04 강

데이터분석방법론2

일반화 선형모형

통계·데이터과학과 이기재 교수

학습목차

학습개요 및 목표

분할표 분석은 반응변수와 설명변수가 범주형일 때 적용했던 연관성 분석법입니다. 이번 강의는 통계모형을 이용한 범주형 자료 분석에 대해서 학습합니다. 통계모형을 이용해서 분석하면 단순한 유의성 검정이 아닌 모수 추정을 통해서 더 많은 정보를 얻을 수 있습니다.

1. 범주형 자료분석에서 통계모형 적용의 필요성에 대해 설명할 수 있다.
2. 일반화선형모형 GLM의 구성요소를 설명할 수 있다.
3. 이항자료와 도수자료에 대한 일반화 선형모형을 설명할 수 있다.

제 4장. 일반화 선형모형

1 GLM의 성분

2 이항자료에 대한 일반화 선형모형

3 도수 자료에 대한 일반화 선형모형
: 포아송 회귀

01

제 4장. 일반화 선형모형

GLM의 성분

1. GLM의 성분 구성요소

GLM (Generalized Linear Model)

- 여러 설명변수들의 효과를 동시에 분석해야 하는 복잡한 상황에 유용함

모형을 이용한 분석에는 모수 추정에 역점을 두며,
단순한 유의성 검정보다 더 많은 정보 제공

- 범주형 및 연속형 반응변수에 대해서
잘 알려진 통계모형들을 포괄하는 일반화된
형태의 통계모형

- 회귀모형, ANOVA 모형, 로지스틱회귀모형 등은
GLM의 특별한 경우임

2. 개요

- GLM
  - 전통적인 회귀모형을 반응변수가
    정규분포를 따르지 않는 경우로 확장한 것임
- GLM을 적용하기 위해서는 반응변수가
  지수족 분포(Exponential Family Of Distributions)를
  따라야만 함
- 모든 GLM의 세 가지 공통 요소
  - 랜덤성분 (Random Component)
  - 체계적 성분 (Systematic Component)
  - 연결함수 (Link Function)

3. GLM의 구성요소

1. 랜덤성분 (Random Component)

- 반응변수 Y를 결정
- $Y_1, Y_2, \cdots, Y_n$을 정규분포, 포아송분포, 이항분포 등에서 추출된 랜덤 표본으로 가정
- $\mu_i = E(Y_i)$가 설명변수들에 의해서 어떻게 영향을 받는지 모형 설정
- Exponential Family 분포

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)]$$

- 예: Poisson 분포

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} = \exp(-\mu)\left(\frac{1}{y!}\right)\exp[y(\log\mu)],\ \ y = 0,1,2,\ \cdots$$

3. GLM의 구성요소

2. 체계적 성분 (Systematic Component)

- 설명변수 $\{x_j\}$의 선형식인 선형예측식(Linear Predictor)으로 구성

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

3. 연결함수

- Y에 대한 기댓값 $\mu = E(Y)$는 설명변수들의 값에 따라 달라짐
- 랜덤성분과 체계적 성분(선형예측식)을 연결하는 함수 $g(\cdot)$를 말함

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

3. GLM의 구성요소

연결함수의 예

- $g(\mu)=\mu$: 항등연결 (Identity Link)

$$\mu = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

반응변수가 정규분포를 따를 때
회귀분석(Ordinary Regression)에 해당함

- $g(\mu)=\log(\mu)$: 로그연결

"평균의 로그를 모형화하는 것으로
빈도와 같이 기대값이 음이 아닌 자료에 적합"

$$\log(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

로그연결을 사용하는 GLM을
로그선형모형(Loglinear Model)이라고 함

3. GLM의 구성요소

연결함수의 예

- $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$: 로짓연결 (Logit Link)

$$\log\left(\frac{\mu}{1-\mu}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

