HỒI QUI
TUYẾN TÍNH ĐA BIẾN

HỒI QUI TUYẾN TÍNH ĐƠN

- Mô hồi qui tuyến tính đa biến
- Phương pháp bình phương tối thiểu
- Hệ số xác định của hồi qui đa biến
- Các giả định của mô hình
- Kiểm định mức ý nghĩa
- Sử dụng mô hình hồi qui ước lượng để ước lượng và dự đoán
- Biến độc lập định tính

MÔ HÌNH
HỒI QUI TUYẾN TÍNH ĐA BIẾN

- Mô hình hồi qui tuyến tính đa biến là phương trình mô tả mối quan hệ giữa biến phụ thuộc y với các biến độc lập $x_1, x_2, \ldots x_p$ và số hạng sai số ε

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

Với:

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ là các tham số, và
ε là biến ngẫu nhiên gọi là số hạng sai số

PHƯƠNG TRÌNH HỒI QUI TUYẾN TÍNH ĐA BIẾN

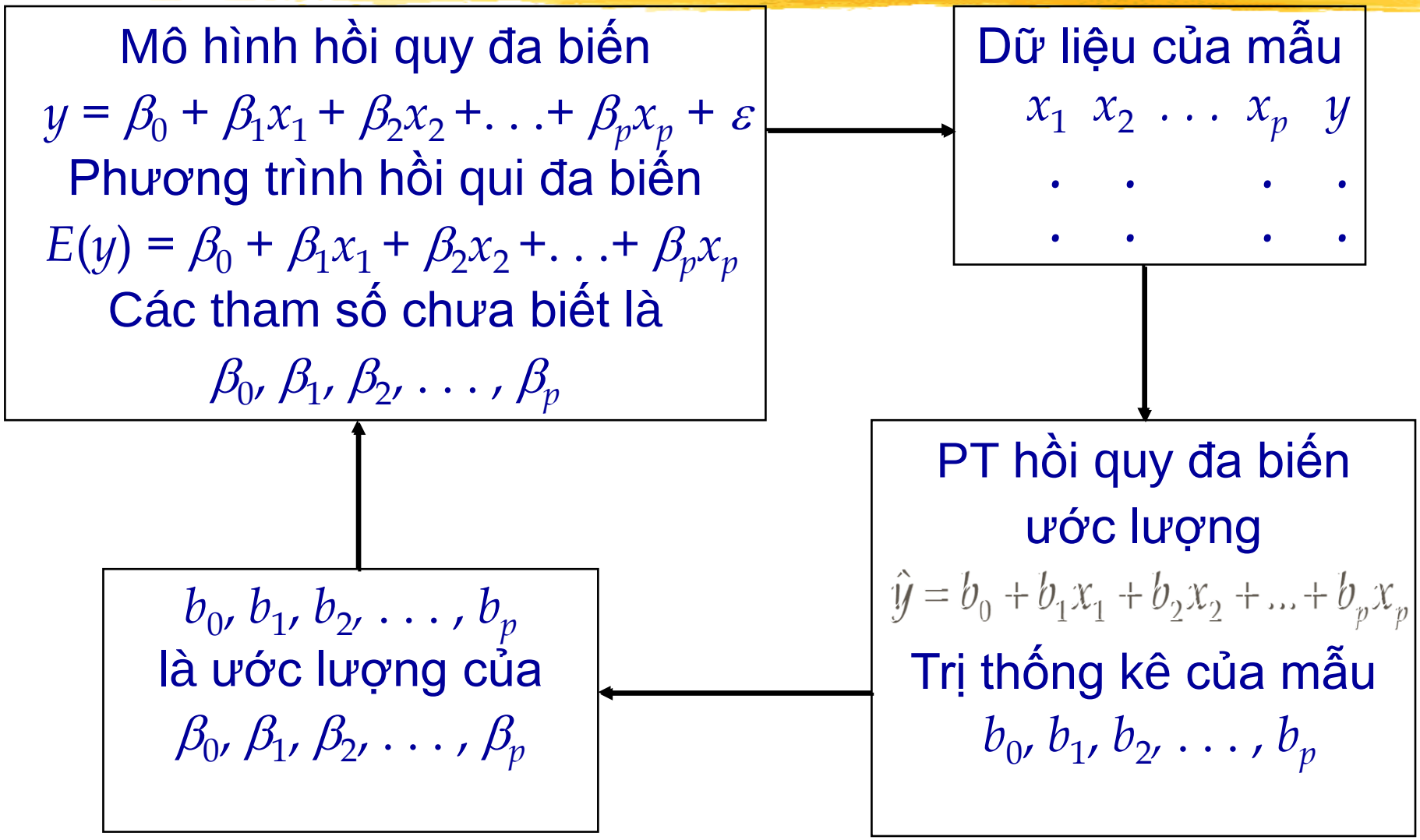
- Phương trình hồi qui tuyến tính đa biến là phương trình mô tả mối quan hệ giữa biến phụ thuộc y với các biến độc lập $x_1, x_2, \ldots x_p$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

