A Brief History of Two Phase Designs

Doctoral Oral Qualifying Exam

Chiara Di Gravio

December 9, 2020

Outline

- Background
- Example
- Select the Most Informative Individuals
- Notation
- Existing Methods
- Proposed Work
- Summary and Future Directions

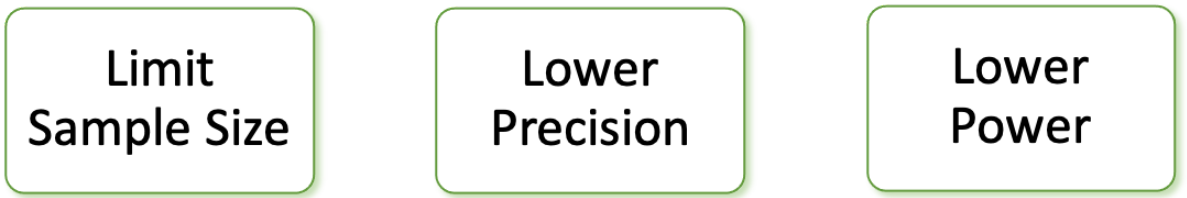
Background

Background

- Electronic health records (EHR) and existing cohort studies provide readibly accessible data on phenotype
- Researchers might be interested in an exposure that is unavailable, and they need to collect additional information

Background

- Electronic health records (EHR) and existing cohort studies provide readibly accessible data on phenotype
- Researchers might be interested in an exposure that is unavailable, and they need to collect additional information
- Problem: the exposure of interest is expensive



Two Phase Design

- We want to use available data in order to identify the most informative subjects for our research question

What we have	What we want
• Outcome • Additional Covariates	• Expensive exposure

- Two phase outcome dependent sampling (ODS) is a retrospective study that assigns different probabilities of being sampled to each individual depending on their observed outcome, or their observed outcome/covariates combination.
- By targeting informative subjects the two phase ODS achieves higher efficiency and power than simple random sampling

Example

Two Phase ODS vs Random Sampling

- Data are from the third and fourth clinical trials of the National Wilms Tumour Study Group
- 4,088 children diagnosed with Wilms tumour. We want to study the relationship between the odds of relapse after chemotherapy, tumour stage and tumour histology

	N (%)
Relapse	571 (14)
Tumour stage	
I	1596 (39)
II	1069 (26)
III	960 (24)
VI	463 (11)
Tumour histology (patient's institution)	
Favourable	3677 (90)
Unfavourable	411 (10)
Tumour histology (NWTSG laboratory)	
Favourable	3623 (89)
Unfavourable	465 (11)

Two Phase ODS vs Random Sampling

- Data are from the third and fourth clinical trials of the National Wilms Tumour Study Group
- 4,088 children diagnosed with Wilms tumour. We want to study the relationship between the odds of relapse after chemotherapy, tumour stage and tumour histology

	N (%)
Relapse	571 (14)
Tumour stage	
I	1596 (39)
II	1069 (26)
III	960 (24)
VI	463 (11)
Tumour histology (patient's institution)	
Favourable	3677 (90)
Unfavourable	411 (10)
Tumour histology (NWTSG laboratory)	
Favourable	3623 (89)
Unfavourable	465 (11)

- A two phase ODS can reduce cost by selecting the most informative individuals for whom a more accurate measure of histology needs to be collected

Model of interest:

$$log\left(\frac{P(\text{relapse})}{1-P(\text{relapse})}\right) = \beta_0 + \beta_1 \text{stage} + \beta_2 \text{histology} + \beta_3 \text{stage * histology}$$

We compared three different analyses:

- Full cohort analysis: consider 4,088 children
- Two phase ODS: sample all cases of relapse, all controls with unfavourable histology, and a random sample of the remaining children such that cases and controls are the same number. This results in 1,142 children
- Random sampling: sample 1,142 children regardless of whether they relapsed

Model of interest:

$$log\left(\frac{P(\text{relapse})}{1-P(\text{relapse})}\right) = \beta_0 + \beta_1 \text{stage} + \beta_2 \text{histology} + \beta_3 \text{stage * histology}$$

We compared three different analyses:

- Full cohort analysis: consider 4,088 children
- Two phase ODS: sample all cases of relapse, all controls with unfavourable histology, and a random sample of the remaining children such that cases and controls are the same number. This results in 1,142 children
- Random sampling: sample 1,142 children regardless of whether they relapsed

	Full Cohort	Two phase ODS	Random Sampling
Intercept	-2.71 (0.11)	-2.72 (0.11)	-2.58 (0.20)
Stage II	0.77 (0.15)	0.78 (0.15)	0.55 (0.27)
Stage III	0.77 (0.15)	0.80 (0.15)	0.83 (0.28)
Stage VI	1.05 (0.17)	1.07 (0.17)	0.66 (0.36)
Unfavourable Histology	1.31 (0.25)	1.46 (0.32)	1.52 (0.43)
Stage II * Unfavourable Histology	0.15 (0.32)	-0.05 (0.44)	-0.13 (0.59)
Stage III * Unfavourable Histology	0.59 (0.32)	0.28 (0.41)	0.02 (0.57)
Stage IV * Unfavourable Histology	1.26 (0.39)	0.91 (0.63)	1.31 (0.71)
Sample Size	4,088	1,142	1,142

