Phương sai thay đổi
(Heteroscedasticity)

Đinh Công Khải

Tháng 04/2016

Nội dung

1. Phương sai thay đổi là gì?
2. Hậu quả của phương sai thay đổi?
3. Làm sao để phát hiện?
4. Các biện pháp khắc phục?

Phương sai thay đổi là gì?

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$
- Giả thiết phương sai bằng nhau của mô hình CLRM

  $Var(Y_i \mid X_i) = Var(u_i \mid X_i) = E(u^2_i) = \sigma^2$ (i=1-n)

- Phương sai thay đổi

  $Var(Y_i \mid X_i) = Var(u_i \mid X_i) = E(u^2_i) = \sigma^2_i$ (i=1-n)

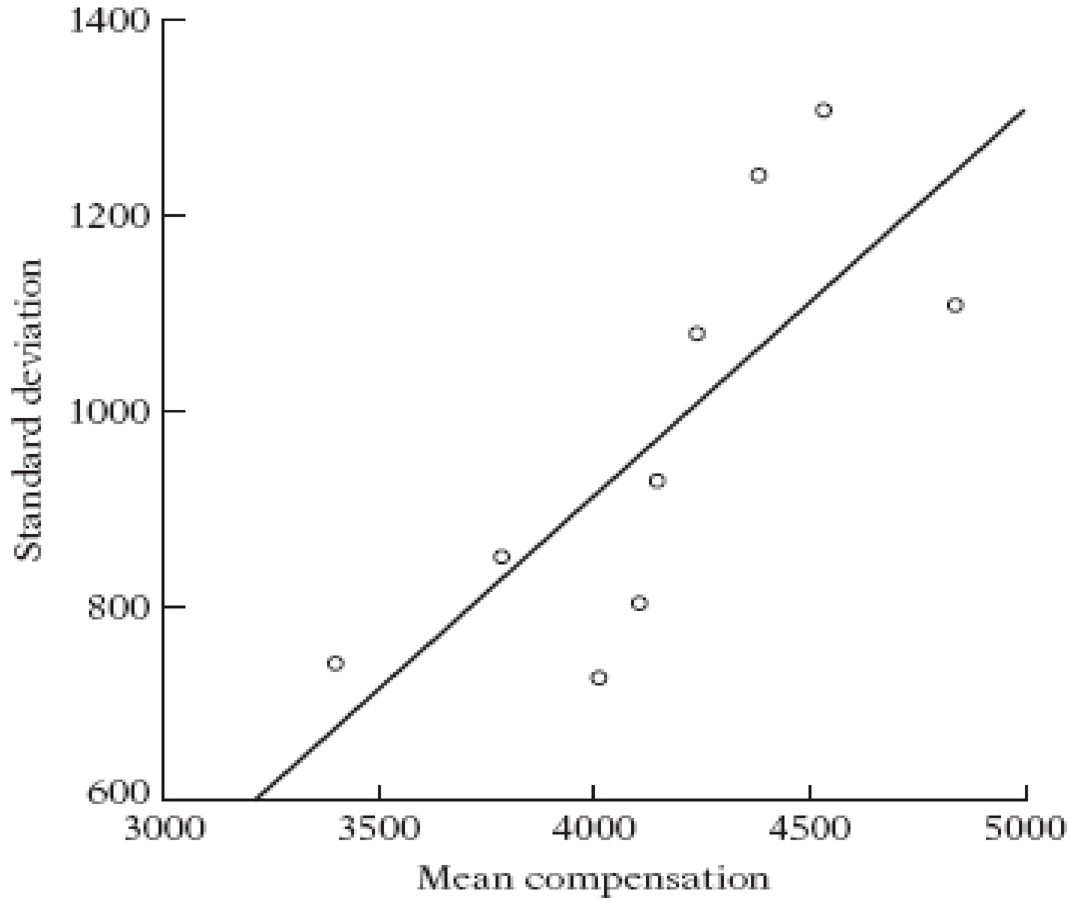
BẢNG 11.1

Mức lương lao động (USD) trong các ngành công nghiệp chế tạo sản phẩm không lâu bền theo quy mô lao động của doanh nghiệp, 1985

