Example: Dental Growth Curves – Initial Plots

We now present some initial plots for the dental data – should not be viewed as comprehensive.

- Initial plots: QQ plots of LS estimates, both univariate (Figure 12) and bivariate (Figure 13).
- Estimates of σ_ϵ: 0.97, 0.59, 0.95, 0.30, 0.58, 0.43, 0.47, 0.19, 0.58, 0.85, 0.89 – not a great deal of variability, so common variance assumption seems reasonable.
- No apparent mean-variance relationship (Figure 14).
- Figure 15 shows that there are clear differences in intercepts, and some variability in slopes.

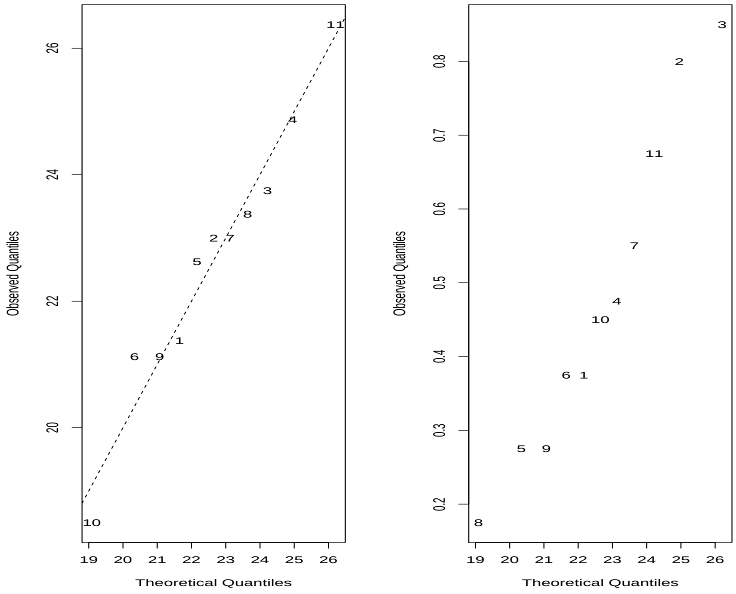

Figure 12: QQ plots of LS estimates: $\widehat{\beta}_0$ (left), $\widehat{\beta}_1$ (right).

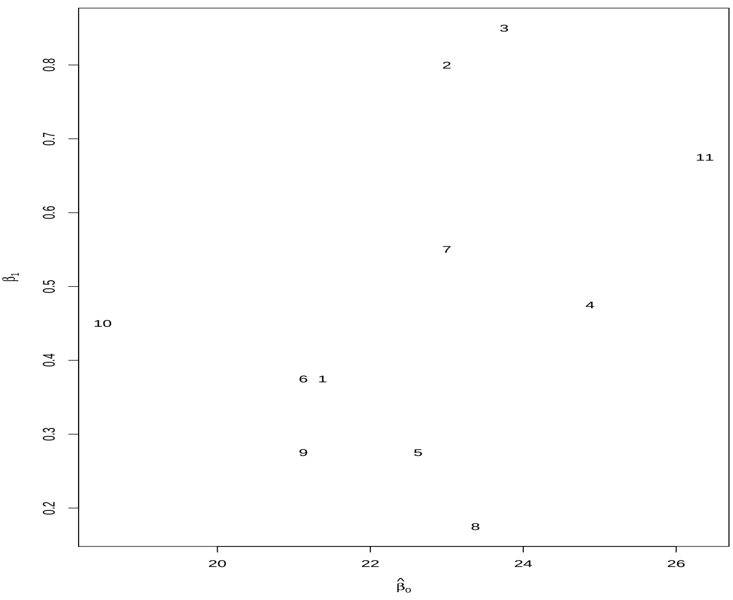

Figure 13: Bivariate plot of LS estimates.

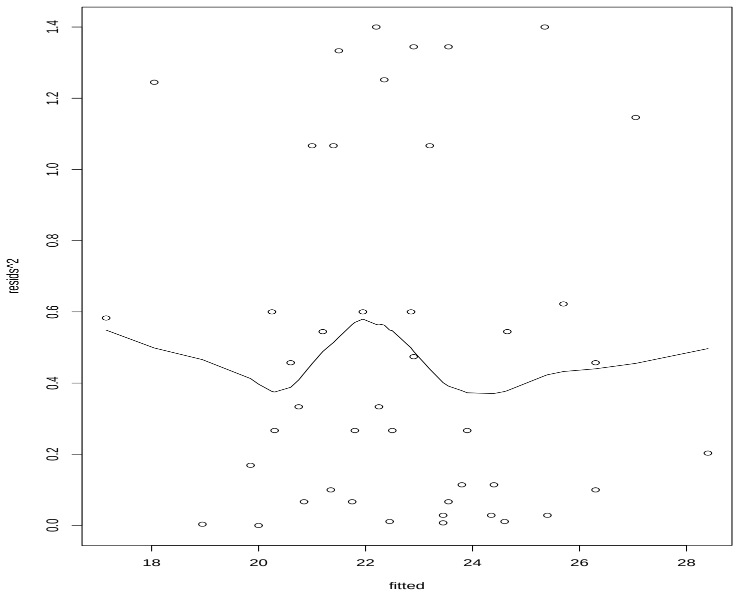

Figure 14: LS residuals versus fitted values

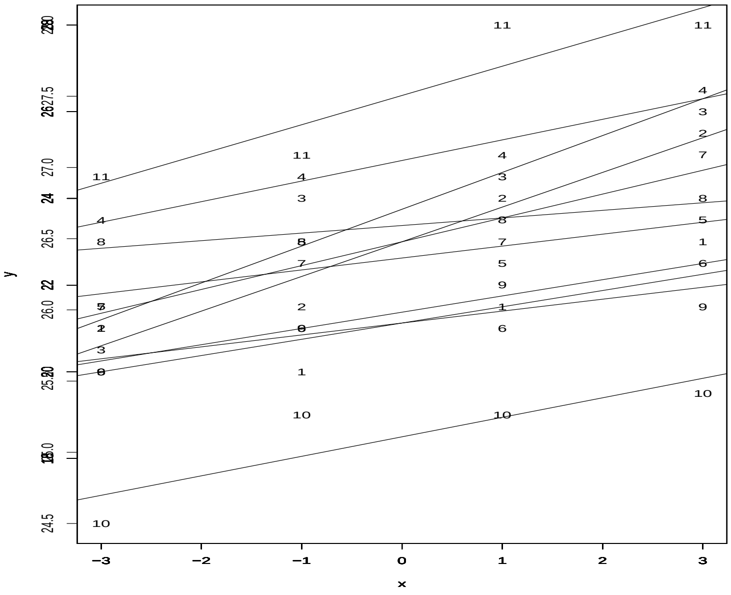

Figure 15: Fitted curves for all data.

Example: Dental Growth Curves – Initial Plots

We now present some residual plots for the dental data.

```
> cnt1 <- rep(4,11); cnt2 <- 1:11
> lme1 <- lme(distance ~ I(age-11), data = Orthgirl, random = ~1 | Subject )
> lmeres1ind <- resid( lme1, level = 1, resType="n") # ind-level resids
> lmefit1ind <- fitted( lme1 )
> plot(I(Orthgirl$age-11),lmeres1,xlab="Centered age",ylab="LME indiv residuals",
    ylim=c(-max(abs(range(lmeres1))),max(abs(range(lmeres1)))),type="n")
> text(I(Orthgirl$age-11),lmeres1,labels=c(rep(cnt2,cnt1))); abline(0,0)
> lines(lowess(I(Orthgirl$age-11),lmeres1))
> qqnorm(lmeres1,main="")
> plot(lmefit1ind,lmeres1ind^2)
> lines(lowess(lmefit1ind,lmeres1ind^2))
```

Figure 16 shows no syystematic deviations between residuals with time. Figure 17 that normality reasonable, and Figure 18 that there is no mean variance relationship.

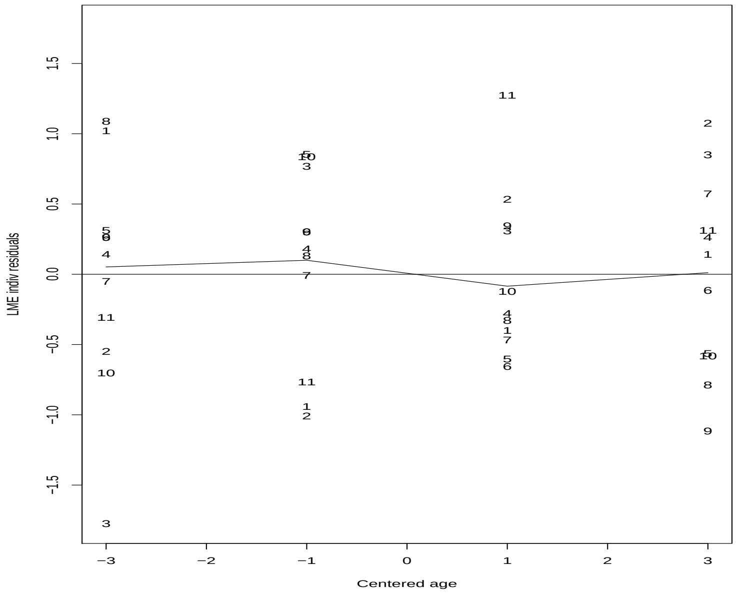

Figure 16: LME normalized residuals versus time.

