

PHÂN TÍCH HỒI QUY TRONG LÂM HỌC

REGRESSION ANALYSIS IN SILVICULTURE

Xem chi tiết chương V trong cuốn sách tham khảo

“Những cách thức và phương pháp xử lý số liệu trong lâm học”

MỤC LỤC

1. Hàm hồi quy tuyến tính đa biến
2. Hàm hồi quy bậc 2
3. Hàm hồi quy Sigmoid và Logit Gauss
4. Hàm hồi quy Logistic
5. Hàm hồi quy Poisson

1. HỒI QUY TUYẾN TÍNH ĐA BIẾN

Hàm hồi quy tuyến tính đa biến (Multiple Linear Regression Function):

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p + e$$

Y = Biến phụ thuộc hay biến phản hồi.

X_i = Biến độc lập hay biến giải thích.

b_0 = Điểm chặn (Hằng số) của Y trên trục tung.

b_i = Hệ số độ dốc đối với mỗi biến giải thích X_i . Hệ số này đo sự thay đổi 1 đơn vị của biến Y khi thay đổi 1 đơn vị của biến X_i .

e = Sai số hay sai lệch của hàm hồi quy.

1.1. Một số áp dụng hồi quy tuyến tính đa biến

- (1) Định lượng phản hồi của các loài cây gỗ và quần thụ với biến động của các nhân tố sinh thái (Khí hậu, Đất, Địa hình, Chỉ số lập địa SI...).
- (2) Xác định quan hệ giữa sản lượng gỗ (M , m³/ha) của một loại rừng với các biến đường kính (D , cm), chiều cao (H , m), tiết diện ngang (G , m²/ha) và mật độ của quần thụ (N , cây/ha).
- (3) Xây dựng các hàm để ước lượng và dự đoán sản lượng gỗ ở mức cây gỗ và quần thụ.

1.2. Các bước phân tích hồi quy tuyến tính đa biến

Bước 1. Xác định rõ ràng các biến phụ thuộc (Y_i). Sau đó thu thập các thông tin của các biến Y_i .

- Bước 2.** Xác định rõ ràng các biến độc lập (X_p). Đó là các biến giải thích quan hệ giữa Y_i với các biến X_p . Sau đó thu thập các thông tin của các biến X_p .
- Bước 3.** Xác định các thống kê mô tả đối với các biến Y_i và X_p (Trung bình, phạm vi biến động (Max-Min), sai lệch (SEE), hệ số biến động (CV%).
- Bước 4.** Phân tích quan hệ giữa các biến Y_i với các biến X_p thông qua ma trận tương quan đơn. Bước này cho biết cường độ và chiều hướng của quan hệ giữa các biến Y_i với các biến X_p .
- Bước 5.** Chọn những biến X_p tồn tại quan hệ chặt chẽ với các biến Y_i .
- Bước 6.** Từ những biến X_p tồn tại quan hệ chặt chẽ với các biến Y_i , xây dựng hàm hồi quy $Y_i = f(X_1, X_2, \dots, X_p)$. Ở đây cần tính hệ số xác định (R^2) và các thành phần sai lệch của mô hình hồi quy ước lượng (SST, SSR, SSE, MAE, MAPE, ME, MPE).
- Bước 7.** Kiểm định sự tồn tại của hàm hồi quy tuyến tính đa biến bằng thống kê F và thống kê T. Thống kê F kiểm định sự tồn tại của hàm hồi quy. Thống kê T kiểm định sự tồn tại của các hệ số hồi quy.
- Bước 8.** Kiểm định cộng tuyến tính giữa các biến X_p . Thống kê cộng tuyến tính (Collinearity Statistics) đo mức độ quan hệ tuyến tính giữa các biến độc lập (X_i). Để đo mức độ cộng tuyến tính, người ta sử dụng thống kê độ chấp nhận (Tolerance) và yếu tố khuếch đại phương sai (Variance Inflation Factor = VIF). Độ chấp nhận của biến thứ p được tính nhằm xác định biến này có ý nghĩa trong mô hình hay không. Tiêu chuẩn độ chấp nhận mặc định là 0,0001. Vì thế, nếu một biến được đưa vào mô hình, thì độ chấp nhận của nó phải lớn hơn 0,0001. Yếu tố khuếch đại phương sai của biến X_i (VIF_i) là phương sai của hệ số hồi quy thứ i. Khi VIF_i lớn, thì hệ số hồi quy thứ i biến động mạnh. Sự biến động mạnh của hệ số hồi quy thứ i cho biết biến này có ý nghĩa kém trong mô hình. **Nói chung, khi một biến độc lập có quan hệ tuyến tính với những biến độc lập khác, thì $VIF \gg 1$ và $CI \gg 15$.** Trong thực hành, việc kiểm định hiện tượng cộng tuyến tính giữa các biến độc lập và chọn những biến độc lập hữu ích nhất để xây dựng hàm hồi quy ước lượng phải trải qua nhiều bước tính toán rất phức tạp. Hiện nay nhiều phần mềm thống kê (IBM SPSS, MINITAB, STATGRAPHICS...) cho phép thực hiện các kiểm định cộng tuyến tính và chọn những biến độc lập hữu ích nhất để xây dựng hàm hồi quy ước lượng.
- Bước 9.** Xây dựng mô hình hồi quy để ước lượng các biến Y . Bước này được thực hiện sau khi chọn được những biến dự đoán (X = độc lập, giải thích) tồn tại quan hệ chặt chẽ với các biến phụ thuộc (Y). Trước khi xây dựng mô hình, chúng ta cần phân tích các biến được đưa vào và loại ra khỏi hàm hồi quy và những thống kê của các biến không được đưa vào mô hình. Khi có nhiều biến độc lập được chọn để xây dựng mô hình, chúng ta cần biết những biến nào tạo

