자료분석법

- 다 변 량 분 석 법 -

서울의과대학 예방의학교실 유 근 영

1. 다변량분석법

1-1. 예방의학 분야에서의 다변량 분석법

예방의학분야에서 다변량분석법이 널리 활용되기는 1980년대 이후라고 생각되며, 역학-환경보건-보건관리의 비교적 다양한 학문영역에서 이용되고 있는 다변량 분석기법은 각 분야마다 약간씩의 차이는 있겠지만 의학 전 분야에서 이용되는 분석기법을 거의 모두 망라하고 있다고 해도 과언은 아닐 것이다. 일반적으로 다변량 분석법은 독립변수가 두가지 이상 혹은 종속변수가 두가지 이상인 경우의 총칭으로 사용하고 있어 예방의학 영역에서 흔히 적용되는 multiple regression, multiple ANOVA, multiple logistic regression 등 의 경우를 지칭하는 용어로 통용되고 있으나, 통계학적으로 보다 엄밀한 의미에서는 MANOVA, polytomous logistic regression 등의 경우와 같이 종속변수가 두가지 이상인 경우를 의미한다.

< 다변량 분석법의 정의 >

일반적 의미 : multiple regression, multiple ANOVA,
multiple logistic regression 등

엄밀한 의미 : MANOVA, polytomous logistic regression 등
종속변수가 두가지 이상인 경우를 의미

다변량 분석법은 일반적인 통계학적 분석법의 경우와 마찬가지로 가설상의 독립변수 및 종속변수의 척도에 따라 다음과 같이 구분되어 적용된다. 단, 아래에서 예로 들은 분석기법들은 대표적인 방법들 만을 열거한 것으로 각 구분마다 실제 적용할 수 있는 기법들은 이외에도 많이 있다.

① 독립변수나 종속변수 모두가 연속적인 변수일 때 → Multiple Regression Analysis
→ Multiple Correlation Analysis

② 독립변수는 비연속이고 종속변수는 연속적인 변수일 때 → Multiple ANOVA
독립변수는 연속 및 비연속 변수가 혼재되어 있고 종속변수는 연속적인 변수일 때
→ Multiple ANACOVA

③ 독립변수와 종속변수가 모두 비연속적인 변수일 때
: 종속변수가 이항분포를 따르는 경우 → Stratified Analysis
→ Multiple Logistic Regression Analysis
: 종속변수가 포아송분포를 따르는 경우 → Multiple Log-Linear Regression Analysis

④ 종속변수가 연속변수와 비연속변수 두가지로 구성되어 있을 때 → Survival Analysis

1-2. 다변량분석의 접근방법

1) 최적모델의 선정

다변량분석을 수행함에 있어서는 다음과 같은 대단히 중요한 사항 몇가지를 고려하지 않으면 안된다. 그 이유는 다변량분석은 분석하고자 하는 변수의 종류가 두가지 이상이 되기 때문인데, 그것이 독립변수이건 종속변수이건 간에 한가지 모델에 포함된 변수들은 예측하기 어려운 정도로 서로 상대방 변수에 영향을 주기 때문이다. 따라서 다변량분석을 시도할 때는 최소한 다음과 같은 사항을 신중히 고려하면서 연구자료에 적합한 모델을 구축하여야 한다.

< 다변량분석시 주의할 점 >

① 분석의 대상이 되는 연구자료는 이용하려 하는 분석법에서 요구하는 분석기법의 가정에 위배되어서는 안된다. 따라서 모든 자료는 분석 이전에 가정 위배여부를 검토해야 한다.
② 분석자료에 가장 적합도가 높은 모델을 선정하여야 한다. 선험적 방법과 통계학적 방법이 있다.
③ 교란변수의 영향은 철저히 통제되어야 한다.
④ 변수의 종류 (연속적 혹은 비연속적) 에 따라 원래 성질에 맞는 형태로 입력되고 해석되어야 한다.
⑤ 변수간의 상호작용을 적절한 방법으로 평가해야 한다.
⑥ 서로 상관성이 높은 변수간에는 collinearity 현상이 있으므로 유의하여야 한다.

