University of California Davis

STA 141 - Fundamentals of Statistical Data Science

Final Project Report

Comprehensive Analysis of Life Expectancy

Authors
Chen Qian mrcqian@ucdavis.edu
Richard Ly rkly@ucdavis.edu
Yingzi Yang yziyang@ucdavis.edu

June 7, 2022

Contents

1 Introduction and Research Question

The standard of living and lifestyle choices varies across different geographical region. Various factors involved in standard of living and lifestyle choices may affect the life expectancy of a person. Our group conducted a study on what factors could affect the life expectancy of people living in different geographical regions. In particular, we examine Eastern Europe, Southern Asia, South-eastern Asia and Central America. Intuitively, regions with higher standards of living and healthier lifestyle choices would have higher life expectancy. If this is found to be true, this suggests that reasonably allocating the resources of a region to improve certain factors involved with standard of living and lifestyle choices may increase life expectancy. This results in the following research questions:

1. Which factors involved in standard of living and lifestyle choices have significant effects on life expectancy?

2. Is life expectancy different between different geographical regions?

In consideration of the questions posed on the proposal, we leave our response here. The data spans over the years 2000-2015 for 193 countries. We use this entire time frame to calculate mean values to fit the clustering and focus on the data of 2014 to fit the regression part. We do not look at the evolution of life expectancy in this report and will not use year as a predictor because of potential autocorrelation considerations. After the clusters were created based on chosen features, we can create thresholds of life expectancy to separate each cluster and classify countries into these clusters. In the submitted proposal, we considered logistic regression as it could be applied for binary or multiple classification. The response variable for the logistic regression is the clusters we created. Then, we go to the clusters created with the logistic regression and check the difference of life expectancy by utilizing the ANOVA model to compare models with new clusters, in a general A/B testing fashion. We later we found this was not appropriate for our data, so we omit the logistic regression and the ANOVA parts of this project in our final report.

2 Data Description

Collected and processed by the Global Health Observatory data repository under World Health Organization (WHO), the Life Expectancy dataset contains data related to the life expectancy of various geographical regions in the world. The dataset consists of 22 columns of features and 2938 rows of observations. The complete dataset includes data between the years of 2000-2015 for 193 countries, but we focus on the year of 2014 and examine the life expectancy of this year. We do not look into the evolution of life expectancy over time in this report, so the predictor variable ‘Year‘ is omitted from the linear regression performed. The 2014 subset contains 183 rows of observations. Aside from the qualitative variables Country and Status, the remaining variables are quantitative variables. We list and define the features in question below.

1. Country - country in which the data is obtained
2. Status - Whether the country is ‘developing‘ or ‘developed‘
3. Life Expectancy - life expectancy, measured in years
4. Adult Mortality - Adult Mortality Rates of both sexes (per 1000 population)
5. Infant Deaths- Number of Infant Deaths per 1000 population
6. Alcohol - Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
7. Percentage Expenditure - Expenditure on health as a percentage of GDP per capita(%)
8. Hepatitis B - Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
9. Measles - number of reported cases per 1000 population
10. BMI - Average Body Mass Index of entire population
11. Under Five Deaths - Number of under-five deaths per 1000 population
12. Polio - Polio (Pol3) immunization coverage among 1-year-olds (%)
13. Total Expenditure - General government expenditure on health as a percentage of total government expenditure (%)
14. Diptheria - Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
15. HIV/AIDS - Deaths per 1 000 live births HIV/AIDS (0-4 years)
16. GDP - Gross Domestic Product per capita (in USD)
17. Population - Population of the country
18. Thinness 1-19 years - Prevalence of thinness among adolescents for Age 10-19 (%)
19. Thinness 5-9 years - Prevalence of thinness among children for Age 5 to 9(%)
20. Income Composition of Resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
21. Schooling - Number of years of Schooling (years)

3 Data Visualization

Before we start fitting a linear regression model, we check the relationships between predictor variables. In our dataset, we have one categorical predictor variable (status), and 19 qualitative predictor variables. The plot of correlation matrix is below, and reveals the relationships between the different qualitative predictor variables. We note that there are a number of highly correlated predictor variables, and we take care to check for variance inflation factor (VIF) values that may indicate multicollinearity. Predictor variables with high VIF are removed as a remedy.

