MDA 9159A Statistical Modelling
Final Project Report

Linear Model on Life Expectancy

Presented by:
Bowen Wang (bwang489, 251139226)
Runcong Wu (rwu252, 251148344)

Introduction

Life expectancy is a statistical measure of the average time in years a person is expected to live. It is a useful benchmark in the financial world, including life insurance, pension planning, social security benefits, to determine life insurance premiums, retirement, and annuity payments. The Worldwide Health Organization (WHO) has published a dataset about people's life expectancy in different countries over the years 2000 to 2015 subject to several society development benchmarks including economics, medical technology, and healthcare research. Therefore, the goal of our project is to predict the life expectancy given various conditions.

Life expectancy is a popular research topic, many researchers try to find relevant factors to improve longevity. Some are focused on economic factors such as income and Gross Domestic Product. Research showed higher income was associated with greater longevity, and life expectancy was correlated with health conditions and geographic factors (Chetty 2016). Some are focused on medical intervention for patients because it is critical to determine the gains in life expectancy as the most important outcome from medical interventions (Wright 1998). Moreover, disease-specific studies on life expectancy are becoming increasingly prevalent in recent years. Neumayer analyzed the effect of HIV/AIDS on life expectancy as well as infant and child survival rates (Neumayer 2004). Wang developed a decision model to predict the cost and life expectancy for follicular lymphoma patients (Wang 2018).

Our dataset from WHO contains 2938 rows and 20 predictors. The 20 predictors were divided into four main categories: immunization related factors, mortality factors, economic factors, and social factors. The description of each variable is listed below:

- Life expectancy: Life Expectancy in age
- Country: the country of the row of data from
- Year: the year of the data

Economic Factors:

- Status: Developed or Developing status of the country
- Percentage Expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita (%)
- Total Expenditure: General government expenditure on health as a percentage of total government expenditure (%)
- GDP: Gross Domestic Product per capita (in USD)
- Income Composition of Resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

Immunization Related Factors:

- Hepatitis B: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- Polio: Polio (Pol3) immunization coverage among 1-year-olds (%)
- Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)

Mortality Factors:

- Adult Mortality: Adult Mortality Rates of both sexes (number of dying between 15 and 60 years per 1000 population)
- Under-five Deaths: Number of under-five deaths per 1000 population
- Infant Deaths: Number of Infant Deaths per 1000 Population
- HIV/AIDS: Deaths per 1000 live births HIV/AIDS (0-4 years)

Social Factors:

- Alcohol: Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
- Measles: Number of reported cases per 1000 population
- BMI: Average Body Mass Index of entire population
- Population: Population of the country
- Thinness 10-19 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 (%)
- Schooling: Number of years of Schooling (years)

Summary Statistics & Data Visualization

In this section, we included tables and figures to visualize the distribution of the response variable (Life expectancy) and predictors. Income composition of resources, Hepatitis B, adult mortality, and schooling were chosen to represent the four main categories of the predictors.

Table I showed the summary statistics of the response variable and the four predictors. The summary statistics included mean, median, standard deviation, maximum, and minimal.

	Mean <dbl>	Median <dbl>	SD <dbl>	Max <dbl>	Min <dbl>
Life expectancy (in years)	69.2997571	71.700	8.7990452	89.000	44.0
Income composition of resources (0-1)	0.6315647	0.673	0.1831331	0.936	0.0
Hepatitis B (%)	79.2143291	89.000	25.6131770	99.000	2.0
Adult Mortality (over 1000 individuals)	168.2149362	148.000	125.3432466	723.000	1.0
Schooling (in years)	12.1198543	12.300	2.7961722	20.700	4.2

Table I: Summary Statistics of Life expectancy, Income composition of resources, Hepatitis B, Adult mortality, and Schooling

Figure I described the distribution of the response variable (life expectancy). We can see the distribution is slightly left-skewed, with the mean life expectancy of 69.3 years and median life expectancy of 71.7 years. It is clear to see the majority of data points ranged from 50 to 85. More than 300 observations' life expectancy was within the range of 72 to 74 years.

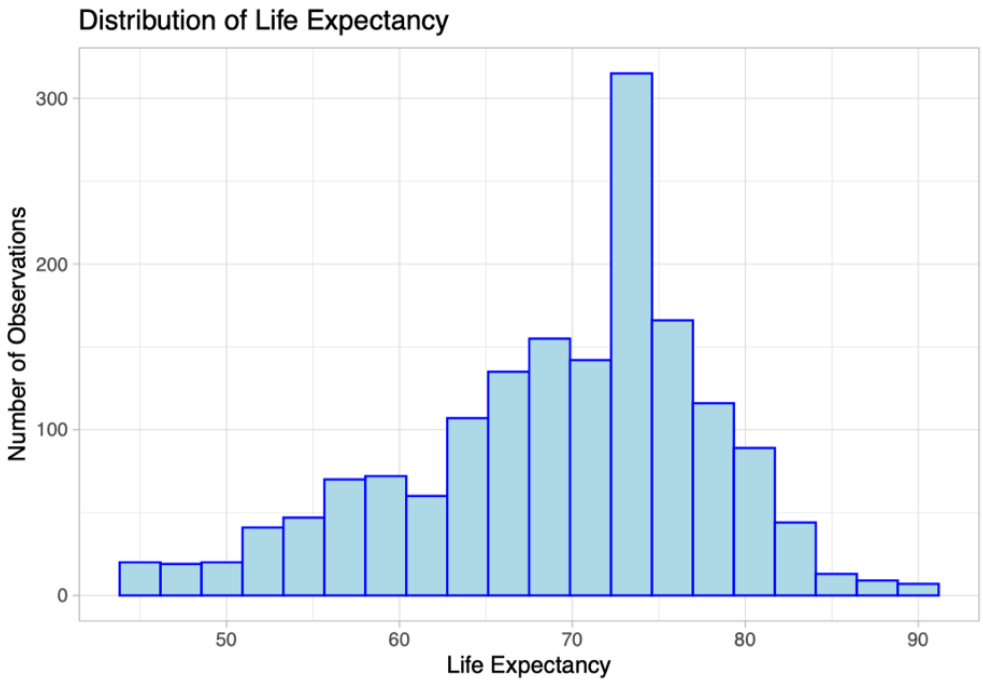


Figure I: Distribution of Life Expectancy

Figure II described the distribution of the income composition of resources. We can see the distribution is slightly left-skewed, with the mean income composition of 63% and median income composition of 67%. The mode occurred at the range from 70% to 75% of income composition, and around 500 observations were in this range.

