GIỚI THIỆU HỒI QUY PHÂN VỊ

Hồi quy phân vị được Koenker và Bassett giới thiệu (1978) nhằm bổ sung cho phân tích hồi quy tuyến tính. Chủ yếu là mở rộng phân vị thông thường từ mô hình định vị sang mô hình tuyến tính tổng quát hơn, trong đó phân vị có điều kiện có dạng tuyến tính (Buchinsky (1998), trang 89). Trong phương pháp bình phương tối thiểu, mục đích đầu tiên là xác định giá trị trung bình có điều kiện của biến ngẫu nhiên Y, đã biết trước các biến giải thích x_i, đạt giá trị kỳ vọng $E[Y| x_i]$. Hồi quy phân vị vượt ra ngoài phạm vi này và cho phép ta tự đặt vấn đề như vậy ở bất kỳ phân vị nào của hàm phân bố có điều kiện.

1. PHÂN VỊ LÀ GÌ

Mô tả phân vị "đơn giản là một giá trị tương ứng với một tỷ lệ cụ thể của một mẫu được sắp xếp của một tổng thể" (Xem Gilchrist, 2001, trang 1). Ví dụ, một phân vị rất hay được sử dụng là giá trị trung vị, bằng tỷ lệ 0,5 của số liệu được sắp xếp. Nó tương ứng với phân vị có xác suất xảy ra bằng 0,5. Phân vị 0,5 này đánh dấu ranh giới của hai phần bằng nhau của hai tập hợp con liên tục của tổng thể (Xem Gilchrist, 2001).

Định nghĩa chính thức của phân vị: Giả sử Y là biến ngẫu nhiên liên tục với hàm phân bố có dạng:

$$F_Y(y) = P(Y \le y) = \tau \qquad (1)$$

Với hàm phân bố $F_Y(y)$ ta có thể xác định cho giá trị y cho trước một xác suất xuất hiện τ. Bây giờ nếu sử dụng phân vị, ta sẽ làm ngược lại, tức là với xác suất xuất hiện τ cho trước ta cần xác định giá trị y tương ứng bằng bao nhiêu. Phân vị bậc τ trong số liệu mẫu ám chỉ xác suất τ của giá trị y.

$$F_Y(y_\tau) = \tau \qquad (2)$$

Lưu ý là có hai trường hợp có thể xảy ra. Một mặt, nếu hàm $F_Y(y)$ tăng đồng biến, thì có thể xác định phân vị cho bất cứ giá trị τ nào nằm trong khoảng (0;1). Tuy nhiên, nếu phân bố của hàm không hoàn toàn đồng biến thì có thể có một số giá trị τ không xác định được giá trị phân vị duy nhất. Trong các trường hợp như vậy ta sẽ sử dụng giá trị nhỏ nhất của y cho xác suất τ. s

2. PHƯƠNG PHÁP BÌNH PHƯƠNG TỐI THIỂU

Trong phân tích hồi quy các nhà nghiên cứu quan tâm tới việc phân tích sự thay đổi của một biến phụ thuộc (y_i), biết trước thông tin về các biến độc lập (x_i) của nó. Phương pháp bình phương tối thiểu là phương pháp chuẩn để cụ thể hoá mô hình hồi quy tuyến tính và ước lượng các thông số chưa biết của nó bằng cách cực tiểu hoá tổng sai số bình phương. Điều này dẫn đến việc lấy xấp xỉ hàm trung bình của phân bố có điều kiện của biến phụ thuộc. Phương pháp bình phương tối thiểu đạt được đặc trưng BLUE. Đó là tốt nhất, tuyến tính và các ước lượng là ước lượng không chệch nếu 4 giả thiết sau đây thỏa mãn:

1. Các biến độc lập x_i không phải là các biến ngẫu nhiên.

2. Kỳ vọng toán của thành phần sai số (ε_i) bằng 0, tức là $E[\varepsilon_i] = 0$

3. Có tính thuần nhất - phương sai của thành phần sai số cố định, tức là $var(\varepsilon_i) = \sigma^2$

4. Không có tự tương quan, tức là $cov(\varepsilon_i, \varepsilon_j) = 0$, $(i \neq j)$

Tuy nhiên, thường một hoặc hai giả thiết trên bị xâm phạm, dẫn đến kết quả là phương pháp bình phương tối thiểu không còn là tốt nhất, tuyến tính và có ước lượng không chệch nữa.

