■ 분산분석 검진(ANOVA Diagnostics)

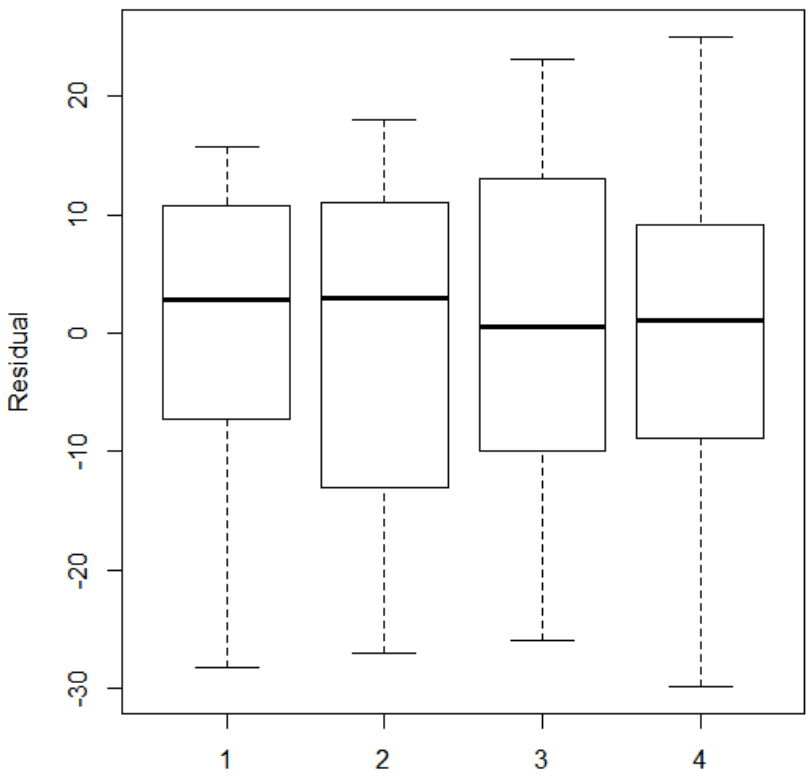
- **기본가정**: $\varepsilon_{ij} \sim$ iid $N(0,\sigma^2)$ ⇦ 잔차분석(residual analysis)
  - 등분산성
  - 독립성 ⇨ 잔차들 간에는 항상 상관관계가 존재
  - 정규성
  - 이상치 유무

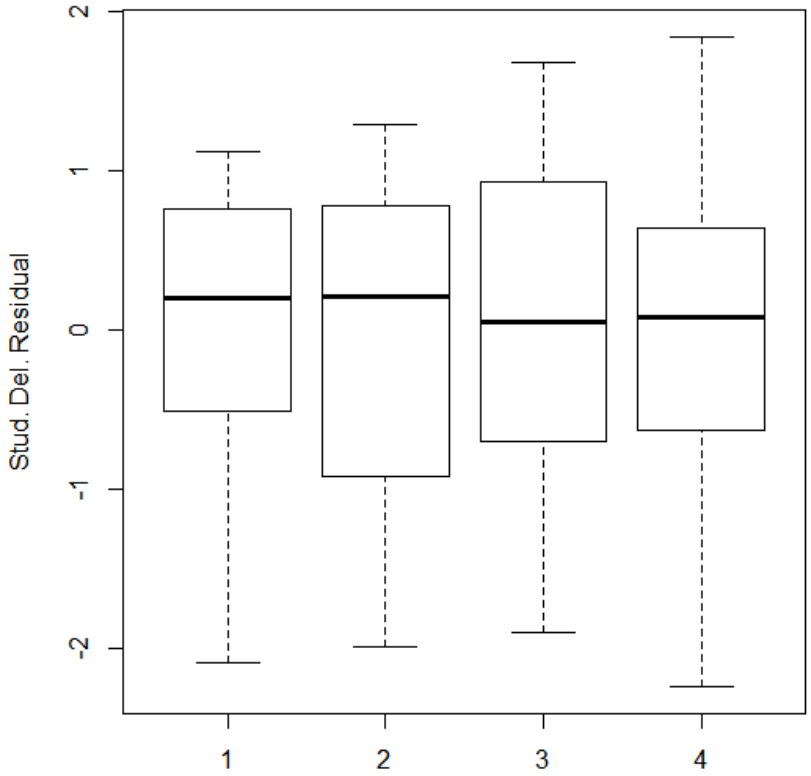
- 잔차그림
  - 잔차 vs 적합값
  - 정규확률그림
  - Box-plot

