Introduction to General and Generalized Linear Models

Mixed effects models - III

Henrik Madsen
Poul Thyregod
Anders Nielsen

April 30, 2012

This lecture

- Bayesian interpretations
- Multivariate fixed and random effects model
- Bias versus Variance of estimators
- General mixed effects models
- Laplace approximation

Bayesian interpretations

Recall the Bayes' Theorem; here related to the conditional density of X for given $Y = y$:

$$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y) f_X(x)}{\int f_{Y|X=x}(y) f_X(x)\, dx} \tag{1}$$

- The distribution $f_X(\cdot)$ of X is called the *prior distribution*
- The conditional distribution with density function $f_{X|Y=y}(x)$ is called the *posterior distribution*.
- The conditional distribution with density $f_{Y|X=x}(y)$ is called the *likelihood function*.

In settings where X is a set off (unknown) parameters this reflects that the parameters in a Bayesian setting are considered as **random variables**. In a Bayesian framework the prior might express a so-called **subjective probability distribution**.

Variance-covariance separations

Recall also the rules relating conditional and marginal moments:

$$\mathrm{E}[\,Y] = \mathrm{E}_X[\,\mathrm{E}[\,Y|X]\,] \tag{2a}$$

$$\mathrm{Var}[\,Y] = \mathrm{E}_X[\,\mathrm{Var}[\,Y|X]\,] + \mathrm{Var}_X[\,\mathrm{E}[\,Y|X]\,] \tag{2b}$$

$$\mathrm{Cov}[\,Y, Z] = \mathrm{E}_X[\,\mathrm{Cov}[\,Y, Z]\,|X\,] + \mathrm{Cov}_X[\,\mathrm{E}[\,Y|X], \mathrm{E}[\,Z|X]\,] \tag{2c}$$

An example: Regression with random X.

Bayesian formulations

Examples where the 'state' x is not (directly) observed arises in many contexts (Hidden Markov Models, State Space Models (Kalman Filters), etc.)
In such cases it is often useful to use a Bayesian framework.

- As seen previously also the one-way random effects model may be formulated in a Bayesian framework, where we may identify the $\mathrm{N}(\cdot, \sigma_u^2)$-distribution of $\mu_i = \mu + U_i$ as the *prior distribution*.
- The statistical model for the data is such that for given μ_i, are the Y_{ij}'s independent and distributed like $\mathrm{N}(\mu_i, \sigma^2)$.
- In a Bayesian framework, the *conditional distribution* of μ_i given $\overline{Y}_i = \overline{y}_i$ is termed the *posterior distribution* for μ_i.

Posterior distribution of μ_i

Theorem (The posterior distribution of μ_i)

Consider again the one-way model with random effects (introduced on page 162)

$$Y_{ij}|\mu_i \sim N(\mu_i, \sigma^2) \tag{3a}$$

$$\mu_i \sim N(\mu, \sigma_u^2) \tag{3b}$$

where μ, σ^2 and σ_u^2 are known.
The posterior distribution of μ_i after observation of $y_{i1}, y_{i2}, \ldots, y_{in}$ is a normal distribution with mean and variance

$$\mathrm{E}[\mu_i \mid \overline{Y}_i = \overline{y}_i] = \frac{\mu/\sigma_u^2 + n_i\bar{y}_i/\sigma^2}{1/\sigma_u^2 + n_i/\sigma^2} = w\mu + (1-w)\overline{y}_i \tag{4a}$$

$$\mathrm{Var}[\mu_i|\overline{Y}_i = \overline{y}_i] = \frac{1}{\frac{1}{\sigma_u^2} + \frac{n}{\sigma^2}} \tag{4b}$$

where

$$w = \frac{\frac{1}{\sigma_u^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_u^2}} = \frac{1}{1+n\gamma} \tag{5}$$

with $\gamma = \sigma_u^2/\sigma^2$.

The multivariate case

The results can rather easily be generalized to the multivariate case (which for instance allow for more advance correlation structures) as shown in this Theorem:

Theorem (Posterior distribution for multivariate normal distributions)

Let $Y \mid \mu \sim N_p(\mu, \Sigma)$ *and let* $\mu \sim N_p(m, \Sigma_0)$*, where* Σ *and* Σ_0 *are of full rank,* p*, say.*
Then the posterior distribution of μ *after observation of* $Y = y$ *is given by*

$$\boldsymbol{\mu \mid Y = y \sim N_p(Wm + (I - W)y, (I - W)\Sigma)} \tag{6}$$

with $W = \Sigma(\Sigma_0 + \Sigma)^{-1}$ *and* $I - W = \Sigma_0(\Sigma_0 + \Sigma)^{-1}$

Multivariate measurements – Fixed effects

Let us now consider the situation where the individual observations are p-dimensional vectors. Let us first consider fixed effects:

- Consider the model

$$\boldsymbol{X}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}, \; \boldsymbol{i = 1, 2, \ldots, k}; \; \boldsymbol{j = 1, 2, \ldots, n_i} \tag{7}$$

  where $\boldsymbol{\mu}, \boldsymbol{\alpha}_i$ and $\boldsymbol{\epsilon}_{ij}$ denotes p-dimensional vectors and where $\boldsymbol{\epsilon}_{ij}$ are mutual independent and normally distributed, $\boldsymbol{\epsilon}_{ij} \in \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, and where $\boldsymbol{\Sigma}$ denotes the $p \times p$-dimensional covariance matrix. For simplicity we will assume that $\boldsymbol{\Sigma}$ has full rank.

