Proceedings of the 3rd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2025.22561

Research on Factors Influencing Income Inequality by Multiple Linear Regression

Yunjia Zhu

Suzhou Foreign Language School, Suzhou, China
zhushengwu9412@ustc.edu

Abstract: This article aims to identify factors that influences income inequality and the method of multiple linear regression is used. The Gini coefficient is employed to measure the degree of income inequality, and multiple variables such as demographic trends, population health, and economic indicators are selected. Through regression analyses of data from China from 2001 to 2020 and data from 30 developed countries in 2020, it is found that in the Chinese model, multiple variables are closely related to the Gini coefficient, with a relatively high level of fitting. In the international model, only the net migration rate has a significant impact on the Gini coefficient, with a lower fitting than the Chinese model. This indicates that there are differences in the factors influencing income inequality between China and developed countries. What holds true for China may not be applicable to the international context, and vice versa. In the Chinese model, the combined effect of variables is more prominent, while in developed countries, the influences of other considered factors are likely to be more prominent.

Keywords: Income inequality, multiple linear regression, Gini coefficient.

1. Introduction

Income inequality has long been a topic of research in the fields of economics. In an economy, income can be generated in production by factors of production including land, labor, capital and entrepreneur. Meanwhile, income will be distributed to different firms and industries [1].

However, the distribution of income generated by these factors of production is far from even. In different firms and industries, the income-earning capabilities could differ significantly. For instance, high-tech industries with abundant capital and advanced technology tend to generate much higher incomes compared to labor-intensive industries. This disparity in income within and between firms and industries gradually accumulates, leading to the issue of income inequality. In addition, on the perspective of society as a whole, income inequality could lead to serious consequences such as poverty, higher crime rates, poorer health level and so on, which make it vital to tackle [2].

Income inequality can be shown by Gini coefficient, a concept introduced by Italian statistician Corrado Gini in 1912, it is a measurement of income inequality which uses numbers between 0 and 1 inclusive to reflect the degree of income inequality in an economy. For example, a Gini coefficient of 0 represents perfect equity, meaning that the income is equally distributed in an economy. Meanwhile, a value of 1 represents maximal inequality [3].

The following paragraphs listed three possible factors of income inequality: inflation, demographic changes, and life expectancy. Firstly, inflation represents a general rise in price level in an economy.

DOI: 10.54254/2753-8818/2025.22561

Inflation reduces the purchasing power of money. When prices surge due to inflation, workers' real wages are likely to decline if their nominal wages do not the same pace with the rising price. Low-income workers are often more vulnerable in this situation. They usually work in jobs with less bargaining power, such as those in the service or unskilled labor sectors. To compare, high-income earners, especially those that are highly demanded, may be more successful in getting more wage to offset inflation. This can widen the income gap between low and high-income groups [4]. Past researches have shown similar results. Bulír found that inflation generally increases income inequality, with a more significant impact in hyper-inflationary countries, and the effect of low inflation on improving income equality is more prominent in low-and high-income countries [5]. Monnin showed that there is a U - shaped relationship between long - run inflation and income inequality in developed economies, where inequality decreases as inflation rises until it reaches about 13% and then increases [4].

Secondly, rapid population growth, especially in developing countries, leads to a significant increase in the labor supply, and when the growth of the labor force exceeds the job vacancies, oversupply of labor occurs [6]. This drives down wages for low-skilled labor, widening the income gap. Immigration can also impact income inequality. Immigrants may take low-skilled and low-paid jobs in host countries, reducing job opportunities for domestic labors. In past researcher, Odusola et al. discovered that whether demographic changes affect income inequality remained ambiguous. At the bivariate level, variables like higher fertility rates and population growth rates are associated with lower inequality. However, at the multivariate level, when control variables are introduced, the population variables often have no effect on inequality [7].

Thirdly, life expectancy is the average number of years a person is expected to live. When life expectancy is longer, individuals are more likely to invest in long-term education and skills training. Researchers have investigated the impact of life cycles on incomes. Strong associations between infant mortality rate and income inequality are shown [8]. Chetty et al. found that higher income was associated with greater longevity throughout the income distribution and inequality in life expectancy increased over time. The reasons behind this are mainly due to health issues like smoking and obesity [9].

To further explore the complex issue of income inequality, this paper will use the method of multiple linear regression to identify the key factors that influence this economic phenomenon. Through this analysis, a deeper understanding of the mechanisms behind income inequality in different economic contexts can be gained, providing a basis for coming up with more targeted and effective policies to address this persistent economic and social problem.

