Hồi quy Tuyến tính Đơn biến (Simple Linear Regression - SLR)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 30 tháng 11 năm 2018

Giới thiệu mô hình SLR

Chúng ta có 2 biến số x và y và muốn tìm hiểu x ảnh hưởng như thế nào đến y. Mô hình đơn giản nhất được viết dưới dạng một hàm số tuyến tính của y theo x:

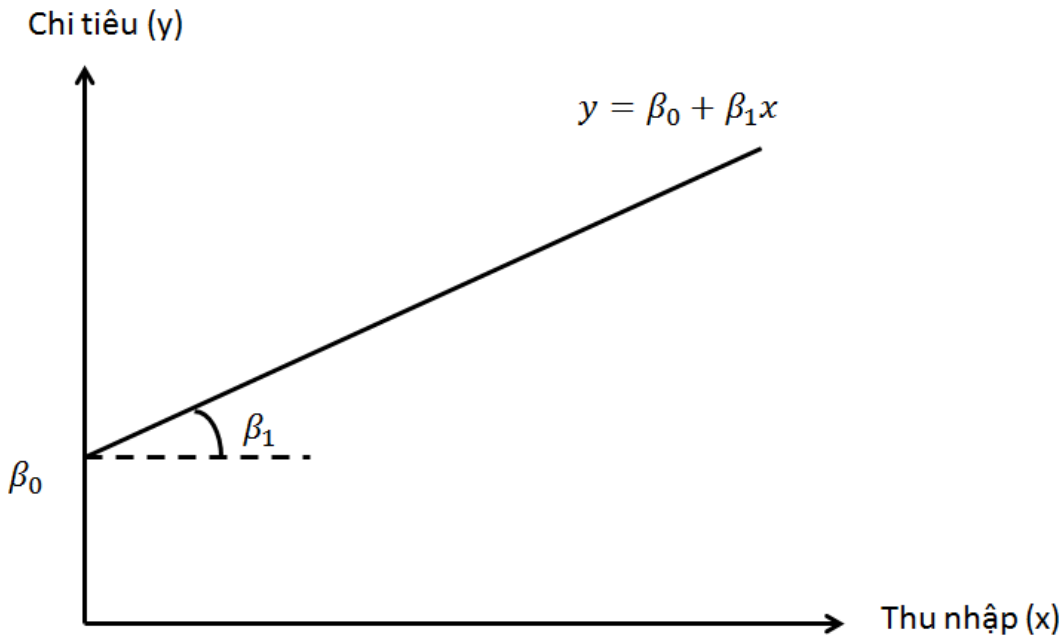
$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- i đại diện cho quan sát thứ i trong tổng thể gồm có n quan sát.
- y gọi là biến phụ thuộc/biến được giải thích/biến phản ứng/biến được dự báo
- x là biến độc lập/biến giải thích/biến kiểm soát/biến dự báo
- u là sai số (số hạng nhiễu), không quan sát được, bao gồm tất cả những yếu tố khác ảnh hưởng đến y nhưng không nằm trong x.
- β_0 và β_1 là các tham số trong mô hình – cần phải ước lượng.

Diễn giải mô hình

- β_0 là tung độ gốc
- β_1 là độ dốc của đường hồi quy
- Nếu các yếu tố khác (u) giữ nguyên không đổi, x tác động tuyến tính tới y thông qua phương trình:

$$\Delta y = \beta_1 \Delta x$$



Hàm hồi quy tổng thể và Hàm hồi quy mẫu

- Với giả định sai số bình quân $E(u)$ trong tổng thể bằng không, $E(u) = 0$, hàm hồi quy tổng thể (Population Regression Function - PFR) được viết dưới dạng:

$$y = \beta_0 + \beta_1 x$$

- Chúng ta không bao giờ biết chính xác giá trị của β_0 và β_1 từ tổng thể.
- Các phương pháp hồi quy sẽ ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ từ dữ liệu, từ đó chúng ta có mô hình hồi quy mẫu (Sample Regression Function - SRF):

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Ví dụ: Ước lượng tác động của tỷ suất sinh lợi của doanh nghiệp lên mức lương của CEO

