Introduction

Throughout history, the extension of the human lifespan has consistently been a fundamental aspect of human endeavour. The research question examined is how life expectancy can be predicted in different regions or countries based on various variables. The numerical predictor variables of the research include the year between 2000 to 2015, adult mortality rate, the number of infant deaths, alcohol consumption, Hepatitis B immunization coverage, BMI (Body Mass Index), HIV/AIDS prevalence, GDP (Gross Domestic Product), Population size, and Schooling years. The categorical predictor variable is the status of a country which was categorized into two levels: developing and developed.

These variables were chosen because they represent health, economic, and social factors that are widely believed to be correlated to the response variable which is human life expectancy. The study done by Meara, Richards, and Cutler (2008) highlighted that from 1981 to 2000, the life expectancy of people with higher education increased significantly compared with those with lower education. This echoes the "Schooling" and "Public Health" variables in our study. This study focuses on non-Hispanic black and white populations and focuses on the 1980s and 1990s. However, a larger and more up-to-date set of data should be observed. After that, Lin, Chen, Chien, and Chan (2012) stated that in less developed countries, life expectancy is largely influenced by GDP, educational resources, nutritional status, and the political system. Among these, GDP and education level are the primary influencing factors. Their research showed a linear progression between life expectancy and status, GDP and education levels. Research by Kabir (2008) found that although increases in GDP, and educational resources are crucial to social development, they do not necessarily directly lead to an increase in life expectancy. Significant increases in life expectancy are closely related to improvements in public health, such as the increase in the number of healthcare professionals. This analysis aims to provide a deeper understanding of existing literature and offers more detailed information on specific aspects of the various factors that influence life expectancy globally.

A multiple linear regression model was chosen to be used because of its ability to predict the relationship between individual variables and life expectancy. As well as evaluating the statistical significance of each variable. To use a multi-linear model, a normal distribution of the response variable Y is needed which is the key assumption of multiple linear regression. Therefore, an unbiased estimation of the standard error could be done. Furthermore, the response variable must exhibit a linear relationship with the predictor variables, which can be demonstrated using scatter plots. Following analysis, this linear relationship will enable the drawing of conclusions regarding the research question.

Method

Firstly, select the dataset, clean the dataset and randomly split it into two groups in R Studio, forming equal-sized training and test datasets. The training dataset will be used to analyze and build our final model.

Fit the model using the selected predictors, indicating the categorical predictor as either 1 for developed or 0 for developing. Using stepwise selection to choose our model.

Start with including all predictor variables, and iteratively remove and add predictors based on AIC values to find the model with the lowest AIC. Evaluate data normality using QQ plots and apply Box-Cox transformation if the plot does not show a straight line. Once normality is established, proceed to check the conditional mean response and conditional mean predictor using scatter plots. If either of these conditions is not met, the residual plots will not yield accurate information, and the patterns observed in them cannot be reliably used to identify violations.

Next, create residual plots for the response variable and each predictor variable. For the categorical variable, display each level's counts and percentages. Using residual plots we can identify the violations of assumptions. The first assumption is linearity, checked by observing whether the roughly straight line form in the residual plots can be corrected using power transformations. The second assumption is constant variance. Check if the points in the residual plots are evenly spread. If they are not, apply variance stabilizing transformations to the response variable Y. Identify the final assumption of uncorrelated errors by checking if cluster patterns exist in the residual plot. Such errors cannot be directly corrected, it is necessary to reevaluate the dataset because it suggests that the predictors may be related. Also the essential to identify the leverage points, outliers and influential points.

After verifying the assumptions, compute the ANOVA table to determine the overall significance of our predictor variables by using F-tests. A larger F-test statistic corresponds to a smaller p-value. A p-value below 0.05 indicates that the predictor variables are significant. Then, conduct partial F-tests multiple times to compare the full model with a reduced model, until eliminate all the insignificant predictors from the full model. Refit the final model and check the assumptions again above. In addition, computing the adjusted R-squared for comparisons between our full model and the reduced model. A higher adjusted R-squared indicates a better model fit. Make sure to check for multicollinearity so that our variables are independent of other variables.

Finally, validate the model by repeating all the previous steps using the test data. Check the criteria including minimal differences in estimated coefficients, predictors are linearly related, similar adjusted R-square, no additional model violations, similar numbers and types of problematic observations, and a similar value of multicollinearity.