ra một mô hình tốt (Sai lệch ước lượng và sai lệch dự đoán đều nhỏ). Những biến đưa vào và loại ra khỏi mô hình được xác định theo phương pháp chọn biến từng bước (Stepwise Method).

Các hàm hồi quy tuyến tính đa biến trong chọn biến từng bước

$$Y = b_0 + b_1X_1 \quad (1)$$

$$Y = b_0 + b_1X_1 + b_2X_2 \quad (2)$$

$$\dots \quad (3)$$

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p \quad (4)$$

Bước 10. Xác định số lượng biến dự đoán tối ưu. Về lý thuyết, khi đưa vào mô hình càng nhiều biến, thì hệ số xác định (R^2) càng cao. Mặt khác, hệ số R^2 sẽ không bao giờ giảm khi các biến dự đoán tiếp tục được đưa vào mô hình. Nhưng một mô hình với nhiều biến dự đoán không có nghĩa là nó sẽ phù hợp hơn với tổng thể. Một mô hình nhiều biến có thể dẫn đến làm tăng sai số chuẩn của ước lượng và kết quả dự đoán trệch. Vì thế, xác định những biến dự đoán tối ưu là cần thiết. Để đạt được mục đích này, người ta sử dụng hệ số R^2 điều chỉnh (Adjusted $R^2 = R^2_a$).

1.3. Áp dụng hồi quy tuyến tính đa biến trong lâm học

Khi áp dụng phương pháp phân tích hồi quy tuyến tính đa biến để xây dựng các hàm phản hồi của cây gỗ và quần thụ đối với sự thay đổi của các biến X_p , nhà lâm học cần lưu ý một số vấn đề dưới đây.

1. Trong lâm học, hầu hết các biến giải thích đều tồn tại quan hệ với nhau. Vì thế, để làm rõ phản hồi của các Y với các biến X_i , mô hình phải bao gồm tất cả các biến X_i . Vai trò của từng biến X_i được xác định bằng cách xây dựng hàm hồi quy tuyến tính đa biến chuẩn hóa. Thứ tự vai trò của các biến X_i từ lớn đến nhỏ được xác định theo giá trị tuyệt đối của các hệ số hồi quy chuẩn hóa ($|b_K|$). Khi xây dựng các mô hình dự đoán Y , nhà lâm học có thể chọn các biến giải thích hữu ích nhất. Đó là các biến giải thích nhận hệ số tương quan riêng phần cao. Việc làm này cho phép giảm bớt các khó khăn trong thu thập số liệu.
2. Về lý thuyết, cây gỗ và quần thụ chịu ảnh hưởng tổng hợp của vô số nhân tố sinh thái khác nhau. Bởi vì nhà lâm học không thể đo đạc được mọi nhân tố sinh thái, nên mô hình hồi quy tuyến tính đa biến chỉ bao gồm một số nhân tố sinh thái có thể đo đạc được.
3. Khi phân tích hồi quy tuyến tính đa biến, nhà lâm học cần phải xác định thứ tự ảnh hưởng của các biến giải thích. Bởi vì ý nghĩa của mỗi biến giải thích phụ thuộc vào thứ tự mà nó được đưa vào mô hình, nên việc chọn biến nào đưa vào trước và sau trong phân tích hồi quy từng bước phải thay đổi tùy theo mục tiêu nghiên cứu. Chẳng hạn: Mục tiêu nghiên cứu là xác định những tháng mà nhiệt độ không khí ảnh hưởng lớn đến tăng trưởng bề rộng vòng năm của Thông ba lá (*Pinus kesiya* Royle ex Gordon). Trong trường hợp này, nhà lâm học phải sử dụng phương pháp chọn biến từng bước để lọc ra các tháng theo thứ tự từ ảnh hưởng thấp đến ảnh hưởng lớn. Trái lại, mục tiêu nghiên cứu là xây dựng hàm phản hồi

