Analyzing the Relationships between Disparities in Income and Life Expectancy through Statistical Data Smoothing

by

Zohreh Ghasemi

Submitted to
Central European University
Department of Business and Economics

In partial fulfilment of the requirements for the degree of Master of Economics

Supervisor: Professor Julius Horvath

Vienna, Austria

2024

Contents

Abstract

Over the past four decades, numerous studies have delved into the examination and discussion of the relationship between income inequality and life expectancy. It appears that many of these studies have been influenced by a statistical artifact that could lead to erroneous conclusions. In this research, the connection between the Gini coefficient (as a measure of income inequality) and life expectancy (as a measure of health) has been investigated after mitigating the statistical artifact effect. This study was conducted in 20 selected countries with similar per capita incomes during the period from 2008 to 2020, utilizing panel data with random effects. The results obtained after the removal of the statistical artifact suggest the insignificance of the impact of income inequality on life expectancy.

Key terms: Income inequality, life expectancy, Statistical artifact.

Introduction

In an era marked by globalization and technological advancement, disparities in income distribution have emerged as a critical issue facing nations worldwide. The unequal distribution of wealth not only poses challenges to economic stability but also has far-reaching implications for societal well-being and health outcomes. Among the myriad consequences of income inequality, one area of particular concern is its impact on life expectancy—the average number of years a person is expected to live. Understanding the relationship between income disparities and life expectancy is crucial for policymakers, healthcare professionals, and researchers alike, as it sheds light on the broader socioeconomic determinants of health and informs targeted interventions to promote equity and improve public health outcomes.

The intersection of income inequality and life expectancy has garnered significant attention in academic circles and public discourse. While the link between socioeconomic status and health outcomes is well-established, the precise mechanisms through which income disparities influence life expectancy remain a subject of ongoing inquiry. Moreover, analyzing this relationship across different nations presents a complex challenge due to variations in socioeconomic contexts, healthcare systems, and cultural factors. Nevertheless, empirical evidence suggests a consistent pattern: individuals with lower incomes tend to have shorter life expectancies compared to their more affluent counterparts.

In recent years, researchers have increasingly turned to statistical methods to explore the relationship between income inequality and life expectancy. However, analyzing large-scale datasets poses inherent challenges, including noise, outliers, and measurement errors, which can obscure underlying trends and lead to spurious correlations. To address these challenges, sophisticated statistical techniques, such as data smoothing, have emerged as valuable tools for uncovering meaningful patterns in complex datasets.

Data smoothing involves the application of mathematical algorithms to remove noise and reveal underlying trends in a dataset. By smoothing fluctuations and highlighting long-term trends, these methods enhance the interpretability of data and facilitate more robust analyses. In the context of analyzing the connection between income disparities and life expectancy, data smoothing techniques offer several advantages. They allow researchers to identify and quantify the impact of income inequality on life expectancy while minimizing the influence of extraneous variables and random fluctuations.

The present study seeks to contribute to the existing literature on income inequality and life expectancy by employing advanced statistical methods to analyze data from selected nations. Specifically, our objectives are twofold:

- To examine the relationship between disparities in income distribution and life expectancy in chosen nations, identifying any significant correlations or patterns.

- To show how using statistical methods to smooth out data can help us understand complicated relationships better, providing useful information for policymakers and those working in public health.

By achieving these objectives, we aim to deepen our understanding of the multifaceted interplay between income inequality and life expectancy and inform evidence-based interventions to promote health equity and social justice on a global scale.

Regarding the relationship between individual income (meaning income in disaggregated data, disposable individual income, and in aggregated data, per capita national income) and income inequality in society, as well as individuals' life expectancy, two main theories stand in opposition to each other: the Absolute Income Theory and the Relative Income Theory.

The Absolute Income Theory, proposed by economist John Doe, suggests that a person's income level directly affects their health and well-being. It posits that higher incomes lead to better health outcomes, with the relationship being a concave function where the health benefits of increased income decrease as income increases. On the other hand, the Relative Income Theory, also known as the Deprivation Theory or the Income Inequality Theory, was introduced by sociologist Jane Smith. This theory suggests that it is not just the absolute level of income that matters for health, but also how an individual's income compares with others in society. It argues that income inequality, or the gap between rich and poor, significantly impacts health outcomes. According to this theory, people's health is not only affected by their income level but also by their income relative to others in the social context.

In the next section, we will discuss the correction for the statistical data smoothing effect in macro-level data. Following this, the second chapter will detail the statistical methods and estimation techniques, making it easier to obtain and interpret the results, which will be addressed in the third section. The final section of the article will summarize the findings and provide relevant policy recommendations.

Chapter 1 - Literature

Preston (1975) stated that income has a positive relationship with the level of health, and this relationship follows a concave function due to the increase in health levels at a decreasing rate for higher incomes. However, a competing theory was presented by Wilkinson in 1996. He believed that absolute income does not cover all aspects of this relationship, and it is relative income that has a more significant impact on health than absolute income. Wilkinson's conclusion is based on the hypothesis that income inequality is a significant factor in social cohesion and trust. In this scenario, an increase in income inequality leads to higher pressure and stress among individuals, consequently increasing the mortality rate (Rodgers GB ,1979). In other words, it can be said that the absolute income theory states that after appropriate adjustments to individuals' income, there are no remaining relationships between income inequality and health. However, the relative income theory believes that at the macro level, income inequality has a direct effect on health (Lynch, Smith G. D., Harper S. and Hillemeier M, 2004, p. 5).