- Xem bộ dữ liệu CEOSAL1.dta.
- Giả sử tiền lương CEO được quyết định do kết quả hoạt động của doanh nghiệp (đại diện bởi tỷ suất sinh lợi trên vốn, *roe*) mang lại:

$$salary = \beta_0 + \beta_1 roe + u$$

- Kỳ vọng gì về giá trị của β_0 và β_1?
- Tìm hiểu bộ dữ liệu:

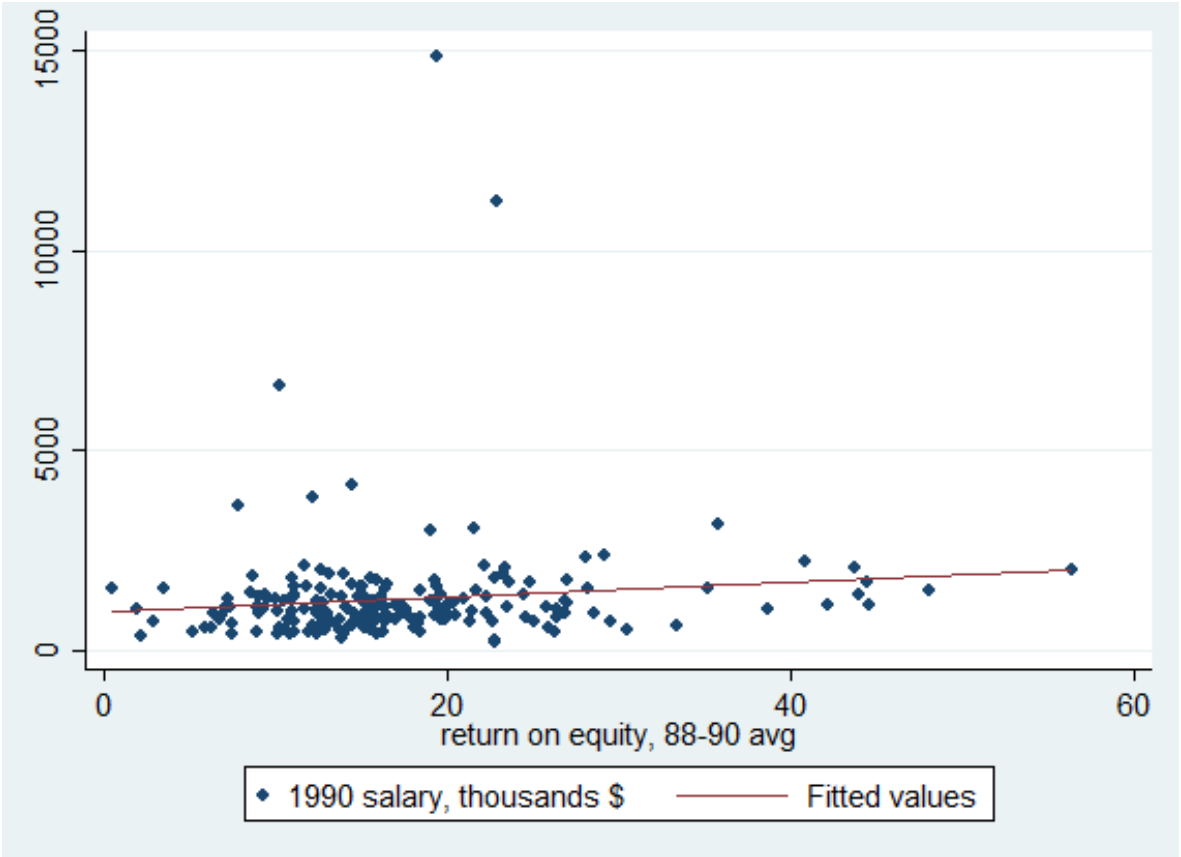
```
. sum salary roe
```

Variable	Obs	Mean	Std. Dev.	Min	Max
salary	209	1281.12	1372.345	223	14822
roe	209	17.18421	8.518509	.5	56.3

```
.
. corr salary roe
(obs=209)
```