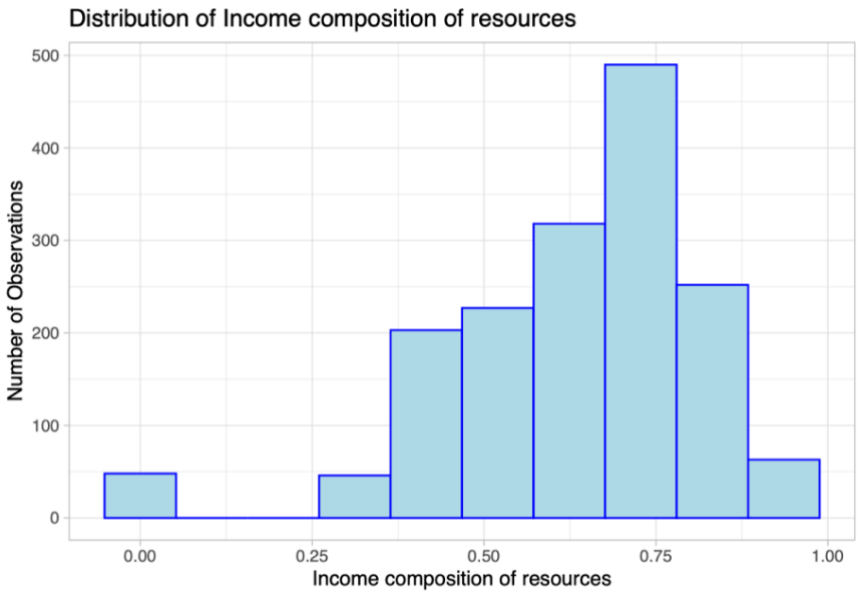


Figure II: Distribution of Income composition of resources

Figure III is a histogram describing the distribution of the Hepatitis B immunization coverage among 1-year-olds. The distribution was left-skewed with the mean coverage percentage of 79% and the median coverage percentage of 89%. The mode occurred at the range from 90% to 100%, and around 800 observations were in this range.

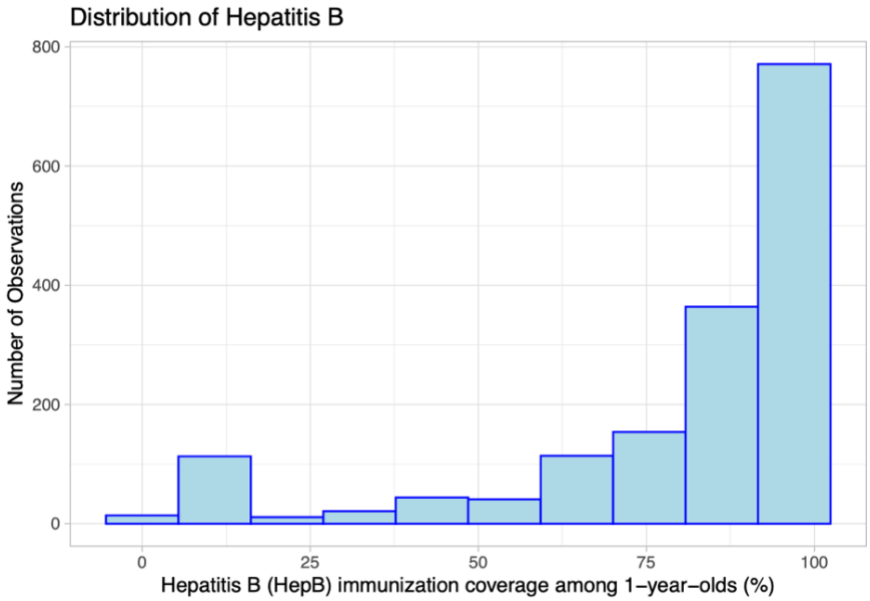


Figure III: Distribution of Hepatitis B

Figure IV is a histogram describing the distribution of adult mortality rates for both sexes over 1000 individuals. The distribution was right skewed with a mean of 168 and a median of 148. The mode occurred at the range from 100 to 200, and almost 600 observations were in this range.

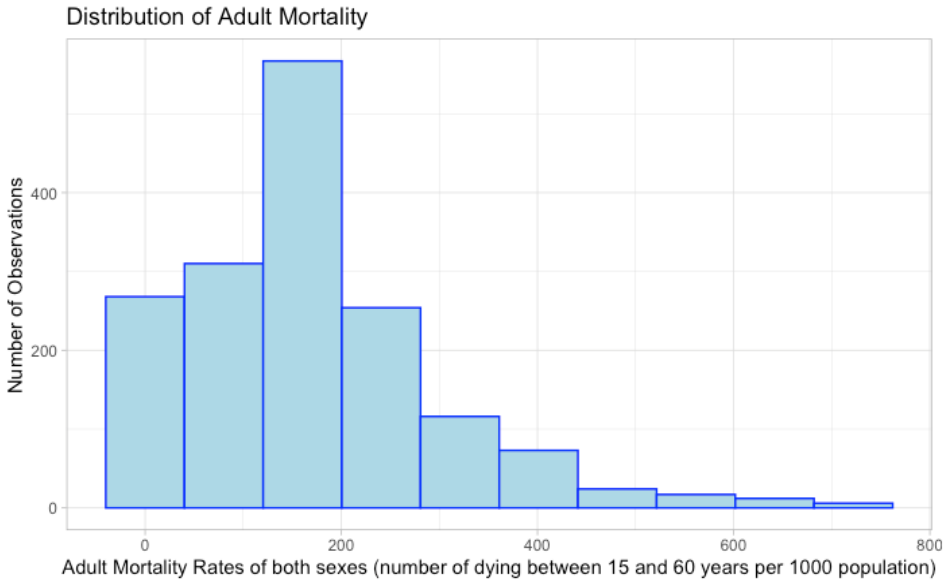


Figure IV: Distribution of Adult mortality

Figure V demonstrated the distribution of schooling in years. The distribution was symmetric with both the mean and median of 12 years. It is interesting to note that around 1300 observations had 10 years or more than 10 years of schooling.

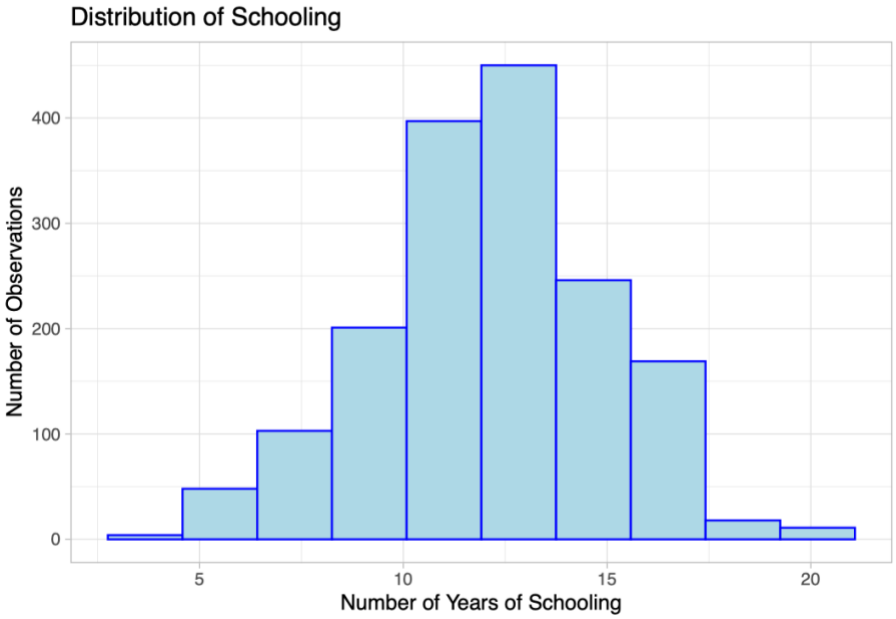


Figure V: Distribution of Schooling

Methodology

According to the general statistic theory and modeling process. We decided to break the experiment into 3 parts: Data processing, Modelling and Validating & Adjusting. The second part and third part form a rotating cycle that once we finish validating & adjusting, we rebuild model again until the model is useful and reasonable. Next, we are going to introduce each part of our methodology in detail.

