Effects of Mortality, Economic and Social Factors on Life Expectancy

Author: *Deep Patel;* ***Contributors:*** *Ken Hora, Jia Yu, Samuel Kanu*

INTRODUCTION

It has been observed that in the past 15 years, there have been huge developments in the health sector resulting in improvement of human mortality rates, especially in the developing nations in comparison to the past 30 years. There have been a lot of studies undertaken in the past on factors affecting life expectancy such as demographic variables, income composition and mortality rates. Overall, this project aims to analyze the effects of immunization factors, mortality factors, economic factors, social factors and other health related factors on life expectancy. Our analysis intends to answer the following questions:

1. Do various predicting factors which have been chosen initially really affect life expectancy? What are the predictors significantly affecting life expectancy?
2. How do infant and adult mortality rates affect life expectancy?
3. How does lifestyle factors and economic factors such as drinking alcohol, government spending on healthcare, and GDP affect life expectancy?
4. What is the impact of schooling on the lifespan of humans?

The data used in this paper consists of 911 observations and 21 predictors. From 21 predictors, 2 categorical predictors were removed, and one categorical predictor was converted to a binary numerical for our analysis resulting in a total 19 numerical predictors.

Data Metrics
Response Variable: Life Expectancy
Number of Predictors: 19 (See Appendix for description)
Observations: 911

METHODOLOGY

A. Data Cleaning

Before creating any linear models with our data, it was necessary to clean it. While some data was missing, it was also clear that other data was entered incorrectly. Unfortunately, there was not always a clear-cut way to filter bad data from good, however we used our best judgement and fairly liberal measures of what counted as good data. First, we removed all rows that contained NA values. For all feature variables we removed rows which contained zeros, as it was apparent that these were placeholders for NA values.

B. Methods

After data cleaning, the distribution of all 19 predictors along with correlation was observed with ggpairs plot. Upon checking the distribution, it was evident that few predictors had moderate correlations with life expectancy, but the distribution was highly skewed. Thus, log of those predictors were taken and the full correlation matrix of new 19 predictors was observed. The linear regression model was applied on the full model with all 19 predictors to observe general

importance of all predictors and the R-squared value. To eliminate irrelevant predictors that have little to no effect on life expectancy, stepwise backward regression was performed and the correlation matrix was observed for all 19 predictors along with analysis of scatterplots of some predictors which narrowed the model down to 6 predictors. On those 6 predictors, another stepwise regression was performed to make sure only important predictors were kept. The residuals plot and normality test was performed to check the normality assumption. Then, two predictors were eliminated using multicollinearity detection. On the final set of predictors, in an effort to improve the normality and remove outliers from the model, methods such as studentized deleted residuals, and influential data points were obtained using DFFITS & DFBETAS. Lastly, to answer our research questions, full vs. reduced models were compared to check if particular predictors truly had significant effects on life expectancy.

DATA ANALYSIS

Before jumping into the analysis, the distribution check was performed on all 19 predictors by using ggpairs() plot function in R (See Appendix). Along, with distribution of all predictors, this also shows the scatterplots and correlations between each pair of variables. Upon analyzing the distribution, it was evident that predictors HIV.AIDS, GDP, and percentage expenditure (PE) had moderate correlation, but the distribution was highly skewed showing a nonlinear curve. Therefore, it made sense to perform log transformation on these predictors.

Afterwards, to briefly visualize the importance of each predictor, the linear regression was performed on all 19 predictors.

```
Call:
lm(formula = Lifeexpectancy ~ ., data = life_select)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7666  -1.6397  -0.0219   1.5418  12.6230

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.166e+01  1.067e+00  48.400  < 2e-16 ***
AdultMortality   -1.271e-02  9.972e-04 -12.744  < 2e-16 ***
Alcohol          -1.593e-01  3.875e-02  -4.112 4.28e-05 ***
HepatitisB       -6.299e-03  4.793e-03  -1.314 0.189101
Measles           6.923e-06  9.230e-06   0.750 0.453411
BMI              -6.566e-03  7.874e-03  -0.834 0.404616
Polio            -5.915e-04  5.657e-03  -0.105 0.916752
Totalexpenditure  6.981e-02  4.804e-02   1.453 0.146538
Diphtheria        1.314e-02  6.531e-03   2.012 0.044517 *
under.fivedeaths -1.752e-02  6.892e-03  -2.541 0.011208 *
thinness1.19years 7.931e-02  4.629e-02   1.713 0.087029 .
thinness5.9years -1.385e-01  4.567e-02  -3.032 0.002502 **
Population        7.876e-10  1.495e-09   0.527 0.598417
infantdeaths      2.237e-02  9.528e-03   2.348 0.019078 *
Income.Comp       3.813e+01  2.402e+00  15.875  < 2e-16 ***
Schooling        -3.612e-01  9.912e-02  -3.644 0.000284 ***
Country_Status   -1.377e+00  4.073e-01  -3.381 0.000754 ***
log_PE            8.766e-01  1.500e-01   5.845 7.09e-09 ***
log_HIV.AIDS     -1.631e+00  9.715e-02 -16.794  < 2e-16 ***
log_GDP          -7.591e-01  1.694e-01  -4.482 8.34e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.01 on 891 degrees of freedom
Multiple R-squared:  0.9011,    Adjusted R-squared:  0.899
F-statistic: 427.3 on 19 and 891 DF,  p-value: < 2.2e-16
```

Figure 1. General Model with a1l 19 predictors.

