

Chiến lược Xây dựng và Chuẩn đoán Mô hình Hồi quy (Regression Diagnostics)

Lê Việt Phú
Chương trình Giảng dạy Kinh tế Fulbright

Ngày 5 tháng 1 năm 2016

Một số đặc điểm đáng lưu ý của các nghiên cứu sử dụng mô hình hồi quy đa biến

1. Xu hướng chọn biến giải thích sao cho có ý nghĩa thống kê mà không quan tâm đến lý thuyết kinh tế học của mô hình ước lượng. 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. Xu hướng sử dụng quá nhiều biến giải thích trong mô hình, kể cả những biến không thực sự liên quan vì khả năng giải thích mô hình (R^2) được tăng lên.
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} ; \bar{R}^2 = 1 - (1 - R^2) \frac{N-1}{N-K}.$$
3. Xu hướng chọn lọc điều chỉnh dữ liệu sao cho mô hình có kết quả phù hợp với định kiến có trước.

Xây dựng và chuẩn đoán mô hình trong nghiên cứu thực nghiệm

1. Thống kê mô tả dữ liệu.
2. Kiểm tra tính tương quan giữa các biến giải thích (multicollinearity). Phát hiện nếu dữ liệu phân phối bất đối xứng (skewed distribution).
3. Chạy thử 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).
5. Hậu hồi quy: rà soát những vấn đề có thể xảy ra và lựa chọn mô hình phù hợp:
 - ▶ Hệ số phóng đại phương sai - Variance Inflation Factors (VIF).
 - ▶ Biến ngoại vi - Outliers.
 - ▶ Đồ thị phần dư - Residuals' plot.

Thực hành

Bộ dữ liệu của chúng ta là bộ dữ liệu điểm số SAT cuối cấp 3 (standard assessment test) của học sinh trung học tại Mỹ. Bộ dữ liệu này có số liệu trung bình của 51 bang. Chúng ta muốn ước lượng mô hình hồi quy giải thích điểm SAT theo các đặc trưng của bang như thu nhập (trung vị) của hộ gia đình, tỉ lệ chi tiêu trung bình cho mỗi học sinh tiểu và trung học, tỷ lệ học sinh thi lấy điểm SAT và các biến giải thích liên quan khác. Trong mô hình này chúng ta tạm thời bỏ qua sự khác biệt về khái niệm quan hệ tương quan với quan hệ nhân quả. Học viên có thể thực hành trên file dữ liệu có tên là `states.dta`.

Mô tả các biến sử dụng

describe csat expense percent income high college
region

Loại biến	Tên biến	Giải thích
Biến phụ thuộc	csat	điểm số SAT trung bình
Biến giải thích	expense percent income	chi phí trung bình cho một học sinh phần trăm học sinh thi lấy điểm SAT thu nhập trung bình hộ gia đình (trung vị)
	high	phần trăm người có bằng tốt nghiệp phổ thông
	college	phần trăm người có bằng tốt nghiệp cao đẳng hoặc đại học

Mô tả dữ liệu

```
summarize csat expense percent income high college  
region
```

Variable	Obs	Mean	Std. Dev.	Min	Max
csat	51	944.098	66.93497	832	1093
expense	51	5235.961	1401.155	2960	9259
percent	51	35.76471	26.19281	4	81
income	51	33.95657	6.423134	23.465	48.618
high	51	76.26078	5.588741	64.3	86.6
college	51	20.02157	4.16578	12.3	33.3
region	50	2.54	1.128662	1	4

Điểm SAT (csat), phần trăm học sinh trung học thi SAT (percent) có thể có phân phối lệch.

Hồi quy đa biến tuyến tính

Bắt đầu bằng mô hình đơn giản nhất, sau đó thêm dần các biến:

	(1)	(2)	(3)
expense	-0.0223*** (0.00367)	0.00335 (0.00478)	-0.00202 (0.00359)
percent		-2.618*** (0.229)	-3.008*** (0.236)
income		0.106 (1.207)	-0.167 (1.196)
high		1.631 (0.943)	1.815 (1.027)
college		2.031 (2.114)	4.671** (1.600)
_lregion_2			69.45*** (18.00)
_lregion_3			25.40* (12.53)
_lregion_4			34.58*** (9.450)
R-sq	0.217	0.824	0.911
adj. R-sq	0.201	0.805	0.894

