# Hồi quy với Biến Định tính
# (Regression with Qualitative Variables)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

# Biến định tính là gì

- Còn được gọi là biến giả (dummy variable)
- Là biến mô tả trạng thái (nam/nữ, đi làm/đi học, làm nông/công chức)
- Có thể là biến nhị phân (có/không) hoặc biến nhóm (categorical variable - có nhiều hơn 2 trạng thái giá trị, ví dụ phương tiện đi lại là ô tô/xe máy/xe đạp/đi bộ)
- Đa số trường hợp các biến định tính không thể xếp được thứ bậc (ví dụ làm việc trong khu vực nhà nước/tư nhân/nước ngoài).
- Một số trường hợp biến định tính có thể xếp được thứ bậc, ví dụ bằng cấp cao nhất có được là gì, từ không có bằng cấp, bằng tiểu học, THCS,THPT, cao đẳng, đại học, thạc sỹ, tiến sỹ.

- Không nhầm lẫn với biến số đếm rời rạc, ví dụ biến số con cái trong gia đình không phải là biến định tính.
- Thống kê mô tả biến định tính khác với biến định lượng.
  - Cần xác định nhóm tham chiếu (baseline/reference group) và nhóm được tham chiếu. Ví dụ với biến giới tính thì có thể đặt nhóm tham chiếu là nữ và nhóm được tham chiếu là nam.
  - Giá trị trung bình diễn giải xác suất xảy ra một sự kiện.
  - Giá trị lớn nhất và nhỏ nhất không có ý nghĩa kinh tế.
  - Sai số chuẩn liên quan đến xác suất quan sát được sự kiện.
  - Hệ số tương quan mẫu (correlation coefficient) không có ý nghĩa.
  - Thường dùng biến định tính để phân tách và so sánh giữa các nhóm, ví dụ nhóm nam và nữ.

# Sử lý biến định tính

Sử dụng lại bộ dữ liệu VHLSS 2010.

- Cần hiểu cách mã hóa biến trong bảng dữ liệu.
- Có thể gộp biến nhóm thành biến nhị phân.
- Có thể tách biến nhóm thành nhiều biến nhị phân.
- Bẫy biến giả (dummy trap): Một biến định tính có $n$ giá trị thì có thể tách ra tối đa là $n-1$ biến giả. Nếu tách làm $n$ biến giả đưa vào mô hình sẽ có hiện tượng đa cộng tuyến hoàn hảo.

# Hồi quy với biến định tính

Ước lượng mô hình tỷ suất thu nhấp của đi học với các biến định tính là có gia đình, học trường công, làm nhà nước, làm nước ngoài, là công chức:

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

# Giải thích ý nghĩa của biến định tính

```
. reg lincome yoeduc yoexper yoexpersq married publicSchool public foreign official
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | 1753.70541 | 8 | 219.213176 | Number of obs | = 7,552 |
| Residual | 4040.86526 | 7,543 | .535710627 | F(8, 7543) | = 409.20 |
| | | | | Prob > F | = 0.0000 |
| | | | | R-squared | = 0.3026 |
| | | | | Adj R-squared | = 0.3019 |
| Total | 5794.57067 | 7,551 | .767391162 | Root MSE | = .73192 |

| lincome | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| yoeduc | .0926075 | .0027428 | 33.76 | 0.000 | .0872309 | .0979841 |
| yoexper | .061687 | .0025081 | 24.60 | 0.000 | .0567705 | .0666035 |
| yoexpersq | -.0012002 | .0000488 | -24.58 | 0.000 | -.0012959 | -.0011044 |
| married | .0352395 | .0221221 | 1.59 | 0.111 | -.0081259 | .078605 |
| publicSchool | -.1145887 | .0423549 | -2.71 | 0.007 | -.1976161 | -.0315613 |
| public | -.1042541 | .0329488 | -3.16 | 0.002 | -.1688429 | -.0396652 |
| foreign | .4499482 | .0363715 | 12.37 | 0.000 | .37865 | .5212464 |
| official | .2705426 | .0359373 | 7.53 | 0.000 | .2000956 | .3409897 |
| _cons | 8.493551 | .0474837 | 178.87 | 0.000 | 8.40047 | 8.586633 |

.

