Maximum Likelihood solution via EM - random intercepts model

- E-step (expectation - "Expected A Posteriori" or Empirical Bayes)

$$\tilde{\upsilon}_i \;=\; \rho_{n_i n_i} \left[\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{\beta}\right]$$

$$\sigma^2_{\upsilon|y_i} \;=\; \sigma^2_\upsilon(1-\rho_{n_i n_i}) \qquad \text{where } \rho_{n_i n_i} = \frac{n_i r}{1+(n_i-1)r} \quad \text{and} \quad r = \frac{\sigma^2_\upsilon}{\sigma^2_\upsilon + \sigma^2}$$

- M-step (maximization - "Maximium Likelihood")

$$\hat{\boldsymbol{\beta}} \;=\; \left(\sum_i^N \boldsymbol{X}'_i\boldsymbol{X}_i\right)^{-1} \sum_i^N \boldsymbol{X}'_i(\boldsymbol{y}_i - \mathbf{1}_i\tilde{\upsilon}_i)$$

$$\hat{\sigma}^2_\upsilon \;=\; \frac{1}{N}\sum_i^N \tilde{\upsilon}_i{}^2 + \sigma^2_{\upsilon|y_i}$$

$$\hat{\sigma}^2 \;=\; (\sum_i^N n_i)^{-1} \sum_i^N (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \mathbf{1}_i\tilde{\upsilon}_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \mathbf{1}_i\tilde{\upsilon}_i) \;+\; n_i\sigma^2_{\upsilon|y_i}$$

- provide starting values for $\boldsymbol{\beta}$, σ^2_υ, and σ^2
- perform E-step, perform M-step, repeat early and often (until convergence)

Wald z or χ^2 tests

$$z = \frac{\hat{\theta}}{se(\hat{\theta})}$$

is approximately normally distributed under the large sample theory of ML estimates

- for single parameter tests
  - *e.g.*, $H_0 : \beta_1 = 0$
- asymptotic test
- not generally recommended for testing variance terms
- can be generalized for multi-parameter tests

Note:

- $z^2 \sim \chi^2_1$, so either z or χ^2 representation
- since asymptotic, t-tests based on sample size have been proposed (in MIXED, the DDFM=KENWARDROGER option on the MODEL statement performs the degrees-of-freedom calculations detailed by Kenward and Roger, Biometrics, 1997)
- this approach also used to test hypotheses regarding random effects $\boldsymbol{v}_i$ (however, multiple tests issue)
- asymptotic confidence intervals for θ formed as

$$\hat{\theta} \pm z_{1-\alpha/2} \times se(\hat{\theta})$$

Likelihood-ratio χ^2 test

Suppose Model I is nested within Model II, then

$$2 \times \log(L_{\text{II}} \,/\, L_{\text{I}}) \;=\; 2 \times (\log L_{\text{II}} - \log L_{\text{I}}) \;\sim\; \chi^2_q$$

where q = number of additional terms in Model II

$-2 \log L$ is called the *deviance*, $D_{\text{I}} - D_{\text{II}} \;\sim\; \chi^2_q$

- for single or multiparameter tests
  e.g., $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$
  or $H_0 : \sigma^2_{\upsilon_0} = \sigma^2_{\upsilon_1} = \sigma_{\upsilon_0\upsilon_1} = 0$ (divide p-value by 2)
- good for comparing models *with same sample size*
  (be careful with missing values on $\boldsymbol{X}$ variables)
- With REML estimation - *cannot* use for testing fixed effects
  (*i.e.*, $\boldsymbol{\beta}$)

Maximum Likelihood Estimation - R.A. Fisher (1922, Foundations of Theoretical Statistics, *Philos. Trans. Royal Society of London, A*, 309-368)

"The object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data."

This object is accomplished by:

- constructing a hypothetical infinite population
- data regarded as constituting a random sample (from this population)
- distribution of hypothetical population is specified by relatively few parameters
- information from sample is used to estimate these parameters

The idea of maximum likelihood is to obtain estimates for these parameters which are most likely given the observed data

Sherlock Holmes as an early proponent of ML

We now join Dr. Watson and Sherlock Holmes in their room at No. 221B, Baker Street, in the opening scene of "The Sign of Four" already in progress

(*Watson remarks*) " . . . you spoke just now of observation and deduction. Surely the one to some extent implies the other."
"Why, hardly," he answered, leaning back luxuriously in his armchair and sending up thick blue wreaths from his pipe. "For example, observation shows me that you have been to the Wigmore Street Post-Office this morning, but deduction lets me know that when there you dispatched a telegram."
"Right!" said I. "Right on both points! But I confess that I don't see how you arrived at it. It was a sudden impulse upon my part, and I have mentioned it to no one."
"It is simplicity itself," he remarked, chuckling at my surprise,

"so absurdly simple an explanation is superfluous; and yet it may serve to define the limits of observation and of deduction. Observation tells me that you have a little reddish mould adhering to your instep. Just opposite the Wigmore Street Office they have taken up the pavement and thrown up some earth, which lies in such a way that it is difficult to avoid treading in it in entering. The earth is of this peculiar reddish tint which is found, as far as I know, nowhere else in the neighbourhood. So much is observation. The rest is deduction."

"How, then, did you deduce the telegram?"

"Why, of course, I knew that you had not written a letter, since I sat opposite to you all morning. I see also in your open desk there that you have a sheet of stamps and a thick bundle of postcards. What could you go into the post-office for, then, but to send a wire? *Eliminate all other factors, and the one which remains must be the truth.*"

For Holmes, deducing (estimating) where Watson went that morning (parameter of interest),

observation = data - red mould on shoe, red earth by post-office

deduction = parameter estimate - that Watson went to the post-office

Deducing (estimating) what Watson did there (other parameter of interest),

observation = data - Watson did not write a letter, Watson has stamps and postcards

deduction = parameter estimate - Watson sent a wire

$\Rightarrow$ *Method of Deduction = Maximum Likelihood since Holmes' recommendation is to choose the option (parameter) with the highest likelihood, given the data*

Likelihood Function $p(\boldsymbol{y}, \boldsymbol{\theta})$ *or* $p(\boldsymbol{y} \mid \boldsymbol{\theta})$

- $\boldsymbol{y}$ represents the observed data
- $\boldsymbol{\theta}$ represents the parameters which characterize hypothetical infinite population, $\boldsymbol{\theta}$ belonging to a set of possible values from $\boldsymbol{\Theta}$
- the likelihood function $p(\boldsymbol{y}, \boldsymbol{\theta})$ is the probability density at $\boldsymbol{y}$ when $\boldsymbol{\theta}$ are the true parameters; expresses plausibility of the parameters after having observed the data
- ML estimates are obtained by maximizing the likelihood function
- obtain $\hat{\boldsymbol{\theta}}$ with the highest plausibility given the data

Definition in Silvey (1975): MLE $\hat{\theta}(y)$ is any element of Θ such that $p\left\{y, \hat{\theta}(y)\right\} = \max_{\theta \in \Theta} p(y, \theta)$

Example: Linear Regression Model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

