선형혼합모형 IV - 제한적 최대가능도 추정법과 고정 효과에 대한 검정

서울시립대 통계학과 이용희

FALL 2019 학부

1 제한적 최대가능도 추정법의 기초

일변량 정규분포에서 평균과 분산의 추정량을 구하는 방법에 대하여 생각해 보자. $x_1, x_2, \ldots, x_n$은 서로 독립이며 정규분포 $N(\mu, \sigma^2)$를 따른다고 하자. 이제 평균과 분산에 대한 최대가능도 추정량을 구하기 위하여 로그가능도함수를 다음과 같이 구할 수 있다.

$$\begin{aligned} l(\mu, \sigma^2|\boldsymbol{x}) &= \log l(\mu, \sigma^2|\boldsymbol{x}) \\ &= \log\left\{\prod_{i=1}^{n} f(x_i|\mu, \sigma^2)\right\} \\ &= \log\left\{\prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\right\} \\ &= \log\left\{(2\pi\sigma^2)^{-n/2} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right]\right\} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

이제 위의 로그가능도 함수를 최대로 하는 추정량은 다음과 같이 쉽게 구할 수 있다.

$$\hat{\mu}_{ML} = \bar{x}, \quad \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \tag{1}$$

여기서 S^2을 다음과 같이 정의하면

$$S^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1} \tag{2}$$

정규분포 이론에 따라서 $\bar{x}$와 S^2은 서로 독립이고 다음과 같은 분포를 얻는다.

$$\bar{x} \sim N(\mu, \sigma^2/n), \quad \frac{\sum_i (x_i - \bar{x})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

따라서 분산에 대한 최대가능도 추정량 $\hat{\sigma}^2_{ML}$은 편이 추정량(biased estimator)이고

$$E(\hat{\sigma}^2_{ML}) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

S^2은 불편추정량(unbiased estimator)이다.

$$E(S^2) = \sigma^2$$

이렇게 분산에 대한 추정에서는 최대 가능도 추정법을 이용하면 편이(bias)를 가지는 추정량을 얻는다. 이러한 최대 가능도 추정법의 단점을 보완하기 위한 방법으로 제한적 최대 가능도 추정법(Restricted Maximum likelihood estimation; REML)이 있다. REML 방법은 평균과 분산에 대한 가능도함수를 평균에 대한 부분을 없애고 오직 분산에 대한 가능도 함수만을 고려하는 방법이다. 가능도함수에서 평균 μ에 대한 부분을 제거하는 방법은 평균의 분포가 $N(\mu, \sigma^2/n)$을 따르는 사실을 이용하여 μ를 포함한 부분을 적분으로 제거하는 방법이다. 다시 말하면 REML 방법은 분산 σ^2 만을 포함하는 가능도함수를 유도하여 이를 이용하여 분산의 추정량을 구하는 방법이다. 분산 σ^2 에 대한 제한적 가능도 함수(Restricted likelihood function)을 $L_R(\sigma^2|\boldsymbol{x})$라고 정의하면 다음과 같이 정의된다.

$$L_R(\sigma^2|\boldsymbol{x}) = \int L(\mu, \sigma^2|\boldsymbol{x})\ d\mu \tag{3}$$

이제 위의 제한적 가능도 함수를 구해보면 다음과 같이 유도된다.

$$\begin{aligned}
L_R(\sigma^2|\boldsymbol{x}) &= \int L(\mu, \sigma^2|\boldsymbol{x})\ d\mu \\
&= \int (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right]\ d\mu \\
&= (2\pi\sigma^2)^{-n/2} \int \exp\left[-\frac{\sum_{i=1}^n (x_i-\bar{x}+\bar{x}-\mu)^2}{2\sigma^2}\right]\ d\mu \\
&= (2\pi\sigma^2)^{-n/2} \int \exp\left[-\frac{\sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right]\ d\mu \\
&= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{2\sigma^2}\right] \int \exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right]\ d\mu \\
&= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{2\sigma^2}\right] (2\pi\sigma^2/n)^{1/2} \int (2\pi\sigma^2/n)^{-1/2} \exp\left[-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right]\ d\mu \\
&= (2\pi\sigma^2)^{-n/2} (2\pi\sigma^2/n)^{1/2} \exp\left[-\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{2\sigma^2}\right] \times 1 \quad (*) \\
&= (2\pi\sigma^2)^{-(n-1)/2} \exp\left[-\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{2\sigma^2}\right]
\end{aligned}$$

