6장 | 모형의 진단

SAS를 이용한 실험 계획과 분산 분석 (자유아카데미)

에러의 전제

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

1. Normality (정규성)
2. Homoscedastic Variance (equal variance) (등분산성)
3. Independence (독립성)

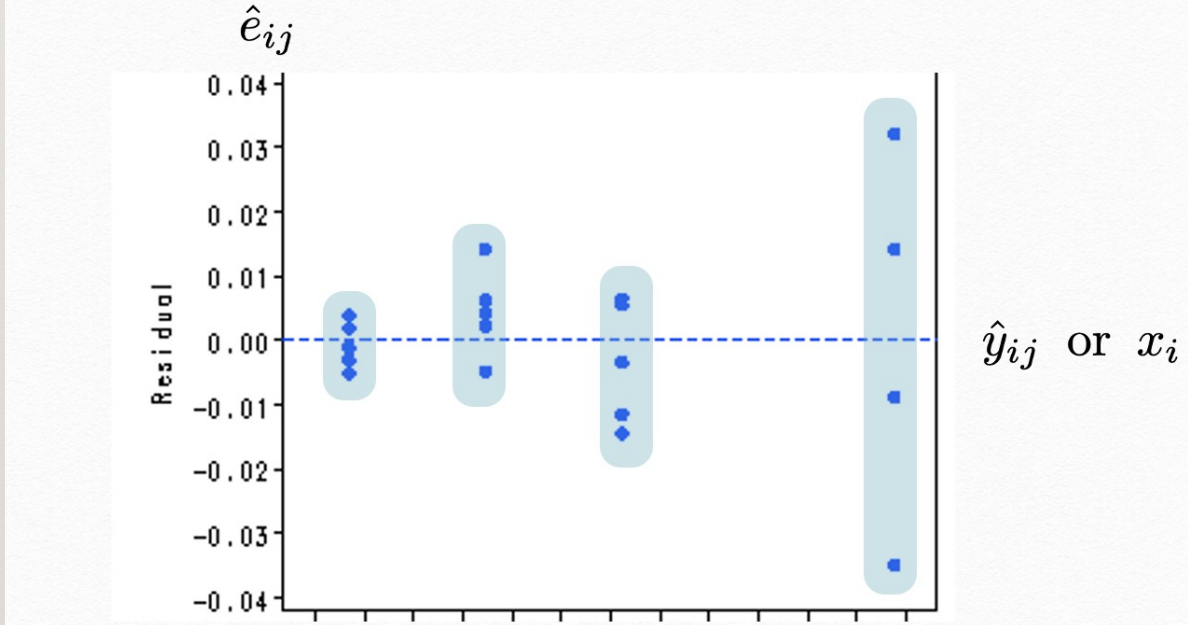
오차 → $e_{ij} =? = y_{ij} - (\mu + \tau_i)$

잔차 ← $\hat{e}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i = y_{ij} - \bar{y}_{i\cdot}$

residual

I. 등분산성 - 시각화

- 잔차산점도(residual plot)



Not equal variances !

I. 등분산성 - F 검정

두 모집단 분산에 대한 F 검정

귀무가설과 대립가설이 아래와 같을 때

$$H_0 : \quad \sigma_1^2 = \sigma_2^2, \qquad \text{vs.} \qquad H_1 : \quad \sigma_1^2 \neq \sigma_2^2, \tag{6.6}$$

각 그룹의 표본분산이 s_1^2, s_2^2 라면

$$F_0 = \frac{s_1^2}{s_2^2} \overset{H_0}{\sim} F_{n_1-1,\, n_2-1} \tag{6.7}$$

이 되어서, 만일

$$F_0 > F_{\frac{\alpha}{2},\, n_1-1,\, n_2-1} \quad \text{혹은} \quad F_0 < F_{1-\frac{\alpha}{2},\, n_1-1,\, n_2-1} \tag{6.8}$$

이면 '두 모집단의 분산은 서로 다르다'고 할 수 있다.

I. 등분산성 – F 검정 (예)

Q 표 1.1의 남녀별 시험성적의 분산이 서로 같다고 할 수 있는지를 F 검정으로 답하라(유의수준 0.05).

