Analyzing key health factors influencing life expectancy using machine learning

Jiahua Sun

University of Newcastle, Callaghan, Australia

c3497122@uon.edu.au

Abstract. With the rapid development of global medical technology and the steady improvement of social living standards, life expectancy per capita has become one of the most important indicators of a country or region's level of development. It not only reflects the quality of life and health of the local people, but also profoundly affects the economic growth and social stability of the country. In underdeveloped countries, life expectancy at birth is generally lower than in more developed countries due to resource constraints and poor medical facilities. This phenomenon has triggered extensive research interest in whether there is some correlation between life expectancy per capita and economic growth, and how improving life expectancy can contribute to a country's economic transformation and development. The purpose of this paper is to explore the impact of life expectancy per capita on economic growth and how to promote the economic transformation and development of countries by increasing life expectancy. Through a comparative study of Asian countries such as China, Singapore and Japan, this paper analyzes the impact of different factors on life expectancy through technical methods such as visual analysis, hypothesis testing, regression modeling, and decision tree modeling, and explores how these factors play a role in economic growth. The results of this study may have some guiding significance for improving life expectancy.

Keywords: life expectancy, hypothesis testing, regression, decision tree.

1. Introduction

This study mainly focuses on the relationship between Life Expectancy and five factors and how much they can influence Life Expectancy. The five factors are one-year-old immunization rates for hepatitis B, diphtheria and polio, alcohol consumption, and the number of measles cases. After exploring the relationship between these five factors and Life Expectancy, some analyses and hypotheses based on scientific knowledge will be put forward. As mentioned in the New England Journal of Medicine, effective analysis of Life Expectancy can provide important information about the population’s health status, which can be used in planning and decision-making [6]. At the same time, it helps identify high-risk groups so that preventive care can be targeted, which is very important to People’s lives are safe.

Through the analysis of the relationship between one-year-old immunization rates for hepatitis B, diphtheria, and polio, alcohol consumption, and the number of measles cases, the country and individuals can take preventive measures in advance after fully understanding the possible relationship between these diseases and Life Expectancy. Cor. test correlation test, regression statistics, residual analysis, and outlier detection were used in this study, and the following conclusions were drawn:

The five factors explored in this study are related to vaccines, alcohol consumption, and measles cases. Sweden has assessed the impact of smoking and alcohol abuse at different levels of education on changes in life expectancy and mortality and concluded that alcohol can affect a person's life expectancy, particularly for men [7]. Therefore, this study will further analyze the association between alcohol and life expectancy, and the extent of the association, to provide better evidence to support alcohol-related policies and actions. Vaccination of one-year-old children is important for the health and safety of children as they grow into adults. Life expectancy and one-year immunization rates for hepatitis B, diphtheria, and polio were examined to test such assumptions. Who and UNICEF launched the Expanded Programme on Immunization (EPI) in 1976 to control six childhood diseases: tuberculosis, diphtheria, whooping cough, tetanus, polio, and measles. Children's vaccination and disease prevention and control are paying attention to health issues all over the world. Studies show that parents' education level and sports level are significantly correlated with children's immunization, and improving parents' education level can improve children's immunization coverage[8]. Based on this research, once the link between life expectancy and vaccination of one-year-old children is confirmed, more research and initiatives will be needed to analyze the reasons behind vaccination rates in order to increase vaccination rates. Measles is one of the most common respiratory infections in children, killing about 2 million children each year, and there is no specific treatment [9]. Therefore, this study intends to explore the relationship between the number of measles cases and life expectancy and observe whether measles has a subtle influence on the life expectancy of the whole population. This is an entirely new hypothesis, and there is not much literature on the relationship between measles and life expectancy[10]. But that is why I want to conduct research to confirm the link between measles and life expectancy in order to raise awareness of measles in society[11].

1) There is almost no relationship between alcohol consumption and life expectancy.

2) There is little relationship between hepatitis B vaccine coverage under one year of age and life expectancy.

3) There was a weak positive correlation between coverage of diphtheria vaccines under one year of age and life expectancy rates.

4) There was a positive correlation between the one-year immunization rate for polio and the life expectancy rate.

5) There was a negative association between the number of measles cases and life expectancy.

2. Literature Review

Reference 1:Alcohol consumption as a predictor of mortality and life expectancy: Evidence from older Chinese males. The literature shows the relationship between alcohol and life expectancy, and the more people drink, the lower their life expectancy[1].

Reference 2: Hepatitis B Virus: Advances in Prevention, Diagnosis, and Therapy. The article focuses on the fact that about 250 million people worldwide are still infected with the viruses that cause hepatitis B. Hepatitis C virus (HCV) and Hepatitis B virus (HBV) are the leading causes of liver cancer and overall mortality globally, surpassing malaria and tuberculosis, and thus the disease is considered to have a significant relationship with life expectancy[2].

Reference 3: The Effect of Measles on Health-Related Quality of Life: A Patient-Based Survey. The results of this article suggest that measles is a highly contagious statutorily-reported disease that may be severe in infants, pregnant women, and immunocompromised individuals, and the conclusions suggest that the short-term impact of measles infection on HRQoL is substantial, both at the level of the individual patient and in terms of the overall burden of disease[3].

Reference 4: Factors Associated with Reduced Quality of Life in Polio Survivors in Korea. The results of this article suggest that people with polio have greater problems with mobility, daily activities, and depression/anxiety. Polio survivors, especially those with more symptoms of pain and fatigue, and those without access to medical care, have a poorer health-related quality of life, which indirectly has an impact on life expectancy[4].

Reference 5: Seroprevalence of antibodies to pertussis and diphtheria among healthy adults in China. The article's findings show that approximately 5% of adults aged 18-50 years were positive for anti-PT IgG antibodies, suggesting that adult pertussis is not uncommon in China. Although a large percentage of the study subjects had protective levels of immunity to diphtheria, antibody levels declined as adults aged, resulting in compromised longevity.[5]

3. Data Source

Data for the Final Research was downloaded from Kaggle. This is the GLOBAL Health Observatory(GHO) data repository under the WTO that tracks the health status of all countries and many other relevant factors for health data analysis purposes, making datasets available to the public. I selected relevant data of five diseases (Hepatitis B, Measles, BMI, Polio, HIV/AIDS) from them, and tried to find evidence of whether these diseases affect Life Expectancy to some extent.