- 이항분포($\mu = \pi$ 이고, $0 < \pi < 1$)의 경우에 주로 사용
- logit = log of odds

3. GLM의 구성요소

연결함수의 예

표준연결함수 (Canonical Link)

- Exponential Family 분포

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)]$$

- Canonical Link : $g(\mu_i) = Q(\theta_i)$

예: $f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} = \exp(-\mu)\left(\frac{1}{y!}\right)\exp[y(\log\mu)],\ \ y = 0,1,2,\ \cdots$

$$g(\mu) = \log(\mu) \rightarrow \log(\mu_i) = \sum_j \beta_j x_{ij},\ \ i = 1,\cdots,n$$

3. GLM의 구성요소

Note 1

GLM은 보통의 회귀분석을
다음 두 가지 관점에서 일반화한 것으로 볼 수 있음

① Y에 대하여 정규분포 이외에 다른 확률분포를 허용함

② 단순한 μ에 대한 모형화가 아닌 μ의 함수인 $g(\mu)$에 대한 모형화가 가능함

3. GLM의 구성요소

Note 2

- GLM의 모수추정 과정에서는 선택된 랜덤성분에 대하여 ML(최대가능도 추정법) 방법을 사용
- 랜덤성분의 확률분포로 정규분포 이외의 분포도 가능함

- 회귀모형, 분산분석 모형, 범주형 자료에 대한 모형들은 GLM의 특별한 경우
- 동일 ML 추정법을 사용하여 모든 GLM에 대한 모형 적합이 가능함

- SAS PROC GENMOD, R glm함수 등을 통해서 분석

02

제 4장. 일반화 선형모형

이항자료에 대한 일반화 선형모형

알아보기

- 반응변수 Y 의 분포는 성공확률, $P(Y=1)=\pi$ 이고,
  실패확률 $P(Y=0)=1-\pi$ 인 경우

- $E(Y)=\pi, \quad Var(Y)=\pi(1-\pi)$

- 설명변수 x 가 변함에 따라 $\pi=\pi(x)$ 가
  영향을 받는 경우의 모형 적합에 대해서 살펴봄

1. 선형확률모형

통계 모형

- $\pi(x) = \alpha + \beta x$
  : 성공확률이 x에 따라 선형적으로 변함

- 이항확률분포에 대하여 항등연결함수를 갖는 GLM
  $(\mu = E(Y) = \pi)$

1. 선형확률모형

단점

- $Var(Y) = \pi(x)(1-\pi(x))$: x 값에 따라 변화함
  ➔ 최소제곱 추정법은 Optimal 안됨
  ➔ 최대가능도 추정법(ML)을 이용하여 GLM을 적합함

- x의 값이 대단히 크거나 작은 경우에는
  $\pi(x) < 0$ 이나 $\pi(x) > 1$ 인 경우가 발생 가능

2. 예제 : 코 골기와 심장병에 관한 연구

개요

- 코 고는 것이 심장병의 위험요인이 될 수 있는 지를 알아보기 위해 2,484명을 대상으로 조사한 자료

- 배우자들의 보고를 근거로 코 고는 정도에 따라 4범주로 분류

Y = 심장병 발병 여부
(1 = 발병, 0 = 발병하지 않음)

x = 코고는 정도
(코 고는 범주에 대하여 (0, 2, 4, 5) 할당)

2. 예제 : 코 골기와 심장병에 관한 연구

코 골기와 심장병과의 관계

코고는 정도	심장병					
	유	무	비율	선형적합	로짓모형	프로빗적합
전혀 아니다.	24	1355	0.017	0.017	0.021	0.020
가끔	35	603	0.055	0.057	0.044	0.046
거의 매일 밤	21	192	0.099	0.093	0.093	0.095
매일 밤	30	224	0.118	0.116	0.132	0.131

<참고> 그룹화(grouped) 또는 그룹화되지 않은(ungrouped) 이항자료 구분

2. 예제 : 코 골기와 심장병에 관한 연구

ML방법에 의한 적합

- 항등연결함수를 사용하는 경우

➔ $\hat{\pi}(x) = 0.017 + 0.0198x$ (PROC GENMOD 이용)

