Determine the optimal study design under a fixed budget constraint

A simulation study

Miaoyan Chen

Introduction

In a clustered randomized trial (CRT), groups or clusters of subjects are randomly assigned to either the intervention/treatment group or the control group. Subjects within the same cluster are expected to exhibit more similar traits and characteristics compared to subjects in other clusters. This introduces both intra-cluster and inter-cluster correlations, which must be considered when designing the experiment. Ideally, independent observations from each participant are preferred over multiple measurements from the same participant. However, when a trial is constrained by budget or time, taking multiple measurements from the same subject may or may not be beneficial, depending on the setup. The purpose of this simulation is to determine the optimal number of clusters and the number of observations within each cluster under a fixed budget constraint B.

Suppose we have our budget fixed at 2000 units. The cost of sampling a new individual is C_1 unit(s), and the cost of taking additional measurements on the same individual is C_2 unit(s), where c_2 is cheaper than c_1. We will explore two clustered scenarios by introducing linear and poisson hierarchical models, incorporating both fixed and random effects.

Setup

To begin, let us consider the setting that Y follows a normal distribution. Let X_i be a binary indicator of whether or not cluster i is assigned to the treatment group ($X = 0$: control, $X = 1$: treatment) and let Y_{ij} be the observed outcome. To estimate the treatment effect, we will assume a hierarchical model for Y_{ij} where:

$$Y_{ij}|\mu_i \sim N(\mu_i, \sigma^2)$$

This implies that each observation Y_{ij} within a cluster i is normally distributed around the cluster mean μ_i, with variance σ^2. σ^2 is the deviation between individual observations Y_{ij} from the cluster mean μ_i. A larger σ^2 indicates more variability within a cluster.

$$\mu_i = \mu_{i0} + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, \gamma^2)$$

Here we have the cluster mean μ_i centered around μ_{i0}, with some variability introduced by random effect ϵ_i that is normally distributed with mean 0 and variance γ^2. γ^2 controls for between cluster variation, with larger γ^2 meaning more heterogeneity across clusters.

$$\mu_{i0} = \alpha + \beta X_i$$

α is the intercept, representing the average mean outcome for the control group. X_i is the binary treatment indicator (0 for control, 1 for treatment) for the $i-$th cluster, therefore β represents the average treatment effect (when $X_i = 1$).

We also consider the poisson distribution for Y with the following structure:

$$\log(\mu_i) \sim \mathcal{N}(\alpha + \beta X_i, \gamma^2),$$

where α is the log-mean intercept, β is the log-treatment effect, and γ^2 is the inter-cluster variability.

$$Y_{ij} \mid \mu_i \sim \text{Poisson}(\mu_i), \quad j = 1, \ldots, R,$$

where R is the number of repeated measurements within cluster i. The aggregated cluster mean is given as:

$$\bar{Y_i} = \sum_{j=1}^{R} Y_{ij} \sim \text{Poisson}(R\mu_i).$$

Simulation

ADEMP Framework

Aim

1. To assess the performance of different study designs for estimating the average treatment effect β under a fixed budget $B = 2000$ by varying the number of clusters (G), observations per cluster (R), and cost ratios (c_1/c_2) in a cluster randomized trial.

2. To examine how the cost ratio and model parameters influence the optimal strategy.

Data-generation mechanism

The data generation process consists of two steps. First, we consider the budget amount and the relative costs of c_1 and c_2 to determine the number of clusters (G) and the number of observations per cluster (R). Data is simulated with a fixed budget of $B = 2000$, and three cases of relative cost ratios are considered in the data generation mechanism such that $C_2 < C_1$, using the budget constraint inequality. The budget constraint is given by:

$$G \cdot c_1 + (G * (R-1)) \cdot c_2 \leq B$$

Case 1: $c_1 = 50, c_2 = 10 \rightarrow c_1/c_2 = 5$ i.e., cost of repeated measures is cheaper than new samples

Case 2: $c_1 = 50, c_2 = 1 \rightarrow c_1/c_2 = 50$ i.e., cost of repeated measures is much cheaper than new samples

Case 3: $c_1 = 50, c_2 = 45 \rightarrow c_1/c_2 = 1.11$ i.e., cost of repeated measures is relatively similar to the cost of new samples

We then generate data for the two aforementioned hierarchical structures based on the cases of relative cost ratios. For both hierarchical structures, we generate cluster-level assignments with $X_i \sim \text{Bern}(0.5)$, reflecting simple randomization with an equal probability of treatment assignment.

