11강

일반화선형모형 (1)

서강대학교 경영학과 이윤동교수

목 차

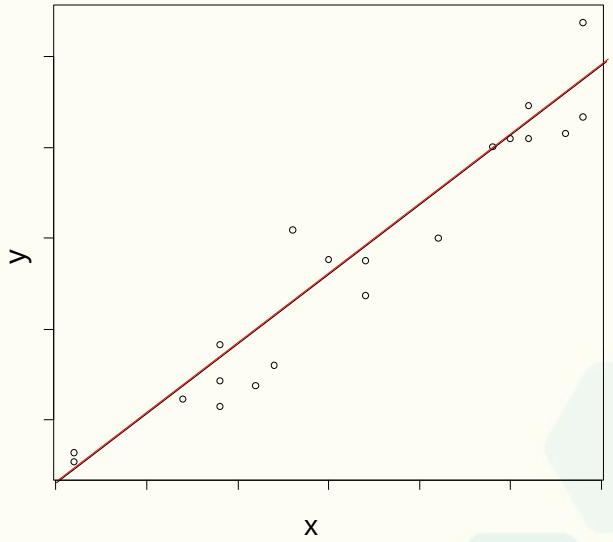
1 일반화선형모형 소개

회귀분석

종속변수 y 의 값들이 독립변수들인 x 의 값들에 따라
어떤 영향을 받아 변하는 지를 알아보는 통계적 방법.

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

$$\epsilon_i \overset{iid}{\sim} \text{Normal}\left(0, \sigma^2\right)$$



선형모형

종속변수 y : 양적변수

회귀분석 : 독립변수 x 가 양적변수인 경우(예 : 몸무게, 광고비)

선형모형 : 독립변수 x 가 질적변수인 경우도 포함 (예 : 성별, 붓꽃의 종류)

y_i : 꽃받침의 길이 x_i : 꽃받침의 폭

z_i : 꽃의 종류

$$y_i = z_i + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

일반화선형모형(GLM)

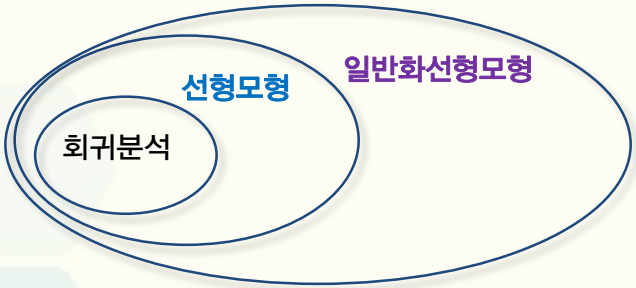
선형모형(Linear Model) : 자료의 분포가 정규분포임을 가정

$$y_i \overset{indep.}{\sim} \text{Normal}\left(z_i + \beta x_i, \sigma^2\right), \quad i = 1, \ldots, n$$

일반화선형모형(Generalized Linear Model) :

자료의 분포가 정규분포인 경우뿐만 아니라, 그 이외의

지수분포족 분포들인 경우들 까지를 대상으로 확장된 통계 모형.



일반화선형모형의 예

AIDS data : Whyte, et.al. 1987 (Dobson, 1990).

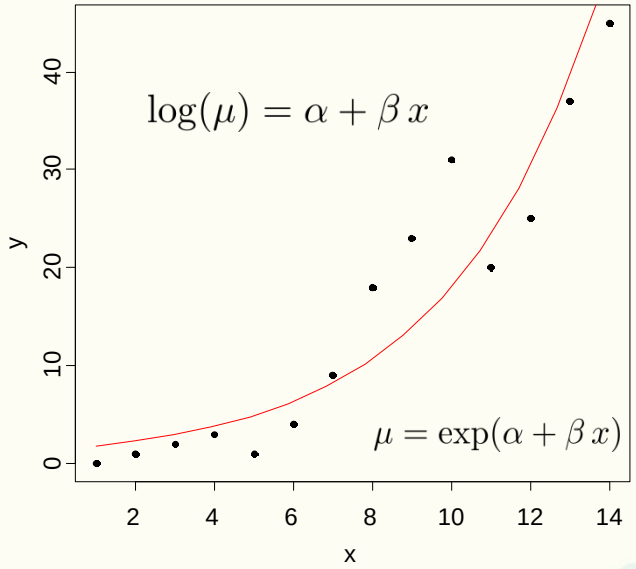
1983~1986년 동안 Australia에서 AIDS로 인한 사망자 수