통계학적 분석의 개념이 잘 정리되어 있지 않은 경우에는 연구자료에 적합한 모델을 선정하는데 소홀하게 되기 쉬우며 따라서 조사된 변수 모두를 그냥 모델에 넣어 컴퓨터에 맡겨버리는 경향이 흔한데, 앞서 주의사항에서도 지적하였듯이 그릇된 모델 (보다 적합한 모델이 있음에도 불구하고 이를 간과하여 버릴 경우) 을 적용하면 엉뚱한 결과를 얻게 될 수도 있음을 명심하여야 한다. 따라서 모델선정과정은 매우 신중히 생각하면서 수행해야 하는 부분이라 할 수 있다. 최적모델을 선정하는 과정은 다음과 같은 단계로 이루어지는게 보통이다. 자료를 설명할 수 있는 가능한 큰 모델을 생각하면서 이 모델에 포함되어야 할 변수들의 목록을 작성한다. 이 모델에 포함되는 변수는 관심있는 변수는 물론 포함되어야 하며 동시에 교란변수로서의 가능성이 있는 변수도 포함되어야 한다. 또한 기존 변수의 다차원형 (예: 연령^2, 신장^2), 대수변환 등의 변수변환 (예:log연령, 신장$^{-1}$), 교호작용 (예: 신장x연령)을 평가하기 위한 변수 등 모든 가능한 형태의 변수들로 이루어 지는 모델을 일차로 선정한다.

변수의 목록이 작성되었으면 실제로 '어떤 변수들의 조합으로 구성된 모델이 연구자료와 가장 잘 적합될 것일까 ?'를 해결해야 한다. 여기에는 ① 경험 및 주관에 의해 실증적으로 모델을 선정하는 방법과 ② 통계학적 이론의 힘을 빌려서 모델을 선정하는 방법이 있다. 자료를 설명할 수 있는 가능한 큰 모델을 일차로 구축하는 과정은 두 방법 모두 동일하지만, 예를 들어 중회귀분석의 경우 전자를 선호하는 연구자는 partial F-test (혹은 Type I SS) 를 이용하여 변수를 차례로 넣고 빼고 함으로써 기존 변수의 영향을 통제한 상태에서 해당 변수의 모델에의 기여도를 중심으로 적합도를 평가하게 되며, 후자를 선호하는 경우에는 가능한 큰 변수로 구성된 모델을 놓고 통계학적 확률론에 입각하여 (예: stepwise selection) 연구자료를 가장 잘 설명하는 모델을 찾아가는 방법을 말한다. 연구자의 견해나 취향에 따라 두가지 방법 중에서 모델선정법을 택하게 되는데 근래에는 전자를 선호하는 경향이 많다. 후자의 방법으로 모델을 선정하는 데는 ① forward selection, ② backward elimination, ③ stepwise, ④ maximum R^2 improvement, ⑤ all possible subsets selection 과 같은 방법이 있다.

2) 다변량 분석시 주의할 점

① 모델에의 적합도 검정 (= test for model assumptions)

【예 제】 회귀분석시 → 「normality assumption 검정」 등

【예 제】 COX'S PROPORTIONAL HAZARDS MODEL 이용한 생존자료 분석시
→ 「proportionality assumption 검정」

② 교란변수의 영향 통제 (= control for confounding effects)

③ 질적 변수의 범주화 (= designing indicator variables)

```
$UNITS 100
$DATA YRS AGE DIS
$READ 1 1 2  1 1 2  1 1 2  1 1 2  1 1 1
$FAC YRS 3 AGE 2
```

← INDICATOR VARIABLE

```
PROC LIFEREG; CLASS NAGE NN NT;
 MODEL MONTH*CENSOR(1) = NAGE NN NT / DIST=WEIBULL;
 RUN;
```

← 변수의 범주화

④ 독립변수간의 상호작용 평가 (= assessing interaction effects)

⑤ 상관성 변수의 다중공선성 통제 (= control for collinearity)

【예 제】
대장암 위험요인으로서의 CALORY INTAKE 과 FAT INTAKE 간의 collinearity 문제

【예 제】
심근경색증 위험요인 분석연구에서 BMI index 와 HEIGHT 혹은 WEIGHT 를 한 모델에 동시에 포함시키는 경우

【예 제】
연령과 혈압과의 관계식을 회귀분석법으로 추정하시오.

연령	수축기혈압	연령	수축기혈압	연령	수축기혈압
39	144	45	138	47	145
.	.	.	.	.	.

second-order polynomial model : [redacted]

⑥ 누락 항목의 처리 (= treatment for missing values)

: 편견이 개입된 결론 (biased conclusion) 을 도출 → 과소평가
: 분석단계에서의 해결방법
- *interpolation* : 기존 자료에서 예측되는 중앙값으로 내삽
- extrapolation : 다른 연구에서의 경험을 바탕으로 누락치를 예측
- elision : 해당 변수의 분석시에만 누락이 있는 변수를 분석대상에서 제외시키는 방법 (list-wise deletion technique)
- replacement : 결과를 귀무의 방향으로 대치시키는 방법 (replacement with values towards the null)

