Upstrap for estimating power and sample size in complex models

Marta Karas[1], Ciprian M. Crainiceanu*[1]

[1] *Department of Biostatistics, Johns Hopkins University*

ccraini1@jhu.edu

Summary

Power and sample size calculation are major components of statistical analyses. The upstrap resampling method introduced by Crainiceanu and Crainiceanu (2018) was proposed as a general solution to this problem but has not been assessed in numerical experiments. We evaluate the power and sample size estimation properties of the upstrap for data sets that are larger or smaller than the original data set. We also expand the scope of upstrap and propose a solution to estimate the power to detect: (1) an effect size observed in the data; and (2) an effect size chosen by a researcher. Simulations include the following scenarios: (a) one- and two-sample t-tests; (b) linear regression with both Gaussian and binary outcomes; (c) multilevel mixed effects models with both Gaussian and binary outcomes. We illustrate the approach using a reanalysis of a cluster randomized trial of malaria transmission. The accompanying software and data are publicly available.

Key words: bootstrap; upstrap; power estimation; sample size calculation; sampling.

*To whom correspondence should be addressed.

1. INTRODUCTION

We consider the context when a test statistic is used on a set of preliminary data. We evaluate the power and sample size estimation properties of the test statistics using the upstrap (Crainiceanu and Crainiceanu, 2018), a general-purpose resampling technique. In summary, the upstrap starts with a sample (observed data) and re-samples with replacement either fewer or more samples than in the original data set. To be even more precise, here is the difference between bootstrap and upstrap.

```
x  <- 1:5                                # Data set
Bx <- sample(x, replace = TRUE)          # Bootstrap resample
r  <- 2                                  # Upstrap sample size multiplier
Nu <- r * length(x)                      # Upstrap sample size
Ux <- sample(x, Nu, replace = TRUE)      # Upstrap resample
```

The estimator of a parameter of interest can be obtained from an upstrap resample for any model, procedure, or algorithm that provides estimators of a target parameter. A collection of these estimators can be obtained by generating multiple upstrap resamples. Such a collection provides a sample from the distribution of the estimator of a parameter of interest when the sample size is either larger or smaller than the original one.

We also expand the scope of the upstrap approach by proposing a solution to estimate the power to detect: (1) an effect size observed in the preliminary data; and (2) effect size chosen by a user. We present "ready to use" code examples to estimate power and sample size to identify a statistically significant effect for an α-level test to achieve the power $1 - \beta$. We evaluate the accuracy of the upstrap to estimate power in an extensive simulation study. We consider both the case when analytic solutions are available (e.g., one-sample t-test) and when they are not (e.g., testing for fixed effect in a multilevel mixed-effects model). We illustrate the approach in a reanalysis of results from a cluster randomized trial (CRT) of the impact of interventions on malaria transmission (Bousema *and others*, 2016), where the treatment was assigned at the cluster (group) level.

1.1 *Related work*

1.1.1 *Nonparametric boostrap for multilevel data.* The nonparametric bootstrap for multilevel data is one of the accepted analytic methods for quantifying variability in mixed effects modeling. For example, Davison and Hinkley (1997) considered hierarchical data with two levels; they proposed to first randomly sample groups with replacement and then to sample within the groups without replacement ("case bootstrap"). Ren *and others* (2010) extended the concept to hierarchical unbalanced data with two and more levels. They showed that sampling units at the highest level and then sampling within these units without replacement is a good strategy. Thai *and others* (2013) evaluated a range of bootstrap methods for estimating the standard error of fixed effect parameters in linear mixed-effect models. They provided evidence that the case bootstrap outperforms the case bootstrap coupled with global/individual residual bootstrap, random effects bootstrap coupled with global/individual residual bootstrap, global/individual residual bootstrap and parametric bootstrap. Huang (2018) used the case bootstrap of two-level unbalanced data for standard error estimation in a context that mimics CRTs. They proposed a modification of separating the "treatment" and "control" clusters into two parts within which the step 1 of the case bootstrap is done independently.

1.1.2 *Power analysis and sample size estimation for multilevel data.* The power analysis and sample size calculations for multilevel data are active areas of research. *A priori* power analysis aims at estimating the sample size needed to achieve specific power $1 - \beta$ for a given significance level α and hypothesized effect size. *Post hoc* power analysis is conducted using available data to estimate the power of a test and the sample size needed for a future study (Hox, 2002).

A class of theoretical results for estimating power in specific multilevel data setups has been proposed; see, for example, references in Lane and Hennes (2018). These results are based on the variance inflation factor correction approach (Kish, 1965, 1987). The approach uses assumptions

about the intra-class correlation coefficient, ρ, or assumes a particular study design (e.g., that the data are balanced).

Another class of approaches is based on simulations. Hox (2002) proposed a general approach that requires specifying a population model for the data, simulate data from the assumed model, and estimate the power via Monte Carlo simulations. There exists software designed for *post-hoc* power analysis of a test given a multilevel data and model (Arnold *and others*, 2011). For example, the `SIMR` R package (Green and MacLeod, 2016; Green *and others*, 2019) is a versatile R software for power analysis of GLMM using simulation. It allows to estimate a power curve for a range of sample sizes by simulating new values for the response variable and refitting the model to the larger data set.

To the best of our knowledge, this is the first time upstrap resampling techniques have been proposed for power analysis and sample size estimation in GLMMs. The remainder of the paper is organized as follows. Section 2 describes the upstrap approach for estimating power and sample size. Section 3 presents the simulation study. Section 4 describes the applications of our approach to the reanalysis of CRT data on the impact of hotspot-targeted interventions on malaria transmission (Bousema *and others*, 2016). We close with a discussion in Section 5. Appendix A demonstrates upstrap method for estimating power in a series of examples with "read, adapt and use" R code. Code for all analyses included in this work is available on GitHub at `https://git.io/J0TH1`.

2. Methods

2.1 *Notation*

Denote by $\mathbf{x}$ a vector of N independent realizations of a random variable X with an unknown distribution F. We assume that $\mathbf{x}$ consists of either scalar-valued, x_i, or vector-valued, $\mathbf{x}_i$, observations. Observations are assumed to be independent across observational units, $i = 1, \ldots, N$.

Throughout the manuscript, the term "sample size" denotes the number of independent units in $\mathbf{x}$. When observations are scalar and independent, the sample size is the number of observations. When observations are vectors with a multilevel dependence structure, the sample size denotes the number of unique units at the highest level of the data structure.

2.2 *The upstrap*

The upstrap method (Crainiceanu and Crainiceanu, 2018) samples with replacement from $\mathbf{x}$ either more or fewer samples than the original sample size, N. For a fraction $r \in (0, \infty)$, the upstrap creates B samples, $\mathbf{x}^{*1}_{rN}, \ldots, \mathbf{x}^{*B}_{rN}$, by sampling with replacement from $\mathbf{x}$ with a sample size rN. The estimators $\hat{\theta}(\mathbf{x}^b_{rN})$, $b = 1, \ldots, B$ can be obtained for any model, procedure, or algorithm that provides estimators of a target parameter θ. The collection of these estimators provides a sample from the distribution of the estimator $\hat{\theta}$ for a dataset of size r times the sample size N of the original data.

2.3 *The upstrap solution to the sample size estimation problem*

We consider the following problem: given $\mathbf{x}$ of sample size N from unknown distribution F, a specific null and alternative hypothesis, a test statistic, and significance level α of the test, estimate a sample size M required to achieve probability $1 - \beta$ of rejecting the null hypothesis when the null is true (i.e., achieve power $1 - \beta$).

2.3.1 *Target effect size as observed in preliminary data* $\mathbf{x}$. In case the target effect size is the one observed in sample $\mathbf{x}$, we propose the following solution:

Step 1: Consider a grid of target sample sizes, M_k, for $k = 1, \ldots, K$;

Step 2: Use the upstrap method to estimate the power for a sample of size M_k, for $k = 1, \ldots, K$:

(a) Generate B resamples of size M_k by sampling with replacement from $\mathbf{x}$, denote $\mathbf{x}^{*1}_{M_k}, \ldots, \mathbf{x}^{*B}_{M_k}$;

(b) Perform hypothesis test on each resample $\mathbf{x}^{*b}_{M_k}$, for $b = 1, \ldots, B$;

(c) Define $\hat{\theta}(\mathbf{x}^{*b}_{M})$ to be equal to 1 if the null hypothesis was rejected for data $\mathbf{x}^{*b}_{M_k}$ and equal to 0 otherwise, for $b = 1, \ldots, B$;

(d) Define $\overline{\hat{\theta}}(\mathbf{x}_{M_k}) = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}(\mathbf{x}^{*b}_{M_k})$ the proportion of resamples when the null hypothesis was rejected as the estimate of power.

Step 3: Identify the smallest sample size M_k for which the desired power $1 - \beta$ is achieved.

Denote by $\hat{\theta}(\mathbf{X}_M)$ the sampling distribution of the statistic $\hat{\theta}$. For a given $\mathbf{x}$ of sample size N, the upstrap-generated values $\hat{\theta}(\mathbf{x}^{*1}_{M}), \ldots, \hat{\theta}(\mathbf{x}^{*B}_{M})$ provide a sample from the distribution of the estimator $\hat{\theta}(\mathbf{X}_M)$; their sample mean $\overline{\hat{\theta}}(\mathbf{x}_M)$ is an estimate of the mean of $\hat{\theta}(\mathbf{X}_M)$ distribution, and the proposed estimate of power.

2.3.2 *Arbitrary target effect size.* A common practical case is when the target effect size is larger or smaller than the one observed in the preliminary data. To address such cases, we propose the following approach:

Step 0: Update the observed outcome values to ensure that the effect size in this updated data is the $\big[$target effect size $\times$ target covariate value$\big]$ instead of the $\big[$observed effect size $\times$ target covariate value$\big]$.

For GLM/GLMM with an identity link function, Step 0 is fairly straightforward to implement. Specifically, we propose to update each observation in sample $\mathbf{x}$ by adding to the outcome the value of [(target effect size $-$ observed effect size) $\times$ target covariate value]. The observed effect size is computed once for $\mathbf{x}$ at the beginning of the procedure, and then the updated sample $\mathbf{x}$ is used in Steps 1-3. See R code examples in the Appendix: A.1, A.2, A.3, A.5.

For other distributions in the GLM/GLMM family (e.g., Gamma, Poisson or Bernoulli), we propose the following modification. For each upstrap resample from Step 2a: (1) use the resample to fit the assumed model; (2) update the obtained model fit by setting the coefficient of interest to [target effect size]; (3) use the updated fit to simulate new outcome values for the resample; and (4) use the updated resample in Step 2b. See R code examples in Appendix A.6. The idea of simulating a new outcome based on an updated R model fit object is shared with the `SIMR` approach. For some samples with independent observations, the procedure can be simplified by simulating new outcome values from the distribution instead of R model fit. Here, the distribution parameters are based on the linear predictor estimated for original sample, $\mathbf{x}$, after updating the coefficient of interest to [target effect size]. See R code examples in Appendix A.4.

2.4 *Sampling with replacement from* $\mathbf{x}$

When the observed sample $\mathbf{x}$ consists of independent scalar observations, the sampling with replacement from $\mathbf{x}$ is done by sampling the observations. When observations are vectors with a multilevel dependence structure we recommend the "case sampling", i.e. sample units at the highest level with replacement, and then sample within these units (a) without replacement at lower level – if the highest level effect is of interest, (b) without replacement at lower level – if lower level effect is of interest.

