계층모형 실습

이용희

2021년 5월 11일

차 례

서문

이 책은 계층모형(hierarchical models)에 대한 예제를 통한 실습, 모형을 적합하는 계산 방법에 대하여 다루고자 합니다.

이 책에서 사용된 기호, 표기법, 프로그램의 규칙과 쓰임은 다음과 같습니다.

- 스칼라(scalar)와 일변량 확률변수는 일반적으로 보통 글씨체의 소문자로 표기한다. 특별한 이유가 있는 경우 대문자로 표시할 것이다.
- 벡터, 행렬, 다변량 확률벡터는 굵은 글씨체로 표기한다.

- 통계 프로그램은 R을 이용하였다. 각 예제에 사용된 R 프로그램은 코드 상자를 열면 나타난다.
- 통계 프로그램은 R에 대한 기초는 저자의 홈페이지에 있는 안내 사이트[1]에서 먼저 학습할 것을 권장한다.

다음은 이 책에서 R 실습을 하기 위하여 필요한 패키지이다.

```
library(tidyr)
library(dplyr)
library(ggplot2)
library(knitr)
library(kableExtra)
library(lme4)
library(lmerTest)
library(brms)
library(sas7bdat)
```

```
library(sampling)
library(stringr)
library(survey)
library(sae)
```

제 1 장

반복측정자료

1.1 개요

반복측정자료(longitudinal data, repeated measurements)는 관측단위안에서 여러 개의 관측값을 측정한 자료의 형식을 말한다.

예를 들어 환자가 병원을 여러 번 방문할 때마다 혈압을 측정하였다면 한 명의 환자에서 반복 측정한 자료는 서로 독립이 아니다. 또한 가구조사(household survey)에서 가구원의 경제덕 상태, 건경 상태 등을 여러 해동안 매년 측정하는 경우가 있는데 이러한 자료를 패널자료(panel data) 또는 longitudinal 자료라고 한다. 이렇게 하나의 관측단위 안에서 반복 측정한 자료들은 서로 독립이 아닌 특징이 있고 자료를 분석하는 경우 이러한 자료들의 종속구조를 고려하는 모형을 사용하는 것이 적절하다.

이렇게 반복측정자료에서 반복자료들의 공분산구조를 설정하는 통계적 방법들은 다양하지만 대표적으로 쉽게 사용할 수 있는 방법이 임의효과를 포함한 혼합모형을 사용하는 방법이다.

1.2 `sleepstudy` 자료

`lme4` 패키지에 자료인 `spleepstudy`는 화물트럭 운전사들에 대한 수면부족 현상에 대하여 연구한 자료이다. 18명의 운전자들이 매일 3시간의 수면(부족한 수면)을 하면서 매일 일정한 동작에 대한 반응시간을 10일 동안 반복적으로 측정한 자료이다. 한명의 운전사에게

10일 동안의 반응에 대한 측정자료 10개가 존재하므로 이는 반복측정 자료이며 이러한 10개의 자료는 독립이 아니다.

혼합모형을 적합할 수 있는 R 패키지는 여러 가지 다양한 패키지가 있지만 최대가능도 추정법에 기반한 lme4 패키지를 먼저 이용할 것이다.

일단 자료의 구조를 살펴보자. 반응변수 Reaction 은 반응시간(ms)를 나타내며 설명변수로서 Days 는 날짜($t = 0, 1, 2, \ldots, 9$), Subject 는 운전자의 고유번호를 나타낸다.

```
str(sleepstudy)
```

```
## 'data.frame':    180 obs. of  3 variables:
##  $ Reaction: num  250 259 251 321 357 ...
##  $ Days    : num  0 1 2 3 4 5 6 7 8 9 ...
##  $ Subject : Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(sleepstudy,n=20)
```

```
##    Reaction Days Subject
## 1  249.5600    0     308
## 2  258.7047    1     308
## 3  250.8006    2     308
## 4  321.4398    3     308
## 5  356.8519    4     308
## 6  414.6901    5     308
## 7  382.2038    6     308
## 8  290.1486    7     308
## 9  430.5853    8     308
## 10 466.3535    9     308
## 11 222.7339    0     309
## 12 205.2658    1     309
## 13 202.9778    2     309
## 14 204.7070    3     309
## 15 207.7161    4     309
## 16 215.9618    5     309
```

```
## 17 213.6303     6      309
## 18 217.7272     7      309
## 19 224.2957     8      309
## 20 237.3142     9      309
```

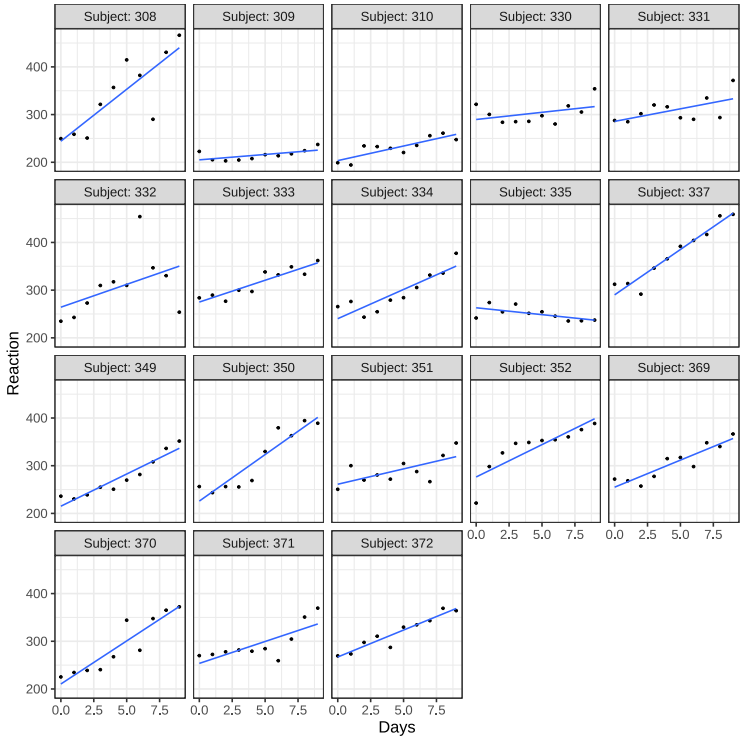
1.3 개별단위 분석 (no pooling)

각 운전자에 대한 10일 간의 반응속도가 시간에 따라 어떻게 변하는 가를 알아보자. 다음은 10명의 운잔사에 대하여 시간과 반응시간에 대한 산포도를 그리고 각 운전사에 대하여 단순 회귀직선을 적합하여 그 결과를 추가한 그림이다.

전반적으로 시간이 지나면서 운전자들의 반응시간이 증가하고 있음을 알 수 있다. 또한 **개인별로 반응 시간의 변화와 패턴이 매우 다르다는 것을 알 수 있다.**

```
ggplot(sleepstudy, aes(x=Days, y=Reaction)) +
     geom_point(size=0.5) +
     stat_smooth(method = "lm",se=F,size=0.5)+
     facet_wrap("Subject", labeller = label_both)+
     theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



각 i 번째 운전자에 대하여 10일간 측정한 반응속도 y_{ij}를 시간을 설명변수로 하는 단순 회귀모형을 적합하면 개인별 회귀직선을 다음과 같이 표시할 수 있다.

$$y_{ij} = \beta_{0i} + \beta_{1i} t_j + e_{ij}, \quad i = 1, 2, \ldots, 18, \quad j = 1, 2, \ldots, 10 \tag{1.1}$$

여기서 오차항 e_{ij}은 서로 독립이며 $N(0, \sigma_e^2)$를 따른다고 가정한다.

행렬식으로는 다음과 같이 나타낼 수 있다.

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \mathbf{e}_i$$

여기서

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{i,10} \end{bmatrix}, \ \mathbf{X}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 9 \end{bmatrix}, \boldsymbol{\beta}_i = \begin{bmatrix} \beta_{0i} \\ \beta_{1i} \end{bmatrix}, \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{i,10} \end{bmatrix}$$

위의 식에서 β_{0i}와 β_{1i}는 i번째 운전사의 반응속도를 설명내는 회귀직선의 절편과 기울기이다. 절편 β_{0i}는 실험 시작때 반응속도를 의미하고 기울기 β_{1i}는 실험이 진행되는 동안 반응속도가 어떻게 변하는 지 변화의 방향과 크기를 보여준다.

함수 `lmList`를 아래와 같이 이용하면 식 (1.1) 을 각 운전사마다 적합시켜 각각의 절편과 기울기를 구할 수 있다.

함수 `lmList`에서 모형식 `Reaction ~ Days | Subject`은 각 `Subject` 별로 반응변수를 `Reaction`으로 하고 설명변수를 `Days`로 하는 단순회귀모형을 적합하라는 의미이다.

```
lmf1 <- lmList(Reaction ~ Days | Subject, sleepstudy)
lmf1
```

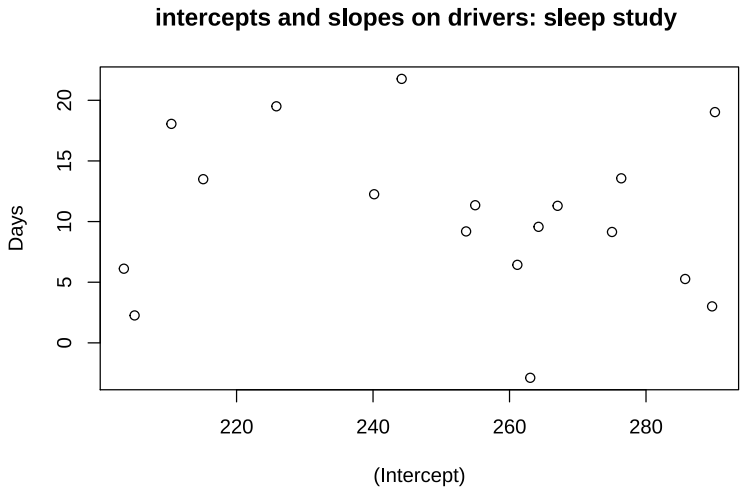
```
## Call: lmList(formula = Reaction ~ Days | Subject, data = sleepstudy)
## Coefficients:
##     (Intercept)      Days
## 308    244.1927 21.764702
## 309    205.0549  2.261785
## 310    203.4842  6.114899
## 330    289.6851  3.008073
## 331    285.7390  5.266019
## 332    264.2516  9.566768
## 333    275.0191  9.142045
## 334    240.1629 12.253141
## 335    263.0347 -2.881034
## 337    290.1041 19.025974
## 349    215.1118 13.493933
## 350    225.8346 19.504017
## 351    261.1470  6.433498
## 352    276.3721 13.566549
## 369    254.9681 11.348109
## 370    210.4491 18.056151
## 371    253.6360  9.188445
## 372    267.0448 11.298073
##
```

```
## Degrees of freedom: 180 total; 144 residual
## Residual standard error: 25.59182
```

```
cor(coef(lmf1))
```

```
##             (Intercept)       Days
## (Intercept)   1.0000000 -0.1375534
## Days         -0.1375534  1.0000000
```

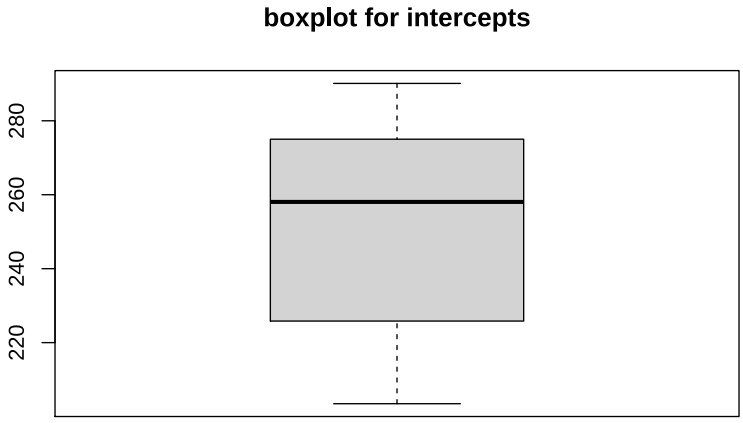
```
plot(coef(lmf1),main="intercepts and slopes on drivers: sleep study ")
```



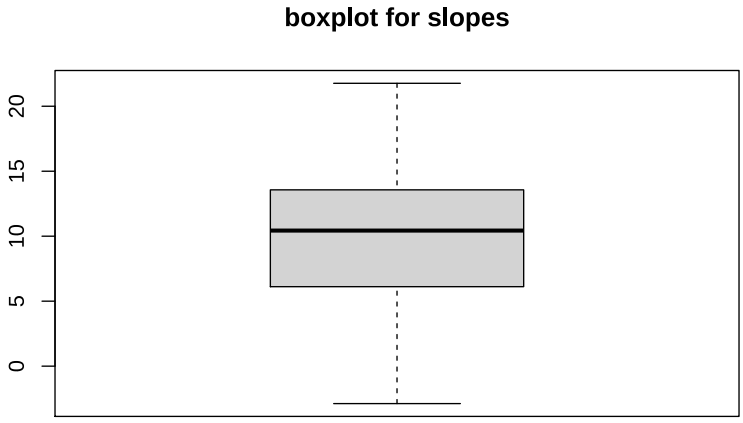
18개의 절편과 기울기는 큰 상관관계는 없는것으로 보이지만 약한 음의 상관계수가 나타났다.

절편과 기울기에 대한 분포를 보기 위하여 상자그림을 그려보면 평균을 중심으로 대칭인 분포를 보이고 있다.

