Prepared by group 15

PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN TUỔI THỌ CỦA CÁC QUỐC GIA BẰNG PHƯƠNG TRÌNH HỒI QUY TUYẾN TÍNH

Team Members

Nguyễn Đức Toàn

23133077

Lương Ngọc Huy

23133028

Nguyễn Đức Thịnh

23133073

Nguyễn Hữu Tâm

23133067

DATASET

Bộ dữ liệu Life Expectancy (WHO) trên Kaggle, do Kumarajarshi chia sẻ, là một nguồn dữ liệu toàn diện về tuổi thọ và các yếu tố ảnh hưởng đến sức khỏe cộng đồng trên toàn cầu (thuộc loại panel data)

- Số lượng quan sát: Hơn 2.000 dòng dữ liệu, mỗi dòng đại diện cho một quốc gia trong một năm cụ thể.
- Thời gian: Dữ liệu từ năm 2000 đến 2015.
- Số lượng quốc gia: Hơn 150 quốc gia.
- Số lượng biến: 22 biến, bao gồm tuổi thọ và các yếu tố kinh tế - xã hội, môi trường và y tế.

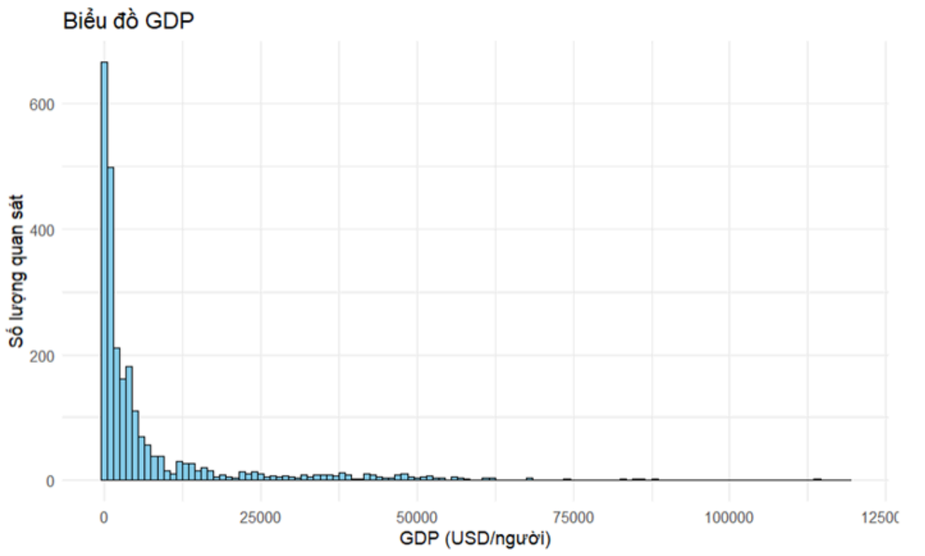
DATASET

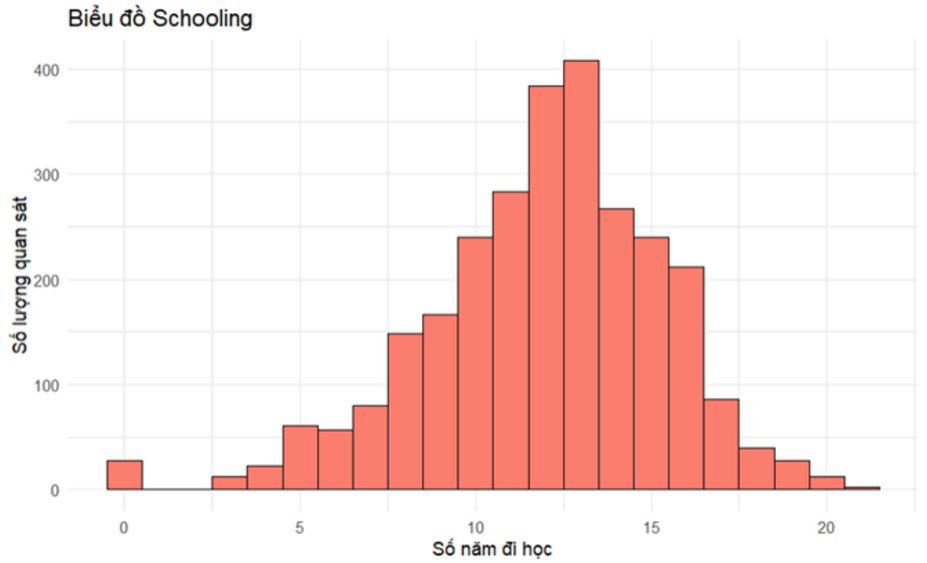
- Tuổi thọ trung bình: Dao động từ 36.3 đến 89 năm, trung bình 69.2 năm → Phản ánh chênh lệch lớn về y tế, kinh tế, điều kiện sống.
- Tử vong:
  - Người lớn: TB 164.8 (1–723) → Chênh lệch rõ rệt giữa các quốc gia.
  - Trẻ dưới 5 tuổi: TB 42.0, cao nhất 2,500 ca → Bất cập trong chăm sóc trẻ em.
- Kinh tế & y tế:
  - GDP: Trung bình 7,483, lệch phải mạnh (1.68–119,172.7).
  - Chi tiêu y tế: TB 738.25, cao nhất 19,479.91 → Đầu tư y tế không đồng đều.
- Tiêm chủng:
  - HepB, Polio, Diphtheria: TB ~80–82%, có quốc gia chỉ 1–3% → Thiếu hụt nghiêm trọng.
- Dinh dưỡng & dịch bệnh:
  - BMI: TB 38.3, có giá trị thấp đến 1.0 → Suy dinh dưỡng hoặc thiếu dữ liệu.
  - HIV/AIDS: TB 1.74%, tối đa 50.6% → Gánh nặng bệnh dịch cao.
- Biến phân loại:
  - Năm (2000–2015) & Tình trạng (phát triển/đang phát triển) → Hỗ trợ phân tích xu hướng & so sánh nhóm quốc gia.