위의 유도식에서 $(*)$는 평균 $\bar{x}$의 분포가 $N(\mu, \sigma^2/n)$ 을 따르는 사실을 이용하여 그 확률밀도함수를 적분하면 1이되는 사실을 이용한 것이다.

이제 σ^2에 대한 제한적 로그 최대 가능도 함수를 구해보면 아래와 같고

$$l_R(\sigma^2|\boldsymbol{x}) = \log L_R(\sigma^2|\boldsymbol{x}) = -\frac{n-1}{2}\log(2\pi) - \frac{n-1}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2\sigma^2} \tag{4}$$

이를 미분하여 0으로 놓으면 다음과 같은 방정식이 얻어진다.

$$\frac{\partial l_R(\sigma^2|\boldsymbol{x})}{\partial\sigma^2} = -\frac{n-1}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2\sigma^4} = 0$$

따라서 분산 σ^2 에 대한 제한적 최대 가능도 추정량 $\hat{\sigma}^2_{REML}$은 불편추정량 S^2과 동일하게 주어진다.

$$\hat{\sigma}^2_{REML} = S^2 = \frac{\sum_i(x_i-\bar{x})^2}{n-1} \tag{5}$$

위와 같이 정규분포에서 평균과 분산을 같이 추정하는 경우 분산에 대한 불편추정량을 구하기 위하여 가능도 함수에서 평균이 포함된 부분을 적분하여 분산만을 포함한 제한적 최대 가능도 함수를 이용하는 방법을 제한적 최대 가능도 추정법(REML)이라고 한다. 여기서 유의할 점은 REML 방법은 분산을 추정할 때만 이용하는 방법이며 평균의 추정은 원래의 가능도 함수를 이용하면 된다. 또한 평균을 추정하는 절차에서는 분산 추정량의 형태가 영향을 미치지 않는다.

2 선형회귀모형에서의 제한적 가능도 추정법

2.1 선형회귀모형의 최소제곱 추정과 최대가능도 추정

반응변수가 y이고 k개의 독립변수 $(x_1, x_2, \cdots, x_k)$가 있다고 가정하고 표본의 크기 n인 자료가 얻어지면 선형회귀식을 행렬로 다음과 같이 표현할 수 있다.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \\ &= \boldsymbol{x}_i^t\boldsymbol{\beta} + e_i, \quad i = 1, 2, \cdots, n \end{aligned}$$

이와 같은 회귀모형을 선형중회귀모형이라 부르며 각 개체에 대한 모형의 방정식을 벡터와 행렬을 이용하여 표현하면 다음과 같다.

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

여기서 회귀분석의 오차항의 가정을 살펴보면 오차항이 서로 독립이고 동일한 분산을 갖는다. 즉, 오차항은 다음의 분포를 따른다. 즉, $\boldsymbol{e} \sim (0, \sigma^2\boldsymbol{I}_n)$. 관측값 벡터 $\boldsymbol{y}$의 평균과 분산을 보면

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}, \quad Var(\boldsymbol{y}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}_n$$

여기서 오차항이 정규분포를 따른다면 [$\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I}_n)$] 관측값 벡터 $\boldsymbol{y}$ 또한 정규분포를 따른다

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

위의 선형 모형 가정하에서, 최소제곱 추정량 $\hat{\boldsymbol{\beta}}$ (least square estimator)는 다음과 같이 주어지며

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{y}$$

분포는 다음과 같다.

