Life Expectancy Analysis

Group member:

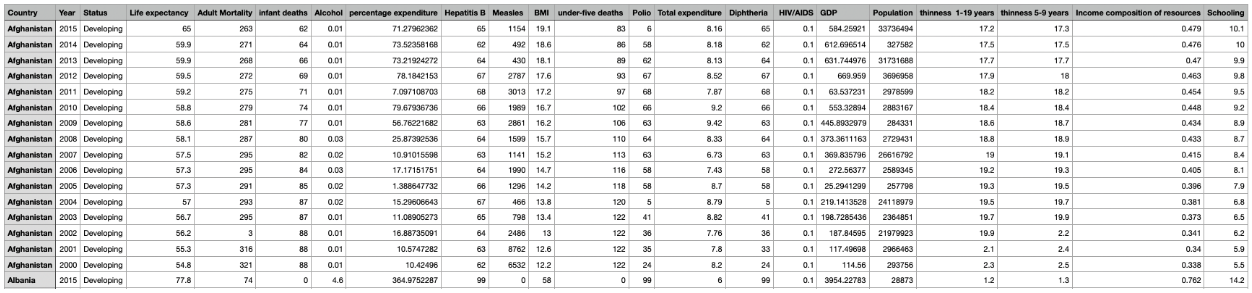
Jinghan Xu(hanxu@ucdavis.edu)
Fengsheng Zhou (fszhou@ucdavis.edu)
Xueyin Zhu (xyzzhu@ucdavis.edu)
Zuer Zheng (zezheng@ucdavis.edu)

Introduction

In the past 15 years, there has been a vast improvement in human mortality rates, especially in the developing nations. However, according to MEDRXIV, life expectancy in many countries like the United States is shorter than before due to the COVID-19 pandemic. Reflecting the number of COVID-19 deaths nationwide, the topic of life expectancy is brought up again. We are thus interested in performing a data analysis on life expectancy to see how life expectancy will be affected by health and social factors. The dataset we are using is collected by WHO and the United Nation website, which reports life expectancy of 193 countries from 2000 to 2015. The data also contains other related factors of life expectancy including country's development status, adult mortality, infant death, alcohol, percentage expenditure, Hepatitis B, Measle, BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources and Schooling. We have a total of 2938 observations with 22 variables, and a preview of data is shown in Fig. 1.

The main purpose of this project is to construct a multiple linear regression model for life expectancy and to check the model's performance through regression diagnostics and simulation study. To achieve this, we will use the model selection criteria of stepwise regression and Akaike Information Criterion, which is denoted as AIC. The estimation procedure for β, or the coefficients of the model, will be least squares with formula $\hat{\beta} = (X^T X)^{-1} X^T y$, where X are our explanatory variables and y is our response variable. The least squares methods we will use to solve for β includes LU decomposition, Cholesky decomposition, and QR decomposition, and we will calculate each method's computation speed to find out the most efficient one in our case.

Besides model construction, we will also perform real analysis for life expectancy. To be specific, we will analyze countries with both high values and low values of life expectancy, as well as countries which are experiencing improvements on life expectancy to find out the real association between life expectancy and explanatory factors.



Country	Year	Status	Life expectancy	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 years	thinness 5-9 years	Income composition of resources	Schooling
Afghanistan	2015	Developing	65	263	62	0.01	71.27962362	65	1154	19.1	83	6	8.16	65	0.1	584.25921	33736494	17.2	17.3	0.479	10.1
Afghanistan	2014	Developing	59.9	271	64	0.01	73.52358168	62	492	18.6	86	58	8.18	62	0.1	612.696514	327582	17.5	17.5	0.476	10
Afghanistan	2013	Developing	59.9	268	66	0.01	73.21924272	64	430	18.1	89	62	8.13	64	0.1	631.744976	31731688	17.7	17.7	0.47	9.9
Afghanistan	2012	Developing	59.5	272	69	0.01	78.1842153	67	2787	17.6	93	67	8.52	67	0.1	669.959	3696958	17.9	18	0.463	9.8
Afghanistan	2011	Developing	59.2	275	71	0.01	7.097108703	68	3013	17.2	97	68	7.87	68	0.1	63.537231	2978599	18.2	18.2	0.454	9.5
Afghanistan	2010	Developing	58.8	279	74	0.01	79.67936736	66	1989	16.7	102	66	9.2	66	0.1	553.32894	2883167	18.4	18.4	0.448	9.2
Afghanistan	2009	Developing	58.6	281	77	0.01	56.76221682	63	2861	16.2	106	63	9.42	63	0.1	445.8932979	284331	18.6	18.7	0.434	8.9
Afghanistan	2008	Developing	58.1	287	80	0.03	25.87392536	64	1599	15.7	110	64	8.33	64	0.1	373.3611163	2729431	18.8	18.9	0.433	8.7
Afghanistan	2007	Developing	57.5	295	82	0.02	10.91015598	63	1141	15.2	113	63	6.73	63	0.1	369.835796	26616792	19	19.1	0.415	8.4
Afghanistan	2006	Developing	57.3	295	84	0.03	17.17151751	64	1990	14.7	116	58	7.43	58	0.1	272.56377	2589345	19.2	19.3	0.405	8.1
Afghanistan	2005	Developing	57.3	291	85	0.02	1.388647732	66	1296	14.2	118	58	8.7	58	0.1	25.2941299	257798	19.3	19.5	0.396	7.9
Afghanistan	2004	Developing	57	293	87	0.02	15.29606643	67	466	13.8	120	5	8.79	5	0.1	219.1413528	24118979	19.5	19.7	0.381	6.8
Afghanistan	2003	Developing	56.7	295	87	0.01	11.08905273	65	798	13.4	122	41	8.82	41	0.1	198.7285436	2364851	19.7	19.9	0.373	6.5
Afghanistan	2002	Developing	56.2	3	88	0.01	16.88735091	64	2486	13	122	36	7.76	36	0.1	187.84595	21979923	19.9	2.2	0.341	6.2
Afghanistan	2001	Developing	55.3	316	88	0.01	10.5747282	63	8762	12.6	122	35	7.8	33	0.1	117.49698	2966463	2.1	2.4	0.34	5.9
Afghanistan	2000	Developing	54.8	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
Albania	2015	Developing	77.8	74	0	4.6	364.9752287	99	0	58	0	99	6	99	0.1	3954.22783	28873	1.2	1.3	0.762	14.2

Figure. 1 Preview of data

Proposed Method

The main statistical methods we are going to use in this project are linear regression model, AIC model selection, least squares, and simulation study. We will introduce these methods in the following sections.

A. *Linear Regression Model*

Linear regression is a modeling approach that models the linear relationship between a response variable and one or more predictor variables. In our case, we are using multiple linear regression because we have more than one explanatory variable.

Suppose we have a $n \times p$ dimension dataset, then the model we will use for linear regression is

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i$$

where vectors x_i are our explanatory variables including the intercept and T means transpose. $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$ are our coefficients for the regression model, and $\beta_0$is the intercept term. The true values for β is unknown, and our goal is to estimate β using the least squares approach, which will be introduced in part C.

