**Final project: model predictors of global mean life expectancy**

Group3: Michael Bania, Catherine Chan, Hugh Ellerman, John Nguyen

**Summary:**

Global mean life expectancy is a metric of population well-being that is under continuous investigation by organizations such as the World Health Organization (WHO). The present study uses the WHO life expectancy dataset to ask the research question: which factors in the WHO dataset best predict global mean life expectancy? We selected variables with an all-possible subsets method to identify the best candidate models for each number of predictors. We then evaluated models according to Akaike’s information criterion, corrected Akaike’s information criterion, Bayesian information criterion, and adjusted R2. After selecting the best candidate model, we calculated variance inflation factor to control for multicollinearity. In the event of a multicollinear model, we repeated the process until we arrived at a valid model. We then evaluated and transformed our model to meet the assumptions of linear models. We propose the model: *Life Expectancy = Adult Mortality + (Hepatitis B)3 + (Polio)3 + Income + log(HIV/AIDS) + log(GDP) + log(Thinness 5-9) + error*

Of the variables in this final model, *Adult Mortality*, *Income, log(HIV/AIDS)*, and *log(Thinness 5-9)* were significant, while *(Hepatitis B)3, (Polio)3* and *log(GDP)* were insignificant. Adult mortality, log(HIV/AIDS) and log(Thinness 5-9) decreased life expectancy on average, with all other variables held constant (β = -0.019, β = -2.35, β = -1.34, respectively). Life expectancy increased with Income on average with all other variables held constant (β = 30.11). These results suggest that mean life expectancy is sensitive to variables that are related to material resources and mortality of youth. Efforts to increase global mean life expectancy would do well to increase availability of vaccines and economic opportunities globally.

**Introduction:**

The factors that lead to long lifespan are elusive and multifarious, yet they are critically important for understanding human population dynamics and well-being. While individual longevity can be attributed to any number of proximate and ultimate causes, the factors that lead to long-lived populations are more tractable. Even so, these factors are not fully understood and are the subject of continuous study. To better understand predictors of population lifespan, the World Health Organization (WHO) collected data globally on several potential determinants of average lifespan in different countries.

The WHO life expectancy dataset consists of 22 variables with nearly 3,000 observations of longevity for 193 countries from 2000-2015. This dataset includes 2 categorical variables: Country and Status, which is a binary predictor variable that indicates whether the country is considered developed or developing. The response variable, average life expectancy, is represented in years. The remainder are numeric predictor variables, described as follows:

Adult mortality: Probability of dying between age 15 and 60 years of age, represented as a number between 0 and 1000, Infant deaths: Number of infant deaths per 1000 population, Alcohol: Recorded per capita consumption of alcohol, in liters of pure alcohol, Percentage expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita, Hepatitis B: Immunization coverage of 1-year olds for Hepatitis B (%), Measles: Number of reported cases per 1000 population, BMI: average body-mass index of population, Under-five deaths: Number of under-five deaths per 1000 population, Polio: Polio vaccine coverage among 1-year olds (%), Total expenditure: General government expenditure on health as a percentage of total government expenditure, Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%), HIV/AIDS: Deaths per 1,000 live births from HIV/AIDS (0-4 years old), GDP: Gross domestic product per capita in USD, Population: Population of country, Thinness 1-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19 (%), Thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9 (%), Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1), Schooling: average number of years in school.

From this dataset, we hope to identify the variables that best predict mean life expectancy globally for the year 2015 in order to better understand the factors that contribute to mean life expectancy.

**Methods:**

*Data cleaning and manipulation*

The data we will be analyzing is a subset of the dataset from WHO. The dataset has been manipulated and filtered to contain countries that contain information on all variables for the year 2015 (in other words, we have no missing data). On initial visual inspection, the data showed some missing values. From R, we dropped three variables: alcohol, total expenditure, and percent expenditure. We were left with 16 predictor variables for life expectancy. The data is available for 130 Countries for the following variables:

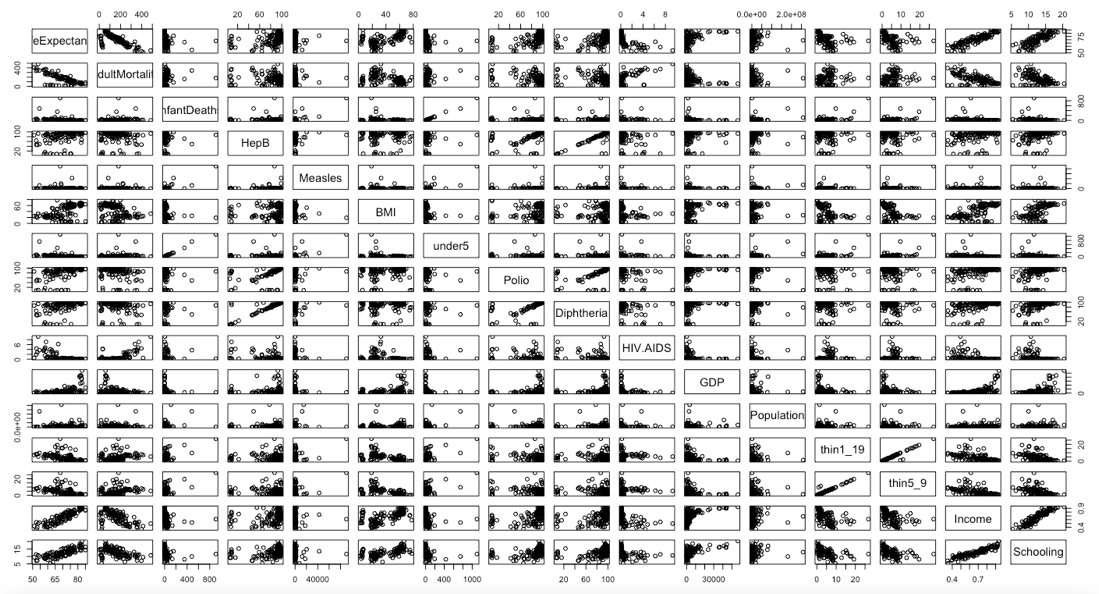
Y = Life Expectancy, X1 = Adult mortality, X2 = Infant deaths, X3 = Hepatitis B, X4 = Measles, X5 = BMI, X6 = Under-five deaths, X7 = Polio, X8 = Diphtheria, X9 = HIV/AIDS, X10 = GDP, X11 = Population, X12 = Thinness 1-19 years, X13 = Thinness 5-9 years, X14 = Income composition of resources, X15 = Schooling, X16 = Status.

The data is available at <https://www.kaggle.com/kumarajarshi/life-expectancy-who>.

*Data visualization*

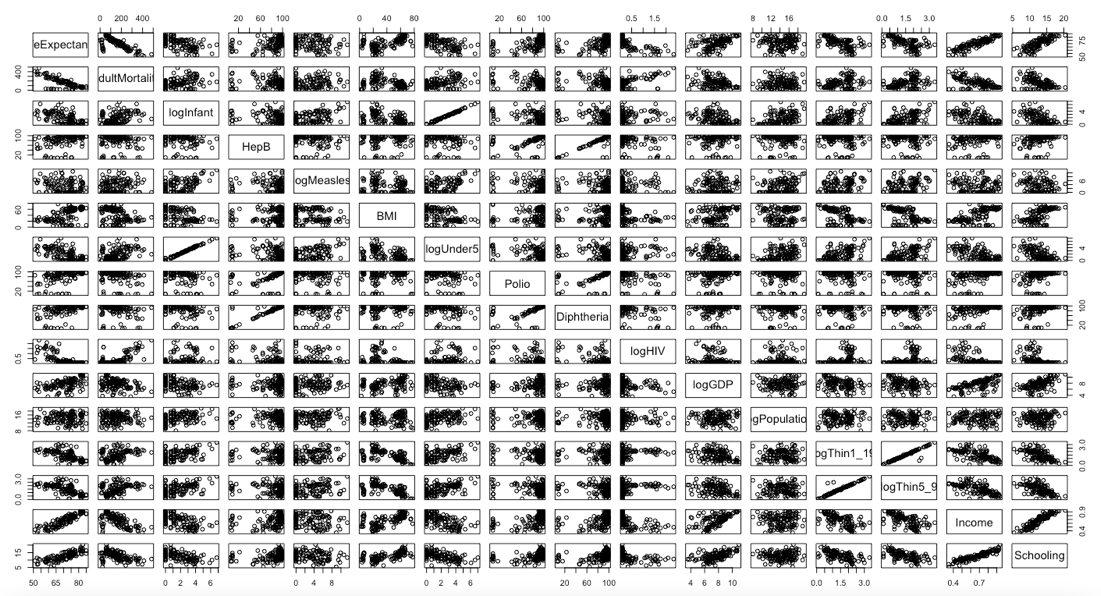
Fitting a regression model to the data with all predictor variables in the model, we get the initial model:

Y = β0 + β1X1 + β2X2 + + β3X3 + β4X4 + β5X5 + β6X6 + β7X7 + β8X8 + β9X9 + β10X10 + β11X11 + β12X12 + β13X13 + β14X14 + β15X15 + β16X16 + e



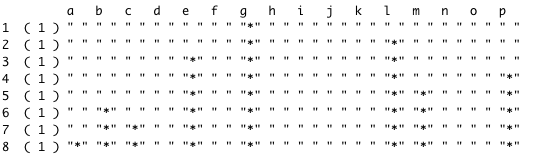
From the data visualization, we see that our data is heavily skewed, so we consider doing early transformations to improve the model. Looking at the pairs() function in R, we notice some obvious log-transformations that we could perform to immediately improve our data. After performing log-transformations on infant deaths, measles, under 5 deaths, HIV/AIDS, GDP, population, thinness 1-19, and thinness 5-9 we propose the model:

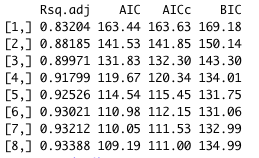
Y = β0 + β1X1 + B2log(X2) + β3X3 + β4log(X4) + β5X5 + β6log(X6) + β7X7 + β8X8 + β9log(X9) + β10log(X10) + β11log(X11) + β12log(X12) + β13log(X13) + β14X14 + β15X15 + β16X16 + e



Using variable selection procedures, we marginally identify all the possible candidate predictors and reduce the number of predictors in our model. Then by diagnostics transformations, we will check whether the assumptions for a valid linear model are satisfied or not.

Performing all possible subsets, we get the selected variables based on R^2 adj , AIC, AICc and BIC for fixed number of predictors.





Using R2 adj, AIC, AICc , and BIC, we see the “best” model is size 8 since it has the largest R2 adj and the smallest AIC, AICc.

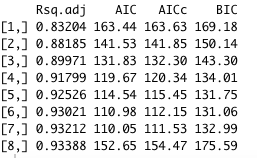
Y = β0 + β1X16 + β2X1 + β3X3 + β4X7 + β5X14 + β6log(X9) + β7log(X10) + β8log(X13) + e

Checking for multicollinearity in the selected model, we see that there is collinearity issue within the model, so we drop the predictor income and re-estimate the model.

Y = β0 + β1X16 + β2X1 + β3X3 + β4X7 + β5log(X9) + β6log(X10) + β7log(X13) + e



Re-estimating the model, we now see that the best model is now size 7.



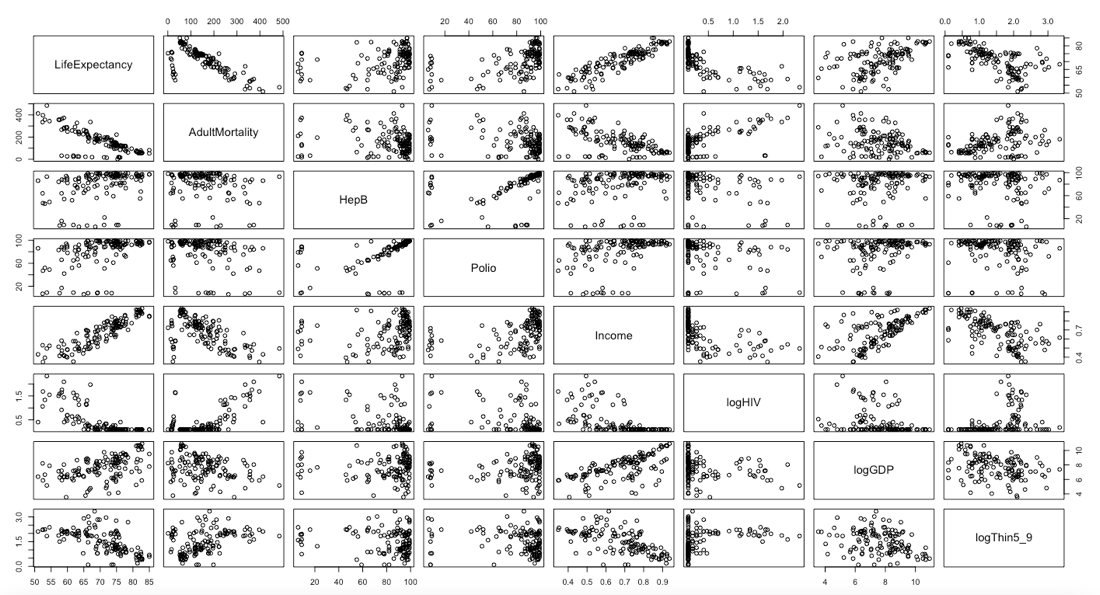
Again, checking for multicollinearity, we see that there are no collinearity issues within the model of size 7.



