

로지스틱 회귀분석

문 재 동

(전남의대 산업의학과)

1. 적용

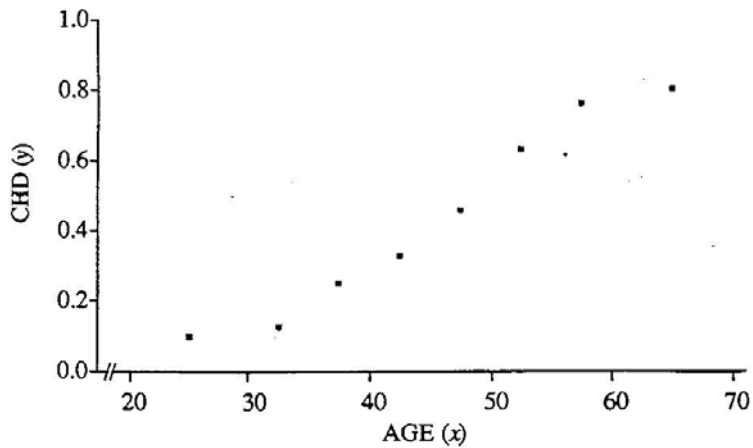
종속변인이 이분(dichotomous) 명목척도로 표현되는 회귀분석으로 혼돈요인(confounder)의 영향을 배제한 후

- 예측인자(predictor)나 진단인자(diagnostic factor)의 선택, 기여도 순위 결정
- 사건발생 가능성(probability), 교차비(odds ratio, OR)의 산출
- 종속변인은 dichotomous, 독립변인은 연속 또는 이산변인 등 상관이 없으며 정규분포의 가정을 전제로 하지 않는다.

2. 이론적 배경

Frequency Table of Age Group by CHD.

Age Group	n	CHD		Mean (Proportion)
		Absent	Present	
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
Total	100	57	43	0.43



Plot of the Mean of CHD in Each Age Group.

$$P = \frac{1}{1 + e^{-z}} = \frac{e^{(z)}}{1 + e^{(z)}}$$

P , probability of event, $z = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$

■ 교차비(odds ratio)의 산출

· Odds: The **ratio of the probability** of occurrence of an event to that of nonoccurrence.

		Disease		
		+	-	
Factor	+	a	b	a+b
	-	c	d	c+d
		a+c	b+d	N

		Disease		
		+	-	
Factor	+	p_1	$1-p_1$	1
	-	p_0	$1-p_0$	1

$$\frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc} = \text{odds ratio} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$$

$$\cdot \text{Relative risk}(RR) = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} \quad \cdot \text{OR or RR} = 1 (?)$$

그러므로 logistic 함수

$$P = \frac{1}{1 + e^{-z}} = \frac{e^{(z)}}{1 + e^z} \text{로 부터}$$

$$P, \text{ probability of event, } z = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

$$\text{Ln}(P/1-P) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p = \text{Ln}(\text{odds}) = \text{Logit}$$

$$\text{Ln}(\text{odds ratio, OR}_{x_1}) = b_1: \text{ regression coefficient of } X_1$$