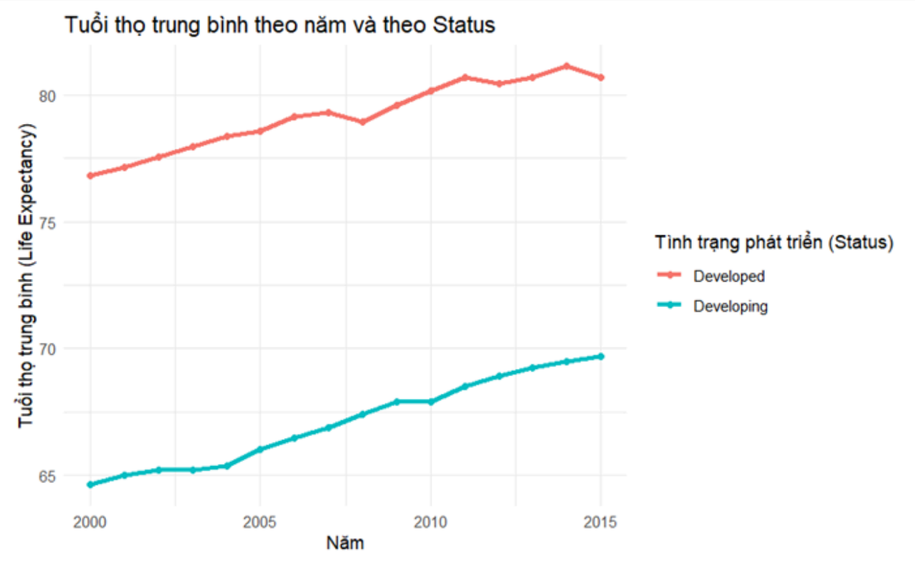
Tổng quan từng cột trong dữ liệu

Column	Min	Max	Mean
Country	NA	NA	NA
Year	2000.00000	2.015000e+03	2.007519e+03
Status	NA	NA	NA
Life.expectancy	36.30000	8.900000e+01	6.922493e+01
Adult.Mortality	1.00000	7.230000e+02	1.647964e+02
infant.deaths	0.00000	1.800000e+03	3.030395e+01
Alcohol	0.01000	1.787000e+01	4.602861e+00
percentage.expenditure	0.00000	1.947991e+04	7.382513e+02
Hepatitis.B	1.00000	9.900000e+01	8.094046e+01
Measles	0.00000	2.121830e+05	2.419592e+03
BMI	1.00000	8.730000e+01	3.832125e+01
under.five.deaths	0.00000	2.500000e+03	4.203574e+01
Polio	3.00000	9.900000e+01	8.255019e+01
Total.expenditure	0.37000	1.760000e+01	5.938190e+00
Diphtheria	2.00000	9.900000e+01	8.232408e+01
HIV.AIDS	0.10000	5.060000e+01	1.742104e+00
GDP	1.68135	1.191727e+05	7.483158e+03
Population	34.00000	1.293859e+09	1.275338e+07
thinness..1.19.years	0.10000	2.770000e+01	4.839704e+00
thinness.5.9.years	0.10000	2.860000e+01	4.870317e+00