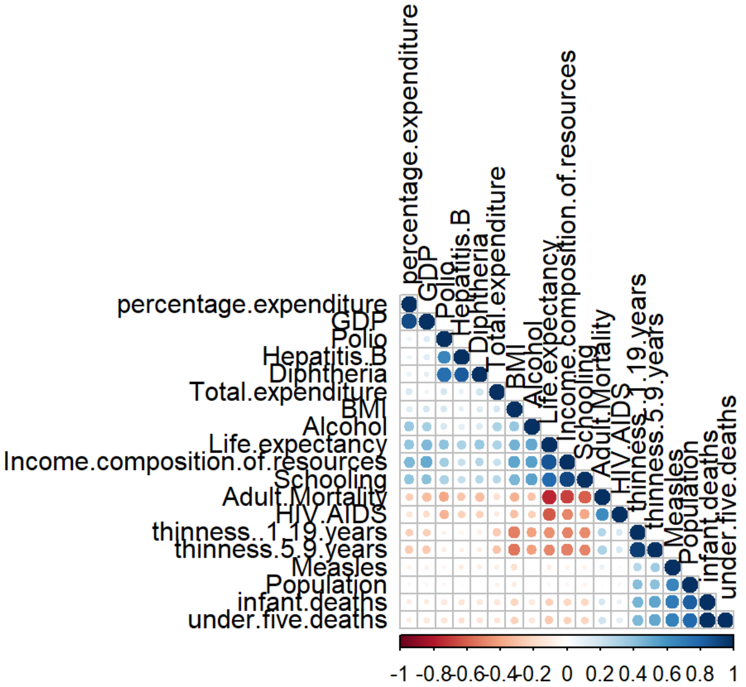


Figure 3.1: Correlation Between Quantitative Predictors

4 Data Analysis

4.1 Linear Regression

Although our original data comes from the years 2000-2015 for 193 countries, we choose data from the year 2014 to avoid data auto correlation and overfitting.

We check for missing values in our 2014 dataset. After looking though the missing values of population and GDP, we found that most of these observations are missing values of population and GDP at the same time. So we choose to remove these observations. After that, we still have some observations which are missing population values. We decide to use the mean population to fill in these missing values. We use the mean of developing countries to fill in developing observation's missing population values, and use means of developed countries to fill in developed observation's missing population values. For the remaining missing values, since the count of missing values are not large, we use mean of the variable to fill in the value.

For modeling, we initially fit the full model, first-order multiple regression model with 20 predictors. It was found to have insignificant predictor variables, as can be seen in Figure 7.1.

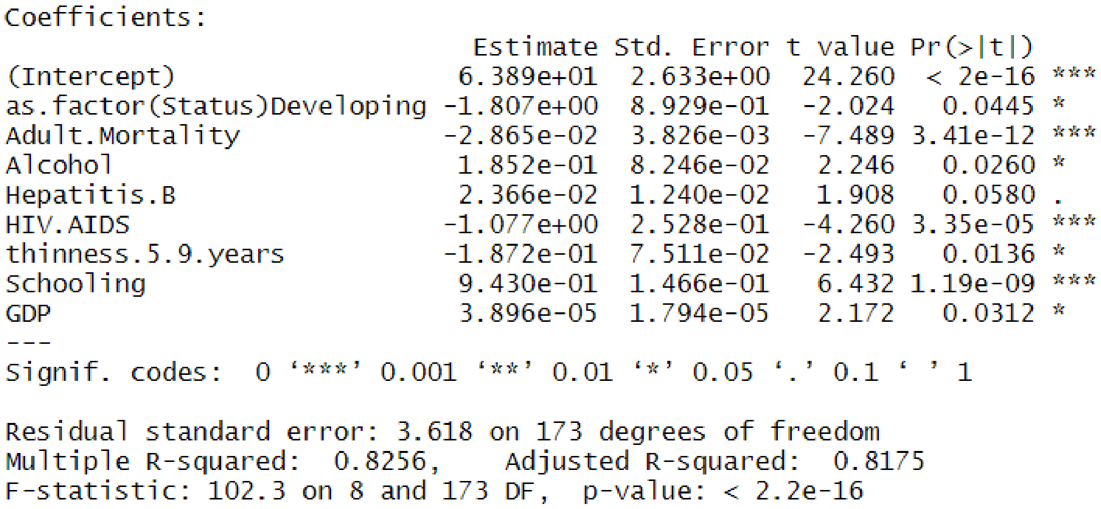
Next we check the plots in Figure 7.2 of model diagnostics. The residuals vs fitted plot does not show a nonlinear pattern, and it appears to not violate assumption of homoescedasticity. Based on the normal Q-Q plot, it shows more probability mass on both tails, but does not severely deviate from the assumption of a normal distribution. Lastly, the leverage plot confirms the model is not significantly affected by influential outliers. Now we check the variance inflation factor for each predictor variable, and we find several groups of intercorrelated predictor variables, such as 'infant.deaths' and 'under.five.deaths', 'thinness..1.19.years' and 'thinness.5.9.years'. To remedy this, we remove one or several of them to reduce multicollinearity.

After removing highly correlated predictor variables, we fit the new model with the remaining variables. However, the summary of the new linear regression fit continues to show variables that are not significant. In order to achieve a good balance of bias and variance, we apply backward stepwise selection with respect to BIC to our model. This yields a models with 8 features. We again check for any violations of the assumptions on the models using the plots in Figure 7.3. The residuals vs fitted values plot does not show any obviously nonlinear pattern. In addition, it appears to not violate the assumption of homoescedasticity. Based on the normal Q-Q plot, it shows lightly heavy-tailed distribution, in comparison to the assumption of normally distributed errors.

The final model is chosen to be the model found using the BIC procedure as it is a less complex model. with adjusted R^2 0.8175 with 173 degrees of freedom. Checking the diagnostic plots of the final model, there were no obvious violation of assumptions as well.

