The differential equation (‡) is of the same form as before (see (*)), but with input function $\mathbf{A}_{(j)}\boldsymbol{\gamma}(\tau) + \mathbf{i}_{(j)}$. Therefore, (**) gives us the form of the solution, we just need to change the input function in (**) and change $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma}_{(j)}$. This solution is

$$\boldsymbol{\gamma}_{(j)}(\tau) = e^{\mathbf{A}\tau}\boldsymbol{\gamma}_{(j)}(0) + e^{\mathbf{A}\tau} * [\mathbf{A}_{(j)}\boldsymbol{\gamma}(\tau) + \mathbf{i}_{(j)}]$$

Cases 1 & 2 combined: putting these two cases together, we can obtain an expression for $\boldsymbol{\gamma}_{(j)}$ no matter which of $\boldsymbol{\gamma}_0$, $\mathbf{A}$, τ, and $\mathbf{i}$ depend on θ_j:

$$\boldsymbol{\gamma}_{(j)}(\tau) = e^{\mathbf{A}\tau}\boldsymbol{\gamma}_{(j)}(0) + e^{\mathbf{A}\tau} * [\mathbf{A}_{(j)}\boldsymbol{\gamma}(\tau) + \mathbf{i}_{(j)}] + \tau_{(j)}[\mathbf{A}\boldsymbol{\gamma}(\tau) + \mathbf{i}(\tau)]$$

- This expression holds for any input function of a bolus (impulse) or continuous infusion/step type.
- Bates and Watts provide details on efficiently computing the convolutions in this expression in Appendix A5 and pseudo-code for doing so in Appendix A3.
- I have programmed this pseudo-code in the R functions formcompmodel() and compmodel(). These functions can be found in the file compmodel.R which can be obtained from the course web site.

Example — Tetracycline

Recall that the transfer matrix in the two compartment model that we considered previously for these data is

$$\mathbf{A} = \begin{pmatrix} -\theta_1 & 0 \\ \theta_1 & -\theta_2 \end{pmatrix}.$$

This is a simple model with a bolus input function, so obtaining the analytic solution for $\boldsymbol{\gamma}(t) = (\gamma_1(t), \gamma_2(t))^T$ is particularly easy.

Recall (top of p.189) that the solution is given by

$$\boldsymbol{\gamma}(t) = \mathbf{U} e^{\Lambda t} \mathbf{U}^{-1} \boldsymbol{\gamma}_0. \qquad (*)$$

The eigenvalues and eigenvectors of $\mathbf{A}$ aren't difficult to calculate here, especially with a symbolic math program like Maple or Mathematica (see tetra1.mws, a Maple worksheet). These calculations lead to

$$\Lambda = \begin{pmatrix} -\theta_2 & 0 \\ 0 & -\theta_1 \end{pmatrix}, \quad \text{and} \quad \mathbf{U} = \begin{pmatrix} 0 & \frac{\theta_2 - \theta_1}{\theta_1} \\ 1 & 1 \end{pmatrix}.$$

We also need $\mathbf{U}^{-1}$:

$$\mathbf{U}^{-1} = \begin{pmatrix} \frac{\theta_1}{\theta_1 - \theta_2} & 1 \\ -\frac{\theta_1}{\theta_1 - \theta_2} & 0 \end{pmatrix}.$$

Plugging into (*), we have

$$\begin{aligned} \boldsymbol{\gamma}(t) &= \begin{pmatrix} 0 & \frac{\theta_2 - \theta_1}{\theta_1} \\ 1 & 1 \end{pmatrix} \begin{pmatrix} e^{-\theta_2 t} & 0 \\ 0 & e^{-\theta_1 t} \end{pmatrix} \begin{pmatrix} \frac{\theta_1}{\theta_1 - \theta_2} & 1 \\ -\frac{\theta_1}{\theta_1 - \theta_2} & 0 \end{pmatrix} \begin{pmatrix} \theta_3 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \theta_3 e^{-\theta_1 t} \\ \frac{\theta_3 \theta_1 (e^{-\theta_1 t} - e^{-\theta_2 t})}{\theta_2 - \theta_1} \end{pmatrix} \end{aligned}$$

- See handout tetra1. Here we fit this two-compartment model in two ways: using the analytic solution $\gamma_2(t)$ on the previous page as the expectation function of the model; and using the matrix exponential approach in which we need not obtain the analytic solution. We also fit the corresponding model with dead time using the latter method.

- In both cases we parameterize the rate constants using an exponential transformation to ensure that the rate constants are positive. That is, use ϕ's to represent the transfer rates rather than θ's. With this notation change the transfer matrix is

$$\mathbf{A} = \begin{pmatrix} -\phi_1 & 0 \\ \phi_1 & -\phi_2 \end{pmatrix}$$

  and we use an unconstrained θ-parameterization where $\phi_i = e^{\theta_i}$, $i = 1, \ldots, 3$. Our model becomes

$$y_i = \frac{e^{\theta_3+\theta_1}(\exp\{-e^{\theta_1}t_i\} - \exp\{-e^{\theta_2}t_i\})}{e^{\theta_2} - e^{\theta_1}} + e_i, \quad i = 1, \ldots, n$$

- In the compmodel() function that implements the matrix exponential approach to computing the expectation function and its derivatives for linear compartment models, the rate constants and initial concentrations are all parameterized as exponential parameters like this. Dead time parameters are not transformed.

- In tetra1.R we code the analytic solution $\gamma_2(t)$ as a function tetramod(). By using the deriv() function, tetramod() will return not only the function value, but the values of its analytic derivatives with respect to the parameters $\theta_1, \theta_2, \theta_3$.

- We will use tetramod() as the expectation function in an nls() fit. But first, we demonstrate that the matrix exponential function compmodel() returns the same value and gradient as tetramod().

- The function formcompmodel() sets up the model from an input matrix **J**. formcompmodel() is always called as the first step in compmodel(), but can also be called on its own. The model is "set up" from a matrix **J** that describes the compartment model. Specifically, it consists of several rows, one for each parameter in the model and/or arrow in the compartment diagram. In addition, **J** has three columns:

  1. the parameter number
  2. the source compartment (0 if the parameter is a dead time parameter)
  3. the destination compartment (0 if the parameter is a dead time parameter or if the destination is excretion; -1 if the parameter is an initial value).

- For example, our 2-compartment tetracyline model is specified as

$$\mathbf{J} = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 2 & 0 \\ 3 & 1 & -1 \end{pmatrix}$$

  Adding a dead time parameter $\theta_4 = t_0$, the model would be specified as

$$\mathbf{J} = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 2 & 0 \\ 3 & 1 & -1 \\ 4 & 0 & 0 \end{pmatrix}$$

- From input **J**, $\boldsymbol{\theta}$, and $\boldsymbol{\gamma}_{0\text{fix}}$, the fixed (known) portion of the initial conditions vector, formcompmodel($\mathbf{J}, \boldsymbol{\theta}, \boldsymbol{\gamma}_{0\text{fix}}$) returns **A**, $\boldsymbol{\gamma}_0$, t_0, $\frac{\partial \mathbf{A}}{\partial \boldsymbol{\theta}}$, $\frac{\partial \boldsymbol{\gamma}_0}{\partial \boldsymbol{\theta}^T}$, and $\frac{\partial t_0}{\partial \boldsymbol{\theta}}$.

- compmodel() takes arguments **J**; $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$; $\boldsymbol{\gamma}_{0\text{fix}}$, a $K \times 1$ vector of initial values for the K compartments (0 should be used for any compartment whose initial value is a parameter to be estimated); **t** a time vector; and k, the compartment from which data are assumed to come and for which a solution is sought.

- compmodel() first calls formcompmodel() to set up the model and then computes and returns the model solution for compartment k. A gradient attribute is also returned containing $\frac{\partial \gamma_k(\tau)}{\partial \boldsymbol{\theta}^T}$ for each value of the time indexing variable τ.

- To demonstrate that compmodel() returns the correct model solution and gradient, we evaluate the model with tetramod() (assigned to analyticeval) and compmodel() (assigned to matrixexptialeval) and print out a portion of both results to verify that they are the same.

- Notice that both the values and gradients agree using the analytic solution given by tetramod() or the matrix exponential evaluation given by compmodel().

- We fit the model with tetramod() first as m1tetra.nls. Then we refit using compmodel() as m2tetra.nls. Starting values can be obtained by exponential peeling. Here, we omit that step and just take $\hat{\boldsymbol{\theta}}^0 = (\log(.2), \log(.4), \log(6))^T$. In each case, the nls() function converges to $\hat{\boldsymbol{\theta}} = (-1.698, -.834, 1.791)^T$, or in the ϕ-parameterization, $\hat{\boldsymbol{\phi}} = \exp(\hat{\boldsymbol{\theta}}) = (.183, .434, 5.994)^T$.

- We refit the model allowing for dead time by changing $\mathbf{J}$ and using compmodel() again. This model is m3tetra.nls. An extra sum of squares test (LRT) of model m2tetra.nls vs. m3tetra.nls tests the necessity of dead time (tests $H_0 : \theta_4 = 0$). We reject this hypothesis and prefer the model with dead time. Residuals from this model look reasonable.

Practical Considerations:

Parameter Transformations:

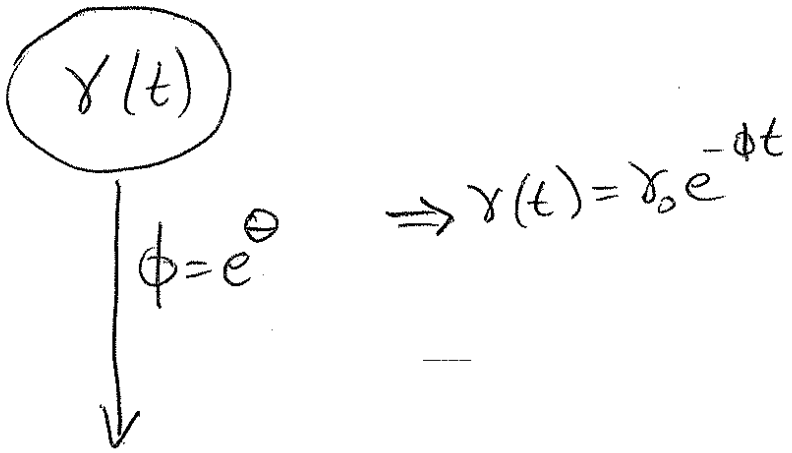
- Rate constants, initial concentrations, and infusion rates must be positive.

A convenient way of ensuring this positivity is with a parameterization in which the constrained rate constant (say) ϕ is parameterized as the exponential of an unconstrained parameter θ:

$$\phi = e^{\theta} \qquad \Rightarrow \qquad \theta = \log(\phi)$$

- This parameter transformation not only imposes the desired constraint, but θ, the log rate constant, has a convenient relationship with $t_{1/2}$, the *half-life* associated with the exchange of material that has rate ϕ.

A single compartment elimination model looks like this



Such a model satisfies the differential equation $\dot{\gamma}(t) = -\phi\gamma(t)$ which has solution $\gamma(t) = e^{-\phi t}\gamma_0$. The half-life $t_{1/2}$ is the time at which half of the initial concentration γ_0 has been eliminated from the compartment. That is $t_{1/2}$ satisfies

$$\frac{\gamma_0 e^{-\phi t_{1/2}}}{\gamma_0} = \frac{1}{2}$$
$$\text{or} \qquad e^{\phi t_{1/2}} = 2$$
$$\text{or} \qquad t_{1/2} = \frac{\log(2)}{\phi}$$

Thus,

$$\log(t_{1/2}) = \log[\log(2)] - \log(\phi) = -.367 - \theta$$

so that the width of a linear approximation interval for $\log(t_{1/2})$ is the same as the width of the interval for $\theta = \log(\phi)$.

Another derived quantity of interest is the *volume of distribution* in a compartment. With a bolus injection, the dose D is known, but the concentration γ_0 is estimated because the volume of the compartment in which that dose is distributed V is unknown. The relationship between initial concentration γ_0, dose D, and initial volume of distribution V is given by

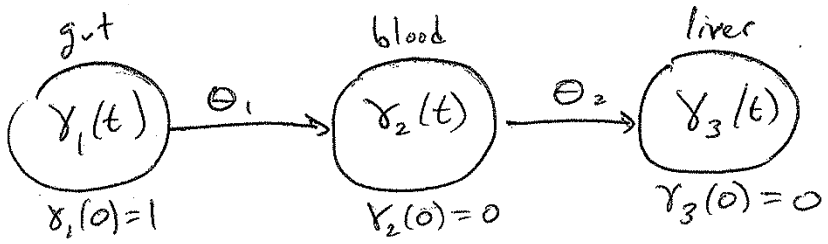
$$\gamma_0 = \frac{D}{V} \quad \Rightarrow \log(\gamma_0) = \log(D) - \log(V)$$

so again, a linear approximation CI on $\log(V)$ will have the same width as that of $\log(\gamma_0)$.

Identifiability.

We have already seen that linear combinations of exponentials can have identifiability problems (recall the biexponential model).

For example, the following three compartment model with initial conditions $\boldsymbol{\gamma}_0 = (1, 0, 0)^T$ and data collected in compartment 3, yields the same $\gamma_3(t)$ curve for the parameter pair $\boldsymbol{\theta} = (a, b)$ as for $\boldsymbol{\theta} = (b, a)$.



- Here (a, b) and (b, a) are exchangeable. In such a case the model is only *locally identifiable* because discrete sets of parameters give the same predicted response.

A more serious situation is *global unidentifiability*, where continuous sets of parameters give the same predictions.

- Identifibaility in compartment models is a big topic with lots of research. We don't attempt to give general results on when a compartment model is identifiable.

- However, a simple way to check identifiability is as follows: fix a set of design times and generate the parameter derivative matrix $\mathbf{V}(\boldsymbol{\theta})$ at a number of different choices of $\boldsymbol{\theta}$. If the matrix $\{\mathbf{V}(\boldsymbol{\theta})\}^T\mathbf{V}(\boldsymbol{\theta})$ is computationslly singular for all choices of $\boldsymbol{\theta}$, then the model can be assumed to be unidentifiable. Note that one should check several choices of $\boldsymbol{\theta}$ because one can get unlucky with a particular choice without the model being unidentifiable.

- Nonidentifiability is a especially problematic in *univariate* compartmental systems in which only one compartment is measured. Multiresponse experiments allow one to fit much richer and more complex compartment models.

Starting Values.