# Diễn giải ý nghĩa của tham số ước lượng đối với biến định tính

- Nếu biến phụ thuộc là **thu nhập** thì tham số ước lượng là tác động tăng thêm của nhóm được tham chiếu so với nhóm tham chiếu.

- Nếu biến phụ thuộc là **log của thu nhập** thì diễn giải tham số ước lượng tùy thuộc vào biến giải thích là biến liên tục hay biến rời rạc.
  - Với **biến liên tục**, ví dụ số năm đi học *yoeduc*, hệ số ước lượng là % tăng thêm của thu nhập. Ví dụ 1 năm đi học làm tăng thu nhập 9.26%.

- Với **biến rời rạc**, ví dụ các biến định tính, hoặc nếu có biến số con trong gia đình, thì:
  - Nếu $\beta$ nhỏ, $\beta$ có thể coi là phần trăm tăng thêm của biến phụ thuộc.
  - Công thức tính chính xác đối với tác động của biến rời rạc lên biến phụ thuộc **log(Y)** là:

$$\frac{Y_1 - Y_0}{Y_0} = e^{\beta} - 1$$

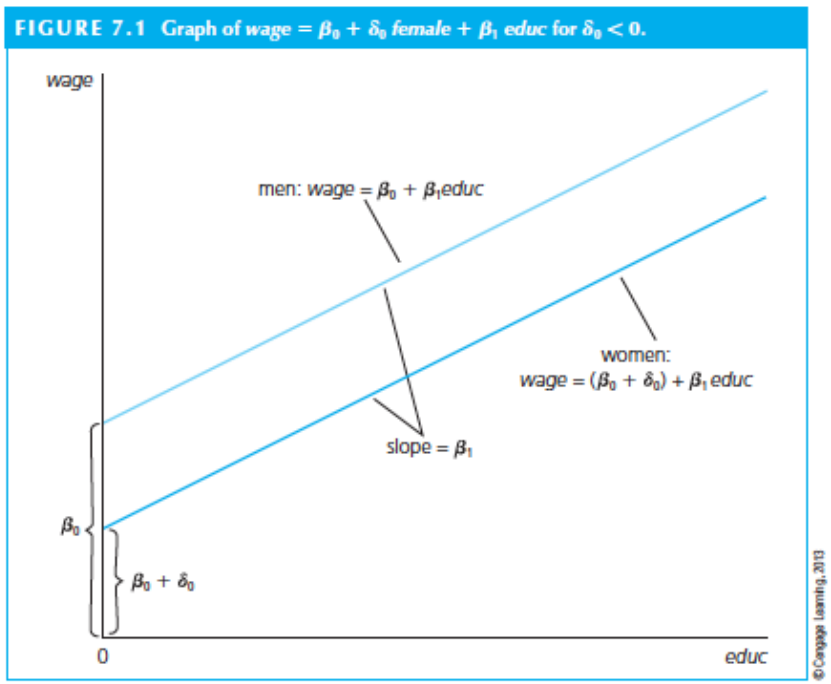
- Trong ví dụ trên:
  - Làm việc trong khu vực nước ngoài thu nhập cao hơn khu vực tư là: $2.718^{.45} - 1 = .5682$ hay 56.82% (chứ không phải là 45%).
  - Làm việc trong khu vực nhà nước thu nhập thấp hơn khu vực tư là: $2.718^{-.1043} - 1 = -.099$ hay 9.9%.
  - Nếu coi *yoeduc* là biến rời rạc thì với mỗi năm học tăng thêm thu nhập là $2.718^{.0926} - 1 = .097$ hay 9.7%.

