**Tóm tắt**

Dữ liệu thành phần mang thông tin quan trọng của chúng trong các mối quan hệ (logratios) giữa các phần thành phần. Bài viết này chỉ ra cách sử dụng nguồn thông tin này trong mô hình hồi quy, trong đó thành phần có thể là biến đáp ứng, biến giải thích, hoặc cả hai. Một bước quan trọng để thiết lập mô hình hồi quy là cách thành phần (các thành phần) được đưa vào mô hình. Ở đây, các tọa độ cân bằng sẽ được xây dựng để hỗ trợ việc diễn giải các hệ số hồi quy và cho phép kiểm định giả thuyết về sự độc lập của các thành phần con. Cả hồi quy cổ điển theo phương pháp bình phương nhỏ nhất và hồi quy MM bền vững đều được xem xét, và chúng được so sánh trong các mô hình hồi quy khác nhau trên một bộ dữ liệu thực từ một dự án lập bản đồ địa hóa học.

**1. Giới thiệu**

Mặc dù phân tích hồi quy là một trong những quy trình thống kê phát triển nhất, nhưng tài liệu về dữ liệu thành phần thì không nhiều (xem, ví dụ, Aitchison 1986; Daunis-i-Estadella et al. 2002; Tolosana-Delgado và van den Boogaart 2011; van den Boogaart và Tolosana-Delgado 2013; Egozcue et al. 2013; Pawlowsky-Glahn et al. 2015; Fišerová et al. 2016; Coenders et al. 2017; Filzmoser et al. 2018; Greenacre 2019). Hồi quy với dữ liệu thành phần có những đặc điểm riêng do thang đo thống kê đặc biệt của các thành phần. Các thành phần thường được định nghĩa (và giới hạn) là các vectơ của các thành phần dương tổng bằng một hằng số, với simplex là không gian lấy mẫu của chúng (Aitchison 1986). Tuy nhiên, từ đầu thế kỷ XXI, người ta ngày càng nhận ra rằng dữ liệu có thể không tuân theo điều kiện tổng hằng số và vẫn được coi là dữ liệu thành phần, điểm quan trọng là tính tương đối của thông tin cung cấp (Aitchison 1997; Barceló-Vidal et al. 2001). Điều này có nghĩa là khi xử lý dữ liệu thành phần, các đại lượng quan trọng được hình thành bởi tỷ lệ giữa các phần thành phần (tức là các biến) hơn là bởi các giá trị dữ liệu được báo cáo trực tiếp.

Có nhiều đề xuất để trích xuất thông tin tương đối này, thường được gọi là các phép biến đổi hoặc tọa độ, với phép biến đổi logratio cộng (alr), phép biến đổi logratio tập trung (clr), và phép biến đổi logratio đẳng hình (ilr) là các đại diện nổi tiếng nhất (van den Boogaart và Tolosana-Delgado 2013; Pawlowsky-Glahn et al. 2015; Filzmoser et al. 2018). Các điểm số thu được từ các phép biến đổi này có thể được gọi là "tọa độ" nhờ cấu trúc không gian Euclid của simplex (Billheimer et al. 2001; Pawlowsky-Glahn và Egozcue 2001a), được mô tả chi tiết hơn ở Phần 2.

Do tính tương đương affine của nó, hồi quy tuyến tính được biết đến là cung cấp kết quả chính xác như nhau dù sử dụng bất kỳ phép biến đổi logratio nào (van den Boogaart và Tolosana-Delgado 2013), mặc dù sử dụng phép biến đổi clr có thể dẫn đến các phức tạp số không cần thiết (nghịch đảo tổng quát của ma trận hiệp phương sai suy biến) khi thành phần đóng vai trò giải thích. Thậm chí hơn nữa, các mô hình hồi quy với dữ liệu thành phần có thể được thiết lập mà không cần sử dụng bất kỳ phép biến đổi logratio nào. Thực tế, điều này cho phép xem xét các đối tượng xuất hiện trong phân tích hồi quy (độ dốc, điểm chặn, gradient, phần dư, dự đoán, v.v.) như là các đối tượng thành phần hoàn toàn nội tại, việc sử dụng một trong những phép biến đổi logratio là lựa chọn cách biểu diễn. Bài báo này chỉ sử dụng biểu diễn logratio đẳng hình do liên kết mật thiết của nó với các kiểm định loại bỏ các thành phần đơn lẻ hoặc các tập hợp con thành phần.

Nói chung, cần phân biệt các loại mô hình hồi quy khác nhau liên quan đến dữ liệu thành phần. Thành phần có thể là phần phản hồi trong mô hình hồi quy, và phần giải thích bao gồm một hoặc nhiều tính năng không phải thành phần. Điều này dẫn đến mô hình hồi quy đa biến, sau này được gọi là mô hình Loại 1. Vấn đề ngược lại, thành phần là phần giải thích, phản hồi không phải thành phần, được gọi là mô hình Loại 2 trong bài viết này; nếu phản hồi chỉ là đơn biến, chúng ta sẽ có mô hình hồi quy tuyến tính bội. Mô hình Loại 3 được đặc trưng bởi một thành phần làm phần giải thích và một thành phần khác làm phần phản hồi, trở thành mô hình hồi quy tuyến tính bội đa biến. Đối với mỗi loại mô hình được xem xét, chúng ta có thể xây dựng mô hình hồi quy khai thác cấu trúc Euclid của simplex, hoặc dưới dạng các tọa độ ilr hoặc các điểm số biến đổi clr. Tuy nhiên, do có vô số khả năng để xây dựng tọa độ ilr (Egozcue et al. 2003), người ta có thể tập trung vào những tọa độ cho phép giải thích mô hình và các hệ số hồi quy tương ứng. Thậm chí hơn nữa, nếu cần kiểm định các giả thuyết cụ thể, điều quan trọng là phải xây dựng các tọa độ hỗ trợ vấn đề kiểm định. Bài viết này tập trung đặc biệt vào các kiểm định gọi là kiểm định độc lập thành phần con, hay đúng hơn là không tương quan. Ở đây, mục tiêu là thiết lập các thành phần con (hoặc các thành phần đơn lẻ) không phụ thuộc hoặc không ảnh hưởng đến các biến đồng biến xem xét.

Một vấn đề khác cũng liên quan đối với hồi quy không phải thành phần truyền thống là ước lượng tham số. Phương pháp được sử dụng rộng rãi nhất là hồi quy bình phương tối thiểu, tối thiểu hóa tổng bình phương phần dư. Khi có các giá trị ngoại lai trong biến giải thích và/hoặc biến phản hồi, người ta biết rằng bộ ước lượng này có thể bị lệch mạnh, và các đối tác mạnh nên được ưu tiên (Maronna et al. 2006). Ở đây, bộ ước lượng MM rất mạnh cho hồi quy được xem xét, cho cả mô hình hồi quy tuyến tính bội và hồi quy tuyến tính đa biến (Van Aelst và Willems 2013). Cũng có kỹ thuật sẵn có cho suy luận thống kê mạnh (Salibian-Barrera et al. 2008). Các kỹ thuật như vậy là rất quan trọng trong dữ liệu thành phần, vì các giá trị nhỏ (gần hoặc dưới giới hạn xác định, số không làm tròn, số không đếm) có thể trở nên ảnh hưởng mạnh sau các phép biến đổi logratio.

Trong bài viết này, chúng tôi định nghĩa rõ ràng các loại mô hình hồi quy khác nhau liên quan đến dữ liệu thành phần, với các diễn giải hình học, và giải thích việc sử dụng chúng trong một nghiên cứu trường hợp. Trong Phần 2, chúng tôi cung cấp thông tin chi tiết hơn về phân tích dữ liệu thành phần, và định nghĩa ba loại mô hình được xem xét. Phần 3 đề cập đến ước lượng bình phương tối thiểu, và tùy thuộc vào mô hình, chúng tôi sẽ chỉ ra cách ước lượng tham số và kiểm định giả thuyết có thể được thực hiện. Ước lượng và kiểm định giả thuyết sử dụng hồi quy MM mạnh được đề cập ở Phần 4. Cả ba mô hình tuyến tính và ước lượng cổ điển cũng như mạnh được minh họa ở Phần 5 với một tập dữ liệu từ dự án GEMAS, một khảo sát địa hóa học đất quy mô lục địa ở Châu Âu. Phần cuối cùng 6 tóm tắt và kết luận.

