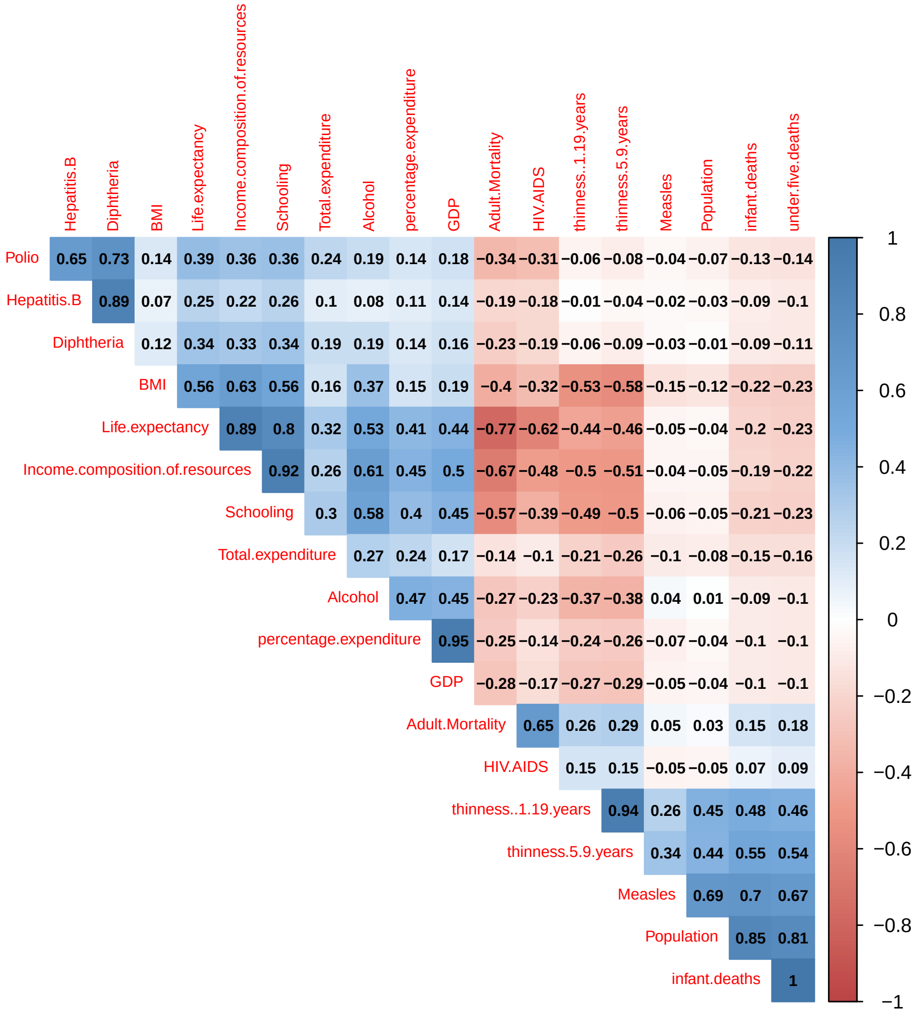
# Primary relationship of interest

Team Orange

2022-10-19

## Primary relationship of interest

Full correlation map (Maybe in appendix):