2. Methods

2.1. Data Source

The datasets used in this paper are available on the Kaggle website. The first multiple linear regression involves the use of three datasets, focusing on China, and the second one involves six, expanding the subject to the international scale.

2.2. Variable Selection

The datasets contain a large amount of data, those that are selected mainly divide into three categories: demographic trends (net migration, change in population), population health (birth rate, life expectancy), and economic indicators (Gini coefficient, change in Gross domestic product (GDP), inflation rate). The list of variables and their definitions are shown below in Table 1.

DOI: 10.54254/2753-8818/2025.22561

Table 1: List of Variables

Logogram	Variable	Definition
Y	Gini coefficient	A measurement of income inequality which uses numbers between 0 and 1 inclusive to reflect the degree of income inequality in an economy
X1	Net migration rate	Reflects the difference between the number of immigrants and emigrants in an area within a certain period
X2	Birth rate	The number of live births per 1,000 people in a population during a given year
X3	Percentage growth in GDP	A measurement of the relative change in a country's Gross Domestic Product over a specific period
X4	Inflation rate	A measurement of the general rise in price over a specific period
X5	Percentage increase in population	A measurement of the relative growth of a population over a specific period
X6	Life expectancy	The average number of years a person is expected to live

2.3. Method Introduction

This paper uses multiple linear regression to explore the factors influencing income inequality, using the Gini coefficient as a measure. This method analyses the relationship between a single dependent variable and various independent variables [10]. The formula of multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \epsilon \quad (1)$$

Where β_0 is a constant term, X is the variable and ϵ is the error term.

This paper is going to conduct two multiple regression analysis. The first one focuses on China, covering the period from 2001 to 2020. The second one expands the scope to the entire world. 30 developed countries in 2020 are selected for the study. The purpose is to investigate whether different results will emerge compared to the analysis of China.

3. Results and Discussion

3.1. Linear Regression Model for China

The relationships among variables are also investigated. As shown in Figure 1 below. Figure 1 shows the correlation between variables. According to the variable correlation heatmap, birth rate is highly related to percentage increase in population and inversely related to life expectancy. However, it can be noticed that the correlation coefficient of all variables listed are approximately 0, indicating that each of them are not strongly correlated to Gini coefficient.

DOI: 10.54254/2753-8818/2025.22561

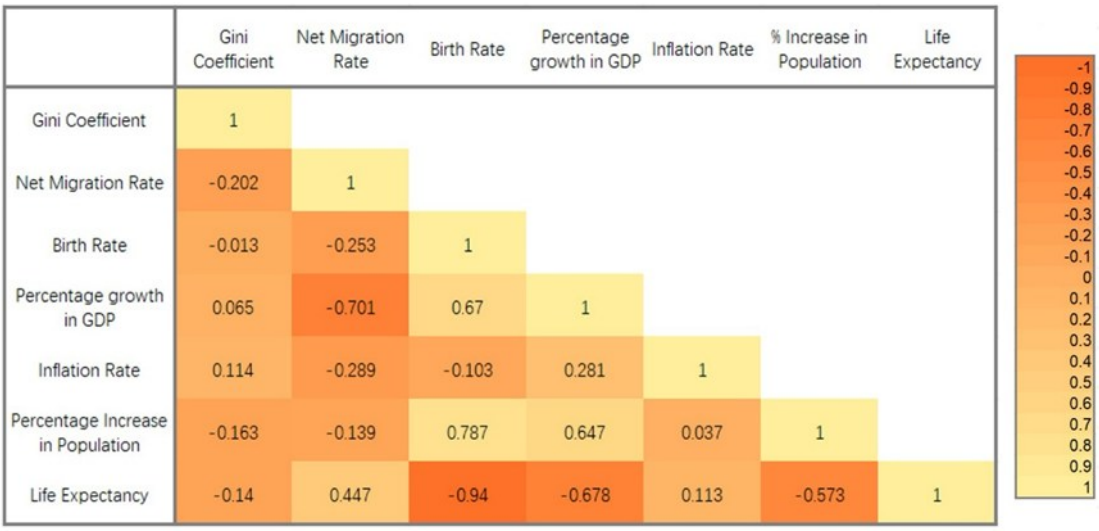

Figure 1: Variables Correlation Heatmap

Given these low individual correlations, it is hypothesized that combining these variables might produce a different result. Therefore, linear regression is going to be conducted. This analysis aims to explore whether the combined effect of these variables on the Gini coefficient is more significant than their individual effects.

