university of groningen / faculty of science and engineering / mathematics and applied mathematics

A Comparative Analysis of Regression Models for Global Health Prediction

Bachelor's Project Mathematics

June 2023

Student: L. Kalve

First supervisor: Prof. Dr. M.A. Grzegorczyk

Second assessor: Dr. W.P. Krijnen

Abstract

With the increasing availability and accessibility of data, it has become crucial to be able to appropriately interpret and analyze it. In this paper, we aimed to address this need by creating and comparing different regression models on a benchmark life expectancy data set. The models considered for comparison were linear regression, stepwise regression, and mixed effects models. To assess their performance and select the most suitable model, we evaluated them using criteria such as AIC, BIC, and cross-validation.

Upon analyzing the results, we found that the regression models and mixed effects models exhibited similar performance in terms of explanatory power, goodness of fit, and prediction accuracy. However, based on careful consideration and several important factors, we advocate for the preference of the mixed effects model for this benchmark data set as it is able to handle nested or hierarchical data structures better.

Contents

1 Introduction

In the contemporary data-driven landscape, statistical literacy and the ability to effectively interpret and analyze data have become indispensable skills. However, a substantial number of individuals lack a profound understanding of statistics, including the limitations of data collection and analysis. This knowledge gap often leads to misinterpretations of data and the propagation of misinformation. [14]

Statistical analysis, particularly regression models, is widely used to predict or explain the values of a given outcome variable with information from explanatory variables. The first type of regression analysis which was rigorously studied was linear regression. The initial conceptualization of linear regression can be traced back to Francis Galton in the late nineteenth-century England. [25]. Linear regression is one of the simplest models as we assume a linear dependency between the outcome variable and explanatory variables, and as such, this model is easier to fit than a model where the parameters are related non-linearly.

Linear regression models can be categorized into fixed effects models, random effects models, and mixed models, depending on whether the coefficients are fixed, random, or a combination of both. Random effects models were introduced in the early twentieth century by Ronald Fisher [11]. Random effects models are considered a special case of mixed models as they assume a fixed overall mean for the observations. [8]. Linear mixed effects models have increased in popularity in the last few decades since including random effects gives us several benefits as it allows us to model structured data with clusters of non-independent hierarchical observations. [12]

In multiple linear regression, the standard approach is to enter all predictor variables at once. An alternative, hierarchical approach is to add predictors in predetermined steps. Stepwise regression is a specific type of hierarchical regression where statistical algorithms determine the predictors included in the model. There are three variations of stepwise regression: forward selection, backward elimination, and stepwise selection. While it is not generally recommended to use stepwise regression as it has several limitations [14], it can be a useful technique for automatic variable selection, reducing model complexity, and exploring potential predictors. [10]

To enhance the effectiveness of regression models, evaluating their performance on benchmark data sets has become crucial. Benchmark data sets are standardized references that have gained wide acceptance within the research community. They provide a consistent and objective basis for analyzing the strengths and weaknesses of different regression models, offering insights into their predictive capabilities.

The assessment of regression models on benchmark data sets holds significant importance for several reasons. Firstly, it enables comparative analysis, facilitating the selection of the most suitable model for specific applications. By evaluating different models on a common benchmark, researchers and practitioners can objectively compare their performance and make informed decisions. Secondly, benchmark data sets enable a fair comparisons across studies, ensuring consistency and reliability in evaluating model performance. This promotes knowledge building and allows for advancements in the field. Moreover, the evaluation process uncovers the factors that influence the performance of regression models, shedding light on the strengths and weaknesses of different modeling approaches.

The World Health Organisations data set containing information about life expectancy will be used in this paper as the benchmark data set. Various models with different regression techniques will be created, and the performances of the models will be compared by means of AIC, BIC, and cross-validation. The results of the models and the implications of them

will be briefly discussed.

This paper holds potential implications and benefits for various domains. By evaluating the performances of different regression models on a benchmark data set, the study can enhance the accuracy and reliability of regression modeling in practical applications, such as in healthcare. Furthermore, the research contributes to the existing knowledge in the field by providing insights into the performance and suitability of different regression models. The findings can guide future studies and inspire the development of enhanced models that address the challenges encountered in real-world data analysis and prediction tasks. Ultimately, this paper aims to advance the understanding of linear regression models' performance on benchmark data set and foster improvements in regression modeling practices.

2 Preliminary Theory

2.1 Linear Regression

The linear regression model function is used to statistically model the relationship between a response variable y and k explanatory variables $x_1, \ldots, x_k$. For observation i the linear regression model can be written in the following form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}. \tag{1}$$

Linear regression is one of the oldest and most basic models in statistics, and as the results are relatively easy to interpret, it is very commonly used [7]. However, the linear model can only give us an approximation of the true relationship between x and y and as such, we also need to include the error term in order to get more meaningful results. The model is the following:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \tag{2}$$

where $\epsilon_i = y_i - \bar{Y}_i$ is the error of prediction which represents the uncertainty in predicting the outcome variable y with the explanatory variable x. We can also represent the model in the following matrix form:

$$y = X\beta + \epsilon \tag{3}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{4}$$

The regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ are in general estimated by the least squares method, yielding the following estimator:

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{5}$$

Assumptions of Linear Regression Models

The simple linear regression model where only one explanatory variable is considered rests on several assumptions which determine how well it operates. The main assumptions for SLR are the following:

1. *Independence.* We assume that the the errors of prediction are statistically independent. This means that we assume the observations to be independent, and that condition is often satisfied by using random sampling.

2. *Constant Variance (homoscedastity).* The variance of errors are assumed to be constant over the distribution of X.

3. *Normality.* We assume that the errors are random variables and that they are normally distributed. Moreover, we assume that the mean is zero which is important for computing the intercept: $\epsilon_i \sim N(0, \sigma^2)$

4. *Linearity.* We assume that there is a linear relationship between Y and X.

The assumptions for multiple linear regression models are the same as for the simple models, with an added assumption of *collinearity* which means that we assume that there is no combination of the explanatory variables X_i which have a perfect association. [14]

Simple Linear Regression

The simple linear regression model is the following:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{6}$$

where y is the response variable (life expectancy), x is the explanatory variable, β_0 is the intercept parameter, β_1 is the slope parameter, and ϵ is the error term [7]. For $j = 1, \ldots, k+1$ we have the following hypothesis:

$$H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0$$

The slope and intercept coefficients are calculated using the *ordinary least squares* (OLS) method, which is the most common method for fitting the regression line [7]. OLS aims to minimize the sum of squared errors (SSE), which is obtained the following way:

$$SSE = \sum \left(y_i - \hat{y_i}\right)^2 = \sum \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2. \tag{7}$$

Since we have that $\hat{\epsilon_i} = y_i - \hat{y_i}$, we can also express SSE as the sum of squared residuals:

$$SSE = \sum \hat{\epsilon_i}^2. \tag{8}$$

When the assumptions of simple linear regression are satisfied, the least squares equation for the estimated slope coefficient has been shown [7] to be optimal for minimizing SSE:

$$\hat{\beta_1} = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum \left(x_i - \bar{x}\right)^2}, \tag{9}$$

and the standard error of the slope coefficient in a simple linear regression model is given as:

$$se\left(\hat{\beta_1}\right) = \sqrt{\frac{\sum \left(y_i - \hat{y_i}\right)/n-2}{\sum \left(x_i - \bar{x}\right)^2}} = \sqrt{\frac{SSE/n-2}{SS(x)}}, \tag{10}$$

where n is the sample size, and $SS(x)$ is the sum of squares of x. A larger SSE results in a larger standard error as the sum of squared errors indicate the variation of the linear regression. Once the standard error of the slope coefficient has been found we can use it to calculate the t-value:

$$t - value = \frac{\hat{\beta_1}}{se(\hat{\beta_1})}. \tag{11}$$

We find the intercept using the slope:

$$\hat{\beta_0} = \bar{y} - \left(\hat{\beta_1}\bar{x}\right). \tag{12}$$

The coefficients for linear regression in this thesis are computed by R, however, as computing these equations can be slow for larger data sets, programs rather employ matrix routines to speed up the computing process. [14]

