Tạp chí Phát triển Khoa học và Công nghệ – Kinh tế-Luật và Quản lý, 5(1):1325-1339

Open Access Full Text Article

Bài nghiên cứu

Sự tương đồng về phân phối xác suất của các biến trong mô hình hồi quy Bayes và ứng dụng

Lê Thanh Hoa[1,*], Phạm Hoàng Uyên[1], Nguyễn Thị Đỗ An[2], Phạm Thế Bảo[3]

Use your smartphone to scan this QR code and download this article

TÓM TẮT

Mô hình hồi quy tuyến tính cũng như mô hình chuỗi thời gian được áp dụng trong nhiều lĩnh vực, trong đó trung bình của biến phụ thuộc là một hàm số của trung bình các biến độc lập. Tuy nhiên, khi xem xét mô hình hồi quy theo phương pháp thống kê cổ điển (thống kê tần suất), tức là các tham số là hằng số, trong nhiều tình huống mô hình hồi quy không mô tả đúng sự biến động của đồng thời biến phụ thuộc và biến độc lập. Bởi vậy, chúng ta cần hiệu chỉnh các tham số không còn dưới dạng hằng số mà dưới dạng biến ngẫu nhiên như mô hình hồi quy trong thống kê Bayes. Mặt khác, khi xem xét các tham số như một biến ngẫu nhiên, các tính toán trong mô hình hồi quy trở nên vô cùng phức tạp, bởi vì chúng ta cần tính toán tích của các phân phối xác suất. Chính vì vậy, chúng ta phải có các đánh giá về sự đa dạng về phân phối xác suất của các biến trong mô hình hồi quy, chứ không chỉ đơn thuần về dạng phân phối như phân phối chuẩn, phân phối Student *t*, phân phối Poisson, phân phối nhị thức... Trong bài báo này, chúng tôi ước lượng dạng phân phối xác suất của biến phụ thuộc trong mô hình hồi quy Bayes đơn trong một số trường hợp thay đổi dạng phân phối xác suất của biến độc lập. Bên cạnh đó, chúng tôi ứng dụng kết quả với dữ liệu giá chứng khoán thực, minh chứng dạng phân phối xác suất phù hợp nhất với dữ liệu là dạng hỗn hợp các phân phối xác suất chứ không phải dạng phân phối chuẩn đơn lẻ.

Từ khoá: Phân phối xác suất, hồi quy Bayes, mô hình tự hồi quy (AR) Bayes

TỔNG QUAN NGHIÊN CỨU

Mô hình hồi quy là các mô hình rất được quan tâm trong các lĩnh vực kinh tế, tài chính, dự báo… trong đó biến phụ thuộc là một hàm số của một hoặc một số biến độc lập. Trong thống kê tần suất, mô hình hồi quy dạng cổ điển cần thỏa mãn các điều kiện bao gồm nhiễu (error ε_i) là nhiễu trắng (white noise $N\left(0,\sigma^2\right)$), các tham số là hằng số (chưa biết), các biến độc lập là phi ngẫu nhiên. Khi đó, tích của các tham số và biến độc lập giống như là hằng số, nhiễu là phân phối chuẩn nên biến phụ thuộc y_i cũng sẽ tuân theo phân phối chuẩn.

Tổng quát trường hợp trên khi nhiễu không còn là phân phối chuẩn (nhiễu trắng), khi đó phụ thuộc vào phân phối xác suất của nhiễu, biến phụ thuộc y_i sẽ tuân theo một số phân phối xác suất khác như phân phối Poisson, phân phối nhị thức (the binomial distribution) hay phân phối có thứ tự (the multinomial distribution, the order distribution), phân phối *t* - Student… …[1,2]. Hiển nhiên, đối với một bộ dữ liệu có thể tồn tại rất nhiều dạng mô hình, chúng ta cần thực hiện các kiểm định nhằm lựa chọn dạng mô hình nào phù hợp nhất với dữ liệu[3].

Rõ ràng, trong thống kê tần suất, các tham số trong mô hình hồi quy được coi như một hằng số (chưa biết). Đặc biệt, trong trường hợp mô hình hồi quy tuyến tính cổ điển, trong đó vừa tuyến tính theo tham số vừa tuyến tính theo biến số, suy ra biến độc lập thay đổi sẽ tác động một lượng không đổi đến biến phụ thuộc, cụ thể khi *x* tăng thì *y* tăng (hoặc giảm) một lượng nhất định phụ thuộc tương ứng tham số dương (hoặc âm). Đây là điều chưa thật sự hợp lý, và do đó trong Thống kê Tần suất đã có sự chỉnh sửa bằng cách đưa thêm các biến "hiệu chỉnh" của các biến độc lập tương ứng nhằm biểu diễn rõ hơn sự tác động của biến độc lập đến biến phụ thuộc một cách phù hợp hơn.

Ngoài ra, các tham số trong mô hình hồi quy thống kê tần suất chỉ có thể xác định được một cách duy nhất dựa vào dữ liệu. Thật sự, trong rất nhiều trường hợp, dữ liệu thu thập được nhiều khi chưa kịp cập nhật thông tin hoặc dữ liệu theo tình huống mới quá ít so với dữ liệu theo tình huống cũ… Ngoài ra, để mô hình phản ánh đúng hơn các tình huống thực tiễn, các tham số được chọn cũng nên dựa vào nhận định của các chuyên gia, tức là trong tình huống tương tự có thể có các khả năng xảy ra, một ví dụ điển hình là trong phân tích dữ liệu lớn (big data). Do đó, một vấn đề đặt ra cần đánh giá các tham số của mô hình dưới dạng linh hoạt, nhằm phù hợp hơn với thực tiễn, đó chính là mô hình hồi quy trong thống kê Bayes.

[1]*Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam*

[2]*Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM, Việt Nam*

[3]*Trường Đại học Sài Gòn, Việt Nam*

Liên hệ

Lê Thanh Hoa, Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam

Email: hoalt@uel.edu.vn

Lịch sử

- Ngày nhận: 23-09-2020
- Ngày chấp nhận: 11-03-2021
- Ngày đăng: 31-03-2021

DOI : 10.32508/stdjelm.v5i1.701

Check for updates

VNU-HCM Press

Trích dẫn bài báo này: Hoa L T, Uyên P H, An N T D, Bảo P T. **Sự tương đồng về phân phối xác suất của các biến trong mô hình hồi quy Bayes và ứng dụng**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 5(1):1325-1339.

Trong thống kê Bayes, các tham số trong mô hình hồi quy được coi như là biến ngẫu nhiên. Các tham số trong mô hình hồi quy Bayes được tính toán dựa vào định lý Bayes nhằm tính ra hàm hậu nghiệm, thông qua các thành phần bao gồm thông tin từ dữ liệu (hàm hợp lý) và thông tin từ các chuyên gia (hàm tiên nghiệm). Các suy luận tiếp theo như ước lượng khoảng của tham số, ước lượng khoảng cho biến phụ thuộc, dự báo cho các quan sát tiếp theo của biến phụ thuộc được dựa vào hàm hậu nghiệm [4,5].

Như vậy, đối với mô hình hồi quy Bayes, chúng ta quan tâm đến cả dạng phân phối xác suất của tham số cũng như dạng phân phối xác suất của biến độc lập nên sẽ dẫn đến sự đa dạng phân phối xác suất của biến phụ thuộc y_i, cụ thể biến phụ thuộc y_i có thể là các phân phối chuẩn đơn lẻ hoặc là hỗn hợp của các phân phối chuẩn [6–8], hoặc là hỗn hợp của các phân phối xác suất khác [9].

Một khó khăn đặt ra là nếu xét mô hình hồi quy trong thống kê Bayes sẽ cần phải tính toán tích của các phân phối xác suất, trong khi việc tính tích của các phân phối xác suất dạng lý thuyết khá khó khăn và chưa tường minh [10]. Do đó, sẽ tiết kiệm thời gian và công sức nếu chúng ta tính toán tích của các phân phối xác suất thông qua mô phỏng ngẫu nhiên [11,12].

