MÔ HÌNH HỒI QUY BOOTSTRAP VỚI CỠ MẪU NGẪU NHIÊN

ON BOOTSTRAPPING REGRESSION MODEL WITH RANDOM RESAMPLE SIZE

Nguyễn Hồng Nhung
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh, Việt Nam

Ngày toà soạn nhận bài 9/11/2016, ngày phản biện đánh giá 7/12/2016, ngày chấp nhận đăng 6/3/2017

TÓM TẮT

Nhiều phương pháp thống kê cổ điển khi tìm khoảng tin cậy cho các hệ số hồi quy cần giả thiết về phân bố tiên nghiệm của các sai số. Với một số giả thuyết nhất định, không cần giả thiết về phân phối của sai số, thủ tục bootstrap có cỡ mẫu lấy lại cố định hoặc ngẫu nhiên có thể thực hiện xấp xỉ bootstrap của phân phối ước lượng bình phương tối thiểu các hệ số hồi quy. Trong bài báo này, tác giả trình bày thuật toán xác định hệ số hồi quy của mô hình hồi quy bootstrap với cỡ mẫu lấy lại là biến ngẫu nhiên N_n. N_n nhận giá trị là các số nguyên dương trên $[m,n]$ với khả năng là như nhau tại mọi giá trị, trong đó m là số nguyên dương nhỏ nhất lớn hơn hoặc bằng $n/4$. Sử dụng phần mềm Matlab xác định hệ số hồi quy bootstrap thực nghiệm và đưa ra nhận xét.

Từ khóa: *Phương pháp bootstrap; hồi quy; lấy lại mẫu; cỡ mẫu ngẫu nhiên; phân phối đều.*

ABSTRACT

To find confidence interval for regression coefficients, classical methods require the distribution of errors. Under mild conditions, without knowing the distribution of errors, the bootstrap approximation with fixed or random resample sizeto estimate the distribution of the least squares is valid. In this paper, the author presents algorithms to determine regression coefficients of the bootstrap regression model with random resample size N_n. N_n is a positive integer-valued in $[m,n]$ with the ability to be the same at all values, where m is the smallest positive integer greater than or equal to $n/4$. Matlab software is used to seek the empirical bootstrap regression coefficients and create analysis comments.

Key words: *bootstrap; regression; resampling; random resample size; uniform distribution.*

1. GIỚI THIỆU

Năm 1979 Efron [1] đưa ra một quá trình tổng quát lấy lại mẫu từ mẫu gốc ban đầu gọi là bootstrap. Coi mẫu gốc $S_n = (X_1, X_2, \dots, X_n)$đóng vai trò là tổng thể mà từ đó nó được rút ra. Từ mẫu ban đầu lấy lại mẫu ngẫu nhiên bằng phương pháp lấy mẫu có hoàn lại. Mẫu lấy lại gọi là mẫu bootstrap ngẫu nhiên $S_n^* = (X_{n1}^*, X_{n2}^*, \dots, X_{nn}^*)$ có cỡ mẫu n .Giả sử $X_1, X_2, \dots, X_n$ độc lập cùng phân phốiFvà$\theta(F)$ là tham số cần quan tâm. GọiF_n là hàm phân phối thực nghiệm của mẫu $S_n, \theta(F_n)$ là một ước lượng của $\theta(F)$.Ứng với mỗi mẫu bootstrap, thống kê của tham số cần quan tâm $\theta(F_n^*)$ được gọi là thống kê bootstrap. Phân phối thực nghiệmF_n^*của thống kê bootstrap được gọi là phân phối bootstrap. Phân phối bootstrap là ước lượng của phân phối thống kê ta đang quan tâm. Phương pháp bootstrap của Efron xấp xỉ phân phối mẫu của $\sqrt{n}\big(\theta(F_n) - \theta(F)\big)$ bởi phân phối mẫu lặp lại $\sqrt{n}\big(\theta(F_n^*) - \theta(F_n)\big)$ dựa trên mẫu bootstrap S_n^* mà trong đó phân phối ban đầu F được thay thế bởi phân phối thực nghiệm F_ndựa trên mẫu gốc $S_n$vàF_n được thay thế bởi phân phối thực nghiệm bootstrap F_n^* dựa trên mẫu bootstrap S_n^*. Enno Mammen [2] giới thiệu quá trình lấy mẫu bootstrap với cỡ mẫu là biến ngẫu nhiên có phân phối Poisson.

Trong [3] Rao, Pathak và Kolt trình bày quá trình lấy mẫu bootstrap là quá trình lấy ngẫu nhiên lần lượt có hoàn lại các phần tử từ S_n cho đến khi có $m = [n(1-e^{-1})] + 1$ phần tử phân biệt trong mẫu gốc. Như vậy, ta thu được mẫu bootstrap $S_{N_n}^* = \left(X_{n1}^*, X_{n2}^*, \dots, X_{nN_n}^*\right)$ có cỡ mẫu N_n là ngẫu nhiên, miễn là trong $X_{n1}^*, X_{n2}^*, \dots, X_{nN_n}^*$ có $m \approx n(1-e^{-1})$ phần tử phân biệt trong mẫu gốc. Cỡ mẫu N_n có thể phân tích thành tổng các biến ngẫu nhiên độc lập như sau:

$$N_n = N_{n1} + N_{n2} + \cdots + N_{nm} \quad (1)$$

trong đó $m = [n(1-e^{-1})] + 1$; $N_1 = 1$ và với mỗi $k, 2 \le k \le m$,

$$P^*(N_{nk} = i) = \left(1 - \frac{k-1}{n}\right)\left(\frac{k-1}{n}\right)^{i-1}, \quad (2)$$

với P^* là ký hiệu xác suất có điều kiện $P(\dots|X_1, \dots, X_n)$.