Với:

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ là các tham số, và
ε là biến ngẫu nhiên gọi là số hạng sai số

QUI TRÌNH ƯỚC LƯỢNG



PHƯƠNG PHÁP BÌNH PHƯƠNG TỐI THIỂU

- Tiêu chí bình phương tối thiểu

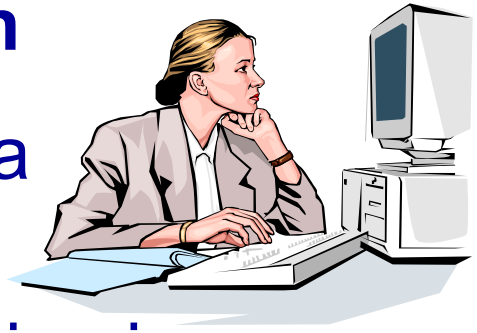
$$\min \sum (y_i - \hat{y}_i)^2$$

- Tính toán các giá trị của hệ số hồi qui

  Các công thức tính toán các hệ số hồi qui b_0, b_1, b_2, … b_p liên quan đến việc sử dụng đại số tuyến tính. Các phần mềm thống kê sẽ thực hiện việc tính toán này.

MÔ HÌNH
HỒI QUI TUYẾN TÍNH ĐA BIẾN

- **Ví dụ: Khảo sát lương lập trình viên**



Một Cty phần mềm thu thập dữ liệu của một mẫu gồm 20 lập trình viên.
Người ta đề nghị sử dụng phân tích hồi qui
Để xác định xem lương có mối liên hệ với số năm kinh nghiệm và điểm thi năng khiếu về lập trình do cty tổ chức hay không?
Số năm kinh nghiệm, điểm thi năng khiếu
Và mức lương hàng năm ($1000s) của 20 lập trình viên được trình bày ở bảng sau:

MÔ HÌNH
HỒI QUI TUYẾN TÍNH ĐA BIẾN

Exper.	Score	Salary	Exper.	Score	Salary
4	78	24.0	9	88	38.0
7	100	43.0	2	73	26.6
1	86	23.7	10	75	36.2
5	82	34.3	5	81	31.6
8	86	35.8	6	74	29.0
10	84	38.0	8	87	34.0
0	75	22.2	4	79	30.1
1	80	23.1	6	94	33.9
6	83	30.0	3	70	28.2
6	91	33.0	3	89	30.0

MÔ HÌNH
HỒI QUI TUYẾN TÍNH ĐA BIẾN

Giả sử chúng ta tin rằng lương hàng năm (y) có mối liên hệ với số năm kinh nghiệm (x_1) và điểm thi năng khiếu (x_2) theo mô hình hồi qui sau:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Với

y = Lương hàng năm($1000)

