

LASSO REGRESSION AND AN APPLICATION IN BREAST CANCER DATA ANALYSIS

Nong Quynh Van*, Tran Dinh Hung

TNU - University of Education

ARTICLE INFO	ABSTRACT
Received: 25/4/2022 Revised: 30/5/2022 Published: 31/5/2022	The LASSO is one of the regularized regression methods proposed by Tibshirani in 1996. The goal of LASSO is to select and estimate parameters in a linear regression model by exactly shrinking some coefficients to zero. In particular, the LASSO is useful in analyzing microarray gene data in which the number of predictors (genes) is much larger than the number of sample observations (number of patients). In this paper, we introduce a brief summary of the LASSO and apply this method to study gene in breast cancer data. The aim was to assess the genes interactions associated with breast cancer microarray data. The results show that the LASSO method performs relatively well in analyzing gene expression levels and indicates genes that related to the breast cancer gene BRCA1 such as genes NBR2, AASDH, KIAA2013, VPS25, NBR1, SEC22C, RPL27, CBLN3, KHDRBS1, XRCC2. In fact, the NBR2 gene is adjacent to BRCA1 on chromosome 17, and two genes share the same promoter region. Thus, breast cancer prognosis determined by regression will help us to better understand the mechanism underlying the occurrence of breast cancer of young women.
KEYWORDS Regression Ordinary least square LASSO L1 regularization Penalized regression Breast cancer	

HỒI QUY LASSO VÀ ỨNG DỤNG TRONG PHÂN TÍCH DỮ LIỆU UNG THƯ VÚ

Nông Quỳnh Vân*, Trần Đình Hùng

Trường Đại học Sư phạm - ĐH Thái Nguyên

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 25/4/2022 Ngày hoàn thiện: 30/5/2022 Ngày đăng: 31/5/2022	Hồi quy LASSO là một trong những phương pháp hồi quy phạt được đề xuất bởi Tibshirani vào năm 1996. Mục tiêu của LASSO là lựa chọn và ước lượng tham số trong mô hình hồi quy tuyến tính bằng cách hiệu chỉnh một số hệ số bằng 0. Đặc biệt, LASSO rất hữu ích trong việc phân tích dữ liệu gen, trong đó số lượng yếu tố dự báo (gen) lớn hơn nhiều so với số lượng quan sát mẫu (số bệnh nhân). Trong bài báo này, chúng tôi sẽ hệ thống lại các kiến thức cơ bản về hồi quy LASSO và áp dụng phương pháp LASSO cho nghiên cứu gen ở bệnh nhân ung thư vú. Mục tiêu của chúng tôi là xác định gen nào ảnh hưởng đến mức độ nghiêm trọng của ung thư vú theo dữ liệu microarray. Kết quả cho thấy, LASSO hoạt động tương đối tốt trong phân tích mức độ biểu hiện gen và chỉ ra được những gen có liên quan tới gen gây ung thư vú BRCA1 là các gen NBR2, AASDH, KIAA2013, VPS25, NBR1, SEC22C, RPL27, CBLN3, KHDRBS1, XRCC2. Trên thực tế, gen NBR2 tiếp giáp với BRCA1 trên nhiễm sắc thể 17 và hai gen này có chung một vùng gen khởi động. Như vậy, tiên lượng ung thư vú xác định bằng hồi quy sẽ giúp chúng ta hiểu rõ hơn về cơ chế tiềm ẩn sự xuất hiện của ung thư vú đối với phụ nữ còn trẻ.
TỪ KHÓA Hồi quy Bình phương tối thiểu LASSO Chính quy hóa L1 Hồi quy phạt Ung thư vú	

