**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ CHỦ NGHĨA XÃ HỘI KHOA HỌC**

**BÁO CÁO CUỐI KÌ**

**PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN TUỔI THỌ**

**CỦA CÁC QUỐC GIA**

**BẰNG PHƯƠNG TRÌNH HỒI QUY TUYẾN TÍNH**

**MÃ MÔN HỌC & MÃ LỚP: RPAN233577\_01**

**GIÁO VIÊN HƯỚNG DẪN: TS.Phan Thị Thể**

**NHÓM THỰC HIỆN: 15**

**Lương Ngọc Huy - 23133028**

**Nguyễn Hữu Tâm 23133067**

**Nguyễn Đức Thịnh 23133073**

**Nguyễn Đức Toàn 23133077**

Tp. Hồ Chí Minh, tháng 5 năm 2025

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc197784365)

[1.1 Lý do chọn đề tài 1](#_Toc197784366)

[1.2 Mục tiêu đề tài 1](#_Toc197784367)

[1.3 Phạm vi đề tài 2](#_Toc197784368)

[1.4 Đối tượng nghiên cứu 2](#_Toc197784369)

[1.5 Phương pháp nghiên cứu 2](#_Toc197784370)

[1.6 Bố cục đề tài 3](#_Toc197784371)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 4](#_Toc197784372)

[2.1. Tổng quan về tuổi thọ và các yếu tố tác động 4](#_Toc197784373)

[2.2. Lý thuyết về mô hình hồi quy 6](#_Toc197784374)

[2.2.1. Hồi quy tuyến tính đơn biến 7](#_Toc197784375)

[2.2.2. Hồi quy tuyến tính đa biến 8](#_Toc197784376)

[2.2.3. Hồi quy tuyến tính logarit 10](#_Toc197784377)

[2.3 Mô hình dữ liệu bảng (Panel Data) 11](#_Toc197784378)

[2.3.1. Hiệu ứng cố định (Fixed Effects) 12](#_Toc197784379)

[2.3.2. Hiệu ứng ngẫu nhiên (Random Effects) 13](#_Toc197784380)

[2.4. Các giả định và kiểm định trong hồi quy 14](#_Toc197784381)

[2.4.1. Đa cộng tuyến, phương sai sai số thay đổi 14](#_Toc197784382)

[2.4.2. Tự tương quan và phân phối chuẩn của phần dư 16](#_Toc197784383)

[CHƯƠNG 3 TIỀN XỬ LÝ DỮ LIỆU VÀ XÂY DỰNG MÔ HÌNH 19](#_Toc197784384)

[3.1 Tổng quan về dữ liệu 19](#_Toc197784385)

[3.2. Quy trình tiền xử lý dữ liệu 22](#_Toc197784386)

[3.2.1. Làm sạch và chuẩn hóa dữ liệu 22](#_Toc197784387)

[3.3 Phương pháp chọn biến độc lập 28](#_Toc197784388)

[3.3.1. Ma trận tương quan 28](#_Toc197784389)

[3.3.2 Stepwise AIC 30](#_Toc197784390)

[3.4 Xây dựng mô hình hồi quy 31](#_Toc197784391)

[3.4.1. Hồi quy OLS 31](#_Toc197784392)

[3.4.2 Mô hình PLM (Panel Linear Model) 32](#_Toc197784393)

[3.5 Kiểm tra giả định các mô hình hồi quy tuyến tính 33](#_Toc197784394)

[3.5.1 Mô hình OLS 33](#_Toc197784395)

[3.5.2 Mô hình PLM 38](#_Toc197784396)

[3.5.3 Cải thiện mô hình 41](#_Toc197784397)

[CHƯƠNG 4: KIỂM THỬ VÀ ĐÁNH GIÁ KÊT QUẢ 44](#_Toc197784398)

[4.1 Kiểm thử mô hình 44](#_Toc197784399)

[4.2 Đánh giá mục tiêu 46](#_Toc197784400)

[CHƯƠNG 5: KẾT LUẬN 50](#_Toc197784401)

[5.1 Kết luận 50](#_Toc197784402)

[5.2 Hạn chế của đề tài 50](#_Toc197784403)

[5.3 Hướng phát triển 51](#_Toc197784404)

[TÀI LIỆU THAM KHẢO 53](#_Toc197784405)

# CHƯƠNG 1: GIỚI THIỆU

## 1.1 Lý do chọn đề tài

Tuổi thọ trung bình là một chỉ số quan trọng thể hiện chất lượng sống và mức độ phát triển kinh tế – xã hội của một quốc gia. Trong bối cảnh toàn cầu hoá và phát triển bền vững ngày càng được quan tâm, việc nghiên cứu các yếu tố ảnh hưởng đến tuổi thọ là cần thiết nhằm hỗ trợ các nhà hoạch định chính sách đưa ra các quyết định đúng đắn, cải thiện điều kiện sống cho người dân.

Có nhiều yếu tố có khả năng tác động đến tuổi thọ trung bình của người dân. Một số yếu tố mang tính kinh tế như thu nhập bình quân đầu người, chi tiêu y tế, hoặc mức độ đô thị hóa. Một số yếu tố khác mang tính xã hội như trình độ học vấn, tỉ lệ biết chữ, tiếp cận thông tin, điều kiện sống. Thêm vào đó, các yếu tố môi trường như mức độ ô nhiễm không khí, chất lượng nguồn nước, và biến đổi khí hậu cũng đóng vai trò quan trọng không kém. Chính sự đa dạng và phức tạp trong mối quan hệ giữa các yếu tố này với tuổi thọ đặt ra yêu cầu cần có một phương pháp định lượng để đánh giá chính xác và khách quan mức độ ảnh hưởng của từng yếu tố.

Để định lượng và đánh giá mức độ ảnh hưởng của từng yếu tố, phương pháp hồi quy tuyến tính là công cụ phù hợp để phân tích mối quan hệ giữa biến phụ thuộc (tuổi thọ) và các biến độc lập liên quan. Do đó, nhóm chúng tôi quyết định lựa chọn đề tài: *“Phân tích các yếu tố ảnh hưởng đến tuổi thọ của các quốc gia bằng phương trình hồi quy tuyến tính”* nhằm cung cấp cái nhìn rõ hơn về tầm quan trọng và mức độ ảnh hưởng của từng yếu tố tới tuổi thọ, từ đó đề xuất các khuyến nghị chính sách phù hợp.

## 1.2 Mục tiêu đề tài

Bài phân tích mang một mục tiêu tổng quát để phân tích và đánh giá các yếu tố tác động đến tuổi thọ trung bình của người dân tại các quốc gia thông qua mô hình hồi quy tuyến tính. Bằng cách đặt ra một loạt câu hỏi về nhiều mặt ảnh hưởng đến mức tuổi thọ của các quốc gia. Những câu hỏi cụ thể:

Thu nhập bình quân đầu người ảnh hưởng như thế nào đến tuổi thọ?

Mức chi tiêu y tế bình quân đầu người ảnh hưởng ra sao đến tuổi thọ?

Tỷ lệ tử vong trẻ sơ sinh có liên hệ như thế nào với tuổi thọ trung bình?

Mức độ giáo dục ảnh hưởng thế nào đến tuổi thọ?

## 1.3 Phạm vi đề tài

Dữ liệu nghiên cứu sẽ bao gồm thông tin thống kê của khoảng hơn 100 quốc gia trên thế giới trong một khoảng thời gian nhất định (2010–2015), được thu thập từ các nguồn dữ liệu công khai từ WHO được tổng hợp trên Kaggle. Dataset bao gồm các thành phần ảnh hưởng đến tình trạng sức khỏe của một quốc gia như tỉ lệ tiêm chủng, tỉ lệ chết, tình trạng kinh tế, xã hội.

## 1.4 Đối tượng nghiên cứu

Các yếu tố kinh tế - xã hội và môi trường ảnh hưởng đến tuổi thọ trung bình tại các quốc gia. Biến phụ thuộc là tuổi thọ trung bình, còn các biến độc lập có thể bao gồm:

GDP bình quân đầu người

Chi tiêu y tế

Tỷ lệ tử vong trẻ sơ sinh/người lớn

Tỷ lệ tiêm chủng

Tỷ lệ nhiễm các bệnh xã hội

Số năm đi học trung bình

## 1.5 Phương pháp nghiên cứu

Để phân tích một cách toàn diện các yếu tố ảnh hưởng đến tuổi thọ trung bình của các quốc gia, đề tài này kết hợp giữa hai phương pháp chính: nghiên cứu định tính và nghiên cứu định lượng. Việc kết hợp hai phương pháp này sẽ giúp đảm bảo tính khách quan, độ tin cậy cũng như chiều sâu trong phân tích.

Phương pháp định tính: Thu thập và phân tích các công trình nghiên cứu trong và ngoài nước liên quan đến các yếu tố ảnh hưởng đến tuổi thọ, bao gồm các tài liệu của Tổ chức Y tế Thế giới (WHO) cũng như các nghiên cứu học thuật được công bố trên các tạp chí khoa học uy tín.Từ những nghiên cứu đã có, đề tài tiến hành xác định các biến độc lập có khả năng ảnh hưởng đến tuổi thọ trung bình.

Phương pháp định lượng: Phương pháp định lượng là trọng tâm của nghiên cứu nhằm lượng hóa mức độ và chiều hướng ảnh hưởng của các biến số độc lập đến tuổi thọ trung bình. vì dữ liệu thu thập được là dữ liệu bảng, do đó áp dụng phương pháp kỹ thuật phân tích dữ liệu bảng với các kiểm định chính là kiểm định đa cộng tuyến, kiểm định phƣơng sai tự thay đổi và kiểm định hiện tương tự tương quan bằng mô hình hồi quy hiệu ứng cố định ( Fixed effects models) và mô hình hồi quy hiệu ứng ngẫu nhiên ( Random effects models), phương pháp GLS.

## 1.6 Bố cục đề tài

Chương 1: Giới thiệu

Chương 2: Cơ sở lý thuyết  
Chương 3 Tiền xử lý dữ liệu và xây dựng mô hình

Chương 4: Phân tích và đánh giá kêt quả

Chương 5: Kết luận

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## 2.1. Tổng quan về tuổi thọ và các yếu tố tác động

Tuổi thọ là một chỉ tiêu thống kê phản ánh số năm sống trung bình mà một cá nhân được kỳ vọng sẽ đạt được từ thời điểm sinh ra, với giả định rằng các điều kiện sống và tỷ lệ tử vong hiện tại vẫn giữ nguyên trong suốt cuộc đời. Đây là một thước đo tổng hợp có ý nghĩa đặc biệt quan trọng trong việc đánh giá chất lượng cuộc sống, hiệu quả của hệ thống y tế, mức độ phát triển kinh tế – xã hội và công bằng xã hội trong một quốc gia hoặc vùng lãnh thổ.

Qua nhiều nghiên cứu thực nghiệm và dữ liệu thống kê trên toàn cầu, người ta đã xác định được nhiều nhóm yếu tố có ảnh hưởng trực tiếp hoặc gián tiếp đến tuổi thọ. Một số nhóm yếu tố tiêu biểu bao gồm:

*Y tế và chăm sóc sức khỏe*

Hệ thống y tế chất lượng đóng vai trò cốt lõi trong việc kéo dài tuổi thọ của dân số. Việc người dân được tiếp cận với các dịch vụ y tế hiện đại, từ chăm sóc sức khỏe ban đầu, tiêm chủng, điều trị bệnh mãn tính đến các chương trình phòng ngừa bệnh tật đều giúp giảm tỷ lệ tử vong. Ngoài ra, chất lượng đội ngũ y bác sĩ, cơ sở vật chất và khả năng chi trả của người dân đối với dịch vụ y tế cũng là những yếu tố quan trọng.

*Trình độ giáo dục*

Giáo dục không chỉ đóng vai trò trong việc nâng cao năng lực lao động mà còn giúp người dân có kiến thức và ý thức hơn về chăm sóc sức khỏe bản thân và gia đình. Người có trình độ học vấn cao thường biết cách phòng tránh bệnh tật, tuân thủ các khuyến cáo y tế, xây dựng chế độ ăn uống khoa học và duy trì lối sống lành mạnh. Nhiều nghiên cứu cho thấy có mối tương quan dương giữa số năm đi học trung bình và tuổi thọ trung bình tại các quốc gia.

*Thu nhập và mức sống*

Mức thu nhập là một yếu tố kinh tế có ảnh hưởng sâu rộng đến nhiều khía cạnh của cuộc sống. Người có thu nhập cao thường có điều kiện tiếp cận dịch vụ y tế tốt, sinh sống trong môi trường ít ô nhiễm, có chế độ ăn uống đầy đủ dinh dưỡng và có khả năng ứng phó tốt hơn với các rủi ro sức khỏe. Ngược lại, nghèo đói làm tăng nguy cơ mắc bệnh, thiếu tiếp cận chăm sóc y tế và giảm khả năng sống thọ.

*Môi trường sống*

Các yếu tố môi trường như ô nhiễm không khí, nguồn nước không đảm bảo vệ sinh, tiếng ồn, hóa chất độc hại,… đều có thể gây ra nhiều bệnh mãn tính và đe dọa đến sức khỏe cộng đồng. Ở các khu vực đô thị lớn, nơi có mật độ dân cư cao và công nghiệp phát triển nhanh, tình trạng ô nhiễm là một thách thức lớn đối với tuổi thọ của cư dân. Chính vì vậy, chính sách bảo vệ môi trường và phát triển bền vững cũng là một phần không thể tách rời trong chiến lược nâng cao tuổi thọ.

*Chế độ dinh dưỡng và lối sống*

Những hành vi và thói quen hàng ngày có ảnh hưởng trực tiếp đến sức khỏe và tuổi thọ. Việc duy trì chế độ ăn uống hợp lý, giàu rau xanh, ít chất béo bão hòa; tập thể dục đều đặn; ngủ đủ giấc; kiểm soát căng thẳng; tránh các hành vi nguy cơ như hút thuốc lá, lạm dụng rượu bia,… là những yếu tố tích cực góp phần kéo dài tuổi thọ. Ngược lại, lối sống không lành mạnh là nguyên nhân phổ biến dẫn đến các bệnh mãn tính như tim mạch, đột quỵ, tiểu đường, ung thư.

Các yếu tố nêu trên thường không tác động riêng lẻ mà có mối liên hệ qua lại và ảnh hưởng đồng thời đến tuổi thọ. Ví dụ, người có thu nhập cao thường cũng có trình độ học vấn cao, điều kiện sống tốt và ý thức chăm sóc sức khỏe cao hơn. Do đó, để hiểu rõ vai trò riêng biệt của từng yếu tố và mức độ ảnh hưởng của chúng đối với tuổi thọ, cần áp dụng các phương pháp phân tích định lượng – trong đó mô hình hồi quy là công cụ phổ biến và hiệu quả.

Mô hình hồi quy giúp lượng hóa mối quan hệ giữa tuổi thọ và các biến giải thích khác nhau, từ đó hỗ trợ đưa ra các khuyến nghị chính sách cụ thể như: ưu tiên đầu tư vào giáo dục, cải thiện hệ thống y tế, kiểm soát ô nhiễm môi trường hay thúc đẩy các chiến dịch truyền thông thay đổi hành vi sức khỏe trong cộng đồng.

## 2.2. Lý thuyết về mô hình hồi quy

Hồi quy là một kỹ thuật thống kê quan trọng được sử dụng để phân tích và mô hình hóa mối quan hệ giữa một biến phụ thuộc (dependent variable) và một hoặc nhiều biến độc lập (independent variables). Trong bối cảnh nghiên cứu về tuổi thọ, hồi quy giúp chúng ta hiểu được các yếu tố như thu nhập bình quân đầu người, trình độ học vấn, chi tiêu y tế, mức độ đô thị hóa, chất lượng môi trường,… ảnh hưởng đến tuổi thọ trung bình của dân cư như thế nào, cũng như định lượng mức độ ảnh hưởng đó.

Mục tiêu của hồi quy là xây dựng một mô hình toán học có khả năng ước lượng hoặc dự đoán giá trị của biến phụ thuộc dựa trên giá trị của các biến độc lập. Thông qua mô hình hồi quy, ta có thể kiểm định giả thuyết, rút ra kết luận về mối quan hệ nhân quả, cũng như xác định yếu tố nào là quan trọng nhất trong việc tác động đến biến nghiên cứu.

Các thành phần chính trong một mô hình hồi quy:

* Biến phụ thuộc (Y): Là đối tượng được nghiên cứu hoặc dự đoán. Trong nghiên cứu này, đó là *tuổi thọ trung bình* của một quốc gia hoặc khu vực.
* Biến độc lập (X): Là những yếu tố có thể ảnh hưởng đến biến phụ thuộc, chẳng hạn như thu nhập bình quân, tỷ lệ biết chữ, chi tiêu y tế, tỷ lệ béo phì, ...
* Hệ số hồi quy (β): Đại diện cho mức độ thay đổi của biến phụ thuộc khi biến độc lập tăng thêm một đơn vị (trong điều kiện các biến khác giữ nguyên). Đây là yếu tố then chốt để đánh giá mức ảnh hưởng của từng biến độc lập.
* Sai số (ε): Phản ánh sự chênh lệch giữa giá trị quan sát và giá trị dự đoán từ mô hình, đại diện cho các yếu tố chưa được đưa vào mô hình hoặc yếu tố ngẫu nhiên.