# Tung độ gốc và hệ số góc trong mô hình hồi quy

Với biến giới tính *male* trong mô hình:

$$log(income) = \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + ... + \sigma_0 \mathbf{male} + u$$

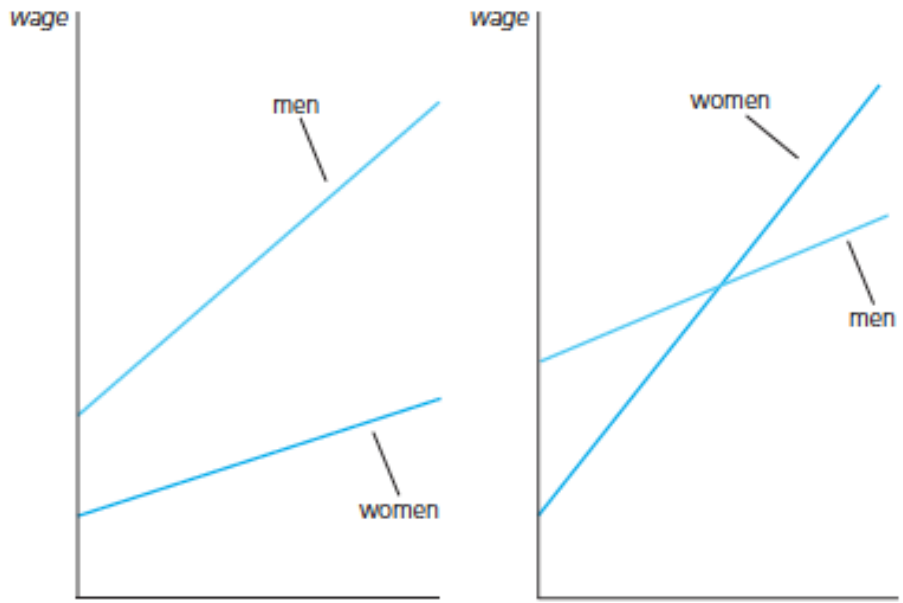
- Tung độ gốc là $\beta_0$ với nhóm nữ, và $\beta_0 + \sigma_0$ với nhóm nam
- Hệ số góc là $\beta_1$ giống nhau với cả hai nhóm (đường hồi quy song song)
- Nếu $\sigma_0 = 0$ thì hai đường hồi quy trùng nhau



# Tung độ gốc và hệ số góc trong mô hình hồi quy với biến tương tác

$$log(income) = \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + ... + \sigma_0 \mathbf{male} + \sigma_1 \mathbf{male} * \mathbf{yoeduc} + u$$

- Tung độ gốc là $\beta_0$ với nhóm nữ, và $\beta_0 + \sigma_0$ với nhóm nam
- Hệ số góc là $\beta_1$ với nhóm nữ, và $\beta_1 + \sigma_1$ với nhóm nam.
- Hai đường hồi quy chỉ trùng nhau khi $\sigma_0$ và $\sigma_1$ đồng thời bằng 0.



# Kiểm định khác biệt theo nhóm

- Tung độ gốc khác nhau $\Rightarrow$ t-test nếu $\sigma_0 = 0$
- Tung độ gốc và hệ số góc khác nhau $\Rightarrow$ F-test nếu $\sigma_0 = \sigma_1 = 0$
- Tất cả các tham số của hai nhóm khác nhau $\Rightarrow$ Chow test

# Ôn tập các loại kiểm định

- Kiểm định đơn: $H_0 : \sigma_0 = 0$

$$t_{\hat{\sigma_0}} \sim t_{n-k-1}$$

- Kiểm định bội: $H_0 : \sigma_0 = \sigma_1 = 0$

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k-1)} \sim F_{q,n-k-1}$$

- Kiểm định khác biệt nhóm (tất cả các tham số):

$$H_0 : \sigma_0 = \sigma_1 = ... = \sigma_k = 0$$

$$F = \frac{[SSR_p - (SSR_1 + SSR_2)]/(k+1)}{[SSR_1 + SSR_2]/(n-2(k+1))} \sim F_{k+1,n-2(k+1)}$$

# Hồi quy với Phương sai thay đổi (Heteroskedasticity)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

# Các giả định của mô hình MLR