```
boxplot(coef(lmf1)[1])
title('boxplot for intercepts')
```




```
boxplot(coef(lmf1)[2])
title('boxplot for slopes')
```



1.4 통합 분석 (complete pooling)

이제 각 운전사에 대하여 회귀식을 따로 적합하지 않고 전체 운전사들의 자료를 모두 합쳐서 하나의 회귀식을 고려할 수 있다. 개체의 특성을 반영하는 모형이 아닌 **전체 집단에 대한 평균적인 모형 (population model)**을 고려하는 것이다.

$$y_{ij} = \beta_0 + \beta_1 t_j + e_{ij}, \quad i = 1, 2 \ldots, 18, j = 1, 2, \ldots, 10 \tag{1.2}$$

여기서 오차항은 서로 독립이며 $N(0, \sigma_e^2)$를 따른다고 가정한다.

모형 (1.2)은 운전자의 특성을 고려하지 않고 전체 운전사 집단의 관측값에 기반한 모형으로서 시간에 따른 반응시간에 대한 관계를 모집단의 평균적 함수 관계를 파악하는 모형이라고 할 수 있다.

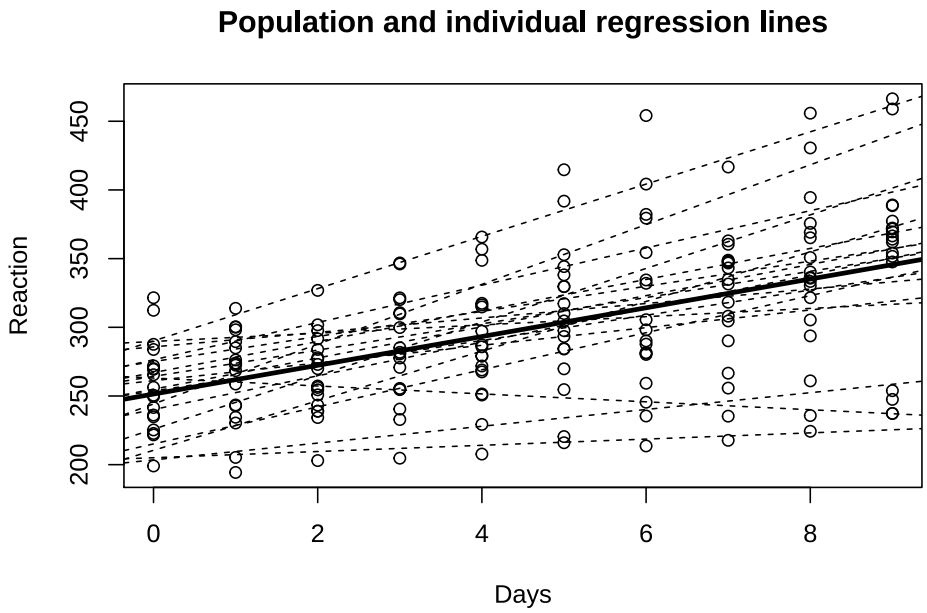
```
lmpop <- lm(Reaction ~ Days, sleepstudy)
summary(lmpop)
```

```
##
## Call:
## lm(formula = Reaction ~ Days, data = sleepstudy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.848  -27.483    1.546   26.142  139.953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  251.405      6.610  38.033  < 2e-16 ***
## Days          10.467      1.238   8.454 9.89e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.71 on 178 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.2825
## F-statistic: 71.46 on 1 and 178 DF,  p-value: 9.894e-15
```

아래 그림에서 굵은 선은 모집단의 평균적 함수 관계를 나타나는 모형 (1.2)을 적합한 결과이며 점선은 개별 운전사의 자료를 적합한 모형 (1.1) 을 나타낸 그림이다.

```
with(sleepstudy, plot(Days, Reaction,main="Population and individual regression lines"))
abline(a=coef(lmpop)[1], b=coef(lmpop)[2],lwd=3)
for ( i in 1:18 ) {
  xx <- as.numeric(coef(lmf1)[i,])
abline(a=xx[1],b=xx[2], lty=2)
```

```
}
```



이제 각 운전사에 대하여 개체별로 적합한 회귀식의 계수들$(\hat{\beta}_{0i}, \hat{\beta}_{1i})$와 전체집단에 적합한 회귀식의 계수 $(\hat{\beta}_0, \hat{\beta}_1)$의 관계를 보면 개체별로 회귀 계수들의 평균이 전체에 적용한 모형의 계수와 매우 가까운 사실을 알 수 있다.

$$\frac{\sum_{i=1}^{n} \hat{\beta}_{0i}}{n} \approx \hat{\beta}_0, \quad \frac{\sum_{i=1}^{n} \hat{\beta}_{1i}}{n} \approx \hat{\beta}_1$$

```
apply(coef(lmf1),2,mean)
```

```
## (Intercept)        Days
##   251.40510    10.46729
```

```
coef(lmpop)
```

```
## (Intercept)        Days
##   251.40510    10.46729
```

1.5 선형 혼합모형(partial pooling)

1.5.1 임의계수모형

앞 절의 모형과 분석에서 알 수 있듯이 한 개체에 대하여 여러 개의 관측값을 측정한 자료에 회귀방정식을 각각 적합시켜보고 또한 개체의 특성을 고려하지않은 전체 모형을 적합해보면 다음과 같은 두 가지 결과를 볼 수 있다.

- 각 개체별 회귀식은 개인의 특성을 반영한다. 즉, 개체에 따라 시간에 따른 반응시간의 변화가 다르게 나타난다.
- 하지만 개인별로 볼 때도 전체적으로는 시간에 따라서 반응시간이 증가하는 경향이 있음을 알 수 있다.
- 전체 자료에 적합한 모형을 보면 개인별로 적합한 모형의 공통적인 성격, 즉 시간에 따른 반응시간의 증가를 알 수 있다.
- 이러한 결과를 보고 각 개인의 변화는 전체적인 변화를 따르면서 각 개인의 특성이 반영되었다고 가정할 수 있다.

위에서 논의하였듯이 전체적인 경향과 게인의 특성을 동시에 고려할 수 있는 모형이 생각할 수 있고 이러한 모형이 다음과 같은 모형이다.

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_j + e_{ij} \tag{1.3}$$

모형 (1.3) 는 절편과 기울기가 두 개의 구성 요소로 더해져서 표현된다.

기울기는 $\beta_1 + b_{1i}$ 로서 나타내어지며 β_1 은 모집단이 가지는 공통적인 경향을 반영하는 모수이고 b_{1i} 는 i 번째 개체의 특성을 반영한 확률변수이다.

절편도 유사한 형식으로 구성된다. 각 개인에 대한 특성을 나타내는 변수 (b_{0i}, b_{1i}) 을 확률변수로 설정하고 이를 모수 (β_0, β_1) (parameter or fixed effect) 와 구별하여 임의효과 (random effect) 라고 한다.

식 (1.3)에서 제시된 모형은 **임의계수모형(random coefficient model)** 이라고 부른다.

18명에 대한 회귀직선의 절편과 기울기를 보면 개인의 차이에 따른 변동을 볼 수 있으며 이러한 각 개인간의 변동을 임의효과 를 이용하여 다음과 같은 모형을 생각해보자.

$$\boldsymbol{\beta}_i = \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} + \begin{bmatrix}b_{0i} \\ b_{1i}\end{bmatrix}, \quad \begin{bmatrix}b_{0i} \\ b_{1i}\end{bmatrix} \sim N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\sigma_{b1}^2 & \rho\sigma_{b1}\sigma_{b2} \\ \rho\sigma_{b1}\sigma_{b2} & \sigma_{b2}^2\end{bmatrix}\right)$$

위의 모형은 각 개인의 회귀직선에서 각 절편과 기울기가 전체평균 β_0와 β_1를 따르며 각 개인의 차이는 전체평균에 임의효과인 b_{0i}와 b_{1i}가 더해져서 나타난다는 것을 의미한다. 이변량 임의효과 b_{0i}와 b_{1i}는 이변량 정규분포를 따르며 각각의 분산과 상관계수가 σ_{b1}^2, σ_{b2}^2, ρ이다.

다른 개체에 대한 임의효과는 서로 독립이며 임의 효과와 오차항은 독립이다. 여기서 오차항은 서로 독립이며 $N(0, \sigma_e^2)$를 따른다고 가정한다.

$$Cov(\mathbf{b}_i, \mathbf{b}_j) = \mathbf{0} \text{ when } i \neq j, \quad Cov(\mathbf{b}_i, e_{jk}) = \mathbf{0} \text{ for all } i, j, k$$

1.5.2 혼합효과 모형

임의계수모형을 각 개인 i에 대하여 행렬식으로 표시하면 다음과 같은 혼합효과모형(mixed effects model)으로 나타낼 수 있다. 혼합효과모형은 반응변수에 영향을 미치는 효과를 고정효과와 임의효과로 나누어 설명하는 모형이다.

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

여기서

$$\mathbf{y}_i = \begin{bmatrix}y_{i1} \\ y_{i2} \\ \vdots \\ y_{i,10}\end{bmatrix}, \mathbf{X}_i = \begin{bmatrix}1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 9\end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix}, \mathbf{Z}_i = \begin{bmatrix}1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 9\end{bmatrix}, \mathbf{b}_i = \begin{bmatrix}b_{0i} \\ b_{1i}\end{bmatrix}, \mathbf{e}_i = \begin{bmatrix}e_{i1} \\ e_{i2} \\ \vdots \\ e_{i,10}\end{bmatrix}$$

위의 각 개인에 대한 모형을 모두 합쳐서 하나의 혼합효과모형으로 나타내면 다음과 같이 표현할 수 있다.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e} \tag{1.4}$$

여기서 반응변수벡터 $\mathbf{y}$와 고정효과 $\boldsymbol{\beta}$에 대한 계획행렬 X는 각 개인의 반응변수벡터 $\mathbf{y}_i$와 $\mathbf{X}_i$를 행으로 쌓아놓은 것으로 표현된다. 오차항에 대한 벡터 $\mathbf{e}$도 동일한 형식의 벡터이다.

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{18} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_{18} \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_{18} \end{bmatrix}$$

임의효과 벡터 $\mathbf{b}$는 각 개인에 대한 임의효과벡터 $\mathbf{b}_i$를 행으로 쌓아놓은것과 같고 임의효과에 대한 계획행렬 $\mathbf{Z}$는 각 개인의 계획행렬 $\mathbf{Z}_i$를 대각원소로 같은 행렬이다.

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_{18} \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \mathbf{Z}_{18} \end{bmatrix}$$

임의효과는 개인의 특성을 설명하는 효과로서 모집단을 구성하는 개인이 표본에 추출되었다고 생각하며 확률분포를 따른다고 가정한다. 반복측정자료에서 임의효과를 공통으로 가지고 있는 관측치는 독립이 아니게 돼며 따라서 같은 개체에서 나온 관측값은 독립이 아니다.

1.5.3 선형혼합모형의 적합

혼합모형 @ref{eq:lme1}은 `lmer()` 함수를 이용하여 적합시켜보자. 모형식에서 `(1 + Days|Subject)` 는 각 개체 `Subject` 에 대하여 절편 `1` 과 기울기 `Days`에 임의효과를 포함한다고 지정한다.