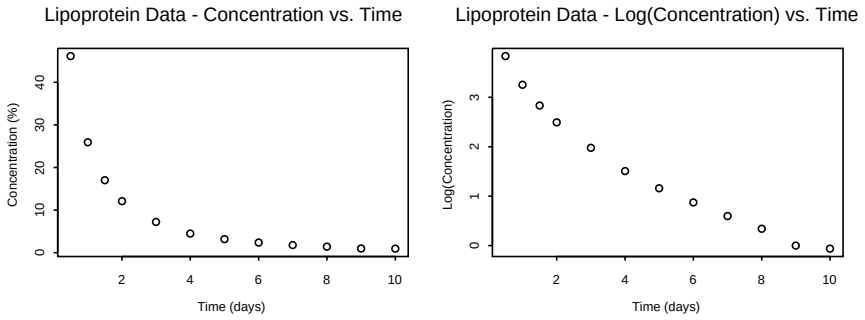
- The method of exponential peeling is widely used to obtain starting values for compartment models.

- A second approach is to build up an appropriate model from a very simple one. At each stage only a small number of parameters (often just one) are added, so that starting values can be easily obtained.

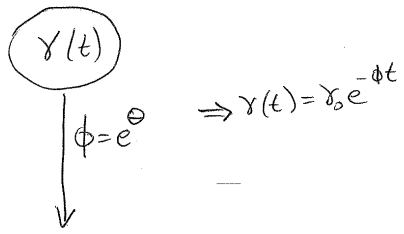
Example — Lipoproteins

- See handout lipo1. Here we fit the models in §5.4 of our text to the lipoprotein data of Appendix 1, §A1.16. The response variable is the percentage concentration of a tracer in the serum of a baboon given a bolus injection at time 0. We assume that the initial concentration is 100% in compartment one and 0 in all other compartments.

- Before fitting any models, we plot the data on both the original and log scale. These plots appear below. From the log-scale plot, it is apparent that at least two compartments will be necessary since this plot is not linear. However, we start by fitting a one compartment model and build up from there.



- We first fit a one compartment elimination model:



- The data near time $t = 0$ may approximately satisfy such a model. According to this model $\gamma_1(t) = \gamma_0 e^{-\phi t} = 100e^{-\phi t}$ or

$$y = 100e^{-\phi t} \quad \Rightarrow \quad \log(y) = \log(100) - \phi t$$
$$\Rightarrow \quad \phi = \frac{\log(100) - \log(y)}{t} = \frac{-\log(y/100)}{t}$$

  At time $t = .5$ we have $y = 46$. Plugging in these data we have

$$\phi = \frac{-\log(.46)}{.5} = 1.55$$

  so we take $\hat{\theta}^0 = \log(1.55)$ as our initial value for θ.

- In lipo1.R we fit this simple model using the analytic solution $100\exp(-e^{\theta}t)$ as the expectation function. This model is assigned to m1Lipo.nls and yields $\hat{\phi} = \exp(\hat{\theta}) = 1.31$ and a residual standard error of 3.48 on 11 residual df.

- The residuals from this model (not shown in lipo1) indicate substantial lack of fit. This was expected and we proceed to a 2 compartment model of the following form:

- Initially, we assume $\phi_2 = \phi_3$ so that we introduce only one additional parameter. The elimination from compartment 1 is now $\phi_1 + \phi_2$ so we set $1.31 = \phi_1 + \phi_2$ and try (somewhat arbitrarily) the initial value $\hat{\boldsymbol{\phi}}^0 = (1.00, 0.31)^T$ or $\hat{\boldsymbol{\theta}}^0 = (\log(1.00), \log(.31))^T$.

- This model is fit as m2aLipo.nls. Note that without the constraint $\phi_2 = \phi_3$ the $\mathbf{J}$ matrix is

$$\mathbf{J} = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \end{pmatrix}$$

  To impose the constraint, we just include two rows for ϕ_2:

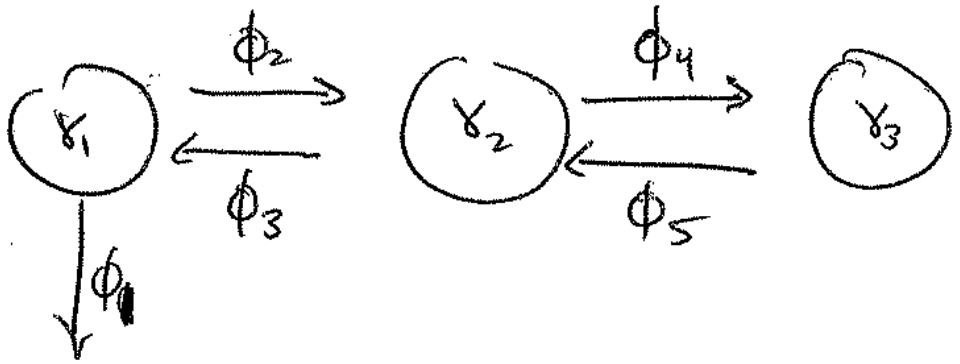
$$\mathbf{J} = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix}$$

  Model m2aLipo.nls yields $\hat{\boldsymbol{\phi}} = (.992, .663)^T$.

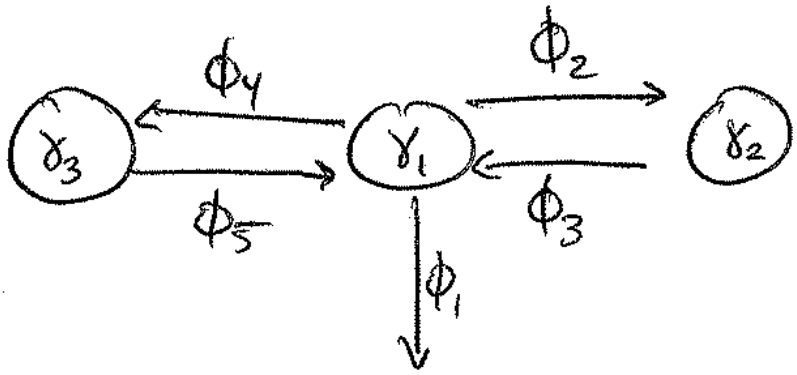
- The model without the constraint $\phi_2 = \phi_3$ is fit as m2bLipo.nls. For starting values we use $\hat{\boldsymbol{\phi}}^0 = (.99, .67, .65)^T$. This leads to $\hat{\boldsymbol{\phi}} = (1.028, .662, .820)^T$ and a residual standard error of .374 on 9 residual df.

- While this 2 compartment model fits substantially better than the one compartment model, the residuals of m2bLipo.nls still clearly indicate that the model is inadequate. To better fit the data we have several choices for extending this model. We choose to try to add a third compartment to the model. This can be done in several ways, but two simple choices are a 3 compartment *catenary* system and a 3 compartment *mamillary* system.

- In the catenary system, compartments are chained together as follows:



- In the mamillary system, there is a central *mother* compartment with which each peripheral compartment exchanges material:



- We will fit both the catenary and mamillary version of the three compartment model. In each case, we could fit fewer than 5 parameters by constraining some rate constants to be equal. Instead, we will fit the full 5-parameter version of these models and then consider reducing the model by equating parameters.

- The catenary model and mamillary models are described by the matrices

$$\mathbf{J} = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \\ 4 & 2 & 3 \\ 5 & 3 & 2 \end{pmatrix}, \qquad \mathbf{J} = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \\ 4 & 1 & 3 \\ 5 & 3 & 1 \end{pmatrix},$$

  respectively. These models are fits as m3acatLipo.nls and m3amamLipo.nls, respectively. For starting values we use $\hat{\boldsymbol{\phi}}^0 = (1.0, .66, .82, .5, .2)^T$ for both models. The first three starting values are taken from the fit of the previous model. Arbitrarily, $\hat{\phi}_4^0$ and $\hat{\phi}_5^0$ are chosen to be smaller than, and distinct from $\hat{\phi}_2^0$ and $\hat{\phi}_3^0$.

- Models m3acatLipo.nls and m3amamLipo.nls converge to $\hat{\boldsymbol{\phi}} = (.990, .763, 1.01, .240, .352)^T$ and $\hat{\boldsymbol{\phi}} = (.990, .531, 1.34, .231, .266)^T$, respectively. Both models have the same residual standard error of .0787 on 7 df. In fact, these two models are equivalent (and not identifiable from one another) when only compartment 1 is measured.

- Finally, we consider reducing these models by constraining some parameters to be equal to one another. In both models parameter estimates for parameters θ_4 and θ_5 are closest to each other relative to their standard errors (as compared with any other pair of $\hat{\theta}_j$'s).

- Therefore, we refit the two three compartment models with $\theta_4 = \theta_5$ in each case. For the constrained models, the **J** matrices become

$$\mathbf{J} = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \\ 4 & 2 & 3 \\ 4 & 3 & 2 \end{pmatrix}, \qquad \mathbf{J} = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \\ 4 & 1 & 3 \\ 4 & 3 & 1 \end{pmatrix},$$

  for m3bcatLipo.nls and m3bmamLipo.nls, respectively. The 4-parameter versions of these models are no longer equivalent, and they yield distinct residual standard errors of .0880 and .0792, respectively.

- Based on extra sum of squares analyses (LRTs) the 4-parameter versions of these models fit as well as the 5-parameter versions, and are therefore preferred based on parsimony. In addition, the residuals from these models look good.

- Based on the residual standard error, the 4-parameter mamillary model fits slightly better than the 4-parameter catenary model. We can also obtain the AIC and BIC values for these model. These criteria also point to model m3bmamLipo.nls. However, the choice of model should be made based on these criteria secondary to biological/pharmacologic considerations.

Growth Models

There are two traditions in the development of models to describe growth.

1. "Statistical" Approach. This is a purely empirical approach in which polynomial models in time (linear models) are fit to the data using multivariate methods.
   - Parameters have no biological interpretation.
   - Models are not necessarily parsimonious.
   - Extrapolation (e.g., prediction of future growth) is always dangerous, but especially so for these models.
   - Models are linear, so methodology, theory are easier.
   - These models are often discussed in multivariate texts. See, for example, Timm (2002, *Applied Multivariate Analysis*) for a good treatment.
2. "Biological" Approach. Models have a mechanistic motivation, although in practice they are often used in a purely empirical way. Models are usually nonlinear, with relatively few, biologically interpretable parameters.

- We concentrate on models in the latter tradition.

Exponential and Monomolecular Models:

The simplest organisms begin to grow by the binary splitting of cells. If we let t denote time and $f(t)$ denote size at time t, then this leads to exponential growth in which the growth rate is proportional to the current size $f(t)$:

$$\frac{\partial f(t)}{\partial t} = \kappa f(t), \qquad \text{or} \qquad f(t) = e^{\kappa(t-\gamma)}. \qquad (*)$$

The *time-power* model

$$f(t) = \alpha t^{\beta},$$

does not increase as fast as the exponential, but is sometimes useful.

- Both of these models imply unlimited growth, which make them unsuitable for many applications (except perhaps as models of early growth).

We can change (*) to imply growth bounded by an upper limit by assuming that the growth rate is proportional to size remaining:

$$\frac{\partial f(t)}{\partial t} = \kappa[\alpha - f(t)] \quad \text{where } \kappa > 0.$$

The solution to this differential equation can be parameterized in a variety of ways including

$$f(t) = \alpha - (\alpha - \beta)e^{-\kappa t}, \quad \alpha > \beta > 0.$$

Here α is the final size (asymptote), β the initial size, and κ dictates the growth rate. Alternative parameterizations are

$$f(t) = \alpha - \beta e^{-\kappa t}$$

and

$$f(t) = \theta_1(1 - e^{-\theta_2(t-\theta_3)}) \qquad \text{(monomolecular growth model)}$$

and

$$f(t) = \phi_1 + \phi_2\phi_3^t \qquad \text{(asymptotic regression model)}$$

Sigmoidal Models:

From the fact that the above model can be reparameterized as the asyptotic regression model, it is clear that this model has an asymptotic form with growth rate decreasing through time.

Since growth rate may increase early in development we may prefer a curve with sigmoidal form.

One way that a sigmoidal curve may be achieved is by assuming that the current growth rate is the product of functions of the current size f and the remaining growth on a transformed scale:

$$\frac{\partial f}{\partial t} \propto g(f)[h(\alpha) - h(f)], \qquad (\dagger)$$

where $g(\cdot)$ and $h(\cdot)$ are increasing functions with $g(0) = h(0) = 0$.

Various choices of g and h lead to the logistic, Gompertz and Von Bertalanffy models.

1. Logistic (Autocatalytic) Model: If we take $g(f) = f$ and $h(f) = f$ then (†) becomes

$$\frac{\partial f}{\partial t} = \frac{\kappa}{\alpha} f[\alpha - f]$$

where $\kappa > 0$ and $0 < f < \alpha$. Here α is the upper limit of growth, and we've chosen to parameterize it so that κ/α is the proportionality constant.

This differential equation has solution

$$f(t) = \frac{\alpha}{1 + e^{-\kappa(t-\gamma)}}, \quad -\infty < t < \infty.$$

This is the 3-parameter (simple) logistic model. It has asymptotes $f = 0$ as $t \to -\infty$ and $f = \alpha$ as $t \to \infty$.

Again, a variety of parameterizations are possible including

$$f(t) = \frac{\alpha}{1 + \beta e^{-\kappa t}}$$

and

$$f(t) = \frac{\alpha}{1 + e^{(\gamma - t)/\beta}} \qquad \text{(the SSlogis function in nlme)}$$

2. Gompertz Model: Here we take $g(f) = f$ and $h = \log$. This leads to a model in which growth is not symmetric about the point of inflection:

$$f(t) = \alpha \exp\{-e^{-\kappa(t-\gamma)}\}.$$

Here, the point of inflection is at time $t = \gamma$ when the size is $f(\gamma) = \alpha/e$. Again, α is the asymptote as $t \to \infty$.

3. Von Bertalanffy Model: von Bertalanffy hypothesized that the growth rate of an animal with weight f is the difference between the metabolic forces of anabolism and catabolism.

Roughly, anabolism is the process of assimilating new material (e.g, eating, breathing) and catabolism is the loss of building material (e.g., excretion, loss of dead cells, etc.).

By a mix of empiricism and theory, he assumed anabolism was proportional to the 2/3 power of weight and catabolism was proportional to weight. Therefore, growth rate is given by

$$\frac{\partial f}{\partial t} = \eta f^{2/3} - \zeta f.$$

Four-Parameter Sigmoidal Models:

4. Richards Model: Richards doubted the theory underlying von Bertalanffy's model, but noted that if we replace the power 2/3 by an unknown parameter δ, then the differential equation leads to a flexible family of curves with arbitrarily placed point of inflection.

The Richards model can be parameterized in a variety of ways including

$$f(t) = \alpha[1 + (\delta - 1)e^{-\kappa(t-\gamma)}]^{1/(1-\delta)}, \quad \delta \neq 1.$$

- Richards' model generalizes models 1–4. It includes the monomolecular model ($\delta = 0$), the von Bertalanffy ($\delta = 2/3$), the logistic ($\delta = 2$), and the Gompertz (limit as $\delta \to 1$).

- The Richards model and all of its special cases are of the monomolecular form for some transformation of size f.

- The Richards model is not to be confused with the so-called Chapman-Richards model used commonly in forestry

$$f(t) = \alpha\{1 - \exp(-\kappa t)\}^{\delta}$$

  which is not a Richards model at all.

