Inference for Variance Components by REML

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

Marginal Likelihood

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

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

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

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

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

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

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

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

Example: Normal linear model

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

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

Hence the marginal likelihood is

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

We know

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

and so

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

to give

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

and

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

REML for LMEM

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

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

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

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

We have

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

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

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

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

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

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

Then

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

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

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

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

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

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

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

is the generalized least squares estimator.

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

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

Then

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

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

We have

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

and

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

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

$$\begin{aligned} p(\boldsymbol{U} \mid \boldsymbol{\alpha}) &= (2\pi)^{-(N-k-1)/2} \frac{\mid \boldsymbol{x}^{\text{T}} \boldsymbol{x} \mid^{1/2} \mid \boldsymbol{V} \mid^{-1/2}}{\mid \boldsymbol{x}^{\text{T}} \boldsymbol{V}^{-1} \boldsymbol{x} \mid^{1/2}} \\ &\times \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G)^{\text{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G) \right\} \end{aligned} \tag{23}$$

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 $4m \times 1$ vector and

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

The variance is given by

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

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

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

with $\sigma^2 = \sigma^2_\epsilon + \sigma^2_0$ and $\rho = \frac{\sigma^2_0}{\sigma^2} = \frac{\sigma^2_0}{\sigma^2_\epsilon + \sigma^2_0}$. Hence the random intercepts model induces a marginal form with constant variances and constant orrelations 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 4 shows the data plotted using a "trellis" plot – note that data are not plotted in the original order.

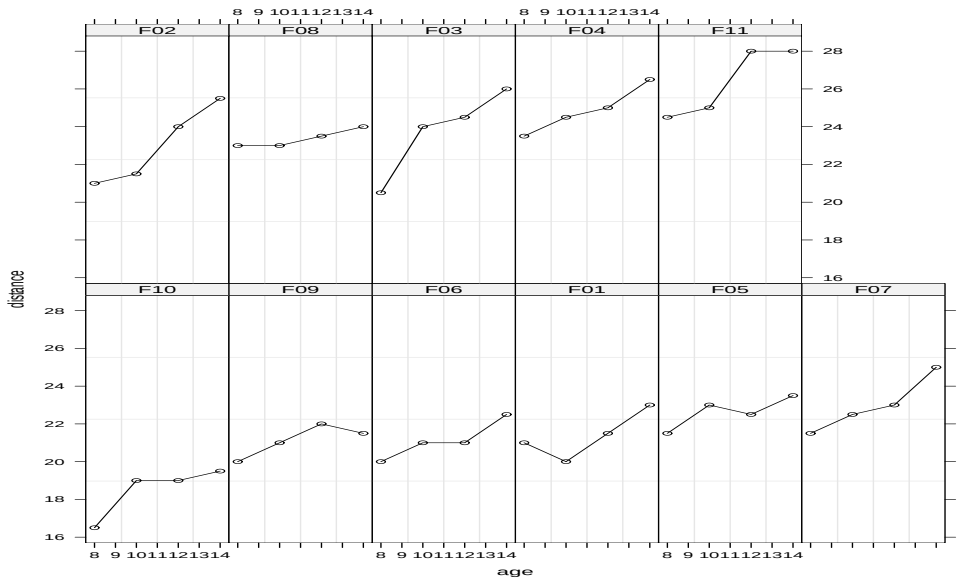

Figure 4: Length versus age (in years) for 11 girls.

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 (19). 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 (19).