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Giải thích mô hình

- ▶ Mô hình 1: chi phí có ý nghĩa thống kê, nhưng chiều hướng tác động không như kỳ vọng.
- ▶ Mở rộng mô hình để kiểm soát các biến khác cho thấy chi phí không còn có ý nghĩa thống kê $\Rightarrow$ mô hình (1) hoặc là không đầy đủ, hoặc là do biến chi phí có tương quan với biến khác trong mô hình đầy đủ.
- ▶ R^2 tăng cao khi kiểm soát thêm các biến trong mô hình (2) và (3) cho thấy sự cần thiết phải mở rộng mô hình.
- ▶ Có thể sử dụng kiểm định F để xác nhận ý nghĩa thống kê của các biến đưa thêm vào mô hình.

test percent=income=high=college=0

F(4, 41) = 99.71

Prob > F = 0.0000

Kiểm tra hệ số tương quan giữa các biến giải thích

Pearson's coefficient of correlation $r = \frac{\text{cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$; $-1 \leq r \leq 1$

`pwcorr csat expense percent income high college, star(0.05) sig`

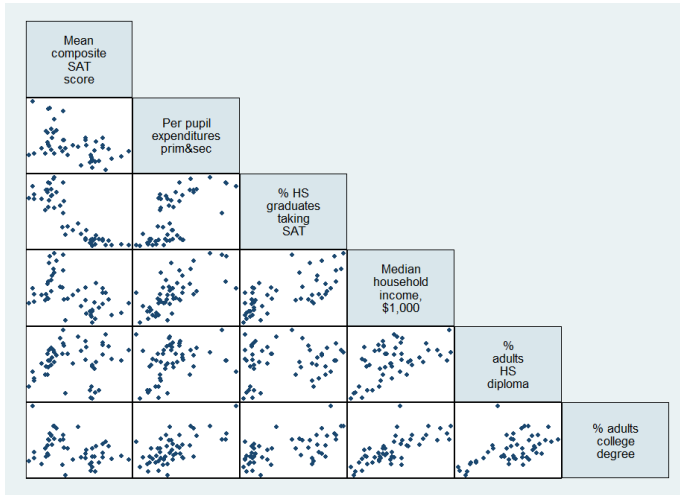
	csat	expense	percent	income	high	college
csat	1.0000					
expense	-0.4663* 0.0006	1.0000				
percent	-0.8758* 0.0000	0.6509* 0.0000	1.0000			
income	-0.4713* 0.0005	0.6784* 0.0000	0.6733* 0.0000	1.0000		
high	0.0858 0.5495	0.3133* 0.0252	0.1413 0.3226	0.5099* 0.0001	1.0000	
college	-0.3729* 0.0070	0.6400* 0.0000	0.6091* 0.0000	0.7234* 0.0000	0.5319* 0.0001	1.0000

* Có ý nghĩa thống kê ở mức 5%

Dấu hiệu tương quan khá rõ rệt giữa các biến giải thích.

Kiểm tra tính tương quan giữa các biến giải thích bằng đồ thị điểm - scatter plots

```
graph matrix csat expense percent income high college, half  
maxis(ylabel(none) xlabel(none))
```



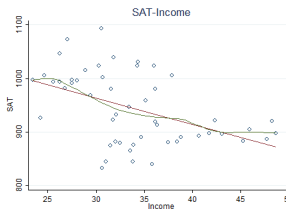
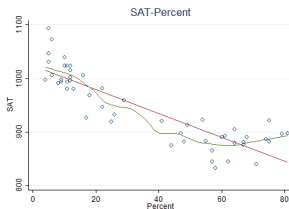
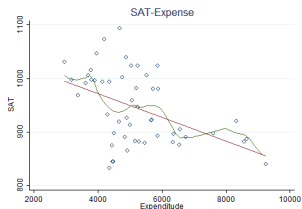
Khi dữ liệu có phân phối lệch (skewed distribution)

- ▶ Các giả định Gauss-Markov và ước lượng sử dụng OLS là BLUE không liên quan đến phân phối của dữ liệu, ngoại trừ phân phối của biến dư là IID chuẩn để kiểm định giả thuyết. Tuy nhiên, phân phối lệch có thể làm sai lệch điều kiện phân phối chuẩn của biến dư hoặc thay đổi phương sai của biến dư.
- ▶ 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.

$$\log INCOME = \beta_0 + \beta_1 * EDUC + \beta_2 * EDUC^2 + \beta_3 * EXPER + u$$

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

- Sử dụng đồ thị phân phối điểm (scatter plot) và hồi quy nội tại (local regression) để chuẩn đoán cấu trúc hàm:



Khả năng phần trăm học sinh thi SAT có quan hệ phi tuyến với điểm SAT. Tại sao lại có quan hệ nghịch biến giữa điểm số SAT và số phần trăm học sinh thi SAT?