Trong bài báo này, chúng tôi đề nghị ước lượng dạng phân phối xác suất của biến phụ thuộc trong mô hình hồi quy Bayes thông qua mô phỏng ngẫu nhiên. Các phần tiếp theo của bài báo chúng tôi trình bày các nội dung:

- Phương pháp nghiên cứu: Mô hình hồi quy Bayes
- Kết quả nghiên cứu: Ước lượng dạng phân phối xác suất của biến phụ thuộc trong mô hình hồi quy Bayes thông qua mô phỏng ngẫu nhiên
- Ví dụ minh họa ứng dụng vào bộ dữ liệu thực
- Thảo luận và kết luận

PHƯƠNG PHÁP NGHIÊN CỨU: MÔ HÌNH HỒI QUY BAYES

Giả sử mô hình hồi quy tuyến tính Thống kê Tần suất thỏa mãn công thức (1):

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i,\ i = 1,2,\cdots,n. \quad (1)$$

trong đó n là số quan sát, các nhiễu ε_i là độc lập và cùng tuân theo một phân phối xác suất chuẩn $N\left(0,\sigma^2\right)$, thỏa mãn công thức (2) [1]:

$$E\left(\varepsilon_i\right) = 0,\ Var\left(\varepsilon_i\right) = \sigma^2. \quad (2)$$

Trong thống kê tần suất, do các biến độc lập được giả định là phi ngẫu nhiên, do đó các tác giả sẽ không quan tâm đến phân phối xác suất của các biến độc lập $x_j, j = \overline{1,k}$. Bên cạnh đó, các tham số $\beta_j, j = (\overline{0,k})$ được giả định như các hằng số. Do đó, tích của tham số và biến độc lập là các hằng số, tổng của các hằng số cũng là hằng số. Chính vì vậy, dạng phân phối xác suất của biến phụ thuộc sẽ bị ảnh hưởng bởi dạng phân phối xác suất của nhiễu. Tuy nhiên, điều này sẽ không còn phù hợp khi các biến độc lập dựa vào kết quả khảo sát luôn được cập nhật theo thời gian, do đó các tham số β_j sẽ thay đổi phụ thuộc vào sự thay đổi của biến độc lập. Hay nói cách khác, các tham số β_j cũng không thể dưới dạng các hằng số được nữa.

Đây chính là sự cần thiết của việc nghiên cứu mô hình hồi quy trong thống kê Bayes. Trong thống kê Bayes, các tham số $\beta_0, \beta_1, \cdots, \beta_k$ được xem xét như là biến ngẫu nhiên [4,5]. Dựa vào dạng phân phối xác suất của các tham số cũng như dạng phân phối xác suất của biến độc lập, chúng ta sẽ có các dạng phân phối xác suất khác nhau của biến phụ thuộc.

Sử dụng phương trình hồi quy (1), chúng ta có thể dự báo cho quan sát tiếp theo y_{n+1} tương ứng quan sát của $x_{n+1} = (x_{1(n+1)}, x_{2(n+1)}, ..., x_{k(n+1)})$ [1], được xác định theo công thức (3) sử dụng định lý Bayes:

$$\begin{aligned} &f\left(y_{n+1}|x_{n+1}, data\right) = \int_{B_0}\int_{B_1}\cdots\int_{B_k} \\ &f\left(y_{n+1}|\beta_0,\beta_1,...,\beta_k,x_{n+1},data\right)\times \\ &\pi\left(\beta_0,\beta_1,...,\beta_k,x_{n+1}.data\right)d\beta_0 d\beta_1 ... d\beta_k. \end{aligned} \quad (3)$$

trong đó $f\left(y_{n+1}|\beta_0,\beta_1,...,\beta_k,x_{n+1},data\right)$ là hàm hợp lý và $\pi\left(\beta_0,\beta_1,...,\beta_k,x_{n+1}.data\right)$ là phân phối hậu nghiệm cho tham số. Do đó, có thể xảy ra các trường hợp cho biến phụ thuộc y_i dưới dạng phân phối chuẩn đơn lẻ hay hỗn hợp các phân phối chuẩn, tức là y_i được biểu diễn theo công thức (4):

$$y_i \sim \sum_{j=1}^{m} \rho_j \times z_j, \quad (4)$$

trong đó $z_j \sim N(\mu_j, \sigma_j^2), \rho_j > 0, \sum_{j=1}^{m}\rho_j = 1$, hay nói cách khác y_i chính là hỗn hợp của các mô hình hồi quy thông thường [7].

KẾT QUẢ NGHIÊN CỨU: ƯỚC LƯỢNG DẠNG PHÂN PHỐI XÁC SUẤT CỦA BIẾN PHỤ THUỘC TRONG MÔ HÌNH HỒI QUY BAYES THÔNG QUA MÔ PHỎNG NGẪU NHIÊN

Thông thường, trong thống kê tần suất, dạng phân phối xác suất của biến phụ thuộc hay được xem như tuân theo phân phối chuẩn. Hay nói cách khác, trong thống kê tần suất, chúng ta không quan tâm đến dạng phân phối xác suất của biến độc lập và luôn xấp xỉ phân phối xác suất của biến phụ thuộc tuân theo phân phối chuẩn.

Nếu phân phối xác suất của biến độc lập tuân theo phân phối chuẩn thể hiện trong Hình 1, thì phân phối xác suất của biến phụ thuộc mới xấp xỉ phân phối chuẩn, trong đó: y=1+2*x+ε, với x$\sim$N(15;4^2),$\varepsilon\sim$N(0;1)

Còn trong trường hợp biến độc lập tuân theo phân phối đều biểu diễn trong Hình 2 và Hình 3, thì phân phối xác suất của biến phụ thuộc không còn xấp xỉ phân phối chuẩn.

Chúng ta nhận thấy trong Hình 2, khi phân phối xác suất của biến độc lập xấp xỉ là phân phối đều thì phân phối xác suất của biến phụ thuộc không tuân theo phân phối chuẩn cũng không tuân theo phân phối đều.

Tương tự như vậy, khi biến độc lập $x \sim 0,4 * N(7;1) + 0,6 * N(2;1)$ có dạng hỗn hợp các phân phối xác suất thì biến phụ thuộc y cũng có dạng hỗn hợp các phân phối xác suất.

Khác với trong thống kê tần suất, các mô hình hồi quy trong thống kê Bayes với giả định các tham số cũng là biến ngẫu nhiên, biến độc lập cũng là biến ngẫu nhiên. Chúng tôi sử dụng phương pháp mô phỏng ngẫu nhiên trong một số trường hợp ước lượng mô hình hồi quy Bayes[11,12]:

a. Tham số và biến độc lập đều tuân theo phân phối chuẩn

Biến phụ thuộc được biểu diễn theo công thức:

$$y = \beta_0 + \beta_1 * x + \varepsilon.$$

Trong đó: tham số β_0 tuân theo phân phối chuẩn $N(1;1)$, β_1 tuân theo phân phối chuẩn $N(-2;1)$, nhiễu ε_i tuân theo phân phối chuẩn $N(0;1)$, biến độc lập x_i tuân theo phân phối chuẩn $N(15;4^2)$.

Khi đó, với số lượng mô phỏng n = 5.000, chúng ta thấy rằng dạng phân phối xác suất của biến phụ thuộc không xấp xỉ phân phối chuẩn. Mặc dù, chúng ta nhận thấy dạng phân phối xác suất của dữ liệu được biểu diễn trong Hình 4 khá giống phân phối chuẩn.

b. Tham số tuân theo phân phối chuẩn, biến độc lập tuân theo phân phối đều trên [0, 1]

Tham số β_0 tuân theo phân phối chuẩn $N(1;1)$, β_1 tuân theo phân phối chuẩn $N(-2;1)$, nhiễu ε_i tuân theo phân phối chuẩn $N(0;1)$, biến độc lập x_i tuân theo phân phối đều trên [0; 1].

Khi đó, với số lượng mô phỏng n = 5.000, chúng ta thấy rằng dạng phân phối xác suất của biến phụ thuộc được biểu diễn trong Hình 5 không xấp xỉ phân phối chuẩn.

c. Tham số tuân theo phân phối chuẩn, biến độc lập tuân theo phân phối đều trên [0; 10]

Tham số β_0 tuân theo phân phối chuẩn $N(1;1)$, β_1 tuân theo phân phối chuẩn $N(-2;1)$, nhiễu ε_i tuân theo phân phối chuẩn $N(0;1)$, biến độc lập x_i tuân theo phân phối đều trên [0; 10].

