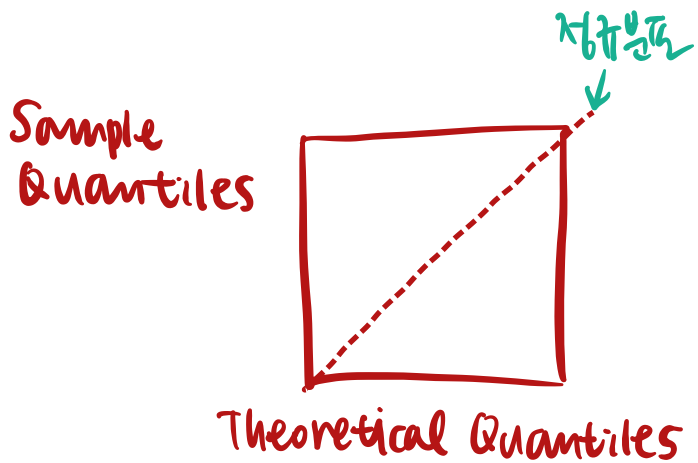
모든 분석에 적용되는 원리 - 가정을 만족하는지 확인 필요

■ 분산분석 검진(ANOVA Diagnostics)

- **기본가정**: $\boxed{\varepsilon_{ij} \sim \text{iid } N(0,\sigma^2)}$ ⇦ 잔차분석(residual analysis)
  - 등분산성
  - 독립성 ⇨ 잔차들 간에는 항상 상관관계가 존재
  - 정규성
  - 이상치 유무

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



- 수치적 방법

  실제계산

$Y_{ij} = \mu_i + \varepsilon_{ij}$, $E(Y_{ij}) = E(\mu_i + \varepsilon_{ij}) = \mu_i + E(\varepsilon_{ij})$
$= \mu_i + 0$

모두 이므로 추정 : $\hat{\mu}_i = \overline{Y}_{i\cdot}$

- 잔차

일반적인 ○ $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \overline{Y}_{i.}$ standard error

$Var(r_{ij}) = Var\left(\frac{e_{ij}}{\widehat{se}(e_{ij})}\right) = \frac{\sigma^2}{\widehat{se}(e_{ij})^2}$

변동성을 위해 ○ studentized 잔차 : $r_{ij} = \frac{e_{ij}}{\widehat{se}(e_{ij})}$ $\hat{\sigma}$

normalize 해줌

- $e_{ij} = \frac{n_i - 1}{n_i} Y_{ij} - \frac{1}{n_i} \sum_{k \neq j} Y_{ik}$ 둘은 독립

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

추정

- $\widehat{se}(e_{ij}) = \sqrt{\frac{(n_i - 1) MSE}{n_i}}$

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

$e_{ij} = Y_{ij} - \frac{1}{n_i} \sum_{k}^{n_i} Y_{ik} = Y_{ij} - \frac{1}{n_i}\left(Y_{ij} - \sum_{k \neq j} Y_{ik}\right)$

$$\text{Stu. } R_{11} = r_{11} = \frac{90-79.2}{\sqrt{\frac{10-1}{10} \times 223.49}} = 0.76$$

$$\text{Stu. Del. } R_{11} = t_{11} = 10.8 \times \left(\frac{40-4-1}{8049.4 \times (1-\frac{1}{10}) - 10.8^2} \right)^{\frac{1}{2}} = 0.76$$

