# An Investigation of Factors Contributing to the Increasing Wealth Gap in South Korea Via Linear Regression: Causes and Policy Suggestions

James Joonsik Cho
Village Christian School

**Abstract**
The wealth gap has been a consistent issue in capitalistic societies, especially in the United States and South Korea. Inequality not only hurts the economy in general but also causes social and health problems in society. Hence, it is worthwhile to look into the causes of the wealth gap to have it reduced in society. In South Korea, experts believe that several factors contribute to the rising wealth gap. Some examples include rising housing prices, education fees, and more non-regular employees compared to full-time employees. This study explores the factors that specifically impact the income gap by using data mining, which is a process of finding major patterns in big data sets. The study involves the use of linear regression - a process involving data utilization to determine the correlation between variables and make predictions based on that relationship. This study utilized regression analysis to organize data as given by economic indicators and compare it to wealth gap indicators, including the gini coefficient and income quintile share ratio. Based on the findings from regression analysis, this study concludes that multiple economic factors, including rising numbers of non-regular employees, housing prices, and education fees, each contributed to the rising income quintile share ratio in South Korea. When these factors were analyzed together, they also showed a strong correlation with the rising Gini coefficient of South Korea. In other words, it is safe to assume that those factors had contributed to the rising income inequality. For future works, it will be meaningful to look into specific local data to analyze the wealth gap by area.

## 1. Introduction

Wealth refers to how many valuable possessions or money one owns. In a capitalist society, individuals can compete to produce and trade in a free market. Since the government minimizes intervention in a free market, individuals of all social classes can participate by producing, trading, manufacturing, transporting, etc. Although such a system has massive social and economic advantages over any rival system, as it allows everyone to compete and benefit

from the market, unbalances often occur, and a wealth gap inevitably appears. The rich get richer through investments of their capital, while the poor struggle to move up the social ladder because they have to spend most of their time working to pay their living costs. According to notable French economist and professor of economics, Thomas Piketty, capitalism ultimately results in an economy controlled by people born with inherited wealth. He also argues in his book that the rate of capital exceeds the speed of economic growth, which guarantees the intensification of wealth inequality in the future [1]. Wealth inequality refers to the unequal share of money among different groups in society. This study will endeavor to explain why the seemingly inevitable wealth gap in society is harmful.

An enormous wealth gap in society has many dangerous effects. Inequality negatively impacts economic growth in wealthy nations. According to a 2014 report by the Organisation for Economic Co-operation and Development (OECD), there was an increasing wealth gap in the U.S. from 1990 to 2010 has influenced GDP per capita. While the rich deposit their wealth in bank accounts, the poor spend most of their income on living costs. Sarah Anderson, the director of the Global Economy Project at the Institute for Policy Studies, analyzed the data and explained that all those dollars the poor spend have more dramatic economic effects than the dollars the rich earn. She reported that every dollar low-wage laborers earn leads to $1.21 in addition to the Gross Domestic Product. Further, there are 39 cents added to the GDP for every dollar the rich make. The widening wealth gap in society also strengthens the political influence of certain people, especially of upper class (high-income) individuals. Oren M. Levin-Waldman, professor of public policy at Metropolitan College of New York, asserted that the unequal distribution of wealth among individuals undermines democracy by allowing procedural inequality where people of disadvantaged households may not have equal access to political officials as opposed to those who are rich [3]. For instance, wealthy individuals or special interest groups try to lobby the government to influence its decisions and policies. One notable right advocacy group is the National Rifle Association. This organization had lobbied an immense amount of money to influence members of Congress on lessening regulations of gun usage in the United States. While the rich can exert influence like this by using their wealth, the poor inevitably have a weaker influence on policy-making because they lack the wealth. The OECD noted that the wealth gap impacts growth by lessening the opportunities for children from disadvantaged households to receive a quality education. The bottom 40 percent of households have trouble paying for their children to receive pricey yet high-quality education. These children tend to become less effective and productive employees compared to others who received more expensive education.

As a result, they would receive lower wages, which reduces the overall participation in the economy. Counterintuitively, even those in more financially secure positions in society are negatively affected by a wide wealth gap [4]. Specifically, owners of large retail or manufacturing businesses hope that as many people as possible can afford and buy their products. If too many people cannot afford their products, their businesses will make less profit. While it seems that a large wealth gap only impacts the economy, it also has bad influences on various areas in society, namely: safety, health, and environment. A study performed by the London School of Economics in 2016 found that areas with higher income gaps experienced more crimes, especially robberies [5]. Also, British researchers Richard Wilkinson and Kate Pickett identified a correlation between higher inequality and health & social problems, including obesity, drug use, homicides, and mental illness [6].