Next, we simulate the cluster-level mean and subject-level outcomes under two settings:

For $Y_{ij} \sim$ normal distribution:

- Cluster-level mean: $\mu_i \mid X_i, \epsilon_i = \alpha + \beta X_i + \epsilon_i, \text{where} \epsilon \sim N(0, \gamma^2)$ We generate the cluster-level mean from a normal distribution with mean $\mu_{i0} = \alpha + \beta X_i$ and variance γ^2.
- Subject-level outcome: $Y_{ij} \mid \mu_i, e_{ij} = \mu_i + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2)$ We generate the subject-level outcome from a normal distribution with mean μ_i and variance σ^2.

For $Y_{ij} \sim$ Poisson distribution:

- Group-level mean: $\log(u_i) \sim N(\alpha + \beta X_i, \gamma^2)$ We generate the cluster-level mean from a normal distribution with mean $\mu_{i0} = \alpha + \beta X_i$ and variance γ^2, and exponentiate it to get the mean counts μ_i
- Subject-level outcome: $Y_{ij} \mid \mu_i \sim Poisson(\mu_i)$ for each subject in the cluster We generate the subject-level outcome from a Poisson distribution with mean μ_i.

The number of simulation repetitions for each parameter combination is set to $n_{sim} = 500$. A seed is set before every data generating mechanism for reproducibility.

Estimand

Our estimand is the cluster average treatment effect β in cluster randomized trial.

Method

1. For the normal distributed outcomes: We fit a linear mixed effects model to estimate β.

2. For the Poisson distributed outcomes: We fit a Poisson generalized linear mixed effects model to estimate β.

We vary parameters $\alpha, \beta, \gamma,$ and σ and relative costs c_1 and c_2 to explore the impact of cost, inter-cluster variability and intra-cluster variability on the optimal G and R trade-off and the study design

Performance measure

The performance metric we looked at is $Var(\hat{\beta})$ the variance of β estimate from the mixed effect regression models, which quantifies the variability of the estimate of the treatment effect when repeated over multiple iterations. A smaller $Var(\hat{\beta})$ indicates that the estimator is more precise, whereas, a large $Var(\hat{\beta})$ would indicate less a reliable estimate of treatment effect.

