Group Project

https://sancaa94.medium.com/life-expectancy-analysis-c93ccc77474f

https://evidencen.com/predictlifeexpectancy/

▼ Data Overview

We have 2938 observations for the whole data set. For year 2014, we have 183 observations with 22 variables.

[1] "Country" "Year" "Status"
[4] "Life.expectancy" "Adult.Mortality" "infant.deaths"
[7] "Alcohol" "percentage.expenditure" "Hepatitis.B"
[10] "Measles" "BMI" "under.five.deaths"
[13] "Polio" "Total.expenditure" "Diphtheria"
[16] "HIV.AIDS" "GDP" "Population"
[19] "thinness..1.19.years" "thinness.5.9.years" "Income.composition.of.resources"
[22] "Schooling"

Source:

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Research Question #1

Research Question #1 will be an inference question. Our group will explore Life Expectancy in the year 2014 as a function of these possible variables:

- *Percentage Expenditure*
- *Hepatitis B*
- *Measels*
- *Polio*
- *HIV/AIDS*
- *Diptheria*
- *GDP*
- *Schooling*
- *Income Composition of Resources*
- *Country Status (Developed/Developing)*
- *Population*

The specific question we will be answer is: “How did major disease, economic, and social factors impact life expectancy around the globe in 2014?”

Research Question #2

Research Question #2 will be predictive. Our group will seek to predict a coutry’s category as developed or developing based on the best-fit model.

The specific question we will answer is: “Can observations of life expectancy-related data such as mortality rates and disease statistics, along with country-specific properties, be used to accurately predict a country’s status as developed or developing?”

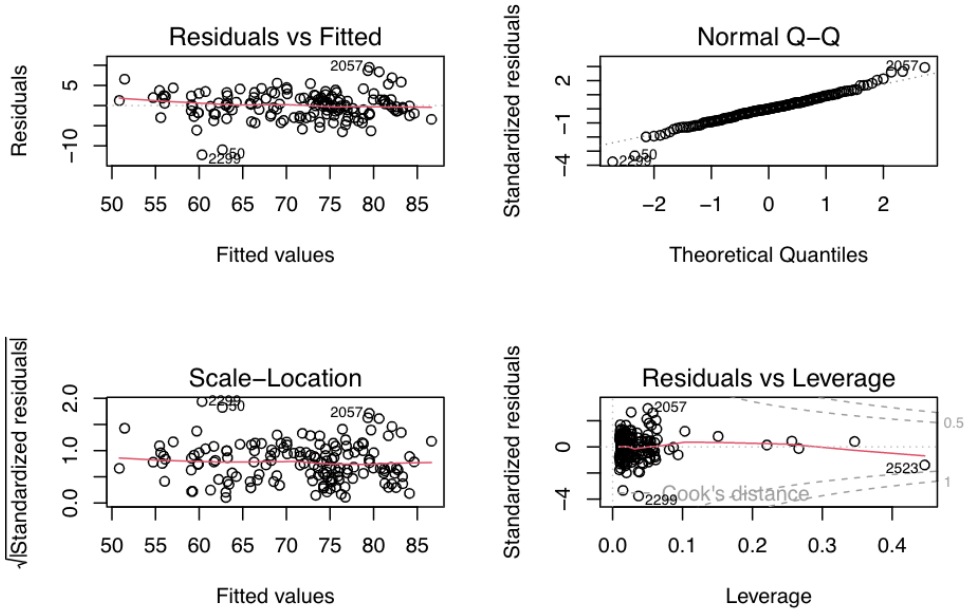
▼ Primary relationship of interest

For year 2014, using GDP, percentage.expenditure, HIV.AIDS rate (number of reported cases per 1000 population), Income Composition of Resources (Income composition is the rate at which society has a higher, middle, or lower-income class [Income composition of resources have the highest correlation coefficient of 0.91 which means that **if a country utilizes its resources productively, it is more likely to see its citizens live longer than expected**
. Human Development Index in terms of income composition of resources (index ranging from 0 to 1)]), and country status (developing or developed) to infer **Life expectancy**

```
```{r, echo=FALSE, results="asis", header=FALSE, message=FALSE, warning=FALSE, cache=TRUE}
fit4 <- lm(Life.expectancy ~ GDP+percentage.expenditure+HIV.AIDS+Income.composition.of.resources+Status, data=le14)
stargazer(fit4,header=FALSE, type="latex",
          no.space = TRUE,
          report = ('vcsp*'), single.row = TRUE,
          column.sep.width = "0.2pt",
          font.size = "small",
          title="Regression Summary")
par(mfrow = c(2, 2))
plot(fit4)
```
```

Table 4: Regression Summary

	Dependent variable:
	Life.expectancy
GDP	−0.00005 (0.00004)
	p = 0.206
percentage.expenditure	0.0004 (0.0002)
	p = 0.145
HIV.AIDS	−1.392 (0.211)
	p = 0.000***
Income.composition.of.resources	42.536 (2.551)
	p = 0.000***
StatusDeveloping	−1.006 (0.880)
	p = 0.256
Constant	44.053 (2.229)
	p = 0.000***
Observations	154
R^2	0.862
Adjusted R^2	0.857
Residual Std. Error	3.315 (df = 148)
F Statistic	184.448*** (df = 5; 148)
Note:	*p<0.1; **p<0.05; ***p<0.01



▼ Other characteristics

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

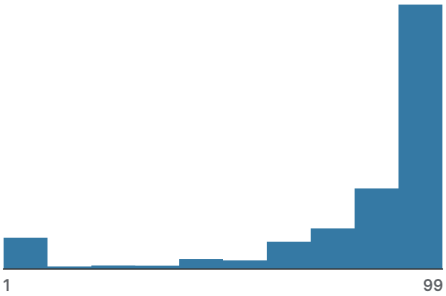
▼ Potential challenges

If we want to include "percentage expenditure" variable, then we need to be cautious for missing values, the data set using “0” as “Sentinel Value”.

Hepatitis B (HepB) immunization coverage data has some missing values

Hepatitis B

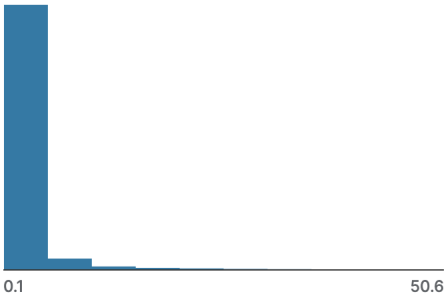
Hepatitis B (HepB) immunization coverage among 1-year-olds (%)



Valid	2385	81%
Mismatched	0	0%
Missing	553	19%
Mean	80.9	
Std. Deviation	25.1	
Quantiles	1	Min
	77	25%
	92	50%
	97	75%
	99	Max

HIV/AIDS

Deaths per 1 000 live births HIV/AIDS (0-4 years)



Valid	2938	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.74	
Std. Deviation	5.08	
Quantiles	0.1	Min
	0.1	25%
	0.1	50%
	0.8	75%
	50.6	Max

Sample size is not very big, we may want to control our variable list.