역학 연수교육 II

역학자료분석에서 통계 패키지 이용
- PC-SAS SYSTEM -

신 해 림
동아대학교 의과대학

PC-SAS를 이용한 역학자료의 분석은 Statistical Methods in Cancer Research Vol I- The analysis of case-control study (Breslow and Day, 1980) 의 부록 I (grouped data from Ille-Vilaine study of oesophageal cancer used for illustraton in chapter 4 and 6)의 자료를 이용하여 Chapter 4 Classical Methods of analysis of Grouped Data 와 Chapter 6 Unconditional Logistic Regression for Large Strata 에서의 표들과 SAS OUTPUT 을 비교하면서 다음의 순서에 따라 설명하고자 한다.

1. STATISTICAL INFERENCE FOR TWO-BY-TWO TABLE (PROC FREQ)

1) Hypothesis test : Pearson's chi-square test
likelihood ratio chi-square test
Fisher's exact test
2) Point estimation : relative risk
odds ratio
3) Interval estimation : logit interval
test-based interval

2. STRATIFIED ANALYSIS (PROC FREQ)

1) Test for homogeneity : Breslow and Day's test
2) Hypothesis test : Cochra-Mantel-Haenzel test
3) Point estimation (common RR / OR)
   : Mantel-Haenszel method
   : logit method
4) Interval estimation : test-based interval
   : logit interval

3. QUALITATIVE ANALYSIS WITH LINEAR LOGISTIC MODEL (PROC LOGISTIC)

1) Model selection : improvement in log-likelihood
   goodness of fit
2) Hypothesis test : likelihood ratio test
   wald test
3) Point estimation : maximum likelihood estimate
4) Interval estimation

The Ille-et-Vilaine study of oesophageal cancer는 프랑스 서북부 반도지역인 Brittany에서 Tuyns et al (1977)등이 1972년 1월부터 1974년 4월까지 그지역의 한 병원에서 식도암으로 진단받은 200명의 남자 환자와 그지역의 각각의 commune에서 작성한 선거명부에서 남자 778명의 표본중 775명을 대조군으로 한 환자-대조군 연구이다.

연구 대상자들로 부터 여러가지 음식, 흡연, 술종류별 음주 등에 대한 면접조사를 하여 자료를 얻었으며, 여기서는 식도암의 위험인자로서 음주부분만 자세히 분석하고자 한다. 원래의 자료에서는 연령(years), 흡연(gram/day) ,음주(gram/day) 등이 모두 연속변수로 측정되었으나 age는 25-34, 35-44, 45-54, 55-64, 65-74, 75+ 의 6개 군, alcohol(g/day)는 0-39, 40-79, 80-119, 120+ 의 4개군, tobacco는 0-9, 10-19, 20-29, 30+의 4개군으로 나눈 grouped data가 사용되었다. 환자군과 대조군의 분포는 표4.1과 같다 (표의 순서는 비교를 위하여 Breslow and Day 책의 표명을 그대로 인용

하기로 한다).

data 입력을 위한 SAS program 과 output은 다음과 같다.

====== PROGRAM 1 ======

```
OPTION PS=300 LS=80;
LIBNAME SHIN 'A:\';
 DATA SHIN.ESOPH;
 INFILE 'A:\ESOPH.DAT';
 INPUT AGE   ALCOHOL   TOBACCO  ESOPH WT @@;
 RUN;

PROC FORMAT;
      VALUE A 1='25-34' 2='35-44' 3='45-54' 4='55-64'
            5='65-74' 6='75+';
PROC FORMAT;
      VALUE D 1='0-39' 2='40-79' 3='80-119' 4='120+';
PROC FORMAT;
      VALUE S 1= '0-9' 2='10-19' 3='20-29' 4='30+';
PROC FORMAT;
      VALUE C 1='case' 0='control';

PROC FREQ;
  TABLES (AGE ALCOHOL TOBACCO) * ESOPH
            /NOROW NOCOL NOPERCENT EXPECTED CHISQ;
  WEIGHT WT;
  FORMAT AGE A.; FORMAT ALCOHOL D.; FORMAT TOBACCO S.;
  FORMAT FORMAT C.;
  TITLE 'DISTRIBUTION OF RISK FACTORS FOR CASES AND
CONTROLS:
  ILLE-VILAINE STUDY OF ESOPHAGEAL CANCER';
RUN;
```