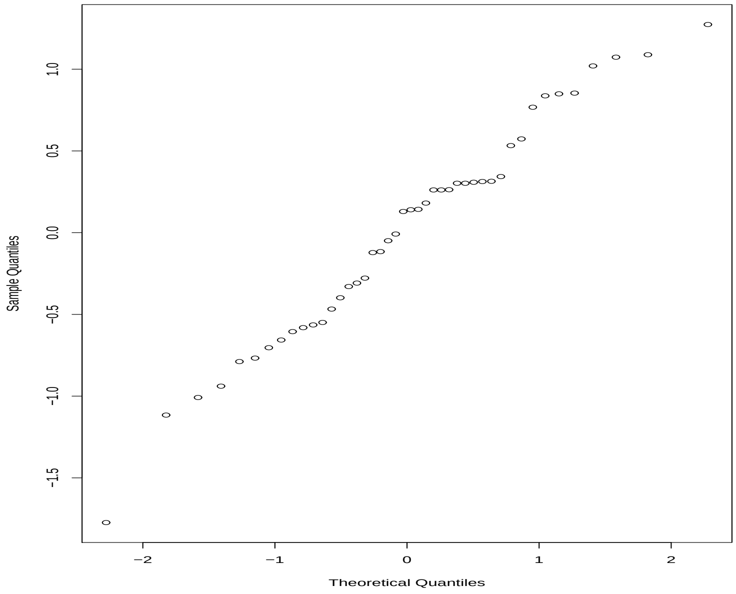

Figure 17: QQ plot of LME normalized residuals.

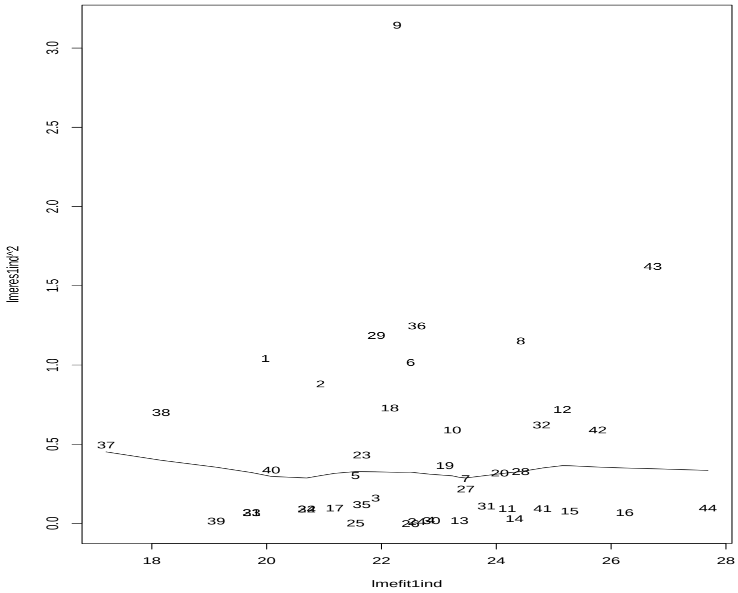

Figure 18: LME normalized residuals versus fitted values.

Assessing Adequacy of the Temporal Covariance Structure

An informal method for assessing whether there is residual temporal dependence is to plot residuals versus time, we now consider more formal tools such as the correlgram and the variogram.

We begin with some definitions.

Consider a stochastic process $Y(t)$ and let

$$\gamma(t,s) = \text{cov}\{Y(t), Y(s)\} = \text{E}[\{Y(t) - \mu(t)\}\{Y(s) - \mu(s)\}],$$

denote the *autocovariance function* of $Y(t)$.

The term *serial dependence* signifies that there is dependence between $Y(t)$ and $Y(s)$ for at least some pairs (s,t) with $s \neq t$.

We write

$$Y(t) = \mu(t) + e(t),$$

where $\mu(t)$ is the deterministic trend component.

Definition: A process $e(t)$ is second-order stationary if $\text{E}[e(t)]$ is constant, for all t, and $\gamma(t,s)$ depends only on $|t-s|$. For a residual process any non-zero constant has been absorbed into $\mu(t)$.

Example: The simplest example of a stationary random sequence is *white noise* which consists of a sequence of mutually independent random variables, each with mean 0 and finite variance σ^2.

There is a fundamental difficulty with trying to decompose $Y(t)$ into the trend and the stochastic component in a single series because the two are unidentifiable without further assumptions.

Is it serial dependence in the residuals, or a high-order polynomial trend for example?

The Autocorrelation Function

For a second-order stationary random process, the autocovariance function is

$$\text{cov}\{Y(t), Y(t+u)\} = \text{cov}\{e(t), e(t+u)\},$$

so that $C(0)$ is the variance of $Y(t)$ for all t.

The autocorrelation function is defined as

$$\rho(u) = \frac{C(u)}{C(0)}.$$

For equally-spaced data we could fit a model and then examine the autocorrelation function (ACF) of the residuals,

$$e_t = \frac{y_t - \widehat{y}_t}{\widehat{\text{var}}(Y_t)^{1/2}}.$$

Consider a stochastic process $e(t)$, and realizations e_t, $t = 1, ..., n$. The *emprical* autocorrelation is defined as

$$\widehat{\rho}(u) = \widehat{\text{corr}}\{e(t), e(t+u)\} = \frac{\sum_{t=1}^{n-u} e_t e_{t+u}/(n-u)}{\sum_{t=1}^{n} e_t^2/n},$$

for $u = 0, 1, ...$

A *correlogram* plot is $\widehat{\rho}(u)$ versus u. If the residuals are a white noise process, we have the asymptotic result

$$\sqrt{n}\ e_t \ \ \rightarrow_d \ \ \text{N}(0, 1),$$

to give confidence bands $\pm 2/\sqrt{n}$.

The Variogram

For unequally-spaced data the ACF is not so convenient, unless we round the observations.

An alternative is provided by the *semi-variogram* which is defined, for a process e_t and $d \geq 0$.

$$\gamma(d) = \frac{1}{2}\text{var}\left(e_t - e_{t-d}\right) = \frac{1}{2}\text{E}\left[\left\{e_t - e_{t-d}\right\}^2\right].$$

Recall that for a second-order stationary process, $\text{E}[e_t] = \mu$ for all t and $\text{cov}(e_t, e_{t-d})$ only depends on the distance d (which implies constant variance).

A smooth process is L_2-continuous, i.e.

$$\text{E}\{(e_t - e_{t-d})^2\} \rightarrow 0$$

as $d \rightarrow 0$. For a second-order stationary smooth process

$$\begin{aligned} \gamma(d) &= \frac{1}{2}\left\{\text{E}[e_t^2] + \text{E}[e_{t-d}^2] - 2\text{E}[e_t e_{t-d}]\right\} \\ &= \sigma_e^2\{1 - \rho(d)\}, \end{aligned}$$

where $\text{var}(e) = \sigma_e^2$.

The semi-variogram is also well-defined for an *intrinsically* stationary process for which $\mathrm{E}[e_t] = \mu$ and for which

$$\mathrm{E}[(e_t - e_{t-d})^2] = 2\gamma(d).$$

As d increases then for observatons far apart in time

$$\gamma(d) \rightarrow \mathrm{var}(e_t) = \sigma_e^2,$$

which (recall) is assumed constant.

Consider measurement error, ϵ_t with $\mathrm{E}[\epsilon_t] = 0$, $\mathrm{var}(\epsilon_t) = \sigma_\epsilon^2$, and

$$Y_t = \mu_t + e_t + \epsilon_t,$$

so that we no longer have a smooth process. Then

$$\gamma(d) = \frac{1}{2}\mathrm{E}\left[\{Y_t - Y_{t-d}\}^2\right] = \sigma_e^2\{1 - \rho(d)\} + \sigma_\epsilon^2,$$

and we have a "nugget" effect σ_ϵ^2s.

The Variogram in Longitudinal Data Analysis

Define the semi-variogram of the population residuals, $e_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta}$, as

$$\gamma_i(d_{ijk}) = \frac{1}{2}\mathrm{E}\left[\{e_{ij} - e_{ik}\}^2\right],$$

for $d_{ijk} =\mid t_{ij} - t_{ik} \mid\geq 0$. We emphasize that we are examining differences on the *same* individual.

The sample semi-variogram uses the empirical halved differences between pairs of population residuals

$$v_{ijk} = \frac{1}{2}(e_{ij} - e_{ik})^2,$$

along with the spacings $u_{ijk} = t_{ij} - t_{ik}$.

