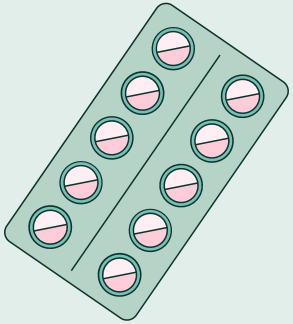
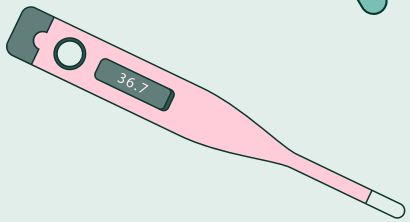





# World Health Organization

# Life Expectancy

Xavier Mojica
xmojica@smu.edu
Alexandra Thibeaux
athibeaux@smu.edu

# Introduction

This case study will focus on the analysis of Life Expectancy data, and on creating predictive models for life expectancy.

1. Build a model in order to identify key relationships along with all necessary testing and interpretations.

2. Compare multiple models in order to develop the best fitting model, that can highly predict life expectancy..

# Data Context

Came from World Health Organization

# Individual Observations: 2938

# of Years: 2000-2015

Response Variable: Life Expectancy

Explanatory Variables:

Adult Mortality, Alcohol, BMI, Diphtheria, GDP, Hepatitis, HIV AIDS, Income Composition of resources, Infant deaths, Measles, Percentage expenditure, Polio, Population, Schooling, Status, Thinness 5-9 years, Thinness 10-19 years, Total expenditure, Under five deaths

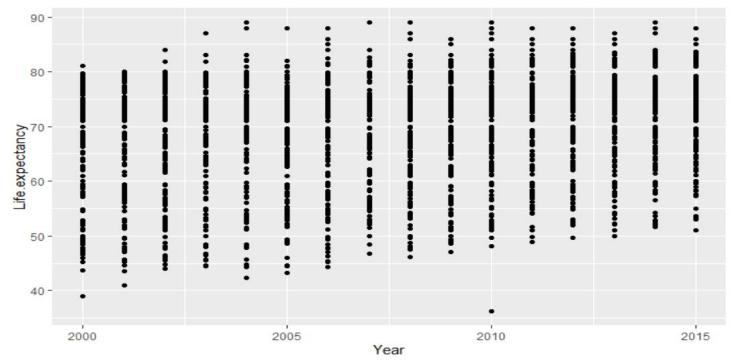
Variance Inflation Factors:

Year: 1.147699
Status: 1.844672
Adult.Mortality: 1.791904
Infant.deaths: 170.003056
Alcohol: 1.908148
Percentage.expenditure: 1.669605
Measles: 1.416355
BMI: 1.760019
Under.five.deaths: 172.685555
Polio: 1.993205
Total.expenditure: 1.204580
HIV.AIDS: 1.486816
log(GDP): 2.081129
Thinness..1.19.years: 8.502136
Thinness.5.9.years: 8.625215
Income.composition.of.resources: 3.206133
Schooling: 3.559931

# Limitations of Data

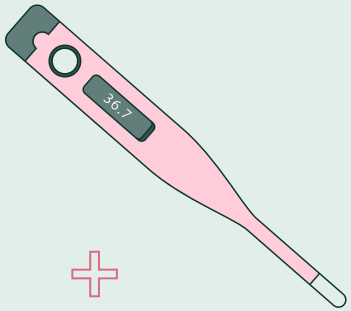
1. Observational study

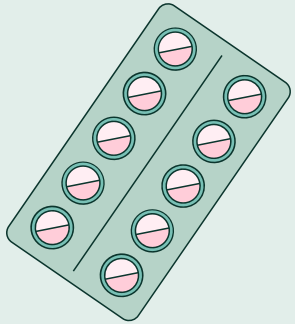
2. Violations of Assumption of Independence (Year)

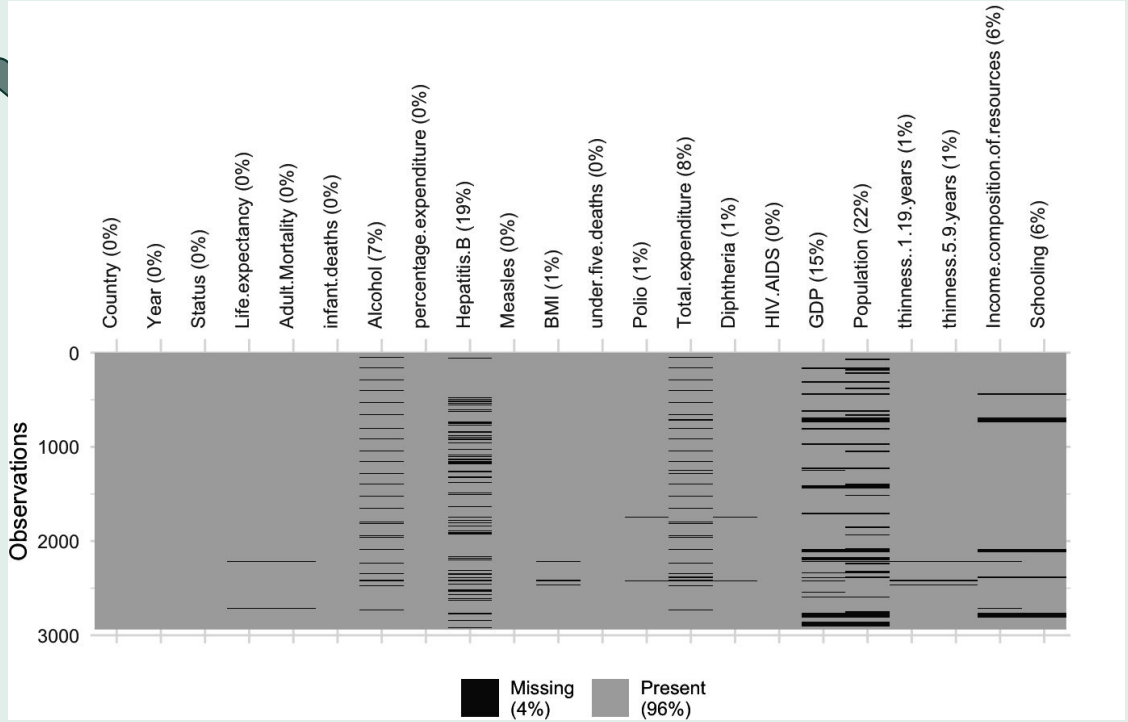



3. Aggregated Attributes:
Income Composition of Resources is a Human Development Index between 0 and 1 based on income and availability of resources

