STA302: FINAL PROJECT

How do socioeconomic and lifestyle factors such as Income Composition of resources, Schooling, Body Mass Index (BMI), HIV/AIDS and Adult Mortality influence the Life Expectancy of an individual?

Introduction

Life expectancy is a measure of the expected number of years a person is expected to live.

This assumes importance for governments since it provides a measure to gauge the overall health of a community. This metric is often used by policy-makers to draft public health-care policies. It also assumes importance for individuals since determining which factors affect live expectancy, would allow individuals to make changes to their lives and improve their health and longevity.

In the past, cross-sectional studies examining the relation between increased urbanisation, income and internet-access have shown to have a positive impact on life expectancy[1,2,3]. Additionally, microsimulation and surveillance models of obesity, HIV and BMI-progression have been able to predict a decrease in life expectancy[1,2,3].

These studies have been limited to mostly small and selected samples with limited scope[1,2,3]. Since there is limited research on large-scale population-based studies across multiple categories, the purpose of our study is to examine the associations among Life expectancy, Income-Composition, Schooling, BMI, Adult Mortality and HIV/AIDS among a significantly representative sample taken from the WHO database[4].

Methods

The data for the analysis has been collected from Kaggle[4] (originally published by WHO[4]). The dataset comprises data related to socioeconomic and health factors collected from over 193 countries.

We start-off by examining the data. This consists of removal of any missing values and plotting boxplots, scatterplots, residual-vs-fitted plots, and Normal QQ-Plots to check for outliers and model assumptions.

Multicollinearity Check

We proceed by fitting a linear model with 'Life Expectancy' as the outcome variable and socioeconomic factors (from EDA) as the predictor variables.

We examine multicollinearity by checking the Variance Inflation Factor and keeping the cut-off to be >5.

If multicollinearity exits, we choose to respecify the model by removing at least one of the correlated predictors. After removing a predictor, we must re-check its effects on multicollinearity.

Model Assumption Check

We then create a residual-vs-fitted plot.

If there are any clusters of residuals which have obvious separation from the rest, the data is not independent and our analysis ends. Furthermore, if there is any systematic pattern in the residuals, such as a curve, we apply the box-cox transformation. Additionally, if there is a fanning pattern where the residuals gradually spread out, we apply the variance stabilizing transformation.

We then plot the Normal QQ-Plot and if violations exist, we apply the Box-Cox transformation. If we apply any of the above transformations, it is better to re-check model-assumptions and multicollinearity with the transformed model.

Influential Points Check

Next, we check for influential observations using Cook's Distance, DFFITS and DFBETA.

If there is a defensible reason for removing the outlier, we remove the point and refit the model, else we fit a different regression model.

It is better to re-check the model-assumptions and multicollinearity after this step.

Variable Selection

We prefer to select models that have high $R^2_{adjusted}$ values while having small AIC, corrected-AIC and BIC values. Thus, we now select variables using AIC and BIC based stepwise selection.

This 'stepwise-selection' accounts for the conditional nature of the model by iterating between Forward and Backwards selection until we are unable to add/delete variables.

Since the selected model obtained from this method is heavily dependent on the data, it may give us biased estimators. Therefore, to determine whether the model is reasonable for prediction purposes, we need to validate it.

Before model validation, we do a quick ANOVA and $R^2_{adjusted}$ check to reaffirm our choice of variable selection.

Model Validation

We validate the model using the cross-validation procedure by randomly splitting the data into k parts. Then we fit the model with k-1 parts (training) and predict the outcomes for the remaining part (test). We use all the k parts as test-set. We then check the prediction accuracy using mean absolute bias or mean squared error. These predictions are then plotted with the observed values to check the accuracy of the estimates visually.

Inference

Once the model has been validated, we can find the coefficients, confidence intervals, error-variance as well as prediction-intervals for our model.

Results

Description of Data

The data consists of 22 columns and 2938 observations. Based on available literature, we select 'Life.expectancy' as our outcome variable and 'Income.composition.of.resources', 'Schooling', 'Adult.Mortality', 'Body Mass Index (BMI)', and 'HIV.AIDS' as our predicting variables.