A 남자 5 명의 시험성적의 표본분산은 $s_1^2 = 360.001$ 이고 여자 5 명의 시험성적의 분산은 $s_2^2 = 321.499$ 이므로

$$F_0 = \frac{360.001}{321.499} = 1.119 \tag{6.9}$$

이므로

$$F_{0.975,\,4,\,4} = 0.104 < F_0 < 9.604 = F_{0.975,\,4,\,4} \tag{6.10}$$

가 되어 남녀 시험성적의 분산은 서로 같다고 할 수 있다.

I. 등분산성 – 바틀렛의 검정

귀무가설과 대립가설이 아래와 같을 때

$$H_0 : \quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2$$

$$H_1 : \quad \text{최소한 한 개의 } (i, j) \text{에 대하여} \quad \sigma_i^2 \neq \sigma_j^2, \quad (1 \leq i \neq j \leq a)$$

각 그룹의 표본분산이 $s_1^2, s_2^2, \cdots, s_a^2$ 이고

$$S_p^2 = \frac{\sum_{i=1}^{a}(n_i - 1)s_i^2}{\sum_{i=1}^{a}(n_i - 1)}$$

이면

$$\chi_0^2 = \frac{\sum_{i=1}^{a}(n_i - 1)\ln S_p^2 - \sum_{i=1}^{a}(n_i - 1)\ln s_i^2}{1 + \frac{1}{3(a-1)}\left(\sum_{i=1}^{a}\frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{a}(n_i - 1)}\right)} \overset{H_0}{\sim} \chi_{a-1}^2$$

이 되어서, 만일

$$\chi_0^2 > \chi_{\alpha,\, a-1}^2 \tag{6.12}$$

이면 a 개 모집단의 분산은 모두 같다고 할 수 없다.

바틀렛의 검정에서 주의할 점은 이 검정

통계량이 자료의 정규성 여부에

민감하다는 것이다. 따라서 관측치가

정규분포를 따르지 않는다면 바틀렛의

검정보다는 다른 방법을 선택할 것을

추천한다

I. 등분산성 – 바틀렛의 검정 (예)

다음 표는 4개의 영업교육 프로그램에 대한 판매 실적 자료이다. 각 프로그램별 판매 실적의 분산이 같은지/다른지 바틀렛의 검정을 사용하여 유의수준 0.05로 답하라.

교육 프로그램1	교육 프로그램2	교육 프로그램3	교육 프로그램4
74	94	62	80
67	82	75	82
83	69	59	75
77	78	79	90
71	68	68	72

	표본분산	합동표본분산
교육 프로그램1	36.796	
교육 프로그램2	113.188	$S_p^2 = 67.365$
교육 프로그램3	71.284	
교육 프로그램4	48.191	

표 6.1: 영업교육 프로그램별 분산과 합동표본분산

그 결과 $\chi_0^2 = 1.319 < 7.814 = \chi_{0.05,3}^2$ 이 되어서 교육 프로그램별 판매실적의 분산은 동일한 것을 알 수 있다.

SAS CODE

```
options ls=80 ps=65;

title1 'Diagnostics Example';

data one;
 infile 'c:\saswork\data\tensile.dat';
 input percent strength time;

proc glm data=one;
 class percent;
 model strength=percent;
 means percent / hovtest=bartlett hovtest=levene hovtest=bf;
 output out=diag p=pred r=res;

proc sort; by pred;
symbol1 v=circle i=sm50; title1 'Residual Plot';
proc gplot;  plot res*pred/frame; run;

proc univariate data=diag normal noprint;
 var res;  qqplot res / normal (L=1 mu=est sigma=est);
 histogram res / normal; run;
```

```
run;

proc sort; by time;
symbol1 v=circle i=sm75;
title1 'Plot of residuals vs time';
proc gplot; plot res*time / vref=0 vaxis=-6 to 6 by 1;
run;

symbol1 v=circle i=sm50;
title1 'Plot of residuals vs time';
proc gplot; plot res*time / vref=0 vaxis=-6 to 6 by 1;
run;
```

SAS OUTPUT

Diagnostics Example

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	475.7600000	118.9400000	14.76	<.0001
Error	20	161.2000000	8.0600000		
Corrected Total	24	636.9600000			

