A Tutorial on Restricted Maximum Likelihood Estimation in Linear Regression and Linear Mixed-Effects Model

Xiuming Zhang
zhangxiuming@u.nus.edu

A*STAR-NUS Clinical Imaging Research Center

October 12, 2015

Summary

This tutorial derives in detail an estimation procedure—restricted maximum likelihood (ReML) [Patterson and Thompson, 1971] [Harville, 1974]—that is able to produce unbiased estimates for variance components of an linear model. We first introduce the concept of bias in variance components by maximum likelihood (ML) estimation in simple linear regression and then discuss a post hoc correction. Next, we apply ReML to the same model and compare the ReML estimate with the ML estimate followed by post hoc correction. Finally, we explain the linear mixed-effects (LME) model for longitudinal analysis [Bernal-Rusiel et al., 2013] and demonstrate how to obtain unbiased estimators of the parameters with ReML.

1 Linear Regression

Familiarity with basic linear regression facilitates the understanding of more complex linear models. We Therefore, start with this and introduce the concept of bias in estimating variance components.

1.1 The Model

The simplest linear regression is of the form $y = \beta_0+\beta_1 x+\epsilon$, where y is named **response** (or dependent variable or prediction), x is named **regressor** (or explanatory variable, independent variable), β's are **regression coefficients**, and ϵ is called **residual** (or

error), which is assumed to distribute as a zero-mean Gaussian with an unknown variance, i.e., $\mathcal{N}(0, \sigma^2)$.

When we have more than one regressor (a.k.a. **multiple linear regression**[1]), the model comes in its matrix form

$$y = X\beta + \epsilon, \tag{1}$$

where y is the response vector, X is the **design matrix** with each its row specifying under what design or conditions the corresponding response is observed (hence the name), β is the vector of regression coefficients, and ϵ is the residual vector distributing as a zero-mean multivariable Gaussian with a diagonal covariance matrix $\mathcal{N}(0, \sigma^2 I_N)$, where I_N is the $N \times N$ identity matrix. Therefore

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_N), \tag{2}$$

meaning that linear combination $X\beta$ explains (or predicts) response y with uncertainty characterized by a variance of σ^2.

As seen, this assumption on the covariance matrix mandates that (i) residuals among responses are independent, and (ii) residuals of all the responses have the same variance σ^2. This assumption makes parameter estimations straightforward, but meanwhile imposes some limitations on the model. For example, the model cannot handle properly intercorrelated responses, such as the longitudinal measurements of one individual. This motivates the usage of linear mixed-effects (LME) model in analyzing longitudinal data [Bernal-Rusiel et al., 2013].

1.2 Parameter Estimation

Under the model assumptions, we aim to estimate the unknown parameters (β and σ^2) from the data available (X and y). Maximum likelihood (ML) estimation is the most common estimator. We maximize the log-likelihood w.r.t. β and σ^2

$$\mathcal{L}(\beta, \sigma^2 \mid y, X) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)$$

and obtain

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

$$\hat{\sigma}^2 = \frac{1}{N}(y - X\hat{\beta})^T(y - X\hat{\beta}), \tag{4}$$

where N is the number of responses. $\hat{\beta}$ is simply the ordinary least squares (OLS) estimator of β, and we compute $\hat{\sigma}^2$ with the value of $\hat{\beta}$. As we will prove in the

[1] A single-response special case of **general linear model**, which itself is a special case of **generalized linear model** with identity link and normally distributed responses.

following section, estimator $\hat{\sigma}^2$ is biased downwards as compared with real value σ^2, because we neglect the loss of **degree of freedom** (DoF) for estimating β.

1.3 Estimation Bias in Variance Component

The **bias** of an estimator refers to the difference between this estimator's expectation (here $\mathrm{E}\{\hat{\sigma}^2\}$) and the true value (here σ^2). To facilitate computing $\mathrm{E}\{\hat{\sigma}^2\}$, we define matrix $A = X(X^TX)^{-1}X^T$, so $Ay = X\hat{\beta} = \hat{y}$. In fact, A is an **orthogonal projection** that satisfies $A^2 = A$ (**idempotent**) and $A^* = A$ (**Hermitian** or self-adjoint or simply symmetric for real matrices). We can then verify that $(I_N - A)^T(I_N - A) = I_N - A$.

