Statistical Modeling of Life Expectancy based on Sociodemographic Factors

ISyE6414 A (Fall 2023) Team 2

Shiven Barbare
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA
sbarbare3@gatech.edu

Priyanka Singh
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA
priyanka.singh@gatech.edu

Karan Nahar
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA
knahar3@gatech.edu

Prasanthi Mounika
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA
ptoram3@gatech.edu

Manikant Thatipalli
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA
mthatipalli3@gatech.edu

***Index Terms*—regression, residual analysis, hypothesis testing, life expectancy, predictive analytics**

I. Introduction

Life expectancy is a statistical measure that estimates the average number of years a person is expected to live depending upon distinct factors. Determinants of life expectancy include gender, ethnicity, medical history, lifestyle, social structure and welfare policies, economic factors such as inflation and per capita income and environmental factors. It provides an indication of the overall health and well-being of a population, and is often used as a key indicator in public health and policy making to assess the quality of life and healthcare services in a given demographic. Advancements in technology, better healthcare facilities, and education have led to positive changes in the lifestyle of people, thereby, increasing the average expected age of a human being in the last few decades.

Data Science and Machine Learning models can be used to explain the factors influencing the rise of life expectancy around the world. These analytical techniques can also indicate methods for improving life expectancy of a distinct population. There have been a lot of regression based studies in the past on factors affecting life expectancy considering demographic variables income composition and mortality rate. However these studies did not take into account the effect of immunization and human development index. Previous research for prediction of life expectancy include multiple linear regression based models on a data set of one year for all the countries.

We aim to resolve both the factors stated previously by building a multiple regression model considering data from years 2000 - 2015 for approximately 193 countries spanning from developing nations to developed countries. We will also consider important immunization factors such as Hepatitis B, Polio and Diphtheria in addition to mortality, economic, social and health related factors. Such a model would be useful for providing practical insights and recommendations to helping developing nations for driving their average life expectancy

II. Data Description

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many related factors for all countries. The data sets are made available to the public for the purpose of health data analysis. The data set related to life expectancy, health factors for 193 countries has been collected from the WHO data repository website and its corresponding economic data was collected from the United Nation website. Among all categories of health related factors only critical and more representative factors were chosen.

The data set consists of 22 columns and 2938 rows. The description of variables available in the data set are as follows.

Response Variable: **Life Expectancy** - A statistical estimate of the average number of years a person is expected to live in the country

Predictors:

- Country: 193 countries across the globe
- Year: Year of data collection (between 2000 and 2015; 5 unique values possible)
- Status: Developed or Developing country
- Mortality: Adult mortality rates for any gender (probability of dying between 15 and 16 years per 1000 population)
- Infant Deaths: Infant deaths per 1000 population
- Alcohol: Per capita consumption of alcohol (in liters) for ages $\geq$15

- Percentage Expenditure: Expenditure on health as a percentage of GDP
- Hepatitis B: HepB immunization coverage among one-year olds
- Measles: Cases reported per 1000 population
- BMI: Body Mass Index (average)
- Under-5 Deaths: Deaths of children aged under 5 per 1000 population
- Polio: Pol3 immunization coverage among one-year olds
- Total Expenditure: Government health expenditure as a percentage of total government expenditure
- Diphtheria: DTP3 immunization coverage among one-year olds
- HIV/AIDS: Death per 1000 live births due to HIV/AIDS (ages 0-4)
- GDP: Gross Domestic Product (per capita in USD) of the country
- Population: Population of the country
- Thinness 10-19 years: Percentage prevalence of thinness among children aged 10-19
- Thinness 5-9 years: Percentage prevalence of thinness among children aged 5-9
- Income Composition of Resources: Human Development Index in terms of income composition of resources (ranges from 0 to 1)
- Schooling: Average number of years of schooling in the population

Note: Only Country and Status are qualitative/categorical variables.

III. Data Cleaning and Pre-processing

Data cleaning involves identifying and correcting errors and inconsistencies in data such as missing values, outliers and duplicates. Common data cleaning techniques include imputation, removal and transformation.

Our original data set consists of 22 columns and 2938 rows with 2 categorical variables - Country and Status. The country variable consists of 193 distinct values. Since having 192 Country factor variables in our final regression model would make model interpretation difficult, we used a country-continent mapping to create another variable Continent.

For certain countries, (namely Tuvalu, San Marino, Saint Kitts and Nevis, Palau, Niue, Nauru, Monaco, Marshall Islands, Dominica, Cook Islands), we had data corresponding to a single year only, so we dropped rows corresponding to these countries.

Certain columns such as Population, GDP, Total expenditure, Hepatitis B, Alcohol, Schooling, Income Composition of Resources were missing values for certain countries. Since these variables are highly specific for a particular country and cannot be imputed using the average values of the whole data set, we dropped columns with missing values.