DOI: <https://doi.org/10.34238/tnu-jst.5901>

* Corresponding author. Email: vannq@tnue.edu.vn

1. Giới thiệu

Các phương pháp phân tích dữ liệu trong Khoa học thống kê giúp chúng ta mô tả và khám phá mối liên quan bí ẩn trong tập dữ liệu. Phân tích hồi quy tuyến tính có lẽ là một trong những phương pháp phân tích dữ liệu thông dụng nhất và giúp chúng ta có nhiều phát hiện quan trọng như khám phá gen, đánh giá hiệu quả của thuốc, xây dựng mô hình dự báo... Kỹ thuật cổ điển nổi tiếng nhất để xây dựng mô hình hồi quy tuyến tính đó chính là phương pháp bình phương bé nhất được Carl F Gauss phát triển vào thế kỷ 18. Tuy nhiên, các ước lượng tham số tìm được từ phương pháp bình phương bé nhất thường có độ chệch lớn và phương sai nhỏ. Điều này ảnh hưởng tới độ chính xác của kết quả dự đoán từ mô hình hồi quy. Đồng thời trong trường hợp số quan sát nhỏ hơn số biến thì việc áp dụng phương pháp bình phương bé nhất không khả thi do tích ma trận quan sát không khả nghịch. Từ đó các phương pháp hiệu chỉnh (hay phương pháp phạt) được đề xuất, xem [1]-[8]. Một trong các phương pháp hiệu chỉnh phổ biến và nổi tiếng nhất hiện nay là phương pháp LASSO (viết tắt của Least absolute and selection operator, hiệu chỉnh trị tuyệt đối bé nhất và lựa chọn biến) [9].

LASSO được đề xuất đầu tiên vào năm 1996 bởi Robert Tibshirani, một nhà thống kê tại Đại học Stanford. Như tên gọi của mình, LASSO khắc phục được các nhược điểm của phương pháp bình phương bé nhất và đồng thời thực hiện "lựa chọn biến" trong mô hình. Đó là, "hiệu chỉnh" một số hệ số ước lượng của biến dự báo bằng 0 và chỉ ra các biến quan trọng nhất của mô hình. Với ưu điểm như vậy, hồi quy LASSO được ứng dụng rộng rãi trong các lĩnh vực y học, kinh tế, công nghiệp, nông nghiệp... và các nghiên cứu khoa học liên ngành [10]. Bài báo này sẽ hệ thống lại một cách khoa học các kiến thức về mô hình hồi quy LASSO và ứng dụng LASSO trong phân tích dữ liệu ung thư vú.

2. Hồi quy LASSO

LASSO hoặc "hiệu chỉnh trị tuyệt đối bé nhất và lựa chọn biến" là một phương pháp để ước lượng các tham số của mô hình hồi quy tuyến tính được đề xuất bởi Tibshirani vào năm 1996. Mục tiêu của LASSO là cực tiểu tổng bình phương các sai số với ràng buộc là tổng trị tuyệt đối của các tham số ước lượng trong mô hình nhỏ hơn một hằng số.

2.1. Định nghĩa

Trong mô hình hồi quy tuyến tính đa biến, ảnh hưởng của các biến dự báo lên biến tiên lượng được mô hình hóa bằng phương trình sau:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i \quad (1)$$

Trong đó, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ là các biến ngẫu nhiên độc lập, cùng phân phối xác suất với kỳ vọng bằng 0 và phương sai bằng hằng số, $y_i (i=1, \dots, n)$ là giá trị của biến tiên lượng, $x_{ij} (j=1, \dots, p)$ là p biến dự báo và $\beta_0, \beta_1, \dots, \beta_p$ là các hệ số (hay các tham số) của mô hình hồi quy.

Bằng cách bỏ qua hệ số chặn β_0 , các hệ số hồi quy LASSO được định nghĩa là nghiệm của bài toán tối ưu có ràng buộc sau:

$$\hat{\beta}^{LASSO} = \arg \min_{\beta} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad (2)$$

với ràng buộc $\sum_{j=1}^p |\beta_j| \leq t$.

Trong đó, $t \geq 0$ được gọi là tham số hiệu chỉnh của mô hình LASSO. Tham số này đóng vai trò quan trọng trong mô hình, quyết định số các hệ số hồi quy được ước lượng bằng 0.

