ỨNG DỤNG PHÂN PHỐI XÁC SUẤT ỔN ĐỊNH PHÂN TÍCH CÁC NHIỄU NGẪU NHIÊN KHÔNG TUÂN THEO LUẬT CHUẨN TRONG XỬ LÝ SỐ LIỆU THỰC NGHIỆM

APPLICATION THE STABLE PROBABILITY DISTRIBUTION FOR RANDOM ERRORS THAT ITS NON-GAUSS IN THE PROCESSING OF EMPIRICAL DATA

Trần Chí Lê

Khoa Khoa học cơ bản, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp

Đến Tòa soạn ngày 02/03/2021, chấp nhận đăng ngày 25/03/2021

Tóm tắt: Khi xét mô hình hồi quy trong xử lý số liệu thực nghiệm, thường đi kèm giả thiết các sai số (nhiễu ngẫu nhiên) tuân theo luật chuẩn, và phân tích mô hình đó bằng T-test và F-test. Trong trường hợp sai số không tuân theo luật chuẩn thì các phân tích trước đó sẽ cho kết quả không chuẩn xác. Bài báo này giới thiệu một lớp phân phối ổn định, phân phối mở rộng của phân phối chuẩn, rất phù hợp để phân tích sai số không tuân theo luật chuẩn. Phương pháp phân tích theo phân phối này cho kết quả chính xác hơn thông qua các kiểm định Kolmogorov-Smirnov và mô hình Bayesian trung bình, các kết quả phân tích được trình bày thông qua các gói lệnh và mã lập trình trên phần mềm xử lý số liệu R.

Từ khóa: Ổn định, Bayesian, R.

Abstract: When looking at the regression model in the processing of empirical data, it is often accompanied by the assumption of the error (residuals) following the normal law, and analyzing the model by T-test and F-test. Incase that its non-gauss then the previous analysis will give inaccurate results. This paper introduces the stable distribution, an extended distribution of the normal distribution that is well suited for analysis the residuals that its non-gauss. This method gives more accurate results by using the Kolmogorov - Smirnov test and the Bayesian model average, and the analysis results are presented through packages and prgramming code on the software R.

Keywords: Stable, Bayesian, R.

1. ĐẶT VẤN ĐỀ

Giả sử cần nghiên cứu một đại lượng *y* trong một hệ thống nào đó. Trong hệ thống ấy, y phụ thuộc vào hai nhóm yếu tố: nhóm yếu tố thứ nhất là các yếu tố độc lập $x_1, x_2, .., x_k$ có thể điều khiển được; nhóm yếu tố thứ hai là nhóm yếu tố ngẫu nhiên không điều khiển được, đại diện bởi biến ngẫu nhiên ξ. Các biến $x_1, x_2, .., x_k$ gọi là các biến vào hay các nhân tố; biến ngẫu nhiên ξ gọi là nhiễu hoặc sai số ngẫu nhiên; *y* gọi là biến ra. Vấn đề là phải tìm ra quan hệ giữa *y* và $(x_1, x_2, .., x_k)$. Giả sử mối quan hệ giữa *y* và $(x_1, x_2, .., x_k)$ có dạng:

$$y = f(x_1, x_2, .., x_k; \theta_1, \theta_2, ..., \theta_m) + \xi, \qquad (1)$$

trong đó dạng hàm f đã biết, và *m* tham số θ_i chưa biết. Thông thường, chúng ta giả thuyết nhiễu $\xi \in N(0; \sigma^2)$, khi đó các bài toán ước lượng các tham số chưa biết $\theta_1, \theta_2, ..., \theta_m$, và

kiểm định, đánh giá mô hình (1) sẽ được tiến hành theo định lý giới hạn trung tâm, xem [3].

Vấn đề đặt ra là, nếu nhiễu $\xi \notin N(0;\sigma^2)$ thì các đánh giá theo hướng cũ không còn phù hợp, thậm chí có thể dẫn đến kết quả sai lệch rất lớn. Vì vậy, để khắc phục nhược điểm này, bài báo giới thiệu đến các ứng dụng của phân phối ổn định, một phân phối phù hợp để phân tích số liệu thực nghiệm trong trường hợp nhiễu không tuân theo luật chuẩn.

Cấu trúc bài báo được trình bày ở các phần tiếp theo như sau. Phần 2 sẽ giới thiệu về phân phối xác suất ổn định và các trường hợp suy biến hay gặp. Phần 3 sẽ xây dựng cơ sở ứng dụng khi phân tích nhiễu không tuân theo luật chuẩn trong xử lý số liệu thực nghiệm, và lấy ví dụ minh họa cũng như so sánh kết quả khi phân tích với giả thuyết nhiễu tuân theo luật chuẩn. Kết luận và các vấn đề ứng dụng sẽ được đưa ra trong Phần 4.

