STAT 574 Linear and Nonlinear Mixed Models

Lecture 2: Linear Mixed Models and Estimation

Chencheng Cai

Washington State University

Weight versus height example

- A dataset contains the heights and weights of 71 people from 18 families.
- A naive linear regression model:

$$W_k = \alpha + \beta H_k + \epsilon_k,$$

where W_k is the weight of the kth person and H_k is his/her height.

Anything wrong with the linear regression assumptions?

LINE Assumptions

- Linear: it might be more realistic to assume
  $$W_k = \alpha + \beta H_k^2 + \epsilon_k.$$
  Justification: Body Mass Index (BMI)
- Independence: weights of people in the same family are highly correlated. Potential confounders: gene, habits, environment.
- Normal: can be diagnosed later.
- Equal variance: may need to take a logarithm of the weight.

LINE Assumptions

- Linear: it might be more realistic to assume

$$W_k = \alpha + \beta H_k^2 + \epsilon_k.$$

  Justification: Body Mass Index (BMI)
- Independence: weights of people in the same family are highly correlated.
  Potential confounders: gene, habits, environment.
- Normal: can be diagnosed later.
- Equal variance: may need to take a logarithm of the weight.

let us focus on the correlation for now.

Model update

In order to incorporate the within-family correlation, we assume

$$W_{ij} = \alpha_i + \beta H_{ij} + \epsilon_{ij}$$

- The index ij refers to the jth person in the ith family.
- Intercept α_i is family-specific.
- ϵ_{ij} is normal-distributed and is independent with other variabes.
- α_i's are I.I.D.

Correlation

- Alice and Bob from the same family:

$$\begin{aligned}\mathrm{Cov}(W_{ij}, W_{ij'}) &= \mathrm{Cov}(\alpha_i + \beta H_{ij} + \epsilon_{ij}, \alpha_i + \beta H_{ij'} + \epsilon_{ij'}) \\ &= \mathrm{Cov}(\alpha_i + \epsilon_{ij}, \alpha_i + \epsilon_{ij'}) \\ &= \mathrm{Var}(\alpha_i) > 0\end{aligned}$$

- Alice and Bob from different families:

$$\begin{aligned}\mathrm{Cov}(W_{ij}, W_{i'j'}) &= \mathrm{Cov}(\alpha_i + \beta H_{ij} + \epsilon_{ij}, \alpha_{i'} + \beta H_{i'j'} + \epsilon_{i'j'}) \\ &= \mathrm{Cov}(\alpha_i + \epsilon_{ij}, \alpha_{i'} + \epsilon_{i'j'}) \\ &= \mathrm{Cov}(\alpha_i, \alpha_{i'}) \\ &= 0\end{aligned}$$

Model reformat

We can also write

$$W_{ij} = \alpha + \beta H_{ij} + b_i \times 1 + \epsilon_{ij}$$

where $\alpha = \mathbb{E}[\alpha_i]$ and $b_i = \alpha_i - \alpha$.
Furthermore, we vectorize it by family:

$$\boldsymbol{W}_i = \alpha + \beta \boldsymbol{H}_i + b_i \boldsymbol{Z}_i + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{Z}_i = (1, 1, \ldots, 1)^T$.

Linear Mixed Effects (LME) Model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad \text{for } i = 1, \ldots, N.$$

- $\boldsymbol{y}_i$: $n_i \times 1$ vector of responses of the ith cluster/group.
- $\boldsymbol{X}_i$: $n_i \times m$ design matrix of fixed effects.
- $\boldsymbol{\beta}$: $m \times 1$ vector of fixed effects coefficents.
- $\boldsymbol{Z}_i$: $n_i \times k$ design matrix of random effects.
- $\boldsymbol{b}_i$: $k \times 1$ vector of random effect coefficients.
- $\boldsymbol{\epsilon}_i$: $n_i \times 1$ vector of error terms.

Examples

- Weight-height relation for different families.
- Social behaviors for people of different ages.
- Stock price movement for companies in different sections/industries.
- Biological studies on animals of different sub-species.

In summary, you should consider a mixed-effects model if you do not believe your model (and the coefficients) are the same for different sub-populations.

Distributional Assumptions

Normal assumptions for the error terms and for the random effect coefficients:

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}), \quad \boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{D})$$

warning! Sometimes, one only assumes $\mathrm{Cov}(\boldsymbol{b}_i) = \sigma^2 \boldsymbol{D}$. But we assume normal here.