From variable selection, we propose the model:

Y = β0 + β1X1 + β2X3 + β3X7 + β4X14 + β5log(X9) + β6log(X10) + β7log(X13) + e

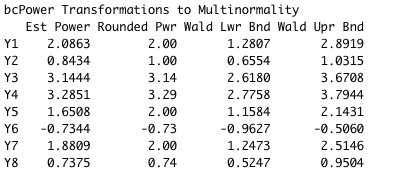
Although we performed log transformations early on, looking at the plots, the data is still skewed.



Based on the R-output by using “powerTransform”, we propose the model

Y = β0 + β1X1 + β2X3^3 + β3X7^3 + β4X14 + β5log(X9) + β6log(X10) + β7log(X13) + e

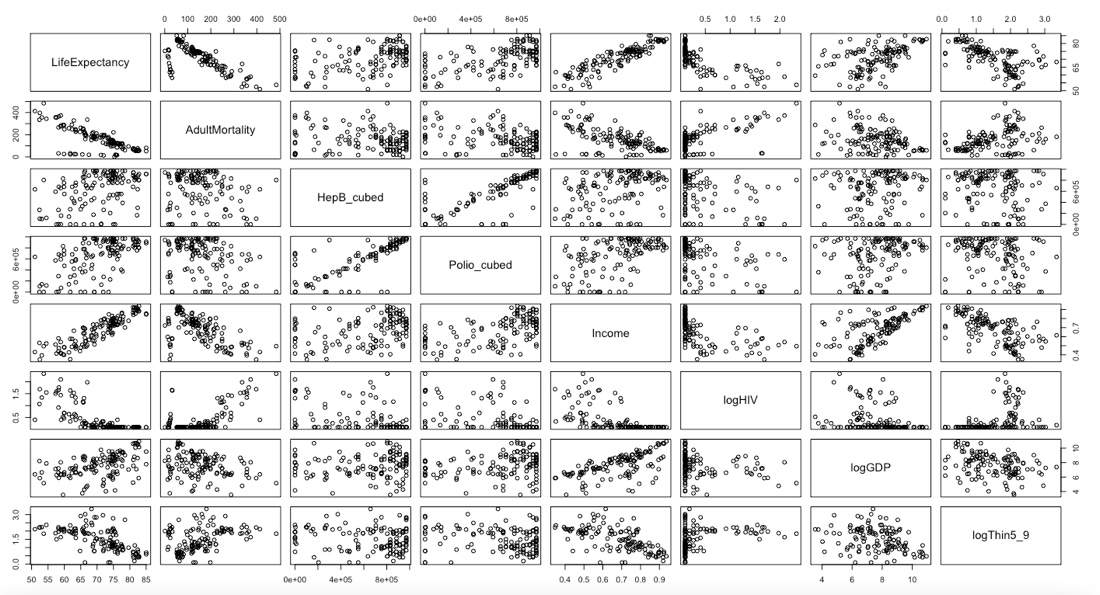
Note: we chose not to transform income since the diagnostic plots already met the assumptions, and we also did not perform any additional transformations to the variables that were log transformed early on. And to retain interpretability of the model and data, we also did not transform life expectancy.



