05 강

데이터분석방법론2

로지스틱회귀모형

통계·데이터과학과 이기재 교수

학습목차

학습개요 및 목표

이번 강의는 일반화선형모형을 적용할 때
통계적 추론과 모형진단 방법에 대해서 공부합니다.
또한 이항자료 분석에서 가장 널리 사용되고 있는 로지스틱회귀모형을
일반화선형모형 관점에서 살펴봅니다.

1. GLM을 적용할 때 통계적 추론과 모형진단 방법을 설명할 수 있다.
2. 로지스틱회귀모형의 통계적 추론 방법을 설명할 수 있다.
3. 로지스틱회귀모형을 적용하여 분석하고 해석할 수 있다.

제 5장. 로지스틱회귀모형

1 통계적 추론과 모형진단(일반화 선형모형)

2 로지스틱회귀모형의 해석

3 로지스틱회귀모형에 대한 추론

01

제 5장. 로지스틱회귀모형

통계적 추론과 모형진단

1. GLM 모수에 대한 추정

- 범주형 자료에 대해서 적용하는 대부분의 GLM에서 모수추정은 ML방법을 통해서 계산됨

- 모수 β에 대한 95% 신뢰구간(Wald 방법)

$$\hat{\beta} \pm Z_{\alpha/2}(SE)$$

- 가설검정

$$H_0 : \beta = 0$$

① Wald 검정

- $Z = \frac{\hat{\beta}}{SE}$ 는 근사적으로 $N(0,1)$을 따름

  또는 $Z^2 = \left(\frac{\hat{\beta}}{SE}\right)^2$은 근사적으로 χ_1^2 을 따름

- SE는 $\hat{\beta}$ 에 대한 표준오차(β의 제약이 없는 상황에서 구함)

1. GLM 모수에 대한 추정

▪ 가설검정

② 가능도비를 이용하는 방법

- l_0 = 귀무가설 하에서 가능도 함수의 최대값($\beta = 0$ 일 때)
- l_1 = 완전모형 하에서 가능도 함수의 최대값(β 에 대한 제약 없음)
- 가능도비 (likelihood-ratio) 검정통계량

$$-2\log(l_0/l_1) = -2[\log(l_0) - \log(l_1)] = -2(L_0 - L_1)$$

(단, L_0 와 L_1 은 각각 귀무가설과
완전모형 하에서의 로그 가능도함수의 최대값)

1. GLM 모수에 대한 추정

▪ 가설검정

③ 스코어 검정

- 귀무가설 $\beta = 0$ 에서 가능도 함수 $L(\beta)$ 에 접하는 선을 그릴 때 그 선의 기울기를 사용함
- 스코어 통계량은 이 기울기 값을 귀무가설의 β 값을 이용하여 계산한 SE로 나눈 비로 정의함
- 스코어 통계량의 제곱은 근사적으로 df=1인 카이제곱분포를 따름