$i = 1 \ldots N$ subjects

or in matrix form,

$$\underset{N\times 1}{\boldsymbol{y}} = \underset{N\times p}{\boldsymbol{X}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{N\times 1}{\boldsymbol{\varepsilon}} \qquad \text{with } \varepsilon_i \sim \mathcal{NID}(0, \sigma^2)$$

$$p(\varepsilon_i \mid \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}\varepsilon_i^2\right]$$

as a result, $y_i \sim \mathcal{NID}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$ and the density of y_i is

$$p(y_i \mid \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right]$$

Since ε_i are independent

$$p(\boldsymbol{\varepsilon} \mid \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right]$$

and the density of $\boldsymbol{y}$ is

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]$$

Need to maximize $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ and σ^2 to give the most plausible parameter values, given the data

$p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is at a maximum (for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$) when

- derivative $p'(\boldsymbol{y} \mid \boldsymbol{\theta}) = 0$
- second derivative $p''(\boldsymbol{y} \mid \boldsymbol{\theta}) < 0$

The likelihood function is given as:

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]$$

maximizing the log-likelihood gives same result as maximizing the likelihood, however,

- logs are our friends & can be natural
- easier to differentiate the log-likelihood

$$\begin{aligned}
\log L &= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sigma^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2}\sigma^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= C - \frac{N}{2}\log(\sigma^2) - \frac{1}{2}\sigma^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})
\end{aligned}$$

Differentiating with respect to σ^2

$$\begin{aligned}\frac{\partial \log L}{\partial \sigma^2} &= -\frac{N}{2}\sigma^{-2} + \frac{1}{2}\sigma^{-4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \frac{1}{2}\sigma^{-2}\left[\sigma^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - N\right]\end{aligned}$$

and setting equal to zero,

$$\hat{\sigma}^2 = \frac{1}{N}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Notice that the above ML estimator yields a biased estimate when sample size is small and/or p is large:

$$\hat{\sigma}^2_{OLS} = \frac{1}{N - p}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$\left(\frac{\partial \log L}{\partial \sigma^2}\right)^2 = \frac{N}{2}\sigma^{-4} - \sigma^{-6}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$= \sigma^{-4}\left[\frac{N}{2} - \sigma^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]$$

and for this to be a maximum,

$$0 > \frac{N}{2} - \sigma^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$> \sigma^2 - \frac{2}{N}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$> \sigma^2 - 2\sigma^2$$

$$> -\sigma^2$$

Similarly, differentiating

$$\log L = C - \frac{N}{2}\log(\sigma^2) - \frac{1}{2}\sigma^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

with respect to $\boldsymbol{\beta}$, yields

$$\underset{p\times 1}{\frac{\partial \log L}{\partial \boldsymbol{\beta}}} = -\frac{1}{2\sigma^2}\ \{\underset{p\times N}{(-\boldsymbol{X})'}\ \underset{N\times 1}{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})} + [\underset{1\times N}{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'}\ \underset{1\times p}{(-\boldsymbol{X})}]'\}$$

and since $(\boldsymbol{A}\boldsymbol{B})' = \boldsymbol{B}'\boldsymbol{A}'$

$$= \frac{1}{2}\sigma^{-2}\left[\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]$$

$$= \sigma^{-2}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

setting equal to zero,

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Notice

$$\left(\frac{\partial \log L}{\partial \boldsymbol{\beta}}\right)^2 = -(\boldsymbol{X}'\boldsymbol{X})$$

which always contains diagonal elements < 0

The mixed-effects regression model for unit i:

$$\underset{n_i\times 1}{\boldsymbol{y}_i} = \underset{n_i\times p}{\boldsymbol{X}_i}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n_i\times r}{\boldsymbol{Z}_i}\ \underset{r\times 1}{\boldsymbol{\upsilon}_i} + \underset{n_i\times 1}{\boldsymbol{\varepsilon}_i}$$

$i = 1 \ldots N$ units ; $\quad j = 1 \ldots n_i$ observations within unit i

with $\boldsymbol{\varepsilon} \sim \mathcal{NID}(0, \sigma^2 \boldsymbol{I}_{n_i})$ and $\boldsymbol{\upsilon} \sim \mathcal{NID}(0, \boldsymbol{\Sigma}_\upsilon)$

As a result, the observations $\boldsymbol{y}$ and random coefficients $\boldsymbol{\upsilon}$ have the joint multivariate normal distribution:

$$\begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{\upsilon}_i \end{bmatrix} \sim \mathcal{NID}\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{Z}_i\boldsymbol{\Sigma}_\upsilon\boldsymbol{Z}_i' + \sigma^2\boldsymbol{I}_{n_i} & \boldsymbol{Z}_i\boldsymbol{\Sigma}_\upsilon \\ \boldsymbol{\Sigma}_\upsilon\boldsymbol{Z}_i' & \boldsymbol{\Sigma}_\upsilon \end{bmatrix} \right)$$

Parameter estimation

- Empirical Bayes (EB) for random-effects $\boldsymbol{\upsilon}_i$
- Maximum (marginal) likelihood (ML) for variance parameters, σ^2 and $\boldsymbol{\Sigma}_\upsilon$, and covariate effects $\boldsymbol{\beta}$

Articles about Estimation

Bock, R.D. (1989). Measurement of human variation: A two-stage model. In R.D. Bock (Ed.), *Multilevel analysis of educational data.* New York: Academic Press.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, **76**: 341-353.

Laird, N. M. & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**: 963 - 974.

Empirical Bayes Estimation

Morris, C.N. (1983) Parametric Empirical Bayes Inference: Theory and Applications. *JASA*, **78**:47-55.

EB inference concerns two stochastic processes

- one for the data $f(y_i \mid \boldsymbol{\upsilon}_i; \boldsymbol{\zeta})$
  - f is the likelihood function (probability density)
  - describes probability of the data associated with the random effects $\boldsymbol{\upsilon}_i$ and parameters $\boldsymbol{\zeta}$ ($\boldsymbol{\beta}$ and σ^2)
- one for the random effects $g(\boldsymbol{\upsilon}; \boldsymbol{\eta})$
  - g is the prior probability density
  - $\boldsymbol{\eta}$ characterizes the distribution ($\boldsymbol{\Sigma}_\upsilon$)

All of the information about $\boldsymbol{\upsilon}_i$ that is available in y_i is conveyed by the posterior probability or probability density of $\boldsymbol{\upsilon}$, given y_i, as expressed in Bayes Theorem:

$$p(\boldsymbol{\upsilon} \mid y_i) = \frac{f(y_i \mid \boldsymbol{\upsilon}_i; \boldsymbol{\zeta})\, g(\boldsymbol{\upsilon}; \boldsymbol{\eta})}{h(y_i)}$$

where

$$h(y_i) = \int_{\boldsymbol{\upsilon}} f(y_i \mid \boldsymbol{\upsilon}_i; \boldsymbol{\zeta})\, g(\boldsymbol{\upsilon}; \boldsymbol{\eta})\, d\boldsymbol{\upsilon}$$

is the marginal probability of the observation

The posterior has all the information about $\boldsymbol{\upsilon}_i$, but what information do you want from the posterior?

