# 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 4, 2023

# Outline

# 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

# 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

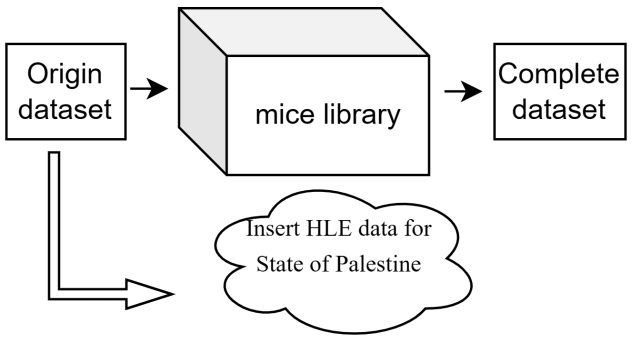
| Aspect | Simple Linear Regression | Multiple Linear Regression |
|---|---|---|
| Model | $Y = \beta_0 + \beta_1 X + \varepsilon$ | $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$ |
| Advantages | Simple and easy interpretation | Captures complex relationships with multiple 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 | | √ |

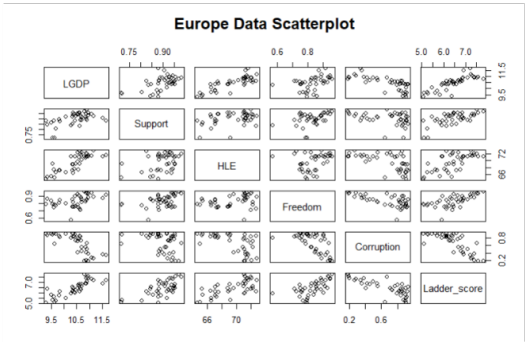
# Constructing the Model – Bring dataset to MLR model

## Step 1: Correlationship

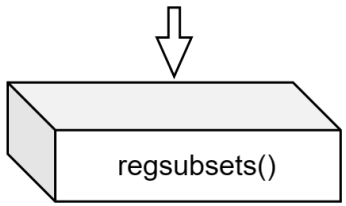
| | LGDP | Support | HLE | Freedom | Corruption | Ladder_score |
|---|---|---|---|---|---|---|
| LGDP | 1.000000000 | 0.52707883 | 0.54785932 | 0.009904757 | 0.07777044 | 0.4917032 |
| Support | 0.527078831 | 1.00000000 | 0.23504723 | 0.074828973 | 0.26647977 | 0.4067194 |
| HLE | 0.547859322 | 0.23504723 | 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

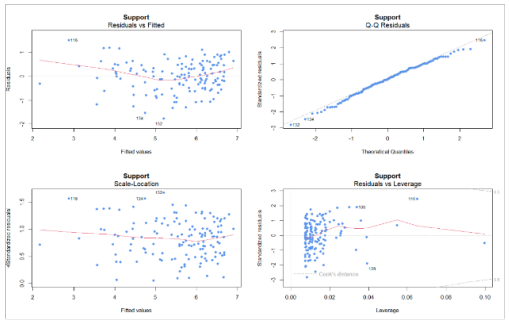



## Step 3: Finding the best subset



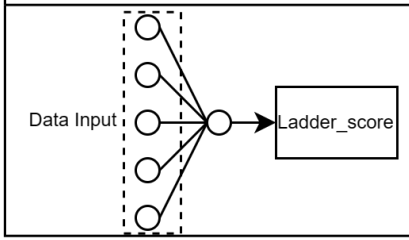

```
  LGDP Support HLE Freedom Corruption
1 ( 1 ) "*"   " "     " " " "  " "
2 ( 1 ) "*"   " "     " " " "  "*"
3 ( 1 ) "*"   "*"     " " " "  "*"
4 ( 1 ) "*"   "*"     " " "*"  "*"
5 ( 1 ) "*"   "*"     "*" "*"  "*"
```