DATASET







- GDP: Phân bố lệch phải, phần lớn quốc gia có GDP thấp → Phản ánh chênh lệch kinh tế lớn, ảnh hưởng đến tuổi thọ.
- Số năm đi học: Tập trung quanh 10–15 năm, nhưng vẫn có quốc gia rất thấp → Tác động tiêu cực đến nhận thức và tiếp cận y tế.
- Tuổi thọ theo năm & tình trạng phát triển:
  - Tăng dần từ 2000–2015 ở cả hai nhóm.
  - Nước phát triển luôn có tuổi thọ cao hơn → Do điều kiện sống và y tế tốt hơn.

Xử lý tiền dữ liệu

1. Xử lý dữ liệu thiếu (NA)

Column	NA_Count
Country	0
Year	0
Status	0
Life.expectancy	10
Adult.Mortality	10
infant.deaths	0
Alcohol	194
percentage.expenditure	0
Hepatitis.B	553
Measles	0
BMI	34
under.five.deaths	0
Polio	19
Total.expenditure	226
Diphtheria	19
HIV.AIDS	0
GDP	448
Population	652
thinness..1.19.years	34
thinness.5.9.years	34
Income.composition.of.resources	167
Schooling	163

- Vấn đề: Nhiều biến số chứa giá trị NA → Gây sai lệch nếu giữ nguyên, mất dữ liệu nếu loại bỏ dòng.
- Giải pháp: Thay NA bằng trung vị theo từng quốc gia (theo nhóm Country).
  - Trung vị được tính từ các giá trị không NA trong cùng quốc gia.
  - Tránh ảnh hưởng của outliers, đại diện tốt hơn so với trung bình cộng.
- Ưu điểm:
  - Giữ lại nhiều dữ liệu hơn.
  - Bảo toàn đặc trưng riêng của từng quốc gia.
  - Tạo bộ dữ liệu đầy đủ cho chuẩn hóa, trực quan hóa và mô hình hóa.

Column	NA_Count
Country	0
Year	0
Status	0
Life.expectancy	0
Adult.Mortality	0
infant.deaths	0
Alcohol	0
percentage.expenditure	0
Hepatitis.B	0
Measles	0
BMI	0
under.five.deaths	0
Polio	0
Total.expenditure	0
Diphtheria	0
HIV.AIDS	0
GDP	0
Population	0
thinness..1.19.years	0
thinness.5.9.years	0
Income.composition.of.resources	0
Schooling	0

Xử lý tiền dữ liệu

2. Chuẩn hóa dữ liệu

```
 [1] "Country"                          "Year"                          "Status"
"Life.expectancy"
 [5] "Adult.Mortality"                  "infant.deaths"                 "Alcohol"
"percentage.expenditure"
 [9] "Hepatitis.B"                      "Measles"                       "BMI"
"under.five.deaths"
[13] "Polio"                            "Total.expenditure"             "Diphtheria"
"HIV.AIDS"
[17] "GDP"                              "Population"
"thinness..1.19.years"             "thinness.5.9.years"
[21] "Income.composition.of.resources" "Schooling"
```

- Tên cột: Chuyển toàn bộ về chữ in hoa → Thống nhất, dễ thao tác.
- Tên quốc gia (COUNTRY): Thay khoảng trắng bằng dấu gạch dưới → Phù hợp khi xử lý và trực quan hóa.

Bolivia (Plurinational State of)	→ Bosnia_and_Herzegovina
Bosnia and Herzegovina	

```
 [1] "COUNTRY"              "YEAR"              "STATUS"
 [5] "ADULT.MORTALITY"      "INFANT.DEATHS"     "ALCOHOL"
 [9] "MEASLES"              "BMI"               "UNDER.FIVE.DEATHS"
[13] "HIV.AIDS"             "GDP"               "POPULATION"
[17] "SCHOOLING"            "THINNESS"          "VACCINATION_RATE"
```

- Biến phân loại (STATUS):
  - "Developed" → 1
  - "Developing" → 0

  → Dễ tích hợp vào mô hình định lượng.
- Gộp biến tương đương:
  - THINNESS..1.19.YEARS & THINNESS.5.9.YEARS → THINNESS
  - HEPATITIS.B, POLIO, DIPHTHERIA → VACCINATION_RATE

  → Giảm đa cộng tuyến, đơn giản mô hình.