1. Tuyến tính theo tham số.
2. Lấy mẫu ngẫu nhiên.
3. Không có cộng tuyến hoàn hảo.
4. $E(u|X) = 0 \Rightarrow$ Ước lượng OLS là không chệch.
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 $\sigma^2$ (independent, identically distributed - *iid*):

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

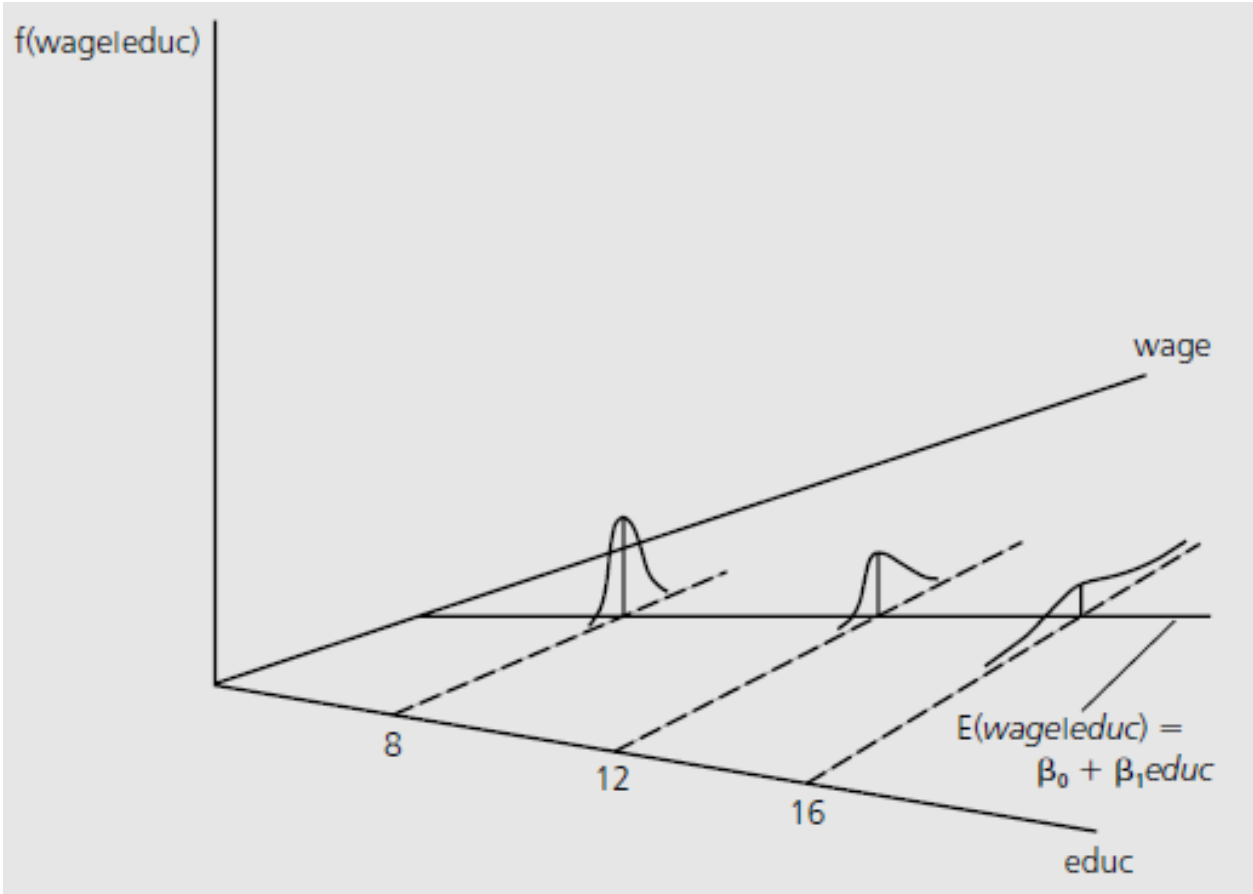
$\Rightarrow$ Mô hình hồi quy tuyến tính cổ điển (CLRM):

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

# Phương sai của sai số thay đổi (heteroskedasticity)

- Vi phạm điều kiện *iid*: $Var(u|X) \neq \sigma^2$
- Với giả định $E(u|X) = 0$ và $cov(u, X) = 0$ thỏa, ước lượng bằng OLS vẫn **không chệch và nhất quán**, tuy nhiên không còn là hiệu quả nhất do sai số của $\hat{\beta}$ không còn là nhỏ nhất.