Bartlett's Test for Homogeneity of strength Variance

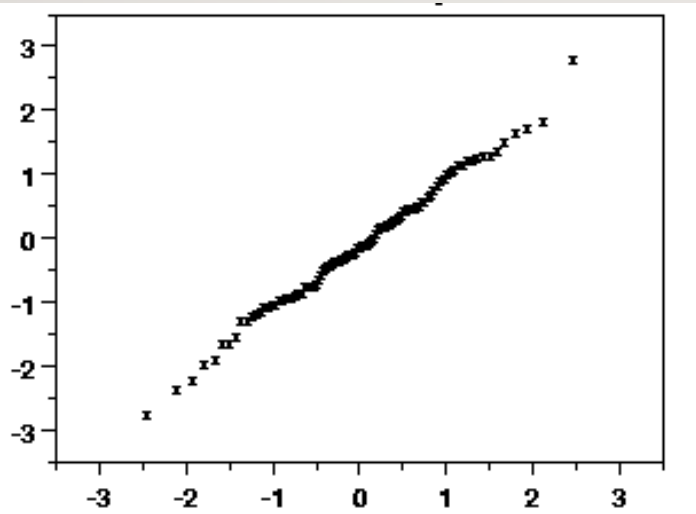
Source	DF	Chi-Square	Pr > ChiSq
percent	4	0.9331	0.9198

Levene's Test for Homogeneity of strength Variance
ANOVA of Squared Deviations from Group Means

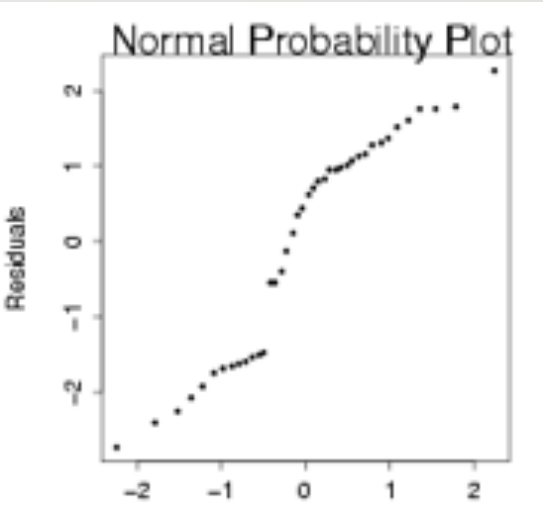
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
percent	4	91.6224	22.9056	0.45	0.7704
Error	20	1015.4	50.7720		

2. 잔차의 정규성

- residual 의 히스토그램 => normal ?
- normal probability plot (정규확률지) => 직선?