Bài toán tối ưu ràng buộc (2) có thể viết dưới dạng hàm phạt như sau:

$$\hat{\beta}^{LASSO} = \arg \min_{\beta} \left[\sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right] \quad (3)$$

hoặc viết dưới dạng ma trận

$$\hat{\beta}^{LASSO} = \arg \min_{\beta} \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right] \quad (4)$$

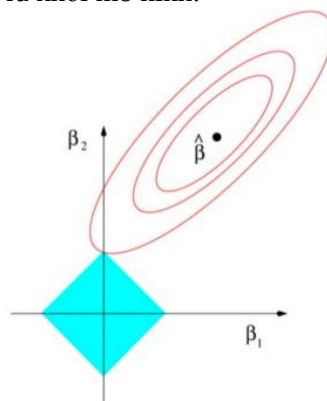
với $\lambda \geq 0$ là tham số Lagrangian có ảnh hưởng tới việc lựa chọn các biến trong mô hình hồi quy LASSO. Như vậy, bài toán (2) và (3) tương đương nhau, theo nghĩa, với mỗi giá trị λ cho trước sẽ tồn tại một tham số hiệu chỉnh t sao cho hai bài toán (2), (3) có cùng tập nghiệm.

Nếu $\lambda = 0$ thì bài toán (3) trở thành bài toán tìm hệ số hồi quy tuyến tính đa biến thông thường (phương pháp bình phương tối thiểu). Mối quan hệ giữa λ và t là mối quan hệ nghịch chiều. Nếu t tiến tới vô cùng thì λ tiệm cận tới 0. Ngược lại, khi t tiến tới 0 thì λ tiến tới vô cùng. λ càng lớn thì $\hat{\beta}^{LASSO}$ càng nhỏ, nghĩa là càng nhiều hệ số hồi quy LASSO bằng 0.

2.2. Ý nghĩa hình học của LASSO

Với tính chất của hàm trị tuyệt đối trong ràng buộc (2), hồi quy LASSO thực hiện lựa chọn biến, theo nghĩa hệ số hồi quy ước lượng của biến loại chính xác bằng 0. Để hình dung rõ hơn tính chất này của LASSO, ta minh họa LASSO trong không gian 2 chiều (chỉ có 2 biến giải thích).

Hình 1 thể hiện ý nghĩa hình học của LASSO với $p = 2$ và $\beta = (\beta_1, \beta_2)^T$. Trong đó, β là ước lượng bình phương tối thiểu của hệ số hồi quy β . Các đường cong ellip minh họa cho đồ thị của hàm tổng bình phương sai số trong (2) và miền ràng buộc $|\beta_1| + |\beta_2| \leq t$ được thể hiện bằng hình vuông xoay có tâm tại gốc tọa độ. Khi đó nghiệm của LASSO chính là điểm tiếp xúc của đường cong ellip với hình vuông. Quan sát hình 1 ta thấy, điểm tiếp xúc đó là một trong các đỉnh của hình vuông, tức là một trong các hệ số hồi quy bằng 0. Trong trường hợp này, với $t = 1$ ta có nghiệm LASSO là $\beta_1 = 0$ và $\beta_2 = 1$. Như vậy, thông qua hồi quy LASSO ta có thể loại biến thứ nhất (ứng với hệ số hồi quy bằng 0) ra khỏi mô hình.



Hình 1. Minh họa LASSO trong không gian 2 chiều

2.3. Các thuật toán tìm nghiệm LASSO

Không giống như trường hợp hồi quy tuyến tính đa biến thông thường, nghiệm của LASSO hay các ước lượng hệ số hồi quy LASSO không có công thức dạng tường minh. Vì vậy, Tibshirani và các nhà nghiên cứu khoa học đã đề xuất một số thuật toán vận dụng trên máy tính để tìm nghiệm LASSO.

Tibshirani xác định rằng công thức LASSO với ràng buộc (2) có thể được thiết lập như một bài toán tìm cực tiểu của hàm mục tiêu dạng toàn phương với các ràng buộc xác định trên một tập

lỗi. Khi đó nghiệm của bài toán sẽ là nghiệm tối ưu toàn cục. Đồng thời Tibshirani cũng chỉ ra rằng nghiệm LASSO là duy nhất nếu X là ma trận trực giao và $X^T X$ có hạng đầy đủ.

