ESTIMATION IN ZERO INFLATION POISSON REGRESSION MODEL WITH RIGHT CENSOR

Van-Trinh Nguyen*, Van-Minh Pham

Fundamental-Basic Faculty, VietNam Maritime University
484 Lach Tray Street, Ngo Quyen District, Hai Phong City

ARTICLE INFO

Received: 11/6/2021
Revised: 29/11/2021
Published: 30/11/2021

KEYWORDS

Excess of zeroes
Count data
MLE
Censor
Poisson model

ABSTRACT

Count data often appears in many fields such as public health, economics, epidemiology... In order to handle this kind of data, some regression models have been developed as Poisson regression, Binomial regression or more generally are generalized linear regression models (GLMs). When count data contains extra of zeroes, zero-inflated (ZI) models are improved to suit. However, when counts are censored, the above models are no longer suitable. Therefore, Saffari and Adnan (2001) mentioned to this model using some simple simulations. However, the authors have not proven the existence, consistency, and asymptotic normality of a maximum likelihood estimator (MLE) yet. With that in mind, this paper develops theory to give out rigorous proof to handle the above problems basing upon the asymptotic normality theory.

ƯỚC LƯỢNG TRONG MÔ HÌNH HỒI QUI POISSON GIÃN NỞ SỐ KHÔNG KIỂM DUYỆT BÊN PHẢI

Nguyễn Văn Trịnh *, Phạm Văn Minh

Khoa cơ sở cơ bản - Đại học Hàng Hải Việt Nam
484 Lạch Tray, Ngô Quyền, Hải Phòng

THÔNG TIN BÀI BÁO

Ngày nhận bài: 11/6/2021
Ngày hoàn thiện: 29/11/2021
Ngày đăng: 30/11/2021

TỪ KHÓA

Giãn nở số không
Dữ liệu đếm
Ước lượng hợp lí cực đại
Kiểm duyệt
Mô hình Poisson

TÓM TẮT

Dữ liệu đếm thường xuất hiện trong nhiều lĩnh vực thực tế như y tế, kinh tế, dịch tễ học... Để xử lý loại dữ liện này, nhiều mô hình hồi quy đã phát triển như hồi quy Poisson, hồi quy Nhị thức hay tổng quát hơn là mô hình hồi quy tổng quát hóa (GLMs). Khi dữ liệu đếm chứa nhiều số không, các mô hình giãn nở số không (ZI) ra đời. Tuy nhiên nếu dữ liệu cần kiểm duyệt thì các mô hình trên không còn phù hợp. Vì vậy, Saffari and Adnan (2001) đã đề cập đến mô hình này bằng nghiên cứu mô phỏng đơn giản. Tuy nhiên, tác giả chưa chứng minh cho sự tồn tại, tính vững và tính tiệm cận chuẩn của đại lượng hợp lí cực đại (MLE). Với nhận định đó, bài báo này phát triển lý thuyết đưa ra chứng minh chặt chẽ cho các vấn đề trên dựa vào lý thuyết tiệm cận chuẩn.

DOI: https://doi.org/10.34238/tnu-jst.4636

*Corresponding author. *Email: trinhnv@vimaru.edu.vn*

1. Giới thiệu

Mô hình thống kê dữ liệu đếm đóng vai trò quan trọng trong các lĩnh vực như nông nghiệp, kinh tế, dịch tễ, công nghiệp hay sức khỏe công cộng.... Mô hình tuyến tính tổng quát [1] là giải pháp phù hợp cho dữ liệu này.

Tuy nhiên, trong nhiều ứng dụng, dữ liệu đếm xuất hiện số không với tần suất nào đó, tức là lượng số không không giải thích được bằng mô hình dựa trên giả thiết về phân phối. Một số công cụ thống kê được nghiên cứu để giải quyết vấn đề này trong đó có mô hình hồi qui giãn nở số không, mô hình này kết hợp một phân phối suy biến với mô hình hồi qui đếm.

Ví dụ, mô hình hồi qui giãn nở số không Poisson (ZIP) đề xuất bởi [2], hay gần đây như [3, 4, 5]. Một số biến đổi của hồi qui ZIP như mô hình ảnh hưởng ngẫu nhiên ZIP [6, 7], mô hình nửa tham số ZIP [8, 9]. Mô hình hồi qui giãn nở số không nhị thức âm (ZINB) được đề xuất bởi [10], hoặc [11, 12].

Bài báo này, chúng tôi nghiên cứu ước lượng trong hồi qui Poisson giãn nở số không khi biến đếm bị kiểm duyệt bên phải. Kiểm duyệt bên phải xảy ra khi chỉ cận dưới của biến tiên lượng được quan sát, hay nói cách khác ta chỉ biết giá trị thực của biến lớn hơn giá trị quan sát.

Saffari và Adnan [13], đề xuất ước lượng hợp lí cực đại (MLE) trong ZIP với giá trị kiểm duyệt bên phải là hằng số. Tuy nhiên, trong nghiên cứu này, tác giả chưa chỉ ra các kết quả tiệm cận của MLE. Với hạn chế và thiếu sót trên, chúng tôi mở rộng mô hình cho trường hợp đại lượng kiểm duyệt là ngẫu nhiên và thực hiện các chứng minh lý thuyết một cách chặt chẽ cho đại lượng MLE của mô hình này.

