로지스틱 회귀분석

가톨릭의과대학 예방의학교실
이 원 철

1) 서론

종속변수도 연속형 자료이고 독립변수도 연속성 자료인 경우에는일반적인 회귀분석을 적용한다. 이에 비하여 종속변수가 범주형 자료이고 독립변수도 범주형 자료인 경우에 X^2 분석을 적용한다.

그런데 종속변수가 범주형 자료인데 독립변수가 연속형 자료인 경우도 얼마든지 가능하다. 예를 들어 연령(독립변수)이 관상동맥질환의 사망(종속변수)에 어떠한 영향을 주는지를 생각해 볼 수 있다.

표 1은 이러한 상황을 가상적으로 만든 자료의 일부이다.

표 1. 100명에 대한 관상동맥질환 발생과 연령 및 연령군

일련번호	연령군	연령	질병발생여부	일련번호	연령군	연령	질병발생여부	일련번호	연령군	연령	질병발생여부
1	1	20	0	35	3	38	0	68	6	51	0
2	1	23	0	36	3	39	0	69	6	52	0
3	1	24	0	37	3	39	1	70	6	52	1
4	1	25	0	38	4	40	0	71	6	53	1
5	1	25	1	39	4	40	1	72	6	53	1
6	1	26	0	40	4	41	0	73	6	54	1
7	1	26	0	41	4	41	0	74	7	55	0

표의 질병 발생 여부에서 0는 발생을 하지 않은 것이고 1은 발생한 경우를 의미한다. 이러한 자료를 회귀분석 개념 그대로 점도표로 그리면 이 점도표는 일반적인 점도표와 많이 다르다. 즉 1) 자료들이 두 개의 평행선상에 위치하고 2) 관상동맥질환 발생과 연령과의 상관성을 보여주지 못하며 3) 각 연령에 대한 관상동맥질환 발생의 변이가 매우 큰 문제들을 지니고 잇다.

이를 해결하기 위한 한 방법이 독립변수에서 구간을 설정하고 각 구간에서의 평균질병발생률을 구하여 이를 점도표로 표현하는 방법이다. 이런 방법으로 표 1의 자료를 병환시킨 것이 표 2 이고 2를 점도표화하면 일반적인 점도표에 근사해진다.

표 2. 관상동맥질환 발생과 연령군에 의한 도수분포표

연령군	대상자수	관상동맥질환 발생		평균(구성비)
		없음	있음	
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
합계	100	57	43	0.43

이러한 변형된 곡선을 충족시켜 주는 함수는 몇가지 종류가 있으나 이들 함수들 중에서 1) 매우 탄력성이 많고 쉽게 사용할 수 있다는 점과 2) 그 결과를 생물학적으로 해석하기에 용이하다는 점 때문에 로지스틱 함수를 사용하게 된다. 이러한 로지스틱 함수는 아래의 형태를 취한다.

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

이와 같이 변형된 도표는 일반적인 회귀 모형에 많이 접근하여 있다. 그러나 일반 회귀모형에서는 X의 범위가 -∞ 에서 +∞에 위치하면 y의 범위 역시 -∞에서 +∞까지 위치하나 위 도형에서는 y의 범위가 0에서 1까지 밖에 위치하지 못하고 있다.

따라서 로지스틱 함수를 아래 식과 같이 로짓변형(logistic fransformation)시키면 결국 일반적인 회귀모형과 같은 형태를 지니게 된다.

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$
$$= \beta_0 + \beta_1 x$$

첫 부분에 언급하였던 바와 같이 지금까지 인용한 자료는 종속변수가 범주형 자료이고 독립변수가 연속형 자료이었기 때문에 일반 회귀모형을 그대로 적용하기가 매우 어려웠으나 이 자료를 가지고 로짓변형시키면 일반적인 회귀모형의 형태를 지니게 되어 분석이 용이하여지게 되었다.

따라서 앞으로 진행하게 되는 설명은 모두 회귀분석 모형에 맞추어 설명과 해석을 진행하게 된다.

2) 로지스틱 회귀모형에서 산출된 회귀계수에 대한 해석

2-1) 단순 2 x 2 표에서의 회귀계수에 대한 해석

일반 회귀모형에서 회귀계수(β)의 의미는 X가 1단위 변화할 때 y는 어느 정도를 변화하는지의 크기를 나타내는 기울기이다. 그렇다면 로지스틱 회귀모형에서의 회귀계수는 어떠한 의미를 지니는 지를 살펴보고자 한다.

우선 위의 연령과 관상동맥질환 발생에 관한 예를 가지고 설명을 하여 보자.

