07 강

데이터분석방법론2

로지스틱회귀모형 (3)

통계·데이터과학과 이기재 교수

학습목차

학습개요 및 목표

이번 강의는 로지스틱회귀모형의 적용과 관련한 모형선택, 모형진단 방법에 대해서 공부하고, 표본크기가 충분하게 크지 않은 경우에 발생하는 희박한 자료(Sparse Data)와 관련한 문제에 대해서 살펴보겠습니다.

1. 로지스틱회귀모형 적합도 검증과 모형 선택을 설명할 수 있다.
2. 로지스틱회귀모형 모형 진단과 잔차 분석을 설명할 수 있다.
3. 희박 자료의 의미와 추정에 미치는 영향을 설명할 수 있다.

제 7장. 로지스틱회귀모형(3)

1 모형선택의 전략

2 모형진단 방법

3 로지스틱 회귀분석의 무한대 추정값

4 베이지안 추론, 벌점가능도 추정법

01

제 7장. 로지스틱회귀모형(3)

모형선택의 전략

1. 모형 선택 고려사항

❑ 모형의 선택 과정에서 고려사항

- 자료에 대한 적합성
  : 모형이 복잡해질수록 유리
- 적합된 모형의 해석의 용이성
  : 모형이 간단할수록 유리

확증적 (Confirmatory) 연구와
탐색적 (Exploratory) 연구

2. 얼마나 많은 예측변수를 사용할 수 있는가?

가이드라인: P.Peduzzi 등, *J. Clin. Epidemiol.*, 49: 1373-1379, (1996)

- The data set should contain at least 10 outcomes of each type for every explanatory variable.

예

① $n = 1000$, $Y = 1$인 경우 30,
$Y = 0$인 경우 970이라면
➔ 모형에 포함되는 예측변수는 3개 이하가 바람직함

② $n = 173$ 참게, $Y = 1$인 경우 111건,
$Y = 0$인 경우 62건
➔ 6개 이하의 예측변수를 사용하는 것이 바람직함

2. 얼마나 많은 예측변수를 사용할 수 있는가?

- **가이드라인:** P.Peduzzi 등, *J. Clin. Epidemiol.*, 49: 1373-1379, (1996)
  - 이와 같은 가이드라인은 다소 보수적이어서 이 조건을 만족하지 않더라도 모형적합 결과를 얻을 수 있음
  - 이 가이드라인을 과도하게 위배하는 경우는 효과에 대한 ML 추정값은 매우 편향되고 표준오차의 추정값도 매우 클 수 있음
  - 여러 개의 예측변수를 갖는 모형은 다중공선성(multi-collinearity)의 문제가 발생할 수 있음

3. 예제: 암참게의 부수체 자료 재분석

- $Y=$ 암 참게가 부수체를 갖고 있는지 여부
  (1 = yes, 0 = no)
- 예측변수
  - 무게(Weight)
  - 너비(Width)
  - 색깔(ML, M, MD, D), c_2, c_3, c_4 의 가변수로 나타냄
  - 등뼈 상태(3개 범주), s_2, s_3 의 가변수로 나타냄

3. 예제: 암참게의 부수체 자료 재분석

- $$logit[P(Y=1)] = \alpha + \beta_1 weight + \beta_2 width + \beta_3 c_2 + \beta_4 c_3 + \beta_5 c_4 + \beta_6 s_2 + \beta_7 s_3$$

