan and Zhou, 2005), globalization (Wan et al., 2006), migration (Park and Wang, 2010) and private ownership of assets (Li et al., 2013). These studies, however, shed little light on the findings of the ISSP survey because some determinants such as globalization might not be relevant to "procedural justice" and for those relevant determinants, cross-country comparisons are impossible without a summary measure.

In this paper, we try to fill this knowledge gap using the theory of equal opportunity (Roemer, 2000, Cohen, 1989, Arneson, 1989), in which society should only concern with inequality due to factors beyond individuals' responsibility ("circumstances") and acknowledge inequality due to factors within individuals' responsibility ("effort"). Inequality caused by circumstances is defined as "inequality of opportunity" (IOP). If China has a higher IOP than other countries, it may explain the public perception of poor "procedural justice" in the country. Implementing the theory of equal opportunity first requires a working definition of individual responsibilities or circumstances. In this paper, we define *observed* factors such as

[1] Quite the opposite, the poverty rate decreased from 85% to lower than 11% during 1980-2012 (The World Bank, 2016b).

[2] ISSP asked respondents to what extent do "coming from a wealthy family", "having well-educated parents", "knowing the right people", "having political connections" and "giving bribes" are important to get ahead in society. (Larsen, 2016) combined these questions into a measure of perceptions on "procedural justice". This measure captures to what extent people need privileges to get ahead in society.

gender, ethnicity and parents' socioeconomic status as "circumstances" and effort as an *unobserved* factor. Furthermore, due to the very short time frame of our dataset, the circumstances variables are treated as *time invariant*, while the effort variable is likely to be time variant.

Equality of opportunity was first conceptualized by John Rawls, who argued that "offices and positions must be open to everyone under conditions of fair equality of opportunity" (Rawls, 1971, p. 302). Based on Rawls's argument, Roemer (2000) proposed a framework to measure IOP. He stands with a classification of people based on types and tranches: those share same circumstances belong to the same type and those exert the same level of effort share the same tranche. IOP can then be measured *ex-post* or *ex-ante*. The ex-post IOP captures the within-tranche inequality, the inequality of a counterfactual income distribution where everyone has the same average tranche (Checchi et al., 2010). Therefore, ex-post IOP is driven entirely by differences in circumstance across the population, conditional on the average level efforts. On the contrary, the ex-ante IOP is the between-type inequality, the inequality of a counterfactual income distribution where everyone in a type has the same type-average income (See Van De Gaer, 1993 and Checchi et al., 2010). Since the ex-post approach demands more data, we use the ex-ante approach in the current study.

Roemer's framework has widely been applied in empirical researches. de Barros et al. (2009) estimated the ex-ante IOP in seven Latin American countries [3]. They found that although Mexico has the highest overall income inequality, the contribution of IOP (20.8%) is the smallest among the seven countries. The biggest share of IOP (37.3%) belong to Guatemala. Using longitudinal data, Pistolesi (2009) showed that rising income inequality in the U.S. during 1968-2001 was not driven by increases in IOP. In fact, they found that IOP in the U.S. has decreased from 43% to 20% over the period. Björklund et al. (2011) differ from many other studies by including individual IQ and body mass index as circumstances in a Swedish study and used the Shapley decomposition to decompose the effects of circumstances. More importantly, they measured the type heterogeneity of effort — the indirect effect of circumstances to income inequality. They found that the share of IOP to total income inequality was less than 30%.

Some studies went beyond measuring IOP and examined the impact of IOP on development. Using data from 42 countries, Ferreira et al. (2014) found IOP to have a negative growth effect but the result is neither conclusive nor robust. In Marrero and Rodrguez (2013), IOP has a negative growth effect in rich countries only, while both IOP and inequality of effort [4] enhance growth in poor countries. Lastly, because these studies use different IOP approaches, inequality measures (e.g. Gini and Theil index), and definitions of circumstances, one should be cautious about

[3]The seven countries are: Brazil, Colombia, Ecuador, Guatemala, Panama, Peru and Mexico

[4]The counterfactual inequality after filtering out the effect of circumstances.

comparing their findings.

For China, Zhang and Eriksson (2010) estimated IOP moved broadly in sync with overall income inequality during 1989 to 2006, with its share of overall income inequality ranged from 46% to 65%. They also found that the IOP was largely due to parental socio-economic status. However, due to lack of information about parents' socio-economic circumstances, most of their estimations were restricted to urban population or state owned enterprise workers which were mostly urban based. Therefore, the study omitted the rural population, which accounted for 55% to 74% of the population during the sample period.[5] In addition, the study measured IOP using Gini coefficient but did not correct for the bias caused by the coefficients *path-dependency* property (Foster and Shneyerov, 2000).

In this paper, we use a representative dataset drawn from the China Family Panel Study (CFPS) which contains 33,600 individual observations for years 2010 and 2012. We measured IOP at the national, the regional and the provincial levels respectively. The data are classified into eight regions for the regional measures and 25 provinces for the provincial measures. In addition, we collected the regional and provincial gross regional product (GRP) from China Statistical Yearbooks (NBS, 2013) to find the relationship between GRP and IOP. Since the data span over three years only, the emphasis is on the cross-regional variation in IOP. Although China as a whole is growing rapidly, the level and change of development and inequality differ vastly across various Chinese regions. The underdeveloped northwest China has relatively high income inequality, while the highly developed southeast has relatively low income inequality. Therefore, a cross-regional comparison in inequality can be used as a vehicle to assess how inequality might change with development and how IOP contributes to that change.

Since the data include samples with no income, we estimated the probability of earning positive income through a hurdle model. In addition, Björklund et al., 2011 showed that type heterogeneity of effort is a source of IOP. They used a non-parametric approach to measure this heterogeneous effect. However, large numbers of types restrict each type with few samples and biased the measure. As a result, this approach only showed the effect of heterogeneity as a whole but failed to reveal the effect of each circumstance. Instead, we took a parametric approach and used maximum-likelihood estimation (MLE) to show the effect of each circumstance variable on the heteroskedasticity.

To better understand the roles of circumstances and effort in driving inequality in China, we conduct two decompositions. First, we follow Björklund et al. (2011) and apply the Shapley decomposition (Shorrocks, 2013) to identify the contributions of circumstances and effort to income inequality. This decomposition technique allows us to use a common inequality index —Gini coefficient— without breaking

[5]China's rural population share has been declining steadily over time. Source of data: World Development Indicators

path dependency (Foster and Shneyerov, 2000).

Second, we apply the Oaxaca decomposition (Oaxaca, 1973) to identify the differential effects of circumstances on income across the advantaged and disadvantaged groups. The IOP measure only provides an overview of the unfair part of inequality, while the Oaxaca decomposition can reveal whether the higher income for the advantaged group is due to their better circumstances or bigger influence of their circumstances on income.

The main contribution of this paper is to evaluate IOP in China at both national and regional levels using a representative, cross-sectional data. Taking advantage of the heterogeneity of Chinese regions, this study sheds light on the contribution of IOP to overall income inequality over various development stages. Moreover, this study is the first one to apply the hurdle model with the Shapley decomposition to include those who receive no income and to show the heterogeneous effect of each circumstance.

We found that at the national level, circumstances account for around 31% of the income inequality in China in 2010 and 43% in 2012. The figures rise around 25% if we include heteroskedasticity between types as parts of IOP. GRP appears to have a negative relationship with income inequality and inequality of effort at the provincial level, but no discernible relationship with the level of IOP. As a result, the share of IOP in the overall inequality rises with the increase of GRP. Lastly, the results from the Oaxaca decomposition showed that getting rich does not require better circumstances per se but the bigger influence of circumstances to income. In addition, the shares of IOPs in the overall inequality are similar across regions.

The rest of this paper is organized as follows. In section 2 we describe the approach to measuring inequality of opportunity. In section 3 we discuss the empirical strategies. Section 4 is the description of data. Section 5 shows the empirical results and section 6 is the conclusion.

2 Measure Inequality of Opportunity

To measure IOP, we followed the approaches introduced by Checchi and Peragine (2010). We first partitioned an income profile into types and tranches. Assume that individuals' income y is determined by a finite set of exogenous and time-invariant circumstances $\mathbf{c}$ and one-dimensional effort e.

$$y = g(\mathbf{c}, e) \tag{1}$$

where $\mathbf{c}$ is a set of variables concerned as circumstances with n finite values. e stands for effort with m different levels. Assume that $C = \{1, \ldots, n\}$ is the set of all types and $E = \{1, \ldots, m\}$ is the set of all tranches so that individuals with the same $\mathbf{c}$ are partitioned into the same type and those with the same e into the same tranche. For $i \in C$ and $j \in E$, we denoted y_i^j as the individual's income given type

i and tranche j. The set of all individual income can be represented by an *income matrix* Y:

$$Y = \begin{bmatrix} y_1^1 & \dots & y_1^j & \dots & y_1^m \\ \vdots & & \vdots & & \vdots \\ y_i^1 & \dots & y_i^j & \dots & y_i^m \\ \vdots & & \vdots & & \vdots \\ y_n^1 & \dots & y_n^j & \dots & y_n^m \end{bmatrix} \in \mathcal{Y}$$

where $\mathcal{Y}$ is the set of all possible individual income matrixes.

An alternative way is to partition them only into types or tranches. We denote $Y = \{Y_1, \dots, Y_n\}$ where $Y_i \in Y$ is the income distribution in type i and $Y = \{Y^1, \dots, Y^m\}$ where $Y^j \in Y$ is the income distribution in tranche j. Let $\mu(Y_i)$ be the average outcome in type i and $\mu(Y^j)$ be the average outcome in tranche j. Since Ex-ante IOP does not require information on effort, we assume that the effort is unobserved. This model also excludes the existence of random components or luck (Lefranc et al., 2008) and interaction between circumstances and effort. Therefore, the following two basic assumptions should be satisfied given the non-observability of effort(Checchi and Peragine, 2010):

Assumption 1. *Function g is monotonically increasing in effort e.*

Assumption 2. *The conditional distribution of effort e is independent of circumstance* **c**

The first assumption indicates that the more effort one exerts, the more income one earns, and the second assumption implies the independence between effort and circumstance. If equation (1) satisfies both assumptions, one can directly measure ex-ante IOP by computing the inequality of a counterfactual income distribution in which the contribution of effort has been eliminated (Ramos and Van de gaer, 2012). We define this counterfactual income distribution as Y_c. An alternative approach —the indirect measure— is to estimate a counterfactual income distribution Y_e by ruling out the contribution of circumstances.

To estimate the counterfactual income distribution, one can decompose the observed income distribution into two — a smoothed between-type distribution, replacing the within-type income with a type-average income; and a standardized within-type distribution that eliminates the difference in the type-average income (Foster and Shneyerov, 2000). Thus, Y_c and Y_e can be denoted as the following vectors:

$$Y_c = \{\mu(Y_1)\mathbf{1}_{N_1}, \dots, \mu(Y_i)\mathbf{1}_{N_i}, \dots, \mu(Y_n)\mathbf{1}_{N_n}\} \tag{2}$$

$$Y_e = \{\tilde{Y_1}, \dots, \tilde{Y_i}, \dots, \tilde{Y_n}\} \tag{3}$$

where $\mathbf{1}_{N_i}$ is the unit vector of length equal to type i's population and $\tilde{Y_i} = \frac{\mu(Y)}{\mu(Y_i)} Y_i$.

To compute the type-average income $\mu(Y_i)$, we relied on a parametric model. Suppose e is unobserved, $y = f(\mathbf{c}) + v$ where v is the error term. $\mu(Y_i)$ is the predicted value when circumstances $\mathbf{c}_i$ corresponds to type i: $\mu(Y_i) = \bar{y_i} = f(\mathbf{c}_i)$.

Therefore, IOP can be measured from Y_c. In this paper, we used two indexes introduced by Ferreira and Gignoux (2008): one for the absolute level of IOP — Inequality of Opportunity Level (IOL) and the other for the share of IOP relative to total income inequality— Inequality of Opportunity Ratio (IOR). The former index is given by:

$$IOL = I(Y_c) \tag{4}$$

where the function $I : \mathbb{R}_+^N \mapsto \mathbb{R}_+$ is an inequality index such as variance, Theil index and Gini coefficients.