# Data Preparation



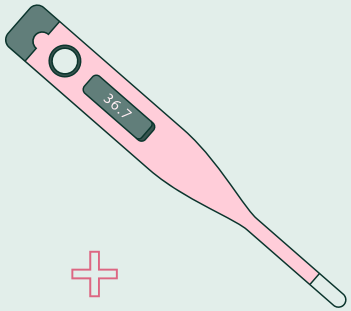



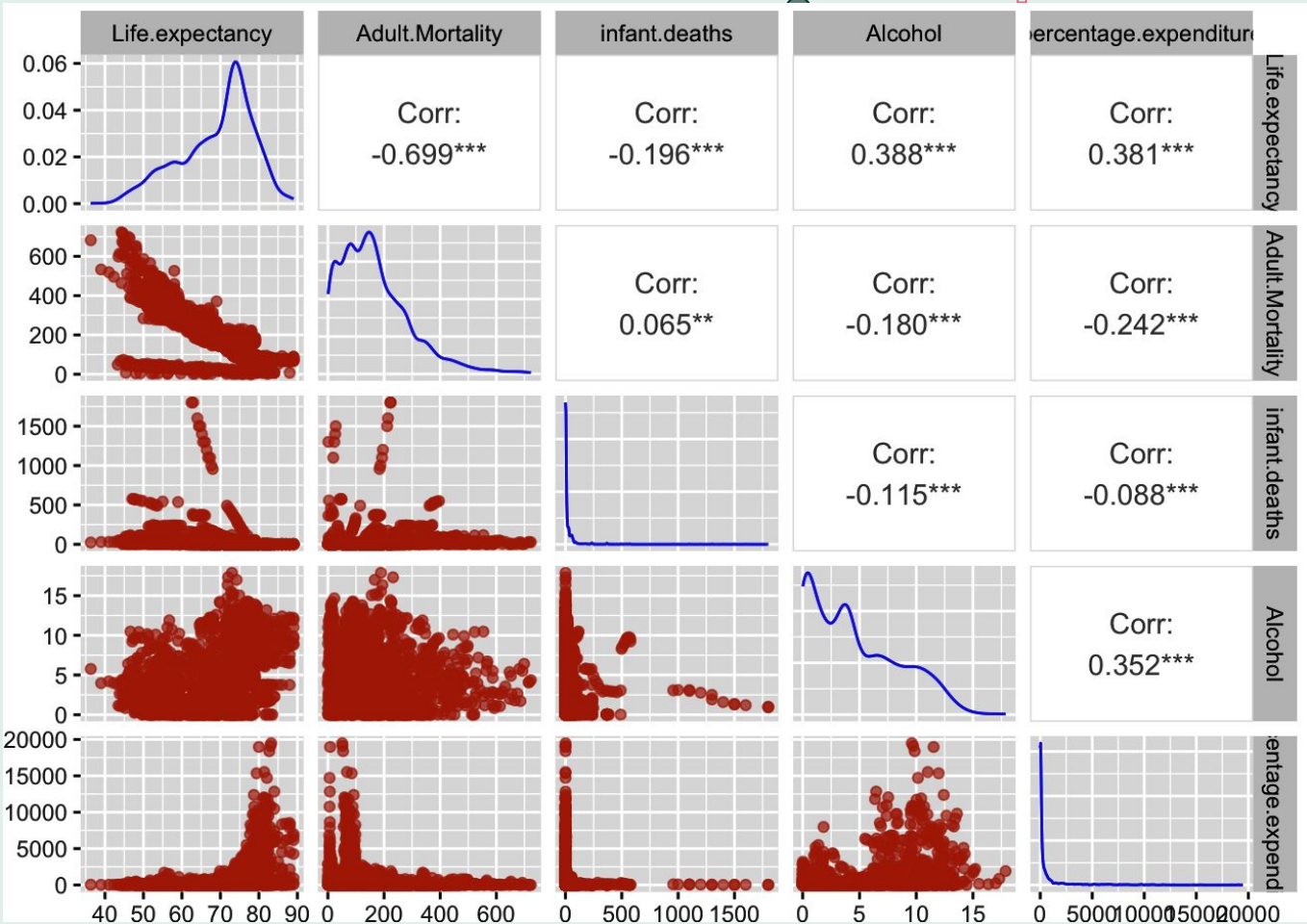


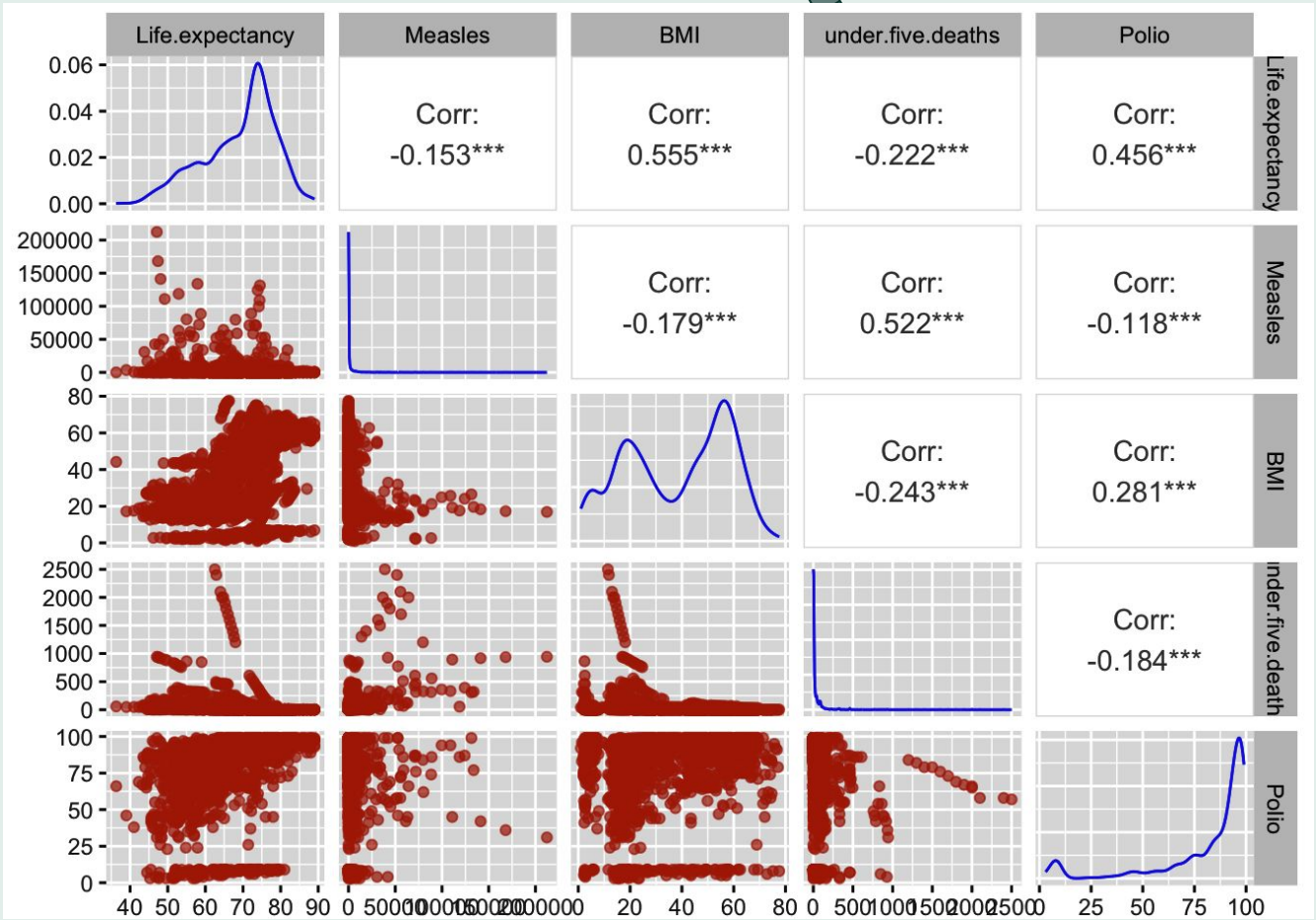

# What do we do?

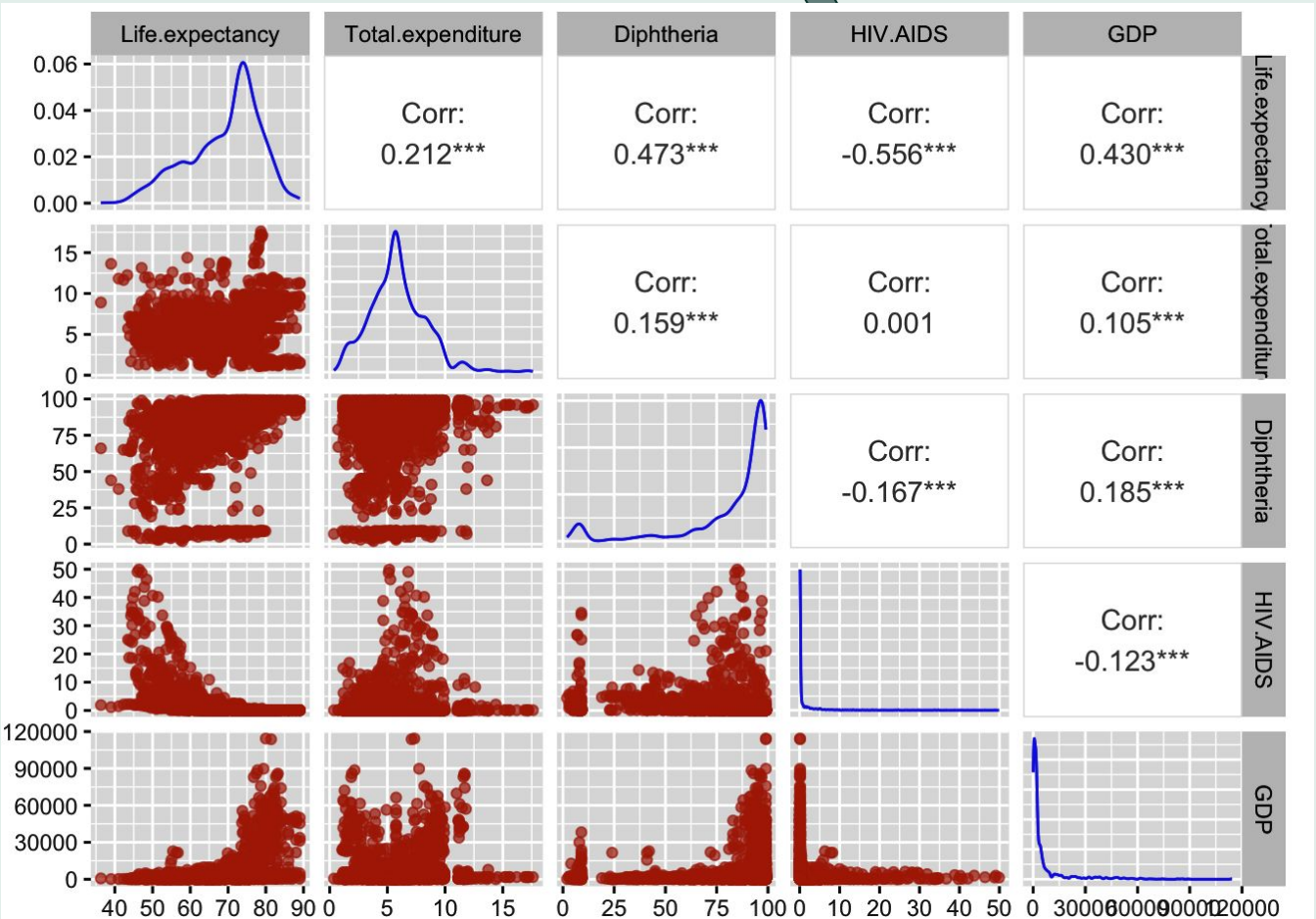
There are 2563 parameter observations that are missing.

