Vấn đề Phương sai của Sai số Thay đổi (Heteroskedasticity)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

09-12/01/2024

Các giả định chính của mô hình tuyến tính đa biến

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + u$$

1. Tuyến tính theo tham số
2. Chọn mẫu ngẫu nhiên
3. Không có cộng tuyến hoàn hảo giữa các biến giải thích

4. $E(u|X) = 0 \Rightarrow$ Ước lượng OLS là không chệch và nhất quán
5. $Var(u|X) = \sigma^2$ (homoskedasticity) $\Rightarrow$ Ước lượng OLS là BLUE
6. Sai số u độc lập với các biến giải thích, có phân phối chuẩn với giá trị trung bình là 0 và phương sai σ^2 (independent, identically distributed - *iid*):

$$u \sim N(0, \sigma^2)$$

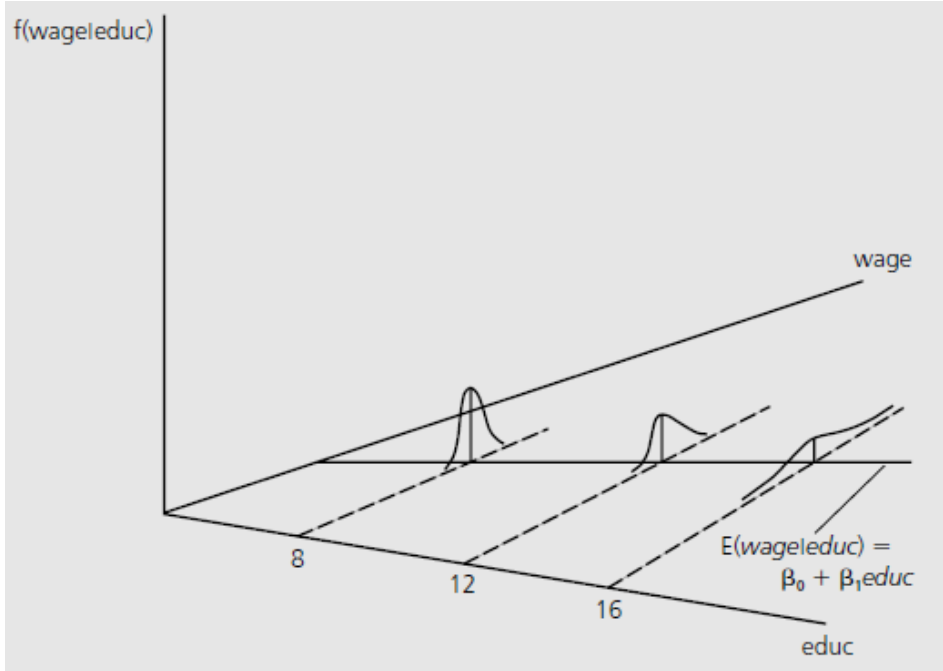
Chúng ta có mô hình hồi quy tuyến tính cổ điển (CLRM) $\Rightarrow$

$$\hat{\beta} \sim N(\beta, Var(\beta))$$

Phương sai của sai số thay đổi (heteroskedasticity)

- Vi phạm điều kiện 5 (và điều kiện 6): $Var(u|X) \neq \sigma^2$
- Ước lượng bằng OLS vẫn là không chệch, nhưng không còn là hiệu quả nhất do sai số của $\hat{\beta}$ không còn là nhỏ nhất
- Các kiểm định t-test, F-test dựa trên phân phối của $\hat{\beta}$ sai do sai số của $\hat{\beta}$ bị sai

Phương sai thay đổi xảy ra khi nào?



- Phương sai của sai số tương quan với biến khác.
  - Ví dụ với người có số năm đi học nhiều thì thường có mức độ dao động của thu nhập càng lớn, dẫn đến tương quan dương giữa phương sai của thu nhập với số năm đi học trong hàm tỷ suất thu nhập của việc đi học.
- Do tương quan chuỗi hoặc tương quan không gian.

- Tương quan chuỗi (auto-correlation): các dữ liệu mang tính phụ thuộc theo thời gian hay chu kỳ.
  - Chi tiêu của mỗi hộ gia đình phụ thuộc vào mức thu nhập hiện tại, thu nhập trong quá khứ, và thu nhập kỳ vọng trong tương lai.
  - Giá chứng khoán, tỷ giá hối đoái, tăng trưởng kinh tế (các dữ liệu mang đặc tính thời gian) đều có vấn đề tương quan chuỗi mạnh.
- Tương quan không gian (spatial correlation): các dữ liệu có tính chất không gian địa lý, dẫn đến hiện tượng các quan sát có vị trí cận kề thường có tương quan lẫn nhau.
  - Các ngôi nhà gần nhau thường có giá bán tương quan nhau.
  - Các học viên ngồi cạnh nhau thường có kết quả học tập tương quan nhau.