- 코를 골지 않는 사람($x = 0$)에 대한 심장병 확률 :

➔ $\hat{\pi}(x) = 0.017 + 0.0198(0)$ (PROC GENMOD 이용)

$= 0.017$

- R 예제 : 교재 95쪽 참고

2. 예제 : 코 골기와 심장병에 관한 연구

보통의 최소제곱법을 이용하는 경우

- 2,484개의 이진형태인 0과 1로 입력한 후
  최소제곱법으로 적합
- $\hat{\pi}(x) = 0.0169 + 0.0200x$
- 모형 적합이 잘 된 경우에
  최소제곱 추정값과 ML 추정값은 비슷하게 됨

2. 예제 : 코 골기와 심장병에 관한 연구

GLM 적합을 위한 PROC GENMOD

```
DATA glm;
INPUT snoring disease total;
CARDS;
0   24   1379
2   35    638
4   21    213
5   30    254
RUN;

PROC GENMOD;
  MODEL disease/total=snoring / dist=bin link=identity;
RUN;
```

2. 예제 : 코 골기와 심장병에 관한 연구

프로그램 수행 결과

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0172	0.0034	0.0105	0.024	25.18	<.0001
snoring	1	0.0198	0.0028	0.0143	0.0253	49.97	<.0001
Scale	0	1	0	1	1		

