Proceedings of the 2023 International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/11/20230418

Life expectancy analysis based on multiple linear regression analysis

Runmeng Liu

Harbin No.3 High School, Harbin 150001, China

631402150103@mails.cqjtu.edu.cn

Abstract. One of the most significant subject matters in contemporary research is life expectancy. The researchers discovered that although the contributing elements of life expectancy have been extensively addressed, there has not yet been any direct connections between them. In order to looking into the relationship between life expectancy and 14 relevant influencing factors as well as the variations in these relationships, this research performed multiple linear regression analysis on the life expectancy data gathered from the website Kaggle. In accordance with the study, several aspects that individuals believed were important but did not in fact affect life expectancy could not be explained by overly specific variables. Researchers can use this finding based on the multiple linear regression method to investigate the similarities and differences between life expectancy and different influencing factors, for the purpose to come up with more effective recommendation to improve the life expectancy and health level of the population.

Keywords: life expectancy, multiple linear regression, OLS.

1. Introduction

The average life expectancy refers to the average age at which the newly born population is expected to survive under the existing economic and health level. As a person's life span is hard to trace and cannot properly reflect his or her health condition, life expectancy becomes a significant indicator of the state of society's health today, which also reflects the level of social development. Since the concept was proposed, life expectancy has attracted the attention of more countries and is one of the topics that scholars of various disciplines are keen to discuss. A more complete understanding of the relationship between life expectancy and its influencing factors could help governments specify better policies and individuals improve their own health.

In order to better analysis life expectancy using regression method, it is of significance to investigates the main factors which have certain impact on the public's health. Primary studies of life expectancy from different aspects have discussed the impacts of various factors might have on life expectancy, which are beneficial in determining the factor worthy to be analysis in this paper. As a comprehensive measure of social progress and population health, life expectancy is influenced by both macro and micro factors [1].

Macroscopically speaking, the average life expectancy has changed over the course of human history, and it has always been closely correlated with the rate of births, deaths, and population change [2]. Approximately two-thirds of all fatalities worldwide are brought on by chronic, non-communicable diseases [3]. The four risk behaviors of smoking, eating poorly, not exercising, and drinking excessively

DOI: 10.54254/2753-8818/11/20230418

are the main causes of these disorders [3]. These evidences indicate that mortality and communicable diseases might have certain effect on people's life expectancy [4]. Besides, social economy is one of the main factors influencing life expectancy, including the distribution pattern of government and personal expenditure, education level, dietary habits, and other factors [5]. What is more, the current state of economic growth and the standard of health service delivery in a country or region can be gauged by qualitative and quantitative analysis of numerous direct and indirect factors impacting life expectancy [2]. Speaking of economics and health service, the population's health can be influenced by investments in the health care system both directly (through primary and secondary medicine prevention; the prevention and treatment of diseases) and indirectly (through the improvement of the standard of living, in the form of an increase in the population's income; improved nutrition; living and housing conditions; improved conditions for work and leisure, etc.) [6]. Life expectancy is also related to the pattern of income distribution [7]. That is to say, expenditure on health, both by government and by individual, should be considered as a factor that might influence life expectancy. Furthermore, some habits and features of individual are also vital to one's health. Alcohol syndrome, severe obesity are important factors as well [8-10].

The field of life expectancy currently is comprehensive, with primary studies focus basically on the tendency of change, or the impact of a few factors on life expectancy. However, there is a relative vacancy of the comparison of the disparate correlations between life expectancy and the factors that can influence it. This paper aims to examine the relationships between various factors and life expectancy with data from different countries in different years, and thereupon compare them using different regression method. This work applies OLS regression to create analysis models for life expectancy. In order to effectively handle the collinearity problem. Moreover, due to the fact that life expectancy is closely related to one's gender, the indicators that vary by gender will be discussed separately in this paper.

2. Methods

2.1. Data source

The paper uses data from the website of Kaggle. The data includes the life expectancy of people in 159 countries from 2000 to 2015, the status of these countries (developing or developed), their adult mortality rate, infant death rate, under-five death rate, per capita alcohol consumption, expenditure on health, the coverage of several communicable diseases, and BMI of residents.

