역학 연수교육 II

역학자료분석에서 통계 패캐지 이용
- EGRET 및 GLIM 시스템 -

유 근 영
서울대학교 의과대학

1. Analytic Strategy for Ordinal Variables
   1) Score test for trend in two-by-k table (EGRET)
   2) Likelihood ratio test for trend (GLIM)

2. Adjust■ent for Confounders by Linear Logistic Model
   1) De■onstration of confounder in a case-control study (PC-SAS)
   2) Strategy for adjust■ent by linear logistic ■odel (GLIM)
   3) Esti■ation of adjusted MLE with intervals (GLIM/EGRET)

3. Quantitative Analysis with Linear Logistic Model
   1) Model selection procedure (GLIM)
   2) Likelihood ratio test (GLIM)
   3) Quantitative interpretation of linear logistic ■odel
   4) Joint odds ratio esti■ation (GLIM)
   5) Quantitative interpretation of non-linear logistic ■odel

4. Conditional Linear Logistic Regression Analysis
   1) Esti■ation of conditional MLE using linear logistic ■odel to fit 1:1 ■atched data (GLIM/EGRET)

5. Application of Log-Linear Model
   1) Log-linear ■odel (GLIM)
   2) Constant hazard ■odel (GLIM)
   3) Proportional hazard ■odel (GLIM)

Analytic Strategy for Ordinal Variables

1) Score Test for Trend

Armitage Test (Cochran 1954; Armitage 1955)

$H_0 : \beta = 0$

$$S^2 = \frac{b}{V(b)}$$

b → (estimated regression coefficient of P_i on e_i)

V(b) → (estimated sampling variance of b on the H_0)

$$= \frac{[\sum_i e_i (a_i - E_i)]^2}{\sum Var(b)} \approx \chi^2, \ df = 1$$

$$_uS^2_{Arm} = \frac{N^3[\sum e_i (a_i - E_i)]^2}{n_1 n_2 [N \sum e_i^2 m_i - (\sum e_i m_i)^2]} \approx \chi^2, \ df = 1$$

【예 제】 단면적 조사의 결과 서울지역 성인 남자의 B형간염 감염율은 아래와 같았다. 이 집단에서 B형 간염 감염율은 연령증가에 따라 증가한다고 할 수 있겠는가?

Age (years)	No. tested	Infected group No. inf.	P.(%)	95% C.I(%)
20-29	131	102	77.9	69.9- 85.1
30-39	255	210	82.4	77.5- 87.3
40-49	159	141	88.7	83.5- 93.9
50-59	81	77	95.1	89.8-100.0
Total	626	530	84.7	81.8- 87.6

< Score Test for Trend by EGRET>

EGRET (R)

Epidemiological GRaphics, Estimation, and Testing package
ANALYSIS MODULE (PECAN), version 0.26.6 ; EPIXACT (R), version 0.03

VARIABLES

1. AGE   2. HBV   3. N

ANALYSIS MODEL:	Traditional 2xK table analysis
EXPOSURE:	Exposure Level Given by Variable: AGE
GROUP SIZE:	Fixed Group Size = 1
OUTCOME SPECIFICATION:	Outcome Variable Name: HBV
REPETITION COUNT:	Repetition Count Given by Variable: N

{X}
{F}

FACTORED VARIABLES:

VARIABLE	#LEVELS	BASE
AGE	4	1

{A}

AGE	HBV 0	HBV 1	UNSTRATIFIED ASYMPTOTIC: ODDS RATIO MLE	CONFIDENCE INTERVAL 95% Cornfield (lower)	CONFIDENCE INTERVAL 95% Cornfield (upper)	TEST FOR OR = 1 chisquare	p
1	29	102					
2	45	210	1.33	.76	2.31	1.12	p=.289
3	18	141	2.23	1.12	4.44	6.17	p=.013
4	4	77	5.47	1.73	19.21	11.21	p<.001

Test for trend: 14.28 p< .001   Test for OR = 1: 14.41 p= .002

Mantel's Extension Test (Mantel 1963)

$H_0 : \beta = 0$

$$S^2_M = \frac{[\Sigma\{T_{1k}-(n_{k1}/n_{k2})T_{2k}\}]^2}{\Sigma[\{n_{k1}n_{k2}(m_kT_{3k}-T_{2k}^2\}/\{m_k^2(m_k-1)\}]} \approx \chi^2 \quad df = 1$$

단, $T_{1k} = \Sigma a_{ki} \cdot e_{ki}$, $T_{2k} = \Sigma m_{ki} \cdot e_{ki}$, $T_{3k} = \Sigma m_{ki} \cdot e_{ki}^2$

【예 제】 위의 예에서 B형간염 감염율의 연령별-출생지역별 분포가 아래와 같을 때, 출생지역을 보정한 상태에서도 연령에 따라 증가한다고 할 수 있겠는가?

	unban		rural	
Age	No. tested	No.(+)	No. tested	No.(+)
1	100	70	31	24
2	151	124	104	97
3	89	76	70	65
4	54	50	27	24
Total	394	320	232	210

< Adjusted Score Test for Trend by EGRET>

EGRET (R)

Epidemiological GRaphics, Estimation, and Testing package
ANALYSIS MODULE (PECAN), version 0.26.6 ; EPIXACT (R), version 0.03

VARIABLES

1. AGE 2. UR 3. HBV 4. N

FACTORED VARIABLES:

VARIABLE	#LEVELS	BASE
AGE	4	1
UR	2	1

{A}

STRATIFICATION VARIABLES

a. UR

STRATIFIED: ASYMPT STATS FOR ALL 2 STRATA COMBINED

AGE	HBV 0	HBV 1	ODDS RATIO M-H	CONFIDENCE INTERVAL 95% Cornfield		TEST OR = 1 chisquare	p-val
1	37	94					
2	34	221	2.26	1.30	4.05	9.56	p=.002
						[Common odds:	p=.270]
3	18	141	2.76	1.42	5.52	10.54	p=.001
						[Common odds:	p=.571]
4	7	74	4.13	1.63	10.76	11.23	p<.001
						[Common odds:	p=.370]

Test for trend: 14.04 p< .001 Test for OR = 1: 18.30 p< .001

2) Likelihood Ratio Test for Trend

< 일 반 원 칙 >

$$\text{logit } P_i(X_k) = \alpha_i$$

$= \alpha_i + \beta_k . X(\text{categorical})$: 각 폭로수준에 따른 OR

$= \alpha_i + \beta_1 . X_k(\text{cotinuous})$: 폭로에 따른 log-RR 의 linear increase 평가

$= \alpha_i + \beta_1 . X_k + \beta_2 . X_k^2$: 폭로에 따른 log-RR 의 quadratic increase 평가

【예 제】 음주와 식도암 발생과의 관련성을 검정하기 위한 환자-대조군 연구결과는 다음과 같았다.

Alcohol intake	CALC (100g/d)	ALCGRP	CASE	CONT	TOTAL
0-39(g/d)	0.2	1	29	386	415
40-79	0.6	2	75	280	355
80-119	1.0	3	51	87	138
120+	1.5	4	45	22	67
T o t a l			200	775	975

< GLIM Program and Output for LR Test for Trend >

```
[o] GLIM 3.77 update 1 (copyright)1985 Royal Statistical Society, London
[o]
[i] ? $INP 8 80
[i] File name? C:\STAT\BRESLOW\ALC2.GLM
[i] $INP? $
[i] $UNIT 88
[i] $DATA AGE ALCGRP TOBGRP CASE CONT
```

```
[i] $READ
[i] 1 1 1 0 40  2 1 1 0 60  3 1 1  1 45  4 1 1 2 47
[i]                         5 1 1  5 43  6 1 1 1 17
    . . . . . . . . . . . . . . . . . . . . . . . . .
[i] $FACT AGE 6 ALCGRP 4 TOBGRP 4
[i] $CAL  N=CASE+CONT : ALC=ALCGRP : TOB=TOBGRP :
[i]       ALC2=ALC*ALC : TOB2=TOB*TOB
[i] $YVAR CASE   $ERR B N
[o]
[i] $FIT  AGE + TOBGRP                $D E
[o] scaled deviance = 210.27 at cycle  5
[o]            d.f. =  79
[o]
[i] $FIT  AGE + TOBGRP + ALC          $D E
[o] scaled deviance =  88.24 at cycle  5
[o]            d.f. =  78
[o]
[i] $FIT  AGE + TOBGRP + ALCGRP       $D E
[o] scaled deviance =  82.34 at cycle  5
[o]            d.f. =  76
[o]          estimate        s.e.      parameter
[o]     1      -6.895       1.083      1
[o]     2       1.981       1.102      AGE(2)
[o]     3       3.776       1.065      AGE(3)
[o]     4       4.335       1.062      AGE(4)
[o]     5       4.896       1.074      AGE(5)
[o]     6       4.827       1.119      AGE(6)
[o]     7      0.4381      0.2283      TOBG(2)
[o]     8      0.5126      0.2730      TOBG(3)
[o]     9       1.641      0.3441      TOBG(4)
[o]    10       1.435      0.2501      ALCG(2)
[o]    11       1.981      0.2848      ALCG(3)
[o]    12       3.603      0.3850      ALCG(4)
[o]     scale parameter taken as  1.000
[o]
[i] $FIT  AGE + TOBGRP + ALC + ALC2  $D E
[o] scaled deviance =  88.14 at cycle  5
[o]            d.f. =  77
[i] $STO
```

Models	No. of parameters	df	G
AGE + TOBGRP	9	79	G_1 = 210.3
AGE + TOBGRP + ALC	10	78	G_2 = 88.2
AGE + TOBGRP + ALCGRP	12	76	G_3 = 82.3
AGE + TOBGRP + ALC + ALC^2	11	77	G_4 = 88.1

① [G_1-G_2] : 'LR test for ALC' = 'LR test for linear trend'
[ALC|AGE, TOBGRP] in log-RR with increasing exposure

$$\chi^2_{trend} = 122.1(1)$$

② [G_1-G_3] : 'LR test for ALCGRP'
[ALCGRP|AGE, TOBGRP]

$$\chi^2_{LR} = 128\ (3)$$

③ [G_2-G_3] : 'Test for departure from the linear model'

$$\chi^2_{departure} = 5.9(2)$$

→ model fitting 때 유용

④ [G_2-G_4] : 'LR test for curvature' in the regression line
→ log-RR 에 미치는 exposure 의 quadratic effect 를 평가

$$\chi^2_{curvature} = 0.1(1)$$

Adjustment for Confounders by Linear Logistic Model

1) Demonstration of Confounder in a Case-Control Study

< 단계 1 > crude ≠ adjusted measures 확인
< 단계 2 > $OR_{ef:D(-)} \neq 1$ 동시에
< 단계 3 > $OR_{df:E(-)} \neq 1$ 일 때

