As discussed in Chapter 2 there are two main methods for modeling such data.:

- In the *conditional* method, the data are modeled conditional upon a vector of unit-specific *random effects*; the distribution of both the data and the random effects are explicitly assumed. In this case a likelihood function is available and either likelihood or Bayesian methods can be used for inference. Such an approach is known as mixed effects modeling, and in this chapter we consider Linear Mixed Effects Models (LMEMs).
- In the *marginal* method, inference is based on specification of an estimating function. This approach is known as Generalized Estimating Equations (GEE).

Historically "profile analysis" was carried out. In this approach, which is applied to balanced data, replication across units is exploited to estimate all n variances $\text{var}(Y_{ij})$, $j = 1, ..., n$, and $n(n-1)/2$ covariances, $\text{cov}(Y_{ij}, Y_{ik})$, $j, k = 1, ..., n$, $j \neq k$. Replication across units implies they are "homogeneous" in some sense, and not for example mixtures of experimental groups/covariate stratum.

Motivation for LMEMs

Consider a longitudinal example (e.g. dental growth). There are two extreme fixed effects approaches:

- Assume a single curve and a standard analysis that would be carried out for independent data.

  Problem: Must acknowledge dependence within units.
- Assume m fixed curves and analyze each separately.

  Problems:
  - No "borrowing of strength", that is, each individual's fit is based only on their own data, and not on those of other individuals. We would hope that if there is *similarity* between the curves, that the totality of data will aid in the estimation of each individual curve. In some instances this may be vital, for example, if $n_i = 1$ for a particular individual, then their own data alone will not allow estimation of parameters.
  - Unable to make formal inference for population.

  If we have abundant data for a specific individual for whom we wish to make inference then these problems not relevant.

The second model is appealing in allowing each individual their own curve, a mixed effects model assumes that at least some parameters of the curves are drawn from a random effects distribution.

Known as a *Conditional Approach*.

Motivation for GEEs

In general the specification of a random effects model is not trivial, and it may be difficult to check that a form that is close to the “truth” has been fitted.

In non-linear models the interpretation of fixed effects also depends on the particular random effects distribution that has been fitted.

An alternative approach that reduces the need for the correct specification of the variance-covariance model (at least for non-small samples), and provides an unambiguous interpretation of fixed effects, is provided by Generalized Estimating Equations.

Known as a *Marginal Approach*.

We illustrate some of the issues with likelihood, Bayesian and GEE inference via an example.

Growth Curve Example

Notation

Let Y_{ij} denote the growth at occasion t_j for girl i, $i = 1, ..., m = 11$, $j = 1, ..., 4$, with $t_1 = 8$, $t_2 = 10$, $t_3 = 12$, $t_4 = 14$, so that $n = 4$ for all i.

Consider the marginal residuals

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

$i = 1, ..., 11; j = 1, ..., 4$. Let

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

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

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

$j = 1, ..., 4$, and

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

$j, k = 1, ..., 4; j \neq k,$.

We empirically estimate these as:

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

where $\widehat{\sigma}_j$ is the empirical standard deviation of the residuals at time t_j, and $\widehat{\rho}_{jk}$ is the empirical correlation between residuals at times t_j and t_j.

Standard deviations appear relatively constant across time.

Correlations approximately constant.

Figure 5(a) shows the raw growth data, and the solid line in (b) is the marginal fit, with (c) showing the residuals from this fit – there is clear dependence across residuals for each individual.

