선형혼합모형 II 반복측정자료

서울시립대 통계학과 이용희

FALL 2019

1 반복측정자료

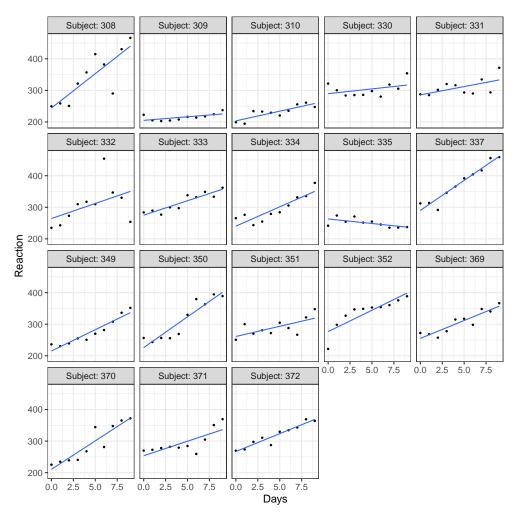
1.1 반복측정자료의 개요와 예제

반복측정자료(longitudinal data, repeated measurements)는 관측단위안에서 여러 개의 관측값을 측정한 자료의 형식을 말한다. 예를 들어 환자가 병원을 여러 번 방문하고 방문시마다 혈압을 측정하였다면한 명의 환자에서 반복 측정한 자료는 서로 독립이 아니다. 또한 가구조사(household survey)에서 가구원의 취업 여부, 건강 상태등을 여러 해동안 매년 측정하는 경우 이러한 자료를 패널자료(panel data) 또는 longitudinal 자료라고 한다. 이렇게 하나의 관측단위 안에서 측정한 자료들은 서로 독립이 아닌특징이 있고 자료를 분석하는 경우 이러한 자료들의 종속구조를 고려하는 모형을 사용하는 것이 적절하다. 이렇게 반복측정자료에서 반복자료들의 공분산구조를 설정하는 통계적 방법들은 다양하지만대표적으로 쉽게 사용할 수 있는 방법이 임의효과를 포함한 혼합모형을 사용하는 방법이다.

에제 1.1 (Sleep study). lme4 패키지에 자료인 spleepstudy는 화물트럭 운전사들에 대한 수면부족 현상에 대하여 연구한 자료이다. l8명의 운전자들이 매일 3시간의 수면(부족한 수면)을 하면서 매일 일정한 동작의 반응시간을 l0일동안 반복적으로 측정한 자료가 있다. 한명의 운전사에게 l0일 동안의 반응에 대한 측정자료 l0개가 존재하므로 이는 반복측정 자료이며 이러한 l0개의 자료는 독립이 아니다. 일단 자료의 구조를 살펴보자. 반응변수 l00 Reaction은 반응시간(l01 분 나타내며 설명변수로서 l02 l03 가요하는 날짜(l02 l03 l03 l04 가요하는 운전자의 고유번호를 나타낸다.

```
library(lme4)
## Loading required package: Matrix
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
         249.6
                  0
                        308
## 2
         258.7
                 1
                        308
## 3
        250.8
                        308
## 4
        321.4
                        308
## 5
        356.9
                 4
                        308
## 6
        414.7
                        308
                 5
## 7
         382.2
                        308
## 8
         290.1
                        308
                  7
## 9
         430.6
                        308
                  8
## 10
         466.4
                        308
## 11
         222.7
                        309
## 12
         205.3
                  1
                        309
         203.0
                        309
## 13
                  2
         204.7
                        309
## 14
                  3
                        309
## 15
         207.7
                  4
## 16
         216.0
                  5
                        309
## 17
         213.6
                        309
## 18
         217.7
                  7
                        309
## 19
         224.3
                  8
                        309
## 20
         237.3
                        309
```

각 운전자에 대한 10일 간의 반응속도가 시간에 따라 어떻게 변하는 가를 알아보자. 전반적으로 시간이 지나면서 운전자들의 반응시간이 증가하고 있음을 알 수 있다. 또한 개인 별로 반응 시간의 변화와 패턴이 다르다는 것을 알 수 있다.