B. *AIC Model Selection*

One important step before we perform the estimation for β is to choose an appropriate number of regressors for the model. The model selection process is necessary to prevent the problem of overfitting, especially if the dataset has a large dimension. Our dataset contains 21 explanatory variables, and we don't want to put them all in the model because the dimension is big and there might be variables that are not significant to life expectancy. One popular criterion to construct model selection is Akaike information criterion (AIC). Given a model with sample size n and p regressors, we have

$$AIC = nlog(\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}) + 2(p + 2)$$

The term $(\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n})$ is equal to MSE, or the mean squared error, and $2(p + 2)$ is a penalty term on the number of regressors in the model. Under AIC procedure, the model with the minimum AIC score will be the best. From the formula we can see that AIC depends on MSE: the larger the MSE is, the larger the AIC is. The model with smaller AIC is better because we want to minimize the difference between true value and estimated value of y. However, AIC also contains a penalty term depending on p. When p, or the number of regressors in the model increase, AIC will increase as well. Therefore, AIC is a procedure to keep balance of the model performance and number of regressors, which is suitable for the purpose of reducing explanatory variables.

One similar model selection approach is called Bayesian information criterion (BIC), where

$$BIC = nlog(\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}) + log(n)(p + 2)$$

The main difference between AIC and BIC is the penalty term. In AIC, the penalty term only depends on the number of regressors, but BIC's penalty term also depends on n. We will not use BIC for the project because we only care about the number of regressors to put in.

C. *Least Squares Methodology*

After selecting the appropriate model, we can continue to estimate the model coefficients. The main idea for least squares is to estimate β by minimizing RSS, or the residual sum of squares, which is equal to $(Y - X\beta)^T(Y - X\beta)$. Under this approach, we solve for β in the form of $Ax = b$, or

$X'Xb = X'y$, where $A = X'X,\ b = X'y$, and $\widehat{\beta} = (X'X)^{-1}X'y$. In the project, we will use and compare three different least squares methods: LU decomposition, Cholesky decomposition, and QR decomposition.

1. LU decomposition (Lower-upper decomposition): Suppose A is a matrix. A is defined to be factored into three parts. A = PLU, where L refers to lower triangular matrix, U refers to upper triangular matrix, and P refers to permutation matrix, which is operated through Gaussian elimination.
2. QR decomposition: Suppose A is a matrix. A is defined to be factored into two parts. A = QR, where Q is an orthogonal matrix and R is an upper triangular matrix, which are found by using the Gram-Schmidt process Householder reflections.
3. Cholesky decomposition: Suppose the matrix, A is a Hermitian, positive-definite. A is defined to be factored into two parts. A = LL’, where L is a lower triangular matrix and L’ is its conjugate transpose.

D. Simulation Study

To check whether our statistical method performs well or not, we will use simulation study. Basically, we generate multiple pseudo-random data with true known parameters and relationship between x and y, then apply our estimation method to the simulated data as well as checking the distribution of estimated parameters.

In our case, we will evaluate our estimation method for β and σ, where we calculate

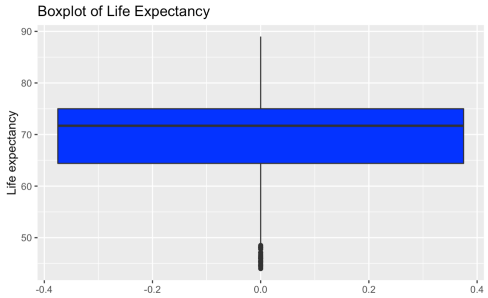
$$\widehat{\beta} = (X'X)^{-1}X'y\ ,\ \widehat{\sigma}^2 = \frac{(y-X\beta)^T(y-X\beta)}{n}$$

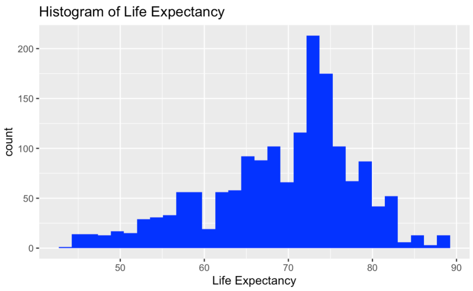
We choose to generate a set of pseudo-random data with the same dimension as our life expectancy dataset and set its estimated parameters to be the true ones for the simulated data. Then we will apply the estimation methods above to get $\widehat{\beta}$ and $\widehat{\sigma}$ and compare them with true β and σ. To do the comparison, we will plot the distribution and mean values for $\widehat{\beta}$ and $\widehat{\sigma}$ to see whether our estimation is biased or not. An unbiased estimation indicates that our estimation method is working efficiently.

Data Analysis Study

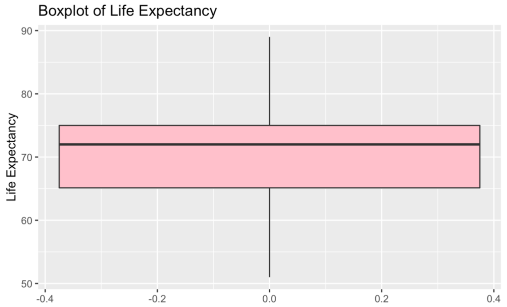
A. Data processing

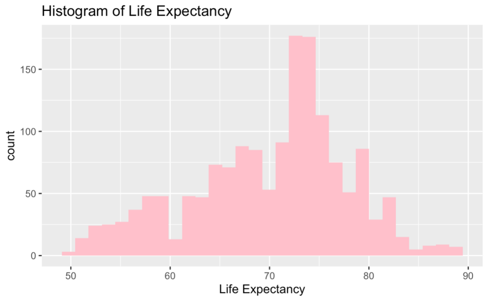
Before analyzing the dataset, we need to process it in advance. Firstly, we find some NA values, which might lead to some serious issues like biased computation and misleading judgment. Therefore, we decide to remove the observations with NA values so that we can ensure the accuracy of the results. After removing the NA values, we have 1649 observations remaining.



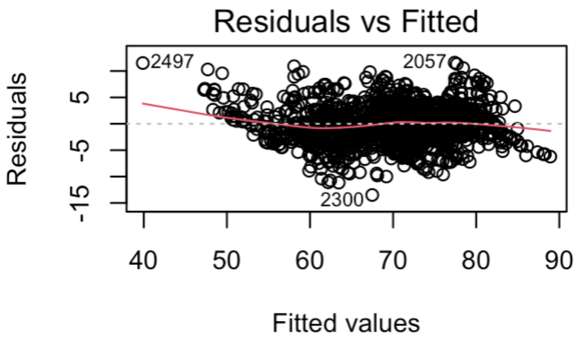


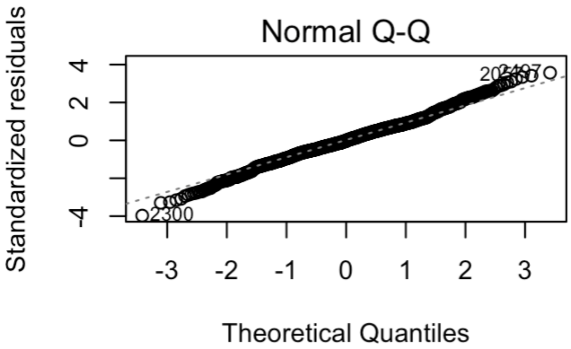
Secondly, we build the boxplot and histogram for life expectancy. From the boxplot(left figure), we can see that the plot size was unevenly distributed and the median value was higher than the mean value of life expectancy. Also, there were some outliers. Meanwhile, the histogram is left-skewed since the outliers exist.