2. PHÂN PHỐI XÁC SUẤT ỔN ĐỊNH VÀ ĐỊNH LÝ GIỚI HẠN TRUNG TÂM SUY RỘNG

2.1. Phân phối xác suất ổn định 1-chiều

Định nghĩa: Cho hai biến ngẫu nhiên độc lập ξ_1; ξ_2 có cùng phân phối với biến ngẫu nhiên ξ. Khi đó, biến ngẫu nhiên ξ được gọi là tuân theo phân phối xác suất ổn định nếu $\forall a_1, a_2 > 0$ luôn tồn tại $c > 0; b \in R$ sao cho:

$$a_1\xi_1 + a_2\xi_2 \overset{d}{=} c\xi + b \qquad (2)$$

ở đây $\overset{d}{=}$ được hiểu là bằng nhau theo phân phối xác suất.

Nếu $b = 0$ ta nói ξ có phân phối hoàn toàn ổn định; nếu $-\xi$ có cùng phân phối với ξ, thì ta nói ξ có phân phối ổn định đối xứng. Hơn nữa, với ξ có phân phối ổn định, luôn tồn tại một số thực $\alpha \in (0;2]$ sao cho: $a_1^\alpha + a_2^\alpha = c^\alpha$, khi đó α được gọi là chỉ số đặc trưng mũ của phân phối ổn định.

Dựa vào biểu diễn Lévy – Khintchine, hàm đặc trưng của phân phối ổn định được xây dựng theo định lý sau, xem [4].

Định lý 1: Hàm đặc trưng của biến ngẫu nhiên ξ tuân theo phân phối xác suất ổn định có dạng:

$$\varphi(t) = \begin{cases} \exp\left\{-\sigma^\alpha.|t|^\alpha.\left[1 - i.\beta.sign(t).\tan\frac{\pi\alpha}{2}\right] + i\mu t\right\} & khi\ \alpha \neq 1 \\ \exp\left\{-\sigma.|t|\ .\left[1 + i.\beta.\frac{2}{\pi}sign(t).\ln|t|\right] + i\mu t\right\} & khi\ \alpha = 1 \end{cases} \qquad (3)$$

trong đó $\alpha \in (0;2]$; $\beta \in [-1;1]$; $\sigma > 0; \mu \in \mathbb{R}$.

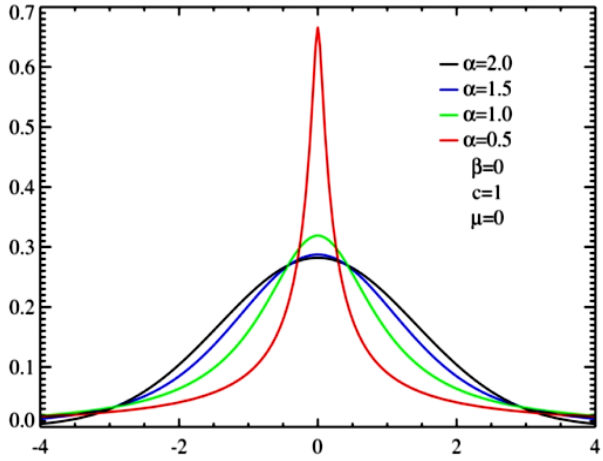
Họ phân phối xác suất ổn định được kí hiệu là $\Gamma(\alpha;\beta;\sigma;\mu)$ phụ thuộc vào 4 tham số đặc trưng, trong đó $\alpha \in (0;2]$ đại diện cho đặc trưng mũ; $\beta \in [-1;1]$ đại diện cho tham số độ lệch (khi $\beta > 0$ mật độ xác suất ở đuôi phải lớn hơn mật độ ở đuôi trái, khi $\beta < 0$ thì mật độ đuôi trái lớn hơn mật độ đuôi phải); $\sigma > 0$ đặc trưng cho tỷ lệ; $\mu \in \mathbb{R}$ đặc trưng cho vị trí.

Tính chất: Với ξ là biến ngẫu nhiên tuân theo phân phối ổn định, khi đó với $1 < \alpha \leq 2$ thì $E(\xi) = \mu$; với $0 < p < \alpha$ thì $E|\xi|^p < +\infty$; với $p > \alpha$ thì $E|\xi|^p = +\infty$, xem [4].

2.2. Một số trường hợp suy biến của phân phối ổn định

Mục này bài báo giới thiệu một số dạng suy biến của phân phối ổn định về các phân phối đã biết như: phân phối chuẩn; phân phối cauchy; phân phối lévy, cụ thể như sau, xem [1]:

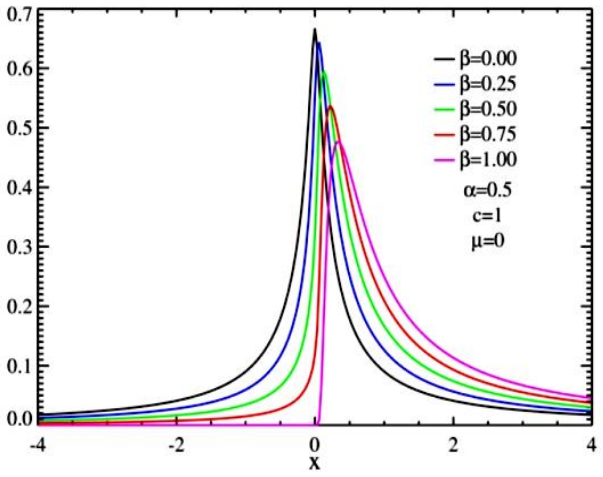
- Khi tham số α thay đổi đồ thị của hàm đặc trưng có dạng như hình 1.



