Stat/Biostat 571 Statistical Methodology: Regression Models for Dependent Data

Jon Wakefield

Departments of Statistics and Biostatistics, UW

Lectures: Monday/Wednesday/Friday 1.30–2.20, T473.

Coursework: (and approximate percentage contribution to final grade) weekly (30%). Examination at mid-term (30%) and final (40%). The mid-term and final will be takehome.

Office Hours:

Jon: Monday 2.30–3.20 and Wednesday 2.30–3.30 (Biostatistics, Health Sciences, 616-6292). Or by appointment (`jonno@u.washington.edu`, Padelford: 616–9388, HS: 616–6292).

TA: Youyi Fong (`yfong@u`); office hours to be arranged.

STAT/BIOSTAT 578 Data Analysis, strongly recommended for Applied Exam (Biostat students). 571 describes methods and not data analysis.

Computing will be carried out using `R` and `WinBUGS`.

Class website: `http://courses.washington.edu/b571/`

Textbooks:

Main Texts

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data, Second Edition*. Oxford University Press.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis*, Wiley.

Background Texts

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*, CRC Press.

Hand, D. and Crowder, M.J. (1996). *Practical Longitudinal Data Analysis*, CRC Press.

Pinheiro, J. and Bates, D.G. (2000). *Mixed-Effects Models in S and S-PLUS*, Springer-Verlag,

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag.

Davison, A.C. (2003). *Statistical Models*. Cambridge University Press.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*, Wiley.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models, Second Edition*, CRC Press.

COURSE OUTLINE

Revision

Motivating Datasets; Benefits and Challenges of Dependent Data; Marginal versus Conditional Modeling. Sandwich Estimation; Ordinary and Weighted Least Squares. Likelihood and Bayesian approaches.

Linear Models

Linear Mixed Effects Models; Frequentist and Bayesian Inference; Equivalence of Marginal and Conditional Modeling.

General Regression Models

Generalized Linear Mixed Models; Frequentist and Bayesian Inference; Non-equivalence of Marginal and Conditional Modeling.

Binary Data Models

Modeling the covariance structure. Mixed Effects approach.

Depending on Time:

Model Selection/Formulation

Spatial Models

Learning Objectives: 571, Winter 2008

By the end of the course the student should be able to:

- Explain why conventional regression models (as reviewed in 570) are inappropriate for dependent data, and explain the likely detrimental effects of their use in the dependent context.
- Describe the key differences between the generalized estimating equations, likelihood and Bayesian approaches to modeling dependent data. In particular the student should be able to describe the differences between the conditional and marginal approaches to modeling, and the advantages and drawbacks of each.
- Suggest appropriate approach(es) to modeling dependent outcomes, critically evaluate the fit of the fitted models, and interpret the estimated regression coefficients.
- Outline implementation strategies for each of the generalized estimating equations, likelihood and Bayesian approaches.

OVERVIEW

Recall: in a *regression analysis* we model a response, Y, as a function of covariates, $\boldsymbol{x}$.

In 570 we considered situations in which responses are *conditionally independent*, that is

$$\begin{aligned} p(Y_1, ..., Y_n|\boldsymbol{\beta}, \boldsymbol{x}) &= p(Y_1|\boldsymbol{\beta}, \boldsymbol{x}_1) \times p(Y_2|Y_1, \boldsymbol{\beta}, \boldsymbol{x}_2) \times ... \times p(Y_n|Y_1, ..., Y_{n-1}, \boldsymbol{\beta}, \boldsymbol{x}_n) \\ &= p(Y_1|\boldsymbol{\beta}, \boldsymbol{x}_1) \times p(Y_2|\boldsymbol{\beta}, \boldsymbol{x}_2) \times ... \times p(Y_n|\boldsymbol{\beta}, \boldsymbol{x}_n) \end{aligned}$$

so that observations are independent *given* parameters $\boldsymbol{\beta}$ and covariates $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

In general, $Y_1, ..., Y_n$ are *never* independent. For example, suppose

$$\mathrm{E}[Y_i|\mu, \sigma^2] = \mu, \quad \mathrm{var}(Y_i|\mu, \sigma^2) = \sigma^2,$$

$i = 1, 2$ and $\mathrm{cov}(Y_1, Y_2|\mu, \sigma^2) = 0$. Then if we are told y_1, this will change the way we think about y_2 so that $p(Y_2|Y_1) \neq p(Y_2)$, and the observations are not independent, however $p(Y_2|Y_1, \mu, \sigma^2) = p(Y_2|\mu, \sigma^2)$, so that we have conditional independence.

Motivating Examples

We distinguish between dependence induced by missing covariates, and that due to contagion (for example, in an infectious disease context) – we will not consider the latter.

