Project Submission #1

Pooja Kabber, Dingkun Yang, Echo Chen, Andrew Kroening

October 21st, 2022

Data Overview

The dataset used for this research is from the World Health Organization (WHO). The WHO compiles data on thousands of variables for as many countries as feasible and presents them for analysis. This particular dataset contains national-level observations of variables related to life expectancy around the globe for a period spanning the early portion of the 21st century.

Overall Characteristics The complete dataset contains observations beginning in the year 2000 and ending in the year 2015. As a complete dataset, there are 2,938 observations for 22 variables. In practical terms, each country has approximately one observation, averaging 183 for each of the 16 years encompassed by the data. The dataset effectively contains 20 variables for each country and year combination, covering major disease, economic, and social factors. The variables are fully described in **Appendix C, Table XXXXX**.

Sample Dataset Characteristics In the context of this analysis, the dataset will be reduced to include only the year 2014. This yields once observation for 183 countries across all of the variables.

Research Questions:

- How did major disease, economic, and social factors impact life expectancy around the globe in 2014?"
- How did disease and mortality rates, along with national economic factors, contribute to a country's development status in 2014?

Primary Relationship of Interest

As mentioned, the data for this analysis will be truncated to 2014 and reduced to certain variables based on their potential relevance to the research questions. 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 opted to use "polio", since it has the highest correlation with Life Expectancy. Similarly, we see extremely high correlation between GDP and percentage expenditure. However, since both of them could have meaningful interpretations and the categories are distinct, we may want to decide which one goes to our final model when we conduct model selection.

For fairly obvious reasons, we know the "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 the "HIV.AIDS" variable (Deaths per 1,000 live births HIV/AIDS (0-4 years). The rationale for this decision is that the mortality and death variables are measuring the fatalities of entire groups of the population, whereas the "HIV.AIDS" variable is a single measurement that might encompass several underlying societal indicators.

We will omit the variables that have low correlation with 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. In doing so, the dataset is reduced in size to 183 observations of the following variables:

> 'Country', 'Status', 'Population', 'Life.expectancy', 'percentage.expenditure', 'Hepatitis.B', 'Measles', 'Polio', ' HIV.AIDS', 'Diphtheria', 'GDP', 'Schooling', 'Income.composition.of.resources'

The research team estimates that these variables will lead to the best model fit for the dual-purpose nature of this analysis. Table 1 describes summary information about each of the variables that the team believes have a potential to be included in a final model.

Table 1: Summary of Variables

Statistic	N	Mean	St. Dev.	Min	Max
Life.expectancy	183	71.537	8.561	48.100	89.000
percentage.expenditure	183	1,001.913	2,553.290	0.000	19,479.910
Hepatitis.B	173	83.116	23.357	2	99
Measles	183	1,831.208	8,770.077	0	79,563
Polio	183	84.727	20.869	8	99
Diphtheria	183	84.082	23.033	2	99
HIV.AIDS	183	0.682	1.388	0.100	9.400
GDP	155	10,015.570	18,484.240	12.277	119,172.700
Population	142	21,062,964.000	112,170,596.000	41.000	1,293,859,294.000
Income.composition.of.resources	173	0.688	0.154	0.345	0.945
Schooling	173	12.887	2.912	4.900	20.400

Other Characteristics

Some important variables are called out below:

- Life.expectancy: Continuous variable of the adult mortality rates of both sexes (probability of dying between 15 and 60 years per 1000 population).
- Status: Categorical variable identifies a country as 'Developing' or 'Developed' in the given year.
- GDP: Continuous variable showing Gross Domestic Product per capita in the given year.
- Population: Continuous variable with the total estimated population for the country in a given year.
- percentage.expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita (%).
- Hepatitis.B: Hepatitis B (HepB) immunization coverage among 1-year-olds (%).
- Measles: Continuous variable of the number of reported cases per 1000 population.
- Polio: Polio immunization coverage among 1-year-olds (%).
- HIV.AIDS: Continuous variable of the deaths per 1,000 live births due to HIV/AIDS for 0-4 year olds.
- Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%).
- Schooling: Continuous variable expressing the average number of years of school completed among the country's population.
- Income.composition.of.resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1).