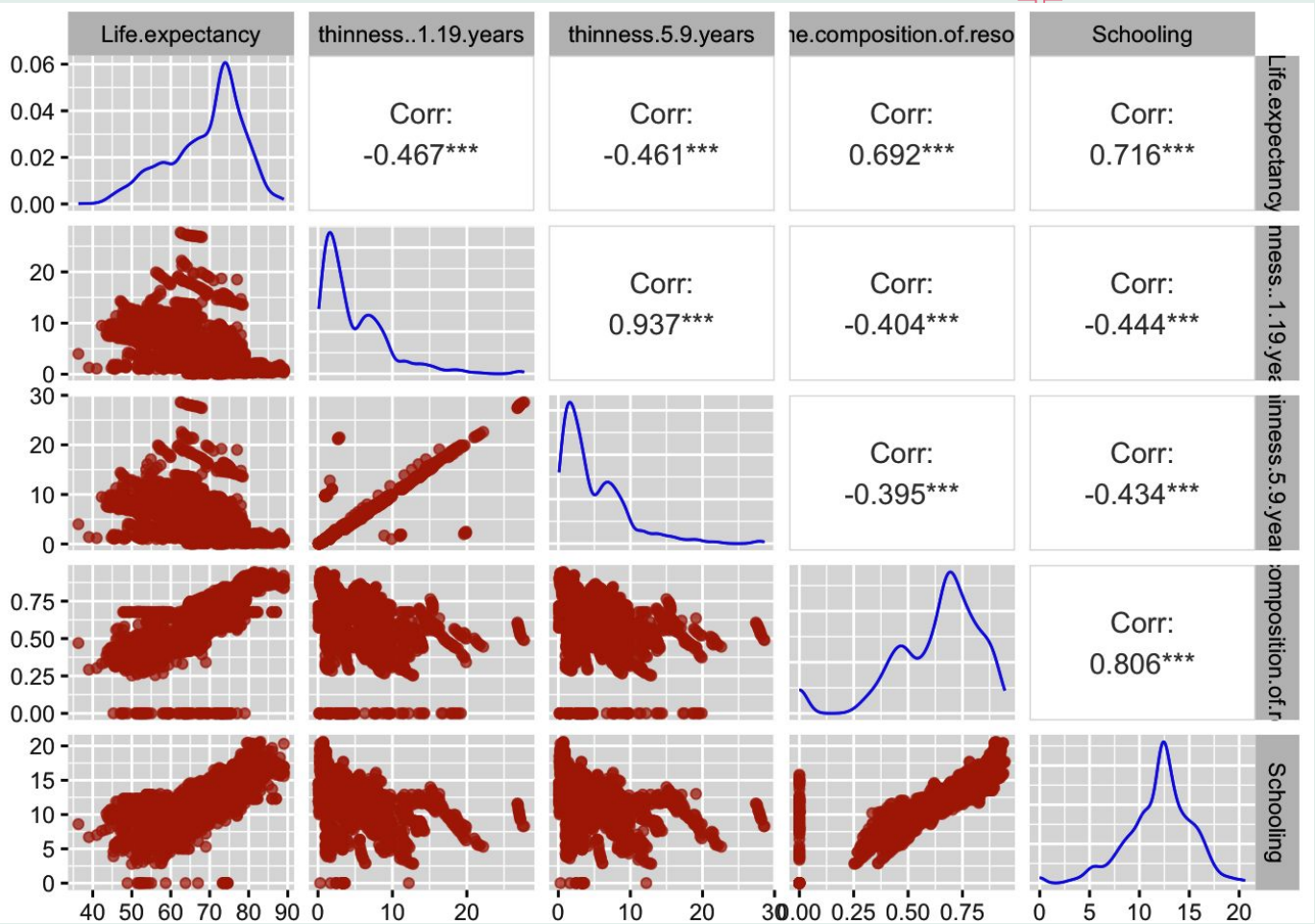
# Data: Visualisation



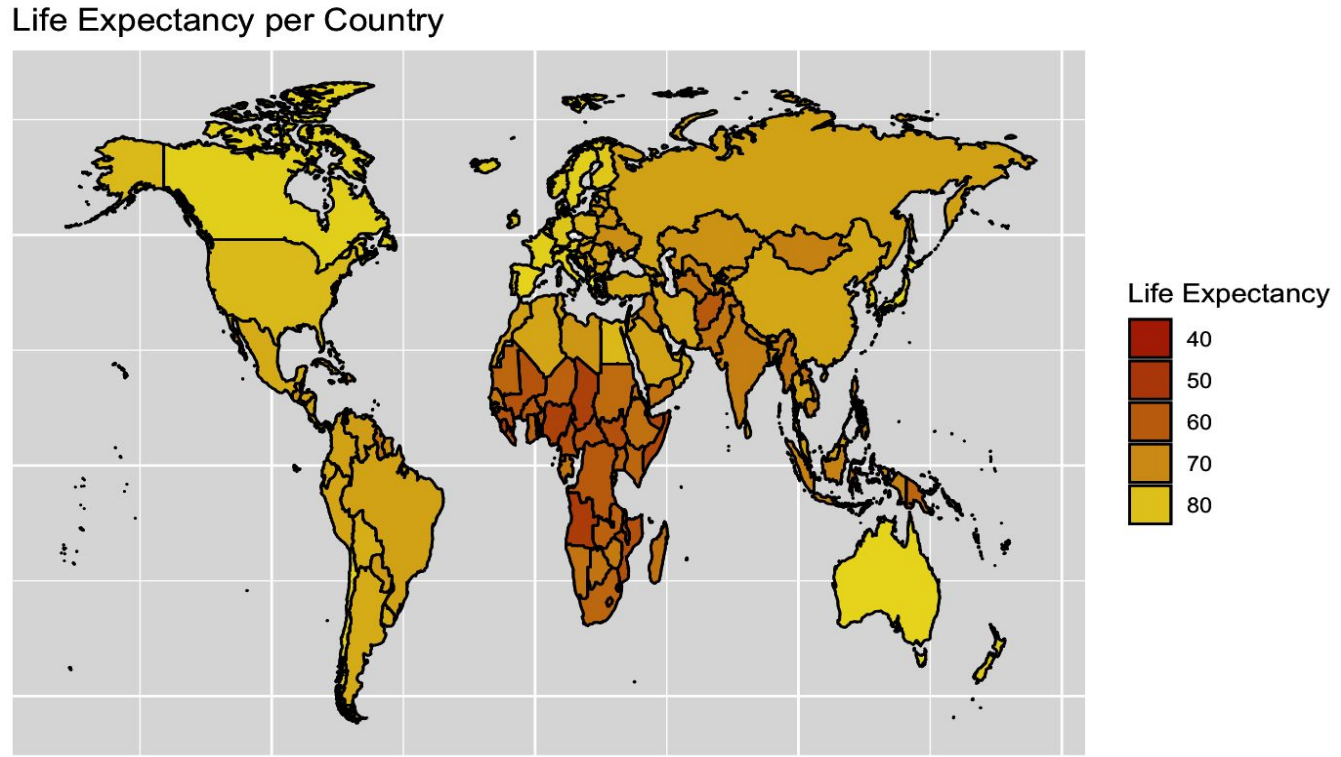

Life.expectancy
Adult.Mortality
infant.deaths
Alcohol
ercentage.expenditur
Life.expectancy
Adult.Mortality
infant.deaths
Alcohol
entage.expend
Corr:
-0.699***
Corr:
-0.196***
Corr:
0.388***
Corr:
0.381***
Corr:
0.065**
Corr:
-0.180***
Corr:
-0.242***
Corr:
-0.115***
Corr:
-0.088***
Corr:
0.352***
0.06
0.04
0.02
0.00
600
400
200
0
1500
1000
500
0
15
10
5
0
20000
15000
10000
5000
0
40 50 60 70 80 90
0 200 400 600
0 500 1000 1500
0 5 10 15
0 5000 10000 15000 20000


Life.expectancy
Measles
BMI
under.five.deaths
Polio
Corr:
-0.153***
Corr:
0.555***
Corr:
-0.222***
Corr:
0.456***
Corr:
-0.179***
Corr:
0.522***
Corr:
-0.118***
Corr:
-0.243***
Corr:
0.281***
Corr:
-0.184***
Life.expectancy
Measles
BMI
under.five.death
Polio

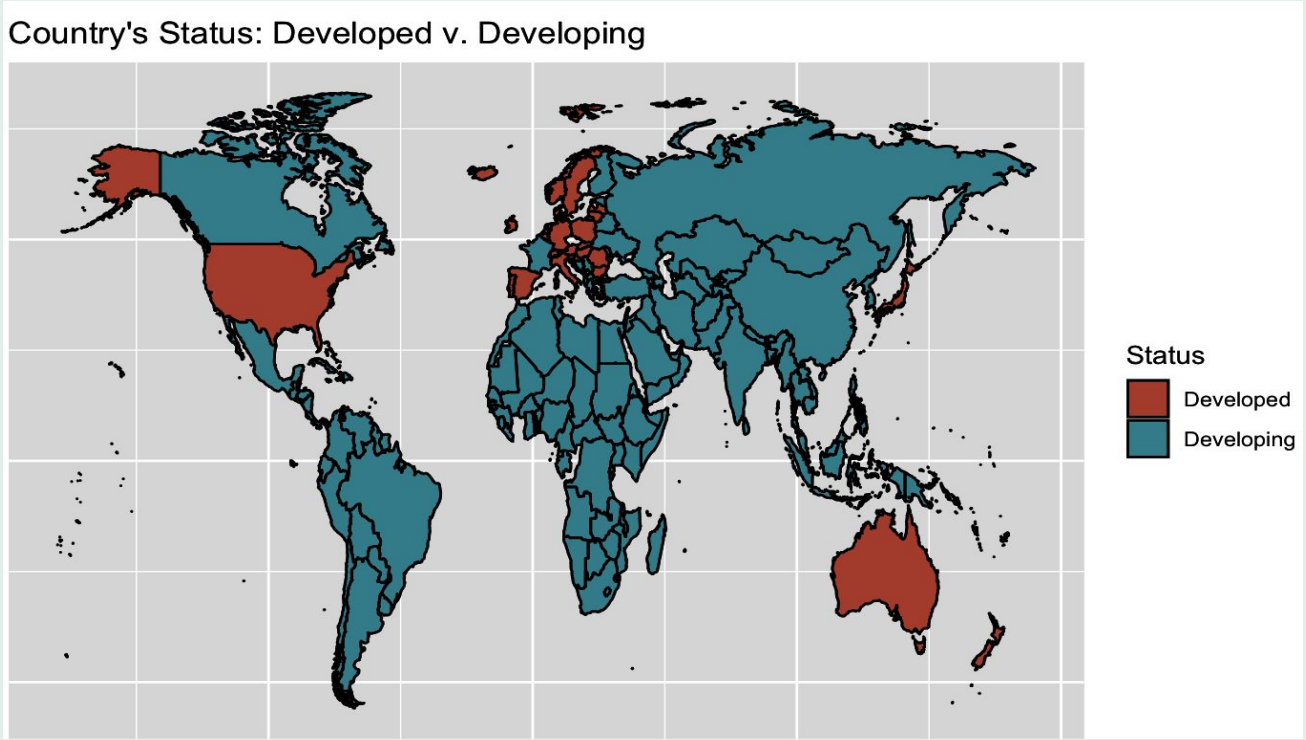

Life.expectancy
Total.expenditure
Diphtheria
HIV.AIDS
GDP
Corr:
0.212***
Corr:
0.473***
Corr:
-0.556***
Corr:
0.430***
Corr:
0.159***
Corr:
0.001
Corr:
0.105***
Corr:
-0.167***
Corr:
0.185***
Corr:
-0.123***
0.06
0.04
0.02
0.00
15
10
5
0
100
75
50
25
0
50
40
30
20
10
0
120000
90000
60000
30000
0
40 50 60 70 80 90
0 5 10 15
0 25 50 75 100
0 10 20 30 40 50
0 30000 60000 90000 120000