Data Processing

To make the dataset consistent and tidy, we firstly want to deal with null values in the dataset. By examining the null values, most of them are occurred in GDP and Population columns. As these values are hard to estimate and not reasonable to simply assign an average value of other data entry, we decide to remove all null values from the dataset. Under the help of *na.omit()* function, our dataset reduces to 1647 rows of consistent full data entries.

As we are interested in a general model instead of a model for a specific country, we decide to remove the Country column. After doing column removal, we notice there are only 2 data entries for year 2015. As we are going to use Year as a dummy predictor, all other year sections have hundreds of entries, concerning too few data may arise variation and bias problems, we decide to remove 2015 data as well.

For Year and Status columns, as they are categorical variables, we decide to treat them as dummy variables. Using *factor()* function to change the variable type so that the *lm()* function can automatically handle them into dummies.

Then we split the 80% of the dataset into training dataset and 20% of the data set into testing dataset. Define the life_expectancy in both training and testing dataset to be y and other variables to X.

After doing all these steps, we have set our dataset and ready for further modeling.

Modelling

We build our first model (Model I) by using all the variables to be our default model and help us to have a general idea about how each predictor works. There are 33 predictors in this model and the adjusted R-Squared comes to 0.8827. However, many of the variables are having extremely high p-values and all of the three assumptions (Linearity, Normality, Equal Variance) are violated, we decide to move forward and find if we can fix the violation.

We assume influential points in this dataset are measurement errors. Applying Cook's Distance method, we get 105 influential observations as influential points. Remove the influential points and rebuild the model on all 33 predictors and get our second model (Model II). Unfortunately, there isn't significant improvement after removing the influential points.

In our third model (Model III), we wonder if Box-Cox Transformation could help us solve the violation problem. We set the interval of lambda as 0 to 3 with 0.05 incremental steps to find the

optimal lambda that result in the highest log-likelihood. And use the optimal lambda build the Box-Cox transformed model. There still minor improvements can be noticed.

In our fourth model (Model IV), we want to find what is the best subset of predictors. So that we apply Stepwise Selection with BIC in this model and build model by using the result from Stepwise Selection. The number of predictors dropped to 8 but the adjusted R-Squared slightly dropped to 0.8405. However, the assumptions still are violated in this model.

While examining the pairs plot of Model IV, we find there is a strong collinearity between *under_five_deaths* and *infant_deaths*, to eliminate the variance and improve model performance, we drop *infant_deaths* and build our fifth model (Model V).

We add polynomial terms (quadratic and cubic) based on Model V to be our Model VI.

Our Model VII is based on Model VI as it has the best performance, and we want to simplify it by removing the redundant predictor. We apply stepwise method on both directions.

In our Model VIII, we use LASSO regression on all first order predictors.

In our Model IX, we use Ridge regression on all first order predictors.

Validating & Adjusting

As the linear regression is based on Normality, Equal Variance, and linearity assumption. We do validation on the three linear regression assumptions for each model we built. To validate the three linear regression assumptions, we use Breusch-Pagan test for equal variance test, and Shapiro-Wilk for normality test, and check the distribution of residuals in residual plot for linearity assumption.

After each model has been built, we use several methods to adjust our model. The methods include modify predictor sets, do transformation on predictors, and change regression models.

Results

Model I (Default Model Based on All Columns)

As our Model I is using every column in the dataset as predictors, the summary of the model and assumption test are shown in Figure VI. The adjusted R-Squared is 0.8438 while the p-value from F-test is less than $2.2*10^{-16}$. The adjusted R-squared indicate that the model explained most of the observations and the p-value indicates that at least 1 predictor is significant in this model. However, there are many non-significant predictors from the summary table includes GDP, Population, etc. The non-significant are redundant and increase the risk of overfitting.

Model I violates equal variance and normality assumptions as the p-values from Breusch-Pagan test and Shapiro-Wilk test are both less than 0.05, which support us to reject their H_0. However, according to the residual plot, the residuals are separated rough evenly on both side of 0, the linearity assumption holds.

```
Call:
lm(formula = Life_expectancy ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12.7939  -2.0709  -0.0313   2.2198  12.5197

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.565e+01  9.994e-01  55.681  < 2e-16 ***
Year2001                       -9.786e-01  6.759e-01  -1.448 0.147856
Year2002                       -1.436e+00  6.611e-01  -2.173 0.029997 *
Year2003                       -1.621e+00  6.283e-01  -2.580 0.009989 **
Year2004                       -1.596e+00  6.296e-01  -2.534 0.011384 *
Year2005                       -1.922e+00  6.256e-01  -3.072 0.002168 **
Year2006                       -1.659e+00  6.153e-01  -2.696 0.007102 **
Year2007                       -1.909e+00  6.086e-01  -3.136 0.001750 **
Year2008                       -2.326e+00  6.166e-01  -3.773 0.000169 ***
Year2009                       -1.713e+00  6.067e-01  -2.824 0.004819 **
Year2010                       -2.196e+00  6.087e-01  -3.608 0.000321 ***
Year2011                       -2.140e+00  6.068e-01  -3.528 0.000434 ***
Year2012                       -2.798e+00  6.118e-01  -4.574 5.26e-06 ***
Year2013                       -2.335e+00  6.089e-01  -3.834 0.000132 ***
Year2014                       -2.501e+00  6.185e-01  -4.044 5.57e-05 ***
StatusDeveloping               -7.691e-01  3.651e-01  -2.107 0.035333 *
Adult_Mortality                -1.707e-02  1.055e-03 -16.180  < 2e-16 ***
infant_deaths                   8.697e-02  1.200e-02   7.249 7.21e-13 ***
Alcohol                        -8.980e-02  3.755e-02  -2.391 0.016926 *
percentage_expenditure          3.845e-04  1.998e-04   1.924 0.054553 .
Hepatitis_B                     1.557e-03  4.900e-03   0.318 0.750668
Measles                        -2.392e-06  1.176e-05  -0.203 0.838913
BMI                             2.747e-02  6.529e-03   4.208 2.75e-05 ***
under_five_deaths              -6.581e-02  8.714e-03  -7.552 8.08e-14 ***
Polio                           5.394e-03  5.498e-03   0.981 0.326745
Total_expenditure               9.858e-02  4.443e-02   2.219 0.026679 *
Diphtheria                      7.529e-03  6.506e-03   1.157 0.247387
HIV_AIDS                       -4.314e-01  2.008e-02 -21.490  < 2e-16 ***
GDP                             1.410e-05  3.113e-05   0.453 0.650806
Population                      5.949e-10  2.230e-09   0.267 0.789701
thinness_1_19_years             3.574e-02  5.554e-02   0.644 0.519997
thinness_5_9_years             -6.637e-02  5.455e-02  -1.217 0.223964
Income_composition_of_resources 9.782e+00  9.063e-01  10.794  < 2e-16 ***
Schooling                       9.290e-01  6.563e-02  14.155  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.469 on 1284 degrees of freedom
Multiple R-squared:  0.8477,    Adjusted R-squared:  0.8438
F-statistic: 216.6 on 33 and 1284 DF,  p-value: < 2.2e-16
```

```
        studentized Breusch-Pagan test

data:  model1
BP = 146.74, df = 33, p-value = 2.741e-16
```

```
        Shapiro-Wilk normality test

data:  resid(model1)
W = 0.99355, p-value = 1.7e-05
```

