

Chương 5.

MỘT SỐ VẤN ĐỀ TRONG MÔ HÌNH HỒI QUY

Chương này đề cập tới ba vấn đề thường xảy ra trong mô hình, vi phạm giả thiết của phương pháp OLS: Đa cộng tuyến, phương sai nhiễu thay đổi, tự tương quan của nhiễu. Đồng thời, trong một chừng mực nào đó, chỉ ra nguyên nhân, phát hiện vấn đề và tìm cách khắc phục, hạn chế những hậu quả không tốt của chúng.

5.1. Đa cộng tuyến

5.1.1. Khái niệm về đa cộng tuyến

a. Xét mô hình hồi quy k biến:

$$Y = a_0 + a_1X_1 + a_2X_2 + \dots + a_{k-1}X_{k-1} + U \quad (5.1)$$

Giả thiết 4 của phương pháp OLS là ma trận

$$X = \begin{pmatrix} 1 & X_{11} & \dots & X_{k-1,1} \\ 1 & X_{12} & \dots & X_{k-1,2} \\ \vdots & \vdots & \dots & \vdots \\ 1 & X_{1n} & \dots & X_{k-1,n} \end{pmatrix}$$

có hạng bằng k, tức là k cột của ma trận này không phải là k véc tơ phụ thuộc tuyến tính. Khi các biến giải thích không tương quan với nhau, mỗi biến chứa đựng những thông tin riêng về Y, không liên quan đến các biến giải thích khác. Khi đó hệ số hồi quy riêng của mỗi biến giải thích cho biết ảnh hưởng của biến này đối với biến phụ thuộc khi các biến khác không đổi. Trong trường hợp này ta nói mô hình không có hiện tượng đa cộng tuyến.

Ta nói mô hình có hiện tượng đa cộng tuyến (*multicollinearity*) nếu tồn tại các hằng số không đồng thời bằng 0: $\lambda_1, \lambda_2, \dots, \lambda_{k-1}$ và biến ngẫu nhiên ε sao cho:

$$\lambda_1 \cdot X_1 + \lambda_2 \cdot X_2 + \dots + \lambda_{k-1} \cdot X_{k-1} = \varepsilon$$

- Khi $\varepsilon \equiv 0$ thì hiện tượng đa cộng tuyến được gọi là đa cộng tuyến hoàn hảo (*perfect multicollinearity*) (Khi đó rõ ràng giả thiết 4 nói trên bị vi phạm)
- Khi $\varepsilon \neq 0$ thì hiện tượng đa cộng tuyến được gọi là đa cộng tuyến không hoàn hảo, (*imperfect multicollinearity*), hay đơn giản là đa cộng tuyến.

b. Nguyên nhân của hiện tượng đa cộng tuyến:

Những nguyên nhân chính là:

* Khi các biến giải thích có mối quan hệ nhân quả cao, tức là có những quan hệ ràng buộc.

Chẳng hạn: trong mô hình hồi quy của Y là lượng điện năng tiêu thụ theo các biến giải thích là: thu nhập X_1 , diện tích nhà ở X_2 , thì sẽ xảy ra hiện tượng đa cộng tuyến vì thu nhập cao thường kéo theo diện tích nhà ở lớn hơn.

* Khi các số liệu quá ít thì chúng vừa không đủ tính đại diện cho tổng thể, lại không xác định được duy nhất các hệ số hồi quy.

* Chọn biến giải thích có độ biến thiên nhỏ.

* Phương pháp chọn mẫu không đủ tính đại diện.

5.1.2. Hậu quả của đa cộng tuyến

Trong thực tế hiện tượng đa cộng tuyến là không tránh khỏi, vấn đề là mức độ đa cộng tuyến là cao hay thấp. Khi mô hình có hiện tượng đa cộng tuyến đáng kể giữa các biến giải thích thì mặc dù tính chất BLUE của các hệ số ước lượng vẫn được bảo toàn, nhưng xuất hiện các hậu quả không tốt sau:

1/ Các hệ số ước lượng có phương sai và hiệp phương sai lớn, nghĩa là các ước lượng này có giá trị thay đổi nhiều từ mẫu này qua mẫu khác, khiến độ chính xác của các ước lượng không cao.

Để thấy rõ điều này, xét mô hình SRF ba biến: $\hat{Y} = \hat{a}_0 + \hat{a}_1X_1 + \hat{a}_2X_2$, ta có:

$$\begin{aligned} var(\hat{a}_1) &= \frac{\sigma^2}{nS^2(X_1).(1-r_{12}^2)}; \quad var(\hat{a}_2) = \frac{\sigma^2}{nS^2(X_2).(1-r_{12}^2)}; \quad (*) \\ cov(\hat{a}_1, \hat{a}_2) &= \frac{-r_{12}\sigma^2}{nS(X_1).S(X_2).(1-r_{12}^2)}; \end{aligned}$$

trong đó r_{12} là hệ số tương quan mẫu giữa X_1, X_2 . Khi mô hình có hiện tượng đa cộng tuyến cao thì $|r_{12}|$ gần đến 1, do đó giá trị tuyệt đối của các biểu thức trên trở nên rất lớn.

2/ Từ hậu quả trên mà khoảng tin cậy cho các hệ số hồi quy rộng hơn, nghĩa là ước lượng có độ chính xác kém đi.

3/ Khi sử dụng thống kê $t = \frac{\hat{b}_j - b_j^*}{se\hat{b}_j}$ để kiểm định giả thuyết $H_0: b_j = b_j^*$, nếu có đa cộng tuyến ở mức độ cao thì các sai số chuẩn của các ước lượng có xu hướng tăng cao, dẫn tới giá trị $|t|$ có xu hướng nhỏ đi, do đó ta có xu hướng chấp nhận giả thuyết H_0 .
4/ Trong khi $|t|$ bé đi thì hệ số xác định R^2 có thể rất cao, dẫn tới những kết luận không phù hợp với thực tế.

5/ Dấu của các hệ số hồi quy ước lượng có thể sai

6/ Các ước lượng $\hat{b}_j$ qua OLS cho các hệ số hồi quy và $se(\hat{b}_j)$ trở nên rất nhạy với những thay đổi nhỏ trong số liệu.

7/ Do các hậu quả trên mà khi thêm vào hay bớt đi các biến cộng tuyến với các biến khác thì mô hình sẽ có sự thay đổi về dấu hoặc độ lớn của các ước lượng.

5.1.3. Cách phát hiện đa cộng tuyến

Như đã chỉ ra, hiện tượng đa cộng tuyến là không tránh khỏi. Người ta đưa ra một số quy tắc kinh nghiệm nhằm phát hiện và đánh giá mức độ đa cộng tuyến như sau.

a/ *Hệ số xác định R^2 cao nhưng giá trị $|t|$ thấp*: đây là một điều mâu thuẫn trong mô hình mà mức độ đa cộng tuyến thấp hoặc không có. Khi $R^2 > 0,8$ thì thường giả thuyết về các hệ số hồi quy đồng thời bằng 0 bị bác bỏ, nhưng khi $|t|$ có giá trị bé thì lại có xu hướng chấp nhận giả thuyết nói trên. Hiện tượng này chỉ thể hiện rõ khi có đa cộng tuyến ở mức độ cao.

b/ *Các cặp biến giải thích có hệ số tương quan cao*: Khi thấy hệ số tương quan cặp giữa các biến giải thích $> 0,8$ thì kinh nghiệm cho thấy hiện tượng đa cộng tuyến trở nên nghiêm trọng (tuy nhiên đây chỉ là điều kiện cần nếu mô hình nhiều hơn 2 biến).

c/ *Sử dụng các hồi quy phụ*: Chạy mô hình hồi quy của một biến giải thích X_j với các biến giải thích còn lại (gọi là hồi quy phụ), ta nhận được hệ số xác định của mô hình này,

ký hiệu là R_j^2 . Theo quy tắc “ngón tay cái” (Rule of Thumb) của Klein, hiện tượng đa cộng tuyến là nghiêm trọng chỉ nếu có hệ số xác định R_j^2 của hồi quy phụ nào đó vượt quá hệ số xác định R^2 của mô hình hồi quy chính của biến phụ thuộc.
d/ *Sử dụng nhân tử phóng đại phương sai VIF*: Nhân tử phóng đại của hồi quy phụ của biến X_j là:

$$VIF_j = \frac{1}{1-R_j^2}$$

Quy tắc kinh nghiệm là khi $VIF_j > 10$ hay $R_j^2 > 0,9$ thì dễ có hiện tượng đa cộng tuyến ở mức độ cao.

5.1.4. Biện pháp khắc phục đa cộng tuyến

Nếu mục tiêu của phân tích hồi quy là dự báo thì trong một số trường hợp, ta không cần khắc phục đa cộng tuyến. Nếu mục tiêu của phân tích hồi quy là xét tác động riêng của từng biến giải thích lên biến phụ thuộc để quyết định chính sách thì đa cộng tuyến trở thành một vấn đề nghiêm trọng. Sau đây là một số biện pháp khắc phục.

a/ *Dùng thông tin tiên nghiệm (A priori information)* Thông tin tiên nghiệm có thể nhận được từ các nghiên cứu thực nghiệm trước đây, hoặc từ các lý thuyết liên quan đến các biến giải thích ta đang xét.

