**Motivation**

Health is the prerequisite of survival and development for human. One important indicator to measure health is life expectancy, defined as the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life (World Bank, 2019). China, as a country with a population of 1.4 billion, had also set it an important goal to prolong citizens’ life expectancy in the 12th and 13th five-year plan. Specifically, the Chinese government aims at increasing citizens’ life to 79 years old in the outline of ‘Healthy China 2030’. In fact, almost every country tries to increase its citizens’ life expectancy and improve their life quality. Therefore, investigating and analyzing factors that affect life expectancy is meaningful and can help governments make decisions to better pursing developments and achieving goals. Meanwhile, it will spontaneously ensure and improve living standards. Considerable research has been done in this field for the reasons aforementioned. Linden and Ray (2017) suggested in their article that people live longer in countries with a higher GDP per capita. From the perspective of the medical area, average life could be increased by 1 year if the infant mortality rate decreases per 9‰ or 10‰ (Shen *et al*., 2015). Many researchers have investigated in this field, but few researchers have done multivariable analysis between life expectancy and macro factors. Therefore, in my study, I aim to find a relationship between life expectancy and demographic structure, economic development, medical treatment level, climate, and health conditions.

**Data Description**

After discussing and experimenting in Python, I finally chose representative types of annual data covering the period 1975 to 2014 (1200 in total) to explore the relationship between life expectancy and demographic structure, economic development, medical treatment level, climate and health conditions in 30 countries around the world. The final 6 regressors are presented below.

**Infant mortality rate (IMR) (‰)**: A measure of a nation’s medical level and health condition. Bhatt and Beck (2018) has proved that a significant decrease in IMR could be resulted from Medicaid expansion.

**Population ages 65 and above (%)**: A measure of a nations’ aging population. The larger aging population comes with a more unbalanced demographic structure. If the aging population is large, the whole country’s life expectancy will be large but their economy would grow extremely slow. Government needs to make a tradeoff between them.

**Female in percentage of total population (%)**: A measure used to reflect whether gender is balanced in the population. A larger female population comes with a more unbalanced gender structure. I suspected that larger female population will increase the country’s life expectancy.

**GDP per capita (constant 2010 US$)**: GDP per capita is GDP divided by midyear population (adjusted for inflation). It is an economic indicator usually used to compare the living standard between countries over time. Higher GDP per capita indicates a higher living standard, which will generally lead to longer life expectancy.

**C****arbon dioxide (CO2) emissions (kt)**: Data used to measure climate situation. If CO2 emissions are high, it may affect the stability of climate and might resulted in the happening of extreme weather, which would have bad effects on the living environment of citizens.

**Prevalence of underweight among adults, BMI<18 (crude estimate) (%)**: A rate which is a measure of health condition. If the rate is high, citizens are more likely to suffer from starving and their quality of standard of living is low, which may resulted in a shorter life expectancy.

Besides the data of prevalence of underweight among adults, which are collected from WHO (<https://www.who.int/en/>), others are all collected from World Bank (<https://data.worldbank.org/>).

List of all variables and data sources are presented below.

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Source |
|  | Infant Mortality Rate (per 1,000 live births) | World Development Indicators Database, World Bank https://databank.worldbank.org/reports.aspx?source=2&series=SP.DYN.IMRT.IN&country= |
|  | Population ages 65 and above (% of total population) | World Development Indicators Database, World Bank https://databank.worldbank.org/reports.aspx?source=2&series=SP.POP.65UP.TO&country= |
|  | Population, female (% of total population) | World Development Indicators Database, World Bank https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS |
|  | GDP per capita (2010 constant US$) | World Development Indicators Database, World Bank https://data.worldbank.org/indicator/NY.GDP.PCAP.KD%C2%A0 |
|  | CO2 emissions (kiloton) | World Development Indicators Database, World Bank https://databank.worldbank.org/reports.aspx?source=2&series=EN.ATM.CO2E.KT&country= |
|  | Prevalence of underweight among adults, BMI<18 (crude estimate) (%) | World Health Organization  http://apps.who.int/gho/data/view.main.NCDBMILT18Cv?lang=enation |
|  | Life Expectancy at Birth, Total (years) | World Development Indicators Database, World Bank https://databank.worldbank.org/reports.aspx?source=2&series=SP.DYN.LE00.IN&country= |

**Empirical model and hypotheses**

I selected data from 30 countries from 1975 to 2014 as samples to find the linear relationship between life expectancy and six independent variables: the proportion of the aging population, IMR, GDP per capita, prevalence of underweight, proportion of female population, and CO2 emissions. My hypotheses are: andhave positive correlations with life expectancy, while and have negative ones. Then I built up the multiple linear regression model:

Starting with the OLS method, I conducted pooled regression on 1200 data to test whether variables were suitable to explain the changes in life expectancy. Since that the order of magnitude of GDP is so large that the coefficient became relatively small, so I take the logarithm of GDP. In the first OLS regression, I observed that the impact of carbon emissions on life expectancy was small and its coefficient was the only insignificant one at 5% significance level. Furthermore, I computed VIF to test multicollinearity on the logarithm of GDP and 5 other variables, and found that there was no significant multicollinearity between every two variable and significant correlation between every two variable was all less than 78%.

