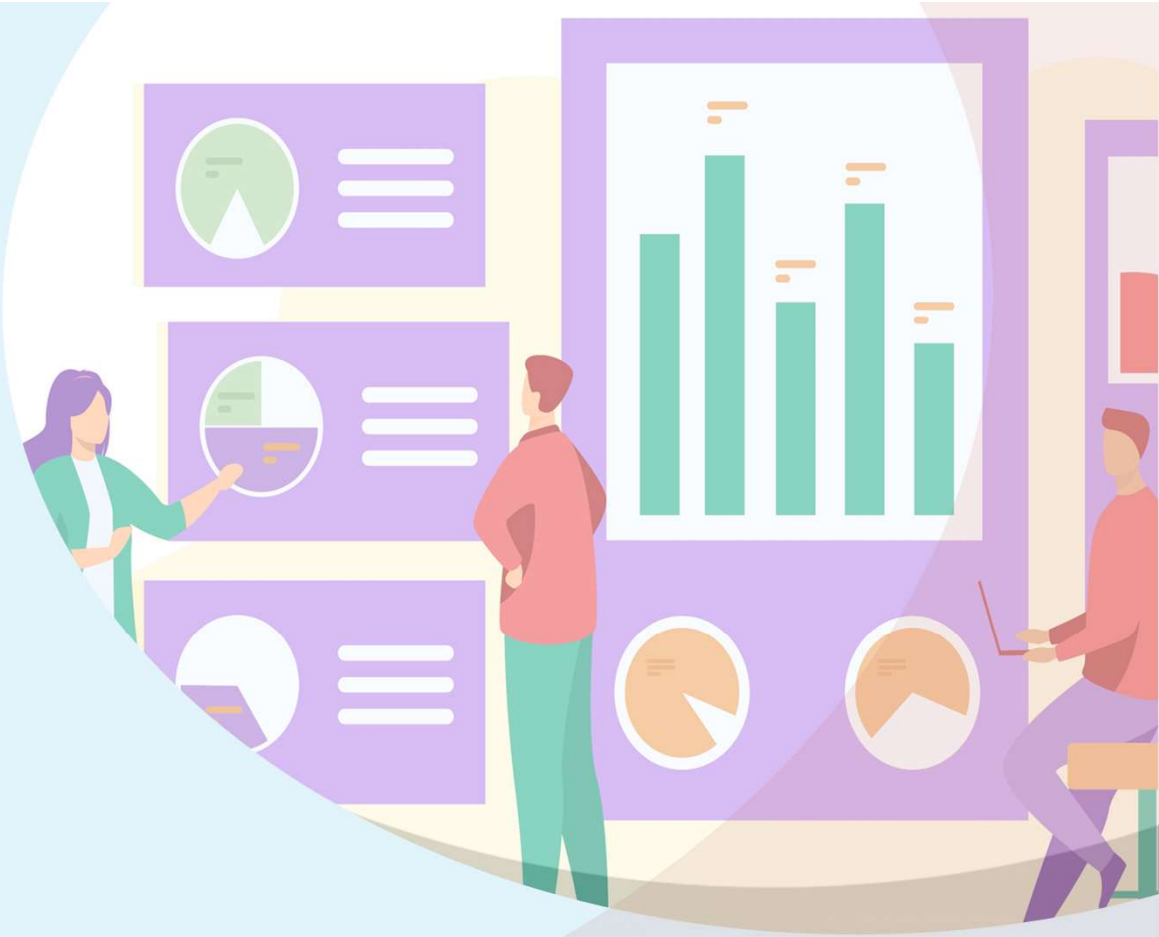
12 강

R 데이터 분석

일반화선형모형 I



서강대학교 이윤동 교수

한국방송통신대학교
Korea National Open University

강의 목차

학습목차

1. 일반화 선형모형 소개
2. 확장지수분포족

회귀분석 :

- 종속변수 y 가 **정규분포**를 따름.
- **선형** 회귀모형, **비선형** 회귀모형
- 독립변수가 명목변수일 때, **가변수**를 이용.

