

가정이 맞아야 분석방법의 결과가 타당하다

## ■ 분산분석 검진(ANOVA Diagnostics)

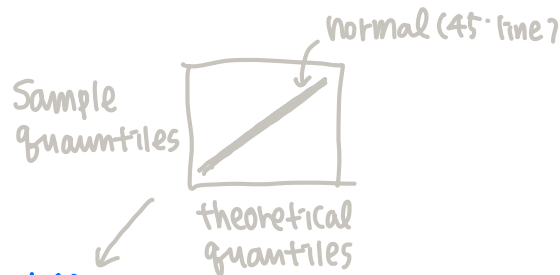
independent & identical distributed

- 기본가정:  $\varepsilon_{ij} \sim \text{iid } N(0, \sigma^2)$   $\Leftrightarrow$  잔차분석(residual analysis)

- 등분산성  $\sigma^2$
- 독립성  $\Rightarrow$  잔차들 간에는 항상 상관관계가 존재  $e_{ij} = y_{ij} - \hat{y}_{ij}$
- 정규성
- 이상치 유무

### ● 잔차그림

- 잔차 vs 적합값
- 정규확률그림 : 분위수-분위수점plot (Q-Q plot)
- Box-plot



- 수치적 방법  $\rightarrow$  식제로 값 계산

추정하기 위해 다른정보를 쓰므로  
독립성이 깨짐

## 잔차

○  $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_i$  ← 모형이  $Y_{ij} = \mu_i + \varepsilon_{ij} \rightarrow E(Y_{ij}) = E(\mu_i + \varepsilon_{ij}) = \mu_i + E(\varepsilon_{ij}) = \mu_i \Rightarrow \bar{Y}_i$

○ studentized 잔차 :  $r_{ij} = \frac{e_{ij}}{\widehat{se}(e_{ij})}$    
 (normalize : 변동성 ↓)   
 $\text{Var}(\sigma_{ij}^2) = \text{Var}\left(\frac{\varepsilon_{ij}}{\widehat{se}(e_{ij})}\right) = \frac{\sigma^2}{\widehat{se}(e_{ij})^2} \rightarrow 1$    
 $\hat{\sigma}$  (standard error) : 변동성에 제한이 있음 (variability를 bound 시킴)

-  $e_{ij} = \frac{n_i - 1}{n_i} Y_{ij} - \frac{1}{n_i} \sum_{k \neq j} Y_{ik} \Leftarrow Y_{ij} - \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik} = Y_{ij} - \frac{1}{n_i} (Y_{ij} + \sum_{k \neq j} Y_{ik})$  : 독립인 두 term으로 만들기 위해

-  $\text{Var}(e_{ij}) = \frac{(n_i - 1)^2}{n_i^2} \sigma^2 + \frac{n_i - 1}{n_i^2} \sigma^2 = \frac{n_i - 1}{n_i} \sigma^2$

-  $\widehat{se}(e_{ij}) = \sqrt{\frac{(n_i - 1) \text{MSE}}{n_i}} \Leftarrow \text{s.e.}(e_{ij}) = \sqrt{\frac{n_i - 1}{n_i} \sigma^2}$

○ studentized deleted 잔차 :  $t_{ij} = e_{ij} \left[ \frac{N - p - 1}{SSE(1 - 1/n_i) - e_{ij}^2} \right]^{1/2}$