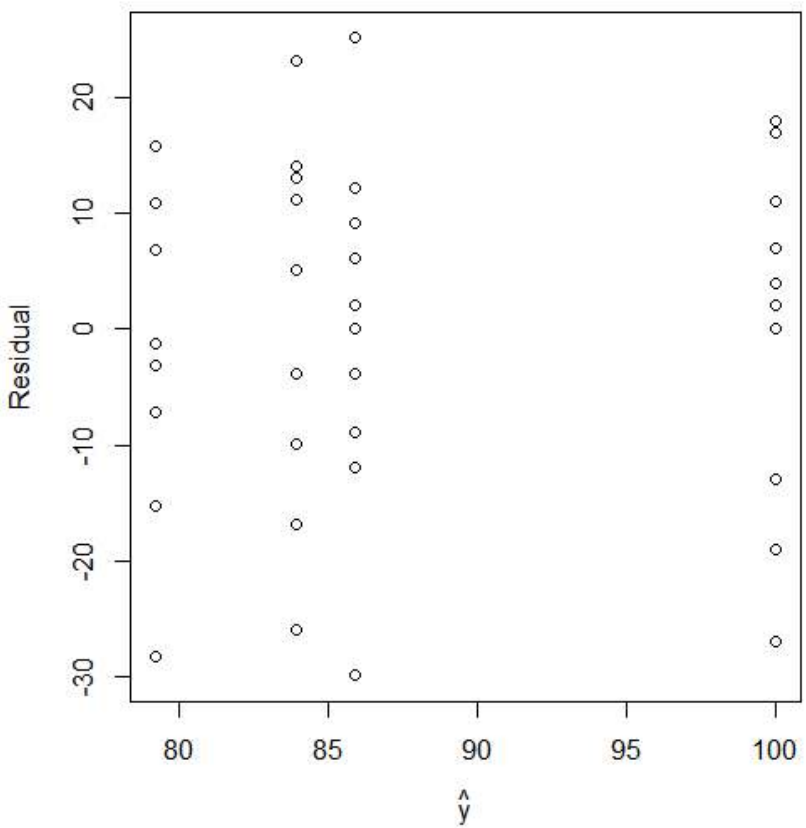
- 수치적 방법

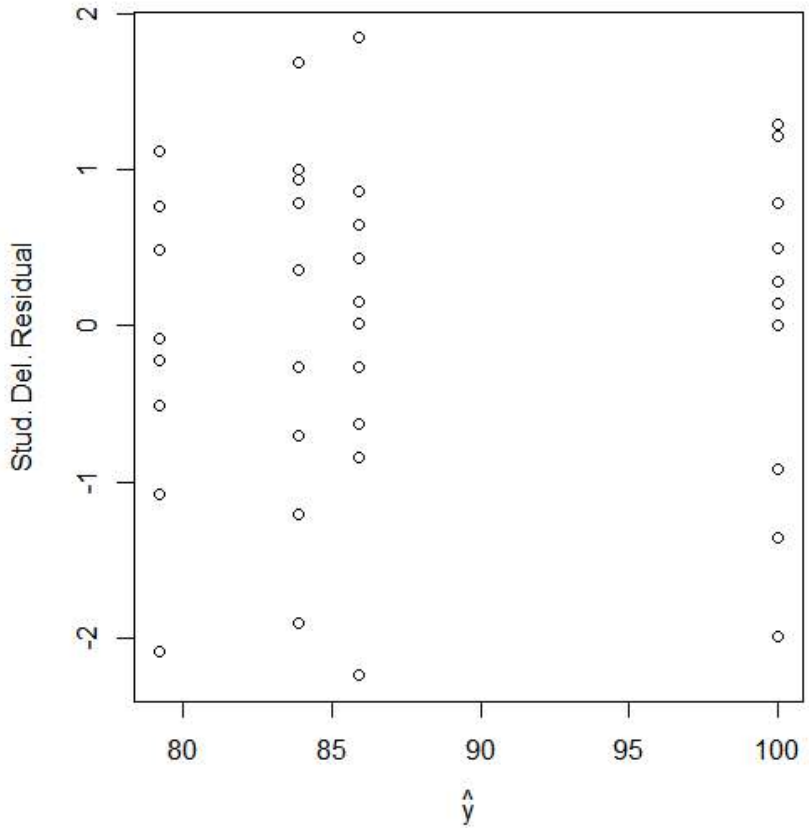
- 잔차
  - $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \overline{Y}_{i.}$
  - studentized 잔차 : $r_{ij} = \dfrac{e_{ij}}{\widehat{se(e_{ij})}}$
    - $e_{ij} = \dfrac{n_i - 1}{n_i} Y_{ij} - \dfrac{1}{n_i} \sum_{k \neq j} Y_{ik}$
    - $Var(e_{ij}) = \dfrac{(n_i - 1)^2}{n_i^2}\sigma^2 + \dfrac{n_i - 1}{n_i^2}\sigma^2 = \dfrac{n_i - 1}{n_i}\sigma^2$
    - $\widehat{se(e_{ij})} = \sqrt{\dfrac{(n_i - 1)MSE}{n_i}}$
  - studentized deleted 잔차 : $t_{ij} = e_{ij}\left[\dfrac{N - p - 1}{SSE(1 - 1/n_i) - e_{ij}^2}\right]^{1/2}$

