HUDM 6122 Final Project

Getong Zhong

2023-05-11

Contents

1 Introduction

The dataset was collected from the Global Health Observatory (GHO) data repository maintained by the World Health Organization (WHO). In this data set, the 20 predicting variables that influence life expectancy were categorized into distinct groups, including immunization-related factors, mortality factors, economic factors, and social factors. The dataset consists of merged individual data files, resulting in a comprehensive dataset with 22 columns and 2938 rows, representing 193 countries. Initial visual inspection revealed missing values, primarily in variables related to population, Hepatitis B, and GDP. Due to difficulties in obtaining data for certain less known countries like Vanuatu, Tonga, Togo, Cabo Verde, these countries were excluded from the final model dataset.

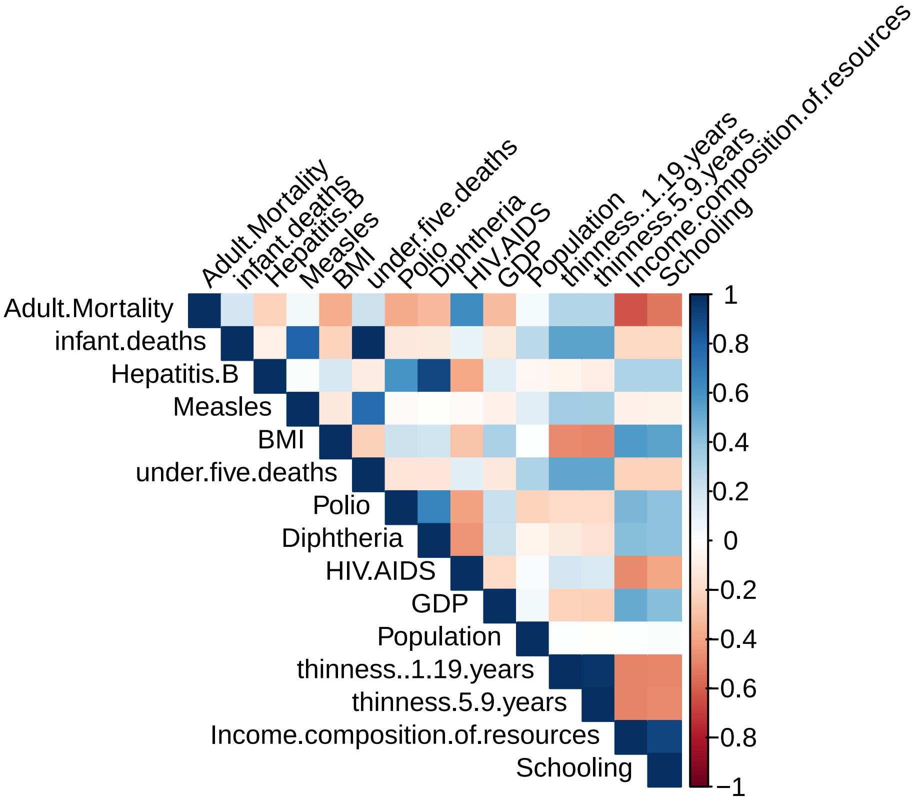
2 Data Cleannig

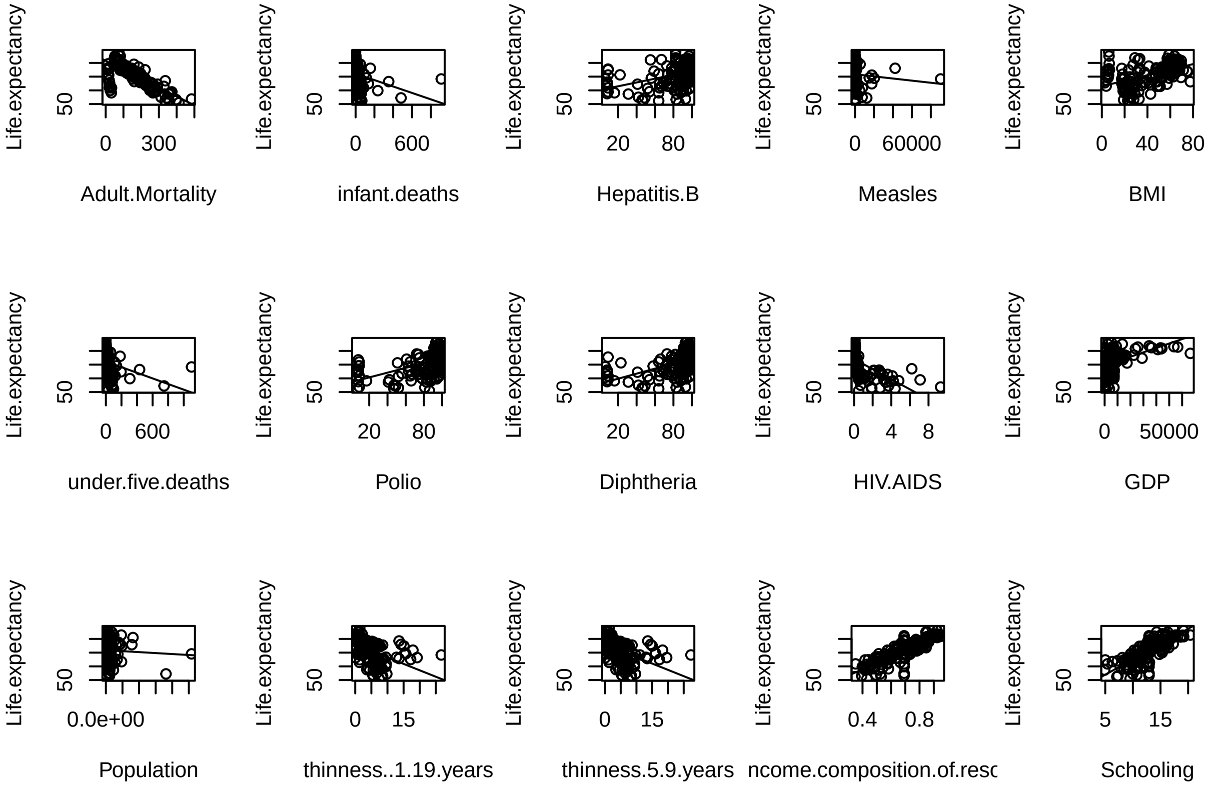
Certain columns with a high number of missing values or predominantly consisting of zeros were deemed unsuitable for analysis and were subsequently dropped from the dataset. Specifically, the “Alcohol” and “Total.expenditure” columns were removed due to a significant number of missing values, and the “percentage.expenditure” and “Income.composition.of.resources” columns were removed due to an excessive number of zero values. Besides, to handle the remaining missing values in other columns, we opted to impute them with the mean value of each respective column.

3 Multivariate Linear Regression

3.1 Relationship between Variables

The pairwise correlation matrix was generated to examine the relationships between variables. Darker shades of blue indicate positive correlations, while darker shades of red indicate negative correlations. It is evident that certain independent variables exhibit strong correlations, which may introduce multicollinearity issues into our analysis. Specifically, “Income composition of resources” and “Schooling”, “Infant death” and “Under five death”, “Infant death” and “Measles”, “thinness 1-19 Years” and “thinness 5-9 years” have Pearson correlation coefficients exceeding 0.7. These high correlations among predictors warrant caution in our analysis. To gain further insight into the relationship between predictors and the response variable, we constructed a scatter plot. It is noteworthy that “Infant death”, “Measles”, “Under five deaths” exhibit clustered values around zero, albeit with some outliers. This pattern is likely to impact our analysis. Notably, “Adult Mortality”, “Income composition of resources”, and “Schooling” demonstrate a linear relationship with the response variable.





