PersonAlyticsPower Simulation Methodology

Stephen Tueller

2019-04-03

Contents

1 Introduction

2 The Basic ICT Model

The ICT model is a multilevel model (MLM) when $N > 1$, which simplifies to an autoregression moving average (ARMA) model when $N = 1$. Using the notatino from Applied Longitudinal Data Anlysis (ALDA; Singer & Willett, 2003), the standard MLM is

$$(1) \quad y_{ij} = \pi_{0i} + \pi_{1i} TIME_j + \varepsilon_{ij}$$

where y is the outcome variable, i indexes the individual ($i = 1, \ldots, n$), j indexes $TIME$ ($j = 1, \ldots, T$), $TIME_1 = 0$, π_{0i} are random intercepts, π_{1i} are random slopes, and ε_{ij} are errors (see ALDA EQ7.1a). The residuals are assumed

$$(2) \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_\varepsilon)$$

The intercepts and slopes can be decomposed into fixed and random effects as

$$(3) \quad \begin{array}{l} \pi_{0i} = \gamma_{00} + \zeta_{0i} \\ \pi_{1i} = \gamma_{10} + \zeta_{1i} \end{array}$$

where the random effects are distributed

$$(4) \quad \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_0 & \sigma_{01} \\ \sigma_1 0 & \sigma^2_1 \end{bmatrix} \right)$$

Where

$$(5) \quad \Sigma_\zeta = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_1 0 & \sigma_1^2 \end{bmatrix}$$

The basic ICT model adds an effect for $PHASE$ and the $Phase \times Time$ interaction:

$$(6) \quad y_{ij} = \pi_{0i} + \pi_{1i} TIME_j + \pi_2 PHASE_j + \pi_3 (PHASE_j \times TIME_j) + \varepsilon_{ij}$$

Substituting the decomposed fixed and random effects gives us

$$(7) \quad y_{ij} = (\gamma_{00} + \zeta_{0i}) + (\gamma_{10} + \zeta_{1i}) TIME_j + \pi_2 PHASE_j + \pi_3 (PHASE_j \times TIME_j) + \varepsilon_{ij}$$

We can rearrange the random effects and the error terms into what is called the composite residual r_{ij}

$$(8) \quad r_{ij} = \zeta_{0i} + \zeta_{1i} TIME_j + \varepsilon_{ij}$$

Substituting r_{ij} the ICT model yields

$$y_{ij} = \gamma_{00} + \gamma_{10} TIME_j + \pi_2 PHASE_j + \pi_3 (PHASE_j \times TIME_j) + r_{ij}$$

The variance can now be written as

$$(9) \quad V(y_{ij}) = \gamma_{10}^2 V(TIME_j) + \pi_2^2 V(PHASE_j) + \pi_3^2 V(PHASE_j \times TIME_j) + V(r_{ij})$$

The terms $V(TIME_j)$, $V(PHASE_j)$, and $V(PHASE_j \times TIME_j)$ are fixed by the study design and the terms γ_{10}, π_2, and π_3 are the fixed effects. It is important to note that at any given time point j, these six terms are constant and do not contribute to the total variance $V(Y_{ij})$, and that all of the variance at a given time point is due to individual differences in the terms in $V(r_{ij})$ which have the subscript i. We will build on this fact in the next section as it facilitates understanding how data can be simulated in ways that ensure comparibility across simulation conditions (e..g, different sample sizes n and time series lengths T). It is also important to note that in single subject ($n = 1$) studies there is no variance at one time point, only across time points.

We now turn to $V(r_{ij})$ which can be expanded to be

$$(10) \quad V(r_{ij}) = \sigma_0^2 + 2\sigma_{01} + TIME_j^2 \sigma_1^2 + \sigma_\varepsilon^2$$

where $V(\varepsilon_{ij}) = \sigma_\varepsilon^2$. It is important to note that $V(r_{ij})$ is heteroscedastic and changes with j, but that this change is proportional to $TIME_j$. Realizing this proportionality allows us to develop simulation study settings that focus on the remaining terms in $V(r_{ij})$ by developing these settings a $TIME_j = 1$ as described in the next section.

3 Data Simulation for Power Analysis

When we simulate data for power analyses, we want to be able to estimate the power to detect one (or more) set of fixed effects across different numbers of time points (e.g., $T = 10$ vs $T = 20$ time points) and/or across different sample sizes (e.g., $n = 5$ on $n = 10$). Since time deterministically impacts the total variance via its effect on random slopes, we can start with $TIME_j = 1$ in which case the variance will be

$$(11) \quad V(r_{ij}) = \sigma_0^2 + 2\sigma_{01} + \sigma_1^2 + \sigma_\varepsilon^2$$

When simulating data, we are interested in the proportion of the error variance σ_ε^2 which is designated as

$$(12) \quad \pi_\varepsilon = \frac{\sigma_\varepsilon^2}{V(r_{ij})}$$

From this we see that

$$(13) \quad \pi_\zeta = 1 - \frac{\sigma_\varepsilon^2}{V(r_{ij})} = \frac{\sigma_0^2 + 2\sigma_{01} + \sigma_1^2}{V(r_{ij})}$$

and consequently, we only need determine the proportion of error variance and the remaining variance at $TIME_j = 1$ can be alloted to the random effects (or vice versa). Once this has been determined, and since σ_ε^2 is constant across j (i.e., it has no subscrpits), the heteroscedastic variance across time points is due to the deterministic effect of $Time$ on the random slopes. As a result a single power analysis will need as inputs:

1. The proportion of error variance π_ε.
2. The proportion of variance left over for the random effects is then fixed $1 - \pi_\varepsilon = \pi_\zeta$, though π_ζ can be divided among σ_0, σ_{01}, and σ_1 as disired by the user. To simplify this process, PersonAlyticsPower allows the user to specify the covariance matrix for the random intercepts and slopes Σ_ζ, then back-calculates σ_ε^2 from the variance in Σ_ζ and π_ε.
3. The number of participants n.
4. The number of time points T.
5. The phase design (e.g., AB or ABA).
6. The values of the fixed effects γ_{00}, γ_{10}, π_2, and π_3.