1. GLM 모수에 대한 추정

▪ 가설검정

④ 검정방법 비교

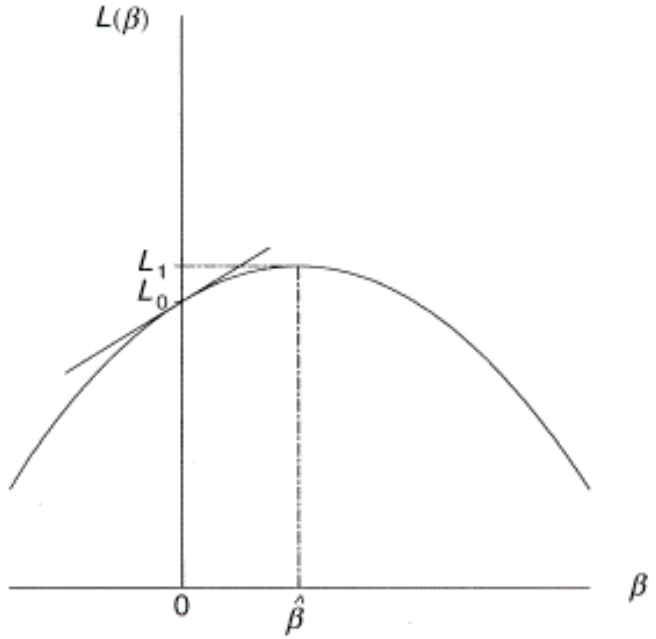


Figure 3.5 Information from the log-likelihood function $L(\beta)$ used in GLM tests of H_0: $\beta = 0$. The Wald test uses $\hat{\beta}$ and the curvature of $L(\beta)$ at $\hat{\beta}$. The likelihood-ratio test uses twice the difference $(L_1 - L_0)$ at $\beta = \hat{\beta}$ and at $\beta = 0$. The score test uses the slope of the line drawn tangent to $L(\beta)$ at $\beta = 0$.

$$L(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i)$$

Score function은 log likelihood의 그레디언트입니다.
($s(\theta)$는 열벡터입니다.)

$$s(\theta) = \nabla_\theta \sum_{i=1}^{n} \log p_\theta(x_i)$$
$$= \sum_{i=1}^{n} \nabla_\theta \log p_\theta(x_i)$$

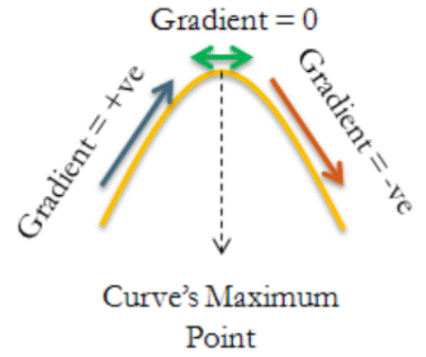


Figure: The gradient of log likelihood function is called ***score***

Score function이 $\log p_\theta(x_i)$의 합이므로, Central Limit Theorem에 따라 score function은 asymtotic하게 정규분포를 따릅니다.

$$s(\theta) \sim \mathcal{N}(0, \Sigma)$$

The information matrix (also called Fisher information matrix) is the matrix of second cross-moments of the score vector.

1. GLM 모수에 대한 추정

Note 1

왈드 방법 신뢰구간 작성

- 귀무가설 $H_0 : \beta = \beta_0$ 에 대한 왈드 통계량: $z = (\hat{\beta} - \beta_0)/SE$
- 95% 신뢰구간 : $\hat{\beta} \pm z_{\alpha/2}(SE)$

1. GLM 모수에 대한 추정

Note 2

가능도비 방법에 의한 β 에 대한 신뢰구간 작성
(프로파일 가능도 신뢰구간)

- 95% 신뢰구간
  = "$H_0 : \beta = \beta_0$ 하의 가능도 검정에서
  P-값이 0.05를 초과하는 모든 β_0 값들로 이루어짐"
- 표본크기 n 이 작거나 설병변수의 효과가 매우 큰 경우는 왈드 통계량보다 가능도비 통계량이 더 검정력이 높고 신뢰할 수 있음

2. 예제 : 암컷 게와 부수체 연구

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	171	55.4531	0.3243
Scaled Deviance	171	55.4531	0.3243
Pearson Chi-Square	171	46.9428	0.2745
Scaled Pearson X2	171	46.9428	0.2745
Log Likelihood		-38.1359	
Full Log Likelihood		-265.3360	
AIC (smaller is better)		534.6719	
AICC (smaller is better)		534.7425	
BIC (smaller is better)		540.9785	

Algorithm converged.

```
PROC GENMOD DATA=crab;
Model satell=width / dist=poi link=log type1;
RUN;
```

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.4178	0.6070	0.2281	2.6075	5.46	0.0195
Width	1	-0.0193	0.0231	-0.0646	0.0259	0.70	0.4026
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	56.1568			
Width	55.4531	1	0.70	0.4015