3.2 Multivariate Linear Model for Life Expectancy

First we construct the initial model after drop some high-correlated term to avoid multicollinearity :

- Life.expectancy = 52.617528 + 0.031560(Polio) + 0.009943(BMI) - 0.598154(HIV.AIDS) + 29.709929(Income.composition.of.resources) - 0.026888(Adult.Mortality) - 0.004114(infant.deaths)

```
##
## Call:
## lm(formula = Life.expectancy ~ Polio + BMI + HIV.AIDS + Income.composition.of.resources +
##     Adult.Mortality + infant.deaths, data = life2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8806  -1.6417   0.0917   1.7066   9.0594
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     52.617528   1.847554  28.480  < 2e-16 ***
## Polio                            0.031560   0.011256   2.804  0.00562 **
## BMI                              0.009943   0.014030   0.709  0.47946
## HIV.AIDS                        -0.598154   0.230141  -2.599  0.01014 *
## Income.composition.of.resources 29.709929   2.405094  12.353  < 2e-16 ***
## Adult.Mortality                 -0.026888   0.003516  -7.648 1.28e-12 ***
## infant.deaths                   -0.004114   0.002905  -1.417  0.15840
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.164 on 176 degrees of freedom
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8483
## F-statistic: 170.6 on 6 and 176 DF,  p-value: < 2.2e-16
```