===== OUTPUT 1 ======

Distribution of risk factors for cases and controls

TABLE OF AGE BY ESOPH

AGE	ESOPH		
Frequency Expected	control	case	Total
25-34	115 92.205	1 23.795	116
35-44	190 158.18	9 40.821	199
45-54	167 169.31	46 43.692	213
55-64	166 192.36	76 49.641	242
65-74	106 127.97	55 33.026	161
75+	31 34.974	13 9.0256	44
Total	775	200	975

FOR TABLE OF AGE BY ESOPH

Statistic	DF	Value	Prob
Chi-Square	5	97.036	0.000
Likelihood Ratio Chi-Square	5	121.045	0.000

Sample Size = 975

TABLE OF ALCOHOL BY ESOPH

ALCOHOL / ESOPH Frequency Expected	control	case	Total
0-39	386	29	415
	329.87	85.128	
40-79	280	75	355
	282.18	72.821	
80-119	87	51	138
	109.69	28.308	
120+	22	45	67
	53.256	13.744	
Total	775	200	975

STATISTICS FOR TABLE OF ALCOHOL BY ESOPH

Statistic	DF	Value	Prob
Chi-Square	3	158.955	0.000
Likelihood Ratio Chi-Square	3	146.498	0.000

TABLE OF TOBACCO BY ESOPH

TOBACCO Frequency Expected	ESOPH control	case	Total
0-9	447	78	525
	417.31	107.69	
10-19	178	58	236
	187.59	48.41	
20-29	99	33	132
	104.92	27.077	
30+	51	31	82
	65.179	16.821	
Total	775	200	975

STATISTICS FOR TABLE OF TOBACCO BY ESOPH

Statistic	DF	Value	Prob
Chi-Square	3	29.357	0.000
Likelihood Ratio Chi-Square	3	27.847	0.000

1. STATISTICAL INFERENCE FOR TWO-BY-TWO TABLE

2X2 표의 분석을 위하여 음주량을 80g/day 미만과 이상인 군으로 나누면 다음과 같다.

	Average daily alcohol consumption		
	80+ g	0-79 g	Total
Cases	96	104	200
Controls	109	666	775
Total	205	770	975

:==== PROGRAM 2 ======

```
DATA ALC; SET SHIN.ESOPH;

 IF ALCOHOL <= 2 THEN ALC=1;
 IF ALCOHOL >= 3 THEN ALC=2;

 PROC FORMAT ;
   VALUE AL 1=' 1-79G' 2='80+G';

 PROC FREQ;
   WEIGHT WT;
  TABLES ALC*ESOPH /NOROW NOPERCENT NOCOL EXPECTED CMH
   CHISQ;
   TITLE ' 2 x 2 TABLE (ALCOHOL CONSUMPTION : DICHOTOMIZED)';
   FORMAT ALC AL.; FORMAT AGE A.; FORMAT ESOPH C.;
   RUN;
```

:==== OUTPUT 2 ======

2 x 2 table (alcohol consumption : dichotomized)

TABLE OF ALC BY ESOPH

ALC	ESOPH		
Frequency¦			
Expected	¦control	¦case	¦ Total
1-79g	¦ 666	¦ 104	¦ 770
	¦ 612.05	¦ 157.95	¦
80+g	¦ 109	¦ 96	¦ 205
	¦ 162.95	¦ 42.051	¦
Total	775	200	975

STATISTICS FOR TABLE OF ALC BY ESOPH

Statistic	DF	Value	Prob
Chi-Square	1	110.255	0.000
Likelihood Ratio Chi-Square	1	96.433	0.000
Continuity Adj. Chi-Square	1	108.221	0.000
Fisher's Exact Test (Left)			1.000
(Right)			1.03E-22
(2-Tail)			1.08E-22

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Bounds	
Case-Control (Odds Ratio)	Mantel-Haenszel	5.640	4.083	7.791
	Logit	5.640	4.001	7.951

The confidence bounds for the M-H estimates are test-based.