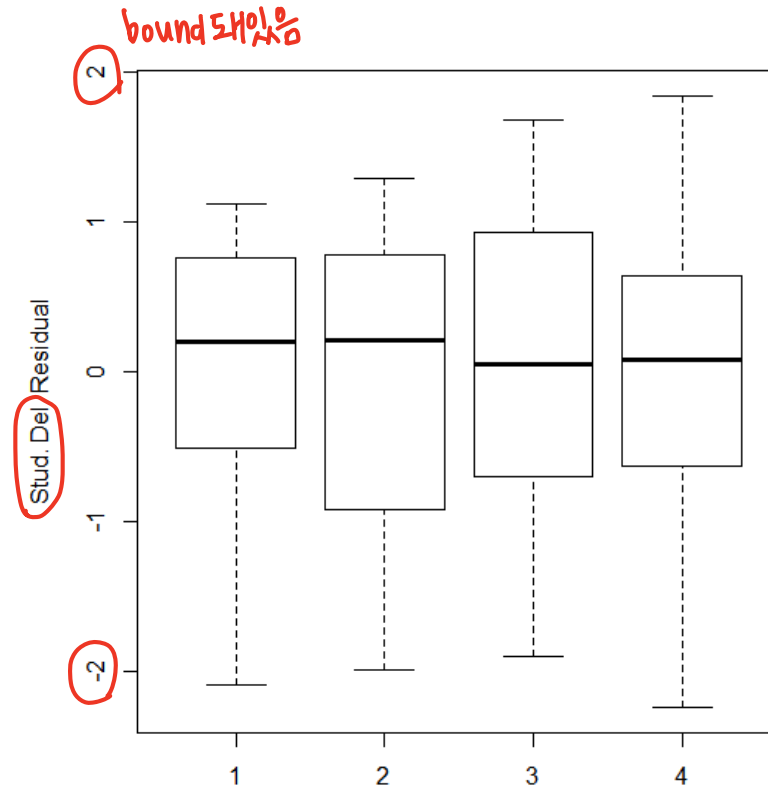
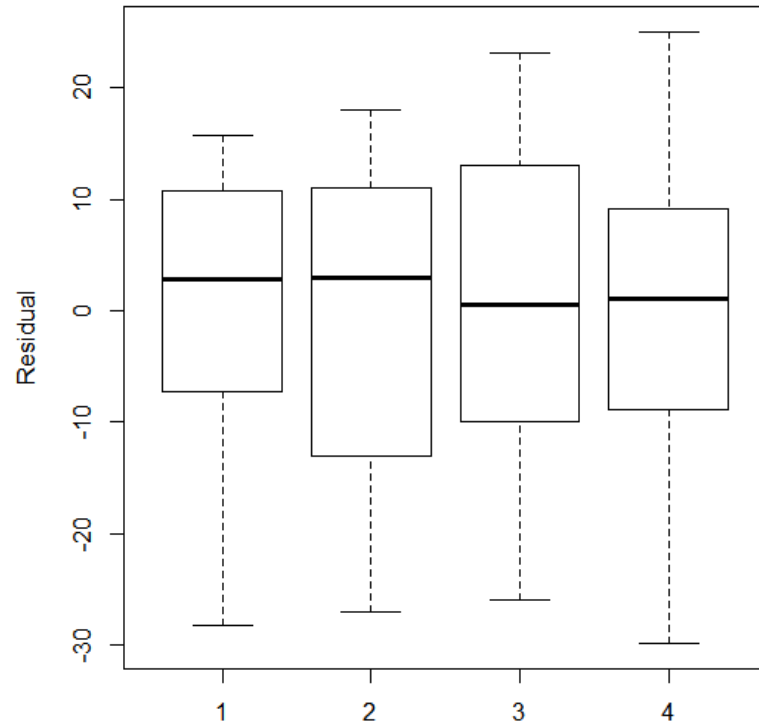
리뷰

$$\text{stud. } R_{11} = r_{11} = \frac{90-79.2}{\sqrt{\frac{10-1}{10} \times 223.59}} = 0.76 \quad \text{stud. del. } R_{11} = t_{11} = 10.8 \times \left( \frac{40-4-1}{8049.4 \times \left( (1-\frac{1}{10}) - 10.8^2 \right)} \right)^{\frac{1}{2}} = 0.76$$

잔차만 보면 얼마나  
거야 별로인지 감이  
안옴  
→ variability를 bound시킨  
누드를 쓰면 소2 눈방에  
있을 때 별로라고  
알 수 있음!