Ở trường hợp này khác với trường hợp b, ở trên là phân phối đều trên một đoạn mở rộng hơn. Khi đó, theo Hình 6, với số lượng mô phỏng n = 5000, chúng ta nhận thấy rằng dạng phân phối xác suất của biến phụ thuộc đã không còn xấp xỉ phân phối chuẩn.

d. Tham số tuân theo phân phối chuẩn, biến độc lập là hỗn hợp các phân phối chuẩn

Tham số β_0 tuân theo phân phối chuẩn $N(1;1)$, β_1 tuân theo phân phối chuẩn $N(-2;1)$, nhiễu ε_i tuân theo phân phối chuẩn $N(0;1)$, biến độc lập x_i tuân theo hỗn hợp hai phân phối chuẩn $N(1;2^2)$ và $N(15;2^2)$ với xác suất lần lượt là 0,8 và 0,2.

Khi đó, với số lượng mô phỏng n = 5.000, theo Hình 7, chúng ta thấy một cách rõ ràng rằng dạng phân phối xác suất của biến phụ thuộc đã không còn xấp xỉ phân phối chuẩn, mà có dạng hỗn hợp các phân phối xác suất.

e. Tham số tuân theo phân phối chuẩn, biến độc lập là hỗn hợp của một phân phối chuẩn và một phân phối đều

Tham số β_0 tuân theo phân phối chuẩn $N(1;1)$, β_1 tuân theo phân phối chuẩn $N(-2;1)$, nhiễu ε_i tuân theo phân phối chuẩn $N(0;1)$, biến độc lập x_i tuân theo hỗn hợp một phân phối chuẩn $N(1;2^2)$ và một phân phối đều $U(8;18)$ với xác suất lần lượt là 0,7 và 0,3.

Khi đó, với số lượng mô phỏng n = 5.000, theo Hình 8, chúng ta thấy rằng dạng phân phối xác suất của biến phụ thuộc đã không còn xấp xỉ phân phối chuẩn, mà có dạng hỗn hợp các phân phối xác suất.

VÍ DỤ MINH HỌA ỨNG DỤNG VỚI DỮ LIỆU THỰC

Mô hình hồi quy Bayes chuỗi thời gian được biểu diễn theo phương trình *ARIMA(p,q)*

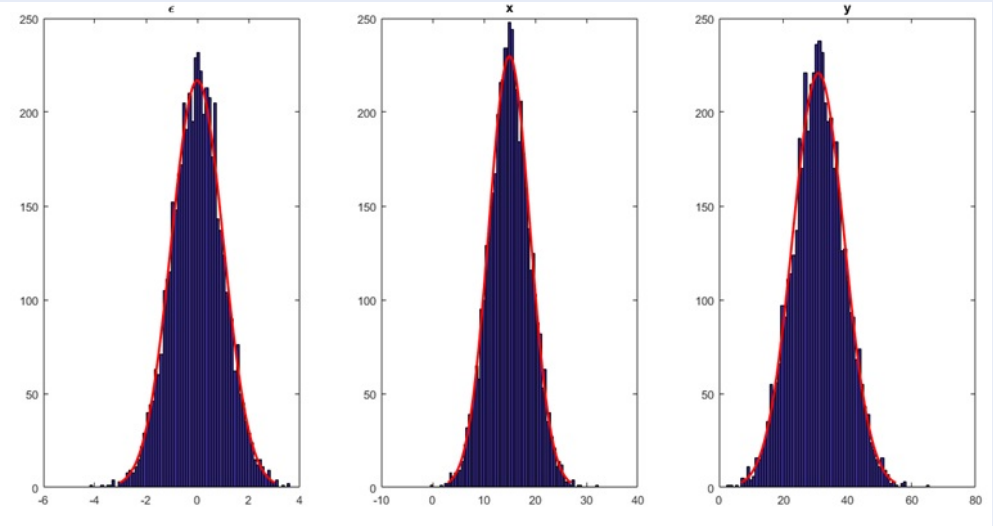
$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_q y_{t-q} + \varepsilon_t + \gamma_1 \varepsilon_{t-1} + \cdots + \gamma_p \varepsilon_{t-p}.$$

Trước hết, chúng tôi thu thập giá đóng cửa của mã chứng khoán AGR, Ngân hàng nông nghiệp và Phát triển nông thôn, tại Trung tâm Nghiên cứu Kinh tế - Tài chính, Trường Đại học Kinh tế - Luật được cung cấp bởi Thomson Reuter.

Trong Hình 9 biểu diễn dữ liệu giá đóng cửa theo chuỗi thời gian từ thời điểm bắt đầu lên sàn đến cuối năm 2017.

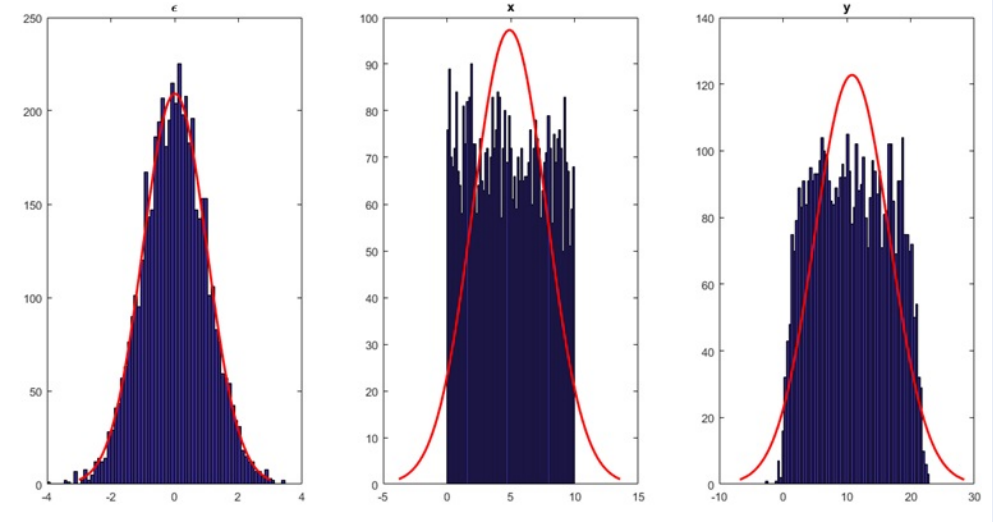
Chúng tôi kiểm định tính dừng của chuỗi dữ liệu AGR dựa vào kiểm định nghiệm đơn vị Dickey-Fuller thông qua kết quả Bảng 1.

Kết quả cho thấy, chuỗi dữ liệu giá đóng cửa không phải là chuỗi dừng. Chính vì vậy, chúng tôi biến đổi



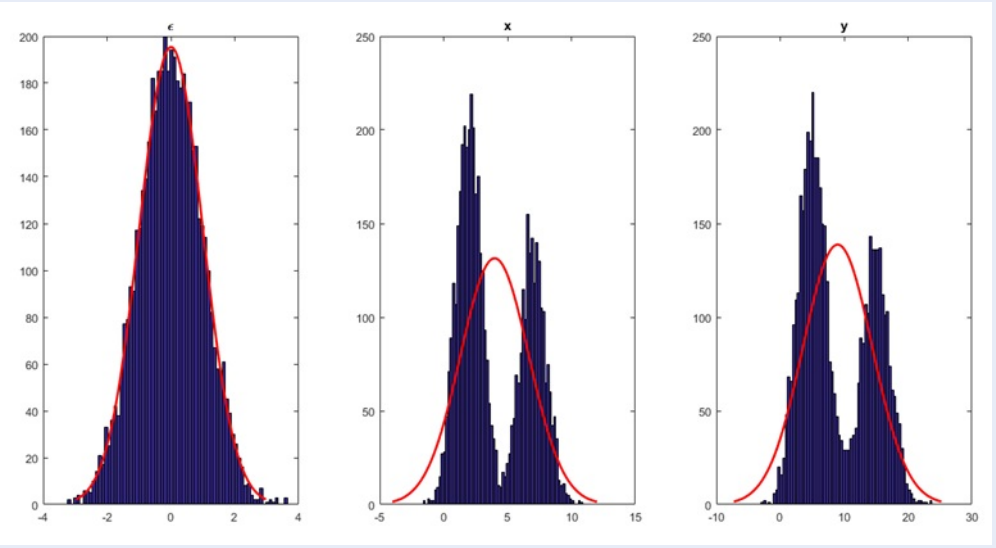
Hình 1: Dạng phân phối xác suất của biến phụ thuộc $y = 1 + 2 * x + \varepsilon$, với $x \sim N(15; 4^2),\ \varepsilon \sim N(0; 1)$[a]