설명을 편하게 하기 위하여 연속변수로 되어 있는 독립변수(연령)을 2군으로 나누어 범주변수로 만들기 위하여 연령이 55세 이상이면 1을 취하고 55세 미만인 경우에는 0을 취하도록 한다. 이에 따라 이분된 연령변수와 CHD의 발생을 가지고 작성한 2 x 2 표가 표 3이다.

표 3. 55세를 기준점으로 이분된 연령과 CHD 발생간의 2 x 2표

CHD	연 령 (X) 55 이상(1)	55 미만(0)	합
있음(1)	21	22	43
없음(0)	6	51	57
합	27	73	100

이 자료를 그대로 이용하여 로지스틱 회귀 프로그램으로 우도(likelihood)를 최대화하는 β_0 와 β_1의 값을 구하면 표4와 같은 회귀계수가 계산된다.

표 4. 표3의 자료를 이용하여 로지스틱 회귀모형으로 분석한 결과

변수	추정회귀계수	표준오차	계수/표준오차	φ
연령	2.094	0.529	3.96	8.1
상수	-0.841	0.255	-3.30	

로지스틱 모형으로 분석한 결과 회귀계수는 2.094로 측정되었다. 그러면 2.094의 의미는 무엇일까? 여기서는 그 과정을 생략하도록 하겠으나 2.094를 antilog 를 취한 값 $e^{2.094}$= 8.1 이 된다.

그런데 이 8.1은 무엇과 같은가 하면 위 표 3의 결과를 가지고 odds ratio 를 구한 값 $\frac{21 \times 51}{22 \times 6} = 8.1$ 과 같은 결과를 나타내었다.

즉 로지스틱 회귀모형에서 추정된 회귀계수에 antilog 를 취하여 나온 값이 곧 odds ratio 이다.

Odds ratio 는 두 변수간의 관련성의 정도를 나타내어 주는 지수로서 역학연구에서 사용되는 중요한 지수들 중의 하나이다. 그러므로 로지스틱 회귀 분석에서 추정된 회귀계수의 antilog 가 바로 odds ratio 의 추정치라는 것 때문에 역학연구에서 로지스틱 회귀분석을 흔히 사용하게 되었다.

참고로 위 표4의 결과를 위한 컴퓨터 프로그램(BMDP)은 아래와 같이 전개된다.

```
/ input file is 'c:\quit\obes6.dat'.
  code is sysusr3.

/ variables use= purecat, age.

/ group codes(case) are 1,0.
  names (case) are case, control.
  cutpoints(age) are 55.
```

```
/ print linesize=80. level=min.

/ trans if (purecat=6 or purecat=9) then case=1.
        if (purecat=15) then case=0.

/ regress dependent is case.
          categ=age.
          dvar=part.
          model = age.
          method=mlr.

/ end
```

3) 범주변수가 다항변수(2xr표)를 이루는 경우

위의 예에서는 연령을 55세 이상과 55세 미만의 2군으로 나누었다. 그러나 어떤 변수는 2개 이상의 수준을 지니는 범주변수인 경우가 있다.

예를 들어 CHD에 대한 연구에서 종족을 나타내는 변수 RACE는 흑인, 백인, 히스패닉, 기타의 4종류로 부호화 되었다고 하자. RACE와 CHD 상태에 의한 분할표는 표 3.5와 같다. 이 자료는 설명을 쉽게하기 위하여 여기서 임의로 작성한 가상적인 자료이다.

변수가 4개 이상으로 범주화 되었다고 하여도 전혀 설명이 다르지 않으므로 이장에서는 4개의 수준으로 설명을 하고자 한다.

표 5. RACE와 CHD 발생에 의한 2 x 4 분할표

CHD	Black	Hispanic	Other	White	합
있음(1)	20	15	10	5	50
없음(0)	10	10	10	20	50
합	30	25	20	25	100
Odds ratio	8.0	6.0	4.0	1.0	
95% CI	(2.3,27.6)	(1.7,21.3)	(1.1,14.9)		
ln(Ψ)	2.08	1.79	1.39	0.0	

표 5의 맨 아래칸에 White(백인)을 기준으로 하였을 경우의 각각의 odds ratio를 산출한 것이 나타나 있다. 예를 들어 hispanic의 odds ratio추정치는 (15x20)/(5x10)=6.0이다.

odds ratio의 log는 표 5의 마지막 줄에 나타나 있다. 이러한 형태의 표는 한 군을 표준군으로 하고 이에 대한 다른 군들의 비교를 행하는 분석의 경우에 해당하는 것이다. 이 자료를 이용하여 로지스틱 회귀분석을 행하면, 적절한 다자인 변수를 사용하였을 경우에 위 표와 동일한 odds ratio를 얻게 된다. 여기서 사용하여야 하는 디자인 변수는 기준이 되는

범주인 White를 0로 놓고 다른 범주들을 1로 부호화하는 방법인데 이러한 방법은 "reference cell coding"으로도 불리는바 BMDPLR에서 partial method를 사용하면 이와 같은 디자인 변수로 분석이 행하여지게 된다.

