ỨNG DỤNG ĐỘ RÁP CỦA HÀM MÔ HÌNH TÌM THAM SỐ PHI TRUNG TÂM CỦA KIỂM ĐỊNH F CHO HỒI QUY TUYẾN TÍNH HAI PHA

Nguyễn Thị Quyên
Khoa Toán và KHTN, Trường Đại học Hải Phòng
Email: quyennt@dhhp.edu.vn

Ngày nhận bài: 11/12/2023
Ngày PB đánh giá: 27/02/2024
Ngày duyệt đăng: 26/3/2024

TÓM TẮT: Trong bài báo này, thông qua *độ ráp của hàm mô hình theo hệ hàm cơ sở cho trước,* chúng tôi đưa ra công thức xác định tham số phi trung tâm của thống kê kiểm định F cho mô hình hồi quy tuyến tính hai pha. Tham số này phụ thuộc vào kích thước mẫu, và tăng theo tỉ số giữa độ ráp và sai số của mô hình. Các kết quả này được sử dụng để tính xác suất mắc sai lầm loại II khi tiến hành kiểm định điểm chuyển cho mô hình hồi quy tuyến tính hai pha. Hơn nữa, dựa vào một số kết quả về sự hội tụ cũng như tốc độ hội tụ của độ ráp theo sự hội tụ yếu của dãy thiết kế, bài báo cũng chỉ ra một số hệ quả và giới hạn của xác suất mắc sai lầm loại II.

Từ khóa: độ ráp, sai lầm loại II, hàm mô hình, điểm chuyển.

APPLICATION OF THE ROUGHNESS OF THE MODEL FUNCTION FINDING NON-CENTRAL PARAMETERS OF F-TEST FOR TWO-PHASE LINEAR REGRESSION

ABSTRACT: In this paper, by using *the roughness of the model function based on the given basis function system*, the author will present an explicit formula for the non-centrality parameter of F-test statistic for two-phase model function. This parameter depends on the sample size and increases with the ratio between the roughness and the error of the model. The findings are used for calculating the hypothesis of the type II error for change-point of two-phase model function. Furthermore, based on some findings of the convergence as well as convergence speed of the roughness based on the weak convergence of the design sequences. The author also points out some consequences and the limit of probability of the type II error.

Keywords: roughness, type II error, the model function, change-point.

---

1. GIỚI THIỆU

Nhiều quá trình trong thực tế tuân theo mô hình hồi quy 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 được gọi là mô hình điểm chuyển - đã được phát triển trong nhiều thập kỉ gần đây và được áp dụng rộng rãi trong nhiều lĩnh vực khác nhau. Có nhiều phương pháp đã được đề xuất để xem xét liệu một hàm mô hình có xảy ra điểm chuyển tại một vị trí nào đó hay không: phương pháp sử dụng tỉ số hợp lý [1], phương pháp sử dụng tiêu chuẩn thông tin [6], sử dụng thống kê CUSUM để phát hiện điểm chuyển được nghiên cứu bởi Berkes, I. và cộng sự [4]. …

Trong trường hợp sai số có phân bố chuẩn với phương sai hằng số, Chow, G [5] đã đưa ra tiêu chuẩn kiểm định cho bài toán hồi quy hai pha dựa vào thống kê F - có phân bố Fisher-Snedecon dưới giả thuyết gốc không có sự chuyển đổi. Để xác định được xác suất mắc sai lầm loại 2 khi kiểm định, việc tìm phân bố của F dưới đối thuyết là một việc cần thiết. Rõ ràng, khi mô hình có sự thay đổi trạng thái thì hàm mô hình đã bị "bẻ gãy", tức là không còn "thẳng, "trơn" nữa. Vì vậy, trong mục 2, chúng tôi giới thiệu tổng quan về độ ráp của hàm mô hình theo hệ hàm cơ sở và một số kết quả về sự hội tụ của nó. Ở mục 3, là kết quả chính của bài báo, chúng tôi chỉ ra độ ráp của hàm mô hình theo hệ hàm cơ sở cho trước không phụ thuộc vào hệ số của pha đầu mà chỉ phụ thuộc vào độ lệch hệ số giữa hai pha. Từ đó, chúng tôi chỉ ra phân bố F dưới đối thuyết là phân bố F phi trung tâm, ở đó tham số phi trung tâm được biểu diễn thông qua kích thước mẫu, độ ráp của hàm mô hình và phương sai của sai số ngẫu nhiên. Trong mục này một số kết quả cũng được đưa ra để đánh giá độ tốt của thống kê kiểm định F được xây dựng thông qua giới hạn của xác suất mắc sai lầm loại II.

