**2nd Talk**

# The Linear Mixed-Effects Probability Model

## So far:

- Micro Introduction
- Examples of different models and datasets

## In this talk:

- Formalize notation
- Introduce new definitions
- Model fitting (as done in the *lme4*-package)

# Recall a basic example of a model
(workers and machines)

**Model:**

$$y_{ijk} = \mu + \beta_j + b_i + \epsilon_{ijk}, \qquad i = 1, ..., 6 \qquad j = 1, 2, 3 \qquad k = 1, 2, 3$$

$\beta_j$ : effect of machine (fixed)

$b_i$: effect of worker (random)

**Assumption:**

$$b_i \sim \mathcal{N}(0, \sigma_b^2) \ \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$$

# Generalization:

Two random variables:

$\mathcal{Y}$: the $n$-dimensional response vector

$\mathcal{B}$: the $q$-dimensional vector of random effects

$$\mathcal{B} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\theta). \tag{1}$$

$$(\mathcal{Y}|\mathcal{B} = \boldsymbol{b}) \sim \mathcal{N}\left(\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\beta, \sigma^2\boldsymbol{I}_n\right) \tag{2}$$

So the *linear predictor* is

$$\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\beta \tag{3}$$

With the *model matrices* $\boldsymbol{Z}$ of dimension $n \times q$

and $\boldsymbol{X}$ of dimension $n \times p$,

where $p$ is the dimension of the *fixed-effects* parameter vector $\boldsymbol{\beta}$

# More definitions:

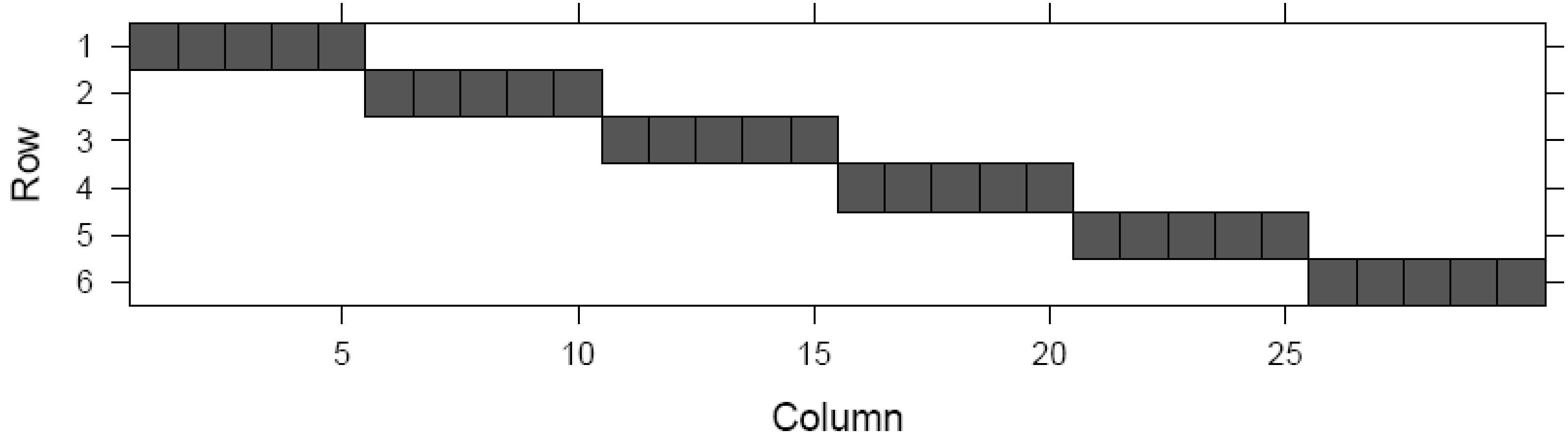
$\boldsymbol{\theta}$ : *variance-component parameter vector*

$\boldsymbol{\Sigma}_{\theta}$ : *variance-covariance matrix*

$\boldsymbol{\sigma}$ : common scale parameter

The form of the random-effects model matrix, $\boldsymbol{Z}$, and the form of the variance-covariance matrix, $\boldsymbol{\Sigma}_{\theta}$, and the method by which $\boldsymbol{\Sigma}_{\theta}$ is determined from the value of $\boldsymbol{\theta}$ are all based on the random-effects terms in the model formula.