Kỳ vọng của cỡ mẫu lấy lại N_ncủa thủ tục bootstrap này là $E(N_n) = n\left[\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-m+1}\right]$.Với$m = n(1-e^{-1})$ suy ra

$$E(N_n) = n + O(1). \quad (3)$$

Rao, Pathak và Kolt đã thiết lập tính vững của lược đồ lấy mẫu này trong [3].

Trong [4] N.V. Toản đã nghiên cứu quá trình bootstrap với cỡ mẫu lấy lại $N_n$không độc lập với mẫu gốc và thỏa mãn điều kiện: có một dãy các số nguyên dương $(k_n)_{1\le n<\infty}$ tiến ra ∞ khi n tiến đến ∞ sao cho dãy $\left(\frac{N_n}{k_n}\right)_{1\le n<\infty}$ hội tụ theo xác suất có điều kiện đến một biến ngẫu nhiên dương v với xác suất 1. Kết quả đạt được cho thấy có thể sử dụng ước lượng bootstrap với cỡ mẫu ngẫu nhiên thay cho ước lượng bootstrap với cỡ mẫu n (?).

Trong trường hợp cỡ mẫu lặp lại là biến ngẫu nhiên nhận giá trị nguyên dương N_nđộc lập với dãy $X_1, X_2, \dots$;và thỏa điều kiện

$$N_n \to_P \infty \text{khi} n \to \infty, \quad (4)$$

thì với hầu hết mọi dãy mẫu $X_1, X_2, \dots$,

$$\left\|F_{N_n}^* - F\right\| \to_P 0 \text{ khi } n \to \infty. \quad (5)$$

Ở đây,

$$\left\|F_{N_n}^* - F\right\| = sup_{-\infty<t<\infty}\left|F_{N_n}^*(t) - F(t)\right|,$$

với $F_{N_n}^*$là phân phối thực nghiệm dựa trên mẫu bootstrap $S_{N_n}^*$ có cỡ mẫu ngẫu nhiên làN_n. Kết quả này N.V. Toản đã chứng minh trong [5] cho thấy quá trình bootstrap thực nghiệm có hiệu lực khi N_n thỏa mãn (4).

Trong [6] N.V. Toản đã đưa ra điều kiện tổng quát cho cỡ mẫu ngẫu nhiên để quá trình bootstrap thực nghiệm tổng quát với cỡ mẫu ngẫu nhiên được đánh dấu bởi một lớp các hàm $\mathcal{F}$ và dựa trên độ đo xác suất P thỏa mãn định lý giới hạn trung tâm.

Mục tiếp theo trình bày thủ tục bootstrap đối với mô hình hồi quy trong trường hợp cỡ mẫu lấy lại là biến ngẫu nhiên nhận giá trị nguyên dương và độc lập với mẫu gốc. Đồng thời trình bày các điều kiện có thể sử dụng ước lượng bootstrap với cỡ mẫu ngẫu nhiên đối với phân phối của ước lượng bình phương bé nhất. Các kết quả này được chứng minh trong các tài liệu [7] và [8].Phần cuối mục, tác giả minh họa ứng dụng của các kết luận lý thuyết bởi quá trình xác định khoảng tin cậy cho hệ số hồi quy thực nghiệm cho mô hình hồi quy bootstrap với cỡ mẫu ngẫu nhiên. Cụ thể, cỡ mẫu lấy lại $N_n$là biến ngẫu nhiên nhận giá trị là các số nguyên dương trên $[m, n]$ với khả năng như nhau tại mọi giá trị, m là số nguyên dương nhỏ nhất lớn hơn hoặc bằng $n/4$. Khi đó,

$$P(N_n = i) = \frac{1}{n-m+1}, i \in \{m, \dots, n\} \quad (6)$$

$$P(N_n \ge k) \to 1 \text{khi } n \to \infty, \text{ với mọi } k. \quad (7)$$

Như vậy N_n là biến ngẫu nhiên nhận giá trị nguyên dương thỏa mãn (4).