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

$$\begin{aligned} \text{cov}_{\widetilde{b},b}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i) &= \text{E}_y[\text{cov}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i \mid \boldsymbol{y})] + \text{cov}_y(\text{E}[\widetilde{\boldsymbol{b}}_i \mid \boldsymbol{y}], \text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}]) \\ &= \text{E}_y[\text{cov}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i \mid \boldsymbol{y})] + \text{cov}_y(\widetilde{\boldsymbol{b}}_i, \widetilde{\boldsymbol{b}}_i) \qquad (24) \\ &= \text{var}(\widetilde{\boldsymbol{b}}_i) \end{aligned}$$

The first term in (24) 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:

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

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

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

The form (25) is not of practical use since it depends on the unknown $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$; instead we use

$$\widetilde{\boldsymbol{b}}_i = \mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}_i] = \widehat{\boldsymbol{D}}\boldsymbol{z}_i^{\mathrm{T}}\widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}). \tag{26}$$

Substitution of $\widehat{\boldsymbol{\beta}}$ is not such a problem (since it is an unbiased estimator, and appears in (25) 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 of (25) can be justified in a number of ways, other than MSE.

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

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

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

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

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

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

but instead the BLUP is the mean of the distribution

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

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

Example: Dental Growth

We again fit a LMEM with random intercepts only.

```
> remlelm <- lme(distance~I(age-11),data = Orthgirl,random = ~1 | Subject)
> summary(remlelm)
 Formula: ~1 | Subject
        (Intercept)  Residual
StdDev:     2.06847 0.7800331
                Value Std.Error DF t-value p-value
(Intercept) 22.647727 0.6346568 32 35.6850       0
I(age - 11)  0.479545 0.0525898 32  9.1186       0
> 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 5).

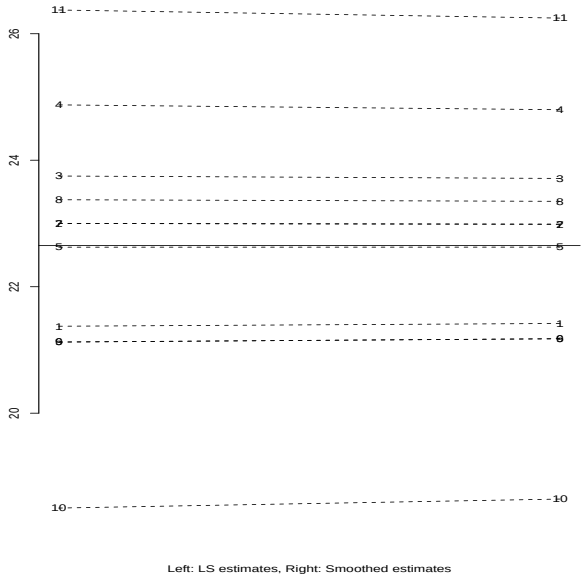

Figure 5: Least squares estimates and smoothed estimates, $\widehat{\beta}_0 + \widetilde{b}_i$.

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

Figure 7 gives a scatter plot of the LS estimates of intercepts and slopes, for boys and girls.

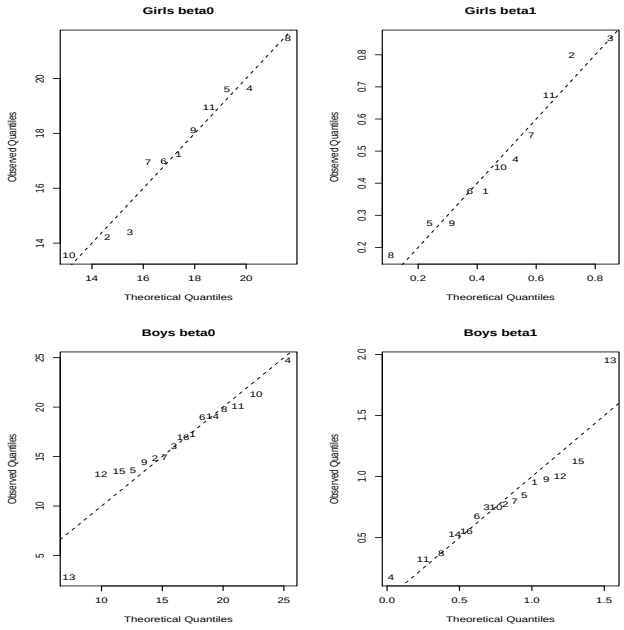

Figure 6: QQ plot of the LS estimates.

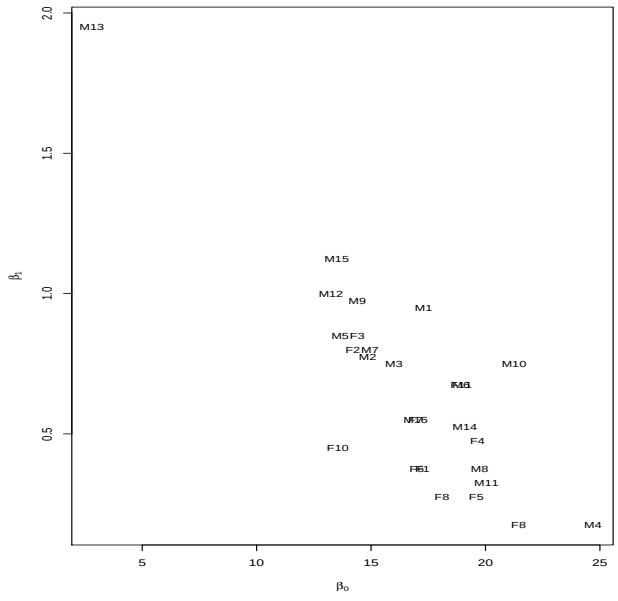

Figure 7: Plot of the LS estimates for boys and girls.

Bayesian Inference for the LMEM

Consider the model

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

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

The form of the posterior follows from exploiting conditional independencies:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b} \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b})\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}) = \prod_{i=1}^m p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}_i)\pi(\boldsymbol{b} \mid \boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\alpha}) \\ &= \prod_{i=1}^m \left\{p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}_i)\pi(\boldsymbol{b}_i \mid \boldsymbol{\alpha})\right\} \pi(\boldsymbol{\beta})\pi(\boldsymbol{\alpha}) \qquad (27) \end{aligned}$$

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

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

where the integrand is giving by the term in curly brackets in (27).

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

Exchangeability

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

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

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

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

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

where

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

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