```
fm1 <- lmer(Reaction ~ 1 + Days + (1 + Days|Subject), sleepstudy)
summary(fm1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
##    Data: sleepstudy
##
```

```
## REML criterion at convergence: 1743.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9536 -0.4634  0.0231  0.4634  5.1793
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 612.10   24.741
##           Days         35.07    5.922   0.07
##  Residual             654.94   25.592
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  251.405      6.825  17.000  36.838  < 2e-16 ***
## Days          10.467      1.546  17.000   6.771 3.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.138
```

위의 혼합모형 적합결과를 살펴보자. 첫째로 고정효과에 대한 추정식은 다음과 같다

```
fixef(fm1)
```

```
## (Intercept)        Days
##   251.40510    10.46729
```

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 251.40510 \\ 10.46729 \end{bmatrix}$$

또한 오차항에 대한 분산 및 임의효과의 분산성분과 상관계수는 다음과 같이 나타난다.

```
VarCorr(fm1)
```

```
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 24.7407
##           Days         5.9221  0.066
##  Residual             25.5918
```

$$\hat{\sigma}_{b1} = 24.7407$$
$$\hat{\sigma}_{b2} = 5.9221$$
$$\hat{\rho} = 0.066$$
$$\hat{\sigma}_{e} = 25.5918$$

1.5.4 임의효과에 대한 예측

이제 임의효과 $\mathbf{b}_i = (b_{0i}, b_{1i})^t$에 대한 예측(prediction)을 생각해보자. 우리는 오직 관측벡터 $\mathbf{y}_i$만을 관측하고 임의효과 $\mathbf{b}_i$는 관측을 할 수 없는 확률변수이다. 하지만 주어진 관측벡터와 추정된 분산으로 임의효과의 값을 예측할 수있으며 그 결과는 다음과 같다.

```
re <- ranef(fm1)$Subject
re
```

```
##     (Intercept)        Days
## 308   2.2585509   9.1989758
## 309 -40.3987381  -8.6196806
## 310 -38.9604090  -5.4488565
## 330  23.6906196  -4.8143503
## 331  22.2603126  -3.0699116
## 332   9.0395679  -0.2721770
## 333  16.8405086  -0.2236361
## 334  -7.2326151   1.0745816
```

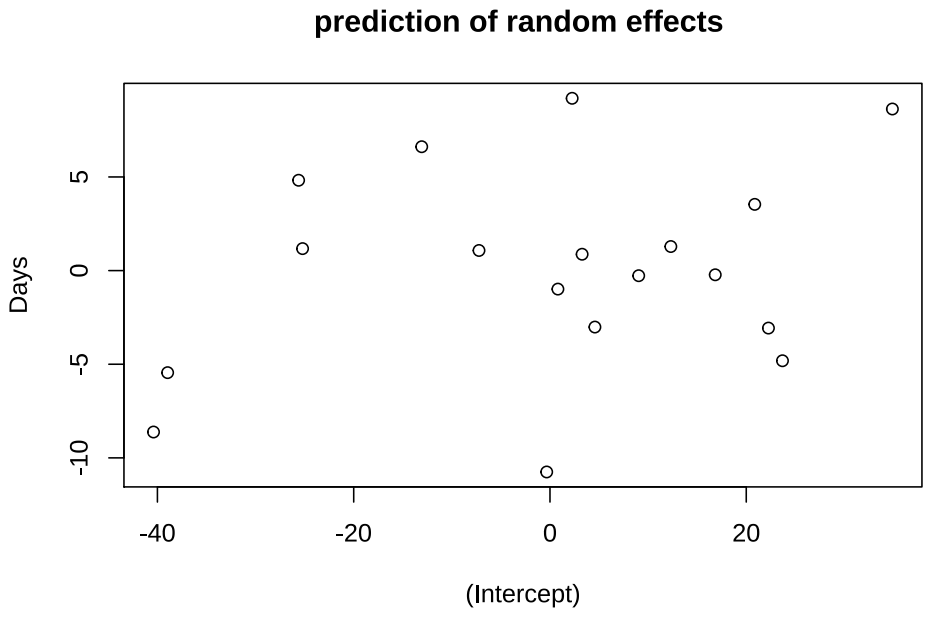
```
## 335  -0.3336684 -10.7521652
## 337  34.8904868   8.6282652
## 349 -25.2102286   1.1734322
## 350 -13.0700342   6.6142178
## 351   4.5778642  -3.0152621
## 352  20.8636782   3.5360011
## 369   3.2754656   0.8722149
## 370 -25.6129993   4.8224850
## 371   0.8070461  -0.9881562
## 372  12.3145921   1.2840221
```

예를 들어 첫 번째 운전자에 대한 절편과 기울기의 임의효과에 대한 예측값은 다음과 같다.

$$\hat{b}_{0i} = 2.2585509, \quad \hat{b}_{1i} = 9.1989758$$

위에서 구한 절편과 기울기에 대한 임의효과들의 산포도를 보면 다음과 같다.

```
plot(re, main ="prediction of random effects ")
```



예측된 각 개인의 절편과 기울기에 대한 임의효과의 예측값 $\hat{b}_{0i}$과 $\hat{b}_{1i}$에 고정효과의 추정량

$\hat{\beta}_0$와 $\hat{\beta}_1$에 각각 더해주면 각 개인의 절편과 기울기에 대한 예측값을 구할 수 있다.

$$\hat{\beta}_{0i} = \hat{\beta}_0 + \hat{b}_{0i}, \quad \hat{\beta}_{1i} = \hat{\beta}_1 + \hat{b}_{1i}$$

```
beta <- matrix(as.numeric(fixef(fm1)),18,2,byrow=T)
beta + re
```

```
##     (Intercept)       Days
## 308    253.6637 19.6662617
## 309    211.0064  1.8476053
## 310    212.4447  5.0184295
## 330    275.0957  5.6529356
## 331    273.6654  7.3973743
## 332    260.4447 10.1951090
## 333    268.2456 10.2436499
## 334    244.1725 11.5418676
## 335    251.0714 -0.2848792
## 337    286.2956 19.0955511
## 349    226.1949 11.6407181
## 350    238.3351 17.0815038
## 351    255.9830  7.4520239
## 352    272.2688 14.0032871
## 369    254.6806 11.3395008
## 370    225.7921 15.2897709
## 371    252.2122  9.4791297
## 372    263.7197 11.7513080
```

예를 들어 선형혼합모형에서 첫 번째 운전자에 대한 절편과 기울기에 대한 추정값은 다음과 같다.

$$\hat{\beta}_{0i} = 253.6637, \quad \hat{\beta}_{1i} = 19.6662617$$

위의 결과를 각 운전자에 대해 개별 회귀직선 (1.1)을 적합시켜서 얻은 18개의 절편과 기울

기와 비교해보자.

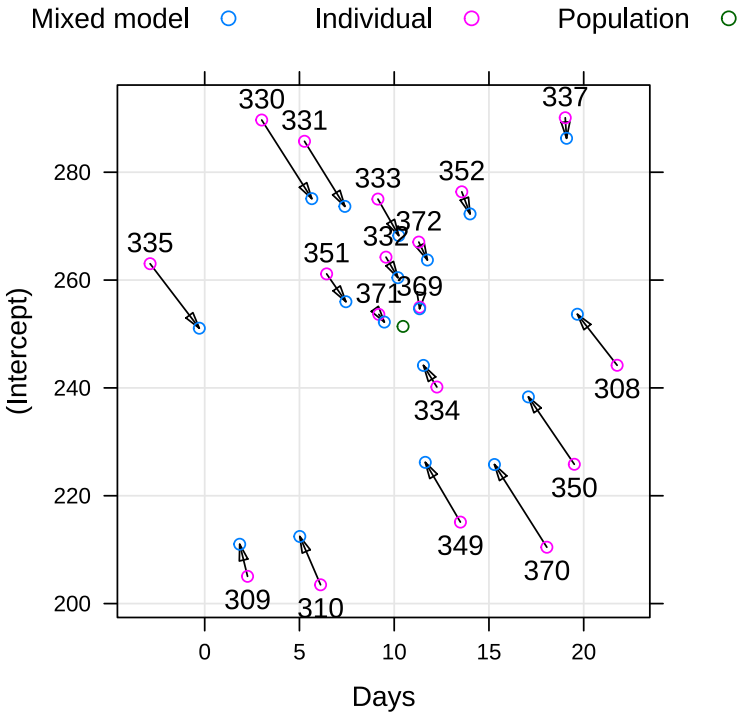
```
coef(lmf1)
```

```
##     (Intercept)      Days
## 308    244.1927 21.764702
## 309    205.0549  2.261785
## 310    203.4842  6.114899
## 330    289.6851  3.008073
## 331    285.7390  5.266019
## 332    264.2516  9.566768
## 333    275.0191  9.142045
## 334    240.1629 12.253141
## 335    263.0347 -2.881034
## 337    290.1041 19.025974
## 349    215.1118 13.493933
## 350    225.8346 19.504017
## 351    261.1470  6.433498
## 352    276.3721 13.566549
## 369    254.9681 11.348109
## 370    210.4491 18.056151
## 371    253.6360  9.188445
## 372    267.0448 11.298073
```

이렇게 혼합모형을 통해서 얻은 각 개인의 절편과 기울기에 대한 예측값과 각각의 개인에 대해서 회귀직선을 따로 적합하여 얻은 절편과 기울기의 관계를 그림으로 그려보면 다음과 같다. 혼합모형을 통해서 얻은 각 개인의 절편과 기울기는 절편과 기울기의 전체평균값 방향으로 축소되는 경향(shrinkage)을 볼수있다.

```
df <- coef(lmf1)
fclow <- subset(df, `(Intercept)` < 251)
fchigh <- subset(df, `(Intercept)` > 251)
cc1 <- as.data.frame(coef(fm1)$Subject)
names(cc1) <- c("A", "B")
```

```
df <- cbind(df, cc1)
ff <- fixef(fm1)
with(df,
     print(xyplot(`(Intercept)` ~ Days, aspect = 1,
                  x1 = B, y1 = A,
                  panel = function(x, y, x1, y1, subscripts, ...) {
                      panel.grid(h = -1, v = -1)
                      x1 <- x1[subscripts]
                      y1 <- y1[subscripts]
                      larrows(x, y, x1, y1, type = "closed", length = 0.1,
                              angle = 15, ...)
                      lpoints(x, y,
                              pch = trellis.par.get("superpose.symbol")$pch[2],
                              col = trellis.par.get("superpose.symbol")$col[2])
                      lpoints(x1, y1,
                              pch = trellis.par.get("superpose.symbol")$pch[1],
                              col = trellis.par.get("superpose.symbol")$col[1])
                      lpoints(ff[2], ff[1],
                              pch = trellis.par.get("superpose.symbol")$pch[3],
                              col = trellis.par.get("superpose.symbol")$col[3])
                      ltext(fclow[,2], fclow[,1], row.names(fclow),
                            adj = c(0.5, 1.7))
                      ltext(fchigh[,2], fchigh[,1], row.names(fchigh),
                            adj = c(0.5, -0.6))
                  },
                  key = list(space = "top", columns = 3,
                  text = list(c("Mixed model", "Individual", "Population")),
                   points = list(col = trellis.par.get("superpose.symbol")$col[1:3],
                   pch = trellis.par.get("superpose.symbol")$pch[1:3]))
                ))
```



1.5.5 모형의 축소

위에서 고려한 임의계수모형 (1.3)에서는 절편과 기울기에 대한 2개의 임의효과 b_{0i} 와 b_{1i} 를 고려하고 더 나아가 두 개의 임의효과가 독립이 아니며 상관계수가 ρ 라고 가정하였다.

앞에서 추정결과에 의하면 두 개의 임의효과의 상관계수의 추정값은 $\hat{\rho} = 0.066$ 으로 거의 0에 가깝다. 이러한 결과에 근거하여 두 임의효과가 독립인 축소모형을 고려해 보자. 즉 임의계수모형 (1.3)에서 임의효과의 상관계수가 $\rho = 0$ 인 임의효과의 분포를 다음과 같이 가정한다.

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{b1}^2 & 0 \\ 0 & \sigma_{b2}^2 \end{bmatrix} \right)$$

이러한 모형을 아래와 같이 적합시키면 추정결과는 다음과 같다. 아래 모형식에서 `(1 + Days||Subject)` 는 각 개체 `Subject` 에 대하여 절편 `1` 과 기울기 `Days`에 임의효과를 포함한다고 지정하며 한개의 바 `|` 대신 두 개의 바 `||` 를 사용하면 임의효과가 독립이라는 것을 지정한다.