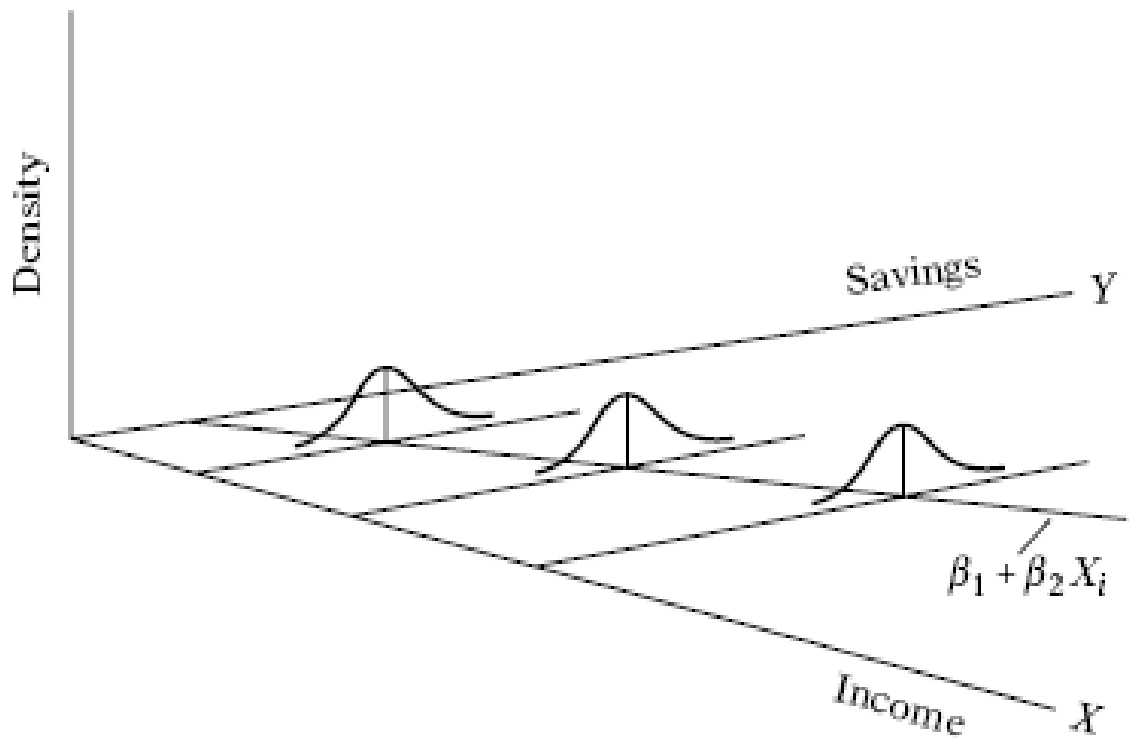
	Qui mô lao động (số công nhân trung bình)								
Ngành công nghiệp	1-4	5-9	10-19	20-49	50-99	100-249	250-499	500-999	1000-2499
Thực phẩm và các sản phẩm tương tự	2,994	3,295	3,565	3,907	4,189	4,486	4,676	4,968	5,342
Sản phẩm thuốc lá	1,721	2,057	3,336	3,320	2,980	2,848	3,072	2,969	3,822
Sản phẩm dệt	3,600	3,657	3,674	3,437	3,340	3,334	3,225	3,163	3,168
May mặc và các sản phẩm liên quan	3,494	3,787	3,533	3,215	3,030	2,834	2,750	2,967	3,453
Giấy và các sản phẩm từ gỗ	3,498	3,847	3,913	4,135	4,445	4,885	5,132	5,342	5,326
In và xuất bản	3,611	4,206	4,695	5,083	5,301	5,269	3,182	5,395	5,552
Hóa chất và các sản phẩm tương tự	3,875	4,660	4,930	5,005	5,114	5,248	5,630	5,870	5,876
Sản phẩm dầu lửa và than	4,616	5,181	5,317	5,337	5,421	5,710	6,316	6,455	6,347
Sản phẩm cao su và nhựa	3,538	3,984	4,014	4,287	4,221	4,539	4,721	4,905	5,481
Da và sản phẩm từ da	3,016	3,196	3,149	3,317	3,414	3,254	3,177	3,346	4,067
Lương lao động trung bình	3,396	3,787	4,013	4,014	4,146	4,241	4,387	4,538	4,843
Độ lệch chuẩn	743,7	851,4	727,8	805,1	929,9	1080,6	1243,2	1307,7	1112,5
Năng suất trung bình	9,355	8,584	7,962	8,275	8,389	9,418	9,795	10,281	11,750

Nguồn: The Census of Manufacturers (Tổng điều tra ngành công nghiệp chế tạo, Bộ Thương mại Hoa Kỳ, 1958 (đã tính toán).