In Appendix A, we demonstrate the upstrap method for estimating power in a series of examples accompanied by "read, adapt and use" R code. The problems include: one-sample t-test, two-sample t-test, testing significance of a coefficient in linear regression, generalized linear regression with binary outcome, multilevel linear regression and generalized multilevel linear regression with binary outcome.

3. SIMULATIONS

An extensive simulation study was used to assess the performance of the upstrap to estimate the power of an α-level test for a sample size smaller or larger than in an observed sample. The problems included estimation of power to detect both arbitrary effect size and effect size observed in the data. The R code for the simulation study is provided on the project's GitHub repository (sub-directory URL: `https://git.io/JsiiA`).

3.1 *Setup of the problems*

Table 1 summarizes the setup for six problems considered in the simulations. The first part considers problems when analytic solutions are available (problems 1-2). The second part of the simulations considers problems when analytic solutions are not available (problems 3-6). The first table column specifies the generative model: (1) linear model with intercept only; (2) linear model with one dichotomous variable; (3) linear model with one dichotomous variable and two other variables (dichotomous and numeric); (4) generalized linear model assuming a logit link function with one dichotomous variable and two other variables (dichotomous and numeric); (5) multilevel linear model with one dichotomous variable; and (6) multilevel generalized linear model assuming logit link function with one dichotomous variable. The specifications of model parameters and generative distributions are provided.

The second table column specifies an effect of interest, β, in each simulation problem. In all simulations, a two-sided test was used to test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ at significance level $\alpha = 0.05$. The third table column describes the assumed observed sample size and target sample sizes. The fourth table column specifies the target effect sizes.

Table 1: Summary of simulation setup across six different problems.

Data-generating model	Effect	Observed/ target sample size	Target effect size

1	$Y_i = \beta_0 + \epsilon_i$ • $i = 1, \ldots, N$ • $\beta_0 = 0.3$ • $\epsilon_i \sim_{\text{iid}} \mathcal{N}(0, 1)$	β_0	50/ $20, 50, \ldots, 200$	$0.3, 0.4,$ observed
2	$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$ • $i = 1, \ldots, 2N$ • $\beta_0 = 0,\ \beta_1 = 0.3$ • $[X_{1i}] = \text{rep}([0, 1], \text{times} = N)$ • $\epsilon_i \sim_{\text{iid}} \mathcal{N}(0, 1)$	β_1	50/ $20, 50, \ldots, 200$	$0.3, 0.4,$ observed
3	$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$ • $i = 1, \ldots, 2N$ • $\beta_0 = 0,\ \beta_1 = 0.5,\ \beta_2 = 1,\ \beta_3 = -1$ • $[X_{1i}] = \text{rep}([0, 1], \text{times} = N)$ • $X_{2i} \sim_{\text{iid}} \text{Bern}(0.5)$ • $X_{3i} \sim_{\text{iid}} \text{Unif}([0, 1])$ • $\epsilon_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$	β_1	50/ $20, 50, \ldots, 200$	$0.5, 1,$ observed
4	$\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$ • $i = 1, \ldots, 2N$ • $\beta_0 = 0,\ \beta_1 = 0.5,\ \beta_2 = 1,\ \beta_3 = -1$ • $[X_{1i}] = \text{rep}([0, 1], \text{times} = N)$ • $X_{2i} \sim_{\text{iid}} \text{Bern}(0.5)$ • $X_{3i} \sim_{\text{iid}} \text{Unif}([0, 1])$ • $Y_i \sim \text{Bern}(\pi_i)$	β_1	50/ $20, 50, \ldots, 200$	$0.5, 1,$ observed
5	$Y_{ij} = b_{0i} + \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \epsilon_{ij}$ • $i = 1, \ldots, 2N$ • $j = 1, \ldots, J;\ J = 3$ • $\beta_0 = 0,\ \beta_1 = 0.5,\ \beta_2 = 1,\ \beta_3 = -1$ • $[X_{1i}] = \text{rep}([0, 1], \text{times} = N)$ • $X_{2i} \sim_{\text{iid}} \text{Bern}(0.5)$ • $X_{3i} \sim_{\text{iid}} \text{Unif}([0, 1])$ • $b_{0i} \sim_{\text{iid}} \mathcal{N}(0, 1)$ • $\epsilon_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$	β_1	50/ $20, 50, \ldots, 200$	$0.5, 1,$ observed
6	$\text{logit}(\pi_{ij}) = b_{0i} + \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij}$ • $i = 1, \ldots, 2N$ • $j = 1, \ldots, J;\ J = 3$ • $\beta_0 = 0,\ \beta_1 = 0.5,\ \beta_2 = 1,\ \beta_3 = -1$ • $[X_{1i}] = \text{rep}([0, 1], \text{times} = N)$ • $X_{2i} \sim_{\text{iid}} \text{Bern}(0.5)$ • $X_{3i} \sim_{\text{iid}} \text{Unif}([0, 1])$ • $b_{0i} \sim_{\text{iid}} \mathcal{N}(0, 1)$ • $Y_{ij} \sim \text{Bern}(\pi_{ij})$	β_1	50/ $20, 50, \ldots, 200$	$0.5, 1,$ observed

In generative models for problems 5 and 6, the covariates X_{1ij}, X_{2ij} and X_{3ij} take the same

value within a subject, i.e. $X_{pij} = X_{pi}$ for $p = 1, 2, 3$ and for any i and j.

3.2 *Comparator methods*

For each of the six simulation problems, the upstrap approach for power estimation was compared with an existing comparator method.

3.2.1 *Simulation problems 1-2.* For simulation problems 1-2, the existing analytic solution for power estimation in one-sample and two-sample t-test was used. Specifically:

- For cases when the target effect size was the effect size observed in the sample $\mathbf{x}$, the comparator approach was to use `power.t.test()` R function with function arguments: (a) `delta`; (b) `sd`, set to: (1) $\mathbf{x}$ sample mean/difference in means; and (2) $\mathbf{x}$ sample standard deviation, respectively;

- For cases when the target effect size was set as an arbitrary value, the comparator approach was to use `power.t.test()` R function with function arguments: (a) `delta`; (b) `sd`, set to: (1) the arbitrary effect size value, (2) $\mathbf{x}$ sample standard deviation, respectively.

3.2.2 *Simulation problems 3-6.* For simulation problems 3-6, the `SIMR` R package (Green and MacLeod, 2016; Green *and others*, 2019) was used as a comparator. As discussed in Green and MacLeod (2016), "in `SIMR`, power is calculated by repeating the following three steps: (i) simulate new values for the response variable using the model provided; (ii) refit the model to the simulated response; (iii) apply a statistical test to the simulated fit." To estimate power for sample size larger than that of the observed sample, the `extend()` function can be used to add (upsample) rows to a model data frame. Then, `powerCurve()` function is provided to estimate power at a range of sample sizes smaller than the "extended" model, including the sizes smaller than in the original sample.

To the best of our understanding, adding the rows in `extend()` is done via upsampling and is performed once before the power estimation procedure starts. Then, in the `powerCurve()` function, the rows removing is done by subsetting the "extended" model data frame to the target sample size deterministic (by keeping the first observations in the sample). Therefore, the same set of observations will be used in the `SIMR` power estimation procedure across all iterations when simulating new values of the response. This is different from the upstrap approach, where data observations are resampled with replacement multiple times within the procedure. Furthermore,

- For cases when the target effect size is the effect size observed in the sample $\mathbf{x}$, `SIMR` uses a model fitted to $\mathbf{x}$ to simulate new responses.

- For cases when the target effect size is set to an arbitrary value, `SIMR` updates a model fitted to $\mathbf{x}$ by modifying its parameters according to the target effect value, and uses this updated model to simulate new responses.

Finally, "`SIMR` is designed to work with any linear mixed model (LMM) or GLMM that can be fit with either `lmer` or `glmer` from `lme4`."

To ensure a fair comparison between `SIMR` and upstrap several precautions have been used. First, both the upstrap and `SIMR` used the same `lmer` function for model coefficient testing in simulation problems 3-6. Second, the number of upstrap resamples (upstrap method's parameter) was set to be equal to the number of simulations in `SIMR` (`SIMR` method's parameter). Third, since there is no feature to maintain balance of covariates explicitly with the `SIMR` `powerCurve()` function, data samples were simulated to ensure that the balance of the dichotomous covariate of interest is preserved when `SIMR` removes samples. Fourth, the authors of the `SIMR` package and the Green and MacLeod (2016) paper were contacted to ensure that `SIMR` is correctly used. One of the authors reported that "has not seen any obvious problems" with how we employed `SIMR`.

3.3 *True power estimation*

True power estimates were obtained by setting the generative model's coefficient equal to the target effect size, and repeatedly ($10,000$ times) sampling independent observations from the target sample size. True power estimate was defined as the proportion of repetitions where null hypothesis was rejected and was denoted as $\hat{\text{pwr}}^{\text{true}}$.

3.4 *Simulation procedure*

The following procedure was performed to estimate and compare power at a range of target sample sizes for upstrap and a comparator approach, for each of the six simulation problems separately.

1. Consider R independent experiment repetitions. For each r-th repetition, $r = 1, \ldots, R$, do:

   (a) Simulate an observed sample size $\mathbf{x}_r$.

   (b) Use the upstrap to estimate power at a range of target sample sizes, $\hat{\text{pwr}}_r^{\text{ups}}$.

   (c) Use the comparator approach to estimate power for a range of target sample sizes, $\hat{\text{pwr}}_r^{\text{cmp}}$,

   (d) Compute percentage error between the upstrap and true power for a range of target sample sizes, $\text{PE}_r^{\text{ups,true}} = 100 \cdot \left[\hat{\text{pwr}}_r^{\text{ups}} - \hat{\text{pwr}}^{\text{true}}\right] / \hat{\text{pwr}}^{\text{true}}$.

   (e) Compute the percent error between comparator and true power for a range of target sample sizes, $\text{PE}_r^{\text{cmp,true}} = 100 \cdot \left[\hat{\text{pwr}}_r^{\text{cmp}} - \hat{\text{pwr}}^{\text{true}}\right] / \hat{\text{pwr}}^{\text{true}}$,

   (f) Compute the percent error between upstrap and comparator for a range of target sample sizes, $\text{PE}_r^{\text{ups,cmp}} = 100 \cdot \left[\hat{\text{pwr}}_r^{\text{ups}} - \hat{\text{pwr}}_r^{\text{cmp}}\right] / \hat{\text{pwr}}_r^{\text{cmp}}$.

   The r-th experiment repetition estimates: $\hat{\text{pwr}}_r^{\text{ups}}$, $\hat{\text{pwr}}_r^{\text{cmp}}$, $\text{PE}_r^{\text{ups,true}}$, $\text{PE}_r^{\text{cmp,true}}$ and $\text{PE}_r^{\text{ups,cmp}}$ are vectors of the length equal to the number of target sample sizes considered.

2. Aggregate estimates from R independent experiment repetitions by computing mean, standard deviation, median, 25th percentile and 75th percentile.

The number of R independent experiment repetitions was set to 1000 for each simulation problem 1-6. The number of upstrap resamples (upstrap method's parameter) and the number of simulations in `SIMR` (`SIMR` method's parameter) were both set to 1000 in each case.

3.5 *Simulation results*

3.5.1 *Simulation problems 1 and 2.* In simulation problems 1-2, the upstrap was compared with `power.t.test()` comparator approach in one-sample and two-sample t-test power estimation, respectively. Figure 1 displays the power trajectories obtained with the upstrap (left plot) and comparator (right plot) approaches in 15 simulations as in problem 1. The case when the target effect size is the effect size observed in the sample is considered. Repetition identifiers are indicated as red numbers. Red lines point at power estimates for a target sample size equal 60 (chosen arbitrarily for presentation purposes). Figure 1 demonstrates that the upstrap power estimates are very similar to those obtained from the `power.t.test()`. The difference is that they are slightly noisier, but this could be addressed by increasing the number of upstrap resamples.