Multiple Linear Regression

Multiple linear regression can be regarded as an extension of simple linear regression as it uses several explanatory variables, instead of one, to predict the response variable. Multiple linear regression model also aims to determine which explanatory variable is strongest associated with the response variable. The standard errors for multiple linear regression models are the following:

$$se\left(\hat{\beta_i}\right) = \sqrt{\frac{\sum \left(y_i - \hat{y_i}\right)^2}{\sum \left(x_i - \bar{x}\right)^2 \left(1 - R_i^2\right)\left(n - k - 1\right)}}, \tag{13}$$

where $(1 - R_i^2)$ is the tolerance, and the R^2 is called an auxiliary regression model, where x_i is taken to be the outcome variable while the other explanatory variables are taken as predictor variables [14]. Alternatively, the standard errors can be found by taking the square roots of the diagonal elements of the following matrix:

$$V = \left(X^T X\right)^{-1} \hat{\sigma}^2, \text{ where } \hat{\sigma}^2 = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n-k} \tag{14}$$

2.2 Stepwise Regression

Stepwise regression is a method for iteratively adding and/or removing variables based on a specified criterion, such as p-value or information criterion of the model, resulting in a simpler model which is easier to interpret. While stepwise regression can successfully automate the variable selection procedure, it can also have its limitations as the method can be sensitive to specific data sets and the chosen criterion which might lead to problems such as overfitting, or biased coefficient estimates. [10]

Backward Selection

Backward selection is a variable selection method where we begin with a full model, that is, a model which includes all explanatory variables, and selectively remove the variables which have statistically insignificant slope coefficients. In the case of backward selection based on p-values, the explanatory variable with the highest p-value is removed and the model gets re-fitted without the removed variable until all variables left in the model are statistically significant, or we reach a pre-defined threshold. For backward selection based on AIC we start with the full model and compute the AIC score for it. Then, we reduce the model by removing one variable, and calculate the AIC score for the reduced model. If it is less than the AIC score for the full model, we replace the current model with the new reduced one, and keep iteratively removing variables until we get the final model. Using the backward selection is a reasonable method for finding the best predictive model while reducing the risk of overfitting. Backward selection may however increase the risk of false negatives as the method might fail to identify variables which have meaningful effects but do not reach the predetermined significance levels. Backward selection also assumes that the full model is appropriate, and if that is not the case the backward selection method can lead to a biased model selection. [14]

Forward Selection

Forward selection is a variable selection method which starts with a model which contains no explanatory variables and gradually adds variables which are determined to be statistically significant for explaining the response variable by fitting separate models for each explanatory variable, and choosing the one which provides the best fit based on either p-values or AIC score. The model is re-fitted after adding each explanatory variable. The stopping criterion is typically a pre-determined significance level or a desired model complexity. As a relatively straightforward variable selection method, forward selection can often provide a clear understanding of how each variable affects the model. However, as variables are added sequentially, we might end up with a suboptimal combination of explanatory variables as some variables might contribute to the model's performance collectively, but be insignificant individually. As the model adds variables constantly, it is biased towards larger models which can create the problem of overfitting. [14]

2.3 Mixed Models

The mixed model is a linear model which contains both fixed effects and random effects. They are designed to separate the variability which is due to differences across individual

units from the variability due to differences across groups. The mixed model can be written as

$$y = X\beta + Z\gamma + \epsilon, \tag{15}$$

where y is a known vector, with $E(y) = X\beta$, β is a vector of fixed effects, γ is a vector of random effects with $E(\gamma) = 0$, $Cov(\gamma) = D$, and $Cov(\gamma, \epsilon) = 0$. X and Z are known matrices relating the observations to β and γ respectively, and ϵ represents the vector of random errors, with $E(\epsilon) = 0$ and $Cov(\epsilon) = R$. Typically, D and R depend on different subsets of some vector parameters. [8]

To improve the computational stability of the model, we can reformulate it by defining a random effects variable U, which has the following distribution:

$$U \sim N(0, \sigma^2 I_q). \tag{16}$$

Given the random effects, the conditional distribution of the response variable is given as

$$(Y|U = u) \sim N(\mu_{Y|U=u}, \sigma^2 W^{-1}), \tag{17}$$

where W is a diagonal matrix of known prior weights, σ is the scale parameter, and

$$\mu_{Y|U=u} = X\beta + Z\Lambda_\theta u + \epsilon \tag{18}$$

where Λ_θ is a singular relative covariance factor depending on the covariance-component parameter vector θ. In order to maximize the likelihood fitting of the model, we repeatedly apply the penalized least-squares method. The goal is to minimize the penalized weighted residual sum of squares, which is given by

$$r^2(\theta, \beta, u) = \rho^2(\theta, \beta, u) + ||u||^2, \tag{19}$$

where

$$\rho^2(\theta, \beta, u) = ||W^{1/2}[y_i - \mu_{Y|U=u}]||^2. \tag{20}$$

We minimize the penalized weighted residual sum of squares over $[u\beta]^T$. It has been shown [5] that we can rewrite the equation in the following way:

$$r^2(\theta, \beta, u) = r^2(\theta) + ||L_\theta^T(u - \mu_{U|Y=y_i}) + R_{ZX}(\beta - \hat{\beta}_\theta)||^2 + ||R_X(\beta - \hat{\beta}_\theta)||^2, \tag{21}$$

where $r^2(\theta)$ is used to replace $r^2(\theta, \hat{\beta}_\theta, \mu_{U|Y=y_i}$, and $\mu_{U|Y=y}$ is the conditional mean of U given $Y = y_i$. The derivations of the matrices L_θ, R_{ZX}, and R_X can be found in [5]. This is an important expression as it relates $r^2(\theta, \beta, u)$ with the minimum value $r^2(\theta)$, which is useful for integration over the random effects to estimate the maximum likelihood function [5]. Moreover, this expression is useful in the theory underlying the **lme4** package in R, which is used for fitting linear mixed-effects models.