These figures are very detailed and accurate because they come from official sources. But some have questioned the accuracy of the data, such as the child mortality column, which is supposed to calculate deaths per 1,000 people, but some numbers are higher than 1,000. This may be an error in the input of the data. However, I believe that due to the large amount of data, these minor errors will not affect the accuracy of the overall analysis. This research will use data on one-year-old immunization rates for hepatitis B, diphtheria, and polio, alcohol consumption, and the number of measles cases, and look for evidence that these vaccinations, diseases, and alcohol somehow affected life expectancy.

Last updated: 2018-02-10

Date created: 2018-02-10

The data description that I will study is:

Hepatitis B (HepB): Hepatitis B immunization coverage of one-year-old children (%).

Measles: Number of reported cases per 1000 population.

Polio: Immunization coverage of one-year-old children with poliomyelitis (Pol3) (%).

Diphtheria: Immunization rate of one-year-old children against diphtheria, tetanus, toxoid, and pertussis (%).

Alcohol: Alcohol, recorded per capita (15+) consumption (liters of pure alcohol).

To analyze the impact of these factors on life expectancy. The study considers life expectancy as the dependent variable to construct different machine learning models, e.g., regression models, decision tree models, etc. Mining potential information between variables through machine learning methods.

4. Methods

4.1. EDA

Detailed interpretation and visual presentation of alcohol data:

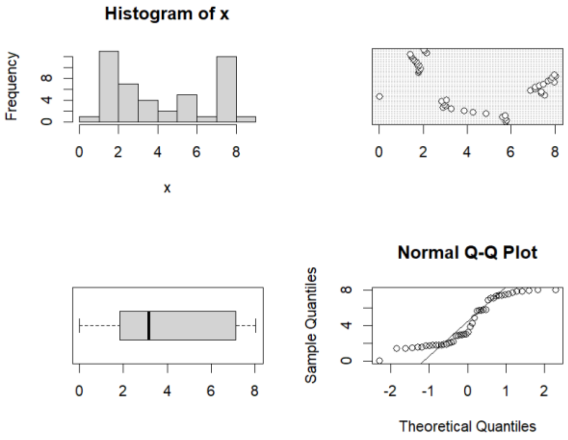


Figure 1. The chart shows the distribution characteristics of variable x through four different graphs

The histogram shows the distribution of alcohol content. The horizontal axis is divided into intervals (0-2, 2-4, 4-6, 6-8) and the vertical axis represents the frequency or frequencies of alcohol levels within each interval. Since the data are concentrated between 0-10, it can be expected that most of the bars will lie within these intervals, especially in the lower intervals (e.g., 0-4).

Scatter plots are used to show the exact location of each data point. In this case, the data points may be concentrated in areas of lower alcohol content (e.g., 0-8), whereas when the alcohol content is greater than 8, the data points may be more widely dispersed, indicating greater variability in these values.

The box plot shows the distribution of the data through quartiles (Q1, Q2, Q3) and outliers. Its top and bottom represent the third quartile (Q3) and first quartile (Q1) of the alcohol variable, while the line in the center (median) will lie inside the rectangle.

The QQ plot is used to determine if a normal distribution is met, and based on the QQ plot results, it is clear that a normal distribution is not met

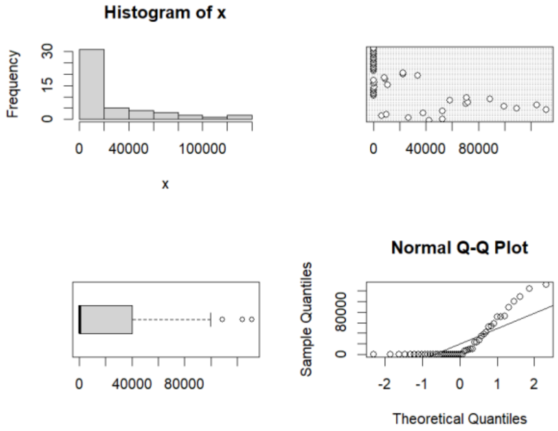


Figure 2. The chart shows the distribution characteristics of variable x through four different graphs

In exploring information on measles variables, histograms, dot plots, box-and-line plots, and QQ plots were conducted to visualize and analyze the characteristics of the data distribution of the number of reported measles cases per 1,000 population. The histogram showed that the measles factor did not satisfy a normal distribution. Dot plots reveal differences in the number of cases across districts. The box plot reveals outliers in the data, suggesting that the number of measles cases in some districts deviates from the norm. In addition, the QQ plot shows that the number of measles cases does not follow a normal distribution, which may be due to natural fluctuations in the number of measles cases or other external factors.

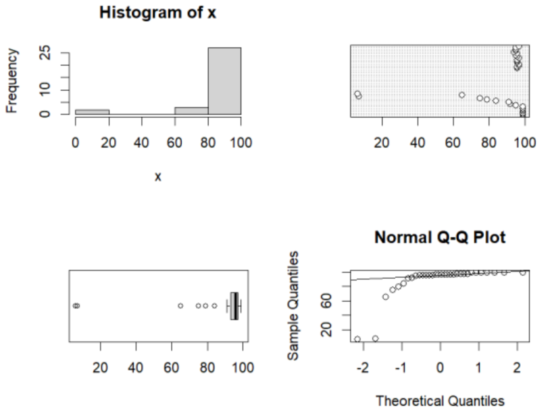


Figure 3. The chart shows the distribution characteristics of variable x through four different graphs

The histogram shows the distribution of hepatitis B immunization coverage among 1-year-old children. The horizontal axis is partitioned into intervals, while the vertical axis indicates the frequency or frequencies within each interval. As can be seen from the histogram, most of the children's hepatitis B immunization coverage is concentrated in the high level interval of 80 to 100.

The dot plot further demonstrates the specific values of hepatitis B immunization coverage for each child.

The box plot provides more information about the distribution of hepatitis B immunization coverage. The upper quartile (Q3) of the box lies around 80%, the median (Q2) may lie higher, and the lower quartile (Q1) may be lower. The results indicate the presence of outliers for hepatitis B.

The QQ plot results indicate that hepatitis B does not conform to a normal distribution.