$$\text{OR} = e^{(\text{regression coefficient})}$$

3. Logistic 회귀분석의 실제 이용

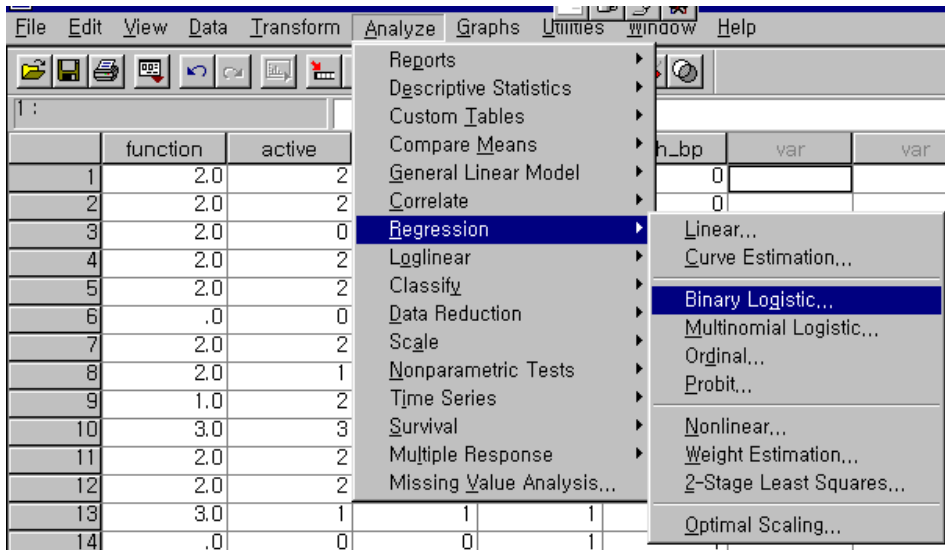
1) 자료/ Kasser

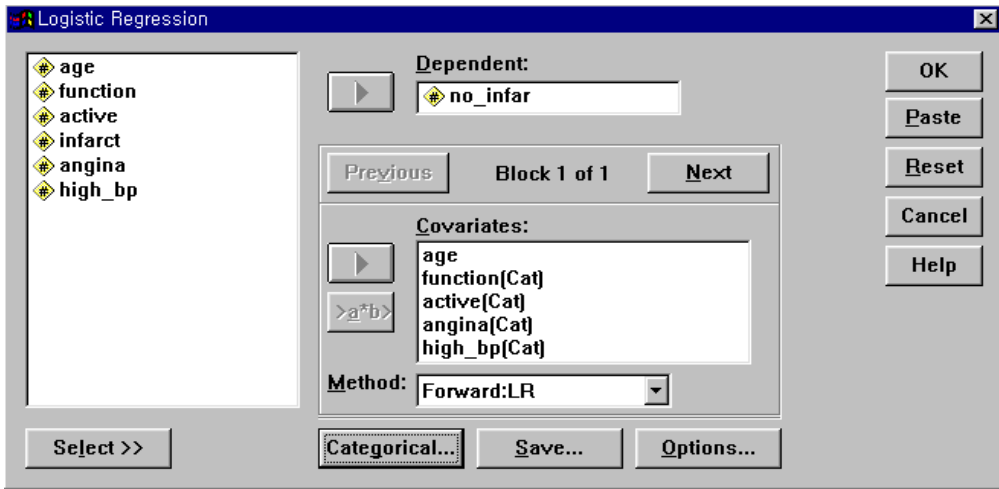
	age	function	active	infarct	angina	high_bp
1	42	2.0	2	1	1	0
2	66	2.0	2	1	1	0
3	56	2.0	0	1	1	0
4	55	2.0	2	1	1	0
5	41	2.0	2	1	1	1
6	62	.0	0	1	0	1
7	46	2.0	2	1	1	1
8	44	2.0	1	0	1	1
9	50	1.0	2	0	1	1
10	73	3.0	3	0	1	0
11	48	2.0	2	1	1	0

2) 분석목적

독립변인 상호간에 영향 배제 후 no infarction의 유발과 관계가 있는 요인들은 무엇이며, 유의한 예측인자들에 폭로시 no infarction의 가능성은 어느 정도인가?

3) 실행 SPSSWIN





■ 모델의 일반적 선택기법

- Forward selection
- Backward elimination
- Stepwise selection
- Fixed model

4) 결과출력

Dependent Variable Encoding

Original Value	Internal Value
.00	0
1.00	1

Categorical Variables Codings

		Frequency	Parameter coding		
			(1)	(2)	(3)
ACTIVE	0	29	.000	.000	.000
	1	9	1.000	.000	.000
	2	60	.000	1.000	.000
	3	19	.000	.000	1.000
FUNCTION	.0	20	.000	.000	.000
	1,0	24	1.000	.000	.000
	2,0	46	.000	1.000	.000
ANGINA	3,0	27	.000	.000	1.000
	0	19	.000	.000	.000
HIGH_BP	1	98	1.000	.000	.000
	0	81	.000	1.000	.000
	1	36	1.000	.000	.000