**2 Compositional Data Analysis**

**2.1 Compositional Geometry  
2.2 Các Mô Hình Tuyến Tính Thành Phần**

Có ba loại mô hình tuyến tính liên quan đến các thành phần đã được định nghĩa (van den Boogaart và Tolosana-Delgado 2013; Pawlowsky-Glahn et al. 2015; Filzmoser et al. 2018): mô hình với biến đáp ứng thành phần (Aitchison 1986; Daunis-i-Estadella et al. 2002), mô hình với biến giải thích thành phần (Aitchison 1986; Tolosana-Delgado và van den Boogaart 2011), và mô hình với cả biến giải thích và biến đáp ứng đều là các thành phần (Egozcue et al. 2013). Các phần tiếp theo sẽ xây dựng hệ thống mỗi mô hình hồi quy chỉ bằng các phép toán hình học và chỉ ra cách các mô hình này được biểu diễn trong phép biến đổi logratio đẳng cự tùy ý.

**2.2.1 Mô Hình Tuyến Tính Với Biến Đáp Ứng Thành Phần (Loại 1)**

Một mô hình với biến đáp ứng thành phần giả định rằng một thành phần ngẫu nhiên \(Y\) là một hàm tuyến tính (theo nghĩa của hình học Aitchison) của một số biến ngẫu nhiên thực giải thích \(X\_0, X\_1, ..., X\_P\), điều này cho giá trị kỳ vọng của một số thành phần phân phối chuẩn,

\[ \hat{Y} = \sum\_{i=0}^P X\_i \cdot b\_i, \quad Y \sim \mathcal{NS}\_D(\hat{Y}, \Sigma\_\epsilon), \tag{8} \]

trong đó \(\mathcal{NS}\_D(\hat{Y}, \Sigma\_\epsilon)\) đại diện cho phân phối chuẩn trên simplex của \(Y\) (Mateu-Figueras và Pawlowsky-Glahn 2008), được tham số hóa theo vector trung bình thành phần và ma trận hiệp phương sai của thành phần ngẫu nhiên trong một biểu diễn ilr nào đó. Điều này phản ánh thực tế rằng phân phối chuẩn trên simplex của một thành phần ngẫu nhiên tương ứng với phân phối chuẩn (thông thường) của biểu diễn ilr của nó. Mô hình hồi quy này hữu ích cho các biến giải thích thuộc loại định lượng (hồi quy), loại phân loại (ANOVA) hoặc kết hợp cả hai (ANCOVA). Lưu ý rằng có thể thiết lập mô hình hồi quy này cho dữ liệu thành phần theo nghĩa bình phương nhỏ nhất (Mood et al. 1974, Chương X), không phụ thuộc vào giả định phân phối chuẩn, bằng cách sử dụng khoảng cách Aitchison [Eq. (2)] như Daunis-i-Estadella et al. (2002) đã đề xuất. Tuy nhiên, giả định phân phối chuẩn là cần thiết trong bối cảnh kiểm định giả thuyết, đây là một trong những đóng góp chính của bài báo này. Cụ thể, nó phục vụ cho việc suy ra phân phối của các thống kê kiểm định trong trường hợp hồi quy (bình phương nhỏ nhất) cổ điển và cũng phục vụ như mô hình tham chiếu cho hồi quy bền vững.

Nếu một phép biến đổi logratio được áp dụng cho mô hình này, điều này sẽ dẫn đến một mô hình hồi quy tuyến tính đa biến thông thường trên các tọa độ

\[ \hat{Y}^\* = \sum\_{i=0}^P X\_i \cdot b^\*\_i, \quad Y^\* \sim \mathcal{N}\_{D-1}(\hat{Y}^\*, \Sigma\_\epsilon). \tag{9} \]

Các tham số của mô hình do đó là các độ dốc \(b^\*\_0, b^\*\_1, ..., b^\*\_P\), và ma trận hiệp phương sai phần dư \(\Sigma\_\epsilon\). Lưu ý rằng thường lấy \(X\_0 \equiv 1\) và sau đó \(b^\*\_0\) thay thế cho hệ số giao điểm của mô hình trong hệ tọa độ logratio đã chọn. Cụ thể trong Eq. (9) có lợi thế là có thể xử lý bằng phần mềm và các phương pháp giải thông thường. Khi các ước lượng của các hệ số vector có sẵn, chúng có thể được biến đổi ngược lại thành các hệ số thành phần, ví dụ \( \hat{b}\_i = \text{ilr}^{-1}(\hat{b}^\*\_i) \) nếu các phép tính được thực hiện trong các tọa độ ilr. Ngoài ra, các tọa độ ilr cũng có thể được chuyển đổi thành các hệ số clr với \( \hat{b}\_{clr\_i} = V \cdot \hat{b}^\*\_i \).

Điều quan trọng cần nhấn mạnh là các dự đoán được cung cấp bởi mô hình hồi quy này là không thiên lệch cả về mặt bất kỳ biểu diễn logratio nào [Eq. (9)], và về mặt thành phần gốc [Eq. (8)] với hình học Aitchison được thảo luận trong Mục 2.1. Điều này trực tiếp từ tính đẳng cự của các phép biến đổi ilr hoặc clr (Egozcue et al. 2012; Pawlowsky-Glahn et al. 2015; Fišerová et al. 2016). Nếu quan tâm đến việc hiểu các tính chất không thiên lệch của các dự đoán (8) với hình học Euclide thông thường của không gian đa biến thực \( \mathbb{R}^D \), tức là về bản chất của giá trị kỳ vọng của \( \hat{Y} - Y \), thì có thể sử dụng tích phân số của mô hình được giải thích bởi Eq. (8), cái cung cấp phân phối điều kiện của \(Y\) cho trước \(\hat{Y}\) (Aitchison 1986).

**2.2.2 Regression with Compositional Explanatory Variable (Type 2)**

**2.2.3 Compositional to Compositional Regression (Type 3)**  
  
**2.3 Độc Lập Thành Phần Con**

Một trong những nhiệm vụ phổ biến nhất của hồi quy là xác nhận một mô hình cụ thể dựa trên dữ liệu, đặc biệt là các mô hình (tuyến tính) độc lập, một phần hoặc hoàn toàn. Trong khung phi thành phần, sự độc lập được xác định bởi một ma trận/vector độ dốc hay gradient hoàn toàn bằng không (độc lập hoàn toàn), hoặc chỉ với một số hệ số bằng không (độc lập một phần). Sự độc lập hoàn toàn đối với các mô hình thành phần cũng được xác định bởi độ dốc bằng không, vector gradient bằng không, hoặc ma trận bằng không của mô hình được thiết lập cho các tọa độ (mỗi độ dốc \( b^\*\_i \), gradient \( b^\* \) hoặc ma trận \( B^\* \)). Tuy nhiên, việc đặt một hệ số của các vector hoặc ma trận này bằng không chỉ buộc sự độc lập của các biến đồng biến với một logratio nhất định, chứ không phải với các thành phần mà logratio này liên quan. Do đó, khái niệm cần thiết trong bối cảnh này là sự độc lập thành phần con, nghĩa là một tập hợp con của các thành phần không ảnh hưởng đến hoặc không bị ảnh hưởng bởi một biến đồng biến. Cần phải phân biệt thêm hai trường hợp, cụ thể là sự độc lập thành phần con nội bộ và ngoại bộ.

Xét các thành phần đầu tiên của một thành phần là độc lập với một biến đồng biến nhất định. Khi đó, có thể xây dựng một cơ sở của \( S\_D \) với ba khối: một cơ sở tùy ý của s-1 vector so sánh các thành phần đầu tiên (thành phần con độc lập), vector cân bằng giữa hai thành phần con (Eq. 7), và một cơ sở tùy ý của r-1 vector so sánh các thành phần cuối cùng \( r = D-1 \) (thành phần con phụ thuộc).

Trong một mô hình hồi quy loại 1 (biến đáp ứng thành phần), sự độc lập nội bộ của một thành phần con nhất định đối với biến đồng biến thứ i \( X\_i \) có nghĩa là biến này không thể thay đổi các quan hệ giữa các thành phần của thành phần con độc lập, tức là \( b^\*\_{1i} = b^\*\_{2i} = ... = b^\*\_{(s-1)i} = 0 \). Sự độc lập ngoại bộ thêm vào giả định rằng tọa độ cân bằng là độc lập với biến đồng biến, \( b^\*\_{si} = 0 \).