For removal of outliers for continuous variables, we removed the corresponding rows containing Z scores greater/lesser than 3.

Finally, we scaled all the continuous variables using min-max scaler in order to transform all variables to [0,1] range for model fitting.

IV. Exploratory Data Analysis

Before modeling, it was essential to get a general sense of what the data looked like, which is why it was important to do some basic exploratory data analysis (EDA).

The first set of analyses focused on looking at the distribution of the individual variables to get a general sense on what range of values does each variable have. This is shown in 1. Life expectancy seems to be distributed normally, which

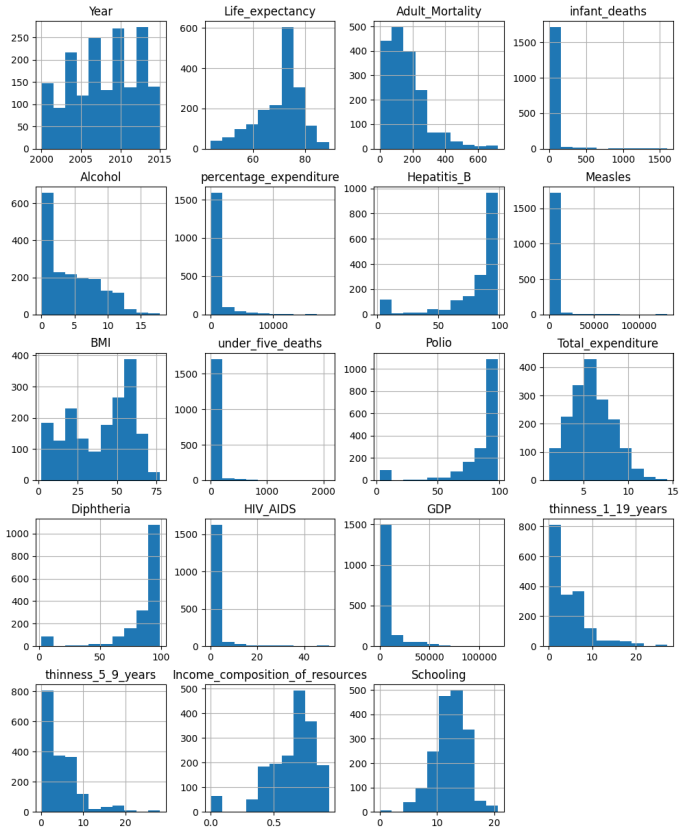

Fig. 1. Distribution of variables

is good for modeling.

One set of analyses focused on understanding the variation of the life expectancy with the categorical variables- status and continent. This variation is shown in 2 and 3.

We can clearly see that some continents like Europe have a higher median value of life expectancy compared to Afirca. Similarly, the life expectancy (median) in developed countries is significantly different from that in developed countries.

Another set of analyses was centered around on how life expectancy varies with the quantitative variables as shown 4, 5 and 6.

The scatter plots show that life expectancy generally increases with Schooling and income composition of resources,