để giải thích ảnh hưởng của nhiệt độ không khí trong năm đến tăng trưởng bề rộng vòng năm của Thông ba lá. Giả sử nhiệt độ không khí tháng 2 và tháng 4 ảnh hưởng tương tự như nhau đến tăng trưởng bề rộng vòng năm của Thông ba lá. Nếu sử dụng phương pháp chọn biến từng bước và đưa nhiệt độ không khí tháng 4 vào trước tháng 2, thì hàm phản hồi không bao gồm tháng 2. Điều đó xảy ra là vì chương trình ưu tiên chọn biến đưa vào trước. Cách chọn biến này là không hợp lý bởi vì cả hai nhân tố này đều ảnh hưởng như nhau đến tăng trưởng bề rộng vòng năm của Thông ba lá. Vì thế, hàm phản hồi phải bao gồm cả nhiệt độ không khí tháng 2 và tháng 4. Khi sử dụng hàm hồi quy để ước lượng và dự đoán tăng trưởng của cây gỗ và quần thụ, thì thứ tự từng biến giải thích trong mô hình là không quan trọng.

4. Báo cáo kết quả phân tích hồi quy tuyến tính đa biến. Các phần mềm thống kê cho phép người sử dụng chọn lựa hai cách báo cáo kết quả phân tích hàm hồi quy tuyến tính đa biến. **Cách thứ nhất:** Báo cáo tóm tắt kết quả phân tích hàm hồi quy thích hợp nhất. Ưu điểm của cách báo cáo này là giúp cho nhà lâm học thu được những hàm phản hồi thích hợp để sử dụng vào việc ước lượng và dự đoán biến phản hồi từ những biến dự đoán hữu ích nhất. **Cách thứ hai:** Báo cáo chi tiết các thông tin của từng bước phân tích hồi quy. Đây là cách báo cáo kết quả rất hữu ích đối với các nghiên cứu trong lâm học, bởi vì những thông tin này giúp ích cho nhà lâm học hiểu rõ về vai trò của từng biến giải thích trong hàm phản hồi. Vì thế, tùy theo mục đích đặt ra, nhà lâm học có thể chọn cách báo cáo.
5. Giải thích phản hồi của cây gỗ và quần thụ đối với sự thay đổi của các biến giải thích. Trong phần thảo luận kết quả nghiên cứu, nhà lâm học cần phải trả lời một số câu hỏi chính dưới đây.
 - (1) Mỗi biến Xi đóng vai trò như thế nào đối với sự thay đổi của biến phản hồi Y? Từ các thông tin trong phân tích hồi quy, nhà lâm học chỉ ra vai trò lớn, nhỏ của từng biến Xi. Thứ tự vai trò lớn, nhỏ của các biến Xi được xác định theo giá trị tuyệt đối của các hệ số hồi quy ($|b_i|$). Vai trò của từng biến Xi còn được đánh giá theo độ lớn của hệ số tương quan riêng phần ($R_{i\text{Change}} = R_{i\text{Thay đổi}}$). Khi hệ số $R_{i\text{Thay đổi}}$ lớn ($P < 0,05$ hay $0,01$), thì sự thay đổi của biến Xi này làm tăng đáng kể hệ số tương quan đa biến (R). Trong báo cáo kết quả, chúng ta chỉ viết ngắn gọn: Biến X_1 đóng vai trò lớn nhất ($|b_1|$); kế đến là biến X_3 ($|b_3|$); nhỏ nhất là biến X_5 ($|b_5|$).
 - (2) Những biến Xi nào ảnh hưởng tốt, còn những biến Xi nào ảnh hưởng xấu đến biến Y?. Những biến Xi ảnh hưởng tốt có hệ số $b_i > 0$, còn những biến Xi ảnh hưởng xấu có hệ số $b_i < 0$. Trong báo cáo kết quả, chúng ta chỉ viết ngắn gọn: Sự gia tăng của các biến $X_1, X_3, \dots, X_5$ ảnh hưởng tốt đến biến Y; trong đó lớn nhất là biến X_3 , nhỏ nhất là biến X_5 . Trái lại, sự gia tăng của các biến $X_2, X_4, \dots, X_6$ ảnh hưởng xấu đến biến Y; trong đó lớn nhất là biến X_4 , nhỏ nhất là biến X_6 .
 - (3) Vì sao sự gia tăng giá trị của các biến Xi lại dẫn đến sự gia tăng Y?. Trái lại, vì sao sự gia tăng giá trị của các biến Xi lại dẫn đến sự suy giảm Y? Đây là hai câu hỏi không dễ tìm ra câu trả lời, nhất là những nghiên cứu chỉ dựa trên số liệu đo đếm từ “Kiểu thu thập số liệu thuần túy” và “Kiểu thu thập số liệu từ các ô mẫu bán định vị”. Để tìm ra câu trả lời gần đúng, nhà lâm học cần phải thiết kế nghiên cứu theo “Kiểu nghiên cứu có khả năng giải thích”.

