Chuẩn đoán Mô hình Hồi quy (Regression Diagnostics)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 7 tháng 1 năm 2018

Xây dựng và chuẩn đoán mô hình hồi quy

1. Thống kê mô tả dữ liệu: phát hiện khác biệt giữa các nhóm, quan sát ngoại vi, phát hiện nếu dữ liệu phân phối bất đối xứng.
2. Kiểm tra tính tương quan giữa các biến giải thích (multicolinearity/correlation).
3. Ước lượng mô hình hồi quy đơn giản và mở rộng.
4. Phát hiện và xử lý nghi vấn về cấu trúc hàm (tuyến tính hoặc phi tuyến, biến tương tác).
5. Hậu hồi quy: rà soát những vấn đề có thể xảy ra và lựa chọn mô hình phù hợp:
   - Thực hiện các loại kiểm định.
   - Hệ số phóng đại phương sai - Variance Inflation Factors (VIF).
   - Đánh giá tác động của quan sát ngoại vi.
   - Đồ thị phần dư - Residuals' plots.

Lưu ý với mô hình hồi quy đa biến

1. Chọn biến giải thích cần dựa trên lý thuyết kinh tế thay vì ý nghĩa thống kê. Với mẫu quan sát lớn, việc tăng số mẫu sẽ làm tăng sự tương quan ngẫu nhiên, mặc dù thực tế không có bất kỳ liên hệ nào giữa các biến đó.

2. Tránh đưa quá nhiều biến giải thích trong mô hình, kể cả những biến không thực sự liên quan nhằm tăng hệ số thích hợp (R^2).
$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\sum_i (\hat{y}_i - \bar{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$; $R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-K}$.

3. Tránh chọn lọc điều chỉnh dữ liệu sao cho mô hình có kết quả phù hợp với định kiến có trước.

Khi dữ liệu có phân phối lệch (skewed distribution)

- Các giả định để ước lượng OLS là BLUE không liên quan đến phân phối của dữ liệu, tuy nhiên, phân phối lệch có thể làm sai lệch điều kiện phân phối chuẩn hoặc phương sai của sai số thay đổi.

- Nếu có phân phối lệch, cần thiết phải kiểm tra ý nghĩa của biến về mặt kinh tế. Ví dụ khi ước lượng mô hình liên quan đến tỷ suất, biến phụ thuộc thường là logarit $\Rightarrow$ chuyển đổi dữ liệu sang hàm log có thể hạn chế được vấn đề phân phối lệch.

Phát hiện và xử lý vấn đề liên quan đến cấu trúc hàm

- Kiểm định giả thuyết bội F và Chow với biến bậc cao, biến tương tác.
- Kiểm định LM về thiếu biến quan trọng trong mô hình.
- Kiểm định Breusch-Pagan và White về phương sai thay đổi và điều chính nếu cần thiết.
- Kiểm định Ramsey về mô hình sai (misspecification test).

Kiểm định mô hình sai - RESET test

Kiểm định Ramsey RESET (Regression Specification Error Test) để kiểm định mô hình sai trong trường hợp tổng quát. Khác với F-test hay Chow-test kiểm định các cấu trúc hàm cho trước (bậc 2, bậc 3...).

- Giả định ta có mô hình hồi quy đa biến sau:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \tag{1}$$

- Kiểm định RESET để nhận biết liệu cấu trúc hàm trên bị sai. Mô hình có thể có thêm các biến giải thích bậc 2, biến tương tác... nhưng không biết chính xác cấu trúc.

Thực hiện kiểm định RESET

- Ước lượng mô hình (1), tính giá trị dự báo $\hat{y}$.
- Đưa giá trị dự báo bình phương và bậc ba vào mô hình gốc và ước lượng lại:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + u \quad (2)$$

- Kiểm định giả thuyết $H_o : \gamma_1 = \gamma_2 = 0$ bằng kiểm định $F_{2,n-k-3}$ hoặc LM với $df = 2$. Nếu bác bỏ H_o thì hàm hồi quy (1) có vấn đề về cấu trúc hàm.

Ví dụ kiểm định RESET

Sử dụng lại mô hình tỷ suất thu nhập với bộ dữ liệu VHLSS 2010.

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexpersq + \beta_4 married \\ &+ \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u \end{aligned}$$

Hậu hồi quy

Hệ số phóng đại phương sai - Variance Inflation Factor (VIF):

- Sử dụng để đo lường độ tương quan giữa các biến. Nếu các biến tự tương quan với nhau được sử dụng trong cùng một mô hình sẽ dẫn đến ước lượng phương sai bị chệch và kiểm định giả thuyết không chính xác.