5. Weibull Model: The Richards model can be obtained by assuming that a transformation of size, namely $f^{1-\delta}$, is monomolecular. The Weibull family is obtained by assuming that size f is monomolecular for some power transformation t^{δ} of time.

   In what follows we provide another derivation in terms of the distribution function of the Weibull probability distribution.

   Since the cumulative distribution function (c.d.f.) of any continuous random variable with a unimodal distribution is sigmoidal, such c.d.f.s are a natural place to start in trying to build a sigmoidal growth curve.

   Let $F(x; \boldsymbol{\nu}) = \Pr(X \leq x)$ denote the c.d.f. of a continuous, unimodal random variable X. Here the distribution is assumed to be described by a possibly vector-valued parameter $\boldsymbol{\nu}$.

   There are a couple of different ways to use F to form a sigmoidal growth curve $f(t)$ that have been used. One is to set

$$f(t) = \alpha F(\kappa(t - \gamma); \boldsymbol{\nu}) \tag{1}$$

   and the other is to set

$$f(t) = \beta + (\alpha - \beta) F(\kappa t; \boldsymbol{\nu}). \tag{2}$$

In (1) the time variable is shifted by γ and rescaled by κ. This shifts and expands/contracts the curve in the horizontal direction. In addition, F is rescaled by α so the asymptote is at α rather than 1.

In (2) the time scale is expanded or contracted by κ, and then the curve is shifted vertically to have asymptote β when $t \to -\infty$ and asymptote α when $t \to \infty$ (assuming $\kappa > 0$).

The one-parameter Weibull distribution has c.d.f.

$$F(x;\delta) = 1 - \exp(-x^\delta), \qquad x > 0.$$

Using method (1) we obtain the model

$$f(t) = \alpha(1 - \exp\{-[\kappa(t-\gamma)]^\delta\}),$$

and using (2) we obtain

$$f(t) = \alpha - (\alpha - \beta)\exp\{-(\kappa t)^\delta\}. \qquad (\ddagger)$$

- These two families of curve are different, and each could be used as a legitimate 4-parameter sigmoidal growth curve. However, (‡) is more commonly used and is the one that is typically meant when people say *the* Weibull growth curve model.

6. Morgan-Mercer-Flodin Model: The M-M-F model is given by

$$f(t) = \alpha - \frac{\alpha - \beta}{1 + (\kappa t)^\delta}.$$

This model can be obtained as was the Weibull model by using (1) with c.d.f.

$$F(x;\delta) = \frac{x^\delta}{1 + x^\delta}, \qquad 0 \le x < \infty.$$

In the parameterization given above α is the horizontal asymptote as $t \to \infty$, $\beta = f(0)$ the size at time 0, and δ and κ are shape and scale parameters, respectively.

7. Four-Parameter Logistic Model: A simple generalization of the simple logistic model is to allow both asymptotes to be parameters of the model (for increasing growth, the simple logistic model assumes the lower asymptote is 0).

   This can be done by adding a fourth parameter to the logistic model. The parameterization used in SSfpl in the nlme software is

$$f(t) = \theta_2 + \frac{\theta_1 - \theta_2}{1 + \exp[(\theta_3 - t)/\theta_4]}.$$

   Here, θ_1 is the horizontal asymptote as $t \to \infty$ and θ_2 is the horizontal asymptote as $t \to -\infty$.

- In general, four-parameter sigmoidal models provide more flexibility to fit the curve closely to the data. However, the flip-side of that statement is that they need more information from the data to do so. That is, they need more data and data that cover the entire nonlinear range of the curve to fit all four parameters with any precision.

- Of the four-parameter models, some studies have suggested that the Richards model has the most parameter-effects nonlinearity in typical applications and is often more difficult to fit than the others.

- The book, *Handbook of Nonlinear Regression Models*, by Ratkowsky (1990) provides more examples of growth curve models, and many different parameterizations of the models I have presented here. He also provides some guidance on which models and parameterizations tend to have less parameter-effects nonlinearity, and he discusses methods of obtaining starting values for most of the models he considers.

The usual approach to fitting growth curve models is to fit a model of the form

$$y_i = f(t_i; \boldsymbol{\theta}) + e_i, \qquad i = 1, \ldots, n. \tag{$*$}$$

For cross sectional data (data from independent units at each separate time point), such a model with independent errors is reasonable.

Often, however, growth data aren't collected cross-sectionally. Instead they are collected longitudinally, where individual units are re-measured at several point through time.

For longitudinal measurements of growth the assumption of independent errors is typically inappropriate. In such situations, we may instead assume $\text{var}(\mathbf{e}) = \sigma^2\Lambda$, try to find an appropriate correlation structure for Λ (e.g., an ARMA model) and then fit model (*) with ML or GLS.

This approach often works reasonably well, but sometimes not. Two difficulties are that

i. we often have relatively short series, so that fitting a high order ARMA model is difficult; and

ii. errors from growth data are seldom stationary (some would say that growth data are inherently nonstationary).

For longitudinal data from multiple subjects (the typical situation), such a model can be fit to the data from each subject separately and then these subject-specific models can be pooled or averaged somehow to estimate population level parameters (see, e.g., Davidian & Giltinan, 1995, Ch.5).

This approach has a long history and is an important technique. However, it suffers from some disadvantages:

- It can be difficult to fit models of the form (*) to the data from each and every subject, especially if the number of observations per subject is small, the model (e.g., the correlation structure) is complex, and/or the data are highly variable.

- Although it is possible to assume a common variance-covariance structure with a common var-cov parameter in separately fit models, to do so requires iterative fitting and is cumbersome.

- This approach tends to *overfit* the data.

- There is not good software to implement this approach easily.

Instead, we concentrate on an approach in which a single nonlinear model is fit to the data from all subjects at once, which takes account of the within subject sorrelation and between subject heterogeneity in an elegant, unified way. This approach, based on the class of nonlinear mixed-effect models, will be discussed next, and for the remainder of the course.

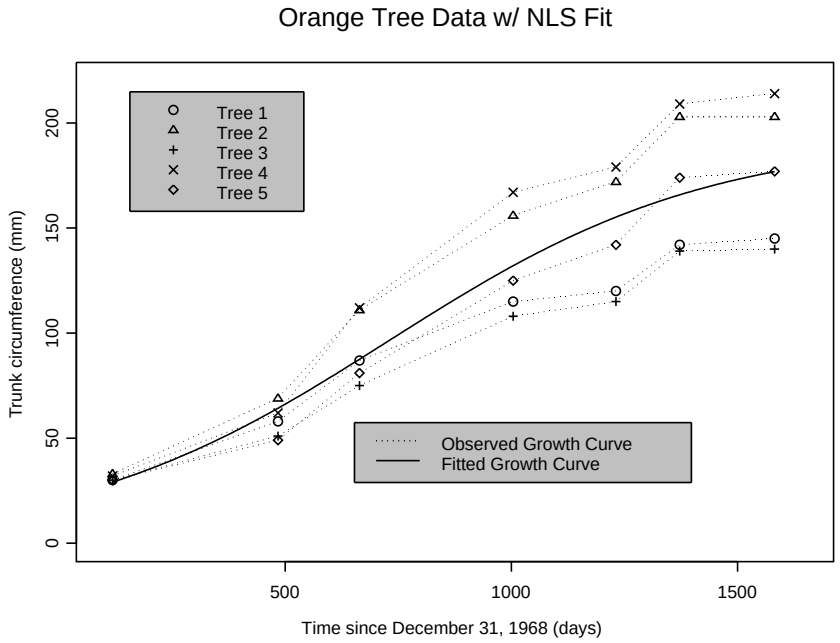
Nonlinear Mixed Effects Models

A Motivating Example — Circumference of Orange Trees

Recall the data in the table below on the circumference of five orange trees over time. We analyzed these data using ordinary (fixed effects) nonlinear regression models in homework #4.

	Tree No.				
Time (days)	1	2	3	4	5
118	30	33	30	32	30
484	58	69	51	62	49
664	87	111	75	112	81
1004	115	156	108	167	125
1231	120	172	115	179	142
1372	142	203	139	209	174
1582	145	203	140	214	177

A plot of the data, with observations from the same tree connected, appears below.



Also displayed in this plot is the fitted curve from a logistic function fit with NLS. That is, if we let y_{ij} = circumference of the i^{th} tree at age t_{ij}, $i = 1, \ldots, 5$, $j = 1, \ldots, 7$, then the fitted model is

$$y_{ij} = \frac{\theta_1}{1 + \exp[-(t_{ij} - \theta_2)/\theta_3]} + e_{ij} \qquad (*)$$

where $\{e_{ij}\} \overset{iid}{\sim} N(0, \sigma^2)$.

- Clearly, model (*) is inadequate.

One obvious deficiency is that, while the fitted curve goes through the center of the combined data from all trees, the growth curves of individual trees, especially large and small trees, are poorly estimated.

- Because the growth curves of the different trees spread out as the trees get older, this misspecification will manifest itself as a cone-shaped residuals vs. fitteds plot suggesting heteroscedasticity.

- In fact, though, it is only (or at least mainly) between-tree variability that is increasing over time. Within-tree error variance looks to be homoscedastic. Simply adding a heteroscedasticity specification to model (*) is not an appropriate solution.

Another deficiency of model (*) is that it treats the observations as independent. There are two obvious potential sources of correlation in these data:

1. Grouping. The data are grouped, or clustered, by tree. Whenever we have grouped data, there is reason to suspect that observations from the same group (tree) will tend to be more similar than observations from different groups. That is, there often is positive within-group or within-cluster correlation and between cluster independence.

    – Minimized by very homogeneous groups.

2. Serial Dependence. As we've noted previously, when data are collected through time, it is often the case that observations close together in time will tend to be correlated more highly than observations far apart in time.

    – Often reduced by long lags between measurements, and/or homogeneous environmental conditions through time.

The first of these sources almost certainly affects the orange tree data and the second may as well.

To deal with the two model deficiencies described above, in homework #4 we fit models in which we allowed the asymptote parameter to differ across trees, and we tried to introduce an appropriate within-tree correlation structure to our model.

That is, we fit a model with 5 separate asymptote parameters, one for each tree:

$$y_{ij} = \frac{\theta_{1i}}{1 + \exp[-(t_{ij} - \theta_2)/\theta_3]} + e_{ij} \qquad (**)$$

and we assumed

$$\text{corr}(\mathbf{e}_i, \mathbf{e}_{i'}) = \begin{cases} \mathbf{0}, & \text{if } i \neq i'; \text{ and} \\ \mathbf{C}(\boldsymbol{\rho}), & \text{if } i = i'. \end{cases}$$

Here $\mathbf{C}$ is an assumed form for the within-group correlation matrix, depending on an unknown parameter $\boldsymbol{\rho}$.

While this approach is clearly an improvement over (*), it has some disadvantages:

A. Number of parameters grows with sample size. In (**) we've introduced a distinct fixed asymptote parameter for each tree. Therefore, if we had measured 500 trees, our model would have 502 regression parameters.

Having the number of parameters increase with the sample size introduces a number of problems:

- Theoretical: in ML and LS estimation, asymptotic arguments establishing consistency, optimality break down.
- Computational: Difficult to optimize a criterion of estimation with respect to many parameters. Hard to form the $\mathbf{V}(\boldsymbol{\theta})$ matrix and solve equations involving large dimension $\{\mathbf{V}(\boldsymbol{\theta})\}^T\mathbf{V}(\boldsymbol{\theta})$ matrix.
- Interpretation: We have 500 separate asymptotes and no single parameter describing the average limit of growth. Do we really care what the limit of growth was for tree #391?
- Conceptual: θ_{1i} is the asympote parameter for tree i. That is, its the fixed theoretical population constant for the limit of growth for tree i. But what's the population? and why is the asymptote of tree i a fixed constant? Wasn't tree i randomly selected from a population of trees? If so, the asymptote of this randomly drawn tree should be regarded as a random variable, not a parameter.

B. Correlation structure. The correlation structure in model (**) accounts for within-group (e.g., within-tree) correlation by modelling source 2 (serial correlation) rather than source 1 (grouping correlation). It is often difficult and unnecessary to model both sources of correlation, but for short time series, modelling 2 is often harder than modelling 1.

That is, it is often not easy to fit an ARMA model to the within-group observations through time. This can be so because of:

- Short series.
- Non-stationary series.
- Unbalanced/missing data and/or irregular or continuous time indexing.

An alternative: A nonlinear mixed-effects model (NLMM) for the orange tree data.

Our fixed effects nonlinear model (**) with 5 separate tree-specific asymptotes is

$$y_{ij} = \frac{\theta_{1i}}{1 + \exp[-(t_{ij} - \theta_2)/\theta_3]} + e_{ij} \qquad (**)$$

Using an ANOVA-type parameterization for θ_{1i} we can write $\theta_{1i} = \bar{\theta}_1 + \tau_i$ where $\sum_{i=1}^{5} \tau_i = 0$. Here $\bar{\theta}_1$ is the average or typical θ_1-value (asymptote) and τ_i is the deviation from the typical value for the i^{th} tree.

Under this parameterization, model (**) becomes

$$y_{ij} = \frac{\bar{\theta}_1 + \tau_i}{1 + \exp[-(t_{ij} - \theta_2)/\theta_3]} + e_{ij} \quad \sum_{i=1}^{5} \tau_i = 0.$$

In the fixed-effects (ordinary) nonlinear regression model, the θ's and the τ's are all considered to be fixed unknown parameters, a.k.a. fixed effects.

In the NLMM, we consider the τ_i's to be random variables, or random effects. τ_i is the deviation from $\bar{\theta}_1$ of the asymptote of the i^{th} tree; it is considered to be random because the tree itself is a randomly selected representative element of the population of trees to which we want to generalize.

Since τ_i is now a random variable, we'd prefer to represent it with a Latin letter rather than a Greek one, so replace the τ_i's with b_i's and the model becomes

$$y_{ij} = \frac{\theta_1 + b_i}{1 + \exp[-(t_{ij} - \theta_2)/\theta_3]} + e_{ij}, \qquad \begin{array}{l} b_1, \ldots, b_5 \overset{iid}{\sim} N(0, \sigma_b^2) \\ \{e_{ij}\} \overset{iid}{\sim} N(0, \sigma^2) \end{array} \qquad (\dagger)$$

Here we've also dropped the bar from $\bar{\theta}_1$.

- Since model (†) contains both fixed effects (the θ's) and random effects (the b_i's) it is called a *mixed-effects* model.

- To completely specify the model we must make distribution assumptions on whatever random variables (error terms, random effects) are in the model. We assume that the random effects are independent normal, with mean 0 (corresponds to assumption that the τ_i's sum to zero) and variance σ_b^2 (distinct from the error variance σ^2).

- In the simplest case, the errors are assumed i.i.d. spherical normal as in the classical nonlinear model. However, this assumption can be relaxed to accommodate heteroscedasticity and.or correlation in the errors.