- For the fixed effects model we further assume

$$\sum_{i=1}^{k} n_i \boldsymbol{\alpha}_i = 0$$

- Given these assumptions we find $\boldsymbol{Z}_i = \sum_j \boldsymbol{X}_{ij} \sim \mathrm{N}_p(\boldsymbol{n_i}(\boldsymbol{\mu} + \boldsymbol{\alpha}_i), \boldsymbol{n_i \Sigma})$.

The separation of variation

- In the case of multivariate observations the variation is described by $p \times p$-dimensional SS matrices.
- Let us introduce the notation

$$\overline{X}_{i+} = \sum_{j=1}^{n_i} X_{ij}/n_i \tag{8}$$

$$\overline{X}_{++} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}/N = \sum_{i=1}^{k} n_i \overline{X}_{i+} / \sum_{i=1}^{k} n_i \tag{9}$$

  as descriptions of the group averages and the total average, respectively.
- Furthermore introduce

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i+})(X_{ij} - \overline{X}_{i+})^T \tag{10}$$

$$SSB = \sum_{i=1}^{k} n_i(\overline{X}_{i+} - \overline{X}_{++})(\overline{X}_{i+} - \overline{X}_{++})^T \tag{11}$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{++})(X_{ij} - \overline{X}_{++})^T \tag{12}$$

  as a description of the variation between groups (SSE), between groups (SSB), and the total variation (SST).
- As previously we have the Pythagorean relation

$$SST = SSE + SSB \tag{13}$$

Random effects

Consider the following model with random effects:

$$\boldsymbol{X_{ij}} = \boldsymbol{\mu} + \boldsymbol{u_i} + \boldsymbol{\epsilon_{ij}}, \ \boldsymbol{i = 1, \ldots, k}; \ \boldsymbol{j = 1, 2, \ldots, n_i}. \tag{14}$$

where u_i now are independent, $\boldsymbol{u_i} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma_0})$, and where ϵ_{ij} is independent, $\boldsymbol{\epsilon_{ij}} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Finally, u and ϵ are independent.

Theorem (The marginal distribution in the case of multivariate p-dimensional observations)

Consider the model introduced in (14). Then the marginal density of $Z_i = \sum_j X_{ij}$ is

$$N_p(n_i\boldsymbol{\mu}, n_i\boldsymbol{\Sigma} + n_i^2\boldsymbol{\Sigma}_0)\text{-distribution} \tag{15}$$

and the marginal density for $\overline{X}_{i+}$ is

$$N_p(\boldsymbol{\mu}, \frac{1}{n_i}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0) \tag{16}$$

Finally, we have that SSE follows a Wishart distribution

$$SSE \in Wis_p(N - k, \boldsymbol{\Sigma}) \tag{17}$$

and SSE is independent of $\overline{X}_{i+}, \ i = 1, 2, \ldots, k$.

MLE estimation

Theorem (MLE for the multivariate random effects model)

Still under the assumptions mentioned previously we find the maximum likelihood estimates (MLEs) for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$ by maximizing the log-likelihood

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0; \overline{x}_{1+}, \ldots, \overline{x}_{k+}) = -\frac{N-k}{2}\log(\det(\boldsymbol{\Sigma})) - \frac{1}{2}\mathrm{tr}((SSE)\boldsymbol{\Sigma}^{-1}) - \sum_{i=1}^{k}\left[\log\left(\det\left(\frac{\boldsymbol{\Sigma}}{n_i} + \boldsymbol{\Sigma}_0\right)\right) + \frac{1}{2}(\overline{x}_{i+} - \boldsymbol{\mu})^T\left(\frac{\boldsymbol{\Sigma}}{n_i} + \boldsymbol{\Sigma}_0\right)^{-1}(\overline{x}_{i+} - \boldsymbol{\mu})\right] \tag{18}$$

with respect to $\boldsymbol{\mu} \in R^p$ and $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$ in the space of non-negative definite $p \times p$ matrices.

Proof.

Omitted, but follows from the fact that SSE follows a $\mathrm{Wis}_p(N-k, \boldsymbol{\Sigma})$-distribution and that SSE and $\overline{X}_{i+}$, $i = 1, 2, \ldots, k$ are independent, and further that $\overline{X}_{i+} \in \mathrm{N}_p(\mu, \boldsymbol{\Sigma}/n_i + \boldsymbol{\Sigma}_0)$ are independent, $i = 1, 2, \ldots, k$. □

Example - Variation between errors for flow meters

	Flow, $[\mathrm{m}^3/\mathrm{h}]$					
	Calibration 1		Calibration 2		Calibration 3	
Meter	0.1	0.5	0.1	0.5	0.1	0.5
41	-2.0	1.0	2.0	3.0	2.0	2.0
42	5.0	3.0	1.0	1.0	2.0	2.0
43	2.0	1.0	-3.0	-1.0	1.0	0.0
44	4.0	4.0	-1.0	2.0	3.0	5.0
45	4.0	2.0	0.0	1.0	-1.0	0.0
46	5.0	9.0	4.0	8.0	6.0	10.0

Table: Results of three repeated calibrations of six flow-meters at two flows.