Results

1. Optimal strategies under different cost ratios

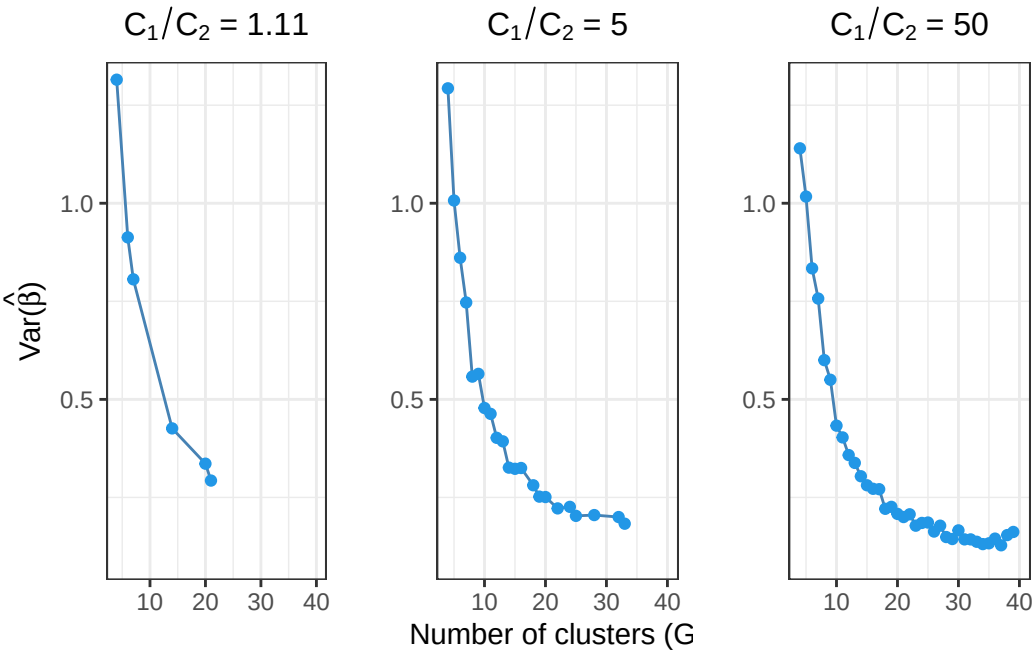

Figure 1: Variance of beta across clusters for cost ratios

When the cost of an additional sample c_2 is relatively close to the cost of the first sample (c_1), i.e., $c_1 \approx c_2$, it is more likely that we would start a new cluster (G) rather than resampling from the same cluster. In the case of slightly higher cost ratio ($c_1/c_2 = 5$), where the first sample is somewhat more expensive then resampling, we tend to balance between G and R to reach a optimal strategy. For the case where $c_2 << c_1$, it is better to maximize R and limit the number of clusters (G). Figure 1 shows the relationship between the $Var(\hat{\beta})$ and number of clusters (G) under 3 different cases of cost ratios. In the left panel, for $c_1/c_2 = 1.11$, the optimal strategy corresponding to the lowest variance is G = 21 Clusters and R = 2. For cost ratio $c_1/c_2 = 5$ and $c_1/c_2 = 50$, we observe a slower reduction in $Var(\hat{\beta})$ as G increases, with the variance curve flattening. hypothetically, under certain constraint, the optimal strategy would align with this flattened region, resembling diminishing returns for additional cluster. However, to fully utilize our given budget, the optimal strategy would be the lowest $Var(\hat{\beta})$ on the curve, which is G = 33, R = 2 for $c_1/c_2 = 5$, and G = 37, R = 5 for $c_1/c_2 = 50$.

2. Parameters and costs impact on study design

2.1 Normal Hierarchical Model

We examined how the cost ratios and the underlying parameters (α, β, γ^2, σ^2) in the linear hierarchical model influence the optimal strategy under a budget constraint. α is the intercept and, therefore, has no effect on β estimation. β is the treatment effect, meaning higher values for β make it much easier to detect the treatment effect due to higher signal-to-noise ratios. Thus, for higher true β values, the variance of $\hat{\beta}$ is lower. In this model, σ^2 represents the intra-cluster variance, while γ^2 represents the inter-cluster variance. Figure 2 illustrates that as γ increases, $Var(\hat{\beta})$ also increases across all scenarios.

