Journal of Science and Transport Technology

University of Transport Technology

Convergence of Estimated Roughness by the Shift Residual Method

Ban Van To[1*], Tuan Anh Đo[2]

[1]University of Transport Technology, Hanoi 100000, Vietnam

[2]Military Technical Academy, Hanoi 100000, Vietnam

Article info

Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2023.vn.3.2.11-17

***Corresponding author:**
E-mail address:
bantv@utt.edu.vn

Received: 01/02/2023
Accepted: 22/05/2023
Published: 28/06/2023

Abstract: The roughness of a function to the given design is introduced. The conditions to ensure the convergence of the roughness of the model functions associated with designs to the roughness of the limit model function are indicated. Since then, the paper confirms the almost sure convergence of the roughness estimated by the method of shift residual to the roughness of the theoretical model function. The conditions to ensure convergence is quite extensive and easy to meet in real data analysis. Simulation studies show the appropriateness of the theoretical conclusions. The quality of the estimate is satisfactory even when the number of observations is relatively small, the roughness of the original model function is not very large, but the variance of the noise needs to be small. As the number of observations increases, the roughness of the original model may decrease, and the variance of the noise may not need to be very small, but the obtained estimate can still be quite satisfactory.

Keywords: Roughness, almost sure convergence, two-phase regression model, change-point.

Tạp chí điện tử
Khoa học và Công nghệ Giao thông
Trường Đại học Công nghệ GTVT

Sự hội tụ của độ ráp ước lượng theo phương pháp phần dư dịch chuyển

Tô Văn Ban[1*], Đỗ Anh Tuấn[2]
[1]Trường Đại học Công nghệ Giao thông Vận tải, Hà Nội 100000, Việt Nam
[2]Học viện Kỹ thuật Quân sự, Hà Nội 100000, Việt Nam

Thông tin bài viết
Dạng bài viết:
Bài báo nghiên cứu

DOI:
https://doi.org/10.58845/jstt.utt.2023.vn.3.2.11-17

***Tác giả liên hệ:**
Địa chỉ E-mail:
bantv@utt.edu.vn

Ngày nộp bài: 01/02/2023
Ngày chấp nhận: 22/05/2023
Ngày đăng bài: 28/06/2023

Tóm tắt: Trong bài báo, độ ráp của hàm số theo thiết kế cho trước được nêu ra. Bài báo chỉ rõ những điều kiện đảm bảo sự hội tụ của độ ráp của hàm mô hình liên kết với dãy thiết kế tới độ ráp của hàm mô hình giới hạn. Từ đó, sự hội tụ hầu chắc chắn của độ ráp ước lượng được theo phương pháp phần dư dịch chuyển tới độ ráp của hàm mô hình lý thuyết được khẳng định. Các điều kiện đảm bảo sự hội tụ là khá tổng quát và dễ thỏa mãn trong những phân tích dữ liệu thực. Nghiên cứu mô phỏng chỉ ra sự phù hợp của kết luận lý thuyết đưa ra. Chất lượng của ước lượng là thỏa đảng kể cả khi số quan sát khá nhỏ, độ ráp của hàm mô hình gốc không lớn lắm song phương sai của nhiễu cần phải bé. Khi số quan sát tăng lên, độ ráp của mô hình gốc có thể giảm đi, và phương sai của nhiễu có thể không cần nhỏ lắm nhưng ước lượng thu được vẫn có thể hoàn toàn thỏa đáng.
Từ khóa: Độ ráp, hội tụ hầu chắc chắn, mô hình hồi quy hai pha, điểm chuyển.