## 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

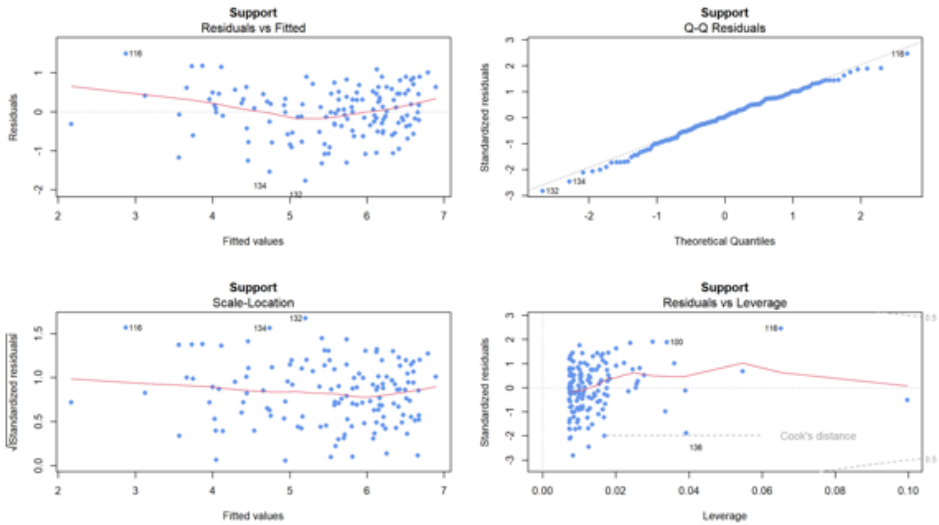
- Interaction between variables
- non-linear transformation of predictors

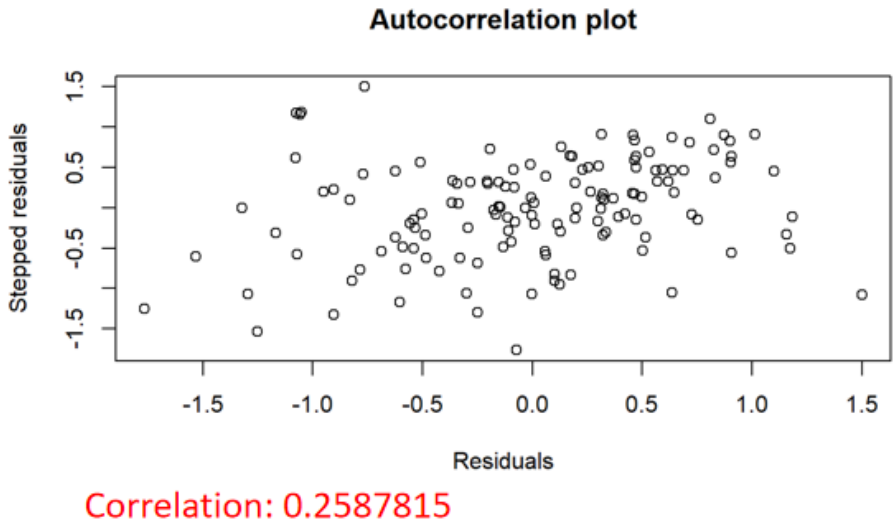
# Outline

# Analysing the model: Simple Linear Regression

## Check Assumptions:

1. Mean = 0
2. Homoskedasticity
3. Independence
4. Normally Distributed



# Analysing the model: Multiple Linear Regression

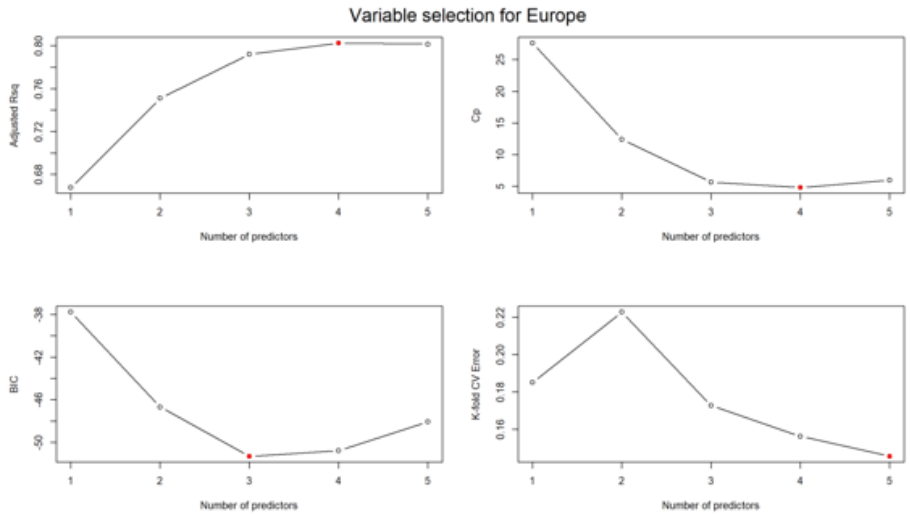
- VIF to check if there is a correlation between predictor variables;
- Example:
  - Africa shows low values
  - Europe shows more of a spread