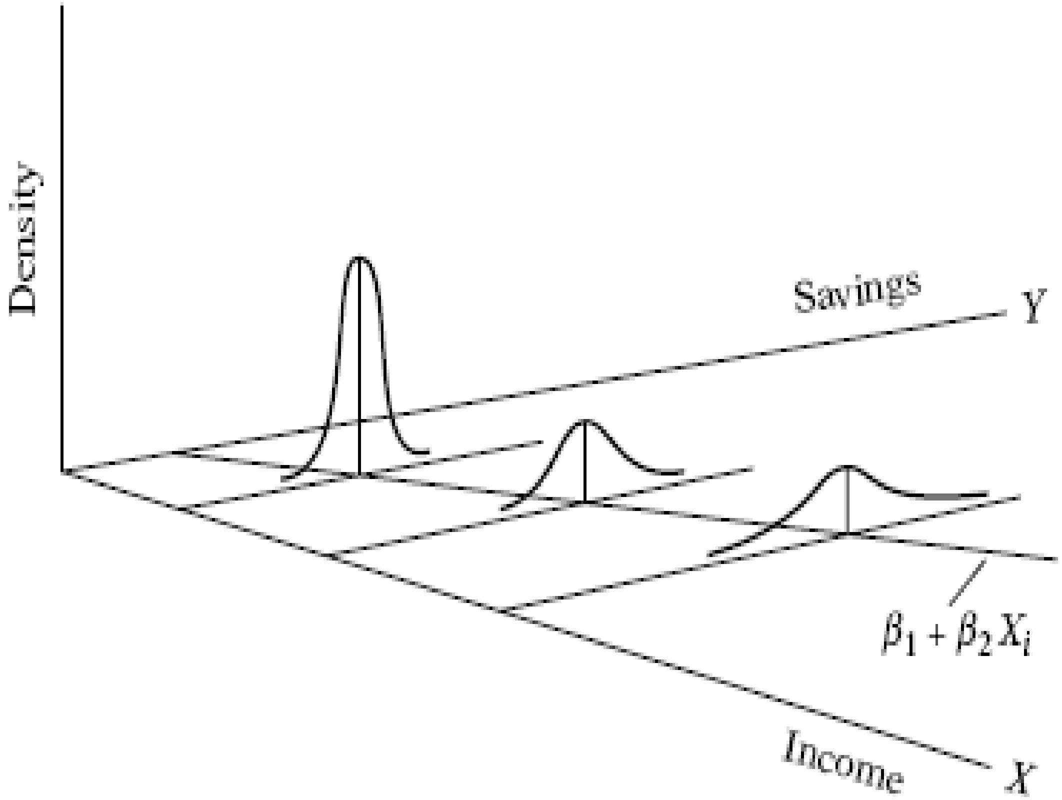
Phương sai thay đổi là gì (tt)



Phương sai bằng nhau



Phương sai thay đổi



Nguyên nhân của phương sai thay đổi là gì?

- Mô hình học tập sai lầm.
- Kỹ thuật thu thập số liệu.
- Do các yếu tố tách biệt (outliers).
- Một số biến X quan trọng bị loại bỏ trong mô hình.
- Phương sai thay đổi thường xuất hiện trong các số liệu chéo.

Ước lượng của OLS khi có phương sai thay đổi

- $Y_i = \beta_1 + \beta_2 X_i + u_i$

- $\hat{\beta}_2$ có phải là ước lượng tuyến tính không thiên lệch tốt nhất?
- $\hat{\beta}_2$ là ước lượng tuyến tính không chệch nhưng nó không phải tốt nhất (nghĩa là phương sai của nó không phải là nhỏ nhất).

Ước lượng của OLS khi có phương sai thay đổi

- Trong trường hợp phương sai bằng nhau

$$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\Sigma x_i^2}$$

- Trong trường hợp phương sai thay đổi

$$\operatorname{var}(\hat{\beta}_2) = \frac{\Sigma x_i^2 \sigma_i^2}{(\Sigma x_i^2)^2}$$

Hậu quả sử dụng OLS khi có phương sai thay đổi

- Khi có phương sai thay đổi nếu vẫn sử dụng OLS thì những kết luận hay sự suy diễn từ quá trình kiểm định thông thường có thể dẫn đến sự sai lầm.

$$Do \quad \text{var}(\hat{\beta}_2{}^{OLS}) \geq \text{var}(\hat{\beta}_2{}^{GLS*})$$

$$t_{\hat{\beta}_2}^{OLS} = \frac{\hat{\beta}_2^{OLS}}{se(\hat{\beta}_2^{OLS})} < t_{\hat{\beta}_2}^{GLS} = \frac{\hat{\beta}_2^{GLS*}}{se(\hat{\beta}_2^{GLS*})}$$

Phương pháp GLS (Generalized Least Square)

Giả sử chúng ta có PRF

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$Y_i = \beta_1 X_{0i} + \beta_2 X_i + u_i$$