Vấn đề sai số thay đổi rất phổ biến trong các dữ liệu và mô hình kinh tế.

Kiểm định hiện tượng phương sai thay đổi

- Kiểm định Breusch-Pagan về phụ thuộc tuyến tính giữa phương sai của sai số và các biến giải thích.
- Kiểm định White trong trường hợp tổng quát.

Kiểm định Breusch-Pagan

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \quad (1)$$

- Giả định $E(u|X) = 0$ và $cov(u, X) = 0$ thỏa $\Rightarrow$ Ước lượng OLS vẫn không chệch và nhất quán.
- Chúng ta muốn kiểm định liệu vấn đề phương sai của sai số thay đổi có xảy ra hay không.

$$H_0 : Var(u|X) = \sigma^2$$

và

$$H_1 : Var(u|X) \neq \sigma^2$$

Các bước thực hiện kiểm định Breusch-Pagan (BP)

Do $E(u|X) = 0$ nên $Var(u|X) = E(u^2) - [E(u)]^2 = E(u^2)$. Do đó kiểm định BP được thực hiện thông qua ước lượng hàm số của $E(u^2)$ theo các biến giải thích. Các bước thực hiện kiểm định BP:

1. Ước lượng mô hình (1) như thông thường
2. Tính giá trị của phần dư $\hat{u}$ và tạo biến phụ thuộc là bình phương của phần dư, $\hat{u}^2$
3. Ước lượng mô hình hồi quy phụ (auxiliary regression) cuả biến $\hat{u}^2$ theo tất cả các biến giải thích:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + ... + \delta_k x_k + v \qquad (2)$$

4. Kiểm định nếu δ_1,..., δ_k đồng thời bằng 0 trong mô hình (2) bằng F-test

$$H_0 : \delta_1 = \delta_2 = ... = \delta_k = 0$$

và

$$H_1 : at\ least\ \delta_j \neq 0$$

Trị kiểm định F được tính từ R_a^2 của mô hình hồi quy phụ:

$$F = \frac{R_a^2/k}{(1 - R_a^2)/(n - k - 1)} \sim F_{k,n-k-1}$$

5. Nếu bác bỏ H_0 chứng tỏ mô hình có hiện tượng phương sai thay đổi

Thực hành kiểm định BP

Ước lượng lại mô hình tỷ suất thu nhập từ bộ dữ liệu VHLSS 2010.

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexpersq + \beta_4 married \\ &+ \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u \end{aligned}$$

- Kiểm định phương sai thay đổi thủ công thông qua kiểm định F
- Thực hiện tự động bằng Stata

Kiểm định phương sai thay đổi trong trường hợp tổng quát bằng kiểm định White

Áp dụng khi cấu trúc hàm của phương sai của sai số u không phải là hàm tuyến tính theo các biến giải thích.

1-2. Tương tự như kiểm định Breusch-Pagan

3. Giả định cấu trúc hàm của phần dư linh hoạt hơn bằng cách thêm bình phương và tương tác giữa các biến giải thích:

$$\begin{aligned}\hat{u}^2 =& \delta_0 + \delta_1 x_1 + ... + \delta_k x_k \\ &+ \sum_{i=1}^{K} \delta_i x_i^2 + \sum_{i=1}^{K} \sum_{j=1}^{K} \delta_{ij} x_i x_j + v\end{aligned} \tag{3}$$

4. Kiểm định bằng F-test nếu tất cả các tham số δ_i và δ_{ij} (loại trừ tung độ gốc δ_0) trong hồi quy phụ (3) bằng 0

Cách thực hiện kiểm định White đơn giản

Trong bước [3-4], tăng số biến trong mô hình sẽ làm giảm số bậc tự do và giảm sức mạnh của kiểm định. Ví dụ mô hình có 3 biến giải thích sẽ có tổng cộng là 9 ràng buộc. Cách thực hiện khác không làm giảm bậc tự do:

1. Ước lượng mô hình như thông thường
2. Ước lượng giá trị dự báo $\hat{y}$, $\hat{y}^2$, và sai số bình phương $\hat{u}^2$
3. Hồi quy $\hat{u}^2$ lên biến $\hat{y}$ và $\hat{y}^2$ trong mô hình phụ:

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + v$$

4. Kiểm định $\delta_1 = \delta_2 = 0$ bằng F-test với 2 ràng buộc
5. Nếu bác bỏ H_0 chứng tỏ mô hình có vấn đề phương sai thay đổi

Thực hành kiểm định White

Ước lượng lại mô hình tỷ suất thu nhập từ bộ dữ liệu VHLSS 2010.