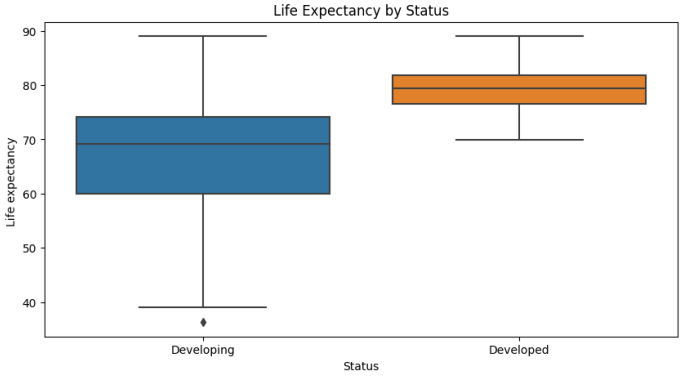

Fig. 2. Variation of Life Expectancy with Status

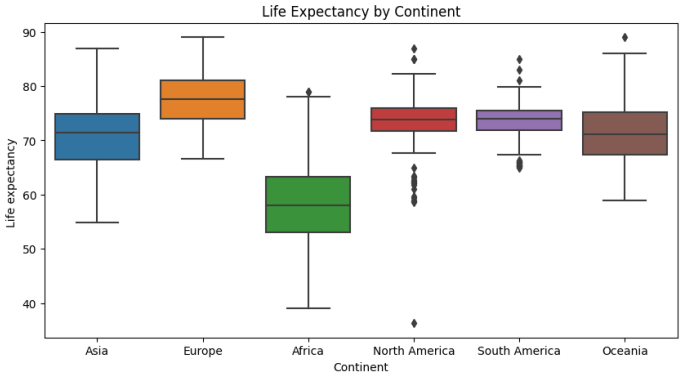

Fig. 3. Variation of Life Expectancy with Continent

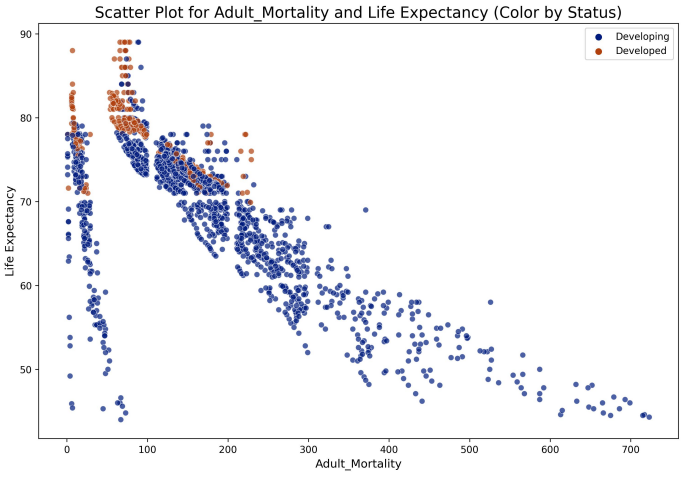

Fig. 4. Variation of Life Expectancy with Adult Mortality

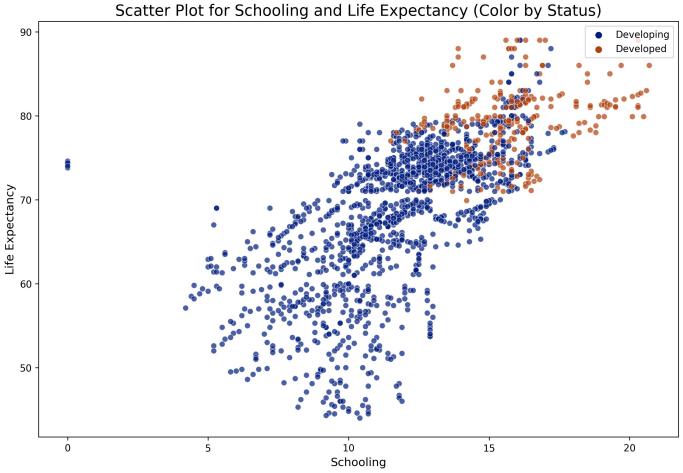

Fig. 5. Variation of Life Expectancy with Schooling

Fig. 6. Variation of Life Expectancy with Income Composition

and decreases with adult mortality, with higher values for developed countries.

Next, it was important to look at the correlation between the variables and flag the predictors that are highly correlated with each other. This forms the basis of multicollinearity analysis discussed later in this report. The correlation matrix is shown in 7. We can see that the variable under_five_deaths and infant_deaths are highly correlated, and so are GDP and percentage_expenditure. So, they need to be examined.

V. Modeling and Diagnostics

The approach for iterative modeling has been described in the flowchart shown in 8. This section focuses on the aspects of model fitting, residual analysis, diagnostics, and model improvement.

