Empirical Report for Linear Econometrics Class

Determinations of Life Expectancy

Introduction

Exploring key factors that influence health is vital for predicting future health trends and life expectancy, a significant measure of a country's living standards. This study aims to identify critical areas for improving life expectancy in a nation's population. Additionally, this research provides me with valuable insights and directions for achieving a longer, healthier life.

The study "Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death" by Kyle J Foreman et al. (2018) suggests that factors such as high BMI, HIV/AIDS, alcohol use, and poverty negatively impact life expectancy, while the main health drivers are expected to show improvement. Our data, sourced from the WHO and United Nations, contains variables highlighted in this study. We intend to use our model to confirm the correlation between these factors and life expectancy as indicated in the paper and to predict life expectancy based on these variables.

Methods

In our research, we consider *life expectancy* as the dependent variable, influenced by various predictors such as a *country's development status, alcohol consumption, BMI, HIV/AIDS mortality, government health expenditure, GDP, adult mortality rate, and population size.* We employ multiple linear regression techniques to examine the interplay between the dependent variable and these predictors. Initially, we create a comprehensive model incorporating all predictors.

$$\begin{aligned} Life.expectancy_i &= \beta_0 + \beta_1 Status_i + \beta_1 Status_i + \beta_2 Alcohol_i + \beta_3 BMI_i + \beta_4 HIV.AIDS_i \\ &+ \beta_5 Total.expenditure_i + \beta_6 GDP_i + \beta_7 Adult.Mortality_i \\ &+ \beta_8 Population_i + \varepsilon_i, (i = 1,2,..n) \end{aligned}$$

In this model, β_i represents the coefficient for the *i-th* predictor, ε_i is the error term, and *'Status'* is a binary variable indicating whether a country is developing (Status = 1) or not (Status = 0).

For model validation, we divide our dataset into training and testing sets, with 80% of the data used for model development and the remaining 20% for testing the model's

efficacy. We utilize t-tests to assess the significance of the regression coefficients and F-tests to evaluate the overall model significance. This approach helps determine the full model's effectiveness and identify any irrelevant predictors.

Variable Selection

To identify the most impactful predictors, we utilize different stepwise selection approaches: backward, forward, and stepwise methods, all oriented around Akaike's Information Criterion (AIC). The backward selection strategy initiates with a complete model, progressively eliminating one predictor at a time to minimize the AIC value in each successive step. In contrast, forward selection starts with a simplistic model, including just the intercept term, and incrementally adds predictors to lower the AIC value at each stage. The stepwise selection method merges aspects of both forward and backward selection, iteratively adding or removing predictors to achieve the model with the lowest AIC value.

Model Diagnostics

Model diagnostics can be performed by verifying model assumptions, checking for multicollinearity, and identifying troublesome data points. By examining the residual plots, we can determine whether the basic principles of linear regression, such as linearity, error homogeneity, error normality, and error uncorrelation, are satisfied.

Multicollinearity, which refers to the distortion or difficulty in estimating the model accurately due to the presence of exact correlation or high correlation between the explanatory variables in a linear regression model, is a key issue in regression modeling. This can be assessed by calculating the variance inflation factor (VIF) for each variable; VIF represents the ratio of the variance of the estimated regression coefficients compared to the variance if no linear correlation is assumed between the independent variables, and if the VIF is greater than 5, then multicollinearity is significant.

Problematic data can manifest as leverage points (points with large residuals), high leverage values (points far from the center of the sample space), and impact points (points that have a large impact on the model and if removed can change the fitted regression equation). When hat values are used for each of these values, if the hat statistic is greater than three times the average hat value, then the observation can be judged to be a high-leverage point. Standardized residuals, if the absolute value of the standardized residuals is >3, the observation is judged to be an outlier, and when the distance from Cook exceeds a specific threshold, this is detected as an influence point.

Model Validation

In the model validation phase, we use the training data to assess the predictive accuracy

of the model and then apply it to the test data. Prediction Mean Square Error (PMSE) is a numerical measure of a model's predictive ability. The smaller the prediction mean square error, the higher the prediction accuracy for unseen data. In addition, we evaluated the adjusted R^2 value of the test data to measure how well the model fits the new data compared to the training set. This comparative analysis allows us to measure the relative effectiveness of the full model versus the final selected model.