Table 2: Regression Coefficient Table for China

	B	S.E.	Beta	t	p	VIF
Constant	83.204	18.209	-	4.569	0.001**	-
X_1	6.58	1.767	2.109	3.725	0.003**	12.222
X_2	-2.508	0.563	-12	-4.455	0.001**	276.679
X_3	0.89	1.721	0.204	0.517	0.614	5.92
X_4	2.738	1.224	0.446	2.238	0.043*	1.513
X_5	337.297	87.056	3.552	3.875	0.002**	32.049
X_6	-0.693	0.151	-10.244	-4.588	0.001**	190.109

Note: * represents 0.1 significance, ** represents 0.05 significance level

Table 2 shows the regression coefficient of all variables. The p values for X_1, X_2, X_4, X_5 and X_6 do not exceed 0.05, indicating that they are in close relationship with the Gini coefficient. Variables with non-zero coefficients show that some factors exacerbate income inequality while others mitigate it. For instance, X_1 and X_5, with positive coefficients, suggest that increase in net migration rate and population growth are associated with a rise in the Gini coefficient. On the contrary, X_2 and X_6 have negative coefficients, meaning that a higher birth rate and longer life expectancy are related to a lower Gini coefficient.

The R^2 value is 0.659, indicating a 65.9% level of fitting. Base on the data, the multiple linear regression equation is shown:

$$Y = 83.204 + 6.580X_1 - 2.508X_2 + 0.890X_3 + 2.738X_4 + 337.297X_5 - 0.693X_6 \quad (2)$$

3.2. Linear Regression on International Level

Although the regression analysis of China's data has listed several possible factors of income inequality, it is insufficient to rely solely on this result to summarize the influencing factors. Therefore,

DOI: 10.54254/2753-8818/2025.22561

another linear regression model will be constructed using data from 30 developed countries. The results from the two models will be compared to see if the previous findings are applicable globally.

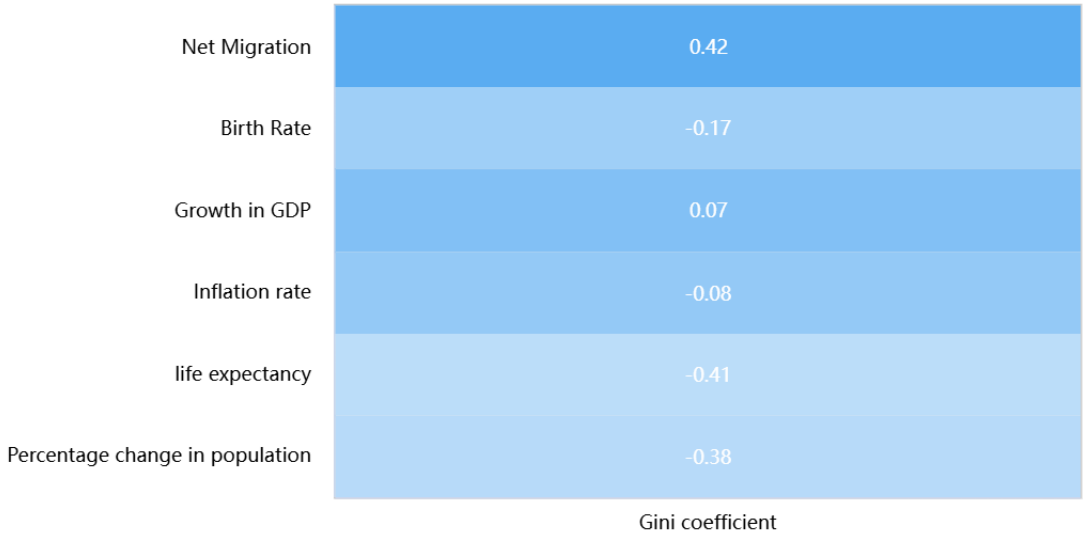

Figure 2: Pearson Correlation Chart

Figure 2 shows the Pearson Correlation Chart of each variable. From it, net migration and life expectation has the main contributions to changes in Gini coefficient with correlation coefficient of approximately 0.4. While others show a relatively lower correlation with Gini coefficient.

