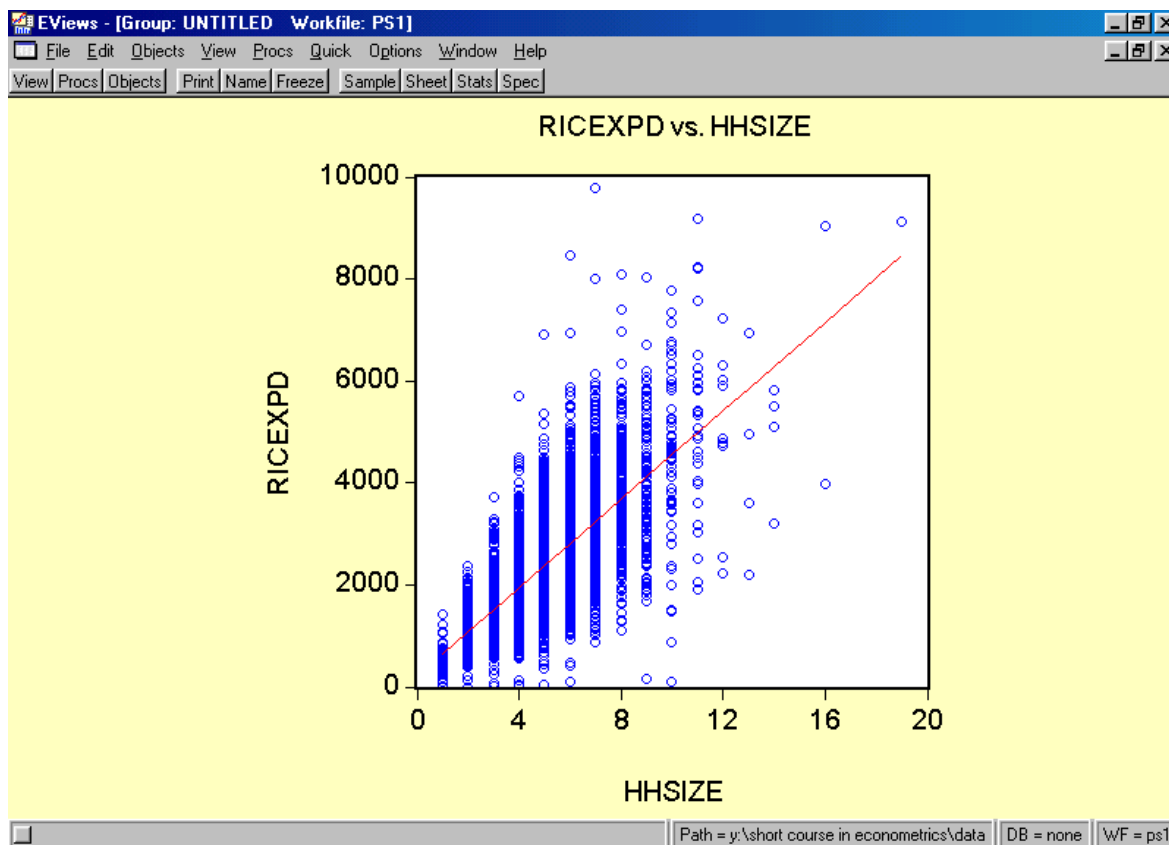


## Mô hình hồi qui tuyến tính chuẩn cổ điển

### Hàm hồi qui tổng thể

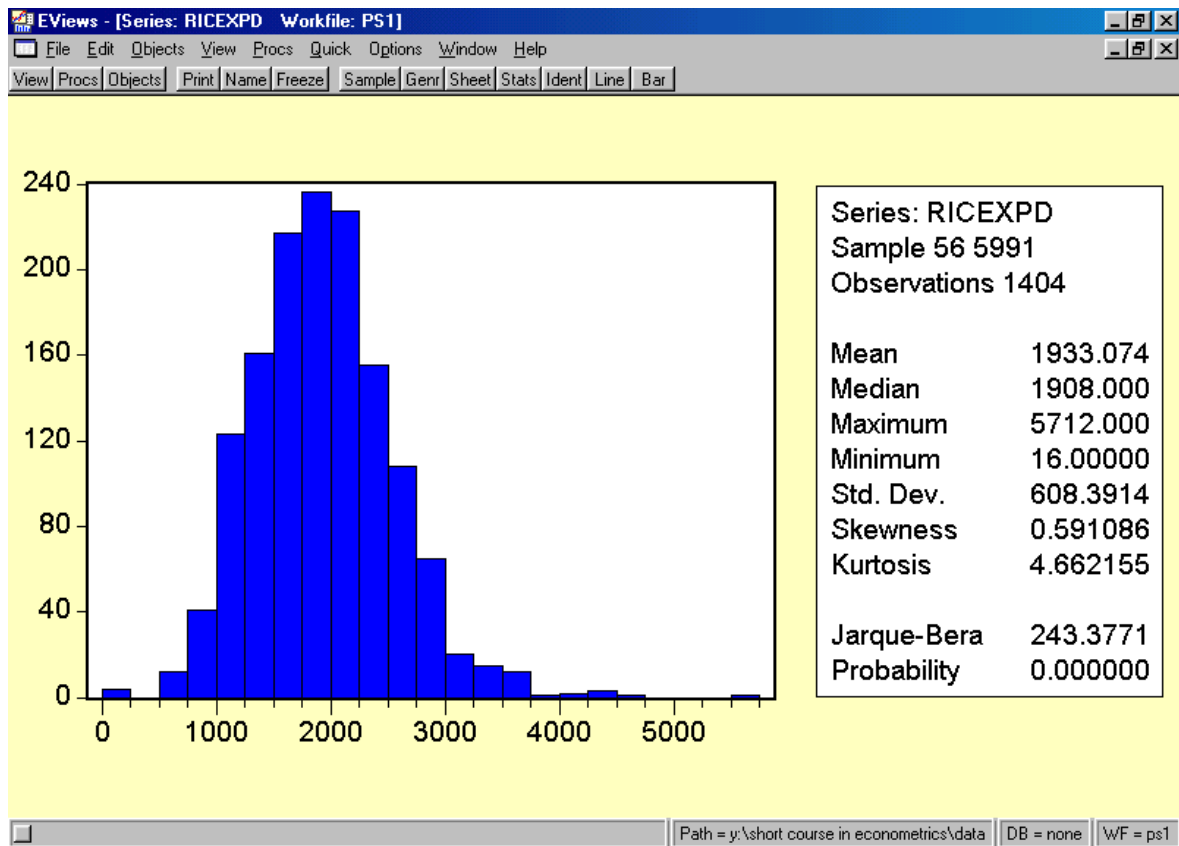
Trong phần trả lời cho Bài tập 2, Anh/Chị đã tạo ra đồ thị sau:



Trong đồ thị này, chúng ta có thể thấy phân bố các quan sát của biến RICEXPD ứng với mỗi giá trị cho trước của HHSIZE.

Trong mô hình hồi qui, chúng ta coi các giá trị của biến giải thích (X) là cho trước trong quá trình chọn mẫu lặp. Chúng ta giả định rằng X là một biến không ngẫu nhiên. Biến phụ thuộc là một biến ngẫu nhiên có điều kiện theo biến giải thích này. Chúng ta giả định điều kiện này là kỳ vọng toán học của biến phụ thuộc là một hàm tuyến tính của biến giải thích.

Đồ thị trên rất tốt dùng để minh họa điều này (nhưng nhớ rằng đồ thị này thể hiện cho một mẫu chứ không phải cho tổng thể). Nếu chúng ta giới hạn mẫu này ở những quan sát với điều kiện  $HHSIZE = 4$ , chúng ta có biểu đồ tần suất cho RICEXPD như sau:



Mối quan hệ tuyến tính giả định được trình bày theo dạng **Hàm hồi qui tổng thể (PRF)**:

$$E[Y_i | X_i] = \beta_1 + \beta_2 X_i$$