### 2.2.1. Hồi quy tuyến tính đơn biến

Hồi quy tuyến tính đơn biến là một mô hình thống kê dùng để phân tích mối quan hệ giữa một biến phụ thuộc (Y) và một biến độc lập (X). Mục tiêu của mô hình là ước lượng ảnh hưởng của X đến Y, đồng thời dự đoán giá trị của Y dựa trên giá trị đã biết của X.

Công thức của mô hình hồi quy tuyến tính đơn biến có dạng:

***Y = β₀ + β₁X + ε***

Trong đó:

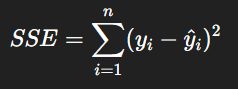
* Y: Biến phụ thuộc
* X: Biến độc lập
* β₀: Tung độ gốc
* β₁: Hệ số góc
* ε: Sai số ngẫu nhiên

*Ý nghĩa của các hệ số*

* β1>0: Mối quan hệ tuyến tính thuận chiều giữa X và Y
* β1<0: Mối quan hệ tuyến tính nghịch chiều
* β1=0: Không tồn tại mối quan hệ tuyến tính giữa X và Y

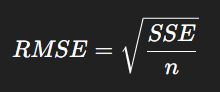
Hiệu quả của mô hình hồi quy được đánh giá thông qua các chỉ số:

*Tổng bình phương sai số (SSE - Sum of Squared Errors):*



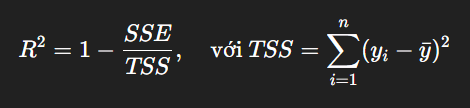
SSE đo lường tổng sai số giữa giá trị thực và giá trị dự đoán. SSE càng nhỏ cho thấy mô hình càng phù hợp.

*Sai số bình phương trung bình gốc (RMSE - Root Mean Squared Error):*



RMSE cho biết độ lệch trung bình giữa giá trị thực tế và giá trị dự báo. RMSE càng nhỏ chứng tỏ mô hình dự đoán càng chính xác. RMSE có cùng đơn vị với biến Y, giúp diễn giải dễ dàng hơn.

*Hệ số xác định (R² - R-squared):*



R² biểu thị tỷ lệ phần trăm biến thiên của biến phụ thuộc được giải thích bởi mô hình. R² gần 1 cho thấy mô hình giải thích tốt dữ liệu.

### 2.2.2. Hồi quy tuyến tính đa biến

Hồi quy tuyến tính đa biến là mô hình mở rộng từ hồi quy tuyến tính đơn biến, cho phép phân tích mối quan hệ giữa một biến phụ thuộc (Y) và nhiều biến độc lập (X₁, X₂, ..., Xₚ). Mục tiêu là xác định mức độ và hướng ảnh hưởng của từng biến độc lập đến biến phụ thuộc.

Công thức của mô hình hồi quy tuyến tính đa biến có dạng:

***Y = β0 + β1​X1 + β2​X2​ + ⋯ + βp​Xp​ + ε***

Trong đó:

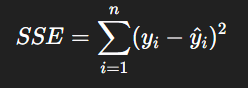
* Y: Biến phụ thuộc (ví dụ: tuổi thọ)
* X1, X2, ..., Xp : Các biến độc lập
* β0: Hệ số chặn
* β1, ..., βp : Hệ số hồi quy tương ứng với các biến độc lập
* ε: Sai số ngẫu nhiên

*Ý nghĩa của các hệ số*

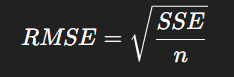
* Mỗi hệ số βi phản ánh mức thay đổi trung bình của Y khi Xi thay đổi một đơn vị, giữ các biến khác không đổi (tác động biên).
* Nếu βi > 0: Biến Xi có quan hệ thuận chiều với Y.
* Nếu βi < 0: Biến Xi có quan hệ nghịch chiều với Y.

Để kiểm tra mức độ phù hợp và độ chính xác của mô hình, người ta sử dụng các chỉ số sau:

* Tổng bình phương sai số phần dư (SSE):

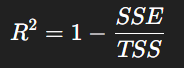


* Sai số bình phương trung bình gốc (RMSE):

​​

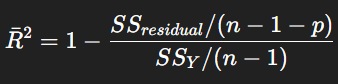
RMSE càng nhỏ, mô hình dự đoán càng chính xác.

* Hệ số xác định (R²):



Cho biết bao nhiêu phần trăm phương sai của biến Y được giải thích bởi mô hình. Tuy nhiên, R² luôn tăng hoặc không đổi khi thêm biến, kể cả biến không quan trọng.

* Hệ số xác định hiệu chỉnh (Adjusted R²):



Trong đó:

* + n: Số quan sát
  + p: Số biến độc lập
  + SSresidual: Tổng bình phương phần dư
  + SSY: Tổng bình phương toàn phần

Adjusted R² điều chỉnh R² theo số biến độc lập, phản ánh mức độ phù hợp thực tế của mô hình. Nếu thêm biến không mang lại lợi ích, Adjusted R² sẽ giảm.

Khi sử dụng mô hình hồi quy tuyến tính đa biến, cần lưu ý một số vấn đề nhằm đảm bảo độ chính xác và ý nghĩa của kết quả phân tích. Trước hết, hiện tượng đa cộng tuyến có thể xảy ra khi các biến độc lập có mối tương quan chặt chẽ với nhau. Điều này làm sai lệch ước lượng các hệ số hồi quy, khiến việc diễn giải kết quả trở nên khó khăn và thiếu tin cậy. Để phát hiện và xử lý vấn đề này, người phân tích thường sử dụng chỉ số VIF (Variance Inflation Factor) – một thước đo mức độ biến động của hệ số do đa cộng tuyến gây ra.

Ngoài ra, cần chú trọng đến việc chọn lọc biến độc lập phù hợp trong mô hình. Việc đưa quá nhiều biến vào mô hình không chỉ làm tăng độ phức tạp mà còn có thể đưa vào những biến không có ý nghĩa thống kê, gây nhiễu cho mô hình. Do đó, các kỹ thuật chọn biến như forward selection, backward elimination hoặc stepwise regression thường được sử dụng nhằm tối ưu hóa mô hình và đảm bảo chỉ giữ lại các biến thực sự có ảnh hưởng đến biến phụ thuộc.

### 2.2.3. Hồi quy tuyến tính logarit

Hồi quy tuyến tính logarit là một dạng biến thể của mô hình hồi quy tuyến tính, trong đó biến phụ thuộc hoặc biến độc lập (hoặc cả hai) được biến đổi bằng hàm logarit (thường là log tự nhiên – ln). Việc biến đổi này thường nhằm mục đích giảm thiểu hiện tượng phân phối lệch, tuyến tính hóa mối quan hệ phi tuyến giữa các biến, đồng thời ổn định phương sai trong mô hình.

Tùy thuộc vào cách biến đổi, mô hình hồi quy logarit có thể được chia thành ba dạng phổ biến:

***Log-Linear (Log-Log)***

Đây là mô hình trong đó cả biến phụ thuộc và biến độc lập đều được logarit hóa, có dạng:

***ln(Y) = β0 + β1ln(X) + ε***

Trong mô hình này, hệ số β1 được gọi là hệ số co giãn, thể hiện mức thay đổi phần trăm của Y ứng với một phần trăm thay đổi trong X.

Cụ thể, nếu X tăng 1%, thì Y sẽ thay đổi trung bình β1%.Mô hình log-log thường được dùng khi cả hai biến đều có tính chất tỷ lệ và có mối quan hệ phi tuyến trên thang đo gốc, nhưng tuyến tính khi đưa về dạng logarit.

***Log-Lin (Semi Log)***

Trong mô hình này, chỉ biến phụ thuộc Y được logarit hóa, còn biến độc lập X vẫn giữ nguyên dạng ban đầu. Phương trình có dạng:

***ln(Y) = β0 + β1X + ε***

Ý nghĩa của hệ số β1 trong mô hình này là: khi biến X tăng thêm 1 đơn vị, thì giá trị trung bình của Y sẽ thay đổi khoảng 100×β1%. Đây là mô hình phù hợp khi giá trị Y có phân phối lệch, hoặc khi ta muốn mô tả mức độ thay đổi tương đối (theo %) của Y thay vì thay đổi tuyệt đối.

***Lin-Log***

Ở dạng này, chỉ biến độc lập X được logarit hóa, còn biến phụ thuộc Y giữ nguyên. Mô hình được viết dưới dạng:

***Y = β0 + β1ln(X) + ε***

Trong mô hình này, β1 cho biết khi X tăng 1%, thì Y thay đổi trung bình khoảng β1/100 đơn vị. Đây là dạng hồi quy thích hợp khi biến độc lập X có sự biến thiên lớn hoặc phân phối lệch, và ta muốn biết tác động của mức tăng tỷ lệ phần trăm trong X đến giá trị tuyệt đối của Y.

## 2.3 Mô hình dữ liệu bảng (Panel Data)

Khái niệm: Dữ liệu bảng là dữ liệu có quy mô về cả thời gian lẫn không gian. Cấu trúc dữu liệu bảng được kết hợp từ 2 thành phần: thành phần dữ liệu chéo (cross - section) và thành phần dữ liệu theo chuỗi thời gian(time series). Việc kết hợp 2 loại dữ liệu có nhiều lợi thế và thuận lợi trong phân tích, đặc biệt khi muốn quan sát, phân tích sự biến động của các nhóm đối tượng nghiên cứu sau các biến cố hay theo thời gian cũng như phân tích sự khác biệt giữa các nhóm đối tượng nghiên cứu.

Dữ liệu bảng nhờ đặc trưng theo dõi các đơn vị quan sát qua thời gian, mang lại nhiều ưu thế nổi bật so với dữ liệu chéo hay chuỗi thời gian thuần túy. Trước hết, dữ liệu bảng cho phép xử lý rõ ràng tính không đồng nhất giữa các đơn vị quan sát thông qua việc đưa vào các biến chuyên biệt theo từng ngân hàng hoặc đối tượng nghiên cứu. Việc kết hợp giữa các yếu tố chuỗi thời gian và dữ liệu chéo không chỉ làm tăng lượng thông tin hữu ích mà còn giúp giảm hiện tượng đa cộng tuyến giữa các biến, tăng bậc tự do và nâng cao hiệu quả ước lượng. Hơn nữa, do bao gồm các quan sát lặp lại theo thười gian, dữ liệu bảng rất phù hợp cho việc phân tích sự thay đổi của các đối tượng nghiên cứu. Khả năng phát hiện và đo lường các yếu tố không quan sát được cũng được cải thiện đáng kể, điều mà dữ liệu chuỗi thời gian hay dữ liệu chéo thuần túy khó thực hiện. Ngoài ra, dữ liệu bảng còn tạo điều kiện thuận lợi cho việc nghiên cứu các mô hình hành vi phức tạp như lợi thế kinh tế theo quy mô hay sự thay đổi công nghệ. Cuối cùng, với khả năng cung cấp dữ liệu từ hàng nghìn đơn vị, dữ liệu bảng giúp giảm thiểu sai lệch do việc gộp nhóm theo các biến tổng hợp, từ đó nâng cao độ chính xác của phân tích.

*Hồi quy dữ liệu bảng trong phân tích cơ bản có hai phương pháp sau: FEM (tác động cố định), REM (tác động ngẫu nhiên).*

### 2.3.1. Hiệu ứng cố định (Fixed Effects)

Mô hình Tác Động Cố Định (Fixed Effects Model - FEM) là một trong những phương pháp phổ biến trong phân tích dữ liệu bảng (panel data), sử dụng để nghiên cứu ảnh hưởng của các biến độc lập đến biến phụ thuộc khi có sự khác biệt giữa các đơn vị trong mẫu nghiên cứu nhưng không thay đổi theo thời gian.

Trong FEM, các yếu tố không quan sát được của từng đơn vị (còn gọi là sai số thành phần chéo) được coi là cố định và có tác động riêng lên biến phụ thuộc. FEM cho phép kiểm soát các yếu tố cố định này để xác định sự khác biệt giữa các đơn vị nghiên cứu (ví dụ: các công ty hoặc cá nhân) qua thời gian.

Đặc điểm của mô hình FEM

Mô hình FEM kiểm soát sự khác biệt không quan sát được giữa các đơn vị nghiên cứu, ví dụ như tính chất hoặc yếu tố nội tại không thay đổi theo thời gian.

Sai số thành phần chéo được cố định với từng đơn vị nhưng có thể thay đổi theo thời gian.

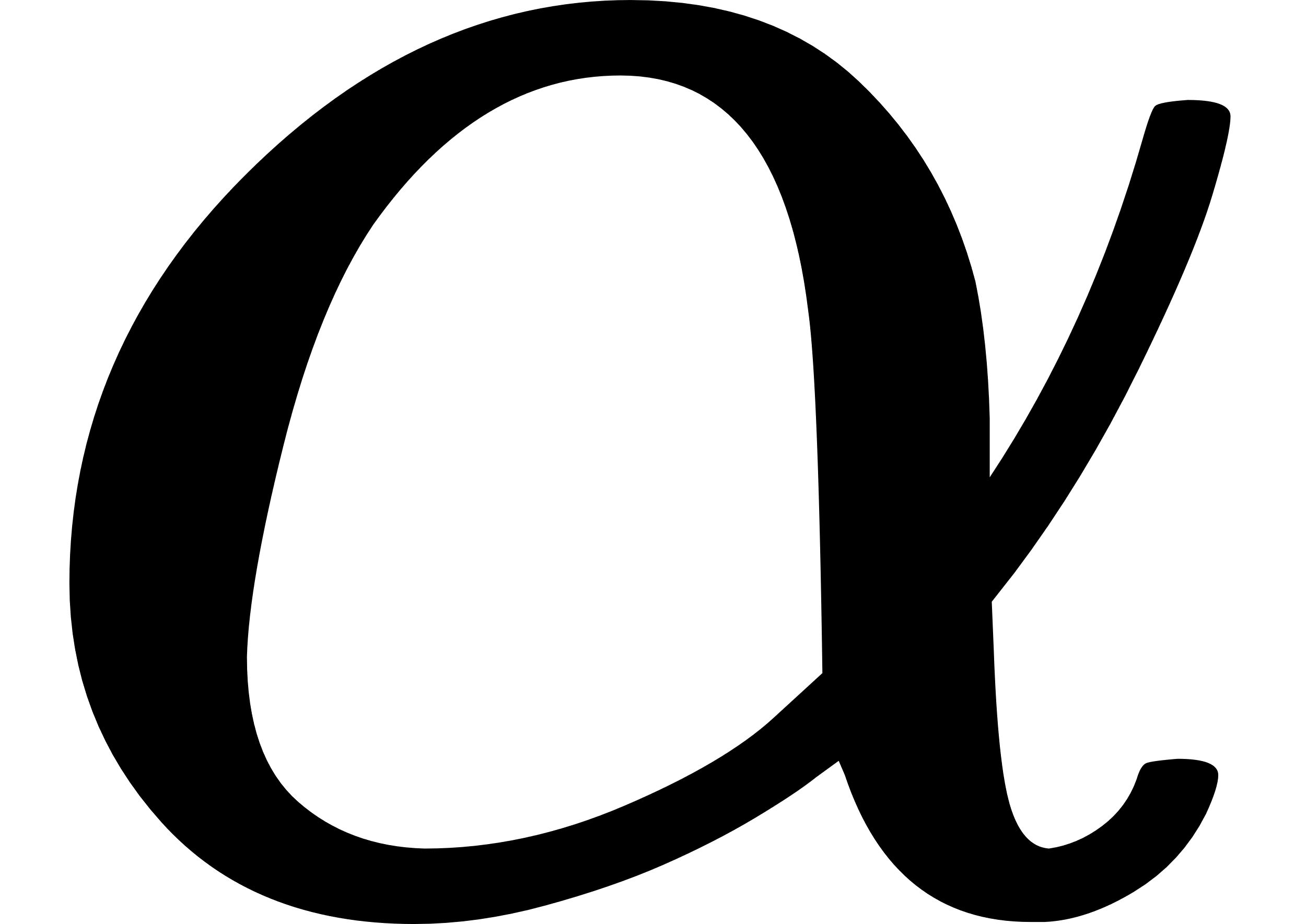
FEM phù hợp khi chúng ta quan tâm đến tác động của các biến độc lập trong từng đơn vị, trong khi giả định rằng các yếu tố không quan sát được là đặc thù và cố định.