1. Giới thiệu

Nhiều quá trình chuyển động tuân theo mô hình tuyến tính hai pha, ở đó các tham số điều khiển mô hình giữ nguyên giá trị trong pha đầu, tại một thời điểm nào đó nó chuyển sang giá trị khác và giữ nguyên trong pha còn lại. Việc nghiên cứu mô hình có thay đổi trạng thái như vậy - còn gọi là mô hình điểm chuyển - đã được phát triển hơn nửa thế kỷ qua và đạt được những thành tựu rực rỡ, được áp dụng rộng rãi trong nhiều lĩnh vực khác nhau. Trong kinh tế, người ta thấy mô hình điểm chuyển bội là phù hợp khi nghiên cứu mối quan hệ giữa lãi suất (interest rate) đối với thay đổi lãi suất chiết khấu (discount rate) quy định bởi FED. Sử dụng mô hình ARCH để nghiên cứu chuỗi thời gian trong miền tần số, người ta đã phát hiện ra sự chuyển đổi của chuỗi thời gian chỉ số chứng khoán, cũng như thị trường ngoại hối liên hệ mật thiết với khủng hoảng tài chính ở châu Á và Liên Xô. Theo Caussinus H., Lyazrhi F., trong giai đoạn nghiên cứu, tổng sản phẩm quốc nội Hoa Kỳ tuân theo mô hình điểm chuyển bội. Mô hình điểm chuyển được áp dụng thành công trong nghiên cứu sự sinh sôi của loài tảo cát liên quan đến ô nhiễm môi trường, trong nghiên cứu địa chấn, nhờ đó đã phân biệt được tín hiệu địa chấn do động đất và tín hiệu địa chấn từ vụ nổ bom nguyên tử. Có thể liệt kê ra hàng loạt áp dụng của mô hình điểm chuyển trong hàng không vũ trụ, biến đổi khí hậu, chế độ thủy văn, lượng mưa, dự báo, tấn công mạng máy tính, nghiên cứu thể thao… Việc nghiên cứu mô hình điểm chuyển là cần thiết và liên tục được phát triển trong những năm gần đây.

Xét mô hình

$$y_i = \begin{cases} \alpha_0 + \alpha_1 x_i + \varepsilon_i & \text{khi } 1 \le i \le k^* \\ \beta_0 + \beta_1 x_i + \varepsilon_i & \text{khi } k^* \le i \le n, \end{cases} \quad (1)$$

trong đó $a \le x_1 < \ldots < x_n \le b$, a, b cố định cho trước, các sai số $\{\varepsilon_i\}$ là ngẫu nhiên, $\alpha_0, \alpha_1, \beta_0, \beta_1, k^*$ cố định chưa biết.

Nếu $\alpha_0 = \beta_0$ và $\alpha_1 = \beta_1$ thì mô hình (1) gọi là

không có chuyển. Ngược lại nếu ít nhất một trong hai đẳng thức này không xảy ra, mô hình được gọi là có chuyển và *k** được gọi là thời điểm chuyển. Đối với mô hình có chuyển, $\alpha_1 \neq \beta_1$, và hai đường thẳng $y = \alpha_0 + \alpha_1 x$ và $y = \beta_0 + \beta_1 x$ cắt nhau tại điểm τ trên nửa khoảng $[x_{k*}, x_{k*+1})$ thì hàm mô hình được gọi là gãy khúc liên tục, mô hình được gọi là liên tục. Trái lại, mô hình được gọi là gián đoạn. Ở đây, chúng ta chỉ xét trường hợp mô hình liên tục. Đặt $h = \beta_1 - \alpha_1$, Mô hình (1) được viết lại dưới dạng $y_i = f(x_i) + \varepsilon_i,\ \ i = 1, \dots, n$, trong đó

$$f(x) = \alpha_0 + \alpha_1 x + h(x - \tau)I(x > \tau) \quad (2)$$

là hàm mô hình và $I(.)$ là hàm chỉ tiêu.

Có nhiều phương pháp để phát hiện sự tồn tại thời điểm chuyển (xem [1], [2], [3],…). Giả sử chúng ta biết rằng thời điểm chuyển tồn tại, cần ước lượng (ƯL) nó. Hãy chia quan sát thành hai nhóm. Nhóm thứ nhất chứa k quan sát đầu $(x_i, y_i), i = 1, \dots, k$ và giả sử $\hat{\alpha}_{0k}, \hat{\alpha}_{1k}$ là ƯL bình phương cực tiểu cho hệ số chặn và hệ số góc của mô hình tuyến tính đơn tương ứng. Nhóm thứ hai chứa $n - k$ quan sát còn lại $(x_i, y_i), i = k + 1, \dots, n$ và giả sử $\hat{\beta}_{0k}, \hat{\beta}_{1k}$ là ƯL bình phương cực tiểu cho hệ số chặn và hệ số góc tương ứng. Yêu cầu tự nhiên là điểm chuyển không được quá gần quan sát đầu cũng như quan sát cuối, vậy ta cần có $k_0 \leq k \leq n - k_0$ với k_0 đủ lớn. Theo [4], [5], xét phần dư dịch chuyển

$$\tilde{e}_{ik} = \begin{cases} y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) & \text{khi} \quad 1 \leq i \leq k, \\ y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 x_i) & \text{khi} \quad k+1 \leq i \leq n. \end{cases} \quad (3)$$

Lưu ý rằng các phần dư dịch chuyển $\tilde{e}_{ik}$ không là phần dư thông thường: Khi tính phần dư cho nhóm quan sát đầu (pha đầu), chúng ta dùng ước lượng tham số của nhóm quan sát sau (pha sau) và ngược lại. Ưu điểm của các phần dư dịch chuyển là chúng gần với phần dư thông thường dưới giả thuyết (khi $(\alpha_0, \alpha_1) = (\beta_0, \beta_1)$), nhưng chúng được phóng đại lên dưới đối thuyết $(\alpha_0, \alpha_1) \neq (\beta_0, \beta_1)$.