Các quan sát cụ thể về biến phụ thuộc ngẫu nhiên lệch ra khỏi giá trị kỳ vọng toán học, nên chúng ta cũng có thể viết PRF dưới dạng như sau:

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

Thành phần sai số  $\varepsilon_i$  được gọi bởi nhiều tên, trong đó hai tên thường gặp nhất là :

- nhiễu ngẫu nhiên (stochastic disturbance)
- nhiễu ngẫu nhiên (random disturbance)

Thành phần nhiễu ngẫu nhiên (stochastic) là tổng của nhiều mục :

- các biến giải thích bị bỏ sót
- sai số khi đo lường biến phụ thuộc
- tính ngẫu nhiên vốn có trong biến phụ thuộc

Chúng ta giả định rằng thành phần nhiễu ngẫu nhiên (stochastic) có hành vi tốt và hành vi của nó có thể được mô tả bằng các giả định cổ điển sau:

### Các giả định cổ điển

Giá trị trung bình zero  $E[\varepsilon_i | X_i] = 0$

Phương sai đồng nhất  $VAR[\varepsilon_i | X_i] = E[\varepsilon_i^2 | X_i] = \sigma^2$

Hệ số tự tương quan zero :  $COV[\varepsilon_i \varepsilon_j | X_i, X_j] = E[\varepsilon_i \varepsilon_j | X_i, X_j] = 0$

Không tương quan với X:  $COV[\varepsilon_i X_j | X_i, X_j] = E[\varepsilon_i X_j | X_i, X_j] = 0$

Nói chung chúng ta sẽ không phiền phức khi viết điều kiện về những biểu thức này.

