UCLouvain KU LEUVEN

LOUVAIN School of Management

Louvain School of Management
and Katholieke Universiteit Leuven

What are the drivers of the income inequality for the women in urban China in 1995 and 2002?

Research Master's Thesis submitted by
Panozzo Marine

With a view of getting the degrees:
Master in Business Engineering
Major in Financial Engineering (LSM)
Major in Data Science and Business Analytics (KUL)

Supervisor :
Prof. Van Keilegom Ingrid

Academic Year [*2019-2020*]

KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

FACULTY OF ECONOMICS AND BUSINESS

What are the drivers of the income inequality for women in urban China in 1995 and 2002?

MASTER OF BUSINESS ENGINEERING

Panozzo Marine
0768158

Thesis submitted to obtain the degree of

Major Data Science and Business Analytics

Promoter: Prof. Van Keilegom Ingrid

Academic year 2019-2020

Abstract

Leuven, May, 2020.

This thesis has shown that the income inequality for urban Chinese women has significantly increased in 7 years. Moreover, to analyze the drivers behind income inequality, the regression-based decomposition and Shapley value decomposition methods have been used. It appeared that the two most important contributors in the explained inequality remained the *Province* and the *Occupation*, for both years. Besides, the *Work Experience* seemed to contribute significantly less into the income inequality in 2002 than in 1995, which is in line with the income inequality literature conclusions for the whole population. On the other hand, the contributions of the *Industry* and the *Ownership* factors have increased significantly. This finding was also found by other papers, considering the whole population in China. Regarding, the *Communist Party in China* factor, it has been analyzed as the last contributor for both years. As regards the *Education*, the factor kept its place of fourth contributor even though, for the whole population, the factor contributed more and more into the income inequality.

Acknowledgment

This Master Thesis is the culmination of my path at the university. Enrolled in a Double Degree program, this thesis marks the final step of my master in Business Engineering both at the KuLeuven and at the Louvain School of Management.

I wish to show my gratitude to my supervisor, Professor Ingrid Van Keilegom who has been good counsel from start to finish and was incredibly available and understanding throughout the whole process. I would not have been able to provide such a work without her persistent help.

These past two and half months have been particularly challenging in different ways. Therefore, I would like to thank my step family that welcomed me and has been able to procure me a calm, safe and hopeful environment for the writing of my Master Thesis. More specifically, I would like to offer my special thanks to Maxime Collard, who has been to me a mentor all along my way.

Contents

Chapter 1

Introduction

Although several years ago it was considered as a developing country, China has then become one of the fastest-growing economies in the world. Indeed, as mentioned in the journal L'Echo in 2019, its GDP has changed from 60 billion dollars in 1960 to 14.217 billion dollars in 2018 which makes China the second-word economic power after the United-States. Nevertheless, this change does not make the country an enviable model. The human rights violation, inequality and the undemocratic organization of power are some of the consequences of such a fast-growth economy (Georis, 2019).

The turning point was in 1979 when China faced a period of economic reforms. The period after 1979 was followed by different reforms in terms of importation/exportation but also in terms of decentralization of power. Indeed, cities began to have the possibility to establish a free market. From 1979 to 2018 China's annual real GDP growth averaged 9.5%. With this GDP growth, China has been able to double the size of its economy in real terms. One consequence of these changes and evolution is that China is on the way to eradicating poverty. (Morrison, 2019)

As said, China became the world's second-largest economy. Over the past century, the country has seen a change in his society, passing from an agrarian population to a more industrial one (Hirst, 2015). China is the leading country in terms of manufacturing, making the country one of the biggest manufacturing centers in the world. Specialized in labor-intensive products, its production is based on the exportation of cheap goods. In terms of growth strategy, China positions itself as a country that assembles and exports low-priced goods worldwide (Eckart, 2016).

On the other hand, in 2018, the International Monetary Fund released a working paper where they focused on the trends, drivers and policy remedies to inequality in China. It comes out of the paper that, over the past two decades, China has experienced a high reduction of its poverty but meanwhile, a considerable increase in inequality has been observed. They also claimed that this increase in inequality can be one of the consequences of the fast-growing economy. Since 1980 inequality in China has increased

that much that China became one of the most unequal countries in the world by 2013. Over the years, inequality has increased although a moderate decline in income inequality has been observed in 2008. (Jain-Chandra and al, 2018)

This sudden rise of inequality has led to an increasing number of research papers and books on the topic. One of them is the book from Wang (2007). The purpose of his work was to analyze and identify the major sources that drive this inequality. Wang made his conclusions based on the survey of the National Bureau of Statistics (NBS) and incorporated the urban population from 1986 to 2000. He found out that factors such as the town of residence, the type of own company and the company sector of the worker have a big impact and drives this inequality. Another one, Xie (2010) showed that the high inequality of incomes in China is mainly driven by the structural forces such as the urban-rural gap and the variation in the different regions in terms of economic well-being.

To analyze the drivers behind the increasing or decreasing trend in income inequality, models have been developed to analyze and quantify the impact of some factors. One of the interesting models has been developed by Fields (2003). This model is called regression-based decomposition and can be used to analyze and quantify the impact of some determinants on the income. Some years later, Shorrocks (2013) introduced a regression-based decomposition based on the Shapley value. The Shapley decomposition is based on the principle of the Shapley value from the cooperative game theory. These two methods will correspond in the rare case there is no correlation between the variables. From this time, many papers focusing on the inequality used these methods to explain the underlying factors behind the increasing trend of inequality. One of them is the paper from Deng and Li (2009) who worked on the China urban population in order to highlight the factors behind earning inequality, from 1988 to 2002. Their conclusions are in line with the one from Wang (2007) and they have proved that the geographical factor impacted the most the inequality, followed by the educational, the ownership, the occupation and the industry factors. These are factors that had an increasing impact between 1988 and 2002. On the other hand, they have found that the work experience contribution is minimal.

As seen in the inequality literature, the inequality has increased for the whole population in China but no one answered the following question: Does the same conclusion apply for the women in urban China between 1995 and 2002? Moreover, how factors that contribute to the income inequality of the whole population have contributed to the increasing/decreasing inequality for urban women? Indeed, many researches have already focused on the inequality in China, but not so many have focused their attention on women's statistics taken separately from the whole population. Therefore, this thesis wants to fill this gap by tackling the subject of women's income inequality in urban China by highlighting the factors behind this inequality. Moreover, to do so, this thesis will be based on the data collected by the National Bureau of Statistics who has performed the Chinese Household income survey (CHIP). This thesis focuses

on CHIP 1995 and CHIP 2002. Besides, the subgroup of urban women will be considered.

To confirm if what was found in the literature can also be applied to women's income inequality, some tools are needed. The tools that will be used are the Lorenz Curve, which will bring a more visual overview of the inequality, and the Gini Index, which will concretely quantify the inequality from a range from 0 to 1. The Gini Index will also allow to confirm significantly the increase of income inequality. To do that, estimators will be calculated thanks to the Davidson's plug in formula for the standard deviation and his unbiased estimator formula. After this analysis, the Concentration Curve will be used to decompose the incomes. These inequality tools will help to have an insight into which source of the incomes is the most unequally distributed. As regards the analysis of the drivers behind the inequality, two important methods will be used : the regression-based decomposition model (Fields, 2003) and the Shapley decomposition (Shorrocks, 2013).

Therefore, the first part of this thesis will be composed of a literature review of the income inequality in China. The chapter number three called "Method" will first begin by a description of the database and the factors considered. Then, the different concepts such as the Lorenz Curve, the Gini Index, the Concentration Curve and the Concentration Index will be described. After that, the two methods used in this thesis to quantify the impact of the factor on the income inequality will be discussed. Finally, to give more statistical value to the results, the boostrapping and the cross-validation method will be explained. The chapter four will focuses on the results and analysis. As mentioned, this thesis aims to answer the question : "What are the drivers of the income inequality for women in urban China in 1995 and 2002?". To answer the question, the analysis part will be divided into two parts. First, the analysis will aim to respond to the question "Is the inequality increasing for the urban Chinese women in 2002 compared to 1995?". To do so, the tools described in the first part of the method chapter will be used. Then, in the case the inequality is significantly increasing/decreasing, the analysis will focus on the second question "How factors that contribute to the income inequality of the whole population have contributed to the increasing/decreasing inequality for urban women in 2002 compared to 1995?". This will be realized thanks to the tools discussed in the second part of the method chapter. Finally, this thesis will end with a conclusion where it will be possible to answer to the research question.

Chapter 2

Literature Review

Initially considered as a poor developing country, China became a major economic power in the last four decades. As explained in the article of Eckart (2016) in the World Economic Forum, the turning point of the Chinese economy took place in 1979 when the country initiated a serie of economic reforms. The period before 1979 had been driven by the leadership of Chairman Mao Zedong that extolled a centrally planned and supervised economy. At that time, the country's economy was mainly directed by the central government, following one main goal: make China a self-sufficient nation. Importations were limited to the goods that could not be made or obtained within China. The fact that the country's economy output was controlled by the central government and the lack of market mechanisms had an impact on the allocation of resources. This situation decreased the incentives for workers, firms and farmers to be more efficient or concerned by the quality of their production. Those actors were following government goals and directives. During the period 1953-1978, China's real GDP grew at an average rate of 6,7%. However, many analysts have questioned the accuracy of this number and suspected exaggeration of the production levels by the government for various political reasons. (Morrison, 2019)

After 1979, China's economy experienced several important reforms on the impulse of the government. The farmers for example were able to sell a portion of their crops on the free market thanks to the decision of the central government to initiate price and ownership incentives. The government completely changed its mind in terms of exportation and importation from this date. Indeed, four economic zones along the coasts were established to boost exports, to attract foreign investments and to import high technology products in China. In these coastal regions and cities, the government implemented reforms and designated them as open cities and development zones which allowed them to develop free-market policies and to offer tax and trade incentives to attract foreign investments. Other reforms such as the decentralization of the economic policy making in several sectors and the trade liberalization were major keys to China's economic success. The implementation of these reforms was done gradually to identify which policies were creating a favorable impact on economic outcomes. Then, those

policies would be implemented in other parts of the country. From 1979 until 2018, China's annual real GDP growth averaged 9.5% meaning that, on average, every eight years, China doubled the size of its economy in real terms (Morrison, 2019).

As already mentioned, China has put in place a wide range of market reforms in 1978. Since this time, reforms have allowed to significantly improve the weighted distribution of the productive forces, leading to a rapid and sustained economic growth. As a consequence, what has been observed, is a tied increase in household income and an improvement of the living conditions. That means, in terms of numbers, that the average annual growth rate of China's GDP was 9.4% from 1978 to 2001 meanwhile the net income per capita in urban households increased by 6.4% annually (Morrison, 2019). For China's history, this period has been the most substantial economic development accompanied by the highest increase in income for the urban as well as rural households. This has also been the time where Chinese people obtained their greatest benefits. Overall, a well-off standard of living has been reached (Hirst, 2015).

However, a sharp increase in inequality has been observed during these 23 years. Indeed, in late 1980, the Gini Index appeared to be around 0.3 then it rose over 0.5 in 2010 (Xie, 2010). Although, those numbers could be even bigger if the statistical errors such as unreported high income and illegal incomes were taken into account. Besides, this continuous increase in inequality has had a negative impact on the economic and social development. For example, the low-income group of the urban population has increased, resulting in an obstruction of the growth in consumption and as a result, on the growth of the nation's economy.

When it comes to talk about inequality in a socio-economic context, it is useful to distinguish the different types of inequalities. In fact, in economics, two main types of inequalities exist: income and wealth inequality. As explained in the book of Cowell (2011), income inequality refers to the magnitude of the difference in revenues within the population. The definition of revenue includes different kinds of revenue such as dividends, wages, salaries, interest on saving accounts and profit or rent from selling companies. On the other hand, wealth inequality refers to the percentage of total wealth possessed by the top percentiles of the population compared to the bottom percentiles. Wealth can be either cars, houses, shares in firms or financial assets. In this thesis, the total income by year will be used and the definition of income encompasses four main revenue types: wage, income from private business or self-employed, property income and transfer income.

In 2004, the Organization for Economic Co-operation and Development released an analysis on the Income Disparities in China. Among others, the paper analyzes the pattern that contributes to the change of the income distribution. It appears in this paper that factors such as the region, the area, the social group and the sector of activity have been increasing inequality among inhabitants since the year of the reforms 1978.

As a summary, the OECD gave a brief overview of the change in inequality caused by the sectors, the social groups and the regions. For example, the increase in inequality has been faster in the coastal areas than in the central and western regions. On the other hand, the increase in inequality in the 16 sectors of the economy has been more rapid in monopolistic sectors than in competitive sectors. Finally, inequality among certain social groups have evolved rapidly. While the top decile of the income distribution for the urban population earns 3.8 times more than the bottom urban decile in 1996, it increased to 5.4 in 2001. In conclusion, inequality has evolved at the same time as the economy and this increase had several causes, such as the region of living, the sector of activity and the social group the inhabitants of China are part of. (OECD, 2004)

Social inequality in China has been one of the most discussed topics in papers published in this nation in recent years. Among the literature about the inequality in China, some of them have tried to explain these inequalities and have looked at some factors that could explain them. The paper from Wang (2007) contributes in a significant way to the literature about the social inequality in China. Wang's paper and analysis are based on China's urban population during the period between 1986-2000. According to Wang (2007), the most important factors that drive and maintain the inequality in China are ownership type, industry, work organization and geographic location. In 1988 for example, gender, education, length and type of employment explain up to 35% of the change in income while adding the province and the town variables allow to explain 47% of the change. In 1988, the locality and the sector of the firms did not influence the inequality. In 1995, on the other hand, if the four first factors are incorporated in the model, they only explain 20% of the change in income. When the locality and sector type of the firm is added into the model, the explanatory power of this one increases by 75% and by 100% if the province is added. Wang also discussed of the limitation of this paper. Indeed, as said before, the paper uses the urban population and not the rural population. Besides, in his paper, the migrant population that comes from rural zones and migrates to the urban zones for a moment, is not taken into account. In fact, those people are often less qualified and earn less for the same job than the locals.