In order to remove outliers, we let the life expectancy be larger than 50. The boxplot of life expectancy showed above. In the new boxplot, we can observe life expectancy ranging from 50 to 90 years, with most counties reaching 65 to 75 years. From the histogram of life expectancy, we can see it is much better than the previous histogram, which seems more bell-shaped, so we can conclude that it does not violate the assumption of normality.

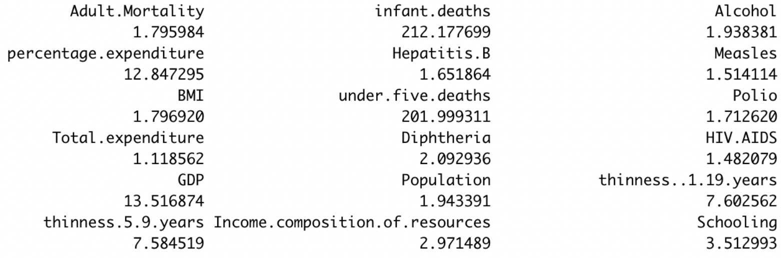




Furthermore, we access the residual plot to test the assumption of linearity and constant variance. From the residual plot from the new regression model, we can see the points nicely fill out this whole space except some outliers stand, and the points distribute pretty evenly around the h=0; so, there is no violation of linearity and constant variance; meanwhile, most of the points are on the QQ-line, and two tails are close to the QQ-line as well, which also means that our data does not violate the normality.

B. Model Selection

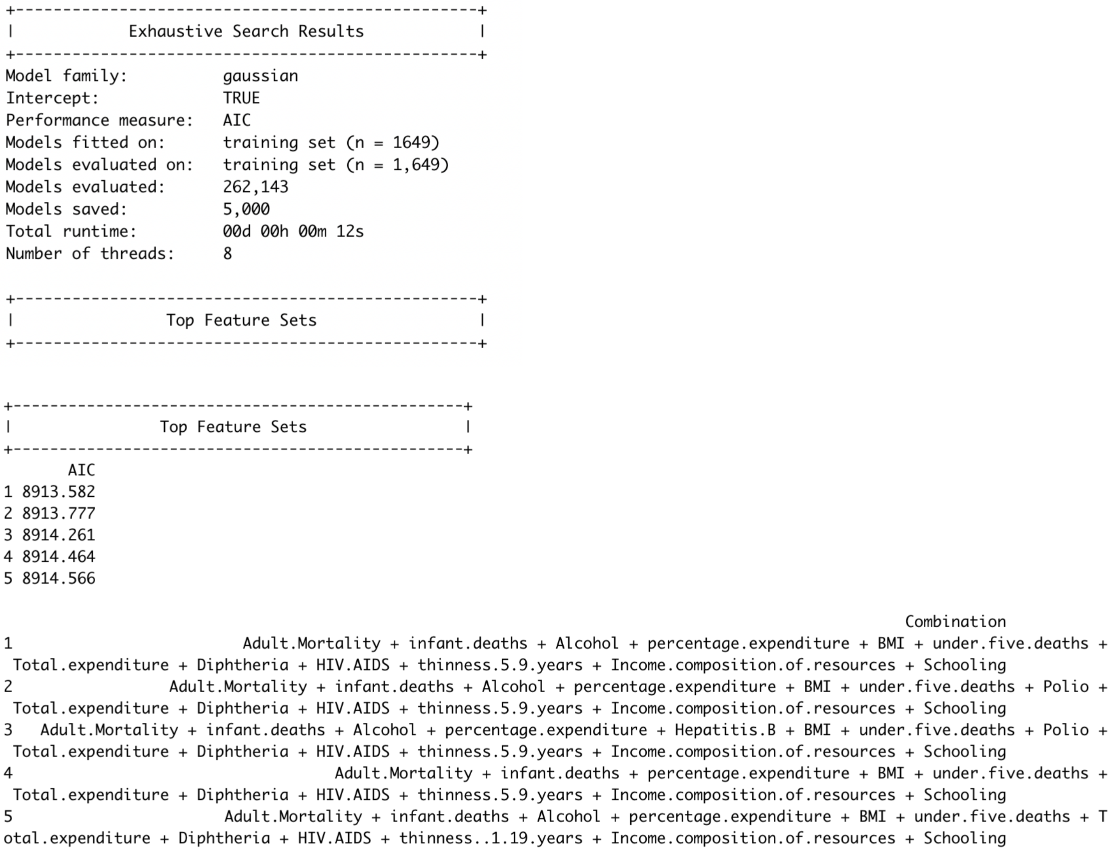
We found that there is existing collinearity between predictors. The variance-inflation factor of infant.deaths(VIF = 212.18), under.five.deaths(VIF = 201.999), GDP(VIF = 13.5), percentage.expenditure(VIF = 12.85) were quite big. It means that those predictors had high correlation efficiency because of the collinearity, which will cause the parameter estimates to become very large, therefore, we built the model by step regression method to do the model selection.



```
       Adult.Mortality                   infant.deaths                         Alcohol 
              1.795984                      212.177699                        1.938381 
percentage.expenditure                     Hepatitis.B                         Measles 
             12.847295                        1.651864                        1.514114 
                   BMI               under.five.deaths                           Polio 
              1.796920                      201.999311                        1.712620 
     Total.expenditure                      Diphtheria                        HIV.AIDS 
              1.118562                        2.092936                        1.482079 
                   GDP                      Population              thinness..1.19.years 
             13.516874                        1.943391                        7.602562 
    thinness.5.9.years Income.composition.of.resources                       Schooling 
              7.584519                        2.971489                        3.512993 
```

In order to reduce the effect of collinearity, we performed model selection by exhaustive search and forward stepwise selection, and selected the best model with an AIC score. The exhaustive search and forward selection gives us the result shown below.