Trong một mô hình hồi quy loại 2 (biến đầu vào thành phần), sự độc lập nội bộ có nghĩa là biến giải thích \( Y \) không thể thay đổi do các biến đổi trong thành phần con độc lập, tức là gradient tọa độ thỏa mãn \( b^\*\_{1} = b^\*\_{2} = ... = b^\*\_{(s-1)} = 0 \). Sự độc lập ngoại bộ thêm vào giả định rằng biến giải thích chỉ phụ thuộc vào các quan hệ trong thành phần con phụ thuộc, do đó thêm vào \( b^\*\_{s} = 0 \).

Sự độc lập thành phần con đối với một mô hình hồi quy loại 3 (thành phần đối thành phần) thừa kế từ các khái niệm được đề cập trước đó. Thành phần con đáp ứng được hình thành bởi các phần đầu tiên của nó là \( s\_y \) phần, là độc lập nội bộ với thành phần con đầu vào được hình thành bởi các phần đầu tiên của nó là \( s\_x \) phần nếu không có sự thay đổi nào trong thành phần sau có thể gây ra bất kỳ thay đổi nào trong thành phần trước, tức là \( b^\*\_{ij} = 0 \) với \( i = 1, 2, ..., (s\_y - 1) \) và \( j = 1, 2, ..., (s\_x - 1) \). Tương tự, sự độc lập ngoại bộ thêm vào giả định \( b^\*\_{ij} = 0 \) với \( i = 1, 2, ..., s\_y \) và \( j = 1, 2, ..., s\_x \), bao gồm cả các hệ số hồi quy liên quan đến các yếu tố cân bằng.

Nếu nghiên cứu trường hợp hoặc câu hỏi hiện tại không đề xuất một thành phần con hoặc các thành phần con để kiểm tra các giả thuyết độc lập thành phần con như vậy, các ứng cử viên có thể được tìm thấy bằng cách sử dụng heuristic sau, được lấy cảm hứng từ phân cụm chế độ Q của các phần thành phần (van den Boogaart và Tolosana-Delgado 2013; Filzmoser et al. 2018). Đối với một vector hệ số hồi quy nhất định \( b^\*\_i \) (một độ dốc, một gradient, một hàng hoặc một cột của một ma trận hệ số loại 3 \( B^\* \)), có thể thu được vector hệ số clr bằng \( b\_{clr\_i} = V \cdot b^\*\_i \). Ý tưởng chính là nếu \( b\_{clr\_{ij}} - b\_{clr\_{ik}} = 0 \) thì log(X\_j / X\_k) không ảnh hưởng đến \( Y\_i \), hoặc \( X\_i \) không ảnh hưởng đến log(Y\_j / Y\_k). Do đó, có thể tính toán ma trận khoảng cách giữa các biến clr \( d^2\_{jk} = (clr\_j(b\_i) - clr\_k(b\_i))^2 \), và áp dụng bất kỳ kỹ thuật phân cụm phân cấp nào, tự nhiên định nghĩa một cơ sở ilr cô lập các thành phần con mà các hệ số clr của chúng tương tự nhất, và do đó có khả năng cao hơn không bị ảnh hưởng bởi hoặc không ảnh hưởng đến biến đồng biến thứ i.

**3 Classical Least Squares (LS) Estimation (**Ước lượng Bình phương nhỏ nhất cổ điển (LS))

3.2 Ước lượng LS trong hồi quy Loại 1 và 3

Khi biến đáp ứng là thành phần, chúng ta thu được các vector điểm số logratio quan sát được \(y^\*\_1, ..., y^\*\_n\) có kích thước \(D - 1\), và các hệ số hồi quy được thu thập trong ma trận \(B^\*\) kích thước \((D - 1) \times (P + 1)\). Cột đầu tiên của ma trận này đại diện cho vector tọa độ giao điểm \(b^\*\_0\). Các cột còn lại có thể được liên kết với \(P\) biến đồng biến thực giải thích (hồi quy Loại 1) hoặc với thành phần giải thích được biến đổi ilr \(P = Dx - 1\) (hồi quy Loại 3). Các vector phần dư là \(r^\*\_i(B^\*) = y^\*\_i - B^\* x\_i\), với \(i = 1, ..., n\).

Xét các ma trận của các biến giải thích và biến đáp ứng \(X^\* = [x^\*\_1, ..., x^\*\_n]\) và \(Y^\* = [y^\*\_1, ..., y^\*\_n]\) (mỗi hàng là một cá nhân, các thành phần được biến đổi ilr), các ước lượng bình phương nhỏ nhất của các tham số mô hình là

\[

\mathbf{B}^\* = [(X^\*)^t \cdot X^\*]^{-1} \cdot (X^\*)^t \cdot Y^\*

\]

và

\[

\Sigma\_\epsilon = \frac{1}{N - P} \sum\_{i=1}^n r^\*\_i(B^\*)^t \cdot r^\*\_i(B^\*).

\]

Cuối cùng, ma trận hiệp phương sai của \(\mathbf{b}^\*\) có thể được ước lượng là

\[

\Sigma\_b = \Sigma\_\epsilon \otimes [(X^\*)^t \cdot X^\*]^{-1},

\]

trong đó \(\otimes\) là tích Kronecker của hai ma trận, và \(\mathbf{B}^\*\) được xếp thành vector theo cột.

**3.3 Kiểm Định LS về Sự Độc Lập Thành Phần Con**

Lý thuyết cổ điển về mô hình hồi quy tuyến tính cung cấp một loạt các kiểm định đối với các tham số hồi quy, cả trong hồi quy đơn biến (Loại 2) và hồi quy đa biến (Loại 1 và 3) (Johnson và Wichern 2007). Trong số đó, chúng tôi đặc biệt quan tâm đến những kiểm định có khả năng xử lý sự độc lập thành phần con (ở cả hai hình thức nội bộ và ngoại bộ, tương ứng) như đã giới thiệu trong Mục 2.3.

Đối với mô hình Loại 2 và sự độc lập thành phần con nội bộ, giả thuyết tương ứng về các tham số hồi quy có thể được biểu diễn như \( Ab^\* = 0 \) với \( A = (0,I) \), trong đó \( I \) là ma trận đơn vị kích thước \( (s - 1) \times (s - 1) \) và \( 0 \) là ma trận \( (s - 1) \times (D - s + 1) \) với tất cả các phần tử bằng không. Trong giả thuyết thay thế, đẳng thức trước đó không đúng. Lưu ý rằng đối với trường hợp sự độc lập thành phần con ngoại bộ, kích thước của các ma trận \( I \) và \( 0 \) sẽ thay đổi thành \( s \times s \) và \( s \times (D - s) \), tương ứng. Trong phần tiếp theo, chỉ sự độc lập thành phần con nội bộ sẽ được xem xét, trường hợp của sự độc lập ngoại bộ có thể được suy ra tương tự.

Dưới các giả định của mô hình bao gồm tính bình thường trên giản đơn và giả thuyết không nêu trên, thống kê kiểm định

\[ T = \frac{(SR - S)/(s - 1)}{S/(n - D)} , (13) \]

trong đó

\[ S = \sum\_{i=1}^{n} r\_i(b^\*), \quad SR = \sum\_{i=1}^{n} r\_i(b^\*\_R), \]

tuân theo phân phối F với bậc tự do \( s-1 \) và \( n-D \). Ở đây, \( b^\*\_R \) biểu diễn các ước lượng LS dưới giả thuyết không (tức là chỉ mô hình con được sử dụng để ước lượng các tham số hồi quy). Giả thuyết về sự độc lập thành phần con nội bộ bị bác bỏ nếu \( t \geq F\_{s-1,n-D}(1 - \alpha) \), tức là phân vị \( (1 - \alpha) \) của phân phối đó. Thống kê kiểm định này trùng khớp với kiểm định tỷ lệ khả năng trên cùng một giả thuyết, có thể dễ dàng tổng quát hóa cho trường hợp hồi quy đa biến. Thống kê cũng có thể được viết dưới dạng

\[ T = \frac{S^\*\_R - S^\*}{s - 1}, \]

với

\[ S^\* = \sum\_{i=1}^{n} \left( \frac{r\_i(b^\*)}{\hat{\sigma}\_\epsilon} \right), \quad S^\*\_R = \sum\_{i=1}^{n} \left( \frac{r\_i(b^\*\_R)}{\hat{\sigma}\_\epsilon} \right). (14) \]

Cuối cùng, lưu ý rằng thường sử dụng thực tế là phân phối của \( (s -1)T \) hội tụ theo luật đến phân phối \( \chi^2 \) với bậc tự do \( s - 1 \) khi \( n \rightarrow \infty \).