The latter index is given by:

$$IOR = \frac{I(Y_c)}{I(Y)} \tag{5}$$

In addition, we defined the inequality of Y_e as the level of inequality of effort (IOE):

$$IOE = I(Y_e) \tag{6}$$

To let IOL be consistent with IOE, we require the sum of both indexes to be the total income inequality:

$$I(Y) = I(Y_c) + I(Y_e) \tag{7}$$

However, Equation (7) holds only if the inequality index $I()$ is path independent[6]. One path independent inequality measure is the mean log deviation (MLD). An alternative is to apply the Shapley decomposition to any inequality index. In this paper, we use the Shapley decomposition with the Gini coefficients because it provides flexibility to choose different inequality indexes and to decompose IOP into each circumstance. We introduce this decomposition technique with other empirical methods in the next section.

3 Empirical Strategy

In this section, we introduce the econometric methodology. To account for zero-incomes for some individuals, we use the lognormal hurdle model (Cragg, 1971). Applying this model, we identify the expected income for each type. However, the within-type income distributions might be heteroskedastic. This heteroskedasticity might imply indirect effects of circumstances on income inequality. To deal with this issue, we used MLE to identify the effects of each circumstance on the mean and

[6] *Path independence* holds when over inequality is the sum of between-group inequality and within-group inequality (Foster and Shneyerov, 2000)

the variance. Implementing the Shapley decomposition, we measure IOL and IOR not only for all circumstances but for each as well. The last part of this section is the Oaxaca decomposition. It is used to examine whether the income gap between types is due to different levels of circumstances or different effects of circumstances on income.

3.1 The Lognormal Hurdle Model

The lognormal hurdle model (Cragg, 1971) takes the zero-income observations into consideration. The model consists of a binary outcome model, which is used to account for the zero-versus-positive income, and a non-linear model, which deals with the positive income.

Assume that income y can be generated as

$$y = wy^* \tag{8}$$

where w is a binary variable that equals to 1 if $y > 0$, and y^* is a continuous variable that equals to y but it is observed only when $w = 1$. y^* is assumed to have a lognormal distribution:

$$y^* = \exp(\mathbf{c}'\boldsymbol{\beta} + u) \tag{9}$$

where $\mathbf{c}$ stands for circumstances and the error term $u|\mathbf{c} \sim Normal(0, \sigma^2)$. So the expectation of y^* given $\mathbf{c}$ is:

$$E(y^*|\mathbf{c}) = \exp(\mathbf{c}'\beta + \frac{\sigma^2}{2}) \tag{10}$$

To estimate the probability of receiving positive income w, we use the logistic model:

$$Pr(w = 1|\mathbf{c}) = \Lambda(\mathbf{c}'\boldsymbol{\gamma}) \tag{11}$$

where Λ is the logistic function and $\boldsymbol{\gamma}$ is the vector of coefficients for the circumstance variables in the logistic model.

Therefore, the between-type income distribution can be represented by the expectation conditional on circumstance variables:

$$y_c = E(y|\mathbf{c}) = \Lambda(\mathbf{c}'\boldsymbol{\gamma}) \times \exp(\mathbf{c}'\boldsymbol{\beta} + \sigma^2/2) \tag{12}$$

where y_c represents the expected income in each type.

Since the effort cannot be observed, we computed the within-type income distribution by rescaling the observed income until the income distribution in each type has the same mean as the overall income distribution:

$$y_e = \frac{y * \bar{y}}{y_c} \tag{13}$$

An alternative approach is to treat the residual of the model as income earned by effort. We didn't use this approach because the residual would sometimes be negative, which might affect the computations of Gini index (Chen et al., 1982).

To apply equation (12), we undertook the estimation into three steps. First, we estimated the binary part using a logistic regression and got the estimator $\hat{\boldsymbol{\gamma}}$. Second, we estimated the continuous part using a log-linear regression and got the estimator $\hat{\boldsymbol{\beta}}$. The last step is to estimate the predicted income $\hat{y}$. Since y is assumed as a log-normal distribution and the true distribution σ^2 could be unknown, we used Duan's (1983) *smearing estimate*.

If u is independent of $\mathbf{c}$, $E(y^*|\mathbf{c}) = E[\exp(u)]\exp(\mathbf{c}\boldsymbol{\beta})$. Let $\tau = E[\exp(u)]$, We can use the estimated smearing factor $\hat{\tau}$ to estimate τ. $\hat{\tau}$ is:

$$\hat{\tau} = N^{-1}\sum_{i}\exp(\hat{u}_i) \tag{14}$$

where $\hat{u}_i$ is the residual of the log-linear regression.

3.2 Type Heterogeneity of Effort

It is possible that the error term u is heteroskedastic (Björklund et al., 2011). In other words, the within-group income distribution is not identical between groups. It might be due to the correlation between circumstances and effort. To include the heteroskedasticity in the measures of inequality of opportunity, we estimated it using maximum-likelihood estimation(MLE). We specified the skedasticity function of income with respect to circumstances $\mathbf{c}$:

$$\sigma_i^2 = \exp(\mathbf{c}_i'\theta) \tag{15}$$

Under the assumption of normal distribution of income, the likelihood function is:

$$f(y|\mathbf{c}) = \left(\frac{1}{\sqrt{2\pi\exp(\mathbf{c}_i'\theta)}}\right)^{n/2} \times \exp\left[-\sum_{i=1}^{n}\frac{(y-\mathbf{c}_i'\beta)^2}{2\exp(\mathbf{c}_i'\theta)}\right] \tag{16}$$

Using MLE, we estimated both β and θ. After identifying γ using the hurdle model, we computed Y_c and Y_e using Equation (12) and Equation (13). The estimators of MLE allow us to further standardized Y_e with respect to its variance:

$$\tilde{Y}_e = Y_e \times \sqrt{\frac{Var(Y_e)}{\hat{\sigma_i^2}}} \tag{17}$$

where $\tilde{Y}_e$ is the homogenized effort which is independent of circumstances and $\hat{\sigma}_i^2$ is the estimator of σ^2 in Equation 15.

3.3 The Shapley Decomposition

Shapley (1952) introduced the Shapley decomposition to solve cooperative games in game theory. Suppose a game has a set of $N = \{1, \ldots, n\}$ players with a characteristic function $v : 2^N \to \mathbb{R}$, mapping all possible coalitions of players to the gains of relative coalitions, the amount player $i \in N$ gaining from the coalitional game can be measured by the Shapley value of i: $\phi_i(v)$.

Similar to the cooperative games, we can decompose income inequality by assuming a set of factors X_k indexed by $K = \{1, \ldots, k, \ldots, m\}$ with a characteristic function (the inequality index) $I : 2^K \to \mathbb{R}$. The set of factors can include both circumstances such as gender, ethnicity and parents' socioeconomic status, and effort. Based on the Shapley decomposition, the factor k's contribution is determined by the Shapley value of k: $\phi_k(I)$ that can be calculated using the following equation:

$$\phi_k(I) = \sum_{S \subseteq K \setminus \{k\}} \frac{|S|!(m - |S| - 1)!}{m!} (I(S \cup \{k\}) - I(S)) \tag{18}$$

where S is the subset of K without k and $|S|$ is the number of factors in S. In this equation. $I(S \cup k) - I(S)$ is the marginal contribution of the factor k to total inequality and $\phi_k(I)$ can be interpreted as the average marginal contribution of all possible permutations in which factor k affects inequality jointly with other factors in the set S.

For a factor in set S, we use the observed value of this factor; otherwise, we take the average of the observed value so that this factor has no effect on inequality.

To measure the Shapley value of each factor, we made use of the alternative equation of Equation (18) and implemented the procedure into three steps.

The alternative equation is:

$$\phi_k(I) = \frac{1}{|K|!} \sum_R [I(P_k^R \cup \{k\}) - I(P_k^R)] \tag{19}$$

where R is an element set from the set $\mathcal{R}$ that contains all permutations of the set K and P_k^R is the subset of R in which all elements precede k in the order R[7]. Additionally, we let $C_k^R = I(P_k^R \cup \{k\}) - I(P_k^R)$ be the marginal contribution of k to income inequality given R. This marginal contribution captures the effect of the factor k to income inequality given the influence of some factors other than k.

Step 1: Choose a factor k and generate all the possible order

The first step is to choose a factor k[8] for measuring its Shapley value. We let that factor be its observed value when measuring $I(S \cup \{k\})$ and fixed to its average

[7]We implemented Equation 19 by the following procedure in the programming language —R (Hofmarcher, 2015).

[8]A factor k can be either a circumstance, a group of circumstances (circumstances in the same group are either observed or fixed to the average simultaneously) or effort

when measuring $I(S)$. Accordingly, we can generate the set of all possible orders of factors K: $\mathcal{R} = \{R \in \mathcal{R}\}$.

Step 2: Compute the marginal contribution C_k^R for all possible order R

In step 2, we pick up one order $R \in \mathcal{R}$. Since factors such as circumstances and effort affect income inequality through income distribution, the marginal contribution of the factor k can be represented by the difference of income inequalities between two income distributions: one given the factor k is observed and the other given k is fixed to its average.

On our empirical implementation, we make use of the lognormal hurdle model to compute C_k^R. Specifically, we estimate three models and computed IOL based on these models respectively. The first model only considere the lognormal model with the positive income (Equation (9)). The second model uses the hurdle model assuming the homogenous effect across type (Equation (12)). The last model additionally includes heteroskedasticity of distribution (Equation (16)).

For the lognormal model, we estimate the counterfactual income distribution based on Equation (9) and obtain estimators of $\hat{\boldsymbol{\beta}}$. Since we assume that effort is unobserved variables in the model, we use Equation (13) to compute the standardized within-type income distribution. To compute $I(P_k^R \cup \{k\})$, we fix the variables which go after k respectively to their average values. To compute $I(P_k^R)$, we additionally fix the factor k to the average level.

For the hurdle model with homoskedasticity, we use the same equations as the lognormal model except that using Equation (12), we include another factor —the probability to have positive income. When this factor is not in P_k^R, the probability is set fixed to the average predicted probability; otherwise, the predicted probability is used.

For the hurdle model with the heteroskedasticity, we estimated variances for each type based on the MLE (Equation (16)) and used the estimators of variances to further standardize Y_e (Equation (17)) so that the effort after standardization is totally independent of circumstances.

We repeat this step until all marginal contributions (C_k^R for all $R \in \mathcal{R}$) are computed.[9]

Step 3: Take the average of all C_k^R computed in step 2

After step 2, we take the average of all C_k^R computed with respect to all possible R. The result is the Shapley value in Equation (19). By implementing this procedure, we are able to compute the contribution of each circumstance and the contribution of effort.

One advantage of the Shapley decomposition is that the sum of the Shapley

[9]If the number of circumstance variables is 5, including the effort, the number of factor is 6. So the number of all possible order is $6! = 720$, which means step 2 is repeated by 720 times until all marginal contributions are computed.

value of each factor is the total contribution of these factors to income inequality. Israeli (2007) showed that the total contributions of the explanatory variables in a simple linear regression is its R-Square if the variance is used as an inequality measurement. In our study, Gini coefficients are used instead of the variance and the effort affected by the residuals of the model is taken into account. Therefore, the sum of the Shapley value for each factor measured by Gini coefficients are equal to total income inequality:

$$I(y) = \sum_k \phi_k(I) \tag{20}$$

3.4 The Oaxaca Decomposition

If population is divided into two groups (e.g. female and male, urban and rural, minority and majority group or under-developed and developed region), circumstances may have different effects on income for each group. To study the group differences, we employ the Oaxaca decomposition (Oaxaca, 1973).

Consider two groups, A and B and income distribution for each group are denoted as Y_A and Y_B, the mean difference between group is:

$$R = E(Y_A) - E(Y_B) \tag{21}$$

where E() is the expected value of income distribution.

We decomposed the between-group difference to three components (Jann et al., 2008):

$$R = EN + CO + INT \tag{22}$$

where EN is the "endowments effect", CO is the contribution of differences in the coefficients and INT is the interaction effect of the former two.

In our model, we assumed that only circumstances can be observed. We specified the model to the following equation:

$$\ln Y = \mathbf{c}'\boldsymbol{\beta} + \epsilon \tag{23}$$

where $\boldsymbol{\beta}$ is the vector of coefficients and ϵ is the error term.

Using this model, the mean difference R becomes

$$R = E(\mathbf{c}_A)'\boldsymbol{\beta}_A - E(\mathbf{c}_B)'\boldsymbol{\beta}_B \tag{24}$$

The first component EN,

$$EN = \{E(\mathbf{c}_A) - E(\mathbf{c}_B)\}'\beta_B \tag{25}$$

captures the group differences in the predictors, i.e. whether the difference in income between groups is due to the difference in circumstances between groups.