x_1 = Số năm kinh nghiệm

x_2 = Điểm thi năng khiếu

MÔ HÌNH
HỒI QUI TUYẾN TÍNH ĐA BIẾN

Dữ liệu

x_1	x_2	y
4	78	24
7	100	43
.	.	.
.	.	.
3	89	30

→ **Sử dụng Phần mềm Để giải Hồi qui Tuyến tính Đa biến** →

Kết quả

$b_0 =$
$b_1 =$
$b_2 =$
$R^2 =$

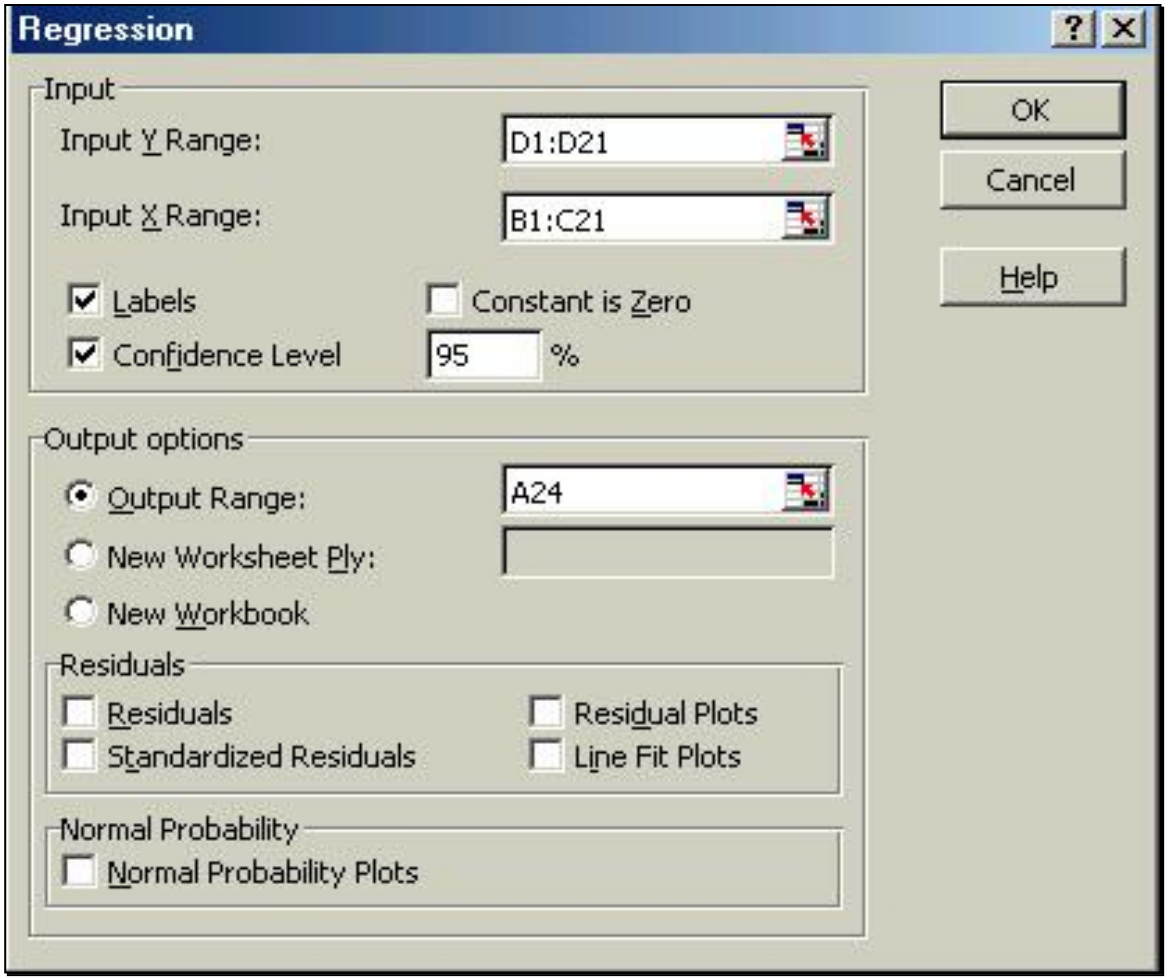
ƯỚC LƯỢNG $\beta_0, \beta_1, \beta_2$

Bảng số liệu trên Excel

	A	B	C	D
1	Programmer	Experience (yrs)	Test Score	Salary ($K)
2	1	4	78	24.0
3	2	7	100	43.0
4	3	1	86	23.7
5	4	5	82	34.3
6	5	8	86	35.8
7	6	10	84	38.0
8	7	0	75	22.2
9	8	1	80	23.1

ƯỚC LƯỢNG $\beta_0, \beta_1, \beta_2$

Hộp thoại hồi qui trên Excel



ƯỚC LƯỢNG $\beta_0, \beta_1, \beta_2$

Kết quả hồi qui trên Excel

	A	B	C	D	E
38					
39		*Coeffic.*	*Std. Err.*	*t Stat*	*P-value*
40	Intercept	3.17394	6.15607	0.5156	0.61279
41	Experience	1.4039	0.19857	7.0702	1.9E-06
42	Test Score	0.25089	0.07735	3.2433	0.00478
43					

PHƯƠNG TRÌNH HỒI QUI ƯỚC LƯỢNG

SALARY = 3.174 + 1.404(EXPER) + 0.251(SCORE)

GIẢI THÍCH CÁC HỆ SỐ HỒI QUI

Trong ohân tích hồi qui đa biến, Mỗi hệ số hồi qui được giải thích như sau:

$\boldsymbol{b_i}$ là một ước lượng cho sự thay đổi của y ứng với sự gia tăng 1 đơn vị của x_i khi tất cả các biến độc lập được giữ không đổi.