2. HÀM HỒI QUY BẬC 2

2.1. Hàm hồi quy bậc 2

$$Y = b_0 + b_1X + b_2X^2 + e$$

Y = Biến phụ thuộc hay biến phản hồi.

X = Biến độc lập hay biến giải thích.

b_0 = Điểm chặn (Hằng số) của Y trên trục tung.

b_1 = Hệ số đo sự thay đổi 1 đơn vị của biến Y khi thay đổi 1 đơn vị của biến X.

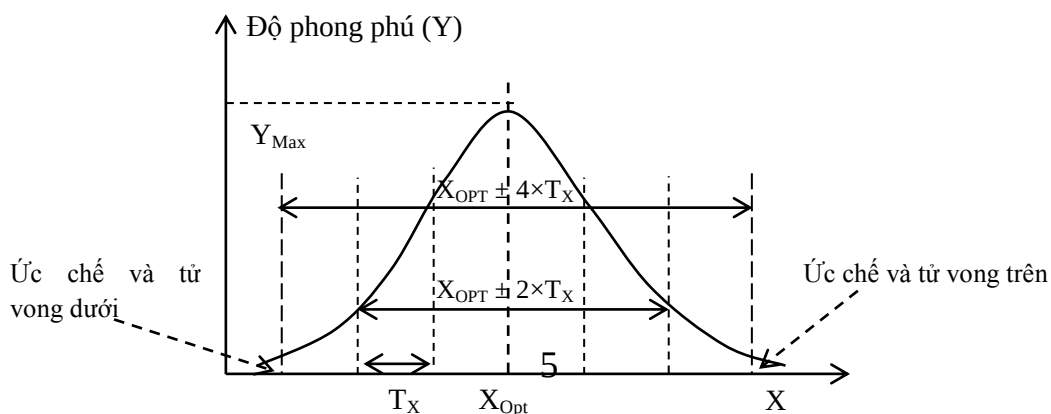
e = Sai số hay sai lệch của hàm hồi quy.

2.2. Điều kiện áp dụng hàm hồi quy bậc 2 trong lâm học

Hàm hồi quy bậc 2 được áp dụng để định lượng phản hồi của các loài cây gỗ và quần thụ (Y = D, H, V, sinh khối...) với biến động của một nhân tố sinh thái nhất định (X). Khi xác định phản hồi của Y với 1 nhân tố sinh thái nhất định (Ví dụ X1), thì các nhân tố sinh thái khác phải ở mức cố định hay hằng số (X_i = Hằng số). Mức cố định của X_i không ở mức giới hạn (Vùng ức chế và tử vong) đối với sự sống của cây gỗ và quần thụ.

2.3. Xác định các tham số sinh thái

Sau khi xây dựng hàm phản hồi, khảo sát hàm bậc 2 để xác định vùng tối ưu sinh thái, biên độ sinh thái (Vùng hoạt động bình thường) và phạm vi chống chịu (Vùng sống) của cây gỗ và quần thụ. Giá trị trung tâm (X_{Bq}) của vùng tối ưu sinh thái là tối ưu sinh thái (Hình 1). Tối ưu sinh thái là giá trị của biến giải thích (X) đảm bảo cho cây gỗ và quần thụ sống tốt nhất. Biên độ sinh thái là phạm vi của biến giải thích đảm bảo cho cây gỗ và quần thụ sống bình thường. Phạm vi chống chịu là ranh giới của biến giải thích đảm bảo cho cây gỗ và quần thụ có thể sống. Các giá trị $X < (X_{Bq} - 4 \times S)$ và $X > (X_{Bq} + 4 \times S)$ là vùng bị ức chế và tử vong. Trong hai vùng này, xác suất bắt gặp cây gỗ và quần thụ còn sống là nhỏ hơn 0,1%. Độ phong phú lớn nhất (Y_{Max}) là giá trị của biến phản hồi cao nhất tại giá trị của biến giải thích tối ưu (X_{Opt}).