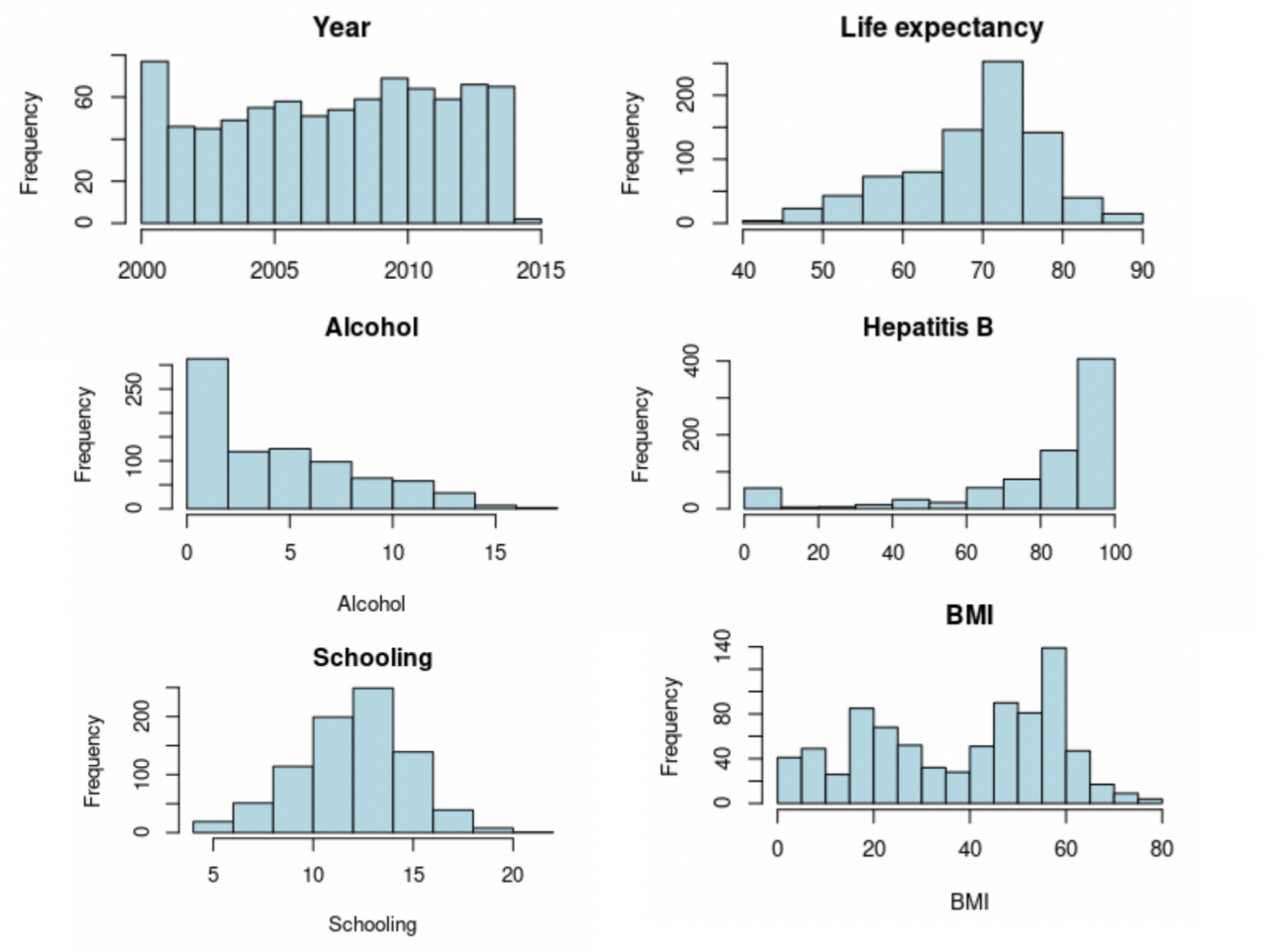
Result

	Min		1^{st} *Quantile*		Median		Mean		3^{rd} *Quantile*		Max	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Life Expectancy	44.00	44.30	64.15	64.50	71.80	71.60	69.33	69.33	75.00	75.00	89.00	89.00
Year	2000	2000	2004	2005	2008	2008	2008	2008	2011	2011	2015	2014
Adult Mortality	1.0	1.00	77.0	77.25	148.0	149.00	167.6	169.65	226.5	229.00	717.0	723.00
Infant Deaths	0.00	0.00	1.00	1.00	3.00	4.00	23.77	24.75	20.50	22.75	536.00	556.00
Alcohol	0.010	0.0100	0.815	0.6925	3.640	3.9800	4.409	4.6931	6.960	7.5950	16.580	17.8700
Hepatitis B	4.00	2.00	74.50	75.00	89.00	91.00	79.56	79.63	96.00	96.00	99.00	99.00
BMI	2.10	2.00	21.10	19.50	44.10	43.80	38.07	38.51	55.40	56.40	76.70	77.10
HIV/AIDS	0.100	0.100	0.100	0.100	0.100	0.100	1.968	2.021	0.800	0.600	49.900	50.600
GDP	11.15	1.68	461.94	466.60	1562.92	1700.38	5469.16	5732.33	4704.97	4848.73	115761.58	119172.74
Population	36	34	178786	213144	1296934	1471644	10068858	11464604	7246643	7735673	255131116	248883232
Schooling	4.20	4.50	10.30	10.40	12.30	12.40	12.03	12.23	13.80	14.20	20.60	20.70

Status	Developing	Developed
Counts	1395	243
Percentage	85.16%	14.84%

Table 1: Summary of numerical and categorical data with train and test data

The cleaned datasets were randomly divided into two groups: approximately 50% was used as training data, and the other half as test data. The dataset spans life expectancy data from 2000 to 2015. Analysis of the data reveals significant differences between the minimum and maximum values of life expectancy. While the mean for adult and infant mortality are relatively low, their maximum values are notably high. Additionally, there are huge differences in the prevalence of HIV/AIDS, as well as huge differences in levels of GDP. Most of the world's countries are still developing. The highest level of education is nearly five times that of the lowest. Notice that the test and training data yield similar results, suggesting that the data was effectively and randomly distributed between these two sets.