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_7 = 0$ 에 대한 가능도비 검정 결과

  $-2(L_0 - L_1) =$ 이탈도의 차이

  $= 225.8 - 185.2 = 40.6$

  $df = 7,\ P-$ 값 < 0.0001

적어도 하나의 예측변수는
효과가 있다는 강한 증거

3. 예제: 암참게의 부수체 자료 재분석

```
------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                     header=TRUE)
> fit <- glm(y ~ weight + width + factor(color) + factor(spine),
+            family=binomial, data=Crabs)

> summary(fit)
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)     -8.06501     3.92855   -2.053    0.0401
weight           0.82578     0.70383    1.173    0.2407
width            0.26313     0.19530    1.347    0.1779
factor(color)2  -0.10290     0.78259   -0.131    0.8954
factor(color)3  -0.48886     0.85312   -0.573    0.5666
factor(color)4  -1.60867     0.93553   -1.720    0.0855 .
factor(spine)2  -0.09598     0.70337   -0.136    0.8915
factor(spine)3   0.40029     0.50270    0.796    0.4259
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 185.20  on 165  degrees of freedom
AIC: 201.2
> 1 - pchisq(225.76-185.20, 172-165) # P-value for test that all beta's = 0
[1] 9.83292e-07
> library(car)
> Anova(fit) # likelihood-ratio tests for individual explanatory variables
              LR Chisq  Df  Pr(>Chisq)
weight          1.4099   1     0.23507
width           1.7968   1     0.18010
factor(color)   7.5958   3     0.05515 .
factor(spine)   1.0091   2     0.60377
------------------------------------------------------------------------------
```

3. 예제: 암참게의 부수체 자료 재분석

참게 자료의 주 효과 모형에 대한 모수 추정값

```
> summary(fit)
                 Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)     -8.06501     3.92855   -2.053    0.0401
weight           0.82578     0.70383    1.173    0.2407
width            0.26313     0.19530    1.347    0.1779
factor(color)2  -0.10290     0.78259   -0.131    0.8954
factor(color)3  -0.48886     0.85312   -0.573    0.5666
factor(color)4  -1.60867     0.93553   -1.720    0.0855 .
factor(spine)2  -0.09598     0.70337   -0.136    0.8915
factor(spine)3   0.40029     0.50270    0.796    0.4259
---
```

3. 예제: 암참게의 부수체 자료 재분석

- 각 예측변수의 효과에 대한 Wald 검정 결과를 보면
  거의 유의하지 않은 것으로 분석됨

예측변수간의 강한 상관성으로 인한
다중공선성(Multicollinearity)으로 인해서
각 모수에 대한 유의성이 없는 것처럼 분석된 것으로 보임

(Weight 와 Width의 상관계수 = 0.89)

➔ Weight 와 Width를 모두 예측변수로 사용할 필요는 없고,
앞으로 Width만 예측변수로 사용함

4. 단계적 변수 선택법

전진선택법
(Forward Selection Procedure)

더 이상 적합이 개선되지 않을 때까지 항(예측변수)을 추가해 모형을 적합하는 방법

후진제거법
(Backward Elimination Procedure)

복잡한 모형에서 시작해서 항을 제거하면서 모형을 적합하는 방법

※ 범주형 예측변수에 대해서는 전체 지시변수(가변수)를 함께 모형에 포함하거나 빼야 함

5. 목적에 따른 설명변수의 선택

- 예측변수를 선택할 때 연구 목표, 상대적 통계적 유의성, 다중공선성 및 잠재적 교란요인 등의 문제를 고려해서 변수 선택 전략을 정함

- Hosmer 등(2013, 4장)에서 제안한 변수 선택 과정

1. 초기 주효과 모형 적합: 잘 알려진 중요 변수, 단일 예측변수 모형에서 어느 정도 연관성(예: P값<0.2)이 있는 변수 이용

2. 후진제거법 수행: 더 엄격한 유의수준에서 유의한 변수, 교락변수 남김

3. 1단계 모형에 포함되지 않았지만, 2단계 이후 유의해진 변수 추가

4. 3단계 모형의 변수들간의 교호작용 여부 확인

5. 후속적인 모형 진단 작업 진행

6. 예제: 참게 자료에서의 변수 선택

참게 자료에 후진제거법의 적용

- 예측변수로 W = width, C = 색깔, S = 등뼈 상태 등을 고려함
- 교호작용을 포함하는 복잡한 모형을 적합함
- 가장 높은 차수의 항 중에서 "가장 덜 유의한"(P-값이 가장 큰) 예측 변수를 제외하고 다시 모형을 적합
- 남은 예측변수가 모두 유의할 때까지 위의 과정을 계속함

6. 예제: 참게 자료에서의 변수 선택

참게 자료에 대한 여러 로지스틱회귀모형 적합 결과

Table 5.1 Results of fitting several logistic regression models to predict horseshoe crab satellites.