2. Sự hội tụ của độ ráp ước lượng

Trước hết chúng ta cần đến định lý sau đã đưa ra ở [6].

Định lý 1. Giả sử xảy ra các giả thiết sau đây:

i) Thiết kế x_i trải đều trên đoạn $[a, b] = [0,1]$, nghĩa là $x_i = i/n, i = 1, \dots, n$.

ii) Hàm mô hình có thể viết dưới dạng

$$f_n(x) = \alpha_0 + \alpha_1 x + h(x - x_{k^*})I(x > x_{k^*}), h \neq 0 \quad (4)$$

iii) Tồn tại $\tau_0 \in (0,1/2)$ sao cho $k_0 < k^* < n - k_0$, trong đó $k_0 = \lfloor n\tau_0 \rfloor + 1$ và $\lfloor l \rfloor$ ký hiệu phần nguyên của số thực l.

iv) $x_{k^*} \to \tau$ khi $n \to \infty$.

v) Các sai số $\{\varepsilon_i\}$ là các biến ngẫu nhiên độc lập, có kỳ vọng không, $E(\varepsilon_i^2) = \sigma_1^2 > 0$ với $i = 1, \dots, k^*, E(\varepsilon_i^2) = \sigma_2^2 > 0$ với $i = k^* + 1, \dots, n$, σ_1^2, σ_2^2 chưa biết.

$$\text{Đặt} \quad \hat{k}_n = \underset{k_0 \leq k \leq n - k_0}{\operatorname{argmax}} \sum_{i=1}^{n} \tilde{e}_{ik}^2$$

Khi đó, $\lim_{n\to\infty} \frac{\hat{k}_n}{n} = \tau$ hầu chắc chắn (h.c.c.)

Hơn nữa,

$$\lim_{n\to\infty} \alpha_{i\hat{k}_n} = \alpha_i, \qquad i = 1,2,$$

$$\lim_{n\to\infty} \beta_{0\hat{k}_n} = \alpha_0 - h\tau, \lim_{n\to\infty} \beta_{1\hat{k}_n} = \alpha_1 + h\tau \ \ (h.c.c)$$

Các giả thiết ở Định lý trên là khá tổng quát và dễ đáp ứng được trong những điều kiện thực tế. Giả thiết (ii) đảm bảo rằng, mô hình là gãy khúc liên tục. Theo giả thiết (iii) ta chỉ cần xét thời điểm chuyển từ k_0 đến $n - k_0$. Giả thiết này đảm bảo sự hội tụ của các tham số ƯL được. Theo [4], chọn k_0 sao cho $k_0 = C_0 n + O(1), C_0 \in (0,0.5)$. Giả thiết (ii) có nghĩa rằng x_{k*} là điểm chuyển của mô hình có *n* quan sát. Tử (iii) rõ ràng rằng $k_0 \to \infty$ khi và chỉ khi $n \to \infty$. Giả thiết (v) rất tổng quát, ở đó các phương sai ở hai pha σ_1^2 và σ_2^2 nói chung khác nhau. Kết luận ở Định lý 1 khẳng định điểm chuyển ước lượng được $\hat{k}_n/n$ sẽ hội tụ hầu chắc chắn, loại hội tụ rất mạnh của lý thuyết xác suất, đến điểm chuyển thực τ.

Hàm *f(x)* càng gồ ghề, càng lệch nhiều so với đường thẳng thì khả năng phát hiện ra điểm chuyển càng lớn. Khái niệm độ ráp được đưa ra để đo mức độ gồ ghề của hàm mô hình.

Định nghĩa 1. Độ ráp của hàm *f(x)* dựa vào thiết kế $\{x_1, \dots, x_n\}$ được ký hiệu bởi $S^2(f, \{x_i\}_n)$ và xác định theo công thức (5)

$$S^2(f,\{x_i\}_n)=\frac{1}{n}\sum_{1}^{n}\big(f(x_i)-(\hat{a}+\hat{b}\,x_i)\big)^2$$

trong đó $\hat{a},\hat{b}$ là ƯL bình phương cực tiểu của hệ số góc và hệ số chặn tương ứng của mô hình hồi quy tuyến tính đơn với tập dữ liệu $(x_i,f(x_i)),i=1,\dots,n.$

Chú ý rằng $S^2(f,\{x_i\}_n)$ là ƯL cho phương sai chung chưa hiệu chỉnh của mô hình tuyến tính thông thường. Tuy nhiên, các dữ liệu $\{(x_i,f(x_i))\}$ không ngẫu nhiên nên ta không nên gọi đây là phương sai chung. Độ ráp là một đặc trưng hình học hay sử dụng trong cơ học, thể hiện mức độ không thẳng, gồ ghề của đường cong $y=f(x)$ khi tiến hành quan sát tại các điểm x_i.