Furthermore, the sociologist Yu Xie wrote a paper in 2010 called "Understanding inequality in China" where he analyzed the evolution of inequality, the forces driving it and its impact in China. He also concluded that the high inequality is mainly explained by the huge difference in the different regions in terms of economic well-being as well as the rural-urban gap.

Some years later, a paper tried to explain this increasing inequality in China and compared it with the inequality in the USA. The paper from Xie and Zhou (2014), called "Income inequality in today's China", has contributed to the literature by testing the hypothesis of Xie (2010) and Wang (2007) that the current rising inequality is mainly driven by the urban-rural gap and the variation among regions. To perform

their analysis, they have used data from the year 2005. Moreover, their analysis was mainly based on the analysis of the bi-variate R-squared and partial R-squared. Thanks to their analysis Xie and Zhou conclusions have confirmed the hypothesis of Xie (2010) and Wang (2007)

In the same period, Knight and Song (2008) wrote a paper about the wage income between 1995 and 2002. They based their findings on the wage function and the decomposition of the increase in mean real wages and a quantile regression. One of their conclusions is that the main wage difference can be explained by the effect of the ownership sector, the province and the Communist Party membership. Moreover, they have looked at the reason why ownership was one of the main causes of the wage gap. They have found out that it because foreign firms selected more productive workers. On the other hand, urban collectives firms were mainly composed of less productive workers.

To analyze the drivers behind the increasing or decreasing trend in income inequality, models have been developed to analyze and quantify the impact of some factors. One of the interesting models has been developed by Fields (2003). This model is called regression-based decomposition and can be used to analyze and quantify the impact of some determinants on the income. Some years later, Shorrocks (2013) introduced a regression-based decomposition based on the Shapley value. The Shapley decomposition is based on the principle of the Shapley value from the cooperative game theory.

In the literature of the regression-based decomposition developed by Fields (2003), a lot of authors have already worked with this model to capture the factor behind the rising inequality. Moreover, very often authors developed the Shapley decomposition method besides the Fields one. One of them, and the most important, is the paper from Li and Deng (2009). They have worked on a regression-based decomposition (Fields, 2003) as well as the Shapley decomposition to identify the factors behind the rising earning inequality. The authors have based their analysis on the Shapley value decomposition developed in the unpublished article of Shorrocks in 1999. Li and Deng (2009) have based their inequality analysis on the urban Chinese population. They concluded that, among all, the regional factor is the biggest contributor in the earning inequality. Moreover, it has been demonstrated that the contributions of ownership status and industry have increased from 1988 to 2002. On the other hand the contribution of the work experience is minimal while the contribution of the education and the occupation in earning inequality have risen. (Deng and Li 2009)

So far, all the papers talked about the inequality in the whole population of China. However, in the literature on income inequality in China, some authors focused on the gender gap in income inequality. This is the case in the papers of Gustafsson and Li (2000) and Chen and al. (2007).

In Gustafsson and Li's (2000) paper, the authors analyzed the evolution of the

gender wage gap in urban China between 1988 and 1995. They discovered that, while the average gender earning gap appears to be smaller in China than in other countries, it has increased between the two periods. In 1988, women used to earn on average 15,6 % less than men but it increased to 17,5 % in 1995. According to the authors, four main factors seem to affect earning in China: geographical location, ownership of the enterprise, age (experience) and education. Concerning the indisputable difference in earnings between men and women, they found it hard to say if it is due to discrimination or for productivity considerations, women being on average less productive than men. One of their key findings is that the increase in gender inequality has been particularly fast in areas strongly influenced by market forces. And this is especially true for young adults and people having followed a shorter education.

On the other hand, Demurger, Fournier and Chen (2007) investigated the impact of economic reforms and market liberalization between the mid-1980s and mid-1990s in China on the gender wage gap and how it affected discrimination against women. They found that the overall gender wage gap has decreased between 1988 and 1995, passing from 18,7 % to 17,7% (which contradicts Gustaffson and Li (2000)). Another of their finding is that the increased liberalization between 1988 and 1995 caused a rise in income inequality in Chinese corporations while it slightly decreased in foreign companies implemented in the country. They also show that, while gender inequality decreased in the higher deciles, it increased for low-wage workers. Concerning women discrimination, they identified that wage differential between men and women was also coming from the fact that while there were proportionally more men in firms paying higher wages, women were much more represented in low-paying jobs. Their findings concluded that the evolution of China's economy didn't benefit as much to women than to men.

In conclusion, there is a growing interest in inequality in China. A lot of authors have focused on the factors that have changed, contributed to and driven the inequality. As seen, some authors focused on wage inequality when others focused on income inequality. As mentioned at the beginning of this section, this thesis will focus on income inequality. Moreover, the population concerned in this analysis is the women urban population. Indeed, this thesis will aim at filling the gap in the literature, regarding women inequality. Even if a lot of papers focus on inequality in China in the whole population, only a tiny proportion focuses on the women population. This thesis aims also at filling the gap in the women inequality literature in China by analyzing the factors behind the increase/decrease of inequality, using a regression-based decomposition. To do so, some statistical concepts will be needed such as the Lorenz Curve, the Gini Index, the Concentration Curve, Concentration Index, the regression-based decomposition and Shapley decomposition. The next part will describe those concepts.

Chapter 3

Method

3.1 Data Description

This thesis relies on data from the National Bureau of Statistics in China. Indeed, the NBS has performed a survey through the country, called: the Chinese Household Income Project (CHIP). Their goal was to measure and estimate the distribution of personal income in both rural and urban areas of the people's republic of China. It exists five different waves of Chinese Household Income Project that have been conducted in 1989, 1996, 2003, 2008 and 2013. Those five waves cover information about expenditure and income and are called CHIP1988, CHIP 1995, CHIP2002, CHIP2007 and CHIP2013, respectively. All the CHIP waves contain at least two separate database about the survey of rural and urban households. However, as in China there was an increase of importance of rural-to-urban migration, from 2002 a third sample of rural-urban migrants has been added. Besides, from 2007, different sampling procedures and survey methods have been applied. The choice has, therefore, be made to select the most two recent CHIP databases that are as similar as possible. This was the case of the 1995 and 2002 CHIP. The two cover almost the same number of province and have been based on almost the same questions and sampling procedure.

The surveys have been carried out through questionnaires and to answer to questions, a face-to-face interview has been made for each household in the sample. As mentioned, for the year 1995 and 2002, it exists a distinction between the rural and the urban population. However, this thesis focuses on the urban samples as the definition of total income in both sample does not coincide. The questionnaire is made up of personal and income information. However, in the urban population, it exists different questions about the income for the whole year such as the wage income, transfer income and total income, for example. On the other hand, in the rural sample, the questions about the income are limited to the gross income by month (Li and al., 2008). It has, therefore, been decided to continue with the sample with the most information, namely the urban one.

The first database that this thesis will approach refers to the urban Chinese Household Income Project in 1995 and has been performed in spring 1996. The second and third refer to the urban Chinese Household Income Project in 2002 and have been also performed in the spring of the next year. There are, indeed, two datasets for urban China in 2002. One with personal information and the second with information regarding the different sources of income. The surveys account for more than 6931 and 6835 different urban households in 1995 and 2002 respectively. Within one household different individuals could be interviewed. Each original woman dataset is composed of 102, 151 and 44 variables respectively for the year 1995, 2002 for the personal information database and 2002 for the income source database. Moreover, the numbers of observations of the original women database are equal to 10936, 10407 and 10458 respectively. However, within the women data base, only the complete information regarding the income analysis will be used. The database for the income function has a total number of observations equal to 6343 and 2890 for 1995 and 2002 using the personal information dataset.

The sample used in the survey has been drawn from a larger one by the National Bureau of Statistics in its annual household survey. For the urban survey, the interviewed households are selected using a two-stage stratified systematic random sampling scheme. In the first stage, counties and cities are selected. Then, the households within the cities and towns selected are chosen. At the first stage, the cities and counties site selected can be described as follow: all the counties and cities sites are divided into 5 categories according to their population size. The different categories are: extremely large, large, medium-sized, small cities and counties site. Then, within the categories the counties site and towns are grouped by their geographical regions (north, east, center, northeast, northwest and southwest). In each geographical region, the counties and cities are divided according to their average wages of their staff and workers with urban hukou (registration). Then, the number of staff and workers in the cities are reckoned up. Finally, the sample of the cities and counties are created using an interval of one million staff and workers. At the second stage, a multiple sampling scheme is used to select the households in each of the samples. For the extra-large and large cities, the procedure is called the 'three-phase' sampling method, referring to the three stages of sampling that can be described as follow: the first stage consists of selecting the sample sub-districts in each city or county town. The second stage consists of selecting the sample resident committees from the sample of the first stage. The last stage consists of selecting the sample households from the sample of the second stage. For the medium-sized and small cities and counties, the sampling method occurs in two stages. The first stage consists of selecting the sample resident committees and then the sample households are selected from the sample of the first stage. The National Bureau of Statistics does not provide an explanation about how the sub-districts, resident committees and households are selected. It assumes that a more or less random selection method is performed (Li and al., 2008) (Li, 2007).

The main goal of the thesis is to quantify the impact of factors in their contribution to the inequality. But the question is: which factors? As the choice of the variables with two database having approximately 100 variables each has already been the subject of different papers, the factors will be chosen by looking through the literature. Indeed the choice of the factors will rely on the factors that have already found to be the most contributors to the income inequality in China. As seen in the literature, the *Region*, the *Province* and the *Sector of the company* have already been detected as leading factors for the income inequality (OECD, 2004). Moreover, it has also been demonstrated that factors such as *Ownership* type of the company and the *Industry* of the company have an impact in explaining the earnings inequality (Wang, 2007).

As mentioned in the literature review, the income is defined as the total gross income an individual has accumulated through the years 1995 and 2002. The total income is composed of the wage income, business income, transfer income and property income. One component of the income is thus the wage income. Interestingly, previous studies have analyzed the factors behind one single component of the income. It will be useful to compare the results. Therefore, those factors will be incorporated into the income function. As mentioned in the literature review, it has been found that *Education* rather than *Work Experience* has been the most important factor for the urban wages in China (Knight and Song, 2008). Moreover, factors such as the *Industry*, *Region* and *Ownership* have also been found as factors that play a role in the wage determination. (Chen and al, 2007). The *Occupation* and membership of the *Communist Party of China* (CPC) also appeared to have a significant impact on the income. Indeed, it has been shown that workers such as professionals, cadres and skills workers tend to have a higher wage than unskilled workers. (Knight and Song, 2003; Yueh, 2004). Moreover, in the paper of Knight and Song (2003) they have also shown that citizens in the *CPC* have a mean earnings 24% higher than citizens who are not in the CPC. Yueh (2004) Knight and Song (2003) Knight and Song (2008)

Thus, this thesis will be based on 7 factors that have to be found such as the most contributor to the inequality in China. Those variables are taken in the two tables below. This variable selection will allow to confirm or not that the conclusions for the whole population of China applies also for urban women in China. The concerning factors are, as summarized in the Table 3.1 and 3.2, the *Province* where the citizen lives, the membership or not to a *Communist party*, the citizen's type of *Education*, the *Work Experience* citizen has, the *Ownership* type of the company, the *Occupation* role of the worker in the company and the *Industry* or sector the worker is working in.

	Levels	Mean	Standard deviation
Income	1	5798.613	4031.461
Province	11	5.933	3.019
CPC	2	1.839981	0.366
Education	7	4.060	1.390
Work Experience	5	2.389	0.965
Ownership	9	2.130853	1.050
Occupation	9	6.094435	2.054
Industry	13	5.339	3.528

Table 3.1 : Descriptive Analysis 1995

	Levels	Mean	Standard deviation
Income	1	9601.180308	7423.4780133
Province	12	6.155363	3.3040899
CPC	4	3.453287	1.1060415
Education	9	5.242561	1.2442968
Work Experience	4	2.303806	0.9052094
Ownership	11	4.292388	3.0165077
Occupation	10	6.569550	2.4120759
Industry	16	6.825952	3.7306999

Table 3.2 : Descriptive Analysis 2002

The two tables above give an insight into the levels of each variable. These levels represent the number of different categories of response for each variable. Indeed, let's take an example. The fourth variable represents the *Education*. One question of the questionnaire has been, therefore, to ask which is the higher degree the citizen has obtained. The answers have been coded as number representing each time one type of degree, for example, secondary or primary. Table 3.1 and 3.2 incorporate also the mean and standard deviation for each variable. The standard deviation tells how numbers are spread out and in that instance, how high the standard deviation is for the variable income in 1995 and 2002. However, as the variables are coded as factor, it does not make sense to interpret their means or standard deviations. Table 3.1 and 3.2 are there to highlight the variables for the two years and compare the difference in levels for each variable.

The first variable relates to the *Province* of the individual. The variable covers 11 provinces (Shanxi, Jiangsu, Beijing, Liaoning, Hunan, Anhui, Guangdong, Sichuan, Hubei, Gansu and Yunnan) for the year 1995 and 12 provinces for the year 2002 because it includes Chongqing as a separate province. In the year 1995 this province was included in the Sichuan province. As seen in Appendix 2, the sample is composed by 8 provinces in central China (Shanxi, Anhui, Henan, Hubei, Chongqing, Yunnan, Gansu and Sichuan) and 4 others from coastal China (Beijing, Liaoning, Jiangsu and Guagdong).

The *Communist of Party of China* (CPC) is the categorical variables gathering the information about the fact that the citizen was part of a CPC or not. However, in 2002, 2 levels are added to the categorical variable to know if citizens are part of other parties or are part of the communist youth league. The *Work Experience* variable accounts for the number of years a citizen has already worked. At the origin, the variable is a continuous one but for a matter of interpretation, the variable has been transformed into a categorical one. For every 10 years of experience one dummy has been created. It appeared that in 1995 there were people with more than 40 years of experience in the sample although the same is not observed in 2002. Moreover, the *Education* variable reveals the type of education the citizen has had. On the other hand, the *Ownership* variable refers to the type of ownership of the company the citizen is working in. The *Occupation* variable refers to the position of the citizen in his work, for example, skilled worker. Finally, the *Industry* variable aims to distinguish the sector the citizen is working in.