- We assume the b_i's are uncorrelated with the e_{ij}'s.

- Now the asymptote for the i^{th} tree is $\theta_1 + b_i$, a random variable because b_i is a random variable. The asymptote for the typical tree is θ_1 (when $b_i = 0$).

- If we write $\theta_{1i} \equiv \theta_1 + b_i$, then we have that the 5 asymptotes are randomly distributed around θ_1: $\theta_{11}, \ldots, \theta_{15} \overset{iid}{\sim} N(\theta_1, \sigma_b^2)$.

Fitting Model (†):

The fact that the random effects $\{b_i\}$ enter into the NLMM (†) nonlinearly complicates the methodology and theory of NLMMs substantially as compared to ordinary NLMs.

To focus on the motivation, interpretation, and basic ideas of NLMMs we temporary skip this material and just assume that the nlme() function in R can fit an NLMM with a "good" method.

- See handout Orange1.

- In this R script, we refit the fixed-effects models (*) as m1Oran.gnls, and (**) as m2Oran.gnls. We then fit the NLMM (†) as m1Oran.nlme using the nlme() function.

- nlme() is called in a manner similar to that used in gnls() with fixed= replacing params=. In addition, nlme() takes an argument random= which is used to specify which parameter(s) should have an associated random effect.

- Notice that the NLME (†) has estimated regression parameter $\hat{\boldsymbol{\theta}} = (191.0, 722.6, 344.2)^T$ similar to the estimated regression parameter in the fixed-effects model (*): $\hat{\boldsymbol{\theta}} = (192.7, 728.8, 353.5)^T$.

- Variability in the asymptotes from tree to tree is captured through b_i, which is assumed normal, mean 0, with estimated variance $\hat{\sigma}_b^2 = (31.48)^2$. The error variance is estimated to be $\hat{\sigma}^2 = (7.85)^2$.

- The NLME (†) has AIC=273.2, BIC=280.9 for 5 estimated parameters: $\theta_1, \theta_2, \theta_3, \sigma_b^2, \sigma^2$. This compares with AIC=324.8, BIC=331.0 for the 4-parameter model (*) and AIC=254.1, BIC=266.5 for the 8-parameter model (**).

- So, the addition of random effects in the asymptote in (*) only costs us 1 df and results in a vast improvement in fit. We can do even better by fitting separate asymptotes to each tree, but that shouldn't be surprising. In (†) we save on df in comparison to (**) by making a parametric assumption on the distribution of the random effects: that they're normal with only an unknown variance to estimate rather. In contrast, model (**) doesn't make any assumption about the tree-to-tree variability in asymptotes, it separately estimates each asymptote.

- Of course the residuals of model (*) looked terrible because the individual trees were poorly fit by the average curve. The residuals of model (**) and model (†) look about equally good.

- In model (†) we only fit one asymptote parameter, $\hat{\theta}_1 = 191.0$. An individual tree's (the i^{th} say) asymptote is $\theta_1 + b_i$. Here, b_i is an unobserved (latent) random variable so we don't usually speak of estimating it. Instead, we *predict* its value from the observed data with a prediction $\hat{b}_i$. Then our fitted model for tree i at time t_{ij} is

$$\hat{y}_{ij} = \frac{\hat{\theta}_1 + \hat{b}_i}{1 + \exp[-(t_{ij} - \hat{\theta}_2)/\hat{\theta}_3]}$$

- The $\hat{b}_i$'s aren't estimated parameters of the model. They're predicted quantities based on the fitted model, the data, and the assumption that $b_1, \ldots, b_5 \overset{iid}{\sim} N(0, \sigma_b^2)$.

- The $\hat{b}_i$'s can be obtained from the fitted model using the ranef() function. ranef(m1Oran.nlme) yields $\hat{\mathbf{b}} = (-29.4, 31.6, -37.0, 40.0, -5.18)^T$ (not shown in the handout) so that the predicted circumference of tree 1, say, at time t_{ij} is given by

$$\hat{y}_{ij} = \frac{191.0 - 29.4}{1 + \exp[-(t_{ij} - 722.6)/344.2]}$$

- The predicted curves for individual trees can be obtained from plot(augPred(m1Oran.nlme,level=0:1)) which yields the plots on p.6 or Orange1. The level=0:1 argument here asks for predictions at level 0 (the population level averaged over all trees — corresponds to $b_i = 0$) and at level 1 (here, the tree-level).

- Finally in Orange1, we examine the ACFs for models (*), (**), and (†). The ACF for (*) is affected by both mean misspecification and var-cov misspecification so is not meaningful as a diagnostic of var-cov structure.

- The ACFs of models (**) and (†) are similar — the two models "account for" the residual correlation structure similarly here. This may be somewhat surprising considering that the two models make different assumptions about the correlation between repeated measures.

In particular, the NLM treats all observations including those from the same group (tree) as independent. According to (**),

$$\text{cov}(y_{ij}, y_{ij'}) = \text{cov}(e_{ij}, e_{ij'}) = 0,$$

which implies independence under normality.

In contrast, model (†) implies

$$\begin{aligned}
\text{cov}(y_{ij}, y_{ij'}) &= \text{cov}\left(\frac{\theta_1 + b_i}{1 + \exp[-(t_{ij} - \theta_2)/\theta_3]} + e_{ij}, \frac{\theta_1 + b_i}{1 + \exp[-(t_{ij'} - \theta_2)/\theta_3]} + e_{ij'}\right) \\
&= \text{cov}\left(\frac{\theta_1 + b_i}{1 + \exp[-(t_{ij} - \theta_2)/\theta_3]}, \frac{\theta_1 + b_i}{1 + \exp[-(t_{ij'} - \theta_2)/\theta_3]}\right) \\
&= \frac{\text{cov}(b_i, b_i)}{\{1 + \exp[-(t_{ij} - \theta_2)/\theta_3]\}\{1 + \exp[-(t_{ij'} - \theta_2)/\theta_3]\}} \\
&= \frac{\sigma_b^2}{\{1 + \exp[-(t_{ij} - \theta_2)/\theta_3]\}\{1 + \exp[-(t_{ij'} - \theta_2)/\theta_3]\}} \\
&\neq 0
\end{aligned}$$

- NLMMs imply that observations that share a random effect are correlated! E.g., observations from the same tree in a model with tree-specific random effects are correlated.

- The corresponding NLM with fixed tree effects assumed that observations from the same tree are independent.

Q: *Then why do the two models have (approximately) the same ACF?*

A: Because both model assume that for a given tree, the observations are independent. The NLMM only accounts for correlation due to tree-to-tree differences, which are also accounted for in the NLM with tree effects.

- We will see, however, that residual correlation can be added in to an NLMM. In addition, more complicated random effects specifications can be made, allowing NLMMs to much more flexible model correlation among clustered data.

- In addition, inclusion of fixed effects for each group/cluster/subject in a fixed effect model is not a feasible or attractive option, in general.

- *If the random effects enter into the model in a linear fashion,* then one can obtain a closed-form expression for the implied marginal correlation matrix. E.g., in our example letting $\mathbf{z}_i$ denote the 7×1 vector with j^{th} element $\{1 + \exp[-(t_{ij} - \theta_2)/\theta_3]\}^{-1}$, we have

$$\text{cov}(\mathbf{y}_i) = \sigma_b^2 \mathbf{z}_i \mathbf{z}_i^T + \sigma^2 \mathbf{I}$$

so that

$$\text{corr}(\mathbf{y}_i) = \text{diag}(\mathbf{v}_i)^{-1/2}(\sigma_b^2 \mathbf{z}_i \mathbf{z}_i^T + \sigma^2 \mathbf{I})\text{diag}(\mathbf{v}_i)^{-1/2},$$

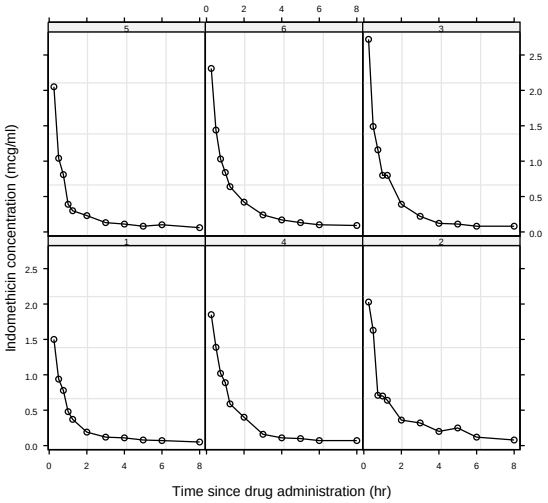
where $\mathbf{v}_i$ denotes the diagonal of $\sigma_b^2 \mathbf{z}_i \mathbf{z}_i^T + \sigma^2 \mathbf{I}$. From our fitted model we can estimate this correlation matrix by plugging in parameter estimates to obtain

$$\hat{\text{corr}}(\mathbf{y}_i) = \begin{pmatrix} 1 & .41 & .45 & .48 & .49 & .49 & .49 \\ & 1 & .70 & .75 & .77 & .77 & .77 \\ & & 1 & .83 & .84 & .84 & .85 \\ & & & 1 & .90 & .90 & .91 \\ & & & & 1 & .92 & .92 \\ & & & & & 1 & .93 \\ & & & & & & 1 \end{pmatrix}$$

Another Example — Pharmacokinetics of Indomethacin

Pinheiro and Bates (2000, §6.2) present and analyze data from a laboratory study by Kwan et al. (1976) on the pharmacokinetics of indomethacin. Six human volunteers received bolus injections of the same dose of indomethacin and had their plasma concentrations of the drug (in mcg/ml) measured 11 times between 15 minutes and 8 hours postinjection. The data are included in the nlme library as a groupedData object called Indometh.

A plot of the data appears below.



Kwan et al. (1976) found that the plasma concentrations for each individual subject were adequately described by a two compartment open model. We will fit two-compartment nonlinear models to these data where we take the model in the sum-of-exponentials form:

$$y_{ij} = \theta_1 \exp[-e^{\theta_2} t_{ij}] + \theta_3 \exp[-e^{\theta_4} t_{ij}] + e_{ij} \quad \begin{matrix} i = 1, \ldots, 6 \\ j = 1, \ldots, 11, \end{matrix} \qquad (1)$$

where y_{ij} = the concentration of indomethacin in plasma at time t_{ij}.

- See handout indometh1.
- In indometh1.R we first plot the concentration over time curves separately by subject in a couple of different ways. It is clear that there is a similar general form across all 6 subjects, but that there is also some subject-to-subject variability in the shape of these curves.

- We fit model (1) with spherical errors as model m1.nls. The third and fourth plots in the handout display the residuals separately by subject. As in the orange tree example, individual curves are poorly fit by a fixed effects model without subject-specific parameters or random effects in the parameters.

- As in the orange tree example, we can account for subject-to-subject heterogeneity by including random effects in the parameters. In this case, though, the decision about which parameters should be modelled with random effects is not as obvious.

- A useful aid for making this decision is to fit model (1) separately by subject (see m1.lis) and then compare the subject-specific estimates of $\theta_1, \ldots, \theta_4$. This can be done graphically via plot(intervals(m1.lis)).

- The only parameter whose subject-specific confidence intervals all overlap is θ_4, but there appears to be significant subject-to-subject variability in $\theta_1, \theta_2, \theta_3$ and (possibly) θ_4. Therefore, we consider a mixed-effects model of the form

$$y_{ij} = \theta_{1i} \exp[-e^{\theta_{2i}} t_{ij}] + \theta_{3i} \exp[-e^{\theta_{4i}} t_{ij}] + e_{ij}, \qquad (2)$$

  where all four of the biexponential parameters have subject-specific random effects:

$$\begin{aligned} \theta_{1i} &= \theta_1 + b_{1i} \\ \theta_{2i} &= \theta_2 + b_{2i} \\ \theta_{3i} &= \theta_3 + b_{3i} \\ \theta_{4i} &= \theta_4 + b_{4i} \end{aligned}$$

- Here, there are four subject-specific random effects in the model. We can think of this as a single 4-dimensional subject-specific random effect $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^T$. Since $\mathbf{b}_i$ is 4-variate, we need to assume a 4-variate distribution for it. Typically, we assume normality:

$$\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_6 \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Psi}).$$

  - Here $\boldsymbol{\Psi}$ is the var-cov matrix of each $\mathbf{b}_i$, and its structure must be assumed as part of the specification of the model.

– Since the four elements of $\mathbf{b}_i$ are each effects specific to the same subject, there is no obvious reason why these effects wouldn't be correlated. So, the minimal assumption is that $\boldsymbol{\Psi}$ is simply a positive-definite matrix, with $4(4+1)/2 = 10$ non-redundant elements to be estimated as parameters of the model:

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} & \psi_{13} & \psi_{14} \\ \psi_{12} & \psi_{22} & \psi_{23} & \psi_{24} \\ \psi_{13} & \psi_{23} & \psi_{33} & \psi_{34} \\ \psi_{14} & \psi_{24} & \psi_{34} & \psi_{44} \end{pmatrix}$$

– Often, a completely general positive-definite form as above results in convergence problems and it is easier to fit a model where $\boldsymbol{\Psi}$ is assumed to be diagonal initially (corresponding to uncorrelated subject-specific random effects). Once convergence has been obtained for the initial model, we may consider relaxing the diagonal $\boldsymbol{\Psi}$ assumption to allow it to be general positive-definite, or possibly some intermediate form (e.g., block-diagonal).

- In indometh1.R, we fit model (2) as m1.nlme. A summary of m1.nlme reveals that the sd of b_{i4} is estimated as $\hat{\psi}_{44} = 3.44 \times 10^{-6}$, very small. This was somewhat expected from our plot of the confidence intervals from m1.lis. Therefore, we consider eliminating the random effect from θ_4.

- The resulting model is m2.nlme. Comparing models with nested random effects structures (like m1.nlme and m2.nlme) via LRTs seems like a reasonable approach. However, this situation is one in which the usual χ^2 reference distribution is no longer appropriate. The problem is a non-standard one not falling under the general theory of Wilks' Theorem. The bottom line is that the naive approach of using anova(m1.nlme,m2.nlme) which compares

  2[logLik(m1.nlme)-logLik(m2.nlme)]

  to a $\chi^2(1)$ distribution, results in a conservative test (overestimated $p-$value). For more on this issue see Pinheiro and Bates (2000, §2.4.1).

- We go ahead and use anova(m1.nlme,m2.nlme) in indometh1.R knowing that the $p-$value will be overestimated. However, the $p-$value from this comparisons ($p = .9512$) is so large that there can be no doubt that model m2.nlme is to be preferred over m1.nlme. The two models have nearly identical loglikelihoods.