We can express the log-likelihood which is to be maximized as

$$\mathcal{L}(\theta, \beta, \sigma^2|y_i) = log f_Y(y_i), \tag{22}$$

where

$$f_Y(y_i) = \frac{|W|^{1/2}|L_\theta|^{-1}}{(2\pi\sigma^2)^{n/2}} \exp\left[\frac{-r^2(\theta) - ||R_X(\beta - \hat{\beta}_\theta)||^2}{2\sigma^2}\right]. \tag{23}$$

The maximum likelihood (ML) criterion is given by

$$-2\mathcal{L}(\theta|y_i) = \log\frac{|L_\theta|^2}{|W|} + n\left(1 + \log\left(\frac{2\pi r^2(\theta)}{n}\right)\right). \tag{24}$$

While the number of columns q in Z and the size of Σ_θ can be large, the expression depends only on θ, which has a small dimension of frequently less than 10. [5]

2.3.1 REML

The restricted maximum likelihood (REML) criterion is a method for estimating the covariance parameters in a linear mixed effects model while taking into account the fixed effects. It provides a way to separate the estimation of the fixed effects from the estimation of the random effects, which can lead to more accurate parameter estimates. The REML criterion can be expressed in the following way:

$$\int f_Y(y_i)d\beta = \frac{|W|^{1/2}|L_\theta|^{-1}}{(2\pi\sigma^2)^{n/2}}\exp\left(\frac{-r^2(\theta)}{2\sigma^2}\right)\int \exp\left(\frac{-||R_X(\beta-\hat{\beta}_\theta)||^2}{2\sigma^2}\right)d\beta. \tag{25}$$

We can use change of variables and the fact that the Jacobian determinant of the transformation from β to $v = R_X(\beta - \hat{\beta}_\theta)$ is $|R_X|$ to simplify the integral to the following form (see [5] for more details):

$$\int f_Y(y_i)d\beta = \frac{|W|^{1/2}|L_\theta|^{-1}|R_X|^{-1}}{(2\pi\sigma^2)^{(n-p)/2}}\exp\left(\frac{-r^2(\theta)}{2\sigma^2}\right). \tag{26}$$

The unprofiled REML criterion involves maximizing the likelihood of the observed data with respect to the random effects and the error term. This approach considers all the random effects in the model simultaneously and estimates their covariance structure, and is obtained by minus twice the log of the integral above:

$$-2\mathcal{L}_R(\theta,\sigma^2|y_i) = \log\frac{|L_\theta|^2|R_X|^2}{|W|} + (n-p)\log(2\pi\sigma^2) + \frac{r^2(\theta}{\sigma^2}. \tag{27}$$

The REML criterion cannot be used to estimate β as it gets integrated out and as such, we rely on the maximum likelihood estimate $\hat{\beta}_{\hat{\theta}}$ at $\theta = \hat{\theta}$. In order to find the profiled REML criterion, which involves maximizing the likelihood of the observed data with respect to the random effects while profiling out the fixed effects, consider the REML estimate of σ^2:

$$\hat{\sigma}^2_\theta = \frac{r^2(\theta)}{n-p}, \tag{28}$$

From this we can find the profiled REML criterion:

$$-2\mathcal{L}_R(\theta|y_i) = \log\frac{|L_\theta|^2|R_X|^2}{|W|} + (n-p)\left(1+\log\left(\frac{2\pi r^2(\theta)}{n-p}\right)\right) \tag{29}$$

In the profiled REML the fixed effects are treated as known or fixed, and their estimated values are used to condition the likelihood function. The profiled REML criterion focuses on estimating the covariance structure of the random effects while removing the influence of the fixed effects. [5]

2.4 Model selection

2.4.1 Information Criterion

Information criterion are designed as a means for model selection. Two of the most common information criteria are Akaike Information Criterion (AIC), formulated by the Japanese statistician Hirotsugu Akaike in 1974, and Bayesian Informarion Criterion (BIC), developed by Gideon E. Schwarz in 1978. The AIC and BIC are defined as follows:

$$AIC = 2k - 2\ln(\hat{L}) \tag{30}$$

$$BIC = k\ln(n) - 2\ln(\hat{L}) \tag{31}$$

where k is the number of estimated parameters in the model, n is the sample size, $\hat{L}$ is the maximized value of the likelihood function of model M: $\hat{L} = p(x|\hat{\theta}, M)$, where $\hat{\theta}$ are the

parameter values which maximize the likelihood function, and x is the observed data. [17]

The AIC estimates the prediction error in terms of MLE. It estimates the amount of information lost by a given model, and the aim is to balance the goodness of fit of a model with the number of parameters used. As such, AIC deals with the risk of overfitting and underfitting. In general, the highest quality models are those which lose the least information, and the models with lower AIC scores are preferred. [17]

The BIC is closely related to AIC as both penalise the number of parameters used in a model, however, for sample sizes greater than 7, the penalty term in BIC is larger than in AIC meaning that it prefers simpler models and aims to avoid overfitting. The models with lower BIC scores are preferred over models with a higher BIC score.

2.4.2 Cross-Validation

Cross-validation is a resampling technique used to evaluate the performance of a model. It is primarily used to assess how well the model generalizes to new, unseen data. The basic idea behind cross-validation is to partition the data set into multiple subsets or "folds." The model is trained on a subset of the data called the training set and then evaluated on the remaining subset, which is called the validation set. This process is then repeated several times, with different subsets of the data used for training and validation as using only one testing set can give us varied results depending on how the data set was partitioned. The performance metrics, such as accuracy or mean squared error, are averaged across all iterations to obtain an overall assessment of the model's performance [22]. There are several metrics which are used to evaluate the accuracy of a given model. The statistical metrics are the following:

1. Root Mean Squared Error (RMSE) is the square root of the averaged squared difference between the actual and predicted target variable value. The formula to compute the RMSE is the following:
$$RMSE = \sqrt{\frac{\sum_i^n}{(} y_i - \hat{y}_i)^2 n}. \tag{32}$$
As RMSE gives us the average prediction error, the model with the lowest RMSE score is preferred.

2. Mean Absolute Error (MAE) gives us the absolute difference between the actual values and the predicted values. It is calculated as follows:
$$MAE = \frac{1}{n}\sum_i^n |y_i - \hat{y}_i|. \tag{33}$$
We can use MAE to evaluate the performance of a model, where we prefer the model with the lowest MAE score.

3. R^2 Error represents the proportion of the variance in the dependent variable that can be explained by the independent variable(s). The formula for calculating the R^2 error is the following:
$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \frac{1}{n}\sum_i^n y_i)^2} \tag{34}$$
The R^2 value tells us how well the model fits the data points on a scale of $0\% - 100\%$, meaning that the higher the R^2 score, the better the model is at explaining the variability in the dependent variable.

3 Methodology

3.1 Data Description

The data set used for this thesis is the Life Expectancy (WHO) data set, which contains life expectancy[19], health, immunization, and economic and demographic information about 179 countries over the span of 16 years, from 2000-2015. The data set has 21 variables and 2864 rows[18]. From figure 1 we can see the average life expectancy across 9 regions.