To add clarity, the descriptions of each level for each variable are listed in Appendix 3.1 for 1995 and Appendix 3.2 for 2002. Both tables have the same variables, however, as said below, the levels in each variable differ from one year to another year. Moreover, in Appendix 3.1 and 3.2, for each level into each variable, the proportion of the population this level represents has been calculated as well as the mean income for each level.

The interesting thing with Appendix 3.1 and 3.2 is that it makes it able to compare the mean income of each level within one variable. For example, in the *Province* variable, the Guangdong province is the one with the highest income in 1995 and 2002. Furthermore, when analyzing the income of the province regarding their location in the country, it is possible to detect higher income in the coastal provinces such as Guangdong, Liaoning and Jiangsu and lower income in central provinces such as Shanxi, Henan, Hubei and Anhui in 1995. Nevertheless, in 2002, even if Guangdong is again the province with a higher income, this higher trend in the coastal province is less evident to detect.

About the mean income for the *CPC* categorical variable, it is interesting to see that the CPC members have a higher average income than the non-CPC members in both years even if the CPC members variables count less individual than the CPC non-members. Regarding the *Work Experience*, the mean income is increasing with the number of years of experience. Indeed, the > 40 level and the [31:40] level have the highest average income in 1995 and 2002 respectively. In the case of the *Education*, the level with the higher mean income is located in the levels which reflect a higher education. Although these tables, it seems that the more educated a Chinese woman is, the higher her mean income is. It would suggest a positive linear correlation between income and education. Indeed, the higher income is located in the "College or above" level in 1995 and in the "College/University" and "Graduate" in 2002.

Figure 3.1: Correlation Matrix 1995 (left) 2002 (right)

Now, regarding the *Ownership* of the company, it is in the foreign-owned enterprise that the income is the highest in both years. Note that the State-Owned Enterprise (SOE) directed by the central government and the SOE controlled by the local government have kept the same magnitude in terms of income. This pattern highlights the increase of monopoly power by the SOE governed by the central government. Indeed, the central SOE has a higher average income in both years. Moreover, the lowest mean income is located in the individual businesses for both years. Besides, this descriptive analysis has allowed to highlight the fact that in the women database, there are no women that work in a company of the type "rural private enterprise" in 2002.

About the *Occupation* factor, the individuals that have the highest income are the "head of institution" and "director". This is a trend that is remarked in both years. On the other hand, the people that have the lowest income are the "unskilled worker" in 1995 following by the "Salesclerk or Service worker" in 2002. Moreover, it appears that, in the database, no women have a position of "farmer" in 2002. Finally, regarding the sector in which the workers are working, it seems that the highest income is located in the "Finance and Insurance" area in 1995 and in the "Scientific Research and Professional Services" area in 2002.

As already mentioned, the total income used in this thesis is composed by the wage income, property income, business income and transfer income. In 1995, two other categories are part of the income: the labor income and the category "other income". Table 3.1 represents the Pearson correlation matrix for the composition of income in 1995 and 2002. When analyzing the correlation between the total income and the different sources of income, it is with the wage income that the coefficient of correlation is the highest. In 1995, the category "other income" has also a high correlation with the total income. On the other hand, this is with the business income that the correlation is the lowest in both years.

The correlations between the income sources are, in general, quite low. However,

in 1995, a correlation of about 0.40, between the labor income and property income, is observed. Such a relation does not hold anymore in 2002. Following these observations, a Pearson coefficient test is performed to know to which extend those coefficients are reliable.

	Income	Wage	Business	Labor	Property	Transfer	Others
Income		0.00	0.00	0.02	0.00	0.00	0.00
Wage	0.00		0.00	0.00	0.00	0.00	0.00
Business	0.00	0.00		0.74	0.45	0.05	0.02
Labor	0.02	0.00	0.74		0.00	0.00	0.26
Property	0.00	0.00	0.45	0.00		0.01	0.00
Transfer	0.00	0.00	0.05	0.00	0.01		0.70
Others	0.00	0.00	0.02	0.26	0.00	0.70	

Table 3.3 : P-value coefficient pearson 1995

The Pearson's significance test confirms that the correlations between the income and wage are significantly different from 0, at significant level of 0.05. This conclusion holds for both years. Moreover, the Pearson correlations between the income and sources of income are also all significantly different from zero, at a 5% significant level. However, it is interesting to see that the correlations with the business income source are, in most of the case, not significant, at a 1% significant level. Moreover, in general, the correlations between the income source are significant, at a 5% significant level.

	Income	Wage	Business	Property	Transfer
Income		0.00	0.00	0.00	0.00
Wage	0.00		0.00	0.00	0.00
Business	0.00	0.00		0.32	0.00
Property	0.00	0.00	0.32		0.00
Transfer	0.00	0.00	0.00	0.00	

Table 3.4 : P-value coefficient pearson 2002

3.2 Concepts

3.2.1 Lorenz Curve

The Lorenz Curve has been introduced in 1905, as being a powerful tool for illustrating the inequality in wealth distribution. The Lorenz Curve (Lorenz, 1905) is very often used to analyze and represent the size distribution of income. The curve corresponds to the cumulative proportion of income units to the cumulative proportion of income with units in an ascending order. The Lorenz Curve can also be obtained by the density function of the income. However, the distribution of income is not known. In the literature one common approach is to fit a well-known density function such as the Pareto or lognormal distribution. However, this approach will not be performed in this thesis because the downside of this method is that it hardly gives a reasonably good fit

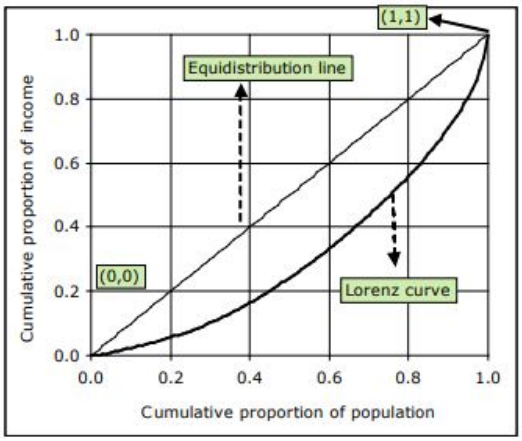


Figure 3.2: Lorenz Curve. Note: Bellù, G.L. and Liberati,P. *Charting income inequality : The Lorenz Curve* (2005)

to the actual data. (Bellù and Liberati, 2005)

Figure 3.2 is a graphical representation of the Lorenz Curve. The vertical axis represents the cumulative income and the horizontal axis represents the population percentile by income. The 45° line on the graph represents "the line of perfect equality", meaning that it reflects a perfectly equal income distribution. One this line every person has the same income. This equality line is represented by the linear equation: y=x. By contrast, the vertical right-hand side line, as seen in the figure, represents the perfectly unequal distribution. This line represents the case where one person has all the income when everyone else has nothing. This line is expressed by the equation y=0. Finally, the curve represents the actual Lorenz Curve. This curve is obtained thanks to the actual distribution of the population. As said before, the Lorenz Curve can be either the line of perfect equality or perfect inequality or in between. (Bellù and Liberati, 2005)

More formally, let's denote X being the income variable that we are interested in. The Cumulative distribution function of X will be $F_X(x) = Pr(X \leqslant x)$ as well the quantile function will be represented as $Q_X(p) = F_X^{-1}(p) = \inf\{x : F_X(x) \geq p\}$ with p $\in$ [0,1]. The x-axis is the cumulative percentile of population ranked by income which can be represented as $\int_{-\infty}^{\infty} x dF_X(x)$. However, the y-axis represents the cumulative income which is defined as $\int_{Q_X^p}^{-\infty} y dF_X(x)$. In the case X is continuous, the Lorenz Curve ordinates can be thus defined as follow (Lambert, 2001) (Hao and Naiman, 2010) :

$$L_X(p) = \frac{\int_{Q_X^p}^{-\infty} y dF_X(x)}{\int_{-\infty}^{\infty} x dF_X(x)} \tag{3.1}$$

Let's take the $L_X(p)$ equation in the finite population form with I(B) representing the binomial function equal to 1 if 1 is true and 0 otherwise such that:

$$L_X(p) = \frac{\sum_{i=1}^{N} X_i I(X_i \leq Q_X(p))}{\sum_{i=1}^{N} X_i} \tag{3.2}$$

Under this form, it is easier to see that each point on the Lorenz Curve represents for each px100 percent of the population, the proportion of total income.

3.2.2 Gini Index

The Gini Index is calculated based on a graphical representation of the income distribution which is the Lorenz Curve. That is why the Lorenz Curve can be also viewed as the logical representation of the Gini Index. Concretely, the Gini Index represents twice the area between the Lorenz Curve and the diagonal. In other words, the Gini Index is a measure of deviation from the absolute equality line. The further the Lorenz Curve is from the absolute equality line, the higher the Gini Index goes. (Arcagni and Porro, 2014)

Indeed, the Gini Index is a coefficient that represents how unequally the distribution of the income is. The index ranges from 0 to 1. The lowest value (0) means that every person in the country has the same income. On the other hand, an index value of 1 is an income distribution where there is high inequality, meaning that one person earns all the income when everyone else earns nothing. It is good to notice that high-income country can have the same Gini Index as low-income country, as long as their income is distributed similarly within the country. (Bellù and Liberati, 2006)

Taking the function F as the cumulative distribution function (CDF) of the income in question and $\mu \equiv \int_0^{\infty} p dF(p)$ is the mean income, the Lorenz Curve can be also defined such as (Arcagni and Porro, 2014) :

$$L(F(x)) = \frac{1}{\mu} \int_0^x p dF(p) \tag{3.3}$$

In this definition, it assumes no negative incomes. From this equation, the Gini Index can be defined as the area below the Lorenz Curves such that:

$$G = 1 - 2 \int_0^1 L(p) d(p) \tag{3.4}$$

where L(p) is the Lorenz Curve. From this geometrical approach, it is easier to understand that the Gini Index is double the area between the Lorenz Curve and the equality line.

Thanks to the concept of mean difference, the geometrical approach described above can be related to a more statistical approach such that (Jedrzejczak, 2008) :

$$G = \frac{\Delta}{2\mu_y} \tag{3.5}$$

where the Δ is the Gini mean difference and can be understood such as "the sum of all possible absolute differences in a population of income receivers." (Jedrzejczak, 2008) and can be expressed by:

$$\Delta = \int_0^\infty \int_0^\infty |x - y| \, dF(x) dF(y) \tag{3.6}$$

where x and y represent two income variables identically distributed. F(x) and F(y) are the corresponding Cumulative distribution function.

Now, let's take the Eq. (3.4) integrated by parts such that :

$$G = -1 + 2 \int_0^1 pL'(p)dp \tag{3.7}$$

and replace the variables by substitution such that F(y) = p. It becomes :

$$G = -1 + 2 \int_0^\infty \frac{yF(y)f(y)dy}{\mu_y} \tag{3.8}$$

with F(y) being the cumulative distribution function, f(y) being the density function and μ_y being the mean income.

Now, let's define the covariance between two random variables with X = income y and Z = F(y). (Jedrzejczak, 2008) That gives :

$$Cov(y, F(y)) = \int_0^\infty yF(y)f(y)du - \frac{\mu_y}{2} \tag{3.9}$$

Moreover, by combining the Eq. 3.8 and 3.9, the definition of the Gini Index can be expressed such as :

$$G = 2\frac{cov[y, F(y)]}{\mu_y} \tag{3.10}$$

The Gini Index can be thus expressed as the covariance between incomes and their cumulative distribution of income. (Bellù and Liberati, 2006)

3.2.3 Davidson Method

Measuring the inequality and compare it through times has been the subject of a lot of papers these years. However, up to now, no parameters have been defined to compare statistically the difference between two Gini Indexes. Indeed, even if from empirical example, it is possible to evaluate two Gini Indexes from two different years, it is impossible to assess if the differences of the Gini Indexes are statistically significant. That is why a test is needed and for that suitable estimator and standard deviation are needed.

Numbers of authors have worked on different ways to evaluate the standard deviation and the estimator of the Gini Index. More specifically, attention has been given to a plug-in formula for the estimation of the standard error without assumptions about the distribution. Some of them have based their variance estimation on the U-statistics (Xu,2007). When, on the other hand, Ogwang (2000) has developed an OLS regression procedure in order to evaluate the Gini Index and analyze a way to use the regression, in order to make easier the computation of the Jackknife standard error. In response to this paper, Giles (2004) has claimed that it was possible to use directly the standard error obtained by the OLS regression for the standard error of the Gini Index. The Jackknife method is among the numerous "computationally intensive" method to evaluate the standard error for the Gini Index and has been proposed by Gastwirth and Modarres (2006). Another way to compute the variance estimator of the Gini Index is the linearization method. This approach uses different kind of techniques to approximate the variance of a non-linear statistics (here, the Gini Index). Indeed, in this method, the development of the first order of the Taylor expansion around a parameter is used. Such a method has been developed by Berger (2008), Davidson (2009), Langel and Tillé (2013), and Arcagni and Porro (2014).

However, it has been proved from empirical analysis that the standard error from the Jackknife method was not a reliable estimator and therefore does not give reliable inference (Davidson, 2009). Moreover, in 2017 Borzadaran and al. have proved that the linearization method developed by Davidson performed better than the Gini estimator based on the U-statistics. Therefore, this thesis will rely on the easy to compute formula of the standard error and unbiased estimator of the Gini Index developed by Davidson (2009).