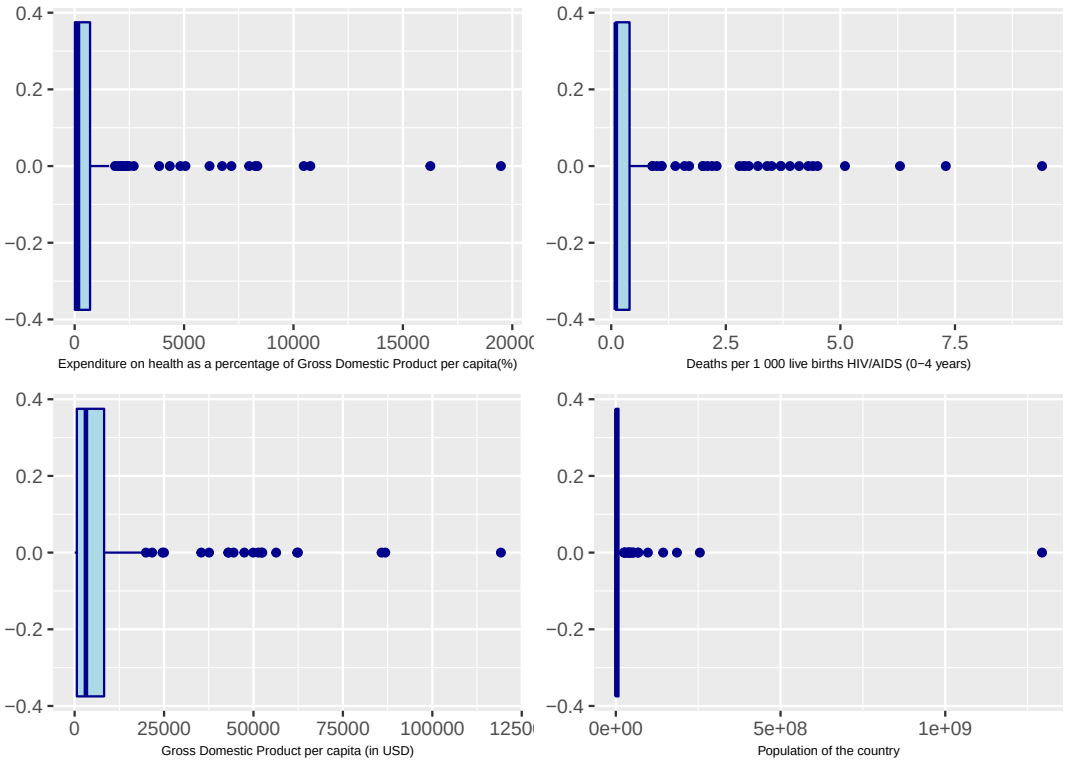
Potential Challenges

Missingness The primary challenge facing the project is missing data. In the 2014 subset of the Life Expectancy dataset, there are 51 observations with a missing value in at least one of the potential predictors:

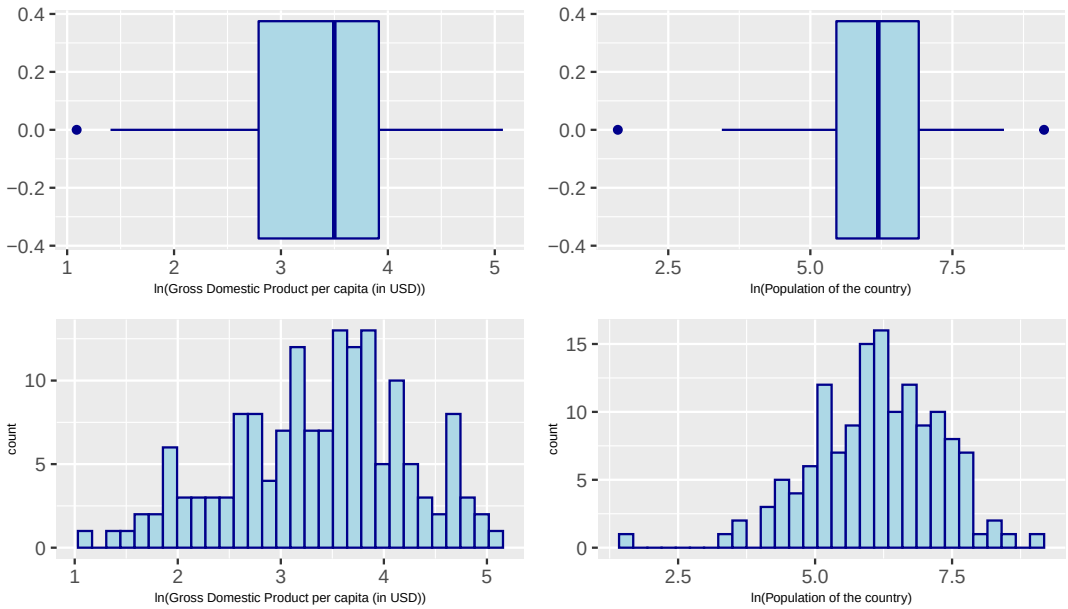
- 41x 'Population' observations
- 10x 'Hep.B' observations
- 28x 'GDP' observations
- 10x 'Schooling' observations
- 10x 'Income.composition.of.resources' observations

The team has proposed three potential avenues for mitigating the effects of these missing values. The observations in question may be dropped or filled with a mean or median value for that variable based on the observed skew.