$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{X_{0i}}{\sigma_i} \right) + \beta_2 \left(\frac{X_i}{\sigma_i} \right) + \left(\frac{u_i}{\sigma_i} \right)$$

$$Y_i^* = \beta_1^* X_{0i}^* + \beta_2^* X_i^* + u_i^*$$

Phương pháp GLS

$$\operatorname{var}(u_i^*) = E(u_i^*)^2 = E\left(\frac{u_i}{\sigma_i}\right)^2$$
$$= \frac{1}{\sigma_i^2}E(u_i^2)$$
$$= \frac{1}{\sigma_i^2}(\sigma_i^2)$$
$$= 1$$

$$\operatorname{var}(\hat{\beta}_2^{**}) = \frac{\sum w_i}{\left(\sum w_i\right)\left(\sum w_i X_i^2\right) - \left(\sum w_i X_i\right)^2}$$

Phương pháp GLS

- $Y_i = \beta_1 + \beta_2 X_i + u_i$

Cho $\beta_1 = 1$, $\beta_2 = 1$, và $u_i \sim N(0, X^{\alpha}_i)$ tính Y_i (PP Monte Carlo)

	Sai số chuẩn của $\hat{\beta}_1$			Sai số chuẩn của $\hat{\beta}_2$		
Giá trị α	OLS	$OLS_{het.}$	GLS	OLS	$OLS_{het.}$	GLS
0,5	0,164	0,134	0,110	0,285	0,277	0,243
1,0	0,142	0,101	0,048	0,246	0,247	0,173
2,0	0,116	0,074	0,0073	0,200	0,220	0,109
3,0	0,100	0,064	0,0013	0,173	0,206	0,056
4,0	0,089	0,059	0,0003	0,154	0,195	0,017

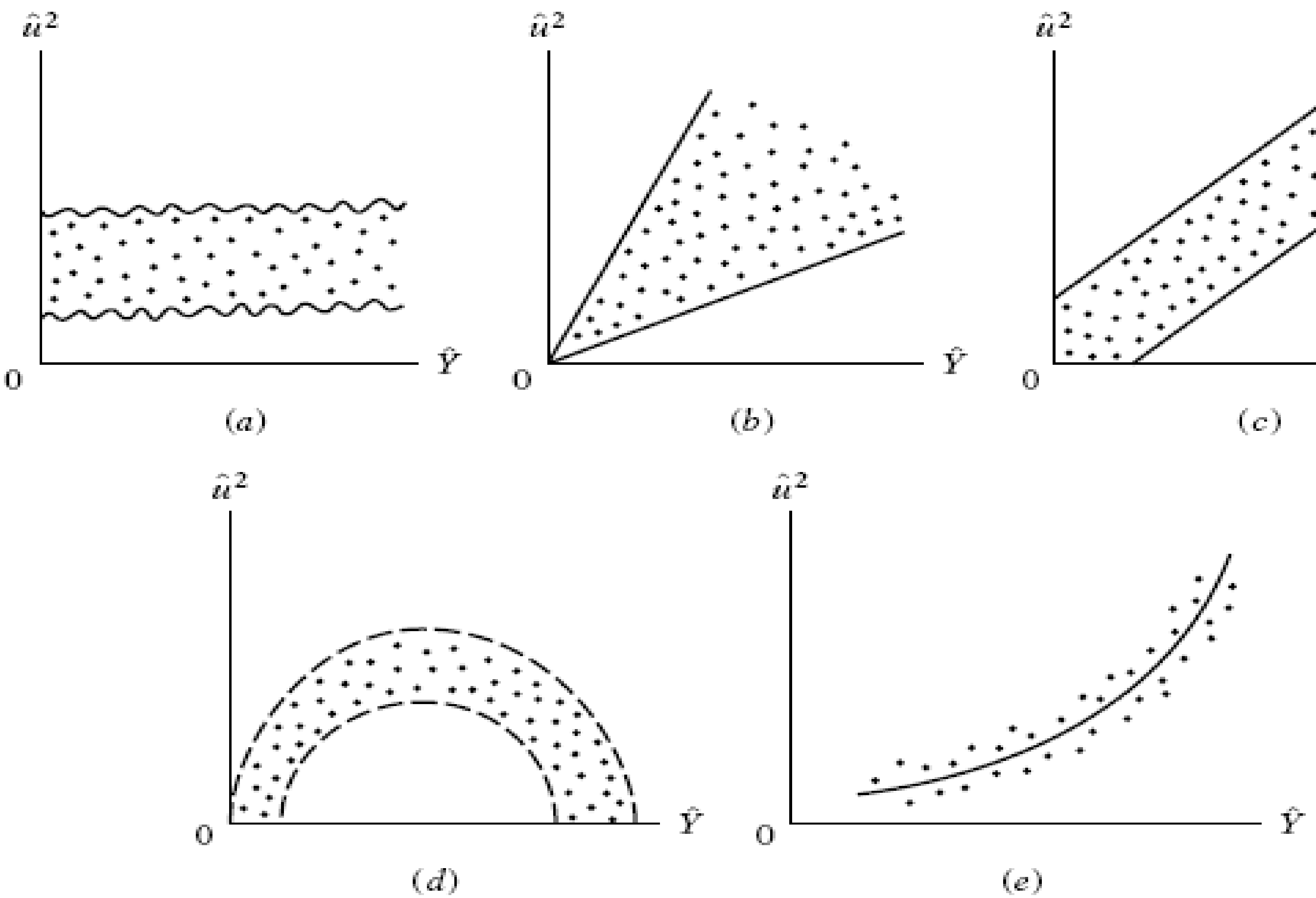
Chú ý: $OLS_{het.}$ có nghĩa là OLS có tính tới phương sai thay đổi.