Based on this, we can consider using cross-validation techniques or other error metrics such as Mean Absolute Error (MAE), a measure of the error between pairs of observations expressing the same phenomenon, for a more comprehensive validation process. In addition, understanding the trade-off between complexity and performance could be discussed, as well as considering the impact of different variable choices on model generalizability. Adding these elements would provide a more reliable assessment of the predictive strength and reliability of the model.

Results

Exploratory Data Analysis

In our Exploratory Data Analysis, we scrutinized a dataset that captures the nuances of life expectancy, health dynamics, economic influences, and population demographics from 193 countries. Initially, the dataset included 2938 data points across 22 variables. Following a thorough cleaning process to exclude any incomplete records and narrowing our focus to key variables, we were left with a streamlined dataset encompassing 2099 data points spanning 9 critical variables. Insights gleaned from this refined data were visually represented in a scatterplot matrix showcased as Figure 1, from which we extracted meaningful conclusions.

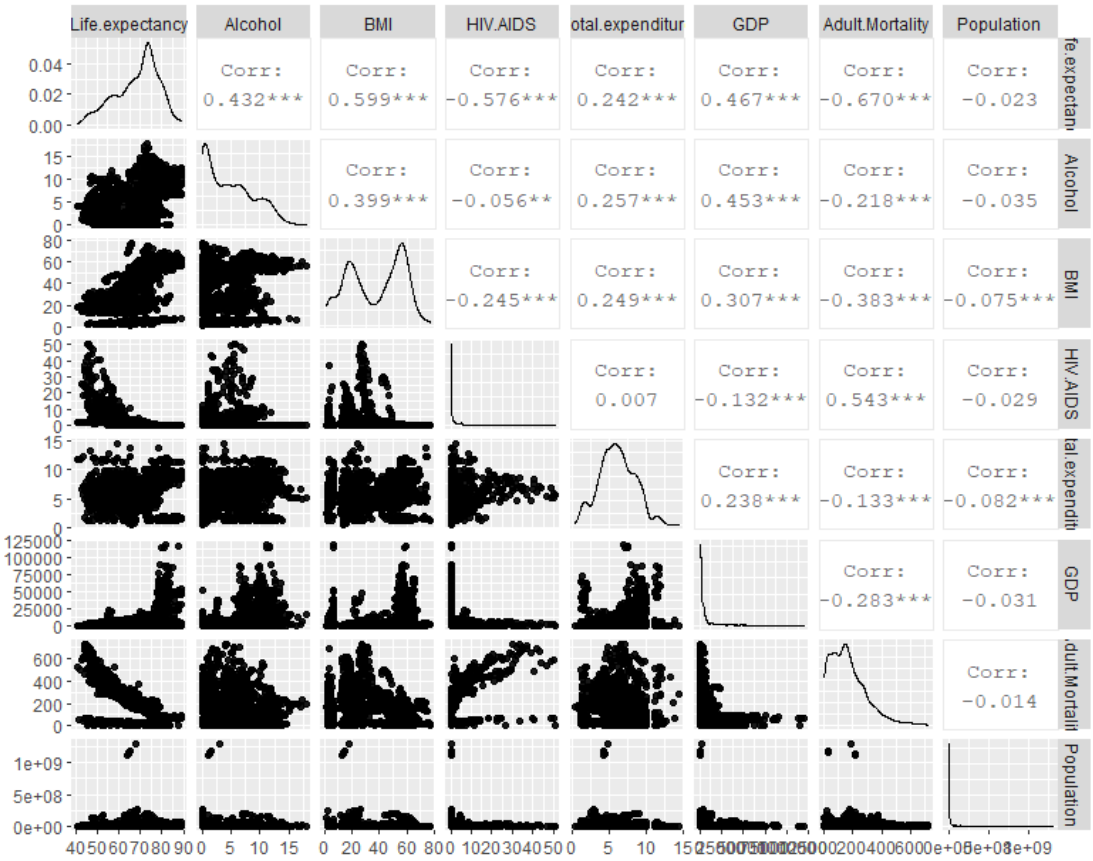

Figure 1: the scatterplot matrix of numerical variables of interest

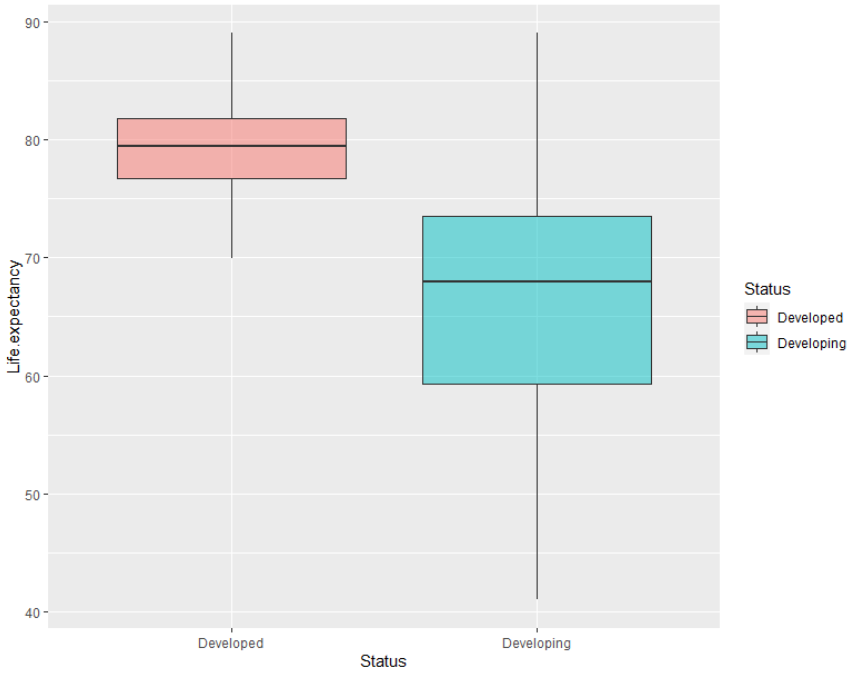

Figure 2: Boxplot of life expectancy versus country status

Our study has uncovered that life expectancy has a direct and positive correlation with factors like BMI, healthcare spending by governments, and a nation's economic output.

This implies that countries investing in health and economic growth tend to enjoy longer average lifespans. Additionally, Figure 2 corroborates the finding that residents of developed nations generally live longer than those in developing nations. However, life expectancy is inversely affected by negative health indicators, such as high rates of HIV/AIDS, large populations, and increased rates of adult mortality.

Process of Obtaining Final Model

The summary of the full model analysis using R software indicates that all the predictors, except for population, significantly contribute to life expectancy based on their t-test results. The overall F-test has a very small p-value, suggesting the model is robust in accounting for a large portion of the variance in life expectancy. Furthermore, an R^2 value of 0.7171 implies that approximately 71.68% of the variance in life expectancy is captured by the model.

Given that population is not a significant predictor, variable selection procedures were applied. Both stepwise and backward selection methods recommended excluding the population variable, while forward selection retained all predictors. For the sake of simplicity, the final chosen model excludes the population predictor.