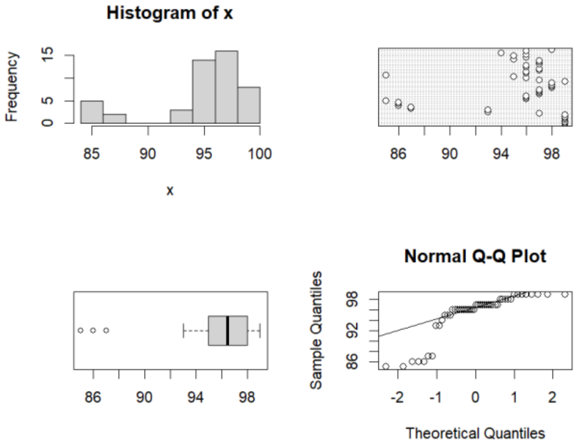


Figure 4. The chart shows the distribution characteristics of variable x through four different graphs

Exploratory analysis of the immunization rates for diphtheria, tetanus, toxoid, and pertussis among one-year-old children, and plot histograms, dot plots, box plots, and QQ for the immunization rates of one-year-old children with diphtheria, tetanus, toxoid, and pertussis. picture. Among them, the horizontal axis labels of the histogram are 85, 90, 95, and 100 respectively. The labels on the vertical axis are 0,5 and 15. The figure shows that the immunization rates of one-year-old children with diphtheria, tetanus, toxoid, and whooping cough are concentrated in the range of 0-10 and 60-100. The immunization rate of one-year-old children with diphtheria, tetanus, toxoid, and whooping cough is relatively divergent from the dot diagram. It can be seen from the box diagram that there are many abnormal values in the immunization rate of one-year-old children with diphtheria, tetanus, toxoid, and whooping cough. The labels of the box diagram are 86, 90,94, and 98, respectively. Through the QQ map, it was found that the immunization rate information of one-year-old children with diphtheria, tetanus, toxoid, and whooping cough did not conform to the normal distribution.

4.2. variable correlation

Using a scatter plot is to use two sets of data to form multiple coordinate points, to examine the distribution of coordinate points, to judge whether there is a certain correlation between two variables, or to summarize the distribution pattern of coordinate points. A scatter chart displays the series as a set of points. The value is represented by the position of the point in the chart. The categories are represented by different marks in the chart. Scatter charts are often used to compare aggregated data across categories.

The relationship between alcohol variables and life expectancy was plotted by scatter plot.Although alcohol consumption is known to be associated with life and health and is associated with many chronic diseases, the data from this study confirm that the relationship between alcohol consumption and life expectancy is very weak. This could be due to imprecise data and the fact that drinking may have different effects on different populations.

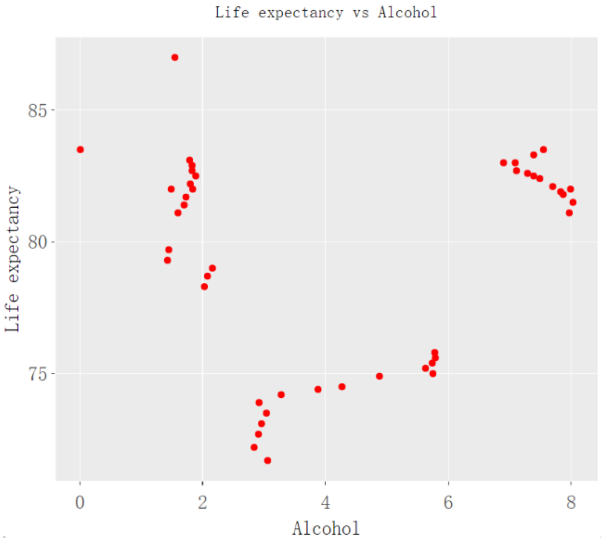


Figure 5. Show the relationship between Life Expectancy and Alcohol. The horizontal axis represents consumption per person per year (unit liters) and the vertical axis represents life expectancy (unit years). Each red dot represents a data sample

The scatter diagram is drawn with the alcohol variable as the horizontal axis and the life expectancy variable as the vertical axis. It can be found from the chart that the two variables have no correlation or the correlation is relatively weak. The scatter points are concentrated between 0 and 5. When the alcohol variable is greater than 6, the scatter diagram is very divergent.

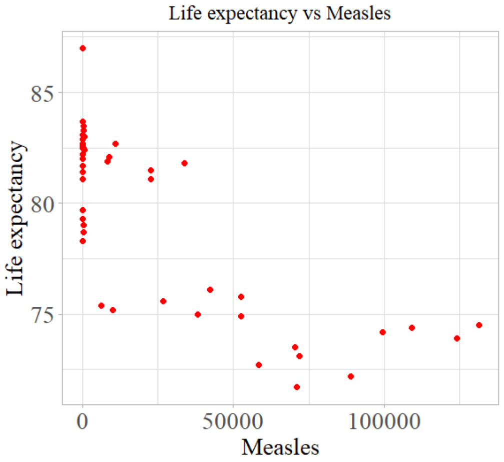


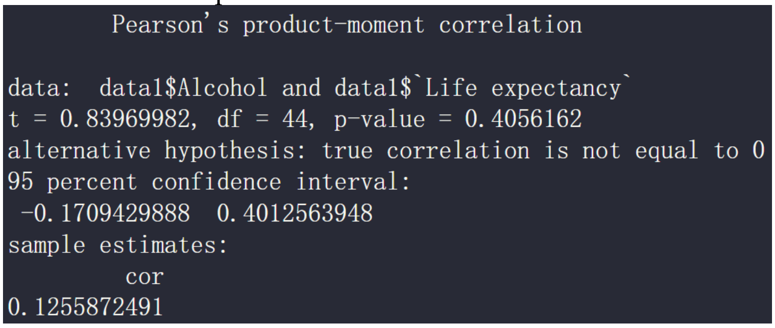
Figure 6. This scatter plot illustrates the relationship between life expectancy and the number of measles cases.

For the decentralized relationship between the number of cases reported per 1000 people and life expectancy, the horizontal axis is the number of cases reported per 1000 people, the vertical axis is the life expectancy, and the data are concentrated around 0. When the number of horizontal axes is greater than 50000, there are almost no scattered points, and the labels of horizontal axes are 0,50000,100000 respectively. When it can be found that there is little relationship between the two variables, Life expectancy is concentrated in places with a small number of cases.