$$\begin{aligned}
\mathrm{E}\{\hat{\sigma}^2\} &= \frac{1}{N}\mathrm{E}\{(y - X\hat{\beta})^T(y - X\hat{\beta})\} \\
&= \frac{1}{N}\mathrm{E}\{(y - Ay)^T(y - Ay)\} \\
&= \frac{1}{N}\mathrm{E}\{y^T(I - A)^T(I - A)y\} \\
&= \frac{1}{N}\mathrm{E}\{y^T(I - A)y\} \\
&= \frac{1}{N}\left(\mathrm{E}\{y^Ty\} - \mathrm{E}\{y^TAy\}\right)
\end{aligned}$$

Theorem 1.1

If $y \sim \mathcal{N}(0, I_N)$, and A is an orthogonal projection, then $y^TAy \sim \chi^2(k)$ with $k = \mathrm{rank}(A)$.

Proof. If A is idempotent, its eigenvalues satisfy $\lambda^2 = \lambda$. If A is also Hermitian, its eigenvalues are real[a]. Hence, in eigendecomposition $A = Q\Lambda Q^{-1} = Q\Lambda Q^{T\,b}$, Λ is a diagonal matrix containing either 0 or 1.
(i) When A is full-rank, $\Lambda = I_N$, A has to be I_N, and $y^TAy = y^Ty$. Then the result follows immediately from the definition of chi-square distribution.
(ii) When A is of rank $k < N$,

$$y^TAy = y^TQ\Lambda Q^Ty = W^T\begin{bmatrix} I_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}W = W_k^TW_k,$$

where $W = Q^Ty$, and W_k denotes the vector containing the first k elements of W. Since Q is orthogonal, $W \sim \mathcal{N}(0, I_N)$, so $W_k \sim \mathcal{N}(0, I_k)$. The result again follows directly from the definition of chi-square distribution. □

[a] See `http://proofwiki.org/wiki/Hermitian_Matrix_has_Real_Eigenvalues`.
[b] Eigenvectors of a real symmetric matrix are orthogonal, i.e., $Q^{-1} = Q^T$.

Recall Equation (2). By Theorem 1.1,

$$\left(\frac{y - X\beta}{\sigma}\right)^T \left(\frac{y - X\beta}{\sigma}\right) \sim \chi^2(N).$$

Hence,

$$\mathrm{E}\{y^T y\} = N\sigma^2 + (X\beta)^T(X\beta).$$

Similarly,

$$\left(\frac{y - X\beta}{\sigma}\right)^T A \left(\frac{y - X\beta}{\sigma}\right) \sim \chi^2(k),$$

where $k = \mathrm{rank}(A)$. Hence,

$$\mathrm{E}\{y^T A y\} = k\sigma^2 + (X\beta)^T(X\beta).$$

Substituting $\mathrm{E}\{y^T y\}$ and $\mathrm{E}\{y^T A y\}$ gives

$$\mathrm{E}\{\hat{\sigma}^2\} = \frac{N - k}{N}\sigma^2 < \sigma^2, \tag{5}$$

where $k = \mathrm{rank}(A)$ is just the number of columns (regressors) in X. Therefore, our estimation of the **variance component** is biased downwards. This bias is especially severe when we have many regressors (a large k), in which case we need to correct this bias by simply multiplying a factor of $N/(N-k)$. Hence, the corrected, unbiased estimator becomes

$$\begin{aligned}\hat{\sigma}^2_{\text{unbiased}} &= \frac{1}{N-k}(y - X\hat{\beta})^T(y - X\hat{\beta}) \\ &= \frac{1}{N-k}(y - X(X^TX)^{-1}X^Ty)^T(y - X(X^TX)^{-1}X^Ty),\end{aligned} \tag{6}$$

which is classically used in linear regression [Verbeke and Molenberghs, 2009].

2 Restricted Maximum Likelihood

In simple problems where solutions to variance components are closed-form (like linear regression above), we can remove the bias post hoc by multiplying a correction factor. However, for complex problems where closed-form solutions do not exist, we need to resort to a more general method to obtain a bias-free estimation for variance components. **Restricted maximum likelihood** (ReML) [Patterson and Thompson, 1971] [Harville, 1974] is one such method.