2. TỔNG QUAN NGHIÊN CỨU

Để xét sự chuyển đổi trạng thái, chúng ta xem xét hàm mô hình f(x) được xác định bởi:

$$f(x)=\begin{cases}\beta_{11}z_1(x)+...+\beta_{1p}z_p(x) & \text{khi} \quad a\le x\le x^*,\\ \beta_{21}z_1(x)+...+\beta_{2p}z_p(x) & \text{khi} \quad x^*<x\le b,\end{cases}$$ (1) ở đó $z_1(x),...,z_p(x)$ đã biết

và các tham số β_{ij} chưa biết.

Chúng ta quan tâm đến bài toán kiểm định giả thuyết xem có sự thay đổi hàm mô hình tại điểm chuyển x^* đã biết hay không, tức cần kiểm định cặp giả thuyết $H_0:\boldsymbol{\beta}_1=\boldsymbol{\beta}_2=\boldsymbol{\beta}$ với $H_1:\boldsymbol{\beta}_1\ne\boldsymbol{\beta}_2$ (2)

Nếu xảy ra H_1, tức là tại thời điểm k* tương ứng với điểm chuyển x*, hàm mô hình không còn giữ nguyên "dáng điệu" cũ mà chuyển sang "dáng điệu" mới, tức là nếu xét với hệ cơ sở $\{z_1(x),...,z_p(x)\}$ thì hàm mô hình đã bị "bẻ cong".

Để nghiên cứu, mức độ bị "bẻ cong" của hàm mô hình sau thời điểm k*, các tác giả trong [2, 3] đã đề xuất khái niệm "độ ráp" của hàm mô hình và đã thu được một số kết quả. Để tiện theo dõi, chúng tôi sử dụng một số kí hiệu: (a_{ij}) là ma trận có phần tử a_{ij} ở dòng i cột j; $\mathbf{I}_n$ là ma trận đơn vị cỡ n; $\mathbf{A}'$ là ma trận chuyển vị của ma trận $\mathbf{A}$; $\|\mathbf{A}\|$ là chuẩn max của ma trận **A**. Hơn nữa, không gian con sinh bởi $\mathbf{d}_1,...,\mathbf{d}_m$ được kí hiệu bởi $L\{\mathbf{d}_1,...,\mathbf{d}_m\}$; phép lấy tổng trực tiếp của hai không gian con bởi $\oplus$; phép chiếu trực giao lên V bởi $\mathbf{P}_V$; dãy hữu hạn $\{x_1,...,x_n\}$ bởi $\{x_i\}_n$. Đối với hàm phân bố F(x) có giá trên I, kí hiệu $<h,g>_F=\int_I h(x)g(x)dF(x)$ là tích trong của hai hàm h, g theo độ đo dF- độ đo Lebesgue-Stieltjes tương ứng với F(x).

Để xấp xỉ hàm $f(x),\ x\in I=[a,b]$ theo hệ hàm $\{z_1(x),...,z_p(x),x\in I\}$ tại các điểm cho trước $x_1,...,x_n\in I$ ta xét mô hình $f(x_i)=\alpha_1z_1(x_i)+...+\alpha_pz_p(x_i)+\gamma_i,\ i=1,...,n,$ (3) với sai số γ_i là xác định. Kí hiệu

$$\mathbf{z}(x)=\begin{pmatrix} z_1(x) \\ \dots \\ z_p(x) \end{pmatrix}, \quad \mathbf{Z}=\begin{pmatrix} \mathbf{z}'(x_1) \\ \dots \\ \mathbf{z}'(x_n) \end{pmatrix}, \quad \boldsymbol{\alpha}=\begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_p \end{pmatrix}, \quad \mathbf{f}=\begin{pmatrix} f(x_1) \\ \dots \\ f(x_n) \end{pmatrix}, \quad \boldsymbol{\gamma}=\begin{pmatrix} \gamma_1 \\ \dots \\ \gamma_n \end{pmatrix}.$$

Khi đó, mô hình (3) được viết lại dưới dạng $\mathbf{f}=\mathbf{Z}\boldsymbol{\alpha}+\boldsymbol{\gamma}$. Ở đây, chúng ta luôn giả sử $\mathrm{Rank}(\mathbf{Z})=p$, nên ước lượng của $\boldsymbol{\alpha}$ làm cho tổng $\sum_{i=1}^{n}\left(f(x_i)-\mathbf{z}'(x_i)\boldsymbol{\alpha}\right)^2$ nhỏ nhất là $\hat{\boldsymbol{\alpha}}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{f}$.

Định nghĩa 2.1. Ước lượng cho sai số của mô hình (3)

$$S^2(f,\mathbf{z},\{x_i\}_n)=\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i)-\mathbf{z}'(x_i)\hat{\boldsymbol{\alpha}}\right)^2$$

được gọi là độ ráp của hàm f(x) đối với hệ hàm $\{z_1(x),...,z_n(x)\}$ dựa trên thiết kế $\{x_i\}_n$.

Định nghĩa 2.2. Thiết kế F(x) được gọi là tương thích với hệ hàm $\{z_1(x),...,z_p(x), x\in I\}$ trên I nếu nó có giá trên I và ma trận $<\mathbf{z},\mathbf{z}'>_F=(<z_i,z_j>_F)$ khả nghịch.