- Life expectancy: represents an individual's life-expectancy in years.
- Income Composition: is the Human Development Index [0-1].
- Schooling: is the number of years of formal education.
- Adult Mortality: is the probability of dying between 15 and 60 years per 1000 population.
- BMI: is the average body mass index of the population.
- HIV/AIDS: is the deaths per 1000 live births from HIV/AIDS.

These subsets of columns contain 202 rows with missing observations which we remove from our data set.

Inspection of boxplots (Appendix Plot-A) reveals the presence of a huge number of outliers for Adult Mortality, HIV/AIDS and Life Expectancy with only a few observed for Schooling and Income Composition.

The scatter plots (Appendix plot-B) reveal a positive association of Life Expectancy with Income Composition, Schooling and BMI while a negative association is observed with Adult Mortality and HIV/AIDS.

The Normal QQ-Plots (Appendix Plot-C) reveals non-normal and skewed data for Income Composition, Adult Mortality, BMI and HIV/AIDS while the residual plots (Appendix plot-C) reveals non-constant variance for Adult Mortality and HIV/AIDS.

Overall, the plots seem to support our variable selection choice and inform us about transformations we might encounter during our analysis.

Presenting the Analysis Process & Goodness of the Final Model

We begin by fitting an MLR model on our outcome and predicting variables.

A check for multicollinearity using VIF produces the values of 2.799692, 2.979326, 1.695767, 1.520382 and 1.403363 for Income Composition, Schooling, Adult Mortality, BMI and HIV/AIDS respectively. Since all the values are <5, we can conclude that there is weak collinearity and proceed with analysis.

Next, we plot a residual-vs-fitted graph, QQ-Plot and a scatter-plot to check model assumptions.

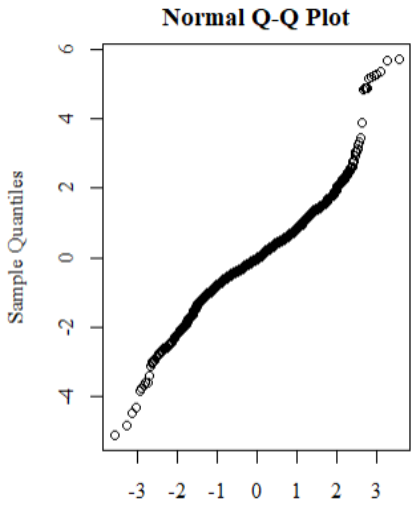


Fig 1.a: Normal QQ Plot of original data

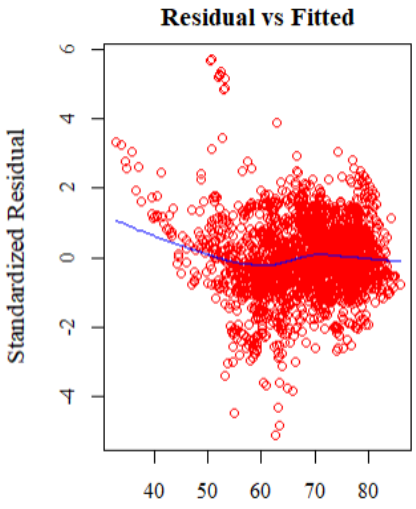


Fig 1.b: Residual vs Fitted Plot of original data

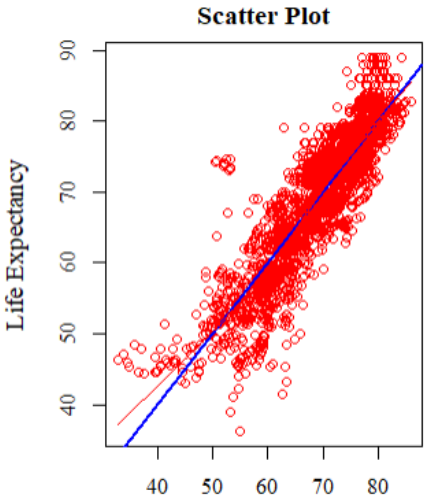


Fig 1.c: Scatter Plot of Original Data

From the scatter plot we see that the linearity assumption has not been violated. Also, in the residual-vs-fitted plot, there are no separated clusters of residuals (so independence assumption holds), there is no systemic-pattern of residuals (so linearity assumption holds), and there is no fanning-pattern amongst the residuals (so homoscedasticity holds). However, looking at the Normal QQ-Plot, we see that the data is not completely normal so we apply Box-Cox transformation. A quick check of the plots with the transformed data does not show any violations.

