

PHÂN TÍCH HỒI QUI LOGISTIC ĐƠN VÀ ĐA BIẾN

Phân tích hồi qui logistic là một kỹ thuật thống kê để xem xét mối liên hệ giữa biến độc lập (biến số hoặc biến phân loại) với biến phụ thuộc là biến nhị phân. Trong hồi qui tuyến tính đơn, biến độc lập x và phụ thuộc y là biến số liên tục liên hệ qua phương trình:

$$y = \alpha + \beta x + \varepsilon$$

Trong hồi qui logistic, biến phụ thuộc y chỉ có 2 trạng thái 1 (ví dụ tử vong) và 0 (ví dụ sống). Muốn đổi ra biến số liên tục người ta tính xác suất của 2 trạng thái này. Nếu gọi p là xác suất để một biến cố xảy ra (ví dụ: tử vong), thì $1-p$ là xác suất để biến cố không xảy ra (ví dụ: sống). Phương trình hồi qui logistic phát biểu:

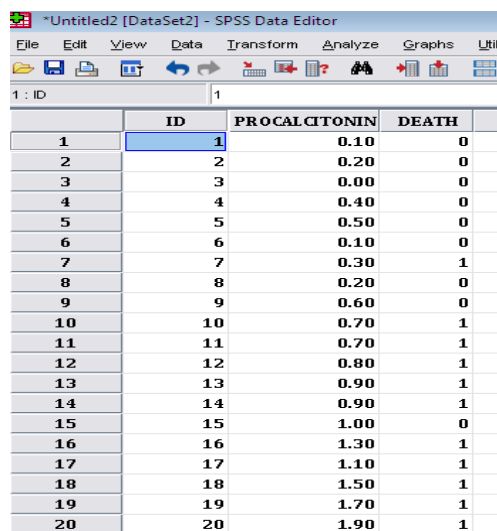
$$\text{Log} \left(\frac{p}{1-p} \right) = \alpha + \beta x + \varepsilon$$

Từ phương trình này, ta có thể tính xác suất tiên đoán tử vong theo trị số của x .

$$\frac{p}{1-p} = e^{\alpha + \beta x}$$

$$p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

Ví dụ: Nghiên cứu sự liên hệ giữa nồng độ procalcitonin máu và tiên lượng tử vong trong nhiễm trùng huyết. Dữ liệu thu thập ở 20 bệnh nhân như sau:

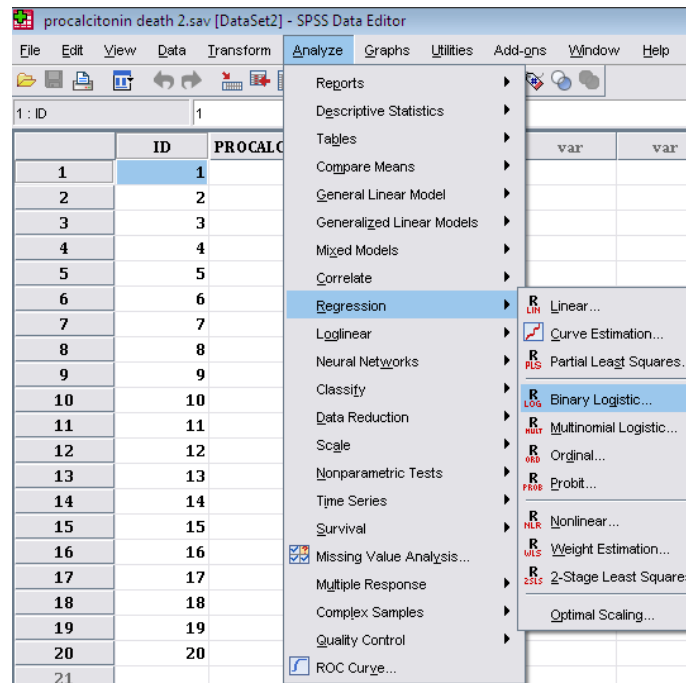


	ID	PROCALCITONIN	DEATH
1	1	0.10	0
2	2	0.20	0
3	3	0.00	0
4	4	0.40	0
5	5	0.50	0
6	6	0.10	0
7	7	0.30	1
8	8	0.20	0
9	9	0.60	0
10	10	0.70	1
11	11	0.70	1
12	12	0.80	1
13	13	0.90	1
14	14	0.90	1
15	15	1.00	0
16	16	1.30	1
17	17	1.10	1
18	18	1.50	1
19	19	1.70	1
20	20	1.90	1