X : 1983년1월 부터 시작한, 3개월 단위 경과기간

Y : 사망자 수

X	Y
1	0
2	1
3	2
4	3
5	1
6	4
7	9

X	Y
8	18
9	23
10	31
11	20
12	25
13	37
14	45



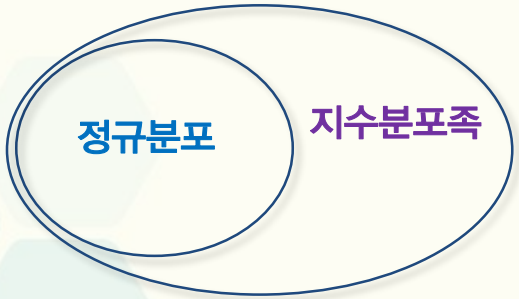
지수분포족

지수분포족 (the exponential family of distributions) :
확률밀도 함수 $f(y;\theta,\varphi)$ 가 다음과 같이 표현되는 분포들.

$$f(y;\theta,\varphi) = \exp\{(y\,\theta - \gamma(\theta))/\varphi + \tau(y,\varphi)\}$$

θ : 정준모수 φ : 산포모수

정규분포, 이항분포, 포아송분포, 감마분포, 역감마분포 등.



지수분포족의 예

정규분포 : 평균이 μ이고 분산이 σ^2인 경우

$$\log f(y;\theta,\varphi) = (y\,\mu - \mu^2/2)/\sigma^2 - (1/2)y^2/\sigma^2 - (1/2)\log(2\pi\sigma^2)$$

$$\theta = \mu \qquad \varphi = \sigma^2$$

포아송분포 : 평균이 λ 인 경우 $\mu = \lambda$

$$\log f(y;\theta,\varphi) = y\,\log\lambda - \lambda - \log(y!)$$

$$\theta = \log\mu \qquad \varphi = 1$$

$\mu = p$

이항분포 : 시행횟수가 n, 성공률이 p, 성공횟수 x, 표본비율 $y = x/n$

$$\log f(y;\theta,\varphi) = n\,\{y\,\log(p/(1-p)) + \log(1-p)\} + (1/2)\log\binom{n}{ny}$$

$$\theta = \log(\mu/(1-\mu)) \qquad \varphi = 1/n$$

선형모형과 일반화선형모형

선형모형 : 정규분포를 가정함

$$y_i \overset{indep.}{\sim} \text{Normal}\left(z_i + \beta x_i, \sigma^2\right), \quad i = 1, \ldots, n$$

$$E(y_i) = z_i + \beta x_i, \quad i = 1, 2, \ldots, n$$

일반화선형모형 : 지수분포족을 가정함

$$y_i \overset{indep.}{\sim} \text{F}\left(\cdot\right), \quad i = 1, \ldots, n$$

$$g\left(E(y_i)\right) = z_i + \beta x_i, \quad i = 1, 2, \ldots, n$$

AIDS data : $\log(\mu_i) = \alpha + \beta\, x_i$

2 연결함수와 분산함수

연결함수

연결함수(link function) :

선형예측치 $\eta = \alpha + \beta x$ 와 평균모수 μ 사이의 관계를
$g(\mu) = \eta$ 가 되도록 연결해주는 함수 $g(\cdot)$.

AIDS data : $\log(\mu) = \alpha + \beta x$

즉, 연결함수는 $g(\mu) = \log(\mu)$

이항분포의 연결함수

로짓 (Logit) : $\eta(p) = \log\left(p/(1-p)\right)$

프라빗 (Probit) : $\eta(p) = \Phi^{-1}(p)$

$\Phi(\cdot)$: **표준정규분포의 누적분포함수**

cloglog (Complementary log–log) :

$$\eta(p) = \log\left(-\log(1-p)\right)$$

정준연결함수

지수족 분포에 대한 확률밀도함수를

$$f(y;\theta,\varphi) = \exp\{(y\,\theta - \gamma(\theta))/\varphi + \tau(y,\varphi)\}$$

라고 표현할 때, $\theta = \theta(\mu)$ 가 **정준연결함수**(canonical link function) 이다.