Total Sample Size = 975

====== INTERPRETATION ======

1) Hypothesis test

종속변수와 독립변수 사이에 연관성이 있는가 (독립적인가 아닌가의 검정)를 알기 위한 가설검정은 위의 결과에서 (Pearson's) chi-square statistic = 110.255 (자유도 1), likelihood rato chi-square statistic = 96.433 (자유도 1), continuity adjusted chi-square statistics = 108.221 (자유도 1)로 각각의 유의수준은 모두 p < 0.000으로 식도암의 유무과 음주와는 유의한 차이가 있다고 할 수 있다. 위의 결과에서의 Fisher's exact test는 '두변수는 서로 무관하다 ({{H_0~ :~ ~=~1}} {{}})' 라는 가정하에서만 성립하는 것으로 어떤 한 cell의 기대값이 5보다 작은 경우에 적용된다.

2) Point estimation

2x2표에서 질병의 비교위험도는 RR = Prob (D=yes|E=no)/Prob (D=yes|E=no)로 정의되며, 환자-대조군 연구에 있어서는 odds ratio로 relative risk 의 추정치를 대신한다. OR= n11 n22/ n12 n21 로 계산된 것으로 위의 결과에서는 Mantel-Haenszel 방법으로 구한 OR=5.64로 이는 하루에 80g 이상의 술을 마시는 사람은 80g 미만을 마시는 사람보다 식도암에 걸릴 위험이 5.64배 높다고 할 수 있다.

3) Interval estimation

신뢰구간은 test-based method에 의한 경우 (4.001, 7.951), logit method에 의한 경우 (4.083, 7.791)로 1을 포함하지 않으므로 유의하다고 하겠다.

2. STRATIFIED ANALYSIS

1) COMBINATION OF RESULTS FROM A SERIES OF 2x2 TABLES ; CONTROL OF CONFOUNDING

confounding factor의 수준에 따른 2x2 표가 여러개 있을 때는 표본을 confounding factor의 값이 같은 수준으로 나누어 각 strata내에서의 separated relative risk 를계산하여야 bias가 없게 된다. 이때는 폭로와 질병의 연관성이 각 strata별로 일정한가를 알아야 하고, 일정하지 않다면 relative risk 가 stratum에 따라 어떻게 변하는가를 아는 것이 중요하다. 이러한 경우 분석을 할 때는 다음의 3가지를 하여야 한다.

(1) a test of the null hypothesis that =1 in all tables
(2) point and interval estimation of assumed to be common to all tables
(3) a test of the homogeneity or no-interaction hypothesis that is constant across tables

여기서 쓰인 예는 거의 모든 암발생에서 potential confounding factor인 연령을 고려하고자 한다. 표4.1에서 환자군의 평균연령은 60.0세, 대조군의 평균연령은 50.2세로 환자군의 연령이 약 10세 가량 많다. 만약 연령이 음주와 관련이 있다면 age-adjusted relative risk는 변화할 것이다. 그러나 표4.2에서 연령과 alcohol 섭취량과의 상관성은 상관계수 -0.02로 약한 역상관관계이므로 이 경우의 confounding effect는 그리 크지 않을 것임을 알 수 있다.

==== PROGRAM 3 ======

```
/* AGE ADJUSTED RELATIVE RISK를 구하기 위한 PROGRAM */

PROC FREQ DATA=ALC;
  WEIGHT WT;
  TABLES  AGE*ALC*ESOPH
        /NOROW NOPERCENT NOCOL EXPECTED CMH;
  RUN;
```