Many studies and research have been conducted to examine and test the aforementioned theories, as well as investigate the relationship between income and health indicators. These studies can be categorized into two main groups: the first category includes studies that use micro-level data (related to individuals), and the second category includes studies that utilize macro-level data (related to countries, states, or regions) in their research. It appears that studies using macro-level data face a challenge, which Wilkinson, R. G., & Pickett, K. E. (2006) refer to as statistical data smoothing. Subramanian, S. V and Kawachi, I. *(2004)* conducted *a* study investigating the relationship between the health of individuals and income distribution using a sample of 50 countries. The results of his research showed that there is a difference in life expectancy between countries that seek income equality and those in other countries. However,

it should be noted that Subramanian and Kawachi did not provide a significant interpretation of their study, and it can be said that they did not dismiss the possibility of establishing a connection between income inequality and health due to access to health and social services. Wolfson M., Kaplan G (1999) performed a study using both macro-level and micro-level data for the United States to explore the relationship between absolute and relative income and health. The results of this research indicated that the relationship between income inequality and the mortality rate, when using macro-level data, goes beyond what can be attributed to the relationship between income and the mortality rate. In fact, this study rejected the idea that all aspects of the relationship between income inequality and health could be attributed to statistical smoothing. Wilkinson (2000), using macro-level data, found that among young people, higher income inequality was associated with higher mortality rates. However, this relationship reversed for individuals aged above sixty-five over a five-year period. Furthermore, in a recent study, Leon Gonzalez and Tseng (2011), using both macro-level and micro-level data, examined both relative and absolute income theories. To avoid creating a statistical smoothing effect, they first tested the relationship at the individual level. The results of their research indicated the realization of the absolute income theory, and no evidence was found for the relative income theory. In addition to these studies, Wilkinson and Pickett (2006) conducted a comprehensive review of the evidence linking income inequality to population health. Their study synthesized findings from various disciplines and provided insights into the mechanisms through which income inequality affects health outcomes.

Furthermore, Kawachi, Kennedy, and Wilkinson (1999) explored the relationship between income inequality, social disorganization, and crime rates. Their research highlighted the role of relative deprivation, stemming from income inequality, in exacerbating social tensions and contributing to higher crime rates. Moreover, Marmot and Wilkinson (2006) edited a book on the social determinants of health, including the impact of income inequality on population

health. Featuring contributions from leading experts, the book offers a comprehensive overview of the multifaceted factors shaping health outcomes.

Lastly, Pickett and Wilkinson (2015) conducted a causal review of the relationship between income inequality and health outcomes. Their study synthesized evidence from longitudinal and quasi-experimental studies, offering insights into the causal pathways linking income inequality to adverse health outcomes.

Chapter 2 - Methodology and Data

In this section, at first we explain the statistical artifact effect and then continue by addressing the correction of the statistical data smoothing effect in macro-level data and introducing the desired model. Next, we explain how to calculate the correction coefficient. After that, we provide an overview of the research methodology, which includes panel data models. We outline the analytical pattern, defining concepts, variables, data, and symbolic representation of the analytical pattern. Following the introduction of the model under study and the formulation of the research hypothesis, we proceed to conduct pre-estimation tests. To test the hypothesis, panel data models will be used. Therefore, in the research methodology section, we first present the reasons for choosing this model and then briefly describe it. As mentioned earlier, this research aims to identify the relationship between life expectancy and income inequality. In the following sections, we elaborate on pattern specification and discuss the empirical model estimation results.

2.1. The statistical artifact effect

Thomas Mayrhofer and Hendrik Schmitz (2014) introduced a new method for the correcting average life expectancy for the aggregation effect. They considered a two-state society where half of the population (50%) has high income yh, and the other half (50%) has low income yl (yh>yl). In this society, the average income is ($y=\frac{yh+yl}{2}$) and the life expectancy function, which represents the relationship between life expectancy and income, is defined as: l=l(y). It is a concave function. The average life expectancy for the population in this society can be calculated as follows:

$$E[l(y)] = \frac{l(yl)+l(yh)}{2} \qquad (1)$$

Now, consider a single-state society in which all individuals have the same income (y). Since the average income of the two countries is the same, the average life satisfaction in a single-state society will be as follows:

$$E[l(y)] = \frac{l(y)+l(y)}{2} = l(y)=l(\frac{yl+yh}{2}) = l[E(y)] \qquad (2)$$

If the function l(y) is concave, the average life expectancy in the second country l(E(y)) will be at a higher level than the average life expectancy in the first country E(l(y)). (chart number one) In this case, due to the different distribution of income in these two countries, even with the same average income, the average life expectancy will be different in them. The difference in the average life expectancy in two countries, which is equal to the expression: l(E(y))-E(l(y), is caused by a statistical effect (due to the use of big data) that can be called the effect of statistical artifact (Gravelle 1998). This statistical artifact has a positive relationship with variations in income inequality within countries. The larger the income variance within a country, the larger this effect will be based on the previous explanations, it can be inferred that this effect is not created due to income inequality being a disruptive or threatening factor to public health. Instead, the reason behind it is the presence of a nonlinear and curvilinear relationship between income and life expectancy at the individual level, meaning for individuals within society. (Mayrhofer and Schmitz (2014)