Model	Explanatory Variables	Deviance	*df*	AIC	Models Compared	Deviance Difference
1	None	225.8	172	227.8		
2	C	212.1	169	220.1	(2) - (1)	13.7 ($df = 3$)
3	S	223.2	170	229.2	(3) - (1)	2.5 ($df = 2$)
4	W	194.5	171	198.5	(4) - (1)	31.3 ($df = 1$)
5	$C + W$	187.5	168	197.5	(5) - (2)	24.6 ($df = 1$)
					(5) - (4)	7.0 ($df = 3$)
6	$C + W + S$	186.6	166	200.6	(6) - (5)	0.9 ($df = 2$)
7	$C + W + C*W$	183.1	165	199.1	(7) - (5)	4.4 ($df = 3$)

Note: C = color, S = spine condition, W = width.

➔ 최종 모형은 너비와 색깔의 주효과만을 갖는 모형

모든 모형은 정확한 모형이라기 보다는 “실제 현상을 단순화시킨 것일 뿐임”

7. AIC와 편향/분산 간의 절충

- 모형 선택: 편향과 분산 사이에 근본적인 절충 문제 발생
- Parsimony (Simplicity) is good
- 모형선택에서 기준(AIC, Akaike Information Criterion)을 이용 할 수 있음 : AIC 값이 최소인 모형을 선택함
- AIC = -2(로그가능도) +2(모형에 있는 모수 개수)
  BIC = -2(로그가능도) +log(n)×(모형에 있는 모수 개수)
  BIC (Bayes Information Criterion), 표본크기가 큰 경우에 유용함(통계적 유의성, 실질적 유의성)
- 탐색적 연구라면 후진제거법과 같은 자동화 방법을 사용 가능
- 각 예측변수에 대하여 반응변수의 각 수준에서 적어도 10개의 관측치가 있는 것이 바람직함

7. AIC와 편향/분산 간의 절충

- AIC를 이용하여 단계적으로 모형선택을 할 수 있음

➔ AIC를 이용하여 후진제거법으로 모형을 선택하는 예

```
-------------------------------------------------------------------------
> fit <- glm(y ~ weight + width + factor(color) + factor(spine),
+            family=binomial, data=Crabs)
> library(MASS)
> stepAIC(fit) # stepwise backward selection using AIC
Start:  AIC=201.2
y ~ weight + width + factor(color) + factor(spine)
Step:  AIC=198.21
y ~ weight + width + factor(color)
Step:  AIC=197.46
y ~ width + factor(color) # AIC now increases if width or color removed
-------------------------------------------------------------------------
```

02

제 7장. 로지스틱회귀모형(3)

모형진단 방법

1. 적합도 검정: 이탈도 통계량 이용

◪ 이탈도 통계량을 이용한 모형비교?

- 해당 모형과 더 복잡한 모형을 비교하는
  가능도비 검정을 통해서 적합결여 여부를 검증하는 방법

더 복잡한 모형을 적합하더라도 현재 고려하고 있는 모형과 비교하여 적합 정도가 개선되지 않는다면 이미 선택된 모형이 적합하다고 할 수 있음

1. 적합도 검정: 이탈도 통계량 이용

예제 : 참게자료 사례

- $logit[\pi(x)] = \alpha + \beta x, \quad x$ 너비
- $logit[\pi(x)] = \alpha + \beta_1 x + \beta_2 x^2$

- 귀무가설 $H_0 : \beta_2 = 0$ 을 검정하는 가능도비
- 검정 통계량은 0.83 이고, $df = 1$ 임
- 따라서 $P-$ 값 $= 0.36$

1. 적합도 검정: 이탈도 통계량 이용

- M : 현재 고려하고 있는 모형

포화모형	각각의 관측치에 대해 개별적인 모수를 갖는 모형

- 모형 M 의 적합도 검정(Goodness of fit test)
  - 포화모형에는 포함되어 있지만
    모형 M 에는 포함되지 않는 모든 모수가
    “0”인지 검정하는 것

1. 적합도 검정: 이탈도 통계량 이용

- GLM에서 적합도 검정을 위한 가능도비 통계량

① 이탈도(Deviance) = $-2[L_M - L_S]$

② 대표본의 경우 근사적으로 카이제곱분포를 따름

③ 검정통계량 값이 크고,
P-값이 작을수록 모형의 적합결여에 대한
강한 증거가 됨

1. 적합도 검정: 이탈도 통계량 이용

❑ 예측변수가 모두 범주형 변수인 경우

- 전체 Data는 예측변수들의 i번째 조합에 대해 분할표 도수로 요약됨
- 예

	1	0	계
⋮			⋮
i	O_{1i}	O_{2i}	n_i
⋮			⋮

[적합값(Fitted Value)]