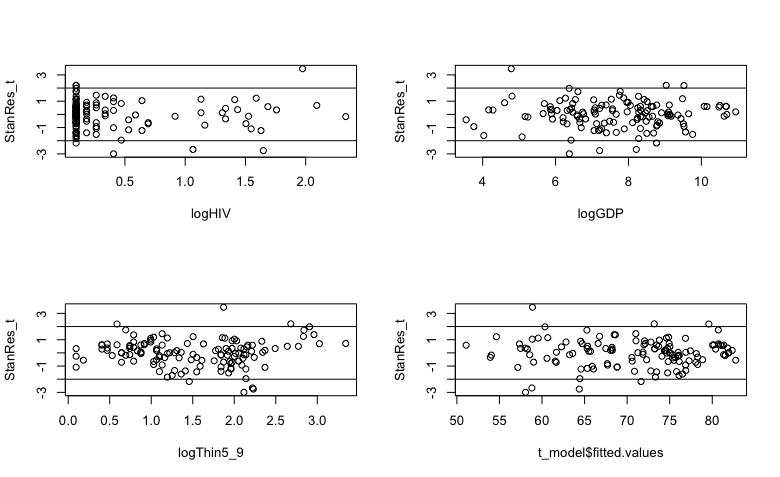
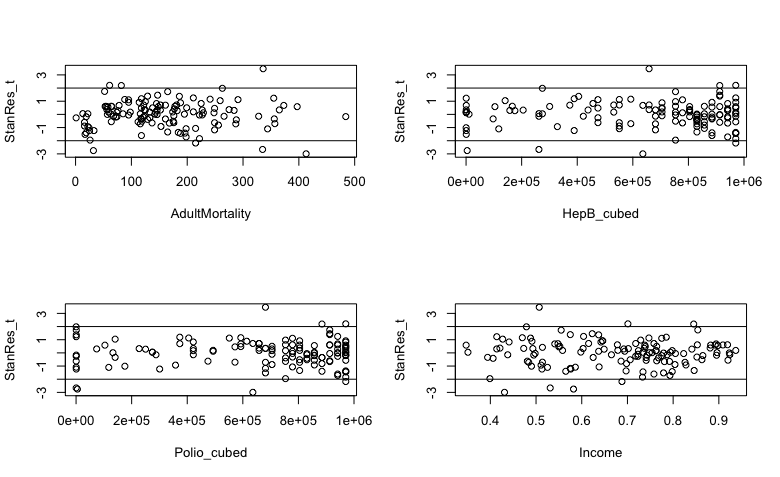
Checking our transformed model for multicollinearity, we see that all VIFs are under 5 so the regression coefficients are adequately estimated.

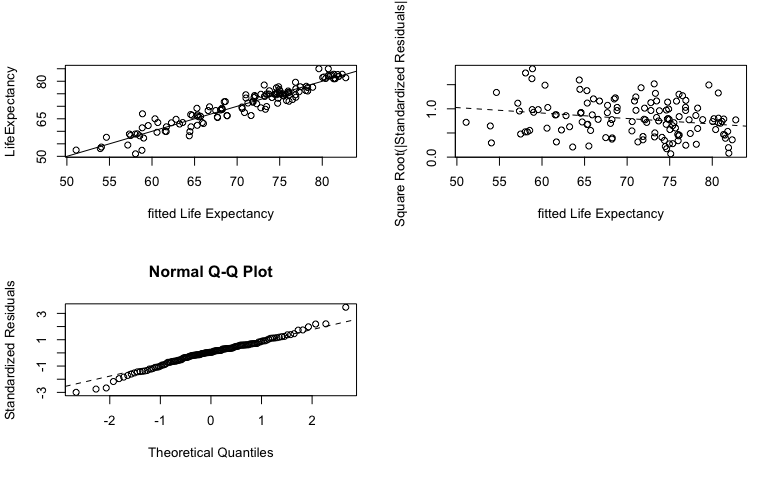


As we can see the data is much better now.



Looking at the diagnostics plots to check the assumptions of a valid multiple linear regression model are met.