Chẳng hạn khi nghiên cứu hàm sản xuất Cobb – Douglas ở Mexico giai đoạn 1955-1974 trong chương trước, ta có mối quan hệ giữa sản lượng Y (đầu ra) phụ thuộc vào các yếu tố đầu tư như lao động X_1 và vốn X_2

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + U \quad (\beta_0 = \ln \alpha)$$

Kết quả thực nghiệm cho thấy giữa vốn và lao động ở Mexico có quan hệ là sản lượng không đổi theo quy mô, tức là: $\beta_1 + \beta_2 = 1$. Nếu sử dụng kết quả thực nghiệm này như là một thông tin tiên nghiệm thì ta có biến đổi mối quan hệ trên về dạng:

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + (1 - \beta_1) \ln X_2 + U,$$

hay: $\ln \left(\frac{Y}{X_2} \right) = \beta_0 + \beta_1 \ln (X_1/X_2) + U$ (là mô hình hồi quy 2 biến)

b/ *Tăng cỡ mẫu (bổ sung thêm số liệu) hoặc lấy thêm mẫu mới*

Với mẫu mới thu được theo cách này, trong nhiều trường hợp người ta hy vọng nó sẽ làm giảm mức độ đa cộng tuyến cao trong mẫu cũ, chẳng hạn trong các hệ thức (*) ở trên, nếu mẫu mới không làm tăng trị tuyệt đối của hệ số tương quan mẫu r_{12} , thì các biểu thức $var(\hat{a}_1)$; $var(\hat{a}_2)$; $cov(\hat{a}_1, \hat{a}_2)$ sẽ nhỏ đi do các phương sai mẫu $S^2(X_1), S^2(X_2)$ tăng lên.

c/ *Kết hợp số liệu chéo và số liệu chuỗi thời gian*

Trong số liệu chuỗi thời gian, thường ẩn chứa vấn đề đa cộng tuyến giữa các biến giải thích. Khi kết hợp thêm số liệu chéo, sẽ khắc phục hay hạn chế bớt mức độ đa cộng tuyến giữa các biến.

d/ *Bỏ bớt biến giải thích trong các biến có cộng tuyến với nhau*

Đây là biện pháp khắc phục khá đơn giản. Tuy nhiên, khi bỏ bớt biến giải thích có thể tránh được đa cộng tuyến cao, nhưng có thể gây nên hậu quả nghiêm trọng là dẫn đến ước lượng chệch nhiều so với giá trị thực của tham số cần ước lượng (trong khi vấn đề đa cộng tuyến không làm thay đổi tính không chệch của ước lượng)

e/ *Sử dụng sai phân cấp 1 (first difference)*

Theo diễn biến của thời gian, các biến kinh tế thường chịu ảnh hưởng của xu hướng nên dễ có tương quan với nhau. Để làm giảm sự tương quan đó, ta có thể sử dụng sai phân cấp 1.

Giả sử theo thời gian t , có mô hình: $Y_t = a_0 + a_1X_{1t} + a_2X_{2t} + U_t$
thì tại thời điểm $t - 1$, ta có: $Y_{t-1} = a_0 + a_1X_{1,t-1} + a_2X_{2,t-1} + U_{t-1}$

Từ đó:

$$Y_t - Y_{t-1} = a_1(X_{1t} - X_{1,t-1}) + a_2(X_{2t} - X_{2,t-1}) + (U_t - U_{t-1}) \quad (5.2)$$

(5.2) được gọi là mô hình sai phân cấp 1, được sử dụng để ước lượng các tham số hồi quy a_1, a_2 . Giữa X_{1t}, X_{2t} nếu có đa cộng tuyến cao thì giữa $(X_{1t} - X_{1,t-1})$, $(X_{2t} - X_{2,t-1})$ có thể không xảy ra đa cộng tuyến cao. Vì thế mô hình sai phân có thể làm giảm mức độ đa cộng tuyến.

Khi sử dụng mô hình sai phân cần lưu ý nhược điểm của nó là bậc tự do giảm đi 1 do giảm đi một quan sát khi chuyển sang mô hình sai phân, nên dễ ảnh hưởng đến kết quả ước lượng khi cỡ mẫu bé; mặc dù U_t có thể không có tự tương quan, nhưng $V_t = (U_t - U_{t-1})$ thì có thể có tự tương quan; hơn nữa việc sử dụng sai phân cấp 1 không thích hợp với số liệu chéo.

f/ *Thay đổi dạng hàm hồi quy*: Nếu ở dạng hàm hồi quy này, các biến giải thích có hiện tượng đa cộng tuyến, thì chuyển sang dạng khác có thể khắc phục được hiện tượng này.

g/ *Một số biện pháp khác*: Ngoài các biện pháp nói trên, để khắc phục vấn đề đa cộng tuyến, tùy vào các trường hợp cụ thể, người ta còn sử dụng các biện pháp khác như: Sử dụng hàm hồi quy độ lệch theo giá trị trung bình trong hồi quy đa thức, hồi quy thành phần chính, hồi quy dạng sóng,...

Khắc phục hiện tượng đa cộng tuyến đòi hỏi các kỹ thuật phức tạp và đôi khi không mang lại hiệu quả như mong muốn. Hơn nữa hầu hết mô hình hồi quy bội đều có tính đa cộng tuyến nhất định nên ta phải thận trọng trong việc xây dựng mô hình và giải thích kết quả.

Ví dụ 5.1: Khi nghiên cứu về quan hệ giữa tiêu dùng nội địa Y (USD), thu nhập X_1 từ lương, thu nhập khác X_2 từ phi nông nghiệp và thu nhập X_3 từ nông nghiệp của nền kinh tế Mỹ từ năm 1928

đến 1950, với số liệu của các năm 1942 đến 1944 bị loại ra khỏi dữ liệu, từ bảng số liệu:

N	Y	X ₁	X ₂	X ₃	N	Y	X ₁	X ₂	X ₃
1928	52.8	39.21	17.73	4.39	1938	63.9	44.16	15.92	4.37
1929	62.2	42.31	20.29	4.6	1939	67.5	47.68	17.59	4.51
1930	58.6	40.37	18.83	3.25	1940	71.3	50.79	18.49	4.9
1931	56.6	39.15	17.44	2.61	1941	76.6	57.78	19.18	6.37
1932	51.6	34	14.76	1.67	1942	86.3	78.97	19.12	8.42
1033	51.1	33.59	13.39	2.44	1946	95.7	73.54	19.76	9.27
1034	54	36.88	13.93	2.39	1947	98.3	74.92	17.55	8.87
1035	57.2	39.27	14.67	5	1948	100.3	74.01	19.17	9.3
1936	62.8	45.51	17.20	3.93	1949	103.2	75.51	20.20	6.95
1937	65	46.06	17.15	5.48	1950	108.9	80.97	22.12	7.15

Bảng 5.1

Klein và Golberger (1995) đã thực hiện hồi quy tiêu dùng Y theo 3 loại thu nhập trên như

sau: $Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + U$

Dependent Variable: Y

Method: Least Squares

Sample: 1 20

Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.304002	8.882885	0.822256	0.4230
X1	1.135052	0.172127	6.594285	0.0000
X2	0.405300	0.645026	0.628347	0.5386
X3	-0.405888	1.105135	-0.367274	0.7182
R-squared	0.954028	Mean dependent var		72.19500
Adjusted R-squared	0.945409	S.D. dependent var		19.34671
S.E. of regression	4.520317	Akaike info criterion		6.031898
Sum squared resid	326.9323	Schwarz criterion		6.231044

Bảng 5.2. Kết quả hồi quy của tiêu dùng theo các loại thu nhập

Kết quả này cho thấy mô hình có tính giải thích cao thể hiện qua $R^2 = 0,954028$ rất cao. Tuy nhiên xuất hiện những vấn đề không phù hợp với ý nghĩa kinh tế, đó là hệ số hồi quy của X3 là $-0,405888 < 0$ và hệ số hồi quy của X1 là 1,135052 cho thấy: khi thu nhập từ lương tăng 1 USD thì bình quân tiêu dùng tăng 1,135052 USD ! Đây là những biểu hiện cho thấy hồi quy này gặp phải hiện tượng đa cộng tuyến và điều này là do các loại thu nhập có xu hướng cùng tăng theo sự phát triển của kinh tế.

Xét ma trận tương quan giữa các biến:

	Y	X1	X2	X3
Y	1.000000	0.975908	0.717164	0.887671
X1	0.975908	1.000000	0.709395	0.918613
X2	0.717164	0.709395	1.000000	0.630607
X3	0.887671	0.918613	0.630607	1.000000

Bảng 5.3

ta thấy hệ số tương quan giữa X1 và X3 là 0.918613, rất cao. Đây cũng là một biểu hiện của hiện tượng đa cộng tuyến cao giữa các biến giải thích.