Skew There is a second challenge posed in some of the variables in the dataset by a skewed distribution. When checking box plots of all variables, we found 4 variables may need transformation before inclusion in the final model.



We may want to use log transformation for population and GDP, because of the magnitude of gaps. But for other two, we need more investigation due to the difficulty of interpretation. The box plots and histograms of the population and GDP variables after log transformation yield a better distribution:

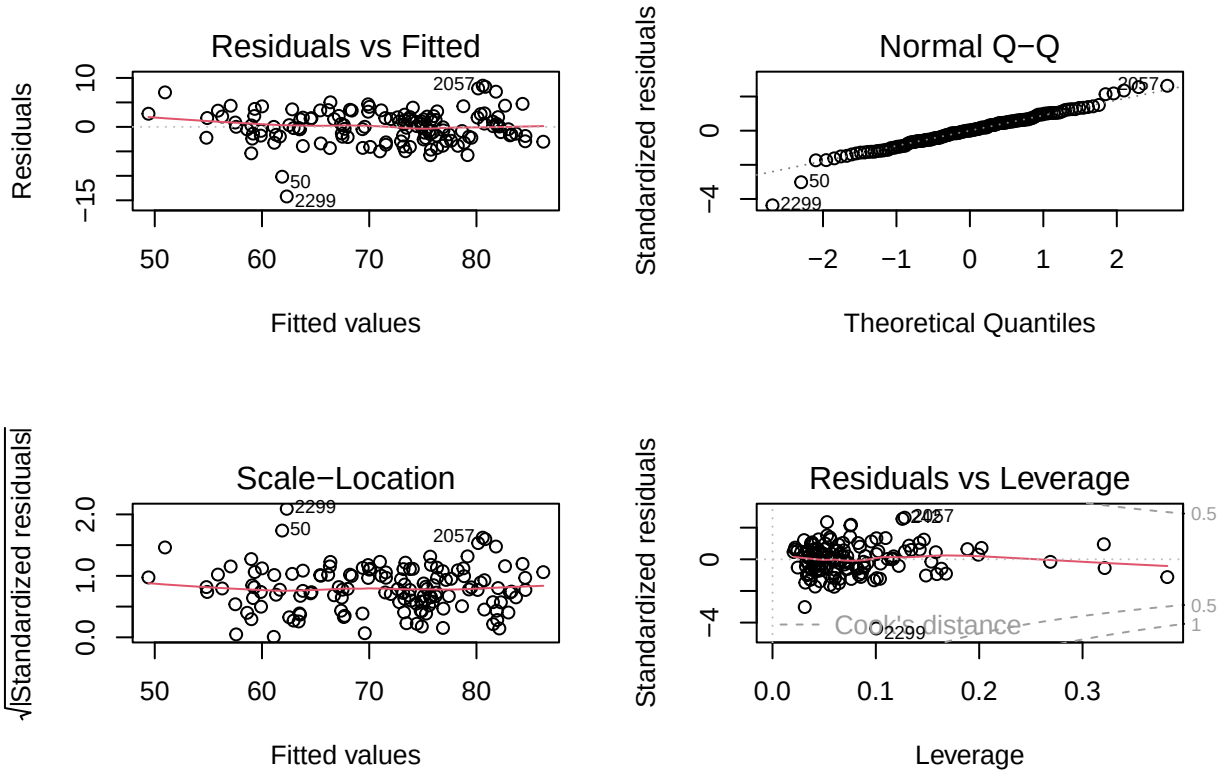

0.4
0.2
0.0
−0.2
−0.4
1
2
3
4
5
ln(Gross Domestic Product per capita (in USD))
2.5
5.0
7.5
ln(Population of the country)
count
10
5
0
15

Appendix A - Regression Models

After selection Model:

Table 2: Regression Summary

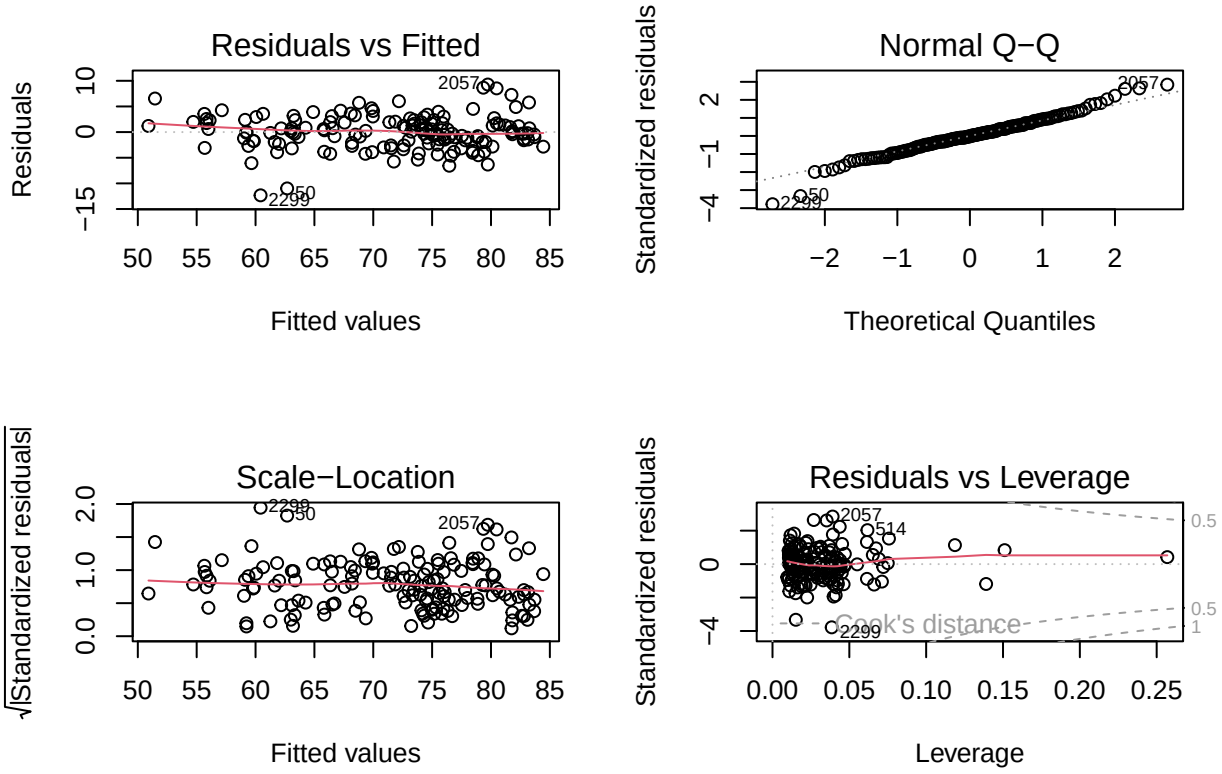
	Dependent variable:
	Life.expectancy
BMI	−0.002 (0.017)
	p = 0.925
log10(GDP)	−0.219 (0.521)
	p = 0.675
percentage.expenditure	0.0001 (0.0001)
	p = 0.463
Polio	0.009 (0.015)
	p = 0.557
HIV.AIDS	−1.339 (0.228)
	p = 0.00000***
Total.expenditure	0.273 (0.120)
	p = 0.025**
log10(Population)	−0.293 (0.262)
	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


Residuals vs Fitted
Residuals
Fitted values
2057
50
2299
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage

Potential Final Model :