One theme of the course will be modeling *residual* dependence, i.e. after we have controlled for covariates.

The obvious situations in which we would expect dependence is in data collected over time or space (but lots of others possible, e.g. families).

Example 1: Growth data

Table 1 records dental measurements of the distance in millimeters from the center of the pituitary gland to the pteryo-maxillary fissure in 11 girls and 16 boys at the ages of 8, 10, 12 and 14 years.

Here we have an example of *repeated measures* or *longitudinal* data.

Figure 1 plots these data and we see that dental growth for each child increases in an approximately linear fashion.

One common aim of such studies is to identify the *within-individual* and *between-individual* sources of variability.

Girls	8	10	12	14
1	21	20	21.5	23
2	21	21.5	24	25.5
3	20.5	24	24.5	26
4	23.5	24.5	25	26.5
5	21.5	23	22.5	23.5
6	20	21	21	22.5
7	21.5	22.5	23	25
8	23	23	23.5	24
9	20	21	22	21.5
10	16.5	19	19	19.5
11	24.5	25	28	28
Boys	8	10	12	14
1	26	25	29	31
2	21.5	22.5	23	26.5
3	23	22.5	24	27.5
4	25.5	27.5	26.5	27
5	20	23.5	22.5	26
6	24.5	25.5	27	28.5
7	22	22	24.5	26.5
8	24	21.5	24.5	25.5
9	23	20.5	31	26
10	27.5	28	31	31.5
11	23	23	23.5	25
12	21.5	23.5	24	28
13	17	24.5	26	29.5
14	22.5	25.5	25.5	26
15	23	24.5	26	30
16	22	21.5	23.5	25

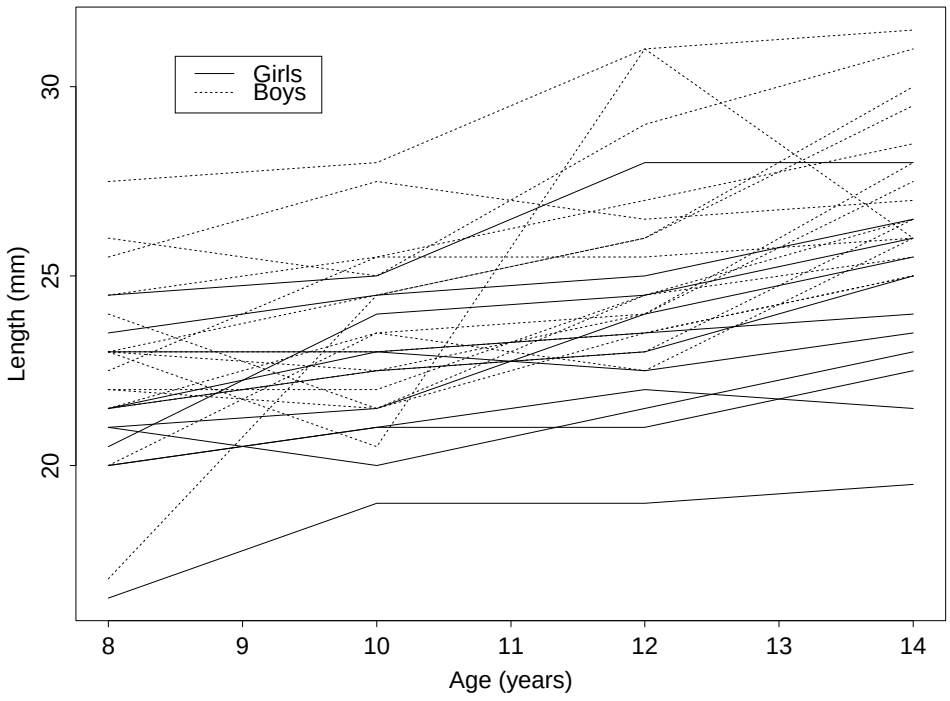

Figure 1: Dental growth data for girls and boys.

Inference

We may be interested in characterizing:

1. the *average* growth curve, or
2. the growth for a *particular* child.

Two types of analysis that will be distinguished are *marginal* and *conditional*. The former is designed for questions of type 1, and the latter may be used for both types, but requires more assumptions.

Even if the question of interest is of type 1, we still have to acknowledge the dependence of responses on the same individual – we do not have 11×4 independent observations on girls and 16×4 independent observations on boys but rather 11 and 16 *sets* of observations on girls and boys.

For either question of interest ignoring the dependence leads to incorrect standard errors and confidence/credible interval coverage.

A **marginal** approach (via Generalized Estimating Equations, GEE) to modeling specifies the moments of the data only, while in a **conditional** (via a Mixed Effects Model, MEM) approach the responses of specific individuals are modeled.