Hồi quy phân vị có thể giải quyết các vấn đề vốn là nhược điểm khi áp dụng OLS trên thực tế: (i) Thường thành phần sai số không phải là không đổi trên toàn bộ phân bố vì thế đã vi phạm tiên đề về tính thuần nhất; (ii) Thông qua việc coi giá trị trung bình là độ đo về vị trí, thông tin về đuôi của phân bố bị mất đi; (iii) OLS rất nhạy cảm với các giá trị ngoại lai có thể làm sai lệch kết quả đáng kể (Xem Montenegro (2001)).

3. PHƯƠNG PHÁP HỒI QUY PHÂN VỊ

Hồi quy phân vị thực chất là chuyển hàm phân bố có điều kiện sang hàm phân vị có điều kiện bằng cách cắt nó ra thành những đoạn (tập hợp con) nhỏ. Các đoạn nhỏ này mô tả phân bố cộng dồn của biến phụ thuộc có điều kiện *Y* biết trước các biến giải thích của nó là x_i có sử dụng các phân vị được định nghĩa bằng phương trình 4.

Đối với biến phụ thuộc *Y* biết trước các biến giải thích X=x của nó và biết trước xác suất xuất hiện bằng τ, $0<\tau<1$, hàm phân vị có điều kiện được định nghĩa là phân vị bậc τ, $Q_{Y|X}(\tau|x)$, của hàm phân bố có điều kiện $F_{Y|X}(y|x)$. Để ước lượng vị trí của hàm phân bố có điều kiện, trung vị có điều kiện, $Q_{Y|X}(0,5|x)$, có thể được sử dụng làm một phương án khác của giá trị trung bình có điều kiện (Xem Lee (2005)).

Ta có thể minh hoạ hồi quy phân vị khi so sánh nó với phương pháp bình phương tối thiểu. Trong phương pháp bình phương tối thiểu, mô hình hoá hàm phân bố có điều kiện của mẫu ngẫu nhiên $(y_1,...,y_n)$ có hàm tham số $\mu(x_i,\beta)$, trong đó x_i là các biến độc lập, β là ước lượng tương ứng và μ là giá trị trung bình có điều kiện, ta có bài toán cực tiểu hóa sau đây:

$$\min_{\beta \in R} \sum_{i=1}^{n} (y_i - \mu(x_i, \beta))^2 \qquad (3)$$

Theo cách ấy ta nhận được hàm kỳ vọng có điều kiện $E[Y|x_i]$. Bây giờ bằng cách tương tự ta có thể tiến hành tương tự trong hồi quy phân vị. Đặc trưng chính bây giờ là hàm ρ_τ, một hàm được coi như là hàm kiểm tra.

$$\rho_\tau(x) = \begin{cases} \tau * x & \text{nếu } x > 0 \\ (\tau - 1) * x & \text{nếu } x < 0 \end{cases} \qquad (4)$$

Hàm kiểm tra này đảm bảo là:

1. Tất cả các ρ_τ đều dương (>0)

2. Đơn vị đo phù hợp với xác suất τ

Hàm số với hai thành phần bổ trợ như vậy phải được đưa ra nếu đề cập tới khoảng cách L1, có thể trở thành âm (<0).

Trong hồi quy phân vị, ta cực tiểu hóa hàm sau:

$$\min_{\beta \in R} \sum_{i=1}^{n} \rho_\tau (y_i - \varepsilon(x_i, \beta)) \qquad (5)$$

Ở đây, trái ngược với phương pháp bình phương tối thiểu, cực tiểu hóa được tiến hành cho từng tập con được định nghĩa bởi ρ_τ, trong đó ước lượng của hàm phân vị bậc τ đạt được với hàm thông số $\varepsilon(x_i, \beta)$ (Koenker và Hallock (2001)).

Các đặc điểm đặc trưng cho hồi quy phân vị và điều làm nó khác với các phương pháp hồi quy khác như sau:

1. Phân bố có điều kiện của biến phụ thuộc Y có thể được đặc trưng thông qua các gía trị khác nhau của τ.

2. Có thể phát hiện ra sự phân tán.

3. Nếu số liệu phân tán, ước lượng hồi quy trung vị có thể hiệu quả hơn ước lượng hồi quy trung bình.

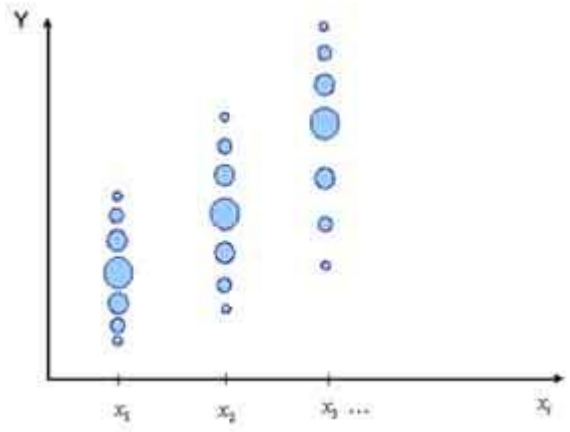
4. Bài toán cực tiểu như trình bày ở phương trình (5) có thể giải được bằng phương pháp quy hoạch tuyến tính, và ước lượng một cách dễ dàng.