4.3. Variable correlation hypothesis test

The analysis results show that the sample correlation coefficient is 0.1255872491. This suggests a positive correlation between alcohol and life expectancy. The correlation coefficient is between -1 and

1, but not close to 1, indicating that the correlation between alcohol and life expectancy is not perfect, that is, not completely consistent. In the following analysis, we will find that this conclusion is consistent with the predicted results of the scatter plot.



```
        Pearson’s product-moment correlation

data:  data1$Alcohol and data1$`Life expectancy`
t = 0.83969982, df = 44, p-value = 0.4056162
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1709429888  0.4012563948
sample estimates:
         cor
0.1255872491
```

Figure 7. This chart presents the results of Pearson's product-moment correlation, evaluating the relationship between alcohol consumption and life expectancy.

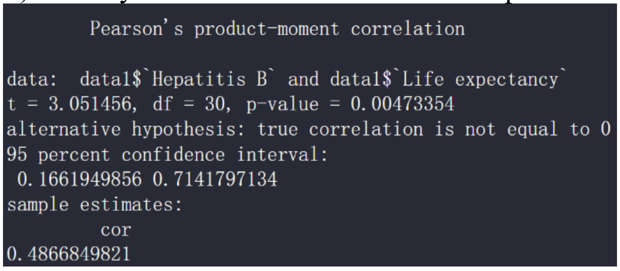
The analysis results show that the sample correlation coefficient is about 0.48668. This suggests that hepatitis B immunization coverage (%) in one-year-old children is positively correlated with life expectancy. 0.48668 is between -1 and 1, but not close to 1, indicating that the correlation between hepatitis B immunization coverage (%) and life expectancy in one-year-old children is not very perfect, that is, not completely consistent or the correlation is low. In the following analysis, we will find that this conclusion is consistent with the predicted results of the scatter plot.

After analysis, the T-Score is 3.05, which is a relatively large number. If the null hypothesis holds, then the deviation between the sample correlation coefficient and the population correlation coefficient is 3.05. Because in the case of a large T-value, the difference will be larger than expected. Therefore, if repeated sampling, the correlation coefficient is reduced by more than 95% (0-1.96SD, 0 + 1.96sd). Therefore, life expectancy may be related to hepatitis B immunization coverage (%) among one-year-old children.

The results showed that the P value was less than 0.05. Because a small p-value means that the probability that our sample gets from the hypothetical population is much less than 5%, the correlation coefficient may not be zero.

The results show that the confidence interval is (0.166,0.714). So the probability that the correlation is in this range is greater than or equal to 95% and, since 0 is not in this range, the null hypothesis is almost certainly not true. Therefore, hepatitis B immunization coverage among 1-year-old children (%)

It is positively correlated with life expectancy and has a linear relationship. In other words, hepatitis B immunization coverage (%) in one-year-old children can have an impact on life expectancy.



```
        Pearson’s product-moment correlation

data:  data1$`Hepatitis B` and data1$`Life expectancy`
t = 3.051456, df = 30, p-value = 0.00473354
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1661949856 0.7141797134
sample estimates:
         cor
0.4866849821
```

Figure 8. This chart presents the results of Pearson's product-moment correlation, evaluating the relationship between hepatitis B immunization rates and life expectancy.

The correlation between these factors is analyzed through the correlation coefficient diagram. The closer the correlation coefficient is to 1, the stronger the positive correlation is; the closer the correlation coefficient is to negative 1, the stronger the negative correlation is between variables.

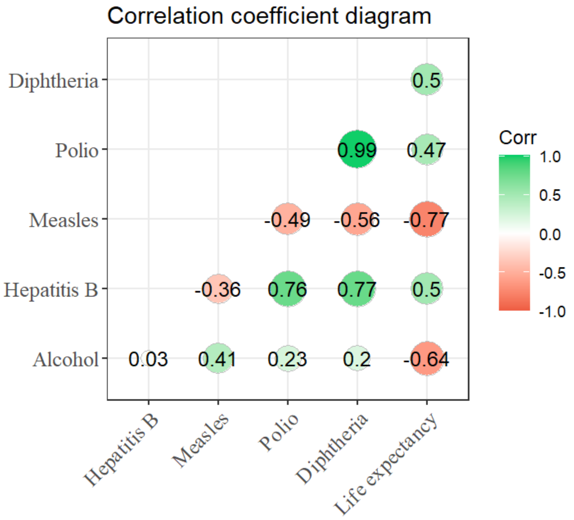


Figure 9. This correlation coefficient diagram illustrates the relationships between immunization rates for hepatitis B, measles, polio, and diphtheria, as well as alcohol consumption and life expectancy.

4.4. Regression models for different countries

Table 1. Model result summary. Taking China as the research object, a regression model was constructed, and the obtained model results are shown in the table

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	62.939	5.919	10.633	0.0000	***
Alcohol	0.296	0.424	0.699	0.5022	
`Hepatitis B`	0.013	0.006	2.009	0.0755	.
Measles	0.000	0.000	0.629	0.5447	
Polio	-0.246	0.260	-0.944	0.3700	
Diphtheria	0.340	0.297	1.144	0.2822	

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*

Residual standard error: 0.4392 on 9 degrees of freedom

Multiple R-squared: 0.9227, Adjusted R-squared: 0.8797

F-statistic: 21.48 on 9 and 5 DF, p-value: 0.0001

We fitted a linear model (estimated using OLS) to predict Life expectancy with Alcohol (formula: `Life expectancy` ~ Alcohol + `Hepatitis B` + Measles + Polio + Diphtheria).

From the p-value of the model, it can be found that it is much less than 0.05, indicating that the model as a whole satisfies significance. The model variance is shown below.

`Life expectancy`=62.94+0.3Alcohol+0.01`HepatitisB`+0Measles - 0.25Polio + 0.34Diphtheria.

The model explains a statistically significant and substantial proportion of variance (R2 = 0.92, F(5, 9) = 21.48, p < .001, adj. R2 = 0.88). The model's intercept, corresponding to Alcohol = 0, is at 62.94 (95% CI [49.55, 76.33], t(9) = 10.63, p < .001). Within this model:

The effect of Alcohol is statistically non-significant and positive (beta = 0.30, 95% CI [-0.66, 1.25], t(9) = 0.70, p = 0.502; Std. beta = 0.30, 95% CI [-0.66, 1.26]).

The effect of Hepatitis B is statistically non-significant and positive (beta = 0.01, 95% CI [-1.63e-03, 0.03], t(9) = 2.01, p = 0.075; Std. beta = 0.32, 95% CI [-0.04, 0.68]).