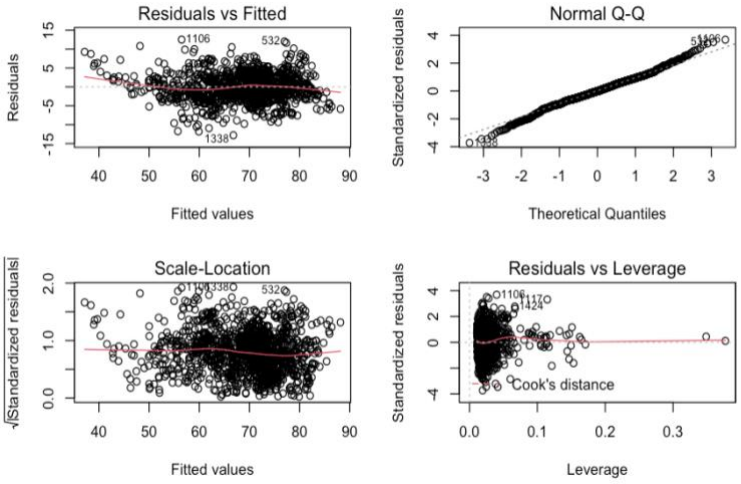


Figure VI: Result of Model I

Model II (All Columns with Removal of Influential Points)

Our Model II has minor improvement comparing to Model I, details are shown in Figure VII. The adjusted R-Squared increased to 0.8827 but there still many predictors are not significant. The redundancy problem still hasn't been solved by removing influential points.

Model II doesn't pass the Breusch-Pagan test and Shapiro-Wilk test under default alpha = 0.05 as well. However, the p-value of Shapiro-Wilk test significantly increased to 0.005, if we set a higher confidence level, say 99.5%, Model II could pass the normality test. Thus, removing the influential points has some positive effect on the normality of data. Then, according to the residual plot, the residuals are separated rough evenly on both side of 0, the linearity assumption holds.

```
Call:
lm(formula = Life_expectancy ~ ., data = train2)

Residuals:
   Min     1Q Median     3Q    Max
-8.623 -1.853 -0.042  2.004  8.015

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.552e+01  8.714e-01  63.706  < 2e-16 ***
Year2001                       -8.036e-01  5.899e-01  -1.362 0.173374
Year2002                       -1.104e+00  5.734e-01  -1.926 0.054396 .
Year2003                       -1.314e+00  5.389e-01  -2.438 0.014931 *
Year2004                       -1.448e+00  5.420e-01  -2.672 0.007633 **
Year2005                       -1.870e+00  5.384e-01  -3.474 0.000531 ***
Year2006                       -1.130e+00  5.308e-01  -2.129 0.033424 *
Year2007                       -1.771e+00  5.233e-01  -3.384 0.000738 ***
Year2008                       -2.100e+00  5.307e-01  -3.957 8.04e-05 ***
Year2009                       -1.556e+00  5.213e-01  -2.985 0.002892 **
Year2010                       -1.768e+00  5.234e-01  -3.378 0.000752 ***
Year2011                       -1.895e+00  5.214e-01  -3.635 0.000290 ***
Year2012                       -2.221e+00  5.261e-01  -4.221 2.61e-05 ***
Year2013                       -2.030e+00  5.238e-01  -3.877 0.000112 ***
Year2014                       -2.295e+00  5.296e-01  -4.333 1.59e-05 ***
StatusDeveloping               -3.151e-01  3.078e-01  -1.024 0.306128
Adult_Mortality                -1.953e-02  9.836e-04 -19.860  < 2e-16 ***
infant_deaths                   9.150e-02  1.137e-02   8.047 2.03e-15 ***
Alcohol                        -7.088e-02  3.187e-02  -2.224 0.026341 *
percentage_expenditure          3.993e-04  1.677e-04   2.382 0.017391 *
Hepatitis_B                    -9.206e-04  4.194e-03  -0.219 0.826302
Measles                         3.687e-07  9.849e-06   0.037 0.970149
BMI                             2.211e-02  5.560e-03   3.976 7.43e-05 ***
under_five_deaths              -6.915e-02  8.286e-03  -8.345  < 2e-16 ***
Polio                           3.286e-03  4.827e-03   0.681 0.496107
Total_expenditure               2.163e-01  3.842e-02   5.630 2.24e-08 ***
Diphtheria                      9.820e-03  5.695e-03   1.724 0.084912 .
HIV_AIDS                       -4.154e-01  2.068e-02 -20.083  < 2e-16 ***
GDP                             4.135e-06  2.575e-05   0.161 0.872461
Population                      4.006e-10  1.978e-09   0.203 0.839527
thinness_1_19_years             2.422e-02  4.695e-02   0.516 0.606071
thinness_5_9_years             -6.655e-02  4.619e-02  -1.441 0.149883
Income_composition_of_resources  1.288e+01  9.228e-01  13.953  < 2e-16 ***
Schooling                       7.258e-01  6.018e-02  12.060  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.837 on 1200 degrees of freedom
Multiple R-squared:  0.8858,    Adjusted R-squared:  0.8827
F-statistic: 282.1 on 33 and 1200 DF,  p-value: < 2.2e-16
```

```
        studentized Breusch-Pagan test

data:  model2
BP = 136.32, df = 33, p-value = 1.632e-14


        Shapiro-Wilk normality test

data:  resid(model2)
W = 0.99638, p-value = 0.00555
```

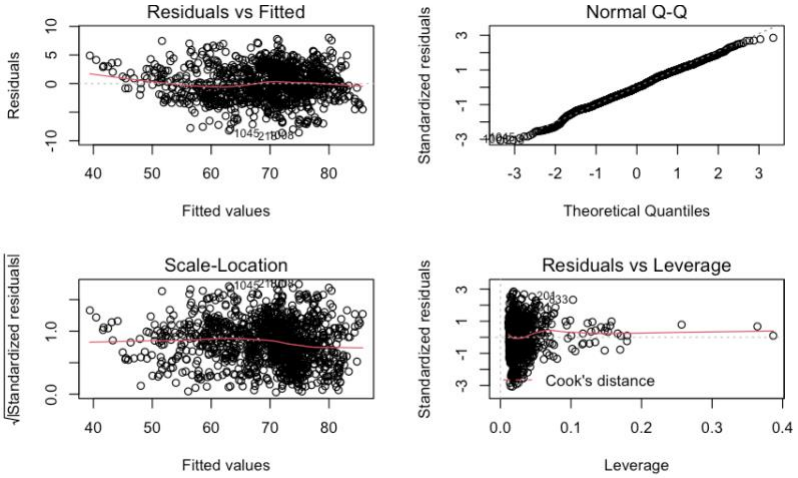
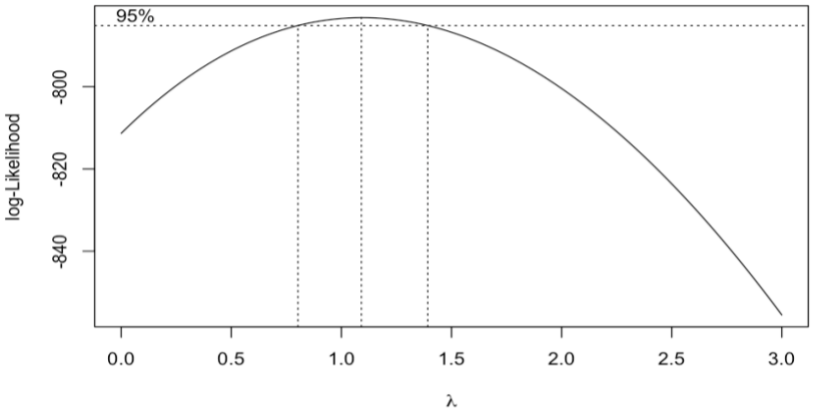


Figure VII: Result of Model II

Model III (All Columns with Box-Cox Transform)

Our Model III also has minor improvement comparing to Model I, details are shown in Figure VIII. As the optimal lambda tends to be around 1, the transform function is roughly g(Y) = Y-1, which is not a useful transformation. The adjusted R-Squared remains 0.8438 and everything in the summary looks same as Model I.