Models

First question is: why not just analyze the data from each child separately? Possible but we wouldn't be able to make formal statements about:

- The average growth rate of teeth for a girl in the age range 8–14 years.
- The between-girl variability in growth rates.

The totality of data on girls may also aid in the estimation of the growth rate for a particular girl – becomes more critical as the number of observations per child decreases. For example, in an extreme case, suppose a particular girl has only one measurement.

At the other extreme we could fit a single curve to the data from all of the girl's data together. The problem with this is that we do not have independent observations, and what if we are interested in inference for a particular child, or for a future chid?

Example 2: Spatial Data

Dependent data may result from studies with a significant spatial component.

Split Plot Data

Example: Three varieties of oats, four nitrogen concentrations.

Agricultural land was grouped into six blocks, each with three plots, and with each plot further sub-divided into four sub-plots. Within each subplot a combination of oats and nitrogen was planted. Hence we have $6 \times 3 \times 4 = 72$ observations.

We would expect observations within the same block to be correlated.

Revision Material: Estimating Functions

Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$, represent n observations from a distribution indexed by a p-dimensional parameter $\boldsymbol{\theta}$, with $\text{cov}(Y_i, Y_j \mid \boldsymbol{\theta}) = 0$, $i \neq j$.

In the following, for ease of presentation, we assume that Y_i, $i = 1, ..., n$ are independent and identically distributed (i.i.d.).

An *estimating function* is a function

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) \tag{1}$$

of the same dimension as $\boldsymbol{\theta}$ for which

$$\text{E}[\boldsymbol{G}_n(\boldsymbol{\theta})] = \mathbf{0} \tag{2}$$

for all $\boldsymbol{\theta}$. The estimating function $\boldsymbol{G}_n(\boldsymbol{\theta})$ is a random variable because it is a function of $\boldsymbol{Y}$.

The corresponding *estimating equation* that defines the estimator $\widehat{\boldsymbol{\theta}}_n$ has the form

$$\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_n) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{G}(\widehat{\boldsymbol{\theta}}_n, Y_i) = \mathbf{0}. \tag{3}$$

Result: Suppose that $\widehat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \mathbf{0},$$

i.e. $\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Then $\widehat{\boldsymbol{\theta}}_n \rightarrow_p \boldsymbol{\theta}$ (consistency) and

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \rightarrow_d \mathrm{N}_p(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{\mathrm{T}-1}) \tag{4}$$

(asymptotic normality) where

$$\boldsymbol{A} \quad = \quad \boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{G}(\boldsymbol{\theta}, Y)\right]$$

and

$$\boldsymbol{B} \quad = \quad \boldsymbol{B}(\boldsymbol{\theta}) = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^{\mathrm{T}}] = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}, Y)\}.$$

The form of the variance in (4) has lead to it being named a **sandwich estimator**.

Example: Least Squares Estimation

For the ordinary least squares/maximum likelihood estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y} \tag{5}$$

with

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2$$

if $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2\boldsymbol{I}$.

Suppose that $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2\boldsymbol{V}$ so that the model from which the estimator (5) was derived was incorrect.

Then this estimator is still unbiased but the appropriate variance estimator is

$$\begin{aligned} \mathrm{var}(\widehat{\boldsymbol{\beta}}) \quad &= \quad (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x})\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1} \\ &= \quad (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2 \end{aligned} \tag{6}$$

Expression (6) can also be derived directly from the estimating function

$$\boldsymbol{G}_n(\boldsymbol{\beta}) = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}),$$

from which we know that

$$(\boldsymbol{A}_n^{-1}\boldsymbol{B}_n\boldsymbol{A}_n^{\mathrm{T}-1})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \rightarrow_d \ \mathrm{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}),$$

(note not iid observations here) where

$$\boldsymbol{B}_n = \mathrm{var}(\boldsymbol{G}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}\sigma^2$$

and

$$\boldsymbol{A}_n = \mathrm{E}\left[\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\beta}}\right] = -\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x},$$

to give

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2.$$

We still need to know $\boldsymbol{V}$ though.

Sandwich estimator with uncorrelated errors

We relax the constant variance assumptions. Consider the estimating function

$$\boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}).$$

The "bread" of the sandwich, $\boldsymbol{A}^{-1}$, remains unchanged since $\boldsymbol{A}$ does not depend on Y.

The "filling" becomes

$$\boldsymbol{B} = \mathrm{var}\{\boldsymbol{G}\} = \boldsymbol{x}^{\mathrm{T}}\mathrm{var}(\boldsymbol{Y})\boldsymbol{x} = \sum_{i=1}^{n}\sigma_i^2\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i, \tag{7}$$

where $\sigma_i^2 = \mathrm{var}(Y_i)$ and we have assumed that the data are uncorrelated.

Unfortunately σ_i^2 is unknown – we now discuss various estimation methods.