1.2 회귀직선모형: 개체와 전체

각 운전자 i 에 대하여 10일간 측정한 반응속도 y_{ij} 를 시간에 대하여 선형모형으로 적합하면 개인별 회귀직선을 다음과 같이 표시할 수 있다.

$$y_{ij} = \beta_{0i} + \beta_{1i}t_j + e_{ij}, \quad i = 1, 2, \dots, 18, \quad j = 1, 2, \dots, 10$$
 (1)

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

행렬식으로는 다음과 같이 나타낼 수 있다.

$$y_i = X_i \beta_i + e_i$$

여기서

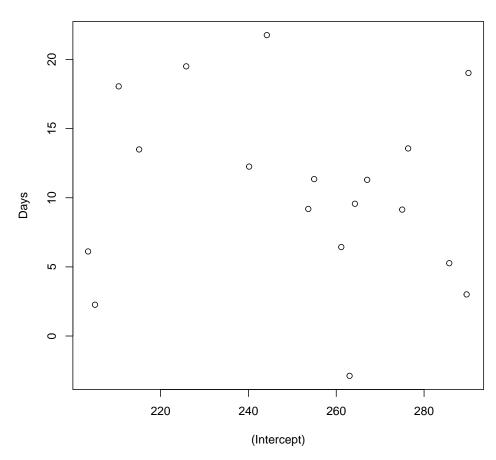
$$oldsymbol{y}_i = \left[egin{array}{c} y_{i1} \ y_{i2} \ dots \ y_{i,10} \end{array}
ight], \; oldsymbol{X}_i = \left[egin{array}{c} 1 & 0 \ 1 & 1 \ dots \ 1 & 9 \end{array}
ight], oldsymbol{eta}_i = \left[egin{array}{c} eta_{0i} \ eta_{1i} \end{array}
ight], oldsymbol{e}_i = \left[egin{array}{c} e_{i1} \ e_{i2} \ dots \ e_{i,10} \end{array}
ight]$$

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 1mList를 아래와 같이 이용하면 식 (1)을 각 운전사마다 적합시켜 각각의 절편과 기울기를 구할 수 있다.

```
lmf1 <- lmList(Reaction ~ Days | Subject, sleepstudy)</pre>
lmf1
## Call: lmList(formula = Reaction ~ Days | Subject, data = sleepstudy)
## Coefficients:
##
       (Intercept)
## 308
             244.2 21.765
## 309
             205.1 2.262
## 310
             203.5 6.115
## 330
             289.7 3.008
## 331
             285.7 5.266
## 332
             264.3 9.567
```

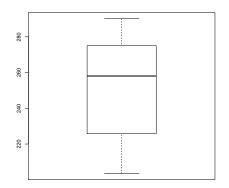
```
## 333
            275.0 9.142
## 334
            240.2 12.253
            263.0 -2.881
## 335
            290.1 19.026
## 337
## 349
            215.1 13.494
## 350
            225.8 19.504
            261.1 6.433
## 351
         276.4 13.567
## 352
         255.0 11.348
## 369
         210.4 18.056
## 370
## 371
          253.6 9.188
           267.0 11.298
## 372
##
## Degrees of freedom: 180 total; 144 residual
## Residual standard error: 25.59
cor(coef(lmf1))
##
             (Intercept) Days
## (Intercept) 1.0000 -0.1376
## Days
                 -0.1376 1.0000
plot(coef(lmf1),main="intercepts and slopes on drivers: sleep study ")
```

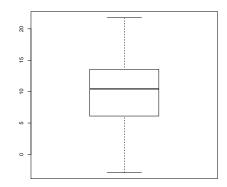
intercepts and slopes on drivers: sleep study



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])
boxplot(coef(lmf1)[2])
```





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.

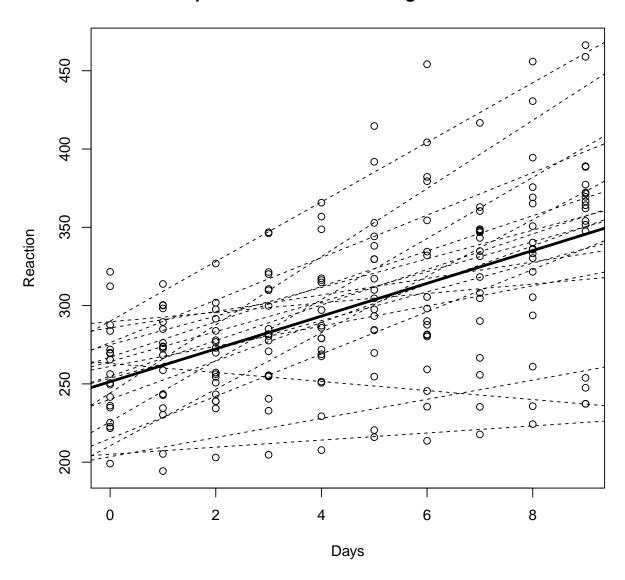
$$y_{ij} = \beta_0 + \beta_1 t_j + e_{ij}, \quad i = 1, 2, \dots, 18, j = 1, 2, \dots, 10$$
 (2)

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

위와 같은 전체 운전사 집단의 관측값을 운전자의 특성을 고려하지 않고 세운 모형으로서 시간에 따른 반응시간에 대한 모집단의 전체적인 평균적 함수 관계를 파악하는 모형이라고 할 수 있다.