[a]



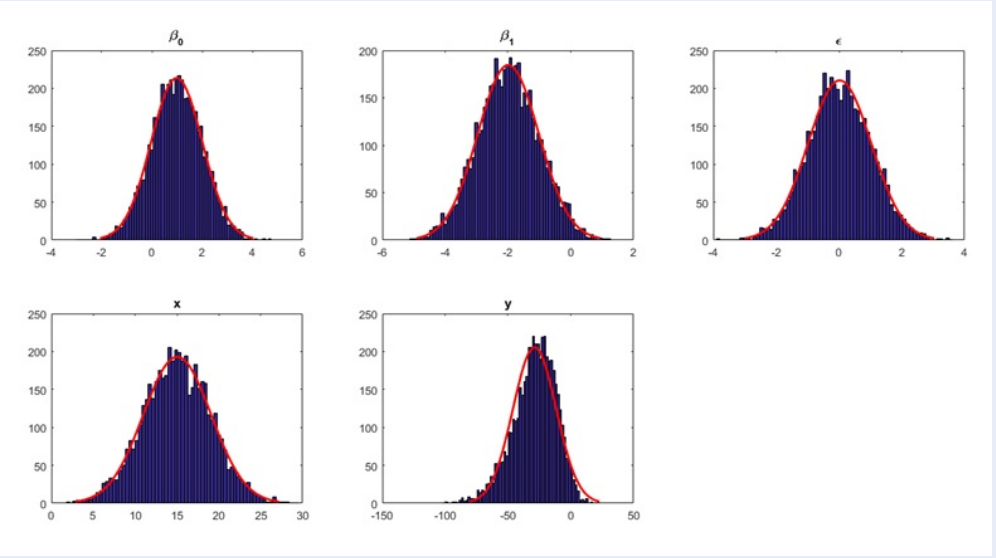
Hình 2: Dạng phân phối xác suất của biến phụ thuộc $y = 1 + 2 * x + \varepsilon$, với $x \sim U(0; 10),\ \varepsilon \sim N(0; 1)$[a]

[a]Nguồn: Kết quả nghiên cứu



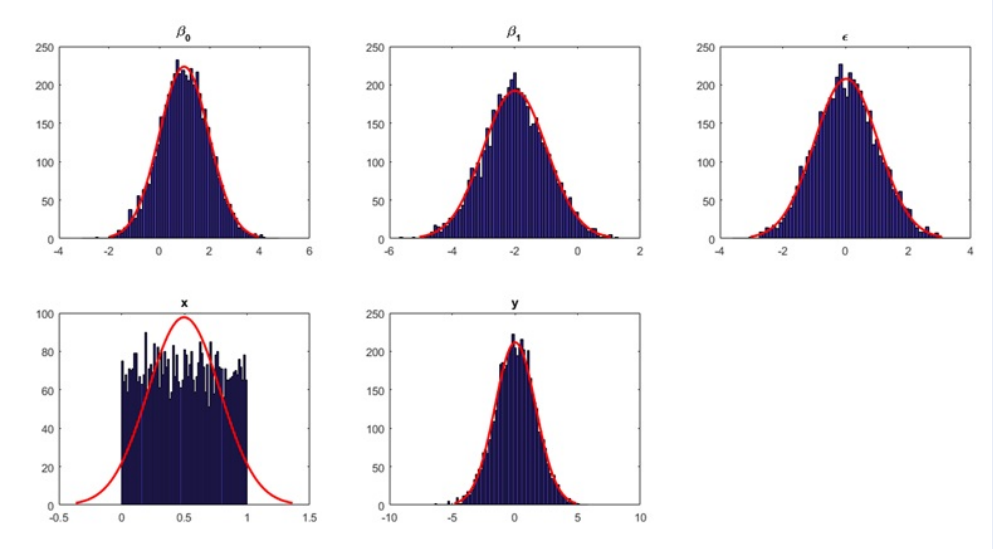
Hình 3: Dạng phân phối xác suất của biến phụ thuộc $y = 1 + 2 * x + \varepsilon$, với $x \sim 0.4 * N(7;1) + 0.6 * N(2;1),\ \varepsilon \sim N(0;1)$[a]

[a]Nguồn: Kết quả nghiên cứu



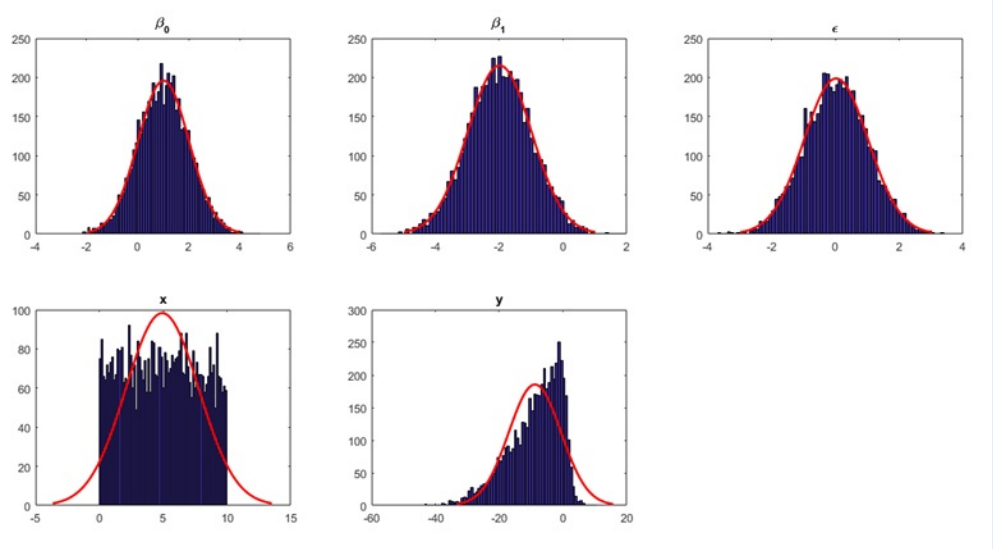
Hình 4: Dạng phân phối xác suất của biến phụ thuộc $\beta_0 \sim N(1;1),\ \beta_1 \sim N(-2;1), \varepsilon \sim N(0;1), x \sim N(15;4^2)$[a]

[a]Nguồn: Kết quả nghiên cứu



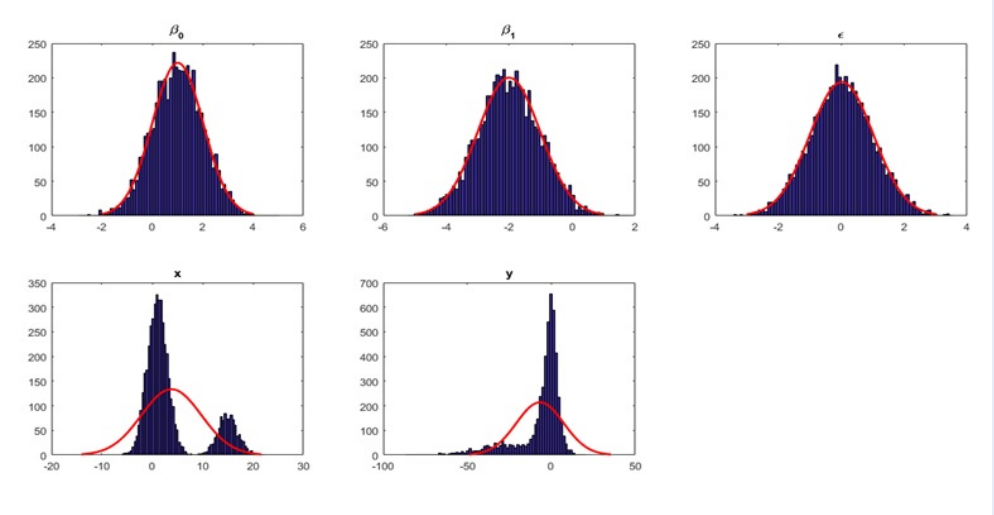
Hình 5: Dạng phân phối xác suất của biến phụ thuộc $\beta_0 \sim N(1;1)$, $\beta_1 \sim N(-2;1)$, $\varepsilon \sim N(0;1)$, $x \sim U(0;1)$[a]

