The effect of ignoring dependence is two-fold. First, standard errors reported from independent data methods are calculated under the assumption of independence and so, if the data are truly dependent, will be inaccurate. Second, and more subtly, models for dependence may control for confounding, e.g. by time in the air pollution example.

While making inference for dependent data is more difficult than for independent data, designs that collect dependent data can be very efficient. For example, in a longitudinal data setting applying different treatments to the same patient over time can be very beneficial since each patient acts as their own control.

While in the Bayesian approach to inference all parameters are viewed as random variables, in the frequentist approach there is a distinction between *fixed effects* (unknown constants) and *random effects* (random variables from a distribution).

For longitudinal data there are two extreme fixed effects approaches. Proceeding naively we could assume a single "marginal" curve for *all* of the data, and carry out a standard analysis assuming independent data.

Example: Dental Growth Data

Suppose $\widehat{\beta}_0^m$ and $\widehat{\beta}_1^m$ are the marginal intercept and slope estimates, and let

$$e_{ij}^m = Y_{ij} - \widehat{\beta}_0^m - \widehat{\beta}_1^m t_j,$$

$i = 1, ..., 11; j = 1, ..., 4$, denote marginal residuals, and

$$\begin{bmatrix} \sigma_1 & & & \\ \rho_{12} & \sigma_2 & & \\ \rho_{13} & \rho_{23} & \sigma_3 & \\ \rho_{14} & \rho_{24} & \rho_{34} & \sigma_4 \end{bmatrix} \tag{9}$$

represent the standard deviation/correlation matrix of the residuals, where

$$\sigma_j = \sqrt{\text{var}(e_{ij}^m)},$$

is the variance of the length at time t_j, $j = 1, ..., 4$, and

$$\rho_{jk} = \frac{\text{cov}(e_{ij}^m, e_{ik}^m)}{\sqrt{\text{var}(e_{ij}^m)\text{var}(e_{ik}^m)}},$$

is the correlation between residual measurements at times t_j and t_k taken on the same girl, $j \neq k, j, k = 1, ..., 4$.

Across girls we may empirically estimate the entries of (9) by

$$\begin{bmatrix} 2.12 & & & \\ 0.83 & 1.90 & & \\ 0.86 & 0.90 & 2.36 & \\ 0.84 & 0.88 & 0.95 & 2.44 \end{bmatrix} \tag{10}$$

illustrating that there is a suggestion that the variance is increasing with the mean, and clear correlation between residuals at different times on the same girl.

The fitting of a single curve, and using methods for independent data, ignores the correlations within each child's data and so standard errors will clearly be inappropriate.

Fitting a marginal model such as this is appealing in one sense, however, since it allows the direct comparison of the average responses in different (in this example the populations of girls at different ages) and forms the basis of the generalized estimating equations (GEE) approach

An alternative fixed effects approach is to assume a fixed curve for each child and analyze each set of data separately.

While providing valid inference for each curve, there is no "borrowing of strength" across children, that is, each girl's fit is based only on their data alone, and not on those of other girls.

We would hope that if there is *similarity* between the curves, and that the totality of data will aid in the estimation of each individual curve. In some instances this may be vital, for example, if $n_i = 1$ for a particular individual, then their own data alone will not allow parameter estimation.

We will also often be interested in making formal inference for the population of girls from which the eleven in the data are viewed as a random sample. This forms the basis of the mixed effects model approach.

Figure 5(b) displays the lines corresponding to each of these fixed effects approaches.

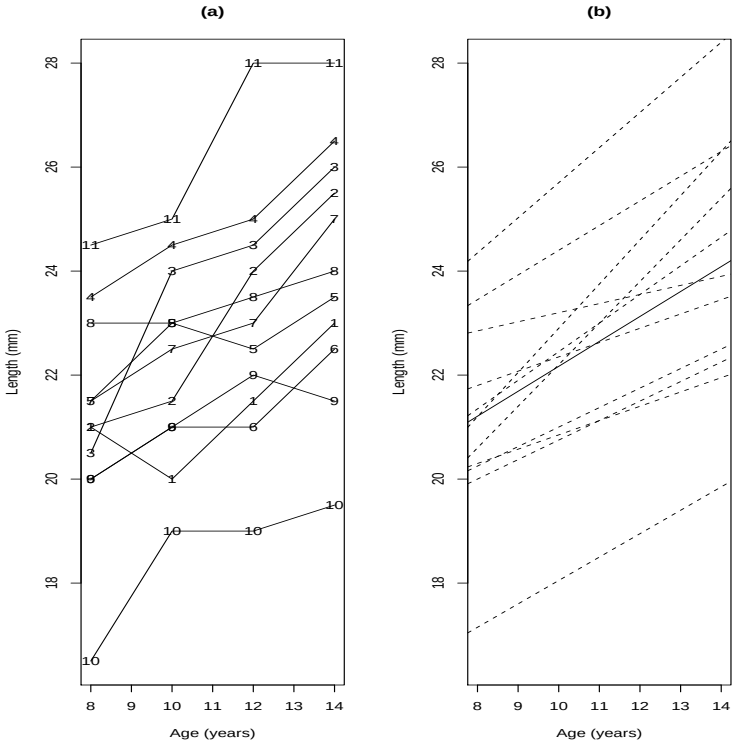

Figure 5: Dental plots for girls only: (a) Individual observed data (with plotting symbol girl index), (b) Individual fitted curves (dashed) and overall fitted curve

Linear Mixed Effects Models

The basic idea behind mixed effects models is to assume that each unit has a regression model characterized by unit-specific parameters, with these parameters being a combination of fixed effects that are common to all units in the population, and then unit-specific perturbations, or random effects (hence "mixed" effects refers to the combination of fixed and random effects).

Given data $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})^{\mathrm{T}}$ on unit i a mixed effects model is characterized by a combination of

- 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 $\boldsymbol{z}_{ij} = (1, z_{ij1}, ..., z_{ijq})^{\mathrm{T}}$.

We then have the following (two stage) Linear Mixed Effects Model (LMEM):
Stage 1: Response model, *conditional* on random effects:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{11}$$

where $\boldsymbol{\epsilon}_i$ is an $n_i \times 1$ zero mean vector of error terms.

Stage 2: Model for random terms:

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

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

The two stages define the marginal model:

$$\begin{aligned} \mathrm{E}[\boldsymbol{y}_i] &= \boldsymbol{\mu}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i\boldsymbol{\beta}, \\ \mathrm{var}(\boldsymbol{y}_i) &= \boldsymbol{V}_i(\boldsymbol{\alpha}) = \boldsymbol{z}_i\boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}} + \boldsymbol{E}_i, \\ \mathrm{cov}(\boldsymbol{y}_i, \boldsymbol{y}_{i'}) &= \mathbf{0}, \quad i \neq i'. \end{aligned}$$

We describe likelihood and Bayesian approaches to inference.

Likelihood Inference

We need to specify a complete probability distribution for the data, and this follows by specifying distributions for $\boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i$, $i = 1, ..., m$. A common model is

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

where

$$\boldsymbol{D} = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01}^2 & ... & \sigma_{0q}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 & ... & \sigma_{1q}^2 \\ ... & ... & ... & ... \\ \sigma_{q0}^2 & \sigma_{q1}^2 & ... & \sigma_{qq}^2 \end{bmatrix}.$$

Here $\boldsymbol{\alpha} = (\sigma_\epsilon^2, \boldsymbol{D})$ denote the variance-covariance parameters. Here $\boldsymbol{V} = \boldsymbol{z}\boldsymbol{D}\boldsymbol{z}^{\mathrm{T}} + \sigma_\epsilon^2 \boldsymbol{I}_N$, where $N = \sum_{i=1}^m n_i$.

Likelihood methods are designed for fixed effects, and so we integrate the random effects from the two-stage model:

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\alpha}) = \int_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \times p(\boldsymbol{b}|\boldsymbol{\beta}, \boldsymbol{\alpha}) \; d\boldsymbol{b}.$$

Exploiting conditional independencies we have:

$$p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\alpha}) = \prod_{i=1}^{m} \int_{\boldsymbol{b}_i} p(\boldsymbol{y}_i|\boldsymbol{b}_i,\boldsymbol{\beta},\sigma^2_\epsilon) \times p(b_i|\boldsymbol{D})\ d\boldsymbol{b}_i.$$

Since a convolution of normals is normal we obtain

$$\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\alpha} \sim \prod_{i=1}^{m} N\{\boldsymbol{\mu}_i(\boldsymbol{\beta}), \boldsymbol{V}_i(\boldsymbol{\alpha})\}.$$

The log-likelihood is

$$\begin{aligned} l(\boldsymbol{\beta},\boldsymbol{\alpha}) = & - \frac{N}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{m} \log|\boldsymbol{V}_i(\boldsymbol{\alpha})| \\ & - \frac{1}{2}\sum_{i=1}^{m} (\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{V}(\boldsymbol{\alpha})_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}). \end{aligned} \tag{12}$$

Example: One-way ANOVA

Consider the simple ANOVA model

$$Y_{ij} = \beta_0 + b_i + \epsilon_{ij},$$

with b_i and ϵ_{ij} independent and distributed as

- $b_i \sim_{ind} \mathrm{N}(0, \sigma^2_0)$,
- $\epsilon_{ij} \sim_{ind} \mathrm{N}(0, \sigma^2_\epsilon)$

for $i = 1, ..., m$, $j = 1, ..., n_i$, with $\sum_{i=1}^{m} n_i = N$. This model can also be written as

$$\boldsymbol{Y}_i = \mathbf{1}_n\beta_0 + \mathbf{1}_n b_i + \boldsymbol{\epsilon}_i,$$

with $\mathrm{E}[\boldsymbol{Y}] = \mathbf{1}_N\beta_0$, $\mathrm{var}(\boldsymbol{Y}) = \boldsymbol{V} = \mathbf{1}_N\mathbf{1}_N^{\mathrm{T}}\sigma^2_0 + \boldsymbol{I}_N\sigma^2_\epsilon = \boldsymbol{J}_N\sigma^2_0 + \boldsymbol{I}_N\sigma^2_\epsilon$.

The marginal variance $\boldsymbol{V}$ is the $N \times N$ matrix

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho & . & . & . & . & 0 & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho & . & . & . & . & 0 & 0 & 0 & 0 \\ \rho & \rho & 1 & \rho & . & . & . & . & 0 & 0 & 0 & 0 \\ \rho & \rho & \rho & 1 & . & . & . & . & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & . & . & . & 1 & \rho & \rho & \rho \\ 0 & 0 & 0 & 0 & . & . & . & . & \rho & 1 & \rho & \rho \\ 0 & 0 & 0 & 0 & . & . & . & . & \rho & \rho & 1 & \rho \\ 0 & 0 & 0 & 0 & . & . & . & . & \rho & \rho & \rho & 1 \end{bmatrix}$$

with $\sigma^2 = \sigma_\epsilon^2 + \sigma_0^2$ and

$$\rho = \frac{\sigma_0^2}{\sigma^2} = \frac{\sigma_0^2}{\sigma_\epsilon^2 + \sigma_0^2}.$$

Here we have a total of 3 regression parameters and variance components $(\beta_0, \sigma_0, \sigma_\epsilon)$, but $m + 3$ if we count the random effects.

A fixed effects model with a separate parameter for each group would have $m + 1$ parameters (and corresponds to the above model with $\sigma_0^2 = \infty$).

In some situations we may have more fixed and random effects than data points, but the random effects have a special status, since they are tied together through a common distribution.

Random effects may be viewed as a means by which dependencies are induced in marginal models.

Inference for Regression Parameters

The score equation for $\boldsymbol{\beta}$ is

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{Y}_i - \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{x}_i \boldsymbol{\beta},$$

and yields the MLE for $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{y}_i \right), \tag{13}$$

which is a weighted least squares estimator. If $\boldsymbol{D} = \boldsymbol{0}$ then $\boldsymbol{V} = \sigma_\epsilon^2 \boldsymbol{I}_N$ and $\widehat{\boldsymbol{\beta}}$ corresponds to the ordinary least squares estimator.

The variance of $\widehat{\boldsymbol{\beta}}$ may be obtained either directly from (13), or from the second derivative of the log-likelihood. Since

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = -\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{x}_i,$$

the observed and expected information matrices coincide with

$$\boldsymbol{I}_{\beta\beta} = -\mathrm{E}\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \right] = \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{x}_i.$$

The estimator, $\widehat{\boldsymbol{\beta}}$ is a linear combination of $\boldsymbol{Y}_i$ and so, under correct specificiation of the model $\widehat{\boldsymbol{\beta}}$ is linear also and

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_{k+1}\left\{ \boldsymbol{\beta}, \left(\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{V}_i^{-1} \boldsymbol{x}_i \right)^{-1} \right\}.$$

In practice, $\boldsymbol{\alpha}$ is never known, but asymptotically, as $m \to \infty$ (it is not sufficient to have m fixed and $n_i \to \infty$ for $i = 1, ..., m$):

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

where $\widehat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$. This result is also relevant if the data and random effects are not normal, so long as the second moment assumptions are correct.

Various t and F-like approaches have been suggested for correcting for the estimation of $\boldsymbol{\alpha}$, see Verbeke and Molenberghs (2000, Chapter 6), but if the sampling size is not sufficiently large for reliable estimation of $\boldsymbol{\alpha}$, we recommend following a Bayesian approach to inference.

So far as the MLE is concerned, the expected information matrix is partitioned as

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{\alpha\alpha} \end{bmatrix}.$$

Standard ML theory gives the asymptotic distribution for the MLE $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}$, as

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} \sim \mathrm{N}_{k+1+r+1}\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}^{-1}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}^{-1}_{\alpha\alpha} \end{bmatrix} \right),$$

where r is the number of distinct elements in $\boldsymbol{D}$.

We have already seen the form of $\boldsymbol{I}_{\beta\beta}$; the form of $\boldsymbol{I}_{\alpha\alpha}$ is not pleasant.

The diagonal form of the expected information has a number of implications. Firstly, we may carry out separate maximization of the log-likelihood with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Secondly, asymptotically we have independence between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$, so any consistent estimator of $\boldsymbol{\alpha}$ will give an asymptotically efficient estimator for $\boldsymbol{\beta}$.

Likelihood ratio tests are available for regression parameters.

Inference for Variance Components by MLE

The MLE of $\boldsymbol{\alpha}$ follows from maximization of (12), and in general there is no closed-form solution.

The maximization may produce a negative variance estimate, in which case this variance is set equal to zero (MLEs must lie in the parameter space).

Maximum likelihood for variance components give estimators that do not acknowledge the estimation of $\boldsymbol{\beta}$.

For the simple linear model, the MLE of σ^2 is RSS$/n$, and not the unbiased version RSS$/(n-k-1)$.

An alternative and often preferable method is provided by restricted maximum likelihood.

Hypothesis tests for variance components

Testing whether random effect variances are zero requires care since the null hypothesis lies on the boundary, and so the usual regularity conditions are not satisfied.

As an example, in the model

$$Y_{ij} = \beta_0 + b_i + \boldsymbol{x}_{ij}\boldsymbol{\beta} + \epsilon_{ij}$$

with $b_i \sim \mathrm{N}(0, \sigma_0^2)$, consider the test of $H_0 : \sigma_0^2 = 0$ versus $H_A : \sigma_0^2 > 0$, where σ_0^2 is a non-negative scalar. In this case the asymptotic null distribution is a 50:50 mixture of χ_0^2 and χ_1^2 distributions, where the former is the distribution that gives probability mass 1 to the value 0.

If the usual χ_1^2 distribution is used then the null would be accepted too often, leading to a variance component structure that is too simple.

Estimating σ_0^2 is equivalent to estimating $\rho = \sigma_0^2/\sigma^2$, and setting equal to zero if the estimated correlation is negative, and under the null this will happen half the time.

Setting $\widehat{\rho} = 0$ gives the null, and so the likelihood ratio will be one.

Inference for Variance Components by REML

Restricted maximum likelihood (REML) is a method that has been proposed as an alternative to ML, there are a number of justifications; we later provide a Bayesian justification, and here provide another based on marginal likelihood.

Marginal Likelihood

Let $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, $\boldsymbol{A}$ be a minimal sufficient statistic where $\boldsymbol{A}$ is ancillary, and for which

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}) &\propto p(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{a} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}) \\ &= p(\boldsymbol{a}) p(\boldsymbol{s}_1 \mid \boldsymbol{a}, \boldsymbol{\lambda}) p(\boldsymbol{s}_2 \mid \boldsymbol{s}_1, \boldsymbol{a}, \boldsymbol{\lambda}, \boldsymbol{\phi}) \end{aligned}$$

where $\boldsymbol{\lambda}$ are parameters of interest and $\boldsymbol{\phi}$ are the remaining (nuisance) parameters.

Inference for $\boldsymbol{\lambda}$ may be based on the *marginal* likelihood

$$L_m(\boldsymbol{\lambda}) = p(\boldsymbol{s}_1 \mid \boldsymbol{a}, \boldsymbol{\lambda}).$$

This is desirable if inference is simplified or if it avoids problems encountered with standard likelihood methods. For example $\dim(\boldsymbol{\phi})$ may increase with n. The marginal likelihood has similar properties to a regular likelihood.

These advantages may outway the loss of efficiency in ignoring the $p(\boldsymbol{s}_2 \mid \boldsymbol{s}_1, \boldsymbol{a}, \boldsymbol{\lambda}, \boldsymbol{\phi})$ term. If there is no ancillary statistic then the marginal likelihood is

$$L_m(\boldsymbol{\lambda}) = p(\boldsymbol{s}_1 \mid \boldsymbol{\lambda}).$$

Example: Normal linear model

Assume $\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2 \sim_{ind} \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$ where $\dim(\boldsymbol{\beta}) = k+1$. Suppose the parameter of interest is $\lambda = \sigma^2$, with remaining parameters $\boldsymbol{\phi} = \boldsymbol{\beta}$. Minimal sufficient statistics are: $s_1 = s^2 = \mathrm{RSS}/(n-k-1)$, and $\boldsymbol{s}_2 = \widehat{\boldsymbol{\beta}}$. We have

$$p(\boldsymbol{y} \mid \sigma^2, \boldsymbol{\beta}) = p(s_1, \boldsymbol{s}_2 \mid \sigma^2, \boldsymbol{\beta}) p(s_1 \mid \sigma^2) p(\boldsymbol{s}_2 \mid \boldsymbol{\beta}, \sigma^2).$$

Hence the marginal likelihood is

$$L_m(\sigma^2) = p(s^2 \mid \sigma^2).$$

We know

$$\frac{(n-k-1)s^2}{\sigma^2} \sim \chi^2_{n-k-1} = \mathrm{Ga}\left(\frac{n-k-1}{2}, \frac{1}{2}\right),$$

and so

$$p(s^2 \mid \sigma^2) = \left(\frac{n-k-1}{2\sigma^2}\right)^{(n-k-1)/2} \frac{\left(s^2\right)^{(n-k-1)/2-1}}{\Gamma\left(\frac{n-k-1}{2}\right)} \times \exp\left[-\frac{(n-k-1)s^2}{2\sigma^2}\right],$$

to give

$$l_m = \log L_m = -(n-k-1)\log\sigma - \frac{(n-k-1)s^2}{2\sigma^2},$$

and

$$\widehat{\sigma}^2 = s^2.$$

REML for LMEM

To use marginal likelihood we need to find a function of the data, $\boldsymbol{U} = f(\boldsymbol{Y})$, whose distribution does not depend upon $\boldsymbol{\beta}$, and then base inference for $\boldsymbol{\alpha}$ on this distribution.

A natural function to choose is the vector of residuals following an ordinary least squares fit:

$$\begin{aligned}\boldsymbol{R} &= \boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_o = \boldsymbol{Y} - \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y} \\ &= (\boldsymbol{I} - \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y},\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_o = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y}$ is the OLS estimator.

We have

$$\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{z}\boldsymbol{b} + \boldsymbol{\epsilon}) = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{z}\boldsymbol{b} + \boldsymbol{\epsilon}),$$

and so the distribution of $\boldsymbol{R}$ does not depend on $\boldsymbol{\beta}$.

Unfortunately the distribution of $\boldsymbol{R}$ is degenerate as it has rank $N - k - 1$.

Consider the $(N - k - 1) \times 1$ random variables

$$\boldsymbol{U} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$$

where $\boldsymbol{B}$ is an $N \times (N - k - 1)$ matrix with $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} = \boldsymbol{I} - \boldsymbol{H}$ and $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{I}$ (such a matrix always exists).

Then

$$\boldsymbol{U} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{R},$$

and $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$ is a linear combination of residuals.

Further $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{0}$, so that

$$\boldsymbol{U} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{z}\boldsymbol{b} + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\epsilon},$$

and the distribution of $\boldsymbol{U}$ does not depend upon $\boldsymbol{\beta}$, and $\mathrm{E}[\boldsymbol{U}] = \boldsymbol{0}$.

We now derive the distibution of $\boldsymbol{U}$. To do this we consider the transformation from $\boldsymbol{Y} \rightarrow (\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_G) = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}, \boldsymbol{G}^{\mathrm{T}}\boldsymbol{Y})$, where

$$\widehat{\boldsymbol{\beta}}_G = \boldsymbol{G}^{\mathrm{T}}\boldsymbol{Y} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y},$$

is the generalized least squares estimator.

We derive the Jacobian of the transformation. To do this we need the following two facts:

1. $\det(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = \det(\boldsymbol{A}^{\mathrm{T}})\det(\boldsymbol{A}) = \det(\boldsymbol{A})^2$.
2. $\left| \begin{array}{cc} \boldsymbol{T} & \boldsymbol{U} \\ \boldsymbol{V} & \boldsymbol{W} \end{array} \right| = \mid \boldsymbol{T} \mid \mid \boldsymbol{W} - \boldsymbol{V}\boldsymbol{T}^{-1}\boldsymbol{U} \mid .$

Then

$$\begin{aligned} \mid \boldsymbol{J} \mid &= \left| \frac{\partial(\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_G)}{\partial \boldsymbol{Y}} \right| = \mid \boldsymbol{B} \;\; \boldsymbol{G} \mid = \left| \left[\begin{array}{c} \boldsymbol{B}^{\mathrm{T}} \\ \boldsymbol{G}^{\mathrm{T}} \end{array} \right] [\boldsymbol{B} \;\; \boldsymbol{G}] \right|^{1/2} \\ &= \left| \left[\begin{array}{cc} \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} & \boldsymbol{B}^{\mathrm{T}}\boldsymbol{G} \\ \boldsymbol{G}^{\mathrm{T}}\boldsymbol{B} & \boldsymbol{G}^{\mathrm{T}}\boldsymbol{G} \end{array} \right] \right|^{1/2} \\ &= \mid \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} \mid^{1/2} \mid \boldsymbol{G}^{\mathrm{T}}\boldsymbol{G} - \boldsymbol{G}^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{G} \mid^{1/2} \\ &= 1 \times \mid \boldsymbol{G}^{\mathrm{T}}\boldsymbol{G} - \boldsymbol{G}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{G} \mid^{1/2} \\ &= \mid \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} \mid^{-1/2} \neq 0 \end{aligned}$$

which implies that $(\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_g)$ is of full rank $(= N)$. The vector $(\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_G)$ is a linear combination of normals and so is normal.