Figure 2 shows power estimates (y-axis) at a range of target sample sizes (x-axis) in simulation problem 1 (one-sample t-test – upper plots panel) and problem 2 (two-sample t-test – lower plots panel). Coloured points represent median and colored lines span from the 25th to the 75th percentiles of the power estimates using the upstrap (purple) and `power.t.test()` (orange). A black circle point indicates the true power estimate. Results are showed separately for three target effect size cases: fixed at 0.3, fixed at 0.4 and the effect size observed in the sample. In each case, the orange and purple points and lines show very close agreement, indicating that the results obtained based on the upstrap are essentially identical to those obtained based on `power.t.test()`. In the case of power estimation for a fixed effect size (first and second plots

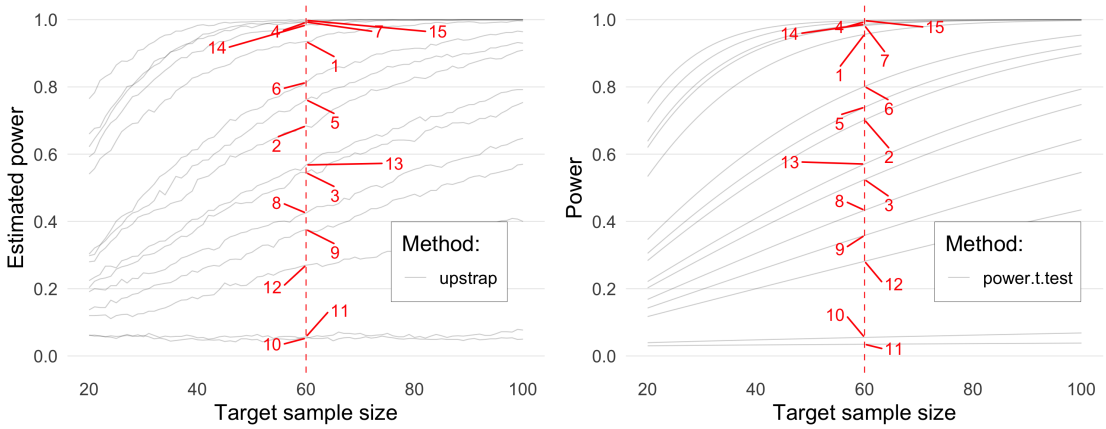

Fig. 1. Power estimation for 15 repetitions from the simulation experiment 1. Left panel: upstrap, right panel: `power.t.test()` R comparator approach. Red numbers denote a specific experiment repetition.

column), the lines spanning from the 25th to the 75th percentiles are very narrow and mostly overlap with the median point, indicating very little variability of power estimates across the simulated samples. Moreover, the medians of power estimates are closely overlapping with true power estimates for both the upstrap and `power.t.test()`; the alignment is slightly better for the latter approach. The variability of power estimates for observed effect size (third plots column) is larger for both upstrap and `power.t.test()`, primarily as a result of the variability of the sample mean across the simulated samples.

Table 4 in Appendix B.1 shows a summary of power estimates and percentage errors (PEs) across simulation problems 1-6. The first and second table columns specify target effect size and target sample size for which the power was estimated; all experiments were done assuming the observed sample size equal 50 (see the parameters setup in Table 1 in Section 3.1). Table columns third through seventh show aggregates (mean and standard deviation) of power estimates and PEs across $R = 1000$ independent experiment repetitions: $\hat{\text{pwr}}^{\text{ups}}$ – upstrap power, $\hat{\text{pwr}}^{\text{cmp}}$ –

Fig. 2. Power estimates (y-axis) at a range of target sample sizes (x-axis) in simulation problem 1 (one-sample t-test – upper plots panel) and problem 2 (two-sample t-test – lower plots panel). Colored points represent median and colored lines span from the 25th to the 75th percentiles of the power estimates from upstrap (purple color) and `power.t.test()` (orange color). Black circle points represent the true power estimate. Results are shown separately for three target effect size cases.

comparator approach power, $\mathrm{PE}^{\mathrm{ups,cmp}}$ – PE between upstrap and comparator, $\mathrm{PE}^{\mathrm{ups,true}}$ – PE between upstrap and true power, $\mathrm{PE}^{\mathrm{cmp,true}}$ – PE between comparator and true power. Overall, the upstrap demonstrated excellent agreement with the `power.t.test()` in terms of estimating power in simulation problems 1 and 2, across target sample sizes and across target effect sizes. For cases when the target effect size was set at a fixed value, the mean PE between the upstrap and comparator ranged between 0% and 1.52%. Both the upstrap and `power.t.test()` tended to overestimate the power when compared to the true power estimates for small target sample

sizes (20, 50; mean PE mostly between 1% and 7%) and showed great agreement for larger target sample sizes ($\geqslant 80$; absolute value of the mean PE mostly below 1%). For cases when the target effect size was the one observed, $\mathrm{PE}^{\mathrm{ups,cmp}}$ ranged between 3.57% to 11.71%, with higher target sample size corresponding to lower $\mathrm{PE}^{\mathrm{ups,cmp}}$.

3.5.2 *Simulation problems 3-6.* In simulation problems 3-6, the upstrap was compared with the `SIMR` for testing the significance of a dichotomous variable coefficient in a regression model: linear regression with both Gaussian and binary outcomes (problems 3-4), multilevel mixed effects models with both Gaussian and binary outcomes (problems 5-6). Figure 3 displays the power estimates for simulation problems 3-6. Coloured points represent median and colored lines span from the 25th to the 75th percentiles of the power estimates using the upstrap (purple color) and `SIMR` (orange color). Black circle points represent the true power estimates. Results are showed separately for three target effect size cases: fixed at 0.5, fixed at 1 and the effect size observed in the sample. Based on visual inspection, the upstrap yields very similar power estimation results to `SIMR`. In most cases, the upstrap/`SIMR` mean power estimates agree closely with the true power estimate; the largest discrepancies were present for simulation problem 4 (GLM with binary outcome) and problem 6 (GLMM with binary outcome) for the smallest target effect size considered (left plots column, 2nd and 4th plots rows).

Based on the summary presented in Table 4 in Appendix B.1, the upstrap and `SIMR` approaches yielded very similar results across experiment repetitions, with absolute value of mean $\mathrm{PE}^{\mathrm{ups,cmp}}$ below 2.6% across all experiment setups, and below 1% for most of them. For simulation problem 3 (LM) and problem 5 (LMM), the mean PE between upstrap and true power estimates was mostly below 1% and had the highest values (1%-4.13%) for small target effect size (0.5) and small target sample size (20, 50) cases. For simulation problem 4 (GLM with binary outcome), both the upstrap and `SIMR` tended to underestimate the power (mean $\mathrm{PE}^{\mathrm{ups,true}}$ be-

tween -6.66% and -0.17%). For simulation problem 6 (GLMM with binary outcome), both the upstrap and `SIMR` tended to overestimate power (mean $\mathrm{PE}^{\mathrm{ups,true}}$ between -0.14% and 9.92%).

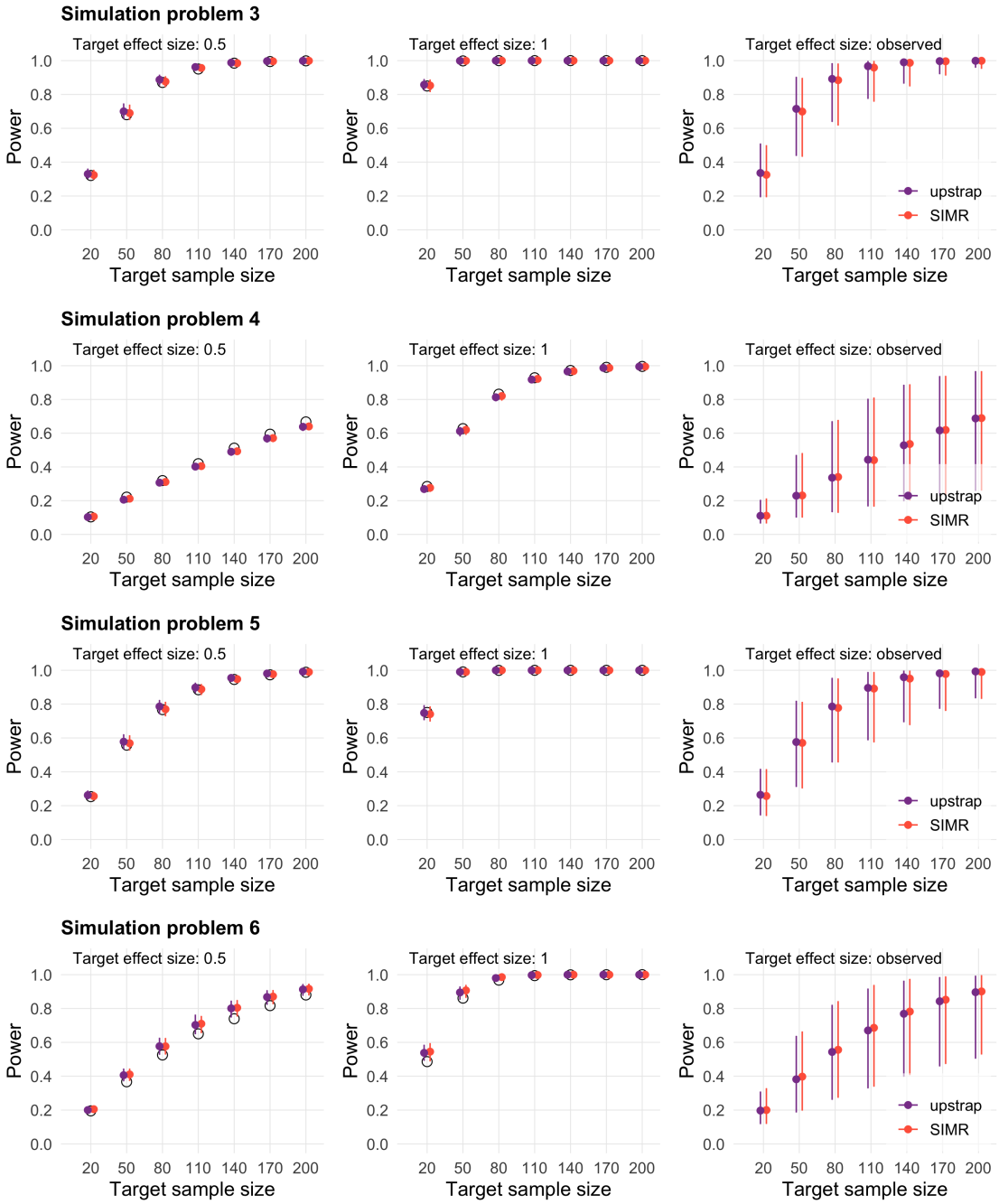

Fig. 3. Power estimates (y-axis) at a range of target sample sizes (x-axis) in simulation problems 3-6: testing significance of coefficient in LM (problem 3), GLM with binary outcome (problem 4), LMM (problem 5), GLMM with binary outcome (problem 6), respectively. Values are showed for upstrap (purple color) and `SIMR` comparator (orange color) approaches.