The base model can be represented as

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

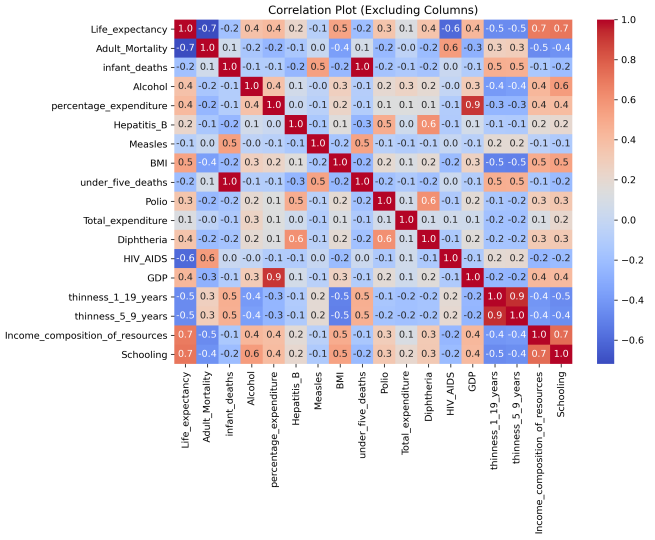

Fig. 7. Correlation Matrix

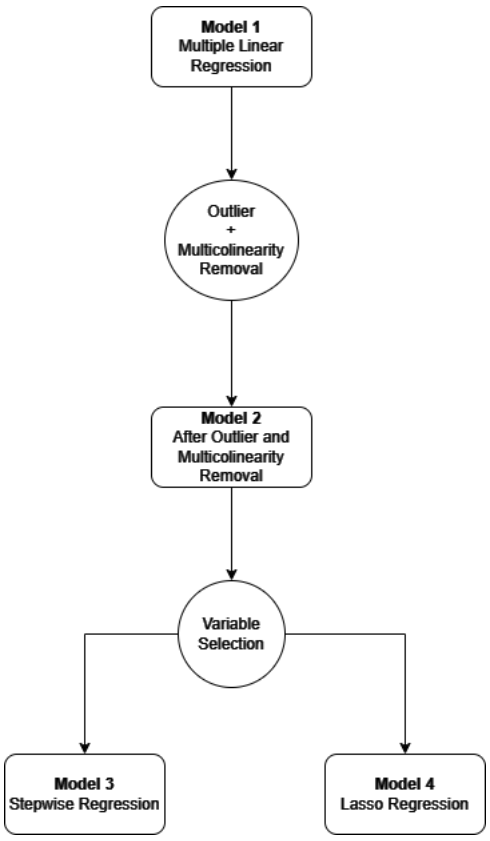

Fig. 8. Modeling Overview

where y is the response (Life Expectancy), β are the coefficients and x are the predictors used in the model.

A. Model 1: Multiple Linear Regression (Full Model)

The base model was built using multiple linear regression (MLR) considering all predictors after initial data cleaning and pre-processing. The summary of the model is shown 9.

```
> summary(model)

Call:
lm(formula = Life_expectancy ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25228 -0.04859  0.00173  0.04390  0.33960

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.2783308  0.0189840  14.661  < 2e-16 ***
Year                           -0.0263871  0.0078803  -3.348 0.000834 ***
Adult_Mortality                -0.2400467  0.0162317 -14.789  < 2e-16 ***
infant_deaths                   1.9634776  0.3753121   5.232 1.94e-07 ***
Alcohol                        -0.0835813  0.0149861  -5.577 2.93e-08 ***
percentage_expenditure          0.1595559  0.0542424   2.942 0.003320 **
Hepatitis_B                     0.0002136  0.0099174   0.022 0.982823
Measles                        -0.0186424  0.0323705  -0.576 0.564772
BMI                             0.0214601  0.0102176   2.100 0.035882 *
under_five_deaths              -1.9789024  0.3617710  -5.470 5.33e-08 ***
Polio                           0.0240513  0.0110924   2.168 0.030307 *
Total_expenditure               0.0122141  0.0126451   0.966 0.334255
Diphtheria                      0.0378243  0.0130900   2.890 0.003917 **
HIV_AIDS                       -0.4166783  0.0219962 -18.943  < 2e-16 ***
GDP                             0.0572455  0.0472656   1.211 0.226045
thinness_1_19_years            -0.0054406  0.0318989  -0.171 0.864596
thinness_5_9_years             -0.0175672  0.0330626  -0.531 0.595274
Income_composition_of_resources  0.1352832  0.0160469   8.431  < 2e-16 ***
Schooling                       0.3540357  0.0266006  13.309  < 2e-16 ***
ContinentAsia                   0.0604243  0.0071592   8.440  < 2e-16 ***
ContinentEurope                 0.0944154  0.0092950  10.158  < 2e-16 ***
ContinentNorth America          0.1282517  0.0086403  14.843  < 2e-16 ***
ContinentOceania                0.0463710  0.0104269   4.447 9.39e-06 ***
ContinentSouth America          0.0915315  0.0103724   8.825  < 2e-16 ***
StatusDeveloping               -0.0429163  0.0082943  -5.174 2.62e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07449 on 1395 degrees of freedom
Multiple R-squared:  0.853,     Adjusted R-squared:  0.8504
F-statistic: 337.2 on 24 and 1395 DF,  p-value: < 2.2e-16
```

Fig. 9. Full Model Summary

We can see that most of the variables are not significant considering a significance level, α of 0.05. The sign of the coefficients makes sense for most, but not all variables so multicollinearity may be an issue here. Clearly there is room for improvement in this model.

The F-value for the overall model is quite high (and corresponding p-value is low), which means that we can reject the null hypothesis that all variables are insignificant, i.e. the model, as a whole, is significant.

Residual Analysis: The residual vs x plots are shown in 10. Since the residuals are randomly distributed about the mean zero line, the model satisfies the assumption of linearity.

The residual vs fitted values plot shows a random pattern with no extreme values of residuals, so our error terms are uncorrelated with constant variance.

The QQ plot and histogram show that the residuals have an approximately normal distributions but a light tail which needs to be analysed. Overall, the model seems to be a reasonably good fit to the data.