Phần còn lại của bài báo được bố cục như sau: mục 2, nhắc lại mô hình hồi qui ZIP, ước lượng hợp lí cực đại với kiểm duyệt ngẫu nhiên được mô tả cùng với việc một số kí hiệu dùng trong bài báo. Mục 3, chúng tôi thiết lập tính vững và tính tiệm cận chuẩn của MLE. Cuối cùng, một số thảo luận và hướng nghiên cứu được thực hiện ở mục 4

2. Mô hình hồi quy Poisson kiểm duyệt

Mục này nhắc lại định nghĩa mô hình ZIP và mô tả ước lượng hợp lí cực đại (MLE) khi biến tiên lượng bị kiểm duyệt bên phải ngẫu nhiên của mô hình ZIP giãn nở số không (CZIP).

2.1. *Ước lượng hợp lí cực đại trong mô hình CZIP*

Mô hình ZIP giả thiết biến tiên lượng Z_i (chỉ số i kí hiệu cho cá thể i) thỏa mãn:

$$Z_i \sim \begin{cases} 0 & \text{với xác suất } \omega_i, \\ \mathscr{P}(\lambda_i) & \text{với xác suất } 1-\omega_i, \end{cases} \tag{1}$$

với $\mathscr{P}(\lambda_i)$ là phân phối Poisson, tham số trung bình $\lambda_i > 0$. Dễ thấy, ZIP trở thành mô hình Poisson nếu $\omega_i = 0$. Trong hồi qui ZIP, xác suất trộn ω_i và tham số trung bình λ_i được xét bởi các mô hình logistic và log-linear tương ứng, cụ thể là:

$$\text{logit}(\omega_i(\gamma)) = \gamma^\top \mathbf{W}_i, \tag{2}$$

và

$$\log(\lambda_i(\beta)) = \beta^\top \mathbf{X}_i, \tag{3}$$

trong đó $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ và $\mathbf{W}_i = (1, W_{i2}, \ldots, W_{iq})^\top$ là các véc tơ ngẫu nhiên biến độc lập, $\beta \in \mathbb{R}^p$ và $\gamma \in \mathbb{R}^q$ là các tham số chưa biết; $\top$ kí hiệu cho toán tử chuyển vị.

Giả sử từ mô hình (1)-(2)-(3) chúng ta quan sát n vec tơ độc lập $(Z_1, \mathbf{X}_1, \mathbf{W}_1), \ldots, (Z_n, \mathbf{X}_n, \mathbf{W}_n)$, xác định trên không gian xác suất $(\Omega, \mathscr{C}, \mathbb{P})$. Khi đó, hàm log-hàm hợp lí của (β, γ) là:

$$\sum_{i=1}^{n} \left(1_{\{Z_i=0\}} \log\left(e^{\gamma^\top \mathbf{W}_i} + e^{-\exp(\beta^\top \mathbf{X}_i)} \right) + 1_{\{Z_i>0\}} \left(Z_i \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Z_i!) \right) - \log\left(1 + e^{\gamma^\top \mathbf{W}_i} \right) \right).$$

Ước lượng hợp lí cực đại (β, γ) thỏa mãn tính vững và tính tiệm cận phân phối chuẩn, xem [14].

Giả sử Z_i kiểm duyệt bên phải, tức là tồn tại một số cá thể i mà ta chỉ quan sát được cận dưới Z_i. Điều này được mô hình bằng cách đưa ra một biến ngẫu nhiên kiểm duyệt C_i. Cá thể i xem như vec tơ $(Z_i^*, \delta_i, \mathbf{X}_i, \mathbf{W}_i)$,

với $Z_i^* = \min(Z_i, C_i)$ và $\delta_i = 1_{\{Z_i < C_i\}}$ (nếu $Z_i = C_i$, ta đặt $Z_i^* = C_i$ và $\delta_i = 0$), đặt $J_i = 1_{\{Z_i^* = 0\}}$. Từ các quan sát $(Z_i^*, \delta_i, \mathbf{X}_i, \mathbf{W}_i)$, $i = 1, \ldots, n$, hàm hợp lí của $\psi := (\beta^\top, \gamma^\top)^\top$ được tính bởi:

$$\begin{aligned} L_n(\psi) &= \prod_{i=1}^{n} \mathbb{P}(Z_i = Z_i^* | \mathbf{X}_i, \mathbf{W}_i)^{\delta_i} \mathbb{P}(Z_i \geqslant Z_i^* | \mathbf{X}_i, \mathbf{W}_i)^{1-\delta_i}; \\ &= \prod_{i=1}^{n} \left(\left(e^{-\lambda_i} \frac{\lambda_i^{Z_i^*}}{Z_i^*!} (1-\omega_i) \right)^{1-J_i} \left(\omega_i + (1-\omega_i) e^{-\lambda_i} \right)^{J_i} \right)^{\delta_i} \\ &\qquad \times \left(1 - \sum_{k=0}^{Z_i^*-1} e^{-\lambda_i} \frac{\lambda_i^k}{k!} (1-\omega_i) - \omega_i \right)^{(1-\delta_i)(1-J_i)}, \end{aligned}$$

Lấy logarit hai vế thu được hàm log-hàm hợp lí $\ell_n(\psi) = \log L_n(\psi)$. Với ω_i và λ_i cho bởi (2) và (3) có:

$$\begin{aligned} \ell_n(\psi) &= \sum_{i=1}^{n} \Big(\delta_i \Big[J_i \log \left(e^{\gamma^\top \mathbf{W}_i} + e^{-\exp(\beta^\top \mathbf{X}_i)} \right) + (1-J_i) \big(Z_i^* \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Z_i^*!) \big) \Big] \\ &\qquad + (1-\delta_i)(1-J_i) \ln \left(1 - \sum_{k=0}^{Z_i^*-1} \frac{e^{-\exp(\beta^\top \mathbf{X}_i) + k\beta^\top \mathbf{X}_i}}{k!} \right) - \log \left(1 + e^{\gamma^\top \mathbf{W}_i} \right) \Big). \end{aligned}$$

Dễ thấy $\ell_n(\psi)$ suy ra log-hàm hợp lí ở trên khi không có yếu tố kiểm duyệt (tức là khi $\delta_i = 1$ với mọi $i = 1, \ldots, n$).