Tương tự, có thể quan tâm xem một số biến giải thích (không thuộc thành phần) có ảnh hưởng đến biến đáp ứng thành phần (Loại 1); trong trường hợp mô hình Loại 3, chúng ta có thể xem xét vấn đề theo khía cạnh độc lập thành phần con một lần nữa. Kiểm định có thể về sự độc lập thành phần con trong đáp ứng thành phần nên được thực hiện trực tiếp với thành phần con liên quan. Đối với ma trận các hệ số hồi quy \( (D - 1) \times (P + 1) \) \( B^\* \), giả thuyết không hiện có thể được biểu diễn như \( AB^\* = 0 \) với giả thuyết thay thế rằng đẳng thức này không đúng. Ma trận \( A \) có cấu trúc giống như trước, chỉ cần thích ứng với ký hiệu mới, do đó có \( s - 1 \) (sự độc lập thành phần con nội bộ) hoặc \( s \) (sự độc lập thành phần con ngoại bộ) cột, tương ứng, và \( D - 1 \) hàng. Chiến lược thông thường là sử dụng kiểm định tỷ lệ khả năng với thống kê (đối với trường hợp sự độc lập thành phần con nội bộ)

\[ \Lambda = \left( \frac{\det(\Sigma\_b)}{\det(\Sigma\_{bR})} \right)^{n/2}, \]

sử dụng dưới dạng biến đổi

\[ TM = \left( n - P - 1 - \frac{1}{2}(D - (s - 1)) \right) \ln \left( \frac{\det(\Sigma\_{bR})}{\det(\Sigma\_b)} \right), (15) \]

trong đó \( \Sigma\_{bR} \) biểu thị ma trận hiệp phương sai ước lượng của ma trận tham số hồi quy ước lượng trong mô hình con, được hình thành dưới giả thuyết không. Khi \( n \rightarrow \infty \) thống kê \( TM \) hội tụ đến phân phối \( \chi^2 \) với bậc tự do \( (D - 1)(s - 1) \) (Johnson và Wichern 2007).

**4 Robust MM Estimation**

Có nhiều đề xuất cho hồi quy bền vững có sẵn trong tài liệu (xem Maronna et al. 2006). Việc lựa chọn một ước lượng phù hợp phụ thuộc vào các tiêu chí khác nhau. Trước hết, ước lượng nên có các tính chất bền vững mong muốn, tức là bền vững trước mức độ ô nhiễm cao, đồng thời có hiệu quả thống kê cao. Các ước lượng MM cho hồi quy có điểm gãy tối đa là 50% (tức là ít nhất 50% mẫu bị ô nhiễm là cần thiết để làm cho ước lượng trở nên vô dụng), và chúng có thể điều chỉnh hiệu quả. Mặc dù các ước lượng hồi quy khác cũng đạt được điểm gãy cao, như ước lượng hồi quy LTS, hiệu quả của chúng có thể khá thấp (Maronna et al. 2006). Một tiêu chí khác cho việc lựa chọn là sự sẵn có của một triển khai phù hợp trong các gói phần mềm. Các ước lượng MM cho hồi quy có sẵn trong môi trường phần mềm R (R Development Core Team 2019). Đối với đáp ứng đơn biến (hồi quy Loại 2), chúng tôi tham khảo hàm lmrob của gói R robustbase (Maechler et al. 2018), đối với đáp ứng đa biến (Loại 1 và 3), có một triển khai trong gói FRB, cũng cung cấp các thống kê suy luận sử dụng bootstrap bền vững nhanh (Van Aelst và Willems 2013).

**4.2 Ước Lượng MM trong Hồi Quy Loại 2**

**4.2 Ước Lượng MM trong Hồi Quy Loại 1 và 3**

Đối với các trường hợp hồi quy Loại 1 và 3, các ước lượng hồi quy MM đa biến có thể được sử dụng như là các đối trọng mạnh mẽ của các ước lượng LS. Với ký hiệu như trong Mục 3.2, các ước lượng MM cho dữ liệu thành phần được định nghĩa như sau:

\[

(B^\*, C) = \argmin\_{B} \sum\_{i=1}^{n} \rho \left( \frac{r^\*\_i(B)^t C^{-1} r^\*\_i(B)}{\hat{\sigma}} \right), (20)

\]

với ước lượng độ lệch chuẩn \(\hat{\sigma} := \det(\Sigma\_S)^{1/(2D-2)}\), trong đó \(\Sigma\_S\) được lấy từ một ước lượng S của hồi quy đa biến (xem Van Aelst và Willems 2013, để biết chi tiết). Ma trận hiệp phương sai của phần dư ước lượng sau đó được cho bởi \(\Sigma\_\epsilon = \hat{\sigma}^2 C\).

**4.3 Kiểm Định MM về Sự Độc Lập Thành Phần Con**

Các kiểm định giả thuyết mạnh mẽ trong hồi quy tuyến tính không đơn giản, bởi vì chúng phải liên quan đến phần dư mạnh mẽ, và một số kiểm định cũng dựa trên ước lượng mạnh mẽ của ma trận hiệp phương sai của các hệ số hồi quy. Trong phần sau, chúng tôi sẽ tập trung vào các kiểm định có thể xử lý sự độc lập thành phần con.

Đối với trường hợp đơn biến (Loại 2), một kiểm định mạnh mẽ tương đương với kiểm định đã đề cập trong Mục 3.3 có sẵn. Đây là một kiểm định loại tỷ lệ khả năng mà, không giống như kiểm định loại Wald, không yêu cầu ước lượng ma trận hiệp phương sai của \( b^\* \). Giả thuyết cần kiểm định giống như đã nêu trong Mục 3.3, cụ thể là \( Ab^\* = 0 \), với \( A = (0,I) \) và \( I \) là ma trận đơn vị kích thước \( s - 1 \). Đối với giả thuyết thay thế \( Ab^\* \neq 0 \). Tương tự như các thuật ngữ trong (14), kiểm định dựa trên

\[ S^\* = \sum\_{i=1}^{n} \rho \left( \frac{r\_i(b^\*)}{\hat{\sigma}\_\epsilon} \right), \quad S^\*\_R = \sum\_{i=1}^{n} \rho \left( \frac{r\_i(b^\*\_R)}{\hat{\sigma}\_\epsilon} \right), (21) \]

trong đó \( \rho(\cdot) \) là một hàm có giới hạn và \( \hat{\sigma}\_\epsilon \) là ước lượng mạnh mẽ của độ lệch chuẩn của phần dư, xem thêm Eq. (16). Với lựa chọn

\[ \xi = \frac{\sum\_{i=1}^{n} \psi \left( \frac{r\_i(b^\*)}{\hat{\sigma}\_\epsilon} \right)}{\sum\_{i=1}^{n} \psi^2 \left( \frac{r\_i(b^\*\_R)}{\hat{\sigma}\_\epsilon} \right)}, \]

trong đó \( \psi = \rho' \), thống kê kiểm định

\[ T = \xi(S^\*\_R - S^\*) (22) \]

xấp xỉ một phân phối \( \chi^2 \) với \( s - 1 \) bậc tự do, \( \chi^2\_{s-1} \) (xem Hampel et al. 1986). Giả thuyết không bị bác bỏ ở mức ý nghĩa \( \alpha \) nếu giá trị của thống kê kiểm định \( t > \chi^2\_{s-1}(1 - \alpha) \).