초경 연령	환자군	대조군	OR
합계			
14세 미만	138	100	1.0
14세 이상	112	150	1.85 (1.28-2.68)
진단시연령 (50세 미만)			
14세 미만	12	54	1.0
14세 이상	33	126	0.85 (0.38-1.86)
진단시연령 (50세 이상)			
14세 미만	126	46	1.0
14세 이상	79	24	0.83 (0.45-1.52)

< 단계 1 > crude OR = 1.85
연령 adjusted OR = OR_{mh} = 0.84
- 연령 stratum-specific OR (1) = 0.85
- 연령 stratum-specific OR (2) = 0.83

< 단계 2 > 대조군에서의 OR (연령 vs 초경연령)
= (54x24)/(126x46) = 0.22 ≠ 1

< 단계 3 > 비폭로군에서의 OR (연령 vs 유암발생)
= (12x46)/(126x54) = 0.08 ≠ 1

2) Strategy for Adjustment by Linear Logistic Model

Logistic model 을 이용한 confounder 보정

① confounder 를 층화변수로 하여 각 층마다 logistic model 을 적용

② 보다 쉬운 방법은 model 내에 교란효과를 수학적으로 흡수

(장점) ┌ 여러 형태 ($X, X^2, \log(X)\ \sqrt{X}\ \cdots$) 의 modelling 이 가능
└ continuous nuisance factors 처리 가능

③ 모델 구축시 무조건 집어 넣는게 아니라 층화분석단계에서 충분히 검토

④ confounder 와 RF 는 동일시 하지 말 것
→ RF 영향을 평가함을 잊지 말 것

⑤ confounder 의 숫자는 경제적으로 잘 조절할 것

⑥ 합리적이고 유의한 변수로 구성되는 'multivariate risk equation'

⑦ 보고자 하는 RR 은 가능한 confounding effect 를 제거한 결과로 산출

⑧ 기존에 알려진 confounding variables 들은 비록 그 변수를 모델에 것을 집어 넣음으로해서 risk variable 의 coefficients 가 변화 된다 하더라도, 그 통계적 유의성에 관계없이 model 에 집어 넣을 것

< GLIM Program and Output for Adjustment for Confounder >

```
[o] GLIM 3.77 update 1 (copyright)1985 Royal Statistical Society, London
[i] File name? C:\PM\YOO\ES02.GLM
[i] $UNIT 88
[i] $DATA AGE ALCGRP TOBGRP CASE CONT
[i] $READ
[i] 1 1 1 0 40  2 1 1 0 60  3 1 1  1 45  4 1 1 2 47
[i]                         5 1 1  5 43  6 1 1 1 17
    . . . . . . . . . . . . . . . . . . . . . . . .

[i] $FACT AGE 6 ALCGRP 4 TOBGRP 4
[i] $CAL  N=CASE+CONT :
[i]       AGE1=AGE :
[i]       AGE2=AGE1**2 : AGE3=AGE1**3 : AGE4=AGE1**4
[i] $YVAR CASE   $ERR B N
[o]
[i] $FIT  TOBGRP + ALCGRP  $D E
[o] scaled deviance = 208.83 at cycle  4
[o]            d.f. =  81
[o]          estimate        s.e.      parameter
[o]     1      -2.814       0.2110     1
[o]     2      0.3862       0.2116     TOBG(2)
[o]     3      0.4251       0.2562     TOBG(3)
[o]     4      0.9877       0.2894     TOBG(4)
[o]     5       1.225       0.2343     ALCG(2)
[o]     6       1.998       0.2643     ALCG(3)
[o]     7       3.180       0.3273     ALCG(4)
[o]     scale parameter taken as  1.000
[o]
[i] $FIT  TOBGRP + ALCGRP + AGE1  $D E
[o] scaled deviance = 101.89 at cycle  4
[o]            d.f. =  80
[o]
[i] $FIT  TOBGRP + ALCGRP + AGE1 + AGE2  $D E
[o] scaled deviance =  84.61 at cycle  4
[o]            d.f. =  79
[o]
[i] $FIT  TOBGRP + ALCGRP + AGE1 + AGE2 + AGE3  $D E
[o] scaled deviance =  83.96 at cycle  5
[o]            d.f. =  78
[o]
[i] $FIT  TOBGRP + ALCGRP + AGE1 + AGE2 + AGE3 + AGE4  $D E
[o] scaled deviance =  83.85 at cycle  5
[o]            d.f. =  77
[i] $STOP
```

Risk category	Type of analysis					
	Unadjusted	Polynomial adjustment for age group				Stratified by age
		Linear	Quadratic	Cubic	Quartic	
Tobacco (g/day)						
0-9	0.0	0.0	0.0	0.0	0.0	0.0
10-19	0.39	0.46	0.44	0.43	0.43	0.44
20-29	0.43	0.55	0.51	0.50	0.50	0.51
30+	0.99	1.52	1.63	1.63	1.64	1.64
Alcohol (g/day)						
0-39	0.0	0.0	0.0	0.0	0.0	0.0
40-79	1.23	1.53	1.44	1.44	1.44	1.44
80-119	2.00	2.17	1.99	2.00	1.99	1.98
120+	3.18	3.60	3.57	3.58	3.59	3.60
Goddness-of-fit						
G	208.8	101.9	84.6	84.0	83.8	82.3
df	81	80	79	78	77	76