표 6은 로지스틱 회귀분석에 의하여 산출된 결과의 일부를 제시한 표이다.
표에서 보는 바와 같이 로지스틱 회귀분석에서 산출한 odds ratio와 2 x 4 분할표에서 산출한 odds ratio는 동일하다.

표 6. 표5의 자료를 이용하여 로지스틱 회귀모형으로 분석한 결과

변수	추정회귀계수	표준오차	계수/표준오차	φ
RACE(1)	2.079	0.633	3.29	8.0
RACE(2)	1.792	0.646	2.78	6.0
RACE(3)	1.386	0.671	2.07	4.0
상수	-1.386	0.500	-2.77	

참고로 표 6의 결과를 위한 프로그램(BMDP)은 아래와 같이 전개된다.

```
/ group codes(case) are 1,0.
  names (case) are case, control.

/ regress dependent is case.
          categ= race
          dvar=part.
          model= race
          method=mlr.
/ end
```

4) 범주변수가 순위변수(ordinal)를 이루는 경우

표 3에서 연령을 55세 이상과 미만의 2군으로 나눈 것을 더 자세히 구분하여 20-34세, 35-44세, 45-54세, 55세 이상의 4군으로 구분하는 경우를 생각하여 볼 수 있다.

이를 2 x 4 표로 작성하면 아래와 같다.

표 7. 4개의 수준으로 구분된 연령과 CHD 발생간의 2 x 4표

CHD	연령군 20-34	35-44	45-54	55-64	Total
있음(1)	3	8	11	21	43
없음(0)	22	19	10	6	57
합	25	27	21	27	100
odds Ratio(φ)	1.0	3.1	8.1	25.7	
ln(φ)	0.0	1.1	2.1	3.2	

표에서 보는 바와 같이 CHD 발생의 위험도가 기준연령군 (reference group)에 비하여 연령이 증가하면서 점차 높아지는 것을 보여주고 있다. 그렇다면 이러한 발생위험도의 증가가 실제로(구조적으로) 연령과 관련되어 있는지를 질문할 수 있다. 이것은 연령에 대한 Trend test와 동일한 의미를 지닌다.

BMDP 프로그램은 이에 대한 검정을 가능하게 하여 준다. 이 프로그램에서는 3가지 종류의 trend test를 지니고 있는데 1) linear 2) quadratic 3) cubic의 3종류로 구성되어 있다.

표 7의 자료를 이용하여 실제로 분석한 후에 나타난 계수들을 열거한 것이 표 8이다.

표 8. 표 7의 자료를 이용하여 선형관계의 유무를 검정한 결과

변수	추정회귀계수	표준오차	계수/표준오차
Age group(1)	2.382	0.534	4.48
Age group(2)	0.150E-01	0.490	0.03
Age group(3)	0.814E-01	0.442	0.18
상수	-0.377E-01	0.245	-0.54

이미 언급하였듯이 위 결과에서 (1)은 linear한 관계를 (2)는 quadratic 한 관계를 (3)은 cubic 한 관계를 나타내고 있는 바 표 8 결과의 맨 오른쪽 항인 '계수/표준오차'의 값이 1.96보다 큰 것은 (1)번의 linear 한 관계이기 때문에 이 자료에서 연령을 CHD 발생과 linear한 trend를 지니고 있는 것으로 해석할 수 있다.

참고로 표 8의 결과를 위한 BMDP 프로그램은 아래와 같다.

```
/ group codes(case) are 1,0.
  names (case) are case, control.
  cutpoints(age) are 34, 44, 54.

/ regress dependent is case.
          interval=age.
          dvar=orth.
          model= age.
          method=mlr.
/ end
```

5) 독립변수가 연속변수일 때

종속변수가 범주형 자료인데 독립변수가 연속형 자료인 경우를 위하여 사용하는 것이 로지스틱 회귀분석이라는 것은 이미 설명한 바와 같다. 이러한 경우에 회귀계수의 의미는

연령이 1세 증가함에 따라서 CHD 발생의 위험도가 얼마나 증가하는가를 알고자 하는 것이다.

예를 들어 범주변수인 비만도와 연속변수인 연령이 CHD의 발생에 미치는 영향을 가상적인 예를 가지고 분석하였더니 다음과 같은 결과가 나왔다고 하자.