2.2. Variable description

Table 1 shows the full name, symbol of the variables used in the study. The range each variable is also calculated to indicate the feature of the data.

Table 1. Name, symbol and description of variables (traditional data).

Full Name	Symbol	Value Range
The average life expectancy of men	Em	[50, 85]
The average life expectancy of women	Ew	[50, 85]
Adult Mortality Rate(men)	AMm%	[50, 100]
Adult Mortality Rate(women)	AMw%	[50, 100]
The number of infant deaths	ID	[0, 1800]
Per capita alcohol consumption (in litres of pure alcohol)	Alco	[0.01, 17.87]
General government expenditure on health as a percentage of total government expenditure	TE%	[0, 17.24]

DOI: 10.54254/2753-8818/11/20230418

Table 1.(continued).

Hepatitis B vaccination coverage in men	HBm%	[18, 75]
Hepatitis B vaccination coverage in women	HBw%	[18, 75]
Number of reported cases of measles	Mea	[0, 212183]
Average Body Mass Index	BMI	[1.4, 87.3]
Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)	Diph	[2, 99]

2.3. Mathematical statistics method

2.3.1. Multiple linear regression model. There are p predictor variables $(x_1, x_2, \cdots, x_p)$ and a single response variable (y_i) and a linear regression can be used to explain the relationship between the response variables y_i and the predictor variables x_i.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (1)$$

Among them, ε_i is random error item, β_0, β_1, ⋯, β_p are regression coefficient, and this is called a multiple linear regression model.

2.3.2. Ordinary least squares (OLS). The multiple linear regression model serves as the foundation for the introduction of the residual sum of squares (RSS). The straight-line gap between the projected value and the actual value is less, and the fitting effect is better as the RSS decreases.

$$RSS = \sum_{i=1}^{n} (y_i - x_i^T b)^2 \quad (2)$$

Among them, y_i is actual value, and the b which minimizes the RSS through the least square method is the corresponding OLS coefficient estimate $\beta = (\beta_0, \beta_1, \cdots, \beta_p)$.

3. Results and discussion

3.1. Descriptive analysis

As shown in Figure 1, it can be found that the span of the life expectancy of men is not large, and its distribution is almost uniform, having a median value of 68 and a range from 50 to 85.

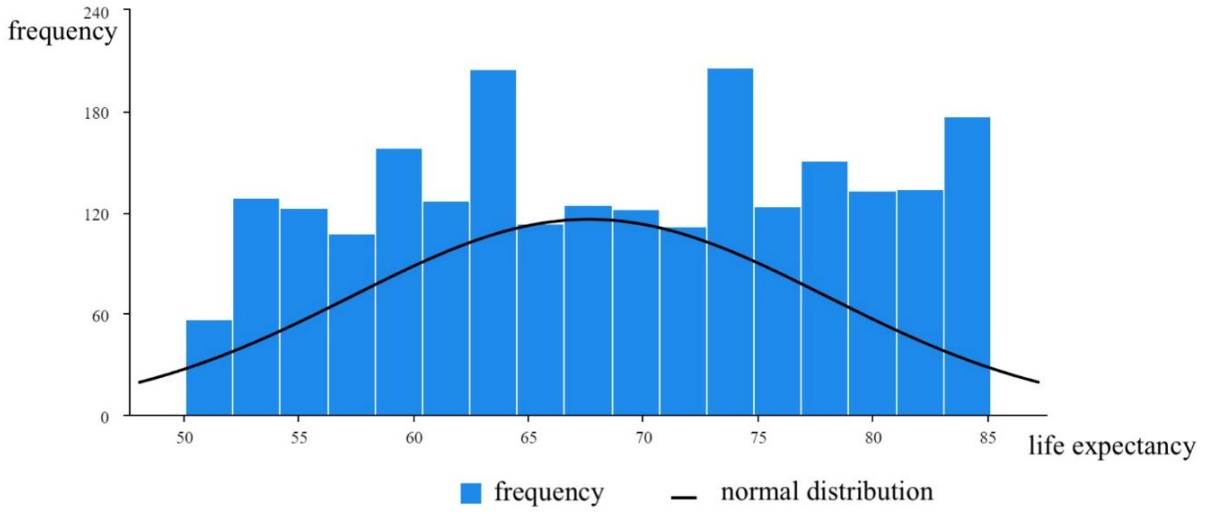

Figure 1. Histogram of life expectancy (men).