Làm thế nào phát hiện phương sai thay đổi?

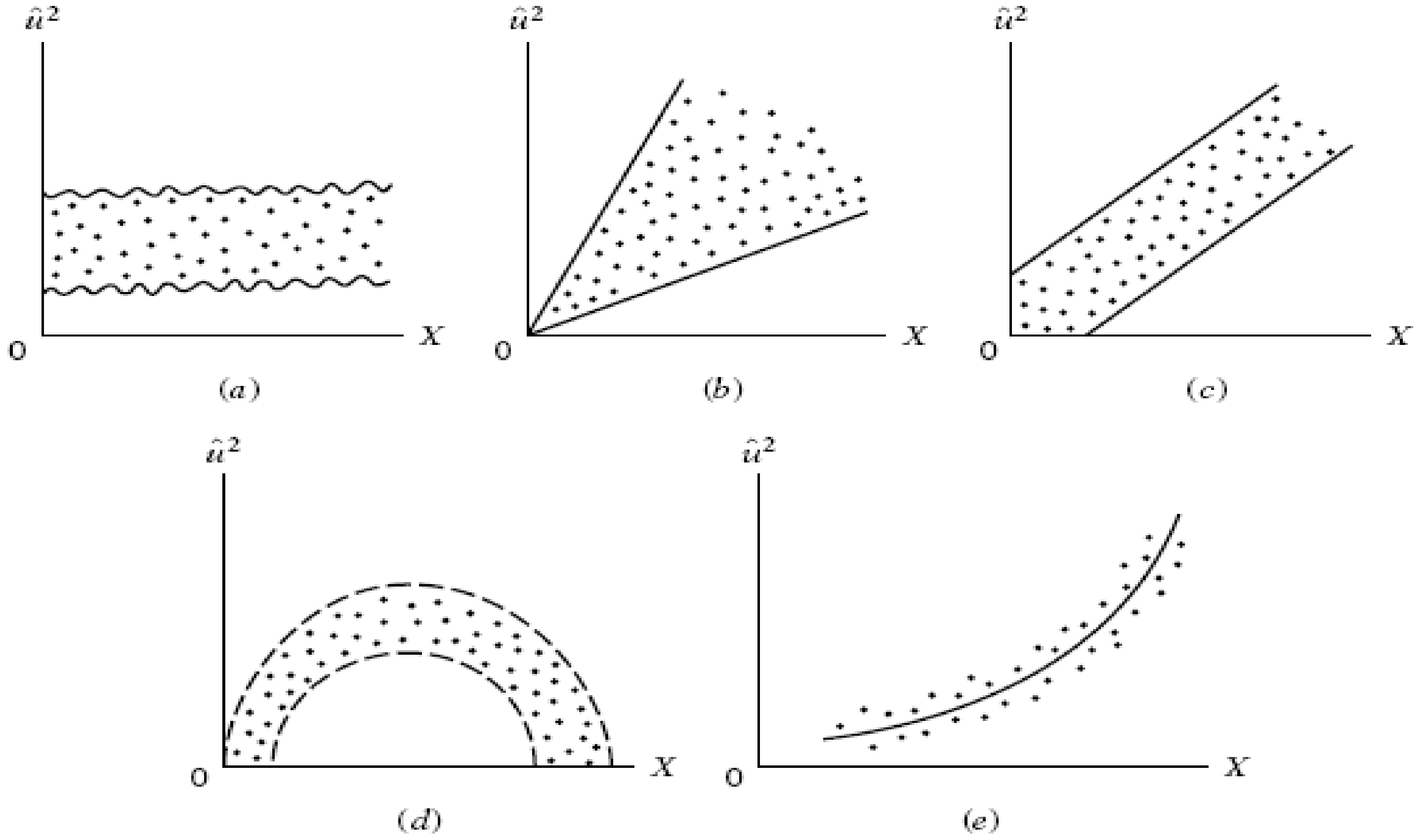
- Phương pháp đồ thị

- Xem mối quan hệ giữa σ^2_i và Y_i
- Do không quan sát được σ^2_i nên chúng ta có thể nghiên cứu $\hat{u}_i^2$
- $\hat{u}_i^2$ được tính từ việc hồi quy mô hình ước lượng với giả thiết không có phương sai thay đổi và quan sát xem chúng có mẫu hình hệ thống không?