Kể từ khi phương pháp tìm nghiệm LASSO của Tibshirani được đề xuất, nhiều phương pháp phức tạp hơn đã được phát triển để giải các bài toán tìm cực trị với nhiều ràng buộc. Đáng chú ý nhất trong số này bao gồm các công trình của Osborne và cộng sự [8] cho phương pháp tuyến tính hóa địa phương, Efron và cộng sự [1] giới thiệu hồi quy góc nhỏ nhất (LARS) - một kỹ thuật giải quyết LASSO và các bài toán tìm cực trị khác rất hiệu quả. Những kỹ thuật này có đặc tính hội tụ nhanh và có thể sử dụng được ngay cả khi số lượng tham số vượt quá số lượng quan sát. Tuy nhiên do tính phức tạp của chúng, các kỹ thuật này không được mô tả chi tiết ở đây.

Đối với dạng hàm phạt, cách hữu hiệu để tìm nghiệm của phương trình (3) là sử dụng các phương pháp lặp như Newton-Raphson hoặc Nelder-Mead. Tuy nhiên, các phương pháp này được chứng minh có độ phức tạp tính toán lớn. Tibshirani có gợi ý về một phương pháp liên quan đến hồi quy Ridge để tìm nghiệm LASSO, nhưng không mô tả cụ thể trong bài báo của mình vì ông cho rằng phương pháp đó không hiệu quả. Một phương pháp với ý tưởng tương tự sau đó đã được đề xuất bởi Fan và Li [2] và được mô tả là hiệu quả do kết hợp với thuật toán Newton-Raphson. Phương pháp đó sử dụng xấp xỉ sau:

$$|\beta_j| \approx \frac{\beta_j^2}{|\beta_j|} \quad (5)$$

Thay giá trị gần đúng này vào công thức hàm phạt LASSO (3) chúng ta thu được biểu thức sau:

$$\beta_{new} = (X^T X + \lambda \cdot \text{diag}(|\beta|)^{-1})^{-1} X^T Y \quad (6)$$

Sử dụng β_{new} và biểu thức (6) sau mỗi vòng lặp ta sẽ tìm được hệ số hồi quy ước lượng mới tốt hơn. $\text{diag}(|\beta|)^{-1}$ là ma trận nghịch đảo tổng quát của ma trận với các phần tử trên đường chéo chính là thành phần của vector β . Việc đưa ma trận nghịch đảo tổng quát vào vòng lặp là cần thiết để loại bỏ các β_j gần bằng 0 ra khỏi mô hình và tránh sự mất ổn định trong thuật toán.

2.4. Lựa chọn tham số hiệu chỉnh

Tùy thuộc vào công thức được sử dụng, xác định giá trị cụ thể của t hoặc λ là cần thiết để tính toán nghiệm LASSO. Như vậy, với mỗi tham số t hoặc λ khác nhau, ta lại có nghiệm hệ số hồi quy LASSO khác nhau. Thông thường, chúng ta muốn chọn được tham số hiệu chỉnh sao cho thu được mô hình hồi quy tối ưu, tức là mô hình có trung bình bình phương sai số (MSE) bé nhất. Tibshirani đề xuất hai phương pháp để lựa chọn mô hình, đó là kiểm định chéo k mức (CV) và kiểm định chéo tổng quát (GCV). Trong phương pháp CV, tập dữ liệu đầy đủ L sẽ được chia ngẫu nhiên thành k tập con có kích thước bằng nhau. Với k mẫu con, s mẫu được chọn làm tập kiểm định, ký hiệu L^s , và $k-s$ mẫu còn lại sử dụng để ước lượng mô hình và được gọi là tập huấn luyện, ký hiệu $L - L^s$, $s=1, \dots, k$. Ước lượng kiểm định chéo của sai số dự đoán được định nghĩa như sau:

$$CV(t) = \sum_{s=1}^k \sum_{(y_i, x_i) \in L^s} (y_i - \hat{y}_i)^2 = \sum_{s=1}^k \sum_{(y_i, x_i) \in L^s} (y_i - x_i^T \hat{\beta}^s)^2 \quad (7)$$

