STA302 group7 project Research report

Haobo Wang
Ricky xu
Jack Zhou
Jiaxuan Song
Beiyang Chen

Introduction:

Life expectancy is a crucial indicator of a country's overall health and well-being (Tulchinsky et al., 2023). Numerous studies have highlighted the different factors that influence life expectancy, such as education and income levels. For example, research indicates that people with higher education levels tend to have better health and longer lifespans compared to their less educated peers (Raghupathi & Raghupathi, 2020), while individuals in higher income groups often live longer due to their ability to afford better health regulations and services (Bayar et al., 2021). Inspired by these insights, our study utilizes a comprehensive dataset from the World Health Organization (WHO) available on Kaggle. This dataset covers life expectancy and related health factors for 193 countries from 2000 to 2015(*Life Expectancy (WHO)*, 2018). The aim of our project is to analyze the impact of various factors on human lifespan using multiple linear regression and to develop a model to predict life expectancy.

However, analyzing life expectancy data may pose several challenges. The complexities of socioeconomic factors, cultural differences, and varying healthcare systems across countries make it difficult to identify causal relationships. Additionally, the potential for multicollinearity among predictors and the presence of outliers in the data may complicate the modeling process. To address these challenges, we can employed techniques such as Variance Inflation Factor (VIF) analysis to detect and mitigate multicollinearity, F-tests to compare nested models and ensure the inclusion of significant variables, and graphical analysis to visually inspect data patterns, outliers, and model fit.

These rigorous methodologies help to build a robust model that can accurately predict life expectancy, providing valuable insights for public health policy and intervention strategies.This research is closely aligned with the STA302 course, as it applies key statistical methods, particularly multiple linear regression, to real-world data.

Methodologies

The cleaning of the dataset

Handling Missing Values:
We used the na.omit() function to remove any rows with missing data, ensuring that only complete and reliable data points were included in our analysis.
Excluding the 'Country' Variable:
Since our goal is to predict overall life expectancy, we excluded the 'Country' variable to focus on global patterns rather than country-specific ones.
Random Sampling:
We randomly selected 1,000 observations from the cleaned dataset to create a manageable subset for analysis, using a random seed (set.seed(302)) to ensure reproducibility.
Data Splitting:
The data was split into 70% for training and 30% for testing, allowing us to build the model and then validate its performance on a separate dataset.

Variable Selection

In this section, we focus on the process of selecting the most relevant variables for predicting life expectancy. The goal was to develop a parsimonious model that includes significant predictors while avoiding multicollinearity and overfitting. The process involved fitting multiple models and using statistical criteria to refine the variable set.

Model 1: Initial Full Model

The first model, Model 1, included a broad set of predictors based on literature and logical reasoning. We considered variables such as *Year, Status (developed or developing), Adult Mortality, Infant Deaths, Hepatitis B, HIV/AIDS, GDP, Population, Thinness (1-19 years), Thinness (5-9 years), Income Composition of Resources, and Schooling.* The rationale for including these variables was based on their potential impact on life expectancy, as suggested by previous research.

Model 2: Refining the Model by Removing Insignificant and high VIF Variables

After analyzing the results of Model 1, predictor variables that may not be significant will be identified. To simplify the model, it is planned to remove variables with p-values greater than 0.05 and VIF values greater than 5. Through this process, the model will be retained to include only variables that may significantly affect life expectancy.

Model 3: Reduce model2

In Model 3, the impact of removing variables with relatively high p-values will be further tested. Although these variables may have been significant in Model 2, their relatively high p-values make it necessary to test whether the explanatory power of the model is significantly

reduced in its absence. The purpose of this step is to determine whether a more streamlined model can be as effective as a model that includes the removal of variables with relatively high p-values.

Model Comparison

In order to make a decision between Model 2 and Model 3, an ANOVA partial F test will be performed. This test is used to compare nested models to determine if the model that includes more variables significantly improves the fit of the model. Based on the results of this test, the most appropriate final model will be selected.