<해석>

① G(unadj) = 208.8(81)
G(adj) = 82.3(76)
] 126.5(5) 로 매우 큰 차이

⇒ age 는 ESOCA 에 매우 중요한 영향

② RR 들을 보면 별로 차이가 없다

⇒ age 의 confounding effect 는 별로 없다
즉, AGE ↔ ALC, AGE ↔ TOB 는 서로 weakly correlated 되어 있어
strong confounder 아님

③ Polynomial term 을 넣음으로써 보다 효과적으로 control 가능
⇒ 일반적으로 "logistic regression adjustment" 가 좋은 방법이다.

3) Estimation of Adjusted MLE with Intervals

< GLIM Program and Output for Estimation of Adjusted MLE >

```
[i] $FACT AGE 6 ALCGRP 4 TOBGRP 4
[i] $CAL  N=CASE+CONT :
[i]       AGE1=AGE :
[i]       AGE2=AGE1**2 : AGE3=AGE1**3 : AGE4=AGE1**4
[i] $YVAR CASE   $ERR B N
[o]
[i] $FIT  TOBGRP + ALCGRP + AGE1  $D E
[o] scaled deviance = 101.89 at cycle  4
[o]            d.f. =  80
[o]
[o]          estimate        s.e.      parameter
[o]     1      -5.904       0.4446     1
[o]     2      0.4567       0.2274     TOBG(2)
[o]     3      0.5483       0.2730     TOBG(3)
[o]     4       1.519       0.3242     TOBG(4)
[o]     5       1.531       0.2538     ALCG(2)
[o]     6       2.173       0.2854     ALCG(3)
[o]     7       3.599       0.3656     ALCG(4)
[o]     8      0.7676      0.08286     AGE1
[o]     scale parameter taken as  1.000
[o]
```

< Estimation of Adjusted MLE by EGRET >

OUTCOME= CASE

TERM	COEFFICIENT	STD ERROR	P-VALUE	ODDS RATIO
%GM	-5.904	(.445)	<.001	.2727E-02
TOBGRP='2'	.4567	(.228)	.045	1.579
TOBGRP='3'	.5483	(.273)	.045	1.730
TOBGRP='4'	1.519	(.324)	<.001	4.569
ALCGRP='2'	1.531	(.254)	<.001	4.624
ALCGRP='3'	2.173	(.286)	<.001	8.783
ALCGRP='4'	3.599	(.366)	<.001	36.56
AGE	.7676	(.829E-01)	<.001	2.155

DEVIANCE ON 80 DF = 101.889

LIKELIHOOD RATIO STATISTIC ON 8 DF = 628.213, p < .001

RESULTS [LR]

OUTCOME= CASE

TERM	ODDS RATIO	95% CONFIDENCE BOUNDS	
%GM	.2727E-02	.1140E-02	.6524E-02
TOBGRP='2'	1.579	1.011	2.466
TOBGRP='3'	1.730	1.013	2.955
TOBGRP='4'	4.569	2.420	8.627
ALCGRP='2'	4.624	2.811	7.606
ALCGRP='3'	8.783	5.019	15.37
ALCGRP='4'	36.56	17.85	74.88
AGE	2.155	1.831	2.535

Table. Adjusted risk of esophageal cancer associated with alcohol drinking

Alcohol drinking (g/d)	No. of cases	No. of controls	Adjusted OR (95% CI)[1)]
0 - 39	29	386	1.0
40 - 79	75	280	4.6 (2.81-7.61)
80 - 119	51	87	8.8 (5.02-15.4)
120+	45	22	36.6 (17.9-74.9)
			χ^2_{trend} = 122.1(1), $p<0.01$

1) Adjusted odds ratio and 95% confidence intervals were derived from regression coefficients and standard error in linear logistic models. Adjustment for age categories of linear term and categorized amount of cigarette smoking was done.

1) Model Selection Procedure

- 모델은 통계학적으로 적합하여야 하지만 생물학적으로도 잘 설명되는 모델이어야 한다

LR stastistic :

$$\Delta G = 2\ \Sigma\ (0 \log \frac{O}{E})$$
$$= 2\ \Sigma\ [n_1 \log \frac{n_1}{NP} + n_0 \log \frac{n_0}{N(1-P)}]$$

Goodness of fit :

$$G = \Sigma\ \frac{(O-E)^2}{E}$$
$$= \Sigma\ [\frac{(n_1-NP^2)}{NP} + \frac{\{n_0-N(1-P)\}^2}{N(1-P)}]$$

단, NP : expected value of cases
N(1-P) : expected value of controls

- fitted cell 이 5 이하만 아니면, 두 값은 서로 매우 근접
  → model 의 goodness of fit 에 이용 가능하다
  ↓
  departure from the model 에 비해서는 덜 민감