====== OUTPUT 3 ======

	ALC	ESOPH		
	Frequency Expected	control	case	Total
AGE 25-34	1-79g	106 105.09	0 0.9138	106
	80+g	9 9.9138	1 0.0862	10
	Total	115	1	116
AGE 35-44	1-79g	164 161.36	5 7.6432	169
	80+g	26 28.643	4 1.3568	30
	Total	190	9	199
AGE 45-54	1-79g	138 124.66	21 34.338	159
	80+g	29 42.338	25 11.662	54
	Total	167	46	213
AGE 55-54	1-79g	139 118.67	34 54.331	173
	80+g	27 47.331	42 21.669	69
	Total	166	76	242
AGE 65-74	1-79g	88 81.64	36 42.36	124
	80+g	18 24.36	19 12.64	37
	Total	106	55	161
AGE 75+	1-79g	31 27.477	8 11.523	39
	80+g	0 3.5227	5 1.4773	5
	Total	31	13	44

SUMMARY STATISTICS FOR ALC BY ESOPH
CONTROLLING FOR AGE

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	85.009	0.000
2	Row Mean Scores Differ	1	85.009	0.000
3	General Association	1	85.009	0.000

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Bounds	
Case-Control	Mantel-Haenszel	5.158	3.639	7.310
(Odds Ratio)	Logit *	5.100	3.512	7.407

The confidence bounds for the M-H estimates are test-based.
* denotes that the logit estimators use a correction
of 0.5 in every cell of those tables that contain a zero.

Breslow-Day Test for Homogeneity of the Odds Ratios
Chi-Square = 9.323 DF = 5 Prob = 0.097

Total Sample Size = 975

```
====== PROGRAM 4 ======
 /* AGE GROUP별로 분석하기 위하여 AGE=1, AGE=2,AGE=3, AGE=4, AGE=5
AGE=6의 */
 /*  6개의 DATA SET을 만들어 분석하는 PROGRAM  */
   DATA AGE1; SET ALC;
      IF AGE=1;
      PROC FREQ;
         WEIGHT WT;
         TABLES ALC*ESOPH/NOROW NOPERCENT NOCOL EXPECTED CMH;
         TITLE 'STRATIFICATION BY AGE GROUP=1';
         FORMAT ALC AL.;  FORMAT ESOPH C.;
         RUN;

====== OUTPUT 4 ======
              Estimates of the Common Relative Risk (Row1/Row2)
                                                              95%
         Type of Study     Method                Value    Confidence Bounds
        ---------------------------------------------------------------
 AGE GROUP=1
       Case-Control      Mantel-Haenszel
         (Odds Ratio)    Logit *              33.632      1.280   883.732
 AGE GROUP=2
       Case-Control      Mantel-Haenszel       5.046      1.428    17.826
         (Odds Ratio)    Logit                 5.046      1.272    20.025
 AGE GROUP=3
       Case-Control      Mantel-Haenszel       5.665      2.906    11.042
         (Odds Ratio)    Logit                 5.665      2.799    11.464
 AGE GROUP=4
       Case-Control      Mantel-Haenszel       6.359      3.552    11.387
         (Odds Ratio)    Logit                 6.359      3.449    11.726
 AGE GROUP=5
       Case-Control      Mantel-Haenszel       2.580      1.229     5.418
         (Odds Ratio)    Logit                 2.580      1.216     5.475
 AGE GROUP=6
       Case-Control      Mantel-Haenszel
         (Odds Ratio)    Logit *              40.765      2.045   812.690

       The confidence bounds for the M-H estimates are test-based.
       To avoid undefined results, some estimates are not computed.

       * denotes that the logit estimators use a correction
            of 0.5 in every cell of those tables that contain a zero.
```

====== INTERPRETATION ======

1) Test for homogeneity : Breslow and Day's test

영가설은 각 strata에서의 odds ratio 는 일정하다는 것으로 Breslow and Day test에서는 각 stratum 의 sample size가 커야한다. 위의 결과는 (Breslow-Day Test for Homogeneity of the Odds Ratios Chi-Square = 9.323 DF = 5 Prob = 0.097) 연령군별로 odds ratio 가 다름을 알 수 있다.