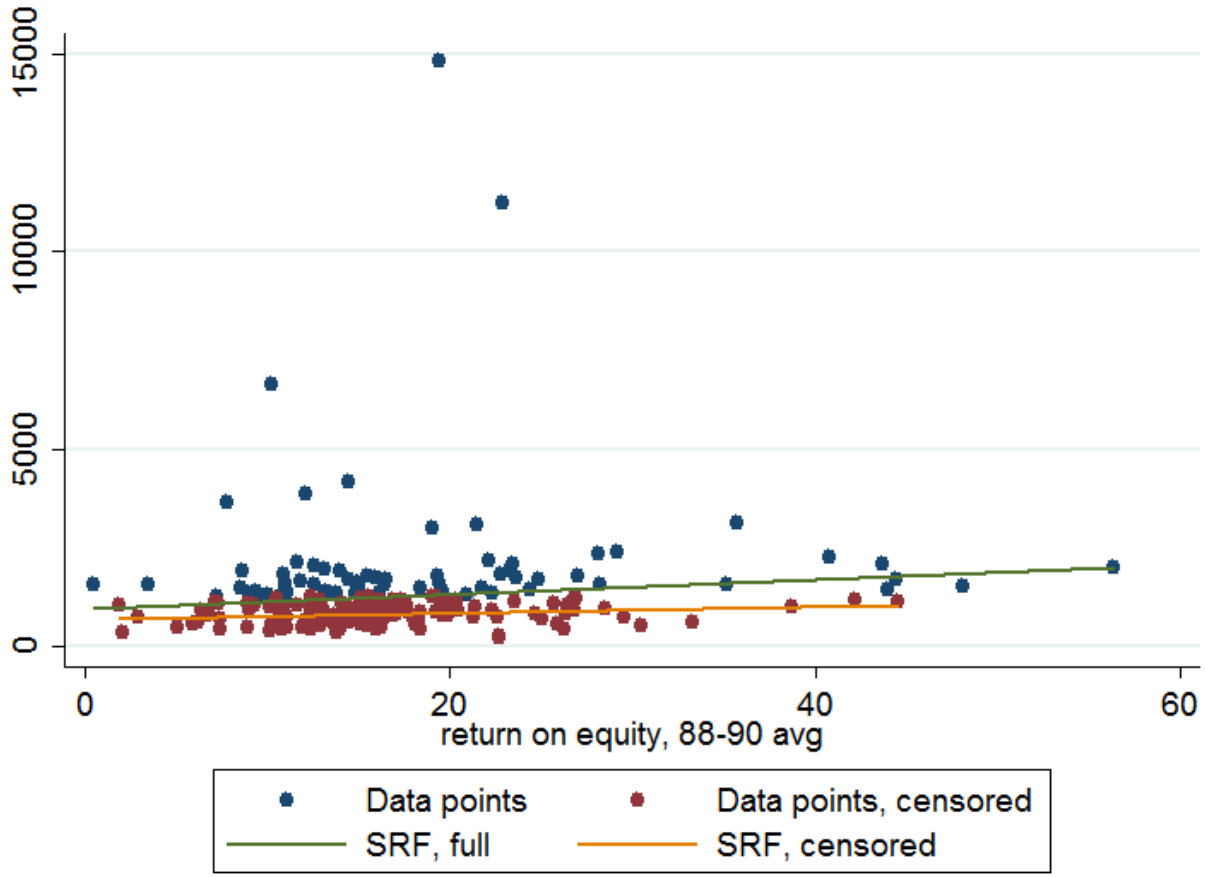
	salary	roe
salary	1.0000	
roe	0.1148	1.0000

Hình dạng đường hồi quy



So sánh đường hồi quy mẫu với tổng thể

Giả sử chúng ta chỉ có dữ liệu của những CEO có mức lương từ trung bình trở xuống (salary < 1.281 triệu đô la/năm). Ước lượng tương ứng với đồ thị màu cam.



$\Rightarrow$ Mục tiêu là ước lượng được $\hat{\beta}_0$ và $\hat{\beta}_1$ của SRF càng gần với β_0 và β_1 của PRF càng tốt.