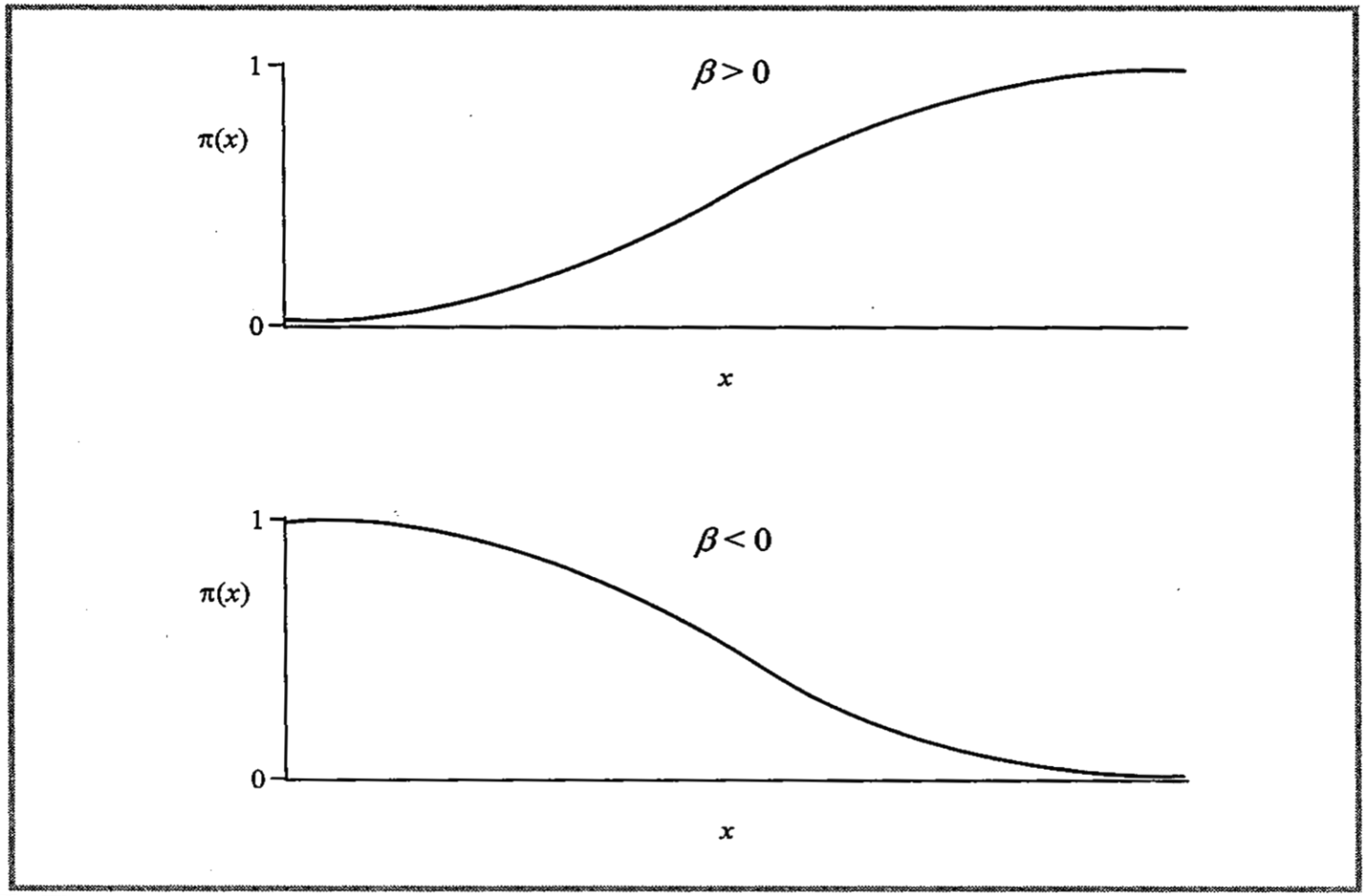
3. 로지스틱 회귀모형

아이디어

- $\pi(x)$와 x간의 관계는 대개 비선형 형태로 볼 수 있음
- x의 일정한 변화량은 π가 구간의 중앙에 있을 때보다 0이나 1에 가까이 있을 때 π에 대한 영향을 덜 미치게 됨
- $$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$
$$\Leftrightarrow \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha+\beta x$$
- $\log(\pi/(1-\pi)) = logit(\pi)$