We have

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_G \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \boldsymbol{J} \mid = p(\boldsymbol{U} \mid \widehat{\boldsymbol{\beta}}_G, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\widehat{\boldsymbol{\beta}}_G \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \boldsymbol{J} \mid$$

and

$$\mathrm{cov}(\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_G) = \mathrm{E}[\boldsymbol{U}(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta})^{\mathrm{T}}] = \boldsymbol{0},$$

and so $\boldsymbol{U}$ and $\widehat{\boldsymbol{\beta}}_G$ are uncorrelated, and since normal therefore independent.

Hence

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{U} \mid \boldsymbol{\alpha}) p(\widehat{\boldsymbol{\beta}}_G \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \boldsymbol{J} \mid .$$

Let $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, be minimal sufficient statistics for which

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}) &\propto p(\boldsymbol{s}_1, \boldsymbol{s}_2 \mid \boldsymbol{\lambda}, \boldsymbol{\phi}) \\ &= p(\boldsymbol{s}_1 \mid \boldsymbol{\lambda}) p(\boldsymbol{s}_2 \mid \boldsymbol{s}_1, \boldsymbol{\lambda}, \boldsymbol{\phi}) \end{aligned}$$

where $\boldsymbol{\lambda}$ is a parameter of interest and $\boldsymbol{\phi}$ are the remaining (nuisance) parameters.

Inference for $\boldsymbol{\lambda}$ may be based on the *marginal* likelihood

$$L_m(\boldsymbol{\lambda}) = p(\boldsymbol{s}_1 \mid \boldsymbol{\lambda}).$$

In the REML context we have $\boldsymbol{s}_1 = \boldsymbol{u}$, $\boldsymbol{s}_2 = \widehat{\boldsymbol{\beta}}_G$, $\boldsymbol{\lambda} = \boldsymbol{\alpha}$, $\boldsymbol{\phi} = \boldsymbol{\beta}$, and $p(\boldsymbol{U} \mid \boldsymbol{\alpha})$ is a marginal likelihood.

Hence

$$p(\boldsymbol{U} \mid \boldsymbol{\alpha}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\widehat{\boldsymbol{\beta}}_G \mid \boldsymbol{\alpha}, \boldsymbol{\beta})} \mid \boldsymbol{J} \mid^{-1} .$$

We have

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = (2\pi)^{-N/2} \mid \boldsymbol{V} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})\right\},$$

and

$$\begin{aligned} p(\widehat{\boldsymbol{\beta}}_G \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) &= (2\pi)^{-(k+1)/2} \mid \boldsymbol{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{x} \mid^{1/2} \\ &\times \exp\left\{-\frac{1}{2}(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{x} (\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta})\right\} \end{aligned}$$

This leads to

$$\begin{aligned} p(\boldsymbol{U} \mid \boldsymbol{\alpha}) &= (2\pi)^{-(N-k-1)/2} \frac{\mid \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x} \mid^{1/2} \mid \boldsymbol{V} \mid^{-1/2}}{\mid \boldsymbol{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{x} \mid^{1/2}} \\ &\times \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G)^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G)\right\} \end{aligned} \tag{14}$$

which does not depend upon $\boldsymbol{B}$, hence we can choose any linear combination of the residuals.

- To summarize: the "data" $\boldsymbol{U}$ (a linear combination of residuals from an OLS fit), has a distribution that depends on $\boldsymbol{\alpha}$ only – this defines a marginal likelihood (the REML likelihood) which may then be maximized as a function of $\boldsymbol{\alpha}$.
- The log marginal (restricted) likelihood is, upto a constant,

  $$l_m(\boldsymbol{\alpha}) = -\frac{1}{2}\log \mid \boldsymbol{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{x} \mid -\frac{1}{2}\log \mid \boldsymbol{V} \mid -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G)^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G).$$

  The profile log-likelihood based on $\boldsymbol{Y}$ is:

  $$l_P(\boldsymbol{\alpha}) = -\frac{1}{2}\log \mid \boldsymbol{V} \mid -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G)^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G),$$

  and so we have the additional term $-\frac{1}{2}\log \mid \boldsymbol{x}^{\mathrm{T}} \boldsymbol{V} \boldsymbol{x} \mid$ that accounts for the degrees of freedom in estimation of $\boldsymbol{\beta}$.
- In terms of computation calculating REML estimators can be carried out with ML code, altered to include the extra term.

- In general, REML estimators have finite sample bias, but they are preferable to ML estimators, particularly for small samples.
- So far as estimation of the variance components are concerned, the asymptotic distribution of the ML/REML estimator is normal, with variance given by Fisher's information.
- Suppose we fit two (nested) models using REML. Different sets of observations are used in each and so we cannot use a likelihood ratio on regression parameters to test whether the smaller model is a valid statistical simplification of the larger model.
- Likelihood ratio tests for variance components are valid.

Implementation of MLE and REML

MLE and REML require iteration between $\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\alpha}}|\widehat{\boldsymbol{\beta}}$.

Originally the *EM algorithm* was used, e.g., Laird and Ware (1982, *Biometrics*). We illustrate for MLE and, for example, suppose $\boldsymbol{E}_i = \boldsymbol{I}_{n_i}\sigma^2$. The "missing data" here are the random effects $\boldsymbol{b}_i$ and the errors $\boldsymbol{\epsilon}_i$.

The M-step: Given $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$, obtain estimates $\widehat{\boldsymbol{\alpha}} = (\widehat{\sigma}^2, \widehat{\boldsymbol{D}})$:

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{\sum_{i=1}^m \boldsymbol{\epsilon}_i^{\mathrm{T}}\boldsymbol{\epsilon}_i}{\sum_{i=1}^m n_i} = \frac{t_1}{N} \\ \widehat{\boldsymbol{D}} &= \frac{1}{m}\sum_{i=1}^m \boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}} = \frac{\boldsymbol{t}_2}{m},\end{aligned}$$

where t_1 and $\boldsymbol{t}_2$ are the sufficient statistics.

The E step: Estimate the sufficient statistics given the current values $\widehat{\boldsymbol{\alpha}}$, via their expected values:

$$\begin{aligned}\widehat{t}_1 &= \mathrm{E}\left[\sum_{i=1}^m \boldsymbol{\epsilon}_i^{\mathrm{T}}\boldsymbol{\epsilon}_i|\boldsymbol{y}_i, \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\alpha}}), \widehat{\boldsymbol{\alpha}}\right] \\ \widehat{\boldsymbol{t}}_2 &= \mathrm{E}\left[\sum_{i=1}^m \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{b}_i|\boldsymbol{y}_i, \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\alpha}}), \widehat{\boldsymbol{\alpha}}\right].\end{aligned}$$

Closed form fixed and random effect estimates are available once we know $\boldsymbol{\alpha}$.

Slow convergence has been reported so that now the *Newton-Raphson method* is more frequently used.

Let $\boldsymbol{\theta}$ be a $p \times 1$ parameter vector containing the variance components, $l(\cdot)$ the log-likelihood, $\boldsymbol{G}$ the $p \times 1$ score vector, and $\boldsymbol{I}^\star(\cdot)$ the $p \times p$ observed information matrix. Then a second order Taylor series expansion of $l(\cdot)$ about $\boldsymbol{\theta}^{(t)}$, the estimate at iteration t gives:

$$\boldsymbol{g}^{(t)}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) + \boldsymbol{G}^{(t)\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^{\mathrm{T}} \boldsymbol{I}^{\star(t)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}),$$

differentiating and setting equal to zero:

$$\frac{\partial \boldsymbol{g}^{(t)}}{\partial \boldsymbol{\theta}} = \boldsymbol{G}^{(t)} + \boldsymbol{I}^{\star(t)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) = \mathbf{0},$$

gives the next estimate

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \{\boldsymbol{I}^{\star(t)}\}^{-1}\boldsymbol{G}^{(t)}.$$

The use of the expected information gives *Fisher's scoring method*.

See Lindstrom and Bates (1988, *JASA*) for details.

Lack of convergence of the algorithm/negative estimates, may sometimes indicate that a poor model is being fitted.

Dental Example

The simplest possible mixed effects model is given by

$$Y_{ij} = \beta_0 + b_i + \beta_1 t_j + \epsilon_{ij},$$

where ϵ_{ij} are iid with $\mathrm{E}[\epsilon_{ij}] = 0$ and $\mathrm{var}(\epsilon_{ij}) = \sigma^2_\epsilon$ and b_i represent random effects with $b_i \sim_{iid} \mathrm{N}(0, \sigma^2_0)$, and represent perturbations for girl i from the population intercept β_0.

Girl-specific intercepts $\beta_{0i} = \beta_0 + b_i$.

We could write b_{0i}, but use b_i for simplicity.

After conditioning on the random effect we have *independent* observations on each girl, we have assumed that allowing the intercepts to vary has removed all within-girl correlation.

The marginal distribution is normal with mean

$$\mathrm{E}[\boldsymbol{Y}|\beta_0, \beta_1, \sigma^2_\epsilon, \sigma^2_0] = \boldsymbol{\mu},$$

where

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_m)^{\mathrm{T}}$$

is $4m \times 1$ vector and

$$\boldsymbol{\mu}_i = (\beta_0 + \beta_1 t_1, \beta_0 + \beta_1 t_2, \beta_0 + \beta_1 t_3, \beta_0 + \beta_1 t_4)^{\mathrm{T}}.$$

The variance is given by

$$\mathrm{var}(\boldsymbol{Y}|\beta_0, \beta_1, \sigma^2_\epsilon, \sigma^2_0) = \boldsymbol{V},$$

where $\boldsymbol{V}$ is the $4m \times 4m$ block diagonal matrix with

$$\boldsymbol{V}_i = \mathrm{var}(\boldsymbol{Y}_i) = \sigma^2[\boldsymbol{J}_{n_i}\rho + \boldsymbol{I}_{n_i}(1-\rho)],$$

with $\sigma^2 = \sigma^2_\epsilon + \sigma^2_0$ and $\rho = \frac{\sigma^2_0}{\sigma^2} = \frac{\sigma^2_0}{\sigma^2_\epsilon + \sigma^2_0}$. Hence the random intercepts model induces a marginal form with constant correlations on measurements on the same child, regardless of the time between observations.

We analyze the dental data using LMEMs. To do this we use the `nlme` package which is described in Pinheiro and Bates (2000) – very flexible, but the syntax is not always obvious...

The `groupedData` function is useful for plotting and modeling (attaches a model function as an attribute to a dataset).

```
> library(nlme)
> data(Orthodont) # Dental data is one of the data sets in the package.
> Orthgirl <- Orthodont[Orthodont$Sex=="Female",]
> trelldat <- groupedData( distance ~ age | Subject, data=Orthgirl )
> plot(trelldat)
```

Figure 6 shows the data plotted using a "trellis" plot – note that data are not plotted in the original order.

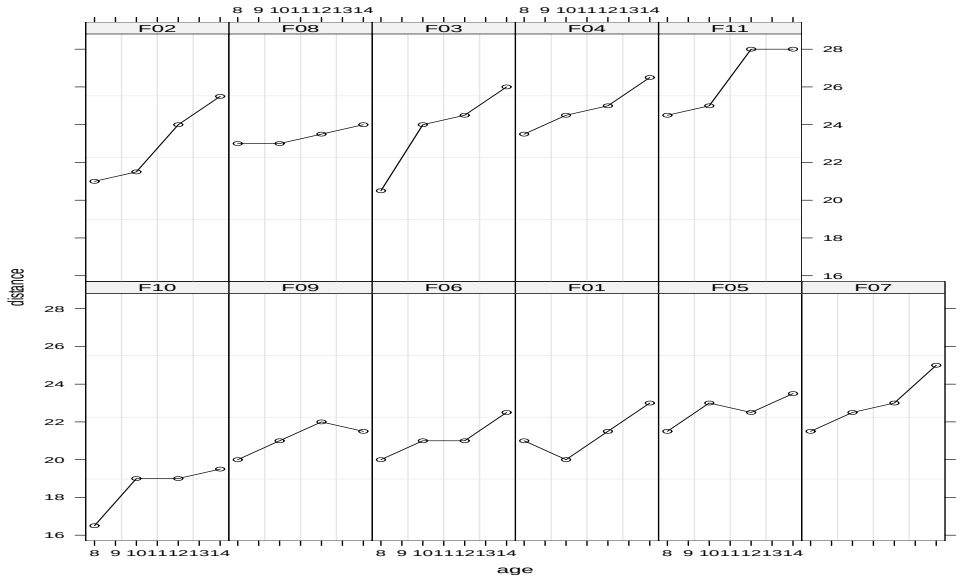

Figure 6: Length versus age (in years) for 11 girls.

We now carry out parameter estimation, first naively, and then using LMEM via REML.

```
> summary(lm(distance~age,data=Orthgirl))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3727     1.6378  10.608 1.87e-13 ***
age           0.4795     0.1459   3.287  0.00205 **
> summary(lme( distance ~ age, data = Orthgirl, random = ~1 | Subject ))
Linear mixed-effects model fit by REML
Random effects:
 Formula: ~1 | Subject
        (Intercept)  Residual
StdDev:     2.06847 0.7800331
Fixed effects: distance ~ age
                Value Std.Error DF   t-value p-value
(Intercept) 17.372727 0.8587419 32 20.230440       0
age          0.479545 0.0525898 32  9.118598       0
```

Notice the standard error for β_1 is smaller for the REML analysis – slopes are being estimated from within-girl comparisons (for the intercept correct standard errors are larger due to dependence).

The REML estimates of the variance components are $\widehat{\sigma}_\epsilon = 0.78$, $\widehat{\sigma}_0 = 2.07$ so that $\widehat{\rho} = 0.875$ which ties in with the empirical correlations (10). The marginal standard deviation is given by $(\widehat{\sigma}_\epsilon^2 + \widehat{\sigma}_0^2)^{1/2} = 2.21$, in agreement with the diagonal elements of (10).

Now for comparison carry out with ML:

```
> summary(lme( distance ~ age, data = Orthgirl, random = ~1 | Subject, method = "ML" )
Linear mixed-effects model fit by maximum likelihood
Random effects:
 Formula: ~1 | Subject
        (Intercept)  Residual
StdDev:    1.969870 0.7681235
Fixed effects: distance ~ age
               Value Std.Error DF   t-value p-value
(Intercept) 17.372727 0.8506287 32 20.423397       0
age          0.479545 0.0530056 32  9.047078       0
```

Note that the MLEs of the variance components are smaller than the REML counterparts. Slight differences in the standard errors of the fixed effects (but not a big difference here).

Bayesian Inference

In the Bayesian approach to inference all *unknown* quantities contained in a probability model for the observed data are treated as random variables.

These unknowns may include, for example, missing data, the true covariate value in an errors-in-variables setting, or the failure time of a censored survival observation.

Inference is made through the *posterior* probability distribution of $\boldsymbol{\theta}$ after observing $\boldsymbol{y}$, and is determined from Bayes theorem:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})}{p(\boldsymbol{y})},$$

where, for continuous $\boldsymbol{\theta}$, the normalizing constant is given by

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta},$$

and is the marginal probability of the observed data given the model (likelihood and prior). Ignoring this constant gives

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \\ \text{Posterior} &\propto \text{Likelihood} \times \text{Prior} \end{aligned}$$

The use of the posterior distribution for inference is very intuitively appealing since it probabilistically combines information on the parameters arising from the data and from prior beliefs.

An important observation is that for all $\boldsymbol{\theta}$ for which $\pi(\boldsymbol{\theta}) = 0$ we have $p(\boldsymbol{\theta} \mid \boldsymbol{y}) = 0$ also, regardless of any realization of the observed data. This has important consequences for prior specification and clearly shows that great care should be taken in excluding parts of the parameter space *a priori*.

Sequential Updating

Suppose first that $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ represent the current totality of data. Then the posterior is given by

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y}_1, \boldsymbol{y}_2)}. \tag{15}$$

Now suppose that we are at a previous time point at which only $\boldsymbol{y}_1$ are available, the posterior in this case is

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1) = \frac{p(\boldsymbol{y}_1 \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y}_1)}.$$

When $\boldsymbol{y}_2$ becomes available, the "prior" for these data corresponds to $p(\boldsymbol{\theta} \mid \boldsymbol{y}_1)$ since it represents the current beliefs concerning $\boldsymbol{\theta}$. We then update via

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{y}_1)}{p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1)}. \tag{16}$$

Identical inference in each case; hence consistent inference is reached regardless of whether we produce the posterior in one stage or two, corresponding to whether all of the data are analyzed simultaneously.

Inference

To summarizes the typically multivariate posterior distribution, $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, marginal distributions for parameters of interest may be considered.

For example the univariate marginal distribution for a component θ_i is given by

$$p(\theta_i \mid \boldsymbol{y}) = \int_{\boldsymbol{\theta}_{-i}} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}_{-i}, \tag{17}$$

where $\boldsymbol{\theta}_{-i}$ is the vector $\boldsymbol{\theta}$ excluding θ_i.

Posterior moments may be evaluated from the marginal distributions; for example the posterior mean is given by

$$\mathrm{E}[\theta_i \mid \boldsymbol{y}] = \int_{\theta_i} \theta_i p(\theta_i \mid \boldsymbol{y}) \, \mathrm{d}\theta_i. \tag{18}$$

Further summarization may be carried out to yield the $100 \times q\%$ quantile, $\theta_i(q)$ $(0 < q < 1)$ by solving

$$\int_{-\infty}^{\theta_i(q)} p(\theta_i \mid \boldsymbol{y}) \, \mathrm{d}\theta_i. \tag{19}$$

In particular, the posterior median, $\theta_i(0.5)$, will often provide an adequate summary of the location of the posterior marginal distribution.

A $100\times p\%$ equi-tailed *credible interval* $(0 < p < 1)$ is provided by $[\ \theta_i\{(1-p)/2\}, \theta_i\{(1+p)/2\}\ ]$.

Such intervals are usually reported though in some cases it which the posterior is skewed one may wish to instead calculate a *highest posterior density* (HPD) interval in which points inside the interval have higher posterior density than those outside the interval (such an interval is also the shortest credible interval).

Another useful inferential quantity is the *predictive* distributions for future observations $\boldsymbol{z}$ which is given, under conditional independence, by

$$p(\boldsymbol{z} \mid \boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y})\ \mathrm{d}\boldsymbol{\theta}. \tag{20}$$

This clearly assumes that the system under study is stable so that the likelihood for future observations is still the relevant data generation mechanism.

Bayesian inference is deceptively simple to describe probabilistically, but there have been two major obstacles to its routine use. The first is how to specify prior distributions and the second is how to evaluate the integrals required for inference, for example, (17)–(20), given that for most models, these are analytically intractable

Simple Example: Suppose we have

$$Y_i|\theta \sim_{i.i.d.} N(\theta, \sigma^2), \quad i = 1, ..., n,$$

with σ^2 assumed known and θ unknown.

Estimation

Recall that the MLE

$$\bar{Y} \sim N\left(\theta, \frac{\sigma^2}{n}\right).$$

Suppose the prior distribution for θ can be described by a normal distribution with mean m and variance v (m and v are known). Then the posterior distribution $p(\theta|\boldsymbol{y})$ is given by

$$N\left(\bar{y} \times w + m \times (1-w), \frac{\sigma^2}{n} \times w\right),$$

where $w = \frac{v}{v+\sigma^2/n}$.

Think about cases: $n = 0$ (recover the prior), $v = 0$ (posterior=prior), $v^{-1} = 0$ (improper prior, frequentist and Bayesian estimates coincide), $n \to \infty$ ($w \to 1$ unless $v = 0$).

One useful way of specifying the prior is as

$$\theta \sim N\left(m, \frac{\sigma^2}{k}\right),$$

in which case k may be regarded as a *prior sample size*. It is 'as if' we carried out an experiment with k observations and we observed a mean of m. This gives $w = n/(n+k)$.

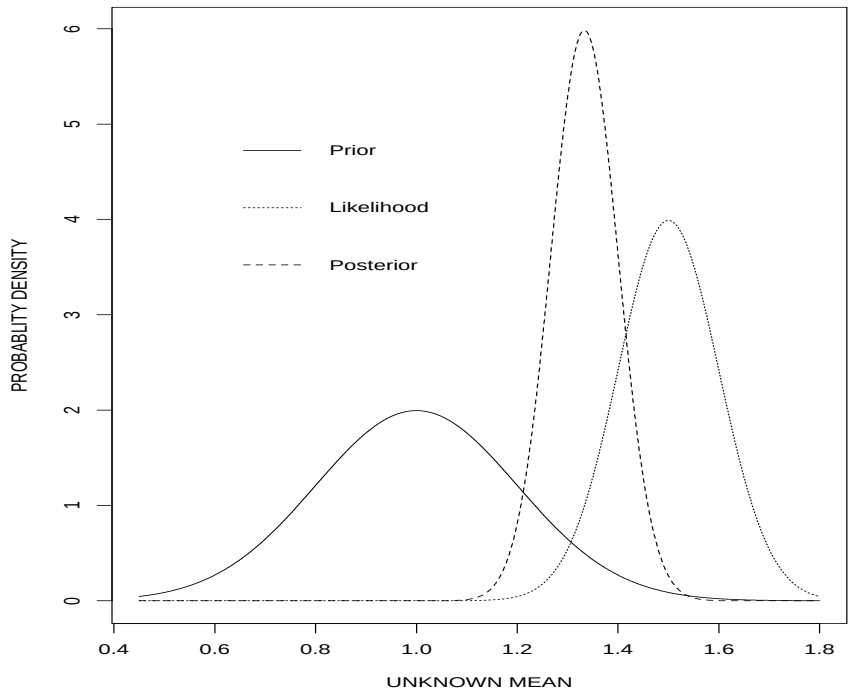

Figure 7: Normal likelihood ($\bar{y}$=1.5,n=10,σ=1), normal prior (m=1, k=5) and the resultant normal posterior.

Prediction

Suppose we wish to obtain the predictive density for a new random variable $Z \sim N(\theta, \sigma^2)$.

Then

$$p(z|\boldsymbol{y}) = \int p(z|\theta) \times p(\theta|\boldsymbol{y}) \mathrm{d}\theta.$$

It may be shown that

$$z|\boldsymbol{y} \sim N\left\{\mathrm{E}[\theta|\boldsymbol{y}], \sigma^2 + \mathrm{var}(\theta|\boldsymbol{y})\right\},$$

so that the mean of the predictive distribution is the posterior mean and the variance is given by the sum of the 'measurement error' and the uncertainty in the posterior mean.

Prior Choice

We distinguish between two prior specification situations. In the first, which we label as a *baseline prior* an analysis is required in which the prior distribution has minimal impact, so that the information in the likelihood dominates the posterior.

The second situation, which we label as a *substantive prior* is one in which it is desired to incorporate more substantial prior information into the analysis.