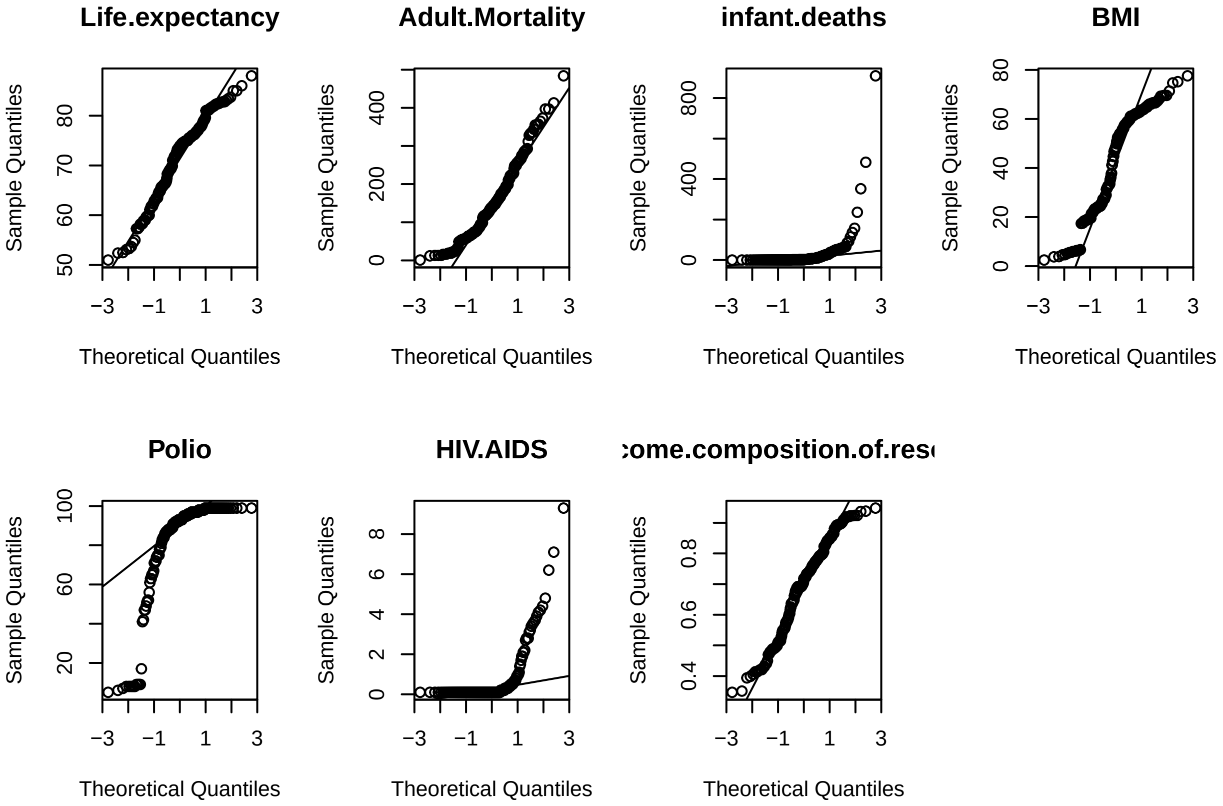
3.3 Assumption Check

To assess the assumption of constant variance, we examined the residual plots and mean squared error (MSE) plot. Upon analyzing these plots, we observed that there were no discernible patterns and the data points appeared to be randomly scattered. In addition, we also perform the Breusch-Pagan Test for Heteroscedasticity. Since the P-value of the test is greater than 0.05, we fail to reject the null hypothesis: : Error variance do not change with the level of the response. Hence, we can conclude that there is no evidence of heteroscedasticity or issues with the assumption of constant variance.

To assess the assumption of normality, we employed a QQ-plot. The majority of the points in the plot adhered closely to the diagonal line, indicating that the residuals approximate a normal distribution. Thus, we can infer that the assumption of normality holds. However, in the multivariate QQ plot, not all variables are stick to the normality line.

Additionally, we investigated the presence of influential observations and outliers. In the Residuals vs. Leverage plot, the majority of the data points fell within the range of ±2 standard deviations, suggesting that there are no severe issues with leverage or influential observations. However, it is worth noting that a few points, specifically observations 183 and 53, deviated from this range. Overall, the model appears to be robust with respect to influential observations and outliers.

```
##
##  studentized Breusch-Pagan test
##
## data:  MLR2
## BP = 28.219, df = 12, p-value = 0.005138
```



3.4 Multicollinearity

In this case, the VIF results reveal that the variables Polio, BMI, HIV.AIDS, infant.deaths, and Adult.Mortality have relatively low correlation with the other variables, as indicated by their VIF values of 1.35, 1.51, 1.74, 1.07, and 2.14, respectively. These values suggest that these variables do not exhibit significant multicollinearity issues. However, the variable Income.composition.of.resources has a VIF of 2.34, indicating a moderate level of correlation with the other variables, though not severe multicollinearity. Overall, these results suggest that there is no severe multicollinearity problem among the independent variables examined.

```
##                            Polio                              BMI
##                         1.354285                         1.514652
##                         HIV.AIDS Income.composition.of.resources
##                         1.741298                         2.339213
##                  Adult.Mortality                    infant.deaths
##                         2.138793                         1.074150
```

3.5 Autocorrelation

Autocorrelation occurs when there is a correlation between the residuals (or errors) of the model at different time points. In this case, the Durbin-Watson test shows a value of 2.126 with a corresponding p-value of 0.8023. Therefore we fail to reject the null hypothesis, which means there is insufficient evidence to conclude the presence of autocorrelation. Therefore, we can assume that the residuals in the regression model are not significantly correlated with each other.

```
##
##  Durbin-Watson test
##
## data:  MLR
## DW = 2.126, p-value = 0.8023
## alternative hypothesis: true autocorrelation is greater than 0
```