Exhaustive search



```
+-------------------------------------------------+
|           Exhaustive Search Results             |
+-------------------------------------------------+
Model family:          gaussian
Intercept:             TRUE
Performance measure:   AIC
Models fitted on:      training set (n = 1649)
Models evaluated on:   training set (n = 1,649)
Models evaluated:      262,143
Models saved:          5,000
Total runtime:         00d 00h 00m 12s
Number of threads:     8

+-------------------------------------------------+
|               Top Feature Sets                  |
+-------------------------------------------------+


+-------------------------------------------------+
|               Top Feature Sets                  |
+-------------------------------------------------+
       AIC
1 8913.582
2 8913.777
3 8914.261
4 8914.464
5 8914.566

                                                                                                Combination
1                        Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + BMI + under.five.deaths + 
 Total.expenditure + Diphtheria + HIV.AIDS + thinness.5.9.years + Income.composition.of.resources + Schooling
2                Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + BMI + under.five.deaths + Polio + 
 Total.expenditure + Diphtheria + HIV.AIDS + thinness.5.9.years + Income.composition.of.resources + Schooling
3   Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Hepatitis.B + BMI + under.five.deaths + Polio + 
 Total.expenditure + Diphtheria + HIV.AIDS + thinness.5.9.years + Income.composition.of.resources + Schooling
4                                  Adult.Mortality + infant.deaths + percentage.expenditure + BMI + under.five.deaths + 
 Total.expenditure + Diphtheria + HIV.AIDS + thinness.5.9.years + Income.composition.of.resources + Schooling
5                        Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + BMI + under.five.deaths + T
otal.expenditure + Diphtheria + HIV.AIDS + thinness..1.19.years + Income.composition.of.resources + Schooling
```

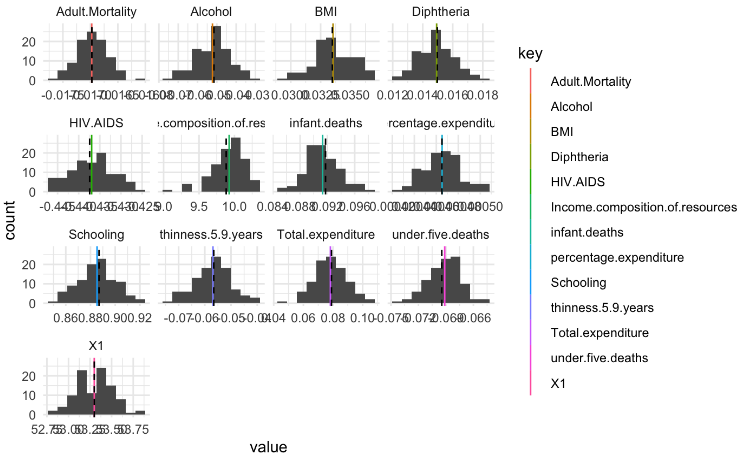
Forward stepwise selection

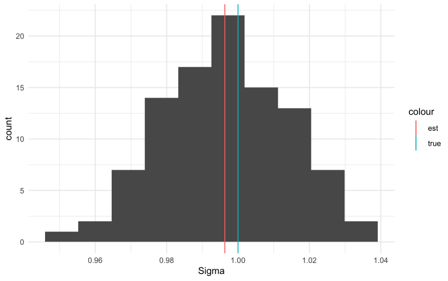
```
##                          Step Df    Deviance Resid. Df   Resid. Dev        AIC
## 1                            NA         NA      1648      127529.31     7172.144
## 2                  + Schooling -1 67519.81536       1647    60009.50     5931.056
## 3                   + HIV.AIDS -1 25626.36300      1646    34383.13    5014.668
## 4            + Adult.Mortality -1  7231.55190      1645    27151.58     4627.288
## 5  + Income.composition.of.resources -1  2815.85465     1644   24335.73     4448.739
## 6        + percentage.expenditure -1   706.09180      1643   23629.63      4402.186
## 7                        + BMI -1    681.26148      1642   22948.37        4355.946
## 8                  + Diphtheria -1   364.78073      1641   22583.59      4331.523
## 9             + under.five.deaths -1   174.74655      1640   22408.85     4320.714
## 10                + infant.deaths -1  1140.69752      1639   21268.15    4236.561
## 11            + thinness.5.9.years -1    54.31031      1638   21213.84      4234.345
## 12             + Total.expenditure -1    45.50034      1637   21168.34     4232.805
## 13                     + Alcohol -1    36.95801      1636   21131.38     4231.923
```

In the exhaustive search model, there always had schooling, Income.composition.of.resources, HIV.AIDS, Adult.Mortality variables, it means these variables were related to life expectancy. the forward selection confirmed this statement. we can see the AIC value from the top to the bottom, schooling had the highest value which means it was related with life expectancy. Meanwhile, the alcohol had the smallest value which means it may not have too much impact on life expectancy.

C. Simulation Study and Model Check

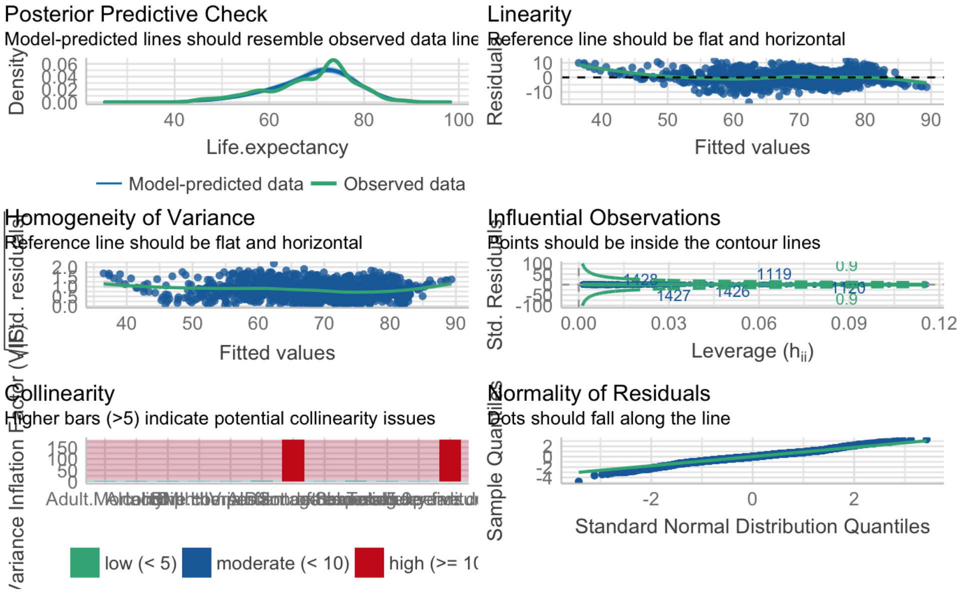
Now we have can use the parameters we chose to create a new linear model $y = X\beta + \varepsilon$. In order to evaluate the model, see how good the estimating is, we can do a simulation to perform compare the estimated $\widehat{\beta}$, $\widehat{\sigma}$ from the true β, σ.





From the left simulation figure above, we can see that the estimate $\hat{\beta}$ is very close to the true β. And from the right figure, we notice difference between the biased estimate for σ and the true σ is very small. This indicates that our estimation method is in good standing.

We can further verify the performance of our model for the life expectancy dataset. We could check our model's performance in terms of linearity of the relationship between X and y, homogeneity of variance, outliers, multicollinearity, and normality of residuals, which are all shown in the following figures. The expected result of each condition is mentioned in its corresponding subplots, and our model overall has a quite good performance because it almost meets all standards. The only potential problem is that we have two variables with relatively high collinearity.