The Most Informative Individuals

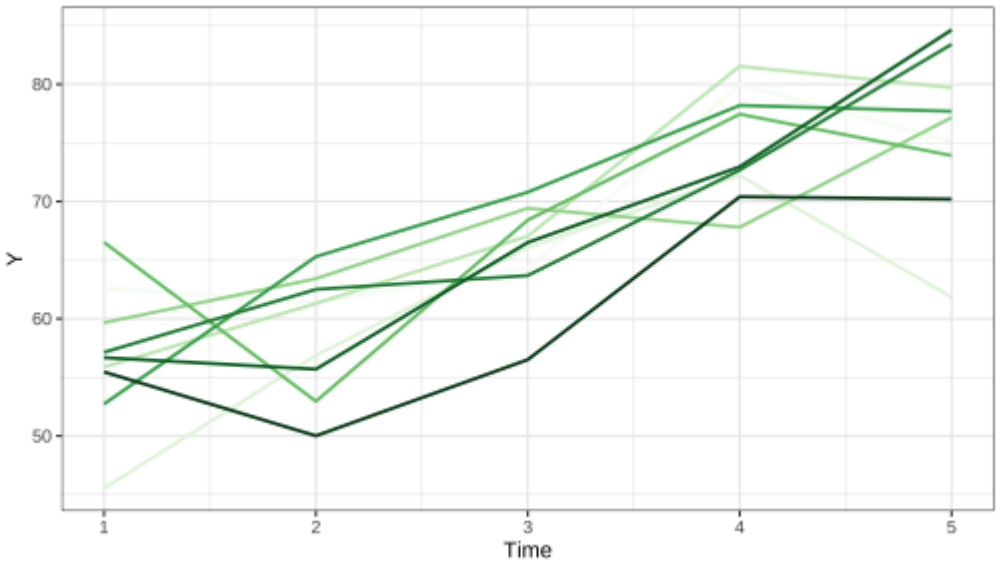
Who Are the Most Informative Subjects?

- Cross-Sectional Outcome
  - Binary Outcome (i.e., relapse/not relapse): the case control study with an equal number of cases and controls is the most efficient design

Who Are the Most Informative Subjects?

- Cross-Sectional Outcome
  - Binary Outcome (i.e., relapse/not relapse): the case control study with an equal number of cases and controls is the most efficient design
  - Continuous Outcome: informative individuals are those with high or low values of the outcome

- Longitudinal Outcome
  - The outcome is repeatedly collected over time. For a subject the outcome is a vector $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$
  - The multiple measures of the outcome allow to separate changes over time within individuals from changes between individuals



- Who the most informative individuals are will depend on whether our interest lies in time-variant covariates or time-invariant covariates

Since individuals have multiple measures of the outcome, the sampling in a two phase ODS is based on a low dimensional summary of the outcome

Since individuals have multiple measures of the outcome, the sampling in a two phase ODS is based on a low dimensional summary of the outcome

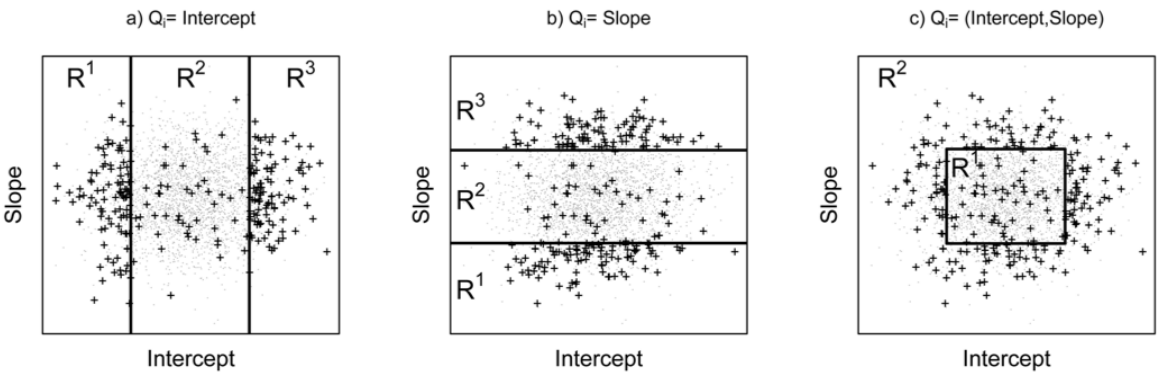
- Binary outcome:
    - Divide the subjects in three groups: those who have never experienced the outcome, those who have always experienced the outcome and those who exhibited response variation.
    - For time-variant covariates the most informative individuals are those who experienced response variation
    - For both time-variant and time-invariant covariates we need to additionally sample individuals who did not report response variation

- Continuous outcome:
  - $E(Y_{ij}) = q_{0i} + q_{1i}t_{ij}$
  - q_{0i} is the subject-specific mean outcome at baseline, q_{1i} is the subject-specific rate of change.
  - For time-variant covariates the most informative individuals are those with extreme values of q_{1i}. For time-invariant covariates the most informative individuals are those with extreme values of q_{0i}

- Continuous outcome:
  - $E(Y_{ij}) = q_{0i} + q_{1i}t_{ij}$
  - q_{0i} is the subject-specific mean outcome at baseline, q_{1i} is the subject-specific rate of change.
  - For time-variant covariates the most informative individuals are those with extreme values of q_{1i}. For time-invariant covariates the most informative individuals are those with extreme values of q_{0i}

How to Select Informative Subjects?

- Sort values of q_{0i} and/or q_{1i} and introduce cutpoints that define sampling strata.



