**Is money the Elixer of mankind? Exploring the Link Between Life Expectancy and National Wealth and the Income Inequality.**

Qianren Pan

1009391846

STA302: Data Analysis

Professor: Antonio Herrera Martin

June 17, 2024

There is a saying: “Health is wealth,” but is it always like this, or could it be the other way around? This project investigates the multifaceted relationship between national wealth and life expectancy and, ultimately, the potential income inequality dilemma. Life expectancy, a fundamental measure of population health, reflects the average lifespan individuals can expect in a particular society. The difference in life expectancy amongst all places can be a consequence of multiple reasons; it could be the medical condition, the race, the society, the economy, etc. However, the average life expectancy is already enough to paint a big picture of the economy’s impact on life expectancy. As technology and the economy progress, people start to realize that there are more and more elders, and people are living longer. While it seems natural that people in wealthier countries should have a longer lifespan, we might wonder if this statement actually holds. This project delves into this critical question by exploring the relationship between three key national indicators: healthcare expenditure, income composition of resources (ICOR), GDP, and income inequality (measured by the Gini index). The previous three predicators are indexes for measuring and understanding a country’s wealth. Whereas the Gini index provides a quantifiable measure of income distribution within a country, offering insights into potential disparities that may influence health outcomes, this question drew my attention because life and death are always important and inevitable portions of human life, and I want to specifically investigate how life expectancy can be impacted. Existing research has established a connection between poverty and life expectancy. For instance, Lawanson and Umar (2021) found a positive relationship between economic growth and life expectancy in Nigeria, suggesting that economic development might contribute to an increased lifespan. Further, Felice, Pujol, and D’Ippoliti (2016) revealed a complex two-way relationship between GDP and life expectancy in Italy and Spain. While economic growth appeared to boost life expectancy, longer lifespans also seemed to have a reciprocal effect on economic growth. This project will specifically analyze data for Canada, exploring whether healthcare expenditure, ICOR, GDP, and Gini index are associated with the lifespan of Canadians. By examining these relationships, this research aims to provide a more nuanced understanding of how economic factors impact population health, shifting the focus from solely relying on GDP as a measure of national well-being. The findings could have implications for policymakers seeking to improve existing health outcomes for Canadians.

For this project, datasets from “Our World in Data” (https://ourworldindata.org/), “Kaggle” (https://www.kaggle.com/), and “World Bank Group” (https://www.worldbank.org/en/home) will be used; further citations will be included on the last page. Three datasets will be used; each dataset will contain at least observations from 2000 to 2015. The first one will be used to determine the relationship between health expenditure and life expectancy; the second one will be used to determine the relationship between ICOR and life expectancy; and the third one will be used to determine the relationship between GDP and life expectancy. Through these three models, one will have a decent understanding of these potential factors in life expectancy. At last, a larger dataset that includes the Gini index will be used to investigate the effects of income inequality. Simple linear regression will be employed to model the relationship between each economic predictor and life expectancy in Canada. Model fit will be assessed using R-squared, adjusted R-squared, and residual analysis. To ensure the models meet the assumptions of linear regression, diagnosis will be performed.

Through linear regression modeling with R, the following data can be collected: in addition, the linearity (multiple R squared), homoscedasticity (no cluster), normality of residuals (residual is normally distributed), and independence of residuals (Durbin Watson test for D-W and p-value) are checked with the aid of R plotting (as an example graph for the health expenditure case will be included in the appendix):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Predictor | Median of residuals | Minimum of residuals | Maximum of residuals | Multiple R squared | F statistic | p-value |
| Health expenditure |  | -0.26252 |  |  |  |  |
| ICOR (cleaned) | -0.05046 | -0.20828 |  | 0.9671 |  |  |
| GDP |  |  |  |  |  |  |

Table1: Data collected for the three predictors’ linear regression model with life expectancy

In the first part, we discuss health expenditure and life expectancy. The dataset for this part contains columns for country, year, life expectancy, health expenditure per capita, and population. We can start by cleaning the data and only keeping the needed columns, such as country, year, life expectancy, and health expenditure per capita. Proceed by manually selecting the data from 2000 to 2021 for Canada and plotting the scatter plot with life expectancy as a dependent variable and health expenditure per capita as an independent variable. 图表, 散点图

描述已自动生成

Graph 1: Linear relationship between life expectancy and health expenditure per capita in Canda from 2000 to 2016

We can see that this is a visually well-fit trendline, which shows the linear proportionality between the two variables. However, it is necessary to dig deeper through linear regression model.