D. Beta calculation — Least-squares method

Based on the previous step, now we know that the model can be explained by some significant features, which are schooling, HIV/AIDS, adult mortality, income composition resources, percentage expenditure, BMI, diphtheria, under-five deaths, infant deaths, thinness 5-9 years, total expenditure and alcohol. Then we are interested in how the speed of operation differs between the different least-squares algorithms (LU, Cholesky, QR) in solving linear equation $y = X\beta$. Since we may get no specific answer from solving the linear regression model, we are interested in solving the equation $(X'X)\beta = Xy$, to get the value of $\hat{\beta}$.

a. LU user system elapsed 0.166 0.044 0.224

LU decomposition gives us a permutation matrix P, a lower matrix L, and an upper matrix U (A = PLU). The cost for solving a linear equation by using LU decomposition approximate to $\frac{2}{3}n^3$ $(O(n^3))$. In this case, the executive speed is 0.224.

b. Cholesky user system elapsed 0.034 0.018 0.055

Using the Cholesky decomposition, the real matrix A can be written as $A = LL^T$. We can use the chol2inv function to solve the equation to avoid computing the inverse matrix, which will save the execution time as well. The cost for solving a linear equation by using cholesky decomposition approximate $\frac{n^3}{3}$ $(O(n^3))$. The executive speed in this case is 0.055

c. QR user system elapsed 0.038 0.024 0.064

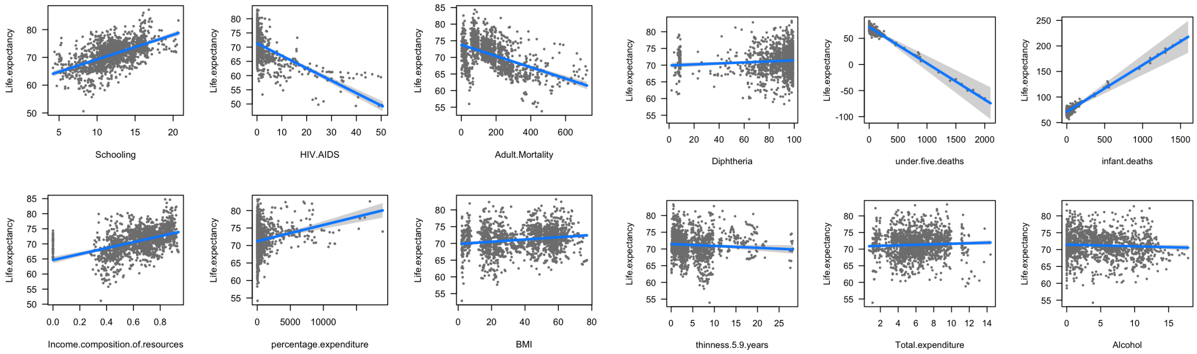
QR decomposition can be factored into orthogonal matrix Q and upper triangular matrix R. The cost of using QR decomposition approximate $2n^2(m - \frac{n}{3})$ $(O(n^3))$ where $m \geq n$. And compared to other methods, QR is more stable. The execution speed is 0.064.

When it is applicable, Cholesky should be twice as fast as the LU decomposition based on the flops count. However, the difference between the real results of these two approaches is not constant. And the QR method is not faster than either of these methods. So if we have a huge dataset, all three methods here might not work.

Further Analysis

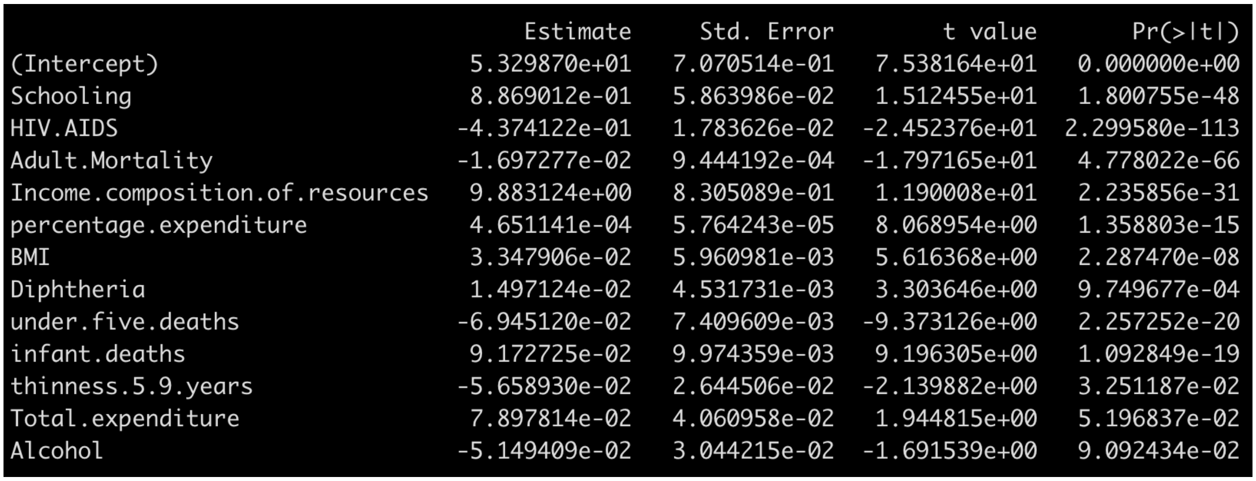
In this section, we will analyze the response variable life expectancy as well as our input variables and how they affect life expectancy in detail.

Firstly, we check the relationship of each input variable and life expectancy. From Fig. x plots we can notice that “Schooling”, “Adult.Mortality”, “Under.five.deaths”, “infant.deaths”, and “Income.composition.of.resources” have a relatively clear association with life expectancy, while other variables don’t show any clear relationships.



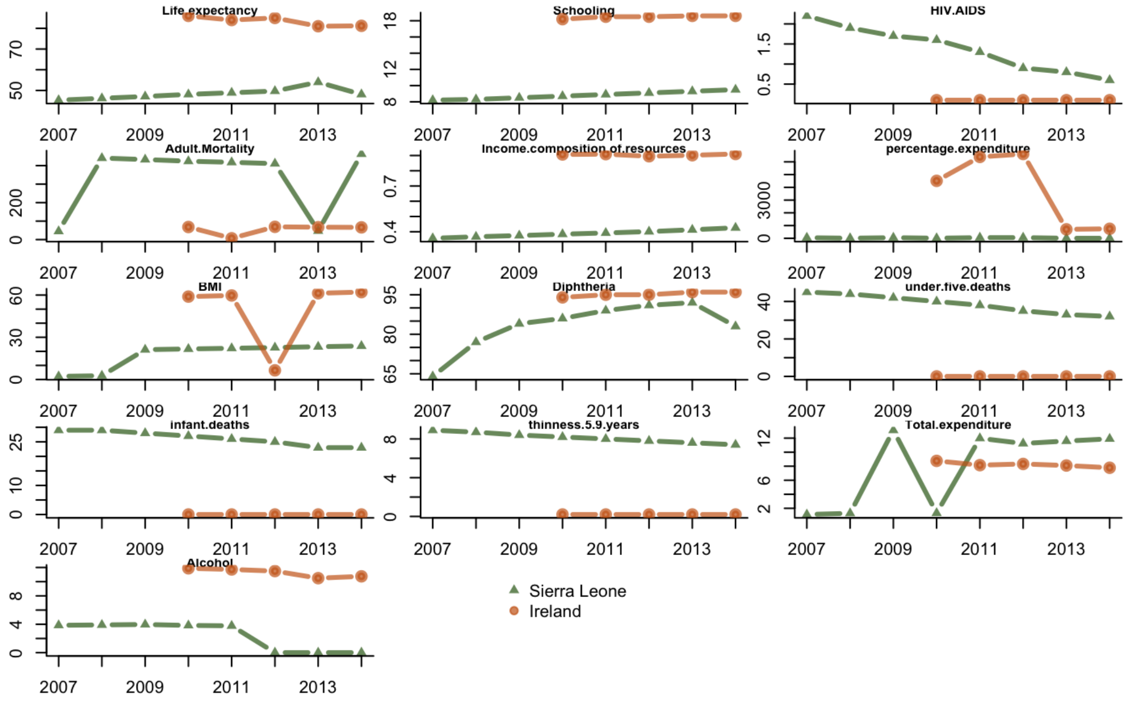
To continue looking at each input variable’s significance, we use our fitted regression model to calculate each variable’s p-value, where the smaller the p-value is, the more significant it will be.