In this case, I decided to keep origin explaining variables except for carbon emissions to carry out OLS regression again and pay attention to the heteroscedasticity of the model. By observing and testing the distribution of residuals, I affirm that the model had strong heteroscedasticity, which was caused by the situation that life expectancy of some samples was mainly affected by omitted variables. However, the variables I selected could not be treated as the main explaining variables for them.

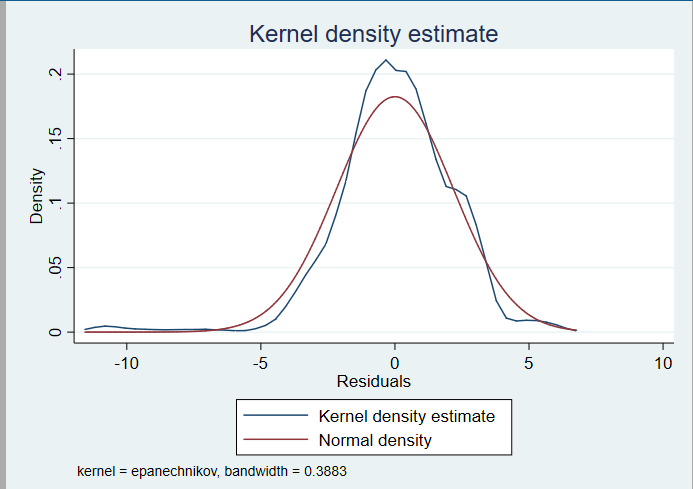


Figure 1 Residuals 1



Since these samples affect the accuracy of this model to a great extent, I need to find and dispose them. I used mathematical methods and successfully received both outliers and leverage influential points. Meanwhile, I found the characteristics of these points through Python: they were mainly concentrated in 5 countries: Brazil, Japan, India, South Africa, and Cameroon and occurred almost every year. To facilitate further study of the panel data, I need to ensure the integrity of data for each country and I directly deleted data of all years for the above 5 countries.

Next, I regressed the rest 1,000 data and found that the model was obviously improved through the test since heteroscedasticity was significantly reduced. Furthermore, the regression coefficients were more accurate. However, heteroscedasticity still existed, I therefore chose “large sample theory” regression method to obtain a more robust standard error.

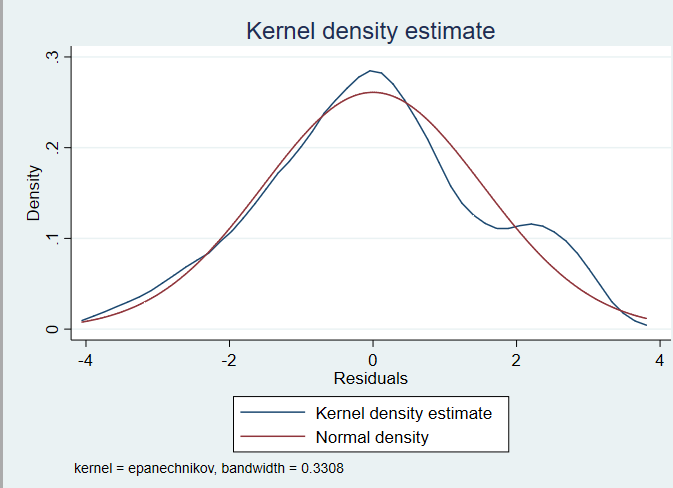


Figure2:residuals2



Although I removed the impact caused by some special countries, there still exist differences between the remaining 25 countries. In other words, even though the regression coefficients were unbiased and identical, the intercepts were respectively different. Moreover, due to the ignored difference, the pooled regression model was not accurate, and I needed to create new models. After Hausman test, I rejected the hypothesis of choosing the random effect model, and after the LSDV regression method, I could safely abandon the pooled regression model. I could conclude that I should use fixed effects regression model to capture all state-specific intercepts by assigning (n-1) binary variables to carry out the regression of life expectancy. Five variables include , , , and . Python results showed that was insignificant and not quite influential to life expectancy and . Therefore, I dropped .

Through the research by Driscoll and Kraay, I knew that cross-sectional correlations between data may exist. For example, the GDP of America may affect that of Cuba in the same year. Therefore, I further optimized the fixed effect by adding Driscoll-Kraay’s regression method to obtain a fixed effect model that considered cross-sectional correlations.