```
lmpop <- lm(Reaction ~ Days, sleepstudy)</pre>
summary(lmpop)
##
## Call:
## lm(formula = Reaction ~ Days, data = sleepstudy)
##
## Residuals:
##
      Min
               1Q Median 3Q
                                     Max
## -110.85 -27.48 1.55 26.14 139.95
##
## Coefficients:
              Estimate Std. Error t value Pr(>|t|)
##
## (Intercept) 251.41 6.61 38.03 < 2e-16 ***
```

Population and individual regression lines



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

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

## (Intercept) Days
## 251.41 10.47
```

```
coef(lmpop)

## (Intercept) Days
## 251.41 10.47
```

1.3 선형혼합모형

앞 절의 모형과 분석에서 알 수 있듯이 한 개체에 대하여 여러 개의 관측값을 측정한 자료에 회귀방 정식을 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 알수 있다.
- 전체 자료에 적합한 모형을 보면 개인별로 적합한 모형의 공통적인 성격, 즉 시간에 따른 반응시 간의 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.

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_j + e_{ij}$$
(3)

모형 (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 반영한확률변수이다. 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. 임의효과는 모집단을 구성하는 개인이 표본에 추출되었다고 생각하며 확률분포를 따른다고가정한다. 반복측정자료에서 인의효과를 공통으로 가지고 있는 관측치는 독립이 아니게 돼며 따라서같은 개체에서 나온 관측값은 독립이 아니다.

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

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

위의 모형은 각 개인의 회귀직선에서 각 절편과 기울기가 전체평균 β_0 와 β_1 를 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 , ρ 이다.

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

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

위와 같은 혼합효과모형(mixed effects model)을 각 개인 i에 대하여 행렬식으로 표시하면 다음과 같다.

$$y_i = X_i \beta + Z_i b_i + e_i$$

여기서

$$oldsymbol{y}_i = \left[egin{array}{c} y_{i1} \ y_{i2} \ dots \ y_{i,10} \end{array}
ight], \; oldsymbol{X}_i = \left[egin{array}{c} 1 & 0 \ 1 & 1 \ dots \ 1 & 9 \end{array}
ight], \; oldsymbol{Z}_i = \left[egin{array}{c} 1 & 0 \ 1 & 1 \ dots \ dots \ 1 & 9 \end{array}
ight], \; oldsymbol{b}_i = \left[egin{array}{c} b_{0i} \ b_{1i} \end{array}
ight], \; oldsymbol{e}_i = \left[egin{array}{c} e_{i1} \ e_{i2} \ dots \ e_{i,10} \end{array}
ight]$$

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

$$y = X\beta + Zb + e \tag{4}$$

여기서 반응변수벡터 y와 고정효과 β 에 대한 계획행렬 X는 각 개인의 반응변수벡터 y_i 와 X_i 를 행으로 쌓아놓은 것으로 표현된다. 오차항에 대한 벡터 e도 동일한 형식의 벡터이다.

$$egin{aligned} oldsymbol{y}_i = \left[egin{array}{c|c} oldsymbol{y}_1 \ oldsymbol{y}_2 \ dots \ oldsymbol{y}_{18} \end{array}
ight], egin{array}{c|c} oldsymbol{X}_1 \ oldsymbol{X}_2 \ dots \ oldsymbol{z} \ oldsymbol{e} \ oldsymbol{e} \ oldsymbol{z} \$$

임의효과 벡터 b는 각 개인에 대한 임의효과벡터 b_i 를 행으로 쌓아놓은것과 같고 임의효과에 대한 계획행렬 Z는 각 개인의 계획행렬 Z_i 를 대각원소로 같은 행렬이다.

$$m{b} = \left[egin{array}{c} m{b}_1 \ m{b}_2 \ dots \ m{b}_{18} \end{array}
ight], \; m{Z} = \left[egin{array}{cccc} m{Z}_1 & 0 & \dots & 0 \ 0 & m{Z}_2 & \dots & 0 \ dots & dots & dots & dots \ 0 & 0 & \dots & m{Z}_{18} \end{array}
ight]$$