Ước lượng hợp lí cực đại $\hat{\psi}_n := (\hat{\beta}_n^\top, \hat{\gamma}_n^\top)^\top$ của ψ là nghiệm của hệ k phương trình

$$\frac{\partial \ell_n(\psi)}{\partial \psi} = 0, \tag{4}$$

ở đó $k = p + q$. Mục tới thiết lập sự tồn tại, tính vững và tính tiệm cận chuẩn của $\hat{\psi}_n$. Trước hết, ta đưa ra một số kí hiệu cần thiết.

2.2. Một số kí hiệu

Kí hiệu $k_i(\gamma) = e^{\gamma^\top \mathbf{W}_i}$ và $L_i(\beta) = e^{-\exp(\beta^\top \mathbf{X}_i)}$, $i = 1, \ldots, n$. Đặt $S_{\lambda_i(\beta)}(u) = \mathbb{P}(\mathscr{P}(\lambda_i(\beta)) \geqslant u)$, $u = 0, 1, \ldots$ là hàm sống sót của phân phối $\mathscr{P}(\lambda_i(\beta))$. Ta có:

$$\begin{aligned} \frac{\partial \ell_n(\psi)}{\partial \beta_\ell} = \sum_{i=1}^{n} X_{i\ell} \Big(&- \delta_i J_i \frac{\lambda_i(\beta) L_i(\beta)}{k_i(\gamma) + L_i(\beta)} + \delta_i (1-J_i) \big(Z_i^* - \lambda_i(\beta) \big) - (1-\delta_i)(1-J_i) \\ &\times \sum_{k=0}^{Z_i^*-1} \frac{L_i(\beta) \lambda_i^k(\beta) (k - \lambda_i(\beta))}{k! S_{\lambda_i(\beta)}(Z_i^*)} \Big), \quad \ell = 1, \ldots, p, \end{aligned} \tag{5}$$

và

$$\frac{\partial \ell_n(\psi)}{\partial \gamma_\ell} = \sum_{i=1}^{n} W_{i\ell} \left(\frac{\delta_i J_i k_i(\gamma)}{k_i(\gamma) + L_i(\beta)} - \frac{k_i(\gamma)}{k_i(\gamma) + 1} \right), \ \ell = 1, \ldots, q. \tag{6}$$

Đặt

$$\begin{aligned} u_i(\psi) &= \frac{\lambda_i(\beta) L_i(\beta)}{(k_i(\gamma) + L_i(\beta))^2} \left[k_i(\gamma) + L_i(\beta) - \lambda_i(\beta) k_i(\gamma) \right], \ i = 1, \ldots, n, \\ v_i(\psi) &= \sum_{k=0}^{Z_i^*-1} \frac{L_i(\beta) \lambda_i^k(\beta)}{k! S_{\lambda_i(\beta)}^2(Z_i^*)} \Big(S_{\lambda_i(\beta)}(Z_i^*) \big((\lambda_i(\beta) - k)^2 - \lambda_i(\beta) \big) \\ &\qquad - \lambda_i(\beta)(k - \lambda_i(\beta)) \mathbb{P}(\mathscr{P}(\lambda_i(\beta)) = Z_i^* - 1) \Big), \ i = 1, \ldots, n, \text{ với } Z_i^* \geqslant 1. \end{aligned}$$

Từ đó thu được:

$$\frac{\partial^2 \ell_n(\psi)}{\partial \beta_\ell \partial \beta_m} = \sum_{i=1}^{n} X_{i\ell} X_{im} \Big(-\delta_i J_i u_i(\psi) - \delta_i (1-J_i) \times \lambda_i(\beta) - (1-\delta_i)(1-J_i) v_i(\psi) \Big), \quad \ell, m = 1, \ldots, p;$$

$$\frac{\partial^2 \ell_n(\psi)}{\partial \beta_\ell \partial \gamma_m} = \sum_{i=1}^{n} X_{i\ell} W_{im} \frac{\delta_i J_i k_i(\gamma) \lambda_i(\beta) L_i(\beta)}{(k_i(\gamma) + L_i(\beta))^2}, \ell = 1, \ldots, p \text{ và } m = 1, \ldots, q;$$

$$\frac{\partial^2 \ell_n(\psi)}{\partial \gamma_\ell \partial \gamma_m} = \sum_{i=1}^{n} W_{i\ell} W_{im} k_i(\gamma) \Big(\frac{\delta_i J_i L_i(\beta)}{(k_i(\gamma) + L_i(\beta))^2} - \frac{1}{(k_i(\gamma)+1)^2} \Big), \quad \ell, m = 1, \ldots, q.$$

Kí hiệu $S_n(\psi) = \partial \ell_n(\psi)/\partial \psi$, $H_n(\psi) = -\partial^2 \ell_n(\psi)/\partial \psi \partial \psi^\top$, $F_n(\psi) = \mathbb{E}(H_n(\psi))$ và I_k là ma trận đơn vị cỡ k. $H_n(\psi)$ được giả thiết xác định dương.

