Again consider the model

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

with $\mathrm{E}[\boldsymbol{b}] = \boldsymbol{0}$.

Consider an estimator $\widetilde{\boldsymbol{b}}$; $\boldsymbol{b}$ is a random variable and so rather than use minimum variance as a criteria (variance is about a fixed point), we use minimum MSE as a criteria. Specifically

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

for non-singular $\boldsymbol{A}$ leads to $\widetilde{\boldsymbol{b}} = \mathrm{E}[\boldsymbol{b} \mid \boldsymbol{y}]$, irrespective of $\boldsymbol{A}$ (see Exercises 4).

Properties of $\widetilde{\boldsymbol{b}}$:

$$\mathrm{E}_Y\{\widetilde{\boldsymbol{b}}\} = \mathrm{E}_y\{\mathrm{E}[\widetilde{\boldsymbol{b}} \mid \boldsymbol{y}]\} = \boldsymbol{0}.$$

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.

If we assume

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

to give

$$\begin{aligned} \widetilde{\boldsymbol{b}}_i = \mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}_i] &= \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}), \\ \mathrm{var}(\boldsymbol{b}_i|\boldsymbol{y}_i) &= \boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\boldsymbol{D}. \end{aligned}$$

Rather than assume normality we could consider estimators that are *linear* in $\boldsymbol{y}$. In Exercises 4 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, $\mathrm{E}[\boldsymbol{b}_i|\boldsymbol{y}_i]$ not necessarily linear in $\boldsymbol{y}$.

Example: One-way ANOVA:

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

$b_i \sim \text{N}(0, \sigma_0^2)$, $\epsilon_{ij} \sim \text{N}(0, \sigma_\epsilon^2)$. Can also write as

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

We show that

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

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.

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

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

We now examine the second moment properties of our estimator via

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

We have

$$\text{cov}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i) = \text{E}_Y[\text{cov}(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_i \mid \boldsymbol{y})] + \text{cov}_Y(\text{E}[\widetilde{\boldsymbol{b}}_i \mid \boldsymbol{y}], \text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}]) = \text{var}(\widetilde{\boldsymbol{b}}_i),$$

so that

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

In lectures we show that

$$\begin{aligned} \text{var}(\widetilde{\boldsymbol{b}}) &= \boldsymbol{D}\boldsymbol{z}^{\text{T}}\boldsymbol{V}^{-1}\{\boldsymbol{V} - \boldsymbol{x}(\boldsymbol{x}^{\text{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}\}\boldsymbol{V}^{-1}\boldsymbol{z}\boldsymbol{D} \\ &= \boldsymbol{D}\boldsymbol{z}^{\text{T}}\{\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{x}(\boldsymbol{x}^{\text{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}\boldsymbol{V}^{-1}\}\boldsymbol{z}\boldsymbol{D}. \end{aligned}$$

We now examine fitted values:

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

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

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

We can also write

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

In the Bayesian approach the posterior distribution for $\boldsymbol{b} \mid \boldsymbol{y}$ may be reported, or one may concentrate on summaries such as posterior quantiles.

The uncertainty in estimation of $\boldsymbol{\beta}, \boldsymbol{\alpha}$ is acknowledged if one approximates the required integrals by analytical, numerical or simulation techniques.

For example, an MCMC approach would sample from the conditional:

$$p(\boldsymbol{b}_i \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \propto p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}_i) \times p(\boldsymbol{b}_i \mid \boldsymbol{D}),$$

which leads to

$$\begin{aligned}\boldsymbol{b}_i \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\alpha} \sim \mathrm{N}_{q+1}\{&(\boldsymbol{z}_i^{\mathrm{T}}\sigma_\epsilon^{-2}\boldsymbol{z}_i + \boldsymbol{D}^{-1})^{-1}\boldsymbol{z}_i^{\mathrm{T}}\sigma_\epsilon^{-2}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}),\\ &(\boldsymbol{z}_i^{\mathrm{T}}\sigma_\epsilon^{-2}\boldsymbol{z}_i + \boldsymbol{D}^{-1})^{-1}\}.\end{aligned}$$

The identity

$$(\boldsymbol{T}+\boldsymbol{U}\boldsymbol{V}^{-1}\boldsymbol{W})^{-1} = \boldsymbol{T}^{-1} - \boldsymbol{T}^{-1}\boldsymbol{U}(\boldsymbol{V}^{-1}+\boldsymbol{W}\boldsymbol{T}^{-1}\boldsymbol{U})^{-1}\boldsymbol{W}\boldsymbol{T}^{-1}$$

gives

$$\begin{aligned}(\boldsymbol{z}_i^{\mathrm{T}}\sigma_\epsilon^{-2}\boldsymbol{z}_i + \boldsymbol{D}^{-1})^{-1} &= \boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}(\boldsymbol{I}\sigma^{-2} + \boldsymbol{z}_i\boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}})^{-1}\boldsymbol{z}_i\boldsymbol{D}\\ &= \boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\boldsymbol{D}\end{aligned}$$

yields

$$\begin{aligned}\boldsymbol{b}_i \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\alpha} \sim \mathrm{N}_{q+1}\{&(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i\sigma_\epsilon^{-2} + \boldsymbol{D}^{-1})^{-1}\boldsymbol{z}_i^{\mathrm{T}}\sigma_\epsilon^{-2}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}),\\ &\boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\boldsymbol{D}\}.\end{aligned}$$

This shows how the variance is changed from the prior uncertainty of $\boldsymbol{D}$ to posterior uncertainty.