Hình 1. Đồ thị khi tham số α thay đổi

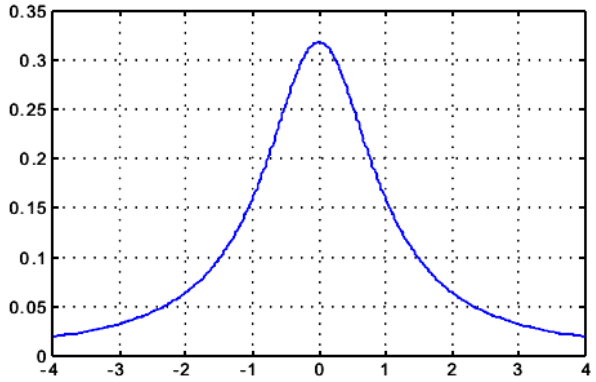
Với tham số α thể hiện tính chất "đuôi dài". Khi α càng nhỏ, phân bố có xu hướng tiệm cận về 0 lâu hơn. Trường hợp đặc biệt, tham số $\alpha = 2$ thì $\Gamma(\alpha;\beta;\sigma;\mu)$ sẽ trùng với phân phối chuẩn với $E(\xi) = \mu$; $D(\xi) = 2\sigma^2$ và $\varphi(t) = \exp\left\{i\mu t - \sigma^2 t^2\right\}$.

- Trong trường hợp tham số β thay đổi đồ thị của hàm đặc trưng có dạng như hình 2:



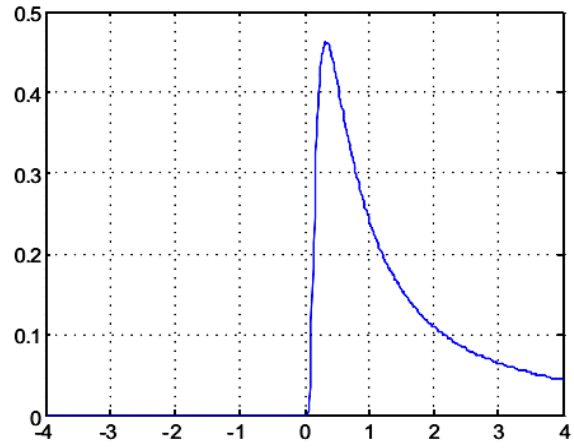
Hình 2. Đồ thị khi tham số β thay đổi

- Nếu tham số $\alpha = 1;\ \beta = 0$ thì $\Gamma(\alpha;\beta;\sigma;\mu)$ sẽ suy biến về phân phối Cauchy với hàm đặc trưng $\varphi(t) = \exp\left\{i\mu t - \sigma|t|\right\}$, trong đó tham số tỷ lệ $\sigma > 0$ và tham số vị trí $\mu \in \mathbb{R}$, xem minh họa hình 3.



Hình 3. Đồ thị khi tham số $\alpha = 1;\ \beta = 0$.

- Nếu tham số $\alpha = 1/2;\ \beta = \pm 1$ thì $\Gamma(\alpha;\beta;\sigma;\mu)$ sẽ trùng với phân phối Lévy với tham số vị trí μ; tham số tỷ lệ σ và hàm đặc trưng có dạng $\varphi(t) = \exp\left\{-\sqrt{\sigma.|t|}\left[1 - i.sign(t)\right] + i\mu t\right\}$, xem hình 4.



Hình 4. Đồ thị khi tham số $\alpha = 1/2;\ \beta = 1$.

2.3. Định lý giới hạn trung tâm suy rộng

Một trong những kết quả quan trọng nhất của lý thuyết xác suất là kết quả về luật phân phối của tổng n - biến ngẫu nhiên ξ_i. Đại ý rằng: Nếu $\{\xi_n\}$ là dãy các biến ngẫu nhiên độc lập có cùng phân phối với kỳ vọng $E(\xi_n) = \mu$ và phương sai $D(\xi_n) = \sigma^2 < +\infty$ hữu hạn, thì tổng của n-biến ngẫu nhiên $S_n = \xi_1 + \xi_2 + ... + \xi_n$ sẽ có phân phối xấp xỉ chuẩn. Một vấn đề đặt ra là, nếu tổng trên mà phương sai $D(\xi_n)$ không hữu hạn, thì tổng S_n

có phân phối như thế nào? Định lý giới hạn trung tâm suy rộng sau đây sẽ trả lời cho câu hỏi này, xem chi tiết trong [4].