Model III doesn't pass the Breusch-Pagan test and Shapiro-Wilk test under default alpha = 0.05 as well. Then, according to the residual plot, the residuals are separated rough evenly on both side of 0, the linearity assumption holds.




```
Call:
lm(formula = ((Life_expectancy^(lambda) - 1)/(lambda)) ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12.7939  -2.0709  -0.0313   2.2198  12.5197

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.465e+01  9.994e-01  54.680  < 2e-16 ***
Year2001                       -9.786e-01  6.759e-01  -1.448 0.147856
Year2002                       -1.436e+00  6.611e-01  -2.173 0.029997 *
Year2003                       -1.621e+00  6.283e-01  -2.580 0.009989 **
Year2004                       -1.596e+00  6.296e-01  -2.534 0.011384 *
Year2005                       -1.922e+00  6.256e-01  -3.072 0.002168 **
Year2006                       -1.659e+00  6.153e-01  -2.696 0.007102 **
Year2007                       -1.909e+00  6.086e-01  -3.136 0.001750 **
Year2008                       -2.326e+00  6.166e-01  -3.773 0.000169 ***
Year2009                       -1.713e+00  6.067e-01  -2.824 0.004819 **
Year2010                       -2.196e+00  6.087e-01  -3.608 0.000321 ***
Year2011                       -2.140e+00  6.068e-01  -3.528 0.000434 ***
Year2012                       -2.798e+00  6.118e-01  -4.574 5.26e-06 ***
Year2013                       -2.335e+00  6.089e-01  -3.834 0.000132 ***
Year2014                       -2.501e+00  6.185e-01  -4.044 5.57e-05 ***
StatusDeveloping               -7.691e-01  3.651e-01  -2.107 0.035333 *
Adult_Mortality                -1.707e-02  1.055e-03 -16.180  < 2e-16 ***
infant_deaths                   8.697e-02  1.200e-02   7.249 7.21e-13 ***
Alcohol                        -8.980e-02  3.755e-02  -2.391 0.016926 *
percentage_expenditure          3.845e-04  1.998e-04   1.924 0.054553 .
Hepatitis_B                     1.557e-03  4.900e-03   0.318 0.750668
Measles                        -2.392e-06  1.176e-05  -0.203 0.838913
BMI                             2.747e-02  6.529e-03   4.208 2.75e-05 ***
under_five_deaths              -6.581e-02  8.714e-03  -7.552 8.08e-14 ***
Polio                           5.394e-03  5.498e-03   0.981 0.326745
Total_expenditure               9.858e-02  4.443e-02   2.219 0.026679 *
Diphtheria                      7.529e-03  6.506e-03   1.157 0.247387
HIV_AIDS                       -4.314e-01  2.008e-02 -21.490  < 2e-16 ***
GDP                             1.410e-05  3.113e-05   0.453 0.650806
Population                      5.949e-10  2.230e-09   0.267 0.789701
thinness_1_19_years             3.574e-02  5.554e-02   0.644 0.519997
thinness_5_9_years             -6.637e-02  5.455e-02  -1.217 0.223964
Income_composition_of_resources 9.782e+00  9.063e-01  10.794  < 2e-16 ***
Schooling                       9.290e-01  6.563e-02  14.155  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.469 on 1284 degrees of freedom
Multiple R-squared:  0.8477,    Adjusted R-squared:  0.8438
F-statistic: 216.6 on 33 and 1284 DF,  p-value: < 2.2e-16
```

```
        studentized Breusch-Pagan test

data:  model3
BP = 146.74, df = 33, p-value = 2.741e-16


        Shapiro-Wilk normality test

data:  resid(model3)
W = 0.99355, p-value = 1.7e-05
```

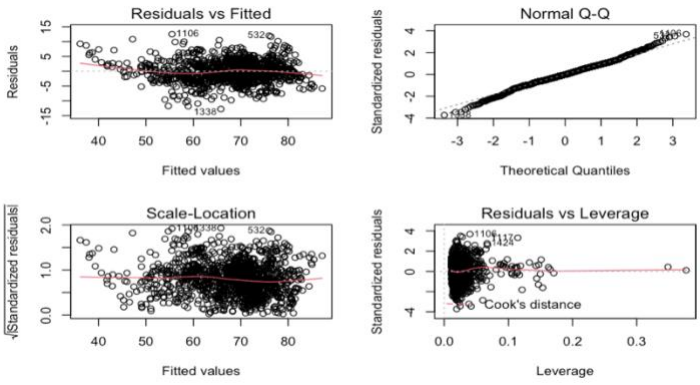


Figure VIII: Result of Model III

Model IV (Stepwise on Both Directions)

By using stepwise selection with BIC on both direction, Model IV is based on less predictors. The finalized predictors are: Schooling, Adult_Mortality, HIV_AIDS, Income_composition_of_resources, Percentage_expenditure, BMI, Under_five_deaths and Infant_deaths. The adjusted R-squared is: 0.8405 while the p-value of the F test is less than $2.2*10^{-16}$. More details are shown in Figure IX.

Model IV doesn't pass the Breusch-Pagan test and Shapiro-Wilk test under default alpha = 0.05 as well. According to the residual plot, the residuals are separated rough evenly on both side of 0, the linearity assumption holds.

```
Call:
lm(formula = Life_expectancy ~ Schooling + Adult_Mortality +
    HIV_AIDS + Income_composition_of_resources + percentage_expenditure +
    BMI + under_five_deaths + infant_deaths, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12.3959  -2.0884   0.0025   2.2437  11.9916

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.440e+01  5.902e-01  92.170  < 2e-16 ***
Schooling                       9.416e-01  6.084e-02  15.476  < 2e-16 ***
Adult_Mortality                -1.795e-02  1.047e-03 -17.134  < 2e-16 ***
HIV_AIDS                       -4.214e-01  1.980e-02 -21.283  < 2e-16 ***
Income_composition_of_resources  9.420e+00  8.787e-01  10.721  < 2e-16 ***
percentage_expenditure           4.784e-04  6.471e-05   7.392 2.56e-13 ***
BMI                              3.181e-02  6.074e-03   5.237 1.90e-07 ***
under_five_deaths               -7.273e-02  8.107e-03  -8.971  < 2e-16 ***
infant_deaths                    9.519e-02  1.096e-02   8.686  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.506 on 1309 degrees of freedom
Multiple R-squared:  0.8414,    Adjusted R-squared:  0.8405
F-statistic: 868.3 on 8 and 1309 DF,  p-value: < 2.2e-16
```

```
        studentized Breusch-Pagan test

data:  model4
BP = 123.66, df = 8, p-value < 2.2e-16
```

```
        Shapiro-Wilk normality test

data:  resid(model4)
W = 0.99281, p-value = 5.067e-06
```