2) Hypothesis test : Cochran-Mantel-Haenszel test

영가설이 any strata내에서도 독립변수와 종속변수사이에 연관성이 없다는 것으로 some strata내에서의 연관성의 pattern이 다른 strata에 의한 pattern과 반대 방향이면 이런 형태의 연관성을 알기는 어렵다. 그래서 CMH statistic가 유의하지 않다는 것은 연관성이 없거나 다른 pattern 보다 우세하다고할 만한 strength or consistency를 가지는 연관성은 없다는 것을 의미한다. 출력결과는 (1) The correlation statistic: Mantel-Haenszel procedure 에 의한 stratum-adjusted correlation statistic (df=1) (2) The ANOVA statistic : Stratum-adjusted ANOVA or Kruskal Wallis test (df=R-1 : R= number of row) (3) The general association statistic : stratum-adjusted Pearson's chi-square statistic (df=(R-1)(C-1)): 의 3가지를 보여준다. 두번째는 최종분석대상 변수가 2x2인 경우에 한해서만 출력되는데, 코호트연구에서 이용되는 relative risk와 환자-대조군 연구에서 사용되는 odds ratio, 이들 지표들의 95%신뢰구간이 계산되어 나온다.

위의 결과에서는 각 연령군별로 3개의 통계량 모두 유의하다.

3) Point estimation (Mantel-Haenszel method, logit method)

age adjusted relative risk estimator는 output 3에서 Mantel- Haenszel 방법으로 추정된 OR= 5.158로 술을 80g/day이상 마시는 사람은 80g/day미만 마시는 사람에 비하여 식도암에 걸릴 위험이 약 5배 높다고 할 수 있다. Logit method에 의한 추정값은 5.100 로 비슷함을 알 수 있다. Logit method에서는 cell의 값이 0(zero)인 경우 1/2을 더하여 계산한다.

각 연령층에서의 odds ratio를 알기 위해서는 연령군별로 자료를 나누어 계산하여야 하며 그 결과는 output 4 이다. 이는 odds ratio 와 confidence interval만 편집한 것으로 실제 SAS Output은 더 많은 통계량들을 모두 출력해준다.

4) Interval estimation : test-based interval, logit interval

Mantel- Haenszel 방법은 test-based interval로 95% CI= 3.639, 7.310이며, Logit intervl은 95% CI= 3.512, 7.407로 비슷하다.

3. QUALITATIVE ANALYSIS WITH LINEAR LOGISTIC MODEL (PROC LOGISTIC)

환자-대조군 연구이든, 코호트 연구이든 역학적 연구를 수행하면서 질병의 위험(risk)에 영향을 미치는 변수들에 대한 자료를 모으게 되면, 각각의 변수들의 수준에 따른 질병발생의 확률을 추정하게 된다. 예를 들어서 55세 된 남자로 30년간 전화회사에 근무하였고, 10대 후반 부터 하루에 담배를 한갑 피운 사람의 폐암에 걸릴 위험은 얼마나 되는가에 대해서 알고자 한다. 이를 알기위해 수학적인 식을 통해 각각의 위험요인들을 조합한 질병의 위험률을 구하는 것을 modelling이라 한다. 모델은 여러가지 위험요인들의 관계를 간단하고 정성적으로 기술하고, 질병의 발생확률을 알게 해준다. 즉 관찰수가 작은 category에서도 risk를 예측할 수 있게 한다. 또한 중요한 것은 결과를 단순하면서도 즉시 해석이 가능하게 relative risk를 제공해주어야 한다. 이러한 요구를 만족하는 모델이 linear logistic model이다.

P를 disease risk 라 하면 logit transform y 는 다음과 같다.
$y = \text{logit } P = \log [P/(1-P)]$, $\text{logit } P(x) = \alpha + \beta x$
이때 $OR = \exp(\beta)$가 된다. (자세한 풀이는 생략함)