Life.expectancy
thinness..1.19.years
thinness.5.9.years
ne.composition.of.reso
Schooling
Corr:
-0.467***
Corr:
-0.461***
Corr:
0.692***
Corr:
0.716***
Corr:
0.937***
Corr:
-0.404***
Corr:
-0.444***
Corr:
-0.395***
Corr:
-0.434***
Corr:
0.806***
Life.expectancy
nness..1.19.yea
inness.5.9.year
omposition.of.r
Schooling

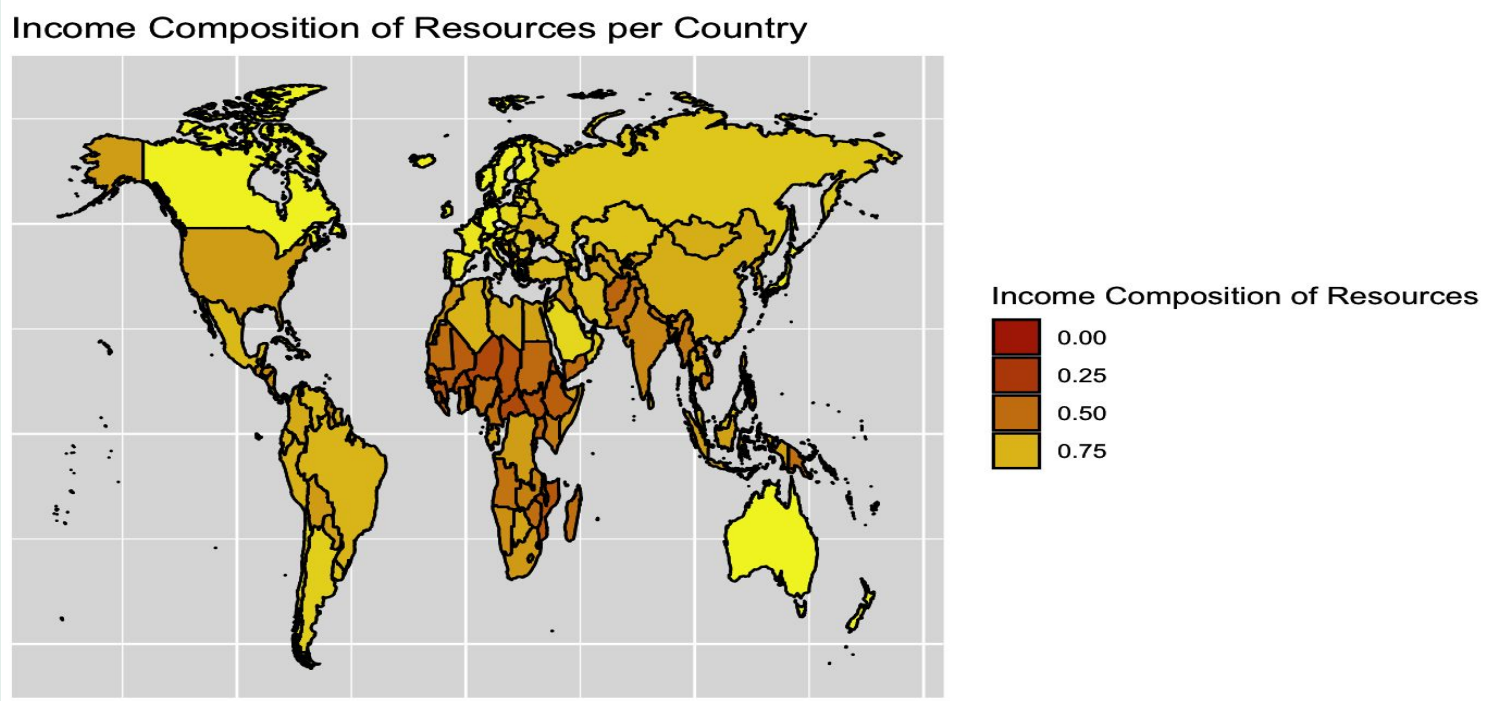
# World Map: Life Expectancy per Country



# World Map: Developed vs. Developing



# World Map: Income Composition of Resources per Country



# Models Used

## 01
### Custom Linear Regression

Eliminated redundant and statistically non-significant variables at alpha = 0.05

## 02
### Feature Selection of Linear Regression

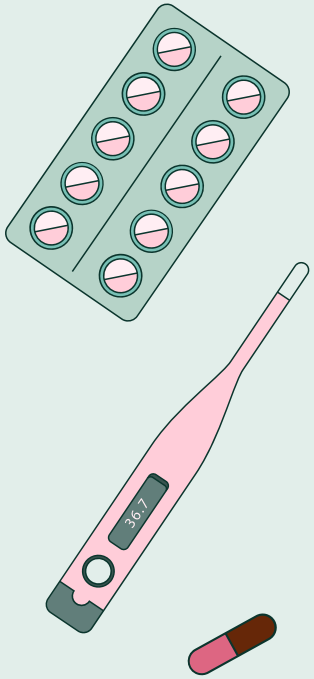
Forward, Backward, Step-wise, and penalized regression

## 03
### Complex Linear Regression

Added log transformations, polynomials, and interaction terms

## 04
### KNN

K Nearest Neighbor (Non-Parametric)

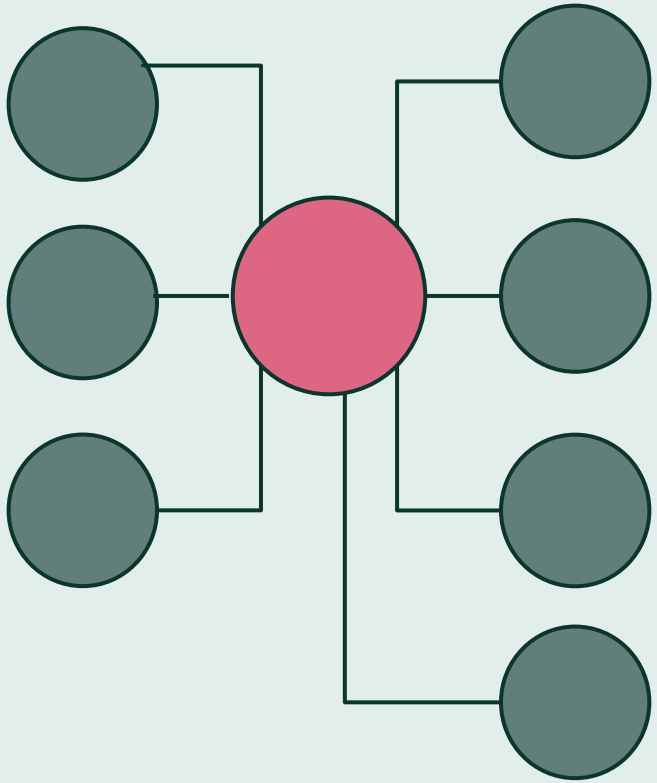


# CUSTOM LINEAR REGRESSION

## BMI
Average Body Mass Index of population

## Diphtheria
DPT3 Immunization Coverage among 1-year-olds

## GDP
Gross Domestic Product per Capita



## HIV AIDS
HIV/AIDS Deaths per 1000 live births

## Income Composition of Resources
Human Development Index in terms of income composition of resources

## Schooling
Years in School

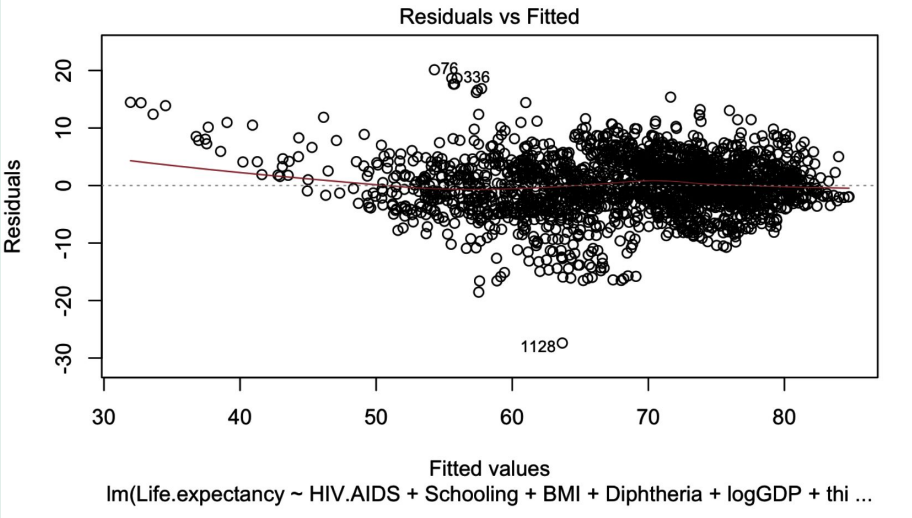
## Thinness
Prevalence of thinness among children ages 10 to 19