The final model is:

$\text{Life\_expectancy} = 63.89 - 1.81 \cdot \mathbb{1}(\text{Status}=\text{Developing}) - .03\text{Adult\_Mortality} + .19\text{Alcohol} + .02\text{Hepatitis} - 1.08\text{HIV\_AIDS} - .19\text{Thinness} + .94\text{Schooling} + .000039\text{GDP}$



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                6.389e+01  2.633e+00  24.260  < 2e-16 ***
as.factor(Status)Developing -1.807e+00  8.929e-01  -2.024   0.0445 *  
Adult.Mortality           -2.865e-02  3.826e-03  -7.489 3.41e-12 ***
Alcohol                    1.852e-01  8.246e-02   2.246   0.0260 *  
Hepatitis.B                2.366e-02  1.240e-02   1.908   0.0580 .  
HIV.AIDS                  -1.077e+00  2.528e-01  -4.260 3.35e-05 ***
thinness.5.9.years        -1.872e-01  7.511e-02  -2.493   0.0136 *  
Schooling                  9.430e-01  1.466e-01   6.432 1.19e-09 ***
GDP                        3.896e-05  1.794e-05   2.172   0.0312 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.618 on 173 degrees of freedom
Multiple R-squared:  0.8256,	Adjusted R-squared:  0.8175 
F-statistic: 102.3 on 8 and 173 DF,  p-value: < 2.2e-16
```

Figure 4.1: Summary of the final model chosen

4.2 Hierarchical Clustering

Again we use the 2014 dataset with missing values removed or filled. Before we apply hierarchical clustering, we scale the data as we use euclidean distance in hierarchical clustering, and the distance will change dramatically for the various units of the predictor variables. After scaling, we can use the dataset in the hierarchical clustering process with complete linkage and Euclidean distance.

We define a function ‘map_cluster(hc.fit,h)‘ in which we cut the resulting dendrogram ‘hc.fit‘(Figure 7.4) at height h=10, resulting in 5 distinct clusters. The boxplot in Figure 4.2 below for the 5 clusters reveal clear differences in median life expectancy. Further, the function ‘map_cluster(hc.fit,h)‘ plots the countries of each cluster, color coded by cluster, to a world map in Figure 4.3 to gain a better spatial understanding of the clusters, and label these clusters with a representative county. This enhanced spatial understanding of the clusters helps us address research question 2. (The full function code and details may be found in section 7.2). We find that countries in Europe and developed countries like Canada and Australia are almost in the cluster 2 which has the highest life expectancy on median with 80.7. Most Countries in South America, Africa and South East Asia are clustered in the second highest life expectancy group with 67.3, except for 2 observations which may be considered outliers. There are 4 countries in Africa that are clustered in the lowest life expectancy group with 52.6 on median.

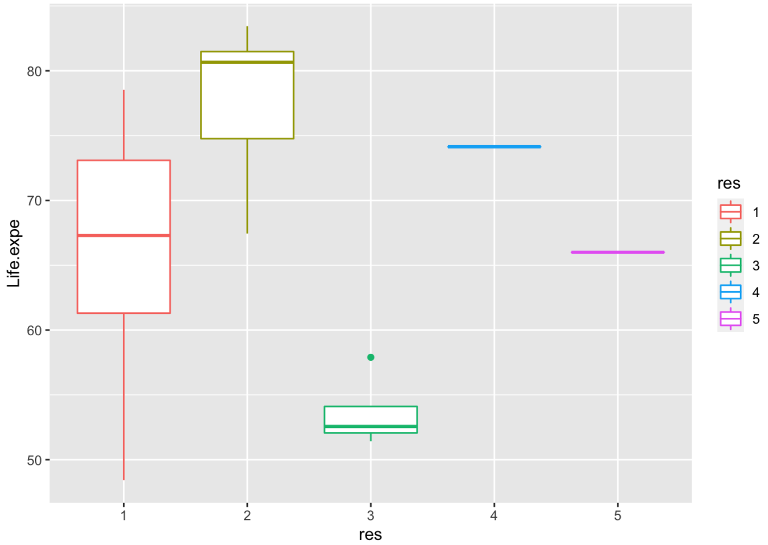


Figure 4.2: Boxplot of Life Expectancy of Clusters

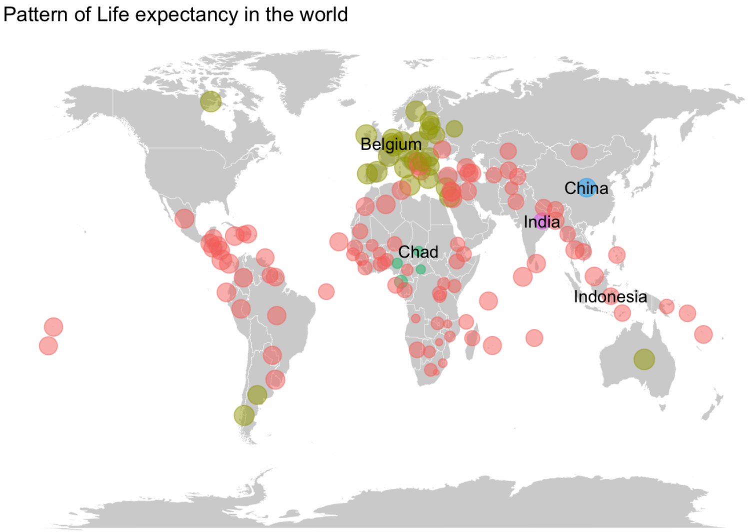


Figure 4.3: Map of Clusters

5 Concluding Thoughts

5.1 Results

5.1.1 Research Question 1

To identify which factors involved in standard of living and lifestyle choices have significant effects on life expectancy, we use linear regression. From fitting a full model with all 20 features, we found there to be many insignificant features. In addition, there were variables with high VIF values found. We remove these insignificant features and features with high multicollinearity. In addition, we use a stepwise procedure to reduce the model to one that best balances bias and variance to obtain the final described in section 4.1. We find that Status, Adult Mortality, Alcohol, Hepatitis B, HIV/AIDS, Thinness of 5-9 year olds, Schooling, and GDP are most significant in predicting life expectancy.

5.1.2 Research Question 2