2.1 The Theory

Generally, estimation bias in variance components originates from the DoF loss in estimating mean components. If we estimated variance components with true mean

component values, the estimation would be unbiased. The intuition behind ReML is to maximize a modified likelihood that is free of mean components instead of the original likelihood as in ML.

Consider a *general* linear regression model

$$y = X\beta + \epsilon,$$

where y is still an N-vector of responses, X is still an $N \times k$ design matrix, but residual ϵ is no longer assumed to distribute as $\mathcal{N}(0, \sigma^2 I_N)$, but rather $\mathcal{N}(0, H(\theta))$, where $H(\theta)$ is a *general* covariance matrix parametrized by θ. For simplicity, $H(\theta)$ is often written as just H. Previously, we have been referring to θ as "variance components."

If vector a is orthogonal to all columns of X, i.e., $a^T X = 0$, then $a^T y$ is known as an **error contrast**. We can find at most $N-k$ such vectors that are linearly independent[2]. Define $A = \begin{bmatrix} a_1 & a_2 & \dots & a_{N-k} \end{bmatrix}$. It follows that $A^T X = \mathbf{0}$ and $\mathrm{E}\{A^T y\} = 0$. $S = I_N - X(X^T X)^{-1} X^T$ is a candidate for A, as $SX = \mathbf{0}$. Furthermore, it can be shown $AA^T = S$ and $A^T A = I_N$.

The error contrast vector

$$w = A^T y = A^T(X\beta + \epsilon) = A^T \epsilon \sim \mathcal{N}(0, A^T H A)$$

is free of β. [Patterson and Thompson, 1971] has proven that in the absence of information on β, no information about θ is lost when inference is based on w rather than on y. We can now directly estimate θ by maximizing a "restricted" log-likelihood function $\mathcal{L}_w(\theta \mid A^T y)$. This bypasses estimating β first and can Therefore, produce unbiased estimates for θ.

Once $H(\theta)$ is known, the **generalized least squares** (GLS) solution to β minimizing squared Mahalanobis length of the residual $(Y - X\beta)^T H^{-1}(Y - X\beta)$ is just

$$\hat{\beta} = (X^T H^{-1} X)^{-1} X^T H^{-1} y. \tag{7}$$

We now derive a convenient expression for $\mathcal{L}_w(\theta \mid A^T y)$ [Harville, 1974].

$$\begin{aligned}
\mathcal{L}_w(\theta \mid A^T y) &= \log f_w(A^T y \mid \theta) \\
&= \log f_w(A^T y \mid \theta) \int f_{\hat{\beta}}(\hat{\beta} \mid \beta, \theta)\, \mathrm{d}\hat{\beta} \\
&= \log f_w(A^T y \mid \theta) \int f_{\hat{\beta}}(G^T y \mid \beta, \theta)\, \mathrm{d}\beta \qquad (\hat{\beta}, \beta \text{ exchangeable here}) \\
&= \log \int f_w(A^T y \mid \theta) f_{\hat{\beta}}(G^T y \mid \beta, \theta)\, \mathrm{d}\beta \\
&= \log \int f_{w,\hat{\beta}}(A^T y, G^T y \mid \beta, \theta)\, \mathrm{d}\beta
\end{aligned}$$

[2]Imagine a plane spanned by $k = 2$ linearly independent vectors in three-dimensional space ($N = 3$). We can find at most $N - k = 1$ vector (passing the origin) orthogonal to the plane.

$$= \log \int f_y \left(\begin{bmatrix} A & G \end{bmatrix}^T y \mid \beta, \theta \right) \mathrm{d}\beta$$
$$= \log \frac{1}{|\det \begin{bmatrix} A & G \end{bmatrix}|} \int f_y \left(y \mid \beta, \theta \right) \mathrm{d}\beta.$$

Interlude 2.1

We express $|\det \begin{bmatrix} A & G \end{bmatrix}|$ in terms of X.