5. Các hàm phân vị đồng thời cũng tương đương với các phép biến đổi đơn. Tức là $Q_{h(Y \mid X)}(x_\tau) = h(Q_{(Y \mid X)}(x_\tau))$, đối với bất kỳ một hàm nào.

6. Phân vị được coi là thô trên góc độ các điểm ngoại lai (Lee (2005)).

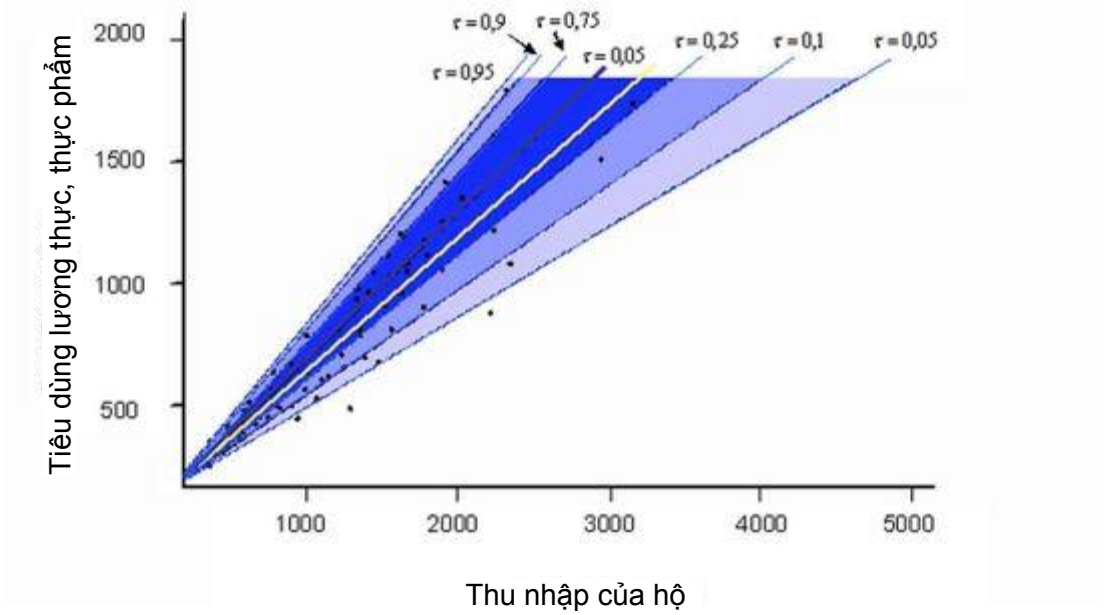
Minh hoạ bằng đồ thị hồi quy phân vị

Trước khi bắt đầu các minh họa bằng số, mục này sẽ trình bày khái niệm hồi quy phân vị bằng đồ thị. Trước tiên, chúng ta xem xét hình 1. Đối với giá trị cho trước của biến giải thích x_i mật độ của biến phụ thuộc Y được biểu thị bằng độ lớn của các vòng tròn. Vòng tròn càng lớn, mật độ càng cao, với ***một*** là nơi mà mật độ cao nhất, thuộc về giá trị của x_i có vòng tròn lớn nhất. Hồi quy phân vị trước tiên là kết nối những vòng tròn cùng kích cỡ, đó chính là các xác suất, thuộc các x_i khác nhau, nhờ vậy cho phép ta tập trung vào mối quan hệ tương tác giữa các biến độc lập x_i và biến phụ thuộc *Y* cho các phân vị khác nhau, như có thể thấy ở hình 2. Các tập con này được đánh dấu bằng các đường phân vị, phản ánh mật độ xác suất của biến phụ thuộc Y cho trước biến độc lập x_i.