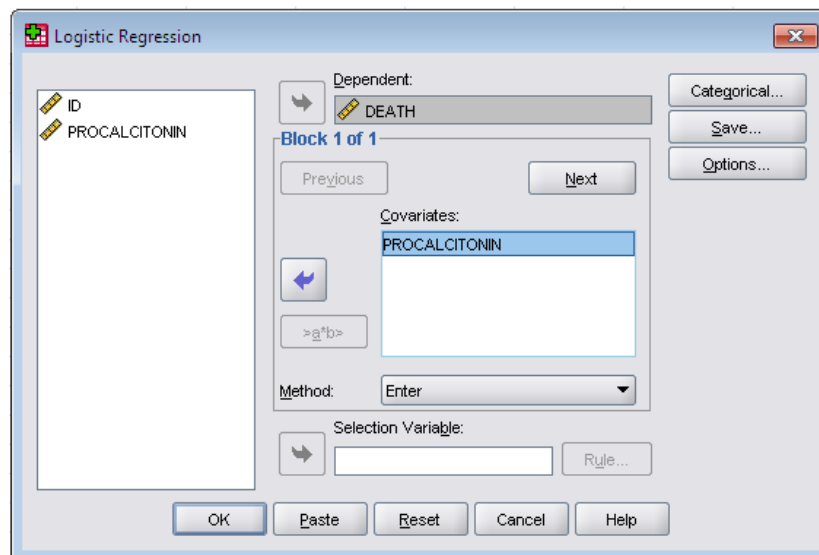
Có 3 cột: ID bệnh nhân, PROCALCITONIN máu (ng/ml), DEATH(0: sống; 1: Tử vong)

Muốn tính p (xác suất tiên đoán tử vong)

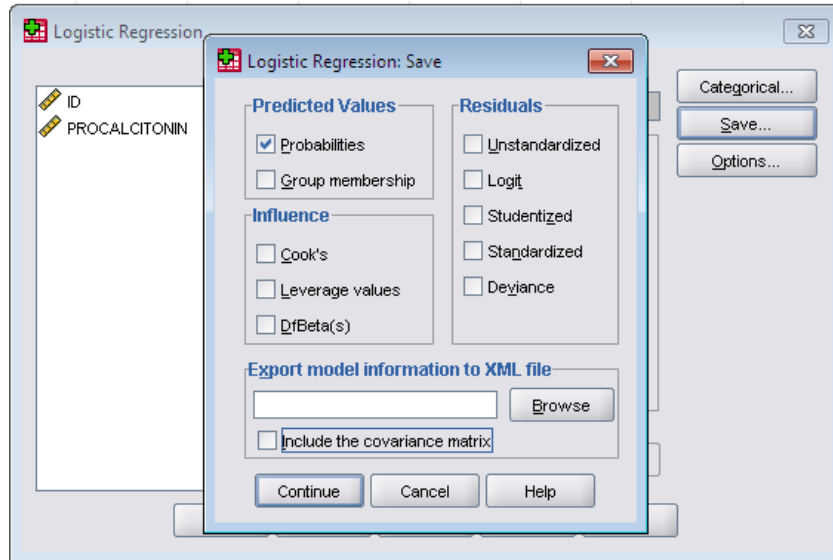
Vào menu: Analyze > Regression > Binary Logistic



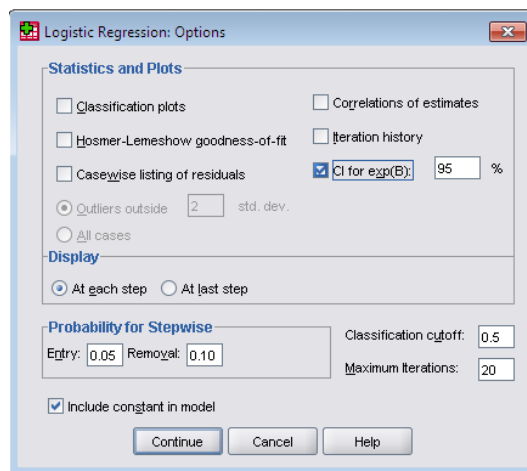
Mở hộp thoại Logistic Regression. Nhấp chuyển biến DEATH (nhị phân) vào ô Dependent và biến PROCALCITONIN (biến số liên tục) vào ô Covariates