When examining wealth inequality, various factors have to be considered for accurate estimation. Globally, the level of income inequality has risen over the past half-century, according to data from the OECD. It reports that the wealthiest 10% earn income nine times higher than the poorest 10% in OECD nations. In China and India, millions of people were freed from poverty due to strong economic growth. However, the benefits of that growth were mainly given to the upper class, which intensified inequality [7]. According to Katherine Schaeffer, a research analyst from the Pew Research Center, the earnings of the wealthiest 20% in the United States in 2018 accounted for more than 50 percent of collective GDP for the U.S. that year. The U.S. is recorded to have the highest income inequality among the G7 nations in a report given by the OECD [8]. The OECD defines the poverty rate as the percentage of people with an income lower than half the nation's median income. It published that the poverty rate of South Korea in 2019 ranks second, right after the U.S.: 17.4 percent and 17.8 percent, respectively. The wealth gap in South Korea has been exacerbated massively over the past few decades. Also, the poverty rate of seniors and youth is significantly higher in Korea compared to other nations, including Japan, Australia, Norway, and France [9].

This study will investigate the rising wealth gap of South Korea and offer recommendations to minimize the damages caused by the ever-growing wealth gap. The first regression will analyze wealth inequality by metropolitan cities and provinces, while the second will look at national disparities. Factors (independent variables) that will be used for the first analysis include the cost of the private education of high schoolers, the number of universities per metropolitan city, and businesses by industry. According to Scott A. Wolla and Jessica Sullivan of the Federal Reserve Bank of St. Louis, there is a strong correlation between income and level of education. They reported that individuals with higher levels of education owned more

properties on average. This is why spendings on education is a key indicator of the wealth gap between households in society. Also, the number of universities by area is important because having more universities leads to improvements in the infrastructure of that area, which would lessen the wealth gap in areas that already have a high median income. Factors (input) for the second analysis are the percentage of non-regular employees, youth unemployment rate, and housing price changes. As mentioned before, the youth unemployment rate is a substantial social issue currently. Also, the rising percentage of non-regular employees and increase in housing prices will help illustrate and analyze Korea's wealth inequality. Hence, the main hypothesis of this paper is that the wealth gap of South Korea has widened due to the increase in the following factors: number of non-regular employees, youth unemployment, average housing prices, and average spendings on private education. The wealth gap in this study is determined by looking at the Gini coefficient (a measure of income inequality) or the income quintile share ratio, which is a ratio of the total income earned by the richest 10% (top quintile) of the nation to the income earned by the poorest 10% (bottom quintile). Each factor will be analyzed separately through linear regression to check how much impact each factor has on the wealth gap. Also, some of the factors will be analyzed together through multi-factor regression, which would show a combination of factors influencing the result (wealth gap). This study predicts that a combination of the aforementioned economic factors would strongly correlate with the growing wealth gap.

**2. Context of the South Korean economy**

As mentioned in the introduction, this study attempts to analyze the wealth gap in South Korea and make suggestions to minimize such a gap. A thorough analysis of the current wealth gap requires one to understand the factors that exacerbated the wealth gap in the first place. During the early 1990s, South Korea saw huge economic developments due to the growth of major companies, including Hyundai and Samsung. At this time, many foreign investors became interested in the Korean market as multiple companies were rapidly developing. Unfortunately, this period of economic growth did not last long. The economy began to collapse as a result of the 1997 Asia Financial Crisis. By then, Korea saw a huge shortage in foreign currency (mainly U.S. dollars), and its government realized the need for foreign currency to subsidize its domestic industries. This led the value of Korean currency (won) to fall very quickly, impacting the economy significantly) [10]. As a result, the Korean government asked for financial assistance from the International Monetary Fund (IMF). During this financial crisis period, more than twenty major company groups went bankrupt, namely KIA group, Daewu group, Ssangyong group, and Haetae group. This crisis marked the beginning

of a high unemployment rate for Korean youths: companies that survived the crash began to employ fewer people. According to Statistics Korea, the youth employment rate fell from 58% to 51.5% after the financial crisis occurred from 1997 to 1998. This low employment rate led to various changes in the labor market. Young people began searching for government jobs because there is a lower risk of being fired once they earn the position. Around the early 2010s, the unemployment rate was showing a decreasing trend, and younger workers were able to find jobs more easily. Unfortunately, this pattern in the labor market did not last long. From 2016 to 2018, the youth unemployment rate reached over 9%, which is much higher than that of 2012 (7.5%). Due to this social issue, many politicians in 2018 made various election promises relevant to the youth employment rate. Jaein Moon, the winning candidate in the 2018 election, planned to bring down the youth unemployment rate by creating thousands of new government jobs. According to an article from Joseon-Ilbo, youth unemployment in Korea continues to be over 10% from February 2021 to May 2021. While it is true that the COVID-19 pandemic has had a major impact socially and economically, it is clear that the issue with the youth unemployment rate started much earlier than the pandemic and is far from being resolved. Lessening the increasing wealth gap requires more debates and plans on decreasing the youth unemployment rate.