This section will thus discuss the linearization method developed by Davidson (2009). Indeed, some years ago, Davidson has presented a procedure for the computation of an asymptotic correct standard error for the Gini Index. The standard error developed is based on a simple formula, that is easy to compute. Besides, the computation is based on the approximation of the Gini Index estimator $\hat{G}$ which is a non-linear statistic. The approximation is expressed as the sum of i.i.d linearized weighted variable V_k such that :

$$\hat{G} - \hat{G} \approx \sum_{k=1}^{n} u_k V_k \tag{3.11}$$

Then, the variance of $\hat{G}$ is approximated by the variance of the normalized sum of a set of i.i.d random variables. Moreover, Davidson (2009) has shown, based on linearization techniques, that $\sqrt{n}(\hat{G} - \hat{G})$ can be approximated by the normalized sum of a set of i.i.d random variables of expectation zero such as (Amini Dehak, 2017) :

$$\sqrt{n}(\hat{G} - \hat{G}) \approx n^{-1/2}\frac{2}{\mu}[\sum_{i=1}^{n} -\frac{E(XF(X))}{\mu} + X_i F(X_i) - E(XI_{|X \leqslant X_i|} - (2E(XF(X)) - \mu)] \tag{3.12}$$

The variance of the Gini Index estimator is thus approximated such as :

$$\hat{\sigma}^2_{\sqrt{n}(\hat{G}-G)} = \frac{1}{(n\hat{\mu})^2}\sum(\hat{Z}_i - \hat{Z})^2 \tag{3.13}$$

where :

- $\sqrt{n}(\hat{G} - G) = \hat{I}_G$ is defined as the bias corrected estimate of the Gini Index which is computed such as : $\frac{n(2\hat{I}/\hat{\mu}-1)}{(n-1)}$ with $\hat{I} = \bar{w}$. $\bar{w}$ is defined such as $w_i \equiv (2i-1)y_{(i)}/2n$. $y_{(i)}$ being the series under interest sorted in a increasing order.

and

- $\hat{Z} = (1/n)\sum_{i=1}^{n}\hat{Z}_i$ is an estimate of $E(Z_i)$ with $\hat{Z}_i = -(\hat{I_G} + 1)y_{(i)} + 2(w_i - v_i)$ and v_i calculated such as $v_i = n^{-1}\sum_{j=1}^{i} y_{(j)}$. Finally, the parameter "n" represents the length of the series in question where $\hat{\mu}$ represents the sample mean. (Davidson, 2009)

The paper of Davidson contributes to the Gini Index literature in different ways. Among others, he developed a bias-corrected estimator for the Gini Index. He derives also a concrete formula for an approximation of the standard error. Besides, he defined the standard error based on the expression of the Gini coefficient expressed as a sum of independent and identically distributed random (iid) variables. Indeed in his paper, Davidson (2009) shows that the plug-in estimate for the Gini Index is asymptotically normal.

Giving two Gini Indexes for two populations, it is often interesting to test whether those two indices are the same or not. In the case of independence of both sample, the unbiased corrected estimators $\tilde{G_1}$ and $\tilde{G_2}$ and the corresponding variances $\sigma_{\hat{G_1}}$ and $\sigma_{\hat{G_2}}$ can be computed. The statistics defined such as :

$$tstat = (\tilde{G_1} - \tilde{G_2})/\sqrt{(\sigma_{\hat{G_1}})^2 + (\sigma_{\hat{G_2}})^2} \tag{3.14}$$

can be thus used for a suitable test in order to assess either or not the means of both population are equals.

3.2.4 Concentration Curve

The Concentration Curve and the Lorenz Curve are two main tools to analyze redistribution and economic inequality. As explained above, the Lorenz Curve provides a degree of inequality from a scale-free viewpoint. Moreover, the change of measurement units will not affect the distribution of the Lorenz Curve and the inequality is independent of the overall outcome level. On the other hand, the Concentration Curve analyses how a covariate can change the shape of the income distribution. The Concentration Curve (Kakwani and al, 1997) can also give an insight into how one variable is distributed across different defined groups in terms of another variable. Initially, the Concentration Curve was used to identify the socioeconomic inequality in the health sector. For example, the Concentration Curves can help to identify whether subsidies in the health sector are better targeted across the poorest or the richest population (Jann, 2016).

The Concentration Curve is based on two key variables. These are, for example, the income variable and another one, being the source of income. The income variable is used to assess the distribution of the subject of interest and the other variable is the one against which the distribution has to be assessed. For this variant of the Lorenz Curve, the analysis can be realized using grouped data, in that case and for each group, the mean value of the type of income in this example needs to be observed. The analysis can be also done using data for each individual, so both variables need to be observed at an individual level, which will be the case in this analysis. In other words, when, for the Lorenz Curve, only the outcome variable X was needed, in this case, another variable Y is used. It follows that the Concentration Curves will represent the cumulative outcome proportions of population members but this time ranked by the variable Y. (Bilger, M. and al., 2011)

To understand the concept, Figure 3.3 represents Concentration Curves, where the distribution to be assessed is the income distribution and the covariates are the transfers income and the capital income. Indeed, this example is based on a household data set, including information on the earnings, capital income and transfer income. The goal is to see how the transfer and capital income are distributed through the households distributions by ranking the households by earnings. In the graph, the vertical axis represents the cumulative percentage of outcome (transfers and capincome) and the x-axis represents the population percentage, ranking by an ascendant order of earnings. The diagonal, or also called, line of equality on the graph, represents the case where everyone has the same value of the outcome and this independently of his living standards. The Concentration Curve can lie above or below the line of equality. In the case of the Concentration Curve is above (below) the diagonal, it means that the outcome variables take higher (lower) value for the poorest people. It follows that the further away above the diagonal the curve is, the more are the outcome variables concentrated among the

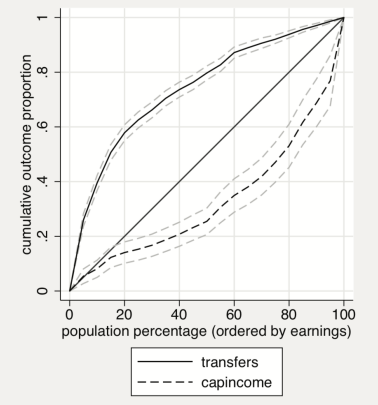


Figure 3.3: Concentration Curves. Note Jann,B. *Estimating Lorenz and concentration curves* - The stata Journal (2016)

poorest variable.

Comparing different curves relative to the same variable is also really meaningful. Indeed, when it comes to compare different curves, the notion of dominance can be used. In the case where a curve (A) lies everywhere closer to the diagonal compared to another curve (B), the curve B is said to be dominated by the curve A, implying that it is possible to rank the curve by degree of inequality. There is non-dominance in the case where Concentration Curves cross (Jann, 2016). In the example above, the transfer income takes higher value for the poorest population. On the other hand, the capital income takes lower value for the poorest population. Leading to the conclusion that the poorest have more transfer income than capital income.

From a mathematics point of view (Bishop and al., 1994) let's denote $Q_Y(p)$ the quantile p of the distribution Y. Moreover, let's denote $f_{XY}(x, y)$ the density of the joint distribution of X and Y. The Concentration Curve of X with respect to Y can be thus represented as :

$$L_{XY}(p) = \frac{\int_{-\infty}^{Q_Y(p)} \int_{-\infty}^{\infty} x f_{XY}(x, y) dx dy}{\int_{-\infty}^{\infty} x dF_X(x)} \tag{3.15}$$

As well as the Lorenz Curve, the Concentration Curve can be re-write in the case of a finite population such that :

$$\frac{\sum_{i=1}^{N} X_i I(Y_i \leqslant Q_Y(p))}{\sum_{i=1}^{N} X_i} \tag{3.16}$$

3.2.5 Concentration Index

The Concentration Index is calculated on the basis of the Concentration Curve explained above. The index represents twice the area between the equality line and the Concentration Curve. The index range from -1 and 1. More formally, the Concentration Index (Kakwani and al, 1997) can be defined as

$$C = 1 - 2\int_0^1 L_{XY}(p)dp \tag{3.17}$$

Let's take the example of the income distribution and the covariates being the wage income, transfer income and property income. In the case where the Concentration Curves lie on the diagonal, the Concentration Indexes are equal to zero, meaning that there are no inequality in the income distribution regarding the wage, transfer and property income. If the Concentration Curves are located above the equality lines, the Concentration Index will be negative, indicating that the income is less concentrated among the people with higher wage income, for example. On the other hand, if the Concentration Curves lie below the line of equality, the Concentration Index will be positive, meaning that the income will be more concentrated among the people with higher wage income, for example. In absolute value, that means that the larger the Concentration Index is the greater the inequality in income there is. (Bilger and al, 2011)

3.3 Models

This thesis aims at analyzing the effects of different pre-selected factors on the dependent variable. The evaluation of the exact contributions of each explanatory variable on the variance of the dependent variable is an important issue in applied economics. In this part, the so-called regression-based decomposition and Shapley decomposition will be developed.

3.3.1 Regression-based decomposition

This thesis aims at analyzing the factors that drive income inequality in China. To resolve this problem, Fields (2003), Morduch and Sicular (2002), Shorrocks (1999) and Bourguignon et al. (2001) have developed a regression-based decomposition. All these authors have developed their ideas in different ways and used different equations, but all had the same goal which is to estimate the effect of factors that contribute to income inequality. However, in this paper, the method of Fields (2003) only will be discussed.

In the regression-based decomposition of Fields (2003), the type of the explanatory variables does not matter, which led to a certain flexibility for the analysis. Moreover, Fields (2003) based his method on the estimation of the semilog income function such that :

$$ln(Y) = \sum_{j=1}^{k} \beta_j X_j + \epsilon \tag{3.18}$$

where Y is the series of income in question, X_j the explanatory variables including, for example, the education, experience, occupation and regions. β_j stands for the coefficient of the explanatory variables and ϵ the disturbance term. Fields (2003). Moreover, the variance of the dependant variable can be decomposed such as :

$$VAR(Y) = \sum_{j=1}^{k} Cov(\beta_j X_j, Y) + Cov(\epsilon, Y) \tag{3.19}$$

Based on the logarithm of income, Fields (2003) has developed the contribution of each factor such as :

$$s_{jFields} = \frac{\widehat{\beta}_j cov(X_j, ln(Y))}{\sigma^2(ln(Y))} \tag{3.20}$$

which is the share of the log variance of income that is attributable to the j-th explanatory variable factor where $\hat{\beta}_j$ is the coefficient of the j-th explanatory factor estimated by a multiple OLS regression, $\sigma^2(ln(y))$ is the variance of the dependent variable and $cov(X_j, ln(y))$ is the covariance between the j-th factor and the dependant variable. The sign of the s_j coefficient will indicate if the contribution of the x_j factor is inequality increasing ($s_j > 0$) or decreasing ($s_j < 0$). However, Fields has also shown that :

$$\sum_{j=1}^{k} s_{jFields} = \frac{\sum_{j=1}^{k} \widehat{\beta}_j cov(X_j, ln(y))}{\sigma^2(ln(y))} = \frac{\sigma^2(ln(\widehat{y}))}{\sigma^2(ln(y))} = R^2 \tag{3.21}$$

Notice that the contributions of the different factors to the linear regression R-squared result also to the percentage contribution to the variance of the dependant variable. In the case the error term of the regression is considered $s_\epsilon = 1 - R^2$, which means that its inequality contribution is equal to the proportion of the inequality, that is not explained by all the variables included in the regression. One drawback of the Fields method is the assumption of functional form for the income-generating function that must be log-linear. Moreover, the Fields (2003) method implies that the constant has no effect on income inequality. Indeed, for the inequality decomposition the β_0 estimate is not meaningful, as it is the same for every observation.

3.3.2 Shaley Decomposition

Some years later, Shorrocks (2013) has introduced another method for the regression-based decomposition using the Shapley value. The Shapley value has been initially used for the cooperative game theory, in order to allocate the costs to each player. Here,

Shorrocks (2013) applied this concept to the income inequality in the sense that the Shapley value is used to determine the exact contribution of each factor to the income inequality.

As just mentioned, Fields (2003) has introduced a way to decompose the variance of the log income into the different components of the regression. Moreover, he showed that if the contribution of the disturbance term is omitted, it is the R-squared that is obtained. However, this method fails to take into account the probable multicollinearity in the regression. Indeed, in the case where two variables are strongly correlated, it is difficult to distinguish their effects. On the other hand, the Shapley decomposition procedure is based on the marginal effect of each variable in the model. It assumes that the marginal effect of the factor should be equal to the contribution of this factor.

The Shapley decomposition approach is, therefore, a way to calculate, for a linear regression, the exact contribution of each explanatory variable to the R-squared. Note that, the contribution of each variable to the R-square is expressed as the percentage of contribution of these variables to the variance of the dependent variable. So, in the case the dependent variable represents the income, this is in reality, the contribution of each explanatory variable in the overall income inequality which is calculated. (Israeli, 2007)

The Shapley decomposition method works such as each marginal contribution of the factor on the R-Squared is calculated by eliminating the factor from the income function. Following this reasoning, the contribution of a factor k (L_k) is calculated such as (Shorrocks, 2013):

$$L_k = R^2[Y = a + \sum_{j \in S} \beta_j X_j + \beta_k X_k + \epsilon] - R^2[Y = a^* + \sum_{j \in S} \beta_j^* X_j + \epsilon^*] \tag{3.22}$$

where S is a subset of the explanatory variables excluding the variable k. The star symbol has been added to the right-hand side of the equation to symbolize the fact that when a variable k is excluded from the equation, the coefficients and the residuals change. Moreover, very often $\beta_j \neq \beta_j^*$,by eliminating this way, it results that the marginal effect takes into account the particular variable effect on the other variables. Notice that the marginal effect changes in the order you eliminate the variable. Thus, in order to avoid this effect, it is the average over all J! possible elimination sequence that is calculated for the contribution of each variable.

The Shapley decomposition and Fields's decomposition method will coincide in the rare case where there is no correlation at all between the explanatory variables. As mentioned before, the Fields method does not take into account the possible correlation between variables and the special case of multicollinearity in the model. Nevertheless, with multicollinearity, the estimated variance tends to be large, which can result in a different coefficient from the population one. This is a serious issue using the Fields

method. With the Shapley decomposition, it is the marginal effect from all combination of a variable which is evaluated, which means that the effect for this variable will be high or low, depending if the other correlated variable is included or not in the model. This procedure expects a similar Shapley value for two highly correlated variable, while the Fields decomposition might result in an inverted signs coefficient for those two variables.

As a reminder, this thesis will contribute to the literature by filling a gap in the Chinese women inequality literature. Indeed, the question is: Are the patterns observed for the whole urban population the same as for women? Are the patterns behind the rising urban women inequality the same as for the whole urban population? In order to find the answers, this thesis will use the regression-based decomposition developed by Fields (2003) and the Shapley decomposition method (Shorrocks, 2013). Moreover, to give more statistical value to the results, the next section introduces the bootstrap method.