2. 예제 : 암컷 게와 부수체 연구

Note

통계소프트웨어에 따라서는
가설 $H_0 : \beta = 0$ 에 대한 유의성 검정방법으로
스코어 통계량(Score statistics)을 제공하기도 함

- 왈드 검정과 가능도비 검정은
  표본크기가 큰 경우에는 유사한 성질을 가짐
- 표본의 크기가 대단히 크지 않은 경우는
  가능도비 검정을 사용하는 것이
  바람직한 것으로 알려짐

3. 이탈도 (Deviance)

- 포화모형(Saturated Model)
  : 각 관측값에 대하여 각각 모수를 갖는 경우로
  $\hat{\mu}_i = y_i$ 을 만족하는 모형을 말함

- 포화모형은 가능한 모형 中
  가장 복잡한(가장 모수가 많은) 형태의 모형으로
  가능도함수(또는 로그 가능도 함수)의 최대값을 가짐

3. 이탈도 (Deviance)

예제

- 코골기와 심장병 자료(4x2 분할표)
  - $M : \pi(x) = \alpha + \beta x$
  - M_s : 코골기의 4수준의 이항관측값에 대하여 각각 다른 모수 사용
    $(\pi(x) = \pi_1,\ \pi(x) = \pi_2,\ \pi(x) = \pi_3,\ \pi(x) = \pi_4)$
  - L_M : 모형 M 에 대해서 구한 로그가능도 함수의 최대값
  - $M : \pi(x) = \alpha + \beta x$
  - L_s : 모형 M_s(포화모형)에 대해서 구한 로그 가능도함수의 최대값
    ➔ 다음가설을 위한 검정 통계량

▪ H_0 : 모형 M을 따름 ▪ H_1 : 포화모형 S를 따름	H_0 : 포화모형에 있는 모수 中 모형 M에 포함되지 않는 모수들은 모두 0임

이탈도 (Deviance) = $-2[L_M - L_s]$

3. 이탈도 (Deviance)

- 이탈도(Deviance)는
  df = (자료 수 - 모형 모수의 개수)인 χ^2 분포로 근사

- 검정통계량 값이 크고, P-값이 작을수록
  모형 M 의 적합결여에 대한 강한 증거가 됨

3. 이탈도 (Deviance)

▪ 코골기 사례

① $\pi(x) = \alpha + \beta x$ 로 적합하는 경우

이탈도(Deviance) = 0.1, df = 4 - 2 = 2, P-값 = 0.97

② $logit[\pi(x)] = \alpha + \beta x$

이탈도(Deviance) = 2.8, df = 2, P-값 = 0.25

→ 두 모형은 자료에 잘 적합됨

4. 관측치와 모형 적합값을 비교하는 잔차

- GLM에 대한 적합결여 통계량
  : 모형이 자료를 잘 적합하는지를 포괄적으로 요약하는 측도임

- 관측도수와 적합값을 비교해 보면 더 자세한 정보를 알 수 있음

잔차 = $y_i - \hat{\mu}_i$ ➡ 대개 μ_i가 커질수록 잔차도 커짐

피어슨 잔차 = $e_i = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\widehat{var}(y_i)}}$

- 예를 들어 포아송 GLM인 경우 $e_i = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$
  - 참고 : 피어슨 통계량

$$\sum_i e_i^2 = \sum_i (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i = X^2$$

4. 관측치와 모형 적합값을 비교하는 잔차

- 표준화잔차 (Standardized Residual)

$$\text{표준화잔차} = \frac{y_i - \hat{\mu}_i}{SE} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{var}(y_i)(1-h_i)}}$$

  - h_i는 i번째 관측값의 leverage라고 함
  - 대체적으로 leverage가 크면 모형적합에 미치는 영향이 커짐

- 표준화잔차가 2 또는 3 정도로 크면 주의 깊게 살펴봐야 함

02

제 5장. 로지스틱회귀모형

로지스틱 회귀모형의 해석

1. 개요

- $Y = 0 \text{ or } 1$

  $\pi = P(Y = 1)$

  $$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x$$

  $$logit[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

- 이항분포를 따르는 Y 에 대해서 "$logit$" link를 사용하는 경우

  $$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x \Leftrightarrow \pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

