2022

Predicting Life Expectancy using Bayesian Techniques

ST406 FINAL PROJECT

S/16/844

DEPARTMENT OF STATISTICS AND COMPUTER SCIENCE | UNIVERSITYOF PERADENIYA

1. Introduction

The term “Life Expectancy” means number of years a person can live. By definition, Life expectancy is based on estimate of average of people’s lives in a population. There are two types of life expectancies which are cohort life expectancy and period life expectancy. The cohort life expectancy is average life expectancy of a cohort. We can track group of people born in a particular year in long decades ago. Then we can track which date they died. From these information we can calculate average life expectancy. This is known as cohort life expectancy. Periodic life expectancy is the alternative method to calculate the average life expectancy. We can calculate this using morality rates observed at one particular period from birth to death. This is much more commonly used in life expectancy metric and periodic life expectancy values are different from cohort life expectancy figures.

In some situations, we have to estimate the life expectancy by looking at prior data. We can get only an average value by life expectancy. But we have to calculate nearest value for age of a human in different kind of situations. Life depends on various components which are diseases, the things that people use and health of the body. By referring these components and prior data we have to estimate the life expectancy of each person. Here we used Bayesian Multiple Linear Regression to estimate the life expectancy of a human.

1.1. Motivation

Life expectancy plays an important role when taking decisions about final phase of life. It is beneficial to individuals, health service providers and the government. For instance, it would make people more aware from their general health like bad habit that are occur to reduce their life expectancy, and life expectancy improvements. Predicting life expectancy give motivation them to make healthier life style. This prediction also use by insurance companies to provide individualized services such as reducing premiums. The government also use these predictions to get idea about social welfare assistance and health care funding. Pharmaceutical companies can coordinate some medical campaigns based on the life expectancy predictions. Therefore life expectancy prediction using statistical methods is more useful thing nowadays.

1.2. Background of the Study

The researchers had been carried out previously to predict the ‘Life Expectancy’ using different types of machine learning and statistical models. The researchers review the evidence on Life Expectancy difference between men diagnosed with PCa and general population. To examine clinician and model predicted life expectancy publicly available life expectancy calculators. The accuracy of clinician-predicted survival is limited. Statistical models offer improvement in discrimination. (Jesse et al.,2015)

In this study the authors aimed to produce a model for predicting the life expectancy of children with severe cystic fibrosis (CF) lung disease. They used survival of 181 children with severe CF lung disease referred for transplantation assessment were studied. Proportional hazards were used to identify assessment measurements that of value in predicting longevity. (Aurora et al.,2000)

Lung cancer survival rate is very limited post-surgery irrespective. The researchers used many machine learning techniques to predict the life expectancy. They used Multi-Layer Perception (MLP), SVM, Naïve Bayes, Decision Tree, Random Forest and Logistic Regression. This study has been carried out by attribute ranking and selection in performing to develop prediction accuracy. (Akshaya et al., 2021)

The statistician has been written lots of books by describing Bayesian theories from top to bottom (Merlise et al.,2021) and they describe how to use them in real world applications with different types of statistical softwares. But most of the statisticians mainly focus on RStudio (Alicia et al.,2021)

1.3. Objectives

- Fitting Bayesian Multiple Linear Regression Model.
- Find the best model using Backward elimination with BIC value.
- Check the model uncertainty.
- Discuss the Posterior means and posterior standard deviations and Credible Intervals Summary.

2. Materials and Methods

2.1. Description of the Sample

The dataset related to life expectancy, health factors for 193 countries. It collects from same WHO data repository website and its corresponding economic data was collected from United Nation website. This dataset has past 15 years data which are 2939 records/people and 20 variables. The individual data files have been merged together to a single data file in “Kaggle” website.

Variable	Description
Status	Developing or Developed
Adult Mortality	Probability of dying between 15 and 60 years per 1000 population
Infant Deaths	Number of infant deaths per 1000 population
Alcohol	Alcohol recorded for capita (15+) consumption (in litres of pure alcohol)
Percentage Expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita (%)

Hepatitis B	Hepatitis B (HepB) immunization coverage among 1 year olds (%)
Measles	Number of reported cases for 1000 population
BMI	Average Body Mass Index of entire population
Under-five Deaths	Number of under five deaths per 1000 population
Polio	Polio (Pol3) immunization coverage among 1 year olds (%)
Total Expenditure	General government expenditure on health as a percentage of on total government expenditure (%)
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1 year olds (%)
HIV/AIDS	Deaths per 1000 live births HIV/AIDS (0-4 years)
GDP	Gross Domestic Product per capita (in USD)
Population	Population of the country
Thinness 1_19 years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
Thinness 5- years	Prevalence of thinness among children for age 5 to 9 (%)
Income Composition of Resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Number of years of schooling (years)
Life Expectancy	Life Expectancy in age

Table 1

2.2. Statistics inference of the sample

- Density plots are used to observe the distribution of a variable in a dataset. The peaks of the density plots help to represent where the values are concentrated with the interval.