Nhấp nút Save, mở hộp thoại Save, đánh dấu nháy vào ô Predicted Values (đây là xác suất tiên đoán tử vong). Nhấp Continue



Nhấp nút Options, mở hộp thoại Options, đánh dấu nháy vào ô CI for exp(B) [khoảng tin cậy của hệ số B=chính là tỉ số odds)



Nhấp Continue, nhấp OK

Kết quả: Trong file dữ liệu, xuất hiện thêm cột PRE-1, đây chính là xác suất tiên đoán tử vong trên trục tung (y)

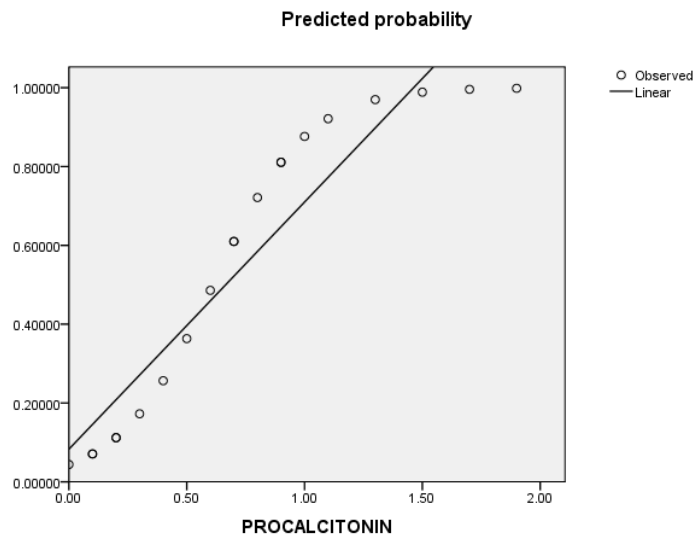
*procalcitonin death 2.sav [DataSet2] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons W

1: ID 1

	ID	PROCALCITONIN	DEATH	PRE_1
1	1	0.10	0	0.07074
2	2	0.20	0	0.11187
3	3	0.00	0	0.04398
4	4	0.40	0	0.25647
5	5	0.50	0	0.36339
6	6	0.10	0	0.07074
7	7	0.30	1	0.17249
8	8	0.20	0	0.11187
9	9	0.60	0	0.48575
10	10	0.70	1	0.60985
11	11	0.70	1	0.60985
12	12	0.80	1	0.72119
13	13	0.90	1	0.81062
14	14	0.90	1	0.81062
15	15	1.00	0	0.87629
16	16	1.30	1	0.96979
17	17	1.10	1	0.92139
18	18	1.50	1	0.98875
19	19	1.70	1	0.99586
20	20	1.90	1	0.99849

Sau khi đã có p, vẽ biểu đồ liên hệ giữa x (procalcitonin) và y (xác suất p tiên đoán tử vong)



Như vậy có sự liên hệ tuyến tính giữa nồng độ procalcitonin (trục x) và xác suất dự đoán tử vong p (trục y)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	15.109 ^a	.462	.619

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table^a

		Predicted		
		DEATH		Percentage Correct
Observed	0	1		
Step 1 DEATH 0	8	1	88.9	
1	1	10	90.9	
Overall Percentage			90.0	

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	PROCALCITONIN	5.037	2.186	5.307	1	.021	153.988	2.120	1.118E4
	Constant	-3.079	1.480	4.327	1	.038	.046		

a. Variable(s) entered on step 1: PROCALCITONIN.