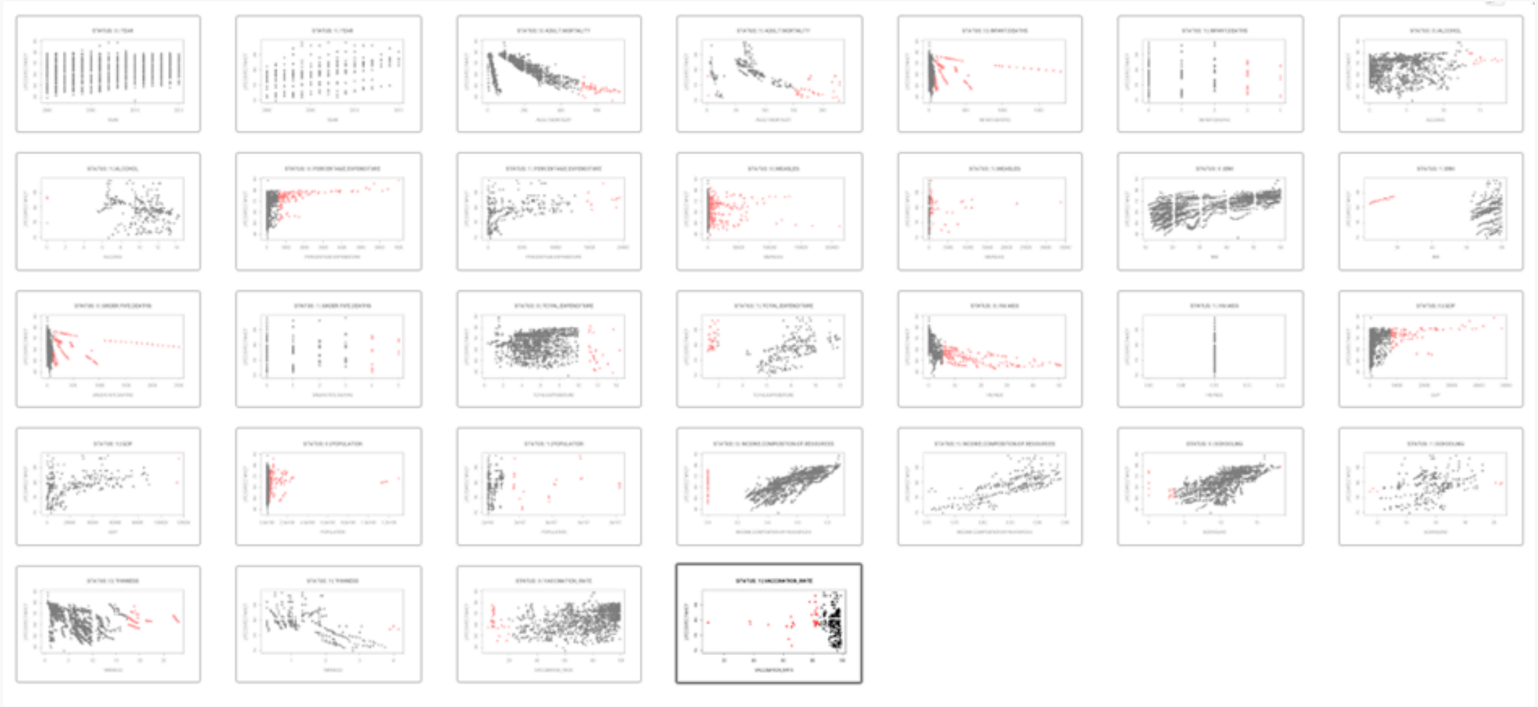
```
"LIFE.EXPECTANCY"
"PERCENTAGE.EXPENDITURE"
"TOTAL.EXPENDITURE"
"INCOME.COMPOSITION.OF.RESOURCES"
```

Xử lý tiền dữ liệu

3. Kiểm tra và xử lý outlier

- Outlier: Giá trị bất thường, có thể do lỗi hoặc hiện tượng đặc biệt → Làm sai lệch kết quả phân tích, đặc biệt với hồi quy tuyến tính.
- Phương pháp IQR (Interquartile Range):
  - Q1 = Phân vị thứ 25%, Q3 = Phân vị thứ 75%
  - IQR = Q3 − Q1
  - Ngưỡng phát hiện:
    - Lower bound = Q1 − 1.5 × IQR
    - Upper bound = Q3 + 1.5 × IQR

  → Giá trị ngoài khoảng này là outlier.
- Ưu điểm: Không bị ảnh hưởng bởi extreme values, phù hợp với dữ liệu lệch.
- Cách thực hiện:
  - Áp dụng riêng cho từng nhóm STATUS (Developed / Developing).
  - Kiểm tra từng biến định lượng (trừ COUNTRY, STATUS, LIFE.EXPECTANCY).
  - Trực quan hóa bằng scatter plot với outlier đánh dấu màu đỏ → Dễ nhận diện ảnh hưởng đến tuổi thọ.