- Next we fit model m3.nlme in which we allow b_{1i}, b_{2i}, b_{3i} to be correlated. That is,

$$\text{var}(\mathbf{b}_i) = \text{var}\begin{pmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{pmatrix} = \boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{12} & \psi_{22} & \psi_{23} \\ \psi_{13} & \psi_{23} & \psi_{33} \end{pmatrix}$$

- An examination of summary(m3.nlme) reveals that the only pair of random effects that are estimated to be highly correlated are b_{1i} and b_{2i}, so we consider a model intermediate between model m2.nlme and m3.nlme in which $\boldsymbol{\Psi}$ is assumed to have block-diagonal structure:

$$\text{var}(\mathbf{b}_i) = \text{var}\begin{pmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{pmatrix} = \boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} & 0 \\ \psi_{12} & \psi_{22} & 0 \\ 0 & 0 & \psi_{33} \end{pmatrix}$$

- This model is fit as m4.nlme. In this case, since the null hypothesis $H_0 : \psi_{13} = \psi_{23} = 0$ doesn't place the parameters on the boundary of their parameter space, we could go ahead and use a LRT to compare models m3.nlme, and m4.nlme. However, an easier and more widely valid approach to selecting the variance-covariance structure is to use use AIC or BIC. Both of these criteria point to model m4.nlme over m3.nlme and m2.nlme.

- The final two plots in indometh1 display the residuals and fitted curves from model m4.nlme. Both indicate that the model fits the data fairly well. The last plot displays the population average curve (solid line) corresponding to $\mathbf{b}_i = \mathbf{0}$ and subject-specific predicted concentration over time curves (dotted lines).

- Finally, note that the parameter estimates from model m4.nlme, $\hat{\boldsymbol{\theta}} = (2.81, .849, .587, -1.11)^T$ are similar to those from the fixed-effects model m1.nls, but standard errors have changed appreciably.

The NLME Model Formulation

By far the most important area of application of NLMEs is for grouped or clustered data, particularly longitudinal or repeated measures data. In describing the NLME model, we first present the single-level-of-grouping and then extend to multilevel data.

- E.g., in an educational context, single level data might be repeated measures through time on each of several students in Mrs. Smith's third grade class at Barrow Elementary School. Multilevel (in this case 3-level) data might be repeated measures through time on each of several students (level 3) in each of several classes (level 2) in each of several schools (level 1) in the Athens-Clarke County School District.

Formulation for Single Level Data:

Let y_{ij} denote the j^{th} observation (e.g., through time) on the i^{th} group (i.e., cluster; e.g, subject) where we have M groups, and n_i observations in the i^{th} group. Let $\mathbf{x}_{ij}$ be a vector of covariates corresponding to response y_{ij}.

Then our NLMM is

$$\begin{aligned} & y_{ij} = f(\boldsymbol{\theta}_{ij}, \mathbf{x}_{ij}) + e_{ij}, \quad \begin{matrix} i = 1, \ldots, M \\ j = 1, \ldots, n_i \end{matrix}, \\ \text{where} \quad & \boldsymbol{\theta}_{ij} = \mathbf{A}_{ij}\boldsymbol{\beta} + \mathbf{B}_{ij}\mathbf{b}_i, \quad \begin{matrix} \mathbf{b}_1, \ldots, \mathbf{b}_M \overset{iid}{\sim} N(0, \boldsymbol{\Psi}) \\ \{e_{ij}\} \overset{iid}{\sim} N(0, \sigma^2) \end{matrix} \end{aligned} \qquad (*)$$

Here, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, and $\mathbf{b}_i$ is a $q \times 1$ vector of random effects specific to the i^{th} group with var-cov matrix $\boldsymbol{\Psi}$. The matrices $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ are model matrices for the fixed and random effects, respectively.

Model (*) can be equivalently expressed in a more succinct matrix form as

$$\begin{aligned} \mathbf{y}_i &= \mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{x}_i) + \mathbf{e}_i, \\ \boldsymbol{\theta}_i &= \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \end{aligned} \qquad (**)$$

for $i = 1, \ldots, M$, where

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix}, \quad \boldsymbol{\theta}_i = \begin{pmatrix} \boldsymbol{\theta}_{i1} \\ \vdots \\ \boldsymbol{\theta}_{in_i} \end{pmatrix}, \quad \mathbf{e}_i = \begin{pmatrix} e_{i1} \\ \vdots \\ e_{in_i} \end{pmatrix}, \quad \mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{x}_i) = \begin{pmatrix} f(\boldsymbol{\theta}_{i1}, \mathbf{x}_{i1}) \\ \vdots \\ f(\boldsymbol{\theta}_{in_i}, \mathbf{x}_{in_i}) \end{pmatrix},$$

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{in_i} \end{pmatrix}, \quad \mathbf{A}_i = \begin{pmatrix} \mathbf{A}_{i1} \\ \vdots \\ \mathbf{A}_{in_i} \end{pmatrix}, \quad \mathbf{B}_i = \begin{pmatrix} \mathbf{B}_{i1} \\ \vdots \\ \mathbf{B}_{in_i} \end{pmatrix}.$$

We assume

$$\mathbf{b}_1, \ldots, \mathbf{b}_M \overset{iid}{\sim} N_q(\mathbf{0}, \boldsymbol{\Psi}), \quad \{\mathbf{e}_i\} \overset{iid}{\sim} N_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i})$$

and the random effects $\{\mathbf{b}_i\}$ are independent of the errors $\{\mathbf{e}_i\}$.

Example — Orange Tree Data

To illustrate the model formulation, we write model (†) that we used for these data in the form (**). Model (†) can be written as

$$y_{ij} = \frac{\theta_{1ij}}{1 + \exp[-(t_{ij} - \theta_{2ij})/\theta_{3ij}]} + e_{ij},$$

$$\text{where} \quad \underbrace{\begin{pmatrix} \theta_{1ij} \\ \theta_{2ij} \\ \theta_{3ij} \end{pmatrix}}_{\boldsymbol{\theta}_{ij}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{A}_{ij}} \underbrace{\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}}_{\mathbf{B}_{ij}} \underbrace{(\, b_{1i} \,)}_{\mathbf{b}_i},$$

where $\mathbf{b}_i = b_i$ is a scalar, so $q = 1$ and

$$b_1, \ldots, b_M \overset{iid}{\sim} N(0, \underbrace{\sigma_b^2}_{\boldsymbol{\Psi}}), \quad \{e_{ij}\} \overset{iid}{\sim} N(0, \sigma^2).$$

In this simple example, the individual coefficients $\boldsymbol{\theta}_{ij}$ and the model matrices $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ are indexed by j but don't vary with j (don't change over time). The var-cov matrix of the random effects $\boldsymbol{\Psi}$ is a scalar variance, σ_b^2.

Another Example — Indomethacin Data

The model that we chose for these data in indometh1 can be written in the form of (*) as follows:

$$y_{ij} = \theta_{1ij}\exp[-e^{\theta_{2ij}}t_{ij}] + \theta_{3ij}\exp[-e^{\theta_{4ij}}t_{ij}] + e_{ij},$$

$$\text{where} \quad \underbrace{\begin{pmatrix}\theta_{1ij}\\ \theta_{2ij}\\ \theta_{3ij}\\ \theta_{4ij}\end{pmatrix}}_{\boldsymbol{\theta}_{ij}} = \underbrace{\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}}_{\mathbf{A}_{ij}}\underbrace{\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\\0&0&0\end{pmatrix}}_{\mathbf{B}_{ij}}\underbrace{\begin{pmatrix}b_{1i}\\ b_{2i}\\ b_{3i}\end{pmatrix}}_{\mathbf{b}_i},$$

where

$$\mathbf{b}_1,\ldots,\mathbf{b}_M \overset{iid}{\sim} N_3\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}\psi_{11}&\psi_{12}&0\\ \psi_{12}&\psi_{22}&0\\ 0&0&\psi_{33}\end{pmatrix}\right), \quad \{e_{ij}\} \overset{iid}{\sim} N(0,\sigma^2).$$

- Again, here the individual coefficients $\boldsymbol{\theta}_{ij}$ and the design matrices $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ do not vary over j (time).

Multilevel Formulation:

A Two-level Example:

The groupedData object Wafer in the nlme library for R/S-PLUS, contains data from an experiment conducted at Lucent Technologies to study different sources of variability in the manufacture of analog MOS(?) circuits.

The intensity of current (in mA) at 0.8, 1.2, 1.6, 2.0, and 2.4 V was measured on manufactured electronic devices. Measurements were made on eight sites in each of ten wafers selected from the same lot (batch of product).

The main objective of the experiment was to construct an empirical model for simulating the behavior of similar circuits.

In this example, there are two nested blocking factors, *wafer* and *site within wafer*. We expect there may be heterogeneity from site to site and wafer to wafer, and consequently correlation for observations obtained from the same site, or from the same wafer.

- It is reasonable to expect that two observations from different wafers will be independent (and hence uncorrelated).
- We expect that two observations from the same wafer, but from different sites, will be correlated.
- Two observations from the same wafer, and the same site within the wafer, we expect to be even more strongly correlated.

A natural way to account for this expected correlation structure is with a NLMM where we have random effects for wafers and random effects for sites within wafers.

- Here, the levels of site are nested within the levels of wafers. That is, site 1 in wafer 1 is not the same site as site 1 in wafer 2.
- Therefore, the site-specific random effects are nested within the wafer-specific random effects in our model.
- This is an example of two-level grouped data.

Our convention on numbering the levels is that level 0 is the population level (averaged over all wafers and sites), level 1 is the wafer level (coarsest level of grouping), and level 2 is the site level (finest level of grouping).

- As an example of > 2 levels of nesting, suppose the experiment was run on devices from several lots. Then we would have a 3-level situation with data grouped by lots (level 1) , wafers within lots (level 2), and sites within wafers within lots (level 3).

For two-level data, let y_{ijk} = the k^{th} observation on the j^{th} second level group (e.g., site), on the i^{th} first level group (e.g., wafer).

We suppose we have M first level groups (e.g., wafers), M_i second level groups within the i^{th} first level group (M_i sites within wafer i), and n_{ij} observations on the j^{th} second level unit within the i^{th} first level unit (n_{ij} observations on the j^{th} site within the i^{th} wafer).

Then the two-level NLMM is

$$y_{ijk} = f(\boldsymbol{\theta}_{ijk}, \mathbf{v}_{ijk}) + e_{ijk}, \quad \begin{array}{l} i = 1, \ldots, M \\ j = 1, \ldots, M_i, \\ k = 1, \ldots, n_{ij} \end{array}$$

$$\text{where} \quad \boldsymbol{\theta}_{ijk} = \mathbf{A}_{ijk}\boldsymbol{\beta} + \mathbf{B}_{i,jk}\mathbf{b}_i + \mathbf{B}_{ijk}\mathbf{b}_{ij}, \quad \begin{array}{l} \mathbf{b}_1, \ldots, \mathbf{b}_M \overset{iid}{\sim} N(0, \boldsymbol{\Psi}_1) \\ \mathbf{b}_{11}, \ldots, \mathbf{b}_{M,M_i} \overset{iid}{\sim} N(0, \boldsymbol{\Psi}_2) \\ \{e_{ijk}\} \overset{iid}{\sim} N(0, \sigma^2) \end{array}$$

- Here the $\mathbf{b}_i$'s, $\mathbf{b}_{ij}$'s and e_{ijk}'s are assumed uncorrelated with eachother.
- Extension to $>$ 2-level NLMMs follows in a straight-forward, but notationally tedious way.
- NLMMs with crossed (rather than nested) random effects can be formulated. However, such models are much more difficult to fit and we will not use them at all in this course. Fortunately, situations for which they are appropriate are much less common then single-level or multilevel nested random effects models.

Estimation and Inference in NLMMs

Over the last 20–25 years, a huge literature has appeared on estimation and inference in NLMMs and in the closely related class of generalized linear mixed effects models (GLMMs). We follow the treatment of Pinheiro and Bates (2000, ch. 7) and restrict attention to methods based on the likelihood function. More comprehensive treatments can be found in the reserve books by Davidian and Giltinan (1995) and Vonesh and Chinchilli (1997).

For ML estimation in the NLMM, as in general, we choose the values of the parameters that maximize the (log)likelihood (the parameter-values under which the data are most likely). To do this maximization, we need to compute the likelihood function, and its derivatives with respect to the parameters.

- Unfortunately, in general this is hard for NLMMs because the random effects enter into the model in a nonlinear fashion.

For a 1-level NLMM the parameters of the model are $\boldsymbol{\beta}, \sigma^2$ and the unknown elements of $\boldsymbol{\Psi}$. So, we will write the likelihood function based on all of the data as $L(\beta, \sigma^2, \boldsymbol{\Psi}; \mathbf{y})$, where $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_M^T)^T$, is the combined data vector from all groups.

$L(\beta, \sigma^2, \boldsymbol{\Psi}; \mathbf{y})$ is just equal to the joint probability density function

$$p(\mathbf{y}; \beta, \sigma^2, \boldsymbol{\Psi}) = \prod_{i=1}^{M} p(\mathbf{y}_i; \beta, \sigma^2, \boldsymbol{\Psi}) \qquad (\ddagger)$$

(since $\mathbf{y}_1, \ldots, \mathbf{y}_M$, the observation vectors from distinct lowest-level groups, are independent).

- Here we've used $p(\cdot)$ to denote a density function rather than $f(\cdot)$ to avoid confusion with the model function f.

We need to translate our model specification into an expression for $p(\mathbf{y}_i; \beta, \sigma^2, \boldsymbol{\Psi})$.

Q: What does our model imply about the density of $\mathbf{y}_i$*?*

A: Directly, nothing. But indirectly, the model assumptions imply the form of $p(\mathbf{y}_i; \beta, \sigma^2, \boldsymbol{\Psi})$.

How?:

First of all, it is clear that the *conditional* density $p(\mathbf{y}_i|\mathbf{b}_i; \beta, \sigma^2, \boldsymbol{\Psi})$ given $\mathbf{b}_i$ is a normal density, inherited from the normality of the vector of error terms $\mathbf{e}_i$.

This is clear from (**), p.238. Conditional on $\mathbf{b}_i$, $\boldsymbol{\theta}_i$ is non-random, so that $\mathbf{y}_i = \mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{x}_i) + \mathbf{e}_i$ has the usual form of a fixed effects NLM, and $\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ implies

$$\mathbf{y}_i|\mathbf{b}_i \sim N\left(\mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{x}_i), \sigma^2\mathbf{I}\right).$$

Therefore, $p(\mathbf{y}_i|\mathbf{b}_i; \beta, \sigma^2, \boldsymbol{\Psi})$ is given by a multivariate normal density with mean $\mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{x}_i)$ and var-cov matrix $\sigma^2\mathbf{I}$.

Secondly, the model directly assumes the density $p(\mathbf{b}_i; \boldsymbol{\Psi})$ of the random effects vector $\mathbf{b}_i$ to be a multivariate normal.

