Chuẩn đoán Mô hình Hồi quy

Lê Việt Phú
Chương trình Giảng dạy Kinh tế Fulbright

Ngày 5 tháng 1 năm 2015

Table of contents

1. Ôn tập lý thuyết hồi quy tuyến tính đa biến và các giả định căn bản

Giả sử chúng ta muốn ước lượng một mô hình tuyến tính đa biến:

$$Y_i = \beta_0 + \beta_1 \times x_i^1 + ... + \beta_K \times x_i^K + \varepsilon_i$$

Dưới dạng ma trận:

$$Y = X\beta + \varepsilon$$

Trong đó Y là ma trận cột Nx1 (N quan sát tương ứng với N dòng và 1 cột); X là ma trận Nxk (N quan sát, mỗi quan sát có k đặc tính); β là ma trận tham số kx1 (k tham số tương ứng với k đặc tính của biến giải thích). ε là ma trận biến dư.

Ước lượng bằng phương pháp bình phương tối thiểu:

$$\hat{\beta} = [X'X]^{-1}X'Y$$

Ôn tập lý thuyết hồi quy tuyến tính đa biến và các giả định căn bản

$\star$ Giả định Gauss-Markov để ước lượng bằng OLS là BLUE (Best Linear Unbiased Estimator):

1. $E[\varepsilon_i] = 0$
2. $Var[\varepsilon_i] = \sigma^2$
3. $Cov[\varepsilon_i, \varepsilon_j] = 0$
4. $Cov[X_i, \varepsilon_i] = 0$
5. Mối quan hệ X và Y là tuyến tính

Một số giả định khác:

6 ε_i độc lập, đồng nhất, và phân phối chuẩn (iid, normally distributed)

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.
   $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\sum_i (\hat{y}_i - \bar{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$ hoặc tối đa hóa $\bar{R}^2$.
3. Xu hướng chọn lọc điều chỉnh dữ liệu sao cho mô hình có kết quả đúng như ý muốn.

2. Các bước chuẩn đoán mô hình trong nghiên cứu thực nghiệm

1. Thống kê mô tả dữ liệu
2. Chạy thử mô hình hồi quy đơn giản và mở rộng
3. Kiểm tra tính tương quan giữa các biến giải thích
4. Phát hiện và xử lý nghi vấn về cấu trúc hàm
5. Hậu hồi quy: rà soát những vấn đề có thể xảy ra và lựa chọn mô hình phù hợp
   - Variance Inflation Factors (VIF)
   - Outliers
   - Residuals' plot
   - DfBeta
   - DfFIT
   - Cook's distance
   - Leverage

Những sự cố hay gặp phải trong mô hình hồi quy đa biến

1. Dữ liệu phân phối bất đối xứng (skewed distribution)
2. Tương quan giữa các biến giải thích (multicolinearity)
3. Quan sát ngoại vi (outliers)
4. Hàm ước lượng phi tuyến (nonlinear functions)

3. Ví dụ thực tế

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 số 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

Giả sử chúng ta quan tâm đến những biến sau:

Loại biến	**Tên biến**	**Giải thích**
Biến phụ thuộc	`csat`	điểm số SAT trung bình
Biến giải thích	`expense`	chi phí trung bình cho một học sinh
	`percent`	phần trăm học sinh thi lấy điểm SAT
	`income`	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

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

Chúng ta bắt đầu bằng mô hình đơn giản nhất, sau đó thêm dần các biến:

	(1)	(2)	(3)
expense	-0.0223***	0.00335	-0.00202
	(0.00367)	(0.00478)	(0.00359)
percent		-2.618***	-3.008***
		(0.229)	(0.236)
income		0.106	-0.167
		(1.207)	(1.196)
high		1.631	1.815
		(0.943)	(1.027)
college		2.031	4.671**
		(2.114)	(1.600)
_Iregion_2			69.45***
			(18.00)
_Iregion_3			25.40*
			(12.53)
_Iregion_4			34.58***
			(9.450)
R-sq	0.217	0.824	0.911
adj. R-sq	0.201	0.805	0.894

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 thông 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.