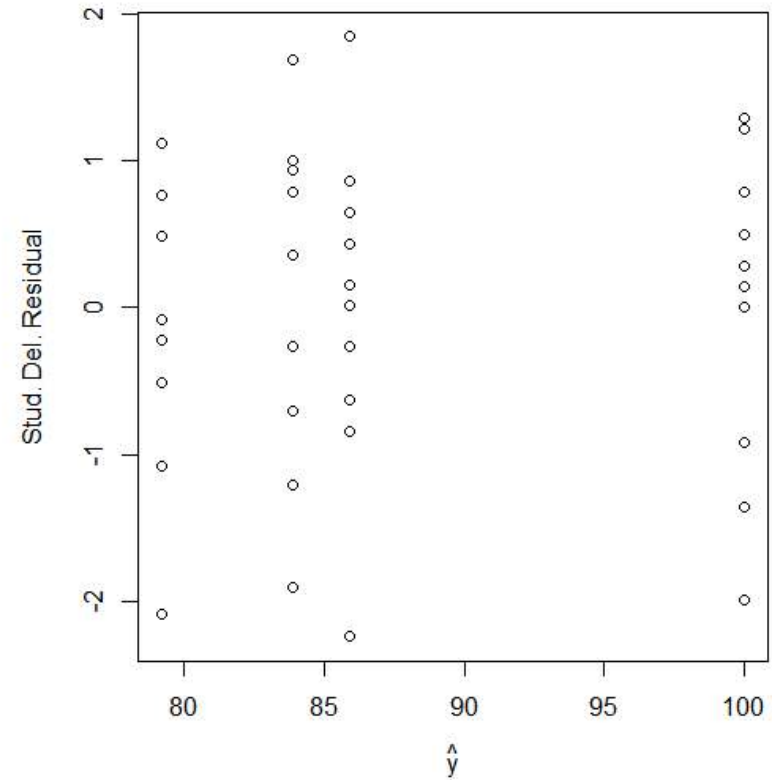
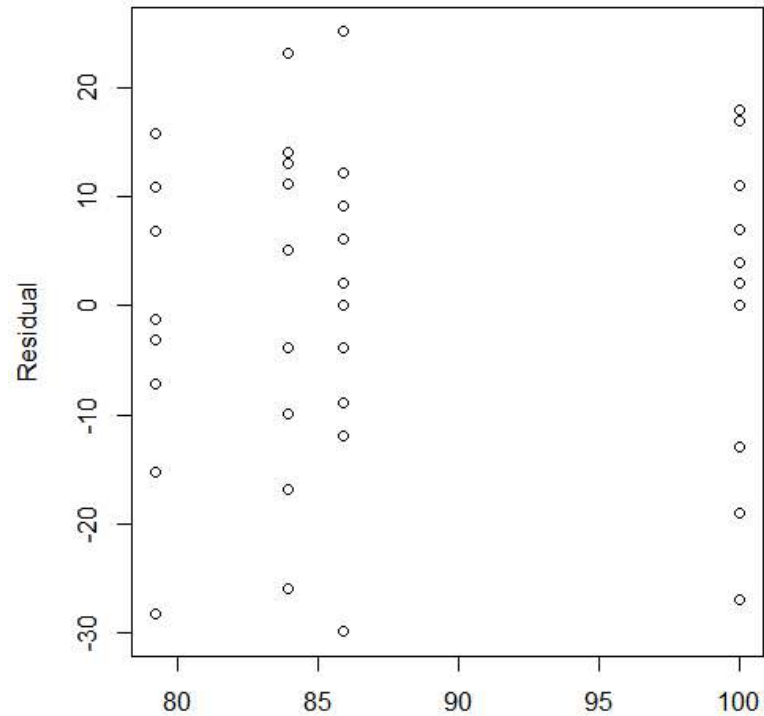
사료 (평균)	내용	1	2	3	4	5	6	7	8	9	10
1 (79.2)	관측값 <small>90-79.2</small>	90	76	90	64	86	51	72	90	95	78
	잔차	10.8	-3.2	10.8	-15.2	6.8	-28.2	-7.2	10.8	15.8	-1.2
	stud. R.	0.76	-0.23	0.76	-1.07	0.48	-1.99	-0.51	0.76	1.11	-0.08
	stud. Del. R.	0.76	-0.22	0.76	-1.08	0.48	-2.09	-0.51	0.76	1.12	-0.08
2 (100)	관측값	73	102	118	104	81	107	100	87	117	111
	잔차	-27	2	18	4	-19	7	0	-13	17	11
	stud. R.	-1.90	0.14	1.27	0.28	-1.34	0.49	0.00	-0.92	1.20	0.78
	stud. Del. R.	-1.99	0.14	1.29	0.28	-1.36	0.49	0.00	-0.92	1.21	0.78
3 (83.9)	관측값 <small>107-83.9</small>	107	95	97	80	98	74	74	67	89	58
	잔차	23.1	11.1	13.1	-3.9	14.1	-9.9	-9.9	-16.9	5.1	-25.9
	stud. R.	1.63	0.78	0.92	-0.27	0.99	-0.70	-0.70	-1.19	0.36	-1.83
	stud. Del. R.	1.68	0.78	0.93	-0.27	1.00	-0.70	-0.70	-1.21	0.36	-1.90
4 (85.9)	관측값	98	74	56	111	95	88	82	77	86	92
	잔차	12.1	-11.9	-29.9	25.1	9.1	2.1	-3.9	-8.9	0.1	6.1
	stud. R.	0.85	-0.84	-2.11	1.77	0.64	0.15	-0.27	-0.63	0.01	0.43
	stud. Del. R.	0.86	-0.84	-2.24	1.84	0.64	0.15	-0.27	-0.63	0.01	0.43

box plot



(등분산성 성립여부  
이상점 존재여부)  $\Rightarrow$  graphically 확인가능!