3. 로지스틱 회귀모형

로지스틱 회귀함수의 형태



3. 로지스틱 회귀모형

로지스틱 회귀모형 적합결과

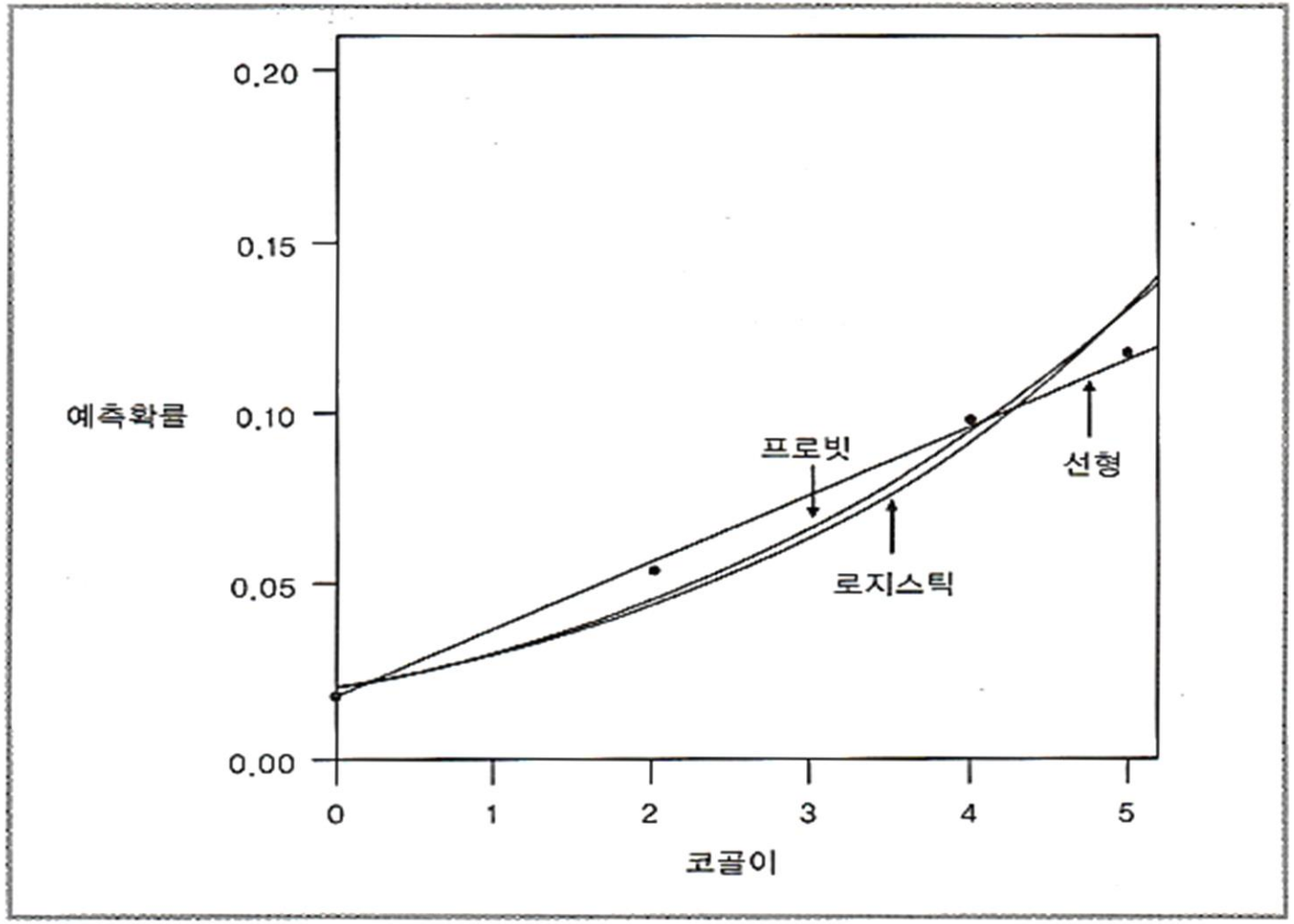
- 코 고는 정도와 심장병과의 연관성 자료에 대한 로지스틱 회귀모형 적합결과

$$\text{logit}\left[\hat{\pi}(x)\right] = \log\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = -3.87 + 0.40x$$

코 고는 정도가 심해질수록 심장병 발병 가능성이 높아짐

3. 로지스틱 회귀모형

모형적합 결과



3. 로지스틱 회귀모형

로지스틱회귀(link = logit 사용)의 적합결과

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.8662	0.1662	-4.1920	-3.5405	541.06	<.0001
snoring	1	0.3973	0.0500	0.2993	0.4954	63.12	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

- R 예제 : 교재 94쪽 프로그램 참고

3. 로지스틱 회귀모형

Note

- 로지스틱 회귀모형에 대해서는 5강에서 자세히 다룸
- 분할표에 대한 검정에서는 H_0 하에서 추정 기대도수를 구하여 X^2, G^2 통계량을 통해 가설 검정
- 주어진 사례의 경우 :
  H_0 : 로지스틱 회귀모형을 따름
  $X^2 = 2.05, \quad G^2 = 1.95, \quad df = 4 - 2 = 2$
  ➔ H_0 를 기각할 만한 뚜렷한 증거는 없음

4. 프로빗 회귀모형

프로빗모형 (Probit Model)

- $probit[\pi(x)] = \Phi^{-1}(\pi(x)),\ \ \Phi(\cdot)$는 표준정규분포의 cdf
- $probit(0.05) = -1.645,\ probit(0.975) = 1.96$
- $probit[\hat{\pi}(x)] = -2.061 + 0.188x$: 모형적합결과