4. Application to a real data problem

The upstrap approach was used to estimate power of α-level tests in our reanalysis of the data from Bousema *and others* (2016). Data are obtained from a CRT on the impact of hotspot-targeted interventions on malaria transmission. Data are available online from the Dryad Digital Repository (`https://bit.ly/3wrahch`; Bousema *and others* (2017)). Our analysis R code is available on GitHub (sub-directory URL: `https://git.io/JsiP8`).

4.1 *Study description*

Bousema *and others* (2016) presented methods and findings from a cluster-randomized controlled trial conducted in western Kenya in 2012. In short, 10 malaria hotspots – areas of serological and parasitological evidence of elevated levels of malaria transmission relative to their surrounding – were identified based on a survey. The hotspots were randomized into $N_1 = 5$ intervention arm hotspots ($2,082$ individuals) and $N_2 = 5$ placebo arm hotspots ($2,468$ individuals). The hotspot-targeted intervention consisted of a combination of "larviciding, distribution of long-lasting insecticide-treated nets, indoor residual spraying, and focal mass drug administration". The placebo hotspots followed Kenyan national policy of malaria control. The parasite prevalence was measured at three time points: baseline, 8 weeks, and 16 weeks post-intervention, across three sections defined based on the distance from the hotspot boundary: hotspot itself, evaluation zone 1 (1–249 meters), evaluation zone 2 (250–500 meters).

The intervention effects were evaluated by comparing cluster-level (hotspot-level) parasite prevalence between control and intervention clusters across each of three evaluation zones. To make the comparison, first, an individual-level logistic regression model was fit with parasite status as an outcome and adjustment variables (baseline prevalence, age group, sex, altitude group, living in a house with open eaves) as covariates. Second, the model was used to estimate the expected outcome at the participant-level, and then to derive expected prevalence (average of

participant-level outcomes) in each cluster. Third, cluster-level residuals (10 values) were calculated as the difference between the observed and expected prevalence. Fourth, the residuals of the control and intervention clusters were compared using a t-test. Results from t-tests conducted on raw prevalence values without covariate adjustment were also reported. The paper concluded that the impact of interventions on parasite prevalence was "modest, transient, and restricted to the targeted hotspot areas"; it also stated that the trial "was not powered to detect subtle effects of hotspot-targeted interventions". Data at participant-level resolution were deposited in the Dryad Digital Repository (Bousema *and others*, 2017).

Below, we reanalyze the Bousema *and others* (2016) data to replicate a subset of the main results. We then use the upstrap approach to estimate power trajectory of the relevant t-tests. Next, we perform an additional new analysis in which we use participant-level data in generalized estimating equations (GEE), and also perform an upstrap power trajectory estimation.

4.2 *Replication of the study analysis results*

A subset of the Bousema *and others* (2016) main results regarding the intervention effect measured 8 weeks post-intervention was considered. The assumed model was $Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$, where $i = 1, \ldots, 2N$, $N = 5$ is the number of clusters in each intervention and control arm, Y_i is the i-th cluster prevalence (unadjusted analysis) or i-th cluster residual of the observed minus expected prevalence (adjusted analysis), X_{1i} is the indicator whether the i-th cluster is in the intervention ($X_{1i} = 1$) or control ($X_{1i} = 0$) arm, $\epsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2)$. A two-sample t-test was used to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, separately within each of three evaluation zones and for both unadjusted and adjusted analyses. In our replication analysis we relied on the R code provided by Naudet *and others* (2018), who replicated the results from Bousema *and others* (2016). Their R code is publicly available at `https://osf.io/7ghfa/`. Our replication R code is available at `https://git.io/J3zSW`.

Table 2 is an adaptation of "Table 3" from Bousema *and others* (2016) comparing estimated parasite prevalence in the intervention and control clusters at two time points (baseline, 8 weeks post-intervention), across three hotspot-specific areas (hotspot itself, evaluation zone 1-249 meters, evaluation zone 250-500 meters). Estimates and p-values from statistical tests comparing the intervention and control groups are provided based on: (1) study's original paper (Bousema *and others* (2016); 3rd column); (2) previous reanalysis (Naudet *and others* (2018), using the hard-coded values found in their code; 4th column); and (3) our reanalysis (5th column). In our covariate-adjusted reanalysis, the t-test rejected the null hypothesis of no treatment effect in hotspot area at the $\alpha = 0.05$ level (p-value=0.047). The hypothesis was not rejected for the other two evaluation zones (p-value equal 0.339 and 0.253, respectively).

4.3 *Estimating power of a t-test using the upstrap*

The upstrap approach was used to estimate power of three statistical tests corresponding to rows: 7, 14, and 21 in Table 2 (adjusted analysis t-tests). A grid $\{5, 6, \ldots, 30\}$ of target sample sizes (number of clusters in each of the two arm groups) was considered, and $B = 10,000$ upstrap replications were used. Top plots panel in Figure 4 displays estimated trajectories of t-test power corresponding to three areas: hotspot, evaluation zone 1-249 meters, and evaluation zone 250-500 meters. Black dots indicate the estimated power (y-axis) for a particular target sample size (x-axis). The horizontal solid blue line represents 0.8 power value; the vertical dashed blue line represents the smallest sample size for which the estimated power is at least 0.8. The number of clusters needed for achieving a power of 0.8 were estimated to be 9, 19, and 14 for the three area-specific tests, respectively. The results indicate that for the $N = 5$ group sample size available in the Bousema *and others* (2016) paper's study, the power of the test was 0.35, 0.14, and 0.19 for the three distance-based areas, respectively.

Table 2. Adaptation of "Table 3" from Bousema *and others* (2016) comparing estimated parasite prevalence in intervention and control clusters, across the two time points (baseline, 8 weeks post-intervention), across three hotspot-specific areas (hotspot itself, evaluation zone 1-249 meters, evaluation zone 250-500 meters). Estimates and p-values from statistical tests comparing intervention and control groups are shown from: study's original paper (Bousema *and others* (2016); 3rd column), previous work reanalysis (Naudet *and others* (2018); 4th column), and this work reanalysis (5th column).

	Time point	Group or value	Bousema *and others* (2016)	Naudet *and others* (2018)	This paper's reanalysis
			nPCR Parasite Prevalence: **Hotspot**		
1	Baseline	Intervention	20.1		21.4
2	Baseline	Control	18.6		18.6
3	8 weeks	Intervention	9.2		9.2
4	8 weeks	Control	19.4		19.4
5	8 weeks	Difference	10.2	10.2	10.2
6	8 weeks	p-value	0.075	0.095	0.095
7	8 weeks	Adj. p-value	0.024	0.047	0.047
			nPCR Parasite Prevalence: **Eval. zone 1-249 m**		
8	Baseline	Intervention	13.2		13.2
9	Baseline	Control	11.7		11.7
10	8 weeks	Intervention	12.3		12.3
11	8 weeks	Control	15.9		15.9
12	8 weeks	Difference	3.6	3.6	3.6
13	8 weeks	p-value	0.219	0.2394	0.2394
14	8 weeks	Adj. p-value	0.216	0.339	0.339
			nPCR Parasite Prevalence: **Eval. zone 250-500 m**		
15	Baseline	Intervention	13.5		13.5
16	Baseline	Control	16.9		16.9
17	8 weeks	Intervention	9.1		9.1
18	8 weeks	Control	12.9		13.1
19	8 weeks	Difference	3.8	4	4
20	8 weeks	p-value	0.199	0.1698	0.1698
21	8 weeks	Adj. p-value	0.187	0.2533	0.2533

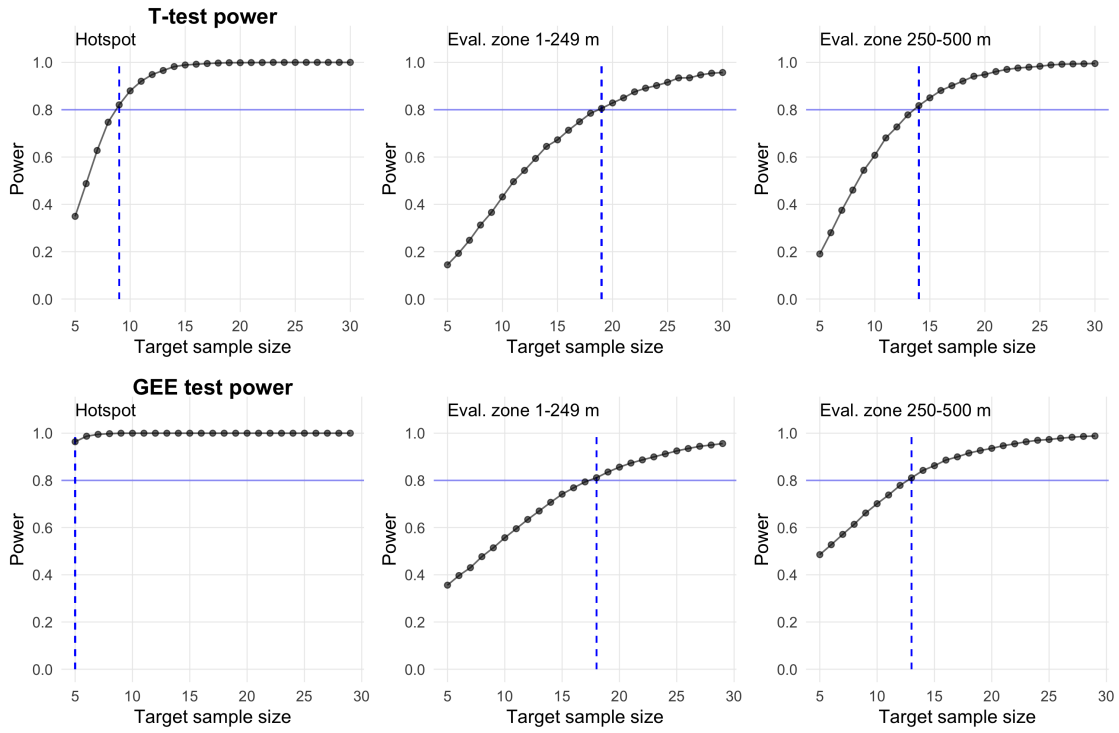

Fig. 4. Top plots panel: estimated t-test power in the reanalysis of Bousema *and others* (2016) paper's data (adjusted t-test for a difference in prevalence between intervention and control groups at 8 weeks post-intervention). Bottom plots panel: estimated power in the additional analysis of Bousema *and others* (2016) paper's data (test of β_1 coefficient from GEE model). The plots correspond to distance-based areas: hotspot itself, evaluation zone 1-249 meters, and evaluation zone 250-500 meters. Black dots show estimated power (y-axis) for a group sample size (x-axis). The horizontal solid blue line represents 0.8 power value; the vertical dashed blue line represents the smallest sample size for which the estimated power is at least 0.8.

4.4 *Estimating power of a GEE-estimated effect test using the upstrap*

Additional analysis of the Bousema *and others* (2016) paper's data was performed using participant-level data. The effect of the intervention was estimated via GEE, where the outcome is the

recorded disease status ($Y_{ij} = 1$ if positive and $Y_{ij} = 0$ otherwise for j-th study participant in the i-th cluster). The linear predictor was modelled as a linear combination of intercept, cluster arm assignment X_{1i} (equal 1 if intervention and 0 if control) and participant's predicted adjusted disease status X_{ij2} (equal 1 if positive and 0 otherwise; see Section 4.1 for details). The models were estimated using the `geeglm` function from `geepack` R package assuming a logit link function and exchangeable correlation structure within each cluster. In this model, the β_1 coefficient of X_{i1} represents a population-averaged effect of the cluster-level intervention (when adjusted by the predicted status). We focused on the 8 weeks post-intervention effect was of interest, and three hotspot distance-based areas (hotspot itself, evaluation zone 1-249 meters, evaluation zone 250-500 meters) were considered separately.