GIẢI THÍCH CÁC HỆ SỐ HỒI QUI

$b_1 = 1.404$

Lương được kỳ vọng tăng $1,404 đối với mỗi 1 năm kinh nghiệm tăng thêm (khi điểm năng khiếu được giữ không đổi).

$b_2 = 0.251$

Lương được kỳ vọng tăng $251 đối với mỗi 1 năm kinh nghiệm tăng thêm (khi số năm kinh nghiệm được giữ không đổi).

HỆ SỐ XÁC ĐỊNH

Mối liên hệ giữa SST, SSR, SSE

$$SST = SSR + SSE$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where:

SST = Tổng bình phương toàn phần

SSR = Tổng bình phương hồi qui

SSE = Tổng bình phương sai số

HỆ SỐ XÁC ĐỊNH

Kết quả hồi qui trên Excel

	A	B	C	D	E	F
32						
33	ANOVA					
34		*df*	*SS*	*MS*	*F*	*Significance F*
35	Regression	2	500.3285	250.1643	42.76013	2.32774E-07
36	Residual	17	99.45697	5.85041		
37	Total	19	599.7855			
38						

SST

SSR

HỆ SỐ XÁC ĐỊNH

$R^2 = SSR/SST$

$R^2 = 500.3285/599.7855 = .83418$

HỆ SỐ XÁC ĐỊNH ĐIỀU CHỈNH

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

$$R_a^2 = 1 - (1 - .834179)\frac{20-1}{20-2-1} = .814671$$

HỆ SỐ XÁC ĐỊNH

Kết quả hồi qui trên Excel

	A	B	C
23			
24	SUMMARY OUTPUT		
25			
26	*Regression Statistics*		
27	Multiple R	0.913334059	
28	R Square	0.834179103	
29	Adjusted R Square	0.814670762	
30	Standard Error	2.418762076	
31	Observations	20	
32			

CÁC GIẢ ĐỊNH VỀ SỐ HẠNG SAI SỐ ε

1. Sai số ε là biến ngẫu nhiên với trung bình bằng 0

2. Phương sai của ε, ký hiệu σ^2, sẽ giống nhau đối với tất cả các giá trị của biến độc lập.

3. Các giá trị của ε là độc lập.

4. Sai số ε là biến ngẫu nhiên tuân theo phân phối chuẩn phản ánh sự biến động của giá trị y và giá trị kỳ vọng của y được xác định bởi $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$.

KIỂM ĐỊNH Ý NGHĨA

- Trong hồi qui tuyến tính đơn biến, kiểm định *F* và *t* *cho cùng kết luận*

- Trong hồi qui tuyến tính đa biến, kiểm định *F* và *t* *có các mục đích khác nhau*

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *F*

- Kiểm định *F* được dùng để xác định có tồn tại mối liên hệ có ý nghĩa giữa biến phụ thuộc và toàn bộ các biến độc lập

- Kiểm định *F* được xem như kiểm định ý nghĩa tổng thể

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *t*

- Nếu kiểm định *F* được xem như kiểm định ý nghĩa tổng thể, thì kiểm định *t* được dùng để xác định xem từng biến độc lập riêng có ý nghĩa hay không

- Kiểm định *t* được thực hiện riêng cho mỗi biến độc lập trong mô hình

- Kiểm định *t* được xem như kiểm định ý nghĩa riêng lẻ

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *F*

Giả thuyết

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_a: Có ít nhất 1 tham số β_i khác 0

Trị kiểm định

$F = MSR/MSE$

Qui tắc bác bỏ

Bác bỏ H_0 nếu p-value $\leq \alpha$ hay nếu $F > F_\alpha$,

Với F_α lấy từ bảng phân phối F

Bậc tự do trên tử số là p và bậc tự do dưới mẫu số là $n - p - 1$.