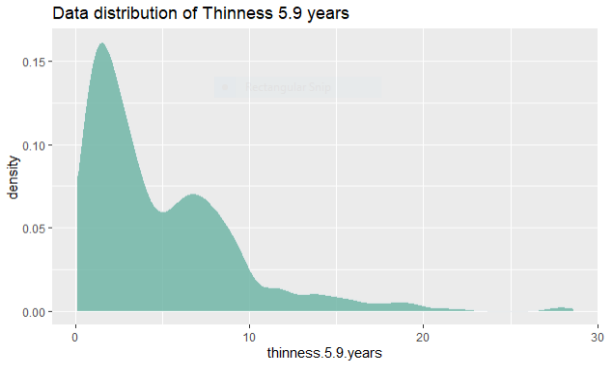

Figure 1

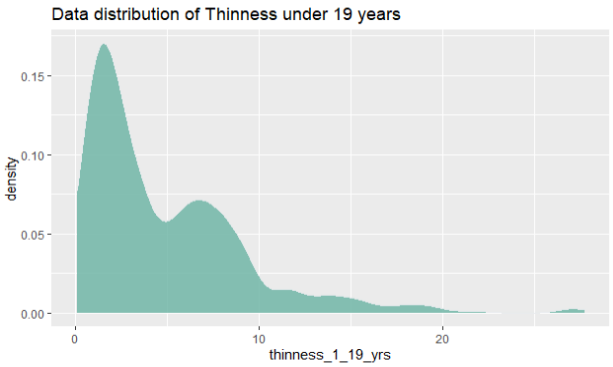

Figure 2

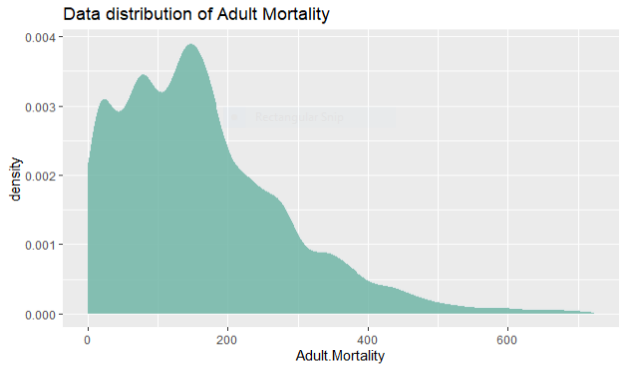

Figure 3

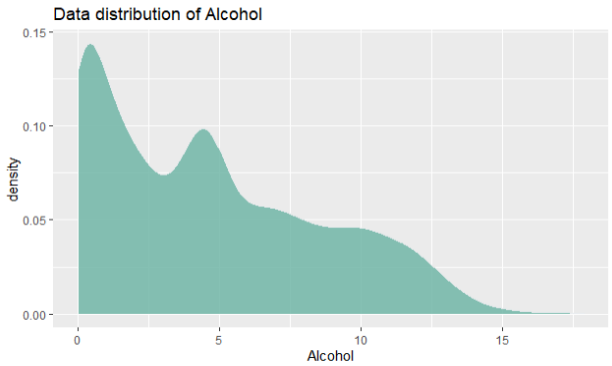

Figure 4

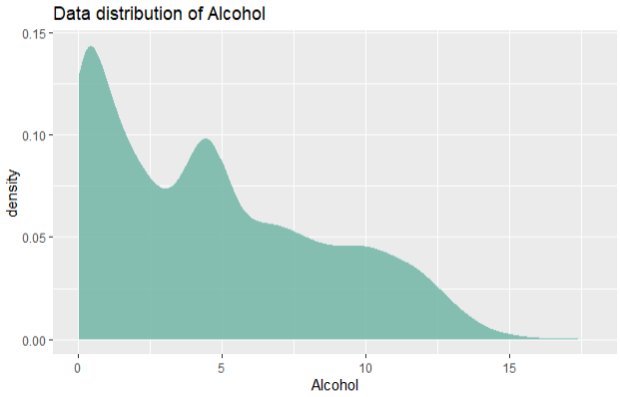

Figure 5

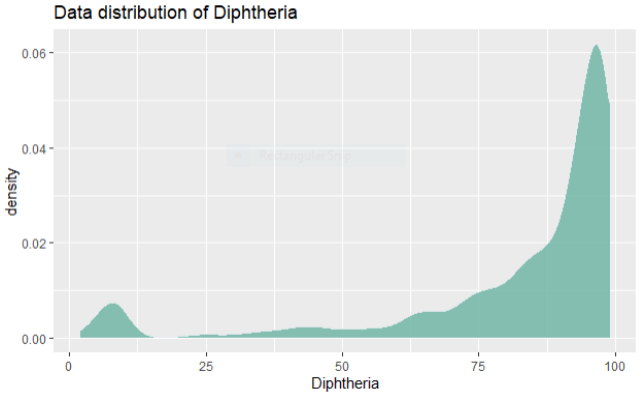

Figure 6

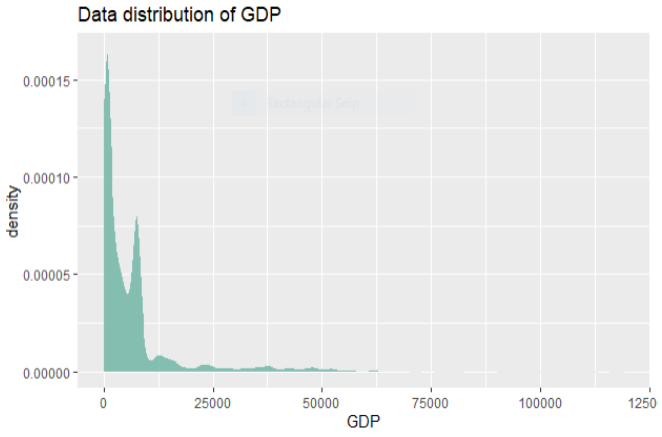

Figure 7

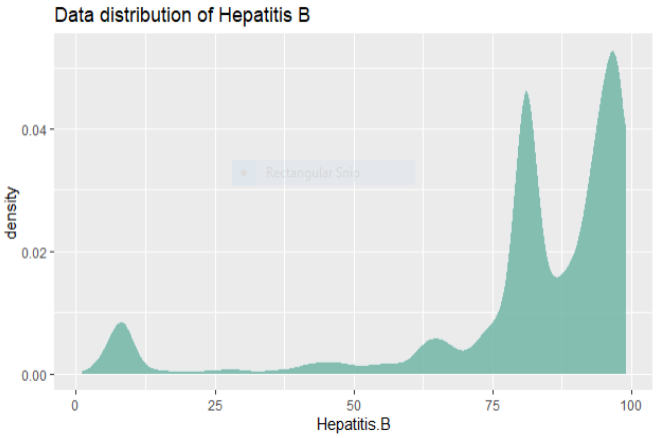

Figure 8

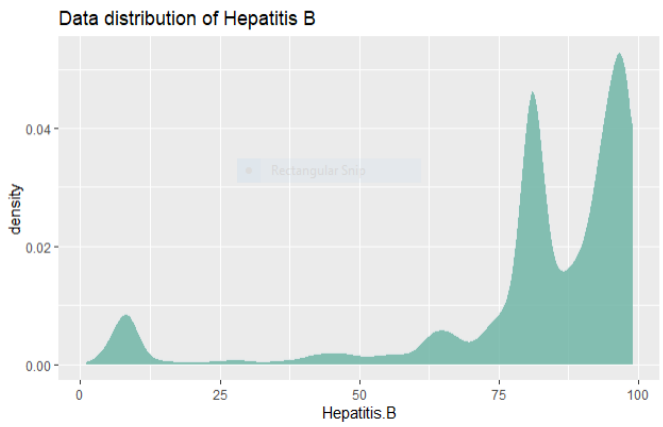

Figure 9

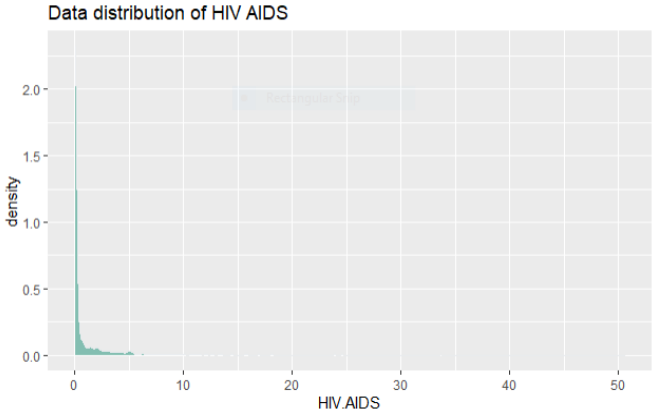

Figure 10

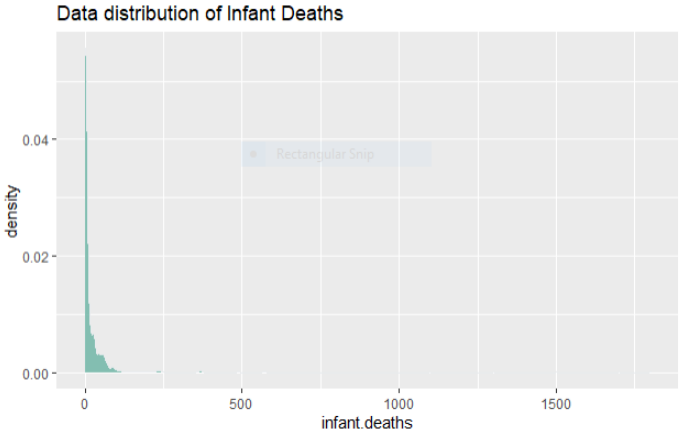

Figure 11

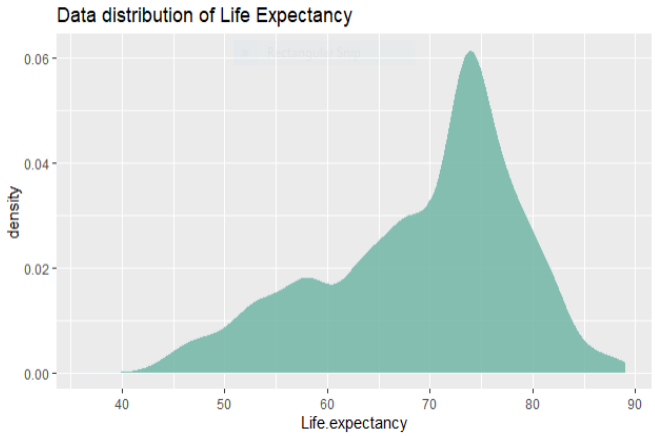