[a]Nguồn: Kết quả nghiên cứu



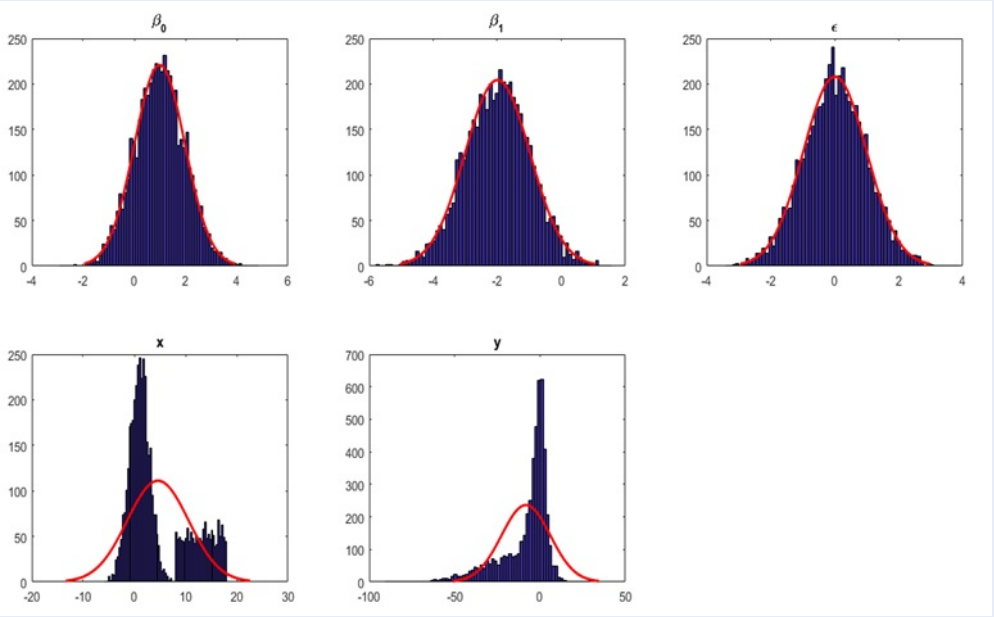
Hình 6: Dạng phân phối xác suất của biến phụ thuộc $\beta_0 \sim N(1;1)$, $\beta_1 \sim N(-2;1)$, $\varepsilon \sim N(0;1)$, $x \sim U(0;10)$[a]

[a]Nguồn: Kết quả nghiên cứu



Hình 7: Dạng phân phối xác suất của biến phụ thuộc $\beta_0 \sim N(1;1)$, $\beta_1 \sim N(-2;1)$, $\varepsilon \sim N(0;1)$, $x \sim 0,8 * N(1;2^2) + 0,2 * N(15;2^2)$[a]

[a]Nguồn: Kết quả nghiên cứu



Hình 8: Dạng phân phối xác suất của biến phụ thuộc $\beta_0 \sim N(1;1)$, $\beta_1 \sim N(-2;1), \varepsilon \sim N(0;1), x \sim 0,7 * N(1;2^2) + 0,3 * U(8;18)$[a]

[a]Nguồn: Kết quả nghiên cứu

Hình 9: Dữ liệu theo thời gian AGR, từ khi bắt đầu lên sàn cuối hết năm 2017[a]

[a]Nguồn: Kết quả nghiên cứu

Bảng 1: Kiểm định nghiệm đơn vị của dữ liệu giá đóng cửa AGR

Giả thuyết: CLOSE có nghiệm đơn vị			
Biến độc lập: Hằng số, Xu thế tuyến tính			
		Giá trị kiểm định t	Xác suất*
Giá trị kiểm định Augmented Dickey-Fuller		-2,524954	0,3159
Giá trị tra bảng:	Mức ý nghĩa 1%	-3,962609	
	Mức ý nghĩa 5%	-3,412043	
	Mức ý nghĩa10%	-3,127932	

*Giá trị xác suất kiểm định một phía MacKinnon (1996).
Nguồn: Kết quả nghiên cứu

Bảng 2: Kiểm định nghiệm đơn vị của sai phân bậc 1 dữ liệu giá đóng cửa AGR

Giả thuyết: D(CLOSE) có nghiệm đơn vị			
Biến độc lập: Hằng số, xu thế tuyến tính			
		Giá trị kiểm định t	Xác suất*
Giá trị kiểm định Augmented Dickey-Fuller		-41,90567	0,0000
Giá trị tra bảng:	Mức ý nghĩa1%	-3,962611	
	Mức ý nghĩa 5%	-3,412044	
	Mức ý nghĩa 10%	-3,127933	

*Giá trị xác suất kiểm định một phía MacKinnon (1996).
Nguồn: Kết quả nghiên cứu

dữ liệu giá đóng cửa dưới dạng sai phân bậc 1 và kết quả kiểm định Dickey-Fuller theo Bảng 2.

Chúng tôi nhận thấy, chuỗi sai phân bậc 1 của dữ liệu giá đóng cửa đã là chuỗi dừng. Kết quả chỉ ra, sai phân bậc 1 của chuỗi dữ liệu giá đóng cửa thực sự là chuỗi dừng. Hơn nữa, dựa vào kết quả của giản đồ tự tương quan trong Bảng 3, chúng tôi lựa chọn mô hình hồi quy tự tương quan bậc 1 cho chuỗi sai phân bậc 1 của dữ liệu giá đóng cửa AGR, do $|AC_1| = max\{|AC_i|, i = 1,2,\ldots,10\}$.

Hình ảnh của dữ liệu sai phân bậc 1 giá đóng cửa AGR theo thời gian được biểu diễn trong Hình 10.

Mặc dù chúng ta nhận thấy sai phân của dữ liệu giá đóng cửa AGR đã là chuỗi dừng, hay nói cách khác là các sai phân này dao động xung quanh giá trị trung bình (xung quanh giá trị 0) và phương sai không đổi. Tuy nhiên, khi biểu diễn rõ hơn về đồ thị của sai phân bậc 1 chuỗi dữ liệu theo Hình 11.

Rõ ràng, theo Hình 11, việc xấp xỉ sai phân bậc 1 của chuỗi dữ liệu theo phân phối chuẩn là không phù hợp. Do đó, trong mô hình AR(1) cho sai phân bậc 1 giá đóng cửa được biểu diễn dưới dạng phương trình (5) chưa mô hình hóa được chính xác dữ liệu thực tế:

$$d(y_t) = \beta_0 + \beta_1 d(y_{t-1}) + \varepsilon_t, \quad (5)$$

Trong đó $\beta_1 < 1$ và ε_t là nhiễu trắng.

Do đó, trong mô hình hồi quy Bayes, chúng ta sẽ mô hình hóa bộ dữ liệu $d(y_{t-1})$ dưới dạng hỗn hợp các phân phối xác suất, các tham số β_0, β_1 dưới dạng phân phối xác suất, ε_t vẫn là nhiễu trắng. Khi đó, vế phải của phương trình (5) có dạng vừa tổng của các biến ngẫu nhiên vừa tổng của tích hai biến ngẫu nhiên. Với mỗi biến ngẫu nhiên, chúng tôi gắn với một phân phối xác suất tương ứng, đây chính là các thông tin về hàm hợp lý dựa vào dữ liệu.

Dựa vào thông tin từ sai phân bậc 1 của dữ liệu giá đóng cửa AGR, chúng tôi ước lượng sai phân của dữ liệu giá đóng cửa AGR thông qua hỗn hợp các phân phối chuẩn, với bảng kết quả của chỉ số AIC khi xấp xỉ hỗn hợp các phân phối chuẩn được biểu diễn theo Bảng 4.

Khi đó, mô hình tốt nhất biểu diễn sai phân bậc 1 của dữ liệu giá đóng cửa AIC thông qua hỗn hợp 3 phân phối chuẩn, dựa vào chỉ số *minAIC*, với tỷ lệ các phân phối chuẩn và trung bình thành phần tương ứng được biểu diễn theo Bảng 5.

Để kiểm tra tính ổn định của việc ước lượng hỗn hợp các phân phối xác suất của dữ liệu thực, chúng tôi so sánh sự khác biệt giữa giá trị thực tế (sai phân của dữ liệu giá AGR) với dữ liệu được mô phỏng từ các tham số ước lượng trong hỗn hợp các phân phối xác suất. Các giá trị này được sắp xếp từ nhỏ đến lớn được biểu diễn trong Hình 12.

Chúng ta nhận thấy các giá trị mô phỏng có hình dáng gần giống với dữ liệu thực, minh họa chi tiết hơn, chúng tôi minh họa biểu đồ hình hộp giữa dữ liệu thực và dữ liệu mô phỏng theo Hình 13.