Hình 1. Đường cong phản hồi của cây gỗ và quần thụ với biến X.

Tối ưu sinh thái, vùng tối ưu sinh thái, vùng tác động bình thường của nhân tố sinh thái phạm vi chống chịu và phản hồi lớn nhất được xác định tương ứng theo công thức 2.1÷2.5; trong đó giá trị $\pm T_X$ là biên độ sinh thái.

$$\begin{array}{l} \text{Tối ưu sinh thái} \\ X_{OPT} = \frac{b_1}{2b_2} \end{array} \quad (2.1)$$

$$\begin{array}{l} \text{Vùng tối ưu sinh thái} \\ X_{68\%} = X_{OPT} \pm T_X \end{array} \quad (2.2)$$

$$\begin{array}{l} \text{Vùng tác động bình thường} \\ X_{95\%} = X_{OPT} \pm 2 \times T_X \end{array} \quad (2.3)$$

$$\begin{array}{l} \text{Phạm vi chống chịu} \\ X_{99,9\%} = X_{OPT} \pm 4 \times T_X \end{array} \quad (2.4)$$

$$\begin{array}{l} \text{Biên độ sinh thái} \\ T_X = \frac{1}{\sqrt{2b_2}} \end{array}$$

$$\begin{array}{l} \text{Phản hồi lớn nhất} \\ Y_{Max} = b_0 + b_1 X_{OPT} - b_2 X_{OPT}^2 \end{array} \quad (2.5)$$

3. HÀM HỒI QUY Sigmoid và Logit Gauss

3.1. Hàm hồi quy Sigmoid

Hàm hồi quy Sigmoid mô tả quan hệ giữa biến phản hồi (Y) ở dạng nhị phân (1 và 0) và các biến giải thích (X) ở dạng định lượng.

Biến phản hồi kỳ vọng là xác suất xuất hiện một sự kiện tương ứng với các biến giải thích.

Trong trường hợp một biến giải thích, hàm phản hồi có dạng như [Hàm 3.1 và 3.2](#). Đồ thị biểu diễn xác suất của [Hàm 3.2](#) có dạng một đường cong hình chữ S. Xác suất lấy giá trị giữa 0 và 1. Các tham số của [Hàm 3.1](#) được ước lượng theo nguyên lý hợp lý tối đa. Khi thay thế thành phần dự đoán tuyến tính ở [Hàm 3.1](#) bằng [Hàm bậc 2](#), chúng ta có hàm Logit Gauss ([Hàm 3.3 và 3.4](#)); trong đó tham số b_2 nhận giá trị âm. Trong trường hợp hai biến dự đoán, hàm Logit Gauss 2 biến số có dạng như [Hàm 3.5 và 3.6](#); trong đó hai tham số b_2 và b_4 nhận giá trị âm.

Khi phân tích ảnh hưởng phối hợp của hai biến dự đoán ($X_1 \times X_2$), hàm Logit Gauss 2 biến số có dạng như [Hàm 3.7 và 3.8](#).

$$\ln[P/(1 - P)] = b_0 + b_1 X \quad (3.1)$$

$$P = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)} \quad (3.2)$$

$$\ln[P/(1 - P)] = b_0 + b_1 X - b_2 X^2 \quad (3.3)$$

$$P = \frac{\exp(b_0 + b_1 X - b_2 X^2)}{1 + \exp(b_0 + b_1 X - b_2 X^2)} \quad (3.4)$$

$$\ln[P/(1 - P)] = b_0 + b_1 X_1 - b_2 X_1^2 + b_3 X_2 - b_4 X_2^2 \quad (3.5)$$

$$P = \frac{\exp(b_0 + b_1X_1 - b_2X_1^2 + b_3X_2 - b_4X_2^2)}{[1 + \exp(b_0 + b_1X_1 - b_2X_1^2 + b_3X_2 - b_4X_2^2)]} \quad (3.6)$$

$$\ln[P/(1 - P)] = b_0 + b_1X_1 - b_2X_1^2 + b_3X_2 - b_4X_2^2 + b_5X_1 \times X_2 \quad (3.7)$$

$$P = \frac{\exp(b_0 + b_1X_1 - b_2X_1^2 + b_3X_2 - b_4X_2^2 + b_5X_1 \times X_2)}{[1 + \exp(b_0 + b_1X_1 - b_2X_1^2 + b_3X_2 - b_4X_2^2 + b_5X_1 \times X_2)]} \quad (3.8)$$

3.2. Ứng dụng hàm hồi quy Logit Gauss trong lâm học

Trong lâm học, hàm hồi quy Logit Gauss được ứng dụng để phân tích biến động độ phong phú của các loài cây gỗ trong quan hệ với biến động của các nhân tố sinh thái. Bằng cách xây dựng các hàm phản hồi ở dạng [Hàm 3.3, 3.5 và 3.7](#), xác định được biên độ sinh thái ($\pm T_i$), tối ưu sinh thái (U_i), vùng tối ưu sinh thái ($U_i + T_i$), vùng tác động bình thường ($U_i \pm 2T_i$), biên độ sống sót ($U_i \pm 4T_i$) và xác suất bất gặp độ phong phú cao nhất (P_{iMax}). Sáu tham số này được xác định theo [Hàm 3.9-3.14](#).