The second component CO,

$$CO = E(\mathbf{c}_B)'(\beta_A - \beta_B) \tag{26}$$

is the difference in the contribution of coefficients. The level of contribution indicates the amount of inequality between groups coming from the effect of circumstances.

The third component INT,

$$INT = \{E(\mathbf{c}_A) - E(\mathbf{c}_B)\}'(\beta_A - \beta_B) \tag{27}$$

is the interaction accounting for both the differences in endowments and coefficients.

4 Data Description

To measure inequality of opportunity in China, we used data from the China Family Panel Studies (CFPS). CFPS is a nationally representative annual longitudinal survey containing not only individual-level data but also household- and community-level data. It has been conducted since 2010 by the Institute of Social Science Survey (ISSS) of Peking University, China. Until 2016, this project has published its surveys for two years— the 2010 baseline survey and the 2012 follow-up survey. Since the survey conducted in 2011 is a maintenance survey and the sample size is small relative to 2010 and 2012, we do not include the 2011 survey into our research.

CFPS covers 16,000 households with more than 33,000 adults and 8,900 youths in 25 provinces/municipalities/autonomous regions in China. It is designed to record changes in the socioeconomic well-being of Chinese people, covering a variety of topics such as economic activities, educational attainment, family relationships and dynamics, migration, and physical and mental health. The design of CFPS was inspired by the authoritative panel study in other countries such as the Panel Study of Income Dynamics (PSID) so that international comparison can be conducted. The original sample sizes in 2010 and 2012 are 33,600 and 35,720 respectively in which 26,393 samples have records in both 2010 and 2012. Of these we focused on individuals between age 21 to 60 because the labour participation rate outside this age range is relatively low. After filtering, the sample size was reduced to 19,736.

Table 1 and 2 present summary statistics of variables we used in the study. Male respondents make up 47% of the sample. Ethnicity is represented by the dummy variable "minority". It is equal to 0 if an individual's ethnicity is the majority group —Han; otherwise, it is equal to 1. The percentage of the minority group is around 8%. In addition, the average age of the respondents is 42.25. 90% of them are married. The percentage of members of the Chinese Communist Party (CCP) are 6% and 7% in 2010 and 2012 respectively.

We also include the number of siblings as one of the circumstance variables. Becker and Lewis (1974) studied the relationship between the number of children and children's outcome such as educational attainment and socioeconomic status. An empirical study conducted in China (Li et al., 2007) also find a negative correlation between the family size and child outcome. In the dataset, the average number of sibling is around 3.

Another circumstance variable we used is regions of residence when the respondent was 12 years old. Two dummy variables were generated as the measurement of regions of residence. One is whether the respondent held a non-agriculture Hukou at 12 years old and the other is whether the respondent lived in coastal provinces at that time. We used these variables because children are unlikely to change these circumstances through their own effort.

Hukou is a system for recording household registration in China. It divides households into agriculture (rural) and non-agriculture (urban) Hukou. The former lives in rural areas and is registered as a rural household and the latter lives in urban areas and is registered as an urban household. Due to the difficulty in changing the Hukou status from agriculture to non-agriculture, lots of rural immigrants hold agriculture Hukou even though they live in urban areas. Individuals normally have the same types of hukou as their parents before they grow up. In our sample, the percentage of individuals who hold non-agriculture (urban) Hukou when they were 12 years old is 15%. We chose Hukou status instead of a place of residence because of the difficulty in changing the Hukou status (Wu and Treiman, 2004).

Coastal provinces are the provinces on the eastern coastline of China. This area is more developed than the inland area. We used a dummy variable to capture whether the respondents who lived in the coastal provinces when they were 12 years old (coastal12). The data show that about 43% of the respondents lived in coastal province when they were 12 years old.

Table 2 shows respondents' *parents' socioeconomic status (SOE) when respondents were 14 years old* including parents' education level, parents' occupation status and parents' political affiliation. For all variables, we only account the higher value between parents.

In terms of parents' education level, it is reported in eight levels in CFPS. We merged them into three levels. (1) Low level: below or equal to junior secondary school; (2) Middle level: high school or vocational school; and (3) High level: above or equal to universities. The change of individuals with low level of education is 64% and the high level makes up 13%.

In terms of parents' occupation, it is divided into 8 big categories including 595 specific occupational codes in CFPS. We regrouped them into three levels: the low level including agricultural workers and workers in manufacture and transportation sectors; the middle level including professionals, clerks, technical staffs and other tertiary sector workers; and the high level including the administrative/management positions, teachers for tertiary education, lawyers and high rank

Table 1: Summary Statistics (Respondents)

Statistic	N	Mean	St. Dev.	Min	Max
Individual income(2010)	19,736	10,575.07	21,520.59	0.00	980,000.00
Individual income(2012)	19,736	14,519.89	30,255.13	0.00	1,804,500.00
Household income per capita(2010)	18,729	9,157.06	15,668.13	1.67	1,000,000.00
Household income per capita(2012)	19,248	12,117.97	16,630.24	0.20	612,700.00
Male	19,736	0.47	0.50	0	1
Minority	19,696	0.08	0.27	0	1
Age	19,736	42.25	10.79	21	60
Urban Hukou at age 12	19,625	0.15	0.35	0	1
Live in Coastal Province at age 12	19,736	0.43	0.50	0	1
Number of Sibling	19,550	3.01	1.88	0	14
Married	19,736	0.90	0.30	0	1
CCP Member in 2010	19,736	0.06	0.24	0	1
CCP Member in 2012	19,736	0.07	0.25	0	1

Note: 1. CCP is the Chinese Communist Party

2. Table was created by stargazer v.5.2 (Hlavac, 2015)

Table 2: Summary Statistics (Respondents' Parents)

Statistic	N	Mean	St. Dev.	Min	Max
Low Occupation	17,309	0.79	0.41	0	1
Mid Occupation	17,309	0.13	0.33	0	1
High Occupation	17,309	0.09	0.28	0	1
CCP member	19,736	0.16	0.37	0	1
Low Education	19,736	0.64	0.48	0	1
Mid Education	19,736	0.22	0.42	0	1
High Education	19,736	0.13	0.34	0	1

Note: 1. All variables refer to characteristics when respondents were 14 years old.

2. All variables only account the higher value within parents.

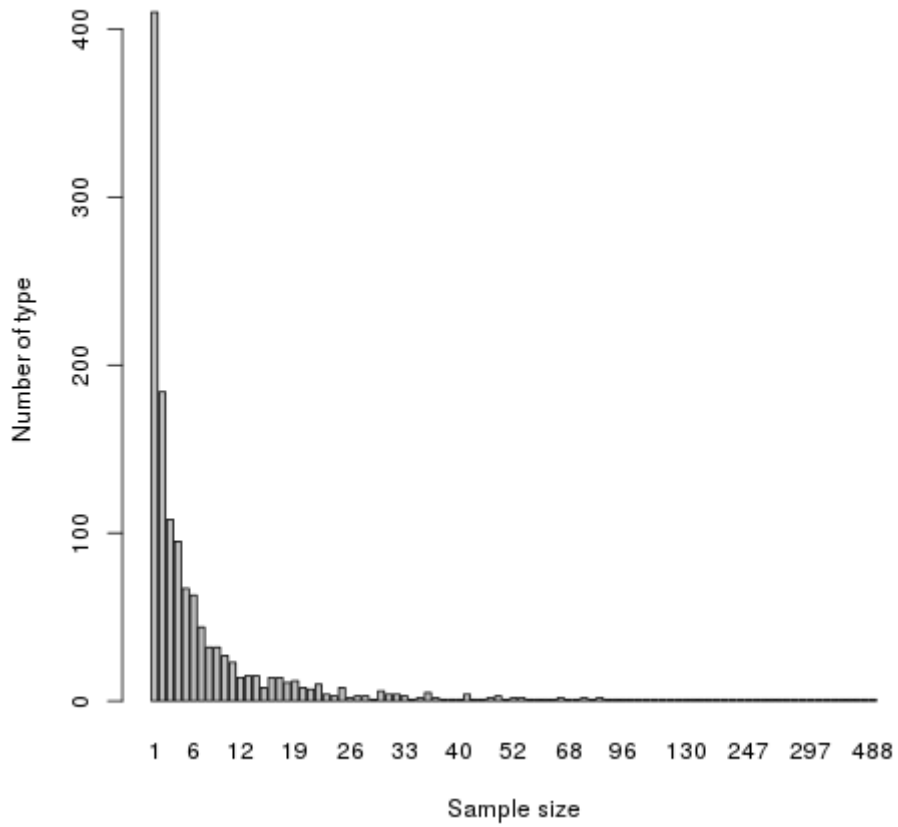
3. Table was created by stargazer v.5.2 (Hlavac, 2015)

military officers. On average, 79% of individuals reported low status of their parents' occupation and 9% reported high status of occupation.

Furthermore, we generated a dummy variable for whether one of the parents is a member of China Communist Party (CCP) when respondents' were 14 years old as a proxy for parents' political affiliation. The percentage of membership of CCP is 16%.

Among the variables we introduced above, we select male, minority, urban hukou at age 12, coastal province at age 12, number of sibling, parents' educational level, parents' occupational level and whether at least one parent is CCP member as circumstances. In total, the data have been divided into 1331 types [10]. Figure 1 is a bar chart showing the number of types for each sample size. Most types have less than 10 samples and more than 400 types even have only 1. Therefore, using non-parametric method to measure inequality of opportunity will result in a large upward bias.

Figure 1: Number of Types for Each Sample Size



For the dependent variables, we use the annual individual income because it is more individually representative than the household income. Household in-

[10]Those types with no observation do not count

come, household consumption and individual labour earnings have also been used in other studies (Ferreira and Gignoux, 2008). The annual individual income is 10,575 Yuan on average in 2010 and 14,519 Yuan in 2012. To construct the income variable, we first computed the labour income by summing up individual wages, awards, allowances, income of working out of town and bonuses. Then we matched each individual to his household's business income (including agricultural and non-agricultural business income), property income, transfer income and other income (including gifts). The individual income is equal to labour income plus income from all sources of non-labour income divided by the family size. The individual income increased by 40% from 2010 to 2012, which is mostly contributed by the increase in the household income per capita.

The household income per capita in 2010 is 9,157 Yuan on average and rises to 12,118 Yuan by 2012. Both the annual individual income and the household income were not adjusted for inflation. This is probably one reason for the dramatic increase in income from 2010 to 2012. Another reason could be the filtering of the sample below 21 and above 60. If we add the excluded sample, the increase of household income per capita is around 20%.

In addition, we found that around 6.8% and 8.1% of the respondents received no income in 2010 and 2012 respectively. Table 3 shows the independent t-test between zero-income and positive income respondents. The coefficients are the mean difference between the zero-income group and the positive income group. A positive coefficient suggests that the zero-income group are more likely to have individuals with a higher value of the relative variable. In terms of the significance, "Male", "coastal province" and "high parents' occupation" are significant in both years. "Urban Hukou status", "number of sibling" are significant in 2010, and "minority", "mid parents' education" and "mid parents' occupation" are significant in 2012. From the sign of coefficients of these variables, the zero-income individual is more likely to be a female, Han ethnicity group(majority), living in coastal provinces with a urban-Hukou status and whose parents have a higher socio-economic status but few sibling.

To measure inequality of opportunity at the regional level, we divided the whole dataset into 8 regions. Figure 2 shows the division. 25 out of 31 provinces in China are included in the dataset, which is coloured in the figure. 8 Different regions are represented by 8 colours. Generally, we grouped the regions by the geographic distance expect the red region which represents metropolitan cities.

The gross regional product (GRP) and the growth rate per capita for each province are shown in table 4. Generally, the east and metropolitan have higher per capita GRP comparing with the rest regions; while the west and north west have higher growth rates comparing with others.

To apply the Oaxaca decomposition, we divided the sample based on individual income, the growth rate and the level of GRP at the provincial level. We treated

Table 3: Zero Income Vs Positive Income (the Independent t-test)

	(1) 2010		(2) 2012	
Male	-0.121***	(-7.95)	-0.123***	(-8.69)
Minority	-0.00141	(-0.17)	-0.0255***	(-3.35)
Coastal province at age 12	0.0729***	(4.82)	0.0989***	(7.00)
Urban Hukou status at age 12	0.0750***	(7.05)	0.0164	(1.64)
Mid education(Parents)	0.0194	(1.51)	0.0405***	(3.36)
High education(Parents)	0.0132	(1.49)	-0.00462	(-0.56)
Mid occupation(Parents)	0.00700	(0.69)	0.0169*	(1.78)
High occupation(Parents)	0.0175**	(2.07)	0.0168**	(2.11)
Member of CCP(Parents)	0.00556	(0.48)	-0.00429	(-0.40)
Number of sibling	-0.165**	(-2.88)	-0.0666	(-1.24)

[1] *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
[2] The standard error is in the parentheses.
[3] The coefficients represent the mean difference between the zero-income group and the positive-income group.