Thirdly, we have the general relationship between a marginal density such as $p(\mathbf{y}_i; \beta, \sigma^2, \boldsymbol{\Psi})$ and a conditonal one:

$$p(\mathbf{y}_i; \beta, \sigma^2, \boldsymbol{\Psi}) = \int p(\mathbf{y}_i|\mathbf{b}_i; \beta, \sigma^2, \boldsymbol{\Psi}) p(\mathbf{b}_i; \boldsymbol{\Psi}) d\mathbf{b}_i.$$

- Here the integral is with respect to $\mathbf{b}_i$ so is a multidimensional integral.

Putting these results together, we have that $p(\mathbf{y}_i; \beta, \sigma^2, \boldsymbol{\Psi})$ is given by the integral of a product of multivariate normal densities. Substituting these normal densities and using (‡), we have

$$p(\mathbf{y}; \beta, \sigma^2, \boldsymbol{\Psi}) = \prod_{i=1}^{M} \int (2\pi\sigma^2)^{-n_i/2} \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{x}_i)\|^2\right] \times (2\pi)^{-q/2}|\boldsymbol{\Psi}|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{b}_i^T\boldsymbol{\Psi}^{-1}\mathbf{b}_i\right] d\mathbf{b}_i$$

To simplify and work with this expression, its convenient to reexpress $\boldsymbol{\Psi}^{-1}$ as follows: Let Δ be a square-root matrix of $\sigma^2\boldsymbol{\Psi}^{-1}$ so that

$$\boldsymbol{\Psi}^{-1} = \sigma^{-2}\Delta^T\Delta.$$

Then, with a fair amount of algebra, we can simplify $p(\mathbf{y}; \beta, \sigma^2, \boldsymbol{\Psi})$ above as

$$p(\mathbf{y}; \beta, \sigma^2, \boldsymbol{\Psi}) = \prod_{i=1}^{M} \frac{|\Delta|}{(2\pi\sigma^2)^{(n_i+q)/2}} \int \exp\left[\frac{\|\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i)\|^2 + \|\Delta\mathbf{b}_i\|^2}{-2\sigma^2}\right] d\mathbf{b}_i.$$

- Here we have changed notation replacing $\mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{x}_i)$ with $\mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i)$ to make explicit the dependence of this quantity on the random effects vector $\mathbf{b}_i$.

Thus, we have an expression for the likelihood function:

$$L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Psi}; \mathbf{y}) = \prod_{i=1}^{M} \frac{|\Delta|}{(2\pi\sigma^2)^{(n_i+q)/2}} \int \exp\left[\frac{\|\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i)\|^2 + \|\Delta\mathbf{b}_i\|^2}{-2\sigma^2}\right] d\mathbf{b}_i.$$

- Unfortunately, because the random effects can enter into the model nonlinearly, this likelihood involves an integral that does not, in general, have a closed form expression.

- The presence of the integral in $L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Psi}; \mathbf{y})$ is what makes NLMMs especially hard to work with, statistically.

Many authors have proposed methods to fit NLMMs that deal with the integral in $L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Psi}; \mathbf{y})$ in various ways. The approaches generally fall in three main categories:

1. Numerical integration (a.k.a., quadrature) methods.

2. Monte Carlo (simulation based) methods.

3. Approximate maximum likelihood/estimating equation methods.

    a. Methods based on approximations to the loglikelihood involving Taylor series/Laplace approximations with expansions taken about $\mathrm{E}(\mathbf{b}_i) = \mathbf{0}$, the mean of the random effects vector.

    b. Methods based on approximations to the loglikelihood involving Taylor series/Laplace approximations with expansions taken about $\hat{\mathbf{b}}_i$, some predicted value of the random effects vector.

We will describe only two methods:

A. The LME Approximation method of Lindstrom and Bates (1990). (This falls in category 3b, and is the method implemented in the nlme software in R).

B. Adaptive Gaussian Quadrature. (This method falls in category 1, and is implemented in SAS' PROC NLMIXED, for 1-level models).

The LME Approximation:

The LME approximation can be derived in several different ways, but the easiest motivation is to obtain the approximate log-likelihood of model (**) by using the loglikelihood of a linear approximation of model (**), p.238.

Recall our single-level NLMM from p.238:

$$\left.\begin{array}{l}\mathbf{y}_i = \mathbf{f}_i(\boldsymbol{\theta}_i, \mathbf{x}_i) + \mathbf{e}_i \equiv \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i) + \mathbf{e}_i, \\ \boldsymbol{\theta}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i,\end{array}\right. \quad i = 1, \ldots, M. \qquad (**)$$

Taking a first-order (linear) Taylor series approximation of $\mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i)$ about $\hat{\mathbf{b}}_i$, a predictor of $\mathbf{b}_i$, we have

$$\mathbf{y}_i \approx \mathbf{f}_i(\boldsymbol{\beta}, \hat{\mathbf{b}}_i) + \hat{\mathbf{Z}}_i(\mathbf{b}_i - \hat{\mathbf{b}}_i) + \mathbf{e}_i$$

where

$$\hat{\mathbf{Z}}_i = \left.\frac{\partial \mathbf{f}_i}{\partial \mathbf{b}_i^T}\right|_{\mathbf{b}_i = \hat{\mathbf{b}}_i}$$

Rearranging, we have

$$\mathbf{y}_i + \hat{\mathbf{Z}}_i\hat{\mathbf{b}}_i = \mathbf{f}_i(\boldsymbol{\beta}, \hat{\mathbf{b}}_i) + \hat{\mathbf{Z}}_i\mathbf{b}_i + \mathbf{e}_i, \qquad (\clubsuit)$$

- Note that now $\mathbf{b}_i$ enters into model ($\clubsuit$) linearly, not nonlinearly.
- Treating $\hat{\mathbf{b}}_i$ as fixed, we can view model ($\clubsuit$) as a nonlinear model with response $\mathbf{y}_i + \hat{\mathbf{Z}}_i\hat{\mathbf{b}}_i$ that is multivariate normal, with mean

  $$\mathrm{E}(\mathbf{f}_i(\boldsymbol{\beta}, \hat{\mathbf{b}}_i) + \hat{\mathbf{Z}}_i\mathbf{b}_i + \mathbf{e}_i) = \mathbf{f}_i(\boldsymbol{\beta}, \hat{\mathbf{b}}_i)$$

  and variance

  $$\begin{aligned}\mathrm{var}(\mathbf{f}_i(\boldsymbol{\beta}, \hat{\mathbf{b}}_i) + \hat{\mathbf{Z}}_i\mathbf{b}_i + \mathbf{e}_i) &= \mathrm{var}(\hat{\mathbf{Z}}_i\mathbf{b}_i + \mathbf{e}_i) \\ &= \hat{\mathbf{Z}}_i\boldsymbol{\Psi}\hat{\mathbf{Z}}_i^T + \sigma^2\mathbf{I} \\ &= \sigma^2(\hat{\mathbf{Z}}_i\Delta^{-1}\Delta^{-T}\hat{\mathbf{Z}}_i^T + \mathbf{I}) \equiv \sigma^2\Sigma_i(\Delta)\end{aligned}$$

Therefore, the likelihood of model (♣), which we take as the LME approximate likelihood of model (**), is

$$\prod_{i=1}^{M}(2\pi\sigma^2)^{-n_i/2}|\Sigma_i(\Delta)|^{-1/2}$$
$$\times \exp\left[-\frac{1}{2\sigma^2}\{\mathbf{y}_i+\hat{\mathbf{Z}}_i\hat{\mathbf{b}}_i-\mathbf{f}_i(\boldsymbol{\beta},\hat{\mathbf{b}}_i)\}^T\Sigma_i(\Delta)^{-1}\{\mathbf{y}_i+\hat{\mathbf{Z}}_i\hat{\mathbf{b}}_i-\mathbf{f}_i(\boldsymbol{\beta},\hat{\mathbf{b}}_i)\}\right].$$

- Note that the above approximate likelihood can also be obtained directly from $L(\boldsymbol{\beta},\sigma^2,\boldsymbol{\Psi};\mathbf{y})$ on p.245, by applying a quadratic Taylor approximation of the integrand around $\hat{\mathbf{b}}_i$.

The LME approximate log-likelihood is

$$-\frac{1}{2}\sum_{i=1}^{M}\Big[n_i\log(2\pi\sigma^2)+\log|\Sigma_i(\Delta)|$$
$$+\{\mathbf{y}_i+\hat{\mathbf{Z}}_i\hat{\mathbf{b}}_i-\mathbf{f}_i(\boldsymbol{\beta},\hat{\mathbf{b}}_i)\}^T\Sigma_i(\Delta)^{-1}\{\mathbf{y}_i+\hat{\mathbf{Z}}_i\hat{\mathbf{b}}_i-\mathbf{f}_i(\boldsymbol{\beta},\hat{\mathbf{b}}_i)\}\Big]. \quad (\heartsuit)$$

This loglikelihood is maximized iteratively, alternating between a step to estimate $\boldsymbol{\beta}$ and obtain the predictor $\hat{\mathbf{b}}_i$ for fixed Δ and a step to estimate Δ for fixed values of $\boldsymbol{\beta}$ and $\hat{\mathbf{b}}_i$.

Step 1 — Estimating $\boldsymbol{\beta}$ and updating the predictor $\hat{\mathbf{b}}_i$:

We obtain these quantities by maximizing ($\heartsuit$) with respect to $\boldsymbol{\beta}$ and $\hat{\mathbf{b}}_i$, $i=1,\ldots,M$, with one simplification: we ignore the dependence of $\Sigma_i(\Delta)$ on $\boldsymbol{\beta}$.

- It can be argued that $\Sigma_i(\Delta)$ will, in general, vary slowly with $\boldsymbol{\beta}$ and therefore the term $\partial\Sigma_i(\Delta)/(\partial\boldsymbol{\beta})$ will be negligible (see Wolfinger & Lin, 1997).

With this simplification, we see that maximizing (♡) with respect to $\boldsymbol{\beta}$ and $\hat{\mathbf{b}}_i$ amounts to minimizing the quantity

$$\sum_{i=1}^{M}\{\mathbf{y}_i+\hat{\mathbf{Z}}_i\hat{\mathbf{b}}_i-\mathbf{f}_i(\boldsymbol{\beta},\hat{\mathbf{b}}_i)\}^T\Sigma_i(\Delta)^{-1}\{\mathbf{y}_i+\hat{\mathbf{Z}}_i\hat{\mathbf{b}}_i-\mathbf{f}_i(\boldsymbol{\beta},\hat{\mathbf{b}}_i)\}$$
$$=\sum_{i=1}^{M}\left[\|\mathbf{y}_i-\mathbf{f}_i(\boldsymbol{\beta},\hat{\mathbf{b}}_i)\|^2+\|\Delta\hat{\mathbf{b}}_i\|^2\right].$$

- Notice from the second form given above, it's appropriate to term this objective function a *penalized nonlinear least squares (PNLS) criterion.*

Step 2 — Estimation of Δ, σ^2:

Maximization of (♡) with respect to Δ and σ^2 is done by first profiling this approximate loglikelihood with respect to these parameters. That is, we substitute $\hat{\boldsymbol{\beta}}(\Delta,\sigma^2)$, the estimator of $\boldsymbol{\beta}$ based on the current Δ,σ^2, into (♡). This profiling yields the objective function

$$-\frac{1}{2}\sum_{i=1}^{M}\left[n_i\log(2\pi\sigma^2)+\log|\Sigma_i(\Delta)|+\{\mathbf{w}_i-\hat{\mathbf{X}}_i\boldsymbol{\beta}\}^T\Sigma_i(\Delta)^{-1}\{\mathbf{w}_i-\hat{\mathbf{X}}_i\boldsymbol{\beta}\}\right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\Delta,\sigma^2)},$$

where

$$\hat{\mathbf{X}}_i=\frac{\partial\mathbf{f}_i(\boldsymbol{\beta},\hat{\mathbf{b}}_i)}{\partial\boldsymbol{\beta}^T},\quad\text{and}\quad\mathbf{w}_i=\mathbf{y}_i-\mathbf{f}_i(\boldsymbol{\beta},\hat{\mathbf{b}}_i)+\hat{\mathbf{X}}_i\boldsymbol{\beta}+\hat{\mathbf{Z}}_i\hat{\mathbf{b}}_i.$$

This objective function is the loglikelihood of a <u>linear</u> mixed-effects model in which the response vector is $\mathbf{w}=(\mathbf{w}_1^T,\ldots,\mathbf{w}_M^T)^T$, and the fixed- and random-effects design matrices are $\hat{\mathbf{X}}=(\hat{\mathbf{X}}_1^T,\ldots,\hat{\mathbf{X}}_M^T)^T$ and $\hat{\mathbf{Z}}=(\hat{\mathbf{Z}}_1^T,\ldots,\hat{\mathbf{Z}}_M^T)^T$, respectively.

This means that step 2 can be accomplished by fitting a linear mixed-effects model using standard techniques for that that kind of a problem.

Thus, the whole LME approximation method iterates between a PNLS step (step 1) and a linear mixed-effects (LME) step (step 2). These two steps are repeated until convergence.

- For extension to the multilevel model, see Pinheiro and Bates (2000, ch. 7).

Adaptive Gaussian Quadrature:

A computationally more intensive, but generally more accurate method to evaluate the likelihood function of an NLMM is adaptive Gaussian quadrature.

The basic idea of (ordinary) Gaussian quadrature is to approximate an integral of the form $\int_a^b g(x)w(x)dx$ by a finite weighted sum of the function $g(\cdot)$ evaluated at a set of x-values chosen to be optimal for the particular form of the weight function $w()$ in the integral.

That is, since an integral of the form $\int_a^b g(x)w(x)dx$ has an interpretation as an infinite weighted sum, approximate it by a finite weighted sum:

$$\int_a^b g(x)w(x)dx \approx \sum_{k=1}^{m} w_i g(x_i) \qquad (\spadesuit)$$

where $w_1, \ldots, w_m$ are weights and $x_1, \ldots, x_m$ are *abscissas* chosen based on the form of $w(\cdot)$.

- In ($\spadesuit$), we use an m-term sum to approximate the integral, so this is called m-point Gaussian quadrature. The accuracy of the approximation increases with m, although not linearly.

In an NLMM context, the integral we require is

$$\int \exp\left[\frac{\|\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i)\|^2 + \|\Delta \mathbf{b}_i\|^2}{-2\sigma^2}\right] d\mathbf{b}_i \qquad (\diamondsuit)$$

(cf. bottom of p.245).

With the change of variable $\mathbf{z} = \sigma^{-1}\Delta\mathbf{b}_i$, this integral can be written as

$$\int \sigma^q |\Delta|^{-1} \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \sigma\Delta^{-1}\mathbf{z})\|^2\right] \exp\left(-\frac{1}{2}\|\mathbf{z}\|^2\right) d\mathbf{z}$$

and we have $\exp\left(-\frac{1}{2}\|\mathbf{z}\|^2\right)$ playing the role of the weight function $w(x)$ in ($\spadesuit$).