Table 3: Regression Summary

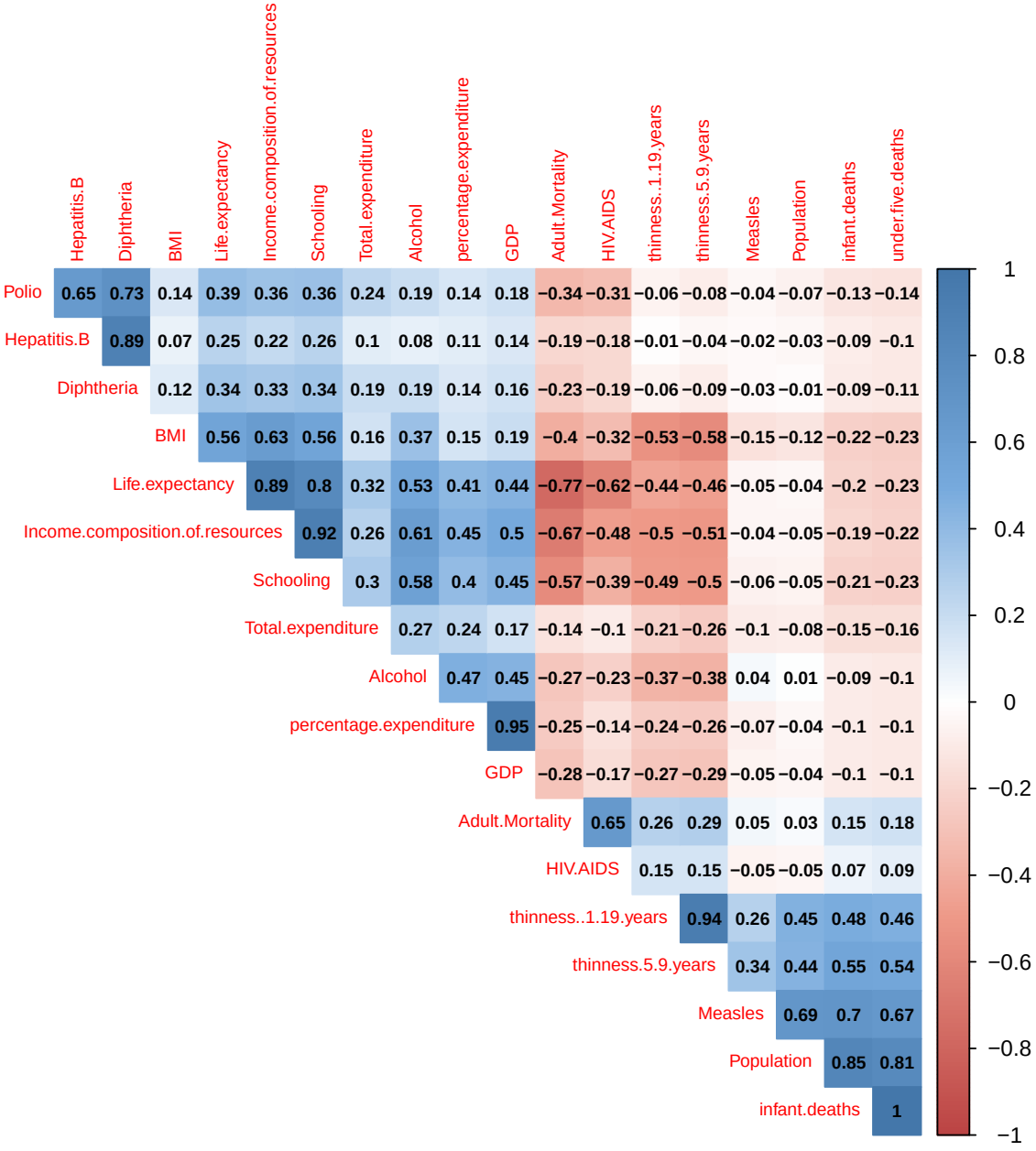
	Dependent variable:
	Life.expectancy
log10(GDP)	−0.201 (0.422)
	p = 0.634
HIV.AIDS	−1.372 (0.214)
	p = 0.000***
Income.composition.of.resources	42.946 (2.947)
	p = 0.000***
StatusDeveloping	−1.259 (0.853)
	p = 0.143
Constant	44.597 (2.177)
	p = 0.000***
Observations	154
R^2	0.860
Adjusted R^2	0.856
Residual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



Appendix B - Supporting Tables and Figures

This appendix is for primary supporting information, tables, or plots that are *valuable for the model that we are keeping.*

"Figure XX - Correlation Matrix"



Appendix C - Variables Description

This appendix is for other information, tables, or plots that are less valuable to the models or will be removed.

Table XXXX - All Variables

Variable	Type	Description
Country	factor	Country name
Year	numeric	Year of the data
Status	factor	Country status of developed or developing
Life_Expectancy	numeric	Life expectancy in age
Adult_Mortality	numeric	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
infant.deaths	numeric	Number of Infant Deaths per 1000 population
Alcohol	numeric	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
percentage.expenditure	numeric	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
Hepatitis.B	numeric	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	numeric	number of reported cases per 1000 population
BMI	numeric	Average Body Mass Index of entire population
under.five.deaths	numeric	Number of under-five deaths per 1000 population
Polio	numeric	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total.expenditure	numeric	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	numeric	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
HIV.AIDS	numeric	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	numeric	Gross Domestic Product per capita (in USD)
Population	numeric	Population of the country
thinness..1.19.years	numeric	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
thinness.5.9.years	numeric	Prevalence of thinness among children for Age 5 to 9(%)
Income.composition.of.resources	numeric	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	numeric	Number of years of Schooling(years)