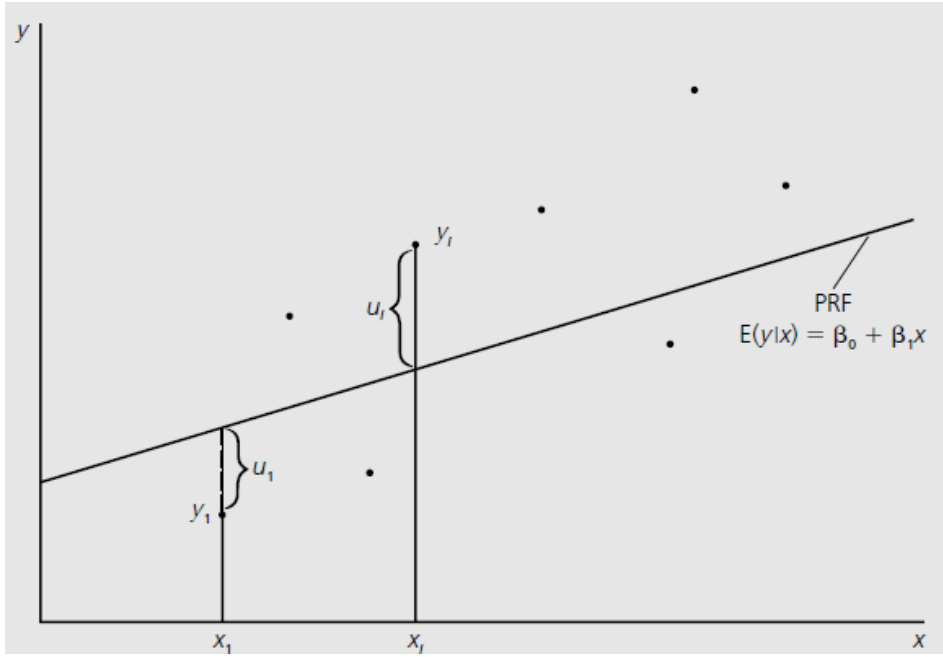
Phương pháp bình phương tối thiểu thông thường (Ordinary Least Square - OLS)

- Ký hiệu i đại diện cho quan sát thứ i của dữ liệu gồm n quan sát. Từ phương trình hồi quy ta có thể viết lại là:

$$u_i = y_i - \beta_0 - \beta_1 x_i$$

- Cơ chế của phương pháp OLS là tìm $\hat{\beta}_0$ và $\hat{\beta}_1$ để tối thiểu hóa tổng bình phương của u_i.

$$U = min \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



- Dựa vào hình vẽ: Bản chất của OLS là tìm phương trình đường thẳng đi qua phân phối điểm của dữ liệu sao cho tổng bình phương khoảng cách từ các điểm dữ liệu đến đường thẳng là tối thiểu. Tại sao phải dùng bình phương của khoảng cách?
- Các phương pháp khác có thể sử dụng giá trị tuyệt đối của khoảng cách.

Cơ chế của phương pháp OLS

Để tìm giá trị $\hat{\beta}_0$ và $\hat{\beta}_1$ để tối thiểu hóa tổng bình phương của u_i, ta sử dụng điều kiện bậc nhất là đạo hàm của hàm mục tiêu bằng không tại các giá trị cực trị:

$$\frac{\partial U}{\partial \beta_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{1}$$

và

$$\frac{\partial U}{\partial \beta_1} = -2 \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{2}$$

Điều kiện của ước lượng OLS

- Hai điều kiện bậc nhất (1) và (2) tương ứng với:

$$\mathbf{E}(\mathbf{u}) = 0$$
$$\mathbf{E}(\mathbf{xu}) = 0$$

  Diễn giải: trung bình của sai số u bằng không và sai số u không tương quan với biến giải thích x.

- Với các điều kiện trên thì ước lượng OLS là không chệch (unbiased), $\mathbf{E}(\hat{\beta}) = \beta$, và nhất quán (consistent), $\mathbf{plim}(\hat{\beta}) \to \beta$ khi cỡ mẫu tiến đến vô cùng.

Giải các điều kiện bậc nhất ta thu được giá trị của $\hat{\beta}_0$ và $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{Cov(X, Y)}{Var(X)}$$

(Lưu ý: ký hiệu X mô tả vector, x là từng giá trị cụ thể)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Sau khi ước lượng được $\hat{\beta}_0$ và $\hat{\beta}_1$, ta có thể tính được các giá trị dự báo của y và u tại các giá trị của x như sau:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

và

$$\hat{u}_i = y_i - \hat{y}_i$$

- $\hat{y}_i$ được gọi là giá trị thích hợp (fitted value) hoặc giá trị dự báo (predicted value) của biến phụ thuộc tại mỗi giá trị của x_i cho trước.
- $\hat{u}_i$ gọi là phần dư (residual).