Another major impact of the financial crisis was the institution of part-time employment opportunities instead of full-time workers to cut input (labor) costs. Around the late 90s and early 2000s, there were increasing protests and strikes by part-time employees (National Worker's Politics Association). The percentage of part-time employees was especially high in a few industries: shipbuilding and car manufacturing. The issue is that the number of part-time employees continues to rise, despite protests and strikes. According to an article from Yonhap News Agency, non-regular employees hit 36% - the highest in twelve years. It reported that there are about 7.5 million non-regular workers as of 2019. Since a vast number of people cannot secure their positions for a long time, this study asserts that the wealth gap would widen among people who were able to find secure jobs and those who were not. The third major impact of the 1997 financial crisis was the beginning of a differential wage gap between small and medium-sized enterprises (SME) and large companies. This widening wage gap contributed to the high wealth gap in Korea.

Over the past two decades, another significant social issue has been present in Korean society: increasing housing prices. Considering inflation over time, it makes sense that housing prices have to increase. However, the housing prices in Korea have skyrocketed since 2019, and they are only continuing to rise [11]. One of the serious problems resulting from increasing housing prices is that the gap between the poor and the

rich is ever increasing. While the poor have to invest more time to earn enough money to afford a house, the rich get wealthier due to an increase in the value of their possessions, including apartments, buildings, etc. Also, the super-rich can reinvest their money into new houses to increase the value of their possessions in the future because experts claim it is clear that housing prices in certain areas, especially Seoul, will not fall for a while [12]. According to an article from Dong-a Ilbo, a popular Korean newspaper, since 1920, the housing prices in Seoul continued to rise even after the government introduced new regulations for the housing market. The major issue is that the housing prices for not just Seoul but cities near Seoul, including Incheon, Seongnam, and Gimpo, are all continuing to increase. In order to decrease the wealth gap in Korean society, the government must pass new policies that will help to bring down the housing prices.,

Within Korea, there is a significant wealth gap by area. While the capital city of South Korea (Seoul) has a solid infrastructure, other cities down south do not. As this gap between cities became noticeable, a new slang that explains this situation went viral on the internet: "The Republic of Seoul." Even though this term went viral recently, there have been constant complaints towards the Korean government's excessive investment in Seoul, neglecting other cities and provinces [13]. Hence, in 2004 the Korean government planned on changing the capital to a different city to activate more economic activities down south. Unfortunately, this plan was not executed, and Seoul continued to remain the capital of South Korea. As a result, a high number of universities and businesses are concentrated in Seoul and near Seoul. According to statistics provided by Statistics Korea (a national agency that arranges and uploads various stats), 48 universities are located in Seoul, and 61 universities are located in Gyeonggi-do (the province that surrounds Seoul). This is about one-third of the total universities in South Korea. In other words, other metropolitan cities and provinces do not have enough universities, despite their big sizes. Numerous benefits come along with having more universities in an area. It will attract thousands of students, and those students will spend money in that area for years until they graduate. As a result, the local economy is stimulated as students visit restaurants to eat and spend money on entertainment, including movies and sports games [14]. When Non-Seoul local economies are stimulated, the wealth gap between Seoul and other cities will lessen.

## 3. Related works

### A. Linear Regression

Regression analysis refers to the utilization of data to determine a correlation between variables and create predictions based on that relationship. Specifically for linear regression, one assumes that the result one estimates is dependent linearly on the data utilized to predict. When a factor is linearly dependent on a

different factor, it increases or decreases concerning the other factor at a constant rate.

There are four assumptions involved: linearity, homoscedasticity, independence, and normality. The first term means that a linear relationship exists between variable X and variable Y. The next assumption notes that residuals have constant variance at every X. Independence means residuals or errors are independent. Lastly, normality means the residuals are distributed normally.

A linear regression model sets up a relation between one dependent variable (Y) and other explanatory variables (X1, X2... X.K.). When a regression model is demonstrated as a mathematical equation, it generally looks like the following:

$Y = \beta 0 + \beta 1X1 + \cdots + \beta KXK + \varepsilon$. ... (eq1), where $\varepsilon$ represents the error term.

**a. Single Linear Regression**

For such a model to be properly used, one has to prepare data of Y values that correspond to Xi values. This broad form of the regression model can be changed when analyzing simple linear regression, which is a model that uses only one explanatory variable. Hence, the equation for this is much simpler:

$$Y = \beta 0 + \beta 1X + \varepsilon. \ldots (eq2)$$

Regression analysis usually allows one to draw the line of best fit. This line is represented by the least-squares criterion. In this line, some points go through the observed data points, but most other points do not match up with the observed data points. The coordinates are usually marked as (X1, Y1)... (X#, Y#). The mathematical model for this line is the following:

$\hat{Y}i = \beta 0 + \beta 1\ Xi$ . ...(eq3) Where the $\hat{Y}i$ represents the predicted value of Y.