```
fm2 <- lmer(Reaction ~ 1 + Days + (1+Days || Subject) , sleepstudy)
summary(fm2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Reaction ~ 1 + Days + (1 + Days || Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1743.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9626 -0.4625  0.0204  0.4653  5.1860
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  Subject   (Intercept) 627.57   25.051
##  Subject.1 Days         35.86    5.988
##  Residual              653.58   25.565
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  251.405      6.885  18.156  36.513  < 2e-16 ***
## Days          10.467      1.560  18.156   6.712 2.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.184
```

상관계수가 0인 모형에 대한 추정 결과는 상관계수가 있는 모형과 크게 다르지 않다.

두 모형, 즉 절편과 기울기에 대한 두 임의효과가 종속인지 또는 독립인지에 대한 두 모형을 AIC(Akaike Information Criteris)와 BIC(Bayesian Information Criteria)로 비교한 결과이다. 두 모형 간의 차이는 거의 없는 것으로 판단된다.

```
c(AIC(fm1) , BIC(fm1))
```

```
## [1] 1755.628 1774.786
```

```
c(AIC(fm2), BIC(fm2))
```

```
## [1] 1753.669 1769.634
```

더나아가 anova 함수를 이용하여 두 모형의 차이를 검정한 결과는 두 모형 간의 차이가 없다는 것이다. 참고할 점은 혼합모형에서의 모형을 비교하는 분산분석에 의한 검정은 효율이 떨어질 수 있기 때문에 주의해야 한다.

```
anova(fm1,fm2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: sleepstudy
## Models:
## fm2: Reaction ~ 1 + Days + (1 + Days || Subject)
## fm1: Reaction ~ 1 + Days + (1 + Days | Subject)
##     npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fm2    5 1762.0 1778.0 -876.00   1752.0
## fm1    6 1763.9 1783.1 -875.97   1751.9 0.0639  1     0.8004
```

1.6 베이지안 추정

임의효과를 이용하는 선형혼합 모형은 모형의 구조상 베이지안 모형과 매우 유사하다. 베이지안 모형에서는 회귀계수와 분산 성분에 대한 사전분포(prior distribution)을 설정하고 모수들의 사후분포(posterior distribution)을 추정한다.

베이지안 모형의 적합은 매우 다양한 패키지나 프로그램을 사용할 수 있다(BUGS,STAN, MCMCglmm, brms). 본 강의에서는 brms 패키지의 brm 함수를 사용하여 베이지안 방법으

로 임의계수 모형을 적합하는 예를 보여주고자 한다. brms 패키지는 lme4 패키지의 모형식을 그대로 사용할 수 있다. 참고로 다른 패키지는 lme4 패키지의 모형식보다 더욱 복잡한 식을 사용해야 한다.

sleepstudy 자료에 대하여 brms 패키지의 brm 함수를 이용하여 베이지안 방법으로 임의계수 모형을 추정하는 프로그램은 아래와 같다.

```
fm3 <- brm(Reaction ~ 1 + Days + (1 + Days|Subject), data = sleepstudy)
```

```
## Compiling Stan program...
```

```
## Start sampling
```

```
##
## SAMPLING FOR MODEL '9911b459d8c4b04b2868a198ead77672' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 7.2e-05 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0.72 se
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 1:
## Chain 1:  Elapsed Time: 1.61174 seconds (Warm-up)
```

```
## Chain 1:                0.570718 seconds (Sampling)
## Chain 1:                2.18246 seconds (Total)
## Chain 1:
##
## SAMPLING FOR MODEL '9911b459d8c4b04b2868a198ead77672' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 2.4e-05 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0.24 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
## Chain 2:
## Chain 2: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 2: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 2: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 2: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 2: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 2: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 2: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 2: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 2: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 2: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 2: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 2: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 2:
## Chain 2:  Elapsed Time: 1.03919 seconds (Warm-up)
## Chain 2:                0.648274 seconds (Sampling)
## Chain 2:                1.68746 seconds (Total)
## Chain 2:
##
## SAMPLING FOR MODEL '9911b459d8c4b04b2868a198ead77672' NOW (CHAIN 3).
## Chain 3:
## Chain 3: Gradient evaluation took 2.2e-05 seconds
## Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 0.22 seconds.
```

```
## Chain 3: Adjust your expectations accordingly!
## Chain 3:
## Chain 3:
## Chain 3: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 3: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 3: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 3: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 3: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 3: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 3: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 3: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 3: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 3: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 3: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 3: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 3:
## Chain 3:  Elapsed Time: 1.1138 seconds (Warm-up)
## Chain 3:                0.518563 seconds (Sampling)
## Chain 3:                1.63236 seconds (Total)
## Chain 3:
##
## SAMPLING FOR MODEL '9911b459d8c4b04b2868a198ead77672' NOW (CHAIN 4).
## Chain 4:
## Chain 4: Gradient evaluation took 2.2e-05 seconds
## Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 0.22 se
## Chain 4: Adjust your expectations accordingly!
## Chain 4:
## Chain 4:
## Chain 4: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 4: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 4: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 4: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 4: Iteration:  800 / 2000 [ 40%]  (Warmup)
```

```
## Chain 4: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 4: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 4: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 4: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 4: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 4: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 4: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 4:
## Chain 4:  Elapsed Time: 1.0575 seconds (Warm-up)
## Chain 4:                0.581012 seconds (Sampling)
## Chain 4:                1.63851 seconds (Total)
## Chain 4:
```

```
fm3
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
##    Data: sleepstudy (Number of observations: 180)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Group-Level Effects:
## ~Subject (Number of levels: 18)
##                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)          27.12      6.95    15.53    42.53 1.00     1761     2389
## sd(Days)                6.66      1.61     4.21    10.31 1.00     1289     1359
## cor(Intercept,Days)     0.09      0.30    -0.47     0.69 1.00      927     1540
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept   251.47      7.59   236.47   266.10 1.00     1668     2446
## Days         10.41      1.76     7.01    13.88 1.00     1173     1341
```

```
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma    25.85      1.53    23.06    29.07 1.00     2827     2619
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

제 2 장

소지역추정

층화추출을 이용하는 표본조사(stratified sampling)에서 층에 배정된 표본수가 적을 경우에는 특성이 유사한 인근의 층의 추출단위들을 결합하여 추출단위들의 그룹을 만들고 그 그룹 내에서 추출단위들은 동일한 특성을 갖는다고 가정할 수 있으면 유사한 인근의 조사결과를 활용해서 좀더 정도높은 추정값을 작성할 수 있다.

이와 같이 비슷한 특성을 갖는 인근의 조사결과나 행정업무자료 또는 센서스 등 다른 통계조사 정보를 조사된 자료에서 얻은 직접 추정값과 결합하여 세부단위에 대한 통계를 작성하는 기법을 **소지역 추정법(Small Area Estimation)**이라고 한다.

소지역 추정법은 지역 수준모형(area-level model)과 단위 수준모형(unit-level model)이 있다.

- 지역 수준모형: 지역과 관련된 직접추정량과 지역의 보조자료를 결합한 모형
- 단위 수준모형: 지역과 관련된 직접추정량과 지역 안의 추출단위에 대한 보조자료를 결합한 모형

이 강의 예제에서는 단위 수준모형만 다룰 것이다.

2.1 단위 수준모형의 개요

2.1.1 모집단에 대한 가정

모집딘 U 가 D 개의 층으로 나누어져 있다고 가정하자, 층 $U_1, U_2, \dots, U_D$ 는 서로 겹치지 않는 작은 세부단위이며 각 층에는 각각 $N_1, N_2, \dots, N_D$ 의 추출단위가 있다고 하자.

이제 Y_{dj} 를 층 d 에 속하는 i 번째 단위가 가진 값이라고 하자. 층 d 에 속하는 모든 반응값들을 벡터로 표시하면 $\mathbf{y}_d = (Y_{d1}, Y_{d2}, \cdots, Y_{dN_d})^t$ 이다.

이제 추정하고 싶은 모수는 층별 모평균 $\bar{Y}_d$ 이다.

$$\bar{Y}_d = \frac{\sum_{j=1}^{N_s} Y_{dj}}{N_d}$$

이제 층별 모평균을 추정하기 위하여 층 U_d 에서 n_d 개의 표본 s_d 를 추출한다. 총 표본의 수 $n = \sum_d n_d$ 는 각 층에서 추출한 표본의 수의 합니다.

소지역에서 표본에 속하지 않는 단위들의 집합을 r_d 로 표기한다.

2.1.2 보조 정보와 계층 모형

이제 추출 단위(sampling unit)에 대한 보보 정보가 있다고 하자. 보조 정보는 추출 단위들의 실제값 y_{dj} 에 대한 정보를 지니고 있으며 p 개의 변수로 구성되어 있다고 가정하자.

추출 단위에 대한 보조 정보 $\mathbf{x}_{dj}$ 는 추출단위의 관심 변수 Y_{dj} 와 다음과 같은 계층보형의 관계를 가진다고 가정한다. 이러한 모형은 (Battese et al., 1988)에 의하여 제안된 선형 혼합효과 모형이다.

$$Y_{dj} = \mathbf{x}_{dj}^t \boldsymbol{\beta} + u_d + e_{dj} \tag{2.1}$$

위 식 (2.1) 에서 u_d 는 임의효과로서 소지역 U_d 에 대한 효과(area effect)를 반영한다. 또한 e_{dj} 는 추출단의의 오차이다. 소지역 효과 u_d 와 오차항 e_{dj} 는 다음과 같이 각각 정규분포를 따른다.

$$u_d \sim N(0, \sigma_u^2), \quad e_{dj} \sim N(0, \sigma_e^2) \tag{2.2}$$

소지역 효과 u_d와 오차항 e_{dj} 는 서로 독립이며 분산성분 σ_u^2 과 σ_e^2 는 모르는 모수로 추정을 해야한다.

2.1.3 추정법

단위 수준모형 (2.1)에서 소지역 평균 $\bar{Y}_d$에 대한 추정량은 다음과 같은 최적 선형불편 예측량(Best Linear Unbiased Prediction; BLUP)으로 주어진다. (Royall, 1970).

$$\tilde{\bar{Y}}_d^{BLUP} = \frac{1}{N_d}\left[\sum_{j\in s_d} Y_{dj} + \sum_{j\in r_d} \tilde{Y}_{dj}\right] \tag{2.3}$$

추정식 (2.3) 에서 표본에 속하지 않은 단위들에 대한 추정량 $\tilde{Y}_{dj}$ 은 다음과 같이 주어진다.

$$\tilde{Y}_{dj} = \mathbf{x}_{dj}^t\tilde{\boldsymbol{\beta}} + \tilde{u}_d$$

위의 식에서 $\tilde{u}_d$ 는 소지역의 임의효과 u_d 에 대한 최적 선형불편 예측량(BLUP)이며 다음과 같이 표시할 수 있다.

$$\tilde{u}_d = \gamma_d(\bar{y}_{ds} - \bar{\mathbf{x}}_{ds}^t\tilde{\boldsymbol{\beta}})$$

여기서

- $\bar{y}_{ds} = \sum_{j\in s_d} Y_{dj}/n_d$ 는 소지역 d 에서 추출된 관심변수 표본의 평균
- $\bar{\mathbf{x}}_{ds} = \sum_{j\in s_d} \mathbf{x}_{dj}/n_d$ 는 소지역 d 에서 추출된 보조변수 표본의 평균

또한 같은 소지역에 소과는 단위들의 상관관계를 의미하는 γ_d는 다음과 같다.

$$\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/n_d}$$

위의 추정식들에서 분산성분 σ_u^2 과 σ_e^2 은 모르는 모수이므로 최대가능도 추정(또는 제한적 최대 가능도 추정량, REML)을 통하여 추정하여야 한다.

