변량 모형

서울시립대 통계학과

2021-03-22

차 례

서문

이번 강의에서는 변량모형(random effect models)에 대하여 알아봅니다.

이번 강의에 필요한 R 패키지는 다음과 같습니다.

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(lme4)
```

제 1 장

변량 모형

제 1 절 고정효과

앞 장에서 하나의 요인있는 일원배치 모형에 대한 추론에 대하여 알아보았다.

$$x_{ij} = \mu + \alpha_i + e_{ij} \tag{1.1}$$

여기서 오차항 e_{ij}는 모두 독립이며 $N(0, \sigma_E^2)$를 따른다.

일원배치 모형 (1.1) 에서 전체 평균 μ 와 처리수준의 효과를 나타내는 $\alpha_1, \alpha_2, \ldots, \alpha_a$는 모두 고정된 값을 가지는 모수(parameter)이다. 식 (1.1)의 오른쪽 항들 중에서 확률변수는 오차항 e_{ij}이 유일하다.

처리수준의 효과 α_i들이 모수이라는 것은 의미는 만약 새로운 실험에서 동일한 실험단위(experiment unit)에 동일한 처리를 적용하면 평균 처리 효과는 α_i로 일정하다는 의미이다.

예를 들어 예제 3.1에서 수행한 실험을 다른 회사에서 동일한 납품업체의 원단(동일한 실험 단위와 처리)을 가지고 새로운 실험을 하면 평균적인 효과는 예제 3.1과 동일하다는 가정을 할 수 있다. 또한 예제 4.1 에 대한 실험에서도 만약 동일한 돼지 품종과 사료를 사용하여 새로운 실험을 수행할 때 처리 효과는 원래 실험과 같다고 가정할 수 있다. 즉, 처리라는 것이 기술적인 의미를 지니고 있어 반복하여 재현할 수 있는 효과이다. 이러한 고정된 모수로서의 효과를 **고정 효과(fixed effect)**라고

부른다.

더 나아가 고정효과를 가지는 모형에서는 고정효과를 추정하고 처리 수준간의 차이가 있는지 추론하는 것이 주 목적이다.

제 2 절 임의효과

이제 고정효과와는 다른 의미를 가지는 몇 가지 실험들을 생각해 보자.

예제 1.1 (화학약품 회사:교과서예제 3.3)**.** 화학약품 회사에서는 매년 원자재의 수백 개의 배치(batch)를 정제하여 순도가 높은 화학약품을 만든다. 품질 관리를 위하여 수백 개의 배치들 중에서 5개를 랜덤하게 선택하고 배치당 3개의 시료를 채취한 후에 순도를 측정하였다.

배치마다 순도가 크게 다르면 품질을 일정하게 유지할 수 없눈 문제가 생긴다. 따라서 실험의 목적은 품질 관리이며 배치 간의 변동과 배치 내의 변동을 알아보는 것이다.

예제 1.2 (학교간의 성적 비교)**.** 학교 간에 성적의 차이를 알아보기 위하여 서울에 있는 603개의 학교에서 20개의 학교를을 임의로 추출하고 추출된 학교에 속한 6학년 학생들 10명을 임의로 추출하여 과학시험을 보게 하여 점수를 얻었다.

이러한 자료에서 학생들의 성적은 가장 점수가 낮은 학생부터 매우 우수한 성적을 낸 학생까지 점수의 변동(variation)이 존재한다. 변동의 요인은 무었일까? 학생의 개인의 차이(예:학생의 지능, 노력 정도, 학습 환경)도 변동의 요인이지만 또한 학교의 차이(예: 교사, 거주 환경)도 변동의 요인이 될 수 있다.

예제 1.3 (Test-ReTest)**.** 새로 개발된 CT 로 만든 영상에 근거하여 의사들이 암의 단계를 점수로 파악하는 방법이 제안되었다. 제안된 방법의 유료성과 안정성을 알아보기 위하여 실험을 진행하였다. 일단 5명의 암환자들에서 CT 영상을 촬영하였다. 다음으로 15명의 의사를 임의로 추출하고 5명의 CT 영상을 본 후 암의 진행단계를 판단할 수 있는 점수를 매기도록 하였다.

실험의 목적은 CT 영상에 근거한 진단이 의사들간에 잘 일치하는지를 알아보는 실험이다. 이 실험에서는 의사와 환자라는 두 가지 요인이 존재한다.

위의 예제에서 **배치, 의사, 학교**는 고정 효과를 가정한 실험에서 고려하는 요인과는 성격이 틀리다. 5개의 배치들은 수백 개의 배치들에서 임의로 추출 되었으며 5명의 의사들은 다수의 의사들 중 임의로 추출되었다. 603개 초등 학교의 모집단에서 20개의 학교가 임의로 추출되었다.

배치, 의사 또는 학교 간의 차이는 잘 설계된 실험의 처리에 대한 고정 효과와는 다르다. 동일한 배치, 학교 또는 의사으로 부터 나온 관측값들은 동일한 처릴 받은 값들이라기 보다는 동일한 **집단(group, cluster)**에서 나온 관측값으로 볼 수 있다. 위의 예제들에서는 식 (1.2) 에서 효과 α_i 의 변동은 모집단을 구성하는 집단들의 변동이라고 할 수 있다.

위에서 언급한 3개 예제는 실험의 목적이 선택된 수준들의 효과의 기술적인 비교가 아니라 모집단이 가지고 있는 여러 가지 변동(variance)에 대하여 추론하는 것이다.

같은 학교에 다니는 학생들은 주거 환경, 교사 등 공통적인 요인에 의하여 영향을 받는다고 가정할 수 있다. 따라서 같은 학교에 다는 학생들의 성적이 독립이 아닐 수도 있다. 동일한 의사가 판단한 5명의 환자에 의 평가 점수들도 독립이라고 가정하기 어렵다.

고정효과처럼 기술적인 처리효과가 아니라 모집단의 구성 단위들의 변동을 기술하는 효과를 **임의효과(random effect, 변량)** 라고 한다. 임의효과를 가진 일원배치 모형을 **변량모형**(random models) 또는 임의효과 모형(random effect models)이라고 부르며 다음과 같이 나타낼 수 있다.

