BÀI TẬP 1 - MÔ HÌNH HOÁ THỐNG KÊ

NHÓM 10

Phạm Thị Hoà - MSHV: 23C23007

Trịnh Quang Trí - MSHV: 23C23

Ngành: Lý Thuyết Xác Suất và Thống Kê Toán Học

Khoá: 2023

Ngày 10 tháng 5 năm 2024

Câu hỏi:

1. Hiện tượng đa cộng tuyến (Multicollinearity) là gì?
2. Tìm hiểu các nhận biết và cách khắc phục
3. Các loại đa cụm tuyến (hoàn hảo hoặc ko hoàn hảo)?
4. Hậu quả (Ví dụ: Không tìm được beta mũ, ...)

Trả Lời

1. Hiện tượng đa cộng tuyến là gì?

Cộng tuyến là thuật ngữ kỹ thuật cho tình huống trong đó một cặp biến được dự đoán có mối quan hệ tương quang đáng kể với nhau. Và nó cũng có

khả năng rằng có mối quan hệ giữa nhiều dự đoán cùng một lúc được gọi là hiện tượng đa cộng tuyến.
Vậy 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.

2. 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.

3. Các loại đa cụm tuyến (hoàn hảo hoặc ko hoàn hảo)?