Để đảm bảo tính đa dạng của dữ liệu, chúng tôi mô phỏng tất cả 100 bộ dữ liệu, với các tham số về tỷ lệ, trung bình và độ lệch chuẩn của các phân phối chuẩn thành phần chính là các giá trị ước lượng ở trên. Số lượng quan sát mô phỏng của mỗi bộ dữ liệu trùng với số lượng quan sát của bộ dữ liệu sai phân bậc 1 của dữ liệu gốc AGR. Sau đó, chúng tôi tính sai số giữa giá trị thực và giá trị mô phỏng thông qua ba công thức tính sai số:

- Trung bình các sai số (Mean error):

$$ME = \frac{\sum_{I=1}^{n} e_i}{n} = \frac{\sum_{I=1}^{n} (y_i - \widehat{y}_i)}{n}.$$

- Trung bình của bình phương các sai số (Mean square error)

$$MSE = \frac{\sum_{I=1}^{n} e_i^2}{n} = \frac{\sum_{I=1}^{n} (y_i - \widehat{y}_i)^2}{n}.$$

Các giá trị của sai số thành phần giữa 100 bộ dữ liệu mô phỏng và bộ dữ liệu thực được biểu diễn theo Hình 14.

Chúng ta nhận thấy các sai số đủ bé chứng tỏ trong 100 bộ dữ liệu mô phỏng là xấp xỉ tốt của dữ liệu thực, trong đó trung bình của các ME là 0.0000, trung bình của các MAE là 0.1059 và trung bình của các MSE là 0.0224.

Bên cạnh đó, trong mô hình hồi quy Bayes còn cần sử dụng thêm các thông tin tiên nghiệm về các tham số β_0, β_1, $d(y_t)$ dựa vào các nghiên cứu trước hay các chuyên gia. Để lựa chọn chính xác thông tin tiên nghiệm phù hợp cần các nghiên cứu dài hơi, phụ thuộc vào mục đích của nhà nghiên cứu như không đánh giá tác động của thông tin tiên nghiệm (tiên nghiệm phi thông tin), đơn giản các tính toán (tiên nghiệm liên hợp), sử dụng tối đa các thông tin (tiên nghiệm entropy cực đại) …

Mô hình hồi quy Bayes sử dụng cho các suy luận cũng như dự báo các giá trị trong tương lai thông qua phân phối hậu nghiệm, là kết quả tính toán dựa vào thông tin hàm hợp lý từ dữ liệu và thông tin tiên nghiệm thông qua công thức Bayes.

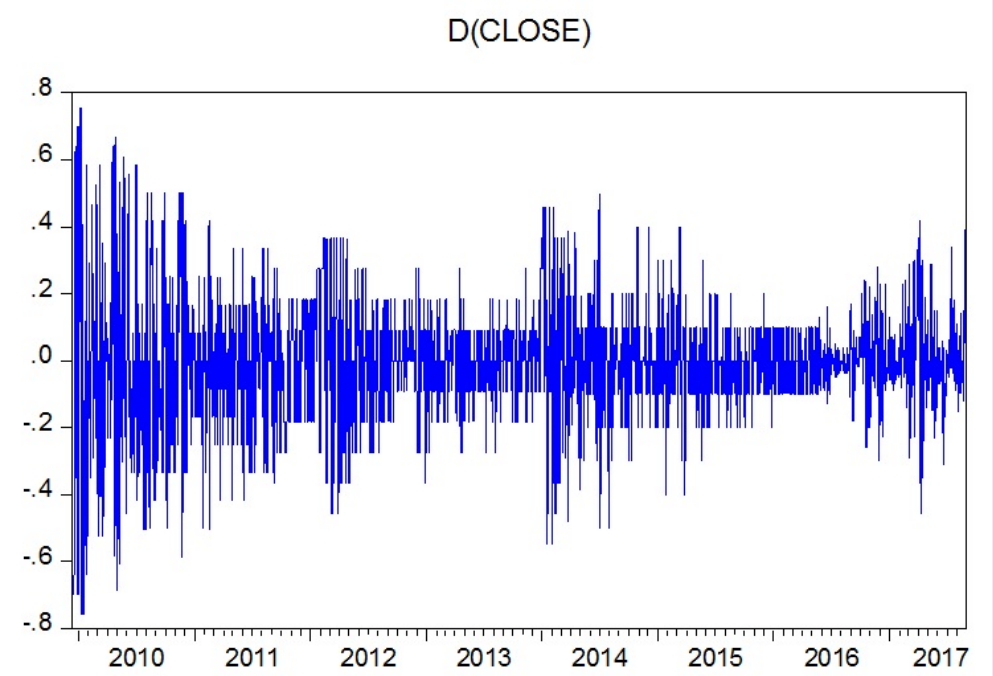
THẢO LUẬN VÀ KẾT LUẬN

Mô hình hồi quy là các mô hình quan trọng trong phân tích tài chính, đặc biệt là trong nghiên cứu giá chứng khoán. Tuy nhiên, dạng phân phối xác suất của biến phụ thuộc là một vấn đề thường bị bỏ qua trong mô hình hồi quy của thống kê tần suất, và như vậy biến phụ thuộc được coi như tuân theo phân phối

Bảng 3: Giản đồ tự tương quan theo thời gian chuỗi dữ liệu sai phân bậc 1 giá đóng cửa AGR

Hệ số tự tương quan	Hệ số tự tương quan riêng phần		AC	PAC	Giá trị kiểm định Q	Xác suất
| |	| |	1	0,071	0,071	10,251	0,001
| |	| |	2	-0,000	-0,006	10,252	0,006
| |	| |	3	-0,023	-0,022	11,283	0,010
| |	| |	4	-0,056	-0,053	17,652	0,001
| |	| |	5	0,011	0,019	17,890	0,003
| |	| |	6	0,018	0,016	18,560	0,005
| |	| |	7	-0,022	-0,027	19,531	0,007
| |	| |	8	0,024	0,025	20,676	0,008
| |	| |	9	0,009	0,008	20,847	0,013
| |	| |	10	-0,015	-0,016	21,295	0,019

Nguồn: Kết quả nghiên cứu



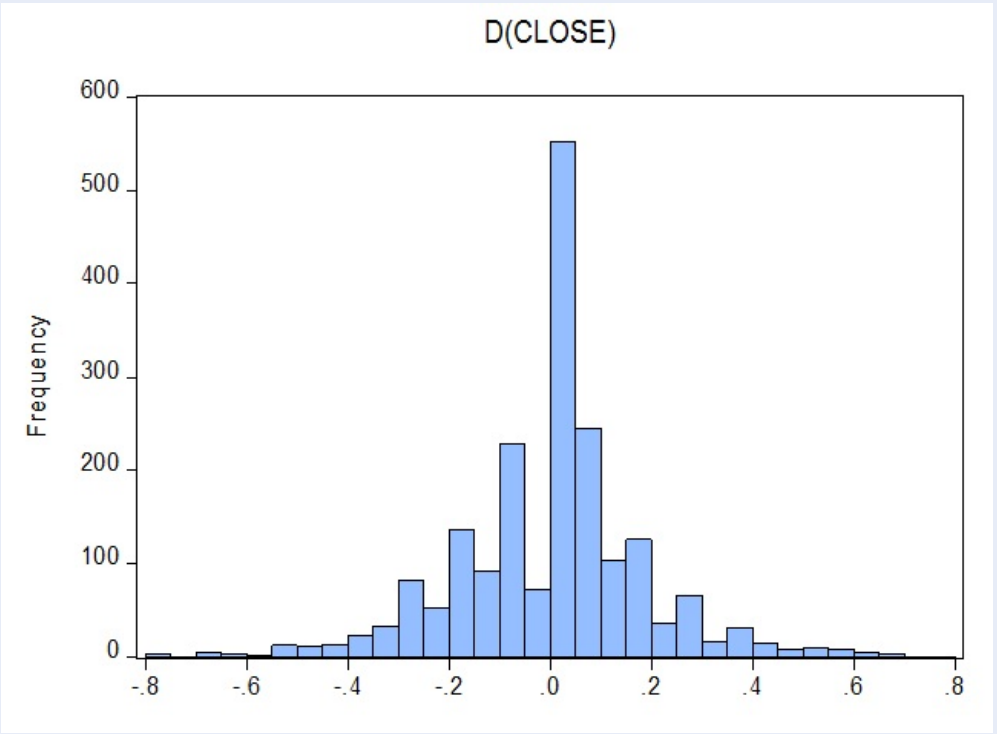
Hình 10: Đồ thị sai phân bậc 1 giá đóng cửa của dữ liệu AGR theo thời gian[a]

