Hồi quy Đa biến
(Multivariate Regression)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 30 tháng 11 năm 2018

Mô hình hồi quy đa biến

Tương tự như mô hình hồi quy đơn biến, tuy nhiên với nhiều biến giải thích. Ví dụ mô hình hồi quy với hai biến giải thích:

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + u_i$$

- i là quan sát thứ i trong mẫu bao gồm n quan sát
- y gọi là biến phụ thuộc/biến được giải thích
- x^1, x^2 là biến độc lập/biến giải thích
- u là sai số, bao gồm tất cả những yếu tố khác ảnh hưởng đến y nhưng không nằm trong x^1, x^2.
- β_0, β_1, β_2 là các tham số trong mô hình – cần phải ước lượng.

Phương pháp bình phương tối thiểu thông thường OLS với hồi quy đa biến

- Tìm $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ để tối thiểu hóa tổng bình phương của sai số u_i:

$$U = min \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i^1 - \beta_2 x_i^2)^2$$

  với ký hiệu i đại diện cho quan sát thứ i.

- $\hat{\beta}_1$ và $\hat{\beta}_2$ là tác động riêng phần của các biến giải thích x^1 và x^2 lên biến phụ thuộc.
- Ý nghĩa của các trị thống kê R^2, SST, SSE, SSR tương tự như mô hình SLR.

Điều kiện của ước lượng OLS

Tương tự như các điều kiện của mô hình SLR:

- Hai điều kiện bậc nhất tương ứng với $\mathbf{E}(\mathbf{u}) = \mathbf{0}$ và $\mathbf{E}(\mathbf{xu}) = \mathbf{0}$ sẽ đảm bảo ước lượng OLS là không chệch (unbiased) và nhất quán (consistent).
- Diễn giải: trung bình của sai số u bằng không và sai số u không tương quan với tất cả các biến giải thích x^1, x^2.

Diễn giải ý nghĩa của hồi quy đa biến

Với hàm hồi quy mẫu:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^1 + \hat{\beta}_2 x^2$$

- $\hat{\beta}_1$ và $\hat{\beta}_2$ là tác động riêng phần của biến x^1 và x^2 lên biến phụ thuộc, *trong điều kiện các yếu tố khác không đổi*.
- $\hat{y}$ là giá trị thích hợp (hoặc giá trị dự báo) của biến phụ thuộc với điều kiện x^1 và x^2 cho trước.
- Phần dư là chênh lệch giữa giá trị thực tế và giá trị dự báo của biến phụ thuộc, $\hat{u} = y - \hat{y}$.

Ví dụ 1: Ước lượng các nhân tố ảnh hưởng đến điểm GPA

Sử dụng bộ dữ liệu GPA1.dta. Ước lượng mô hình điểm GPA học đại học *colGPA* với một và hai biến giải thích là điểm GPA cho giai đoạn học trung học *hsGPA* và điểm thành tích *ACT*.

```
. reg colGPA hsGPA
```

Source	SS	df	MS	Number of obs	=	141
				F(1, 139)	=	28.85
Model	3.33506006	1	3.33506006	Prob > F	=	0.0000
Residual	16.0710394	139	.115618988	R-squared	=	0.1719
				Adj R-squared	=	0.1659
Total	19.4060994	140	.138614996	Root MSE	=	.34003

colGPA	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsGPA	.4824346	.0898258	5.37	0.000	.304833	.6600362
_cons	1.415434	.3069376	4.61	0.000	.8085635	2.022304

```
.
. reg colGPA hsGPA ACT
```

Source	SS	df	MS	Number of obs	=	141
				F(2, 138)	=	14.78
Model	3.42365506	2	1.71182753	Prob > F	=	0.0000
Residual	15.9824444	138	.115814814	R-squared	=	0.1764
				Adj R-squared	=	0.1645
Total	19.4060994	140	.138614996	Root MSE	=	.34032

colGPA	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsGPA	.4534559	.0958129	4.73	0.000	.2640047	.6429071
ACT	.009426	.0107772	0.87	0.383	-.0118838	.0307358
_cons	1.286328	.3408221	3.77	0.000	.612419	1.960237