Consequently,

$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{X}_i \boldsymbol{\beta}, \sigma^2 \boldsymbol{V}_i), \qquad \text{for } i = 1, \ldots, N$$

with $\boldsymbol{V}_i = \boldsymbol{I} + \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T$.

A more compact formula

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

where

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_N \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_N \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_N \end{bmatrix}$$

and $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N)$. Also

$$\mathrm{Cov}(\boldsymbol{\eta}) = \sigma^2 \mathrm{diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_N)$$

Special Model: random intercepts model

$$y_{ij} = \alpha_i + \boldsymbol{\gamma}'\boldsymbol{u}_{ij} + \epsilon_{ij}.$$

where $\alpha_i = \alpha + b_i$ and $b_i \sim \mathcal{N}(0, \sigma^2 d)$.

Special Model: balanced random-coefficient model

We assume all clusters have the same size $n_i = n$ and

$$\boldsymbol{Z} = \boldsymbol{X}_i = \boldsymbol{Z}_i, \quad i = 1, \ldots, N.$$

Then, LME becomes:

$$\boldsymbol{y}_i = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N.$$

A more compact form:

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta}\mathbf{1}^T + \boldsymbol{E}.$$

with $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N]$ and $\boldsymbol{E}$ has I.I.D. columns with covariance $\sigma^2(\boldsymbol{I} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T)$.

Linear Growth Curve Model

We start with a linear regression model with random coefficients:

$$\boldsymbol{y}_i = \boldsymbol{Z}_i \boldsymbol{a}_i + \boldsymbol{\epsilon}_i.$$

Furthermore, we assume the coefficients are linear combinations of other covariates:

$$\boldsymbol{a}_i = \boldsymbol{A}_i \boldsymbol{\beta} + \boldsymbol{b}_i$$

with $\mathbb{E}(\boldsymbol{b}_i) = 0$ and $\mathrm{Cov}(\boldsymbol{b}_i) = \sigma^2 \boldsymbol{D}$.
Combine them:

$$\boldsymbol{y}_i = \boldsymbol{Z}_i \boldsymbol{A}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i.$$

Example:

- Index i: person. index j: time.
- y: health indicator. $\boldsymbol{Z}$: health-related covariates.
- $\boldsymbol{A}$: whether the person takes a medicine or a placebo. $\boldsymbol{\beta}$: the effect of the medicine.

Log-likelihood Function

- Distribution:
$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_i), \quad \boldsymbol{V}_i = \boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T.$$

- Probability:
$$p(\boldsymbol{y}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{D}, \sigma^2) = \frac{1}{(2\pi)^{n_i/2}\sqrt{|\sigma^2\boldsymbol{V}_i|}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\}$$

- Log-likelihood:
$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2) &= \sum_{i=1}^{N}\left\{-\frac{n_i}{2}\log 2\pi\sigma^2 - \frac{1}{2}\log|\boldsymbol{V}_i| - \frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\} \\ &= -\frac{1}{2}\left\{N_T\log\sigma^2 + \sum_{i=1}^{N}\left[\log|\boldsymbol{V}_i| + \frac{1}{\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right]\right\} + C\end{aligned}$$

where $N_T = \sum_{i=1}^{N} n_i$ and $C = -\frac{N_T}{2}\log 2\pi$.

MLE for LME Models

$$\max_{(\boldsymbol{\beta},\boldsymbol{D},\sigma^2)\in\boldsymbol{\Theta}} \underbrace{-\frac{1}{2}\left\{N_T \log\sigma^2 + \sum_{i=1}^{N}\left[\log|\boldsymbol{V}_i| + \sigma^{-2}\boldsymbol{e}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{e}_i\right]\right\}}_{\ell(\boldsymbol{\beta},\boldsymbol{D},\sigma^2)}$$

with $\boldsymbol{V}_i = \boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T$ and $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}$.

Optimization algorithms:

- Newton–Raphson — optimized gradient descent.
- Fisher scoring — a stable version of NR
- Expectation–Maximization — for missing value problems

The Parameter Space

- Nonnegative Definite Parameter Space:

$$\mathbf{\Theta} = \{(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{\sigma}^2) : \boldsymbol{\beta} \in \mathbb{R}^m, \sigma^2 > 0, \boldsymbol{D} \succeq 0\}$$

- Dimension: $\dim(\mathbf{\Theta}) = m + 1 + k(k+1)/2$.
- Drawback: difficult to enforce nonnegativeness.
- Another parameter space:

$$\mathbf{\Theta} = \{(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{\sigma}^2) : \boldsymbol{\beta} \in \mathbb{R}^m, \sigma^2 > 0, \boldsymbol{V}_i \succ 0 \text{ for } i = 1, \ldots, N\}$$

- Benefits: $\ell(\theta) \to -\infty$ on boundary.