Baseline Priors

On first consideration it would seem that the specification of a baseline prior is straightforward, one simply takes the choice

$$\pi(\boldsymbol{\theta}) \propto 1 \tag{21}$$

so that the posterior distribution depends solely on the data through the likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$.

There are two difficulties with this.

The first difficulty is that the prior (21) is improper (it does not integrate to a positive constant $< \infty$) unless the range of each element of $\boldsymbol{\theta}$ is finite.

In some instances this is not a problem since the posterior corresponding to the prior is proper. Philosophically a posterior arising from an improper prior may be justified as a limiting case of proper priors. More practically we may instead assume that the prior is integrable over its support but is "locally uniform", so that the likelihood dominates.

For nonlinear models in particular, care must be taken to ensure that the posterior corresponding to a particular prior choice is proper. Some general guidelines are available, for example, improper priors for the regression parameters in a generalized linear model will usually lead to a proper posterior although not for some pathological cases.

For example suppose $Y \mid p \sim \text{Binomial}(n, p)$, and a uniform prior is used on the logit of p, $\log\{p/(1-p)\}$ which implies the prior on p is $\pi(p) = [p(1-p)]^{-1}$. Then an improper posterior results if $y = 0$ (or $y = n$) since the non-integrable spike at $p = 0$ (or $p = 1$) remains in the posterior. For $n = 1$ one of these events will always occur and so an improper posterior always results.

To illustrate the non-propriety in another non-linear situation consider the model

$$Y_i \mid \theta \sim_{ind} N\{\exp(-\theta x_i), \sigma^2\}, \tag{22}$$

$i = 1, ..., n$, with $\theta > 0$ and σ^2 assumed known. With an improper uniform prior on θ we have the posterior

$$p(\theta \mid \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathrm{e}^{-\theta x_i})^2\right\}.$$

As $\theta \to \infty$,

$$p(\theta \mid \boldsymbol{y}) \to \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right\},$$

a constant, so that the posterior is improper.

The second is that if we reparameterize the model in terms of $\boldsymbol{\phi} = \boldsymbol{g}(\boldsymbol{\theta})$ where $\boldsymbol{g}(\cdot)$ is a one-one mapping, then the prior for $\boldsymbol{\phi}$ corresponding to (21) is given by

$$\pi(\boldsymbol{\phi}) = \left| \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\boldsymbol{\phi}} \right|,$$

which, unless $\boldsymbol{g}$ is linear, is not constant.

As an example, consider a variance σ^2, the prior $\pi(\sigma^2) \propto 1$ corresponds to a prior for the standard deviation of $\pi(\sigma) \propto \sigma$; the problem is that we cannot be "flat" on different scales.

This indicates that a desirable property in constructing baseline priors is there invariance to parameterization, so that we obtain the same prior regardless of the starting parameterization. In the example just considered suppose the data are normally distributed with variance σ^2. The improper prior

$$\pi(\sigma) \propto \frac{1}{\sigma}$$

has a number of justifications including invariance to parameterization.

Example: Normal linear regression, variance unknown

Suppose we have $Y_i \mid \boldsymbol{\beta}, \sigma^2 \sim_{ind} N(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2)$, $i = 1, ..., n$. $\dim(\boldsymbol{\beta}) = p$.

MLE: $\widehat{\boldsymbol{\beta}} \sim t_p(\boldsymbol{\beta}, (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}s^2, n-p)$, a Student t distribution with $n-p$ degrees of freedom.

Improper prior: $\pi(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$.

Marginal posterior:

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) = \int p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) d\sigma^2,$$

where

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) \propto l(\boldsymbol{\beta}, \sigma^2) \times \pi(\boldsymbol{\beta}, \sigma^2).$$

Hence

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{y}) &= \int \frac{(2\pi\sigma^2)^{-n/2}}{\sigma^2} \exp\left\{ -\frac{[(n-p)s^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})]}{2\sigma^2} \right\} d\sigma^2 \\ &\propto \int (\sigma^2)^{-(n/2+1)} \exp\left\{ -\frac{c}{2\sigma^2} \right\} d\sigma^2 \end{aligned}$$

where

$$c = (n-p)s^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

We have the kernel of an inverse Gamma distribution $\text{IGa}(n/2, c)$.

An inverse gamma r.v. X has density

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp(-\beta/x), \quad x > 0.$$

Hence

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{y}) &\propto \left(\frac{c}{2}\right)^{-n/2} \\ &\propto \{(n-p)s^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^{-n/2} \\ &\propto \left\{1 + \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{(n-p)s^2}\right\}^{[-(n-p)+p]/2} \\ &= \left\{1 + \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \Sigma^{-1} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{n-p}\right\}^{[-(n-p)+p]/2} \end{aligned}$$

where $\Sigma = (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x})^{-1} s^2$.

Hence the posterior

$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim t_p(\widehat{\boldsymbol{\beta}}, (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x})^{-1} s^2, n-p).$$

A p dimensional multivariate Student's t r.v. $\boldsymbol{X}$ with degrees of freedom d has density

$$p(\boldsymbol{x}) = \frac{\Gamma\{(d+p)/2\}}{\Gamma(d/2)(d\pi)^{p/2}} \mid \boldsymbol{\Sigma} \mid^{-1/2} \times \left[1 + (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})/d\right]^{-(d+p)/2}.$$

Bayesian Justification for REML

Another justification is to assign a flat improper prior to the regression coefficients and then integrate these from the model.

Example: Normal Linear Model

Consider the linear regression for independent data: $\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2 \sim N(\boldsymbol{x\beta}, \boldsymbol{I}_n\sigma^2)$, with $\dim(\boldsymbol{\beta}) = k+1$.

Consider

$$p(\boldsymbol{y}|\sigma^2) = \int p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta})d\boldsymbol{\beta},$$

and assume $\pi(\boldsymbol{\beta}) \propto 1$.

Hence

$$\begin{aligned} p(\boldsymbol{y}|\sigma^2) &= \int (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{x\beta})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x\beta})\right] d\boldsymbol{\beta} \\ &= (2\pi\sigma^2)^{-n/2} \int \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{x}\widehat{\boldsymbol{\beta}}+\boldsymbol{x}\widehat{\boldsymbol{\beta}}-\boldsymbol{x\beta})^{\mathrm{T}}\right. \\ &\times \left.(\boldsymbol{y}-\boldsymbol{x}\widehat{\boldsymbol{\beta}}+\boldsymbol{x}\widehat{\boldsymbol{\beta}}+\boldsymbol{x\beta})\right] d\boldsymbol{\beta} \\ &= (2\pi\sigma^2)^{-(n-k-1)/2} \exp\left[-\frac{RSS}{2\sigma^2}\right] |\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}|^{-1/2} \end{aligned}$$

where the residual sum of squares

$$RSS = (\boldsymbol{y}-\boldsymbol{x}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}\widehat{\boldsymbol{\beta}}).$$

Maximization of $l(\sigma^2) = \log p(\boldsymbol{y}|\sigma^2)$ yields the unbiased estimator

$$\widehat{\sigma}^2 = \frac{RSS}{n-k-1}.$$

Example: LMEM

Again obtain the distribution of the data as a function of $\boldsymbol{\alpha}$ only, by integrating $\boldsymbol{\beta}$ from the model, and assuming an improper flat prior for $\boldsymbol{\beta}$.

We have

$$p(\boldsymbol{y}|\boldsymbol{\alpha}) = \int_{\boldsymbol{\beta}} p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\alpha}) \times \pi(\boldsymbol{\beta}) \, \mathrm{d}\boldsymbol{\beta},$$

leading to

$$\begin{aligned} l(\boldsymbol{\alpha}) &= \log p(\boldsymbol{y}|\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{m} \log |\boldsymbol{V}_i(\boldsymbol{\alpha})| \\ &- \frac{1}{2}\sum_{i=1}^{m} \log |\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\boldsymbol{\alpha}) \boldsymbol{x}_i| - \frac{1}{2}\sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{V}^{-1}(\boldsymbol{\alpha})(\boldsymbol{y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}), \end{aligned}$$

which differs from the "usual" likelihood by the term

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

This expression as the same as that which results from the maximization of the distribution of the residuals.

Estimates of $\boldsymbol{\beta}$ change since they are a function of $\widehat{\boldsymbol{\alpha}}$.

Inference for Random Effects

Examples:

- Pharmacokinetics: individualization of a profile.
- Dairy herds: genetic merit of a particular bull – data are in the form of the milk yields of his daughters.
- Psychology: inference for the IQ of an individual from a set of test scores.
- Industrial applications: operating characteristics of a particular machine.

From a frequentist perspective, inference for random effects is often viewed as *prediction* rather than estimation, since $\boldsymbol{b}$ are random variables.

The usual frequentist optimality criteria for a fixed effect $\boldsymbol{\theta}$, are based upon unbiasedness:

$$\mathrm{E}[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta} = \mathbf{0},$$

where $\boldsymbol{\theta}$ is a fixed constant, and upon the variance of the estimator

$$\mathrm{var}(\widehat{\boldsymbol{\theta}}).$$

These need to be adjusted when inference is required for a random effect $\boldsymbol{b}$.

We wish to find a predictor $\widetilde{\boldsymbol{b}} = f(\boldsymbol{Y})$ of $\boldsymbol{b}$.

An unbiased predictor $\widetilde{\boldsymbol{b}}$ is such that

$$\mathrm{E}_{y,b}[\widetilde{\boldsymbol{b}} - \boldsymbol{b}] = \mathrm{E}[\widetilde{\boldsymbol{b}} - \boldsymbol{b}] = \boldsymbol{0},$$

to give

$$\mathrm{E}[\widetilde{\boldsymbol{b}}] = \mathrm{E}[\boldsymbol{b}]$$

so that the expectation of the predictor is equal to the expectation of the random variable that it is predicting.

The variance of a random variable is defined with respect to a fixed number, the mean. In the context of prediction of a random variability, a more relevant summary of the variability is

$$\mathrm{var}(\widetilde{\boldsymbol{b}} - \boldsymbol{b}) = \mathrm{var}(\widetilde{\boldsymbol{b}}) + \mathrm{var}(\boldsymbol{b}) - 2\mathrm{cov}(\widetilde{\boldsymbol{b}}, \boldsymbol{b}).$$

There are many different criteria that may be used to find a predictor.

Since we are predicting a random variable it is natural to use minimum mean squared error (MSE) as a criteria, rather than minimum variance.

The MSE of $\widetilde{\boldsymbol{b}}$ is given by

$$\mathrm{MSE}(\widetilde{\boldsymbol{b}}) = \mathrm{E}_{y,b}[(\widetilde{\boldsymbol{b}} - \boldsymbol{b})^{\mathrm{T}} \boldsymbol{A} (\widetilde{\boldsymbol{b}} - \boldsymbol{b})],$$

for non-singular $\boldsymbol{A}$.

This leads to $\widetilde{\boldsymbol{b}} = \mathrm{E}[\boldsymbol{b} \mid \boldsymbol{y}]$, irrespective of $\boldsymbol{A}$ (see Exercises 2). Hence the best prediction is that which estimates the random variable by its conditional mean.

We now examine properties of $\widetilde{\boldsymbol{b}}$.

Unbiasedness

We have

$$\mathrm{E}_y[\widetilde{\boldsymbol{b}}] = \mathrm{E}_y\{\mathrm{E}_{b|y}[\boldsymbol{b} \mid \boldsymbol{y}]\} = \mathrm{E}_b[\boldsymbol{b}]$$

where we first step follows on substitution of $\widetilde{\boldsymbol{b}}$ and the second from iterated expectation. (Note: $\mathrm{E}_u[U] = \mathrm{E}_{u,v}[U] = \mathrm{E}_v\{\mathrm{E}_{u|v}[U|V]\}$.)

Variability

Recall an appropriate measure of variability:

$$\text{var}(\widetilde{\boldsymbol{b}}_i - \boldsymbol{b}_i) = \text{var}(\widetilde{\boldsymbol{b}}_i) + \text{var}(\boldsymbol{b}_i) - 2\text{cov}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i).$$

We have

$$\begin{aligned} \text{cov}_{\widetilde{b},b}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i) &= \text{E}_y[\text{cov}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i \mid \boldsymbol{y})] + \text{cov}_y(\text{E}[\widetilde{\boldsymbol{b}}_i \mid \boldsymbol{y}], \text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}]) \\ &= \text{E}_y[\text{cov}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i \mid \boldsymbol{y})] + \text{cov}_y(\widetilde{\boldsymbol{b}}_i, \widetilde{\boldsymbol{b}}_i) \qquad (23) \\ &= \text{var}(\widetilde{\boldsymbol{b}}_i) \end{aligned}$$

The first term in (23) is the covariance between a constant $\text{E}[\widetilde{\boldsymbol{b}} \mid \boldsymbol{y}]$ (since $\boldsymbol{y}$ is conditioned upon), and $\widetilde{\boldsymbol{b}}$, and so is zero (because the covariance between a constant and any quantity is zero). In the second term we have used $\text{E}[\widetilde{\boldsymbol{b}}_i \mid \boldsymbol{y}] = \text{E}[\text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}] \mid \boldsymbol{y}] = \widetilde{\boldsymbol{b}}_i$.

Hence

$$\text{var}(\widetilde{\boldsymbol{b}}_i - \boldsymbol{b}_i) = \text{var}(\boldsymbol{b}_i) - \text{var}(\widetilde{\boldsymbol{b}}_i).$$

Application to the LMEM

The predictor, $\widetilde{\boldsymbol{b}} = \text{E}[\boldsymbol{b} \mid \boldsymbol{y}]$, is a random variable, since it a function of $\boldsymbol{y}$, and so we need to know something about $p(\boldsymbol{b} \mid \boldsymbol{y})$ in order to derive its form.

Definitions: Suppose $\boldsymbol{U}$ is an $n \times 1$ vector of random variables, and $\boldsymbol{V}$ is an $m \times 1$ vector of random variables. Then $\text{cov}(\boldsymbol{U}, \boldsymbol{V}) = \boldsymbol{C}$ is an $n \times m$ matrix with (i,j)-th element $\text{cov}(U_i, V_j)$, $i = 1, ..., n; j = 1, ..., m$. Also $\text{cov}(\boldsymbol{V}, \boldsymbol{U}) = \boldsymbol{C}^{\text{T}}$. Now suppose $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{U}$ where $\boldsymbol{A}$ is an $m \times n$ matrix. Then $\text{cov}(\boldsymbol{U}, \boldsymbol{A}\boldsymbol{U}) = \boldsymbol{W}\boldsymbol{A}^{\text{T}}$ where $\boldsymbol{W} = \text{cov}(\boldsymbol{U})$, and $\text{cov}(\boldsymbol{A}\boldsymbol{U}, \boldsymbol{U}) = \boldsymbol{A}\boldsymbol{W}$.

Consider the LMEM

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{z}\boldsymbol{b} + \boldsymbol{\epsilon},$$

and assume $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are independent and $\boldsymbol{b}_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{\epsilon} \sim \text{N}(\boldsymbol{0}, \sigma^2_\epsilon \boldsymbol{I})$ then, using the above results:

$$\begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{y}_i \end{bmatrix} \sim \text{N}_{q+1+n_i} \left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{x}_i\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{D} & \boldsymbol{D}\boldsymbol{z}_i^{\text{T}} \\ \boldsymbol{z}_i\boldsymbol{D} & \boldsymbol{V}_i \end{bmatrix} \right).$$

since

$$\text{cov}(\boldsymbol{b}_i, \boldsymbol{y}_i) = \text{cov}(\boldsymbol{b}_i, \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i) = \text{cov}(\boldsymbol{b}_i, \boldsymbol{z}_i\boldsymbol{b}_i) = \boldsymbol{D}\boldsymbol{z}_i^{\text{T}},$$

and similarly $\text{cov}(\boldsymbol{y}_i, \boldsymbol{b}_i) = \boldsymbol{z}_i\boldsymbol{D}$.

Using properties of the multivariate normal distribution, the predictor takes the form:

$$\widetilde{\boldsymbol{b}}_i = \mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}_i] = \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) \tag{24}$$

This is known as the best linear unbiased predictor (BLUP), where unbiased refers to it satisfying $\mathrm{E}[\widetilde{\boldsymbol{b}}_i] = \mathrm{E}[\boldsymbol{b}_i]$.

The random effect predictor is a shrinkage estimator since it pulls the data towards zero, as we see in examples later.

The form (24) is not of practical use since it depends on the unknown $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$; instead we use

$$\widetilde{\boldsymbol{b}}_i = \mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}_i] = \widehat{\boldsymbol{D}}\boldsymbol{z}_i^{\mathrm{T}}\widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}). \tag{25}$$

Substitution of $\widehat{\boldsymbol{\beta}}$ is not such a problem (since it is an unbiased estimator, and appears in (24) in a linear fashion), but $\widehat{\boldsymbol{\alpha}}$ is more problematic.

The uncertainty in the prediction is given by

$$\mathrm{var}(\widetilde{\boldsymbol{b}}_i - \boldsymbol{b}_i) \quad = \quad \mathrm{var}(\boldsymbol{b}_i) - \mathrm{var}(\widetilde{\boldsymbol{b}}_i) = \boldsymbol{D} - \mathrm{var}(\widetilde{\boldsymbol{b}}_i)$$

We have

$$\widetilde{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}) = \boldsymbol{K}_i(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}),$$

and

$$\mathrm{var}(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}) = \mathrm{var}(\boldsymbol{Y}_i) + \boldsymbol{x}_i\mathrm{var}(\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^{\mathrm{T}} - 2\mathrm{cov}(\boldsymbol{Y}_i, \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}).$$

Since

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{Y}_i,$$

we have

$$\mathrm{cov}(\boldsymbol{Y}_i, \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}) = \boldsymbol{x}_i(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{x}_i\mathrm{var}(\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^{\mathrm{T}},$$

and so

$$\mathrm{var}(\widetilde{\boldsymbol{b}}_i) = \boldsymbol{K}_i[\mathrm{var}(\boldsymbol{Y}_i) - \boldsymbol{x}_i\mathrm{var}(\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^{\mathrm{T}}]\boldsymbol{K}_i^{\mathrm{T}} = \boldsymbol{K}_i[\boldsymbol{V}_i - \boldsymbol{x}_i\mathrm{var}(\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^{\mathrm{T}}]\boldsymbol{K}_i^{\mathrm{T}}$$

to give

$$\mathrm{var}(\widetilde{\boldsymbol{b}}_i - \boldsymbol{b}_i) = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\boldsymbol{D} + \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{x}_i(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\boldsymbol{D}.$$

The variability of the prediction does not acknowledge the uncertainty in $\widehat{\boldsymbol{\alpha}}$.

We now examine fitted values:

$$\begin{aligned}\widehat{\boldsymbol{Y}}_i &= \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_i\widehat{\boldsymbol{b}}_i \\ &= \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_i\{\boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})\} \\ &= (\boldsymbol{I}_{n_i} - \boldsymbol{z}_i\boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1})\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_i\boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{Y}_i,\end{aligned}$$

a weighted combination of the population profile, and the unit's data.

Note that if $\boldsymbol{D} = \boldsymbol{0}$ we obtain $\widehat{\boldsymbol{Y}}_i = \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}$.

We can also write

$$\widehat{\boldsymbol{Y}}_i = \sigma_\epsilon^2\boldsymbol{V}_i^{-1}\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + (\boldsymbol{I}_{n_i} - \sigma_\epsilon^2\boldsymbol{V}_i^{-1})\boldsymbol{Y}_i$$

so that as $\sigma_\epsilon^2 \to 0$, $\widehat{\boldsymbol{Y}}_i \to \boldsymbol{Y}_i$.

Example: One-way ANOVA

For the simple balanced ANOVA model previously considered

$$\widetilde{b}_i = \frac{n\sigma_0^2}{\sigma_\epsilon^2 + n\sigma_0^2}(\overline{y}_i - \beta_0).$$

In practice we have an estimate $\widehat{\beta}_0$, and the predictor is a weighted combination of the distance $\overline{y}_i - \widehat{\beta}_0$ and zero. Hence for finite n the predictor is biased towards zero (recall our definition of unbiasedness is in terms of $\boldsymbol{b}$).

As $n \to \infty$, $\widetilde{b}_i \to \overline{y}_i - \widehat{\beta}_0$, so that

$$\widehat{\beta}_0 + \widetilde{b}_i \to \overline{y}_i \to \mathrm{E}[Y_i].$$

The form of (24) can be justified in a number of ways, other than MSE.

Rather than assume normality we could consider estimators that are *linear* in $\boldsymbol{y}$. In Exercises 2 we show that this again leads to

$$\widetilde{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}).$$

Hence the best linear predictor is identical to the best predictor under normality.

For general distributions, $\mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}_i]$ is not necessarily linear in $\boldsymbol{y}$. Once we plug $\boldsymbol{\alpha}$ into the BLUP we don't even have a linear predictor.

The BLUP is an empirical Bayes estimator. We should be considering $\mathrm{E}[\boldsymbol{b} \mid \boldsymbol{y}]$, with

$$p(\boldsymbol{b} \mid \boldsymbol{y}) = \int\int p(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{\alpha} \mid \boldsymbol{y})\ d\boldsymbol{\beta}d\boldsymbol{\alpha} = \int\int p(\boldsymbol{b} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{y})p(\boldsymbol{\beta}, \boldsymbol{\alpha} \mid \boldsymbol{y})\ d\boldsymbol{\beta}d\boldsymbol{\alpha},$$

but instead the BLUP is the mean of the distribution

$$p(\boldsymbol{b} \mid \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \boldsymbol{y}),$$

so that rather than integrating over $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, estimates have been conditioned upon.

Example: Dental Growth

We again fit a LMEM with random intercepts only.

```
> remlelm <- lme(distance~I(age-11),data = Orthgirl,random = ~1 | Subject)
> summary(remlelm)
 Formula: ~1 | Subject
        (Intercept)  Residual
StdDev:     2.06847 0.7800331
                Value Std.Error DF t-value p-value
(Intercept) 22.647727 0.6346568 32 35.6850       0
I(age - 11)  0.479545 0.0525898 32  9.1186       0
> coef(remlelm)
    (Intercept) I(age - 11)
F10    18.64240   0.4795455
F09    21.17728   0.4795455
F06    21.17728   0.4795455
F01    21.41869   0.4795455
F05    22.62578   0.4795455
F07    22.98791   0.4795455
F02    22.98791   0.4795455
F08    23.35003   0.4795455
F03    23.71216   0.4795455
F04    24.79853   0.4795455
F11    26.24704   0.4795455
```

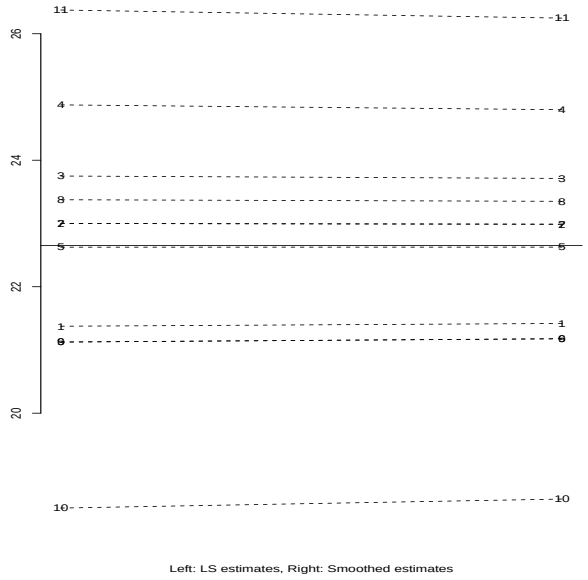

Figure 8: Least squares estimates and smoothed estimates, $\widehat{\beta}_0 + \widetilde{b}_i$; not much shrinkage here since $\widehat{\sigma}_0$ is large relative to $\widehat{\sigma}_\epsilon$.