From Table x, we can see that “thinness.5.9 years”, “Total.expenditure”, and “Alcohol” are not significant if we choose a significance level of 0.05, for their p-values are larger than 0.05. This indicates that these three variables may not be influential for life expectancy and the real influential variables are the remaining ones.



```
                                   Estimate     Std. Error      t value      Pr(>|t|)
(Intercept)                     5.329870e+01   7.070514e-02   7.538164e+01   0.000000e+00
Schooling                       8.869012e-01   5.863986e-02   1.512455e+01   1.800755e-48
HIV.AIDS                       -4.374122e-01   1.783626e-02  -2.452376e+01  2.299580e-113
Adult.Mortality                -1.697277e-02   9.444192e-04  -1.797165e+01   4.778022e-66
Income.composition.of.resources  9.883124e+00  8.305089e-01   1.190008e+01   2.235856e-31
percentage.expenditure          4.651141e-04   5.764243e-05   8.068954e+00   1.358803e-15
BMI                             3.347906e-02   5.960981e-03   5.616368e+00   2.287470e-08
Diphtheria                      1.497124e-02   4.531731e-03   3.303646e+00   9.749677e-04
under.five.deaths              -6.945120e-02   7.409609e-03  -9.373126e+00   2.257252e-20
infant.deaths                   9.172725e-02   9.974359e-03   9.196305e+00   1.092849e-19
thinness.5.9.years             -5.658930e-02   2.644506e-02  -2.139882e+00   3.251187e-02
Total.expenditure               7.897814e-02   4.060958e-02   1.944815e+00   5.196837e-02
Alcohol                        -5.149409e-02   3.044215e-02  -1.691539e+00   9.092434e-02
```

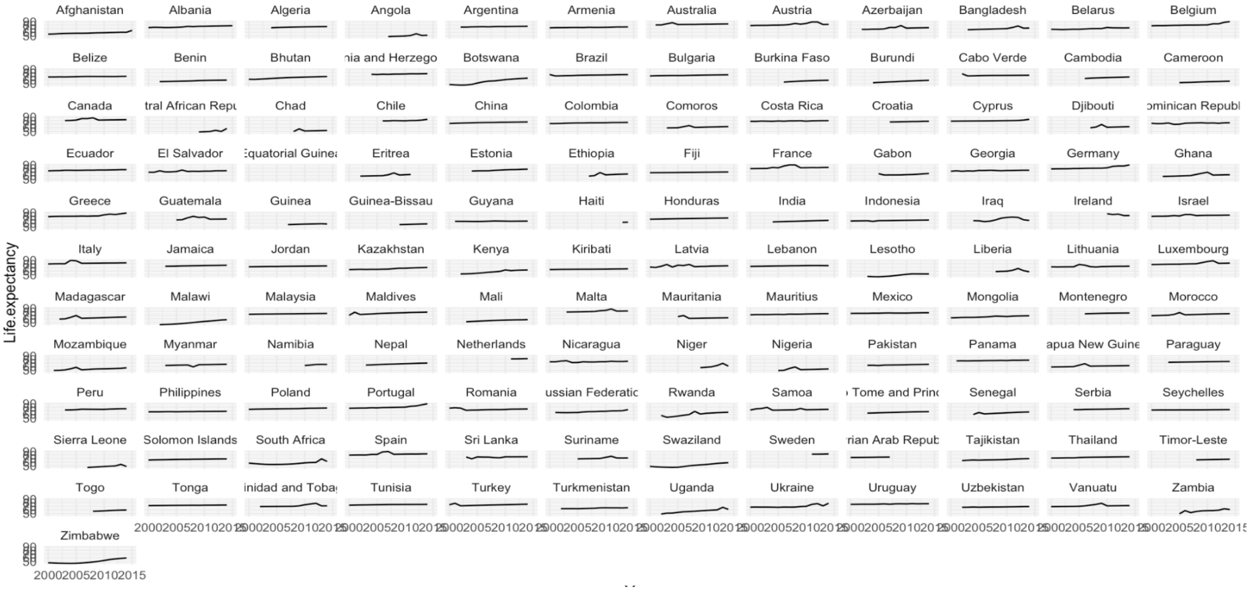
To check this implication drawn from p-values, we compare the performance between Ireland and Sierra Leone, the country with the highest average life expectancy and the one with the lowest respectively. We plot each country's performance on the input variables across time. The orange line represents countries with above-average life expectancy, while the green line represents the opposite.



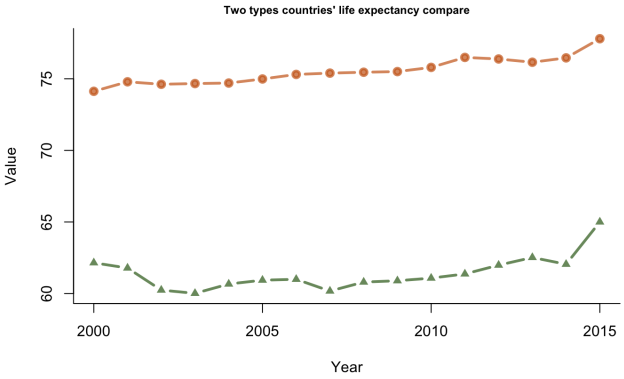
From the above plot, we can indeed notice some striking differences between Ireland and Sierra Leone. For example, Ireland has much higher years of schooling and percentages of expenditure on health compared to Sierra Leone, and this might suggest that having more schooling years and more expenditure on health as a percentage of GDP can help improve life expectancy. On the other hand, it seems that

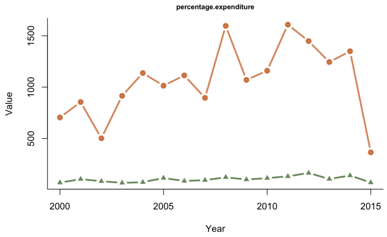
“Total.expenditure” and “Alcohol” are less influential. From the graph, Ireland has more alcohol consumption than Sierra Leone, but we can not simply conclude that having more alcohol can improve life expectancy. Besides, Sierra Leone’s government has a higher expenditure on health as a percentage of total government expenditure than Ireland beginning around 2011, but it still has a much lower life expectancy. It is very likely that there are more diseases in Sierra Leone, so its government has to spend more on health. As a result, neither “Total.expenditure” nor “Alcohol” can contribute to the difference in life expectancy between Ireland and Sierra Leone. This agrees with our previous conclusions from the p-values.