Model Validation

In this section, we will focus on validating the final model that was developed through the variable selection process. The purpose of this validation is to ensure that the model not only fits the training data well but also generalizes effectively to new, unseen data. To achieve this, we will employ a range of model validation techniques, including performance evaluation on both the training and testing datasets.

Evaluating Model Performance on the Training Data

The first step in the validation process involves evaluating the model's performance using the training dataset. Key metrics such as the adjusted R-squared, residual sum of squares (SSres), AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) will be calculated. These metrics provide insight into how well the model fits the training data, while also accounting for model complexity and the risk of overfitting.

Testing Model Generalization on the Testing Data

To ensure the model's robustness, we will evaluate its performance on the testing dataset, which was not used during the model-building phase. This step is crucial to assess the model's generalization ability and to detect any potential overfitting that may have occurred during training.

- **Fit Comparison**: The final model's performance on the testing data will be compared against its performance on the training data. We will look for consistency between the two sets, which would indicate that the model generalizes well.
- **Assessment of Limitations**: If the model's performance on the testing data does not meet expectations, we will identify and discuss potential limitations. This may include issues such as model bias, variance, or the need for further refinement of the model's structure.

Through these validation steps, we aim to confirm that the final model is both accurate and reliable when applied to new data, ensuring that it is suitable for predicting life expectancy in various contexts.

Model Diagnostics

To ensure the validity of the model, several diagnostic checks were performed using residual plots and QQ plots:

1. **Linearity**: The residual plot should show no pattern. If there is a curve pattern, the linearity assumption is violated.
2. **Constant Variance**: The residual plot should show no pattern. If the variance changes with the level of the fitted values, the assumption of homoscedasticity is violated.
3. **Normality**: The QQ plot should show points aligned along a straight line. Deviations from this line suggest violations of the normality assumption.
4. **Independence**: The residual plot should show no pattern. Clustering or patterns indicate a violation of the independence assumption.

Additionally, Outliers and influence diagnostics were performed:

- **Outliers**: Points with residuals $r > 4$ or $r < -4$ were flagged as potential outliers.
- **Influential Points**: Points with high Cook's Distance D_i were noted as influential and examined further.

These diagnostic checks are essential to ensure that the final model meets all necessary assumptions for valid inference and reliable predictions.

Furthermore, If there are significant patterns in the residual plots, we will either square some of the more influential variables or introduce interaction terms to improve the model.

Result

Description of Data

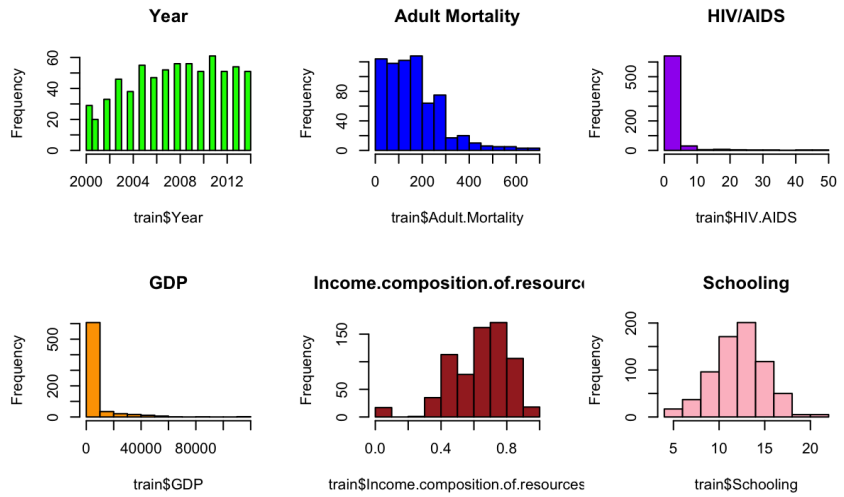
Data sources and size: The source of the data is a dataset on Life Expectancy Data found in the kaggle website. The overall data totaled 2939 rows and 22 columns.