Xử lý tiền dữ liệu

3. Kiểm tra và xử lý outlier

Xử lý Outlier theo từng quốc gia:

- Lý do: Mỗi quốc gia có đặc điểm kinh tế – xã hội khác nhau → Không nên xử lý theo chuẩn chung.
- Quy trình:
  - Nhóm dữ liệu theo COUNTRY.
  - Với mỗi biến định lượng:
    - Tính Q1, Q3, IQR, xác định ngưỡng outlier.
    - Phát hiện các giá trị bất thường (kể cả giá trị bằng 0 không hợp lý).
    - Thay thế outlier bằng trung vị của biến đó trong cùng quốc gia.
- Ưu điểm:
    - Làm sạch dữ liệu hiệu quả.
    - Giữ nguyên đặc trưng quốc gia, đảm bảo phân tích tuổi thọ chính xác.

MÔ HÌNH HỒI QUY TUYẾN TÍNH

Mô hình hồi quy tuyến tính đa biến

Hiện tượng đa cộng tuyến
Overfitting

MÔ HÌNH HỒI QUY TUYẾN TÍNH ĐA BIẾN

Phương pháp OLS (Ordinary Least Squares)

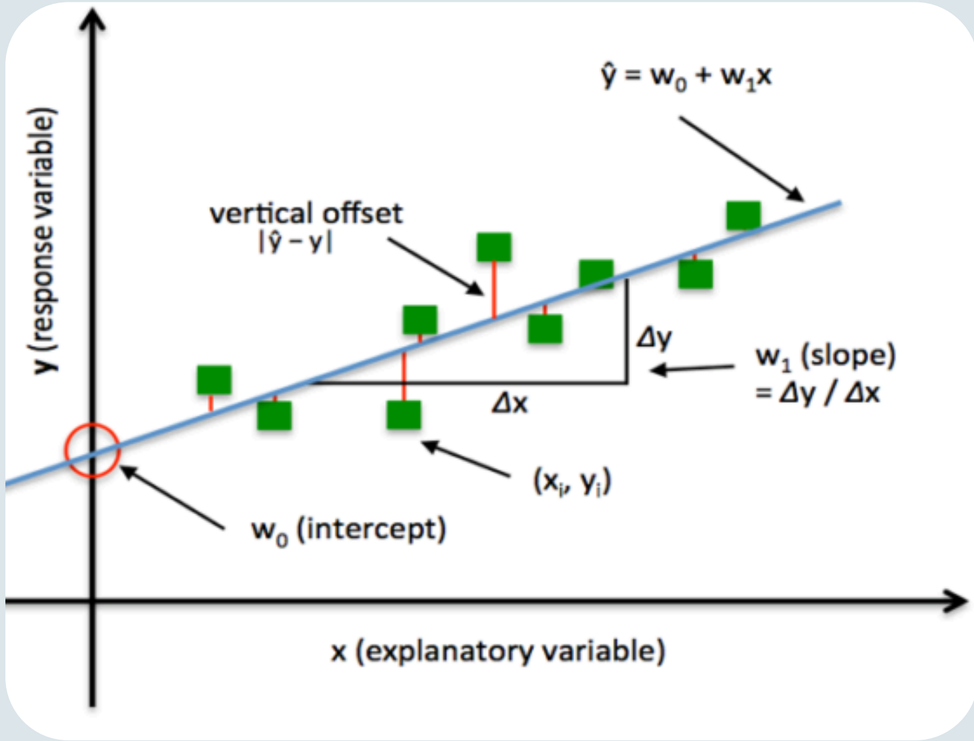
- Mục đích của việc sử dụng phương pháp OLS nhằm **giảm thiểu sai số dự đoán**

- Phương pháp OLS tìm ra các hệ số β sao cho **Tổng bình phương sai số (Residual Sum of Squares - RSS)** là nhỏ nhất:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