- the mean or expected value $\tilde{\boldsymbol{\upsilon}}_i$

$$\tilde{\boldsymbol{\upsilon}}_i = \int_{\boldsymbol{\upsilon}} \boldsymbol{\upsilon}\, p(\boldsymbol{\upsilon} \mid y_i)\, d\boldsymbol{\upsilon}$$

- the variance-covariance matrix $\boldsymbol{\Sigma}_{\upsilon|y_i}$

$$\boldsymbol{\Sigma}_{\upsilon|y_i} = \int_{\boldsymbol{\upsilon}} (\boldsymbol{\upsilon} - \tilde{\boldsymbol{\upsilon}}_i)(\boldsymbol{\upsilon} - \tilde{\boldsymbol{\upsilon}}_i)'\, p(\boldsymbol{\upsilon} \mid y_i)\, d\boldsymbol{\upsilon}$$

How do we get these quantities?

- posterior is conditional distribution of $\boldsymbol{\upsilon}$ given $\boldsymbol{y}_i$
- $\boldsymbol{\upsilon}$ and $\boldsymbol{y}_i$ jointly multivariate normally distibuted
- use property of multivariate normal distribution (Rao, Bock, *etc.*,), conditional distribution of $\boldsymbol{a}_2$ given $\boldsymbol{a}_1$ is $\mathcal{NID}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{a}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$

Using this property, the posterior is given by

$$\boldsymbol{\upsilon} \mid \boldsymbol{y}_i \sim \mathcal{NID}\Big[\boldsymbol{\Sigma}_\upsilon \boldsymbol{Z}_i'(\boldsymbol{Z}_i\boldsymbol{\Sigma}_\upsilon\boldsymbol{Z}_i' + \sigma^2\boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \ ,$$

$$\boldsymbol{\Sigma}_\upsilon - \boldsymbol{\Sigma}_\upsilon\boldsymbol{Z}_i'(\boldsymbol{Z}_i\boldsymbol{\Sigma}_\upsilon\boldsymbol{Z}_i' + \sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\boldsymbol{\Sigma}_\upsilon\Big]$$

The posterior mean provides the EB estimate of $\boldsymbol{\upsilon}_i$

The posterior variance-covariance matrix provides the EB estimate of uncertainty about $\boldsymbol{\upsilon}_i$

however, the forms above require inversion of $n_i \times n_i$ matrices

Using the really cool property (Dempster *et al.*, 81; Bock, 83)

$$\boldsymbol{\Sigma}_{\upsilon}\boldsymbol{Z}_i'(\boldsymbol{Z}_i\boldsymbol{\Sigma}_{\upsilon}\boldsymbol{Z}_i'+\sigma^2\boldsymbol{I}_{n_i})^{-1} = (\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i+\boldsymbol{\Sigma}_{\upsilon}^{-1})^{-1}\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}$$

which only requires inversion of $r \times r$ matrices

Now, the mean of the posterior distribution of $\boldsymbol{\upsilon}$, given $\boldsymbol{y}_i$, is

$$\begin{aligned}\tilde{\boldsymbol{\upsilon}}_i &= (\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i + \boldsymbol{\Sigma}_{\upsilon}^{-1})^{-1}\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\ &= (\boldsymbol{Z}_i'\boldsymbol{Z}_i + \sigma^2\boldsymbol{\Sigma}_{\upsilon}^{-1})^{-1}\boldsymbol{Z}_i'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\end{aligned}$$

with the variance-covariance matrix as

$$\boldsymbol{\Sigma}_{\upsilon|y_i} = (\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i + \boldsymbol{\Sigma}_{\upsilon}^{-1})^{-1}$$

These are the Empirical Bayes or EAP estimates ("Expected A Posteriori") of the random effects with corresponding posterior variance terms (sometimes PSDs are given instead - positive square roots of the diagonal elements of $\mathbf{\Sigma}_{\upsilon|y_i}$)

From RIESBYM2.SAS, add `S` (or `SOLUTION`) on `RANDOM` statement:

```
PROC MIXED METHOD=ML;
     CLASS id;
     MODEL hamd = week / SOLUTION ;
     RANDOM INT week / SUBJECT=id TYPE=UN G GCORR S;
```

Solution for Random Effects

Obs	Effect	id	Estimate	StdErr Pred	DF	tValue	Probt
1	Intercept	101	1.0290	2.0432	243	0.50	0.6150
2	week	101	-2.1035	0.6995	243	-3.01	0.0029
3	Intercept	103	3.6392	2.0432	243	1.78	0.0761
4	week	103	-0.4743	0.6995	243	-0.68	0.4983
5	Intercept	104	2.6386	2.0432	243	1.29	0.1978
6	week	104	-1.4891	0.6995	243	-2.13	0.0343
.	.	.	.	.	.	.	.

- `Estimate` column provides $\tilde{\boldsymbol{\upsilon}}_i$ (posterior means)
- `StdErr Pred` column provides square root of diagonal elements of $\boldsymbol{\Sigma}_{\upsilon|y_i}$ (posterior std devs)

Suppose a random-intercepts models

$$\begin{aligned}
\tilde{v}_i &= (\mathbf{1}_i'\mathbf{1}_i + \sigma^2/\sigma_v^2)^{-1}\, \mathbf{1}_i'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\
&= (n_i + \sigma^2/\sigma_v^2)^{-1}\, \mathbf{1}_i'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\
&= \left(\frac{n_i\sigma_v^2 + \sigma^2}{\sigma_v^2}\right)^{-1} \mathbf{1}_i'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\
&= \left(\frac{\sigma_v^2}{\sigma_v^2 + \sigma^2/n_i}\right)\, \frac{1}{n_i}\, \mathbf{1}_i'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\
&= \left(\frac{n_i\sigma_v^2}{n_i\sigma_v^2 + \sigma^2}\right)\, \frac{1}{n_i}\, \mathbf{1}_i'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})
\end{aligned}$$

$$= \left(\frac{n_i\sigma_\upsilon^2}{\sigma_\upsilon^2+\sigma^2}\right)\left(\frac{1}{1+(n_i-1)\frac{\sigma_\upsilon^2}{\sigma_\upsilon^2+\sigma^2}}\right)\frac{1}{n_i}\,\mathbf{1}_i'(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})$$

$$= \left(\frac{n_i r}{1+(n_i-1)r}\right)\frac{1}{n_i}\,\mathbf{1}_i'(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})$$

$$= \rho_{n_in_i}\,\frac{1}{n_i}\,\mathbf{1}_i'(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}) \;=\; \rho_{n_in_i}\,\frac{1}{n_i}\sum_j^{n_i}(y_{ij}-\boldsymbol{x}_{ij}'\boldsymbol{\beta})$$

$$= \rho_{n_in_i}\,\left[\bar{\boldsymbol{y}}_{i.}-\bar{\boldsymbol{x}}_{i.}'\boldsymbol{\beta}\right] \;=\; \rho_{n_in_i}\,\hat{\upsilon}_{i\,OLS}$$

where $\rho_{n_in_i} = n_i r/\left[1+(n_i-1)r\right]$ is Spearman-Brown reliability

suppose $r = .5$

$$\rho_{n_i n_i} = \frac{.5 n_i}{.5[2 + (n_i - 1)]} \;=\; n_i/(n_i + 1)$$

if $n_i = 1 \;\Rightarrow\; \rho_{n_i n_i} = .5$ \qquad while if $n_i = 5 \;\Rightarrow\; \rho_{n_i n_i} = 5/6$

