PHÂN TÍCH HỒI QUI TUYẾN TÍNH ĐA BIẾN

Mô hình hồi qui tuyến tính đơn biến dùng để xem xét mối quan hệ tuyến tính giữa biến phụ thuộc y (biến kết cục) và biến độc lập x (biến dự đoán). Phương trình tuyến tính (đường thẳng) đơn biến có dạng:

$$y= \alpha + \beta x_i+\varepsilon_i$$

Trong đó α là điểm cắt trên trục tung, β là độ dốc (trong thống kê gọi là hệ số hồi qui) và ε là phần dư. Ví dụ tìm sự liện hệ giữa lượng cholesterol máu (x) và bề dày lớp nội trung mạc (NTM) động mạch cảnh (y).

Tuy nhiên trong y sinh học, thường có rất nhiều yếu tố khác nhau dẫn đến một hiện tượng. Ví dụ như hiện tượng dày lớp NTM động mạch cảnh không chỉ do lượng cholesterol máu mà còn do nhiều yếu tố khác như di truyền, chủng tộc, mắc bịnh tim mạch, tuổi, giới, BMI, tăng huyết áp, đái tháo đường....Vì vậy cần phải có mô hình hồi qui tuyến tính đa biến. Trong ví dụ này ta không đề cập các yếu tố di truyền, chủng tộc, giới, mắc bịnh tim mạch... mà chỉ lưu ý đến các biến số như: tuổi, cholesterol, glucose, huyết áp tâm thu và BMI.

Ví dụ: dữ liệu đo bề dày lớp NTM động mạch cảnh trên siêu âm ở 100 bệnh nhân có bệnh tim mạch như sau:

Dữ liệu này có 7 cột: ID bệnh nhân, tuổi (năm), BMI (kg/m2), huyết áp tâm thu (mmHg), glucose máu (mmol/L), cholesterol (mmol/L), bề dày lớp NTM trên siêu âm (mm). *(Ghi chú: có thể copy dữ liệu này vào Excel, sau đó chuyển vào SPSS để thực tập)*

Bảng 1. Dữ liệu bề dày NTM của 100 bệnh nhân

ID	TUOI	BMI	HA	GLUCOSE	CHOLESTEROL	BEDAYNTM
1	56	21	160	14.0	6.00	1.95
2	76	18	150	12.0	4.97	1.33
3	63	16	160	4.4	6.39	.83
4	78	20	100	4.0	7.00	2.00
5	87	20	110	4.6	4.10	1.30
6	76	19	150	4.6	2.74	1.16
7	55	31	160	5.5	4.60	1.00
8	74	22	100	6.8	5.04	1.00
9	81	21	120	5.8	4.75	.80
10	77	24	160	5.4	6.94	1.60
11	29	20	120	3.8	4.84	.65
12	71	22	160	3.3	6.63	1.00
13	77	21	160	5.1	4.93	.97
14	59	18	150	6.0	4.55	.73
15	58	27	130	6.9	6.70	1.10
16	34	19	130	4.5	3.20	1.10
17	74	22	100	10.6	4.30	1.10
18	61	19	170	18.0	6.80	.80
19	53	20	130	25.0	5.50	.99
20	65	28	140	6.5	6.80	1.00
21	80	19	160	4.8	5.74	1.13
22	71	25	160	6.2	6.90	1.00
23	90	24	160	4.7	7.00	1.70