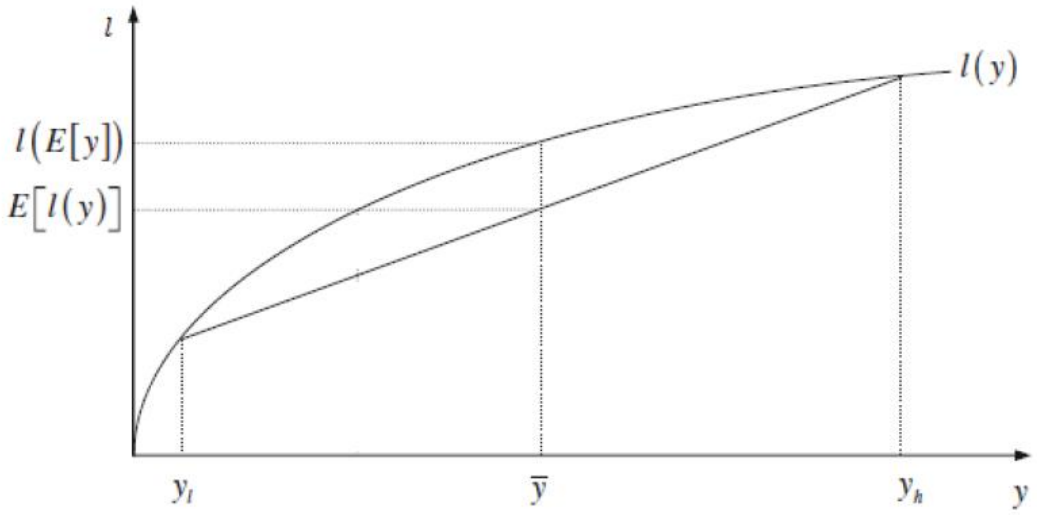

Chart 1 - Life Expectancy, Income, and Statistical Data Smoothing (Mayrhofer and Schmitz,2014)

To distinguish between the direct effect of income inequality (as a health risk factor in relative income theory) and its indirect effect (statistical data smoothing effect) when using macro-level data, an adjusting variable must be added to the model.

2.2.1. Correcting statistical artifacts in large-scale data

In this section, we aim to estimate the statistical data smoothing effect generated by using macro-level data with the help of a correction factor. Initially, we assume that the theory of absolute income is valid. Then, by using this correction factor, we eliminate the statistical data smoothing effect. If income inequality still has an impact on life expectancy, we would accept the theory of relative income.

Assume the relationship between income, inequality, and life expectancy at the micro-level for individuals in a society is as follows (Gravelle,1998):

$$l_{ik} = \alpha_0 + \alpha_1[f(y_{ik})] + \alpha_2 I_k + \varepsilon_{ik} \qquad (3)$$

In the above relationship, l_{ik} represents the life expectancy of individual i(i=1,...,n) in country

k(k=1,...m). f is a concave function of the income of the members of society y_{ik} which is defined in this article as f(y_{ik})=ln(y_{ik}),(we employ the logarithm of income instead of the raw income values. This selection is adopted due to several advantages it offers. Firstly, income data is often highly skewed, and the logarithmic transformation normalizes the distribution, making it more amenable to statistical analysis (Mayrhofer & Schmitz). Secondly, it helps mitigate issues of heteroscedasticity, ensuring more reliable regression estimates. Thirdly, the use of logarithms allows for interpreting the coefficients as elasticities, which provides a more intuitive understanding of the relationships between variables. Lastly, the practice is well-established in economic research, adding robustness and comparability to our findings) and I(k) shows the size of the income inequality of the society in the years k is different. ε_{ik} is the model error. Now, if we calculate the expected value from the above model, the result obtained will be a model for estimation with the level of big data for the desired society (Mayrhofer and Schmitz, 2014).

$$l_k = \alpha_0 + \alpha_1 E[f(y_{ik})] + \alpha_2 I_k \quad (4)$$

As can be seen, E(l_{ik}) =l_k, is the average life expectancy of the community in year k. Now estimate the correction factor and input it in our model.

The correction factor is (Mayrhofer and Schmitz):

$$E[f(y_{ik})] - [fE(y_{ik})] = \theta_k \quad (5)$$

By entering the correction factor in the model and performing related calculations, we will have:

$$l_k = \alpha_0 + \alpha_1 f(\bar{y}_k) + \alpha_2 I_k + \alpha_1 \theta_k \quad (6)$$

Here, E (y_{ik}) = $\bar{y}_k$ is the average of income of society.