Để nghiên cứu trường hợp giới hạn, chúng ta coi mỗi hàm phân bố có giá trên I và giá đó chứa ít nhất p điểm phân biệt là một thiết kế trên I.

Chúng ta xét một số giả thiết sau.

(A1) Hàm mô hình $f(x),\ x\in I$ liên tục từng khúc và bị chặn, tồn tại đạo hàm $f'(x)$ liên tục, bị chặn, trừ một số hữu hạn điểm.

(A2) Hệ hàm $z_1(x),...,z_p(x),\ x\in I$ liên tục.

(A3) Thiết kế được sử dụng tương thích với hệ $\{z_1(x),...,z_p(x), x\in I\}$.

Đôi khi, chúng ta có thể thay (A2) bởi:

(A2') Hệ các hàm $z_1(x),...,z_p(x),\ x\in I$ là khả vi liên tục.

Nếu các giả thiết (A1)-(A3) thỏa mãn thì vector tham số $\boldsymbol{\alpha}$ làm cực tiểu $\int_I\left(f(x)-\mathbf{z}'(x)\boldsymbol{\alpha}\right)^2 dF(x)$ được ước lượng bởi $\hat{\boldsymbol{\alpha}}=\left(\langle\mathbf{z},\mathbf{z}'\rangle_F\right)^{-1}<\mathbf{z},f>_F$,

với $<\mathbf{z},f>_F=(<z_1,f>_F,...,<z_p,f>_F)'$.

Định nghĩa 2.3. Với các giả thiết (A1)-(A3), $S^2(f,\mathbf{z},F)=\int_I\left(f(x)-\mathbf{z}'(x)\hat{\boldsymbol{\alpha}}\right)^2 dF(x)$ được gọi là độ ráp của hàm mô hình f(x) đối với hệ hàm $\{z_1(x),...,z_p(x),\ x\in I\}$ dựa trên thiết kế F(x).

Sự hội tụ của độ ráp được chỉ ra trong [2], và tốc độ hội tụ được chỉ ra trong [3] thông qua định lý sau.

Định lý 2.4 ([2]) Với giả thiết (A1)-(A3), dãy thiết kế $\{G_n(x)\}$ hội tụ yếu về thiết kế G(x). Gọi $\hat{\boldsymbol{\alpha}}^{(n)}$, $\hat{\boldsymbol{\alpha}}$, $S^2(f,\mathbf{z},G_n)$, $S^2(f,\mathbf{z},G)$ là ước lượng của tham số và độ ráp của hàm mô hình f(x) dựa trên thiết kế $G_n(x)$, $G(x)$ tương ứng. Giả sử $(dG)(D_f)=0$, với D_f là tập điểm gián đoạn của f(x), khi đó

(i) $\lim\limits_{n\to\infty} \hat{\boldsymbol{\alpha}}^{(n)} = \hat{\boldsymbol{\alpha}},$

(ii) $\lim\limits_{n\to\infty} S^2(f,\mathbf{z},G_n) = S^2(f,\mathbf{z},G)$.

Định lý 2.5 ([3]) Giả sử (A1), (A2') và phân bố đều rời rạc $F_n(x)$ có giá tại các điểm $\{x_i = a+(b-a)i/n,\ i=1,...,n\}$ và phân bố đều $F_{[a,b]}(x)$ tương thích với hệ $\{z_1(x),...,z_p(x)\}$. Kí hiệu $\hat{\boldsymbol{\alpha}}^{(n)}$, $\hat{\boldsymbol{\alpha}}$, $S^2(f,\mathbf{z},\{x_i\}_n)$, $S^2(f,\mathbf{z},F_{[a,b]})$ tương ứng là ước lượng của tham số và độ ráp của hàm mô hình f(x) dựa trên thiết kế $\{x_i\}_n$ và $F_{[a,b]}(x)$.

Khi đó (i) $\|\hat{\boldsymbol{\alpha}}^{(n)} - \hat{\boldsymbol{\alpha}}\| = O\left(\frac{1}{n}\right),$ (4)

(ii) $S^2(f,\mathbf{z},\{x_i\}_n) = S^2(f,\mathbf{z},F_{[a,b]}) + O\left(\frac{1}{n}\right).$ (5)