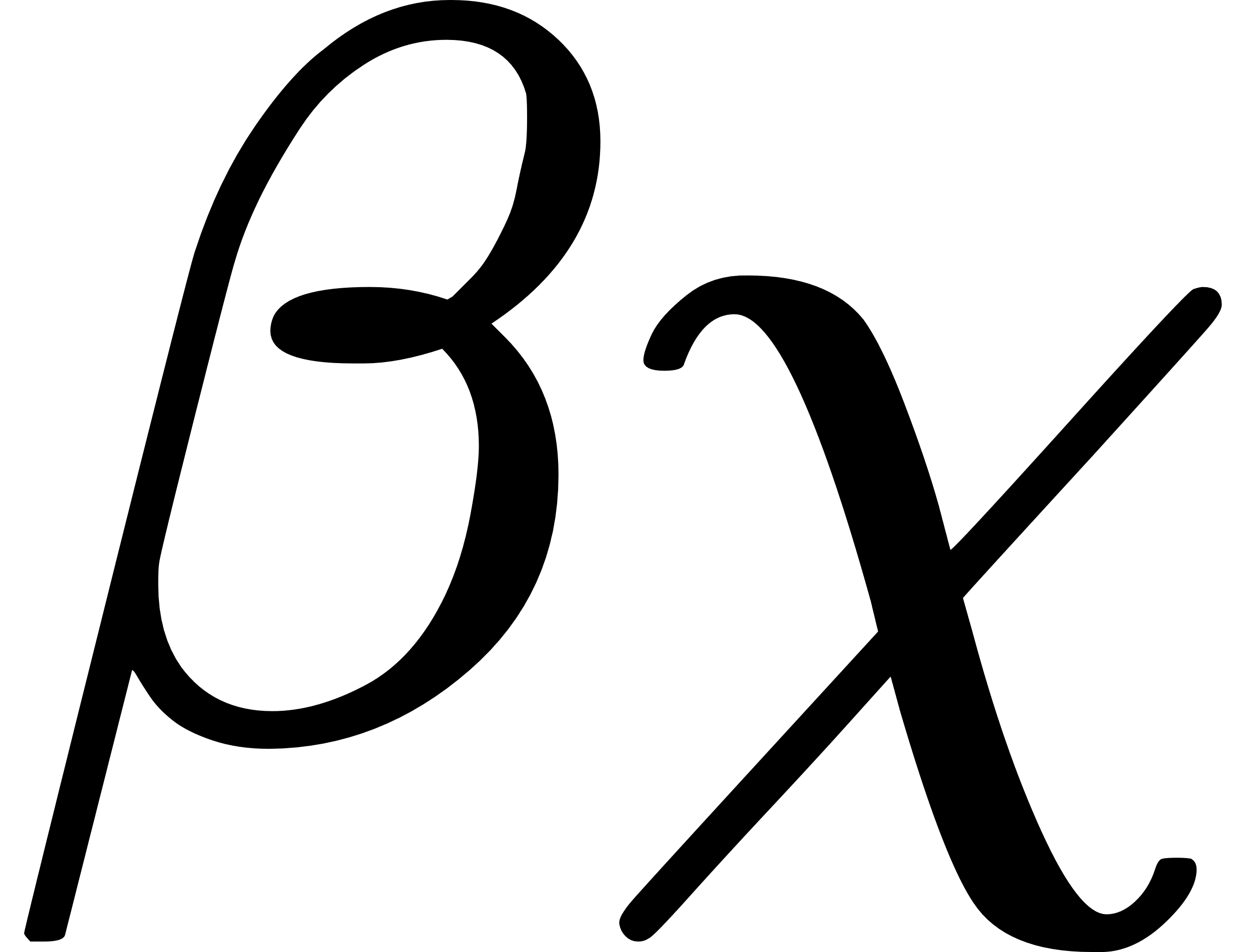
Công thức của mô hình FEM:

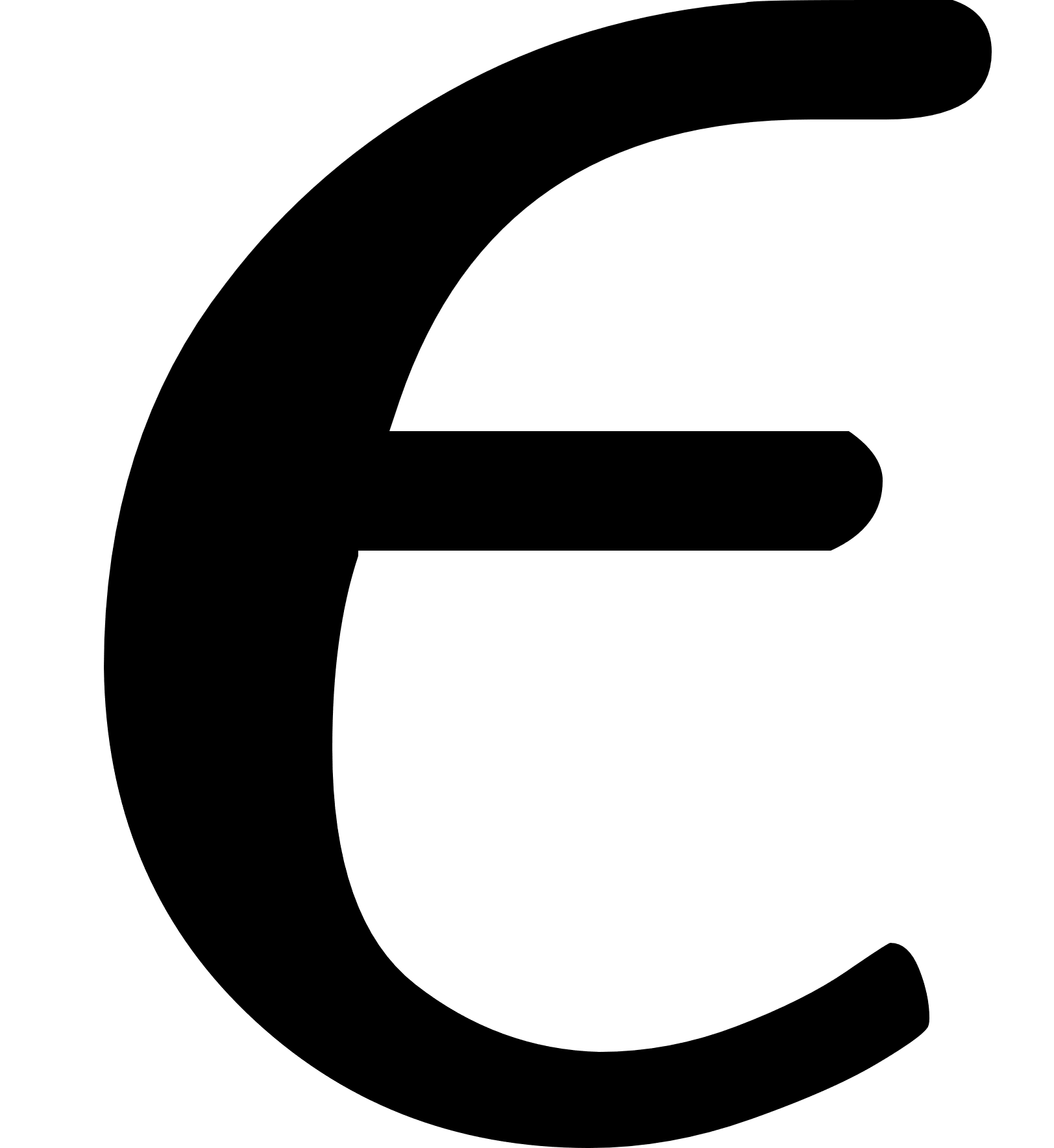


Trong đó:

- Yit : Giá trị của biến phụ thuộc tại đơn vị i và thời điểm t.

- I :  Sai số thành phần cố định riêng cho từng đơn vị i.

- it : Hệ số hồi quy, biểu diễn mức độ ảnh hưởng của biến độc lập lên biến phụ thuộc  
- Xit: Giá trị của biến độc lập tại đơn vị i và thời điểm t.

- it:  Sai số ngẫu nhiên với trung bình bằng 0 và phương sai không đổi.

Ưu điểm và nhược điểm của mô hình

- Ưu điểm: FEM giúp kiểm soát các yếu tố không quan sát được giữa các đơn vị, giảm thiểu sai lệch trong kết quả phân tích. Phù hợp khi các yếu tố không quan sát được này có thể gây nhiễu đến mối quan hệ giữa biến độc lập và biến phụ thuộc.

- Nhược điểm: FEM không phân tích các biến không thay đổi theo thời gian, vì các biến cố định bị loại bỏ trong quá trình ước lượng. Phương pháp này có thể làm giảm sự phong phú của dữ liệu, vì các biến không thay đổi theo thời gian trong từng đơn vị không được tính vào mô hình.

### 2.3.2. Hiệu ứng ngẫu nhiên (Random Effects)

Mô hình tác động ngẫu nhiên (Random Effects Model - REM) là một phương pháp phân tích dữ liệu bảng (panel data), trong đó các yếu tố ảnh hưởng đến dữ liệu không được xem là cố định mà là ngẫu nhiên. Mô hình này đặc biệt phù hợp khi các đặc điểm riêng biệt của từng đối tượng nghiên cứu được coi là ngẫu nhiên và không tương quan với các biến giải thích trong mô hình.

Công thức:



Trong đó:

- Yit : Giá trị của biến phụ thuộc tại đơn vị i và thời điểm t.

- C :  Hằng số chung cho tất cả các đối tượng.

- wpsit : Hệ số hồi quy, biểu diễn mức độ ảnh hưởng của biến độc lập lên biến phụ thuộc  
- Uit: Sai số riêng của từng quan sát tại từng thời điểm

- wpsit:  Sai số ngẫu nhiên với trung bình bằng 0 và phương sai không đổi.

Ưu điểm của mô hình

- Ưu điểm: Mô hình REM có thể sử dụng tốt khi số lượng đối tượng trong dữ liệu lớn và các đặc điểm ngẫu nhiên của đối tượng có thể thay đổi theo thời gian. Khi không có sự tương quan giữa các yếu tố ngẫu nhiên và các biến giải thích, REM cho ra kết quả ước lượng hiệu quả và không chệch.

## 2.4. Các giả định và kiểm định trong hồi quy

### 2.4.1. Đa cộng tuyến, phương sai sai số thay đổi

**a. Đa cộng tuyến**   
 \* *Định nghĩa:*

Cộng tuyến và đa cộng tuyến là hai thuật ngữ thường được sử dụng trong thống kê, đặc biệt là trong phân tích hồi quy. Cộng tuyến và 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.

- Cộng tuyến (Collinearity): Cộng tuyến xảy ra khi hai biến độc lập có mối tương quan tuyến tính quá mạnh với nhau.

- Đa cộng tuyến (Multicollinearity): Đa cộng tuyến xảy ra khi ba biến độc lập trở lên có tương quan tuyến tính quá mạnh với nhau. Nó là một trường hợp tổng quát hơn của cộng tuyến.

*\* Nguyên nhân:*

- Do khi khảo sát chúng ta sử dụng nhưng thuộc tính không khác biết nhau về tính chất ý nghĩa.

- Đặc điểm của môi trường khảo sát có thể dẫn đến hiện tượng đa cộng tuyến. Cùng một bảng khảo sát, môi trường thứ nhất không gây đa cộng tuyến nhưng môi trường thứ hai lại xuất hiện hiện tượng này. Dù theo lý luận ban đầu các nhân tố không tương quan cao, nhưng trong thực tế tại môi trường thứ hai, chúng lại phát sinh mối tương quan mạnh, dẫn đến đa cộng tuyến khi thu thập dữ liệu.

*\* Hậu quả*

- Hệ số hồi quy thiếu ổn định.

- Sai số chuẩn của hệ số hồi quy tăng.

- Kết quả kiểm định ý nghĩa của hệ số hồi quy không đáng tin cậy.

- Gây khó khăn trong việc xác định biến độc lập quan trọng.

- Có thể làm thay đổi chiều tác động của biến độc lập.

*\* cách khắc phục:* Sử dụng các kiểm định như VIF để Đo lường mức độ cộng tuyến giữa các biến độc lập. Nếu VIF của một biến vượt quá 10, có thể có đa cộng tuyến. Bằng cách loại bỏ hoặc kết hợp các biến có sự tương quan mạnh.

**b. Phương sai sai số thay đổi**

*\* Định nghĩa*

Phương sai sai số thay đổi là hiện tượng xảy ra khi phần dư (residuals) hoặc sai số (e) trong mô hình hồi quy không có phương sai đồng nhất và không phân bố ngẫu nhiên. Nói cách khác, phương sai của sai số thay đổi giữa các quan sát.

*\* Nguyên nhân*

- Sai số của mô hình không đồng đều trên toàn bộ tập dữ liệu.

- Mức độ phân tán của biến phụ thuộc khác nhau giữa các nhóm dữ liệu.

- Tác động của các biến bị bỏ sót nhưng có ảnh hưởng đến biến phụ thuộc.

- Tồn tại tương quan giữa các biến độc lập.

- Sự khác biệt về mức độ ảnh hưởng của từng quan sát lên biến phụ thuộc.

*\* Hậu quả*

- Phương sai thay đổi gây ảnh hưởng tiêu cực đến mô hình hồi quy và quá trình ước lượng.

- Dù không làm mất tính không chệch và nhất quán của ước lượng OLS, nhưng làm giảm hiệu quả của nó.

- Ước lượng OLS không còn là ước lượng tuyến tính không chệch tốt nhất (BLUE), chỉ còn là LUE.

- Kiểm định t và F trở nên không đáng tin cậy nếu có phương sai thay đổi.

- Dẫn đến nguy cơ đưa ra kết luận sai về ý nghĩa thống kê của các hệ số hồi quy.

*\* cách khắc phục:* thông qua các kiểm định Breusch-Pagan, White để Kiểm tra phương sai sai số thay đổi bằng cách biến đổi dữ liệu, robust standard errors để cải thiện độ chính xác và giảm phương sai sai số thay đổi.

### 2.4.2. Tự tương quan và phân phối chuẩn của phần dư

**a. Tự tương quan**

*\* Định nghĩa:*

Tự tương quan là hiện tượng mà các phần dư (residuals) của mô hình hồi quy có tương quan với nhau, tức là sai số tại thời điểm này có thể bị ảnh hưởng bởi sai số ở thời điểm trước. Hiện tượng này phổ biến trong các dữ liệu chuỗi thời gian.

*\* Nguyên nhân*

- Dữ liệu có tính chuỗi thời gian, trong đó các giá trị có xu hướng phụ thuộc vào các giá trị trước đó.

- Mô hình bị thiếu biến quan trọng, làm cho phần dư mang thông tin của các yếu tố chưa được đưa vào mô hình.

- Sai số mô hình không được xác định đúng (misspecification).

*\* Hậu quả*

- Làm sai lệch sai số chuẩn của các ước lượng.

- Các kiểm định t và F trở nên không đáng tin cậy.

- Có thể dẫn đến kết luận sai về ý nghĩa thống kê của các hệ số hồi quy.

- Mô hình không còn hiệu quả, giảm độ tin cậy trong dự báo. *\* Cách khắc phục: Thông qua k***iểm định Breusch-Godfrey** cho phép kiểm tra tự tương quan bậc cao. Để khắc phục, có thể áp dụng:

- Sử dụng phương pháp hồi quy tổng quát nhất quán (Generalized Least Squares - GLS) hoặc Cochrane-Orcutt để điều chỉnh mô hình.

- Dùng phương pháp ước lượng sai số chuẩn vững (robust standard errors).

**b. Phân phối chuẩn của phần dư**

*\* Định nghĩa*

Phân phối chuẩn của phần dư là giả định rằng các phần dư trong mô hình hồi quy tuyến tính phải tuân theo phân phối chuẩn. Giả định này quan trọng cho các kiểm định suy diễn như kiểm định t và F.

*\* Nguyên nhân*

- Dữ liệu có nhiều giá trị ngoại lai (outliers) hoặc bị lệch.

- Sai số không được phân bố ngẫu nhiên.

- Dữ liệu không đủ lớn (sample size nhỏ) nên phân phối không tiệm cận chuẩn.

- Thiếu biến quan trọng hoặc mô hình sai dạng hàm.

*\* Hậu quả*

- Các kiểm định t, F không còn chính xác vì dựa trên giả định phân phối chuẩn.

- Có thể dẫn đến việc đưa ra các kết luận sai về ý nghĩa thống kê.

- Mô hình mất tính chính xác trong việc suy luận thống kê, nhất là với mẫu nhỏ. *\* Cách khắc phục:* Khi kiểm định Shapiro-Wilk cho thấy phần dư không tuân theo phân phối chuẩn, có thể áp dụng các biện pháp sau để cải thiện:

- Tăng kích thước mẫu (với luật số lớn, phần dư tiệm cận phân phối chuẩn).

- Loại bỏ hoặc xử lý các giá trị ngoại lai.

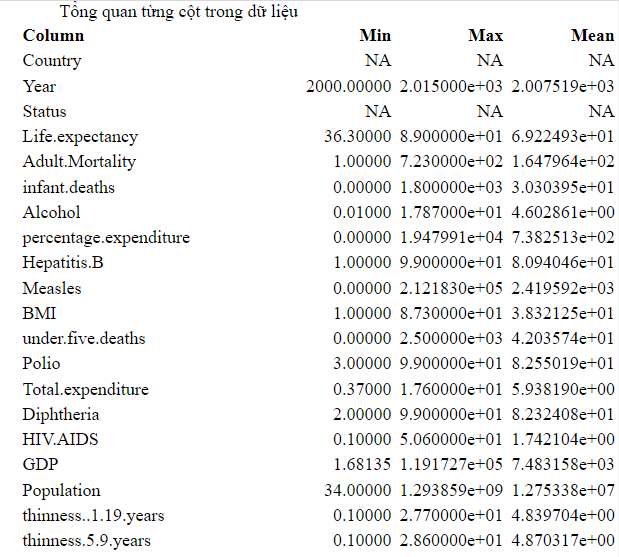
- Thực hiện biến đổi dữ liệu (như log, căn bậc hai, Box-Cox).