혼합모형 (4)를 lmer() 함수를 이용하여 적합시켜보자. 모형에서 (1 + Days|Subject) 이 개체에 대하여 절편과 기울기에 대한 임의효과를 지정한다.

```
fm1 <- lmer(Reaction ~ 1 + Days + (1 + Days|Subject), sleepstudy)
summary(fm1)
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
##
     Data: sleepstudy
##
## REML criterion at convergence: 1744
##
## Scaled residuals:
         1Q Median 3Q
     Min
## -3.954 -0.463 0.023 0.463 5.179
##
## Random effects:
   Groups
          Name
                   Variance Std.Dev. Corr
   Subject (Intercept) 611.9 24.74
##
##
            Days
                       35.1 5.92
                                      0.07
  Residual
                       654.9 25.59
## Number of obs: 180, groups: Subject, 18
##
## Fixed effects:
              Estimate Std. Error t value
```

```
## (Intercept) 251.41
                                     36.84
                              6.82
## Days
                 10.47
                              1.55
                                     6.77
##
## Correlation of Fixed Effects:
        (Intr)
## Days -0.138
methods(class="merMod")
   [1] anova
                      as.function
                                    coef
                                                     confint
                                                                    cooks.distance
  [6] deviance
                      df.residual
                                    drop1
                                                     extractAIC
                                                                    family
## [11] fitted
                      fixef
                                      formula
                                                     fortify
                                                                    getL
## [16] getME
                      hatvalues
                                     influence
                                                     isGLMM
                                                                    isLMM
## [21] isNLMM
                      isREML
                                                     model.frame
                                                                    model.matrix
                                      logLik
## [26] ngrps
                      nobs
                                      plot
                                                     predict
                                                                    print
## [31] profile
                      qqmath
                                      ranef
                                                     refit
                                                                    refitML
## [36] rePCA
                      residuals
                                      rstudent
                                                     show
                                                                    sigma
## [41] simulate
                       summary
                                                     update
                                                                    VarCorr
                                      terms
## [46] vcov
                      weights
## see '?methods' for accessing help and source code
```

위에서 method()는 선형혼합모형을 적함시킨 추정 결과를 보여주는 함수들이다. 위의 혼합모형 적합결과를 살펴보자. 첫째로 고정효과에 대한 추정식은 다음과 같다

```
fixef(fm1)

## (Intercept) Days
## 251.41 10.47
```

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

```
VarCorr(fm1)

## Groups Name Std.Dev. Corr

## Subject (Intercept) 24.74

## Days 5.92 0.07

## Residual 25.59
```

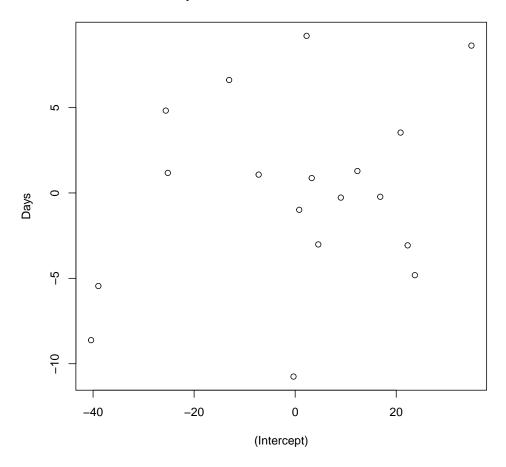
이제 임의효과 b에 대한 예측(prediction)을 생각해보자. 우리는 오직 관측벡터 y만을 관측하고 임의효과 b는 관측을 할 수 없는 확률변수이다. 하지만 주어진 관측벡터와 추정된 분산으로 임의효과의 값을 예측할 수있으며 그 결과는 다음과 같다.

```
re <- ranef(fm1)$Subject
re
##
      (Intercept) Days
## 308
         2.2575 9.1993
      -40.3943 -8.6205
## 309
## 310
      -38.9564 -5.4496
        23.6889 -4.8141
## 330
## 331
       22.2585 -3.0697
        9.0388 -0.2721
## 332
## 333
      16.8390 -0.2234
## 334
        -7.2320
                 1.0745
## 335
        -0.3327 -10.7525
## 337
         34.8865 8.6290
## 349
         -25.2080 1.1731
## 350
         -13.0694
                 6.6142
## 351
         4.5777 -3.0153
## 352
          20.8615 3.5364
## 369
         3.2751 0.8723
         -25.6111 4.8223
## 370
## 371
        0.8071 -0.9882
```

```
## 372 12.3133 1.2842

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

prediction of random effects



예측된 각 개인의 절편과 기울기에 대한 임의효과 b_{0i} 과 b_{1i} 에 고정효과의 추정량 $\hat{\beta}$ 를 더해주면 각 개인의 절편과 기울기에 대한 예측값을 구할 수 있다.

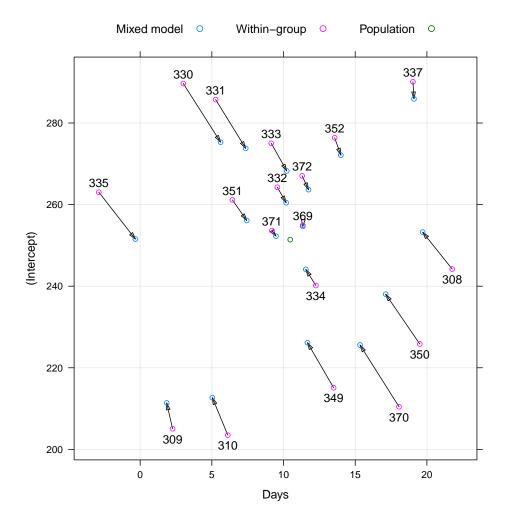
```
## 330
            275.1 5.6531
## 331
            273.7 7.3976
## 332
            260.4 10.1952
            268.2 10.2439
## 333
## 334
            244.2 11.5418
            251.1 -0.2852
## 335
            286.3 19.0963
## 337
## 349
            226.2 11.6404
## 350
           238.3 17.0815
## 351
           256.0 7.4520
## 352
           272.3 14.0037
## 369
           254.7 11.3396
           225.8 15.2895
## 370
## 371
            252.2 9.4791
## 372
            263.7 11.7515
```

위의 결과를 각 개체에 대해 별도의 회귀직선을 적합시켜서 얻은 18개의 절편과 기울기와 비교해 보자.

