Power and Sample Size for Longitudinal Models in R – The longpower Package and Shiny App

---

Replies to Reviewers' Comments

RESPONSE TO EDITOR

Comment: Comments to Author.

Comments. Thank you for submitting your "Power and Sample Size for Longitudinal Models in R - The longpower Package and Shiny App" to the R Journal. The reviews request major revisions as described below.

We would appreciate a revised version and point-by-point response to the reviewers comments within 2 months. Remember, that when responding to the reviewer's comments, your job is to persuade me, the editor, that either you've dealt with the issue, or that it's not relevant. To this end, please produce a single document that includes all the reviewers comments mingled with your responses.

Response. We are grateful for the opportunity to respond to the reviewers' comments.

Point-by-point responses have been provided to address all comments.

RESPONSE TO REVIEWER #3

Comment: General comment

Comments. This manuscript introduces the longpower package and a shiny application for calculating the power and sample size for the linear mixed-effects model and repeated measures mixed models. It provides functions for computing power and sample size for linear models of longitudinal data based on the formula due to Liu and Liang (1997) and Diggle et al (1994). The manuscript looks interesting and may contribute the reader of R Journal, especially with the fact that it is a study that includes a shiny application, however it must be improved with the following minor and major items.

Response. We appreciate the positive feedback on our manuscript. We have attempted to improve the manuscript based on the major and minor suggestion made.

Comment: Minor suggestion #1

Comments. There is a "Summary" section consist nothing. I don't understand why this section is in the manuscript.

Response. Thank you for pointing this out. This has been removed.

Comment: Minor suggestion #2

Comments. The title of the section "Illustration of the power package" should be revised as "... longpower ...".

Response. The correction has been made.

Comment: Minor suggestion #3

Comments. In page 6, there is an overfrow that need fixing.

Response. Corrected.

Comment: Minor suggestion #4

Comments. The information about the data after the author affiliations may be given in the appendix.

Response. We have moved this information to a fresh page.

Comment: Major suggestions: #1

Comments. A discussion part must be added to "Introduction" part about the counter- part packages in the CRAN. The some packages implemented for similar manner are given in the references. In this part, the pros and cons of the longpower package also should be emphasized over the alternatives.

Response. We are grateful to the reviewer for the suggestions and the references. We have reviewed the references and added the following paragraph in the introduction part of the manuscript.

> "Several R packages can be found on CRAN to compute sample size based on mixed-effect models and other specific designs depending on the area of applications. For example, Martin et. al (2011) proposed a simulation-based power calculation and an R package "pamm" for random regression models, a specific form

of mixed-effect model that detects significant variation in individual or group slopes. In their approach, a power analysis was performed to detect a specified level of individual and environmental interactions within evolution and ecology applications. This is achieved by simulating power to detect a given covariance structure. Other simulation-based packages for power analysis are the "SIMR" by Green and MacLeod (2016) for linear and generalized linear mixed model and "clusterPower" by Kleinman (2021) for cluster-randomized and cross-over designs. Schoenfeld (2019) developed a power and sample size package called "LPower" (Diggle, Liang, and Zeger, 1994) to perform power analysis for longitudinal design accounting for attrition and different random effect specification. The approach requires the specification of a design matrix, and the variance-covariance matrix of the repeated measures (Yi and Panzanella, 2002). In pharmacokinetic study designs, Kloprogge and Tarning (2015) developed the "PharmPow" power calculation package for mixed study designs including crossover and parallel designs. Quite recently, other R packages for performing power analysis exist for different designs; for example: "powerMediation" (Qiu, 2021) for mediation effect, mean change for longitudinal study with 2-time points, the slope for simple Poisson regression, etc.; "powerEQTL" (Dong et. al, 2021) for unbalanced one-way ANOVA in a Bulk Tissue and Single-Cell eQTL Analysis; "WebPower" (Zhang, Mai and Yang, 2021) for basic and advanced power correlation, proportion, t-test, one-way ANOVA, two-way ANOVA, linear regression, logistic regression, Poisson regression, mediation analysis, longitudinal data analysis, structural equation modeling, and multilevel modeling; among others. These packages were tailored towards very specific areas of application although the methods can be adapted and utilized for other disciplines."