Using all predictors in a model resulted in a high R-squared value of 0.901. R-squared value, known as the coefficient of determination, is a proportion of variation of dependent variable which is explained by independent variables. R-squared value ranges from 0 to 1, with 0 being the weakest suggesting no variation and 1 being the strongest suggesting a perfect model. However, in some cases, very high R-squared values can be an indicator of a problem affecting

the regression model. Some of the possible problems include overfitting the regression model, including too many predictors which may include misleading coefficients, and including same variables in different forms could also lead to high R-squared values. To check these problems, different models were built to investigate if an inflated R-squared value is due to too many predictors in the model having misleading coefficients and removing the same kind of variables in different forms such as infant deaths and under five deaths.

A. Stepwise Regression- Full model (19 predictors)

In order to be more efficient with our approach of removing predictors, stepwise backward regression was performed which tests each predictor and removes one predictor at a time. The stepwise backward elimination was chosen over forward and both stepwise regression as it had the lowest AIC of 2020.44.

```
Call:
lm(formula = Lifeexpectancy ~ AdultMortality + Alcohol + Diphtheria +
    under.fivedeaths + thinness1.19years + thinness5.9years +
    infantdeaths + Income.Comp + Schooling + Country_Status +
    log_PE + log_HIV.AIDS + log_GDP, data = life_select)

Coefficients:
      (Intercept)     AdultMortality            Alcohol         Diphtheria
        52.037786          -0.012846          -0.151554           0.009578
 under.fivedeaths  thinness1.19years   thinness5.9years       infantdeaths
        -0.019882           0.074996          -0.137637           0.026419
      Income.Comp          Schooling     Country_Status              log_PE
        37.071276          -0.334911          -1.355742           0.923409
     log_HIV.AIDS            log_GDP
        -1.616945          -0.799327
```

Figure 2. Final output of Stepwise Backward Regression (19 predictors)
**See Appendix for full final output*

B. Correlation Matrix/Heatmap

Additionally, the variables that highly correlated with the response variable were considered. This was done using the cor() function in R.

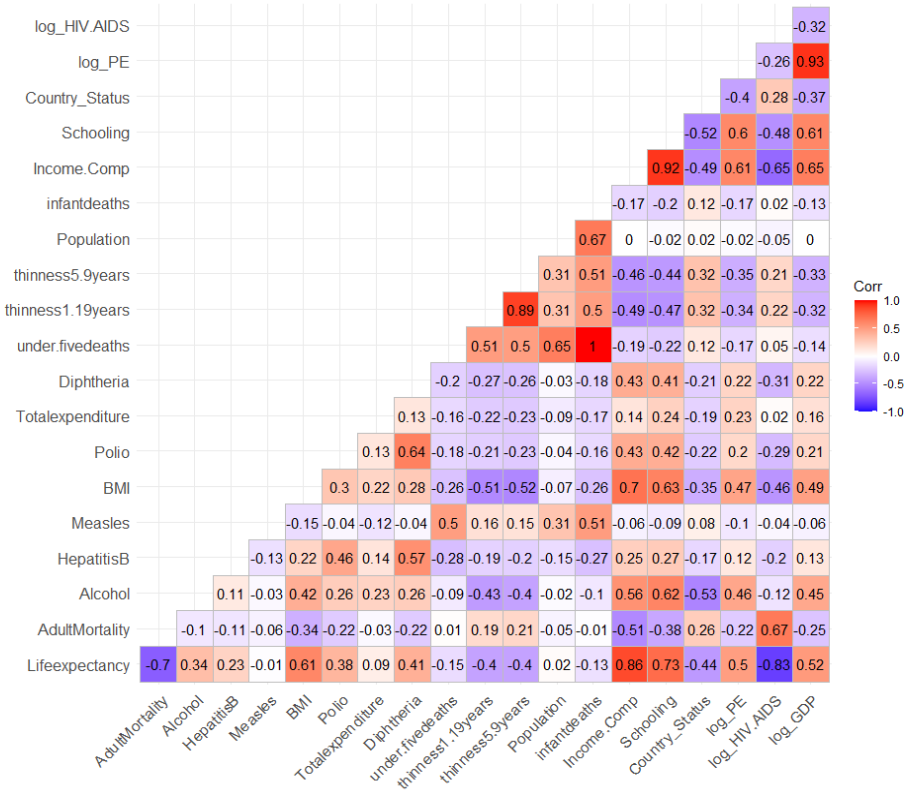
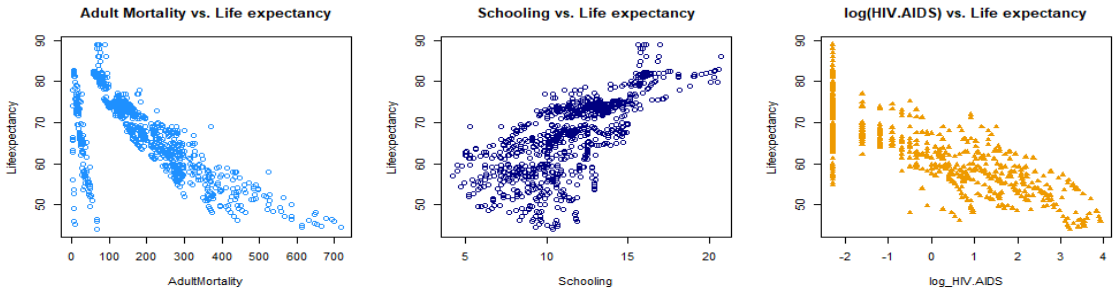