$$\begin{aligned}
|\det \begin{bmatrix} A & G \end{bmatrix}| &= \left(\det \begin{bmatrix} A & G \end{bmatrix}^T \begin{bmatrix} A & G \end{bmatrix} \right)^{\frac{1}{2}} \\
&= \left(\det \begin{bmatrix} A^T G & A^T G \\ G^T A & G^T G \end{bmatrix} \right)^{\frac{1}{2}} \\
&= \left(\det A^T A \right)^{\frac{1}{2}} \left(\det G^T G - G^T A (A^T A)^{-1} A^T G \right)^{\frac{1}{2}} \\
&= \left(\det I \right)^{\frac{1}{2}} \left(\det G^T G - G^T A I^{-1} A^T G \right)^{\frac{1}{2}} \\
&= \left(\det G^T G - G^T S G \right)^{\frac{1}{2}} \\
&= \left(\det X^T X \right)^{-\frac{1}{2}}
\end{aligned}$$

We continue deriving

$$\begin{aligned}
\mathcal{L}_w(\theta \mid A^T y) &= \log \frac{1}{|\det \begin{bmatrix} A & G \end{bmatrix}|} \int f_y \left(y \mid \beta, \theta \right) \mathrm{d}\beta \\
&= \log \left(\det X^T X \right)^{\frac{1}{2}} \int f_y \left(y \mid \beta, \theta \right) \mathrm{d}\beta \\
&= \log \left(\det X^T X \right)^{\frac{1}{2}} \int \frac{1}{\sqrt{(2\pi)^N \det H}} \exp \left(-\frac{1}{2} (y - X\beta)^T H^{-1} (y - X\beta) \right) \mathrm{d}\beta \\
&= \log \left(\det X^T X \right)^{\frac{1}{2}} (2\pi)^{-\frac{N}{2}} (\det H)^{-\frac{1}{2}} \int \exp \left(-\frac{1}{2} (y - X\beta)^T H^{-1} (y - X\beta) \right) \mathrm{d}\beta
\end{aligned}$$

Interlude 2.2

We can decompose $(y - X\beta)^T H^{-1} (y - X\beta)$ into

$$(y - X\hat{\beta})^T H^{-1} (y - X\hat{\beta}) + (\beta - \hat{\beta})^T (X^T H^{-1} X)(\beta - \hat{\beta})$$

with Equation (7).

We resume

$$\begin{aligned}
\mathcal{L}_w(\theta \mid A^T y) &= \log \left(\det X^T X \right)^{\frac{1}{2}} (2\pi)^{-\frac{N}{2}} (\det H)^{-\frac{1}{2}} \int \exp \left(-\frac{1}{2} (y - X\beta)^T H^{-1} (y - X\beta) \right) \mathrm{d}\beta \\
&= \log \left(\det X^T X \right)^{\frac{1}{2}} (2\pi)^{-\frac{N}{2}} (\det H)^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (y - X\hat{\beta})^T H^{-1} (y - X\hat{\beta}) \right)
\end{aligned}$$

$$
\begin{aligned}
&\int \exp\left(-\frac{1}{2}(\beta-\hat{\beta})^T(X^TH^{-1}X)(\beta-\hat{\beta})\right)\mathrm{d}\beta \\
&= \log\left(\det X^TX\right)^{\frac{1}{2}}(2\pi)^{-\frac{N}{2}}(\det H)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y-X\hat{\beta})^TH^{-1}(y-X\hat{\beta})\right) \\
&(2\pi)^{\frac{k}{2}}(\det X^TH^{-1}X)^{-\frac{1}{2}} \qquad \text{(a Gaussian integral)} \\
&= \log(2\pi)^{-\frac{1}{2}(N-k)}\left(\det X^TX\right)^{\frac{1}{2}}(\det H)^{-\frac{1}{2}}(\det X^TH^{-1}X)^{-\frac{1}{2}} \\
&\exp\left(-\frac{1}{2}(y-X\hat{\beta})^TH^{-1}(y-X\hat{\beta})\right) \\
&= -\frac{1}{2}(N-k)\log(2\pi)+\frac{1}{2}\log\det X^TX-\frac{1}{2}\log\det H-\frac{1}{2}\log\det X^TH^{-1}X \\
&-\frac{1}{2}(y-X\hat{\beta})^TH^{-1}(y-X\hat{\beta}), \qquad (8)
\end{aligned}
$$

where

$$\hat{\beta} = (X^TH^{-1}X)^{-1}X^TH^{-1}y. \qquad (9)$$

With this convenient expression, we can maximize the restricted log-likelihood $\mathcal{L}_w(\theta \mid A^Ty)$ w.r.t. variance components θ to obtain an unbiased estimate for the covariance matrix $H(\hat{\theta})$ and the corresponding regression coefficient estimates $\hat{\beta}$. Newton-Raphson method is usually employed. For more computational details, see [Lindstrom and Bates, 1988].