Đối với hồi quy Loại 1 và 3, chúng ta có thể sử dụng kiểm định tương đương mạnh mẽ của kiểm định tỷ lệ khả năng đã đề cập trong Mục 3.3. Theo Eq. (15), ma trận hiệp phương sai của ma trận tham số hồi quy ước lượng là cần thiết. Điều này có thể đạt được bằng phương pháp bootstrap như sau. Trong gói R FRB của họ, Van Aelst và Willems (2013) cung cấp chức năng cho các thống kê suy luận trong hồi quy MM đa biến bằng cách sử dụng ý tưởng bootstrap nhanh và mạnh mẽ (Salibian-Barrera và Zamar 2002). Một quy trình bootstrap thông thường sẽ không phù hợp với các ước lượng mạnh mẽ, vì có thể xảy ra rằng một tập dữ liệu bootstrap chứa nhiều điểm ngoại lai hơn so với tập gốc do sự đại diện quá mức của các quan sát ngoại lai, dẫn đến sự sụp đổ của ước lượng. Hơn nữa, việc tính toán lại các ước lượng mạnh mẽ cho mỗi mẫu sẽ rất tốn thời gian. Ý tưởng của bootstrap nhanh và mạnh mẽ (FRB) là ước lượng các tham số chỉ cho dữ liệu gốc. Giả sử \( \theta \) chứa tất cả các ước lượng \( B \) và \( \Sigma\_\epsilon \) dưới dạng vector, và kí hiệu \( \Omega\_\Theta \) là tập hợp các giá trị có thể của tham số mô hình dưới dạng vector này. Các ước lượng MM có thể được viết dưới dạng một hệ phương trình điểm cố định, tức là nhờ một hàm \( g: \Omega\_\Theta \rightarrow \Omega\_\Theta \) sao cho \( \theta = g(\theta) \). Thực tế, nếu hàm \( g \) được biết, người ta có thể ước lượng \( \theta \) như là điểm cố định của phương trình. Hàm \( g \) phụ thuộc vào mẫu, do đó đối với một mẫu bootstrap, chúng ta có được một hàm khác \( g\_b \). Ý tưởng là sử dụng ước lượng gốc và phương trình điểm cố định cho mẫu bootstrap, thu được \( \theta\_{1b} := g\_b(\theta) \). Điều này dẫn đến một ước lượng xấp xỉ của các ước lượng bootstrap \( \theta\_b \) sẽ được thu được trực tiếp từ mẫu bootstrap, tức là giải \( \theta\_b = g\_b(\theta\_b) \). Áp dụng khai triển Taylor, một ước lượng cải thiện \( \theta\_{Ib} \) có thể được suy ra, ước lượng cùng phân phối giới hạn như \( \theta\_b \), và nhất quán với \( \theta \). Để biết thêm chi tiết về bootstrap nhanh và mạnh mẽ cho ước lượng MM trong hồi quy, xem Salibian-Barrera et al. (2008).

**5. Nghiên cứu Trường hợp: Bộ Dữ liệu GEMAS**

**5.1 Thông tin Chung**

Chiến dịch khảo sát địa hóa học đất GEMAS ("Geochemical Mapping of Agricultural and grazing land Soil") được thực hiện ở cấp độ châu Âu, do EuroGeoSurveys, hiệp hội của các Khảo sát Địa chất châu Âu, điều phối (Reimann et al. 2014a, b). Nó bao gồm 33 quốc gia và tập trung vào các mục đích sử dụng đất quan trọng cho sản xuất thực phẩm. Khu vực này được lấy mẫu với mật độ khoảng 1 điểm trên 2.500 km². Các mẫu được lấy từ đất nông nghiệp (0 đến 20 cm) và đất chăn thả (0 đến 10 cm). Tại mỗi điểm, 5 mẫu tại các góc và trung tâm của một hình vuông 10 x 10 m được thu thập và mẫu hợp nhất được phân tích. Khoảng 60 nguyên tố hóa học được thu thập trong các mẫu của cả hai loại đất. Thành phần kết cấu đất cũng được phân tích, tức là tỷ lệ trọng lượng (%) của cát, bùn và sét. Một số thông số mô tả khí hậu (loại khí hậu, nhiệt độ trung bình hoặc lượng mưa trung bình hàng năm) và nền địa chất (loại đá) cũng có sẵn. Cụ thể hơn, lượng mưa trung bình hàng năm và nhiệt độ trung bình hàng năm tại các vị trí lấy mẫu được lấy từ Reimann et al. (2014a) và có nguồn gốc từ www.worldclim.org. Sự phân chia khu vực dự án GEMAS thành các vùng khí hậu bắt nguồn từ Baritz et al. (2005).

Từ các biến số có sẵn, chúng tôi tập trung vào ảnh hưởng giữa thành phần đất (dù là hóa học của nó hay kết cấu cát-bùn-sét) và các biến đồng biến: lượng mưa trung bình hàng năm, pH đất (cả hai là biến liên tục) và các vùng khí hậu (là biến danh mục, với kích thước mẫu tương ứng; các danh mục là Địa Trung Hải (Medi, 438), Ôn đới (Temp, 1,102), Ôn đới–Boreal (BoTe, 352) và Siêu Boreal (Spbo, 203)). Hình 1 hiển thị một tập hợp các biểu đồ mô tả của các biến số và thành phần này. Tổng cộng có n = 2095 mẫu dữ liệu GEMAS được sử dụng, bao phủ gần như toàn bộ châu Âu, ngoại trừ Romania, Moldova, Belarus, miền Đông Ukraine và Nga (Hình 2). Từ so sánh giữa các panel A và B (Hình 1), có thể kết luận rằng cần lấy logarit của Lượng Mưa Hàng Năm để xử lý tiếp. Mặc dù tính đối xứng hoặc chuẩn không đạt được, thậm chí với logarit (cả hai giá trị p của kiểm định Anderson-Darling cho tính chuẩn (Anderson và Darling 1952) đều bằng không), ít nhất là cái nhìn theo bốn nhóm khí hậu gợi ý rằng các sai lệch từ đối xứng là từ trung bình đến nhẹ (Hình 1c), không ảnh hưởng tiêu cực đến kết quả hồi quy tiếp theo.

Như đã chỉ ra ở trên, dữ liệu này có thiết kế khá không cân đối với các khu vực khí hậu (Hình 1d), đặc biệt do sự thống trị của khí hậu ôn đới, chiếm hơn 50% mẫu, xem thêm Hình 2.

**\*\*Hình 1\*\*** Biểu đồ mô tả của các tập hợp biến số được sử dụng. Hiển thị các biểu đồ tần số của Lượng Mưa Hàng Năm trên thang đo ban đầu (a) và logarit (b). Hộp đồ thị của Lượng Mưa Hàng Năm trên thang đo logarit (c) và biểu đồ tần số kích thước mẫu (d) theo nhóm khí hậu. Các nhóm khí hậu được sử dụng để tô màu các thành phần cát-bùn-sét trong biểu đồ tam giác (e). Cuối cùng, cấu trúc dữ liệu đa biến của các thành phần hóa học được thể hiện bằng biểu đồ biplot logratio tập trung (f).

**\*\*Hình 2\*\*** Vị trí mẫu. Màu sắc theo vùng khí hậu: đỏ cho Địa Trung Hải, xanh cho Ôn đới, xanh lá cho Ôn đới–Boreal và tím cho Siêu Boreal.

Thành phần kết cấu cát-bùn-sét được thể hiện trong Hình 1e như một biểu đồ tam giác, với các màu theo bốn vùng khí hậu: chúng cho thấy sự kiểm soát nhất định về lượng sét, điều này sẽ được khám phá sau. Về thành phần oxit chính bao gồm SO3 và LOI (mất khi đốt cháy), điều này được thể hiện trong Hình 1f như một biểu đồ biplot hiệp phương sai logratio tập trung, như thường lệ trong phân tích dữ liệu thành phần (Aitchison 1997; Aitchison và Greenacre 2002). Điều này cho thấy một tập dữ liệu khá đồng nhất mà không có bất kỳ nhóm mạnh nào có thể ảnh hưởng tiêu cực đến chất lượng của các bước hồi quy tiếp theo.

**### 5.2 Thành Phần Kích Thước Hạt So Với Lượng Mưa (Hồi Quy Loại 1)**

Nhiệm vụ đầu tiên là biểu diễn thành phần cát-bùn-sét (biến phụ thuộc) như một hàm số của lượng mưa, trên thang logarit (biến giải thích), sử dụng mô hình hồi quy từ mục 2.2.1. Hình 3 hiển thị sự phụ thuộc này bằng cách vẽ mỗi tỷ số logarit có thể có trên trục tung so với logarit của lượng mưa hàng năm trên trục hoành. Dữ liệu có sẵn được biến đổi theo ma trận sau đây hoặc các tọa độ ilr này:

\[

V\_t = \begin{bmatrix}

-\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\

-\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0

\end{bmatrix}

\]

\[

y^\*\_1 = \frac{1}{\sqrt{6}} \ln \frac{y\_{clay}}{y\_{silt} \cdot y\_{sand}}, \quad y^\*\_2 = \frac{1}{\sqrt{2}} \ln \frac{y\_{silt}}{y\_{sand}}

\]

Một mô hình dạng Eq. (9) được ước lượng bằng phương pháp LS. Bảng 1 hiển thị các hệ số logarit, cũng như giá trị của chúng sau khi được chuyển đổi ngược lại. Lưu ý rằng các giá trị chuyển đổi ngược lại sẽ giống hệt nhau, bất kể phương pháp logarit nào được sử dụng cho các tính toán.