Picture taken directly from Schildcrout et al (2013)

Notation

Data are generated from:

$$f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})dG(\boldsymbol{X}|\boldsymbol{Z})dH(\boldsymbol{Z})$$

- $\boldsymbol{Y}$ is the outcome, $\boldsymbol{X}$ is the expensive covariate and $\boldsymbol{Z}$ is the inexpensive covariate

Goal: Estimate $\boldsymbol{\theta}$

Data are generated from:

$$f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})dG(\boldsymbol{X}|\boldsymbol{Z})dH(\boldsymbol{Z})$$

- $\boldsymbol{Y}$ is the outcome, $\boldsymbol{X}$ is the expensive covariate and $\boldsymbol{Z}$ is the inexpensive covariate

Goal: Estimate $\boldsymbol{\theta}$

- N **:** number of subjects in phase one.

Data are generated from:

$$f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})dG(\boldsymbol{X}|\boldsymbol{Z})dH(\boldsymbol{Z})$$

- $\boldsymbol{Y}$ is the outcome, $\boldsymbol{X}$ is the expensive covariate and $\boldsymbol{Z}$ is the inexpensive covariate

Goal: Estimate $\boldsymbol{\theta}$

- N **:** number of subjects in phase one.
- n_V **:** number of subjects in phase two $(n_V < N)$

Data are generated from:

$$f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})dG(\boldsymbol{X}|\boldsymbol{Z})dH(\boldsymbol{Z})$$

- $\boldsymbol{Y}$ is the outcome, $\boldsymbol{X}$ is the expensive covariate and $\boldsymbol{Z}$ is the inexpensive covariate

Goal: Estimate $\boldsymbol{\theta}$

- N **:** number of subjects in phase one.
- n_V **:** number of subjects in phase two $(n_V < N)$
- R_i is the indicator variable on whether subject i has been sampled for phase two
- $V = \{i : R_i = 1\}$ index set of all subjects sampled for phase two
- $\bar{V} = \{i : R_i = 0\}$ index set of all subjects not sampled for phase two

Data are generated from:

$$f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})dG(\boldsymbol{X}|\boldsymbol{Z})dH(\boldsymbol{Z})$$

- $\boldsymbol{Y}$ is the outcome, $\boldsymbol{X}$ is the expensive covariate and $\boldsymbol{Z}$ is the inexpensive covariate

Goal: Estimate $\boldsymbol{\theta}$

- N **:** number of subjects in phase one.
- n_V **:** number of subjects in phase two $(n_V < N)$
- R_i is the indicator variable on whether subject i has been sampled for phase two
- $V = \{i : R_i = 1\}$ index set of all subjects sampled for phase two
- $\bar{V} = \{i : R_i = 0\}$ index set of all subjects not sampled for phase two
- $\pi(\boldsymbol{Y_i},\boldsymbol{Z_i})$ **:** probability that subject i is sampled for phase two

Existing Methods

Why Don't We Use Standard Methods?

$$f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})dG(\boldsymbol{X}|\boldsymbol{Z})dH(\boldsymbol{Z})$$

- Estimators of $\boldsymbol{\theta}$ based on $f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})$ are generally biased.
- The probability that a subject is observed depends on $\boldsymbol{Y}$

Why Don't We Use Standard Methods?

$$f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})dG(\boldsymbol{X}|\boldsymbol{Z})dH(\boldsymbol{Z})$$

- Estimators of $\boldsymbol{\theta}$ based on $f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})$ are generally biased.
- The probability that a subject is observed depends on $\boldsymbol{Y}$
- Exception: if sampling depends on a binary outcome only, we can use logistic regression

Summary of Methods

Methods that use phase two data only

- Complete data likelihood
- Weighted likelihood
- Semiparametric empirical likelihood (SELE)

Methods that use phase one and phase two data

- Semiparametric likelihood
- Estimated pseudolikelihood
- Pseudoscore estimator
- Maximum estimated likelihood estimator (MELE)
- Maximum likelihood
- Semiparametric maximum likelihood estimator (SMLE)
- Imputation methods

Complete Data Likelihood

- Estimate $\boldsymbol{\theta}$ by explicitly conditioning on a subject being sampled in phase two

$$
\begin{aligned}
L_{C0}(\boldsymbol{\theta}, G) &= \prod_{i \in V} P(Y_i, \boldsymbol{X_i} | \boldsymbol{Z_i}, R_i = 1) \\
&= \prod_{i \in V} \left[\frac{f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i} | \boldsymbol{Z_i}) \pi(Y_i, \boldsymbol{Z_i})}{P(R_i = 1 | \boldsymbol{Z_i}; \boldsymbol{\theta})} \right]
\end{aligned}
$$

Complete Data Likelihood

- Estimate $\boldsymbol{\theta}$ by explicitly conditioning on a subject being sampled in phase two

$$\begin{aligned} L_{C0}(\boldsymbol{\theta}, G) &= \prod_{i \in V} P(Y_i, \boldsymbol{X_i} | \boldsymbol{Z_i}, R_i = 1) \\ &= \prod_{i \in V} \left[\frac{f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i} | \boldsymbol{Z_i}) \pi(Y_i, \boldsymbol{Z_i})}{P(R_i = 1 | \boldsymbol{Z_i}; \boldsymbol{\theta})} \right] \end{aligned}$$

- The scaling factor $P(R_i = 1 | \boldsymbol{Z_i}; \boldsymbol{\theta})$ includes both $\boldsymbol{\theta}$ and $dG(\boldsymbol{X_i} | \boldsymbol{Z_i})$.