We proceed to check for influential points such as leverage-points and outliers. Using Cook's Distance, DFFITS and DFBETAS we detect 243 influential points. Since there is not a defensible reason for keeping these points, we remove these points and refit the model.

We again create plots with the modified data to check for any violations.

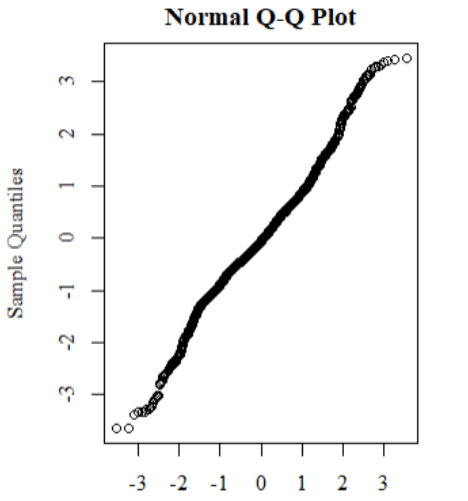


Fig 2.a: Normal QQ-Plot (without influential obs)

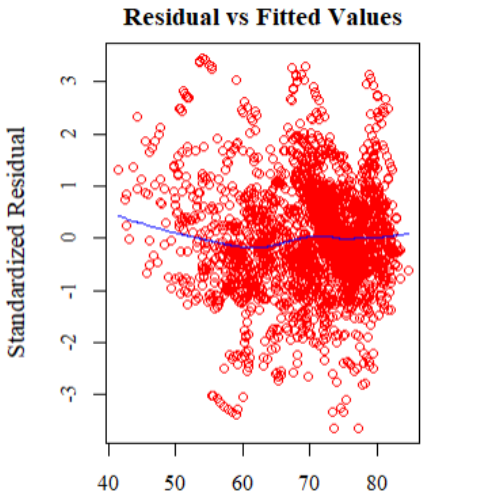


Fig 2.b: Residual vs Fitted Plot (without influential obs

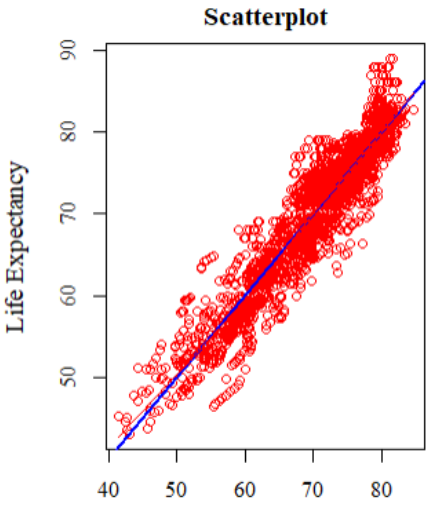


Fig 2.c: Scatter Plot (without influential obs)

The Normal QQ-Plot, shows that the points have become more uniform. There is a marked improvement in the spread of the residuals in the residual-vs-fitted plot as well as an improvement in the linear-spread of points in the scatter plot.

Next, both the AIC and BIC based variable-selection methods agree on the same set of original predictors – Income Composition, Schooling, Adult Mortality, BMI and HIV/AIDS.

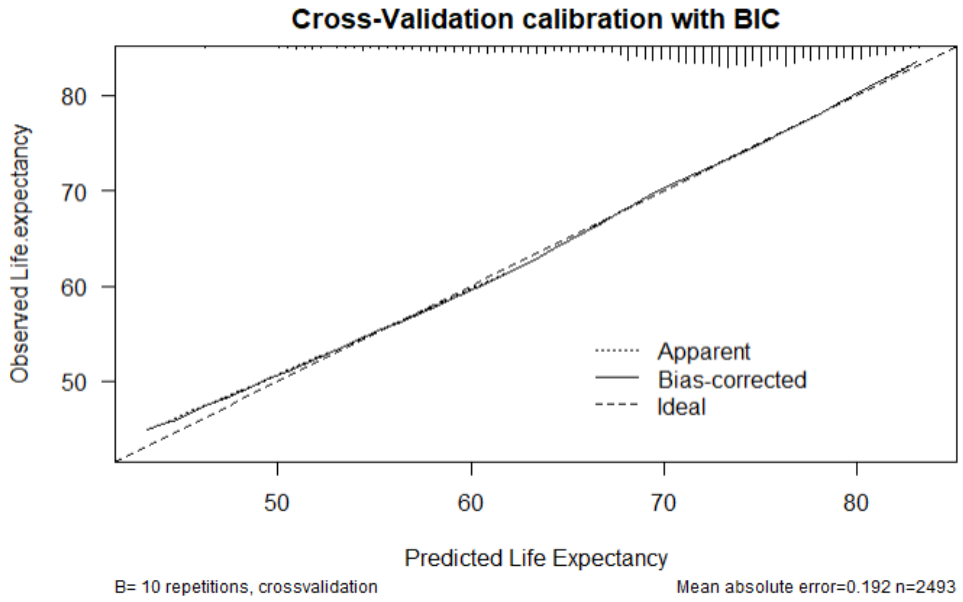
Furthermore, the conducted ANOVA test and adjusted R^2 results enforce that our model/variable selection is viable.