$\boldsymbol{Z}$ can be large, but it is sparse (i.e. most elements in the matrix are zero).



# More definitions:

$\mathbf{\Lambda}_\theta$: relative covariance factor, defined so that

$$\mathbf{\Sigma}_\theta = \sigma^2 \mathbf{\Lambda}_\theta \mathbf{\Lambda}_\theta^T$$

where $\sigma^2$ is the same variance parameter as in $(Y|B = \boldsymbol{b})$.

With the spherical random effects:

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 I_q)$$

we get:

$$\mathcal{B} = \mathbf{\Lambda}_\theta \mathcal{U}$$

So we really have:

$$E[\mathcal{B}] = \mathbf{\Lambda}_\theta E[\mathcal{U}] = \mathbf{\Lambda}_\theta 0 = 0$$

and:

$$
\begin{aligned}
Var(\mathcal{B}) &= E[(\mathcal{B} - E[\mathcal{B}])(\mathcal{B} - E[\mathcal{B}])^T] = E[\mathcal{B}\mathcal{B}^T] \\
&= E[\mathbf{\Lambda}_\theta \mathcal{U}\mathcal{U}^T \mathbf{\Lambda}_\theta^T] = \mathbf{\Lambda}_\theta E[\mathcal{U}\mathcal{U}^T]\mathbf{\Lambda}_\theta^T = \mathbf{\Lambda}_\theta Var(\mathcal{U})\mathbf{\Lambda}_\theta^T \\
&= \mathbf{\Lambda}_\theta \boldsymbol{\sigma}^2 I_q \mathbf{\Lambda}_\theta^T = \boldsymbol{\sigma}^2 \mathbf{\Lambda}_\theta \mathbf{\Lambda}_\theta^T = \mathbf{\Sigma}_\theta
\end{aligned}
$$

In our discussion we will concentrate on $\boldsymbol{\Lambda}_\theta$ (not $\boldsymbol{\Sigma}_\theta$) and $\mathcal{U}$ (not $\mathcal{B}$):

So we look at:

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 I_q). \tag{4}$$

$$(\mathcal{Y}|\mathcal{U} = \boldsymbol{u}) \sim \mathcal{N}\left(\boldsymbol{Z}\boldsymbol{\Lambda}_\theta u + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n\right) \tag{5}$$

So the *linear predictor* becomes:

$$\gamma = \boldsymbol{Z}\boldsymbol{\Lambda}_{\theta}u + \boldsymbol{X}\boldsymbol{\beta} \tag{6}$$

And the *conditional mean* of $\mathcal{Y}$, given $\mathcal{U} = \boldsymbol{u}$:

$$\mu = E[\mathcal{Y}|\mathcal{U} = \boldsymbol{u}] \tag{7}$$

Note: For a linear mixed model, we obviously have $\mu = \gamma$.

In other forms of mixed models this may not be the case anymore.

# Conditional Distribution:

**Notation:**

$y_{obs}$ : an observed data vector (an actual realization of $\mathcal{Y}$)

$y$ : an arbitrary value of $\mathcal{Y}$

Now we are interested in the conditional distribution of $(\mathcal{U}|\mathcal{Y} = \boldsymbol{y})$

Our model parameters are $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$.

The likelihood of those parameters, given the observed data, $\boldsymbol{y}_{obs}$, is the probability density of $\mathcal{Y}$, evaluated at $\boldsymbol{y}_{obs}$.

Parameters fixed, y varying

y fixed at $y_{obs}$, parameters varying

## Natural approach for evaluating the likelihood:

1. Determine marginal distribution of $\mathcal{Y}$

- Determine joint density of $\mathcal{U}$ and $\mathcal{Y}$: $f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}, \boldsymbol{u})$
- Integrate this density wrt. $\boldsymbol{u}$ to get the marginal density: $f_{\mathcal{Y}}(\boldsymbol{y})$

2. Evaluate that density at $\boldsymbol{y}_{obs}$.

But here we choose a different order of the steps, that is:

Evaluate the joint density at $y_{obs}$ to produce an intermediate function $h(\boldsymbol{u})$.

And then integrate this function $h(\boldsymbol{u})$ along $\boldsymbol{u}$.