The difference between the predicted value $\hat{Y}i$ and the observed value Yi represents as an error (residual):

$ei = Yi - \hat{Y}i = Yi - (\beta 0 + \beta 1Xi)$. ...(eq4)

Here, linear regression makes use of $\beta 0$ and $\beta 1$ to minimize the value of the sum of squared errors. Conclusively, out of all the possible lines that can be drawn, regression chooses to use the one that minimizes the sum of squared errors:

SSE: $\sum_{i=1}^{n} e(i)^2$... (*eq5*)

To minimize SSE, $\beta 0$ and $\beta 1$ should be $\frac{\sum Yi - \beta 1 \sum Xi}{n}$, $\frac{\sum YiXi - \bar{y}\sum Xi}{\sum (Xi * Xi) - \bar{x}\sum Xi}$. ... (*eq6,7*)

**Proof**

First, we can say the target function is (*eq5*): $\sum ei^2 = \sum (Yi - aXi - b)^2 = S$

To find which $\beta 0$ and $\beta 1$ can make S meet the local extreme points, we partially differentiate S by $\beta 0$ and $\beta 1$. We start with $\beta 0$ as it is the constant of (*eq3*):

<u>The partial derivative of $\beta 0$</u>

$\frac{\partial \sum ei2}{\partial \beta 0}$= 2 ∑(-1)(Yi - β1Xi - β0)

The goal is to find where the partial derivative of β0 is 0.

0 = -2 (∑Yi - β1∑Xi - ∑β0) = -2 (∑Yi - β1∑Xi - n*β0), $\because \sum_{i=1}^{n}$ β0 = n * β0, 0 = ∑Yi - β1∑Xi - n*β0

n*β0 is sent to the left side to find the value of β0.

n*β0 = ∑Yi - β1∑Xi , β0 = $\frac{\sum Yi - \beta 1 \sum Xi}{n}$

When $\sum Yi$is divided by n, it adds all the y values present and divides how many numbers of y values there are. This equals the average value of y. $\bar{y}$ = average value of y,

β0 = $\bar{y}$ - β1$\bar{x}$

<u>The partial derivative of β1</u>

$\frac{\partial \sum ei2}{\partial \beta 1}$= (-1) 2 ∑(Yi - β1Xi - β0) (Xi)

The goal is to find where the partial derivative of β1 is 0 on the left side.

0 = -2 [∑YiXi - β1∑($Xi^2$) - β0∑Xi], β1∑($Xi^2$) = ∑YiXi - β0∑Xi

Substitute ($\bar{y}$ - β1$\bar{x}$) for β0, β1 is sent to the left side.

β1 [∑($Xi^2$) - $\bar{x}$∑Xi] = ∑YiXi - $\bar{y}$ ∑Xi = $\frac{\sum YiXi - \bar{y}\sum Xi}{\sum(Xi*Xi) - \bar{x}\sum Xi}$

**Example 1**: <*table 1*> assumes we can predict the local GDP per capita via only a variable 'private education.'

<*table 1*>

| **i (regions)** | **Xi** | **Yi** |
|---|---|---|
| Average 전체 | 36.5 | 37,530 |
| Seoul 서울 | 55.6 | 44,865 |
| Busan 부산 | 36 | 27,409 |
| Daegu 대구 | 37.5 | 23,744 |
| Incheon 인천 | 36.3 | 30,425 |
| Gwangju 광주 | 29.7 | 27,548 |
| Daejeon 대전 | 36.4 | 28,364 |
| Ulsan. 울산 | 30.2 | 65,352 |
| Sejong 세종 (2012 년 설립) | 37.3 | 35,826 |
| Gyeonggi-do 경기 | 41.4 | 36,133 |
| Gangwon-do 강원 | 21.9 | 32,061 |
| Chungcheongbuk-do 충북 | 23.6 | 42,653 |
| Chungcheongnam-do 충남 | 23.8 | 52,402 |
| Jeollabuk-do 전북 | 26.7 | 28,740 |
| Jeollanam-do 전남 | 19.7 | 43,323 |
| Gyeongsangbuk-do 경북 | 21.6 | 40,272 |
| Gyeongsangnam-do 경남 | 24.4 | 33,690 |
| Jeju-do 제주 | 28.1 | 30,720 |

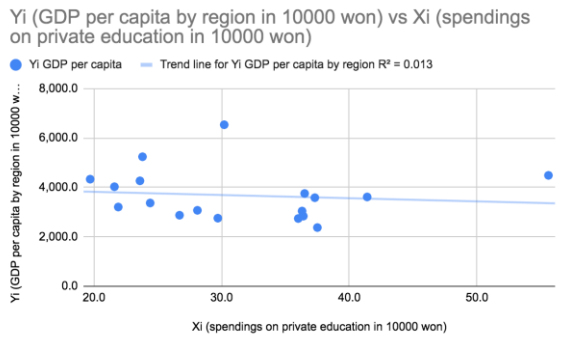
Using linear regression, Yi(GDP per capita by region in 1000 won) can be predicted using random Xi(spendings on private education in 10000 won).

<*figure 1*> represents the scatter graph of <*table 1*> with its linear regression model (a line graph). By applying eq6,7 into $Y\hat{}i = \beta0 + \beta1\ Xi$,

we get $\beta0$ = 4,082 and B1 = -13.