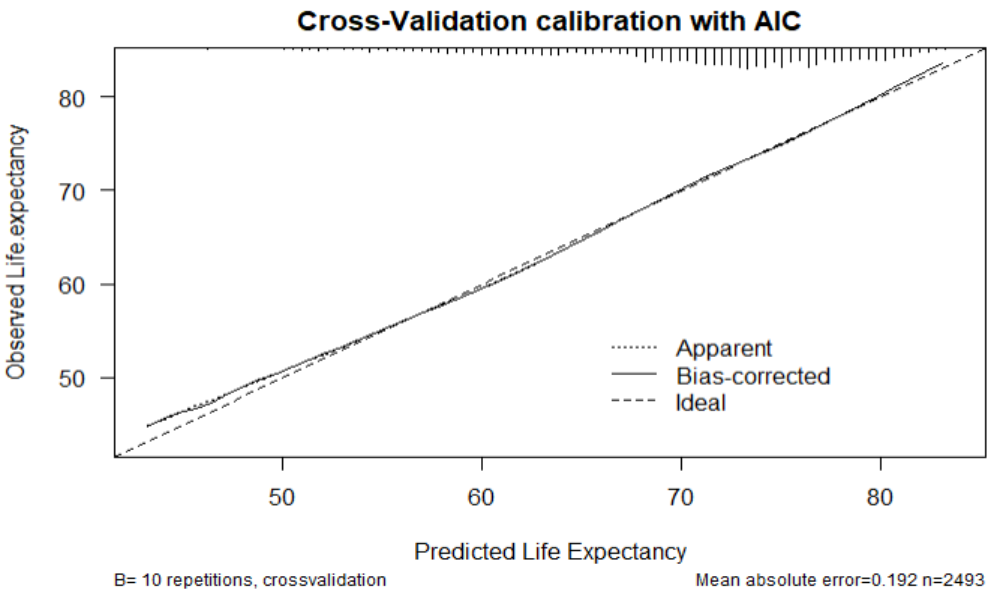
Fig 3: The summary of ANOVA for the Multiple linear regression model

term	df	sumsq	meansq	statistic	p.value
Income.composition.of.resources	1	133252.4305	1.332524e+05	15333.21488	0
Schooling	1	268.4948	2.684948e+02	30.89541	0
Adult.Mortality	1	18176.0043	1.817600e+04	2091.49340	0
BMI	1	280.9095	2.809095e+02	32.32396	0
HIV.AIDS	1	4155.7094	4.155709e+03	478.19304	0
Residuals	2487	21613.1318	8.690443e+00	NA	NA

The p-values displayed in the ANOVA table are significant and $R^2_{adjusted}$ was found to be 0.8782 (close to 1), informing that the model is a good-fit.

Now, performing a cross-validation calibration with AIC/BIC based selection, yields the following results –





Since the lines in both the graphs have a high degree of overlapping, we see that cross-validation was successful.