This does not work generally, it could even happen that the joint density does not exist (think of a joint distribution that is discrete wrt. $\boldsymbol{y}$ and continous wrt. $\boldsymbol{u}$)

We define:

$$h(u) = f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{obs}, \boldsymbol{u}) \tag{9}$$

the *unnormalized conditional density.*

We see that:

$$f_{\mathcal{U}|\mathcal{Y}}(\boldsymbol{u}|\boldsymbol{y}_{obs}) = \frac{h(\boldsymbol{u})}{\int_{R^q} h(\boldsymbol{u})\, d\boldsymbol{u}} \tag{10}$$

and thus the likelihood is:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma|\boldsymbol{y}_{obs}) = \int_{R^q} f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{obs}, \boldsymbol{u})\, d\boldsymbol{u} = \int_{R^q} h(\boldsymbol{u})\, d\boldsymbol{u}. \tag{11}$$

## Tools to evaluate the likelihood in general situations:

$\tilde{\boldsymbol{u}}$ : the conditional mode of $\boldsymbol{u}$, given $\mathcal{Y} = \boldsymbol{y}_{obs}$:

$$\tilde{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}} f_{\mathcal{U}|\mathcal{Y}}(\boldsymbol{u}|\boldsymbol{y}_{obs}) = \arg\max_{\boldsymbol{u}} h(\boldsymbol{u}) = \arg\max_{\boldsymbol{u}} f_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{y}_{obs}|\boldsymbol{u}) f_{\mathcal{U}}(\boldsymbol{u}) \tag{12}$$

Recall (4) and (5):

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 I_q). \tag{4}$$

$$(\mathcal{Y}|\mathcal{U} = \boldsymbol{u}) \sim \mathcal{N}\left(\boldsymbol{Z\Lambda_\theta} u + \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n\right) \tag{5}$$

Thus we have:

$$f_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{y}|\boldsymbol{u}) = \frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z\Lambda_\theta}\,\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{n/2}} \tag{13}$$

$$f_{\mathcal{U}}(\boldsymbol{u}) = \frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{q/2}} \tag{14}$$

And so the product:

$$h(\boldsymbol{u}) = \frac{\exp(-\left[\|\boldsymbol{y}_{obs} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\,\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2\right]/(2\sigma^2))}{(2\pi\sigma^2)^{(n+q)/2}} \tag{15}$$

Looking at the negative log density, we get:

$$-2\log(h(\boldsymbol{u})) = (n+q)\log(2\pi\sigma^2) + \frac{\|\boldsymbol{y}_{obs} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\,\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2}{\sigma^2} \tag{16}$$

So we get:

$$\tilde{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} \|\boldsymbol{y}_{obs} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\, \boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 \tag{17}$$

penalty for high complexity

sum of squared residuals

The expression to be minimized $\|\boldsymbol{y}_{obs} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\, \boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2$ is called the *objective function*, here it is a *penalized residual sum of squares* (**PRSS**).

The minimizer $\tilde{\boldsymbol{u}}$ is called the *penalized least squares* (**PLS**) solution

We think of the **PRSS** criterion as a function of the parameters, given the data, ie.:

$$\boldsymbol{r}^2_{\theta,\beta} = \min_{\boldsymbol{u}} \left[ \|\boldsymbol{y}_{obs} - \boldsymbol{X\beta} - \boldsymbol{Z}\Lambda_\theta \, \boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 \right] \tag{18}$$

We can also minimize this expression wrt $\beta$.
And we will see that this can even be done simultaniously wrt $\boldsymbol{u}$ and $\beta$ without using iterations. The minimum value we get is:

$$\boldsymbol{r}^2_{\theta} = \min_{\boldsymbol{u},\beta} \left[ \|\boldsymbol{y}_{obs} - \boldsymbol{X\beta} - \boldsymbol{Z}\Lambda_\theta \, \boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 \right] \tag{19}$$

$\tilde{\beta}$: *conditional estimate* of $\beta$

the value of $\beta$ for which the minimum in (19) is attained.

One way to determine the solution is to rephrase it as a linear least squares problem for an extended residual vector

$$\tilde{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} \left\| \begin{bmatrix} \boldsymbol{y}_{obs} - \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z}\boldsymbol{\Lambda}_\theta \\ \boldsymbol{I}_q \end{bmatrix} \boldsymbol{u} \right\|^2 \tag{20}$$

This is called a pseudo-data approach because we create the effect of the penalty term, $\|\boldsymbol{u}\|^2$, by adding "pseudo-observations" to the response vector and to the predictor (adding zeros and $I_q\boldsymbol{u}$).