Trong đó, $\hat{\beta}^s$ là ước lượng của β trên tập huấn luyện. Khi đó ta chọn tham số hiệu chỉnh t sao cho tại t hàm $CV(t)$ đạt cực tiểu, tức là:

$$\hat{t} = \arg \min_t [CV(t)] \quad (8)$$

3. Ứng dụng LASSO trong phân tích dữ liệu ung thư vú

3.1. Mô tả dữ liệu

Bộ dữ liệu “bcTCGA” được lấy từ các mẫu mô ung thư vú của dự án Bản đồ bộ gen ung thư (TCGA), Hiệp hội ung thư Hoa Kỳ và được tải trực tiếp trên trang web <https://myweb.uiowa.edu/pbreheny/data/bcTCGA>. Các kết quả của tập dữ liệu thu được bằng cách sử dụng kỹ thuật microarray (vi mạch) phân tích biểu hiện gen.

Tên “BRCA” là chữ viết tắt của “gen BReast Cancer - gen ung thư vú”. BRCA1 và BRCA2 là hai gen khác nhau được phát hiện có tác động đến khả năng phát triển ung thư vú của một người. Mỗi người đều có cả gen BRCA1 và BRCA2. Không giống với tên gọi của mình, các gen BRCA không gây ra ung thư vú. Trên thực tế, những gen này thường đóng một vai trò quan trọng trong việc ngăn ngừa ung thư vú. Chúng giúp sửa chữa các đứt gãy DNA có thể dẫn đến ung thư và sự phát triển không kiểm soát của các khối u. Do đó, gen BRCA còn được gọi là gen ức chế khối u.

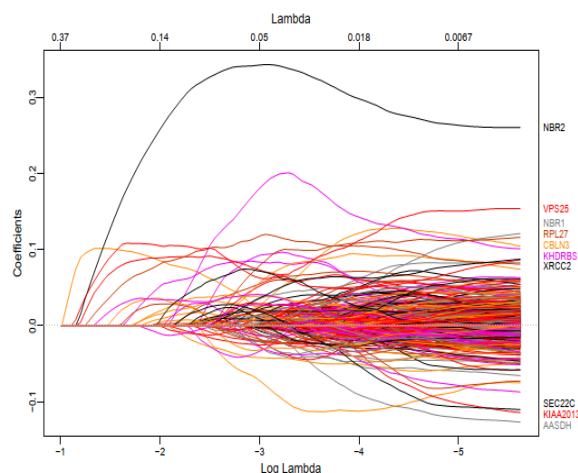
Tuy nhiên, ở một số người, các gen ức chế khối u này không hoạt động bình thường. Khi một gen bị thay đổi hoặc bị hỏng, nó sẽ không hoạt động chính xác. Đây được gọi là đột biến gen. Khi gen BRCA bị đột biến, gen này có thể không còn hiệu quả trong việc sửa chữa các DNA bị hỏng và giúp ngăn ngừa ung thư vú. Do đó, những người có đột biến gen BRCA dễ bị ung thư vú và dễ mắc bệnh ở độ tuổi trẻ hơn. Người mang gen đột biến cũng có thể truyền đột biến gen cho con cái của họ.

Trong hai gen, BRCA1 là gen đầu tiên được xác định làm tăng nguy cơ ung thư vú khởi phát sớm. Vì BRCA1 có khả năng tương tác với nhiều gen khác, bao gồm các chất ức chế khối u và các chất điều hòa chu kỳ phân chia tế bào nên việc tìm ra các gen có mức độ biểu hiện liên quan đến BRCA1 rất quan trọng trong y học lâm sàng và trong các nghiên cứu y khoa. Do đó, mục tiêu nghiên cứu của chúng tôi trong bài báo này là tìm các gen có ảnh hưởng nhất và liên quan nhất tới gen BRCA1 bằng phương pháp LASSO. Như vậy, biến tiên lượng (y) trong mô hình LASSO là biến BRCA1, các biến dự báo là cấu hình biểu hiện gen của khối u ung thư vú đối với từng bệnh nhân.