Bias Variance trade off

In the previous lecture, we consideret ML versus REML estimation.
Say something - use the black board ...

Reminder: Multivariate normal distribution

The density for a k-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is:

$$L(\boldsymbol{z}) = \frac{1}{(2\pi)^{k/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right]$$

We write $\boldsymbol{Z} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The log is:

$$\ell(\boldsymbol{z}) = -\frac{1}{2}\left\{k\log(2\pi|\boldsymbol{\Sigma}|) + (\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right\}$$

General mixed effects models

- In the previous part of this course you have seen how to handle
  - General linear models
  - Generalized linear models
  - General linear mixed models
- For the rest of the course we shall look at nonlinear and non-normal mixed effects models.
- In general it is impossible to obtain closed form solutions and hence numerical methods must be used.
- Estimation and inference will be based on likelihood principle.

General mixed effects models

The general mixed effects model can be represented by its likelihood function:

$$L_M(\boldsymbol{\theta}; \boldsymbol{y}) = \int_{\mathbb{R}^q} L(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y}) d\boldsymbol{u}$$

- $\boldsymbol{y}$ is the observed random variables
- $\boldsymbol{u}$ is the q unobserved random variables
- $\boldsymbol{\theta}$ is the model parameters to be estimated

The likelihood function L is the joint likelihood of both the observed and the unobserved random variables.

The likelihood function for estimating $\boldsymbol{\theta}$ is the marginal likelihood L_M obtained by integrating out the unobserved random variables.

Notice we have already seen this

- In the Poisson distribution the variance is equal to the mean, which is an assumption that is not always valid.
- Consider the model:

$$Y \sim \text{Pois}(\lambda), \quad \text{where} \quad \lambda \sim \Gamma\left(n, \frac{1-\phi}{\phi}\right) \quad 0 < \phi < 1$$

- It can be shown (next lecture) that:

$$Y \sim \text{Nbinom}(n, \phi)$$

- Notice:
  - No λ in marginal likelihood for Y
  - Analytical integration is not the typical case

General mixed effects models

- The integral shown on the previous slide is generally difficult to solve if the number of unobserved random variables is more than a few, i.e. for large values of q.
- A large value of q significantly increases the computational demands due to the product rule which states that if an integral is sampled in m points per dimension to evaluate it, the total number of samples needed is m^q, which rapidly becomes infeasible even for a limited number of random effects.
- The likelihood function gives a very broad definition of mixed models: the only requirement for using mixed modeling is to define a joint likelihood function for the model of interest.
- In this way mixed modeling can be applied to any likelihood based statistical modeling.
- Examples of applications are linear mixed models (LMM) and nonlinear mixed models (NLMM), generalized linear mixed models, but also models based on Markov chains, ODEs or SDEs.

Hierarchical models

As for the Gaussian linear mixed models it is useful to formulate the model as a *hierarchical model* containing a *first stage model*

$$f_{Y|u}(\boldsymbol{y}; \boldsymbol{u}, \boldsymbol{\beta})$$

which is a model for the data given the random effects, and a *second stage model*

$$f_U(\boldsymbol{u}; \boldsymbol{\Psi})$$

which is a model for the random effects. The total set of parameters is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Psi})$. Hence the joint likelihood is given as

$$L(\boldsymbol{\beta}, \boldsymbol{\Psi}; \boldsymbol{u}, \boldsymbol{y}) = f_{Y|u}(\boldsymbol{y}; \boldsymbol{u}, \boldsymbol{\beta}) f_U(\boldsymbol{u}; \boldsymbol{\Psi})$$

Hierarchical models

To obtain the likelihood for the model parameters $\boldsymbol{\theta}$ the unobserved random effects are again integrated out.

The likelihood function for estimating $\boldsymbol{\theta}$ is as before the marginal likelihood

$$L_M(\boldsymbol{\theta}; \boldsymbol{y}) = \int_{\mathbb{R}^q} L(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y}) d\boldsymbol{u}$$

where q is the number of random effects, and $\boldsymbol{\theta}$ contains all parameters to be estimated.

The Laplace approximation

- We need to calculate the difficult integral

$$L_M(\boldsymbol{\theta}, \boldsymbol{y}) = \int_{\mathbb{R}^q} L(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) d\boldsymbol{u}$$

- So we set up an approximation of $\ell(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) = \log L(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})$

$$\ell(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) \approx \ell(\boldsymbol{\theta}, \hat{\boldsymbol{u}}_{\boldsymbol{\theta}}, \boldsymbol{y}) - \frac{1}{2}(\boldsymbol{u} - \hat{\boldsymbol{u}}_{\boldsymbol{\theta}})^t \left(-\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}}\right) (\boldsymbol{u} - \hat{\boldsymbol{u}}_{\boldsymbol{\theta}})$$

- Which (for given $\boldsymbol{\theta}$) is the 2. order Taylor approximation around:

$$\hat{\boldsymbol{u}}_{\boldsymbol{\theta}} = \underset{\boldsymbol{u}}{\text{argmax}}\ L(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})$$

