Weighted Linear Mixed-Effects Models

Developed by Paul Bailey*†

February 14, 2018

Introduction {introduction}

Obtaining population estimates for official statistics or doing social science research with hopes of a direct appeal to external validity requires that a population be sampled. That usually means multi-stage probability sampling, with groups drawn, potentially groups in groups, and then individuals. These samples can be analyzed with weighted linear models and variance estimators applied Wolter (year). Another common approach to nested models is to use hierarchical linear models, or a subset of mixed models where the units are entirely nested in groups.

While any likelihood that can be written down can be maximized, efficent calculation is always appreciated by analyists. In particular, the weighted hierarchical linear model has nested integrals that are difficult to compute efficently. For unweighted data, Bates and Pinheiro (1998) show a method of calculation that leaves only a probability density function in the integral, which can be easily integrated as the multiplicative identity, obviating the need for numerical integration.

This paper generalizes the method in Bates and Pinheiro (1998) and the `lme4` package (Bates, Maechler, Bolker, and Walker, 2015) to the weighted case. I am unaware of a description of this and will detail the information available in existing statistical software documentation which tends to simply cite Rabe-Heskety and not provide additional detail on how weighted estimates are formed.

This paper develops the method used to more quickly estimate linear models. These are implemented in the `WeMix` R package. After the derivation, a few example growth curve models are fit with `WeMix`.

Model {model}

The model of interest is a linear mixed model where units are nested inside groups, which may themselves be nested in groups, a model specified by Rabe-Hesketh and Skrondal (2006) and Rabe-Hesketh, Skrondal, and Pickles (2002) for the GLLAMM software. The advantage of fitting the model in nested levels is that survey sampling weights can be incorporated; the pseudo-likelihood,[1] at the top level, is an integral that sums across the top-level groups, weighting each group or unit (in a two-level model) according to sampling weights.

At the highest level, the likelihood ($\mathcal{L}$) is a function of the parameters for the random-effect covariance matrix ($\boldsymbol{\theta}$), the fixed-effect estimates ($\boldsymbol{\beta}$), and the outcomes $\boldsymbol{y}$. This equation relates the overall likelihood $[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{y})]$ to the integral of the likelihoods of the integrals of the top-level units $[\mathcal{L}_j^{(L)}(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{y}|\boldsymbol{u}^{(L)})$, indexed by $j]$; the top-level unit likelihoods have a superscript (L) to indicate they are for the Lth (top) level. Here we omit reference to the fixed-effect design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times k_x}$ and the random-effect design matrix $\boldsymbol{Z} \in \mathbb{R}^{n \times k_z}$, but

*This publication was prepared for NCES (National Center for Education Statistics) under Contract No. ED-IES-12-D-0002 with American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. government.

†The authors would like to thank Mike Cohen for reviewing this document.

[1]For inverse probability of selection survey data, the likelihood is an estimated population likelihood and not an observed likelihood; the typical phrase for this is *pseudo-likelihood*, which applies to every likelihood in this document.

the likelihood is conditional on those terms as well.

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{y}) = \int g^{(L)}(\boldsymbol{u}^{(L)}; \boldsymbol{\theta}^{(L)}) \prod_j \left[\mathcal{L}_j^{(L)}(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{y} | \boldsymbol{u}^{(L)}) \right]^{w_j^{(L)}} d\boldsymbol{u}^{(L)} \tag{1}$$

where the $\boldsymbol{u}$ terms are the random effects, which are marginalized by integrating over them;[2] the g function plays a role similar to a prior, but has its variance (or covariance matrix) fit using covariance parameter vector $\boldsymbol{\theta}$, while j represents the indexes of all the top-level units, which have their likelihood raised to the power of their weight $w_j^{(L)}$. Because $\boldsymbol{u}$ may be a vector—for example, if there is a random slope and intercept—the covariance matrix between the $\boldsymbol{u}$ terms ($\boldsymbol{\Sigma}^{(L)}$) may be diagonal (no covariance) or dense. In any case, the covariance matrix is parameterized by a vector of values $\boldsymbol{\theta}$; at the Lth level, the relevant elements are denoted by $\boldsymbol{\theta}^{(L)}$.

The conditional likelihood at each level from L to two—those above the first, or lowest level—is then recursively defined, for the jth unit, at level l ($\mathcal{L}_j^{(l)}$; Rabe-Hesketh et al., 2002, eq. 3) as:

$$\mathcal{L}_j^{(l)}(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{y} | \boldsymbol{U}^{(l+1)}) = \int g^{(l)}(\boldsymbol{u}^{(l)}; \boldsymbol{\theta}^{(l)}) \prod_{j'} \left[\mathcal{L}_{j'}^{(l-1)}(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{y} | \boldsymbol{U}^{(l)}) \right]^{\boldsymbol{w}_{j'}^{(l-1)}} d\boldsymbol{u}^{(l)} \quad l = 2, \ldots, L-1 \tag{2}$$

where the subscript j' that the product is indexed over indicates that the likelihood $\mathcal{L}_{j'}^{(l-1)}(\cdot)$ is for the units of level $l-1$ nested in unit j, and $\boldsymbol{U}^{(l)}$ is all of the random effects at or above level l, so that $\boldsymbol{U}^{(l)}$ includes $\left\{\boldsymbol{u}^{(l)}, \boldsymbol{u}^{(l+1)}, \ldots, \boldsymbol{u}^{(L)}\right\}$. This equation reflects the nested nature of the data; a group (e.g., school), annotated as j, has an individual likelihood that is the product of the likelihoods of the units or groups (e.g., students or classrooms, annotated as j') nested inside of it.

In the case of a Gaussian residual, the likelihood ($\mathcal{L}^{(1)}$) at the individual unit level is given by the equation

$$\mathcal{L}_i^{(1)}(\boldsymbol{\theta}, \boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{U}^{(2)}) = \frac{1}{\sigma} \phi \left(\frac{\hat{e}_i^{(1)}}{\sigma} \right) \tag{3}$$

where the subscript i is used to indicate that this is the likelihood of the i^{th} individual, the superscript (1) on $\hat{\boldsymbol{e}}_i^{(1)}$ is used to indicate that it is an unpenalized residual—where the penalized residual will be introduced in the next section—$\phi(\cdot)$ is the standard normal density function and σ is the residual variance (a scalar), and the residuals vector $\hat{\boldsymbol{e}}^{(1)}$ represents the residuals $\hat{\boldsymbol{e}}_i^{(1)}$ for each individual and is given, in vector form, by

$$\hat{\boldsymbol{e}}^{(1)} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \sum_{l=2}^{L} \boldsymbol{Z}^{(l)} \hat{\boldsymbol{u}}^{(l)} \tag{4}$$

where $\boldsymbol{Z}^{(l)} \in \mathbb{R}^{n \times k_z^{(l)}}$ are block matrixes that form $\boldsymbol{Z}$

$$\boldsymbol{Z} = \left[\boldsymbol{Z}^{(2)}, ..., \boldsymbol{Z}^{(L)} \right] \tag{5}$$

When solved with the above integrals (as in Rabe-Hesketh et al., 2002), σ is fit as a parameter and there is no direct equation for it.

[2]The $\boldsymbol{u}$ terms are integrated out and so they are not, strictly speaking, estimated. Because of this, we do not call them empirical Bayes estimates. However, the 'WeMix' package does provide a value for them and, because the variance of the distribution that acts as a prior is estimated, these values could reasonably be regarded as an empirical Bayes estimate of the random effects.

Meaning of weights {meaning-of-w

The weights used in these models are typically not the inverse selection probabilities. This would fit the random effect variances as if the data had been imputed to a sample the size of the population.

Instead the weights are often adjusted to capture both the unequal selection probabilities and total sample size. These methods are documented in Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash (1998) and Rabe-Hesketh & Skrondal (2006), and other places.

Typically, the difference between weights that are interpretable as sampling and precision weights shows up only in variance estimation. However, I'll document later that the `lme4' and`WeMix`weights result in different estimates. This is because`WeMix‘ uses the formulas in this paper that weight both the units as well as the groups.

Existing literature {existing-lit

The SAS manual is often a source of exhaustive documentation of implementation but this author cannot find any indication that these methods are used by SAS at in the PROC GLIMMIX procedure documentation, "Pseudo-likelihood Estimation for Weighted Multilevel Models."