Khi chuyển sang trường hợp có vô hạn điểm thiết kế, ta coi mỗi hàm phân bố *F(x)* có giá $J\subset[a,b]$ chứa ít nhất hai điểm là một thiết kế suy rộng trên *J.* Độ đo xác suất ứng với hàm phân bố *F(x)* ký hiệu là *(dF)*.

Định nghĩa 2. Độ ráp của hàm *f(x)* dựa vào thiết kế *F(x)* có giá trên *J* được ký hiệu bởi $S^2(f,F)$ và xác định theo công thức (6)

$$S^2(f,F)=\min_{(a,b)\in\mathbb{R}^2}\int_J(f(x)-(a+bx))^2dF(x)$$

Đặt

$$z_1(x)=1,\ z_2(x)=x,$$

$$<z,z^T>_F=\big[<z_i,z_j>_F\big]=\begin{bmatrix}<z_1,z_1>_F & <z_1,z_2>_F\\ <z_2,z_1>_F & <z_2,z_2>_F\end{bmatrix},$$

$$<z,f>_F=\begin{bmatrix}<z_1,f>_F\\ <z_2,f>_F\end{bmatrix},\quad(7)$$

trong đó $\langle k,\ell\rangle_F=\int_J k(x)\ell(x)dF(x).$

Chúng ta chỉ xét những thiết kế mà ma trận $<z,z^T>_F$ khả nghịch. Theo phương pháp bình phương cực tiểu, cực tiểu ở (6) tồn tại và đạt được tại

$$(\hat{a},\hat{b})^T=(<z,z^T>_F)^{-1}<z,f>_F\quad(8)$$

Mỗi thiết kế rời rạc $\{x_i,i=1,\dots,n\}$ có ít nhất hai điểm phân biệt là một thiết kế suy rộng $F_{x_1,\dots,x_n}(x)$ là hàm phân bố mẫu của mẫu quan sát $x_1,\dots,x_n$. Dễ thấy rằng (5) là trường hợp đặc biệt của (6). Người ta cũng đưa ra khái niệm độ ráp dựa vào họ đường cong tổng quát hơn như họ đường bậc hai, bậc ba,… Các tính chất của độ ráp có thể tham khảo ở [7].

Giả sử đối với mô hình (1), chúng ta tìm được ước lượng cho thời điểm chuyển là $\hat{k}_n$ và ước lượng tương ứng cho tham số ở pha đầu và pha sau lần lượt là $\hat{\alpha}_{0\hat{k}_n},\hat{\alpha}_{1\hat{k}_n}$ và $\hat{\beta}_{0\hat{k}_n},\hat{\beta}_{1\hat{k}_n}$. Hỏi rằng độ ráp của hàm mô hình ƯL được

$$\hat{f}_n(x)=\begin{cases}\hat{\alpha}_{0\hat{k}_n}+\hat{\alpha}_{1\hat{k}_n}x, & 0\le x\le\hat{k}_n/n,\\ \hat{\beta}_{0\hat{k}_n}+\hat{\beta}_{1\hat{k}_n}x, & \hat{k}_n/n<x<1\end{cases}\quad(9)$$

có hội tụ về độ ráp của hàm *f(x)* xác định bởi (2) hay không? Nếu điều này được khẳng định thì với *n* đủ lớn, độ ráp $S^2(\hat{f}_n,\{x_i\}_n)$ sẽ xấp xỉ độ ráp $S^2(f,x)$, và do đó, nếu $S^2(\hat{f}_n,\{x_i\}_n)$ là lớn, ta có thể tin tưởng những kết luận thống kê đã đưa ra. Trái lại, nếu $S^2(\hat{f}_n,\{x_i\}_n)$ tương đối nhỏ, các kết luận về giá trị của các tham số $\hat{k}_n,\hat{\alpha}_{0\hat{k}_n},\hat{\alpha}_{1\hat{k}_n},\hat{\beta}_{0\hat{k}_n},\hat{\beta}_{1\hat{k}_n}$ có độ tin tưởng thấp.

Câu trả lời là khẳng định. Trước hết ta đưa ra định lý sau đây.