Moreover, I considered that life expectancy and explaining variables could be affected by the same time trend, which may lead to the misjudge that the linear relationship to be significant, which actually is not. By observing the relationship between the data of each variable and time, I roughly believed that there was no non-stationary time series for each variable. Therefore, I introduced the time trend of years into my fixed effect model as an explaining variable to eliminate the “spurious regression” as mentioned above. Besides, I dropped which was not significant anymore. With this, my model was completed.

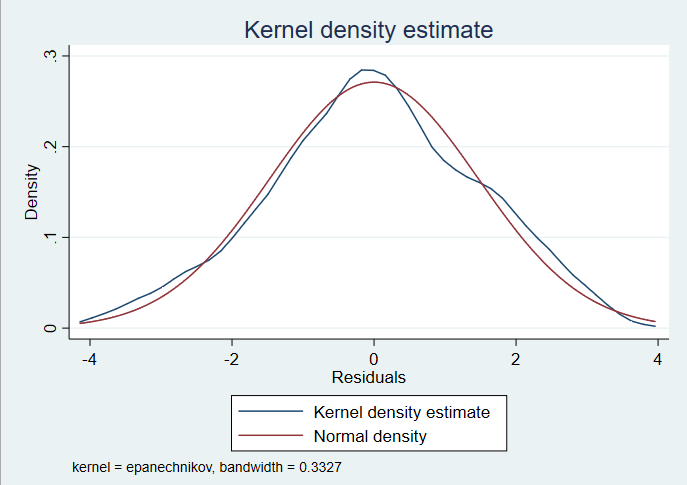


Figure3:residuals3

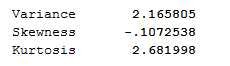




Table 1 Linear Regression Models for Life Expectancy

**Presentation of results and inference**

Therefore, the Fix effects regression line for life expectancy is

The unbiased estimate of the coefficient on () is -77.67429, which means one percent increase on will reduce life expectancy by 0.7767429 years, and a clearly negative impact can be concluded. The coefficient of the estimator (2) equals 26.12217. Therefore, it is believed to have a positive impact on life expectancy, and one unit change on will cause life expectancy to change 26.12217 years in the same direction. The third coefficient of the unbiased estimator () is 1.06739. With one unit positive change on , life expectancy is expected to increase 1.06739 years. The estimate of the coefficient on ) is 0.143618, indicating that life expectancy will increase 0.1436185 year every single year. The goodness of fit () is how much the dependent variable can be explained by regressors in the regression. of this regression model is 0.9472, which is close to 1, indicating relative sound goodness of fit. Root MSE is about 0.00372, indicating good accuracy of the model.

From the standardized coefficient method by Python attached in the appendix, the larger the corresponding absolute value of coefficient, the greater the influence of variable on life expectancy.

is believed to be the most influential one among three variables as I do not consider the time trend as an explaining variable. The result received is similar to that obtained by Hiram (2019), the increasing of the aging population increases national total life expectancy to a great extent. The second influential one is and the least powerful variable is . reflects a country’s health status and medical level to a great extent. Regarding the reason why is the powerful one, as Mizuno and Yakita (2013) mentioned, an increase in the proportion indicates a general postponement of retirement age and a rise in labor participation. Therefore, GDP is expected to increase and medical level development can be promoted. Simultaneously, will reduce.

Estimated life expectancy from 1975 to 1984 and 2005 to 2014 are derived to make comparisons. In each time period, the remaining 25 examined countries are divided into 9 middle-income countries and 16 high-income countries based on the World Bank’s classification standard (2019).

From Python results attached in the appendix, average life expectancy in middle-income countries increased by almost 10.756%, and that in high-income ones grew around 10.484% in 30 years. However, the gap between life expectancy has widened by nearly 7.404%.



Figure4



Figure5



Figure6

Three graphs regarding average , GDP and are attached above. Referring to them and what has been mentioned before, is the least influential variable, although its rate of change is relative fast, its actual influence is limited. Besides, between high-income and middle-income countries widens faster than , and is a more powerful determinant of life expectancy. The average life expectancy of middle-income countries during 2005-2014 finally catches up with that of high-income ones in 1975-1984.

To determine whether the three explaining variables and time trends have different impacts on different groups of countries, the continued process compared the most effective variable in high-income countries and middle-income countries during two time periods, 1975-1984 and 2005-2014. From Python results, it is clear that in middle-income countries, is the most powerful variable to affect life expectancy in 1975-1984, and the most influential one changes to in 2005-2014. For high-income countries, the variable is always . One reason for this difference might be the dissimilarity of medicine and economy development speed in high and middle-income countries. In 1975-1984, the medical level in middle-income countries may be low, therefore is not the most effective factor in this decade. Probably it has been improved in 20 years and reached a superior level in 2005-2014, which may have a large impact on life expectancy. In high-income countries, the medical level had already been at a relatively high level and is less likely to make notable improvements. Subsequently, has the biggest influence on life expectancy in high-income countries constantly.