< GLIM Program and Output for Model Selection >

```
[i] File name? C:\PM\YOO\ESO3.GLM
[i] $FACT AGE 6 ALCGRP 4 TOBGRP 4
[i] $CAL  N=CASE+CONT
[i] $YVAR CASE   $ERR B N
[o]
[i] $FIT    AGE + ALCGRP + TOBGRP  $D E
[o] scaled deviance =  82.34 at cycle  5
[o]              d.f. =  76
[i] $PRINT 'PEARSON GOODNESS OF FIT = ' %X2
[o] PEARSON GOODNESS OF FIT =    86.46
[o]
[i] $FIT    AGE + TOBGRP  $D E
[o] scaled deviance = 210.27 at cycle  5
[o]              d.f. =  79
[i] $PRINT 'PEARSON GOODNESS OF FIT = ' %X2
[o] PEARSON GOODNESS OF FIT =    347.8
[o]
[i] $FIT    AGE + ALCGRP  $D E
[o] scaled deviance = 105.88 at cycle  5
[o]              d.f. =  79
[i] $PRINT 'PEARSON GOODNESS OF FIT = ' %X2
[o] PEARSON GOODNESS OF FIT =    117.4
[o]
[i] $FIT    AGE            $D E
[o] scaled deviance = 246.91 at cycle  5
[o]              d.f. =  82
[i] $PRINT 'PEARSON GOODNESS OF FIT = ' %X2
[o] PEARSON GOODNESS OF FIT =    362.5
```

Table. Summary of model fitting

M o d e l	Log-likelihood	Goodness of fit
AGE + ALCGRP + TOBGRP	82.34 (76) ←	86.46
AGE + TOBGRP	210.27 (79)	347.80
AGE + ALCGRP	105.88 (79)	117.40
AGE	246.91 (82)	362.50

2) Likelihood Ratio Test

: '그 변수가 들어있는 모델이 결과변수를 설명하는 능력이
그 변수가 들어있지 않는 모델이 결과변수를 설명하는 능력에
비해 더 많은 기여를 하고 있는가 ?'

$\triangle G$ = change in scaled deviance

= [LR without the variable] - [LR with the variable]

$$= [-2\log \frac{\ell(\beta) \text{ of the current model}}{\ell(\beta) \text{ of the saturated model}}] \text{ without the variable}$$

$$- [-2\log \frac{\ell(\beta) \text{ of the current model}}{\ell(\beta) \text{ of the saturated model}}] \text{ with the variable}$$

$$= [-2\log \frac{\ell(\beta) \text{ without the variable}}{\ell(\beta) \text{ with the variable}}]$$

* LR :
  - ① 특정 분포를 따르지 않음
  - ② goodness of fit 에 중요
  - ③ scaled deviance (GLIM), G (Breslow-Day)

* LR statistic :
  - ① χ^2 분포를 따름
  - ② goodness of fit 에 이용하려면 data 가 커야 함
    → nature of departure 가 더 민감한 방법
  - ③ change in scaled deviance (GLIM),
    $\triangle G$ (Breslow-Day)

Wald test

: standardized regression coefficient

H_0 : $\beta=0$ 일때

$$W = \frac{\beta_i}{SE(\beta_i)} \approx Nd$$

$$W^2 = (\frac{\beta_i}{SE(\beta_i)})^2 \approx \chi^2(1)$$

< GLIM Program and Output for Likelihood Ratio Test >

```
[i] File name? C:\PM\YOO\ES04.GLM
[o]
[i] $FIT    AGE + ALCGRP + TOBGRP  $D E
[o] scaled deviance =  82.34 at cycle  5
[o]             d.f. =  76
[i] $FIT  - ALCGRP  $D E
[o] scaled deviance = 210.27 (change =  +127.9) at cycle  5
[o]             d.f. =  79    (change =    +3  )
[o]
[i] $FIT    AGE + ALCGRP + TOBGRP  $D E
[o] scaled deviance =  82.34 at cycle  5
[o]             d.f. =  76
[i] $FIT  - TOBGRP  $D E
[o] scaled deviance = 105.88 (change =  +23.54) at cycle  5
[o]             d.f. =  79    (change =   +3   )
[o]
[i] $FIT    AGE + ALCGRP  $D E
[o] scaled deviance = 105.88 at cycle  5
[o]             d.f. =  79
[i] $FIT  - ALCGRP  $D E
[o] scaled deviance = 246.91 (change =  +141.0) at cycle  5
[o]             d.f. =  82    (change =    +3  )
[o]
[i] $FIT    AGE + TOBGRP  $D E
[o] scaled deviance = 210.27 at cycle  5
[o]             d.f. =  79
[i] $FIT  - TOBGRP  $D E
[o] scaled deviance = 246.91 (change =  +36.64) at cycle  5
[o]             d.f. =  82    (change =   +3   )
[o]
[i] $FIT  - AGE  $D E
[o] scaled deviance = 367.95 (change =  +121.0) at cycle  3
[o]             d.f. =  87    (change =    +5  )
[o]
[i] $STOP
```

Table. Summary of likelihood ratio tests

Model	Log-likelihood	ΔG (Δdf)	Interpretation
AGE + ALCGRP + TOBGRP	G_1 = 82.34 (76)	G_2-G_1 = 128(3)	ALCGRP ¦ AGE, TOBGRP
AGE + TOBGRP	G_2 = 210.27 (79)	G_3-G_1 = 24(3)	TOBGRP ¦ AGE, ALCGRP
AGE + ALCGRP	G_3 = 105.88 (79)	G_4-G_3 = 141(3)	ALCGRP ¦ AGE
AGE	G_4 = 246.91 (82)	G_4-G_2 = 37(3)	TOBGRP ¦ AGE
Null	G_5 = 367.95 (87)	G_5-G_4 = 121(5)	AGE

3) Quantitative Interpretation of Linear Logistic Model

【예 제】 음주와 식도암 발생과의 관련성을 검정하기 위한 환자-대조군 연구결과는 다음과 같았다.