An obvious estimator is given by

$$\widehat{\boldsymbol{B}}_n = \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i (Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}})^2, \tag{8}$$

and its use provides a consistent estimator of (7), if the data are uncorrelated.

For linear regression the estimator

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}})^2 = \frac{1}{n} \sum_{i=1}^{n} \widehat{\sigma}_i^2,$$

is downwardly biased, with bias $-p\sigma^2/n$.

The sandwich estimator is therefore also downwardly biased.

Using

$$\widetilde{\sigma}_i^2 = \frac{n}{n-p} (Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}})^2 \tag{9}$$

provides a simple correction, but in general the estimator of the variance has finite bias since the bias in $\widehat{\sigma}^2$ changes as a function of the design points $\boldsymbol{x}_i$ – various corrections have been suggestions (see Kauermann and Carroll, 2001, *JASA*).

Likelihood Methods

A special case of the estimating function methodology occurs when the estimating equation

$$\boldsymbol{G} = \frac{\partial l}{\partial \boldsymbol{\theta}}$$

is a score equation (derivative of the log-likelihood). Then $\widehat{\boldsymbol{\theta}}$ is the MLE and

$$\sqrt{n}\ (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \rightarrow_d \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{I}^{-1}) \tag{10}$$

(asymptotic normality) where $\boldsymbol{I}$ is the expected information matrix:

$$\boldsymbol{I} = \boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{G}(\boldsymbol{\theta}, Y)\right] = \boldsymbol{B}(\boldsymbol{\theta}) = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^{\mathrm{T}}] = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}, Y)\}.$$

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*.

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

$$p(\theta_i \mid \boldsymbol{y}) = \int_{\boldsymbol{\theta}_{-i}} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}_{-i}, \tag{11}$$

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

$$\mathrm{E}[\theta_i \mid \boldsymbol{y}] = \int_{\theta_i} \theta_i p(\theta_i \mid \boldsymbol{y}) \, \mathrm{d}\theta_i. \tag{12}$$

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

$$\int_{-\infty}^{\theta_i(q)} p(\theta_i \mid \boldsymbol{y}) \, \mathrm{d}\theta_i. \tag{13}$$

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

$$p(\boldsymbol{z} \mid \boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}. \tag{14}$$

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, (11)–(14), given that for most models, these are analytically intractable

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 $\mathrm{IGa}(n/2, c)$.

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

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

Note the similarity with frequentist inference.

LINEAR MODELS

We now begin thinking about specific situations, starting with linear models. Clearly, in general, ignoring dependence will give inappropriate standard errors.

While making inference for dependent data is more difficult than for independent data, designs that collect dependent data can be very efficient. For example (as we see shortly), 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).

Design Implications of a Longitudinal Study

To examine the implications of carrying out a longitudinal study, as compared to a cross-sectional study, we consider a very simple situation in which we wish to compare two treatments, coded as -1 and +1, and we have a linear model.

Cross-Sectional Study:

A single measurement is taken on each of $m = 4$ individuals where

$$Y_{i1} = \beta_0 + \beta_1 x_{i1} + \epsilon_{i1},$$

$i = 1, ..., m = 4$, ϵ_{i1} iid with $\text{var}(\epsilon_{i1}) = \sigma^2$ and

$$x_{11} = -1, x_{21} = -1, x_{31} = 1, x_{41} = 1$$

Note: $\text{E}[Y_1|x = 1] - \text{E}[Y_1|x = -1] = 2\beta_1$.

Using least squares:

$$\widehat{\beta}_0^c = \frac{\sum_{i=1}^4 Y_{i1}}{4}, \quad \widehat{\beta}_1^c = \frac{Y_{31} + Y_{41} - (Y_{11} + Y_{21})}{4},$$

and

$$\text{var}(\widehat{\beta}_0^c) = \text{var}(\widehat{\beta}_1^c) = \frac{\sigma^2}{4}.$$

Longitudinal Study:

We assume the model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \delta_{ij},$$

with b_i and δ_{ij} independent and with $\text{var}(b_i) = \sigma_0^2$, $\text{var}(\delta_{ij}) = \sigma_\delta^2$. We therefore have marginally:

$$\text{var}(Y_{ij}|\beta_0, \beta_1) = \sigma_0^2 + \sigma_\delta^2 = \sigma^2,$$

and

$$\text{cov}(Y_{i1}, Y_{i2}) = \sigma_0^2.$$

We let $\rho = \sigma_0^2/\sigma^2$, represent the correlation on observations on the same individual.