- R 예제 : 교재 95쪽 참고, link=probit

4. 프로빗 회귀모형

코골이 수준 $x = 0$ 인 경우

- $probit[\hat{\pi}(0)] = -2.061$

➔ $\hat{\pi}(0) = \Phi(-2.061) = 0.020$

- $probit[\hat{\pi}(5)] = -2.061 + 0.188(5) = -1.121$

➔ $\hat{\pi}(5) = \Phi(-1.121) = 0.131$

4. 프로빗 회귀모형

Note

- 자료에 대한 프로빗 곡선과 로지스틱회귀곡선은 유사함
- 프로빗모형은 1934년 독성학 연구에서 처음 도입
- 오늘날 로지스틱회귀모형이 프로빗 모형에 비해서 더 많이 활용
  ➔ 로지스틱회귀모형의 모수는 오즈비와 연관됨
- 사례
  : 대조 연구 자료에 적용하여 오즈비를 추정할 수 있음

03

제 4장. 일반화 선형모형

도수자료에 대한 일반화 선형모형: 포아송 회귀

1. 포아송 분포 (Poisson Distribution)

포아송 분포

- 도수와 같이 음이 아닌 임의의 정수값을 취할 때 보통 포아송 분포를 가정하여 분석함

- $P(y) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, \ldots$
- $E(Y) = Var(Y) = \mu, \quad \sigma(Y) = \sqrt{\mu}$

1. 포아송 분포 (Poisson Distribution)

포아송 분포의 성질

- 도수가 커질수록 분산도 커짐
- 평균이 증가할수록 치우친 정도는 감소하여 점차 좌우대칭인 종모양을 나타냄
- 실제 데이터 분석에서 종종 $\sigma^2 > \mu$인 경우를 볼 수 있는 데 이를 과대산포(Overdispersion)라고 함

2. 포아송 회귀모형

◪ 포아송 회귀모형이란?

- Y는 포아송 분포를 따르고, x를 설명변수로 가정함

- 모형

① 항등연결함수 : $\mu = \alpha + \beta x$

② 로그연결함수 : $\log(\mu) = \alpha + \beta x$

➔ 로그선형 모형(Loglinear Model)이라고 함 (Part 7의 내용)

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}$$

3. 예제 : 암 참게와 부수체에 관한 연구

연구 개요

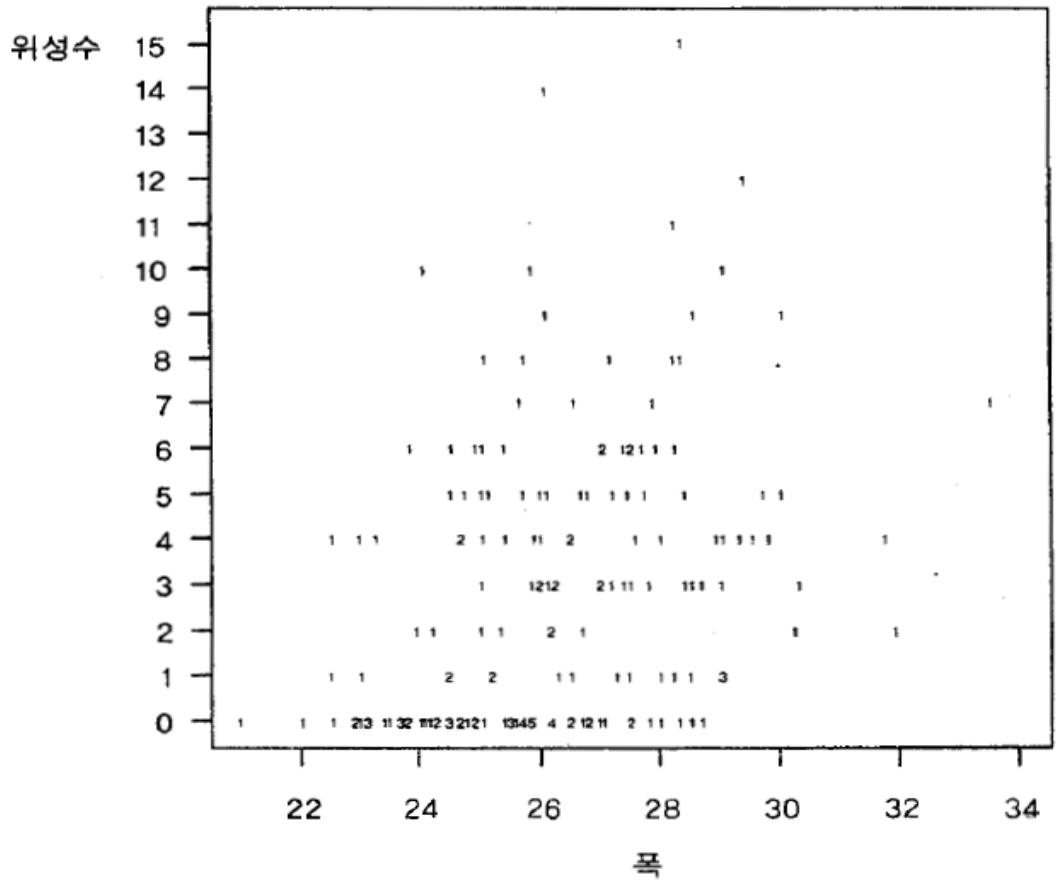
- 각 암 참게는 집을 갖고 있으며
  이 집에 붙어사는 숫 참게를 가지고 있음
- 이 숫 참게를 부수체(Satellite)라고 부름

