Statistica Sinica **35** (2025), 1-20
doi:https://doi.org/10.5705/ss.202022.0414

CONSTRAINED D-OPTIMAL DESIGN FOR PAID RESEARCH STUDY

Yifei Huang[1], Liping Tong[2] and Jie Yang*[1]

[1] *University of Illinois at Chicago and* [2] *Advocate Aurora Health*

Abstract: We consider constrained sampling problems in paid research studies or clinical trials. When qualified volunteers are more than the budget allowed, we recommend a D-optimal sampling strategy based on the optimal design theory and develop a constrained lift-one algorithm to find the optimal allocation. Unlike the literature which mainly deals with linear models, our solution solves the constrained sampling problem under fairly general statistical models, including generalized linear models and multinomial logistic models, and with more general constraints. We justify theoretically the optimality of our sampling strategy and show by simulation studies and real-world examples the advantages over simple random sampling and proportionally stratified sampling strategies.

Key words and phrases: Constrained sampling, D-optimal design, generalized linear model, lift-one algorithm, multinomial logistic model.

1. Introduction

We consider a constrained sampling problem frequently arising in paid research studies or clinical trials, especially when recruiting volunteers via the Internet or emails, which could gather attention widely and quickly. For example, some investigators plan to conduct a research study to evaluate the effect of a new treatment on anxiety. Besides the treatment cost, the investigators also need to prepare certain compensation for participants' time. Due to limited funding, the investigators could only support up to n participants while there are $N > n$ eligible volunteers. The question is how they select n participants out of N to evaluate the treatment effect most accurately. Noted that the goal of the sampling problem in this paper is not the mean of response but the treatment effect or regression coefficients of an underlying statistical model.

A straightforward approach is to use the *simple random sampling without replacement* (SRSWOR) (Lohr, 2019, Ch. 2), which randomly chooses an index set $1 \le i_1 < i_2 < \cdots < i_n \le N$ such that each index set of n distinct subjects has the equal chance $n!(N-n)!/N!$ to be chosen. This can be applied if the investigators know nothing about the volunteers except contact information, or the covariate information provided by the volunteers does not seem relevant to the treatment effect.

*Corresponding author. E-mail: jyang06@uic.edu