Now, calculating the coefficients, confidence-intervals (95%) and the error variance of our model yields the results –

Fig 5.a: The Coefficients of the model

names	x
(Intercept)	52.8528706
Income.composition.of.resources	22.2630761
Schooling	0.4832496
Adult.Mortality	-0.0213692
BMI	0.0174879
HIV.AIDS	-0.4401883

Fig 5.b: The 95% Confidence Intervals of the Coefficients

	2.5 %	97.5 %
(Intercept)	52.0933639	53.6123774
Income.composition.of.resources	20.5460405	23.9801117
Schooling	0.3965963	0.5699028
Adult.Mortality	-0.0229048	-0.0198336
BMI	0.0100281	0.0249476
HIV.AIDS	-0.4796609	-0.4007156

Error variance is found to be: 8.690443

Discussion

Final Model Interpretation and Importance

From the results of the model, we can interpret the Income Composition coefficient (from the coefficient table) to mean that keeping all the other variables constant, a 1 unit increase in Income composition would result in a 22.2630761 unit increase in Life Expectancy. Furthermore, from our confidence-intervals above, we are 95% confident that this increase would be between 20.5460405 and 23.9801117 units.

Similarity, we can interpret the Adult Mortality coefficient to mean that keeping all the other variables constant, a 1 unit increase in Adult Mortality would result in a -0.0213692 unit decrease in Life Expectancy. Furthermore, from our confidence intervals, we are 95% confident that this decrease would be between -0.0229048 and -0.0198336 units.

In summary, the model tells us that Income composition has a strong positive association with Life Expectancy while Schooling and BMI have a weak positive association with Life Expectancy. On the other hand, Adult Mortality and HIV/AIDS have a weak negative association with Life Expectancy.

This information helps answer our research question since we now know that an individual must focus on increasing his Income composition to significantly increase his life expectancy. Furthermore, increased Schooling or BMI can help to some extent. Furthermore, an individual must try to reduce Adult Mortality and prevent HIV/AIDS to increase his life expectancy, however, the overall affect in this case would not be much significant.

Limitations

A limitation of our model is that we are removing any rows for which we have missing data. This may cause us to lose out on relevant modelling data for other predictors, which also gets removed upon the elimination of the row. We might be able to correct this by applying appropriate techniques for predicting missing data.

If there was collinearity present in the model, we would have been required to respecify the model by removing at least one of the correlated predictors. This has its limitations since there is no obvious choice for which predictor to remove. Further, if the wrong predictor is removed, it may reduce the ability of the model to accurately make predictions and there is always the possibility that the final model would have just as much collinearity as the original data.

References

[1] Kim, J. I., & Kim, G. (2016). Country-Level Socioeconomic Indicators Associated with Healthy Life Expectancy: Income, Urbanization, Schooling, and Internet Users: 2000–2012. Social Indicators Research, 129(1), 391–402. https://doi.org/10.1007/s11205-015-1107-2

[2] Lung, T., Jan, S., Tan, E.J.*et al.*Impact of overweight, obesity and severe obesity on life expectancy of Australian adults.*Int J Obes***43**, 782–789 (2019). https://doi.org/10.1038/s41366-018-0210-2