Figure 2: The Regional Division of China

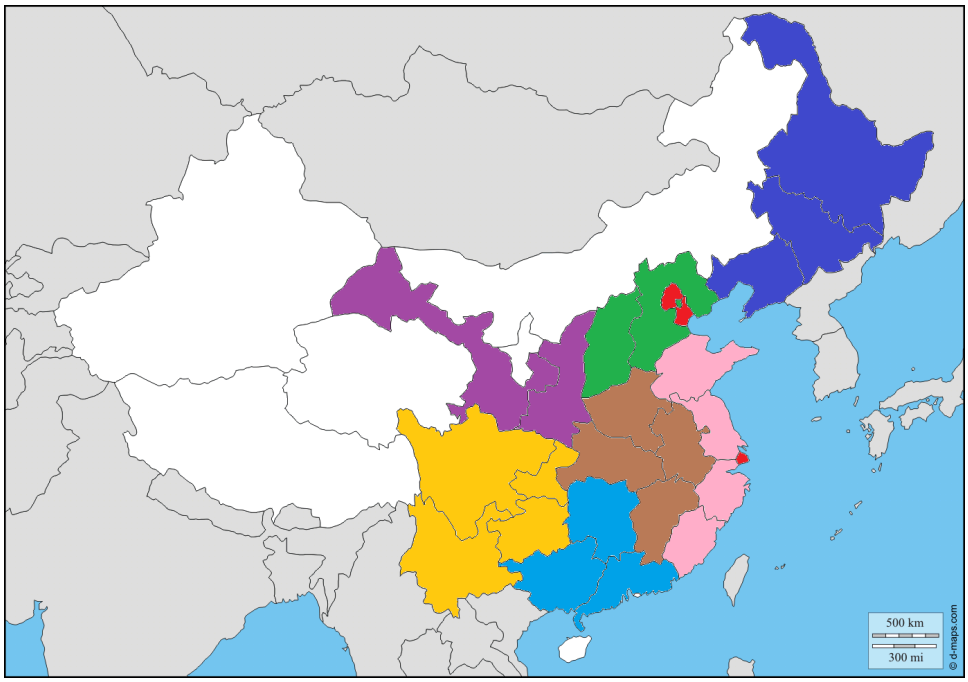


Table 4: Per Capita Gross Regional Product and Indices

		Per Capita GRP(Yuan)			Indices (preceding year=100)		
Province	Region	2010	2011	2012	2010	2011	2012
Fujian	East	40025	47377	52763	113.2	111.6	110.5
Jiangsu	East	52840	62290	68347	112	110.3	109.7
Shandong	East	41106	47335	51768	111.3	109.9	109.2
Zhejiang	East	51711	59249	63374	109.5	107.2	107.7
Tianjin	Metropolitan	72994	85213	93173	111.7	110.9	109.2
Shanghai	Metropolitan	76074	82560	85373	106.4	105	105.7
Beijing	Metropolitan	73856	81658	87475	104.8	103.8	104.9
Shanxi	Mid-North	26283	31357	33628	111.2	110.4	109.6
Hebei	Mid-North	28668	33969	36584	110.6	109.7	108.9
Anhui	Mid-South	20888	25659	28792	118.8	112.6	111.8
Hubei	Mid-South	27906	34197	38572	114.7	113.5	110.7
Jiangxi	Mid-South	21253	26150	28800	113.2	111.8	110.4
Henan	Mid-South	24446	28661	31499	112.6	112.5	110.1
Jilin	North	31599	38460	43415	113.6	113.5	111.9
Liaoning	North	42355	50760	56649	113.4	111.6	109.4
Heilongjiang	North	27076	32819	35711	112.6	112.2	110.1
Shaanxi	Northwest	27133	33464	38564	114.4	113.7	112.6
Gansu	Northwest	16113	19595	21978	111.6	112.3	112.2
Guangxi	South	20219	25326	27952	113.9	112	110.4
Hunan	South	24719	29880	33480	112.9	111.2	110.7
Guangdong	South	44736	50807	54095	109.5	108	107.4
Chongqing	West	27596	34500	38914	116.2	115.1	112.4
Sichuan	West	21182	26133	29608	115.7	115.9	112.3
Guizhou	West	13119	16413	19710	114.7	116.1	113.5
Yunnan	West	15752	19265	22195	111.6	112.9	112.3

Source: China Statistical Yearbook NBS (2013)

Table 5: Income difference in dichotomous data (Part 1)

	(1)		(2)		(3)	
	Rich	Poor	Slow growth	Fast growth	Under-developed	Developed
HHincome(2010)	16055.9	6146.5	9217.6	9056.4	7290.5	11597.3
	(25127.0)	(6877.7)	(15409.0)	(16090.1)	(15940.3)	(14958.6)
INDincome(2010)	28921.4	2632.1	10962.8	10072.9	8211.8	13787.9
	(32154.1)	(2694.1)	(22354.5)	(20062.0)	(18983.2)	(24092.5)
HHincome(2012)	21043.6	7877.8	13759.9	9587.3	9836.1	15887.5
	(23922.6)	(8960.1)	(18589.9)	(12635.5)	(14309.4)	(19298.2)
INDincome(2012)	38667.6	3221.0	17453.0	10411.2	10914.0	20906.9
	(44562.4)	(3753.0)	(34894.7)	(21350.7)	(23083.1)	(39028.2)

Notes:1. HHincome is the household income per capita.

2. INDincome is the individual income per capita.

3. The values in parentheses are the standard error.

Source: CFPS and authors' calculation

Table 6: Income difference in dichotomous data (Part 2)

	(1)		(2)		(3)	
	Female	Male	Majority	Minority	Rural	Urban
HHincome(2010)	9155.3	9236.0	9395.9	6850.0	6305.6	12921.4
	(14198.7)	(17250.9)	(16040.2)	(11279.2)	(11009.9)	(19615.4)
HHincome(2012)	11885.0	12396.2	12478.3	8096.8	8848.5	16358.8
	(15677.6)	(17981.0)	(17197.3)	(10859.4)	(10012.6)	(22046.7)
INDincome(2010)	7472.9	14252.8	11034.3	6682.5	7301.1	15044.2
	(15414.2)	(26359.5)	(22090.6)	(14046.5)	(14777.3)	(27395.6)
INDincome(2012)	10737.5	19139.4	15217.5	8924.9	8968.3	22128.9
	(19151.9)	(39425.7)	(31690.4)	(16118.4)	(15814.9)	(41802.9)

Notes:1. HHincome is the household income per capita

2. INDincome is the individual income per capita

3. The values in parentheses are the standard error.

Source: CFPS and authors' calculation

those with income higher than average as the "rich" group and the others as the "poor" group. We also collected the GRP data from the Chinese Statistical Yearbooks and sorted them by the growth rate and the level of GRP respectively. Since there are 25 provinces in our dataset, we denoted the 12 provinces with the highest growth rate as the "fast growth" group, and the rest the "slow growth" group, and likewise the 12 provinces with the higher GRP the "developed" group and the rest the "under-developed" group. Thus, we generated two dummy variables: one for the growth rate and one for the level of GRP, treating the "fast growth" ("developed") group equal to "1" and the "slow growth" ("under-developed") group equal to "0".

The household and individual income statistics for these six groups are listed in table 5. The "rich" have more than 10 times individual incomes compared to the "poor". The differences shrink to less than 3 times for household incomes.

For the "developed" group, the household income and the individual income are over 50% higher than that in the "under-developed" group in both years. On the contrary, income in fast growth regions is similar to income in slow growth regions in 2010 but lower than income in slow growth regions in 2012. Slow growth regions appear to have more increase in income from 2010 to 2012. This is because higher income regions in 2012 are grouped into the "slow growth" region. It indicates that in 2010 the developed and under-developed regions have similar growth rates while in 2012 the under-developed regions grows faster than the developed regions.

In addition, we divided the sample by gender, Hukou status and ethnicity. The mean and standard deviation are listed in table 6. Male's individual income is almost doubled compared to female's, while male's household income per capita just slightly exceeds female's. Household income of the majority group is 37% more than that of the minority group in 2010, which is similar in terms of individual's

income. Income in urban areas is twice as much as income in rural areas.

5 The Results

The empirical results are presented into four sections. The first section presents inequality of opportunity at the national level; the second is at the regional level; the third presents the provincial IOP and its relationship with GRP. Results from Oaxaca decomposition are in the final part.

5.1 Inequality of Opportunity at the National Level

To measure inequality of opportunity at the national level, we ran regressions based on the hurdle model for the data in 2010 and 2012 respectively. The results are shown in Table 7. Column (1) and column (3) are the results from the logistic regressions and the estimators are presented in odd ratios. Column (2) and column (4) are the results from the linear regressions in the hurdle model.

In Table 7, the coefficients for male, minority, urban Hukou and coastal province are significant at 1% level except for minority in 2010 and urban Hukou in 2012 in the logistic regression. The contributions for each variable to total income range from 28.8% to 134%. The coefficients of parents' socioeconomic status are mostly significant in the linear regression for both years but not significant in the logistic regression except for the mid-level education in 2012. In the linear regression, the significant coefficients of parents' socioeconomic status range from 15.2% to 29.8%. Therefore, gender, ethnicity and geographic characteristics such as province and Hukou status seem to contribute more to income inequality than parents' SOE. These demographic characteristics largely affect not only income but the decision in labour participation as well. Parents' SOE, on the other hand, affects the amount of income earned but has less implication in the labour participation.

In addition, based on the results from the logistic regressions, we predicted the probabilities of earning incomes. Figure 3 shows the distribution of that probabilities. The distributions for both years range from about 0.85 to 1, which suggests that the circumstances imply little on whether individuals will work or not.

In terms of the heteroskedasticity, Table 8 shows the results of MLE using equation (16). The first two columns are the estimators of the mean and the last two are of the variance. Compared this table with Table 7, the coefficients of the mean are similar. Therefore, when computing the Shapley values, the results might be similar if replacing the coefficients of the mean from OLS with that from MLE.

As is shown in the last two columns in Table 8, the heteroskedasticity has a significant effect on income inequality through a variety of factors such as gender, parents' backgrounds and personal geographies. Male's income has a higher variation than female's (7.8% in 2010 and 5.7% in 2012). This difference in income variance might be due to the higher level of effort male spent than female did. In

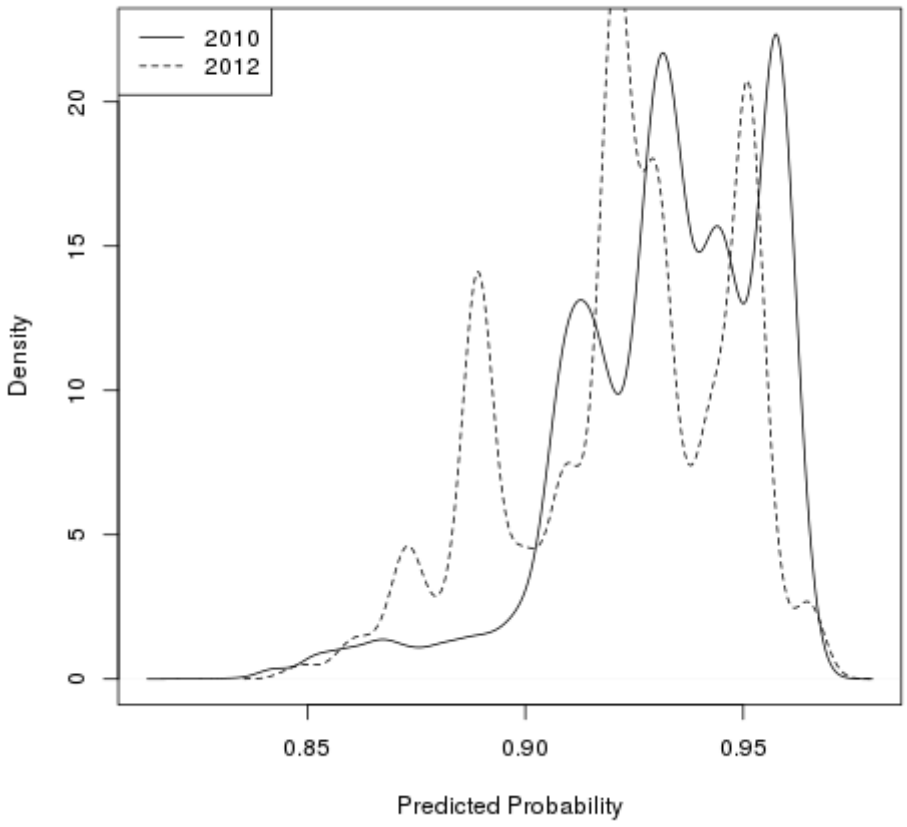
Table 7: The Hurdle Model at the National Level