3. Kết quả chính

Trong mục này, ta thiết lập tính vững và tiệm cận chuẩn của $\hat{\psi}_n$. Gọi $\mathbb{R}^k$ là không gian các vec tơ k chiều có chuẩn Euclidean $\|\cdot\|_2$ và $M_{k\times k}(\mathbb{R})$ là không gian các ma trận thực, vuông cấp k với chuẩn ma trận $|||A|||_2 := \sup_{\|x\|_2=1} \|Ax\|_2$ (để đơn giản, ta sử dụng $\|\cdot\|$ cho cả hai chuẩn. Nhắc lại rằng, nếu ma trận thực đối xứng A cỡ $(k \times k)$ có các giá trị riêng $\lambda_1, \ldots, \lambda_k$, thì $\|A\| = \max_i |\lambda_i|$ ($\lambda_{\min}(A)$ và $\lambda_{\max}(A)$ kí hiệu cho giá trị riêng nhỏ nhất và lớn nhất của A tương ứng).

Trước hết, ta phát biểu các điều kiện chính tắc:

D1 Các biến độc lập bị chặn, tức là tồn tại các tập compact $\mathscr{X} \subset \mathbb{R}^p$ và $\mathscr{W} \subset \mathbb{R}^q$ sao cho $\mathbf{X}_i \in \mathscr{X}$ và $\mathbf{W}_i \in \mathscr{W}$, $i = 1, 2, \ldots$

D2 Giá trị thực $\psi_0 = (\beta_0^\top, \gamma_0^\top)^\top$ thuộc miền trong của tập compact và lồi $\mathscr{C}$ nào đó, $\mathscr{C} = \mathscr{B} \times \mathscr{G} \subset \mathbb{R}^k$ (ở đó $\mathscr{B} \subset \mathbb{R}^p$ và $\mathscr{G} \subset \mathbb{R}^q$ là các không gian tham số β và γ).

D3 Tồn tại đại lượng dương không đổi c_1 sao cho $n/\lambda_{\min}(F_n(\psi_0)) \leqslant c_1$ với mỗi $n = 1, 2, \ldots$

D4 Biến ngẫu nhiên kiểm duyệt $C_i, i = 1, 2, \ldots$ dương, bị chặn bởi hằng số $M < \infty$.

D1-D3 là các điều kiện trong mô hình tuyến tính tổng quát cổ điển và các mô hình hồi qui tuyến giãn nở số không [15]. Điều kiện D4 là yêu cầu cho giá trị kiểm duyệt.

Với mỗi $n = 1, 2, \ldots$ và $\varepsilon > 0$, xét lân cận của ψ_0: $N_n(\varepsilon) = \{\psi \in \mathscr{C} : (\psi - \psi_0)^\top F_n (\psi - \psi_0) \leqslant \varepsilon^2\}$, ở đó F_n kí hiệu cho $F_n(\psi_0)$.

Kết quả đầu tiên phát biểu rằng nghiệm của phương trình (4) tồn tại, và vững trong lân cận $N_n(\varepsilon)$ của ψ_0 khi n đủ lớn, nhưng trước hết ta chỉ ra một bổ đề kĩ thuật.

Bổ đề 3.1. *Giả sử điều kiện D1-D4 đúng. Khi đó* $\sup_{\psi \in N_n(\varepsilon)} \|F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}} - I_k\|$ *hội tụ theo xác suất tới* 0 *khi* $n \to \infty$.

Chứng minh. Ta có

$$\begin{aligned} \|F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}} - I_k\| &= \|F_n^{-\frac{1}{2}} (H_n(\psi) - F_n) F_n^{-\frac{1}{2}}\|, \\ &\leqslant \frac{1}{\lambda_{\min}(F_n)} \|H_n(\psi) - F_n\|, \\ &\leqslant c_1 \left\| \frac{1}{n} (H_n(\psi) - \mathbb{E}(H_n(\psi))) \right\| + c_1 \left\| \frac{1}{n} (\mathbb{E}(H_n(\psi)) - F_n) \right\|, \end{aligned}$$

(vì D3). Do đó, bổ đề được chứng minh nếu ta chỉ ra cả hai số hạng bên vế phải của bất đẳng thức cuối hội tụ đều theo xác suất tới 0 trong $\psi \in N_n(\varepsilon)$ khi $n \to \infty$. Do đó, ta cần $\sup_{\psi \in N_n(\varepsilon)} \|\frac{1}{n}(H_n(\psi) - \mathbb{E}(H_n(\psi)))\|$ hội tụ theo xác suất tới 0 khi $n \to \infty$. Ta sẽ chỉ ra $\sup_{\psi \in N_n(\varepsilon)} |\frac{1}{n}(H_{n,(\ell,m)}(\psi) - \mathbb{E}(H_{n,(\ell,m)}(\psi)))|$ hội tụ tới 0 với mỗi (ℓ, m), $\ell, m = 1, \ldots, k$, ở đó $H_{n,(\ell,m)}$ kí hiệu (ℓ, m) là phần tử của H_n. Ta chứng minh cho trường hợp $l, m \in \{1, \ldots, p\}$, với $H_{n,(\ell,m)}(\psi) = -\partial^2 \ell_n(\psi)/\partial \beta_\ell \partial \beta_m$ (các trường hợp khác chứng minh tương tự). Thật vậy,

$$\left|\frac{1}{n}(H_{n,(\ell,m)}(\psi)-\mathbb{E}(H_{n,(\ell,m)}(\psi)))\right| \leqslant \left|\frac{1}{n}\sum_{i=1}^{n}\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi)-\mathbb{E}\left[X_{i\ell}X_{im}\delta_i J_i u_i(\psi)\right]\}\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\left\{X_{i\ell}X_{im}\delta_i(1-J_i)\lambda_i(\beta)-\mathbb{E}\left[X_{i\ell}X_{im}\delta_i(1-J_i)\lambda_i(\beta)\right]\right\}\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\left\{(1-\delta_i)(1-J_i)v_i(\psi)-\mathbb{E}\left[(1-\delta_i)(1-J_i)v_i(\psi)\right]\right\}\right|.$$