Summary of the training data:

Table 1: Summary the data of the numerical variable

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Year	2000	2005	2008	2008	2011	2014
Adult Mortality	1	75	145	163.3	224	693
HIV/AIDS	0.1	0.1	0.1	1.878	0.7	49.1
GDP	11.15	488.56	1612.05	5902.71	5138.32	119172.7
Income Composition of Resources	0	0.5078	0.679	0.6391	0.7595	0.93
Schooling	4.2	10.4	12.3	12.21	14.2	20.6

For this numerical table, we can find numbers such as the maximum and minimum values and the median and mean. For example, the maximum value of the year is 2014 and the minimum value is 2000.



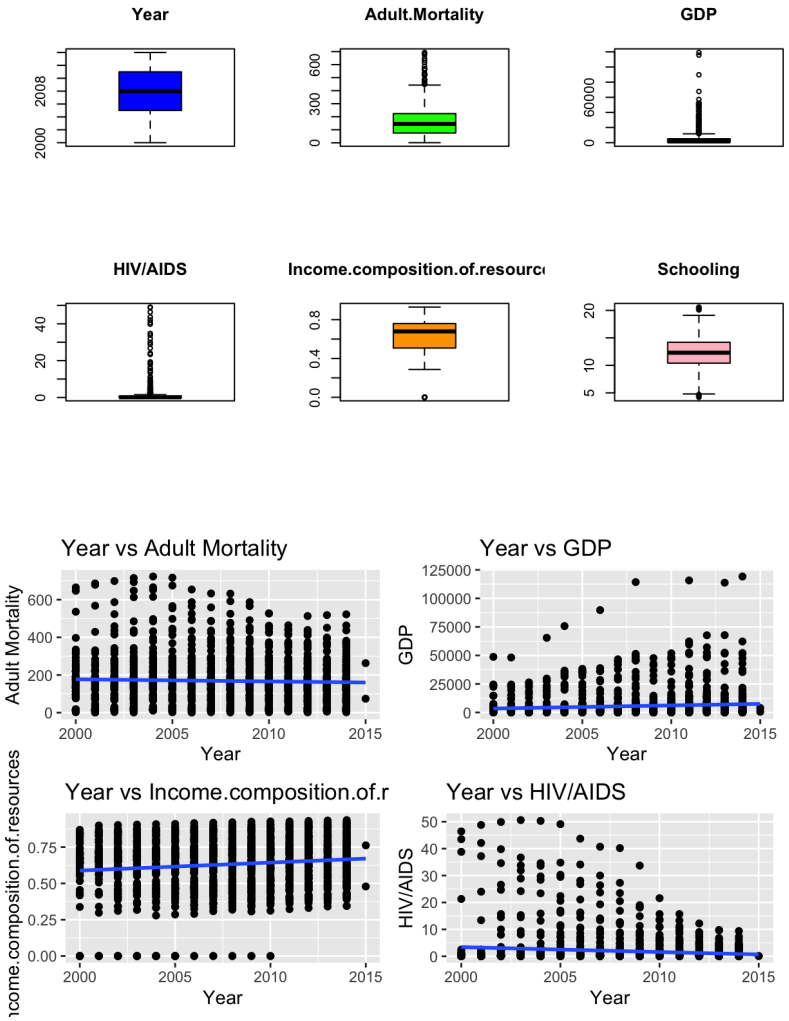


Figure1: histograms, boxplot and scatter plots of numerical variable

For the above three kinds of graphs, first of all, the top one is histogram, through which we can observe the distribution pattern, range and its concentration of data. For example, Year: distribution type is roughly uniform distribution. Range is 2000 to 2015 and data is evenly distributed, with no noticeable skew. In addition, in the figure we can find that HIV/AIDS and also GDP are highly right-skewed.For the box plot, we can use it to observe the outliers as well as the distributional characteristics of the data, in which we can clearly find high outliers for GDP and HIV, as well as a higher concentration of years. Finally, based on the scatterplot, we can find out how the variables change over the years.