	Dependent variable:			
	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.664^{***}	0.681^{***}	1.667^{***}	0.730^{***}
	(0.064)	(0.028)	(0.060)	(0.027)
Minority	0.939	-0.288^{***}	1.416^{***}	-0.518^{***}
	(0.116)	(0.052)	(0.125)	(0.050)
Urban Hukou at age 12	0.598^{***}	0.801^{***}	0.973	1.340^{***}
	(0.083)	(0.043)	(0.085)	(0.042)
Coastal Province at age 12	0.768^{***}	0.336^{***}	0.687^{***}	0.499^{***}
	(0.062)	(0.028)	(0.058)	(0.027)
Mid education(Parents)	0.995	0.224^{***}	0.861^{**}	0.255^{***}
	(0.078)	(0.036)	(0.071)	(0.035)
High education(Parents)	0.917	0.022	1.067	0.029
	(0.104)	(0.049)	(0.103)	(0.048)
Mid occupation(Parents)	1.137	0.200^{***}	0.904	0.298^{***}
	(0.100)	(0.045)	(0.091)	(0.045)
High occupation(Parents)	0.939	0.164^{***}	0.825^{*}	0.275^{***}
	(0.114)	(0.054)	(0.106)	(0.054)
Member of CCP(Parents)	1.066	0.171^{***}	1.124	0.152^{***}
	(0.087)	(0.039)	(0.082)	(0.039)
Number of sibling	1.031^{*}	-0.033^{***}	1.005	-0.061^{***}
	(0.017)	(0.008)	(0.016)	(0.007)
Constant	12.462^{***}	7.663^{***}	11.493^{***}	7.816^{***}
	(0.082)	(0.037)	(0.077)	(0.036)
Observations	17,009	15,852	17,009	15,672

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

The standard error is in the parenthesis.

Figure 3: The Distributions of Predicted Probabilities of Earning Incomes



Note:The distributions are a kernel density estimation using a Gaussian kernel
The bandwidths are chosen based on a Silverman's "rule of thumb" (Silverman, 1986)

Source: The predicted probabilities are based on CFPS and authors' calculation.

2012, holding a Hukou status decreases income variances by 63% while it almost has no effect on income variance in 2010. Living in the coastal province also contributes the increase in income variances by 34.2% in 2010 and 6.5% in 2012. Note that variables such as Hukou status with a large difference in coefficients of variance over two periods also have large difference in coefficients of mean. The former difference seems to be offset by the latter one. Therefore, the Shapley decomposition might differ between 2010 and 2012 if assuming the homoskedasticity but the difference might be offset when we consider the heteroskedasticity.

Table 8: The MLE with Type Heteroskedasticity at the National Level

	Mean		Variance	
	2010	2012	2010	2012
Constant	7.651^{***}	7.831^{***}	1.017^{***}	1.113^{***}
	(0.036)	(0.036)	(0.031)	(0.031)
Male	0.677^{***}	0.668^{***}	0.078^{***}	0.057^{**}
	(0.027)	(0.027)	(0.023)	(0.023)
Minority	-0.251^{***}	-0.499^{***}	-0.203^{***}	0.047
	(0.045)	(0.051)	(0.042)	(0.042)
Urban Hukou at age 12	0.803^{***}	1.377^{***}	0.004	-0.630^{***}
	(0.044)	(0.035)	(0.035)	(0.035)
Coastal Province at age 12	0.318^{***}	0.510^{***}	0.342^{***}	0.065^{***}
	(0.028)	(0.027)	(0.023)	(0.023)
Mid education(Parents)	0.214^{***}	0.205^{***}	0.066^{**}	0.068^{**}
	(0.036)	(0.034)	(0.029)	(0.030)
High education(Parents)	0.032	0.062	−0.017	−0.026
	(0.047)	(0.045)	(0.040)	(0.040)
Mid occupation(Parents)	0.190^{***}	0.284^{***}	0.049	−0.052
	(0.046)	(0.041)	(0.037)	(0.037)
High occupation(Parents)	0.143^{**}	0.261^{***}	0.191^{***}	0.058
	(0.057)	(0.051)	(0.044)	(0.045)
Member of CCP(Parents)	0.172^{***}	0.146^{***}	-0.058^{*}	-0.053^{*}
	(0.038)	(0.036)	(0.032)	(0.032)
Number of sibling	-0.026^{***}	-0.056^{***}	-0.048^{***}	-0.019^{***}
	(0.007)	(0.007)	(0.006)	(0.006)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

The standard error is in the parenthesis.

To implement the Shapley decomposition, we grouped 10 explanatory variables into 5 factors — gender, ethnicity, geographic characteristics, parents' socioeconomic status (parents' SOE) and the number of siblings. Geographic characteristics contain Hukou status and coastal province; and parents' socioeconomic status includes parents' educational level, occupational level and member of CCP. When the Shapley values are being computed, the variables belonging to same factors are treated as observed or fixed to the average at the same time.

Table 9 shows the decomposition of IOR at the national level. The first two columns are the decomposition of the results from the linear regression. Only samples with positive income are included. The middle two columns are the decomposition using the whole hurdle model with the homoskedasticity. The last two columns are the decomposition using the hurdle model with the heteroskedasticity.

Comparing the model without zero income with the first hurdle model, we found that the inclusion of zero income only slightly changed the Shapley value for each factor. In total, IOR decreased over 1% in both years if considering zero-income individuals. This slight decline might indicate that those who have advantages in circumstances might be more likely to receive zero income.

In terms of the contribution of each factor, gender and geographic characteristics are two main sources. They together contribute more than 20% of total income inequality if assuming the homoskedasticity. This figure increases to more than 40% when the error are heteroskedastic. Parents' socioeconomic status accounts for around 5% to 7% for the homoskedasticity and more than 10% for the heteroskedasticity; sibling number and ethnicity makes up less than 6% of total income inequality for the homoskedasticity and around 7% to 8% for the heteroskedasticity. To sum up all these factors, we found that IORs were 31.68% in 2010 and 43.49% in 2012 for the linear regression model. When we accounted for the samples with zero income, IORs reduced to 30.45% in 2010 and 41.53% in 2012. However, when we included heteroskedasticity, IORs increased to around 60% in both years.

The difference in IORs between the homoskedasticity and the heteroskedasticity implies that circumstances also largely affect income inequality indirectly through effort. Among these differences in IORs, we found that around 4-5% income inequality is due to the indirect effect of gender, about 8% to 13% is due to geographic characteristics and about 4% is due to parents' SOE. Comparing the results over two periods, we found that the difference in IOR is mainly from the geographic characteristics. It is offset when the heteroskedasticity is taken into account, which is in line with what we found in the results of MLE.

We also did a sensitive analysis for the IOR measures shown in Table 9. The results are presented in the Appendix (Table 15, 16, and 17). In the sensitive analysis, the first test changed the sibling number into a three-level variable: with no sibling, 1 sibling, and 2 or more sibling. The number of types was reduced to 585. The second test dropped the types with less than 5 samples. The number of types was dropped to 534. The last test dropped the types with less than 10 samples. The number of types was further reduced to 296. The measures in the first two tests (Table 15 and 16) are similar to our main results (Table 9). The measures slightly decreased when types with less than 10 samples were dropped. These results from the sensitive analysis show that our main results in Table 9 are robust and not affected by the reduction of samples and types.

Zhang and Eriksson (2010) measured IOR in nine provinces in China from 1989

Table 9: The Shapley decomposition at the National Level

	OLS		Hurdle model 1		Hurdle model 2	
	2010	2012	2010	2012	2010	2012
Gender	10.45	9.72	10.13	9.42	15.19	13.76
Ethnicity	0.98	1.52	0.94	1.40	1.63	1.63
Geographic	12.81	21.99	12.19	20.81	25.65	28.56
Parents' SOE	5.48	6.71	5.19	6.27	10.66	10.67
Sibling_number	1.96	3.55	1.78	3.25	5.52	6.45
Income: +/0			0.24	0.38	0.19	0.37
IOE	68.32	56.51	69.55	58.47	41.15	38.57
IOR	31.68	43.49	30.45	41.53	58.85	61.43

[1] OLS is the regression without zero-income. Hurdle model 1 is the regression using the hurdle model with type homoskedasticity. Hurdle model 2 is the regression using the hurdle model with type heteroskedasticity.
[2] The "Geographic" factor includes individuals' Hukou status when they were 12 years old.
[3] Parents' SOE is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[4] Income: +/0 is the contribution of probability to have a positive income.
[5] All values are presented in percentage.
[6] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

to 2006 excluding individuals with no income. Their results ranged from 46% in 1989 to 63% in 2006. The paper did not use the Shapley decomposition and treated the predicted income from the linear regression as between-type income inequality. If circumstances $\mathbf{c}$ is observable and effort e is unobserved, the relationship between income inequality and circumstances can be modeled by $y = \mathbf{c}\boldsymbol{\beta} + \epsilon$ where ϵ is the residual. Let the predicted income be $\hat{y} = \mathbf{c}\hat{\boldsymbol{\beta}}$, $IOR = I(\hat{y})/I(y)$. [11] Applying the same method as Zhang and Eriksson (2010), we found that IOR is 37.18% in 2010 and 53.64% in 2012. The results are higher than those using our method.

We further implemented Shapley decomposition on IOR computed from predicted income to identify the contribution of each factor. The results are shown in Table 10. The first two columns are the results using Zhang and Eriksson (2010)'s method and the last two are the results from our method. It shows that almost all factors have a higher contribution if using Zhang and Eriksson (2010)'s method.

Considering the results using the predicted income, IOR reduces from 63% in 2006 as reported in Zhang and Eriksson (2010) to around 50% in 2012 as reported in 10. In addition, Zhang and Eriksson (2010) estimated the contributions of each circumstance to total income. They found that parents' socioeconomic status is the most important factor in circumstances while we fount that the most important factors are geographic characteristics including whether living in a coastal province and Hukou status.

Comparing this study to studies on other countries, we found that IOR in China is relatively higher. If we only consider individuals with positive incomes and

[11] The same method can also be found in Manna et al. (2012) if using variances as the inequality index, $I(\hat{y})$ is equal to the coefficient of determination R^2 (Israeli, 2007).

Table 10: The Shapley Decomposition of the Predicted Income at the National Level

	Predicted Income		Observed Income	
	2010	2012	2010	2012
Gender	13.97	11.24	10.45	9.72
Ethnicity	0.98	1.75	0.98	1.52
Geographic	14.07	28.98	12.81	21.99
Parents' SOE	6.36	8.10	5.48	6.71
Sibling Number	1.8	3.58	1.96	3.55
IOR	37.18	53.64	31.68	43.49

[1] The first two columns are the results using Zhang and Eriksson (2010)'s method and the last two are the results from our method.
[2] The "Geographic" factor includes individuals' Hukou status when they were 12 years old.
[3] Parents' SOE is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[4] All values are presented in percentage, representing the contribution of the relative factor to total income inequality.
[5] IOR represents the proportion of income inequality due to circumstances.

assume homoskedasticity across type, IOR in China (31.68% in 2010 and 53.64% in 2012) is higher than most of the Latin American countries (20.8% - 37.3% from de Barros et al. (2009)'s study) and U.S. (around 20% in 2001 (Pistolesi, 2009)). Using Shapley decomposition, Björklund et al. (2011) reported IOR higher than 30% in Sweden including heteroskedasticity while in China IOR is more than 55% if taking heteroskedasticity into account. The higher IOR in China might explain why Chinese respondents have the lowest "feeling of procedural justice" in ISSP survey.