$$x_{ij} = \mu + \alpha_i + e_{ij} \quad \text{where } \alpha_i \sim N(0, \sigma_A^2), \;\; e_{ij} \sim N(0, \sigma_E^2) \tag{1.2}$$

위의 식에서 $\alpha_1, \alpha_2, \ldots, \alpha_a$를 임의 효과라고 부르며 서로 독립인 확률 변수로서 분포는 $N(0, \sigma_A^2)$을 따른다. 또한 임의 효과 α_i와 오차항 e_{ij}은 서로 독립이다.

임의효과가 가지는 분산을 σ_A^2을 분산성분(variance component)라고 하며 집단 간의 변동을 의미한다. σ_A^2이 크면 모집단을 구성하고 있는 단위들의 변동이 크다고

할 수 있다. 반면 σ_A^2이 작으면 단위들간의 변동이 작아진다.

제 3 절 변량모형의 성질

3.1 총변동의 분해

일원배치 변량 모형 (1.2)을 따르는는 반응변수 x_{ij}의 평균과 분산은 다음과 같다.

$$\begin{aligned} E(x_{ij}) &= E(\mu + \alpha_i + e_{ij}) \\ &= E(\mu) + E(\alpha_i) + E(e_{ij}) \\ &= \mu + 0 + 0 \\ &= \mu \end{aligned}$$

$$\begin{aligned} V(x_{ij}) &= Var(\mu + \alpha_i + e_{ij}) && (1.3) \\ &= V(\alpha_i) + V(e_{ij}) && (1.4) \\ &= \sigma_A^2 + \sigma_E^2 && (1.5) \end{aligned}$$

식 (1.5)에서 나타난 분해는 다음과 같이 의미로 표현할 수 있다.

$$\underbrace{V(x_{ij})}_{\text{total variation}} = \underbrace{\sigma_A^2}_{\text{variation between groups}} + \underbrace{\sigma_E^2}_{\text{variation within group}}$$

3.2 관측값의 종속성

식 (1.2) 로 표현된 변량모형의 가장 큰 특징 중에 하나는 같은 집단에 속하는 관측치들은 서로 독립이 아니며 양의 상관관계가 있는 것이다. 예를 들어 위의 학교간의 성적 비교 예제에서 두 학생 x_{ij}와 x_{ik}이 같은 학교 i에 속한다면

$$\begin{aligned} Cov(x_{ij}, x_{ik}) &= Cov(\mu + \alpha_i + e_{ij}, \mu + \alpha_i + e_{ik}) \\ &= Cov(\alpha_i, \alpha_i) + Cov(\alpha_i, e_{ik}) + Cov(e_{ij}, \alpha_i) + Cov(e_{ij}, e_{ik}) \\ &= Cov(\alpha_i, \alpha_i) + 0 + 0 + 0 \\ &= V(\alpha_i, \alpha_i) \\ &= \sigma_A^2 \end{aligned}$$

따라서

$$\begin{aligned} corr(x_{ij}, x_{ik}) &= \frac{Cov(x_{ij}, x_{ik})}{\sqrt{V(x_{ij})\ V(x_{ik})}} \\ &= \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} \\ &= \rho \end{aligned}$$

위의 상관계수(교과서애서 기여율)는 보통 급내 **상관계수(Intra Class Correlation, ICC)**라고 부른다. 그룹 변동의 크기를 나타내는 분산성분 σ_A^2가 그룹 내 변동을 나타내는 오차항의 분산 σ_E^2보다 상대적으로 클수록 급내 상관계수가 1에 가까와진다.

보통 σ_A^2을 집단간 변동(between-group variance)라 하고 σ_E^2를 집단내 변동(within-group variance)라고 한다. 따라서 σ_A^2와 σ_E^2의 상대적인 크기의 차이에 따라 그룹내 관측값의 상관관계가 달라진다.

3.3 제곱합의 기대값

일원배치 변량 모형 (1.2)은 고정효과 모형 (1.1)과 동일한 분산분석(ANOVA) 표를 사용한다. 분산분석 표의 제곱합을 이용하여 σ_A^2와 σ_E^2에 대한 추정량을 얻을 수 있다.

첫 째로 분산분석 표에서 SS_E의 기대값을 구해보자.

먼저 다음과 같은 분해를 고려하자.

$$\begin{aligned} x_{ij} - \bar{x}_{i.} &= (\mu + \alpha_i + e_{ij}) - \frac{\sum_{j=1}^{r}(\mu + \alpha_i + e_{ij})}{r} \\ &= (\mu + \alpha_i + e_{ij}) - \left(\mu + \alpha_i + \frac{\sum_{j=1}^{r} e_{ij}}{r}\right) \\ &= (\mu + \alpha_i + e_{ij}) - (\mu + \alpha_i + \bar{e}_{i.}) \\ &= e_{ij} - \bar{e}_i \end{aligned}$$

이므로 오차제곱합 SS_E의 기대값은 다음과 같이 구해진다.