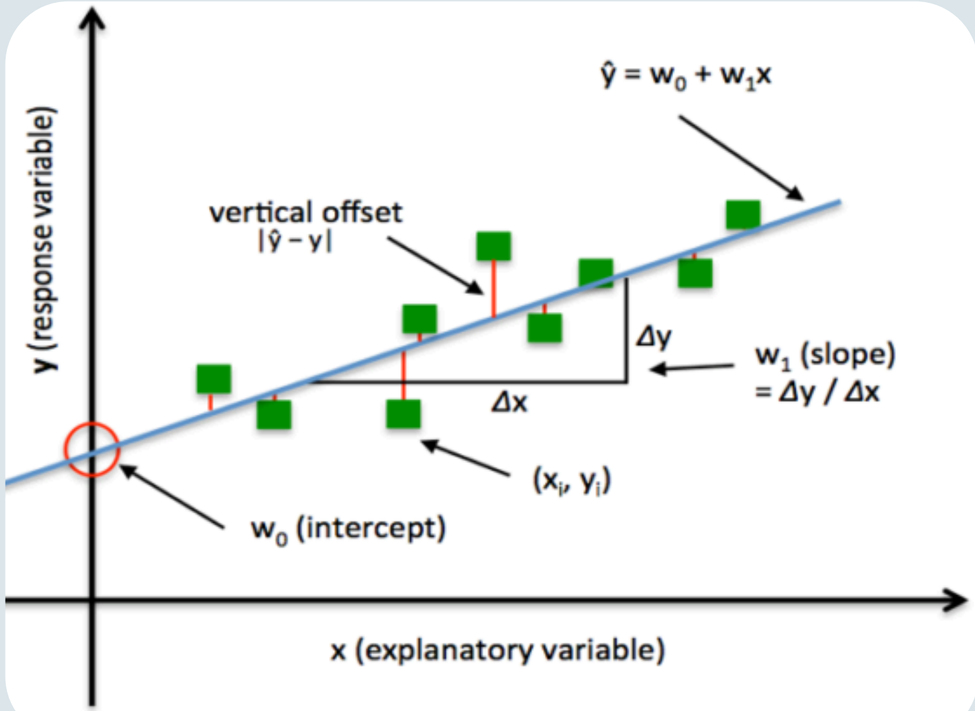
Trong đó:

- y_i : giá trị thực tế
- $\hat{y}_i$: giá trị dự đoán từ mô hình
- OLS giúp mô hình phù hợp nhất với dữ liệu hiện có
- Trong ngôn ngữ R, phương pháp OLS được sử thực hiện đơn giản bằng cách sử dụng hàm lm()



Phương pháp Stepwise (AIC)

- Mục đích: Phương pháp Stepwise được sử dụng nhằm mục đích lọc biến tự động trong hồi quy, nó sử dụng tiêu chí AIC để tự động chọn các biến đầu vào tốt nhất cho mô hình hồi quy.



Tóm lại vai trò của OLS và Stepwise trong mô hình hồi quy

- OLS là phương pháp ước lượng các hệ số hồi quy sao cho sai số nhỏ nhất.
- Stepwise(AIC) là kỹ thuật chọn biến đầu vào 1 cách tự động dựa trên tiêu chí thống kê (AIC)

→ Kết hợp hai phương pháp này giúp xây dựng mô hình chính xác hơn.

Model:

PLM

Fix Effect

1 Khái niệm

Fixed Effects (FE) giả định rằng mỗi đối tượng trong dữ liệu (ví dụ: quốc gia, công ty, cá nhân) có một hiệu ứng riêng biệt không thể thay đổi qua thời gian, và những hiệu ứng này có thể ảnh hưởng đến biến phụ thuộc $Y_{it} = \alpha_i + \beta X_{it} + \varepsilon_{it}$

alpha: là hiệu ứng riêng biệt của từng đối tượng

2 Ý nghĩa

- Phân tích các yếu tố biến thiên trong thời gian mà không bị ảnh hưởng bởi các yếu tố không quan sát được (fixed effects)

PLM Random Effect

1 Khái niệm

Random Effects (RE) giả định rằng các hiệu ứng không quan sát được giữa các đối tượng (quốc gia, công ty, cá nhân) là ngẫu nhiên và được phân phối theo một phân phối xác suất

$$Y_{it} = \alpha + \beta X_{it} + u_i + \varepsilon_{it}$$

α: Hằng số chung cho tất cả các đối tượng.

u_i: **Hiệu ứng ngẫu nhiên** riêng cho từng đối tượng i (giống như "tính cách" của từng quốc gia).