- Sử dụng các phương pháp phi tham số hoặc ước lượng vững khi phân phối chuẩn không đạt được.

# CHƯƠNG 3 TIỀN XỬ LÝ DỮ LIỆU VÀ XÂY DỰNG MÔ HÌNH

## 3.1 Tổng quan về dữ liệu

Dữ liệu được sử dụng trong nghiên cứu này có nguồn gốc từ tổ chức y tế thế giới (WHO), chứa thông tin về các chỉ số y tế, kinh tế, môi trường và xã hội của nhiều quốc gia trong khoảng thời gian từ năm 2000 đến 2015. Mục tiêu chính của việc sử dụng bộ dữ liệu này là để tìm hiểu các yếu tố có ảnh hưởng đến tuổi thọ trung bình của một quốc gia.



*Bảng tổng quan về dữ liệu của dataset*

Theo bảng trên có thể thấy được đặc điểm của tập dữ liệu như sau:

*Tuổi thọ trung bình (Life.expectancy):*

* Dao động từ 36.3 năm đến 89 năm, với giá trị trung bình là 69.2 năm. Sự chênh lệch lớn này cho thấy sự khác biệt đáng kể về điều kiện sống, y tế và kinh tế giữa các quốc gia.

*Các chỉ số tử vong:*

* Tử vong người lớn (Adult.Mortality): Trung bình 164.8 (từ 1 đến 723), phản ánh tình trạng sức khỏe cộng đồng không đồng đều.
* Tử vong trẻ dưới 5 tuổi (Under.five.deaths): Trung bình 42.0 nhưng có quốc gia lên đến 2,500 ca, cho thấy thách thức về chăm sóc sức khỏe trẻ em ở một số khu vực.

*Chỉ số kinh tế và y tế:*

* GDP: Trung bình 7,483 nhưng chênh lệch cực lớn (từ 1.68 đến 119,172.7), điều này giải thích sự phân bố lệch phải trong biểu đồ tần suất GDP, nơi phần lớn quốc gia tập trung ở mức thấp.
* Chi tiêu y tế (Percentage.expenditure): Trung bình 738.25, nhưng có quốc gia chi tiêu lên đến 19,479.91, thể hiện sự đầu tư không đồng đều vào hệ thống y tế.

*Tỷ lệ tiêm chủng:*

* Hepatitis B, Polio, Diphtheria có trung bình khoảng 80-82%, nhưng một số quốc gia có tỷ lệ rất thấp (min = 1-3%). Điều này cho thấy nỗ lực tiêm chủng chưa đồng đều, cần xử lý giá trị thiếu và outliers để đảm bảo phân tích chính xác.

*Dinh dưỡng và dịch bệnh:*

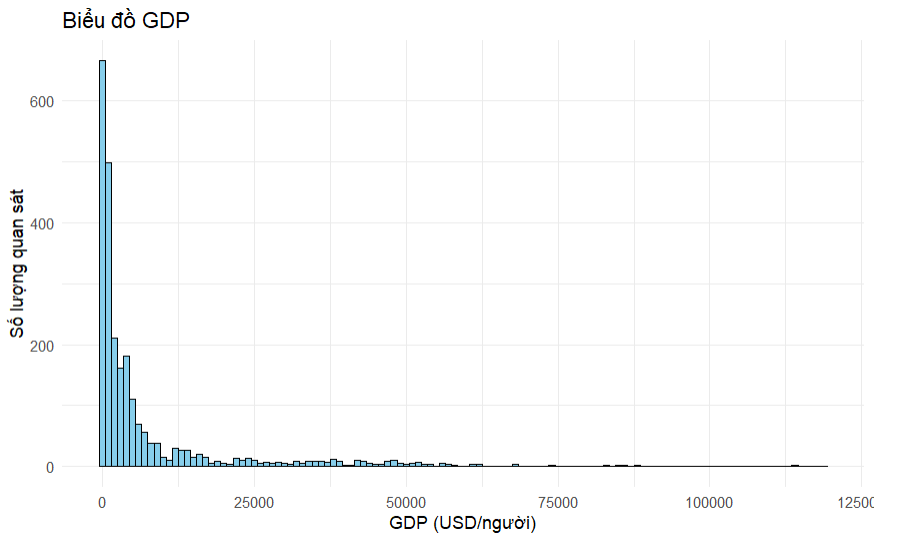
* BMI trung bình là 38.3, nhưng có quốc gia chỉ đạt 1.0, phản ánh tình trạng suy dinh dưỡng hoặc thiếu dữ liệu.
* HIV/AIDS trung bình 1.74%, nhưng có quốc gia lên đến 50.6%, cho thấy gánh nặng dịch bệnh nghiêm trọng ở một số khu vực.

*Biến phân loại:*

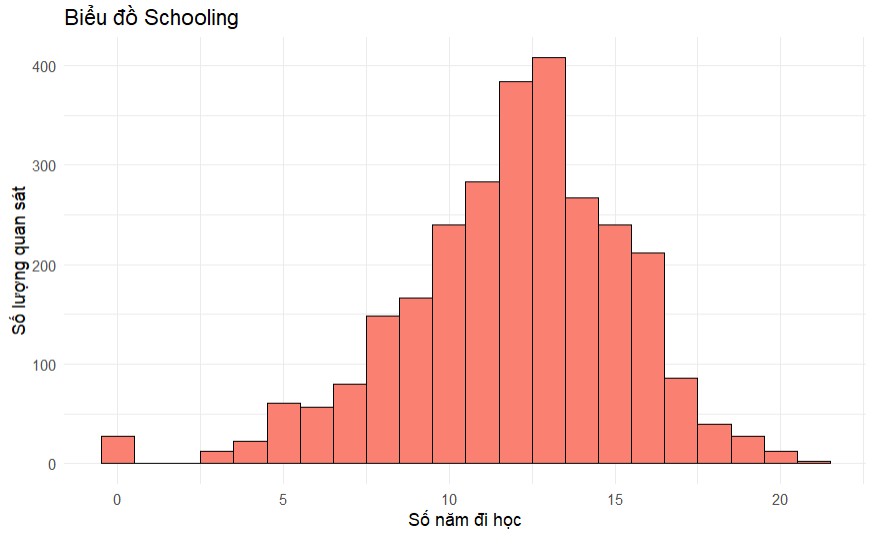
* Year (2000–2015) và Status (phát triển/đang phát triển) giúp phân tích xu hướng theo thời gian và nhóm quốc gia.

Tổng thể, dữ liệu có phạm vi rộng với hơn 2.000 dòng quan sát từ hơn 150 quốc gia khác nhau, đại diện cho cả nhóm nước phát triển và đang phát triển, bao quát nhiều khía cạnh của đời sống xã hội và sức khỏe cộng đồng. Tuy nhiên, tập dữ liệu vẫn tồn tại một số vấn đề như giá trị thiếu (NA), nhiễu và ngoại lệ (outlier), do đó cần được xử lý làm sạch và chuẩn hóa cẩn thận trước khi đưa vào phân tích mô hình.

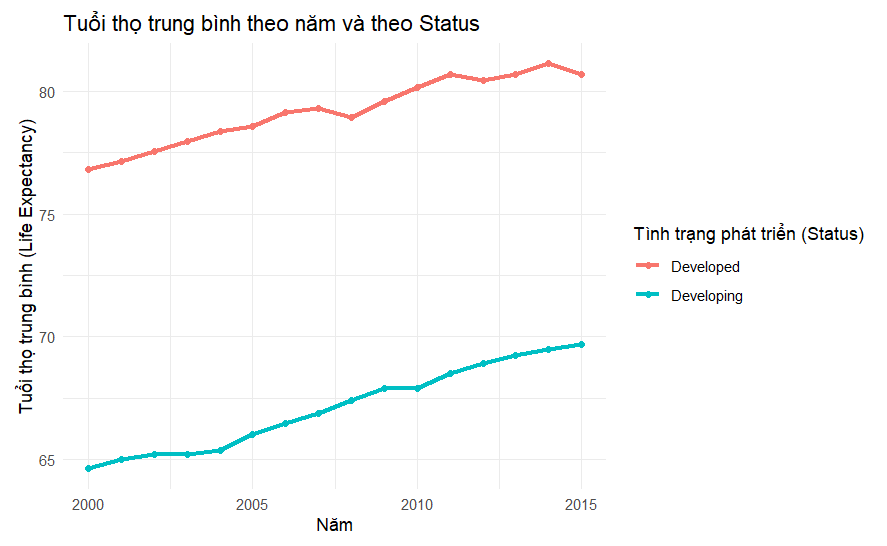
Một trong những phương pháp giúp hiểu rõ hơn về phân phối và sự phân bố của các biến trong dữ liệu là vẽ biểu đồ. Chúng ta sẽ vẽ biểu đồ cho một số biến tiêu biểu trong dữ liệu để khám phá sự phân bố và nhận diện các đặc điểm quan trọng của các yếu tố như GDP, số năm đi học, và các yếu tố khác có ảnh hưởng đến tuổi thọ trung bình của các quốc gia. Việc này giúp ta có cái nhìn trực quan về dữ liệu và hỗ trợ trong việc xử lý các vấn đề về giá trị thiếu hay ngoại lệ trước khi tiếp tục với các bước phân tích sâu hơn.



Biểu đồ tần suất GDP cho thấy sự phân bố không đồng đều giữa các quốc gia, với phần lớn quốc gia tập trung ở mức GDP thấp. Điều này phản ánh sự chênh lệch lớn về kinh tế giữa các nước trong tập dữ liệu, và có thể là một yếu tố ảnh hưởng đáng kể đến tuổi thọ trung bình.



Biểu đồ tần suất số năm đi học (Schooling) cho thấy hầu hết quốc gia có trung bình khoảng 10–15 năm học. Tuy nhiên, vẫn tồn tại một số quốc gia có mức giáo dục trung bình rất thấp, điều này có thể ảnh hưởng tiêu cực đến nhận thức về sức khỏe, tiếp cận dịch vụ y tế, và qua đó ảnh hưởng đến tuổi thọ.



Biểu đồ đường thể hiện sự thay đổi tuổi thọ trung bình qua các năm, phân theo tình trạng phát triển của quốc gia. Nhìn chung, tuổi thọ có xu hướng tăng dần trong giai đoạn 2000–2015 ở cả hai nhóm nước phát triển và đang phát triển. Tuy nhiên, nhóm nước phát triển duy trì mức tuổi thọ cao hơn rõ rệt, phản ánh sự khác biệt về điều kiện sống và hệ thống y tế.

## 3.2. Quy trình tiền xử lý dữ liệu

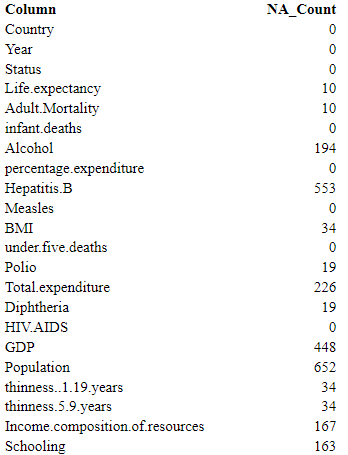
### 3.2.1. Làm sạch và chuẩn hóa dữ liệu

Sau khi thực hiện tổng quan bộ dữ liệu, có thể nhận thấy rằng dữ liệu ban đầu còn tồn tại một số vấn đề như giá trị thiếu (NA), tên biến chưa đồng nhất, định dạng chưa chuẩn hóa và sự xuất hiện của các giá trị ngoại lai (outliers). Những vấn đề này nếu không được xử lý kỹ lưỡng sẽ ảnh hưởng trực tiếp đến hiệu quả của các mô hình phân tích và dự đoán, đặc biệt là mô hình hồi quy tuyến tính.

Do đó, để đảm bảo tính chính xác và tin cậy cho mô hình, cần tiến hành quy trình tiền xử lý dữ liệu bao gồm các bước: làm sạch dữ liệu (xử lý giá trị thiếu, loại bỏ hoặc thay thế outliers), chuẩn hóa tên biến và định dạng, mã hóa biến phân loại, cũng như gộp các biến có ý nghĩa tương đồng. Các bước này không chỉ giúp dữ liệu trở nên nhất quán và đầy đủ hơn mà còn hỗ trợ quá trình huấn luyện mô hình đạt kết quả tối ưu.

***Đầu tiên, chúng ta sẽ xử lý dữ liệu thiếu (NA)***

Trong bộ dữ liệu ban đầu, nhiều biến dạng số có chứa giá trị bị thiếu (NA), đặc biệt là các chỉ số kinh tế, y tế hoặc dân số theo từng năm và từng quốc gia. Việc giữ nguyên các giá trị NA có thể làm sai lệch kết quả phân tích, trong khi loại bỏ toàn bộ các dòng có NA lại gây mất mát dữ liệu nghiêm trọng, dẫn đến giảm độ tin cậy của mô hình.



*Số lượng giá trị NA của từng biến trước xử lý*

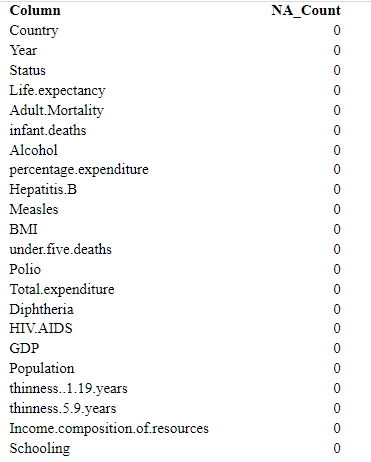
Để giải quyết vấn đề này một cách hợp lý, nhóm em đã chọn phương pháp thay thế các giá trị thiếu bằng trung vị (median) theo từng quốc gia.

* Dữ liệu được nhóm theo biến Country, đảm bảo việc xử lý được thực hiện độc lập trên từng quốc gia.
* Với mỗi biến dạng số trong nhóm đó, nếu có giá trị bị thiếu, hệ thống sẽ tính trung vị của toàn bộ giá trị hiện có (không NA) trong cùng quốc gia để thay thế cho các NA.
* Việc dùng trung vị thay vì trung bình cộng giúp giảm ảnh hưởng của các giá trị ngoại lai, vốn có thể làm lệch trung bình và dẫn đến sự thay thế không chính xác.

Phương pháp này có một số ưu điểm:

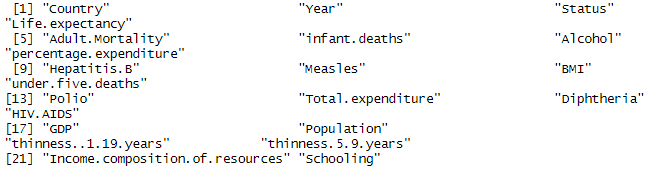
* Giữ lại nhiều dữ liệu hơn so với việc loại bỏ dòng.
* Bảo toàn đặc trưng riêng của từng quốc gia, vì dữ liệu được xử lý nội bộ theo từng nhóm.
* Tạo ra một bộ dữ liệu đầy đủ và có thể sử dụng ngay cho các bước phân tích tiếp theo như chuẩn hóa, trực quan hóa và huấn luyện mô hình hồi quy.

Sau khi xử lý, các giá trị NA của dataset đã được lấp đầy một cách có kiểm soát, đảm bảo cân bằng giữa tính đầy đủ dữ liệu và tính đại diện thống kê.



*Số lượng giá trị NA của từng biến sau xử lý*

***Tiếp theo chúng ta sẽ chuẩn hóa dữ liệu*** nhằm đảm bảo tính nhất quán và thuận tiện cho việc phân tích, trực quan hóa cũng như xây dựng mô hình hồi quy.



*Tên các biến trước chuẩn hóa*

*Chuẩn hóa tên các cột*

Tên các biến trong tập dữ liệu ban đầu có định dạng không đồng nhất (chữ thường, chữ hoa, ký tự đặc biệt,...). Điều này gây khó khăn trong quá trình truy xuất và xử lý biến. Vì vậy, toàn bộ tên biến được chuyển đổi thành chữ in hoa (uppercase) để đảm bảo thống nhất và dễ thao tác sau này.

*Chuẩn hóa tên quốc gia*

Tên các quốc gia trong cột COUNTRY chứa khoảng trắng (ví dụ: "South Korea", "New Zealand"), gây bất tiện khi xử lý hoặc vẽ biểu đồ. Do đó, tên quốc gia được chuẩn hóa bằng cách thay thế toàn bộ khoảng trắng bằng dấu gạch dưới (\_), ví dụ: "South\_Korea", "New\_Zealand".

*Chuẩn hóa biến phân loại*

Biến STATUS ban đầu có hai giá trị phân loại là "Developed" và "Developing". Để thuận tiện cho các mô hình phân tích định lượng, biến này được mã hóa lại thành biến nhị phân:

* "Developed" được gán giá trị 1
* "Developing" được gán giá trị 0

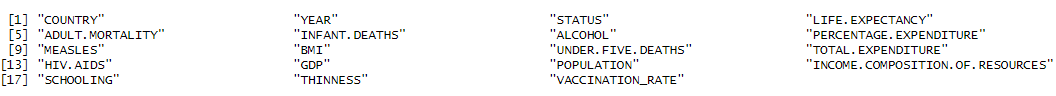
Việc chuyển đổi này giúp mô hình hồi quy có thể xử lý biến phân loại như một biến số.

