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

1. Introduction
    - Intro to Data Set and its Context
2. Statistical Modeling
    - Model Explanation
    - Model Construction
3. Conclusions and Analysis
    - Results and Analysis
    - Conclusions
4. Future Work

Outline

1. Introduction
   - Intro to Data Set and its Context
2. Statistical Modeling
   - Model Explanation
   - Model Construction
3. Conclusions and Analysis
   - Results and Analysis
   - Conclusions
4. Future Work

What is Happiness?

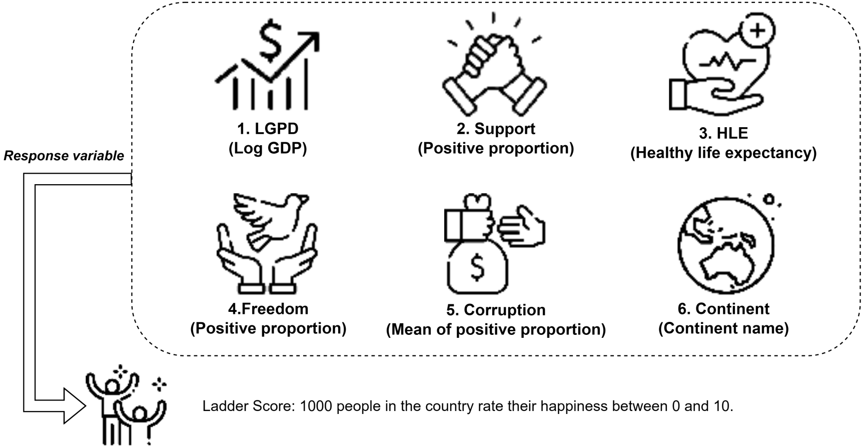
“Happiness Ladder”

The higher a country is on the happiness ladder, the happier, on average, its people tend to be.

Affecting factors?

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)

4.Freedom
(Positive proportion)

5. Corruption
(Mean of positive proportion)

6. Continent
(Continent name)

Response variable

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

Outline

1 Introduction
- Intro to Data Set and its Context

2 Statistical Modeling
- Model Explanation
- Model Construction

3 Conclusions and Analysis
- Results and Analysis
- Conclusions

4 Future Work

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

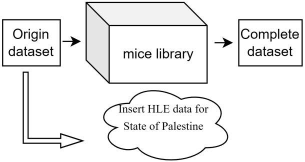
Outline

1. Introduction
   - Intro to Data Set and its Context
2. Statistical Modeling
   - Model Explanation
   - Model Construction
3. Conclusions and Analysis
   - Results and Analysis
   - Conclusions
4. Future Work

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

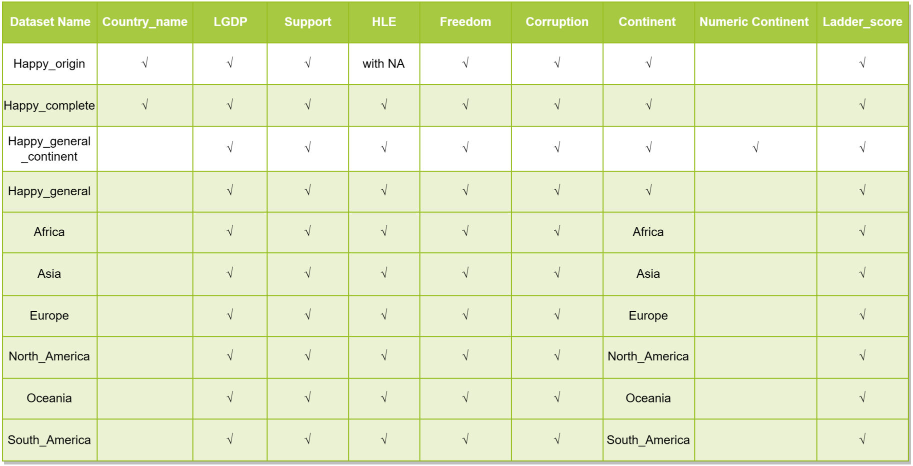
"North America" = 1
"South America" = 2
"Europe" = 3
"Asia" = 4
"Africa" = 5
"Oceania" = 6

Serperate Dataset according to Continent

Our Dataset

- Happy_general
- Happy_general_continent
- Africa
- Asia
- Europe
- ...
- ...