Định lý 2. Giả sử xảy ra các điều kiện sau đây:

1) Dãy thiết kế $F_n(x)$ hội tụ yếu đến thiết kế $F(x)$: $F_n\Rightarrow F$.

2) $g_n(x),g(x)$ là những hàm đo được, bị chặn đều trên [0,1]:

Tồn tại $M>0$ để $|g(x)|,|g_n(x)|<M\ \forall x\in J,\forall n.$

3) $(dF)(E_g)=0$, trong đó *(dF)* là độ đo xác suất ứng với hàm phân bố *F(x)*, $E_g=\{t\in J:\exists\{t_n\}\subset\mathbb{R},\ t_n\to t,\ g_n(t_n)\not\to g(t)\}$.

Khi đó $S^2(g_n,F_n)\to S^2(g,F)$.

Chứng minh. Các hàm $z_i(x)$ liên tục và bị chặn, $F_n\Rightarrow F$, vậy

$$<z_i,z_j>_{F_n}=\int_0^1 z_i(x)z_j(x)dF_n(x)\to\int_0^1 z_i(x)z_j(x)dF(x)\quad(i,j=1,2).$$

Hơn nữa, các ma trận $<z,z^T>_{F_n},<z,z^T>_F$ khả nghịch, vậy

$\det(< z, z^T >_{F_n}) \to \det(< z, z^T >_F) \neq 0.$

Từ đó mỗi dãy các phần tử của ma trận $(< z, z^T >_{F_n})^{-1}$ hội tụ đến phần tử tương ứng của ma trận $< z, z^T >_F$.

Rõ ràng các hàm $z_i(x)g_n(x)$, $z_i(x)g(x)$ là đo được, bị chặn; $E_{z_i \times g} \subset E_g, (dF)(E_{z_i \times g}) \leq (dF)(E_g) = 0$ từ điều kiện (3). Theo Định lý 5.5 trong [8] thì

$$< z_i, g_n >_{F_n} = \int_0^1 z_i(x)\, g_n(x)\, dF_n(x) \to \int_0^1 z_i(x)\, g(x)\, dF(x).$$

Suy ra

$$(\hat{a}_n, \hat{b}_n)^T = (< z, z^T >_{F_n})^{-1} < z, g_n >_{F_n} \to (< z, z^T >_F)^{-1} < z, g >_F = (\hat{a}, \hat{b})^T. \quad (10)$$

Từ chỗ

$$S^2(g_n, F_n) = \int_0^1 \left(g_n(x) - (\hat{a}_n + \hat{b}_n x)\right)^2 dF_n(x),$$

khai triển vế phải thành tổng, sử dụng (10) và lập luận tương tự như trên ta được

$$S^2(g_n, F_n) \to \int_0^1 \left(g(x) - (\hat{a} + \hat{b}x)\right)^2 dF(x).$$

Lưu ý rằng giới hạn nhận được chính là

$$S^2(g, F) = \min_{(a,b)\in\mathbb{R}^2} \int_J (f(x) - (a + bx))^2 dF(x).$$

Hệ quả 3. Giả sử hàm phân bố mẫu $F_{x_1,\dots,x_n}(x)$ của mẫu $\{x_i\}_n$ hội tụ yếu đến *F(x)* là thiết kế trên $[0,1]$, hàm *F(x)* liên tục tại $\upsilon \in (0,1)$. Giả sử *các hàm mô hình* $h(x), h_n(x)$ cho bởi

$$h_n(x) = \begin{cases} a_{0n} + a_{1n}x, 0 \leq x \leq \upsilon_n, \\ b_{0n} + b_{1n}x, \upsilon_n < x \leq 1 \end{cases},$$

$$h(x) = \begin{cases} a_0 + a_1 x, 0 \leq x \leq \upsilon, \\ b_0 + b_1 x, \upsilon < x \leq 1. \end{cases}$$

sao cho

$$\lim_{n\to\infty} a_{in} = a_i, \quad \lim_{n\to\infty} b_{in} = b_i, \quad i = 0,1, \quad \lim_{n\to\infty} \upsilon_n = \upsilon.$$

Khi đó $\lim_{n\to\infty} S^2(h_n, \{x_i\}_n) = S^2(h, F).$

Chứng minh. Rõ ràng các hàm $h_n(x), h(x)$ là các hàm đo được, bị chặn đều trên $[0,1]$, $\lim_{n\to\infty} h_n(x) = h(x), \forall x \neq \upsilon, x \in [0,1]$, vậy $E_h \subset \{\upsilon\}$. Theo giả thiết, điểm gián đoạn duy nhất có thể của *h(x)* là ν, từ đó $(dF)(E_h) \leq (dF)\{\nu\} = 0$. Áp dụng Định lý 2 ta được

$$\lim_{n\to\infty} S^2\left(g_n, \{x_i\}_n\right) = \lim_{n\to\infty} S^2\left(g_n, F_{x_1,\dots,x_n}\right) = S^2(g, F)$$