24	44	24	120	6.0	3.40	.90
25	91	27	150	6.1	4.92	.89
26	75	22	160	6.2	6.08	.80
27	60	24	140	4.7	6.25	.81
28	51	22	150	4.8	5.40	1.20
29	91	29	120	4.2	6.54	.82
30	45	24	170	4.9	3.91	.89
31	62	24	140	5.4	5.30	1.19
32	65	19	150	12.0	2.60	.97
33	70	22	160	3.6	6.85	.97
34	56	27	150	5.7	3.75	.97
35	51	19	120	4.7	5.84	.88
36	75	18	140	10.1	6.91	.97
37	58	32	160	4.7	5.01	.90
38	61	19	160	5.2	4.00	.89
39	72	18	120	22.2	4.88	.80
40	82	25	90	8.2	4.20	1.13
41	95	24	120	11.0	4.47	1.20
42	56	21	160	4.9	6.90	.90
43	36	28	130	4.5	4.71	.81
44	67	18	100	5.5	5.70	.80
45	64	22	130	6.2	3.00	.74
46	81	22	140	5.0	5.06	2.66
47	46	22	160	3.3	4.61	.89
48	56	22	150	4.1	4.15	.79
49	60	24	140	7.1	5.30	.80
50	35	19	120	7.4	4.10	.56
51	55	21	160	5.4	3.00	.80
52	70	20	150	6.2	2.57	1.20
53	64	23	130	5.7	6.78	.82
54	64	19	160	5.9	5.62	.90
55	58	27	160	26.0	8.07	1.00
56	73	23	140	5.6	3.00	1.15
57	41	24	110	10.0	3.31	1.16
58	74	23	100	5.3	4.73	.97
59	21	23	160	5.0	4.00	.80
60	67	25	150	3.5	3.60	1.67
61	57	23	140	6.4	5.30	1.06
62	69	21	120	7.6	6.00	1.10
63	53	34	140	8.1	6.49	.80
64	58	23	160	9.0	7.00	1.70
65	54	29	130	6.4	7.48	.99
66	49	17	130	6.3	5.19	1.16
67	59	22	140	7.0	3.00	.62
68	65	23	150	5.9	6.70	1.00
69	42	22	150	3.9	7.00	.82
70	75	24	100	6.4	6.60	1.00
71	72	21	140	11.0	5.75	1.70
72	82	24	190	18.0	4.70	2.30
73	70	18	160	3.3	4.61	.89
74	42	24	160	6.0	6.30	.97
75	32	19	140	4.0	2.00	.70
76	61	21	140	5.2	2.50	1.10
77	60	26	130	11.3	4.79	1.01
78	76	19	160	5.1	5.31	1.15
79	78	27	120	4.9	3.80	.92
80	71	26	150	6.6	7.13	1.10
81	49	24	140	4.3	5.50	.80
82	36	23	140	4.3	4.20	.70
83	74	21	140	17.0	3.30	1.00
84	53	21	140	5.6	5.90	.80
85	56	19	140	4.1	4.73	.89
86	60	20	160	4.9	3.00	.60
87	83	21	120	7.9	5.88	1.50
88	68	23	130	4.0	5.39	.70
89	69	19	100	4.4	6.15	1.10
90	31	21	120	4.1	3.94	.81
91	34	21	140	6.7	3.83	.70
92	41	20	120	2.7	4.93	.71
93	72	22	160	6.4	7.00	2.70
94	54	22	170	6.2	8.18	1.13
95	54	28	150	4.2	8.16	1.70
96	55	24	160	5.0	7.20	.90
97	76	15	140	3.1	5.24	1.16
98	70	25	180	4.0	4.40	1.00
99	85	21	160	5.2	5.20	.97

100	87	22	130	9.0	5.20	2.30	

Mô hình hồi qui tuyến tính đa biến có dạng:

$$y= \alpha + \beta_1 x_{1i}+ \beta_2 x_{2i} +\beta_3 x_{3i} + ...+.....+ \beta_k x_{ki} + \varepsilon_i$$

Với ví dụ trên ta có thể viết phương trình hồi qui tuyến tính đa biến với 5 yếu tố (x_1=tuổi, x_2=cholesterol, x_3=glucose, x_4=huyết áp TT và x_5= BMI) như sau:

Bề dày NTM= α + β_1(tuổi) + β_2(cholesterol) + β_3(glucose) + β_4(huyết áp TT) + β_5 (BMI) + ...+ ε

Trong phân tích hồi qui tuyến tính đa biến, ta cần biết mức độ ảnh hưởng của từng yếu tố lên biến kết cục y (bề dày lớp NTM trong ví dụ này). Muốn biết mức độ ảnh hưởng cần lưu ý đến các trị số sau:

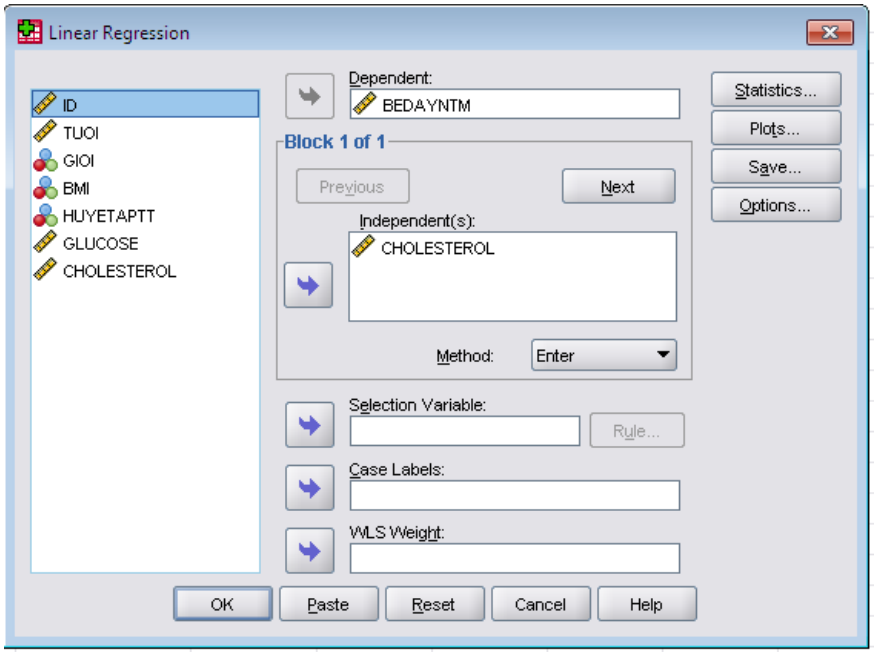
1. Hệ số tương quan R (coefficient of correlation): yếu tố nào có R càng lớn thì ảnh hưởng càng nhiều
2. Bình phương của R (R square): yếu tố nào có R^2 càng lớn thì mối quan hệ giữa yếu tố đó và biến y càng chặt chẽ.
3. Hệ số hồi qui β (regression coefficient): yếu tố nào có β cao thì ảnh hưởng nhiều hơn, tuy nhiên các yếu tố có đơn vị khác nhau (tuổi, mmol/L, mmHg....) nên không thể so sánh mức ảnh hưởng giữa các yếu tố. Nếu muốn so sánh phải đổi các yếu tố có cùng đơn vị là độ lệch chuẩn, lúc đó ta có hệ số hồi qui chuẩn hóa: $\beta_S= \beta. \frac{Sx}{Sy}$ (Với Sx là độ lệch chuẩn của x tương ứng và Sy là độ lệch chuẩn của y). Dựa vào công thức trên, ta tính được hệ số hồi qui chuẩn của tuổi (0,42), cholesterol (0,22), glucose (0,14), huyết áp (0,10) và BMI (0,02)
4. Trị số p (p value): càng nhỏ mức ảnh hưởng càng mạnh.

Vào SPSS, phân tích từng đơn biến để xem các trị số này:

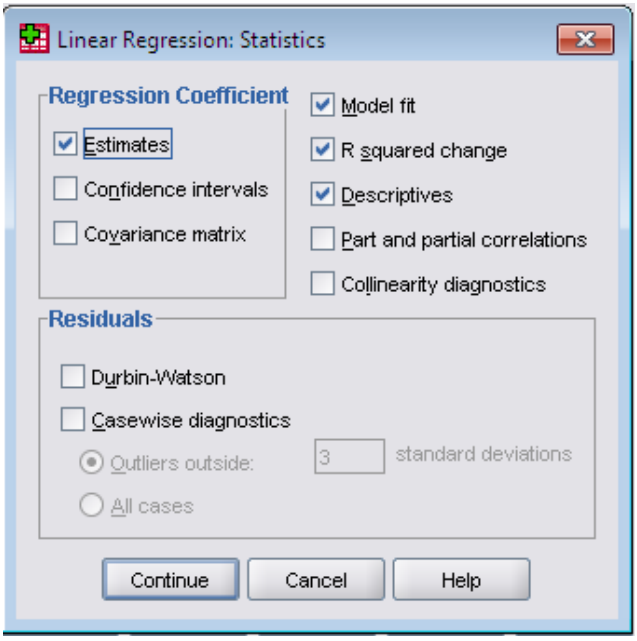
Phân tích hồi qui tuyến tính đơn biến trong SPSS.

Vào Analyse> Regression > Linear

Mở hộp thoại Linear Regression. Nhắp chuyển BEDAYNTM vào ô Dependent (trục tung) và cholesterol vào ô Independent (trục hoành).



Nhấn vào nút Statistics, đánh dấu nháy vào các ô: Estimates, Model fit, R squared change và Descriptives.