The data set originally had some missing values and inaccurate data. The updated data about the population, GDP, and Life Expectancy was updated according to World Bank data. The data about vaccinations for Measles, Hepatitis B, Polio, and Diphtheria was collected from the public data sets of the World Health Organization. Information about alcohol consumption, BMI, HIV incidents, mortality rates, and thinness for children aged 5-9 and 10-19 was also collected from the WHO public data sets. Data about schooling was gathered from the Our World in Data, which is a University of Oxford project.[18]

For the missing values in the data set, a few strategies were applied. If a country was missing a value in any year, the gap was filled with the closest average over a three-year period. If, however, a country had missing values for all the years, then the data was filled using the average of the region. If a country was missing more than 4 data columns, which was the case for countries such as Sudan, South Sudan, and North Korea, then the country was omitted from the database.[18]

The values in the data set represent the following:

- Life Expectancy is defined as the average number a years a newborn could expect to live for a specific year and in a given country. It is based on sex- and age specific death rates and is derived from life tables.[1]
- Infant deaths are calculated per 1000 live births, where an infant is considered between birth and 11 months.[16]
- The values for under five deaths represents the probability of dying by age 5 per 1000 deaths.[26]
- Adult mortality - probability of dying between 15 and 60 years per 1000 population.[2]

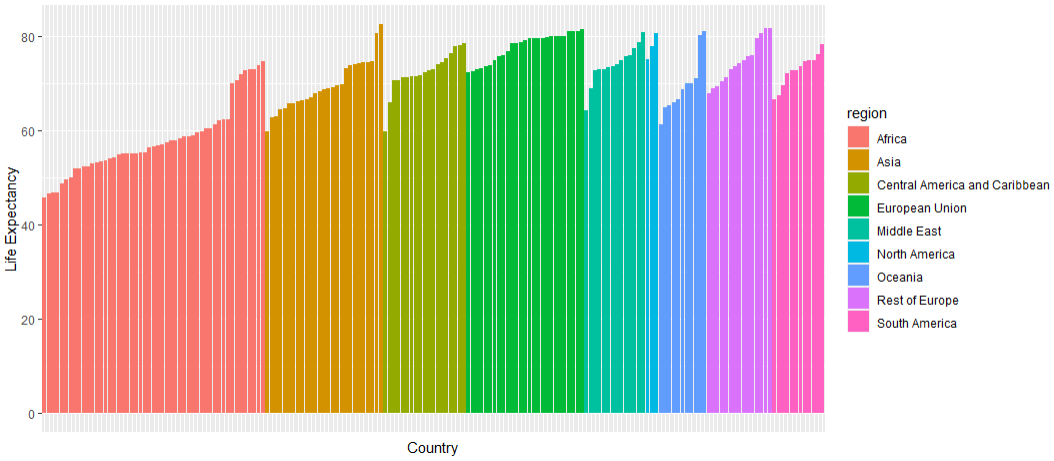

Figure 1: The average life expectancy

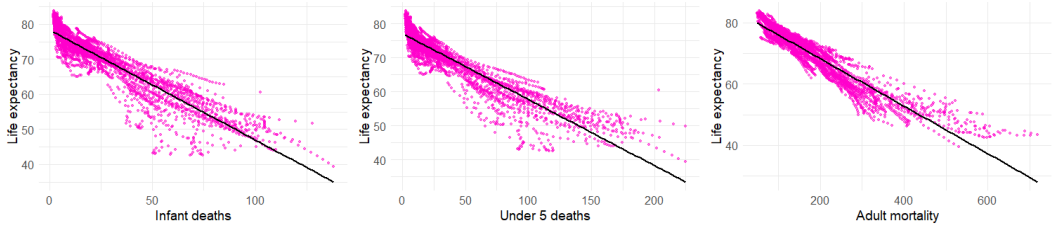

Figure 2: Mortality rates for three age groups: infants, under five years old, and adults

- Alcohol consumption is recorded in liters of pure alcohol per capita (15+) consumption.[3]
- Hepatitis B [13], measles [21], polio [23], and diphtheria [9] - immunization coverage among 1-year olds.
- BMI shows the mean body mass index (kg/m^2) age-standardized estimate.[20]
- HIV - incidents of HIV per 1000 population, aged 15 to 49.[15]
- Thinness shows the percentage of defined population with a BMI < 2 standard deviations below the median for age groups 5-9 and 10-19.[24]
- Schooling shows the average number of years people over the age of 25 participated in formal education.[4]
- Since each country defines itself as developed or developing, categorising the countries can be challenging. This categorisation of the countries has been done according to a 2014 list by UN, where countries are classified as developed, developing, or in transition for analytical purposes. The countries whose economic status is in transition have similar characteristics to the developed, or developing countries. Countries have been divided into four income groups according to their gross national income per capita: high-income, higher-middle-income, lower-middle-income, and low-income. In order to ensure comparability, the levels of gross domestic income are set by the world bank. [6]

Plotting the data

In order to get a better understanding of the data set, we begin by plotting how a single explanatory variable[1] globally affects the life expectation as plotting the data early and often is seen as a good statistical practice[14]. The plots for simple linear regression were generated using RStudio package ggplot2. From figure 2 we can see that for each explanatory variable, the relationship is indeed linear and therefore it makes sense to use this model. Furthermore, we get the expected results that a higher mortality rate over various age groups lowers the average life expectancy.

The plots for Hepatitis-B, Polio, and Diphtheria are all similar as we can see from figure 3. This can be explained by the fact that vaccinations for all these diseases were given to newborns at the same time frame. From the plots we can see that a higher vaccination rate for the three diseases results in a higher life expectancy, however, the linear dependency between vaccinations and life expectancy is not immediately clear from the figure.

[1] country, region, year, infant deaths, under five deaths, adult mortality, alcohol consumption, hepatitis B, measles, BMI, polio, diphtheria, HIV, GDP per capita, population size, thinness, schooling, and economy status

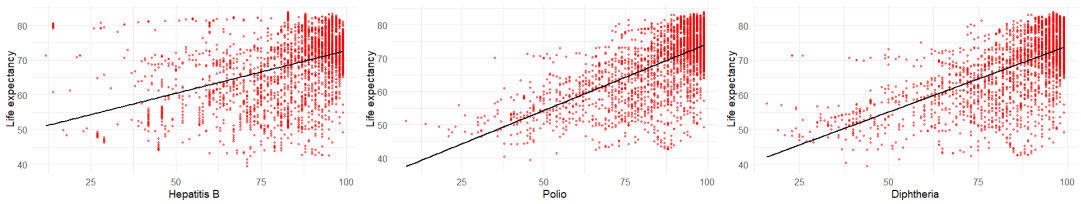

Figure 3: Vaccination for hepatitis B, polio, and diphtheria

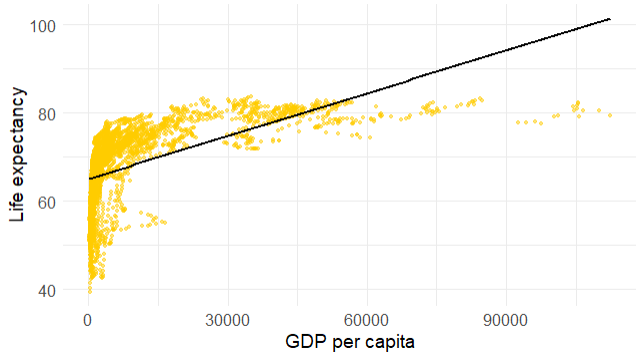

Figure 4: GDP per Capita

From figure 4 we can see that GDP (gross domestic product) per capita and life expectancy do not seem to have a linear dependency, and as such, the fourth assumption of linearity potentially fails. In this case, a different model should be considered in order to get more accurate approximations of the relation between GDP and life expectancy.

It seems from figure 5 that both alcohol consumption and a higher BMI (body mass index) positively affect the life expectancy. That is, according to a simple linear regression, the more alcohol one consumes and the higher one's BMI, the longer they live. The normal range for the BMI is considered to be between $18.5 - 24.9$, BMI of $25 - 29.9$ is considered overweight, and individuals with a BMI of $30 - 34.9$ are considered obese. On the plot representing the BMI values we also notice groups or clusters of data points forming. Those represent the BMI values for a country over the time period when the data was collected (2000-2015).

With the data set, various models are created and compared. All the models will be compared by their information criteria scores, and by the means of cross-validation. We assume that the assumptions for linear regression are met.