3.3.3 Bootstrap

The purpose of this work is to analyze the factors behind the increasing inequality from 1995 compared to 2002, with the notion of statistical significance. A bootstrap based algorithm applied to the sample statistics of the Shapley value is introduced in this thesis and will be used to perform different tests. These tests will allow to determine if changes in coefficient are either significant or not. The bootstrap method is among the re-sampling techniques. Indeed, it uses re-sampling techniques to have information about spread, shape and center of the distribution of interest. This method will be thus used to test the null hypothesis that two Shapley values from 2 different years are equals.

Firstly introduced by Efron in 1979 the bootstrap method is still commonly used to estimate confidence intervals (Efron, 1979). The re-sampling method such as bootstrap is usually used for the computation of the Shapley value's or Fields's confidence interval (Israeli, 2006) (Huettner and Sunder, 2011). The re-sampling method works as follow: it considers an observed sample as a finite population and then generates random samples from it to have an estimation of the population characteristics such as mean, shape and spread. Therefore, it allows to make inferences about the sample population. (Booth and al., 1993)

The Monte Carlo method estimates the distribution of a population by re-sampling. The bootstrap method is one class of the Monte Carlo technique. Moreover, it exists two types of bootstrap method, the parametric or non-parametric one. In the case of Monte-Carlo simulations, it requires sampling from a known probability function. In this case, this is called a parametric bootstrap. In the case the distribution is not specified, the pseudo population is used. The pseudo population is the sample that represents the distribution of the finite population. (Boos, 2003)

More formally it works as followed (Mills, 1997), suppose S, the parameter that needs to be evaluated, and $\hat{S}$ the estimator of S. The bootstrap estimate of $\hat{S}$ is thus obtained such as:

1. An estimate of $\hat{S}$ and a sample $X_1, ..., X_n$ of size n are considered
2. Bootstrap sample of size n with replacement from $X_1, ..., X_n$ are drawn
3. The estimator for each R is calculated, R estimators are thus obtained such that $\hat{S}_1^*, ..., \hat{S}_R^*$

These estimators are then used to calculate the variance of the original estimator. Thus, it means that the sample variance of the estimators is then used to estimate the variance of the original statistic via bootstrap variance estimation. Let's consider $\bar{S}^* = \frac{1}{R}\sum_{r=1}^{R} \hat{S}_r^*$. The bootstrap variance of $\hat{S}$ is thus estimated such as:

$$\hat{\sigma^2}_{Boots} = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{S}_r^* - \bar{S}^*) \tag{3.23}$$

Thanks to this variance, it will be possible to construct a confidence interval for each statistic. Indeed, here, the principle has been shown using one statistic but the same reasoning can be used for multiple statistics. Moreover, as said at the beginning of this part, this variance will help to test the null hypothesis of equal means.

3.4 Model Validation

As seen in the regression-based decomposition (Fields, 2003), a log-linear relation is assumed for the income regression. This assumption will be maintained for the Shapley decomposition. Moreover, the income function will be based on pre-selected factors. Before applying the methods to the model, a model validation will performed. The model validation will allow to know how much of the inequality are explained by the factors. Besides, the model validation approach yields to determine if the model will function sufficiently well in its intended operating environment. Model validation is the surest method to determine if the predicted values from the model will accurately predict the values that were not incorporated into the model. Model validation is commonly used to choose the best model among different ones.

In the case of multivariate regression, the R-squared is a good measure for the model predictive power. It is also very convenient for the comparison among different models. Unfortunately, the R-squared is biased. The index is, indeed, very influenced by the number of explanatory variables in the model. However, the adjusted R-squared is not

biased and solves thus this problem. Nevertheless, it has been proven that method such as data-splitting, cross-validation and bootstrap leads to unbiased R-squared (Frank and Harrell, 2001).

As mentioned, data splitting is one of the model validation and is the simplest one. It requires two data sets: training data and testing data. The training data is there to develop the model and the testing data set to validate the model. In order to appraise the model, the MSE is commonly used as well as the R-squared of the testing model. The MSE is the mean squared error and is calculated such as :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \tag{3.24}$$

with n being the number of observations, $\widehat{Y}_i$ the dependant variable for each observation i estimated, thanks to the train set and Y_i the dependant variable for each observation i in the test set (Frank and Harrell, 2001). Moreover, a drop of the R-squared between the training data set and testing data set indicates overfitting.

The data-splitting method has the capacity to test the model and evaluate his performance with the test data set. However, data-splitting has several disadvantages. Indeed, the estimate of predictive accuracy differs regarding the way data have been split. Besides, to have the same precision of accuracy than cross-validation, data-splitting requires a larger sample. Moreover, the data-splitting method validates the test model which is not the final model. Indeed, in the end, this is the combination of the test model and training model that are combined and this model is not validated. Finally, the data-splitting method requires a split before the first analysis of the data. On the other hand, with other methods, this is possible to carry out the analysis with the whole data set. (Frank and Harrell, 2001)

Hopefully, the cross-validation method is a category of data-splitting that solves some of the problems listed above. More specifically, the leave-one-out cross-validation is a way to execute the cross-validation. In this case, the predictions are made using n-1 observations to predict the omitted observation. This operation is executed n times to obtain an average accuracy. However, Efron (1983) has claimed that grouped cross-validation gave more accurate results. This method is commonly called K-fold cross-validation. As mentioned, in that case, it is a group of k observations that is omitted at a time. (Cook and Picard, 1984).

The K-folds cross-validation divides the data set into k subsets. At each step, the first k-1 folds are used to develop the model and the remaining folds are used to assess the accuracy. This procedure is repeated at least k times to get an average of k indexes such as R2. That implies that, at each step, one of the k subsets is used as the test set and the other k-1 set is used as the training set. In the end, the average error is computed. The fact that in this method, the data are divided into k folds avoids bias

due to the way how the data are divided. In this thesis, k will be equal to 10 because it has been proven relatively good properties compared to other values. (Bunke and Droge, 1984)

As said at the beginning, methods used in this thesis will be based on a model of pre-selected factors. The model validation will therefore not be used to select the best models but to determine for one model the prediction and explanatory power of the inequality. Indeed, as explained, with the cross-validation method, R-squared is more accurately assessed. Besides, as this thesis evaluates two models with the same pre-selected factors for two different years, the cross-validation approach will be used to assess how the predictive and explanatory power changes from 1995 to 2002.

Chapter 4

Analysis and Results

4.1 Lorenz Curve and Gini Index

According to the literature about China inequality, income inequality in China has increased through the years. The first step of the analysis has been thus to confirm this hypothesis for the urban women and for the years 1995 and 2002. This has been based on the Lorenz Curve and the Gini Index. Moreover, a suitable test for the comparison of the two indices has been performed. From a technical point of view, the Lorenz Curves have been evaluated in R using the function "LC" from the package *ineq*. The Gini Indexes have been evaluated thanks to the function "Gini" from the *Desctools* package.

Figure 4.1 represents the Lorenz Curve for the Chinese urban Women in 1995 as well as in 2002. The difference between the two curves is small but the Lorenz Curve in 1995 is closer to the equality line than the Lorenz Curve in 2002. In order to have an idea about the difference between the two years, the unbiased corrected estimator of the Gini Indexes have been computed. Then, thanks to the methodology developed by Davidson (2009), the standard deviation of both indexes and the confidence intervals have been calculated. All those results have been incorporated in Table 4.1.

	1995	**2002**
Gini Index	0.349671	0.3868482
Standard deviation	0.005818	0.006441
Confidence Interval	[0.3382679 : 0.3610741]	[0.3754451 : 0.3994726]

Table 4.1 : Gini Index (per year)

When looking at the Gini Indexes, the income inequality seems to have increased in 2002 compared to 1995. Indeed, the unbiased estimator of the Gini Index equals 0.35 when it equals 0.39 in 2002. However, until now, there is no proof that both Gini Indexes are statistically different from each other. To fill this gap, the standard errors and the confidence intervals have been computed. The confidence intervals have been

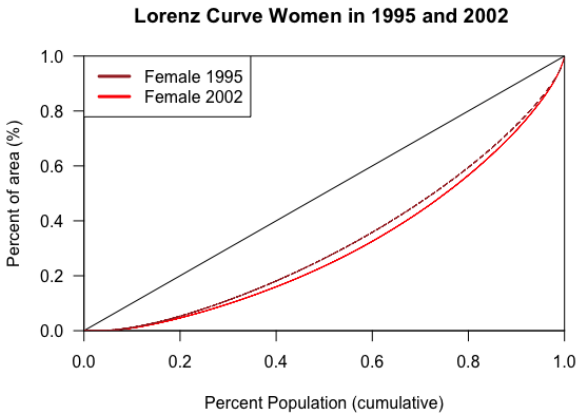


Figure 4.1: Women Lorenz Curve in 1995 and 2002

constructed thanks to the critical value of the normal distribution at a 5% significant level. Then, a suitable test has been performed using the easy-to-compute formula of Davidson (2009) for the standard deviation. The critical value has been computed thanks to the Equation (3.14) and equals 4.28330 in absolute value. Moreover, the degree of freedom has been calculated such as $n_1 + n_2 - 2$ where n_1 and n_2 are the length of the 1995 and 2002 sample respectively. The corresponding p-value equals 0.0000184142 which is smaller than the significant level of 0.05. The null hypothesis is, therefore, rejected at a 0.05 significance level. The Gini Indexes are significantly different from each other. Moreover, both Gini Indexes are significantly different from zero at a 0.05 significance level.

In conclusion, this part has statistically proven that income inequality has also increased for urban women in China in 2002 compared to 2002.

4.2 Concentration Curve and Concentration Index

The second part of the analysis consisted to breaking down the income by its different sources. This has been done with the help of the Concentration Curve and Concentration Index. The definition of income encompasses different revenues such as wage, income from a private business or self-employed, transfer income and property income. However, in 1995, other revenues are encompassed into the definition of income. Indeed, those are the labor income and the source called “other income”. As a matter of comparison, the two figures of the Concentration Curves have been based on the common income sources. However, the Concentration Curve with all the income source for 1995 is situated in Appendix 1. So, in this section, the total income has been evaluated using its different components such as transfer income, wage income, business

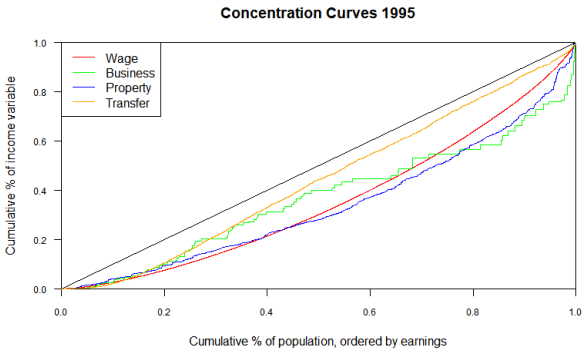


Figure 4.2: Concentration Curve 1995

income and property income. This decomposition will allow to analyze the different sources of distribution against the total income distribution.

Figures 4.2 and 4.3 have been evaluated, thanks to the function "curveContent" from the *IC2* package in R. Regarding the Concentration Curves of 1995, all are located below the equality line. That means that the concentration of income is higher when the source of income is higher. In addition, the transfer income dominates all the other components. Besides, in general, the property income is dominated by the wage and the transfer income. Regarding the business component, the interpretation is not clear in comparison with the property and wage income. However, the business income is dominated by the transfer income.

In conclusion, all the curves lie below the diagonal, which means that for all the components of income, the income takes a lower value for the poorest. Moreover, it is possible to rank them by degree of inequality. Note that when Concentration Curves cross, there is no dominance. What is clear from the figure is that the most equal distribution is the transfer source. On the other hand, if the cross between the transfer and wage is neglected, the property income is the most unequal distribution overall. Nevertheless, the business income dominates neither the property or wage income.

As explained, the figure in Appendix 1 represents the Concentration Curves with all the components of income in 1995. In this case, the labor income dominates all the other sources of income. The labor source of income is the most equal component of the income. On the other hand, the component "other income" is dominated by all the other components. That makes this component the most unequal source of income.

As regards the plot of 2002, the transfer income and the business income dominate the wage and property source of income. However, there is no dominance between the

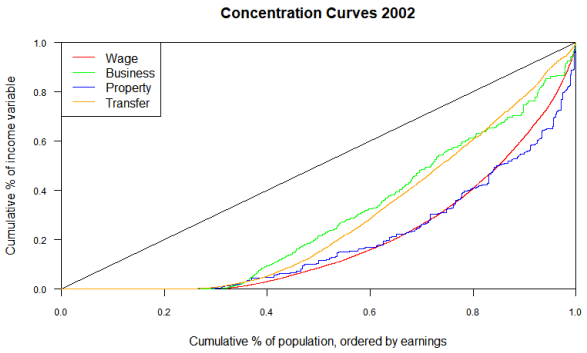


Figure 4.3: Concentration Curve 2002

business and the transfer income. The same conclusion holds for the wage and property income. Indeed, in this case from a visual point of view, it is not possible to see which source of income dominates the other.

Besides, when comparing the Concentration Curves of the two years, from a visual point of view, the two curves business and transfer income get closer as well as the wage and transfer income. Moreover, the two pairs get further apart.

Table 4.2 contains the Concentration Index in 1995 and 2002 for the wage income, business income, transfer income and property income. Remember that a positive Concentration Index means a higher income when the source of income is higher. This is a pattern that is observed for all the sources of income, meaning that the income is more concentrated when the wage, the business, the transfer and the property income are higher. Moreover, the contribution of the wage, the business, the transfer and the property to the income inequality has increased by 51%, 34%, 70% and 43% between 1995 and 2002. As the sources are the components of the total income, it appears that the two most unequally distributed sources of income are the property income and wage income. These are the two sources that contribute the most in the unequal distribution of the total income inequality.

	Wage	Business	Transfer	Property
1995	0.2945	0.2740	0.1295	0.3536
2002	0.6015	0.4181	0.4395	0.6092

Table 4.2 : Concentration Index

4.3 Income function

As demonstrated in the Lorenz Curve and Gini Index part, the women inequality has increased in 2002 compared to 1995. What this thesis aims at, is to explain the factors

that drive this inequality trend, thanks to a regression-based decomposition and the Shapley value decomposition method. However, before doing that, the specification and estimation of the income-generating function is needed. The income function that will be used for both methods is, as explained in the method part, composed by 7 explanatory categorical variables. Each of these variables have been coded as a factor such as the R program converts all the categorical variables in dummies. In this OLS regression, this is the logarithm of the income and not the income itself that represents the dependant variables. Indeed, this is one of the assumptions in the Fields decomposition method and it will be maintained for the Shapley value decomposition.