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

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

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

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

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

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

Hence

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

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

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

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

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

as $N \to \infty$.

Implications

The interpretation of this theorem is of great significance:

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

In conventional language, we have the *likelihood function*

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

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

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

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

for $1 \leq m < n$ where

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

and

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

Proof

Write

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

and then use the previous result on numerator and denominator.

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

Further results

General Representation Theorem:

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

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

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

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

Relevance of Exchangeability

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

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

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

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

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

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

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

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

Bayesian Computation

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

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

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

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

Example: Binomial Likelihood with Weird Functions of Interest

Suppose we have

$$Y_j \mid p_j \sim \text{Binomial}(n_j, p_j)$$

$j = 1, 2$, with independent priors

$$p_j \sim U(0, 1)$$

The posteriors are available analytically as

$$p_j \mid y_j \sim \text{Beta}(y_j + 1, n_j - y_j + 1)$$

but suppose we are interested in inference for the odds ratio

$$\phi = \frac{p_1}{1 - p_1} / \frac{p_2}{1 - p_2}$$

and for the relative risk

$$\theta = \frac{p_1}{p_2}$$

The following is `R code` to simulate from

$$p_1 \mid y_1, y_2 \text{ and } p_2 \mid y_1, y_2$$

and

$$\phi \mid y_1, y_2 \text{ and } \theta \mid y_1, y_2$$

when

$$n_1 = 35, n_2 = 45, y_1 = 30, y_2 = 10$$

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


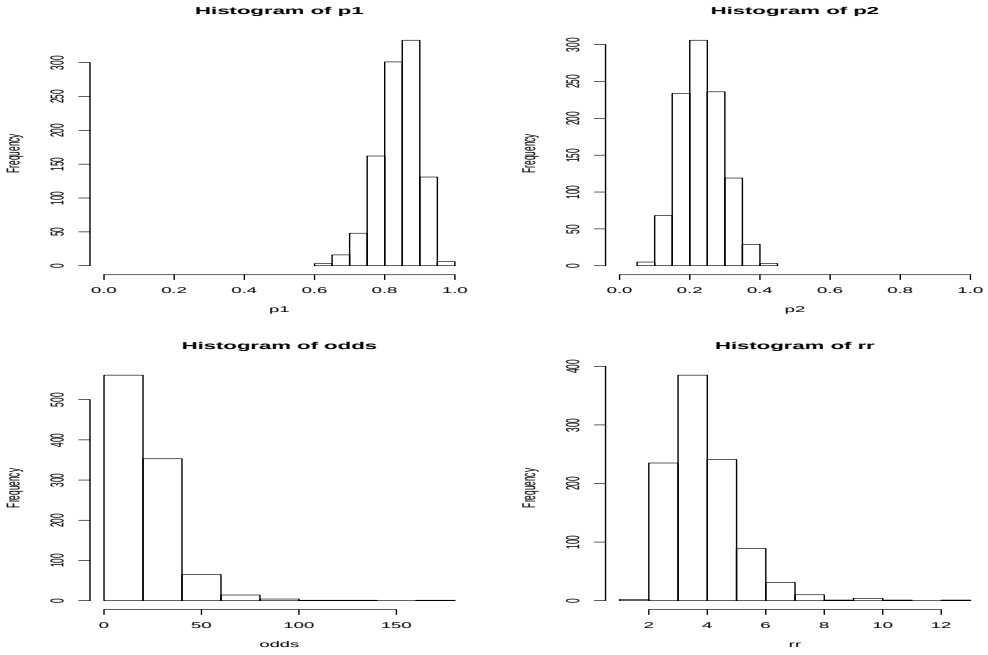

Figure 8: Posterior distributions for p_1, p_2, the odds ratio $\frac{p_1}{1-p_1} / \frac{p_2}{1-p_2}$ and for the relative risk $\theta = \frac{p_1}{p_2}$.

The Composition Method

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

Write the joint posterior distribution for θ_1, θ_2 as

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

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

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

Markov chain Monte Carlo

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

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

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

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

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

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

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

Markov chains over a discrete parameter space

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

Then the stationary distribution π is defined by

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

where π is a $1 \times K$ row vector.

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

Gibbs Sampling