Suppose $q = 1$. Then the integral that we wish to approximate is 1-dimensional, and we can apply the Gaussian quadrature approximation given by (♠). In this case, we have

$$(\diamondsuit) \approx \sigma|\Delta|^{-1} \sum_{j=1}^{m} \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \sigma\Delta^{-1}z_j)\|^2\right] w_j$$

where z_j, w_j $j = 1, \ldots, m$, are abscissas and weights, respectively, for the $\exp(-\frac{1}{2}z^2)$ weight function. These are known quantities, that can be looked up in tables or calculated with computer programs.

If $q > 1$, then the integral we wish to approximate is multidimensional. In this case, the (ordinary) Gaussian quadrature approximations becomes a q-term nested sum:

$$(\diamondsuit) \approx \sigma^q|\Delta|^{-1} \sum_{j_1=1}^{m} \cdots \sum_{j_q=1}^{m} \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \sigma\Delta^{-1}\mathbf{z}_j)\|^2\right] \prod_{k=1}^{m} w_{j_k}$$

where now $\mathbf{z}_j$ is a vector of abscissa values: $\mathbf{z}_j = (z_{j_1}, \ldots, z_{j_q})^T$.

The adaptive Gaussian quadrature approximation is much the same, except that the weighted sum approximation is evaluated at abscissas-values centered around $\hat{\mathbf{b}}_i$, our predictor of the random effects vector $\mathbf{b}_i$ rather than around $\mathrm{E}(\mathbf{b}_i) = \mathbf{0}$ as in ordinary Gaussian quadrature. In addition, the abscissas are rescaled.

This recentering and rescaling is accomplished by the change of variable $\mathbf{z} = \sigma^{-1}(\hat{\mathbf{Z}}_i^T\hat{\mathbf{Z}}_i + \Delta^T\Delta)^{1/2}(\mathbf{b}_i - \hat{\mathbf{b}}_i)$ (rather than the one given at the bottom of the previous page). This leads to the adaptive Gaussian quadrature approximation

$$\begin{aligned}(\diamondsuit) \approx{}& \sigma^q|\hat{\mathbf{Z}}_i^T\hat{\mathbf{Z}}_i + \Delta^T\Delta|^{-1/2} \\ &\times \sum_{j_1=1}^{m} \cdots \sum_{j_q=1}^{m} \exp\Big\{ -\frac{1}{2\sigma^2}\big[\|\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \hat{\mathbf{b}}_i + \sigma(\hat{\mathbf{Z}}_i^T\hat{\mathbf{Z}}_i + \Delta^T\Delta)^{-1/2}\mathbf{z}_j)\|^2 \\ &+ \|\Delta\{\hat{\mathbf{b}}_i + \sigma(\hat{\mathbf{Z}}_i^T\hat{\mathbf{Z}}_i + \Delta^T\Delta)^{-1/2}\mathbf{z}_j\}\|^2\big] + \|\mathbf{z}_j\|^2/2\Big\} \prod_{k=1}^{m} w_{j_k}\end{aligned}$$

- Although the adaptive Gaussian quadrature approach can, in principle, be extended to the multilevel case, it is much more computationally demanding than the LME approximation, and only the 1-level case is practical.

- In fact, even in the one-level case, models with $q > 2$ random effects (e.g., our indomethacin example where the final model had $q = 3$) are very difficult to fit in PROC NLMIXED. Only one-level models with $q \leq 2$ are really feasible currently with this methodology.

- It is true, though, that the adaptive Gaussian quadrature approach is more accurate than the LME approximation, and is the preferred method for feasible problems.

Example — CO_2 Uptake in Grass

- See §8.2.2 in Pinheiro and Bates (2000) for more details on this example.

The groupedData object CO2 in the nlme library in R contains data from an experiment to investigate the effects of cold temperatures on the CO_2 uptake of the grass species, *Echinochloa crus-galli*. A total of 12 four-week-old plants, 6 from Québec and 6 from Mississippi, were divided into two groups: control plants that were kept at 26°C and chilled plants that were subject to 14 h of chilling at 7°C. After 10 h of recovery at 20°C, CO_2-uptake was measured for each plant at seven concentrations of ambient CO_2.

- See handout CO2. A plot of the data appears on p.5 of the handout. This plot suggests an asymptotic relationship between uptake and concentration.

We follow Potvin et al. (1990), who originally presented these data, in considering models based on the asymptotic regression model SSasympOff. In its fixed-effects incarnation, this model has expectation function

$$f(\boldsymbol{\theta}, x_{ij}) = \theta_1 \left[1 - \exp\{-e^{\theta_2}(x_{ij} - \theta_3)\}\right] \qquad (*)$$

where y_{ij} and x_{ij} are the CO_2-uptake and ambient CO_2 concentration, respectively, for the j^{th} observation made on the i^{th} plant.

In building a mixed-effects version of (*), we must account for plant-to-plant heterogeneity and also differences in the mean response from one experimental group to the next. Here the experimental groups are the crossing of Type of plant (Québec vs. Mississippi) with Treatment (non-chilled vs. chilled).

It is natural to include random effects in one or more of the model parameters $\theta_1, \theta_2, \theta_3$ in (*) to account for plant-to-plant heterogeneity. But we will also want to include fixed-effects in our model to account for group-to-group differences.

However, there are several decisions to make: Which parameters should be modelled with random-effects? In which parameters should we include fixed-effects to account for group differences. How should we test for significance of fixed-effects? How should we test for significance of random effects? and in what order?

There is not one uniquely correct strategy of model building for NLMMs. We will take the strategy advocated by Pinheiro and Bates (2000, §8.2):

1. Start with the basic NLMM involving no covariates (no fixed-effects for group differences) with all parameters mixed.

2. Drop any obviously non-significant random effects from the model. (By non-significant we mean random effects whose variance components are non-significant).

   – Testing for significance of random effects is complicated by the non-chisquare limiting distribution of the LRT. Using information criteria for nested models is more appropriate, although LRTs are informative if we keep in mind that chi-square-based $p-$values (e.g., as given by the anova() function in R) are overestimated.

3. Plot predicted random effects versus covariate-values (potential explanatory variable values) to determine which parameters' variability is accounted for by the covariate(s) and thus which parameter should be modelled with covariates and additional fixed-effects.

   – Significance of new fixed-effects is assessed with Wald-type tests.

4. After covariates are added, we reconsider the random-effects structure. It may be possible to remove random effects (or, possibly but unlikely, necessary to add random effects) once covariates have been added to the model.

   – Again significance of random effects is assessed with information criteria and LRTs for nested models.

- In the CO2 handout, we first fit the SSasympOff model (*) to the data from each plant separately using the nlsList() function. An intervals() plot of the plant-specific parameters would be helpful in deciding which parameters should be mixed, but instead we go ahead and fit the model with all parameters mixed:

$$y_{ij} = \theta_{1i}\left[1 - \exp\{-e^{\theta_{2i}}(x_{ij} - \theta_{3i})\}\right] + e_{ij}$$

$$\boldsymbol{\theta}_i = \begin{pmatrix} \theta_{1i} \\ \theta_{2i} \\ \theta_{3i} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{pmatrix} = \boldsymbol{\beta} + \mathbf{b}_i$$

$$\{\mathbf{b}_i\} \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Psi}), \quad \{e_{ij}\} \overset{iid}{\sim} N(0, \sigma^2)$$

- The above model is fit as m1CO2.nlme and can be fit directly from the nlsList() model m1CO2.lis. A summary of m1CO2.nlme reveals that b_{1i} and b_{3i} are almost perfectly correlated. This suggests that the model is overparameterized, and we try dropping b_{3i} from the model.

- The resulting model, m2CO2.nlme fits better than m1CO2.nlme according to both information criteria. Model m2CO2.nlme has a reasonable looking residual plot. We now go on and consider the explanatory factors Type and Treatment.

- We plot predicted random effects from m2CO2.nlme against the experimental groups given by the crossing of Type and Treatment on p.7. It appears that the random effects corresponding to θ_1 and those corresponding to θ_2 both change across experimental groups in a way that suggests an interaction between Type and Treatment. Therefore, next we consider model m3CO2.nlme, in which

$$\begin{aligned} \theta_{1i} &= \beta_{10} + \beta_{11}x_{1i} + \beta_{12}x_{2i} + \beta_{13}x_{1i}x_{2i} + b_{1i} \\ \theta_{2i} &= \beta_{20} + \beta_{21}x_{1i} + \beta_{22}x_{2i} + \beta_{23}x_{1i}x_{2i} + b_{2i} \\ \text{and} \quad \theta_{3i} &= \beta_3 \end{aligned}$$

  where

$$x_{1i} = \begin{cases} 0, & \text{if Type=Québec} \\ 1, & \text{if Type=Mississippi} \end{cases}$$

$$x_{2i} = \begin{cases} 0, & \text{if Treatment=non-chilled} \\ 1, & \text{if Treatment=chilled} \end{cases}$$

- This model is fit as m3CO2.nlme. Significance of $\beta_{11}, \beta_{12}, \beta_{13}$ (the main effect of Type, main effect of Treatment, and interaction in the asymptote parameter) is assessed with a Wald test via anova(m3CO2.nlme,Terms=2:4). Similarly we test the significance of $\beta_{21}, \beta_{22}, \beta_{23}$ (the main effects and interaction in the rate parameter) via anova(m3CO2.nlme,Terms=6:8). In both cases, these terms are significant and we retain them in the model.

- Finally, we consider removing one or more of the the random effects b_{1i}, b_{2i} now that explanatory variables have been incorporated into θ_{1i}, θ_{2i}. A comparison of the estimated standard deviation of b_{1i} ($\sqrt{\hat{\psi}_{11}} = 2.35$) versus $|\hat{\beta}_{10}| = 32.34$ and a comparison of the estimated standard deviation of b_{2i} ($\sqrt{\hat{\psi}_{22}} = .080$) versus $|\hat{\beta}_{20}| = 4.51$ reveals that the variance component associated with b_{2i} is relatively smaller than that associated with b_{1i} Therefore, we consider dropping b_{2i} from the model. This yields model m4CO2.nlme, which fits better than m3CO2.nlme according to the information criteria.

- We also consider dropping b_{1i} from the model (see m5CO2.gnls), but this change results in a significant reduction in the quality of the fit.

- So, the final model for these data is

$$y_{ij} = \theta_{1i}\left[1 - \exp\{-e^{\theta_{2i}}(x_{ij} - \theta_{3i})\}\right] + e_{ij}$$

$$\boldsymbol{\theta}_i = \begin{pmatrix} \theta_{1i} \\ \theta_{2i} \\ \theta_{3i} \end{pmatrix} = \begin{pmatrix} \beta_{10} + \beta_{11}x_{1i} + \beta_{12}x_{2i} + \beta_{13}x_{1i}x_{2i} + b_{1i} \\ \beta_{20} + \beta_{21}x_{1i} + \beta_{22}x_{2i} + \beta_{23}x_{1i}x_{2i} \\ \beta_3 \end{pmatrix}$$

$$\{\mathbf{b}_i\} \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Psi}), \quad \{e_{ij}\} \overset{iid}{\sim} N(0, \sigma^2)$$

- A plot of the observed and predicted responses for each plant is included on the final page of the handout. As you can see, the model appears to fit the data well.

- To illustrate the use of PROC NLMIXED and its implementation of adaptive Gaussian quadrature. See CO2.sas and CO2.lst. In these programs we refit the final model above. As you'll see, the results are quite similar but not identical to those from nlme() using the LME approximation.

An example of a multilevel NLMM fit in nlme is given by Pinheiro and Bates (2000, §8.2.3).

Restricted Maximum Likelihood

In the linear mixed effects model, ML estimators of variance components are generally not preferred because they are biased. To take a simple example, consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

where $\mathbf{X}$ is an $n \times p$ full rank model matrix.

The ML estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2,$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the MLE/LSE of $\boldsymbol{\beta}$ and $\mathbf{x}_i^T$ is the i^{th} row of $\mathbf{X}$.

It is well known and easy to show that $\hat{\sigma}^2$ is biased. A bias-corrected and generally preferred estimator is the mean squared error:

$$s^2 = \frac{1}{n-p}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2 = \frac{n}{n-p}\hat{\sigma}^2.$$

- s^2 is unbiased.
- the divisor $n-p$ in s^2 is df_E, the degrees of freedom due to error in the model, which is also equal to $\text{df}_T - \text{df}_{\text{Reg}} = n-p$, the total degrees of freedom (n) minus the degrees of freedom due to estimating $\boldsymbol{\beta}$ (p). Since the divisor in s^2 is $\text{df}_T - \text{df}_{\text{Reg}}$ rather than just df_T as used in $\hat{\sigma}^2$, it is often said that s^2 accounts for degrees of freedom lost from having to estimate the regression parameter $\boldsymbol{\beta}$.

Another example, a simple linear mixed model. Consider the one-way anova model with random effects:

$$y_{ij} = \mu + b_i + e_{ij}, \quad i = 1, \ldots, a, j = 1, \ldots, m.$$

Here, μ is a (fixed) mean, $b_1, \ldots, b_a \overset{iid}{\sim} N(0, \sigma_b^2)$ are random effects, and the e_{ij}'s are i.i.d. $N(0, \sigma^2)$ error terms assumed independent of the b_i's.

- Such a model might be appropriate for an example in which we are trying to estimate the calcium content of the leaves of a certain plant and we have calcium measurements on several (m) leaves from each of several (a) plants. In such a situation, we're interested in μ, the mean calcium concentration, but the data are grouped by plant, so we want plant-specific random effects b_i in the model to account for heterogeneity from plant-to-plant and correlation in the observations taken from the same plant.

In this model, the total variance of a response y_{ij} is

$$\text{var}(y_{ij}) = \sigma_b^2 + \sigma^2.$$

Therefore, σ_b^2 and σ^2 are called variance components.

The ANOVA table for this model is as follows:

Source of Variation	Sum of Squares	df	Mean Squares	F
Groups	SS_{Trt}	$a-1$	$MS_{Grps} = SS_{Grps}/\text{df}_{Grps}$	$\frac{MS_{Grps}}{MS_E}$
Error	SS_E	$n-a$	$MS_E = SS_E/\text{df}_E$	
Total	SS_T	$n-1$		

The MLEs of the variance components are given by

$$\hat{\sigma}_b^2 = \frac{1}{n}\left(\frac{a-1}{a}MS_{Grps} - MS_E\right) \quad \text{and} \quad \hat{\sigma}^2 = MS_E, \qquad \text{if } \hat{\sigma}_b^2 \geq 0$$

$$\hat{\sigma}_b^2 = 0 \quad \text{and} \quad \hat{\sigma}^2 = \frac{SS_T}{n}, \qquad \text{if } \hat{\sigma}_b^2 < 0$$

These are not the usual variance component estimators taught in courses like STAT 8200. Instead, the usual estimators are derived from the ANOVA table by equating mean squares to their expected values and solving for σ^2 and σ_b^2. This leads to the "ANOVA" estimators

$$\tilde{\sigma}_b^2 = \frac{1}{n}(MS_{Grps} - MS_E) \quad \text{and} \quad \tilde{\sigma}^2 = MS_E, \qquad \text{if } \tilde{\sigma}_b^2 \geq 0$$

$$\tilde{\sigma}_b^2 = 0 \quad \text{and} \quad \tilde{\sigma}^2 = \frac{SS_T}{n-1}, \qquad \text{if } \tilde{\sigma}_b^2 < 0$$

- The ANOVA estimators, like s^2, the MSE in the previous example, are less biased than the corresponding ML estimators because they account for degrees of freedom lost in having to estimate fixed effects are taken into account.