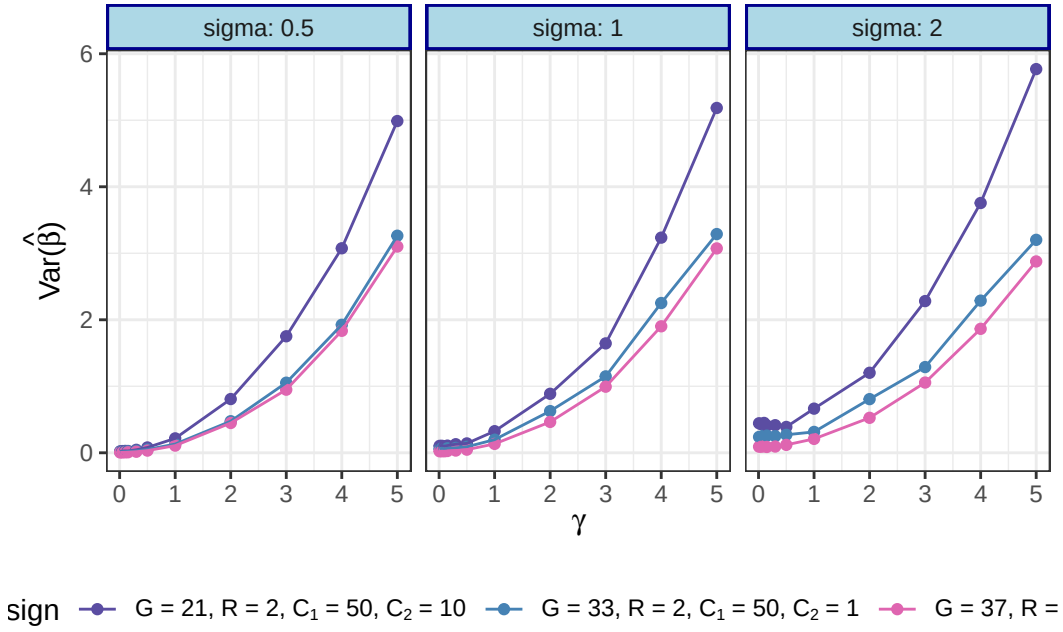

Figure 2: Variation of beta by gamma and sigma

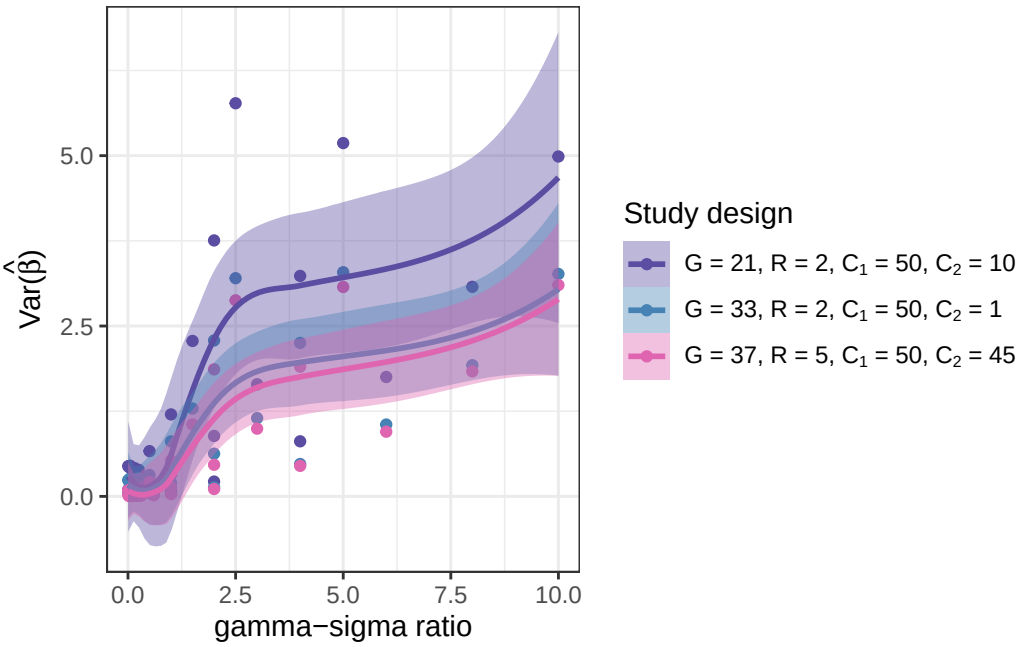

Figure 3: Variance of beta vs. gamma/sigma ratio

Figure 3 shows the relationship between the γ/σ ratio and the variance of $\hat{\beta}$ for different study designs. The pink study design consistently shows the lowest variance across all gamma-sigma ratios. Meanwhile the purple study design shows highest variance, especially at higher gamma-sigma ratios, indicating that fewer clusters and subjects lead to less precise estimates of treatment effect. The γ/σ ratio is a measure of the relative importance of inter-cluster variability (γ) to intra-cluster variability (σ). A higher γ/σ ratio indicates that inter-cluster variability is more significant or dominant than intra-cluster variability. We observe that the variance of $\hat{\beta}$ increases with the γ/σ ratio, indicating that the precision of the treatment effect estimate decreases as the relative importance of inter-cluster variability increases.