Các mẫu hình giả thiết của phần dư bình phương ước lượng



Các mẫu hình giả thiết của phần dư bình phương ước lượng



Kiểm định phương sai thay đổi

- Phương pháp chính thức

1) <u>Kiểm định Breusch-Pagan-Godfrey</u>

- Giả thiết

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \ldots + \alpha_m Z_{mi} + v_i$$

- H_0: $\alpha_2 = \alpha_3 = \ldots = \alpha_m = 0$ ➔ $\sigma^2_i = \alpha_1$ (phương sai bằng nhau)

Kiểm định phương sai thay đổi (tt)

2) <u>Kiểm định Glejser</u>

- Giả thiết:

$$|\hat{u}_i| = \alpha_1 + \alpha_2 X_i + v_i$$

$$|\hat{u}_i| = \alpha_1 + \alpha_2 \sqrt{X_i} + v_i$$

$$|\hat{u}_i| = \alpha_1 + \alpha_2 \frac{1}{X_i} + v_i$$

$$|\hat{u}_i| = \alpha_1 + \alpha_2 \frac{1}{\sqrt{X_i}} + v_i$$

$$|\hat{u}_i| = \sqrt{\alpha_1 + \alpha_2 X_i} + v_i$$

$$|\hat{u}_i| = \sqrt{\alpha_1 + \alpha_2 X_i^2} + v_i$$

- H_0: $\alpha_2 = 0$ ➔ $\sigma^2_i = \alpha_1$ (phương sai bằng nhau)

3) Kiểm định Harvey-Godfrey

- Giả thiết

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$
$$\ln(\sigma_i^2) = \alpha_1 + \alpha_2 Z_{2i} + ... + \alpha_m Z_{mi} + v_i$$

- H_0: $\alpha_2 = \alpha_3 = \ldots = \alpha_m = 0$ ➔ $\ln(\sigma^2_i) = \alpha_1$ (phương sai bằng nhau)

Kiểm định phương sai thay đổi

4) Kiểm định Park

- Giả thiết:

$$\sigma_i^2 = \sigma^2 X_i^{\alpha} e^{v_i}$$

$$\ln \sigma_i^2 = \ln \sigma^2 + \alpha \ln X_i + v_i$$

$$\ln \hat{u}_i^2 = \alpha_0 + \alpha_1 \ln X_i + v_i$$

- Các bước kiểm định

i) Ước lượng bằng OLS, tính

ii) Thực hiện hồi quy phụ

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 Z_{2i} + ... + \alpha_m Z_{mi} + v_i$$
$$|\hat{u}_i| = \alpha_1 + \alpha_2 Z_{2i} + ... + \alpha_m Z_{mi} + v_i$$
$$\ln(\hat{u}_i^2) = \alpha_1 + \alpha_2 Z_{2i} + ... + \alpha_m Z_{mi} + v_i$$

iii) Với H_0: không có phương sai thay đổi

$LM = nR_{hqp}^{\;2} \sim_{asy} \chi^2_{df}$ với bậc tự do bằng với số biến độc lập

➔ Nếu nR^2 vượt giá trị Chi-bình phương tới hạn với mức ý nghĩa đã chọn (vd là 5%) hay p-value < 5%, ta bác bỏ H_0

Làm thế nào phát hiện phương sai thay đổi (tt)

5) <u>Kiểm định Goldfeld-Quandt</u>

- Giả thiết: H_0: Không có phương sai thay đổi

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$
$$\sigma_i^2 = \sigma^2 X_i^2$$

- Các bước kiểm định

i. Sắp thứ tự từ thấp đến cao các quan sát theo các giá trị X_i

ii. Loại bỏ c quan sát ở giữa, và chia (n-c) quan sát còn lại thành 2 nhóm bằng nhau

iii) Thiết lập hồi quy OLS cho 2 nhóm, tính RSS_1 và RSS_2 tương ứng với bậc tự do

[(n-c)/2] –k   hoặc   (n-c-2k)/2

iv) Tính tỷ lệ

$$\lambda = \frac{RSS_2 / df}{RSS_1 / df}$$

Nếu u_i có phân phối chuẩn và phương sai bằng nhau thì λ tuân theo phân phối F với bậc tự do tử số và mẫu số [(n-c)/2] –k hay (n-c-2k)/2

- ➔ Nếu $\lambda > F_{\text{tới hạn}}$ tại mức ý nghĩa chọn trước thì bác bỏ giả thiết phương sai không thay đổi (tức là, tồn tại khả năng có phương sai thay đổi)
- ➔ Chú ý: Năng lực kiểm định tùy thuộc vào việc chọn c.