2. MÔ HÌNH HỒI QUY BOOTSTRAP VỚI CỠ MẪU NGẪU NHIÊN

2.1 Mô hình hồi quy

Xét mô hình tuyến tính bội

$$Y(n) = X(n)\beta + \varepsilon(n). \quad (8)$$

Trong phương trình này β là một $p \times 1$ vectơ của các tham số chưa biết được ước lượng từ dữ liệu. $Y(n)$ là một $n \times 1$ vectơ dữ liệu, $Y(n)$ là vectơ ngẫu nhiên hay biến đáp

ứng. $X(n)$ là một $n \times p$ ma trận dữ liệu có hạng là $p \leq n$, $X(n)$ còn được gọi là ma trận thiết kế. $\varepsilon(n)$ là một $n \times 1$ vectơ không quan sát được, $\varepsilon(n)$ được gọi là sai số ngẫu nhiên, phần dư hay nhiễu. Dữ liệu quan sát có dạng $(X(n), Y(n))$ và ta gọi $X(n)$ là tập hợp các điểm thiết kế của mô hình. (X_i, Y_i) là hàng thứ $i, 1 \leq i \leq n$, của $(X(n), Y(n))$. Ta gọi (8) là mô hình hồi quy nếu các phân tích được đưa ra dựa trên các điểm thiết kế $X(n)$.

Giả thiết (8) thỏa mãn các điều kiện:

(A1) $X(n)$ là không ngẫu nhiên.

(A2) Trong mô hình (8) các thành phần $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ của $\varepsilon(n)$ là độc lập có cùng phân phối F với trung bình bằng 0 và phương sai σ^2. Cả F và σ^2 đều chưa biết.

Ước lượng bình phương bé nhất cho β là

$$\hat{\beta}(n) = (X(n)^T X(n))^{-1} X(n)^T Y(n). \quad (9)$$

Vectơ Y được khảo sát là giá trị quan sát của vectơ ngẫu nhiên $X(n)\beta + \varepsilon(n)$. Khi đó $\hat{\beta}(n)$ có trung bình β và ma trận hiệp phương sai $\sigma^2\{X(n)^T X(n)\}^{-1}$. Giả sử

(A3) $\frac{1}{n}\{X(n)^T X(n)\} \to V$ xác định dương.

Đồng thời giả sử rằng các phần tử của $X(n)$ đều bé so với $\sqrt{n}$. Khi đó $\sqrt{n}\big(\hat{\beta}(n) - \beta\big)$ tiệm cận chuẩn với trung bình 0 và ma trận hiệp phương sai $\sigma^2 V^{-1}$. Đặc biệt, phân phối của $\{X(n)^T X(n)\}^{-1/2}\{\hat{\beta}(n) - \beta\}/\sigma$ tiệm cận chuẩn với trung bình 0 và ma trận hiệp phương sai là ma trận đơn vị cấp p.

Nếu ta thêm giả thiết, các phần dư ε_i có cùng phân phối chuẩn $N(0, \sigma^2)$, tức là $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ có phân phối chuẩn $N(0, \sigma^2 I_n)$. Khi đó ta có thể xác định khoảng tin cậy cho các hệ số hồi quy β_i và thực hiện các kiểm định về hệ số hồi quy. Trong [9] N.H. Dư đã chỉ ra khi ε có phân phối chuẩn $N(0, \sigma^2 I_n)$ thì $U = (X^T X)^{1/2}\big(\hat{\beta} - \beta\big)$ có phân phối chuẩn $\big(0, \sigma^2 I_p\big)$; $\big(\hat{\beta} - \beta\big)$ có phân phối chuẩn $N(0, \sigma^2 (X^T X)^{-1})$. Nếu như điều kiện về phân phối chuẩn của mô hình không chỉ ra được thì quá trình lấy mẫu bootstrap sẽ là một lựa chọn để giải quyết các bài toán thuộc dạng này.

2.2 Mô hình hồi quy bootstrap

Giả thiết rằng mô hình hồi quy (8) thỏa mãn các điều kiện A(1-3). Ta xem $X(n)$ là n hàng đầu tiên của một dãy vô hạn các hàng. Tương tự, xem $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ là n phần tử đầu tiên của dãy vô hạn các biến ngẫu nhiên độc lập cùng phân phối F. Từ mẫu gốc $(X(n), Y(n))$ ta tính được ước lượng bình phương bé nhất của β là $\hat{\beta}(n)$. Từ đó, ta xác định được vectơ phần dư $\hat{\varepsilon}(n)$ xác định bởi

$$\hat{\varepsilon}(n) = Y(n) - X(n)\hat{\beta} \quad (10)$$

Gọi $\hat{F}_n$ là phân phối thực nghiệm của $\hat{\varepsilon}(n)$, có trung tâm tại kỳ vọng, nên $\hat{F}_n$ đặt trọng lượng $1/n$ tại $\hat{\varepsilon}_i(n) - \hat{\mu}_n$ và $\int x\, d\hat{F}_n^x = 0$. Theo E. Mammen [10], thực hiện quá trình lấy mẫu bootstrap từ tập các phần dư trung tâm $\{\hat{\varepsilon}_1 - \hat{\varepsilon}_\cdot\}$, trong đó $\hat{\varepsilon}_\cdot = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i$ ta được các bootstrap sai số $\hat{\varepsilon}^*(n)$ là n vectơ mà thành phần thứ i là $\hat{\varepsilon}_i^*$; giả sử $\hat{\varepsilon}_1^*, \dots, \hat{\varepsilon}_n^*$ độc lập có điều kiện cùng phân phối $\hat{F}_n$. Đặt