Figure 12

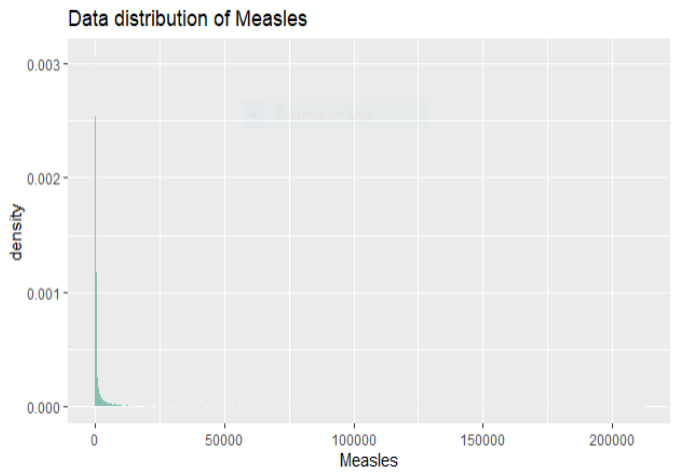

Figure 13

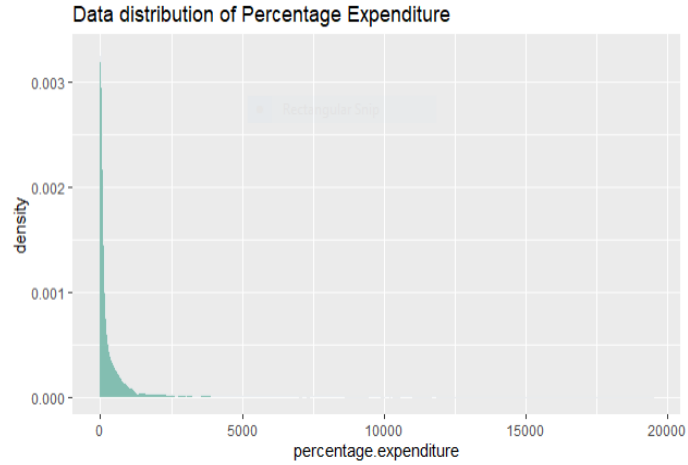

Figure 14

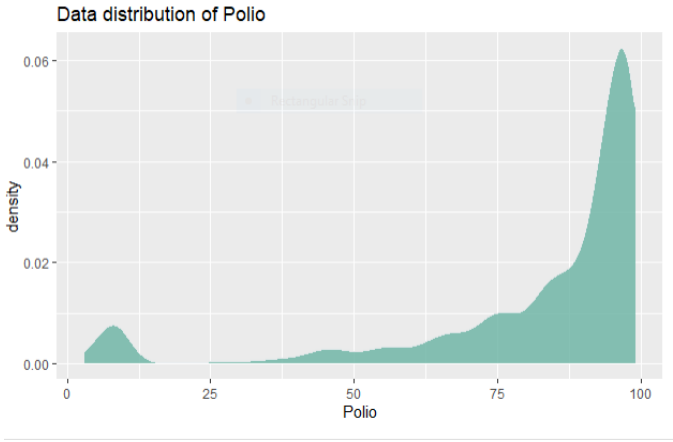

Figure 15

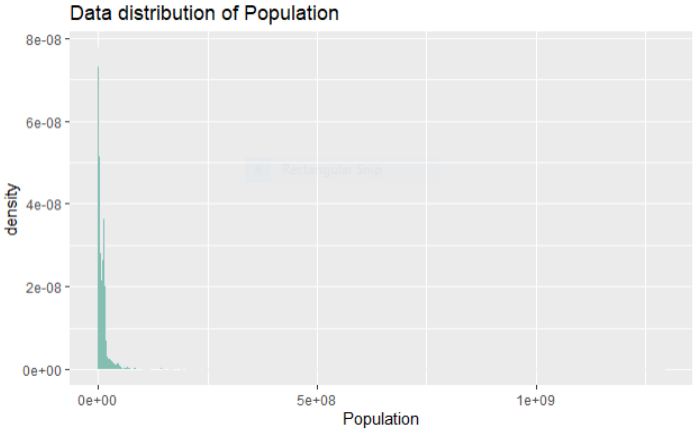

Figure 16

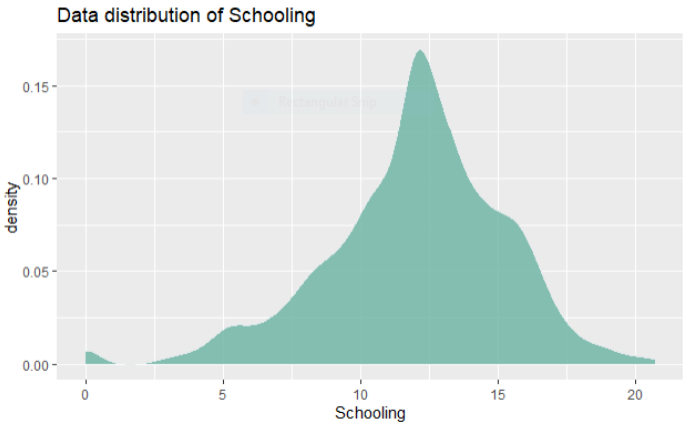

Figure 17

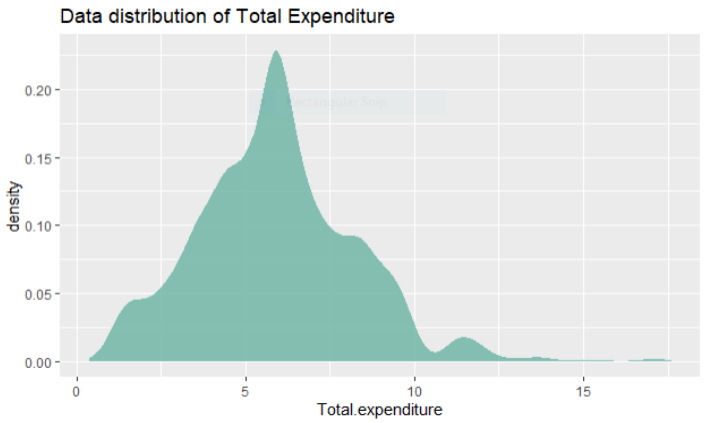

Figure 18

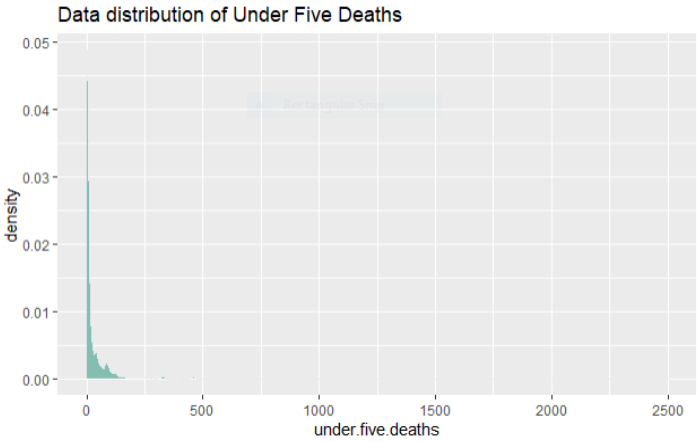

Figure 19

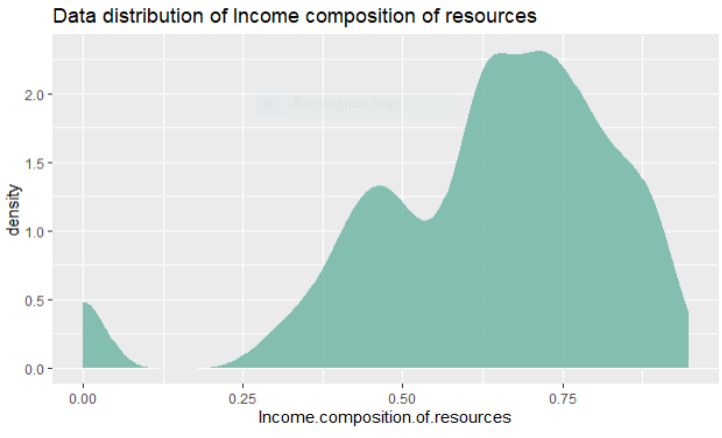

Figure 20

- *Figure 1* to *Figure 20* show how the distribution of each variables.
- Pie charts can represent the distribution of the categorical variables.