Từ bảng kết quả trên ta có thể viết phương trình hồi qui logistic:

$$\text{Log} \left(\frac{p}{1-p} \right) = -3,079 + 5,037 * \text{procalcitonin}$$

Như vậy cứ procalcitonin tăng lên 1 ng/ml thì log của tỉ lệ xác suất chết/sống sẽ tăng thêm 5,037 lần. Để dễ diễn dịch, ta có thể viết:

$$\left(\frac{p}{1-p} \right) = e^{-3,079 + 5,037 \cdot \text{procalcitonin}}$$

Ta biết $\frac{p}{1-p}$ chính là odd, vì vậy có thể viết:

$$\text{odd} = e^{-3,079 + 5,037 \cdot \text{procalcitonin}}$$

Nếu gọi odd₀ khi procalcitonin=0 thì odd₀ = e^{-3,079}

Và gọi odd₁ khi procalcitonin tăng lên 1 đơn vị thì odd₁ = e^{-3,079 + 5,037}

$$\text{Và tỉ số của 2 odds chính là odds ratio: } \frac{\text{Odd}_1}{\text{Odd}_0} = \frac{e^{-3,079 + 5,037}}{e^{-3,079}} = e^{5,037} = 153,98$$

(Xem cột Exp(B) ở bảng trên)

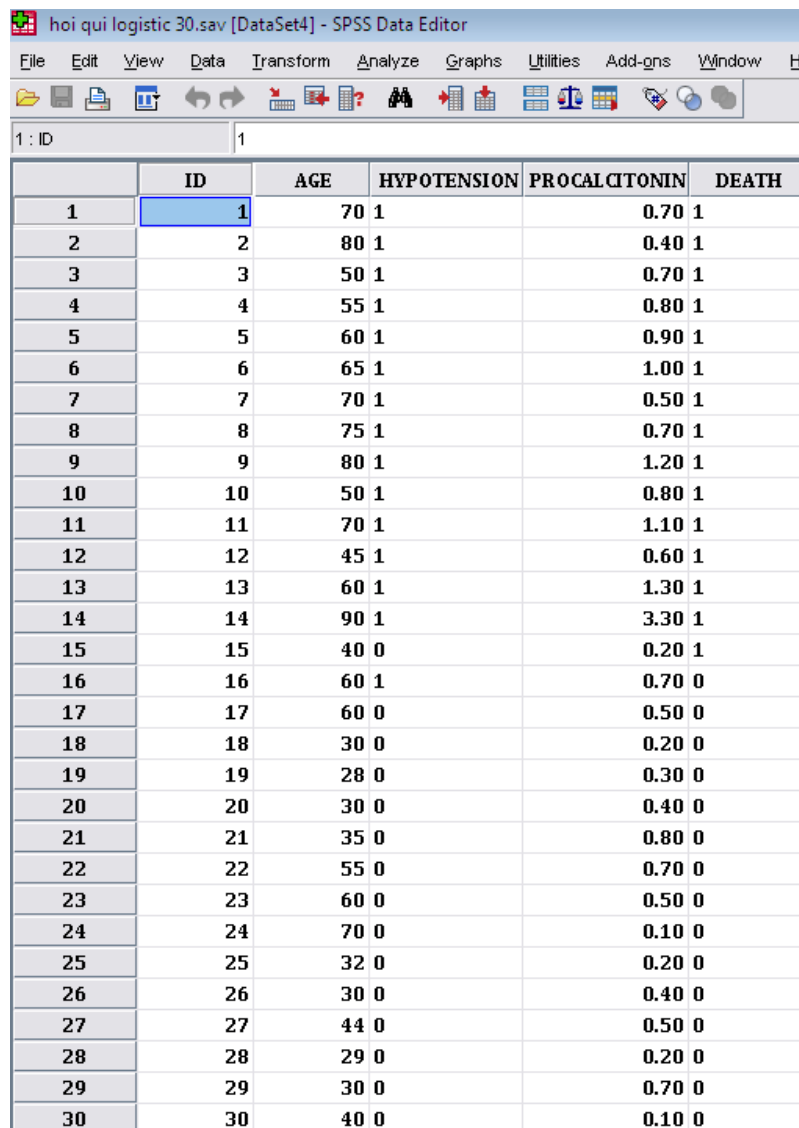
Lúc này ta có thể diễn dịch, cứ procalcitonin tăng lên 1 ng/ml thì nguy cơ tử vong tăng lên 154 lần (kết quả exp (B) ở bảng 1, với e^{5,037} = 153,988) hoặc đúng hơn là cứ procalcitonin tăng thêm 0,1 ng/ml thì nguy cơ tử vong tăng lên 15,4 lần.