There are some outliers in the data as we can see from the plot of cook's distance in 10. These need to be examined further before removing them.

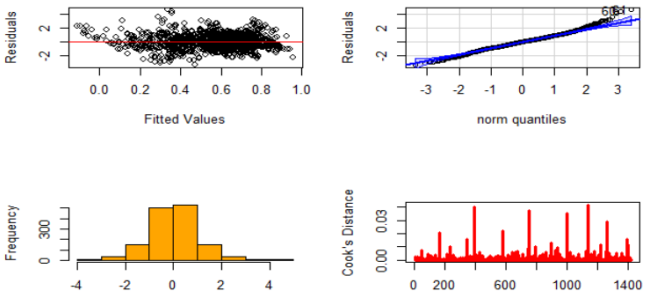

Fig. 10. Residual Analysis and Cook's Distance for Model 1

Box Cox Transformations: To check whether a transformation of the response variable is essential, we used the `boxcox` function in R to estimate the transformation parameter λ using maximum likelihood estimation. The result is shown in 11.

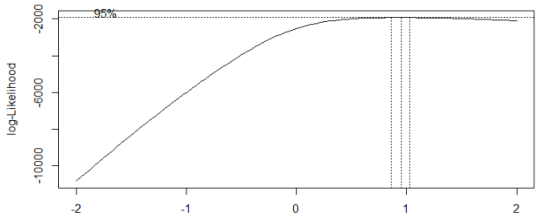

Fig. 11. Estimating transformation parameter, λ

Multicollinearity: The variance inflation factor (VIF) for each predictor is summarized in 12. Clearly there are two variables infant_deaths and under_five_deaths that high extremely high VIF values, indicating multicollinearity may be an issue here.

	GVIF	Df	GVIF^(1/(2*Df))
Year	1.178341	1	1.085514
Adult_Mortality	2.073796	1	1.440068
infant_deaths	287.857535	1	16.966365
Alcohol	2.904313	1	1.704205
percentage_expenditure	5.933048	1	2.435785
Hepatitis_B	1.760308	1	1.326766
Measles	1.654748	1	1.286370
BMI	1.876686	1	1.369922
under_five_deaths	280.876871	1	16.759382
Polio	1.730943	1	1.315653
Total_expenditure	1.269440	1	1.126694
Diphtheria	2.131901	1	1.460103
HIV_AIDS	1.706070	1	1.306166
GDP	5.924118	1	2.433951
thinness_1_19_years	7.203346	1	2.683905
thinness_5_9_years	7.358569	1	2.712668
Income_composition_of_resources	2.958380	1	1.719994
Schooling	3.737043	1	1.933143
Continent	8.698719	5	1.241497
Status	2.306234	1	1.518629

Fig. 12. VIF for Model 1

B. Model 2: Post Outlier and Multicollinearity Removal

The base model was built using multiple linear regression (MLR) with outliers removed along with highly correlated variables removed as well. The summary of the model is shown 13. Though many variables still seem insignificant (by p-values), the exclusion of outliers and columns with multicollinearity enhanced the R^2 from 85

```
> summary(model2)

Call:
lm(formula = Life_expectancy ~ ., data = data_subset)

Residuals:
      Min        1Q    Median        3Q       Max
-0.184609 -0.044029  0.003507  0.041291  0.231706

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.234630   0.017065  13.749  < 2e-16 ***
Year                            -0.029073   0.006845  -4.247 2.32e-05 ***
Adult_Mortality                 -0.268213   0.015055 -17.816  < 2e-16 ***
Alcohol                         -0.101024   0.013056  -7.738 2.01e-14 ***
percentage_expenditure           0.162300   0.047672   3.405 0.000683 ***
Hepatitis_B                      0.003072   0.008773   0.350 0.726303
Measles                          0.032558   0.036588   0.890 0.373718
BMI                              0.019254   0.009008   2.137 0.032745 *
under_five_deaths               -0.090548   0.031958  -2.833 0.004676 **
Polio                            0.027614   0.010074   2.741 0.006206 **
Total_expenditure                0.042046   0.011471   3.666 0.000257 ***
Diphtheria                       0.037699   0.011987   3.145 0.001698 **
HIV_AIDS                        -0.441847   0.023487 -18.813  < 2e-16 ***
GDP                              0.036143   0.040481   0.893 0.372104
thinness_1_19_years              0.019731   0.028605   0.690 0.490460
thinness_5_9_years              -0.016892   0.029534  -0.572 0.567446
Income_composition_of_resources  0.182271   0.015410  11.828  < 2e-16 ***
Schooling                        0.355300   0.024881  14.280  < 2e-16 ***
ContinentAsia                    0.054109   0.006076   8.906  < 2e-16 ***
ContinentEurope                  0.086090   0.008025  10.727  < 2e-16 ***
ContinentNorth America           0.115307   0.007402  15.579  < 2e-16 ***
ContinentOceania                 0.042422   0.009131   4.646 3.73e-06 ***
ContinentSouth America           0.085396   0.008776   9.730  < 2e-16 ***
StatusDeveloping                -0.034812   0.007202  -4.834 1.50e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06311 on 1314 degrees of freedom
Multiple R-squared:  0.8813,    Adjusted R-squared:  0.8793
F-statistic: 424.3 on 23 and 1314 DF,  p-value: < 2.2e-16
```