5.2 Inequality of Opportunity at the Regional Level

Table 11 and 12 show the measures of inequality of opportunity at the regional level using the hurdle model assuming the homoskedasticity across types in 2010 and 2012 respectively (the results of regressions are listed in the Appendix from Table 18 to 25). We gave up capturing the heteroskedastic model because at the regional level, most coefficients of the mean and variance from MLE are not significant, which results in a large bias in the Shapley decomposition. Since some regions contain all coastal provinces and some contain all inland provinces, we removed the "coastal province" dummy in the regressions.

In general, IOR varies from 22.84% to 29.61% in 2010 and from 28.92% to 37.89% in 2012. These results are lower than those at the national level. It is probably because the regional disparity contributes to IOP at the national level. In particular, IOR is the highest in the south region over two periods and the lowest in the mid-south region in 2010 and in the east region in 2012. The differences between the highest and the lowest are around 7% in 2010 and 9% in 2012, which indicates a regional disparity in inequality of opportunity in China.

In terms of the Shapley decomposition, gender, Hukou and parents' socioeconomic status are three main sources of income inequality for all regions. This result

is in line with that at the national level. However, the contributions of these three sources vary across regions. In the metropolitan, gender accounts for 10.40% in 2010 and 4.74% in 2012 of total income inequality; while in the mid-north, it made up 16.17% in 2010 and 16.20% in 2012 of total income inequality. The difference between these two regions is more than 6%.

The contributions of Hukou and parents' SOE also show huge difference across regions. Hukou status contributes more than 12% of income inequality in 2010 in the northern west and more than 15% in 2012 in the north and the northern west; while it contributes 0.22% to income inequality in the east in 2010. This slight contribution might due to the smaller rural-urban income gap in the east. In our dataset, the average income in the rural east region is 12,436 Yuan and the urban income is 17,741 Yuan; while in the northern west the rural samples earn 5,335 Yuan and the urban earn 24,587 on average. In terms of parents' SOE, the south region is the highest for both years. It accounts for more than 11% over two periods. The lowest contribution is in the northern west(3.03%) in 2010 and the north(4.05%) in 2012.

To conclude, we find that regional disparities exist not only in income inequality but also in its sources. Rich regions like the metropolitan region have a lower level of income inequality but higher IOR; while poor regions have a higher level of income inequality but lower IOR. Specifically, gender, Hukou and parents' socioeconomic status are three main sources of income inequality. Their Shapley values vary from regions to regions, which indicates large regional heterogeneity in each source of income inequality.

5.3 Provincial Inequality and GRP per capita

We also estimated inequality of opportunity at the provincial level. The IORs at the provincial level are computed using the Shapley decomposition with the hurdle model. We only assumed type homoskedasticity because the sample size for each province is too small to cover all types.

Figure 4 demonstrates the relationship between GRP per capita and inequality at the provincial level. The upper two graphs show GRP per capita with the observed Gini coefficients and IOR respectively. Provinces with higher GRP per capita clearly have lower Gini coefficients but higher IOR. To interpret this difference, we graphed IOE alongside with IOL in the lower panel. The graph shows that IOE has negative relationship with GRP, dropping from around 50% when GRP per capita is lower than 20,000 Yuan to below 30% when GRP per capita is close to 100,000 Yuan if assuming type homoskedasticity, while IOL only increases slightly (from 20% to 22%).

To sum up, IOE clearly has a decreasing trend; while IOL does not show a clear trend. This finding is not consistent with Marrero and Rodrguez (2013), who found that inequality of opportunity is negatively related to growth and inequality

Table 11: Inequality of Opportunity at the Regional Level(2010)

	Metropolitan	Mid-North	North	East	Mid-South	South	West	Northern West
GRP	74308	27476	33677	46421	23624	29891	20161	20126
Observed Gini	0.5729	0.6630	0.6305	0.6415	0.6623	0.7119	0.6506	0.6887
Gender	10.40	16.17	13.91	12.34	10.73	8.78	9.28	10.67
Ethnicity	0.02	2.84	0.31	0.03	0.50	0.61	3.57	0.47
Hukou	6.10	5.10	4.43	0.22	3.71	7.66	4.68	12.53
Parents' SOE	7.15	4.04	3.72	8.18	5.78	11.68	4.07	3.03
Sibling_number	3.11	0.15	2.89	3.09	1.67	0.03	1.09	0.72
Income: +/0	0.65	1.06	0.59	0.63	0.45	0.85	0.47	-0.06
IOE	72.57	70.65	74.16	75.51	77.16	70.39	76.84	72.64
IOR	27.43	29.35	25.84	24.49	22.84	29.61	23.16	27.36

[1] Parents' SOE is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[2] Income: +/0 is the contribution of probability to have a positive income.
[3] All values are presented in percentage.
[4] GRP denotes the Gross Regional Product.
[5] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

Table 12: Inequality of Opportunity at the Regional Level(2012)

	Metropolitan	Mid-North	North	East	Mid-South	South	West	Northern West
GRP	88674	35106	45259	59063	31916	38509	28174	29137
Observed Gini	0.5132	0.6854	0.6183	0.6608	0.6454	0.6730	0.7013	0.7036
Gender	4.74	16.20	9.56	11.84	12.97	7.61	10.39	12.79
Ethnicity	-0.00	1.74	0.06	0.11	0.41	0.21	4.02	0.74
Hukou	11.25	7.01	16.29	4.87	9.07	14.39	8.55	15.53
Parents' SOE	6.90	7.92	4.05	4.39	8.27	11.49	5.15	5.23
Sibling_number	5.92	0.47	2.89	7.04	3.95	3.20	1.58	0.28
Income: +/0	2.27	0.65	0.91	0.68	0.20	1.00	0.07	0.58
IOE	68.92	66.00	66.23	71.08	65.13	62.11	70.24	64.86
IOR	31.08	34.00	33.77	28.92	34.87	37.89	29.76	35.14

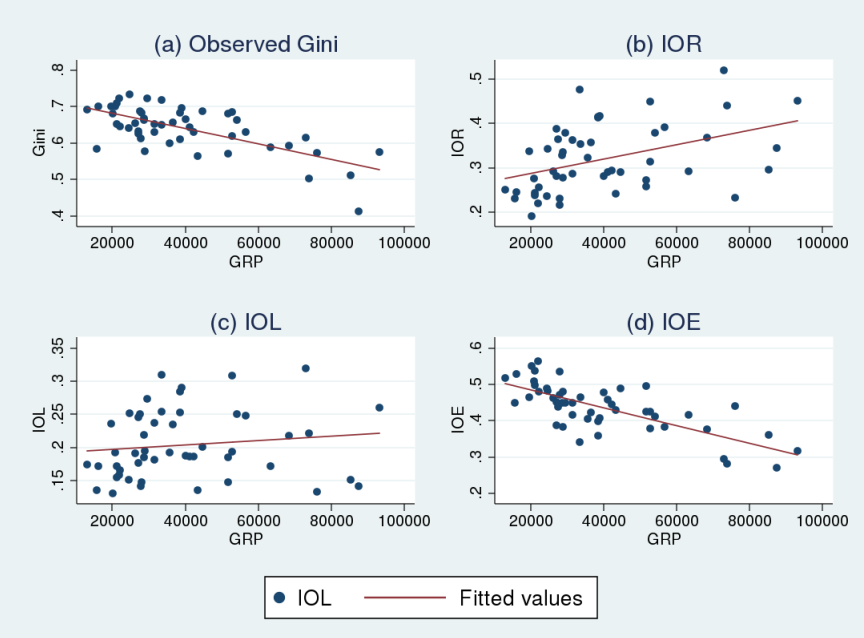
[1] Parents' SOE is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[2] Income: +/0 is the contribution of probability to have a positive income.
[3] All values except GRP and Gini are presented in percentage.
[4] GRP denotes the Gross Regional Product.
[5] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

of effort is positively related.

The difference in the results might be due to the fact that our research focus on a developing country— China; while Marrero and Rodrguez (2013) studied a developed country— United States. At the early stages of development, an increase of effort (e.g. decision for the peasants to find a job in urban areas) might make huge difference in income; while at the late stage, the same amount of increase in effort might make no difference (e.g. an urban job is not as easy to find as 20 years ago for a rural peasant.)

In summary, the results indicate that income inequality reduces from about 0.7 to 0.5 when GRP per capita rises from below 20,000 Yuan to more than 90,000 Yuan. This reduction seems mostly due to the decrease in IOE, which might imply that a poor province has a more diverse distribution of effort or a bigger influence of effort on income inequality.

Figure 4: Provincial Inequality and GRP per capita



Note: 1. GRP is the Gross Regional Product.

2. IOR is the proportion of income inequality due to circumstances.

3. IOL stands for the level of income inequality due to circumstances.

4. IOE represents the proportion of income inequality due to efforts.

Source: GRP is collected from the China statistical yearbook (NBS, 2013).

Observed Gini, IOL, IOE and IOR are based on authors' calculation.

5.4 Results from Oaxaca Decomposition

Table 13 and 14 are the results from Oaxaca decomposition. We separated the dataset by individual's income, growth rate of provinces, GRP per capita of provinces, gender, Hukou status and ethnicity. In both tables, "Advantage" represents the predicted income of the advantage groups: the group with higher income, fast growth, high GRP, male, urban Hukou or majority(Han—the dominant ethnic group in China); while "Disadvantage" represents the predicted income of the disadvantage groups. Based on the predicted income, we decomposed the expected income difference between groups into three components: endowments, coefficients and interaction.

The results show that the predicted income of the advantaged groups is all higher than the disadvantaged except "growth" in 2012. The highest difference comes from the comparison between the high-income group and the low-income group. Most of them can be explained by the difference in coefficients even though the endowments effect is also significant. The rich has slightly better circumstances but their incomes greatly benefit from their circumstances. This result that those who have got ahead take better advantage of their circumstances.

In terms of other divisions of groups, being a male brings no advantage in other circumstances so the endowments in both 2010 and 2012 are close to 0 and not significant. We found that almost all the income difference between gender comes from coefficients. The contributions of coefficients account for 60% to 70% of the expected income difference for Hukou and ethnicity. However, the urban Hukou holders and the majority group are more likely to have advantages in other circumstances. The difference in endowments roughly accounts for 25% to 30% of the expected income difference between urban and rural Hukou holders and it makes up about 30% between the minority and majority group.

With respect to the regional dummies — growth and developed, although the real individual income in developed regions is almost doubled compared with underdeveloped regions (see Table 5), the predicted income is only increased by around 7%. It indicates that most income inequality between developed and under developed regions are not due to circumstances, which is consistent with what we find in the regional and provincial studies.

Table 13: Oaxaca Decomposition(2010)

	(1)	(2)	(3)	(4)	(5)	(6)
	Rich	Growth	Developed	Male	Urban	Majority
Differential						
Disadvantage	7.375***	8.200***	8.015***	7.901***	7.981***	7.816***
	(0.0157)	(0.0198)	(0.0185)	(0.0209)	(0.0167)	(0.0481)
Advantage	10.07***	8.276***	8.537***	8.588***	9.133***	8.263***
	(0.00908)	(0.0256)	(0.0271)	(0.0228)	(0.0357)	(0.0165)

Difference	-2.697***	-0.0765**	-0.522***	-0.686***	-1.152***	-0.446***
	(0.0181)	(0.0324)	(0.0328)	(0.0309)	(0.0394)	(0.0509)
Decomposition						
Endowments	-0.139***	0.0404**	-0.0313	-0.0135	-0.354***	-0.152***
	(0.00822)	(0.0154)	(0.121)	(0.00878)	(0.0397)	(0.0177)
Coefficients	-2.668***	-0.167***	-0.0716	-0.682***	-0.762***	-0.339***
	(0.0242)	(0.0329)	(0.0933)	(0.0300)	(0.162)	(0.0538)
Interaction	0.109***	0.0501**	-0.419**	0.00868	-0.0359	0.0442*
	(0.0182)	(0.0170)	(0.150)	(0.00533)	(0.162)	(0.0257)
Observations	12724	12724	12724	12724	12711	12724

Advantage is the predicted income when the dummy variable listed in column equal to 1.

Disadvantage is the predicted income when the dummy variable listed in column equal to 0.