Figure 3. Correlation Heatmap for all predictors.

From this matrix, it can be seen that the predictors that are moderately to strongly correlated are: Adult Mortality (-0.70), BMI (0.61), Income.Comp (0.86), Schooling (0.73), log_PE (0.5), log_HIV.AIDS (0.83), log_GDP (0.52). Notice, when the stepwise backward regression was performed on all predictors, BMI was eliminated in the final output. Therefore, scatterplots were obtained to visualize the highly correlated features and some predictors in the final output of stepwise backward regression that shows poor correlation in correlation matrix. These scatterplots are shown below. Notice there is a clear positive relationship with Life expectancy for predictors Schooling, Income Composition of Resources, log(PE) and log(GDP). There is a clear negative relationship with Life expectancy for predictors Adult Mortality, log(HIV.AIDS). The predictors; alcohol and infant deaths do not show any relationship.



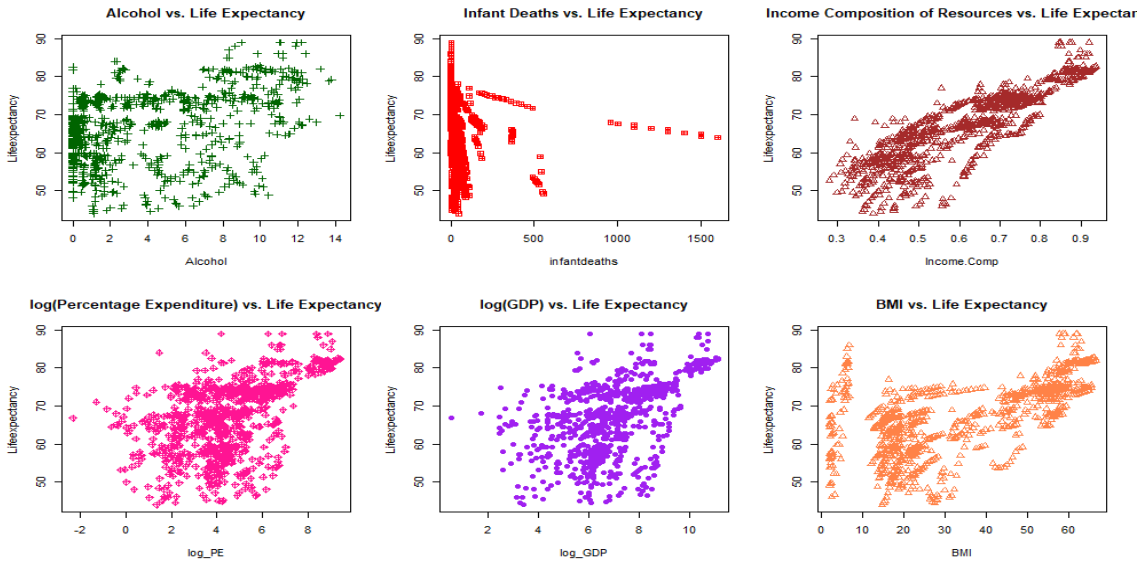


Figure 4. Scatterplots of predictors of interest.

BMI as a predictor was discarded as it was eliminated in stepwise backward regression, despite having moderate correlation with life expectancy. This comes as no surprise after observing the scatterplot of BMI vs. Life Expectancy; notice that much of the data points don't make sense; the smallest values are in the single digits and the highest are above 60. According to BMI charts, BMIs under 15 are considered severely underweight and BMIs over 40 are considered severely obese. Additionally, it seems unrealistic for human bodies to be able to survive at BMIs below 10, and even in this decade it is hard to believe that any nation's average BMI is over 40. A quick internet search finds that average BMIs by nation range from 20 to 32. Thus, it was determined that best predictors for further analysis are Adult Mortality, Income.Comp, Schooling, log_PE, log_HIV.AIDS and log_GDP. It was clear that these 6 predictors were the most relevant, per the stepwise approach, as performing the backward stepwise regression one more time did not reduce the number of predictors, as shown below.

```
Call:
lm(formula = Lifeexpectancy ~ ., data = life_select3)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4873  -1.6022  -0.0004   1.6355  12.4422

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    50.805761   0.736078  69.022  < 2e-16 ***
AdultMortality -0.013596   0.001003 -13.552  < 2e-16 ***
Income.Comp    37.003910   2.206559  16.770  < 2e-16 ***
Schooling      -0.331337   0.095309  -3.476 0.000532 ***
log_PE          0.964182   0.148004   6.515 1.21e-10 ***
log_HIV.AIDS   -1.755394   0.094385 -18.598  < 2e-16 ***
log_GDP        -0.853483   0.169598  -5.032 5.84e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.085 on 904 degrees of freedom
Multiple R-squared:  0.8946,    Adjusted R-squared:  0.8939
F-statistic:  1278 on 6 and 904 DF,  p-value: < 2.2e-16
```