Fig. 13. Model 2 Summary

Residual Analysis: The residual analysis for model 2 is shown in 14.The residual vs fitted values curve shows randomly distributed residuals with no pattern, implying uncorrelated errors and constant variance.

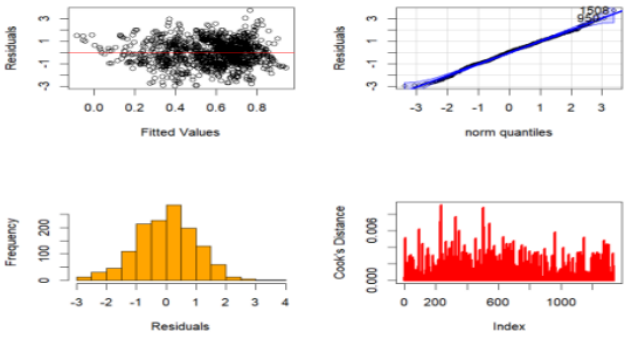

Fig. 14. Residual Analysis and Cook's Distance for Model 2

The QQ plot and histogram show that the residuals have an approximately normal distribution.

The cook's distance plot clearly shows that the outliers from Model 1 have been removed.

Multicollinearity: The VIF values for the predictors are shown in 15. We can see that all VIF values are small, implying that multicollinearity is no longer an issue.

C. Model 3: Reduced Model- Stepwise Regression

Model 2 has 21 predictors. To improve the explainability of the model we implemented variable selection approach

```
> car::vif(model2)
                                    GVIF Df GVIF^(1/(2*Df))
Year                            1.178104  1        1.085405
Adult_Mortality                 2.046944  1        1.430715
Alcohol                         2.868407  1        1.693637
percentage_expenditure          5.927073  1        2.434558
Hepatitis_B                     1.758063  1        1.325920
Measles                         1.563036  1        1.250214
BMI                             1.876454  1        1.369837
under_five_deaths               2.025955  1        1.423361
Polio                           1.728026  1        1.314544
Total_expenditure               1.269439  1        1.126694
Diphtheria                      2.106591  1        1.451410
HIV_AIDS                        1.702043  1        1.304624
GDP                             5.911290  1        2.431314
thinness_1_19_years             7.202999  1        2.683840
thinness_5_9_years              7.302090  1        2.702238
Income_composition_of_resources 2.951878  1        1.718103
Schooling                       3.737043  1        1.933143
Continent                       8.044876  5        1.231833
Status                          2.305856  1        1.518505
```

Fig. 15. VIF for Model 2

using backward stepwise regression was used to arrive at a reduced model (Model 3) consisting of the most important variables based on the Akaike Information Criterion (AIC). The summary of the model is shown 16.

```
> summary(backward_stepwise)

Call:
lm(formula = Life_expectancy ~ Year + Adult_Mortality + Alcohol +
    percentage_expenditure + BMI + under_five_deaths + Polio +
    Total_expenditure + Diphtheria + HIV_AIDS + Income_composition_of_resources +
    Schooling + Continent + Status, data = data_subset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18471 -0.04366  0.00305  0.04147  0.23117

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.236337   0.016265  14.530  < 2e-16 ***
Year                            -0.028736   0.006761  -4.250 2.29e-05 ***
Adult_Mortality                 -0.268714   0.015004 -17.909  < 2e-16 ***
Alcohol                         -0.102170   0.012986  -7.868 7.48e-15 ***
percentage_expenditure           0.198716   0.023516   8.450  < 2e-16 ***
BMI                              0.019803   0.008770   2.258 0.024109 *
under_five_deaths               -0.077014   0.024211  -3.181 0.001502 **
Polio                            0.028909   0.009927   2.912 0.003650 **
Total_expenditure                0.040143   0.011288   3.556 0.000389 ***
Diphtheria                       0.039312   0.010733   3.663 0.000259 ***
HIV_AIDS                        -0.441398   0.023354 -18.900  < 2e-16 ***
Income_composition_of_resources  0.182631   0.015208  12.009  < 2e-16 ***
Schooling                        0.355919   0.024779  14.364  < 2e-16 ***
ContinentAsia                    0.054878   0.005930   9.254  < 2e-16 ***
ContinentEurope                  0.085665   0.007878  10.873  < 2e-16 ***
ContinentNorth America           0.114670   0.007244  15.829  < 2e-16 ***
ContinentOceania                 0.040851   0.008739   4.674 3.25e-06 ***
ContinentSouth America           0.084710   0.008595   9.856  < 2e-16 ***
StatusDeveloping                -0.035100   0.007140  -4.916 9.96e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06304 on 1319 degrees of freedom
Multiple R-squared:  0.8811,    Adjusted R-squared:  0.8795
F-statistic: 543.3 on 18 and 1319 DF,  p-value: < 2.2e-16

> |
```