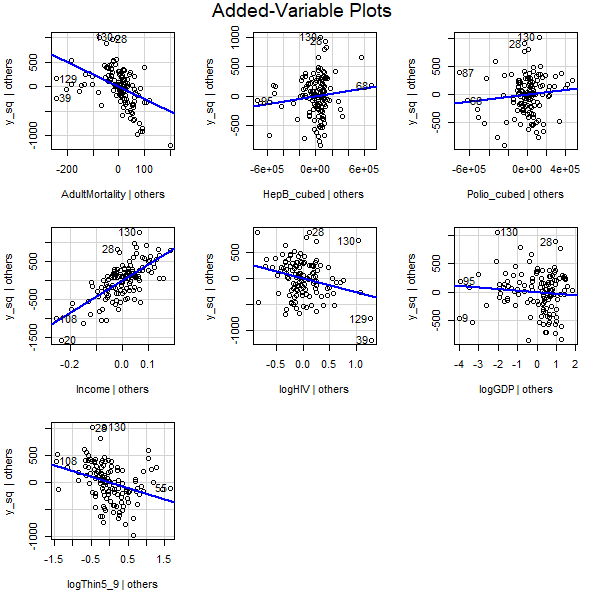


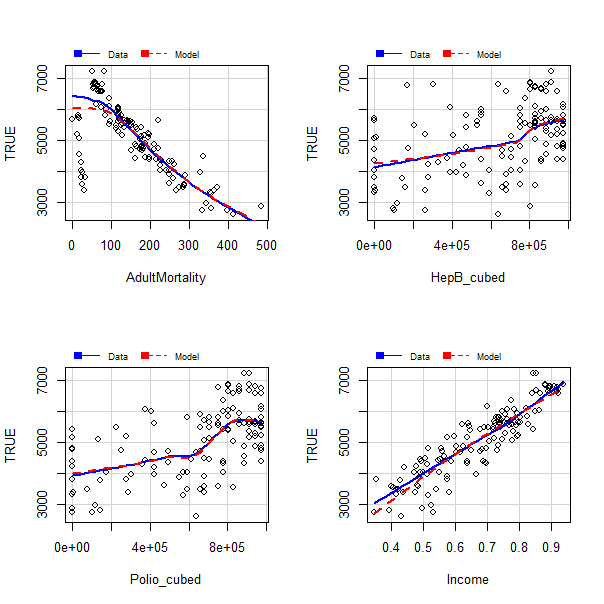
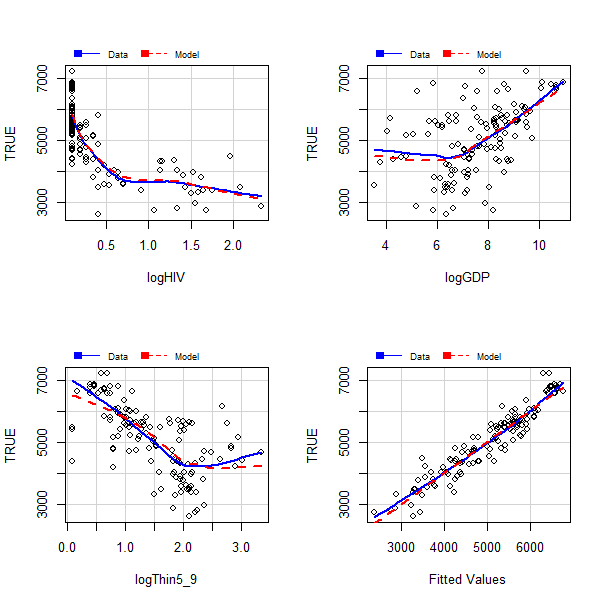


There seems to be no obvious pattern in most standardized residual plots. The scatter plot for Life Expectancy shows a good fit. Residuals show near-random behavior but a slightly decreasing trend. The scale location plot shows constant variance. The QQ plot shows the random error is close to normal distribution besides for a few points that skew away from the fit towards both tail ends. Except for some small issues, the diagnostics plots indicate that the current model is valid.

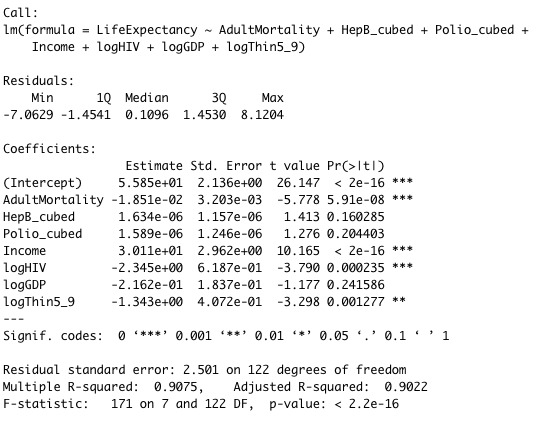
*Model validation*

The included added variable plots display the effect of each predictor variable given the others. These plots also visually display those observations that influentially affect the slope coefficient estimators of each predictor variable. Based on the Marginal Model plots, our model is accurate at capturing the mean of life expectancy. The lines for the data and model follow a very similar trajectory, hence providing evidence that the model is valid. Some plots display minor deviation in the tail ends such as in adult mortality, income, logGDP, and logThin5\_9. Other plots also show deviation in the mid-range areas, particularly where there are changes in slope of the data fit. Overall, the marginal model plots show that our selected model sufficiently represents the mean of life expectancy.