- Các kiểm định (t-test, F-test) không có hiệu lực do ước lượng 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 số năm học nhiều thì mức độ dao động của thu nhập càng lớn.
- Do tương quan chuỗi hoặc tương quan không gian.

# 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

Giả sử chúng ta ước lượng mô hình:

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

Chúng ta muốn kiểm định nếu phương sai của sai số không đổi.

Do $E(u|X) = 0$ nên:

$$Var(u|X) = E(u^2) - [E(u)]^2 = E(u^2)$$

Do đó giả thuyết $H_0$ được viết như sau:

$$H_0 : E(u^2) = \sigma^2$$

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

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

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

4. Kiểm định giả thuyết $H_0 : \delta_1,..., \delta_k$ đồng thời bằng 0 trong mô hình (2) bằng F-test. Trị kiểm định được tính từ $R_a^2$ của mô hình 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 $\hat{u}^2$ phụ thuộc vào một trong các biến giải thích, hay phương sai của sai số của mô hình không phải là hằng số.

# Ví dụ kiểm định BP

- Ước lượng lại mô hình tỷ suất thu nhập từ bộ dữ liệu VHLSS 2010.
- Kiểm định phương sai thay đổi thủ công.
- Thực hiện tự động bằng Stata.

# Kiểm định White đối với hiện tượng phương sai thay đổi trong trường hợp tổng quát

3. Tương tự như kiểm định Breusch-Pagan ở bước 1-2, nhưng ở bước 3 giả định cấu trúc hàm của sai số linh hoạt hơn bằng cách thêm bình phương và tương tác giữa các biến giải thích trong hồi quy phụ:

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

4. Kiểm định bằng F-test nếu tất cả các tham số $\delta$ (loại trừ tung độ gốc $\delta_0$) trong hồi quy phụ bằng 0.
5. Bác bỏ giả thuyết $H_0$ có nghĩa là mô hình có hiện tượng phương sai thay đổi.

# Kiểm định White đối với hiện tượng phương sai thay đổi - thực hiện đơn giản

Kiểm định White trong trường hợp tổng quát sẽ làm giảm số bậc tự do trong mô hình, ví dụ mô hình có 3 biến tự do sẽ có tổng cộng là 9 ràng buộc. Một hình thức khác có thể kiểm định bằng cách tính $\hat{u}$ và $\hat{y}$ từ bước 1 và 2 tương tự như trong kiểm định BP.

3. Hồi quy $\hat{u}^2$ lên biến $\hat{y}$ và $\hat{y}^2$ trong mô hình hồi quy 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. Bác bỏ giả thuyết $H_0$ có nghĩa là mô hình có hiện tượng phương sai thay đổi.

# Ví dụ kiểm định White

- Ước lượng lại mô hình tỷ suất thu nhập từ bộ dữ liệu VHLSS 2010.
- Kiểm định phương sai thay đổi thủ công thông qua F statistic.