2. 연속변수의 다변량 분석법

2-1. 분산분석법의 응용

1) 분산분석법의 기본 가정

- 세 개 이상의 평균치를 동시에 비교할 경우
- Student's t-test를 세 번 반복적용하면 통계학적 오류 (특히 α-error) 증가

- 가 정

> ① 독립변수, 즉 세 가지 치료군은 반드시 서로 독립적이어야 함
> ② 오류없이 측정된 각 X에 대하여 가지는 Yi 값(치료효과)들은 반드시 정규분포를 따라야 함
> ③ 각 X에 대한 Yi 값들의 분포의 분산은 반드시 같아야 함

(참 고) Unbalanced data 의 분산분석 접근

PROC ANOVA : 처치군별로 관찰한 반복수가 같은 경우 (balanced data) 에만
PROC GLM : missing values 로 처치군별 반복관찰수가 같지 않을 때 사용 가능

LOG
```
WARNING: PROC ANOVA has determined that the number of observations in each
         cell is not equal.  PROC GLM may be more appropriate.
```

2) 독립변수가 연속변수인 경우의 접근법 (공분산분석: ANACOVA)

【 예 제 】
임신 중의 mycoplasma infection 감염이 있었던 산모와 없었던 산모에서 태어난 아기의 출생시체중을 비교하려 한다. 출생시체중에 영향을 미치는 다른 여러 요인들 (산모의 신장, 재태기간 등) 의 복합적인 영향을 보정한 상태에서 감염여부 단독에 의한 출생시 체중을 비교하려면 어떤 모델을 이용하여야 하겠는가 ?

Y(출생시체중) $= \beta_0 + \beta_1 X_1$(감염여부) $+ \beta_2 X_2$(신장) $+ \beta_3 X_3$(재태기간)

【 예 제 】

Colchicine (0.25%, 20ul/d) 를 국소점안할 경우 안압의 변화를 관찰하기 위하여 투여군과 비투여군에서 안압을 처치전과 처치 1주 후에 측정한 결과는 다음과 같았다. 두 군간에 유의한 차이가 인정되는가 ?

3) 종속변수가 둘 이상인 경우의 접근법 (다변량 분산분석: MANOVA)

【 예 제 】

수유중인 산모에게 DMPA (medroxyprogesterone acetate) 와 norgestrel 의 두가지 hormonal contraceptives 를 투여한 후 시간경과에 따른 milk volume 의 변화를 대조집단과 비교한 자료는 다음과 같다. 종속변수인 milk volume 에 영향을 미치는 산모의 연령과 출산력을 보정한 상태에서도 약제의 종류에 따른 milk volume 에 차이가 인정되는가 ?

치료군	연령	출산력	B1	B2
CONT	24	3	115	135
DMPA	32	3	180	150
NORG	29	4	215	220

치료군	연령	출산력	B1	B2
CONT	21	4	190	210
DMPA	24	2	225	265
NORG	23	2	215	235

치료군	연령	출산력	B1	B2
CONT	24	2	80	105
DMPA	21	2	155	160
NORG	24	2	145	145

【 프로그램 】

```
PROC GLM DATA=MILK;
  CLASS GRP NAGE NPAR;
  MODE[redacted]
[redacted]
QUIT;
```

2-2. 회귀분석법의 응용 (중회귀분석)

중회귀분석은 선형회귀분석을 독립변수가 둘 이상의 경우로 확장시킨 것이다. 그러나 독립변수가 둘 이상으로 구성되는 중회귀모델은 단변수 모델의 경우에 비해 다음과 같은 이유로 인해 상당히 복잡하다. 즉, ① 변수가 여러가지로 가능성있는 경우의 수가 다양해 측정자료를 가장 잘 반영해주는 모델을 설정하는데 있어 어려움이 따른다. ② 측정자료에의 적합도를 가시적으로 구성하는데 있어 삼차원 이상의 도표를 작성해야 하므로 이해가 어렵다. ③ 가장 잘 반영하는 모델을 찾는다 해도 실제 상황에서 그 결과의 의미를 해석하는데 어려움이 있다. ④ 계산을 하는데 특별한 기능을 가진 고속컴퓨터나 통계용 프로그램의 힘을 빌어야만 한다.

1) 중회귀모델의 기본가정

중회귀모델을 이용한 회귀분석의 가정은 원칙적으로는 단변수 선형회귀모델의 경우와 동일하다.

2) 중회귀 분석법

자료분석을 위한 중회귀모델이 설정되면 여기서 구한 추정치를 통해 독립변수(들)이 종속변수 Y 를 예측하는데 얼마나 도움이 되는지 알아야 한다. 중회귀모델을 이용하여 가설검정을 하면 이와같은 문제를 해결할 수 있는데 보통 다음과 같은 세가지 검정방법이 있다.

① Test for significant overall regression : 모델에 포함된 독립변수 모두가 Y 를 유의하게 예측할 수 있는가?