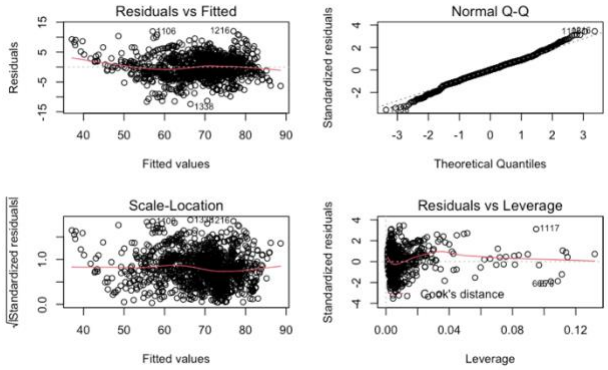
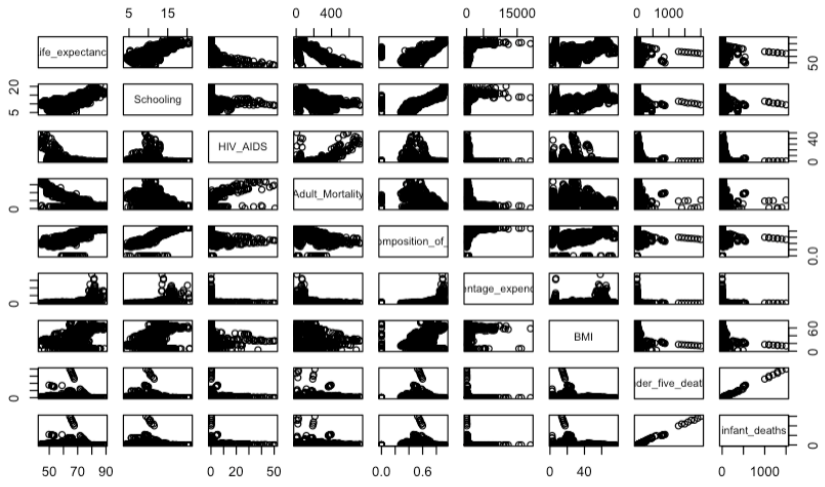


Figure IX: Result of Model IV

From the pairs plot, there is an obvious linear relationship between Under_five_deaths and Infant_deaths. We assume there could be a collinearity between them. To verify our assumption, we calculate Variance Inflation Factor of the Model IV, Under_five_deaths and Infant_deaths have extremely large VIF. Thus, in next Model we remove one of them.




```
                                Schooling
                                 3.061629
Income_composition_of_resources
                                 2.833374
                       under_five_deaths
                                170.381802
```

```
                      Adult_Mortality
                             1.944878
              percentage_expenditure
                             1.245016
                       infant_deaths
                           169.927707
```

Figure X: Result of Pairs Plot

Model V (Drop One Collinearity Predictor)

Figure XI shows the result of Model V. Comparing to Model IV, the adjusted R-squared drops to 0.8314. Other benchmarks are similar.

Model V doesn't pass the Breusch-Pagan test and Shapiro-Wilk test under default alpha = 0.05 as well. According to the residual plot, the residuals are separated rough evenly on both side of 0, the linearity assumption holds.

```
Call:
lm(formula = Life_expectancy ~ Schooling + HIV_AIDS + Adult_Mortality +
    Income_composition_of_resources + percentage_expenditure +
    BMI + under_five_deaths, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3713  -2.1066   0.0831   2.3501  11.6679

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.438e+01  6.067e-01  89.634  < 2e-16 ***
Schooling                       9.459e-01  6.254e-02  15.125  < 2e-16 ***
HIV_AIDS                       -4.233e-01  2.036e-02 -20.793  < 2e-16 ***
Adult_Mortality                -1.940e-02  1.063e-03 -18.249  < 2e-16 ***
Income_composition_of_resources 9.901e+00  9.015e-01  10.983  < 2e-16 ***
percentage_expenditure          4.391e-04  6.636e-05   6.617 5.33e-11 ***
BMI                             2.979e-02  6.239e-03   4.774 2.01e-06 ***
under_five_deaths              -2.536e-03  6.675e-04  -3.799 0.000152 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.604 on 1310 degrees of freedom
Multiple R-squared:  0.8323,    Adjusted R-squared:  0.8314
F-statistic: 928.8 on 7 and 1310 DF,  p-value: < 2.2e-16
```

```
        studentized Breusch-Pagan test

data:  model5
BP = 152.95, df = 7, p-value < 2.2e-16
```

```
        Shapiro-Wilk normality test

data:  resid(model5)
W = 0.99244, p-value = 2.789e-06
```

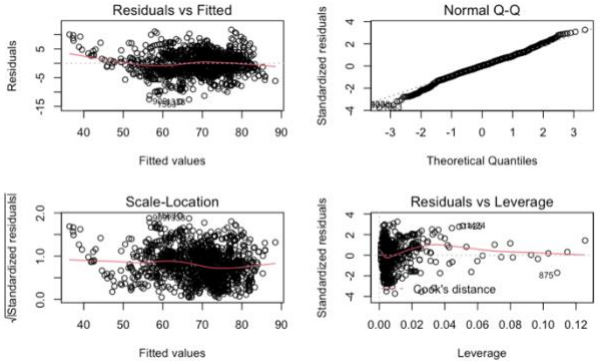


Figure XI: Result of Model V

Model VI (Add Polynomial Terms)

We add polynomial terms (quadratic + cubic) based on Model V. The result comes to 21 predictors and the adjusted R-squared is 0.8903, the set of predictors are significant to the model.

However, the scenario of assumptions doesn't have any change. The model is still violating equal variance and normality assumption. As for linearity assumption, the residual plot tends to be closer to an even separate, the linearity assumption holds.

```
Call:
lm(formula = Life_expectancy ~ I(Schooling^2) + I(HIV_AIDS^2) +
    I(Adult_Mortality^2) + I(Income_composition_of_resources^2) +
    I(percentage_expenditure^2) + I(BMI^2) + I(infant_deaths^2) +
    I(Schooling^3) + I(HIV_AIDS^3) + I(Adult_Mortality^3) + I(Income_composition_of_resources^3) +
    I(percentage_expenditure^3) + I(BMI^3) + I(infant_deaths^3) +
    Schooling + HIV_AIDS + Adult_Mortality + Income_composition_of_resources +
    percentage_expenditure + BMI + infant_deaths, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6109  -1.7706   0.0058   1.7328  11.3844

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         5.687e+01  2.922e+00  19.461  < 2e-16 ***
I(Schooling^2)                     -2.417e-01  6.949e-02  -3.478 0.000522 ***
I(HIV_AIDS^2)                       1.832e-02  5.936e-03   3.087 0.002067 **
I(Adult_Mortality^2)               -1.747e-04  1.835e-05  -9.516  < 2e-16 ***
I(Income_composition_of_resources^2)  1.058e+02  1.555e+01   6.802 1.57e-11 ***
I(percentage_expenditure^2)        -1.288e-07  4.341e-08  -2.966 0.003067 **
I(BMI^2)                            2.966e-03  1.463e-03   2.027 0.042832 *
I(infant_deaths^2)                  8.013e-06  8.875e-06   0.903 0.366775
I(Schooling^3)                      5.909e-03  1.915e-03   3.085 0.002075 **
I(HIV_AIDS^3)                      -2.090e-04  9.162e-05  -2.281 0.022699 *
I(Adult_Mortality^3)                1.864e-07  2.046e-08   9.108  < 2e-16 ***
I(Income_composition_of_resources^3) -4.478e+01  1.089e+01  -4.113 4.16e-05 ***
I(percentage_expenditure^3)         3.868e-12  1.917e-12   2.018 0.043841 *
I(BMI^3)                           -2.415e-05  1.304e-05  -1.851 0.064349 .
I(infant_deaths^3)                 -3.453e-09  5.015e-09  -0.688 0.491305
Schooling                           2.956e+00  7.885e-01   3.749 0.000185 ***
HIV_AIDS                           -7.044e-01  9.458e-02  -7.448 1.73e-13 ***
Adult_Mortality                     2.340e-02  4.018e-03   5.823 7.27e-09 ***
Income_composition_of_resources    -4.433e+01  5.651e+00  -7.844 9.07e-15 ***
percentage_expenditure              1.092e-03  2.343e-04   4.660 3.48e-06 ***
BMI                                -9.589e-02  4.684e-02  -2.047 0.040844 *
infant_deaths                      -5.751e-03  3.412e-03  -1.685 0.092149 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.907 on 1296 degrees of freedom
Multiple R-squared:  0.892,     Adjusted R-squared:  0.8903
F-statistic: 509.9 on 21 and 1296 DF,  p-value: < 2.2e-16
```

```
        studentized Breusch-Pagan test

data:  model6
BP = 98.071, df = 21, p-value = 6.32e-12


        Shapiro-Wilk normality test

data:  resid(model6)
W = 0.9884, p-value = 9.884e-09
```