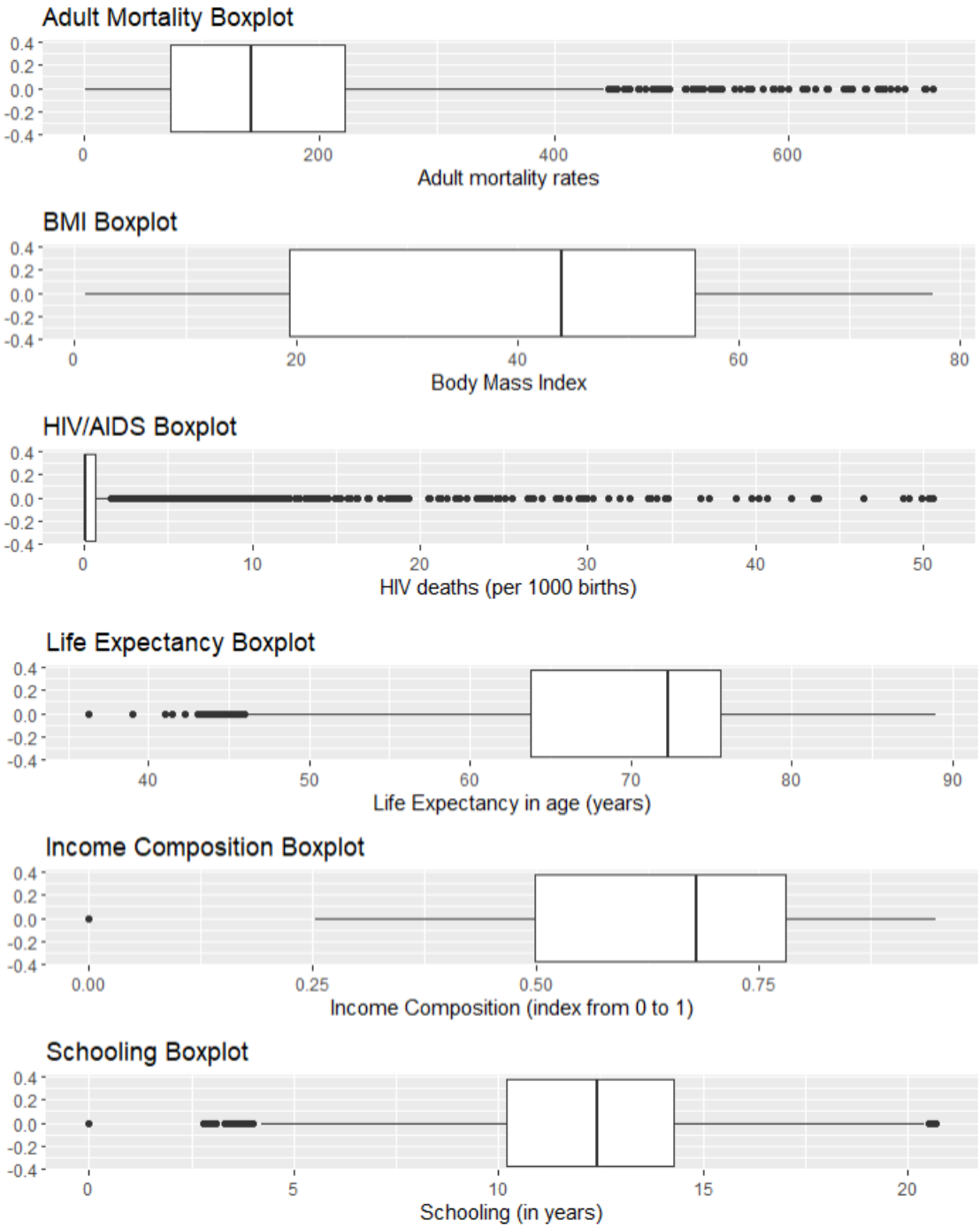
[3] Harrison, K. M., Song, R. P., & Zhang, X. P. (2010, January). Life Expectancy After HIV Diagnosis Based on National HIV Surveillance Data From 25 States, United States. *JAIDS Journal of Acquired Immune Deficiency Syndromes, 53*(1), 124-130. doi:10.1097/QAI.0b013e3181b563e7

[4] *Kaggle*. (n.d.). Retrieved from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

APPENDIX

Plot A

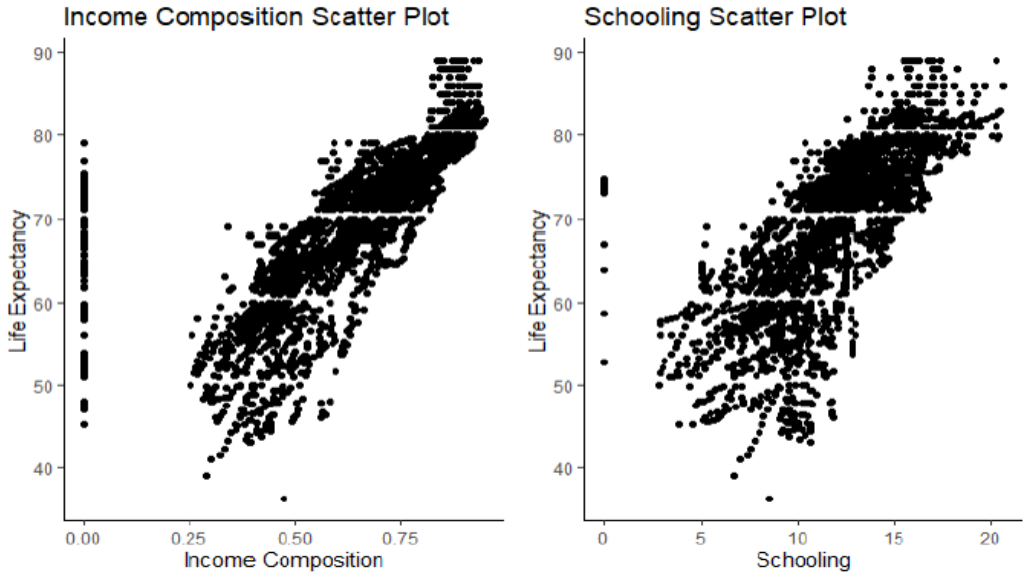
Boxplots of all the variables under consideration