Alcohol intake	CALC (100g/d)	ALCGRP	CASE	CONT	TOTAL
0-39(g/d)	0.2	1	29	386	415
40-79	0.6	2	75	280	355
80-119	1.0	3	51	87	138
120+	1.5	4	45	22	67
T o t a l			200	775	975

↑ midpoint coding (CALC)   ↑ qualitative coding (ALCGRP)

logit P_i = - 7.271
+ 1.851 AGE(2) + 3.610 AGE(3) + 4.160 AGE(4)
+ 4.696 AGE(5) + 4.665 AGE(6)
+ 2.548 CALC   (SE_{ALC} = 0.2504)
+ 0.409 CTOB   (SE_{TOB} = 0.08718)

< 해 석 > exp(2.548) = 12.78 (per 100g/d)
→ ALC 의 단위 증가량 증가 (10g/d) 에 따른 위험도는
= exp(2.548/10) = 1.29
95% CI = exp(2.548/10 ± 1.96*0.2504/10)
= 1.22-1.36

exp (0.409) = 1.51 (per 10g/d)
→ TOB 의 단위 증가량 (10g/d) 에 따라 1.51 배씩 RR 증가
95% CI = exp(0.409 ± 1.96*0.08718)
= 1.27-1.79

4) Joint Odds Ratio Estimation

< GLIM Program and Output for Joint Odds Ratio Estimation >

```
[i] File name? C:\PM\YOO\ES06.GLM
[i] $FACT AGE 6 ALCGRP 4 TOBGRP 4
[i] $CAL  N=CASE+CONT : CALC=ALCGRP : CTOB=TOBGRP
[i] $YVAR CASE    $ERR B N
[o]
[i] $FIT    AGE + ALCGRP + TOBGRP   $D E V C
[o] scaled deviance =  82.34 at cycle  5
[o]            d.f. =  76
[o]
[o]           estimate        s.e.      parameter
[o]     1       -6.895       1.083      1
[o]     2        1.981       1.102      AGE(2)
[o]     3        3.776       1.065      AGE(3)
[o]     4        4.335       1.062      AGE(4)
[o]     5        4.896       1.074      AGE(5)
[o]     6        4.827       1.119      AGE(6)
[o]     7        1.435       0.2501     ALCG(2)
[o]     8        1.981       0.2848     ALCG(3)
[o]     9        3.603       0.3850     ALCG(4)
[o]    10       0.4381       0.2283     TOBG(2)
[o]    11       0.5126       0.2730     TOBG(3)
[o]    12        1.641       0.3441     TOBG(4)
[o]     scale parameter taken as  1.000
[o]
[o] (Co)variances of parameter estimates
[o]   7      0.06253
[o]   8      0.04266     0.08109
[o]   9      0.04406     0.04401      0.1482
[o]  10  -0.0004504   -0.005430   -0.001091     0.05213
[o]  11   -0.004501   -0.001079   -0.003470     0.02038     0.07451
[o]  12    0.005281    0.005291    0.008115     0.02107     0.02136     0.1184
[o]              7           8           9          10          11         12
[o]     scale parameter taken as  1.000
[i] $STOP
```

ALCGRP(2) 이고 TOBGRP(3) 인 경우의 joint RR 은 ?

$$\text{logit } P_i = -6.895 + 1.981\ AGE(2) + 3.776\ AGE(3) + 4.335\ AGE(4) + 4.896\ AGE(5) + 4.827\ AGE(6) + \underline{1.435\ ALCGRP(2)} + 1.981\ ALCGRP(3) + 3.603\ ALCGRP(4) + 0.438\ TOBGRP(2) + \underline{0.513\ TOBGRP(3)} + 1.641\ TOBGRP(4)$$

$\underline{SE_{ALCGRP(2)}} = 0.2501$, $\underline{SE_{TOBGRP(3)}} = 0.2730$

$COV_{ALCGRP(2)*TOBGRP(3)} = -0.004501$

RR (ALCGRP=2, TOBGRP=3 ; all)

$$= \exp[\beta_{ALCGRP(2)} \cdot (X_2^* - X_2) + \beta_{TOBGRP(3)} \cdot (X_3^* - X_3)]$$
$$= \exp[1.435 \times 1 + 0.5126 \times 1]$$
$$= 7.0$$

$$Var = SE[\beta_{ALCGRP(2)}]^2 + SE[\beta_{TOBGRP(3)}]^2 + 2 \cdot COV[\beta_{ALCGRP(2)} \cdot \beta_{TOBGRP(3)}]$$
$$= 0.2501^2 + 0.2730^2 + 2 \cdot (-0.004501)$$
$$= 0.128$$

$$SE[\beta_{ALCGRP(2)} \cdot \beta_{TOBGRP(3)}] = \sqrt{0.128} = 0.3578$$

$$95\%\ C.I. = \exp[1.9476 \pm 1.96 * 0.3578] = [3.48 \sim 14.14]$$

5) Quantitative Interpretation of Non-linear Logistic Model

logit Pi = ...	
$- 0.2808 \log(DUBF+1)$	$(SE_{\log(DUBF+1)} = 0.1225)$
$+ 1.6560$ FHX	$(SE_{FHX} = 1.1300)$
$+ 0.0634 \log(FFTP+1)$	$(SE_{\log(FFTP+1)} = 0.0316)$
$+ 0.0681$ CLVB	$(SE_{CLVB} = 0.1454)$