* Nếu bỏ bớt biến X3, ta có hồi quy:

Dependent Variable: Y

Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.710141	8.586595	0.897928	0.3818
X1	1.080588	0.085136	12.69245	0.0000
X2	0.423208	0.626601	0.675402	0.5085
R-squared	0.953641	Mean dependent var		72.19500
Adjusted R-squared	0.948187	S.D. dependent var		19.34671

Bảng 5.4

theo đó ta vẫn chưa khắc phục được hiện tượng hệ số hồi quy không phù hợp với lý thuyết kinh tế: hệ số hồi quy của X1 là **1.080588** > 1.

* Nếu dùng hồi quy sai phân cấp 1, ta nhận được kết quả từ Eviews như sau:

Dependent Variable: Y-Y(-1)

Method: Least Squares

Sample (adjusted): 2 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X1-X1(-1)	0.339464	0.143755	2.361398	0.0312
X2-X2(-1)	1.515549	0.518401	2.923507	0.0099
X3-X3(-1)	0.728987	0.667778	1.091661	0.2911
R-squared	0.460379	Mean dependent var		2.952632
Adjusted R-squared	0.392926	S.D. dependent var		4.153896
S.E. of regression	3.236505	Akaike info criterion		5.330805
Sum squared resid	167.5995	Schwarz criterion		5.479927
Log likelihood	-47.64264	Hannan-Quinn criter.		5.356042
Durbin-Watson stat	1.014878			

Bảng 5.5

theo đó mô hình ít phù hợp với số liệu ($R^2 = 0,460379$), mặt khác hệ số hồi quy của (X2-X2(-1)) là 1,515549 > 1.

Do vậy đối với mô hình này, để khắc phục những hiện tượng trên, ta phải kết hợp các biện pháp khác nhau: bổ sung thêm số liệu, kết hợp thêm các số liệu chéo, bỏ bớt biến trong các biến có đa cộng tuyến cao, thay đổi mô hình,.... Để khắc phục hiện tượng này, ta sẽ trở lại ví dụ 5.1 trong phần sau, khi thay đổi dạng hàm hồi quy sang tuyến tính log.

5.2. Phương sai của nhiễu thay đổi

5.2.1. Khái niệm về phương sai thay đổi

Giả thiết 2 của mô hình hồi quy tuyến tính cổ điển yêu cầu phương sai của nhiễu không thay đổi qua các quan sát. Do trung bình của nhiễu bằng 0 nên yêu cầu này có nghĩa là:

$$\text{var}(U_i) = E(U_i^2) = \sigma^2$$

Trong thực tế sai số nhiễu có thể tăng, giảm khi giá trị của các biến giải thích thay đổi, tức là:

$$\text{var}(U_i) = E(U_i^2) = \sigma_i^2 \quad (5.3)$$

Khi đó ta nói có hiện tượng phương sai nhiễu thay đổi (*heteroscedasticity*). Hiện tượng phương sai thay đổi thường gặp ở dữ liệu chéo và dữ liệu bảng.

Có thể chỉ ra những lý do sau đây:

* Do việc tích lũy kinh nghiệm hay do học được hành vi trong quá khứ mà sai số theo thời gian ngày càng giảm. Chẳng hạn đối với thợ học việc, khi số giờ thực hành càng nhiều thì số phế phẩm càng nhỏ và càng ít biến động. Trong trường hợp này phương sai nhiễu có xu hướng giảm theo thời gian.

* Do bản chất của mối liên hệ mà có nhiều mối quan hệ kinh tế đã chứa đựng hiện tượng này, khi biến kinh tế tăng kéo theo sai số nhiễu cũng tăng. Chẳng hạn khi thu nhập tăng

người ta có nhiều lựa chọn hơn trong tiêu dùng. Khi đó trong hồi quy của tiết kiệm theo thu nhập thì phương sai nhiều có xu hướng tăng theo thu nhập.

- * Khi cải thiện phương pháp và kỹ thuật thu thập số liệu thì sai số càng giảm.
- * Khi trong mẫu có các số liệu vượt trội (quá lớn hoặc quá bé so với tập số liệu) cũng khiến cho phương sai thay đổi.
- * Không xác định đúng dạng mô hình, thiếu biến quan trọng.
- * Trong mô hình sử dụng số liệu chéo cũng khiến cho phương sai không đồng đều.

5.2.2. Hậu quả của phương sai thay đổi

- * Các ước lượng OLS tuy vẫn còn tính chất tuyến tính không chệch, nhưng không còn là ước lượng hiệu quả nữa.
- * Phương sai của sai số bị tính sai nên việc dùng thống kê t và thống kê F để kiểm định giả thuyết không còn đáng tin cậy nữa (thống kê t không chắc có phân phối student), các trị của t – stat và sai số chuẩn của hệ số ước lượng do phần mềm cung cấp trở nên vô dụng.
- * Kết quả dự báo không hiệu quả khi dựa trên các ước lượng OLS có phương sai không nhỏ nhất.

5.2.3. Cách phát hiện phương sai nhiều thay đổi

Việc phát hiện ra có hiện tượng này trong thực tế không đơn giản vì ta chỉ có thể dựa vào mẫu chứ không thể có toàn bộ thông tin về tổng thể. Vì thế ta không thể có một phương pháp chắc chắn để phát hiện ra phương sai thay đổi, mà chỉ có thể dựa vào một số công cụ sau đây để chẩn đoán giúp ta phát hiện ra hiện tượng này:

a/ *Bản chất của vấn đề nghiên cứu*: Bản chất của vấn đề nghiên cứu khiến ta phải nghĩ tới khả năng xảy ra hiện tượng này, chẳng hạn khi ta dùng các số liệu chéo liên quan đến các đơn vị không thuần nhất, khác nhau về quy mô.

b/ *Xem xét đồ thị của phần dư*: Đó là đồ thị của sai số của hồi quy (hay phần dư) đối với biến giải thích X nào đó hoặc đối với giá trị ước lượng $\hat{Y}$. Phương sai của phần dư được chỉ ra bằng độ rộng của biểu đồ phân rải của phần dư khi X hoặc $\hat{Y}$ tăng. Nếu độ rộng này tăng hoặc giảm thì giả thiết về phương sai không đổi có thể bị vi phạm.

- Đối với mô hình hồi quy bội, người ta thường khảo sát đồ thị phần dư $\hat{U}^2$ đối với $\hat{Y}$.

c/ *Dùng các phương pháp kiểm định*:

c1/ *Kiểm định Park*: Kiểm định Park dựa trên cơ sở giả định rằng phương sai nhiều thay đổi dưới dạng hàm lũy thừa của biến giải thích X :

$$\sigma_i^2 = \sigma^2 \cdot X_i^\beta \cdot e^{V_i} \quad (5.4)$$

lấy log hai vế ta nhận được:

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \cdot \ln X_i + V_i \quad (5.5)$$

Vì σ_i^2 chưa biết nên Park thay σ_i^2 bởi $\hat{U}_i^2$ (có được từ hồi quy gốc) trong (5.5), nhận được:

$$\ln \hat{U}_i^2 = \alpha + \beta \cdot \ln X_i + V_i \quad (\alpha = \ln \sigma^2) \quad (5.6)$$

Khi đó kiểm định Park gồm các bước sau:

B1. Thực hiện hồi quy gốc: $Y = a + b \cdot X + U$, nhận được các ước lượng: $\hat{Y}_i$ và $\hat{U}_i$.

B2. Thực hiện hồi quy: $\ln \hat{U}_i^2 = \alpha + \beta \cdot \ln X_i + V_i$.

B3. Tiến hành kiểm định giả thuyết

$$H_0: \beta = 0 (\text{phương sai không đổi}), H_1: \beta \neq 0 (\text{phương sai thay đổi})$$

Chú ý:

* Đối với mô hình hồi quy bội, các bước tiến hành là tương tự như đối với hồi quy đơn, trong đó có thể hồi quy $\ln \hat{U}_i^2$ theo mỗi biến độc lập hoặc theo $\hat{Y}_i$.

* Trong kiểm định Park, nhiều V_i phải thỏa mãn các giả thiết cổ điển.

c2. *Kiểm định White*: Kiểm định White khảo sát phần dư $\hat{U}_i^2$ theo các biến độc lập. Kiểm định này không đòi hỏi nhiều U_i phải có phân phối chuẩn.

Giả sử ta đang xét mô hình hồi quy gốc:

$$Y = a_0 + b_1 X_1 + b_2 X_2 + U \quad (5.7)$$

Kiểm định White gồm các bước sau:

B1. Hồi quy mô hình gốc (5.7), tìm được các phần dư $\hat{U}_i$.

B2. Hồi quy mô hình phụ:

$$U_i^2 = \alpha_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{2i}^2 + \beta_5 X_{1i} X_{2i} + V_i \quad (5.8)$$

Từ đó nhận được hệ số xác định của mô hình này, ký hiệu là: R_{aut}^2

Mô hình phụ có thể có số mũ cao hơn và nhất thiết phải có hệ số chặn α_0 , bất kể mô hình gốc có hay không có hệ số chặn a_0 .