추정된 분산성분을 다시 추정식 @res(eq:blup1) 에 대입하고 최종적으로 정리하면 소지역 추정량은 다음과 같이 주어진다.

$$\hat{\bar{Y}}_d^{BLUP} = f_d \bar{y}_{ds} + (\bar{\mathbf{X}}_d - f_d \bar{\mathbf{x}}_{ds})^t \hat{\boldsymbol{\beta}} + (1 - f_d)\hat{u}_d \tag{2.4}$$

여기서 $f_d = n_d/N_d$ 로서 표본 비율이다. 최종 소지역 추정식 @res(eq:blup2) 를 보면 보조 변수의 모집단 평균 $\bar{\mathbf{X}}_d$rk 필요한 것을 알 수 있다.

위의 소지역 추정에 대한 요약은 (Molina and Marhuenda, 2015)에 더 자세한 내용이 있다.

2.2 예제

예제에서는 자영업이 많은 부분을 차지하고 있는 음식점에 대한 경제자료를 이용하여 표본평균을 이용한 단순추정량과 소지역 모형을 사용한 추정량을 비교하려고 한다.

분석 자료는 A지역의 사업체를 한식음식점을 대상으로 제한하여 사용하였다.

- 표본설계는 동(소지역)을 층으로 하는 단순층화추출을 사용하였다.
- 관심변수는 매출액을 고려하였다.
- 2개의 보조변수는 재료구입비와 연간급여액을 고려하였다. 연간급여액과 재료구입비는 고용 보험 및 신용카드 내역을 통해 모든 사업체에 대해여 얻을 수 있는 자료로 지역 추정 모형에서 보조변수로 활용하였다.

두 보조자료를 바탕으로 매출액을 추정하는 소지역 추정 모형을 세워 단순 추정과 비교하고자 한다. 분석 자료에서 쓰이는 변수의 단위는 모두 '백만원'이다.

분석에 사용되는 모집단 자료의 개수는 총 10167 개로 A지역은 총 326개의 동(소지역, 층)로 나뉜다.

모의실험에서 표본 자료는 1489 개의 음식점을 추출하여 사용하였으며, 이를 추출하기 위하여 모집단의 각 동의 수에 비례하는 층화추출 방법을 사용하였다.

분석에서 사용되는 변수와 의미는 다음과 같다.

- `SLE_ALL_AMT"`: 매출액
- `MNF_PCST` : 재료구입비
- `SURV_PHS_SLR_SUM` : 연간급여액
- `AD_CD` : 지역코드

2.2.1 자료 입력 및 정리

```
# SEED 고정
set.seed(321)

# 자료읽기
data <- read.sas7bdat("/Users/ylee19067/Dropbox/working/teaching/kimhj/sas/derived/fin_data.sas7b

# 필요한 변수 선택
data <- data[c("SURV_PHS_SLR_SUM", "SLE_ALL_AMT", "MNF_PCST" , "AD_CD")]

# 자료를 지역번호 순으로 정렬
data <- data[order(data$AD_CD), ]

# 지역코드를 범주화 변수로 변환
data$AD_CD <- as.factor(data$AD_CD)

head(data, n=10)
```

```
##      SURV_PHS_SLR_SUM SLE_ALL_AMT MNF_PCST   AD_CD
## 42                 14          103       23 1101053
## 44                 13           40       16 1101053
## 45                  4           40       15 1101053
## 878                12          100       38 1101053
## 901                 2           20        8 1101053
## 909                 7           36       14 1101053
## 910                 1           30        9 1101053
## 920                 7           15        6 1101053
## 921                 2           15        6 1101053
## 1028                5           56       15 1101053
```

2.2.2 모집단 정보 생성

```
# 모집단에서 각 동에 속한 사업체의 수를 계산
gu <-  unique(data$AD_CD)
gugu <- table(data$AD_CD)
head(gugu, n=20)
```

```
##
## 1101053 1101054 1101055 1101056 1101057 1101058 1101060 1101061 1101063 1101064
##     111      23       7      22       3      11      33     274      60      37
## 1101065 1101066 1101067 1101068 1101069 1101070 1101071 1101072 1102052 1102054
##      38       4      21      14       4      14      18      16     125      85
```

2.2.3 사업체의 수가 10개 미만인 동을 제거

표본 추출할 경우 사업체으 수가 극단적으로 작은 동은 표본의 배정이 사업체의 수보다 많아질 수 있으므로 사업체의 수가 10 미만인 동을 모집단에서 제거한다.

```
# 사업체의 수가 10 미만인 동을 제거
smalldong <- names(gugu[gugu < 10])
data <- data %>% rowwise() %>% filter(!AD_CD %in% smalldong)
data$AD_CD <- droplevels(data$AD_CD)
# 모집단에서 각 동에 속한 사업체의 수를 다시 계산
gugu = table(data$AD_CD)
head(gugu, n=20)
```

```
##
## 1101053 1101054 1101056 1101058 1101060 1101061 1101063 1101064 1101065 1101067
##     111      23      22      11      33     274      60      37      38      21
## 1101068 1101070 1101071 1101072 1102052 1102054 1102055 1102057 1102058 1102059
##      14      14      18      16     125      85     188      41      24      61
```

2.2.4 최종 모집단의 수

```
dim(data)
```

```
## [1] 10167     4
```

- 최종 모집단에서 음식점의 수는 10167 개이다.
- 최종 모집단에서 소지역(동)의 개수는 326 개이다.

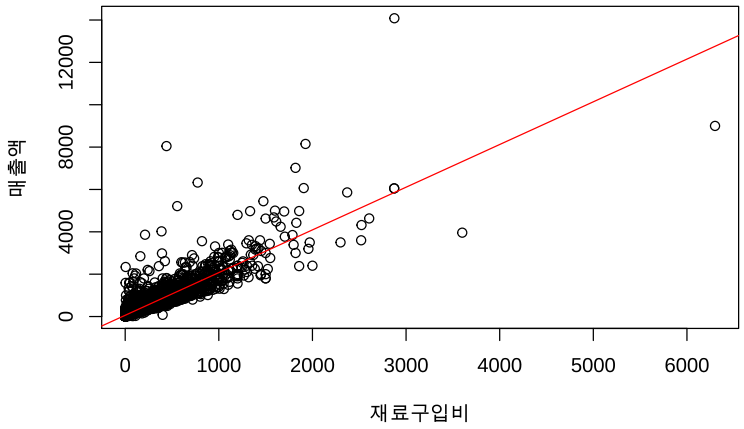
2.2.5 관심변수와 보조변수의 관계

이제 모집단에서 매출액과 재료 구입비에 대한 관계를 회귀분석으로 분석해 보자.

```
line1 <- lm(data$SLE_ALL_AMT ~ data$MNF_PCST)
summary(line1)
```

```
##
## Call:
## lm(formula = data$SLE_ALL_AMT ~ data$MNF_PCST)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3759.3   -52.0   -25.8    20.7  8228.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   58.92009    2.66787   22.09   <2e-16 ***
## data$MNF_PCST  2.01593    0.01018  197.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 230.2 on 10165 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.794
## F-statistic: 3.918e+04 on 1 and 10165 DF,  p-value: < 2.2e-16
```

```
plot(data$SLE_ALL_AMT ~ data$MNF_PCST, xlab='재료구입비', ylab='매출액')
abline(line1, col='red')
```




```
equatiomatic::extract_eq(line1, use_coefs = TRUE)
```

$$\text{data\$SLE\_ALL\_AMT} = 58.92 + 2.02(\text{data\$MNF\_PCST})$$

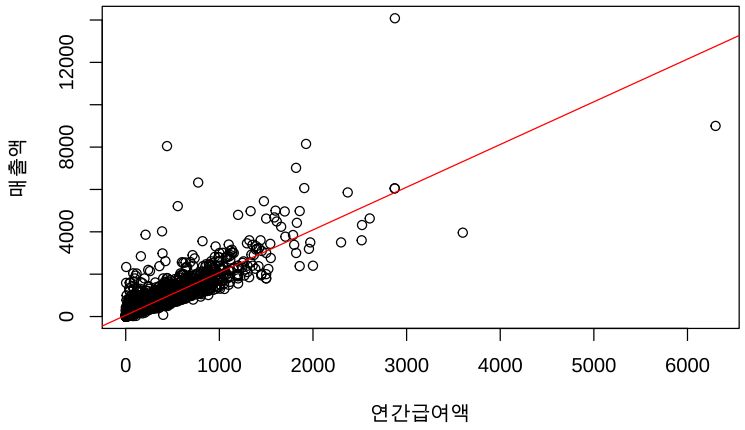
이제 모집단에서 매출액과 연간급여액에 대한 관계를 분석해 보자.

```
line2 <- lm(data$SLE_ALL_AMT ~ data$MNF_PCST)
summary(line2)
```

```
##
## Call:
## lm(formula = data$SLE_ALL_AMT ~ data$MNF_PCST)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3759.3   -52.0   -25.8    20.7  8228.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   58.92009    2.66787   22.09   <2e-16 ***
## data$MNF_PCST  2.01593    0.01018  197.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 230.2 on 10165 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.794
## F-statistic: 3.918e+04 on 1 and 10165 DF,  p-value: < 2.2e-16
```

```
plot(data$SLE_ALL_AMT ~ data$MNF_PCST, xlab='연간급여액', ylab='매출액')
abline(line2, col='red')
```




```
equatiomatic::extract_eq(line2, use_coefs = TRUE)
```

$$\text{data\$SLE\_ALL\_AMT} = 58.92 + 2.02(\text{data\$MNF\_PCST})$$

2.2.6 표본 배정과 표본 가중치을 위한 함수

```
#표본 배정을 위한 함수
calsamspleize <- function(samplesize, gugu) {
  ceiling(gugu/sum(gugu)*samplesize)
}
```

```
#표본 가중치을 위한 함수
calweight2 <- function(samplesize, gugu){
  xx <- as.data.frame(gugu)
  yy <- as.data.frame(samplesize)
  w = xx$Freq/ yy$Freq
  data.frame(AD_CD = xx$Var1, popN = xx$Freq, samplesize = yy$Freq, weight=w)
}
```

2.2.7 표본 배정

```
#먼저 1000 개의 표본을 비례할당
samplesize = 1000
sample.result <- calsamspleize(samplesize, gugu)

# 최소 표본을 1개로 지정
sample.result <- sample.result+1

# 표본 개수 다시 계산
samplesize <- sum(sample.result )

# 표본의 총수
sum(samplesize)
```

```
## [1] 1489
```

```
# 최종 표본 할당
head(sample.result, n=20)
```

```
##
## 1101053 1101054 1101056 1101058 1101060 1101061 1101063 1101064 1101065 1101067
##      12       4       4       3       5      28       7       5       5       4
## 1101068 1101070 1101071 1101072 1102052 1102054 1102055 1102057 1102058 1102059
```

```
##        3        3        3        3       14       10       20        6        4        7
```

2.2.8 표본 가중치 계산 및 모집단 정보 생성

```
popinfo <- calweight2(sample.result, gugu)
head(popinfo)
```

```
##     AD_CD popN samplesize   weight
## 1 1101053  111         12 9.250000
## 2 1101054   23          4 5.750000
## 3 1101056   22          4 5.500000
## 4 1101058   11          3 3.666667
## 5 1101060   33          5 6.600000
## 6 1101061  274         28 9.785714
```