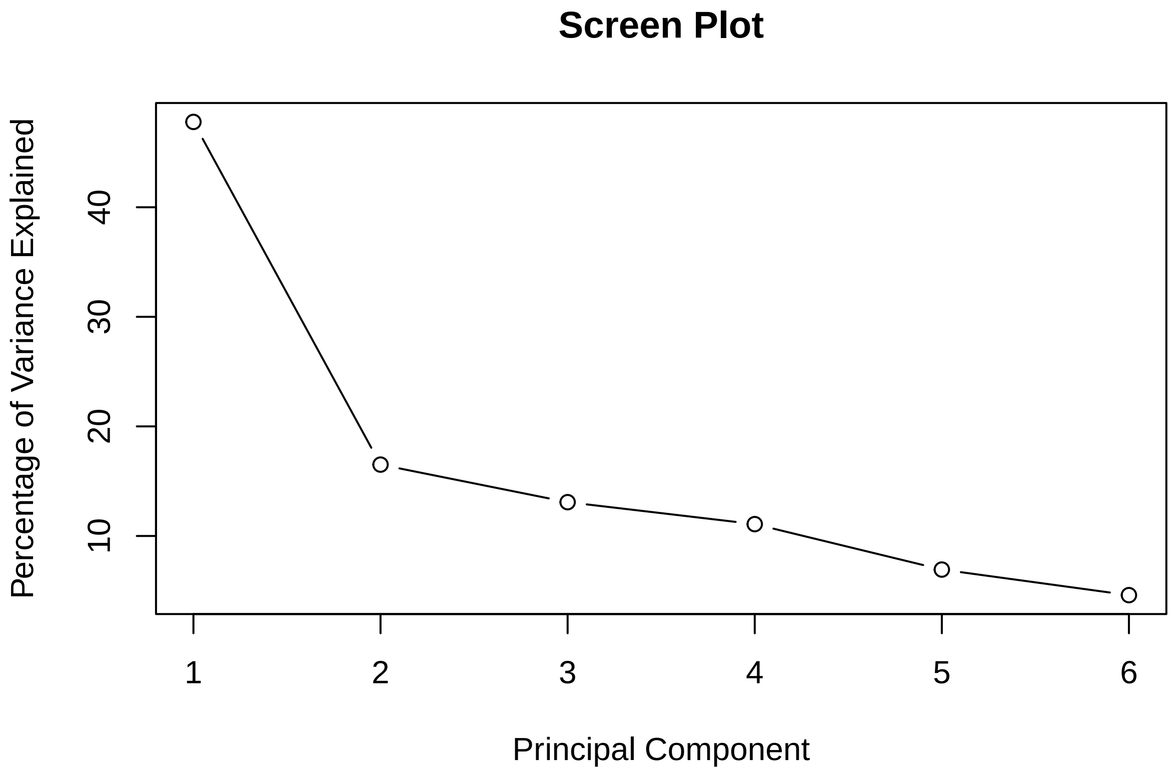
3.6 Discussion

While addressing multicollinearity and considering the clustering pattern observed in certain variables, we constructed a robust regression model. The significant predictors such as “Polio,” “BMI,” “HIV/AIDS,” “Income composition of resources,” “Adult Mortality,” and “Infant Deaths” contribute to our understanding of the complex dynamics influencing life expectancy.