The linear regression model can also be calculated:

Referring to Table 1, the simple linear regression model for health expenditure per capita in Canada between 2000 and 2016 demonstrated a strong linear relationship (R-squared = 0.9808), indicating a good fit. The small median and range of residuals suggest accurate predictions. The high F-statistic and low p-value confirm the statistical significance of the model, meaning that health expenditure is a significant predictor of life expectancy. Specifically, the model predicts an increase of  years (approximately 0.29 days) in life expectancy for every one unit ($1000) increase in healthcare expenditure per capita. While this model suggests a strong association, it's important to note that other factors beyond healthcare expenditure can also influence lifespan. We also need to validate this model by splitting the data. With the aid of R, the values of RMSE = 0.1422 and R-squared = 0.9953 are obtained. Since RMSE is a small value and the R-squared value is close to 1, the model is valid.

In the second part, another factor that could affect the lifespan will be discussed, which is the income composition of resources (ICOR). Similar to the previous analysis on health expenditure, we start by checking the plot of life expectancy as a dependent variable and ICOR as an independent variable.图表, 散点图

描述已自动生成

Graph 2: Linear relationship between life expectancy and ICOR in Canda from 2000 to 2015 with potential outliers

Through visually scanning the plot, there are some significant outliers in this graph which is located at year 2006, 2007 and 2008. We can proceed with R to check if these points are influential by performing DFFITS test; we can say a data point is influential if its absolute value is greater than the following:

.

Since the corresponding DFFIT values are 0.5354, 0.5122, and 0.9285, the data are influential; thus, we can get rid of them. Alternatively, if we can perform the linear regression model, performing a linear regression model on R, we are able to know that the multiple R squared is not closed to 1, the range of minimum and maximum residuals is large, the F statistics are small, and the p value is greater than 0.05. Therefore, remove the data for 2008 and redo the analysis. The scatter plot shows:  
图表, 散点图

描述已自动生成

Graph 3: Linear relationship between life expectancy and ICOR in Canda from 2000 to 2015 (influential points removed)

This trendline is now visually more fitting. To determine if it is more fitting, a linear regression model will be needed (refer to Table 1). Through this linear regression model, one can first determine that it is a well-fit model since the multiple R squared value is close to 1, but it is not as well-fit as the last model. The maximum residual further proves this point because it indicates that there is at least one observation that the model overpredicted life expectancy for more than half of a year, which is very likely to be 2005. The fact that the F statistic is high and the p value is very small indicates that the model is statistically significant, which means that the ICOR predictor is a responsible predictor. Moreover, the simple linear regression equation implies that the increase of one unit in ICOR will result in an increase of 58.294 years in life expectancy. Again, validation needs to be tested. By the same method, the RSME obtained is 0.3090, which is acceptable, and the R-squared value is 0.9923, which is very close to 1. The model is valid.

In the third part, the relationship between GDP and life expectancy will be investigated. Start by plot independent variable of GDP and dependent variable of life expectancy:

图表, 散点图

描述已自动生成

Graph 4: Linear relationship between life expectancy and GDP in Canda from 2000 to 2021

As the graph shows, although the plots are not as fitted as the health expenditure graph, they still demonstrate a relatively tight connection with the trendline. To check the details of the linear regression model, we can use the data collected by R (table 1). Therefore, we notice that this is a good model that demonstrates a strong linear relationship since the multiple R squared value is close to 1. The maximum residual still demonstrates that there is at least one point where the prediction is deviated for more than half of a year. The high F statistic and low p value again show the model is highly statistically significant. Finally, the model suggests that at the starting point of 64.40 years of life expectancy, for one unit increase in GDP, an increase of 0.0004 years (0.15 days) of lifespan will be expected. Again, model validation needs to be performed. The RMSE obtained is 0.3609, which is reasonable, and the R-squared value is 0.8670, which is descent. The model is validated.