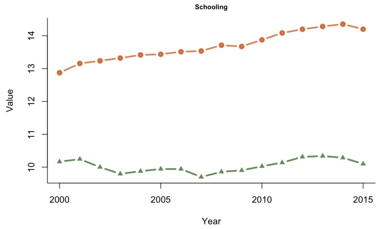
Secondly, we are interested in observing the overall life expectancy for all countries. We can notice that most countries have a very stable life expectancy trend. Although there might be some fluctuation for some countries, the overall trend of the life expectancy for every country is positive. However, by comparing the average life expectancy curves of each country, we still see that there are some countries where people live longer as time goes by. For example, the life expectancies of countries tend to go upward such as Botswana, Zimbabwe, Malawi, and Uganda. Without a doubt, there are also exceptional countries that have a negative life expectancy trend such as Liberia, and Ireland.

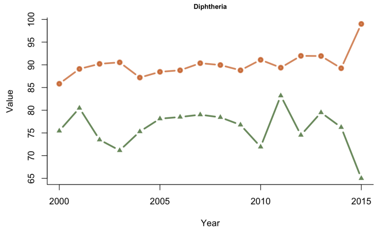


Therefore, we can divide the different countries into two different groups, one is the countries that have a life expectancy above the overall average, and the other is below the average. We are interested in observing which significant external features (such as schooling, percentage expenditure, Diphtheria, etc.) may affect life expectancy directly, and compare the results within two groups. The following picture is the display of two groups' average life expectancy trends between 2000 to 2015.









We can see that the orange line is higher than the green line in all these four features overall.

Percentage expenditure represents the expenditure on health as a percentage of Gross Domestic Product per capita(%). The higher the number, the more people in that country are spending on health care. This feature variates very frequently. Two of the largest changes were in 2001-2002 and 2014-2015. During these time periods, the expenditure on health was in sharp decline. Surprisingly, the life expectancy was still going upward for both groups. So we guess that the Percentage expenditure did not affect people's life expectancy directly.

Schooling represents the Number of Years of Schooling in the country. This variable is relatively flat and coincides with the curve of life expectancy. The countries where people have above average life expectancy have at least 13 years of educational background, while people from the remaining countries have only around 10 years. We can infer that countries with higher levels of education will pay more attention to their health issues, which will lead to an increase in life expectancy.

Diphtheria represents the Diphtheria tetanus toxoid and pertussis immunization coverage percentage among 1-year-olds. The Diphtheria vaccination rates for newborns in countries with longer life expectancy are at least 85%, while those in the remaining countries are 80% or less. The requirements for this vaccine are depending on the health policy of each country. However, in this case, we speculate that a higher vaccination rate would lead to a reduction in infant mortality for newborns, thus achieving the goal of higher average life expectancy.

Conclusion

In this project, we construct a multiple linear regression model for life expectancy and to check the model's performance through regression diagnostics and simulation study by using the model selection criteria of stepwise regression Akaike Information Criterion, and real analysis.

In part Proposed Method, we introduce the main methods that we used for this project are linear regression model, AIC model selection, least squares, and simulation study. In this section, all the result indicate that the methods that we chose for the project worked efficiently,

In part Data Analysis Study, the first thing we did was to process the dataset in advance. Following that, we performed model selection by exhaustive search and forward stepwise selection, and got the result which showed that different variables have different impacts on life expectancy. Later on, we started our Beta Calculation by using LU decomposition, Cholesky, and QR decomposition. In the result, the executive speed for LU decomposition is 0.224, the execution speed for Cholesky decomposition is 0.0055, and the executive speed for QR decomposition is 0.064. From these results, we can conclude that Cholesky has the fastest execution speed at 0.055 s.

In the last section, Further Analysis, we analyzed the response variable life expectancy. We found that “Schooling”, “Adult.Mortality”, “Under.five.deaths”, “infant.deaths”, and “Income.composition.of.resources” have a relatively clear association with life expectancy, while other variables don’t show any clear relationships. After that, we also concluded that most countries have a very stable life expectancy trend even though there are countries that have some fluctuations. Based on the results, we divided the countries into two groups. One is that countries have a life expectancy that is above the overall average, while the other’s life expectancy is lower than the overall life expectancy. We conclude that Percentage expenditure represents the expenditure on health as a percentage of Gross Domestic Product per capita(%). The higher the number, the more people in that country are spending on health care. Schooling represents the Number of Years of Schooling in the country. This variable is relatively flat and coincides with the curve of life expectancy. Diphtheria represents the Diphtheria tetanus toxoid and pertussis immunization coverage percentage among 1-year-olds. The Diphtheria vaccination rates for newborns in countries with longer life expectancy are at least 85%, while those in the remaining countries are 80% or less.

Lasso: Least absolute shrinkage and selection operator(Lasso) is a method of regression analysis that selects the best model by operating variable selection and regularization to improve the accuracy of the model performance. Even though it is created for linear regression models, the regularization of Lasso can also be applied to other statistical models such as proportional hazards models, generalized estimating equations, and generalized linear models. How Lasso selects the subset is based on the constraint form. *(https://en.wikipedia.org/wiki/Lasso_(statistics))*

```
library(performance)
library(ggplot2)
library(dplyr)

library(visreg)

life <- 路径

#summary
summary(life_x)

life_x <- life[,4:22]
```

#Data Analysis Study

A. Data Processing

```
#check na
sum(is.na(data))
#remove na
life_x=na.omit(life_x)
```

```
dim(life_x)
#boxplot of life_expantancy
ggplot(data=life_x,aes(y=life_x$Life.expectancy))+
  geom_boxplot(fill="blue")+
  ggtitle("Boxplot of Life Expectancy") +
  ylab("Life expectancy")

#histogram of life_expantancy
ggplot(data=life_x,aes(x=life_x$Life.expectancy))+
   geom_histogram(fill="blue")+
  ggtitle("Histogram of Life Expectancy") +
  xlab("Life Expectancy")

#remove outlier
life_x<-life_x[life_x$Life.expectancy>50, ]

#boxplot of new dataset
ggplot(data=life_x,aes(y=life_x$Life.expectancy))+
  geom_boxplot(fill="pink")+
  ggtitle("Boxplot of Life Expectancy") +
  ylab("Life Expectancy")

#histogram of new dataset
ggplot(data=life_x,aes(x=life_x$Life.expectancy))+
   geom_histogram(fill="pink")+
  ggtitle("Histogram of Life Expectancy") +
  xlab("Life Expectancy")
```