Tiếp tục với bài toán (2), không mất tính tổng quát, ta xét hàm mô hình f(x) tại các quan sát $x_i \in I = [a, b]$, i=1,...,n và $x_1 \le x_2 \le ... \le x_{k^*} \le x^* < x_{k^*+1} \le ... \le x_n$, khi đó mô hình được viết dưới dạng:

$$y_i = f(x_i) + \varepsilon_i = \begin{cases} \beta_{11}z_1(x_i) + ... + \beta_{1p}z_p(x_i) + \varepsilon_i, & i = 1,...,k^*, \\ \beta_{21}z_1(x_i) + ... + \beta_{2p}z_p(x_i) + \varepsilon_i, & i = k^*+1,...,n, \end{cases} \quad (6)$$

với giả thiết $\{\varepsilon_i\}$ là dãy các biến ngẫu nhiên độc lập cùng phân bố $N(0,\sigma^2)$, ở đó $\sigma^2 > 0$ chưa biết. Kí hiệu

$$\mathbf{Z}_1 = \begin{pmatrix} \mathbf{z}'(x_1) \\ ... \\ \mathbf{z}'(x_{k^*}) \end{pmatrix},\ \mathbf{Z}_2 = \begin{pmatrix} \mathbf{z}'(x_{k^*+1}) \\ ... \\ \mathbf{z}'(x_n) \end{pmatrix},\ \mathbf{Y}_1 = \begin{pmatrix} y_1 \\ ... \\ y_{k^*} \end{pmatrix},\ \mathbf{Y}_2 = \begin{pmatrix} y_{k^*+1} \\ ... \\ y_n \end{pmatrix}, \boldsymbol{\beta}_i = \begin{pmatrix} \beta_{i1} \\ ... \\ \beta_{ip} \end{pmatrix},\ i = 1, 2,$$

$$\boldsymbol{\varepsilon}_1 = \begin{pmatrix} \varepsilon_1 \\ ... \\ \varepsilon_{k^*} \end{pmatrix},\quad \boldsymbol{\varepsilon}_2 = \begin{pmatrix} \varepsilon_{k^*+1} \\ ... \\ \varepsilon_n \end{pmatrix},\ \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix},\quad \mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix},\quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}.$$

Khi đó, mô hình (7) được viết lại

$$\mathbf{Y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,\ i = 1, 2 \text{ hay } \mathbf{Y} = \begin{pmatrix} \mathbf{Z}_1\boldsymbol{\beta}_1 \\ \mathbf{Z}_2\boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix} = \mathbf{f} + \boldsymbol{\varepsilon}. \quad (7)$$

Để thực hiện kiểm định, chúng ta luôn giả sử ma trận $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}$ có đủ hạng, tức là $\text{Rank}(\mathbf{Z}_1) = \text{Rank}(\mathbf{Z}_2) = \text{Rank}(\mathbf{Z}) = p.$

Dưới H_0, ước lượng bình phương nhỏ nhất cho $\boldsymbol{\beta} = (\beta_1,...,\beta_p)'$ được xác định

$\hat{\boldsymbol\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$ và tổng bình phương phần dư tương ứng là $SSE_T = \sum_{i=1}^{n}\left(y_i - \mathbf{z}'(x_i)\hat{\boldsymbol\beta}\right)^2$.

Dưới H_1, ước lượng bình phương nhỏ nhất cho $\boldsymbol\beta_1$, $\boldsymbol\beta_2$ dựa trên k^* quan sát đầu và $n-k^*$ quan sát sau được xác định: $\hat{\boldsymbol\beta}_1 = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Y}_1$, $\quad \hat{\boldsymbol\beta}_2 = (\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\mathbf{Y}_2$, (8) và tổng bình phương phần dư tương ứng là

$$SSE_1 = \sum_{i=1}^{k^*}\left(y_i - \mathbf{z}'(x_i)\hat{\boldsymbol\beta}_1\right)^2,\ SSE_2 = \sum_{i=k^*+1}^{n}\left(y_i - \mathbf{z}'(x_i)\hat{\boldsymbol\beta}_2\right)^2.$$

Trong Chow, G. [5], Lehman và cộng sự [7], đã chỉ ra dưới H_0, thống kê kiểm định $F = \frac{\left(SSE_T - (SSE_1 + SSE_2)\right)/p}{(SSE_1 + SSE_2)/(n-2p)}$ (9) có phân bố $F(p, n-2p)$. Do đó, với mức ý nghĩa α, giả thuyết H_0 bị bác bỏ nếu $F > f_\alpha(p, n-2p)$. (10)

3. KẾT QUẢ NGHIÊN CỨU

Để đánh giá bài toán kiểm định (2), ta viết lại hàm mô hình (1) dưới dạng $f(x) = \mathbf{Z}(x)\boldsymbol\beta_1 + \mathbf{Z}(x)(\boldsymbol\beta_2 - \boldsymbol\beta_1)I_{(x^*,b]}(x)$.

Đặt $\boldsymbol\delta = \boldsymbol\beta_2 - \boldsymbol\beta_1$, khi đó ta viết $f_{\boldsymbol\beta_1,\boldsymbol\delta,x^*}(x)$ thay cho $f(x) = \mathbf{Z}(x)\boldsymbol\beta_1 + \mathbf{Z}(x)\boldsymbol\delta.I_{(x^*,b]}(x)$

.