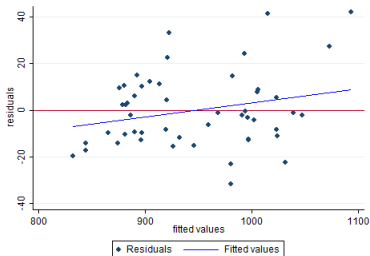
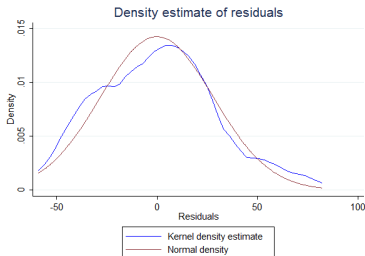
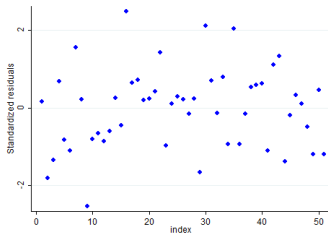
Điều chỉnh mô hình

$$csat_i = \beta_0 + \beta_1 expense_i + \beta_2 percent_i + \beta_3 income_i + \beta_4 high_i + \beta_5 college_i + \sum_j \alpha_j Region_j + \beta_6 percent_i^2 + \varepsilon_i$$

	(1)	(2)	(3)	(4)
expense	-0.0223***	0.00335	-0.00202	0.00141
percent		-2.618***	-3.008***	-5.945***
income		0.106	-0.167	-0.914
high		1.631	1.815	1.869
college		2.031	4.671**	3.418**
_lregion_2			69.45***	5.077
_lregion_3			25.40*	5.209
_lregion_4			34.58***	19.25*
percent²				0.0460***
R-sq	0.217	0.824	0.911	0.940
adj. R-sq	0.201	0.805	0.894	0.927

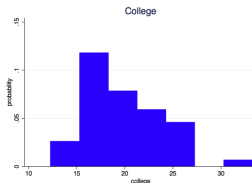
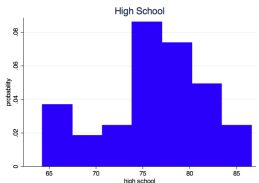
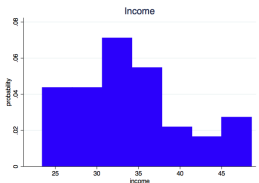
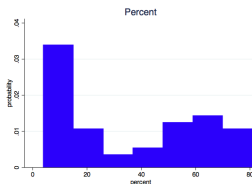
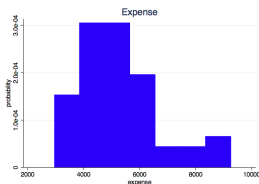
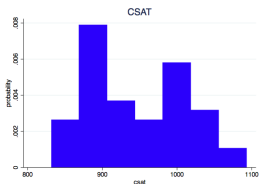
Hậu hồi quy

- ▶ Đồ thị phân phối của phần dư - residuals' plots: Kiểm tra khả năng phương sai thay đổi.
- ▶ Bỏ sót biến quan trọng trong mô hình.



Biến ngoại vi - Outliers

- ▶ Dựa vào thống kê mô tả và đồ thị phân phối
- ▶ Bỏ các quan sát ngoại vi và ước lượng lại mô hình



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.

$$csat_i = \beta_0 + \beta_1 expense_i + \beta_2 percent_i + \beta_3 income_i + \beta_4 high_i \\ + \beta_5 college_i + \sum_j \alpha_j Region_j + u_i$$

- ▶ 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,

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

Hệ số phóng đại phương sai

- Nếu biến X_i tự tương quan với các biến khác thì R_i^2 có giá trị cao, dẫn đến VIF lớn. Nguyên tắc chung là $VIF > 10$ chứng tỏ biến X_i có độ tương quan cao với các biến khác.

Variable	VIF
income	4.78
high	4.71
college	4.34
_lregion_3	4.18
percent	3.88
_lregion_2	3.57
expense	3.18
_lregion_4	1.8
Mean VIF	3.81

- Dự đoán điều gì xảy ra nếu sử dụng bình phương của phần trăm số học sinh thi SAT trong mô hình hồi quy?

Một số loại kiểm định khác

- ▶ Kiểm định RAMSEY RESET về mô hình bị thiếu biến quan trọng:

Ramsey RESET test using powers of the fitted values of csat

Ho: model has no omitted variables

$F(3, 38) = 2.15$

Prob > F = 0.1096

- ▶ Kiểm định Durbin-Watson về tương quan chuỗi:

Durbin-Watson d-statistic(9, 50) = 2.36287

- ▶ Kiểm định Breusch-Pagan về phương sai của sai số thay đổi (kỳ sau).

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of csat

$\chi^2(1) = 1.40$

Prob > χ^2 = 0.2375