- "1"의 개수 = $n_i \cdot \hat{\pi}_i$
- "0"의 개수 = $n_i \cdot (1 - \hat{\pi}_i)$

1. 적합도 검정: 이탈도 통계량 이용

- $G^2(M) = 2\sum$ 관찰값 $[\log$ (관찰값/적합값)$]$
- $X^2(M) = \sum$ (관찰값-적합값)^2 / 적합값
- 모든 적합도수가 5 이상일 때,
  $G^2(M)$와 $X^2(M)$은 근사적으로 카이제곱분포를 따름

  자유도 = (포화모형에서의 모수의 수) - (해당 모수의 수)

- $X^2(M)$ 값과 $G^2(M)$ 값이 클수록
  적합이 결여된 것을 의미함

2. 예제: 마리화나 사용 조사

마리화나 사용 자료

- Data

인종(X)	성별(Z)	마리화나 사용(Y)	
		예(1)	아니오(0)
백인	여자	420	620
	남자	483	579
다른 인종	여자	25	55
	남자	32	62

- 로지스틱회귀모형 적합

$$\text{logit}(\hat{\pi}) = -0.8303 + 0.2026 \times \text{GENDER} + 0.4437 \times \text{RACE}$$

2. 예제: 마리화나 사용 조사

```
------------------------------------------------------------------------------
> Marijuana
   race gender yes  no
1 white female 420 620
2 white   male 483 579
3 other female  25  55
4 other   male  32  62
> fit <- glm(yes/(yes+no) ~ gender + race, weights=yes+no, family=binomial,
+            data=Marijuana)
> fit$deviance; fit$df.residual
[1] 0.05798  # residual deviance goodness-of-fit statistic
[1] 1        # residual df
> 1 - pchisq(fit$deviance, fit$df.residual)
[1] 0.80972 # P-value for deviance goodness-of-fit test
> fitted(fit)
      1       2       3       4
0.40453 0.45413 0.30357 0.34802 # estimated prob's of marijuana use
> fit.yes <- n*fitted(fit); fit.no <- n*(1 - fitted(fit))
> attach(Marijuana)
> data.frame(race, gender, yes, fit.yes, no, fit.no)
   race  gender  yes    fit.yes   no     fit.no
1 white  female  420  420.71429  620  619.28571
2 white    male  483  482.28571  579  579.71429
3 other  female   25   24.28571   55   55.71429
4 other    male   32   32.71429   62   61.28571
------------------------------------------------------------------------------
```

3. 적합도 검정: 그룹, 비그룹, 연속형

◪ 그룹화·비그룹화된 자료와 연속성 예측변수

- 예측변수가 범주형인 경우 자료 파일의 구분

그룹화된 자료	각 예측변수의 조합에서 관측된 성공과 실패의 총합자료(분할표 형식으로 요약된 경우)
비그룹화된 자료	0과 1로 표현된 관측값으로 분할표 등으로 요약되기 전의 원자료

3. 적합도 검정: 그룹, 비그룹, 연속형

◪ 그룹화·비그룹화된 자료와 연속성 예측변수

모수의 ML추정값	위의 두 가지 형태의 자료에 대해서 동일함
적합도 검정	그룹화된 자료의 경우에만 적용할 수 있음

- G^2과 X^2는 적합도수가 5 이상인 분할표에 대해서 적용

3. 적합도 검정: 그룹, 비그룹, 연속형

◪ 그룹화·비그룹화된 자료와 연속성 예측변수

- 연속형 또는 연속형에 가까운 예측변수를 갖는 경우의 로지스틱회귀 모형의 적합도 검정 방법

검정방법

① 각 예측변수를 범주화하여(예: 사분위수를 이용하여 4개의 범주로 구분) 그룹화된 자료의 관찰도수와 적합 도수에 대해 G^2과 X^2를 적용함

② 예측된 확률 $(\hat{\pi}_i)$을 크기 순에 따라 나열하고, 자료를 그룹화하여 관찰값과 적합값을 구하여 검정하는 방법 : Hosmer-Lemeshow 검정

- ➲ 설명변수의 모든 가능한 조합의 수 J가 전체 표본 크기 n과 같거나 유사한 경우로 가정
- ➲ 예측확률이 작은 값부터 큰 값 순서로 개체들을 배열한 후 백분위점을 기준으로 n/g개씩 개체들로 하나의 범주를 구성함 ($g \times 2$ 분할표로 변환됨)