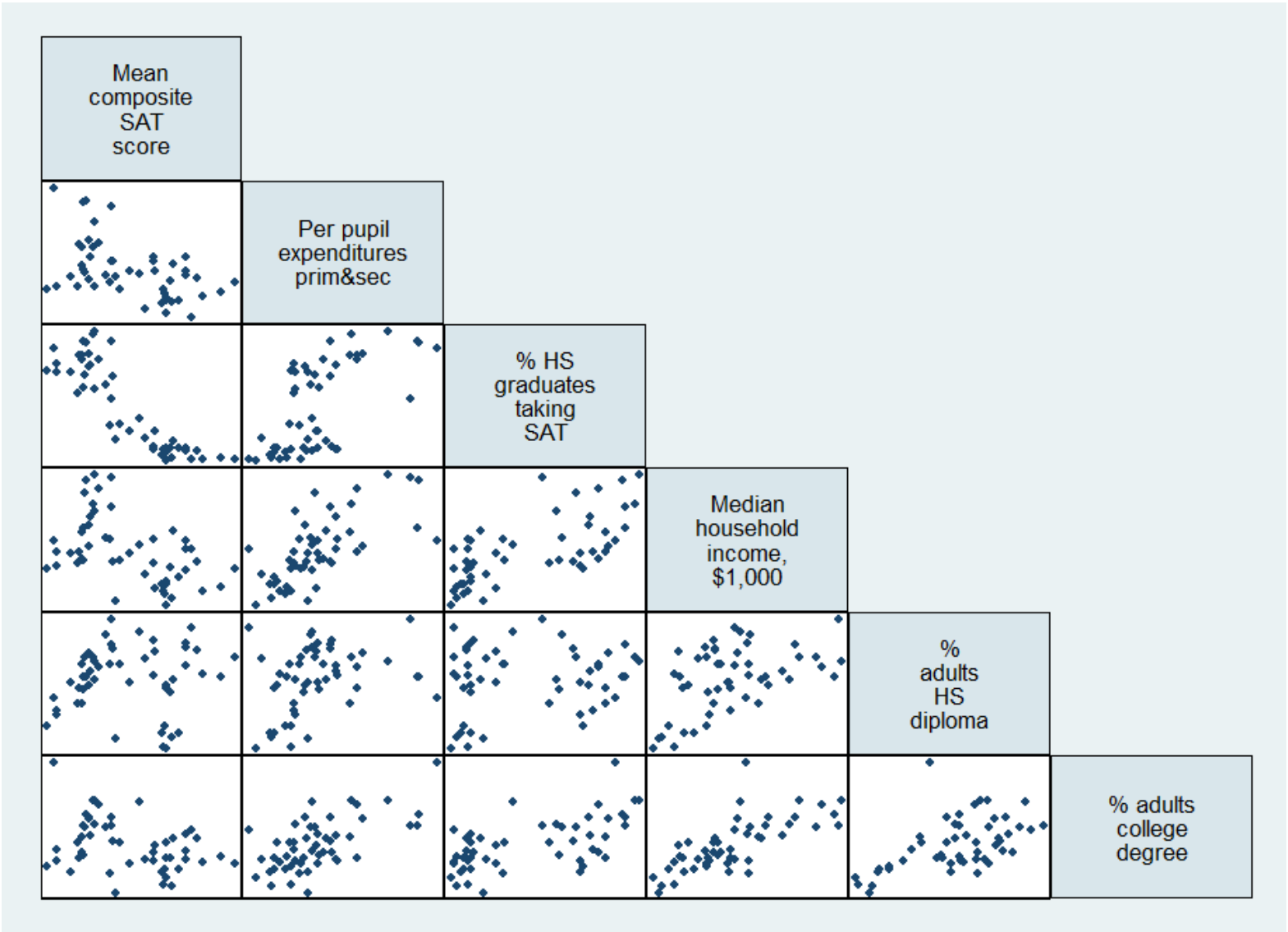
Kiểm tra tính tương quan giữa các biến

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

* 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



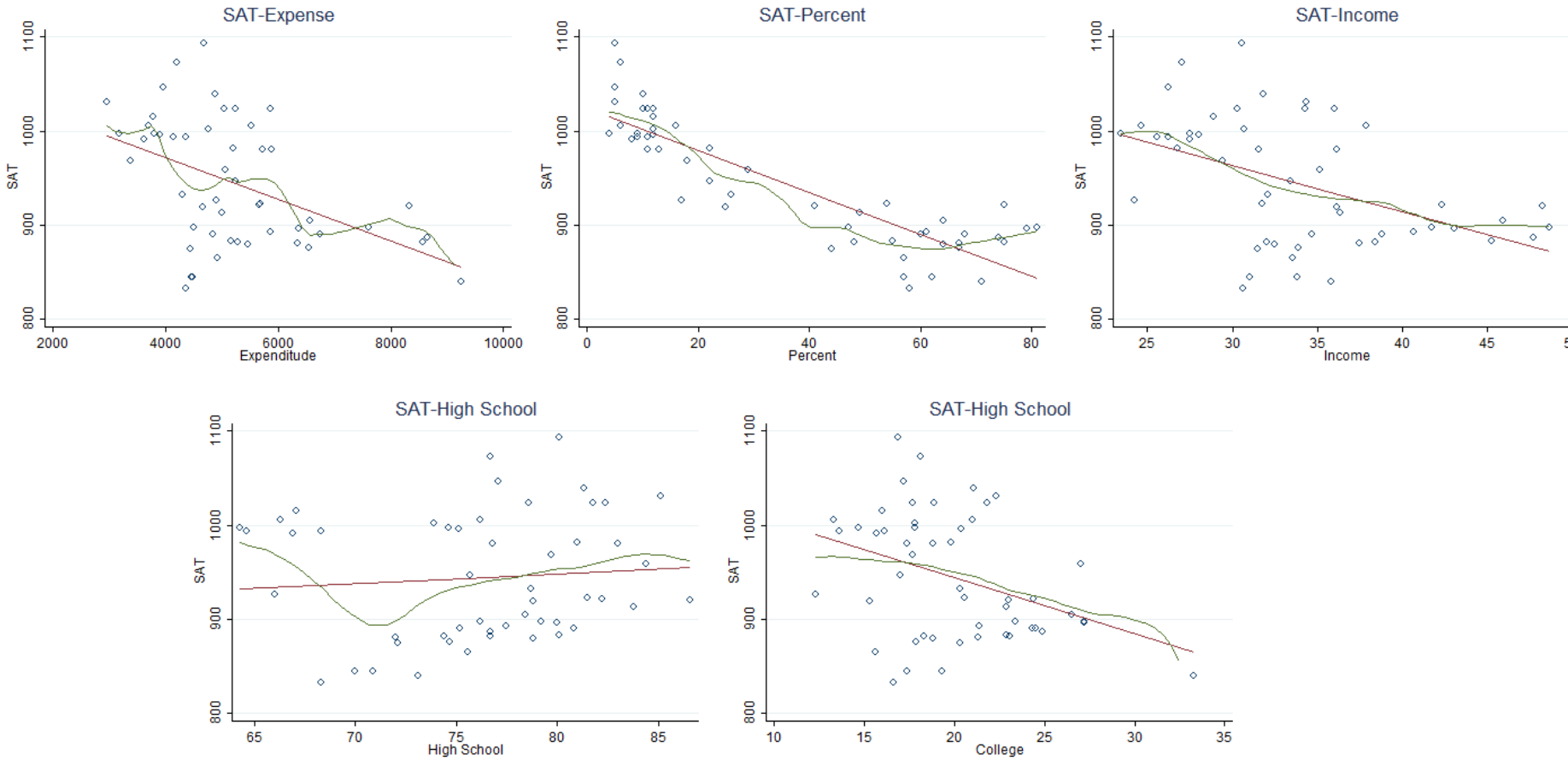
Xử lý thế nào khi dữ liệu có phân phối lệch?

- 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 đơn vị của dữ liệu sang hàm log có thể hạn chế được vấn đề phân phối lệch.

$$logY = X\beta + \varepsilon$$

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 hệ số góc âm?

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
	(0.00367)	(0.00478)	(0.00359)	(0.38)
percent		-2.618***	-3.008***	**-5.945*** **
		(0.229)	(0.236)	(-9.28)
income		0.106	-0.167	-0.914
		(1.207)	(1.196)	(-0.94)
high		1.631	1.815	1.869
		(0.943)	(1.027)	(2.01)
college		2.031	4.671**	3.418**
		(2.114)	(1.600)	(2.98)
_Iregion_2			69.45***	5.077
			(18.00)	(0.24)
_Iregion_3			25.40*	5.209
			(12.53)	(0.50)
_Iregion_4			34.58***	19.25*
			(9.450)	(2.37)
percent^2				**0.0460*** **
				(4.52)
R-sq	0.217	0.824	0.911	0.940
adj. R-sq	0.201	0.805	0.894	0.927