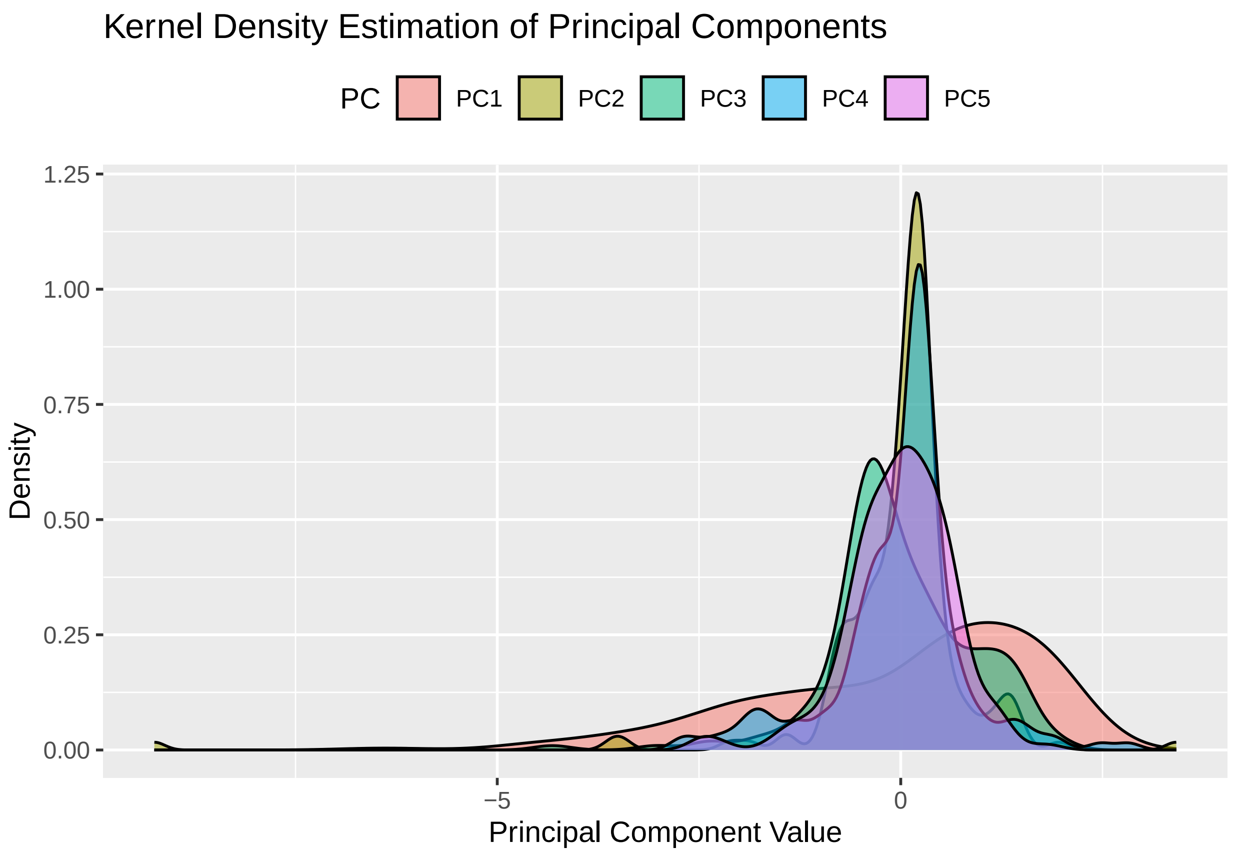
4 Principal Component Analysis

The results of the Principal Component Analysis (PCA) provide valuable insights into the underlying structure and variability of the data. The PCA output reveals that the first few principal components (PCs) carry significant importance in explaining the variability in the dataset. PC1 exhibits the highest standard deviation (1.6933) and accounts for a substantial proportion of the variance (47.79%). This indicates that PC1 captures the primary patterns and trends in the data. As we consider subsequent PCs, the standard deviation and proportion of variance decrease, implying that these components explain progressively smaller portions of the overall variability. However, it is important to note that the cumulative proportion of variance increases as we move along the PCs, demonstrating that the combined effect of multiple components provides a comprehensive understanding of the data. The cumulative proportion reaches 95.4% by PC5, indicating that these eight components explain a significant majority of the total variance. These results suggest that the dataset exhibits inherent structure that can be effectively summarized using a reduced set of principal components.

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5    PC6
## Standard deviation     1.6933 0.9954 0.8861 0.8154 0.64500 0.5254
## Proportion of Variance 0.4779 0.1651 0.1309 0.1108 0.06934 0.0460
## Cumulative Proportion  0.4779 0.6430 0.7739 0.8847 0.95400 1.0000
```



In the Kernel Density Estimation plot of different principal components, we observe that the density curves for PC2 and PC4 are similar, PC3 and PC5 are similar. Except PC1, kernel density cure for other PCs are symmetric This suggests that PC1 captures a significant portion of the variation in the data. Besides, we did not observe the density curve forany PCs to be bimodal, suggesting there is no presence of two distinct groups or clusters within the data. Additionally, we observe that there is some overlap between almost all the curves, indicating a degree of correlation or shared information between those principal components.



## Regression with Principal Components We estimated regression coefficient with the first 5 principal components, as they explained most of the variance in the data:

- Life Expectancy = 71.6169 + 4.2775(PC1) - 0.3653(PC2) - 0.4130(PC3) - 0.2203(PC4) - 2.6724(PC5)