Hình 1: Xác suất xuất hiện của từng biến giải thích

Thí dụ được sử dụng trong hình 2 nguyên gốc là của Koenker và Hallock (2000), và nó minh họa ứng dụng thực nghiệm của Ernst Engel's (1857). Ông đã nghiên cứu mối quan giữa tiêu dùng thực phẩm - được coi là biến phụ thuộc, và thu nhập là biến độc lập. Trong hồi quy phân vị hàm có điều kiện của $Q_{Y \mid X}(\tau|X)$ được phân ra theo phân vị bậc τ. Trong phân tích, phân vị bậc τ, $\tau \in \{0{,}05; 0{,}1; 0{,}25, 0{,}5; 0{,}75; 0{,}9; 0{,}95\}$, được đánh dấu bằng các đường nét mảnh. Chúng được tách bạch bằng các vạt đậm nhạt khác nhau, các thông tin được đặt phía trên. Trung vị có điều kiện ($\tau = 0{,}5$) được đánh dấu bằng đường nét dày. Đường trung bình có điều kiện được đánh dấu bằng đường mầu trắng. Các vạt mầu trình bày các tập số liệu được khái quát hóa bởi các phân vị.



Hình 2: Đường cong với giá trị trung vị được biểu thị bằng đường đen dày, còn đường trung bình được biểu thị bằng đường trắng mảnh.

Hình 2 có thể hiểu là sơ đồ đường viền trình bày đồ thị ba chiều, với chi tiêu cho lương thực thực phẩm và thu nhập, tương ứng với trục tung (y) và trục hoành (x). Chiều thứ ba là mật độ xác suất của các giá trị tương ứng. Mật độ của một giá trị được biểu thị bằng độ xậm của mầu. Mầu càng xậm, xác suất xuất hiện càng cao. Ví dụ, ở đường biên ngoài, nơi mà mầu rất nhạt, mật độ xác suất cho tập số cho trước tương đối thấp, vì chúng được đánh dấu bằng các phân vị từ 0,05 đến 0,1 và từ 0,9 đến 0,95. Một điều quan trọng cần phải lưu ý là hình 2 trình bày cho từng tập hợp con xác suất xuất hiện riêng. Tuy nhiên, phân vị sử dụng xác suất cộng dồn của hàm điều kiện. Ví dụ, giá trị τ = 0,05 có nghĩa là 5% số quan sát nằm phía dưới đường này, τ = 0,25 có nghĩa là 25% số quan sát nằm dưới đường này và đường 0,1.

Đồ thị trong hình 2 cho thấy là sự biến động của sai số không trải đều trong phân bố. Sự phân tán của tiêu dùng lương thực, thực phẩm tăng khi thu nhập của các hộ tăng. Đồng thời số liệu lệch về bên trái, được biểu thị bằng khoảng cách của các đường phân vị giảm phía trên đường trung vị. Đường này nằm phía trên đường trung bình. Điều đó cho thấy là tiên đề mà phương pháp bình phương tối thiểu dựa vào là sự đồng nhất bị xâm phạm. Vì vậy các nhà thống kê đã khuyên sử dụng một phương pháp phân tích khác, đó là phương pháp hồi quy phân vị, phương pháp có khả năng giải quyết vấn đề này.

KẾT LUẬN

Đối với hàm phân bố $F_Y(y)$ ta có thể xác định cho giá trị của y một xác suất xuất hiện τ. Bây giờ, đối với phân vị ta làm tương tự nhưng theo chiều ngược lại. Tức là, đối với xác suất τ cho trước của tập hợp số liệu con, ta cần xác định giá trị tương ứng của y.

(tiếp theo trang 23)

GIỚI THIỆU HỒI QUY PHÂN VỊ *(tiếp theo trang 14)*

Ở phương pháp bình phương tối thiểu, mục tiêu đầu tiên của ta là xác định giá trị trung bình có điều kiện của biến Y, biết trước giá trị của các biến giải thích x_i, $E[Y|x_i]$. Hồi quy phân vị tiến xa hơn điều đó. Nó cho phép ta nghiên cứu vấn đề đó ở mọi phân vị của hàm phân bố có điều kiện. Nó tập trung vào phân tích mối quan hệ giữa biến phụ thuộc và các biến độc lập của nó ở phân vị cho trước. Hồi quy phân vị khắc phục một số vấn đề mà phương pháp bình phương tối thiểu mắc phải. Thông thường, thành phần sai số không cố định trong phân bố, vì vậy xâm phạm tiên đề về tính đồng nhất. Đồng thời thông qua việc coi giá trị trung bình là độ đo vị trí, thông tin về đuôi phân bố sẽ bị mất đi. Và cuối cùng phương pháp bình phương tối thiểu rất nhậy cảm với các giá trị ngoại lai, một giá trị có thể làm sai lệch đáng kể kết quả■

Phan Ngọc Trâm *lược dịch từ: Statistics: Numerical Methods/Quantile regression. From Wikibooks, the open-content textbooks collection. Statistic: Numerical Methods*