② Partial F-test : 독립변수 하나를 기존의 독립변수들에 추가할 때 그 변수의 첨가로 인해 Y 의 예측이 유의하게 좋아지는가?

③ Multiple-partial F-test : 독립변수 몇개를 기존의 독립변수들에 추가할 때 그 변수의 첨가로 인해 Y 의 예측이 유의하게 좋아지는가?

3. 비연속변수의 분석법

3-1. 비연속자료 분석법

의학분야에는 WBC differential count나 치료의 반응정도 (grade I/II/III) 처럼 평균을 구하기 애매한 경우가 많이 있다. 연속적(continuous)이지 못한 변수는 평균치와 표준편차 형태로 요약될 수 없기 때문에 (보다 구체적으로는 정규분포를 가정할 수 없기 때문에), 앞서 실습한 Student's t-test나, 분산-회귀분석법 등을 적용할 수 없다. 연속변수의 성질을 가지지 못하는 변수를 비연속 혹은 범주형 (categorical, qualitative) 변수라 한다. 성별, 흡연여부, 환자군/대조군, 수술후 생존/사망여부 등은 둘로 갈라지는 양분성 (dichotomous) 변수의 좋은 예이며, ABO 혈액형, 농도를 달리한 약물 투여군, -/+/++/+++ 식의 병리학적 분류, 항체 titer 등 세가지 이상으로 갈라지는 양적인 변수를 polychotomous 변수라 한다. 물론 원천적으로는 연속적 변수라 하더라도 분석을 위해 분류하는 과정에서 얼마든지 비연속변수로 변환될 수 있다. 비 연속변수는 자료의 형태에 따라서 다음과 같이 구분되어 분석된다.

① Two-by-two table : Fisher's exact test, Pearson's chi-square test, chi-square test with Yate's correction, unadjusted likelihood ratio test relative risk or odds ratio, confidence limits phi coefficient,

② Two-by-k table : global chi-square test, score test for trend, unadjusted likelihood ratio test for trend

③ R-by-C table : Pearson's chi-square test, score test for trend, ridit analysis Cramer's V, gamma ststistic, Kendall's Tau-b, Stuart's Tau-c, Spearman's rank correlation

④ Stratified table	:	BD test for homogeneity, adjusted global test, adjusted score test for trend, adjusted likelihood ratio test, adjusted likelihood ratio test for trend adjusted relative risk (MH or logit estimator), confidence limits
⑤ Multivariate analysis	:	linear or polynomial logistic regression, log-linear regression, polychotomous logistic regression, conditional logistic regression

지금부터는 단일 예제를 이용하여 "통계학적 분석법의 종류에 따라 연구결과 및 해석이 어떻게 달라지는가?" 를 비연속자료의 분석을 통해 보여주기로 하겠다.

3-1. 이분성 자료의 분석법

【예 제 2-1】 모유의 수유가 유암의 발생을 보호하는 효과가 있음은 아직도 학문적 논쟁거리이다. 다음은 모유 수유 여부와 유암 발현과의 관련성 증명하기 위하여 시도된 환자-대조군 연구의 결과이다. 모유 수유 여부와 유암과의 관련성에 관하여 통계적으로 분석하시오.

	모유수유경험(+)	모유수유경험(-)
환자군	388	133
대조군	426	95

【 예 제 풀 이 】

< 관측값 >

위험요인에의 폭로여부

질 병	E(+)	E(-)	
D(+)	388	133	521
D(-)	426	95	521
	814	228	1,042

< 기대값 >

	E(+)	E(-)	
D(+)	407	114	521
D(-)	407	114	521
	814	228	1,042

$$\therefore S^2 = \frac{(388\times95-133\times426)^2\times1042}{521\times521\times814\times228} = 8.11 > \chi^2(1) = 3.84 \rightarrow \text{가설 기각}$$

3-2. 층화분석법

1) 교란변수의 확인 및 분석

a. Follow-up Study 의 경우 (no interaction)

< 단계 1 > crude 와 adjusted measures 간에 차이가 있음을 확인
< 단계 2 > $RR_{ef} \neq 1$ 임이 확인되고 동시에
< 단계 3 > $RR_{df|E(-)} \neq 1$ 임이 확인될 때

【예 제 : 임상시험】
Renal calculi 에 대한 percutaneous nephrolithotomy 의 치료효과에 관한 임상시험 결과는 다음과 같았다. 어떤 교란변수의 영향이 예상되는가 ? 교란변수의 영향을 제거한 후 치료효과는 어떻게 변했는가 ?