- With this approximation we can calculate:

$$
\begin{aligned}
L_M(\boldsymbol{\theta}, \boldsymbol{y}) &= \int_{\mathbb{R}^q} L(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) d\boldsymbol{u} \\
&\approx \int_{\mathbb{R}^q} e^{\ell(\boldsymbol{\theta}, \hat{\boldsymbol{u}}_{\boldsymbol{\theta}}, \boldsymbol{y}) - \frac{1}{2}(\boldsymbol{u}-\hat{\boldsymbol{u}}_{\boldsymbol{\theta}})^t \left(-\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}}\right)(\boldsymbol{u}-\hat{\boldsymbol{u}}_{\boldsymbol{\theta}})} d\boldsymbol{u} \\
&= L(\boldsymbol{\theta}, \hat{\boldsymbol{u}}_{\boldsymbol{\theta}}, \boldsymbol{y}) \int_{\mathbb{R}^q} e^{-\frac{1}{2}(\boldsymbol{u}-\hat{\boldsymbol{u}}_{\boldsymbol{\theta}})^t \left(-\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}}\right)(\boldsymbol{u}-\hat{\boldsymbol{u}}_{\boldsymbol{\theta}})} d\boldsymbol{u} \\
&= L(\boldsymbol{\theta}, \hat{\boldsymbol{u}}_{\boldsymbol{\theta}}, \boldsymbol{y}) \sqrt{\frac{(2\pi)^q}{|\left(-\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}}\right)|}}
\end{aligned}
$$

- In the last step we remember the normalizing constant for a multivariate normal, and that $|A^{-1}| = 1/|A|$.
- Taking the logarithm we get:

$$
\ell_M(\boldsymbol{\theta}, \boldsymbol{y}) \approx \ell(\boldsymbol{\theta}, \hat{\boldsymbol{u}}_{\boldsymbol{\theta}}, \boldsymbol{y}) - \frac{1}{2}\log(|\left(-\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}}\right)|) + \frac{q}{2}\log(2\pi)
$$

The Laplace approximation

- The Laplace likelihood only approximates the marginal likelihood for mixed models with nonlinear random effects and thus maximizing the Laplace likelihood will result in some amount of error in the resulting estimates.
- It can be shown that joint log-likelihood converges to a quadratic function of the random effect for increasing number of observations per random effect and thus that the Laplace approximation is asymptotically exact.
- In practical applications the accuracy of the Laplace approximation may still be of concern, but often improved numerical approximation of the marginal likelihood (such as Gaussian quadrature) may easily be computationally infeasible to perform.
- Another option for improving the accuracy is Importance sampling.

Two-level hierarchical model

- For the two-level or hierarchical model it is readily seen that the joint log-likelihood is

$$\ell(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) = \ell(\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{u}, \boldsymbol{y}) = \log f_{Y|u}(\boldsymbol{y}; \boldsymbol{u}, \boldsymbol{\beta}) + \log f_U(\boldsymbol{u}; \boldsymbol{\Psi})$$

  which implies that the Laplace approximation becomes

$$\ell_{M,LA}(\boldsymbol{\theta}, \boldsymbol{y}) = \log f_{Y|u}(\boldsymbol{y}; \tilde{\boldsymbol{u}}, \boldsymbol{\beta}) + \log f_U(\tilde{\boldsymbol{u}}; \boldsymbol{\Psi}) - \frac{1}{2} \log \left| \frac{\boldsymbol{H}(\tilde{\boldsymbol{u}})}{2\pi} \right|$$

  where $\boldsymbol{H}(\tilde{\boldsymbol{u}}) = -\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}}$.
- It is clear that as long as a likelihood function of the random effects and model parameters can be defined it is possible to use the Laplace likelihood for estimation in a mixed model framework.

Gaussian second stage model

- Let us assume that the second stage model is zero mean Gaussian, i.e.

$$\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{\Psi})$$

  which means that the random effect distribution is completely described by its covariance matrix $\boldsymbol{\Psi}$.
- In this case the Laplace likelihood in becomes

$$\begin{aligned}\ell_{M,LA}(\boldsymbol{\theta}, \boldsymbol{y}) = \log f_{Y|u}(\boldsymbol{y}; \tilde{\boldsymbol{u}}, \boldsymbol{\beta}) - \frac{1}{2}\log|\boldsymbol{\Psi}| \\ - \frac{1}{2}\tilde{\boldsymbol{u}}^T\boldsymbol{\Psi}^{-1}\tilde{\boldsymbol{u}} - \frac{1}{2}\log|\boldsymbol{H}(\tilde{\boldsymbol{u}})|\end{aligned}$$

  where it is seen that we still have no assumptions on the first stage model $f_{Y|u}(\boldsymbol{y}; \boldsymbol{u}, \boldsymbol{\beta})$.

Gaussian second stage model

- If we furthermore assume that the first stage model is Gaussian

$$\boldsymbol{Y} \,|\, \boldsymbol{U} = \boldsymbol{u} \sim N(\boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{u}), \boldsymbol{\Sigma})$$

  then the Laplace likelihood can be further specified.
- For the hierarchical Gaussian model it is rather easy to obtain a numerical approximation of the Hessian $\boldsymbol{H}$ at the optimum, $\tilde{\boldsymbol{u}}$

$$\boldsymbol{H}(\tilde{\boldsymbol{u}}) \approx \boldsymbol{\mu}'_u \boldsymbol{\Sigma}^{-1} {\boldsymbol{\mu}'_u}^T + \boldsymbol{\Psi}^{-1}$$

  where $\boldsymbol{\mu}'_u$ is the partial derivative with respect to $\boldsymbol{u}$.
- The approximation in is called Gauss-Newton approximation
- In some contexts estimation using this approximation is also called the First Order Conditional Estimation (FOCE) method.