We consider two situations, both with two observations on two individuals:

Constant treatment for each individual:

$$x_{11} = x_{12} = -1, \quad x_{21} = x_{22} = 1.$$

Changing treatment for each individual:

$$x_{11} = x_{22} = 1, \quad x_{12} = x_{21} = -1.$$

Using Generalized Least Squares we have

$$\widehat{\boldsymbol{\beta}}^l = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{Y},$$

and

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}^l) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{x})^{-1}\sigma^2,$$

where

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{bmatrix}.$$

In lectures we will show that

$$\mathrm{var}(\widehat{\beta}_1^l) = \frac{\sigma^2(1-\rho^2)}{4 - 2\rho(x_{11}x_{12} + x_{21}x_{22})}.$$

The *efficiency* e is given by

$$e = \frac{\mathrm{var}(\widehat{\beta}_1^l)}{\mathrm{var}(\widehat{\beta}_1^c)} = \frac{(1-\rho^2)}{1 - \rho(x_{11}x_{12} + x_{21}x_{22})/2}.$$

Usually we have $\rho > 0$.

For the constant treatment longitudinal study

$$e = 1 + \rho,$$

so that the cross-sectional study is preferable since we have lost information due to the correlation.

For the changing treatment longitudinal study

$$e = 1 - \rho,$$

so that the longitudinal study is more efficient, because each individual is acting as their own control, that is, we are making within-individual comparisons.

If $\rho = 0$ the designs have the same efficiency.

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

$$\begin{bmatrix} \sigma_1 & & & \\ \rho_{12} & \sigma_2 & & \\ \rho_{13} & \rho_{23} & \sigma_3 & \\ \rho_{14} & \rho_{24} & \rho_{34} & \sigma_4 \end{bmatrix} \tag{15}$$

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 (15) by

$$\begin{bmatrix} 2.12 & & & \\ 0.83 & 1.90 & & \\ 0.86 & 0.90 & 2.36 & \\ 0.84 & 0.88 & 0.95 & 2.44 \end{bmatrix} \tag{16}$$

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.

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

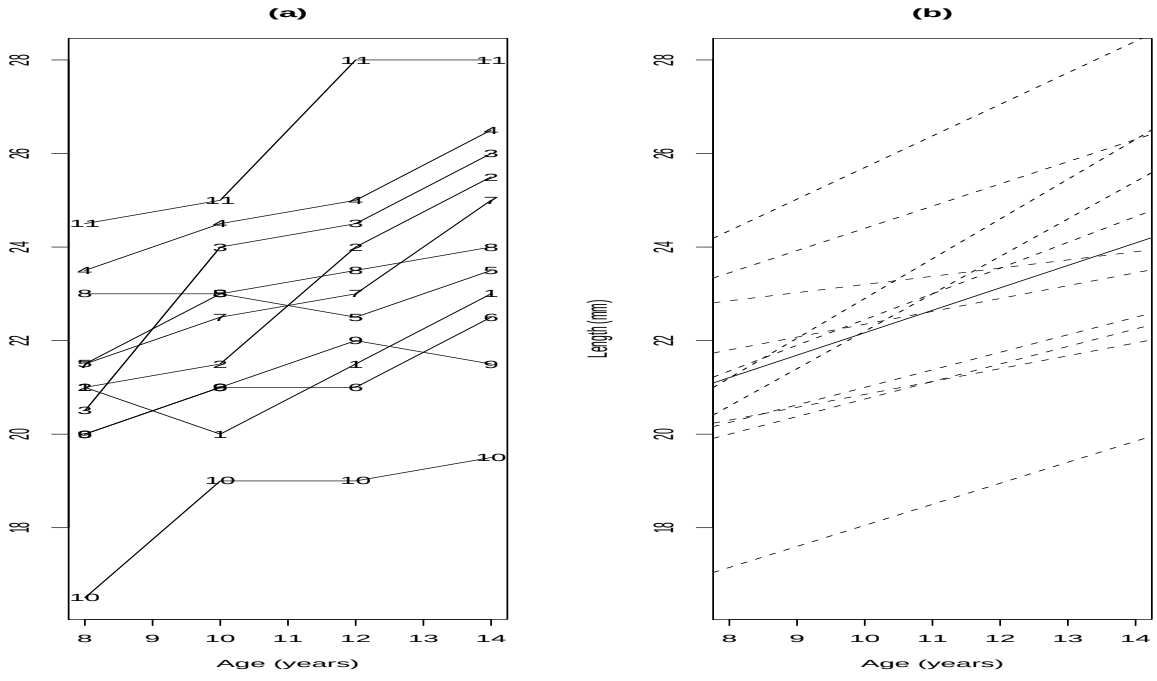

Figure 2: Dental plots for girls only: (a) Individual observed data (with plotting symbol girl index), (b) Individual fitted curves (dashed) and overall fitted curve (solid).