Tập dữ liệu chứa các phép đo biểu hiện của 17814 gen từ 536 bệnh nhân; tất cả các phép đo được ghi lại theo thang đo logarit.

3.2. Kết quả phân tích dữ liệu

Để hỗ trợ phân tích dữ liệu, chúng tôi sử dụng ngôn ngữ lập trình R đã tích hợp một số gói lệnh phù hợp với mô hình hồi quy tuyến tính và mô hình lựa chọn dữ liệu. Kết quả phân tích được thể hiện trong Hình 2 và Hình 3.



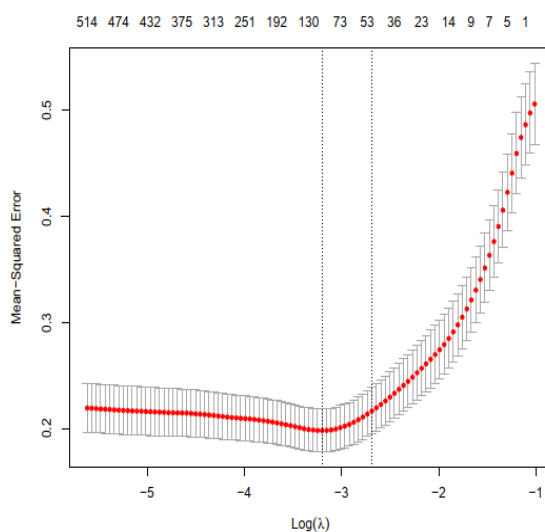
Hình 2. Lựa chọn biến

Hình 1 thể hiện mối liên hệ giữa tham số λ và các hệ số ứng với các biến của mô hình hồi quy. Mỗi đường cong đại diện cho một trong các biến dự báo và thể hiện vai trò của biến đó với biến tiên lượng BRCA1. Đường cong của biến nào càng thu hẹp dần tới 0 thì hệ số của biến đó trong mô hình gần bằng 0, tức là ta có thể loại biến đó ra khỏi mô hình. Phân tích biểu đồ trong Hình 2, chúng ta có thể nói rằng hai biến có ảnh hưởng nhiều nhất đến mô hình là biến NBR2 và biến AASDH. Do đó, chúng ta có thể giả định rằng hai biểu hiện gen này có liên quan để đánh giá xem một mẫu mới có khối u hay không.

Thực tế, NBR2 tiếp giáp với BRCA1 trên nhiễm sắc thể 17, và gần đây nhiều bằng chứng thực nghiệm chỉ ra rằng hai gen này có chung một vùng gen khởi động (là vùng trình tự trên gen cần thiết trong quá trình phiên mã các mRNA ở sinh vật nhân thực) [11]. Vì vậy sự xuất hiện của gen NBR2 trong mô hình hồi quy LASSO thể hiện phân tích của chúng tôi có ý nghĩa thực tiễn.

Cần lưu ý rằng, NBR2 không phải là gen đầu tiên xuất phát trong chùm đường cong LASSO, tức là, nó không phải là gen có liên kết biên cao nhất với BRCA1. Điều này minh họa sức mạnh của phương pháp hồi quy LASSO so với các phương pháp kiểm tra liên kết gen đơn lẻ để xác định các yếu tố sinh học quan trọng nhất từ một khối lượng lớn dữ liệu nhiễu.

Ngoài ra, chúng ta có thể lựa chọn được các biến quan trọng khác bằng cách xem xét xu hướng của chúng, chẳng hạn như: VPS25, KIAA2013, NBR1, SEC22C,... Các biến còn lại dường như ít quan trọng hơn đối với mô hình. Chú ý, trong hình 2 chỉ có 10 biến quan trọng nhất, liên quan đến gen BRC1 mới được kể tên.