| | Hepatitis.B | Diphtheria | BMI | Life.expectancy | Income.composition.of.resources | Schooling | Total.expenditure | Alcohol | percentage.expenditure | GDP | Adult.Mortality | HIV.AIDS | thinness..1.19.years | thinness.5.9.years | Measles | Population | infant.deaths | under.five.deaths |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Polio | 0.65 | 0.73 | 0.14 | 0.39 | 0.36 | 0.36 | 0.24 | 0.19 | 0.14 | 0.18 | −0.34 | −0.31 | −0.06 | −0.08 | −0.04 | −0.07 | −0.13 | −0.14 |
| Hepatitis.B | | 0.89 | 0.07 | 0.25 | 0.22 | 0.26 | 0.1 | 0.08 | 0.11 | 0.14 | −0.19 | −0.18 | −0.01 | −0.04 | −0.02 | −0.03 | −0.09 | −0.1 |
| Diphtheria | | | 0.12 | 0.34 | 0.33 | 0.34 | 0.19 | 0.19 | 0.14 | 0.16 | −0.23 | −0.19 | −0.06 | −0.09 | −0.03 | −0.01 | −0.09 | −0.11 |
| BMI | | | | 0.56 | 0.63 | 0.56 | 0.16 | 0.37 | 0.15 | 0.19 | −0.4 | −0.32 | −0.53 | −0.58 | −0.15 | −0.12 | −0.22 | −0.23 |
| Life.expectancy | | | | | 0.89 | 0.8 | 0.32 | 0.53 | 0.41 | 0.44 | −0.77 | −0.62 | −0.44 | −0.46 | −0.05 | −0.04 | −0.2 | −0.23 |
| Income.composition.of.resources | | | | | | 0.92 | 0.26 | 0.61 | 0.45 | 0.5 | −0.67 | −0.48 | −0.5 | −0.51 | −0.04 | −0.05 | −0.19 | −0.22 |
| Schooling | | | | | | | 0.3 | 0.58 | 0.4 | 0.45 | −0.57 | −0.39 | −0.49 | −0.5 | −0.06 | −0.05 | −0.21 | −0.23 |
| Total.expenditure | | | | | | | | 0.27 | 0.24 | 0.17 | −0.14 | −0.1 | −0.21 | −0.26 | −0.1 | −0.08 | −0.15 | −0.16 |
| Alcohol | | | | | | | | | 0.47 | 0.45 | −0.27 | −0.23 | −0.37 | −0.38 | 0.04 | 0.01 | −0.09 | −0.1 |
| percentage.expenditure | | | | | | | | | | 0.95 | −0.25 | −0.14 | −0.24 | −0.26 | −0.07 | −0.04 | −0.1 | −0.1 |
| GDP | | | | | | | | | | | −0.28 | −0.17 | −0.27 | −0.29 | −0.05 | −0.04 | −0.1 | −0.1 |
| Adult.Mortality | | | | | | | | | | | | 0.65 | 0.26 | 0.29 | 0.05 | 0.03 | 0.15 | 0.18 |
| HIV.AIDS | | | | | | | | | | | | | 0.15 | 0.15 | −0.05 | −0.05 | 0.07 | 0.09 |
| thinness..1.19.years | | | | | | | | | | | | | | 0.94 | 0.26 | 0.45 | 0.48 | 0.46 |
| thinness.5.9.years | | | | | | | | | | | | | | | 0.34 | 0.44 | 0.55 | 0.54 |
| Measles | | | | | | | | | | | | | | | | 0.69 | 0.7 | 0.67 |
| Population | | | | | | | | | | | | | | | | | 0.85 | 0.81 |
| infant.deaths | | | | | | | | | | | | | | | | | | 1 |