치 료 종 류	환자수	성공	실패	성공률(%)
합계 (n=700)				
open surgery	350	273	77	78
percutaneous	350	289	61	83
Stone < 2cm (n=357)				
open surgery	87	81	6	93
percutaneous	270	234	36	87
Stone >= 2cm (n=343)				
open surgery	263	192	71	73
percutaneous	80	55	25	69

< 단계 1 > crude rate = 1.06
adjusted rate = 0.94
- stratum-specific rate (1) = 0.94
- stratum-specific rate (2) = 0.95

< 단계 2 > RRef = (87/263)/(270/80) = 0.1 ≠ 1
< 단계 3 > RRdf|E(-) = (81/6)/(192/71) = 4.99 ≠ 1

b. Case-Control Study 의 경우 (no interaction)

< 단계 1 > crude 와 adjusted measures 간에 차이가 있음을 확인
< 단계 2 > $OR_{ef|D(-)} \neq 1$ 임이 확인되고 동시에
< 단계 3 > $OR_{df|E(-)} \neq 1$ 임이 확인될 때

【예 제 : 환자-대조군 연구】

유방암의 발생위험은 초경연령이 빠를수록 증가하는 것으로 알려져 있다. 이같은 사실을 확인하기 위하여 계획되었던 연구의 결과는 다음과 같다. 예상되는 교란변수의 존재여부를 연구자료에서 확인해 보시오.

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)

가설 I [초경연령 → 유암발생] ⋯⋯→ [에스트로겐에의 총 폭로기간]

< 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

가설 II [연령 → 초경연령], [연령 → 유암발생] ⋯⋯→ 교란변수 : [진단당시의 연령]

2) 층화분석법 (Stratified Analysis)

관찰표본을 교란변수에 대해 층화하여 각 strata 내에서는 교란변수가 homogenous 하게 분포하도록 한 후 각 층에서 관찰한 값을 수학적으로 합해 줌으로써 교란변수의 영향을 보정하는 방법

* 분석전략 : strata 마다의 RR 을 비교하여

RR 들이 constant 하다면 →
- summary RR 계산 (estimation)
- test of significance
- 95% C.I.

└ RR 들이 constant 하지 않으면
→ confounder 와 RR 간의 interaction
→ detailed analysis and multivariate modelling 으로 해결!

【예 제 3-3】 다음은 모유 수유와 유암과의 관련성을 증명하려는 환자-대조군 연구의 결과로 이들간의 관련성을 역학적으로 증명하는데 만삭분만시 연령 (3 범주) 의 교란영향을 의심하여 이를 제거하려고 작성한 층화 표이다. 이를 이용하여 각 층간의 상호관련도 지표를 각각 산출하고 이들 지표간의 동질성을 검정하시오.

	만삭분만시 연령 (i)								
	24세 이하			25-29세			30세 이상		
	모유 수유			모유 수유			모유 수유		
	1+	0		1+	0		1+	0	
환자군	171	13	184	172	21	193	41	8	49
대조군	212	10	222	179	10	189	30	9	39
	383	23	406	351	31	382	71	17	88

【예 제 풀 이】

$$x^2_{B-D} = \sum \frac{[a_i - A_i(\psi)]^2}{Var(a_i:\psi)}$$

$$= \frac{(171-171.06)^2}{5.34} + \frac{(172-173.99)^2}{6.784} + \frac{(41-37.96)^2}{3.175}$$

$$= 3.495$$

☞ 통계학적으로는 "이들 각 층은 서로 동질적인 OR 을 가진다" 고 결론

[공식 3-6] Cochran-Mantel-Haenszel chi-square test statistic

$$x^2_{CMH} = \frac{[|\sum_i a_i - \sum_i A_i(1)| - \frac{1}{2}]^2}{\sum_i Var(a_i:\psi=1)} \approx x^2(1)$$

$$\therefore \; x^2_{CMH} = \frac{\{|384-390.45| - \frac{1}{2}\}^2}{15.95} = 2.22 \; < \; x^2(1)=3.84$$

☞ '만삭분만시 연령 (3 범주) 로 교란변수를 간주하여 그 영향을 통제한 결과 모유 수유는 유암과 유의한 상관관계에 있지 못함

3-3. 선형로짓회귀분석법

[선형로짓회귀분석법의 장점]

① 층화해야 할 교란변수의 수가 많아지면 층에 따라 세분화하는 과정에서 특정 층의 자료가 소실되는 문제 (breakdown) 가 발생된다. 즉, 층 내의 관찰 수가 적어지게 되어 어떤 칸은 환자군만 있게 되며 어떤 칸은 대조군만 있게 되는 경우가 초래되어 결국은 관찰된 모든 정보를 분석에 이용하지 못하게 된다.