정규분포 : 평균이 μ 이고 분산이 σ^2인 경우, $\theta(\mu) = \mu$

이항분포 : 시행횟수가 n, 성공률이 p인 이항분포 경우, $y = \hat{p}$

$$\theta(\mu) = \log(\mu/(1-\mu)) \qquad \mu = p$$

포아송분포 : 평균이 λ 인 경우, $\theta(\mu) = \log\mu \qquad \mu = \lambda$

[표 6.1] 분포족과 연결함수. ★ 는 정준연결함수[5)]

	분포족 이름				
연결함수	이항분포	감마분포	정규분포	역정규분포	포아송분포
logit	★				
probit	●				
cloglog	●				
identity		●	★		●
inverse		★			
log		●			★
1/mu^2				★	
sqrt					●

분산함수

지수분포족 분포들의 분산은, 평균 μ에 대한 함수로 다음과 같이 표현된다.

$$E(y) = \mu \qquad Var(y) = \varphi V(\mu)$$

이때 평균과 분산사이의 관계를 설명하는 함수 $V(\mu)$를 **분산함수**(variance function)라 한다.

정규분포 : $V(\mu) = 1$

이항분포 : $V(\mu) = \mu(1-\mu)$

포아송분포 : $V(\mu) = \mu$

〈표 6.2〉 정준연결함수와 분산함수[6)]

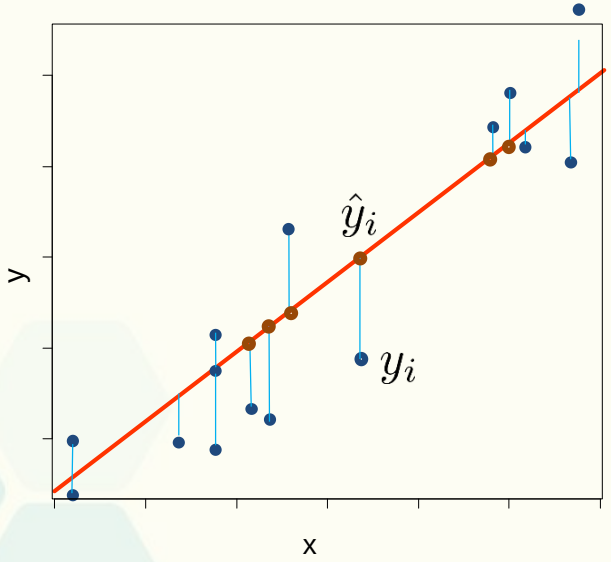
분포족	정준연결함수	이름	분산함수	이름
이항분포 binomial	$\log(\mu/(1-\mu))$	logit	$\mu(1-\mu)$	mu(1−mu)
감마분포 Gamma	$-1/\mu$	inverse	μ^2	mu^2
정규분포 gaussian	μ	identity	1	constant
역정규분포 inverse.gaussian	$-2/\mu^2$	1/mu^2	μ^3	mu^3
포아송분포 poisson	$\log\mu$	log	μ	mu

3 최대우도법

최소자승법

$$y_i \overset{indep.}{\sim} \text{Normal}\left(\alpha + \beta x_i, \sigma^2\right), \quad i = 1, \ldots, n$$

$$\theta = (\alpha, \beta) \qquad \mu_i(\theta) = \alpha + \beta x_i \qquad \hat{\mu}_i = \mu_i(\hat{\theta}) = \hat{\alpha} + \hat{\beta} x_i$$



잔차제곱합 함수 :

$$SSE(\theta) = \sum_{i=1}^{n} (y_i - \mu_i(\theta))^2$$

잔차제곱합 :

$$SSE(\hat{\theta}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$$

최대우도법

$$y_i \overset{indep.}{\sim} \text{Normal}\left(\alpha + \beta x_i, \sigma^2\right), \quad i = 1, \ldots, n$$

$$\theta = (\alpha, \beta, \sigma)$$

$$f(\mathbf{y}; \theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i(\theta))^2\right\}$$

(두 배의) 음로그우도 함수 :

$$l(\theta) = -2\log f(\mathbf{y}; \theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i(\theta))^2 + constant$$