3. 적합도 검정: 그룹, 비그룹, 연속형

◪ 그룹화·비그룹화된 자료와 연속성 예측변수

- 연속형 또는 연속형에 가까운 예측변수를 갖는 경우의 로지스틱회귀 모형의 적합도 검정 방법

검정방법

② 예측된 확률 $(\hat{\pi}_i)$ 을 크기 순에 따라 나열하고, 자료를 그룹화하여 관찰값과 적합값을 구하여 검정하는 방법 : Hosmer-Lemeshow 검정

➲ Hosmer-Lemeshow 검정통계량

$$\hat{c} = \sum_{k=1}^{g} \frac{(O_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

➲ $\hat{c}$ 는 자유도 $g-2$ 인 카이제곱분포로 근사됨

➲ $\hat{c}$ 값이 클수록 적합결여에 대한 강한 증거임

```
PROC LOGISTIC ;
    MODEL y=width / lackfit ;
```

4. 로지스틱 모형의 잔차, 표준화 잔차, 이탈도 잔차

- 범주형 예측변수에 대해서,
  관측도수와 적합도수를 비교하기 위하여 잔차를 사용함

- Pearson 잔차
  - $e_i = \dfrac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$ (SAS GENMOD에서 Reschi 로 표현)

    - y_i : "성공"한 도수
    - n_i : 전체 시행횟수
    - π_i : 적합된 모형으로부터 구한 π_i 의 예측값

  - $e_i \sim N(0, v), \quad v < 1$

* 포아송 GLM에서
- Pearson 잔차: $e_i = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$
- 잔차의 제곱합 = 적합도 카이제곱 통계량
  → 이탈도 잔차의 제곱합으로 분해

4. 로지스틱 모형의 잔차, 표준화 잔차, 이탈도 잔차

표준화 잔차(Standardized Pearson Residual)

- $r_i = \dfrac{y_i - n_i\hat{\pi}_i}{SE} = \dfrac{e_i}{\sqrt{1-h_i}}$ (StReschi로 표시)

  h_i : 관측값의 레버리지(Leverage)를 나타내며 첫 행렬(Hat Matrix)의 대각원소임

- 관측값의 레버리지가 클수록, 모형적합에 미치는 잠재적인 영향력이 커짐
  - 근사적으로 $r_i \sim N(0,1)$
  - $|r_i| > 2$ or 3 이면 모형의 적합결여를 시사함

5. 로지스틱 모형에서 영향점 진단

- 로지스틱 회귀분석의 영향 측도들은 그룹화된 자료 파일의 관측값들에 대해서 유용하게 사용됨
- 영향력을 측정하기 위한 측도: 대개 전체 자료에서 한 관측값을 제거했을 때에 추정에 미치는 효과를 나타냄
  - 표준화잔차 및 이를 기반하는 측도들(예: Cook의 거리)
  - 모형의 각 모수에 대해서 한 관측값을 제거했을 때, 모수의 추정값에 발생하는 변화량,(예: Dfbeta(변화량을 표준오차로 나눈 값))
  - 관측값을 제거했을 때의 이탈도 감소량

6. 예제: 심장병과 혈압

Table 5.3 Diagnostic measures for logistic regression model fitted to heart disease data.

Blood Pressure	Sample Size	Observed Disease	Fitted Disease	Standardized Residual	*Dfbeta*	Deviance Decrease
111.5	156	3	5.2	−1.11	0.49	1.39
121.5	252	17	10.6	2.37	−1.14	5.04
131.5	284	12	15.1	−0.95	0.33	0.94
141.5	271	16	18.1	−0.57	0.08	0.34
151.5	139	12	11.6	0.13	0.01	0.02
161.5	85	8	8.9	−0.33	−0.07	0.11
176.5	99	16	14.2	0.65	0.40	0.42
191.5	43	8	8.4	−0.18	−0.12	0.03

Source: J. Cornfield, *Fed. Proc. 21, Suppl.* **11**: 58–61 (1962). Data are in `HeartBP` data file at text website.