Table 3 summarizes the estimated 8 weeks post-intervention effect across three hotspot distance-based areas. In each case, the $\hat{\beta}_1$ point estimate is negative; the statistical hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ is rejected at significance level $\alpha = 0.05$ for the hotspot (p-value < 0.0001), but not for either of the two evaluation zones (p-value equal 0.199 and 0.076, respectively).

The upstrap was used to estimate power of three statistical tests $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ from the new analysis. Similarly as earlier, a grid $\{5, \ldots, 30\}$ of target sample sizes (size of each of the two arm groups) was considered, and $B = 10,000$ upstrap replications were used. Top plots panel in Figure 4 shows estimated trajectories of statistical power. The smallest group sample sizes for for which the estimated power is at least 0.8 were estimated as $\leqslant 5$, 18, and 13 for the three area-specific tests, respectively: hotspot itself, evaluation zone 1-249 meters, evaluation zone 250-500 meters. For the second and third area-specific tests, the upstrap results using participant-level data therefore closely aligned with the ones for tests using cluster-level aggregate data in original analysis; for the test regarding the hotspot area, a smaller sample size was expected to be needed in our additional analysis ($\leqslant 5$ versus 9).

Table 3. Estimated post-intervention effect across three hotspot distance-based areas using participant-level data. The β_1 coefficient represents a population-averaged effect of the cluster-level intervention (when adjusted by the participant's predicted status).

	Time point	Model fit parameter	This paper's additional analysis
			Hotspot
1	8 weeks	$\hat{\beta}_1$; $\exp(\hat{\beta}_1)$	-0.768; 0.464
2	8 weeks	$\hat{\beta}_1$ se	0.175
3	8 weeks	$\hat{\beta}_1$ p-value	< 0.0001
			Eval. zone 1-249 m
4	8 weeks	$\hat{\beta}_1$; $\exp(\hat{\beta}_1)$	-0.237; 0.789
5	8 weeks	$\hat{\beta}_1$ se	0.185
6	8 weeks	$\hat{\beta}_1$ p-value	0.199
			Eval. zone 250-500 m
7	8 weeks	$\hat{\beta}_1$; $\exp(\hat{\beta}_1)$	-0.409; 0.664
8	8 weeks	$\hat{\beta}_1$ se	0.231
9	8 weeks	$\hat{\beta}_1$ p-value	0.076

5. Discussion

Power estimation and sample size calculation are major components of statistical analyses. We proposed the upstrap-based methods to estimate power to detect both the effect size observed in the data and an effect size chosen by a researcher. The most important contribution of this paper is to evaluate the accuracy of the upstrap method for power estimation in simple and complex simulation scenarios.

First, our results indicate that for one- and two-sample t-test, the upstrap performs essentially identical to the well-established power estimation solutions. Second, in complex scenarios, the upstrap performed similarly the existing method from `SIMR` R package. Both approaches demonstrated very high agreement with the true power estimates. Thus, the upstrap provides an easy to understand and implement solution for complex model scenarios. Third, the upstrap

method is "read-and-use", as it can be implemented by any analyst who is familiar with software allowing to (a) resample data, (b) run the statistical test of interest. Fourth, the R code for our simulations and real data application example is provided.

The upstrap is particularly useful when the model might be misspecified, as re-sampling of independent units is done under weak modeling assumptions. For example, up-sampling and down-sampling independent units in a longitudinal data analysis will preserve the true correlation structures in the data. The `SIMR` approach requires a true model for the correlation structure. Moreover, the upstrap can be used to preserve covariate class proportions and distribution. There is currently no feature to maintain balance or impose specific covariate class proportions in the `SIMR` R package.

Alas, our work is not without limitations. First, the upstrap computational time can be substantial. Steps that can speed up the process include employing parallel and/or cluster-based computing, as well as algorithmic improvements; for example, see our implementation of "rolling p-value" for one-sample t-test (`https://git.io/JGqmh`) and two-sample t-test (`https://git.io/JGqOe`). Second, this work has not covered a wide range of models, including longitudinal data or functional data regression models; further work is required to evaluate upstrap for power and sample size estimation in these settings.

Acknowledgments

Conflict of Interest: None declared.

References

Arnold, B. F., Hogan, D. R., Colford, J. M. and Hubbard, A. E. (2011). Simulation methods to estimate design power: an overview for applied research. *BMC Medical Research Methodology* **11**, 94.

Bousema, Teun, Stresman, Gillian, Baidjoe, Amrish Y., Bradley, John, Knight, Philip, Stone, William, Osoti, Victor, Makori, Euniah, Owaga, Chrispin, Odongo, Wycliffe, China, Pauline, Shagari, Shehu, Doumbo, Ogobara K., Sauerwein, Robert W., Kariuki, Simon, Drakeley, Chris, Stevenson, Jennifer *and others*. (2016, 04). The Impact of Hotspot-Targeted Interventions on Malaria Transmission in Rachuonyo South District in the Western Kenyan Highlands: A Cluster-Randomized Controlled Trial. *PLOS Medicine* **13**(4), 1–25.

Bousema, Teun, Stresman, Gillian, Baidjoe, Amrish Y., Bradley, John, Knight, Philip, Stone, William, Osoti, Victor, Makori, Euniah, Owaga, Chrispin, Odongo, Wycliffe, China, Pauline, Shagari, Shehu, Doumbo, Ogobara K., Sauerwein, Robert W., Kariuki, Simon, Drakeley, Chris, Stevenson, Jennifer *and others*. (2017). Data from: The impact of hotspot-targeted interventions on malaria transmission in Rachuonyo south district in the western Kenyan highlands: a cluster-randomized controlled trial.

Crainiceanu, Ciprian M and Crainiceanu, Adina. (2018, 09). The upstrap. *Biostatistics* **21**(2), e164–e166.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Green, Peter, MacLeod, Catriona and Alday, Phillip. (2019). *simr: Power Analysis for Generalised Linear Mixed Models by Simulation*. R package version 1.0.5.

Green, Peter and MacLeod, Catriona J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution* **7**(4), 493–498.

Hox, Joop. (2002). Multilevel analysis techniques and applications. Chapter 10. Lawrence Erlbaum Associates Publishers, pp. 173–196.

Huang, Francis L. (2018). Using Cluster Bootstrapping to Analyze Nested Data With a Few Clusters. *Educational and Psychological Measurement* **78**(2), 297–318.

Kish, Leslie. (1965). *Survey Sampling*. Wiley.

Kish, Leslie. (1987). *Statistical Design for Research*. Wiley.

Lane, Sean P. and Hennes, Erin P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships* **35**(1), 7–31.

Naudet, Florian, Sakarovitch, Charlotte, Janiaud, Perrine, Cristea, Ioana, Fanelli, Daniele, Moher, David and Ioannidis, John P A. (2018). Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine. *BMJ* **360**.

Ren, Shiquan, Lai, Hong, Tong, Wenjing, Aminzadeh, Mostafa, Hou, Xuezhang and Lai, Shenghan. (2010). Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics* **37**(9), 1487–1498.

Thai, Hoai-Thu, Mentré, France, Holford, Nicholas H.G., Veyrat-Follet, Christine and Comets, Emmanuelle. (2013). A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical Statistics* **12**(3), 129–140.

APPENDIX

A. Methods code examples

We demonstrate the upstrap method for estimating power in a series of examples accompanied by "read, adapt and use" R code. The problems include: one-sample t-test, two-sample t-test, testing significance of a coefficient in: linear regression, generalized linear regression with binary outcome, multilevel linear regression and generalized multilevel linear regression with binary outcome. All statistical tests assume significance level $\alpha = 0.05$.

A.1 *One-sample t-test*

Consider a random sample from normal distribution. We are interested in estimating the power of the one-sample t-test.

We define simulation parameters and simulate a sample of size $N = 30$.

```
# simulation parameters
N <- 30
R_boot <- 1000
# simulate sample
set.seed(1)
x <- rnorm(n = N, mean = 0.3, sd = 1)
```

We get the observed effect size.

```
mean(x)
# [1] 0.3824582
```

We estimate test the power via upstrap.

- **Case**: target sample size $M = N = 30$, target effect size as observed in the sample ("observed power" case).

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  x_rr <- sample(x, replace = TRUE)
  out[rr] <- (t.test(x_rr)$p.value < 0.05)
}
```

```
mean(out)
# [1] 0.593
```

- **Case**: target sample size $M = 40$, target effect size as observed in the sample.

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  x_rr <- sample(x, size = 40, replace = TRUE)
  out[rr] <- (t.test(x_rr)$p.value < 0.05)
}
mean(out)
# [1] 0.716
```

- **Case**: target sample size $M = 40$, target effect size set to 0.5.

```
# update sample x to represent the target effect size
x_upd <- x + (0.5 - mean(x)) * 1
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  x_rr <- sample(x_upd, size = 40, replace = TRUE)
  out[rr] <- (t.test(x_rr)$p.value < 0.05)
}
mean(out)
# [1] 0.917
```

A.2 *Two-sample t-test*

Consider two random samples (not paired) from normal distribution. We are interested in estimating power of two-sample t-test.

We define simulation parameters and simulate two samples, each of size $N = 30$.

```
# simulation parameters
N <- 30
R_boot <- 1000
# simulate sample
set.seed(1)
x1 <- rnorm(N, mean = 0, sd = 1)
x2 <- rnorm(N, mean = 0.3, sd = 1)
```

We get the observed effect size.

```
mean(x2) - mean(x1)
# [1] 0.3503164
```

We estimate test power via upstrap.

- **Case**: target sample size $M = N = 30$, target effect size as observed in the sample ("observed power" case).

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  x1_rr <- sample(x1, replace = TRUE)
  x2_rr <- sample(x2, replace = TRUE)
  pval_rr <- t.test(x1_rr, x2_rr, paired = FALSE, var.equal = TRUE)$p.value
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.349
```

- **Case**: target sample size $M = 40$, target effect size as observed in the sample.

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  x1_rr <- sample(x1, size = 40, replace = TRUE)
  x2_rr <- sample(x2, size = 40, replace = TRUE)
  pval_rr <- t.test(x1_rr, x2_rr, paired = FALSE, var.equal = TRUE)$p.value
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.44
```

- **Case**: target sample size $M = 40$, target effect size set to 0.5.

```
# update sample to represent the target effect size
x1_upd <- x1
x2_upd <- x2 + (0.5 - (mean(x2) - mean(x1))) * 1
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  x1_rr <- sample(x1_upd, size = 40, replace = TRUE)
  x2_rr <- sample(x2_upd, size = 40, replace = TRUE)
  pval_rr <- t.test(x1_rr, x2_rr, paired = FALSE, var.equal = TRUE)$p.value
  out[rr] <- (pval_rr < 0.05)
```

```
}
mean(out)
# [1] 0.752
```

A.3 *Testing for significance of a coefficient in the linear model (LM)*

Consider a random sample with $N = 40$ independent observations (e.g., 40 subjects, 1 observation per subject). Assume a continuous outcome Y and three covariates: dichotomous X_1, dichotomous X_2, continuous X_3. We are interested in estimating power of test for significance of the coefficient β_1 in linear model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$, where $i = 1, \ldots, N$ and $\varepsilon_i \sim_{\text{iid}} N(0, \sigma^2)$. We define simulation parameters and simulate a sample of size $N = 40$.

```
# simulation parameters
N       <- 40
coef_x0 <- 0
coef_x1 <- 0.6
coef_x2 <- 0.3
coef_x3 <- -0.1
sigma2  <- 1
R_boot  <- 1000
# simulate sample
set.seed(1)
subjid_i  <- 1 : N   # subject ID unique in data set
x1_i      <- rbinom(n = N, size = 1, prob = 0.5)
x2_i      <- rbinom(n = N, size = 1, prob = 0.5)
x3_i      <- runif(n = N, min = 18, max = 100)
eps_i     <- rnorm(N, sd = sqrt(sigma2))
y_i       <- coef_x0 + (coef_x1 * x1_i) + (coef_x2 * x2_i) +  (coef_x3 * x3_i) + eps_i
dat       <- data.frame(y = y_i, x1 = x1_i, x2 = x2_i, x3 = x3_i, subjid = subjid_i)
head(dat, 3)
# y           x1 x2       x3 subjid
# 1 -2.662590  0  1 53.64208      1
# 2 -7.381860  0  1 76.42620      2
# 3 -3.590214  1  1 50.79954      3
```

We get the observed effect size.

```
fit  <- lm(y ~ x1 + x2 + x3, data = dat)
coef(fit)["x1"]
# 0.3949275
```

We estimate the test power via the upstrap.