Profiling LLH Function

Consider the one-step optimization:

$$\max_{\sigma^2} \; -\frac{1}{2}\left\{N_T \log \sigma^2 + \sum_{i=1}^{N}\left[\log|\boldsymbol{V}_i| + \sigma^{-2}\boldsymbol{e}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{e}_i\right]\right\}$$

Partial derivative:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N_t}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}\boldsymbol{e}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{e}_i.$$

Setting above to zero, we have:

$$\hat{\sigma}^2 = \frac{1}{N_T}\sum_{i=1}^{N}\boldsymbol{e}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{e}_i$$

profiled log-likelihood function:

$$\ell_p(\boldsymbol{\beta}, \boldsymbol{D}) = \ell(\boldsymbol{\beta}, \boldsymbol{D}, \hat{\sigma}^2) = -\frac{1}{2}\left\{N_T \log \sum_{i=1}^{N}\boldsymbol{e}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{e}_i + \sum_{i=1}^{N}\log|\boldsymbol{V}_i|\right\} + C$$

Profiling LLH Function

One step further:

$$\max_{\boldsymbol{\beta}} \underbrace{-\frac{1}{2}\left\{N_T \log \sum_{i=1}^{N} \boldsymbol{e}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{e}_i + \sum_{i=1}^{N} \log |\boldsymbol{V}_i|\right\}}_{\ell_p(\boldsymbol{\beta}, \boldsymbol{D})}$$

Equivalent to:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \boldsymbol{e}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{e}_i$$

Set the partial derivative to zero:

$$2 \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta})^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i = 0$$

Profiling LLH Function

Solution:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i\right)$$

The above solution is the Generalized Least Squares (GLS) estimator.
Further profiled log-likelihood function:

$$\ell_p(\boldsymbol{D}) = \ell_p(\hat{\boldsymbol{\beta}}, \boldsymbol{D}) = -\frac{1}{2}\left\{N_T \log\left(s_{yy} - \boldsymbol{s}_{xy}^T \boldsymbol{S}_{xx}^{-1} \boldsymbol{s}_{xy}\right) + \sum_{i=1}^{N} \log|\boldsymbol{V}_i|\right\}$$

where $s_{yy} = \sum_i \boldsymbol{y}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i$, $\boldsymbol{s}_{xy} = \sum_i \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i$ and $\boldsymbol{S}_{xx} = \sum_i \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i$.

We can maximize $\ell_p(\boldsymbol{D})$ instead of $\ell(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2)$ to speed up computation!

Profiling LLH Function

The final optimization:

$$\max_{\boldsymbol{D}} \; -\frac{1}{2}\left\{N_T \log\left(s_{yy} - \boldsymbol{s}_{xy}^T \boldsymbol{S}_{xx}^{-1} \boldsymbol{s}_{xy}\right) + \sum_{i=1}^{N} \log|\boldsymbol{V}_i|\right\}$$

Unfortunately, it usually has no analytical solution.

Let $\hat{\boldsymbol{D}}$ be the optimum and $\hat{\boldsymbol{V}}_i = \boldsymbol{I} + \boldsymbol{Z}_i \hat{\boldsymbol{D}} \boldsymbol{Z}_i^T$. Then, we have the MLE for the other two parameters:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{y}_i\right)$$

$$\hat{\sigma}^2 = \frac{1}{N_t} \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}})^T \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}})$$

Optimizating $\ell_p(\boldsymbol{D})$

Using Woodbury identity $V_i^{-1} = \boldsymbol{I} - \boldsymbol{Z}_i(\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i^T$, we have

$$\begin{aligned}
s_{yy} &= \sum \boldsymbol{y}_i^T\boldsymbol{y}_i - \sum \boldsymbol{y}_i^T\boldsymbol{Z}_i(\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i^T\boldsymbol{y}_i \\
\boldsymbol{s}_{xy} &= \sum \boldsymbol{X}_i^T\boldsymbol{y}_i - \sum \boldsymbol{X}_i^T\boldsymbol{Z}_i(\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i^T\boldsymbol{y}_i \\
\boldsymbol{S}_{xx} &= \sum \boldsymbol{X}_i^T\boldsymbol{X}_i - \sum \boldsymbol{X}_i^T\boldsymbol{Z}_i(\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i^T\boldsymbol{X}_i
\end{aligned}$$