Ví dụ 2: Ước lượng mô hình tiền lương

Sử dụng bộ dữ liệu WAGE1.dta. Ước lượng tác động của số năm đi học *educ*, số năm thâm niên *exper*, số năm kinh nghiệm làm việc hiện tại *tenure* lên tiền lương *lwage*.

```
. reg lwage educ exper tenure
```

Source	SS	df	MS	Number of obs	=	526
				F(3, 522)	=	80.39
Model	46.8741776	3	15.6247259	Prob > F	=	0.0000
Residual	101.455574	522	.194359337	R-squared	=	0.3160
				Adj R-squared	=	0.3121
Total	148.329751	525	.28253286	Root MSE	=	.44086

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.092029	.0073299	12.56	0.000	.0776292	.1064288
exper	.0041211	.0017233	2.39	0.017	.0007357	.0075065
tenure	.0220672	.0030936	7.13	0.000	.0159897	.0281448
_cons	.2843595	.1041904	2.73	0.007	.0796756	.4890435

Ví dụ 3: Ước lượng mô hình tiền lương với tác động phi tuyến của giáo dục

Cũng với mô hình trên, nhưng giả sử số năm đi học có tác động phi tuyến (bình phương) lên thu nhập.

```
. gen educsq = educ^2

. reg lwage educ educsq exper tenure
```

Source	SS	df	MS	Number of obs	=	526
				F(4, 521)	=	65.87
Model	49.8213265	4	12.4553316	Prob > F	=	0.0000
Residual	98.5084249	521	.189075672	R-squared	=	0.3359
				Adj R-squared	=	0.3308
Total	148.329751	525	.28253286	Root MSE	=	.43483

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	-.0316271	.0321443	-0.98	0.326	-.0947755	.0315213
educsq	.0052535	.0013306	3.95	0.000	.0026394	.0078676
exper	.0037126	.0017028	2.18	0.030	.0003673	.0070579
tenure	.0216263	.0030534	7.08	0.000	.0156279	.0276247
_cons	.9776996	.2034733	4.81	0.000	.5779708	1.377428

Tác động biên của học thêm một năm lên thu nhập là (%):

$$\frac{\Delta y}{\Delta educ} \approx \beta_1 + 2\beta_2 \times educ$$

Những vấn đề cần lưu ý với hồi quy đa biến

- Chọn biến số đưa vào mô hình theo tiêu chí gì?
- Hậu quả gì nếu đưa biến không liên quan vào mô hình?
- Hậu quả gì nếu bỏ sót biến quan trọng trong mô hình?
- Hậu quả gì nếu đưa các biến tương quan nhau vào mô hình?

Chọn biến đưa vào mô hình

- R^2 luôn luôn tăng khi đưa thêm biến vào mô hình, kể cả những biến không liên quan.
- Do đó, để tránh lạm dụng đưa quá nhiều biến vào mô hình, sử dụng R^2-điều chỉnh:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

  với n và k là số quan sát và số biến giải thích trong mô hình.
- R^2_{adj} có thể tăng hoặc giảm khi đưa biến mới vào mô hình.

Ví dụ 4: Ước lượng mô hình tiền lương với nhiều biến giải thích

Sử dụng bộ dữ liệu WAGE1.dta. Ước lượng mô hình lần lượt với các biến giải thích là (1) số năm đi học, số năm đi học bình phương, kinh nghiệm; (2) thêm biến màu da, giới tính, và hôn nhân; (3) thêm biến số người phụ thuộc. Kiểm tra R^2 và R^2_{adj} thay đổi như thế nào khi thêm biến.

Regression Results

	Model 1 b/se	Model 2 b/se	Model 3 b/se
educ	-0.0340	-0.0136	-0.0134
	(0.0336)	(0.0314)	(0.0315)
educsq	0.0056***	0.0043**	0.0043**
	(0.0014)	(0.0013)	(0.0013)
exper	0.0098***	0.0072***	0.0072***
	(0.0015)	(0.0015)	(0.0016)
nonwhite		-0.0089	-0.0094
		(0.0609)	(0.0612)
female		-0.3097***	-0.3098***
		(0.0377)	(0.0378)
married		0.1404***	0.1394**
		(0.0409)	(0.0422)
numdep			0.0016
			(0.0156)
Constant	0.9571***	1.0261***	1.0219***
	(0.2128)	(0.1987)	(0.2031)
Obs	526.0000	526.0000	526.0000
R2	0.2719	0.3792	0.3793
R2-adj	0.2678	0.3721	0.3709
df(r)	522.0000	519.0000	518.0000
SSR	107.9936	92.0757	92.0739

* p<0.05, ** p<0.01, *** p<0.001

Sử dụng hệ số phóng đại phương sai (Variance Inflation Factor) để lựa chọn biến

Hệ số *VIF* dùng để kiểm tra mức độ tương quan của một biến giải thích với các biến còn lại. Biến số càng ít tương quan với các biến khác càng tốt.