Định lý 2: Nếu $\xi_n \in \Gamma(\alpha;\beta;\sigma;\mu);\ i=1..n$ là dãy các biến ngẫu nhiên độc lập thì:

$$\frac{\xi_1+\xi_2+...+\xi_n}{n} \in \Gamma(\alpha;\beta^*;\sigma^*;\mu^*) \qquad (4)$$

với các tham số $\beta^* = \beta.n^{1-\alpha};\ \sigma^* = \sigma.n^{\frac{1-\alpha}{\alpha}}$; và

$$\mu^* = \begin{cases} \mu & khi\ \alpha \neq 1; \\ \mu + \frac{2}{\pi}\beta.\sigma.\ln(n) & khi\ \alpha = 1. \end{cases}$$

Việc chứng minh chi tiết định lý trên có thể tìm được trong [4]. Về mặt ứng dụng, chúng ta có thể hiểu đơn giản rằng: tổng của n - biến ngẫu nhiên độc lập cùng phân phối $S_n = \xi_1+\xi_2+...+\xi_n$ sẽ xấp xỉ về phân phối ổn định, nếu điều kiện phương sai hữu hạn không được kiểm chứng.

3. ỨNG DỤNG TRONG PHÂN TÍCH NHIỄU KHÔNG TUÂN THEO LUẬT CHUẨN

3.1. Mô hình Bayesian Model Average (BMA)

BMA là phương pháp tìm tất cả các mô hình hồi quy khả dĩ và lọc ra các mô hình tối ưu, dựa trên đánh giá xác suất ảnh hưởng của các biến và bộ dự liệu mẫu ban đầu. Phương pháp này hiệu quả hơn so với các phương pháp truyền thống (T-test; F-test) khi nhiễu ξ không tuân theo luật chuẩn. Bởi vì, khi nhiễu không tuân theo luật chuẩn thì các kiểm định T-test để đánh giá sự có ý nghĩa của các biến trong mô hình sẽ cho kết quả không được chính xác. Hơn thế nữa, khi nhiễu tuân theo luật chuẩn thì phân tích theo BMA cho kết quả tương đương với các phương pháp truyền thống, xem [5].

Cho $M = (M_1, M_2, ..., M_k)$ là tập tất cả các mô hình được xét. Một mô hình có thể được xác định bởi một hoặc nhiều thuộc tính, chẳng hạn như tập hợp con của các biến giải thích trong mô hình hoặc phân tích phương sai phần dư của mô hình. Nếu Δ là đại lượng cần quan tâm, chẳng hạn một tham số trong mô hình, thì phân phối hậu nghiệm của Δ khi đã có dữ liệu Z được xác định bởi:

$$P(\Delta \mid Z) = \sum_{i=1}^{k} P(\Delta \mid Z, M_i).P(M_i \mid Z) \qquad (5)$$

Trong đó, xác suất hậu nghiệm cho mô hình M_i xét với dữ liệu Z là:

$$P(M_i \mid Z) = \frac{P(Z \mid M_i).P(M_i)}{\sum_{i=1}^{k} P(Z \mid M_i).P(M_i)} \qquad (6)$$

với $P(Z \mid M_i) = \int P(Z \mid \theta_i, M_i).P(\theta_i \mid M_i)d\theta_i$ là hàm hợp lý của mô hình M_i, còn θ_i là vectơ các tham số của mô hình M_i, và $P(\theta_i \mid M_i)$ là mật độ tiên nghiệm của các tham số xét trên mô hình M_i, xem [5].

Khi xét với dữ liệu Z và mỗi mô hình cụ thể M_i, ta sẽ tính được xác suất khả dĩ cho mỗi mô hình đó theo công thức (6), hơn nữa ta có thể tính được xác suất ảnh hưởng của các biến trong mỗi mô hình đang xét theo công thức (5). Khi đó, ta sẽ ưu tiên chọn mô hình khả dĩ nhất (mô hình có xác suất hậu nghiệm lớn nhất) và xác định được những biến nào có ảnh hưởng; những biến nào không ảnh hưởng trong mô hình khả dĩ đó.

Việc sử dụng phương pháp BMA khi xử lý số liệu, được thực hiện với các gói lệnh (packages) trên các phần mềm thống kê (***SPSS, MATLAB, R…***) sẽ cho kết quả thuận tiện hơn. Cụ thể, trong bài báo này tác giả sẽ sử dụng các gói lệnh cũng như các mã (code) lập trình trên phần mềm ***R,*** xem [2].