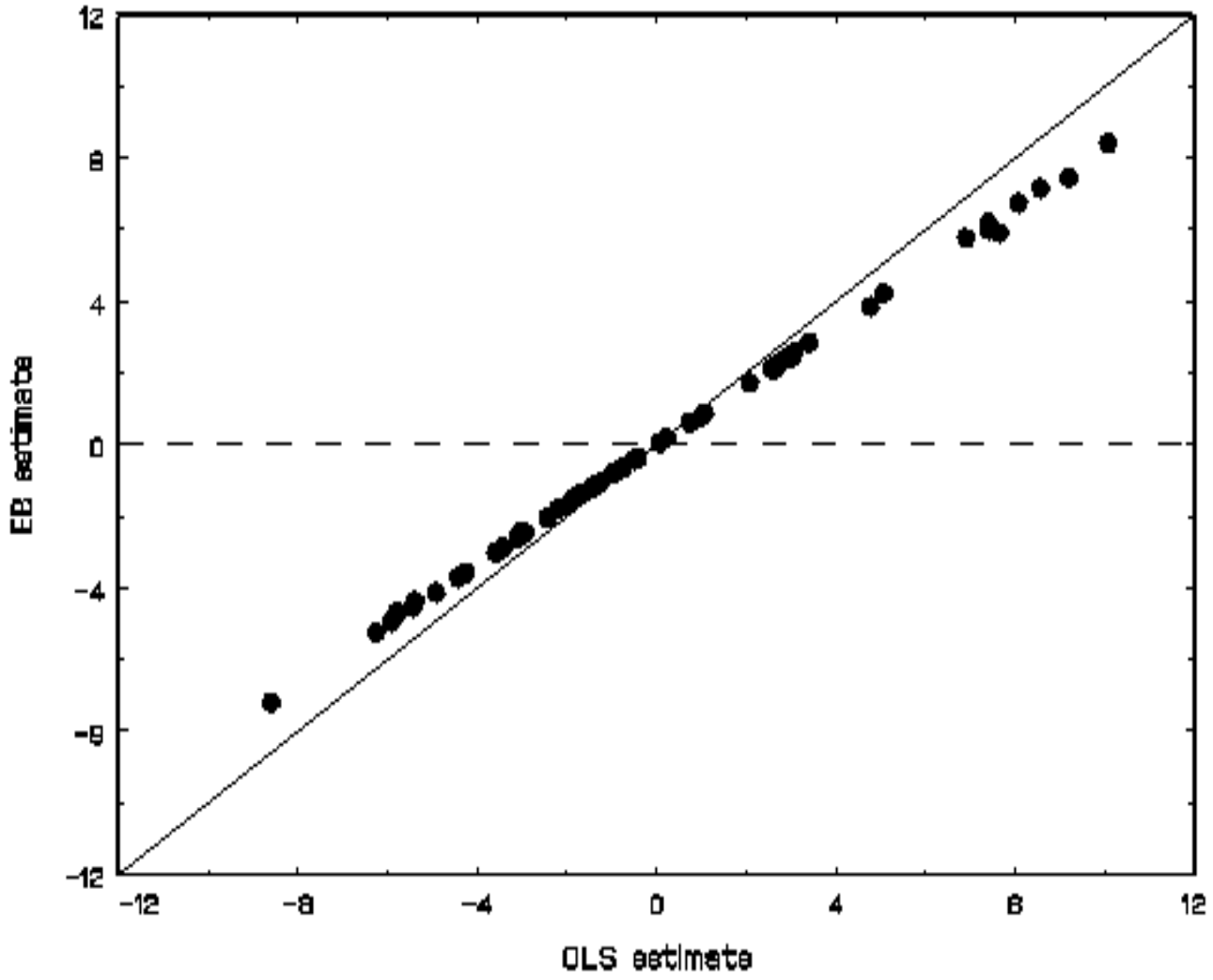
as $n_i \uparrow$ then $\rho_{n_i n_i} \approx 1$

suppose $r = .75$

$$\rho_{n_i n_i} = \frac{n_i}{4/3 + (n_i - 1)} \;=\; n_i/(n_i + 1/3)$$

if $n_i = 1 \;\Rightarrow\; \rho_{n_i n_i} = 3/4$ \qquad while if $n_i = 5 \;\Rightarrow\; \rho_{n_i n_i} = 5/5.33$

as $r \uparrow$ then $\rho_{n_i n_i}$ also goes to 1



Plot of the EB estimates versus their OLS counterparts for a random-intercept model of the Reisby data

Similarly,

$$\begin{aligned}
\sigma^2_{v|y_i} &= \left(\mathbf{1}_i'(\sigma^2 \boldsymbol{I}_{n_i})^{-1}\mathbf{1}_i + \sigma_v^{-2}\right)^{-1} \\
&= \left(n_i\sigma^{-2} + \sigma_v^{-2}\right)^{-1} \\
&= \left(\frac{n_i}{\sigma^2} + \frac{1}{\sigma_v^2}\right)^{-1} \\
&= \left(\frac{n_i\sigma_v^2 + \sigma^2}{\sigma^2\,\sigma_v^2}\right)^{-1} \\
&= \sigma_v^2\left(\frac{\sigma^2}{n_i\sigma_v^2 + \sigma^2}\right)
\end{aligned}$$

$$= \sigma_\upsilon^2 \left(\frac{\sigma^2/n_i}{\sigma_\upsilon^2 + \sigma^2/n_i} \right)$$

$$= \sigma_\upsilon^2 \left(\frac{\sigma_\upsilon^2 + \sigma^2/n_i}{\sigma_\upsilon^2 + \sigma^2/n_i} - \frac{\sigma_\upsilon^2}{\sigma_\upsilon^2 + \sigma^2/n_i} \right)$$

$$= \sigma_\upsilon^2 (1 - \rho_{n_i n_i})$$

where

$$\rho_{n_i n_i} = \frac{\sigma_\upsilon^2}{\sigma_\upsilon^2 + \sigma^2/n_i} = \frac{n_i r}{1 + (n_i - 1)r}$$