The effect of Measles is statistically non-significant and positive (beta = 3.95e-06, 95% CI [-1.03e-05, 1.82e-05], t(9) = 0.63, p = 0.545; Std. beta = 0.12, 95% CI [-0.31, 0.55]).

The effect of Polio is statistically non-significant and negative (beta = -0.25, 95% CI [-0.83, 0.34], t(9) = -0.94, p = 0.370; Std. beta = -1.17, 95% CI [-3.96, 1.63]). The effect of Diphtheria is statistically non-significant and positive (beta = 0.34, 95% CI [-0.33, 1.01], t(9) = 1.14, p = 0.282; Std. beta = 1.63, 95% CI [-1.59, 4.85]).

Based on the regression model results, a variable importance histogram was obtained using the vip package of R language. The importance of variables can be intuitively compared from the bar chart. For China, the hepatitis B variable is of the highest importance.

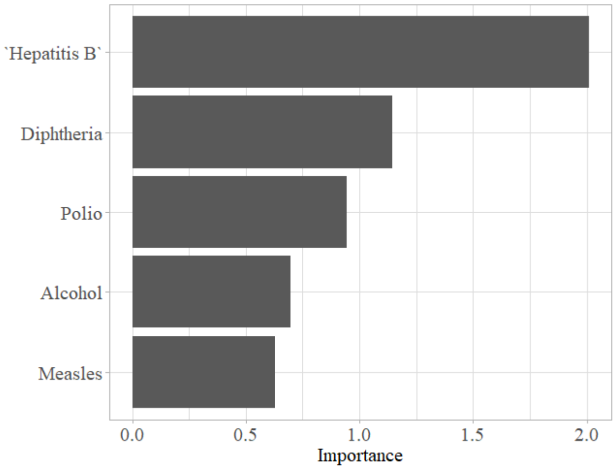


Figure 10. This chart displays the importance of different variables in influencing life expectancy.

The R function car package vif() can be used to detect multicollinearity in the regression model, and the test results are shown in the following table. It can be found that there is multicollinearity.

Table 2. Collinearity test results table. This table presents the Variance Inflation Factor (VIF) values for each variable, used to assess multicollinearity issues

variable	vif
Alcohol	21.00547
`Hepatitis B`	2.948548
Measles	4.23248
Polio	178.0845
Diphtheria	235.7375

Taking Japan as the research object, a regression model was constructed, and the obtained model results are shown in the table below.

Table 3. Model result summary. This table presents the summary of linear regression model results,

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	72.946	7.105	10.267	0.0000	***
Alcohol	-0.126	0.068	-1.864	0.0920	.
Measles	-0.000	0.000	-1.024	0.3301	
Polio	0.042	0.038	1.099	0.2974	
Diphtheria	0.068	0.049	1.402	0.1911	

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*

Residual standard error: 0.4773 on 10 degrees of freedom

Multiple R-squared: 0.692, Adjusted R-squared: 0.5688

F-statistic: 5.617 on 10 and 4 DF, p-value: 0.0124

We fitted a linear model (estimated using OLS) to predict Life expectancy with Alcohol (formula: `Life expectancy` ~ Alcohol + Measles + Polio + Diphtheria).

From the p-value of the model, it can be found that the p-value is 0.0124, which is less than 0.04, indicating that the overall significance of the model is satisfied. The obtained model equation is as follows.

`Life expectancy`=72.95-0.13Alcohol+0Measles+0.04Polio+ 0.07Diphtheria

The model explains a statistically significant and substantial proportion of variance (R2 = 0.69, F(4, 10) = 5.62, p = 0.012, adj. R2 = 0.57). The model's intercept, corresponding to Alcohol = 0, is at 72.95 (95% CI [57.12, 88.78], t(10) = 10.27, p < .001).

Within this model:

The effect of Alcohol is statistically non-significant and negative (beta = -0.13, 95% CI [-0.28, 0.02], t(10) = -1.86, p = 0.092; Std. beta = -0.34, 95% CI [-0.75, 0.07]). The effect of Measles is statistically non-significant and negative (beta = -1.99e-05, 95% CI [-6.32e-05, 2.34e-05], t(10) = -1.02, p = 0.330; Std. beta = -0.29, 95% CI [-0.94, 0.35]). The effect of Polio is statistically non-significant and positive (beta = 0.04, 95% CI [-0.04, 0.13], t(10) = 1.10, p = 0.297; Std. beta = 0.28, 95% CI [-0.28, 0.83]) .

The effect of Diphtheria is statistically non-significant and positive (beta = 0.07, 95% CI [-0.04, 0.18], t(10) = 1.40, p = 0.191; Std. beta = 0.31, 95% CI [-0.18, 0.80]).

Based on the regression model results, a variable importance histogram was obtained using the vip package of R language. The importance of variables can be intuitively compared from the bar chart. The importance of variables can be intuitively compared from the bar chart.

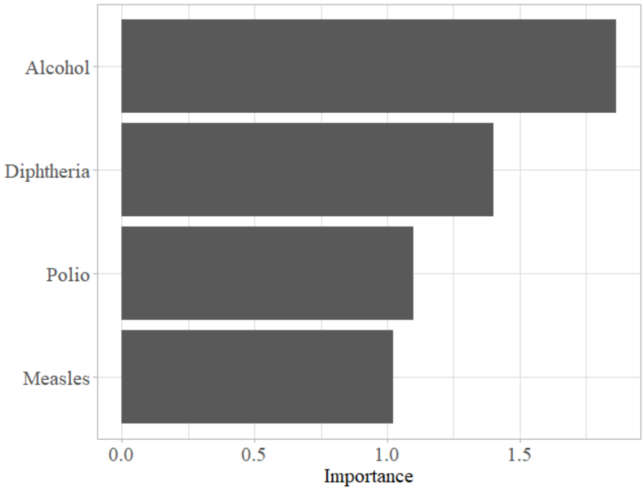


Figure 11. This chart displays the importance of different variables in influencing life expectancy.

Table 4. The results of the multicollinearity test of the national regression model for Japan are shown in the table.

variable	vif
Alcohol	1.104908
Measles	2.686382
Polio	2.042280
Diphtheria	1.571231

The graph shows that there is no multicollinearity.

Table 5. Taking Singapore as the research object, a regression model was constructed, and the obtained model results are shown in the table.