2. 선형근사 해석

β의 부호해석

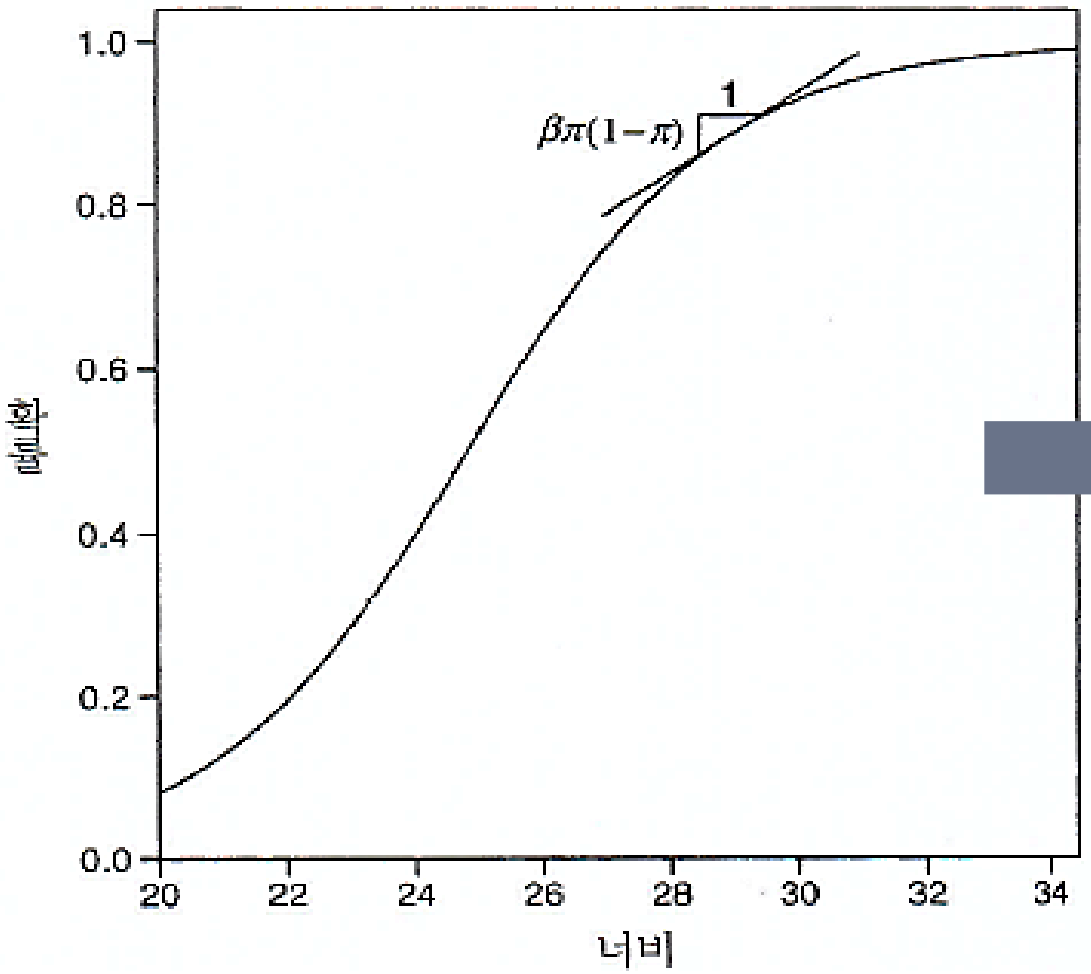
① $\beta > 0 \Leftrightarrow \pi(x) \uparrow as\ x \uparrow$

② $\beta < 0 \Leftrightarrow \pi(x) \downarrow as\ x \uparrow$

③ $\beta = 0 \Leftrightarrow \pi(x) = \frac{e^{\alpha}}{1+e^{\alpha}}$ 상수 $as\ x \uparrow$