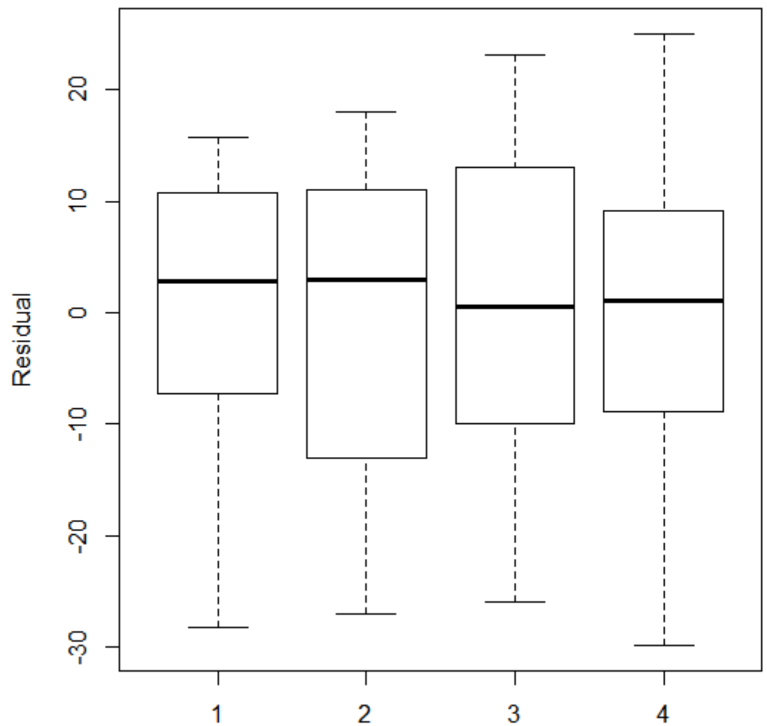
잔차보다 낫다. 감잡기 가능
(variability는 bound시켰으니)
→ 절댓값 2 넘으면
꽤 크다고 간주

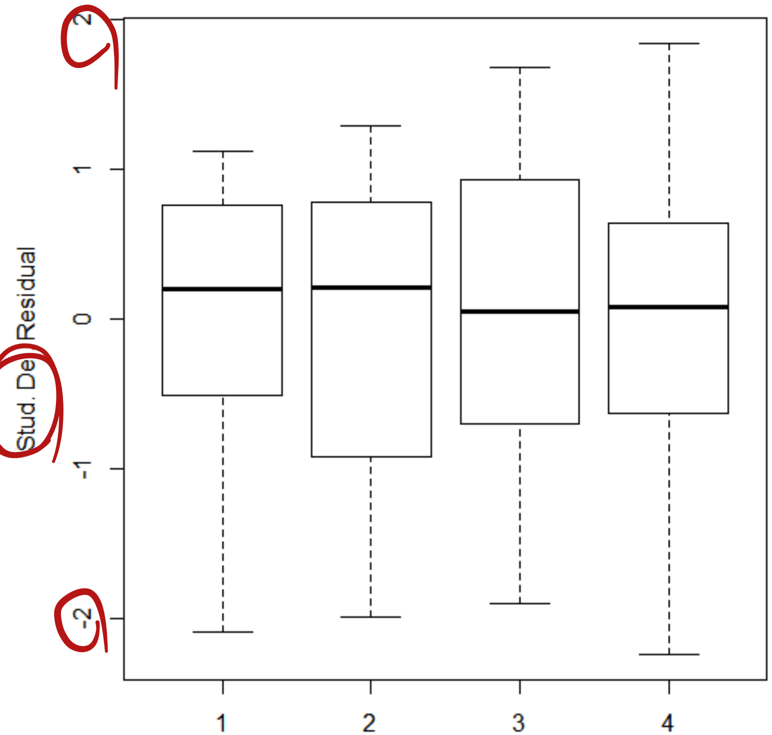
90−79.2

사료 (평균)	내용	1	2	3	4	5	6	7	8	9	10
1 (79.2)	관측값	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	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.