I would commend the authors of the Stata manuals for routinely producing extensive documentation on methods and formulas of their software. Uncharacteristicly, in the case of weighted HLM they simply state, "Weighted estimation is achieved through incorporating w_j and $w_{i|j}$ into the matrix decomposition methods detailed above to reflect replicated clusters for w_j and replicated observations within clusters for $w_{i|j}$. (Stata, page 534)"[3]

Unweighted case {unweighted-c

This document describes how `WeMix` uses symbolic integration that relies on a mix of both Bates and Pinheiro (1998) and the `lme4` package (Bates, Maechler, Bolker, and Walker, 2015), obviating the need for numerical integration in a weighted, nested model.

The basic model in `lme4` is of the form (Bates et al., 2015, eqs. 2 and 3):

$$(\boldsymbol{y}|\boldsymbol{U}=\boldsymbol{u}) \sim N(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}, \sigma^2 W^{-1}) \quad (6) \quad \{\texttt{eq:lme4A}\}$$

$$\boldsymbol{U} \sim N(0, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \quad (7) \quad \{\texttt{eq:lme4B}\}$$

where $N(\cdot,\cdot)$ is the multivariate normal distribution, and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive semidefinite—allowing, for example, a variance parameter to be exactly zero—that is parameterized by a vector of parameters $\boldsymbol{\theta}$.

The likelihood is maximized by integrating (symbolically) over $\boldsymbol{U}$ (Bates et al., 2015, eqs. 20, 21, 22, and 24):

$$f_{\boldsymbol{y}}(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{\theta}) = \int f_{\boldsymbol{y}|\boldsymbol{U}=\boldsymbol{u}}(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{u}) \cdot f_{\boldsymbol{U}}(\boldsymbol{u}) d\boldsymbol{u} \quad (8) \quad \{\texttt{eq:lme4C}\}$$

where the $f_{\boldsymbol{U}}$ term is analogous to $g(\boldsymbol{u}^{(l)})$ in the Rabe-Hesketh et al. (2006) formulation but is intentionally represented in a non-nested structure to allow for crossed terms.

In this document we show how `WeMix` uses a derivation similar to that in Bates et al. (2015) and Bates and Pinheiro (1998) to fit a sample-weighted mixed model, avoiding the integration necessary in GLLAMM. Comparing `WeMix` to `lmer`, the latter is more general in the sense of allowing crossed terms, while `WeMix` allows for sampling weights; unweighted, the models should agree.

The next section follows Bates et al. (2015) and shows the `lme4` and `WeMix` model without weights—the only case where they are identical. The subsequent section then shows the adaptations to the likelihood for the

[3] See "Methods and Formulas" for "mixed—Multilevel mixed-effects linear regression" in "[ME] Stata Multilevel mixed-effects reference manual: release 17."

sample weighted case. Notably, `lme4` has unit-level weights, but they are precision weights and not level-1 sample weights, so even when only the level-1 weights are nontrivial, the models disagree. The final section describes the application of the profile likelihood (again, following `lme4`) used to estimate linear models in `WeMix`.

Following the logic of Bates et al. (2015), the unweighted (unit weight) case simplifies eqs. 57 and 58 to

$$(\boldsymbol{y}|\boldsymbol{U}=\boldsymbol{u}) \sim T_1 \tag{9}$$ {eq:WeMixAnoW

$$\boldsymbol{U} \sim T_2 * T_3 * \cdots * T_L \tag{10}$$ {eq:WeMixBnoW

where eqs. 59 and 60 are simplified to

$$T_1 \sim N(\boldsymbol{X\beta} + \boldsymbol{Zu}, \sigma^2 I) \tag{11}$$ {eq:WeMixA2no

$$T_l \sim N(0, \boldsymbol{\Sigma_{ll}}) \quad l = 2, \cdots, L \tag{12}$$ {eq:WeMixB2no

the random-effect vector $\boldsymbol{U}$ is rewritten as the product of a square root-covariance matrix $\boldsymbol{\Lambda}$ and an *iid* normal vector $\boldsymbol{\upsilon}$:

$$\boldsymbol{U} = \boldsymbol{\Lambda\upsilon} \tag{13}$$ {eq:Uf}

$$\boldsymbol{\upsilon} \sim N(0, \sigma^2 \boldsymbol{I}) \tag{14}$$

When $\boldsymbol{\Lambda}$ is a *square root matrix* of $\boldsymbol{\Sigma}$, meaning

$$\frac{1}{\sigma^2}\boldsymbol{\Sigma} = \frac{1}{\sigma^2}\boldsymbol{\Lambda}^T\boldsymbol{\Lambda} \tag{15}$$ {eq:root}

it follows that $\boldsymbol{U}$ has the distribution shown in eq. 12, but the equations can use the much easier to work with $\boldsymbol{\upsilon}$. Note that eq. 15 implies

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}_{22} & 0 & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{\Sigma}_{LL} \end{bmatrix} \tag{16}$$

$$= \begin{bmatrix} \boldsymbol{\Lambda}_{11}^T & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Lambda}_{22}^T & 0 & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{\Lambda}_{LL}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_{11} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Lambda}_{22} & 0 & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{\Lambda}_{LL} \end{bmatrix} \tag{17}$$

$$= \boldsymbol{\Lambda}^T\boldsymbol{\Lambda} \tag{18}$$

Note that $\boldsymbol{\Lambda}$ (and thus $\boldsymbol{\Sigma}$) will be parameterized by a set of parameters $\boldsymbol{\theta}$, so eq. 15 could be written

$$\frac{1}{\sigma^2}\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\left(\boldsymbol{\Lambda}(\boldsymbol{\theta})\right)^T\boldsymbol{\Lambda}(\boldsymbol{\theta}) \tag{19}$$

and will be written this way to keep track of the parameters involved. These parameters vary by model. For a two-level model with a random intercept and nothing else, $\boldsymbol{\theta}$ would be a scalar with the relative precision of the random effect. The matrix $\boldsymbol{\Lambda}(\boldsymbol{\theta})$ would then be diagonal and of the form $\boldsymbol{\Lambda}(\boldsymbol{\theta}) = \theta\boldsymbol{I}$, with $\boldsymbol{I}$ being an identity matrix with the same number of rows and columns as there are groups (random effects). For a two-level model with a random slope and intercept, the $\boldsymbol{\theta}$ would have length three and would parameterize the three elements of a covariance matrix. In this special case, the parameterization could be $\boldsymbol{\Lambda}(\boldsymbol{\theta}) = \begin{bmatrix} \theta_1\boldsymbol{I} & \theta_2\boldsymbol{I} \\ 0 & \theta_3\boldsymbol{I} \end{bmatrix}$; a block matrix that allows the slope and intercept term to covary with each other, within a group.

Then, the residual (penalized) sum of squares is (Bates et al., 2015, eqs. 14 and 15)

$$r^2(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\upsilon}) = ||\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z\Lambda(\theta)\upsilon}||^2 + ||\boldsymbol{\upsilon}||^2 \tag{20}$$ {eq:pseudoSS}

where $||\boldsymbol{v}||^2$ is the sum of squares for the vector $\boldsymbol{v}$; as an equation, $||\boldsymbol{v}||^2 = \sum v_i^2$.

Notice that the penalty function ($||\boldsymbol{v}||^2$) and the residual both are assumed to be independent identically distributed normal distributions with variance σ^2; this allows for both the regression and the random effects to be estimated in one regression where the pseudo-data outcome for the random effects is a vector of zeros (Bates et al., 2015, eq. 16). This rewrites r as a sum of squares, adding a vector of zeros below $\boldsymbol{y}$—the *pseudo-data*:

$$r^2(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{v}) = \left|\left| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{X} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\beta} \end{bmatrix} \right|\right|^2 \tag{21}$$ {eq:WeMixR0}

Unlike Bates et al. (2015, eq. 16), we proceed by taking a QR decomposition (Trefethen and Bau, 1997) of

$$\boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{X} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \tag{22}$$ {eq:uwA}

Plugging eq. 22 into eq. 21 and finding the least squares solution to

$$\boldsymbol{A} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \tag{23}$$ {eq:lsqnolsq}

using the QR decomposition on $\boldsymbol{A}$, which rewrites $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ for an orthogonal matrix $\boldsymbol{Q}$ (So, $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$) and an upper triangular matrix $\boldsymbol{R}$,

$$\boldsymbol{Q}\boldsymbol{R} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \tag{24}$$

where $\boldsymbol{R}$ can be written in block form as

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{0} & \boldsymbol{R}_{22} \end{bmatrix} \tag{25}$$ {eq:Rblock}

in which $\boldsymbol{R}_{11}$ is also upper triangular, square, and conformable with $\boldsymbol{v}$, and $\boldsymbol{R}_{22}$ is similarly upper triangular, square, and conformable with $\boldsymbol{\beta}$, while $\boldsymbol{R}_{12}$ is rectangular and (potentially) dense.