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	61.399	61.658	0.996	0.3428	
Alcohol	-3.198	3.764	-0.850	0.4153	
`Hepatitis B`	2.448	1.257	1.948	0.0800	.
Measles	-0.015	0.010	-1.546	0.1532	
Polio	-5.135	3.439	-1.493	0.1662	
Diphtheria	2.972	3.225	0.922	0.3784	

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*

Residual standard error: 1.796 on 10 degrees of freedom

Multiple R-squared: 0.5445, Adjusted R-squared: 0.3168

F-statistic: 2.391 on 10 and 5 DF, p-value: 0.1128

It can be observed that the overall p-value of the model is very large, so stepwise regression optimization is carried out to optimize the model. The results of the stepwise regression model are shown in the table below.

Table 6. The results of the stepwise regression model.

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	92.780	52.268	1.775	0.1012	
`Hepatitis B`	1.552	0.931	1.666	0.1215	
Measles	-0.014	0.005	-2.891	0.0136	*
Polio	-1.654	0.839	-1.972	0.0721	.

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*

Residual standard error: 1.785 on 12 degrees of freedom

Multiple R-squared: 0.4601, Adjusted R-squared: 0.3251

F-statistic: 3.409 on 12 and 3 DF, p-value: 0.0531

We fitted a linear model (estimated using OLS) to predict Life expectancy with Hepatitis B (formula: `Life expectancy` ~ `Hepatitis B` + Measles + Polio).

The final regression model equation was obtained.

`Life expectancy`=92.78+1.55`HepatitisB`-0.01Measles- 1.65Polio

The model explains a statistically not significant and substantial proportion of variance (R2 = 0.46, F(3, 12) = 3.41, p = 0.053, adj. R2 = 0.33). The model's intercept, corresponding to Hepatitis B = 0, is at 92.78 (95% CI [-21.10, 206.66], t(12) = 1.78, p = 0.101).

Within this model:

The effect of Hepatitis B is statistically non-significant and positive (beta = 1.55, 95% CI [-0.48, 3.58], t(12) = 1.67, p = 0.121; Std. beta = 0.63, 95% CI [-0.19, 1.46]).

The effect of Measles is statistically significant and negative (beta = -0.01, 95% CI [-0.02, -3.33e-03], t(12) = -2.89, p = 0.014; Std. beta = -0.66, 95% CI [-1.16, -0.16]).

The effect of Polio is statistically non-significant and negative (beta = -1.65, 95% CI [-3.48, 0.17], t(12) =-1.97, p = 0.072; Std. beta = -0.78, 95% CI [-1.64, 0.08]).

Using the vip package in R language, a variable importance histogram was obtained, from which it can be found that the measles variable has the highest importance.

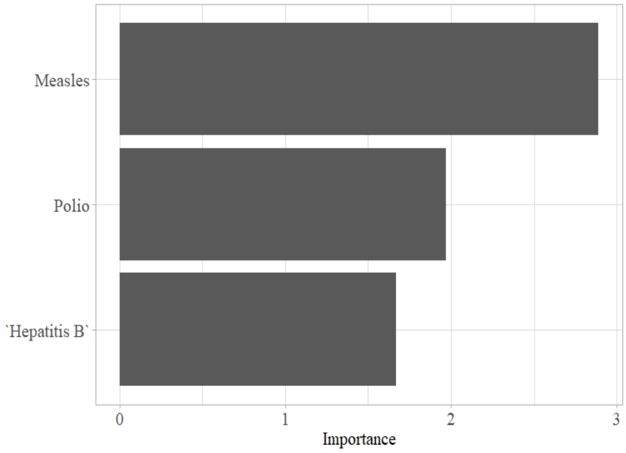


Figure 12. The x-axis represents the importance scores of the variables, and the y-axis lists the variables

The results of the multicollinearity test for the Singapore national regression model are shown in the table below. It can be found that the vif values are relatively small, so it can be considered that there is no multicollinearity.

Table 7. Collinearity test results table. This table presents the Variance Inflation Factor (VIF) values for each variable, used to assess multicollinearity issues.

variable	vif
`Hepatitis B`	3.199763
Measles	1.173143
Polio	3.476662

4.5. Multivariate Model

Multiple regression model is a mathematical model used for regression analysis. The regression model containing only one regression variable is called univariate regression model, otherwise it is called multiple regression model. The variables determined by the regression model are related. Under a large number of observations, they will show a certain regularity, which can be expressed with the help of functional relationship. This function is called regression function or regression equation. The five factors were hepatitis B, diphtheria and poliomyelitis immunization coverage, consumption of alcohol and number of measles cases. The relationship between these five factors and life expectancy was explored. Establish multiple regression model.

Table 8. Model result summary. This table presents the summary of linear regression model results,

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	51.649	14.369	3.594	0.0014	**
Alcohol	-1.624	0.315	-5.156	0.0000	***
`Hepatitis B`	0.027	0.024	1.117	0.2747	
Measles	-0.000	0.000	-1.834	0.0785	.
Polio	0.454	0.890	0.510	0.6144	
Diphtheria	-0.141	0.970	-0.145	0.8855	

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*

Residual standard error: 1.735 on 25 degrees of freedom

Multiple R-squared: 0.8524, Adjusted R-squared: 0.8228

F-statistic: 28.86 on 25 and 5 DF, p-value: 0.0000

We fitted a linear model (estimated using OLS) to predict Life expectancy with Alcohol (formula: `Life expectancy` ~ Alcohol + `Hepatitis B` + Measles + Polio + Diphtheria). The model explains a statistically significant and substantial proportion of variance (R2 = 0.85, F(5, 25) = 28.86, p < .001, adj. R2 = 0.82). The model's intercept, corresponding to Alcohol = 0, is at 51.65 (95% CI [22.06, 81.24], t(25) = 3.59, p = 0.001).