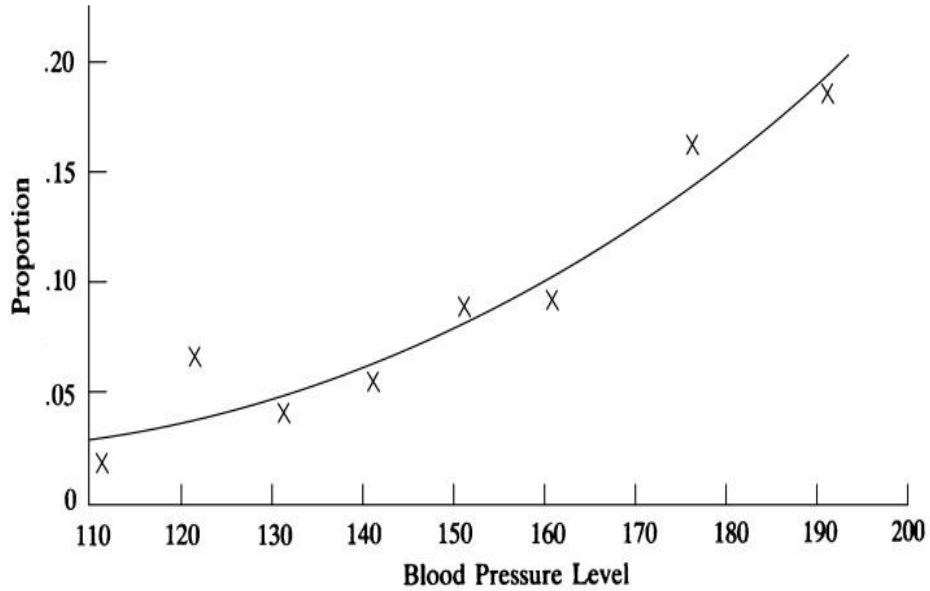

Figure 5.1 Observed proportion (x) and fitted probability of heart disease (curve) for the linear logistic model.

03

제 7장. 로지스틱회귀모형(3)

로지스틱 회귀분석의 무한대 추정값

1. 완전과 준완전 분리: 완전한 분류

양적 예측변수에서 무한대 효과의 추정

- $y = 0 \text{ for } x < 50 \text{ and } y = 1 \text{ for } x > 50$
  (완전한 판별식의 경우)
- $\text{logit}[P(Y=1)] = \alpha + \beta x$
  ➔ $\hat{\beta} = \infty$
- GLM에서 모수 적합은 Fisher의 스코어 알고리즘을 사용하면 대개 잘 수렴함. 하지만 성공이 완전히 판별되는 경우, 성공이나 실패 중 하나만 관찰된 경우는 ML 추정값이 무한대 또는 존재하지 않을 수 있음

2. 희박한 자료 (Sparse Data)

희박한 자료(Sparse Data)란?

- 작은 도수를 갖는 칸들이 많은 분할표의 경우를 말함
- 예측변수가 많거나
  다수의 수준으로 분류된 분할표에서 흔히 발생함
- 표집영(Sampling Zero)

이론상으로 그 칸에 속한 관측값이 가능함
But 현재 Data 상으로는 해당 칸의 도수가 0인 경우
(표본의 크기가 충분히 커지면 양의 도수를 가질 수 있음)

2. 희박한 자료 (Sparse Data)

표집영 (Sampling Zero)

- 모형에 따라 표집영은 모형의 모수에 대한 ML추정값이 무한대가 되는 원인이 될 수 있음

예

	S	F
1	8	2
0	10	0

- 모형 $\log[\frac{P(S)}{P(F)}] = \log[\frac{\pi}{1-\pi}] = \alpha + \beta x$
- $e^{\hat{\beta}} = odds\ ratio = \frac{8 \times 0}{2 \times 10} = 0$

$\hat{\beta} = \log^{[odds\ ratio]} = -\infty$

3. 사례

희박한 자료를 가진 임상시험 자료

센터	처리	반응변수		분할표	
		성공	실패	성공	실패
1	Active drug	0	5	0	14
	Placebo	0	9		
2	Active drug	1	12	1	22
	Placebo	0	10		
3	Active drug	0	7	0	12
	Placebo	0	5		
4	Active drug	6	3	8	9
	Placebo	2	6		
5	Active drug	5	9	7	21
	Placebo	2	12		
XY 분할표	Active drug	12	36		
	Placebo	4	12		