2. 선형근사 해석

로지스틱회귀 곡선에 대한 선형근사



- $\pi(x) = \frac{1}{2}$ 인 x 값에서의 접선 기울기가 가장 큼
- 그 때의 x 값은 $x = -\frac{\alpha}{\beta}$ 임

3. 참게 예제

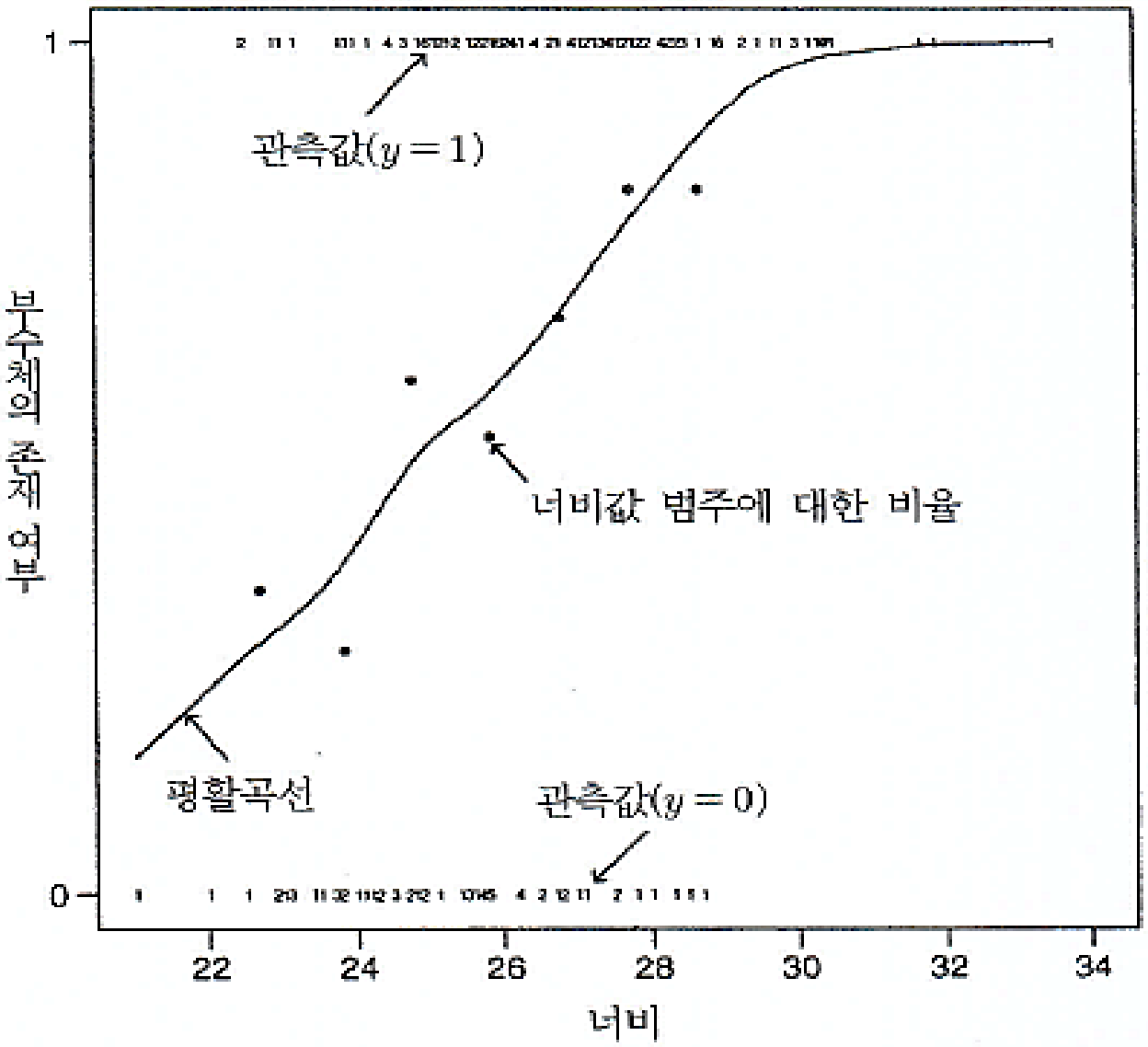
$$Y=\begin{cases}1\text{ , 한 마리 이상의 부수체를 보유한 경우}\\ 0\text{ , 부수체가 없는 경우}\end{cases}$$