DOI: 10.54254/2753-8818/11/20230418

As shown in Figure 2, the distribution of average life expectancy of women is resemblance to that of the men, with a similar median value of 67 and a same range.

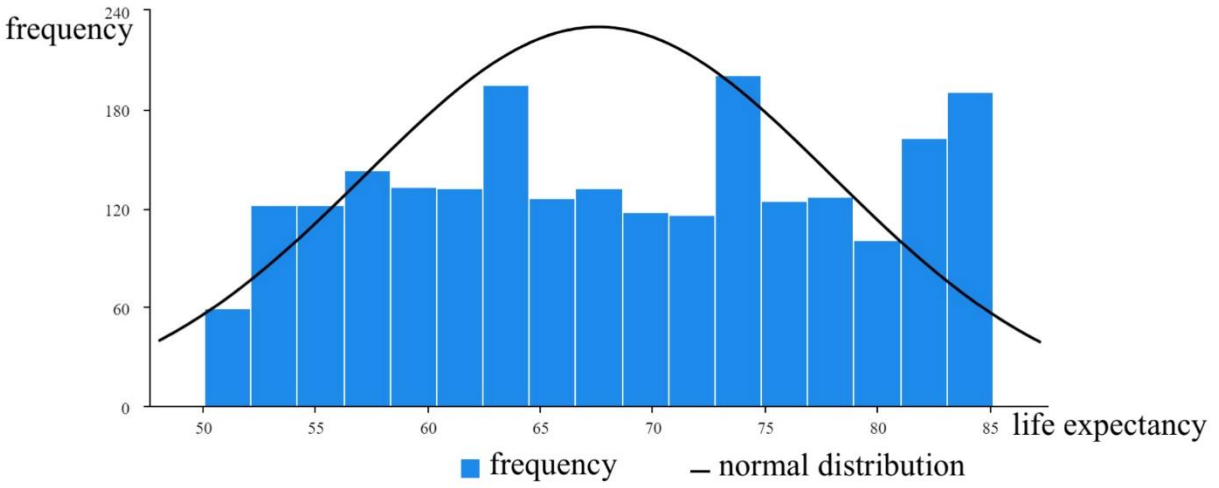

Figure 2. Histogram of life expectancy (women).

What is more, it can be seen from Figure 1 and Figure 2 that both groups of data about life expectancy conform to the normal distribution law, which indicates that they are effective for being used to analyse the relationship between life expectancy and its influencing factors.

As indicated in Figure 3, the range and IQR value between clusters are almost the same through both means of classification. It is also noticed that there is a small gap between the mean value of life expectancy of men and women in developed countries, while the two values is almost the same in developing countries.