2.2 Applied to Simplest Linear Regression

We have seen the estimation bias in θ by ML from Equation (5). In the simplest form of linear regression where we assume $H = \sigma^2 I_N$, estimation $\hat{\sigma}^2$ is closed-form (Equation (4)), allowing us to correct the bias simply with a multiplicative factor. In this section, we verify that, in the simplest form of linear regression, the ReML method produces exactly the same solutions as ML method followed by the post hoc correction. Set

$$
\begin{aligned}
\frac{d}{d\sigma^2}\mathcal{L}_w(\sigma^2 \mid A^Ty) &= \frac{d}{d\sigma^2}-\frac{1}{2}\log\det H-\frac{1}{2}\log\det X^TH^{-1}X-\frac{1}{2}(y-X\hat{\beta})^TH^{-1}(y-X\hat{\beta}) \\
&= \frac{d}{d\sigma^2}-\frac{1}{2}(N-k)\log\sigma^2-\frac{1}{2\sigma^2}(y-X\hat{\beta})^T(y-X\hat{\beta}) \\
&= 0
\end{aligned}
$$

We obtain exactly the same result as Equation (6), produced by post hoc correction.

It is worth noticing that in this simplest linear regression case, the mean estimate $\hat{\beta}$ is independent of the variance component θ (Equation (3)). This implies although the ML and ReML estimates of $\hat{\sigma}^2$ are different, the estimates of $\hat{\beta}$ are the same. This is no longer true for more complex regression models, such as the linear mixed-effects model, as to be seen in the next section. Thus, for those complex models, we have a

ReML estimate of θ and also a "ReML" estimate of β, both being different from their ML estimates.

3 Linear Mixed-Effects Model

Longitudinal data are (usually non-uniformly) ordered in time, and missing data are very common. Furthermore, serial measurements of one subject are positively correlated, and between-subject variance is not constant over time due to possible diverging trajectories [Bernal-Rusiel et al., 2013]. The linear mixed-effects (LME) model [Laird and Ware, 1982] is a suitable model to handle such data.

3.1 The Model

The N_i serial measurements y_i of subject i are modeled as

$$y_i = X_i\beta + Z_i b_i + \epsilon_i,$$

where X_i is the $n_i \times p$ subject design matrix for the fixed effects (e.g., gender, education, clinical group), β is a p-vector of fixed effects regression coefficients to be estimated, Z_i is a $N_i \times q$ design matrix for the random effects, b_i is a q-vector of random effects, and ϵ is a N_i-vector of residuals. Z_i's columns are a subset of X_i's, linking random effects b_i to Y_i. That is, any component of β can be allowed to vary randomly by simply including the corresponding columns of X_i in Z_i [Bernal-Rusiel et al., 2013]. For example, to allow each subject to have their own trajectory intercepts, we set Z_i to be a N_i-vector of 1's. The following distributional assumptions are made

$$\begin{aligned} &b_i \sim \mathcal{N}(0, D) \\ &\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{N_i}) \\ &\epsilon_1, \ldots, \epsilon_M, b_1, \ldots, b_M \text{ independent,} \end{aligned}$$

where D and $\sigma^2 I_N$ are covariance matrices of multivariate Gaussian distributions, M is the total number of subjects, b_i reflects how the subset of regression coefficients for subject i deviates from those of the population, and ϵ_i represents residuals not explained by fixed or random effects. This allows subjects to deviate from the population, accounting for **inter-subject variability**.

Intuitively, the LME model extends simple multiple regression $y = X\beta + \epsilon$ by allowing for the additions of zero Gaussian noise b_i to a subset of the regression coefficients. More formally, the introduction of random effects helps distinguish the conditional (subject-specific) mean $\mathrm{E}\{y_i \mid b_i\}$ and marginal (population-average) mean $\mathrm{E}\{y_i\}$:

$$\mathrm{E}\{y_i \mid b_i\} = X_i\beta + Z_i b_i$$

$$\mathrm{E}\{y_i\} = X_i\beta$$

as well as subject-specific covariance $\mathrm{Cov}\{y_i \mid b_i\}$ and population-average covariance $\mathrm{Cov}\{y_i\}$:

$$\begin{aligned}\mathrm{Cov}\{y_i \mid b_i\} &= \mathrm{Cov}\{\epsilon_i\} = \sigma^2 I_{N_i} \\ \mathrm{Cov}\{y_i\} &= \mathrm{Cov}\{Z_i b_i\} + \mathrm{Cov}\{\epsilon_i\} = Z_i D Z_i^T + \sigma^2 I_{N_i},\end{aligned}$$

which is *not* a diagonal matrix (cf. the diagonal covariance matrix $\sigma^2 I_N$ in linear regression). Therefore, Z_i and D give a "structure" to the originally diagonal matrix, in turn allowing us to model **intra-subject measurement correlations**.

Finally, for each subject, we have

$$y_i \sim \mathcal{N}(X_i\beta, H_i(\theta)),$$

where $H_i(\theta) = Z_i D Z_i^T + \sigma^2 I_{N_i}$. See [Bernal-Rusiel et al., 2013] and [Verbeke and Molenberghs, 2009] for real-world applications of the LME model.

3.2 Estimation by Restricted Maximum Likelihood

We aim to estimate fixed effects regression coefficients β as well as model parameters σ^2 and D from the data. The solutions to variance components θ (i.e., σ^2 and D) are not closed-form. Therefore, we need to perform ReML rather than ML followed by post hoc correction to obtain unbiased estimates of θ.

To estimate these quantities, which are shared across all subjects, we need to stack up all subjects' data

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon,$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} \quad \mathbf{Z} = \begin{bmatrix} Z_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & Z_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & Z_M \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_M \end{bmatrix}.$$

That is,

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\beta, \mathbf{H}(\theta) = \begin{bmatrix} H_1(\theta) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & H_2(\theta) & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & H_M(\theta) \end{bmatrix}\right),$$

where $\mathbf{H}(\theta)$ (or $\mathbf{H}$ for short) is a block diagonal matrix.

Substitute $y = \mathbf{y}$, $X = \mathbf{X}$, and $H = \mathbf{H}$ into Equation (8) and drop constant terms

$$\begin{aligned}
\mathcal{L}_w(\theta \mid A^T y) &= -\frac{1}{2}\log\det H - \frac{1}{2}\log\det X^T H^{-1} X - \frac{1}{2}(y - X\hat{\beta})^T H^{-1}(y - X\hat{\beta}) \\
&= -\frac{1}{2}\log\det \mathbf{H} - \frac{1}{2}\log\det \mathbf{X}^T\mathbf{H}^{-1}\mathbf{X} - \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\beta})^T\mathbf{H}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \\
&= -\frac{1}{2}\log\prod_{i=1}^{M}\det H_i - \frac{1}{2}\log\prod_{i=1}^{M}\det X_i^T H_i^{-1} X_i - \frac{1}{2}\sum_{i=1}^{M}(y_i - X_i\hat{\beta})^T H_i^{-1}(y_i - X_i\hat{\beta}) \\
&= -\frac{1}{2}\sum_{i=1}^{M}\log\det H_i - \frac{1}{2}\sum_{i=1}^{M}\log\det X_i^T H_i^{-1} X_i - \frac{1}{2}\sum_{i=1}^{M}(y_i - X_i\hat{\beta})^T H_i^{-1}(y_i - X_i\hat{\beta}),
\end{aligned}$$

where

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}^{-1}\mathbf{y} \\
&= \left(\sum_{i=1}^{M} X_i^T H_i^{-1} X_i\right)^{-1}\sum_{i=1}^{M} X_i^T H_i^{-1} y_i.
\end{aligned}$$

We can then maximize $\mathcal{L}_w(\theta \mid A^T y)$ w.r.t. β, σ^2, and D. Computational details are found in [Lindstrom and Bates, 1988]. Details on hypothesis testing can be found in [Bernal-Rusiel et al., 2013].

References

[Bernal-Rusiel et al., 2013] Bernal-Rusiel, J. L., Greve, D. N., Reuter, M., Fischl, B., and Sabuncu, M. R. (2013). Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *NeuroImage*, 66:249–260.

[Harville, 1974] Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383–385.

[Laird and Ware, 1982] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

[Lindstrom and Bates, 1988] Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

[Patterson and Thompson, 1971] Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

[Verbeke and Molenberghs, 2009] Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.