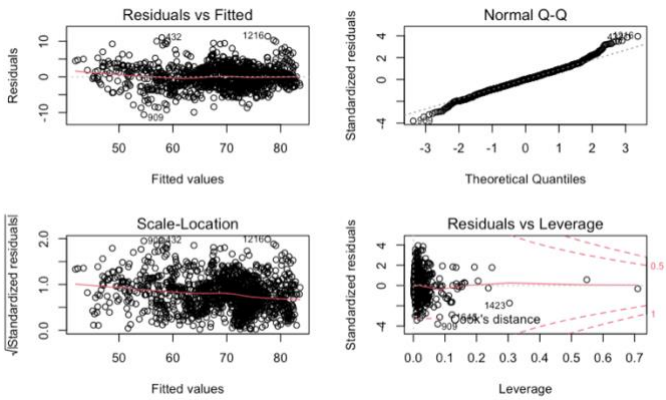


Figure XII: Result of Model VI

Model VII (Stepwise on Polynomial Model)

From the summary of Model VII shown in Figure XIII, the Model VII is significantly simplified that reduce the number of predictors to 10 and remain the adjusted R-squared as 0.8881. All the predictors are significant to the model and has a relatively low AIC: 6592.491 and BIC: 6654.698.

```
Call:
lm(formula = Life_expectancy ~ I(HIV_AIDS^2) + I(Adult_Mortality^2) +
    I(Income_composition_of_resources^2) + I(percentage_expenditure^2) +
    I(Adult_Mortality^3) + I(Income_composition_of_resources^3) +
    HIV_AIDS + Adult_Mortality + Income_composition_of_resources +
    percentage_expenditure, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7899 -1.8736 -0.0253  1.8119 10.7293

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         6.719e+01  5.165e-01 130.090  < 2e-16 ***
I(HIV_AIDS^2)                       4.967e-03  1.253e-03   3.964 7.78e-05 ***
I(Adult_Mortality^2)               -1.933e-04  1.709e-05 -11.314  < 2e-16 ***
I(Income_composition_of_resources^2)  1.311e+02  1.004e+01  13.059  < 2e-16 ***
I(percentage_expenditure^2)        -4.080e-08  1.045e-08  -3.904 9.95e-05 ***
I(Adult_Mortality^3)                2.123e-07  1.835e-08  11.573  < 2e-16 ***
I(Income_composition_of_resources^3) -6.426e+01  7.217e+00  -8.903  < 2e-16 ***
HIV_AIDS                           -5.209e-01  5.384e-02  -9.674  < 2e-16 ***
Adult_Mortality                     2.594e-02  3.872e-03   6.699 3.11e-11 ***
Income_composition_of_resources    -5.281e+01  3.872e+00 -13.637  < 2e-16 ***
percentage_expenditure              7.094e-04  1.346e-04   5.271 1.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.936 on 1307 degrees of freedom
Multiple R-squared:  0.8889,    Adjusted R-squared:  0.8881
F-statistic:  1046 on 10 and 1307 DF,  p-value: < 2.2e-16
```

Figure XIII: Result of Model VII

Model VIII (LASSO Regression)

To use LASSO regression, first we call *cv.glmnet()* function to get the best lambda. The best lambda is 0.004508736. Applying the best lambda to the LASSO regression, we get the coefficients are shown in Figure XIV. We expected LASSO would perform feature selection. Unfortunately, it doesn't penalize any parameters to zero using the best lambda.

	s0
(Intercept)	5.515811e+01
Year2001	-5.575169e-01
Year2002	-1.031305e+00
Year2003	-1.214434e+00
Year2004	-1.178606e+00
Year2005	-1.503521e+00
Year2006	-1.265405e+00
Year2007	-1.504157e+00
Year2008	-1.941900e+00
Year2009	-1.308221e+00
Year2010	-1.807306e+00
Year2011	-1.749774e+00
Year2012	-2.381572e+00
Year2013	-1.932240e+00
Year2014	-2.113525e+00
StatusDeveloping	-7.691302e-01
Adult_Mortality	-1.736019e-02
infant_deaths	6.724992e-02
Alcohol	-9.388304e-02
percentage_expenditure	3.837181e-04
Hepatitis_B	1.123709e-03
Measles	2.861533e-06
BMI	2.801229e-02
under_five_deaths	-5.149254e-02
Polio	5.964807e-03
Total_expenditure	9.456753e-02
Diphtheria	8.573485e-03
HIV_AIDS	-4.305115e-01
GDP	1.325765e-05
Population	9.787122e-10
thinness_1_19_years	1.932137e-02
thinness_5_9_years	-4.593153e-02
Income_composition_of_resources	9.829235e+00
Schooling	9.305198e-01

Figure XIV: Result of Model VIII

Model IX (Ridge Regression)

To use Ridge regression, first we call *cv.glmnet()* function to get the best lambda. The best lambda is 0.6391016. Since the best lambda is small, the regularization effect is relatively small as well. Applying the best lambda to the Ridge regression, we get the coefficients shown in Figure XV:

	s0
(Intercept)	5.474474e+01
Year2001	2.465619e-01
Year2002	-2.582338e-01
Year2003	-3.674438e-01
Year2004	-3.160629e-01
Year2005	-6.147947e-01
Year2006	-4.457582e-01
Year2007	-6.509172e-01
Year2008	-1.114207e+00
Year2009	-4.423231e-01
Year2010	-9.310190e-01
Year2011	-8.739038e-01
Year2012	-1.341440e+00
Year2013	-9.584482e-01
Year2014	-1.150686e+00
StatusDeveloping	-8.787125e-01
Adult_Mortality	-1.798531e-02
infant_deaths	1.510109e-03
Alcohol	-8.103859e-02
percentage_expenditure	2.792412e-04
Hepatitis_B	1.159538e-03
Measles	2.041932e-05
BMI	3.358098e-02
under_five_deaths	-3.640304e-03
Polio	9.912337e-03
Total_expenditure	8.299976e-02
Diphtheria	1.316459e-02
HIV_AIDS	-4.061238e-01
GDP	3.017722e-05
Population	2.662321e-09
thinness_1_19_years	-2.852371e-03
thinness_5_9_years	-2.577132e-02
Income_composition_of_resources	9.791622e+00
Schooling	8.283926e-01

Figure XV: Result of Model IX

Results Summary

To validate the performance of each model, we use each model to make prediction on both our training dataset and testing dataset. Calculate the MSE of each model and each dataset. Collect each MSE and do comparison, the lower MSE, the better model. The results are shown in Figure XVI. Model VII has the second lowest MSE in both Testing dataset prediction and Training dataset prediction, which is very close to the lowest MSE from Model VI, but Model VII is much simpler.