- Giả định: $u_i \sim N(0, \sigma_u^2)$
- Không tương quan với X_{it}

ε_{it}: nhiễu ngẫu nhiên (idiosyncratic error)

2 Ý nghĩa

- Đặc điểm riêng của từng đối tượng không liên quan đến các biến giải thích (ví dụ như GDP hay giáo dục không liên quan đến yếu tố "quốc gia" trong mô hình)

Fixed effects

1 Lệnh

```
fixef_values <- fixef(plm_model)
```

	COUNTRY <chr>	FIXED_EFFECT <dbl>
80	Sierra_Leone	39.63825
50	Lesotho	40.77764
54	Malawi	41.91442
2	Angola	42.18016
98	Zimbabwe	42.86379
17	Central_African_Republic	43.61022
66	Nigeria	43.83089
97	Zambia	44.51886
18	Chad	45.26920
83	South_Africa	46.50190

	COUNTRY <chr>	FIXED_EFFECT <dbl>
36	Germany	64.84808
12	Bosnia_and_Herzegovina	64.87424
22	Costa_Rica	65.85382
16	Canada	65.91338
23	Cyprus	65.95013
84	Spain	66.08804
57	Malta	66.14330
5	Austria	66.21305
34	France	66.54609
52	Luxembourg	66.82494

HIỆN TƯỢNG ĐA CỘNG TUYẾN OVERFITTING

Hiện tượng đa cộng tuyến

1 Khái niệm

Đa cộng tuyến là hiện tượng trong mô hình hồi quy tuyến tính khi có mối quan hệ tuyến tính mạnh mẽ giữa các biến độc lập. Điều này dẫn đến việc các biến độc lập không còn độc lập với nhau, gây khó khăn cho việc ước lượng chính xác các tham số của mô hình.

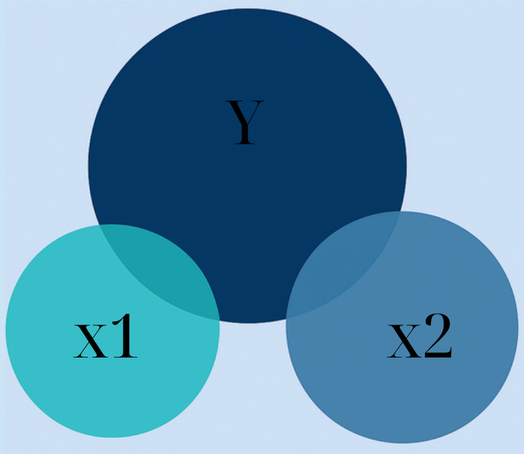
2 Nguyên Nhân

- Do các biến độc lập có quan hệ tuyến tính chặt chẽ với nhau.
- Do sử dụng nhiều biến mô tả cùng một đặc tính của dữ liệu.
- Do biến giả (dummy variables) được tạo không hợp lý.

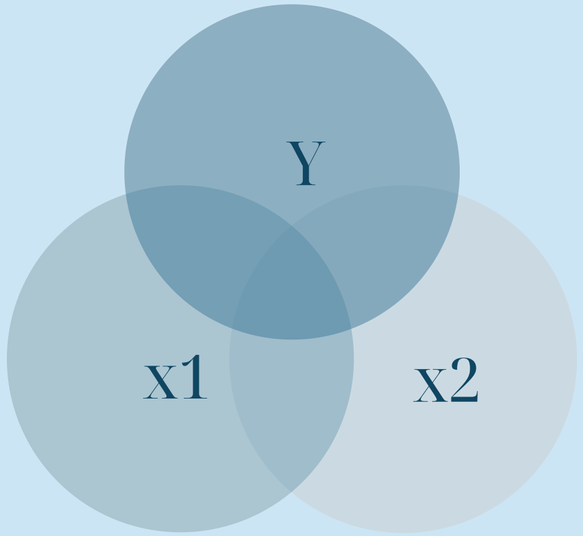
Hiện tượng đa cộng tuyến

3 Phân loại

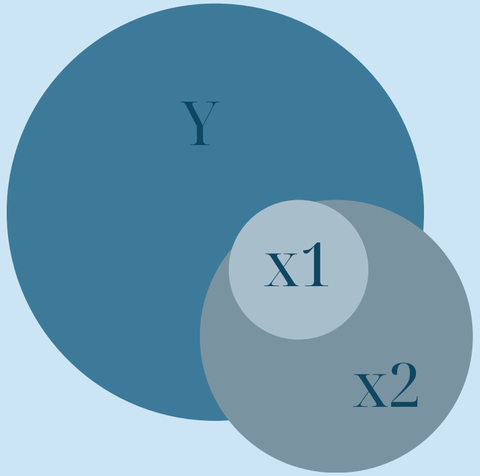
Không đa cộng tuyến



Đa cộng tuyến



Đa cộng tuyến hoàn hảo



Hiện tượng đa cộng tuyến

4 Hậu quả và dấu hiệu nhận biết

Hậu quả

- Hệ số hồi quy không ổn định: Sai số chuẩn (Standard Error) tăng cao, dẫn đến kiểm định t không còn ý nghĩa.
- Khó diễn giải mô hình: Không thể tách biệt tác động của từng biến.
- Mô hình dễ bị overfitting nếu không xử lý