$$Y^*(n) = X(n)\hat{\beta}(n) + \hat{\varepsilon}^*(n). \quad (11)$$

Bây giờ ta có bộ số liệu đánh dấu sao để ước lượng tham số. Ước lượng bootstrap của $\hat{\beta}(n)$ là

$$\hat{\beta}^*(n) = \big(X(n)^T X(n)\big)^{-1} X(n)^T Y^*(n) \quad (12)$$

Nguyên lý bootstrap cho rằng phân phối của $\sqrt{n}\big(\hat{\beta}^* - \hat{\beta}\big)$, mà ta có thể tính trực tiếp từ dữ liệu, xấp xỉ phân phối của $\sqrt{n}\big(\hat{\beta} - \beta\big)$. Freedman [7] đã chứng minh rằng xấp xỉ này là rất tốt khi n lớn và $\sigma^2 p.\, trace(X^T X)^{-1}$ nhỏ.

Trong [7] Freedman đã phát triển một số định lý xấp xỉ ứng dụng trong mô hình hồi quy bootstrap của Efron với cỡ mẫu lấy lại là m khác với n là cỡ mẫu ban đầu. Dữ liệu đánh dấu sao sinh bởi

$$\underset{m \times 1}{Y^*(m)} = \underset{m \times p}{X(m)}\ \underset{p \times 1}{\hat{\beta}(n)} + \underset{m \times 1}{\varepsilon^*(m)} \quad (13)$$

với $\hat{\varepsilon}_1^*, \dots, \hat{\varepsilon}_m^*$ độc lập có điều kiện cùng phân phối $\hat{F}_n$. Bây giờ $\hat{\beta}^*(m)$ là tham số ước lượng dựa trên dữ liệu đánh dấu sao:

$$\underset{p \times 1}{\hat{\beta}^*(m)} = \underset{p \times p}{\big(X(m)^T X(m)\big)^{-1}}\ \underset{p \times m}{X(m)^T}\ \underset{m \times 1}{Y^*(m)} \quad (14)$$

$\sqrt{m}\left(\hat{\beta}^*(m) - \widehat{\beta}\,(n)\right)$ là xấp xỉ phân phối rất tốt của $\sqrt{n}\left(\hat{\beta} - \beta\right)$ khi m lớn và $\sigma^2 p.trace(X^T X)^{-1}$ nhỏ. Trong [7] Freedman đã khẳng định hầu chắc chắn của tiệm cận khi m và n tiến tới ∞. Trong [8] N.V. Toản đã chứng minh quá trình bootstrap có hiệu lực với mô hình hồi quy nếu cỡ mẫu bootstrap N_n là biến ngẫu nhiên nhận giá trị nguyên dương, độc lập với $Y_1, Y_2, \dots, Y_n$ và thỏa mãn (4).

2.3 Mô hình hồi quy bootstrap với cỡ mẫu ngẫu nhiên

Giả sử mô hình hồi quy (8) thỏa mãn A(1-3). Theo hầu hết các dãy mẫu, cho $Y_1, Y_2, \dots, Y_n$, N. V. Toản trong [8] đã chứng minh được khi n tiến tới ∞:

(B1) Phân phối có điều kiện của $\sqrt{N_n}\{\hat{\beta}^*(N_n) - \hat{\beta}(n)\}$ hội tụ yếu đến phân phối chuẩn với trung bình 0 và ma trận hiệp phương sai $\sigma^2 V^{-1}$.

(B2) Phân phối có điều kiện của $\hat{\sigma}^*_{N_n}$ hội tụ đến điểm có khối lượng tại σ.

(B3) Phân phối có điều kiện của $\{X(N_n)^T X(N_n)\}^{-1/2}\{\hat{\beta}^*(N_n) - \hat{\beta}(n)\}/\hat{\sigma}^*_{N_n}$ hội tụ đến phân phối chuẩn trong $\mathbb{R}^p$.

Để minh họa cho các kết quả đã được chứng minh trong lý thuyết, tác giả xây dựng quá trình xác định hệ số hồi quy bootstrap thực nghiệm với cỡ mẫu thực nghiệm là một biến ngẫu nhiên. Các bước thực hiện quá trình lấy lại mẫu bootstrap từ mẫu gốc ban đầu và xác định hệ số hồi quy của mô hình hồi quy bootstrap với cỡ mẫu ngẫu nhiên được trình bày như sau:

Bước 1: Từ số liệu gốc ban đầu (X_i, Y_i) trong đó $1 \le i \le n$ ta tính được ước lượng bình phương bé nhất $\hat{\beta}(n)$ của β trong mô hình hồi quy (8) theo công thức (9).

Bước 2: Xác định các thành phần của vectơ phần dư $\hat{\varepsilon}(n)$là $\hat{\varepsilon}_i = Y_i - X_i\hat{\beta}, 1 \le i \le n$.

Bước 3: Xác định một giá trị ngẫu nhiên của biến ngẫu nhiên N_n.Lấy ngẫu nhiên lần lượt có hoàn lại từ tập các phần dư trung tâm $\{\hat{\varepsilon}_1 - \hat{\varepsilon}.\}$, trong đó $\hat{\varepsilon}. = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i$, ta được các bootstrap sai số $\hat{\varepsilon}^*(N_n)$ là N_n vectơ mà thành phần thứ i là $\hat{\varepsilon}^*_i$.