Let's assume that we want to predict Yx when Xx = 35:

Yx = 4082 + -13 x 35 = 3627




<*figure 1*>

**b. Evaluation of the Fitness**

There are a few more terms and notations to take into consideration when using regression analysis. Some examples include D.F. (degree of freedom), S.S. (sum for squares), and M.S. (mean square). Mean squares equal the sum of squares, which accounts for variability, divided by the degree of freedom.

$$R^2 = \frac{explained\ variability}{total\ variability} = \frac{SSR}{SST}$$

The equation above shows that $R^2$ value is a ratio of the explained variability to the total variability.

$$SST = \sum_{i=1}^{n} (Yi - \bar{y})^2$$

The observed dependent variables are represented as $Y_1,\ldots,Y_n$ and their sample mean is $\bar{y}$. In this case, the total sum of squares is defined as the equation above.

$$Y\hat{}i = B0 + \beta1\ Xi$$

The equation above is similar to sample variance, except that n-1 is not yet divided. $Y\hat{}i$ refers to the y prediction value based on the assumption value of Xi. While B0 is an estimate of the intercept, $\beta1$ is the slope given by the linear regression.

$$SSR = \sum_{i=1}^{n} (Y\hat{}i - \bar{y})^2$$

The equation above defines the regression of the sum of squares (explained variability from $R^2$ definition).

$$ei = Yi - Y\hat{}i$$

The i-th residual above is defined as the difference between the real value of y and the predicted value of y.

$$SSE = \sum_{i=1}^{n} (ei^2)$$

The error of the sum of squares is defined as above.

$$SSR + SSE = SST$$

The addition of explained variation and unexplained variation leads to the total variation.

$$R^2 = \frac{SSR}{SST}$$

The Total S.S. indicates the total variability in what is being measured, while the Regression S.S. shows explained variation. When SSR is divided by SST, it equals the coefficient of determination or R-square. It's important to keep in mind that a high R-square value implies strong explanatory power while a low R-squared value indicates the opposite. For simple linear regression, a high R-square value indicates a strong relationship between the two variables.

The maximum value of SSR is SST when the error value SSE becomes 0. In other words, the value of R-square should be between 0 and 1. Generally, the standard of 'good' value of R-square is from 0.45. Even though the value of R-square is not 'good enough, it doesn't imply the corresponding factor has no relation at all. The value might get higher with other combinations of the factors.

**Example 2**: In **Example 1**, we find a linear line in <*figure 1*> for data set in <*table1*>.

However, the linear line does not demonstrate a strong relationship between x and y, which is why the $R^2$ value is quite low. Since $R^2$ value is low by 0.013, there is not much significance in assuming and predicting a value. It infers that private education fee as a single factor does not affect GDP so much.

**c. Multi-linear regression**

In section b, we covered single linear regression. Mining data considering a factor with linear regression definitely gives us some useful facts. But some indicators like wealth gap indicators usually are affected by a bunch of factors. In an extension of coefficients,

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_x x_x + \epsilon_i$$

The formulas below are used in multi-linear regression:

$$b_1 = \frac{\sum x_{2i}^{*2} \sum y_i^* x_{1i}^* - \sum x_{1i}^* x_{2i}^* \sum y_i^* x_{2i}^*}{\sum x_{1i}^{*2} \sum x_{2i}^{*2} - \left(\sum x_{1i}^* x_{2i}^*\right)^2}$$

,,

$$b_2 = \frac{\sum x_{1i}^{*2} \sum y_i^* x_{2i}^* - \sum x_{1i}^* x_{2i}^* \sum y_i^* x_{1i}^*}{\sum x_{1i}^{*2} \sum x_{2i}^{*2} - \left(\sum x_{1i}^* x_{2i}^*\right)^2}$$

...

To prove this, we need just to consider more partial deviations.

d. **Modules in Python**

We used Python to implement linear regression. Many powerful libraries can be applied via Python. In this research, linear regression fit functions are from the Sklearn module. Numpy is used to format the data.

**a. Sklearn**

Sklearn in Python is one of the most useful libraries for machine learning that can be accessed in Python. It includes various tools for both machine learning and statistical modeling like regression. Since linear regression is used as

an important modeling technique in this paper, sklearn was used.

**b. NumPy**

NumPy is a library in Python used for scientific computing. Its system allows advanced mathematical equations and operations to be executed more easily because fewer codes are involved than Python's built-in sequence. For this reason, this library was also used to facilitate computing and other operations involved in regression.