선형모형 :

- 종속변수 y 가 **정규분포**를 따름.
- **선형** 회귀모형
- 독립변수로 **명목변수**가 사용 가능함을 **강조**

선형모형의 한계:

종속변수 (y) : 수술 폐암환자 5년 **생존** (1/0) 여부

독립변수 (x) : 수술 시점 환자의 나이

일반화 선형모형 :

- 선형모형을 확장한 모형

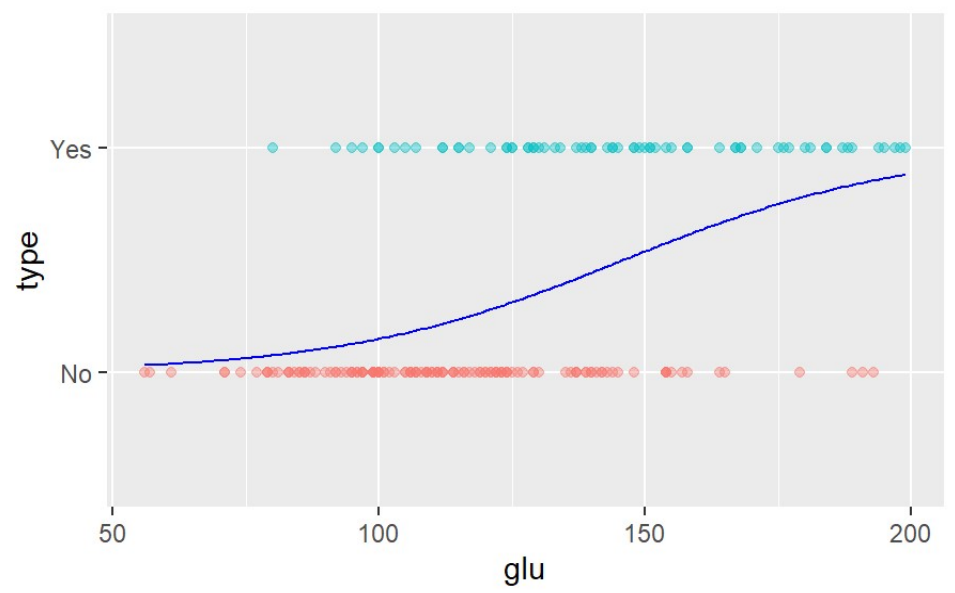
- 종속변수 y 가 **정규분포 이외의 분포**인 경우 고려

일반화 선형모형의 예 1:

Pima Indian diabetes data:

미국 피마 인디언 여성의 혈당과 당뇨판정

glu: 구강검사에 의한 혈당
type: WHO 기준 당뇨판정



일반화 선형모형의 예 2:

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

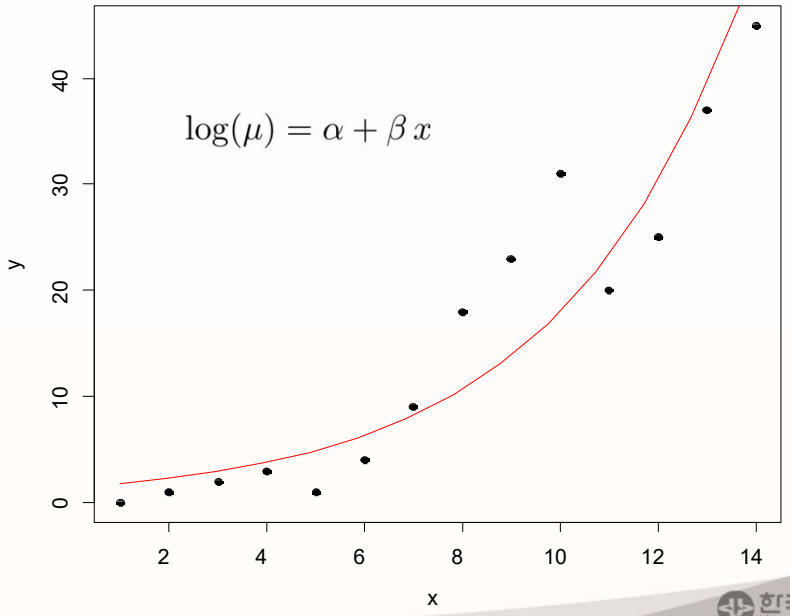
1983~1986년 동안 Australia에서 AIDS로 인한 사망자 수

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



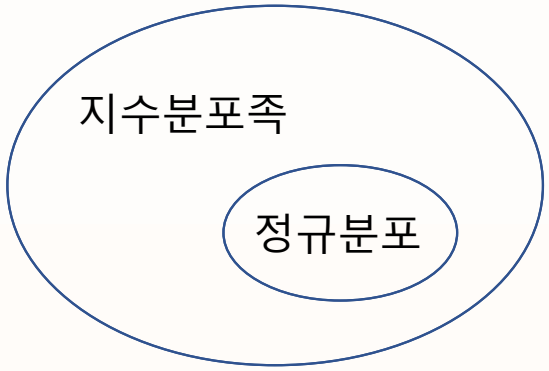
일반화 선형모형 :

- 종속변수 y 가 (**확장**) **지수분포족**임을 가정

- **정준모수** θ 를 모형화

지수분포족 :

- 정규분포를 확장한 분포족
- 유용하고 주요한 분포들이 대부분 포함
- **정규분포**, 지수분포, **이항분포**, **포아송분포**, ...