2.2.9 자료와 모집단 정보의 결합

```
data1 <- left_join(data,popinfo,by="AD_CD")
head(data1, n=20)
```

```
## # A tibble: 20 x 7
## # Rowwise:
##    SURV_PHS_SLR_SUM SLE_ALL_AMT MNF_PCST AD_CD    popN samplesize weight
##               <dbl>       <dbl>    <dbl> <fct>   <int>      <dbl>  <dbl>
##  1               14         103       23 1101053   111         12   9.25
##  2               13          40       16 1101053   111         12   9.25
##  3                4          40       15 1101053   111         12   9.25
##  4               12         100       38 1101053   111         12   9.25
##  5                2          20        8 1101053   111         12   9.25
##  6                7          36       14 1101053   111         12   9.25
##  7                1          30        9 1101053   111         12   9.25
##  8                7          15        6 1101053   111         12   9.25
##  9                2          15        6 1101053   111         12   9.25
```

```
## 10          5          56       15 1101053   111          12   9.25
## 11          7          22        8 1101053   111          12   9.25
## 12          9          54       21 1101053   111          12   9.25
## 13         11         100       70 1101053   111          12   9.25
## 14         15          76       24 1101053   111          12   9.25
## 15         13          43       10 1101053   111          12   9.25
## 16          7          43        9 1101053   111          12   9.25
## 17          7          43        9 1101053   111          12   9.25
## 18         13          70       32 1101053   111          12   9.25
## 19         16         100       38 1101053   111          12   9.25
## 20         11          90       35 1101053   111          12   9.25
```

2.2.10 매출액의 지역별 모집단 평균계산

```
popMeanT <- data1 %>% group_by(AD_CD)   %>%
  summarise(SLE_ALL_AMT_mean  = mean(SLE_ALL_AMT, na.rm=TRUE)) %>% as.data.frame()
head(popMeanT)
```

```
##     AD_CD SLE_ALL_AMT_mean
## 1 1101053         474.3514
## 2 1101054         583.4348
## 3 1101056         479.0000
## 4 1101058         254.6364
## 5 1101060         319.1818
## 6 1101061         373.5182
```

2.2.11 표본 추출

```
sample.final = sampling::strata(data1 , 'AD_CD', size=sample.result, method="srswor")
sampledata = data1[sample.final$ID_unit, ]
sampledata <- as.data.frame(sampledata )
head(sampledata, n=20)
```

```
##    SURV_PHS_SLR_SUM SLE_ALL_AMT MNF_PCST   AD_CD popN samplesize weight
## 1                 7          22        8 1101053  111          12   9.25
## 2                 7          43        9 1101053  111          12   9.25
## 3                 1           3        1 1101053  111          12   9.25
## 4                 3          12        5 1101053  111          12   9.25
## 5                43         298       99 1101053  111          12   9.25
## 6                73         298       69 1101053  111          12   9.25
## 7                16         210      120 1101053  111          12   9.25
## 8                90         781      429 1101053  111          12   9.25
## 9               100         600       20 1101053  111          12   9.25
## 10               10          24        7 1101053  111          12   9.25
## 11              192         520      180 1101053  111          12   9.25
## 12              222        1208      733 1101053  111          12   9.25
## 13               61         420      189 1101054   23           4   5.75
## 14              138         800      320 1101054   23           4   5.75
## 15              288        1020      357 1101054   23           4   5.75
## 16              768        2380      360 1101054   23           4   5.75
## 17               20         144       70 1101056   22           4   5.50
## 18                5         117       60 1101056   22           4   5.50
## 19               10         155       30 1101056   22           4   5.50
## 20              151         365      150 1101056   22           4   5.50
```

2.2.12 표본자료와 표본 추출정보 결합

```
strat_design <- svydesign(id = ~1, strata = ~AD_CD, weights = ~weight, data = sampledata)
```

2.2.13 단순 추정량 계산

```
res0 <- survey::svyby(~SLE_ALL_AMT, ~AD_CD, strat_design, svymean)
head(res0, n=20)
```

```
##           AD_CD SLE_ALL_AMT         se
```

```
## 1101053 1101053   334.91667 109.349485
## 1101054 1101054  1155.00000 426.722002
## 1101056 1101056   195.25000  57.143642
## 1101058 1101058    76.66667   8.333333
## 1101060 1101060   260.40000  77.576156
## 1101061 1101061   493.14286  99.047270
## 1101063 1101063   233.57143  39.544343
## 1101064 1101064   243.60000  58.795918
## 1101065 1101065   228.40000  58.707410
## 1101067 1101067   215.25000  81.787708
## 1101068 1101068    75.00000  24.637370
## 1101070 1101070    59.33333  16.895101
## 1101071 1101071   360.00000  83.266640
## 1101072 1101072   756.33333 345.904772
## 1102052 1102052   459.71429 100.088720
## 1102054 1102054   319.90000 149.510271
## 1102055 1102055   682.40000 133.417082
## 1102057 1102057   211.66667  66.352928
## 1102058 1102058   499.00000 235.047158
## 1102059 1102059   389.42857 151.791501
```

2.2.14 보조변수에 대한 층별 모평균 계산

```
popMean <- data1 %>% group_by(AD_CD)   %>%
  summarise(MNF_PCST_mean  = mean(MNF_PCST, na.rm=TRUE),  SURV_PHS_SLR_SUM_mean=mean(SU
head(popMean)
```

```
##     AD_CD MNF_PCST_mean SURV_PHS_SLR_SUM_mean
## 1 1101053      214.2883              85.09009
## 2 1101054      217.2609             144.69565
## 3 1101056      205.8636             121.54545
## 4 1101058      130.5455              33.81818
## 5 1101060      135.4848              76.81818
```

```
## 6 1101061       157.2007              77.51095
```

2.2.15 층별 사업체 수 생성

```
popN <- popinfo[,1:2]
head(popN)
```

```
##     AD_CD popN
## 1 1101053  111
## 2 1101054   23
## 3 1101056   22
## 4 1101058   11
## 5 1101060   33
## 6 1101061  274
```

2.2.16 소지역 모형 적합

```
res <- eblupBHF(SLE_ALL_AMT ~ MNF_PCST + SURV_PHS_SLR_SUM, dom =AD_CD, meanxpop = popMean,
                popnsize = popN, data = sampledata)
```

2.2.17 모수 추정 결과

$$\hat{\sigma}_b^2 = 49.65, \quad \hat{\sigma}_e^2 = 17216.15$$

```
res$fit$summary
```

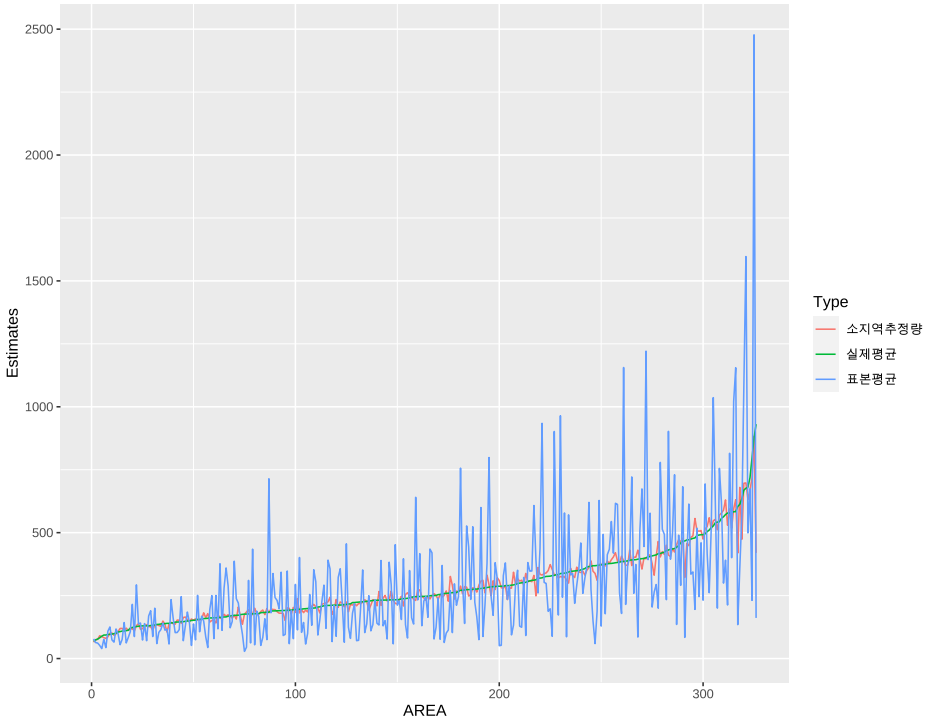
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ys ~ -1 + Xs + (1 | dom)
##
## REML criterion at convergence: 18756.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -6.3258 -0.2029 -0.0063  0.2039 20.8385
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  dom      (Intercept)    49.66   7.047
##  Residual             17216.15 131.210
## Number of obs: 1489, groups:  dom, 326
##
## Fixed effects:
##                     Estimate Std. Error t value
## Xs(Intercept)       5.74487    4.00677   1.434
## XsMNF_PCST          1.24440    0.02233  55.721
## XsSURV_PHS_SLR_SUM  2.45568    0.04966  49.455
##
## Correlation of Fixed Effects:
##             Xs(In) XMNF_P
## XsMNF_PCST  -0.159
## XSURV_PHS_S -0.191 -0.771
```

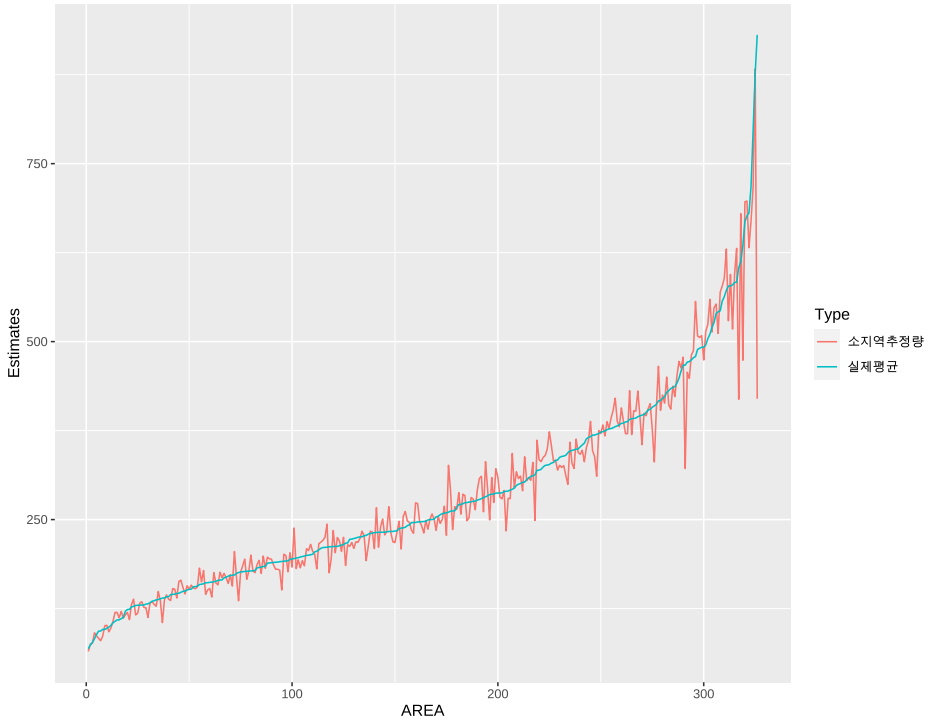
2.2.18 추정량의 비교

```
finalres1 <- popMeanT
finalres1$samplemean <- res0$SLE_ALL_AMT
finalres1$sae <- res$eblup$eblup
finalres1 <- finalres1[order(finalres1$SLE_ALL_AMT_mean), ]
finalres1$area <- 1:(dim(finalres1)[1])
colnames(finalres1) <- c("AD_CD", "실제평균", "표본평균", "소지역추정량", "AREA")
finalres1 <- finalres1[,2:5]
finalres2 <- gather(finalres1,  Type, Estimates,"실제평균", "표본평균", "소지역추정량")
```

```
ggplot(finalres2, aes(x=AREA, y=Estimates, color=Type)) +geom_line()
```




```
finalres3 <- finalres2 %>% filter(Type != "표본평균")
ggplot(finalres3, aes(x=AREA, y=Estimates, color=Type)) +geom_line()
```


750
500
250
Estimates
0
100
200
300
AREA
Type
소지역추정량
실제평균

제 3 장

계층모형

이제 계층모형에 대한 전현적인 예제와 분석을 살펴보고자 한다. 아래에서 사용할 자료와 분석 내용은 교과서 (Finch et al., 2019)를 따른 것이다.

3.1 학생 성취도 자료

학생 선취도 자료는 교과서 (Finch et al., 2019) 의 홈페이지[1]에서 다운로드 받을 수 있다.