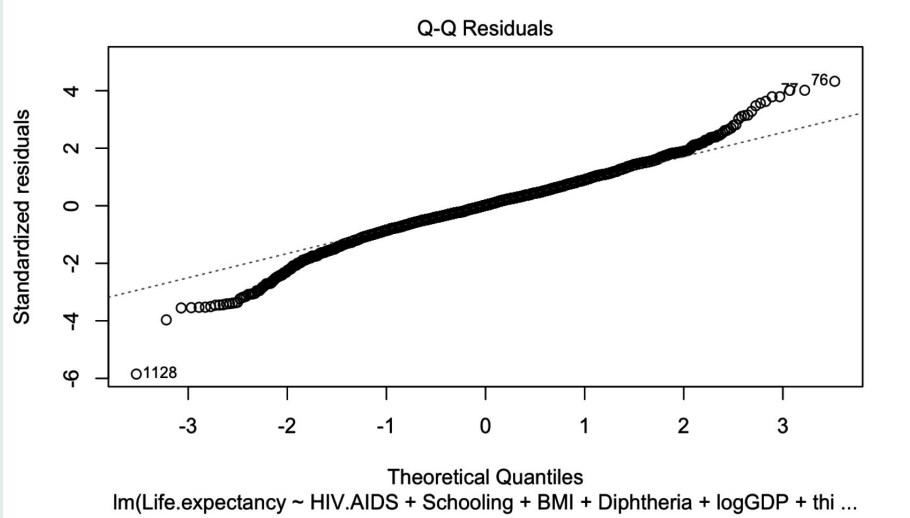
# Interpretation of Coefficients

| | |
|---|---|
| **Intercept** | We estimate that the median Life Expectancy for an individual with none of the variables present would theoretically be 43.6420 years old (p-value < 2e-16). We are 95% confident that the Life Expectancy is between 42.43 and 44.85 years. |
| **HIV AIDS** | For every 1,000 live births, there is an increase in HIV.AIDS, where we estimate that the mean of Life Expectancy will decrease by 68.0044 (p-value < 2e-16) holding all other variables constant. We are 95% confident that the true estimate is between 64.10 and 71.90 children who are born with HIV AIDS. |
| **Schooling** | For each year increase in Schooling, the mean of Life Expectancy will increase by 0.8440 (p-value < 2e-16) holding all other variables constant. We are 95% confident that the true estimate is between 0.7377 and 0.9505. |
| **BMI** | For each average kg/m^2 increase in BMI for the population, the mean of Life Expectancy will increase by 0.0561 (p-value < 2e-16) holding all other variables constant. We are 95% confident that the true estimate is between 0.0437 and 0.0685. |

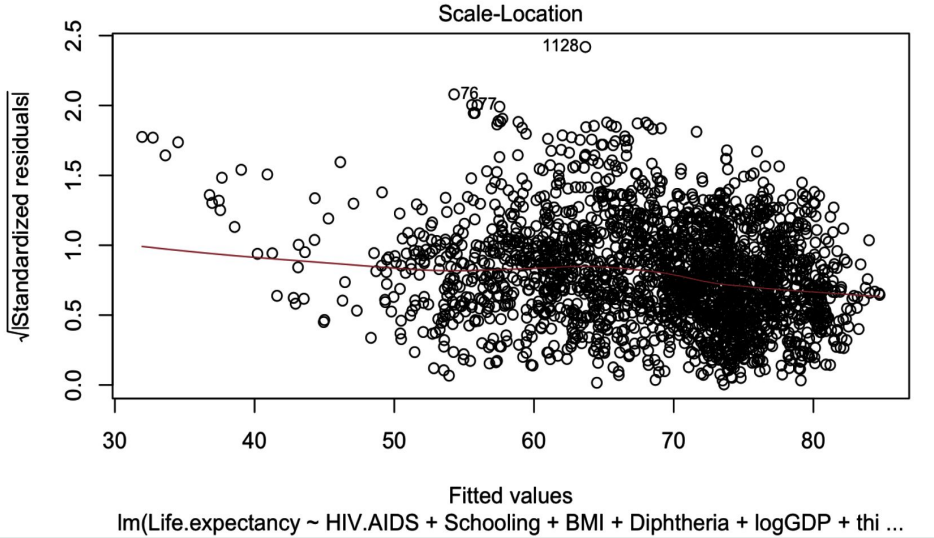
| | |
|---|---|
| **Diphtheria** | For one percent immunization coverage among 1-year olds vaccinated using DTP3 for Diphtheria, the mean of Life Expectancy will increase by 0.0675 (p-value < 2e-16) holding all other variables constant. We are 95% confident that the true estimate is between 0.0587 and 0.0762. |
| **logGDP** | For a doubling of logGDP per capita, the mean of Life Expectancy will change by 0.599579*ln (2) = 0.4155965 (p-value < 2e-16) holding all other variables constant. We are 95% confident that the true multiplicative increase is between 0.46073069 *ln(2) = 0.3193542 and 0.73842704 *ln(2) = 0.5118386. |
| **Thinness** | For 1 percent increase in Thinness (10-19 years) among children in adolescence, the mean of Life Expectancy will decrease by 0.1237 (p-value 4.58e-06). We are 95% confident that the true estimate is between 0.1765 and 0.0709. |
| **Income Composition of Resources** | For every unit increase in Income Composition of Resources, the mean of Life Expectancy will increase by 7.9522 (p-value < 2e-16). We are 95% confident that the true estimate is between 6.3250 and 9.5793. |