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexpersq + \beta_4 married \\ &+ \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u \end{aligned}$$

- Kiểm định phương sai thay đổi thủ công thông qua kiểm định F.
- Thực hiện tự động bằng Stata.

Chỉnh sửa mô hình khi xảy ra hiện tượng phương sai thay đổi

Sai số được điều chỉnh xử lý vấn đề phương sai thay đổi được gọi là sai số vững, heteroskedasticity-robust standard errors hoặc robust standard errors.

- Phương pháp White-Huber.
- Có thể chỉnh sửa vấn đề heteroscedasticity thủ công dựa vào giả định cấu trúc hàm của phương sai của sai số:
  - Nếu biết cấu trúc hàm của $Var(u|X) \Rightarrow$ Hồi quy với trọng số (Weighted Least Squares - WLS)
  - Nếu không biết cấu trúc hàm của $Var(u|X) \Rightarrow$ Phương pháp bình phương tối thiểu tổng quát khả thi (Feasible Generalized Least Squares - FGLS)

Phương sai biết cấu trúc hàm

- Sử dụng phương pháp hồi quy bình phương tối thiểu có trọng số (Weighted Least Squares - WLS). Giả định phương sai của sai số là một hàm số của x:

$$Var(u|X) = \sigma^2 h(x)$$

- Thực hiện chuyển đổi dữ liệu trước khi ước lượng:

$$\frac{y}{\sqrt{h(x)}} = \beta_0 + \beta_1 \frac{x_1}{\sqrt{h(x)}} + \beta_2 \frac{x_2}{\sqrt{h(x)}} + ... + \frac{u}{\sqrt{h(x)}} \quad (4)$$

- Ước lượng (4) bằng phương pháp OLS có tính chất BLUE.

Phương sai không biết cấu trúc hàm

- Sử dụng phương pháp bình phương tối thiểu tổng quát khả thi (Feasible Generalized Least Squares - FGLS). Thông thường giả định phương sai của sai số là hàm mũ nào đó của biến giải thích X:

$$Var(u|X) = \sigma^2 e^{\delta_0 + \delta_1 x_1 + ... + \delta_k x_k}$$

- Phương pháp FGLS sẽ ước lượng hàm của $Var(u|X)$ để làm trọng số trong phương pháp WLS.

Các bước thực hiện FGLS

1. Hồi quy y theo các biến giải thích, và ước lượng phần dư $\hat{u}$.
2. Tạo biến $log(\hat{u}^2)$.
3. Ước lượng hồi quy $log(\hat{u}^2)$ lên các biến giải thích, và ước lượng giá trị dự báo (fitted value), $\widehat{log(\hat{u}^2)}$.
4. Lấy lũy thừa cơ số e của giá trị dự báo ở bước 3, $\widehat{h(x)} = e^{\widehat{log(\hat{u}^2)}}$.
5. Ước lượng lại mô hình ban đầu bằng WLS, với trọng số là $1/\widehat{h(x)}$.

Thực hành ước lượng và so sánh các mô hình với sai số vững theo phương pháp White, WLS và FGLS

Ước lượng lại mô hình tỷ suất thu nhập từ bộ dữ liệu VHLSS 2010.