**c. Implementation**

| part | code |
|---|---|
| 1 | from sklearn import linear_model<br>import numpy as np<br>import pandas as pd<br>import matplotlib<br>import matplotlib.pyplot as plt |
| 2 | data = {'x': [13,19,16,14,15,14],'y':[40,83,62,48,58,43]}<br>data = pd.DataFrame(data)<br>data.plot(kind="scatter",x="x", y="y",figsize=(5,5), color = "black") |
| 3 | linear_regression = linear_model.LinearRegression()<br>linear_regression.fit(X=pd.DataFrame(data['x']), y = data["y"]) |
| 4 | prediction = linear_regression.predict(X=pd.DataFrame(data['x']))<br>print(linear_regression.intercept_, linear_regression.coef_) |
| 5 | residuals = data['y'] - prediction<br>SSE = (residuals**2).sum()<br>SST = ((data['y']-data['y'].mean())**2).sum()<br>R_squared = 1 - (SSE/SST)<br>print(R_squared) |
| 6 | data.plot(kind="scatter",x="x", y="y",figsize=(5,5), color = "black")<br>plt.plot(data['x'],prediction,color = 'red')<br>plt.show() |

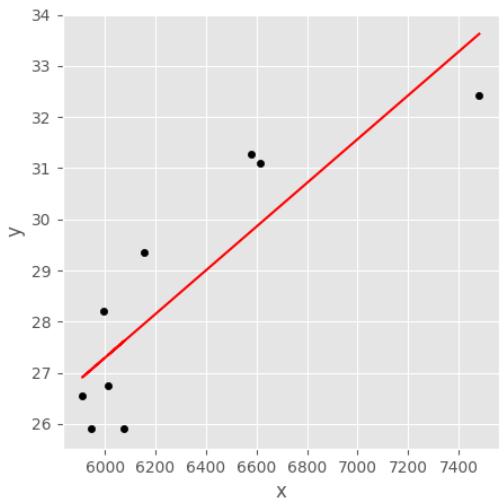
In Part 1, a few codes are added to activate the sklearn model to prepare linear regression on the collected data. Part 2 is where the data of x and y are added. To test out multiple sets of data, this part can be altered accordingly. In part 3, two codes are selected to prepare for linear

regression and data input. The first code in part 4 functions to show the prediction of linear regression on the final result (graph). This prediction is made based on the pattern and relationship between points of x and y from part 2. In part 5, multiple codes are entered to get the value of R squared ($R^2$). This process is necessary because the R squared value presents whether the values of x and y are linearly dependent on each other or not. The higher the R squared value means x and y are more strongly related or dependent. To get the value of R squared, the values for residuals, SSE, and SST have to be determined beforehand since R squared equals 1 - SSE/SST. At last, two codes are put in to plot the data points on a graph in black and demonstrate the prediction of this regression in red.

**4. Prediction and Analysis**

**A. Single Factor**

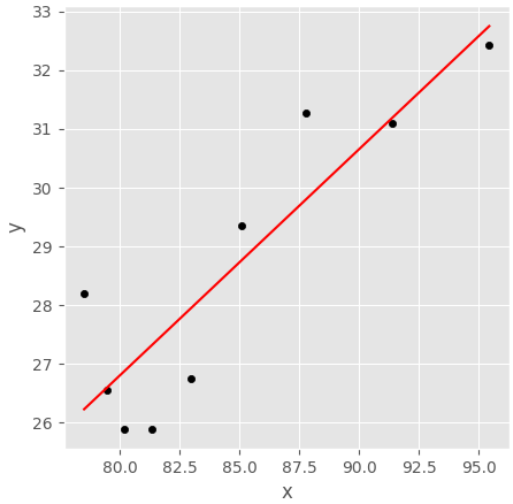
a. number of non-regular employees per year as x, income quintile share as y




<number of non-regular employees per year vs. income quintile share ratio>

This is a linear regression graph that indicates the linear relationship between the number of non-regular employees per year in South Korea and the income quintile share ratio, which is a ratio of the total income earned by the richest 10% (top quintile) of the nation to the income earned by the poorest 10% (bottom quintile). In other words, the higher the income quintile share ratio means, the higher the income gap between the rich and poor. The x coordinates refer to the number of non-regular employees of a specific year, and the y coordinates indicate the income quintile share ratio of that year. The $R^2$ of this linear relationship is 0.75, which indicates that a strong correlation exists between x and y. This is unlike the example on page 10. With strong correlation, a prediction value of y, which is earned by assuming a random number for x, becomes more accurate. Hence, this relationship implies that the increasing number of non-regular employees each year led to an increased wage gap between the rich and the poor.

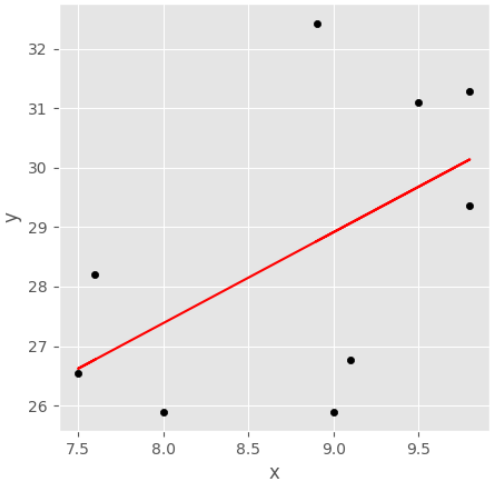
b. average housing prices per year as x, income quintile share as y