- Hồi quy lần lượt biến x^j lên các biến còn lại. Tính hệ số thích hợp R_j^2.
- Tính hệ số *VIF*:

$$VIF_j = \frac{1}{1 - R_j^2}$$

- Nếu R_j^2 lớn chứng tỏ biến x^j tương quan nhiều với các biến giải thích khác.
- Quy tắc chung: Loại biến có $VIF > 10$

Ví dụ 5: Chọn biến sử dụng hệ số *VIF*

Ước lượng lại ví dụ (4), tính VIF và giải thích.

- Nếu có một biến cộng tuyến hoàn hảo trong mô hình thì VIF của biến đó là bao nhiêu?

Đưa biến không liên quan vào mô hình

- Giả sử mô hình chuẩn là $\mathbf{Y} = \tilde{\beta}_\mathbf{0} + \tilde{\beta}_\mathbf{1}\mathbf{x^1}$, nhưng chúng ta ước lượng mô hình $\mathbf{Y} = \hat{\beta}_\mathbf{0} + \hat{\beta}_\mathbf{1}\mathbf{x^1} + \hat{\beta}_\mathbf{2}\mathbf{x^2}$.
- Mối quan hệ giữa $\tilde{\beta}_1$ và $\hat{\beta}_1$ là:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\sigma_1}$$

  Với $\tilde{\sigma_1}$ là hệ số góc của hàm hồi quy của biến x^2 lên biến x^1.
- Nếu biến x^2 không quan trọng, $\hat{\beta}_2 = 0$, do đó $\tilde{\beta}_1$ vẫn không chệch, $\tilde{\beta}_1 = \hat{\beta}_1$.
- Phương sai của các ước lượng sẽ thay đổi!

Thiếu biến quan trọng trong mô hình

- Giả sử mô hình chuẩn là $\mathbf{Y} = \hat{\beta}_\mathbf{0} + \hat{\beta}_\mathbf{1}\mathbf{x^1} + \hat{\beta}_\mathbf{2}\mathbf{x^2}$, nhưng chúng ta ước lượng mô hình $\mathbf{Y} = \tilde{\beta}_\mathbf{0} + \tilde{\beta}_\mathbf{1}\mathbf{x^1}$.
- Từ công thức:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\sigma}_1$$

- Mức độ chệch của ước lượng khi xảy ra vấn đề thiếu biến quan trọng là:

$$Bias(\tilde{\beta}_1) = \tilde{\beta}_1 - \hat{\beta}_1 = \hat{\beta}_2\tilde{\sigma}_1$$

Đánh giá hướng chệch trong mô hình thiếu biến quan trọng

- Nếu $\beta_2 = 0$ (nghĩa là biến x_2 không phải là biến quan trọng) thì ước lượng của $\tilde{\beta}_1$ không chệch.
- Nếu $\tilde{\sigma_1} = 0$ (nghĩa là x_1 và x_2 không tương quan) thì $\tilde{\beta}_1$ cũng không chệch.
- Nếu không phải 2 trường hợp trên, $\tilde{\beta}_1$ chệch, với hướng và mức độ chệch tùy thuộc vào giá trị của β_2 và $\tilde{\sigma_1}$.

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Nếu nghi ngờ mô hình thiếu biến thì khi giải thích kết quả phải nhận định hướng chệch của tác động!