Rewriting r^2 as a deviation from the least squares solution, eq. 21 becomes

$$r^2(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{v}) = \left|\left| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \boldsymbol{A} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\beta} \end{bmatrix} \right|\right|^2 \tag{26}$$

$$= \left|\left| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \boldsymbol{A} \begin{bmatrix} \boldsymbol{v} - \hat{\boldsymbol{v}} + \hat{\boldsymbol{v}} \\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}} \end{bmatrix} \right|\right|^2 \tag{27}$$

$$= \left|\left| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \boldsymbol{A} \begin{bmatrix} \hat{\boldsymbol{v}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} - \boldsymbol{A} \begin{bmatrix} \boldsymbol{v} - \hat{\boldsymbol{v}} \\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \end{bmatrix} \right|\right|^2 \tag{28}$$

Where $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{\beta}}$ are the least squares solution to eq. 23. Using the identity that for any vector $\boldsymbol{v}$ the sum of squares is also just the inner product, so $||\boldsymbol{v}||^2 = \boldsymbol{v}^T\boldsymbol{v}$,

$$r^2(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{v}) = \left[\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \boldsymbol{A} \begin{bmatrix} \hat{\boldsymbol{v}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} - \boldsymbol{A} \begin{bmatrix} \boldsymbol{v} - \hat{\boldsymbol{v}} \\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \end{bmatrix} \right]^T \left[\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \boldsymbol{A} \begin{bmatrix} \hat{\boldsymbol{v}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} - \boldsymbol{A} \begin{bmatrix} \boldsymbol{v} - \hat{\boldsymbol{v}} \\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \end{bmatrix} \right] \tag{29}$$ {eq:finalT}

Defining $\hat{\boldsymbol{e}}$ to be the penalized least squares residual,

$$\hat{\boldsymbol{e}} \equiv \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \boldsymbol{A} \begin{bmatrix} \hat{\boldsymbol{v}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} \tag{30}$$

then eq. 29 can be rewritten

$$r^2(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\upsilon}) = \left[\hat{\boldsymbol{e}} - \boldsymbol{A}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}\right]^T\left[\hat{\boldsymbol{e}} - \boldsymbol{A}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}\right] \tag{31}$$

$$= \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} - 2\hat{\boldsymbol{e}}^T\left[\boldsymbol{A}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}\right] + \left[\boldsymbol{A}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}\right]^T\left[\boldsymbol{A}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}\right] \tag{32}$$ {eq:quad}

Since $\hat{\boldsymbol{e}}$, the uniquely minimized residuals to the least squares problem is in the null of $\boldsymbol{A}$, while $\boldsymbol{Ax}$ is in the span of $\boldsymbol{A}$ for any vector $\boldsymbol{x}$, then $\hat{\boldsymbol{e}}^T\boldsymbol{Ax} = 0$ for any $\boldsymbol{x}$. Thus, $\hat{\boldsymbol{e}}^T\left[\boldsymbol{A}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}\right] = \boldsymbol{0}$ and eq. 32 becomes

$$r^2(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\upsilon}) = \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} + \begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}^T\boldsymbol{A}^T\boldsymbol{A}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix} \tag{33}$$

$$= \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} + \begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}^T[\boldsymbol{QR}]^T[\boldsymbol{QR}]\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix} \tag{34}$$

$$= \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} + \begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}^T\boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{QR}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix} \tag{35}$$

Then, because $\boldsymbol{Q}$ is orthonormal, $\boldsymbol{Q}^T = \boldsymbol{Q}^{-1}$ and

$$r^2(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\upsilon}) = \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} + \begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}^T\boldsymbol{R}^T\boldsymbol{R}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix} \tag{36}$$

$$= \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} + \left|\left|\boldsymbol{R}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}\right|\right|^2 \tag{37}$$ {eq:r2F}

Notice that $\hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}}$ is the value of r^2 evaluated at the least squares solution (denoted by adding hats to $\boldsymbol{\beta}$ and $\boldsymbol{\upsilon}$), so that

$$r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) = \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} \tag{38}$$ {eq:r2tau}

Plugging eqs. 38 and 25 into eq. 37 (Bates et al., 2015, eq. 19),

$$r^2(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\upsilon}) = r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) + \left|\left|\begin{bmatrix}\boldsymbol{R}_{11} & \boldsymbol{R}_{12}\\ \boldsymbol{0} & \boldsymbol{R}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}\right|\right|^2 \tag{39}$$

$$= r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) + \left|\left|\boldsymbol{R}_{11}(\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}) + \boldsymbol{R}_{12}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2 + \left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2 \tag{40}$$ {eq:r2sub}

From the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\upsilon}$ (Bates et al., 2015, eqs. 20, 21, and 22),

$$(\boldsymbol{y},\boldsymbol{\upsilon}) \sim N(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Z\upsilon},\sigma^2\boldsymbol{I}_n) * N(\boldsymbol{\upsilon},\sigma^2\boldsymbol{I}_{k_z}) \tag{41}$$

and the probability density function of the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\upsilon}$ is

$$f_{\boldsymbol{y},\boldsymbol{\upsilon}}\left(\boldsymbol{y},\boldsymbol{\upsilon},\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\left[\frac{-(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Z\upsilon})^2}{2\sigma^2}\right]\cdot\frac{1}{\left(2\pi\sigma^2\right)^{\frac{k_z}{2}}}\exp\left[\frac{-||\boldsymbol{\upsilon}||^2}{2\sigma^2}\right] \tag{42}$$

pluggin in eq. 20

$$f_{\boldsymbol{y},\boldsymbol{\upsilon}}\left(\boldsymbol{y},\boldsymbol{\upsilon},\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2\right) = = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\left[\frac{-r^2(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\upsilon})+||\upsilon||^2}{2\sigma^2}\right]\cdot\frac{1}{\left(2\pi\sigma^2\right)^{\frac{k_z}{2}}}\exp\left[\frac{-||\boldsymbol{\upsilon}||^2}{2\sigma^2}\right] \tag{43}$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n+k_z}{2}}}\exp\left[\frac{-r^2(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\upsilon})}{2\sigma^2}\right] \tag{44}$$

This can then be integrated over $\boldsymbol{\upsilon}$ to get the likelihood of the model (Bates et al., 2015, eqs. 25 and 26):

$$\mathcal{L}\left(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2;\boldsymbol{y}\right) = \int f_{\boldsymbol{y},\boldsymbol{\upsilon}}\left(\boldsymbol{y},\boldsymbol{\upsilon},\boldsymbol{\beta},\boldsymbol{\theta}\right)d\boldsymbol{\upsilon} \tag{45}$$

$$= \int \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n+k_z}{2}}}\exp\frac{-r^2(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\upsilon})}{2\sigma^2}d\boldsymbol{\upsilon} \tag{46}$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n+k_z}{2}}}\exp\frac{-r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) - \left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}\int\exp\frac{-\left|\left|\boldsymbol{R}_{11}(\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}})+\boldsymbol{R}_{12}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}d\boldsymbol{\upsilon} \tag{47}$$ {eq:lnlInt}

This can be solved with a change of variables (Bates et al., 2015, eq. 27):

$$\boldsymbol{\gamma} = \boldsymbol{R}_{11}(\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}) + \boldsymbol{R}_{12}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \tag{48}$$ {eq:gamma}

$$\frac{d\boldsymbol{\gamma}}{d\boldsymbol{\upsilon}} = \boldsymbol{R}_{11} \tag{49}$$ {eq:Jgamma}

Using the change of variables formula,[4] we add the inverse determinant of $\boldsymbol{R}_{11}$ when plugging eq. 48 into 47, and using eq. 49 (Bates et al., 2015, eq. 28):

$$\mathcal{L}\left(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2;\boldsymbol{y}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n+k_z}{2}}}\exp\frac{-r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) - \left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}\int\exp\frac{-\left|\left|\boldsymbol{\gamma}\right|\right|^2}{2\sigma^2}\left|\det(\boldsymbol{R}_{11})\right|^{-1}d\boldsymbol{\gamma} \tag{50}$$

$$= \frac{1}{\left|\det(\boldsymbol{R}_{11})\right|\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\frac{-r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) - \left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}\left\{\frac{1}{\left(2\pi\sigma^2\right)^{\frac{k_z}{2}}}\int\exp\frac{-\left|\left|\boldsymbol{\gamma}\right|\right|^2}{2\sigma^2}d\boldsymbol{\gamma}\right\} \tag{51}$$

and because the term in curly braces is now of a probability density function, it integrates to unity

$$\mathcal{L}\left(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2;\boldsymbol{y}\right) = \frac{1}{\left|\det(\boldsymbol{R}_{11})\right|\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\frac{-r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) - \left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2} \tag{52}$$ {eq:WeMixL0}

Having solved the integral symbolically, this expression for the likelihood no longer has an explicit integral and has been calculated exactly, without use of numerical quadrature. This derivation is incredibly close to Bates et al. (2015), with the only modification being that we use the QR decomposition of $\boldsymbol{A}$ where they used the Cholesky decomposition of $\boldsymbol{A}^T\boldsymbol{A}$.