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^t \boldsymbol{X})^{-1})$$

오차항의 분산에 대한 추정은 정규분포 가정을 오차항에 대한 정규분포를 가정하고 다음과 같은 잔차제곱합의 분포에 대한 결과를 이용하면

$$SSE(\hat{\boldsymbol{\beta}}) \sim \sigma^2 \chi^2 (n-k-1)$$

오차항의 분산 σ^2의 불편추정량 S^2을 구할 수 있다.

$$S^2 = \frac{SSE(\hat{\boldsymbol{\beta}})}{(n-k-1)} = \frac{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}})^2}{(n-k-1)}$$

따라서

$$E(S^2) = \sigma^2 \tag{6}$$

여기서 자유도 $n-k-1$은 자료의 개수 n에서 절편을 포함한 회귀계수의 개수 $k+1$를 뺀 수이다.

위에서 분포가정을 이용하여 회귀계수와 오차항의 분산에 대한 최대가능도 추정법(Maximum likelihoodestimation)을 고려할 수 있다. 선형모형과 정규분포 가정 하에서

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

i번째 관측치 y_i는 정규분포 $N(\boldsymbol{x}_i^t \boldsymbol{\beta}, \sigma^2)$를 따르고 서로 독립이므로 관측치의 가능도함수(likelihood function) $L = L(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y})$는 다음과 같다.

$$L = L(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right] \tag{7}$$

모든 관측값은 독립적이기 때문에 그들의 결합확률밀도는 그들의 주변 밀도의 곱이다. 위의 식으로부터 로그가능도함수(Log likelihood function) $l = l(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y})$는 다음과 같이 된다.

$$l = \log L = const - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

로그가능도함수를 최대로 하는 최대 가능도 추정량은 다음과 같이 주어진다.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}SSE(\hat{\boldsymbol{\beta}})$$

그러므로 최대가능도 추정량 $\hat{\boldsymbol{\beta}}$는 최소제곱 추정량과 같다. 여기서 오차 분산의 최대가능도추정량 $\hat{\sigma}^2_{ML}$는 편향되어진다(biased estimator).

$$E(\hat{\sigma}^2_{ML}) = \frac{n-k-1}{n}\sigma^2 \neq \sigma^2 \tag{8}$$

2.2 선형회귀모형에 대한 제한적 최대가능도함수 추정

위의 식 (8)에서 본 것과 같이 선형회귀모형에서 오차항의 분산 σ^2의 최대가능도 추정량은 편이 추정량(biased estimator)이다. 하지만 최소제곱법에서 잔차제곱합의 분포를 이용하여 유도한 S^2은 식 (6)에서 보듯이 불편추정량이다.

이제 앞 절에서 같이 오차항의 분산 σ^2에 대한 제한적 최대 가능도 추정(REML)을 생각해 보자. 일단 다음과 같은 분해가 가능하다.

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^t[\boldsymbol{X}^t\boldsymbol{X}](\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\end{aligned}$$

위에서 교차합 $(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})$은 다음과 같이 0이 된다.

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) &= \boldsymbol{y}^t(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) - \hat{\boldsymbol{\beta}}^t\boldsymbol{X}(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^t\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}^t\boldsymbol{X}\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}\boldsymbol{\beta} \\ &\quad - \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} + \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} + \boldsymbol{y}^t\boldsymbol{X}\boldsymbol{\beta} \\ &= 0\end{aligned}$$

이제 위에서 주어진 제곱합 $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$의 분해를 식 (7) 에 주어진 가능도 함수에 대입하여 계수벡터 $\boldsymbol{\beta}$에 대하여 적분하면 σ^2에 대한 제한적 최대 가능도 함수 $L_R(\sigma^2|\boldsymbol{y})$가 얻어진다.