지수분포족 :

확률밀도함수/확률질량함수의 형태가 다음과 같은 분포

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

정규분포 :

$$y \sim N(\mu, \sigma^2)$$

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

12 일반화 선형모형 I

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

$$\log f(y;\mu,\sigma^2) = -(1/2)\log(2\pi\sigma^2) + \frac{1}{\sigma^2}(y\mu - \mu^2/2 - y^2/2)$$

$$\log f(y;\mu,\sigma^2) = (y\mu - \mu^2/2)/\sigma^2 - (1/2)\left(y^2/\sigma^2 + \log(2\pi\sigma^2)\right)$$

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

$$\theta = \mu \qquad \varphi = \sigma^2 \qquad \gamma(\theta) = \mu^2/2$$

포아송분포 :

$$y \sim \text{Poisson}(\lambda)$$

$$\mu = E(y) = \lambda$$

$$f(y;\lambda) = e^{-\lambda}\frac{\lambda^y}{y!}, \quad y = 0, 1, 2, \ldots$$

12 일반화 선형모형 I

$$f(y;\lambda) = e^{-\lambda}\,\frac{\lambda^y}{y!}$$

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

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

$$\theta = \log\lambda \qquad \varphi = 1 \qquad \gamma(\theta) = \lambda = e^{\theta}$$

이항분포 :

$$y \sim \text{Bin}(1, p), \quad y = 0, 1$$

$$\mu = E(y) = p$$

$$f(y; p) = p^y (1 - p)^{(1-y)}$$

$$f(y;p) = p^y(1-p)^{(1-y)}$$

$$\log f(y;p) = \left(y \log \frac{p}{1-p} + \log(1-p) \right)$$

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

$$\theta = \log \frac{p}{1-p} \qquad \varphi = 1 \qquad \gamma(\theta) = -\log(1-p) = \log(1+e^{\theta})$$

정규분포:

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

$$\theta = \mu$$

이항분포:

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

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

포아송분포:

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

$$\theta = \log\mu \qquad \mu = \lambda$$

정준모수, 산포모수 :

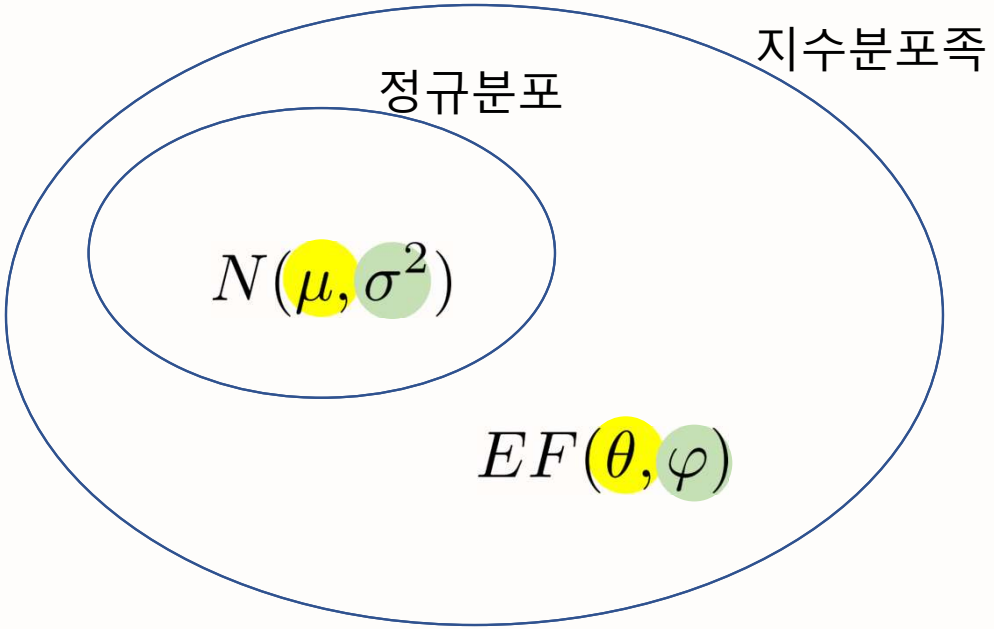
지수분포족에 속하는 분포들에 대하여,

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

θ : **정준모수**, 정규분포에서는 **평균**

φ : **산포모수**, 정규분포에서는 **분산**

12 일반화 선형모형 I



선형모형 :

