Introduction

The question of whether a statistically significant linear relationship exists between a country's economic, healthcare, and social characteristics and the average life expectancy of its population is highly relevant in today's society: given specific values of these characteristics, the ability to predict the effect that changes to these factors have on average life expectancy could aid significantly with public health and policy planning. *Our goal for this project is to answer this question by developing a linear regression model which predicts the effect of several economic, health, and social attributes on a country's average life expectancy*. Previous studies indicate that similar models are feasible to produce: for example, Roffia et al. produced a multiple regression model to predict life expectancy at birth for 36 OECD countries [1]. Many more papers attempting to develop such models have been published in the past [2][3]. Our study *largely agrees* with the methods employed by the existing literature, since we plan to follow the *same statistical methods* (e.g. backward stepwise selection, ANOVA, residual analysis) used by the above studies in order to achieve a *similar goal* (constructing a regression model which predicts the effect of economic/health/social attributes on a country's mean life expectancy).

Methods

Initial Variable Selection

We found our data on Kaggle. The initial dataset was a .csv file containing a grand total of 22 variables; to make our analysis more relevant for our research question, we focused on the following 11 variables as a starting point:

Variable Name	Units	Variable Description	Role in Model
Adult.Mortality	Person/1000	Adult Mortality Rates of both sexes per 1000 population	Predictor Variable
Alcohol	Litres/person	Per capita consumption, in Litres	Predictor Variable
BMI	Body Mass Index	Average Body Mass Index of entire population	Predictor Variable
under.five.deaths	Person/1000	Number of under-five deaths per 1000 population	Predictor Variable
Total.expenditure	%	General government expenditure on health as a percentage of total government expenditure (%)	Predictor Variable
HIV.AIDS	Deaths/(1000 live births)	Deaths per 1000 live births to HIV/AIDS	Predictor Variable
GDP	USD	Gross domestic product per capita (measured in USD)	Predictor Variable
Population	People	Population of the country	Predictor Variable
Income.composition.of.resources	Number from 0 to 1	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)	Predictor Variable
Schooling	Years	Average number of years of schooling	Predictor Variable
Life.expectancy	Years	Average life expectancy in the country	Response Variable

Figure 1: Variables of interest in the initial dataset and their roles in the model we plan to construct.

To find a final model for the data, we will follow the following procedure in the order indicated by the numbers on the left of each step. To make marking easier, the statistical tools, assumptions, and diagnostics we used are highlighted throughout the procedure with purple, blue, and red respectively.

Procedure Phase 1: Data Initialization, Exploratory Data Analysis, Data Splitting

1. Initialize the dataset in R by using R's read.csv function, and remove all observations which have missing values using R's na.omit function. Since the data is sorted by year, there is a possibility of temporal autocorrelation existing between different years, which we will avoid by only considering data from the year 2012 (this is a limitation of the model, which will be further discussed in the Discussion section).
2. Perform an exploratory data analysis. We will do this by:
    - Using R's summary function to summarize the data. The summary function returns key statistics such as the mean/median/mode and standard deviation of each variable we are considering.
    - Running a scatterplot matrix to visualize relationships between each predictor and the dependent variable. If the plot of a predictor and the response appears to follow a straight line, this might support a linear relation between the two variables.
    - Plotting a Pearson correlation matrix to determine which variables are strongly correlated to life expectancy (which might support including the variable in the model), and which variables are

strongly correlated to each other (in which case we should not use both of them as this would induce multicollinearity in the model).
- Plotting a histogram of the proportion of the response variable. This will help illustrate the distribution of the response variable, which can help to identify outliers.

While we will not use our EDA to jump to any formal conclusions, this tool will help us make heuristic judgements on whether the data appears to satisfy the aforementioned assumptions, and will also help us identify suspicious outliers (which enables us to clean any suspicious values).

3. Once Steps 1 and 2 have been completed, split the data into a training dataset and a testing dataset by randomly allocating 50% of the data to the training dataset and 50% of the data to the testing dataset.
   - Splitting data improves the model's ability to generalize to new data, since it helps avoid overfitting and enables us to test the model's performance on previously unseen data (via the testing dataset) [4]. Thus, the technique of splitting data will help us make more accurate conclusions about our model's ability to predict average life expectancies of countries.