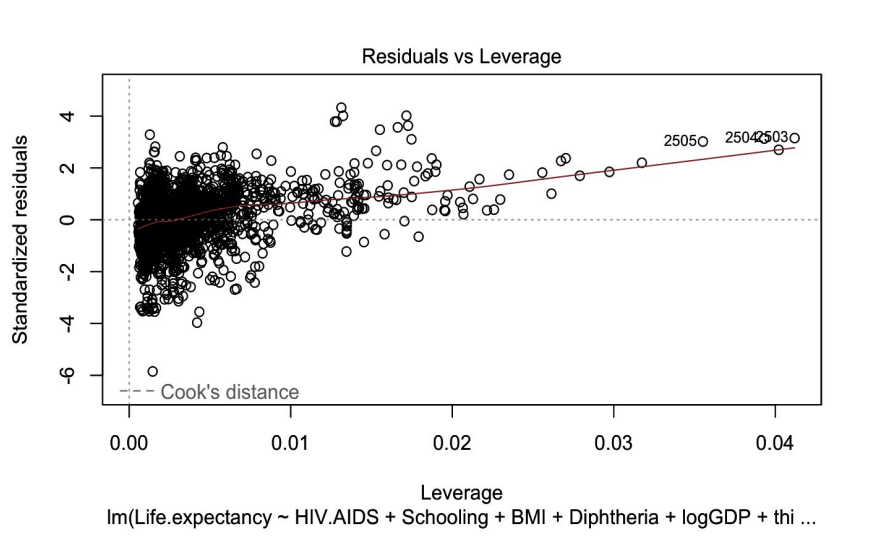
# Custom Model Residuals



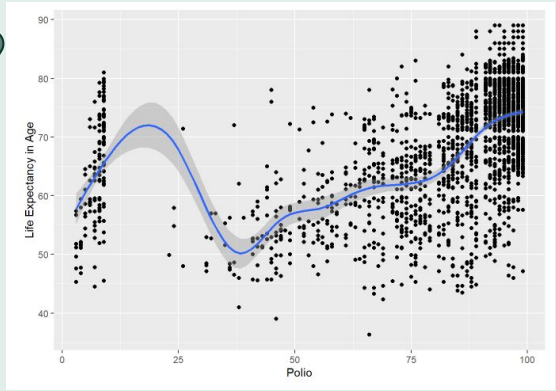



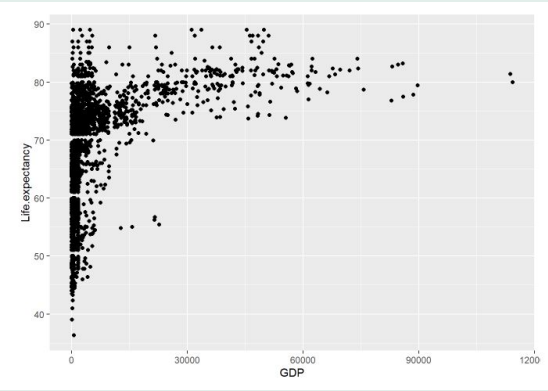
# Custom Model Residuals

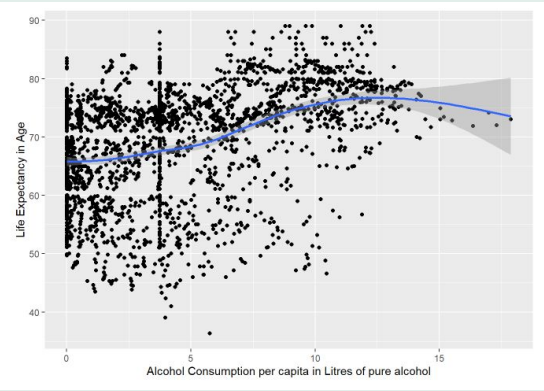


# Complex Linear Regression




## Added Polynomial Terms

Alcohol, BMI, Diphtheria, Polio, Schooling

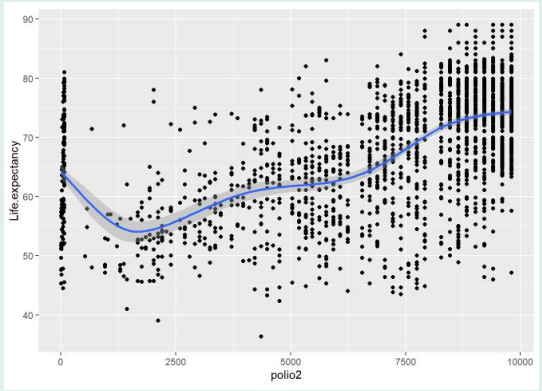
## Log-Transformed Variables

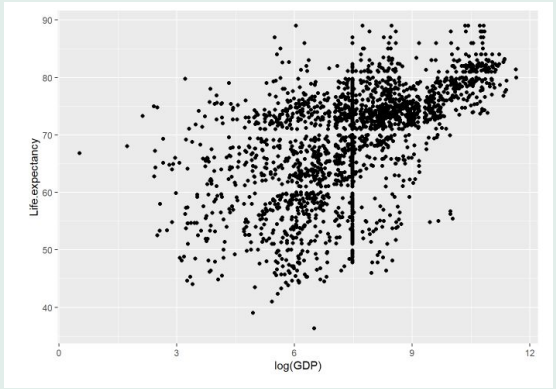
GDP, HIV AIDS, Thinness

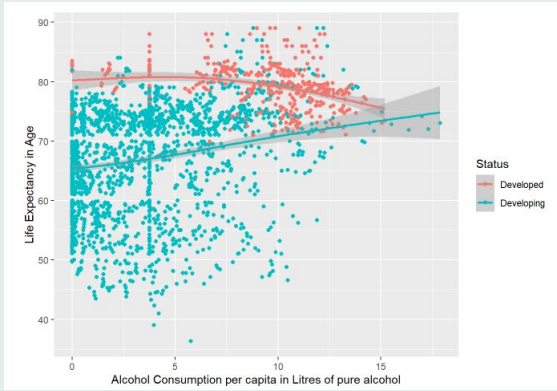
## Added Interactions

Status with each variable

# Complex Linear Regression




## Added Polynomial Terms

Alcohol, BMI, Diphtheria, Polio, Schooling

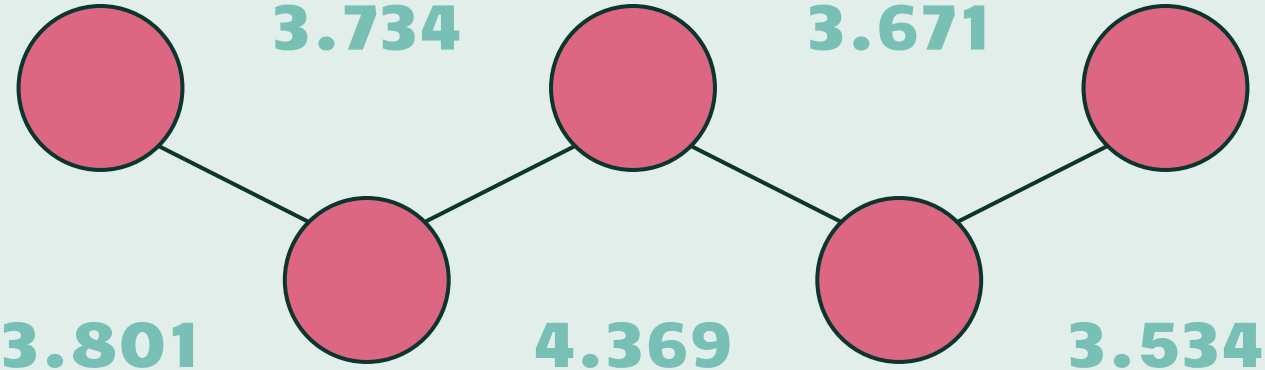
## Log-Transformed Variables

GDP, HIV AIDS, Thinness

## Added Interactions

Status with each variable

# JOURNEY OF RMSE (CLR)



Logged Variable & Septic Regression

**3.734**

Logged & Quadratic Model w/Interactions

**3.671**

**3.801**

Logged Variable & Quadratic Regression

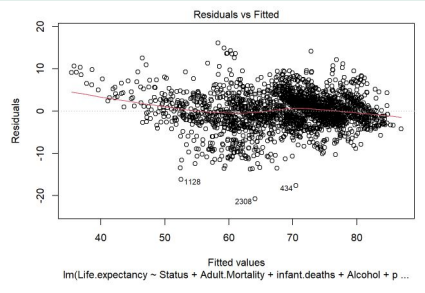
**4.369**

Variables with Interaction Term

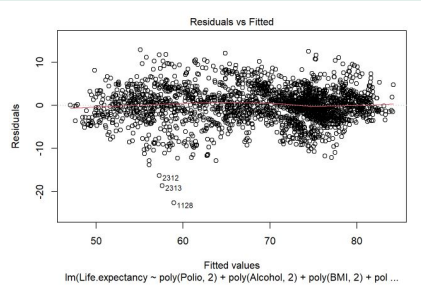
**3.534**

Logged & Septic Model with Interactions

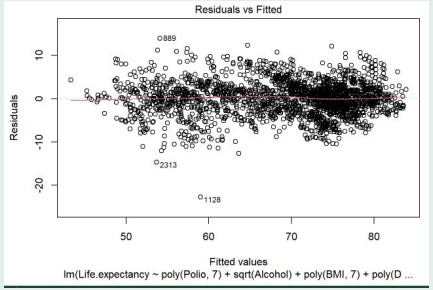
# Residuals



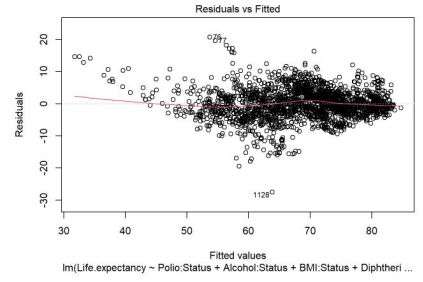

Full Model



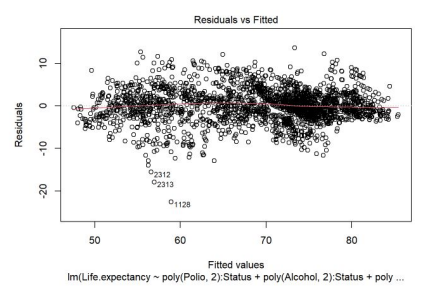

Quadratic Model



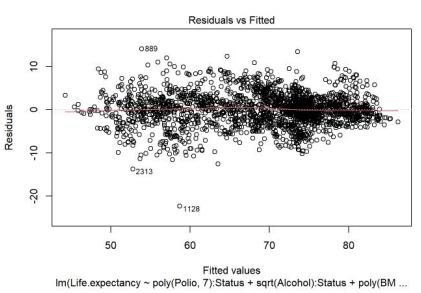

Septic Model




Interaction Term

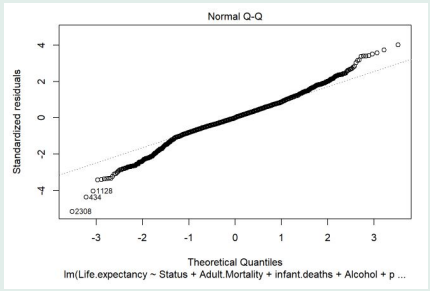



Quadratic Model + Interaction Term

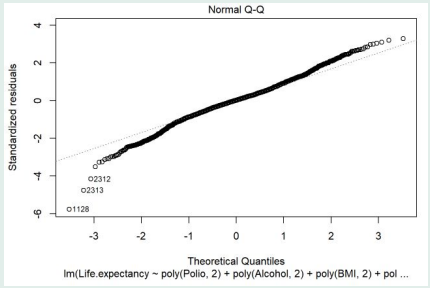



Septic Model + Interaction Term

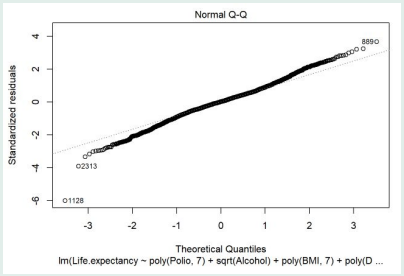
# QQ Plots



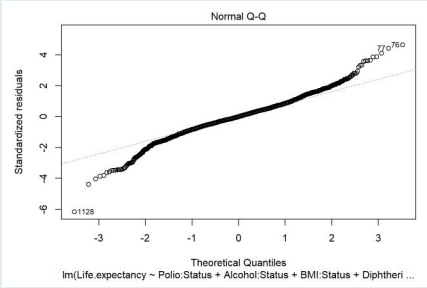

Full Model



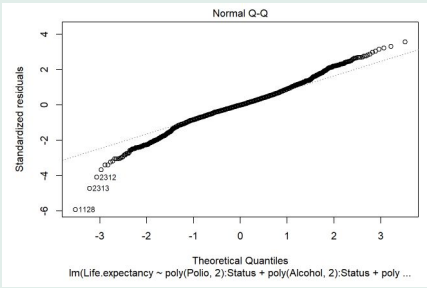

Quadratic Model



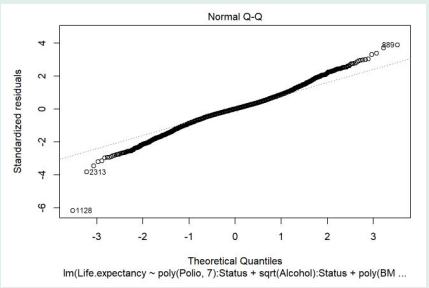

Septic Model




Interaction Term

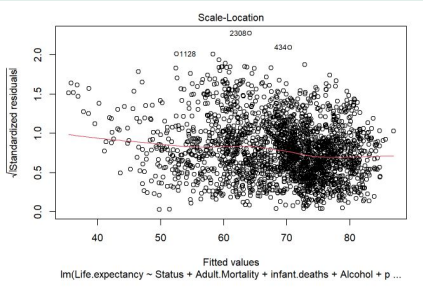



Quadratic Model + Interaction Term

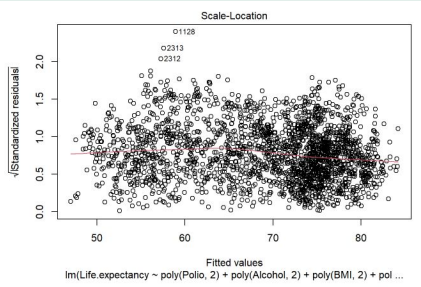



Septic Model + Interaction Term

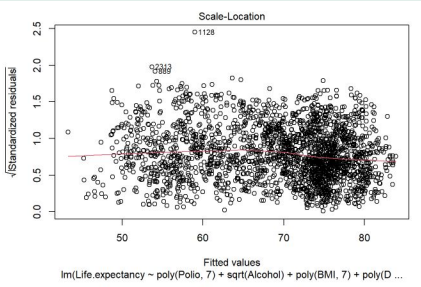