EXP(COEF) TERM	회귀계수	표준오차	계수/표준오차	EXP(계수)
비만도	0.3217E-01	0.148	0.217	1.03(0.772-1.38)
연령	0.3427E-01	0.673E-02	5.10	1.03(1.02-1.05)
CONSTANT	-2.649	0.435	-6.09	

위 결과는 연령이 1세 늘어갈수록 CHD 발생위험도가 1.03배 증가한다는 것을 의미한다. 그런데 이를 1세마다의 증가율로 표현하기 보다는 5, 10 등 5의 배수로 변화를 나타내면 대부분 이해가 쉬워진다. 여기서 한가지 조심할 것은 연령이 10세 늘어갈때 CHD 발생위험도가 1.03 x 10 = 10.3 이 되는 것이 아니고 아래와 같은 방식에 의하여 연령이 10세 증가할 때 CHD 발생가능성이 1.41배 증가하는 것으로 해석하여야 한다는 점이다.

eg) OR(10) = exp (0.03427 x 10) = 1.41

exp (0.03427 x 10 ± 1.96 x 10 x 0.00673) = (1.61, 1.23)

(for Age Variable)

이를 위하여 사용하는 BMDP 프로그램은 아래와 같다.

```
/ group codes(case) are 1,0.
  names (case) are case, control.
  cutpoints(BMI) are 26.9.

/ regress dependent is case.
          categ=BMI.
          interval=age.
          dvar=part.
          model = BMI, age
          method=mlr.

/ end
```

6) 중 로지스틱 회귀분석(multiple logistic regression)

지금까지 우리는 모형에 독립변수 한개(single variable)만이 포함되어 있는 경우의 로지스틱 회귀계수 추정치의 해석에 대하여 논하였다. 여러가지 변수들을 한개씩만 포함시킨 모형이 그 자료에 대한 적절한 분석을 제공하여 주는 경우는 드물게 되는데 그 이유는 몇가

지 독립변수들이 함께 영향을 주고 있으며, 동시에 종속변수(outcome variable)의 각 수준마다 각각 다른 분포를 나타낼 수 있기 때문이다. 따라서 우리는 자료를 더 포괄적으로 나타내 줄 수 있는 모형을 위한 다변량 분석을 고려하게 된다. 이러한 분석의 목표중의 하나는 모형에 포함된 각 변수의 추정시에 다른 독립변수들의 분포가 다른 것과 다른 독립변수들 간의 관련성에 대하여 통계적으로 보정(statistically adjust)한 후에 효과를 추정하고자 하는 것이다. 이러한 개념을 지닌 다중 로지스틱 회귀모형에 적용하면 개개의 계수추정치는 모형에 포함된 다른 모든 변수들의 영향을 조정한 후의 추정치를 나타내게 된다.

이러한 다중회귀(multiple regression)의 곡선을 설정하는데 있어 명심해야 할 것은 (1) 독립변수들을 설정하는 기본계획을 수립해야 하며 (2) 각각의 변수의 적합성 및 전체방정식의 타당성이다. 여기서 적합성(adequacy)이라 함은 단순한 통계적 의미 뿐만 아니라 각 변수들의 의미가 경험적으로나 상식적으로 납득할 수 있어야 함을 뜻한다.

다중회귀 모형에 포함되는 변수를 선정하는 원칙 및 절차는 다음과 같다.

회귀방정식의 변수를 선택하는데 있어서 일반적인 원칙은 독립변수의 갯수가 적을수록 좋다는, 이른바 최소화의 원칙(Rationale for minimizing the number of variables)이다. 이 이론의 핵심은, 독립변수의 갯수가 적을수록 수리적으로 안정되고 방정식의 일반적 해석이 쉬워진다는 것이다. 또한 변수의 수가 많아지면 표준오차(standard error)가 커지므로 자료의 크기에 따라 계수들의 유의성이 잘 변동하게 된다는 것이다.

최근에 이 최소화의 원칙과 반대되는 이른바 최대화의 원칙이라고 부를 수 있는 이론도 제시되었다. 이러한 주장의 근거는 한 변수만을 모형에 포함시켰을 때 잘 나타나지 않던 교란요인의 효과가 여러가지 집합적으로 작용하면서 교란효과를 나타낼 수 있다는 것에 의한다. 로지스틱 회귀방정식(logistic regression model)에서 변수를 설정하는 데는 다음의 절차가 일반적으로 통용된다.