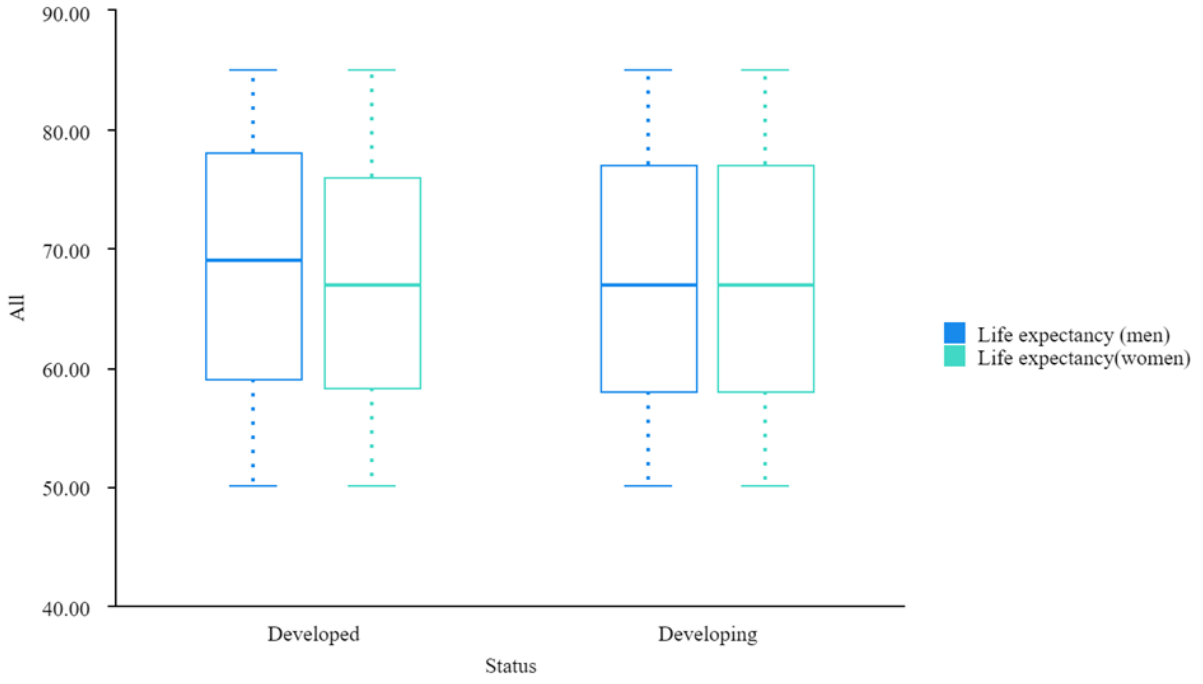

Figure 3. Box-plot of life expectancy in developed and developing countries.

DOI: 10.54254/2753-8818/11/20230418

Figure 4 display the correlation between 14 different factors and life expectancy, in which the shade of the colours shows a slight difference between the impacts of these factors on life expectancy of men and women.

Figure 4. Correlation coefficient diagram between 15 factors and life expectancy.

Figure 5 shows the change of life expectancy of countries around the world over fifteen years time. It is noticed that despite some fluctuations, the mean value of average life expectancy in 159 countries around the world is generally on the rise, with a slight increase from 2000 to 2015.