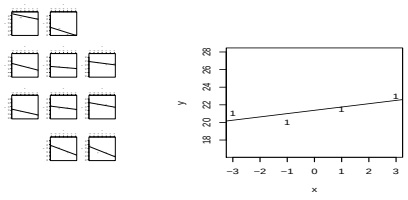

Figure 9: Individual fits: solid line is LS, broken line from LMEM.

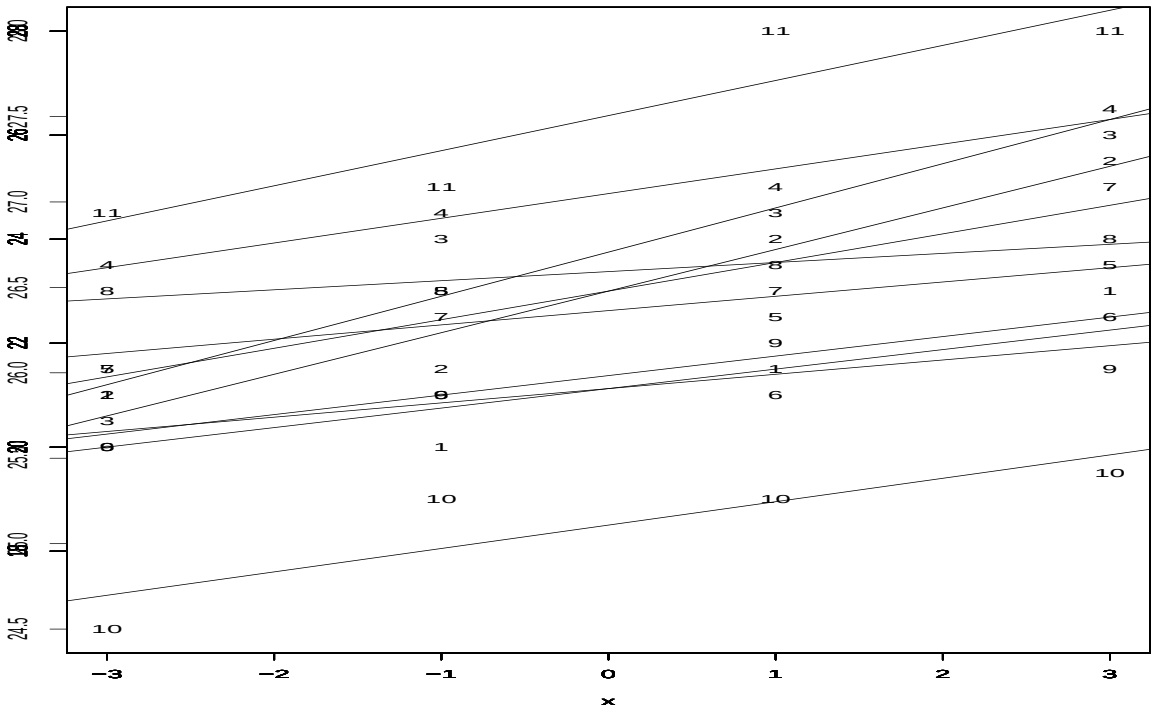

Figure 10: Individual fits: Solid lines are LS, broken line from LMEM.

Dental Example: Boys and Girls Joint Analyses

Table 2 describes a variety of LMEMs to the dental data and Table 3 results.

Model	Description
1	Separate fits, random intercepts
2	Separate fits, random intercepts and slopes, uncorrelated
3	Separate fits, random intercepts and slopes, correlated
4	Combined fit, separate intercepts, common slope, random intercepts
5	Combined fit, separate intercepts and slopes, random intercepts
6	Combined fit, separate intercepts and slopes, random intercepts and slopes, uncorrelated
7	Combined fit, separate intercepts and slopes, random intercepts and slopes, correlated

Table 2: Various LMEMs.

	Boys						Girls					
Model	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\sigma}_0$	$\widehat{\sigma}_1$	$\widehat{\rho}_{01}$	$\widehat{\sigma}_\epsilon$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\sigma}_0$	$\widehat{\sigma}_1$	$\widehat{\rho}_{01}$	$\widehat{\sigma}_\epsilon$
1	25.0	0.78	1.63	–	–	1.68	22.7	0.48	2.07	–	–	0.78
2	25.0	0.78	1.64	0.19	–	1.61	22.6	0.48	2.08	0.16	–	0.67
3	25.0	0.78	1.64	0.19	-0.01	1.61	22.6	0.48	2.08	0.16	0.53	0.67
4	25.0	0.66	1.81	–	–	1.43	22.6	0.66	1.81	–	–	1.43
5	25.0	0.78	1.82	–	–	1.39	22.6	0.48	1.82	–	–	1.39
6	25.0	0.78	1.83	0.18	–	1.31	22.6	0.48	1.83	0.18	–	1.31
7	25.0	0.78	1.83	0.18	0.21	1.31	22.6	0.48	1.83	0.18	0.21	1.31

Table 3: Various LMEM analyses.

R code for models

```
# Set parameterization (to corner point)
> options(contrasts=c("contr.treatment","contr.poly"))
# Separate fits - intercept only, model 1
> remlF <- lme( distance ~ I(age-11), data = Orthgirl, random = ~1 )
> remlM <- lme( distance ~ I(age-11), data = Orthboy, random = ~1 )
# Separate fits - intercept and age, diagonal, model 2
> remlF2d <- lme( distance ~ I(age-11), data = Orthgirl,random = pdDiag(~I(age-11)))
> remlM2d <- lme( distance ~ I(age-11), data = Orthboy,random = pdDiag(~I(age-11)))
# Separate fits - intercept and age, non-diagonal, model 3
> remlF2 <- lme( distance ~ I(age-11), data = Orthgirl, random = ~I(age-11))
> remlM2 <- lme( distance ~ I(age-11), data = Orthboy, random = ~I(age-11))
# Combined fit - common slope, intercept only, model 4
> remlMF <- lme( distance ~ I(age-11)+Sex, data = Orthodont, random = ~1 )
# Combined fit - seperate intercepts and slopes, intercept only - model 5
> remlMFi <- lme( distance ~ I(age-11)+Sex+I(age-11):Sex, data = Orthodont,
              random = ~1 )
# Combined fit -sep intercepts and slopes, uncor random intercepts and slopes - model 6
> remlMF2 <- lme( distance ~ I(age-11)+Sex+I(age-11):Sex, data = Orthodont,
                random=pdDiag(~I(age-11)) )
# Combined fit - sep intercepts and slopes, cor random intercepts and slopes - model 7
> remlMF3 <- lme( distance ~ I(age-11)+Sex+I(age-11):Sex, data = Orthodont,
                random=~I(age-11) )
```


Example of Output (model 4)

```
> summary(remlMF)
Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    1.807425 1.431592
Fixed effects: distance ~ I(age - 11) + Sex
               Value Std.Error DF  t-value p-value
(Intercept) 24.968750 0.4860008 80 51.37595  0.0000
I(age - 11)  0.660185 0.0616059 80 10.71626  0.0000
SexFemale   -2.321023 0.7614168 25 -3.04829  0.0054
 Correlation:
            (Intr) I(-11)
I(age - 11)  0.000
SexFemale   -0.638  0.000
Number of Observations: 108
Number of Groups: 27
```

Figure 11 gives normal QQ plots of the LS estimates of intercepts and slopes, for boys and girls.

Figure 12 gives a scatter plot of the LS estimates of intercepts and slopes, for boys and girls.

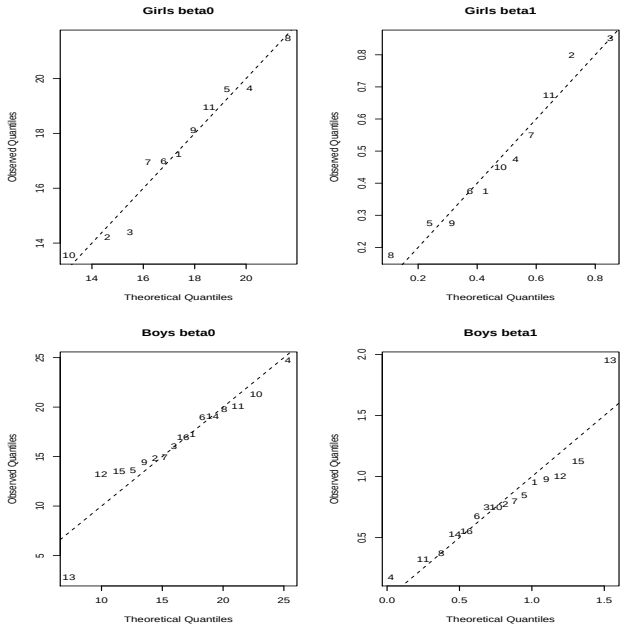

Figure 11: QQ plot of the LS estimates.

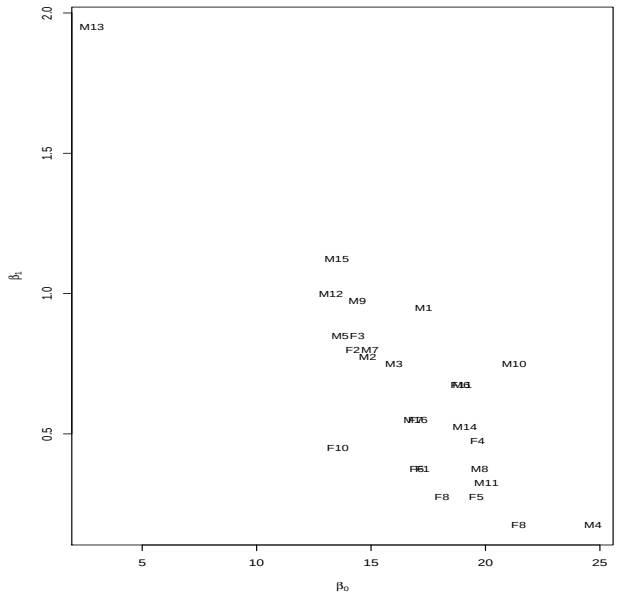

Figure 12: Plot of the LS estimates for boys and girls.

Bayesian Inference for the LMEM

Consider the model

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

with $\boldsymbol{b}_i \sim_{iid} \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{\epsilon}_i \sim_{ind} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_{n_i}\sigma^2_\epsilon)$, with $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ independent.

The form of the posterior follows from exploiting conditional independencies:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b} \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b})\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}) = \prod_{i=1}^m p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}_i)\pi(\boldsymbol{b} \mid \boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\alpha}) \\ &= \prod_{i=1}^m \left\{p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}_i)\pi(\boldsymbol{b}_i \mid \boldsymbol{\alpha})\right\} \pi(\boldsymbol{\beta})\pi(\boldsymbol{\alpha}) \qquad (26) \end{aligned}$$

Alternatively, we can derive the posterior for $\boldsymbol{\beta}, \boldsymbol{\alpha}$ directly:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\alpha} \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha})\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^m p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha})\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &= \prod_{i=1}^m \int p(\boldsymbol{y}_i, \boldsymbol{b}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha})\, d\boldsymbol{b}_i \pi(\boldsymbol{\beta}, \boldsymbol{\alpha}) \end{aligned}$$

where the integrand is giving by the term in curly brackets in (26).

The prior on $\boldsymbol{b}_i$ is justified by the context, formally via *exchangeability*.

Exchangeability

Definition: A finite set $Y_1, ..., Y_n$ of random variables is said to be *exchangeable* if every permutation $(Y_1, ..., Y_n)$ has the same joint distribution as every other permutation. An infinite collection is exchangeable if every finite subcollection is exchangeable.

Every collection of independent and identically distributed random variables is exchangeable.

Theorem: *De Finetti's representation Theorem for 0/1 random variables.*

If $Y_1, Y_2, ...$ is an infinitely exchangeable sequence of 0/1 random variables, there exists a distribution $\pi(\cdot)$ such that the joint mass function $\Pr(y_1, ..., y_n)$ has the form

$$\Pr(y_1, ..., y_n) = \int_0^1 \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}\pi(\theta)\, \mathrm{d}\theta,$$

where

$$\int_0^\theta \pi(u)\, \mathrm{d}u = \lim_{n\to\infty} \Pr\left(\frac{Z_n}{n} \leq \theta\right),$$

with $Z_n = Y_1 + ... + Y_n$, and $\theta = \lim_{n\to\infty} Z_n/n$.

Proof: See Bernardo and Smith (1994) for more details.

Let $z_n = y_1 + ... + y_n$ be the number of 1's (which we label "successes") in the first n observations. Then, due to exchangeability,

$$\Pr(y_1 + ... + y_n = z_n) = \left(\begin{array}{c} n \\ z_n \end{array} \right) \Pr(Y_{\pi(1)}, ..., Y_{\pi(n)}),$$

for all permutations π of $\{1, ..., n\}$ such that $y_{\pi(1)} + ... + y_{\pi(n)} = z_n$. Then we can embed the event $y_1 + ... + y_n = z_n$ within a sequence, $y_1, ..., y_N$, $N \geq n$, and

$$\begin{aligned} \Pr\left(\sum_{i=1}^{n} y_i = z_n\right) &= \sum_{Z_N = z_n}^{N-(n-z_n)} \Pr(y_1 + ... + y_n = z_n, y_1 + ... + y_N = z_N) \\ &= \sum_{z_N = z_n}^{N-(n-z_n)} \Pr(y_1 + ... + y_n = z_n \mid y_1 + ... + y_N = z_N) \\ &\times \Pr(y_1 + ... + y_N = z_N). \end{aligned}$$

To obtain the conditional probability we observe that it is as if we have a population of N people of which z_N are successes, and $N - z_N$ failures, from which we draw n people, the probability of z_n successes is then hypergeometric.

Hence

$$\Pr(y_1 + ... + y_n = z_n) = \sum_{z_N = z_n}^{N-(n-z_n)} \frac{\left(\begin{array}{c} z_N \\ z_n \end{array} \right) \left(\begin{array}{c} N - z_N \\ n - z_n \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} \Pr(z_N)$$

Here $\Pr(z_N)$ is the "prior" belief in the number of successes out of N.

Let $N \to \infty$ and by the strong law of law numbers $\theta = \lim_{N \to \infty} z_N/N$.

The hypergeometric tends to a binomial with parameters n and θ, and the prior $\Pr(z_N)$ is translated into a prior for θ, $\pi(\theta)$. Hence we have

$$\Pr(y_1 + ... + y_n = z_n) \to \left(\begin{array}{c} n \\ z_n \end{array} \right) \int \theta^{z_n} (1 - \theta)^{n - z_n} \pi(\theta) \, \mathrm{d}\theta,$$

as $N \to \infty$.

Implications

The interpretation of this theorem is of great significance:

- We may view the Y_i to be independent, Bernoulli random variables, conditional on a random variable θ.
- θ is itself assigned a probability distribution $\pi()$.
- π may be interpreted as 'beliefs about the limiting relative frequency of 1's'.

In conventional language, we have the *likelihood function*

$$p(Y_1, ..., Y_n|\theta) = \prod_{i=1}^{n} p(Y_i|\theta) = \prod_{i=1}^{n} \theta^{Y_i}(1-\theta)^{1-Y_i},$$

where the *parameter* θ is assigned a *prior distribution* $\pi(\theta)$.

Corollary: If $Y_1, Y_2, ...$ is an infinitely exchangeable sequence of 0/1 random variables, then we have the conditional probability function

$$p(y_{m+1}, ..., y_n \mid y_1, ..., y_m) = \int_0^1 \prod_{i=m+1}^{n} \theta^{Y_i}(1-\theta)^{1-Y_i} \pi(\theta \mid y_1, ..., y_m) \, \mathrm{d}\theta,$$

for $1 \leq m < n$ where

$$\pi(\theta \mid y_1, ..., y_m) = \frac{\prod_{i=1}^{m} \theta^{y_i}(1-\theta)^{1-y_i}\pi(\theta)}{\int_0^1 \prod_{i=1}^{m} \theta^{y_i}(1-\theta)^{1-y_i}\pi(\theta) \, \mathrm{d}\theta}$$

and

$$\int_0^\theta \pi(u) \, \mathrm{d}u = \lim_{n\to\infty} \Pr\left(\frac{z_n}{n} \leq \theta\right).$$

Proof

Write

$$\Pr(y_{m+1}, ..., y_n \mid y_1, ..., y_m) = \frac{\Pr(y_1, ..., y_n)}{\Pr(y_1, ..., y_m)},$$

and then use the previous result on numerator and denominator.

Interpretation: the *prior distribution* $\pi(\theta)$ for θ has been revised, via *Bayes' Theorem*, into the *posterior distribution* $\pi(\theta|y_1, ..., y_m)$.

Further results

General Representation Theorem:

If $Y_1, Y_2, ...$ is an infinitely exchangeable sequence of random variables with probability measure P, there exists a distribution function Q such that the joint mass function $p(Y_1, ..., Y_n)$ has the form

$$p(Y_1, ..., Y_n) = \int \prod_{i=1}^{n} p(Y_i|\boldsymbol{\theta})\pi(\theta)\mathrm{d}\boldsymbol{\theta},$$

with $p(\cdot|\boldsymbol{\theta})$ denoting the density function corresponding to the 'unknown parameter' $\boldsymbol{\theta}$.

Further assumptions on $Y_1, Y_2, ...$ are required to identify $p(\cdot|\boldsymbol{\theta})$.

Relevance of Exchangeability

If we believe *a priori* that $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m$ are exchangeable (and are considered within a hypothetical infinite sequence of such random variables), then it can be shown using representation theorems that the prior can be written in the form

$$p(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m) = \int \prod_{i=1}^{m} p(\boldsymbol{\theta}_i|\boldsymbol{\phi})\pi(\boldsymbol{\phi})\ \mathrm{d}\boldsymbol{\phi},$$

that is, they are conditionally independent, given *hyperparameters* $\boldsymbol{\phi}$, with the hyperparameters having a *hyperprior* distribution.

Hence we have a two-stage (hierarchical) prior:

Stage A: $\boldsymbol{\theta}_i|\boldsymbol{\phi} \sim_{iid} p(\cdot|\boldsymbol{\phi})$, $i = 1, ..., m$.

Stage B: $\boldsymbol{\phi} \sim_{iid} \pi(\cdot)$.

Parametric choices for $p(\cdot|\boldsymbol{\phi})$ and $\pi(\cdot)$ are usually made for computational convenience.

Contrast with the sampling theory approach in which the random effects are assumed to be a random sample from a hypothetical infinite population.

Bayesian Computation

We have seen that to summarize posterior distributions integration is required and, in all but the simplest (conjugate) models, these integrals are not analytically tractable.

Integration is also required to integrate out the random effects in nonlinear mixed effects models, to obtain the likelihood, and later we will review a number of analytical and numerical approaches, for now we concentrate on Markov chain Monte Carlo (MCMC).

The first key idea is the duality between densities and samples from that density: given a density we can always generate samples, and given samples we can reconstruct the density.

Simulation-based techniques have revolutionized Bayesian statistics, by allowing the fitting of very complex models.

Example

Suppose we have $Y_j \mid p_j \sim \text{Binomial}(n_j, p_j)$, $j = 1, 2$, with independent priors $p_j \sim U(0, 1)$. The posteriors are available analytically as $p_j \mid y_j \sim \text{Beta}(y_j + 1, n_j - y_j + 1)$, but suppose we are interested in inference for the odds ratio $\phi = \frac{p_1}{1-p_1} / \frac{p_2}{1-p_2}$ and for the relative risk $\theta = \frac{p_1}{p_2}$.

The following is `R code` to simulate from $\phi \mid y_1, y_2$ when $n_1 = 35, n_2 = 45$, $y_1 = 30, y_2 = 10$:

```
> n1 <- 35; n2 <- 45; y1 <- 30; y2 <- 10
> nsamp <- 1000
> p1 <- rbeta(nsamp,y1+1,n1-y1+1); p2 <- rbeta(nsamp,y2+1,n2-y2+1)
> odds <- (p1/(1-p1))/(p2/(1-p2));  rr <- p1/p2
> par(mfrow=c(2,2))
> hist(p1,xlim=c(0,1))
> hist(p2,xlim=c(0,1))
> hist(odds)
> hist(rr)
> sum(odds[odds>10])/sum(odds) # Posterior prob that odds ratio is > than 10
[1] 0.945683
```

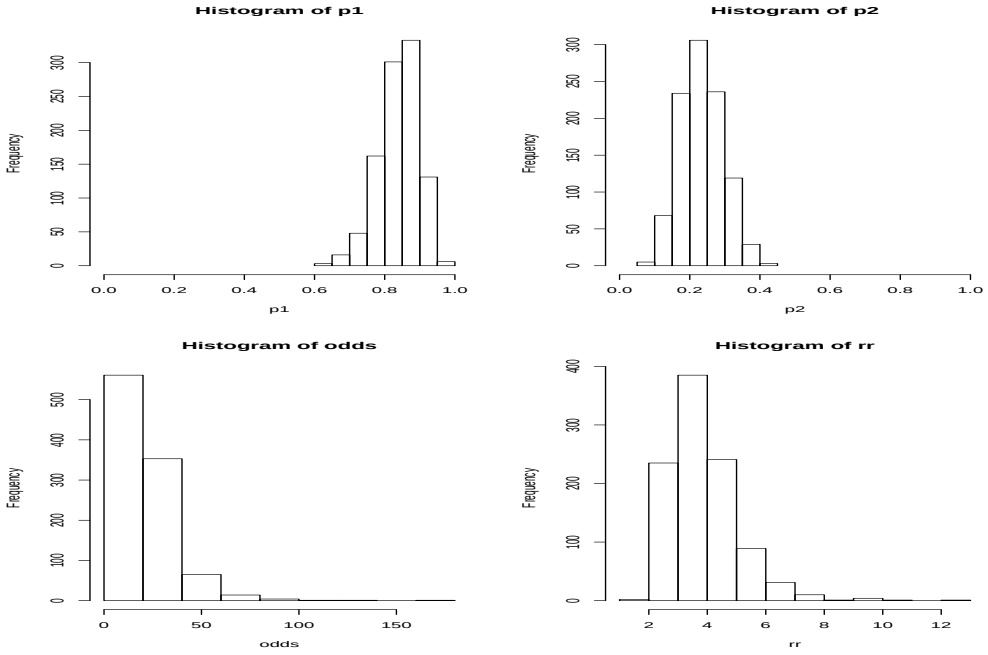

Figure 13: Posterior distributions for p_1, p_2, the odds ratio $\frac{p_1}{1-p_1}/\frac{p_2}{1-p_2}$ and for the relative risk $\theta = \frac{p_1}{p_2}$.