Biên độ sinh thái:	$T_i = 1/\sqrt{2b_2}$	(3.9)
--------------------	-----------------------	-------

Vùng tối ưu sinh thái:	$U_i \pm T_i$	(3.10)
------------------------	---------------	--------

Tối ưu sinh thái:	$U_i = (b_1/2b_2)$	(3.11)
-------------------	--------------------	--------

Vùng tác động bình thường:	$U_i \pm 2T_i$	(3.12)
----------------------------	----------------	--------

Biên độ sống sót:	$U_i \pm 4T_i$	(3.13)
-------------------	----------------	--------

Xác suất lớn nhất	$P_{iMax} = \frac{\exp(b_0 + b_1U_i + b_2U_i^2)}{1 + \exp(b_0 + b_1U_i + b_2U_i^2)}$	(3.14)
-------------------	--	--------

Cây gỗ và quần thụ chịu ảnh hưởng của nhiều nhân tố sinh thái khác nhau. Vì thế, mô hình hồi quy Logit Gauss cần được xây dựng với nhiều biến giải thích khác nhau. Phương pháp phân tích hàm Logit Gauss với nhiều biến số giúp cho nhà lâm học phân tích ảnh hưởng tổng hợp của nhiều nhân tố sinh thái đến cây gỗ và quần thụ; trong đó biến phản hồi ở dạng bất gặp và không bất gặp, còn biến giải thích ở dạng định lượng. Bằng cách chọn biến từng bước, nhà lâm học có thể phân tích và lọc ra những tập hợp biến giải thích hữu ích nhất. Đây là phương pháp nghiên cứu được áp dụng rộng rãi trong lâm học.

3.3. Những lưu ý khi áp dụng hàm hồi quy Logit Gauss trong lâm học

1. Khi phân tích biến động độ phong phú của các loài cây gỗ (Y) trong quan hệ với biến động của một nhân tố sinh thái nhất định (X) bằng hàm hồi quy Logit Gauss, thì các nhân tố sinh thái khác phải ở dạng ổn định ($Z_i = \text{Hằng số}$).
2. Trong điều kiện ở rừng tự nhiên, các nhân tố sinh thái Z_i không phải là hằng số. Lí do là vì điều kiện môi trường là không đồng nhất. Ví dụ: Khi xem xét phản hồi của cây tái sinh Dầu song nòng (Y) với biến động của độ ẩm đất dưới tán rừng tự nhiên (X), thì các nhân tố sinh thái khác (Độ tàn che, cây bụi, thảm tươi, địa hình...) là không đồng nhất. Để đảm bảo điều kiện cơ bản trong phân tích hàm hồi quy Logit Gauss, nhà lâm học phải giả định các nhân tố sinh thái khác là đồng nhất. Ở ví dụ này, nhà lâm học phải giả định độ tàn che, cây bụi, thảm tươi, địa hình... trong một trạng thái rừng nhất định là ổn định.

3. Khi các nhân tố sinh thái Z_i không phải là hằng số, thì hàm $Y = f(X)$ có thể ở dạng hàm bậc k ($k = 3, 4, \dots$). Trong trường hợp này xảy ra, nhà lâm học phải phân chia các nhân tố sinh thái Z_i thành từng cấp tương đối đồng nhất. Ví dụ: Nếu một trạng thái rừng phân bố trên nhiều loại đất, thì nhà lâm học phải phân chia trạng thái rừng này theo từng loại đất. Ở mỗi trạng thái rừng trên 1 loại đất, nhà lâm học phân tích phản hồi của cây tái sinh với độ ẩm đất.