- Thực hiện tự động bằng Stata.

# Ước lượng mô hình khi có hiện tượng phương sai thay đổi

1. Sử dụng phương pháp White để điều chỉnh sai số của tham số:

$$\widehat{Var}(\hat{\beta}) = \frac{\sum_{i=1}^{n} \hat{r_{ij}}^2 \hat{u_i^2}}{SSR_j^2}$$

Sai số điều chỉnh cách thức này còn được gọi là:

- Heteroskedasticity-robust standard errors.
- White-Huber standard errors.
- Robust standard errors.

# So sánh mô hình tỷ suất thu nhập có và không điều chỉnh phương sai thay đổi

Regression Results

| | Homo.<br>b/se | Robust<br>b/se |
|---|---|---|
| yoeduc | 0.0926*** | 0.0926*** |
| | (0.0027) | (0.0031) |
| yoexper | 0.0617*** | 0.0617*** |
| | (0.0025) | (0.0032) |
| yoexpersq | -0.0012*** | -0.0012*** |
| | (0.0000) | (0.0001) |
| married | 0.0352 | 0.0352 |
| | (0.0221) | (0.0217) |
| publicSchool | -0.1146** | -0.1146* |
| | (0.0424) | (0.0465) |
| public | -0.1043** | -0.1043* |
| | (0.0329) | (0.0423) |
| foreign | 0.4499*** | 0.4499*** |
| | (0.0364) | (0.0328) |
| official | 0.2705*** | 0.2705*** |
| | (0.0359) | (0.0430) |
| Constant | 8.4936*** | 8.4936*** |
| | (0.0475) | (0.0539) |
| Obs | 7552.0000 | 7552.0000 |
| R2 | 0.3026 | 0.3026 |
| R2-adj | 0.3019 | 0.3019 |
| df(r) | 7543.0000 | 7543.0000 |
| SSR | 4040.8653 | 4040.8653 |

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

# Kiểm định giả thuyết bội khi có hiện tượng phương sai thay đổi

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

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

- Sử dụng phương sai có điều chỉnh theo phương pháp White.
- Sử dụng trị kiểm định heteroskedasticity-robust F.
- Kết quả sẽ khác so với khi giả định phương sai không đổi.

# Các phương pháp khác ước lượng mô hình khi có hiện tượng phương sai thay đổi

2. Khi biết cấu trúc của phương sai (WLS).
3. Không biết cấu trúc của phương sai (FGLS).

# Phương sai biết cấu trúc hàm - Phương pháp WLS

WLS - Weighted Least Square: Hồi quy bình phương tối thiểu có quyền số.

- Giả định phương sai của sai số là một hàm số của $x$:

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

- Có thể thực hiện chuyển đổi dữ liệu trước khi ước lượng:

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

- Ước lượng mô hình hồi quy dựa trên các biến số đã chuyển đổi sẽ có đặc tính BLUE.

# Phương sai không biết cấu trúc hàm - FGLS

FGLS: Feasible Generalized Least Square - Bình phương tối thiểu tổng quát khả thi.

- Thông thường giả định phương sai của sai số là hàm mũ của $x$, nhưng không biết cấu trúc hàm:

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

- Phương pháp FGLS sẽ ước lượng hàm của $Var(u|x)$ để làm quyền 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 quyền số là $1/\widehat{h(x)}$.

# Ví dụ hồi quy WLS và FGLS

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

- Ước lượng WLS với giả định phương sai của sai số tỷ lệ thuận với tổng thu nhập:

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

- Ước lượng FGLS cho trường hợp phương sai thay đổi và không biết cấu trúc hàm.
- So sánh các ước lượng với giả định Homoskedasticity, Heteroskedasticity-robust standard errors, WLS, và FGLS.

Regression Results

| | Homo<br>b/se | Robust<br>b/se | WLS<br>b/se | FGLS<br>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