At last, by plotting the data on life expectancy and Gini index into a scatter plot, one cannot obtain a well-fit linear regression model (refer to graph 5 in the appendix). Therefore, one can proceed to analyze through ANOVA and Tukey’s HSD test by categorizing low Gini index in range (30–32), medium Gini index in range (32–34), and high Gini index in range (34–36). By running ANOVA and Tukey's HSD test on R, one can find that the p-value for the ANOVA test is 0.00215, which is significantly less than 0.05. Thus, we can reject the null hypothesis and conclude that there is a statistically significant difference in mean life expectancy across the three Gini index categories in Canada. The Tukey’s HSD test finds that there is a significant difference between the medium and low inequality groups since the adjusted p-value is very small. The "medium" group has a higher mean life expectancy than the "low" group by 2.3178 years. To have a better understanding, a box plot can be used (refer to Graph 6). The box plot shows a noticeable difference in the median life expectancy between the "medium" and "low" Gini index categories. The "medium" group's median is higher, indicating a higher average life expectancy in those years with medium income inequality. In addition, the "High" group's box overlaps considerably with both the "Medium" and "Low" groups, suggesting there's no strong visual evidence of a difference in life expectancy for the "High" category compared to the others.

To conclude, we can dive into the project again and analyze each predictor. Our analysis revealed a strong positive relationship between healthcare expenditure per capita and life expectancy in Canada between 2000 and 2016. The model, with a high R-squared value of 0.9808, indicates that health expenditure accounts for a significant portion of the variation in life expectancy during this period. This suggests that increased investment in healthcare services has a substantial impact on Canadians' lifespan, supporting the notion that a healthy population is essential for a thriving society. Further analysis indicated a statistically significant positive association between ICOR and life expectancy in Canada. However, although the relationship is statistically significant, it is not as robust as the association with healthcare expenditure. It suggests that a greater share of national income dedicated to labor income (higher ICOR) tends to be associated with a higher life expectancy. This finding aligns with the concept that a more equitable income distribution could lead to improved health outcomes. Moreover, GDP and life expectancy are strongly correlated. This finding highlights the relationship between economic growth in macro-scope and health outcomes. It suggests that GDP is a significant indicator of economic prosperity, but it is not the sole indicator of life expectancy. In the income inequality case, the analysis of the Gini index indicates that there is a statistically significant difference in life expectancy between categories of income inequality. A low Gini index means the income is more evenly distributed; in other words, there is less inequality. The finding that the "medium" Gini group has a higher life expectancy than the "low" group highlights the complexity of the relationship between income inequality and life expectancy. It demonstrates that when income inequality reaches a threshold, it might be beneficial for health outcomes. As it is counterintuitive, this could be a consequence of the tax mechanism. However, in these cases, since life expectancy has many factors, one predicator does not define the trend of the health outcomes of Canadians. Through this analysis, we also discovered that health expenditure is the most important or strong influencer of Canadians’ lifespan, followed by ICOR, then GDP. Overall, the finding suggests that investment in healthcare largely and efficiently promotes life expectancy; policies that promote the nation’s economy also increase life expectancy; and measures to stabilize income inequality are vital for extending the lifespan of Canadians as well. Despite the analysis, there are limitations throughout the project, which include the limitation of available data (datasets are not big) and some weird but influential outliers, which result in the linear regression model for ICOR not being very well-fit. Neglect the potential error created by the limitations; the analysis is a rough bridge that connects the economy and health in Canada. This analysis highlights the importance of national investments in health expenditure, income patterns, the general economy, and income inequality. Based on this analysis, further research can explore how different policies or events interact with the economy and health of Canadians. To conclude, this analysis proposes that as people are going through the growth and shrink of the economy, it is important to figure out a method to leverage this great impact on the health outcomes created for Canadians.

References

Felice, E., Andreu, J. P., & D’Ippoliti, C. (2016). GDP and life expectancy in Italy and Spain over the long run: A time-series approach. Demographic Research, 35(1), 813–866. https://doi.org/10.4054/DemRes.2016.35.28

Gini Index - Canada. World Bank Open Data. (n.d.). https://data.worldbank.org/indicator/SI.POV.GINI?locations=CA

KumarRajarshi. (2018, February 10). Life expectancy (WHO). Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/discussion?sort=hotness&page=2

Lawanson, O. I., & Umar, D. I. (2021). The life expectancy–economic growth nexus in Nigeria: the role of poverty reduction. SN Business & Economics, 1(10), 127–127. https://doi.org/10.1007/s43546-021-00119-9