Fig. 16. Backward Stepwise Regression Model Summary

The summary shows that the adjusted R^2 has also increased slightly compared to Model 2.

Residual Analysis: The plots for residual analysis can be seen in 17. The residual vs fitted values curve shows randomly distributed residuals with no pattern, implying uncorrelated errors and constant variance.

The QQ plot and histogram show that the residuals have an approximately normal distribution.

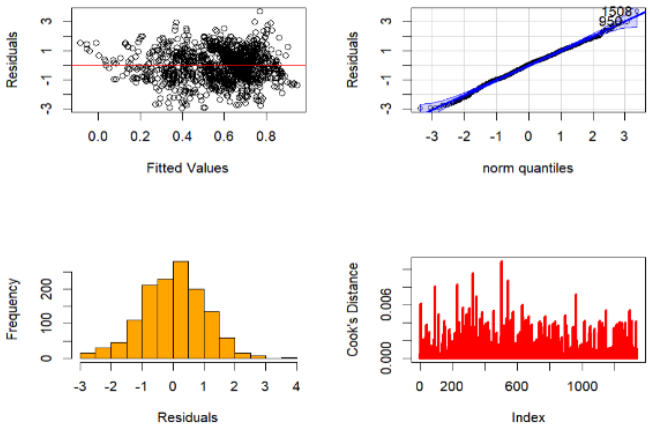

Fig. 17. Residual Analysis and Cook's Distance for Model 3

Cook's distances are smaller than the $\frac{4}{n}$ threshold so we do not have outliers.

D. Model 4: Reduced Model- Lasso Regression

We employed Lasso regression to perform variable selection. The relative importance of variables based on singular value of predictors is shown in 18. A new multiple linear regression model was built using these variables. 19 shows the summary of the model, which suggests that all the variables included are statistically significant. The F-value for the overall model is quite high (and corresponding p-value is low), which means that we can reject the null hypothesis that all variables are insignificant, i.e. the model, as a whole, is significant.

Variable	Importance
infant_deaths	2.961
HIV_AIDS	0.502
Schooling	0.374
Adult_Mortality	0.285
Income_composition_of_resources	0.159
percentage_expenditure	0.136

Fig. 18. Variable Selection with Lasso

Residual Analysis: The residual vs x plots are shown in 20. The residuals are NOT randomly distributed about the mean zero line for adult_mortality and HIV_AIDS, so the model does not satisfy the assumption of linearity for these predictors.

The residual vs fitted values plot shows a random pattern with no extreme values of residuals, so our error terms are uncorrelated with constant variance.

The QQ plot and histogram show that the residuals have an approximate normal distribution with the presence of outliers.

There are not many outliers in the data as we can see from the plot of cook's distance in 21.

VI. Model Performance Comparison

To compare the performance of models, we tabulated and compared four performance metrics- Adjusted R^2, Mean

```
> summary(red_model)

Call:
lm(formula = Life_expectancy ~ infant_deaths + HIV_AIDS + Schooling +
    Adult_Mortality + Income_composition_of_resources + percentage_expenditure,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33515 -0.04410  0.00136  0.05373  0.46952

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       0.27406    0.01376  19.923  < 2e-16 ***
infant_deaths                    -0.14480    0.02929  -4.944 8.57e-07 ***
HIV_AIDS                         -0.49542    0.02377 -20.844  < 2e-16 ***
Schooling                         0.43965    0.02590  16.972  < 2e-16 ***
Adult_Mortality                  -0.33308    0.01788 -18.630  < 2e-16 ***
Income_composition_of_resources   0.18444    0.01744  10.576  < 2e-16 ***
percentage_expenditure            0.20618    0.02768   7.448 1.65e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08605 on 1413 degrees of freedom
Multiple R-squared:  0.8013,    Adjusted R-squared:  0.8004
F-statistic: 949.7 on 6 and 1413 DF,  p-value: < 2.2e-16
```