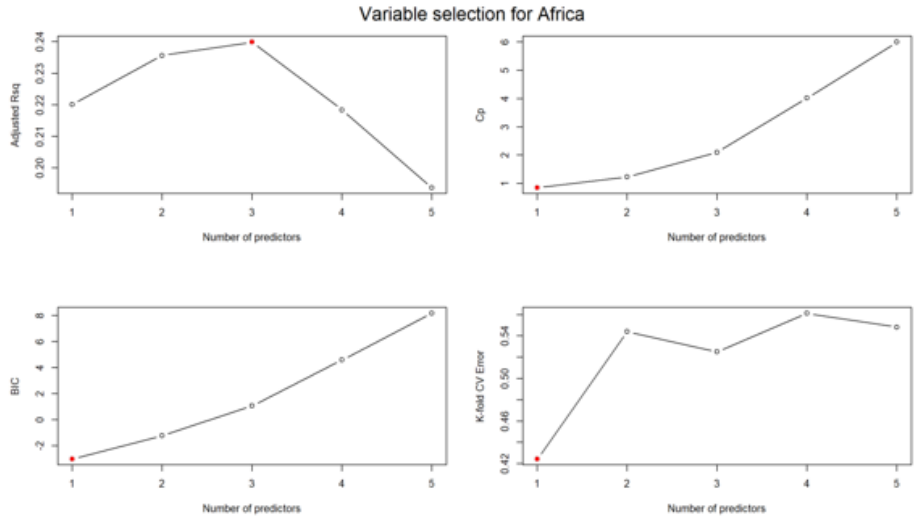
| LGDP | Support | HLE | Freedom | Corruption |
|---|---|---|---|---|
| 1.897157 | 1.514884 | 1.578191 | 1.117253 | 1.162777 |

| LGDP | Support | HLE | Freedom | Corruption |
|---|---|---|---|---|
| 4.777991 | 1.612113 | 3.034882 | 1.823342 | 2.865431 |

# Best Subset Selection

- Europe shows agreement between the range of models.
- Africa shows a larger range of possible values with far more variable graphical shapes.



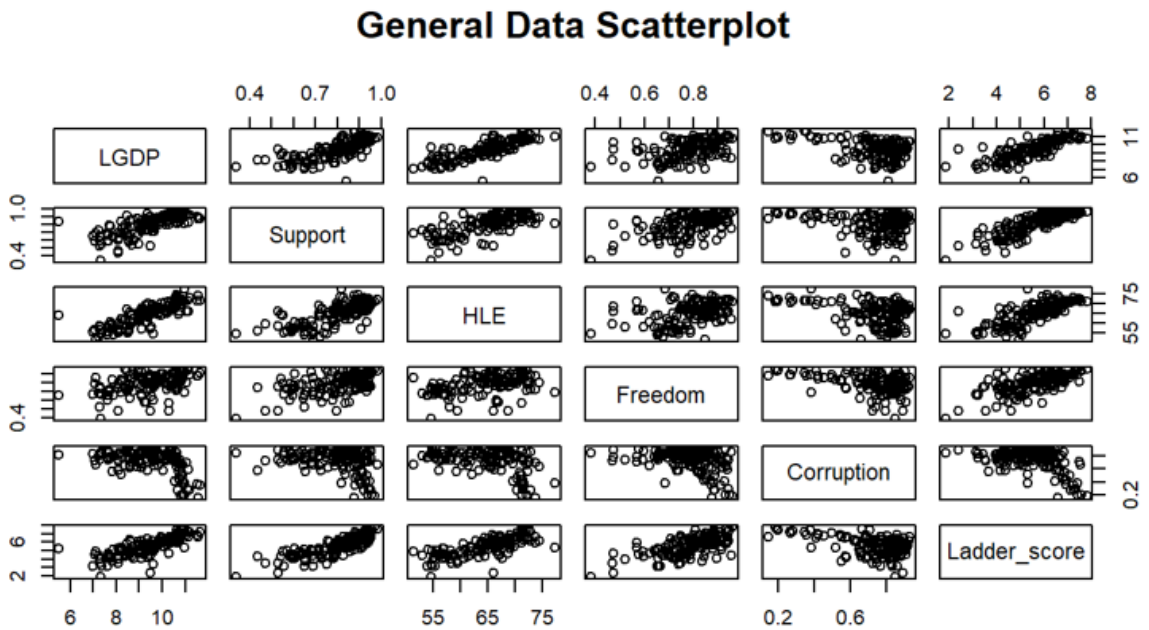
# Outline

# Correlation

Can use this to investigate the correlation between variables and measure if any correlation is evident.