```
##
## Call:
## lm(formula = Life.expectancy ~ pca$x[, 1:5], data = life2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7419  -1.6596   0.3153   1.6887   9.4115
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       71.6169     0.2402 298.217  < 2e-16 ***
## pca$x[, 1:5]PC1   4.2775     0.1422  30.078  < 2e-16 ***
## pca$x[, 1:5]PC2  -0.3653     0.2419  -1.510    0.133
## pca$x[, 1:5]PC3  -0.4130     0.2718  -1.520    0.130
## pca$x[, 1:5]PC4  -0.2203     0.2953  -0.746    0.457
## pca$x[, 1:5]PC5  -2.6724     0.3733  -7.158  2.1e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.249 on 177 degrees of freedom
## Multiple R-squared:  0.8445, Adjusted R-squared:  0.8401
## F-statistic: 192.2 on 5 and 177 DF,  p-value: < 2.2e-16
```

4.1 Autocorrelation

Autocorrelation occurs when there is a correlation between the residuals (or errors) of the model at different time points. In this case, the Durbin-Watson test shows a value of 2.0622 with a corresponding p-value of 0.662 Therefore we fail to reject the null hypothesis, which means there is insufficient evidence to conclude the presence of autocorrelation. Therefore, we can assume that the residuals in the regression model are not significantly correlated with each other.

```
##
##  Durbin-Watson test
##
## data:  MLR_pca
## DW = 2.0622, p-value = 0.662
## alternative hypothesis: true autocorrelation is greater than 0
```

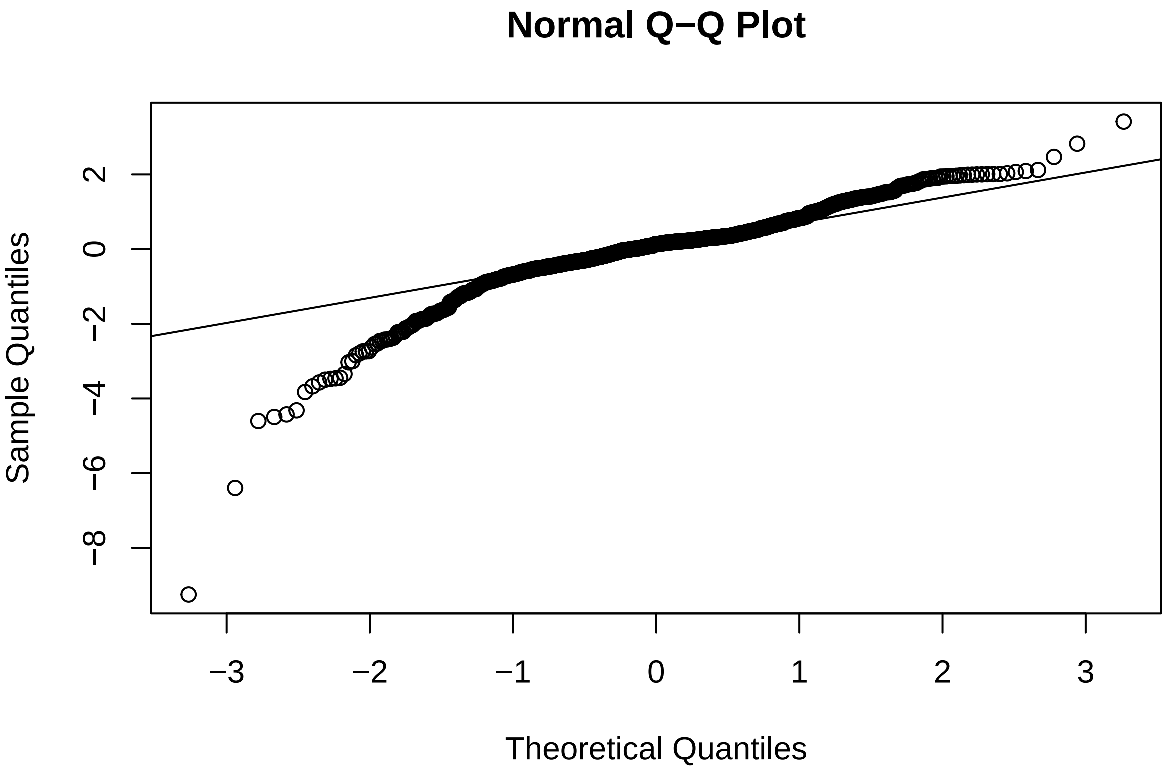
4.2 Multicollinearity

Multicollinearity is not an issue with a PCA model. Since all Principal Components are linearly independent

4.3 Normality

To check normality, we create QQplot and perform shapiro-wilk test for the first 5 PCs. Combined the two results we can would say that the points are not strictly stick to the normality line, especially the left tail of the graph clearly shows a little skewed, which matched with the shapiro-wilk test result, P-value less than 0.05, which means residuals are not normally distributed.