With highly-irregular sampling times the variogram can be estimated from the pairs (u_{ijk}, v_{ijk}), $i = 1, ..., m$, $j < k = 1, ..., n_i$, with the resultant plot being smoothed.

The marginal distribution of each v_{ijk} is χ_1^2, and this large variability can make the variogram difficult to interpret.

The total variance is estimated as the average of $\frac{1}{2}(e_{ij} - e_{lk})^2$, for $i \neq l$, since

$$\frac{1}{2}\text{E}\left[(e_{ij} - e_{lk})^2\right] = \frac{1}{2}\left\{\text{E}[e_{ij}^2] + \text{E}[e_{lk}^2]\right\} = \sigma^2,$$

assuming that observations on different individuals are independent (and the variance is constant over time, and for different individuals).

Consider the interpretation of the variogram for the model

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i + \delta_{ij} + \epsilon_{ij},$$

where $b_i \sim_{ind} N(0, \sigma_0^2)$ (note, univariate), $\epsilon_{ij} \sim_{ind} N(0, \sigma_\epsilon^2)$, and δ_{ij} represent error terms with serial dependence.

A simple and commonly-used form for serial dependence is the AR(1) model given by

$$\text{cov}(\delta_{ij}, \delta_{ik}) = \sigma_\delta^2 \rho^{|t_{ij} - t_{ik}|}.$$

Under this model

$$\text{var}(Y_{ij}|\boldsymbol{\beta}) = \sigma^2 = \sigma_0^2 + \sigma_\delta^2 + \sigma_\epsilon^2.$$

Consider the theoretical variogram for the residuals

$$e_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta} = b_i + \delta_{ij} + \epsilon_{ij},$$

$i = 1, ..., m; j = 1, ...n_i$, with the AR(1) model.

For differences in residuals on the same individual

$$e_{ij} - e_{ik} = b_i + \delta_{ij} + \epsilon_{ij} - b_i - \delta_{ik} - \epsilon_{ik} = \delta_{ij} + \epsilon_{ij} - \delta_{ik} - \epsilon_{ik},$$

and so

$$\gamma_i(d_{ijk}) = \frac{1}{2}\text{E}\left[(e_{ij} - e_{ik})^2\right] = \sigma_\delta^2(1 - \rho^{d_{ijk}}) + \sigma_\epsilon^2. \qquad (42)$$

As $d_{ijk} \rightarrow 0$, $\gamma_i(d_{ijk}) \rightarrow \sigma_\epsilon^2$ and b_i is the mean of e_{ij} and so its variance does not appear in (42).

Figure 19 shows the theoretical semi-variogram under this model and for the population residuals.

The variogram is limited in its use for *population* residuals for the LMEM, as we now illustrate.

Consider, the mixed effects model with random intercepts and independent random slopes:

$$b_{i0} \sim N(0, \upsilon_{00}^2), \quad b_{i1} \sim N(0, \upsilon_{11}^2)$$

leads to non-constant marginal variance

$$\text{var}(Y_{ij}|\boldsymbol{\beta}) = \upsilon_{00}^2 + 2\upsilon_{11}^2 t_{ij}^2,$$

so that we would not want to look at a variogram of population residuals because we do not have second-order stationarity. However, we could look at individual residuals after the random intercepts and slopes model has been fitted.

In my experience the variogram is often dominated by sampling variability (and there can be strong dependence in the plot since each residual contributes many points).

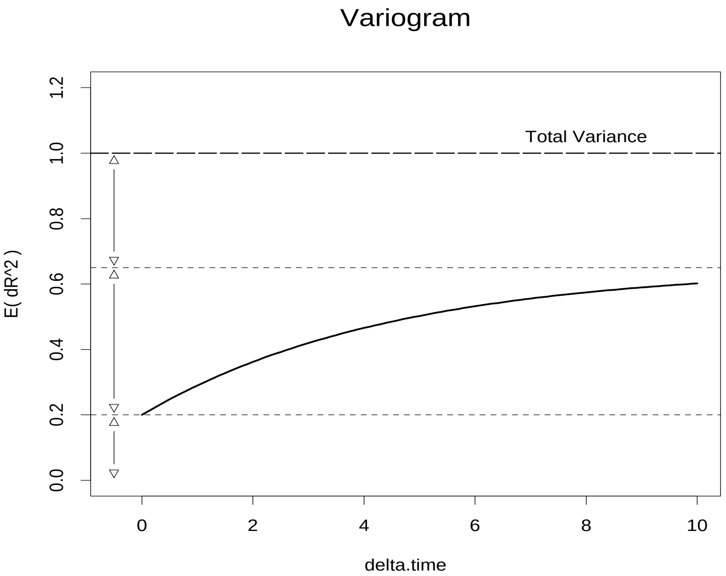

Figure 19: Theoretical variogram for a model with a random intercept, serial correlation, and measurement error.

Example: Air Pollution Data

We illustrate the correlogram and variogram for the air pollution data.

We fit a Poisson log-linear regression model in PM_{10} and ozone.

In Figure 20 we clearly see strong dependence in the Pearson residuals, hence the quasi-likelihood standard errors quoted earlier will be wrong.

The dependence is confirmed by the dependence in the variogram in Figure 21. In the left-hand panel we have only plotted 1000 of the 53301 (327 × 326/2) points.

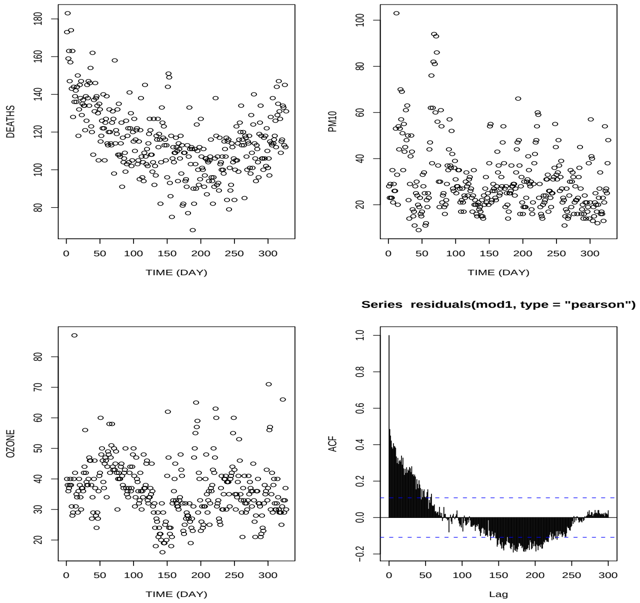

Figure 20: Time series plots and correlogram of residuals for air pollution data.

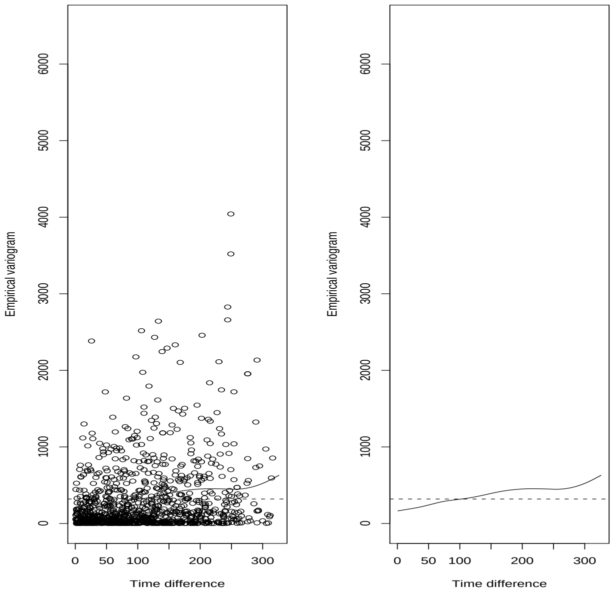

Figure 21: Variogram of residuals for air pollution data.

CHAPTER 9: GENERAL REGRESSION MODELS

We begin by considering the class of *non-linear mixed effects models* (NLMEMs) before turning to *Generalized Linear Mixed Models* (GLMMs).

In this chapter we will again consider both a *conditional* approach to modeling, via the introduction of random effects, and a *marginal* approach using GEEs. Likelihood and Bayesian methods will be used for inference in the conditional approach.

Non-linear Mixed Effects Models

Example: Pharmacokinetics of Indomethacin

Six human volunteers received bolus intravenous doses (of the same size) of Indomethacine, and subsequently 11 blood samples were taken, and the drug concentrations recorded.

Figure 22 shows the concentration-time data – the curves follow a similar pattern but there is clearly person to person variability.