Appendix 4.1 and 4.2 represent the estimates of the income function for 1995 and 2002 respectively. Notice that, since the regression has one continuous variable as dependant variable and factors variables as explanatory variables, the OLS regression has to be analyzed in a specific way. Indeed, each level into each variable has to be analyzed by taking the first level as reference. Indeed, for each categorical variable, R does not evaluate each dummy for significance since, in that case, their coefficient are not uniquely defined. This is, therefore, the contribution of all dummies corresponding to the same categorical variable that is evaluated.

First of all, when looking at the statistics of the overall significance of the model, the null hypothesis for the two years must be rejected. The null hypothesis is that all the coefficient regressions are equals to zero. Moreover, the adjusted R-squared is equal to 0.36 and 0.33 for the year 1995 and 2002, respectively. This means that 36% and 33% of the variability of income is explained by the model.

The income function estimation confirms what can be observed in the descriptive statistics tables. Indeed, the factor *Province* seems to affect significantly the income. In 1995, the province Guangdong has a 37% higher income than the reference province being Beijing. On the other hand, all the other provinces have a lower income compared to Beijing. In 2002, most of the provinces have significantly lower income than Beijing. Moreover, the central and western provinces such as Henan and Hubei have a lower income than the eastern regions such as Jiangsu and Guangdong.

Income is also significantly affected by the *Communist Party in China*. Indeed, as discussed in the descriptive part, non-members of CPC earn less than members of the CPC. The gap between the two seems to widen over time. Indeed, in 1995 the non-members earn 8.5% less than the CPC members compared to 9.6% in 2002. As regards the factor *Education*, the same pattern as observed in the descriptive part can be deducted. In comparison to the level "[0,10]" years of work experience, the more work experience the individual has, the more the income increases significantly. However, in 2002, individuals having between 31 to 40 years of experience seemed to earn less than those in 1995. Thus, the more educated people are the more they significantly earn. In 1995, people with a primary school education earned 44% less

than those having a university diploma. In 2002, the academics earned 51% more than those without any education.

In the *Ownership* variable, SOEs governed by local government earns significantly less than the reference category which is the "central SOEs". The magnitude increases over time, passing from 16% to 18%. In 1995, foreign-owned enterprises earn up to 45% more than the central SOEs. However, in 2002, the sino-foreign and foreign enterprises didn't have significantly higher income than the central SOEs. Regarding to the occupation workers have in a company, it is clear that unskilled workers earn significantly less than the owner of a private or individual enterprise in 1995. However, unskilled workers do not have, in a significant way, less income than private owner enterprise in 2002. Finally, the industry has, as the reference category, the "agriculture, forestry, husbandry and fishing" sector. In 1995, the "insurance and finance" sector has significantly 27% more income than the the reference category. Nevertheless, in 2002, no sector seems to have significantly higher or lower income compared to the reference category.

Until now, the analysis focused on the description of each variable in comparison with the referenced category and the variation between the groups. However, the power of the model has not been assessed. Indeed, how do the pre-selected factors perform to explain inequality? Until now, this is missing information. In most of the papers in the literature, very often, authors do not look at the power of predictability of the model. Therefore, this thesis analyzes it thanks to a K-folds cross-validation. For both years, the specific k value has been chosen to be equal to 10. This number has been found through experimentation to be a good number to have a model estimate with modest variance and low bias.

Table 4.3 summarizes the R-squared and the Root Mean Squared Error obtained for the years 1995 and 2002. As seen below, the power of prediction of the model turns around 34% when the RMSE is about 54%, which is quite high. As seen in Table 4.3, the R-squared is higher in 1995 than in 2002, meaning that the factors explained more the variations of the log income in 1995 than in 2002.

	1995	**2002**
R-Squared	0.3600501	0.326319
RMSE	0.5306811	0.5617648

Table 4.3 : Cross Validation

4.4 Factor contributions to inequality

4.4.1 Fields Decomposition

After having estimated the income function, this part focuses on the decomposition of the income inequality using the regression-based decomposition developed by Fields

(2003) and the Shapley decomposition method. Results for the Fields method and Shapley method have been included in table 4.4 and 4.5 respectively. The results of the year 1995 are in the two first columns and the results of the year 2002 are in the two last columns. For each method, the factor inequality weight has been calculated as well as the percentage contribution in the explained inequality.

Remember that the s_j weights for the Fields method have been calculated as the ratio of the covariance between the dependant variable and the j-th factor multiplied by the j-th coefficient of the variable estimated by an OLS multiple regression, and the variance of the dependant variable. As demonstrated by Fields (2003), the sum of the s_j weights lead to the R-Squared of the OLS multivariate regression. The R-squared for the Fields method equal to 36.82 in 1995 and 34.94 in 2002. The remaining part is attributable to the residuals. A large proportion of the inequality is not explained by the variables in the model for both years. Moreover, the constant of the regression is only a multiplier that does not make a contribution to income inequality.

	s_j x100	ln %	s_j x100	ln %
Province	13.98	37.97	13.45	38.49
CPC	0.93	2.53	0.98	2.80
Education	5.11	13.88	4.25	12.16
Work Experience	6.87	18.66	3.43	9.82
Ownership	3.82	10.37	5.33	15.25
Occupation	4.48	12.17	4.84	13.85
Industry	1.64	4.45	2.65	7.58
Explained inequality	36.82	100	34.94	100
Unexplained inequality	63.18		65.06	
Total	100.0		100.0	

Table 4.4: Regression based decomposition using Fields 1995 (left) 2002 (right)

In 1995, the variables that contribute the most into the explained income inequality are, in the ascending order, the *Province* with 37.97%, the *Work Experience* with 18.66%, the *Education* with 13.88%, the *Occupation* with 12.17 % and the *Ownership* with 10.37%. Smaller weights are attributed to the *Communist Party in China* with 2.53% and *Industry* with 4.45%.

In 2002, the variable *Province* is still the one that contributes the most to the explained income inequality. Moreover, in 2002, this is the *Ownership* variable with 15.25% and the *Occupation* with 13.85% followed by 12.16% of the *Education* contribution that are the three most important contributors to the explained income inequality after the *Province*. Smaller contributions are detected for the *Communist Party in China*, the *Industry* and the *Work Experience*.

4.4.2 Shapley decomposition

The Shapley decomposition has also been applied to the function income in 1995 and 2002. Note that, the Shapley values would be equal to the Fields decomposition if there were no correlations between the variables. The decomposition results are based on the variance of the logarithm. As seen in Table 4.5, the contribution of the factors to the income inequality in absolute term (s_jx100) and the percentage of the observed inequality (ln %) have been calculated for each variable and for each year.

	s_j x100	ln %	s_j x100	ln %
Province	14.06	40.19	12.87	37.00
CPC	1.35	3.86	1.40	4.02
Education	4.45	12.72	3.85	11.07
Work Experience	4.73	13.52	3.05	8.77
Ownership	3.54	10.12	4.95	14.23
Occupation	4.89	13.98	5.52	15.87
Industry	1.92	5.49	3.12	8.97
Explained Inequality	34.98	100	34.78	100
Unexplained Inequality	65.02		65.22	
Total	100		100	

Table 4.5: Regression based decomposition using Shapley decomposition 1995 (left) 2002 (right)

In comparison to the coefficient obtained in Table 4.4, Table 4.5 differs slightly. That means that there were some correlations between the variables and/or multicollinearity. As explain in the description of the Shapley decomposition, in the case two variables are highly correlated, they tend to have the same weight. This is something that shows up for the *Education* and the *Work Experience* in both years. Indeed, as previously detected in the description analysis, the variables seem to be correlated. With the Shapley decomposition, the *Province* remains the most important contributor to the explained inequality. However, its total contribution decreases from one year to the other. Moreover, the three most important contributors in 1995 with the Fields Method are not the same with the Shapley value. Indeed, after the *Province*, it is the *Occupation* factor with 13.98%, closely followed by the *Work Experience* with 13.52% and the *Education* with 12.72%. The fifth contributor remains the *Ownership* with 10.12% of contribution in the explained income inequality. The low weight contributors are the *Industry* and the *CPC* in 1995.

Compared to the other method, the Shapley value results differ also slightly. But still, it is the *Province* factor that remains the main contributor. However, it is followed by the *Occupation* with 15.87 percentage of contribution and the *Ownership* with 14.23%. The fourth place is still occupied by the *Education*. Nevertheless, the *Industry* seems to have more contribution with the Shapley value method than with the Fields method. The remaining factors contribution sum up to 12.79 % and is composed by

the *Work Experience* with 8.77% and *CPC* factors.

As the Shapley decomposition gives more reliable results by taking into account the multicollinearity effect, the bootstrap method will be applied using the Shapley statistics. It will allow to confirm statistically that the weights for each factor are significantly different from each other.

4.4.3 Bootstrapping Estimators

The bootstrap method described at the beginning of this thesis will be used in order to have the estimators for each statistic. In this case, the statistics are the weight for each factor calculated with the Shapley decomposition. As the method required a high CPU and is computer-intensive, bootstrapping has been realized using 100 re-samplings. The statistics, bias and standards errors have been obtained with the "boot" function from the *boot* package in R incorporated with the Shapley statistics. The confidence intervals, in Table 4.6 and 4.7, have been calculated manually with a significant level of 0.05 and the critical value from the normal distribution. Indeed, following the central limit theorem, it is reasonable to assume that the Shapley value can be approximated by a normal distribution (Castro and al., 2009).

	Statistics	Bias	Standard Errors	Confidence Intervals
Province	0.1406	-0.0006	0.0073	[0.1257 : 0.1544]
CPC	0.0135	0.0004	0.0019	[0.0102 : 0.0176]
Education	0.0445	0.0004	0.0042	[0.0369 : 0.0535]
Work Experience	0.0473	0.0003	0.0038	[0.0402 : 0.0552]
Ownership	0.0354	0.0016	0.0041	[0.0289 : 0.0449]
Occupation	0.0489	0.0011	0.0039	[0.0424 : 0.0575]
Industry	0.0192	0.0017	0.0023	[0.0164 : 0.0255]

Table 4.6: Shapley values estimators, standard errors and confidence intervals in 1995

	Statistics	Bias	Standard Errors	Confidence Intervals
Province	0.1287	0.0012	0.0045	[0.1211 : 0.1387]
CPC	0.0141	0.0010	0.0031	[0.0291 : 0.0499]
Education	0.0385	0.0010	0.0053	[0.0090 : 0.0211]
Work Experience	0.0304	-0.0002	0.0045	[0.0215 : 0.0390]
Ownership	0.0495	0.0025	0.0057	[0.0408 : 0.0632]
Occupation	0.0552	0.0017	0.0067	[0.0438 : 0.0700]
Industry	0.0312	0.0045	0.0046	[0.0267 : 0.0447]

Table 4.7: Shapley values estimators, standard errors and confidence intervals in 2002

As explained in the data description, the two samples have been drawn independently. A suitable two-sample test has, therefore, been used. Moreover, the variances for each year are known. The critical values, in Table 4.8, have been thus calculated such that

$\frac{\hat{S}_1-\hat{S}_2}{\sqrt{\hat{\sigma}^2_{S_1}+\hat{\sigma}^2_{S_2}}}$ where $\hat{S}_1$ and $\hat{S}_2$ are the estimators of the Shapley value for one common variable in 1995 and 2002, respectively. Besides, $\hat{\sigma}^2_{S_1}$ and $\hat{\sigma}^2_{S_2}$ are the corresponding variance. Moreover, the degree of freedom has been calculated such as n_1+n_2-2 where n_1 and n_2 are the length of the 1995 and 2002 sample respectively. Table 4.8 resumes for each factor the calculated critical value as well as the theoretical critical value from a normal distribution. These tests have been based on a 5% significant level. As seen in the table, less than half of the factors have to reject the null hypothesis. The null hypothesis, for each factor, is that the Shapley values are equals for the two years. More than half of the factors are, therefore, not significantly different from one year to the other. Indeed, only the *Work Experience*, the *Ownership* and the *Industry* factors have shown a significantly change in 2002 compared to 1995. The next section will analyze how these findings will change the conclusions.

Factors	Stat	Critical Value	Reject/Not Reject
Province	1.1762	1.96	Not Reject
CPC	-0.3301	1.96	Not Reject
Education	0.8321	1.96	Not Reject
Work Experience	2.9634	1.96	Reject
Ownership	-2.1485	1.96	Reject
Occupation	-0.8958	1.96	Not Reject
Industry	-2.8539	1.96	Reject

Table 4.8: Two sample t-test

4.4.4 Comparison

As seen in Table 4.7, the two main contributors for both years are the *Province* followed by the *Occupation*, which is in line with what can be found in the Chinese inequality literature for the whole population. Indeed, the geographical position of the citizen remains the factor that contributes the most to the income inequality. However, even if the *Province* relative contribution has decreased from 1995 to 2002, the change is not significantly different. At the same time, the *Occupation* has remained the second most important contributor. Its relative contribution has also not significantly increased.

Moreover, for both years, the factor *Communist party in China* remains the last contributor. It is interesting to draw a parallel between these findings and the descriptive tables and income function estimated. Indeed, even if the gap between the CPC member and non-CPC members is widening, the contribution of the factor in the income inequality is stable over time.

Among the other factors, high relative changes in the weight contributions can be detected. Indeed, factor such as *Work Experience* that was among the highest contributors in 1995 has seen its net contribution percentage in the explained inequality decreased significantly by 45.8% in 2002. The *Work Experience* has thus become one of

the smallest contributor in the explained income inequality. This observation can be interpreted in two ways. Either there is a socio-economic reason behind the fact that *Work Experience* is really something that contributes less to income inequality, or its decrease is explained by the increasing contribution of other factors. On the other hand, the *Industry* and *Ownership* have seen their net contribution significantly increased by 38.78% and 28.88% respectively in 2002. Regarding the *Education*, its contribution has remained significantly the same and make it one of the fourth main contributor in the explained income inequality for both years.