[a]Nguồn: Kết quả nghiên cứu

Bảng 4: Các tỷ lệ thành phần của hỗn hợp các phân phối xác suất của sai phân bậc 1 dữ liệu giá đóng cửa AGR

Số thành phần hỗn hợp phân phối chuẩn	1	2	3	4	5
AIC	-946,1066	-1183,9227	-1187,3642	-1171,7926	-1164,4863

Nguồn: Kết quả nghiên cứu



Hình 11: Biểu đồ của sai phân bậc 1 chuỗi dữ liệu giá đóng cửa AGR theo thời gian[a]

[a]Nguồn: Kết quả nghiên cứu

Bảng 5: Các trung bình các phân phối xác suất thành phần của sai phân bậc 1 dữ liệu giá đóng cửa AGR

Tỷ lệ	**0,2697**	**0,3635**	**0,3668**
Trung bình	0,0137	-0,0285	0,0063

Nguồn: Kết quả nghiên cứu

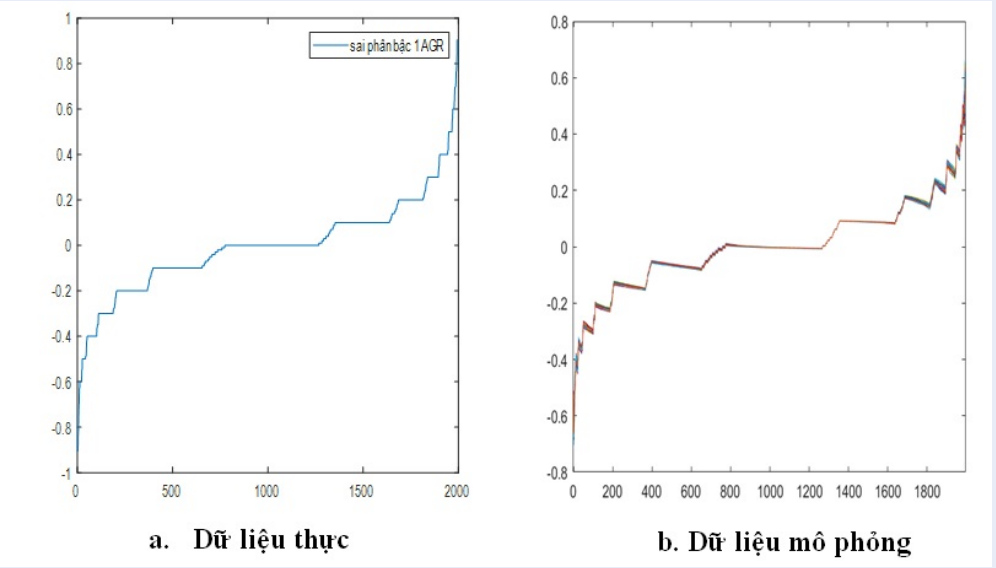
chuẩn. Hơn nữa, trong mô hình hồi quy thống kê tần suất, các tham số được coi như là một hằng số, do đó sẽ không còn phù hợp vì một khi biến độc lập thay đổi dẫn đến các biến phụ thuộc sẽ thay đổi nhưng không phải chỉ theo một hằng số cố định.

Thật vậy, trong mô hình hồi quy thống kê tần suất, khi biến độc lập tuân theo phân phối chuẩn thì biến phụ thuộc cũng tuân theo phân phối chuẩn, khi biến độc lập tuân theo phân phối đều thì biến phụ thuộc cũng tuân theo phân phối đều, khi biến độc lập tuân theo dạng hỗn hợp các phân phối xác suất thì biến phụ thuộc cũng tuân theo dạng hỗn hợp các phân phối xác suất. Tuy nhiên, trong mô hình hồi quy thống kê tần suất như mô hình tự hồi quy phân tích giá chứng khoán, với biến độc lập là hỗn hợp các phân phối xác suất còn biến phụ thuộc lại xấp xỉ phân phối chuẩn, do đó không thể tồn tại tham số dưới dạng hằng số để thỏa mãn dạng phân phối xác suất của hai vế mô hình hồi quy.

Chính vì vậy, chúng ta cần nghiên cứu mô hình hồi quy Bayes, trong đó các tham số dưới dạng các phân phối xác suất nhằm đảm bảo đa dạng hóa dạng phân phối xác suất của biến phụ thuộc, và hơn thế phù hợp với sự thay đổi cập nhật liên tục của biến độc lập. Trong bài báo, chúng tôi đã minh họa chi tiết các phân phối xác suất trong các trường hợp của tham số, của biến độc lập cũng như biến phụ thuộc; đồng thời chúng tôi mở ra triển vọng ứng dụng trong phân tích giá chứng khoán thực. Hướng nghiên cứu tiếp theo chúng tôi sẽ sử dụng các mô hình hồi quy Bayes trong dự báo các giá trị trong tương lai.

LỜI CẢM ƠN

Nghiên cứu được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (ĐHQG-HCM) trong khuôn khổ Đề tài mã số C2019-34-04.



Hình 12: Dữ liệu về sai phân bậc 1 dữ liệu gốc AGR được sắp xếp từ nhỏ đến lớn và dữ liệu mô phỏng hỗn hợp các phân phối chuẩn theo tỷ lệ và trung bình ước lượng từ dữ liệu thực.[a]

[a]Nguồn: Kết quả nghiên cứu

XUNG ĐỘT LỢI ÍCH

Nhóm tác giả xin cam đoan rằng không có bất kì xung đột lợi ích nào trong công bố bài báo

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Tác giả 1, Lê Thanh Hoa, chịu trách nhiệm về thiết kế các nội dung đưa vào bài báo, về ước lượng dạng phân phối xác suất của biến phụ thuộc trong mô hình hồi quy Bayes thông qua mô phỏng ngẫu nhiên.

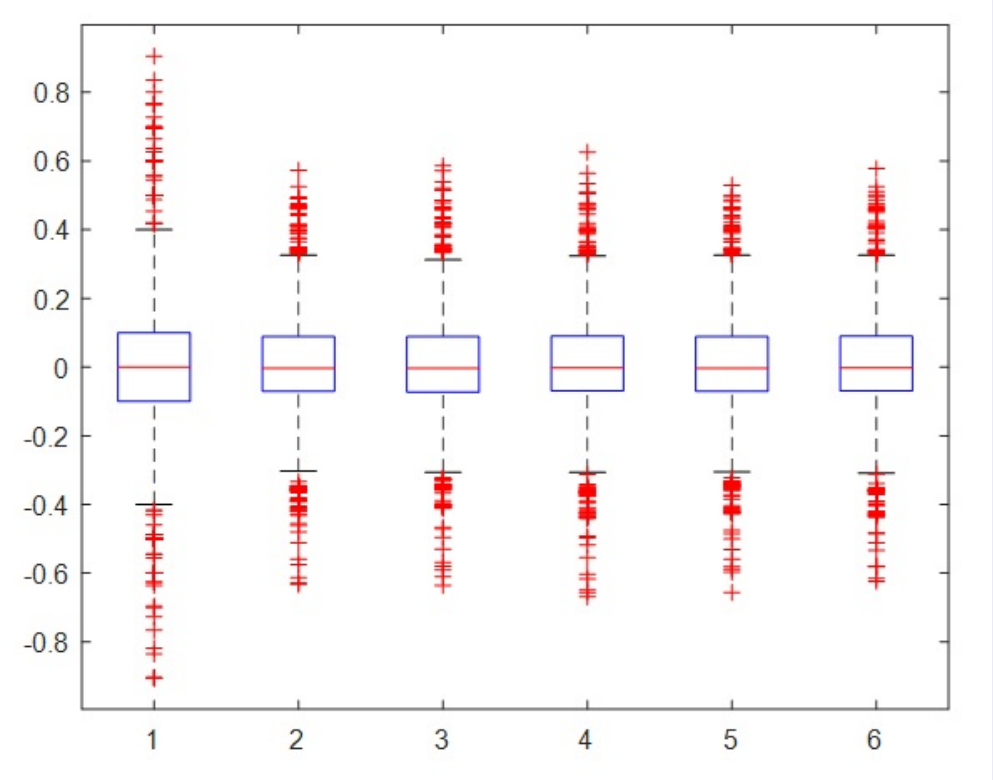
Tác giả 2, Phạm Hoàng Uyên, chịu trách nhiệm về các lý thuyết của mô hình hồi quy Bayes.