반응변수 : 암 참게의 부수체 수
설명변수 : 암 참게의 등딱지 너비
(암 참게의 크기)

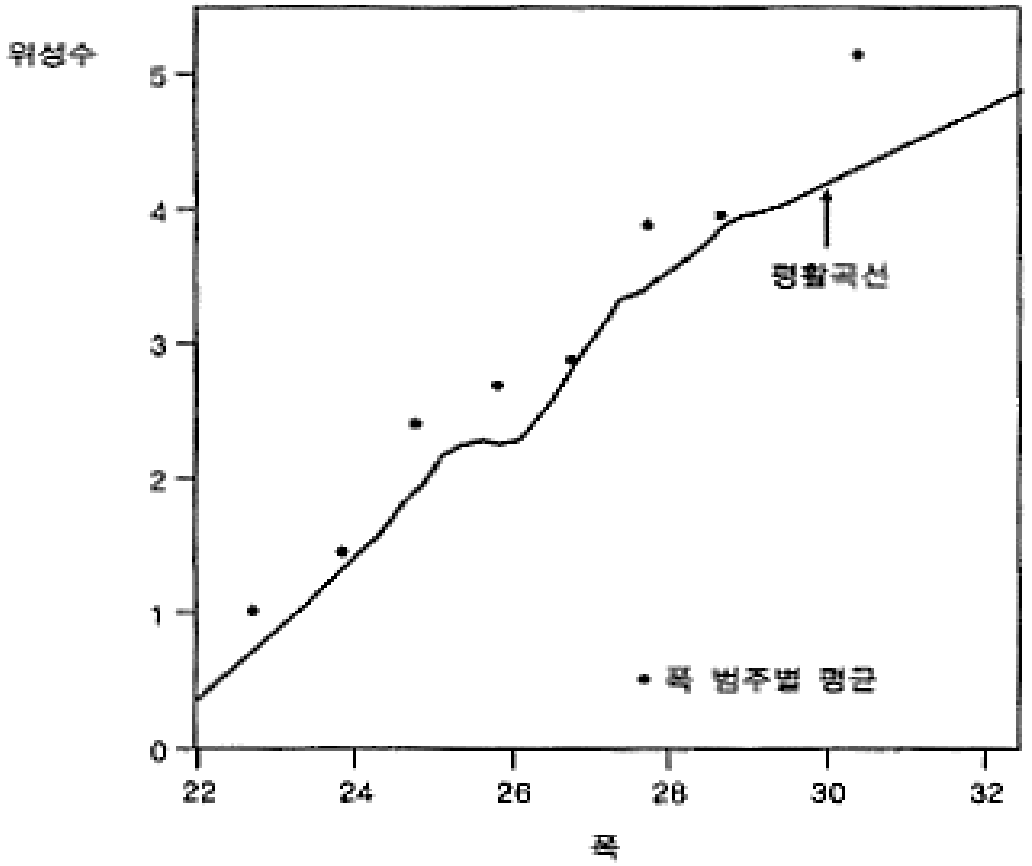
3. 예제 : 암 참게와 부수체에 관한 연구

암 참게의 등딱지 너비와 부수체 수 관계

[암컷 게의 등딱지 폭에 따른 부수체 수]



[암컷 게의 부수체 수에 대한 평활]



3. 예제 : 암 참게와 부수체에 관한 연구

포아송 로그 선형 적합

- $\log(\mu) = \alpha + \beta x$ 을 ML방법으로 적합
  (SAS PROC GENMOD 이용)

- R 예제 : 교재 99쪽 참고

- $\log(\hat{\mu}) = \hat{\alpha} + \hat{\beta} x = -3.305 + 0.164x$
  ($\hat{\beta}$ 의 ASE = 0.020)

➔ $\hat{\beta} > 0$ 이므로 등딱지 너비가 증가함에 따라 부수체수도 증가함

➔ $\hat{\mu} = \exp(-3.305 + 0.164x)$

평균적으로 너비가 대략 2cm 증가하면 부수체 수가 하나씩 증가함

3. 예제 : 암 참게와 부수체에 관한 연구

◪ 포아송 회귀모형 : 항등연결

- $\mu = \alpha + \beta x$ 을 이용

➔ $\hat{\mu} = \hat{\alpha} + \hat{\beta} x = -11.5 + 0.550x$

($\hat{\beta}$ 의 ASE=0.550)

평균적으로 너비가 대략 2cm 증가하면
부수체 수가 하나씩 증가함

3. 예제 : 암 참게와 부수체에 관한 연구

SAS 프로그램

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