Ví dụ ước lượng tác động của tỷ suất thu nhập lên tiền lương của CEO

Sử dụng bộ dữ liệu CEOSAL1.dta. Chúng ta muốn ước lượng tiền lương của CEO theo tỷ suất thu nhập trên vốn, *roe*. Giả sử hai điều kiện về sai số và không tương quan được thỏa.

Source	SS	df	MS	Number of obs	=	209
				F(1, 207)	=	2.77
Model	5166419.04	1	5166419.04	Prob > F	=	0.0978
Residual	386566563	207	1867471.32	R-squared	=	0.0132
				Adj R-squared	=	0.0084
Total	391732982	208	1883331.64	Root MSE	=	1366.6

salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
roe	18.50119	11.12325	1.66	0.098	-3.428196	40.43057
_cons	963.1913	213.2403	4.52	0.000	542.7902	1383.592

	salary	roe	salaryhat	uhat
1	1095	14.1	1224.058	-129.0581
2	1001	10.9	1164.854	-163.8543
3	1122	23.5	1397.969	-275.9692
4	578	5.9	1072.348	-494.3483
5	1368	13.8	1218.508	149.4923
6	1145	20	1333.215	-188.2151
7	1078	16.4	1266.611	-188.6108

Thực hành ước lượng OLS theo các bước

Tạo bộ dữ liệu mô phỏng và mô hình hồi quy thực. Ước lượng các tham số hồi quy dựa trên công thức:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{Cov(X, Y)}{Var(X)}$$

và

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

So sánh kết quả với mô hình hồi quy thực.

Vai trò của các giả định trong mô hình OLS

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

1. Tuyến tính theo tham số
2. Lấy mẫu ngẫu nhiên
3. Có sự thay đổi trong các giá trị của biến giải thích
4. Sai số u không tương quan với biến giải thích x, $E(u|x) = 0$

Bằng toán học, giả định (4) tương đương với:

$$\mathbf{E}(\mathbf{u}) = 0 \qquad (4.1)$$

$$\mathbf{E}(\mathbf{xu}) = 0 \qquad (4.2)$$

Vai trò của các giả định trong mô hình OLS

- Giả định (4.2) là giả định quan trọng nhất trong mô hình OLS. Rất khó chứng minh trong thực tế. Cần thiết phải hiểu sâu về lý thuyết kinh tế và quá trình thu thập dữ liệu để giải thích.
- Nếu giả định (4.2) bị vi phạm, ước lượng OLS sẽ không nhất quán.
- Toàn bộ nội dung của môn KTL 2 chỉ tập trung để giải quyết vấn đề này.

Một số ví dụ về tính hợp lý của giả định sai số không tương quan với biến giải thích

- Ước lượng mô hình tỷ suất thu nhập của việc đi học với biến giải thích là số năm đi học
- Ước lượng mô hình năng suất nông nghiệp với biến giải thích là lượng phân bón tiêu thụ
- Ước lượng hiệu quả hoạt động của doanh nghiệp với chi phí không chính thức (hối lộ)
- Ước lượng mô hình hàm cầu tiêu thụ xăng dầu với biến giải thích là giá

Lựa chọn biến và cấu trúc hàm trong mô hình hồi quy

- Cách sử dụng biến số ảnh hưởng đến ý nghĩa của mô hình.
- Sử dụng đơn vị (level), logarithm, hay tỷ lệ thay đổi được quyết định bởi mô hình kinh tế.
- Có thể lấy logarithm của biến số khi dữ liệu có phân phối lệch.

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Đánh giá độ thích hợp của các mô hình hồi quy

Dựa trên tổng bình phương (SST, còn được gọi là tổng biến thiên), tổng bình phương được giải thích (SSE), và tổng bình phương phần dư:

$$SST = \sum (y_i - \bar{y})^2$$

$$SSE = \sum (\hat{y}_i - \bar{y})^2$$

$$SSR = \sum \hat{u}_i^2$$

và

$$SST = SSE + SSR$$

Hệ số thích hợp R-bình phương được tính bằng tỷ số giữa biến thiên được giải thích và tổng biến thiên:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Hiểu thế nào về hệ số thích hợp R^2?