Định lý 4. Giả sử các giả thiết ở Định lý 1 được thỏa mãn. Khi đó

$$\lim_{n\to\infty} S^2(\hat{f}_n, \{x_i\}_n) = S^2(f, U) \qquad (h.c.c)$$

trong đó f(x) xác định theo (2), $\hat{f}_n$ theo (9) và *U(x)* là hàm phân bố đều trên [0,1].

Chứng minh. Trước hết ta thấy $F_{x_1,\dots,x_n} \Rightarrow U$. Các hàm $\hat{f}_n(x)$ và *f(x)* đo được, bị chặn đều h.c.c., *f(x)* liên tục. Đặt $\hat{\tau}_n = \hat{k}_n/n$, Theo Định lý 1,

$$\lim_{n\to\infty} \hat{\tau}_n = \tau, \quad \lim_{n\to\infty} \alpha_{i\hat{k}_n} = \alpha_i, \quad i = 1,2,$$

$$\lim_{n\to\infty} \hat{\beta}_{0\hat{k}_n} = \alpha_0 - h\tau = \beta_0, \quad \lim_{n\to\infty} \hat{\beta}_{1\hat{k}_n} = \alpha_1 + h = \beta_1$$

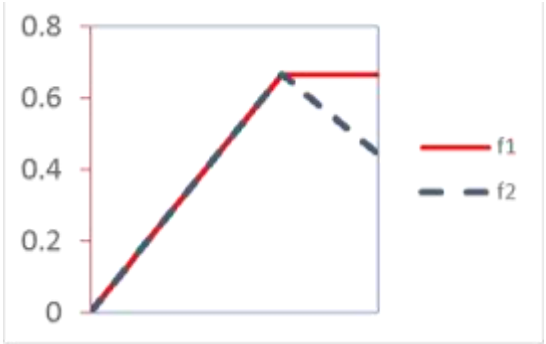
(giới hạn h.c.c.). Vì *U(x)* liên tục trên (0,1), áp dụng Hệ quả 3 ta nhận được kết quả cần chứng minh.

3. Nghiên cứu mô phỏng

Xét hai hàm mô hình

$$f_1(x) = \begin{cases} x, & 0 \leq x \leq \frac{2}{3} \\ \frac{2}{3}, & \frac{2}{3} < x \leq 1, \end{cases}$$

$$f_2(x) = \begin{cases} x, & 0 \leq x \leq \frac{2}{3} \\ -\frac{2}{3}x + \frac{10}{9}, & \frac{2}{3} < x \leq 1. \end{cases}$$



Hình 1. Đồ thị các hàm mô hình f_1 và f_2

Trước hết cần tính $S^2(f_i, U) = \min_{a,b} \int_0^1 (f_i(x) - (a + bx))^2 dx$. Sử dụng kết quả ở [9, tr 65]:

$$S^2(f_i, U) = det \begin{pmatrix} \int_0^1 dx & \int_0^1 xdx & \int_0^1 f_i(x)dx \\ \int_0^1 xdx & \int_0^1 x^2dx & \int_0^1 xf_i(x)dx \\ \int_0^1 f_i(x)dx & \int_0^1 xf_i(x)dx & \int_0^1 f_i^2(x)dx \end{pmatrix}$$

$$\times\ det^{-1}\begin{pmatrix}\int_0^1 dx & \int_0^1 xdx\\ \int_0^1 xdx & \int_0^1 x^2dx\end{pmatrix},$$

chúng ta nhận được $S(f_1,U)=0.060767$, $S(f_2,U)=0.101279$.

Quá trình mô phỏng bao gồm $N=2000$ lần lặp, mỗi lần lặp độc lập gồm một số bước.

Bước 1: Sinh quan sát mô phỏng. Ở mỗi bước lặp độc lập, đối với kích thước mẫu n và với $k^*=\lfloor 2n/3\rfloor$, sinh k^* biến ngẫu nhiên độc lập $\varepsilon_1,\dots,\varepsilon_{k^*}$ từ phân bố chuẩn $N(0,\sigma_1^2)$, và $n-k^*$ biến ngẫu nhiên độc lập $\varepsilon_{k^*+1},\dots,\varepsilon_n$ từ phân bố chuẩn $N(0,\sigma_2^2)$.