Procedure Phase 2: Construction of Initial Model, Initial Model Diagnostics/Validation

4. Use R's lm function to perform multiple regression and generate a least squares model for the training dataset.
5. Use R's summary function to summarize key features of the model. This function displays key attributes of the least squares model which will help us identify (and conclude) which predictor variables don't have a significant linear relation with the dependent variable (life expectancy). Of particular importance are the following attributes:
   - The R^2 (and adjusted R^2) value of the model for the training data. If these values are close to 1, this indicates a strong linear relationship between the predictor variables and the response variable
   - The estimates for the variable coefficients, as well as their corresponding p-values. These will help us conclude what the values of the coefficients will be in our model.
6. Identify problematic observations in the training data.
   - We will do this by writing R code to identify outliers, leverage points, and influential points of the model. For any observations that fall into the above categories (standardized residuals $\notin$ [-2,2], h_ii > 2p/n, Cook's D >4/(#observations)), we conclude that they are problematic.
   - While problematic observations can hamper the fit of our model, we do not believe it is justified to remove them just to achieve a better fit, since removing points which legitimately came from the training dataset might make our model worse at generalizing to unseen data. As such, we will not remove a problematic observation unless examining its recorded attributes reveals that it is impossible for the observation to be legitimate (e.g., if we find a data point which claims the GDP of a country is $-75, we can remove the point since it must be faulty).
7. If any observations were removed from the training data after Step 6, refit the model to the new training data which does not contain these observations.
8. Check for multicollinearity by computing the Variance Inflation Factor (VIF) of the model.
   - While multicollinearity is typically difficult to avoid in regression models, the presence of problematic multicollinearity can hamper our model's ability to make predictions by making it difficult to determine the exact contributions of each individual predictor to the dependent variable. Thus, we want to ensure that problematic multicollinearity is not present in our model to maintain the accuracy and interpretability of the regression coefficients.
   - If the VIF is greater than 5 for any variable, we conclude that problematic multicollinearity exists within the model [5]. If such a VIF is found, we will consider (but not necessarily move forward with) removing or transforming the variable which has this VIF until the model no longer has problematic multicollinearity.
9. Perform residual analysis in order to check whether the data satisfies the assumptions of *i)* linearity, *ii)* homoscedastic (i.e. constant variance) errors, *iii)* normally distributed errors, and *iv)* independent errors.

- We will check if *i), ii)* and *iv)* are satisfied through the *diagnostic of plotting the residuals against the fitted values*. If the data points appear to be randomly scattered in such a plot, then we conclude that linearity, homoscedasticity and error independence are satisfied.
- We will check if *iii)* is satisfied through the *diagnostic of plotting a normal Q-Q plot of the residuals* using R's qqnorm function. If the 45-degree reference line of the Q-Q plot (which can be plotted with R's qqline function) appears to fit the residuals well, then error normality is satisfied. If the residuals deviate excessively from this line, then error normality is not satisfied.
- We can also evaluate *iii)* normality of the residuals using the *Shapiro-Wilk test*. If the W-statistic is close to 1 and the p-value is above 0.05, then we fail to reject the null hypothesis of the residuals being normally distributed, which will help us conclude that the residuals are normally distributed.

10. Handle any assumption violations that were found in step 8.
    - If a violation of the assumption in *i)* or *iv)* occurs, we can either handle it by performing backward elimination to see if linearity holds after excluding insignificant variables, *or* we can try to apply a variable transformation on predictor variables so that the residual vs. fitted values plot looks randomly scattered after the transformation.
    - If a violation of the assumption in *ii)* or *iii)* occurs, we will attempt to apply variance stabilization by transforming the response variable in a way that helps us equalize the variance across different levels of the predictor variables. For example, we can consider logarithmic, square root, inverse, or Box-Cox transformations if necessary.

