Mixed models with balanced random effects are equivalent to standard ANOVA models

VEGARD NYGAARD, EINAR ANDREAS RØDLAND*, EIVIND HOVIG

Summary

There are different methods for analysing mixed random effects models, and differing opinions as to which of these methods are more appropriate. Even when all modelling assumptions are met, many of these methods rely on approximations, eg Satterthwaite's method for estimating appropriate effective degrees of freedom for use with F distributions. However, in some cases where the random effects are balanced, they may be eliminated from the model without loss of information, potentially reducing a random effects model to a simple model. We demonstrate that this is the case for case when each subject in a trial is measured a fixed number of times, each time with a random error due to eg techical variation.

1 Introduction

We investigate the statistical model

$$X_{rk} = \alpha \cdot A_r + \epsilon_r + \delta_{rk} \tag{1}$$

where $r = 1, \ldots, n$ represent the subjects, and $k = 1, \ldots, K$ enumerates K independent measurements per subject, $\alpha = [\alpha_1, \ldots, \alpha_p$ is a vector of model parameters, and A_r is the design vector for subject r. We assume independent error terms $\epsilon_r \sim N(0, \sigma_\epsilon^2)$ per subject, and $\delta_{rk} \sim N(0, \sigma_\delta^2)$ per measurement. This model may be thought of as a traditional linear model per subject, but with K independent measurements or replicates per subject, each with its own additional error term.

A particular case of this model is

$$X_{ijrk} = \alpha + \beta_i + \gamma_j + \epsilon_{ijr} + \delta_{ijrk} \tag{2}$$

where i are different batches, j are different study groups, $r = 1, \ldots, n_{ij}$ are subjects per batch and group, and $k = 1, \ldots, K$ enumerate K independent measurements per subject.

All three alternative anlyses proposed by Towfic et al. in their Letter to the Editor correspond, explicitly or implicitly, to model (2). The two mixed random effects models, using `aov` and `lmer`, explicitly specify this model (although for their implementation of the `aov` model the syntax was incorrect). The linear model using `lmFit` with `duplicateCorrelation` implicitly corresponds to the same model: model (2) could alternernatively be specified as $X_{rk} = \alpha + \beta_i + \gamma_j + \epsilon_{ijrk}$ where $\epsilon_{ijrk} \sim N(0, \sigma_\epsilon^2 + \sigma_\delta^2)$ and correlation $\mathrm{corr}(\epsilon_{ijrk}, \epsilon_{ijrk'}) = \sigma_\epsilon^2/(\sigma_\epsilon^2 + \sigma_\delta^2)$ for two different measurements on the same subject.

We demonstrate that model (1) reduces to a simpler model for the average, X_r, of the K measurements X_{rk}:

$$X_r = \alpha \cdot A_r + \epsilon'_r \tag{3}$$

with $\epsilon'_r = \epsilon_r + \delta_r$ where δ_r is the average of the error terms δ_{rk} across the measurements. Furthermore, the conditial distribution of X_{rk} given X_r depends on σ_δ^2 only, not on α or σ_ϵ^2, and so does not contain any additional information that can be used in the estimation of α.

This result also applies if a prior distribution is imposed on the model parameters, α, making it apply to more complex mixed random effects models as well as Bayesian models: eg if the batch effects in model (2) are assumed to follow some prior distribution.

Please note that we do not claim any form of originality. Indeed, the stated results are generally known within the statistical community. However, since we refer to this result in our letter of reply, we found it natural to substantiate and elaborate on it.

*To whom correspondence should be addressed: `einarro@ifi.uio.no`

2 Splitting model into subject and measurement effects

Let $X_r = \frac{1}{K}\sum_{k=1}^{K} X_{rk}$ and $\delta_r = \frac{1}{K}\sum_{k=1}^{K}\delta_{rk}$ be the averages across measurements on the same subject, r. By averaging both sides of model (1), we have

$$X_r = \alpha \cdot A_r + \epsilon'_r \tag{4}$$

where $\epsilon'_r = \epsilon_r + \delta_r \sim N(0, \sigma_\epsilon^2 + \sigma_\delta^2/K)$. While this proves that model (3) is correct, it is not clear that no information about the model parameters, α, has been lost in the process.

The remaining information in the data may be encoded by

$$\Delta X_{rk} = X_{rk} - X_r = \delta_{rk} - \delta_r \sim N(0, \sigma_\delta^2 V) \tag{5}$$

where $V = I - 1^T 1/K$ is the $K \times K$ matrix of rank $K-1$ that projects a vector $u = [u_1, \ldots, u_K]$ to $\Delta u = [u_1 - \Delta u, \ldots, u_K - \Delta u]$ for $\Delta u = \frac{1}{K}\sum_{k=1}^{K} u_k$.