여기서는 Breslow and Day (1980)의 제 6장에 있는 표6.1을 이용한 분석을 통하여 설명하고자 한다.

```
====== PROGRAM 5 ======

  DATA TAB6_1;
  INPUT AGE EXPOSURE CASES TOTAL @@;
  CARDS;
  1 1 1 10 1 0 0 106 2 1 4 30 2 0 5 169
 3 1 25 54 3 0 21 159 4 1 42 69 4 0 34 173
 5 1 19 37 5 0 36 124 6 1 5 5 6 0 8 39
;

  DATA  T1; SET TAB6_1 ;
  IF AGE=2 THEN AGEST1=1; ELSE AGEST1=0;
  IF AGE=3 THEN AGEST2=1; ELSE AGEST2=0;
  IF AGE=4 THEN AGEST3=1; ELSE AGEST3=0;
  IF AGE=5 THEN AGEST4=1; ELSE AGEST4=0;
  IF AGE=6 THEN AGEST5=1; ELSE AGEST5=0;
   AGEALC=AGE*EXPOSURE;

  PROC LOGISTIC DATA=T1 ;
     MODEL CASES/TOTAL=AGEST1 AGEST2 AGEST3 AGEST4 AGEST5 ;
     TITLE 'MODEL 1';
     OUTPUT OUT T1_OUT1  STDXBETA=STDB  XBETA = B ;    RUN;
     PROC PRINT DATA=T1_OUT1;RUN;

  PROC LOGISTIC DATA=T1 ;
     MODEL CASES/TOTAL=AGEST1 AGEST2 AGEST3 AGEST4 AGEST5
EXPOSURE ;
     TITLE 'MODEL 2';
     OUTPUT OUT= T1_OUT2 STDXBETA=STDB XBETA=B;
     PROC PRINT DATA=T1_OUT2;

  PROC LOGISTIC DATA=T1 ;
      MODEL CASES/TOTAL=AGEST1 AGEST2 AGEST3 AGEST4 AGEST5
EXPOSURE  AGEALC;
      TITLE 'MODEL 3';
      OUTPUT OUT= T1_OUT3 STDXBETA=STDB XBETA=B;
      PROC PRINT DATA=T1_OUT3;
  RUN;
```

====== OUTPUT 5 =====

/* MODEL 1 OUTPUT : 연령군만 포함된 모델 */

Data Set: WORK.T1
Response Variable (Events): CASES
Response Variable (Trials): TOTAL
Number of Observations: 12
Link Function: Logit

Response Profile

Ordered Value	Binary Outcome	Count
1	EVENT	200
2	NO EVENT	775

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	991.488	880.444	.
SC	996.371	909.739	.
-2 LOG L	989.488	868.444	121.045 with 5 DF (p=0.0001)
Score	.	.	97.036 with 5 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	-4.7449	1.0043	22.3206	0.0001	.
AGEST1	1.6951	1.0607	2.5541	0.1100	0.376869
AGEST2	3.4556	1.0180	11.5216	0.0007	0.787621
AGEST3	3.9637	1.0138	15.2849	0.0001	0.944467
AGEST4	4.0888	1.0180	16.1329	0.0001	0.837438
AGEST5	3.8759	1.0573	13.4387	0.0002	0.443815

Antilogs of ML estimates with confidence intervals

Variable	exp(beta)	95 percent confidence interval		90 percent confidence interval	
AGEST1	5.4472	0.6812	43.5571	0.9464	31.3513
AGEST2	31.6773	4.3073	232.9631	5.9056	169.9153
AGEST3	52.6518	7.2185	384.0405	9.8841	280.4709
AGEST4	59.6682	8.1134	438.8159	11.1239	320.0573
AGEST5	48.2261	6.0714	383.0663	8.4263	276.0122

OBS	AGE	EXPOSURE	CASES	TOTAL	AGEST1	AGEST2	AGEST3	AGEST4	AGEST5	AGEALC	B	STDB
1	1	1	1	10	0	0	0	0	0	1	-4.74493	1.00433
2	1	0	0	106	0	0	0	0	0	0	-4.74493	1.00433
3	2	1	4	30	1	0	0	0	0	2	-3.04980	0.34114
4	2	0	5	169	1	0	0	0	0	0	-3.04980	0.34114
5	3	1	25	54	0	1	0	0	0	3	-1.28935	0.16651
6	3	0	21	159	0	1	0	0	0	0	-1.28935	0.16651
7	4	1	42	69	0	0	1	0	0	4	-0.78125	0.13850
8	4	0	34	173	0	0	1	0	0	0	-0.78125	0.13850
9	5	1	19	37	0	0	0	1	0	5	-0.65611	0.16618
10	5	0	36	124	0	0	0	1	0	0	-0.65611	0.16618
11	6	1	5	5	0	0	0	0	1	6	-0.86904	0.33043
12	6	0	8	39	0	0	0	0	1	0	-0.86904	0.33043