## 잔차 vs 정확한값



$\hat{y}$

↑ 원배치분산분석에서의  $\bar{y}_i$  4개

(등분산성 성립여부  
이상점 존재여부)  $\Rightarrow$  graphically (퍼진 정도를 보고) 확인 가능!

graphically보다정확하게

## □ 등분산 검정

- 반복수가 같은 경우 동일한 분산을 가진다는 가정을 약간 어기는 경우  
분산분석 방법은 robust함 (등분산)
- 반복수가 다르거나 어떤 한 분산이 다른 분산들보다 상당히 큰 경우 분산분석  
방법은 robust하지 않음  $\Rightarrow$  분산들이 같은지 다른지를 검정필요  
(만분산)  $\rightarrow$  결과값이 막 바뀐다!
- 가설:  $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_p^2$  VS  $H_1$  : 최소한 하나 이상의 분산은 다름

# 레지션스

## ○ Hartley 검정

- 동일 반복수  $n$
- 검정통계량 :  $H^* = \frac{\max(S_i^2)}{\min(S_i^2)} \sim H(p, n-1)$  → Hartley's Fmax table 이용
  - $S_i^2 = \sum (Y_{ij} - \bar{Y}_{i.})^2 / (n_i - 1)$
- 기각역 :  $H^* > H(1 - \alpha, p, n - 1) \Rightarrow \frac{\max}{\min}$  이 충분히 등분산성이 깨짐

## ◎ 주 성장량

- $S_1^2 = 192.84, S_2^2 = 229.11, S_3^2 = 246.77, S_4^2 = 225.66$
- $H = \frac{246.77}{192.84} = 1.280 < H(0.95, 4, 9) = 6.31 \Rightarrow$  등분산을 만족함

## ○ Brown-Forsythe 검정

- 절대편차를 먼저 계산

$$D_{ij} = |Y_{ij} - \tilde{Y}_i| = |\text{observed 값} - \text{그룹의 중앙값}|$$

짝수인 경우 (2n개):  $\frac{n\text{번째} + (n+1)\text{번째}}{2}$

- $\tilde{Y}_i$ :  $i$  번째 그룹의 중앙값  
(4개나)에 대한 것

- 검정통계량 :  $F_{BF}^* = \frac{MSTR^*}{MSE^*} \simeq F_{p-1, N-p}$

## ◎ 주 성장량

- 중앙값:  $\tilde{Y}_1 = 82, \tilde{Y}_2 = 103, \tilde{Y}_3 = 84.5, \tilde{Y}_4 = 87$
- $\bar{D}_{1.} = 11, \bar{D}_{2.} = 11.4, \bar{D}_{3.} = 13.3, \bar{D}_{4.} = 10.9, \bar{D}_{..} = 11.65$   $\rightarrow$  각 집단의 평균
- TSS=2804.6, SSTR=37.7, SSE=2766.9
- $MSTR^* = 12.567, MSE^* = 76.858 \Rightarrow F_{BF}^* = 0.164 < F_{0.05, 3, 36} = 2.866$   
 $= SSTR/3 \quad = SSE/36$   
 $= 37.7/3 \quad = 2766.9/36$   
 $\Rightarrow H_0$  기각 불가, 즉 등분산성립



○ Bartlett 검정 : 분산에 대한 검정

기분전조

● 검정통계량 :  $\chi_0^2 = 2.3026 \frac{q}{c} \sim \chi_{p-1}^2$

○  $q = (N-p) \log_{10} MSE - \sum_{i=1}^p (n_i - 1) \log_{10} S_i^2$

○  $c = 1 + \frac{1}{3(p-1)} \left\{ \sum_{i=1}^p \frac{1}{n_i - 1} - \frac{1}{N-p} \right\}$