Analysis process and results

Step 1: We first utilize Year, Status, Adult.Mortality, infant.deaths, Hepatitis.B, HIV.AIDS, GDP, Population, thinness of 1 ~ 19 years, thinness of 5~9 years, Income composition of resources,Schooling, these variables are used as x-variables in model 1, through which we

study their effect on Life expectancy, which is our y-variable. In addition, we used train data, that is, we chose 70% of the dataset for training. Thereby, we get our model 1.

Step 2: According to the data in Model 1, we deleted the variables with P-value greater than 5. This is because these variables do not reflect their effect on Life expectancy well. Therefore, based on this screening, we are left with Year, Adult Mortality, HIV/AIDS, GDP , Income composition of resources, Schooling as our X. This leads us to our Model II. In addition, based on the data from Model II we find that none of the variables in Model II has a p-value of more than 5, indicating that all of them are able to analyze Life expectancy well. We also get an Adjusted R-squared of 0.8283, which shows that this model can fit Life expectancy well.

Step 3: We try to delete the variable GDP to form model 3, however, we find that we get an Adjusted R-squared of 0.822, which is not as good as the value of model 2, so we abandon model 3.

Step 4: we attempted an ANOVA partial F test to compare models 2 and 3. The reason for this maneuver was that we wanted to see the likelihood of model 2 continuing to simplify. As a result, we found that the P-value obtained was less than 0.05, so we chose model 2, which is the full model, to be our model.

AIC BIC for model2, model3

Model 2 - AIC: 3824.238

Model 2 - BIC: 3860.646

Model 3 - AIC: 3848.528

Model 3 - BIC: 3880.386

Hence, by ANOVA partial F-test and the AIC and BIC values for model2 and model3, we can conclude that model 2 is better.

True model for model2:

Life Expectancy = β0 +β1× Year + β2 × Adult Mortality + β3 × HIV/AIDS + β4 × GDP + β5 × Income Composition of Resources + β6 × Schooling + ε

Table 2: Data for model 2:

$\widehat{Life\ Expectancy}$ = b0 + b1× Year + b2 × Adult Mortality + b3 × HIV/AIDS + b4 × GDP + b5 × Income Composition of Resources + b6 × Schooling				
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	3.795e+02	7.104e+01	5.342	1.25e-07
Year	-1.626e-01	3.541e-02	-4.592	5.22e-06
Adult.Mortality	-1.961e-02	1.456e-03	-13.466	2e-16
HIV.AIDS GDP	-4.451e-01 6.854e-05	2.750e-02 1.331e-05	-16.184 5.150	2e-16 3.40e-07
Income composition of resources	1.335e+01	1.367e+00	9.765	2e-16
Schooling	9.613e-01	8.505e-02	11.303	2e-16
Multiple R-squared: 0.8298				
Adjusted R-squared: 0.8283				