(1) 다중회귀방정식에 포함될 각 독립변수들이, 우선 단순(univariate) 회귀방정식에 있어서 종속변수와 유의한 인과관계에 있어야 한다. 범주변수, 순위변수, 그리고 몇가지 종류의 수에 한정된 연속변수인 경우에는 독립변수를 K 수준으로 변형하여 2 X K 분할표에 의한 분석을 행하는 것이 바람직하다.

단순회귀분석에서 t값이 유의한 변수들은 일단 다중회귀분석에 포함될 자격을 갖춘 것으로 인정한다.

(2) 분석하고자 하는 변수들의 단순회귀분석을 끝낸 후 다중회귀분석을 위한 변수선정에 들어간다. p-value 가 0.25 이하인 변수들은 일단 다중회귀분석에 이용될 자격이 생긴다고 볼 수 있다. 일반적인 유의수준(0.05)을 기준으로 했을 때는 중요한 변수들이 누락될 가능성이 있기 때문에 이 과정에서는 유의수준을 0.25를 기준으로 하는 것이 널리 인정되고 있다.

개개변수에 대한 검정이 끝나면 이 변수들 모두를 포함하는 다중회귀 모형이 과연 자료에 적합한지를 검정하여 보면 포함된 변수들 중 일부는 다중회귀모형에서 통계적으로 유의하지 않게 된다. 다중회귀 모형에서 통계적으로 유의하지 않게 나온 변수가 의학적으로 종속변수에 매우 유의한 영향을 준다고 알려져 있으면 비록 유의하지 않다고 하더라도 계속 모형에 포함시키는 경우도 있으며 그렇지 않을 경우에는 이 변수를 제거한 후의 다중 로지스틱 모형을 가지고 분석을 반복하게 된다. 이와 같은 과정을 거쳐서 모든 변수가 유의하게 남아있게 되면 이러한 변수들로 구성된 모형을 선택한다.

변수선정의 또 다른 방법으로는 stepwise 방법이 있으나 이는 연구자가 변수를 선택하기 보다는 컴퓨터 통계프로그램의 변수를 선택하는 편에 가깝기 때문에 적은 수의 변수들을 가지고 변수선정을 하는 경우에는 추천할 만한 방법이 되지 못한다.

변수선택의 또 다른 방법으로서 소규모 변수군 선정에 의한 선별(best subsets selection)을 들 수 있는데 이것은 2,3,4, .. 개의 변수를 포함하는 모델들을 여러개 조합하여 여러가지 통계적 기준에 따라 그중 가장 적합한 것을 골라내는 방법이다.

이 모든 방법들을 동원하여 적합한 모델을 선정하는 과정에서 우리가 꼭 간과할 수 없는 것은, 이 여러가지 방법들 중 그 어떤 방법도 모든 변수들을 완전히 다 검토하기는 어렵다는 점이다.

(3) 위의 단계를 거쳐서 다중회귀방정식을 설정한 후에는 interaction 변수를 포함시킬지를 결정하여야 한다. 어떠한 모델에 있어서든 두 변수간의 interaction은 한 변수의 영향이 다른 변수의 수준에 따라 유동적(not constant)라는 것이다. 예를 들어 성(sex)와 연령(age)의 interaction 이란 연령(age)의 회귀계수(slope coefficient)가 남성(male)과 여성(female)에 따라 다른 경우에 상호반응(interaction)관계가 존재한다고 인정할 수 있다.

Interaction 변수를 다중회귀모델에 포함시킬 지의 여부는 그 interaction variable 을 첨가했을 때 다른 변수들의 계수의 유의성을 감소시키지 않으면서 그 interaction 변수가 다중회귀방정식에서 유의한 계수를 갖는지의 여부에 따라 정해진다.

7) 결론

이상으로 종속변수가 범주형 자료이면서 독립변수가 연속형 자료인 경우에 사용 가능한 로지스틱 회귀모형에 대하여 설명하였다. 위에서 사용한 예에서 알 수 있듯이 종속변수가 범주형 자료이면서 독립변수도 범주형 자료이거나 또는 종속변수가 범주형 자료이면서 독립변수가 연속형 자료와 범주형 자료가 혼합되어 있는 경우에도 물론 로지스틱 회귀모형의 응용이 가능하다. 이러한 로지스틱 회귀모형의 장점은 이 모형에서 추정된 회귀계수에 antilog를 취하여 odds ratio를 추정할 수 있다는 점에 있다.

참 고 문 헌

1. Hosmer, D.W., Lemeshow, S. (1989). Applied logistic regression. John Wiley & Sons, New York.
2. SAS Institute Inc. (1988). SAS/STAT User's Guide, Releds 6.03 Edition. Cary. NC: SAS Institute Inc.