Gói lệnh phân tích theo mô hình BMA trên phần mềm ***R*** như sau:

```
#nhập BMA vào môi trường R
> library(BMA)
# nhập dữ liệu các biến độc lập
> z=data.frame(x1,x2,…,xn)
# nhập dữ liệu cho biến phụ thuộc
> y=c(y1,y2,…,yn)
# phân tích BMA
> BMA=bicreg(z,y,trict=FALSE,OOR=20)
> summary(BMA)
```

3.2. Tiêu chuẩn Kolmogorov – Smirnov kiểm định giả thuyết về phân phối

Kiểm định Kolmogorov-Smirnov: là kiểm định phi tham số đối với các phân phối xác suất nhận giá trị liên tục. Kiểm định này sử dụng để so sánh phân phối của một mẫu với một phân phối xác suất cho trước, thông qua khoảng cách giữa hàm phân phối thực nghiệm của mẫu với hàm phân phối tích lũy của phân phối cần so sánh.

Giả sử $X_1, X_2, .., X_n$ là các quan sát độc lập cùng phân phối với hàm phân phối tích lũy F, xét bài toán kiểm định giả thuyết $H_0 : F = F_0$ và đối thuyết $H_1 : F \neq F_0$. Tiêu chuẩn kiểm định Kolmogorov-Smirnov được xác định bởi thống kê sau:

$$T_n = \sup_{t \in R} \left(n^{1/2} . | F_n - F_0 | \right)$$

trong đó $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty;t]}(X_i)$ là hàm phân phối thực nghiệm của mẫu $X_1, X_2, .., X_n$, với $I_{[-\infty;t]}$ là hàm chỉ số được cho bởi: $I_{[-\infty;t]}(X_i) = 1 \; khi \; X_i < t$.

$\& I_{[-\infty;t]}(X_i) = 0 \; khi \; X_i \geq t$; và F_0 là phân phối xác suất cho trước (như phân phối chuẩn, phân phối ổn định,…). Gọi T_{1-p} là phân vị mức $1-p$ của phân phối T_n, khi đó ta sẽ bác bỏ giả thuyết nếu $T_n > T_{1-p}$, hoặc P-value < 0.05, xem [1].

Gói lệnh để kiểm định tiêu chuẩn Kolmogorov-Smirnov trên phần mềm ***R*** như sau:

```
#chèn gói lệnh kolmim và gói lệnh stable
> library(kolmim)
>library(stable)
# Ước lượng các tham số của sai số
>E=Yi-Yi^
>mean(E)
>sd(E)
> stable_mle_fit(E)
>beta=
>alpha=
# nhập quy luật phân phối y cần so sánh
> y=norm or stable
#nhập các ước lượng của tham số tương ứng
>th.so=(mean(E),sd(E))
 or (loc,scale,beta,alpha)
# kiểm định Kolmogorov-Smirnov
> ks.test(x, y,th.so)
```

3.3. Phân tích sai số trong mô hình hồi quy

Xét lại mô hình hồi quy dạng (1): $y = f(x_1, x_2, .., x_k; \theta_1, \theta_2, ..., \theta_m) + \xi$, bằng các phân tích theo phương pháp BMA ta sẽ xác định được các biến tham gia trong mô hình, sau đó lập mô hình hồi quy dựa trên số biến này trong hai trường hợp: trường hợp giả thuyết nhiễu ξ tuân theo luật chuẩn và trường hợp nhiễu tuân theo luật phân phối ổn định. Từ đó tính được sai số $e_i = y_i - y_i$ tương ứng cho nhiễu ξ trong hai trường hợp, với y_i là các giá trị xác định từ dữ liệu mẫu, còn y_i được xác định qua ước tính hàm hồi quy tương ứng.

Gói lệnh lập mô hình hồi quy trong hai trường hợp giả thuyết về nhiễu và tính sai số tương ứng trên phần mềm ***R*** như sau, xem [6]:

```
#phân tích hồi quy với nhiễu tuân theo chuẩn
> reg1=lm(solieu)
>summary(reg1)
#ước tính sai số hồi quy 1
>E1=resid(reg1)
#phân tích hồi quy với nhiễu tuân theo phân
phối ổn định
> reg2=stable_lm(y~z,data=solieu)
>print(reg)
#ước tính sai số 2
>yi=solieu[,1]
>yi^=predit(reg2)
>E2=yi-yi^
```

Với các giá trị sai số thu được, ta tiến hành phân tích mẫu E1 và E2 theo tiêu chuẩn kiểm định Kolmogorov-Smirnov để rút ra các kết luận về nhiễu.

Các bước phân tích được tiến hành như sau:

Bước 1: Phân tích mô hình BMA

Sử dụng phân tích mô hình BMA lựa chọn mô hình tối ưu theo tiêu chí có các biến tham gia với xác suất hậu nghiệm cao nhất, từ đó ta xác định được số lượng các biến tham gia trong mô hình, và lập mô hình hồi quy tương ứng để phân tích theo Bước 2.

Bước 2: Ước tính sai số cho mô hình hồi quy

Với số lượng biến tham gia trong mô hình hồi quy nhận được từ Bước 1, ta lập hai mô hình hồi quy tương ứng với hai trường hợp của nhiễu. Từ đó tính được hai bộ mẫu đặc trưng cho sai số, để tiến hành phân tích theo Bước 3.