Procedure Phase 3: Model Tweaking, Comparison, & Final Model Diagnostics/Validation

11. If insignificant variables were found in Step 5, perform backward elimination to drop them from the model.
    - We will use R's step function to automatically discard variables from the model whose p-values are too high for us to conclude a significant linear relationship exists between such variables and the dependent variable.
    - After all irrelevant variables have been dropped, repeat Steps 8-10 in order to ensure our new model does not violate any of the regression assumptions.
12. Compare the reduced model with the initial model to determine which model is a better fit.
    - We can do this by running an ANOVA test using R's anova function. ANOVA compares the reduced model to the initial model and returns an F-statistic whose p-value we can assess to determine if the initial model's performance is significantly better ($p < 0.05$).
    - We can also compare the adjusted R^2 values for the initial model and the reduced model to help solidify our choice. As per Prof. Herrera's notes [6], we conclude the model with the higher adjusted R^2 should be preferred, and if the reduced model has a slightly lower R^2 we should still prefer the reduced model since it has fewer variables.
13. If the coefficient estimates are still too insignificant, repeat Steps 11-13 until we arrive at a model which satisfies the assumptions and achieves as high of a confidence in the coefficient estimates as we can manage. This will be our final model.
14. Validate the final model by comparing results of applying the final model to the testing dataset versus results of applying the final model to the training dataset. In order for us to conclude our model can generalize to previously unseen data successfully, we should witness the following outcomes:
    - There should be minimal differences in the estimated regression coefficients, similar adjusted goodness-to-fit values, similar problematic observations (if they exist), similar multicollinearity (if it exists), the same significant predictor variables, and similar impact of problematic observations on the model [7].

    If one or more of these outcomes does not occur, we conclude that our model *fails* to generalize successfully to new, previously unseen data.

Results

Description of Data

A preliminary inspection of the data revealed several observations that *didn't make sense*, as there were several countries with average BMIs of more than 60, which is impossible (an adult is obese if their BMI is >30 [8]). We chose not to consider any observations whose impossibility we were POSITIVE about, but did not remove observations that seemed implausible but still possible (since we did not want to remove valid data). Furthermore, we also found *missing values* in several of the rows, so we *decided to exclude observations with missing values from our analysis (as per Step 1)* using the na.omitall function in R. This caused the size of the dataset to be reduced to 129 observations from the initial size of 183.

Once data cleaning was done, we performed an exploratory data analysis (as per Step 2) to *summarize each variable numerically and visually* through a table of summary statistics, scatterplot matrix, correlation matrix, and response variable histogram.

```
 Adult.Mortality    Alcohol            BMI         under.five.deaths Total.expenditure    HIV.AIDS
 Min.   :  2.0   Min.   : 0.010   Min.   : 2.1   Min.   :   0.00   Min.   : 1.100    Min.   : 0.100
 1st Qu.:121.0   1st Qu.: 0.010   1st Qu.: 6.9   1st Qu.:   3.00   1st Qu.: 3.700    1st Qu.: 0.100
 Median :214.0   Median : 0.010   Median :18.4   Median :  29.00   Median : 5.110    Median : 0.600
 Mean   :207.5   Mean   : 2.502   Mean   :16.9   Mean   :  87.42   Mean   : 5.569    Mean   : 1.853
 3rd Qu.:285.0   3rd Qu.: 3.100   3rd Qu.:22.7   3rd Qu.:  79.00   3rd Qu.: 7.110    3rd Qu.: 1.900
 Max.   :513.0   Max.   :16.350   Max.   :29.0   Max.   :1400.00   Max.   :12.600    Max.   :12.200
      GDP             Population          Income.composition.of.resources   Schooling
 Min.   :   52.35   Min.   :      385   Min.   :0.3310                  Min.   : 5.10
 1st Qu.:  561.90   1st Qu.:   513979   1st Qu.:0.4630                  1st Qu.: 9.60
 Median :  973.47   Median :  2567666   Median :0.5330                  Median :10.80
 Mean   : 3999.05   Mean   : 22265626   Mean   :0.5629                  Mean   :11.17
 3rd Qu.: 3158.59   3rd Qu.: 15572753   3rd Qu.:0.6660                  3rd Qu.:12.40
 Max.   :49474.76   Max.   :248883232   Max.   :0.9210                  Max.   :18.50
```

```
> describe(filtered_data)
                                vars  n        mean          sd      median     trimmed        mad    min
Adult.Mortality                    1 57      207.54      127.17      214.00      203.45     114.16   2.00
Alcohol                            2 57        2.50        4.20        0.01        1.64       0.00   0.01
BMI                                3 57       16.90        8.07       18.40       17.19       7.12   2.10
under.five.deaths                  4 57       87.42      215.79       29.00       40.11      41.51   0.00
Total.expenditure                  5 57        5.57        2.53        5.11        5.38       2.31   1.10
HIV.AIDS                           6 57        1.85        2.76        0.60        1.27       0.74   0.10
GDP                                7 57     3999.05     9362.38      973.47     1814.28     888.01  52.35
Population                         8 57 22265626.12 49516329.87 2567666.00 9611414.09 3605640.20 385.00
Income.composition.of.resources    9 57        0.56        0.15        0.53        0.55       0.13   0.33
Schooling                         10 57       11.17        2.79       10.80       10.99       2.37   5.10
                                         max        range  skew kurtosis         se
Adult.Mortality                       513.00       511.00  0.19    -0.74      16.84
Alcohol                                16.35        16.34  1.69     1.90       0.56
BMI                                    29.00        26.90 -0.52    -1.12       1.07
under.five.deaths                    1400.00      1400.00  4.62    23.13      28.58
Total.expenditure                      12.60        11.50  0.72     0.18       0.33
HIV.AIDS                               12.20        12.10  1.89     2.96       0.37
GDP                                 49474.76     49422.41  4.04    16.43    1240.08
Population                      248883232.00 248882847.00  2.91     8.24 6558598.11
Income.composition.of.resources         0.92         0.59  0.80    -0.31       0.02
Schooling                              18.50        13.40  0.59     0.18       0.37
```