Comment: Major suggestions: #2

Comments. Users can calculate power using the package or the shiny application developed. However, the study should include a section that includes power comparisons of the calculation methods given in the package. Thus, a broad perspective is presented that users can have an idea about the superiority of the methods against each other. For fair power comparisons, I suggest to use penalized power instead of the classical power of the test (Cavus et al., 2019).

Response. We appreciate the this suggestion from the reviewer. However, we feel this is beyond the scope of the present manuscript. The objective of our manuscript is to present a package and shiny application for previously published

sample size computation methods for longitudinal studies. Many of the formulas covered in the package can be made to provide equivalent results via different but equivalent model parameterizations. An example of this is demonstrated in the package vignette, which shows the equivalence of formulas by Liu and Liang (1997), and Diggle, et al (2002). The former formula covers a broader range of linear model design specifications (e.g. imbalanced groups), but can be made to cover the more specific parameterization of the latter. Generally, the formulas in the package are either exactly equivalent when provided with appropriately equivalent pilot parameters, or they cover different use cases. In either case, performance comparisons are not really appropriate.

Comment: Major suggestions: #3

Comments. To the graphic window given in the shiny app, should be improved to be able to compare the multiple method to support the point mentioned above. In this way, users can compare the power of the methods under the conditions they want.

Response. We refer to our response to the earlier suggestions.

RESPONSE TO REVIEWER #4

Comment: Major suggestions: #1

Comments. The paper discusses implementation of methods to perform sample size calculations for longitudinal studies in the R package "longpower". Unfortunately, the manuscript is full of typos, varying and undefined notation, and other errors that make it very difficult to work out what is going on, or what the software is doing. As far as I can tell the manuscript identifies 5 "methods":

"Diggle 1" sample size for test of difference in population means "Diggle 2" sample size for test of difference in slope parameter with respect to time in a linear model "Liu & Liang" not stated but I believe sample size for test of difference in population means "Edland" sample size for test of difference in slope parameter with respect to time in a linear model (note the cited reference for this method, Edland 2009, does not appear to exist) "MMRM" not stated but I believe sample size for test of difference in population means based on the reference

Response. We thank the reviewer for the good summary. We have attempted to address the typos, notation, and errors that we have identified.

You are correct regardging Diggle 1 and 2, and the Edland method. The method of Liu and Liang computes power for difference in average response as can be inferred from the model specification. We have explicitly stated this hypothesis in the following statement

"Another sample size computation approaches for correlated data is derived by Liu and Liang to detect differences in the average response between two groups."

We have updated the reference of the Edland 2009 to Ard and Edland (2011).

The MMRM approach tests for a difference in mean response at a specific time point. We have captured this in the following statement.

"$\ldots \delta_J$ is the effect size (difference in mean response between the two groups) at the last time point, J $\ldots$"

Comment: Comment: #2

Comments. Method 1 does not appear to be implemented in the software, Method 3 is just a generalisation of Method 1, Method 4 also appears to be a generalisation but of Method 2, Method 5 is similarly a generalisation of Method 1. However, all methods are nominally implemented in the software. Methods 1/2, 3, and 4 give the same answer for sample size based on the software, but they appear to be estimating different things. It does not state anywhere in the methods what the null hypothesis is that the sample size is for. However, given the problems with the notation it is difficult to work out exactly what the software is implementing.

Response. Indeed, method 1 is a special case of method 3 and therefore is implicitly captured by the software.

We have added explicit indications that "... the null hypothesis is that there is no difference between groups."