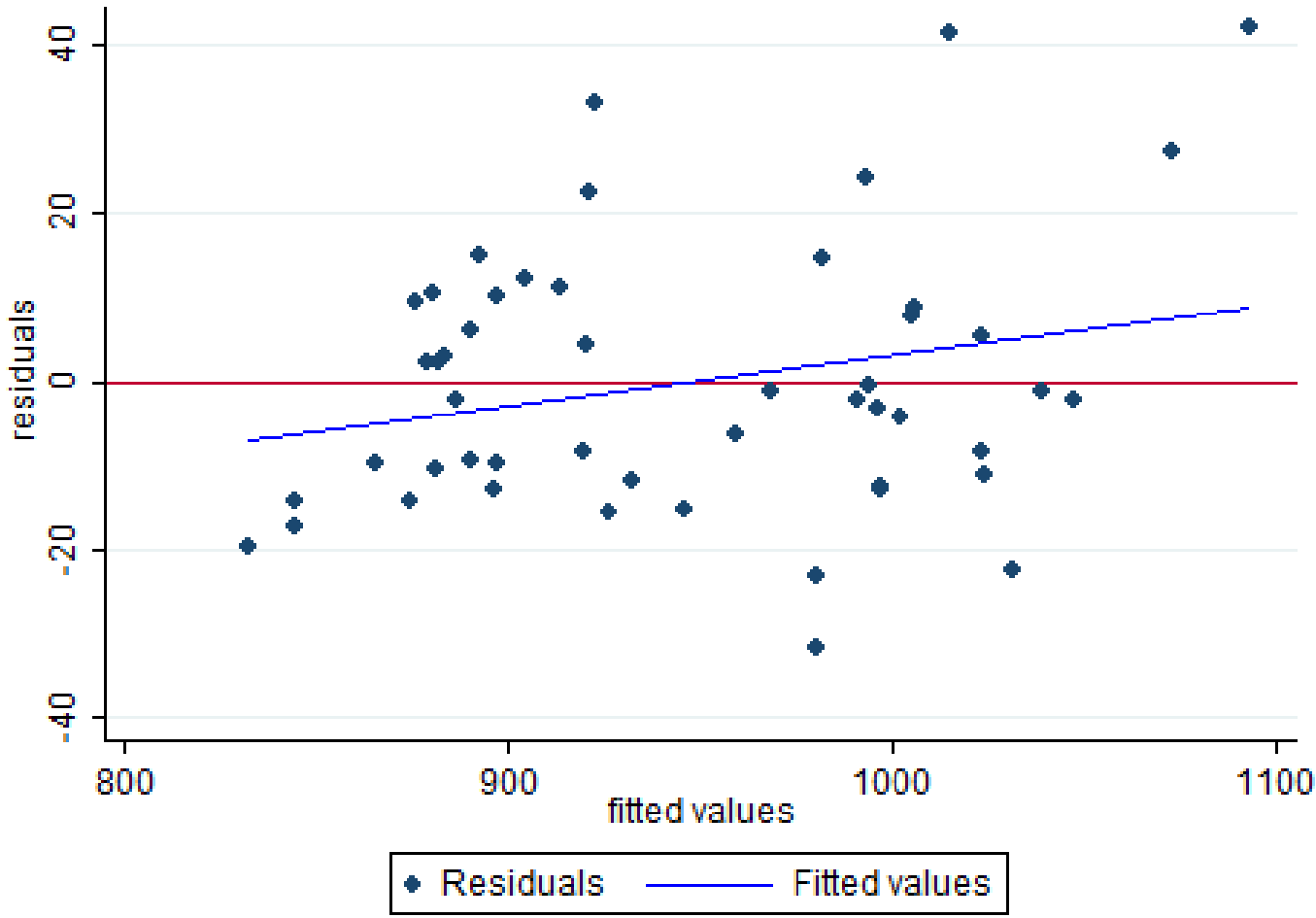
Ý nghĩa của tham số β_2 và β_6 là gì?

Hậu hồi quy: kiểm tra tính phù hợp của các biến giải thích

- Residuals' plots
- Outliers
- Variance Inflation Factors (VIF)
- DfBeta
- DfFIT
- Cook's distance
- Leverage
- Bias vs effiency tradeoff