Định lý sau sẽ chỉ ra độ ráp của hàm mô hình f(x) không phụ thuộc và hệ số của pha đầu mà chỉ phụ thuộc vào độ lệch hệ số giữa hai pha $\boldsymbol\beta_2 - \boldsymbol\beta_1$.

Định lý 3.1 Với giả thiết (A2)-(A3) thì $S^2(f_{\boldsymbol\beta_1,\boldsymbol\delta,x^*}, \mathbf{z}, F) = S^2(f_{\boldsymbol\delta,x^*}, \mathbf{z}, F)$ ở đó, $f_{\boldsymbol\delta,x^*}(x) = f_{\mathbf{0},\boldsymbol\delta,x^*}(x) = \mathbf{z}'(x)\boldsymbol\delta.I_{(x^*,b]}(x)$.

Chứng minh. Gọi $\hat{\boldsymbol\alpha}$ và $\hat{\boldsymbol\alpha}_0$ là các tham số ước lượng làm cực tiểu

$$\int_I (f_{\boldsymbol\beta_1,\boldsymbol\delta,x^*}(x) - \mathbf{z}'(x)\boldsymbol\alpha)^2\, dF(x) \text{ và } \int_I (f_{\boldsymbol\delta,x^*}(x) - \mathbf{z}'(x)\boldsymbol\alpha)^2\, dF(x)$$

Khi đó,

$$\begin{aligned} S^2(f_{\boldsymbol\beta_1,\boldsymbol\delta,x^*}, \mathbf{z}, F) &= \int_I (f_{\boldsymbol\beta_1,\boldsymbol\delta,x^*}(x) - \mathbf{z}'(x)\hat{\boldsymbol\alpha})^2\, dF(x) \\ &= \int_I (f_{\boldsymbol\delta,x^*}(x) - \mathbf{z}'(x)(\hat{\boldsymbol\alpha} - \boldsymbol\beta_1))^2\, dF(x) \\ &\geq \int_I (f_{\boldsymbol\delta,x^*}(x) - \mathbf{z}'(x)\hat{\boldsymbol\alpha}_0)^2\, dF(x) = S^2(f_{\boldsymbol\delta,x^*}, \mathbf{z}, F). \end{aligned}$$

Tương tự, $S^2(f_{\boldsymbol\delta,x^*}, \mathbf{z}, F) \geq S^2(f_{\boldsymbol\beta_1,\boldsymbol\delta,x^*}, \mathbf{z}, F)$, từ đó ta có điều phải chứng minh. □

Hơn nữa, cũng bằng kí hiệu trên, từ việc xét bài toán kiểm định điểm chuyển cho hàm mô hình f_{β_1,δ,x^*} ta chuyển về bài toán kiểm định điểm chuyển cho hàm mô hình f_{δ,x^*}. Khi đó, bài toán (2) trở thành $H_0 : \boldsymbol{\delta} = \mathbf{0}$ với $H_1 : \boldsymbol{\delta} \neq 0$ (11)

Tiếp theo, chúng tôi sẽ chỉ ra phân bố F dưới H_1 thông qua độ ráp bởi định lý sau.

Định lý 3.2. Với giả thiết (A2)-(A3), dưới H_1 thống kê kiểm định F xác định bởi (12) có phân bố Fisher phi trung tâm $F(\lambda, p, n-2p)$ với tham số phi trung tâm $\lambda = nS^2(f_{\delta,x^*}, \mathbf{z}, \{x_i\}_n) / \sigma^2$. (12)

Chứng minh. Để đơn giản, chúng ta vẫn kí hiệu ước lượng hệ số cho pha đầu (hệ số **0**) là $\hat{\boldsymbol{\beta}}_1$ và $\hat{\boldsymbol{\beta}}$ là ước lượng hệ số dựa trên toàn bộ số liệu, chúng ta ước lượng hệ số cho pha sau (hệ số $\boldsymbol{\delta}$) bởi $\hat{\boldsymbol{\delta}}$. Đặt $W_1 = SSE_1 + SSE_2$, $W_2 = SSE_T - (SSE_1 + SSE_2)$.

Rõ ràng $W_1 = SSE_1 + SSE_2 = \left\|\mathbf{Y}_1 - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1\right\|^2 + \left\|\mathbf{Y}_2 - \mathbf{Z}_2\hat{\boldsymbol{\delta}}\right\|^2$.