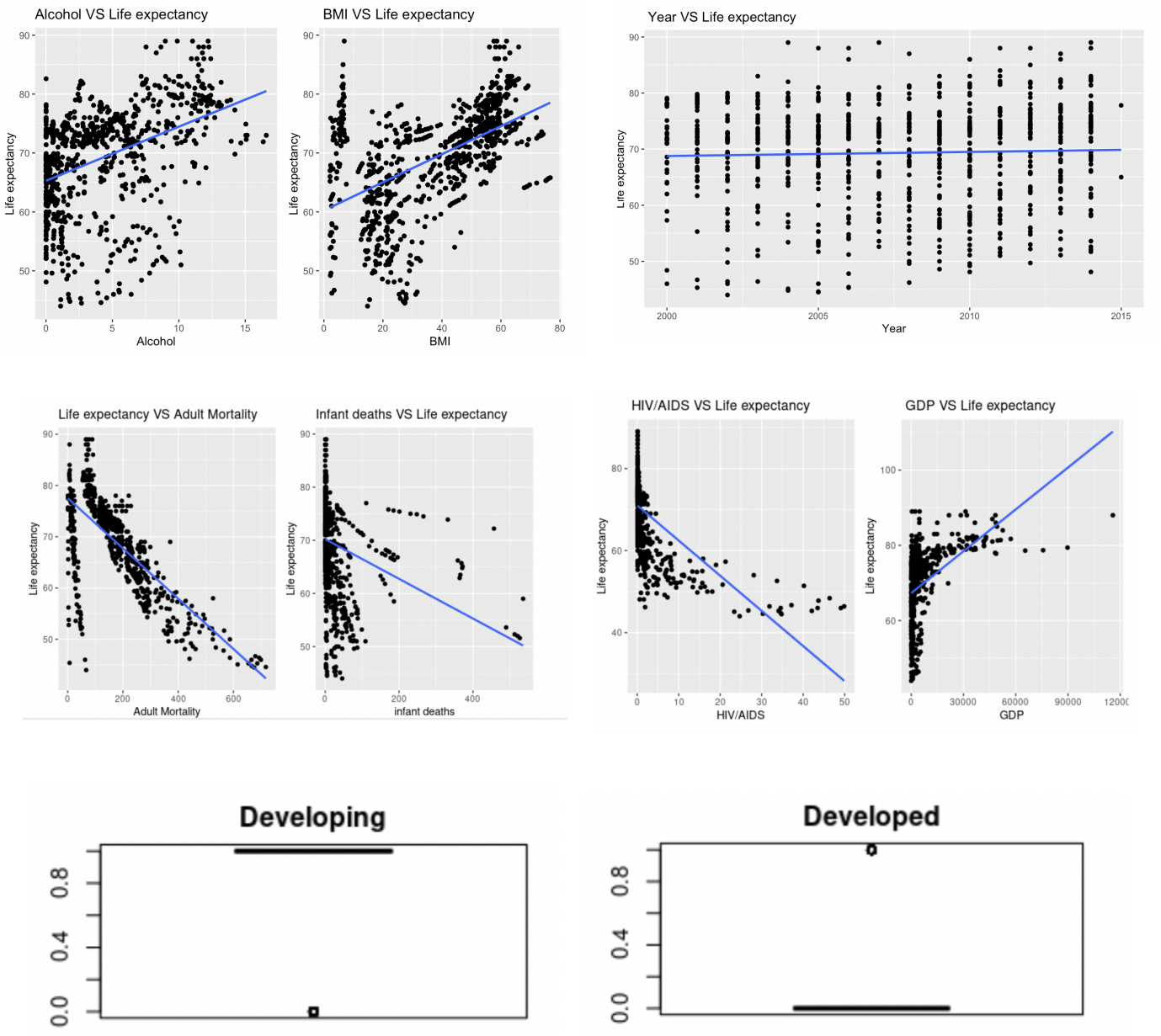


Figure 1: Histograms & Scatter plots of numerical variables, and boxplots of categorical variables.

The figures and plots presented above illustrate the data distribution of predictor variables and the response variable. The histograms of life expectancy display a relatively normal distribution, with the distribution of schooling appearing to be the most normally distributed. The scatter plots show a linear relationship between life expectancy and adult mortality the most. Additionally, the boxplots of the categorical variable indicate that most countries are in developing stages.

Construct Model

Step 1: Based on the three articles and common sense about various factors that affect life expectancy.

Life Expectancy = 264.0 − 0.1044 × Year − 0.01933 ×Adult Mortality − 0.003777 ×Infant Deaths − 0.06793 × Alcohol + 0.005994 × Hepatitis B + 0.0473 × BMI − 0.4364 × HIV/AIDS + 0.00009274 × GDP + 5.585 ×10 − 9 × Population +1.416 × Schooling + *I* status{Developing = 0, Developed = 1}

Step 2: Choosing the most significant variables from step 1 to form model 2. The significant variables have p-values smaller than 0.05.

Life Expectancy = 247.0 - 0.09616 × Year - 0.01955 ×Adult Mortality + 0.0492 × BMI - 0.4384 × HIV/AIDS + 0.00009340 × GDP + 1.425 × Schooling + *I*status{Developing = 0, Developed = 1}

Step 3: Implement stepwise selection to identify the optimal model. Calculate the AIC, AICc, and BIC of the above linear model. Select the model that demonstrates the lowest AIC, AICc, and BIC values.