Complete Data Likelihood

- Estimate $\boldsymbol{\theta}$ by explicitly conditioning on a subject being sampled in phase two

$$\begin{aligned} L_{C0}(\boldsymbol{\theta}, G) &= \prod_{i \in V} P(Y_i, \boldsymbol{X_i} | \boldsymbol{Z_i}, R_i = 1) \\ &= \prod_{i \in V} \left[\frac{f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i} | \boldsymbol{Z_i}) \pi(Y_i, \boldsymbol{Z_i})}{P(R_i = 1 | \boldsymbol{Z_i}; \boldsymbol{\theta})} \right] \end{aligned}$$

- The scaling factor $P(R_i = 1 | \boldsymbol{Z_i}; \boldsymbol{\theta})$ includes both $\boldsymbol{\theta}$ and $dG(\boldsymbol{X_i} | \boldsymbol{Z_i})$.
- $dG(\boldsymbol{X_i} | \boldsymbol{Z_i})$ needs to be included in the maximisation procedure even if it does not depend on $\boldsymbol{\theta}$.

Weighted Likelihood

$$\prod_{i \in V} \frac{1}{\pi(Y_i, \boldsymbol{Z_i})} f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta})$$

- Robust to model misspecification
- Estimated $\boldsymbol{\theta}$ is unbiased, but it can be inefficient when $\pi(Y_i, \boldsymbol{Z_i})$ are highly variable

Methods that use phase one and phase two data

Methods that require categorical phase one outcome data

- Semiparametric likelihood
- Estimated pseudolikelihood

Methods that use categorical and continuous phase one outcome data

- Pseudoscore estimator
- MELE
- Maximum likelihood
- SMLE
- Imputation methods

Pseudoscore & MELE

$$L(\boldsymbol{\theta}, G) = \prod_{i \in V} f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i} | \boldsymbol{Z_i}) \prod_{j \in \bar{V}} \int_{\mathcal{X}} f(Y_j | \boldsymbol{x}, \boldsymbol{Z_j}) dG(\boldsymbol{x} | \boldsymbol{Z_j})$$

Pseudoscore & MELE

$$L(\boldsymbol{\theta}, G) = \prod_{i \in V} f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i} | \boldsymbol{Z_i}) \prod_{j \in \bar{V}} \int_{\mathcal{X}} f(Y_j | \boldsymbol{x}, \boldsymbol{Z_j}) dG(\boldsymbol{x} | \boldsymbol{Z_j})$$

Pseudoscore Estimator

1) Get the score function $\frac{\partial log L(\boldsymbol{\theta}, G)}{\partial \theta}$

2) Substitute $\boldsymbol{dG(X_i | Z_i)}$ in the score function with its empirical estimate

3) Estimate $\boldsymbol{\theta}$ using iterated reweighted algorithm

Pseudoscore & MELE

$$L(\boldsymbol{\theta}, G) = \prod_{i \in V} f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i} | \boldsymbol{Z_i}) \prod_{j \in \bar{V}} \int_{\mathcal{X}} f(Y_j | \boldsymbol{x}, \boldsymbol{Z_j}) dG(\boldsymbol{x} | \boldsymbol{Z_j})$$

Pseudoscore Estimator

1) Get the score function $\frac{\partial log L(\boldsymbol{\theta}, G)}{\partial \theta}$

2) Substitute $\boldsymbol{dG(X_i|Z_i)}$ in the score function with its empirical estimate

3) Estimate $\boldsymbol{\theta}$ using iterated reweighted algorithm

MELE

1) Substitute $\boldsymbol{dG(X_i|Z_i)}$ in the likelihood with its empirical estimate

2) Estimate $\boldsymbol{\theta}$ using Newton-Rapson

- The pseudoscore estimator can be used when some subjects have zero probability of being sampled

- The pseudoscore estimator can be used when some subjects have zero probability of being sampled
- The pseudoscore estimator and the MELE can accommodate categorical inexpensive covariates
- MELE cannot be extended to continuous inexpensive covariates

- The pseudoscore estimator can be used when some subjects have zero probability of being sampled
- The pseudoscore estimator and the MELE can accommodate categorical inexpensive covariates
- MELE cannot be extended to continuous inexpensive covariates
- The MELE and the pseudoscore estimator are more efficient than methods using phase two data only
- The MELE is slightly less efficient than the pseudoscore estimator

Maximum Likelihood

- The MELE and the pseudoscore estimator are based on approximations of the full likelihood computed using a consistent estimate of $dG(\boldsymbol{X}_i|\boldsymbol{Z}_i)$.
- The MELE and pseudoscore estimator are not fully efficient
- Fully efficient methods estimate $dG(\boldsymbol{X}_i|\boldsymbol{Z}_i)$ and $\boldsymbol{\theta}$ simultaneously

Maximum Likelihood

- The MELE and the pseudoscore estimator are based on approximations of the full likelihood computed using a consistent estimate of $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$.
- The MELE and pseudoscore estimator are not fully efficient
- Fully efficient methods estimate $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$ and $\boldsymbol{\theta}$ simultaneously
- Initially methods that estimate $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$ and $\boldsymbol{\theta}$ simultaneously did not account for inexpensive covariates

$$L(\boldsymbol{\theta}, G) = \prod_{i \in V} f(Y_i|\boldsymbol{X_i};\boldsymbol{\theta}) dG(\boldsymbol{X_i}) \prod_{j \in \bar{V}} \int_{\mathcal{X}} f(Y_j|\boldsymbol{x}) dG(\boldsymbol{x})$$

