Group Number: 8
1155205212 TAN Xinyi
1155201146 HAO Yitan
1155201551 ZHU Zhixuan
1155205754 YUAN Liang

Statistical Analysis on Influence Factors of Life Expectancy

I. Introduction

The influence factors of life expectancy have always been one of the hottest research topics. In this project, apart from analyzing the factors that affect life expectancy, the comparison between developed countries and developing countries would also be considered. In addition, some special cases which are different from the overall analysis would be discussed in detail, some interesting discoveries would be present in the project. The data comes from WHO and United Nation Website including life expectancy, health factors, and economic data for 193 countries from year 2000 to 2015. The multiple linear regression, LASSO and machine learning will be used to explore the impact of different factors on life expectancy, and the extent of their impact. Finally, time series will also be performed to predict life expectancy.

II. Dataset and Data Cleaning

i. Data Description

The dataset is collected from the WHO data repository website[1], the United Nation website[2] and the World Bank website[3]. The dataset is related to life expectancy of 193 countries from 2000-2015, consisting of 22 columns and 2918 rows. The dependent variable is life expectancy. The independent variables are divided into several broad categories: immunisation-related factors, mortality factors, economic factors and social factors, such as Hepatitis B immunization coverage, adult mortality rates, GDP, schooling and so on. The detailed description of data is shown in the data appendix.

ii. Data Cleaning

In the data cleaning process, the first step is to handle duplicate data in the dataset. To achieve this, Python is used to detect duplicates, and the result shows that there are no duplicate data. And since the dataset is ranged from 2000 to 2015, those countries whose data is less than 16 years are found and dropped. The next step is to handle missing values. Different methods are adopted for different situations. For those data with a few missing values, which are below 50, the overall mean of these variables are chosen to impute since their impact is relatively small. For variables with missing values greater than 100, a more complicated method is adopted. First, a heatmap is used to identify the variable with the highest correlation coefficient with the variable to be imputed. Then, different intervals are divided based on the distribution features of the scatter plot, and the mean of the variable in those intervals is calculated to fill the corresponding missing values. The same method is adopted for other variables with missing values greater than 100, except for population and GDP, where some obvious mistakes are found. False data are found in the population and GDP variables, and since there are a lot of missing values in these two variables, an official dataset is collected from the World Bank to replace the original population and GDP data. After completing these steps, all the missing values are appropriately imputed, and the dataset is ready for analysis.

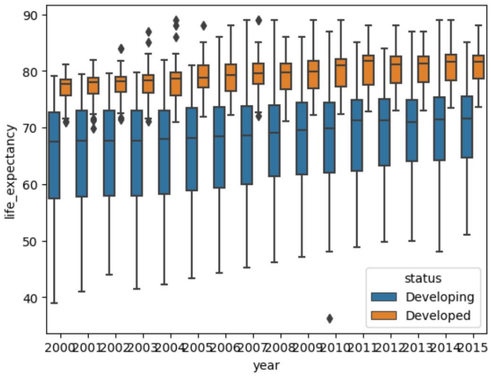
III. Research Questions

The study focuses on the following five questions.

A. Compare life expectancy differences between developed and developing countries.

The box plot of life expectancy in developed and developing countries from 2000 to 2015 is drawn for illustration (see Figure 1). As can be seen, the median lines of developed country boxes lie outside the developing country boxes, indicating that developed countries have a higher level of life expectancy than developing countries. Furthermore, the inter-quantile ranges in developed countries are narrower than those in developing countries, suggesting that the life expectancy is less dispersed in developed countries and more dispersed in developing countries.

Figure 1. Life expectancy in developed and developing countries from 2000 to 2015



B. What is the relationship between life expectancy and other factors? How to predict life expectancy with other factors?

The data needs to be further processed before regression. The binary variable status is labeled 0 for developing countries and 1 for developed countries. The dataset is split into 70% training set and 30% test samples.

i. Ordinary Least Squares (OLS) Regression

The OLS method finds the coefficients that minimize the sum of squares of the difference between actual and predicted y values. Modeling on the training set, the R-squared is 0.837, indicating a good fitting. The global F-test is conducted to test joint hypotheses. Since the p-value is less than 0.1, it is concluded that the overall model is significant. Then T-tests are performed to test the significance of individual variables. In conclusion, six variables are not statistically significant for regression: status, alcohol, percentage expenditure, hepatitis B, total expenditure, and thinness 5-9 years. In addition, eight variables have a significant positive impact on life expectancy: infant death, BMI, polio, diphtheria, GDP, population, income composition of resources, and schooling. Five variables have a significant negative effect on life expectancy: adult mortality, measles, under-five deaths, HIV/AIDS, and thinness 1-19 years. By fitting the model to the test sample, the in-sample MSE is 3.8576.