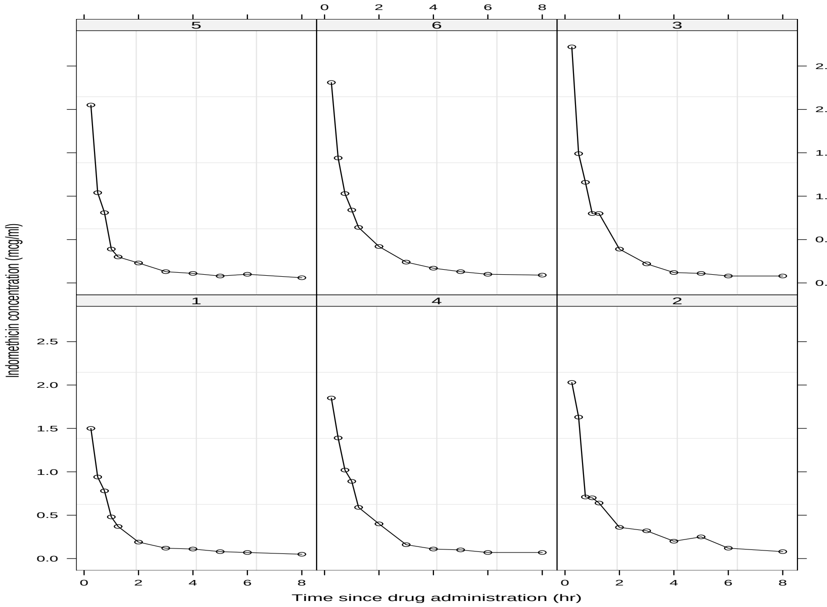

Figure 22: Concentration time data for Indomethacin.

Non-Linear Mixed Effects Models

Consider the two-stage model:

Stage 1: Response model, *conditional* on random effects:

$$\boldsymbol{y}_i = \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_{ij}) + \boldsymbol{\epsilon}_i, \tag{43}$$

where $\boldsymbol{f}_i = [f_{i1}, ..., f_{in_i}]^{\mathrm{T}}$ are a set of functions that are non-linear in the parameters $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, and $\boldsymbol{\epsilon}_i$ is an $n_i \times 1$ zero mean vector of error terms.

Stage 2: Model for random terms:

$$\begin{aligned} \mathrm{E}[\boldsymbol{\epsilon}_i] &= \boldsymbol{0}, \quad \mathrm{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{E}_i(\boldsymbol{\alpha}), \\ \mathrm{E}[\boldsymbol{b}_i] &= \boldsymbol{0}, \quad \mathrm{var}(\boldsymbol{b}_i) = \boldsymbol{D}(\boldsymbol{\alpha}), \\ \mathrm{cov}(\boldsymbol{b}_i, \boldsymbol{\epsilon}_i) &= \boldsymbol{0} \end{aligned}$$

where $\boldsymbol{\alpha}$ is the vector of variance-covariance parameters.

A common model assumes

$$\boldsymbol{\epsilon}_i \sim_{ind} N(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}_{n_i}), \quad \boldsymbol{b}_i \sim_{iid} N(\boldsymbol{0}, \boldsymbol{D}),$$

A particular form that covers a lot of longitudinal situations is to assume $f_i(\eta_{ij}, t_{ij})$ where

$$\eta_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i,$$

where

- a $(k+1) \times 1$ vector of fixed effects, $\boldsymbol{\beta}$,
- a $(q+1) \times 1$ vector of random effects, $\boldsymbol{b}_i$, with $q \leq k$.
- $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i})^{\mathrm{T}}$, the design matrix for the fixed effect with $\boldsymbol{x}_{ij} = (1, x_{ij1}, ..., x_{ijk})^{\mathrm{T}}$, and
- $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, ..., \boldsymbol{z}_{in_i})^{\mathrm{T}}$, and design matrix for the random effects with $\boldsymbol{z}_{ij} = (1, z_{ij1}, ..., z_{ijq})^{\mathrm{T}}$.

Let $\boldsymbol{\alpha}$ represent σ_ϵ^2 and the parameters of $\boldsymbol{D}$ and $N = \sum_i n_i$.

Example: Pharmacokinetics of Indomethacin

Let Y_{ij} represent the concentration of drug on individual i at time t_{ij}, $i = 1, ..., 6$, $j = 1, ..., 11$ The compartmental model that has previously been used for this drug is the two-compartment bi-exponential model:

$$\mathrm{E}[Y_{ij} \mid \beta] = A_{1i} \exp\{-\alpha_{1i} t_{ij}\} + A_{2i} \exp\{-\alpha_{2i} t_{ij}\},$$

where Y_{ij} is concentration and $A_{1i}, A_{2i}, \alpha_{1i}, \alpha_{2i} > 0$.

An obvious NLMEM would take

$$\begin{aligned} \log A_{1i} &= \beta_1 + b_{1i} \\ \log A_{2i} &= \beta_2 + b_{2i} \\ \log \alpha_{1i} &= \beta_3 + b_{3i} \\ \log \alpha_{2i} &= \beta_4 + b_{4i} \end{aligned}$$

with $\boldsymbol{b}_i = [b_{1i}, b_{2i}, b_{3i}, b_{4i}]^{\mathrm{T}} \sim_{iid} \mathrm{N}_4(\mathbf{0}, \boldsymbol{D})$.

Likelihood Inference

See Pinheiro and Bates (2000, Chapter 7).

The likelihood is, as usual, obtained by integrating out the random effects:

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\alpha}) &= (2\pi\sigma_\epsilon^2)^{-N/2}(2\pi)^{-m/2}|\boldsymbol{D}|^{-m/2} \\ &\times \prod_{i=1}^{m}\int \exp\left[-\frac{(\boldsymbol{y}_i-\boldsymbol{f}_i)^{\mathrm{T}}(\boldsymbol{y}_i-\boldsymbol{f}_i)}{2\sigma_\epsilon^2}-\frac{\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{b}_i}{2}\right]\, d\boldsymbol{b}_i. \end{aligned}$$

where $\boldsymbol{f}_i$ is made up of terms $f(\eta_{ij},t_{ij})$, $i=1,...,m$, $j=1,...,n_i$.

Difficulties

1. The first difficulty is how to calculate the required integrals, which for non-linear models are analytically intractable, recall for linear models they were available in closed form. Even the first two moments are not available in closed form in general:

$$\begin{aligned} \mathrm{E}[Y_{ij}\mid\boldsymbol{\beta},\boldsymbol{\alpha}] &= \mathrm{E}_{b_i|D}[f(\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{x}_{ij})] \neq f(\boldsymbol{\beta},\boldsymbol{0},\boldsymbol{x}_{ij}) \\ \mathrm{var}(Y_{ij}\mid\boldsymbol{\beta},\boldsymbol{\alpha}) &= \sigma_\epsilon^2+\mathrm{var}_{b_i|D}[f(\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{x}_{ij})] \\ \mathrm{cov}(Y_{ij},Y_{ij'}\mid\boldsymbol{\beta},\boldsymbol{\alpha}) &= \mathrm{cov}_{b_i|D}(f(\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{x}_{ij}),f(\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{x}_{ij'})] \\ \mathrm{cov}(Y_{ij},Y_{i'j'}\mid\boldsymbol{\beta},\boldsymbol{\alpha}) &= 0, \quad i\neq i' \end{aligned}$$

   Note that

$$\mathrm{E}_{b_i|D}[f(\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{x}_{ij})] \neq f(\boldsymbol{\beta},\boldsymbol{0},\boldsymbol{x}_{ij})$$

   we had equality for the linear model.

   The data do not have a known marginal distribution.

2. How do we then maximize the resultant likelihood? For the linear model we used EM or Newton-Raphson algorithms.

Overview of Integration Techniques

We describe a number of generic integration techniques, in particular:

- Laplace approximation (an analytical approximation).
- Quadrature (numerical integration).
- Importance sampling (a Monte Carlo method).

Before the MCMC revolution these techniques were used in a Bayesian context.

Laplace Approximation

Let

$$I = \int \exp\{ng(\theta)\}\mathrm{d}\theta,$$

denote a generic integral of interest and suppose m is the maximum of $g(\cdot)$.

We have

$$ng(\theta) = n\sum_{k=0}^{\infty} \frac{(\theta - m)^k}{k!} g^{(k)}(m),$$

where $g^{(k)}(m)$ represents the $k-$th derivative of g evaluated at m. Hence

$$\begin{aligned} I &= \int \exp\left\{ n\sum_{k=0}^{\infty} \frac{(\theta - m)^k}{k!} g^{(k)}(m) \right\} \mathrm{d}\theta \\ &\approx e^{ng(m)} \int \exp\left\{ \frac{(\theta - m)^2}{2/[ng^{(2)}(m)]} \right\} \mathrm{d}\theta \\ &= e^{ng(m)}(2\pi v)^{1/2}n^{-1/2} \end{aligned}$$

where$v = -1/[g^{(2)}(m)]$, and we have ignored terms in cubics or greater in the Taylor series.

Laplace Approximation in the NLMEM

See Pinheiro and Bates, Chapter 7.