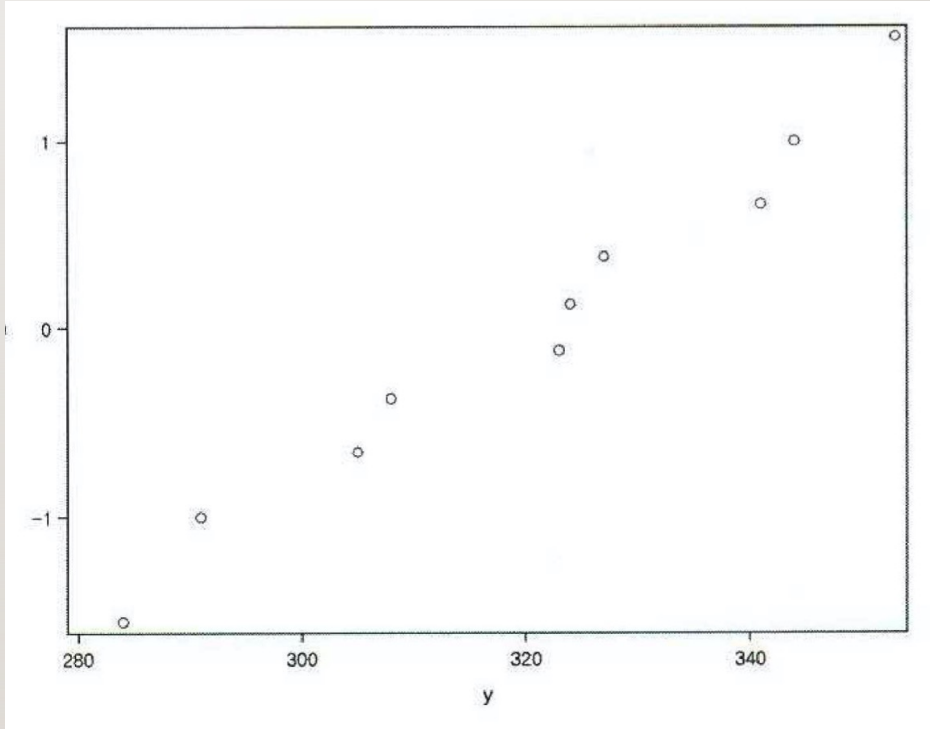
Normal



Non-normal

2. 잔차의 정규성

```
proc glm data=a;
   class x;
   model y = x;
   output out=aout r=res;
run;
```



2. 잔차의 정규성

- 샤피로-윌크 검정
- 앤더슨-달리 검정

- 콜모고로프-스미르노프 검정
- 크래머-폰미제스 검정

3. 잔차의 독립성

- 검정은 아직 불가능
- 자기 상관계수 (skip)나 그래프를 통해 가능

  - Plot $\hat{\epsilon}_{ij}$ vs time/space
  - Plot $\hat{\epsilon}_{ij}$ vs variable of interest

4. 해결

- Box-Cox 변환을 통해 정규성이나 등분산성, 혹은 독립성 문제를 해결 할 수 있다.

- Y ⟶ g(Y)

관측치의 기하평균이

$$\tilde{y} = \left(\Pi_{i=1}^{a}\Pi_{j=1}^{n_i} y_{ij}\right)^{\frac{1}{N}},$$

$$N = \sum_{i=1}^{a} n_i \tag{6.20}$$

라고 정의되면

$\lambda \neq 0$ 이면

$$y'_{ij} = \frac{y_{ij}^{\lambda} - 1}{\lambda \tilde{y}^{\lambda-1}} \tag{6.21}$$

$\lambda = 0$ 이면

$$y'_{ij} = \tilde{y} \ln y_{ij} \tag{6.22}$$

라고 하고 y'_{ij} 의 최소제곱추정량을 $\hat{y}'_{ij}$ 이라고 했을 때

$$\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} \left(y'_{ij} - \hat{y}'_{ij}\right)^2 \tag{6.23}$$

을 최소화시키는 $\hat{\lambda}$ 을 구하면 된다.

4. BOX-COX TRANSFORMATION

이렇게 얻은 $\hat{\lambda}$을 가지고 편의상

1) $\hat{\lambda} \neq 0$이면

$$y_{ij}^{*} = \frac{y_{ij}^{\hat{\lambda}} - 1}{\hat{\lambda}} \tag{6.24}$$

2) $\hat{\lambda} = 0$이면

$$y_{ij}^{*} = \ln y_{ij} \tag{6.25}$$

로 변환하면 된다.

Remark 박스-칵스 변환으로 구한 $\hat{\lambda}$는 정규성-등분산성-독립성 가정하에 SSE를 최소화하는 값이지 이렇게 구한 $\hat{\lambda}$가 반드시 이 세 가지 가정을 만족시킨다는 보장은 없다. 따라서 박스-칵스 변환한 자료를 가지고 분산분석을 한 후에도 반드시 다시 잔차를 살펴봐야 한다.

예

Example Heart-Lung 펌프의 회전속도(speed)에 따른 혈액양(blood) 자료에서 잔차가 정규성-등분산성-독립성을 갖기 위해 어떤 박스-칵스 변환이 필요한지를 살펴보고 적당한 변환을 해 보자.

```
proc transreg details data=a;
model boxcox(blood / convenient lambda=-2 to 2 by 0.01)
    = class(speed) ; run;
```

그 결과 그림 6.6과 같은 로그-가능도함수(log-likelihood function)를 최대로 하는 $\hat{\lambda}$은 0이 되어서 이 자료를 일원배치법으로 분석할 경우는 로그변환($\ln y$)을 추천한다.