The deviance function ($D(\cdot) \equiv -2\ell(\cdot)$),

$$D\left(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2;\boldsymbol{y}\right) = 2\ln\left|\det(\boldsymbol{R}_{11})\right| + n\ln\left(2\pi\sigma^2\right) + \frac{r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) + \left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{\sigma^2} \tag{53}$$ {eq:WeMixD0}

Using the profile likelihood, the deviance is minimized when $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ because $\boldsymbol{\beta}$ only appears inside of a sum of squares that can be minimized (set to zero) using $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ (Bates et al., 2015, eqs. 30 and 31). The profile deviance then becomes

$$D\left(\boldsymbol{\theta},\sigma^2;\boldsymbol{y}\right) = 2\ln\left|\det(\boldsymbol{R}_{11})\right| + n\ln\left(2\pi\sigma^2\right) + \frac{r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}})}{\sigma^2} \tag{54}$$

Similarly, the value of σ^2 can be found by taking the derivative of the profile deviance with respect to σ^2 and setting it equal to zero. This yields

$$\widehat{\sigma^2} = \frac{r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}})}{n} \tag{55}$$

[4] See the section, "Substitution for Multiple Variables" in "Integration by Substitution" (Wikipedia, n.d.).

giving a profile deviance that is a function only of the parameter $\boldsymbol{\theta}$:

$$D\left(\boldsymbol{\theta};\boldsymbol{y}\right)=2\ln|\det(\boldsymbol{R}_{11})|+n\left(\ln\left(2\pi\widehat{\sigma^2}\right)+1\right) \tag{56}$$

the profiled deviance can then be numerically maximixed over the parameters in $\boldsymbol{\theta}$ and the plug in estimators can then be used for $\hat{\sigma^2}$ and $\boldsymbol{\beta}$.

Weighted Case

{weighted-cas

Generally, the Rabe-Hesketh et al. (2006) model can be rewritten in a form very similar to `lme4` as

$$(\boldsymbol{y}|\boldsymbol{U}=\boldsymbol{u})\sim T_1^{\omega^{(1)}} \tag{57}$$ {eq:WeMixA}

$$\boldsymbol{U}\sim T_2^{\omega^{(2)}}*T_3^{\omega^{(3)}}*\cdots*T_L^{\omega^{(L)}} \tag{58}$$ {eq:WeMixB}

where T^k represents the convolution of k instances of T and $*$ represents the convolution of two distributions, with a likelihood that is their product; $w^{(l)}$ are the weights assigned to units ($l=1$) or groups ($l>1$) that are ideally[5] the inverse (unconditional) probability of selecting the unit or group; and the Ts have distribution

$$T_1\sim N(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u},\sigma^2 I) \tag{59}$$ {eq:WeMixA2}

$$T_l\sim N(0,\boldsymbol{\Sigma_{ll}})\quad l=2,\cdots,L \tag{60}$$ {eq:WeMixB2}

where $\boldsymbol{\Sigma}$ is block diagonal, disallowing nonzero prior correlations across levels, with the lth block being written $\boldsymbol{\Sigma}_{ll}$.

The purpose of `WeMix` is to estimate the likelihood of the weighted model (eqs. 57 and 58). In this section we derive the weighted estimator that is analogous to the estimator used by `lme4` and is similar to Henderson (1982) but we include a deviance (and likelihood), which Henderson omits.

The first difference is in the penalized sum of squares, which now weights the residuals by the unit-level weights ($\boldsymbol{\Omega}$) and the random-effect penalties by the group-level weights ($\boldsymbol{\Psi}$):

$$r^2(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\upsilon})=\left|\left|\boldsymbol{\Omega}^{\frac{1}{2}}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\sum_{l=1}^{L}\boldsymbol{Z}_l\Lambda_{ll}(\theta)\upsilon_l\right)\right|\right|^2+\sum_{l=2}^{L}\left|\left|(\boldsymbol{\Psi}_{ll})^{\frac{1}{2}}\,\boldsymbol{\upsilon}_l\right|\right|^2 \tag{61}$$ {eq:r2sep}

where $\boldsymbol{\Omega}$ and $\boldsymbol{\Psi}_{ll}$ are diagonal matrices with unconditional inverse probability of selection for each unit ($\boldsymbol{\Omega}$) or group ($\boldsymbol{\Psi}_{ll}$) along its diagonal. The unconditional probability that a unit or group was selected can be readily calculated as the product of a probability of its own probability of selection and the unconditional probability of the group to which it belongs.

Then, the weighted pseudo-data notation combines the two terms in eq. 61, adding a vector of *pseudo-data* to the end of the $\boldsymbol{y}$ vector:

$$r^2(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\upsilon})=\left|\left|\begin{bmatrix}\boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{y}\\ \boldsymbol{0}\end{bmatrix}-\begin{bmatrix}\boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{X}\\ \boldsymbol{\Psi}^{\frac{1}{2}} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\upsilon}\\ \boldsymbol{\beta}\end{bmatrix}\right|\right|^2 \tag{62}$$ {eq:WeMixWr2}

$$=\left|\left|\boldsymbol{\Omega}^{\frac{1}{2}}\left(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{\upsilon}-\boldsymbol{X}\right)\right|\right|^2+\left|\left|\boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{\upsilon}\right|\right|^2 \tag{63}$$

where $\boldsymbol{Z}$ is now a block matrix that incorporates all of the $\boldsymbol{Z}$ matrices for the various levels:

$$\boldsymbol{Z}=\begin{bmatrix}\boldsymbol{Z}_1 & \boldsymbol{Z}_2 & \cdots & \boldsymbol{Z}_L\end{bmatrix} \tag{64}$$

$\boldsymbol{\Lambda}(\boldsymbol{\theta})$ is a block diagonal matrix with elements $\boldsymbol{\Lambda}_{ll}(\boldsymbol{\theta})$, $\boldsymbol{\Psi}$ is a block diagonal matrix with elements $\boldsymbol{\Psi}_{ll}$, and

$$\boldsymbol{\upsilon}=\begin{bmatrix}\boldsymbol{\upsilon}_1\\ \boldsymbol{\upsilon}_2\\ \vdots\\ \boldsymbol{\upsilon}_L\end{bmatrix} \tag{65}$$

[5]This weight is only ideally this because of how weights are adjusted for total nonresponse (see Gelman 2007).

The likelihood of $\boldsymbol{y}$, conditional on $\boldsymbol{v}$ is then

$$f_{\boldsymbol{y}|\boldsymbol{v}=\boldsymbol{v}}(\boldsymbol{y},\boldsymbol{v}) = \prod_{i=1}^{n}\left[\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left(-\frac{||\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{v}||^2}{2\sigma^2}\right)\right]^{\boldsymbol{\Omega}_{ii}} \tag{66}$$

$$= \prod_{i=1}^{n}\frac{1}{(2\pi\sigma^2)^{\frac{\boldsymbol{\Omega}_{ii}}{2}}}\exp\left(-\frac{\left|\left|\boldsymbol{\Omega}_{ii}^{\frac{1}{2}}\left(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{v}\right)\right|\right|^2}{2\sigma^2}\right) \tag{67}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{\sum_i \boldsymbol{\Omega}_{ii}}{2}}}\exp\left(-\frac{\left|\left|\boldsymbol{\Omega}^{\frac{1}{2}}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{v}\right)\right|\right|^2}{2\sigma^2}\right) \tag{68}$$ {eq:WeMixWfyu}

And the unconditional density of $\boldsymbol{v}$ is

$$f_{\boldsymbol{v}}(\boldsymbol{v}) = \prod_{j=1}^{k_z}\left[\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left[-\frac{||\boldsymbol{v}_j||^2}{2\sigma^2}\right]\right]^{\boldsymbol{\Psi}_{jj}} \tag{69}$$

$$= \prod_{j=1}^{k_z}\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}||\boldsymbol{\Psi}_{jj}||_F}}\exp\left[-\frac{\left|\left|\boldsymbol{\Psi}_{jj}^{\frac{1}{2}}\boldsymbol{v}_j\right|\right|^2}{2\sigma^2}\right] \tag{70}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}||\boldsymbol{\Psi}_{jj}\,||_F}}\exp\left[-\frac{\left|\left|\boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{v}\right|\right|^2}{2\sigma^2}\right] \tag{71}$$ {eq:WeMixWfu}

where $||\cdot||_F$ is the Frobenius norm of a matrix, or the sum of all its elements.