Bây giờ, ta chứng minh

$$\sup_{\psi\in N_n(\varepsilon)}\left|\frac{1}{n}\sum_{i=1}^{n}\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi)-\mathbb{E}\left[X_{i\ell}X_{im}\delta_i J_i u_i(\psi)\right]\}\right|$$

hội tụ theo xác suất tới 0 khi $n\to\infty$ (hai số hạng còn lại làm tương tự). Để chỉ ra điều này ta cần khẳng định lớp $\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi):\psi\in\mathscr{C}\}$ là Donsker (và do đó theo Glivenko-Cantelli nó sẽ hội tụ đều (theo ψ)). Lớp $\{X_{i\ell}X_{im}\delta_i J_i\}$ hiển nhiên là Donsker. Dưới điều kiện D1 và D2, các lớp $\{\beta^\top\mathbf{X}_i:\beta\in\mathscr{B}\}$ và $\{\gamma^\top\mathbf{W}_i:\gamma\in\mathscr{G}\}$ là Donsker. Hàm mũ là Lipschitz trên các tập compact và do đó các lớp $\{e^{\beta^\top\mathbf{X}_i}:\beta\in\mathscr{B}\}$, $\{e^{-\exp(\beta^\top\mathbf{X}_i)}:\beta\in\mathscr{B}\}$ và $\{e^{\gamma^\top\mathbf{W}_i}:\gamma\in\mathscr{G}\}$ cũng là Donsker. Hơn nữa, tích và tổng của các lớp Donsker bị chặn là Donsker, do đó, lớp $\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi):\psi\in\mathscr{C}\}$ là Donsker. Vì vậy,

$$\sup_{\psi\in\mathscr{C}}|\frac{1}{n}\sum_{i=1}^{n}\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi)-\mathbb{E}\left[X_{i\ell}X_{im}\delta_i J_i u_i(\psi)\right]\}|$$

hội tụ theo xác suất tới 0 khi $n\to\infty$. Vì $N_n(\varepsilon)\subset\mathscr{C}$, nên

$$\sup_{\psi\in N_n(\varepsilon)}|\frac{1}{n}\sum_{i=1}^{n}\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi)-\mathbb{E}\left[X_{i\ell}X_{im}\delta_i J_i u_i(\psi)\right]\}|$$

cũng hội tụ tới 0 khi $n\to\infty$.

Định lý 3.1 (Tồn tại và duy nhất)**.** *Giả sử các điều kiện D1-D4 đúng. Khi đó, xác suất $\hat{\psi}_n$ tồn tại và nằm trong $N_n(\varepsilon)$ dần tới 1 khi $n\to\infty$. Hơn nữa, $\hat{\psi}_n$ hội tụ theo xác suất tới ψ_0 khi $n\to\infty$.*

Chứng minh. Chứng minh của chúng tôi dựa theo [15] nhưng kĩ thuật chi tiết thì khác. Hơn nữa, một số lập luận dẫn tới chứng minh trực tiếp hơn.

a) Trước hết, chứng minh tính tồn tại tiệm cận của $\hat{\psi}_n$. Ta sẽ chỉ ra rằng, với mỗi $\eta>0$, tồn tại $\varepsilon>0$ và $n_1\in\mathbb{N}$ sao cho

$$\mathbb{P}\left(\ell_n(\psi)-\ell_n(\psi_0)<0,\ \forall\psi\in\partial N_n(\varepsilon)\right)\geqslant 1-\eta,\ \text{với } n\geqslant n_1, \tag{7}$$

ở đó $\partial N_n(\varepsilon)=\{\psi\in\mathscr{C}:(\psi-\psi_0)^\top F_n(\psi-\psi_0)=\varepsilon^2\}$ là biên của $N_n(\varepsilon)$. Điều này suy ra tồn tại cực đại địa phương của ℓ_n trong $N_n(\varepsilon)$. Tính xác định dương của H_n và tính lồi của $\mathscr{C}$ khẳng định cực đại đó là toàn cục và duy nhất.