Figure 5. Model with selected 6 predictors.

However, the value of R-squared decreased by a small margin to 0.8946 which is understandable as R-squared was inflated in full model with 19 predictors due to the same kind of predictors and due to the reasons discussed earlier in the report.

```
Call:
lm(formula = Lifeexpectancy ~ AdultMortality + Income.Comp +
    Schooling + log_PE + log_HIV.AIDS + log_GDP, data = life_select3)

Coefficients:
   (Intercept)  AdultMortality      Income.Comp        Schooling           log_PE    log_HIV.AIDS
      54.74178        -0.01599        10.02801         0.62187          0.87134         -2.26562
       log_GDP
      -0.48711
```

Figure 6. Final output of Stepwise Backward Regression with selected 6 predictors.
**See Appendix for full final output*

The model obtained for our backwards stepwise regression with 0 steps is the same as our linear model with 6 predictors:

$$Y = 54.74 - 0.01599 * AdultMorality + 10.03 * Income.Comp + 0.6219 * Schooling + 0.8713 * \log\_PE - 2.266 * \log\_HIV.AIDS - 0.4871 * log\_GDP$$

C. Residual Plot and Normality Test

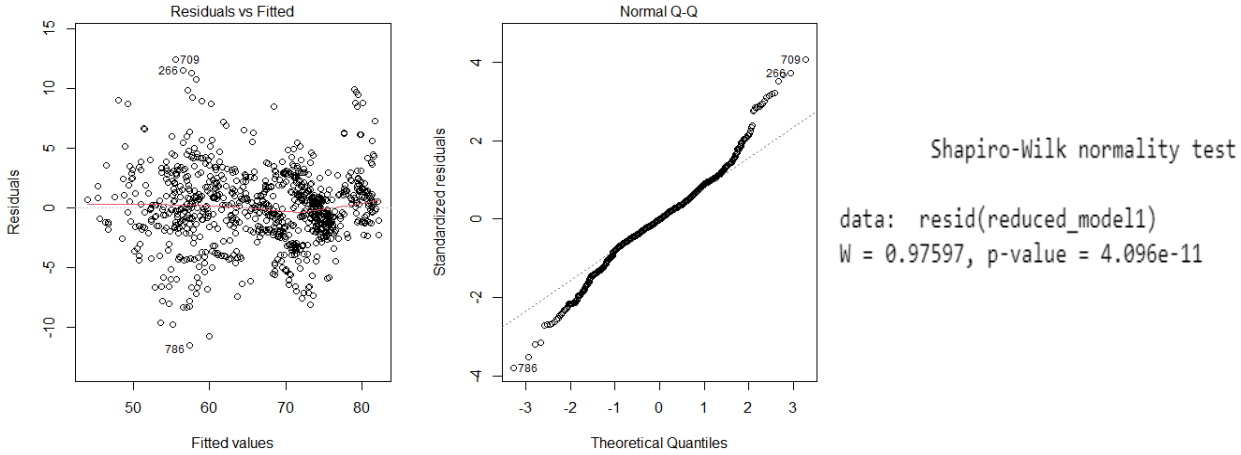


Figure 7. Residual Plot, Normality plot and Shapiro-Wilk Normality Test.

The Residual vs. Fitted value plot shows that residuals have almost equal variance. However, the normality plot on the right has heavy tails deviating from the normality line suggesting residuals may not be normally distributed. To further confirm this, Shapiro-Wilk test was performed on the residuals. As the p-value is very low, it rejects the normality assumption indicating there is sufficient evidence to conclude that residuals are not normally distributed.