Example: Orange tree

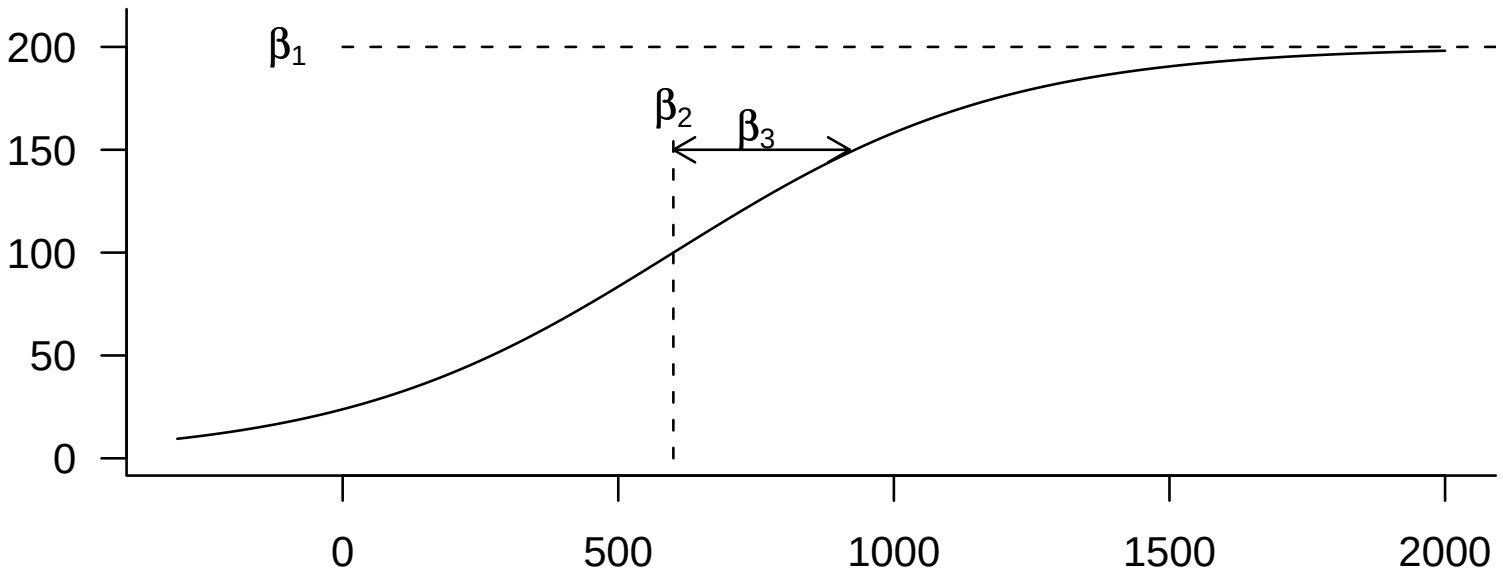
```
> library(datasets)
> data(Orange)
> head(Orange)
```

```
  Tree  age circumference
1    1  118            30
2    1  484            58
3    1  664            87
4    1 1004           115
5    1 1231           120
6    1 1372           142
```

Simple Orange data model

$$y_{ij} = \frac{\beta_1}{1 + \exp[-(t_{ij} - \beta_2)/\beta_3]} + \epsilon_{ij} \ , \quad i = 1...5, \quad j = 1...7,$$