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:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{17}$$

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

Stage 2: Model for random terms:

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

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

The two stages define the marginal model:

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

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

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

where

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

Here $\boldsymbol{\alpha} = (\sigma^2_\epsilon, \boldsymbol{D})$ denote the variance-covariance parameters. Here $\boldsymbol{V} = \boldsymbol{z}\boldsymbol{D}\boldsymbol{z}^{\mathrm{T}} + \sigma^2_\epsilon \boldsymbol{I}_N$, 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

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = & - \frac{N}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^m \log |\boldsymbol{V}_i(\boldsymbol{\alpha})| \\ & - \frac{1}{2} \sum_{i=1}^m (\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{V}(\boldsymbol{\alpha})_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}). \end{aligned} \tag{18}$$

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

- $b_i \sim_{ind} \mathrm{N}(0, \sigma_0^2)$,
- $\epsilon_{ij} \sim_{ind} \mathrm{N}(0, \sigma_\epsilon^2)$

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 $\mathrm{E}[\boldsymbol{Y}] = \mathbf{1}_N \beta_0$, $\mathrm{var}(\boldsymbol{Y}) = \boldsymbol{V} = \mathbf{1}_N \mathbf{1}_N^{\mathrm{T}} \sigma_0^2 + \boldsymbol{I}_N \sigma_\epsilon^2 = \boldsymbol{J}_N \sigma_0^2 + \boldsymbol{I}_N \sigma_\epsilon^2$, where $\boldsymbol{J}_N$ is the $N \times N$ matrix of 1's.

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

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho & \dots & 0 & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho & \dots & 0 & 0 & 0 & 0 \\ \rho & \rho & 1 & \rho & \dots & 0 & 0 & 0 & 0 \\ \rho & \rho & \rho & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & \rho & \rho & \rho \\ 0 & 0 & 0 & 0 & \dots & \rho & 1 & \rho & \rho \\ 0 & 0 & 0 & 0 & \dots & \rho & \rho & 1 & \rho \\ 0 & 0 & 0 & 0 & \dots & \rho & \rho & \rho & 1 \end{bmatrix}$$

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 $(\beta_0, \sigma_0^2, \sigma_\epsilon^2)$, 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

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{y}_i \right), \tag{19}$$

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 (19), 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 $i = 1, ..., m$):

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

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

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

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 (18), 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 (REML). In general REML is carried out to reduce bias in estimation of variance components by "accounting for the estimation of $\boldsymbol{\beta}$".

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 $b_i \sim \text{N}(0, \sigma_0^2)$, 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.

Intuition: 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.

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.

Inference for Variance Components by REML

Restricted (or residual) 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

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

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

Hence

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

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

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

and

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

This leads to

$$\begin{aligned} p(\boldsymbol{U} \mid \boldsymbol{\alpha}) &= (2\pi)^{-(N-k-1)/2} \frac{\mid \boldsymbol{x}^{\text{T}} \boldsymbol{x} \mid^{1/2} \mid \boldsymbol{V} \mid^{-1/2}}{\mid \boldsymbol{x}^{\text{T}} \boldsymbol{V}^{-1} \boldsymbol{x} \mid^{1/2}} \\ &\times \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G)^{\text{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G) \right\} \end{aligned} \tag{20}$$

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:

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

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 a $4m \times 1$ vector and

$$\boldsymbol{\mu}_1 = ... = \boldsymbol{\mu}_m = (\beta_0 + \beta_1 t_1, \beta_0 + \beta_1 t_2, \beta_0 + \beta_1 t_3, \beta_0 + \beta_1 t_4)^{\mathrm{T}}.$$

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 variances and 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 3 shows the data plotted using a "trellis" plot – note that data are not plotted in the original order.