Thật vậy (7) tương đương với: $\eta>0$, tồn tại $\varepsilon>0$ và $n_1\in\mathbb{N}$ sao cho

$$\mathbb{P}\left(\ell_n(\psi)-\ell_n(\psi_0)\geqslant 0\text{ với }\psi\in\partial N_n(\varepsilon)\right)\leqslant\eta,\ \text{với } n\geqslant n_1,$$

Xét khai triển Taylor

$$\ell_n(\psi)-\ell_n(\psi_0) = (\psi-\psi_0)^\top S_n(\psi_0)-\frac{1}{2}(\psi-\psi_0)^\top\times H_n(\tilde{\psi})(\psi-\psi_0) := (\psi-\psi_0)^\top S_n(\psi_0)-Q_n(\psi),$$

với $\tilde{\psi}=a\psi+(1-a)\psi_0$ $(0\leqslant a\leqslant 1)$, đặt $0<c<\frac{1}{2}$. Ta có:

$$\mathbb{P}\left(\ell_n(\psi)-\ell_n(\psi_0)\geqslant 0,\ \text{với }\psi\in\partial N_n(\varepsilon)\right) = \mathbb{P}\Big((\psi-\psi_0)^\top S_n(\psi_0)\geqslant Q_n(\psi)\text{ và }Q_n(\psi)>c\varepsilon^2,\ \text{với }\psi\in\partial N_n(\varepsilon)\Big) + \mathbb{P}\Big((\psi-\psi_0)^\top S_n(\psi_0)\geqslant Q_n(\psi)\text{ và }Q_n(\psi)\leqslant c\varepsilon^2,\ \text{với }\psi\in\partial N_n(\varepsilon)\Big) \leqslant \mathbb{P}(A)+\mathbb{P}(B),$$

ở đó $A = \{(\psi - \psi_0)^\top S_n(\psi_0) > c\varepsilon^2, \text{ với } \psi \in \partial N_n(\varepsilon)\}$ và $B = \{Q_n(\psi) \leqslant c\varepsilon^2, \text{ với } \psi \in \partial N_n(\varepsilon)\}$ tương ứng. Đặt $u_n(\psi) = \frac{1}{\varepsilon} F_n^{\frac{1}{2}}(\psi - \psi_0)$. Khi đó

$$\begin{aligned} A &= \{u_n(\psi)^\top F_n^{-\frac{1}{2}} S_n(\psi_0) > c\varepsilon, \text{ với } \psi \in \partial N_n(\varepsilon)\}, \\ &\subseteq \{\sup_{\|u_n(\psi)\|=1} |u_n(\psi)^\top F_n^{-\frac{1}{2}} S_n(\psi_0)| > c\varepsilon\}, \\ &= \{\|F_n^{-\frac{1}{2}} S_n(\psi_0)\| > c\varepsilon\}. \end{aligned}$$

Suy ra $\mathbb{P}(A) \leqslant \mathbb{P}(\|F_n^{-\frac{1}{2}} S_n(\psi_0)\| > c\varepsilon)$. Từ Định lý 1.5 của [16], $\mathbb{E}\|F_n^{-\frac{1}{2}} S_n(\psi_0)\|^2 = k$ và bất đẳng thức Chebyshev suy ra

$$\mathbb{P}(A) \leqslant \frac{k}{c^2\varepsilon^2}.$$

Cuối cùng, đặt $\varepsilon = \sqrt{\frac{2k}{\eta c^2}}$ suy ra $\mathbb{P}(A) \leqslant \eta/2$. Lại có:

$$\begin{aligned} B &= \left\{\frac{1}{2}(\psi - \psi_0)^\top H_n(\tilde{\psi})(\psi - \psi_0) \leqslant c\varepsilon^2, \text{ với } \psi \in \partial N_n(\varepsilon)\right\}, \\ &= \left\{\frac{1}{2} u_n(\psi)^\top F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}} u_n(\psi) \leqslant c, \text{ với } \psi \in \partial N_n(\varepsilon)\right\}, \\ &\subseteq \left\{\frac{1}{2}\lambda_{\min}\left(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}\right) u_n(\psi)^\top u_n(\psi) \leqslant c, \text{ với } \psi \in \partial N_n(\varepsilon)\right\}, \\ &= \left\{\frac{1}{2}\lambda_{\min}\left(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}\right) \leqslant c, \text{ với } \psi \in \partial N_n(\varepsilon)\right\}. \end{aligned}$$

nên $\mathbb{P}(B) \leqslant \mathbb{P}(\exists\, \psi \in \partial N_n(\varepsilon) : \lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}) \leqslant 2c)$. Theo Bổ đề 3.1 ở trên, $F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}}$ hội tụ đều theo xác suất tới I_k trong $\psi \in N_n(\varepsilon)$, khi $n \to \infty$. Do đó, theo [17], $\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}})$ hội tụ đều theo xác suất tới 1 trong $\psi \in N_n(\varepsilon)$, khi $n \to \infty$.
Nếu $\tilde{\psi} = a\psi + (1-a)\psi_0$ $(0 \leqslant a \leqslant 1)$ và $\psi \in N_n(\varepsilon)$ thì

$$\begin{aligned} \|F_n^{\frac{1}{2}}(\tilde{\psi} - \psi_0)\| &= \|F_n^{\frac{1}{2}}(a\psi + (1-a)\psi_0 - \psi_0)\| = a\|F_n^{\frac{1}{2}}(\psi - \psi_0)\|, \\ &\leqslant \|F_n^{\frac{1}{2}}(\psi - \psi_0)\| \leqslant \varepsilon, \end{aligned}$$

vì vậy $\tilde{\psi} \in N_n(\varepsilon)$. Từ đó suy ra $\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}})$ hội tụ theo xác suất tới 1 khi $n \to \infty$, vì

$$|\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}) - 1| \leqslant \sup_{\psi \in N_n(\varepsilon)} |\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}}) - 1|.$$

Do đó, với n đủ lớn (ví dụ, $n \geqslant n_1$), $\mathbb{P}(\exists\, \psi \in \partial N_n(\varepsilon)$ sao cho $\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}) \leqslant 2c) \leqslant \eta/2$, vì $2c < 1$. Điều này suy ra $\mathbb{P}(B) \leqslant \eta/2$. Từ đó,

$$\mathbb{P}\left(\ell_n(\psi) - \ell_n(\psi_0) \geqslant 0, \text{ với } \psi \in \partial N_n(\varepsilon)\right) \leqslant \mathbb{P}(A) + \mathbb{P}(B) \leqslant \eta,$$