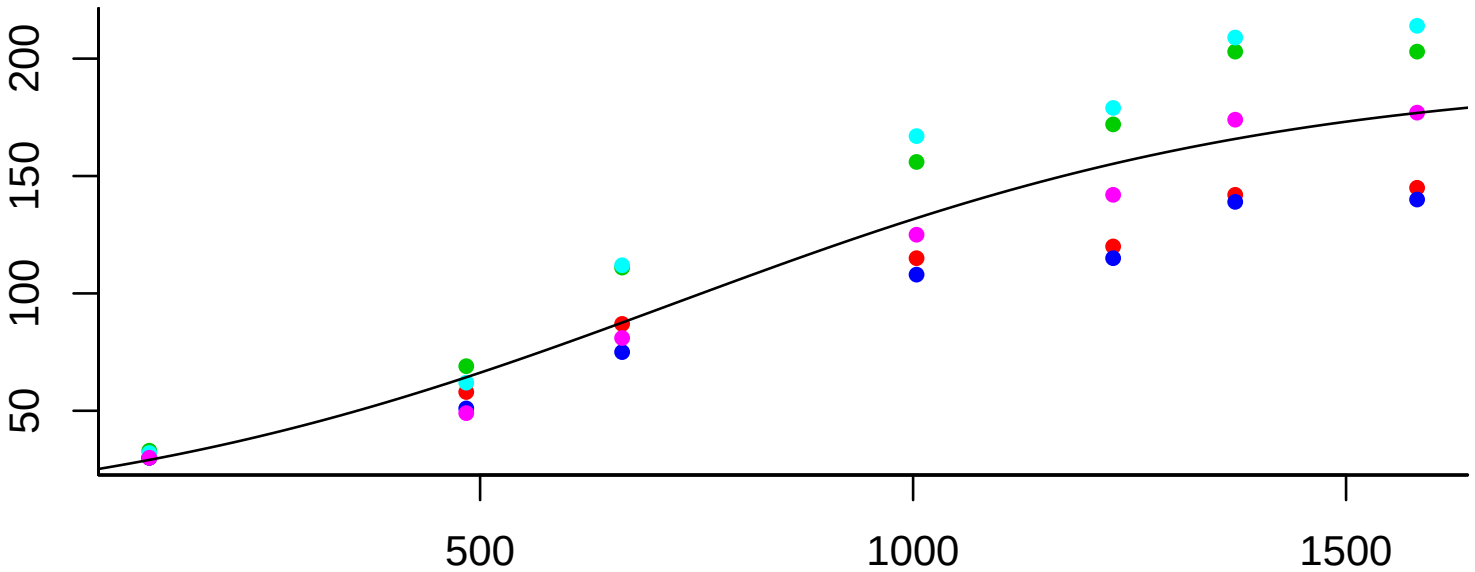
$$\epsilon_{ij} \sim N(0, \sigma^2)$$



Fit of simple Orange data model

Fit using `nls()`

```
      Estimate Std. Error t value Pr(>|t|)
Asym    192.69      20.24   9.518 7.48e-11 ***
xmid    728.75     107.30   6.792 1.12e-07 ***
scal    353.53      81.47   4.339 0.000134 ***
Residual standard error:  23.37 on 32 degrees of freedom
'log Lik.' -158.3987 (df=4)
```



Example: Orange tree

- First we wish to run the model:

$$\text{cir}_i = \frac{\beta_1 + U(\text{tree}_i)}{1 + \exp(-(\text{age}_i - \beta_2)/\beta_3)} + \epsilon_i$$

- where $\epsilon_i \sim N(0, \sigma^2)$ and $U(\text{tree}_i) \sim N(0, \sigma_U^2)$ independent.
- Let's first set up a function to calculate the mean for a given observation.

```
> mu <- function(beta, u, tree, age) {
+     (beta[1] + u[tree])/(1 + exp(-(age - beta[2])/beta[3]))
+ }
```

Example: Orange tree

- The joint log likelihood is:

$$\ell(\text{cir}, u, \theta) = \ell_{\text{cir}|u}(\text{cir}, u, \beta, \sigma) + \ell_U(u, \sigma_U)$$

- The part $\ell_U(u, \sigma_U)$ is implemented by:

```
> l.u <- function(u, s.u) {
+     sum(dnorm(u, mean = 0, sd = s.u, log = TRUE))
+ }
```

- The part $\ell_{\text{cir}|u}(\text{cir}, u, \beta, \sigma)$ is implemented by:

```
> l.cir <- function(cir, u, b, s) {
+     mv <- mu(b, u, Orange$Tree, Orange$age)
+     sum(dnorm(cir, mean = mv, sd = s, log = TRUE))
+ }
```

- And the joint negative log likelihood $-\ell(\text{cir}, u, \beta)$ is:

```
> nl <- function(th, u, cir) {
+     -l.cir(cir, u, th[1:3], exp(th[4])) - l.u(u, exp(th[5]))
+ }
```

Example: Orange tree

- Now we can set up the Laplace approximation

```
> library(numDeriv)
> l.LA <- function(th) {
+     u.init <- rep(0, nlevels(Orange$Tree))
+     obj <- function(u) nl(th, u, Orange$cir)
+     est <- nlminb(u.init, obj)
+     lval <- est$obj
+     u <- est$par
+     H <- hessian(obj, u)
+     lval + 0.5 * log(det(H)) - length(u)/2 * log(2 * pi)
+ }
```

- And optimize w.r.t. the model parameters in θ

```
> fit <- nlminb(c(300, 700, 200, 0, 0), l.LA)
> H <- hessian(l.LA, fit$par)
```

Example: Orange tree

```
> fit

$par
[1] 192.053210 727.906549 348.073223   2.059623   3.454621

$objective
[1] 131.5719

$convergence
[1] 0

$iterations
[1] 54

$evaluations
function gradient
      60      307

$message
[1] "relative convergence (4)"

> cbind(est = fit$par, sd = sqrt(diag(solve(H))))

             est         sd
[1,] 192.053210 15.6576723
[2,] 727.906549 35.2486970
[3,] 348.073223 27.0798192
[4,]   2.059623  0.1290996
[5,]   3.454621  0.3242544
```

Laplace approximation work flow

0. Initialize $\boldsymbol{\theta}$ to some arbitrary value $\boldsymbol{\theta}_0$
1. With current value for $\boldsymbol{\theta}$ optimize joint likelihood w.r.t. $\boldsymbol{u}$ to get $\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}$ and corresponding Hessian $H(\hat{\boldsymbol{u}}_{\boldsymbol{\theta}})$.
2. Use $\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}$ and $H(\hat{\boldsymbol{u}}_{\boldsymbol{\theta}})$ to approximate $\ell_M(\boldsymbol{\theta})$
3. Compute value and gradient of $\ell_M(\boldsymbol{\theta})$
4. If the gradient is "$>\epsilon$" set $\boldsymbol{\theta}$ to a different value and go to 1.