The Composition Method

A useful technical for simulating from joint posterior distributions is the following.

Write the joint posterior distribution for θ_1, θ_2 as

$$p(\theta_1, \theta_2 \mid \boldsymbol{y}) = p(\theta_1 \mid \boldsymbol{y}) p(\theta_2 \mid \theta_1, \boldsymbol{y})$$

Then a simulating algorithm to produce independent samples from $p(\theta_1, \theta_2 \mid \boldsymbol{y})$ is, for $s = 1, ..., S$:

1. Simulate $\theta_1^{(s)} \sim_{ind} p(\theta_1 \mid \boldsymbol{y})$.
2. Simulate $\theta_2^{(s)} \sim_{ind} p(\theta_2 \mid \theta_1^{(s)}, \boldsymbol{y})$.

Markov chain Monte Carlo

MCMC is a very general technique that has revolutionized practical Bayesian statistics.

In the usual derivation of Markov chains over a discrete sample space we are given a transition matrix and the aim is to find the stationary distribution (if it exists). Probabilities of movement depend on the current state only, hence the name.

In the context of sampling from a distribution $\pi(\cdot)$, the aim is to construct a Markov chain whose stationary distribution is π.

Samples $\boldsymbol{\theta}^{(s)}$, $s = 1, ..., S$, produced by a Markov chain "look" more and more like *dependent* samples from π as $S \rightarrow \infty$. The dependency does not cause a problem in terms of estimation since

$$\frac{1}{S}\sum_{s=1}^{S} f(\boldsymbol{\theta}^{(s)}) \rightarrow \mathrm{E}\{f(\boldsymbol{\theta})\},$$

as $S \rightarrow \infty$ (provided the expectation exists).

The only difficulty with the dependency is establishing an appropriates Monte Carlo error on the resultant estimator. We discuss two (related) Markov chains – the Gibbs sampler, and the Metropolis-Hastings algorithm.

Markov chains over a discrete parameter space

Consider a random variable that may take on K values, and consider a Markov chain defined by a $K \times K$ transition matrix $\boldsymbol{P}$.

Then the stationary distribution π is defined by

$$\pi = \pi \boldsymbol{P},$$

where π is a $1 \times K$ row vector.

Roughly speaking, if $\boldsymbol{P}$ is *irreducible* and *aperiodic* (i.e. ergodic) then the stationary distribution is unique.

Gibbs Sampling

Consider a two-parameter problem in which the (intractable) posterior is:

$$\pi(\theta_1, \theta_2|\boldsymbol{y}) \propto l(\theta_1, \theta_2) \times \pi(\theta_1, \theta_2).$$

We have

$$\pi(\theta_1, \theta_2|\boldsymbol{y}) = p(\theta_1|\boldsymbol{y}) \times p(\theta_2|\theta_2, \boldsymbol{y}),$$

but $p(\theta_1|\boldsymbol{y})$ will typically be unavailable.

Gibbs sampling proceeds by iterating between the steps:

$$\theta_1^{(s)} \sim p(\theta_1|\theta_2^{(s-1)}, \boldsymbol{y}),$$

and

$$\theta_2^{(s)} \sim p(\theta_2|\theta_1^{(s)}, \boldsymbol{y}),$$

to produce the sequence

$$(\theta_1^{(0)}, \theta_2^{(0)}), (\theta_1^{(1)}, \theta_2^{(1)}), ..., (\theta_1^{(s)}, \theta_2^{(s)}), ...$$

which may be viewed as a draw from $\pi(\theta_1, \theta_2|\boldsymbol{y})$

Gibbs Sampling over a discrete parameter space

Let $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and suppose that the parameters θ_1 and θ_2 can each take one of the two values, 0 and 1. The posterior distribution is given in Table 4.

$p(\theta_1, \theta_2|\boldsymbol{y})$	$\theta_2 = 0$	$\theta_2 = 1$
$\theta_1 = 0$	π_{00}	π_{01}
$\theta_1 = 1$	π_{10}	π_{11}

Table 4: Joint posterior distribution.

In this case the Gibbs sampler defines a 4×4 transition matrix $\boldsymbol{P}$. The elements of this matrix are given by

$$\begin{aligned}
\Pr\{(i,j),(k,l)\} &= \Pr\{\boldsymbol{\theta}^{(s)} = (k,l)|\boldsymbol{\theta}^{(s-1)} = (i,j)\} \\
&= \Pr(\theta_1^{(s)} = k|\theta_2^{(s)} = j) \Pr(\theta_2^{(s)} = l|\theta_1^{(s)} = k) \\
&= \frac{\pi_{kj}}{\pi_{+j}} \times \frac{\pi_{kl}}{\pi_{k+}}
\end{aligned}$$

It is straigtforward to show that $\boldsymbol{P}$ is such that $\pi = \pi\boldsymbol{P}$.

Example: Normal likelihood, unknown mean and variance

Likelihood:

$$Y_i|\boldsymbol{\beta}, \sigma^2 \sim N(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2), i = 1, ..., n.$$

Prior:

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{V}), \sigma^{-2} \sim \mathrm{Ga}(a, b).$$

Posterior

$$\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto l(\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta})\pi(\sigma^2),$$

is intractable unless $p(\boldsymbol{\beta})$ is improper uniform and the prior for σ^2 is inverse gamma.

Gibbs sampling iterates between $\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2$ and $\sigma^{-2}|\boldsymbol{y}, \boldsymbol{\beta}$ where

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2) &\propto l(\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}) \\ &\sim N(\boldsymbol{\mu}^*, \boldsymbol{V}^*), \\ p(\sigma^{-2}|\boldsymbol{y}, \boldsymbol{\beta}) &\propto l(\boldsymbol{\beta}, \sigma^2)\pi(\sigma^{-2}) \\ &\sim \mathrm{Ga}\left(a + \frac{n}{2}, b + \frac{(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})}{2}\right). \end{aligned}$$

where

$$\boldsymbol{\mu}^* = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\sigma^{-2} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{V}^{-1})^{-1}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\hat{\boldsymbol{\beta}}\sigma^{-2} + \boldsymbol{b}\boldsymbol{V}^{-1}),$$

and

$$\boldsymbol{V}^* = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\sigma^{-2} + \boldsymbol{V}^{-1})^{-1}.$$

Metropolis Algorithm – discrete parameter space

Suppse we have a discrete sample space Ω and we wish to construct a Markov chain whose stationary distribution is $\pi(\cdot)$.

Let $\boldsymbol{Q}$ be an irreducible transition matrix on Ω, satisfying the symmetry condition

$$\boldsymbol{Q}(x, y) = \boldsymbol{Q}(y, x), \quad x, y \in \Omega.$$

We may then define a Markov chain $\{\theta^{(s)}, s = 0, 1, 2, ...\}$ via the following steps.

- Suppose we are currently at state x.
- Generate a proposal from $\boldsymbol{Q}(x, y)$.
- Accept $\theta^{(s+1)} = y$ with probability

$$\min\left(1, \frac{\pi(y)}{\pi(x)}\right),$$

  otherwise stay at x.

This results in the transition matrix

$$\boldsymbol{P}(x, y) = \boldsymbol{Q}(x, y) \times \min\left(1, \frac{\pi(y)}{\pi(x)}\right).$$

Metropolis Algorithm – continuous parameter space

Suppose the stationary distribution is $\pi(\boldsymbol{\theta})$ and consider the *symmetric* probability density function

$$g(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) = g(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a).$$

Suppose $\boldsymbol{\theta}^{(0)}$ denotes the initial point. The Metropolis algorithm then consists of, at iteration s

- Sample $\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s-1)} \sim g(\cdot|\boldsymbol{\theta}^{(s-1)})$.
- Calculate $r = \pi(\boldsymbol{\theta}^*)/\pi(\boldsymbol{\theta}^{(s-1)})$.
- Set

$$\boldsymbol{\theta}^{(s)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min(r, 1), \\ \boldsymbol{\theta}^{(s-1)} & \text{otherwise.} \end{cases}$$

At iteration s the transition density $P(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^{(s-1)})$ is a mixture of $g(\cdot|\boldsymbol{\theta}^{(s-1)})$ and the point $\boldsymbol{\theta}^{(s-1)}$.

Important point: the calculation of r does not depend on the normalizing constant of the target density π.

Metropolis-Hastings Algorithm

Generalizes the Metropolis algorithm to allow a non-symmetric proposal density.

Suppose $\boldsymbol{\theta}^{(0)}$ denotes the initial point. The Metropolis-Hastings algorithm then consists of, at iteration s:

- Sample $\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s-1)} \sim g(\cdot|\boldsymbol{\theta}^{(s-1)})$.
- Calculate
$$r = \frac{\pi(\boldsymbol{\theta}^*)/g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s-1)})}{\pi(\boldsymbol{\theta}^{(s-1)})/g(\boldsymbol{\theta}^{(s-1)}|\boldsymbol{\theta}^*)}.$$
- Set
$$\boldsymbol{\theta}^{(s)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min(r,1), \\ \boldsymbol{\theta}^{(s-1)} & \text{otherwise.} \end{cases}$$

Issues:

- Convergence of the Markov chain?
- Parameterization.

Convergence

- Early iterations $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(m)}$ reflect the (arbitrary) starting value $\theta^{(0)}$.
- These iterations are called the *burn-in*.
- Chain will gradually 'forget' its initial state and converge to the unique stationary distribution which is independent of $\theta^{(0)}$.
- Burn-in samples should be ignored when summarizing the samples for posterior inference via Monte Carlo integration, i.e.
$$E[g(\theta)] \approx \frac{1}{n-m}\sum_{s=m+1}^{n} g(\theta^{(s)})$$

Convergence Diagnosis

- Strictly speaking, convergence is only achieved for $n = \infty$.
- But we only need Markov chain to be 'approaching' convergence for Monte Carlo integration to yield a consistent estimate of the true expectation.
- How do we determine m, the number of 'burn-in' iterations?
- Informal examination of time series plots and running of multiple chains is a must.
- Two issues: have we 'found' the posterior? Do we have enough samples to answer the inferential questions? Some chains may be very slow mixing (examination of autocorrelation is important).

Parameterization

The Markov chain will display better mixing properties if the parameters are approximately independent in the posterior.

In an extreme case, if we have independence then

$$p(\theta_1, ..., \theta_k|\boldsymbol{y}) = \prod_{i=1}^{k} p(\theta_i|\boldsymbol{y}),$$

and Gibbs sampling via the conditional distributions $p(\theta_i|\boldsymbol{y}), i = 1, ..., n$, is equivalent to direct sampling from the posterior.

In general it is better to sample 'blocks' of parameters that are approximately independent.

Hyperpriors

Consider the LMEM

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

with $\boldsymbol{b}_i \sim \mathrm{N}_{q+1}(\boldsymbol{0}, \boldsymbol{D})$, and $\boldsymbol{\epsilon}_i \sim \mathrm{N}_{n_i}(\boldsymbol{0}, \sigma^2_\epsilon \boldsymbol{I}_{n_i})$, $i = 1, ..., m$. A Bayesian analysis requires prior distributions on $\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2_\epsilon$; it is common to assume independent priors

$$\pi(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2_\epsilon) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{D})\pi(\sigma^2_\epsilon).$$

For $\boldsymbol{\beta}$ a multivariate normal distribution and for σ^2_ϵ an inverse gamma distribution are often specified since they lead to conditional distributions of convenient form for Gibbs sampling, but other choices are possible.

If $\boldsymbol{D}$ is a diagonal matrix with elements σ^2_k, $k = 0, 1, ..., q$, then a prior that leads to conjugate conditional distributions in a Gibbs sampling algorithm is

$$\pi(\sigma^2_0, ..., \sigma^2_q) = \prod_{k=0}^{q} \mathrm{IGa}(a_k, b_k),$$

where $\mathrm{IGa}(a_k, b_k)$ denotes the inverse gamma distribution with pre-specified parameters a_k, b_k, $k = 0, ..., q$.

The Wishart Distribution

A prior for a non-diagonal $\boldsymbol{D}$ is more troublesome; there are $(q+2)(q+1)/2$ elements, with the restriction that the resultant matrix is positive definite.

The inverse Wishart distribution is the conjugate choice, and is the only distribution for which any great practical experience has been gained.

Suppose $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_r \sim_{iid} \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{S})$, with $\boldsymbol{S}$ a non-singular variance-covariance matrix, and let

$$\boldsymbol{W} = \sum_{j=1}^{r} \boldsymbol{Z}_j \boldsymbol{Z}_j^{\mathrm{T}}. \tag{27}$$

Then $\boldsymbol{W}$ follows a Wishart distribution, denoted $\mathrm{W}_p(r, \boldsymbol{S})$, and

$$p(\boldsymbol{w}) = c^{-1} \mid \boldsymbol{w} \mid^{(r-p-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{w}\boldsymbol{S}^{-1})\right\}$$

where

$$c = 2^{rp/2}\Gamma_p(r/2) \mid \boldsymbol{S} \mid^{r/2}, \tag{28}$$

with

$$\Gamma_p(r/2) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma((r+1-j)/2)$$

the generalized gamma function, and $r \geq p$ for a proper density.

The mean is given by

$$\mathrm{E}[\boldsymbol{W}] = r\boldsymbol{S}.$$

The Wishart distribution is a multivariate version of the gamma distribution. Taking $p = 1$ yields

$$p(w) = \frac{(2S)^{-r/2}}{\Gamma(r/2)} w^{r/2-1} \exp(-w/2S),$$

for $w > 0$, the gamma distribution $\mathrm{Ga}(r/2, 1/(2S))$. Further, taking $S = 1$ gives a χ^2_r random variable, which is clear from (27).

The Inverse Wishart Distribution

If $\boldsymbol{W} \sim \mathrm{W}_p(r, \boldsymbol{S})$, the distribution of $\boldsymbol{D} = \boldsymbol{W}^{-1}$ is known as the inverse Wishart distribution, and is given by

$$p(\boldsymbol{d}) = c^{-1} \mid \boldsymbol{d} \mid^{-(r+p+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{d}^{-1}\boldsymbol{S})\right\}$$

where c is again given by (28). The mean is given by

$$\mathrm{E}[\boldsymbol{D}] = \frac{\boldsymbol{S}^{-1}}{r-p-1}$$

and is defined for $r > p + 1$. If $p = 1$ we recover the inverse gamma distribution $\mathrm{IGa}(r/2, 1/2S)$ with $\mathrm{E}[D] = 1/[s(r-2)]$ and $\mathrm{var}(D) = 1/[S^2(r-2)(r-4)]$ (so that small r gives a larger spread).

Thinking ahead to application in the LMEM if $\boldsymbol{W} \sim \mathrm{W}_{q+1}(r, \boldsymbol{R}^{-1})$, then

$$\mathrm{E}[\boldsymbol{W}] = r\boldsymbol{R}^{-1},$$

and

$$\mathrm{E}[\boldsymbol{D}] = \boldsymbol{R}/(r-q-1-1),$$

so that $\boldsymbol{R}$, may be scaled to be a prior estimate of $\boldsymbol{D}$, with r acting as a strength of belief in the prior.

Issues with the Wishart Prior

- A problem with the Wishart distribution is that it is deficient in second moment parameters since there is only a single degrees of freedom parameter r. So, for example, it is not possible to have differing levels of certainty in the tightness of the prior distribution for different elements of $\boldsymbol{D}$. With diagonal $\boldsymbol{D}$ and independent inverse gamma priors we have a precision parameter for each variance.
- The form of the conditional distribution suggests that it may be better to err on the side of picking $\boldsymbol{R}$ too small (if m small, prior always influential).
- Intuition: as if our prior data for the precision consists of observing r normal random variables with variance-covariance matrices $\boldsymbol{R}$.
- We need to take $r \geq q+1$ for a proper prior, with the flattest prior corresponding to $r = q+1$. A proper prior is required to ensure propriety of the posterior distribution.
- Figure 14 displays samples from the Wishart distribution $W_2\{20, (20\boldsymbol{S})^{-1}\}$ where $\boldsymbol{S} = \begin{bmatrix} 0.4 & 0 \\ 0 & 1.0 \end{bmatrix}$. The mean is $\mathrm{E}[\boldsymbol{W}] = \boldsymbol{S}^{-1} = \begin{bmatrix} 2.5 & 0 \\ 0 & 1.0 \end{bmatrix}$.

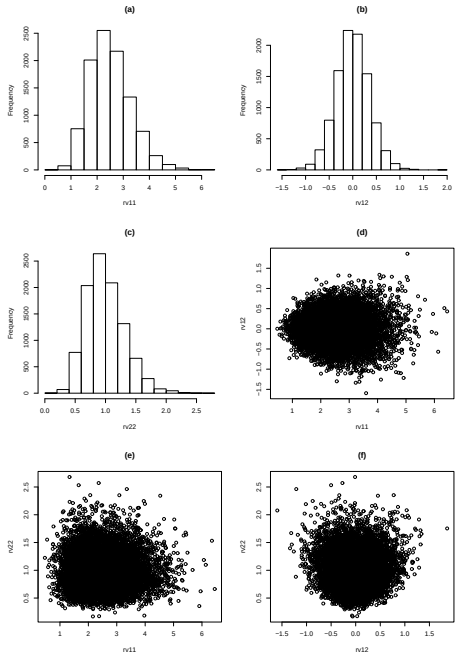

Figure 14: Histograms of (a) w_{11}, (b) w_{12}, (c) w_{22}, scatterplots of (d) w_{11}, w_{12}, (e) w_{11}, w_{22}, w_{12}, w_{22}

Conditional Conjugacy

We now consider a Gibbs sampling scheme and assume for simplicity that $\boldsymbol{x}_i = \boldsymbol{z}_i$. It is computational more convenient to reparameterize in terms of the set $\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \tau, \boldsymbol{\beta}, \boldsymbol{W}\}$ where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{b}_i$, $\tau = \sigma_\epsilon^{-2}$, $\boldsymbol{W} = \boldsymbol{D}^{-1}$.

The joint posterior is

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

with priors:

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

and derive the required conditional distributions:

- $p(\boldsymbol{\beta} \mid \tau, \boldsymbol{W}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{y})$
- $p(\tau \mid \boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{y})$
- $p(\boldsymbol{W} \mid \boldsymbol{\beta}, \tau, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{y})$
- $p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \tau, \boldsymbol{W}, \boldsymbol{y})$, $i = 1, ..., m$.

Conditional for $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{W} \sim \mathrm{N}_{q+1} \left\{ \left(m\boldsymbol{W} + \boldsymbol{V}_0^{-1} \right)^{-1} \left(\boldsymbol{W} \sum_{i=1}^{m} \boldsymbol{\beta}_i + \boldsymbol{V}_0^{-1} \boldsymbol{\beta}_0 \right), \left(m\boldsymbol{W} + \boldsymbol{V}_0^{-1} \right)^{-1} \right\}$$

Conditional for τ

$$\tau \mid \boldsymbol{\beta}_i, \boldsymbol{y} \sim \mathrm{Ga} \left(a_0 + \frac{\sum_{i=1}^{m} n_i}{2}, b_0 + \frac{1}{2} \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}_i)^{\mathrm{T}} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}_i) \right)$$

Conditional for $\boldsymbol{\beta}_i$

$$\boldsymbol{\beta}_i \mid \tau, \boldsymbol{W}, \boldsymbol{y} \sim \mathrm{N}_{q+1} \left\{ (\tau \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i + \boldsymbol{W})^{-1} (\tau \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{y}_i + \boldsymbol{W}\boldsymbol{\beta}), (\tau \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i + \boldsymbol{W})^{-1} \right\}$$

Note the way that the conditional independencies have been exploited so that in each case we condition on only a subset of the parameters.

Conditional for $\boldsymbol{W}$

First note that

$$(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{\beta}_i - \boldsymbol{\beta}) = \mathrm{tr}((\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{\beta}_i - \boldsymbol{\beta})) = \mathrm{tr}(\boldsymbol{W} (\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}}).$$

Then

$$\begin{aligned}
\boldsymbol{W} \mid \boldsymbol{y}, \boldsymbol{\beta}_i, \boldsymbol{\beta} &\propto \prod_{i=1}^{m} p(\boldsymbol{\beta}_i \mid \boldsymbol{W}) \times \pi(\boldsymbol{W}) \\
&\propto \mid \boldsymbol{W} \mid^{(m+r-q-1-1)/2} \exp\left\{ -\frac{1}{2} \left[\sum_{i=1}^{m} (\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{\beta}_i - \boldsymbol{\beta}) + \mathrm{tr}(\boldsymbol{W}\boldsymbol{R}) \right] \right\} \\
&= \mid \boldsymbol{W} \mid^{(m+r-q-1-1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr} \left(\boldsymbol{W} \left[\sum_{i=1}^{m} (\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} + \boldsymbol{R} \right] \right) \right\}
\end{aligned}$$

Hence the conditional distribution is

$$\boldsymbol{W} \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{\beta}, \boldsymbol{y} \sim \mathrm{W}_{q+1} \left\{ r + m, \left(\boldsymbol{R} + \sum_{i=1}^{m} (\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} \right)^{-1} \right\}.$$

Example: Dental Data for Girls

Three-Stage Hierarchical Model:

First Stage:

$$y_{ij} = \beta_{0i} + \beta_{1i}(t_j - 11) + \epsilon_{ij},$$

with $\epsilon_{iid} \sim \mathrm{N}(0, \tau^{-1})$, $j = 1, ..., 4$, $i = 1, ..., 11$.

Second Stage: Let

$$\boldsymbol{\beta}_i = \begin{bmatrix} \beta_{0i} \\ \beta_{1i} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \boldsymbol{D} = \begin{bmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{bmatrix},$$

and then

$$\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D} \sim \mathrm{N}_2(\boldsymbol{\beta}, \boldsymbol{D}),$$

$i = 1, ..., m$.

Third Stage:

$$\pi(\tau, \boldsymbol{\beta}, \boldsymbol{D}^{-1}) \propto \mathrm{Ga}(0, 0) \times \mathrm{N}_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10^6 & 0 \\ 0 & 10^6 \end{bmatrix} \right) \times W_2(r, \boldsymbol{R}^{-1}).$$

Results below are for priors, with prior mean

$$\mathrm{E}[\boldsymbol{D}] = \frac{1}{r-q-2}\boldsymbol{R} = \frac{1}{r-3}\boldsymbol{R} = \begin{bmatrix} 1.0 & 0 \\ 0 & 0.1 \end{bmatrix}$$

(since $q = 1$) and different degrees of freedom r.

We see sensitivity to the prior in inference for $\boldsymbol{D}$, but not for $\boldsymbol{\beta}$.