We wish to evaluate

$$p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) = (2\pi\sigma^2)^{-n_i/2}(2\pi)^{-(q+1)/2} \mid \boldsymbol{D} \mid^{-1/2} \int \exp\{n_i g(\boldsymbol{b}_i)\}\ d\boldsymbol{b}_i,$$

where

$$-2n_i g(\boldsymbol{b}_i) = [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i)]^{\mathrm{T}}[\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i)]/\sigma_\epsilon^2 + \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{b}_i.$$

A Laplace approximation is a second-order Taylor series expansion of g about

$$\widehat{\boldsymbol{b}}_i = \arg\min_{\boldsymbol{b}_i} -g(\boldsymbol{b}_i)$$

which will not be available in closed form for a non-linear model.

Gaussian Quadrature

A general method of integration is provided by quadrature (numerical integration) in which an integral

$$I = \int f(u)\ \mathrm{d}u,$$

is approximated by

$$\widehat{I} = \sum_{i=1}^{n_w} f(u_i)w_i,$$

for design points $u_1, ..., u_{n_w}$ and weights $w_1, ..., w_{n_w}$. Different choices of (u_i, w_i) lead to different integration rules.

In mixed model applications we have integrals with respect to a normal density, *Gauss-Hermite* quadrature is designed for problems of this type.

Specifically, it provides exact integration of

$$\int_{-\infty}^{\infty} g(u)\mathrm{e}^{-u^2}\ \mathrm{d}u,$$

where $g(\cdot)$ is a polynomial of degree $2n_w - 1$.

The design points are the zeroes of the so-called Hermite polynomials. Specifically, for a rule of n_w points, u_i is the $i-$th zero of $H_{n_w}(u)$, the Hermite polynomial of degree n_w, and

$$w_i = \frac{w^{n_w-1} n_w! \sqrt{\pi}}{n_w^2 [H_{n_w-1}(u_i)]^2}.$$

Now suppose $\boldsymbol{\theta}$ is two-dimensional and we wish to evaluate

$$I = \int f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \int \int f(\theta_1, \theta_2) \mathrm{d}\theta_2 \mathrm{d}\theta_1 = \int f^*(\theta_1) \mathrm{d}\theta_1,$$

where

$$f^*(\theta_1) = \int f(\theta_1, \theta_2) \mathrm{d}\theta_2.$$

Now form

$$\widehat{I} = \sum_{i=1}^{m_1} w_i \widehat{f}^*(\theta_{1i}),$$

where

$$\widehat{f}^*(\theta_{1i}) = \sum_{j=1}^{m_2} u_j f(\theta_{1i}, \theta_{2j}).$$

Then we have

$$\widehat{I} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_i u_j f(\theta_{1i}, \theta_{2j}),$$

which is known as the *Cartesian Product.*

Scaling and reparameterization

To implement this method the function must be centered and scaled in some way, for example we could center and scale by the current estimates of the mean, $\boldsymbol{m}$, and variance-covariance matrix, $\boldsymbol{V}$ – known as adaptive quadrature.

We then form

$$\boldsymbol{X} = \boldsymbol{L}(\boldsymbol{\theta} - \boldsymbol{m})$$

where $\boldsymbol{L}'\boldsymbol{L} = \boldsymbol{V}^{-1}$ and carry out integation in the space of $\boldsymbol{X}$.

There is no guarantee that the most efficient rule is obtained by scaling in terms of the posterior mean and variance, but we note that the 'best' normal approximation to a density (in terms of Kullbach-Leibler divergence) has the same mean and variance.

Gauss-Hermite Code in R

Nodes and weights for $n = 4$:

```
> n <- 4
> quad <- gauss.quad(n,kind="hermite")
> quad$nodes
[1] -1.6506801 -0.5246476  0.5246476  1.6506801
> quad$weights
[1] 0.08131284 0.80491409 0.80491409 0.08131284
```

Nodes and weights for $n = 5$:

```
> n <- 5
> quad <- gauss.quad(n,kind="hermite")
> quad$nodes
[1] -2.0201829 -0.9585725  0.0000000  0.9585725  2.0201829
> quad$weights
[1] 0.01995324 0.39361932 0.94530872 0.39361932 0.01995324
```

Importance Sampling

Rather than deterministically selecting points we may randomly generate points from some density $h(\boldsymbol{\theta})$.

We have

$$I = \int f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \int \frac{f(\boldsymbol{\theta})}{h(\boldsymbol{\theta})} h(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \mathrm{E}[w(\boldsymbol{\theta})],$$

where $w(\boldsymbol{\theta}) = f(\boldsymbol{\theta})/h(\boldsymbol{\theta})$.

Hence we have the obvious estimator

$$\widehat{I} = \sum_{i=1}^{m} w(\boldsymbol{\theta}_i),$$

where $\theta_i \sim_{iid} h(\cdot)$. We have $\mathrm{E}[\widehat{I}] = I$ and

$$V = \mathrm{var}(\widehat{I}) = \frac{1}{m} \mathrm{var}\{w(\boldsymbol{\theta})\}.$$

From this expression it is clear that a good $h(\cdot)$ produces an approximately constant $w(\boldsymbol{\theta})$.

We may estimate V via

$$\widehat{V} = \frac{1}{m}\sum_{i=1}^{m}\frac{f^2(\boldsymbol{\theta}_i)}{h^2(\boldsymbol{\theta}_i)} - \frac{1}{m}\widehat{I}^2,$$

and (appealing to the central limit theorem) $\hat{I}$ is asymptotically normal and so a $100(1-\alpha)\%$ confidence interval is given by

$$\hat{I} \pm Z_{\alpha/2}\hat{V}^{1/2}$$

where $Z_{\alpha/2}$ is the $\alpha/2$ point of an $N(0,1)$ random variable.

Hence the accuracy of the approximation may be directly assessed, providing an advantage over analytical approximations and quadrature methods.

Notes on Importance Sampling

- We require an $h(\cdot)$ with heavier tails than the integrand. We can carry out importance sampling with any h but if the tails are lighter we will have an estimator with infinite variance (and hence an inconsistent procedure). Many suggestions for h have been made including Student t distributions and mixtures of Student t distributions.
- Iteration may again be used to obtain an estimator with good properties.

Notes on Implementation

- If the number of parameters is small then numerical integration techniques (e.g. quadrature) are highly efficient in terms of the number of function evaluations required. Hence if, for example, obtaining a point on the likelihood surface is computationally expensive (as occurs if a large simulation is required) then such techniques are preferable to Monte Carlo methods.
- The method employed will depend on whether it is for a one-off application, in which case ease-of-implementation is a consideration, or for a great deal of use, in which case an efficient method may be required.
- In general it is difficult to assess the accuracy of Laplace/numerical integration techniques.
- For simulation methods we note that independent samples are ideal for assessing Monte Carlo error since standard errors on expectations of interest may be simply calculated.
- Evans and Swartz (1995, Statistical Science) provide a good review of integration techniques.

The nlme algorithm

Within nlme an algorithm, introduced by Lindstrom and Bates (1990) is used.

The algorithm alternates between two steps:

Penalized Non-linear Least Squares (PNLS)

Condition on the current estimates of $\widehat{\boldsymbol{D}}$ and $\widehat{\sigma}_\epsilon^2$ and then minimize

$$\frac{1}{\widehat{\sigma}_\epsilon^2}\sum_{i=1}^m (\boldsymbol{y}_i - \boldsymbol{f}_i)^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{f}_i) + \boldsymbol{b}_i \widehat{\boldsymbol{D}}^{-1}\boldsymbol{b}_i,$$

to obtain estimates $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}_1, ..., \widehat{\boldsymbol{b}}_m$, which may be viewed as finding the posterior mode for $\boldsymbol{\beta}$ and $\boldsymbol{b}_1, ..., \boldsymbol{b}_m$.

Linear Mixed Effects (LME)

Carry out a first-order Taylor series of $\boldsymbol{f}_i$ about $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}_i$.

This results in a linear mixed effects model which can be maximized to obtain estimates of $\boldsymbol{D}$ and σ_ϵ^2.

We have likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}) =\mid \boldsymbol{D} \mid^{-m/2} \sigma_\epsilon^{-N} \int \exp\left\{ -\frac{1}{2}\sum_{i=1}^m (\boldsymbol{y}_i - \boldsymbol{f}_i)^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{f}_i) - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{b}_i \right\} d\boldsymbol{b}_i$$

where $\boldsymbol{f}_i = \boldsymbol{f}(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i)$, $i = 1, ..., m$.

Carry out a first-order Taylor series expansion of $\boldsymbol{f}_i$ about the estimates, obtained in the PNLS step at iteration k, of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, call these $\widehat{\boldsymbol{\beta}}^{(k)}$ and $\widehat{\boldsymbol{b}}_i^{(k)}$.