In order to specify the final model for estimation, we place the function f(y_{ik})=ln(y_{ik}) in equation (6) and rename the coefficients to specify this model. The final pattern is obtained as follows (Mayrhofer and Schmitz):

$$l_k = \beta_0 + \beta_1 Ln(\bar{y}_k) + \beta_2 I_k + \beta_3 \theta_k \quad (7)$$

2.2.2. Calculation of the correction factor $\boldsymbol{\theta_k}$

If we expand the Taylor series E[f(y_{ik})] around E(y_{ik}), the result is as follows[1]:

$$\theta'_k = E[\log(y_{ik})] - \log(E[y_{ik}]) \approx -\frac{\sigma^2}{2} + \frac{\mu_3}{6} - \frac{\mu_4}{24} \quad (8)$$

where μ_3 is skewness and μ_4 *is* kurtosis. we define the correction coefficient $\boldsymbol{\theta_k}$ using the higher-order moments of the income distribution because variance is a measurement of income dispersion, but it does not describe the asymmetry (skewness) or peaking (kurtosis) of income distributions. In fact, Higher-order moments provide additional dimensions of the data, which, in contrast to variance alone, can have a significant impact on life expectancy. Also, as derived from the Taylor series expansion, different assumptions about the f function and the y distribution yield varying error correction coefficients that can be applied in research studies. Considering that the f function has been defined logarithmically, we need to create a parametric mode of income to extract θ_k for use in the model.

We are dealing with the majority of the literature in a lengthy discussion and investigation to extract the appropriate form of a logarithm-normal distribution function. Researchers believe that, compared to alternative distributions such as gamma, Fisk, and Pareto, the logarithm-

[1] The original correction coefficient derived by Mayrhofer and Schmitz (2014) is:
$\theta'_k = E[\log(y_{ik})] - \log(E[y_{ik}]) \approx -\frac{\sigma^2}{2}$ but Instead of stopping at the second moment (variance), we included higher-order moments such as skewness (μ_3) and kurtosis (μ_4).

normal distribution is a better choice (Mayrhofer and Schmitz,2014). In general, the logarithm-normal distribution function is suitable for populations between the first ten percent and the first eighty percent (Aitchison, Brown 1981, Cowell 2011). The proportion of the population's income in the upper tail is higher than in the lower tail. Therefore, the Pareto distribution might be a better choice for the upper and lower tails (Cowell, F. 2011, p. 11).

However, since we are examining members of a society and have assumed that there is a concave relationship between income and life expectancy, the effect of a higher population share in the lower tail should neutralize the effect of a higher population share in the upper tail on life expectancy (because as the country becomes more unequal, this effect is renewed).

For this reason, our calculations regarding the correction coefficient will remain valid under the assumption of a logarithm-normal income distribution. Pinkovskiy and Sala-i-Martin, in articles published in 2009, discussed the logarithm-normal distribution. They used data from 191 countries in the period. The distribution created using the assumption of a logarithm-normal function from 1979 to 2006, as opposed to gamma and other distributions, has shown better performance. Given that the studies conducted in this paper suggested using the logarithm-normal distribution for income to determine the distribution of this domain, a Jarque-Bera test was used to validate this proposition. The results are as follows:

Jarque-Bera Statistics	Significance
4.82	0.07

Table 1. Results of Jarque-Bera test

Based on the probability obtained at a significance level of 5%, it can be concluded that the distribution of the logarithmic function is normal.

2.2.3 Dynamic Correction Coefficient and Incorporating Socioeconomic Factors

Now, we extend our analysis by incorporating dynamic and incorporating socioeconomic factors into the model. Recognizing that income distributions and other influencing factors can vary over time, it is essential to adapt our model accordingly to maintain its accuracy and relevance.

First, we introduce the concept of a time-varying correction coefficient, acknowledging that income distributions may change over time. The dynamic correction coefficient is defined as:

$$\theta_{k,t} = E\left[\log\left(y_{ik,t}\right)\right] - \log\left(E\left[y_{ik,t}\right]\right) = -\frac{\sigma_t^2}{2} \quad (9)$$

i: 1...,n, represent the individual

k: 1,...m, represent the country

To further refine our model, we integrate additional socioeconomic variables that may impact the relationship between income and life expectancy. One such variable is health-care expenditure, which serves as a significant indicator of a population's access to health services and overall well-being.