Vậy đã chứng minh (7), tức là tồn tại duy nhất cực đại toàn cục của ℓ_n trên $N_n(\varepsilon)$.

b) Trở lại với tính vững của $\hat{\psi}_n$. Ta có:

$$\begin{aligned} \lambda_{\min}(F_n)\|\hat{\psi}_n - \psi_0\|^2 &= (\hat{\psi}_n - \psi_0)^\top \lambda_{\min}(F_n) I_k (\hat{\psi}_n - \psi_0), \\ &\leqslant (\hat{\psi}_n - \psi_0)^\top F_n (\hat{\psi}_n - \psi_0), \\ &= \|F_n^{\frac{1}{2}}(\hat{\psi}_n - \psi_0)\|^2 \leqslant \varepsilon^2, \end{aligned}$$

với xác suất dần tới 1 khi $n \to \infty$, theo *a*). Từ điều kiện D3, $\lambda_{\min}(F_n)$ dần tới ∞ khi $n \to \infty$. Do đó $\|\hat{\psi}_n - \psi_0\|$ hội tụ tới 0 với xác suất dần tới 1 khi $n \to \infty$, dẫn tới điều phải chứng minh.

Kết quả thứ hai là:

Định lý 3.2 (Tiệm cận chuẩn). *Giả sử các điều kiện D1-D4 đúng. Khi đó* $F_n^{\frac{1}{2}}(\hat{\psi}_n - \psi_0)$ *hội tụ theo phân phối tới vector Gaussian* $\mathcal{N}(0, I_k)$, *khi* $n \to \infty$.

Chứng minh. Chứng minh của chúng tôi dựa theo chứng minh tiệm cận chuẩn của MLE trong hồi qui Poisson tổng quát giãn nở số không không kiểm duyệt [18]. Tuy nhiên, tác giả sử dụng điều kiện định lí giới hạn trung tâm Lyapunov, còn chúng tôi dựa vào điều kiện yếu hơn Lindeberg, điều này mang lại chứng minh ngắn hơn nhiều.

Trước hết, ta chứng minh tính tiệm cận chuẩn của vector chuẩn hóa $F_n^{-\frac{1}{2}} S_n$, ở đây S_n kí hiệu cho $S_n(\psi_0)$. Đặt u là vector bất kì trong $\mathbb{R}^k$, ta chỉ ra rằng $u^\top F_n^{-\frac{1}{2}} S_n$ hội tụ theo phân phối tới $\mathcal{N}(0, u^\top u)$ (không mất tính tổng quát, ta giả sử $\|u\| = 1$). Từ (5) và (6), ta chú ý rằng S_n có thể viết lại thành một tổng các vector ngẫu nhiên độc lập k-chiều $S_{n,i} = (S_{n,i,1}, \ldots, S_{n,i,k})^\top$, $S_n = \sum_{i=1}^n S_{n,i}$. Dễ thấy dưới điều kiện D1, D2 và D4, các thành phần của $S_{n,i}$ bị chặn bởi đại lượng dương không đổi c_2 nào đó, tức là, $|S_{n,i,\ell}| < c_2$, $\ell = 1, \ldots, k$. Do đó, $\|S_{n,i}\|^2 < c_3 := k c_2^2$.

Đặt

$$u^\top F_n^{-\frac{1}{2}} S_n = u^\top F_n^{-\frac{1}{2}} \sum_{i=1}^n S_{n,i} := \sum_{i=1}^n S_{n,i}^*.$$

Khi đó $\mathbb{E}(S_{n,i}^*) = 0$ và $\text{var}(\sum_{i=1}^n S_{n,i}^*) = 1$. Bây giờ, ta sẽ xác nhận điều kiện Lindeberg, cụ thể:

$$\text{với } \varepsilon > 0, \sum_{i=1}^n \mathbb{E}\left(S_{n,i}^{*2} 1_{\{|S_{n,i}^*| > \varepsilon\}}\right) \to 0 \text{ khi } n \to \infty.$$

Xét $\varepsilon > 0$, ta có:

$$\sum_{i=1}^n \mathbb{E}\left(S_{n,i}^{*2} 1_{\{|S_{n,i}^*| > \varepsilon\}}\right) \leqslant \sum_{i=1}^n \mathbb{E}\left(\|u\|^2 \|F_n^{-\frac{1}{2}}\|^2 \|S_{n,i}\|^2 1_{\{|S_{n,i}^*| > \varepsilon\}}\right) \leqslant \frac{c_1 c_3}{n} \sum_{i=1}^n \mathbb{E}(1_{\{|S_{n,i}^*| > \varepsilon\}}),$$

do điều kiện D3. Vì $\{|S_{n,i}^*| > \varepsilon\}$ suy ra rằng $\{\lambda_{\min}(F_n) < c_3/\varepsilon^2\}$, do đó, $1_{\{|S_{n,i}^*| > \varepsilon\}} \leqslant 1_{\{\lambda_{\min}(F_n) < c_3/\varepsilon^2\}}$ vì vậy,

$$\sum_{i=1}^n \mathbb{E}\left(S_{n,i}^{*2} 1_{\{|S_{n,i}^*| > \varepsilon\}}\right) \leqslant \frac{c_1 c_3}{n} \sum_{i=1}^n 1_{\{\lambda_{\min}(F_n) < c_3/\varepsilon^2\}} = c_1 c_3 1_{\{\lambda_{\min}(F_n) < c_3/\varepsilon^2\}}.$$