# Standardized Residuals



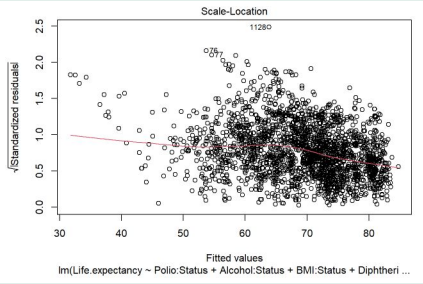

Full Model



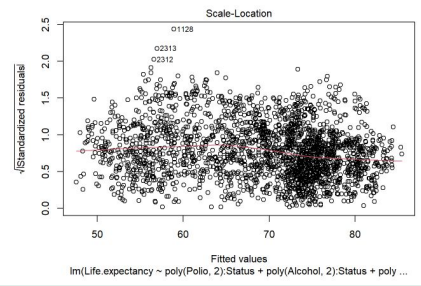

Quadratic Model



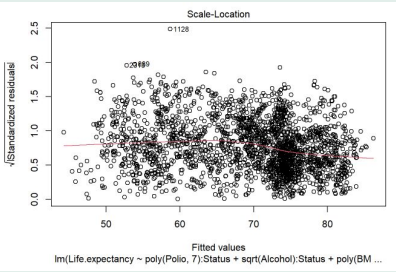

Septic Model




Interaction Term

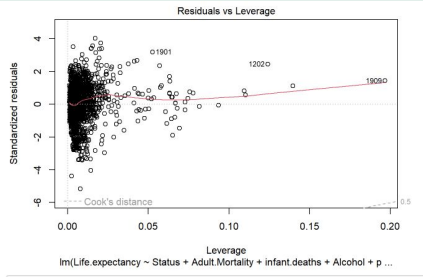



Quadratic Model + Interaction Term

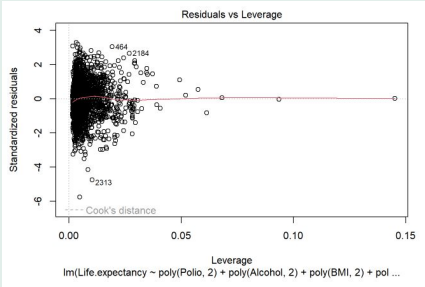



Septic Model + Interaction Term

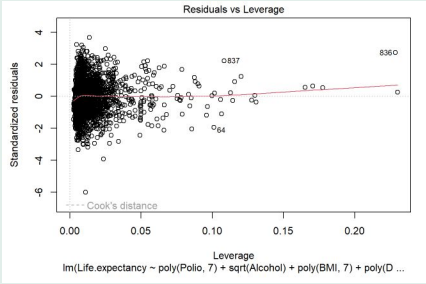
# Residuals vs Leverage



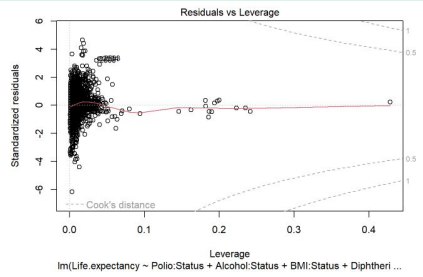

Full Model



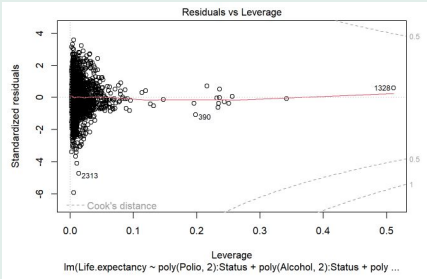

Quadratic Model



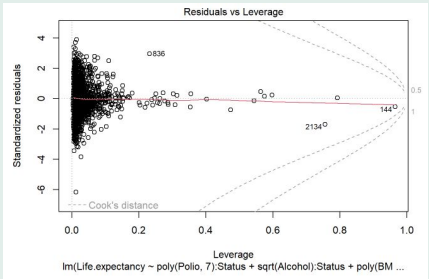

Septic Model




Interaction Term




Quadratic Model + Interaction Term




Septic Model + Interaction Term

# Feature Selection

## Forward

Selected 10/15 Parameters

67%

## Backward

Selected 13/15 Parameters

87%

## Stepwise

Selected 9/15 Parameters

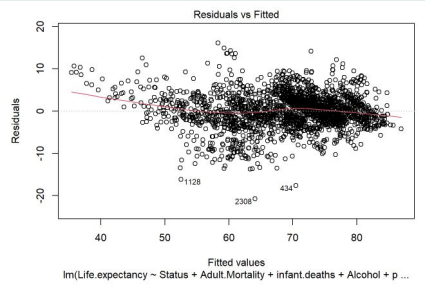
60%

## Penalized Regression

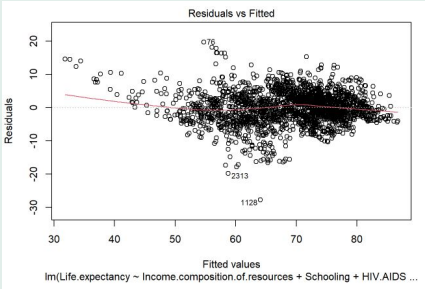
Selected 14/15 Parameters

93%

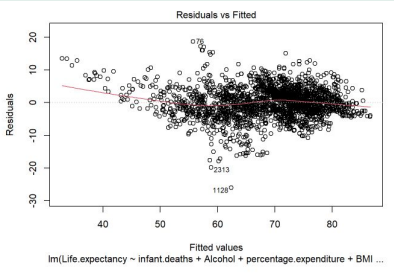
# Residuals



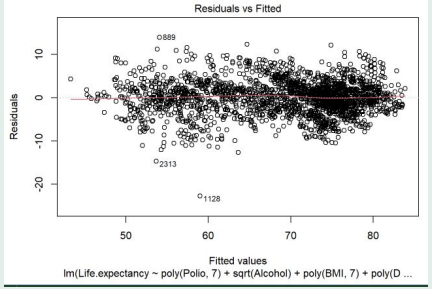

Full Model



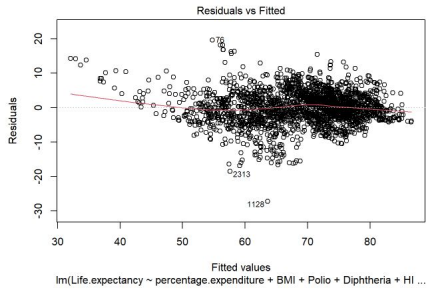

Forward Selection



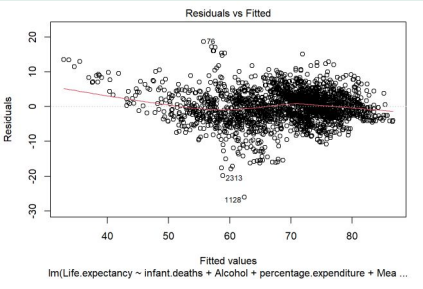

Backward Selection




Septic Model

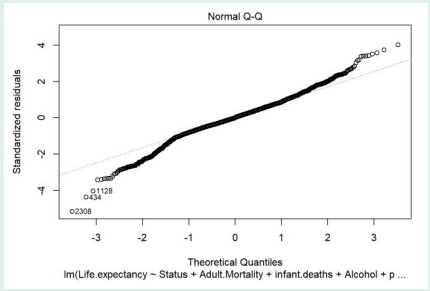



Stepwise Selection

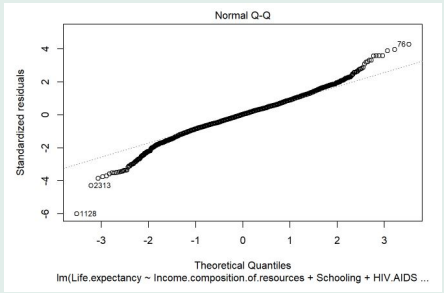



Penalized Regression

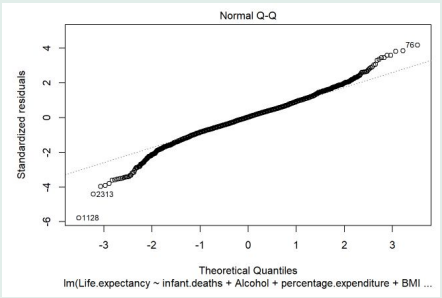
# QQ Plots



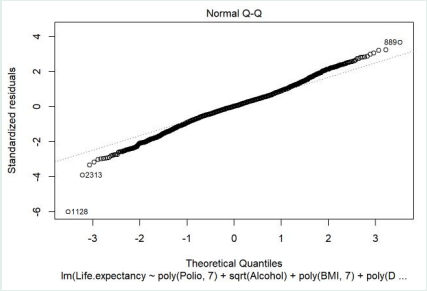

Full Model



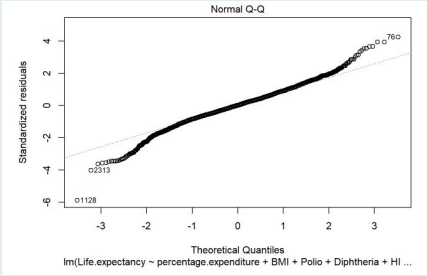

Forward Selection



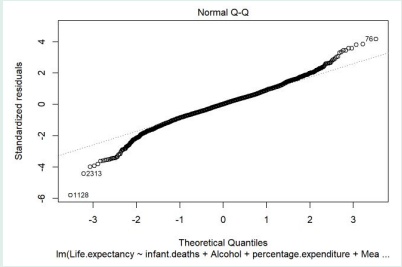

Backward Selection




Septic Model

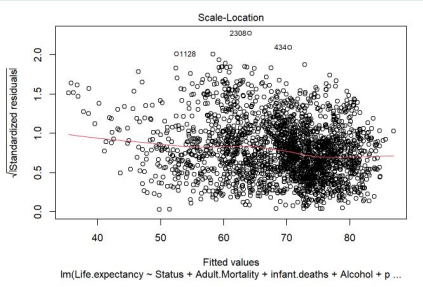



Stepwise Selection

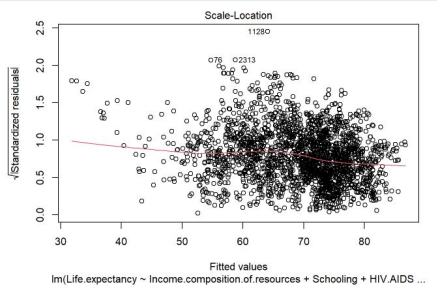



Penalized Regression

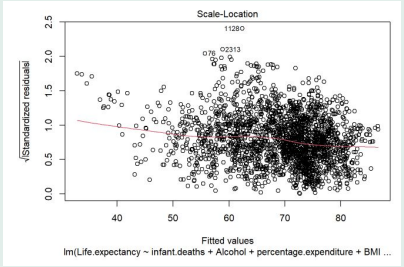
# Standardized Residuals



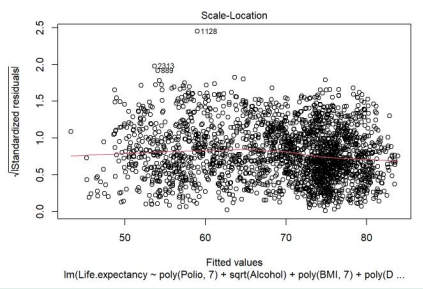

Full Model



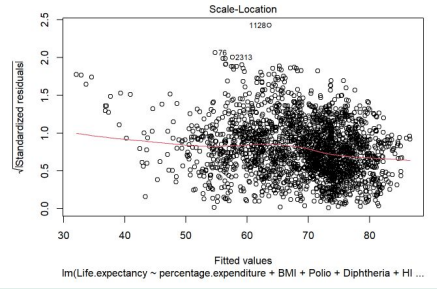

Forward Selection



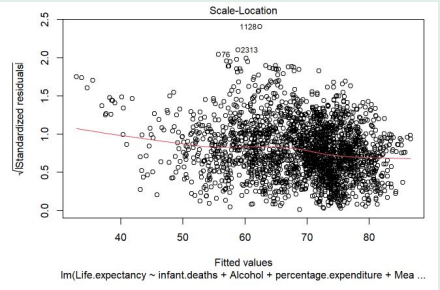