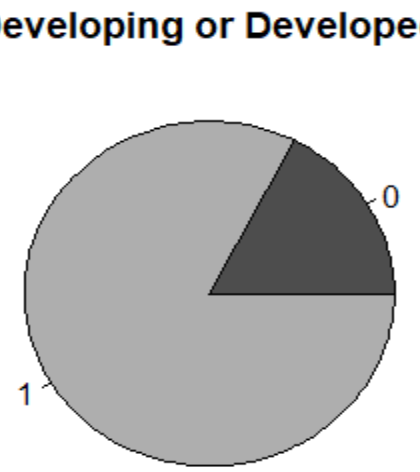

Figure 21

- 'Developing' represent by label 1 and 'Developed' represents by label 0 in *Figure 21.*
- Correlation heat map is a two dimensional plot. It represents amount of correlation (dependency between two variables) between variables by colors.

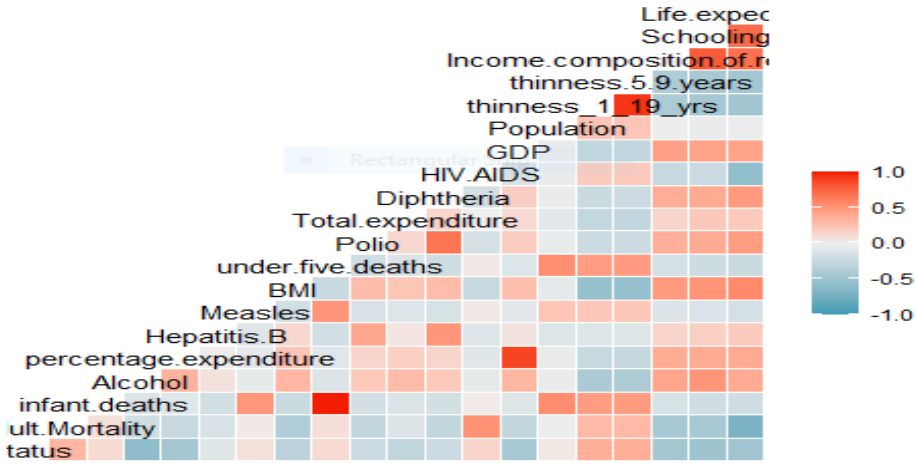

Figure 22

- *Figure 21* represents the correlation between each variable to each variable by colors. It shows correlation of the most of the variables between 0 – 0.5 and 0 – (-0.5) . Therefore we can conclude that there are no high correlations between the variables. It is advantage of this dataset because high correlation between the variables can affect for the final results of regression analysis.

2.3.Statistical Analysis

Statistical analysis was done by R programming language using Rstudio. Both graphical and inference methods were used for preliminary analysis.

In this project 'Life Expectancy' is predicted by applying Bayesian Multiple Linear Regression. This analysis discuss about Bayesian Model selection using BIC method, posterior means and posterior standard deviation summary, credible intervals summary, visualizing model uncertainty, finding most needed variables to the model by visualizing. Before fitting the models and doing these Bayesian based analysis, I found the missing values of the dataset and did missing value imputation by replacing mean of the each corresponding variables.

2.3.1. Missing values imputation

There can be missing values because of many different reasons. But we can't remove them row wise because it removes original values of other variables also. There for we can do missing value prediction or missing value imputation. I used missing values imputation by corresponding variable mean.

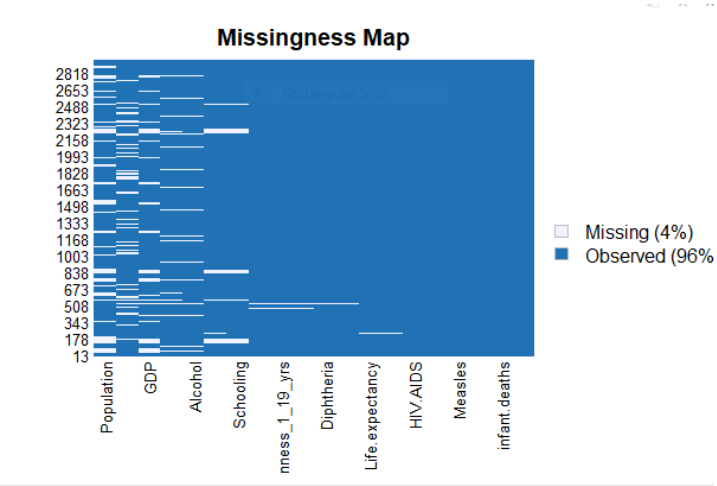

Figure 23

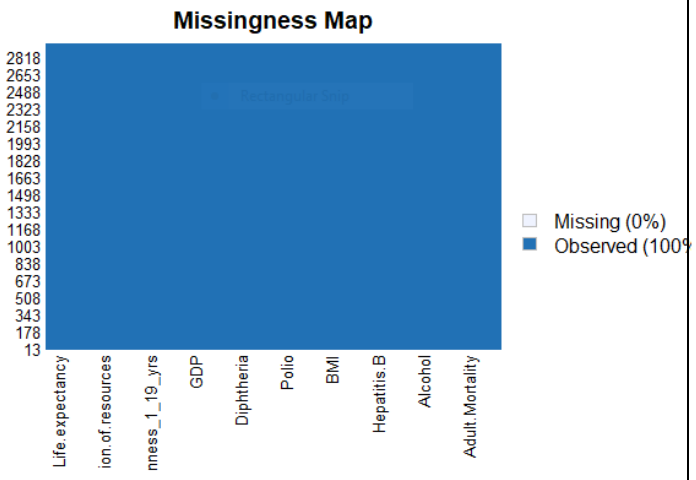

Figure 24

Figure 23 represents, there are 4% of missing values in the original dataset. Therefore we have to do missing values imputation before doing the

analysis. *Figure 24* represents, there are no missing values after doing missing values imputation.

2.3.2. Bayesian Multiple Linear regression

Bayesian version shows that under reference prior, we will obtain the posterior distribution of α and β_i analogous. Dependent variable is "Life Expectancy". Others are independent variables and you can view it from *Table 1*.

$$Y_i = \alpha + \beta_1 + \beta_2 + \ldots\ldots\ldots\ldots + \beta_{19} + \xi_i \qquad \textit{Equation 1}$$

From *Equation 1* we can implement the Bayesian multiple regression model. α represents the intercept of the model and β_i s represent the coefficients of each independent variables. We assume that ξ_i is independent and identically distributed with Normal distribution.

2.3.3. Posterior means and posterior standard deviations and Credible Intervals Summary

We can compare the estimates the coefficient values of OLS method and the spread of the distribution related to the standard errors by posterior mean and posterior standard deviations.

Credible interval is different from OLS confidence interval from the interpretation. Since we have obtained the distribution of each coefficient we can construct the credible interval. It provides the probability that a specific coefficient falls into this credible interval.