Bước 3: Kiểm định Kolmogorov-Smirnov với sai số

Sử dụng tiêu chuẩn kiểm định Kolmogorov-Smirnov kiểm định cho hai bộ mẫu đặc trưng về sai số.

- Với bộ mẫu dựa trên giả thuyết nhiễu tuân theo phân phối chuẩn, nếu kết luận là chấp nhận giả thuyết thì sử dụng các kết quả của mô hình hồi quy tương ứng để phân tích và suy luận, nếu kết luận là bác bỏ thì tiến hành kiểm định trên giả thuyết nhiễu tuân theo phân phối ổn định.

- Với bộ mẫu dựa trên giả thuyết nhiễu tuân theo phân phối ổn định, nếu kết luận là bác bỏ thì cần kiểm tra, sàng lọc lại nguồn lấy mẫu vì sai số trong trường hợp này bị ảnh hưởng bởi yếu tố không ngẫu nhiên gây ra. Nếu kết luận là chấp nhận giả thuyết, thì sử dụng mô hình hồi quy tương ứng để phân tích về: giá trị ước lượng, khoảng ước lượng tin cậy và các tham số của nhiễu ξ, xem [4] .

3.4. Ví dụ minh họa

Xét mối liên hệ tăng giảm (%) so với quy chuẩn của cường độ tín hiệu wifi Y và các yếu tố ảnh hưởng: X1: nguồn phát; X2: khoảng cách; X3: lượng truy cập, với số liệu thí nghiệm thu được từ việc mô phỏng trên phần mềm ***R*** ở bảng 1, và mức ý nghĩa $\alpha = 0,05$ được sử dụng cho tất các kết luận thống kê.

Bảng 1. Số liệu thí nghiệm

N	X1	X2	X3	Y
1	0.09	3.66	1.77	-6.12
2	-1.23	0.52	2.38	8.69
3	-0.72	3.58	2.79	13.43
4	-2.01	-1.12	1.74	11.23
5	4.24	-4.43	1.06	-7.97
6	5.64	-1.65	2.59	-5.43
7	-2.81	1.62	1.03	13.34
8	-4.96	-8.06	3.45	5.34
9	-1.86	4.01	3.57	12.96
10	1.92	-3.68	-5.91	-25.95
11	-2.31	0.47	2.72	13.22
12	-1.96	6.35	4.22	9.02
13	4.62	3.92	3.14	4.98
14	-1.86	-7.52	3.35	7.12
15	0.62	2.81	2.61	11.12
16	2.53	2.98	2.79	9.07
17	1.74	-10.08	3.79	-2.56
18	0.93	-1.61	3.09	6.09
19	1.99	2.77	3.92	50.67
20	0.71	0.74	13.41	41.89
21	-0.59	4.07	0.88	3.24
22	1.41	2.76	2.39	12.49
23	-1.03	6.85	3.33	21.29
24	0.34	-4.49	-0.29	-7.41
25	0.96	-1.68	2.48	3.39
26	-2.1	1.28	1.08	9.92
27	2.81	1.44	2.99	5.96
28	6.12	24.65	3	26.61
29	-5.5	-5.86	2.55	14.21
30	-2.19	13.84	3.08	30.77
31	2.71	3.16	4.11	7.64
32	-1.04	2.53	2.87	13.76
33	0.09	3.66	1.77	-6.12

Bài báo sẽ trình bày các kết quả phân tích của ví dụ này theo các bước trong mục 3.3, cụ thể như sau:

Bước 1: Phân tích mô hình BMA: sử dụng gói lệnh phân tích trên phần mềm *R* theo mục 3.1 ta thu được kết quả phân tích:

Bảng 2. Phân tích BMA

Call: bicreg(x=xvar, y=yv, strict=FALSE, OR=20) 2 models were selected, Best 2 models			
	P!=0	Model1	Model2
Const	100.0	-0.2043	-0.4730
X1	73.4	-1.3288	.
X2	100.0	1.0330	0.8749
X3	100.0	3.2897	3.3589
nVar		3	2
post prob		0.734	0.266

*(Nguồn: kết quả xử lý bằng **R**).*

Từ bảng 2 có hai mô hình khả dĩ là Model1 và Model2, trong đó mô hình Model1 có xác suất đáp ứng là 0.734 (post prob) so với 0.266 của mô hình Model2. Với Model1 có sự tham gia của cả 3 biến (n Var =3) với các xác suất ảnh hưởng lần lượt là: 100%; 73.4%; 100%; 100% (cột thứ 2 trong bảng 2). Vậy bằng phương pháp phân tích BMA ta xác định được mô hình khả dĩ nhất có sự tham gia của cả 3 biến đều có ý nghĩa thống kê.