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

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

We have

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

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

Gibbs sampling proceeds by iterating between the steps:

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

and

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

to produce the sequence

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

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

Gibbs Sampling over a discrete parameter space

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

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

Table 4: Joint posterior distribution.

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

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

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

Example: Normal likelihood, unknown mean and variance

Likelihood:

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

Prior:

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

Posterior

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

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

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

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

where

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

and

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

Metropolis Algorithm – discrete parameter space

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

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

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

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

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

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

  otherwise stay at x.

This results in the transition matrix

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

Metropolis Algorithm – continuous parameter space

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

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

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

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

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

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

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

Metropolis-Hastings Algorithm

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

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

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

$$r = \frac{\pi(\boldsymbol{\theta}^*)/g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s-1)})}{\pi(\boldsymbol{\theta}^{(s-1)})/g(\boldsymbol{\theta}^{(s-1)}|\boldsymbol{\theta}^*)}.$$

- Set

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

Issues:

- Convergence of the Markov chain?
- Parameterization.

Convergence

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

$$E[g(\theta)] \approx \frac{1}{n-m} \sum_{s=m+1}^{n} g(\theta^{(s)})$$

Convergence Diagnosis

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

Parameterization

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

In an extreme case, if we have independence then

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

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

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

Hyperpriors

Consider the LMEM

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

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

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

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

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

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

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

The Wishart Distribution

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

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

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

$$\boldsymbol{W} = \sum_{j=1}^{r} \boldsymbol{Z}_j \boldsymbol{Z}_j^{\mathrm{T}}. \tag{28}$$

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

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

where

$$c = 2^{rp/2}\Gamma_p(r/2) \mid \boldsymbol{S} \mid^{r/2}, \tag{29}$$

with

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