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *F*

Giả thuyết

H_0: $\beta_1 = \beta_2 = 0$

H_a: Có ít nhất 1 tham số β_i khác 0

Qui tắc bác bỏ

Với $\alpha = 5\%$ và Bậc tự do là 2 và 17

Tra bảng $F_{.05} = 3.59$

Bác bỏ H_0 nếu p-value $\leq .05$ hay $F \geq 3.59$

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *F*

Kết quả hồi qui trên Excel

	A	B	C	D	E	F
32						
33	ANOVA					
34		*df*	*SS*	*MS*	*F*	*Significance F*
35	Regression	2	500.3285	250.1643	42.76013	2.32774E-07
36	Residual	17	99.45697	5.85041		
37	Total	19	599.7855			
38						

p-value được dùng để kiểm định ý nghĩa tổng thể

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *F*

Trị kiểm định

F = MSR/MSE
= 250.16/5.85 = 42.76

Kết luận

p-value $\leq .05$, vì vậy có thể bác bỏ H_0.
(cũng vậy, $F = 42.76 \geq 3.59$)

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *t*

Giả thuyết

H_0: $\beta_i = 0$

H_a: β_i khác 0

Trị kiểm định

$t = b_i / S_{bi}$

Qui tắc bác bỏ

Bác bỏ H_0 nếu p-value $\leq \alpha$ hay

nếu $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$ với $t_{\alpha/2}$

Được lấy từ bảng phân phối t

Với bậc tự do là $n - p - 1$

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *t*

Kết quả hồi qui trên Excel

	A	B	C	D	E
38					
39		*Coeffic.*	*Std. Err.*	*t Stat*	*P-value*
40	Intercept	3.17394	6.15607	0.5156	0.61279
41	Experience	1.4039	0.19857	7.0702	1.9E-06
42	Test Score	0.25089	0.07735	3.2433	0.00478
43					

Trị thống kê t và *p*-value được dùng để kiểm định ý nghĩa riêng của biến"Experience"

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *t*

Kết quả hồi qui trên Excel

	A	B	C	D	E
38					
39		*Coeffic.*	*Std. Err.*	*t Stat*	*P-value*
40	Intercept	3.17394	6.15607	0.5156	0.61279
41	Experience	1.4039	0.19857	7.0702	1.9E-06
42	Test Score	0.25089	0.07735	3.2433	0.00478
43					

Trị thống kê t và *p*-value được dùng để kiểm định ý nghĩa riêng của biến"Test Score"

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *t*

Giả thuyết

H_0: $\beta_i = 0$

H_a: β_i khác 0

Qui tắc bác bỏ

Với $\alpha = .05$ và bậc tự do = 17, $t_{.025} = 2.11$

Bác bỏ H_0 nếu p-value $\leq .05$ hay $t \geq 2.11$

KIỂM ĐỊNH Ý NGHĨA: KIỂM ĐỊNH *t*

Trị kiểm định

$t = b_1/S_{b1} = 1.4039/0.1986 = 7.07$

$t = b_2/S_{b2} = 0.25089/0.07735 = 3.24$

Kết luận

Bác bỏ cả $H_0: \beta_1 = 0$ và $H_0: \beta_2 = 0$.

Cả hai biến độc lập đều có ý nghĩa

KIỂM ĐỊNH Ý NGHĨA: ĐA CỘNG TUYẾN

Thuật ngữ đa cộng tuyến liên quan đến sự tương quan giữa các biến độc lập. Đa cộng tuyến thường xảy ra khi các biến độc lập có tương quan mạnh **(|*r* | > .7)**

Hậu quả của ĐCT:

- Khi có ĐCT hoàn hảo **(|*r* | = 1)**
  Chúng ta không thể ước lượng được mô hình
- Sai số chuẩn của các hệ số sẽ lớn $\mathbf{S_{bi}}$
- R^2 rất cao cho dù thống kê t ít ý nghĩa
- Các ước lượng sẽ không chính xác
- Dấu vài hệ số sẽ khác với kỳ vọng

KIỂM ĐỊNH Ý NGHĨA: ĐA CỘNG TUYẾN

- Qui trình ước lượng y trong hồi qui đa biến cũng tương tư như trong hồi qui đơn biến.

- Chúng ta thay thế các biến $x_1, x_2, \ldots, x_p$ vào phương trình hồi qui ước lượng thay vì chỉ sử dụng 1 biến độc lập x trong hồi qui đơn biến.

KIỂM ĐỊNH Ý NGHĨA: ĐA CỘNG TUYẾN

- Nếu phương trình hồi qui ước lượng được dùng cho mục đích dự báo thì ĐCT không gây ra vấn đề nghiêm trọng gì.

- Để hạn chế ĐCT, ta không đưa các biến độc lập có tương quan mạnh vào phương trình hồi qui đa biến.

BIẾN ĐỘC LẬP ĐỊNH TÍNH

- Trong nhiều tình huống thực tiễn chúng ta phải sử dụng các biến định tính như giới tính (Nam, Nữ); Vùng miền (Bắc, Trung, Nam)

- Ví dụ, x_2 có thể đại diện cho giới tính với $x_2 = 0$ để chỉ Nam và $x_2 = 1$ để chỉ Nữ.