Figure 2: (L) Summary statistics for variables of interest, (R) Further statistics for variables of interest

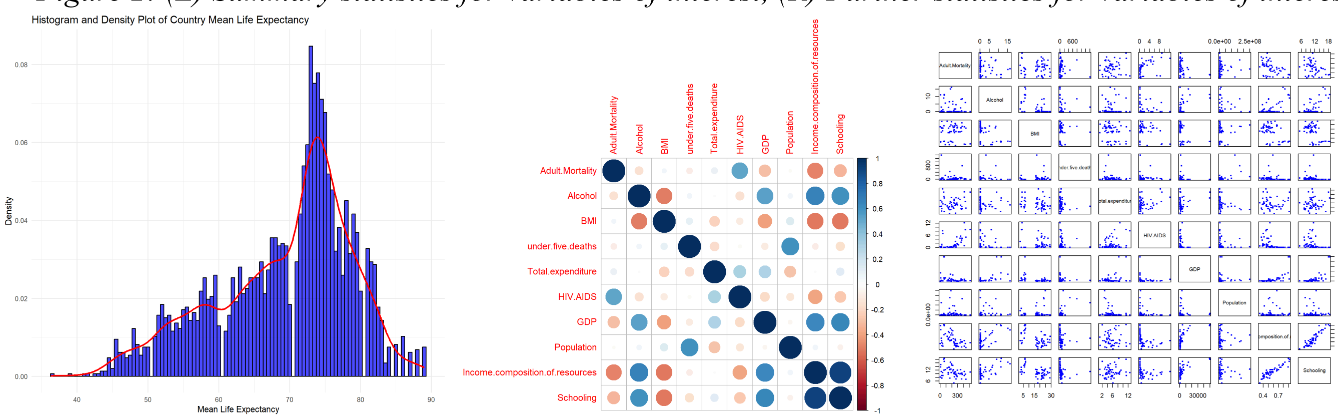


Figure 3: (L) Response histogram, (M) correlation matrix, scatterplot matrix (R) for variables of interest

From Figure 2(R), we can see that the Population has a high *spread,* with a standard deviation of over 49 million. This factor combined with its high mean relative to other variables introduces the potential that this variable could dominate any regression model we make.

After our EDA, we split our data 50-50 into testing and training datasets (as per Step 3), and then fitted an initial model to the training dataset using R's lm function (Step 4). The summary of this model revealed that 7 of the predictors (all except "Total.expenditure", "HIV.AIDS", and "Income.composition.of.resources") were not significant. After identifying problematic observations (leverage points, outliers, and influential points), we discovered several *outliers* (points whose standardized residuals fell outside the range [-2, 2]; plot shown in [1] in Appendix) and decided to inspect them more closely to determine whether they were legitimately collected. However, our inspection revealed nothing suspicious about them (none of the attributes associated with them seemed to have impossible values), so we did not remove them in accordance with Step 6.