Block 0: Beginning Block

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	-.543	.192	8.014	1	.005	.581

Variables not in the Equation

Step	Variables	Score	df	Sig.
0	AGE	.309	1	.578
	FUNCTION	.804	3	.849
	FUNCTION(1)	.007	1	.932
	FUNCTION(2)	.127	1	.722
	FUNCTION(3)	.177	1	.674
	ACTIVE	1.684	3	.640
	ACTIVE(1)	1.483	1	.223
	ACTIVE(2)	.163	1	.687
	ACTIVE(3)	.261	1	.609
	ANGINA(1)	4.287	1	.038
	HIGH_BP(1)	.540	1	.462
Overall Statistics		9.634	9	.381

Classification Table^a

Observed	Predicted	NO_INFAR		Percentage Correct
		.00	1.00	
Step 1 NO_INFAR	.00	74	0	100.0
	1.00	43	0	.0
Overall Percentage				63.2

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 ANGINA(1)	1.302	.662	3.871	1	.049	3.677	1.005	13.454
Constant	-1.674	.629	7.078	1	.008	.188		

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

Model if Term Removed

Variable	Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1 ANGINA	-76,942	4,778	1	,029

Variables not in the Equation

Step	Variables	Score	df	Sig.
1	AGE	,251	1	,616
	FUNCTION	1,763	3	,623
	FUNCTION(1)	,378	1	,539
	FUNCTION(2)	,259	1	,611
	FUNCTION(3)	,661	1	,416
	ACTIVE	3,218	3	,359
	ACTIVE(1)	2,543	1	,111
	ACTIVE(2)	,340	1	,560
	ACTIVE(3)	,616	1	,432
	HIGH_BP(1)	,586	1	,444
Overall Statistics		5,625	8	,689

5) 결과 해석

표7-2 Factor affecting the probability of **no** myocardial infarction

Factor	OR (95% CI)*
Angina	
No	1.0
Yes	3.68 (1.01 - 13.5)

* Odds ratio (95% confidence interval)

상기 독립변인들 가운데 유의수준 $\alpha = 0.05$ 에서 no infarction과 통계적으로 유의하게 관계가 있는 예방인자는 angina 경험여부이며($P = 0.049$) odds ratio가 3.68(95% confidence interval 1.005 - 13.6)로서 angina를 경험한 경우 경험하지 않은 군에 비해 약 3.7배 infarct 예방효과가 있는데, angina를 경험한 경우 no infarction의 가능성은 0.408이다.

■ Crude odds ratio

		No infarct	Infarct
Angina	+	40	58
	-	3	16

$$\text{Crude OR} = \frac{40 \times 16}{58 \times 3} = 3.678$$

참 고 문 헌

- Clayton D, Hills M (1993): Statistical Models in Epidemiology. Oxford: Oxford University Press.
- Dawson B, Trapp RG (2001): Basic & Clinical biostatistics, 3rd ed. New York: McGraw_Hill.
- Dixon WJ (1992): BMDP Statistical Software Manual. Berkeley: University of California Press.
- Hosmer DW, Lemeshow S (1989): Applied Logistic Regression. New York: John Wiley & Sons.
- Kleinbaum DG, Kupper LL, et al (1988): Applied Regression Analysis and Other Multivariable Methods. Boston: PWS-KENT Publishing Company.
- Kleinbaum DG (1994): Logistic Regression. A Self-Learning Text. New York: Springer-Verlag.
- McCullagh P, Nelder JA (1989): Generalized Linear Models, 2nd ed. London: Chapman & Hall.
- SPSS (1999). Version 10.0. SPSS Inc. Chicago, IL.
- Vogt WP (1993): Dictionary of Statistics and Methodology. A Nontechnical Guide for The Social Sciences.