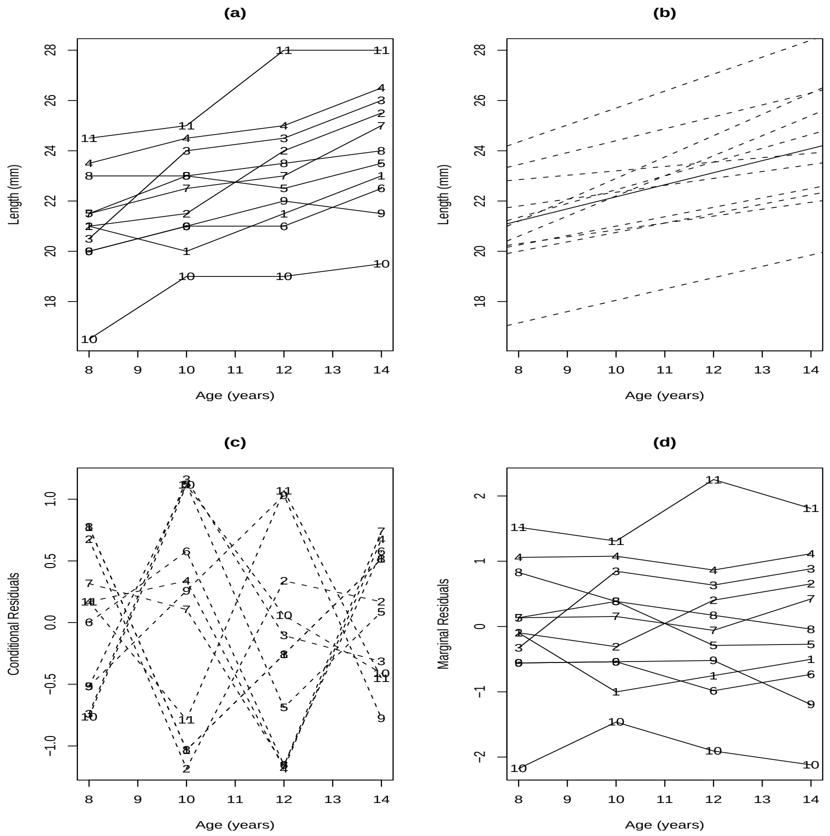

Figure 5: Dental plots for girls only: (a) Individual observed data, (b) Individual (conditional) fitted curves (dashed) and overall (marginal) fitted curve (solid), (c) Conditional residuals, (d) Marginal residuals.

Conditional correlations/s.d.'s of the conditional residuals

$$e_{ij}^c = \frac{Y_{ij} - \widehat{\beta}_{i0}^c - \widehat{\beta}_{i1}^c t_j}{\widehat{\sigma}_i^c},$$

$i = 1, ..., 11; j = 1, ..., 4$, are given by:

$$\begin{bmatrix} 0.59 & & & \\ -0.90 & 0.92 & & \\ -0.13 & -0.31 & 0.81 & \\ 0.66 & -0.26 & -0.84 & 0.53 \end{bmatrix}$$

Weird patten of correlations due to dependence within residuals; we only have two independent pieces of information for each fit.

If we considered boys and girls together and ignored gender we would have induced dependence also.

Conditional Model

We have

$$Y_{ij} = \mathrm{E}[Y_{ij}|\beta_0, \beta_1, b_i] + \epsilon_{ij},$$

where the ϵ_{ij} are iid with $\mathrm{E}[\epsilon_{ij}] = 0$ and $\mathrm{var}(\epsilon_{ij}) = \sigma_\epsilon^2$, and

$$\mathrm{E}[Y_{ij}|\beta_0, \beta_1, b_i] = \beta_0 + b_i + \beta_1 t_j,$$

where $b_1, ..., b_m$ represent *random effects* that are assigned a distribution.

We then assume the b_i are iid with $\mathrm{E}[b_i] = 0$ and $\mathrm{var}(b_i) = \sigma_0^2$. We also assume that the ϵ_{ij} and b_i are independent.

The intercepts (or equivalently baseline length) are assumed to be girl-specific, and given by $\beta_{0i} = \beta_0 + b_i$, but the growth rate, β_1, is the same for all girls.

Note that 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.

From the conditional model we can derive the marginal model by integrating over the random effect:

$$p(\boldsymbol{y}|\beta_0, \beta_1, \sigma_\epsilon^2, \sigma_0^2) = \int_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{b}, \beta_0, \beta_1, \sigma_\epsilon^2, \sigma_0^2) \times p(\boldsymbol{b}|\beta_0, \beta_1, \sigma_\epsilon^2, \sigma_0^2) \, d\boldsymbol{b}$$

Exploiting conditional independencies we have:

$$p(\boldsymbol{y}|\beta_0, \beta_1, \sigma_\epsilon^2, \sigma_0^2) = \prod_{i=1}^m \int_{b_i} p(\boldsymbol{y}_i|b_i, \beta_0, \beta_1, \sigma_\epsilon^2) \times p(b_i|\sigma_0^2) \, db_i.$$

To specify a likelihood we require distributional assumptions on the random variables ϵ_{ij}, b_i.