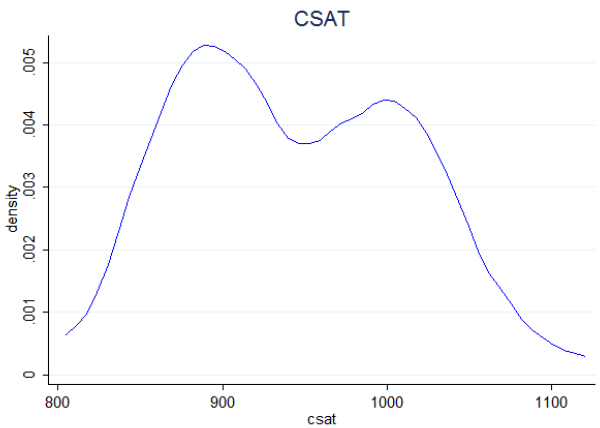
Residuals' plots

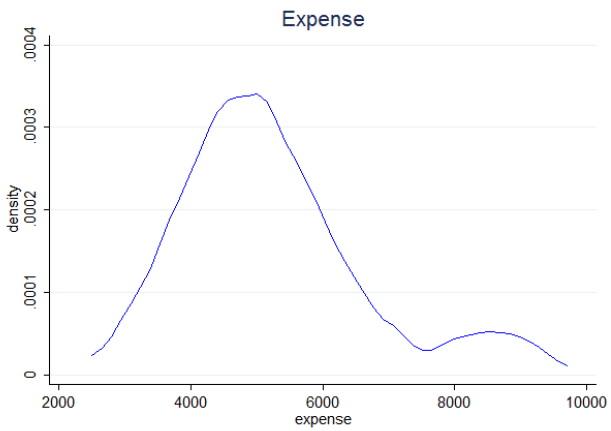
- Kiểm tra khả năng phương sai thay đổi
- Bỏ sót biến quan trọng trong mô hình
- Định dạng hàm sai

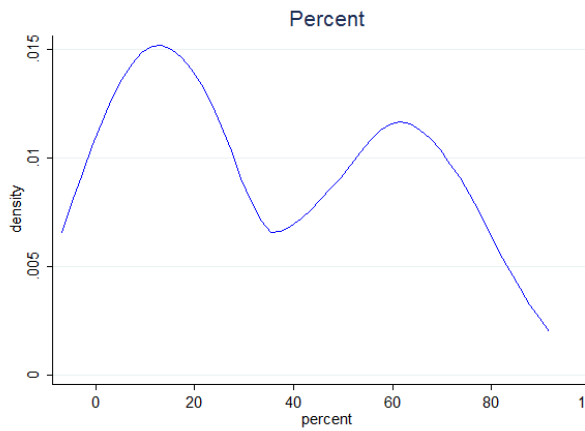


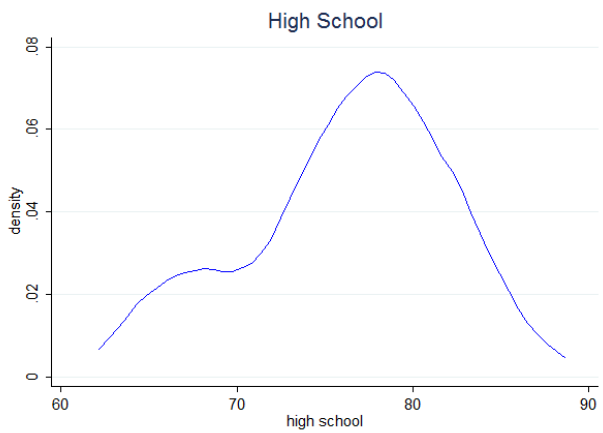
Biến ngoại vi

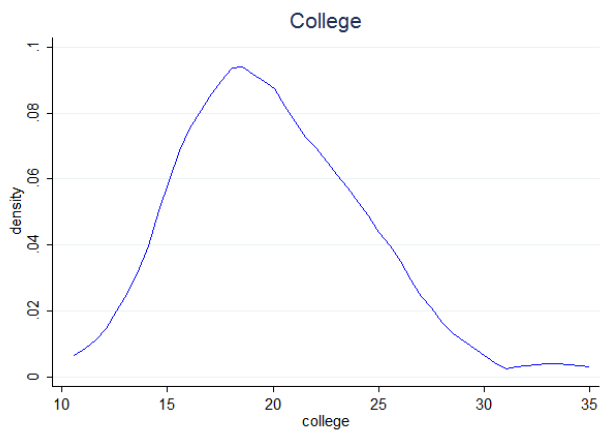
- Dựa vào thống kê mô tả và đồ thị phân phối
- Bỏ các quan sát ngoại vi và ước lượng lại mô hình











Variance Inflation Factor (VIF)

Sử dụng để đo lường độ tương quan giữa các biến. Nếu các biến tự tương quan được sử dụng trong cùng một mô hình sẽ dẫn đến ước lượng phương sai chệch và kiểm định thống kê không chính xác.
Mô hình ban đầu:

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

VIF được tính bằng cách hồi quy mỗi biến giải thích X_i dựa vào các biến khác,

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

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.

VIF

Variable	VIF
income	4.78
high	4.71
college	4.34
_Iregion_3	4.18
percent	3.88
_Iregion_2	3.57
expense	3.18
_Iregion_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 ước lượng?

Các công cụ khác

- DfBeta: kiểm tra liệu ước lượng của một tham số có bị ảnh hưởng bởi một quan sát ngoại vi nào đó.
- DfFIT: Kiểm tra liệu có một quan sát ngoại vi nào đó ảnh hưởng đến ước lượng của mô hình hay không.
- Cook's distance, leverage: các kiểm định về ảnh hưởng của biến ngoại vi.