$$
\begin{aligned}
L_R(\sigma^2|\boldsymbol{y}) &= \int L(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})d\boldsymbol{\beta} \\
&= \int (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right] d\boldsymbol{\beta} \\
&= \int (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^t[\boldsymbol{X}^t\boldsymbol{X}](\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\}\right] d\boldsymbol{\beta} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\right] \int \exp\left[-\frac{1}{2\sigma^2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t[\boldsymbol{X}^t\boldsymbol{X}](\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right] d\boldsymbol{\beta} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}}(2\pi\sigma^2)^{(k+1)/2} det(\boldsymbol{X}^t\boldsymbol{X})^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\right] \\
&= (2\pi\sigma^2)^{-\frac{n-k-1}{2}} det(\boldsymbol{X}^t\boldsymbol{X})^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\right]
\end{aligned}
$$

위 식에서 적분은 회귀계수 추정량의 분포가 다변량 정규분포 $N(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1})$을 따르는 사실과 을 $\boldsymbol{X}^t\boldsymbol{X}$의 차원이 $k+1$임을 이용하면 다음과 같은 값을 가지므로 결과가 얻어지게 된다.

$$
\begin{aligned}
&\int \exp\left[-\frac{1}{2\sigma^2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t[\boldsymbol{X}^t\boldsymbol{X}](\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right] d\boldsymbol{\beta} \\
&= det(2\pi\sigma^2[\boldsymbol{X}^t\boldsymbol{X}]^{-1})^{\frac{1}{2}} \int det(2\pi\sigma^2[\boldsymbol{X}^t\boldsymbol{X}]^{-1})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t\frac{[\boldsymbol{X}^t\boldsymbol{X}]}{\sigma^2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right] d\boldsymbol{\beta} \\
&= det(2\pi\sigma^2[\boldsymbol{X}^t\boldsymbol{X}]^{-1})^{\frac{1}{2}} \\
&= (2\pi\sigma^2)^{(k+1)/2} det(\boldsymbol{X}^t\boldsymbol{X})^{-1/2}
\end{aligned}
$$

위의 식으로부터 제한적 로그가능도함수 $l_R = l_R(\sigma^2|\boldsymbol{y})$는 다음과 같이 된다.

$$
l_R = \log l_R = const - \frac{n-k-1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})
$$

제한적 로그가능도함수를 최대로 하는 오차항의 추정량은 다음과 같이 주어지고 이는 불편추정량이다.

$$
\hat{\sigma}^2_{REML} = \frac{1}{n-k-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) = S^2
$$

3 예제

`lme4` 패키지에서 로그가능도추정(ML)과 제한적 로그가능도추정(REML)에 대한 선택은 `method=FALSE` 또는 `method=TRUE`로 지정할 수 있다. 선택을 지정하지 않으면 제한적 로그가능도추정(REML)으로 분산을 추정한다.

3.1 학교 과학 성적의 실험

서울시 A구에 초등학교가 20개 있다고 하자. 각 학교에 대하여 과학 교육에 대한 두 가지 교수법중 하나를 임의로 선택하여 할당하였다. 10개의 학교는 A교수법이, 다른 10개의 학교에는 B 교수법이 적용되었다. 20개의 학교중 6개의 학교를 임의로 추출하고 (각각 A 교수법 적용 학교 3개, B 교수법 적용 학교 3개 추출) 추출된 학교에 속한 모든 6학년 학생중 임의로 추출된 20명에게 과학시험을 보게하여 점수를 얻었다.

다음과 같은 모형을 고려할 수 있다.

$$y_{ij} = \mu + \tau_{k(i)} + A_i + e_{ij} \tag{9}$$

다음과 같은 가정으로 모형을 가정하자.