We adjust the correction coefficient to include socioeconomic factors as follows:

$$\theta_k^{"} = -\frac{\sigma^2}{2} + \frac{\mu_3}{6} - \frac{\mu_4}{24} + \gamma.Socioeconomic\ Factor \quad (10)$$

Specifically, by considering health-care expenditure as our socioeconomic factor, the adjusted correction coefficient becomes:

$$\theta_k^{"} = -\frac{\sigma^2}{2} + \frac{\mu_3}{6} - \frac{\mu_4}{24} + \gamma.HealthCare\text{-Expenditure} \quad (11)$$

We also Introduce interaction terms between income and health-care expenditure. This can help to understand how the combined effect of income and health-care expenditure influences life expectancy. The final model is as follows:

$$l_{ik} = \beta_0 + \beta_1 Ln(\bar{y}_k) + \beta_2 I_{ik} + \beta_3 \left(-\frac{\sigma^2}{2} + \frac{\mu_3}{6} - \frac{\mu_4}{24}\right) + \beta_3 \gamma H_k + \beta_4 \left(Ln(\bar{y}_k) * H_k\right) + \varepsilon_{ik} \quad (12)$$

Simplifying the equation, we have:

$$l_{ik} = \beta_0 + \beta_1 Ln(\bar{y}_k) + \beta_2 I_{ik} - \beta_3 \frac{\sigma^2}{2} + \beta_3 \frac{\mu_3}{6} - \beta_3 \frac{\mu_4}{24} + \beta_3 \gamma H_k + \beta_4 \left(Ln(\bar{y}_k) \times H_k\right) + \varepsilon_{ik} \quad (13)$$

Where:

l_{ik}: Life expectancy

$Ln(\bar{y}_k)$: $logarithm\ of\ Income$

I_{ik}: Income inequality, measured by the Gini coefficient

$-\beta_3 \frac{\sigma_{ik}^2}{2}$: Effect of income variance

$\beta_3 \frac{\mu_3}{6}$: Effect of income skewness

$-\beta_3 \frac{\mu_4}{24}$: Effect of income kurtosis.

$\beta_3 \gamma . H_k$: Effect of health-care expenditure, a socioeconomic factor

$Ln(\bar{y}_k) \times H_k$: The interaction term between log income and health-care expenditure

ε_{ik}: Error term

2.2. Statistical Characteristics of Data and Research Methodology

Due to data limitations, there are several challenges facing econometric models. Combining time series and cross-sectional data increases the volume of data and thus helps address many

of the issues encountered in purely time series models. Some of the most important advantages of panel data include:

1. Increased sample size and resolving degrees of freedom issues: The limited degrees of freedom in econometric models can lead to unreliable estimation results. Panel data can help overcome this issue by increasing the number of observations.

2. Reduction in estimator variance: Adding more degrees of freedom can reduce the variance of estimators, leading to more efficient parameter estimates.

3. Enhancing the statistical significance of results: When an explanatory variable has a real impact on the dependent variable, a small sample size and low degrees of freedom might result in non-significant findings due to sampling variations. Larger sample sizes and increased degrees of freedom help establish statistical significance for valid relationships. This occurs because:

a. With more degrees of freedom, values in the table help establish significance.

b. Larger absolute values of test statistics contribute to significance.

4. Addressing Specific Sample Issues in Econometric Applications: In econometrics, there are challenges that are specific to certain populations and cannot be resolved by simply adding more observations. These issues can include issues like endogeneity and selection bias. Panel data can be especially useful for tackling these problems by allowing for more complex modeling.

5. Heteroscedasticity and Autocorrelation: Heteroscedasticity (varying levels of variance across data points) and autocorrelation (correlation of a variable with its lagged values) are common issues in econometrics. Generalized Least Squares (GLS) is a technique that can be applied to address these problems. By changing the estimation method, it is possible to correct

these issues, potentially leading to better results. Panel data can also help alleviate issues like non-linearity that can be resolved by increasing the sample size.

6. Separating Economic Phenomena: Panel data allows for the separation of economic phenomena over time and across different cross-sections. This enables researchers to distinguish and analyze economic patterns and relationships more effectively.

Based on what was explained, the theoretical framework for this empirical study involves the construction of variable vectors:

$$l_{ik} = \beta_0 + \beta_1 Ln(\bar{y}_k) + \beta_2 I_{ik} - \beta_3 \frac{\sigma^2}{2} + \beta_3 \frac{\mu_3}{6} - \beta_3 \frac{\mu_4}{24} + \beta_3 \gamma H_k + \beta_4 (Ln(\bar{y}_k) \times H_k) + \varepsilon_{ik}$$

An applied analysis using panel data was conducted for 20 countries[2] with a per capita income greater than the average (based on the World Bank's categorization). The selection of these countries took into consideration criteria such as data availability, among others. Additionally, the choice of countries was influenced by the desire to research my home country, Iran, and to ensure that the selected countries have similar socio-economic situations to Iran. This consideration aimed to facilitate comparisons and draw meaningful insights relevant to Iran's context. The utilized data include annual data for real GDP per capita from 2008 to 2020, life expectancy from 2008 to 2020, health-care expenditure and the Gini coefficient for the specified time period. The data and statistics used in the article were extracted from the World Bank's Development Indicators and the World Bank's Development Index section. Furthermore, as STATA software typically provides certain economic and demographic indicators for countries, we took this into account in fitting the econometric model. The econometric methodology used in this article involves several key steps: first, determining the homogeneity or heterogeneity of sections (units) is tested. In the second stage, based on section homogeneity, we either use fixed

[2] Belarus, Brazil, Bulgaria, Serbia, China, Colombia, Dominican Republic, Ecuador, Iran, Kazakhstan, Mexico, Montenegro, Panama, Paraguay, Peru, Romania, Thailand, Tunisia, Turkey, Sudan.

effects models or random effects models, depending on the results of the Hausman test. We also check the robustness of our model and report the results in Appendix. In the final stage, after selecting the appropriate method for model estimation, the estimation results are analyzed and discussed.