Nhấn nút Continue, sau đó nhấn OK

Kết quả như sau:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change		
					R Square Change	F Change	
1	.219[a]	.048	.038	.39606	.048	4.919	

a. Predictors: (Constant), CHOLESTEROL

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.772	1	.772	4.919	.029[a]
	Residual	15.373	98	.157		
	Total	16.144	99			

a. Predictors: (Constant), CHOLESTEROL

b. Dependent Variable: BEDAYNTM

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.748	.151		4.954	.000
	CHOLESTEROL	.062	.028	.219	2.218	.029

a. Dependent Variable: BEDAYNTM

Tương tự thay thế cholesterol bằng yếu tố khác: tuổi, glucose....ta có thể tóm tắt các trị số trong bảng sau:

Bảng 2. Tổng hợp các trị số trong phân tích hồi qui tuyến tính đơn

Các yếu tố	β	α	R	R^2	p
Tuổi	0,011	0,406	0,408	0,166	0,000
Cholesterol	0,062	0,748	0,219	0,048	0,020
Glucose	0,013	0,981	0,140	0,020	0,165
HA tâm thu	0,002	0,771	0,107	0,011	0,290
BMI	0,003	0,999	0,027	0,001	0,788

β: Hệ số hồi qui; α :điểm cắt trên trục tung; R: hệ số tương quan; p: ý nghĩa thống kê

Ta có thể viết các phương trình tuyến tính đơn sau:

1. Bề dày NTM= 0,748 + 0,062 (cholesterol)
2. Bề dày NTM= 0,406 + 0,011 (tuổi)
3. Bề dày NTM= 0,981 + 0,062 (glucose)
4. Bề dày NTM= 0,771 + 0,002 (Huyết áp TT)

5. Bề dày NTM= 0,999 + 0,003 (BMI)

Nhìn vào bảng phân tích đơn biến ta thấy hệ số tương quan R của tuổi cao nhất (0,408) rồi đến >cholesterol (0,219)>Glucose (0,140) >HA tâm thu (0,107)>BMI (0,027). Như vậy tuổi và cholesterol là 2 yếu tố có ảnh hưởng nhiều nhất làm dày lớp NTM.

Trị số R^2 cũng tương tự như vậy. Trong thống kê trị số R^2 được diễn dịch như sau:

Sự thay đổi của tuổi giải thích được 16,6% sự thay đổi của bề dày lớp NTM

Sự thay đổi của cholesterol giải thích được 4,8% sự thay đổi của bề dày lớp NTM.

Tương tự glucose giải thích được 2%, HA tâm thu giải thích được 1,1%, và

BMI chỉ giải thích được 0,1% sự thay đổi lớp bề dày NTM. Tóm lại nếu cộng lại tất cả các yếu tố trên chỉ giải thích được khoảng 24,6% sự thay đổi bề dày của lớp NTM. Như vậy có thể còn có nhiều yếu tố khác làm tăng bề dày lớp NTM như do di truyền, chủng tộc, giới tính hoặc mắc bệnh tim mạch (không thuộc phạm vi bài nghiên cứu này).

Hệ số hồi qui β (nếu +: tỉ lệ thuận và – tỉ lệ nghịch).Trong thống kê hệ số hồi qui được diễn dịch như sau:

Cứ tuổi tăng lên 1 đơn vị (1 tuổi) thì bề dày lớp NTM tăng lên 0,011 mm

Cứ cholesterol tăng lên 1 đơn vị (1mmol/L) thì bề dày lớp NTM tăng lên 0,062 mm

Cứ glucose tăng lên 1 đơn vị (1 mmol/L) thì bề dày lớp NTM tăng lên 0,013 mm

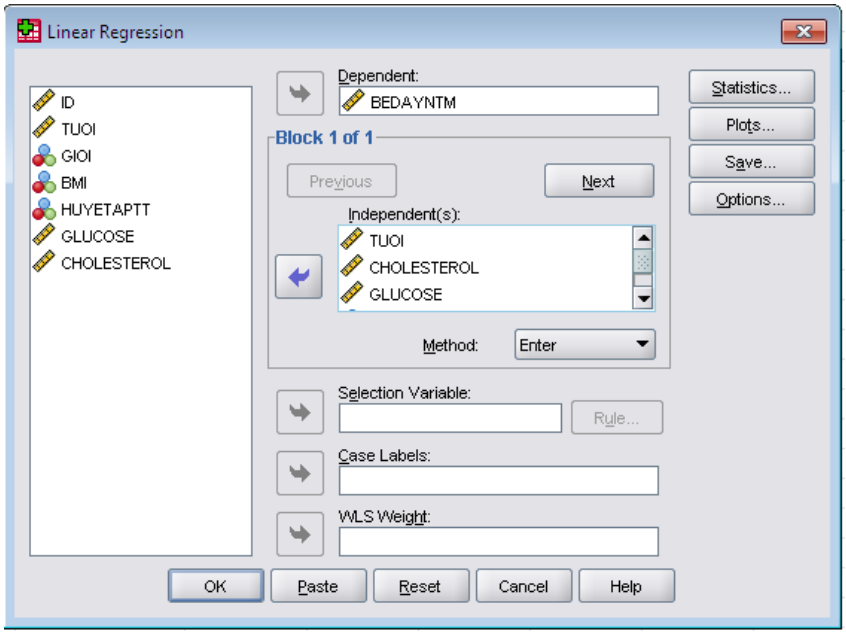
Cứ HA tâm thu tăng lên 1 đơn vị (1 mmHg) thì bề dày lớp NTM tăng lên 0,002 mm