Regression diagnosis is the test and analysis of the assumptions and data in regression analysis. It usually contains two aspects: (1) Check whether the assumptions in the regression analysis are reasonable. For example, in a linear regression model, it is usually assumed that the random errors are independent, the expectation is zero, and the variance is the same, or they are further assumed to obey the normal distribution. One of the problems to be solved by regression diagnosis is to test whether these assumptions are reasonable. If these assumptions are not Reasonable, what kind of corrections can be made to the data to make them satisfy or approximately satisfy these assumptions. (2) Diagnose the data, check whether there is abnormal data in the observation value, and how to deal with it when there is abnormal data. The results of model diagnosis on this model are shown below. The following conclusions were drawn from the diagnosis results. Residual error and fitted value (upper left). The data points between residual and fitted values are evenly distributed on both sides of y=0, showing a random distribution. The red line shows a smooth curve without obvious shape characteristics. In the residual QQ graph (upper right), the data points are arranged in a diagonal straight line, tending to a straight line, and are directly crossed by the diagonal, which intuitively conforms to the normal distribution. Standardized residual square root and fitted value (bottom left), the data points are evenly distributed on both sides of y=0, showing a random distribution and the red line shows a smooth curve without obvious shape characteristics. Standardized residuals and leverage values (bottom right), there are red contour lines, indicating that there are abnormal points in the data that particularly affect the regression results.

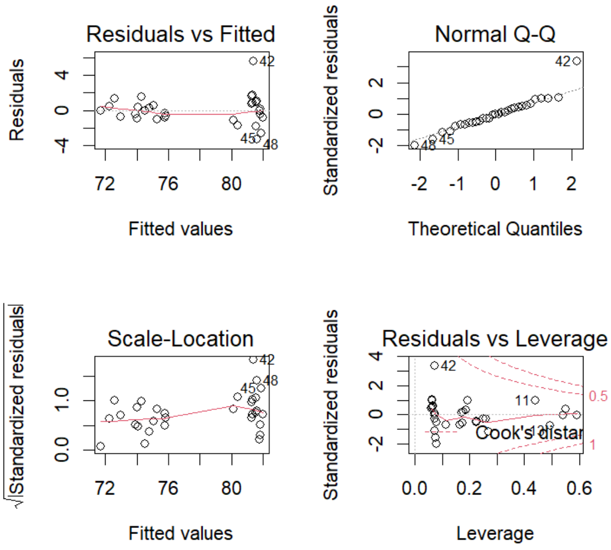


Figure 13. The plot is used to evaluate the fit quality and residual behavior of the regression model

D-W test: The null assumption of the DW test is that the errors are not correlated. According to the test result, the P-value is less than 0.05, so the null hypothesis is rejected, that is, the error is considered to be correlated.

```
lag Autocorrelation D-W Statistic p-value
   1      0.2122339153      1.398986169      0.038
 Alternative hypothesis: rho != 0
```

Using the VIP package, obtain the importance of variables and display the importance of different variables on the dependent variable by drawing a bar chart.

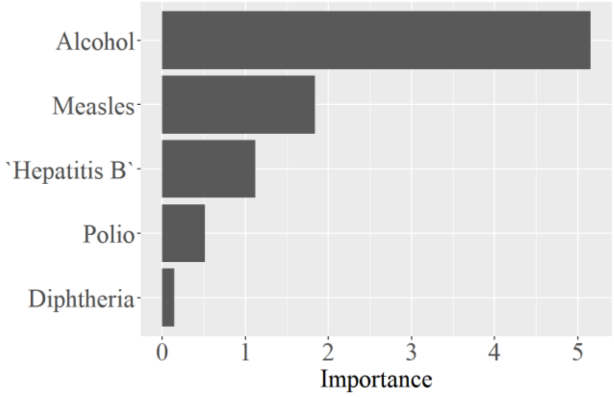


Figure 14. This chart displays the importance of different variables in influencing life expectancy.

A decision tree is a basic classification and regression method, and a regression decision tree mainly refers to the CART algorithm. The values of internal node features are "yes" and "no", and it is a binary tree structure. The so-called regression is to determine the corresponding output values based on the feature vectors, and the regression tree is to divide the feature space into several units, each of which has a specific output. Because each node is judged by "yes" and "no", the boundary of the partition is parallel to the coordinate axis. For test data, classify it into a certain unit according to its characteristics to obtain the corresponding output value. Based on the principle of the decision tree algorithm, a machine learning model is constructed to predict and analyze the lifespan of Asian countries.

A visual decision tree model showcases the branching process of the model.

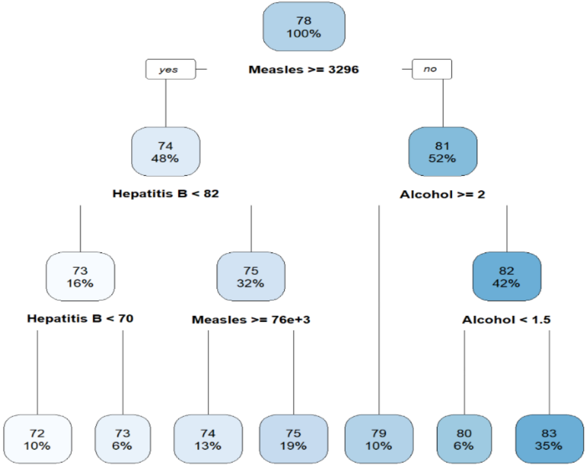


Figure 15. This chart presents a decision tree model used to predict life expectancy.

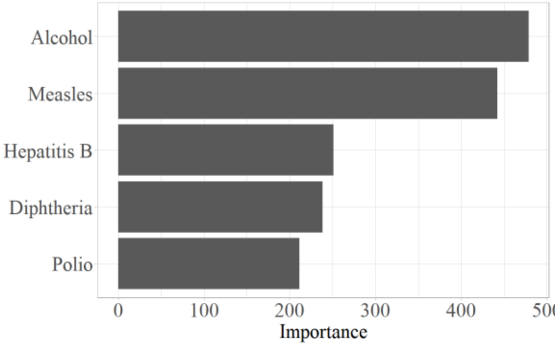


Figure 16. Based on the Figure 15, draw a bar chart of variable importance.

4.6. Model Comparison

Two models were constructed, including a multiple regression model and a decision tree model, and the models were validated on a test set. The following are the evaluation indicators for the model.

MAE (Mean absolute error):

$$\mathrm{MAE} = \frac{|p_1 - a_1| \ + \cdots + |p_n - a_n|}{n}$$

MSE (Mean squared error):

$$\mathrm{MSE} = \frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}$$

Root relative squared error:

$$\mathrm{RMSE} = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$$

R square:

$$RSS = \frac{\sum_{i=1}^{n}(p_i - \bar{a})^2}{\sum_{i=1}^{n}(a_i - \bar{a})^2}$$

Table 9. (Comparison of model results)

	R2	MAE	MSE	RMSE
Linear regression	0.8523515931	1.101700212	2.428098794	1.558235795
rpart model	0.9473812594	0.5043010753	0.8653225806	0.9302271662

5. Conclusions

According to the results of the study, the correlation between "Alcohol", "Hepatitis B", "Measles", "Polio" and "Diphtheria" and life expectancy was analyzed.