② 독립변수가 연속변수일 경우에도 유사한 문제가 발생되는데, 층화를 위해서 연속성인 독립변수를 범주형으로 자르는 과정에서 원래의 연속성이 상실되는 문제 (intrapolation 혹은 extrapolation) 가 발생되며, 그래서 너무 잘게 자르면 마찬가지로 특정 칸의 소실이 발생할 수 있다.

③ '범주형 자료의 경우 각 범주의 수준을 어떻게 정하느냐?'에 따라 결과가 일률적이지 못한 문제가 있으며, 동시에 기준이 되는 수준 (reference level) 을 무엇으로 하는가에 따라 결과가 달라지는 문제가 있다.

④ 한가지 이상 몇 가지 위험요인의 복합작용 (joint effect) 에 의한 결과를 산정하는 과정이 일반적으로 복잡하다.

⑤ 독립변수들 간의 교호작용 (interaction effect) 을 파악 할 수 없다는 심각한 문제를 내포하고 있다.

【예 제 4-2】 다음은 유암과 모유 수유와의 관련성에 관한 환자-대조군 연구자료를 이용하여 선형로짓분석법을 수행한 결과이다. 연령 (NAGE) 및 첫 만삭분만시 연령 (NFTP)의 영향을 보정한 상태에서 모유 수유 (NNBF) 와 유암과는 관련성이 있다고 할 수 있겠는가?

【예 제 풀 이】

< 표 4-7 > 우도비 검정법을 이용한 가설검정법의 요약 (EGRET 시스템)

Model	Log-likelihood	△G (△df)	*P*	Interpretation
NAGE + NNBF + NFTP	G_1 = 1205.67 (864)	G_2-G_1 = 2.54(1)	0.11	NNBF ¦ NAGE, NFTP
NAGE + NFTP	G_2 = 1208.21 (865)	G_3-G_1 = 3.31(2)	0.19	NFTP ¦ NAGE, NNBF
NAGE + NNBF	G_3 = 1208.98 (866)	G_4-G_3 = 3.27(1)	0.07	NNBF ¦ NAGE
NAGE	G_4 = 1212.26 (867)	G_4-G_2 = 4.04(2)	0.13	NFTP ¦ NAGE
Null model	G_5 = 1213.74 (875)	G_5-G_4 = 1.48(8)	0.99	NAGE

☞ 선형로짓모델을 이용한 가설검정 결과, 진단시 연령과 첫 만삭분만시 연령의 영향을 보정한 상태에서 모유 수유는 유암 위험을 유의하게 변동시키지 못함

3-4. 순위변수의 경향분석법

◉ 모유 수유와 유암 연구의 통계학적 비판

◉ Score test statistics for linear trend

【예 제 5-1】 모유 수유에 의한 유암 보호효과를 역학적으로 증명하기 위하여 만삭분만의 경험이 있는 여성 중 426명의 유암 환자와 450명의 대조군을 대상으로 환자-대조군 연구를 수행하였다. 설문을 통해 자녀당 평균 수유기간 (DUBF, 0 - 12개월) 에 관한 자료를 수집하였으며, 이를 4가지로 분류되는 범주형 변수 (NDBF, 4 범주) 로 처리하여 '자녀당 평균 수유기간이 길면 길수록 유암을 보호하는 효과가 클 것인가?' 의 가설을 증명하시오.

자녀당 평균 수유기간 (DUBF)	자녀당 평균 수유기간 (NDBF)	환자군	대조군	계
0 개월	0	42	29	71
1-3	1	83	83	166
4-6	2	67	61	128
7-9	3	34	48	82
10-12	4	200	229	429
합 계		426	450	876

【예 제 풀 이】

$$\text{unconditional } uS^2_{Arm} = \frac{N^3[\sum X_k(a_k - e_k)]^2}{n_1 n_0[N\sum X_k{}^2 m_k - (\sum X_k m_k)^2]}$$

$$= \frac{876^3 \times (1119 - 1159.331)^2}{426 \times 450 \times (876 \times 8268 - 2384^2)} = 3.658$$

$$\text{conditional } cS^2_{Arm} = \frac{N^2(N-1)[\sum X_k(a_k - e_k)]^2}{n_1 n_0[N\sum X_k{}^2 m_k - (\sum X_k m_k)^2]}$$

$$= \frac{876^2 \times 875 \times (1119 - 1159.331)^2}{426 \times 450 \times (876 \times 8268 - 2384^2)} = 3.654$$

☞ '모유 수유의 평균기간과 유암과는 5% 유의수준에서는 직선적 관계에 있지 않다' 혹은 '10% 유의수준에서는 직선적 관계에 있다'

◉ 선형로짓경향분석법 (Likelihood ratio test for trend)

【예 제 5-3】 모유 수유에 의한 유암 보호효과를 역학적으로 증명하기 위하여 만삭분만의 경험이 있는 여성을 대상으로 수행된 환자-대조군 연구의 결과는 아래와 같다. '자녀당 평균 수유기간이 길면 길수록 유암을 보호하는 효과가 클 것인가?' 의 가설을 첫 만삭분만시 연령의 영향을 보정하면서 로짓분석을 이용해 증명해 보시오.