Using Sylvester's identity $|\boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T| = |\boldsymbol{I} + \boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{Z}_i|$, we have

$$\log|\boldsymbol{V}_i| = \log(|\boldsymbol{D}||\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^T\boldsymbol{Z}_i) = \log|\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^T\boldsymbol{Z}_i| - \log|\boldsymbol{D}^{-1}|$$

Benefits:

- Complicated computation (determinant, inverse) on $n_i \times n_i$ matrices are now on $k \times k$ matrices.
- Many quantities can be pre-computed.
- Optimization can be done with respect to $\boldsymbol{D}^{-1}$.

Optimizating $\ell_p(\boldsymbol{D}^{-1})$

For algorithms, it is necessary to know the partial derivative of $\ell_p(\boldsymbol{D}^{-1})$.

$$
\begin{aligned}
d\log|\boldsymbol{V}_i| &= \operatorname{tr}\left[(\boldsymbol{D}^{-1}+\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}d\boldsymbol{D}^{-1}\right]-\operatorname{tr}\left[\boldsymbol{D}d\boldsymbol{D}^{-1}\right]\\
ds_{yy} &= \sum \boldsymbol{y}_i^T\boldsymbol{Z}_i(\boldsymbol{D}^{-1}+\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}d\boldsymbol{D}^{-1}(\boldsymbol{D}^{-1}+\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i^T\boldsymbol{y}_i\\
d\boldsymbol{s}_{xy} &= \sum \boldsymbol{X}_i^T\boldsymbol{Z}_i(\boldsymbol{D}^{-1}+\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}d\boldsymbol{D}^{-1}(\boldsymbol{D}^{-1}+\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i^T\boldsymbol{y}_i\\
d\boldsymbol{S}_{xx} &= \sum \boldsymbol{X}_i^T\boldsymbol{Z}_i(\boldsymbol{D}^{-1}+\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}d\boldsymbol{D}^{-1}(\boldsymbol{D}^{-1}+\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i^T\boldsymbol{X}_i
\end{aligned}
$$

Optimizating $\ell_p(\boldsymbol{D}^{-1})$

Eventually, we have

$$\frac{\partial \ell_p(\boldsymbol{D}^{-1})}{\partial \boldsymbol{D}^{-1}} = -\frac{1}{2}\left\{\sum_{i=1}^{N}\boldsymbol{G}_i - N\boldsymbol{D} + \frac{N_T}{s_{yy} - \boldsymbol{s}_{xy}^T\boldsymbol{S}_{xx}\boldsymbol{s}_{xy}}\sum_{i=1}^{N}\boldsymbol{G}_i\boldsymbol{Z}_i^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{Z}_i\boldsymbol{G}_i\right\}$$

where

$$\boldsymbol{G}_i = (\boldsymbol{D}_i^{-1} + \boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}$$
$$\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{S}_{xx}^{-1}\boldsymbol{s}_{xy}$$

Restricted MLE

The MLE for σ^2 is estimated from $\hat{\boldsymbol{e}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ directly:

$$\hat{\sigma}^2 = \frac{1}{N_t}\sum_{i=1}^{N} \hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{e}}_i$$

$\hat{\sigma}^2$ is in general biased for σ^2 because $\hat{\boldsymbol{e}}_i$ involves $\hat{\boldsymbol{\beta}}$.
The likelihood function (after certain modifications) that disentangles $\hat{\boldsymbol{\beta}}$ from other estimators is called the **restricted likelihood** function.
The method that maximizes the restricted likelihood function is called **restricted maximum likelihood (REML)**.