The joint distribution of $\boldsymbol{v}$ and $\boldsymbol{y}$ is then the product of eqs. 68 and 71:

$$f_{\boldsymbol{y},\boldsymbol{v}}\left(\boldsymbol{y},\boldsymbol{v},\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2\right) = f_{\boldsymbol{y}|\boldsymbol{v}=\boldsymbol{v}}(\boldsymbol{y},\boldsymbol{v})\cdot f_{\boldsymbol{v}}(\boldsymbol{v}) \tag{72}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{\sum_i \boldsymbol{\Omega}_{ii}}{2}}}\exp\left[-\frac{\left|\left|\boldsymbol{\Omega}^{\frac{1}{2}}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{v}\right)\right|\right|^2}{2\sigma^2}\right]\cdot\frac{1}{(2\pi\sigma^2)^{\frac{\sum_j \boldsymbol{\Psi}_{jj}}{2}}}\exp\left[-\frac{\left|\left|\boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{v}\right|\right|^2}{2\sigma^2}\right] \tag{73}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{\sum_i \boldsymbol{\Omega}_{ii}+\sum_j \boldsymbol{\Psi}_{jj}}{2}}}\exp\left[-\frac{\left|\left|\boldsymbol{\Omega}^{\frac{1}{2}}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{v}\right)\right|\right|^2+\left|\left|\boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{v}\right|\right|^2}{2\sigma^2}\right] \tag{74}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{\sum_i \boldsymbol{\Omega}_{ii}+\sum_j \boldsymbol{\Psi}_{jj}}{2}}}\exp\left[-\frac{r^2(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{v})}{2\sigma^2}\right] \tag{75}$$

Using the same logic for the results in eq. 40, r^2 can be written as a sum of the value at the optimum ($\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{v}}$) and deviations from that:

$$f_{\boldsymbol{y},\boldsymbol{v}}\left(\boldsymbol{y},\boldsymbol{v},\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{\frac{\sum_i \boldsymbol{\Omega}_{ii}+\sum_j \boldsymbol{\Psi}_{jj}}{2}}}\exp\left[-\frac{r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{v}})-\left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2-\left|\left|\boldsymbol{R}_{11}(\boldsymbol{v}-\hat{\boldsymbol{v}})+\boldsymbol{R}_{12}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}\right] \tag{76}$$

Now, finding the integral of this over $\boldsymbol{\upsilon}$,

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2;\boldsymbol{y}) = \int f_{\boldsymbol{y},\boldsymbol{\upsilon}}\left(\boldsymbol{y},\boldsymbol{\upsilon},\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2\right) d\boldsymbol{\upsilon} \tag{77}$$

$$= \int \frac{1}{(2\pi\sigma^2)^{\frac{\sum_i \boldsymbol{\Omega}_{ii}+\sum_j \boldsymbol{\Psi}_{jj}}{2}}} \exp\left[-\frac{r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) - \left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2 - \left|\left|\boldsymbol{R}_{11}(\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}) + \boldsymbol{R}_{12}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}\right] d\boldsymbol{\upsilon} \tag{78}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{\sum_i \boldsymbol{\Omega}_{ii}+\sum_j \boldsymbol{\Psi}_{jj}}{2}}} \exp\left[-\frac{r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) - \left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}\right] \int \exp\left[-\frac{\left|\left|\boldsymbol{R}_{11}(\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}) + \boldsymbol{R}_{12}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}\right] d\boldsymbol{\upsilon} \tag{79}$$ {eq:tmpLA}

Notice that while the unweighted integral has k_z dimensions, this weighted integral has $\sum_j \boldsymbol{\Psi}_{jj}$ dimensions—the number of (population) individuals values to integrate out.

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2;\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{\sum_i \boldsymbol{\Omega}_{ii}}{2}}} \exp\left[-\frac{r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) - \left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}\right] \tag{80}$$

$$\left\{\frac{1}{(2\pi\sigma^2)^{\frac{\sum_j \boldsymbol{\Psi}_{jj}}{2}}} \int \exp\left[-\frac{\left|\left|\boldsymbol{R}_{11}(\boldsymbol{\upsilon}-\hat{\boldsymbol{\upsilon}}) + \boldsymbol{R}_{12}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}\right] d\boldsymbol{\upsilon}\right\} \tag{81}$$ {eq:tmpLB}

However, while we know there are $\sum_j \boldsymbol{\Psi}_{jj}$ dimensions to integrate out (the number of population cases), a change of variables must maintain the dimensionality of the integration, so it is not clear how to proceed. Instead, we name the term inside the integral α and use a different methodology to derive its value. Then,

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2;\boldsymbol{y}) = \alpha(\boldsymbol{\theta};\boldsymbol{\Omega},\boldsymbol{\Psi}) \frac{1}{(2\pi\sigma^2)^{\frac{\sum_i \boldsymbol{\Omega}_{ii}}{2}}} \exp\left[-\frac{r^2(\boldsymbol{\theta},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\upsilon}}) - \left|\left|\boldsymbol{R}_{22}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|\right|^2}{2\sigma^2}\right] \tag{82}$$ {eq:tmpL2}

where α is a constant for a fixed $\boldsymbol{\theta}$ and set of weights $\boldsymbol{\Omega}$ and $\boldsymbol{\Psi}$.

While these formulas allow for estimation of a likelihood function that allows for estimation of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma^2}$ via profiling, they do not depend on α because $\boldsymbol{\theta}$ appears as a parameter of α; optimizing the log-likelihood with respect to $\boldsymbol{\theta}$ requires all of the terms in the log-likelihood to be calculated, including α.

Calculation of α

Bates and Pinheiro (1998) offer an unweighted method of calculating α that we extend here to admit weighting. The essential insight of Bates and Pinheiro is that the variables must be remapped, per group, using an orthogonal transform that separates out the q_l random effects associated with group g at level l. In what follows we describe a three-level case, but the methods readily generalize to the L-level case.

The integral is slightly re-expressed using $\boldsymbol{u}$ instead of υ, but instead of using $\boldsymbol{\Lambda}$ it uses the $\boldsymbol{\Delta}$ matrix, which is an individual block of $\boldsymbol{\Lambda}$, so $\boldsymbol{\Delta}$ is defined, implicitly, by

$$\boldsymbol{\Lambda}(\boldsymbol{\theta}) = \boldsymbol{\Delta}(\boldsymbol{\theta}) \otimes \boldsymbol{I} \tag{83}$$

where $\otimes$ is the Kronecker product.

Using this notation, the likelihood is then given by

$$\mathcal{L} = \int \prod_g \left[\int f_{y|U}(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{u}_{2g}, \cdots, \boldsymbol{\upsilon}_{Lg}) f_U(\boldsymbol{\theta}, \boldsymbol{u}_{2g}) d\boldsymbol{u}_{2g} \right] f_U(\boldsymbol{\theta}, \boldsymbol{u}_{3g}) d\boldsymbol{u}_{3g} \tag{84}$$

$$\propto \int \prod_g \left[\int \exp \left[\left|\left| \Omega_{gg}^{\frac{1}{2}} (\boldsymbol{y}_g - \boldsymbol{X}_g \boldsymbol{\beta} - \boldsymbol{Z}_g \boldsymbol{u}) \right|\right|^2 + \left|\left| \boldsymbol{u}_{2g}^T \boldsymbol{\Delta}^T \boldsymbol{\Delta} \boldsymbol{u}_{2g} \right|\right|^2 \right] d\boldsymbol{u}_{2g} \right] f_U(\boldsymbol{\theta}, \boldsymbol{u}_{3g}) d\boldsymbol{u}_{3g} \tag{85}$$