$$SSE(\theta) = \sum_{i=1}^{n}(y_i - \mu_i(\theta))^2$$

포아송분포의 우도

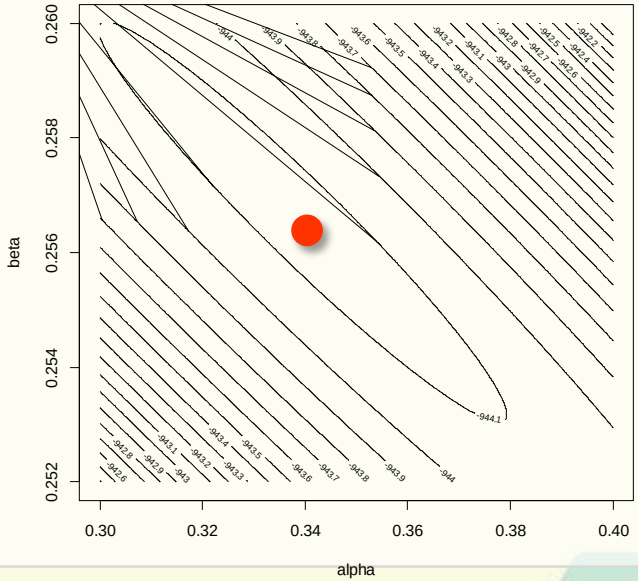
AIDS data :

$$l(\theta) = -2\sum_{i=1}^{n}(y_i \log \lambda_i - \lambda_i - \log(y_i!))$$

$$\log \lambda_i = \alpha + \beta x_i$$

$$\hat{\alpha} = 0.3396$$

$$\hat{\beta} = 0.2565$$



모형과 우도

$$SSE(M) = SSE(M; \hat{\theta})$$

$$l(M) = l(M; \hat{\theta})$$

$$M_1 : \quad y_i \overset{indep.}{\sim} \text{Normal}\left(\alpha, \sigma^2\right), \quad i = 1, \ldots, n$$

$$M_2 : \quad y_i \overset{indep.}{\sim} \text{Normal}\left(\alpha + \beta x_i, \sigma^2\right), \quad i = 1, \ldots, n$$

$$SSE(M_2) \leq SSE(M_1)$$

$$l(M_2) \leq l(M_1)$$

포화모형

모수 개수의 증가 ➡ $l(M)$ 감소

모형 예측값들이 관측값들과 모두 동일해지도록, 즉

$$y_i = \mu_i(\hat{\theta}), \quad i = 1, 2, \ldots, n$$

이 되도록, 모수의 수를 증가시킨 모형.

M_∞ : 포화모형, $l(M_\infty) \leq l(M)$

정규분포: $l(M_\infty) = 0$ (비정규 분포 X)

4 잔차와 이탈도

이탈도

이탈도(deviance) : 포화모형에서의 값이 0 이 되도록 하고
두 배의 음로그우도를 보정하여, 산포모수를 곱한 값.

$$D(M) = \varphi \cdot \{l(M) - l(M_\infty)\}$$

- $D(M) \geq 0$ 최소값 0
- $SSE(M) \geq 0$
- 정규분포 : 잔차제곱합 = 이탈도

이탈도 잔차

잔차 $r_i = y_i - \hat{\mu}_i$ → 제곱합 → $SSE(M)$: 잔차제곱합

이탈도 잔차 d_i → $D(M)$: 이탈도

$$d_i = \text{sign}(y_i - \hat{\mu}_i) \cdot \sqrt{2 \sum_{i=1}^{n} \left\{ (y_i\, \theta_{\infty,i} - \gamma(\theta_{\infty,i})) - (y_i\, \hat{\theta}_i - \gamma(\hat{\theta}_i)) \right\}}$$

$$D(M) = \sum_{i=1}^{n} d_i^2$$

피어슨 잔차

피어슨(Pearson) 잔차 : 잔차 $r_i = y_i - \hat{\mu}_i$ 를 분산함수의 제곱근으로 나누어 얻은 값. 즉,

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

- $Var(y_i - \mu_i) = \varphi V(\mu_i)$
- $Var(r_i^p) \approx \varphi$

AIDS data : 잔차

Y	X	예측치(Muhat)	Y-Muhat	이탈도잔차	피어슨잔차
0	1	1.815	-1.815	-1.905	-1.347
1	2	2.346	-1.346	-0.993	-0.879
2	3	3.032	-1.032	-0.632	-0.593
3	4	3.919	-0.919	-0.484	-0.464
1	5	5.064	-4.064	-2.21	-1.806
4	6	6.545	-2.545	-1.073	-0.995
9	7	8.46	0.540	0.184	0.186
18	8	10.933	7.067	1.953	2.137
23	9	14.131	8.869	2.161	2.359
31	10	18.263	12.737	2.708	2.98
20	11	23.603	-3.603	-0.762	-0.742
25	12	30.506	-5.506	-1.029	-0.997
37	13	39.427	-2.427	-0.391	-0.386
45	14	50.956	-5.956	-0.851	-0.834