With these inputs, PersonAlyticsPower simulates data using the model

$$(14) \quad y_{ij} = (\gamma_{00} + \zeta_{0i}) + (\gamma_{10} + \zeta_{1i})TIME_j + \pi_2 PHASE_j + \pi_3(PHASE_j \times TIME_j) + \varepsilon_{ij}$$

Additionally, the user can specify an ARMA model for the ε_{ij}. Note that the choice of ARMA parameters will not impact the value of σ_ε^2 because after simulating the ε_{ij}, they are rescaled to have the variance as determined in step 2 above.

4 Appendix A - Total Variance Calculations

Converting user specified effect sizes requires the expected total variance for long format data. Since the data are simulated in wide format according to Equation 7, we must find the relationship between the means and variances at each time point and the total variance when the data are stacked, which we will define as

$$(15) \quad \mathbf{y} = vec(y_{ij})$$

TODO: add description of how to get an asymptotic check for this in the software.

4.1 The Expected Means at Each Time Point

$$(16) \quad \mu_j = E[y_{ij}] = E[(\gamma_{00} + \zeta_{0i}) + (\gamma_{10} + \zeta_{1i})TIME_j + \pi_2 PHASE_j + \pi_3(PHASE_j \times TIME_j) + \varepsilon_{ij}]$$

Noting that the expected value of a sum is the sum of expected values, and that $E[\zeta_{0i}] = 0$, $E[\zeta_{1i}] = 0$, $E[\varepsilon_{ij}] = 0$, we can simplify Equation 15 as

$$(16) \quad \mu_j = E[Y_{ij}] = \gamma_{00} + \gamma_{10}TIME_j + \pi_2 PHASE_j + \pi_3(PHASE_j \times TIME_j)$$

4.2 The Expected Mean Across All Time Points

The mean across all time points is simple the mean of the μ_j defined in Equation 15:

$$\mu_{\mathbf{y}} = \frac{1}{T}\sum_{j=1}^{T}\mu_j \quad (17)$$

4.3 The Expected Variances at Each Time Point

The variance of each time point is given in Equeation 10, repeated here for convenience

$$\sigma_j^2 = V(r_{ij}) = \sigma_0^2 + 2\sigma_{01} + TIME_j^2\sigma_1^2 + \sigma_\varepsilon^2 \quad (18)$$

4.4 The Expected Variance Across all Time Points

First we define

$$N = n * T \quad (19)$$

$$\begin{aligned}
V[\mathbf{y}] =& \frac{1}{N-1}\sum_{i,j}^{T}(y_{ij} - \mu_{\mathbf{y}})^2 \\
=& \frac{1}{N-1}\sum_{i,j}^{T}(y_{ij} - \mu_j + \mu_j - \mu_{\mathbf{y}})^2 \\
=& \frac{1}{N-1}\sum_{i,j}^{T}[(y_{ij} - \mu_j)^2 + (\mu_j - \mu_{\mathbf{y}})^2 + 2(y_{ij} - \mu_j)(\mu_j - \mu_{\mathbf{y}})] \\
=& \frac{1}{N-1}\left[\sum_{i,j}^{T}(y_{ij} - \mu_j)^2 + \sum_{j}^{T}(\mu_j - \mu_{\mathbf{y}})^2 + 2\sum_{i,j}^{T}(y_{ij} - \mu_j)\sum_{j}^{T}(\mu_j - \mu_{\mathbf{y}})\right]
\end{aligned}$$

Noting that $\sum_j^T(\mu_j - \mu_{\mathbf{y}}) = 0$, that $1 = \frac{(n-1)}{(n-1)}$, and expanding the summations we have

$$\begin{aligned}
V[\mathbf{y}] =& \frac{1}{N-1}\left[\frac{(n-1)}{(n-1)}\sum_{j}^{T}\sum_{i}^{n}(y_{ij} - \mu_j)^2 + \sum_{i}^{n}\sum_{j}^{T}(\mu_j - \mu_{\mathbf{y}})^2\right] \\
=& \frac{1}{N-1}\left[(n-1)\sum_{j}^{T}\frac{1}{(n-1)}\sum_{i}^{n}(y_{ij} - \mu_j)^2 + n\sum_{j}^{T}(\mu_j - \mu_{\mathbf{y}})^2\right] \quad (20) \\
=& \frac{1}{N-1}\left[(n-1)\sum_{j}^{T}\sigma_j^2 + n\sum_{j}^{T}(\mu_j - \mu_{\mathbf{y}})^2\right]
\end{aligned}$$

where the components of Equation 20 are defined in Equations 16, 17, 18, 19.