D. Detecting & Removing Multicollinearity

If the model contains two predictors that are highly correlated with each other, there could be multicollinearity issues which may be the reason for altering normality. Variation Inflation Factor (VIF) was used to detect the presence of multicollinearity in the model which shows variance of regression coefficient for each predictor. If the predictor has a factor of greater than 4, then it suggests that there may be correlation between the two predictors to warrant investigation. If the VIF is greater than 10, then it suggests there is high correlation between two predictors and indicates serious signs of multicollinearity. The VIF for selected 6 predictors is shown below which shows Income.Comp predictor has the highest VIF followed by log(GDP), log(PE) and Schooling. Thus, the Income.Comp predictor was removed and the new model was run.

```
> vif(reduced_model1)
AdultMortality    Income.Comp      Schooling         log_PE   log_HIV.AIDS        log_GDP
      1.839017      11.846299       7.936051       8.119402       2.783107       8.620951
```

Figure 8. Multicollinearity Detection with Variation Inflation Factor (6 predictors)

Upon removing Income.Comp, the Schooling VIF decreased but log(GDP) and log(PE) VIF remained around 8. As GDP does have correlation with PE as shown in correlation matrix, the removal of one of those predictors was necessary. Therefore, log(GDP) was removed along with Income.Comp which resolved the multicollinearity issue.

```
> vif(reduced_model3)
AdultMortality      Schooling         log_PE   log_HIV.AIDS
      1.818624       1.888778       1.560337       2.009533
```

Figure 9. Multicollinearity Detection with Variation Inflation Factor (4 predictors)

E. Assessing the Remaining Four Predictors with Reduced Models

To determine if the remaining four predictors are each worth keeping in the model, we perform F-tests on each of them individually. Each one requires a reduced model where the term in question is dropped from the full model.

Full Model: $Y=\beta 0+\beta 1*$Adult_Mortality $+\beta 2*$Schooling $+\beta 3*$log_PE $+\beta 4*$log_HIV.AIDS

Hypothesis for each term:

$H_0: \beta i = 0$

$H_a: \beta i \neq 0$

The F-statistic for each term is calculated by:

F=[(SSE(R)-SSE(F))/(df(R)-df(F))]/MSE(F)*

F-critical is the same for all terms and is calculated by:

F-critical = qf(1-0.05, 1, 906) = 3.85

Decision Rule:

If F <= F-critical conclude* H_0

If F > F-critical conclude* H_a

Table 1 below shows the F-test results based on Full versus Reduced Model, where Full model includes 4 predictors and Reduced model includes one less predictor for each F-test

Table 1. F-test Results.

Term	F*	F-critical	Conclusion
Adult Mortality	179	3.85	$Reject\ H_0,\ \beta 1 \neq 0$

Schooling	393	3.85	$Reject\ H_0,\ \beta 2 \neq 0$
log_HIV.AIDS	766	3.85	$Reject\ H_0,\ \beta 3 \neq 0$
log_PE	172	3.85	$Reject\ H_0,\ \beta 4 \neq 0$

F. Efforts to Improve Normality (On Best Model)

Using studentized deleted residual approach, outliers check was performed, but there were no detected outliers. DFFITS is the difference in fits which shows the number of standard deviations that the fitted value changes for each case when the particular case is omitted. If absolute value is greater than $2 \cdot \sqrt{p/n}$ for a large dataset, then those observations are deemed influential unduly affecting the regression analysis. The cutoff value for possibly influential observations for this data was 0.148. Upon computing DFFITS, a total of 85 observations out of 911 were found to be influential. When these observations were removed, the normality plot obtained from the new model appeared more normal and it barely passed the normality test according to the Shapiro-Wilk Normality test as shown below. The computation of DFBETAS for each predictor also suggested 75 observations for each predictor was deemed influential.

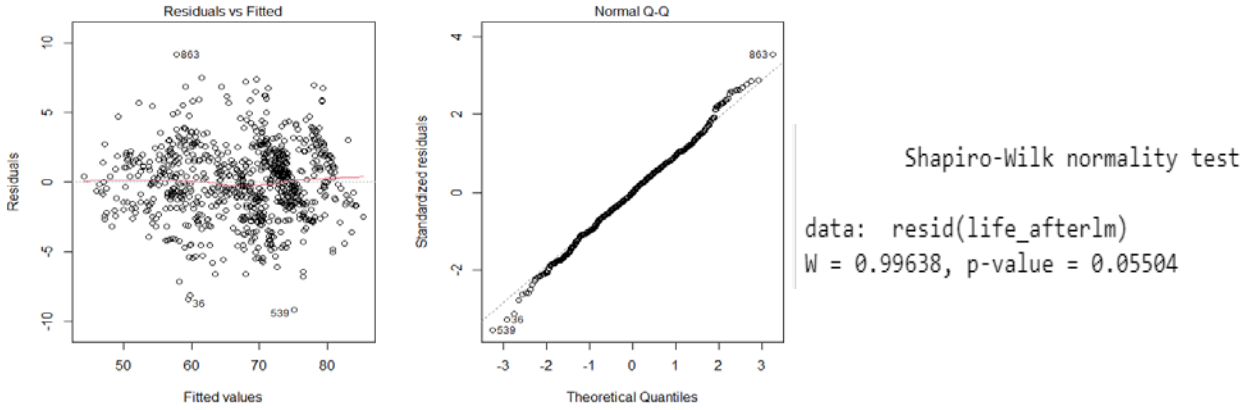


Figure 10. Residual plot, Normality plot, and Shapiro-Wilk Normality test (after removing influential points)

The importance of each final four predictors was reassessed after removing influential points by performing F-test on each of them. The results remained the same concluding each selected predictor (Adult Mortality, Schooling, log(HIV.AIDS), log(PE)) has a significant effect on life expectancy.