Goodness of the Final Model

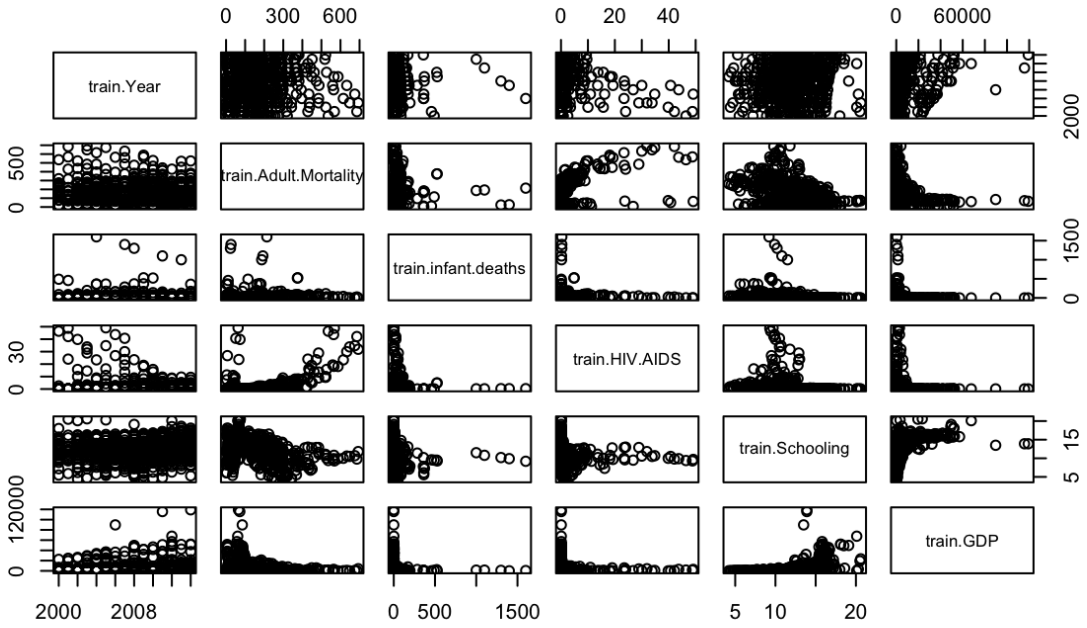


Figure 3. Pair Plot of Year, Adult Mortality, HIV/AIDS, GDP , Income composition of resources, Schooling.

There are no very significant biases or outliers in this graph, and the trend suggests that higher GDP and education levels are generally associated with better health outcomes.

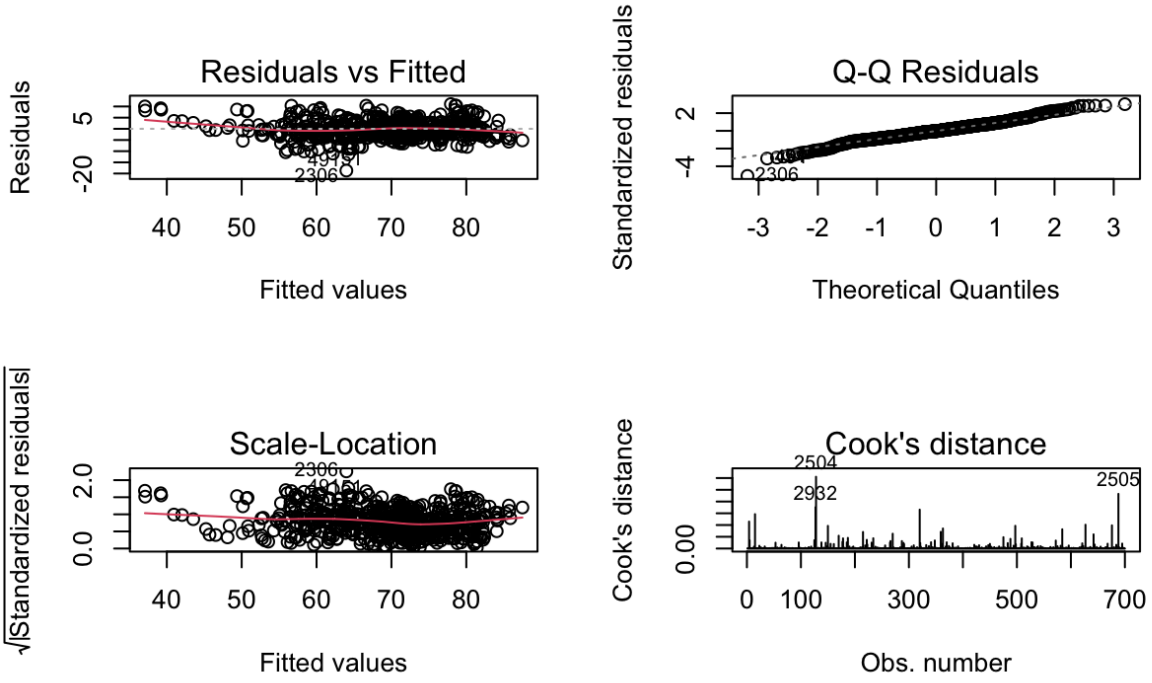


Figure 4.Residuals vs Fitted Plot & Q-Q Plot & Scale-Location Plot & Cook's Distance Plot