Specifically

$$\boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i) \quad \approx \quad \boldsymbol{f}_i\left(\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{b}}_i^{(k)}\right) + \widehat{\boldsymbol{x}}_i^{(k)}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(k)}\right) + \widehat{\boldsymbol{z}}_i^{(k)}\left(\boldsymbol{b}_i - \widehat{\boldsymbol{b}_i}^{(k)}\right)$$

where

$$\begin{aligned} \widehat{\boldsymbol{x}}_i^{(k)} &= \left.\frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right|_{\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{b}}_i^{(k)}} \\ \widehat{\boldsymbol{z}}_i^{(k)} &= \left.\frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{b}_i^{\mathrm{T}}}\right|_{\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{b}}_i^{(k)}} \end{aligned}$$

This gives

$$\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i) \quad \approx \quad \boldsymbol{y}_i^{(k)} - \widehat{\boldsymbol{x}}_i^{(k)}\boldsymbol{\beta} - \widehat{\boldsymbol{z}}_i^{(k)}\boldsymbol{b}_i$$

where

$$\boldsymbol{y}_i^{(k)} \quad = \quad \boldsymbol{y}_i - \boldsymbol{f}_i\left(\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{b}}_i^{(k)}\right) + \widehat{\boldsymbol{x}}_i^{(k)}\widehat{\boldsymbol{\beta}}^{(k)} + \widehat{\boldsymbol{z}}_i^{(k)}\widehat{\boldsymbol{b}}_i^{(k)}$$

The integral can now be evaluated in closed-form to give the log-likelihood

$$l(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{m} \log|\widehat{\boldsymbol{V}}_i| - \frac{1}{2}\sum_{i=1}^{m} (\boldsymbol{y}_i^{(k)} - \widehat{\boldsymbol{x}}_i^{(k)}\boldsymbol{\beta})^{\mathrm{T}}\widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{Y}_i - \widehat{\boldsymbol{x}}_i\boldsymbol{\beta})$$

where

$$\widehat{\boldsymbol{V}}_i = \widehat{\boldsymbol{z}}_i^{(k)}\boldsymbol{D}\widehat{\boldsymbol{z}}_i^{(k)\mathrm{T}} + \sigma_\epsilon^2\boldsymbol{I}_i,$$

which may be maximized to give ML estimates. REML estimates are obtained by adding the term

$$-\frac{1}{2}\sum_{i=1}^{m} \log \mid \widehat{\boldsymbol{x}}_i^{(k)\mathrm{T}}\widehat{\boldsymbol{V}}_i(\boldsymbol{\alpha})\widehat{\boldsymbol{x}}_i^{(k)} \mid$$

The Laplace approximation is generally more accurate than the LB algorithm, it is, however, more computationally expensive.

Asymptotic Inference

Under the LB algorithm, the asymptotic distribution of the REML estimator $\widehat{\boldsymbol{\beta}}$ is

$$\left(\sum_{i=1}^{m} \widehat{\boldsymbol{x}}_i^{\mathrm{T}} \widehat{\boldsymbol{V}}_i^{-1} \widehat{\boldsymbol{x}}_i \right)^{1/2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\to_d\; \mathrm{N}_{p+1}(\boldsymbol{0}, \boldsymbol{I}_{p+1}),$$

where $\widehat{\boldsymbol{x}}_i = \widehat{\boldsymbol{x}}_i^{(k)}$ with k the final iteration, $i = 1, ..., m$

Similarly, the asymptotic distribution of $\boldsymbol{\alpha}$ is based on the information as calculated from the linear approximation to the likelihood.

The LB estimator is inconsistent if the n_i's are fixed and $m \to \infty$.

Empirical Bayes estimates for the random effects are available, but caution should be given to using these for checking assumptions since they are strongly influenced by the assumption of normality being correct. If n_i is large then this will be less of a problem.

Approaches for NLMEMs

Various other approaches to likelihood inference have been suggested, we briefly summarize.

In general we need to carry out m integrals of dimension $q + 1$ for each likelihood evaluation, so with large m and q this can be computationally expensive.

First-Order Approximation

Let $\boldsymbol{\beta}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{b}_i$, and then carry out a first-order Taylor series about $\mathrm{E}[\boldsymbol{b}_i] = \boldsymbol{0}$ to give

$$\boldsymbol{y}_i = \boldsymbol{f}_i(\boldsymbol{\beta}_i) + \boldsymbol{\epsilon}_i \approx \boldsymbol{f}_i(\boldsymbol{x}_i\boldsymbol{\beta}) + \frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{\beta}_i} \frac{\partial \boldsymbol{\beta}_i}{\partial \boldsymbol{b}_i} \boldsymbol{b}_i + \boldsymbol{\epsilon}_i.$$

In contrast to the LB algorithm which considered an expansion about the subject-specific mean, the expansion here is about the population-averaged mean. The first-order estimator is inconsistent and has bias even if n_i and m go to infinity, see Demidenko (2004, Chapter 8)

Adaptive Gaussian quadrature may also be used.

Example: Pharmacokinetics of Indomethacin

The compartmental model that has previously been used for this drug is the two-compartment bi-exponential model:

$$\mathrm{E}[Y] = A_1 \exp\{-\alpha_1 t\} + A_2 \exp\{-\alpha_2 t\},$$

where Y is concentration, and t is time, and $A_1, A_2, \alpha_1, \alpha_2 > 0$.

Note: this model is unidentifiable since the parameter set $(A_1, \alpha_1, A_2, \alpha_2)$ gives the same fitted curve (and hence likelihood) as the set $(A_2, \alpha_2, A_1, \alpha_1)$. If this is a practical problem for a particular dataset (say $\alpha_1 \approx \alpha_2$) then we may parameterize in terms of α_1 and $\alpha_2 - \alpha_1$.

Figure 23 gives the log concentrations versus time – such a plot can be useful for picking the number of exponentials (and modeling the log concentration can provide initial estimates). Certainly not linear in time so more than a single exponential needed.

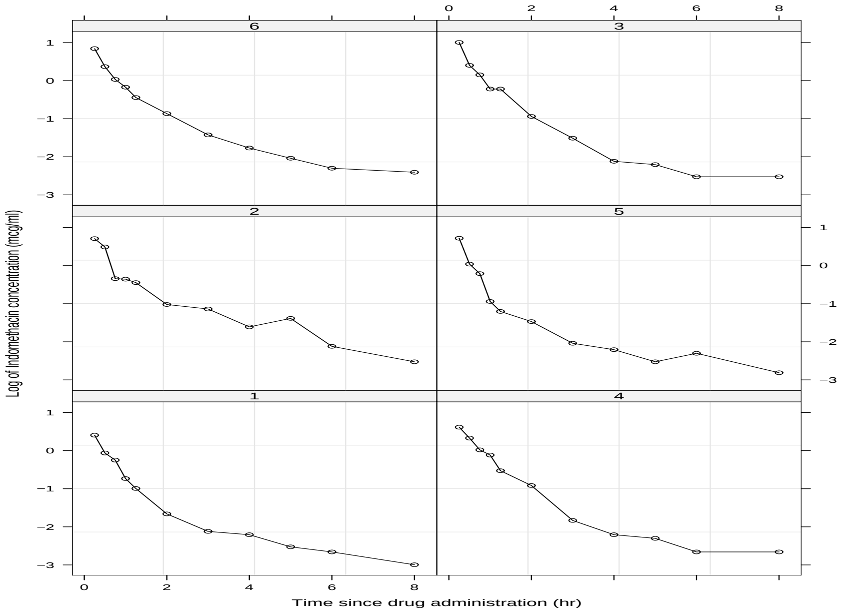

Figure 23: Log concentration time data for Indomethacin.

Individual fits

Let Y_{ij} be the drug concentration at time t_{ij} on indvidual i, $j = 1, ..., 11$, $i = 1, ..., 6$. We first fit bi-exponential models to each individual, using non-linear least squares.

We parameterize as

$$\mathrm{E}[Y_{ij} \mid \boldsymbol{\beta}_i] = \beta_{1i} \exp\{-\mathrm{e}^{\beta_{3i}} t_{ij}\} + \beta_{2i} \exp\{-\mathrm{e}^{\beta_{4i}} t_{ij}\},$$

for $i = 1, ..., 6$.

Even though the data are balanced, the standard errors are different for different individuals, as we see in Figure 24.