\*\*Hình 3\*\* Thành phần kích thước hạt như một hàm của (log) lượng mưa hàng năm; dữ liệu GEMAS quan sát được (chấm) và các mô hình khớp (đỏ: cổ điển; xanh: bền vững). Kích thước biểu tượng (trong các bảng bên dưới bên trái) tỷ lệ nghịch với trọng số được tính trong hồi quy bền vững.

\*\*Bảng 1\*\* Hồi quy thành phần kích thước hạt so với (log) lượng mưa hàng năm, sử dụng hồi quy LS và MM

| | \(y^\*\_1\) | \(y^\*\_2\) | p-value \(y^\*\_1\) | p-value \(y^\*\_2\) | Sand | Silt | Clay |

|----------------|------------|----------|-------------------|-------------------|------|------|------|

| LS: Intercept | 0.931 | -0.402 | 0.023 | 0.234 | 0.255| 0.144| 0.600|

| LS: log(AnnPrec)| -0.298 | 0.019 | 0.000 | 0.718 | 0.366| 0.376| 0.258|

| MM: Intercept | 1.813 | 0.656 | 0.000 | 0.014 | 0.055| 0.139| 0.806|

| MM: log(AnnPrec)| -0.405 | -0.125 | 0.000 | 0.004 | 0.418| 0.350| 0.233|

Các cột biểu thị các tham số ước lượng cho các tọa độ ilr, các giá trị p tương ứng và các hệ số hồi quy đã chuyển đổi ngược lại.

\*\*Hình 4\*\* Biểu đồ tán xạ (trái: mặt phẳng ilr; phải: biểu đồ tam giác) của dữ liệu và các dự đoán của cả hai mô hình LS (đỏ) và MM (xanh). Các đường kẻ biểu thị các mô hình ngoại suy ra ngoài phạm vi lượng mưa hàng năm quan sát được. Kích thước biểu tượng tỷ lệ nghịch với trọng số được tính bởi hồi quy MM bền vững.

Bảng 1 liệt kê các hệ số cho các tọa độ ilr được định nghĩa trong Eq. (23), các giá trị p tương ứng và các hệ số đã chuyển đổi ngược lại, sử dụng các ước lượng LS và MM. Các ước lượng LS cho thấy tỷ lệ bùn-so với-cát không bị ảnh hưởng bởi lượng mưa hàng năm, trong khi mối quan hệ của chúng với sét phụ thuộc vào biến đồng biến này. Ngược lại, với các ước lượng MM, cả hai tọa độ đều bị ảnh hưởng bởi lượng mưa hàng năm. Hình 3 hiển thị cả hai mô hình LS và MM, được diễn giải lại trong mỗi cặp tỷ số logarit có thể. Lưu ý rằng độ dốc và giao điểm cho tọa độ \(y^\*\_2\) trong Bảng 1 tương ứng với bảng (2,1) của Hình 3 này. Các giao điểm và độ dốc cho mỗi bảng khác có thể được tính bằng cách chuyển đổi các hệ số (ysand, ysilt, yclay) tương ứng.

Hình 4 cho thấy các dự đoán mô hình cho mô hình cổ điển (đỏ) và mô hình bền vững (xanh). Biểu đồ bên trái trình bày các dự đoán trong tọa độ ilr, như chúng được sử dụng trong các mô hình hồi quy, và biểu đồ bên phải hiển thị các dự đoán cho thành phần ban đầu. Kích thước biểu tượng tỷ lệ nghịch với trọng số từ hồi quy MM bền vững, và tại đây rõ ràng rằng do các giá trị rất nhỏ của sét (các giá trị được làm tròn), các hiện tượng bất thường trong dữ liệu được tạo ra trong các tọa độ ilr, nhưng những quan sát này bị giảm trọng lượng bởi hồi quy MM. Đây là lý do chính cho sự khác biệt giữa mô hình LS và MM.

**### 5.3 Thành Phần Kích Thước Hạt So Với Khí Hậu (Hồi Quy Loại 1)**

Hồi quy thành phần kích thước hạt (biến phụ thuộc) với các vùng khí hậu (biến giải thích) cần xem xét rằng các vùng khí hậu được sắp xếp theo trình tự rõ ràng từ Địa Trung Hải (Medi), Ôn Đới (Temp), Ôn Đới–Boreal (BoTe) đến Cận Boreal (Spbo), từ Nam đến Bắc. Điều này rõ ràng thấy ở Hình 5, cho thấy tỷ lệ cát/bùn trung bình tương đối ổn định giữa các vùng khí hậu, nhưng có một xu hướng đơn điệu rõ ràng của tỷ lệ cát/sét và bùn/sét về phía Bắc. Xu hướng này cũng được thể hiện qua các trung tâm thành phần (Pawlowsky-Glahn et al. 2015) cho các danh mục khí hậu tương ứng, xem Bảng 2.

Do đó, giả thuyết không tương quan sau đây có vẻ hợp lý:

1. Cân bằng giữa cát và bùn không tương quan với khí hậu (tức là thành phần phụ cát-bùn không tương quan nội bộ với khí hậu).

2. Cân bằng giữa sét và hai loại khác phụ thuộc vào khí hậu chỉ theo các điều kiện tuyến tính, như được giải thích trong đoạn sau.

Với các giả thuyết này, các tọa độ ilr tương tự như trong mục trước (Eq. 23) sẽ được sử dụng ở đây.

Trong R—gói stats; (R Development Core Team 2019)—, một mô hình hồi quy với một yếu tố có thứ tự gồm 4 mức độ yêu cầu xây dựng một ma trận thiết kế phụ trợ (n×3), trong đó mỗi hàng được lấy làm hàng tương ứng của Bảng 3. Các nhãn L—“Tuyến tính”, Q—“Bậc hai” và C—“Bậc ba” biểu thị loại xu hướng giữa bốn danh mục khớp với dữ liệu, L ngụ ý rằng sự khác biệt giữa hai danh mục liên tiếp là không đổi (Simonoff 2003).

Bảng 4 tóm tắt đầu ra số của mô hình hồi quy này, bao gồm các hệ số ước lượng (giao điểm và các hiệu ứng L, Q và C) cho mỗi một trong hai cân bằng, các giá trị p của các giả thuyết hệ số không, và các hệ số đã chuyển đổi ngược lại. Các kết quả này được cung cấp cho cả hồi quy cổ điển (LS) và hồi quy bền vững (MM). Hồi quy LS cổ điển cho thấy rằng các hiệu ứng C và Q có thể bị loại bỏ đối với \(y^\*\_2\) nhưng không phải là hiệu ứng L, tức là giả thuyết đầu tiên (không tương quan nội bộ của thành phần phụ cát-bùn với khí hậu) phải bị từ chối. Đối với giả thuyết thứ hai, việc làm vô hiệu các hệ số cho các hiệu ứng L và Q trên \(y^\*\_1\) khác biệt đáng kể so với không (giá trị p nhỏ hơn mức quan trọng 0.05), điều này ngụ ý rằng giả thuyết thứ hai cũng sai. Tuy nhiên, các hiệu ứng C có thể bị loại bỏ. Một kiểm tra toàn cầu theo phong cách được giải thích trong Mục 3.3 cho giá trị p bằng không cho giả thuyết không có hiệu ứng Q hoặc C, do đó hỗ trợ cho các kết luận này. Hồi quy bền vững đưa ra một bức tranh tương tự, ngoại trừ rằng ở đây tất cả các hiệu ứng đều có ý nghĩa đối với \(y\_2\).

\*\*Bảng 3\*\* Các vectơ hàng để xây dựng ma trận thiết kế liên quan đến biến danh mục khí hậu

| L | Q | C |

|--------|--------|--------|

| Medi | -0.6708 | 0.5000 | -0.2236 |

| Temp | -0.2236 | -0.5000 | 0.6708 |

| BoTe | 0.2236 | -0.5000 | -0.6708 |

| SpBo | 0.6708 | 0.5000 | 0.2236 |

Các con số này là kết quả của việc áp dụng hàm contrasts của R trên biến yếu tố có thứ tự khí hậu. L là viết tắt của hiệu ứng tuyến tính, Q cho hiệu ứng bậc hai và C cho hiệu ứng bậc ba.