Can see a relationship is evident between *Ladder_score* and all these other predictor variables.

The most notable ones are LGDP and suppport



# Equations

| Dataset | Variables selected | Equation |
|---|---|---|
| Happy_general | LGDP, Support, Freedom, Corruption | $Ladder\_score = -2.90 + 2.96(LGDP) + 7.27(Support) + 5.24(Freedom) - 3.01(Corruption)$ |
| Africa | LGDP | $Ladder\_score = -0.58 + 0.29(LGDP)$ |
| Asia | LGDP, Support, Freedom | $Ladder\_score = -2.18 + 0.39(LGDP) + 5.04(Support) + 2.36(Freedom)$ |
| Europe | LGDP, Support, Freedom, Corruption | $Ladder\_score = -3.70 + 0.40(LGDP) + 3.34(Support) + 1.52(Freedom) - 0.78(Corruption)$ |
| North America | LGDP, Freedom | $Ladder\_score = -1.39 + 0.27(LGDP) + 5.07(Freedom)$ |
| South America | Can't be analysed | *N/A* |
| Oceania | Can't be analysed | *N/A* |

# Government Advice

- By looking at the general, overall dataset clear support is the most important metric and so the variable to focus on improving first.

- However, through best subset selection it is clear that a range of variables all come into this, and so focus on purely support will no yield the best results, although it should be prioritised.

# Advice per Continent

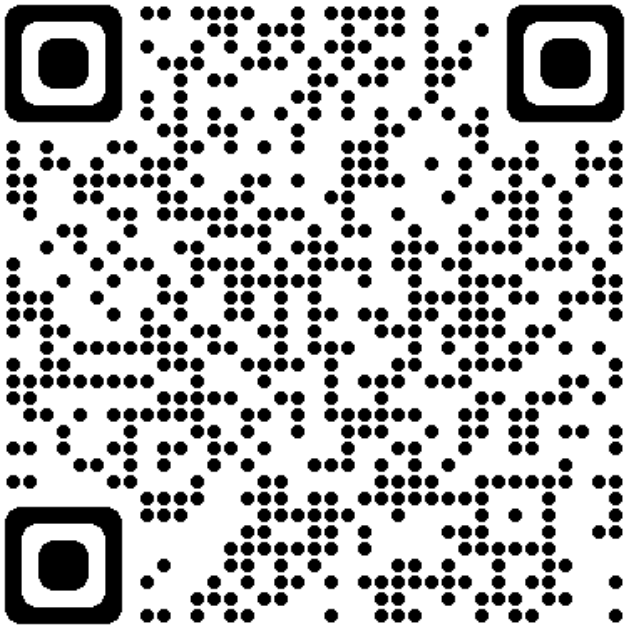
For each continent, there are certain differences that stood out and should be highlighted for governments within them:

1. Africa: LGDP is the most important metric and to such an extent that ladder score can almost be modelled directly off it. Thus to improve happiness, LGDP is key to focus on;
2. Europe: Similar to the general dataset and shows a range of factors will need to be focused on, starting with LGDP;
3. Asia: Focus on all factors, however with a prioritisation of LGDP, Support and Freedom;
4. North America: A focus on LGDP and Freedom is crucial to improving the Happiness extent;
5. Data for South America and Oceania is not sufficient to currently draw conclusions from.

# Next Steps

1. **Improve the accuracy of the model**
   - Interaction between variables (e.g. $HLE \times Support$).
   - Try non-linear transformation of the predictors (e.g. $log(HLE)$, $HLE^2$, $HLE^3$).
2. **Bug fixing**
   - The data for North America was anomalous and independent variables could not be selected by linear regression analysis. Our team will next process anomalous data from the North American dataset.
   - There are only two countries in Oceania, and when they are brought into the analytical model, they report an error and cannot perform linear regression. We will try to combine the Oceania data into other similar continents for analysis.

# Thanks for watching!



ISDS Group 10
December 4, 2023