To address whether life expectancy is different between different geographical regions, we use hierarchical clustering. By cutting the resulting dendrogram at a height of 10, we obtained 5 distinct clusters. We plot the life expectancy of each of these clusters in a side-by-side boxplot and are able to compare the median and spread of life expectancy for each cluster in this way. Each country in these clusters were plotted on a world map, and color coded according to their cluster. We find evidence that life expectancy may be different for different geographical regions, as 4 of the 5 clusters are centralized in a general geographic region. However, it is important to note that the red cluster, containing the largest number of countries, is spread throughout the regions studied. Therefore, it can not be definitively concluded that life expectancy is different between different geographical regions.

5.2 Conclusion

With respect to the first research question proposed, we were able to find a conclusive answer. From fitting a linear regression model to the 2014 dataset, we were able to find a subset of significant predictor variables among the original set of features. With this result, we were able to identify the 8 most significant features that contribute towards life expectancy among the countries that were included in the dataset. From the finalized model, the coefficients shows that life expectancy increases by 0.19 for every unit increase in pure alcohol consumption ; by 0.02 every percent increase in Hepatitis B immunization coverage among 1-year-olds; by 0.94 for every unit increase in number of years of schooling; and very lightly increase in GDP. And the coefficients shows that life expectancy decreases by 1.81 if the country is developing country; by 0.03 for every unit increase in adult mortality rates; by 1.08 for every unit increase in death per 1000 born with HIV; by 0.19 for every percent increase in prevalence of thinness of children of 5-9.

As for the second research question proposed, we are not able to give a definitive answer. While we were able to cluster the countries in the data according to their features and observe clear difference in median life expectancies, we were not able to conclude that life expectancy may be dependent on geographic region. While 4 of the 5 clusters are centralized in a geographic region, the spread of the 5th and most prevalent cluster provides great uncertainty in answering this research question.

5.3 Discussion

The countries in our data were clustered into 5 clusters based on immunization, mortality, economic, social and other health related factors. Regions with highly immunization rate, good economic indicators such as GDP, and low mortality rate, such as developed countries like Europe or Canada, have higher life expectancy. The world map in Figure 4.3 shows that most high life expectancy countries are distributed in Europe and that most lower life expectancy countries are found in Africa. This suggests that African countries could improve life expectancy by allocating resources towards the subset of features, found through linear regression, to be significant with respect to life expectancy. Factors such as GDP, however, are not so easily improved upon by an individual country, and perhaps could receive assistance from higher life expectancy countries with higher GDP's in order to obtain greater equality in life expectancy.

Interestingly, 2 outliers were found based after hierarchical clustering; They are the two most populous countries in the world. China has a moderate high life expectancy and India has a moderate low life expectancy. This leaves an open research question to be explored in the future for these two countries: Why do the clusters of China and India have such a narrow interval of life expectancy? What factors contribute towards this? Such questions could be addressed in a future project.

6 References

[1] Source of data: https://www.kaggle.com/kumarajarshi/life-expectancy-who

7 Appendix

7.1 Individual Contributions

Richard Ly - Contributed to obtaining, analyzing, and organizing results from both linear regression and hierarchical clustering. Contributed largely to the written report.

Chen Qian - Contributed to performing hierarchical clustering and generating related plots. Contributed to the written report.

Yingzi Yang - Contributed to the data visualization section. Contributed to fitting linear regression models and identifying a final model, as well as checking model diagnostics. Contributed to the written report.

7.2 Self-Written Function