Poliomyelitis is a disease that can lead to severe disability and even death. The incidence of the disease can be significantly reduced through immunization, which has a direct impact on life expectancy. This direct health benefit may make the relationship between polio immunization and life expectancy even more significant.

Alcohol consumption may lead to health problems that reduce life expectancy, but the data may not capture the full extent or intensity of this effect.

Measles vaccine has significantly reduced measles morbidity and mortality. Thus, the effect of measles on life expectancy may have been offset by the vaccine, making it difficult to observe a statistically significant association.

Weak or no Polio correlation was observed, possibly due to data quality issues and generally high vaccination rates resulting in very low disease incidence.

In this study, data from China, Singapore, and Japan were selected to represent Asian data, and correlation testing, regression statistics, residual analysis, and outlier detection were used to analyze the relationship between longevity and other factors in Asian countries. And the following conclusions were drawn: 1) There is almost no relationship between alcohol consumption and life expectancy. 2) There is little relationship between hepatitis B vaccine coverage under one year of age and life expectancy. 3) There was a weak positive correlation between coverage of diphtheria vaccines under one year of age and life expectancy rates. 4) There was a positive correlation between the one-year immunization rate for polio and the life expectancy rate. 5) There was a negative association between the number of measles cases and life expectancy.

In comparing linear regression and random forest models, multiple evaluation indicators such as the r square was selected, and it was found that the decision tree model had a higher r square. Therefore, it is recommended to choose the decision tree model for research in the field of predicting people's lifespan in Asian countries.

In predicting life expectancy, decision tree models are better than linear regression models in some cases for the following reasons:

Nonlinear relationship:

The relationship between life expectancy and its influencing factors may be non-linear. Linear regression models assume that there is a linear relationship between variables, which may lead to poor prediction results in the presence of nonlinear relationships. The decision tree model is able to handle nonlinear relationships naturally because it works by recursively splitting the data set into different subsets, each with its own prediction rules.

Robustness to outliers and missing values:

Decision tree models are less sensitive to the distribution and outliers of the data, so they are more robust. When dealing with data sets that contain outliers or missing values, decision tree models are generally able to give more stable results. Linear regression models may require more complex pre-processing steps or additional assumptions when dealing with such data.

The decision tree model is insensitive to missing values and can process irrelevant feature data.

6. Discussion

In this study, in fact, half of the results did not match the hypothesis.

Using the literature as an example, accurate life expectancy prediction at the national and regional level, the study aims to explore trends in regional life expectancy differences in China, divided by province, urban and rural population, which makes the investigation of life expectancy not limited to the national level. The results found that, across all regions, the differences in life expectancy at birth and age 65 are largely driven by differences in life expectancy among rural populations in each region, with life expectancy at birth in the least developed regions catching up with developing regions by 2040.

The results do not match the assumptions, possibly for several reasons:

Regional differences: There are significant differences in economic conditions, medical facilities, social support, etc. in different regions, and these factors may have a greater impact on life expectancy than the disease factors we studied. For example, poor and remote areas may have a longer life expectancy due to a slower pace of life and less environmental stress, even though they face a higher risk of disease.

Population differences: Different social groups (such as rich, poor, urban residents, rural residents, etc.) have significant differences in health behaviors and access to medical resources. For example, while wealthier people in cities may be at higher risk for alcohol consumption, they often have access to better medical care, which helps mitigate the negative impact of disease on life expectancy.

Therefore, if the study continues, I hope to be able to refine specific areas and specific populations, so as to conduct more targeted studies on the relationship between these factors and life expectancy.

References

[1] Yu, D., Lu, B., & Piggott, J.R. (2022). Alcohol consumption as a predictor of mortality and life expectancy: Evidence from older Chinese males. The Journal of the Economics of Ageing.

[2] Nguyen, M. H., Wong, G., Gane, E., Kao, J. H., & Dusheiko, G. (2020). Hepatitis B Virus: Advances in Prevention, Diagnosis, and Therapy. Clinical microbiology reviews, 33(2), e00046-19. https://doi.org/10.1128/CMR.00046-19.

[3] Thorrington, D., Ramsay, M., van Hoek, A. J., Edmunds, W. J., Vivancos, R., Bukasa, A., & Eames, K. (2014). The effect of measles on health-related quality of life: a patient-based survey. PloS one, 9(9), e105153. https://doi.org/10.1371/journal.pone.0105153.

[4] Yang, E. J., Lee, S. Y., Kim, K., Jung, S. H., Jang, S. N., Han, S. J., Kim, W. H., & Lim, J. Y. (2015). Factors Associated with Reduced Quality of Life in Polio Survivors in Korea. PloS one, 10(6), e0130448. https://doi.org/10.1371/journal.pone.0130448.

[5] Zhang, Q., Han, F., Nie, Q., Ren, H., Zhang, B., Liu, Q., He, Q., & Shao, Z. (2011). Seroprevalence of antibodies to pertussis and diphtheria among healthy adults in China. The Journal of infection, 63 6, 441-6 .

[6] Katz, S., Branch, L. G., Branson, M. H., Papsidero, J. A., Beck, J. C., & Greer, D. S. (1983). Active life expectancy. New England journal of medicine, 309(20), 1218-1224.

[7] Östergren, O., Martikainen, P., & Lundberg, O. (2018). The contribution of alcohol. consumption and smoking to educational inequalities in life expectancy among Swedish men and women during 1991–2008. International journal of public health, 63(1), 41-48.

[8] Siddiqi, N., Khan, A., Nisar, N., & Siddiqi, A. E. (2007). Assessment of EPI. (expanded program of immunization) vaccine coverage in a peri-urban area. Jpma, 57(8), 391-395.

[9] Hussey, G. D., & Klein, M. (1990). A randomized, controlled trial of vitamin A in children with severe measles. New England journal of medicine, 323(3), 160-164.

[10] Jakovljevic Mihajlo,Ogura Seiritsu. Editorial: Insights in health economics: 2021 [J]. Frontiers in Public Health,2022,10.

[11] Li, J., Bateman, H., & Liu, K.Z. (2014). Regional Differences in Life Expectancy in China.