$$\begin{aligned} E\left[\sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i.})^2\right] &= E\left[\sum_{i=1}^{a}\sum_{j=1}^{r}(e_{ij} - \bar{e}_{i.}))^2\right] \\ &= (r-1)\sum_{i=1}^{a} E\left[\frac{\sum_{j=1}^{r}(e_{ij} - \bar{e}_{i.}))^2}{r-1}\right] \\ &= (r-1)\sum_{i=1}^{a}\sigma_E^2 \\ &= a(r-1)\sigma_E^2 \end{aligned}$$

또한 SS_A의 기대값을 구하기 위하여

$$\begin{aligned} \bar{x}_{i.} - \bar{\bar{x}} &= (\mu + \alpha_i + \bar{e}_{i.}) - (\mu + \bar{\alpha} + \bar{\bar{e}}) \\ &= (\alpha_i - \bar{\alpha}) + (\bar{e}_{i.} - \bar{\bar{e}}) \end{aligned}$$

이므로 SS_A의 기대값은 다음과 같이 구해진다.

$$
\begin{aligned}
E\left[\sum_{i=1}^{a}\sum_{j=1}^{r}(\bar{x}_{i.}-\bar{\bar{x}})^2\right] &= E\left[\sum_{i=1}^{a}\sum_{j=1}^{r}\{(\alpha_i-\bar{\alpha})+(\bar{e}_{i.}-\bar{\bar{e}})\}^2\right] \\
&= \sum_{i=1}^{a}\sum_{j=1}^{r}E[(\alpha_i-\bar{\alpha})^2]+\sum_{i=1}^{a}\sum_{j=1}^{r}E[(\bar{e}_{i.}-\bar{\bar{e}})^2] \\
&= r(a-1)E\left[\frac{\sum_{i=1}^{a}(\alpha_i-\bar{\alpha})^2}{a-1}\right]+r(a-1)E\left[\frac{\sum_{i=1}^{a}(\bar{e}_{i.}-\bar{\bar{e}})^2}{a-1}\right] \\
&= r(a-1)\sigma_A^2+r(a-1)\frac{\sigma_E^2}{r} \\
&= (a-1)(r\sigma_A^2+\sigma_E^2)
\end{aligned}
$$

위의 계산에서 이용한 사실은 α_i는 서로 독립으로 $N(0,\sigma_A^2)$를 따르고 $\bar{e}_{i.}$는 서로 독립으로 $N(0,\sigma_E^2/r)$를 따른다는 것이다.

위의 제곱합의 기대값을 정리해보면 다음과 같은 두 방병식을 얻는다.

$$
E(SS_A)=(a-1)(r\sigma_A^2+\sigma_E^2),\quad E(SS_E)=a(r-1)\sigma_E^2 \tag{1.6}
$$

위의 모수 방정식에 적률추정법(methods of moment)을 적용하면 다음과 같은 방정식을 얻고

$$
SS_A=(a-1)(r\hat{\sigma}_A^2+\hat{\sigma}_E^2),\quad SS_E=a(r-1)\hat{\sigma}_E^2 \tag{1.7}
$$

위의 방정식을 풀면 σ_A^2와 σ_E^2의 불편 추정량을 구할 수 있다. 여기서 유의할 사항은 σ_A^2에 대한 추정량은 0보다 작은 값이 나올 수 있으므로 이럴 경우 0으로 지정한다.

$$
\begin{aligned}
s_E^2=\hat{\sigma}_E^2&=\frac{SS_E}{a(r-1)}=MS_E \\
s_A^2=\hat{\sigma}_A^2&=\max\left[0,\frac{SS_A/(a-1)-\hat{\sigma}_E^2}{r}\right]=\max\left[0,\frac{MS_A-MS_E}{r}\right]
\end{aligned}
$$

제 4 절 가설 검정

고정효과 모형에서 요인 A의 수준간에 차이가 있는 지를 검정하는 경우 귀무가설은 $H_0 : \alpha_1 = \cdots = \alpha_a = 0$ 이었다. 이제 변량 모형에서는 집단 간의 변동이 없는지 검정하는 것이므로 다음과 같은 가설을 고려한다.

$$H_0 : \sigma_A^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_A^2 > 0 \tag{1.8}$$

분산성분 σ_A^2가 0 이라는 의미는 모든 α_i 가 0이고 이는 집단 간의 차이가 없는 상황을 의미한다. 위의 **가설을 검정하는 방법은 고정효과 모형과 동일하다**. 즉 다음과 같은 조건이 만족되면 귀무가설을 기각한다.

$$\text{reject } H_0 \text{ if } F_0 = \frac{MS_A}{MS_E} > F[1-\alpha, a-1, a(r-1)]$$

제 2 장

예제 3.3

교과서 59 페이지에 있는 예제를 R 프호그램을 사용하여 풀어보자.

화학약품 회사에서는 매년 원자재의 수백 개의 배치(batch)를 정제하여 순도가 높은 화학약품을 만든다. 품질 관리를 위하여 수백 개의 배치들 중에서 5개를 랜덤하게 선택하고 배치당 3개의 시료를 채취한 후에 순도를 측정하였다.

배치마다 순도가 크게 다르면 품질을 일정하게 유지할 수 없눈 문제가 생긴다. 따라서 실험의 목적은 품질 관리이며 배치 간의 변동과 배치 내의 변동을 알아보는 것이다.

제 1 절 자료

다음과 같이 자료를 만들자