- Available methods use the EM algorithm or a mixed Newton algorithm to estimate the parameters

Problems with Maximum Likelihood

- If inexpensive covariates are available at phase one, maximum likelihood incurs in efficiency loss: we are not including information about the inexpensive covariates in the estimation of $\boldsymbol{\theta}$
- If sampling for phase two is related to any inexpensive covariate, the maximum likelihood methods do not reflect the correct sampling mechanism and lead to biased results

SMLE

- A fully efficient method that accounts for the presence of categorical and continuous inexpensive covariates and allows phase two sampling to depend on phase one data in any manner

SMLE

- A fully efficient method that accounts for the presence of categorical and continuous inexpensive covariates and allows phase two sampling to depend on phase one data in any manner

$$L(\boldsymbol{\theta}, G) = \prod_{i \in V} f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i} | \boldsymbol{Z_i}) \prod_{j \in \bar{V}} \int_{\mathcal{X}} f(Y_j | \boldsymbol{x}, \boldsymbol{Z_j}) dG(\boldsymbol{x} | \boldsymbol{Z_j})$$

- Approximate $\boldsymbol{dG(X_i|Z_i)}$ using the methods of sieve with B-splines basis
- Use an EM algorithm to estimate $\boldsymbol{\theta}$

SMLE

- A fully efficient method that accounts for the presence of categorical and continuous inexpensive covariates and allows phase two sampling to depend on phase one data in any manner

$$L(\boldsymbol{\theta}, G) = \prod_{i \in V} f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) dG(\boldsymbol{X_i} | \boldsymbol{Z_i}) \prod_{j \in \bar{V}} \int_{\mathcal{X}} f(Y_j | \boldsymbol{x}, \boldsymbol{Z_j}) dG(\boldsymbol{x} | \boldsymbol{Z_j})$$

- Approximate $\boldsymbol{dG(X_i|Z_i)}$ using the methods of sieve with B-splines basis
- Use an EM algorithm to estimate $\boldsymbol{\theta}$
- SMLE is more efficient that maximum likelihood methods and pseudoscore estimator

Imputation Methods

- In a two phase design we have all the information necessary to understand why expensive covariates $\boldsymbol{X}$ are missing:

$$f(\boldsymbol{X_i}|\boldsymbol{Z_i}, \boldsymbol{Y_i}, R_i = 0) = f(\boldsymbol{X_i}|\boldsymbol{Z_i}, \boldsymbol{Y_i}) = f(\boldsymbol{X_i}|\boldsymbol{Z_i}, \boldsymbol{Y_i}, R_i = 1)$$

Imputation Methods

- In a two phase design we have all the information necessary to understand why expensive covariates $\boldsymbol{X}$ are missing:

$$f(\boldsymbol{X_i}|\boldsymbol{Z_i}, \boldsymbol{Y_i}, R_i = 0) = f(\boldsymbol{X_i}|\boldsymbol{Z_i}, \boldsymbol{Y_i}) = f(\boldsymbol{X_i}|\boldsymbol{Z_i}, \boldsymbol{Y_i}, R_i = 1)$$

- Since the missing data mechanism is ignorable, we can impute $\boldsymbol{X}$ for unsampled individuals without accounting for having a biased sample
    - Use available data to construct a model for $f(\boldsymbol{X_i}|\boldsymbol{Z_i}, \boldsymbol{Y_i}, R_i = 0)$
    - Fill-in the missing observations by sampling from the constructed model
    - Repeat the process M times and pool the results together using Rubin's rule

- Two methods are available for cases with longitudinal continuous outcome data and a binary expensive covariate.

$$\frac{P(X_i = 1|\boldsymbol{Z_i}, \boldsymbol{Y_i}, R_i = 0)}{P(X_i = 0|\boldsymbol{Z_i}, \boldsymbol{Y_i}, R_i = 0)} = \frac{\boldsymbol{f(Y_i|X_i = 1, Z_i}, R_i = 1; \boldsymbol{\theta})}{\boldsymbol{f(Y_i|X_i = 0, Z_i}, R_i = 1; \boldsymbol{\theta})} \times \frac{P(X_i = 1|\boldsymbol{Z_i}, R_i = 1)}{P(X_i = 0|\boldsymbol{Z_i}, R_i = 1)}$$

- Two methods are available for cases with longitudinal continuous outcome data and a binary expensive covariate.

$$\frac{P(X_i = 1|\boldsymbol{Z_i}, \boldsymbol{Y_i}, R_i = 0)}{P(X_i = 0|\boldsymbol{Z_i}, \boldsymbol{Y_i}, R_i = 0)} = \frac{\boldsymbol{f(Y_i|X_i = 1, Z_i}, R_i = 1; \boldsymbol{\theta})}{\boldsymbol{f(Y_i|X_i = 0, Z_i}, R_i = 1; \boldsymbol{\theta})} \times \frac{P(X_i = 1|\boldsymbol{Z_i}, R_i = 1)}{P(X_i = 0|\boldsymbol{Z_i}, R_i = 1)}$$

Indirect Approach

- Estimate $\frac{\boldsymbol{f(Y_i|X_i=1,Z_i},R_i=1;\boldsymbol{\theta})}{\boldsymbol{f(Y_i|X_i=0,Z_i},R_i=1;\boldsymbol{\theta})}$ and $\frac{P(X_i=1|\boldsymbol{Z_i},R_i=1)}{P(X_i=0|\boldsymbol{Z_i},R_i=1)}$ separately.
- Need to be tweaked every time we change design