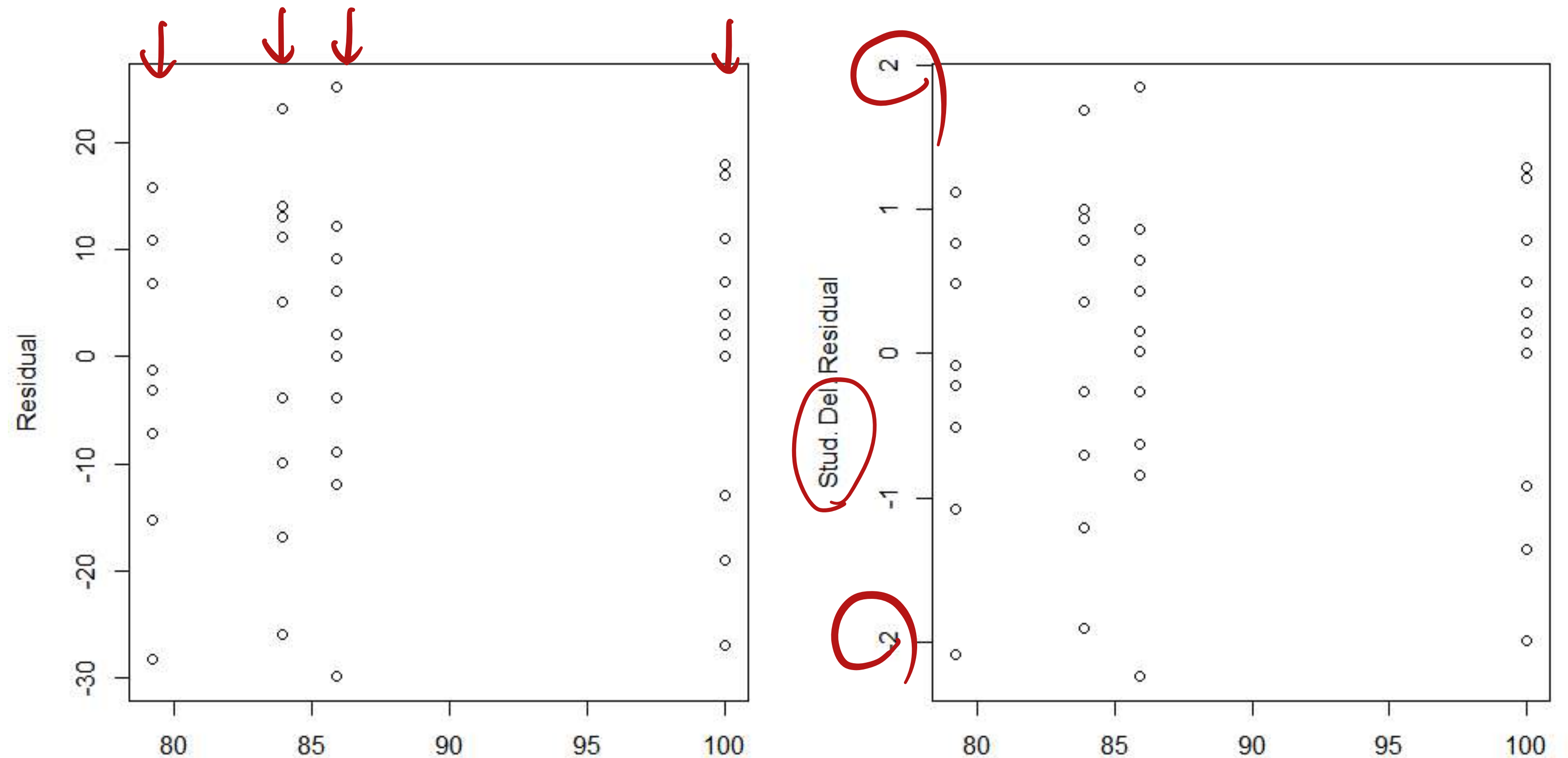
bound 돼있음.