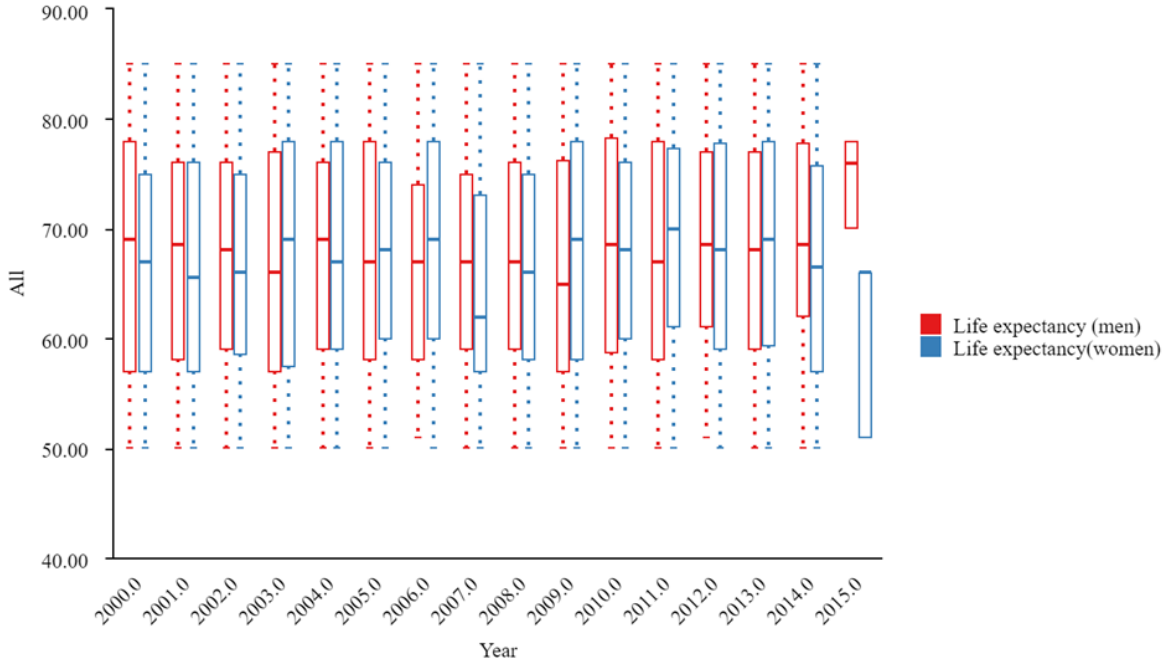

Figure 5. Box-plot of life expectancy over years from 2000 to 2015.

DOI: 10.54254/2753-8818/11/20230418

3.2. Inferential analysis

In this paper, the author tries to analyse the relationship between life expectancy and its influencing factors by multiple linear regression model. So as to make this process succinct and clear, the author chooses 8 specific factors which are representative, including adult mortality rate, infant death rate and other six ones shown below in Tables 2-3.

The OLS coefficient estimation was obtained through SPSSAU, as shown in the second column to fourth of the Tables 2-3. Next, the t value and p value are also calculated to determine whether the analysis items present significance. After that, a collinearity diagnostic is also conducted through gaining the VIF and tolerability values.

Table 2. Indicators from OLR for LEm.

Variables	Unstanderdized coefficient		Standerdized coefficient	t	p	collinearity diagnostics	
	B	Std. Error	Beta			VIF	tolerability
AMm%	-0.026	0.015	-0.037	-1.765	0.078	1.006	0.994
ID%	-0.003	0.002	-0.038	-1.533	0.126	1.408	0.710
TE%	-0.007	0.094	0.002	-0.076	0.940	1.120	0.893
HBm%	0.001	0.013	0.001	0.069	0.945	1.006	0.994
Mea%	-0.000	0.000	-0.011	-0.440	0.660	1.367	0.732
Diph%	-0.001	0.010	-0.003	-0.122	0.903	1.125	0.889
Alco%	0.022	0.059	0.008	0.364	0.716	1.228	0.814
BMI	0.014	0.012	0.028	1.196	0.232	1.250	0.800

Dependent Variable: Life expectancy (men)

* $p<0.05$ ** $p<0.01$

According to Table 2, which uses Life expectancy (men) as the dependent variable for a linear regression analysis and Adult Mortality (men), Infant Number, Total Expenditure, Hepatitis B (men), Measles, Diphtheria, Alcohol, and BMI as the independent variables, men's life expectancy is calculated with a R-square value of 0.005. This means that the causes of the 0.5% change in men's life expectancy may be explained by adult mortality (men), infant mortality, total expenditure, hepatitis B (men), measles, diphtheria, alcohol, and BMI. However, the model was found to have failed the F-test ($F = 1.413$, $p = 0.186 > 0.05$), which means that Adult Mortality (men), Infant death, Total expenditure, and Hepatitis B (men), Measles, Diphtheria, Alcohol, BMI have no influence relationship on expectancy(men).

According to Table 3, which uses Life expectancy (women) as the dependent variable for a linear regression analysis and Adult Mortality (women), Infant Number, Total Expenditure, Hepatitis B (women), Measles, Diphtheria, Alcohol, and BMI as the independent variables, men's life expectancy is established with a R-square value of 0.002. This means that the causes of the 0.2% change in men's life expectancy may be explained by adult mortality (men), infant mortality, total expenditure, hepatitis B (men), measles, diphtheria, alcohol, and BMI. However, the model was found to have failed the F-test ($F = 1.462$, $p = 0.883 > 0.05$), which means that Adult Mortality (women), Infant death, Total expenditure, and Hepatitis B (women), Measles, Diphtheria, Alcohol, BMI do not have any influence relationship on expectancy(women).