EB estimates - multiple random effects

$$\tilde{\boldsymbol{\upsilon}}_i = \left[\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i + \boldsymbol{\Sigma}_{\upsilon}^{-1}\right]^{-1}\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

note, using the property $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$

$$\left[\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i + \boldsymbol{\Sigma}_{\upsilon}^{-1}\right]^{-1}$$

$$= \left\{\left[\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right]\left[\boldsymbol{I}_r + \left(\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right)^{-1}\boldsymbol{\Sigma}_{\upsilon}^{-1}\right]\right\}^{-1}$$

$$= \left\{\left[\boldsymbol{\Sigma}_{\upsilon} + \left(\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right)^{-1}\right]\boldsymbol{\Sigma}_{\upsilon}^{-1}\right\}^{-1}\left[\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right]^{-1}$$

$$= \boldsymbol{\Sigma}_{\upsilon}\left[\boldsymbol{\Sigma}_{\upsilon} + \left(\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right)^{-1}\right]^{-1}\left[\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right]^{-1}$$

$$= \boldsymbol{R}\left[\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right]^{-1}$$

where $\boldsymbol{R} = \boldsymbol{\Sigma}_{\upsilon}\left[\boldsymbol{\Sigma}_{\upsilon} + \left(\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right)^{-1}\right]^{-1}$ is the "multivariate analogue of reliability" (Bock, 1966). Thus,

$$\begin{aligned}
\tilde{\boldsymbol{\upsilon}}_i &= \boldsymbol{R}\left[\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right]^{-1}\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\
&= \boldsymbol{R}\left[\boldsymbol{Z}_i'\boldsymbol{Z}_i\right]^{-1}\boldsymbol{Z}_i'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\
&= \boldsymbol{R}\hat{\boldsymbol{\upsilon}}_i \quad \text{where } \hat{\boldsymbol{\upsilon}}_i = \text{ OLS estimator}
\end{aligned}$$

Similarly,

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\upsilon|y_i} &= (\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i + \boldsymbol{\Sigma}_{\upsilon}^{-1})^{-1} \\
&= \boldsymbol{R}\left[\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right]^{-1} \\
&= \boldsymbol{R}\left\{\left[\boldsymbol{\Sigma}_{\upsilon} + \left(\boldsymbol{Z}_i'(\sigma^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\right)^{-1}\right]\boldsymbol{\Sigma}_{\upsilon}^{-1}\boldsymbol{\Sigma}_{\upsilon} - \boldsymbol{\Sigma}_{\upsilon}\right\} \\
&= \boldsymbol{R}\left(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_{\upsilon} - \boldsymbol{\Sigma}_{\upsilon}\right) \\
&= (\boldsymbol{I} - \boldsymbol{R})\boldsymbol{\Sigma}_{\upsilon}
\end{aligned}
$$

Properties of EB estimates

- $\tilde{\boldsymbol{\upsilon}}_i$ "shrunken" towards assumed mean of υ_i (0)

- degree of "shrunkenness" depends on $\rho_{n_i n_i}$

- number of obs. from individual i, n_i, and intra-unit correlation r influence $\tilde{\boldsymbol{\upsilon}}_i$
  - as $n_i \uparrow$ and/or $r \uparrow$, then $\rho_{n_i n_i} \rightarrow 1$ and $\Rightarrow \tilde{\boldsymbol{\upsilon}}_i$ approaches average deviation of the individual (LS estimate)
  - as $n_i \downarrow$ and/or $r \downarrow$, then $\rho_{n_i n_i} \rightarrow 0$ and $\Rightarrow \tilde{\boldsymbol{\upsilon}}_i$ approaches assumed mean of υ_i, namely 0

- n_i and r have similar effect on the uncertainty about the random effect, $\sigma^2_{\upsilon|y_i}$

As a result,

- calculated variance of $\tilde{\boldsymbol{\upsilon}}_i$ estimates will be less than
  - the calculated variance based on the average individual deviations
  - the estimated population variance $\hat{\sigma}_{\upsilon}^2$

Maximum (marginal) Likelihood Solution

The ML solution uses maximum likelihood estimation in the distribution obtained by integrating over the distribution of $\boldsymbol{\upsilon}$, that is, the marginal distribution:

$$h(\boldsymbol{y}_i) = \int_{\boldsymbol{\upsilon}} f(\boldsymbol{y}_i \mid \boldsymbol{\upsilon}; \boldsymbol{\beta}, \sigma^2)\, g(\boldsymbol{\upsilon}; \boldsymbol{\Sigma}_\upsilon)\, d\boldsymbol{\upsilon}$$

where

$$g(\boldsymbol{\upsilon}; \boldsymbol{\Sigma}_\upsilon) = (2\pi)^{-r/2} \left|\boldsymbol{\Sigma}_\upsilon\right|^{-1/2} \exp\left[-1/2\, \boldsymbol{\upsilon}'\, \boldsymbol{\Sigma}_\upsilon^{-1}\, \boldsymbol{\upsilon}\right]$$

$$\begin{aligned} f(\boldsymbol{y}_i \mid \boldsymbol{\upsilon}; \boldsymbol{\beta}, \sigma^2) = {} & (2\pi)^{-n_i/2} \left|\sigma^2 \boldsymbol{I}_{n_i}\right|^{-1/2} \exp[-1/2 \\ & (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{\upsilon})'(\sigma^2 \boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{\upsilon})] \end{aligned}$$

Need to sum marginal log-likelihoods over the sample:

$$\log L = \sum_{i=1}^{N} \log h(\boldsymbol{y}_i)$$

and find out when $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\Sigma}_{\upsilon}$ maximize this function

For this, denote the posterior density p_i, likelihood function f_i, prior density g, and marginal log-likelihood h_i

$$\begin{aligned} \log L &= \sum_{i=1}^{N} \log h_i \\ &= \sum_{i=1}^{N} \log \left[\int_{\boldsymbol{\upsilon}} f_i \cdot g \, d\boldsymbol{\upsilon} \right] \end{aligned}$$

and $p_i = f_i \cdot g / h_i$

For the coefficient vector $\boldsymbol{\beta}$ of the fixed covariates we obtain:

$$
\begin{aligned}
\frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{N} \frac{\partial \log h_i}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{N} \frac{1}{h_i} \frac{\partial \left[\int_{\boldsymbol{\upsilon}} f_i \cdot g \, d\boldsymbol{\upsilon}\right]}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{N} \frac{1}{h_i} \int_{\boldsymbol{\upsilon}} \frac{\partial f_i}{\partial \boldsymbol{\beta}} \, g \, d\boldsymbol{\upsilon} \\
&= \sum_{i=1}^{N} \int_{\boldsymbol{\upsilon}} \frac{f_i \cdot g}{h_i} \frac{\partial \log f_i}{\partial \boldsymbol{\beta}} \, d\boldsymbol{\upsilon} \qquad (\text{since } f_i = \exp \log f_i) \\
&= \sum_{i=1}^{N} \int_{\boldsymbol{\upsilon}} p_i \, \boldsymbol{X}_i' \left(\sigma^2 \boldsymbol{I}_{n_i}\right)^{-1} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{Z}_i \boldsymbol{\upsilon}\right) d\boldsymbol{\upsilon} \\
&= \sigma^{-2} \sum_{i=1}^{N} \boldsymbol{X}_i' \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{Z}_i \tilde{\boldsymbol{\upsilon}}_i\right) \quad (\text{since } \tilde{\boldsymbol{\upsilon}}_i = \int_{\boldsymbol{\upsilon}} \boldsymbol{\upsilon} \, p_i \, d\boldsymbol{\upsilon})
\end{aligned}
$$