- Bartlett's 검정 통계량은 정규 가정에 매우 민감하기 때문에 정규 가정이 의심스러우면 사용할 수 없음

◎ 쥐 성장량

○  $q = 0.0614, \quad c = 1.0463 \quad \Rightarrow \quad \chi_0^2 = 0.135 < \chi_{3,0.05}^2 = 7.81 \quad \Rightarrow \quad H_0 \text{ 기각 불가, 즉 등분산성립}$

$$\begin{cases} q = (40-4) \log_{10} 223.59 - 9 \times (\log_{10} 192.84 + \log_{10} 229.11 + \log_{10} 246.77 + \log_{10} 225.66) = 0.0614 \\ c = 1 + \frac{1}{3 \times (4-1)} \left( \frac{4}{9} - \frac{1}{36} \right) = 1.0463 \end{cases}$$

## □ 정규성 검정

- Shapiro-Wilk test, Kolmogorov-Smirnov test, Cramer-von Mises test, Anderson-Darling test
- Jarque-Bera test

이 값이 크면 비정규성, 정규성 X

$$\underline{JB} = \frac{n}{6} \left( b_1 + \frac{1}{4} (b_2 - 3)^2 \right) \sim \chi^2(2) \text{ 카이스퀘어 자유도 2}$$

○  $\sqrt{b_1}$  : 왜도(skewness) =  $E\left[\left(\frac{x-M}{\sigma}\right)^3\right]$

○  $b_2$  : 첨도(kurtosis) =  $E\left[\left(\frac{x-M}{\sigma}\right)^4\right]$

비정규성

## □ 문제 발생 시 해결방안

### ① 변환(transformation)

등분산성 성립이 안된 경우

- 분산상수화변환(variance stabilizing transformation, 분산안정화 변환)
  - 잔차그림에서 잔차의 표준편차(분산)이  $\hat{Y}$ 의 값과 연관성을 보이는 경우
  - 분산을 상수화시키기 위한 변환을 찾는 방법
    - $\sigma_i^2 = \text{Var}(Y_{ij})$  와  $\mu_i = E(Y_{ij})$  사이에 함수관계가 존재하는 경우:  
↳  $\sigma_i^2$ 가  $\mu_i$ 에 따라 달라짐  
$$\sigma_i^2 = f(\mu_i)$$
    - 예:  $\sigma_i^2 = c\mu_i^2$  ( $\sigma_i = c\mu_i$ ),  $\sigma_i^2 = c\mu_i$  ( $\sigma_i = \sqrt{\mu_i}$ )
    - $g(Y_{ij})$ 의 분산이  $\mu_i$ 에 영향을 받지 않게 하는 함수  $g(\cdot)$ 를 찾는 방법: 등분산성이 성립하도록
      - 함수  $g(Y_{ij})$ 를  $\mu_i$ 에 대한 1차 테일러전개

ex)  $\hat{Y}$ 가 커질수록 더 많이 퍼지는 경우

다) 1차 테일러전개

$$f(x) = f(x_0) + (x - x_0)f'(x_0)$$

$$g(Y_{ij}) \simeq g(\mu_i) + (Y_{ij} - \mu_i)g'(\mu_i)$$

-  $g(Y_{ij})$ 의 분산

$$\begin{aligned} \text{Var}[g(Y_{ij})] &\simeq \text{Var}[g(\mu_i) + (Y_{ij} - \mu_i)g'(\mu_i)] \quad \begin{array}{l} \text{상수} \\ \downarrow \end{array} \quad \because \text{Var}(X+c) = \text{Var}(X) \\ &= \text{Var}[(Y_{ij} - \mu_i)g'(\mu_i)] = (g'(\mu_i))^2 \text{Var}(Y_{ij} - \mu_i) \\ &= \{g'(\mu_i)\}^2 \text{Var}(Y_{ij}) = \{g'(\mu_i)\}^2 f(\mu_i) \end{aligned}$$

-  $g(Y_{ij})$ 의 분산이  $\mu_i$ 와 무관한 상수가 되려면  $c \simeq \{g'(\mu_i)\}^2 f(\mu_i)$   
constant

$$\Rightarrow g'(\mu_i)^2 \propto \frac{1}{f(\mu_i)} \quad \Rightarrow g'(\mu_i) \propto \frac{1}{\sqrt{f(\mu_i)}}$$

$$\Rightarrow \text{변환함수: } g(x) \propto \int \frac{1}{\sqrt{f(x)}} dx \quad \leftarrow \text{적분 (original 함수)}$$

①  $\sigma_i^2 = c\mu_i^2$  ( $\sigma_i = c\mu_i$ )  $\Rightarrow$  자연로그변환인  $\log(Y_{ij})$ 를 이용

