Choosing good subsamples for regression models

Tong Chen[1, 2] and Thomas Lumley[2]

1 Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Melbourne, Australia
2 Department of Statistics, University of Auckland, Auckland, New Zealand

Introduction

- A common problem in health research is that we have a large database with many variables measured on a large number of individuals.
- We are interested in measuring **additional variables** on a subsample; these measurements may be newly available, or expensive, or simply not considered when the data were first collected.
- The intended use for the new measurements is to fit a regression model generalisable to the whole cohort (and to its source population).
- This is a two-phase sampling problem. It measures variables of interest on a subcohort where the outcome and covariates are readily available or cheap to collect on all individuals in the cohort.
- We aimed to choose a good phase-two subsample to minimise the variance of parameters of interest in the regression model.
- We focus on deriving the optimal sampling for design-based estimators.

From sums to parameters

- In classical design and analysis, researchers were interested in estimating totals.
- A unifying concept in translating the classical results from sums to regression parameters is the ***influence function***.
- Influence function shows the behavior of the target estimator under slight perturbations of the empirical distribution.
- Suppose β are the regression parameters in the model of interest, an asymptotically linear estimator satisfies:

$$\sqrt{N}(\hat{\beta} - \beta) = \frac{1}{\sqrt{N}}\sum \mathbf{h}_i(\beta) + o_p(1)$$

where $\mathbf{h}_i(\beta)$ is a function of the ith observation.

- The mean of influence function for an asymptotically linear estimator gives the linear approximation of the estimator
- For design-based estimators, Demnati and Rao (2004) showed that the influence functions can be computed as the derivative of the estimator of regression parameters with respected to weights on each observation

On optimal designs

- Suppose the number of strata is K, we want to sample n individuals from a cohort of size N.
- **Optimal design for sums:**

Under stratified random sampling for estimation of the population totals of variable Y, the optimal allocation is Neyman allocation :

$$n_k \propto N_k \sigma_k$$

where σ_k is the standard deviation of the variable Y.

- **Optimal design for regression with IPW estimation**

Since the regression estimates are asymptotically equivalent to the estimated population total of influence functions. The optimal design for analysis via the IPW estimator is to apply Neyman allocation to influence functions:

$$n_k \propto N_k \sqrt{\text{var}(\mathbf{h}_i(\beta) \mid \text{stratum } k)}$$

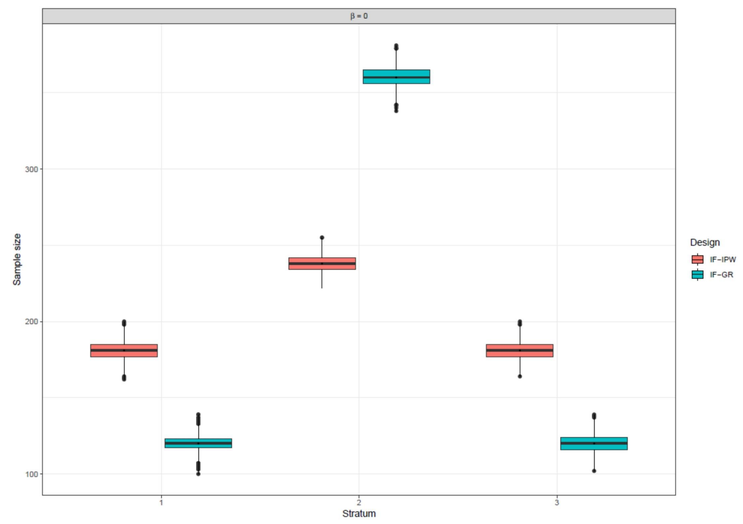
In the case of binary variables, McIsaac and Cook (2015) arrived at this design rule by direct optimisation using Lagrange multipliers

- **Optimal design for regression with AIPW estimation**

Chen and Lumley (2022) showed the optimal design for analysis via the AIPW / Generalised raking estimator was to apply Neyman allocation to the projections of influence functions

$$n_k \propto N_k \sqrt{\text{var}(\mathbf{h}_i(\beta) - \mathbf{h}_i(\beta^*)\theta) \mid \text{stratum } k)}$$

where $\mathbf{h}_i(\beta)$ are the influence functions and $\mathbf{h}_i(\beta^*)$ are the best estimates of influence functions we have of them.