REML

- Consider a linear regression model $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$.
- The observation can be decomposed into two orthogonal parts: $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$.
- Why orthogonal?

$$\mathrm{Cov}(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{e}}) = \mathrm{Cov}\left(\boldsymbol{H}\boldsymbol{y}, (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\right) = \boldsymbol{H}\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0}.$$

- Therefore, we can write:

$$\ell(\hat{\boldsymbol{e}}, \boldsymbol{V}) = \ell(\boldsymbol{y}, \boldsymbol{V}) - \ell(\hat{\boldsymbol{\beta}}, \boldsymbol{V}) = -\frac{1}{2}\left\{\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| + \log|\boldsymbol{V}| + \hat{\boldsymbol{e}}^T\boldsymbol{V}^{-1}\hat{\boldsymbol{e}}\right\}$$

- Furthermore, we can have the restricted log-likelihood:

$$\ell_R(\boldsymbol{\beta}, \boldsymbol{V}) == -\frac{1}{2}\left\{\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| + \log|\boldsymbol{V}| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\}$$

REML — Bayesian Perspective

We take a non-informative prior on $\boldsymbol{\beta}$ (i.e. uniform). Then the profiled likelihood function is

$$L_R(\boldsymbol{V}) = \int_{\mathbb{R}^m} L(\boldsymbol{\beta}, \boldsymbol{V}) d\boldsymbol{\beta} \propto |\boldsymbol{V}|^{-1/2} |\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \right\}$$

- Why non-informative?
- Why independent?

REML for LME

- Back to linear mixed models, we have

$$\ell_R(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2) = -\frac{1}{2}\Big\{(N_T - m)\log\sigma^2 + \log\left|\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right| + \sum_{i=1}^{N}\left[\log|\boldsymbol{V}_i| + \sigma^{-2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right]\Big\}$$

- $\log\left|\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right|$ v.s. $\sum_{i=1}^{N}\log|\boldsymbol{V}_i|$
- $(N_T - m)\log\sigma^2$

Profiled REML

- Maximize $\ell_R(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2)$:

$$\hat{\sigma}_R^2 = \frac{1}{N_T - m} \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

- Therefore the profiled restricted LLH is

$$\begin{aligned}\ell_{Rp}(\boldsymbol{\beta}, \boldsymbol{D}) = -\frac{1}{2}\Big\{ (N_T - m) \log \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\ + \log \left| \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right| + \sum_{i=1}^{N} [\log |\boldsymbol{V}_i|] \Big\}\end{aligned}$$

REML summary

- likelihood function based on the transformed data.
- likelihood function independent of fixed effect coefficients.
- unbiased estimators for variance components.
- R functions, including **lme** and **lmer**, support both REML and ML.

Balanced Random-Coefficient Model

- Assumption 1: $\boldsymbol{Z} = \boldsymbol{X}_i = \boldsymbol{Z}_i$ for $i = 1, \ldots, N$. (so $\boldsymbol{V}_i = \boldsymbol{V}$)
- Assumption 2: $n_i = n$ for $i = 1, \ldots, N$.
- Interpretation:
$$\boldsymbol{y}_i = \boldsymbol{Z}(\boldsymbol{\beta} + \boldsymbol{b}_i) + \boldsymbol{\epsilon}_i,$$
where $\boldsymbol{\beta}$ is the expectation of coefficients and $\boldsymbol{b}_i$ is the random part.
- Example: repeated measure of patients.

Balanced Random-Coefficient Model

- Fixed-effect coefficients:

$$\hat{\boldsymbol{\beta}}_{GLS} = \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\bar{\boldsymbol{y}}$$

  where $\boldsymbol{y} = N^{-1}\sum_i \boldsymbol{y}_i$.

- Variance:

$$\hat{\sigma}^2_{ML} = \hat{\sigma}^2_{RML} = \frac{1}{N(n-m)}\sum_{i=1}^{N}\boldsymbol{y}_i^T(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T)\boldsymbol{y}_i$$

$$\hat{\boldsymbol{D}}_{ML} = \frac{1}{N\hat{\sigma}^2_{ML}}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\hat{\boldsymbol{E}}\hat{\boldsymbol{E}}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} - (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$$

$$\hat{\boldsymbol{D}}_{RML} = \frac{1}{(N-1)\hat{\sigma}^2_{ML}}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\hat{\boldsymbol{E}}\hat{\boldsymbol{E}}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} - (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$$

  where $\hat{\boldsymbol{E}}\hat{\boldsymbol{E}}^T = \sum_i(\boldsymbol{y}_i - \boldsymbol{Z}\hat{\boldsymbol{\beta}})(\boldsymbol{y}_i - \boldsymbol{Z}\hat{\boldsymbol{\beta}})^T$.