②  $\sigma_i^2 = c\mu_i$  ( $\sigma_i = \sqrt{\mu_i}$ )  $\Rightarrow$  제곱근변환인  $\sqrt{Y_{ij}}$ 를 이용

$$\left( \begin{array}{l} \textcircled{1} \int \frac{1}{(x^2)^{1/2}} dx = \int \frac{1}{x} dx = \ln|x| + c \\ \textcircled{2} \int \frac{1}{(x)^{1/2}} dx = 2\sqrt{x} + c \end{array} \right.$$

내 데이터가 normal을 따르지 않는 경우 normal이 아닌 변수들을 normal type으로 바꿀

- Box-Cox transformation (1964)

- 최대가능도 추정에 의한 변환선택

원래지널 버전  $\Rightarrow g(x, \lambda) = \begin{cases} (x^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(x), & \lambda = 0. \end{cases}$

- Yeo-Johnson transformation (2000)

종더정교하게  $\Rightarrow g(x, \lambda) = \begin{cases} ((x+1)^\lambda - 1)/\lambda, & \lambda \neq 0, \quad x \geq 0 \\ \log(x+1), & \lambda = 0, \quad x \geq 0 \\ -((-x+1)^{2-\lambda} - 1)/(2-\lambda), & \lambda \neq 2, \quad x < 0 \\ -\log(-x+1), & \lambda = 2, \quad x < 0. \end{cases}$

- Modulus transformation (2000)

$$g(x, \lambda) = \begin{cases} \text{sign}(x) \frac{(|x| + 1)^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \text{sign}(x) \log(|x| + 1), & \lambda = 0. \end{cases}$$

$\text{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$

이전에 배웠던 선형모델의  $Y_{ij}$ 는 연속형  $\rightarrow$  보통 정수형 가정  
 하지만 더 넓은 범위의 분포를 따르는 경우가 많기 때문에 일반화

## ② 일반화선형모형(generalized linear models, GLM)

- $Y_{ij}$ 의 분포 - 지수족(exponential family)

- 정규분포, 이항분포, 음의 이항분포, 포아송분포, 감마분포(지수분포), ...

- 구조식

- $E(Y_{ij}) = \mu_i$

- $g(\mu_i) = x_{ij}^T \beta$  ←  $\beta$ 는 매개변수

link func. 'g' 이용  
 ( $\mu_i$ 를 직접 연결함수 X)

- $g$ : 연결함수(link function)

- logit link:  $\log(\mu_i / (1 - \mu_i))$ ,  $0 < \mu_i < 1$   $\rightarrow$   $Y$ 가 베르누이, 이항분포를 따를 때

- log link:  $\log(\mu_i)$ ,  $\mu_i > 0$   $\rightarrow$   $Y$ 가 포아송분포를 따를 때

- 최대가능도법을 이용하여 모수추정( $\beta$ )

\* 기본방식:  $\mu_i = x_{ij}^T \beta = x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + x_{ip}\beta_p \rightarrow (-\infty, \infty)$

① 만약 내 데이터가 베르누이분포를 따른다면  $(0, 1)$   
 $\rightarrow$  equality 성립을 위해 굉장히 많은 constraint 필요

$\rightarrow$  logit link 취하면  $(-\infty, \infty)$ 로 변경가능

② 포아송데이터는 보통 count에서 사용, 즉 0 이상 정수

$\rightarrow$  log link 취하면  $(-\infty, \infty)$ 로 범위변경

③ 비모수적 방법 : 가정에 덜 강한 방법  
대신 robust함 (∴ 정보↓)

• 자료의 값 대신 순위(rank)를 사용

- 자료를 정렬한 후 해당 자료의 순위를 구함
- tie가 있는 경우 순위의 중간값 사용

Y 대신  
rank R 사용

- $TSS = \sum_i \sum_j (R_{ij} - \bar{R}_{..})^2$
- $SSE = \sum_i \sum_j (R_{ij} - \bar{R}_{i.})^2$
- $SSTR = \sum_i n_i (\bar{R}_{i.} - \bar{R}_{..})^2$

- 검정통계량

$$F_0 = \frac{SSTR / (p-1)}{SSE / (N-p)} = \frac{MSTR}{MSE} \sim F_{p-1, N-p}$$

똑같이 계산!

(T도 사용가능, rank 이용 외의 차이점 없음)