It is interesting to compare these conclusions with those found for the inequality in the whole population of China. Indeed, remember the paper from Deng and Li (2009), that has analyzed the earning inequality in the urban population of China. The authors based their analysis on the year 1988, 1995 and 2002. They have also applied the Fields methods as well as the Shapley decomposition method. However, they did not perform suitable tests for significant changes between the years. They found that the geographical location was also the most important contributor to the inequality in China and it remains one for the urban women in China. Moreover, they have also found that the *Industry* and the *Ownership* factor have increased their contributions, from 1988 until 2002. This something that has been also detected here. This thesis also proved that the observable increases were significant at a 5% significant level. Besides, they have also observed that the *Work Experience* was minimal and decreasing overtimes. Regarding the *Education*, its contribution has remained significantly the same in 1995 and 2002. Nevertheless, in Deng and Li's paper (2009), they have observed that the *Education* has had an increasing role in earning inequality. Besides, they have also found out that the *Occupation* has had an increasing role in the earning inequality although this thesis proved that this increase, in 2002, is not significantly different from the weight in 1995.

Chapter 5

Conclusion

This thesis aimed at analyzing the factors behind income inequality in 1995 compared to 2002 for women in urban China. However, to answer the question, the thesis decomposed it into two research questions. First, have the income inequality increased/decreased significantly for the urban women in China in 2002 compared to 1995? Secondly, how factors that contribute to the income inequality of the whole population have contributed to the increasing/decreasing inequality for urban women?

Indeed, the first part of the analysis has tackled the first sub-question, with the help of the Lorenz Curve and the Gini Index. As expressed in the literature, the income inequality has been increasing for the whole population. Therefore, it seemed logical to observe the same trend for the subgroup this thesis is analyzing. In order to do that, the Davidson unbiased Gini's estimator and the plug-in formula of the Gini's standard deviation have been used. It has been shown that the Gini Index has moved from 0.35 to 0.39 in 7 years. Besides, it has been proven that the two Gini Indexes were significantly different. The income inequality are, therefore, significantly increased for the urban women in China.

Then, to have an idea of how the components of the income contribute to income inequality, the different sources of income have been analyzed with the help of the Concentration Curve. Besides the Concentration Curve, the Concentration Index helped to draw conclusions about the different sources of income. It has been observed that the distributions of wage, business, transfer and property, in terms of the total income, have been more unequally distributed in 2002 than in 1995. In particular, the wage and property income were the two largest sources of income with the most unequal distribution. This is an interesting result since, in the descriptive part, a significantly high correlation between the wage and the total income has been found. The wage income has, therefore, a high impact on the distribution of the income.

Following these observations, the income function has been identified with respect to factors that were proven to have an impact on the inequalities in China. Those factors

have been selected from the literature. Thanks to the cross-validation, it has been shown that those factors explained 36% and 33% of the income inequality, respectively for each year. The explanatory power of the pre-selected factors have thus decreased in 2002 compared to 1995. The conclusions about the factors have thus to be taken in perspective of their explanatory power.

Then, to answer the second question: "how factors that contribute to the income inequality of the whole population have contributed to this increasing/decreasing inequality?", two models have been used. The first model was the one introduced by Fields and is called Regression-based decomposition. The second is called Shapley value decomposition. The comparison of the results of the two methods have shown some multicollinearity in the income function. Indeed, the results differed and some coefficients were almost equals. Moreover, due to this change in weights, the two methods resulted in different conclusions. Indeed, the Shapley decomposition approach allows to measure the exact inequality contribution of factors taking into account their interaction with other factors. For this reason, the decision has been made to continue the analysis based on the results of the Shapley decomposition.

Based on these observations, the estimators of the Shapley values, the standard errors, and the confidence intervals have been computed, thanks to a bootstrap re-sampling technique using 100 re-samplings. Moreover, a two-sample test has been realized to compare the statistical significance of the weight variation through the years. It comes out of these tests that not all factors were significantly different in 2002 compared to 1995.

Weight and net percentage of contribution in the explained inequality for each factor and for each year have been represented in Table 4.5. The conclusions of the contribution of the factors have been thus based on this latter table in regards to the conclusion of Table 4.8. From these tables, it has been shown that the two highest contributors on the explained inequality were the geographical place of the citizen, in other words, the *Province* and the *Occupation* of the workers in the company. This conclusion holds for the two years analyzed. Moreover, the net percentage contribution for each factor has remained statistically constant in 2002 compared to 1995. This is not surprising, since, as explained in the literature review, reforms in 1978 have allow to center the economic activity in the coastal regions, creating disparities between provinces. This is also a pattern that was observed in the descriptive analysis where the coastal regions had a higher mean income in 1995. Moreover, the same conclusions applied in 2002, even if this distinction was less obvious. Besides, the worker position in the company is the second most important factor that explained the income inequality for the urban women in China. This can be explained by the wage gap between different occupations in a company. As seen in the descriptive part and in the income function analysis, jobs that required limited skill pay less than those in the director department.

The *Communist Party in China* membership has been the factor that contributes the less into the explained income inequality in 1995 and also in 2002. The change in the weight factor is not significantly different from one year to the other. However, as seen in the income function analysis, it has been shown that, for both years, a non-CPC member earn significantly less than a CPC member.

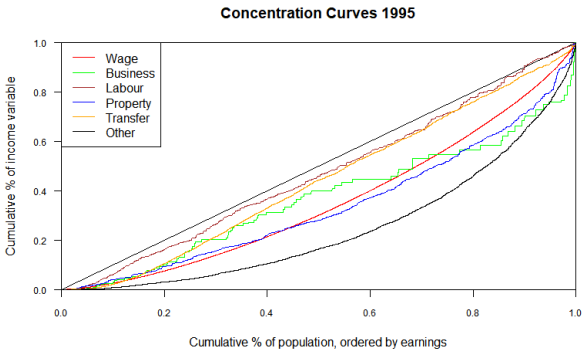
Another interesting finding that this thesis has tackled, regards the contribution of the *Education* factor. Indeed, when for other inequality analysis, considering the whole population of China, *Education* contributed more and more in the income gap. On the other hand, for urban women in China, this factor does not show a significant difference in 2002 compared to 1995. This could be explained by the remaining imbalance in the educational opportunities for women in China (Carpenter and Liu, 2005).

Regarding the other factors, they have changed their relative importance from one year to the other. Indeed, the contribution of the *Work Experience* decreased significantly by 45.8%. This decrease makes the *Work Experience* factor one of the smallest contributors in 2002. It confirms the hypothesis of Deng and Li (2009) that the *Work Experience* contribution is minimal for the whole population but also for urban women. On the other hand, factors such as the *Industry* and *Ownership* have increased significantly their contribution to the explained inequality by 38.78% and 28.88% respectively. It confirms the findings of Wang (2007) and extends it to the urban women population.

Concretely, the income inequality in 1995 was much lower and was mainly driven by the *Province*, *Work Experience* and *Occupation* factors when, in 2002, income inequality has risen and was mainly driven by the *Province*, the *Ownership* and the *Occupation*. Thus, in 1995, the inequality has been driven mostly by the place of living, the years of experience of the workers and the position of the workers in the company. However, seven years later, *Work Experience* for women in China has lost significant importance in the inequality instead of the type of *Ownership* of the company that has seen a significant increase in the contribution of inequality. In other words, when in 1995, the *Work Experience* could influence the gap between the income, in 2002, this is the type owner of the company and the position of the worker in the company that explained the most the gap between the income. As mentioned in the literature, that can be explained by the fact that the foreign firms required more productive workers when on the other hand urban collective firms required less productive workers widening therefore the wage gap. Besides, the underlying sources of income that contributed the most in the unequal distribution of the total income were the property income and wage income. This is in line with the conclusions as, among the most important contributors, the *Occupation* and the *Ownership* factors are directly linked to the woman wage income.

Chapter 6

Appendix



Appendix 1 : Concentration Curve 1995

Appendix 2 . : China Province

	Proportion	Income(Yuan)
Province		
Beijing (11)	0.075	7554.218
Shanxi (14)	0.090	4149.433
Liaoning (21)	0.101	5079.263
Jiangsu (32)	0.114	6259.055
Anhui (34)	0.077	4590.447
Henan (41)	0.089	4191.204
Hubei (42)	0.101	5860.551
Guangdong (44)	0.086	10148.604
Sichuan (51)	0.119	5616.522
Yunnan (53)	0.097	5736.440
Gansu (62)	0.052	4199.366
CPC		
CPC members	0.16	7376.214
Non-CPC members	0.84	5498.076
Work Experience		
$\leqslant 10$	0.206	4451.691
[11:20]	0.328	5757.363
[21:30]	0.346	6367.776
[31:40]	0.109	6548.102
> 40	0.010	7021.154
Education		
College or above	0.048	7824.069
Professional college	0.107	7346.164
Middle Level professional, technical or vocational school	0.164	6638.682
Senior middle school	0.238	5548.517
Junior middle school	0.326	5304.296
Primary school	0.092	4446.739
Below Primary	0.024	3490.175
Ownership		
Central SOEs	0.224	6740.942
Local SOEs	0.530	5851.895
Urban Collective	0.210	4574.620
Private enterprise	0.003	4443.050
Individual enterprise	0.014	5213.533
Sino-foreign joint venture	0.010	8462.468
Foreign owned	0.001	14943.857
Township and Village enterprise	0.000	7450.000
Others	0.007	4920.891

	Proportion	Income (Yuan)
Occupation		
Owner of private or individual enterprise	0.014	5212.136
Owner and manager of private enterprise	0.001	6737.556
Professional or Technical Worker	0.215	7184.235
Head of institution	0.016	8459.554
Division head institution	0.040	8132.000
Office worker	0.215	6004.698
Skilled worker	0.185	5131.562
Unskilled worker	0.238	4577.079
Others	0.076	5030.150
Industry		
Agriculture, forestry, husbandry, fishing	0.014	6130.523
Manufacturing	0.420	5214.613
Mining and geological prospecting	0.009	5759.810
Construction	0.025	5537.219
Transport, communications, posts, telecommunications	0.039	6275.903
Commerce, restaurants, Material supply, warehousing	0.166	5298.317
Real estate, public utilities, personal and consulting services	0.045	5762.014
Health, physical culture and social welfare	0.058	6894.019
Education, culture, art and technical services	0.083	6823.081
Scientific Research and Technical services	0.020	6923.161
Finance and insurance	0.019	7289.264
Government and Party organs, social organizations	0.093	7092.393
Others	0.009	5063.714
Observations : 6343		
Mean income (Yuan) :5798.614		

Appendix 3.1 : Descriptive table 1995

	Proportion	Income(Yuan)
Province		
Beijing (11)	0.085	15057.043
Shanxi (14)	0.063	7836.814
Liaoning (21)	0.123	8326.889
Jiangsu (32)	0.110	9147.987
Anhui (34)	0.070	7689.218
Henan (41)	0.092	6575.092
Hubei (42)	0.092	8618.477
Guangdong (44)	0.104	15633.505
Sichuan (50)	0.038	9347.050
Chongqing (51)	0.090	8180.764
Yunnan (53)	0.079	9191.063
Gansu (62)	0.054	7714.855
CPC		
CPC members	0.16	12242.648
Other Parties	0.008	9016.248
Communist Youth League	0.051	9842.404
No party	0.781	9050.624
Work Experience		
$\leqslant 10$	0.221	8126.524
[11:20]	0.337	9280.362
[21:30]	0.360	10056.483
[31:40]	0.083	12870.649
Education		
Never Schooled	0.002	4581.200
Classes for eliminating illiteracy	0.002	3142.800
Elementary School	0.031	6449.123
Junior middle school	0.272	7811.623
Senior middle school (including professional middle school)	0.350	9288.577
Technical secondary school	0.134	10681.833
Junior College	0.170	11822.242
College/University	0.037	14592.296
Graduate	0.001	17221.333
Ownership		
Central/Provincial SOEs	0.132	11779.250
Local SOEs	0.280	9729.089
Urban Collective	0.133	7658.628
Private enterprise	0.066	8340.324
Individual enterprise	0.136	7375.244
Sino-foreign joint venture	0.023	15018.900
Foreign owned	0.008	14429.183
State share-holding company	0.046	12486.926
Other share-holding company	0.122	10356.425
Rural private enterprise	NA	NA
Rural individual enterprise	0.000	4800.000
Others	0.054	8339.507

	Proportion	Income (Yuan)
Occupation		
Owner of private enterprise	0.004	9145.475
Self employed	0.059	8173.648
Professional	0.152	12156.330
Director of government agent, institution and enterprise	0.006	13817.078
Department director of government agent, instit. enterprise	0.025	17278.450
Office Worker	0.182	11607.492
Skilled worker	0.158	9113.337
Unskilled worker	0.149	7897.828
Salesclerk or Service worker	0.233	7634.486
Farmer	NA	NA
Others	0.033	6736.328
Industry		
Agriculture, forestry, husbandry, fishing	0.007	9591.310
Mining	0.013	7809.873
Manufacturing	0.351	9410.856
Electricity, gas and water supply facilities	0.035	13100.534
Construction	0.032	11328.513
Geological prospecting, irrigation administration	0.004	15263.364
Transportation, storage, post office and communication	0.065	11892.169
Wholesale, retail and food services	0.218	8475.238
Finance and insurance	0.035	13316.163
Real estate	0.010	10758.049
Social Services	0.145	7959.559
Health, sports and social welfare	0.021	10016.437
Education, culture and arts, mass media and entertainment	0.018	10484.243
Scientific Research and professional services	0.007	16462.283
Government agents, party organisations and social groups	0.011	11466.844
Others	0.029	9258.275
Observations : 2890		
Mean income (Yuan) : 9601.18		