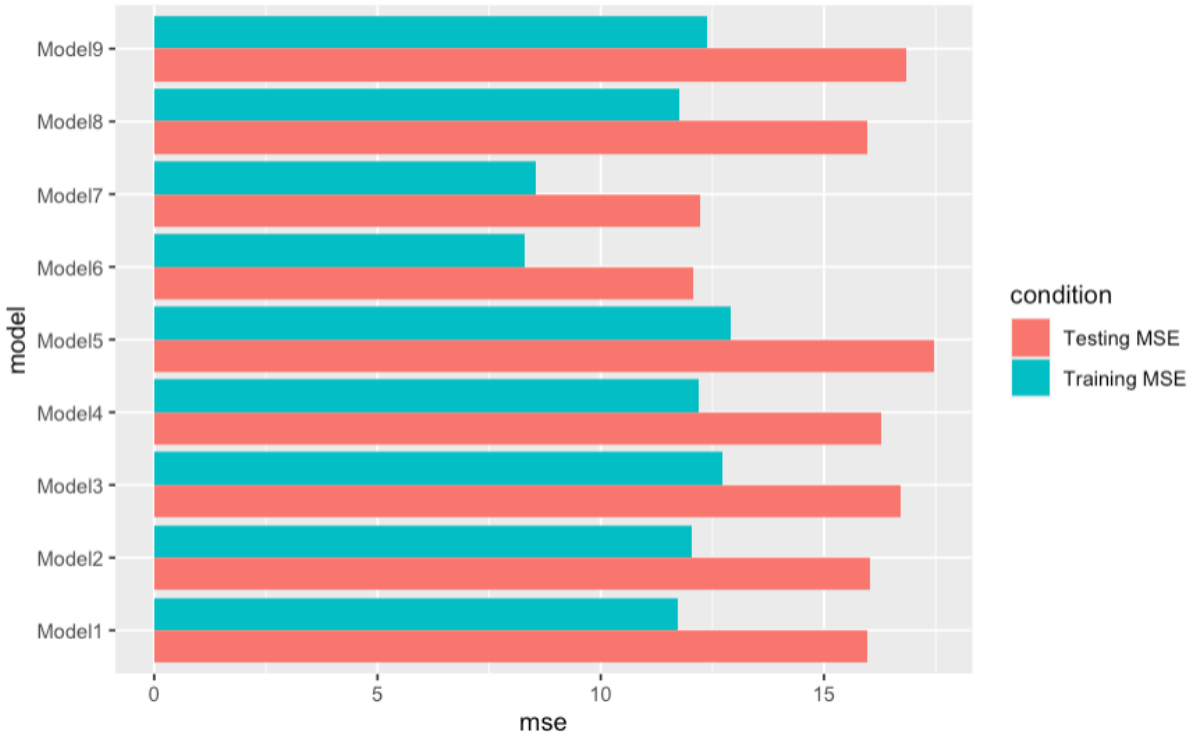


Figure XVI: MSE Plot

As for other metrics, we monitor AIC, BIC, Adjusted R-square of each model exclude Model VIII and Model IX. The results are shown in the Table II below (optimal solution are indicated in red).

	AIC	BIC	Adj R-square	PRESS
Model I	7054.709	7236.144	0.8437891	12.392372
Model II	6110.988	6290.118	0.8826912	7.731163
Model III	7054.709	7236.144	0.8437891	12.392372
Model IV	7057.884	7109.722	0.8404641	12.446803
Model V	7129.741	7176.396	0.8313977	13.139570
Model VI	6577.318	6696.547	0.8902829	8.680979
Model VII	6592.491	6654.698	0.8880941	8.773330
Model VIII	3316.076	3492.328	0.8433512	11.77426
Model IX	3383.016	3559.268	0.8351896	12.38771

Table II: Evaluation Metrics

Discussion & Limitations

After our experiment, we decide to pick Model VII as our optimal model. By using this model, 88.81% of the observation could be explained. The significant predictors are:

- HIV_AIDS
- Adult_mortality
- Incom_composition_of_resources
- Percentage_expenditure
- HIV_AIDS^2
- Adult_Mortality^2
- Income_composition_of_resources^2
- percentage_expenditure^2
- Adult_Mortality^3
- Income_composition_of_resources^3

Even we have tried several methods to transform our dataset and our model has a good performance on predicting, the equal variance and normality assumptions are still being violated. However, "In the large data sets typical in public health research, most statistical methods rely on the Central Limit Theorem, which states that the average of a large number of independent random variables is approximately Normally distributed around the true population mean." (Lumley 2002). Our dataset has more than 1500 observations which can make our model robust to the violations. We finally decided not to worry about the assumption violations in our regression models.

During the experiment, we are always alert to overfitting. Fortunately, our model doesn't have much overfitting issue since there isn't a big gap between R-square and adjusted R-square and the difference of MSE from training dataset and testing dataset is acceptable. Thus, we believe our current model is close to the balance point of the bias/variance trade-off.

Future Work

Even our model achieves a 0.8881 adjustive R-Square, we are still not satisfied as we have a large and complete dataset. We are planning to add some interaction terms in our next experimental stage. As our current model doesn't suffering overfitting issue, we are pretty safe to add more predictors from different statistical approaches. Adding interaction term may help our model include more relationship between the predictors and explain larger portion of variance of y. All of the predictors in our current model are significant which indicates the interaction term is likely to be significant as well.

If the result from the above method still doesn't have any significant improvement. We plan to use some other machine learning and deep learning methods such as Decision Tree Regression, Neural Networks, etc. to build a more advanced model to see if there could be a significant improvement.

Besides trying different model, adding more data to our current model could also be a good way to improve the model performance. The current model contains data up to 2014, if we can obtain the data up to 2020, we believe our model could have better performance.

Reference

Chetty, S. (2016). The Association Between Income and Life Expectancy in the United States, 2001-2014. JAMA: the Journal of the American Medical Association, 315(16), 1750–1766. https://doi.org/10.1001/jama.2016.4226

Wright, W. (1998). Gains in Life Expectancy from Medical Interventions — Standardizing Data on Outcomes. The New England Journal of Medicine, 339(6), 380–386. https://doi.org/10.1056/NEJM199808063390606

Neumayer, E. (2004). HIV/AIDS and Cross-National Convergence in Life Expectancy. Population and Development Review, 30(4), 727–742. https://doi.org/10.1111/j.1728-4457.2004.00039.x

Wang, R. (2018). A Generic Model for Follicular Lymphoma: Predicting Cost, Life Expectancy, and Quality-Adjusted-Life-Year Using UK Population–Based Observational Data. Value in Health, 21(10), 1176–1185. https://doi.org/10.1016/j.jval.2018.03.007

Lumley, T. (2002). The Importance of the Normality Assumption in Large Public Health Data Sets. Annual Review of Public Health, Vol. 23:151-169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546