Balanced Random-Coefficient Model

- Log-likelihood

$$\ell = -\frac{N}{2}\left\{ n\log\sigma^2 + \log|\boldsymbol{I} + \boldsymbol{Z}^T\boldsymbol{D}\boldsymbol{Z}| + \frac{1}{N\sigma^2}\sum_i (\boldsymbol{y}_i - \boldsymbol{Z}\boldsymbol{\beta})^T(\boldsymbol{I} + \boldsymbol{Z}^T\boldsymbol{D}\boldsymbol{Z})^{-1}(\boldsymbol{y}_i - \boldsymbol{Z}\boldsymbol{\beta}) \right\}$$

- Use

$$\boldsymbol{V}_i^{-1} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{D}^{-1} + \boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$$
$$\boldsymbol{V}_i^{-1}\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{D}^{-1} + \boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{D}^{-1}$$

Fitting Linear Mixed-effect Models in R

- Datasets
  - Download from the author's GitHub repository.
  - https://github.com/eugenedemidenko/mixedmodels
  - datasets are stored in .txt files in **Data/MixedModels** folder.
- Packages
  - `lme` function from `nlme` library.
  - `lmer` function from `lme4` library.
  - `lme` supports more covariance structures.
  - `lmer` has better scalability.

Load Dataset

Load the height-weight datasets from `Family.txt` file.

```r
data = read.table("./Data/MixedModels/Chapter02/Family.txt",
                  header=T, stringsAsFactors=F)
head(data)
```

	Height	Weight	Sex	ParentChild	Age	FamilyID
1	67.0	215	1	1	75	1
2	64.0	155	0	1	63	1
3	63.5	145	0	0	29	1
4	71.0	227	1	0	26	1
5	61.0	120	0	0	24	1
6	68.0	220	1	0	22	1

Fit LME with `lme()` Function

```r
library(nlme)
fit.lme = lme(fixed=Weight~Height, random=~1|FamilyID, data=data)
fit.lme
```

- `fixed` argument specifies the fixed effect model. In this example, it is the linear regression of `Weight` against `Height`.
- `random` argument specifies the random effect model.
  - `~1` specifies that the random effect is on the intercept.
  - `FamilyID` specifies the group variable.

Fit LME with `lme()` Function

```
Linear mixed-effects model fit by REML
Data: data
Log-restricted-likelihood: -331.6369
Fixed: Weight ~ Height
(Intercept)      Height
-206.832149    5.345309

Random effects:
Formula: ~1 | FamilyID
        (Intercept) Residual
StdDev:    14.07057  24.7059

Number of Observations: 71
Number of Groups: 18
```

- Default estimation method: REML
- Coefficients from the fixed-effect model.
- Standard deviation for the random-effect coefficients.

Fit LME with `lme()` Function

To use ML, need to specify the `method` argument:

```
fit.lme = lme(fixed=Weight~Height, random=~1|FamilyID,
              method="ML", data=data)
fit.lme
```

Fit LME with `lme()` Function

```
Linear mixed-effects model fit by maximum likelihood
  Data: data
  Log-likelihood: -334.7041
  Fixed: Weight ~ Height
(Intercept)      Height
-205.015367    5.319309

Random effects:
 Formula: ~1 | FamilyID
        (Intercept) Residual
StdDev:    13.34261 24.50155

Number of Observations: 71
Number of Groups: 18
```

- (a) Different likelihood values.
- (b) Different fixed-effect coefficients.
- (c) Different variance parameters.

Why?

Fit LME with `lme()` Function

```
Linear mixed-effects model fit by maximum likelihood
  Data: data
  Log-likelihood: -334.7041
  Fixed: Weight ~ Height
(Intercept)       Height
-205.015367     5.319309

Random effects:
 Formula: ~1 | FamilyID
        (Intercept) Residual
StdDev:    13.34261 24.50155

Number of Observations: 71
Number of Groups: 18
```

- (a) Different likelihood values.
- (b) Different fixed-effect coefficients.
- (c) Different variance parameters.

Why?

- (a) and (c): the use of restricted likelihood.
- (b): same formula for $\hat{\beta}_{GLS}$ but with different $\hat{\boldsymbol{D}}$'s.