<average housing prices per year vs. income quintile share ratio>

The graph above indicates the relationship between the average housing prices of Korea per year and the income quintile share ratio. The average housing prices are estimated by Statistics Korea each year, and they are represented as x coordinates on the graph above. Similar to the first graph, this graph also displays a strong linear relationship between x and y because the $R^2$ value is quite high: 0.786. A relationship with a strong $R^2$ value is safe to use prediction values from it because they tend to be more accurate than predictions from graphs or relations with low $R^2$ values. Hence, it would be safe to assume that the increasing housing prices of Korea led to an increased income gap between the rich and the poor.

c. yearly youth unemployment rate as x, income quintile share as y

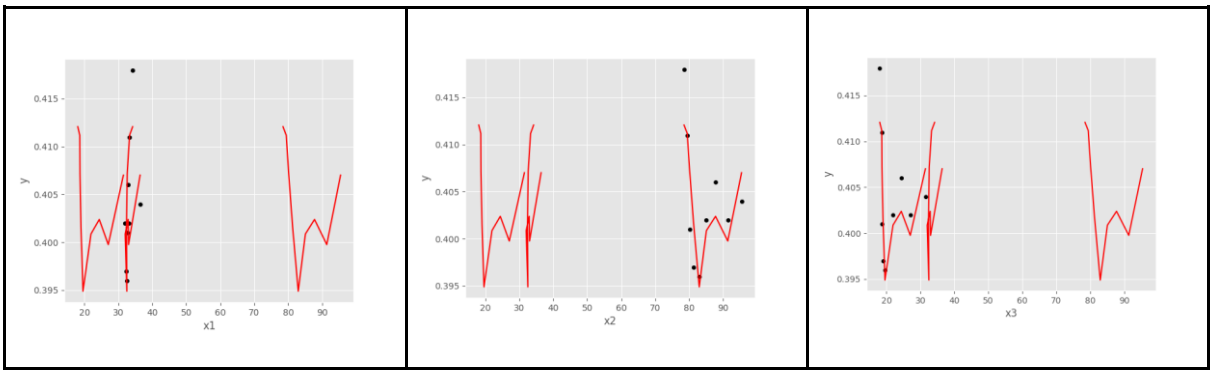



<yearly youth unemployment rate vs. income quintile share ratio>

The graph above displays the relationship between South Korea's yearly youth unemployment rate and the income quintile share ratio. The annual youth unemployment rate is released by Statics South Korea, and those data are represented as x coordinates on the graph. The y coordinates are the income quintile share ratios of the same year x coordinate is representing. Unlike the two charts above, this does not show a strong correlation between x and y, which is why the $R^2$ value is lower than the ones above(0.293). In other words, getting a prediction value of y by assuming a random number for x would not be accurate or significant since the relationship holds no strong ties.

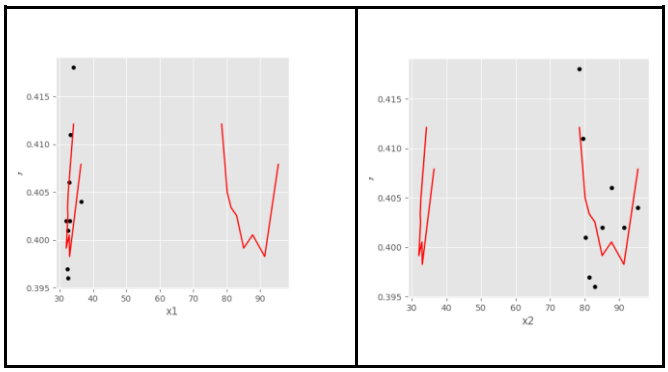
**B. Multi-Factor**

a. Youth unemployment rate as x1, housing prices as x2, average education fee as x3, Gini coefficient as y




The three graphs above show a regression involving multiple x variables, unlike the ones with a single factor. For this regression, the three x variables used are the following: annual youth unemployment rate, average housing prices, and average private education fee. The y variable used is the yearly Gini coefficient, which is a measure of income equality in South Korea. Graphs of a multi-factor regression do not display a linear line because they require more than two dimensions to properly demonstrate the relationship between x variables and the y variable. This particular regression requires four dimensions because it has three different x variables and one y variable. Hence, it is more meaningful to look at the $R^2$ value rather than the graphs above. Since the $R^2$ value is pretty big(0.672), this specific regression shows a strong correlation between x variables and the y variable. In other words, the increase in the annual youth unemployment rate, average housing prices, and average spendings on private education every year all contributed to the increase in the Gini coefficient of Korea.

b. Youth unemployment rate as x1, average housing prices as x2,
Gini coefficient as y




The two graphs above demonstrate a multi-factor regression involving two independent variables: the annual youth unemployment rate and average housing prices per year of Korea. The y variable of this regression is yearly Gini coefficient. Similar to what happened with the

regression above, the graphs for this regression do not show linear lines because they need three dimensions: there are two different x variables and one y variable involved. Hence, the $R^2$ value should be analyzed more deeply. Even though 0.446 is not a low value, it is still not high enough to assume that a strong correlation exists between x variables and the y variable. It would be safe to state that a weak correlation exists between independent and dependent variables. One important thing to take away from this regression is that only one factor (average spendings on private education) was removed from the previous multi-factor regression, but the $R^2$ value decreased by 0.226.