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

The mean is given by

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

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

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

for $w > 0$, the gamma distribution $\mathrm{Ga}(r/2, 1/(2S))$. Further, taking $S = 1$ gives a χ^2_r random variable, which is clear from (28).

The Inverse Wishart Distribution

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

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

where c is again given by (29). The mean is given by

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

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

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

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

and

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

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

Issues with the Wishart Prior

- A problem with the Wishart distribution is that it is deficient in second moment parameters since there is only a single degrees of freedom parameter r. So, for example, it is not possible to have differing levels of certainty in the tightness of the prior distribution for different elements of $\boldsymbol{D}$. With diagonal $\boldsymbol{D}$ and independent inverse gamma priors we have a precision parameter for each variance.
- The form of the conditional distribution suggests that it may be better to err on the side of picking $\boldsymbol{R}$ too small (if m small, prior always influential).
- Intuition: as if our prior data for the precision consists of observing r normal random variables with variance-covariance matrices $\boldsymbol{R}$.
- We need to take $r \geq q+1$ for a proper prior, with the flattest prior corresponding to $r = q+1$. A proper prior is required to ensure propriety of the posterior distribution.
- Figure 9 displays samples from the Wishart distribution $W_2\{20, (20\boldsymbol{S})^{-1}\}$ where $\boldsymbol{S} = \begin{bmatrix} 0.4 & 0 \\ 0 & 1.0 \end{bmatrix}$. The mean is $\mathrm{E}[\boldsymbol{W}] = \boldsymbol{S}^{-1} = \begin{bmatrix} 2.5 & 0 \\ 0 & 1.0 \end{bmatrix}$.

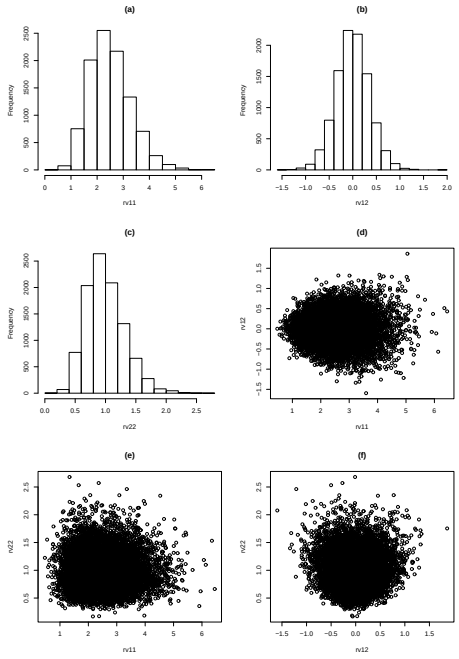

Figure 9: Histograms of (a) w_{11}, (b) w_{12}, (c) w_{22}, scatterplots of (d) w_{11}, w_{12}, (e) w_{11}, w_{22}, w_{12}, w_{22}

Conditional Conjugacy

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

The joint posterior is

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

with priors:

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

and derive the required conditional distributions:

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

Conditional for $\boldsymbol{\beta}$

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

Conditional for τ

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

Conditional for $\boldsymbol{\beta}_i$

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

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

Conditional for $\boldsymbol{W}$

First note that

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

Then

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

Hence the conditional distribution is

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

Example: Dental Data for Girls

Three-Stage Hierarchical Model:

First Stage:

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

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

Second Stage: Let

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

and then

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

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

Third Stage:

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

Results below are for priors, with prior mean

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

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

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

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

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

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

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

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

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

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

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