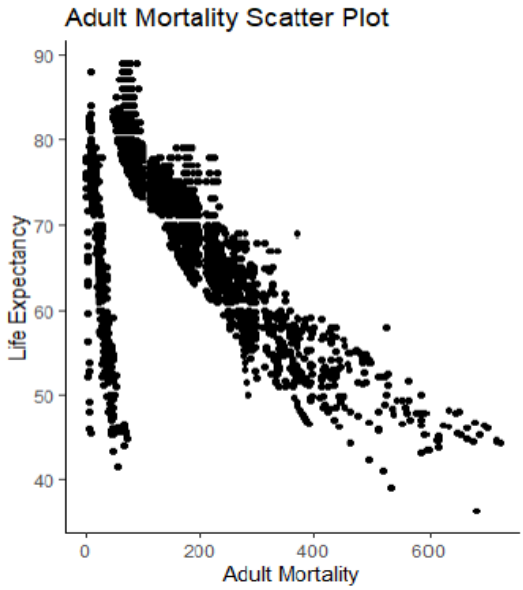


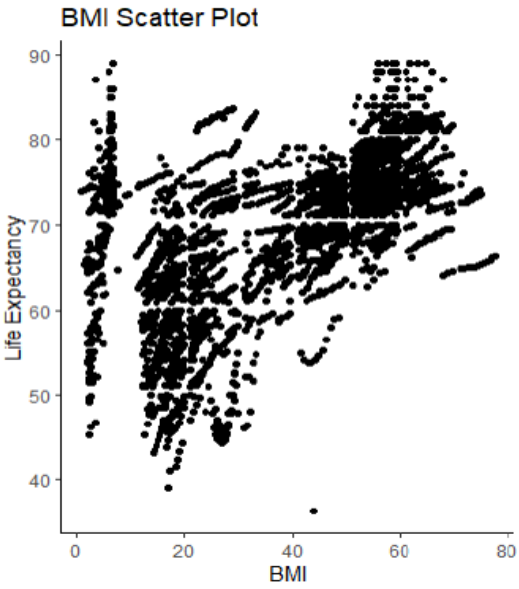
Appendix

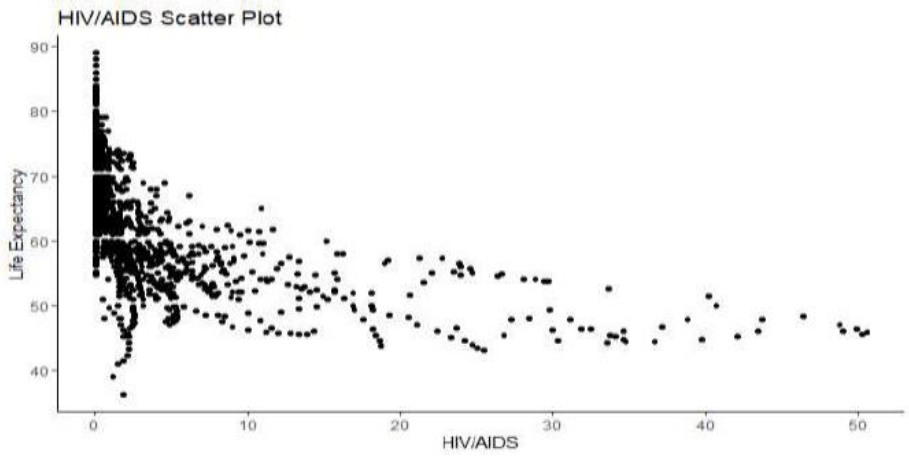
Plot B

Scatter plots of Life Expectancy vs Predictor variables









Appendix

Plot C

Normal QQ-Plots and Residual-vs-Fitted Plots of the predictor variables