- **Case**: target sample size $M = N = 40$, target effect size as observed in the sample ("observed power" case).

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  dat_rr_idx <- sample(1 : nrow(dat), replace = TRUE)
  dat_rr  <- dat[dat_rr_idx, ]
  fit_rr  <- lm(y ~ x1 + x2 + x3, data = dat_rr)
  pval_rr <- summary(fit_rr)$coef["x1", 4]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.286
```

- **Case**: target sample size $M = 80$, target effect size as observed in the sample.

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  dat_rr_idx <- sample(1 : nrow(dat), size = 80, replace = TRUE)
  dat_rr  <- dat[dat_rr_idx, ]
  fit_rr  <- lm(y ~ x1 + x2 + x3, data = dat_rr)
  pval_rr <- summary(fit_rr)$coef["x1", 4]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.511
```

- **Case**: target sample size $M = 80$, target effect size set to 0.5.

```
# update the outcome in the sample to represent the target effect size
dat_upd <- dat
dat_upd$y <- dat_upd$y + (0.5 - coef(fit)["x1"]) * dat_upd$x1
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  dat_rr_idx <- sample(1 : nrow(dat_upd), size = 80, replace = TRUE)
  dat_rr  <- dat_upd[dat_rr_idx, ]
  fit_rr  <- lm(y ~ x1 + x2 + x3, data = dat_rr)
  pval_rr <- summary(fit_rr)$coef["x1", 4]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.686
```

A.4 *Testing for significance of coefficient in generalized linear model (GLM)*

Consider a random sample with $N = 80$ independent observations (e.g. 80 subjects, 1 observation per subject). Assume binary outcome Y and three covariates: dichotomous X_1, dichotomous X_2, continuous X_3. We are interested in estimating power of testing the coefficient β_1 in generalized linear model of a form $g(\pi_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$, where $i = 1 \ldots, N$, $E(Y_i) = \pi_i$ and $g(\cdot)$ is a logit link function.

We define simulation parameters. We simulate a sample of size $N = 80$.

```
# simulation parameters
N       <- 80
coef_x0 <- -0.2
coef_x1 <- 0.5
coef_x2 <- 0.1
coef_x3 <- -0.01
sigma2  <- 1
R_boot  <- 1000
# simulate sample
set.seed(1)
subjid_i <- 1 : N
x1_i     <- rep(c(0, 1), times = N/2)
x2_i     <- rbinom(n = N, size = 1, prob = 0.5)
x3_i     <- runif(n = N, min = 18, max = 100)
XB_i     <- coef_x0 + (coef_x1 * x1_i) + (coef_x2 * x2_i) + (coef_x3 * x3_i)
p_i      <- 1/(1 + exp(-XB_i))
y_i      <- rbinom(n = length(p_i), size = 1, prob = p_i)
dat      <- data.frame(y = y_i, x1 = x1_i, x2 = x2_i, x3 = x3_i, subjid = subjid_i)
head(dat, 3)
#   y x1 x2       x3 subjid
# 1 0  0  0 53.64208      1
# 2 0  1  0 76.42620      2
# 3 0  0  1 50.79954      3
```

We get the observed effect size.

```
fit  <- glm(y ~ x1 + x2 + x3, data = dat,
            family = binomial(link = "logit"))
coef(fit)["x1"]
# 0.7354476
```

We estimate test power via upstrap.

- **Case**: target sample size $M = N = 80$, target effect size as observed in the sample ("ob-

served power" case).

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  dat_rr_idx <- sample(1 : nrow(dat), replace = TRUE)
  dat_rr  <- dat[dat_rr_idx, ]
  fit_rr  <- glm(y ~ x1 + x2 + x3, data = dat_rr,
                 family = binomial(link = "logit"))
  pval_rr <- summary(fit_rr)$coef[2, 4]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.356
```

- **Case**: target sample size $M = 120$, target effect size as observed in the sample.

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  dat_rr_idx <- sample(1 : nrow(dat), size = 120, replace = TRUE)
  dat_rr  <- dat[dat_rr_idx, ]
  fit_rr  <- glm(y ~ x1 + x2 + x3, data = dat_rr,
                 family = binomial(link = "logit"))
  pval_rr <- summary(fit_rr)$coef[2, 4]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.441
```

- **Case**: target sample size $M = 120$, target effect size set to 1.0.

```
dat_upd <- dat
# define link term value, assuming target effect size
dat_upd$link_orig <- predict(fit, type = "link")
dat_upd$link_upd  <- dat_upd$link_orig  + (1 - coef(fit)["x1"]) * dat_upd$x1
dat_upd$res_upd   <- 1/(1 + exp(-dat_upd$link_upd))
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  dat_rr_idx <- sample(1 : nrow(dat_upd), size = 80, replace = TRUE)
  dat_rr   <- dat_upd[dat_rr_idx, ]
  # simulate response on resampled data, assuming target effect size
  dat_rr$y <- rbinom(n = nrow(dat_rr), size = 1, prob = dat_rr$res_upd)
  fit_rr   <- glm(y ~ x1 + x2 + x3, data = dat_rr,
                  family = binomial(link = "logit"))
  pval_rr  <- summary(fit_rr)$coef[2, 4]
```

```
  out[rr]  <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.565
```

A.5 *Testing for significance of a coefficient in the linear mixed model (LMM)*

Consider a random sample with $N = 40$ subjects, $n_i = 3$ observations per each subject. Assume continuous outcome Y and two covariates: dichotomous X_1, continuous X_2. We are interested in estimating power of test for significance of the subject-level (group-level) coefficient β_1 in linear mixed model $Y_{ij} = b_{0i} + \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_{ij}$, where $i = 1, \ldots, N$, $j = 1, 2, 3$, $b_{0i} \sim_{\text{iid}} \mathcal{N}(0, \tau^2)$, and $\varepsilon_i \sim_{\text{iid}} N(0, \sigma^2)$.

We define simulation parameters and simulate a sample of size $N = 40$.

```
# simulation parameters
N       <- 40
ni      <- 3
coef_x1 <- 1.2
coef_x2 <- 0.01
tau2    <- 1
sigma2  <- 1
R_boot  <- 1000
# simulate sample
set.seed(1)
subjid_i      <- 1:N                        # subject ID unique in data set
x1_i          <- rep(c(0, 1), times = N/2) # indicator of a "trial arm"
x2_i          <- runif(n = N, min = 18, max = 100)
b0_i          <- rnorm(n = N, mean = 0, sd = tau2)
eps_ij        <- rnorm(n = N * ni, mean = 0, sd = sigma2)
subjid_ij     <- rep(subjid_i, each = ni)
x1_ij         <- rep(x1_i, each = ni)
x2_ij         <- rep(x2_i, each = ni)
b0_ij         <- rep(b0_i, each = ni)
y_ij          <- b0_ij + coef_x1 * x1_ij + eps_ij
dat           <- data.frame(y = y_ij, x1 = x1_ij, x2 = x2_ij, subjid = subjid_ij)
head(dat, 4)
#           y x1       x2 subjid
# 1 3.3205951  0 39.77171      1
# 2 0.8797374  0 39.77171      1
# 3 1.6087167  0 39.77171      1
# 4 1.8101385  1 48.51416      2
```

We get the observed effect size.

```
library(lme4)
library(lmerTest)
fit  <- lmer(y ~ x1 + x2 + (1 | subjid), data = dat)
fixef(fit)["x1"]
# 0.6890758
```

We estimate test power via upstrap.

- **Case**: target sample size $M = N = 40$, target effect size as observed in the sample ("observed power" case).

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  print(rr)
  dat_rr_subj_id <- sample(unique(dat$subjid), replace = TRUE)
  dat_rr  <- lapply(dat_rr_subj_id, function(subjid_tmp) dat[dat$subjid == subjid_tmp,
  dat_rr  <- do.call("rbind", dat_rr)
  # make new subject ID so as to treat subjects resampled >1 as unique ones
  dat_rr$subjid <- rep(1 : N, each = ni)
  fit_rr  <- lmer(y ~ x1 + x2 + (1 | subjid), data = dat_rr)
  pval_rr <- summary(fit_rr)$coef["x1", 5]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.536
```

- **Case**: target sample size $M = 60$, target effect size as observed in the sample.

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  print(rr)
  dat_rr_subj_id <- sample(unique(dat$subjid), size = 60, replace = TRUE)
  dat_rr  <- lapply(dat_rr_subj_id, function(subjid_tmp) dat[dat$subjid == subjid_tmp,
  dat_rr  <- do.call("rbind", dat_rr)
  # make new subject ID so as to treat subjects resampled >1 as unique ones
  dat_rr$subjid <- rep(1 : 60, each = ni)
  fit_rr  <- lmer(y ~ x1 + x2 + (1 | subjid), data = dat_rr)
  pval_rr <- summary(fit_rr)$coef["x1", 5]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.714
```

- **Case**: target sample size $M = 60$, target effect size set to 1.2.

```
# update the outcome in the sample to represent the target effect size
dat_upd <- dat
dat_upd$y <- dat_upd$y + (1.2 - fixef(fit)["x1"]) * dat_upd$x1
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  print(rr)
  dat_rr_subj_id <- sample(unique(dat_upd$subjid), size = 60, replace = TRUE)
  dat_rr  <- lapply(dat_rr_subj_id, function(subjid_tmp) dat_upd[dat_upd$subjid == subji
  dat_rr  <- do.call("rbind", dat_rr)
  # make new subject ID so as to treat subjects resampled >1 as unique ones
  dat_rr$subjid <- rep(1 : 60, each = ni)
  fit_rr  <- lmer(y ~ x1 + x2 + (1 | subjid), data = dat_rr)
  pval_rr <- summary(fit_rr)$coef["x1", 5]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.994
```

A.6 *Testing for significance of coefficient in generalized linear mixed model (GLMM)*

Consider a random sample with $N = 80$ subjects, $n_i = 3$ observations per each subject. Assume binary outcome Y and two covariates: dichotomous X_1, continuous X_2. We are interested in estimating power of test for significance of the subject-level (group-level) coefficient β_1 in linear mixed model $g(\pi_{ij}) = b_{0i} + \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$, where $i = 1, \ldots, N$, $j = 1, 2, 3$, $b_{0i} \sim_{\text{iid}} \mathcal{N}(0, \tau^2)$, $E(Y_{ij}) = \pi_{ij}$ and $g(\cdot)$ is a logit link function.