For this linear least squares problem, we can give the solution by solving the normal equations. So we get that the solution satisfies:

$$(\Lambda_\theta^T \boldsymbol{Z}^T \boldsymbol{Z}\Lambda_\theta + \boldsymbol{I}_q)\tilde{\boldsymbol{u}} = \Lambda_\theta^T \boldsymbol{Z}^T(\boldsymbol{y}_{obs} - \boldsymbol{X}\boldsymbol{\beta}) \tag{21}$$

We want fast evaluation of $\tilde{\boldsymbol{u}}$ for different inputs, so we form the *sparse Cholesky fatcor*, $\boldsymbol{L}_\theta$.
It is a lower $q$ x $q$ matrix with:

$$\boldsymbol{L}_\theta \boldsymbol{L}_\theta^T = (\Lambda_\theta^T \boldsymbol{Z}^T \boldsymbol{Z} \Lambda_\theta + \boldsymbol{I}_q) \tag{22}$$

In order to get a sparse Cholesky factor $\boldsymbol{L}_\theta$ we might want to permutate the columns of our data.

This is done through a so-called *Permutation matrix* $\boldsymbol{P}$.

We also call them *fill-reducing permutations* as we want to avoid positions in the factor getting filled, where the matrix being decomposed is zero.

(22) thus becomes:

$$\boldsymbol{L}_\theta \boldsymbol{L}_\theta^T = \boldsymbol{P}(\Lambda_\theta^T \boldsymbol{Z}^T \boldsymbol{Z} \Lambda_\theta + \boldsymbol{I}_q)\boldsymbol{P}^T \tag{23}$$

The pseudo-data representation in (20) becomes:

$$\tilde{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} \left\| \begin{bmatrix} \boldsymbol{y}_{obs} - \boldsymbol{X\beta} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z\Lambda_\theta P}^T \\ \boldsymbol{P}^T \end{bmatrix} \boldsymbol{Pu} \right\|^2 \tag{24}$$

And the system of linear equations for $\tilde{\boldsymbol{u}}$ accordingly:

$$\boldsymbol{L}_\theta \boldsymbol{L}_\theta^T \boldsymbol{P}\tilde{\boldsymbol{u}} = \boldsymbol{P}(\Lambda_\theta^T \boldsymbol{Z}^T(\boldsymbol{y}_{obs} - \boldsymbol{X\beta}))\boldsymbol{P}^T\boldsymbol{P}\tilde{\boldsymbol{u}} = \boldsymbol{P}\Lambda_\theta^T \boldsymbol{Z}^T(\boldsymbol{y}_{obs} - \boldsymbol{X\beta}) \tag{25}$$

Note: Once we evaluate $L_\theta$ it is straight forward to solve (25) for $\tilde{\boldsymbol{u}}$. Thus this step is very crucial, and the ability to evaluate $L_\theta$ rapidly for many different values of $\boldsymbol{\theta}$ is what makes the methods in `lme4` feasible.

# Back to the evaluation of the likelihoood:

We've seen in (11) and (15) that:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma | \boldsymbol{y}_{obs}) = \int_{R^q} f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{obs}, \boldsymbol{u})\, d\boldsymbol{u} = \int_{R^q} h(\boldsymbol{u})\, d\boldsymbol{u}.$$

$$h(\boldsymbol{u}) = \frac{\exp(-\left[||\boldsymbol{y}_{obs} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\, \boldsymbol{u}||^2 + \|\boldsymbol{u}\|^2\right]/(2\sigma^2))}{(2\pi\sigma^2)^{(n+q)/2}}$$

We can now write the **PRSS** for general $\boldsymbol{u}$ as:

$$||\boldsymbol{y}_{obs} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\, \boldsymbol{u}||^2 + ||\boldsymbol{u}||^2 = r^2_{\theta,\beta} + ||\boldsymbol{L}_\theta^T(\boldsymbol{u} - \tilde{\boldsymbol{u}})||^2 \qquad (26)$$

Plugging this into the definition of $h(\boldsymbol{u})$ and using the change-of-variable:

$$z = \frac{\boldsymbol{L}_\theta^T(\boldsymbol{u} - \tilde{\boldsymbol{u}})}{\sigma} \qquad (27)$$