and iteratively rewritten to symbolically integrate out the lowest level random effects, starting with level 2 and increasing until there are no remaining integrals. When this is done, we will note the change of variable associated with a weighted model and use that for α in eq. 82. Because of that, the portions unrelated to the change of variable were dropped from relation 85. Thus, notice that this goal is just for units in a particular group (g) at level 2. These results of several integrals have to be combined to solve the integral across all groups and calculate α. The value of α will be calculated by level and then summed; for a three-level model:

$$\alpha = \alpha_2 + \alpha_3 \tag{86}$$

While the residual sum of squares (r^2) has been, up until now, treated for the entire data set, the notion is to think of the residual sum of squares for just group g (at level l). Bates and Pinheiro (1998) also use the notion of r_g^2, or the r^2 contribution for group g, which is defined as

$$r_g^2(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\upsilon}) = \left|\left| \begin{bmatrix} \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{y}_g \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{Z}_g & \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} X_g \\ \boldsymbol{\Delta}_l' & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_g \\ \boldsymbol{\beta} \end{bmatrix} \right|\right|^2 \tag{87}$$

where $\boldsymbol{y}_g$, $\boldsymbol{\upsilon}_g$, $\boldsymbol{X}_g$, $\boldsymbol{Z}_g$, and $\boldsymbol{\Omega}_{gg}$ are the rows of the $\boldsymbol{y}$ vector, the $\boldsymbol{\upsilon}$ vector, $\boldsymbol{X}$ matrix, $\boldsymbol{Z}$ matrix, and $\boldsymbol{\Omega}$ matrix that are associated with group g, respectively; while $\boldsymbol{\Lambda}(\boldsymbol{\theta})$ is the full $\boldsymbol{\Lambda}(\theta)$ matrix, $\boldsymbol{\Delta}_l'$ is a block matrix:

$$\boldsymbol{\Delta}_l' \equiv \begin{bmatrix} \boldsymbol{\Delta}_l & \boldsymbol{0} \end{bmatrix} \tag{88}$$

where $\boldsymbol{\Delta}_l$ is the portion of $\boldsymbol{\Delta}$ associated with level l, and the $\boldsymbol{0}$ matrix is entirely zeros and conforms to the portion of $\boldsymbol{u}_g$ that is associated with level 3 of the model.

Expanding $\boldsymbol{Z}_g$ gives

$$r_g^2(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\upsilon}) = \left|\left| \begin{bmatrix} \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{y}_g \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{Z}_{2g} & \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{Z}_{3g} & \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} X_g \\ \boldsymbol{\Delta}_2 & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\upsilon}_{2g} \\ \boldsymbol{\upsilon}_{3g} \\ \boldsymbol{\beta} \end{bmatrix} \right|\right|^2 \tag{89}$$

Starting at level 2, a change of variables is chosen via the (full) QR decomposition,

$$\begin{bmatrix} \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{Z}_{2g} \\ \boldsymbol{\Delta}_2 \end{bmatrix} = \boldsymbol{Q}_{2g} \begin{bmatrix} \boldsymbol{R}_{12g} \\ \boldsymbol{0} \end{bmatrix} \tag{90}$$

where the subscript on $\boldsymbol{R}_{12g}$ indicates that it is the top submatrix (1), at level 2, and for group g. Notice that the blocks are different shapes on the left- and right-hand sides of eq. 90; on the left-hand side, the top block ($\boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{Z}_{2g}$) has as many rows as there are observations in group g and the bottom block ($\boldsymbol{\Delta}_2$) has as many rows as there are random effects at level 2, while the right-hand side is flipped. The top block ($\boldsymbol{R}_{12g}$) has as many rows as there are random effects at level 2, while the bottom block ($\boldsymbol{0}$) has as many rows as there are observations in group g. The reason for this is that the change makes the top block on the right-hand side a square, upper triangular matrix, which will allow the change of variables to proceed.

Because $\boldsymbol{Q}_{2g}$ is orthogonal by construction, $\boldsymbol{Q}_{2g}^T \boldsymbol{Q}_{2g} = I$, one can freely premultiply the inside of the sum of squares in eq. 89 by $\boldsymbol{Q}_{2g}^T$ without changing the sum of squares:[6]

$$r_g^2(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\upsilon}) = \left|\left| \boldsymbol{Q}_{2g}^T \left\{ \begin{bmatrix} \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{y}_g \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{Z}_{2g} & \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{Z}_{3g} & \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{X}_g \\ \boldsymbol{\Delta}_2 & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\upsilon}_{2g} \\ \boldsymbol{\upsilon}_{3g} \\ \boldsymbol{\beta} \end{bmatrix} \right\} \right|\right|^2 \tag{91}$$

[6]Recalling $||\boldsymbol{x}||^2 = \boldsymbol{x}^T \boldsymbol{x}$, so that, for orthogonal matrix $\boldsymbol{Q}$, it is easy to see $||\boldsymbol{Q}\boldsymbol{x}||^2 = \boldsymbol{x}^T \boldsymbol{Q}^T \boldsymbol{Q} \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{x}$.

Then, defining

$$\boldsymbol{Q}_{2g}^T \begin{bmatrix} \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{y}_g \\ \boldsymbol{0} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{R}_{1yg} \\ \boldsymbol{R}_{2yg} \end{bmatrix} \qquad \boldsymbol{Q}_{2g}^T \begin{bmatrix} \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{Z}_{3g} \\ \boldsymbol{0} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{R}_{13g} \\ \boldsymbol{R}_{23g} \end{bmatrix} \qquad \boldsymbol{Q}_{2g}^T \begin{bmatrix} \boldsymbol{\Omega}_{gg}^{\frac{1}{2}} \boldsymbol{X}_g \\ \boldsymbol{0} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{R}_{1Xg} \\ \boldsymbol{R}_{2Xg} \end{bmatrix} \tag{92}$$ {eq:Qtdef}

similar to eq. 90, and with the same dimensions, the blocks are of different shapes on the left and right of the equation.

Multiplying the $\boldsymbol{Q}_{2g}^T$ through and plugging the eq. 92 equations into eq. 91,

$$r_g^2(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}) = \left|\left| \begin{bmatrix} \boldsymbol{R}_{1yg} \\ \boldsymbol{R}_{2yg} \end{bmatrix} - \begin{bmatrix} \boldsymbol{R}_{12g} & \boldsymbol{R}_{13g} & \boldsymbol{R}_{1Xg} \\ \boldsymbol{0} & \boldsymbol{R}_{23g} & \boldsymbol{R}_{2Xg} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_{2g} \\ \boldsymbol{u}_{3g} \\ \boldsymbol{\beta} \end{bmatrix} \right|\right|^2 \tag{93}$$

it is now possible to simplify the integral, do a change of variables and integrate out level 2 random effects for group g, and solve the first integral in eq. 85 symbolically. In particular, this rewriting of the terms means there are q_2 terms with $\boldsymbol{u}_{2g}$ in them and $n_g - q_2$ terms that are redefined to be orthogonal to those two terms. It is this orthogonality that allows the other terms to be removed from the integral. So, since we have just shown

$$||\Omega_{gg}^{\frac{1}{2}}(\boldsymbol{y}_g - \boldsymbol{X}_g\boldsymbol{\beta} - \boldsymbol{Z}_g\boldsymbol{\Lambda}_{ll}(\boldsymbol{\theta})\boldsymbol{u})||^2 + ||\boldsymbol{u}||^2 = \left|\left| \begin{bmatrix} \boldsymbol{R}_{1yg} \\ \boldsymbol{R}_{2yg} \end{bmatrix} - \begin{bmatrix} \boldsymbol{R}_{12g} & \boldsymbol{R}_{13g} & \boldsymbol{R}_{1Xg} \\ \boldsymbol{0} & \boldsymbol{R}_{23g} & \boldsymbol{R}_{2Xg} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_{2g} \\ \boldsymbol{u}_{3g} \\ \boldsymbol{\beta} \end{bmatrix} \right|\right|^2 \tag{94}$$

$$\begin{aligned} &= ||\boldsymbol{R}_{1yg} - \boldsymbol{R}_{12g}\boldsymbol{u}_{2g} - \boldsymbol{R}_{13g}\boldsymbol{u}_{3g} - \boldsymbol{R}_{1Xg}\boldsymbol{\beta}_g||^2 \\ &\quad + ||\boldsymbol{R}_{2yg} - \boldsymbol{R}_{23g}\boldsymbol{u}_{3g} - \boldsymbol{R}_{2Xg}\boldsymbol{\beta}_g||^2 \end{aligned} \tag{95}$$

which can now be substituted into eq. 85, allowing a change of variables:

$$\boldsymbol{\gamma}_{2g} = \boldsymbol{R}_{1yg} - \boldsymbol{R}_{12g}\boldsymbol{u}_{2g} - \boldsymbol{R}_{13g}\boldsymbol{u}_{3g} - \boldsymbol{R}_{1Xg}\boldsymbol{\beta}_g \tag{96}$$

$$\frac{d\boldsymbol{\gamma}_{2g}}{d\boldsymbol{u}_{2g}} = \boldsymbol{R}_{12g} \tag{97}$$