Direct Approach

- Take the logarithm of the equation above and use algebra to find the terms of the imputation model
- Try to relate a time-invariant $\boldsymbol{X}$ with time-variant $\boldsymbol{Y}$
- Need to be tweaked every time we add/remove $\boldsymbol{Z}$

Proposed Work

We aim to develop an imputation method that solves the problems of both direct and indirect imputation

$$\boldsymbol{Y_i^t V_i^{-1}(\mu_{1,i} - \mu_{0,i})} - \frac{1}{2}(\boldsymbol{\mu_{1,i}^T V_i^{-1} \mu_{1,i} - \mu_{0,i}^T V_i^{-1} \mu_{0,i}}) + log\left[\frac{P(X_i = 1|\boldsymbol{Z_i})}{P(X_i = 0|\boldsymbol{Z_i})}\right]$$

where $\boldsymbol{\mu_{x,i}} = E[\boldsymbol{Y_i}|X_i = 1, \boldsymbol{Z_i}]$ and $\boldsymbol{V_i} = Cov(\boldsymbol{Y_i}|X_i, \boldsymbol{Z}_i)$

- The imputation model is an offseted logistic regression

We aim to develop an imputation method that solves the problems of both direct and indirect imputation

$$\boldsymbol{Y_i^t V_i^{-1}(\mu_{1,i} - \mu_{0,i})} - \frac{1}{2}(\boldsymbol{\mu_{1,i}^T V_i^{-1} \mu_{1,i} - \mu_{0,i}^T V_i^{-1} \mu_{0,i}}) + log\left[\frac{P(X_i = 1|\boldsymbol{Z_i})}{P(X_i = 0|\boldsymbol{Z_i})}\right]$$

where $\boldsymbol{\mu_{x,i}} = E[\boldsymbol{Y_i}|X_i = 1, \boldsymbol{Z_i}]$ and $\boldsymbol{V_i} = Cov(\boldsymbol{Y_i}|X_i, \boldsymbol{Z}_i)$

- The imputation model is an offseted logistic regression

Problem: the elements of $\boldsymbol{V}^{-1}$ and $\boldsymbol{\mu}$ are not known

Solution: estimate $\boldsymbol{V}^{-1}$ and $\boldsymbol{\mu}$ and the offset iteratively

1) Fit the linear mixed effects model of interest on sampled subjects only

2) Take the estimated parameters and compute the offset

Solution: estimate $\boldsymbol{V}^{-1}$ and $\boldsymbol{\mu}$ and the offset iteratively

1) Fit the linear mixed effects model of interest on sampled subjects only

2) Take the estimated parameters and compute the offset

3) Fit the logistic imputation model using the offset in calculated 2)

$$\boldsymbol{Y_i^t V_i^{-1}(\mu_{1,i} - \mu_{0,i})} - \frac{1}{2}(\boldsymbol{\mu_{1,i}^T V_i^{-1} \mu_{1,i} - \mu_{0,i}^T V_i^{-1} \mu_{0,i}}) + log\left[\frac{P(X_i = 1|\boldsymbol{Z_i})}{P(X_i = 0|\boldsymbol{Z_i})}\right]$$

4) For unsampled subjects impute their expensive covariate using the results of the model in 3)

Solution: estimate $\boldsymbol{V}^{-1}$ and $\boldsymbol{\mu}$ and the offset iteratively

1) Fit the linear mixed effects model of interest on sampled subjects only

2) Take the estimated parameters and compute the offset

3) Fit the logistic imputation model using the offset in calculated 2)

$$\boldsymbol{Y_i^t V_i^{-1}(\mu_{1,i} - \mu_{0,i})} - \frac{1}{2}(\boldsymbol{\mu_{1,i}^T V_i^{-1} \mu_{1,i} - \mu_{0,i}^T V_i^{-1} \mu_{0,i}}) + log\left[\frac{P(X_i = 1|\boldsymbol{Z_i})}{P(X_i = 0|\boldsymbol{Z_i})}\right]$$

4) For unsampled subjects impute their expensive covariate using the results of the model in 3)

5) Fit the linear mixed effects model of interest on everyone

6) Repeat steps 2) to 5), and burn the first few iterations

7) Combine the results using Rubin's rule

Summary and Future Directions

Summary and Future Directions

- When exposure ascertaiment cost limits the sample size, it is desirable to target a sample of informative subjects. The two phase design aims to sample informative individuals for a specific research question.
- Given a fixed sample size the two phase design leads to higher precision than simple random sampling.
- Introduce likelihood based methods to estimate the parameter of interest that account for phase two data only, and phase one and phase two data. Methods comparison showed that including all available data results in more efficient estimate. Present a new imputation approach

Summary and Future Directions

- When exposure ascertaiment cost limits the sample size, it is desirable to target a sample of informative subjects. The two phase design aims to sample informative individuals for a specific research question.
- Given a fixed sample size the two phase design leads to higher precision than simple random sampling.
- Introduce likelihood based methods to estimate the parameter of interest that account for phase two data only, and phase one and phase two data. Methods comparison showed that including all available data results in more efficient estimate. Present a new imputation approach
- Look more into convergence of our imputation model
- Extend the algorithm to binary longitudinal and other type of exposure
- Code an EM algorithm for parameter estimation under a two phase design with longitudinal data

Thank you!

Selected Reference

[1] Chatterjee et al, A pseudoscore estimator for regression problems with two-phase sampling, JAMA 98 (2003), no. 461, 158-168.