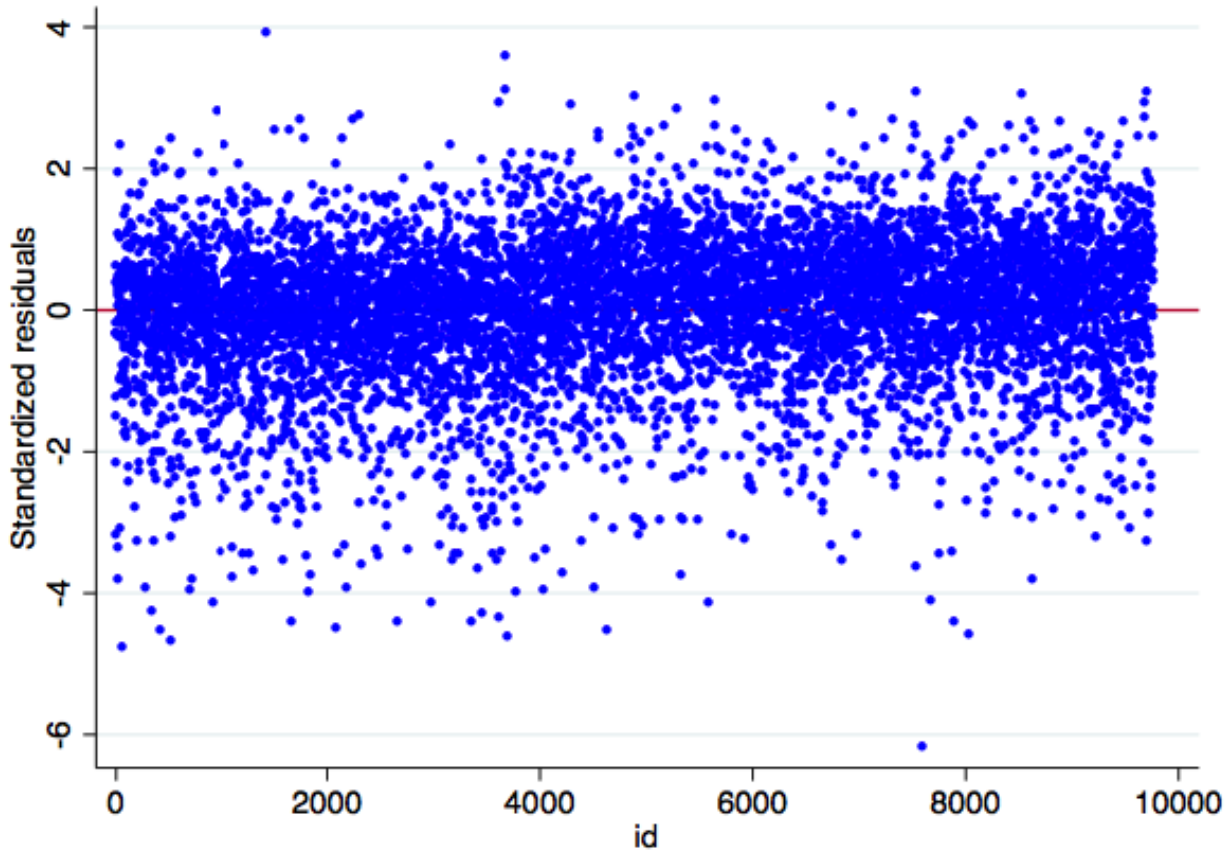
- Cần lọc ra những biến quan trọng nhất (về mặt thống kê). VIF được tính bằng cách hồi quy mỗi biến giải thích X_i dựa vào các biến khác,

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Quy ước bỏ biến có $VIF > 10$.

Đồ thị phân phối của phần dư - residuals' plots:

- Kiểm tra quan sát ngoại vi
- Kiểm tra phương sai thay đổi



Quan sát ngoại vi - Outliers

- Phát hiện dựa vào thống kê mô tả và đồ thị phân phối
- Điều chỉnh mô hình theo trọng số (phương pháp WLS)
- Bỏ các quan sát ngoại vi và ước lượng lại mô hình
- Phương pháp trị tuyệt đối tối thiểu - Least absolute deviation (LAD)

Các vấn đề liên quan đến dữ liệu

- Dữ liệu không ngẫu nhiên, hoặc dữ liệu bị chặn $\Rightarrow$ Vấn đề lựa chọn mẫu trong hồi quy (sample selection problem):
  - Ước lượng có thể bị chệch và không nhất quán.

- Dữ liệu bị thiếu:
  - Thiếu ngẫu nhiên hay thiếu hệ thống?
  - Loại bỏ quan sát bị thiếu thông tin
  - Ghép thông tin (data imputation)