~~~
```{r}
map_clusters <- function(hc.fit,h=10){
  res = cutree(hc.fit, h = 10) # We can choose 10 as our height to cut

res = as.data.frame(res)
res$region <- rownames(res)
world <- map_data("world")
loc = world[world$region%in% rownames(res),] %>% group_by(region) %>%
summarise_at(vars("long", "lat"),list(name = mean))
res$res = as.factor(res$res)
res$Life.expe = df[,"Life.expectancy_name"]

res = merge(x=res,y=loc,by="region",all.x=TRUE)

# replace the long and lati manually for regions contain missing in
# these 2 variables
res[res$region == "Trinidad␣and␣Tobago",]$long_name = -61.311951
res[res$region == "Trinidad␣and␣Tobago",]$lat_name = 10.536421

res[res$region == "Syrian␣Arab␣Republic",]$long_name = 36.278336
res[res$region == "Syrian␣Arab␣Republic",]$lat_name = 33.510414
~~~

```
res[res$region == "Russian␣Federation",]$long_name = 37.618423
res[res$region == "Russian␣Federation",]$lat_name = 55.751244

res[res$region == "Cabo␣Verde",]$long_name = -22.807835
res[res$region == "Cabo␣Verde",]$lat_name = 16.095011

# We choose the median for robust the outliers.

ggplot(res, aes(x=res, y=Life.expe, color=res)) +
  geom_boxplot()

life_median = res %>%group_by(res) %>% summarise_at(vars(Life.expe),
list(name = median))
life_median # Table 1

# World map

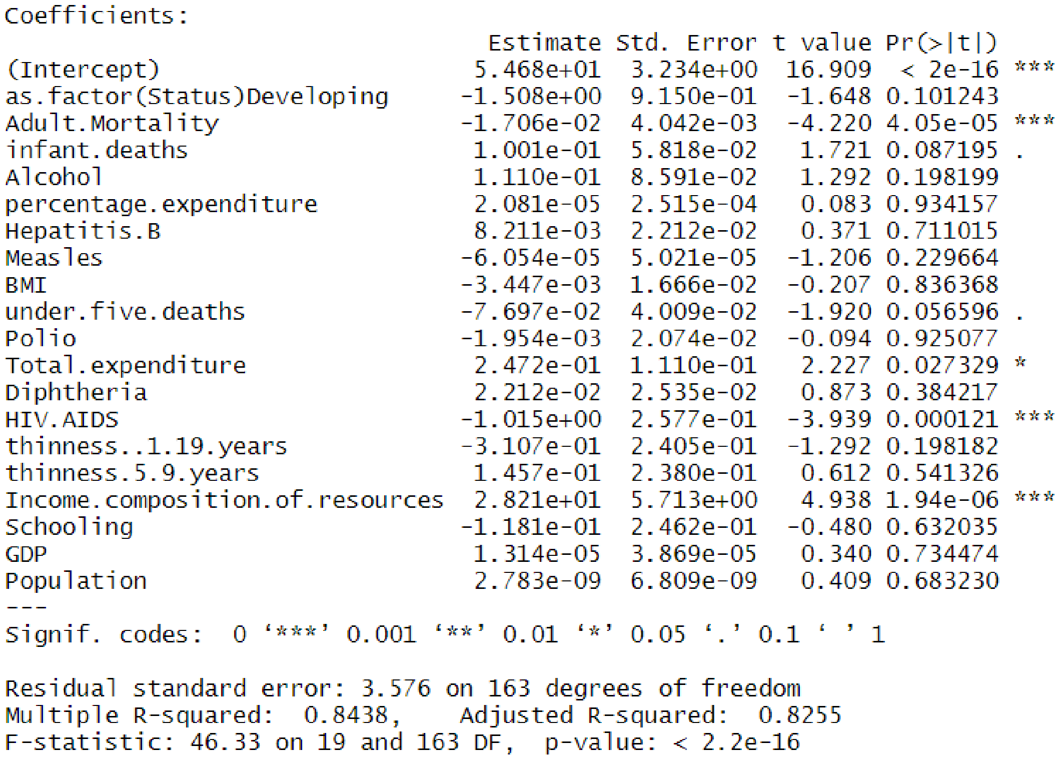
# choose the representative for each cluster
res %>% filter(Life.expe %in% (life_median$name))
res[res$res == 1,][which(abs(res[res$res == 1,]$Life.expe -
life_median$name[1])== min(abs(res[res$res == 1,]$Life.expe -
life_median$name[1])))),]
res[res$res == 3,][which(abs(res[res$res == 3,]$Life.expe
- life_median$name[3]) == min(abs(res[res$res == 3,]$Life.expe -
life_median$name[3])))),]


ggplot() +
  geom_map(
    data = world, map = world,
    aes(long, lat, map_id = region),
    color = "white", fill = "lightgray", size = 0.1
  ) +


  geom_point(
    data = res,
    aes(long_name, lat_name,
        color = res,
        size=Life.expe),
    alpha = 0.5
  ) +
  #labs(x = NULL, y = NULL, color = NULL)+
```

~~~
  theme_void() +
  theme(legend.position = "none")+
  labs(title="Pattern␣of␣Life␣expectancy␣in␣the␣world") +
  geom_text(data=res[res$region %in% c("China", "India",
  "Belgium", "Indonesia", "Chad"),], aes(long_name, lat_name, label=region))
}
```
~~~

7.3 Additional Figures and Plots