Cứ BMI tăng lên 1 đơn vị (kg/m^2) thì bề dày lớp NTM tăng lên 0,003 mm.

Nếu nhìn trị số p (p value) ta thấy chỉ có 2 yếu tố có ý nghĩa thống kê là tuổi và cholesterol.

Bây giờ nếu đưa tất cả các biến vào cùng lúc để phân tích đa biến.

Vào **Analyse> Regression> Linear Regession.** Mở hộp thoại Linear Regession. Nhắp chuyển tất cả 5 biến (TUOI, CHOLESTEROL, HUYETAPTT, GLUCOSE, BMI) vào ô Independent (s) và nhắp chuyển BEDAYNTM vào ô Dependent.



Nhấn OK, kết quả như sau:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change		
					R Square Change	F Change	
1	.457[a]	.209	.166	.36869	.209	4.953	

a. Predictors: (Constant), BMI, TUOI, HUYETAPTT, GLUCOSE, CHOLESTEROL

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.366	5	.673	4.953	.000[a]
	Residual	12.778	94	.136		
	Total	16.144	99			

a. Predictors: (Constant), BMI, TUOI, HUYETAPTT, GLUCOSE, CHOLESTEROL

b. Dependent Variable: BEDAYNTM

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.108	.388		-.277	.782
	TUOI	.010	.002	.382	4.068	.000
	CHOLESTEROL	.038	.027	.135	1.403	.164
	GLUCOSE	.008	.009	.081	.876	.383
	HUYETAPTT	.002	.002	.110	1.179	.241
	BMI	.000	.011	-.002	-.020	.984

a. Dependent Variable: BEDAYNTM

Lúc này phương trình hồi qui tuyến tính đa biến có thể viết như sau:

Bề dày NTM= -0,108 + 0,01(tuổi) + 0,038 (cholesterol) + 0,008 (glucose) + 0,002 (huyết áp)

(Bỏ BMI vì hệ số hồi qui gần bằng 0)

Hệ số tương quan chung là R=0,457 và tất cả 5 yếu tố này chỉ giải thích được gần 21% (R^2=0,209) sự thay đổi của bề dày NTM.

Nhìn vào trị số p (cột cuối cùng-Sig.), chỉ có tuổi là có giá trị thống kê với p=0,000.

Như vậy trong phân tích từng biến riêng lẻ (phân tích đơn biến) thì tuổi và cholesterol có ý nghĩa thống kê. Trong phân tích đa biến chỉ còn tuổi là có ý nghĩa độc lập giải thích sự thay đổi bề dày lớp NTT. Thật vậy, giữa tuổi và bề dày lớp NTM có tương quan cao nhất (R=0,408) và giải thích 16,6% sự thay đổi của bề dày NTM. Bốn yếu tố còn lại giải thích vào khoảng 5% sự tăng bề dày lớp NTM.

Kết luận: Chỉ có tuổi là ***yếu tố độc lập*** có ý nghĩa dự đoán bề dày lớp NTM động mạch cảnh, các yếu tố khác như cholesterol, glucose, huyết áp, BMI không có hoặc có rất ít ảnh hưởng trên bề dày lớp NTM.

TS Nguyễn Ngọc Rạng, email:rangbvag@yahoo.com
Website:bvag.com.vn

Tài liệu tham khảo:

1. Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010 Nov;107(44):776-82.
2. Tripepi G, Jager KJ, Stel VS, Dekker FW, Zoccali C. How to Deal with Continuous and Dichotomic Outcomes in Epidemiological Research: Linear and Logistic Regression Analyses. Nephron Clin Pract. 2011 Feb 23;118(4):c399-c406.
3. Hoàng Trọng và Chu Nguyễn Mộng Ngọc, Phân tích dữ liệu nghiên cứu với SPSS. Nhà xuất bản thống kê năm 2005