DUBF 1명 증가에 따른 위험도의 증가는 ?   95% 신뢰구간은 ?

$$RR = (DUBF + 1)^{-0.2808}$$
$$= 2^{-0.2808} = 0.8231$$

$$95\%\ CI = 2^{(-0.2808 \pm 1.96*0.1225)}$$
$$= [0.70 \sim 0.97]$$

CLVB 5명 증가에 따른 위험도의 증가는 ? 95% 신뢰구간은 ?

$$RR = \exp [5 \times 0.0681] = 1.40$$

$$95\%\ CI = \exp [5 \times 0.0681 \pm 1.96 \times 5 \times 0.1454]$$
$$= [0.34 \sim 5.84]$$

Conditional Linear Logistic Regression Analysis

1) Estimation of Conditional MLE using Linear Logistic Model to Fit 1:1 Matched Data

1:1 로 짝지은 이분성인 폭로요인 X 에 대해 변수 X_{i0} 가 환자군에서 관찰되며 동시에 변수 X_{i1} 가 대조군에서 관찰되는 조건부 확률은 다음과 같다 (Holford 등 1978).

$$\frac{\exp[(X_{i0} - X_{i1})*\beta]}{1 + \exp[(X_{i0} - X_{i1})*\beta]}$$

이 조건부 확률은 다음과 같은 수학적 변환을 거치면 linear logistic model 의 형태를 따르게 된다.

① covariates are difference between case and control : $(X_{i0} - X_{i1})$
② conditioning of every outcome to be success : Y = 1
③ dropping intercept term from the model

< GLIM Program and Output for Conditional Linear Logistic Regression >

```
$UNIT 75
$DATA GR1 SEX1 MRA1 QBY1 STL1 MAG1 FAG1
      GR2 SEX2 MRA2 QBY2 STL2 MAG2 FAG2
$READ
1 1 1 1 0 19 28 2 1 1 1 0 19 28
1 1 1 1 0 32 37 2 1 1 1 0 24 25
. . . . .  .  . . . . . .  .  .
$CAL STLB = STL1 - STL2 :
     MAGE = MAG1 - MAG2 : FAGE = FAG1 - FAG2 :  N = 1
$YVAR N    $ERR  B N
$FIT - %GM + STLB + MAGE + FAGE     $D E
$FIT - STLB                         $D E
$STOP
```

< Conditional Linear Logistic Regression by EGRET>

Data file name: C:\PM\YOO\RHABD

8 variables and 166 observations

VARIABLES

1. MATCH	2. DISEASE	3. SEX	4. RACE	5. QBYR
6. STLBN	7. MAGE	8. FAGE		

ANALYSIS MODEL: Conditional logistic regression
RISK TYPE: Relative risk (multiplicative)
MATCHING: Match on the Values of Variable: MATCH
OUTCOME SPECIFICATION: Outcome Variable Name: DISEASE

FACTORED VARIABLES:

VARIABLE	#LEVELS	BASE
STLBN	2	0

REGRESSION TERMS [CLR]

a. STLBN b. MAGE c. FAGE

[Using 83 matched sets]

RESULTS [CLR]

OUTCOME= DISEASE

TERM	COEFFICIENT	STD ERROR	P-VALUE	ODDS RATIO
STLBN='1'	-1.385	(.502)	.006	.2503
MAGE	.6404E-01	(.553E-01)	.247	1.066
FAGE	-.6473E-01	(.520E-01)	.213	.9373
	DEVIANCE =	102.785		

LIKELIHOOD RATIO STATISTIC ON 3 DF = 12.278, p = .006

RESULTS [CLR]