2.3.4. Model selection using BIC value

Bayesian Informative Criteria (BIC) is one of the most popular Bayesian criteria is used for Bayesian model selection. The BIC is defined as,

$$BIC = -2\ ln\ (likelihood) + (p+1)\ ln(n) \qquad \textit{Equation 2}$$

Here n is the number of observations in the model, and p is the number of predictors and (p+1) is number of total parameters and likelihood is estimated value.

$$Likelihood = p\ (data \mid \Theta, M) = L(\Theta, M) \qquad \textit{Equation 3}$$

The maximized value of likelihood, estimated value of likelihood, can be achieved by some special value of the parameter Ɵ, it is the estimated value of original Ɵ.

I used 'Backward Elimination with BIC' for model selection. At the first step, let's fit all the variables to the model. Then remove each variable from the full model and find which model has the lowest BIC value. Then we choose that model which has the smallest BIC value. In the next step, let's remove each variable from the selected model and find the BIC value of each model. Again, we choose the model which has the smallest BIC value after dropping the each variable. We continue this process till the BIC value is higher than the selected model of the previous step. That means we reach the best model.

2.3.5. Bayesian Model Uncertainty

Backward model selection with BIC gives the best model from the fitted models but it is not necessarily be the best fitting model for the given dataset. Therefore, we should perform model checking after model selection. I performed external validation. It divides the dataset in two parts. Then use one part to build a model and the other part to validate the model.

We can also visualize the model uncertainty using R 'image' function. It obtains a clearer view of model comparison. Y axis represents the predictors including the intercept. X axis represents each different model. Each vertical column represents one model. The variables that are not represented in the model represents black cells in each corresponding vertical column.

3. Results and Discussion

3.1. Discussion of Posterior Means, Posterior Standard Deviation and Credible Interval of the full model

	Post Mean	Post SD	Post p (B ! = 0)
Intercept	6.922e+01	7.472e-02	1
Status	-1.600e+00	2.693e-01	1
Adult Mortality	-1.987e-02	7.934e-04	1
Infant Deaths	9.990e-02	8.422e-03	1
Alcohol	6.389e-02	2.582e-02	1
Percentage Expenditure	8.930e-05	8.435e-05	1
Hepatitis B	-1.475e-02	3.913e-03	1
Measles	-1.938e-05	7.650e-06	1
BMI	4.439e-02	4.929e-03	1

Under Five Deaths	-7.480e-02	6.173e-03	1
Polio	2.860e-02	4.465e-03	1
Total Expenditure	6.400e-02	3.402e-03	1
Diphtheria	4.013e-02	4.704e-03	1
HIV AIDS	-4.698e-01	1.754e-02	1
GDP	3.283e-05	1.296e-05	1
Population	2.573e-10	1.692e-09	1
Thinness_1_19_yrs	-8.230e-02	5.037e-02	1
Thinness.5.9.years	7.445e-03	4.964e-02	1
Income composition of resources	5.727e+00	6.351e-01	1
Schooling	6.555e-01	4.173e-02	1

Table 2

Table 2 represents how the posterior mean and standard deviation vary with each variable in the full model. In the last column of Table 2, we can see the probability of coefficients are non-zero and always 1 because we gave an argument as 'include.always = ~ . ' . It forces the model to include all the variables. Under this 'centered' model and reference prior, the posterior mean of the intercept is now the sample mean of the response variable 'Life Expectancy'. Let's visualize the coefficient of the predictor variables as follows;

Figure 25 to *Figure 29* represent distributions all center the posterior distributions at their respective estimates with the spread of the distribution related to standard errors.