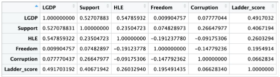
Constructing the Model – Dataset Inspectation



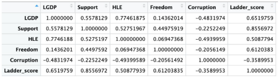
Dataset Name	Country_name	LGDP	Support	HLE	Freedom	Corruption	Continent	Numeric Continent	Ladder_score
Happy_origin	√	√	√	with NA	√	√	√		√
Happy_complete	√	√	√	√	√	√	√		√
Happy_general _continent		√	√	√	√	√	√	√	√
Happy_general		√	√	√	√	√	√		√
Africa		√	√	√	√	√	Africa		√
Asia		√	√	√	√	√	Asia		√
Europe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

Constructing the Model – Bring dataset to MLR model

Step 1: Correlationship

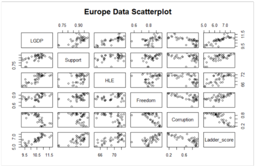


	LGDP	Support	HLE	Freedom	Corruption	Ladder_score
LGDP	1.000000000	0.52707883	0.54785932	0.009904757	0.07777044	0.4917032
Support	0.527078831	1.00000000	0.23504723	0.074828973	0.26647977	0.4067194
HLE	0.547859322	0.23504723	1.00000000	−0.191237780	−0.09175306	0.2603294
Freedom	0.009904757	0.07482897	−0.19123778	1.000000000	−0.14779236	0.1954914
Corruption	0.077770437	0.26647977	−0.09175306	−0.147792362	1.00000000	0.0662834
Ladder_score	0.491703192	0.40671942	0.26032940	0.195491435	0.06628340	1.0000000

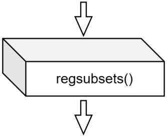


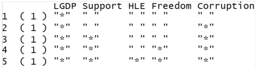
	LGDP	Support	HLE	Freedom	Corruption	Ladder_score
LGDP	1.0000000	0.5578129	0.77461875	0.14362014	−0.4831974	0.6519759
Support	0.5578129	1.0000000	0.52751967	0.44975919	−0.2252249	0.8556972
HLE	0.7746188	0.5275197	1.00000000	0.06947368	−0.4939959	0.5087794
Freedom	0.1436201	0.4497592	0.06947368	1.00000000	−0.2056149	0.6120383
Corruption	−0.4831974	−0.2252249	−0.49399589	−0.20561492	1.0000000	−0.3589953
Ladder_score	0.6519759	0.8556972	0.50877939	0.61203835	−0.3589953	1.0000000

Step 2: Scatterplot



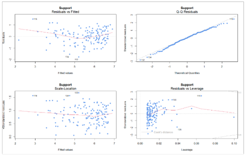
Step 3: Finding the best subset





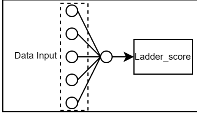
```
         LGDP Support HLE Freedom Corruption
1  ( 1 ) "*"  " "     " " " "     " "
2  ( 1 ) "*"  " "     " " " "     "*"
3  ( 1 ) "*"  "*"     " " " "     "*"
4  ( 1 ) "*"  "*"     " " "*"     "*"
5  ( 1 ) "*"  "*"     "*" "*"     "*"
```

Step 4: Compared with Single Linear Regression



Step 5: Regression Analytics

Calculate the model



Model Evaluation

Variance Inflation Factor Check

Autocorrelation -- Residuals

Adjust R squared

Future Analytics

Interaction between variables

non-linear transformation of predictors

Outline

1. Introduction
   - Intro to Data Set and its Context
2. Statistical Modeling
   - Model Explanation
   - Model Construction
3. Conclusions and Analysis
   - Results and Analysis
   - Conclusions
4. Future Work

Model Results and Analysis

Outline

1. Introduction
   - Intro to Data Set and its Context
2. Statistical Modeling
   - Model Explanation
   - Model Construction
3. Conclusions and Analysis
   - Results and Analysis
   - Conclusions
4. Future Work

Conclusions and Insights

Next Steps