Note the greater shrinkage to the prior mean for the second and third priors.

r	$\boldsymbol{R}$	β_0	β_1
4	1.0 0 0 0.1	22.6 (21.4,23.8)	0.48 (0.33,0.63)
7	4.0 0 0 0.4	22.6 (21.5,23.7)	0.48 (0.31,0.65)
28	25 0 0 2.5	22.6 (21.8,23.5)	0.48 (0.28,0.67)

Table 5: Posterior medians and 95% intervals for population means, under three priors.

r	Diag $\boldsymbol{R}$	D_{00}	D_{01}	D_{11}
4	1.0 0.1	3.48 (1.66, 8.75)	0.13 (-0.10,0.54)	0.03 (0.01,0.10)
7	4.0 0.4	2.97 (1.51, 6.63)	0.10 (-0.14,0.46)	0.05 (0.02,0.12)
28	25 2.5	1.78 (1.14, 2.97)	0.04 (-0.10,0.20)	0.08 (0.05,0.14)

Table 6: Posterior medians and 95% intervals for population variances, under two priors.

The code below is for the analysis with $r = 4$, BUGS parametrizes the Wishart in terms of $\boldsymbol{R}^{-1}$ and r.

```
model
{
for( i in 1 : N ) {
    for( j in 1 : T ) {
        Y[i , j] ~ dnorm(mu[i , j],eps.tau)
        mu[i , j] <- beta[i,1] + beta[i,2] * (x[j]-11)
    }
    beta[i,1:2] ~ dmnorm(beta.mu[1:2],iSigma[1:2,1:2])
}
beta.mu[1:2] ~ dmnorm(mean[1:2], prec[1:2, 1:2])
iSigma[1:2, 1:2] ~ dwish(R[1:2, 1:2], r)
Sigma[1:2, 1:2] <- inverse(iSigma[1:2, 1:2])
eps.tau <- exp(logtau)
logtau ~ dflat()
sigma <- 1 / sqrt(eps.tau)
}
```



```
list(x = c(8,10,12,14), N = 11, T = 4,
Y = structure(
.Data =   c(21,20,21.5,23,
21,21.5,24,25.5,
20.5,24,24.5,26,
23.5,24.5,25,26.5,
21.5,23,22.5,23.5,
20,21,21,22.5,
21.5,22.5,23,25,
23,23,23.5,24,
20,21,22,21.5,
16.5,19,19,19.5,
24.5,25,28,28),
.Dim = c(11,4)),mean = c(0, 0),r=4,
R = structure(.Data = c(1, 0, 0, 0.1),
.Dim = c(2, 2)),
prec = structure(.Data = c(1.0E-6, 0,0,1.0E-6),
.Dim = c(2, 2))))
list(beta = structure(.Data = c(18,.5,18,.5,18,.5,18,.5,18,.5,18,.5,18,.5,18,
.5,18,.5,18,.5,18,.5), .Dim=c(11,2)), beta.mu = c(18,.5),
     iSigma = structure(.Data = c(1, 0, 0, 0.1), .Dim = c(2, 2)), logtau = 0)
```

Quasi-Likelihood

We now describe a method for inference, generalized estimating equations, that attempts to make minimal assumptions about the data-generating process.

We begin with a recap of the related quasi-likelihood procedure, which is an alternative to MLE, when we do not wish to commit to specifying the full distribution of the data and we can assume independent data. The resultant estimators are known as quasi-MLE (QMLE).

The approach is based on specifying the first two moments of the data only, and assuming they take the form:

$$\begin{aligned} \mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{\beta}] &= \boldsymbol{\mu}(\boldsymbol{\beta}) \\ \mathrm{cov}(\boldsymbol{Y} \mid \boldsymbol{\beta}) &= \alpha \boldsymbol{V}\{\boldsymbol{\mu}(\boldsymbol{\beta})\} \end{aligned}$$

where $\boldsymbol{\mu}(\boldsymbol{\beta}) = [\mu_1(\boldsymbol{\beta}), ..., \mu_n(\boldsymbol{\beta})]^{\mathrm{T}}$ represents the regression function and $\boldsymbol{V}$ is a diagonal matrix (so the observations are uncorrelated), with

$$\mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha V\{\mu_i(\boldsymbol{\beta})\},$$

and $\alpha > 0$ a scalar which is independent of $\boldsymbol{\beta}$.

Consider the sum of squares

$$(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})/\alpha, \tag{29}$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta})$. To minimize this sum of squares there are two ways to proceed.

First approach: differentiate and obtain

$$-2\boldsymbol{D}^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})/\alpha + (\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}} \frac{\partial \boldsymbol{V}^{-1}}{\partial \boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{\mu})/\alpha,$$

where $\boldsymbol{D}$ is the $n \times p$ matrix of derivatives with elements $\partial \mu_i / \partial \beta_j$, $i = 1, ..., n; j = 1, ..., p$. Unfortunately the expectation of this expression is not zero, and so an inconsistent estimator of $\boldsymbol{\beta}$ will result.

Second approach: pretend $\boldsymbol{V}$ is not a function of $\boldsymbol{\beta}$, so that $\widehat{\boldsymbol{\beta}}$ is the root of:

$$\boldsymbol{D}(\widehat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{V}(\widehat{\boldsymbol{\beta}})^{-1} \{\boldsymbol{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\}/\alpha = \boldsymbol{0}.$$

As shorthand we write this estimating function as

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\{\boldsymbol{Y}-\boldsymbol{\mu}\}/\alpha. \tag{30}$$

This estimating function is linear in the data and so its properties are straightforward to evaluate. In particular:

1. $\mathrm{E}[\boldsymbol{U}(\boldsymbol{\beta})] = \boldsymbol{0}$.
2. $\mathrm{cov}\{\boldsymbol{U}(\boldsymbol{\beta})\} = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha$.
3. $-\mathrm{E}\left[\frac{\partial U}{\partial \beta}\right] = \mathrm{cov}\{\boldsymbol{U}(\boldsymbol{\beta})\} = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha$.

Applying the earlier result on properties of estimators arising from estimating functions:

$$(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \rightarrow_d \ N_p(\boldsymbol{0}, \alpha\boldsymbol{I}_p),$$

where we have so far assumed that α is known.

Since the root of (30) does not depend on α, $\widehat{\boldsymbol{\beta}}$ is consistent regardless. For appropriate standard errors we require an estimator of α however.

Unknown α

Since

$$\mathrm{E}[(\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})] = n\alpha,$$

an unbiased estimator of α would be

$$\widehat{\alpha} = (\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})/n,$$

a degrees of freedom corrected (but not in general, unbiased) estimate is given by the Pearson statistic divided by its degrees of freedom:

$$\widehat{\alpha} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i-\widehat{\mu}_i)^2}{V(\widehat{\mu}_i)},$$

where $\widehat{\mu}_i = \widehat{\mu}_i(\widehat{\boldsymbol{\beta}})$.

The asymptotic distribution that is used in practice is therefore given by

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{D}}/\widehat{\alpha})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \rightarrow_d \ N_p(\boldsymbol{0}, \boldsymbol{I}_p),$$

In general we may use sandwich estimation with quasi-likelihood. We have

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathrm{var}(\boldsymbol{Y})\boldsymbol{V}^{-1}\boldsymbol{D}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\alpha^2,$$

and $\mathrm{var}(\boldsymbol{Y})$ may be estimated by the diagonal matrix with elements $(Y_i-\widehat{\mu}_i)^2$.

Why "Quasi"

Integration of the quasi-score (30) gives

$$l(\mu, \alpha) = \int_y^\mu \frac{y-t}{\alpha V(t)} \mathrm{d}t$$

which, if it exists, behaves like a log-likelihood. As an example, for the model $\mathrm{E}[Y] = \mu$ and $\mathrm{var}(Y) = \alpha\mu$ we have

$$l(\mu, \alpha) = \int_y^\mu \frac{y-t}{\alpha t} \mathrm{d}t = \frac{1}{\alpha}[y \log \mu - \mu + c],$$

where $c = -y \log y - y$ and $y \log \mu - \mu$ is the log likelihood of a Poisson random variable.

The word "quasi" refers to the fact that the score may or not correspond to a probability function.

For example, the variance function $\mu^2(1-\mu)^2$ does not correspond to a probability distribution.

Example: Air Pollution Data

We examine the association between daily mortality, Y_i, and the daily value of PM_{10} (particulate matter less than 10 micrometers, which is about 0.0004 inches, in diameter), x_i, with $i = 1, ..., 335$, indexing the 335 days on which there are no missing PM_{10} is to be investigated.

Figure 15 shows the association between log daily counts and PM_{10}.

Assume the model

$$\mathrm{E}[Y_i \mid \boldsymbol{\beta}] = \exp(\boldsymbol{x}_i\boldsymbol{\beta}), \quad \mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha E[Y_i \mid \boldsymbol{\beta}].$$

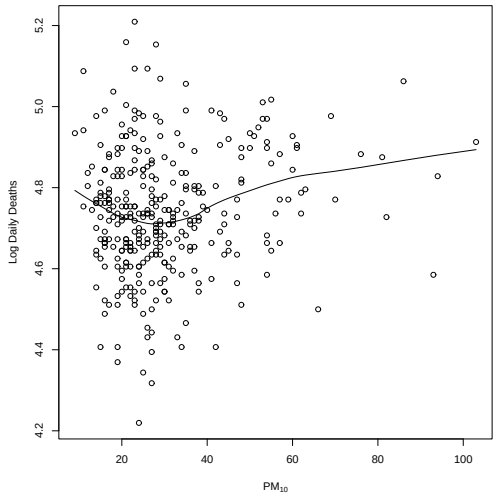

Figure 15: Log daily deaths versus PM_{10}.

Fitting the quasi-likelihood model yields $\widehat{\boldsymbol{\beta}} = (4.71, 0.0015)^{\mathrm{T}}$ and $\widehat{\alpha} = 2.77$ so that the quasi-likelihood standard errors are $\sqrt{\widehat{\alpha}} = 1.67$ times larger than the Poisson model-based standard errors.

The variance-covariance matrix is given by

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{D}})^{-1}\widehat{\alpha} = \begin{bmatrix} 0.019^2 & \star \\ -0.89 \times 0.019 \times 0.00056 & 0.00056^2 \end{bmatrix}.$$

Standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are 0.019 and 0.00056.

Asymptotic 95% confidence interval for β_1 is given by $(0.00040, 0.0026)$.

A more useful summary is a confidence interval for the relative risk associated with a 10-unit increase in PM_{10}, which is

$$(e^{0.00040\times 10}, e^{0.026\times 10}) = (1.004, 1.026)$$

so that the interval suggests that the increase in daily mortality associated with a 10-unit increase in PM_{10} is between 0.4% and 2.6%.

Extension to Quasi-Likelihood

Suppose we have

$$\begin{aligned} \mathrm{E}[Y_i \mid \boldsymbol{\beta}] &= \mu_i(\boldsymbol{\beta}) \\ \mathrm{var}(Y_i \mid \boldsymbol{\beta}) &= V_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) \end{aligned}$$

where $\boldsymbol{\alpha}$ is a $k \times 1$ vector of parameters that appear only in the variance model.

Previously, in quasi-likelihood method, we had "separable" mean and variance models, that is, $\mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha V_i(\mu_i)$ (which is why we obtained a consistent estimator even if the form of the variance was wrong).

Let $\widehat{\boldsymbol{\alpha}}_n$ be a consistent estimator of $\boldsymbol{\alpha}$. We state without proof the following result. The estimator $\widehat{\boldsymbol{\beta}}_n$ that satisfies

$$\boldsymbol{G}(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\alpha}}_n) = \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_n)^{\mathrm{T}} \boldsymbol{V}^{-1}(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n) \left\{ \boldsymbol{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}_n) \right\} \tag{31}$$

has asymptotic distribution

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}} \widehat{\boldsymbol{V}}^{1/2} \widehat{\boldsymbol{D}})^{-1} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \rightarrow_d \ N_p(\boldsymbol{0}, \boldsymbol{I}_p) \tag{32}$$

where $\widehat{\boldsymbol{D}} = \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_n)$ and $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n)$. Sandwich estimation may be used to obtain empirical standard errors which are correct even if the variance model is wrong, so long as we have a consistent estimator of α.

Computation

Previously we assumed $\mathrm{var}(Y_i) = \alpha V_i(\mu_i)$, and the estimating function did not depend on α and so, correspondingly, $\widehat{\boldsymbol{\beta}}$ did not depend on α, though the standard errors did.

In general iteration is convenient to simultaneously estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Let $\widehat{\boldsymbol{\alpha}}^{(0)}$ be an initial estimate.

Then set $j = 0$ and iterate between

1. Solve $\boldsymbol{G}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}^{(j)}) = \boldsymbol{0}$ to give $\widehat{\boldsymbol{\beta}}^{(j+1)}$,
2. Estimate $\widehat{\boldsymbol{\alpha}}^{(j+1)}$ with $\widehat{\mu}_i = \mu_i \left(\widehat{\boldsymbol{\beta}}^{(j+1)} \right)$. Set $j \rightarrow j + 1$ and return to 1.

Example: Air Pollution Data

Consider the random effects formulation:

$$\mathrm{E}[Y_i \mid \boldsymbol{\beta}, \theta_i] = \mathrm{var}(Y_i \mid \boldsymbol{\beta}, \theta_i) = \mu_i(\boldsymbol{\beta})\theta_i \tag{33}$$

with

$$\mathrm{E}[\theta_i] = 1, \quad \mathrm{var}(\theta_i) = 1/\alpha. \tag{34}$$

Assuming $\theta_i \sim_{iid} \mathrm{Ga}(\alpha, \alpha)$, we could derive the marginal distribution of the data (which is negative binomial) and proceed with likelihood.

As an alternative we consider the model

$$\begin{aligned} \mathrm{E}[Y_i \mid \boldsymbol{\beta}] &= \mu_i(\boldsymbol{\beta}) \\ \mathrm{var}(Y_i \mid \alpha, \boldsymbol{\beta}) &= \mu_i(\boldsymbol{\beta})\{1 + \mu_i(\boldsymbol{\beta})/\alpha\}. \end{aligned} \tag{35}$$

that are the marginal first two moments of the data given (33) and (34).

The form (35) suggests the estimating function for $\boldsymbol{\beta}$ (with α assumed known):

$$\sum_{i=1}^{n} \boldsymbol{D}(\boldsymbol{\beta})_i^{\mathrm{T}} \boldsymbol{V}_i^{-1}(\alpha, \boldsymbol{\beta})\{y_i - \mu_i(\boldsymbol{\beta})\}$$

For a fixed α we can solve this estimating equation to obtain an estimator $\widehat{\boldsymbol{\beta}}$.

We describe a method-of-moments estimator for α for the *quadratic* variance model we have

$$\mathrm{var}(Y_i \mid \boldsymbol{\beta}, \alpha) = \mathrm{E}[(Y_i - \mu_i)^2] = \mu_i(1 + \mu_i/\alpha),$$

and so

$$\alpha^{-1} = \mathrm{E}\left[\frac{(Y_i - \mu_i)^2 - \mu_i}{\mu_i^2}\right],$$

$i = 1, ..., n$, leading to the method-of-moments estimator

$$\widehat{\alpha} = \left\{\frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i - \widehat{\mu}_i)^2 - \widehat{\mu}_i}{\widehat{\mu}_i^2}\right\}^{-1}. \tag{36}$$

If we have a consistent estimator $\widehat{\alpha}$, and the mean is correctly specified then valid inference follows from

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}}\widehat{\boldsymbol{V}}(\widehat{\alpha})^{-1}\widehat{\boldsymbol{D}})^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \rightarrow_d N(\mathbf{0}, \boldsymbol{I}_p).$$

We fit this model to the air pollution data.

The estimates (standard errors) are $\widehat{\beta}_0 = 4.71$ (0.018) and $\widehat{\beta}_1 = 0.0014$ (0.00056).

The moment-based estimator is $\widehat{\alpha} = 65.20$.

This analysis therefore produces virtually identical inference with the quasi-likelihood approach in which the variance was a linear function of the mean.

In Figure 16 we plot the linear and quadratic variance functions (over the range of the mean for these data) and we see that they are very similar.

Examination of the residuals did not clearly indicate the superiority of either variance model; it is typically very difficult to distinguish between the two models, unless the mean of the data has a large spread.

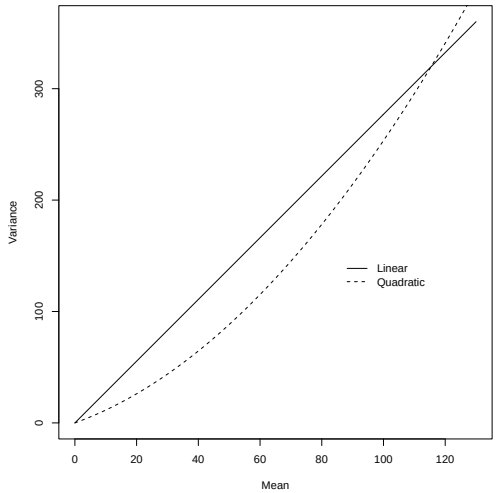

Figure 16: Linear and quadratic variance functions for the air pollution data.

Example: Rcode for Quasi-Poisson Regression

We run the Poisson regression and then evaluate the method-of-moments estimator of α "by hand".

```
> mod1 <- glm(ynew~x1new,family=poisson)
> summary(mod1)
Coefficients:
                  Value   Std. Error   t value
(Intercept) 4.705062304 0.0113962988 412.85880
      x1new 0.001458115 0.0003348748   4.35421
(Dispersion Parameter for Poisson family taken to be 1 )
    Null Deviance: 927.372 on 334 degrees of freedom
Residual Deviance: 908.6531 on 333 degrees of freedom
Number of Fisher Scoring Iterations: 3
Correlation of Coefficients:
      (Intercept)
x1new -0.8949913
> resid1 <- (ynew - mod1$fit)/sqrt(mod1$fit)
> alphahat <- sum(resid1 * resid1)/(length(ynew) - 2)
> alphahat
[1] 2.772861
```


We now fit the Quasi-Likelihood model with

$$\mathrm{E}[Y_i|\boldsymbol{\beta}] = \mu_i = \exp(\beta_0 + \beta_1 x_i)$$

and

$$\mathrm{var}(Y_i|\boldsymbol{\beta}) = \alpha\mu_i = \alpha\exp(\beta_0 + \beta_1 x_i).$$

```
> mod2 <- glm(ynew~x1new,quasi(link=log,variance=mu))
> summary(mod2)
Coefficients:
                  Value  Std. Error    t value
(Intercept) 4.705062304 0.018976351 247.943468
      x1new 0.001458115 0.000557611   2.614932
(Dispersion Parameter for Quasi-likelihood
    family taken to be 2.772667 )
    Null Deviance: 927.372 on 334 degrees of freedom
Residual Deviance: 908.6531 on 333 degrees of freedom
Number of Fisher Scoring Iterations: 3
Correlation of Coefficients:
      (Intercept)
x1new -0.8949913
```

The standard errors are multiplied by $\sqrt{\hat{\alpha}}$ (=1.67 here), but the estimates are unchanged.

Example: Rcode for Quadratic Variance Model

The `glm.nb` function carries out MLE for the negative binomial model (it is part of the `MASS` library).

We find the MLE of α, and then use this as a starting value for the iterative strategy in which a method-of-moments estimator is used.

```
> library(MASS)
> modnegbinmle <- glm.nb(y~x)
> summary(modnegbinmle)
Call:
glm.nb(formula = y ~ x, init.theta = 67.7145, link = log)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.7055974  0.0188269 249.941   <2e-16 ***
x           0.0014405  0.0005577   2.583   0.0098 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Correlation of Coefficients:
  (Intercept)
x -0.90
              Theta:  67.71
          Std. Err.:  8.27
> alphahat <- 67.61
```


Now iterate to a solution by estimating $\boldsymbol{\beta}$ for fixed α, and then re-estimating α.

```
> alphanew <- 0
> counter <- 0
> for (i in 1:5){
        fit <- glm(y~x,family=negative.binomial(alphahat))
        mu <- fit$fitted
        alphanew <- 1/(sum(((y-mu)^2-mu)/mu^2)/(length(y)-2))
        alphahat <- alphanew
        cat("Iteration ",i,alphahat,"\n")
}
Iteration  1 65.19642
Iteration  2 65.19649
Iteration  3 65.19649
Iteration  4 65.19649
Iteration  5 65.19649
> summary(fit)
Call:
glm(formula = y ~ x, family = negative.binomial(alphahat))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.705605   0.019071 246.747   <2e-16 ***
x           0.001440   0.000565   2.549   0.0112 *
(Dispersion parameter for Negative Binomial(65.1965)
family taken to be 1.001560)
```

Generalized Estimating Equations

Suppose we assume

$$\mathrm{E}[\boldsymbol{Y}_i \mid \boldsymbol{\beta}] = \boldsymbol{x}_i\boldsymbol{\beta},$$

and consider the $n_i \times n_i$ *working* variance-covariance matrix:

$$\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{W}_i.$$

To motivate GEE we begin by assuming that $\boldsymbol{W}_i$ is known. In this case the GLS estimator minimizes

$$\sum_{i=1}^{m}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}),$$

and is given by the solution to the estimating function

$$\sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}),$$

which is

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

We now examine the properties of this estimator.

We have

$$\mathrm{E}[\widehat{\boldsymbol{\beta}}] = \left(\sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\boldsymbol{x}_i\right)^{-1}\sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\mathrm{E}[\boldsymbol{Y}_i] = \boldsymbol{\beta},$$

so long as the mean is correctly specified.

If the information about $\boldsymbol{\beta}$ grows with increasing m, then $\widehat{\boldsymbol{\beta}}$ is consistent.

The variance, $\mathrm{var}(\widehat{\boldsymbol{\beta}})$, is given by

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

If the assumed variance-covariance matrix is correct, i.e. $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i$, then

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\boldsymbol{x}_i\right)^{-1},$$

and a Gauss-Markov Theorem shows that, in this case, the estimator is efficient amongst linear estimators.

If m is large then a multivariate central limit theorem shows that $\widehat{\boldsymbol{\beta}}$ is asymptotically normal.

We now suppose that $\text{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is of so that $\boldsymbol{\alpha}$ are parameters in the variance-covariance model. The regression parameters are contained in $\boldsymbol{W}_i$ to allow, mean-variance relationships, e.g.

$$\begin{aligned} \text{var}(Y_{ij} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \alpha_1 \mu_{ij}^2 \\ \text{cov}(Y_{ij}, Y_{ik} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \alpha_1 \alpha_2^{|t_{ij} - t_{ik}|} \mu_{ij} \mu_{ik} \end{aligned}$$

where

- $\mu_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta}$,
- α_1 is the variance (which is assumed constant across time and across individuals), and
- α_2 is the correlation (which is assumed to be the same for all individuals), and
- $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$.

For known $\boldsymbol{\alpha}$ we would minimize

$$\sum_{i=1}^{m} (\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\text{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}, \boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}),$$

with solution given by the root of the estimating equation

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\text{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}, \boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) = \boldsymbol{0}.$$

In general the roots of this equation are not available in closed form (because $\boldsymbol{\beta}$ occurs in $\boldsymbol{W}$).