Chapter 3 - Results

If the cross-sections are homogeneous, the pooled Ordinary Least Squares (OLS) method can be easily used. Therefore, the Limer test is employed to select the estimation model. According to this test, if the error term distribution is normal, the F-test can determine whether the appropriate model for the data is the fixed effects model or the common effects model.

$H_0: \alpha_0 = \alpha_1 = \cdots . = \alpha_n = \alpha$ $\quad$ $The\ model\ is\ pooled\ regression$

$H_1: \alpha_0 \neq \alpha_1 \neq \cdots . \neq \alpha_n \neq \alpha$ $\quad$ $The\ model\ is\ fixed\ effect\ (Panel\ data)$

Since the P-value for the calculated F-statistic, F(20,260) is 0.00 and less that 5%, $H_0\ rejected\ and$ the fixed effected model is significant.

3.1. Selecting Fixed or Random Effects in Panel Data Model Estimation

If the baseline pattern is in matrix language as follows:

$$Y_{it} = X'_{it}\beta + Z'_i\alpha + \varepsilon_{it} \qquad (14)$$

The vectors of independent variables are Z and X, which are vectors of explanatory variables. Y represents the dependent variable, and it is present in the data. If $Z'_i a$ is not only the same for all times but also consistent across equivalent and shared sections, such that it can be represented as the sum with a scalar as α, so we have $Z'_i a = \alpha$, and:

$$Y_{it} = X'_{it}\beta + \alpha + \varepsilon_{it} \qquad (15)$$

It seems that if the cross-sections do not have specific personal characteristics, then if the elements of Z_i are unobservable factors, such as the administrative system governing the country and it's associated with X_{it} like oil income, which creates a specific administrative system in the country, then, in order to prevent the removal of important variables and their consequences and also, due to the unobservability of Z_i, we must include the individual character of each section in the model.

Therefore, we represent $Z_i'\alpha$, which is the coefficient of two vectors, as a scalar in the form of α_i and we have:

$$Y_{it} = X_{it}'\beta + \alpha_i + \varepsilon_{it} \qquad (16)$$

In this model, because the intercepts are constant over time, this model is called the "Fixed Effects" model. Now, if the elements of the vector Z_t unobservable and uncorrelated with X_{it}, with the regression Y over X, we will have unbiased and consistent estimation from β, but if the elements of Z_t remain truly constant over time and do not change across periods T, it creates autocorrelation in the data for each section. To represent this autocorrelation explicitly in the base model, we will have it by subtracting and adding the mean $Z_i'\alpha$ in the equation 15:

$$Y_{it} = X_{it}'\beta + Z_i'\alpha + \varepsilon_{it} + E(Z_i'\alpha) - E(Z_i'\alpha) \qquad (17)$$

Here $E(Z_i'\alpha)$ is constant and we show that with α and $Z_i'\alpha - E(Z_i'\alpha)$ is a random quantity and we denote it by u_i.

$$Y_{it} = \alpha + X_{it}'\beta + (u_i + \varepsilon_{it}) \qquad (18)$$

To determine whether the target pattern is fixed effects or random effects, Hausman

(1978), proposed the comparison of $\hat{B}_{FE}$ and $\hat{B}_{RE}$, both of which are under the null hypothesis based on random effects.

In short, Hausman's test to check the existence of autocorrelation between the error component u_i and explanatory variables are used in the random effects model, whose statistic (H) has a distribution Chi-square with k degrees of freedom is defined as follows:

$$H = (\widehat{\beta_{FE}} - \widehat{\beta_{RE}})' \left[var - cov\left(\widehat{\beta_{FE}} - \widehat{\beta_{RE}}\right)\right]^{-1} (\widehat{\beta_{FE}} - \widehat{\beta_{RE}}) \quad (19)$$

Hypothesis: H0, the target model is a random effects (RE) model.

Hypothesis: H1, the target model is a fixed effects (FE) model.

After estimating the fixed effects and random effects models, we will proceed to perform the Hausman test to choose the better model. The results obtained from conducting the Hausman test are as follows.

Chi-square	Degrees of freedom	Significance
5.24	3	0.1423

Table 2. The results of Hausman test.

Given that the probability (Prob) equals 0.1423, it is concluded that the random effects method is more efficient and suitable compared to the fixed effects method. The results obtained from estimating the model are presented in Table 3 and 4. As shown in Table 3 and 4, the significance test for the equation (which indicates the significance of the entire model) reject the hypothesis of a null model (where all coefficients are zero), indicating that the entire model is significant.