```
##
##  Shapiro-Wilk normality test
##
## data:  pca$x[, 1:5]
## W = 0.89423, p-value < 2.2e-16
```



4.4 Discussion

We try to construct a linear model for "Life Expectancy with the PCA's that we found in previous section. Keeping track of which original variable is contributing model much is a hard work to do when the predictors are Principal Components. And in our PC's, we can not give a clear information such: This variable contribute these specific PC's. Besides, there are some cautions about the PC model such as violation of normality for the residuals of PC.

5 Canonical Correlation Analysis

Canonical analysis is a statistical technique that allows us to assess the magnitude of linear relationships between sets of variables. The primary objective of this analysis is to examine the correlation between a linear combination of variables in one set and a linear combination of variables in another set. The process involves identifying the pair of linear combinations that exhibits the highest correlation, known as canonical variables, and quantifying their correlation, referred to as canonical correlations. These canonical correlations serve as indicators of the strength of association between the two sets of variables. By exploring the relationships among the canonical variables, we gain valuable insights into the underlying patterns and connections between the sets of variables under investigation.

5.1 Grouping Variables

Recall the variables that include in the data set, we can divide them in to the following group factors: immunization factors, mortality factors, economic factors, social factors and other health related factors. In particular, our grouping for this data set are:

- Immunization factors:
  - Hepatitis B: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
  - Polio: Polio (Pol3) immunization coverage among 1-year-olds (%)
  - Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- Mortality factors:
  - Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
  - Infant deaths: Number of Infant Deaths per 1000 population
  - under-five deaths: Number of under-five deaths per 1000 population
- Economic factors:
  - GDP: Gross Domestic Product per capita (in USD)
  - Income.composition.of.resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- Social factors:
  - Population: Population of the country
  - Schooling: Number of years of Schooling(years)
- Other health related factors:
  - Measles - number of reported cases per 1000 population
  - BMI: Average Body Mass Index of entire population
  - HIV.AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years)
  - thinness.1.19.years: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
  - thinness.5.9.years: Prevalence of thinness among children for Age 5 to 9(%)

```
##                      immunization mortality  economic    social
## immunization            1.0000000 0.5202029 0.4927223 0.5055755
## mortality               0.5202029 1.0000000 0.5055755 0.6376555
## economic                0.4927223 0.5055755 1.0000000 0.5830705
## social                  0.5055755 0.6376555 0.5830705 1.0000000
## other health related    0.6376555 0.5830705 0.8901233 0.9184800
##                      other health related
## immunization                    0.6376555
## mortality                       0.5830705
## economic                        0.8901233
## social                          0.9184800
## other health related            1.0000000
```

5.2 Discussion

In this section, we delve into the key aspects of obtaining the canonical variables and examining their correlations. The table presented showcases the highest squared values of the canonical variate pairs, shedding light on the extent of variation explained by each group. Notably, within the realm of immunization factors, we observe that 52% of the variation can be attributed to mortality factors. Similarly, economic factors account for 49% of the variation in this group. Social factors explaining 50% of the variation, while other health-related factors exhibit a remarkable explanatory power of 64%. Turning our attention to economic factors, we find that they account for 51% of the variation, while social factors explain 63% of the variation, and other health-related factors contribute to 58% of the observed variability. Furthermore, when considering variables related to the economic situation, social and other health-related factors offer additional explanatory capabilities, contributing to 58% and 89% of the variation, respectively. Lastly, within the realm of social factors, apart from immunization, mortality, and economic factor variables, a substantial portion of the variation (92%) can be attributed to other health-related factors within this dataset. Therefore variation in life expectancy is highly understandable with other health related factors.