Fit LME with lme() Function

To get more output information, we can call

```r
summary(fit.lme)
```

```
Linear mixed-effects model fit by maximum likelihood
  Data: data
       AIC      BIC    logLik
  677.4082 686.4589 -334.7041

Random effects:
 Formula: ~1 | FamilyID
        (Intercept) Residual
StdDev:    13.34261 24.50155

Fixed effects:  Weight ~ Height
                Value Std.Error DF   t-value p-value
(Intercept) -205.01537  54.08819 52 -3.790390   4e-04
Height         5.31931   0.78212 52  6.801126   0e+00
 Correlation:
       (Intr)
Height -0.997

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.10329358 -0.54475601 -0.08698002  0.36755416  3.64634517

Number of Observations: 71
Number of Groups: 18
```

Fit LME with `lme()` Function

- height coefficient is random

Fit LME with `lme()` Function

- height coefficient is random

```
lme(fixed=Weight~Height, random=~1+Height|FamilyID, data=data)
```

- group indicated by FamilyID and Sex

Fit LME with `lme()` Function

- height coefficient is random

```
lme(fixed=Weight~Height, random=~1+Height|FamilyID, data=data)
```

- group indicated by FamilyID and Sex

```
lme(fixed=Weight~Height, random=~1|FamilyID/Sex, data=data)
```

Fit LME with lmer() Function

```r
library(lme4)
fit.lme = lmer(Weight~Height+(1|FamilyID), data=data)
fit.lme
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Weight ~ Height + (1 | FamilyID)
    Data: data
REML criterion at convergence: 663.2737
Random effects:
  Groups   Name        Std.Dev.
  FamilyID (Intercept) 14.07
  Residual             24.71
Number of obs: 71, groups:  FamilyID, 18
Fixed Effects:
(Intercept)       Height
-206.832        5.345
```

Fit LME with `lmer()` Function

Do not foget "()" for random effects

```
fit.lme.wrong = lmer(Weight~Height+1|FamilyID, data=data)
fit.lme.wrong
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Weight ~ Height + 1 | FamilyID
    Data: data
REML criterion at convergence: 697.5127
Random effects:
  Groups   Name        Std.Dev. Corr
  FamilyID (Intercept) 33.4422
           Height       0.5802  -1.00
  Residual             33.5836
Number of obs: 71, groups:  FamilyID, 18
Fixed Effects:
(Intercept)
    160.9
optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 2 lme4 warnings
```

Fit LME with lmer() Function

Use REML argument to choose the estimation method.

```r
fit.lme = lmer(Weight~Height+(1|FamilyID), REML=F, data=data)
fit.lme
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Weight ~ Height + (1 | FamilyID)
   Data: data
      AIC       BIC    logLik  deviance  df.resid
 677.4082  686.4589 -334.7041  669.4082        67
Random effects:
 Groups   Name        Std.Dev.
 FamilyID (Intercept) 13.34
 Residual             24.50
Number of obs: 71, groups:  FamilyID, 18
Fixed Effects:
(Intercept)       Height
-205.015        5.319
```

Fit LME with lmer() Function

- height coefficient is random

```
lmer(Weight~Height+(1+Height|FamilyID), data=data)
```

- Two-way group structure

```
lmer(Weight~Height+(1|FamilyID/Sex), data=data)
```

- Force independent random effects

```
lmer(Weight~Height+(1+Height||FamilyID), data=data)
```

Maximization Algorithms

To maximize $\ell(\boldsymbol{\theta})$, the iterative algorithms update the value of $\boldsymbol{\beta}$ by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \lambda_t \boldsymbol{H}_t^{-1} \nabla_{\boldsymbol{\theta}} \ell.$$

- Newton Raphson: $\boldsymbol{H} = -\nabla \ell \nabla^T$ is the negative Hessian matrix of ℓ.
- Fisher scoring: $\boldsymbol{H} = -\mathbb{E}[\nabla \ell \nabla]$ is the negative information matrix.
- EM algorithm: $\boldsymbol{H}$ is some positive definite matrix.

A Note on EM Algorithm

EM algorithm is used to maximize the marginal likelihood function when missing data exists.

$$\max_{\boldsymbol{\theta}} \; \ell(X \mid \boldsymbol{\theta}) = \log \int L(X, Y \mid \boldsymbol{\theta}) dY$$

- Expectation step: compute the expected complete log-likelihood function given the observed data X and the parameter.

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)}) = \int \ell(X, Y \mid \boldsymbol{\theta}) p(Y \mid X, \boldsymbol{\theta}^{(i)}) dY$$

- Maximization step: maximize the expected log-likelihood function.

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \; Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i)})$$