ii. Least Absolute Shrinkage and Selection Operator (LASSO) Regression

The LASSO method estimates a linear regression with a large set of candidate x variables. It fits the model by shrinking coefficients and, importantly, shrinking some of them to zero[4]. The grid search cross-validation (GridSerachCV) algorithm is used to search for the best turning parameter based on the training set. The best turning parameter obtained by the algorithm is 0.00001. By fitting the model to the test sample, the in-sample MSE is 3.8584.

iii. CART Model

The CART model is a decision tree-based algorithm that recursively splits the dataset. However, CART models tend to overfit the data. To address the issue, pruning techniques were applied. Pruning involves reducing the complexity of the tree, which helps improve the model's generalization performance on unseen data. Before pruning, the MSE calculated by this method is 8.402. After pruning, the MSE becomes 7.814, which has dropped significantly.

iv. Random Forest

Random Forest is an ensemble learning method that combines multiple CART models to make predictions. It's able to capture complex interactions between variables and reduce overfitting compared to individual CART models. The MSE output for this method is 3.628, which shows that RF performs better than CART.

v. Comparison of Models

Several statistical methods are used in this analysis, and RMSE is used to compare the performance of different models.

Table 1. Comparison of Different Models

Model	**RMSE**
OLS	1.96408062
LASSO	1.96427922
CART(before pruning)	2.89854334
CART(after pruning)	2.79542782
RF	1.90483084

The following conclusions can be drawn from the table. The Random Forest model exhibits the smallest RMSE, indicating its superior performance on the dataset. However, it requires a longer training time and lackes interpretability compared to linear models. Linear models, such as OLS and LASSO, perform similarly and outperform the CART model in terms of predictive performance. Despite the improvement achieved through pruning, the performance of the CART model remains suboptimal.

C. In the above regression analysis, are there some results that are contrary to normal cognition? Why?

The coefficient for population is positive and statistically significant (p-value 0.017). The result may be counterintuitive because population size itself does not directly lead to an increase in life expectancy. To further explore the impact of population on life expectancy, the GDP per capita variable was introduced. The model shows a very significant coefficient for GDP per capita, which is 80.776, indicating that an increase in GDP per capita is significantly associated with an increase in life expectancy. High per capita GDP may reflect a higher level of national development, including better healthcare and social services, which in turn could lead to higher life expectancy.

D. How to perform time series regression?

i. Selection of data and model for analysis

The ARIMA(p, d, q) Model is utilized to demonstrate the time series regression analysis of the life expectancy of different countries. Considering that the life expectancy data for each country constitutes an independent set of variables, presenting them together in a single chart would lead to chaos. Consequently, a careful selection of 10 representative countries out of the total 181 has been made. These countries are China, Denmark, Canada, Egypt, France, the Dominican Republic, Brazil, the United States of America, India, and Japan.

ii. Stepwise ARIMA(p, d, q) model building

The visualization of the time series data first reveals a general increase in life expectancy over time in most countries. However, there are notable instances of sharp fluctuations within short periods that require further analysis.

To determine the stationary of the time series, the Augmented Dickey-Fuller (ADF) test is employed. The p-value obtained from the test results is compared against a predetermined significance level (e.g., 0.05). If the p-value is lower than the significance level, the null hypothesis is rejected, indicating the presence of stationary data. The p-values for all ten countries are above 0.05 by now, suggesting non-stationary for most countries. Therefore, the first difference transformation is applied to reduce the variance of the time series. After this transformation, the p-values for most countries fall below 0.05, allowing the rejection of the null hypothesis and treating the differenced data as stationary. The value of the parameter d is set to 1.

The Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) charts are plotted to determine the optimal orders p and q for the ARIMA model. The ACF charts reveal characteristics of first-order censoring or tailing, while the PACF charts exhibit corresponding first-order censoring. Based on the analysis of these charts, two models are established: ARIMA(1,1,0) and ARIMA(1,1,1).

iii. Construction and evaluation of the ARIMA models

The Bayesian Information Criterion (BIC) is used to evaluate the constructed ARIMA models. BIC values are calculated for each model, and lower values indicate better model fit. It is observed that the BIC values differ for each country under the two models. For more accurate prediction results, it is recommended to analyze the data for each country separately.

IV. Conclusion

Based on previous research, the following key findings have been obtained. First, the boxplot is used to explore the differences in life expectancy between developed and developing countries. Developed countries have a higher level of life expectancy. Then three techniques (multiple linear regression, LASSO, and machine learning) are used to investigate the influence factors of life expectancy and further explore their relationships. Random Forest fits the best with the lowest RSME of 1.9048. In multiple linear regression, some results are contrary to normal cognition. The interaction term is used to explain. Last but not least, the ARIMA model is used to perform the time series regression.

V. References

[1] Data Retrieved from https://www.who.int/

[2] Data Retrieved from https://www.un.org/

[3] Data Retrieved from https://data.worldbank.org/

[4] Békés, G., & Kézdi, G. (2021). *Data analysis for business, economics, and policy* (pp.407-408). Cambridge University Press.