Appendix 3.2 : Descriptive table 2002

	Estimate
Province	
Intercept	8.867062***
Beijing (11)	-
Shanxi (14)	-0.544961***
Liaoning (21)	-0.355947***
Jiangsu (32)	-0.035704
Anhui (34)	-0.368764***
Henan (41)	-0.492495***
Hubei (42)	-0.194778***
Guangdong (44)	0.369408***
Sichuan (51)	-0.280014***
Yunnan (53)	-0.222550***
Gansu (62)	-0.500305***
CPC	
CPC members	-
Non-CPC members	-0.085377***
Work Experience	
$\leqslant 10$	-
[11:20]	0.359726***
[21:30]	0.433813***
[31:40]	0.433429***
> 40	0.416610***
Education	
College or above	-
Professional college	-0.041892
Middle Level professional, technical or vocational school	-0.118861***
Senior middle school	-0.179698***
Junior middle school	-0.226116***
Primary school	-0.442223***
Below Primary	-0.617211***
Ownership	
Central SOEs	-
Local SOEs	-0.162727***
Urban Collective	-0.304236***
Private enterprise	-0.322255**
Individual enterprise	-0.166955*
Sino-foreign joint venture	0.111518
Foreign owned	0.456544*
Township and Village enterprise	0.021659
Others	-0.254523**

	Estimate
Occupation	
Owner of private or individual enterprise	-
Owner and manager of private enterprise	0.036737
Professional or Technical Worker	0.044152
Head of institution	0.042991
Division head institution	0.035353
Office worker	-0.076919
Skilled worker	-0.102280
Unskilled worker	-0.204608**
Others	-0.168094*
Industry	
Agriculture, forestry, husbandry, fishing	-
Manufacturing	-0.013702
Mining and geological prospecting	0.009321
Construction	0.016968
Transport, communications, posts, telecommunications	0.069821
Commerce, restaurants, Material supply, warehousing	0.005808
Real estate, public utilities, personal and consulting services	-0.013222
Health, physical culture and social welfare	0.087718
Education, culture, art and technical services	0.079605
Scientific Research and Technical services	0.032501
Finance and insurance	0.267505***
Government and Party organs, social organizations	0.099500
Others	-0.130810
Adj. R-Squared : 0.3633	
F-Stat : 74.85	
P-value : <2.2e-16	

Appendix 4.1 : Income function Estimate in 1995

Note : ***,**,* represents statistical significance at the 1%, 5% and 10% levels, respectively

	Estimate
Intercept	9.484163
Province	
Beijing (11)	-
Shanxi (14)	-0.687271***
Liaoning (21)	-0.541084***
Jiangsu (32)	-0.457905***
Anhui (34)	-0.683595***
Henan (41)	-0.764870***
Hubei (42)	-0.614053***
Guangdong (44)	-0.037133
Sichuan (50)	-0.504844***
Chongqing (51)	-0.565638***
Yunnan (53)	-0.513593***
Gansu (62)	-0.657836***
CPC	
CPC members	-
Other Parties	0.013206
Communist Youth League	-0.107264
Non-CPC members	-0.096791**
Work Experience	
$\leqslant$10	-
[11:20]	0.189994***
[21:30]	0.254187***
[31:40]	0.377845***
Education	
Never Schooled	-
Classes for eliminating illiteracy	-0.440269
Elementary School	0.022420
Junior middle school	0.163383
Senior middle school (including professional middle school)	0.253105
Technical secondary school	0.310363
Junior College	0.401057
College/University	0.506762
Graduate	0.427321
Ownership	
Central/Provincial SOEs	-
Local SOEs	-0.179773***
Urban Collective	-0.299702***
Private enterprise	-0.249451***
Individual enterprise	-0.303738***
Sino-foreign joint venture	0.125712
Foreign owned	-0.025372
State share-holding company	0.080900
Other share-holding company	-0.089728*
Rural private enterprise	NA
Rural individual enterprise	-0.304684
Others	-0.316494***

	Estimate
Occupation	
Owner of private enterprise	-
Self employed	-0.006895
Professional	-0.014016
Director of government agent, institution and enterprise	-0.012463
Department director of government agent, instit. enterprise	0.157480
Office Worker	-0.051039
Skilled worker	-0.085418
Unskilled worker	-0.255940
Salesclerk or Service worker	-0.227846
Farmer	NA
Others	-0.406869*
Industry	
Agriculture, forestry, husbandry, fishing	-
Mining	-0.243819
Manufacturing	-0.150341
Electricity, gas and water supply facilities	0.197801
Construction	-0.133843
Geological prospecting, irrigation administration	0.230492
Transportation, storage, post office and communication	-0.001208
Wholesale, retail and food services	-0.173445
Finance and insurance	-0.006113
Real estate	-0.133301
Social Services	-0.172725
Health, sports and social welfare	-0.132130
Education, culture and arts, mass media and entertainment	-0.072927
Scientific Research and professional services	0.057175
Government agents, party organisations and social groups	-0.175834
Others	-0.152950
Adj. R-Squared : 0.33	
F-stat : 25.76	
P-value : <2.2e-16	

Appendix 4.2 : Income function Estimate in 2002

Note : ***,**,* represents statistical significance at the 1%, 5% and 10% levels, respectively

Bibliography

Abdel-Salam, G. and Verma, J. (2019), *Testing Statistical Assumptions in Research*, Wiley, chapter 4. Assumptions in Parametric Tests.

Amini Dehak, M., Mirzaei, S. and Mohtashami Borzadaran, G. R. (2017), 'A comparative study of the gini coefficient estimators based on the linearization and u-statitstics methods', *Revista Colombiana de Estadistica* **40**(2), 205–221.

Arcagni, A. and Porro, F. (2014), 'The graphical representation of inequality', *Revista Colombiana de Estadistica* **37**(2), 419–436.

Barrett, G. and Donald, S. G. (2009), 'Statistical inference with generalized gini indices of inequality, poverty and welfare', *Journal of Business Economic Statistics* **27**(1), 1–17.

Bellù, L. G. and Liberati, P. (2006), 'Inequality analysis : The gini index', *FAO EASYpol Module 40* .
URL: *http://www.fao.org/3/a-am352e.pdf*

Bellù, L. and Liberati, P. (2005), 'Charting income inequality : The lorenz curve', *FAO, EASYpol Module 00* .
URL: *http://www.fao.org/3/a-am391e.pdf*

Berger, Y. G. (2008), 'A note on the asymptotic equivalence of jackknife and linearization variance estimation for the gini coefficient', *Journal of official statistics* **24**(1), 541–555.

Bilger, M., Lokshin, M., Sajaia, Z. and Wagstaff, A. (2011), *Health Equity and Financial Protection*, The World Bank, Washington, D.C.

Bishop, J. A., Chow, K. V. and Formby, J. P. (1994), 'Testing for marginal changes in income distributions with lorenz and concentration curves', *International Economic Review* **35**(2), 479–488.

Boos, D. (2003), 'Introduction to the bootstrap world', *Statistical Science* **18**(2), 168–174.

Booth, G. J., Hall, P. and Wood, A. T. A. (1993), 'Balanced importance resampling for the bootstrap', *The annals of Statistics* **21**(1), 286–298.

Bourguignon, F., Fournier, M. and Gurgand, M. (2001), 'Fast development with a stable income distribution : Taïwan, 1979-94', *Review of Income and Wealth* **47**(2), 139–163.

Bunke, O. and Droge, B. (1984), 'Bootstrap and cross-validation estimates of the predictor error for linear regression models', *The Annals of Statistics* **12**(4), 1400–1424.

Carpenter, M. and Liu, J. (2005), 'Trends and issues of women's education in china', *The Clearing House: A Journal of Educational Strategies, Issues and Ideas* **78**(6), 277–281.

Castro, J., Gomez, D. and Tejada, J. (2009), 'Polynomial calculation of the shapley value based on sampling', *Computers and Operations Research* **36**(5), 1726–1730.

Chen, Y., Demurger, S. and Fournier, M. (2007), 'The evolution of gender earnings gaps and discrimination in urban china : 1988-95', *The Developing Economies* **45**(1), 97–121.

Cook, D. and Picard, R. (1984), 'Cross-validation of regression models', *Journal of the American Statistical Association* **79**(9), 575–583.

Cowell, F. (2011), *Measuring Inequality*, Oxford University Press, chapter First Principles.

Davidson, R. (2009), 'Reliable inference for the gini index', *Journal of Econometrics* **150**(1), 30–40.

Deng, Q. and Li, S. (2009), 'What lies behind rising earnings inequality in urban china? regression-based decompositions', *CESifo Economic Studies* **55**(3-4), 598–623.

Eckart, J. (2016), '8 things you need to know about china's economy', *World Economic Forum* .

Efron, B. (1979), 'Bootstrap methods, another look at the jacknife', *The annals of Statistics* **7**(1), 1–26.

Efron, B. (1983), 'Estimating the error rate of a prediction rule: Improvement on cross-validation', *Journal of the American Statistical Association* **78**(382), 316–331.

Fields, G. (1987), 'Measuring inequality change in an economy with income growth', *Journal of Development Economics* **26**(1), 357–374.

Fields, G. (2003), 'Accounting for income inequality and its change: A new method, with application to the distribution of earnings in the united states', *Research in labor economics* **22**, 66.

Frank E. Harrell, J. (2001), *Regression modeling strategies*, Second Edition, Springer, chapter Describing, Resampling, Validating and Simplifying the Model.

Gastwirth, J. and Modarres, R. (2006), 'A cautionary note on estimating the standard error of the gini index of inequality', *Oxford Bulletin of Economics and Statistics* **68**(3), 385–90.

Georis, V. (2019), 'La chine montre ses muscles pour cacher ses divisions internes', *L'Echo* .

Giles, D. E. A. (2004), 'Calculating a standard error for the gini coefficient : some further results', *Oxford Bulletin of Economics and Statistics* **66**(3), 425–33.

Gustafsson, B. and Li, S. (2000), 'Economic transformation and the gender earnings gap in urban china', *Journal of Population Economics* **13**(2), 305–329.

Hao, L. and Naiman, D. Q. (2010), *Assessing Inequality*, Thousand Oaks, Los Angeles : Sage, chapter PDFs, CDFs, Quantile Functions, and Lorenz Curves.

Hirst, T. (2015), 'A brief history of china's economic growth', *World Economic Forum* .

Huettner, F. and Sunder, M. (2011), 'Decomposing r^2 with the owen value', *University of Leipzig, Faculty of Economics and Management Science* .

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

Jain-Chandra, S., Khor, N., Mano, R., Schauer, J., Wingnder, P. and Zhuang, J. (2018), 'Imf working paper : Inequality in china : Trends, drivers and policy remedies'.

Jann, B. (2016), 'Estimating lorenz and concentration curves', *The stata Journal* **16**(4), 837–866.

Jedrzejczak, A. (2008), 'Decomposition of the gini index by sources of income', *International Advances in Economic Research* **14**(4), 441–447.

Kakwani, N., Wagstaff, A. and Van Doorslaer, E. (1997), 'Socioeconomic inequalities in health: Measurement, computation and statistical inference.', *Journal of Econometrics* **77**(1), 87–103.

Knight, J. and Song, L. (2003), 'Increasing urban wage inequality in china : Extent, elements and evaluation', *Economics of Transition* **11**(4), 597–619.

Knight, J. and Song, L. (2008), *China's Emerging Urban Wage Structure, 1995-2002*, Cambridge University Press.

Lambert, P. J. (2001), *The distribution and redistribution of income*, 3 edn, Manchester University Press.

Langel, M. and Tillé, Y. (2013), 'Variance estimation of the gini index : revisiting a result several times published', *Journal of Royal Statistics Society-Series A* **179**(2), 521–540.

Li, S. (2007), 'Changes in the distribution of wealth in china 1995-2002', *World Institute for Development Economic Research (UNU-WIDER), Working Papers* .

Li, S. (2008), 'Chinese household income project, 2002', *Inequality and Public Policy in China* pp. 337–354.

Li, S., Luo, C., Wei, Z. and Yue, X. (2008), Appendix: The 1995 and 2002 household surveys: Sampling methods and data description, *in* B. A. Gustafsson, L. Shi and T. Sicular, eds, 'Inequality and Public Policy in China', Cambridge University Press, p. 337–354.

Li, S., Renwei, Z. and Riskin, C. (2010), 'Chinese household income project, 1995', *Inter-university Consortium for Political and Social Research* .

Lorenz, M. O. (1905), 'Methods of measuring the concentration of wealth.', *Journal of the American statistical Association* **9**(70), 209–219.

Mills, J. A. and Zandvakili, S. (1997), 'Statistical inference via bootstrapping for measures on inequality', *Journal of Applied Econometrics* **12**(2), 133–150.

Morduch, J. Sicular, T. (2002), 'Rethinking inequality decomposition, with evidence from rural china', *Economic Journal* **112**(476), 93–106.

Morrison, W. (2019), 'China's economic rise : History, trends, challenges, and implications for the united states', *Washington, DC : Congressional Research Service* .

OECD (2004), *Income Disparities in China : An OECD Perspective*, OECD Publishing Paris.

Ogwang, T. (2000), 'A convenient method of computing the gini index and its standard error', *Oxford Bulletin of Economics and Statistics* **62**(1), 123–29.

Ostertagovà, E., O. O. (2013), 'Methodology and application of one-way anova', *American Journal of Mechanical Engineering* **1**(7), 256–261.

Serfling, R. J. (2002), 'Approximation theorems of mathematical statistics', *John Wiley Sons, New York* .

Shorrocks, A. (1999), 'Decomposition procedures for distributional analysis : a unified framework based on the shapley value', *Department of Economics, University of Essex* .

Shorrocks, A. (2013), 'Decomposition procedures for distributional analysis: a unified framework based on the shapley value', *Journal of Economic Inequality* **11**(1), 1–28.

Wang, F. (2007), *Rising inequality in post-socialist urban China*, Stanford University Press.

Xie, Y. (2010), 'Understanding inequality in china', *Society* **30**(3), 1–20.

Xie, Y. and Zue, X. (2014), 'Income inequality in today's china', *Proceedings of the National Academy of Sciences of the United States of America* **111**(19), 6928–6933.

Xu, K. (2007), 'U-statistics and their asymptotic results for some inequality and poverty measures', *Economics Reviews* **26**(5), 567–77.

Yitzhaki, S. and Schechtman, E. (2013), *The Gini Methodology : A primer on a statistical Methodology*, Springer, New York.

Yueh, L. (2004), 'Wage reforms in china during the 1990s', *Asian Economic Journal* **18**(2), 149–164.

FACULTY OF BUSINESS AND ECONOMICS
Naamsestraat 69 bus 3500
3000 LEUVEN, België
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/lsm**