Linear Regression

The first model we consider is the linear regression model, with **life expectancy** as the response variable. The covariates used can be found in Table 1. Two of the variables are categorical, country and region, and 17 are numerical. For numerical variables we can fit the multiple linear regression model in a straightforward way, however, for categorical variables we need a different approach. To include the variables **country** and **region** in the model in a meaningful way, we need to transform them into factors. This is done in RStudio using the following command:

```
dataset$Country <- as.factor(dataset$Country)
```

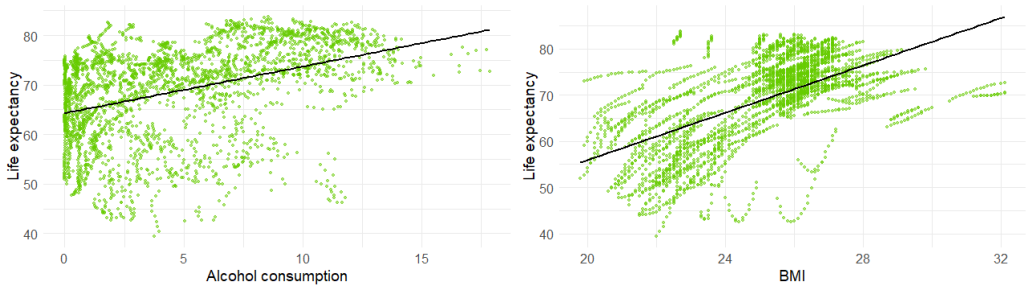

Figure 5: Alcohol consumption and BMI

```
dataset$Region <- as.factor(dataset$Region)
```

This R code assigns a unique integer value to each distinct category in the variable, which allows us to treat the variable as a discrete categorical variable rather than a continuous numeric variable. We call each category of a variable a level, and as we have 179 countries, there are 179 levels where each level represents a country. For the variable region, we have 9 levels. Now that the categorical values have been factored, we can use the following function to create the linear regression model:

```
lm(response_variable ~ covariate1 + covariate2 + ..., data = dataset)
```

Where the variables are as discussed in the beginning of this section.

Variable Selection

In the model discussed in the previous section we included all the covariates, however, if we are interested in which variables affect life expectancy, we need to perform variable selection. This has been done in three ways: via backward selection, forward selection, and both. The

Variable	Class	Min	Max
Country	character	NA	NA
Region	character	NA	NA
Year	numeric	2000	2015
Infant Deaths	numeric	1.8	138.1
Under five Deaths	numeric	2.3	224.9
Adult Mortality	numeric	49.4	719.4
Alcohol Consumption	numeric	0	17.9
Hepatitis B	numeric	12	99
Measles	numeric	10	99
BMI	numeric	19.8	32.1
Polio	numeric	8	99
Diphtheria	numeric	16	99
Incidents of HIV	numeric	0	21.9
GDP per Capita	numeric	148	112418
Population	numeric	0.1	1379.9
Thinness 5-9 year olds	numeric	0.1	28.6
Thinness 10-19 year olds	numeric	0.1	27.7
Schooling	numeric	1.1	14.1
Economy Status	numeric	0	1

Table 1: Variables used in the models and their class, minimum value, and maximum value

function used to compute the stepwise regression model was part of the `olsrr` package in RStudio. The functions used for backward selection were the following:

```
ols_step_backward_p(model) ols_step_backward_aic(model)
```

The functions used for forward selection were:

```
ols_step_forward_p(model) ols_step_forward_aic(model)
```

And the functions which include both backward and forward selection were the following:

```
ols_step_both_p(model) ols_step_both_aic(model)
```

These functions use different methods to choose which variables are significant for predicting life expectancy. The argument *model* corresponds with the linear regression model, which was discussed in the previous section.

Mixed Effects Model

In the mixed effects model we include the numerical variables as fixed, and the categorical variables as random. Three mixed effects models were created using the following function in R:

```
MixedModel <- lme(response_variable ~ covariate1 + covariate2 + ...  , data
        = dataset, random=~1|RandomEff1/RandomEff2, method="REML")
```

Where the response variable is life expectancy, the covariates are the fixed effects and random effects are defined separately. We use the restricted maximum likelihood (REML) method. With this function, three mixed effects models were created: the first had only country as random effect, the second model had only region as random effect, and the third model included both country and region as random effects. The fixed effects remained the same for all three cases.

3.2 Comparison

The models will first be compared based on information criteria. We will be comparing both the AIC and BIC scores for each of the models presented, namely, for the linear regression model, the models obtained via stepwise regression, and the mixed effects models. The AIC and BIC scores of the mixed effects models will be compared independently of the other regression models. This is because the AIC and BIC are calculated based on the maximum likelihood function, and since the likelihood functions differ, the information criteria values are not comparable.

We will also assess the performances of the models by means of cross-validation. More specifically, the leave-one-out cross validation technique will be used. The R code used for this can be found in Appendix B.

Variable	Simple LR	Multiple LR
Year	**0.356 (0.038)**	**0.143 (0.007)**
Infant Deaths	**-0.314 (0.003)**	-0.009 (0.007)
Under five Deaths	**-0.194 (0.002)**	**-0.044 (0.004)**
Adult Mortality	**-0.077 (0.000)**	**-0.042 (0.001)**
Alcohol Consumption	**0.943 (0.405)**	**-0.027 (0.012)**
Hepatitis B	**0.24568 (0.010)**	0.002 (0.000)
Measles	**0.247 (0.008)**	0.002 (0.001)
BMI	**2.566 (0.064)**	**-0.415 (0.064)**
Polio	**0.340 (0.009)**	0.001 (0.003)
Diphtheria	**0.380 (0.009)**	**0.009 (0.003)**
Incidents of HIV	**-2.184 (0.615)**	**0.155 (0.025)**
GDP per Capita	**0.000 (0.000)**	**0.000 (0.000)**
Population	0.002 (0.001)	0.002 (0.001)
Thinness among 5-9 year olds	**-0.952 (0.035)**	-0.013 (0.007)
Thinness among 10-19 year olds	**-0.991 (0.035)**	**-0.014 (0.007)**
Schooling	**2.172 (0.038)**	**-0.098 (0.029)**
Economy Status	**12.164 (0.370)**	**8.232 (0.725)**

Table 2: Comparison of SLR and MLR variable coefficients, statistically significant variables are in bold

4 Results

4.1 Model Performance

The results for the multiple linear regression can be found in appendix A, table 4. Note that the variables **country** and **region** were included in the model, but have been left out of the results table due to the large size of it. We can see the difference in the coefficients between simple and multiple linear regression in table 2. When including all the explanatory variables in the model, the coefficients change drastically, and several variables which were significant in the simple model are no longer significant. In general, the multiple linear regression model is preferred over simple linear regression as it gives a better overview of the relation the explanatory variables have with the response variable. The AIC score for the linear regression model was 4184.9, and the BIC score was 5353.1. The leave-one-out cross-validation scores for the linear regression model were the following: R^2 was 0.979, RMSE was 1.362, and MAE was 1.081.

The next models were created using the backward selection method, the results for the backward selection based on p-values are shown in Table 6 and the results for the backward selection based on AIC can be found in Table 5. The AIC score for the backward selection model based on p-values was 4199.1 and the BIC score was 5337.3. The R^2 score was 0.979, RMSE score was 1.361, and MAE score was 1.078. The backward selection based on AIC scores performed slightly better: the AIC score for that model was 4180.7 and the BIC score was 5331.0. The R^2 score was 0.997, RMSE score was 0.508, and MAE score was 0.340 which indicates that backward selection based on AIC would be preferred to backward selection based on p-values.