/* MODEL 2 OUTPUT : 모델에 연령군외에 EXPOSURE 변수를 포함시킨 결과 */

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	991.488	802.922	.
SC	996.371	837.099	.
-2 LOG L	989.488	788.922	200.567 with 6 DF (p=0.0001)
Score	.	.	183.628 with 6 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	-5.0543	1.0094	25.0718	0.0001	.
AGEST1	1.5423	1.0659	2.0937	0.1479	0.342889
AGEST2	3.1988	1.0231	9.7745	0.0018	0.729085
AGEST3	3.7135	1.0185	13.2928	0.0003	0.884853
AGEST4	3.9669	1.0231	15.0344	0.0001	0.812463
AGEST5	3.9622	1.0650	13.8405	0.0002	0.453696
EXPOSURE	1.6699	0.1896	77.5695	0.0001	0.375352

Antilogs of ML estimates with confidence intervals

Variable	exp(beta)	95 percent confidence interval		90 percent confidence interval	
AGEST1	4.6753	0.5788	37.7681	0.8054	27.1407
AGEST2	24.5031	3.2987	182.0126	4.5298	132.5440
AGEST3	40.9970	5.5691	301.7986	7.6368	220.0875
AGEST4	52.8205	7.1109	392.3585	9.7648	285.7208
AGEST5	52.5729	6.5195	423.9434	9.0698	304.7373
EXPOSURE	5.3116	3.6630	7.7023	3.8848	7.2626

regression coefficients

age strata	beta	SD of beta
1	-5.05435	1.00942
2	-3.51205	0.35649
3	-1.85559	0.19383
4	-1.34086	0.16452
5	-1.08747	0.18417
6	-1.09216	0.34422

/* MODEL 3 OUTPUT : */
/* 연령군과 음주의 INTERACTION (agealc=age*exposure)을 포함시킨 결과 */

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	991.488	804.490	.
SC	996.371	843.549	.
-2 LOG L	989.488	788.490	200.998 with 7 DF (p=0.0001)
Score	.	.	190.854 with 7 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	-5.1823	1.0357	25.0388	0.0001	.
AGEST1	1.5664	1.0701	2.1426	0.1433	0.348244
AGEST2	3.2818	1.0362	10.0318	0.0015	0.748016
AGEST3	3.8483	1.0457	13.5446	0.0002	0.916979
AGEST4	4.1332	1.0614	15.1631	0.0001	0.846518
AGEST5	4.1271	1.1010	14.0523	0.0002	0.472578
EXPOSURE	2.1499	0.7534	8.1434	0.0043	0.483242
AGEALC	-0.1246	0.1891	0.4344	0.5099	-0.105317

Antilogs of ML estimates with confidence intervals

Variable	exp(beta)	95 percent confidence interval		90 percent confidence interval	
AGEST1	4.7894	0.5880	39.0091	0.8193	27.9961
AGEST2	26.6237	3.4933	202.9079	4.8166	147.1614
AGEST3	46.9132	6.0420	364.2615	8.3553	263.4082
AGEST4	62.3772	7.7901	499.4682	10.8253	359.4266
AGEST5	61.9979	7.1645	536.4965	10.0790	381.3624
EXPOSURE	8.5840	1.9606	37.5835	2.4764	29.7554
AGEALC	0.8828	0.6094	1.2789	0.6462	1.2061

regression coefficients

age strata	beta	SD of beta
1	-5.18232	1.03566
2	-3.61595	0.39725
3	-1.90051	0.20773
4	-1.33401	0.16447
5	-1.04916	0.19144
6	-1.05524	0.34506

====== INTERPRETATION ======

1) MODEL SELECTION

model statement에서 option으로 backward, forward, none, stepwise를 선택하면 regression analysis에서와 같은 방법으로 유의한 변수를 선택하여 최종모델을 만들게 된다. default는 none이다. 그러나 모델을 구축하기 위해서는 모든 변수에 대하여 층화분석을 실시하여 충분히 검토하고, 기존의 알려진 counfounding variable인 경우에는 그 변수가 들어가서 coefficient가 조금이라도 변화되면 통계적 유의성에 관계없이 모델에 포함시켜야 한다.