X : 암 참게의 등딱지 너비

- 일반화가법모형(Generalized Additive Models)에 기초한 평활방법

너비와 부수체의 비율 간의 관계에 대해
특정함수의 형태를 가정하지 않고 자료를 평활하여
일반적인 추세를 조사하는 방법

3. 참게 예제


1
0
부수체의 존재 여부
관측값($y=1$)
너비값 범주에 대한 비율
평활곡선
관측값($y=0$)
22
24
26
28
30
32
34
너비

3. 참게 예제

❑ 로지스틱회귀모형 적용

- $\pi(x)$: 너비가 x 인 암 참게가 부수체를 가질 확률
- 선형확률모형 $\pi(x) = \alpha + \beta x$ 을 적용하는 경우
- $\hat{\pi}(x) = -1.776 + 0.092x$

- $x = 33.5$ (최대 너비)에서 예측확률값
  : $-1.766 + 0.092(33.5) = 1.3$
  ➔ “예측 확률값이 1보다 크게 나타남”

3. 참게 예제

로지스틱회귀모형 적용

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = -12.351 + 0.497x$$

$$\Leftrightarrow \quad \hat{\pi}(x) = \frac{\exp(-12.351 + 0.497x)}{1+\exp(-12.351 + 0.497x)}$$

$\hat{\beta} > 0 \Leftrightarrow$ 폭이 커질수록 예측확률 $\hat{\pi}$ 가 커짐

$$\hat{\pi}(21.0) = 0.129, \quad \hat{\pi}(33.5) = 0.987$$

3. 참게 예제

PROC LOGISTIC 적합

```
DATA crab;
INPUT  color   spine   width   satell   weight;
IF satell>0 THEN y=1;
IF satell=0 THEN y=0;
CARDS;
2  3  28.3  8  3.05
3  3  22.5  0  1.55
1  1  26.0  9  2.30

.....중간 생략.....
1  1  28.0  0  2.63
4  3  27.0  0  2.63
2  2  24.5  0  2.00
;
PROC LOGISTIC descending;
   MODEL y= width/covb lackfit;
```

PROC GENMOD 이용

```
PROC GENMOD;
   MODEL  y= width/dist=bin link=logit;
```

4. 오즈비 해석

- $\log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x \Leftrightarrow \frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta x}$
- $\frac{\pi(x+1)}{1-\pi(x+1)} = e^{\alpha+\beta(x+1)} = e^{\beta}e^{\alpha+\beta x} = e^{\beta}\frac{\pi(x)}{1-\pi(x)}$

즉, $x+1$에서의 오즈는 x에서의 오즈의 e^{β}를 곱한 것과 같다.

4. 오즈비 해석

- $\beta = 0 \Leftrightarrow e^{\beta} = 1$, x가 변하더라도 오즈는 변하지 않음
- $\mathrm{logit}[\pi(x)] = -12.35 + 0.497x$

부수체에 대한 오즈는 너비가 1cm 증가함에 따라
$e^{\hat{\beta}} = e^{0.497} = 1.64$배 증가함

5. 후향적 연구에서의 로지스틱 회귀모형

- 반응변수 Y가 랜덤이 아니고
  설명변수 X가 랜덤인 경우에도
  로지스틱회귀모형을 적용할 수 있음

  → 주로 후향적 표본추출 설계에서 얻어짐

- 사례-대조 연구에서
  $Y = 1$(“사례”)과 $Y = 0$(“대조”)인
  개체로 구성된 표본으로부터
  X값이 관측되는 예제에 적용 가능함