Check Model Assumptions
Check Conditions 1 & 2

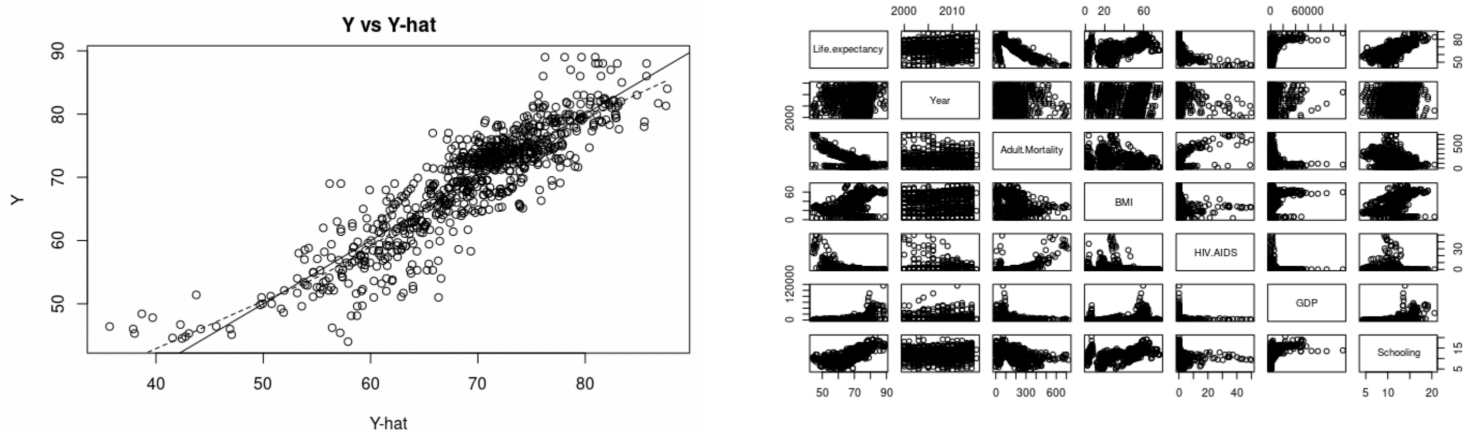


Figure 2: Plots of Y vs $\widehat{Y}$ & pair ggplots for all predictor variables.

The plot on the left demonstrates the linear relationship between variable Y and $\widehat{Y}$, with points closely aligned around the line, indicating that Condition 1 (linearity) is satisfied. The plots on the right, which display the numerical predictor variables, also show an almost linear relationship, suggesting that Condition 2 is met.

Check linear model assumption

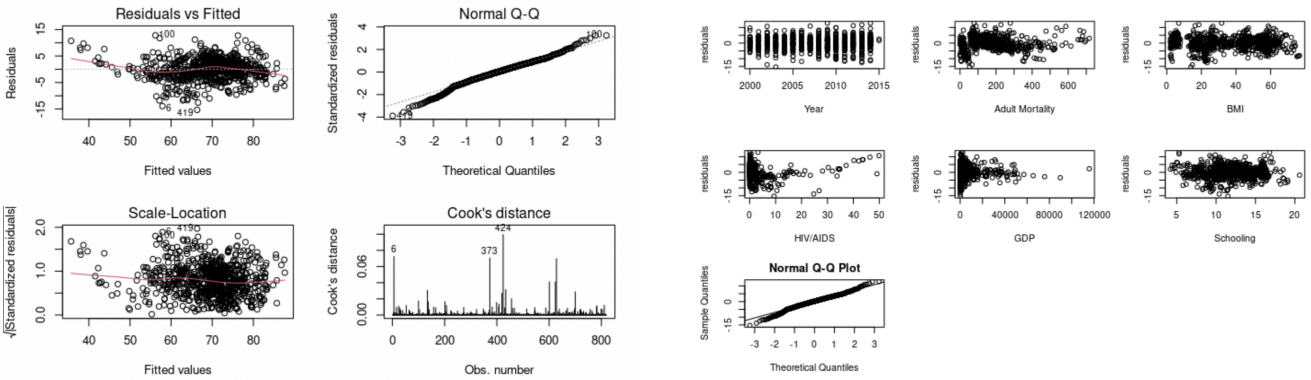


Figure 3: Residual plots & QQ plot

The residual plot shows no systematic pattern and no curved lines, suggesting a linear relationship. There are no fanning patterns or increasing variance shows constant variance. Along with the lack of clustered patterns in the plot, indicates there is no uncorrelated error which shows a good fit for the linear model. The QQ plot shows lines that form almost a straight line, providing evidence that normality exists. The Cook's distance plot gives possible problematic points, where the unusual peaks are located. Hence all four assumptions of the linear model are satisfied.

Check Problematic Observations

The model gives output that there are no outliers in the model. There are 47 leverage points which have the potential to shift the regression line. There are 57 influential points in m2 which are influential on their own fitted values. For different beta, there are 58 influential points for β_0, 66 influential points for β_1, 47 for β_2, 39 for β_3, 25 for β_4, and 58 for β_5.

Final Model

Comparing models to determine which one performs better, use F-tests and partial F-tests. These tests evaluate if a reduced model is more effective than the full model. In this process, the adjusted R^2 is calculated for each model. Furthermore, ANOVA testing reveals which predictor has the lowest p-value. According to the above tests, these results indicate that the model from the second step of model selection is the most suitable choice.