B3. Tiến hành kiểm định

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 (\text{phương sai không thay đổi})$$

Trên cơ sở H_0 đúng thì người ta chỉ ra được rằng: $n \cdot R_{aut}^2$ có phân phối xấp xỉ $\chi^2(df)$, với bậc tự do $df =$ số tham số của mô hình phụ (5.8), không kể hệ số chặn (trong trường hợp này $df = 5$). Vì thế:

- Nếu $n \cdot R_{aut}^2 > \chi_{\alpha}^2(df)$ thì bác bỏ H_0 .

c3. *Kiểm định Glejser*: Tương tự như kiểm định Park, kiểm định Glejser coi nhiều có thể thay đổi theo biến độc lập X, nhưng theo một trong các dạng hàm:

$$|\hat{U}_i| = \alpha_0 + \alpha_1 X_i + V_i; \quad |\hat{U}_i| = \alpha_0 + \alpha_1 \sqrt{X_i} + V_i; \quad (5.9)$$

$$|\hat{U}_i| = \alpha_0 + \alpha_1 \frac{1}{X_i} + V_i; \quad |\hat{U}_i| = \alpha_0 + \alpha_1 \frac{1}{\sqrt{X_i}} + V_i; \quad (5.10)$$

$$|\hat{U}_i| = \sqrt{\alpha_0 + \alpha_1 X_i} + V_i; \quad |\hat{U}_i| = \sqrt{\alpha_0 + \alpha_1 X_i^2} + V_i \quad (5.11)$$

Kiểm định giả thuyết phương sai thay đổi ở đây là kiểm định giả thuyết: $H_0: \alpha_1 = 0$, đối thuyết $H_1: \alpha_1 \neq 0$.

Lưu ý:

* Kiểm định Glejser yêu cầu nhiều V_i thỏa mãn các giả thiết cổ điển.

* Các mô hình (5.11) không phải là mô hình tuyến tính nên không dùng được phương pháp OLS.

c4. *Kiểm định Goldfeld – Quandt*:

Nếu ta phát hiện phương sai nhiều tương quan thuận với một biến giải thích X nào đó dưới dạng: $\sigma_i^2 = \sigma^2 \cdot X_i^2$ (σ^2 là hằng số) thì sử dụng kiểm định Goldfeld – Quandt, theo các bước sau:

B1. Sắp xếp số liệu theo thứ tự tăng dần của X

B2. Loại bỏ c quan sát nằm ở giữa, (n – c) quan sát còn lại chia làm 2 nhóm, mỗi nhóm có (n – c)/2 quan sát.

B3. Thực hiện hồi quy OLS đối với mô hình gốc: $Y_i = a + bX_i + U$ với $(n - c)/2$ quan sát đầu ta được RSS_1 (gọi là nhóm phương sai nhỏ) và với $(n - c)/2$ quan sát cuối ta được RSS_2 (gọi là nhóm phương sai lớn) và chúng đều có bậc tự do là $df = (n - c - 2k)/2$ (k là số tham số trong mô hình)

B4. Để xác minh phương sai của hai nhóm có sự khác biệt đáng kể hay không, ta tiến hành kiểm định F với giả thiết H_0 : *phương sai không đổi* như sau: Trên cơ sở H_0 là đúng, người ta chỉ ra được đại lượng: $F = \frac{RSS_2/df}{RSS_1/df}$ có phân phối F với các bậc tự do (df, df) .

Do đó nếu $F > F_{\alpha}(df, df)$ thì bác bỏ H_0 , tức là chấp nhận phương sai có thay đổi.

Lưu ý:

* Mặc dù độ tin cậy của kết luận phụ thuộc vào c , nhưng ta lại không có quy tắc nào để xác định giá trị c cho tốt nhất. Theo kinh nghiệm, người ta thường chọn c như sau:

- Nếu n xấp xỉ 30 thì chọn $c = 4$ hoặc $c = 8$,
- Nếu n xấp xỉ 60 thì chọn $c = 10$ hoặc $c = 16$.

* Kiểm định *Goldfeld – Quandt* thích hợp với những mẫu nhỏ.

* Đối với mô hình hồi quy bội, ta có thể sắp xếp các quan sát theo một biến bất kỳ trong các biến giải thích của mô hình. Khi không có thông tin tiên nghiệm để biết biến giải thích nào là thích hợp, ta có thể thực hiện kiểm định Park đối với mỗi biến giải thích.

5.2.4. Biện pháp khắc phục

Do hậu quả của phương sai thay đổi, biện pháp khắc phục là hết sức cần thiết. Việc khắc phục được chia ra hai trường hợp: biết hay chưa biết σ_i^2 . Trước khi đi vào các biện pháp khắc phục, ta trình bày các phương pháp bình phương bé nhất có trọng số và phương pháp bình phương bé nhất tổng quát.

1. Phương pháp bình phương bé nhất có trọng số

Xét mô hình hai biến: $Y_i = a + b.X_i + U_i$

Trước đây, để nhận được các ước lượng, phương pháp OLS nhằm cực tiểu tổng bình phương các phần dư: $\sum_{i=1}^n U_i^2 = \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}.X_i)^2$ (5.12)

Bây giờ ta đặt cho mỗi phần dư U_i^2 một trọng số: $W_i = \frac{1}{\sigma_i^2}$, (trong đó $var(U_i|X_i) = var(Y_i|X_i) = \sigma_i^2$) với lý do là: khi có hiện tượng phương sai nhiều thay đổi thì ta không thể đặt mức độ tin cậy các quan sát như nhau, quan sát nào ít sai lệch thì mức độ tin cậy sẽ cao hơn.

Để nhận được các ước lượng cho a , b , theo phương pháp bình phương bé nhất có trọng số, ta cực tiểu hóa tổng bình phương các phần dư có trọng số:

$$\sum_{i=1}^n W_i . U_i^2 = \sum_{i=1}^n W_i (Y_i - a^* - b^*.X_i)^2 \rightarrow \min \quad (5.13)$$

Về trái (5.13) là hàm bậc 2 đối với các biến a^* , b^* nên việc cực tiểu hóa hàm này cho ta các ước lượng:

$$b^* = \frac{\sum_{i=1}^n W_i . \sum_{i=1}^n W_i . X_i . Y_i - \sum_{i=1}^n W_i . X_i . \sum_{i=1}^n W_i . Y_i}{\sum_{i=1}^n W_i . \sum_{i=1}^n W_i . X_i^2 - (\sum_{i=1}^n W_i . X_i)^2}; a^* = \bar{Y}^* - b^* \bar{X}^*; \quad (5.14)$$

(trong đó: $\bar{X}^* = \sum_{i=1}^n W_i . X_i / \sum_{i=1}^n W_i$; $\bar{Y}^* = \sum_{i=1}^n W_i . Y_i / \sum_{i=1}^n W_i$)

2. Phương pháp bình phương bé nhất tổng quát GLS (Generalized Least Squares)

Xét mô hình hai biến: $Y_i = a + b.X_i + U_i$ (5.15)

trong đó tất cả các giả thiết của mô hình hồi quy tuyến tính cổ điển đều được thỏa mãn, trừ giả thiết phương sai nhiễu không đổi bị vi phạm.

Mục đích của phương pháp GLS biến đổi từ mô hình có phương sai nhiễu thay đổi sang mô hình có phương sai nhiễu không thay đổi.

Đặt $X_{0i} = 1, \forall i$ và chia 2 vế của (5.15) cho σ_i ta nhận được mô hình:

$$Z_i = a \cdot D_{0i} + b \cdot D_i + V_i \quad (5.16)$$

(trong đó: $Z_i = \frac{Y_i}{\sigma_i}$; $D_{0i} = \frac{X_{0i}}{\sigma_i}$; $D_i = \frac{X_i}{\sigma_i}$; $V_i = \frac{U_i}{\sigma_i}$)

Mô hình (5.16) thỏa mãn tất cả các giả thiết của mô hình tuyến tính cổ điển, với phương sai nhiễu không đổi ($\text{var}(V_i) = \text{var}\left(\frac{U_i}{\sigma_i}\right) = 1$). Vì thế dùng phương pháp OLS cho mô hình (5.16) ta nhận được các ước lượng không chệch tuyến tính tốt nhất cho a và b là $\hat{a}^*$ và $\hat{b}^*$ sau đây:

$$\hat{b}^* = \frac{\sum_{i=1}^n W_i \cdot \sum_{i=1}^n W_i \cdot X_i \cdot Y_i - \sum_{i=1}^n W_i \cdot X_i \cdot \sum_{i=1}^n W_i \cdot Y_i}{\sum_{i=1}^n W_i \cdot \sum_{i=1}^n W_i \cdot X_i^2 - (\sum_{i=1}^n W_i \cdot X_i)^2}; \hat{a}^* = \bar{Z} - \hat{b}^* \cdot \bar{D}. \quad (5.17)$$

(để ý: $\text{var} \hat{b}^* = \frac{\sum_{i=1}^n W_i}{\sum_{i=1}^n W_i \cdot \sum_{i=1}^n W_i \cdot X_i^2 - (\sum_{i=1}^n W_i \cdot X_i)^2}$.)