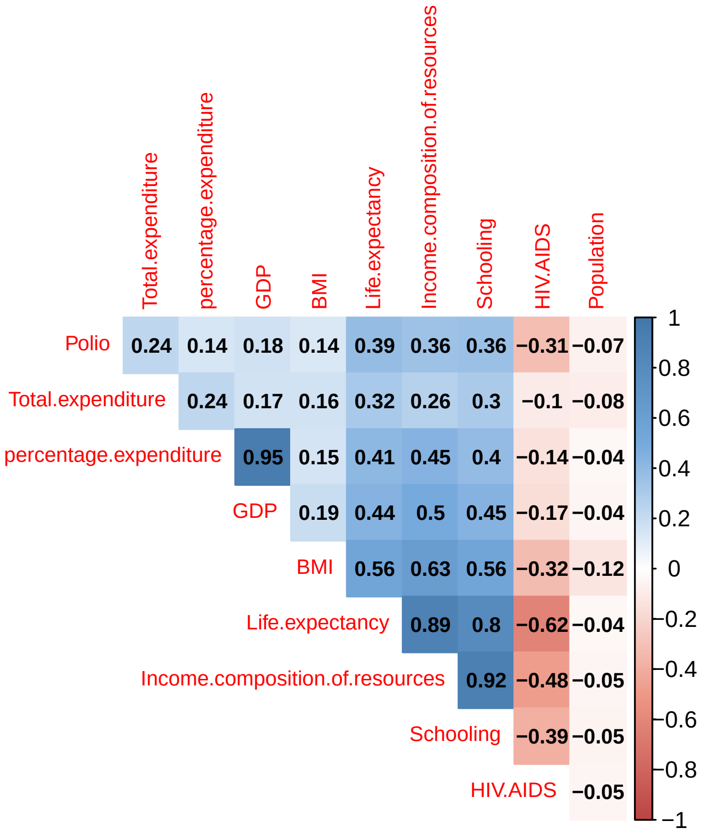
**After piror selection:**

When building a model, it is not wise use all variables when they are highly correlated with each other. To represent immunization coverage, among “Hepatitis.B”,“Polio”, “Diphtheria tetanus toxoid and pertussis (DTP3)”, we decide to use “polio”, since it has the highest correlation between Life Expectancy. Similarly, we see extremely high correlation between GDP and percentage expenditure, but both of them would have meaningful interpretation, we may want to decide which one goes to our final model when we conduct model selection.

On the one hand, with domain knowledge, we know “Adult.Mortality”, “infant.deaths” and “under.five.deaths” variables are directly correlated to Life Expectancy, we choose to drop them from the predictor variable list; on the other hand, we are interested in “HIV.AIDS” variable (Deaths per 1 000 live births HIV/AIDS (0-4 years), . . . . (please add some reasons here)

We would like to omit the variables have Low correlation between Life Expectancy: “Measles”; however, we do want to include “population” because of our interest.

As for categorical variable, we would like to keep country status (developing/developed) as one of the predictors



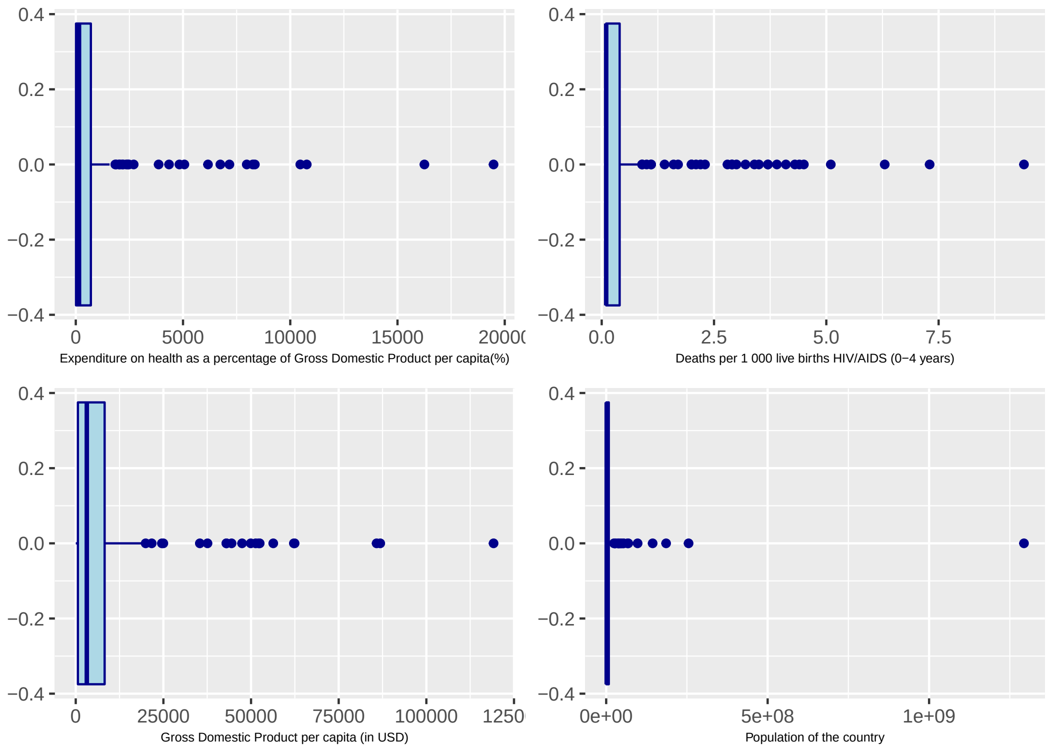
**Potential predictor variables:**

- *Percentage Expenditure* or *GDP*
- *BMI*
- *Polio*
- *HIV/AIDS*
- *Total.expenditure*