OUTCOME= DISEASE

TERM	ODDS RATIO	95% CONFIDENCE BOUNDS	
STLBN='1'	.2503	.9356E-01	.6697
MAGE	1.066	.9566	1.188
FAGE	.9373	.8465	1.038

Application of Log-Linear Model

< GLIM Program and Output for Log-Linear Regression Model >

```
[i] File name? C:\PM\YOO\FRAMING.GLM
[o]
[i] $UNITS 32
[i] $DATA BP CHOL DIS FREQ
[i] $READ
[i] 1 1 2   2   2 1 2   3   3 1 2  3   4 1 2  4
[i] 1 2 2   3   2 2 2   2   3 2 2  0   4 2 2  3
    . . .   .   . . .   .   . . .  .   . . .  .
[i] $FACTOR BP 4 CHOL 4 DIS 2
[i] $YVAR FREQ
[i] $ERR  P
[o]
[i] $FIT BP*CHOL*DIS - BP.CHOL.DIS      $D E
[o] scaled deviance = 8.0762 at cycle  4
[o]            d.f. = 9
[o]
[o]          estimate         s.e.      parameter
[o]     1       4.749      0.09226      1
[o]     2     0.04237       0.1286      BP(2)
[o]     3     -0.8879       0.1692      BP(3)
[o]     4      -1.588       0.2184      BP(4)
[o]     5     -0.2962       0.1411      CHOL(2)
[o]     6     0.04280       0.1283      CHOL(3)
[o]     7     -0.5562       0.1503      CHOL(4)
[o]     8      -3.482       0.3486      DIS(2)
[o]     9     0.08645       0.1945      BP(2).CHOL(2)
[o]    10      0.5092       0.1701      BP(2).CHOL(3)
[o]    11      0.3674       0.1988      BP(2).CHOL(4)
[o]    12      0.1547       0.2512      BP(3).CHOL(2)
[o]    13      0.3124       0.2235      BP(3).CHOL(3)
[o]    14      0.5553       0.2462      BP(3).CHOL(4)
[o]    15      0.1911       0.3200      BP(4).CHOL(2)
[o]    16      0.5233       0.2757      BP(4).CHOL(3)
[o]    17      0.8491       0.2936      BP(4).CHOL(4)
[o]    18    -0.04146       0.3037      BP(2).DIS(2)
[o]    19      0.5324       0.3324      BP(3).DIS(2)
[o]    20       1.200       0.3269      BP(4).DIS(2)
[o]    21     -0.2080       0.4664      CHOL(2).DIS(2)
[o]    22      0.5622       0.3508      CHOL(3).DIS(2)
[o]    23       1.344       0.3430      CHOL(4).DIS(2)
[o]     scale parameter taken as  1.000
[o]
[i] $STOP
```

< GLIM Program and Output for Constant Hazard Model >

```
[i] File name? C:\PM\YOO\MIRATE.GLM
[o]
[i] $UNITS 15
[i] $DATA FAIL PT
[i] $READ
[i] 2 4056  4 2892  2 1728  2 4172  3 3467   1 2183   7 4616  6 3835
[i] 6 3102  2  888  5 1101  2 1674  6 2456  10 3740  46 7269
[i] $CAL SMOK =%GL(5,3)   :   LEUK=%GL(3,1) :
[i]      LPT=%LOG(PT) : RATE =1000*FAIL/PT : LRATE=%LOG(RATE)
[i] $FACT SMOK 5 LEUK 3
[o]
[i] $TPRINT (S=1) RATE;FAIL;PT LEUK;SMOK
[o]             +-----------------------------------------------+
[o]        SMOK !    1        2        3        4        5     !
[o]   LEUK      !                                              !
[o] +-----------+-----------------------------------------------+
[o] !    1 RATE !  0.4931   0.4794   1.5165   2.2523   2.4430  !
[o] !      FAIL !   2.000    2.000    7.000    2.000    6.000  !
[o] !        PT !  4056.0   4172.0   4616.0    888.0   2456.0  !
[o] -----------------------------------------------------------
[o] !    2 RATE !  1.3831   0.8653   1.5645   4.5413   2.6738  !
[o] !      FAIL !   4.000    3.000    6.000    5.000   10.000  !
[o] !        PT !  2892.0   3467.0   3835.0   1101.0   3740.0  !
[o] -----------------------------------------------------------
[o] !    3 RATE !  1.1574   0.4581   1.9342   1.1947   6.3282  !
[o] !      FAIL !   2.000    1.000    6.000    2.000   46.000  !
[o] !        PT !  1728.0   2183.0   3102.0   1674.0   7269.0  !
[o] +-----------+-----------------------------------------------+
[o]
[i] $YVAR  FAIL
[i] $ERR   P
[i] $OFFSET LPT                                          ←  OFFSET
[o]
[i] $FIT SMOK + LEUK  $D E
[o] scaled deviance =  9.704 at cycle  4
[o]             d.f. =  8
[o]
[o]          estimate        s.e.      parameter
[o]     1      -7.246       0.3917     1
[o]     2     -0.4329       0.5402     SMOK(2)
[o]     3      0.5313       0.4219     SMOK(3)
[o]     4      0.8205       0.4901     SMOK(4)
[o]     5       1.405       0.3842     SMOK(5)
[o]     6      0.2890       0.2992     LEUK(2)
[o]     7      0.6470       0.2763     LEUK(3)
[o]     scale parameter taken as  1.000
[o]
[i] $STOP
```

< GLIM Program and Output for Proportional Hazard Model >

```
[i] File name? C:\PM\YOO\COX.GLM
[o]
[i] $UNIT 135
[i] $DATA TIME                                    ←  FAILURE TIME
[i] $READ
[i] 157 134 212 77 71 130 21 22 101
[i] 139 110 136 68 63  72 17 18  63
      .   .   .   .   .   .   .   .   .
[o]
[i] $DATA DTHS                                    ←  NUMBER OD DEATH
[i] $READ
[i]  9 12 42 5 4 28 1 1 19
[i]  2  7 26 2 3 19 1 1 11
     .  .  .  . .  .  . . .
[i] $CAL STAG=%GL(3,1) : HIST=%GL(3,3) : FLUP=%GL(15,9) :
[i]      LTIM=%LOG(TIME) :
[i]      WT=%GT(TIME,0)
[i] $FACT STAG 3 HIST 3 FLUP 15
[i] $YVAR DTHS
[i] $ERR  P
[i] $OFFSET LTIM                                  ←  OFFSET
[i] $WEIGHT WT
[o]
[i] $FIT STAG*HIST + FLUP  $D E
[o]
[i] $STOP
```