Vì $\left\|\mathbf{Y}_1 - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1\right\|^2 = \sigma^2\chi^2(k^* - p)$, $\left\|\mathbf{Y}_2 - \mathbf{Z}_2\hat{\boldsymbol{\delta}}\right\|^2 = \sigma^2\chi^2(n - k^* - p)$ và chúng độc lập nên $W_1 = \sigma^2\chi^2(n-2p)$. Mặt khác,

$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{f} + \boldsymbol{\varepsilon}) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{f} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon},$ (13) với kí hiệu $\mathbf{f} = \begin{pmatrix} \mathbf{0} \\ \mathbf{Z}_2\boldsymbol{\delta} \end{pmatrix}$

Dễ dàng kiểm tra được $\left(\mathbf{Y}_1 - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1\right)'\left(\mathbf{Z}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{Z}_1\hat{\boldsymbol{\beta}}\right) + \left(\mathbf{Y}_2 - \mathbf{Z}_2\hat{\boldsymbol{\delta}}\right)'\left(\mathbf{Z}_2\hat{\boldsymbol{\delta}} - \mathbf{Z}_2\hat{\boldsymbol{\beta}}\right) = 0$ nên

$$SSE_T = \left\|\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}\right\|^2 = \left\|\begin{matrix} \mathbf{Y}_1 - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1 \\ \mathbf{Y}_2 - \mathbf{Z}_2\hat{\boldsymbol{\delta}} \end{matrix}\right\|^2 + \left\|\begin{matrix} \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{Z}_1\hat{\boldsymbol{\beta}} \\ \mathbf{Z}_2\hat{\boldsymbol{\delta}} - \mathbf{Z}_2\hat{\boldsymbol{\beta}} \end{matrix}\right\|^2$$

Do đó, $W_2 = SSE_T - (SSE_1 + SSE_2) = \|\boldsymbol{\phi}\|^2$, với $\boldsymbol{\phi} = \begin{pmatrix} \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{Z}_1\hat{\boldsymbol{\beta}} \\ \mathbf{Z}_2\hat{\boldsymbol{\delta}} - \mathbf{Z}_2\hat{\boldsymbol{\beta}} \end{pmatrix}$.

Hơn nữa, từ (7), (8) và (13)

$$\begin{aligned} \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{Z}_1\hat{\boldsymbol{\beta}} &= \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\boldsymbol{\varepsilon}_1 - \mathbf{Z}_1(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{f} + \boldsymbol{\varepsilon}) \\ &= -\mathbf{Z}_1(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{f} + \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\boldsymbol{\varepsilon}_1 - \mathbf{Z}_1(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}. \end{aligned}$$

Tương tự,

$$\mathbf{Z}_2\hat{\boldsymbol\delta} - \mathbf{Z}_2\hat{\boldsymbol\beta} = \mathbf{Z}_2\boldsymbol\delta - \mathbf{Z}_2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{f} + \mathbf{Z}_2(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\boldsymbol\varepsilon_2 - \mathbf{Z}_2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol\varepsilon.$$

Nên

$$\boldsymbol\phi = \begin{pmatrix} \mathbf{Z}_1\hat{\boldsymbol\beta}_1 - \mathbf{Z}_1\hat{\boldsymbol\beta} \\ \mathbf{Z}_2\hat{\boldsymbol\beta}_2 - \mathbf{Z}_2\hat{\boldsymbol\beta} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{Z}_2\boldsymbol\delta \end{pmatrix} - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{f} + \begin{pmatrix} \mathbf{P}_1\boldsymbol\varepsilon_1 \\ \mathbf{P}_2\boldsymbol\varepsilon_2 \end{pmatrix} - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol\varepsilon$$

$$= \mathbf{f} - \mathbf{Z}\hat{\boldsymbol\alpha} + \begin{pmatrix} \mathbf{P}_1\boldsymbol\varepsilon_1 \\ \mathbf{P}_2\boldsymbol\varepsilon_2 \end{pmatrix} - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol\varepsilon = \boldsymbol\ell + \begin{pmatrix} \mathbf{P}_1\boldsymbol\varepsilon_1 \\ \mathbf{P}_2\boldsymbol\varepsilon_2 \end{pmatrix} - \mathbf{P}\boldsymbol\varepsilon.$$

Với $\boldsymbol\ell = \mathbf{f} - \mathbf{Z}\hat{\boldsymbol\alpha} \in \mathbb{R}_n$, và $\mathbf{P}_i = \mathbf{Z}_i(\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'$, $\mathbf{P} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'$ tương ứng là các phép chiếu trực giao lên không gian con sinh bởi các cột của ma trận $\mathbf{Z}_i$, $\mathbf{Z}$. Với j=1,…,p, đặt

$\mathbf{u}_j = (z_j(x_1),...,z_j(x_{k^*}),0,...,0)'$, $\mathbf{v}_j = (0,...0,z_j(x_{k^*+1}),...,z_j(x_n))'$,

$\mathbf{e}_j = \mathbf{u}_j + \mathbf{v}_j = (z_j(x_1),...,z_j(x_n))'$, $j = 1,...,p$,

$$L_1 = L\{\mathbf{u}_1,..., \mathbf{u}_p\},\ L_2 = L\{\mathbf{v}_1,..., \mathbf{v}_p\},\ L = L\{\mathbf{e}_1,..., \mathbf{e}_p\},\ K = L_1 \oplus L_2.$$