  Cỡ mẫu n=30 ➔ c=8 (4 theo Judge et al)

  Cỡ mẫu n=60 ➔ c=16 (10 theo Judge et al)

6) Kiểm định White

- Giả thiết (các biến độc lập đồng thời gây ra phương sai thay đổi)

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$\sigma_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i$$

- H_0: $\alpha_2 = \alpha_3 = \ldots = \alpha_m = 0$ ➔ $\sigma^2_i = \alpha_1$ (phương sai bằng nhau)
- Các bước kiểm định

i. Ước lượng bằng OLS, tính

ii) Thực hiện hồi quy phụ

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{31} + v_i$$

iii) Với H_0 ta có

$$nR_{hqp}^{\;2} \sim_{asy} \lambda_{df}^2$$

với bậc tự do bằng với số biến độc lập (vd: df=5)

➔ Nếu nR^2 vượt giá trị Chi-bình phương tới hạn với mức ý nghĩa đã chọn hoặc p-value < 5%, ta bác bỏ H_0 và kết luận rằng có phương sai thay đổi.

Các biện pháp khắc phục

1) Nếu σ^2_i biết trước:

Mô hình ban đầu

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

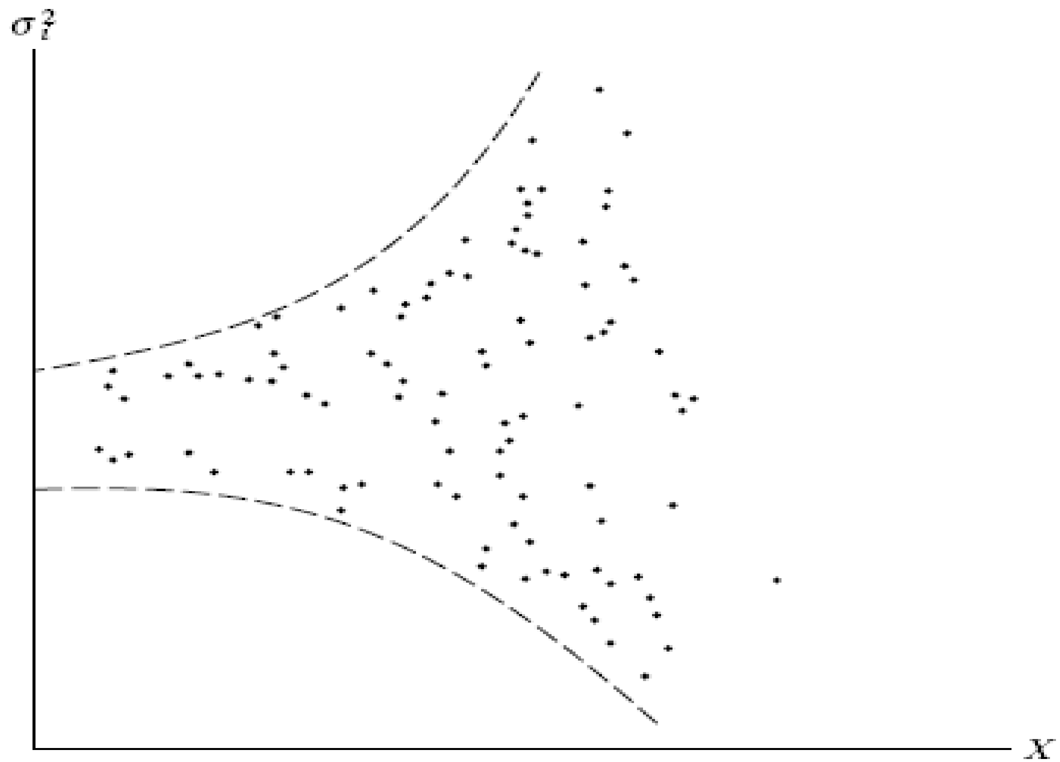
➔ Dùng PP bình phương tối thiểu có trọng số (WLS)

$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{1}{\sigma_i}\right) + \beta_2 \left(\frac{X_i}{\sigma_i}\right) + \left(\frac{u_i}{\sigma_i}\right)$$

$$E\left(\frac{u_i}{\sigma_i}\right)^2 = \frac{1}{\sigma_i^{\,2}} E(u_i^2) = \frac{1}{\sigma_i^{\,2}} \sigma_i^{\,2} = 1$$