- Trong trường hợp này x_2 được gọi là biến giả, biến chỉ thị hay biến thuộc tính.

BIẾN ĐỘC LẬP ĐỊNH TÍNH

- **Ví dụ: Khảo sát lương lập trình viên**

  - Như một sự mở rộng vấn đề khảo sát lương lập trình viên.

Giả sử về mặt quản lý, người ta tin rằng lương hàng năm có liên quan đến cá nhân có bằng tốt nghiệp về khoa học máy tính hay hệ thống thông tin.

Dữ liệu về Số năm kinh nghiệm, Điểm thi năng khiếu, Bằng cấp chuyên môn và lương hàng năm ($1000) của mẫu gồm 20 lập trình viên được trình bày như sau:

BIẾN ĐỘC LẬP ĐỊNH TÍNH

Exper.	Score	Degr.	Salary	Exper.	Score	Degr.	Salary
4	78	No	24.0	9	88	Yes	38.0
7	100	Yes	43.0	2	73	No	26.6
1	86	No	23.7	10	75	Yes	36.2
5	82	Yes	34.3	5	81	No	31.6
8	86	Yes	35.8	6	74	No	29.0
10	84	Yes	38.0	8	87	Yes	34.0
0	75	No	22.2	4	79	No	30.1
1	80	No	23.1	6	94	Yes	33.9
6	83	No	30.0	3	70	No	28.2
6	91	Yes	33.0	3	89	No	30.0

ƯỚC LƯỢNG PHƯƠNG TRÌNH HỒI QUY

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

Với:

y = Lương hàng năm ($1000)

x_1 = Số năm kinh nghiệm

x_2 = Điểm thi năng khiếu

x_3 = 0 nếu <u>không</u> có bằng cấp chuyên môn

1 nếu <u>có</u> bằng cấp chuyên môn

x_3 là biến giả

BIẾN ĐỘC LẬP ĐỊNH TÍNH

Kết quả hồi qui trên Excel

	A	B	C
23			
24	SUMMARY OUTPUT		
25			
26	*Regression Statistics*		
27	Multiple R	0.920215239	
28	R Square	0.846796085	
29	Adjusted R Square	0.818070351	
30	Standard Error	2.396475101	
31	Observations	20	
32			

BIẾN ĐỘC LẬP ĐỊNH TÍNH

Kết quả hồi qui trên Excel

	A	B	C	D	E	F
32						
33	ANOVA					
34		*df*	*SS*	*MS*	*F*	*Significance F*
35	Regression	3	507.896	169.2987	29.47866	9.41675E-07
36	Residual	16	91.88949	5.743093		
37	Total	19	599.7855			
38						

BIẾN ĐỘC LẬP ĐỊNH TÍNH

Kết quả hồi qui trên Excel

	A	B	C	D	E
38					
39		*Coeffic.*	*Std. Err.*	*t Stat*	*P-value*
40	Intercept	7.94485	7.3808	1.0764	0.2977
41	Experience	1.14758	0.2976	3.8561	0.0014
42	Test Score	0.19694	0.0899	2.1905	0.04364
43	Grad. Degr.	2.28042	1.98661	1.1479	0.26789
44					

Không có ý nghĩa

BIẾN ĐỘC LẬP ĐỊNH TÍNH

- Nếu biến định tính có k thuộc tính thì sẽ sử dụng k – 1 biến giả. Mỗi biến giả sẽ được mã hóa là 0 và 1.

- Ví dụ, một biến định tính có 3 thuộc tính A, B và C có thể được đại diện bằng 2 biến x_1 và x_2 với các giá trị (0, 0) cho A, (1, 0) cho B, and (0,1) cho C.

- Lưu ý: Phải cẩn thận trong việc định nghĩa và giải thích biến giả

BIẾN ĐỘC LẬP ĐỊNH TÍNH

- Ví dụ, một biến định tính về trình độ học vấn có thể được trình bày bằng biến x_1 và x_2 với các giá trị như sau:

Bằng cấp Cao nhất	x_1	x_2
Cử nhân	0	0
Thạc sĩ	1	0
Tiến sĩ	0	1

More Complex Qualitative Variables

For example, a variable indicating level of education could be represented by x_1 and x_2 values as follows:

Highest Degree	x_1	x_2
Bachelor's	0	0
Master's	1	0
Ph.D.	0	1