Dấu hiệu:

- Hệ số hồi quy có giá trị lớn bất thường hoặc có dấu không hợp lý.
- Khi thêm hoặc bớt biến độc lập, hệ số hồi quy thay đổi đáng kể.

5 Giải pháp

- Loại bỏ biến dư thừa: Chọn biến có ý nghĩa nhất hoặc dùng phương pháp chọn biến
- Chỉ phân tích thành phần chính: Giảm chiều dữ liệu bằng cách gộp các biến tương quan vào thành phần mới.

Hiện tượng Overfitting

1 Khái niệm

Overfitting (quá khớp) là một hành vi máy học không mong muốn xảy ra khi mô hình máy học đưa ra dự đoán chính xác cho dữ liệu đào tạo nhưng không cho dữ liệu mới. Một mô hình overfit có thể đưa ra dự đoán không chính xác và không thể thực hiện tốt cho tất cả các loại dữ liệu mới.

2 Nguyên nhân

- Sử dụng quá nhiều biến độc lập hoặc mô hình quá phức tạp.
- Dữ liệu huấn luyện có nhiễu nhưng mô hình lại học theo cả những yếu tố nhiễu đó.
- Số lượng dữ liệu huấn luyện quá ít so với số lượng biến đầu vào.

Hiện tượng Overfitting

3 Hậu quả và dấu hiệu nhận biết

Hậu quả:

- Hiệu suất trên tập huấn luyện rất tốt, nhưng trên tập kiểm tra rất tệ.
- Mô hình không tổng quát hóa được .

Dấu hiệu

- Sai số trên tập huấn luyện rất nhỏ nhưng sai số trên tập kiểm tra cao.
- Mô hình có độ chính xác rất cao nhưng không phản ánh đúng xu hướng tổng thể.
- Độ lệch bias thấp nhưng phương sai (variance) cao.

4 Giải Pháp

- Giảm số lượng biến: Chỉ chọn các biến thực sự có ý nghĩa.
- Tăng kích thước dữ liệu huấn luyện: Giúp mô hình học tổng quát hơn.
- Kiểm tra hiệu suất mô hình trên nhiều tập con của dữ liệu để phát hiện overfitting.

Kiểm định

1 Kiểm định Hausman

Giúp lựa chọn giữa Fix Effects và Random Effects

- H0: Không có tương quan → RE phù hợp.
- H1: Có tương quan → Chọn FE.
- p-value < 0.05 → Bác bỏ H0 → Chọn Fixed Effects.
- p-value ≥ 0.05 → Không bác bỏ H0 → Chọn Random Effects.

2 Tự tương quan

- Kiểm tra xem phần dư (residuals) có bị liên hệ theo thời gian hay không.
- Nếu phần dư ở thời điểm t liên quan đến phần dư ở t−1 → các ước lượng vẫn đúng trung bình (unbiased) nhưng không hiệu quả
- p-value < 0.05 → Có tự tương quan → cần điều chỉnh.
- p-value ≥ 0.05 → Không có vấn đề.

Kiểm định

3. Robust Standard Errors

khắc phục sai số chuẩn bị sai do các vấn đề như:

- Phương sai thay đổi.
- Tự tương quan.

Nếu phát hiện sai số chuẩn bị sai, thì p-value và kiểm định t sẽ không đáng tin → cần dùng robust standard error

```
        studentized Breusch-Pagan test

data:  plm_model
BP = 328.68, df = 4, p-value < 2.2e-16


t test of coefficients:

                                 Estimate Std. Error t value Pr(>|t|)
Income.composition.of.resources  4.1187020  1.1027550  3.7349 0.000195 ***
Adult.Mortality                 -0.0039254  0.0012623 -3.1097 0.001909 **
Schooling                        0.7834820  0.2448152  3.2003 0.001402 **
log(BMI)                        -0.0219628  0.1001513 -0.2193 0.826450
---
```

THANK YOU FOR YOUR ATTENTION