Phân tích hồi qui logistic đa biến.

Bây giờ nếu ta xét đến nhiều yếu tố khác có thể có thể làm tăng nguy cơ tử vong trong nhiễm trùng huyết, ví dụ như tuổi bệnh nhân, hạ huyết áp, có bệnh nền hoặc suy giảm miễn dịch... Trong ví dụ này ta chỉ xét đến 3 yếu tố: procalcitonin (biến số), hạ huyết áp (biến nhị phân) và tuổi bệnh nhân (biến số).

Dữ liệu 30 bệnh nhân nhiễm trùng huyết như sau:

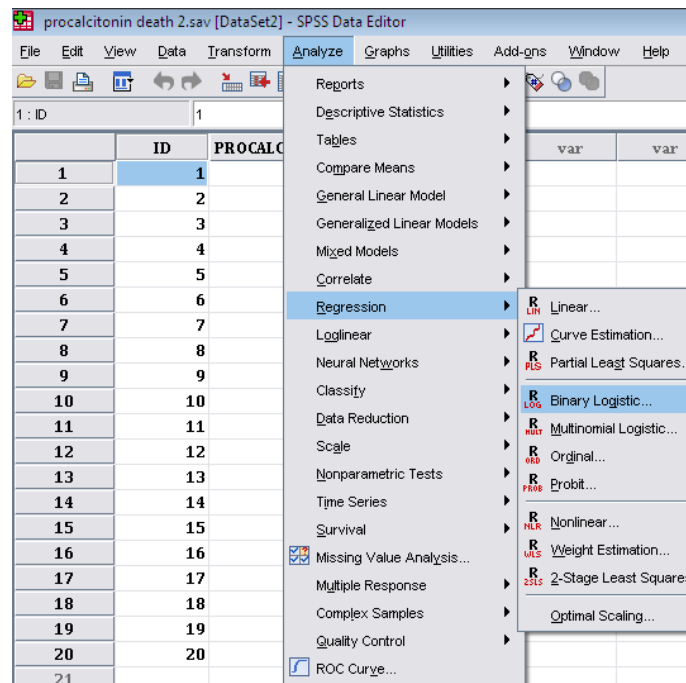
Có 5 cột: ID, Age (tuổi), Hypotension (hạ huyết áp), Procalcitonin (ng/ml) và Death (tử vong)



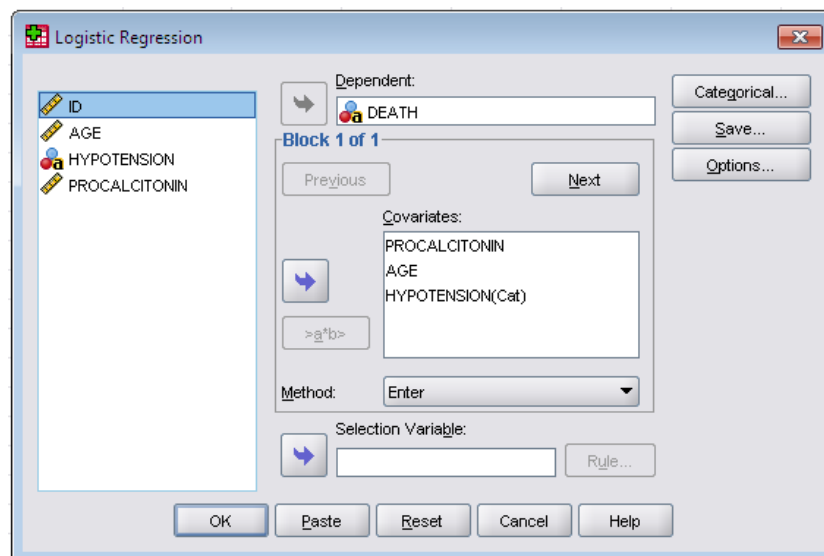
The screenshot shows the SPSS Data Editor window for a file named 'hoi qui logistic 30.sav'. The data is organized into a table with 30 rows and 5 columns. The columns are labeled ID, AGE, HYPOTENSION, PROCALCITONIN, and DEATH. The data points are as follows:

	ID	AGE	HYPOTENSION	PROCALCITONIN	DEATH
1	1	70	1	0.70	1
2	2	80	1	0.40	1
3	3	50	1	0.70	1
4	4	55	1	0.80	1
5	5	60	1	0.90	1
6	6	65	1	1.00	1
7	7	70	1	0.50	1
8	8	75	1	0.70	1
9	9	80	1	1.20	1
10	10	50	1	0.80	1
11	11	70	1	1.10	1
12	12	45	1	0.60	1
13	13	60	1	1.30	1
14	14	90	1	3.30	1
15	15	40	0	0.20	1
16	16	60	1	0.70	0
17	17	60	0	0.50	0
18	18	30	0	0.20	0
19	19	28	0	0.30	0
20	20	30	0	0.40	0
21	21	35	0	0.80	0
22	22	55	0	0.70	0
23	23	60	0	0.50	0
24	24	70	0	0.10	0
25	25	32	0	0.20	0
26	26	30	0	0.40	0
27	27	44	0	0.50	0
28	28	29	0	0.20	0
29	29	30	0	0.70	0
30	30	40	0	0.10	0

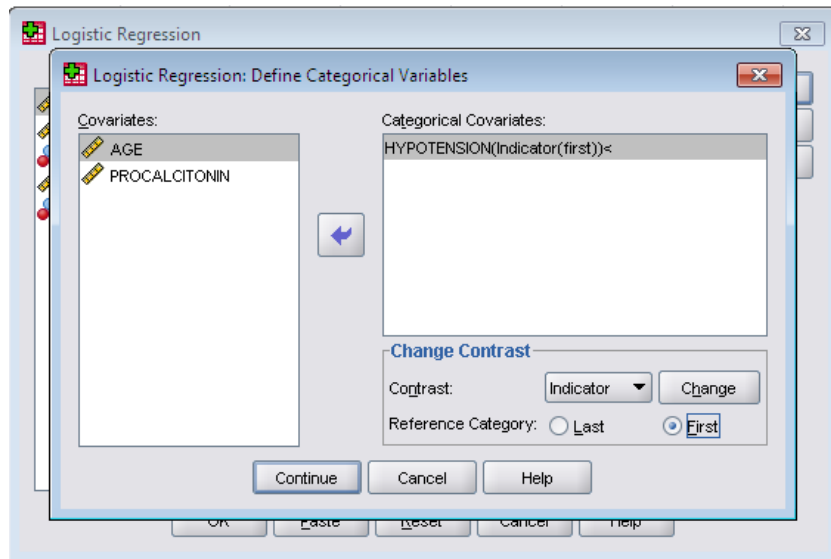
Vào menu: Analyze > Regression > Binary Logistic



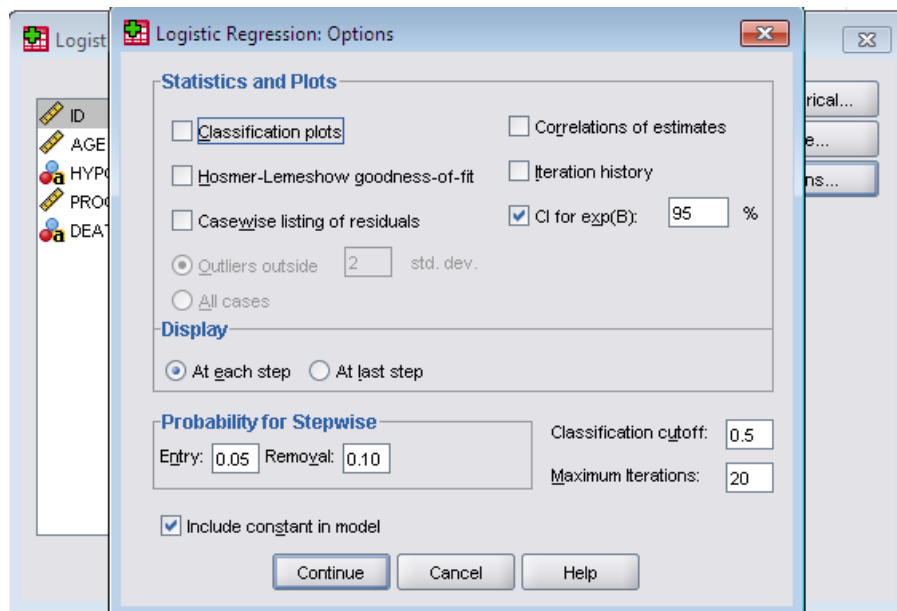
Mở hộp thoại Logistic Regression, Nhấp chuyển biến DEATH vào ô Dependent. Nhấp cả 3 biến AGE, HYPOTENSION và PROCALCITONIN vào ô Covariates.