**5. Conclusion and Future works**

Effective policies are brought about by breaking down social and economic issues into smaller parts, analyzing these in detail, and considering how they can synergize with other factors. Thus, making effective policies is similar to producing good results in data mining as these policies are proven to be effective or not if they result in growth rates and wealth distribution. Good results in data mining indicate that findings are accurate and match human intuitions. In this paper, linear regression, a method of data mining, was used to do factor analysis and find the connection between factors and the wealth gap in South Korea.

Multiple factors (number of non-regular employees, youth unemployment, average housing prices, and average spendings on private education) were analyzed in this study to see whether these had an impact on the rising income gap. According to the single factor regressions and multi-factor regressions from section 4, some factors correlated with the growing wealth gap, which means the hypothesis of this paper is partially correct. As has been demonstrated, as the number of non-regular employees increases, the wealth gap also increases. Therefore the Korean government must do its utmost to protect workers' rights and ensure that more employees have access to full-time, contracted long-term work contracts. On the other hand, one factor that shows a weak connection with the wealth gap is the yearly unemployment rate. For multi-factor regression, which involves the following three factors; youth unemployment rate, average housing prices, and average spending on private education, there is a strong connection with the rising Gini coefficient, proving that the hypothesis is correct. Based on the findings from section 4, explicit factors contributed to the increasing wealth gap of South Korea. In other words, when new policies are made to deal with these economic factors, the Gini coefficient or the income quintile share ratio will likely fall. Those factors include increasing numbers of non-regular employees, average housing prices, and spendings on private education. Making practical policy suggestions calls on being sensible to look into specific policies that deal with similar issues in the past. For example, the United States saw huge economic growth in the

1990s, but it had a high wealth gap compared to other OECD nations: the Gini coefficient was 36.8, which was 6 higher than the OECD average. The U.S. government in 1964 declared a "war on poverty." The Johnson administration passed policies that increased spendings on public education, medical care, and programs that dealt with rural poverty. The main point of those policies was for the poor to have more access to education to escape poverty and move up the social ladder. Although America was not very successful in reducing poverty through those policies, this example is worthwhile to note because those policies can be modified to fit the current situation in Korea. As mentioned earlier, the average spending on education has increased in Korea. However, the issue is that while the rich can spend more money on children's education, the poor continue to struggle to pay the increasing costs of private education. Hence, the Korean government should pass new policies that expand educational opportunities for the lower class. For example, the Seoul Metropolitan government started a new program that gave lower-class students access to online tutoring and courses, which helped them to prepare for the Korean standardized test. This will help disadvantaged students to get accepted to prestigious universities and earn scholarships from organizations. As of now, this program is only available to students in Seoul. The Korean government should open access to this program to disadvantaged students in all areas of Korea.

In this work, we analyzed the gap between rich and poor with a focus on the individual. Income inequality can be seen as a dependent result on its own, but broadly speaking, it can be seen as a contributing factor in the disparity between rich and poor in the whole country.

To analyze this, it is necessary first to define the indicators of the gap between the rich and the poor between regions. The gap between the rich and the poor between regions can be expressed in numerous indicators. For example, the Korean government provided multiple rounds of Covid-19 emergency relief funds to citizens. The fifth round was given to people in the bottom 88 percent of the nation's income range. About 70% of citizens in Seoul received the relief fund, while 90% of people in other cities or provinces received the money. This indicates that people with high income are more concentrated in Seoul than in any other place.

For future works, this study recommends the investigation of local economic data to find out if specific factors contribute to the gap in GDP by region. Those indicators can then be used to compare with the Gini coefficient. Moreover, we can predict the future wealth gap by estimating the data of annual economic indicators that affect the wealth gap.

**Reference**

[1] "Capital in the Twenty-First Century" by Thomas Piketty

[2] "Wall Street Bonuses and the Minimum Wage" by Sarah Anderson

[3] "How Inequality Undermines Democracy" by Oren M. Levin-Waldman

[4] "How rising inequality hurts everyone, even the rich" by Christopher Ingraham

[5] "Social Disadvantage, Crime and Punishment" by Tim Newburn

[6] "The Spirit Level: Why More Equal Societies Almost Always Do Better" by Kate Pickett and Richard Wilkinson

[7] Inequality by OECD

puses" by Ji-hye Jeong

[8] "6 facts about economic inequality in the U.S." by Katherine Schaeffer

[9] "Korea has 2nd-highest income gap in OECD" by Yon-se Kim

[10] "The Korean Financial Crisis - Causes, Effects and Solutions" by Terry Black and Susan Black

[11] "House prices are running again… Will the 2019 situation repeat?" by Hyeon-gyu Hwang

[12] "The real reason apartment prices don't fall" by Hyung-seok Shim

[13] "We live in the Republic of Seoul" by Sungkyunkwan university newspaper

[14] "Seoul to revive commercial areas around university cam