We define simulation parameters. We simulate a sample of size $N = 80$.

```
# simulation parameters
N       <- 80
ni      <- 3
coef_x1 <- 0.8
coef_x2 <- 0.01
tau2    <- 1
sigma2  <- 1
R_boot  <- 1000
# simulate sample
set.seed(1)
subjid_i  <- 1:N                        # subject ID unique in data set
x1_i      <- rep(c(0, 1), times = N/2) # indicator of e.g. a "trial arm"
```

```
x2_i      <- runif(n = N, min = 18, max = 100)
b0_i      <- rnorm(n = N, mean = 0, sd = tau2)
subjid_ij <- rep(subjid_i, each = ni)
x1_ij     <- rep(x1_i, each = ni)
x2_ij     <- rep(x2_i, each = ni)
b0_ij     <- rep(b0_i, each = ni)
XB_ij     <- b0_ij + coef_x1 * x1_ij
p_ij      <- 1/(1 + exp(-XB_ij))
y_ij      <- rbinom(n = length(p_ij), size = 1, prob = p_ij)
dat       <- data.frame(y = y_ij, x1 = x1_ij, subjid = subjid_ij)
dat       <- data.frame(y = y_ij, x1 = x1_ij, x2 = x2_ij, subjid = subjid_ij)
head(dat, 4)
#   y x1       x2 subjid
# 1 0  0 39.77171      1
# 2 1  0 39.77171      1
# 3 1  0 39.77171      1
# 4 1  1 48.51416      2
```

We get the observed effect size.

```
library(lme4)
library(lmerTest)
fit  <- glmer(y ~ x1 + x2 + (1 | subjid), data = dat, family = binomial)
fixef(fit)["x1"]
# 0.8574488
```

We estimate test power via upstrap.

- **Case**: target sample size $M = N = 80$, target effect size as observed in the sample ("observed power" case).

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  print(rr)
  dat_rr_subj_id <- sample(unique(dat$subjid), replace = TRUE)
  dat_rr  <- lapply(dat_rr_subj_id, function(subjid_tmp) dat[dat$subjid == subjid_tmp,
  dat_rr  <- do.call("rbind", dat_rr)
  # make new subject ID so as to treat subjects resampled >1 as unique ones
  dat_rr$subjid <- rep(1 : N, each = ni)
  fit_rr  <- glmer(y ~ x1 + x2 + (1 | subjid), data = dat_rr, family = binomial)
  pval_rr <- summary(fit_rr)$coef["x1", 4]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# 0.652
```

- **Case**: target sample size $M = 100$, target effect size as observed in the sample.

```
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  print(rr)
  dat_rr_subj_id <- sample(unique(dat$subjid), size = 100, replace = TRUE)
  dat_rr  <- lapply(dat_rr_subj_id, function(subjid_tmp) dat[dat$subjid == subjid_tmp, ]
  dat_rr  <- do.call("rbind", dat_rr)
  # make new subject ID so as to treat subjects resampled >1 as unique ones
  dat_rr$subjid <- rep(1 : 100, each = ni)
  fit_rr  <- glmer(y ~ x1 + x2 + (1 | subjid), data = dat_rr, family = binomial)
  pval_rr <- summary(fit_rr)$coef["x1", 4]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.794
```

- **Case**: target sample size $M = 100$, target effect size set to 1.2.