- $\mu = 70$
- $\tau_1 = 5$, $\tau_1 = 0$
- $\sigma_a = 3$
- $\sigma_e = 3$

위의 모수를 이용하여 6개의 학교에서 20명의 성적을 임의로 추출하여 자료를 만드는 프로그램은 아래와 같다. 혼합모형을 추정하는 패키지는 `lme4`이다.

```
library(lme4)
set.seed(31313111)
#parameter
tau <- 5
mu <- 70
I <- 6
J <- 20
siga <- 3
sige <- 3
#generationg effect and make response variable
AA <- rnorm(I,0,siga)
```

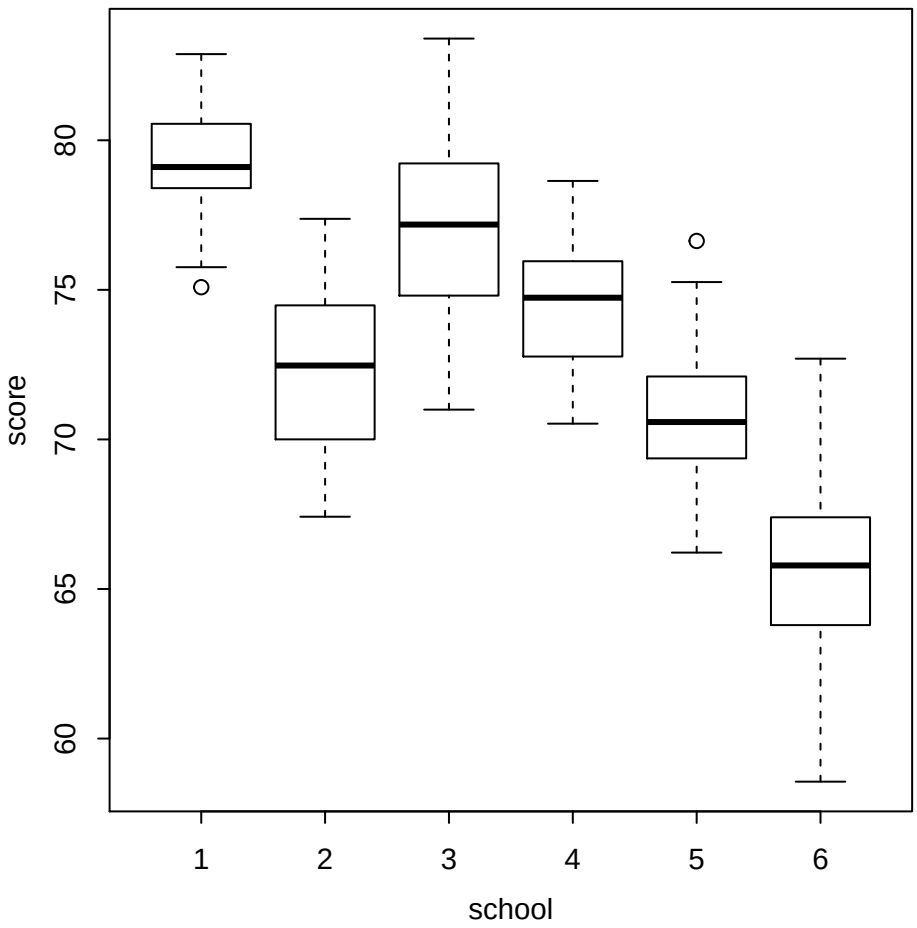
```r
AA <- rep(AA,each=J)
ee <- rnorm(I*J,0,sige)
tt<- c(rep(tau,J*I/2),rep(0,J*I/2))
score <- mu+tt+AA+ee
school <- factor(rep(1:6,each=J))
teach <- factor(c(rep("A",J*I/2),rep("B",J*I/2)))
#making data
data1 <- data.frame(score, school,teach)
head(data1)
```

```
##   score school teach
## 1 80.83      1     A
## 2 79.53      1     A
## 3 75.09      1     A
## 4 79.01      1     A
## 5 80.27      1     A
## 6 79.18      1     A
```

```r
tail(data1)
```

```
##     score school teach
## 115 65.84      6     B
## 116 63.83      6     B
## 117 70.81      6     B
## 118 67.65      6     B
## 119 63.75      6     B
## 120 66.56      6     B
```

```r
boxplot(score~school,data=data1,xlab="school",ylab="score")
```



