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, A420.

Coursework: (and approximate percentages) weekly (30%). Examination at mid-term (30%) and final (40%).

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: Cecilia Cotton (`ccotton@u`); office hour H657 11–1 Tuesdays, phone: 616–2767.

STAT/BIOSTAT 578 Data Analysis, strongly recommended for Applied Exam. 571 teaches 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.

Model Selection/Formulation

Types of analysis: descriptive, confirmatory, predictive. Causality and confounding.

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

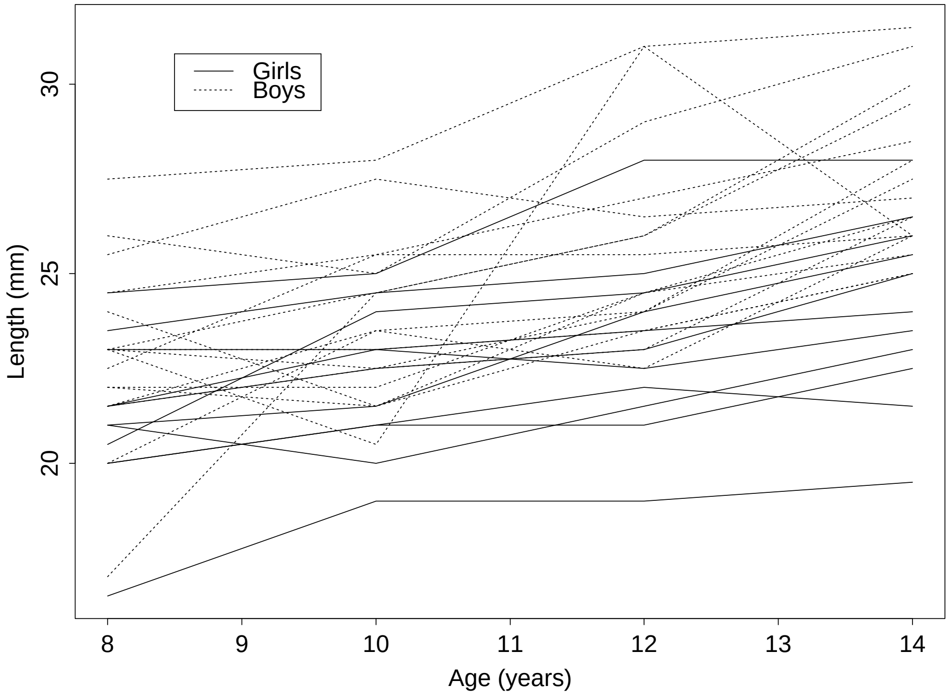

Figure 1: Dental growth data for girls and boys.

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 interval coverage.

A marginal approach to modeling specifies the moments of the data only, while in a conditional 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?

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

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

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

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

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

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

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

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

(asymptotic normality) where

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

and

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

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 $\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y}$ 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 was derived was incorrect.

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

$$\begin{aligned} \mathrm{var}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x})\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1} \\ &= (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2 \qquad (5) \end{aligned}$$

Expression (5) can also be derived directly from the estimating function

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

from which we know that

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

(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

$$\boldsymbol{B} = \mathrm{var}\{\boldsymbol{G}\} = \boldsymbol{x}^{\mathrm{T}}\mathrm{var}(\boldsymbol{Y})\boldsymbol{x} = \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i, \qquad (6)$$

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

$$\widehat{\boldsymbol{B}}_n = \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i (Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}})^2, \qquad (7)$$

and its use provides a consistent estimator of (6), 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

$$\widetilde{\sigma}_i^2 = \frac{n}{n-p} (Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}})^2 \qquad (8)$$

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

$$\sqrt{n}\ (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \rightarrow_d \mathrm{N}_p(\mathbf{0}, \boldsymbol{I}^{-1}) \qquad (9)$$

(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

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

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

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y}_1, \boldsymbol{y}_2)}. \tag{10}$$

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

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{y}_1)}{p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1)}. \tag{11}$$

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

$$p(\theta_i \mid \boldsymbol{y}) = \int_{\boldsymbol{\theta}_{-i}} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}_{-i}, \tag{12}$$

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

$$\mathrm{E}[\theta_i \mid \boldsymbol{y}] = \int_{\theta_i} \theta_i p(\theta_i \mid \boldsymbol{y}) \, \mathrm{d}\theta_i. \tag{13}$$

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

$$\int_{-\infty}^{\theta_i(q)} p(\theta_i \mid \boldsymbol{y}) \, \mathrm{d}\theta_i. \tag{14}$$

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

$$p(\boldsymbol{z} \mid \boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}. \tag{15}$$

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, (12)–(15), given that for most models, these are analytically intractable

Example: Normally Distributed Data 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 $w = n/(n+k)$.