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexpersq + \beta_4 married \\ &+ \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u \end{aligned}$$

1. Ước lượng mô hình với giả định phương sai của sai số không đổi
2. Ước lượng mô hình có sai số vững theo phương pháp White-Huber
3. Ước lượng WLS nếu giả định phương sai của sai số tuân theo:

$$Var(u|X) = \sigma^2 income$$

4. Ước lượng FGLS cho trường hợp phương sai thay đổi và không biết cấu trúc hàm

	Homoskedas~y b/se	White Robust b/se	WLS b/se	FGLS b/se
yoeduc	0.0926***	0.0926***	0.0909***	0.0993***
	(0.0027)	(0.0031)	(0.0043)	(0.0026)
yoexper	0.0617***	0.0617***	0.1088***	0.0681***
	(0.0025)	(0.0032)	(0.0029)	(0.0028)
yoexpersq	-0.0012***	-0.0012***	-0.0019***	-0.0013***
	(0.0000)	(0.0001)	(0.0000)	(0.0001)
married	0.0352	0.0352	0.0966**	0.0073
	(0.0221)	(0.0217)	(0.0339)	(0.0206)
publicSchool	-0.1146**	-0.1146*	-0.0153	-0.1262**
	(0.0424)	(0.0465)	(0.0530)	(0.0435)
public	-0.1043**	-0.1043*	-0.3122***	-0.0938*
	(0.0329)	(0.0423)	(0.0489)	(0.0472)
foreign	0.4499***	0.4499***	0.9413***	0.4529***
	(0.0364)	(0.0328)	(0.0735)	(0.0286)
official	0.2705***	0.2705***	0.8263***	0.2296***
	(0.0359)	(0.0430)	(0.0614)	(0.0475)
Constant	8.4936***	8.4936***	6.9670***	8.4044***
	(0.0475)	(0.0539)	(0.0567)	(0.0492)
Obs	7552.0000	7552.0000	7552.0000	7552.0000
R2	0.3026	0.3026	0.3249	0.3460
R2-adj	0.3019	0.3019	0.3242	0.3453
df(r)	7543.0000	7543.0000	7543.0000	7543.0000
SSR	4040.8653	4040.8653	9608.9369	3408.4380

* p<0.05, ** p<0.01, *** p<0.001

Kiểm định giả thuyết khi xảy ra hiện tượng phương sai của sai số thay đổi

Kiểm định nếu số năm kinh nghiệm và số năm kinh nghiệm bình phương đồng thời bằng không.

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexpersq + \beta_4 married \\ &+ \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u \end{aligned}$$

- Do phương sai thay đổi, trị kiểm định t và F sẽ thay đổi (theo hướng giảm so với ước lượng bằng OLS) $\Rightarrow$ sức mạnh của kiểm định giảm.
- Nếu không chỉnh sửa vấn đề phương sai thay đổi khi có hiện tượng này sẽ dẫn đến kết luận sai về ý nghĩa thống kê của các tham số ước lượng theo hướng có tác động trong khi trên thực tế là không.

Chuẩn đoán mô hình hồi quy (Regression Diagnostics)

Xây dựng và chuẩn đoán mô hình hồi quy

1. Thống kê mô tả dữ liệu: phát hiện khác biệt giữa các nhóm, quan sát ngoại vi, phát hiện nếu dữ liệu phân phối bất đối xứng
2. Kiểm tra tính tương quan giữa các biến giải thích (multicolinearity/correlation)
3. Ước lượng mô hình hồi quy đơn giản và mở rộng
4. Phát hiện và xử lý nghi vấn về cấu trúc hàm (tuyến tính hoặc phi tuyến, biến tương tác)
5. Hậu hồi quy: rà soát những vấn đề có thể xảy ra và lựa chọn mô hình phù hợp:
    - Thực hiện các loại kiểm định
    - Hệ số phóng đại phương sai - Variance Inflation Factors (VIF)
    - Đánh giá tác động của quan sát ngoại vi
    - Đồ thị phần dư

Lưu ý với mô hình hồi quy đa biến

- Chọn biến giải thích cần dựa trên lý thuyết kinh tế thay vì ý nghĩa thống kê. Với mẫu quan sát lớn, việc tăng số mẫu sẽ làm tăng sự tương quan ngẫu nhiên, mặc dù thực tế không có bất kỳ liên hệ nào giữa các biến đó.

- Tránh đưa quá nhiều biến giải thích trong mô hình, kể cả những biến không thực sự liên quan nhằm tăng hệ số thích hợp (R^2). Sử dụng R^2_{adj} để lựa chọn biến phù hợp.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\sum_i (\hat{y}_i - \bar{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

$$R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-K}$$

- Tránh chọn lọc điều chỉnh dữ liệu sao cho mô hình có kết quả phù hợp với định kiến có trước.

Khi dữ liệu có phân phối lệch (skewed distribution)

- Mặc dù các giả định để ước lượng OLS là BLUE (giả định 1-5) không liên quan đến phân phối của **dữ liệu**, tuy nhiên, dữ liệu phân phối lệch *có thể* làm cho phương sai của **sai số** thay đổi hoặc điều kiện phân phối chuẩn của **sai số** (giả định 6 để có mô hình CLRM) bị vi phạm.