The results for the forward selection model based on p-values are shown in Table 8, and the results for the forward selection based on AIC can be found in Table 7. The AIC score of the model based on p-values was 4199.0, and the BIC score was 5350.1. The R^2 score was 0.997, the RMSE was 0.514, and the MAE was 0.343. The results for the model based on AIC were the following: the AIC score was 4181.0, the BIC was 5331.3, and the R^2 score was 0.997, the RMSE was 0.510, and the MAE was 0.341. While there is a significant difference in the

	Model	df	AIC	BIC
rndeffc	1	20	5271.228	5390.301
rndeffr	2	20	9387.785	9506.859
rndeffcr	3	21	5273.228	5398.255
model.fixed	4	19	10068.486	10181.606

Table 3: Mixed Effects Models Scores

information criteria scores of these two models, the results from cross-validation indicate that the two models have a similar error in predicting the data.

In the last stepwise regression model which includes both forward and backward selection, we have omitted the categorical variables from the selection based on p-values as the function cannot distinguish between the individual effects of two or more predictor variables if they are highly correlated. The results for the stepwise regression based on p-values can be found in Table 10 and the results for the regression based on AIC are shown in Table 9. Without the categorical variables, we get much higher AIC score of 9888.6 and BIC score of 10012.4. The R^2 score for the model was 0.979, RMSE was 1.374, and MAE was 1.095. For the stepwise regression based on AIC scores we include all the explanatory variables (table 1), and the resulting model was the same as the one obtained via backward selection based on AIC. As such the results for the two models were identical: the AIC score was 4180.7, the BIC score was 5331.0, R^2 was 0.997, RMSE was 0.508, and MAE was 0.340.

The last model considered was the mixed effects model. We tested three models, model a with only **country** as a random effect, model b with only **region**, and model c with both included. We compared these models against the linear regression model which contained the same fixed effects as the three mixed models (model d). The summaries of the information criteria scores for these models can be found in Table 3. The summary of model a can be found in Table 11, of model b in Table 12, and of model c in Table 13. The results from the cross-validation for model a were the following: R^2 was 0.997, RMSE was 0.515, and MAE was 0.344. model b had a lower R^2 value at 0.984, and a higher RMSE and MAE values, which were 1.202 and 0.954 respectively. The third model, which included both **country** and **region** as random variables had an R^2 value of 0.997, RMSE was 0.513, and MAE was 0.342. This means that as models 1 and 3 gave more accurate predictions and should thus be preferred over model b, while models 1 and 3 produced similar results.

4.2 Interpretation of the Results

When analysing the relationship between a response variable and a single explanatory variable, we found that all but one explanatory variables have a significant effect on the response variable, **life expectancy**. The only coefficient which had a p-value of greater than 0.05 was **population**. However, interpreting these results requires caution and consideration of additional factors as the observed association between the explanatory variable and the response variable could be influenced by confounding variables. For instance, access to healthcare might have an impact on both the explanatory variable (such as BMI or incidents of HIV) and the response variable life expectancy.

Most of the coefficients obtained from the multiple linear regression differ from the simple linear regression results. This is due to the fact that multiple linear regression accounts for the influence of the other variables as well while the simple linear regression does not. Moreover, some of the variables have a high correlation (for example, thinness among 5-9 year olds and thinness among 10-19 year olds), which means that when we include both of them in the model, only one will appear as significant while an individual analysis of such variables would give us the result that both are significant.

All the models we obtained via the stepwise selection method based on AIC outperformed the models which were based on the p-values. There are several reasons why this might be the case. One explanation could be that since we are dealing with multicollinearity, p-values can be unreliable for variable selection. Stepwise regression with AIC considers the overall improvement in model fit when deciding which variables to include or exclude. It can help identify the most important predictors while accounting for their interrelationships, leading to models which are more robust and better interpretable.

The mixed effects model is the most appropriate when our data includes groupings such as repeated measurements on the same subjects or observations clustered within certain categories, such as by country or region. By accounting for the clustering structure, we can estimate the effects of the categorical variables on the outcome variable and thus reduce potential bias and provide more accurate and reliable results. The mixed effects model with both **country** and **region** performed as well as the model which only included **country** as a random effect, while both of these models were significantly better than mixed effects model with **region** as random effect. Therefore we can conclude that the variability explained by the random effect **region** does not contribute significantly to the model's performance beyond the random effect of **country**. The higher AIC and BIC, MAE and RMSE and lower R^2 for the model with region as a random effect indicate that the additional complexity of including region does not justify the improvement in model fit.

5 Discussion and Conclusion

5.1 Discussion

We have found that when evaluating the regression models, there were six linear regression models with similar R-squared, root mean squared error, and mean absolute error metrics. The backward selection based on AIC, forward selection based on p-values and on AIC, stepwise regression based on AIC, mixed effects model with country as random effect, and the mixed effects model with country and region as random effects. This shows that using hierarchical regression produces models with better predictive qualities.

One possible explanation for the absence of differences in the evaluation metrics could be attributed to the specific type of cross-validation employed in this study. In this research, we utilised leave-one-out cross-validation (LOOCV), which is a common technique. However, considering the clustered nature of our data, this choice of cross-validation may have led to inflated results, suggesting very strong predictive capabilities of the models.

A possible solution to this is to employ an alternative cross-validation technique, such as leaving out one country at a time or utilising a cross-validation strategy that accounts for the clustering structure. With this method, we might have obtained a more accurate assessment of the true performance of the models. This alternative approach could provide a better understanding of the models' generalizability and predictive power beyond the specific data set used for training and evaluation.

While the stepwise regression models and the mixed effect model produced similar metrics, we may still prefer the latter over the former because mixed effects models are specifically designed to handle nested or hierarchical data structures. Using a mixed effects model allows us to explicitly model the within-group correlations and account for the variability across different levels. This model is also often preferred over stepwise regression because the mixed effects models can provide more stable estimates. Stepwise regression can result in unstable variable selection, where the inclusion or exclusion of variables may change with minor variations in the data or model assumptions.

There are several other possible modelling approaches we could consider for the life expectancy data set. For example, as the data is collected over time, we could use time series analysis techniques to capture temporal dependencies. That method is particularly useful for identifying patterns or trends.

5.2 Conclusion

With increasing amounts of data becoming easily accessible and available, it is important to know how to analyse it in a meaningful way. In this paper we considered and compared several linear regression models for the life expectancy data set and examined the relationships between several explanatory variables, and life expectancy. We created linear regression, stepwise regression, and mixed effects models and evaluated the predictive qualities of these models via information criterion, and cross-validation.

It was found that when employing stepwise selection, the models which were based on AIC values performed better than the stepwise selection models based on p-values. Due to the nature of our data, the regression models based on p-values can be unreliable for variable selection as we are dealing with multicollinearity.

The initial results from the cross-validation indicate that the stepwise regression models based on AIC and the forward selection based on p-values have very similar predictive qualities to the mixed effects models with country or country and region as random effects.

However, we might still prefer the mixed effects models over stepwise regression since the former is designed specifically for clustered data as we had in our data set. Moreover, due to the leave-one-out cross-validation, we might be dealing with inflated results.