First, the Residuals vs Fitted Plot demonstrates that the residuals for all of these values are clustered around the fitted value, which indicates a good model fit. Second, in the Q-Q Plot, most of the points are close to a straight line and the residuals are normally distributed. Third, in the Scale-Location Plot, the variances of the residuals are generally consistent. Fourth, in Cook's Distance Plot, there are three influential points, but the impact is not significant.

Conclusion:

The Final Model is:

$\widehat{Life\ Expectancy}$ =379.5 –0.1626× Year -0.01961× Adult Mortality -0.4451 × HIV/AIDS + 0.00006854 × GDP + 13.35 × Income Composition of Resources + 0.9613 × Schooling.

Train model VS test model:

	Intercept	Year	Adult Mortality	HIV/AIDS	GDP	Income Composition of Resources	Schooling
Training Model	379.5	-0.1626	-0.01961	-0.4451	6.854e-05	13.35	0.9613
Test Model	266.2	-0.1049	-0.02047	-0.471	7.512e-05	8.131	1.048

The minimal differences between the coefficients in the training and test models indicate that the model demonstrates good generalization and stability, with no significant evidence of overfitting.

According to our final model, we can clearly see that Year, Adult Mortality, HIV/AIDS, GDP, Income Composition of Resources and Schooling are closely correlated to Life expectancy. In general, each coefficient (the number multiplying a variable) represents the expected change in the outcome for a one-unit increase in that variable, holding all other variables constant. To be more specific, with **Income Composition of Resources** increasing by **one unit**, while **other variables stay constant**, the **life expectancy could increase by 13.35 units on average.** The model suggests that factors like income composition of resources, GDP and schooling have positive effects on the outcome, while factors like year, adult mortality and HIV/AIDS have negative effects.

Based on our final model, our research findings are consistent with the results of the two studies mentioned in the introduction. We found that individuals with higher education levels generally have better health and longer lifespans, which aligns with the findings of Raghupathi and Raghupathi (2020). Additionally, our study also shows that people in higher income groups tend to live longer, a conclusion that echoes the findings of Bayar et al. (2021).

Limitation:
There are some limitations in our project. Specifically, our model has not been validated, and there are outliers present. For instance, our "Cook's Distance plot" reveals three influential points (2504, 2932, 2505). The outliers in boxplot could potentially affect the validity of our predictions. Additionally, in our final model, we only considered a few variables, which means we may have overlooked other critical factors that influence lifespan, such as genetic influences, medical conditions, and the natural environment.

Acknowledgements:

We would like to express our sincere gratitude to our Professor and all the TAs, for their invaluable guidance and support throughout this project. We also extend our appreciation to all team members for their hard work and collaboration. All team members did almost the same amount of work, Ricky xu and Haobo Wang did the modeling and plotting parts and some of the written word , Jack Zhou, Jiaxuan Song and Beiyang Chen did the complete introductions, conclusions and other written work.

Reference:

Life expectancy. Life Expectancy - an overview | ScienceDirect Topics. (n.d.-a). https://www.sciencedirect.com/topics/social-sciences/life-expectancy

Raghupathi, V., & Raghupathi, W. (2020, April 6). *The influence of education on health: An empirical assessment of OECD countries for the period 1995–2015 - archives of public health*. BioMed Central. https://archpublichealth.biomedcentral.com/articles/10.1186/s13690-020-00402-5

Bayar, Y., Gavriletea, M. D., Pintea, M. O., & Sechel, I. C. (2021, December 14). *Impact of environment, life expectancy and real GDP per capita on health expenditures: Evidence from the EU member states*. MDPI. https://www.mdpi.com/1660-4601/18/24/13176

Life expectancy (WHO). (2018, February 10). Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who