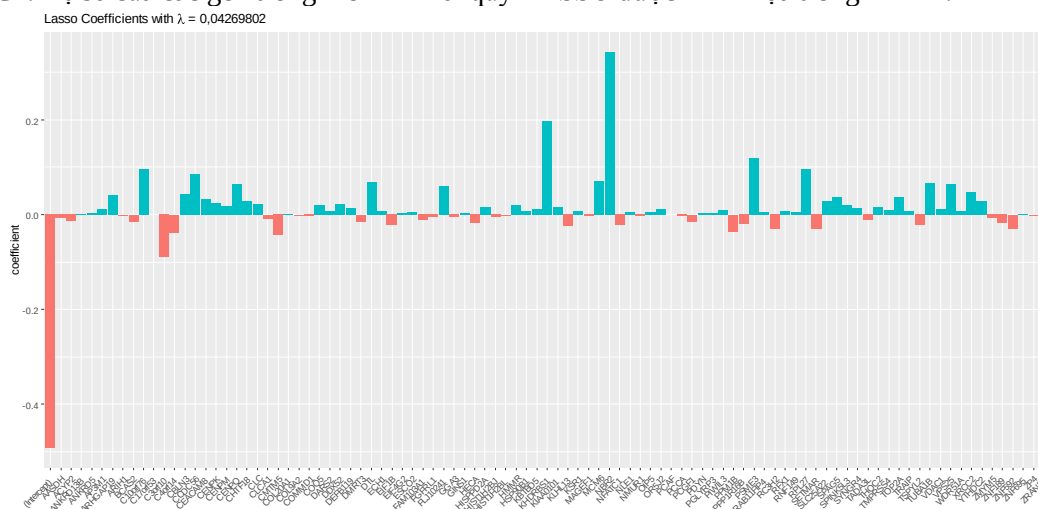
**Hình 3.** Lựa chọn tham số λ

Bước tiếp theo là chọn giá trị nào của λ để xem xét và xây dựng mô hình, hay nói cách khác là lựa chọn mô hình tối ưu. Trên phương diện thống kê, mô hình tối ưu được hiểu theo nghĩa mô hình có ít biến số nhất nhưng cung cấp nhiều thông tin nhất.

Gói lệnh **cv.glmnet** trong R giúp chúng ta chọn được giá trị thích hợp nhất cho λ bằng thuật toán kiểm định chéo. Kết quả kiểm định mô hình được minh họa trong Hình 3. Hình vẽ thể hiện mối liên quan giữa $\log(\lambda)$ và MSE (trung bình bình phương sai số). Giá trị MSE đại diện cho hiệu suất của mô hình, MSE càng nhỏ thì mô hình dự báo càng chuẩn xác. Vì vậy, ta mong muốn sẽ lựa chọn được giá trị λ làm MSE đạt giá trị nhỏ nhất.

Phần trên của biểu đồ là số biến dự báo có ảnh hưởng đến mô hình (tức là các biến có hệ số hồi quy khác không). Các đường nét đứt chỉ ra các giá trị λ_{min} và λ_{1se} , hai giá trị được chọn cho λ . Cụ thể, λ_{min} ứng với mô hình có MSE nhỏ nhất, còn λ_{1se} đưa ra mô hình có MSE nằm trong

một sai số tiêu chuẩn của MSE nhỏ nhất. Tùy từng bài toán cụ thể, ta có thể chọn một trong hai giá trị λ cho phù hợp với vấn đề nghiên cứu. Với dữ liệu của chúng ta, thuật toán kiểm định chéo đưa ra kết quả $\lambda_{min} = 0,04269802$ và $\lambda_{1se} = 0.06798732$. Hơn nữa, giá trị λ_{min} cho kết quả 73 gen được chọn trong 17814 gen, nghĩa là các gen này có ảnh hưởng nhất tới gen tiên lượng ung thư vú BRC1. Hệ số của các gen trong mô hình hồi quy LASSO được minh họa trong Hình 4.