Equating to zero yields,

$$\sum_{i=1}^{N} \boldsymbol{X}_i' \, \boldsymbol{X}_i \boldsymbol{\beta} \; = \; \sum_{i=1}^{N} \boldsymbol{X}_i' \, (\boldsymbol{y}_i - \boldsymbol{Z}_i \tilde{\boldsymbol{\upsilon}}_i)$$

Thus,

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{N} \boldsymbol{X}_i' \, \boldsymbol{X}_i \right]^{-1} \left[\sum_{i=1}^{N} \boldsymbol{X}_i' \, (\boldsymbol{y}_i - \boldsymbol{Z}_i \tilde{\boldsymbol{\upsilon}}_i) \right]$$

as $\tilde{\boldsymbol{\upsilon}}_i \rightarrow 0$, then $\hat{\boldsymbol{\beta}} \rightarrow$ OLS estimates ignoring subject effects

For the residual variance, with $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\tilde{\boldsymbol{\upsilon}}_i$

$$\frac{\partial \log L}{\partial \sigma^2} = \sum_{i=1}^{N} \frac{1}{h_i} \int_{\boldsymbol{\upsilon}} \frac{\partial f_i}{\partial \sigma^2}\, g\, d\boldsymbol{\upsilon}$$

$$= \sum_{i=1}^{N} \int_{\boldsymbol{\upsilon}} \frac{f_i \cdot g}{h_i} \frac{\partial \log f_i}{\partial \sigma^2}\, d\boldsymbol{\upsilon}$$

$$= \sum_{i=1}^{N} \int_{\boldsymbol{\upsilon}} p_i \left[-\frac{n_i}{2}\sigma^{-2} + \frac{1}{2}\sigma^{-4}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{\upsilon})' \times (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{\upsilon}) \right] d\boldsymbol{\upsilon}$$

$$= \frac{1}{2}\sigma^{-4} \sum_{i=1}^{N} \left(-n_i\sigma^2 + \boldsymbol{e}_i'\boldsymbol{e}_i + \operatorname{tr}\left[\boldsymbol{Z}_i \boldsymbol{\Sigma}_{\upsilon|y_i} \boldsymbol{Z}_i' \right] \right)$$

The mysterious last step is because of:

$$\int_{\boldsymbol{v}} p_i \, (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{v})'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{v}) d\boldsymbol{v}$$

$$= \int_{\boldsymbol{v}} p_i \, [(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\tilde{\boldsymbol{v}}_i) - \boldsymbol{Z}_i(\boldsymbol{v} - \tilde{\boldsymbol{v}}_i)]' \times [(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\tilde{\boldsymbol{v}}_i) - \boldsymbol{Z}_i(\boldsymbol{v} - \tilde{\boldsymbol{v}}_i)] \, d\boldsymbol{v}$$

$$= (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\tilde{\boldsymbol{v}}_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\tilde{\boldsymbol{v}}_i) + \mathrm{tr}\boldsymbol{Z}_i\boldsymbol{\Sigma}_{v|y_i}\boldsymbol{Z}_i'$$

using

$$\boldsymbol{\Sigma}_{v|y_i} = \int_{\boldsymbol{v}} (\boldsymbol{v} - \tilde{\boldsymbol{v}}_i)(\boldsymbol{v} - \tilde{\boldsymbol{v}}_i)' \, p_i \, d\boldsymbol{v}$$

Equating to zero yields (with $\hat{\boldsymbol{e}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{Z}_i\tilde{\boldsymbol{\upsilon}}_i$),

$$\hat{\sigma}^2 = \left(\sum_i^N n_i\right)^{-1} \sum_i^N \hat{\boldsymbol{e}}_i'\hat{\boldsymbol{e}}_i \; + \; \text{tr}\left[\boldsymbol{Z}_i\boldsymbol{\Sigma}_{\upsilon|y_i}\boldsymbol{Z}_i'\right]$$

Notice similarity of above to ML estimator in multiple regression model:

$$\hat{\sigma}^2 = \left(\sum_i^N n_i\right)^{-1} \sum_i^N \hat{\boldsymbol{e}}_i'\hat{\boldsymbol{e}}_i$$

as $\tilde{\boldsymbol{\upsilon}}_i \to 0$, then $\boldsymbol{\Sigma}_{\upsilon|y_i} \to 0$, then $\hat{\sigma}^2$ goes to ML error estimate ignoring subjects (which is approximately equal to OLS error estimate, *i.e.*, MSE)

Now things get really ugly!

For the solution for $\boldsymbol{\Sigma}_v$ we will use the "vec" and "vech" notation (Henderson & Searle, 1979) and

$$\frac{\partial \log |\boldsymbol{\Sigma}_v|}{\partial \text{vec}\boldsymbol{\Sigma}_v} = \text{vec}\boldsymbol{\Sigma}_v^{-1}$$

$$\begin{aligned}\frac{\partial\, \boldsymbol{v}'\boldsymbol{\Sigma}_v^{-1}\boldsymbol{v}}{\partial\, \text{vec}\boldsymbol{\Sigma}_v} &= \frac{\partial\, \text{tr}\boldsymbol{\Sigma}_v^{-1}\boldsymbol{v}\boldsymbol{v}'}{\partial\, \text{vec}\boldsymbol{\Sigma}_v} \\ &= -\text{vec}\, \boldsymbol{\Sigma}_v^{-1}\boldsymbol{v}\boldsymbol{v}'\boldsymbol{\Sigma}_v^{-1}\end{aligned}$$

the "vec" of an $n \times n$ matrix $\boldsymbol{A}$ is the column vector $[\boldsymbol{a}_1', \boldsymbol{a}_2', \ldots, \boldsymbol{a}_n']'$, or in words, the "vec" operator stacks the column vectors of a matrix one on top of each other to form one large column vector

the "vech" of an $n \times n$ symmetric matrix $\boldsymbol{A}$ is the $n(n+1)/2$ column vector $[a_{11}, a_{21}, a_{22}, a_{31}, a_{32}, a_{33}, \ldots, a_{nn}]'$, or in words, the "vech" operator stacks the on or above diagonal elements of a square matrix on top of each other to form one column vector

We also need:

$$\begin{aligned} \int_{\boldsymbol{\upsilon}} p_i \, \boldsymbol{\upsilon}\boldsymbol{\upsilon}' d\boldsymbol{\upsilon} &= \int_{\boldsymbol{\upsilon}} p_i \left[(\boldsymbol{\upsilon} - \tilde{\boldsymbol{\upsilon}}_i) + \tilde{\boldsymbol{\upsilon}}_i\right] \left[(\boldsymbol{\upsilon} - \tilde{\boldsymbol{\upsilon}}_i) + \tilde{\boldsymbol{\upsilon}}_i\right]' d\boldsymbol{\upsilon} \\ &= \boldsymbol{\Sigma}_{\upsilon|y_i} + \tilde{\boldsymbol{\upsilon}}_i \tilde{\boldsymbol{\upsilon}}_i{}' \end{aligned}$$