- Nếu có phân phối lệch, cần thiết phải kiểm tra ý nghĩa của biến số về mặt kinh tế. Ví dụ khi ước lượng mô hình liên quan đến tỷ suất, biến phụ thuộc thường là logarit $\Rightarrow$ chuyển đổi dữ liệu sang hàm log có thể hạn chế được vấn đề phân phối lệch.

Phát hiện và xử lý vấn đề liên quan đến cấu trúc hàm

- Kiểm định giả thuyết bội F và Chow với biến bậc cao, biến tương tác.
- Kiểm định Breusch-Pagan và White về phương sai thay đổi và điều chỉnh nếu cần thiết.
- Kiểm định Ramsey về cấu trúc hàm sai (misspecification test).

Kiểm định mô hình sai - RESET test

Kiểm định Ramsey RESET (Regression Specification Error Test) để kiểm định mô hình sai trong trường hợp tổng quát, khác với F-test hay Chow-test kiểm định các cấu trúc hàm cho trước (bậc 2, bậc 3...):

- Giả định ta có mô hình hồi quy đa biến sau:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \tag{5}$$

- Kiểm định RESET dùng để nhận biết liệu cấu trúc hàm trên bị sai

Các bước thực hiện kiểm định RESET

1. Ước lượng mô hình (5), tính giá trị dự báo $\hat{y}$
2. Đưa giá trị dự báo bình phương và bậc ba vào mô hình gốc và ước lượng hồi quy phụ:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + u$$

3. Kiểm định giả thuyết $H_0 : \gamma_1 = \gamma_2 = 0$ bằng kiểm định $F_{2,n-k-3}$ với $df = 2$. Nếu bác bỏ H_0 thì hàm hồi quy (5) có vấn đề về cấu trúc hàm

Thực hành kiểm định RESET

Sử dụng lại mô hình tỷ suất thu nhập với bộ dữ liệu VHLSS 2010

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexpersq + \beta_4 married \\ &+ \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u \end{aligned}$$

- Kiểm định liệu cấu trúc hàm trên có sai không?
- Tuy nhiên kiểm định này (và tất cả các loại kiểm định nói chung) không cho phép tìm mô hình chuẩn. Nếu mô hình bị sai thì có thể chỉnh sửa bằng cách thêm các biến bậc 2, bậc 3, biến tương tác, biến giải thích khác...
- Chọn lựa mô hình tối ưu thường phải do lý thuyết kinh tế quyết định thay vì chỉ dựa các thủ thuật kiểm định thống kê

Hậu hồi quy (Post-regression diagnostics)

Hệ số phóng đại phương sai - Variance Inflation Factor (VIF):

- Sử dụng để đo lường độ tương quan giữa các biến. Nếu các biến tự tương quan với nhau được sử dụng trong cùng một mô hình sẽ dẫn đến ước lượng phương sai bị chệch và kiểm định giả thuyết không chính xác.