종속변수가 정규분포 임을 가정

$$y \sim N(\mu, \sigma^2)$$

평균 μ 를 모형화

예) $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

일반화 선형모형 :

종속변수가 지수분포족임을 가정

$$y \sim EF(\theta, \varphi)$$

정준모수 θ 를 모형화

예) $\theta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

정규분포:

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

$$\theta(\mu) = \mu$$

이항분포:

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

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

포아송분포:

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

$$\theta(\mu) = \log\mu \qquad \mu = \lambda$$

정준연결함수 canonical link function

- 평균 μ 의 함수로써의 정준모수, 즉 $\theta(\mu)$

- 각 분포마다, 대응하는 **정준연결함수** 가 있음

정준연결함수 :

- 정규분포: **항등함수** $\theta(\mu) = \mu$
- 이항분포: **로짓함수** $\theta(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
- 포아송분포: **로그함수** $\theta(\mu) = \log\mu$

연결함수 link function

- 정준연결함수 대신 사용할 수 있는 함수
- 부드러운 단조함수 $g(\mu)$
- 특별한 이유가 없다면 정준연결함수를 사용
- 예외적으로, 각 분포에 따라 다양한 형태의 적절한 연결함수 사용이 가능

포아송분포의 연결함수 :

- 로그함수: $g(\mu) = \log(\mu)$ 정준연결함수

- 항등함수: $g(\mu) = \mu$

- 제곱근함수: $g(\mu) = \sqrt{\mu}$

12 일반화 선형모형 I

표 6-2 분포족과 연결함수(★는 정준연결함수)

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

이항분포의 연결함수 :

- **로짓함수**: $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$

- 보충로그로그함수: $\text{cloglog}(p) = \log\left(-\log(1-p)\right)$

- 프라빗함수: $\text{probit}(p) = \Phi^{-1}(p)$

$\Phi(\cdot)$: cdf of $N(0,1)$

비율(proportion): 전체 중에서 해당하는 정도,

예: 6승 4패, 승률 p = 0.6

오즈(odds): 비해당 경우에 대한 해당 경우의 비율,

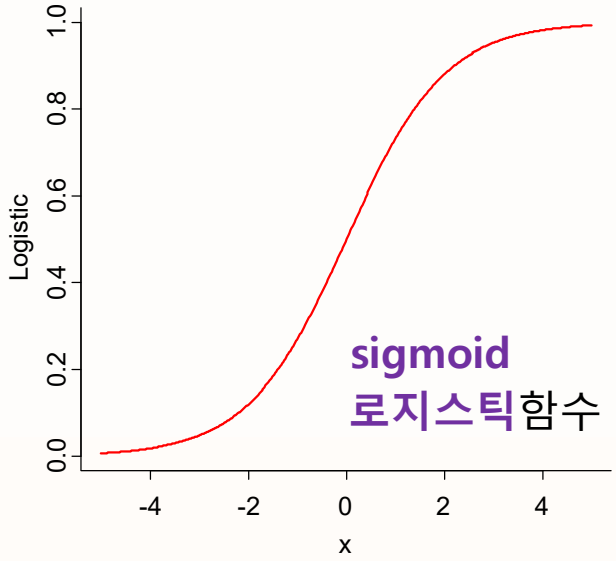
예: 6승 4패, 승패율 d = 6/4 =1.5

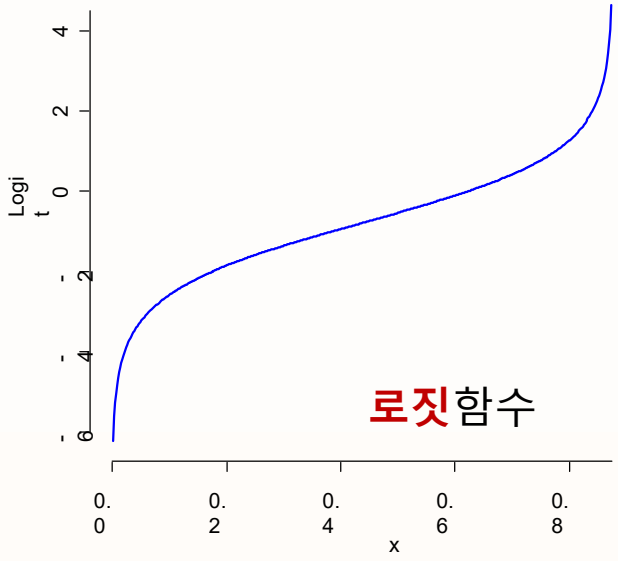
$$d = p/(1-p)$$

로짓함수: $\text{logit}(p) = \log d = \log \frac{p}{1-p}$

로지스틱함수:

- **로짓함수**의 **역함수**, $\text{logistic}(x) = \frac{e^x}{1+e^x}$





선형모형 : **정규분포**를 가정, **평균**을 모형화

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

일반화 선형모형 : **지수분포족**을 가정, **정준모수**를 모형화

$$\theta(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

학습목차

1 일반화 선형모형 소개

2 **확장지수분포족**

평균과 분산 :

- 정규분포: $Var(y) = \sigma^2 = \varphi \cdot 1, \qquad \varphi = \sigma^2$

- 이항분포: $Var(y) = p(1-p) = 1 \cdot \mu(1-\mu), \qquad \varphi = 1$

- 포아송분포: $Var(y) = \lambda = 1 \cdot \mu, \qquad \varphi = 1$

분산함수 :

- 평균과 분산 사이의 관계를 나타내는 함수 $V(\mu)$
- $Var(y) = \varphi \cdot V(\mu)$

분산함수 예 :

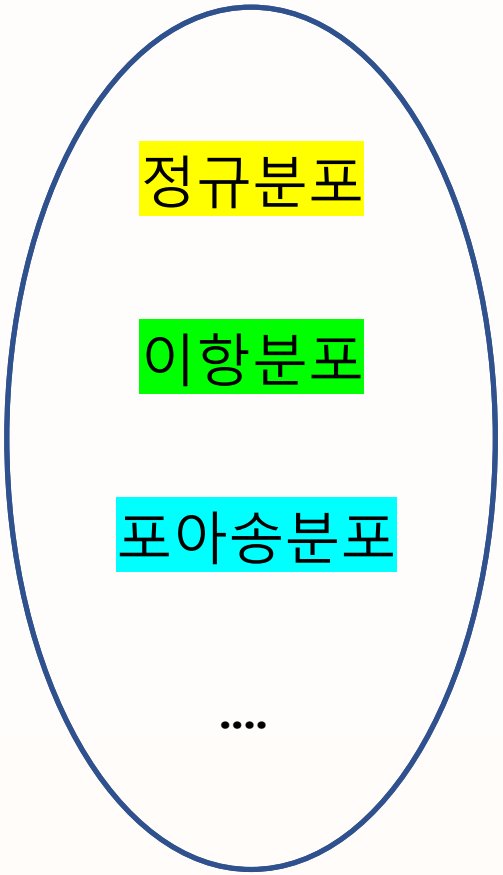
- 정규분포: $Var(y) = \varphi \cdot 1$ $V(\mu) = 1$

- 이항분포: $Var(y) = 1 \cdot \mu(1-\mu)$ $V(\mu) = \mu(1-\mu)$

- 포아송분포: $Var(y) = 1 \cdot \mu$ $V(\mu) = \mu$

12 일반화 선형모형 I

$$y \sim EF(\theta, \varphi)$$



분포의 특정

- 자료분석에서 분포족을 정확하게 특정할 수 있을까?

- 분포족을 특정할 수 없다면 대안은 뭘까?

분산함수에 의한 분포족 특정

- 관찰된 자료로부터, **분포족** 특정은 불가능
- 관찰된 자료로부터, **분산함수** 특정은 가능
- **분포족** 특정 대신, **분산함수**만을 특정하자

확장지수분포족

- 지수분포족의 범위를 확장한 개념
- **분산함수**만을 특정하여, 분포족 특정을 대신함
- 다양한 **산포모수**를 가질 수 있고, 추정할 수 있음
- 과산포분포, 미산포분포가 포함됨

과산포 분포, 미산포 분포

$$Var(y) = \varphi \cdot V(\mu)$$

- 과산포 이항분포, 과산포 포아송분포 $\varphi > 1$

- 미산포 이항분포, 미산포 포아송분포 $\varphi < 1$

학습요약

1. 일반화 선형모형 소개
2. 확장지수분포족

학습 내용 요약

선형모형 : **정규분포**를 가정, **평균**을 모형화

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

일반화 선형모형 : **지수분포족**을 가정, **정준모수**를 모형화

$$\theta(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

학습 내용 요약

분산함수 :

분포의 평균과 분산의 관계를 나타내는 함수

확장지수분포족 :

분산함수로 특정되는 지수분포족

13 강

다음시간 안내

일반화선형모형 II

수고하셨습니다!