The analysis revealed the impact of collinearity, the influence of categorical variables, and the significance of model selection criteria. Additionally, the use of appropriate cross-validation techniques to account for data clustering is crucial. We also identified limitations, such as the potential for inflated results due to specific cross-validation methods.

References

[1] URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years).

[2] *Adult mortality rate (probability of dying between 15 and 60 years per 1000 population)*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/adult-mortality-rate-(probability-of-dying-between-15-and-60-years-per-1000-population).

[3] *Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/alcohol-recorded-per-capita-(15-)-consumption-(in-litres-of-pure-alcohol).

[4] *Average years of schooling*. URL: https://ourworldindata.org/grapher/mean-years-of-schooling-long-run.

[5] Douglas Bates et al. "Fitting Linear Mixed-Effects Models Using lme4". In: *ArXiv e-prints* arXiv:1406 (June 2014). DOI: 10.18637/jss.v067.i01.

[6] George Casella and Roger L. Berger. "Simple linear regression". In: *Statistical inference*. Brooks/Cole Cengage Learning, 2021, pp. 539–562.

[7] RONALD CHRISTENSEN. *Plane answers to complex questions: The theory of Linear Models*. SPRINGER, 2021.

[8] Ronald Christensen. *Advanced linear modeling: Statistical learning and dependent data*. Springer, 2019.

[9] *Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/diphtheria-tetanus-toxoid-and-pertussis-(dtp3)-immunization-coverage-among-1-year-olds-(-).

[10] Annette J. Dobson and Adrian G. Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2018.

[11] Ronald Aylmer Fisher. *The correlation between relatives on the supposition of Mendelian Inheritance*. Royal Society of Edinb., 1918.

[12] Xavier A. Harrison et al. "A brief introduction to mixed effects modelling and multimodel inference in ecology". In: *PeerJ* 6 (2018). DOI: 10.7717/peerj.4794.

[13] *Hepatitis B (HepB3) immunization coverage among 1-year-olds*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/hepatitis-b-(hepb3)-immunization-coverage-among-1-year-olds-(-).

[14] John P. Hoffmann. *Linear regression models applications in R*. CRC Press, Taylor amp; Francis Group, 2022.

[15] *Incidence of HIV, ages 15-49 (per 1,000 uninfected population ages 15-49)*. URL: https://data.worldbank.org/indicator/SH.HIV.INCD.ZS.

[16] *Infant mortality rate (between birth and 11 months per 1000 live births)*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/infant-mortality-rate-(probability-of-dying-between-birth-and-age-1-per-1000-live-births).

[17] Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer, 2008.

[18] Lasha. *Life expectancy (WHO) fixed*. Mar. 2023. URL: https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated.

[19] *Life expectancy at birth (years)*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years).

[20] *Mean BMI (kg/m^2) (age-standardized estimate)*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/mean-bmi-(kg-m-)-(age-standardized-estimate).

[21] *Measles-containing-vaccine first-dose (MCV1) immunization coverage among 1-year-olds*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/measles-containing-vaccine-first-dose-(mcv1)-immunization-coverage-among-1-year-olds-(-).

[22] Richard R. Picard and R. Dennis Cook. "Cross-Validation of Regression Models". In: *Journal of the American Statistical Association* 79.387 (1984), pp. 575–583. ISSN: 01621459. URL: http://www.jstor.org/stable/2288403 (visited on 07/10/2023).

[23] *Polio (Pol3) immunization coverage among 1-year-olds*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/polio-(pol3)-immunization-coverage-among-1-year-olds-(-).

[24] *Prevalence of thinness among children and adolescents, BMI ¡ -2 standard deviations below the median (crude estimate)*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/prevalence-of-thinness-among-children-and-adolescents-bmi--2-standard-deviations-below-the-median-(crude-estimate)-(-).

[25] Jeffrey M. Stanton. "Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors". In: *Journal of Statistics Education* 9.3 (2001), null. DOI: 10.1080/10691898.2001.11910537. eprint: https://doi.org/10.1080/10691898.2001.11910537. URL: https://doi.org/10.1080/10691898.2001.11910537.

[26] *Under-five mortality rate (per 1000 live births) (SDG 3.2.1)*. URL: who.int/data/gho/data/indicators/indicator-details/GHO/under-five-mortality-rate-(probability-of-dying-by-age-5-per-1000-live-births).

A Regression models

List of Tables

	Dependent variable:
	Life_expectancy
Year	$0.143^{***}(0.007)$
BMI	$-0.415^{***}(0.064)$
Schooling	$-0.098^{***}(0.029)$
Alcohol Consumption	$-0.027^{**}(0.012)$
Diphtheria	$0.009^{***}(0.003)$
Adult Mortality	$-0.042^{***}(0.001)$
GDP per Capita	$0.00003^{***}(0.00001)$
Hepatitis B	0.002(0.00)
Incidents HIV	$0.155^{***}(0.025)$
Measles	0.002(0.001)
Polio	0.001(0.003)
Thinness 5-9 years	$-0.013^{*}(0.007)$
Thinness 10-19 years	$-0.014^{**}(0.007)$
Under five Deaths	$-0.044^{***}(0.004)$
Infant Deaths	−0.009(0.007)
Population	0.0002(0.001)
Economy Status Developed	$8.232^{***}(0.725)$
Constant	$-203.009^{***}(12.483)$
Observations	2,864
R^2	0.979
Adjusted R^2	0.979
Residual Std. Error	1.357 (df = 2846)
F Statistic	$7{,}925.419^{***}$ (df = 17; 2846)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Multiple Linear Regression Results

Variable	AIC	R-Sq
Full Model	4202.942	0.99751
Region	4186.942	0.99751
Economy Status Developed	4184.942	0.99751
Population	4182.964	0.99751
Polio	4181.051	0.99751
Infant Deaths	4180.699	0.99751

Table 5: Backward Elimination based on AIC Summary

Step	Removed Variable	R-Square	AIC	RMSE
1	Population	0.9975	4200.9637	0.4859
2	Polio	0.9975	4199.0514	0.4858

Table 6: Backward Elimination based on p-value Summary

Variable	AIC	R-Sq
Country	12671.486	0.95127
Adult Mortality	7489.688	0.99202
Infant Deaths	5499.942	0.99602
Year	4529.547	0.99717
Under five Deaths	4379.524	0.99731
BMI	4316.007	0.99737
Diphtheria	4264.558	0.99742
Incidents HIV	4230.155	0.99746
GDP per Capita	4213.686	0.99747
Thinness 10-19 years	4197.700	0.99749
Schooling	4187.495	0.99750
Alcohol Consumption	4183.930	0.99750
Thinness 5-9 years	4182.361	0.99751
Measles	4181.013	0.99751

Table 7: Forward Selection based on AIC Summary

Step	Entered Variable	R-Square	AIC	RMSE
1	Country	0.9966	5038.8079	0.5636
2	Region	0.9974	4359.2492	0.5005
3	Schooling	0.9974	4292.7048	0.4946
4	Adult Mortality	0.9974	4258.6169	0.4916
5	Polio	0.9975	4239.7138	0.4899
6	Under five Deaths	0.9975	4219.9406	0.4881
7	Infant Deaths	0.9975	4205.4328	0.4868
8	Year	0.9975	4201.8184	0.4864
9	BMI	0.9975	4200.2373	0.4862
10	Incidents HIV	0.9975	4198.9360	0.4860
11	Thinness 10-19 years	0.9975	4198.9637	0.4859