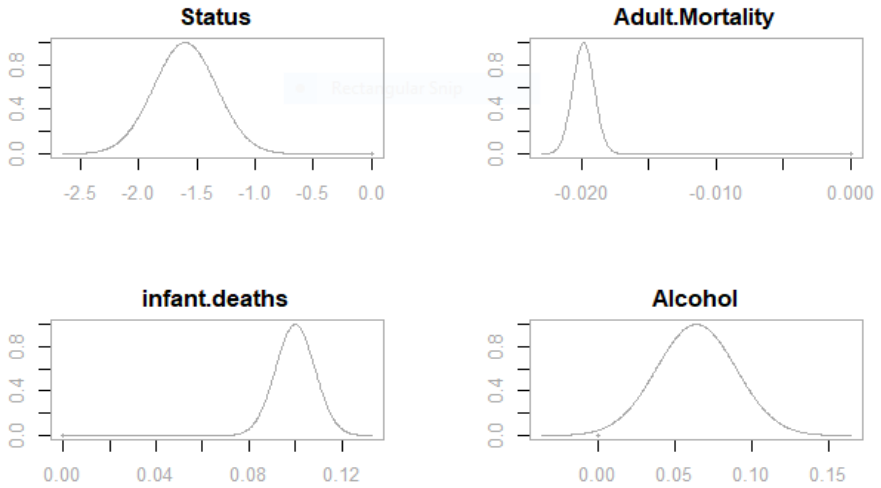

Figure 25

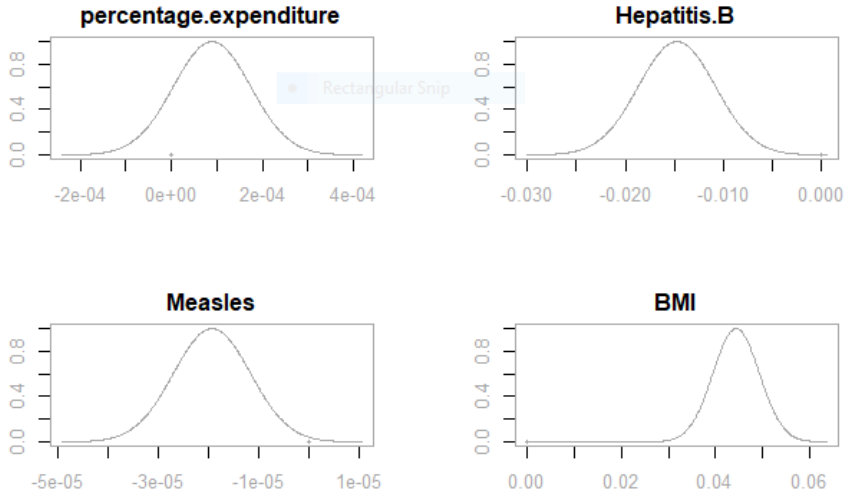

Figure 26

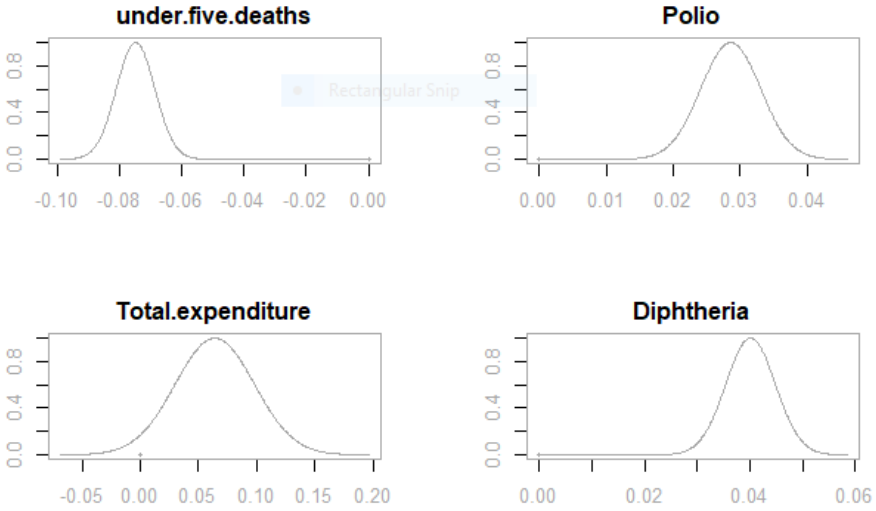

Figure 27

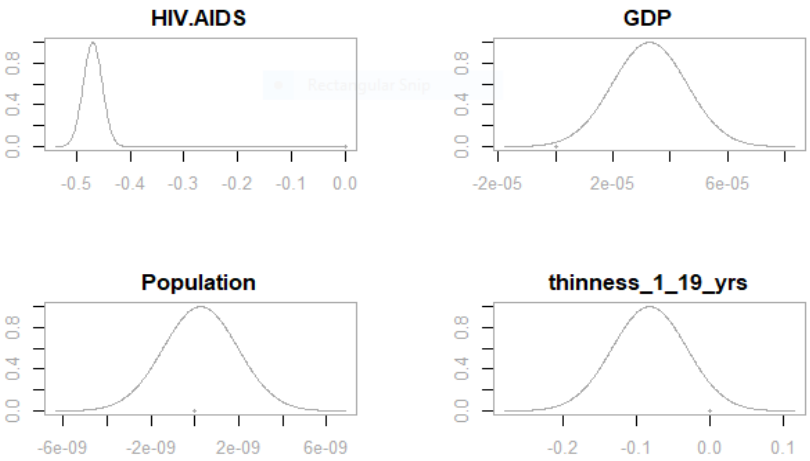

Figure 28

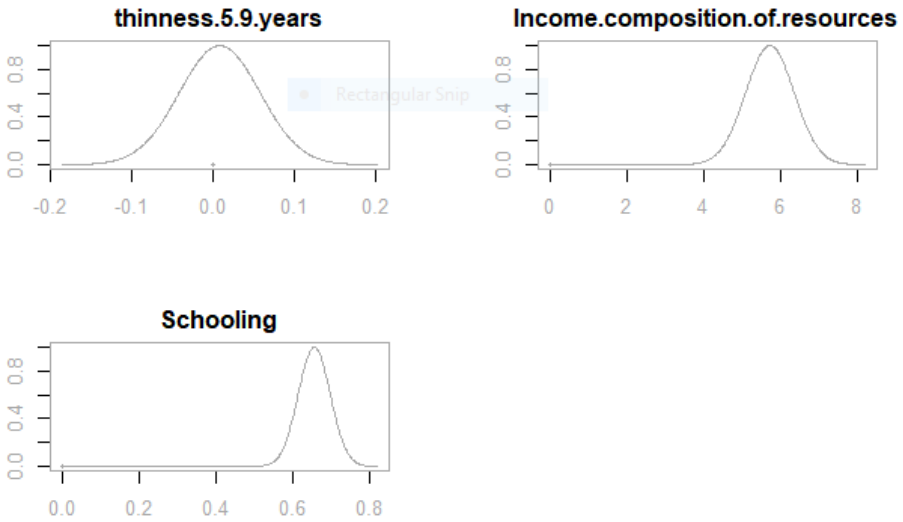

Figure 29

Table 3 gives more clearer and more useful summary from the 95% credible intervals of the coefficients of all the predictors. These intervals centered at the posterior mean with width given by appropriate t quantile with (n-p-1) degrees of freedom times posterior standard deviation. Let's interpret 'Infant death' variable for an example. There is a 95% chance that 'Life Expectancy' increases by 0.08 to 0.12 with one additional increase of the 'Infant death' variable. Some of the variables include 0, which represents that we should improve this model and they are less useful variables for the predictions.

	Post mean	Post Std	2.5%	97.5%
Intercept	69.22	0.07	69.08	69.37
Status	-1.6	0.27	-2.13	-1.07
Adult Mortality	-0.02	0.00	-0.02	-0.02
Infant Deaths	0.1	0.01	0.08	0.12
Alcohol	0.06	0.03	0.01	0.11
Percentage Expenditure	0.00	0.00	0.00	0.00
Hepatitis B	-0.01	0.00	-0.02	-0.01
Measles	0.00	0.00	0.00	0.00
BMI	0.04	0.00	0.03	0.05
Under five deaths	-0.07	0.01	-0.09	-0.06
Polio	0.03	0.00	0.02	0.04
Total expenditure	0.06	0.03	0.00	0.13
Diphtheria	0.04	0.00	0.03	0.05

HIV AIDS	-0.47	0.02	-0.50	-0.44
GDP	0.00	0.00	0.00	0.00
Population	0.00	0.00	0.00	0.00
Thinness_1_19_yrs	-0.08	0.05	-0.18	0.02
Thinness.5.9.years	0.01	0.05	-0.09	0.10
Income Composition of resources	5.73	0.64	4.48	6.97
Schooling	0.66	0.04	0.57	0.74

Table 3

3.2 Bayesian Model Selection using Backward BIC value

As we explained earlier we can get best predictor variables for creating a Bayesian multiple linear regression model. Our full model as follows;

Life.expectancy = Status + Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP + Population + thinness_1_19_yrs + thinness.5.9.years + Income.composition.of.resources + Schooling *Equation 4*

AIC should be represented as BIC in the RStudio output. Full model AIC value is 8358.24 . Then we should do the process described as in the methodology part till get the models which are do not less than previous best model AIC value. In RStudio this process run automatically and obtain the best model.

Life.expectancy = Status + Adult.Mortality + infant.deaths + Hepatitis.B + BMI + under.five.deaths + Polio + Diphtheria + HIV.AIDS + GDP + thinness_1_19_yrs + Income.composition.of.resources + Schooling *Equation 5*

We get the *Equation 5* as the best model and its AIC value is 8330.01. If we remove each variable one by one from this model, we can't get the model which has lower AIC value than *Equation 5* model. *Table 4* represents the final AIC results. From that we can conclude that there is no models which are suitable than *Equation 5* model.

Variable	AIC
Thinness_1_19_yrs	8337.2
Hepatitis B	8338.6
Polio	8363.9
GDP	8365.0
Status	8389.8
Diphtheria	8399.2
Income composition of resources	8399.9
BMI	8410.3
Infant deaths	8463.9

Under five deaths	8470.8
Schooling	8586.5
Adult Mortality	8883.5
HIV AIDS	8956.8

Table 4

3.3 Check the Model Uncertainty

As we explained in the methodology part we obtain the best model from the fitted models. But we have to check the *Equation 5* model uncertainty before finalized it as the best model. We used model validation technique to check the accuracy of the *Equation 5* model. We checked the accuracy of last three models which gave from backward elimination method. Equation 5 , Equation 6, Equation 7 represent the last three models. We choosed these models because they have the least AIC value.

Life.expectancy = Status + Adult.Mortality + infant.deaths + Alcohol + Hepatitis.B + BMI + under.five.deaths + Polio + Diphtheria + HIV.AIDS + GDP + thinness_1_19_yrs + Income.composition.of.resources + Schooling *Equation 6*

Life.expectancy = Status + Adult.Mortality + infant.deaths + Alcohol + Hepatitis.B + Measles + BMI + under.five.deaths + Polio + Diphtheria + HIV.AIDS + GDP + thinness_1_19_yrs +Income.composition.of.resources + Schooling *Equation 7*

I used Python programming language and jupyter notebook to obtain the accuracy of these models. I used cross validation technique to 10 boxes and got the average value that gave from 10 accuracies.

Model	Accuracy
Equation 5	0.7841
Equation 6	0.7790
Equation 7	0.7774

Table 5

Table 5 represents *Equation 5* which is the best model gave from the backward elimination method has the highest accuracy.

We can check the model uncertainty using visualization also.