\*\*Bảng 4\*\* Hệ số khớp và giá trị p của các mô hình hồi quy thành phần kích thước hạt so với khí hậu

| | \(y^\*\_1\) | \(y^\*\_2\) | p-value \(y^\*\_1\) | p-value \(y^\*\_2\) | Sand | Silt | Clay |

|---------------|-----------|-----------|-------------------|-------------------|-------|-------|-------|

| LS: intercept | -1.238 | -0.308 | 0.000 | 0.000 | 0.548 | 0.355 | 0.097 |

| LS: L | -1.195 | -0.142 | 0.000 | 0.001 | 0.493 | 0.403 | 0.103 |

| LS: Q | -0.120 | -0.019 | 0.002 | 0.594 | 0.354 | 0.344 | 0.302 |

| LS: C | 0.045 | 0.009 | 0.165 | 0.765 | 0.325 | 0.329 | 0.346 |

| MM: intercept | -1.143 | -0.223 | 0.000 | 0.000 | 0.515 | 0.376 | 0.109 |

| MM: L | -1.202 | -0.157 | 0.000 | 0.000 | 0.498 | 0.399 | 0.102 |

| MM: Q | -0.270 | -0.085 | 0.000 | 0.006 | 0.390 | 0.346 | 0.264 |

| MM: C | 0.017 | 0.054 | 0.728 | 0.022 | 0.319 | 0.344 | 0.338 |

Tất nhiên, các tương phản khác có thể được sử dụng cho phân tích này, tùy thuộc vào bản chất của các giả thuyết phụ thuộc mà chúng ta quan tâm đến việc kiểm tra. Nếu, ví dụ, ai đó muốn kiểm tra liệu đất từ các vùng khí hậu khác nhau có cùng kết cấu đất trung bình hay không, người đó có thể đã sử dụng hàm constr.treatment của R để buộc loại so sánh này.

Một cách hoặc cách khác, trong một mô hình hồi quy danh mục như thế này, giao điểm có thể được diễn giải như một loại giá trị trung bình toàn cầu bù đắp cho sự thiếu cân bằng giữa bốn danh mục. Trong khi trung tâm thành phần thông thường là [cát; bùn; sét] = [52.39%; 35.27%; 12.34%], hồi quy bình phương tối thiểu đưa ra một ước lượng [54.83%; 35.48%; 9.69%] và hồi quy bền vững là [51.54%; 37.60%; 10.86%], cả hai đều giảm tầm quan trọng của sét. Lưu ý rằng giao điểm này không phụ thuộc vào bộ tương phản nào được chọn để nắm bắt biến danh mục.

\*\*Hình 6\*\* Hệ số hồi quy bình phương tối thiểu (trái) và MM (phải) của thành phần chính chuyển đổi clr so với logarit lượng mưa hàng năm.

**### 5.4 Thành Phần Oxit Chính So Với Lượng Mưa (Hồi Quy Loại 1)**

Các giả thuyết phong phú hơn có thể được kiểm tra nếu thành phần sử dụng có nhiều phần. Để minh họa điều này, hồi quy của các oxit chính so với (log) lượng mưa hàng năm sẽ được thực hiện. Một câu hỏi ban đầu tự nhiên là liệu hóa học đất có bị ảnh hưởng bởi lượng mưa hay không. Để làm điều này, Hình 6 cho thấy các hệ số clr được ước lượng bằng phương pháp bình phương tối thiểu cổ điển và hồi quy bền vững: các hệ số clr không nên được diễn giải một cách riêng lẻ; đúng hơn, sự khác biệt giữa chúng có thể được hiểu như ảnh hưởng đến một tỷ số log của một cặp cụ thể. Do đó, chúng ta đang tìm kiếm những sự khác biệt nhỏ nhất giữa các hệ số vì chúng xác định các cặp biến mà cân bằng của chúng không bị ảnh hưởng bởi biến giải thích. Khi một tập hợp các tỷ số log của các cặp xác định một (phụ) thành phần, điều này cung cấp thông tin về các phụ thành phần có thể tiềm năng không phụ thuộc nội bộ vào biến đồng biến, chẳng hạn như:

– TiO2–Fe2O3–MnO

– Al2O3-LOI (với Na2O trong hồi quy bình phương tối thiểu, hoặc MgO trong hồi quy bền vững)

– SiO2–K2O

Một tập hợp tọa độ ilr được chọn để chứa các cân bằng giữa các phụ thành phần này. Ma trận dấu hiệu để xây dựng các cân bằng này được cho trong Bảng 5. Hãy nhớ rằng trong bảng dấu hiệu, +1 biểu thị các biến xuất hiện trong tử số của cân bằng, -1 biểu thị các biến trong mẫu số, và 0 biểu thị các biến không tham gia vào cân bằng cụ thể đó. Ví dụ, cân bằng giữa các phụ thành phần TiO2–Fe2O3–MnO và Al2O3–Na2O–LOI là \(y^\*\_4\), và các cân bằng (\(y^\*\_7\), \(y^\*\_8\)) mô tả sự biến đổi nội bộ trong phụ thành phần TiO2–Fe2O3–MnO.

Sử dụng tập hợp các cân bằng này, một mô hình hồi quy với biến giải thích (log) lượng mưa hàng năm được điều chỉnh, với hồi quy LS và MM. Kết quả được báo cáo trong Bảng 6. Chú ý đến các giá trị p của các độ dốc của hai mô hình, chúng ta kết luận rằng phụ thành phần Al2O3–Na2O–LOI (\(y^\*\_7\), \(y^\*\_8\)) là không phụ thuộc nội bộ vào lượng mưa hàng năm (cả hai phương pháp cổ điển và bền vững đều đồng ý về điều này). Nói một cách lỏng lẻo, điều tương tự áp dụng cho các cân bằng SiO2/K2O (\(y^\*\_{10}\)) và MgO so với tất cả các thành phần khác (\(y^\*\_1\)). Cuối cùng, cân bằng TiO2/Fe2O3 (\(y^\*\_6\)) dường như không tương quan với lượng mưa hàng năm chỉ từ quan điểm của hồi quy bình phương tối thiểu.

\*\*Bảng 5\*\* Bảng dấu hiệu để xây dựng các cân bằng cho thành phần oxit chính trong một bài toán hồi quy so với (log) lượng mưa hàng năm

| | \(y^\*\_1\) | \(y^\*\_2\) | \(y^\*\_3\) | \(y^\*\_4\) | \(y^\*\_5\) | \(y^\*\_6\) | \(y^\*\_7\) | \(y^\*\_8\) | \(y^\*\_9\) | \(y^\*\_{10}\) |

|---------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|--------------|

| SiO2 | +1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | +1 |

| TiO2 | +1 | 0 | +1 | 0 | 0 | +1 | 0 | 0 | 0 | 0 |

| Al2O3 | +1 | 0 | 0 | -1 | 0 | 0 | +1 | 0 | 0 | 0 |

| Fe2O3 | +1 | 0 | 0 | +1 | 0 | -1 | 0 | 0 | 0 | 0 |

| MnO | +1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| MgO | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| CaO | +1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| Na2O | +1 | 0 | 0 | -1 | 0 | 0 | 0 | 0 | 0 | 0 |

| K2O | +1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 |

| P2O5 | +1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| LOI | +1 | 0 | 0 | -1 | 0 | 0 | 0 | 0 | 0 | 0 |

\*\*Bảng 6\*\* Các hệ số ước lượng và giá trị p cho giao điểm (int) và độ dốc (slp), đối với hồi quy bình phương tối thiểu (LS) và hồi quy bền vững (MM)

| | LS.int | MM.int | LS.slp | MM.slp | LS.int.p | MM.int.p | LS.slp.p | MM.slp.p |

|---------|--------|--------|--------|--------|----------|----------|----------|----------|

| \(y^\*\_1\) | 0.596 | 1.229 | 0.013 | -0.086 | 0.053 | 0.000 | 0.776 | 0.053 |

| \(y^\*\_2\) | -7.722 | -7.469 | 0.822 | 0.789 | 0.000 | 0.000 | 0.000 | 0.000 |

| \(y^\*\_3\) | 3.301 | 3.454 | -0.218 | -0.233 | 0.000 | 0.000 | 0.000 | 0.000 |

| \(y^\*\_4\) | -3.450 | -3.937 | 0.140 | 0.218 | 0.000 | 0.000 | 0.000 | 0.000 |

| \(y^\*\_5\) | 2.876 | 3.192 | -0.082 | -0.130 | 0.000 | 0.000 | 0.003 | 0.000 |

| \(y^\*\_6\) | -1.011 | -0.654 | -0.035 | -0.086 | 0.000 | 0.000 | 0.085 | 0.000 |

| \(y^\*\_7\) | -1.151 | -0.339 | 0.024 | -0.076 | 0.002 | 0.361 | 0.662 | 0.157 |

| \(y^\*\_8\) | 1.752 | 1.176 | -0.004 | 0.074 | 0.000 | 0.000 | 0.911 | 0.089 |

| \(y^\*\_9\) | -1.243 | -0.242 | 0.451 | 0.314 | 0.036 | 0.000 | 0.761 | 0.000 |

| \(y^\*\_{10}\) | 2.825 | 2.791 | -0.044 | -0.042 | 0.000 | 0.000 | 0.111 | 0.073  
  