Table 3: Regression Coefficient Table for Selected Countries

	B	S.E.	Beta	t	p	VIF
Constant	0.824	0.471	-	1.751	0.093	-
X_1	0	0	0.571	3.558	0.002**	1.248
X_2	-0.003	0.01	-0.075	-0.347	0.732	2.278
X_3	0.005	0.003	0.28	1.675	0.108	1.356
X_4	-0.008	0.01	-0.146	-0.778	0.445	1.698
X_5	-3.987	2.578	-0.44	-1.547	0.136	3.924
X_6	-0.004	0.005	-0.179	-0.698	0.492	3.18

Note: ** represents 0.05 significance level

Table 3 shows the regression coefficient of all variables for the selected countries. Base on the data, the multiple linear regression equation is shown:

$$Y = 0.824 + 0X_1 - 0.003X_2 + 0.005X_3 - 0.008X_4 - 3.987X_5 - 0.004X_6 \quad (3)$$

Only variable X1 has p value smaller than 0.05. This indicates that in the international sample, fewer variables have a significant impact on income inequality. According to the Beta value, only X3 shows a positive relationship with the change in Gini coefficient, which is different to the previous result. Additionally, the R^2 value of 0.526 is lower than that of the Chinese model, suggesting that the model is less able to explain the variation in the Gini coefficient in this international context.

DOI: 10.54254/2753-8818/2025.22561

3.3. Comparison Between Two Models

For China, the variable correlation heatmap presented earlier indicated that there were no significant individual relationships between each variable. However, the results from the regression model shows that the combined effect of these variables on the Gini coefficient is substantial. This finding shows that seemingly unrelated variables can interact and jointly impact economic phenomena.

While comparing the two models, the one for China not only shows a stronger relationship between Gini Coefficient and each variable, but also a higher level of fitting. One possible reason for this difference is that the international model does not include several other crucial factors. In developed economies, factors such as technological innovation, globalization, and tax policies might have a more significant impact on income inequality, and these were not considered in the current model.

4. Conclusion

This study employed multiple linear regression to explore the factors influencing income inequality, using data from China from 2001 to 2020 and 30 developed countries in 2020. The findings reveal significant differences in the relationships between variables and income inequality in the Chinese and international models.

In the Chinese model, despite the low individual correlations between variables and the Gini coefficient as shown in the correlation heatmap, the regression results indicate that multiple variables, including net migration rate, birth rate, inflation rate, percentage increase in population, and life expectancy, are closely associated with the Gini coefficient. In contrast, in the international model, only the net migration rate has a significant impact on the Gini coefficient. The lower R^2 value suggests that the model is less capable of explaining the variation in income inequality compared to the previous one. Therefore, what is observed in the Chinese context may not be true on the international level, and vice versa.

Nevertheless, the study has limitations. High Variance Inflation Factor (VIF) values for many variables in the Chinese regression model, such as birth rate and life expectancy, imply strong collinearity among independent variables. This collinearity reduces the accuracy of assessing each variable's contribution. Moreover, the economic theories used rely on idealized assumptions such as perfect competition, overlooking real-world complexities such as information failure and monopolies. As a result, there is a difference between the theoretical analysis and real-world situations.

Future research could focus on addressing the collinearity issue. For example, by using more advanced statistical techniques or selecting variables more carefully. Additionally, incorporating real-world complexities into the theoretical work would enhance the understanding of income inequality and its underlying economic mechanisms, contributing to the development of more effective policies to mitigate it.

In conclusion, while this study has laid a foundation for understanding income inequality, there is still much more to be done. Through these findings, researchers can contribute to the development of more comprehensive and effective strategies for reducing income inequality both in China and globally.

References

[1] *Kakwani, N.C. (1980) Income inequality and poverty. New York: World Bank.*
[2] *Kawachi, I. and Kennedy, B.P. (1999) Income inequality and health: pathways and mechanisms. Health services research, 34(1), 215.*
[3] *Catalano, M.T., Leise, T.L. and Pfaff, T.J. (2009) Measuring resource inequality: The Gini coefficient. Numeracy, 2(2), 4.*
[4] *Monnin, P. (2014) Inflation and income inequality in developed economies. CEP Working Paper Series.*
[5] *Bulíř, A. (2001). Income inequality: does inflation matter?. IMF Staff papers, 48(1), 139-159.*

DOI: 10.54254/2753-8818/2025.22561

[6] *Mazumdar, D. (1989) Microeconomic issues of labor markets in developing countries: analysis and policy implications. World Bank Publications, 40.*
[7] *Odusola, A., Mugisha, F., Workie, Y. and Reeves, W. (2017) Income inequality and population growth in Africa. Income Inequality Trends in Sub-Saharan Africa: Divergence, Determinants and Consequences.*
[8] *Bocoum, I., Macombe, C. and Revéret, J.P. (2015) Anticipating impacts on health based on changes in income inequality caused by life cycles. The International Journal of Life Cycle Assessment, 20, 405-417.*
[9] *Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., et al. (2016) The association between income and life expectancy in the United States, 2001-2014. Jama, 315(16), 1750-1766.*
[10] *Uyanık, G.K. and Güler, N. (2013) A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences, 106, 234-240.*