```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    5.468e+01  3.234e+00  16.909  < 2e-16 ***
as.factor(Status)Developing   -1.508e+00  9.150e-01  -1.648 0.101243    
Adult.Mortality               -1.706e-02  4.042e-03  -4.220 4.05e-05 ***
infant.deaths                  1.001e-01  5.818e-02   1.721 0.087195 .  
Alcohol                        1.110e-01  8.591e-02   1.292 0.198199    
percentage.expenditure         2.081e-05  2.515e-04   0.083 0.934157    
Hepatitis.B                    8.211e-03  2.212e-02   0.371 0.711015    
Measles                       -6.054e-05  5.021e-05  -1.206 0.229664    
BMI                           -3.447e-03  1.666e-02  -0.207 0.836368    
under.five.deaths             -7.697e-02  4.009e-02  -1.920 0.056596 .  
Polio                         -1.954e-03  2.074e-02  -0.094 0.925077    
Total.expenditure              2.472e-01  1.110e-01   2.227 0.027329 *  
Diphtheria                     2.212e-02  2.535e-02   0.873 0.384217    
HIV.AIDS                      -1.015e+00  2.577e-01  -3.939 0.000121 ***
thinness..1.19.years          -3.107e-01  2.405e-01  -1.292 0.198182    
thinness.5.9.years             1.457e-01  2.380e-01   0.612 0.541326    
Income.composition.of.resources  2.821e+01  5.713e+00   4.938 1.94e-06 ***
Schooling                     -1.181e-01  2.462e-01  -0.480 0.632035    
GDP                            1.314e-05  3.869e-05   0.340 0.734474    
Population                     2.783e-09  6.809e-09   0.409 0.683230    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.576 on 163 degrees of freedom
Multiple R-squared:  0.8438,	Adjusted R-squared:  0.8255 
F-statistic: 46.33 on 19 and 163 DF,  p-value: < 2.2e-16
```

Figure 7.1: Summary of the full model

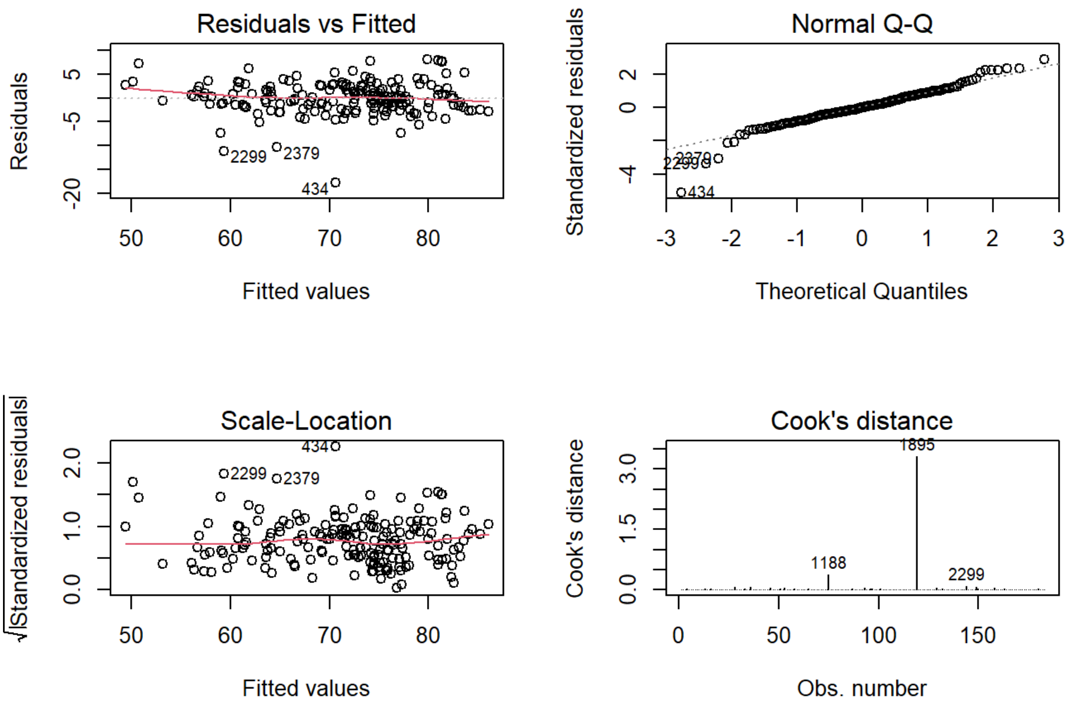


Figure 7.2: Model diagnostic plots of the full model

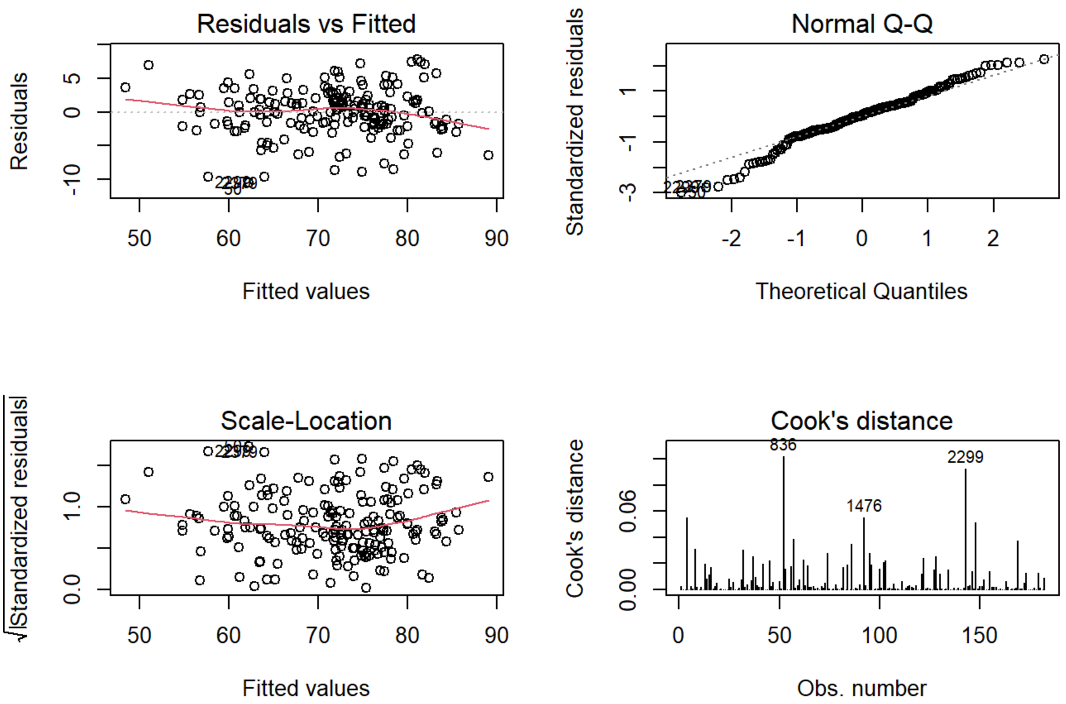


Figure 7.3: Model diagnostic plots of the final model

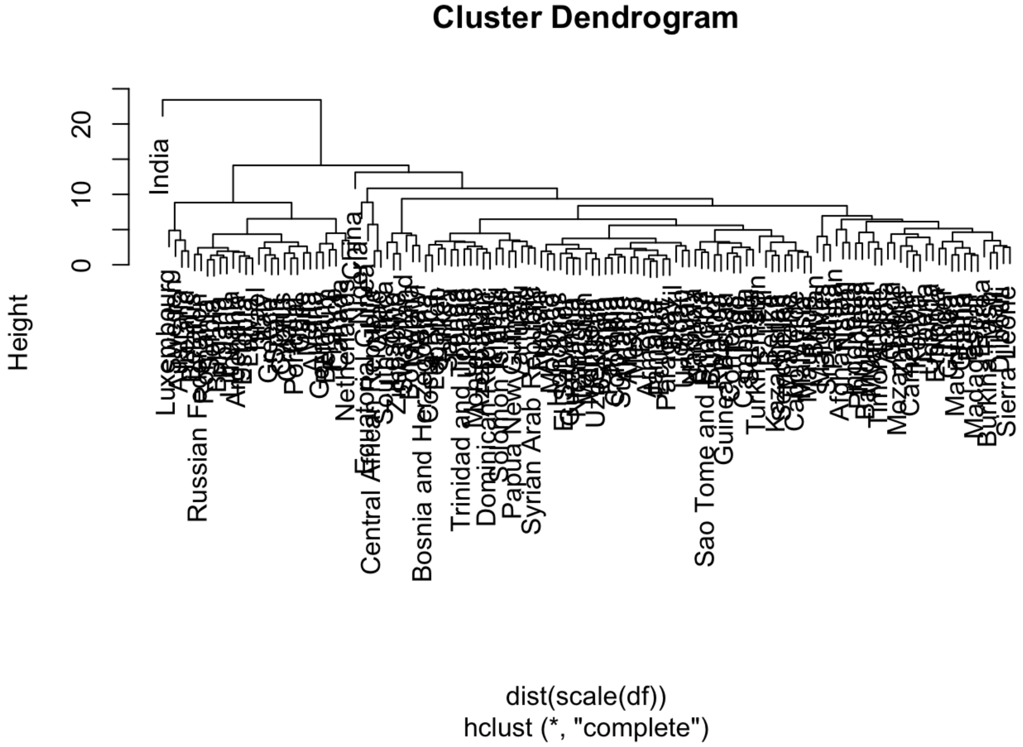


Figure 7.4: Dendrogram from Hierarchical Clustering

7.4 Code