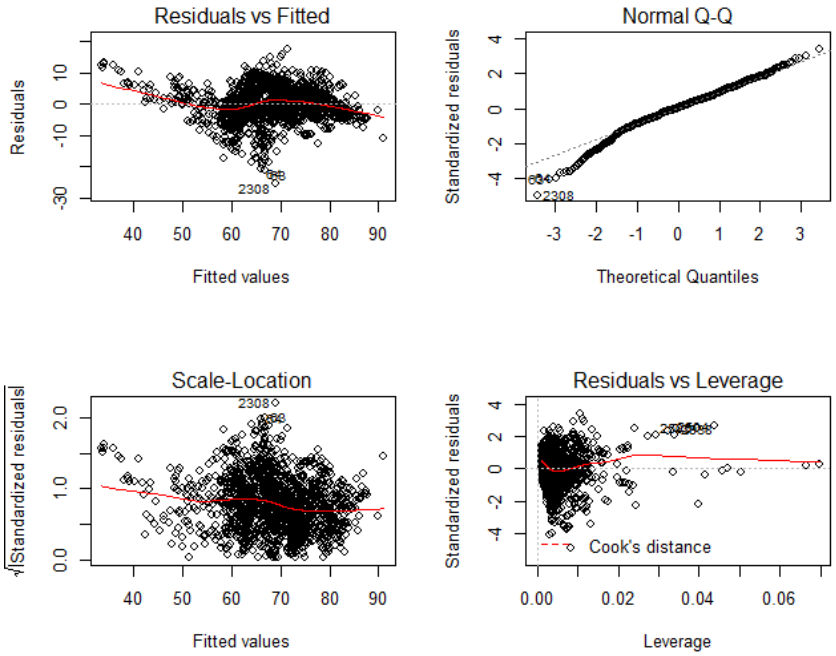

Figure 3: The residuals plots of selected (final) model

Table 1 presents the estimates, standard errors, and t-test results for the coefficients of the final model. It is clear from the table that all predictors are significant, as indicated by their p-values being less than 0.001.

Table 1: Summary table of the coefficients of final model

	Estimate	Std. Error	t value	P value	Significance
(Intercept)	6.760e+01	7.008e-01	96.466	< 2e-16	***
StatusDeveloping	-2.875e+00	4.609e-01	-6.238	5.61e-10	***
Alcohol	2.204e-01	4.156e-02	5.303	1.29e-07	***
BMI	1.410e-01	7.554e-03	18.666	< 2e-16	***
HIV.AIDS	-5.081e-01	2.699e-02	-18.826	< 2e-16	***
Total.expenditure	2.213e-01	5.707e-02	3.878	0.000109	***
GDP	9.800e-05	1.109e-05	8.838	< 2e-16	***
Adult.Mortality	-2.223e-02	1.280e-03	-17.372	< 2e-16	***
Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**					

Goodness of Model

As shown in Figure 3, the assumptions of selected model do hold. In the plot of residuals versus fitted values, there is no clear pattern to appear where residuals are randomly distributed around the horizontal dash line. It shows the assumption of linearity is not violated. The normal QQ plot of residuals shows that the normality of residuals is satisfied since most of points are on the straight line. The independence of errors may be hard to assess but we have known the observations are independently collected and there are not timing pattern about the residuals. The assumption of common variance of errors seems to be reasonable.

Based on the hat values of observations, we see that there are about 8.76% of observations in the training dataset are with high leverage. The observation indexed by 2308 are identified as an outlier, because the absolute value of standardized residual is greater than 4. There is no any influential points that affect the model if they are removed.

To validate the performance of model, the PMSE of final model on the testing data is recorded as 30.4152, which is slightly higher than that of full model. It indicates variable selection is effective for model improvements. The out-of-sample adjusted R squared is 0.2919196, which is lower than in-sample adjusted squared R 0.7156. It shows the model is under the risk of overfitting, but we are still confident in the predictive power of established model.

Discussion

In this project, we explore the relationship between life expectancy of people for 197 countries and other factors including the development, demographics, expenditure on health etc. The results show that the average life expectancy of people in developing countries is lower than that in developed countries. Holding all else constant, every 1

unit increase in BMI leads to around 0.14 years longer in lifespan, or every 1% more government funds places on the health, the life expectancy of people tends to extend 0.22 years. A country with lower HIV infection rate and adult mortality rate tends to have a higher life expectancy for people. But it is surprising that there is high correlation between the life expectancy and the alcohol consumption.

However, there are some limitations in this project. We select 9 variables of interest into modelling based on the prior knowledge instead of all 22 variables. In the future work, we can complete this task for unveiling the complete relationship between the response and all potential predictor variables. The interaction terms of some predictors should be taken into consideration since the effect of one predictor on the response may depend on another predictor variable.

Reference list

1. KumarRajarshi. Life expectancy (WHO) [Internet]. Kaggle. 2018 [cited 2021Oct22]. Available from: https://www.kaggle.com/kumarajarshi/life-expectancy-who

2. “Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death”
   Foreman KJ, Marquez N, Dolgert A, Fukutaki K, Fullman N, McGaughey M, et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: Reference and alternative scenarios for 2016–40 for 195 countries and Territories. The Lancet. 2018;392:2052–90.