It is convenient to assume:

$$\begin{aligned} \epsilon_{ij} &\sim_{iid} N(0, \sigma_\epsilon^2) \\ b_i &\sim_{iid} N(0, \sigma_0^2), \end{aligned}$$

since in this case the marginal distribution of the data, which is simply a function of the fixed effects $\beta_0, \beta_1, \sigma_\epsilon^2, \sigma_0^2$, is normal (since a convolution of normals is normal); hence all we need are the moments.

First moment:

$$\mathrm{E}[\boldsymbol{Y}|\beta_0, \beta_1, \sigma_\epsilon^2, \sigma_0^2] = \boldsymbol{\mu},$$

where

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

is $4m \times 1$ vector and

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

Second moment:

$$\text{var}(\boldsymbol{Y}|\beta_0, \beta_1, \sigma_\epsilon^2, \sigma_0^2) = \boldsymbol{V},$$

where $\boldsymbol{V}$ is the $4m \times 4m$ matrix,

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

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

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

Likelihood

We write

$$\boldsymbol{V} = \sigma^2 \begin{bmatrix} \boldsymbol{R}_1 & \boldsymbol{0} & ... & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_2 & ... & \boldsymbol{0} \\ ... & ... & ... & ... \\ \boldsymbol{0} & \boldsymbol{0} & ... & \boldsymbol{R}_m \end{bmatrix}$$

where $\sigma^2 = \sigma_\epsilon^2 + \sigma_0^2$ and $\boldsymbol{R}_i$ is an $n_i \times n_i$ correlation matrix with 1's on the diagonal, and off-diagonal elements $\rho = \sigma_0^2/(\sigma_\epsilon^2 + \sigma_0^2)$.

We saw that the data followed a normal distribution which gives the likelihood function

$$L(\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_0^2) = p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{V}) = \prod_{i=1}^{m} p(\boldsymbol{y}_i|\boldsymbol{\mu}_i, \boldsymbol{V}_i),$$

with

$$\boldsymbol{y}_i|\boldsymbol{\mu}_i, \boldsymbol{V}_i \sim_{ind} N(\boldsymbol{\mu}_i, \boldsymbol{V}_i),$$

and $\boldsymbol{\mu}_i$ the $n_i \times 1$ mean vector, $i = 1, ..., m$.

This may now be maximized to obtain the MLEs of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_0^2)$.

The explicit form of the likelihood is

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

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

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

which yields

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

To derive the score equations for the variance components $\sigma_\epsilon^2, \sigma_0^2$ we first list some identities, see Searle, Casella and McCulloch (1992), *Variance Components*, Wiley, Appendices M(e) and M(f), p. 456–457.

$$\begin{aligned} \frac{\partial}{\partial \sigma_\delta^2}\log|\boldsymbol{V}_i| &= \mathrm{tr}\left(\boldsymbol{V}_i^{-1}\frac{\partial \boldsymbol{V}_i}{\partial \sigma_\delta^2}\right), \\ \frac{\partial}{\partial \sigma_\delta^2}\boldsymbol{V}_i^{-1} &= -\boldsymbol{V}_i^{-1}\frac{\partial \boldsymbol{V}_i}{\partial \sigma_\delta^2}\boldsymbol{V}_i^{-1}, \end{aligned}$$

where $\delta = \epsilon$ or 0.

Hence we have

$$\begin{aligned} \frac{\partial l}{\partial \sigma_0^2} &= -\frac{1}{2}\sum_{i=1}^m \mathrm{tr}\left(\boldsymbol{V}_i^{-1}\frac{\partial \boldsymbol{V}_i}{\partial \sigma_0^2}\right) \\ &+ \frac{1}{2}\sum_{i=1}^m (\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}_i^{-1}\frac{\partial \boldsymbol{V}_i}{\partial \sigma_0^2}\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}), \end{aligned}$$

and similarly for $\frac{\partial l}{\partial \sigma_\epsilon^2}$.