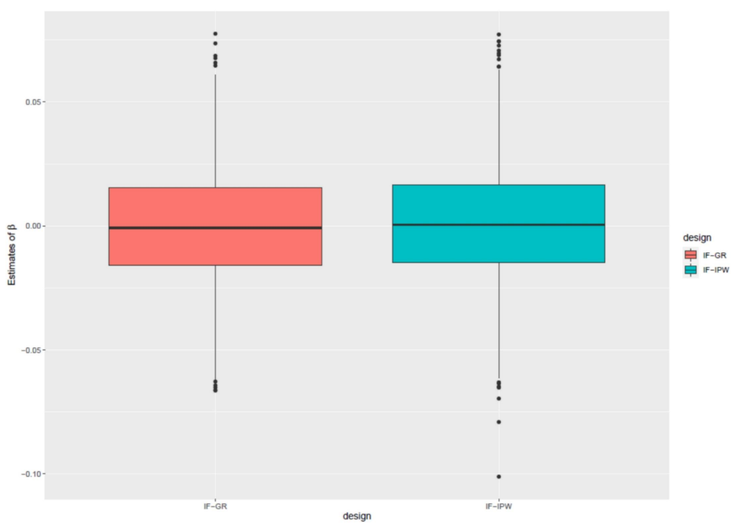
Empirical comparison of sample sizes between the optimal design for analysis via the IPW estimator (IF-IPW) and those for analysis via generalized raking estimators (IF-GR).

- The two designs end up with very similar efficiency under generalised raking analysis
- The two designs can be very different.
- Broadly the same results were found for a wide range of scenarios in Chen and Lumley (2022).

Simulation studies

- Generated 2000 phase-one samples with size 4000 and sampled 600 individuals at phase two.
- Suppose both the variables of interest X and outcome Y were continuous, and $X \sim N(0,1)$.
- An error-prone variable $\tilde{X}$ was generated from $X+U$, where $U \sim N(0, 0.5^2)$.
- $Y = 1 + 0 \times X + Z_1 + Z_2 + \epsilon$, where $Z_1 \sim \text{Ber}n\ (0.5), Z_2 \sim N(0,1)$, and $\epsilon \sim N(0,1)$.
- We defined 3 strata based on the cut-off points at the 20^{th} and 80^{th} percentiles of $\tilde{X}$.
- We compared the optimal design for analysis via the IPW estimator (IF-IPW) with that for analysis via the generalised raking estimators (IF-GR) under generalised raking estimations.

Results:



Empirical comparison of parameter β estimated from raking analysis between IF-IPW and IF-GR.

Conclusions and future work

Conclusions:

- We derived closed-form solutions for the optimal design for analysis via the IPW and generalised raking estimators.
- In practice, it is hard to approximate the optimal design for analysis by generalised raking estimators since it needs to estimate the influence functions and their best estimates.
- Through simulation studies, we found that the two designs can be very different, but they often end up with similar efficiency.
- Lack of improvement is desired in practice.

Future work:

- The semiparametric maximum likelihood estimators are more efficient than design-based estimators, but they are not robust to model misspecification. Even if the model is only slightly misspecified, the design-based estimators can be more efficient in some settings.
- It would be interesting to study and understand the efficiency gap between the model-based and design-based estimators.

References

Chen, T., and Lumley, T. (2022). Optimal sampling for design-based estimators of regression models. Statistics in Medicine, 41, 1482–1497.
Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. Survey Methodology, 30, 17–26.
McIsaac, M.A., and Cook, R.J. (2015). Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. Statistics in Medicine, 34, 2899–2912