- Mô hình gồm có phần quan sát được x và phần không quan sát được u.
- Phần quan sát được giải thích được càng nhiều các nhân tố ảnh hưởng đến y càng tốt. Ví dụ $R^2 = 0.5$ có nghĩa là mô hình giải thích được 50% độ biến thiên của mẫu.
- $\hat{y}_i$ và $\hat{u}_i$ sẽ có quan hệ nghịch biến vì tổng biến thiên là cố định đối với mỗi mẫu.

$$0 \leq R^2 \leq 1$$

- Trên thực tế, hệ số xác định luôn $0 < R^2 < 1$.
- *Câu hỏi: Nếu $R^2 = 0$ hoặc $R^2 = 1$ thì hình dạng đường hồi quy mẫu sẽ như thế nào?*

Ví dụ mô hình tiền lương của CEO

So sánh hai mô hình với biến phụ thuộc lần lượt là tiền lương và logarithm của tiền lương. Mô hình nào phù hợp hơn? Giải thích.

Lưu ý về hệ số thích hợp R^2

- Nhìn chung những người mới nghiên cứu hay có xu hướng chọn mô hình hay biến số để tăng R^2. Điều này không sai nhưng không được khuyến khích để xây dựng mô hình.
- Sử dụng R^2 để chọn biến có thể dẫn đến những sai sót rất nghiêm trọng, đặc biệt khi biến giải thích là không ngẫu nhiên.
- Không có tiêu chí để xác định R^2 khi nào cao hay thấp.
- Với hồi quy đa biến, tăng số biến số trong mô hình làm tăng R^2, do đó cần phải cân đối giữa số biến với độ thích hợp của mô hình.

Ví dụ mô hình giá nhà

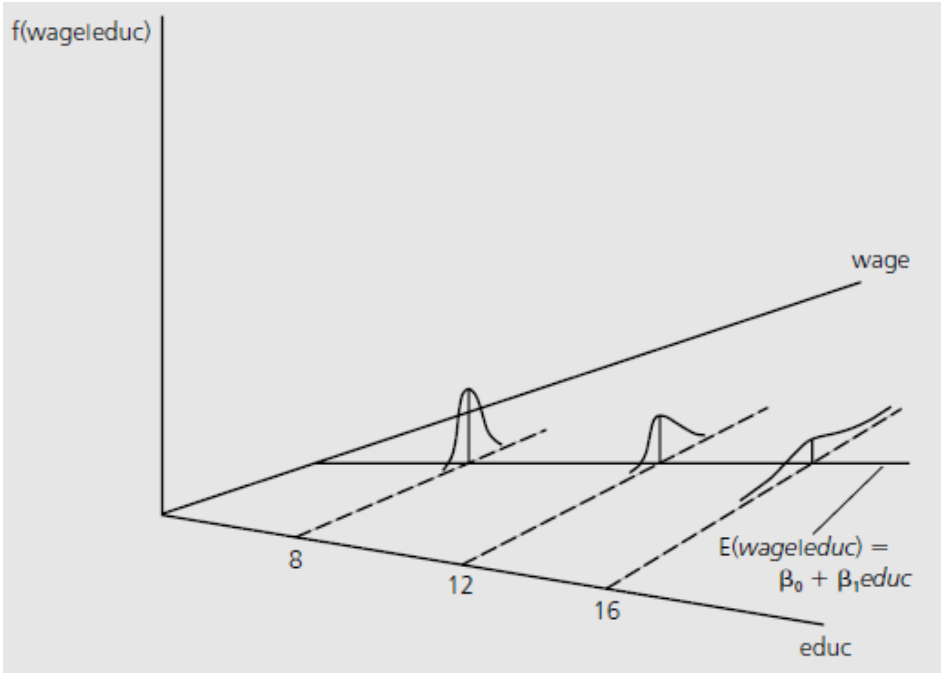
Sử dụng bộ dữ liệu hprice1.dta.

Hãy lựa chọn một mô hình hồi quy đơn biến giải thích các nhân tố ảnh hưởng đến giá nhà. Biến số nào giải thích tốt nhất? Cấu trúc hàm nào phù hợp nhất?

Giả định 5: Phương sai của sai số trong mô hình hồi quy

Nếu phương sai của sai số là $Var(u) = \sigma^2$ là một hằng số, không phụ thuộc vào các biến giải thích x, khi này ta có mô hình hồi quy đơn biến với phương sai của sai số không đổi (homoskedasticity).

- Phương sai không đổi là gì?



- Ước lượng bằng OLS có tính chất đặc biệt gọi là ước lượng tuyến tính không chệch hiệu quả nhất (Best Linear Unbiased Estimator - BLUE).