Nhấp nút Categorical để khai báo biến HYPOTENSION là biến phân loại (hoặc nhị phân)
Và khai báo giá trị tham khảo của HYPOTENSION là 1 (có hạ huyết áp): Reference Category là First , Nhấp Continue



Sau đó nhấp nút Options, vào hộp thoại Options, đánh dấu nháy vào ô CI for Exp(B): để xem tỉ số odds và khoảng tin cậy 95% của odds.



Nhấp Continue, sau đó nhấp OK. Kết quả như sau:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	14.627 ^a	.593	.791

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table^a

Observed		Predicted		Percentage Correct
		DEATH		
		0	1	
Step 1	DEATH 0	14	1	93.3
	1	1	14	93.3
Overall Percentage				93.3

a. The cut value is .500

Variables in the Equation

Step	Variable	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
1 ^a	PROCALCITONIN	.137	1.875	.005	1	.942	1.147	.029	45.204
	AGE	.013	.058	.049	1	.825	1.013	.904	1.135
	HYPOTENSION(1)	4.915	1.968	6.240	1	.012	136.326	2.882	6447.900
	Constant	-3.232	2.634	1.506	1	.220	.039		

a. Variable(s) entered on step 1: PROCALCITONIN, AGE, HYPOTENSION.

Chỉ lấy 3 bảng cuối cùng và được diễn dịch như sau:

Trị số -2 Log likelihood chỉ độ phù hợp của mô hình. Trị số này càng nhỏ thì mô hình càng phù hợp. Trong mô hình này với -2 Log likelihood = 14.627 là không cao lắm, như vậy có độ phù hợp khá tốt với mô hình tổng thể.

Với kết quả này mô hình hồi qui logistic được viết là:

$$\text{Log}\left(\frac{p}{1-p}\right) = -3,232 + 0,137 \cdot \text{PROCALCITONIN} + 0,013 \cdot \text{AGE} + 4,915 \cdot \text{HYPOTENSION}$$

$$\text{Hoặc: } \frac{p}{1-p} = e^{-3,232 + 0,137 \cdot \text{PROCALCITONIN} + 0,013 \cdot \text{AGE} + 4,915 \cdot \text{HYPOTENSION}}$$

Như diễn dịch ở phần trên:

- Tỷ số odds của procalcitonin: $OR_1 = e^{0,137} = 1,147$
- Tỷ số odds của tuổi: $OR_1 = e^{0,013} = 1,013$
- Tỷ số odds của hạ huyết áp: $e^{4,915} = 136$

(Xem 3 cột cuối cùng ở bảng trên, sẽ thấy tỷ số odds và khoảng tin cậy 95%)

Ta có thể tóm tắt trong bảng sau:

Yếu tố	Odds ratio (OR)	Khoảng tin cậy 95%	Ý nghĩa thống kê p
Procalcitonin	1,14	0,03 - 45,20	0,94
Tuổi	1,01	0,90 - 1,13	0,82
Hạ huyết áp	136	3 – 6447	0,01

Kết luận: Chỉ có hạ huyết áp là yếu tố độc lập dẫn đến tử vong ở bệnh nhân bị nhiễm trùng huyết với OR=136 (KTC 95%: 3 – 6447; p=0,01)

BS Nguyễn Ngọc Rạng, email:rangbvag@yahoo.com

Website: bvag.com.vn

Tài liệu tham khảo:

1. Nguyễn Văn Tuấn. 2007. Phân tích hồi qui logistic trong: Phân tích số liệu và tạo biểu đồ bằng R. Nhà Xuất bản Khoa học và Kỹ thuật. trang 215-218.
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