```
set.seed(1)
out <- rep(NA, R_boot)
for (rr in 1 : R_boot){
  print(rr)
  dat_rr_subj_id <- sample(unique(dat$subjid), size = 100, replace = TRUE)
  dat_rr  <- lapply(dat_rr_subj_id, function(subjid_tmp) dat[dat$subjid == subjid_tmp, ]
  dat_rr  <- do.call("rbind", dat_rr)
  # make new subject ID so as to treat subjects resampled >1 as unique ones
  dat_rr$subjid <- rep(1 : 100, each = ni)
  # fit model, update coefficient for the size effect of interest
  fit_rr_sim_y  <- glmer(y ~ x1 + x2 + (1 | subjid), data = dat_rr, family = binomial)
  fixef(fit_rr_sim_y)["x1"] <- 1.2
  # simulate new outcome
  dat_rr$y <- simulate(fit_rr_sim_y, nsim = 1, newdata = dat_rr)[[1]]
  # fit "final" model
  fit_rr  <- glmer(y ~ x1 + x2 + (1 | subjid), data = dat_rr, family = binomial)
  pval_rr <- summary(fit_rr)$coef["x1", 4]
  out[rr] <- (pval_rr < 0.05)
}
mean(out)
# [1] 0.957
```

B. Results

B.1 *Simulations results*

Table 4: Summary of power estimates and percentage errors (PEs) across simulation problems 1-6. The first and second table columns specify target effect size and target sample size for which power was estimated; all experiments were done assuming the observed sample size equal 50. Table columns third to seventh show aggregates (mean and standard deviation) of power estimates and PE across $R = 1000$ independent experiment repetitions: $\hat{\text{pwr}}^{\text{ups}}$ – upstrap power, $\hat{\text{pwr}}^{\text{cmp}}$ – comparator approach power, $\text{PE}^{\text{ups,cmp}}$ – percentage error between upstrap and comparator, $\text{PE}^{\text{ups,true}}$ – percentage error between upstrap and true power, $\text{PE}^{\text{cmp,true}}$ – percentage error between comparator and true power.

Simulation problem 1

	Target effect size	*Target sample size*	$\hat{\text{pwr}}^{\text{ups}}$ *mean (sd)*	$\hat{\text{pwr}}^{\text{cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,true}}$ *mean (sd)*	$\text{PE}^{\text{cmp,true}}$ *mean (sd)*
1	0.3	20	0.26 (0.05)	0.25 (0.04)	1.07 (9.36)	5.95 (20.24)	4.82 (17.62)
2	0.3	50	0.57 (0.08)	0.56 (0.08)	1.33 (2.87)	3.63 (15.23)	2.29 (14.92)
3	0.3	80	0.77 (0.08)	0.76 (0.08)	1.11 (1.96)	1.71 (10.44)	0.62 (10.36)
4	0.3	110	0.88 (0.06)	0.87 (0.06)	0.73 (1.50)	0.63 (6.70)	-0.08 (6.74)
5	0.3	140	0.94 (0.04)	0.94 (0.04)	0.45 (1.05)	-0.40 (4.11)	-0.84 (4.18)
6	0.3	170	0.97 (0.02)	0.97 (0.02)	0.26 (0.72)	-0.25 (2.47)	-0.50 (2.54)
7	0.3	200	0.99 (0.01)	0.98 (0.01)	0.14 (0.47)	-0.02 (1.44)	-0.16 (1.50)
8	0.4	20	0.41 (0.07)	0.41 (0.07)	0.72 (5.33)	4.05 (18.15)	3.28 (17.04)
9	0.4	50	0.80 (0.07)	0.79 (0.07)	1.06 (2.03)	0.80 (9.38)	-0.23 (9.31)
10	0.4	80	0.94 (0.04)	0.94 (0.04)	0.45 (1.18)	-0.34 (4.12)	-0.78 (4.14)
11	0.4	110	0.98 (0.02)	0.98 (0.02)	0.15 (0.56)	-0.27 (1.65)	-0.42 (1.67)
12	0.4	140	1.00 (0.01)	0.99 (0.01)	0.06 (0.25)	-0.23 (0.62)	-0.29 (0.63)
13	0.4	170	1.00 (0.00)	1.00 (0.00)	0.02 (0.14)	-0.09 (0.24)	-0.11 (0.23)
14	0.4	200	1.00 (0.00)	1.00 (0.00)	0.00 (0.07)	-0.02 (0.10)	-0.02 (0.08)
15	observed	20	0.29 (0.18)	0.28 (0.18)	5.23 (16.80)	-	-
16	observed	50	0.54 (0.29)	0.54 (0.29)	3.09 (11.09)	-	-
17	observed	80	0.67 (0.31)	0.67 (0.31)	2.06 (8.78)	-	-
18	observed	110	0.74 (0.30)	0.74 (0.30)	1.76 (7.82)	-	-
19	observed	140	0.79 (0.29)	0.79 (0.29)	1.34 (7.18)	-	-
20	observed	170	0.82 (0.28)	0.82 (0.28)	1.40 (7.91)	-	-
21	observed	200	0.84 (0.27)	0.84 (0.27)	1.30 (7.84)	-	-

Simulation problem 2

	Target effect size	*Target sample size*	$\hat{\text{pwr}}^{\text{ups}}$ *mean (sd)*	$\hat{\text{pwr}}^{\text{cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,true}}$ *mean (sd)*	$\text{PE}^{\text{cmp,true}}$ *mean (sd)*
1	0.3	20	0.16 (0.02)	0.15 (0.02)	2.73 (9.35)	2.09 (13.86)	-0.57 (10.38)

2	0.3	50	0.33 (0.04)	0.32 (0.04)	1.66 (4.88)	6.75 (13.62)	5.03 (12.60)
3	0.3	80	0.48 (0.06)	0.48 (0.05)	1.61 (3.25)	3.26 (12.08)	1.64 (11.55)
4	0.3	110	0.61 (0.06)	0.60 (0.06)	1.45 (2.44)	2.55 (10.38)	1.10 (10.17)
5	0.3	140	0.72 (0.06)	0.71 (0.06)	1.24 (1.99)	0.99 (8.72)	-0.23 (8.57)
6	0.3	170	0.80 (0.06)	0.79 (0.06)	1.08 (1.67)	0.75 (7.15)	-0.32 (7.12)
7	0.3	200	0.86 (0.05)	0.85 (0.05)	0.88 (1.37)	0.70 (5.78)	-0.16 (5.81)
8	0.4	20	0.24 (0.03)	0.24 (0.03)	1.52 (7.14)	2.92 (14.05)	1.37 (11.85)
9	0.4	50	0.52 (0.06)	0.51 (0.06)	1.48 (3.11)	2.50 (11.64)	1.01 (11.16)
10	0.4	80	0.72 (0.06)	0.71 (0.06)	1.14 (2.04)	0.78 (8.62)	-0.34 (8.49)
11	0.4	110	0.85 (0.05)	0.84 (0.05)	0.88 (1.42)	0.93 (6.07)	0.06 (6.04)
12	0.4	140	0.92 (0.04)	0.91 (0.04)	0.52 (0.99)	0.67 (4.04)	0.15 (4.05)
13	0.4	170	0.96 (0.02)	0.95 (0.02)	0.35 (0.71)	0.29 (2.55)	-0.06 (2.60)
14	0.4	200	0.98 (0.02)	0.98 (0.02)	0.23 (0.50)	-0.02 (1.60)	-0.25 (1.61)
15	observed	20	0.21 (0.15)	0.20 (0.15)	11.71 (23.27)	-	-
16	observed	50	0.39 (0.28)	0.38 (0.28)	7.61 (19.74)	-	-
17	observed	80	0.50 (0.32)	0.50 (0.32)	6.16 (17.40)	-	-
18	observed	110	0.58 (0.34)	0.57 (0.34)	5.38 (16.53)	-	-
19	observed	140	0.63 (0.34)	0.62 (0.35)	4.80 (15.17)	-	-
20	observed	170	0.67 (0.34)	0.66 (0.35)	3.88 (13.42)	-	-
21	observed	200	0.70 (0.34)	0.69 (0.34)	3.57 (12.89)	-	-

Simulation problem 3

	Target effect size	*Target sample size*	$\hat{\text{pwr}}^{\text{ups}}$ *mean (sd)*	$\hat{\text{pwr}}^{\text{cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,true}}$ *mean (sd)*	$\text{PE}^{\text{cmp,true}}$ *mean (sd)*
1	0.5	20	0.33 (0.04)	0.33 (0.05)	1.70 (8.61)	3.38 (13.90)	2.09 (14.32)
2	0.5	50	0.70 (0.06)	0.69 (0.07)	1.06 (3.20)	2.81 (9.55)	1.80 (9.61)
3	0.5	80	0.88 (0.05)	0.87 (0.05)	1.10 (1.87)	1.36 (5.26)	0.29 (5.48)
4	0.5	110	0.96 (0.03)	0.95 (0.03)	0.57 (1.07)	0.74 (2.66)	0.18 (2.77)
5	0.5	140	0.99 (0.01)	0.98 (0.01)	0.26 (0.60)	0.01 (1.22)	-0.25 (1.34)
6	0.5	170	0.99 (0.01)	0.99 (0.01)	0.11 (0.37)	0.06 (0.55)	-0.06 (0.62)
7	0.5	200	1.00 (0.00)	1.00 (0.00)	0.05 (0.23)	-0.04 (0.25)	-0.09 (0.30)
8	1	20	0.85 (0.05)	0.85 (0.06)	0.77 (3.48)	0.52 (5.96)	-0.15 (6.52)
9	1	50	1.00 (0.00)	1.00 (0.00)	0.02 (0.24)	-0.06 (0.30)	-0.08 (0.32)
10	1	80	1.00 (0.00)	1.00 (0.00)	-0.00 (0.03)	-0.00 (0.02)	-0.00 (0.02)
11	1	110	1.00 (0.00)	1.00 (0.00)	0.00 (0.01)	0.00 (0.00)	-0.00 (0.01)
12	1	140	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
13	1	170	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
14	1	200	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
15	observed	20	0.36 (0.21)	0.36 (0.21)	2.19 (10.79)	-	-
16	observed	50	0.65 (0.28)	0.64 (0.28)	1.15 (6.76)	-	-
17	observed	80	0.77 (0.26)	0.77 (0.26)	1.35 (5.53)	-	-
18	observed	110	0.84 (0.24)	0.83 (0.24)	1.02 (4.41)	-	-
19	observed	140	0.87 (0.22)	0.87 (0.22)	0.87 (4.11)	-	-
20	observed	170	0.90 (0.20)	0.89 (0.21)	0.69 (3.74)	-	-
21	observed	200	0.92 (0.19)	0.91 (0.19)	0.71 (4.18)	-	-

Simulation problem 4

	Target effect size	*Target sample size*	$\hat{\text{pwr}}^{\text{ups}}$ *mean (sd)*	$\hat{\text{pwr}}^{\text{cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,true}}$ *mean (sd)*	$\text{PE}^{\text{cmp,true}}$ *mean (sd)*
1	0.5	20	0.10 (0.01)	0.10 (0.02)	-0.61 (17.4)	-1.32 (13.1)	1.06 (15.4)
2	0.5	50	0.21 (0.02)	0.21 (0.02)	-1.91 (8.25)	-6.38 (8.29)	-4.23 (8.37)
3	0.5	80	0.31 (0.02)	0.31 (0.02)	-0.83 (6.79)	-4.06 (7.67)	-3.06 (7.45)
4	0.5	110	0.40 (0.03)	0.40 (0.03)	-0.54 (5.36)	-4.48 (6.89)	-3.83 (6.82)
5	0.5	140	0.49 (0.03)	0.49 (0.03)	-0.39 (4.51)	-4.69 (6.38)	-4.23 (6.22)
6	0.5	170	0.57 (0.03)	0.57 (0.03)	-0.40 (3.90)	-4.87 (5.78)	-4.42 (5.74)
7	0.5	200	0.63 (0.04)	0.64 (0.03)	-0.47 (3.46)	-4.84 (5.38)	-4.35 (5.20)
8	1	20	0.27 (0.03)	0.27 (0.04)	-1.79 (11.73)	-6.66 (11.64)	-4.08 (13.62)
9	1	50	0.61 (0.04)	0.62 (0.04)	-1.36 (3.62)	-3.12 (6.41)	-1.71 (6.54)
10	1	80	0.81 (0.04)	0.81 (0.04)	-0.60 (2.23)	-2.62 (4.23)	-2.02 (4.23)
11	1	110	0.91 (0.02)	0.92 (0.02)	-0.31 (1.35)	-1.56 (2.58)	-1.24 (2.60)
12	1	140	0.96 (0.01)	0.96 (0.01)	-0.13 (0.87)	-0.87 (1.52)	-0.74 (1.50)
13	1	170	0.98 (0.01)	0.99 (0.01)	-0.06 (0.58)	-0.56 (0.83)	-0.50 (0.83)
14	1	200	0.99 (0.00)	0.99 (0.00)	-0.03 (0.35)	-0.17 (0.46)	-0.14 (0.44)
15	observed	20	0.15 (0.12)	0.16 (0.12)	-0.22 (16.72)	-	-
16	observed	50	0.31 (0.25)	0.32 (0.25)	-0.29 (11.79)	-	-
17	observed	80	0.41 (0.30)	0.41 (0.30)	0.10 (10.26)	-	-
18	observed	110	0.48 (0.33)	0.49 (0.33)	0.05 (9.13)	-	-
19	observed	140	0.54 (0.34)	0.54 (0.34)	-0.12 (8.98)	-	-
20	observed	170	0.58 (0.35)	0.58 (0.35)	0.09 (7.93)	-	-
21	observed	200	0.61 (0.35)	0.61 (0.35)	0.27 (7.05)	-	-

Simulation problem 5

	Target effect size	*Target sample size*	$\hat{\text{pwr}}^{\text{ups}}$ *mean (sd)*	$\hat{\text{pwr}}^{\text{cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,true}}$ *mean (sd)*	$\text{PE}^{\text{cmp,true}}$ *mean (sd)*
1	0.5	20	0.26 (0.04)	0.26 (0.04)	2.37 (10.20)	4.13 (14.85)	2.18 (14.42)
2	0.5	50	0.58 (0.06)	0.57 (0.06)	1.51 (3.92)	3.71 (11.60)	2.26 (11.61)
3	0.5	80	0.78 (0.06)	0.77 (0.06)	1.67 (2.56)	1.90 (7.94)	0.28 (8.06)
4	0.5	110	0.89 (0.04)	0.88 (0.05)	1.11 (1.72)	1.04 (5.04)	-0.05 (5.16)
5	0.5	140	0.95 (0.03)	0.94 (0.03)	0.72 (1.13)	0.52 (2.97)	-0.18 (3.15)
6	0.5	170	0.98 (0.02)	0.97 (0.02)	0.42 (0.77)	0.42 (1.69)	0.01 (1.87)
7	0.5	200	0.99 (0.01)	0.99 (0.01)	0.24 (0.52)	0.13 (0.94)	-0.11 (1.06)
8	1	20	0.75 (0.06)	0.74 (0.07)	1.11 (4.37)	-0.52 (8.48)	-1.48 (8.89)
9	1	50	0.99 (0.01)	0.99 (0.01)	0.10 (0.58)	-0.28 (1.10)	-0.39 (1.14)
10	1	80	1.00 (0.00)	1.00 (0.00)	0.01 (0.11)	-0.01 (0.11)	-0.02 (0.12)
11	1	110	1.00 (0.00)	1.00 (0.00)	0.00 (0.02)	-0.00 (0.02)	-0.00 (0.02)
12	1	140	1.00 (0.00)	1.00 (0.00)	-0.00 (0.01)	-0.00 (0.01)	-0.00 (0.00)
13	1	170	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
14	1	200	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
15	observed	20	0.30 (0.19)	0.30 (0.19)	2.30 (12.12)	-	-
16	observed	50	0.56 (0.29)	0.55 (0.29)	0.90 (7.23)	-	-
17	observed	80	0.69 (0.30)	0.68 (0.30)	1.41 (6.37)	-	-
18	observed	110	0.76 (0.29)	0.75 (0.29)	0.96 (5.39)	-	-

19	observed	140	0.80 (0.28)	0.80 (0.28)	0.97 (5.18)	-	-
20	observed	170	0.83 (0.27)	0.83 (0.27)	0.67 (4.54)	-	-
21	observed	200	0.85 (0.26)	0.85 (0.26)	0.61 (4.26)	-	-

Simulation problem 6

	Target effect size	*Target sample size*	$\hat{\text{pwr}}^{\text{ups}}$ *mean (sd)*	$\hat{\text{pwr}}^{\text{cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,cmp}}$ *mean (sd)*	$\text{PE}^{\text{ups,true}}$ *mean (sd)*	$\text{PE}^{\text{cmp,true}}$ *mean (sd)*
1	0.5	20	0.20 (0.02)	0.20 (0.02)	-2.60 (9.79)	1.32 (10.50)	4.74 (12.50)
2	0.5	50	0.40 (0.05)	0.41 (0.05)	-0.66 (5.64)	9.92 (14.36)	10.88 (14.89)
3	0.5	80	0.57 (0.07)	0.57 (0.07)	-0.06 (4.18)	8.69 (13.87)	8.90 (14.21)
4	0.5	110	0.70 (0.08)	0.70 (0.08)	-0.40 (2.92)	7.01 (12.07)	7.50 (12.19)
5	0.5	140	0.79 (0.07)	0.79 (0.07)	-0.38 (2.44)	6.70 (10.12)	7.14 (10.13)
6	0.5	170	0.85 (0.07)	0.86 (0.07)	-0.32 (1.86)	4.60 (8.20)	4.95 (8.14)
7	0.5	200	0.90 (0.06)	0.90 (0.06)	-0.25 (1.59)	2.36 (6.41)	2.63 (6.34)
8	1	20	0.53 (0.06)	0.54 (0.08)	-0.42 (6.79)	9.71 (13.00)	10.86 (16.13)
9	1	50	0.88 (0.06)	0.89 (0.06)	-1.11 (1.63)	2.48 (6.90)	3.65 (6.96)
10	1	80	0.97 (0.03)	0.98 (0.02)	-0.52 (0.79)	0.37 (2.79)	0.89 (2.53)
11	1	110	0.99 (0.01)	0.99 (0.01)	-0.20 (0.40)	-0.14 (1.07)	0.05 (0.88)
12	1	140	1.00 (0.00)	1.00 (0.00)	-0.07 (0.20)	-0.10 (0.42)	-0.03 (0.32)
13	1	170	1.00 (0.00)	1.00 (0.00)	-0.02 (0.10)	-0.03 (0.18)	-0.01 (0.12)
14	1	200	1.00 (0.00)	1.00 (0.00)	-0.01 (0.04)	-0.01 (0.06)	-0.01 (0.05)
15	observed	20	0.24 (0.16)	0.25 (0.17)	-1.23 (12.01)	-	-
16	observed	50	0.42 (0.27)	0.44 (0.28)	-2.04 (9.49)	-	-
17	observed	80	0.54 (0.31)	0.55 (0.31)	-1.07 (9.33)	-	-
18	observed	110	0.61 (0.32)	0.62 (0.32)	-0.61 (9.52)	-	-
19	observed	140	0.66 (0.32)	0.67 (0.32)	-0.34 (9.19)	-	-
20	observed	170	0.70 (0.32)	0.70 (0.32)	-0.07 (9.67)	-	-
21	observed	200	0.72 (0.31)	0.73 (0.32)	-0.12 (9.72)	-	-