Introduction to Statistic for Data Science

Group Mini-Project Presentation: Happiness Ladder

ISDS Group 10

Christopher Barrow, Zehao Qian, Mengyuan Zhu, Hithein Augustine

Department of Natural Sciences
Durham University, England, UK

December 3, 2023

Outline

Data Set and its Context

“Happiness Ladder” — Data collected from 137 countries

1. LGPD
(Log GDP)

2. Support
(Positive proportion)

3. HLE
(Healthy life expectancy)

Response variable

4.Freedom
(Positive proportion)

5. Corruption
(Mean of positive proportion)

6. Continent
(Continent name)

Ladder Score: 1000 people in the country rate their happiness between 0 and 10.

Outline

Multiple Linear Regression Model

- **Model Target:** Some socio-economic indexes are used to predict Ladder Score to assist government decision-making.
- **Independent Variables:** LGDP, Support, HLE, Freedom, Corruption, Continent
- **Dependent Variables:** Ladder Score
- **Model Function:**
  $input(LGDP, Support, ...) \Rightarrow output(Ladder\ Score)$
  $LadderScore = \beta_0 + \beta_1 * LGDP + \beta_2 * Support + ... + \epsilon$
- **Optimization Target:** Making the model with the best subset and higher Adjusted R-squared.
  $$\begin{cases} Select\ the\ independent\ variables \\ Update\ \beta\ and\ \epsilon\ to\ minimize\ the\ residual\ sum\ of\ squared \end{cases}$$

Why we choose Multiple Linear Regression Model

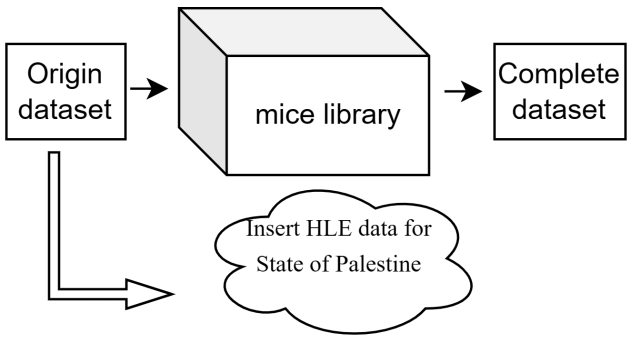
Aspect	Simple Linear Regression	Multiple Linear Regression
Model	$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$
Advantages	Simple and easy interpretation	Captures complex relationships with multiple predictors
	Suitable for examining two-variable relationships	Considers multiple factors, offering a comprehensive view
	Less prone to overfitting with fewer predictors	Analyzes independent effects of each predictor
Disadvantages	Limited to two-variable relationships	More complex, challenging interpretation
	Assumes a linear relationship	Susceptible to multicollinearity with correlated predictors
	May not capture real-world complexity	More assumptions (linearity, independence, normality)
		Risk of overfitting, especially with many predictors

Outline

Constructing the Model – Data Processing

Data Pre-process

Missing Value Handling



Country area error

Fix it with our common sense knowledge

Ecudor → South America

Uganda → Africa

Dataset Segmentation

Continent Mapping

"North America" = 1
"South America" = 2
"Europe" = 3
"Asia" = 4
"Africa" = 5
"Oceania" = 6

Serperate Dataset according to Continent

Our Dataset
Happy_general
Happy_general_continent
Africa
Asia
Europe
...
...

Constructing the Model – Dataset Inspectation

Dataset Name	Country_name	LGDP	Support	HLE	Freedom	Corruption	Continent	Numeric Continent	Ladder_score
Happy_origin	√	√	√	with NA	√	√	√		√
Happy_complete	√	√	√	√	√	√	√		√
Happy_general _continent		√	√	√	√	√	√	√	√
Happy_general		√	√	√	√	√	√		√
Africa		√	√	√	√	√	Africa		√
Asia		√	√	√	√	√	Asia		√
Europe		√	√	√	√	√	Europe		√
North_America		√	√	√	√	√	North_America		√
Oceania		√	√	√	√	√	Oceania		√
South_America		√	√	√	√	√	South_America		√

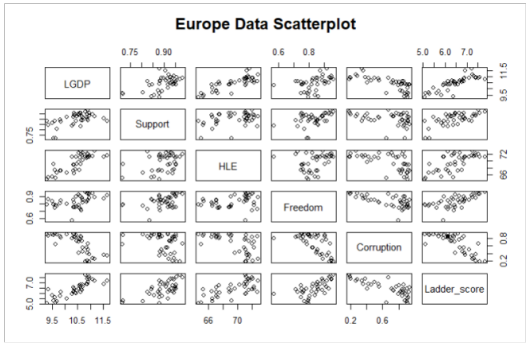
Constructing the Model – Bring dataset to MLR model

Correlationship

	LGDP	Support	HLE	Freedom	Corruption	Ladder_score
LGDP	1.0000000	0.7380688	0.8315083	0.4514389	-0.4369608	0.7843673
Support	0.7380688	1.0000000	0.7260243	0.5416301	-0.2724904	0.8345317
HLE	0.8315083	0.7260243	1.0000000	0.4087733	-0.3996796	0.7416618
Freedom	0.4514389	0.5416301	0.4087733	1.0000000	-0.3837863	0.6629244
Corruption	-0.4369608	-0.2724904	-0.3996796	-0.3837863	1.0000000	-0.4719105
Ladder_score	0.7843673	0.8345317	0.7416618	0.6629244	-0.4719105	1.0000000

	LGDP	Support	HLE	Freedom	Corruption	Ladder_score
LGDP	1.0000000	0.5428825	0.8091949	0.4229237	-0.7147429	0.8224668
Support	0.5428825	1.0000000	0.3615463	0.4133283	-0.3393648	0.6509029
HLE	0.8091949	0.3615463	1.0000000	0.2637575	-0.5905998	0.6732940
Freedom	0.4229237	0.4133283	0.2637575	1.0000000	-0.6113704	0.6156773
Corruption	-0.7147429	-0.3393648	-0.5905998	-0.6113704	1.0000000	-0.7614405
Ladder_score	0.8224668	0.6509029	0.6732940	0.6156773	-0.7614405	1.0000000

Scatterplot



Finding the best subset

Regression Analytics

Outline

Model Results and Analysis

Outline

Conclusions and Insights

Next Steps