Ngoài các giả định cổ điển, chúng ta thường đưa thêm vào giả định phân phối chuẩn :

Phân phối chuẩn :  $\varepsilon_i \sim N(0, \sigma^2)$

Như chúng ta thấy, những giả định này là quan trọng trong việc hình thành các phân phối chọn mẫu của các ước lượng và các trị thống kê kiểm định mà chúng ta sẽ sử dụng trong phân tích hồi qui.

### Hàm hồi qui mẫu

Hàm hồi qui tổng thể (PRF) mang tên này vì nó chỉ ra mối quan hệ tuyến tính giữa trung bình tổng thể của biến phụ thuộc và biến giải thích (biến hồi qui).

Do không có dữ liệu về tổng thể, nên chúng ta không biết giá trị trung bình tổng thể của biến phụ thuộc là đúng tới mức độ nào, vì thế chúng ta không thể tính được các giá trị thực của tung độ gốc và độ dốc của PRF. Chúng ta cần dựa vào dữ liệu mẫu để ước lượng những điều này.

Xét biểu đồ phân tán của hai biến sau

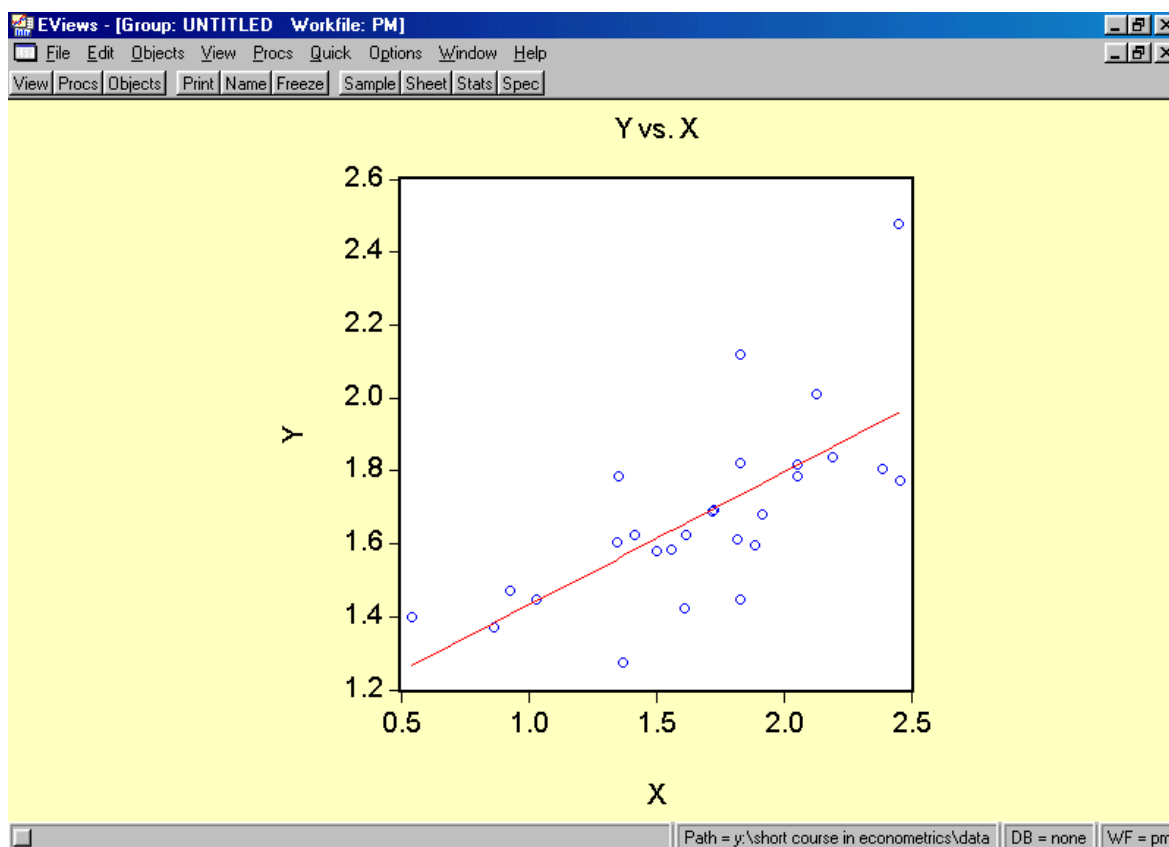
$$Y = \log(\text{output} / \text{labor}) \text{ đối với } X = \log(\text{capital} / \text{labor})$$

cho dữ liệu của 27 hãng chế tạo các kim loại sơ cấp trong file dữ liệu PM.wf1.