The value of α_2 in the unweighted case is now clear:

$$\alpha_{2u} = \sum_{g=1}^{n_2} |\det(\boldsymbol{R}_{12g})|^{-1} \tag{98}$$

where α_{2u} is the unweighted alpha. This formula can be weighted simply by applying the replicate weights to the individual terms:

$$\alpha_2 = \sum_{g=1}^{n_2} \boldsymbol{\Psi}_{gg} |\det(\boldsymbol{R}_{12g})|^{-1} \tag{99}$$

However, for the three-level model, the likelihood still has level 3 integrals. The level 3 integral can also be removed. We cannot restart this process with the original $\boldsymbol{Z}$ and $\boldsymbol{X}$ matrices and other components because they change with the components inside the level 2 integral. However, the $\boldsymbol{R}_{2**}$ matrices are the portions of the higher level $\boldsymbol{u}_{3g}$ that were not integrated out, and they can be used independent of $\boldsymbol{u}_{2g}$.

We continue on with these remapped variables, starting the unweighted case (only at level 2), and now using g' to indicate that this is a different group (at level 3), with $n_{g'}$ subgroups in it. Each group (labeled i) contributes an outcomes matrix $\boldsymbol{R}_{2yi}$, a matrix per level 2 group $\boldsymbol{R}_{23i}$ and a matrix per fixed effects regressor $\boldsymbol{X}_{2Xi}$, for $i = 1, \cdots, n_{g'}$. Combining these, the residual sum of squares at level 3 for the group g' is

$$r_{g'}^2(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\upsilon}) = \left|\left| \begin{bmatrix} \boldsymbol{R}_{2y1} \\ \vdots \\ \boldsymbol{R}_{2yn_{g'}} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{R}_{231} & \boldsymbol{R}_{2X1} \\ \vdots & \vdots \\ \boldsymbol{R}_{23n_{g'}} & \boldsymbol{R}_{2Xn_{g'}} \\ \boldsymbol{\Delta}_3 & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_{3g} \\ \boldsymbol{\beta} \end{bmatrix} \right|\right|^2 \tag{100}$$ {eq:RunW}

Following the example at level 2 above, the QR decomposition is then used:

$$\begin{bmatrix} \boldsymbol{R}_{231} \\ \vdots \\ \boldsymbol{R}_{23n_{g'}} \\ \boldsymbol{\Delta}_3 \end{bmatrix} = \boldsymbol{Q}_{3g',u} \begin{bmatrix} \boldsymbol{R}_{13g',u} \\ \boldsymbol{0} \end{bmatrix} \tag{101}$$

where a subscript u is used to indicate that $\boldsymbol{Q}_{3g',u}$ and $\boldsymbol{R}_{13g',u}$ are unweighted. The remaining steps are then identical to the level 2 case; $\boldsymbol{R}_{13g',u}$ is used as the change of variables, so that $\alpha_{3u} = \sum_{g'=1}^{n_2} |\det(\boldsymbol{R}_{13g',u})|^{-1}$, and the other $\boldsymbol{R}_{3**}$ matrices can be used to integrate out level-4 cases and so on.

When there are level 2 conditional weights—non-unit probabilities of selection for the groups at level 2, conditional on the selection of the level 3 unit—each matrix in eq. 100 could be replicated $\boldsymbol{\Psi}_{ii}$ times. Equivalently, each matrix can be weighted by the conditional probability of selection, so that eq. 100 becomes

$$r^2_{g'}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\upsilon}) = \left\| \begin{bmatrix} \boldsymbol{\Psi}_{11}^{\frac{1}{2}} \boldsymbol{R}_{2y1} \\ \vdots \\ \boldsymbol{\Psi}_{g'g'}^{\frac{1}{2}} \boldsymbol{R}_{2yn_{g'}} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Psi}_{11}^{\frac{1}{2}} \boldsymbol{R}_{231} & \boldsymbol{\Psi}_{11}^{\frac{1}{2}} \boldsymbol{R}_{2X1} \\ \vdots & \vdots \\ \boldsymbol{\Psi}_{g'g'}^{\frac{1}{2}} \boldsymbol{R}_{23n_{g'}} & \boldsymbol{\Psi}_{g'g'}^{\frac{1}{2}} \boldsymbol{R}_{2Xn_{g'}} \\ \boldsymbol{\Delta}_3 & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_{3g} \\ \boldsymbol{\beta} \end{bmatrix} \right\|^2 \tag{102}$$ {eq:R}

This change leads to the same QR decomposition as the replicated case:

$$\begin{bmatrix} \boldsymbol{\Psi}_{11}^{\frac{1}{2}} \boldsymbol{R}_{231} \\ \vdots \\ \boldsymbol{\Psi}_{g'g'}^{\frac{1}{2}} \boldsymbol{R}_{23n_{g'}} \\ \boldsymbol{\Delta}_3 \end{bmatrix} = \boldsymbol{Q}_{3g'} \begin{bmatrix} \boldsymbol{R}_{13g'} \\ \boldsymbol{0} \end{bmatrix} \tag{103}$$ {eq:QRg3}

The weighted value of α_3 for the third level is then

$$\alpha_3 = \sum_{g'=1}^{n_3} \boldsymbol{\Psi}_{g'g'} \, |\boldsymbol{R}_{13g'}|^{-1} \tag{104}$$

Implementation Details

The actual implementation of the calculation is slightly different from what is above. First, when the QR decomposition is taken (eqs. 90 and 103), it is possible to stop the decomposition as is described in Bates and Pinheiro (1998). It is also possible to continue the QR decomposition for the other levels of $\boldsymbol{Z}$. Using the $\boldsymbol{QR}$ on the entire matrix still results in an orthogonal component in the submatrices, and so meets the goals of the decomposition while obviating the need to form the $\boldsymbol{Q}$ matrix explicitly.

Also note that, in the above, the value of r^2 was never used, so the components relating to $\boldsymbol{y}$ and $\boldsymbol{X}$ need not be formed.

Estimation {estimation}

Continuing to follow `lme4`, the estimation uses the profile likelihood. Since $\boldsymbol{\beta}$ appears only in the final term in quadratic form, it is immediately evident that the maximum likelihood estimator (MLE) for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$, making eq. 82 profile to

$$D\left(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}\right) = 2\ln\left(\alpha(\boldsymbol{\theta}; \Psi, \Omega)\right) + \left(\sum_{\mathrm{i}} \Omega_{\mathrm{ii}}\right) \ln\left(2\pi\sigma^2\right) + \frac{\mathrm{r}^2(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\upsilon}})}{\sigma^2} \tag{105}$$ {eq:WeMixPB}

Then, the value of σ^2 can also be profiled out by taking the derivative of the deviance with respect to σ^2 and setting it equal to zero (Bates et al., 2015, eq. 32):

$$0 = \frac{\sum_i \boldsymbol{\Omega}_{ii}}{\widehat{\sigma^2}} - \frac{r^2(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\upsilon}})}{\widehat{\sigma^4}} \tag{106}$$

rearranging

$$\widehat{\sigma^2} = \frac{r^2(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\upsilon}})}{\sum_i \boldsymbol{\Omega}_{ii}} \tag{107}$$ {eq:WeMixMaxs

Eq. 107 can then be plugged into eq. 105 to give

$$D\left(\boldsymbol{\theta}; \boldsymbol{y}\right) = 2\ln\left(\alpha(\boldsymbol{\theta}; \Psi, \Omega)\right) + \left(\sum_{\mathrm{i}} \Omega_{\mathrm{ii}}\right)\left[\ln\left(2\pi\widehat{\sigma^2}\right) + 1\right] \tag{108}$$ {eq:WeMixP}

This function is then minimized numerically with respect to $\boldsymbol{\theta}$, using the profile estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\upsilon}$ (eq. 62) and $\widehat{\sigma^2}$ (eq. 108).

The estimated values are then the $\boldsymbol{\theta}$ that maximize eq. 108, the σ^2 value from eq. 107, and the $\boldsymbol{\beta}$ values from solving the system of equations in eq. 62.