Hình 4. Hệ số của mô hình hồi quy LASSO ứng với $\lambda_{min} = 0,04269802$

Như vậy, sử dụng phân tích này, chúng tôi đã thu được các gen có liên quan nhất trong việc phát hiện bệnh ung thư vú bao gồm các gen sau: NBR2, AASDH, KIAA2013, VPS25, NBR1, SEC22C, RPL27, CBLN3, KHDRBS1, XRCC2.

4. Kết luận

Một khó khăn trong phân tích mô hình hồi quy tuyến tính là tìm được một mô hình tối ưu. Mô hình tối ưu ở đây hiểu theo nghĩa ít biến dự báo nhất, nhưng dự đoán chính xác nhất. Đồng thời, trong trường hợp đa cộng tuyến xảy ra, nghĩa là 2 hay nhiều hơn 2 biến dự báo có liên quan mật thiết với nhau, việc xây dựng mô hình hồi quy càng trở nên khó khăn. Vì vậy, trong bài báo này chúng tôi tập trung nghiên cứu một trong các phương pháp giải quyết vấn đề trên, đó là phương pháp hồi quy LASSO.

Hồi quy LASSO là một phương pháp thay thế cho kỹ thuật hồi quy tiêu chuẩn và kỹ thuật lựa chọn biến.

Thu hẹp kích thước của hệ số hồi quy và lựa chọn biến là những mục tiêu quan trọng trong phân tích dữ liệu microarray, trong đó số lượng mẫu được thu thập nhỏ hơn nhiều so với số lượng gen trên mỗi chip, tức là số lượng biến dự báo lớn hơn nhiều so với kích thước mẫu. Việc ước lượng hệ số hồi quy trong trường hợp này có thể được thực hiện bằng cách sử dụng LASSO.

Chúng tôi đề xuất sử dụng phương pháp LASSO trong xây dựng mô hình hồi quy tuyến tính biểu thị mối quan hệ giữa gen BRC1 và các gen khác của bệnh nhân ung thư vú. Mục tiêu của chúng tôi là xác định gen nào ảnh hưởng đến mức độ nghiêm trọng của ung thư vú theo dữ liệu microarray. Hồi quy LASSO cho hiệu suất dự đoán tương đối tốt và hợp lý. Tóm lại, tập hợp các hệ số LASSO khác không đại diện cho các gen có ảnh hưởng nhất tới gen BRC1. Đó là các gen NBR2, AASDH, KIAA2013, VPS25, NBR1, SEC22C, RPL27, CBLN3, KHDRBS1, XRCC2.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least Angle Regression (with discussion)," *Annals of Statistics*, vol. 32, pp. 407-499, 2004.

-
- [2] J. Fan and R. Li, "Variable selection via non concave penalized likelihood and its oracle properties," *Journal of American Statistical Association*, vol. 96, pp. 1348–1360, 2001.
 - [3] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, pp. 101-148, 2010.
 - [4] W. J. Fu, "Penalized regression: The bridge versus the LASSO," *Journal of Computational and Graphical Statistics*, vol. 7, pp. 397-416, 1998.
 - [5] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society Series B*, vol. 67, pp. 301-320, 2005.
 - [6] H. Zou, "The adaptive lasso and its oracle properties," *Journal of American Statistical Association*, vol. 101, pp. 1418-1429, 2006.
 - [7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer: New York, NY, USA, 2008.
 - [8] M. R. Osborne, B. Presnell, A. Brevin, and B. Turlach, "On the LASSO and its dual," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 319-338, 2000.
 - [9] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society Series B*, vol. 58, pp. 267-288, 1996.
 - [10] A. S. Amusan and I. O. Adeshina, "Multicollinearity Regularization Using Lasso and Ridge Regression on Economic Data," *Kasu Journal of Mathematical Sciences*, vol. 2, no. 2, pp. 43-54, 2021.
 - [11] A. Brown, F. Xu, H. Nicolai, B. Griffiths, A. Chambers, D. Black, and E. Solomon, "The 5' end of the BRCA1 gene lies within a duplicated region of human chromosome 17q21," *Oncogene*, vol. 12, pp. 2507-2513, 1996.