G. Confidence Intervals for Predictor Betas

After confirming that the remaining terms are significant and removing influential points, we now want to determine confidence intervals for the beta of each term at the 95% significance level. This is done by using a Bonferroni joint confidence interval:

$$b_i + B \cdot s(b_i) \leq b_i \leq b_i + B \cdot s(b_i)$$
$$B = t(1 - \alpha/2g;\ n - 2)$$

Table 2. Confidence Intervals for Predictor Betas

Term	b_i	$s(b_i)$	Lower limit	Upper limit
Intercept	54.32	0.4793	53.08	55.56
Adult Mortality	-0.01755	0.00103	-0.02021	-0.01489
Schooling	1.025	0.04332	0.9131	1.137
log_HIV.AIDS	-2.382	0.08467	-0.1958	-4.568
log_PE	0.4979	0.05955	0.3441	0.6517

CONCLUSION

Our analysis has shown that many predictors in this data set can't be said to have a significant impact on life expectancy. Backwards stepwise regression removed most of the predictors, and analyzing correlation matrix left us with adult mortality, schooling, HIV.AIDS deaths and percentage expenditure (PE). Surprisingly, some predictors which we thought would remain in our model, such as alcohol and infant mortality, were dropped. It was also evident that R-squared was inflated due to too many predictors and having the same kind of predictors. GDP was dropped due to multicollinearity as it was correlated with PE. The number of years of school had a surprisingly significant effect; each additional year of schooling was shown to increase life expectancy by one year when all other factors were held constant. Similarly, adult mortality had a significant effect on life expectancy; the more the adult mortality, the lower the life expectancy. HIV.AIDS deaths also had a significant effect on Life expectancy as the removal of that predictor decreased R-squared value drastically from 0.86 to 0.74. Although the removal of PE, which represents expenditure on health, did not decrease R-squared by much, the F-test on each term rejected the null assumption indicating each predictor has a significant effect on life expectancy.

Although appropriate approaches were taken for data cleaning, this dataset had many deficiencies. More accurate data collection might have improved the data analysis resulting in fewer outliers or influential points. Some possible procedures that can be applied include using ridge regression instead of eliminating predictors which is also one of the options to handle multicollinearity. Also, testing other kinds of relationships besides linear relationship may also provide a clearer understanding of the dropped predictors which, for our MLR models, showed no effect on life expectancy.

REFERENCES

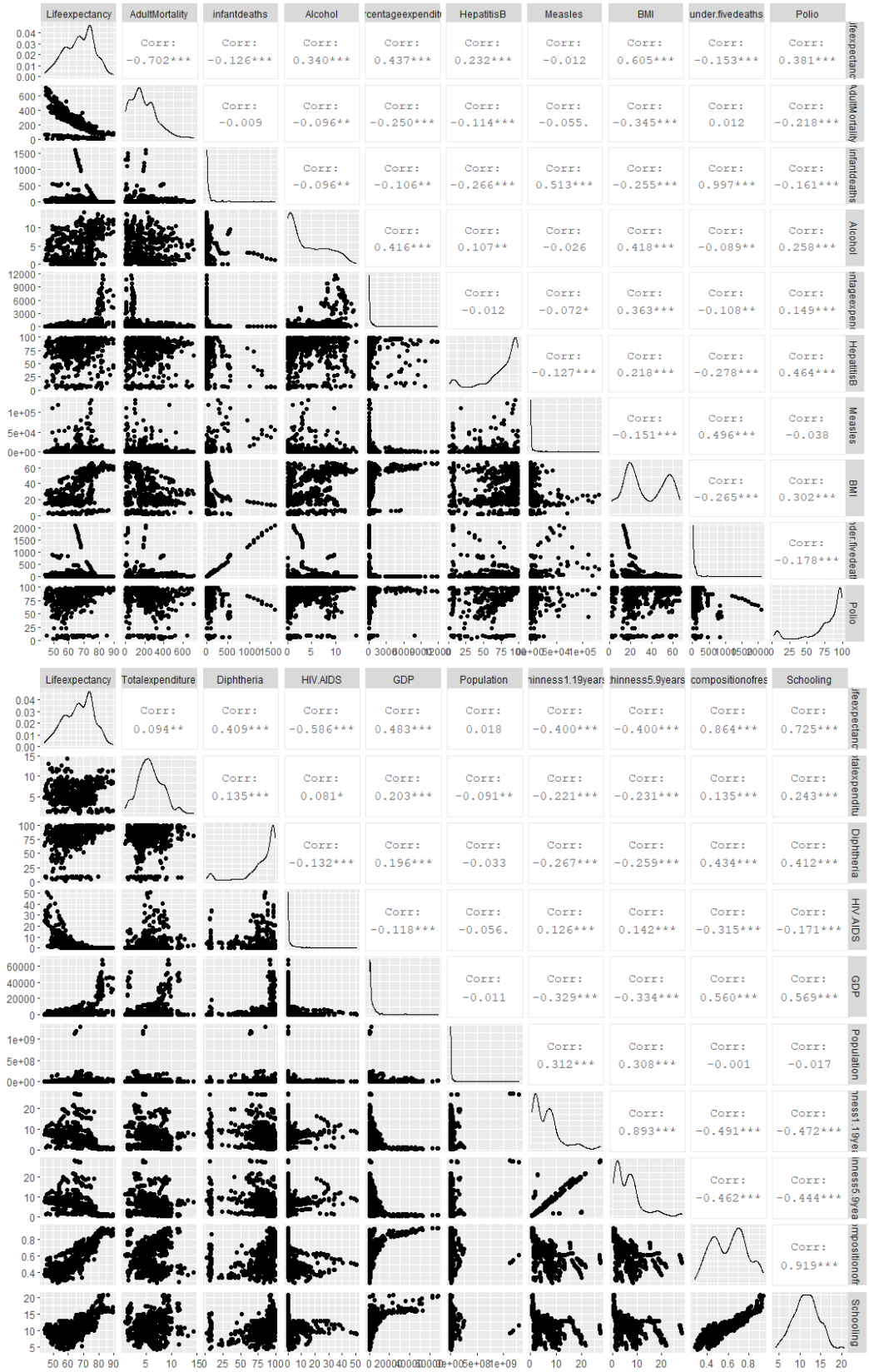
Kumar, Rajashri. 2018. Life Expectancy (WHO)- Statistical Analysis on Factors influencing Life Expectancy. Retrieved from https://www.kaggle.com/kumarajarshi/life-expectancy-who