Variance Estimation

{variance-est

The inverse Hessian of $\boldsymbol{\beta}$ is given by (Bates et al., 2015, eq. 54):

$$\widehat{\mathrm{Var}}(\boldsymbol{\beta}) = \widehat{\sigma^2}\boldsymbol{R}_{22}^{-1}\left(\boldsymbol{R}_{22}^{T}\right)^{-1} \tag{109}$$

with $\boldsymbol{R}_{22}$ coming from eq. 82. This variance estimator assumes that the weights are information weights and so is inappropriate for survey weights.

A robust (sandwich) variance estimator is given by (Binder, 1983) is appropriate:

$$\left(\widehat{\sigma^2}\boldsymbol{R}_{22}^{T}\right)^{-1}\boldsymbol{J}\left(\widehat{\sigma^2}\boldsymbol{R}_{22}^{T}\right)^{-1} \tag{110}$$ {eq:sandwich}

where $\boldsymbol{J}$ is the sum of outer products of the Jacobian matrix

$$\boldsymbol{J} = \frac{n_L}{n_L - 1}\sum_{g=1}^{n_L}\frac{\partial(\ell_g)}{\partial\boldsymbol{\beta}} \tag{111}$$

where n_L is the number of level-L (top-level) groups, g indexes level-L groups, and ℓ_g is the log-likelihood for group g and all groups and units nested inside of g. The log-likelihood of the full model is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2; \boldsymbol{y}) = \ln\left[\alpha(\boldsymbol{\theta}; \Omega, \Psi)\right] - \frac{\sum_{\mathrm{i}}\Omega_{\mathrm{ii}}}{2}\ln\left(2\pi\sigma^2\right) - \frac{\mathrm{r}^2\left(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\upsilon}}\right)}{2\sigma^2} - \frac{\left|\left|\mathrm{R}_{22}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right|\right|^2}{2\sigma^2} \tag{112}$$ {eq:Vlnl}

where we are allowing $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ to vary but are fixing σ^2 at the estimated value of $\hat{\sigma}^2$. This could have been annotated by making $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ because $\hat{\boldsymbol{\beta}}$ is the estimated value conditional on $\boldsymbol{\theta}$ and appears in the equation separate from the value of $\boldsymbol{\beta}$, but that is not shown here.

While it would be convenient if eq. 112 could be directly broken up into a portion attributable to each group, and some encouragement appears when the first three terms can be, the final term has dependencies across multiple groups. A distinct likelihood is needed that depends only on the data in that group. This is achieved by noting that data for a particular group is also valid data for a mixed model of the same type as the global

mixed model, and so eq. 112 can be used on a single group's data to get the group log-likelihood; thus a group log-likelihood can be written using the notion of the fitted value of $\boldsymbol{\beta}$ in the group ($\hat{\boldsymbol{\beta}}_g$)

$$\ell_g(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2; \boldsymbol{y}_g) = \ln\left[\alpha_{\mathrm{g}}(\boldsymbol{\theta}; \Omega, \Psi)\right] - \frac{\sum_{\mathrm{i}\in\mathrm{g}} \Omega_{\mathrm{ii}}}{2} \ln\left(2\pi\sigma^2\right) - \frac{\mathrm{r}^2\left(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_{\mathrm{g}}, \hat{\boldsymbol{v}}_{\mathrm{g}}\right)}{2\sigma^2} - \frac{\left|\left|\mathrm{R}_{22\mathrm{g}}\left(\hat{\boldsymbol{\beta}}_{\mathrm{g}} - \boldsymbol{\beta}\right)\right|\right|^2}{2\sigma^2} \tag{113}$$ {eq:Vlnlg}

where α_g is the α term for group g and any groups nested in it, the sum for Ω is just over i terms associated with group g, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{v}}$ are the values fitted only on this group, and $\boldsymbol{R}_{22g}$ is the result of a QR on a version of $\boldsymbol{A}$ performed on just data ($\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{y}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Omega}$) associated with group g, while the values of σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$ are the values from the value the function is being evaluated at globally. Then,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2; \boldsymbol{y}) = \sum_g \ell_g(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2; \boldsymbol{y}_g) \tag{114}$$

A few notes are required at this point on how, exactly, this is calculated in degenerate cases. When the matrix $\boldsymbol{A}_g$ is singular for a group (e.g., when there is only one unit in the group), then the inestimable values of $\boldsymbol{\beta}_g$ are set to zero when forming $\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}$. Similarly, $\boldsymbol{R}_{22g}$ may not have enough rows to form a upper-triangular matrix. The portion that can be formed (including all columns) is then used—this does not affect the computation of $\boldsymbol{R}_{22g}\left(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}\right)$.

Using these formulas the Jacobian matrix can now be calculated numerically and the robust variance estimator formed with eq. 110.

So far this section has regarded only $\boldsymbol{\beta}$ but similar methods apply to the estimation of the variance of the random effect variance estimates ($\boldsymbol{\theta}$ and σ). These variance terms have their variance estimated assuming that they are uncorrelated with the $\boldsymbol{\beta}$ terms. At each level the variance is calculated, including a term for σ, as

$$\mathrm{Var}\left(\boldsymbol{\theta}, \sigma\right) = \left(-\boldsymbol{H}\right)^{-1} \boldsymbol{J}_{\boldsymbol{\theta},\sigma} \left(-\boldsymbol{H}\right)^{-1} \tag{115}$$ {vartheta}

where $\boldsymbol{H}$ is the Hessian of the likelihood (eq. 112) with respect to $\boldsymbol{\theta}$ and σ while $\boldsymbol{J}_{\boldsymbol{\theta},\sigma}$ is the portion of the Jacobian that regards $\boldsymbol{\theta}$ and σ. The estimated value for the variance of σ from the lowest level group (level 2) is used to form the standard error of the residual variance.

However, the variance estimates are not simply the values of $\boldsymbol{\theta}$ and σ but transformations of that (eq. 15). To estimate the variances of the variance estimates, the Delta method is used so that

$$\mathrm{Var}\left(\boldsymbol{\Sigma}, \sigma^2\right) = \left[\nabla(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})\right]^T \mathrm{Var}\left(\boldsymbol{\theta}, \sigma^2\right) \left[\nabla(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})\right] \tag{116}$$

where the gradient ($\nabla(\cdot)$) is taken with respect to the elements of $\boldsymbol{\Sigma}$ and σ^2, and $\mathrm{Var}\left(\boldsymbol{\theta}, \sigma^2\right)$ is from eq. 115.

Model Evaluation: Wald Test

{model-evalua

We can use the a Wald test to test both fixed effects parameters (β) and variance of the random parameters (Λ).

The Wald test compares estimated parameters with null hypothesis values. In the default case the null hypothesis is that value of the parameters is 0.

In this default case, if the test fails to reject the null hypothesis, removing the variables from the model will not substantially harm the fit of that model.

One advantage of the Wald test is that it can be used to test multiple hypotheses about multiple parameters simultaneously.

To test q hypotheses on p estimated parameters, let $\hat{P}$ be the vector of estimated coefficients, R be a q x p hypothesis matrix (this matrix has 1 row per coefficient being tested with a value of 1 in the column

corresponding to that coefficient), $\hat{V}$ be the estimated covariance matrix for $\hat{P}$, and r be the vector of hypothesized values for $\hat{\beta}$.

Then the Wald test statistic for multiple parameters is equal to:

$$W = (R\hat{\beta} - r)'(R\hat{V}R')^{-1}(R\hat{\beta} - r)$$

The resulting test statistic can be tested against a chi-square distribution. For this test, the degrees of freedom is the number of parameters that are tested.

$$W \sim \chi^2(p)$$

References

{references}

Bates, D., Machler, M., Bolker, B. M., & Walker, S. C. (2015), Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(*1*), 1–48.

Bates, D., & Pinheiro, J. C. (1998). *Computational methods for multilevel modelling.* Bell Labs Report.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*(*3*), 279–292.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(*2*), 153–164.

Henderson, C. R. (1982). Analysis of covariance in the mixed model: higher-level, nonhomogeneous, and random regressions. *Biometrics*, *38*(*3*), 623–640.

Integration by Substitution. (n.d.). In Wikipedia. Retrieved February 13, 2019, from https://en.wikipedia.org/wiki/Integration_by_substitution

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society*. Series A (Statistics in Society), *169*(*4*), 805–827.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal*, *2*, 1–21.

Trefethen, L. N., & Bau, D. (1997). *Numerical linear algebra.* Philadelphia, PA: SIAM.