첫 만삭분만시 연 령 (NFTPX)	자녀당 평균 수유기간 (NDBF)	환자군	대조군	합 계
29세 이하	0	34	20	54
	1	72	75	147
	2	59	58	117
	3	31	40	71
	4	181	218	399
소 계		377	411	788
30세 이상	0	8	9	17
	1	11	8	19
	2	8	3	11
	3	3	8	11
	4	19	11	30
소 계		49	39	88

자료명 [PAR.DAT] : Yoo et al. (1992)

M o d e l	No. of parameters	df	Log-likelihood (G)
(1) NAGE1--NAGE9 + NFTPX2	10	9	G_1 = 3.254
(2) NAGE1--NAG29 + NFTPX2 + NDBF	11	10	G_2 = 7.333
(3) NAGE1--NAGE9 + NFTPX2 + NDBF1--NDBF4	14	13	G_3 = 9.519
(4) NAGE1--NAGE9 + NFTPX2 + NDBF + NNDBF	12	11	G_4 = 7.663

$$\chi^2_{trend} = [G_1 - G_2] = 4.079\ (1)$$

☞ χ^2_{trend} = 4.079 (1), p=0.04

☞ '만삭분만시 연령의 교란영향을 로짓모델로 보정한 상태에서 볼 때, 모유수유의 기간이 길면 길수록 유암 위험은 직선적으로 낮아지는 경향이 관찰되었으며 이러한 현상은 통계적으로도 유의하였다'

3-5. 선형로짓모델 분석법

1) 선형로짓모델을 이용한 교란변수 보정

① confounder 를 층화변수로 하여 각 층마다 logistic model 을 적용

② 보다 쉬운 방법은 model 내에 교란효과를 수학적으로 흡수

(장점) ┌ 여러 형태 ($X, X^2, \log(X)$ 나 X …) 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 집어 넣음으로해서 risk variable 의 coefficients 가 변화된다 하더라도 통계적 유의성에 관계없이 model 에 집어 넣을 것

2) 선형로짓모델을 이용한 adjusted MLE 추정

```
PROC LOGISTIC DATA=CDA.PAR ORDER=DATA;
 MODEL GRP = NAG2 NAG3 NAG4 NAG5 NAG6 NAG7 NAG8 NAG9 NFTP2 NFTP3
             NDBF1 NDBF2 NDBF3 NDBF4;
 RUN;
```

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	0.1239	0.4554	0.0740	0.7856	.
NAG2	0.0749	0.4351	0.0296	0.8633	0.011350
NAG3	-0.0652	0.4062	0.0258	0.8725	-0.012661
NAG4	0..0438	0.3978	0.0121	0.9124	0.009438
NAG5	0..1077	0.3990	0.0728	0.7873	0.022892
NAG6	0.0806	0.4153	0.0377	0.8461	0.014500
NAG7	0.3040	0.4238	0.5144	0.4732	0.051898
NAG8	0.1739	0.4538	0.1469	0.7015	0.024045
NAG9	0.0823	0.4487	0.0336	0.8545	0.011727
NFTP2	0.1932	0.1453	1.7661	0.1839	0.052838
NFTP3	0.3550	0.2435	2.1246	0.1450	0.058861
NDBF1	-0.2998	0.2926	1.0502	0.3055	-0.064819
NDBF2	-0.2172	0.3032	0.5132	0.4738	-0.042327
NDBF3	-0.6670	0.3339	3.9903	0.0458	-0.107179
NDBF4	-0.4743	0.2701	3.0830	0.0791	-0.130790

【예 제 풀 이】

adjusted OR of NDBF1 | NAGE, NFTP = exp(−0.2998) = 0.7410

adjusted OR of NDBF2 | NAGE, NFTP = exp(−0.2172) = 0.8048

adjusted OR of NDBF3 | NAGE, NFTP = exp(−0.6670) = 0.5132

adjusted OR of NDBF4 | NAGE, NFTP = exp(−0.4743) = 0.6223

$$95\%\ CI\ (OR_{NDBF1|NAGE,\ NFTP}) = \exp[\beta_{NDBF1} \pm Z_{a/2} \times s.e.(\beta_{NDBF1})]$$
$$= \exp[-0.2998 \pm 1.96 \times 0.2926]$$
$$= (0.4176 - 1.3148)$$

$$95\%\ CI\ (OR_{NDBF2|NAGE,\ NFTP}) = \exp[\beta_{NDBF2} \pm Z_{a/2} \times s.e.(\beta_{NDBF2})]$$
$$= \exp[-0.2172 \pm 1.96 \times 0.3032]$$
$$= (0.4442 - 1.4580)$$