2.2 Poisson Hierarchical Model

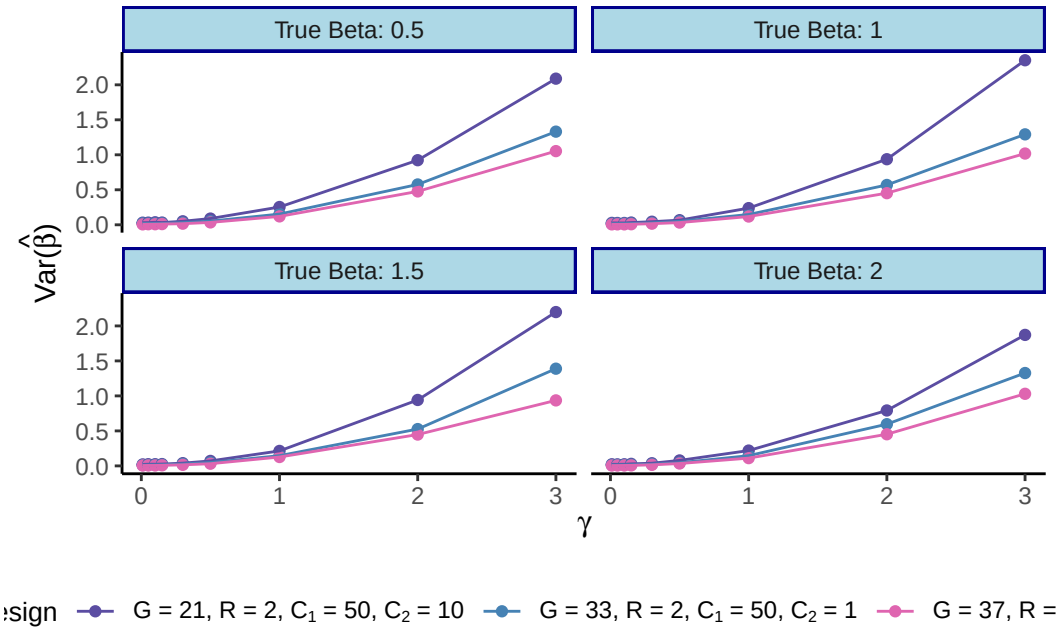

Figure 4: Impact of Gamma on Variance of Beta

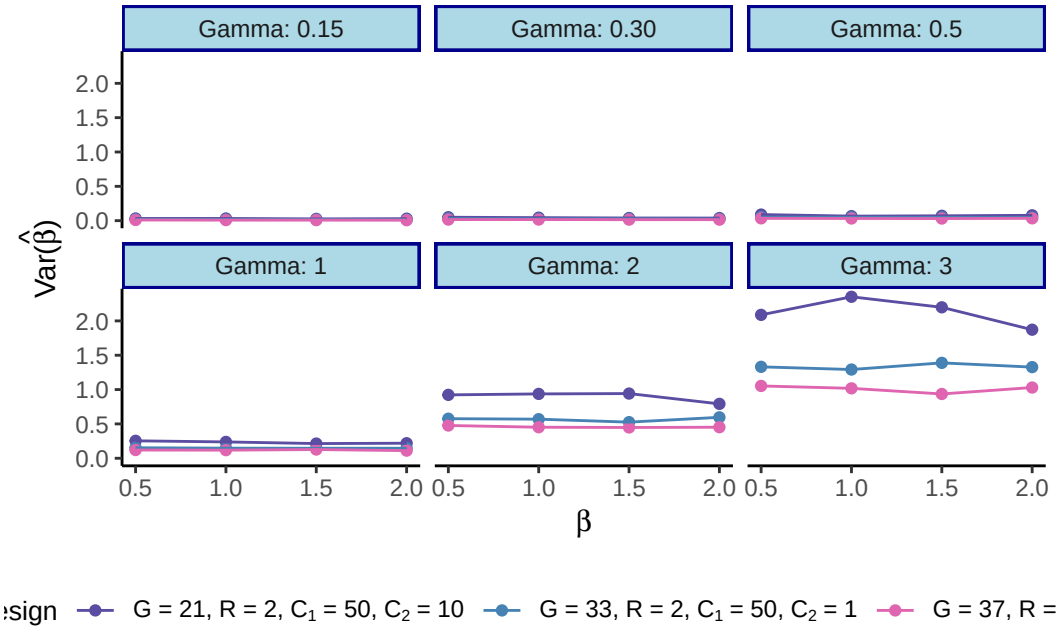

Figure 5: Impact of Beta on Variance

Figure 4 shows the trend of $Var(\hat{\beta})$ across different values of γ for the Poisson hierarchical model. We observe a positive trend in $Var(\hat{\beta})$ as γ increases, indicating that the variance of the treatment effect estimate increases with higher inter-cluster variability. This trend is consistent across different true β values, with higher β values exhibiting lower variance in $\hat{\beta}$ estimates. Study designs with larger G and R have lower $\text{Var}(\hat{\beta})$, with $G = 37, R = 5, c_1 = 50, c_2 = 45$ yielding the most precise estimates, while designs with fewer clusters and subjects (e.g., $G = 21, R = 2, c_1 = 50, c_2 = 10$) result in higher variance. There's also some fluctuation in $Var(\hat{\beta})$ across β values as γ gets larger.

Discussion and limitations

For small γ (e.g., $\gamma < 1$), the ideal study design prioritizes more observations (R) and fewer clusters (G). This is because the treatment effect is relatively homogeneous across clusters, and adding new clusters does not capture additional variability, and therefore does not meaningfully improve the precision of the treatment effect estimate. In this case, increasing R (repeated measurements within a cluster) is more cost-effective.

In contrast, when γ is large (e.g., $\gamma > 1$), the optimal strategy is to increasing the number of clusters (G) to better capture the variability across clusters. In this scenario, adding more observations (R) within the same cluster becomes less effective, as repeated measurements are highly correlated and provide limited additional information. Therefore, reducing R and allocating resources to increase G is a more efficient strategy for improving the precision of the treatment effect estimate.