```
coef(lmf1)
      (Intercept) Days
## 308
            244.2 21.765
## 309
       205.1 2.262
         203.5 6.115
## 310
## 330
          289.7 3.008
## 331
           285.7 5.266
## 332
            264.3 9.567
## 333
            275.0 9.142
## 334
            240.2 12.253
## 335
            263.0 -2.881
            290.1 19.026
## 337
```

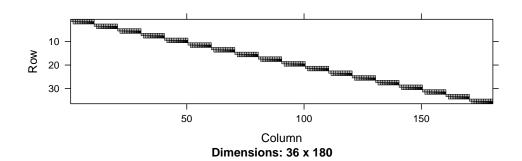
```
## 349
            215.1 13.494
            225.8 19.504
## 350
## 351
            261.1 6.433
## 352
            276.4 13.567
## 369
            255.0 11.348
## 370
            210.4 18.056
            253.6 9.188
## 371
## 372
            267.0 11.298
```

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



여기서 혼합모형의 식 (4)의 임의효과에 대한 계획행렬 Z의 구조를 살펴보자. R의 적합된 결과에서 getME함수를 이용하여 계획행렬 Z의 전치행렬(transpose matrix, Z^t)을 얻을 수 있다. 계획행렬 Z는 그 값의 많은 부분이 0으로 구성되어 있어서 성김행렬(saprse matrix)라고 부르며 이런 행렬은 특별한 형식으로 저장되어 있다.

image(getME(fm1,"Zt"))



또한 혼합모형의 식 (4)에서 임의효과의 상관계수가 $\rho = 0$ 인 경우의 모형을 고려해 보자

$$\left[\begin{array}{c}b_{0i}\\b_{1i}\end{array}\right] \sim N\left(\left[\begin{array}{c}0\\0\end{array}\right], \left[\begin{array}{cc}\sigma_{b1}^2&0\\0&\sigma_{b2}^2\end{array}\right]\right)$$

이러한 모형을 아래와 같이 적합시키면 추정결과는 다음과 같다.

```
fm2 <- lmer(Reaction ~ 1 + Days + (1|Subject) + (0+Days|Subject), sleepstudy)
summary(fm2)

## Linear mixed model fit by REML ['lmerMod']

## Formula: Reaction ~ 1 + Days + (1 | Subject) + (0 + Days | Subject)

## Data: sleepstudy
##</pre>
```

```
## REML criterion at convergence: 1744
## Scaled residuals:
## Min 1Q Median 3Q Max
## -3.963 -0.463 0.020 0.465 5.186
## Random effects:
## Groups Name Variance Std.Dev.
## Subject (Intercept) 627.5 25.05
## Subject.1 Days 35.9 5.99
## Residual
                    653.6 25.57
## Number of obs: 180, groups: Subject, 18
##
## Fixed effects:
       Estimate Std. Error t value
## (Intercept) 251.41 6.89 36.51
## Days 10.47 1.56 6.71
##
## Correlation of Fixed Effects:
## (Intr)
## Days -0.184
```