Tác giả 3, Nguyễn Thị Đỗ An, chịu trách nhiệm về ví dụ minh họa các dữ liệu thực.

Tác giả 4, Phạm Thế Bảo, chịu trách nhiệm về giới thiệu tổng quan các vấn đề nghiên cứu.

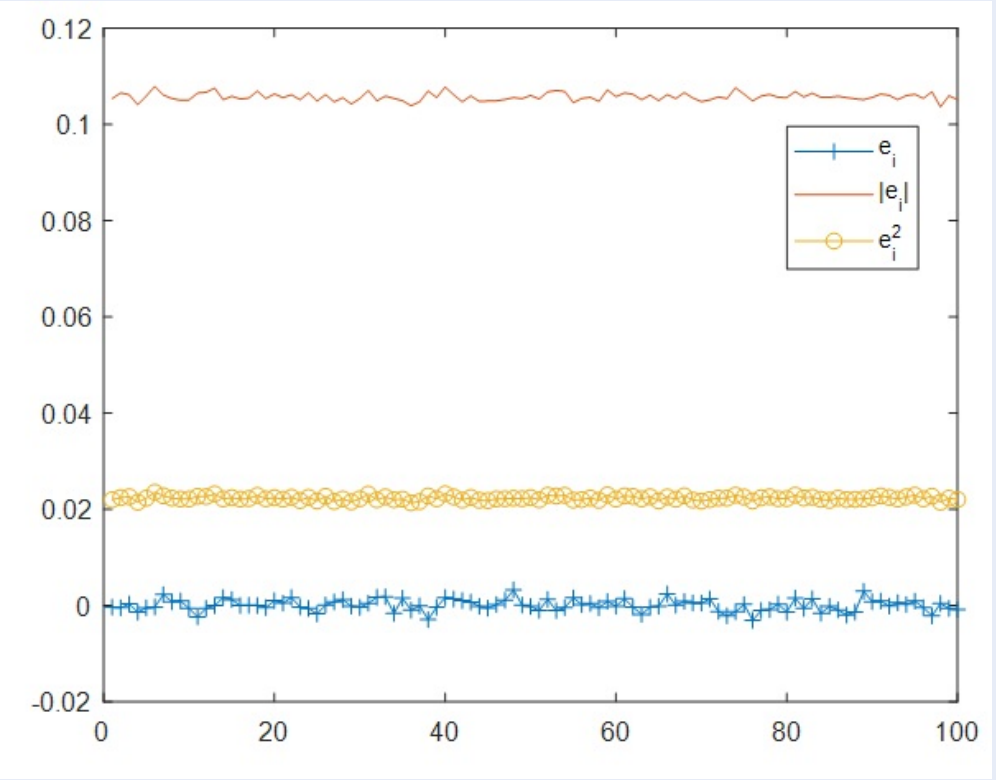
TÀI LIỆU THAM KHẢO

1. Fahrmeir L, Kneib T, et al. Regression: Models, Methods and Applications, Springer Science & Business Media. 2013;PMID: 23893691. Available from: https://doi.org/10.1007/978-3-642-34333-9.
2. Frees EW. Regression Modeling with Actuarial and Financial Applications. Cambridge University Press. 2009;Available from: https://doi.org/10.1017/CBO9780511814372.
3. Harrell FE. Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression. GET ADDRESS: Springer. 2001;Available from: https://doi.org/10.1007/978-1-4757-3462-1.
4. Bolstad WM, Curran JM. Introduction to Bayesian Statistics. John Wiley & Sons. 2016;Available from: https://doi.org/10.1002/9781118593165.
5. Gelman A, et al. Bayesian Data Analysis. Chapman and Hall/CRC. 2013;Available from: https://doi.org/10.1201/b16018.
6. Fallah A, Mohammadzadeh M. Bayesian Regression Analysis with Linked Data Using Mixture Normal Distributions. Statistical Papers. 2010;51(2):421–430. Available from: https://doi.org/10.1007/s00362-009-0208-x.
7. Lee KJ, et al. Bayesian Variable Selection for Finite Mixture Model of Linear Regressions. Computational Statistics & Data Analysis. 2016;95:1–16. Available from: https://doi.org/10.1016/j.csda.2015.09.005.
8. Cancho VG, et al. Bayesian Nonlinear Regression Models with Scale Mixtures of Skew-Normal Distributions: Estimation and Sase Influence Diagnostics. Computational Statistics & Data Analysis. 2011;55(1):588–602. Available from: https://doi.org/10.1016/j.csda.2010.05.032.
9. Yang F, Yuan H. A Non-Iterative Bayesian Sampling Algorithm for Linear Regression Models with Scale Mixtures of Normal Distributions. Computational Economics. 2017;49(4):579–597. Available from: https://doi.org/10.1007/s10614-016-9580-5.
10. Glen AG, et al. Computing the Distribution of the Product of Two Continuous Random Variables," Computational Statistics & Data Analysis. 2004;44(3):451–464. Available from: https://doi.org/10.1016/S0167-9473(02)00234-7.
11. Le H, Pham U, Nguyen P, Bao PT. Improvement on Monte Carlo Estimation of HPD intervals. Communications in Statistics - Simulation and Computation. 2018;p. 1–17. Available from: https://doi.org/10.1080/03610918.2018.1513141.
12. Hoa LT, Uyên PH, Đình Thiên N. Một phương pháp mới tìm khoảng mật độ hậu nghiệm cao nhất và ứng dụng. Tạp chí Phát triển Kinh tế. 2017;28(10):79–120.



Hình 13: Biểu đồ hình hộp về sai phân bậc 1 dữ liệu AGR tương ứng bộ dữ liệu 1 và các dữ liệu mô phỏng tương ứng bộ dữ liệu 2 đến 6[a]

[a]Nguồn: Kết quả nghiên cứu



Hình 14: Các trung bình các sai số ME thông qua e_i trung bình tuyệt đối các sai số MAE thông qua $|e_i|$, trung bình bình phương các sai số MSE thông qua e_i^2 giữa giá trị mô phỏng và giá trị thực.[a]

[a]Nguồn: Kết quả nghiên cứu

Science & Technology Development Journal – Economics - Law and Management, 5(1):1325-1339

Open Access Full Text Article

Research Article

The similarity about the probability distributions of variables in the Bayesian regression model and application

Hoa Le[1,*], Uyen Pham[1], Nguyen Thi Do An[2], Pham The Bao[3]

ABSTRACT

The linear regression model as well as the time series model is applied in many fields, in which the mean of the dependent variable is one function of the mean of the independent variables. However, to consider the regression model following in the Classical Statistics (the Frequent Statistics), it means that the parameters are the constants, in many situations, the regression model does not describe the fluctuation of both the dependent variable and the independent variables. Therefore, we need to modify the parameters following the random variable form, not the constant form, like as the regression in Bayesian Statistics. The other side, when the parameters considered as the random variables, computations in the regression model becomes very complex, because we need to compute the product of the probability distributions. So, we must evaluate about to vary of the variables' probability distributions not only the normal distribution, the Student distribution t, the Poisson distribution, the binomial distribution… In this paper, we estimated the dependent variable's probability distribution form through the simple Bayesian regression model in cases having many the probability distribution forms of the independent variable. In addition, we apply the results to real stock price data, proving that the most appropriate probability distribution with the data is a mixture of probability distributions, not a single normal distribution.

Key words: The probability distribution, The Bayesian regression model, The Bayesian autoregression model (AR)

[1]*University of Economics and Law, VNU-HCM, Vietnam*

[2]*University of Science, VNU-HCM, Vietnam*

[3]*Sai Gon University, Vietnam*

Correspondence

Hoa Le, University of Economics and Law, VNU-HCM, Vietnam

Email: hoalt@uel.edu.vn

History

- Received: 23-09-2020
- Accepted: 11-03-2021
- Published: 31-03-2021

DOI : 10.32508/stdjelm.v5i1.701

Cite this article : Le H, Pham U, An N T D, Bao P T. **The similarity about the probability distributions of variables in the Bayesian regression model and application**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 5(1):1325-1339.