Đường thẳng là Hàm hồi qui mẫu (SRF) được định nghĩa như sau:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

Những thành phần trong phương trình này là các ước lượng cho các thành phần tương ứng trong PRF. Câu hỏi là : chúng ta tính chúng như thế nào ?



### Phương pháp bình phương tối thiểu

Chúng ta định nghĩa độ lệch giữa giá trị ước lượng  $\hat{Y}_i$  và giá trị quan sát  $Y_i$  là phần dư  $e_i$ . Để có được đường hồi qui “thích hợp” nhất, chúng ta chọn các ước lượng của tung độ gốc  $\hat{\beta}_1$  và độ dốc  $\hat{\beta}_2$  sao cho phần dư là nhỏ.

Trong các bài tập ước lượng trước đây, trực quan của chúng ta đã có ích vì chúng ta đã quyết định ước lượng giá trị trung bình tổng thể bằng cách sử dụng trung bình mẫu và chúng ta đã quyết định ước lượng phương sai tổng thể bằng cách sử dụng phương sai mẫu. Chúng ta đã phát hiện ra là những ước lượng này có các tính chất thống kê tốt.

Trực quan của chúng ta không có ích lắm ở đây. Trong trường hợp này, chúng ta cần một kỹ thuật được gọi là phương pháp bình phương tối thiểu và chúng ta sẽ phát hiện ra là nó cho ta các ước lượng với những tính chất thống kê tốt.

Mục tiêu của chúng ta là : chọn  $\hat{\beta}_1$  và  $\hat{\beta}_2$  để  $\sum_{i=1}^n e_i^2$  có giá trị nhỏ nhất.

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

Các điều kiện bậc nhất được rút ra là :

$$\frac{\partial \left( \sum_{i=1}^n e_i^2 \right)}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = -2 \sum_{i=1}^n e_i = 0$$

$$\frac{\partial \left( \sum_{i=1}^n e_i^2 \right)}{\partial \hat{\beta}_2} = -2 \sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i = -2 \sum_{i=1}^n e_i X_i = 0$$

Hãy nhớ rằng những điều kiện này đi kèm với các giả định cổ điển.

Giải phương trình thứ nhất tìm ra tung độ gốc ước lượng:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

Thay biểu thức này vào phương trình thứ hai và biến đổi đại số khá nhiều, ta có:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Để tiện ký hiệu, chúng ta thể hiện các độ lệch giữa giá trị quan sát và giá trị trung bình là các chữ thường và viết lại các ước lượng trên bằng các biểu thức dưới đây:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

Chính người đọc nên tự thuyết phục mình là những kết quả này là chính xác khi họ tự chứng minh.

## Tóm tắt

Vẽ một đồ thị thể hiện PRF và SRF. Nhấn mạnh vào các khác biệt giữa:

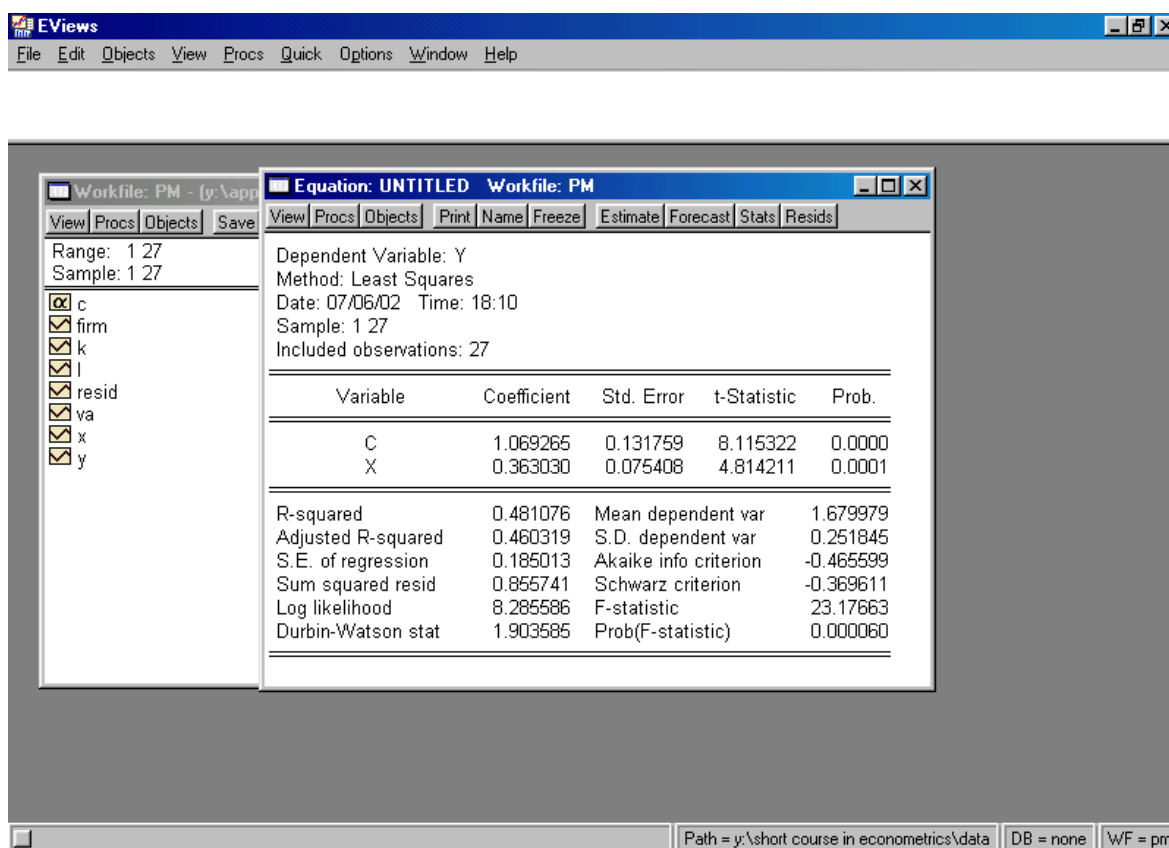
$$E[Y_i] \text{ với } Y_i \text{ với } \hat{Y}_i$$

và  $\varepsilon_i$  với  $e_i$

và  $\beta_1$  và  $\beta_2$  với  $\hat{\beta}_1$  và  $\hat{\beta}_2$

### Minh hoạ bằng kết quả từ EViews

Kỹ thuật để đạt được điều này được giải thích trong tài liệu *Giới thiệu EViews*. Đây là kết quả hồi qui tương ứng với biểu đồ phân tán đã nêu ở trang 4.



Sự giải thích thì dễ dàng: hệ số co giãn của Y theo X là 0,3630. Hãy nhớ lại rằng chúng ta đã định nghĩa Y và X như sau

Y là LOG(Giá trị gia tăng trên đơn vị lao động )  
X là LOG(Tỉ lệ vốn/lao động).

Nhớ là có rất ít con số trong bảng kết quả. Để hiểu chúng có ý nghĩa như thế nào và từ đó chúng ta có thể giải thích chúng cho những người khác, bây giờ chúng ta xét phân phối chọn mẫu đối với các ước lượng của chúng ta về tung độ gốc và độ dốc.

## Phân phối chọn mẫu của các hệ số hồi qui

Ước lượng cho hệ số độ dốc là

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

Bằng biến đổi đơn giản chúng ta có

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$$

Bây giờ thay  $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$  và đơn giản hóa một chút:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n x_i (\beta_1 + \beta_2 X_i + \varepsilon_i)}{\sum_{i=1}^n x_i^2} = \beta_2 + \frac{\sum_{i=1}^n x_i \varepsilon_i}{\sum_{i=1}^n x_i^2}$$

Đây là một biểu thức rất hữu ích .

Chúng ta có thể dễ dàng thiết lập các kết quả sau :

$$E[\hat{\beta}_2] = \beta_2$$

$$E[(\hat{\beta}_2 - \beta_2)(\hat{\beta}_2 - \beta_2)] = \text{VAR}[\hat{\beta}_2] = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

$$\hat{\beta}_2 \sim N(\beta_2, \sigma_{\hat{\beta}_2}^2)$$

Từ đó, chúng ta thấy rằng ước lượng bình phương tối thiểu cho hệ số độ dốc là không chệch và có phân phối chuẩn. Cho trước thông tin này, bây giờ chúng ta đã biết cách làm thế nào xây dựng các khoảng tin cậy và kiểm định giả thuyết nếu chúng ta biết giá trị thực của phương sai sai số.

Một vấn đề còn lại là liệu ước lượng của chúng ta có tính hiệu quả không. Chúng ta sẽ nghiên cứu vấn đề này trong bài giảng tới.