등분산성 성립여부
이상점존재여부) graphically 확인가능

잔차 vs 정확한값



일원배치분산분석에서 $\hat{y} = \bar{Y}_{i.}$

→ 퍼진정도를 보고 등분산성, 이상점 판단가능

□ 등분산 검정

- 반복수가 같은 경우 동일한 분산을 가진다는 가정을 약간 어기는 경우 분산분석 방법은 robust함 둔감

- 반복수가 다르거나 어떤 한 분산이 다른 분산들보다 상당히 큰 경우 분산분석 방법은 robust하지 않음 ⇨ 분산들이 같은지 다른지를 검정필요

  ㄴ 민감하다 (결과값이 막 바뀐다)

- 가설: $H_0 \;:\; \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2$ VS H_1 : 최소한 하나 이상의 분산은 다름

○ Hartley 검정

- 동일 반복수 n
- 검정통계량 : $H^{*} = \dfrac{\max(S_i^2)}{\min(S_i^2)} \sim H(p, n-1)$

Hartley's Fmax Table 이용

  - ○ $S_i^2 = \sum (Y_{ij} - \overline{Y}_{i.})^2 / (n_i - 1)$
- 기각역 : $H^{*} > H(1-\alpha, p, n-1)$

$\dfrac{\max}{\min}$ 이 클수록 등분산성이 깨짐.

◉ 쥐 성장량

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

○ Brown-Forsythe 검정

짝수인경우 (2개)

$\frac{n\text{번째} + (n+1)\text{번째}}{2}$

- 절대편차를 먼저 계산

$$D_{ij} = \left| Y_{ij} - \widetilde{Y}_i \right|$$ = | observed 값 - 그룹의 중앙값 |

- ○ $\widetilde{Y}_i$: i 번째 그룹의 중앙값

D에 대해서

D에 대한

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

◉ 쥐 성장량

- ○ 중앙값: $\widetilde{Y}_{1.} = 82$, $\widetilde{Y}_{2.} = 103$, $\widetilde{Y}_{3.} = 84.5$, $\widetilde{Y}_{4.} = 87$
- ○ $\overline{D}_{1.} = 11$, $\overline{D}_{2.} = 11.4$, $\overline{D}_{3.} = 13.3$, $\overline{D}_{4.} = 10.9$, $\overline{D}_{..} = 11.65$ 각 집단의 평균
- ○ TSS=2804.6, SSTR=37.7 , SSE=2766.9
- ○ $MSTR^* = 12.567$, $MSE^* = 76.858$ ⇨ $F^*_{BF} = 0.164$ $< F_{0.05,3,36} = 2.866$

  = SSTR / 3 = SSE / 36

↳ H_0 기각불가, 즉 등분산 성립

○ Bartlett 검정

카이제곱분포이용 → 등분산검정

- $\boxed{\text{검정통계량 : } \chi_0^2 = 2.3026\dfrac{q}{c} \sim \chi_{p-1}^2}$

$$\left\{ \begin{aligned} &\checkmark \quad q = (N-p)\log_{10} MSE - \sum_{i=1}^{p}(n_i - 1)\log_{10} S_i^2 \\ &\checkmark \quad c = 1 + \frac{1}{3(p-1)}\left\{\sum_{i=1}^{p}\frac{1}{n_i - 1} - \frac{1}{N-p}\right\} \end{aligned} \right.$$

- Bartlett's 검정 통계량은 **정규 가정**에 매우 민감하기 때문에 정규 가정이 의심스러우면 사용할 수 없음

$$\chi_0^2 = 2.3026 \times \frac{0.0614}{1.0463}$$

◉ 쥐 성장량

- $q = 0.0614, \quad c = 1.0463 \quad \Rightarrow \quad \chi_0^2 = 0.135$ $< \chi^2_{3, 0.05} = 7.81$ → H_0 기각불가, 즉 등분산

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

□ 정규성 검정

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

$$JB = \frac{n}{6}\left(b_1 + \frac{1}{4}(b_2 - 3)^2\right) \sim \chi^2_2 \text{ (자유도 2)}$$

이 값이 크면 H_0 기각

→ 정규성 X

  - $\sqrt{b_1}$: 왜도(skewness)
  - b_2 : 첨도(kurtosis)

$$\begin{cases} \sqrt{b_1} = E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right) \\ b_2 = E\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) \end{cases}$$