When the cost of initiating a new cluster (c_1) is higher than the cost of taking repeated measurements (c_2), the study design should balance G and R based on the value of γ. For small γ, it is more cost-effective to allocate the budget toward increasing R. Conversely, for large γ, the budget should be allocated to maximize G, as capturing variability across clusters becomes more critical to improving the precision of the treatment effect estimate.

There is a limitation to the simulation study, as we only considered a fixed costs for c_1 and c_2 for all the clusters. In practice, the cost of sampling and repeated measures may vary across clusters. Future studies could consider a more flexible cost structure that allows for varying costs across clusters.

In addition, the number of simulation iterations could be increased to improve the robustness of the results. A larger number of iterations, such as $n_{sim} = 1000$ would provide more accurate estimates of $\hat{\beta}$ and the variance of $\hat{\beta}$.

A key limitation is that the model fitting procedure is associated with the "boundary (singular)" warning in mixed-effects models, indicating that the model has likely encountered singular fits when the random effects parameters are near zero. Other warnings, such as non-convergence, were also encountered in the simulation study and could not be resolved by adjusting the model specifications for this simulation study.

Conclusion

This simulation demonstrates the impact of underlying parameters (γ, σ, α, β) and study design parameters (number of clusters G, subjects per cluster R, and costs c_1, c_2) on the precision of $\widehat{\beta}$, the treatment effect estimate. The findings suggest that increasing G (number of clusters) is generally more effective than increasing R (subjects per cluster) for improving precision, particularly when γ is large.