Notice the huge number of — possibly high dimensional — optimizations that are required.

Other models..

- (5.111) (plus random component for sampling occasion (j))

$$y_{ij} = \frac{\beta_1 + u_{1i} + u_{2j}}{1 + \exp[-(t_j - \beta_2)/\beta_3]} + \epsilon_{ij} \tag{19}$$

- (5.112) (seasonal variation):

$$y_{ij} = \frac{\beta_1 + u_{1i}}{1 + \exp[-((t_j - \beta_2)/\beta_3 + s_j\beta_4)]} + \epsilon_{ij} \tag{20}$$

- (5.113) (First order AR)

$$\mathrm{cov}(\epsilon_{ij}, \epsilon_{ij'}) = \sigma^2 \exp(-\phi|t_{j'} - t_j|/(365/2)), \quad \phi \geq 0 \tag{21}$$

ie. the full covariance matrix is block diagonal.

Table: Parameter estimates (and standard errors) and log-likelihoods for models estimated for the orange tree data.

Model	β_1	β_2	β_3	β_4	σ	σ_{u1}	σ_{u2}	ρ	$\log(L)$
(??)	192.1	727.9	348.1		7.84	31.6			-131.57
	(15.7)	(35.3)	(27.1)						
(19)	196.2	748.4	352.9		5.30	32.6	10.5		-125.45
	(19.4)	(62.3)	(33.3)						
(20)	217.1	857.5	436.8	0.322	4.79	36.0			-116.79
	(18.1)	(42.0)	(24.5)	(0.038)					
(19) + (21)	192.4	730.1	348.1		6.12	32.7	12.0	0.773	-118.44
	(19.6)	(63.8)	(34.2)						
(20) + (21)	216.2	859.1	437.8	0.330	5.76	36.7		0.811	-106.18
	(17.6)	(30.5)	(21.6)	(0.022)					

Beetles exposed to ethylene oxide

Ten groups beetles were exposed to different concentrations of ethylene oxide and it was recorded how many died.

```
> conc <- c(24.8, 24.6, 23, 21, 20.6, 18.2, 16.8, 15.8, 14.7, 10.8)
> n <- c(30, 30, 31, 30, 26, 27, 31, 30, 31, 24)
> y <- c(23, 30, 29, 22, 23, 7, 12, 17, 10, 0)
```

The natural model is a binomial, and we wish to setup a logit-linear model as a function of the logarithm of the concentrations

$$y_i \sim \text{Bin}(n_i, p_i) \ , \ \text{where}$$
$$\text{logit}(p_i) = \mu + \beta \log(\text{conc}_i)$$

```
> resp <- cbind(y, n - y)
> fit <- glm(resp ~ I(log(conc)), family = binomial())
```

Beetles exposed to ethylene oxide

We get:

```
> fit

Call:  glm(formula = resp ~ I(log(conc)), family = binomial())

Coefficients:
 (Intercept)  I(log(conc))
     -17.867         6.265

Degrees of Freedom: 9 Total (i.e. Null);  8 Residual
Null Deviance:      138
Residual Deviance: 36.44  AIC: 68.02

> 1 - pchisq(fit$deviance, fit$df.residual)

[1] 1.456223e-05
```

Grouping structures and nested effects

- For nonlinear mixed models where no closed form solution to the likelihood function is available it is necessary to invoke some form of numerical approximation to be able to estimate the model parameters.
- The complexity of this problem is mainly dependent on the dimensionality of the integration problem which in turn is dependent on the dimension of $\boldsymbol{U}$ and in particular the *grouping structure* in the data for the random effects.
- These structures include a *single grouping*, *nested grouping*, *partially crossed* and *crossed random effects*.
- For problems with only one level of grouping the marginal likelihood can be simplified as

$$L_M(\boldsymbol{\beta}, \boldsymbol{\Psi}; \boldsymbol{y}) = \prod_{i=1}^{M} \int_{\mathbb{R}^{q_i}} f_{Y|u_i}(\boldsymbol{y}; \boldsymbol{u}_i, \boldsymbol{\beta}) f_{U_i}(\boldsymbol{u_i}; \boldsymbol{\Psi})\, d\boldsymbol{u}_i$$

where q_i is the number of random effects for group i and M is the number of groups.

Grouping structures and nested effects

- Instead of having to solve an integral of dimension q it is only necessary to solve M smaller integrals of dimension q_i.
- In typical applications there is often just one or only a few random effects for each group, and this thus greatly reduces the complexity of the integration problem.
- If the data has a nested grouping structure a reduction of the dimensionality of the integral similar to that shown on the previous slide can be performed.
- An example of a nested grouping structure is data collected from a number of schools, a number of classes within each school and a number of students from each class.

Grouping structures and nested effects

- If the nonlinear mixed model is extended to include any structure of random effects such as crossed or partially crossed random effects it is required to evaluate the full multi-dimensional integral
- Estimation in these models can efficiently be handled using the multivariate Laplace approximation, which only samples the integrand in one point common to all dimensions.