□ 문제 발생 시 해결방안

① 변환(transformation)

ex) $\hat{y}$가 커질수록 더 많이 퍼진다던가

- 분산상수화변환(variance stablizing transformation, 분산안정화 변환)
  - ○ 잔차그림에서 잔차의 표준편차(분산)이 $\hat{Y}$의 값과 연관성을 보이는 경우
  - ○ 분산을 상수화시키기 위한 변환을 찾는 방법
    - $\sigma_i^2 = Var(Y_{ij})$ 와 $\mu_i = E(Y_{ij})$ 사이에 함수관계가 존재하는 경우:

처리에따라 달라진다는뜻

$$\sigma_i^2 = f(\mu_i)$$

    - 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 μ_i에 영향을 받지 않게 하는 함수 $g(\cdot)$를 찾는 방법

등분산성을 성립하도록

    - 함수 $g(Y_{ij})$를 μ_i에 대한 1차 테일러전개

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

테일러전개

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

- $g(Y_{ij})$의 분산

상수

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

여기서 $\sigma_i^2 = f(\mu_i)$

$$= \{g'(\mu_i)\}^2 \, Var(Y_{ij}) = \{g'(\mu_i)\}^2 f(\mu_i)$$

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

constant 해야한다

⇨ $g'(\mu_i)^2 \propto \dfrac{1}{f(\mu_i)}$ ⇨ $g'(\mu_i) \propto \dfrac{1}{\sqrt{f(\mu_i)}}$

적분

⇨ 변환함수: $g(x) \propto \displaystyle\int \frac{1}{\sqrt{f(x)}} dx$

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

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

① $\displaystyle\int \frac{1}{(x^2)^{1/2}} dx = \int \frac{1}{x} dx = \ln|x| + C$ | ② $\displaystyle\int \frac{1}{(x)^{1/2}} dx = 2\sqrt{x} + C$

내 데이터가 nomal을 따르지 않는 경우
nomal이 아닌 변수들을 nomal type으로 바꿈

- Box-Cox transformation (1964)
  - 최대가능도 추정에 의한 변환선택

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

  - Yeo-Johnson transformation (2000)

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

  - Modulus transformation (2000)

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

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

이전에 배웠던 선형 모델의 Yij는 연속형 -> 정규성 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

  기존) $\mu_i = x_{ij}^T \beta = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots x_{ip}\beta_p$ $(-\infty, \infty)$

  ← 모수 β
  - $E(Y_{ij}) = \mu_i$
  - $g(\mu_i) = x_{ij}^T \beta$ link function 'g' 이용
  - g: 연결함수(link function)
    - logit link: $\log(\mu_i/(1-\mu_i))$, $0 < \mu_i < 1$
    - log link: $\log(\mu_i)$, $\mu_i > 0$ (로그특성)
- 최대가능도법을 이용하여 모수추정

Y가 포아송분포를 따를때 많이 쓴다.

① 이때 만약 내 데이터가 베르누이분포를 따른다면 (0 ~ 1) 이므로 equality 성립을 위해 굉장히 많은 constraint (제한조건) 가 들어가야함

↳ log link ($\log(\mu_i)$) 를 사용하면 $(-\infty, \infty)$ 로 쉽게 변경가능

② 포아송데이터는 보통 count에서 사용, 즉 0이상정수 연결함수의 도움없이는 매우 불편

log link 이용시 $(-\infty, \infty)$ 로 범위 변경 용이

「비모수통계학」 과목에서 배울내용.

③ **비모수적 방법** : 가정에 둔감한 방법, 대신 robust함 (정보↓)

- 자료의 값 대신 순위(rank)를 사용
  - 자료를 정렬한 후 해당 자료의 순위를 구함
  - tie가 있는 경우 순위의 중간값 사용
  - $TSS = \sum_i \sum_j (R_{ij} - \overline{R}_{..})^2$
  - $SSE = \sum_i \sum_j (R_{ij} - \overline{R}_{i.})^2$
  - $SSTR = \sum_i n_i (\overline{R}_{i.} - \overline{R}_{..})^2$
  - 검정통계량

Y 대신 rank R로 사용

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

똑같이 계산. 차이는 rank 이용하는 것밖에 없음.