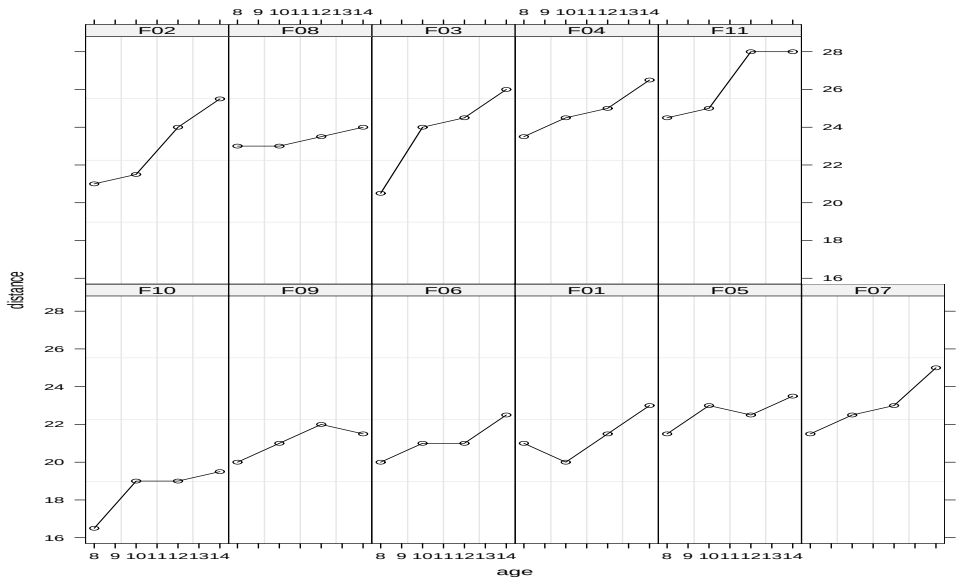

Figure 3: Length versus age (in years) for 11 girls.

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.

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 (16). 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 (16).

Now for comparison we fit the LMEM 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 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

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

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

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

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

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

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

This leads to $\widetilde{\boldsymbol{b}} = \text{E}[\boldsymbol{b} \mid \boldsymbol{y}]$, 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

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

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

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

$$\begin{aligned} \text{cov}_{\widetilde{b},b}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i) &= \text{E}_y[\text{cov}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i \mid \boldsymbol{y})] + \text{cov}_y(\text{E}[\widetilde{\boldsymbol{b}}_i \mid \boldsymbol{y}], \text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}]) \\ &= \text{E}_y[\text{cov}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i \mid \boldsymbol{y})] + \text{cov}_y(\widetilde{\boldsymbol{b}}_i, \widetilde{\boldsymbol{b}}_i) \qquad (21) \\ &= \text{var}(\widetilde{\boldsymbol{b}}_i) \end{aligned}$$

The first term in (21) 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, $\widetilde{\boldsymbol{b}} = \mathrm{E}[\boldsymbol{b} \mid \boldsymbol{y}]$, 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 $\mathrm{cov}(\boldsymbol{U}, \boldsymbol{V}) = \boldsymbol{C}$ is an $n \times m$ matrix with (i,j)-th element $\mathrm{cov}(U_i, V_j)$, $i = 1, ..., n; j = 1, ..., m$. Also $\mathrm{cov}(\boldsymbol{V}, \boldsymbol{U}) = \boldsymbol{C}^{\mathrm{T}}$. Now suppose $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{U}$ where $\boldsymbol{A}$ is an $m \times n$ matrix. Then $\mathrm{cov}(\boldsymbol{U}, \boldsymbol{A}\boldsymbol{U}) = \boldsymbol{W}\boldsymbol{A}^{\mathrm{T}}$ where $\boldsymbol{W} = \mathrm{cov}(\boldsymbol{U})$, and $\mathrm{cov}(\boldsymbol{A}\boldsymbol{U}, \boldsymbol{U}) = \boldsymbol{A}\boldsymbol{W}$.

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

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

since

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

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

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

$$\widetilde{\boldsymbol{b}}_i = \mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}_i] = \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) \tag{22}$$

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 fixed effect estimator towards zero, as we see in examples later.

The form (22) is not of practical use since it depends on the unknown $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$; instead we use

$$\widetilde{\boldsymbol{b}}_i = \mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}_i] = \widehat{\boldsymbol{D}}\boldsymbol{z}_i^{\mathrm{T}}\widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}). \tag{23}$$

Substitution of $\widehat{\boldsymbol{\beta}}$ is not such a problem (since it is an unbiased estimator, and appears in (22) in a linear fashion), but $\widehat{\boldsymbol{\alpha}}$ is more problematic.

The uncertainty in the prediction is given by

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

We have

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

and

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

Since

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

we have

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

and so

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

to give

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

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

We now examine fitted values:

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

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

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

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 the above form.

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
> b0hat <- b1hat <- NULL
> for (i in 1:11){
    x <- Orthgirl$age[seq((i-1)*4+1,(i-1)*4+4)]-11
    y <- Orthgirl$distance[seq((i-1)*4+1,(i-1)*4+4)]
    mod <- lm(y~x)
    b0hat[i] <- mod$coef[1]
    b1hat[i] <- mod$coef[2]
}
> index <- c(10,9,6,1,5,7,2,8,3,4,11)
> LSb0hat <- b0hat[index]; LSb1hat <- b1hat[index]
```


Shrinkage of Intercepts