*Gộp các biến tương đương về mặt ý nghĩa*

Trong bộ dữ liệu ban đầu, có một số cặp biến thể hiện cùng một khái niệm nhưng dưới nhiều nhóm tuổi hoặc dạng khác nhau. Để đơn giản hóa mô hình và giảm đa cộng tuyến, các biến này được gộp lại thành một biến trung bình duy nhất. Cụ thể:

* Hai biến THINNESS..1.19.YEARS và THINNESS.5.9.YEARS được gộp thành một biến mới là THINNESS, đại diện cho mức độ suy dinh dưỡng chung của trẻ em.
* Ba biến về tỷ lệ tiêm chủng là HEPATITIS.B, POLIO và DIPHTHERIA được gộp lại thành một biến VACCINATION\_RATE, phản ánh tỷ lệ tiêm chủng tổng hợp.

Sau khi gộp, các cột ban đầu không còn cần thiết nên đã được loại bỏ khỏi tập dữ liệu để tránh trùng lặp thông tin và giảm số chiều của dữ liệu.



*Tên các biến sau chuẩn hóa*

***Cuối cùng chúng ta sẽ kiểm tra và xử lý các giá trị outlier***

Outlier (giá trị ngoại lai) là những điểm dữ liệu có giá trị khác biệt lớn so với phần lớn các quan sát trong tập dữ liệu. Sự xuất hiện của outlier có thể bắt nguồn từ lỗi nhập liệu, sai số trong đo lường hoặc đơn giản là phản ánh các hiện tượng bất thường trong thực tế. Nếu không được phát hiện và xử lý đúng cách, các outlier có thể gây sai lệch trong phân tích thống kê, làm giảm độ chính xác và độ tin cậy của các mô hình dự báo, đặc biệt là mô hình hồi quy tuyến tính.

Để phát hiện outlier, nghiên cứu sử dụng phương pháp dựa trên khoảng tứ phân vị (Interquartile Range – IQR). Đây là một phương pháp thống kê phổ biến, được sử dụng để đo độ phân tán của dữ liệu và xác định các điểm dữ liệu bất thường.

Cụ thể:

* Q1 (First Quartile): Là giá trị tại vị trí 25% dữ liệu đầu tiên khi sắp xếp theo thứ tự tăng dần.
* Q3 (Third Quartile): Là giá trị tại vị trí 75% dữ liệu.
* IQR: Là độ dài khoảng tứ phân vị, được tính bằng công thức:

***IQR = Q3 − Q1***

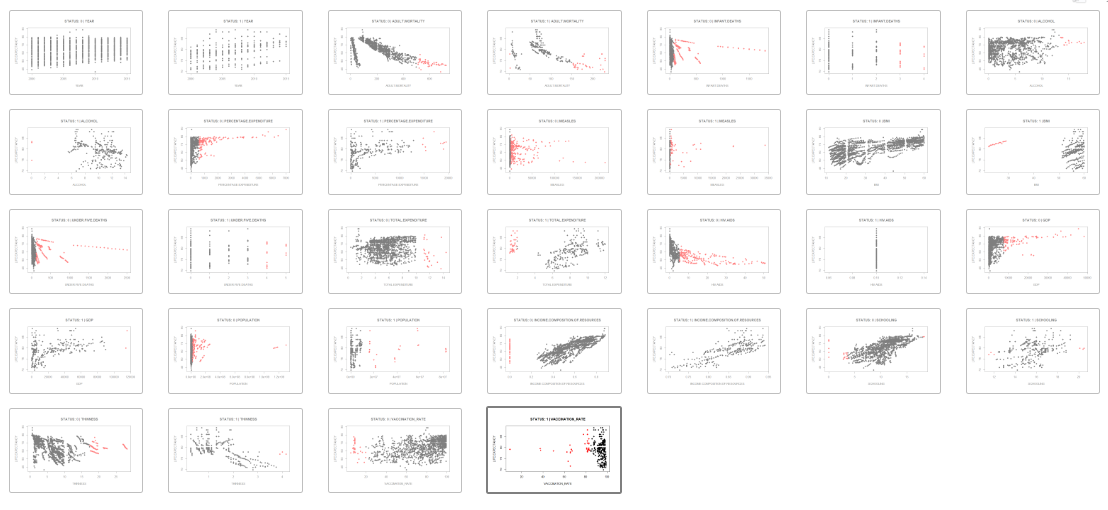
* Ngưỡng phát hiện outlier:
  + Giới hạn dưới (Lower bound): Q1 − 1.5 × IQR
  + Giới hạn trên (Upper bound): Q3 + 1.5 × IQR

Mọi giá trị nằm ngoài khoảng từ giới hạn dưới đến giới hạn trên sẽ được xem là outlier.

Phương pháp IQR có ưu điểm là không bị ảnh hưởng bởi outlier vì nó dựa trên tứ phân vị (thay vì trung bình và độ lệch chuẩn), nên rất phù hợp để phát hiện những giá trị ngoại lai trong dữ liệu lệch.

Trong dataset này, để kiểm soát chặt chẽ hơn, các biến được kiểm tra outlier riêng biệt theo từng nhóm STATUS (Developed hoặc Developing). Cụ thể:

* Với mỗi biến định lượng (loại trừ COUNTRY, STATUS, và biến mục tiêu LIFE.EXPECTANCY), thực hiện:
  + Phân nhóm dữ liệu theo biến STATUS.
  + Tính Q1, Q3 và IQR cho từng nhóm.
  + Xác định giới hạn dưới và trên để phát hiện các outlier.
  + Tạo scatter plot trực quan giữa từng biến và LIFE.EXPECTANCY, trong đó các outlier được đánh dấu màu đỏ, giúp người phân tích dễ dàng nhận diện mối liên hệ và mức độ ảnh hưởng.



*Tổng quan về biểu đồ scatter plot trực quan giữa từng biến và LIFE.EXPECTANCY*

Sau khi kiểm tra outlier, quá trình xử lý outlier được thực hiện theo chiến lược riêng biệt cho từng quốc gia. Lý do là vì mỗi quốc gia có điều kiện kinh tế, xã hội, y tế khác nhau, việc thay đổi dữ liệu theo một chuẩn chung có thể làm sai lệch đặc điểm nội tại của từng nước.

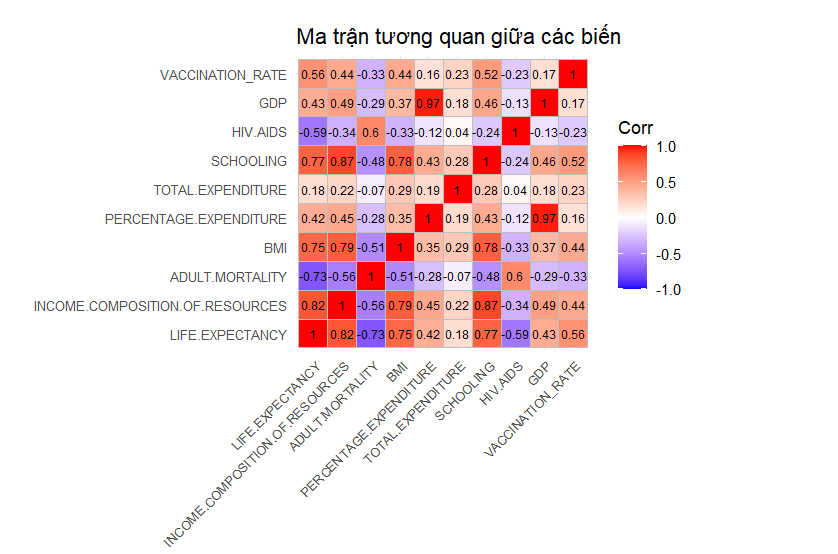
Quy trình cụ thể như sau:

* Với mỗi biến định lượng, nhóm dữ liệu theo COUNTRY.
* Trong từng quốc gia:
  + Tính Q1, Q3, IQR và ngưỡng phát hiện outlier.
  + Xác định các giá trị bất thường (bao gồm cả các giá trị bằng 0 nếu không hợp lý).
  + Thay thế outlier bằng giá trị trung vị (median) của biến đó trong chính quốc gia đó.

Cách xử lý này giúp làm sạch dữ liệu mà vẫn giữ nguyên các đặc trưng của từng quốc gia, đặc biệt quan trọng trong phân tích tuổi thọ – một biến chịu ảnh hưởng bởi rất nhiều yếu tố.

## 3.3 Phương pháp chọn biến độc lập

### 3.3.1. Ma trận tương quan



Biến phụ thuộc LIFE.EXPECTANCY cho thấy mối tương quan đáng kể (|r| > 0.3) với nhiều biến độc lập, phản ánh các yếu tố ảnh hưởng quan trọng đến tuổi thọ. Cụ thể:

* INCOME.COMPOSITION.OF.RESOURCES (r = 0.82): Có tương quan dương mạnh với tuổi thọ, cho thấy thu nhập đóng vai trò then chốt trong việc cải thiện chất lượng sống và sức khỏe cộng đồng.
* SCHOOLING (r = 0.77): Cũng có tương quan dương cao, phản ánh vai trò của giáo dục trong việc nâng cao nhận thức và hành vi chăm sóc sức khỏe.
* BMI (r = 0.75): Chỉ số dinh dưỡng có liên hệ tích cực với tuổi thọ, cho thấy tầm quan trọng của chế độ ăn uống và tình trạng thể chất.
* ADULT.MORTALITY (r = -0.67): Có tương quan âm mạnh, điều này phù hợp vì tỷ lệ tử vong cao làm giảm tuổi thọ trung bình.
* HIV.AIDS (r = -0.73): Tương quan âm rõ rệt, phản ánh tác động tiêu cực của dịch bệnh tới sức khỏe cộng đồng và tuổi thọ.
* Các biến khác như GDP (r = 0.43), VACCINATION\_RATE (r = 0.56) và PERCENTAGE.EXPENDITURE (r = 0.42) có tương quan dương từ trung bình đến yếu, nhưng vẫn thể hiện vai trò hỗ trợ trong việc cải thiện tuổi thọ.
* Một số biến như TOTAL.EXPENDITURE (r = 0.18) có tương quan rất yếu với tuổi thọ, do đó có thể được xem xét loại bỏ khỏi mô hình để giảm nhiễu.

Ngoài ra, tồn tại một số cặp biến độc lập có tương quan cao với nhau, có khả năng gây ra hiện tượng đa cộng tuyến trong mô hình hồi quy. Cần lưu ý:

* SCHOOLING và INCOME.COMPOSITION.OF.RESOURCES (r = 0.87): Các quốc gia có thu nhập cao thường đi đôi với đầu tư vào giáo dục, dẫn đến sự chồng lấn về ảnh hưởng giữa hai biến.
* SCHOOLING và GDP (r = 0.80): GDP cao thường gắn liền với hệ thống giáo dục phát triển, tạo nguy cơ đa cộng tuyến.

Những cặp biến này cần được kiểm tra thêm qua các chỉ số như VIF để quyết định giữ lại hay loại bỏ trong mô hình cuối cùng.

### 3.3.2 Stepwise AIC

**Stepwise Regression** là một kỹ thuật trong phân tích thống kê, đặc biệt hữu ích trong quá trình xây dựng mô hình hồi quy. Phương pháp này liên quan đến việc lựa chọn một tập hợp con của các biến giải thích, từ một tập hợp lớn hơn các biến tiềm năng, để tạo ra một mô hình hồi quy mạnh mẽ và hiệu quả.

*Định nghĩa về Stepwise Regression*

Stepwise Regression là một phương pháp chọn biến trong hồi quy tuyến tính, trong đó các biến độc lập được thêm vào hoặc loại bỏ khỏi mô hình dựa trên các tiêu chí thống kê nhất định như AIC (Akaike Information Criterion) hoặc BIC (Bayesian Information Criterion), nhằm tối ưu hóa độ phù hợp của mô hình. Phương pháp này có thể được thực hiện thông qua ba kỹ thuật chính: **Forward Selection** (lựa chọn tiến), **Backward Elimination** (loại bỏ ngược), và **Bidirectional Elimination** (kết hợp cả hai). Với Forward Selection, quá trình bắt đầu từ mô hình rỗng (không có biến nào), sau đó thêm từng biến một dựa trên mức ý nghĩa thống kê. Ngược lại, Backward Elimination khởi đầu với mô hình đầy đủ, rồi dần loại bỏ các biến ít quan trọng để kiểm tra ảnh hưởng của chúng đến kết quả tổng thể. Phương pháp kết hợp sẽ thực hiện cả thêm và loại biến trong cùng một quy trình. Mặc dù Stepwise Regression giúp đơn giản hóa mô hình và loại bỏ các biến không cần thiết, nhưng nó cũng tồn tại những nhược điểm, chẳng hạn như dễ dẫn đến **overfitting**, bỏ sót các tương tác quan trọng giữa các biến, hoặc kết quả bị ảnh hưởng mạnh bởi các thay đổi nhỏ trong dữ liệu. Do đó, đây là một phương pháp mang tính kinh nghiệm và cần được áp dụng cẩn trọng, kết hợp với hiểu biết chuyên môn và phân tích thống kê sâu hơn.

*Tầm quan trọng của việc chọn lựa các biến*

Việc lựa chọn đúng đắn các biến giải thích trong mô hình hồi quy đóng vai trò quan trọng vì nó ảnh hưởng trực tiếp đến hiệu quả, độ chính xác và tính khả thi của mô hình. Nếu mô hình bao gồm quá nhiều biến không cần thiết, nguy cơ xảy ra hiện tượng overfitting sẽ tăng cao, khiến mô hình hoạt động kém hiệu quả khi áp dụng vào dữ liệu mới. Ngược lại, việc loại bỏ những biến quan trọng có thể làm giảm khả năng dự đoán và độ chính xác của mô hình. Trong bối cảnh đó, Stepwise Regression trở thành một công cụ hữu ích giúp xác định các biến có ảnh hưởng đáng kể đến biến phụ thuộc. Bằng cách tự động thêm vào hoặc loại bỏ các biến dựa trên các tiêu chí thống kê, phương pháp này hỗ trợ các nhà nghiên cứu và nhà phân tích xây dựng những mô hình vừa tinh gọn vừa có giá trị khoa học và thực tiễn cao.

## 3.4 Xây dựng mô hình hồi quy

### 3.4.1. Hồi quy OLS

Phương pháp OLS là một phương pháp thống kê được sử dụng để tìm ra mô hình hồi quy tuyến tính tốt nhất cho dữ liệu. Nó áp dụng nguyên lý tối thiểu hóa sai số bình phương giữa giá trị dự đoán và giá trị thực tế của biến phụ thuộc. Trong OLS, chúng ta cố gắng tìm ra đường hồi quy tuyến tính mà tổng các sai số bình phương là nhỏ nhất.

***Ý nghĩa và ứng dụng của phương pháp OLS trong nghiên cứu***

Phương pháp OLS có vai trò quan trọng trong việc phân tích và dự đoán dữ liệu. Nó được sử dụng rộng rãi trong các lĩnh vực như kinh tế học, tài chính, khoa học xã hội và nghiên cứu y học. Với khả năng xác định mối quan hệ giữa biến độc lập và biến phụ thuộc, OLS giúp chúng ta hiểu rõ hơn về tác động của các yếu tố khác nhau và dự đoán kết quả trong tương lai.

*Các bước thực hiện phương pháp OLS*

- **Thu thập dữ liệu:** Bước đầu tiên là thu thập dữ liệu chính xác và đầy đủ. Việc này đảm bảo rằng chúng ta có đủ thông tin để xây dựng mô hình hồi quy tuyến tính và đánh giá kết quả.

- **Xây dựng mô hình hồi quy tuyến tính:** Sau khi có dữ liệu, chúng ta xây dựng mô hình hồi quy tuyến tính bằng cách tìm đường hồi quy tuyến tính tốt nhất. Điều này đảm bảo rằng mô hình của chúng ta phù hợp với dữ liệu đã thu thập và có khả năng dự đoán tốt.

- **Kiểm định các giả định của phương pháp OLS:** Phương pháp OLS đòi hỏi các giả định đặc biệt về dữ liệu, bao gồm sự độc lập tuyệt đối giữa các quan sát, tính tuyến tính, sai số đồng nhất và không tồn tại sai số thừa thảTrong bước này, chúng ta cần kiểm tra xem các giả định này có được thỏa mãn hay không.

- **Đánh giá và chọn lựa mô hình phù hợp:** Cuối cùng, chúng ta đánh giá mô hình hồi quy tuyến tính và chọn lựa mô hình phù hợp nhất. Điều này liên quan đến việc xem xét các chỉ số đánh giá như R-square, F-statistic và các giá trị p-value để đảm bảo rằng mô hình của chúng ta có khả năng giải thích dữ liệu tốt.

Lợi ích và hạn chế của phương pháp OLS

*Lợi ích của phương pháp OLS trong việc phân tích dữ liệu*

Phương pháp OLS mang lại nhiều lợi ích trong việc phân tích dữ liệu. Đầu tiên, nó là một phương pháp đơn giản và dễ hiểu, cho phép chúng ta xác định mối quan hệ giữa các biến một cách rõ ràng. Ngoài ra, OLS cũng cung cấp các ước lượng cho các tham số trong mô hình hồi quy, giúp chúng ta hiểu rõ hơn về tác động của các yếu tố khác nhau.