Table 8: Forward Selection based on p-value Summary

Variable	Method	AIC	R-Sq
Country	addition	12671.486	0.95127
Adult Mortality	addition	7489.688	0.99202
Infant Deaths	addition	5499.942	0.99602
Year	addition	4529.547	0.99717
Under five Deaths	addition	4379.524	0.99731
BMI	addition	4316.007	0.99737
Diphtheria	addition	4264.558	0.99742
Incidents HIV	addition	4230.155	0.99746
Infant Deaths	removal	4229.915	0.99745
GDP per Capita	addition	4213.519	0.99747
Thinness 5-9 years	addition	4197.466	0.99749
Schooling	addition	4187.670	0.99750
Alcohol Consumption	addition	4184.553	0.99750
Thinness 10-19 years	addition	4182.803	0.99750
Hepatitis B	addition	4181.163	0.99751
Measles	addition	4180.699	0.99751

Table 9: Stepwise Selection based on AIC Summary

Step	Variable	Added/Removed	R-Square	AIC	RMSE
1	Schooling	addition	0.974	10494.5551	1.5096
2	Adult Mortality	addition	0.977	10162.8106	1.4244
3	Polio	removal	0.977	10160.8133	1.4241
4	Under five Deaths	addition	0.978	10024.7851	1.3905
5	Infant Deaths	addition	0.979	9979.7702	1.3793
6	Economy Status Developed	addition	0.979	9945.0098	1.3707
7	Polio	addition	0.979	9926.7270	1.3661
8	GDP per Capita	addition	0.979	9911.6000	1.3623
9	Alcohol Consumption	addition	0.979	9898.4616	1.3589
10	BMI	addition	0.979	9888.5692	1.3564

Table 10: Stepwise Selection based on p-values Summary

Variable	Value	Std.Error	DF	t-value	p-value
(Intercept)	-117.66227	9.790430	2669	-12.01809	0.0000
Diphtheria	0.00746	0.002765	2669	2.69928	0.0070
Year	0.09817	0.005268	2669	18.63514	0.0000
BMI	-0.09504	0.051490	2669	-1.84575	0.0650
Schooling	0.00733	0.026807	2669	0.27356	0.7844
Alcohol Consumption	-0.03307	0.011778	2669	-2.80763	0.0050
Adult Mortality	-0.04277	0.000645	2669	-66.35252	0.0000
GDP per Capita	0.00004	0.000006	2669	7.40718	0.0000
Hepatitis B	0.00211	0.001383	2669	1.52352	0.1277
Incidents HIV	0.15942	0.024131	2669	6.60652	0.0000
Measles	0.00322	0.001374	2669	2.34625	0.0190
Polio	0.00140	0.002757	2669	0.50664	0.6124
Thinness 5-9 years	-0.01386	0.007191	2669	-1.92796	0.0540
Thinness 10-19 years	-0.01566	0.007275	2669	-2.15242	0.0315
Under five Deaths	-0.04376	0.003775	2669	-11.59376	0.0000
Infant Deaths	-0.01236	0.006964	2669	-1.77523	0.0760
Population	0.00019	0.000918	2669	0.21188	0.8322
Economy Status Developed	3.54127	0.463950	177	7.63288	0.0000

Table 11: Mixed Effects model a Summary

Variable	Value	Std.Error	DF	t-value	p-value
(Intercept)	15.756785	10.366067	2838	1.52003	0.1286
Diphtheria	-0.008575	0.005232	2838	-1.63909	0.1013
Year	0.033991	0.005176	2838	6.56772	0.0000
BMI	-0.134440	0.020081	2838	-6.69495	0.0000
Schooling	0.101453	0.016555	2838	6.12829	0.0000
Alcohol Consumption	-0.004769	0.010307	2838	-0.46272	0.6436
Adult Mortality	-0.046678	0.000554	2838	-84.28878	0.0000
GDP per Capita	0.000020	0.000002	2838	9.40045	0.0000
Hepatitis B	-0.007625	0.002312	2838	-3.29733	0.0010
Incidents HIV	0.093929	0.016334	2838	5.75052	0.0000
Measles	0.002031	0.001544	2838	1.31598	0.1883
Polio	0.009701	0.005171	2838	1.87589	0.0608
Thinness 5-9 years	0.025189	0.014989	2838	1.68052	0.0930
Thinness 10-19 years	-0.037680	0.015102	2838	-2.49501	0.0127
Under five Deaths	-0.051146	0.003546	2838	-14.42292	0.0000
Infant Deaths	-0.052484	0.005632	2838	-9.31850	0.0000
Population	-0.000226	0.000180	2838	-1.25527	0.2095
Economy Status Developed	2.479128	0.148047	2838	16.74556	0.0000

Table 12: Mixed Effects model b Summary

Variable	Value	Std.Error	t-value	p-value
(Intercept)	−117.66228	9.790430	−12.01809	0.0000
Diphtheria	0.00746	0.002765	2.69928	0.0070
Year	0.09817	0.005268	18.63514	0.0000
BMI	−0.09504	0.051490	−1.84576	0.0650
Schooling	0.00733	0.026807	0.27356	0.7844
Alcohol Consumption	−0.03307	0.011778	−2.80763	0.0050
Adult Mortality	−0.04277	0.000645	−66.35252	0.0000
GDP per Capita	0.00004	0.000006	7.40718	0.0000
Hepatitis B	0.00211	0.001383	1.52352	0.1277
Incidents HIV	0.15942	0.024131	6.60652	0.0000
Measles	0.00322	0.001374	2.34625	0.0190
Polio	0.00140	0.002757	0.50664	0.6124
Thinness five nine years	−0.01386	0.007191	−1.92796	0.0540
Thinness ten nineteen years	−0.01566	0.007275	−2.15242	0.0315
Under five Deaths	−0.04376	0.003775	−11.59376	0.0000
Infant Deaths	−0.01236	0.006964	−1.77523	0.0760
Population mln	0.00019	0.000918	0.21188	0.8322
Economy Status Developed	3.54127	0.463950	7.63288	0.0000

Table 13: Mixed Effects model c Summary

B R code

```
library(lme4)

# Convert categorical variables to factors to use as fixed effects
dataset$Country <- as.factor(dataset$Country)
dataset$Region <- as.factor(dataset$Region)

# Define your regression model formula
formula <- Life_expectancy ~ (1|random_effect) + covariate1 + ...

# Perform LOOCV
loocv <- lapply(1:nrow(dataset), function(i) {
  # Remove the i-th observation from the dataset
  training_data <- dataset[-i, ]
  testing_data <- dataset[i, ]

  # Fit the model using training data
  model <- lmer(formula, data = training_data)

  # Predict on the left-out observation
  prediction <- predict(model, newdata = testing_data)

  # Return the prediction
  return(prediction)
})

# Combine the LOOCV predictions into a single vector
loocv_predictions <- unlist(loocv)

# Calculate LOOCV performance metrics (e.g., R-squared, RMSE, MAE)
actual_values <- dataset$Life_expectancy
r_sq <- 1  - sum((actual_values - loocv_predictions)^2) / sum((actual_values
               - mean(actual_values))^2)
rmse <- sqrt(mean((actual_values - loocv_predictions)^2))
mae <- mean(abs(actual_values - loocv_predictions))

# Print the performance metrics
cat("LOOCV R-squared:", r_sq, "\n")
cat("LOOCV RMSE:", rmse, "\n")
cat("LOOCV MAE:", mae, "\n")
```