- *Schooling*
- *Income Composition of Resources*
- *Country Status (Developed/Developing)*
- *Population*

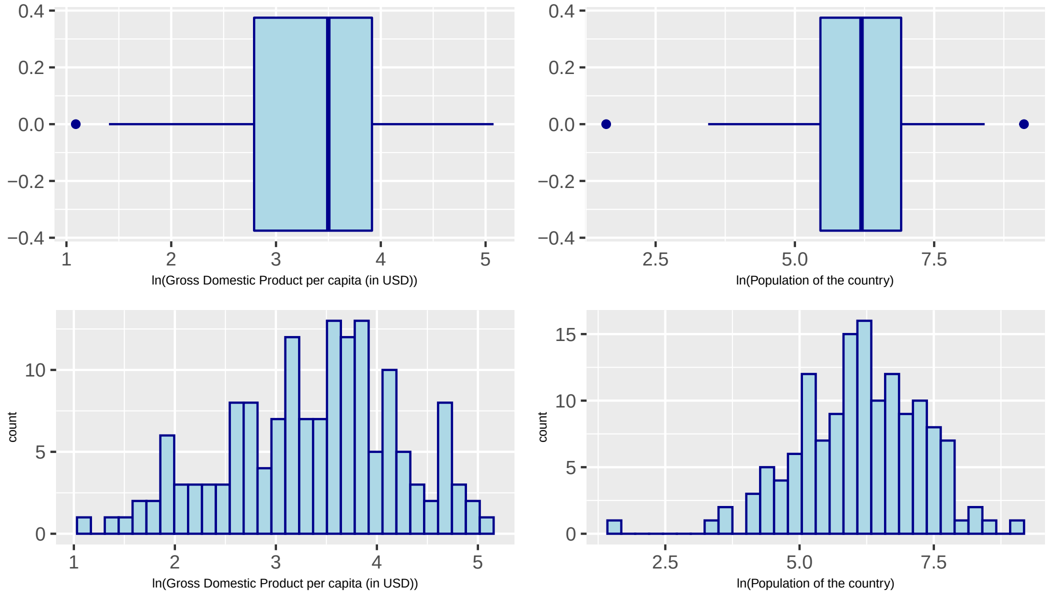
### Transformation if needed when modeling

When checking box plots of all variables, we find 4 variables may need some transformation.



We may want to use log transformation for population and GDP, since the magnitude of gaps are huge. But for other two, we need more investigation due to the difficulty of interpretation.

The box plots and histograms of population and GDP after log transformation:

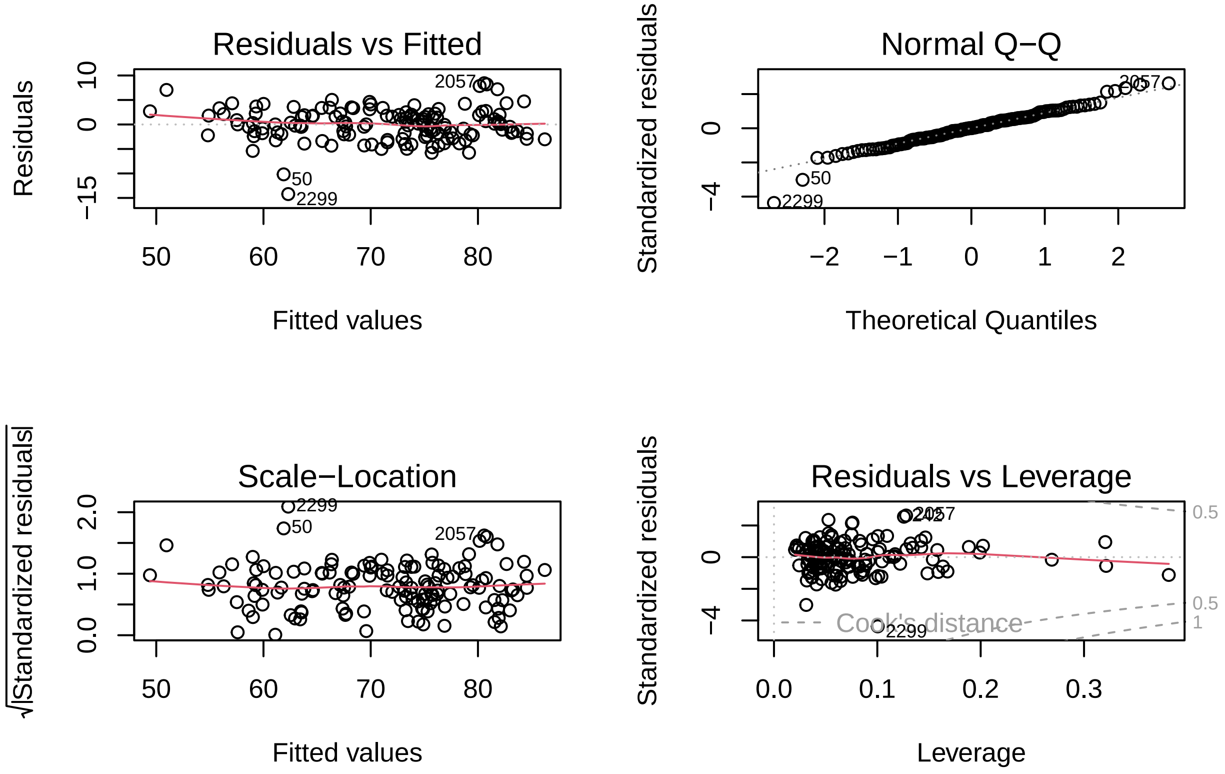


(For our reference)

After selection Model:

Table 1: Regression Summary

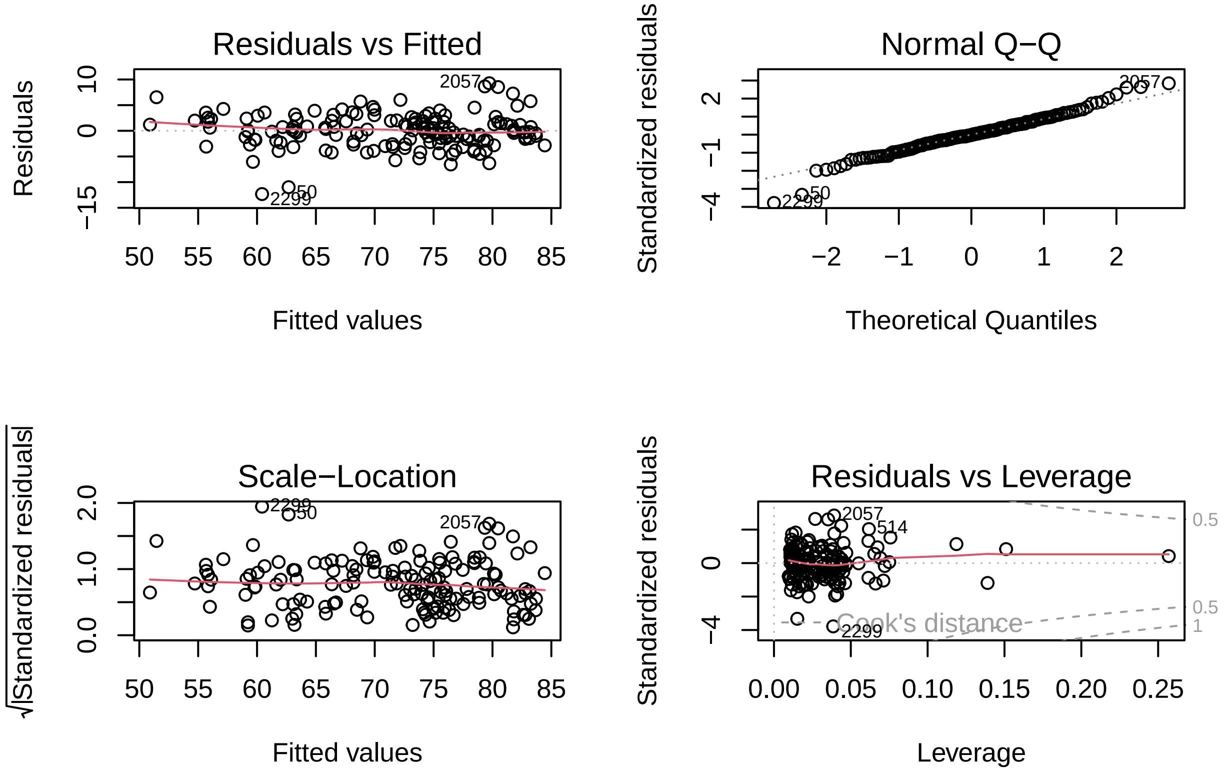
| | *Dependent variable:* |
|---|---|
| | Life.expectancy |
| BMI | −0.002 (0.017) |
| | p = 0.925 |
| log10(GDP) | −0.219 (0.521) |
| | p = 0.675 |
| percentage.expenditure | 0.0001 (0.0001) |
| | p = 0.463 |
| Polio | 0.009 (0.015) |
| | p = 0.557 |
| HIV.AIDS | −1.339 (0.228) |
| | p = 0.00000$^{***}$ |
| Total.expenditure | 0.273 (0.120) |
| | p = 0.025$^{**}$ |
| log10(Population) | −0.293 (0.262) |
| | p = 0.267 |
| Income.composition.of.resources | 43.377 (5.893) |
| | p = 0.000$^{***}$ |
| StatusDeveloping | −0.560 (1.010) |
| | p = 0.580 |
| Schooling | −0.069 (0.271) |
| | p = 0.800 |
| Constant | 44.058 (3.245) |
| | p = 0.000$^{***}$ |
| Observations | 139 |
| $R^2$ | 0.863 |
| Adjusted $R^2$ | 0.852 |
| Residual Std. Error | 3.428 (df = 128) |
| F Statistic | 80.544$^{***}$ (df = 10; 128) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |



Potential Final Model :

Table 2: Regression Summary

| | *Dependent variable:* |
|---|---|
| | Life.expectancy |
| log10(GDP) | −0.201 (0.422) |
| | p = 0.634 |
| HIV.AIDS | −1.372 (0.214) |
| | p = 0.000$^{***}$ |
| Income.composition.of.resources | 42.946 (2.947) |
| | p = 0.000$^{***}$ |
| StatusDeveloping | −1.259 (0.853) |
| | p = 0.143 |
| Constant | 44.597 (2.177) |
| | p = 0.000$^{***}$ |
| Observations | 154 |
| $R^2$ | 0.860 |
| Adjusted $R^2$ | 0.856 |
| Residual Std. Error | 3.326 (df = 149) |
| F Statistic | 228.657$^{***}$ (df = 4; 149) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |


Residuals vs Fitted
Residuals
10
0
−15
50
55
60
65
70
75
80
85
Fitted values
2057
50
2299
Normal Q−Q
Standardized residuals
2
−1
−4
−2
−1
0
1
2
Theoretical Quantiles
2057
50
2299
Scale−Location
√|Standardized residuals|
2.0
1.0
0.0
50
55
60
65
70
75
80
85
Fitted values
2299
50
2057
Residuals vs Leverage
Standardized residuals
0
−4
0.00
0.05
0.10
0.15
0.20
0.25
Leverage
2057
514
2299
Cook's distance
0.5
0.5
1