3. 사례

희박한 자료를 가진 임상시험 자료

- $\mathrm{logit}[P(Y=1)] = \alpha + \beta x + \beta_k^z$
- 센터 1과 3에서 성공한 경우가 없음
  ➔ β_1^z와 β_3^z의 ML추정값은 $-\infty$가 됨

참고

- 성공이나 실패가 한 번도 없는 센터들은 모비율의 차이와 같은 모수를 추정하는 데는 유용함
- 로지스틱회귀모형에서 오즈비를 추정하거나 처리효과가 있는지 여부를 알아보고자 할 때는 도움이 되지 않음

04

제 7장. 로지스틱회귀모형(3)

베이지안 추론, 벌점가능도 추정법

1. 베이지안 모형화: 사전분포 명시

- ML 추정값이 무한대일 때 편향성이 적고, 유한한 추정값을 구할 수 있는 수정된 가능도 기반의 방법 중 하나
- 로지스틱 회귀분석에서 베이지안 추론은 보통 $\{\beta_j\}$를 평균이 0이고, 서로 독립인 정규분포의 확률변수로 취급하여 추론함
- 보통 정규분포의 표준편차 값을 아주 크게 하여 사전분포가 결과적으로 추정 결과에 거의 영향을 미치지 않도록

Table 5.5 Results of Bayesian and frequentist fitting of models to the endometrial cancer data-set of Table 5.4 .

Analysis	$\hat{\beta}_1$ (SD)	Interval[a]	$\hat{\beta}_2$ (SD)	$\hat{\beta}_3$ (SD)
ML	∞ (—)	(1.3, ∞)	−0.42 (0.44)	−1.92 (0.56)
Bayes, $\sigma = 10$	9.12 (5.10)	(2.1, 21.3)	−0.47 (0.45)	−2.14 (0.59)
Bayes, $\sigma = 1$	1.65 (0.69)	(0.3, 3.0)	−0.22 (0.33)	−1.77 (0.43)

[a] Profile-likelihood interval for ML and equal-tail posterior interval for Bayes.

2. 로지스틱 모형에서 벌점가능도(penalized likelihood)

- 매우 희박한 분할표와 같이 잠재적으로 추정이 불안정한 상황에서 ML 방법을 수정하여 합리적인 추정을 하는 방법
- 모형의 모수 β에 대한 로그 가능도 함수 $L(\beta)$가 있는 모형에 대해서 다음 식을 최대로 하는 추정값을 찾는 방법.

$$L^*(\beta) = L(\beta) - s(\beta)$$

여기서, $s(\cdot)$는 β가 더 평활해(smoother)질수록 $s(\beta)$가 감소하는 함수

" 이 평활화 방법은 ML 추정값을 0으로 축소시킴.

ML 추정값이 무한대 값을 갖거나 다중공선성이 있는 경우

좀더 안정적인 추정결과를 제공함"

2. 로지스틱 모형에서 벌점가능도(penalized likelihood)

- 자궁내막암의 위험인자 분석 사례

```
> fit <- glm(HG ~ NV + PI + EH, family=binomial, data=Endo)
> summary(fit)
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)    4.305       1.637    2.629    0.0086
NV            18.186    1715.751    0.011    0.9915 # true estimate = infinity
PI            -0.042       0.044   -0.952    0.3413
EH            -2.903       0.846   -3.433    0.0006
---
     Null deviance: 104.903  on 78  degrees of freedom
Residual deviance:  55.393  on 75  degrees of freedom
```

Firth's 방법 :
Jeffery's prior를 s(.)로 이용하는 방법

```
--------------------------------------------------------------------------------
> library(logistf) # can implement Firth's penalized likelihood method
> fit.penalized <- logistf(HG ~ NV2 + PI2 + EH2, family=binomial, data=Endo)
> summary(fit.penalized)
Confidence intervals and p-values by Profile Likelihood
                coef  se(coef)  lower 0.95 upper 0.95   Chisq          p
(Intercept)  0.3080    0.8006     -0.9755     2.7888   0.169  6.810e-01
NV2          2.9293    1.5508      0.6097     7.8546   6.798  9.124e-03
PI2         -0.3474    0.3957     -1.2443     0.4045   0.747  3.875e-01
EH2         -1.7243    0.5138     -2.8903    -0.8162  17.759  2.507e-05
--------------------------------------------------------------------------------
```

09 강

다음시간안내

로지스틱회귀모형 (4)

수고하셨습니다.