2) Nếu σ^2_i chưa biết

Giả thiết 1: $E(u_i^2) = \sigma^2 X_i^2$



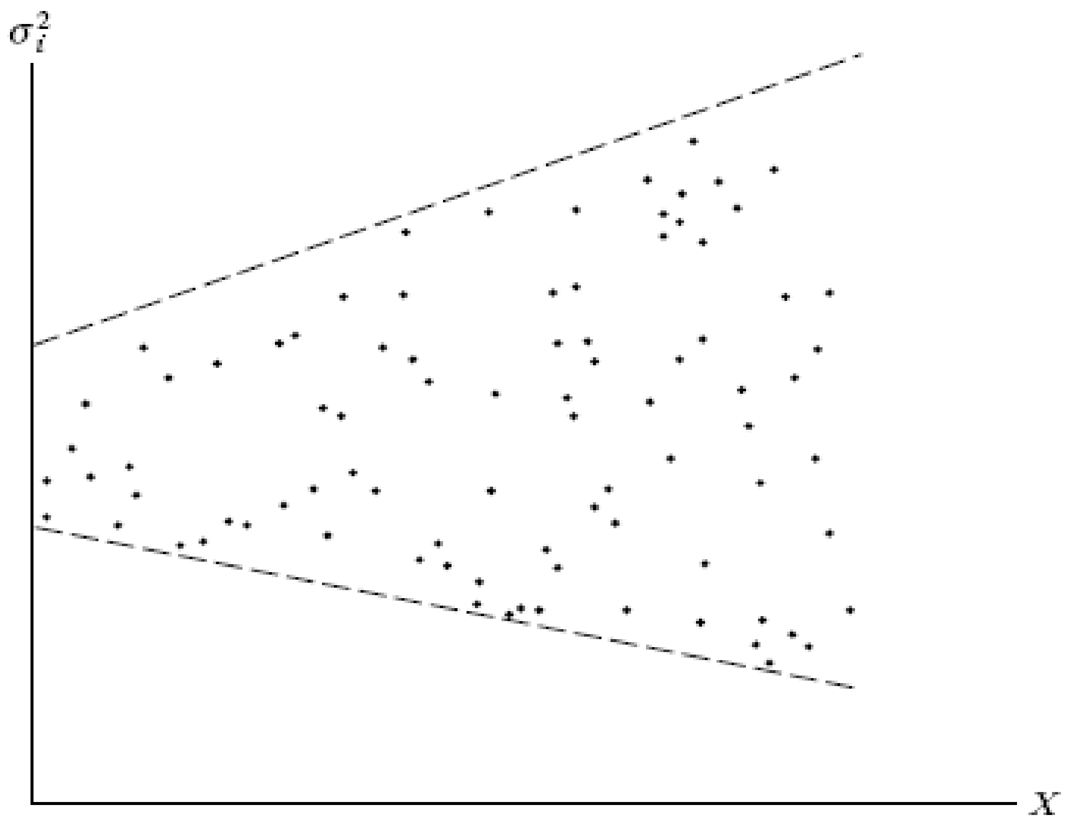
➔Biến đổi mô hình gốc

$$\frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \beta_2 + \frac{u_i}{X_i}$$

$$E(\frac{u_i}{X_i})^2 = \frac{1}{X_i^{\ 2}} E(u_i^2) = \sigma^2$$

Muốn trở lại mô hình ban đầu phải nhân phương trình đầu cho X_i

Giả thiết 2: $E(u_i^2) = \sigma^2 X_i$



→Biến đổi mô hình gốc

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \frac{u_i}{\sqrt{X_i}}$$

$$E\left(\frac{u_i}{\sqrt{X_i}}\right)^2 = \frac{1}{X_i}E(u_i^2) = \sigma^2$$

Muốn trở lại mô hình ban đầu phải nhân phương trình đầu cho

$(X_i)^{1/2}$

- Giả thiết 3: $E(u_i^2) = \sigma^2[E(Y_i)]^2$

➔Biến đổi mô hình gốc

$$\frac{Y_i}{E(Y_i)} = \frac{\beta_1}{E(Y_i)} + \beta_2 \frac{X_i}{E(Y_i)} + \frac{u_i}{E(Y_i)}$$

$$E(Y_i) \approx \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$$E\left(\frac{u_i}{E(Y_i)}\right)^2 = \frac{1}{E(Y_i)^2} E(u_i^2) = \sigma^2$$

Muốn trở lại mô hình ban đầu phải nhân phương trình đầu cho $E(Y_i)$

- Giả thiết 4:

→ Sử dụng

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i$$

thay vì

$$Y_i = \beta_1 + \beta_2 \ln X_i + u_i$$

- Các biện pháp khắc phục HET theo các kiểm định ở trên (xem Ramanathan, Chương 8, trang 360-362).