APPENDIX

1) Description of Predictors

- **Adult Mortality:** Number of adults that die between age 15-60 per 1000 people.
- **Infant Deaths:** Number of infant deaths per 1000 people.
- **Alcohol:** recorded per capita (15+) consumption (in liters of pure alcohol).
- **Percentage Expenditure:** Expenditure on health as a percentage of Gross Domestic Product per capita(%).
- **Hepatitis B:** Hep B immunization coverage among 1-year-olds (%).
- **Measles:** Number of reported cases per 1000 people.
- **BMI:** average body mass index for entire population
- **Under Five deaths:** Number of under five deaths per 1000 people.
- **Polio:** (Pol3) immunization 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 by HIV/AIDS among 0-4 years old per 1000 live births.
- **GDP:** Gross Domestic Product per capita in USD.
- **Population**- Population of the country.
- **Country Status-** Developing (1) or developed (0).
- **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.
- **Schooling:** Number of years of schooling per person.

2) Distributions and Correlation of all original predictors in the dataset.



3) Full Final outputs of Stepwise Backward Regression

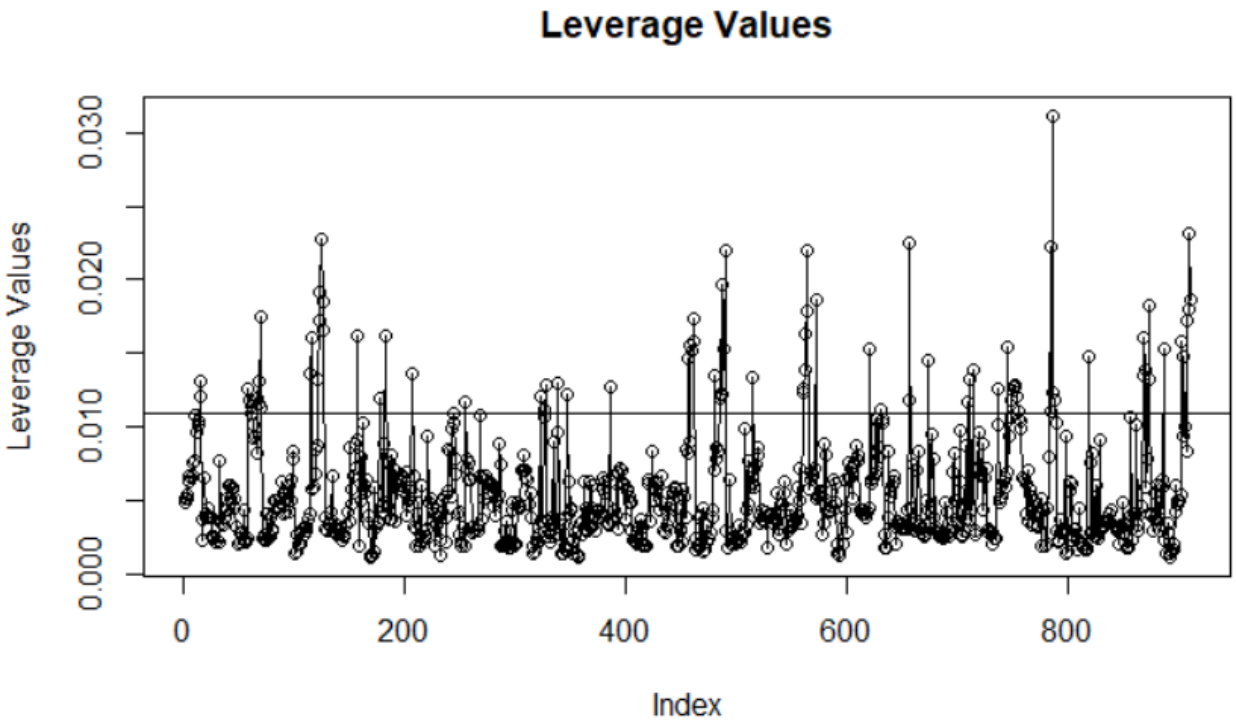
```
Step:  AIC=2020.44
Lifeexpectancy ~ AdultMortality + Alcohol + Diphtheria + under.fivedeaths +
    thinness1.19years + thinness5.9years + infantdeaths + Income.Comp +
    Schooling + Country_Status + log_PE + log_HIV.AIDS + log_GDP

                   Df Sum of Sq     RSS    AIC
<none>                          8116.3 2020.4
- thinness1.19years  1     24.08  8140.4 2021.1
- Diphtheria         1     32.50  8148.8 2022.1
- infantdeaths       1     79.84  8196.1 2027.3
- under.fivedeaths   1     82.12  8198.4 2027.6
- thinness5.9years   1     85.38  8201.7 2028.0
- Country_Status     1    101.37  8217.7 2029.7
- Schooling          1    110.85  8227.2 2030.8
- Alcohol            1    140.87  8257.2 2034.1
- log_GDP            1    206.28  8322.6 2041.3
- log_PE             1    357.27  8473.6 2057.7
- AdultMortality     1   1512.32  9628.6 2174.1
- Income.Comp        1   2449.47 10565.8 2258.7
- log_HIV.AIDS       1   2534.33 10650.6 2266.0

Call:
lm(formula = Lifeexpectancy ~ AdultMortality + Alcohol + Diphtheria +
    under.fivedeaths + thinness1.19years + thinness5.9years +
    infantdeaths + Income.Comp + Schooling + Country_Status +