```
Achieve <- read.csv("Achieve.csv", header=T)
head(Achieve,n=3)
```

```
##   row id region corp school class gender age race geread gevocab gereadcm
## 1   1  1      2  940    767     1      2 104    5    3.5     3.1      3.2
## 2   2  2      2  940    767     1      2 106    5    1.2     2.8      2.0
## 3   3  3      2  940    767     1      2 112    5    2.1     1.7      1.9
##   gelang gelangmc gelangcm gemath gemathcp gemathcm getotal npanverb npamem
## 1    2.3      4.3      3.0    3.1      2.7      2.9     3.0       16     56
## 2    1.8      1.7      1.6    3.2      4.3      3.8     2.7       39     79
## 3    2.5      2.2      2.4    3.6      3.2      3.4     2.7       37     41
##   npaverb npatotal csi multi clenroll classize ptratio ptia locale chapter1
```

[1] http://www.mlminr.com/

```
## 1      46       26  90     2       16        1      16    2      7        1
## 2      37       47  99     2       16        1      16    2      7        1
## 3      34       35  94     2       16        1      16    2      7        1
##    ses context calender senroll sattend white1 black1 hispanc1 asian1 aindian1
## 1 80.4     929        1     463    95.6   99.8      0        0      0        0
## 2 80.4     929        1     463    95.6   99.8      0        0      0        0
## 3 80.4     929        1     463    95.6   99.8      0        0      0        0
##   multi1 total1 noteach avgage1 avgexp1 avgsal1 spert thrdclss thrdstud passla1
## 1    0.2    0.2    28.5    46.2    20.6   43517  16.3        4       60       63
## 2    0.2    0.2    28.5    46.2    20.6   43517  16.3        4       60       63
## 3    0.2    0.2    28.5    46.2    20.6   43517  16.3        4       60       63
##   passmth1 passbth1 tmnnce1 rmdnce1 lamdnce1 mmdnce1 tmdnce1 avgcsi1 geog
## 1       74       52    60.5    57.6       55      61    59.7      99    2
## 2       74       52    60.5    57.6       55      61    59.7      99    2
## 3       74       52    60.5    57.6       55      61    59.7      99    2
##   totepp cenroll cattend freelnch lep speced minority white2 black2 hispanc2
## 1   5956    3115    95.9       16 0.2   14.4      1.2   3077      1       12
## 2   5956    3115    95.9       16 0.2   14.4      1.2   3077      1       12
## 3   5956    3115    95.9       16 0.2   14.4      1.2   3077      1       12
##   asian2 aindian2 multi2 total2 thrdadm thrdtech avgage2 avgexp2 avgsal2
## 1      3        1     21     38     208       11    45.2    18.4   41096
## 2      3        1     21     38     208       11    45.2    18.4   41096
## 3      3        1     21     38     208       11    45.2    18.4   41096
##   thrdaide passla2 passmth2 passbth2 tmnnce2 rmdnce2 lamdnce2 mmdnce2 tmdnce2
## 1        3      67       70       52    59.4    57.8     59.8    60.6    60.8
## 2        3      67       70       52    59.4    57.8     59.8    60.6    60.8
## 3        3      67       70       52    59.4    57.8     59.8    60.6    60.8
##   rmediate
## 1       18
## 2       18
## 3       18
```

성취도 자료 Achieve는 160개 학교를 추출하고 각 학교마다 작게는 11명, 크게는 162명의 학생들을 추출하여 학교의 여러 가지 정보와 학생들의 다양한 성적을 수집한 자료이다.

성취도 자료 Achieve는 계층적 자료이며 다음과 같은 계층구조를 가지고 있다.

- 계층 1 : 학생
- 계층 2 : 학교

분석의 목적은 학교의 특성과 학생들의 다양한 성적들(예를 들어 어휘 능력, vocabulary scores)이 학생들의 읽기 성취도(general reading achievement)에 어떤 영향을 미치는지 분석하는 것이다.

이제 i 번째 학교에 속한 j 번째 학생의 읽기 성취도 점수를 y_{ij} 라고 하자.

3.2 단순 계층모형

가장 단순한 계층 모형으로서 읽기 성취도 점수에 대하여 학교 school 이 임의효과인 모형을 고려해 보자.

$$y_{ij} = \beta_0 + b_{0i} + e_{ij} \tag{3.1}$$

위의 식에서 β_0 는 전체 평균 점수를 나타내는 모수이며 학교에 대한 임의효과 b_{0i} 와 오차항 e_{ij} 는 서로 독립이며 다음과 같은 분포를 따른다.

$$b_{0i} \sim N(0, \sigma^2_{b0}), \quad e_{ij} \sim N(0, \sigma^2_e)$$

이제 lmer 함수로 모형 (3.1)을 적합시켜 보자.

```
model1 <- lmer(gread~1 +(1|school), data=Achieve)
summary(model1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: gread ~ 1 + (1 | school)
##    Data: Achieve
##
## REML criterion at convergence: 46268.3
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3229 -0.6378 -0.2138  0.2850  3.8812
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 0.3915   0.6257
##  Residual             5.0450   2.2461
## Number of obs: 10320, groups:  school, 160
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)   4.30675    0.05498 158.53888   78.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
equatiomatic::extract_eq(model1, use_coefs = TRUE)
```

$$\begin{aligned} \hat{\text{read}}_i &\sim N\left(4.31_{\alpha_{j[i]}}, \sigma^2\right) \\ \alpha_j &\sim N\left(0, 0.63\right), \text{ for school j} = 1, \dots, \text{J} \end{aligned}$$

추정결과를 보면 학생들의 읽기 성취도의 전체 평의 추정량은 $\hat{\beta}_0 = 4.3067534$ 이다. 임의효과의 분산성분 σ_{b0} 과 오차항의 분산 σ_e 의 추정량은 다음과 같다.

$$\hat{\sigma}_{b0} = 0.6257119, \quad \hat{\sigma}_e = 2.2461096$$

지난 강의에서 언급한 그룹내 상관계수(ICC)의 값을 구해보면 0.072 로서 같은 학교에 속한 학생들의 성적들의 상관계수를 의미한다.

$$\text{ICC} = \frac{\hat{\sigma}^2_{b0}}{\hat{\sigma}^2_{b0} + \hat{\sigma}^2_e} = 0.072$$

3.3 계층 1 설명변수가 있는 모형

이제 학생들의 어휘능력 성적 gevocab를 설명 변수(x_{ij1})로 포함하는 모형을 고려해 보자. 첫 번째 계층의 구성원인 학생들에 대한 성적이므로 **계층 1 설명변수(level 1 covariate)** 라고 부른다. 일단 어휘능력 성적은 고정 효과로서 모집단 전체에 대한 회귀 계수를 나타낸다.

$$y_{ij} = (\beta_0 + b_{0i}) + \beta_1 x_{ij1} + e_{ij} \tag{3.2}$$

```
model11 <- lmer(geread~gevocab +(1|school), data=Achieve)
summary(model11)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: geread ~ gevocab + (1 | school)
##    Data: Achieve
##
## REML criterion at convergence: 43137.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0823 -0.5735 -0.2103  0.3207  4.4334
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 0.09978  0.3159
##  Residual             3.76647  1.9407
## Number of obs: 10320, groups:  school, 160
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 2.023e+00  4.931e-02 7.582e+02   41.03   <2e-16 ***
## gevocab     5.129e-01  8.373e-03 9.801e+03   61.26   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr)
## gevocab -0.758
```

학생들의 어휘능력 성적은 읽기 성취도를 예측하는데 유의한 변수임이 t-검정로 나타난다 (t-통계량 = 61.26). 또한 다음의 추정식과 같이 어휘능력 성적이 1점 증가하면 평균적으로 읽기 성취도은 0.5점 증가한다.

```
equatiomatic::extract_eq(model11, use_coefs = TRUE)
```

$$\hat{\text{geread}}_i \sim N\left(2.02_{\alpha_{j[i]}} + 0.51_{\beta_1}(\text{gevocab}), \sigma^2\right)$$
$$\alpha_j \sim N\left(0, 0.32\right), \text{ for school j} = 1, \dots, \text{J}$$

3.4 계층 2 설명변수가 있는 모형

이제 학교의 규모를 나타내는 등록 학생의 수 senroll를 설명 변수(x_{i2})로 포함하는 모형을 고려해 보자. 두 번째 계층의 구성원인 학교들에 대한 정보이므로 **계층 2 설명변수(level 2 covariate)**라고 부른다. 일단 등록 학생의 수은 고정 효과로서 모집단 전체에 대한 회귀계수를 나타낸다.

$$y_{ij} = (\beta_0 + b_{0i}) + \beta_1 x_{ij1} + \beta_2 x_{i2} + e_{ij} \tag{3.3}$$

```
model12 <- lmer(geread~gevocab +senroll +(1|school), data=Achieve)
summary(model12)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: geread ~ gevocab + senroll + (1 | school)
##    Data: Achieve
##
## REML criterion at convergence: 43152.1
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0834 -0.5729 -0.2103  0.3212  4.4336
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 0.1003   0.3168
##  Residual             3.7665   1.9408
## Number of obs: 10320, groups:  school, 160
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  2.075e+00  1.140e-01  2.373e+02   18.20   <2e-16 ***
## gevocab      5.129e-01  8.373e-03  9.798e+03   61.25   <2e-16 ***
## senroll     -1.026e-04  2.051e-04  1.652e+02   -0.50    0.618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) gevocb
## gevocab -0.327
## senroll -0.901 -0.002
```

학교의 규모는 읽기 성취도를 예측하는데 유의하지 않다(t-통계량의 p-값 =0.618). 다른 변수에 대한 추정값은 거의 변하지 않았다.

```
equatiomatic::extract_eq(model12, use_coefs = TRUE)
```

$$\hat{\text{ge}}\text{read}_i \sim N\left(2.07_{\alpha_{j[i]}} + 0.51_{\beta_1}(\text{gevocab}), \sigma^2\right)$$
$$\alpha_j \sim N\left(0_{\gamma_1^\alpha}(\text{senroll}), 0.32\right), \text{ for school j} = 1, \dots, \text{J}$$

3.5 계층 간의 상호작용

계층 모형에서는 서로 다른 계층의 설명 변수이 상호 작용을 가지는 경우가 매우 중요한 이슈이다. 또한 같은 계층 안에 속하는 변수들의 상호 작용도 중요하다.

이제 다음과 같은 두 모형을 고려한다.

- 계층 내 상호작용이 있는 모형: 학생의 어휘성적과 연령
- 계층 간 상호작용이 있는 모형: 학생의 어휘성적과 학교의 규모

3.5.1 계층 내 상호작용

```
model21 <- lmer(geread~gevocab + age + gevocab*age + (1|school), data=Achieve)
summary(model21)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: geread ~ gevocab + age + gevocab * age + (1 | school)
##    Data: Achieve
##
## REML criterion at convergence: 43143.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0635 -0.5706 -0.2108  0.3191  4.4467
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 0.09875  0.3143
##  Residual             3.76247  1.9397
## Number of obs: 10320, groups:  school, 160
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  5.187e+00  8.668e-01  1.031e+04   5.984 2.24e-09 ***
```

```
## gevocab      -2.808e-02  1.881e-01  1.030e+04  -0.149 0.881373
## age          -2.937e-02  8.035e-03  1.031e+04  -3.655 0.000258 ***
## gevocab:age   5.027e-03  1.750e-03  1.030e+04   2.873 0.004072 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) gevocb age
## gevocab      -0.879
## age          -0.998  0.879
## gevocab:age   0.877 -0.999 -0.879
```

학생의 연령과 어휘 능력과의 상호작용은 읽기 성취도를 예측하는데 유의하다. 다른 변수에 대한 추정값은 거의 변하지 않았다. 반면 어휘 능력의 유의성은 사라졌다. 이러한 결과는 연령에 따라서 어휘 능력의 기여도가 달라진다는 것을 의미하며 연령이 증가하면 어휘능력의 효과가 커진다.