  
Bây giờ, các bài kiểm tra tổng quát về sự độc lập nội bộ và bên ngoài của Al2O3–Na2O–LOI đối với lượng mưa hàng năm đã được thực hiện theo phương pháp trong Mục 3.3, và đưa ra các giá trị p lần lượt là 0.884 và 0. Các kết quả này có phần mâu thuẫn với hiểu biết chung về quá trình phong hóa là một quá trình làm giàu Al2O3 (và có lẽ là LOI) với chi phí của Na2O (và CaO). Lượng mưa hàng năm, một trong những yếu tố của phong hóa hóa học, không cho thấy bất kỳ ảnh hưởng đáng kể nào đến tỷ lệ log của Al2O3/Na2O. Mặt khác, bài kiểm tra tổng quát bền vững đưa ra các kết quả có ý nghĩa: trong cả hai trường hợp, các giá trị p đều bằng không.

\*\*Hình 7\*\* Độ dốc hồi quy của pH so với thành phần oxit chính, được biểu diễn bằng các hệ số clr: ước lượng bình phương tối thiểu (LS, bên trái) và ước lượng bền vững (MM, bên phải)

**6 Kết luận**

Mục đích của đóng góp này là phác thảo khái niệm phân tích hồi quy cho dữ liệu thành phần, và chỉ ra cách thực hiện phân tích này trong thực tế với dữ liệu thực. Chúng tôi đã phân biệt ba loại mô hình hồi quy: Loại 1, trong đó biến đáp ứng là một thành phần và biến giải thích là biến thực không phải thành phần; Loại 2 với thành phần là biến giải thích và biến đáp ứng là biến thực, và Loại 3, trong đó cả biến đáp ứng và biến giải thích đều là các thành phần. Lưu ý rằng cũng có thể xem xét trường hợp hồi quy trong một thành phần, bằng cách chia các phần thành phần thành một nhóm tạo thành các biến đáp ứng và một nhóm đại diện cho các biến giải thích. Trường hợp này không được đề cập ở đây vì nó đòi hỏi một mô hình gọi là "lỗi trong biến số," xem Hrůzová et al. (2016) để biết chi tiết.

Đối với cả ba loại mô hình, cách xử lý thành phần trong mô hình hồi quy là rất quan trọng. Một cách tiếp cận hợp lý về mặt hình học là sử dụng các tọa độ trực chuẩn, gọi là cân bằng, có thể được xây dựng để giải thích các hệ số hồi quy và kiểm định các giả thuyết khác nhau. Nếu mối quan tâm không phải là suy luận thống kê mà chỉ là độ phù hợp của mô hình và các giá trị ước lượng, bất kỳ tọa độ logratio nào cũng phù hợp để đại diện cho thành phần. Lưu ý rằng phép biến đổi clr sẽ không phù hợp cho các mô hình hồi quy Loại 2 hoặc Loại 3, vì ma trận dữ liệu kết quả là suy biến, dẫn đến vấn đề trong ước lượng tham số khi thành phần đóng vai trò là biến giải thích.

Hồi quy cổ điển theo phương pháp bình phương nhỏ nhất (LS) cũng như hồi quy MM bền vững đã được xem xét để ước lượng các tham số hồi quy và các giá trị p tương ứng cho các kiểm định giả thuyết. Nếu các yêu cầu của mô hình được thỏa mãn, ước lượng hồi quy LS là ước lượng tuyến tính không chệch tốt nhất (BLUE) với các tính chất tối ưu tương ứng (xem, ví dụ, Johnson và Wichern 2007), nhưng trong trường hợp này hồi quy MM cũng dẫn đến một ước lượng có hiệu quả thống kê cao. Tuy nhiên, trong trường hợp vi phạm mô hình, ví dụ, do dữ liệu ngoại lệ, các tính chất tối ưu này không còn hiệu lực. Tuy vậy, ước lượng MM vẫn đáng tin cậy vì nó rất bền vững trước các ngoại lệ, cả trong các biến giải thích và biến đáp ứng. Trong các ứng dụng thực tế, không phải lúc nào cũng rõ ràng liệu có ngoại lệ trong dữ liệu hay không. Trong trường hợp này, có thể khuyến nghị thực hiện cả hai loại phân tích và so sánh kết quả. Đặc biệt, có thể kiểm tra các biểu đồ chẩn đoán từ hồi quy bền vững (như đã làm trong Mục 5.5) để xác định các ngoại lệ tiềm năng có thể đã ảnh hưởng đến ước lượng LS, xem Maronna et al. (2006).

Các loại hồi quy và ước lượng khác nhau đã được áp dụng cho một bộ dữ liệu ví dụ từ dự án GEMAS (Reimann et al. 2014a, b). Tất cả các ví dụ được trình bày chỉ nhằm mục đích minh họa, nhưng chúng chỉ ra cách xây dựng các cân bằng và cách kiểm định giả thuyết. Đối với các ước lượng bền vững, các hàm có sẵn trong các gói R robustbase (Maechler et al. 2018) và FRB (Van Aelst và Willems 2013). Điều quan trọng cần lưu ý là không chỉ các tham số hồi quy được ước lượng một cách bền vững với các gói này, mà việc ước lượng bền vững cũng được thực hiện cho việc ước lượng sai số chuẩn và kiểm định giả thuyết, cho phương sai dư, thước đo R2 bội, v.v. Chúng tôi trình bày các khả năng chẩn đoán hồi quy trong Mục 5.5. Trong hầu hết các ví dụ, so sánh hồi quy LS và MM đã được cung cấp.

Một vấn đề quan trọng trong bối cảnh hồi quy là vấn đề lựa chọn biến, hoặc sự độc lập của các thành phần con. Đặc biệt đối với Loại 2 và 3, nơi các biến giải thích xuất phát từ một thành phần, không dễ dàng để có được "tập hợp con tốt nhất" của các phần thành phần mà không chứa các phần không thông tin và vẫn tạo ra một mô hình có sức mạnh dự đoán tương tự như mô hình đầy đủ. Có các phương pháp có sẵn trong tài liệu để giảm số lượng thành phần, xem, ví dụ, Pawlowsky-Glahn et al. (2011), Hron et al. (2013), Mert et al. (2015) và Greenacre (2019). Tuy nhiên, không có phương pháp nào về sự độc lập của các thành phần con hoạt động tương đương với các phương pháp không phải thành phần, chẳng hạn như lựa chọn biến tiến tới hoặc lùi; chỉ một cái nhìn sơ lược về những phương pháp trong bối cảnh thành phần đã được phác thảo trong Filzmoser et al. (2018). Những phương pháp đó sẽ được xử lý trong nghiên cứu tương lai của chúng tôi.

Lời cảm ơn: Karel Hron và Peter Filzmoser xin chân thành cảm ơn sự hỗ trợ của Quỹ Khoa học Séc GA19-01768S.

Tài trợ: Quỹ truy cập mở do Đại học Khoa học Ứng dụng ZHAW Zurich cung cấp.

Truy cập Mở: Bài viết này được cấp phép theo Giấy phép Creative Commons Attribution 4.0 International, cho phép sử dụng, chia sẻ, điều chỉnh, phân phối và sao chép dưới bất kỳ hình thức hoặc định dạng nào, miễn là bạn ghi công tác giả gốc và nguồn, cung cấp liên kết đến giấy phép Creative Commons, và chỉ ra nếu có sự thay đổi. Các hình ảnh hoặc tài liệu của bên thứ ba khác trong bài viết này được bao gồm theo giấy phép Creative Commons của bài viết, trừ khi có ghi chú khác trong phần tín dụng cho tài liệu. Nếu tài liệu không được bao gồm trong giấy phép Creative Commons của bài viết và việc sử dụng của bạn không được phép theo quy định pháp luật hoặc vượt quá việc sử dụng cho phép, bạn sẽ cần phải xin phép trực tiếp từ bên sở hữu bản quyền.