Phương pháp tìm các ước lượng $\hat{a}^*$ và $\hat{b}^*$ vừa chỉ ra gọi là *Phương pháp bình phương bé nhất tổng quát (GLS)* (*phương pháp OLS là trường hợp riêng của GLS*)

3. Biện pháp khắc phục: Ta chia các trường hợp để khắc phục hiện tượng này như sau:

a. Khi biết σ_i^2 : Sử dụng phương pháp GLS nói trên.

b. Khi chưa biết σ_i^2 : Ta vẫn sử dụng phương pháp GLS, nhưng đòi hỏi phải có những giả thiết nhất định sau đây về phương sai tổng thể:

Giả thiết 1: Phương sai tổng thể tỷ lệ với bình phương của biến giải thích:

$$\text{var}(U_i) = E(U_i^2) = \sigma^2 \cdot X_i^2 \quad (5.18)$$

Khi đó từ mô hình gốc $Y_i = a + b \cdot X_i + U_i$, ta đưa về mô hình:

$$\frac{Y_i}{X_i} = a \cdot \frac{1}{X_i} + b + \frac{U_i}{X_i} \quad (5.19)$$

(5.19) có phương sai nhiễu: $\text{var}\left(\frac{U_i}{X_i}\right) = \frac{\text{var} U_i}{X_i^2} = \sigma^2, \forall i$

(Lưu ý rằng phép lấy mẫu đối với X là không ngẫu nhiên mà xác định trước nên các thành phần mẫu X_i xem là các hằng số). Trong thực tế ta dùng $\hat{U}_i$ để ước lượng cho U_i , vì thế người ta thường khảo sát $\hat{U}_i^2$ theo X_i . Đòi hỏi quy bội, có thể dùng đồ thị biểu diễn $\hat{U}_i^2$ theo từng biến giải thích, hoặc sử dụng hồi quy phụ $\hat{U}_i^2$ theo bình phương của từng biến giải thích, qua đó đánh giá được biến giải thích nào thích hợp với giả thiết 1 nhiều nhất để tiến hành biến đổi trên biến giải thích này. Tuy nhiên cần đề phòng trường hợp biến đổi mô hình gốc theo một biến nào đó dẫn đến vi phạm một giả thiết cổ điển khác.

Giả thiết 2: Phương sai tổng thể tỷ lệ với biến độc lập, tức là:

$$\text{var}(U_i) = E(U_i^2) = \sigma^2 \cdot X_i \quad (5.20)$$

Khi đó: từ mô hình gốc $Y_i = a + b \cdot X_i + U_i$, ta đưa về mô hình:

$$\frac{Y_i}{\sqrt{X_i}} = a \cdot \frac{1}{\sqrt{X_i}} + b \cdot \sqrt{X_i} + V_i \quad (5.21)$$

trong đó $V_i = \frac{U_i}{\sqrt{X_i}}$, có $varV_i = var\left(\frac{U_i}{\sqrt{X_i}}\right) = \sigma^2$

Mô hình (5.21) có phương sai nhiều không thay đổi và là mô hình hồi quy tuyến tính qua gốc. Sau khi chạy hồi quy mô hình này, ta có mô hình ước lượng cho mô hình gốc bằng cách nhân 2 vế của mô hình nhận được với $\sqrt{X_i}$.

Giả thiết 3: Phương sai của nhiễu tỷ lệ với bình phương của kỳ vọng của Y, tức là: $var(U_i) = E(U_i^2) = \sigma^2 \cdot (EY_i)^2$

Khi đó: từ mô hình gốc $Y_i = a + b \cdot X_i + U_i$, ta đưa về mô hình:

$$\frac{Y_i}{EY_i} = a \cdot \frac{1}{EY_i} + b \cdot \frac{X_i}{EY_i} + \frac{U_i}{EY_i} \quad (5.22)$$

(5.22) là mô hình tuyến tính cổ điển có phương sai nhiều: $varV_i = var\left(\frac{U_i}{EY_i}\right) = \sigma^2$

Tuy nhiên trong mô hình (5.22) ta chưa biết được EY_i (do a, b chưa biết), ta sẽ thay EY_i bằng một ước lượng của nó. Ta tiến hành theo các bước sau:

B1. Chạy hồi quy mô hình gốc bằng phương pháp OLS, thu được $\hat{Y}_i$ là một ước lượng vững cho EY_i . Dùng $\hat{Y}_i$ đưa mô hình gốc về dạng:

$$\frac{Y_i}{\hat{Y}_i} = a \cdot \frac{1}{\hat{Y}_i} + b \cdot \frac{X_i}{\hat{Y}_i} + \frac{U_i}{\hat{Y}_i} \quad (5.22a)$$

B2. Chạy hồi quy mô hình (5.22a), từ đó nhận được mô hình hồi quy gốc.

Lưu ý: Vì trong (5.22a), ta xấp xỉ EY_i bằng ước lượng vững $\hat{Y}_i$ của nó, nên khi cỡ mẫu khá lớn thì sai số trong xấp xỉ này sẽ bé và mô hình là chấp nhận được.

Giả thiết 4: Dùng mô hình tuyến tính log thay thế:

$$\ln Y_i = a + b \cdot \ln X_i + V_i \quad (5.23)$$

Ví dụ 5.2: Bảng 5.6 dưới đây cho số liệu về chi phí đầu tư Y (triệu USD) cho việc nghiên cứu và phát triển của 18 ngành công nghiệp ở Mỹ trong năm 1988, trong đó nhóm các ngành công nghiệp được đánh số thứ tự từ 1 đến 18, X_2 (triệu USD) là số liệu về doanh thu, X_1 (triệu USD) là lợi nhuận. Ta muốn xét tác động của doanh thu đối với đầu tư cho phát triển như thế nào qua việc ước lượng mô hình hồi quy sau:

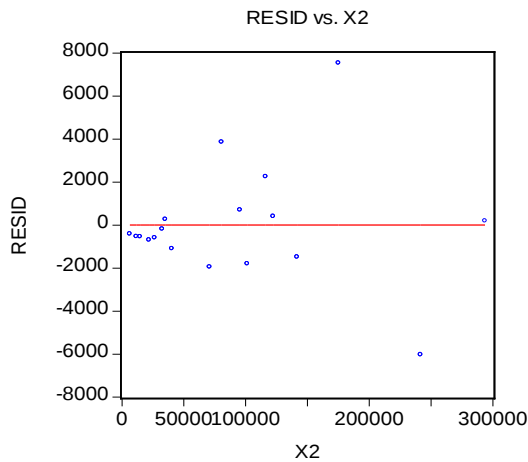
$$Y_i = \alpha + \beta \cdot X_2 + U_i$$

với hy vọng khi doanh thu tăng thì đầu tư cho nghiên cứu và phát triển cũng sẽ tăng, mà việc nghiên cứu và phát triển có ảnh hưởng tích cực đối với các nhóm ngành nên làm tăng lợi nhuận, tức là giữa Y và X_2 có mối quan hệ đồng biến.

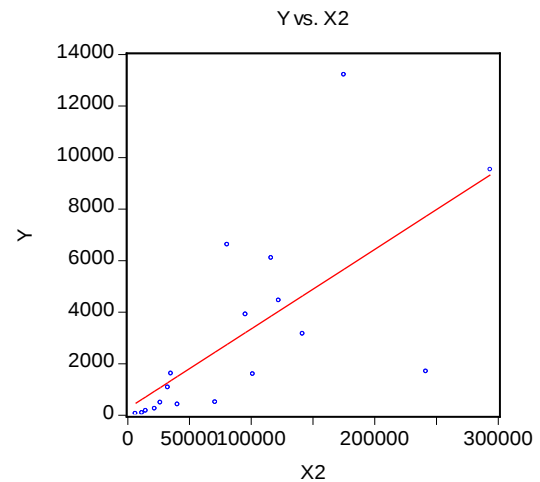
STT	Y	X_2	X_1	STT	Y	X_2	X_1
1	62.5	6375.3	185.1	10	6620.1	80552.8	13869.9
2	92.9	11626.4	1569.5	11	3918.6	95294	4487.8
3	178.3	14655.1	276.8	12	1595.3	101314.1	10278.9
4	258.4	21869.2	2828.1	13	6107.5	116141.3	8787.3
5	494.7	26408.3	2225.9	14	4454.1	122315.7	16438.8
6	1083	32405.6	3751.9	15	3163.8	141649.9	9761.4
7	1620.6	35107.7	2884.1	16	13210.7	175025.8	19774.5
8	421.7	40295.4	4645.7	17	1703.8	241434.8	23168.5
9	509.2	70761.6	5036.4	18	9528.2	293543	18415.4

