MÔ HÌNH HOÁ THỐNG KÊ - BÀI TẬP 1
Hiện tượng đa cộng tuyến

Nhóm 10
Trịnh Quang Trí - 23C23011
Phạm Thị Hòa - 23C23007

1 Định nghĩa

Trong mô hình hồi quy bội:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \tag{1}$$

Hiện tượng đa cộng tuyến (multicollinearity) là tình trạng trong phân tích hồi quy khi một hoặc nhiều biến độc lập có mức độ tương quan cao với nhau. Hiện tượng này có thể gây ra rất nhiều vấn đề nghiêm trọng trong mô hình này.
Có hai loại đa cộng tuyến:

Đa cộng tuyến hoàn hảo (Perfect Multicollinearity) Đa cộng tuyến hoàn hảo xảy ra khi có một mối tương quan tuyến tính chính xác (tức là tương quan có hệ số tuyến tính bằng ± 1) giữa hai hoặc nhiều biến độc lập trong mô hình. Trong trường hợp này, một hoặc nhiều biến độc lập có thể được biểu diễn tuyến tính hoàn toàn bằng các biến độc lập khác, và do đó, không thể ước lượng các hệ số của chúng một cách riêng lẻ. Đa cộng tuyến hoàn hảo gây ra sự không thể ước lượng hợp lý (unidentifiability) và làm cho mô hình trở nên không hợp lý.

Đa cộng tuyến không hoàn hảo (Imperfect Multicollinearity) Đa cộng tuyến không hoàn hảo xảy ra khi có mức độ tương quan cao giữa các biến độc lập, nhưng không phải là tuyến tính hoàn toàn. Trong trường hợp này, mặc dù không có một tương quan tuyến tính chính xác giữa

các biến, nhưng vẫn có một mối tương quan đáng kể giữa chúng. Đa cộng tuyến không hoàn hảo có thể gây ra các vấn đề như không ổn định của ước lượng hệ số và mất tính diễn giải của mô hình.

2 Nguyên nhân

Hiện tượng đa cộng tuyến có thể xảy ra bởi môt số nguyên nhân sau:

- Thiết kế thí nghiệm và quản lí lượng quan trắc không tốt trong quá trình thu thập dữ liệu.
- Phát sinh các biến phụ thuộc mới từ các biến ban đầu.
- Có sự trùng lặp giữa các biến phụ thuộc
- Số lượng các quan trắc không đủ lớn
- ...

3 Tìm hiểu các nhận biết và cách khắc phục

- Kiểm tra trực quan: Sử dụng biểu đồ heatmap, scatter plot cho nhiều biến để nhìn được sự tương quan của các cột dữ liệu với nhau
- Tính toán hệ số tương quan: Tính toán ma trận tương quan giữa các biến độc lập. Nếu có một số cặp biến có hệ số tương quan cao (gần 1 hoặc -1), đó là dấu hiệu của đa cộng tuyến.
- PCA: Có thể được sử dụng để mô tả mức độ của vấn đề khi tập dữ liệu có quá nhiều yếu tố dự đoán để có thể trực quan. Nếu thành phần chính đầu tiên chiếm một lượng lớn phần trăm phương sai, điều này ngụ ý rằng có ít nhất một nhóm yếu tố dự đoán đại diện cho cùng một thông tin. Tải PCA có thể được sử dụng để hiểu cái nào các yếu tố dự đoán được liên kết với từng thành phần để làm sáng tỏ mối quan hệ này.
- Lạm phát phương sai (variance inflation factor - VIF) trong hồi quy tuyến tính: VIF là một phép đo thống kê được sử dụng để đánh giá mức độ của đa cộng tuyến. Giá trị VIF cao (thường là hơn 10) cho biết mức độ đa cộng tuyến cao và cần phải được xử lý.

Cách khắc phục - các bước:

1. Tính toán ma trận tương quan của các yếu tố dự đoán.
2. Xác định hai yếu tố dự đoán liên quan đến cặp tuyệt đối lớn nhất mối tương quan (gọi chúng là yếu tố dự đoán A và B).
3. Xác định mối tương quan trung bình giữa A và các biến khác. Làm tương tự với yếu tố dự đoán B.
4. Nếu A có hệ số tương quan trung bình lớn hơn thì loại bỏ nó; mặt khác, loại bỏ yếu tố dự đoán B.
5. Lặp lại các bước 2–4 cho đến khi không có mối tương quan tuyệt đối nào vượt quá ngưỡng.

Giả sử chúng ta muốn sử dụng một mô hình đặc biệt nhạy cảm với mối tương quan giữa các yếu tố dự đoán, chúng ta có thể áp dụng ngưỡng 0,75. Điều này có nghĩa là chúng ta muốn loại bỏ số lượng dự đoán tối thiểu để đạt được tất cả theo cặp hệ số tương quan nhỏ hơn 0,75.

4 Ảnh hưởng

Trong công thức (1), các tham số β thể hiện cho mức độ biến thiên của biến phụ thuốc x lên biến dự báo Y. Khi có sự hiện diện của đa cộng tuyến cộng mô hình, các giá trị β này sẽ biến thiên gần như nhau, điều này gây khó khăn cho quá trình diễn giải ý nghĩa của mô hình. Ngoài ra, trong công thức tính phương sai của ước lượng cho các tham số β,

$$\texttt{Var}(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)S_{x_j}^2} \quad j = 1, 2 \tag{2}$$

khi hiện tượng đa cộng tuyến xảy ra, giá trị tử số $(1 - R_j^2)$ trong công thức (2) sẽ tiến gần về 0, làm cho giá trị phương sai sẽ rất lớn, do đó mô hình hồi quy sẽ không còn phù hợp.
Đa cộng tuyến còn ảnh hưởng rất lớn đến quá trình ước lượng tham số β

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3}$$

như công thức trên, để xác định được các tham số $\hat{\beta}$, thì $\mathbf{X}'\mathbf{X}$ phải lấy được nghịch đảo $\rightarrow \det(\mathbf{X}'\mathbf{X}) \neq 0$, nhưng khi có đa cộng tuyến giá trị của định thức sẽ bằng $0 \rightarrow$ Không thể xác định được ước lượng $\hat{\beta}$.

Tài liệu

[1] Simon J. Sheather.*A Modern approach to Regression with R*

[2] Gareth James et al, *An Introduction to Statistical Learning with Applications in R*

[3] Max Kuhn, Kjell Johnsonl, *Applied Predictive Modeling*