~~~
```{r,eval=FALSE}
#load data and choose a fixed year
data0<-read.csv("D://data//Life␣Expectancy␣Data.csv")
data<-data0[which(data0$Year==2014),]
data<-data[,-2]
sort(sapply(data, function(x) sum(is.na(x))),decreasing = TRUE)

#deal with missing values
#delete rows that population and GDP are missing
data<-data[-which(is.na(data$Population) & is.na(data$GDP)),]

sort(sapply(data, function(x) sum(is.na(x))),decreasing = TRUE)
#fill population with categorical mean values
pop_mean_developing=mean(data[which(data$Status=='Developing'),]$Population,
na.rm = T)
~~~

```
pop_mean_developed=mean(data[which(data$Status=='Developed'),]$Population,
na.rm = T)

data[which(is.na(data$Population) & data$Status=='Developing'),'Population']
<- pop_mean_developing
data[which(is.na(data$Population) & data$Status=='Developed'),'Population']
<- pop_mean_developed

#fill the rest missing values with col mean
fillin<-function(data){
  for(i in 1:ncol(data)){
    data[is.na(data[,i]), i] <- mean(data[,i], na.rm = TRUE)
  }
  return(data)
}
data<-fillin(data)


#correlation matrix
library(corrplot)
corrplot(cor(data[,4:21]), type = "lower", order = "hclust", tl.col = "black")
#high correlated variables
for (i in 1:length(corr)){
  cat(colnames(corr)[i],'and␣', rownames(corr)[which(abs(corr[,i]) > 0.8
  & abs(corr[,i]) < 1)],'\n')
}

fit.e1<-lm(Life.expectancy~., data=data)
length(fit.e1$coefficients)
summary(fit.e1)
par(mfrow = c(2, 2))
plot(fit.e1,which = 1:4)
car::vif(fit.e1)

fit.e2 <-
  lm(Life.expectancy~as.factor(Status)+Adult.Mortality+Alcohol
     +Hepatitis.B+Measles+BMI+Polio
     +Total.expenditure+HIV.AIDS+thinness.5.9.years
     + Schooling + GDP + Population, data=data)
length(fit.e2$coefficients)
summary(fit.e2)
plot(fit.e2)
car::vif(fit.e2)
```

~~~
#No variable's vif>5

#stepwise regression
#AIC selection
library(MASS)
none_mod <- lm(Life.expectancy ~ 1, data = data)
fit.e2aic <-stepAIC(none_mod, scope = list(upper = fit.e2, lower = ~1),
direction = "both", k = 2, trace = FALSE)
summary(fit.e2aic)

fit.e2bic <- stepAIC(fit.e2, scope = list(upper = fit.e2, lower =~1),
direction = "both", k = log(n), trace = FALSE)
summary(fit.e2bic)

#r2(aic)>r2(bic)
fit.e2best <- fit.e2aic
car::vif(fit.e2best)
par(mfrow = c(2, 2))
plot(fit.e2best,which = 1:4)

#exam y outliers
p <- length(fit.e2best$coefficients)
stu.res.del <- studres(fit.e2best)
head(sort(abs(stu.res.del), decreasing = TRUE))
qt(1-.1/(2*n), n-p-1)
#exam x outliers
h <- influence(fit.e2best)$hat
c <- 2 * p/n
sort(h[which (h > c)], decreasing = TRUE)
#see cook's distance
par(mfrow = c(1, 1))
plot(fit.e2best,which = 4)

fit.e2best <- lm(formula = Life.expectancy ~ as.factor(Status) +
Adult.Mortality + Alcohol + Hepatitis.B + HIV.AIDS + thinness.5.9.years +
Schooling + GDP, data = data,subset=setdiff(rownames(data),"434"))
summary(fit.e2best)

par(mfrow = c(2, 2))
plot(fit.e2best,which = 1:4)
```

```{r include=F}
~~~

~~~~
knitr::opts_chunk$set(warning = FALSE, message = FALSE, fig.pos = 'H')
options(repos="https://cran.rstudio.com")
install.packages("maps")
library(ggplot2)
library(dplyr)
library(maps)
life <- read.csv("/Users/chenqian/Downloads/Final_project_141A/
LifeExpectancyData.csv")
life = na.omit(life)
life$Country = as.factor(life$Country)
```

```{r}
df = life %>%group_by(Country) %>%
summarise_at(vars(colnames(life)[-c(1,2,3)]),list(name = mean))
df = as.data.frame(df)
rownames(df) <- df[,1]
df <- df[,-1]


hc.fit = hclust(dist(scale(df)),method = "complete")
plot(hc.fit) # dendrogram

res = cutree(hc.fit, h = 10) # We can choose 10 as our height to cut

res = as.data.frame(res)
res$region <- rownames(res)
world <- map_data("world")
loc = world[world$region%in% rownames(res),] %>% group_by(region) %>%
summarise_at(vars("long", "lat"),list(name = mean))
res$res = as.factor(res$res)
res$Life.expe = df[,"Life.expectancy_name"]

res = merge(x=res,y=loc,by="region",all.x=TRUE)

# replace the long and lati manually for regions contain missing in
these 2 variables
res[res$region == "Trinidad␣and␣Tobago",]$long_name = -61.311951
res[res$region == "Trinidad␣and␣Tobago",]$lat_name = 10.536421

res[res$region == "Syrian␣Arab␣Republic",]$long_name = 36.278336
res[res$region == "Syrian␣Arab␣Republic",]$lat_name = 33.510414

res[res$region == "Russian␣Federation",]$long_name = 37.618423
~~~~

```
res[res$region == "Russian␣Federation",]$lat_name = 55.751244

res[res$region == "Cabo␣Verde",]$long_name = -22.807835
res[res$region == "Cabo␣Verde",]$lat_name = 16.095011

# We choose the median for robust the outliers.

ggplot(res, aes(x=res, y=Life.expe, color=res)) +
  geom_boxplot()

life_median = res %>%group_by(res) %>% summarise_at(vars(Life.expe),
list(name = median))
life_median # Table 1

# World map

# choose the representative for each cluster
res %>% filter(Life.expe %in% (life_median$name))
res[res$res == 1,][which(abs(res[res$res == 1,]$Life.expe -
life_median$name[1])== min(abs(res[res$res == 1,]$Life.expe -
life_median$name[1]))),]
res[res$res == 3,][which(abs(res[res$res == 3,]$Life.expe
- life_median$name[3]) == min(abs(res[res$res == 3,]$Life.expe -
life_median$name[3]))),]


ggplot() +
  geom_map(
    data = world, map = world,
    aes(long, lat, map_id = region),
    color = "white", fill = "lightgray", size = 0.1
  ) +


  geom_point(
    data = res,
    aes(long_name, lat_name,
        color = res,
        size=Life.expe),
    alpha = 0.5
  ) +
  #labs(x = NULL, y = NULL, color = NULL)+
  theme_void() +
  theme(legend.position = "none")+
  labs(title="Pattern of Life expectancy in the world") +
  geom_text(data=res[res$region %in% c("China", "India", "Belgium",
  "Indonesia", "Chad"),], aes(long_name, lat_name, label=region))
```

‘‘‘