3-6. 비선형로짓모델 분석법

【예 제 7-3】 모유 수유가 유방암을 보호하는 효과가 있음을 환자-대조군 연구로 평가하기 위하여 연구자료를 분석한 결과, 각 층별 OR 값들이 직선적인 관계에 있지 않는 것으로 의심되었다. 이 자료를 비선형 로짓모델에 적합시켜 분석하시오.

NDBF	aOR	(95% CI)
자녀당 평균 모유수유 기간		
0	1.0	
1-3	0.74	(0.42-1.31)
4-6	0.80	(0.44-1.46)
7-9	0.51	(0.27-0.99)
10-12	0.62	(0.37-1.06)

```
OPTIONS PS=60 LS=80;
LIBNAME CDA 'C:\CDA\';

PROC LOGISTIC DATA=CDA.PAR ORDER=DATA;
 MODEL GRP = NAG2 NAG3 NAG4 NAG5 NAG6 NAG7 NAG8 NAG9 FFTP [redacted]; RUN;
```

$$\text{logit } P_i = -1.1442 + 0.1028\,NAG2 - 0.0356\,NAG3 + 0.0945\,NAG4 + 0.1698\,NAG5 + 0.1389\,NAG6 + 0.4042\,NAG7 + 0.2554\,NAG8 + 0.1925\,NAG9 + 0.0531\,FFTP \quad (SE_{FFTP} = 0.0212) - 0.2072 \log(DUBF+1) \quad (SE_{\log(DUBF+1)} = 0.0912)$$

$$OR = (DUBF+1)^{-0.2072} = 2^{-0.2072} = 0.8662$$

$$95\%\ CI = 2^{(-0.2072 \pm 1.96 \times 0.0912)} = [0.765 \approx 0.981]$$

< 표 7-1 > Adjusted risk of breast cancer related with the duration of breast feeding per child among parous women in a case-control study

Average duration of breast feeding per child (month)	No. of cases	controls	Adjusted OR (95% CI)[1]
0	42	29	1.0
1 - 3	83	83	0.74 (0.42-1.31)
4 - 6	67	61	0.80 (0.44-1.46)
7 - 9	34	48	0.51 (0.27-0.99)
10 - 12	200	229	0.62 (0.37-1.06)
			χ^2_{trend}[2] = 5.211(1), p=0.022

[1] Adjusted odds ratio and 95% confidence intervals were derived from regression coefficients and standard error in linear logistic models. Adjustment for age intervals and the categorized age at first full term pregnancy was done.

[2] Chi-square value of likelihood ratio test for trend [redacted] in the logit risk with continuous exposure to the risk factors.

<< 참 고 문 헌 >>

1. 자료분석론 일반

- 안윤옥, 유근영, 박병주. 실용의학통계론. 서울대학교출판부, 서울, 1996.
- 유근영, 박병주, 김 헌, 이무송. 의·약·보건학을 위한 PC-SAS. 한울아카데미, 서울, 1995.
- Daniel WW. Applied nonparametric statistics. Houghton Mifflin Company. Boston, 1978.
- Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. Biostatistics in Clinical Medicine. MacMillan Publishing Co., Inc., New York, 1987.

2. 연속변수의 다변량 분석법

- Kleinbaum DG, Kupper LL, Muller KE. Applied regression analysis and other multivariable methods. PWS-KENT Publishing Company. Boston, 1988.
- Yamane T. Statistics. An introductory analysis. Harper International Edition. New York, 1973.

3. 비연속변수의 다변량분석법

- 유근영. 의학·보건학을 위한 범주형 자료분석론. 선형로짓분석법을 중심으로. 서울대학교출판부 서울, 1996.
- Breslow NE, Day NE. Statistical methods in cancer research. Vol. 1-The analysis of case-control studies. IARC, Lyon, 1980.
- Aitkin M, Anderson D, Francis B, Hinde J. Statistical modelling system in GLIM. Oxford Science Publications. Oxford, 1990.
- Hosmer DW, Lemeshow S. Applied logistic regression. John Wiley & Sons. New York, 1989.
- Cody RP, Smith JK. Applied statistics and the SAS programming language. North-Holland. New York, 1991.
- Statistics and Epidemiology Research Corporation. EGRET. SERC Inc. Seattle, 1990.
- SAS Institute Inc. SAS/STAT guide for personal computers. SAS Institute Inc. Cary, 1987.