Bước 4: Đặt $Y^*(N_n) = X(N_n)\hat{\beta}(n) + \hat{\varepsilon}^*(N_n)$ với thành phần thứ $i, 1 \le i \le N_n$ là $Y^*_i = X_i\hat{\beta} + \hat{\varepsilon}^*_i$.

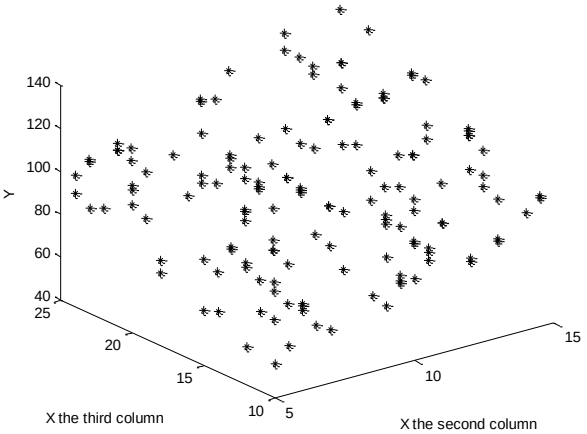
Bước 5: Với mỗi dữ liệu sao $\left(X(N_n), Y^*(N_n)\right)$ ta tính được ước lượng bootstrap của $\hat{\beta}(n)$ là

$$\hat{\beta}^*(N_n) = \left(X(N_n)^T X(N_n)\right)^{-1} X(N_n)^T Y^*(N_n) \quad (15)$$

là một vectơ $p \times 1$.

Ta xét một ví dụ minh họa về mô hình $Y = X\beta + \varepsilon$ có vectơ tham số $\beta = (\beta_1, \beta_2, \beta_3)^T$ chưa biết đang cần ước lượng; vectơ dữ liệu Y cấp 150×1; ma trận thiết kế X cấp 150×3 và vectơ sai số $\varepsilon = (\varepsilon_1, \dots, \varepsilon_{150})^T$ không quan sát được.

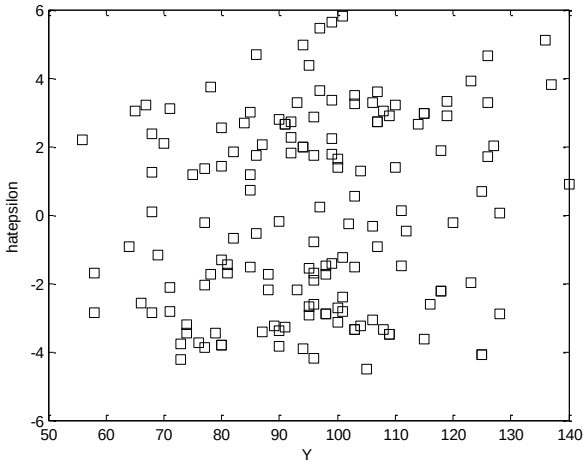
Đầu tiên ta khảo sát đồ thị của các dữ liệu.



Hình 1. *Đồ thị phân tán biểu diễn mối quan hệ giữa X và Y*

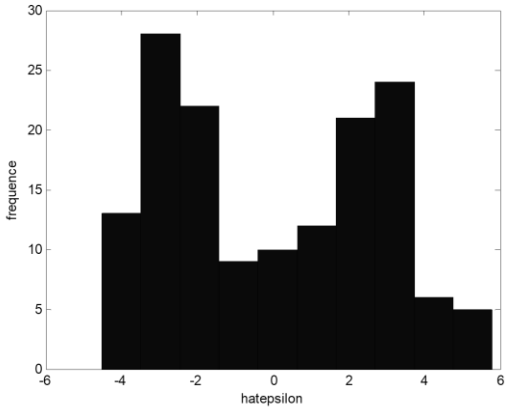
Theo hình 1, các điểm tập trung gần một mặt phẳng nên ta dự đoán có thể sử dụng mô hình hồi quy tuyến tính để biểu diễn mối quan hệ giữa X và Y.

Từ các sai số $\hat{\varepsilon}_i$ tính được ta vẽ đồ thị phân tán của $\hat{\varepsilon}_i$ theo giá trị dự đoán y_i, được hình 2. Xu thế trong đồ thị sẽ chứng tỏ các sai số $\hat{\varepsilon}_i$có độc lập hay phụ thuộc với y_i.



Hình 2. *Đồ thị phân tán của các sai số $\hat{\varepsilon}_i$ và giá trị dự đoán y_i*

Trong hình 2 ta thấy không có xu thế nào của chùm điểm thể hiện mối quan hệ giữa sai số $\hat{\varepsilon}_i$ và giá trị dự đoán y_i nên ta chấp nhận giả thuyết độc lập giữa sai số ε và biến dự đoán Y. Mặt khác ta thấy khoảng rộng của độ lệch gần như là như nhau tại mọi phần của đồ thị nên ta chấp nhận giả thuyết phương sai của sai số ε là không đổi. Như vậy, bộ số liệu thỏa mãn A(1-3).