*Nhược điểm và hạn chế của phương pháp OLS*

Mặc dù phương pháp OLS có nhiều ưu điểm, nhưng cũng có một số hạn chế. Đầu tiên, OLS yêu cầu các giả định đặc biệt về dữ liệu, và nếu các giả định này không được thỏa mãn, kết quả của mô hình có thể không chính xác. Thứ hai, OLS không phù hợp khi dữ liệu có tính chất phi tuyến tính hoặc khi có nhiễu lớn trong dữ liệu.

### 3.4.2 Mô hình PLM (Panel Linear Model)

Mô hình PLM (Panel Linear Model) là một phương pháp hồi quy tuyến tính được sử dụng để phân tích dữ liệu bảng (panel data), tức là dữ liệu được thu thập từ nhiều đối tượng (như cá nhân, doanh nghiệp, quốc gia) qua nhiều thời điểm khác nhau. Mô hình này cho phép kiểm soát sự dị biệt không quan sát được giữa các đối tượng và đánh giá tác động của các biến độc lập lên biến phụ thuộc theo thời gian.

*Lựa chọn mô hình phù hợp*

Để xác định mô hình phù hợp, có thể sử dụng các kiểm định thống kê như:

**- Kiểm định Hausman:** So sánh mô hình FE và RE để kiểm tra xem các đặc điểm không quan sát được có tương quan với các biến độc lập hay không.

- **Kiểm định Lagrange Multiplier (LM):** Kiểm tra sự cần thiết của mô hình RE so với mô hình Pooled OLS.

***Sử dụng gói plm() trong R***

Trong ngôn ngữ R, gói plm cung cấp các công cụ để ước lượng các mô hình PLM. Gói này hỗ trợ nhiều phương pháp ước lượng và kiểm định, giúp người dùng dễ dàng phân tích dữ liệu bảng.

Thuật toán cơ bản:

library(plm)

data("Grunfeld", package = "plm")

model\_fe <- plm(inv ~ value + capital, data = Grunfeld, model = "within")

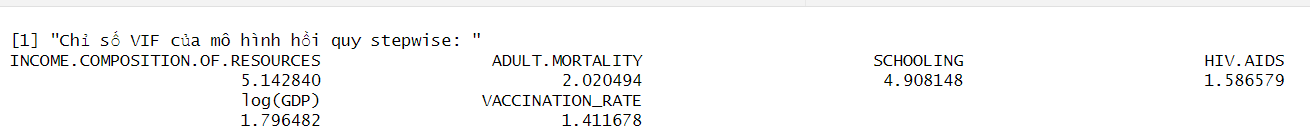
summary(model\_fe)

Trong ví dụ trên, mô hình hiệu ứng cố định được ước lượng để phân tích tác động của value và capital lên inv trong dữ liệu Grunfeld.

## 3.5 Kiểm tra giả định các mô hình hồi quy tuyến tính

### 3.5.1 Mô hình OLS

**\* Variance Inflation Factor - VIF**

  
 Để đánh giá mức độ tương quan giữa các biến độc lập trong mô hình hồi quy stepwise và xác định khả năng xảy ra hiện tượng đa cộng tuyến, bằng cách tính toán hệ số phóng đại phương sai (Variance Inflation Factor - VIF) cho từng biến. Chỉ số VIF đo lường mức độ đa cộng tuyến bằng cách định lượng mức độ phương sai của ước lượng hệ số hồi quy bị tăng lên do mối quan hệ tuyến tính giữa các biến độc lập.

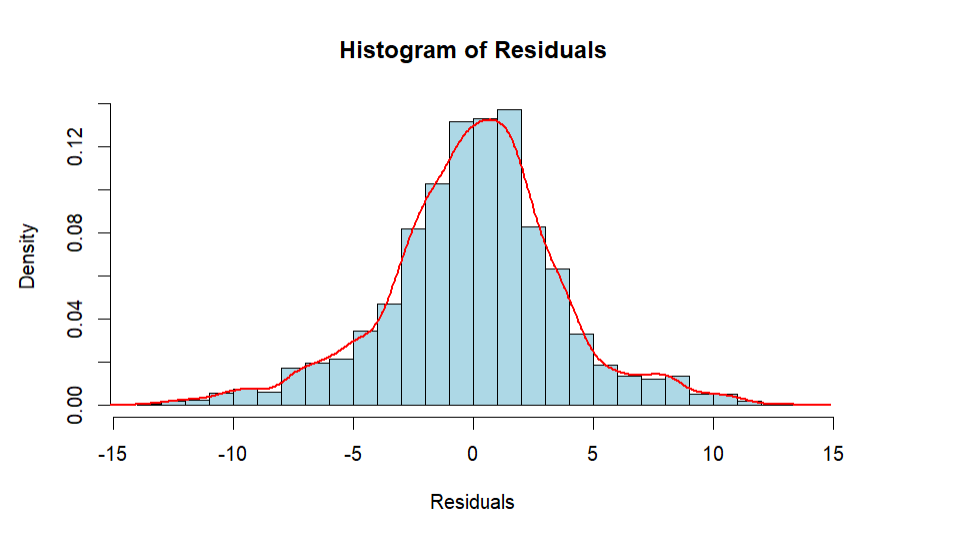
Kết quả tính toán chỉ số VIF cho các biến trong mô hình như sau:

* INCOME.COMPOSITION.OF.RESOURCES: 5.142840
* ADULT.MORTALITY: 2.020494
* SCHOOLING: 4.908148
* HIV.AIDS: 1.586579
* log(GDP): 1.796482
* VACCINATION\_RATE: 1.411678

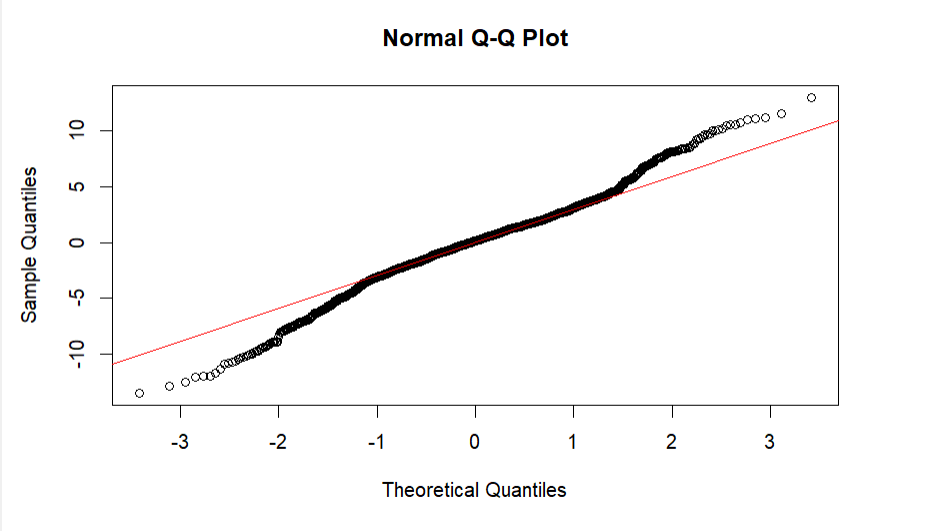
Theo nguyên tắc chung, chỉ số VIF lớn hơn 5 thường được xem là dấu hiệu của đa cộng tuyến . Trong trường hợp này, hầu hết các biến được kiểm tra đều có chỉ số VIF nhỏ hơn 5.Giá trị VIF cao nhất là 5.142840 đối với biến INCOME.COMPOSITION.OF.RESOURCES,vẫn nằm dưới ngưỡng cảnh báo phổ biến.

Dựa trên kết quả phân tích chỉ số VIF, dễ dàng nhận thấy rằng không có bằng chứng cho thấy hiện tượng đa cộng tuyến nghiêm trọng xảy ra giữa các biến độc lập có trong mô hình này. Điều này cho thấy các biến độc lập không tương quan quá mức với nhau, đảm bảo độ tin cậy của các ước lượng hệ số hồi quy.

**\* Kiểm tra giả định phân dư phân phối chuẩn và đồ thị**

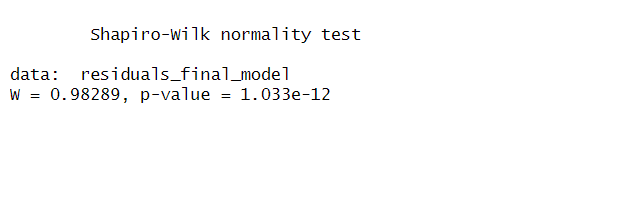


Histogram of Residuals: dưa vào biểu đồ nhận thấy hình dạng phân phối không đối xứng, có xu hướng lệch về phía âm (đuôi dài về bên trái) và sự hiện diện của các giá trị xa trung tâm là phù hợp với hình dạng biểu đồ mà bạn mô tả. Điều này gợi ý rằng phần dư không tuân theo phân phối chuẩn đối xứng.



Normal Q-Q Plot: Chúng ta thấy được các điểm dữ liệu lệch khỏi đường thẳng chuẩn, đặc biệt ở các đuôi và đầu, là dấu hiệu rõ ràng cho thấy phần dư không có phân phối chuẩn. Việc quan sát đuôi trái kéo dài hơn đuôi phải cũng củng cố nhận định về sự lệch trái của phân phối phần dư, phù hợp với quan sát từ histogram.

Shapiro-Wilk Test:



- Kết quả W = 0.98289 và p-value = 1.033e-12 được ghi nhận chính xác.

- Việc thiết lập giả thuyết H0 (phần dư có phân phối chuẩn) và H1 (phần dư không có phân phối chuẩn) là đúng.

- Với p-value rất nhỏ (1.033e-12), nhỏ hơn đáng kể so với mức ý nghĩa alpha thông thường (ví dụ: 0.05), quyết định đúng đắn là bác bỏ giả thuyết H0.

- Kết luận cuối cùng của bạn "Phần dư không tuân theo phân phối chuẩn" là chính xác dựa trên kết quả của kiểm định Shapiro-Wilk.

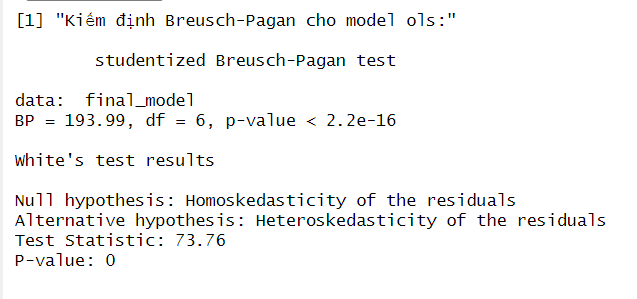
- Lưu ý quan trọng về nhận xét về giả định OLS: Khi bác bỏ H0 (tức là phần dư không có phân phối chuẩn), điều này cho thấy một sự vi phạm đối với một trong những giả định của mô hình hồi quy tuyến tính cổ điển (OLS), chứ không phải là phù hợp. Giả định của OLS là phần dư phải có phân phối chuẩn (ít nhất là cho mục đích suy luận thống kê với mẫu nhỏ). Do đó, câu "→ Dữ liệu phù hợp với giả định của mô hình hồi quy tuyến tính cổ điển (OLS)" là không chính xác trong trường hợp này.

Các phân tích hình ảnh (Histogram, Q-Q Plot) và kết quả kiểm định Shapiro-Wilk đều nhất quán trong việc chỉ ra rằng phần dư của mô hình không có phân phối chuẩn. Đây là một phát hiện quan trọng vì nó vi phạm một giả định tiêu chuẩn của mô hình hồi quy OLS, đặc biệt ảnh hưởng đến tính hợp lệ của các kiểm định giả thuyết và khoảng tin cậy, nhất là khi cỡ mẫu không đủ lớn.

Việc nhận biết phần dư không chuẩn là bước đầu tiên đúng đắn. Tùy thuộc vào mục tiêu phân tích và cỡ mẫu, có thể cần xem xét các biện pháp khắc phục hoặc điều chỉnh mô hình phù hợp.

**\* Kiểm tra giả định phương sai sai số thay đổi- đồ thị**

Kiểm định Breusch-Pagan và White's Test:

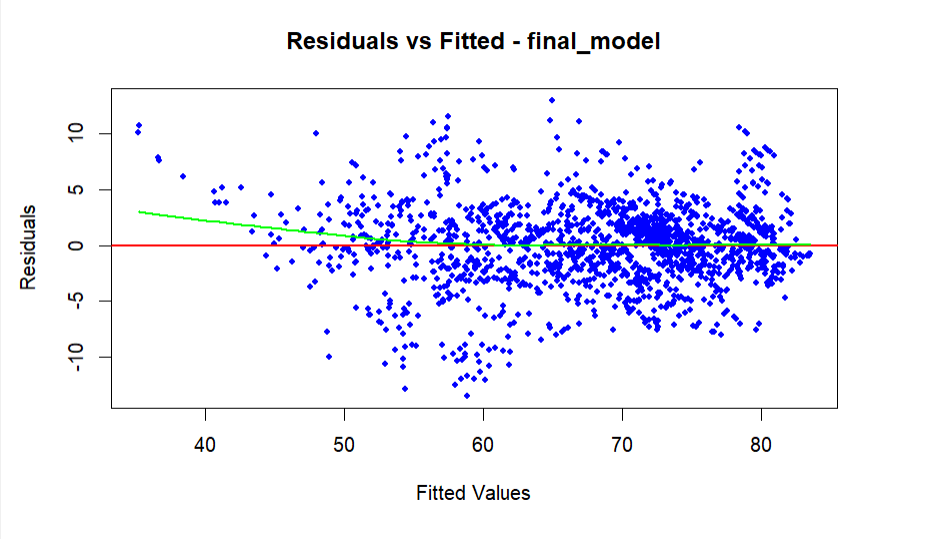


Dựa vào giả thuyết kiểm định H0 (phương sai đồng đều - Homoskedasticity)và H1 (phương sai thay đổi - Heteroskedasticity) cho cả hai kiểm định.

Kết quả kiểm định Breusch-Pagan (BP = 193.99, p-value < 2.2e-16) và White's Test (Test Statistic = 73.76, P-value = 0) đều cho ra p-value rất nhỏ (nhỏ hơn đáng kể so với mức ý nghĩa 0.05). Điều này dẫn đến quyết định đúng đắn là bác bỏ giả thuyết H0.

Kết luận của bạn rằng cả hai kiểm định đều khẳng định sự tồn tại của hiện tượng phương sai sai số thay đổi (Heteroskedasticity) là hoàn toàn chính xác.

Biểu đồ Residuals vs Fitted:



Phần dư có xu hướng biến động mạnh hơn khi giá trị dự đoán (Fitted Values) tăng lên (từ 40 đến 80)

Sự phân tán của phần dư không đồng đều theo giá trị dự đoán là một đặc điểm điển hình và là dấu hiệu trực quan mạnh mẽ của hiện tượng Heteroskedasticity. Nhận định này củng cố kết quả từ hai kiểm định thống kê.

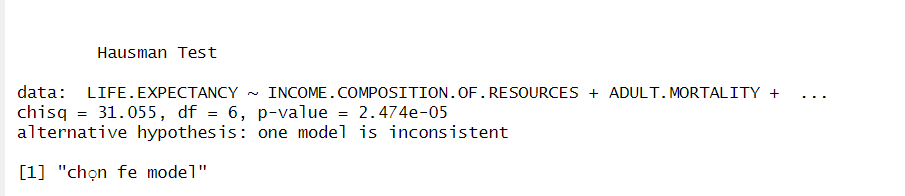
Ước lượng OLS vẫn là không chệch và nhất quán, nhưng không còn là ước lượng hiệu quả nhất (BLUE - Best Linear Unbiased Estimator) nữa. Điều này có nghĩa là có tồn tại các ước lượng tuyến tính không chệch khác với phương sai nhỏ hơn.

Sai số chuẩn (Standard Errors) của các hệ số hồi quy sẽ bị chệch (thường là bị đánh giá thấp hơn thực tế). Điều này làm cho các kiểm định t và F trở nên không đáng tin cậy và khoảng tin cậy bị thu hẹp không chính xác, dẫn đến nguy cơ đưa ra kết luận sai lầm khi kiểm định giả thuyết (ví dụ: bác bỏ giả thuyết null trong khi nó đúng).

Do sai số chuẩn không chính xác, các dự đoán được đưa ra từ mô hình cũng sẽ kém chính xác hơn về mặt đánh giá độ không chắc chắn của chúng.