Bước 2: Ước tính sai số: sử dụng gói lệnh phân tích trên phần mềm *R* theo mục 3.3 ta thu được kết quả phân tích:

▪ Với giả thuyết nhiễu tuân theo quy luật chuẩn, kết quả ước lượng mô hình:

Bảng 3. Mô hình với giả thuyết chuẩn

Y ~ X1 + X2 + X3				
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.20	2.20	-0.09	0.9265
X1	-1.33	0.58	-2.29	0.0296 *
X2	1.03	0.25	4.07	0.0003 ***
X3	3.30	0.58	5.6	5.4e-06 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Multiple R-squared: 0.6693, Adjusted R-squared: 0.6339 F-statistic: 18.89 on 3 and 28 DF, p-value: 6.741e-07				

*(Nguồn: kết quả xử lý bằng **R**).*

Từ bảng 3, ta nhận được mô hình hồi quy tương ứng là:

$$\hat{Y} = -0.20 - 1.33X1 + 1.03X2 + 3.29X3 \qquad (7)$$

Khi đó, tính được sai số $e_i = y_i - y_i$ theo mô hình (7), thu được mẫu:

E1={-15.40; -1.11; -0.20; 4.20; -1.04; -4.55; 4.75; -4.07; -5.20; 0.05; 0.92; -13.82; -3.06; 1.60; 0.66; 0.38; -2.10; -0.97; 37.76; -1.84; -4.44; 3.85; 2.09; -1.16; -1.55; 2.46; -1.43; -0.39; 4.77; 3.64; -5.34; 0.53}

▪ Với giả thuyết nhiễu tuân theo quy luật phân phối ổn định, kết quả ước lượng mô hình:

Bảng 4. Mô hình với giả thuyết ổn định

Y ~ X1 + X2 + X3			
	Estimate	left.conf	right.conf
Intercept	0.71	-0.07	1.50
X1	-2.07	-2.26	-1.88
X2	1.16	1.09	1.22
X3	3.10	2.95	3.24

*(Nguồn: kết quả xử lý bằng **R**).*

Từ bảng 4, ta nhận được mô hình hồi quy tương ứng là:

$$\hat{Y} = 0.71 - 2.07X1 + 1.16X2 + 3.10X3 \qquad (8)$$

khi đó, tính được sai số $e_i = y_i - y_i$ theo mô hình (8), thu được mẫu:

E2={ 16.38; 2.55; 1.57; -2.26; -1.95; 0.58; -1.74; 6.98; 7.32; 0.10; 1.25; 16.20; 0.45; -0.90; -0.34; -1.49; -0.28; 0.41; -38.71; -0.22; 6.14; -4.09; -0.18; 1.31; 1.07; -0.03; -0.13; -0.67; -1.01; 0.08; 3.87; 0.94}

Bước 3: Kiểm định Kolmogorov-Smirnov: sử dụng gói lệnh phân tích trên phần mềm *R* theo mục 3.2 ta thu được kết quả phân tích:

- Với mẫu E1:

Bảng 5. Kiểm định mẫu E1

	One-sample Kolmogorov-Smirnov test
Norm	data: E1 D = 0.24978, p-value = 0.03027 alternative hypothesis: two-sided
Stable	$par alpha beta scale loc 1.38 -0.14 2.30 -0.35
	One-sample Kolmogorov-Smirnov test data: E1 D = 0.068036, p-value = 0.9961 alternative hypothesis: two-sided

*(Nguồn: kết quả xử lý bằng **R**).*

Qua bảng 5 ta có kết quả: Xét với phân phối chuẩn (Norm), kiểm định phân phối nhận được p-value=0,03027<0,05 dẫn tới bác bỏ giả thuyết sai số tuân theo luật chuẩn. Khi xét với phân phối ổn định (Stable) trên bộ tham số ước lượng (alpha=1.38; beta=–0.14; scale=2.30; loc=–0.35), kiểm định phân phối nhận được p-value = 0,9961 > 0,05 dẫn tới giả thuyết nhiễu tuân theo phân phối ổn định được chấp nhận.

- Với mẫu E2:

Bảng 6. Kiểm định mẫu E2

	One-sample Kolmogorov-Smirnov test
Stable	$par alpha beta scale loc 1.1 0.39 1.21 0.03
	One-sample Kolmogorov-Smirnov test data: E2 D = 0.17584, p-value = 0.2454 alternative hypothesis: two-sided

*(Nguồn: kết quả xử lý bằng **R**).*

Qua bảng 6 ta có kết quả: Xét với phân phối ổn định (Stable) trên bộ tham số ước lượng (alpha=1.1; beta=0.39; scale=1.21; loc=0.03), kiểm định phân phối nhận được p-value = 0,2454 > 0,05 dẫn tới chấp nhận giả thuyết nhiễu tuân theo phân phối ổn định.