Ví dụ 6: Ước lượng phương trình tiền lương theo số năm đi học

Sử dụng bộ dữ liệu WAGE1.dta

- Giả sử mô hình chuẩn có hai biến là giáo dục (educ) và tố chất cá nhân (ability):

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

- Chúng ta không quan sát được tố chất cá nhân, do đó chúng ta chỉ ước lượng được mô hình:

$$log(wage) = \beta_0 + \beta_1 educ + u$$

- Ước lượng của β_1 có bị chệch không? và chệch theo hướng nào?

```
. reg lwage educ
```

Source	SS	df	MS	Number of obs	=	526
				F(1, 524)	=	119.58
Model	27.5606288	1	27.5606288	Prob > F	=	0.0000
Residual	120.769123	524	.230475425	R-squared	=	0.1858
				Adj R-squared	=	0.1843
Total	148.329751	525	.28253286	Root MSE	=	.48008

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0827444	.0075667	10.94	0.000	.0678796	.0976091
_cons	.5837727	.0973358	6.00	0.000	.3925563	.7749891

- Tỷ suất thu nhập của một năm đi học ước lượng được là 8.3%.

Mô hình thiếu biến quan trọng trong trường hợp tổng quát

- Mô hình tổng quát với nhiều biến giải thích:

$$Y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + ... + \beta_k x^k + u$$

- Nếu thiếu một biến quan trọng nào đó, tất cả các ước lượng $\hat{\beta}$ đều bị chệch.
- Xác định hướng chệch khó hơn nhiều do tương quan giữa các biến giải thích với biến bị thiếu, và giữa các biến giải thích với nhau.

Tóm tắt các giả định đối với hồi quy đa biến

Tương tự như các điều kiện của hồi quy đơn biến:

1. Tuyến tính theo tham số.
2. Chọn mẫu ngẫu nhiên.
3. **Không có cộng tuyến hoàn hảo.**
4. Trung bình có điều kiện của sai số bằng 0:

$$E(u|x^1, ..., x^k) = 0$$

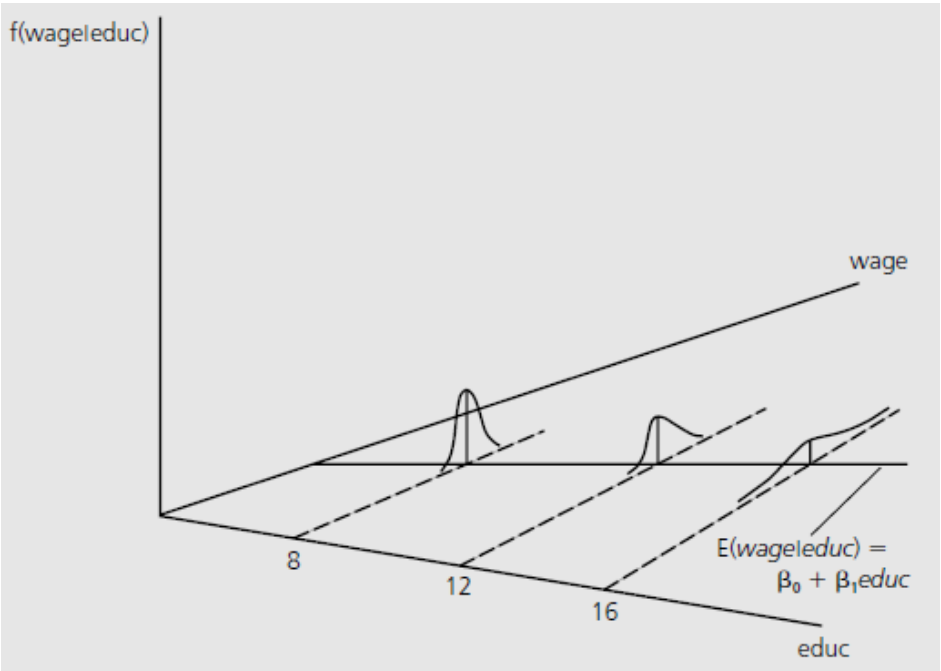
$\Rightarrow$ Ước lượng OLS của các tham số β là không chệch.

$$E(\hat{\beta}) = \beta$$

Giả định phương sai của sai số không đổi (homoskedasticity)

5. Với các giá trị của các biến giải thích cho trước, phương sai của sai số là một hằng số:

$$Var(u|x_1, ..., x_k) = \sigma^2$$



Đặc tính của ước lượng OLS

- Với các giả định 1-5, ước lượng của OLS là ước lượng tuyến tính, không chệch, và hiệu quả nhất (Best Linear Unbiased Estimator - BLUE)
  - Trong tất cả các ước lượng tuyến tính, OLS có phương sai của ước lượng là nhỏ nhất.
  - Không chệch.