For Step 7 (checking multicollinearity), we computed the VIF of the predictors and found that 2 of them ("Income composition of resources" and "Schooling") had a VIF greater than 5. After referring to the correlation matrix in Figure 3(M), we realized that these variables were highly correlated (the matrix indicated their correlation was close to 1); thus, we decided it would make sense to revise our model to only consider the "Income.composition.of.resources" variable, which had a higher significance in the model. While doing this reduces the number of variables, it was nevertheless a necessary decision since we want to ensure that each individual predictor is stable and interpretable (and keeping multiple coefficients with high correlation prevents this).

Our diagnostics for the regression assumptions (Step 8) revealed that all the assumptions were satisfied. The plot of the residuals vs. fitted values appeared to be randomly scattered, indicating that the linearity, constant error variance, and error independence assumptions were satisfied. With regard to residual normality, the Q-Q plot showed the residuals aligning well with the 45-degree reference line, and the Shapiro test yielded W = 0.98498, p = 0.6075 which supported residual normality. These results are shown in the figure below.

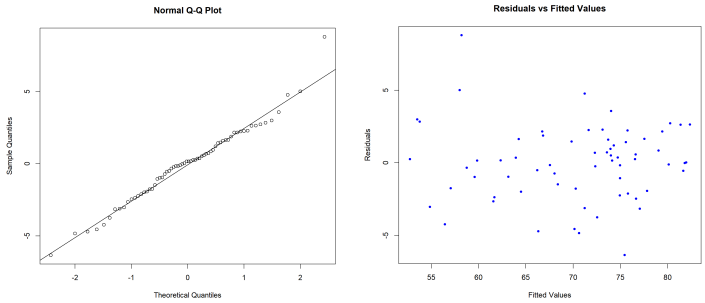


Figure 4: (L) Normal Q-Q plot & (R) residuals vs. fitted values plot of final model (on training data)

Since several predictors in the initial model were found to be insignificant, we performed backward elimination as per Step 11, and arrived at a reduced model of the form *"Life.expectancy = 42.4358 + 0.5214(Total.expenditure) + -1.2135(HIV.AIDS) + 39.8965(Income.composition.of.resources)"*. As a result, all the remaining predictor variables were deemed significant by R's summary function, with p-values of 0.000946, 2.98e-09, and <2e-16 for the estimates of 0.5214, -1.2135, and 39.8965 respectively. The R^2 was 0.9053. Upon repeating Steps 8-10 to ensure the new model did not violate assumptions, we found that all predictors had an acceptable VIF (less than 5), indicating no problematic multicollinearity in the new model; furthermore, repeating the diagnostics in Step 9 revealed that all the assumptions were once again satisfied.