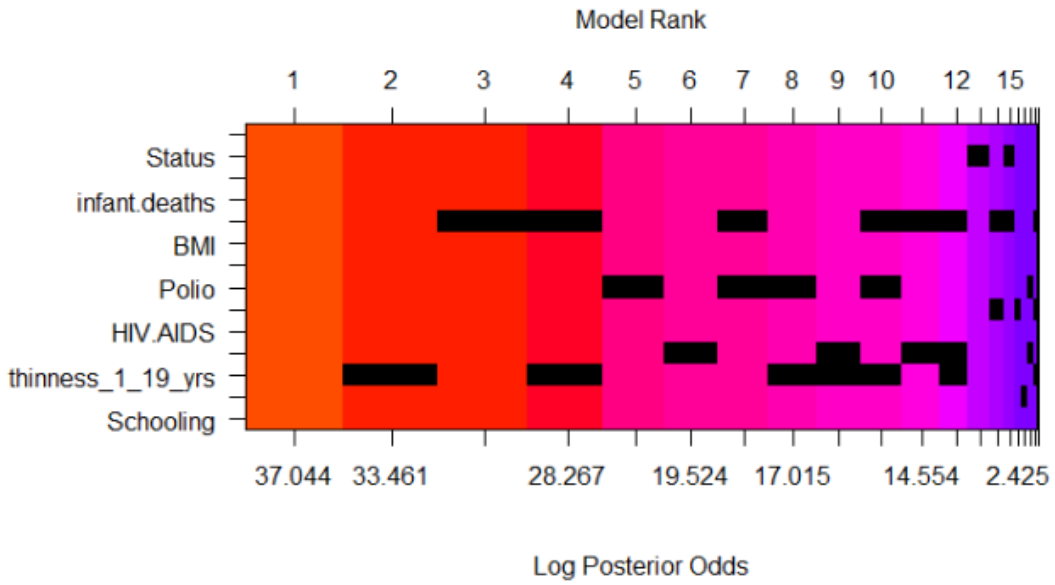

Figure 30

We know that black boxes represent variables which don't include in the models. Model 1 includes all the variables that contain the Equation 5 (best model was taken from backward elimination method). These models ordered according to the log of posterior odd over the null model.

From *Figure 31* concludes that variables which are in the *Equation 5* are highly useful to predict the Life expectancy because all the spikes are colored as blue and there aren't light blue color spikes.

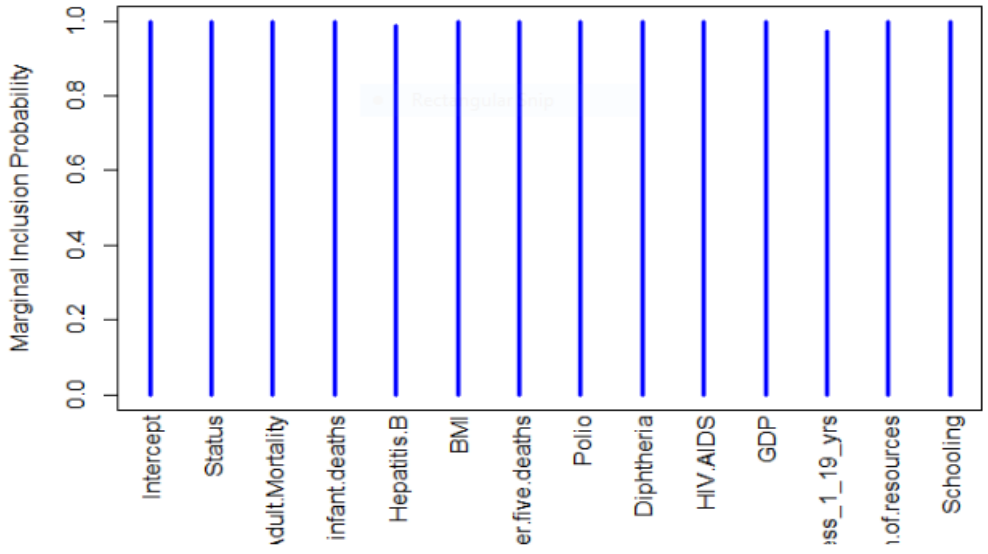

Figure 31

4. Conclusion

According to the results obtained in this study, it can be concluded as Bayesian Multiple Linear Regression give better accuracy to predict Life expectancy, backward elimination method with Bayesian Informative Criteria gives the best variables that help to build the regression model and posterior means, posterior standard deviations, model checking techniques give better idea about the generated Bayesian model. Finally we can conclude that, using Status, Adult mortality, infant deaths, Hepatitis B, BMI, under five deaths, Polio, Diphtheria, HIV AIDS, GDP, Thinness 10-19 years, Income composition of resources, Schooling variables we can predict Life expectancy with 78.34% accuracy from Bayesian Multiple Linear regression Model.

5. References

Jesse, D.S., Firas, A., Anthony, D., Matthew, G., Alexander, H., Nazareno, S., Andrew, V., Quoc-Dien, T. (2015). *Predicting Life Expectancy in Men Diagnosed with Prostate Cancer. European Urology.* https://doi.org/10.1016/j.eururo.2015.03.020

Aurora, P., Wade, A., Whitmore, P., Whitehead, B. (2000). *A model for predicting life expectancy of children with cystic fibrosis. European Respiratory Journal.* https://erj.ersjournals.com/content/16/6/1056

Akshaya, R., Krutika, M., Shreya, A., Suresh, S. (2021*). Post Thoracic Surgery Life Expectancy Prediction Using Machine Learning. International Journal of Healthcare Information Systems and Informatics.* 10.4018/IJHISI.20211001.oa32

Merlise, C., Mine, C., Colin, R., David, B., Christine, C., Lizzy, H. & Lizzy, H. (2021). An *Introduction to Bayesian Thinking.* (Rev. ed.). Coursera. https://statswithr.github.io/book/

Alicia, A.J., Miles, Q.O. & Mine, D. (2021). *Bayes Rules! An Introduction to Applied Bayesian Modeling.* (Rev. ed.). Github. https://www.bayesrulesbook.com/index.html

6. Appendices

6.1 Dataset

The dataset has 2938 records and 20 variables and you can view it from This Link. I attached part of the dataset in *Figure 32*.

Status	Adult Mor	infant dea	Alcohol	percentag	Hepatitis	Measles	BMI	under-five	Polio	Total expe	Diphtheria	HIV/AIDS	GDP	Populatio	thinness_	thinness	Income cc	Schooling	Life expectancy
Developir	263	62	0.01	71.27962	65	1154	19.1	83	6	8.16	65	0.1	584.2592	33736494	17.2	17.3	0.479	10.1	65
Developir	271	64	0.01	73.52358	62	492	18.6	86	58	8.18	62	0.1	612.6965	327582	17.5	17.5	0.476	10	59.9
Developir	268	66	0.01	73.21924	64	430	18.1	89	62	8.13	64	0.1	631.745	31731688	17.7	17.7	0.47	9.9	59.9
Developir	272	69	0.01	78.18422	67	2787	17.6	93	67	8.52	67	0.1	669.959	3696958	17.9	18	0.463	9.8	59.5
Developir	275	71	0.01	7.097109	68	3013	17.2	97	68	7.87	68	0.1	63.53723	2978599	18.2	18.2	0.454	9.5	59.2
Developir	279	74	0.01	79.67937	66	1989	16.7	102	66	9.2	66	0.1	553.3289	2883167	18.4	18.4	0.448	9.2	58.8
Developir	281	77	0.01	56.76222	63	2861	16.2	106	63	9.42	63	0.1	445.8933	284331	18.6	18.7	0.434	8.9	58.6
Developir	287	80	0.03	25.87393	64	1599	15.7	110	64	8.33	64	0.1	373.3611	2729431	18.8	18.9	0.433	8.7	58.1
Developir	295	82	0.02	10.91016	63	1141	15.2	113	63	6.73	63	0.1	369.8358	26616792	19	19.1	0.415	8.4	57.5
Developir	295	84	0.03	17.17152	64	1990	14.7	116	58	7.43	58	0.1	272.5638	2589345	19.2	19.3	0.405	8.1	57.3
Developir	291	85	0.02	1.388648	66	1296	14.2	118	58	8.7	58	0.1	25.29413	257798	19.3	19.5	0.396	7.9	57.3
Developir	293	87	0.02	15.29607	67	466	13.8	120	5	8.79	5	0.1	219.1414	24118979	19.5	19.7	0.381	6.8	57
Developir	295	87	0.01	11.08905	65	798	13.4	122	41	8.82	41	0.1	198.7285	2364851	19.7	19.9	0.373	6.5	56.7
Developir	3	88	0.01	16.88735	64	2486	13	122	36	7.76	36	0.1	187.846	21979923	19.9	2.2	0.341	6.2	56.2
Developir	316	88	0.01	10.57473	63	8762	12.6	122	35	7.8	33	0.1	117.497	2966463	2.1	2.4	0.34	5.9	55.3
Developir	321	88	0.01	10.42496	62	6532	12.2	122	24	8.2	24	0.1	114.56	293756	2.3	2.5	0.338	5.5	54.8
Developir	74	0	4.6	364.9752	99	0	58	0	99	6	99	0.1	3954.228	28873	1.2	1.3	0.762	14.2	77.8
Developir	8	0	4.51	428.7491	98	0	57.2	1	98	5.88	98	0.1	4575.764	288914	1.2	1.3	0.761	14.2	77.5
Developir	84	0	4.76	430.877	99	0	56.5	1	99	5.66	99	0.1	4414.723	289592	1.3	1.4	0.759	14.2	77.2