어떤 모델이 가장 잘 fit하는 가를 보는 방법으로 우선 Criteria for assessing model fit의 결과인 AIC (Akaike Information Criterion), SC (Schwarz Criterion), -2 Log Likelihood 등 통계량이 3가지가 있으며, 같은 자료로 여러개의 모델을 구축했을 때 이들의 통계값이 작을수록 desirable model이 된다.

2) Hypothesis test : likelihood ratio test, Wald test

model fitting 후에 '과연 model에 들어있는 변수들이 유의한가?'를 판정하기 위한 검정으로 위의 예에서 exposure(alcohol)변수가 들어있는 모델이 식도암을 설명하는 능력이 exposure변수가 들어 있지 않은 모델의 식도암을 설명하는 능력에 비해 많은 기여를 하고 있는가를 알아보면 된다. 이때에 쓰이는 방법으로 가장 powerful한 것은 Likelihood ratio test (LR test) 로 이는 change in scaled deviance의 값으로 검종을 한다. 이는 특정분포를 따르지 않으며 Goodness of fit에 매우 중요하다.

표 6.2에 쓰인 GLIM 의 Goodness-of-fit statistics와는 통계량이 다르나, 어느 모델

이 더 잘 적합을 하는가를 알기 위한 두모델사이의 log-likelihood의 차이인 LR statistic (Scaled deviance) 는 같다.

GLIM 통계량과 SAS OUTPUT의 비교

Goodness-of-fit Statistics (Log likelihood)						
	from GLIM			from SAS		
	df	G		df	-2 log L	x2
model 1	6	90.56	79.52(df=1)	5	868.44	121.05 / 79.52(df=1)
model 2	5	11.04	0.43(df=1)	6	788.922	200.57 / 0.43(df=1)
model 3	4	10.61		7	788.49	200.998

model 1 : age group covariate
model 2 : age group + alcohol drinking
model 3 : age group + alcohol + age*alcohol

여기서는 연령군을 보정한 음주량에 따른 위험도를 구하는 모델(모델 2)이 가장 잘 적합함을 알 수 있다.

Wald test는 standardized regression coefficient로 평가하는 방법으로 OUTPUT 5 중에서 모델 2에 있는 exposure(alcohol drinking)변수의 wald chi-square는 77.57 ($p < 0.001$)로 모델 1과 모델 2의 scaled deviance와 비슷함을 알 수 있다.

3) Point estimate : maximum likelihood estimate

여기서 구하여진 paremeter estimate는 maximum likelihood method로 구한 것으로 연령을 보정한 음주의 비교위험도는 exp (1.67) = 5.31이다.

```
/* PROC LOGISTIC의 OUPUT에서는 PARAMETER ESTIMATE만 나오며, Antilogs of ML   */
/* estimates with confidence intervals을 구하기 위해서는 FLOG.EXE라는 FILE을   */
/* 이용하여  C:\SAS> FLOG  INFILE OUTFILE [SWITCH] 의 명령어를 실행하면        */
/* 추정값들의 ANTILOG를 구할 수 있다.                                          */
/*     INFILE  : THE NAME OF THE SAS PRINT FILE TO MODIFY                     */
/*     OUTFILE : THE NAME OF THE FILE TO HOLD THE MODIFIED PRINTOUT           */
/*     SWITCH  : (OPTIONAL) SWITCH THE SIGNS, ETC.                            */
```

4) Interval estimation

95% 신뢰구간은 exp(1.67-1.96*0.19), exp (1.67+1.96*0.19)으로 3.66,7.70 이다.