DSL 심화스터디 <기초반>

ANOVA /
logistic regression

발표자 : 5기 박채은

ANOVA

ANOVA?

ANOVA = Analysis of variance = 분산 분석
: 3개 이상 모집단의 평균을 비교

ANOVA 예) 자동차 충돌 실험: 자동차 그룹에 따른 마네킹 인형의 파손 정도는?
⇒자동차 등급 별로 파손 평균에 차이가 있는지를 알아보기
Response variable: 마네킹 인형의 파손 정도
Predictor variable(=explanatory variable) : 자동차 등급
처리/level: 대형, 중형, 소형

ANOVA

Anova의 predictor variable(=설명변수)은 범주형 설명 변수
범주형 설명변수(categorical explanatory variable) = 요인 변수(factor variable)

하나의 factor variable에 대해 그룹(treatment, level)이 나눠짐.

Factors	자동차 등급	Sound	Font size
Levels (group/treatment)	대형 중형 소형	Music No music	Small Medium large
Level 수	3개	2개	3개

ANOVA

우리가 배웠던 ANOVA 방식

1. 가설세우기

H_0 : $\mu_1 = \mu_2 = \mu_3$

H_1 : not H_0(적어도 하나는 다르다)

2. 처리 내 변동(그룹 내 변동)과 처리 간 변동(그룹 간 변동) 비교

처리 간 변동 >> 처리 내 변동 => 귀무가설 기각

강의에서는 모델에 더 초점!

ANOVA

일원분산분석(one-way anova) vs 이원분산분석(two-way anova)
요인(factor)의 개수에 따라 구분
요인이 한 가지이면 일원분산분석 예) 자동차 충돌 실험
요인이 두 가지이면 이원분산분석 예) 온라인 마케팅에 영향 미치는 요인

예) 온라인 쇼핑몰 사이트 매출에 영향을 미치는 two factors

Factors	sound	Font size
levels	Music No music	Small Medium large

여섯 가지의 처리 조합 가능

one factor일 때의 모델

G levels을 갖는 one explanatory variable의 모델

1.

$$y_i | g_i, \underset{\sim}{\mu}, \sigma^2 \overset{ind}{\sim} N(\mu_{g_i}, \sigma^2)$$

$$g_i \in \{1, \dots, G\}$$

$$i = 1, \dots, n$$

one factor일 때의 모델

G levels을 갖는 one explanatory variable의 모델

2. (R에서 디폴트인 방식)

$$E(y_i) = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_{G-1} X_{G-1,i}$$

이 두 방식은 모두 모델에서 mean을 나타낼 때 G개의 parameter을 가짐.

ANOVA

등분산을 가정하기 어렵다면?

분산분석에서는 등분산을 가정함.

그러나 등분산 가정을 하기 어려운 경우에는?
서로 다른 그룹에 대해 서로 다른 분산을 가지는 것을 고려하기

분산분석 가정사항
1. 정규성 가정
2. 독립성 가정
3. 등분산 가정

Two factors일 때의 모델

1. cell-means model
- Different mean for each treatment combination

		B		
		1	2	3
A	1	$\mu_{1,1}$	$\mu_{1,2}$	$\mu_{1,3}$
	2	$\mu_{2,1}$	$\mu_{2,2}$	$\mu_{2,3}$

Parameter 수 6개!

Two factors일 때의 모델

2. Additive model
- Variable 사이에 interaction이 없을 때
- 6개보다 적은 4개의 parameter 사용

		B		
		1	2	3
A	1	$\mu_{1,1}$	$\mu_{1,2}$	$\mu_{1,3}$
	2	$\mu_{2,1}$	$\mu_{2,2}$	$\mu_{2,3}$

$$E(y_i) = \underbrace{\mu}_{\text{baseline}} + \alpha_2 I_{(a_i=2)} + \beta_2 I_{(b_i=2)} + \beta_3 I_{(b_i=3)}$$

Logistic regression

Logistic regression

Response variable y가 binary(0, 1)일 때

Normal likelihood를 사용하지 않고 Bernoulli likelihood를 사용

모델

$$y_i \mid \phi_i \overset{ind}{\sim} Bern(\phi_i) \quad i = 1, \ldots, n$$

(ϕ_i: 성공확률)

$$E(y_i) = \phi_i$$

~~$E(y_i) = \phi_i = \beta_0 + \beta_1 x_{1,i}$~~

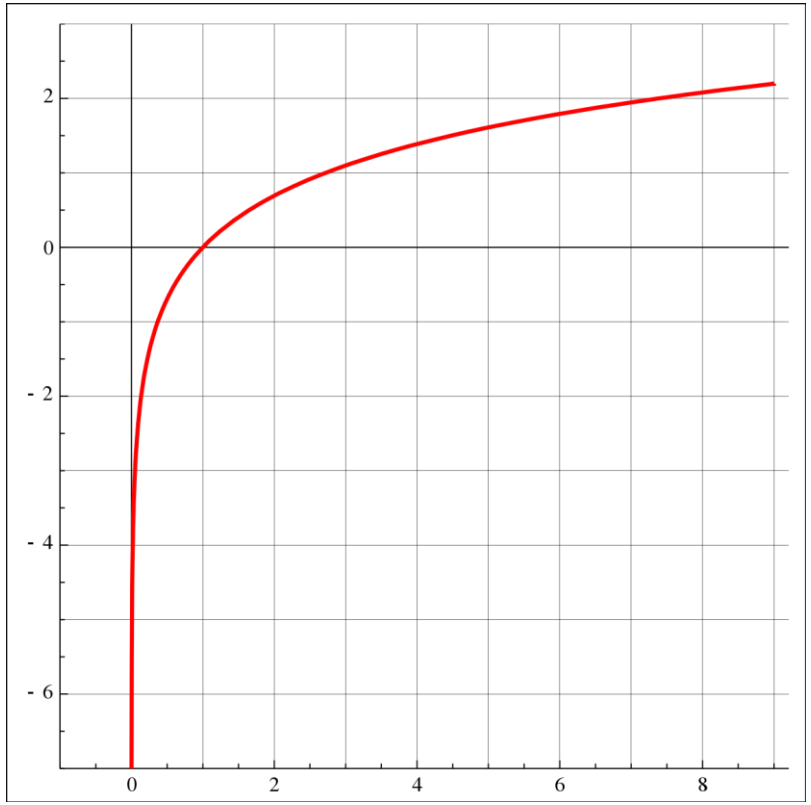
Logistic regression

제약을 두는 대신 link func을 사용!

Logit link(=logistic link): $\log(\frac{\emptyset}{1-\emptyset})$

$$\text{Logit}(\phi_i) = \log(\frac{\phi_i}{1-\phi_i}) = \beta_0 + \beta_1 X_{1,i}$$

$$E(y_i) = \phi_i = \frac{e^{\beta_0+\beta_1 X_{1,i}}}{1+e^{\beta_0+\beta_1 X_{1,i}}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1,i})}}$$



Logistic regression

Logistic regression model에서 Prediction

새로 관측된 X 값이 있을 때, 모델에 X 값을 넣어서 phat을 예측

$$E(y_i) = \phi_i = \frac{e^{\beta_0+\beta_1 x_{1,i}}}{1+e^{\beta_0+\beta_1 x_{1,i}}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1,i})}}$$

Phat 값을 통해 y 값을 분류: 0 또는 1

예) Phat < 0.3이면 0으로 분류
phat > 0.3이면 1로 분류