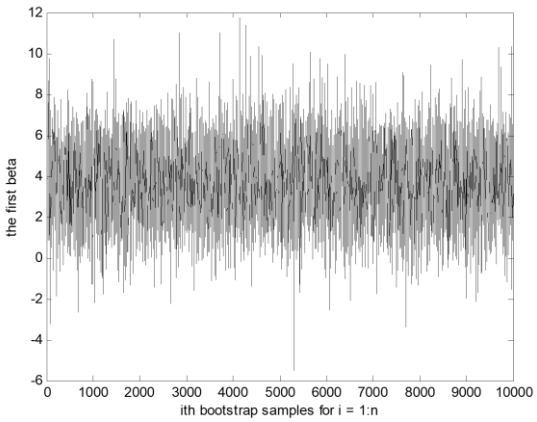


Hình 3*. Biểu đồ mô phỏng phân phối của các sai số $\hat{\varepsilon}_i$*

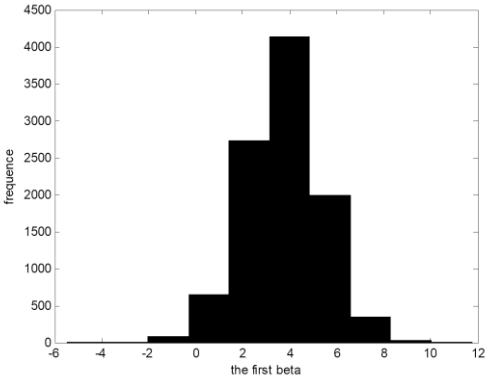
Hình 3 cho thấy sai số ε không có phân phối chuẩn và ta cũng chưa biết dạng phân phối của các sai số ε. Như vậy với số liệu này ta không thể sử dụng các phương pháp xác định hệ số hồi quy truyền thống.

Trong bài báo này tác giả sử dụng phần mềm Matlab để phân tích số liệu. Sau đây là thuật toán tìm khoảng tin cậy 95% của tham số hồi quy β bằng cách sử dụng quá trình bootstrap với cỡ mẫu lấy lại là biến ngẫu nhiên có phân phối đều trên $[n/4; n]$.

```
>>[n p]=size(X); # Xác định cỡ ma trận X
>>hatbeta=inv(X'*X)*X'*Y # Ước lượng hợp lý cực đại của β.
hatbeta = [3.7457  4.0935  2.9579]^T
>>hatepsilon=Y-X*hatbeta; # Vectơ ε̂(n)
>>data=hatepsilon-(sum(hatepsilon)/n)*ones(n,1);# Ma trận phần dư trung tâm đóng vai trò là mẫu gốc để lấy lại mẫu.
>>betaB=zeros(p,10000);
>> r=randi(n,1,10000); # Dãy 10000 số nguyên dương ngẫu nhiên có giá trị 1 đến n.
>>forI =1:10000rs=r(1,i);
Whilers<(n/4)rs=randi(n,1,1);end
Es=zeros(rs,1); # Ma trận phần dư bootstrap
Xs=zeros(rs,p);# Ma trận X gồm các hàng tương ứng với các ε̂_i lấy lại từ mẫu gốc.
while det(Xs'*Xs)=0 rb=randi(n,1,rs);
for j=1:rs
k=rb(1,j);Es(j,1)=data(k,1);Xs(j,:)=X(k,:);
end
Ys=Xs*hatbeta+Es;
betaB(:,i)=inv(Xs'*Xs)*Xs'*Ys; # Hệ số β bootstrap tương ứng mẫu lấy lại thứ i.
end
```

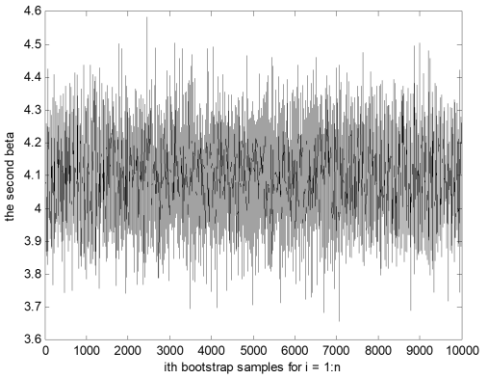


Hình 4. *Đồ thị các hệ số β_1 bootstrap với cỡ mẫu ngẫu nhiên có phân phối đều trên $[n/4; n]$.*

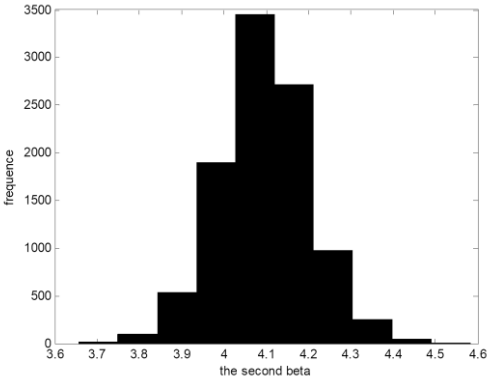


***Hình 5**. Biểu đồ mô phỏng phân phối của các hệ số β_1 bootstrap với cỡ mẫu ngẫu nhiên có phân phối đều trên* $[n/4; n]$.