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


Residual
20
10
0
-10
-20
-30
1
2
3
4


Stud. Del. Residual
2
1
0
-1
-2
1
2
3
4


Residual
20
10
0
-10
-20
-30
80
85
90
95
100
ŷ


Stud. Del. Residual
2
1
0
-1
-2
80
85
90
95
100
ŷ

□ 등분산 검정

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

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

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

◉ 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

- 절대편차를 먼저 계산

$$D_{ij} = \left| Y_{ij} - \tilde{Y}_i \right|$$

  - $\tilde{Y}_i$: i 번째 그룹의 중앙값

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

◉ 쥐 성장량

  - 중앙값: $\tilde{Y}_{i.} = 82$, $\tilde{Y}_{2.} = 103$, $\tilde{Y}_{3.} = 84.5$, $\tilde{Y}_{4.} = 87$
  - $\overline{D}_{1.} = 11$, $\overline{D}_{2.} = 11.4$, $\overline{D}_{3.} = 13.3$, $\overline{D}_{4.} = 10.9$, $\overline{D_{..}} = 11.65$
  - TSS=2804.6, SSTR=37.7 , SSE=2766.9
  - $MSTR^{*} = 12.567$, $MSE^{*} = 76.858$ ⇨ $F_{BF}^{*} = 0.164$

○ Bartlett 검정

- 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 가정**에 매우 민감하기 때문에 정규 가정이 의심스러우면 사용할 수 없음

◉ 쥐 성장량

- $q = 0.0614, \quad c = 1.0463 \quad \Rightarrow \chi_0^2 = 0.135$

□ 정규성 검정

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

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

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

□ 문제 발생 시 해결방안

① 변환(transformation)

- 분산상수화변환(variance stablizing transformation, 분산안정화 변화)
  - ○ 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:

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

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

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

- $g(Y_{ij})$의 분산

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

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

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

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

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

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

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

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

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

  - Yeo-Johnson transformation (2000)

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

  - Modulus transformation (2000)

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

② 일반화선형모형(generalized linear models, GLM)

- Y_{ij}의 분포 – 지수족(exponential family)
  - 정규분포, 이항분포, 음의 이항분포, 포아송분포, 감마분포(지수분포), ...
- 구조식
  - $E(Y_{ij}) = \mu_i$
  - $g(\mu_i) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}$
  - g: 연결함수(link function)
    - logit link: $\log(\mu_i/(1-\mu_i))$, $0 < \mu_i < 1$
    - log link: $\log(\mu_i)$, $\mu_i > 0$
- 최대가능도법을 이용하여 모수추정

③ 비모수적 방법

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

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