최대가능도추정(ML)으로 혼합모형 (9)을 적합시켜보자.

```
# mixed effect model with ML
fm1ML <- lmer(score~teach+(1|school), data=data1,REML = FALSE)
summary(fm1ML)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: score ~ teach + (1 | school)
##    Data: data1
##
##      AIC      BIC   logLik deviance df.resid
##    606.8    618.0   -299.4    598.8      116
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7308 -0.7019 -0.0097  0.6351  2.5110
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 10.04    3.17
##  Residual              7.28    2.70
## Number of obs: 120, groups:  school, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    76.32       1.86   40.98
## teachB         -5.95       2.63   -2.26
##
## Correlation of Fixed Effects:
##        (Intr)
## teachB -0.707
```

σ_a의 ML 추정값은 3.1693 이고 σ_e의 ML 추정값은 2.6975 이다. 교수법의 차이에 대한 검정에 대해서는 p-value가 주어지지 않았지만 t-검정통계량 값은 -2.2595이다.

제한적 최대가능도추정(REML)으로 혼합모형 (9)을 적합시켜보자.

```
# mixed effect model with REML
fm2ML <- lmer(score~teach+(1|school), data=data1, REML = TRUE)
summary(fm2ML)

## Linear mixed model fit by REML ['lmerMod']
## Formula: score ~ teach + (1 | school)
##    Data: data1
##
## REML criterion at convergence: 592.3
```

```
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.711 -0.716 -0.006  0.631  2.531
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 15.25    3.9
##  Residual              7.28    2.7
## Number of obs: 120, groups:  school, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    76.32       2.28   33.46
## teachB         -5.95       3.23   -1.84
##
## Correlation of Fixed Effects:
##        (Intr)
## teachB -0.707
```

σ_a의 REML 추정값은 3.9049 이고 σ_e의 REML 추정값은 2.6975 이다. 앞에서 ML 추정량과 비교하면 σ_e의 추정은 동일하고 σ_a은 REML 이 ML 추정량보다 크다.

참고할 점은 고정효과에 대한 추정량은 ML 과 REML 모두 동일하지만 분산성분의 추정량이 다르기 때문에 그에 대한 표준오차가 다르고 따라서 t-값도 다르다. REML을 사용한 경우 t-검정통계량 값은 -1.8449이다.

3.2 Sleep study

강의노트(선형혼합모형 II - 반복측정자료)에서 소개된 Sleep study 자료에 대하여 ML 과 REML을 각각 적용하여 분산성분을 추정해 보았다.

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

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
##    Data: sleepstudy
##
##      AIC      BIC   logLik deviance df.resid
##     1764     1783     -876     1752      174
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.942 -0.466  0.029  0.464  5.179
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 565.5    23.78
##           Days         32.7     5.72    0.08
##  Residual             654.9    25.59
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   251.41       6.63   37.91
## Days           10.47       1.50    6.97
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.138

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

```
summary(fm2)

## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1744
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -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       6.82   36.84
## Days           10.47       1.55    6.77
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.138
```

4 혼합모형에서 고정효과에 대한 검정

아래의 주어진 혼합모형에서

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{e}$$

회귀계수 $\boldsymbol{\beta}$에 대한 가설 검정(hypothesis test)은 최대가능도 추정에서 가능도비(likelihood ratio test)를 이용한다. 보통의 선형회귀분석에서는 t-검정이나 F-검정을 사용하는 것과 다르게 혼합모형에서의 가능도비 검정은 카이제곱 검정(χ^2-test, Wald test) 또는 F-검정을 사용한다.