Khoảng ước lượng bootstrap với cỡ mẫu ngẫu nhiên của hệ số β_1với độ tin cậy 95% là (0.4768;6.9116).

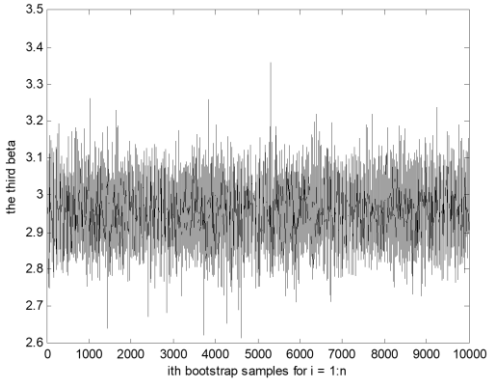


***Hình 6.** Đồ thị các hệ số β_2 bootstrap với cỡ mẫu ngẫu nhiên có phân phối đều trên* $[n/4; n]$.

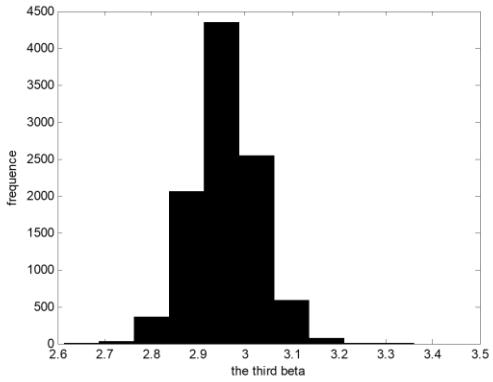


***Hình 7**. Biểu đồ mô phỏng phân phối của các hệ số β_2 bootstrap với cỡ mẫu ngẫu nhiên có phân phối đều trên* $[n/4; n]$.

Khoảng ước lượng bootstrap với cỡ mẫu ngẫu nhiên của hệ số β_2với độ tin cậy 95% là (3.8840; 4.3157).



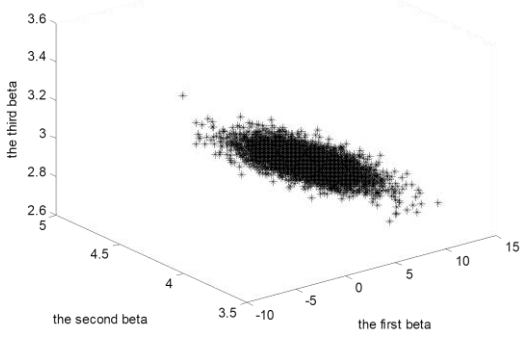
***Hình 8.** Đồ thị các hệ số β_3 bootstrap với cỡ mẫu ngẫu nhiên có phân phối đều trên* $[n/4; n]$.



***Hình 9**. Biểu đồ mô phỏng phân phối của các hệ số β_3 bootstrap với cỡ mẫu ngẫu nhiên có phân phối đều trên* $[n/4; n]$.

Khoảng ước lượng bootstrap với cỡ mẫu ngẫu nhiên của hệ số β_3với độ tin cậy 95% là (2.8223; 3.0974).

Ta bác bỏ các giả thuyết $\beta_1 = 0$; $\beta_2 = 0$; $\beta_3 = 0$; vì các khoảng ước lượng bootstrap với cỡ mẫu ngẫu nhiên của các hệ số này không chứa 0.



***Hình 10**. Đồ thị phân tán của các hệ số β_i bootstrap với cỡ mẫu ngẫu nhiên có phân phối đều trên* $[n/4; n]$.

Để so sánh kết quả giữa các phương pháp bootstrap, ta xác định hệ số hồi quy bootstrap trong trường hợp cỡ mẫu lấy lại cố định bằng cỡ mẫu gốc n; hoặc bằng $m = [0.625n] < n$ hay bằng $M = 2n$; và trường hợp cỡ mẫu bootstrap ngẫu nhiên sao cho mẫu lấy lại có đúng $m \approx n(1 - e^{-1}) \approx 0.632n$ phần tử phân biệt của mẫu gốc. Tổng hợp các kết quả từ quá trình phân tích số liệu ta có bảng 1, từ đó ta có cùng kết luận là các hệ số hồi quy bootstrap của mô hình này khác 0.