DOI: 10.54254/2753-8818/11/20230418

Table 3. Indicators from OLR for LEw.

Variables	Unstanderdized coefficient		Standerdized coefficient	t	p	collinearity diagnostics	
	B	Std. Error	Beta			VIF	tolerability
AMm%	-0.004	0.015	-0.005	-0.238	0.812	1.005	0.995
ID%	-0.002	0.002	0.025	1.205	0.305	1.410	0.709
TE%	-0.048	0.095	-0.001	-0.504	0.615	1.121	0.892
HBm%	-0.003	0.013	-0.005	-0.260	0.795	1.005	0.995
Mea%	0.000	0.000	0.007	0.302	0.763	1.367	0.731
Diph%	-0.006	0.010	-0.014	-0.628	0.530	1.125	0.889
Alco%	0.008	0.060	0.003	0.133	0.894	1.234	0.811
BMI	0.015	0.012	0.028	1.219	0.223	1.250	0.800

Dependent Variable: Life expectancy (women)
* $p<0.05$ ** $p<0.01$

4. Conclusion

The object of this study is to analyze the relationships between life expectancy and the 8 factors that might influence it and make a comparison between these relationships from a general perspective. Using data from the website Kaggle, a multiple linear regression analysis is generated along with a t-test, P-value and a collinearity diagnostic. Descriptive statistics and OLS regression were performed initially, indicating some basic features about the relationship between life expectancy and other variables. However, the p-tests and collinearity diagnostics reveal that the variables chosen don't have influence on life expectancy. This might be simply due to the fact that the variables chosen in this study are too specific that no direct relationship can be found like what is conducted in other studies in a certain field. The result of this paper might help researchers make a better choice when finding variables to analyse life expectancy, since a specific variable included in a certain genre might not be an influencing factor as well. Take, for example, the number of reported cases of measles is a part of the communicable diseases, but the value itself don't influence life expectancy.

References

[1] Sun Q, Lv J and Li L 2021 Development and application of life expectancy related indicators. *Chinaepi* 42(9), 1677-1682.
[2] Wei X 2020 Start with life expectancy. *Quality and Standardization* 06, 7-9.
[3] World Health Organization, Clobal Status Report on Noncommunicable Diseases 2010. Description of the Clobal Burden of NCDs, TheiRisk Factors and Determinants, WHO official website, 2010, https: //ww. who. int/nmh/publications/ncd_ report2010/en.
[4] Kennelly B, Shea E and Garvey E 2003 Socialcapital, life expectancy and mortality: a cross-national examination. Social Science & Medicine 56(12). 2367-2377.
[5] Wu Z 2013 Study on the influence of social and economic factors on life expectancy in China-Quantitative analysis based on data from the fifth and sixth national censuses. *MA thesis. SASS* 1-70.
[6] Kolosnitsena M T, Kossova T B and Sheruentsova M A 2022 Factors influencing life expectancy growth: a cluster analysis of world countries. *Social Science* 10, 212-225.

DOI: 10.54254/2753-8818/11/20230418

[7] Xi J, Lin X, Liang B, Gu Q and Hao Y 2023 Study on macro factors of life expectancy of elderly population in Guangzhou. *Chinese Journal of Health Information Management* 02, 191-196
[8] Nguyen X T and Jonsson E 2016 Life Expectancy of People with Alchohol Syndrome. *Journal of Population Therapeutics and Clinical Pharmacology* 23(1), 53-59.
[9] Walls H, Backholer K, Proietto J and McNeil J 2012 Obesity and Trends in Life Expectancy. *Journal of Obesity*.
[10] Peeters A, Barendregt J and Willekens F 2003 Obesity in Adulthood and Its Consequences for Life Expectancy: A LifeTable Analysis. *Annals of internal medicine* 138(1), 24-32.