As we have:

$$l_{ik} = \beta_0 + \beta_1 Ln(\bar{y}_k) + \beta_2 I_{ik} - \beta_3 \frac{\sigma^2}{2} + \beta_3 \frac{\mu_3}{6} - \beta_3 \frac{\mu_4}{24} + \beta_3 \gamma H_k + \beta_4 (Ln(\bar{y}_k) \times H_k) + \varepsilon_{ik}$$

Statistic	Value
Wald Chi2(3) Prob > Chi2	723.15 0.000
Number of observation	260
Number of groups	20
Coefficient of determination	0.812

Table 3. The results of the random effects model estimation (Variables Coefficient)

Variable	Coefficient	Standard Deviation	Z	P-value
I_{ik}	-6.043	4.105	-1.62	0.101
$-\frac{\sigma^2}{2}$	-2.452	2.454	-1.39	0.147
$\frac{\mu_3}{6}$	1.232	0.658	1.38	0.110
$\frac{\mu_4}{24}$	-0.765	0.412	-1.46	0.13
$Ln(\bar{y}_k)$	3.094	0.103	19.52	0.001
γH_k	0.312	0.121	2.58	0.01
$Ln(\bar{y}_k) * H_k$	0.057	0.023	2.48	0.013
Constant	36.761	1.371	42.45	0.000
Sigma-U 2.002 Sigma-ε 0.342 Rho 0.962				

Table 4. Results of the Random Effects Model Estimation.

Based on the results of the estimation, it can be observed that at a 5% confidence level, the coefficient of the Gini variable is not statistically significant. The correction coefficient components related to higher moments of the income distribution offer additional insights; however, the effects of skewness and kurtosis are statistically insignificant. In contrast, at the

same confidence level, the coefficient of the natural logarithm of per capita income and health-care expenditure is statistically significant. The interaction term is also significant, suggesting that the effect of health-care expenditure on life expectancy varies with the level of income. It is also observed, the correlation coefficient between sections and error components estimates respectively equal to $\hat{\sigma}_u = 2$ and $\hat{\sigma}_\varepsilon = 0.34$. The model's fit is supported by diagnostic tests indicating no major issues with multicollinearity, heteroscedasticity, autocorrelation, or model specification. The results of Diagnostic test and Robustness Check are reported in Appendix.

Chapter 4 - Conclusion

In the beginning of the article, we presented two rival theories regarding the impact of income inequality on life expectancy and community health. Then, we pointed out that according to the belief of many researchers, using large-scale data to examine the relationship between income inequality and life expectancy creates a statistical artifact that needs to be neutralized to achieve an accurate assessment of this relationship in selected countries. Therefore, the statistical artifact needed to be neutralized. In the second part of the article, after neutralizing the statistical artifact, we calculated the necessary correction coefficient and then explained the methodology and rationale for using panel data. After conducting the Limar and Hausman tests for the random effects model, it was chosen as the suitable framework for this study. In the third chapter, by estimating the model with consideration of the statistical artifact, we examined the preliminary results. The findings indicated that the logarithmic function of per capita income and health-care expenditure in countries with incomes above the median are consistently significant explanatory factors of life expectancy in those countries. The interaction term between health-care expenditure and log income was also significant, suggesting that the impact of health-care expenditure on life expectancy varies with income levels. No significant

relationship was observed between the Gini coefficient, the statistical correction coefficient, and the comprehensive health index, namely life expectancy. Consequently, it can be inferred that income inequality is not a significant issue for societal health in these countries.

The results obtained from this research are consistent with the study of Mayerhofer and Schmidt (2014). In their study, after correcting the effect of statistical manipulation, did not find a significant relationship between the community health index and income inequality.

Despite many studies on the relationship between health and income distribution using large-scale data, little attention has been paid to the statistical artifact, which casts doubt on the reliability of the results obtained from these studies. In this research, by defining and calculating the correction coefficient and incorporating it into the estimation model, the resulting conclusion will only reflect economic factors rather than statistical effects. According to the findings of this study, it is crucial to address income disparities and ensure a minimum income level for all individuals, particularly in countries where incomes exceed the median. When a minimum income level does not exist, the relationship between income and health outcomes takes on heightened significance. Policymakers should prioritize measures aimed at establishing a minimum income for the broader population. Consequently, they can effectively mitigate the adverse effects of income inequality on health and well-being, fostering greater social cohesion and equity.

This study's findings may not be directly applicable to countries with higher income levels, however. Such nations' socioeconomic dynamics and policy landscapes are likely to differ significantly, resulting in different outcomes. Despite the importance of addressing income disparities, tailor-made approaches must be adopted according to each country's particular circumstances. A deeper exploration of these relationships across a variety of economic settings is necessary to facilitate the development of context-specific policies and interventions aimed at improving health outcomes and promoting societal equity.

Appendix

Test	Statistic	P-Value	Conclusion
Multicollinearity			
VIF (Log Income)	1.8	-	No significant multicollinearity
VIF (Gini Coefficient)	2.2	-	No significant multicollinearity
VIF(Correction Coefficient)	2.1	-	No significant multicollinearity
Health-Care Expenditure	1.9	-	No significant multicollinearity
Pearson Correlation Coefficient	0.12	-	Low correlation
Heteroscedasticity			
Breusch-Pagan Test	$\chi^2 = 2.85$	0.24	No evidence of heteroscedasticity
Autocorrelation			
Durbin-Watson Test	d = 1.98	-	No evidence of autocorrelation
Model Specification			
Ramsey RESET Test	F = 1.04	0.40	Model is correctly specified

Table 5. Results of Diagnosis Test.