Fig. 19. Lasso Regression Model Summary

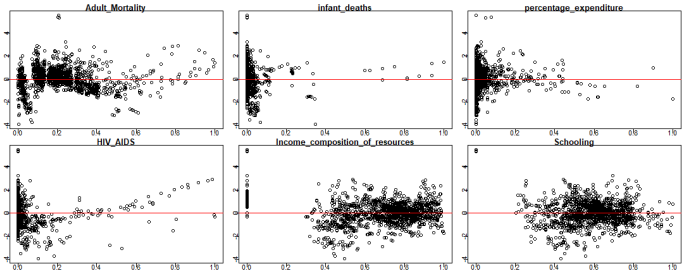

Fig. 20. Residual vs predictors plot for Model 4

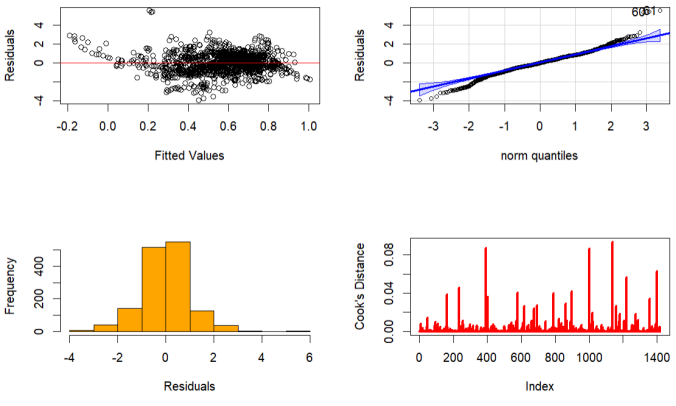

Fig. 21. Residual Analysis and Cook's Distance for Model 4

Squared Error (MSE), Mean Absolute Prediction Error (MAE), and Precision Error (PM). Model performance was estimated on a test data set comprising of 80% points from random sub-sampling. The results are shown in 22.

	Adjusted R^2	MSPE	MAE	PM
Model I	0.8504	0.00565	0.054641	0.14665
Model II	0.8793	0.00596	0.055251	0.15468
Best → Model III	0.8795	0.006	0.055435	0.15570
Model IV	0.8023	0.0074	0.060923	0.19433

Fig. 22. Model Performance Comparison

Clearly Model 3 outperforms the others based on all these metrics, meaning that this model is the best of the four alternatives.

VII. Conclusions

Based on the analysis described above, Model 3 is the best performing model among the four alternatives. We can conclude the following:

- 14 out of the 21 predictors affect life expectancy significantly.
- Life expectancy is significantly correlated with macroeconomic factors, health parameters as well as geography.
- Government expenditure on healthcare as well as education can significantly improve life expectancy.
- Viral outbreaks, like HIV-AIDS can negatively impact life expectancy.
- Developed countries have higher life expectancy probably due to improved quality of life and accessibility to healthcare resources.

VIII. Further Scope

While we could model the problem of life expectancy prediction with a reasonably accurate model, we acknowledge that our model is not perfect and that there is still scope for improvement.

- The data is limited in the sense that it does not capture all relevant health/macroeconomic factors; including more parameters and data points can improve the model's prediction accuracy
- Data is for the years 2000 - 2015 and hence needs to be collected for more recent years for the study to be relevant
- Multiple Linear Regression was used since the data set was reasonably small. For a larger data set, we could use models such as Random Forest and Gradient Boosting.

Acknowledgment

We would like to express our heartfelt gratitude to Professor Gamze Tokol-Goldsman, whose insightful lectures and commitment to fostering an active learning environment have contributed greatly to our understanding of regression and statistical analysis . A special thanks to our teaching assistant,

Veronica Lee for providing valuable assistance and constructive feedback throughout the course. Last but not the least, we appreciate the collaborative spirit, lively discussion and shared interest in the topic from our classmates, that greatly enhanced our learning experience.

REFERENCES

[1] Lakshmanarao, A. (2022). Life Expectancy Prediction through Analysis of Immunization and HDI Factors using Machine Learning Regression Algorithms. International Journal of Online & Biomedical Engineering, 18(13).

[2] Mathias, J. S., Agrawal, A., Feinglass, J., Cooper, A. J., Baker, D. W., & Choudhary, A. (2013). Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. Journal of the American Medical Informatics Association, 20(e1), e118-e124.

DATA SOURCE

Kumarrajsrshi. "Life Expectancy (WHO): Statistical Analysis on factors influencing Life Expectancy", Kaggle, 2017, https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/.

LIST OF FIGURES

APPENDIX

The R code for the statistical analysis and prediction of life expectancy can be found at the following link.

This report is part of the semester project required for completion of the course. ISyE 6414 (A): Statistical Modeling and Regression Analysis at Georgia Institute of Technology. The contributors are members of Team 2 of the Fall 2023 version of this course.