[2] Lawless et al, Semiparametric methods for response-selective and missing data problems in regression, JRSS B Stat. Methodol. 61 (1999), no. 2, 413-438.

[3] Schildcrout et al, Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics, Biometrics 69 (2013), no. 2, 405-16.

[4] Song R et al, A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome, Biometrika 96 (2009), no. 1, 221-228.

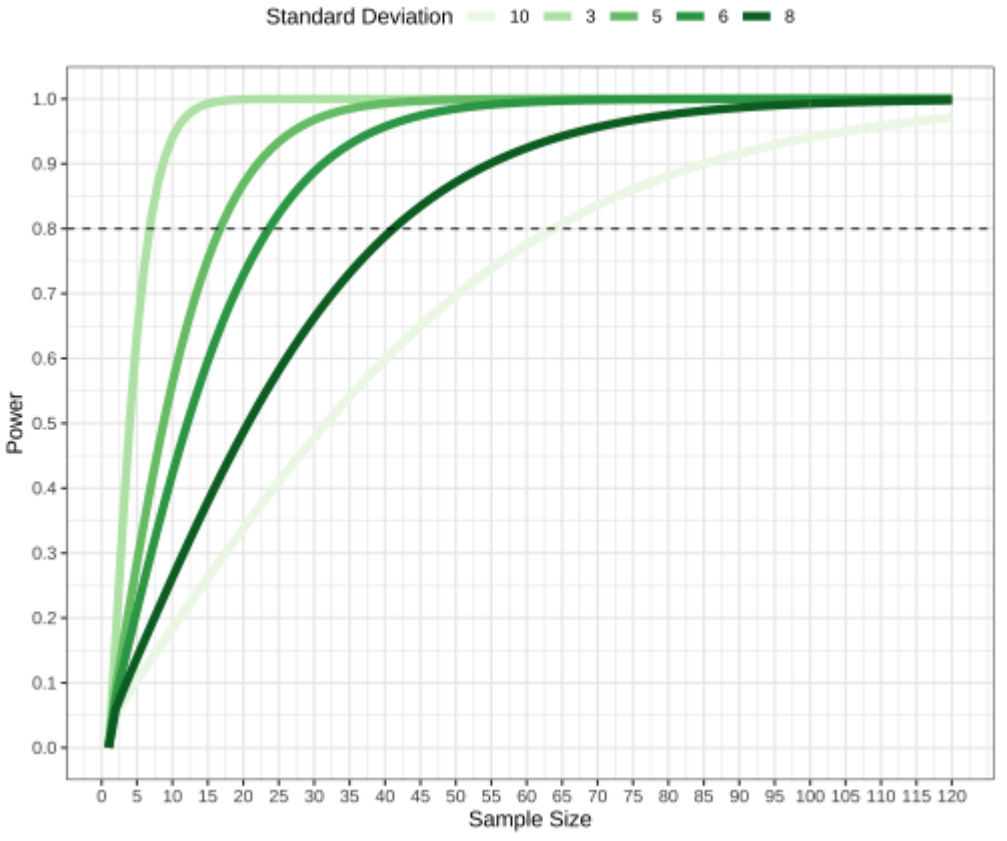
[5] Tao R et al, Efficient Semiparametric Inference Under Two-Phase Sampling With Applications to Genetic Association Studies, JASA 112 (2017), no. 520, 1468-1476.

[6] Weaver et al, An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling, JASA 100 (2005), no. 470, 459-469.

[7] Zhou et al, An efficient sampling and inference procedure for studies with a continuous outcome, Epidemiology 18 (2017), no. 4, 461-468.

Supplementary Slides

Power Curve


Standard Deviation
10
3
5
6
8
Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120
Sample Size

The Scaling Factor

Complete Data Likelihood

- Assume there are no inexpensive covariates $\boldsymbol{Z}$. We consider a continuous outcome $\boldsymbol{Y}$ and a continuous expensive covariate $\boldsymbol{X}$, then

$$\begin{aligned} P(R=1) &= \int\int P(R=1|y,\boldsymbol{x})f(y,\boldsymbol{x})d\boldsymbol{x}dy \\ &= \int\int P(R=1|y)f(y|\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x}dy \end{aligned}$$

Complete Data Likelihood

By conditioning on $\boldsymbol{X}$ the complete data likelihood can be re-written as:

$$\begin{aligned} L_{C1}(\boldsymbol{\theta}, G) &= \prod_{i \in V} P(Y_i | \boldsymbol{Z_i}, \boldsymbol{X_i}, R_i = 1) \\ &= \prod_{i \in V} \left[\frac{f(Y_i | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta}) \pi(Y_i, \boldsymbol{Z_i})}{P(R_i = 1 | \boldsymbol{X_i}, \boldsymbol{Z_i}; \boldsymbol{\theta})} \right] \end{aligned}$$

where

$$P(R = 1 | \boldsymbol{Z}) = \int P(R = 1 | y, \boldsymbol{x}, \boldsymbol{z}) f(y | \boldsymbol{x}, \boldsymbol{z} : \boldsymbol{\theta}) dy$$

Semiparametric empirical likelihood (SELE)

- Assume no inexpensive covariates $\boldsymbol{Z}$ and a continuous outcome $\boldsymbol{Y}$

Semiparametric empirical likelihood (SELE)