Derivatives with respect to the $r(r+1)/2$ unique elements of $\boldsymbol{\Sigma}_{\upsilon}$:

$$\begin{aligned}
\frac{\partial \log L}{\partial \mathrm{vech}\boldsymbol{\Sigma}_{\upsilon}} &= \sum_{i=1}^{N} \frac{1}{h_i} \int_{\boldsymbol{\upsilon}} \frac{\partial g}{\partial \mathrm{vech}\boldsymbol{\Sigma}_{\upsilon}} f_i \, d\boldsymbol{\upsilon} \\
&= \sum_{i=1}^{N} \int_{\boldsymbol{\upsilon}} \frac{f_i \cdot g}{h_i} \frac{\partial \log g}{\partial \mathrm{vech}\boldsymbol{\Sigma}_{\upsilon}} \, d\boldsymbol{\upsilon} \\
&= \mathsf{G}' \mathrm{vec} \sum_{i}^{N} \int_{\boldsymbol{\upsilon}} p_i \left[-\frac{1}{2}\boldsymbol{\Sigma}_{\upsilon}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_{\upsilon}^{-1}\boldsymbol{\upsilon}\boldsymbol{\upsilon}'\boldsymbol{\Sigma}_{\upsilon}^{-1} \right] \\
&= \frac{1}{2}\mathsf{G}' \sum_{i}^{N} \mathrm{vec}\, \boldsymbol{\Sigma}_{\upsilon}^{-1} \left(-\boldsymbol{\Sigma}_{\upsilon} + \boldsymbol{\Sigma}_{\upsilon|y_i} + \tilde{\boldsymbol{\upsilon}}_i \tilde{\boldsymbol{\upsilon}}_i{}' \right) \boldsymbol{\Sigma}_{\upsilon}^{-1}
\end{aligned}$$

where the matrix G transforms the vech of a square matrix into the vec (McCulloch, 1982), that is, $\mathrm{vec}\boldsymbol{A} = \mathsf{G}\ \mathrm{vech}\boldsymbol{A}$ for a square matrix $\boldsymbol{A}$

setting this monster equal to zero, we find

$$\hat{\boldsymbol{\Sigma}}_{\upsilon} = \frac{1}{N} \sum_{i}^{N} \left(\tilde{\boldsymbol{\upsilon}}_i \tilde{\boldsymbol{\upsilon}}_i{}' + \boldsymbol{\Sigma}_{\upsilon|y_i} \right)$$

e.g., Subject variance for random intercepts model

$$\hat{\sigma}^2_{\upsilon} = \frac{1}{N} \sum_{i}^{N} \tilde{\upsilon}_i{}^2 + \sigma^2_{\upsilon|y_i}$$

- Average of EB individual effect squared ($\approx$ sample estimate of variance of EB effect) plus the EB estimate of the uncertainty about the individual effect
- $\sigma^2_{\upsilon}, \tilde{\upsilon}_i, \sigma^2_{\upsilon|y_i}$ are one happy family (they move together!)
  - as $\tilde{\upsilon}_i$ and $\sigma^2_{\upsilon|y_i}$ go to 0, so does σ^2_{υ}

– as covariates included in model $\Rightarrow \tilde{v}_i$ gets smaller, thus so does σ_v^2

- $\hat{\sigma}_v^2$ is larger than the calculated variance of $\tilde{v}_i$
  – since calculated variance equals

$$\sum_i^N (\tilde{v}_i - \bar{x}_{\tilde{v}_i})^2/(N-1)$$

  and $\bar{x}_{\tilde{v}_i} \approx 0$

  note: this difference is diminished as all n_i get large, then $\sigma^2_{v|y_i} \longrightarrow 0$ and $\hat{\sigma}_v^2$ is approximately equal to the calculated variance of $\tilde{v}_i$

Within-subjects and between-subjects components

random intercepts model

Within-subjects model - level 1

$$y_{ij} = b_{0i} + \boldsymbol{\beta}'_{(1)}\boldsymbol{x}_{(1)\,ij} + \varepsilon_{ij}$$

i = $1 \ldots N$ subjects
j = $1 \ldots n_i$ observations for subject i

Between-subjects models - level 2

$$b_{0i} = \beta_0 + \boldsymbol{\beta}'_{(2)}\boldsymbol{x}_{(2)\,i} + \upsilon_i$$

Total model

$$y_{ij} = \boldsymbol{\beta}' \boldsymbol{x}_{ij} + \upsilon_i + \varepsilon_{ij}$$

where

$$\boldsymbol{x}'_{ij} = \begin{bmatrix} \mathbf{1} & \vdots & \boldsymbol{x}_{(1)\,ij} & \vdots & \boldsymbol{x}_{(2)\,i} \end{bmatrix}$$

and

$$\boldsymbol{\beta}' = \begin{bmatrix} \beta_0 & \vdots & \boldsymbol{\beta}'_{(1)} & \vdots & \boldsymbol{\beta}'_{(2)} \end{bmatrix}$$

$$\varepsilon_{ij} \sim \mathcal{NID}(0, \sigma^2) \qquad \upsilon_i \sim \mathcal{NID}(0, \sigma^2_\upsilon)$$

Adding in $\boldsymbol{x}_{(1)\,ij}$ can reduce σ^2 and σ^2_υ

Adding in $\boldsymbol{x}_{(2)\,i}$ can reduce σ^2_υ

Estimation via EM algorithm: opposite process of "I do cocaine so I can work more, so I can do more cocaine, so I can work more, *etc.*, "

Cocaine		*Work*	*Health*
do cocaine	$\rightarrow$	work more	declines
do more cocaine	$\rightarrow$	work more	declines more
do even more cocaine	$\rightarrow$	work even more	declines even more
...	...	...	...
do a ton of cocaine	$\rightarrow$	always working	death

M-Step (ML)		*E-Step (EB)*	*Estimation*
starting values $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}_\upsilon$	$\rightarrow$	estimate $\tilde{\boldsymbol{\upsilon}}_i$ $\boldsymbol{\Sigma}_{\upsilon\|y_i}$	improves
re-estimate $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}_\upsilon$	$\rightarrow$	re-estimate $\tilde{\boldsymbol{\upsilon}}_i$ $\boldsymbol{\Sigma}_{\upsilon\|y_i}$	improves more
re-re-estimate $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}_\upsilon$	$\rightarrow$	re-re-estimate $\tilde{\boldsymbol{\upsilon}}_i$ $\boldsymbol{\Sigma}_{\upsilon\|y_i}$	improves even more
...	...	...	...
RE-estimate $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}_\upsilon$	$\rightarrow$	RE-estimate $\tilde{\boldsymbol{\upsilon}}_i$ $\boldsymbol{\Sigma}_{\upsilon\|y_i}$	convergence

$\Rightarrow$ EM is better than cocaine since EM leads to convergence and not death

EM solution - shifts between

- estimation of individual parameters $\tilde{\boldsymbol{\upsilon}}_i$ and $\boldsymbol{\Sigma}_{\upsilon|y_i}$
- estimation of the structural $\boldsymbol{\beta}$ and σ^2 and population parameters $\boldsymbol{\Sigma}_{\upsilon}$

- convergence by EM can be slow and torturous
- need solution which utilizes second derivatives
  - Newton-Raphson procedure
  - Fisher-scoring procedure
    - replaces second derivatives with their expectations (negative of the information matrix)
    - multiplying the vector of first derivatives by the inverse of the information matrix provides the corrections at each iteration