```
> cbind(LSb0hat,LSb1hat,rcoef)
    LSb0hat LSb1hat (Intercept) I(age - 11)
F10  18.500   0.450    18.64240   0.4795455
F09  21.125   0.275    21.17728   0.4795455
F06  21.125   0.375    21.17728   0.4795455
F01  21.375   0.375    21.41869   0.4795455
F05  22.625   0.275    22.62578   0.4795455
F07  23.000   0.550    22.98791   0.4795455
F02  23.000   0.800    22.98791   0.4795455
F08  23.375   0.175    23.35003   0.4795455
F03  23.750   0.850    23.71216   0.4795455
F04  24.875   0.475    24.79853   0.4795455
F11  26.375   0.675    26.24704   0.4795455
```

Note ordering difference in coefficients from `lme`, and the slight shrinkage here towards the overall mean of 22.65; not much shrinkage here since $\widehat{\sigma}_0$ is large relative to $\widehat{\sigma}_\epsilon$ (see Figure 4).

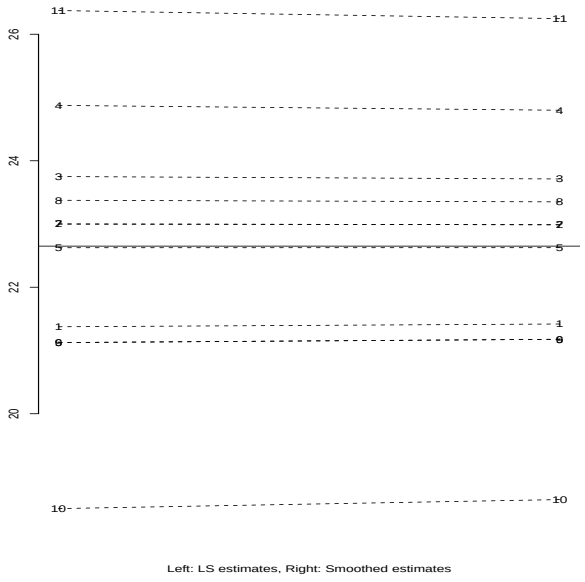

Figure 4: Least squares estimates and smoothed estimates, $\widehat{\beta}_0 + \widetilde{b}_i$.

Dental Example: Boys and Girls Joint Analyses

Table 2 describes LMEMs applied 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 5 gives normal QQ plots of the LS estimates of intercepts and slopes, for boys and girls.

Figure 6 gives a scatter plot of the LS estimates of intercepts and slopes, for boys and girls.

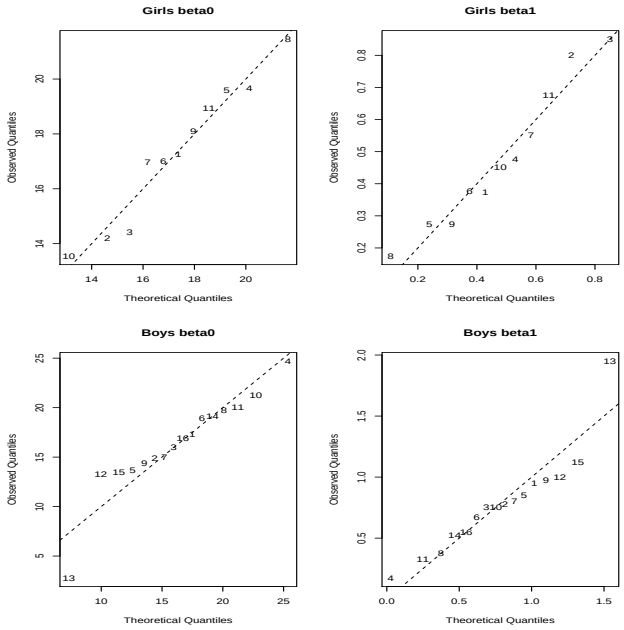

Figure 5: QQ plot of the LS estimates.

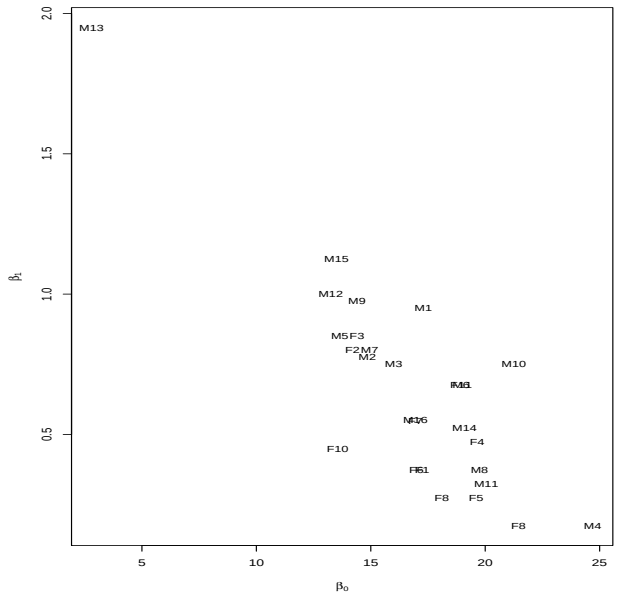

Figure 6: Plot of the LS estimates for boys and girls.