Figure 32

6.2 R Codes

- library(tidyverse)
- library(skimr)
- library(psych)
- library(Amelia)
- library(mice)
- library(janitor)
- library(Hmisc)
- library(reshape2)
- library(reshape)
- library(stats)
- df <- read.csv('LifeExp.csv')
  head(df)
- clean_names(df)
- glimpse(df)
- describe(df)
- df$Status <- ifelse(df$Status == "Developing",1,0)
- str(df)
- skim(df)
- missmap(df)
- colnames(df)
- mean_val <- colMeans(df,na.rm = TRUE)
  for(i in colnames(df))
  df[,i][is.na(df[,i])] <- mean_val[i]

- skim(df)
- missmap(df)
- df %>%
  ggplot(aes(x=Life.expectancy)) +
  geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+

  ggtitle("Data distribution of Life Expectancy")

- pie(table(df$Status), col=grey.colors(3), main="Developing or Developed")
- scormatrix = rcorr(as.matrix(df), type='spearman')
  scormatrix
- library(GGally)
  ggcorr(df, method = c("everything", "pearson"))
- n = nrow(df)
  df.lm = lm(Life.expectancy ~ ., data=df)
  df.step = step(df.lm, k=log(n))
- df.coef = coef(df.BIC)
  df.coef
- par(mfrow = c(2, 2), col.lab = "darkgrey", col.axis = "darkgrey", col = "darkgrey")
  plot(df.coef, subset = 2:20, ask = F)
- confint(df.coef, parm = 2:20)
- out = confint(df.coef)[, 1:2]
  names = c("posterior mean", "posterior std", colnames(out))
  out = cbind(df.coef$postmean, df.coef$postsd, out)
  colnames(out) = names
  round(out, 2)
- df.bestBIC = bas.lm(Life.expectancy ~ ., data = df,
  prior = "BIC", n.models = 1, # We only fit 1 model
  bestmodel = bestgamma, # We use bestgamma to indicate variables
  modelprior = uniform())

  df.coef = coef(df.bestBIC)

  out = confint(df.coef)[, 1:2]

  coef.BIC = cbind(df.coef$postmean, df.coef$postsd, out)
  names = c("post mean", "post sd", colnames(out))
  colnames(coef.BIC) = names
  coef.BIC
- model1 = bas.lm(Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
  Hepatitis.B + BMI + under.five.deaths + Polio + Diphtheria +
  HIV.AIDS + GDP + thinness_1_19_yrs + Income.composition.of.resources +

```
    Schooling, data = df,
            prior = "BIC", modelprior = uniform())
plot(model1, which = 4, ask = F, caption = "", sub.caption = "",
     col.in = "blue", col.ex = "darkgrey", lwd = 3)
```

- image(model1, rotate = F)
  model1.coef = coef(model1)
  par(mfrow = c(2, 2), col.lab = "darkgrey", col.axis = "darkgrey", col = "darkgrey")
  plot(model1.coef, subset = 2:14, ask = F)

6.3 Python Codes

- import pandas as pd
- df = pd.read_csv('LifeExp.csv')
  df.head()
- df_one = pd.get_dummies(df["Status"])
- df_two = pd.concat((df_one, df), axis=1)
- df_two = df_two.drop(["Status"], axis=1)
- df_two = df_two.drop(["Developing"], axis=1)
- df_new = df_two.rename(columns={"Developed": "Status"})
- df_new.isnull().sum()
- df_new['Adult Mortality'].fillna((df_new['Adult Mortality'].mean()), inplace=True)
- df_new['Alcohol'].fillna((df_new['Alcohol'].mean()), inplace=True)
- df_new['Hepatitis B'].fillna((df_new['Hepatitis B'].mean()), inplace=True)
- df_new[' BMI '].fillna((df_new[' BMI '].mean()), inplace=True)
- df_new['Polio'].fillna((df_new['Polio'].mean()), inplace=True)
- df_new['Total expenditure'].fillna((df_new['Total expenditure'].mean()), inplace=True)
- df_new['Diphtheria '].fillna((df_new['Diphtheria '].mean()), inplace=True)
- df_new['GDP'].fillna((df_new['GDP'].mean()), inplace=True)
- df_new['Population'].fillna((df_new['Population'].mean()), inplace=True)
- df_new[' thinness_1_19_yrs'].fillna((df_new[' thinness_1_19_yrs'].mean()), inplace=True)
- df_new[' thinness 5-9 years'].fillna((df_new[' thinness 5-9 years'].mean()), inplace=True)
- df_new['Income composition of resources'].fillna((df_new['Income composition of resources'].mean()), inplace=True)
- df_new['Schooling'].fillna((df_new['Schooling'].mean()), inplace=True)

- df_new['LifeExpectancy '].fillna((df_new['LifeExpectancy '].mean()), inplace=True)
- X_full = df_new.drop(['LifeExpectancy '],axis=1)
  Y = df_new['LifeExpectancy ']
- from sklearn import linear_model
- from sklearn.model_selection import train_test_split
- from sklearn.model_selection import cross_val_score
- from sklearn.model_selection import StratifiedKFold
- scores_full=cross_val_score(linear_model.BayesianRidge(), X_full, Y,cv=10)
- import numpy as np
- np.average(scores_full)
- model1 = df_new[['Status' ,'Adult Mortality' , 'infant deaths' , 'Alcohol' , 'Hepatitis B' , ' BMI ' , 'under-five deaths ' , 'Polio' , 'Diphtheria ' , ' HIV/AIDS' , 'GDP' , ' thinness_1_19_yrs' ,'Income composition of resources' , 'Schooling']]
- model2 = df_new[['Status' ,'Adult Mortality' , 'infant deaths' , 'Alcohol' , 'Hepatitis B' , ' BMI ' , 'under-five deaths ' , 'Polio' , 'Diphtheria ' , ' HIV/AIDS' , 'GDP' , ' thinness_1_19_yrs' ,'Income composition of resources' , 'Schooling','Measles ']]
- model3 = df_new[['Status' ,'Adult Mortality' , 'infant deaths' , 'Alcohol' , 'Hepatitis B' , ' BMI ' , 'under-five deaths ' , 'Polio' , 'Diphtheria ' , ' HIV/AIDS' , 'GDP' , ' thinness_1_19_yrs' ,'Income composition of resources' , 'Schooling','Measles ','Total expenditure']]
- scores1=cross_val_score(linear_model.BayesianRidge(), model1, Y,cv=10)
- np.average(scores1)
- scores2=cross_val_score(linear_model.BayesianRidge(), model2, Y,cv=10)
- np.average(scores2)
- scores3=cross_val_score(linear_model.BayesianRidge(), model3, Y,cv=10)
- np.average(scores3)