4. HÀM HỒI QUY LOGISTIC

4.1. Khái quát về hàm hồi quy Logistic

Hồi quy Logistic được sử dụng để làm phù hợp mô hình hồi quy; trong đó biến phụ thuộc (Y) đặc trưng cho một sự kiện với chỉ 2 kết quả. Đối với các biến Y ở dạng tỷ lệ của một sự kiện nào đó, hàm Logistic có dạng hàm tuyến tính (Công thức 4.1); trong đó hệ số b_i mô tả mối quan hệ giữa Y với các biến dự đoán (X_i). Đối với các biến ngẫu nhiên Y ở dạng nhị phân (1, 0), hàm Logistic có dạng hàm tuyến tính (Công thức 4.2); trong đó $P_{(Y=1)}$ là xác suất bắt gặp sự kiện Y , b_i là hệ số mô tả mối quan hệ giữa $P_{(Y=1)}$ với các biến dự đoán (X_i).

Để đo đặc ảnh hưởng của các biến X_i đến xác suất bắt gặp Y ở hai hay nhiều đối tượng khác nhau, người ta sử dụng số đo tỷ lệ sai khác (Odds Ratio = OR). Sự sai khác (Odds) là tỷ lệ giữa xác suất bắt gặp và không bắt gặp một sự kiện nào đó (Công thức 4.3).

Tỷ lệ sai khác (OR) giữa hai đối tượng là tỷ lệ của hai sự sai khác. Tỷ lệ sai khác giữa 2 đối tượng A và B (OR_{A-B}) được tính theo công thức 4.4.

$$E_{(Y)} = \frac{1}{1 + \exp[-(b_0 + \sum_{i=1}^k b_i X_i)]} \quad (4.1)$$

$$P_{(Y=1)} = \frac{1}{1 + \exp[-(b_0 + \sum_{i=1}^k b_i X_i)]} \quad (4.2)$$

$$\text{Odds}(D) = \frac{P_{(D)}}{1 - P_{(D)}} \quad (4.3)$$

$$OR_{(A-B)} = \text{Odds}(D_A) / \text{Odds}(D_B) = \left[\frac{P_{(DA)}}{1 - P_{(DA)}} \right] / \left[\frac{P_{(DB)}}{1 - P_{(DB)}} \right] \quad (4.4)$$

Khi biến đổi ở dạng Logit, hàm Logistic 4.2 được viết ở dạng Hàm 4.5. Hàm Logistic có thể được sử dụng để so sánh tỷ lệ sai khác giữa các nhóm dựa trên một tập hợp các biến dự đoán $X_1, X_2, \dots, X_k$. Nhóm A bao gồm các biến dự đoán $X_A = (X_{A1}, X_{A2}, \dots, X_{Ak})$, còn nhóm B bao gồm các biến dự đoán $X_B = (X_{B1}, X_{B2}, \dots, X_{Bk})$. Khi $k = 3$, hàm Logit có dạng như Hàm 4.6. Tỷ lệ sai khác giữa hai nhóm được tính theo công thức 4.7. Tỷ lệ sai khác giữa hai đối tượng theo mỗi biến giải thích được xác định theo công thức 4.8.

$$\text{Logit (Odds)} = \text{Ln} \left[\frac{P_{(Y=1)}}{1 - P_{(Y=1)}} \right] = b_0 + \sum_{i=1}^k b_i X_i \quad (4.5)$$

$$\text{Logit}(P/1 - P) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \quad (4.6)$$

$$\text{OR}_{(X_A - X_B)} = \exp[\sum_{i=1}^k (X_A - X_B)b_i] \quad (4.7)$$

$$\text{OR}_{(X_A - X_B)} = \exp(b_i) \quad (4.8)$$

Nói chung, hồi quy Logistic được sử dụng cho 2 kiểu số liệu: (1) Biến phụ thuộc (Y) là tỷ lệ xuất hiện của một sự kiện nào đó; (2) Biến phụ thuộc (Y) được đo bằng hai biến nhị phân (1 và 0).

Khi biến phụ thuộc là tỷ lệ của một sự kiện nào đó, thì các hệ số của hàm Logistic được ước lượng bằng phương pháp bình phương nhỏ nhất có trọng số (Weighted Least Squares = WLS).

Khi biến phụ thuộc là hai biến nhị phân, thì các hệ số của hàm Logistic được ước lượng bằng phương pháp hợp lý tối đa (Maximum Likelihood = MLE).

4.2. Ứng dụng hàm hồi quy Logistic trong lâm học

4.2.1. Trường hợp biến phụ thuộc ở dạng tỷ lệ

Trong lâm học, một số đặc tính của cây gỗ và quần thụ có thể được đo đặc bằng những giá trị tỷ lệ. Chẳng hạn: (a) Tỷ lệ số cây bị bệnh ở những cấp tuổi hay cấp kích thước khác nhau; (b) Tỷ lệ số cây gỗ quý trong các cấp D; (c) Tỷ lệ số cây tái sinh bị cụt ngọn trong các cấp tuổi hay cấp H... Đối với những trường hợp này, nhà lâm học có thể sử dụng hàm Logistic để mô tả quan hệ giữa tỷ lệ xuất hiện một sự kiện và biến giải thích.