```
equatiomatic::extract_eq(model21, use_coefs = TRUE)
```

$$
\begin{aligned}
\hat{\text{geread}}_i &\sim N\left(\mu, \sigma^2\right) \\
\mu &= 5.19_{\alpha_{j[i]}} - 0.03_{\beta_1}(\text{gevocab}) - 0.03_{\beta_2}(\text{age}) + 0.01_{\beta_3}(\text{age} \times \text{gevocab}) \\
\alpha_j &\sim N\left(0, 0.31\right), \text{for school j} = 1, \dots, \text{J}
\end{aligned}
$$

3.5.2 계층 간 상호작용

```
model22 <- lmer(geread~gevocab + senroll + gevocab*senroll + (1|school), data=Achieve)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
summary(model22)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: geread ~ gevocab + senroll + gevocab * senroll + (1 | school)
##    Data: Achieve
##
## REML criterion at convergence: 43163.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1228 -0.5697 -0.2090  0.3188  4.4359
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 0.1002   0.3165
##  Residual             3.7646   1.9403
## Number of obs: 10320, groups:  school, 160
##
## Fixed effects:
##                    Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)      1.748e+00  1.727e-01  1.058e+03  10.118   <2e-16 ***
## gevocab          5.851e-01  2.986e-02  9.766e+03  19.592   <2e-16 ***
## senroll          5.121e-04  3.186e-04  8.402e+02   1.607   0.1084
## gevocab:senroll -1.356e-04  5.379e-05  9.849e+03  -2.520   0.0118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) gevocb senrll
## gevocab     -0.782
## senroll     -0.958  0.735
## gevcb:snrll  0.752 -0.960 -0.766
```

```
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

학교의 규모와 어휘 능력과의 상호작용은 읽기 성취도를 예측하는데 유의하다. 또한 어휘 능력도 유의하다. 이러한 결과는 학교의 규모에 따라서 어휘 능력의 기여도가 달라진다는 것을 의미하며 학교의 규모가 커지면 어휘능력의 효과가 감소한다(buffering or inhibitory effect).

```
equatiomatic::extract_eq(model22, use_coefs = TRUE)
```

$$\hat{\text{gereadi}}_i \sim N\left(1.75_{\alpha_{j[i]}} + 0.59_{\beta_1}(\text{gevocab}), \sigma^2\right)$$
$$\alpha_j \sim N\left(0_{\gamma_1^\alpha}(\text{senroll}) + 0_{\gamma_2^\alpha}(\text{gevocab} \times \text{senroll}), 0.32\right), \text{ for school j} = 1, \ldots, \text{J}$$

3.6 임의계수 모형

이제 어휘능력에도 학교에 대한 임의효과가 들어가는 임의계수 모형을 고려해 보자.

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})x_{ij1} + e_{ij} \tag{3.4}$$

```
model3 <- lmer(geread~gevocab +  (1+gevocab |school), data=Achieve)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0191462 (tol = 0.002, component 1)
```

```
summary(model3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: geread ~ gevocab + (1 + gevocab | school)
##    Data: Achieve
##
## REML criterion at convergence: 42992.9
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7096 -0.5674 -0.2079  0.3177  4.6765
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  school   (Intercept) 0.28050  0.5296
##           gevocab     0.01922  0.1386   -0.86
##  Residual             3.66613  1.9147
## Number of obs: 10320, groups:  school, 160
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)   2.00575    0.06097 154.62243   32.90   <2e-16 ***
## gevocab       0.52032    0.01440 145.40535   36.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr)
## gevocab -0.866
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.0191462 (tol = 0.002, component 1)
```

위의 적합된 결과를 보면 어휘능력은 유의한 설명변수이다. 이제 학생들의 읽기 성취도에 대한 총변동의 분해를 보면 다음과 같다.

$$Var(y_{ij}) = \hat{\sigma}^2_{b0} + \hat{\sigma}^2_{b1} + \hat{\sigma}^2_e = 0.28 + 0.02 + 3.67$$

학교의 변동을 설명하는임의효과에 대한 분산성분 σ^2_{b0} 과 σ^2_{b1} 의 추정치는 각각 0.28과 0.02로서 학생 개인들의 변동에 대한 분산 σ^2_e 의 추정치 3.67에 비하여 매우 작다. 따라서 학생들의 읽기 성취도는 학교 요인보다 학생들의 개인 요인이 더 크게 기여한다.

```
equatiomatic::extract_eq(model3, use_coefs = TRUE)
```

$$\hat{\text{geread}}_i \sim N\left(2.01_{\alpha_{j[i]}} + 0.52_{\beta_{1j[i]}}(\text{gevocab}), \sigma^2\right)$$
$$\begin{pmatrix} \alpha_j \\ \beta_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.53 & -0.86 \\ -0.86 & 0.14 \end{pmatrix}\right), \text{for school j} = 1, \dots, \text{J}$$

참고로 모형 (3.4)를 lmer 로 적합할 때 계산에 대한 경고가 나타났다. 이와 같은 계산에 대한 경고는 계층모형에서 매우 흔하게 나타난다. 모형이 너무 복잡하여 계산에 문제가 있거나 계산에서 사용되는 여러 가지 조건이 충분하지 않아서 발생한다. 이러한 경고가 나오면 추정 결과를 면밀하게 검토하고 다른 모형들에 대한 고려도 해야 한다.

이제 두 임의효과가 독립인 축소모형을 적합해 보자.

```
model31 <- lmer(geread~gevocab +  (1+gevocab || school), data=Achieve)
summary(model31)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: geread ~ gevocab + (1 + gevocab || school)
##    Data: Achieve
##
## REML criterion at convergence: 43045.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3662 -0.5719 -0.2095  0.3280  4.4650
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 0.03646  0.19096
##  school.1 gevocab     0.00571  0.07556
##  Residual             3.70688  1.92533
## Number of obs: 10320, groups:  school, 160
```

```
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)   2.03042    0.04543 490.48961   44.70   <2e-16 ***
## gevocab       0.50979    0.01069 441.29210   47.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr)
## gevocab -0.688
```

```
equatiomatic::extract_eq(model31, use_coefs = TRUE)
```

$$\hat{\text{geread}}_i \sim N\left(2.03_{\alpha_{j[i]}} + 0.51_{\beta_{1j[i]}}(\text{gevocab}), \sigma^2\right)$$

$$\begin{pmatrix} \alpha_j \\ \beta_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.19 & 0 \\ 0 & 0.08 \end{pmatrix}\right), \text{for school j} = 1, \dots, \text{J}$$

두 임의효과가 독립이 아닌 경우 나타난 경고는 나오지 않는다. 또한 고정효과에 대한 결과는 변하지 않았으나 분산성분의 추정은 다소 변화가 있다.

3.7 가장 복잡한 모형

이제 위의 결과를 이용하여 연령과 학교규모를 고정효과로 보고 상호작용도 추가함 모형을 살펴보자. 임의효과는 학교에 대한 항만 고려한다.

```
model4 <- lmer(geread~gevocab + age + senroll + gevocab*senroll  + gevocab*age +  (1| s
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
summary(model4)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: geread ~ gevocab + age + senroll + gevocab * senroll + gevocab *
##     age + (1 | school)
##    Data: Achieve
##
## REML criterion at convergence: 43169.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1294 -0.5687 -0.2121  0.3174  4.4477
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 0.0991   0.3148
##  Residual             3.7606   1.9392
## Number of obs: 10320, groups:  school, 160
##
## Fixed effects:
##                   Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)      4.918e+00  8.816e-01  1.024e+04   5.578 2.49e-08 ***
## gevocab          4.164e-02  1.901e-01  1.031e+04   0.219 0.826611
## age             -2.944e-02  8.033e-03  1.031e+04  -3.665 0.000249 ***
## senroll          5.150e-04  3.181e-04  8.412e+02   1.619 0.105762
## gevocab:senroll -1.361e-04  5.376e-05  9.840e+03  -2.532 0.011362 *
## gevocab:age      5.053e-03  1.749e-03  1.029e+04   2.889 0.003876 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) gevocb age    senrll gvcb:s
```

```
## gevocab     -0.875
## age         -0.981  0.869
## senroll     -0.184  0.110 -0.004
## gevcb:snrll  0.143 -0.145  0.004 -0.766
## gevocab:age  0.861 -0.988 -0.879  0.005 -0.006
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

```
equatiomatic::extract_eq(model4, use_coefs = TRUE)
```

$$\hat{\text{geread}}_i \sim N\left(\mu, \sigma^2\right)$$

$$\mu = 4.92_{\alpha_{j[i]}} + 0.04_{\beta_1}(\text{gevocab}) - 0.03_{\beta_2}(\text{age}) + 0.01_{\beta_1}(\text{age} \times \text{gevocab})$$

$$\alpha_j \sim N\left(0_{\gamma_1^\alpha}(\text{senroll}) + 0_{\gamma_2^\alpha}(\text{gevocab} \times \text{senroll}), 0.31\right), \text{ for school j} = 1, \ldots, \text{J}$$

3.8 설명변수의 중심화

계층모형에서는 설명 변수를 중심화하는 것(centering, 변수의 평균이 0)이 모형의 결과를 해석하는데 편리하다. 이유는 변수들의 효과가 전체 평균을 기준으로 높거나 낮은 경향으로 나타나므로 해석이 용이하다.

이제 계층 1 의 설명변수들인 연령과 어휘능력을 중심화하여 다시 가장 복잡한 모형을 적합해보자.

```
Achieve$Cgevocab <- Achieve$gevocab - mean(Achieve$gevocab)
Achieve$Cage <- Achieve$age - mean(Achieve$age)
model5 <- lmer(geread~Cgevocab + Cage + senroll + Cgevocab*senroll  + Cgevocab*Cage +
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
summary(model5)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: geread ~ Cgevocab + Cage + senroll + Cgevocab * senroll + Cgevocab *
##     Cage + (1 | school)
##    Data: Achieve
##
## REML criterion at convergence: 43169.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1294 -0.5687 -0.2121  0.3174  4.4477
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 0.0991   0.3148
##  Residual             3.7606   1.9392
## Number of obs: 10320, groups:  school, 160
##
## Fixed effects:
##                     Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)       4.381e+00  1.074e-01  1.927e+02  40.807  < 2e-16 ***
## Cgevocab          5.850e-01  2.985e-02  9.762e+03  19.597  < 2e-16 ***
## Cage             -6.733e-03  3.916e-03  1.031e+04  -1.719  0.08561 .
## senroll          -9.663e-05  2.043e-04  1.653e+02  -0.473  0.63682
## Cgevocab:senroll -1.361e-04  5.376e-05  9.840e+03  -2.532  0.01136 *
## Cgevocab:Cage     5.053e-03  1.749e-03  1.029e+04   2.889  0.00388 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Cgevcb Cage   senrll Cgvcb:
```

```
## Cgevocab    -0.008
## Cage         0.002  0.019
## senroll     -0.954  0.010  0.000
## Cgvcb:snrll  0.009 -0.960 -0.004 -0.011
## Cgevocab:Cg  0.012  0.012  0.205  0.001 -0.006
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

```
equatiomatic::extract_eq(model5, use_coefs = TRUE)
```

$$
\begin{aligned}
\hat{\text{gereading}}_{i} &\sim N\left(\mu, \sigma^2\right) \\
\mu &= 4.38_{\alpha_{j[i]}} + 0.59_{\beta_1}(\text{Cgevocab}) - 0.01_{\beta_2}(\text{Cage}) + 0.01_{\beta_1}(\text{Cage} \times \text{Cgevocab}) \\
\alpha_j &\sim N\left(0_{\gamma_1^{\alpha}}(\text{senroll}) + 0_{\gamma_2^{\alpha}}(\text{Cgevocab} \times \text{senroll}), 0.31\right), \text{ for school j} = 1, \dots, \text{J}
\end{aligned}
$$

참고 문헌

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Finch, W. H., Bolin, J. E., and Kelley, K. (2019). *Multilevel modeling using R*. Crc Press.

Molina, I. and Marhuenda, Y. (2015). sae: An r package for small area estimation. *R J.*, 7(1):81.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2):377–387.