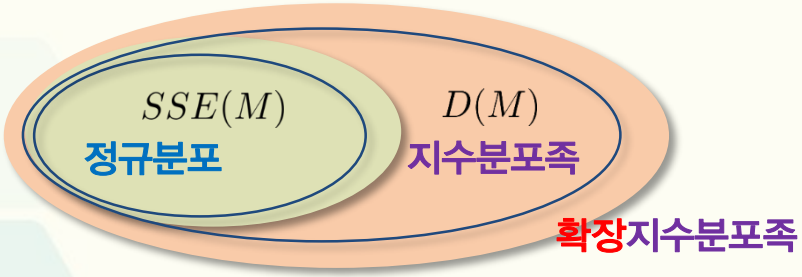
확장지수분포족

지수분포족 : 확률밀도함수로 분포를 특정.

확장지수분포족 : 평균과 분산사이의 관계만으로 분포를 특정.

의사포아송(quasi poisson)분포 : $V(\mu) = \mu$

의사이항(quasi binomial)분포 : $V(\mu) = \mu(1-\mu)$



AIC

잔차제곱합, 이탈도 : 특정 모형에 대한 자료의 적합/부적합 척도

좋은 모형: 자료 적합도가 높고, 가능한 단순한 모형

AIC (Akaike Information Criterion) :

$$AIC = l(\theta) + 2k$$

$$l(\theta) = -2\log f(\mathbf{y};\theta)$$

k : 모형 모수의 개수

5 R에서의 GLM 적용 방법

함수

lm() : 선형모형에 특화된 함수.

glm() : 일반화선형모형 전체에 적용되는 함수.

```
glm(formula, family, data)
```

AIDS data:

```
x <- 1:14
y <- c(0,1,2,3,1,4,9,18, 23,31,20,25,37,45)

glm( y~x, family="poisson" )

aids<- data.frame(x=x,y=y)
glm(y~x, family="poisson", data=aids)
```

formula 인자

formula 인자 : (일반화)선형모형에서 모형을 설정하는 인자.

예 : 종속변수 이름이 “y”라고 할 때,

$$g(\mu(x)) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

glm(y~ 1+x1+x2, ….)
glm('y~ 1+x1+x2', ….)
glm("y~ 1+x1+x2", ….)

절편없는 모형

```
x <-  1:10
y <-  1+2*x+rnorm(10)
# (A, B) and (C, D) show the same results
lm(y ~ x)    #  A
lm(y ~ 1+x)  #  B
lm(y ~ 0+x)  #  C
lm(y ~-1+x)  #  D
```

〈예 6.1〉 formula 전달인자 설정 방법

요인변수의 처리

```
sex<-factor(rep(c("M","F"), e=8))
type<-factor(rep(c("A","B","O","AB"), 4))
score<- 10*c(rnorm(8,0.5), rnorm(8,1))
lm(score ~ sex + type )                  # 교호작용이 없는 모형
lm(score ~ sex + type + sex:type )       # 교호작용이 있는 모형
```

〈예 6.2〉 교호작용이 있는 모형과 없는 모형

lm(score ~ sex * type)

lm(score ~ sex/height)

특별한 표현

lm(score ~ x + I(1/x))

$$\mu(x) = \alpha + \beta_1 x + \beta_2 (1/x)$$

lm(score ~ x + offset(0.1/x))

$$\mu(x) = \alpha + \beta_1 x + (0.1/x)$$

결과 사용 예

```
> aids.out<-glm(y~x, family="poisson",data=aids)
> class(aids.out)
[1] "glm" "lm"
> anova(aids.out)
Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    13    207.272
x     1   177.62        12     29.654
```

glm() 클래스에 대한 메쏘드

anova 분산분석표를 순차적으로 제시하거나, 몇 개의 계층적인 모형을 비교

coef(혹은 coefficients) (일반화)선형모형의 계수를 제시

deviance 잔차제곱합 제시

fitted(혹은 fitted.values) 적합된 값 제시

print 단순결과 제시

predict 새로운 자료에 대한 평균값을 예측하거나, 표준오차 선택적 제공

plot 진단 그림.

resid(혹은 residuals) 잔차 확인

plot 진단 그림.

update 모형의 재적합

MASS 패키지 함수

addterm 기존 적합된모형에 하나의 항을 추가한 모형들을 제시

dropterm 기존 적합된모형에 하나의 항을 뺀 모형들을 제시

stepAIC AIC를 기준으로 한 stepwise 모형 선택

vcov 모수 추정값들에 대한 분산 공분산 행렬

다음시간 안내

일반화선형모형 (2)