Ta có $L \subset K$ nên viết được $K = L \oplus L^\perp$, ở đó $L^\perp$ là phần bù trực giao của L trong K. Từ $\dim(L_1) = \dim(L_2) = \dim(L) = p$ suy ra $\dim(K) = 2p$, và $\dim(L^\perp) = p$. Hơn nữa,

$$\begin{pmatrix} \mathbf{P}_1\boldsymbol\varepsilon_1 \\ \mathbf{P}_2\boldsymbol\varepsilon_2 \end{pmatrix} = (\mathbf{P}_{L_1} + \mathbf{P}_{L_2})\boldsymbol\varepsilon = \mathbf{P}_K\boldsymbol\varepsilon, \quad \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}\boldsymbol\varepsilon = \mathbf{P}_L\boldsymbol\varepsilon,$$

nên $\boldsymbol\phi = \boldsymbol\ell + (\mathbf{P}_K\boldsymbol\varepsilon - \mathbf{P}_L\boldsymbol\varepsilon) = \boldsymbol\ell + \mathbf{P}_{L^\perp}\boldsymbol\varepsilon = \sigma\left(\dfrac{\boldsymbol\ell}{\sigma} + \dfrac{1}{\sigma}\mathbf{P}_{L^\perp}\boldsymbol\varepsilon\right).$

Do đó $W_2 = \|\boldsymbol\phi\|^2 = \sigma^2\chi^2\left(\dfrac{1}{\sigma^2}\|\boldsymbol\ell\|^2, p\right)$, với $\chi^2(\lambda, k)$ kí hiệu của phân bố khi bình hương phi trung tâm với k bậc tự do và tham số phi trung tâm λ. Bởi vì W_1, W_2 độc lập nên $F = \dfrac{W_2 / p}{W_1 / (n-2p)} = F\left(\dfrac{1}{\sigma^2}\|\boldsymbol\ell\|^2, p, n-2p\right)$, với

$$\frac{1}{\sigma^2}\|\boldsymbol\ell\|^2 = \frac{1}{\sigma^2}\|\mathbf{f} - \mathbf{Z}\hat{\boldsymbol\alpha}\|^2 = \frac{nS^2(f_{\delta,x^*}, \mathbf{z}, \{x_i\}_n)}{\sigma^2}.$$

Từ đó ta có điều phải chứng minh. □

Nhận xét 3.3. Công thức (12) cho thấy tham số phi trung tâm phụ thuộc vào độ ráp - thể hiện sự "không thẳng" hoặc "không trơn" của hàm mô hình theo hệ hàm cơ sở cho trước. Hơn nữa, ta có

$$\sigma^2\lambda = \sum_{i=1}^{n}(\mathbf{Z}_{i.}\boldsymbol\beta - \mathbf{Z}_{i.}\hat{\boldsymbol\beta})^2 - \sum_{i=1}^{k^*}(\mathbf{Z}_{i.}\boldsymbol\beta - \mathbf{Z}_{i.}\hat{\boldsymbol\beta})^2 - \sum_{i=k^*+1}^{n}(\mathbf{Z}_{i.}\boldsymbol\beta - \mathbf{Z}_{i.}\hat{\boldsymbol\beta})^2.$$

Thực vậy, $VP = \left\|\begin{pmatrix}\mathbf{0}\\ \mathbf{Z}_2\boldsymbol{\delta}\end{pmatrix} - \mathbf{Z}\hat{\boldsymbol{\beta}}\right\|^2 - \left(\left\|-\mathbf{Z}_1\hat{\boldsymbol{\beta}}_1\right\|^2 + \left\|\mathbf{Z}_2\boldsymbol{\delta} - \mathbf{Z}_2\hat{\boldsymbol{\delta}}\right\|^2\right).$

Lại có,

$$\left\|\begin{pmatrix}\mathbf{0}\\ \mathbf{Z}_2\boldsymbol{\delta}\end{pmatrix} - \mathbf{Z}\hat{\boldsymbol{\beta}}\right\|^2 = \left\|\mathbf{f} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{f}+\boldsymbol{\varepsilon})\right\|^2 = \left\|(\mathbf{I}_n - \mathbf{P})(\mathbf{f}) + \mathbf{P}(\boldsymbol{\varepsilon})\right\|^2$$

$$= \left\|(\mathbf{I}_n - \mathbf{P})(\mathbf{f})\right\|^2 + \left\|\mathbf{P}(\boldsymbol{\varepsilon})\right\|^2,$$

$$\left\|\mathbf{Z}_1\hat{\boldsymbol{\beta}}_1\right\|^2 + \left\|\mathbf{Z}_2\boldsymbol{\delta} - \mathbf{Z}_2\hat{\boldsymbol{\delta}}\right\|^2 = \left\|-\mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\boldsymbol{\varepsilon}_1)\right\|^2 +$$

$$+\left\|\mathbf{Z}_2\boldsymbol{\delta} - \mathbf{Z}_2(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\mathbf{Z}_2'(\mathbf{Z}_2\boldsymbol{\delta}+\boldsymbol{\varepsilon}_2)\right\|^2 = \left\|\mathbf{P}_1(\boldsymbol{\varepsilon}_1)\right\|^2 + \left\|\mathbf{P}_2(\boldsymbol{\varepsilon}_2)\right\|^2 = \left\|\mathbf{P}(\boldsymbol{\varepsilon})\right\|^2$$

nên $\sigma^2.VP = \left\|(\mathbf{I}_n - \mathbf{P})(\mathbf{f})\right\|^2 = \left\|\mathbf{f} - \mathbf{Z}\hat{\boldsymbol{\alpha}}\right\|^2 = \sigma^2\lambda$. □