Nhận xét: Từ ví dụ trên ta thấy với mô hình (7) thu được từ Bảng 3, các kết quả kiểm định mô hình thông qua F-test=18,89 (p-value rất nhỏ) đều cho kết luận là phù hợp để diễn tả mối liên hệ giữa các nhân tố X1, X2, X3 và Y. Nhưng giả thuyết nhiễu tuân theo luật chuẩn đã bị bác bỏ, nên các kết quả này không còn được chính xác, dẫn tới mô hình (7) không có ý nghĩa thống kê. Với mô hình hồi quy (8) thông qua các kiểm định ở bảng 5 và 6 đều cho kết quả là mô hình phù hợp hơn và mô tả chính xác hơn mối liên hệ giữa các nhân tố X1, X2, X3 và Y. Hơn nữa qua bảng 4 còn nhận được các ước lượng của nhiễu $\xi \in \Gamma(\alpha = 1.10;\ \beta = -0.39;\ \sigma = 1.21;\ \mu = 0.03)$. Từ ước lượng này, việc sử dụng mô hình (8) để ước lượng, dự đoán hiệu suất cường độ tín hiệu Wifi sẽ cho kết quả chính xác, có ý nghĩa thống kê hơn so với với mô hình (7).

4. KẾT LUẬN

Trong bài toán phân tích nhiễu của mô hình hồi quy trong xử lý số liệu thực nghiệm, bài báo đã giới thiệu một phương pháp phân tích mới dựa theo luật phân phối xác suất ổn định, phương pháp cho kết quả chính xác hơn trong trường hợp nhiễu không tuân theo luật phân phối chuẩn. Với phương pháp này, dựa trên phân tích BMA để lựa chọn số biến ảnh hưởng trong mô hình và kiểm định Kolmogorov-Smirnov trong giả thuyết nhiễu tuân theo phân phối chuẩn, cùng với ước lượng các hệ số của các nhân tố và các tham số của nhiễu sẽ cung cấp cho chúng ta mô

hình ước lượng có ý nghĩa thống kê so với phương pháp truyền thống.

Nhược điểm của việc phân tích nhiễu theo phân phối ổn định là các công thức xây được xây dựng và tính toán dựa trên cơ sở của lý thuyết xác suất chuyên ngành (như hàm đặc trưng, định lý giới hạn trung tâm suy rộng, xác suất hậu nghiệm...), dẫn đến khi thực nghiệm trên mẫu số liệu cụ thể sẽ gặp khó khăn về khối lượng và thời gian tính toán. Tất cả các nhược điểm này, gần đây đã được khắc phục triệt để dựa trên các gói lệnh và các mã lập trình mở trên các phần mềm thống kê, như phần mềm **R** chẳng hạn**.** Đặc biệt, trong bài báo này gói lệnh phân tích hồi quy với phân phối ổn định stabreg được cập nhật mới nhất ngày 06/06/2019 bởi hai tác giác: Oleg Kopylow-Sebastian Ament cho kết quả chính xác và rút bớt thời gian tính toán hơn rất nhiều.

Với phương pháp phân tích nhiễu tuân theo phân phối ổn định này, bài báo sẽ bổ sung thêm một phương pháp mới trong các phương pháp phân tích sai số (nhiễu) của mô hình hồi quy ở các tài liệu giảng dạy xử lý số liệu thực nghiệm trong các trường đại học, đặc biệt là tài liệu xử lý số liệu thực nghiệm ở Trường Đại học Kinh tế - Kỹ thuật Công nghiệp, giảng dạy cho học viên cao học. Qua đó, giúp học viên cập nhật thêm công cụ mới, để xử lý các tình huống gặp phải trong quá trình học tập chuyên ngành, cũng như khi làm luận văn và đề tài nghiên cứu khoa học liên quan tới tính toán số liệu thực nghiệm.

TÀI LIỆU THAM KHẢO

[1] Bùi Quảng Nam, Vũ Đình Ba, Hồ Đăng Phúc *"Vận dụng phân phối xác suất ổn định vào phân tích tín hiệu GPS",* Tạp chí nghiên cứu Khoa học & Công nghệ quân sự, số 39, trang 90-96, (2015).

[2] Nguyễn Văn Tuấn, *"Phân tích dữ liệu với R",* NXB Tổng hợp TP Hồ Chí Minh, 2020.

[3] Trần Chí Lê, "Nghiên cứu ứng dụng phương pháp P-giá trị cho bài toán kiểm định sự phù hợp của mô hình hồi quy thông qua hệ số xác định hiệu chỉnh R2 trong xử lý số liệu thực nghiệm", Tạp chí Khoa học & Công nghệ - Trường Đại học Kinh tế - Kỹ thuật Công nghiệp, số 22, trang 91-96 (2020).

[4] Nolan J, "*Stable Distributions Models for Heavy Tailed Data*" American University, W.D.C (2005).

[5] Liang, F.M., Troung, Y., and Wong, W.H. "Automatic Bayesian model averaging for linear regession and applications in Bayesian curve fitting", Statistica Sinaca, 2001.

[6] Oleg Kopylow-Sebastian Ament, (2019). Package *"stabreg",* https://cran.r-project.org/web/packages/stabreg/stabreg.pdf

Thông tin liên hệ: **Trần Chí Lê**

Điện thoại: 0912954359 - Email: tcle@uneti.edu.vn

Khoa Khoa học cơ bản, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.