Từ điều kiện D3 suy ra, $\lambda_{\min}(F_n) \to \infty$ khi $n \to \infty$ nên $\sum_{i=1}^n \mathbb{E}(S_{n,i}^{*2} 1_{\{|S_{n,i}^*| > \varepsilon\}}) \to 0$ khi $n \to \infty$. Từ đó suy ra với mỗi $u \in \mathbb{R}^k$, $u^\top F_n^{-\frac{1}{2}} S_n$ hội tụ theo phân phối tới $\mathcal{N}(0,1)$, theo định lí Cramer-Wold, $F_n^{-\frac{1}{2}} S_n$ hội tụ theo phân phối tới $\mathcal{N}(0, I_k)$.

Tính hội tụ yếu của $F_n^{\frac{1}{2}}(\hat{\psi}_n - \psi_0)$ có được bằng cách khai triển $S_n := S_n(\psi_0)$ quanh $\hat{\psi}_n$. Phần còn lại của chứng minh tương tự như Định lí 3 của [15] do đó bỏ qua.

4. Kết luận

Bài báo này nghiên cứu ước lượng hợp lí cực đại trong mô hình hồi qui giãn nở số không khi biến tiên lượng kiểm duyệt ngẫu nhiên bên phải. Chúng tôi thiết lập tính vững và tiệm cận chuẩn của đại lượng MLE cho mô hình bằng các chứng minh chặt chẽ.

Từ kết quả bài báo, một số vấn đề có thể tiến hành trong tương lai như nghiên cứu số và áp dụng cho số liệu thống kê y tế công cộng của Việt Nam. Ngoài ra ước lượng hồi quy khi biến tiên lượng kiểm duyệt ngẫu nhiên bên trái, hai bên hay tổng quát hơn như mô hình giãn nở số không nửa tham số cũng đáng được quan tâm .... Các vấn đề này sẽ là công việc tiếp theo của chúng tôi.

Lời cảm ơn

Nghiên cứu này được thực hiện dưới sự tài trợ bởi trường Đại học Hàng hải Việt Nam trong đề tài mã số: DT20-21.91.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] P. McCullagh and J.A. Nelder, *Generalized linear models (Second edition).Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1989.

[2] D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1-14, 1992.

[3] E. Dietz and Bohning, "On estimation of the Poisson parameter in zero-modified Poisson models, " *Computational Statistics & Data Analysis* , vol. 34, no. 4, pp. 441-459, 2000.

[4] H. K. Lim, W. K. Li, and P. L.H. Yu,"Zero-inflated Poisson regression mixture model," *Computational Statistics & Data Analysis*, vol. 71, pp.151-158, 2014.

[5] A. Monod, "Random effects modeling and the zero-inflated Poisson distribution,"*Communications in Statistics. Theory and Methods*, vol. 43, no. 4, pp. 664-680, 2014.

[6] D. B. Hall, "Zero-inflated Poisson and binomial regression with random effects: a case study," *Biometrics*, vol. 56, no. 4, pp. 1030-1039, 2000.

[7] Y. Min and A. Agresti, "Random effect models for repeated measures of zero-inflated count data," *Statistical Modelling*, vol. 5, no. 1, pp. 1-19, 2005.

[8] K. F. Lam, H. Xue, and Y. B. Cheung, "Semiparametric analysis of zero-inflated count data,"*Biometrics*, vol. 62, no. 4, pp. 996-1003, 2006.

[9] J. Feng, and Z. Zhu, "Semiparametric analysis of longitudinal zero-inflated count data," *Journal of Multivariate Analysis*, vol. 102, no. 1, pp. 61-72, 2011.

[10] M. Ridout, J. Hinde, and C. G. B. Demetrio, "A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives,"*Biometrics*, vol. 57, no. 1, pp. 219-223, 2001.

[11] A. Moghimbeigi, M. R. Eshraghian,K. Mohammad, and B. McArdle, "Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros,"*Journal of Applied Statistics*, vol. 35, no. 9, pp. 1193-1202, 2008.

[12] S. M. Mwalili, E. Lesaffre, and D. Declerck, "The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research," *Statistical Methods in Medical Research*, vol. 17, no. 2, pp. 123-139, 2008.

[13] S. E. Saffari, and R. Adnan, "Zero-inflated Poisson regression models with right censored count data," *Matematika*, vol. 27, no. 1, pp. 21-29, 2001.

[14] C. Czado, V. Erhardt, A. Min, and S. Wagner, "Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates," *Statistical Modelling*, vol. 7, no. 2, pp.125-153, 2007.

[15] L. Fahrmeir and H. Kaufmann, "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models,"*The Annals of Statistics*, vol. 13, no. 1, pp. 342-368, 1985.

[16] G. A.F.Seber and A. J. Lee, *Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 2012.

[17] R. A. Maller, "Asymptotics of regressions with stationary and nonstationary residuals," *Stochastic Processes and their Applications*, vol. 105, no. 1, pp. 33-67, 2003.

[18] C. Czado and A. Min, "Consistency and asymptotic normality of the maximum likelihood estimator in a zero-inflated generalized Poisson regression," Collaborative Research Center 386, Discussion Paper 423 , *Ludwig-Maximilians-Universität, München*, 2005.