The uncertainty over $\boldsymbol{\beta}, \boldsymbol{\alpha}$ is acknowledged over the Gibbs iterates. Note: could derive immediately from joint distribution of $\boldsymbol{y}_i, \boldsymbol{b}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}$.

Restricted Maximum Likelihood Estimation (REML)

We assume the model

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

with $\boldsymbol{b}^{\mathrm{T}} = (\boldsymbol{b}_1, ..., \boldsymbol{b}_m)$, $\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I})$ and $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are independent;

- $\boldsymbol{Y}$ is $N \times 1$,
- $\boldsymbol{x}$ is $N \times (k+1)$,
- $\boldsymbol{\beta}$ is $(k+1) \times 1$,
- $\boldsymbol{z}$ is $N \times (q+1)$,
- $\boldsymbol{b}$ is $(q+1) \times 1$
- $\boldsymbol{\epsilon}$ is $N \times 1$.

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

REML is a method that has been proposed to rectify this problem – there are a number of justifications; we have already seen a Bayesian justification, we now provide another based on marginal likelihood.

The overall rationale is: 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

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

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 $N-k-1$.

Hence we take $N-k-1$ linearly independent residuals:

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

where $\boldsymbol{B}$ is any $N \times (N-k-1)$ matrix with $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} = \boldsymbol{I}-\boldsymbol{H}$ and $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{I}$ so that

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

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 so 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 first 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. $\begin{vmatrix} \boldsymbol{T} & \boldsymbol{U} \\ \boldsymbol{V} & \boldsymbol{W} \end{vmatrix} =\mid \boldsymbol{T} \mid\mid \boldsymbol{W} - \boldsymbol{V}\boldsymbol{T}^{-1}\boldsymbol{U} \mid .$

Then

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

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

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

and

$$\mathrm{E}[\boldsymbol{U}] = \boldsymbol{0},$$

and

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

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

Hence

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

Recall the definition of marginal likelihood. Let $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, be minimal sufficient statistics for which

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

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

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

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

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

Hence

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

We have

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

and

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

This leads to

$$\begin{aligned} p(\boldsymbol{U} \mid \boldsymbol{\alpha}) &= (2\pi)^{-(N-k-1)/2} \frac{\mid \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x} \mid^{1/2} \mid \boldsymbol{V} \mid^{-1/2}}{\mid \boldsymbol{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{x} \mid^{1/2}} \\ &\times \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G)^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_G) \right\} (11) \end{aligned}$$

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

Notes on REML

- 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 likelihood (the REML likelihood) which may then be maximized as a function of $\boldsymbol{\alpha}$.
- The log restricted likelihood is, upto a constant,

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

  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 to test whether the smaller model is a valid statistical simplification of the larger model.
- Likelihood ratio tests for variance components are valid.

Hypothesis tests for variance components

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

As an example, consider the test of $H_0 : D = 0$ versus $H_A : D = \sigma_0^2$, where σ_0^2 is a non-negative scalar. In this case the asymptotic null distribution is a 50:50 mixture of χ_0^2 and χ_1^2 distributions, where the former is the distribution that gives probability mass 1 to the value 0.

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

The intuition on the null distribution is that, under the null, half of the time the correlation will be estimated as $\widehat{\rho} = 0$.

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 longer sequence and

$$\begin{aligned} \Pr\left(\sum_{i=1}^{n} y_i = z_n\right) &= \sum_{Z_N=z_n}^{N-(n-z_n)} \Pr(y_1 + ... + y_n = z_n, Y_1 + ... + Y_N = z_N) \\ &= \sum_{Z_N=z_n}^{N-(n-z_n)} \Pr(y_1 + ... + y_n = z_n \mid Y_1 + ... + Y_N = z_N) \\ &\times \quad \Pr(Y_1 + ... + Y_N = z_N). \end{aligned}$$

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

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

Here $\Pr(z_N)$ is the "prior" belief in the number of successes out of N. We now let $N \rightarrow \infty$ such that $z_N/N = \theta$. Then 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

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

as $N \rightarrow \infty$.

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 $\pi()$.
- By the strong law of large numbers, $\lim_{n\to\infty} Z_n/n$, so that 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: $\boldsymbol{\phi} \sim_{ind} \pi(\cdot).$

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.

Predictive distributions in the context of hierarchical models

Predictive distributions may be obtained for random variables at any level of the hierarchy.

For example in the hierarchy:

Stage 1: $Y_i|\theta_i,$, $i = 1, ..., m$,

Stage 2: $\theta_i|\phi$, $i = 1, ..., m$,

Stage 3: ϕ,

we can obtain a predictive distribution for the parameters of a new unit, θ^*, assumed to be from the same population (at stage 2) via

$$p(\theta^*|y) = \int p(\theta^*|\phi)\pi(\phi|y)\mathrm{d}\phi,$$

and for an observation, y^*, for this new unit via:

$$p(y^*|y) = \int p(y^*|\theta^*)\pi(\theta^*|y)\mathrm{d}\theta^*.$$

Simulation from these predictive distributions is straightforward given samples from the posterior. For example:

$$p(\theta^\star|y) \approx \frac{1}{S}\sum_{s=1}^{S} p(\theta^\star|\phi^{(s)}),$$

where $\phi^{(s)} \sim \pi(\phi|y)$.