Bảng 5.6

Khảo sát các biểu đồ phân tán:

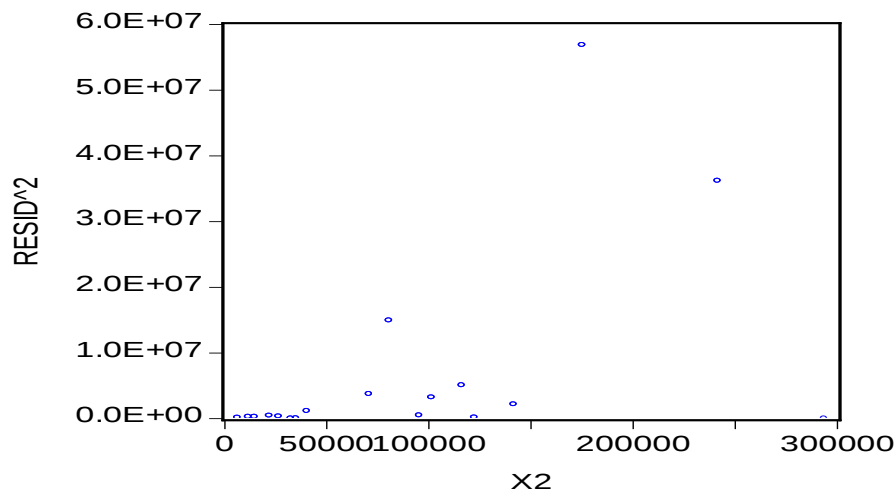


Hình 5.1a



Hình 5.1b

Hình 5.1a cho thấy các điểm phân tán có xu thế đi lên theo chiều tăng của X_2 và khi X_2 càng lớn thì các điểm phân tán càng dần rộng ra. Điều này cho thấy khi doanh thu tăng thì bình quân đầu tư cũng tăng và phương sai nhiều (đo mức độ phân tán) cũng tăng, tức là phương sai thay đổi. Điều này có thể được lý giải bởi số liệu sử dụng là số liệu chéo, từ các ngành nghề khác nhau với quy mô và đặc điểm khác nhau.



Hình 5.1c

Để thấy rõ hơn hiện tượng này, ta khảo sát hình 5.1b biểu diễn sự biến thiên của phần dư theo doanh thu, hình 5.1c biểu diễn sự biến thiên của bình phương phần dư theo doanh thu.

Chạy hồi quy đầu tư và phát triển theo doanh thu (mô hình 2 biến), nhận được:

Dependent Variable: Y
Method: Least Squares
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	266.1917	1002.961	0.265406	0.7941
X2	0.030878	0.008346	3.699582	0.0019
R-squared	0.461042	Mean dependent var		3056.856
Adjusted R-squared	0.427357	S.D. dependent var		3705.973
S.E. of regression	2804.428	Akaike info criterion		18.82023
Sum squared resid	1.26E+08	Schwarz criterion		18.91916
Log likelihood	-167.3820	Hannan-Quinn criter.		18.83387
F-statistic	13.68690	Durbin-Watson stat		3.020747
Prob(F-statistic)	0.001944			

Bảng 5.7. Kết quả hồi quy đầu tư và phát triển theo doanh thu

Chạy hồi quy của đầu tư và phát triển Y theo X_1 và X_2 (mô hình 3 biến), ta có kết quả sau:

Dependent Variable: Y
Method: Least Squares
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.644362	1013.043	-0.002610	0.9980
X1	0.251118	0.207017	1.213031	0.2439
X2	0.010947	0.018375	0.595769	0.5602
R-squared	0.509189	Mean dependent var		3056.856
Adjusted R-squared	0.443747	S.D. dependent var		3705.973
S.E. of regression	2764.002	Akaike info criterion		18.83776
Sum squared resid	1.15E+08	Schwarz criterion		18.98615
Log likelihood	-166.5398	Hannan-Quinn criter.		18.85822
F-statistic	7.780819	Durbin-Watson stat		3.170338
Prob(F-statistic)	0.004807			

Bảng 5.8. Hồi quy Đầu tư theo doanh thu và lợi nhuận

Tiến hành kiểm định White đối với mô hình 3 biến:

Heteroskedasticity Test: White

F-statistic	20.18959	Prob. F(5,12)	0.0000
Obs*R-squared	16.08761	Prob. Chi-Square(5)	0.0066
Scaled explained SS	23.57634	Prob. Chi-Square(5)	0.0003

Bảng 5.9. Kết quả kiểm định White về phương sai thay đổi

Từ bảng này ta có: giá trị $p - value = 0,0066 < 0,05$ nên bác bỏ H_0 . Vậy ta chấp nhận có hiện tượng phương sai thay đổi.

- Cách khắc phục: Lần lượt theo một trong các giả thiết:

$$a/ E(U_i^2) = \sigma_i^2 = \sigma^2 \cdot X_{2i}^2$$

Với giả thiết này, chạy hồi quy ước lượng cho mô hình: $\frac{Y}{X_2} = \alpha_0 + \alpha_1 \cdot \frac{1}{X_2} + \alpha_2 \cdot \frac{X_1}{X_2} + V$ ta nhận được kết quả sau:

Dependent Variable: Y/X2				
Method: Least Squares				
Sample: 1 18				
Included observations: 18				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.022272	0.015723	1.416528	0.1771
1/X2	-197.8795	142.5285	-1.388351	0.1853
X1/X2	0.141385	0.140051	1.009526	0.3287
R-squared	0.220286	Mean dependent var		0.029063
Adjusted R-squared	0.116324	S.D. dependent var		0.023120
S.E. of regression	0.021734	Akaike info criterion		-4.668853
Sum squared resid	0.007086	Schwarz criterion		-4.520458
Log likelihood	45.01968	Hannan-Quinn criter.		-4.648392
F-statistic	2.118911	Durbin-Watson stat		2.665118
Prob(F-statistic)	0.154709			

Bảng 5.10. Điều chỉnh mô hình để khắc phục

Theo đó nhận được mô hình ước lượng:

$$\left(\frac{Y}{X_2}\right) = 0.022272 - 197.8795 \cdot \frac{1}{X_2} + 0.141385 \cdot \frac{X_1}{X_2}$$

Đối với mô hình vừa nhận được, dùng kiểm định White có số hạng tích chéo:

Heteroskedasticity Test: White			
F-statistic	1.821390	Prob. F(5,12)	0.1830
Obs*R-squared	7.766404	Prob. Chi-Square(5)	0.1696
Scaled explained SS	3.258843	Prob. Chi-Square(5)	0.6601

Bảng 5.11

Ta có: $p - value = 0,1696 > 0,05$ nên chấp nhận H_0 . Vậy trong mô hình vừa nhận được không còn hiện tượng phương sai thay đổi.

$$b/ E(U_i^2) = \sigma_i^2 = \sigma^2 \cdot X_{2i}$$

Đối với trường hợp này, chạy hồi quy ước lượng cho mô hình:

$$\frac{Y}{\sqrt{X_2}} = \frac{\alpha_0}{\sqrt{X_2}} + \alpha_1 \cdot \frac{X_1}{\sqrt{X_2}} + \alpha_2 \cdot \sqrt{X_2} + V$$

ta nhận được kết quả sau:

Dependent Variable: Y/SQR(X2)
Method: Least Squares
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
1/SQR(X2)	-243.4290	367.5355	-0.662328	0.5178
SQR(X2)	0.011638	0.017296	0.672900	0.5112
X1/SQR(X2)	0.272748	0.174788	1.560450	0.1395
R-squared	0.445020	Mean dependent var		8.850796
Adjusted R-squared	0.371023	S.D. dependent var		8.837239
S.E. of regression	7.008643	Akaike info criterion		6.883177
Sum squared resid	736.8161	Schwarz criterion		7.031572
Log likelihood	-58.94859	Hannan-Quinn criter.		6.903639
Durbin-Watson stat	3.035036			

Bảng 5.12. Điều chỉnh mô hình để khắc phục

Dùng kiểm định White có số hạng tích chéo:

Heteroskedasticity Test: White

F-statistic	5.746517	Prob. F(5,12)	0.0062
Obs*R-squared	12.69712	Prob. Chi-Square(5)	0.0264
Scaled explained SS	10.32900	Prob. Chi-Square(5)	0.0664

Bảng 5.13

$p - value = 0.0264 < 0.05$ nên bác bỏ H_0 : vẫn còn hiện tượng phương sai thay đổi trong mô hình ba biến. Điều này cho thấy giả thiết $E(U_i^2) = \sigma_i^2 = \sigma^2$. X_{2i} là không thích hợp.

c/ $var(U_i) = E(U_i^2) = \sigma^2 \cdot (EY_i)^2$

Với giả thiết này, ước lượng cho mô hình: $\frac{Y}{\hat{Y}} = a_0 \cdot \frac{1}{\hat{Y}} + a_1 \cdot \frac{X_1}{\hat{Y}} + a_2 \cdot \frac{X_2}{\hat{Y}} + V$,

ta nhận được kết quả sau:

Dependent Variable: Y/YDB
Method: Least Squares
Sample: 1 18
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
1/YDB	-144.6701	92.52261	-1.563619	0.1388
X1/YDB	0.091016	0.123207	0.738726	0.4715
X2/YDB	0.024967	0.011741	2.126469	0.0505
R-squared	0.140061	Mean dependent var		0.869221
Adjusted R-squared	0.025402	S.D. dependent var		0.598553
S.E. of regression	0.590902	Akaike info criterion		1.936678
Sum squared resid	5.237474	Schwarz criterion		2.085073
Log likelihood	-14.43010	Hannan-Quinn criter.		1.957140
Durbin-Watson stat	2.563496			

Bảng 5.14

Theo đó ta có mô hình ước lượng:

$$\frac{Y}{\hat{Y}} = -144.6701 \cdot \frac{1}{\hat{Y}} + 0.091016 \cdot \frac{X_1}{\hat{Y}} + 0.024967 \cdot \frac{X_2}{\hat{Y}} + \hat{V}$$

Dùng kiểm định White đối với mô hình này, nhận được kết quả:

Heteroskedasticity Test: White

F-statistic	0.655352	Prob. F(5,12)	0.6636
Obs*R-squared	3.860875	Prob. Chi-Square(5)	0.5696
Scaled explained SS	1.282882	Prob. Chi-Square(5)	0.9367

Bảng 5.15

Kết quả trên cho thấy p – value = 0,5696 > 0,05, vậy ta chấp nhận giả thuyết H_0 : mô hình điều chỉnh nhận được không còn hiện tượng phương sai nhiều thay đổi.
d. Nếu dùng mô hình tuyến tính log thay thế:

$$\ln Y = a + b \ln X_1 + c \ln X_2 + V$$

ta nhận được kết quả hồi quy ước lượng:

Dependent Variable: LOG(Y)

Method: Least Squares

Sample: 1 18

Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6.553704	2.411367	-2.717838	0.0159
LOG(X1)	0.173952	0.352393	0.493631	0.6287
LOG(X2)	1.113761	0.441872	2.520548	0.0235
R-squared	0.793743	Mean dependent var		7.109987
Adjusted R-squared	0.766242	S.D. dependent var		1.606119
S.E. of regression	0.776535	Akaike info criterion		2.483063
Sum squared resid	9.045107	Schwarz criterion		2.631458
Log likelihood	-19.34756	Hannan-Quinn criter.		2.503524
F-statistic	28.86237	Durbin-Watson stat		2.464834
Prob(F-statistic)	0.000007			

Bảng 5.16

Có SRF ngẫu nhiên:

$$\ln Y = -6.553704 + 0.173952 \ln X_1 + 1.113761 \ln X_2 + \hat{V}$$

Dùng kiểm định White đối với mô hình này, ta có:

Heteroskedasticity Test: White

F-statistic	0.699894	Prob. F(5,12)	0.6340
Obs*R-squared	4.064039	Prob. Chi-Square(5)	0.5402
Scaled explained SS	2.078399	Prob. Chi-Square(5)	0.8382

Bảng 5.17

Theo đó p – value = 0,5402 > 0,5. Vậy ta chấp nhận giả thuyết H_0 , tức là mô hình thay thế này không còn hiện tượng phương sai nhiều thay đổi.

Nhận xét: Trong 4 trường hợp giả thiết trên, mô hình tuyến tính log không có hiện tượng phương sai nhiều thay đổi và tỏ ra phù hợp hơn cả vì có hệ số xác định $R^2 = 0.793743$ là cao nhất.

Ví dụ 5.3: Xét tập số liệu trong ví dụ 5.1, bỏ đi biến X_3 , thay đổi sang mô hình tuyến tính Lin-log:

$$Y = a_0 + a_1 \ln X_1 + a_2 \ln X_2 + U$$

ta nhận được kết quả hồi quy:

Dependent Variable: LOG(Y)

Method: Least Squares

Sample: 1 20

Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.990484	0.250201	3.958756	0.0010
LOG(X1)	0.810357	0.054504	14.86794	0.0000
LOG(X2)	0.028181	0.125057	0.225343	0.8244
R-squared	0.965967	Mean dependent var	4.247404	
Adjusted R-squared	0.961963	S.D. dependent var	0.255259	
S.E. of regression	0.049783	Akaike info criterion	-3.024800	
Sum squared resid	0.042132	Schwarz criterion	-2.875440	
Log likelihood	33.24800	Hannan-Quinn criter.	-2.995643	
F-statistic	241.2589	Durbin-Watson stat	1.405147	
Prob(F-statistic)	0.000000			

Bảng 5.18. Điều chỉnh mô hình để khắc phục

Theo đó mô hình có hệ số xác định $R^2 = 0,965967$ là rất cao, hơn nữa các hệ số hồi quy không có dấu hiệu bất thường. Với kết quả nhận được, dùng kiểm định White, ta có:

Heteroskedasticity Test: White

F-statistic	1.324761	Prob. F(5,14)	0.3097
Obs*R-squared	6.423455	Prob. Chi-Square(5)	0.2672
Scaled explained SS	14.89457	Prob. Chi-Square(5)	0.0108

Bảng 5.29

Theo đó: $p - \text{value} = 0,2672 > 0,05$, ta chấp nhận giả thuyết H_0 : mô hình tuyến tính log không có hiện tượng phương sai nhiều thay đổi.

5.3. Tự tương quan của nhiễu

5.3.1. Khái niệm về tự tương quan

Trong mô hình h.quy tuyến tính cổ điển, giả thiết 2 còn yêu cầu không có tương quan giữa các phần dư, tức là:

$$\text{cov}(U_i, U_j) = 0, \forall i \neq j, \text{ hay: } E(U_i \cdot U_j) = 0, \forall i \neq j$$

(do g. thiết: $EU_i = 0, \forall i$)

Ý nghĩa thực tế của yêu cầu này là: Nhiễu của quan sát này không bị ảnh hưởng bởi nhiễu của các quan sát khác.

Tuy nhiên trong thực tế đối với một chuỗi số liệu thì yêu cầu này dễ bị vi phạm. Hiện tượng này được gọi là tự tương quan (*Autocorrelation*) của nhiễu, đó là sự tương quan giữa các thành phần của dãy quan sát theo thời gian hoặc không gian. Như vậy hiện tượng tự tương quan có nghĩa là:

$$\exists i, j, i \neq j \text{ sao cho: } cov(U_i, U_j) \neq 0$$

Có thể chỉ ra các nguyên nhân sau đây:

1. Nguyên nhân khách quan:

- *Tính chất quán tính của dãy số liệu*: hầu hết số liệu chuỗi thời gian trong kinh tế đều có tính chất quán tính. Chẳng hạn số liệu theo thời gian về chỉ số giá, tỷ lệ thất nghiệp, GNP,... thường có tính chu kỳ và do đó trong hồi quy chuỗi thời gian thì các quan sát kế tiếp nhau có nhiều khả năng tương quan với nhau;

- *Sự tác động trễ (Lags) trong chuỗi thời gian*: số liệu tại thời điểm t chịu tác động bởi số liệu tại thời điểm $t - 1$ trước đó

- *Hiện tượng mạng nhện (Cobweb phenomenon)*: Khi lượng cung của một số mặt hàng phản ứng lại trước sự thay đổi của giá trễ hơn một khoảng thời gian vì các quyết định cung đòi hỏi phải có thời gian để thực hiện.

2. Nguyên nhân chủ quan:

- *Việc xử lý, làm trơn số liệu*: Trước khi sử dụng, số liệu thô thường được xử lý, làm trơn (chẳng hạn dùng phương pháp trung bình di động). Sự làm trơn này có thể dẫn tới sai số hệ thống trong các nhiễu và gây ra tự tương quan giữa chúng.

- *Phép nội suy (interpolation) và ngoại suy (extrapolation)* có thể gây ra sai số có tính chất hệ thống.

- *Định dạng hồi quy chưa phù hợp, đưa không đủ biến hay bỏ sót biến quan trọng trong mô hình.*

5.3.2. Hậu quả của hiện tượng tự tương quan

1. Các hệ số hồi quy ước lượng theo OLS không chệch nhưng không hiệu quả, tức là không còn tính chất BLUE.

2. Ước lượng của phương sai bị chệch nên các kiểm định t , F không còn tin cậy.

3. Ước lượng của hệ số R^2 tăng quá cao.