```
response <- c( 74, 76, 75,
               68, 71, 72,
               75, 77, 77,
               72, 74, 73,
               79, 81, 79)
batch <- factor(rep(1:5, each=3))
df <- data.frame(batch, response)
df
```

```
##    batch response
## 1      1       74
## 2      1       76
## 3      1       75
## 4      2       68
## 5      2       71
## 6      2       72
## 7      3       75
## 8      3       77
## 9      3       77
## 10     4       72
## 11     4       74
## 12     4       73
## 13     5       79
## 14     5       81
## 15     5       79
```

제 2 절 추정과 가설검정

변량모형을 적합시키기 위해서는 lme4 패키지 가 필요하다. 일원배치 변량모형을 적합시키는 함수는 lmer이며 다음과 같이 사용한다. 아래 모형식에서 1은 평균 μ를 나타내고 (1|batch)는 배치에 대한 임의 효과 α_i를 나타낸다.

```
res <- lmer(response ~ 1 + (1|batch), data=df )
summary(res)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: response ~ 1 + (1 | batch)
##    Data: df
##
## REML criterion at convergence: 62.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.90384 -0.53153  0.00484  0.61386  1.16817
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  batch    (Intercept) 11.71    3.422
##  Residual              1.80    1.342
## Number of obs: 15, groups:  batch, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   74.867      1.569   47.71
```

위의 결과에서 다음과 같은 추정값을 얻는다.

- $\hat{\mu} = 74.867$
- $\hat{\sigma}_A^2 = 11.71$
- $\hat{\sigma}_E^2 = 1.80$

따라서 급내 상관계수(기여율)의 추정값은 다음과 같다.

$$\hat{\rho} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_E^2} = \frac{11.7}{11.7 + 1.8} = 0.867$$

위의 $\hat{\rho} = 0.867$을 기여율로 해석하면 총변동 중에서 배치 간의 변동이 차지하는 비율이 86.7% 이라는 것이다.

또한 $H_0 : \sigma_A^2 = 0$ 에 대한 검정은 다음과 같이 `aov`함수를 사용하여 수행할 수 있다.

```
summary(aov(response ~ batch, data=df ))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## batch        4  147.7   36.93   20.52 8.25e-05 ***
## Residuals   10   18.0    1.80
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-값이 유의수준 5% 보다 매우 작으므로 H_0를 기각한다. 배치 간 변동이 유의하

다고 할 수 있다. 따라서 품질이 배치 간에 따라서 크게 다르다.