Now that all the irrelevant variables had been appropriately handled, we used ANOVA to compare our reduced model to our initial model as per Step 13. The F-statistic returned by ANOVA did not give evidence that the initial model was a significantly better fit than the reduced one (p-value = 0.918), and the adjusted R^2 of the reduced model (0.9007) was higher than the initial one (0.8931). As such, we followed the guidelines of Step 13 and concluded the reduced model was a better fit since it had less variables and a higher adjusted R^2. We thus considered the reduced model to be our final model, and proceeded to validate it by testing its performance on the testing dataset in accordance with Step 15. The result was that the estimated regression coefficients were similar, the adjusted R^2 was similar, the problematic observations were similar, and the VIFs were similar. However, we witnessed a significant limitation, since one of the predictors that was initially identified as significant now had a p-value of 0.93 (indicating non-significance). This discrepancy led us to conclude that our model was *not* generalizable to real-world data (in accordance with the guidelines in Step 15). These results are shown in the figure below.

```
> anova(reduced_model1, model)
Analysis of Variance Table

Model 1: Life.expectancy ~ Total.expenditure + HIV.AIDS + Income.composition.of.resources
Model 2: Life.expectancy ~ Adult.Mortality + Alcohol + BMI + under.five.deaths +
    Total.expenditure + HIV.AIDS + GDP + Population + Income.composition.of.resources +
    Schooling
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     62 438.25
2     55 418.73  7    19.515 0.3662  0.918
```

```
Call:
lm(formula = Life.expectancy ~ Total.expenditure + HIV.AIDS +
    Income.composition.of.resources, data = trainData)

Residuals:
     Min       1Q   Median       3Q      Max
-6.3469 -1.7677  0.1643  1.6351  8.8029

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      42.4358     1.7593  24.121  < 2e-16 ***
Total.expenditure                 0.5214     0.1502   3.472 0.000946 ***
HIV.AIDS                         -1.2135     0.1754  -6.918 2.98e-09 ***
Income.composition.of.resources  39.8965     2.7208  14.664  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.659 on 62 degrees of freedom
Multiple R-squared:  0.9053,    Adjusted R-squared:  0.9007
F-statistic: 197.6 on 3 and 62 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Life.expectancy ~ Total.expenditure + HIV.AIDS +
    Income.composition.of.resources, data = testData)

Residuals:
     Min       1Q   Median       3Q      Max
-9.9988 -1.8354  0.1651  1.8976  9.9565

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     44.88245    2.46396  18.216  < 2e-16 ***
Total.expenditure                0.01632    0.18454   0.088     0.93
HIV.AIDS                        -1.54212    0.26702  -5.775 3.04e-07 ***
Income.composition.of.resources 39.95198    3.29846  12.112  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.51 on 59 degrees of freedom
Multiple R-squared:  0.8436,    Adjusted R-squared:  0.8356
F-statistic: 106.1 on 3 and 59 DF,  p-value: < 2.2e-16
```

Figure (5): (L) Results of ANOVA comparison between reduced and initial model, (M) Summary statistics of final model applied to training data, (R) Summary statistics of final model applied to testing data

Discussion

Final Model Interpretation and Importance

We interpret the "Income.composition.of.resources" coefficient *in the context of our research question* as follows. Due to its low p-value (p < 2e-16), as well as this coefficient remaining stable when the model was applied to both the testing and training dataset (as shown in Figure 5), we are strongly confident that *an increase of 1/39.8965 in the Human Development Index (in terms of income composition of resources) of a*

country will result in an increase of 1 year in the average life expectancy of this country (assuming all other variables are held fixed). A *general summary* of what the model tells us about the relation between the predictors and response is as follows: (1) Since the R^2 value of the model on the testing data is 0.8436, our model explains 84.36% of the variability in the response variable (average life expectancy of a country). (2) When all predictor variables are zero, we expect an average life expectancy of 42.4358 for the country. (3) General government expenditure on health (as a percentage of total government expenditure has a positive impact on life expectancy, HIV/AIDS death rates have a negative impact, and the Human Development Index has a positive impact (the coefficient is inflated since this variable can only range from 0 to 1).

We emphasize that our final model contributes to answering the initial research question, which is supported by some of its predictors being extremely significant. Indeed, the p-values of <2e-16 and 2.98e-09 for the "Income.composition.of.resources" and "HIV.AIDS" predictors (when applying the model to testing data) suggest a strongly significant linear relationship between these variables and average life expectancy; this contributes partially towards answering our research question (whether a linear relationship exists between economic+healthcare+social characteristics of a country and average life expectancy). Specifically, we were able to find a significant linear relation between Economic/Social factors and avg. life expectancy through the former predictor, and a significant relation between Healthcare factors through the latter (since HIV/AIDS death rates are relevant to healthcare, especially in poorer countries where the disease is more prevalent). However, there still remain limitations of the model that prevent our conclusions from being completely accurate/useful.

Limitation 1: Model's significant predictors were not the same for the testing and training dataset

One limitation is that one of the predictors ("Total.expenditure") was deemed significant when we applied our final model to the training dataset, but after the model was applied to the testing dataset it was no longer significant. This negatively impacts the usefulness of the final model, because it is clear that our model's ability to generalize to new data is more limited if at least one of its predictors is non-significant. We emphasize that we couldn't correct this issue because we had already constructed our final model; attempting to change the model after applying it to testing data would only cause us to lose the ability to ascertain whether our model would be generalizable to unseen data.