**Results and Discussion:**



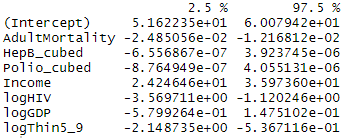
Based on our selection criteria, we propose the model:

Life Expectancy = β0 + β1Adult Mortality + β2(Hepatitis B)3 + β3(Polio)3 + β4Income + β5log(HIV/AIDS) + β6log(GDP) + β7log(Thinness 5-9) + error

The model summary output details the significant variables which include adult mortality, income, log(HIV/AIDS), and log(Thinness 5-9). The model has an R2 of 0.9022 and a p-value less than 0.001.

Through our calculations and model selection, we were able to find variables that were most associated with life expectancy for the year 2015. Income had a relatively large effect on life expectancy with a coefficient of 30.11. Influential variables with negative effects on life expectancy were adult mortality, income, log(HIV/AIDS), log(GDP), and log(Thinness 5-9). Other small positive effects were variables (Hepatitis B)3 and (Polio)3. Although we were able to find the predictor variables of life expectancy and their respective influences, there are several other factors to consider when interpreting the model.

Below are the confidence intervals for all coefficients within our final model. Hence, we are 95% confident that the true coefficient for AdultMortality is (-0.02485, -0.01217). We are also 95% confident that the true coefficient for Income is (24.27, 35.97).



**Discussion:**

According to our analysis, we found the most important predictor variables associated with life expectancy are Adult mortality, income, log(HIV), and log(Thinness5-9). Nonsignificant variables included (Hepatitis B)3, (Polio)3, and log(GDP).

The model indicated that immunization to disease has an important effect on life expectancy since Polio\_cubed, HepB\_cubed, and HIV/AIDS were variables included in the final model. A reason for why immunization of these diseases would have a statistically significant impact on life expectancy could be that the people who are impacted by the diseases in our model are typically younger. According to the WHO, people who are at risk for polio are children under the age of 5. Countries with a lower percentage of one year-old vaccinated for polio will have a greater number of children die from the disease and thus will see a greater childhood mortality rate as a result. This will decrease the population’s average life expectancy. Hepatitis B also impacts infants and small children heavily as 80-90% of infants who become infected with Hepatitis B during the first year of their life will develop chronic infections and 30-50% of children infected before turning 6 will develop chronic infections (WHO, 2019). These numbers help provide context for why lack of immunization can decrease a population’s life expectancy and highlight the importance that vaccines have with regards to public health. Increased access to vaccines in countries with higher mortality rates from Hepatitis B and Polio would increase mean life expectancy globally.

Our model also shows that economic factors such as income composition and GDP played a role in the global life expectancy. GDP measures the total monetary value of all products and services produced within a country. One reason why an increase in GDP would result in a higher life expectancy is that as a country increases their economic production, the amount of spendable income also increases. People begin to have their basic needs fulfilled such as clean water, nutritious food, and sanitary living conditions. These measures increase the life expectancy of an individual. Along with covering their basic needs, countries with higher GDP and spendable income may have more resources to devote to medical care and health care facilities.

The prevalence of thinness among children aged 5-9 is a significant factor in our model. One reason for this finding could be that individuals who are considered thin are malnourished and are not receiving the necessary nutrients to maintain a healthy life.

Adult Mortality is a statistically significant predictor in our final model and as the adult mortality increased, the life expectancy of that country decreased. Since a higher adult mortality rate means that a higher percentage of people aged 15-60 are dying, this lowers the average life expectancy for the entire population within that country.

The dataset used in this analysis resulted in some challenges in identifying variables most associated with life expectancy in year 2015. Although the year was specific to our question, using only one year limits our understanding of how consistent the findings are, and the overall purpose of better understanding the determinants of population life expectancy. It is possible that this year was not representative of life expectancy predictors. Another challenge in this dataset was missing data for variables potentially associated with life expectancy such as Alcohol, Total Expenditure, and Percent Expenditure. As a result, the effects of these are unknown in this study.

In addition, challenges included variables that have disproportionate data and effects on life expectancy. These include infant deaths, under age five deaths, and adult mortality as they directly influence life expectancy. Variables that may be better suited to our question would be those causes of deaths. Although two of these variables were not included in our final model, the process of model selection included these and may have had an influence on the other predictors.