Optimal Study Designs in Cluster Randomized Trials – Combining Insights from Theory and Simulation

Daniel Posmik[1]
Brown University School of Public Health, Department of Biostatistics, 121 S Main St, Providence RI 02903
E-Mail: daniel_posmik@brown.edu | GitHub: https://github.com/posmikdc/php2550-projects/tree/main/project3

Overview

When designing optimal study designs in the presence of clustering, researchers must weigh the benefit the value of sampling fewer or more clusters. We find that higher cluster numbers generally minimize the variance of treatment effect estimator. This result is most sensitive to varying individual-level sampling cost.

Background

- In many CRT settings, researchers seek to optimize their budget efficacy across individual-level and cluster-level costs.
- **Setting:**
- The first individual is sampled at cost c_1, the rest at cost c_2:

$$T = G[c_1 + (r-1)c_2]$$

- **Question:**
- Given T, what combination of individual and cluster-level samples (R^*, G^*) optimizes the variance of our estimator?
- **Parameter of interest:**
- The parameter of interest is the treatment effect β

Results

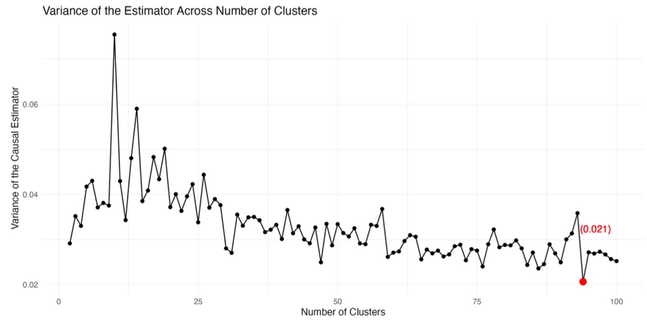
- Interesting theory comes from Raudenbush (1997). We derive closed-form solutions for (R^*, G^*) as such:

$$R^* - 1 = \frac{\sigma}{\gamma}\sqrt{\frac{c_1}{c_2}} \implies R^* = \lfloor \frac{\sigma}{\gamma}\sqrt{\frac{c_1}{c_2}} + 1 \rfloor$$

$$G^* = \frac{T}{\frac{\sigma}{\gamma}\cdot\sqrt{c_1 \cdot c_2} + c_1} = \lfloor \frac{T}{(R^*-1)c_2 + c_1} \rfloor$$

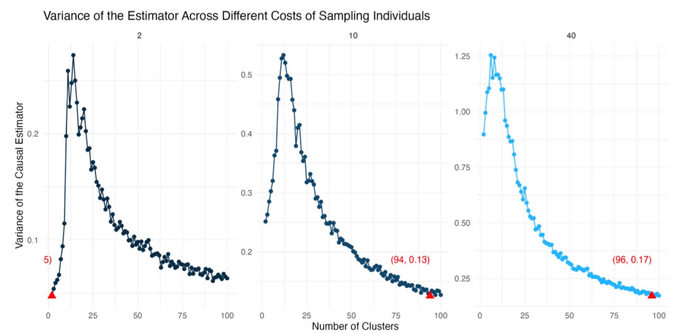
$$\mathrm{Var}(\hat{\beta}) = \frac{4(\gamma^2 + \sigma^2/R^*)}{R^* \cdot G^*}$$

- The theory supports our simulation results. In the following, we simulate the variance of β across different G, keeping fixed all other parameters. Notably, budget $B = b, c_1 = 50, c_2 = 2$.

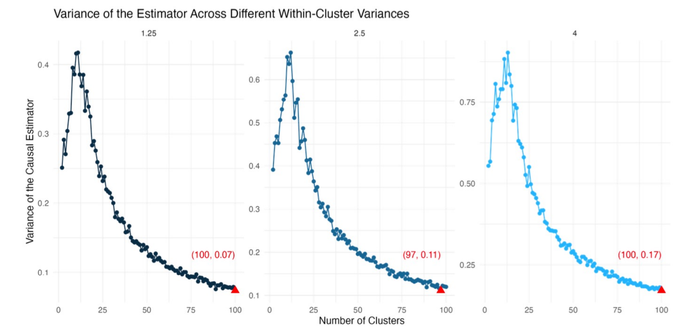


Sensitivity to Parameters

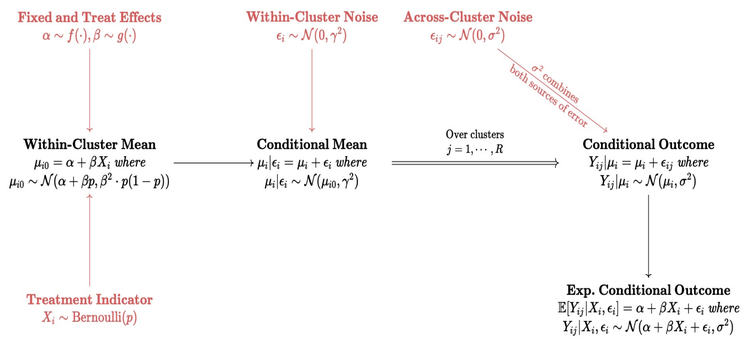
- We explored which parameters affect optimal study design.
- **Robustness across c_2**:



- **Robustness across intracluster correlation ($\rho = (\gamma^2)/(\gamma^2 + \delta^2)$):**

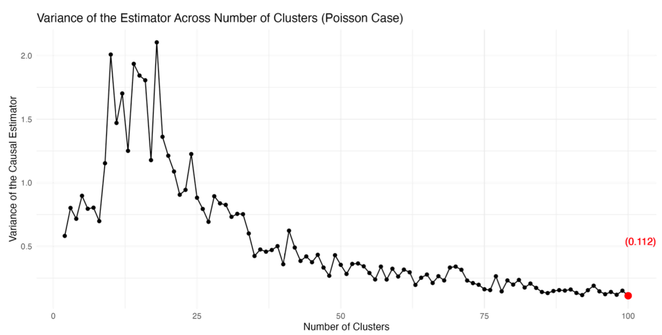


DGP[1]s and Data Structure



Robustness across DGPs

- Suppose our outcomes are Poisson-generated count data. Does our result hold beyond the normal case?



Conclusion

- Generally, sampling from more clusters yields better experimental designs.
- **High sensitivity to c_2** can be observed. When individual-level sampling cost is low, we are able to obtain a variance-minimizing amount of clusters much earlier, i.e. on the left side of the peak
- **Low sensitivity to intracluster correlation** is surprisingly observed in our simulations. This contrasts with the theoretical results of Raudenbush (1997)
- This above results hold beyond the normal case, specifically when the outcome is contrasts with data.

Footnotes: (1) DGP: Data generating process

This project was completed for PHP2550 (Practical Data Analysis), Fall 2024. Instructor: Dr. Alice Paul