`R` code for fitting individual models:

```
> indiv.lis <- nlsList( conc ~ SSbiexp(time,A1,lrc1,A2,lrc2),data=Indometh )
> indiv.lis
Call:
Model:conc~SSbiexp(time,A1,lrc1,A2,lrc2)|Subject
   Data: Indometh
Coefficients:
        A1      lrc1        A2       lrc2
1 2.029277 0.5793887 0.1915475 -1.7877849
4 2.198132 0.2423124 0.2545223 -1.6026859
2 2.827673 0.8013195 0.4989175 -1.6353512
5 3.566103 1.0407660 0.2914970 -1.5068522
6 3.002250 1.0882119 0.9685230 -0.8731358
3 5.468312 1.7497936 1.6757522 -0.4122004
Degrees of freedom: 66 total; 42 residual
Residual standard error: 0.0755502
> plot( intervals(indiv.lis) )
```

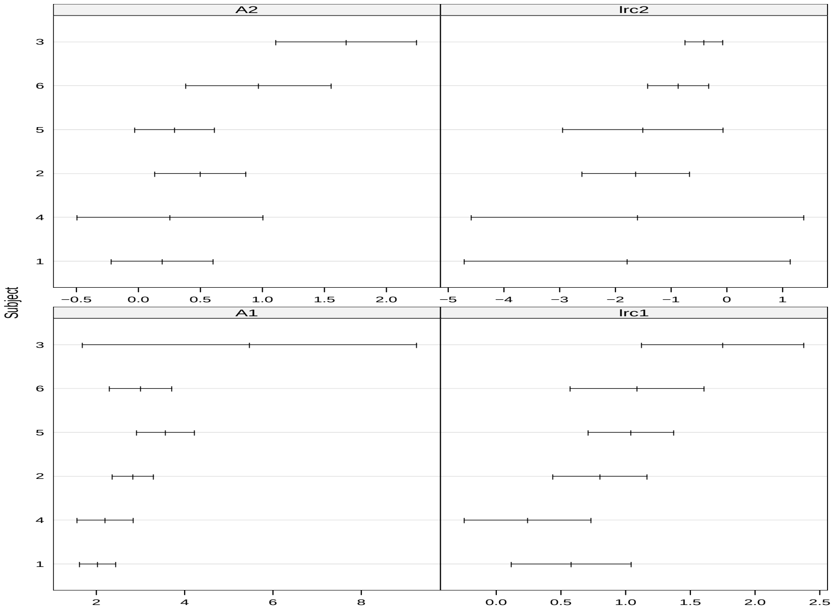

Figure 24: Asymptotic 95% CIs for elements of $\boldsymbol{\beta}_i$, $i = 1, ..., 6$.

Now we fit some NLMEMs, we first assume a diagonal $\boldsymbol{D}$ with random effects for first three elements only.

```
> nlme.indo <- nlme( indiv.lis,random=pdDiag(A1+lrc1+A2~1))
> summary(nlme.indo)
Nonlinear mixed-effects model fit by maximum likelihood
  Model: conc ~ SSbiexp(time, A1, lrc1, A2, lrc2)
Random effects:
 Formula: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1)
 Level: Subject
 Structure: Diagonal
             A1      lrc1        A2   Residual
StdDev: 0.57135 0.1581214 0.1115283 0.08149631
Fixed effects: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1, lrc2 ~ 1)
          Value Std.Error DF   t-value p-value
A1    2.8276029 0.2639744 57 10.711656   0e+00
lrc1  0.7732529 0.1100086 57  7.029021   0e+00
A2    0.4610197 0.1127560 57  4.088648   1e-04
lrc2 -1.3450041 0.2313139 57 -5.814627   0e+00
 Correlation:
     A1     lrc1   A2
lrc1  0.055
A2   -0.102  0.630
lrc2 -0.139  0.577  0.834
```

Now assume a non-diagonal $\boldsymbol{D}$ for all four parameters.

```
> nlme2.indo2 <- update( nlme.indo, random=A1+lrc1+A2+lrc2~1)
> summary(nlme.indo2)
  Model: conc ~ SSbiexp(time, A1, lrc1, A2, lrc2)
Random effects: Formula: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1, lrc2 ~ 1)
 Structure: General positive-definite, Log-Cholesky parametrization
         StdDev     Corr
A1       0.77583020 A1    lrc1  A2
lrc1     0.26863662 0.963
A2       0.38707000 0.459 0.682
lrc2     0.48253192 0.153 0.414 0.948
Residual 0.06962038
Fixed effects: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1, lrc2 ~ 1)
          Value Std.Error DF   t-value p-value
A1    2.8531611 0.3485825 57  8.185039   0e+00
lrc1  0.8755645 0.1253269 57  6.986245   0e+00
A2    0.6357872 0.1715520 57  3.706091   5e-04
lrc2 -1.2757709 0.2161119 57 -5.903288   0e+00
 Correlation:
     A1    lrc1  A2
lrc1 0.907
A2   0.411 0.676
lrc2 0.108 0.378 0.912
```

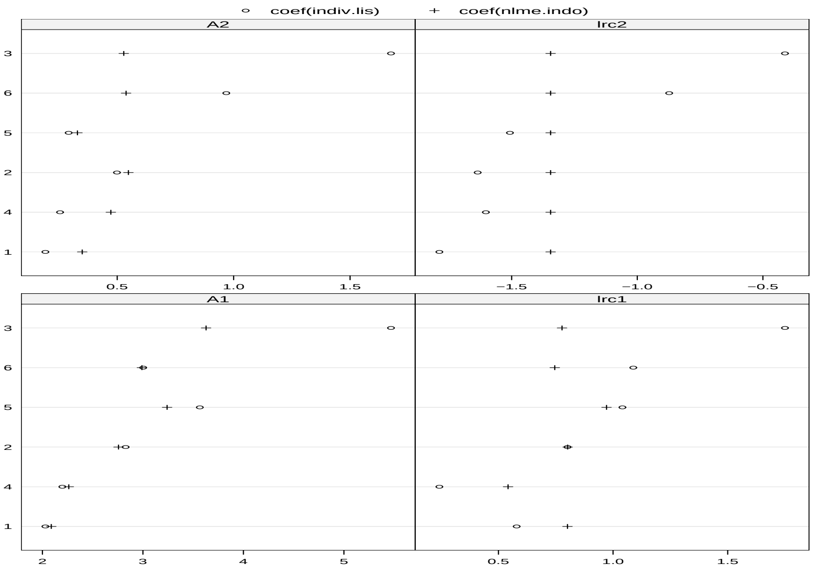

Figure 25: Comparison of non-linear LS and nlme estimates, with the latter from the model `nlme.indo` Created using the command `plot(compareFits(coef(indiv.lis),coef(nlme.indo)))`.

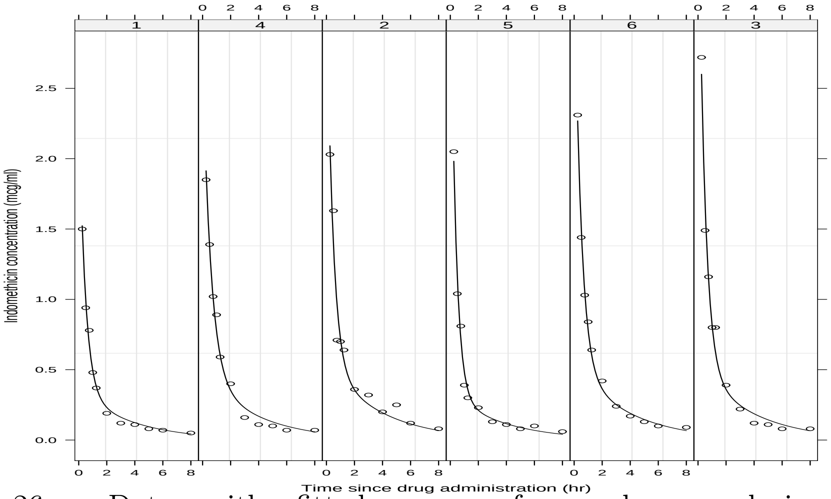

Figure 26: Data with fitted curves from nlme analysis superimposed from the model `nlme.indo`. Created with the command `plot(augPred(nlme.indo),aspect=’’xy’’,grid=T)`.

The following commands produced Figures 27–29.

```
> plot(nlme.indo,resid(.,type="n")~fitted(.),id=0.05,adj=-1) # id=0.05 gives
# outliers outside of 95% of distn, adj=-1 adjusts the text which
# labels these outliers
> plot(nlme.indo,resid(.,type="n")~time,id=0.05,adj=-1)
> qqnorm(nlme.indo)
> plot(augPred(nlme.indo,level=0:1)) # Obtain predictions at population and
# individual level of hierarchy
```

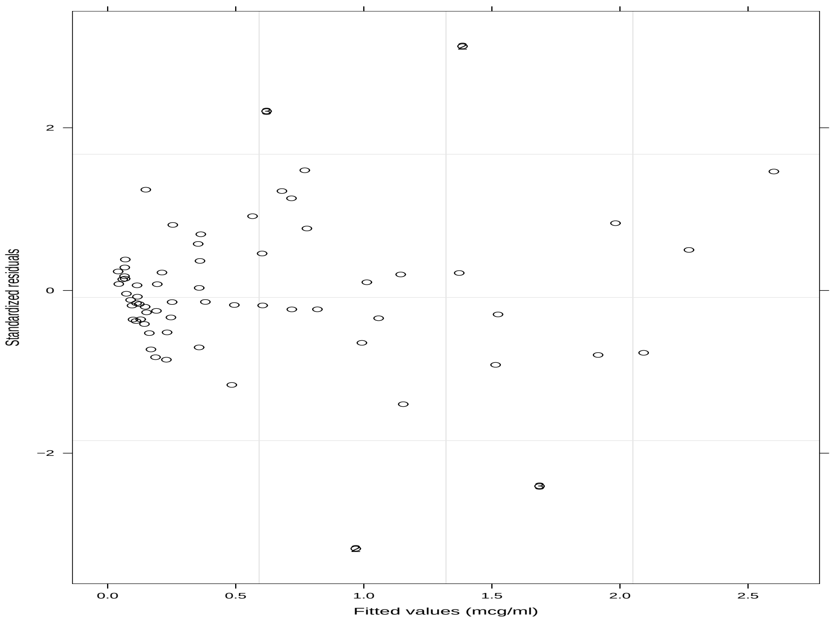

Figure 27: Standardized residuals versus fitted values.