The MLEs for $\sigma_\epsilon^2, \sigma_0^2$ are only available in closed form in rare cases and so numerical methods are turned to, for example the function `lme` in `R`.

Prediction of Random Effects

In the frequentist approach, since random effects have a distinct status from fixed effects, their estimation is often viewed as a *prediction* problem.

One development is as follows: to predict $\boldsymbol{b} = (b_1, ..., b_m)^{\mathrm{T}}$ using an estimator $\tilde{\boldsymbol{b}}(\boldsymbol{y}) = (\tilde{b}_1(\boldsymbol{y}), ..., \tilde{b}_m(\boldsymbol{y}))^{\mathrm{T}}$ we may wish to minimize the mean-squared prediction error:

$$\mathrm{E}_{\boldsymbol{y},\boldsymbol{b}}\left[\{\tilde{\boldsymbol{b}}(\boldsymbol{y}) - \boldsymbol{b}\}^{\mathrm{T}}\{\tilde{\boldsymbol{b}}(\boldsymbol{y}) - \boldsymbol{b}\}\right] =$$

$$\int_{\boldsymbol{y}}\int_{\boldsymbol{b}}\{\tilde{\boldsymbol{b}}(\boldsymbol{y}) - \boldsymbol{b}\}^{\mathrm{T}}\{\tilde{\boldsymbol{b}}(\boldsymbol{y}) - \boldsymbol{b}\}p(\boldsymbol{b} \mid \boldsymbol{y})p(\boldsymbol{y})\ d\boldsymbol{b}\ d\boldsymbol{y}$$

and to minimize wrt to $\tilde{\boldsymbol{b}}$ we only need to minimize the inner integral.

Minimizing with respect to $\tilde{\boldsymbol{b}}(\boldsymbol{y})$ we obtain

$$\widehat{\boldsymbol{b}}(\boldsymbol{y}) = \mathrm{E}[\boldsymbol{b}|\boldsymbol{y}],$$

which has various names including the Best Linear Unbiased Predictor (BLUP).

We assume for the moment that $\boldsymbol{\theta}$ is known. Then since

$$\mathrm{cov}(b_i, \boldsymbol{y}_i) = \mathrm{cov}(b_i, \boldsymbol{x}_i\boldsymbol{\beta} + \mathbf{1}_{n_i}b_i) = \mathbf{1}_{n_i}\sigma_\alpha^2,$$

where $\mathbf{1}_{n_i}$ is the $n_i \times 1$ vector of 1's, we have

$$\left[\begin{array}{c} b_i \\ \boldsymbol{y}_i \end{array}\right] \sim N\left(\left[\begin{array}{c} 0 \\ \boldsymbol{x}_i\boldsymbol{\beta} \end{array}\right], \left[\begin{array}{cc} \sigma_0^2 & \mathbf{1}_{n_i}^{\mathrm{T}}\sigma_0^2 \\ \mathbf{1}_{n_i}\sigma_0^2 & \boldsymbol{V}_i \end{array}\right]\right).$$

We have

$$b_i|\boldsymbol{y}_i \sim N\{\mathbf{1}_{n_i}^{\mathrm{T}}\sigma_0^2\boldsymbol{V}_i(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}), [1 - \mathbf{1}_{n_i}^{\mathrm{T}}\sigma_0^2\boldsymbol{V}_i^{-1}\mathbf{1}_{n_i}]\sigma_0^2\}.$$

Hence the BLUP is

$$\widehat{b_i} = \mathrm{E}[b_i|\boldsymbol{y}_i] = \mathbf{1}_{n_i}^{\mathrm{T}}\sigma_0^2\boldsymbol{V}_i(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}).$$

In practice an estimator for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_0^2, \sigma_\epsilon^2)$ is plugged in, to give

$$\widehat{b_i} = \mathrm{E}[b_i|\boldsymbol{y}_i, \widehat{\boldsymbol{\theta}}] = \mathbf{1}_{n_i}^{\mathrm{T}}\widehat{\sigma}_0^2\widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})$$

but the variance of the BLUP does not reflect the uncertainty in $\widehat{\boldsymbol{\theta}}$.