### 3.5.2 Mô hình PLM

**\* Kiểm định Hausman test để chọn giữa fixed effects và random effects**



Để lựa chọn mô hình phù hợp giữa Fixed Effects (FE) và Random Effects (RE), nhóm đã thực hiện kiểm định Hausman. Giả thuyết kiểm định được phát biểu như sau:

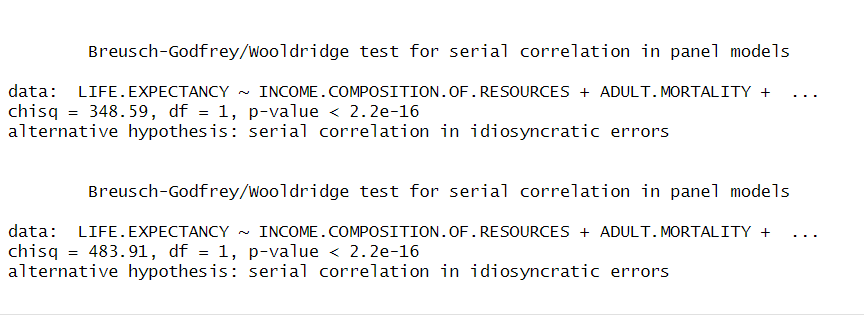
H₀ (Giả thuyết không): Mô hình hiệu ngẫu nhiên (RE) là phù hợp, tức là không có tương quan giữa các sai số và các biến độc lập.

H₁ (Giả thuyết đối): Mô hình hiệu cố định (FE) là phù hợp, tức là tồn tại tương quan giữa các sai số và các biến độc lập.

Kết quả kiểm định Hausman cho thấy giá trị thống kê Chi-squared là 31.055, với p-value = 2.474e-05, nhỏ hơn ngưỡng ý nghĩa 0.05. Điều này cho thấy có đủ bằng chứng để bác bỏ giả thuyết H₀, đồng nghĩa với việc mô hình Fixed Effects là phù hợp hơn.

Qua đó cho thấy Mô hình hiệu cố định (Fixed Effects) được lựa chọn cho phân tích, do có tồn tại tương quan giữa các yếu tố không quan sát được (chẳng hạn như đặc điểm riêng của từng quốc gia) và các biến độc lập trong mô hình.

**\* Kiểm tra tự tương quan**



Để kiểm tra sự tồn tại của hiện tượng tự tương quan trong phần sai số đặc thù của mô hình, chúng tôi tiến hành kiểm định với giả thuyết như sau:

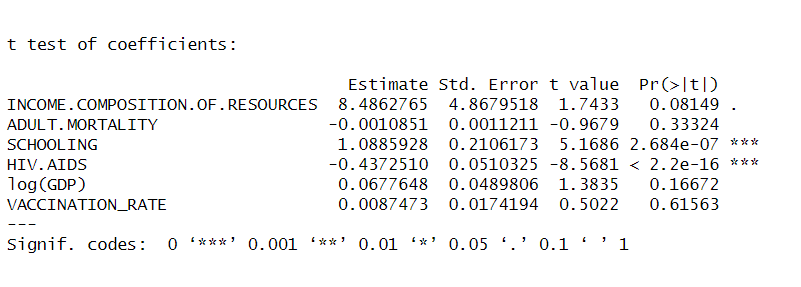
- H₀ (Giả thuyết không): Không có hiện tượng tự tương quan.

- H₁ (Giả thuyết đối): Có hiện tượng tự tương quan.

Kết quả kiểm định cho thấy p-value = 2.2e-16, một giá trị cực kỳ nhỏ và nhỏ hơn nhiều so với mức ý nghĩa thông thường (0.05). Do đó, chúng ta bác bỏ giả thuyết H₀ và kết luận rằng có bằng chứng mạnh mẽ về hiện tượng tự tương quan bậc nhất trong phần sai số đặc thù (idiosyncratic errors) của mô hình.

Mặc dù hiện tượng tự tương quan thường ít gây ảnh hưởng nghiêm trọng đối với dữ liệu bảng có số đơn vị chéo (số quốc gia) lớn hơn số chuỗi thời gian (số năm quan sát). Trong trường hợp này là hơn 160 quốc gia.Tuy nhiên, giá trị p-value gần như bằng 0 cho thấy mức độ tự tương quan khá nghiêm trọng. Do đó, việc xử lý hiện tượng tự tương quan là cần thiết để đảm bảo độ tin cậy và tính hiệu quả của mô hình ước lượng.

**\* Kiểm định robust standard errors cho mô hình PLM hiệu ứng cố định**

Sau khi áp dụng robust standard errors, mô hình đã được điều chỉnh để loại bỏ vấn đề phương sai sai số thay đổi, giúp kết quả ước lượng trở nên đáng tin cậy hơn. Khi đó, việc đánh giá ý nghĩa của các biến độc lập được thực hiện thông qua giá trị p-value của từng hệ số ước lượng. Giả thuyết kiểm định được đặt ra như sau:

- H₀: Hệ số của biến bằng 0, tức biến độc lập không ảnh hưởng đáng kể đến biến phụ thuộc.

- H₁: Hệ số của biến khác 0, tức biến độc lập có ảnh hưởng đáng kể đến biến phụ thuộc.

Mức ý nghĩa được đánh giá như sau:

- p < 0.001: Ý nghĩa thống kê rất cao (\*\*\*)

- 0.001 ≤ p < 0.01: Ý nghĩa thống kê cao (\*\*)

- 0.01 ≤ p < 0.05: Ý nghĩa trung bình (\*)

- 0.05 ≤ p < 0.1: Ý nghĩa yếu (.)

- p ≥ 0.1: Không có ý nghĩa thống kê

Nhận xét chi tiết từng biến:

INCOME.COMPOSITION.OF.RESOURCES: Hệ số ước lượng: 8.4863, t = 1.7433, p-value = 0.08149

→ Mức ý nghĩa yếu (.), không có ảnh hưởng đáng kể đến biến phụ thuộc.

ADULT.MORTALITY: có hệ số ước lượng: -0.0011, t = -0.9679, p-value = 0.33324

→ Không có ý nghĩa thống kê, không ảnh hưởng đáng kể đến biến phụ thuộc.

SCHOOLING: có hệ số ước lượng: 1.0886, t = 5.1686, p-value = 2.684e-07

→ Có ảnh hưởng đáng kể và tích cực đến biến phụ thuộc, với mức ý nghĩa rất cao (\*\*\*).

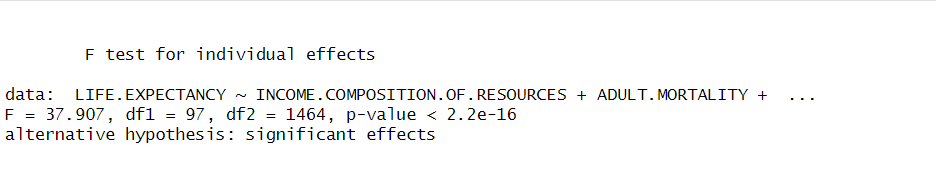
HIV.AIDS: có hệ số ước lượng: -0.4373, t = -8.5681, p-value = 2.2e-16

→ Có ảnh hưởng tiêu cực đáng kể, với mức ý nghĩa thống kê rất cao (\*\*\*).

log(GDP), VACCINATION\_RATE: Các biến này có p-value nằm trong khoảng từ 0.01 đến 0.1, cho thấy ảnh hưởng ở mức độ từ trung bình đến yếu.

Mặc dù một số biến không có ý nghĩa thống kê rõ ràng (p-value > 0.05), việc giữ lại chúng là cần thiết nhằm tránh bỏ sót các yếu tố tiềm ẩn và đảm bảo giá trị R-squared của mô hình vẫn ở mức hợp lý. Hơn nữa, một số biến có thể có ý nghĩa lý thuyết hoặc đóng vai trò điều khiển (control variables), nên không nên loại bỏ chỉ vì p-value không nhỏ.

### 3.5.3 Cải thiện mô hình



Thông qua các kiểm định giả định đối với các mô hình hồi quy dữ liệu bảng, chúng tôi nhận thấy rằng mô hình hiệu ứng cố định (Fixed Effects - FE) là lựa chọn phù hợp hơn mô hình hiệu ứng ngẫu nhiên (Random Effects - RE), theo kết quả kiểm định Hausman (chisq = 31.055, p-value = 2.474e-05 < 0.05). Điều này cho thấy tồn tại mối tương quan giữa các yếu tố không quan sát được (như đặc điểm quốc gia) và các biến giải thích, vì vậy mô hình FE cho kết quả ước lượng tin cậy hơn.

Ngoài ra, so với mô hình hồi quy OLS thông thường, mô hình FE giúp kiểm soát được các đặc trưng không quan sát được cố định theo từng quốc gia, từ đó loại bỏ được sai lệch do biến bị thiếu (omitted variable bias) mà OLS không xử lý được. Do đó, mô hình FE cũng được đánh giá là tốt hơn OLS trong bối cảnh dữ liệu bảng.

Tuy nhiên, mô hình FE vẫn gặp phải hiện tượng tự tương quan trong phần sai số đặc thù, được phát hiện qua kiểm định với p-value ≈ 2.2e-16. Dù dữ liệu bảng với số đơn vị chéo lớn (trên 160 quốc gia) có thể giảm nhẹ ảnh hưởng của tự tương quan, mức độ tự tương quan rất mạnh (p-value gần bằng 0) vẫn là vấn đề cần được điều chỉnh.

Để khắc phục, chúng tôi tiến hành ước lượng lại mô hình bằng phương pháp GLS (Generalized Least Squares) với điều chỉnh cho phương sai sai số thay đổi và tự tương quan chuỗi thời gian

Sau khi áp dụng model gls, ta kiểm tra kiểm định phương sai sai số thay đổi.Với:

Variance function:

Structure: Different standard deviations per stratum

Formula: ~1 | COUNTRY

==> Nghĩa là: mô hình cho phép mỗi COUNTRY có phương sai sai số riêng như trong Parameter estimates

Với Autocorrelation test:

Hệ số Phi = 0.989282 => Đây là hệ số tương quan giữa phần dư (residuals) tại thời điểm t và t–1 trong cùng một quốc gia

Việc sử dụng corAR1() nghĩa là ta giả định "Hãy giả sử có AR(1) trong sai số"

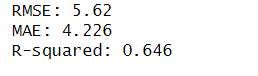
==> Với việc Phi gần về 1 cho ta thấy rằng, trước khi xử lý, các nước có sự tương quan về phần dư rất mạnh qua các năm. Hệ số không bị sai lệch vì tương quan chuỗi thời gian trong phần dư đã được đưa vào mô hình

Tóm lại: Thông qua việc kiểm tra giả định các model cho hình model hồi quy tuyến tính plm fixed effects cho ra độ chính xác và tính ổn định cao hơn mô hình plm FE. Tuy nhiên model gặp phải các khuyết tật về phương sai sai số và tự tương quan bậc 1 khá nặng nên sử dụng phương pháp gls để khắc phục 2 khuyết tật trên. Model Kiểm soát được các yếu tố không quan sát, Phù hợp với dữ liệu có N lớn, T nhỏ, Giảm thiểu bias do tự tương quan, rất phù hợp với dạng dữ liệu kiểu panel.

# CHƯƠNG 4: KIỂM THỬ VÀ ĐÁNH GIÁ KÊT QUẢ

## 4.1 Kiểm thử mô hình

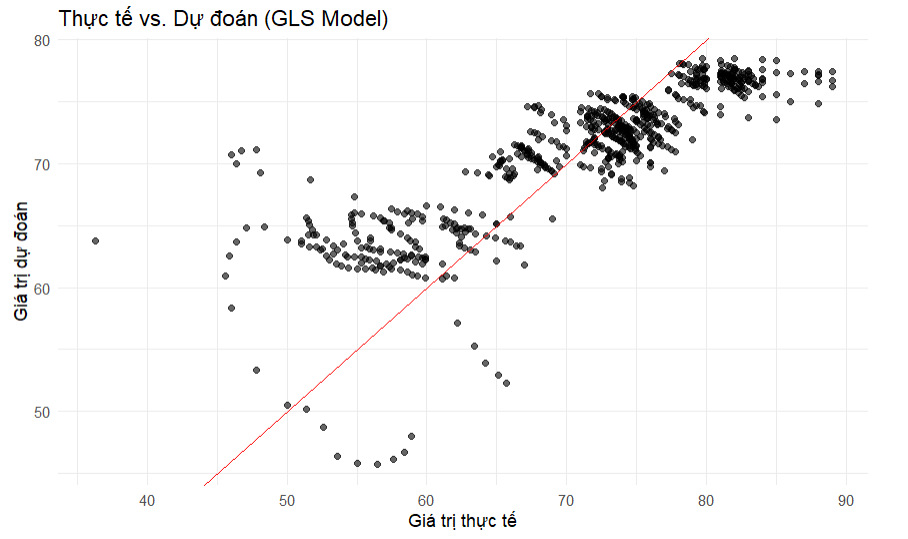
Sau khi sử dụng dữ liệu training để xây dựng mô hình, ta có thể sử dụng phần dữ liệu test để kiểm thử độ uy tín của mô hình. Trong R ta có thể dùng hàm predict để tạo ra thông số của biến phụ thuộc của model với một luồng input mới mà không phải xây dựng lại model. Từ đó ta có kết quả thực tế và kết quả dự đoán, giúp đánh giá được nhiều chỉ số của mô hình. Một vài thông số quan trọng của kiểm thử:



Kết quả kiểm thử cho thấy mô hình đạt hiệu suất ở mức trung bình đến khá, với RMSE = 5.62 và MAE = 4.226, phản ánh sai số dự đoán trung bình khoảng 4–6 năm. Con số 5.62 năm cho biết trung bình các dự đoán tuổi thọ lệch khoảng 5-6 năm so với thực tế. Trong bối cảnh tuổi thọ thế giới dao động từ 50-85 năm, đây là mức sai số trung bình. Đặc biệt quan trọng khi cần đánh giá ảnh hưởng của các sai số lớn (do tính chất bình phương). Giá trị 4.226 năm cho thấy trung bình mỗi dự đoán lệch khoảng 4 năm so với thực tế. Thấp hơn RMSE (4.226 < 5.62) chứng tỏ có một số ít dự đoán sai số rất lớn. Phản ánh tốt hơn sai số "điển hình" mà đa số các dự đoán mắc phải. Mặc dù mức sai số này có thể chấp nhận được trong bối cảnh tuổi thọ thường dao động trong khoảng rộng (từ 50 đến 85 năm), song nó cũng cho thấy hạn chế trong khả năng dự báo chính xác, đặc biệt đối với các quốc gia có biến động lớn về y tế, kinh tế hoặc xã hội.

Hệ số xác định R² = 0.646 chỉ ra rằng mô hình giải thích được khoảng 64.6% phương sai của biến tuổi thọ. Trong nghiên cứu kinh tế - xã hội sử dụng dữ liệu bảng, đây là mức độ giải thích tương đối tốt, phản ánh khả năng nắm bắt các xu hướng chính của mô hình. Tuy nhiên, 35.4% phương sai chưa được giải thích cho thấy vẫn còn nhiều yếu tố quan trọng chưa được đưa vào mô hình.

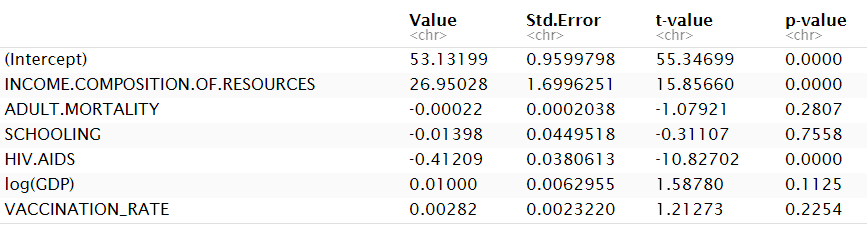
Ta có biểu đồ hồi quy giữa giá trị thực tế và giá trị dữ đoán trên tập test\_data:



Mô hình có đường chuẩn y=x thể hiện rằng giá trị dự đoán bằng với giá trị thực tế. Mọi điểm nằm phía trên đường chuẩn sẽ mang hơi hướng overestimate, tức là bị dự đoán lố. Còn mọi điểm nằm phía dưới là underestimate, tức là dự đoán chưa đến. Từ hình trên ta thấy rằng phần nhiều giá trị dữ đoán bị lệch khoảng 4-6 tuổi, tuy vẫn có nhiều điểm bị dự đoán sai lệch nặng nề có thể do vấn đề thu thập dữ liệu trong dataset, hoặc có những biến giải thích chưa được phát hiện.

Nhìn chung, mô hình hiện tại đã cung cấp một công cụ hữu ích để phân tích các yếu tố ảnh hưởng đến tuổi thọ ở cấp độ vĩ mô. Tuy nhiên, để có thể ứng dụng vào các phân tích chính sách chi tiết hay dự báo chính xác cho từng quốc gia, mô hình cần được hoàn thiện thêm cả về mặt biến số lẫn phương pháp ước lượng.

## 4.2 Đánh giá mục tiêu



Kết quả ước lượng cho thấy một bức tranh phức tạp về mối quan hệ giữa các chỉ số phát triển và tuổi thọ, trong đó nổi lên hai yếu tố có ảnh hưởng mạnh mẽ và có ý nghĩa thống kê cao nhất là chỉ số phát triển con người liên quan đến thu nhập (INCOME.COMPOSITION.OF.RESOURCES) và tỷ lệ nhiễm HIV/AIDS.