Discussion

Coefficient	Model 1(Train)	**Model 2(Train)**	Model 3(Train)	Model 4(Train)	Model 5(Train)	Model 6(Train)	Model 7(Test)
Intercept	2.640e+02	**2.470e+02**	5.410e+01	3.214e+02	69.732694	46.252791	2.634e+02
Year	-1.044e-01	**-9.616e-02**		-1.363e-01			-1.035e-01
Adult Mortality	-1.933e-02	**-1.955e-02**	-1.976e-02		-0.030859		-2.057e-02
Infant Deaths	-3.777e-03						
Alcohol	-6.793e-02						
Hepatitis B	5.994e-03						
BMI	4.730e-02	**4.920e-02**	5.084e-02		0.144988	0.065247	4.304e-02
HIV/AIDS	-4.364e-01	**-4.384e-01**	-4.290e-01	-6.686e-01	-0.380882	-0.637346	-4.805e-01
GDP	9.274e-05	**9.340e-05**	1.459e-05	1.113e-04			7.232e-05
Population	5.585e-09						
Schooling	1.416e+00	**1.425e+00**	6.716e-02	1.850e+00		1.815454	1.327e+00
Multiple R^2	0.8032	**0.8015**	0.7994	0.7448	0.6346	0.7412	0.8128
Adjusted R^2	0.8005	**0.8**	0.7982	0.7435	0.6332	0.7402	0.8115
AIC	2.270256e+03	**2.265249e+03**	2.271597e+03	2.467110e+03	2.758966e+03	2.476563e+03	2.194075+e03
AICc	2.270852e+03	**2.265471e+03**	2.271774e+03	2.467248e+03	2.759070e+03	2.476667e+03	2.194297e+03
BIC	2.344877e+03	**2.311622e+03**	2.313261e+03	2.504066e+03	2.791215e+03	2.508812e+03	2.240437e+03

Table 2: Summary of the model based on training and test data

Based on the model selection and multiple partial F-tests, the final model is model 2.
Life Expectancy = 247.0 - 0.09616 × Year - 0.01955 ×Adult Mortality + 0.0492 × BMI - 0.4384 × HIV/AIDS + 0.00009340 × GDP + 1.425 × Schooling + *I*status{Developing = 0, Developed = 1} + *I*status{Developing = 0, Developed = 1} * 1.425 × Schooling

The model shows that holding other predictor constants, each extra year of schooling adds 1.416 years to life expectancy. Each unit rise in GDP adds 0.0000934 years. Life expectancy increases by 0.4384 years for each additional HIV/AIDS unit and by 0.0492 years for each BMI unit. Conversely, each unit increase in adult mortality and each additional year decreases life expectancy by 0.4384 and 0.09616 years, respectively. Being in a developed country adds 1.435 years to life expectancy. Because the developed countries off better education.
Comparing the coefficients of Model 2 and Model 7, which represent the test data, it is evident that each corresponding coefficient is closely similar, exhibiting only minor differences from one another. Additionally, upon removing all the insignificant predictor variables, the result is Model 2, which exhibits the highest adjusted R^2 value and lowest AIC, AICc, and BIC value compared to other reduced models, which indicates the best model. The model created using the test data also shows very similar results regarding the coefficients of predictor variables and adjusted R^2.

Limitation

Given the presence of problematic points in the model, when we fit the test data using the same predictor variables, the final model does not achieve a perfect fit. This issue is primarily due to the existence of leverage and influential points. To ensure the accuracy of our dataset and the integrity of our methodology, we should refrain from removing any of these problematic points.

Conclusion

Based on our analysis, we have identified that the most significant factors influencing life expectancy are the level of education, adult mortality rate, GDP, and the year which are factors that we expected. Additionally, indicators such as HIV/AIDS and BMI are reflective of a country's overall public health status. The categorical variable shows that developed countries generally have longer life expectancy compared with developing countries. Consequently, we predict that regions characterized by higher levels of education, GDP, and BMI, coupled with better healthcare and lower adult mortality rates, are likely to experience longer life expectancies which also matches the research of three peer-reviewed articles mentioned in the introduction.