Importance sampling

- Importance sampling is a re-weighting technique for approximating integrals w.r.t. a density f by simulation in cases where it is not feasible to simulate from the distribution with density f.
- Instead it uses samples from a different distribution with density g, where the support of g includes the support of f.
- For general mixed effects models it is possible to simulate from the distribution with density proportional to the second order Taylor approximation

$$\widetilde{L}(\boldsymbol{\theta}, \hat{\boldsymbol{u}}_\theta, \boldsymbol{Y}) = e^{\ell(\boldsymbol{\theta}, \hat{\boldsymbol{u}}_\theta, \boldsymbol{Y}) - \frac{1}{2}(\boldsymbol{u}-\hat{\boldsymbol{u}}_\theta)^T(-\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{Y})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_\theta})(\boldsymbol{u}-\hat{\boldsymbol{u}}_\theta)}$$

which, apart from a normalization constant, it is the density $\phi_{\hat{u}_\theta, \hat{V}_\theta}(\boldsymbol{u})$ of a multivariate normal with mean $\hat{\boldsymbol{u}}_\theta$ and covariance

$$\hat{\boldsymbol{V}}_\theta = \boldsymbol{H}^{-1}(\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}) = (-\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{Y})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_\theta})^{-1}.$$

Importance sampling

The integral to be approximated can be rewritten as:

$$L_M(\boldsymbol{\theta}, \boldsymbol{Y}) = \int L(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{Y}) d\boldsymbol{u} = \int \frac{L(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{Y})}{\phi_{\hat{u}_\theta, \hat{V}_\theta}(u)} \phi_{\hat{u}_\theta, \hat{V}_\theta}(\boldsymbol{u}) d\boldsymbol{u}.$$

So if $\boldsymbol{u}^{(i)}$, $i = 1, \ldots, N$ is simulated from the multivariate normal distribution with mean $\hat{\boldsymbol{u}}_\theta$ and covariance $\hat{\boldsymbol{V}}_\theta$, then the integral can be approximated by the mean of the importance weights

$$L_M(\boldsymbol{\theta}, \boldsymbol{Y}) = \frac{1}{N} \sum \frac{L(\boldsymbol{\theta}, \boldsymbol{u}^{(i)}, \boldsymbol{Y})}{\phi_{\hat{u}_\theta, \hat{V}_\theta}(\boldsymbol{u}^{(i)})}$$

Two-level hierarchical model

- For the two-level or hierarchical model it is readily seen that the joint log-likelihood is

$$\ell(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}) = \ell(\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{u}, \boldsymbol{y}) = \log f_{Y|u}(\boldsymbol{y}; \boldsymbol{u}, \boldsymbol{\beta}) + \log f_U(\boldsymbol{u}; \boldsymbol{\Psi})$$

  which implies that the Laplace approximation becomes

$$\ell_{M,LA}(\boldsymbol{\theta}, \boldsymbol{y}) = \log f_{Y|u}(\boldsymbol{y}; \tilde{\boldsymbol{u}}, \boldsymbol{\beta}) + \log f_U(\tilde{\boldsymbol{u}}; \boldsymbol{\Psi}) - \frac{1}{2} \log \left| \frac{\boldsymbol{H}(\tilde{\boldsymbol{u}})}{2\pi} \right|$$

- It is clear that as long as a likelihood function of the random effects and model parameters can be defined it is possible to use the Laplace likelihood for estimation in a mixed model framework.

Gaussian second stage model

- Let us assume that the second stage model is zero mean Gaussian, i.e.

$$\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{\Psi})$$

  which means that the random effect distribution is completely described by its covariance matrix $\boldsymbol{\Psi}$.
- In this case the Laplace likelihood in becomes

$$\begin{aligned}\ell_{M,LA}(\boldsymbol{\theta}, \boldsymbol{y}) = \log f_{Y|u}(\boldsymbol{y}; \tilde{\boldsymbol{u}}, \boldsymbol{\beta}) - \frac{1}{2}\log|\boldsymbol{\Psi}| \\ - \frac{1}{2}\tilde{\boldsymbol{u}}^T\boldsymbol{\Psi}^{-1}\tilde{\boldsymbol{u}} - \frac{1}{2}\log|\boldsymbol{H}(\tilde{\boldsymbol{u}})|\end{aligned}$$

  where it is seen that we still have no assumptions on the first stage model $f_{Y|u}(\boldsymbol{y}; \boldsymbol{u}, \boldsymbol{\beta})$.

Gaussian second stage model

- If we furthermore assume that the first stage model is Gaussian

$$\boldsymbol{Y} \,|\, \boldsymbol{U} = \boldsymbol{u} \sim N(\boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{u}), \boldsymbol{\Sigma})$$

  then the Laplace likelihood can be further specified.
- For the hierarchical Gaussian model it is rather easy to obtain a numerical approximation of the Hessian $\boldsymbol{H}$ at the optimum, $\tilde{\boldsymbol{u}}$

$$\boldsymbol{H}(\tilde{\boldsymbol{u}}) \approx \boldsymbol{\mu}'_u \boldsymbol{\Sigma}^{-1} {\boldsymbol{\mu}'_u}^T + \boldsymbol{\Psi}^{-1}$$

  where $\boldsymbol{\mu}'_u$ is the partial derivative with respect to $\boldsymbol{u}$.
- The approximation in is called Gauss-Newton approximation
- In some contexts estimation using this approximation is also called the First Order Conditional Estimation (FOCE) method.