        중간생략

1  1  28.0  0  2.63
4  3  27.0  0  2.63
2  2  24.5  0  2.00
;
PROC GENMOD;
    MODEL satell = width/dist=poi  link=log type1;
```

4. 과대산포 : 예측된 것보다 큰 분산

과대 산포

- 포아송 분포를 따를 경우 평균과 분산은 같음

$$E(Y) = Var(Y) = \mu$$

- 모형의 랜덤성분에 의해 예측되는 분산보다 더 큰 분산을 갖는 현상을 과대산포(Overdispersion)라고 함

4. 과대산포 : 예측된 것보다 큰 분산

부수체 수의 표본평균과 표본 분산

너비	경우 수	부수체 수	표본평균	표본분산
< 23.25	14	14	1.00	2.77
23.25 - 24.25	14	20	1.43	8.88
24.25 - 25.25	28	67	2.39	6.54
25.25 - 26.25	39	105	2.69	11.38
26.25 - 27.25	22	63	2.86	6.88
27.25 - 28.25	24	93	3.87	8.81
28.25 - 29.25	18	71	3.94	16.88
> 29.25	14	72	5.14	8.29

4. 과대산포 : 예측된 것보다 큰 분산

과대 산포 결과

- 개체들간의 이질성으로 발생할 수 있음
- 암컷의 부수체 수에 너비, 무게, 색깔, 등뼈의 상태 등이 모두 영향을 줄 때 너비만 고려한 모형을 적용하면 분산이 커질 수 있음
- 과대산포는 도수에 포아송 회귀모형을 적용할 때 흔히 나타남
  ➔ 이는 포아송 분포에서 분산이 평균과 같기 때문임

05 강

다음시간안내

로지스틱회귀모형

수고하셨습니다.