    log_PE + log_HIV.AIDS + log_GDP, data = life_select)

Coefficients:
      (Intercept)     AdultMortality            Alcohol         Diphtheria   under.fivedeaths
        52.037786          -0.012846          -0.151554           0.009578          -0.019882
thinness1.19years   thinness5.9years       infantdeaths        Income.Comp          Schooling
         0.074996          -0.137637           0.026419          37.071276          -0.334911
   Country_Status              log_PE       log_HIV.AIDS             log_GDP
        -1.355742           0.923409          -1.616945          -0.799327
```

Stepwise Backward Regression on all 19 predictors

```
Start:  AIC=4122
Lifeexpectancy ~ AdultMortality + Income.Comp + Schooling + log_PE +
    log_HIV.AIDS + log_GDP

                Df Sum of Sq   RSS    AIC
<none>                      20004 4122.0
- log_GDP        1     144.0 20148 4131.8
- log_PE         1     600.2 20604 4168.6
- Schooling      1    1622.9 21627 4248.2
- Income.Comp    1    1988.5 21992 4275.8
- AdultMortality 1    3558.5 23562 4389.2
- log_HIV.AIDS   1   11128.3 31132 4847.2

Call:
lm(formula = Lifeexpectancy ~ AdultMortality + Income.Comp +
    Schooling + log_PE + log_HIV.AIDS + log_GDP, data = life_select3)

Coefficients:
   (Intercept)  AdultMortality     Income.Comp       Schooling          log_PE    log_HIV.AIDS
      54.74178        -0.01599        10.02801         0.62187         0.87134        -2.26562
       log_GDP
      -0.48711
```

Stepwise Backward Regression on selected 6 predictors

4) Leverage values plot



5) R-squared comparison among different combinations of predictors

Number of Predictors	R-squared	R-squared (Adjusted)
19 predictors	0.9011	0.899
6 selected predictors	0.8946	0.8939
5 predictors (no Income.Comp)	0.8617	0.8610
4 predictors (no Income.Comp & no logGDP)	0.8615	0.8609
3 predictors (Schooling, log(PE),log(HIV.AIDS))	0.8341	0.0.8336
3 predictors (Schooling, log(PE),Adult Mortality)	0.7444	0.7436
3 predictors (Schooling, Adult Mortality, log(HIV.AIDS))	0.8522	0.8517
3 predictors (Adult Mortality, log(PE),log(HIV.AIDS))	0.8013	0.8007
2 predictors (log(PE), log(HIV.AIDS))	0.7375	0.7369
2 predictors (Adult Mortality, Schooling)	0.7623	0.762
1 predictor (Adult Mortality)	0.4932	0.4927
1 predictor (log(HIV.AIDS))	0.6452	0.645
1 predictor (Schooling)	0.5311	0.5308
1 predictor (log(PE))	0.2486	0.2477