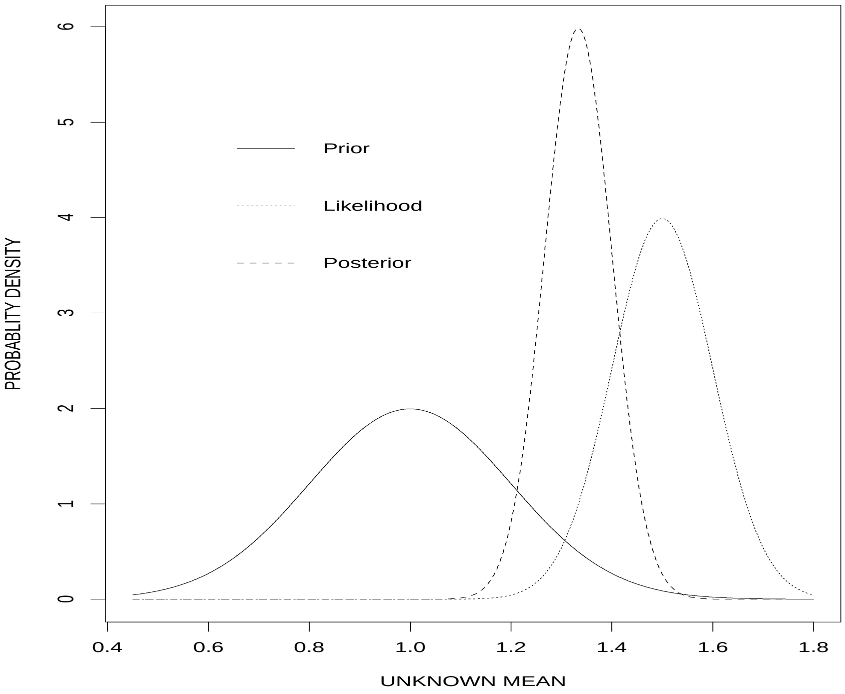

Figure 2: Normal likelihood ($\bar{y}$=1.5,n=10,σ=1), normal prior (m=1, k=5) and the resultant normal posterior.

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

$$\pi(\boldsymbol{\theta}) \propto 1 \tag{16}$$

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 (16) 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.

Example: Binomial Likelihood

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.

Example: Non-Linear Model

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

$$Y_i \mid \theta \sim_{ind} N\{\exp(-\theta x_i), \sigma^2\}, \tag{17}$$

$i = 1, ..., n$, 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 $\theta \rightarrow \infty$,

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

a constant, so that the posterior is improper.

The second aspect of priors 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 (16) 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

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

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, 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

$$\begin{bmatrix} \sigma_1 & & & \\ \rho_{12} & \sigma_2 & & \\ \rho_{13} & \rho_{23} & \sigma_3 & \\ \rho_{14} & \rho_{24} & \rho_{34} & \sigma_4 \end{bmatrix} \tag{18}$$

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 (18) by

$$\begin{bmatrix} 2.12 & & & \\ 0.83 & 1.90 & & \\ 0.86 & 0.90 & 2.36 & \\ 0.84 & 0.88 & 0.95 & 2.44 \end{bmatrix} \qquad (19)$$

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

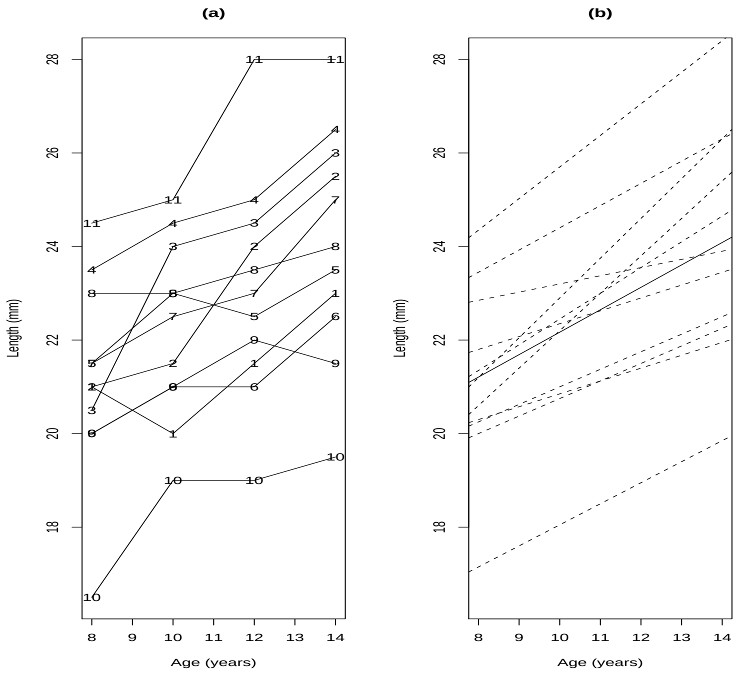

Figure 3: Dental plots for girls only: (a) Individual observed data (with plotting symbol girl index), (b) Individual fitted curves (dashed) and overall fitted curve (solid).

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:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{20}$$

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

Stage 2: Model for random terms:

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

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

The two stages define the marginal model:

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

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

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

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 $\boldsymbol{V} = \boldsymbol{z}\boldsymbol{D}\boldsymbol{z}^{\text{T}} + \sigma_\epsilon^2 \boldsymbol{I}_N$, 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:

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

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

$$\begin{aligned} l(\boldsymbol{\beta},\boldsymbol{\alpha}) = &- \frac{N}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{V}_i(\boldsymbol{\alpha})| \\ &- \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\alpha})_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}). \qquad (21) \end{aligned}$$

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

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

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

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

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

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

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{y}_i \right), \qquad (22)$$

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 (22), 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

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

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

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

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

$$\boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \left[\begin{array}{cc} \boldsymbol{I}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{\alpha\alpha} \end{array}\right].$$

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

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

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 (21), 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.

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.