For a parameter vector $\boldsymbol{\Theta}$, the estimated values for iteration $\iota + 1$ are obtained by

Newton-Raphson

$$\boldsymbol{\Theta}_{\iota+1} = \boldsymbol{\Theta}_{\iota} - \left[\frac{\partial^2 \log L}{\partial \boldsymbol{\Theta}_{\iota}\, \partial \boldsymbol{\Theta}_{\iota}'}\right]^{-1} \frac{\partial \log L}{\partial \boldsymbol{\Theta}_{\iota}}$$

Fisher scoring

$$\boldsymbol{\Theta}_{\iota+1} = \boldsymbol{\Theta}_{\iota} - \mathcal{E}\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\Theta}_{\iota}\, \partial \boldsymbol{\Theta}_{\iota}'}\right]^{-1} \frac{\partial \log L}{\partial \boldsymbol{\Theta}_{\iota}}$$

where the information matrix has the form:

$$I \;=\; -\mathcal{E}\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\Theta}_{\iota}\, \partial \boldsymbol{\Theta}_{\iota}'}\right] \;=\; \begin{bmatrix} I(\boldsymbol{\beta}) & & \\ 0 & I(\boldsymbol{\Sigma}_{\upsilon}) & \\ 0 & I(\sigma^2, \boldsymbol{\Sigma}_{\upsilon}) & I(\sigma^2) \end{bmatrix}$$

The information matrix - random intercepts model

$$I = \begin{bmatrix} I(\boldsymbol{\beta}) & 0 & 0 \\ 0 & I(\sigma_v^2) & I(\sigma^2, \sigma_v^2) \\ 0 & I(\sigma^2, \sigma_v^2) & I(\sigma^2) \end{bmatrix}$$

where

$$I(\boldsymbol{\beta}) = \sigma^{-4} \sum_{i=1}^{N} \boldsymbol{X}_i'(\sigma^2 \boldsymbol{I}_{n_i} - \sigma_{v|y_i}^2 \mathbf{1}_i \mathbf{1}_i') \boldsymbol{X}_i$$

$$I(\sigma_v^2) = \frac{1}{2}\sigma_v^{-8} \sum_{i=1}^{N} (\sigma_v^2 - \sigma_{v|y_i}^2)^2$$

$$I(\sigma^2) = \frac{1}{2}\sigma^{-8} \sum_{i=1}^{N} n_i (\sigma^2 - \sigma_{v|y_i}^2)^2$$

$$I(\sigma^2, \sigma_v^2) = \frac{1}{2}\sigma^{-4}\sigma_v^{-4} \sum_{i=1}^{N} n_i \sigma_{v|y_i}^4$$

- At convergence, I^{-1} provides asymptotic variances and covariances of ML estimates
- For standard errors, take the square root of the diagonal elements of I^{-1}
- single parameter hypothesis test $H_0 : \theta = 0$

$$z = \frac{\hat{\theta}}{\hat{se}(\hat{\theta})}$$

Reparameterization of the Variance terms

- Fisher scoring estimation of $\boldsymbol{\Sigma}_v$ is problematic when diagonal elements get small (near-zero variances)
- reparameterize $\boldsymbol{\Sigma}_v$ in terms of the Gaussian factorization of a symmetric matrix
- use exponential transformation for diagonal matrix $\boldsymbol{D}$ corresponding to variance parameters

Specifically, consider $\boldsymbol{\Sigma}_v = \boldsymbol{LDL'} = \boldsymbol{L}\ \exp(\boldsymbol{\Pi})\ \boldsymbol{L'}$

e.g., , if $r = 3$,

$$\boldsymbol{\Sigma}_v = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} \mathrm{e}^{\pi_1} & 0 & 0 \\ 0 & \mathrm{e}^{\pi_2} & 0 \\ 0 & 0 & \mathrm{e}^{\pi_3} \end{bmatrix} \begin{bmatrix} 1 & l_{21} & l_{31} \\ 0 & 1 & l_{32} \\ 0 & 0 & 1 \end{bmatrix}$$

$\Rightarrow$ estimate $\boldsymbol{L}$ and $\boldsymbol{\Pi}$

Derivatives - Magnus (1988); Magnus & Neudecker (1988)

diagonal elements in $\mathbf{\Pi}$

w($\boldsymbol{A}$) denotes vector which contains only the diagonal elements of a square matrix $\boldsymbol{A}$

$$\frac{\partial \log L}{\partial(\mathrm{w}(\mathbf{\Pi}))} =$$

$$\frac{1}{2}\boldsymbol{D}^{-1}\mathsf{H}_r(\boldsymbol{L}^{-1} \otimes \boldsymbol{L}^{-1}) \left\{ \left(\sum_i^N \mathrm{vec}\left[\mathbf{\Sigma}_{\upsilon|y_i} + \tilde{\boldsymbol{\upsilon}}_i \tilde{\boldsymbol{\upsilon}}_i{}' \right] \right) - N \ \mathrm{vec}\, \mathbf{\Sigma}_{\upsilon} \right\}$$

where $\otimes$ denotes the Kronecker product, and where the $r \times r^2$ matrix H_r eliminates the off-diagonal elements from the vec of a $r \times r$ matrix

elements of $\boldsymbol{L}$ *below the main diagonal*

$\tilde{\mathrm{v}}(\boldsymbol{A})$ denotes vector which contains only the elements below the main diagonal of a square matrix $\boldsymbol{A}$

$$\frac{\partial \log L}{\partial \left(\tilde{\mathrm{v}}(\boldsymbol{L})\right)} = \tilde{\mathsf{J}}_r(\boldsymbol{L}^{-1} \otimes \boldsymbol{\Sigma}_{\upsilon}^{-1}) \left(\sum_{i}^{N} \mathrm{vec} \left[\boldsymbol{\Sigma}_{\upsilon|y_i} + \tilde{\boldsymbol{\upsilon}}_i \tilde{\boldsymbol{\upsilon}}_i{}' \right] \right)$$

where the $r \times r^2$ matrix $\tilde{\mathsf{J}}_r$ eliminates the elements of the vec which are on or above the main diagonal

Restricted Maximum Likelihood (REML)

- introduced by Patterson & Thompson (1971) *Biometrika*
- object: to remove bias in ML estimates of variance terms
- *e.g.*, in linear regression

$$\hat{\sigma}^2_{ML} = SSE/N \qquad \hat{\sigma}^2_{REML} = SSE/(N-p)$$

  $\Rightarrow$ important as N is small and/or p is large
  (here, p includes intercept)

- idea: in ML estimation of variance terms treats $\boldsymbol{\beta}$ as known, however $\boldsymbol{\beta}$ are estimated
- difference occurs as p gets relatively large $\Rightarrow$ MLEs of variances and their standard errors are downward biased
- for fixed p, methods asymptotically equivalent as either or both of N and n_i tend to infinity

- relatively simple modification of ML algorithm (DLZ p. 64-68)
- REML estimators are preferable, however
- **the REML deviance ($-2\log L$) can be used only for comparison of models with identical regression parts**
- in SAS, use METHOD=REML (the default)