Table 2: Linear Regression Model for Life Expectancy of High-income and Middle-income Countries

**Discussion**

Since life expectancy is an important indicator of health status in a country (World Bank, 2019), the regression model shows a clear relationship between the medical level, the aging of population, economic status and health status. Different from other researches, this regression focuses on not only one variable but four variables. It especially gives a new perspective on observing the effects that GDP has on life expectancy, and the point is not commonly done before. However, there are still some limitations. Firstly, only data from the rest 25 countries are used in the regression and all of them are either high-income or middle-income countries, ignoring low-income countries. Secondly, the model still has some ineradicable multicollinearity and heteroscedasticity. Also, when I were dealing with the panel data, I only established an entity fixed effect model without considering time fixed effect as I could not imagine that there were some omitted variables affecting life expectancy that only changed with time. Moreover, when I added the time trend as explaining variable to the model to estimate “spurious regression”, I did not give full consideration to whether each variable is stationary or non-stationary. There are still other influential factors excluded and need further research.

Based on the statistical results and inference, some advice is provided to the government to enhance life expectancy. First of all, improving the medical level can reduce the mortality rate in not only infant groups but also other ones. Increasing fiscal spending on medical facility and treatment, and setting up expenses reimbursement system on medical expenditure are both effective methods. Also, the government should release pro-natalist policy to encourage birth. Positive effects will be noticed on optimizing population structure, enlarging labor force and increasing GDP in the long term (Mizuno and Yakita, 2013).

**Appendix**

1. **Reference**

Bhatt, C. B. and Beck-Sagué, C. M. (2018) ‘Medicaid Expansion and Infant Mortality in the United States’, *American Journal of Public Health*, 108(4), pp. 565–567. doi: 10.2105/AJPH.2017.304218 [Accessed 7 Dec. 2019].

Data.worldbank.org. (2019). High income | Data. [online] Available at: <https://data.worldbank.org/income-level/high-income?view=chart> [Accessed 7 Dec. 2019].

Data.worldbank.org. (2019). Middle income | Data. [online] Available at: <https://data.worldbank.org/income-level/middle-income?view=chart> [Accessed 7 Dec. 2019].

Data.worldbank.org. (2019). Mortality rate, neonatal (per 1,000 live births) | Data. [online] Available at: <https://data.worldbank.org/indicator/SH.DYN.NMRT?view=chart> [Accessed 7 Dec. 2019].

Data.worldbank.org (2019). GDP per capita (constant 2010 US$)| Data. [online] Available at: <https://data.worldbank.org/indicator/NY.GDP.PCAP.KD> [Accessed 7 Dec. 2019].

Data.worldbank.org. (2019). Life expectancy at birth, total (years) | Data. [online] Available at: <https://data.worldbank.org/indicator/SP.DYN.LE00.IN?view=chart> [Accessed 7 Dec. 2019].

Hiram Beltrán-Sánchez and S.V. Subramanian (2019) ‘Period and cohort-specific trends in life expectancy at different ages: Analysis of survival in high-income countries’, SSM: Population Health. doi: 10.1016/j.ssmph.2019.100422. [Accessed 7 Dec. 2019].

Hoechle, D. *et al.* (2007). ‘Robust standard errors for panel regressions with cross-sectional dependence’, *The Python journal*, 7(3), pp. 281-312. [[https://www.Python-journal.com/sjpdf.html?articlenum=st0128](https://www.stata-journal.com/sjpdf.html?articlenum=st0128)](https://www.stata-journal.com/sjpdf.html?articlenum=st0128) [Accessed 7 Dec. 2019]

Linden, M. and Ray, D. (2017) ‘Aggregation bias-correcting approach to the health–income relationship: Life expectancy and GDP per capita in 148 countries, 1970–2010’, *Economic Modelling*, 61, pp. 126–136. doi: 10.1016/j.econmod.2016.12.001. [Accessed 7 Dec. 2019].

Mizuno, M. and Yakita, A. (2013) ‘Elderly labor supply and fertility decisions in aging-population economies’, *Economics Letters*, 121(3), pp. 395–399. doi: 10.1016/j.econlet.2013.09.022. [Accessed 7 Dec. 2019].

‘Outliers and influences’, (2019) Available at: <https://www.real-statistics.com/multiple-regression/outliers-and-influencers/> (Accessed: 7 Dec. 2019)