B. Model Selection

```
#define empty model

empty_model <- lm(Life.expectancy ~ 1, data=life_x)

#define model with all predictors

full_model <- lm(Life.expectancy ~ ., data=life_x)

#check assumption

summary(full_model)

vif(full_model)

step(full_model)

#exhaustive search

Exh_test <- ExhaustiveSearch(Life.expectancy ~ ., data = life2, family = "gaussian")

#forward selection

forward_model <- step(empty_model, direction='forward', scope=formula(all), trace=0)

forward_model$anova
```

```
# C. simulation study and model check
```

```
library(tidyverse)
library(ggplot2)
n = nrow(X)
p = ncol(X)
set.seed(141)

lm_c <-lm(Life.expectancy~.,data=forward_model$model)
beta_ <- lm_c$coefficients
X<- as.matrix(subset(xy, select = -Life.expectancy))
y_bar = X%*%as.matrix(beta_)
num_datasets = 100
Y_gen = sapply(1:num_datasets,function(i){y_bar+rnorm(n,mean = 0, sd = 1)})

Beta = solve(crossprod(X))%*%t(X)%*%Y_gen
Sigma = sqrt(colSums((Y_gen-X%*%Beta)**2)/n)

Beta_df = data.frame(t(Beta))
gather_df <- Beta_df %>% gather()

name<-data_frame(colnames(subset(xy, select = -Life.expectancy)))
name[,2] <- as.double(beta_);name[1,1] <- 'X1'
colnames(name) <- c("key","value")

estLine <- gather_df %>% group_by(key) %>%summarize(mean_x = mean(value))
trueLine<- name %>% gather()%>%group_by(key) %>%summarize(mean_x = mean(value))

ggplot(gather_df) +
  aes(value) +
  facet_wrap(~key, scales = 'free_x') +
  geom_histogram(bins = 10) +
  theme_minimal() +
  geom_vline(data  = estLine, aes(xintercept = mean_x, colour = key))+
  geom_vline(data  = trueLine, aes(xintercept = mean_x),linetype =2)
ggplot() +
    geom_histogram(aes(Sigma),bins = 10) +
    theme_minimal() +
    geom_vline(aes(xintercept = 1, color = 'true'))+
    geom_vline(aes(xintercept = mean(Sigma), color = 'est'))

#model check

check_model(model) #overall performance


# D. Beta calculation — Least-squares method
(xy <- add_column(forward_model$model,1,.after = 1))
y <- xy$Life.expectancy
X<- as.matrix(subset(xy, select = -Life.expectancy))
```

```
#LU
library(Matrix)
t1 <- proc.time()
luX <- lu(crossprod(X))
elu <- expand(luX)
P <- elu$P
L <- elu$L
U <- elu$U
# P%*%L%*%U
(beta_lu <- solve(P%*%L%*%U, crossprod(X,y)))
proc.time()-t1

#cholesky
t2 <- proc.time()
x_chol = chol(crossprod(X)) # X'X = LL'
(beta_cholesky <- solve(crossprod(x_chol),crossprod(X,y)))
proc.time()-t2

#QR
t3 <- proc.time()
QR <- qr(X)
Q <- qr.Q(QR)
R <- qr.R(QR)
(beta_QR <- backsolve(R,crossprod(Q,y)))
proc.time()-t3
```

```
# Further Analysis
#get coefficients and p-values
model <- lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
    Income.composition.of.resources + percentage.expenditure +
    BMI + Diphtheria + under.five.deaths + infant.deaths + thinness.5.9.years +
    Total.expenditure + Alcohol, data = life_x)
summary(model)$coefficients

par(mfrow=c(2,3))
visreg(model) #check each variable's performance in the model
```

```
# PART 1
#calculate each country's mean life expectancy
country_mean<- life %>% group_by(Country) %>%
         summarise(mean(Life.expectancy))
#get top countries with the lowest values of mean life expectancy
country_mean[order(country_mean$`mean(Life.expectancy)`),]

#get top countries with the highest values of mean life expectancy
country_mean[order(country_mean$`mean(Life.expectancy)`,decreasing = TRUE),]

#give year and country column to xy
top <- as.data.frame(cbind(life[,1:2, drop=FALSE],xy))
#select Ireland and Sierra Leone (lowest & highest Life.expec )
SL <- top %>%
  filter(Country== "Sierra Leone")
Ireland <- top %>%
  filter(Country == "Ireland")

# Plots for comparing Ireland and SL in a for loop

c <- colnames(SL)

par(oma = c(0.2,0.2,0.2,0.2),mfrow = c(5, 3),mar = c(2, 2, 0.3, 0.3))

for (i in 3:15) { # Loop over loop.vector

  # store data in column.i

  s <- SL[,i]

  ir <- Ireland[,i]

  # Plot histogram of x

  plot(s~SL$Year,

     type="b",

     bty="l",

     xlab="Year",

     ylab="Income.composition.of.resources",

     col=rgb(0.2,0.4,0.1,0.7),

     lwd=3,

     pch=17,

     ylim=c(min(s,ir),max(s,ir)), #reset y range

     main = paste(c[i]),

     cex.main=0.75)

  lines(ir ~ Ireland$Year,

        col=rgb(0.8,0.4,0.1,0.7) ,

        lwd=3 ,

        pch=19 ,
     type="b" )

}




# PART 2

# separate all countries into 2 groups

feat_<-which(life$Life.expectancy >= mean(life$Life.expectancy))

high_half<- life[feat_,]

lower_half <- life[-feat_,]

high_mean <- aggregate(high_half[, 4:length(high_half)], list(Year = high_half$Year), mean)

lower_mean <- aggregate(lower_half[, 4:length(high_half)], list(Year = lower_half$Year), mean)

mean_<- high_half %>%group_by(Year) %>%summarize(h_life_avg = mean(Life.expectancy))

mean_ <- cbind(mean_,

          lower_half %>%group_by(Year) %>%summarize(l_life_avg = mean(Life.expectancy))

          %>%select(l_life_avg))

# plot of average life expectancy

plot(mean_$l_life_avg ~ mean_$Year,

  type="b",

  bty="l",

  xlab="Year",

  ylab="Value",

  col=rgb(0.2,0.4,0.1,0.7),

  lwd=3,

  pch=17,

  ylim=c(min(mean_$h_life_avg,mean_$l_life_avg),

      max(mean_$h_life_avg,mean_$l_life_avg)), #reset y range

  main ="Two types countries' life expectancy compare" ,

  cex.main=0.75)

lines(mean_$h_life_avg ~ mean_$Year,
      col=rgb(0.8,0.4,0.1,0.7) ,
      lwd=3 ,
      pch=19 ,
      type="b")

# plot of different features
mean_name<-colnames(high_mean)
for (i in 3:length(high_mean)) {
  plot(lower_mean[,i]~lower_mean$Year,
   type="b",
   bty="l",
   xlab="Year",
   ylab="Value",
   col=rgb(0.2,0.4,0.1,0.7),
   lwd=3,
   pch=17,
   ylim=c(min(lower_mean[,i],high_mean[,i]),
       max(lower_mean[,i],high_mean[,i])), #reset y range
   main = mean_name[i],
   cex.main=0.75)
   lines(high_mean[,i]~high_mean$Year,
       col=rgb(0.8,0.4,0.1,0.7) ,
       lwd=3 ,
       pch=19 ,
       type="b")
}
```

Wrok cited

Medrixv https://www.medrxiv.org/content/10.1101/2022.04.05.22273393v1