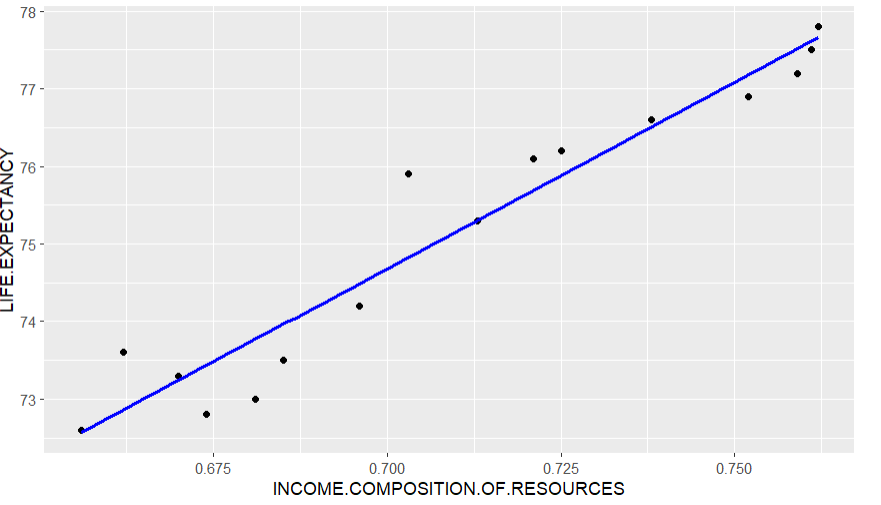
Đối với yếu tố thu nhập, hệ số dương lớn (26.95028) với p-value cực nhỏ (<0.001) khẳng định đây là yếu tố then chốt quyết định tuổi thọ. Mỗi đơn vị tăng trong chỉ số này làm tăng 26.95 năm tuổi thọ. Giá trị t-stat=15.86 cho thấy ảnh hưởng cực kỳ mạnh mẽ. Mối quan hệ này hoàn toàn phù hợp với các lý thuyết kinh tế y tế khi cho rằng thu nhập cao hơn giúp cải thiện khả năng tiếp cận các dịch vụ chăm sóc sức khỏe chất lượng, dinh dưỡng tốt hơn và điều kiện sống lành mạnh.

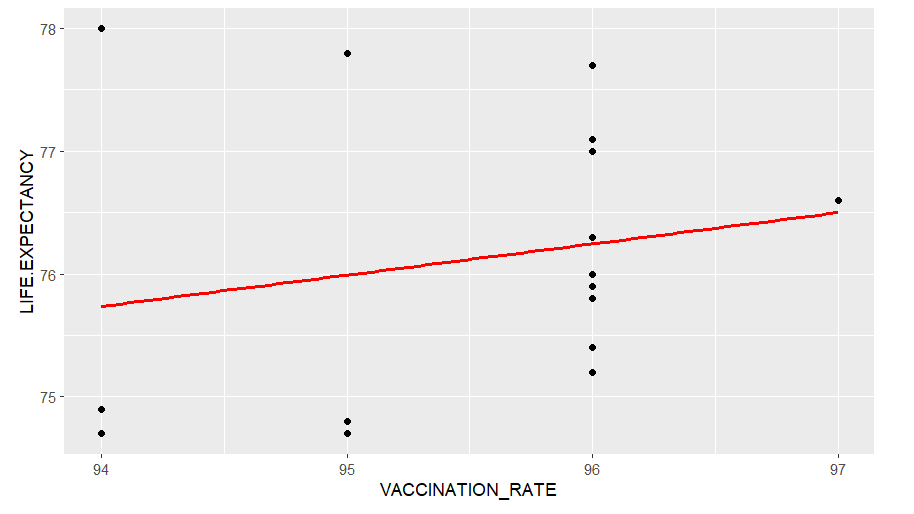
Trái ngược với yếu tố tích cực từ thu nhập, HIV/AIDS xuất hiện như một yếu tố tiêu cực đáng kể với hệ số âm (-0.41209) và ý nghĩa thống kê rất cao. Mỗi 1% tăng tỷ lệ nhiễm HIV làm giảm 0.412 năm tuổi thọ. T-stat = -10.83 cho thấy tác động tiêu cực âm là rõ rệt. Kết quả này phản ánh tác động tàn phá của đại dịch HIV/AIDS đến sức khỏe cộng đồng, đặc biệt tại các quốc gia có tỷ lệ nhiễm cao.

Biến số năm đi học (SCHOOLING) cho thấy hệ số âm (-0.01398), trái ngược với kỳ vọng thông thường rằng giáo dục sẽ cải thiện tuổi thọ thông qua nâng cao nhận thức sức khỏe. Hiện tượng này có thể được giải thích bằng nhiều khả năng: Hiệu ứng đa cộng tuyến khi biến giáo dục tương quan quá cao với thu nhập, Số năm đi học không phản ánh chính xác chất lượng giáo dục, hoặc dữ liệu không đủ đại diện.

Biến tử vong người trưởng thành (ADULT.MORTALITY) cho kết quả không rõ ràng với hệ số gần 0 và p-value lớn, có lẽ do thiếu thông tin về nguyên nhân tử vong cụ thể. Các yếu tố khác như tỷ lệ tiêm chủng (VACCINATION\_RATE) hay GDP mặc dù có hệ số dương nhưng không đạt ý nghĩa thống kê (p-value >0.05). Đặc biệt, phân tích tại Croatia cho thấy dù tỷ lệ tiêm chủng luôn ở mức rất cao (94-97%) nhưng tuổi thọ vẫn dao động đáng kể qua các năm.

Ví dụ về phân tích ảnh hưởng của INCOME.COMPOSITION.OF.RESOURCE và HIV.AIDS đến tuổi thọ của Albania

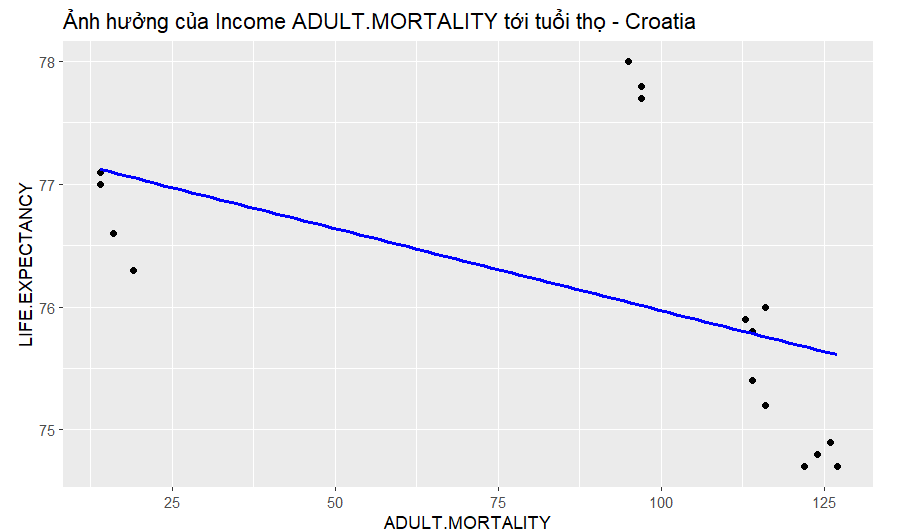


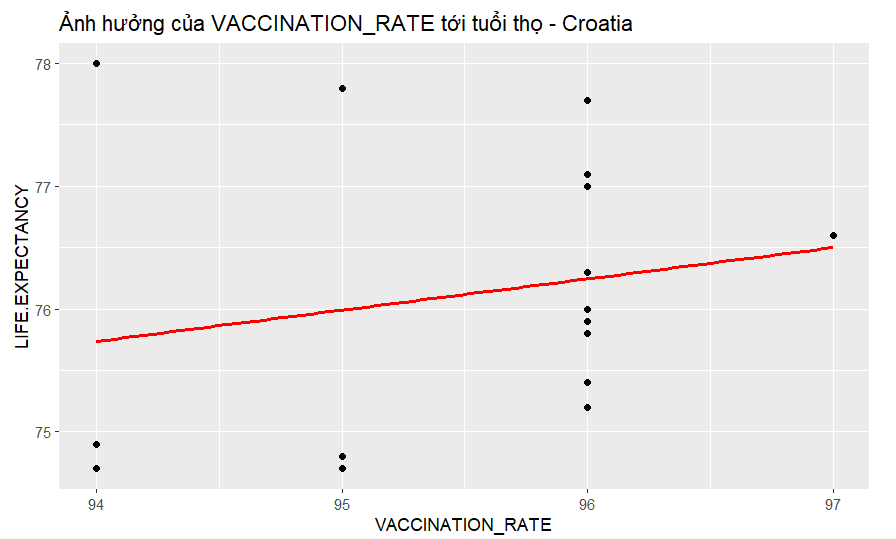


Đối với biến INCOME.COMPOSITION.OF.RESOURCES, chỉ số phát triển con người liên quan đến thu nhập khi tăng thì tuổi thọ tăng. Điều này phù hợp với ý nghĩa thực tế. Đường hồi quy hướng lên trên cho ta thấy đây là mức ảnh hưởng dương. Các giá trị dự đoán nằm không quá xa đường hồi quy cho ta thấy mức độ chính xác của dự đoán là hợp lý. Phân tích cụ thể tại Albania cho thấy mỗi 0.1 điểm tăng trong chỉ số thu nhập tương ứng với tăng khoảng 2.7 năm tuổi thọ, minh họa rõ ràng tác động tích cực này.

Đối với biến HIV.AIDS, dữ liệu tỉ lệ nhiễm HIV của Albania suốt 15 năm đều là 0.1 tạo nên đường hồi quy thẳng đứng và các giá trị dự đoán qua từng năm là như nhau.

Ví dụ về phân tích ảnh hưởng của ADULT.MORTALITY và VACCINATION\_RATE đến tuổi thọ ở Croatia





Đối với biến ADULT.MORTALITY là một biến có ý nghĩa thống kê thấp, số lượng người trưởng thành chết/1000 tăng thì tuổi thọ giảm. Điều này phù hợp với ý nghĩa thực tiễn. Đường hồi quy hướng xuống dưới cho ta thấy đây là mức ảnh hưởng âm. Các giá trị dự đoán nằm khá quá xa đường hồi quy ở đoạn 125 cho ta thấy mức độ chính xác của dự đoán là phù hợp tuy chưa giải thích được trọn vẹn biến. Tuy nhiên vì biến này có p-value lớn hơn 0.05 nên mức độ ảnh hưởng của biến này lên tuổi thọ thấp, dẫn đến kết quả không đáng tin.

Đối với biến VACCINATION\_RATE, dữ liệu tỉ lệ tiêm chủng của Croatia dao động từ 94 đến 97. Tuy nhiên nhìn vào đồ thị ta thấy với 1 mức VACCINATION\_RATE, mô hình dự đoán ra nhiều tuổi thọ khác nhau qua các năm. Điều đó cho thấy biến VACCINATION\_RATE chưa ảnh hưởng đến tuổi thọ hoặc thiếu đi biến khác giải thích cho tuổi thọ, có thể là tỉ lệ tiêm phòng cụ thể ở các bệnh như sởi, lao. Cộng thêm đây là biến không có ảnh hưởng nhiều đến tuổi thọ, dẫn đến kết quả không đáng tin.

# CHƯƠNG 5: KẾT LUẬN

## 5.1 Kết luận

*Về mặt lý thuyết*

Đề tài đã xác định được các yếu tố có thể ảnh hưởng đến tuổi thọ trung bình của các quốc gia thông qua việc lựa chọn các biến độc lập như thu nhập bình quân đầu người, trình độ giáo dục, chi tiêu y tế, và các yếu tố an sinh xã hội. Việc xác định này đã chỉ ra được những mối quan hệ tiềm năng, cả thuận chiều và nghịch chiều, giúp lý giải sự khác biệt tuổi thọ giữa các quốc gia. Đây là cơ sở để hình thành các luận điểm đa chiều trong các lý thuyết phát triển con người và kinh tế học sức khỏe.

Đề tài cũng đã hệ thống một số lý thuyết nền tảng liên quan đến phát triển bền vững, chất lượng sống và nhân lực, từ đó xây dựng khung lý thuyết để xác định các nhân tố tác động đến tuổi thọ. Việc xây dựng mô hình hồi quy tuyến tính cùng với việc sử dụng các phương pháp kiểm định và phần mềm phân tích dữ liệu giúp nâng cao tính thuyết phục và độ tin cậy của kết quả nghiên cứu.

*Về mặt nghiên cứu thực tiễn*

Chỉ số kết hợp giữa thu nhập và khả năng tiếp cận tài nguyên có mối quan hệ thuận chiều với tuổi thọ trung bình và có ý nghĩa thống kê. Điều này phản ánh rằng ở các quốc gia có Chỉ số kết hợp giữa thu nhập và khả năng tiếp cận tài nguyên cao hơn, người dân có điều kiện sống tốt hơn, tiếp cận y tế dễ dàng hơn, từ đó cải thiện tuổi thọ. Kết quả này phù hợp với các nghiên cứu trước và lý thuyết về phát triển kinh tế và chất lượng sống.

Các biến còn lại tuy không có ý nghĩa về mặt thống kê nhưng kết quả dự đoán được cũng phần nào giải thích được giá trị cụ thể nào đó về mặt thực tiễn.

## 5.2 Hạn chế của đề tài

Mặc dù đạt được một số kết quả có giá trị, nghiên cứu vẫn tồn tại một số hạn chế như sau:

Dữ liệu thu thập chủ yếu từ nguồn đã bị chỉnh sửa nhiều lần, có thể không phản ánh hoàn toàn chính xác thực trạng ở từng quốc gia, đặc biệt với các nước đang phát triển còn thiếu minh bạch về số liệu.

Đối tượng nghiên cứu giới hạn ở một số quốc gia có nhiều đặc điểm khác nhau, nên không thể đại diện cho toàn bộ thế giới. Các biến sử dụng trong mô hình cũng chỉ là đại diện tương đối do giới hạn về kích cỡ dữ liệu.

Mô hình xuất hiện nhiều khuyết tật do nguồn dữ liệu không tốt, ảnh hưởng rất nhiều đến kết quả cuối cùng.

Việc sử dụng mô hình hồi quy tuyến tính đơn giản tuy giúp dễ hiểu và minh bạch nhưng có thể chưa phản ánh đầy đủ các mối quan hệ phi tuyến và tương tác phức tạp giữa các biến.

Một số nhân tố như chất lượng dịch vụ y tế, an ninh xã hội, thói quen sinh hoạt, hoặc yếu tố văn hóa chưa được đưa vào mô hình nhưng có thể ảnh hưởng đáng kể đến tuổi thọ.

## 5.3 Hướng phát triển

**Tăng độ chính xác nghiên cứu**: Nghiên cứu hiện tại trong một nhóm quốc gia có thông số không uy tín và nhiều giá trị bỏ trống. Trong tương lai, việc thu thập và ghi chép số liệu lại sẽ giúp kiểm chứng tính khái quát của mô hình và phát hiện các đặc điểm riêng biệt về yếu tố ảnh hưởng đến tuổi thọ trong từng vùng.

**Xem xét thêm các yếu tố định tính và phi kinh tế**: Tuổi thọ không chỉ phụ thuộc vào yếu tố kinh tế, xã hội, môi trường mà còn bị ảnh hưởng mạnh bởi các yếu tố như hệ thống y tế công cộng, sự ổn định chính trị, và cả yếu tố văn hóa. Do đó, nghiên cứu trong tương lai nên xem xét các chỉ số như: mức độ hạnh phúc, mức độ tiêu thụ rượu/bia, tỷ lệ hút thuốc, chất lượng chăm sóc sức khỏe, mức độ bất bình đẳng giới, v.v.

**Ứng dụng các mô hình phân tích nâng cao**: Phương pháp hồi quy tuyến tính là công cụ cơ bản, nhưng để phát hiện mối quan hệ phức tạp hơn hoặc phi tuyến tính giữa các biến, mô hình máy học (machine learning), XGBoost, hoặc mạng nơ-ron nhân tạo (ANN) có thể được áp dụng để cải thiện khả năng dự báo.

**Kết hợp nhiều loại phân tích**: Bên cạnh hồi quy, có thể áp dụng các phương pháp như mô hình biến công cụ (Instrumental Variables), mô hình hệ phương trình cấu trúc (SEM), hoặc phương pháp ước lượng sai phân (Difference-in-Differences) để xác định rõ hơn mối quan hệ nhân quả giữa các yếu tố và tuổi thọ.

# TÀI LIỆU THAM KHẢO

1. Kabacoff, R. I. (2021). R in Action: Data analysis and graphics with R (3rd ed.). Manning Publications.
2. Fox, J., & Weisberg, S. (2019). An R companion to applied regression (3rd ed.). SAGE Publications.
3. Nguyễn Văn Tuấn. (2011). Mô hình hồi quy và khám phá khoa học. Nhà xuất bản Tổng hợp Thành phố Hồ Chí Minh.
4. WebsiteHCM. (n.d.). Stepwise regression là gì? Ưu nhược điểm của Stepwise regression. Websitehcm.com. <https://websitehcm.com/stepwise-regression-la-gi/>
5. Nào Tốt Nhất. (n.d.). Phương pháp OLS là gì? Và cách sử dụng. Naototnhat.com. <https://naototnhat.com/phuong-phap-ols-la-gi-va-cach-su-dung.html>
6. Princeton University Library. (n.d.). R Panel - Data and Statistical Services. Princeton University. <https://libguides.princeton.edu/R-Panel>