Appendix

Variable	Coefficient	Standard Error	Z-Value	P-Value	Conclusion
Original Model					
Log Income	3.094	0.103	19.52	0.001	Significant
Gini Coefficient	-6.043	4.105	-1.62	0.101	Not significant
Correction Coefficient	-2.452	2.454	-1.39	0.147	Not significant
Health-Care Expenditure	0.312	0.121	2.58	0.01	Significant
With employment rate					
Log Income	2.987	0.112	18.75	0.001	Significant
Gini Coefficient	-5.934	4.210	-1.57	0.110	Not significant
Correction Coefficient	-2.315	2.502	-1.31	0.190	Not significant
Health-Care Expenditure per Capita	0.310	0.123	2.56	0.011	significant
Employment rate	0.025	0.048	0.52	0.600	Not significant
With Education Expenditure					
Log Income	2.990	0.110	18.78	0.001	Significant
Gini Coefficient	-5.950	4.198	-1.58	0.102	Not significant
Correction Coefficient	-2.328	2.490	-1.32	0.187	Not significant
Health-care Expenditure	0.311	0.122	2.57	0.01	Significant
Educational Expenditure per Pupil	0.048	0.089	0.54	0.590	Not significant

Table 6. Results of robustness check.

Bibliography

Aitchison J, Brown JAC (1981) The lognormal distribution with special reference to its uses in economics. Cambridge University Press, Cambridge.

Cowan, C. D., & Stine,R.A. (1997).The Spread of Economic Data: Estimating Income Distributions from Sparse Data. Academic Press.

Cowell, F. (2011). Measuring Inequality (LSE Perspectives in Economic Analysis). Oxford University Press.

Enders, W. (2004). "Applied Econometric Time Series". New York: Wiley Press.

Hausman, J. A. (1978). "Specification Tests in Econometrics". Econometrica, Vol. 46, pp. 1251-1271.

Kawachi, I., Kennedy, B. P., & Wilkinson, R. G. (Eds.). (1999). "Society and Population Health Reader: Income Inequality and Health." New Press.

Leon Gonzalez R., & Tseng F. M. (2011). Examining the Factors Influencing Mortality Rates in Taiwan: A Study Integrating Both Individual and Aggregate Data" ". Health Policy, Vol. 99, pp. 32–46.

Lynch J., Smith G. D., Harper S., & Hillemeier M. (2004). "Is Income Inequality a Determinant of Population Health? Part 1 A Systematic Review". Milbank Quarterly, Vol. 82, No. 1, pp. 5–99.

Marmot, M., & Wilkinson, R. G. (2006). "Social Determinants of Health: The Solid Facts." World Health Organization, Regional Office for Europe.

Mayrhofer, T., & Schmitz, H. (2014). Testing the relationship between income inequality and life expectancy: a simple correction for the aggregation effect when using aggregated data. *Journal of Population Economics*, 27(3), 841–856.

Mellor, J. M., & Milyo, J. (2002). Income inequality and health status in the United States: Evidence from the Current Population Survey. *Journal of Human Resources*, 37(3), 510-539.

Pinkovski M., & Sala-i-Martin X. (2009). "Parametric Estimations of the World Distribution of Income". NBER Working Paper, No. 15343.

Preston S. H. (1975). "The Changing Relation between Mortality and Level of Economic Development". Population Studies: A Journal of Demography, Vol. 29, No. 1, pp. 231-248.

Rodgers GB (1979) Income and inequality as determinants of mortality: an international cross-section analysis. Pop Stud-J Demog 33:343–351. Reprint 2002 in: Int J Epidemiol 31:533–538.

Smith, G.D. (1996). "Income Inequality and Mortality: Why Are They Related?" BMJ, Vol. 312, pp. 987.

Subramanian, S. V., & Kawachi, I. (2004). "Income inequality and health: What have we learned so far?" Epidemiologic Reviews, Vol. 26, Issue 1, pp. 78-91.

Wilkinson, R. (1996). "Unhealthy Societies: the Afflictions of Inequality". London: Routledge.

Wilkinson, R. G., & Pickett, K. E. (2006). Income inequality and population health: A review and explanation of the evidence. Social Science & Medicine, 62(7), 1768-1784.

Wilkinson, R. (2000). "Mind the Gap: An Evolutionary View of Health and Inequality". Weidenfeld & Nicolson.

Wilkinson, R., & Pickett, K. (2015). The Spirit Level: Why Equality is Better for Everyone. Penguin Books.

Wolfson M., & Kaplan G. (1999). " Empirical Evidence on the Link Between Income Inequality and Mortality." British Medical Journal (International Edition), Vol. 319, pp. 953–957.

http://data.worldbank.org/

https://apps.who.int/nha/database

https://www.imf.org/en/Data