We get after a calculation (Bates[10], ch. 5.4.2 - available on http://lme4.r-forge.r-project.org/book/):

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma | \boldsymbol{y}_{obs}) = \frac{\exp(-\frac{r^2_{\theta,\beta}}{2\sigma^2})}{(2\pi\sigma^2)^{n/2}|\boldsymbol{L}_\theta|}$$

So the deviance (negative twice the log-likelihood) becomes:

$$d(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma | \boldsymbol{y}_{obs}) = -2\log(L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma | \boldsymbol{y}_{obs})) = n\log(2\pi\sigma^2) + \frac{r^2_{\boldsymbol{\theta},\boldsymbol{\beta}}}{\sigma^2} + 2\log(|\boldsymbol{L}_\theta|^2)$$

So the maximum-likelihood estimates for the parameters are those that minimize this deviance.

We can even further simplify this expression by using the facts, that $\beta$ only occurs in $r^2_{\boldsymbol{\theta},\boldsymbol{\beta}}$ and minimizing this expression wrt $\beta$ for any value of $\theta$ goes back to the penalized least square problems.

So let $\hat{\beta}_\theta$ be the value of $\beta$ that minimizes PRSS wrt to $\beta$ and $u$.

And $r^2_\theta$ the PRSS at these minimizing values.

Furthermore let $\hat{\sigma_\theta}^2 = r^2_\theta/n$, the value of $\sigma^2$ that minimizes the above deviance or a given $r^2_\theta$.

Then the *profiled deviance*, which is now only a function of $\theta$, becomes:

$$\tilde{d}(\theta|\boldsymbol{y}_{obs}) = 2\log|\boldsymbol{L}_\theta| + n\left[1 + \log\left(\frac{2\pi r^2_\theta}{n}\right)\right]$$

Now minimization of $\tilde{d}(\theta|\boldsymbol{y}_{obs})$ wrt $\theta$ determines the MLE, $\tilde{\theta}$. The MLEs for $\hat{\beta}$ and $\hat{\sigma}$ then are the corresponding conditional estimates evaluated at $\hat{\theta}$.

Simultaniously evaluating $\tilde{\boldsymbol{u}}$ and $\beta_\theta$ uses the same approach we've already seen in (20), that is to rephrase the PLS problem into a linear least square problem.

Thus we rewrite:

$$\tilde{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} \left\| \begin{bmatrix} \boldsymbol{y}_{obs} - \boldsymbol{X\beta} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z\Lambda}_\theta \\ \boldsymbol{I}_q \end{bmatrix} \boldsymbol{u} \right\|^2 \tag{20}$$

as

$$\begin{bmatrix} \tilde{\boldsymbol{u}} \\ \hat{\beta}_\theta \end{bmatrix} = \arg\min_{\boldsymbol{u},\beta} \left\| \begin{bmatrix} \boldsymbol{y}_{obs} \\ 0 \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z\Lambda}_\theta \boldsymbol{P}^T & \boldsymbol{X} \\ \boldsymbol{P}^T & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{Pu} \\ \beta \end{bmatrix} \right\|^2 \tag{28}$$

Which now yields the equation:

$$\begin{bmatrix} \boldsymbol{P}(\boldsymbol{\Lambda}_\theta^T \boldsymbol{Z}^T \boldsymbol{Z} \boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q)\boldsymbol{P}^T & \boldsymbol{P}\boldsymbol{\Lambda}_\theta^T \boldsymbol{Z}^T \boldsymbol{X} \\ \boldsymbol{X}^T \boldsymbol{Z} \boldsymbol{\Lambda}_\theta \boldsymbol{P}^T & \boldsymbol{X}^T \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{P}\tilde{\boldsymbol{u}} \\ \hat{\beta}_\theta \end{bmatrix} = \begin{bmatrix} \boldsymbol{P}\boldsymbol{\Lambda}_\theta^T \boldsymbol{Z}^T \boldsymbol{y}_{obs} \\ \boldsymbol{X}^T \boldsymbol{y}_{obs} \end{bmatrix}$$

The Matrix on the LHS can be decomposed into a '*Cholesky-like*' decomposition.

^

This way we also found a fast way to get $\tilde{\boldsymbol{u}}$ and $\hat{\beta}_\theta$, thus we have found all MLE estimators of the parameters.

That is we fitted the model to the actual data.

## Outlook: Complete Analysis of several data examples