Backward Selection




Septic Model

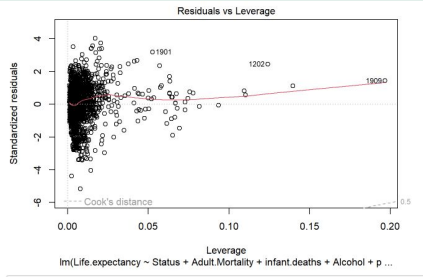



Stepwise Selection

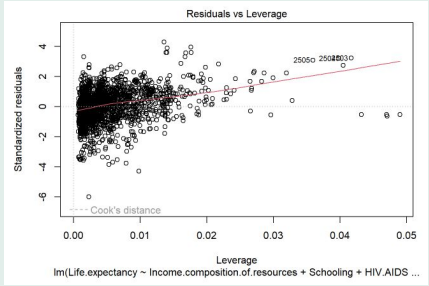



Penalized Regression

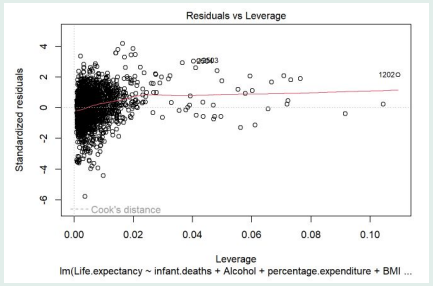
# Residuals vs Leverage



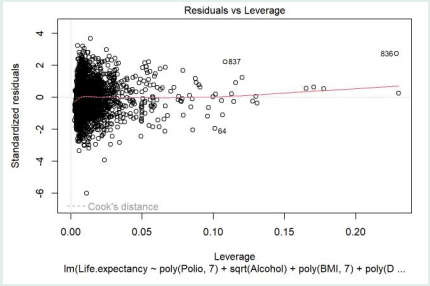

Full Model



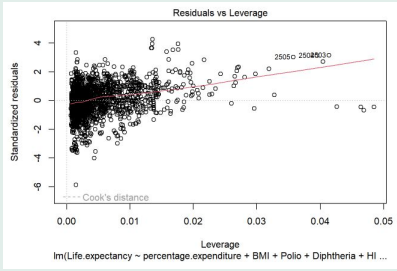

Forward Selection



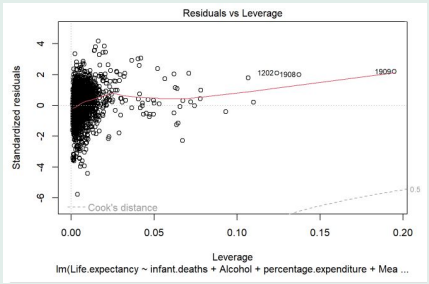

Backward Selection




Septic Model

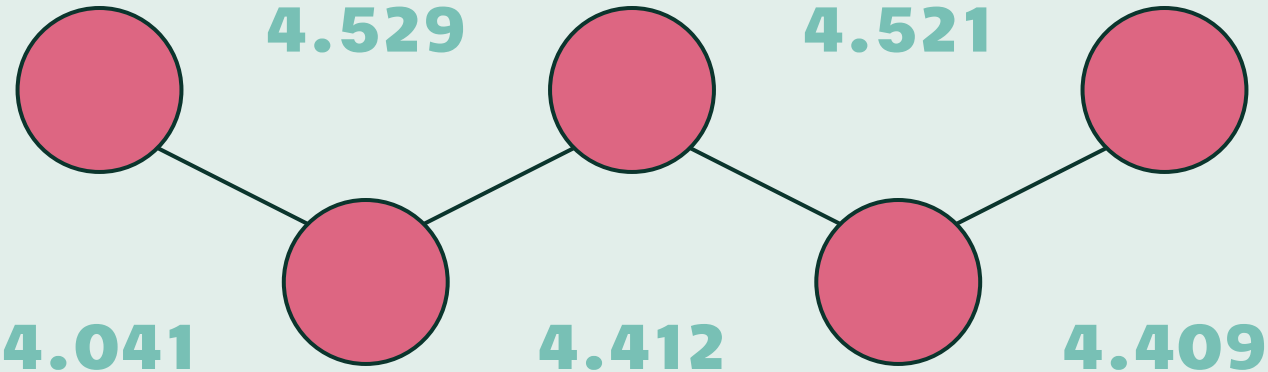



Stepwise Selection

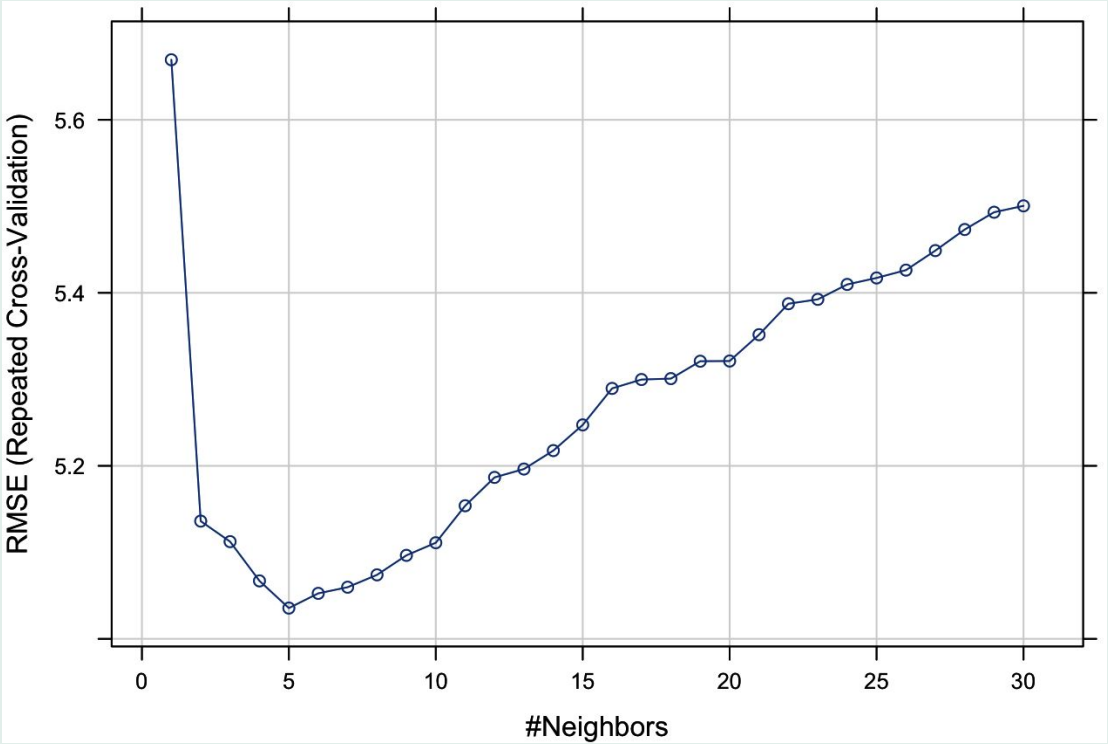



Penalized Regression

# JOURNEY OF RMSE (Feature Selection)



# KNN




RMSE:
4.840354

# Conclusion: Best RMSE of Models

## 4.650

### Custom

7 Variables

## 3.534

### Complex

11 Logged or Septic Function Variables with interactions

## 4.409

### Feature Selection

15 Variables,
Penalized Regression

## 4.840

### KNN

16 Variables, k = 5

# Final Thoughts

What Objective 1 means.

What Objective 2 means.

# Scope of Inference

## Repeated Measures

## Time Series

# Thank you!


## Xavier Mojica

xmojica@smu.edu



## Alexandra Thibeaux

athibeaux@smu.edu


# CONTENTS OF THIS TEMPLATE

This is a slide structure based on a multi-medical presentation
You can delete this slide when you're done editing the presentation

| | |
|---|---|
| **FONTS** | To view this template correctly in PowerPoint, download and install the fonts we used |
| **USED AND ALTERNATIVE RESOURCES** | An assortment of graphic resources that are suitable for use in this presentation |
| **THANKS SLIDE** | You must keep it so that proper credits for our design are given |
| **COLORS** | All the colors used in this presentation |
| **INFOGRAPHIC RESOURCES** | These can be used in the template, and their size and color can be edited |
| **CUSTOMIZABLE ICONS** | They are sorted by theme so you can use them in all kinds of presentations |

For more info:
SLIDESGO | SLIDESGO SCHOOL | FAQS

You can visit our sister projects:
FREEPIK | FLATICON | STORYSET | WEPIK | VIDFY