Đối với mỗi hàm mô hình $f(x)=f_i(x)$, lập mẫu mô phỏng (x_i,y_i) với

$$x_i=\frac{i}{n},\quad y_i=f\left(\frac{i}{n}\right)+\varepsilon_i,\qquad i=1,\dots,n.$$

Bước 2: Tìm ƯL của điểm chuyển

+ Với mỗi $k\in\left\{\left\lfloor\frac{n}{6}\right\rfloor,\left\lfloor\frac{n}{6}\right\rfloor+1,\dots,\left\lfloor\frac{5n}{6}\right\rfloor\right\}=K$, gọi:

$\hat\alpha_{0k},\hat\alpha_{1k}$ là ƯL bình phương cực tiểu của hệ số chặn và hệ số góc của mô hình tuyến tính đơn đối với k quan sát đầu tiên (pha đầu): $(1/n,y_1),\dots,(k/n,y_k)$.

$\hat\beta_{0k},\hat\beta_{1k}$ là ƯL bình phương cực tiểu của hệ số chặn và hệ số góc của mô hình tuyến tính đơn đối với $n-k$ quan sát cuối cùng (pha sau):

$$\left(\frac{k+1}{n},y_{k+1}\right),\dots,\left(\frac{n}{n},y_n\right).$$

+ Lập các phần dư dịch chuyển $\hat e_{ik},i=1,\dots,n$:

$$\tilde e_{ik}=\begin{cases}y_i-\left(\hat\beta_{0k}+\hat\beta_{1k}\frac{i}{n}\right) & \text{với}\quad 1\le i\le k\\ y_i-\left(\hat\alpha_{0k}+\hat\alpha_{1k}\frac{i}{n}\right) & \text{với}\ \ k+1\le i\le n.\end{cases}$$

+ Tính sai số bình phương trung bình của các phần dư dịch chuyển:

$$s_{nk}^2=\frac{1}{n-2}\sum_{i=1}^{n}\hat e_{ik}^2.$$

+ Tìm k để sai số bình phương trung bình các phần dư dịch chuyển lớn nhất:

$$\hat k_n=\arg\max_{k\in K}s_{nk}^2.$$

Bước 3. Tìm độ nháp của hàm mô hình ước lượng

+ Lập hàm mô hình ƯL

$$f_n(x)=\begin{cases}\hat\alpha_{0\hat k_n}+\hat\alpha_{1\hat k_n}x, & 0\le x\le \hat k_n/n\\ \hat\beta_{0\hat k_n}+\hat\beta_{1\hat k_n}x, & \hat k_n/n<x\le 1.\end{cases}$$

+ Lập các dự báo của hàm mô hình ƯL:

$$\hat f_{ni}=\begin{cases}\hat\alpha_{0\hat k_n}+\hat\alpha_{1\hat k_n}\frac{i}{n} & \text{với}\quad 1\le i\le\lfloor\hat k_n/n\rfloor\\ \hat\beta_{0\hat k_n}+\hat\beta_{1\hat k_n}\frac{i}{n} & \text{với}\quad \lfloor\hat k_n/n\rfloor<i\le n.\end{cases}$$

+ Lọc bằng mô hình hồi quy tuyến tính đơn đối với dữ liệu $((i/n),\hat f_{ni},),i=1,\dots,n$, tìm sai số chuẩn (Standard error) $SE_n=\sqrt{(1/(n-2))\sum_{i=1}^n\varepsilon_i^2}$.

Bảng 1. Giá trị trung bình $\overline{SE}_n$ và độ lệch chuẩn s_{SE_n} của độ ráp của hàm mô hình ước lượng theo mô phỏng

$Sigma$	n	$f=f_1$			$f=f_2$		
		$S(f,U)$	$\overline{SE}_n$	s_{SE_n}	$S(f,U)$	$\overline{SE}_n$	s_{SE_n}
$(1,0.5)$	20	0.060767	0.349127	0.139681	0.101279	0.355916	0.144979
	50	0.060767	0.232269	0.089166	0.101279	0.238314	0.087734
	100	0.060767	0.170687	0.061752	0.101279	0.177822	0.063905
$(0.4,0.2)$	20	0.060767	0.144138	0.059700	0.101279	0.099449	0.036686
	50	0.060767	0.100117	0.036995	0.101279	0.087827	0.026550
	100	0.060767	0.075539	0.025693	0.101279	0.086222	0.022090
$(0.2,0.1)$	20	0.060767	0.080605	0.032294	0.101279	0.154684	0.062447
	50	0.060767	0.060830	0.020777	0.101279	0.111987	0.039481
	100	0.060767	0.054021	0.016421	0.101279	0.094985	0.030688

Tổng hợp kết quả. Tính trung bình mẫu $\overline{SE_n}$ và độ lệch chuẩn mẫu s_{SE_n} của 2000 giá trị SE_n nhận được. Đưa kết quả tổng hợp vào Bảng 1.