However, if $\boldsymbol{W}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{W}_i(\boldsymbol{\alpha})$ we have

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\text{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}) \boldsymbol{x}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{x}_i^{\text{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}) \boldsymbol{Y}_i.$$

Finally, suppose that $\boldsymbol{\alpha}$ is unknown but we have a method by which a consistent estimator $\widehat{\boldsymbol{\alpha}}$ is produced (e.g. method of moments).

We then solve the estimator function

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

In general iteration is needed to simultaneously estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Let $\widehat{\boldsymbol{\alpha}}^{(0)}$ be an initial estimate, then set $t = 0$ and iterate between

1. Solve $\boldsymbol{G}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}^{(t)}) = \mathbf{0}$ to give $\widehat{\boldsymbol{\beta}}^{(t+1)}$,
2. Estimate $\widehat{\boldsymbol{\alpha}}^{(t+1)}$ with $\widehat{\mu}_i = \mu_i\left(\widehat{\boldsymbol{\beta}}^{(t+1)}\right)$. Set $t \rightarrow t+1$ and return to 1.

We have

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathrm{N}_{k+1}\left(\mathbf{0}, \boldsymbol{I}\right),$$

where

$$\begin{aligned}
\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) &= \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})\boldsymbol{x}_i\right)^{-1} \\
&\times \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})\mathrm{var}(\boldsymbol{Y}_i)\boldsymbol{W}_i^{-1}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})\boldsymbol{x}_i\right) \\
&\times \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})\boldsymbol{x}_i\right)^{-1}.
\end{aligned}$$

We have assumed that $\mathrm{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}) = 0$ for $i \neq i'$, and this is required for the asymptotic distribution to be appropriate.

The final element of GEE is sandwich estimation of $\text{var}(\widehat{\boldsymbol{\beta}})$. In particular $\text{cov}(\boldsymbol{Y}_i)$ is estimated by

$$(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^{\mathrm{T}},$$

may be multiplied by $N/(N-p)$ to account for estimation of $\boldsymbol{\beta}$ ($N = \sum_i n_i$).

Empirical would be a better word than *robust* (which is sometimes used) for the estimator of the variance – not robust to sample size, in fact could be highly unstable.

We can write the $(k+1) \times 1$ estimating function as

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta})$$

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}[\boldsymbol{x}_{i1}\cdots\boldsymbol{x}_{in_i}]\begin{bmatrix} W_i^{11} & \cdots & W_i^{1n_i} \\ \cdots & \cdots & \cdots \\ W_i^{n_i1} & \cdots & W_i^{n_in_i} \end{bmatrix}\begin{bmatrix} Y_{i1} - \boldsymbol{x}_{i1}\boldsymbol{\beta} \\ \cdots \\ Y_{in_i} - \boldsymbol{x}_{in_i}\boldsymbol{\beta} \end{bmatrix}$$

where W_i^{ij} denotes entry (i,j) of the inverse $\boldsymbol{W}_i$. We use the middle form since this emphasizes that the basic unit of replication is indexed by i.

Example: Suppose for simplicity that we have a balanced design, with $n_i = n$ for all i, and assume a working variance-covariance matrix with

$$\begin{aligned} \text{var}(Y_{ij}) &= \text{E}[(Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta})^2] = \text{E}[\epsilon_{ij}^2] = \alpha_1 \\ \text{cov}(Y_{ij}, Y_{ik}) &= \text{E}[(Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta})(Y_{ik} - \boldsymbol{x}_{ik}\boldsymbol{\beta})] = \text{E}[\epsilon_{ij}\epsilon_{ik}] = \alpha_1\alpha_{2jk}, \end{aligned}$$

for $i = 1, ..., m; j, k = 1, ..., n; j \neq k$. Hence we have $n + n(n-1)/2$ elements of $\boldsymbol{\alpha}$.

Letting

$$e_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\widehat{\boldsymbol{\beta}},$$

method-of-moments estimators are given by

$$\widehat{\alpha}_1 = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} e_{ij}^2,$$

and

$$\widehat{\alpha}_1\widehat{\alpha}_{2jk} = \frac{1}{m}\sum_{i=1}^{m} e_{ij}e_{ik}.$$

Generalized Estimating Equation (GEE) Summary

We have:

- Regression parameters (of primary interest) $\boldsymbol{\beta}$ and,
- Variance-covariance parameters $\boldsymbol{\alpha}$.

We have considered the GEE

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

where

- $\mu_i = \mu_i(\boldsymbol{\beta}) = \boldsymbol{x}_i \boldsymbol{\beta}$.
- $\boldsymbol{D}_i = \boldsymbol{D}_i(\boldsymbol{\beta}) = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \boldsymbol{x}_i^{\mathrm{T}}$,
- $\boldsymbol{W}_i = \boldsymbol{W}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the "working" covariance model,

Three important ideas:

1. Separate estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.
2. Sandwich estimation of $\mathrm{var}(\widehat{\boldsymbol{\beta}})$.
3. Replication across units in order to estimate covariances – so we have assumed that observations on different units are independent.

Notes:

- We have seen the first and second ideas in independent data situations – e.g. estimation of the α parameter in the quadratic negative binomial model.
- We may use method of moments estimators for $\boldsymbol{\alpha}$ (or set up another estimating equation, see later).
- We could go with model-based standard errors:

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right)^{-1}. \tag{37}$$

If we have an independence working model ($\boldsymbol{W}_i = \boldsymbol{I}$) then no iteration necessary (since no $\boldsymbol{\alpha}$ in the GEE) – in this case we'd want to use sandwich estimation, however.

Dental Example

Look at various estimators of $\boldsymbol{\beta}$ for girls only. Note here that we might question the asymptotics for GEE since we only have replication across $m = 11$ units (girls) (check with simulation – see coursework).

Start with ordinary least squares – unbiased estimator for $\boldsymbol{\beta}$, but standard errors are wrong because independence is assumed.

```
> summary(lm(distance~age,data=Orthgirl))

Call:
lm(formula = distance ~ age, data = Orthgirl)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3727     1.6378  10.608 1.87e-13 ***
age           0.4795     0.1459   3.287  0.00205 **
Residual standard error: 2.164 on 42 degrees of freedom
Multiple R-Squared: 0.2046,     Adjusted R-squared: 0.1856
F-statistic:  10.8 on 1 and 42 DF,  p-value: 0.002053
```


Now implement GEE with working independence – the following is an `R` implementation.

```
> library(nlme); data(Orthodont); Orthgirl <- Orthodont[Orthodont$Sex=="Female",]
> install.packages("geepack")
> library(geepack)
> summary(geese(distance~age,id=Subject,data=Orthgirl,corstr="independence"))
Call:
geese(formula = distance ~ age, id = Subject, data = Orthgirl,corstr = "independence")
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian
 Coefficients:
              estimate    san.se      wald            p
(Intercept) 17.3727273 0.7819784 493.56737 0.000000e+00
age          0.4795455 0.0666386  51.78547 6.190604e-13
Scale Model:
 Scale Link:                identity
 Estimated Scale Parameters:
            estimate   san.se     wald           p
(Intercept) 4.470403 1.373115 10.59936 0.001131270
Correlation Model:
 Correlation Structure:     independence
Returned Error Value:    0
Number of clusters:   11   Maximum cluster size: 4
```

Next we examine an exchangeable correlation structure in which all pairs of observations on the same unit have a common correlation:

```
> summary(geese(distance~age,id=Subject,data=Orthgirl,corstr="exchangeable"))
geese(formula = distance ~ age, id = Subject, data = Orthgirl,
    corstr = "exchangeable")
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian
 Coefficients:
              estimate    san.se      wald            p
(Intercept) 17.3727273 0.7819784 493.56737 0.000000e+00
age          0.4795455 0.0666386  51.78547 6.190604e-13
Scale Model:
 Scale Link:                identity
 Estimated Scale Parameters:
            estimate   san.se     wald           p
(Intercept) 4.470403 1.373115 10.59936 0.001131270
Correlation Model:
 Correlation Structure:     exchangeable
 Correlation Link:          identity
 Estimated Correlation Parameters:
       estimate    san.se     wald            p
alpha 0.8680178 0.1139327 58.04444 2.564615e-14
Number of clusters:   11   Maximum cluster size: 4
```


Notes:

- Independence estimates are always identical to OLS because we have assumed working independence, which means that the estimating equation is the same as the normal equations.
- Standard error for β_1 is smaller with GEE because regressor (time) is changing within an individual.
- Here we obtain the same estimates for exchangeable as working independence but only because balanced and complete (i.e. no missing) data.

Finally we look at AR(1) and unstructured errors – this time we see slight differences in estimates and standard errors.

```
> summary(geese(distance~age,id=Subject,data=Orthgirl,corstr="ar1"))
geese(formula = distance ~ age, id = Subject, data = Orthgirl, corstr = "ar1")
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian
 Coefficients:
             estimate     san.se      wald            p
(Intercept) 17.3049830 0.85201953 412.51833 0.000000e+00
age          0.4848065 0.06881228  49.63692 1.849965e-12
Scale Model:
 Scale Link:                identity
 Estimated Scale Parameters:
            estimate   san.se   wald            p
(Intercept) 4.470639 1.341802 11.101 0.0008628115
Correlation Model:
 Correlation Structure:     ar1
 Correlation Link:          identity
 Estimated Correlation Parameters:
      estimate     san.se     wald p
alpha 0.9298023 0.07164198 168.4403 0
Number of clusters:   11   Maximum cluster size: 4
```


Now delete last two observations from girl 11 to illustrate that identical answers before were consequence of balance and completeness of data.

```
> Orthgirl2<-Orthgirl[1:42,]
> summary(lm(distance~age,data=Orthgirl2))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.0713     1.5102  11.966 8.56e-15 ***
age           0.3963     0.1357   2.921  0.00571 **
Residual standard error: 1.964 on 40 degrees of freedom
> summary(geese(distance~age,id=Subject,data=Orthgirl2,
corstr="independence"))
Coefficients:
             estimate     san.se      wald            p
(Intercept) 18.0713312 0.82603439 478.61250 0.000000e+00
age          0.3962971 0.06934195  32.66253 1.096304e-08
Scale Model:
 Scale Link:                identity
 Estimated Scale Parameters:
            estimate   san.se     wald           p
(Intercept) 3.674926 1.317669 7.778294 0.005287771
Correlation Model:
 Correlation Structure:     independence
Returned Error Value:    0
Number of clusters:   11   Maximum cluster size: 4
```

```
> summary(geese(distance~age,id=Subject,data=Orthgirl2,corstr="exchangeable"))
Call:
geese(formula = distance ~ age, id = Subject, data = Orthgirl2,
    corstr = "exchangeable")
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian
 Coefficients:
             estimate     san.se      wald            p
(Intercept) 17.6050097 0.79007168 496.52320 0.000000e+00
age          0.4510122 0.06641218  46.11913 1.112765e-11
Scale Model:
 Scale Link:                identity
 Estimated Scale Parameters:
            estimate   san.se    wald           p
(Intercept) 3.706854 1.320019 7.88589 0.004982194
Correlation Model:
 Correlation Structure:     exchangeable
 Correlation Link:          identity
 Estimated Correlation Parameters:
       estimate     san.se     wald p
alpha 0.7968515 0.09367467 72.36198 0
Returned Error Value:    0
Number of clusters:   11   Maximum cluster size: 4
```


Comparison of Analyses

In Table 7 summaries are presented under likelihood, Bayesian and GEE analyses.

Two Bayesian models were fitted, a normal model:

$$\begin{aligned}
\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D} &\sim_{iid} \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{D}), \quad \mathrm{var}(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D}) = \boldsymbol{D} \\
\boldsymbol{D}^{-1} &\sim \mathrm{W}(r, \boldsymbol{R}^{-1}), \quad \mathrm{E}[\mathrm{var}(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D})] = \frac{\boldsymbol{R}}{r-3} \\
\boldsymbol{R} &= \begin{bmatrix} 1.0 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad r = 4
\end{aligned}$$

and a Student t_4 model:

$$\begin{aligned}
\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D} &\sim_{iid} \mathrm{St}_4(\boldsymbol{\beta}, \boldsymbol{D}), \quad \mathrm{var}(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D}) = 2\boldsymbol{D} \\
\boldsymbol{D}^{-1} &\sim \mathrm{W}(r, \boldsymbol{R}_t^{-1}), \quad \mathrm{E}[\mathrm{var}(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D})] = 2\frac{\boldsymbol{R}_t}{r-3} \\
\boldsymbol{R}_t &= \begin{bmatrix} 0.5 & 0 \\ 0 & 0.05 \end{bmatrix}, \quad r = 4
\end{aligned}$$

Approach	$\widehat{\beta}_0$	s.e.($\widehat{\beta}_0$)	$\widehat{\beta}_1$	s.e.($\widehat{\beta}_1$)
LMEM ML	22.65	0.62	0.480	0.065
LMEM REML	22.65	0.63	0.479	0.066
Bayes Normal	22.65	0.60	0.479	0.075
Bayes t_4	22.65	0.58	0.475	0.073
GEE Independence	22.65	0.55	0.480	0.067
GEE AR(1)	22.64	0.58	0.485	0.069

Table 7: Summaries for fixed effects.

- Overall, the analyses are in good correspondence.

Approach	$\widehat{\text{var}}(\beta_{0i})$	$\widehat{\text{var}}(\beta_{1i})$	$\widehat{\text{corr}}(\beta_{0i}, \beta_{1i})$	$\widehat{\sigma}_\epsilon$
LMEM ML	1.98	0.15	0.55	0.67
LMEM REML	2.08	0.16	0.53	0.67
Bayes Normal	1.93 (1.29,2.96)	0.18 (0.10,0.31)	0.39 (-0.32,0.85)	0.70 (0.52,0.93)
Bayes t_4	2.06 (1.18,3.46)	0.20 (0.11,0.35)	0.42 (-0.34,0.88)	0.71 (0.54,0.95)

Table 8: Summaries for variance components.

GEE with working independence gives $\alpha_1 = 4.47$.

GEE with working AR(1) gives $\alpha_1 = 4.47$, $\alpha_2 = 0.93$.

The parameterization adopted for the linear model changes the interpretation of $\boldsymbol{D}$. For example:

Model 1: $(\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_j$, $\boldsymbol{b}_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{D})$.

Model 2: $(\gamma_0 + b^\star_{0i}) + (\gamma_1 + b^\star_{1i})(t_j - \overline{t})$, $\boldsymbol{b}^\star_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{D}^\star)$.

Giving $\beta_0 = \gamma_0 - \gamma_1\overline{t}$, $\beta_1 = \gamma_1$.

$b_{0i} = b^\star_{0i} - \overline{t}b^\star_{1i}$, $b_{1i} = b^\star_{1i}$.

Moral: $\boldsymbol{D} \neq \boldsymbol{D}^\star$; $D_{00} = D^\star_{00} - 2\overline{t}D^\star_{01} + \overline{t}^2 D^\star_{11}$, $D_{01} = D^\star_{01} - \overline{t}D_{11}$, $D_{11} = D^\star_{11}$.

Covariance Models for Clustered Data

Whether we take a GEE or LME approach (with inference from the likelihood or from the posterior) we require flexible yet parsimonious covariance models.

In GEE we require a working covariance model

$$\text{cov}(\boldsymbol{Y}_i) = \boldsymbol{W}_i,$$

$i = 1, ..., m$.

With LME we have so far assumed the model

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{38}$$

with $\boldsymbol{b}_i \sim_{ind} \text{N}(\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{\epsilon}_i \sim_{ind} \text{N}(\boldsymbol{0}, \boldsymbol{E}_i)$, with $\boldsymbol{E}_i = \boldsymbol{I}_{n_i}\sigma^2$.

With $\boldsymbol{z}_i\boldsymbol{b}_i = \boldsymbol{1}_{n_i}b_i$ we obtained an *exchangeable* (also known as compound symmetry):

$$\text{var}(\boldsymbol{Y}_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

This model is particularly appropriate for clustered data with no time ordering (e.g. ANOVA).

An obvious extension for longitudinal data is to assume

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i,$$

with:

- Random effects $\boldsymbol{b}_i \sim_{ind} \text{N}(\boldsymbol{0}, \boldsymbol{D})$.
- Serial correlation $\boldsymbol{\delta}_i \sim_{ind} \text{N}(\boldsymbol{0}, \boldsymbol{R}_i\sigma^2_\delta)$, with $\boldsymbol{R}_i$ an $n_i \times n_i$ correlation matrix with elements

  $$R_{ijj'} = \text{corr}(Y_{ij}, Y_{ij'}|\boldsymbol{b}_i),$$

  $j, j' = 1, ..., n_i$.
- Measurement error $\boldsymbol{\epsilon}_i \sim_{ind} \text{N}(0, \boldsymbol{I}_{n_i}\sigma^2_\epsilon)$.

In general it is difficult to identify all three sources of variability – but the above provides a useful conceptual model.

See DHLZ, Chapter 5; Verbeke and Molenberghs, Chapter 10; Pinheiro and Bates, Chapter 5.

Within-Unit Covariance Models

Autoregressive errors

A widely-used time series model is the autoregressive, AR(1), process

$$\delta_{ij} = \rho\delta_{i,j-1} + u_{ij}, \tag{39}$$

for $j \geq 2$, $|\rho| \leq 1$ where $u_{ij} \sim_{iid} \mathrm{N}(0, \sigma_u^2)$ and are independent of δ_{ik}, $k > 0$. For LMEM we require a likelihood and hence the joint distribution of $\boldsymbol{\delta}_i$, for GEE the first two moments.

Repeated application of (39) gives, for $k > 0$,

$$\delta_{ij} = u_{ij} + \rho u_{i,j-1} + \rho^2 u_{i,j-2} + ... + \rho^{k-1} u_{j-k+1} + \rho^k \delta_{i,j-k}. \tag{40}$$

Assume the process has been running since $j = -\infty$ and that it is 'stable' so that $|\rho| < 1$ and the δ_{ij} all have the same distribution.

Then, from (40)

$$\mathrm{var}(\delta_{ij}) = \sigma_u^2(1 + \rho^2 + \rho^4 + ... + \rho^{2(k-1)}) + \rho^{2k}\mathrm{var}(\delta_{i,j-k}).$$

As $k \to \infty$, since $\sum_{l=1}^{\infty} x^{l-1} = 1/(1-x)$,

$$\mathrm{var}(\delta_{ij}) = \frac{\sigma_u^2}{(1-\rho^2)} = \sigma_\delta^2,$$

and, by substitution of (40),

$$\mathrm{cov}(\delta_{ij}, \delta_{i,j-k}) = \mathrm{E}[\delta_{ij}\delta_{i,j-k}] = \frac{\sigma_u^2\rho^k}{(1-\rho^2)} = \sigma_\delta^2\rho^k.$$

Hence under this model we have

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & \rho & \rho^2 & ... & \rho^{n_i-1} \\ \rho & 1 & \rho & ... & \rho^{n_i-2} \\ \rho^2 & \rho & 1 & ... & \rho^{n_i-3} \\ ... & ... & ... & ... & ... \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & ... & 1 \end{bmatrix}$$

as the correlation matrix for $\boldsymbol{\delta}_i$.

Often this model is written in the form

$$\mathrm{cov}(Y_{ij}, Y_{ik}) = \sigma_\delta^2 \exp(-\phi d_{ijk}),$$

($\rho = \mathrm{e}^{\phi}$) with $d_{ijk} = |t_{ij} - t_{ik}|$ which is valid for unequally-spaced times also.

Toeplitz: Unstructured correlation:

$$\text{var}(\boldsymbol{Y}_i) = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

Heterogeneous versions with non-constant variance can also be fitted.

For example, the heterogenenous exchangeable model is given by:

$$\text{var}(\boldsymbol{Y}_i) = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4 \\ \rho\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho\sigma_4\sigma_1 & \rho\sigma_4\sigma_2 & \rho\sigma_4\sigma_3 & \sigma_4^2 \end{bmatrix}$$

Note that we should be careful when specifying the covariance structure – identifiability problems may arise if we try to be too flexible.

Assessment of Assumptions

Each of the approaches to modeling that we have described depend upon assumptions concerning the structure of the data; to ensure that inference is appropriate we need to attempt to check that these assumptions are valid.

We first recap the assumptions:

GEE

Model:

$$\boldsymbol{Y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{e}_i,$$

with working covariance model $\text{var}(\boldsymbol{e}_i) = \boldsymbol{W}_i(\boldsymbol{\alpha})$, $i = 1, ..., m$.

G1 Marginal model $\text{E}[\boldsymbol{Y}_i] = \boldsymbol{x}_i\boldsymbol{\beta}$ is appropriate.

G2 m is sufficiently large for asymptotic inference to be appropriate.

G3 m is sufficiently large for robust estimation of standard errors.

G4 The working covariance $\boldsymbol{W}_i(\boldsymbol{\alpha})$ is not far from the "true" covariance structure; if this is the case then the analysis will be very inefficient (standard errors will be much bigger than they need to be).

LMEM via Likelihood Inference

Model:

$$\boldsymbol{Y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

with $\boldsymbol{b}_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{\epsilon}_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{E}_i)$, $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ independent ($\boldsymbol{E}_i$ may have complex structure depending on both independent and dependent terms), $i = 1, ..., m$.

L1 Mean model for fixed effects $\boldsymbol{x}_i\boldsymbol{\beta}$ is appropriate.

L2 Mean model for random effects $\boldsymbol{z}_i\boldsymbol{b}_i$ is appropriate.

L3 Variance model for $\boldsymbol{\epsilon}_i$ is correct.

L4 Variance model for $\boldsymbol{b}_i$ is correct.

L5 Normality of $\boldsymbol{\epsilon}_i$.

L6 Normality of $\boldsymbol{b}_i$.

L7 m is sufficiently large for asymptotic inference to be appropriate.

LMEM via Bayesian Inference

Model as for LMEM, plus priors for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Each of L1–L6 (asymptotic inference is not required if, for example, MCMC is used, though "appropriate" priors are needed).

Overall strategy

Before any formal modeling is carried out the data should be examined, in table and plot form, to see if the data have been correctly read in and to see if there are outliers.

For those individuals with sufficient data, individual-specific models should also be fitted, to allow examination of the appropriateness of initially hypothesized models in terms of the:

- linear component (which covariates, including transformations and interactions),
- and assumptions about the errors, such as constant variance and serial correlation.

Following fitting of marginal, mixed models, the assumptions should then be re-assessed, primarily through residual analysis.

Residual Analysis

Residuals may be defined with respect to different levels of the model.

A vector of unstandardized *population-level* (marginal) residuals is given by

$$\boldsymbol{e}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}.$$

A vector of unstandardized *unit-level* (Stage One) residuals is given by

$$\boldsymbol{\epsilon}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{b}_i.$$

The vector of random effects, $\boldsymbol{b}_i$, is also a form of (Stage Two) residual.

Estimated versions of these residuals are given by

$$\begin{aligned}\widehat{\boldsymbol{e}}_i &= \boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\epsilon}}_i &= \boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_i\widehat{\boldsymbol{b}}_i\end{aligned}$$

and $\widehat{\boldsymbol{b}}_i$, $i = 1, ..., m$.