4.2.2. Trường hợp biến phụ thuộc ở dạng nhị phân

Trong lâm học, một số đặc tính của cây gỗ và quần thụ có thể được đo đặc bằng độ bất gặp (1, 0). Chẳng hạn: (a) Độ bất gặp loài cây tái sinh của những loài cây gỗ quý trong những lớp H khác nhau; (b) Độ bất gặp loài cây gỗ quan tâm trong quan hệ với một số nhân tố sinh thái...

Khi nghiên cứu về quan hệ giữa rừng và đất, nhà lâm học có thể đặt ra một số câu hỏi. Một là độ bất gặp loài cây gỗ quan tâm có thay đổi theo loại đất hay không? Hai là tính chất của mỗi loại đất thay đổi ảnh hưởng đến độ phong phú của loài cây gỗ quan tâm như thế nào? Đối với trường hợp này, hàm Logistic là một công cụ hữu ích để giúp nhà lâm học trả lời rõ hai câu hỏi đặt ra trên đây. Dưới đây dẫn ví dụ áp dụng hàm Logistic để phân tích quan hệ giữa rừng và đất.

5. HÀM HỒI QUY POISSON

5.1. Khái quát về hàm hồi quy Poisson

Hồi quy Poisson mô tả quan hệ giữa biến phụ thuộc Y với các biến dự đoán X_i . Biến Y là các số nguyên không âm. Các biến X_i có thể là biến nhị phân (1, 0) hoặc biến định lượng.

Mô hình phân bố xác suất Poisson có dạng như [Hàm 5.1](#); trong đó Y_i = tần số xuất hiện sự kiện Y thứ i, T_i = dung lượng mẫu quan sát sự kiện Y_i , λ_i = tỷ lệ của biến phụ thuộc (Y_i) thứ i ($\lambda_i = Y_i/T_i$), $\lambda_i T_i$ = giá trị trung bình của Y_i .

Ở [Hàm 5.1](#), hai thành phần Y_i và T_i đã biết. Tham số λ_i được ước lượng theo các biến dự đoán X_i .

Tham số λ_i và các biến dự đoán X_i có quan hệ với nhau theo dạng Logarit tuyến tính ([Hàm 5.2](#)). Các tham số b_i ở [Hàm 5.2](#) phản ánh tỷ lệ (Phần mười hay %) thay đổi của biến Y_i khi biến X_i thay đổi 1 đơn vị.

Khi biết λ_i , thì biến phụ thuộc $Y_i = \lambda_i T_i$. Tỷ lệ thay đổi của biến phụ thuộc Y_i theo mỗi đơn vị của biến dự đoán X_i được xác định theo [công thức 5.3](#); trong đó RR_i = tỷ lệ thay đổi (Rate Ratio), λ_{i1} = tỷ lệ bắt gặp, λ_{i0} = tỷ lệ không bắt gặp.

$$\frac{\exp(-\lambda_i T_i)(\lambda_i T_i)^{Y_i}}{Y_i!} \quad (5.1)$$

$$\ln(\lambda) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k \quad (5.2a)$$

$$\lambda = \exp(b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k) \quad (5.2b)$$

$$RR_i = \frac{\lambda_{i1}}{\lambda_{i0}} = \exp(b_i) \quad (5.3)$$

Khi xây dựng hàm hồi quy Poisson, các tham số b_i được ước lượng theo phương pháp Maximum Likelihood. Hàm hồi quy Poisson có thể xây dựng ở dạng bậc 1 và bậc 2. Dạng bậc 1 chỉ bao gồm ảnh hưởng chính. Dạng bậc 2 bao gồm ảnh hưởng chính của các biến giải thích và sự tương tác giữa 2 biến giải thích. Khi mô hình bao gồm nhiều biến giải thích, thì hồi quy Poisson có thể bao gồm tất cả các biến đưa vào mô hình hoặc chỉ bao gồm một số biến giải thích có ý nghĩa lớn. Ở trường hợp sau, những biến giải thích có ý nghĩa lớn xác định theo phương pháp phân tích hồi quy tương quan từng bước.

5.2. Ứng dụng hàm hồi quy Poisson trong lâm học

Trong lâm học, hồi quy Poisson được sử dụng để mô tả sự xuất hiện của những sự kiện hiếm gặp trong các QXTV rừng. Chẳng hạn: (a) Số cây tái sinh của một loài cây gỗ hiếm xuất hiện trong những QXTV; (b) Số cây bị bệnh rỗng ruột ở các cấp D khác nhau.

Nguyễn Văn Thêm