As may be seen, the averaged model (4) depends on the parameters α and $\sigma^2 = \sigma_\epsilon^2 + \sigma_\delta^2/K$, while the model (5) for the difference from the average depends only on σ_δ^2. Since σ^2 and σ_δ^2 are linearly independent (in terms of σ_ϵ^2 and σ_δ^2), they may be estimated independently. The only relation between these two models is that by requiring that both σ_ϵ^2 and σ_δ^2 be non-negative, we get the restriction that $\sigma^2 \geq \sigma_\delta^2$: this is enforced in some implementations of mixed random effects models (eg `lmer`), while not in others (eg `aov` with random effects). If $\sigma^2 \geq \sigma_\delta^2$ is enforced, the estimates of σ^2 and σ_δ^2 may be affected, but estimates of α will remain unaffected and dependend on the averages X_r only.

2.1 Alternative derivation

An alternative derivation may be obtained from writing down the likelihood function of model (1) and split this into the product of two independent likelihood functions: one corresponding to model (4) and one to model (5). In terms of likelihood functions, the equivalent result, which we will show, is

$$L(x..|\alpha, \sigma_\epsilon^2, \sigma_\delta^2) = L(x.|\alpha, \sigma^2) \times L(x..|x., \sigma_\delta^2), \tag{6}$$

from which we again see that the complete data, $x..$, with all measurements only depends on the parameters α through the per subject averages, $x.$, and so does not give any additional information about α.

Let $x.. = (x_{rk})_{r=1,\ldots,n;k=1,\ldots,K}$ and $x_{r\cdot} = (x_{rk})_{k=1,\ldots,K}$ refer to the data for all subjects and per subject, respectively. The likelihood function may then be expressed as follows:

$$\begin{aligned}
L(x..|\alpha, \sigma_\epsilon^2, \sigma_\delta^2) &= \prod_{r=1}^{n} L(x_{r\cdot}|\alpha, \sigma_\epsilon^2, \sigma_\delta^2) \\
&= \prod_{r=1}^{n} \int L(\epsilon_r|\sigma_\epsilon^2)\, L(x_{r\cdot}|\epsilon_r, \alpha, \sigma_\delta^2)\, d\epsilon_r \\
&= \prod_{r=1}^{n} \int \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left(-\frac{\epsilon_r^2}{2\sigma_\epsilon^2}\right) \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma_\delta^2}} \exp\left(-\frac{(x_{rk} - \alpha \cdot a_r - \epsilon_r)^2}{2\sigma_\delta^2}\right) d\epsilon_r \\
&= C \prod_{r=1}^{n} \int \exp\left(-\frac{\epsilon_r^2}{2\sigma_\epsilon^2} - K \cdot \frac{(x_r - \alpha \cdot a_r - \epsilon_r)^2}{2\sigma_\delta^2}\right) \prod_{k=1}^{K} \exp\left(-\frac{\Delta x_{rk}^2}{2\sigma_\delta^2}\right) d\epsilon_r
\end{aligned} \tag{7}$$

where $C = \frac{1}{(2\pi\sigma_\epsilon^2)^{n/2}(2\pi\sigma_\delta^2)^{nK/2}}$ and $\Delta x_{rk} = x_{rk} - x_r$ with $x_r = \frac{1}{K}\sum_{k=1}^{K} x_{rk}$. Using $y_r = x_r - \alpha \cdot a_r$ to simplify, we continue:

$$\begin{aligned}
&= C \prod_{r=1}^{n} \int \exp\left(-\frac{\epsilon_r^2}{2\sigma_\epsilon^2} - \frac{K y_r^2}{2\sigma_\delta^2}\right) d\epsilon_r \cdot \prod_{k=1}^{K} \exp\left(-\frac{\Delta x_{rk}^2}{2\sigma_\delta^2}\right) \\
&= C \prod_{r=1}^{n} \int \exp\left[-\frac{1}{2s^2}\left(\epsilon_r - \frac{K s^2 y_r}{\sigma_\epsilon^2}\right)^2 - \frac{y_r^2}{2\sigma^2}\right] d\epsilon_r \cdot \exp\left(-\sum_{k=1}^{K} \frac{\Delta x_{rk}^2}{2\sigma_\delta^2}\right)
\end{aligned} \tag{8}$$