Life expectancy vs. GDP per capita. Our World in Data. (n.d.-a). https://ourworldindata.org/grapher/life-expectancy-un-vs-gdp-per-capita-wb?country=~CHE

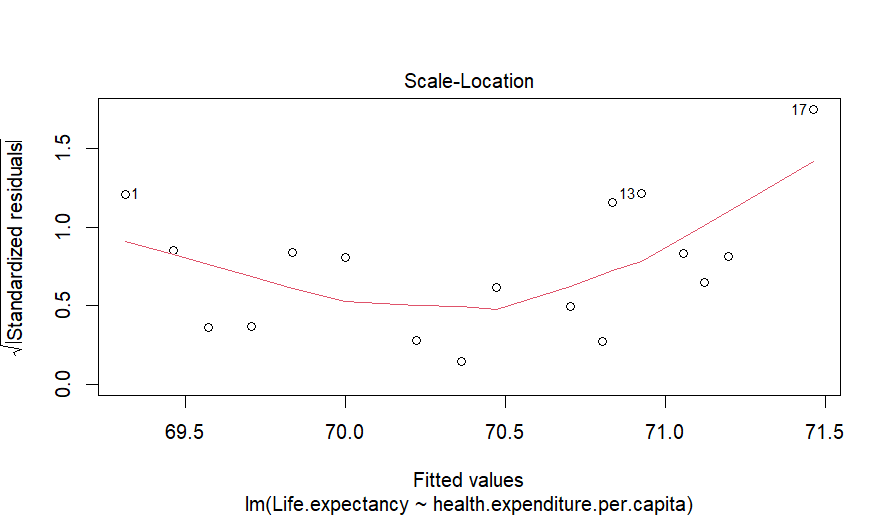
Life expectancy vs. healthcare expenditure. Our World in Data. (n.d.-b). https://ourworldindata.org/grapher/life-expectancy-vs-healthcare-expenditure

Appendix:

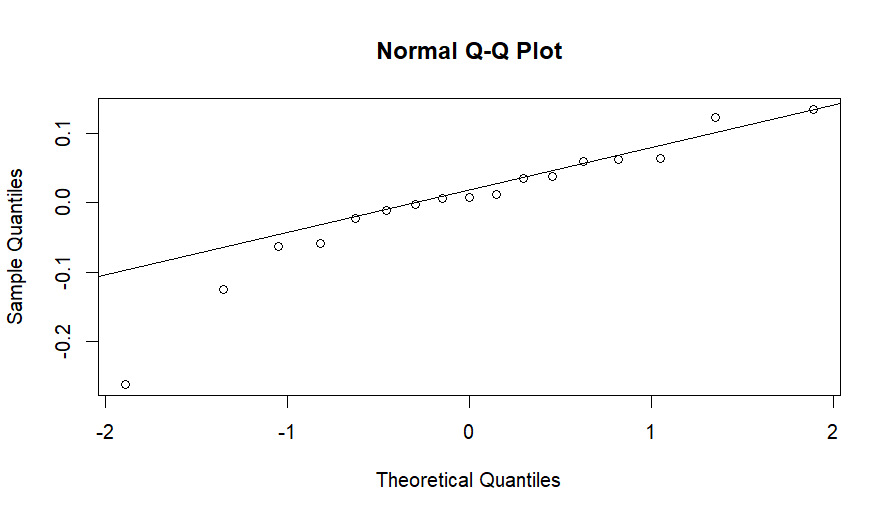
图表, 箱线图

描述已自动生成

Graph 6: This boxplot illustrates the distribution of life expectancy across three categories of income inequality (Low, Medium, High). The median life expectancy for the 'Medium' inequality group is notably higher than that of the 'Low' inequality group.



Graph 7: test for homoscedasticity in the health expenditure case



Graph 8: test for normality of residuals for the health expenditure case

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Predictor | Median of residuals | Minimum of residuals | Maximum of residuals | Multiple R squared | F statistic | p-value |
| Health expenditure |  | -0.26252 |  |  |  |  |
| ICOR (cleaned) | -0.05046 | -0.20828 |  | 0.9671 |  |  |
| GDP |  |  |  |  |  |  |

Table1: Data collected for the three predictors’ linear regression model with life expectancy