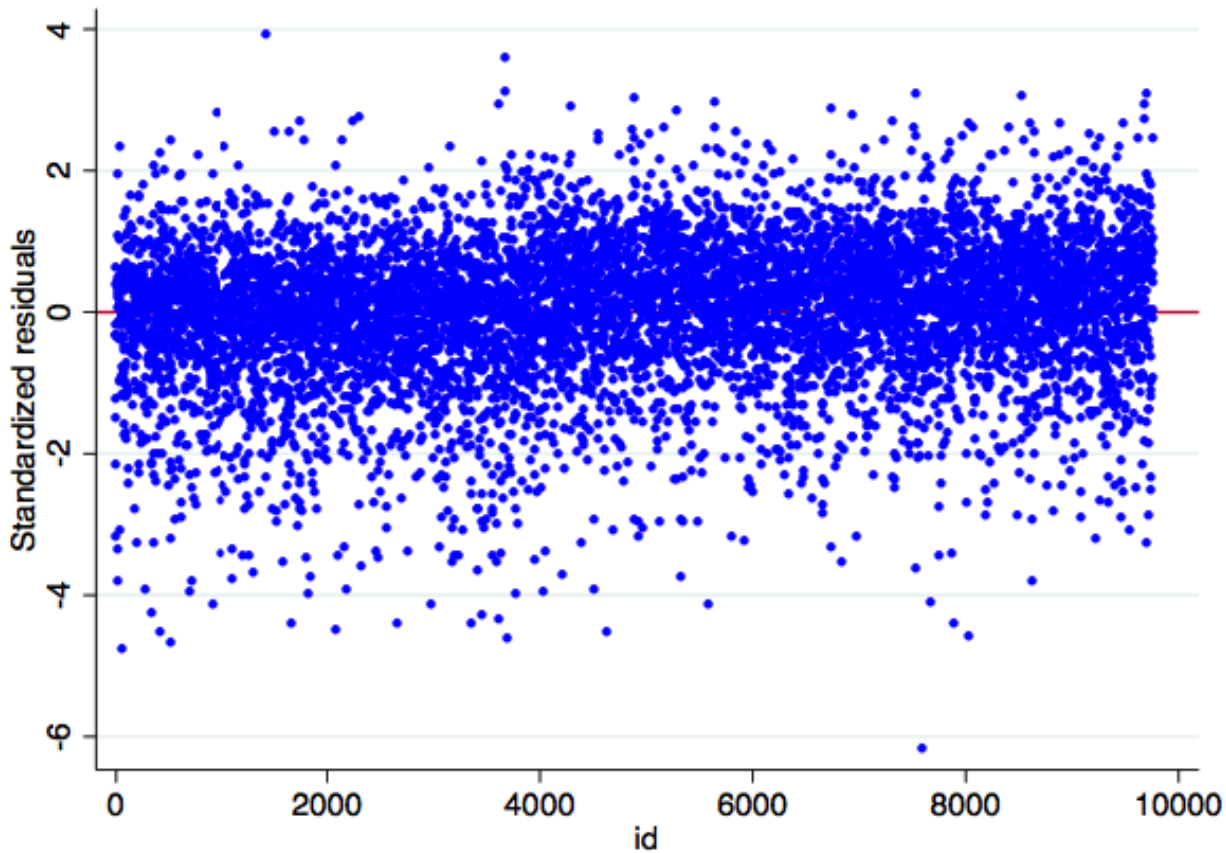
- Cần lọc ra những biến quan trọng nhất (về mặt thống kê).
  VIF được tính bằng cách hồi quy mỗi biến giải thích X_i dựa vào các biến khác,

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Quy ước bỏ biến có $VIF > 10$.

Kiểm tra đồ thị phân phối của phần dư:

- Kiểm tra có tồn tại quan sát ngoại vi hay không?
- Kiểm tra liệu có vấn đề phương sai thay đổi, tự tương quan hoặc tương quan chuỗi hay không?



Quan sát ngoại vi - Outliers

- Phát hiện dựa vào thống kê mô tả và đồ thị phân phối
  - Vẽ đồ thị boxplot hoặc histogram để xác định liệu có quan sát ngoại vi
  - Lấy logarithm của dữ liệu có phân phối lệch có thể xử lý được vấn đề quan sát ngoại vi (nếu phù hợp với lý thuyết kinh tế)
  - Bỏ các quan sát ngoại vi và ước lượng lại mô hình xem kết quả có biến động lớn không
- Điều chỉnh mô hình theo trọng số bằng phương pháp WLS

Các vấn đề liên quan đến dữ liệu

- Dữ liệu không ngẫu nhiên (vấn đề lựa chọn mẫu/sample selection) có thể xảy ra do quá trình kinh tế hay quá trình điều tra dẫn đến thông tin quan sát được không đảm bảo đại diện cho quần thể.
  - Để nhận diện vấn đề lựa chọn mẫu cần phải hiểu sâu về bối cảnh nghiên cứu và cách thức điều tra thu thập thông tin.
  - Nếu dữ liệu không ngẫu nhiên, ước lượng bằng OLS sẽ bị chệch. Cần nhận diện OLS bị chệch theo hướng nào.
  - Để xử lý vấn đề lựa chọn mẫu cần có kỹ thuật phức tạp (Định lượng ứng dụng).

- Dữ liệu bị thiếu/missing values:
  - Thiếu ngẫu nhiên hay thiếu có hệ thống?
  - Khi nào thì loại bỏ quan sát bị thiếu thông tin không ảnh hưởng đến kết quả mô hình?
  - Ghép thông tin (data imputation)
  - Thiếu thông tin quan trọng: Cần kỹ thuật hoặc thiết kế nghiên cứu phức tạp để xử lý (Định lượng ứng dụng).