5. 후향적 연구에서의 로지스틱 회귀모형

- 사례 - 대조 연구에서 로지스틱회귀모형을 적용하여 관심 있는 설명변수의 효과를 추정할 수 있음

- 모형에서 절편을 나타내는 α는 $Y=1$일 때와 $Y=0$일 때의 도수에 대한 상대적인 값이므로 의미를 갖지 못함

03

제 5장. 로지스틱회귀모형

로지스틱 회귀모형에 대한 추론

1. 이항자료의 그룹화와 비그룹화

그룹화된 이항자료

- 예 : 코골기와 심장병 예제

  - 코골기와 심장병 예제에서 254명이 매일 코를 골며

    그 중 30명이 심장병이 있음

    ➔ 표본크기 254명 중 30명이 심장병이 있다고 표시 가능

비그룹화된 이항자료

- 예 : 암게의 부수체 자료

  - 설명변수가 연속형일 때의 이항자료

  - 그룹화 이항자료는 비그룹화 자료로 표시 가능

2. 효과에 대한 신뢰구간

- $\mathrm{logit}[\pi(x)] = \alpha + \beta x$에서 모수 β에 대한 신뢰구간
  $\hat{\beta} \pm z_{\alpha/2}(SE)$ (wald 방법)
- e^{β}에 대한 신뢰구간
  $(e^{\hat{\beta} - z_{\alpha/2}(SE)}, e^{\hat{\beta} + z_{\alpha/2}(SE)})$

설명변수 x에서 한 단위 증가할 때
오즈에 미치는 효과에 대한 신뢰구간

2. 효과에 대한 신뢰구간

참게 예제

- $\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = -12.351 + 0.497x$
- $\hat{\beta} = 0.497\,,\ SE = 0.102$

[β에 대한 95% 신뢰구간]

$$0.49 \pm 1.96(0.102) \Leftrightarrow (0.298, 0.697)$$

2. 효과에 대한 신뢰구간

- e^{β}에 대한 95% 신뢰구간 : $(e^{0.308}, e^{0.709}) = (1.36, 2.03)$

암 참게가 부수체를 소유할 오즈는
너비가 1cm 증가함에 따라
적어도 36%에서 두 배까지 증가함

Note

- 표본크기 n이 작은 경우는 Wald 신뢰구간 보다는 가능도비 (likelihood ratio) 신뢰구간을 사용하는 것이 바람직함
- SAS GENMOD 절차에서 LRCI 옵션을 이용하여 구할 수 있음

3. 유의성 검정

- $H_0 : \beta = 0 \Leftrightarrow$ “Y와 X는 서로 독립”
- 검정통계량 $Z = \frac{\hat{\beta}}{SE}$

  : 귀무가설 하에서 표준정규분포를 따름

3. 유의성 검정

가능도비 검정 (귀무가설 $H_0 : \beta = 0$)

- L_0 : $\beta = 0$일 때 ($\pi(x)$가 모든 x 값에 대해서 동일)의 로그가능도 함수의 최대값
- L_1 : β에 대한 아무런 제한 조건이 없는 경우에 로그가능도 함수의 최대값

검정통계량 : $-2(L_0 - L_1) \sim \chi^2(1)$
(SAS GENMOD에서 TYPE3 옵션을 통해서 구할 수 있음)

3. 유의성 검정

Note

로지스틱회귀모형이 실제로 만족한다면,
모형으로부터 구한 확률의 추정량 $\hat{\pi}(x)$는
표본비율보다 더 좋은 추정량이 됨

- 모형을 이용한 추정방법은 주어진 x 값에 해당하는 자료만을 이용하는 것이 아님
- 모든 자료를 사용하여 더 정확한 추정결과를 제공함

06 강

다음시간안내

로지스틱회귀모형 (2)

수고하셨습니다.