Since the methods are an aggregation of those in the literature, we struggled with making the notation consistent within the manuscript versus consistent with the original publications. However, we agree that it would be best to strive for internal consistency and adjustments are described below.

Comment: Comment: #3

Comments. The scope of the software is not stated, and in various places it discusses "longitudinal studies", "clinical trials", "Alzheimer's disease trials", and other types of study. However, I believe the software implements sample size calculations for linear mixed models with a dichotomous "treatment" and the references to the Alzheimer's data are both for illustration but also a particular objective of the package to support sample size calculations for studies in this area. I also believe that the methods are limited to the number of individuals in a study and do not include the length of time or number of time periods of observation as an additional output. Some clarification of the scope, use cases, and limitations of the package is required.

Response. We have used the terms; longitudinal studies, clinical trials, and Alzheimer's diseases trials interchangeably. Typically, Alzheimer's disease clinical trials adopt longitudinal or repeated measures designs. Indeed, the sample size method for longitudinal study design applies to other areas of application. In this study, we emphasize the use of this design commonly used in AD clinical trials.

In this application, sample size and power are the main outputs. The number of time periods is considered an input parameter to be specified by the user.

Comment: Comment: #4

Comments. In the statistical section, I believe much notation has been copied verbatim from references and not adjusted to be consistent throughout. A non-exhaustive list of problems in the methods section is: The number of individuals is sometimes given as m *and sometimes as* N β *is sometimes used as a parameter and sometimes as type II error rate interchangeably in the same sections The sample size is sometimes "per group" and sometimes "per arm" although it is not clear what study design these are referring to The parameters* δ *and* Δ *are both used to define an effect size of interest, which varies by model, but is only defined with respect to a model and/or hypothesis test once. The variable* t *is various written with subscripts* t_{ij}, t_i *and* t_j *so it is not clear for whom it is varying, particularly as j is also supposed to represent the time of the observation The groups/arms between which comparisons are made are sometimes referred to as groups* A *and* B *and sometimes with a subscript a and as groups 1 and 2. a is also defined as the attrition rate in the same section.* Σ_a *is given as a covariance term but is actually supposed to be a matrix and I believe the same problem arises for R.*

Response. We have ensured some consistency in the use of the notations. For example, m is used to represent sample size 'per arm', while N is used for the total sample size. We now use J throughout for the number of time-points and Z_P to denote the standard normal quantile associated with the level of power, P. All sample size formula are provided in terms of total sample size N, doubling the prior displays for N per group. Also, we have used δ to represent the effect size comparing mean response between the two groups and Δ to represent the effect size comparing the mean rate of change (slope) between the two groups.

The parameter t_{ij} always denotes the jth observation for individual i. When t_{ij} is assumed to be same value for each subject i, we drop i and use t_j. We do not use t_i anywhere.

We now use groups A and B consistently throughout.

We have changed the attrition rate from a_{aj} to ζ_{aj}.

We state that Σ_a is a variance-covariance matrix and R a correlation matrix.

Comment: Comment: #5

Comments. The authors don't survey any currently existing packages in R or for Shiny that are relevant for the topic of the paper. Some examples might include: https://clusterrcts.shinyapps.io/rshinyapp/ https://strengejacke.github.io/sjstats/reference/samplesize_mixed.html The SIMR package (https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12504)

Response. We have added some literature on existing packages in R following a similar comments by Reviewer #3. We have also now included some shiny packages in the following statement.

> "In terms of software applications, attempts have been made in recent times to implement power and sample size calculations in Shiny applications to facilitate easy usage. For example, Hemming et. al (2020) developed an R Shiny App to conduct power analysis for several cluster designs including parallel, cross-over, and stepped-wedge designs. Schoemann, Boulton, and Shorts (2017) shiny application conducts power analysis for mediation model and Hu and Qu (2021) performs sample size and power calculations for a random coefficient regression model (RCRM) and a two-stage mixed effects model."