Bảng 1 chỉ ra giá trị trung bình $\overline{SE_n}$ (cột 4, 7) và độ lệch chuẩn mẫu s_{SE_n} (cột 5, 8) tính toán dựa trên các hàm mô hình f_1 và f_2 trong các trường hợp $n = 20, 50, 100$ và khi $(\sigma_1, \sigma_2) = (1, 0.5)$, $(0.4, 0.2)$, $(0.2, 0.1)$. Để tiện theo dõi, cột 3 và cột 6 được đưa thêm để chỉ căn bậc hai của độ ráp của mô hình gốc tương ứng.

Chất lượng của ước lượng phụ thuộc vào độ lớn của độ chệch $|S(f,U) - \overline{SE_n}|$ và s_{SE_n}. Các giá trị này càng nhỏ, chất lượng ƯL càng cao; trái lại, các giá trị này càng lớn, chất lượng của ƯL càng giảm. Các kết quả ở Bảng 1 chỉ ra rằng, những kết quả phát hiện ra ở phân tích lý thuyết nêu trên là thỏa đáng. Khi kích thước mẫu n lớn, độ ráp của hàm mô hình gốc lớn và phương sai của nhiễu khá nhỏ, chúng ta sẽ có những ƯL tốt nhất. Khi số quan sát còn tương đối nhỏ $(n \approx 20)$, và sai số nhiễu không lớn, $Max(\sigma_1, \sigma_2) \leq 0.2$, chúng ta vẫn có ƯL tốt kể cả hàm mô hình gốc không ráp lắm $(S(f,U) \approx 0.07)$. Khi sai số nhiễu lớn lên, $Max(\sigma_1, \sigma_2) \approx 0.4$, cần có kích thước mẫu n lớn hơn và (hoặc) độ ráp của hàm mô hình gốc lớn hơn. Nói chung, khi kích thước mẫu n tăng lên, khi độ lệch tiêu chuẩn σ_1, σ_2 giảm đi, và khi độ ráp của hàm mô hình gốc $S(f,U)$ tăng lên thì chất lượng của ƯL sẽ tăng lên.

4. Kết luận

Với một số giả thiết dễ dàng thực hiện trong những điều kiện của thực tế, độ ráp ước lượng được của hàm mô hình khi dùng phương pháp phần dư dịch chuyển hội tụ hầu chắc chắn tới độ ráp của mô hình thực. Từ đó, khi *n* đủ lớn, độ ráp ước lượng được cho ta thông tin hữu ích: Nếu độ ráp lớn, ta có cơ sở để tin tưởng các kết luận đưa ra; nếu độ ráp nhỏ, các kết luận thu được có độ tin tưởng thấp. Nghiên cứu mô phỏng chỉ ra sự phù hợp của các kết luận đưa ra.

Tài liệu tham khảo

[1] Chen C. W. S., Chan J. S. K., Gerlach R., Hsieh W. Y. L., *A comparison of estimators for regression models with change points,* Stat. Comput., 21, pp. 395-414 (2011).

[2] Kirch C., *Bootstrapping sequential change-point tests,* Sequential Anal., 27, pp. 330-349 (2008).

[3] Nosek K., *Schwarz information criterion based testsfor a changepoint in regression models,* Stat. Papers, 51, pp. 915-929 (2010).

[4] Liu Z., Qian L., *Changepoint estimation in a segmented linear regression via empirical likelihood,* Communications in Statistics-Simulation and Computation, 39, pp. 85-100 (2009).

[5] Zhao H., Chen H., Wu X., *Changepoint analysis by modified empirical likehood method in two-phase linear regression models,* Opend Journal of Applied Sciences, 3, pp. 1-6 (2013).

[6] V.B.To, T.Q.Nguyen, *Estimating a change-point in two-phases regression model based on the shift of parameter estimates,* Theoretical Mathematics and Applications, vol.6, no.4, pp. 33-52 (2016).

[7] V.B.To, N.T.Nguyen, T.H.Phan, *The roughness of model function to the basis function,* Journal of Mathematics and System Sciences, 3, pp. 385-390 (2013).

[8] Billingsley P., *Convergence of probability measures,* John Wiley (1968).

[9] K.A.Pham. *Giải tích số.* Nxb Đại học Quốc gia Hà Nội (1998).