where $\frac{1}{s^2} = \frac{1}{\sigma_\epsilon^2} + \frac{K}{\sigma_\delta^2}$ and $\sigma^2 = \sigma_\epsilon^2 + \sigma_\delta^2/K$

$$= D \prod_{r=1}^{n} \exp\left(-\frac{y_r^2}{2\sigma^2}\right) \cdot \exp\left(-\sum_{k=1}^{K} \frac{\Delta x_{rk}^2}{2\sigma_\delta^2}\right) \tag{9}$$

where $D = C \cdot (2\pi s^2)^{n/2}$

$$\begin{aligned} &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\sum_{r=1}^{n} \frac{y_r^2}{2\sigma^2}\right) \times \frac{1}{(2\pi\sigma_\delta^2)^{n(K-1)/2}} \exp\left(-\sum_{r=1}^{n}\sum_{k=1}^{K} \frac{\Delta x_{rk}^2}{2\sigma_\delta^2}\right) \qquad (10)\\ &= \prod_{r=1}^{n} L(x_r|\alpha, \sigma^2) \times \prod_{r=1}^{n}\prod_{k=1}^{K} L(x_{rk}|x_r, \sigma_\delta^2) \\ &= L(x_\cdot|\alpha, \sigma^2) \times L(x_{\cdot\cdot}|x_\cdot, \sigma_\delta^2) \end{aligned}$$

2.2 Extension to Bayesian models or model restrictions on parameters

Assume a probability distribution $L(\alpha|\theta)$ for the parameters α: note that a parametrisation $\alpha = \alpha(\theta)$ is a special case of this. The likelihood function now becomes

$$L(x_{\cdot\cdot}, \alpha|\sigma_\epsilon^2, \sigma_\delta^2, \theta) = L(\alpha|\theta) \times L(x_\cdot|\alpha, \sigma^2) \times L(x_{\cdot\cdot}|x_\cdot, \sigma_\delta^2) = L(x_\cdot, \alpha|\sigma^2, \theta) \times L(x_{\cdot\cdot}|x_\cdot, \sigma_\delta^2) \tag{11}$$

where integration over all α gives

$$L(x_{\cdot\cdot}|\sigma_\epsilon^2, \sigma_\delta^2, \theta) = L(x_\cdot|\sigma^2, \theta) \times L(x_{\cdot\cdot}|x_\cdot, \sigma_\delta^2). \tag{12}$$

This demonstrates that the result also applies to a larger range of models.

In the corresponding empirical Bayesian model for the parameters α, or Bayesian case if there is no parameter θ, the equivalence follows from

$$L(\alpha|x_{\cdot\cdot}, \sigma_\epsilon^2, \sigma_\delta^2, \theta) = \frac{L(x_{\cdot\cdot}, \alpha|\sigma_\epsilon^2, \sigma_\delta^2, \theta)}{L(x_{\cdot\cdot}|\sigma_\epsilon^2, \sigma_\delta^2, \theta)} = \frac{L(x_\cdot, \alpha|\sigma^2, \theta)}{L(x_\cdot|\sigma^2, \theta)} = L(\alpha|x_\cdot, \sigma^2, \theta). \tag{13}$$

3 Conclusion

The derivations demonstrate that when there is a fixed number of technical replicates per subject, estimation of the parameters, α, depend on the per subject averages only, ie X_r rather than all of X_{rk} for $k = 1, \ldots, K$.

The only influence of the per subject replicates is to put a lower bound on the variance of the error term of X_r: ie $\mathrm{Var}X_r = \sigma^2 \geq \sigma_\delta^2/K$ where σ_δ^2 is the within subject variance. This is enforced by `lmer`, but not by `aov`.

The two mixed random effects models presented by Towfic et al. in their Letter to the Editor are directly covered by these results. Their results differ because of misspecification of one model, and an inapproriate approximation in the other which may be asymptotically correct but fails badly when samples sizes are insufficient.

They also present one approach which uses `duplicateCorrelation` to estimate the correlation between replicates, which corresponds to $\rho = \sigma_\epsilon^2/(\sigma_\epsilon^2 + \sigma_\delta^2)$. If the different experiments (ie probes) had been analysed individually, and the correlation had been estimated per experiment, this would also be directly covered by the results presented here. It differs, however, in that it estimates a single correlation across all expreriments, without taking any account of the fact that ρ, σ_ϵ^2, and σ_δ^2 all may vary substantically between experiments, followed by variance reestimation using `eBayes` applied to what are strictly speaking no longer independent experiments.