Both of these examples of "bias-corrected" alternatives to ML estimators of variance components are special case of what are known as **restricted maximum likelihood (REML)** estimators. In REML, the goal is to produce better estimators of variance components by constructing an objective function that does not involve the fixed effects. That is, REML estimators maximize a likelihood-like function in which the nuisance parameter $\boldsymbol{\beta}$ has been eliminated, or concentrated out of the likelihood.

The linear mixed model for grouped (two-level) data with spherical errors takes the form

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad i = 1, \ldots, M,$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_M \overset{iid}{\sim} N_{n_i}(0, \sigma^2\mathbf{I})$, $\mathbf{b}_1, \ldots, \mathbf{b}_M \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Psi})$, where the $\mathbf{e}_i$'s and $\mathbf{b}_i$'s are independent of one another.

This model can be written more succinctly as

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

where

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_M \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_M \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_M \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_M \end{pmatrix}$$

and $\mathbf{Z} = \text{blkdiag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_M)$. Now $\mathbf{b} \sim N(\mathbf{0}, \text{blkdiag}(\boldsymbol{\Psi}, \ldots, \boldsymbol{\Psi}))$ and $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$. For simplicity, assume that $\mathbf{X}$ is $n \times p$ of rank p.

Note that this model implies

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\Psi}, \sigma^2)), \quad \mathbf{V}(\boldsymbol{\Psi}, \sigma^2) = \text{blkdiag}(\mathbf{V}_1(\boldsymbol{\Psi}, \sigma^2), \ldots, \mathbf{V}_M(\boldsymbol{\Psi}, \sigma^2)),$$

where $\mathbf{V}_i(\boldsymbol{\Psi}, \sigma^2) = \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^T + \sigma^2\mathbf{I}$.

- Note that the distribution of $\mathbf{y}$ depends upon the fixed effects $\boldsymbol{\beta}$.

To eliminate $\boldsymbol{\beta}$ from the objective function, in REML we work not with the likelihood function corresponding to the joint density of $\mathbf{y}$'s, but instead we work with the joint density of a set of linearly independent *error contrasts* of $\mathbf{y}$'s.

- A linear combination $\mathbf{a}^T\mathbf{y}$ is called an error contrast if $\text{E}(\mathbf{a}^T\mathbf{y}) = 0$ for all $\boldsymbol{\beta}$.

Suppose we can find $n-p$ linearly independent error contrasts

$$\begin{aligned} w_1 &= \mathbf{a}_1^T \mathbf{y} \\ w_2 &= \mathbf{a}_2^T \mathbf{y} \\ &\vdots \\ w_{n-p} &= \mathbf{a}_{n-p}^T \mathbf{y} \end{aligned}$$

or

$$\mathbf{w} = \mathbf{A}^T \mathbf{y},$$

where $\mathbf{A}$ has columns $\mathbf{a}_1, \ldots, \mathbf{a}_{n-p}$.

Then

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\Psi}, \sigma^2))$$

implies

$$\mathbf{w} \sim N_{n-p}(\mathbf{0}, \mathbf{A}^T \mathbf{V}(\boldsymbol{\Psi}, \sigma^2)\mathbf{A}).$$

- It is clear now that the loglikelihood of $\mathbf{w}$ does not depend upon $\boldsymbol{\beta}$.

The log density of $\mathbf{w}$ is taken as $\ell^R(\boldsymbol{\Psi}, \sigma^2; \mathbf{y})$, the restricted loglikelihood of the variance parameters $\boldsymbol{\Psi}$ and σ^2, based on $\mathbf{y}$:

$$\ell^R(\boldsymbol{\Psi}, \sigma^2; \mathbf{y}) = \sum_{i=1}^{M} \Big[-\frac{1}{2} \log |\mathbf{A}^T \mathbf{V}(\boldsymbol{\Psi}, \sigma^2)\mathbf{A}| - \frac{1}{2} \mathbf{w}^T \{\mathbf{A}^T \mathbf{V}(\boldsymbol{\Psi}, \sigma^2)\mathbf{A}\}^{-1} \mathbf{w} + \text{constant} \Big].$$

In the REML approach, the loglikelihood is replaced by the "restricted loglikelihood" which is defined as the loglikelihood based upon a set of linear independent error contrasts in the original response. It can be shown that this restricted loglikelihood does not depend upon which set of error contrasts is chosen. All choices lead to the same restricted loglikelihood, which, ignoring constant terms, is given by

$$\ell^R(\boldsymbol{\Psi}, \sigma^2; \mathbf{y}) = -\frac{1}{2}\sum_{i=1}^{M}\{\log|\mathbf{V}_i(\boldsymbol{\Psi}, \sigma^2)| + \log|\mathbf{X}_i^T\mathbf{V}_i^{-1}(\boldsymbol{\Psi}, \sigma^2)\mathbf{X}_i| + [\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}(\boldsymbol{\Psi}, \sigma^2)]^T\mathbf{V}_i^{-1}(\boldsymbol{\Psi}, \sigma^2)[\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}(\boldsymbol{\Psi}, \sigma^2)]\}$$

where $\hat{\boldsymbol{\beta}}(\boldsymbol{\Psi}, \sigma^2) = \{\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\Psi}, \sigma^2)\mathbf{X}\}^{-1}\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\Psi}, \sigma^2)\mathbf{y}$.

In contrast, the profile likelihood for $\boldsymbol{\Psi}, \sigma^2$ corresponding to ML estimation is

$$p\ell(\boldsymbol{\Psi}, \sigma^2; \mathbf{y}) = -\frac{1}{2}\sum_{i=1}^{M}\{\log|\mathbf{V}_i(\boldsymbol{\Psi}, \sigma^2)| + [\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}(\boldsymbol{\Psi}, \sigma^2)]^T\mathbf{V}_i^{-1}(\boldsymbol{\Psi}, \sigma^2)[\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}(\boldsymbol{\Psi}, \sigma^2)]\}$$

That is, with respect to the variance parameters $\boldsymbol{\Psi}$ and σ^2, the REML objective function differs from that of ML as follows

$$\ell^R(\boldsymbol{\Psi}, \sigma^2; \mathbf{y}) = p\ell(\boldsymbol{\Psi}, \sigma^2; \mathbf{y}) - \frac{1}{2}\sum_{i=1}^{M}\log|\mathbf{X}_i^T\mathbf{V}_i^{-1}(\boldsymbol{\Psi}, \sigma^2)\mathbf{X}_i|.$$

- Consequently, REML is sometimes called a *penalized likelihood* method.

REML estimates of $\boldsymbol{\Psi}, \sigma^2$ are obtained by solving the equations obtained by differentiating $\ell^R(\boldsymbol{\Psi}, \sigma^2; \mathbf{y})$ with respect to these parameters and setting equal to zero.

- REML estimators are not, in general, unbiased, but they are typically less biased the ML estimators and are preferred in most contexts.

- An important caveat concerning REML estimation is that the objective function is not a true loglikelihood, and cannot be treated as such for all aspects of statistical inference. In particular, models with different fixed effects specifications cannot be compared via restricted likelihood ratio type tests or via model selction criteria such as AIC and BIC in which the loglikelihood has been replaced by the restricted loglikelihood.

REML estimation doesn't easily generalize to a nonlinear model context, because the idea of using an error contrast to eliminate the fixed effects only works when the fixed effects enter into the model in a linear way.

However, the LME approach to fitting the NLMM does lend itself nicely to an approximate REML-type procedure. Recall that the LME approach consisted of two steps:

1. A PNLS step in which we minimize the penalized nonlinear least squares criterion given on the top of p. 249 with respect to $\boldsymbol{\beta}$ and $\hat{\mathbf{b}}_i$, and

2. An LME step in which we maximize the function

$$\ell_{\mathrm{LME}}(\hat{\boldsymbol{\beta}}(\Delta), \sigma^2, \Delta; \mathbf{y}) = -\frac{1}{2} \sum_{i=1}^{M} \Big[n_i \log(2\pi\sigma^2) + \log|\Sigma_i(\Delta)| + \{\mathbf{w}_i - \hat{\mathbf{X}}_i\boldsymbol{\beta}\}^T \Sigma_i(\Delta)^{-1} \{\mathbf{w}_i - \hat{\mathbf{X}}_i\boldsymbol{\beta}\} \Big]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\Delta,\sigma^2)}, \qquad (*)$$

   where

$$\hat{\mathbf{X}}_i = \frac{\partial \mathbf{f}_i(\boldsymbol{\beta}, \hat{\mathbf{b}}_i)}{\partial \boldsymbol{\beta}^T}, \quad \text{and} \quad \mathbf{w}_i = \mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \hat{\mathbf{b}}_i) + \hat{\mathbf{X}}_i\boldsymbol{\beta} + \hat{\mathbf{Z}}_i\hat{\mathbf{b}}_i.$$

- Recall that step 2 corresponds to fitting a linear mixed effects model with ML estimation.

The approximate REML version of the LME approach to fitting an NLMM just fits the linear mixed effects model in step two with REML rather than ML. That is, the objective function (*) is replaced by

$$\ell^R_{\text{LME}}(\sigma^2, \Delta; \mathbf{y}) = \ell_{\text{LME}}(\hat{\boldsymbol{\beta}}(\Delta), \sigma^2, \Delta; \mathbf{y}) \underbrace{-\frac{1}{2}\sum_{i=1}^{M} \log \left| \sigma^{-2} \hat{\mathbf{X}}_i^T \Sigma_i(\Delta)^{-1} \hat{\mathbf{X}}_i \right|}_{\text{REML penalty term}}.$$

- The idea here is that the LME approximate ML procedure consists of iteratively fitting a LMM with ML, so the LME approximate REML procedure is done by iteratively fitting the LMM with REML.

Inference on Fixed Effects Using the LME Approximation:

- We present these results for the two-level model only.

Under the LME approximation, the distribution of the approximate (restricted) maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ of the fixed effects is

$$\hat{\boldsymbol{\beta}} \dot{\sim} N\left(\boldsymbol{\beta}, \sigma^2 \left[\sum_{i=1}^{M} \hat{\mathbf{X}}_i^T \Sigma_i^{-1} \hat{\mathbf{X}}_i\right]^{-1}\right), \qquad (\clubsuit)$$

where $\Sigma_i = \hat{\mathbf{Z}}_i \Delta^{-1} \Delta^{-T} \hat{\mathbf{Z}}_i^T + \mathbf{I}$.

- In practice $\text{var}(\hat{\boldsymbol{\beta}})$ is estimated with

$$\hat{\text{var}}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left[\sum_{i=1}^{M} \hat{\mathbf{X}}_i^T \Sigma_i^{-1} \hat{\mathbf{X}}_i\right]^{-1} \Bigg|_{\Delta=\hat{\Delta}, \sigma^2=\hat{\sigma}^2},$$

  where $\hat{\Delta}$ and $\hat{\sigma}^2$ are the ML or REML estimates of these parameters.

Standard errors of $\hat{\beta}_i$, the j^{th} component of $\hat{\boldsymbol{\beta}}$ are obtained as the square root of the j^{th} diagonal element of $\hat{\text{var}}(\hat{\boldsymbol{\beta}})$.

The distributional result (♣) suggests that approximate Wald tests can be used for inference on $\boldsymbol{\beta}$. In particular, an approximate α-level test of a hypothesis of the form $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ where $\mathbf{A}$ is $k \times \dim(\boldsymbol{\beta})$ reject H_0 if

$$(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})^T \{\mathbf{A}[\hat{\text{var}}(\hat{\boldsymbol{\beta}})]^{-1}\mathbf{A}^T\}^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}) > \chi^2_{1-\alpha}(k)$$

where $\chi^2_{1-\alpha}(k)$ is the upper α^{th} critical value of a chi-square distribution on k df.

As a special case of this result, an approximate z test of $H_0 : \hat{\beta}_j = 0$ versus $H_0 : \hat{\beta}_j \neq 0$ rejects H_0 if

$$\frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)} > z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of a standard normal distribution. In addition, an approximate $100(1 - \alpha)\%$ CI for β_j is given by

$$\hat{\beta}_j \pm z_{1-\alpha/2}\text{s.e.}(\hat{\beta}_j).$$

- These Wald-based inferences are "approximately asymptotic". That is, their validity depends upon the accuracy of the LME approximation as an approximate version of ML (or REML) estimation, *and* on the usual asymptotic arguments for ML estimation that justify Wald-based inference as approximately valid in finite samples.

- Vonesh and Carter (1992) and Pinheiro and Bates (2000) have suggested that more accurate inferences can be accomplished by using F and t reference distributions in place of the χ^2 and z distributions given above.

- The idea here is to account for the fact that we're using $\hat{\text{var}}(\hat{\boldsymbol{\beta}})$ instead of $\text{var}(\hat{\boldsymbol{\beta}})$ to form our test statistics, and this substitution should introduce additional error into the sampling distributions of the test statistics.

An approximate F test of $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ is based on the test statistic

$$\frac{(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})^T \{\mathbf{A} \left[\hat{\text{var}}(\hat{\boldsymbol{\beta}})\right]^{-1} \mathbf{A}^T\}^{-1} (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})}{k} \dot{\sim} F(k, \nu).$$

In addition, $H_0 : \hat{\beta}_j = 0$ can be tested via the test statistic

$$\frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)} \dot{\sim} t(\nu).$$

What is the appropriate choice for the denominator d.f. ν in these tests?

Unfortunately, there is not a definitive answer to that question yet.

- For the two level model (i.e., single level of clustering), Vonesh and Carter (1992) sugggested $\nu = M - \dim(\boldsymbol{\beta})$.
- For the two level model SAS' PROC NLMIXED uses $M - q$, where q is the dimension of the random effects vector $\mathbf{b}_i$.
- Pinheiro and Bates (2000, p.322) suggest that the same procedure as used to compute denominator dffor tests in the LMM be used in the NLMM. That procedure is described on p.91 of their book. However, in the nlme() function, the procedure from the LMM is not followed exactly. In the two-level model, the denominator d.f. computed by nlme() will always be $n - M - (p - 1)$.
- None of these choices for ν can be justified rigorously, and it is not at all clear which is the best approach to use in practice. The Wald based inferences are perhaps easiest to justify theoretically, but they will tend to be liberal (reject often, tight intervals) relative to the other methods.