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$

Table 14: Oaxaca Decomposition(2012)

	(1)	(2)	(3)	(4)	(5)	(6)
	Rich	Growth	Developed	Male	Urban	Majority
Differential						
Disadvantage	6.734***	7.981***	7.641***	7.382***	7.501***	7.315***
	(0.0270)	(0.0332)	(0.0271)	(0.0331)	(0.0254)	(0.0752)
Advantage	10.37***	7.627***	8.190***	8.350***	9.019***	7.879***
	(0.00842)	(0.0332)	(0.0455)	(0.0331)	(0.0555)	(0.0251)
Difference	-3.639***	0.354***	-0.549***	-0.967***	-1.518***	-0.564***
	(0.0283)	(0.0470)	(0.0530)	(0.0469)	(0.0610)	(0.0793)
Decomposition						
Endowments	-0.148***	0.0888	-0.203	-0.0145	-0.508***	-0.174***
	(0.00785)	(0.0803)	(0.169)	(0.0100)	(0.0617)	(0.0231)
Coefficients	-3.768***	-0.0833	-0.210**	-0.955***	-1.594***	-0.428***
	(0.0441)	(0.0653)	(0.0844)	(0.0461)	(0.245)	(0.0845)
Interaction	0.277***	0.349***	-0.136	0.00170	0.584**	0.0391
	(0.0351)	(0.0924)	(0.182)	(0.00837)	(0.245)	(0.0404)
Observations	13561	13561	13561	13561	13545	13561

Advantage is the predicted income when the dummy variable listed in column equal to 1.

Disadvantage is the predicted income when the dummy variable listed in column equal to 0.

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$

In summary, we find that the rich have much higher predicted income than the

poor if only circumstances matter. However, most of the income gap is due to a larger effect of circumstances on incomes of the rich. In addition, circumstances have similar effects on different regions, no matter whether a region is poor or rich with fast or slow growth. Gender inequality can be fully explained by the coefficient effect, while for some disadvantaged groups such as the minority and rural Hukou origins, inequality of opportunity is not only contributed by the coefficient effect but also due to the endowment of circumstances.

6 Conclusion

In this paper, we used the data from CFPS and computed IOR in 2010 and 2012 respectively. Taking advantage of the heterogeneity of regional development in China, we grouped 25 provinces into 8 regions. At the national level, we found that IOR is around 32% in 2010 and 43% in 2012 if assuming the homoskedasticity across types. The figures are higher than the findings of most Latin American countries and U.S. IOR rises up to over 55% if considering the heteroskedasticity. This increase might suggest a large indirect effect of circumstances to income inequality.

We also found evidence of relationships between regional development and inequality. GRP, as a proxy for regional development, has a negative relationship with the observed income inequality measured by Gini coefficients but a positive relationship with IOR. More specifically, income inequality due to effort decreases comparing the rich regions to the poor while that due to circumstances does not show a clear trend. As a result, the overall observed income inequality decreases with regional development and with the rise of IOR.

On the one hand, the results shed light on how income inequality is driven by circumstances and effort. On the other hand, we are aware of the bias existed in the conventional approaches to inequality of opportunity. Using the hurdle model is an attempt to correct the bias from the exclusion of the samples with zero income. Although IORs change little after including zero-income samples, a larger sample size might improve the robustness and the representativeness of the results.

MLE aims to correct another bias—type heteroskedasticity. Using it, we are able to show the indirect contribution of each circumstance to total income inequality. After concerning type heteroskedasticity, IORs show more consistency over two periods than the results without type heteroskedasticity.

We are also aware of the different role circumstances play in income for different cohorts. For example, the disadvantage of minority and rural population is partly due to the endowment of their other circumstances and the disadvantage of female compared to male totally depends on the heterogeneous effects of similar circumstances. More importantly, we found that the rich might only have slightly more advantageous circumstances than the poor. However, the rich seem to benefit much more from their circumstances in income.

Applying these econometric techniques, we are able to relex the assumption of independence between circumstances and effort and measure the indirect effect of effort on circumstances through the heteroskedastic model. Other advanced models such as models for panel data can be applied in future study when the data for more years are available in CFPS.

References

Arneson, R. (1989). Equality and equal opportunity for welfare. *Philosophical Studies*, 56(1):77–93.

Becker, G. S. and Lewis, H. G. (1974). Interaction between quantity and quality of children. In *Economics of the family: Marriage, children, and human capital*, pages 81–90. UMI.

Björklund, A., Jantti, M., and Roemer, J. E. (2011). Equality of Opportunity and the Distribution of Long-Run Income in Sweden. In *IZA Discussion Paper No.5466*, number 5466.

Checchi, D. and Peragine, V. (2010). Inequality of opportunity in Italy. *Journal of Economic Inequality*, 8(4):429–450.

Checchi, D., Peragine, V., and Serlenga, L. (2010). Fair and Unfair Income Inequalities in Europe. IZA Discussion Papers 5025, Institute for the Study of Labor (IZA).

Chen, C.-N., Tsaur, T.-W., and Rhai, T.-S. (1982). The gini coefficient and negative income. *Oxford Economic Papers*, 34(3):473–478.

Cohen, G. A. (1989). On the currency of egalitarian justice. *Ethics*, 99(4):906–944.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844.

de Barros, R. P., Ferreira, F. H. G., Vega, J. R. M., and Chanduvi, J. S. (2009). *Measuring Inequality of Opportunities in Latin America and the Caribbean*. Number 2580 in World Bank Publications. The World Bank.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383):605–610.

Ferreira, F. H. G. and Gignoux, J. (2008). The measurement of inequality of opportunity: theory and an application to Latin America. *World Bank Policy Research Working Paper ....*

Ferreira, F. H. G., Lakner, C., Lugo, M. A., and Ozler, B. (2014). Inequality of opportunity and economic growth : a cross-country analysis. Policy Research Working Paper Series 6915, The World Bank.

Foster, J. E. and Shneyerov, A. A. (2000). Path independent inequality measures. *Journal of Economic Theory*, 91(2):199 – 222.

Hlavac, M. (2015). stargazer: well-formated regression and summary statistics tables. r package version 5.2. htt. *CRANR-projectorg/package= stargazer*.

Hofmarcher, P. (2015). An introduction to r for quantitative economics. *Journal of Statistical Software*, 67(1):1–3.

Israeli, O. (2007). A shapley-based decomposition of the r-square of a linear regression. *The Journal of Economic Inequality*, 5(2):199–212.

Jann, B. et al. (2008). The blinder-oaxaca decomposition for linear regression models. *The Stata Journal*, 8(4):453–479.

John Knight, L. S. (1993). The spatial contribution to income inequality in rural china. *Cambridge Journal of Economics*, 17(2):195–213.

Larsen, C. A. (2016). How three narratives of modernity justify economic inequality. *Acta Sociologica*, 59(2):93–111.

Lefranc, A., Pistolesi, N., and Trannoy, A. (2008). Inequality of opportunities vs. inequality of outcomes: Are western societies all alike? *Review of Income and Wealth*, 54:513–546.

Li, H., Zhang, J., and Zhu, Y. (2007). The Quantity-Quality Tradeoff of Children in a Developing Country: Identification Using Chinese Twins. IZA Discussion Papers 3012, Institute for the Study of Labor (IZA).

Li, S., Sato, H., and Sicular, T. (2013). Inequality in focus vol. 2(no. 2).

Manna, R., Regoli, A., et al. (2012). Regression-based approaches for the decomposition of income inequality in italy, 1998–2008. *Rivista di Statistica ufficiale*, (1).

Marrero, G. A. and Rodrguez, J. G. (2013). Inequality of opportunity and growth. *Journal of Development Economics*, 104(0):107 – 122.

NBS (2013). *China Statistical Yearbook 1996-2012*. National Bureau of Statistics of China.

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

Park, A. and Wang, D. (2010). Migration and urban poverty and inequality in china. *China Economic Journal*, 3(1):49–67.

Pistolesi, N. (2009). Inequality of opportunity in the land of opportunities, 19682001. *Journal of Economic Inequality*, 7(4):411–433.

Program, I. S. S. (2009). Issp 2009 "social inequality iv" - za no. 5400.

Ramos, X. and Van de gaer, D. (2012). Empirical Approaches to Inequality of Opportunity : Principles , Measures , and Evidence. In *IZA Discussion Paper No.6672*, number 6672.

Rawls, J. (1971). *A Theory Of Justice (Orig Edn)*. Harvard paperback. Harvard University Press.

Roemer, J. E. (2000). *Equality of Opportunity*. Harvard University Press.

Shapley, L. S. (1952). A value for n-person games.

Shorrocks, A. F. (2013). Decomposition procedures for distributional analysis: a unified framework based on the shapley value. *Journal of Economic Inequality*, pages 1–28.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

The World Bank (2016a). Gini index.

The World Bank (2016b). Poverty and equity in china.

Van De Gaer, D. (1993). *Equality of Opportunity and Investment in Human Capital*. Faculteit der Economische en Toegepaste Economische Wetenschappen, Katholieke Universiteit Leuven. Kath. Univ.

Wan, G., Lu, M., and Chen, Z. (2006). Globalization and Regional Income Inequality: Empirical evidence from within China. Technical report.

Wan, G. and Zhou, Z. (2005). Income inequality in rural china: Regression-based decomposition using household data. *Review of Development Economics*, 9(1):107–120.

Wu, X. and Treiman, D. J. (2004). The household registration system and social stratification in china: 1955–1996. *Demography*, 41(2):363–384.

Xie, Y. and Zhou, X. (2014). Income inequality in todays china. *Proceedings of the National Academy of Sciences*, 111(19):6928–6933.

Zhang, Y. and Eriksson, T. (2010). China Economic Review Inequality of opportunity and income inequality in nine Chinese provinces ,. *China Economic Review*, 21(4):607–616.

Appendices

Table 15: The Measures of Inequality of Opportunity at the National Level (2-level Sibling Number)

	OLS		Hurdle model 1		Hurdle model 2	
	2010	2012	2010	2012	2010	2012
Gender	10.49	9.86	10.16	9.53	15.29	13.96
Ethnicity	1.00	1.56	0.95	1.43	1.71	1.70
Geographic	12.81	22.22	12.18	21.00	26.05	28.97
Parents' SOE	5.63	7.09	5.32	6.62	11.32	11.53
Sibling_number	1.55	2.32	1.42	2.15	4.29	4.63
Income: +/0			0.25	0.45	0.20	0.44
IOE	68.52	56.94	69.72	58.82	41.15	38.76
IOR	31.48	43.06	30.28	41.18	58.85	61.24

[1] OLS is the regression without zero-income. Hurdle model 1 is the regression using the hurdle model with type homoskedasticity. Hurdle model 2 is the regression using the hurdle model with type heteroskedasticity.
[2] The "Geographic" factor includes individuals' Hukou status when they were 12 years old.
[3] Parents' SOE is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[4] Income: +/0 is the contribution of probability to have a positive income.
[5] All values are presented in percentage.
[6] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

Table 16: The Measures of Inequality of Opportunity at the National Level(Dropping Types with less than 5 Samples)

	OLS		Hurdle model 1		Hurdle model 2	
	2010	2012	2010	2012	2010	2012
Gender	10.76	10.23	10.45	9.93	15.68	14.39
Ethnicity	0.84	1.46	0.81	1.34	1.36	1.58
Geographic	12.13	20.62	11.56	19.46	25.02	26.96
Parents' SOE	5.47	6.37	5.19	5.96	11.16	10.47
Sibling_number	2.13	3.86	1.95	3.55	5.58	7.28
Income: +/0			0.25	0.45	0.19	0.43
IOE	68.66	57.46	69.79	59.30	41.01	38.88
IOR	31.34	42.54	30.21	40.70	58.99	61.11

[1] OLS is the regression without zero-income. Hurdle model 1 is the regression using the hurdle model with type homoskedasticity. Hurdle model 2 is the regression using the hurdle model with type heteroskedasticity.
[2] The "Geographic" factor includes individuals' Hukou status when they were 12 years old.
[3] Parents' SOE is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[4] Income: +/0 is the contribution of probability to have a positive income.
[5] All values are presented in percentage.
[6] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

Table 17: The Measures of Inequality of Opportunity at the National Level(Dropping Types with less than 10 Samples)

	OLS		Hurdle model 1		Hurdle model 2	
	2010	2012	2010	2012	2010	2012
Gender	11.41	11.05	11.08	10.73	16.36	15.69
Ethnicity	0.81	1.43	0.77	1.31	1.33	1.54
Geographic	9.87	17.74	9.38	16.73	22.46	24.55
Parents' SOE	5.53	5.67	5.26	5.30	10.24	9.32
Sibling_number	1.76	3.63	1.59	3.32	5.40	7.20
Income: +/0			0.31	0.43	0.23	0.41
IOE	70.63	60.48	71.61	62.18	43.97	41.29
IOR	29.37	39.52	28.39	37.82	56.02	58.71

[1] OLS is the regression without zero-income. Hurdle model 1 is the regression using the hurdle model with type homoskedasticity. Hurdle model 2 is the regression using the hurdle model with type heteroskedasticity.
[2] The "Geographic" factor includes individuals' Hukou status when they were 12 years old.
[3] Parents' SOE is the parents' socioeconomic status which include parents' educational level, occupational status and political affiliations.
[4] Income: +/0 is the contribution of probability to have a positive income.
[5] All values are presented in percentage.
[6] IOE stands for the proportion of income inequality due to effort and IOR represents the proportion of income inequality due to circumstances.