Tiếp theo, ta đặt $R = S^2(f_{\delta,x^*}, \mathbf{z}, F)/\sigma^2$ và $R_n = S^2(f_{\delta,x^*}, \mathbf{z}, \{x_i\}_n)/\sigma^2$. Khi đó, dưới H_1, thống kê kiểm định F có phân bố $F(nR_n, p, n-2p)$ nên từ (10) ta hoàn toàn tính được xác suất mắc sai lầm loại 2 tương ứng là

$$P_{II}(f_{\delta,x^*}, \mathbf{z}, \{x_i\}_n) = P\left[F(nR_n, p, n-2p) < f_\alpha(p, n-2p)\right].$$

Hệ quả 3.4 Với giả thiết (A2’), (A3), ta có

$$P_{II}(f_{\delta,x^*}, \mathbf{z}, \{x_i\}_n) = P\left\{F\left(nR + \frac{1}{\sigma^2}O(1), p, n-2p\right) < f_\alpha(p, n-2p)\right\}, \quad (14)$$

với $x_i = a + (b-a)i/n,\ i = \overline{1,n}$, và $R = S^2(f, \mathbf{z}, F_{[a,b]})/\sigma^2$.

Hơn nữa, nếu $S^2(f, \mathbf{z}, F_{[a,b]}) > 0$ thì $\lim\limits_{n\to\infty} P_{II}(f, \mathbf{z}, \{x_i\}_n) = 0$.

Chứng minh. Từ (5),

$$R_n = \frac{S(f, \mathbf{z}, \{x_i\}_n)}{\sigma^2} = \frac{S(f, \mathbf{z}, F_{[a,b]})}{\sigma^2} + \frac{1}{\sigma^2}O\left(\frac{1}{n}\right) = R + \frac{1}{\sigma^2}O\left(\frac{1}{n}\right).$$

Nên ta suy được ngay (14). Phần kết luận còn lại là hiển nhiên. □

4. KẾT LUẬN

Bài báo đưa ra khái niệm về độ ráp của hàm mô hình và đã chỉ ra được tính chất khá thú vị của nó đối với mô hình hồi quy tuyến tính hai pha: không phụ thuộc vào hệ số của pha đầu mà chỉ phụ thuộc vào độ lệch hệ số của hai pha. Điều này sẽ giúp việc tính toán, nghiên cứu độ ráp tiện lợi hơn khi các hàm mô hình được "phân lớp" dựa trên độ lệch hệ số của hai pha. Từ đó, bài báo đã chỉ ra được công thức tính tham số phi trung tâm của tiêu chuẩn kiểm định F và xác suất mắc sai lầm loại II cho bài toán kiểm định sự tồn tại điểm chuyển của mô hình hồi quy tuyến tính hai pha. Cùng với một số kết quả về sự hội tụ và tốc độ hội tụ của độ ráp, bài báo cũng đã đánh giá được giới hạn về xác suất mắc sai lầm loại II đó.

TÀI LIỆU THAM KHẢO

1. Aue, A., Horvath, L., Huskova, M., Kokoszka, P (2008), "Testing for changes in polynomial regression*", Bernoulli,* 14, 637-660.

2. Ban, T.V., Quyen, N.T., Ha, P. T.: "The roughness of model function to the basis function"*, J. of Math. and System Science*, 3, pp. 385-390.

3. Tô Văn Ban, Nguyễn Thị Quyên, (2015), "Độ ráp của hàm mô hình và ứng dụng vào bài toán điểm chuyển", *Tạp chí Ứng dụng toán học*, 13 (1), tr. 17-22

4. Berkes, I., Horvath, L., Schauer, J. (2011), "Asymptotics of trimmed CUSUM statistics*", Bernoulli*, 17, pp. 1344-1367.

5. Chow, G.C. (1960), "Tests of equality between sets of coefficients in two linear regressions"*, Econometrica*, 28, pp. 591-605.

6. Koul, H. L., Qian, L. (2002), "Asymptotics of maximum likelihood estimator in a two-phase linear regression model"*, J. of Statistical Planning and Inference*, 108, pp. 99-119.

7. Lehmann, E.L., Romano, J. P. (2005)," Testing statistical hypotheses", 3th edition, Springer, pp. 277-282.