Since the estimator is of the form $\mathrm{E}[\boldsymbol{b}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}]$ it is known as an *empirical Bayes* estimator.

Note also that $\widehat{b_i}$ is a weighted combination of the residuals of the $i-$th individual's data.

Bayesian Approach

Recall $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_0^2)$.

The posterior for $\boldsymbol{\theta}, \boldsymbol{b}$ is given by

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{b}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{b})p(\boldsymbol{\theta}, \boldsymbol{b})}{p(\boldsymbol{y})} \\ &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{b})p(\boldsymbol{b}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \end{aligned}$$

Alternatively we may work with the marginal likelihood:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}.$$

For the random effects we have

$$\begin{aligned} p(\boldsymbol{b}|\boldsymbol{y}) &= \int p(\boldsymbol{b}, \boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta} \\ &= \int p(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta} \qquad (1) \end{aligned}$$

and

$$\mathrm{E}[\boldsymbol{b}|\boldsymbol{y}] = \int \left\{ \int \boldsymbol{b} p(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\theta}) \; d\boldsymbol{b} \right\} p(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta} = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\{\mathrm{E}[\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\theta}]\}$$

showing that we should be averaging over $\boldsymbol{\theta}$, rather than conditioning on $\widehat{\boldsymbol{\theta}}$, as is done in empirical Bayes. If the posterior for $\boldsymbol{\theta}|\boldsymbol{y}$ is peaked about $\widehat{\boldsymbol{\theta}}$ then conditioning will be a good approximation.

Marginal Model

As an alternative approach we could just specify (say) the first two moments of the data, for example, the forms assumed on the previous page, but without the assumptions on the data, or the use of random effects.

We begin with the following assumptions:

(a) $\mathrm{E}[\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{V}] = \boldsymbol{x}\boldsymbol{\beta}$ and

(b) $\mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{V}) = \boldsymbol{V}$,

where $\boldsymbol{Y} = (\boldsymbol{Y}_1, ..., \boldsymbol{Y}_m)$, $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})$, $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_m)$ is $N \times (k+1)$ with $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i})^{\mathrm{T}}$, $\boldsymbol{x}_{ij} = (1, x_{ij1} \; ... \; x_{ijk})$, $N = \sum_i n_i$ and $\boldsymbol{\beta}$ is $(k+1) \times 1$.

Estimating Function:

Consider the estimating function

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

where $\boldsymbol{W}$ is an $N \times N$ *weight matrix.*

It follows immediately that

$$\mathrm{E}[\boldsymbol{G}(\boldsymbol{\beta})] = \boldsymbol{0}$$

and defining

$$\boldsymbol{G}(\widehat{\boldsymbol{\beta}}_W) = \boldsymbol{0}$$

yields the generalized least squares estimator

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

so that

$$\mathrm{E}[\widehat{\boldsymbol{\beta}}_W] = \boldsymbol{\beta}$$

regardless of $\boldsymbol{W}$ (so long as $\boldsymbol{W}$ is non-singular).

We also have

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}_W) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}^{-1}\boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}^{-1}\boldsymbol{x})^{-1}.$$

Special Cases:

If $\boldsymbol{W} = \boldsymbol{V}$ then

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}_V) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}.$$

An extension of the Gauss-Markov Theorem shows that $\widehat{\boldsymbol{\beta}}_V$ is the most *efficient* linear estimator.

Note that $\widehat{\boldsymbol{\beta}}_V$ is identical to the MLE derived earlier under the assumption that $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{x}\boldsymbol{\beta}, \boldsymbol{V})$.

If $\boldsymbol{W} = \boldsymbol{I}$ (known as *working independence*) then

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}_I) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}.$$

In both cases we need to know $\boldsymbol{V}$ – later we will see how this may be empirically estimated, using the fact that we have replication across i.

Note: clearly wrong to say that

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