Ethics Discussion

In the analysis, automated selection methods. The dataset in the analysis brings out many possible subsets which are cumbersome. There are 11 predictors in total which construct subsets. The subsets provide a list of models that need a systematic way to select models from a large number of predictors. When finding the preferred model using R Code, the coefficients were manually changed between the predictors. Although it might not be giving the best model of fit, it could give an idea of the preferred model which has fewer differences from the best. To get the preferred model, all predictors were used initially in the R code. Then they were deleted and added back rationally to make sure they were matched with the specific coefficient and got the largest R^2, the smallest AIC, AICc, and BIC. This step including forward, and backward selection and calculating the data takes great effort if done manually. As a result, Model 2 contains the usage of the predictors except for Infant Deaths, Alcohol, Hepatitis B, and Population. Using manual selection would be reckless and negligent when doing a large dataset analysis. It is possible to check the models manually, but the possible subsets of all models that could ever be made would never be checked one by one.

References

Kabir, M. (2008). Determinants of life expectancy in developing countries. *The Journal of Developing Areas, 41*(2), 185–204. https://doi.org/10.1353/jd a.2008.0013

Lin, R.-T., Chen, Y.-M., Chien, L.-C., & Chan, C.-C. (2012). Political and social determinants of life expectancy in less developed countries: a longitudinal study. *BMC Public Health, 12*(1), 85–85. https://doi.org/10.1186/14 71-2458-12-85

Meara, E. R., Richards, S., & Cutler, D. M. (2008). The gap gets bigger: Changes in mortality and life expectancy, by education, 1981 - 2000. *Health Affairs, 27*(2), 350–360. https://doi.org/10.1377/hlt haff.27.2.350

Appendix:

```
 24  25  26  43  44  45  46  53  54  81 112 136 138 249 273 289 290 302 316 318 319 396 397 398 423 427 439 451 468
 24  25  26  43  44  45  46  53  54  81 112 136 138 249 273 289 290 302 316 318 319 396 397 398 423 427 439 451 468
519 521 523 558 605 665 666 667 672 684 705 772 773 775 808 815 816 817 818
519 521 523 558 605 665 666 667 672 684 705 772 773 775 808 815 816 817 818
```

```
[1] "beta 0"
 53  81 108 112 132 149 189 197 200 205 212 263 269 289 302 316 334 339 380 396 408 418 419 434 435 464 468 519 523
 53  81 108 112 132 149 189 197 200 205 212 263 269 289 302 316 334 339 380 396 408 418 419 434 435 464 468 519 523
552 553 554 605 620 621 622 642 672 705 731 760 770 775
552 553 554 605 620 621 622 642 672 705 731 760 770 775
[1] "beta 1"
 53  81 108 112 132 149 189 197 200 205 212 220 263 269 289 302 316 334 339 380 396 408 418 419 434 435 464 468 519
 53  81 108 112 132 149 189 197 200 205 212 220 263 269 289 302 316 334 339 380 396 408 418 419 434 435 464 468 519
523 552 553 554 605 620 621 622 642 672 705 731 760 770 775
523 552 553 554 605 620 621 622 642 672 705 731 760 770 775
[1] "beta 2"
 24  25  26  67 108 112 132 136 138 177 197 249 263 266 269 301 302 316 318 319 380 384 385 386 396 397 398 430 438
 24  25  26  67 108 112 132 136 138 177 197 249 263 266 269 301 302 316 318 319 380 384 385 386 396 397 398 430 438
451 468 501 519 523 554 555 557 558 628 629 630 631 646 665 666 667 672 703 704 705 762 772 773 774 775 808 815 816
451 468 501 519 523 554 555 557 558 628 629 630 631 646 665 666 667 672 703 704 705 762 772 773 774 775 808 815 816
817 818
817 818
[1] "beta 3"
 15  21  30  51 180 189 232 233 286 289 290 302 334 350 351 359 370 396 397 398 399 400 401 427 438 451 464 472 493
 15  21  30  51 180 189 232 233 286 289 290 302 334 350 351 359 370 396 397 398 399 400 401 427 438 451 464 472 493
515 517 519 558 605 642 667 687 705 725 726 757 772 773 797 799 808
515 517 519 558 605 642 667 687 705 725 726 757 772 773 797 799 808
[1] "beta 4"
 24  25  26 108 112 138 249 422 423 451 519 521 523 558 665 666 667 672 677 678 679 703 704 705 772 773 808 815 816
 24  25  26 108 112 138 249 422 423 451 519 521 523 558 665 666 667 672 677 678 679 703 704 705 772 773 808 815 816
817 818
817 818
[1] "beta 5"
 43  44  45  46  47  49  54  55  57  81 167 249 263 273 289 290 291 380 433 436 437 439 472 546 628 684
 43  44  45  46  47  49  54  55  57  81 167 249 263 273 289 290 291 380 433 436 437 439 472 546 628 684
```

```
          Year `Adult Mortality`              BMI        `HIV/AIDS`              GDP         Schooling
      1.053159          1.670931          1.583647          1.409018          1.310839          1.938014
```