Limitation 2: Potential violation of the error independence assumption

Another limitation is that the error independence assumption underlying our regression model could be violated, since the dataset groups observations by country, and countries that are geographically or culturally close may exhibit cultural/governmental similarities that lead to correlated errors. This lack of independence in the residuals can bias the standard errors, reducing our model's overall utility by making its inferences less reliable. However, we justify our inability to correct this limitation by noting that 'grouping by country' is a problem inherent in the dataset, so every model derived from the dataset will suffer from the same problem. Moreover, correcting such an issue would not be feasible solely by using the methods taught in this course; for instance, solutions such as mixed-effects models or generalized least squares require sophisticated implementation methods which are not within the scope of STA302.

Limitation 3: Data only taken from one year (2012)

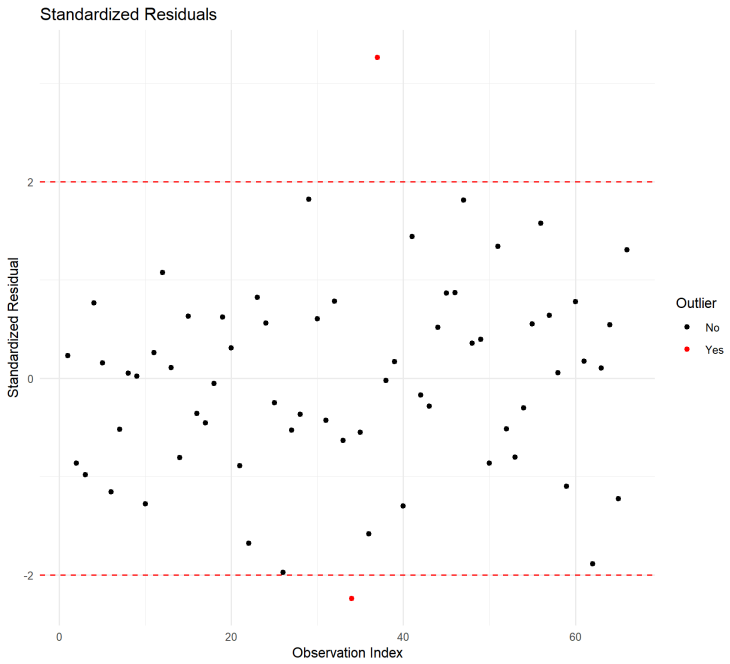
Lastly, our model is limited by our choice to only include country data taken from the year 2012. This negatively impacts the usefulness of the model, since data from 2012 may not accurately reflect the current reality in 2024, where salaries in many countries (e.g. Canada) have gone up due to inflation [9], which inflates the value of the economic variables in our model (such as Total.expenditure). Thus, while our model might have been useful for predicting life expectancy in 2013, a 12-year difference might make our model much less reliable in the present day. This problem could have been better resolved by developing a time-series model for the data (which would increase the model's long-term predictive capabilities), instead of only modelling based on data from a single year. While including time-series techniques might have improved the predictive ability of our model, we justify not using them by the fact that we were only given a brief introduction to time series regression in this course. Thus, a time-series approach would require us to employ methods outside the scope of what was taught by Prof. Herrera (which we wished to avoid, since the focus of this project is on showcasing methods taught to us in STA302).

References

[1] Roffia et al., https://link.springer.com/article/10.1007/s10754-022-09338-5
[2] Chandirasekeran et al., https://link.springer.com/article/10.1007/s40009-022-01118-6
[3] Amit et al., https://link.springer.com/chapter/10.1007/978-981-19-3391-2_30
[4] Antonio Herrera, STA302 Lecture 8 handout, pg. 6
[5] Antonio Herrera, STA302 Lecture 7 handout, pg. 3
[6] Antonio Herrera, STA302 Lecture 8 handout, pg. 3
[7] Antonio Herrera, STA302 Lecture 8 handout, pg. 7
[8] https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html
[9] Aaron O'Neill, https://www.statista.com/statistics/271247/inflation-rate-in-canada/

Appendix

[1]



Standardized residuals plot for the initial model; outliers are highlighted in red.