Recall from consideration of the ordinary linear model that estimated residuals have dependencies induced by the estimation procedure; in the dependent data context the situation is much worse as the "true" residuals have dependencies due to the dependent error terms of the models used.

Hence standardization is essential to remove the dependence.

Standardized Population Residuals

If $\boldsymbol{V}_i(\boldsymbol{\alpha})$ is the true error structure then

$$\text{var}(\boldsymbol{e}_i) = \boldsymbol{V}_i, \quad \text{and} \quad \text{var}(\widehat{\boldsymbol{e}}_i) \approx \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}}),$$

so that the residuals are dependent under the model, which means that it is not possible to check whether the covariance model is correctly specified (both form of the correlation structure and mean-variance model).

Plotting $\widehat{e}_{ij}$ versus x_{ij} may also be misleading due to the dependence within the residuals.

As an alternative, let $\widehat{\boldsymbol{V}}_i = \boldsymbol{L}_i\boldsymbol{L}_i^{\text{T}}$ be the Cholesky decomposition of $\widehat{\boldsymbol{V}}_i = \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})$, the estimated variance-covariance matrix.

We can use this decomposition to form

$$\widehat{\boldsymbol{e}}_i^\star = \boldsymbol{L}_i^{-1}\widehat{\boldsymbol{e}}_i = \boldsymbol{L}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}).$$

so that $\text{var}(\boldsymbol{e}_i^\star) \approx \boldsymbol{I}_{n_i}$. We have the model

$$Y_i^\star = \boldsymbol{x}_i^\star\boldsymbol{\beta} + \boldsymbol{e}_i^\star$$

where $\boldsymbol{Y}_i^\star = \boldsymbol{L}_i^{-1}\boldsymbol{Y}_i$, $\boldsymbol{x}_i^\star = \boldsymbol{L}_i^{-1}\boldsymbol{x}_i$, $\boldsymbol{e}_i^\star = \boldsymbol{L}_i^{-1}\boldsymbol{e}_i$.

Hence plots of $\widehat{e}_{ij}^\star$ against columns of $\boldsymbol{x}_{ij}^\star$ should not show systematic patterns, *if* the assumed form is correct.

QQ plots of $\widehat{e}_{ij}^\star$ versus the expected residuals from a normal distribution can be used to assess normality (normal residuals are not required for GEE, but will help asymptotics).

Unstandardized versions will still be normally distributed if the $\boldsymbol{e}_i$ are (since the $e_{ij}^\star$ are linear combinations of $\boldsymbol{e}_i$), though the variances may be non-constant, and there may be strong dependence between different points.

The correctness of the mean-variance relationship can be assessed via examination of $e_{ij}^{\star 2}$ versus $\widehat{\mu}_{ij}^\star = \boldsymbol{x}_{ij}^\star\widehat{\boldsymbol{\beta}}$.

Local smoothers can be added to plots to aid interpretation. Plotting symbols also useful – unit number, or observation number.

Stage One Residuals

If $\boldsymbol{\epsilon}_i \sim \text{N}(\boldsymbol{0}, \sigma_i^2\boldsymbol{I}_{n_i})$ then residuals

$$\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_i\widehat{\boldsymbol{b}}_i$$

may be formed. Standardized versions are given by $\widehat{\boldsymbol{\epsilon}}_i/\widehat{\sigma}_i$.

The standardized versions should be used if the σ_i are unequal across i. Some uses:

- Plot residuals against covariates. Departures may suggest adding in covariates, both to $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$.
- To provide QQ plots – mean-variance relationship is more important to detect than lack of normality (so long as sample size is not small).
- assess constant variance assumption – one useful plot is versus $\widehat{\mu}_{ij} = \boldsymbol{x}_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}\widehat{\boldsymbol{b}}_i$.
- assess if serial correlation present in residuals

may be plotted against covariates to assess the form of the model, with QQ plots assessing normality of the measurement errors.

If $\boldsymbol{\epsilon}_i \sim \text{N}(\boldsymbol{0}, \sigma_\epsilon^2\boldsymbol{R}_i)$ with $\boldsymbol{R}_i$ a correlation matrix then the residuals should be standardized, as with population residuals.

Stage Two Residuals

Predictions of the random effects $\widehat{\boldsymbol{b}}_i$ may be used to assess assumptions associated with the random effects distribution, in particular:

- Are the random effects normally distributed?
- If we have assumed independence between random effects, does this appear reasonable?
- Is the variance of the random effects independent of covariates $\boldsymbol{x}_i$?

It should be born in mind that interpretation of random effects predictions is more difficult since they are functions of the data.

Recall that $\widehat{\boldsymbol{b}}_i$ are shrinkage estimators, and hence assumptions about $\boldsymbol{b}_i$ may not be reflected in $\widehat{\boldsymbol{b}}_i$.

We may fit curves for particular individuals with n_i large, and then check the assumptions from these.

For the LMEM it is better to examine first and second stage residuals – population residuals are a mixture so if something wrong not clear at which stage there is trouble.

Model refinement

A big problem is over-fitting in which models become too dataset-specific as they are refined on the basis of the examination of diagnostics.

In practice, if refinement is carried out through the fitting of alternative models (e.g. transformation of covariates, choice of distribution for the responses), then interval estimates will often be too narrow since they are produced by conditioning on the final model, and hence do not reflect the mechanism by which the model was selected.

From a frequentist standpoint estimators and test statistics should be examined via their long-run behaviour *given* the model-fitting process, including refinement. To be more explicit, let P denote the procedure by which a final model M is decided upon. Then suppose it is of interest to examine the bias of a statistic T,

$$\mathrm{E}[T|P] = \mathrm{E}_{M|P}\{\mathrm{E}[T|M]\}. \tag{41}$$

In general it will be incorrect to report $\mathrm{E}[T \mid \hat{M}]$ where $\hat{M}$ is the final model chosen, since this does not reflect the procedure by which $\hat{M}$ was chosen, but rather acts as if the final model is the "truth".

From a Bayesian standpoint the same problem exists because the posterior distribution should reflect all sources of uncertainty and *a priori* all possible models that may be entertained should be explicitly stated, with prior distributions being placed upon different likelihoods and the parameters of these likelihoods; model averaging should then be carried out across the different possibilities.

One solution to this difficulty is to never refine the model for a given data set. This approach is operationally pure but pragmatically dubious (unless one is in the context of a randomized experiment) since we may obtain appropriate inference for a model that is a very poor description of the phenomenon under study.

The philosophy suggested here is to think as carefully as possible about the initial model class before the analysis proceeds, but after fitting to carry out model checking and refine the model in the face of *clear* model misspecification, with refinement ideally being carried out within distinct *a priori* known classes.

So that, for example, examining quantile-quantile plots for different t distributions and picking the one that produces the straightest line would not be a good idea. Inference then proceeds as if the final model were the one that were chosen initially. This is clearly a subjective procedure but can be informally justified via either philosophical approaches.

Under a frequentist approach inference follows from the behaviour of an estimator under repeated sampling from the true model, and if an initial model is clearly wrong on the basis of a residual plot (say), then it is very unlikely to be close to the "true" model and hence it is more appropriate to obtain properties of estimators under the assumed model. With reference to (41), if a model is chosen because it is clearly superior to the alternatives, then it may be reasonable to assume that $\mathrm{E}[T \mid P] \approx \mathrm{E}[T \mid \hat{M}]$, because $\hat{M}$ would be consistently chosen in repeated sampling under these circumstances.

In a similar vein, under a Bayesian approach the above procedure is consistent with model-averaging but with the posterior model weight being concentated upon the chosen model (since alternative models are only rejected on the basis of clear inadequacy). The aim is to provide probability statements, from either philosophical standpoints that are "honest" representations of uncertainty. The above approach is relevant to analyses that are more confirmatory in their outlook, as opposed to being used for prediction, or for more exploratory purposes (for example, to gain clues to models that may be appropriate for future data analyses).

If aim of analysis is simply exploratory, then we can do what we like (as soon as we quote CIs or significance levels ... trouble!).

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 17) and bivariate (Figure 18).
- 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 19).
- Figure 20 shows that there are clear differences in intercepts, and some variability in slopes.

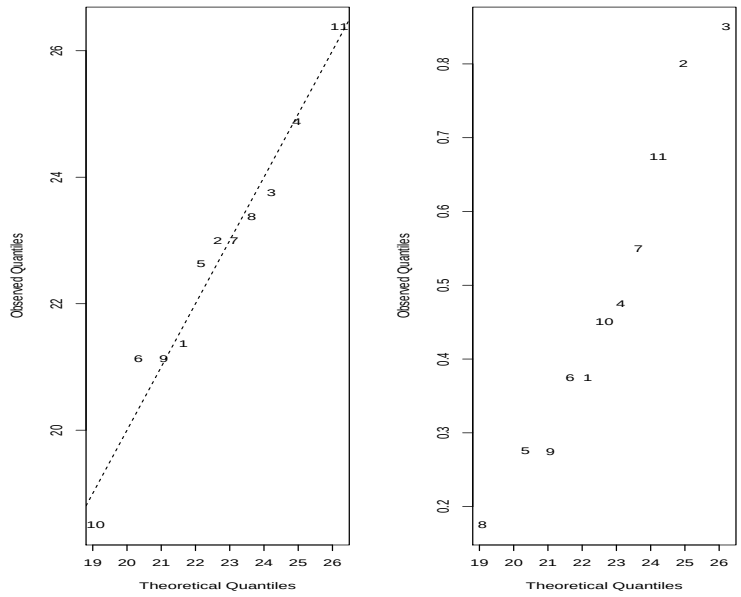

Figure 17: QQ plots of LS estimates: $\widehat{\beta}_0$ (left), $\widehat{\beta}_1$ (right).

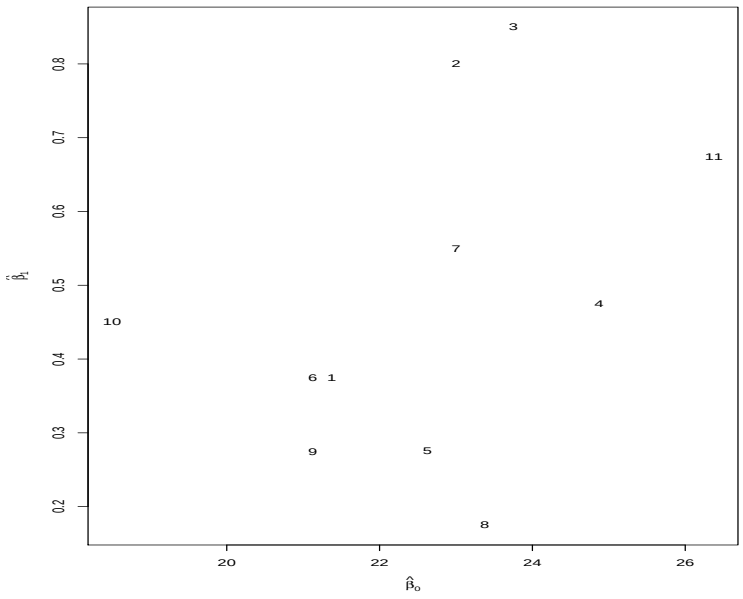

Figure 18: Bivariate plot of LS estimates.

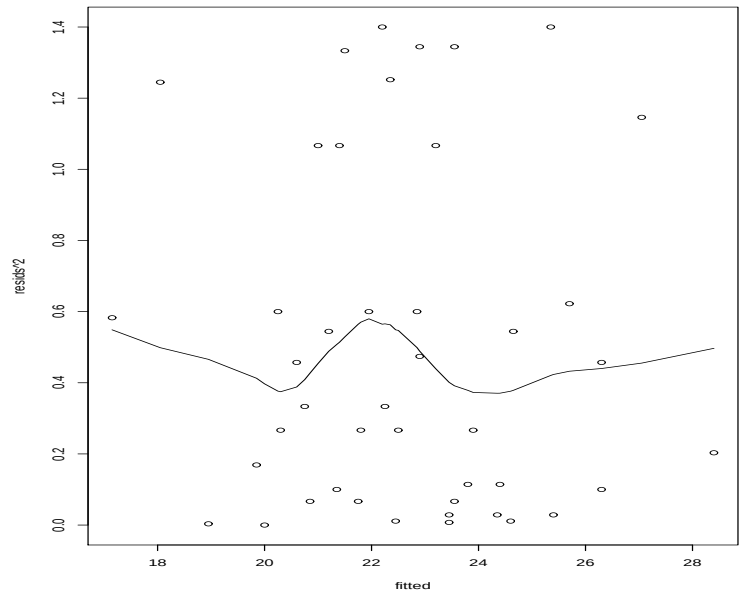

Figure 19: LS residuals versus fitted values

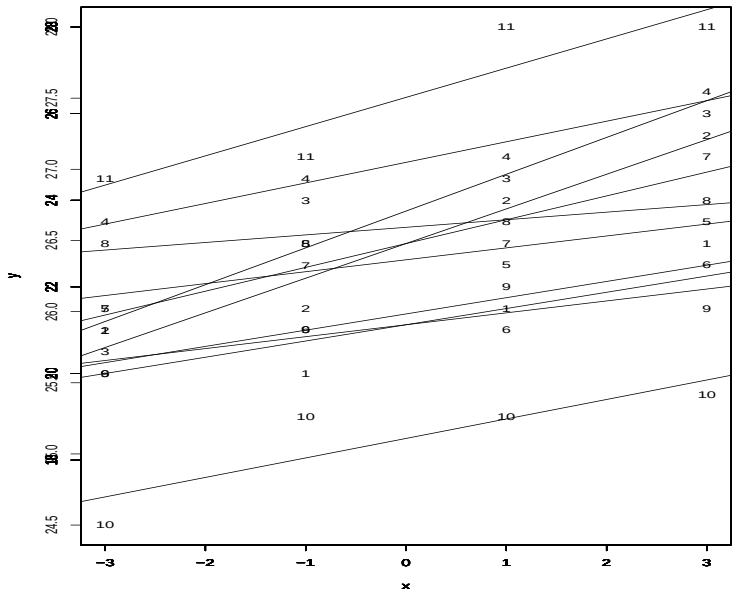

Figure 20: Fitted curves for all data.

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 21 shows no syystematic deviations between residuals with time. Figure 22 that normality reasonable, and Figure 23 that there is no mean variance relationship.

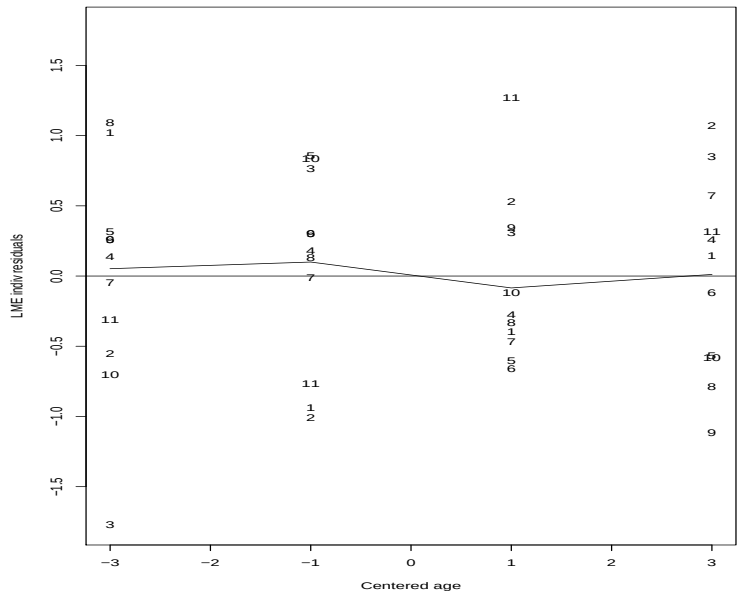

Figure 21: LME normalized residuals versus time.

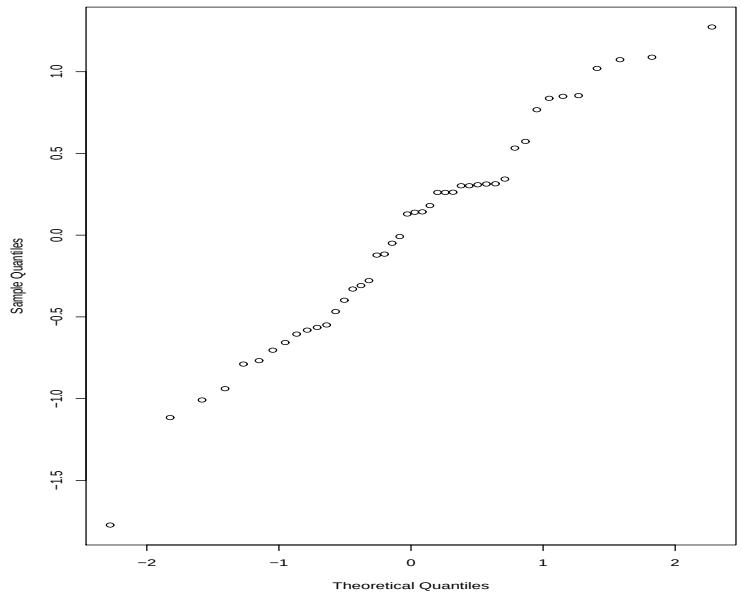

Figure 22: QQ plot of LME normalized residuals.

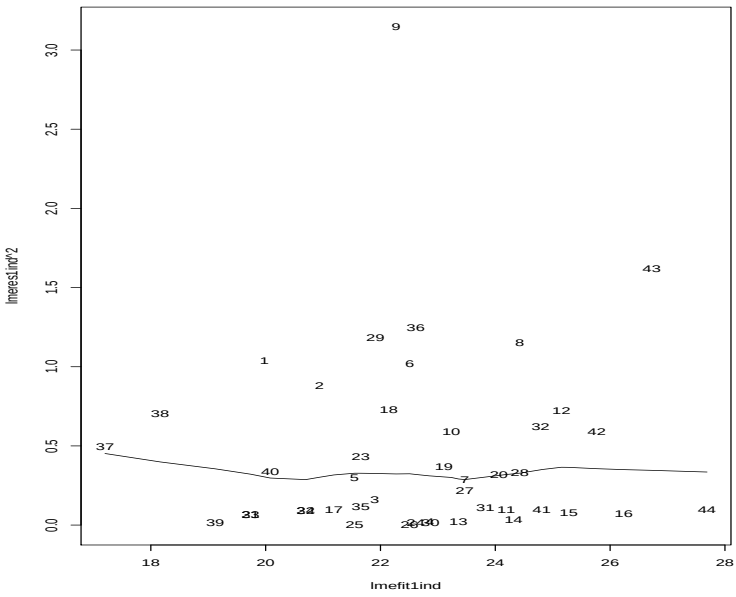

Figure 23: LME normalized residuals versus fitted values.

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 $\mathrm{E}[e(t)]$ 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

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

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,

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

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 $d \geq 0$.

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

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^2_1, 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

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

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

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

Under this model

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

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

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

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

Figure 24 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).

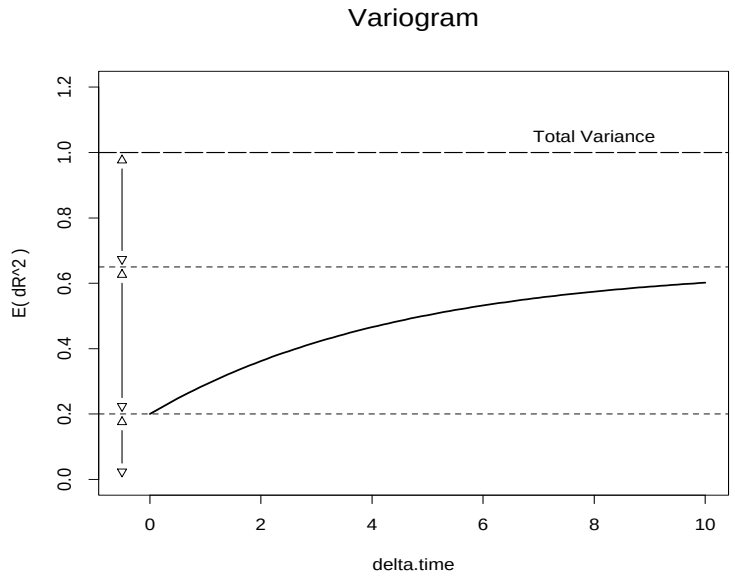

Figure 24: Theoretical variogram for a model with a random intercept, serial correlation, and measurement error.

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 25 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 26. In the left-hand panel we have only plotted 1000 of the 53301 ($327 \times 326/2$) points.

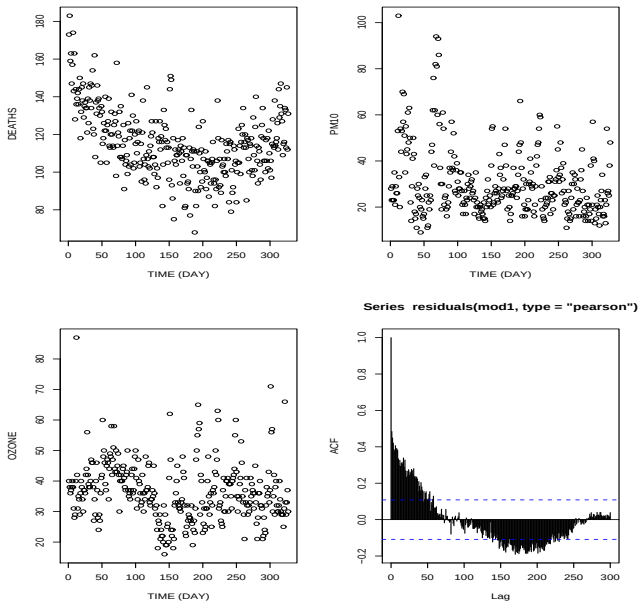

Figure 25: Time series plots and correlogram of residuals for air pollution data.

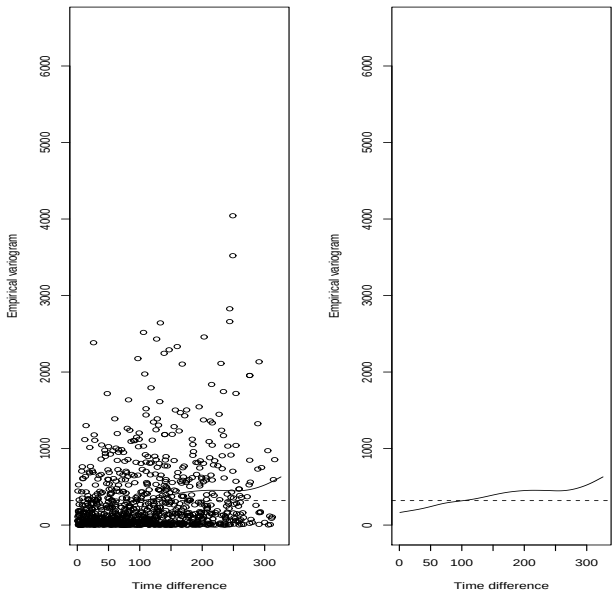

Figure 26: Variogram of residuals for air pollution data.