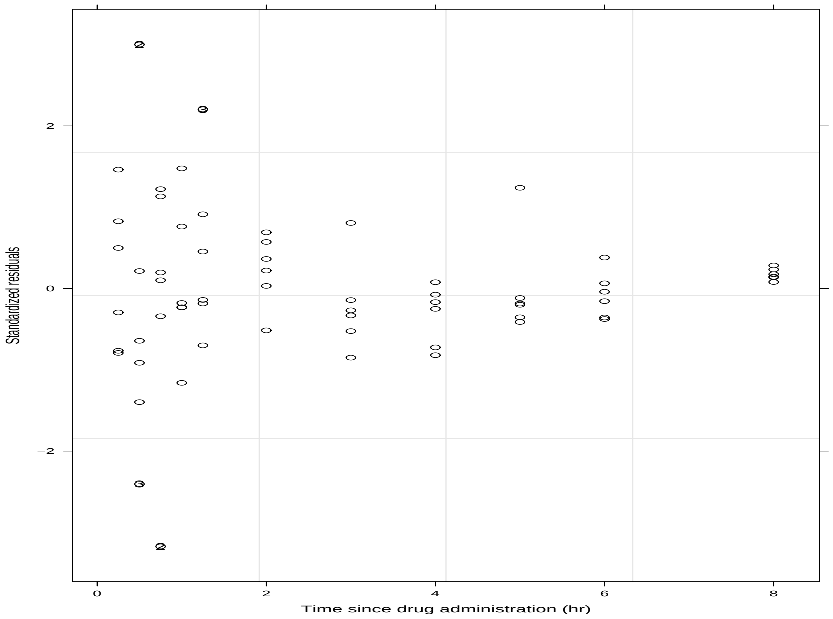

Figure 28: Standardized residuals versus time.

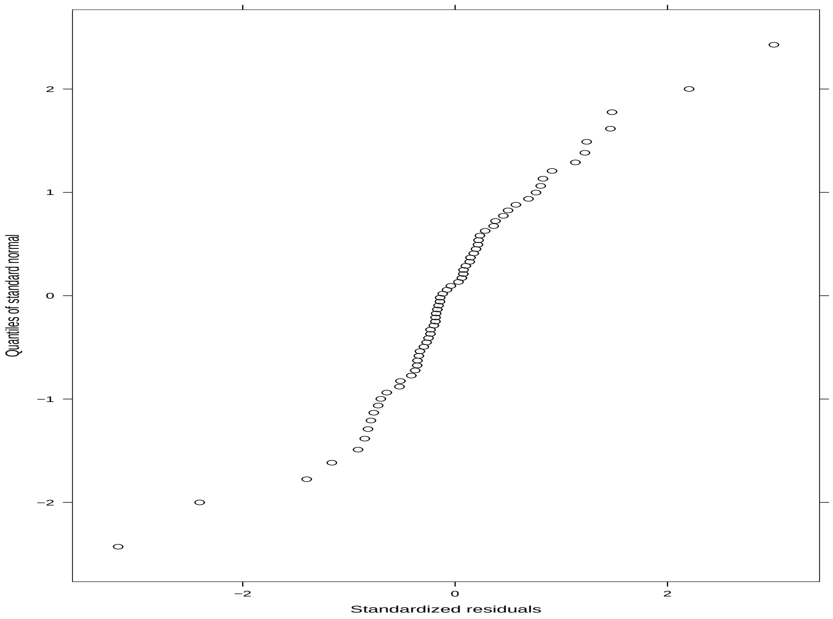

Figure 29: QQ plot of normalized residuals.

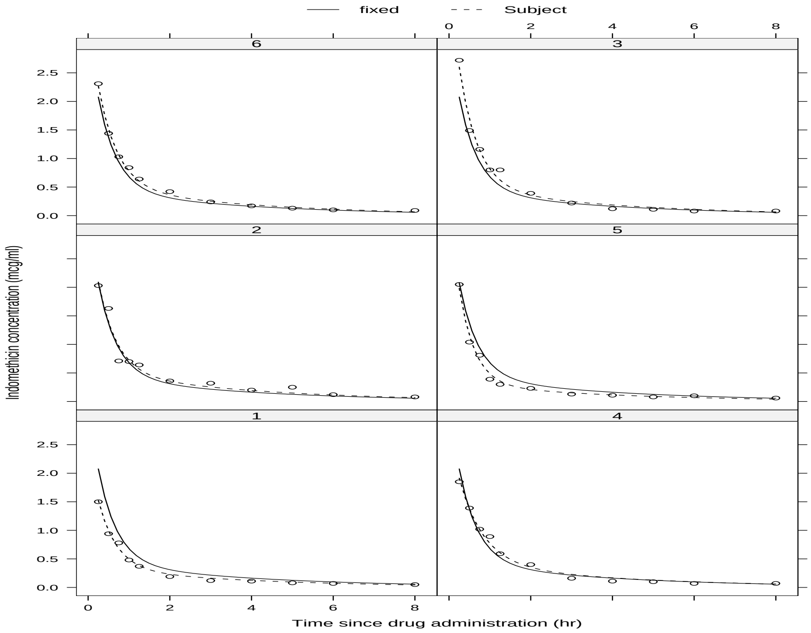

Figure 30: Solid lines are population predictions, dashed lines individual predictions.

Bayesian Approach

A Bayesian approach adds a prior distribution for $\boldsymbol{\beta}, \boldsymbol{\alpha}$, to the likelihood $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$. As with the linear model proper prior is required for the matrix $\boldsymbol{D}$. In general a proper prior is required for $\boldsymbol{\beta}$ also, to ensure the propriety of the posterior distribution. Closed-form inference is unavailable, but MCMC is almost as straightforward as in the LMEM case. The joint posterior is

$$p(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \tau, \boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{y}) \propto \prod_{i=1}^{m} \{p(\boldsymbol{y}_i \mid \boldsymbol{\beta}_i, \tau) p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{W})\} \, \pi(\boldsymbol{\beta})\pi(\tau)\pi(\boldsymbol{W}).$$

Suppose we have priors:

$$\begin{array}{rcl} \boldsymbol{\beta} & \sim & \mathrm{N}_{q+1}(\boldsymbol{\beta}_0, \boldsymbol{V}_0) \\ \tau & \sim & \mathrm{Ga}(a_0, b_0) \\ \boldsymbol{W} & \sim & \mathrm{W}_{q+1}(r, \boldsymbol{R}^{-1}) \end{array}$$

The conditional distributions for $\boldsymbol{\beta}$, τ, $\boldsymbol{W}$ are unchanged from the linear case. There is no closed form conditional distribution for $\boldsymbol{\beta}_i$, which is given by:

$$p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \tau, \boldsymbol{W}, \boldsymbol{y}) \propto p(\boldsymbol{y}_i \mid \boldsymbol{\beta}_i, \tau) \times p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{W})$$

but a Metropolis-Hastings step can be used.

Generalized Estimating Equations

If interest lies in population parameters then we may use the estimator $\widehat{\boldsymbol{\beta}}$ that satisfies

$$\boldsymbol{G}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}) = \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0},$$

where $\boldsymbol{D}_i = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}$, $\boldsymbol{W}_i = \boldsymbol{W}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})$ is the working covariance model, $\mu_i = \mu_i(\boldsymbol{\beta})$ and $\widehat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$. Sandwich estimation may be used to obtain an empirical estimate of the variance, $\boldsymbol{V}_\beta$:

$$\left(\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right)^{-1} \left\{ \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \mathrm{cov}(\boldsymbol{Y}_i) \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right\} \left(\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right)^{-1}.$$

We then have

$$\boldsymbol{V}_\beta^{-1/2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}).$$

In practice an empirical estimator of $\mathrm{cov}(\boldsymbol{Y}_i)$ is substituted to give $\widehat{\boldsymbol{V}}_\beta$.

GEE has not been extensively used in a non-linear (non-GLM) setting. This is probably because in many settings (e.g. pharmacokinetic/pharmacodynamic) interest focuses on understanding between individual-variability, and explaining this in terms of individual-specific covariates.