Table 18: The Hurdle Model at the Regional Level (Metropolitan)

	Dependent variable:			
	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.708**	0.585***	1.304	0.264***
	(0.225)	(0.092)	(0.202)	(0.067)
Minority	0.774	−0.198	354,787.100	−0.005
	(1.080)	(0.569)	(438.076)	(0.391)
Hukou at age 12	1.110	0.352***	2.991***	0.536***
	(0.236)	(0.099)	(0.250)	(0.072)
Mid education(Parents)	0.629*	0.195*	0.489***	0.195**
	(0.262)	(0.112)	(0.240)	(0.082)
High education(Parents)	0.433**	0.392**	0.618	0.129
	(0.327)	(0.160)	(0.343)	(0.115)
Mid occupation(Parents)	0.956	0.079	1.316	0.183**
	(0.268)	(0.117)	(0.272)	(0.084)
High occupation(Parents)	1.069	0.037	1.564	0.133
	(0.368)	(0.149)	(0.362)	(0.108)
Member of CCP(Parents)	1.855*	0.186	1.026	0.056
	(0.326)	(0.120)	(0.277)	(0.088)
Number of sibling	1.026	−0.047*	0.853***	−0.076***
	(0.066)	(0.028)	(0.057)	(0.020)
Constant	12.629***	8.992***	12.713***	9.822***
	(0.271)	(0.116)	(0.249)	(0.085)
Observations	1,468	1,374	1,468	1,354
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 19: The Hurdle Model at the Regional Level (Mid-North)

	Dependent variable:			
	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	2.325***	0.938***	2.309***	1.111***
	(0.209)	(0.083)	(0.208)	(0.085)
Minority	0.780	−1.222***	2.801*	−0.768***
	(0.350)	(0.173)	(0.596)	(0.172)
Hukou at age 12	0.459**	1.129***	0.498*	1.403***
	(0.394)	(0.211)	(0.377)	(0.218)
Mid education(Parents)	1.084	0.123	0.791	0.270**
	(0.229)	(0.102)	(0.217)	(0.105)
High education(Parents)	1.907	0.083	1.202	0.380**
	(0.435)	(0.147)	(0.370)	(0.151)
Mid occupation(Parents)	2.212*	0.027	0.747	0.457***
	(0.423)	(0.147)	(0.299)	(0.154)
High occupation(Parents)	3.018**	0.196	1.861	0.128
	(0.547)	(0.174)	(0.492)	(0.179)
Member of CCP(Parents)	0.820	0.209*	1.367	0.174
	(0.269)	(0.117)	(0.289)	(0.118)
Number of sibling	1.115*	−0.004	1.096	−0.012
	(0.059)	(0.025)	(0.059)	(0.025)
Constant	7.106***	7.423***	8.041***	7.394***
	(0.229)	(0.108)	(0.232)	(0.110)
Observations	1,868	1,747	1,868	1,745

Note: *p<0.1; **p<0.05; ***p<0.01

Table 20: The Hurdle Model at the Regional Level (North)

	Dependent variable:			
	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.550***	0.758***	1.931***	0.575***
	(0.130)	(0.067)	(0.145)	(0.060)
Minority	1.396	0.093	1.128	−0.022
	(0.256)	(0.117)	(0.252)	(0.106)
Hukou at age 12	0.622***	0.354***	0.942	0.978***
	(0.141)	(0.081)	(0.161)	(0.072)
Mid education(Parents)	1.097	0.092	0.852	0.009
	(0.151)	(0.082)	(0.160)	(0.073)
High education(Parents)	1.071	0.028	1.146	0.061
	(0.228)	(0.122)	(0.261)	(0.108)
Mid occupation(Parents)	0.995	0.179*	0.738	0.188**
	(0.183)	(0.101)	(0.193)	(0.091)
High occupation(Parents)	0.897	0.237*	0.752	0.300***
	(0.219)	(0.124)	(0.246)	(0.111)
Member of CCP(Parents)	0.965	0.064	1.493**	0.078
	(0.166)	(0.091)	(0.197)	(0.081)
Number of sibling	1.081**	−0.047***	1.017	−0.046***
	(0.034)	(0.017)	(0.036)	(0.015)
Constant	6.113***	8.142***	7.780***	8.412***
	(0.161)	(0.088)	(0.174)	(0.078)
Observations	2,654	2,366	2,654	2,415

Note: *p<0.1; **p<0.05; ***p<0.01

Table 21: The Hurdle Model at the Regional Level (East)

	Dependent variable:			
	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.594**	0.690***	1.406**	0.813***
	(0.192)	(0.108)	(0.162)	(0.096)
Minority	566,745.500	−0.108	956,604.000	0.421
	(484.012)	(0.705)	(480.644)	(0.615)
Hukou at age 12	1.000	−0.294	0.556*	1.130***
	(0.449)	(0.258)	(0.329)	(0.237)
Mid education(Parents)	0.894	0.485***	0.782	0.261*
	(0.252)	(0.150)	(0.211)	(0.134)
High education(Parents)	0.835	0.051	1.400	0.041
	(0.315)	(0.195)	(0.320)	(0.170)
Mid occupation(Parents)	1.022	0.335*	1.256	−0.143
	(0.322)	(0.187)	(0.282)	(0.167)
High occupation(Parents)	1.076	0.447**	1.466	0.113
	(0.364)	(0.204)	(0.314)	(0.181)
Member of CCP(Parents)	1.297	−0.174	0.745	0.277**
	(0.276)	(0.148)	(0.206)	(0.134)
Number of sibling	1.008	−0.047	0.952	−0.117***
	(0.050)	(0.029)	(0.043)	(0.026)
Constant	9.538***	8.081***	8.612***	8.350***
	(0.215)	(0.129)	(0.191)	(0.115)
Observations	1,662	1,535	1,662	1,481

Note: *p<0.1; **p<0.05; ***p<0.01

Table 22: The Hurdle Model at the Regional Level (Mid-South)

	Dependent variable:			
	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.672***	0.644***	1.826***	0.835***
	(0.171)	(0.062)	(0.176)	(0.064)
Minority	1.474	−1.140***	0.698	−1.128***
	(0.729)	(0.234)	(0.535)	(0.250)
Hukou at age 12	0.985	0.475***	0.537***	0.966***
	(0.275)	(0.105)	(0.235)	(0.112)
Mid education(Parents)	0.971	0.245***	0.938	0.243***
	(0.205)	(0.078)	(0.205)	(0.081)
High education(Parents)	0.792	0.046	0.960	0.011
	(0.258)	(0.106)	(0.280)	(0.110)
Mid occupation(Parents)	1.099	0.184*	0.919	0.360***
	(0.272)	(0.100)	(0.260)	(0.105)
High occupation(Parents)	0.836	−0.163	0.625*	0.235*
	(0.288)	(0.117)	(0.269)	(0.123)
Member of CCP(Parents)	0.953	0.272***	0.947	0.273***
	(0.233)	(0.090)	(0.229)	(0.094)
Number of sibling	1.032	−0.031	1.047	−0.072***
	(0.050)	(0.019)	(0.052)	(0.019)
Constant	12.537***	7.755***	14.257***	7.948***
	(0.212)	(0.083)	(0.221)	(0.086)
Observations	2,773	2,613	2,773	2,619

Note: *p<0.1; **p<0.05; ***p<0.01

Table 23: The Hurdle Model at the Regional Level (South)

	Dependent variable:			
	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.948***	0.665***	1.493***	0.616***
	(0.165)	(0.087)	(0.127)	(0.089)
Minority	4.519**	0.339*	1.491	−0.132
	(0.721)	(0.197)	(0.341)	(0.204)
Hukou at age 12	0.499***	0.840***	1.132	1.293***
	(0.217)	(0.141)	(0.203)	(0.141)
Mid education(Parents)	1.077	0.373***	1.183	0.519***
	(0.210)	(0.114)	(0.166)	(0.117)
High education(Parents)	1.304	−0.190	1.666**	−0.229*
	(0.251)	(0.133)	(0.215)	(0.134)
Mid occupation(Parents)	1.632*	0.271**	0.912	0.267*
	(0.280)	(0.138)	(0.205)	(0.142)
High occupation(Parents)	0.717	0.396**	0.514***	0.610***
	(0.264)	(0.167)	(0.214)	(0.174)
Member of CCP(Parents)	1.194	0.487***	1.603**	0.114
	(0.244)	(0.131)	(0.208)	(0.134)
Number of sibling	0.960	0.002	1.025	−0.059**
	(0.043)	(0.024)	(0.035)	(0.025)
Constant	9.374***	7.504***	4.330***	8.028***
	(0.204)	(0.116)	(0.162)	(0.120)
Observations	2,203	2,019	2,203	1,898

Note: *p<0.1; **p<0.05; ***p<0.01

Table 24: The Hurdle Model at the Regional Level (West)

	Dependent variable:			
	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.331	0.543***	1.473*	0.698***
	(0.181)	(0.068)	(0.205)	(0.078)
Minority	0.955	−0.286***	1.353	−0.396***
	(0.189)	(0.072)	(0.226)	(0.082)
Hukou at age 12	0.756	0.820***	0.572	1.361***
	(0.371)	(0.163)	(0.353)	(0.191)
Mid education(Parents)	1.395	0.380***	1.301	0.403***
	(0.292)	(0.105)	(0.303)	(0.120)
High education(Parents)	0.810	0.002	0.712	0.109
	(0.285)	(0.122)	(0.302)	(0.140)
Mid occupation(Parents)	0.645	−0.167	0.564*	0.151
	(0.326)	(0.137)	(0.335)	(0.159)
High occupation(Parents)	0.599	−0.186	0.707	−0.179
	(0.402)	(0.175)	(0.433)	(0.200)
Member of CCP(Parents)	1.473	0.249**	1.102	0.224*
	(0.306)	(0.108)	(0.308)	(0.125)
Number of sibling	1.035	0.032*	1.149**	−0.033
	(0.049)	(0.019)	(0.057)	(0.021)
Constant	11.628***	7.601***	10.386***	7.646***
	(0.209)	(0.082)	(0.223)	(0.094)
Observations	2,062	1,926	2,062	1,953

Note: *p<0.1; **p<0.05; ***p<0.01

Table 25: The Hurdle Model at the Regional Level (Northern West)

	Dependent variable:			
	2010		2012	
	logistic	*OLS*	*logistic*	*OLS*
	(1)	(2)	(3)	(4)
Male	1.384	0.598***	2.222***	0.810***
	(0.309)	(0.062)	(0.217)	(0.065)
Minority	0.164***	−0.391**	0.377**	−1.173***
	(0.457)	(0.196)	(0.424)	(0.204)
Hukou at age 12	0.340***	1.412***	1.176	1.849***
	(0.417)	(0.136)	(0.399)	(0.141)
Mid education(Parents)	0.814	0.137	0.568**	0.310***
	(0.379)	(0.085)	(0.238)	(0.090)
High education(Parents)	0.605	−0.076	0.657	−0.047
	(0.577)	(0.139)	(0.398)	(0.146)
Mid occupation(Parents)	2.322	−0.031	0.932	0.167
	(0.660)	(0.127)	(0.356)	(0.135)
High occupation(Parents)	1.424	−0.026	1.220	0.115
	(0.601)	(0.145)	(0.437)	(0.153)
Member of CCP(Parents)	0.511*	0.200**	0.737	0.145
	(0.374)	(0.089)	(0.257)	(0.094)
Number of sibling	1.210**	−0.014	0.989	−0.006
	(0.092)	(0.017)	(0.054)	(0.018)
Constant	36.700***	7.418***	19.359***	7.416***
	(0.357)	(0.075)	(0.238)	(0.079)
Observations	2,319	2,272	2,319	2,207

Note: *p<0.1; **p<0.05; ***p<0.01