***Bảng 1**. Khoảng tin cậy 95% của các hệ hồi quy bootstrap.*

	Khoảng ước lượng bootstrap với độ tin cậy 95%	
Cỡ mẫu lấy lại N_n là biến ngẫu nhiên có phân phối đều trên $[n/4; n]$.	β_1	(0.4768;6.9116)
	β_2	(3.8840; 4.3157)
	β_3	(2.8223; 3.0974)
Cỡ mẫu lấy lại cố định bằng cỡ mẫu gốc n.	β_1	(1.3645; 6.0832)
	β_2	(3.9397; 4.2466)
	β_3	(2.8591; 3.0600)
Cỡ mẫu lấy lại cố định là m nhỏ hơn cỡ mẫu gốc n.	β_1	(0.7706; 6.7809)
	β_2	(3.9035; 4.2915)
	β_3	(2.8316; 3.0852)
Cỡ mẫu lấy lại cố định là $M = 2n$ lớn hơn cỡ mẫu gốc n.	β_1	(2.1015; 5.3848)
	β_2	(3.9857; 4.2048)
	β_3	(2.8890; 3.0285)
Cỡ mẫu lấy lại là ngẫu nhiên sao cho có đúng $m \approx n(1 - e^{-1})$ phần tử phân biệt của mẫu gốc.	β_1	(1.8359; 5.1854)
	β_2	(3.9903; 4.2081)
	β_3	(2.9003; 3.0438)

3. KẾT LUẬN

Quá trình phân tích thực nghiệm đã minh họa được cụ thể quá trình xác định khoảng tin cậy cho hệ số hồi quy cho mô hình hồi quy bootstrap với cỡ mẫu cố định và trường hợp cỡ mẫu lấy lại là ngẫu nhiên. Trong bài báo này, tác giả đã thực hiện được quá trình xác định hệ số hồi quy bootstrap thực nghiệm với cỡ mẫu lấy lại là biến ngẫu nhiên có phân phối đều $[n/4; n]$. Qua đó làm phong phú thêm các phương pháp xác định các hệ số hồi quy bootstrap.

Kết quả phân tích thực nghiệm cho thấy nếu cỡ mẫu lấy lại tăng thì độ dài của khoảng ước lượng giảm. Tuy nhiên, khi cỡ mẫu gốc ban đầu là n lớn nếu ta lấy cỡ mẫu lấy lại là bằng cỡ mẫu gốc hoặc bằng $M = 2n$ thì số lần lấy phần tử từ mẫu gốc khi lấy b mẫu bootstrap là nb hay $2nb$ sẽ rất lớn, làm tốn thời gian cho quá trình phân tích số liệu. Trường hợp cỡ mẫu lấy lại là biến ngẫu nhiên mà cụ thể là biến ngẫu nhiên có phân phối đều trên trên $[n/4; n]$ thì số lần lấy phần tử trung bình là $E(N_n)b = \frac{1}{2}\left(\frac{n}{4} + n\right)b = 0.625nb$ sẽ tiết kiệm thời gian hơn cho quá trình phân tích số liệu.

Trong [11] N.V. Toản đã chỉ ra tốc độ hội tụ của xấp xỉ bootstrap của phân phối trung bình mẫu với cỡ mẫu lấy lại là biến ngẫu nhiên N_n. Hướng nghiên cứu tiếp theo có thể thực hiện việc xác định tốc độ hội tụ của xấp xỉ bootstrap của phân phối ước lượng bình phương bé nhất trong mô hình hồi quy có cỡ mẫu bootstrap cố định hay là biến ngẫu nhiên N_n. Biến ngẫu nhiên N_n là số nguyên dương thuộc $[a, b]$ hoặc N_n là số lần lấy phần tử từ mẫu gốc cho đến khi xuất hiện $m \approx n(1 - e^{-1}) \approx 0.632n$ phần tử phân biệt trong mẫu gốc.

TÀI LIỆU THAM KHẢO

[1] Bradley Efron. *Bootstrap method: Another look at the Jacknife*. Ann. Statist. 7. (1979).
[2] E. Mammen. *Bootstrap, wild bootstrap, and asymptotic normality*. Probab. Theory Relat. Fields 93, 439–455 (1992).

[3] C. R. Rao, P.K. Pathak, and V. I. Koltchinskii. *Bootstrap by sequential resampling*. J. Statist. Plan. Inference 64.(1997).
[4] Toan, N.V. *On the asymptotic distribution of the bootstrap estimate with random resample size*. Vietnam J. Math. 33:3, 261–270 (2005).
[5] Toan, N.V. *Rate of convergence in bootstrap approximations with random sample size*. Acta Mathematica Vietnammica, 25. 161-179 (2000).
[6] Toan, N.V. *On Weak Convergence of the Bootstrap General Empirical Process with Random Resample Size*. Vietnam J. Math, 42, 233–245 (2014).
[7] D. A. Freedman. *Bootstrap regression models*. Ann. Statist. 9. (1981).
[8] Toan, N.V. *On bootstrapping regression and correlation models with random resample size*. Vietnam J. Math, 37, 443–456 (2009).
[9] Hữu, N.V and Dư, N.H. *Phân tích thống kê và Dự báo.* NXB ĐH Quốc Gia Hà Nội. (2003).
[10] E. Mammen. *When does bootstrap work*. Springer-Verlag New York, Inc. (1992).
[11] Toan, N.V. *On weak convergence of the bootstrap empirical process with random resample size*.Vietnam J. Math. 28:2, 153–158 (2000).

Tác giả chịu trách nhiệm bài viết:

Nguyễn Hồng Nhung
Trường Đại học Sư phạm Kỹ thuật Tp. HCM
Email: nhungnh@hcmute.edu.vn