figure 6: problematic observations and multicollinearity for test data

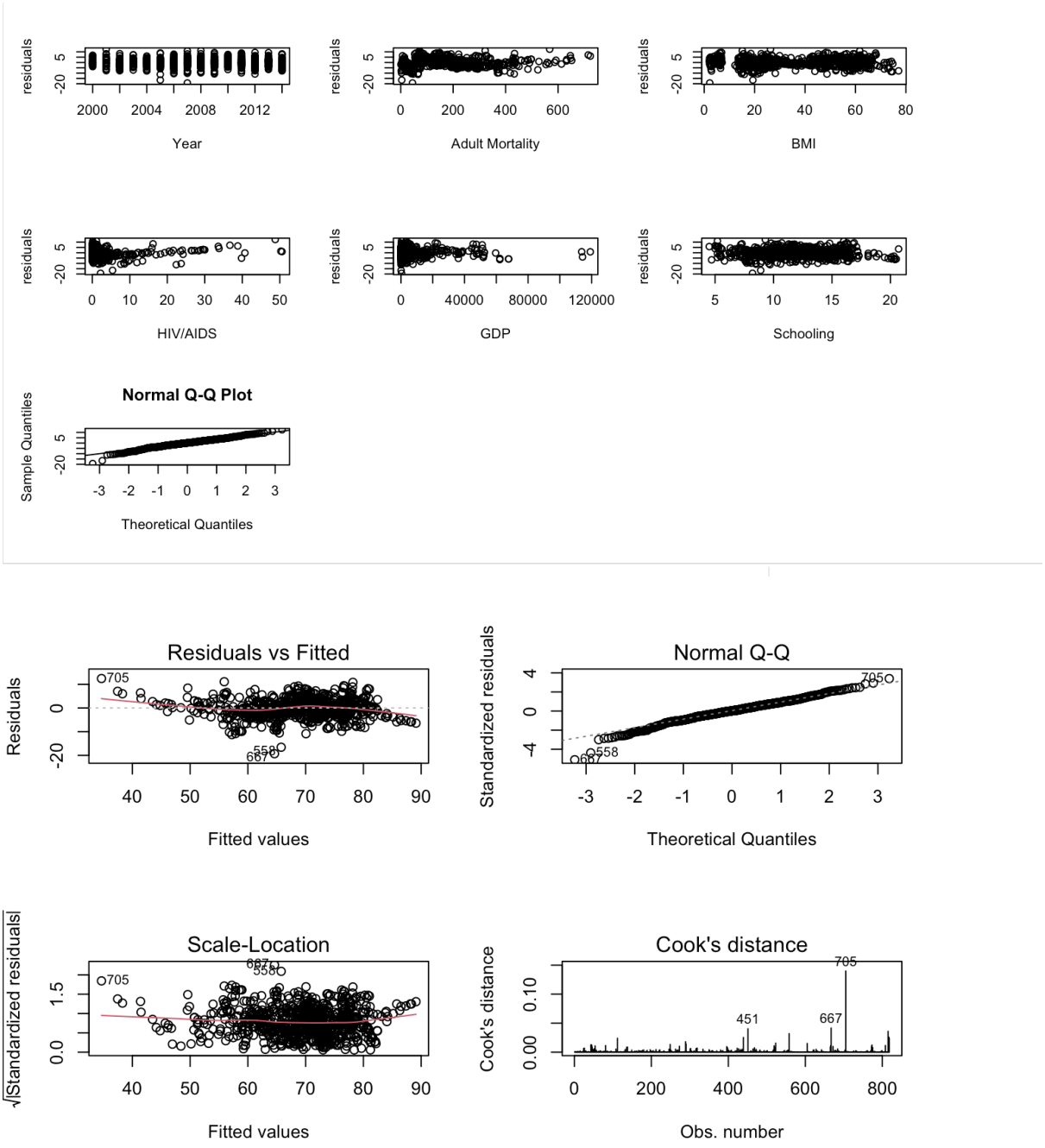


Figure 7: plots for test data