예를 들어 앞 절에서 살펴본 학생들의 과학시험에 대한 모형에서

$$y_{ij} = \mu + \tau_{k(i)} + A_i + e_{ij}$$

교수법의 효과에 대한 검정을 다음과 같이 고려할 수 있다.

$$H_0 : \tau_1 = \tau_2 \quad \text{vs.} \quad H_0 : \tau_1 \neq \tau_2$$

여기서 $\tau_{k(i)}$는 i번째 학교가 A 교수법으로 가르치면 $\tau_{k(i)} = \tau_1$, B 교수법으로 가르치면 $\tau_{k(i)} = \tau_2$가 되는 모수이다. 실제로 `lme4`로 자료를 적합시키면 A 교수법의 효과는 0으로 고정하고 ($\tau_1 = 0$) A 교수법의 상대 효과(τ_2, `teachB`) 를 추정한다. 앞에서 볼 수 있듯이 각 효과에 대한 t-값만 제시되고 그에 대한 검정의 결과(예를 들어 p-value)는 제공되지 않는다. 이는 가능도비 검정이 `lme4`패키지에서 제공되고 있지 않기 때문이다. 고정효과에 대한 가능도비 검정은 `car` 패키지에 있는 `Anova`함수를 이용하여 실행할 수 있다.

앞에서 고려한 교수법의 효과에 대한 검정(카이제곱 검정)을 실시해보자. ML과 REML로 적합한 두 결과에 대하여 교수법의 효과가 있는지에 대한 검정을 아래와 같이 실시하였다. 검정에 대한 p-값은 `Pr(>Chisq)`로 주어진다. ML의 결과를 이용하면 교수법의 차이가 유의하게 다르고 (H_0 기각) REML의 결과를 이용하면 교수법의 차이가 유의하지 않다 (H_0 기각 못함). 두 결과가 다른 이유는 분산성분 σ_b^2에 대한 추정값이 ML 과 REML에 따라서 다르기 때문이다. 일반적인 경우 분산성분의 불편 추정량을 제공해주는 REML의 결과를 이용하여 가설검정을 한다.

```
anova(fm1ML)

## Analysis of Variance Table
##       Df Sum Sq Mean Sq F value
## teach  1   37.1    37.1    5.11
```

```r
anova(fm2ML)
```

```
## Analysis of Variance Table
##       Df Sum Sq Mean Sq F value
## teach  1   24.8    24.8     3.4
```

```r
library(car)
```

```
## Loading required package: carData
## Registered S3 methods overwritten by 'car':
##  method                           from
##  influence.merMod                 lme4
##  cooks.distance.influence.merMod lme4
##  dfbeta.influence.merMod         lme4
##  dfbetas.influence.merMod         lme4
```

```r
anova(fm1ML)
```

```
## Analysis of Variance Table
##       Df Sum Sq Mean Sq F value
## teach  1   37.1    37.1    5.11
```

```r
Anova(fm1ML)
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: score
##       Chisq Df Pr(>Chisq)
## teach  5.11  1      0.024 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(fm2ML)
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: score
##       Chisq Df Pr(>Chisq)
## teach   3.4  1      0.065 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5 부록: 행렬대수

5.1 다변량 정규분포

k 차원 다변량 정규분포 $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$의 확률밀도함수가 다음과 같다.

$$f(\boldsymbol{y}) = (2\pi)^{-k/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]$$

여기서 $|\boldsymbol{A}|$는 행렬 $\boldsymbol{A}$의 행렬식이다.

5.2 행렬식(Determinant)

$$|\boldsymbol{A}^t| = |\boldsymbol{A}|$$

$$|c\boldsymbol{A}| = c^k|\boldsymbol{A}|$$

여기서 k는 행렬 $\boldsymbol{A}$의 차원이다.

$$|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$$

$$|\boldsymbol{A}^{-1}| = |\boldsymbol{A}|^{-1}$$

$$|a\boldsymbol{I}_n + b\mathbf{1}_n\mathbf{1}_n^t| = a^{n-1}(a+nb)$$