4. Các giá trị dự báo không còn đáng tin cậy.

5.3.3. Cách phát hiện có tự tương quan

1. Dựa vào biểu đồ phân tán

Trong mô hình hồi quy tuyến tính cổ điển, giả thiết không có tự tương quan gắn với các nhiễu U_t không quan sát được. Ta chỉ quan sát được các phần dư $\hat{U}_t = Y_t - \hat{Y}_t$. Mặc dù $\hat{U}_t$ không hoàn toàn giống U_t , nhưng nó là ước lượng của U_t nên quan sát các phần dư $\hat{U}_t$ có thể gợi ý cho ta những nhận xét về U_t . Vì thế để có thông tin về tự tương quan của nhiễu U , ta có thể khảo sát một trong các biểu đồ phân tán sau:

a/ Biểu đồ phân tán $(\hat{U}_t, t)$ của $\hat{U}_t$ (hoặc của $\hat{U}_t^2$) theo thời gian.

b/ Biểu đồ phân tán $(\frac{\hat{U}_t}{\hat{\sigma}}, t)$ của phần dư chuẩn hóa $\frac{\hat{U}_t}{\hat{\sigma}}$ theo thời gian t

Lưu ý rằng $U_t \sim N(0, \sigma^2)$ nên $\frac{U_t}{\sigma} \sim N(0, 1)$. Vì thế khi cỡ mẫu n khá lớn thì $\frac{\hat{U}_t}{\sigma}$ có phân phối xấp xỉ phân phối $N(0, 1)$.

c/ Biểu đồ phân tán $(\hat{U}_t, \hat{U}_{t-1})$ của $\hat{U}_t$ theo $\hat{U}_{t-1}$ (gọi là lược đồ AR(1))

- Nếu biểu đồ phân tán có dạng ngẫu nhiên thì không có tự tương quan, nếu biểu đồ phân tán có dạng không ngẫu nhiên, biểu thị xu hướng biến thiên có tính chất hệ thống thì nhận định có tự tương quan.

2/ Kiểm định Durbin - Watson

a. Xét thống kê:
$$d = \frac{\sum_{t=2}^n (\hat{U}_t - \hat{U}_{t-1})^2}{\sum_{t=1}^n \hat{U}_t^2} \quad (5.24)$$

Người ta chỉ ra được rằng khi n đủ lớn thì: $d \approx 2(1 - \hat{\rho})$

trong đó: $\hat{\rho} = \frac{\sum_{t=2}^n \hat{U}_t \cdot \hat{U}_{t-1}}{\sum_{t=1}^n \hat{U}_t^2}$ (hệ số tự tương quan mẫu bậc nhất) (5.25)

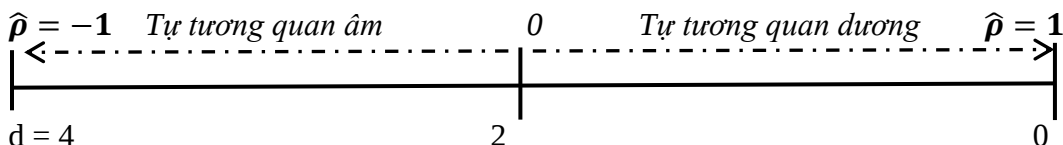
$\hat{\rho}$ là ước lượng của hệ số tự tương quan bậc nhất ρ trong mô hình tự hồi quy bậc nhất (hay tự tương quan bậc nhất):

$$U_t = \rho \cdot U_{t-1} + \varepsilon_t \quad (-1 \leq \rho \leq 1) \quad (AR(1)) \quad (5.26)$$

với ε_t là nhiễu ngẫu nhiên thỏa:

$$E\varepsilon_t = 0, cov(\varepsilon_t, \varepsilon_s) = 0, t \neq s, var\varepsilon_t = \sigma^2, \forall t \quad (5.27)$$

Nhận xét: từ $-1 \leq \rho, \hat{\rho} \leq 1$ suy ra: $0 \leq d \leq 4$



Hình 5.2. Hệ số tự tương quan bậc nhất và giá trị tổng kê d tương ứng

- Khi $d = 4$ hoặc $= 0$, ta có tự tương quan hoàn hảo. Khi $d = 2$ thì không có tự tương quan.

Bảng thống kê Durbin – Watson chỉ ra các giá trị tới hạn d_U, d_L dựa vào ba tham số: mức ý nghĩa α , số quan sát n , số biến độc lập k' .

b. Quy tắc kiểm định Durbin – Watson: Kiểm định giả thuyết mô hình có tự tương quan chính là $H_0: \rho = 0$, đối thuyết $H_1: \rho > 0 / \rho < 0 / \rho \neq 0$

Tính giá trị của thống kê d qua số liệu, so sánh d với các giá trị tới hạn để bác bỏ hay chấp nhận H_0 theo quy tắc sau:

* $H_0: \rho = 0, H_1: \rho > 0$, mức ý nghĩa α

0 (có tự tương quan dương)	d_U (không có tự tương quan dương)	4
----------------------------	--------------------------------------	---

* $H_0: \rho = 0, H_1: \rho < 0$, mức ý nghĩa α

0 (không có tự tương quan âm)	$4 - d_U$ (có tự tương quan âm)	4
-------------------------------	---------------------------------	---

* $H_0: \rho = 0, H_1: \rho \neq 0$, mức ý nghĩa 2α

0 (có ttq dương)	d_U (không có tự tương quan)	$4 - d_U$ (có ttq âm)	4
------------------	--------------------------------	-----------------------	---

Chú ý:

a/ Điều kiện để sử dụng kiểm định Durbin – Watson:

- * Mô hình hồi quy phải có hệ số bị chặn. Nếu mô hình không có hệ số bị chặn thì phải ước lượng mô hình có hệ số bị chặn để tính $RSS = \sum \hat{U}_i^2$, sau đó tiến hành kiểm định.
- * Việc lấy mẫu các biến độc lập là lấy mẫu xác định (không phải mẫu ngẫu nhiên).
- * Các nhiễu có tương quan bậc nhất: $U_t = \rho \cdot U_{t-1} + \varepsilon_t$ (AR(1))
- * Mô hình không có dạng tự hồi quy, tức là không xét mô hình dạng:

$$Y_t = a + b_1 X_t + b_2 Y_{t-i} + U_t$$

- * Không có quan sát bị mất trong dữ liệu.
- b/ Nhược điểm của kiểm định Durbin- Watson:

- * Khi cỡ mẫu n lớn thì không có trong bảng tra,
- * Có một số mâu thuẫn khi tra bảng tìm d_U , d_L (chẳng hạn khi $n = 9$, $k' = 3$, $\alpha = 5\%$ thì $4 - d_U < d_U$)

c/ Đôi khi người ta sử dụng tiêu chuẩn kiểm định Durbin-Watson theo kinh nghiệm như sau:

0	(có ttq dương)	1	(không có ttq)	3	(có tự tương quan âm)	4
---	----------------	---	----------------	---	-----------------------	---

3. Kiểm định Breusch – Godfrey (BG)

Xét mô hình hồi quy: $Y_t = a + bX_t + U_t$ (5.28)

với thành phần nhiễu có tự tương quan bậc p (AR(p)):

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + \dots + \rho_p U_{t-p} + \varepsilon_t$$
 (5.29)

trong đó ε_t là nhiễu ngẫu nhiên thỏa các giả thiết OLS.

Khi đó giả thuyết không có sự tương quan bậc p là: $H_0: \rho_1 = \rho_2 = \dots = \rho_p = 0$

Thủ tục kiểm định BG như sau:

B1: Chạy hồi quy mô hình (5.28) theo OLS và tìm được phần dư $\hat{U}_t$.

B2: Chạy hồi quy mô hình:

$$\hat{U}_t = \alpha + \beta \cdot X_t + \gamma_1 \hat{U}_{t-1} + \gamma_2 \hat{U}_{t-2} + \dots + \gamma_p \hat{U}_{t-p} + V_t$$
 (5.30)

từ đó tính được hệ số xác định của mô hình (5.30), ký hiệu là $R_{(5.30)}^2$.

B3: Nếu $(n - p) \cdot R_{(5.30)}^2 > \chi_{\alpha}^2(p)$ thì bác bỏ H_0 , tức là thừa nhận có tự tương quan bậc p.

Chú ý:

- * Kiểm định BG áp dụng cho cỡ mẫu lớn và mở rộng cho mô hình nhiều biến.
- * Kiểm định BG có thể áp dụng cho mô hình tự hồi quy (mô hình có biến giải thích Y_{t-1} , Y_{t-2} , ..., tức là có biến trễ).
- * Kiểm định BG áp dụng cho tự tương quan với bậc bất kỳ.
- * Kiểm định BG đòi hỏi phải xác định trước bậc của tự tương quan p. Trong thực tế người ta phải kiểm định với nhiều giá trị p khác nhau.
- * Kiểm định BG có thể được áp dụng cho mô hình có nhiễu U được tạo ra theo tiến trình trung bình động bậc q (MA(q): q^{th} – order Moving Average),
tức là: $U_t = \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + \dots + \lambda_q \varepsilon_{t-q}$

trong đó ε là nhiễu ngẫu nhiên với kỳ vọng bằng 0 và phương sai không đổi.

4. Kiểm định các đoạn mạch (hay kiểm định chuỗi dấu(Runs test)) (Tham khảo)

Số hạng nhiễu có giá trị khi âm, khi dương, do đó nếu sự thay đổi về dấu của số hạng nhiễu diễn ra mang tính hệ thống, theo một xu thế nào đó thì biểu hiện có sự tự tương