- Assume no inexpensive covariates $\boldsymbol{Z}$ and a continuous outcome $\boldsymbol{Y}$
- Group subjects in $\boldsymbol{K}$ strata $(\mathcal{S}_1, \ldots, \mathcal{S}_K)$ based on their outcome $\boldsymbol{Y}$.
- $\boldsymbol{Y_{ki}}$ the outcome for subject $\boldsymbol{i}$ in stratum $\boldsymbol{k}$

Semiparametric empirical likelihood (SELE)

- Assume no inexpensive covariates $\boldsymbol{Z}$ and a continuous outcome $\boldsymbol{Y}$
- Group subjects in $\boldsymbol{K}$ strata $(\mathcal{S}_1, \ldots, \mathcal{S}_K)$ based on their outcome $\boldsymbol{Y}$.
- $\boldsymbol{Y_{ki}}$ the outcome for subject $\boldsymbol{i}$ in stratum $\boldsymbol{k}$

$$\left[\prod_{i \in V} f(Y_i | \boldsymbol{X_i}; \boldsymbol{\theta})\right] \left[\prod_{i \in V} dG(\boldsymbol{X_i})\right] \left[\prod_{k=1}^{K} P(Y_{ki} \in \mathcal{S}_k)^{-n_k}\right]$$

- Similar to the complete data likelihood with the scaling factor being the probability that a subject $\boldsymbol{i}$ is in a specific stratum rather than the probability of being sampled for phase two.

Semiparametric empirical likelihood (SELE)

- Assume no inexpensive covariates $\boldsymbol{Z}$ and a continuous outcome $\boldsymbol{Y}$
- Group subjects in $\boldsymbol{K}$ strata $(\mathcal{S}_1, \ldots, \mathcal{S}_K)$ based on their outcome $\boldsymbol{Y}$.
- $\boldsymbol{Y_{ki}}$ the outcome for subject $\boldsymbol{i}$ in stratum $\boldsymbol{k}$

$$\left[\prod_{i \in V} f(Y_i | \boldsymbol{X_i}; \boldsymbol{\theta})\right] \left[\prod_{i \in V} dG(\boldsymbol{X_i})\right] \left[\prod_{k=1}^{K} P(Y_{ki} \in \mathcal{S}_k)^{-n_k}\right]$$

- Similar to the complete data likelihood with the scaling factor being the probability that a subject $\boldsymbol{i}$ is in a specific stratum rather than the probability of being sampled for phase two.
- $\boldsymbol{\theta}$ is estimated by modelling $\boldsymbol{dG(X)}$ nonparametrically

Categorical phase one data

$$\prod_{j=1}^{K}\left[\prod_{i\in V} f(Y_i|\boldsymbol{X_i},\boldsymbol{Z_i};\boldsymbol{\theta})dG(\boldsymbol{X_i}|\boldsymbol{Z_i})\right] Q_k(\boldsymbol{\theta},G)^{N_k-n_{V_k}}$$

where $Q_k(\boldsymbol{\theta},G)^{N_k-n_{V_k}} = pr\{(Y,\boldsymbol{X},\boldsymbol{Z}\in\mathcal{S}_k)\}$

Estimated pseudolikelihood

1) Substitute $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$ with a consistent estimate based on the empirical conditional distribution function of $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$

2) Use Newton-Rapson to estimate $\boldsymbol{\theta}$

Categorical phase one data

$$\prod_{j=1}^{K}\left[\prod_{i\in V} f(Y_i|\boldsymbol{X_i},\boldsymbol{Z_i};\boldsymbol{\theta})dG(\boldsymbol{X_i}|\boldsymbol{Z_i})\right] Q_k(\boldsymbol{\theta},G)^{N_k-n_{V_k}}$$

where $Q_k(\boldsymbol{\theta},G)^{N_k-n_{V_k}} = pr\{(Y,\boldsymbol{X},\boldsymbol{Z}\in\mathcal{S}_k)\}$

Estimated pseudolikelihood

1) Substitute $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$ with a consistent estimate based on the empirical conditional distribution function of $dG(\boldsymbol{X_i}|\boldsymbol{Z_i})$

2) Use Newton-Rapson to estimate $\boldsymbol{\theta}$

Semiparametric likelihood

1) Fix $\boldsymbol{\theta}$ and solve for an empirical likelihood estimate $\widehat{dG}(\boldsymbol{X}|\boldsymbol{Z})$ from a constrained likelihood function assuming that $dG(\boldsymbol{X}|\boldsymbol{Z})$ is a probabability mass function over $\boldsymbol{X}$

2) Plug $\widehat{dG}(\boldsymbol{X}|\boldsymbol{Z})$ into the likelihood

3) Use Newton-Rapson to estimate $\boldsymbol{\theta}$

Iterated Reweighted Algorithm

Let $\boldsymbol{\theta}^{(m-1)}$ be the value of the parameter at step $m-1$, then at step m:

1) For each subject j not sampled, build the filled-in data $\{(\boldsymbol{Y_j}, \boldsymbol{X_i}, \boldsymbol{Z_j})\}$ using all observed combinations of $(\boldsymbol{X_i}, \boldsymbol{Z_i})$ with $\boldsymbol{Z_i} = \boldsymbol{Z_j}$

2) For each filled-in observation $\{(\boldsymbol{Y_j}, \boldsymbol{X_i}, \boldsymbol{Z_j})\}$ calculate its associate weight:

$$\omega_{ij}\left(\boldsymbol{\theta}^{(m-1)}\right) = \frac{h^{\hat{\pi}}_{\theta^{(m-1)}}(\boldsymbol{Y_j}, \boldsymbol{X_i}, \boldsymbol{Z_j})}{\sum_l h^{\hat{\pi}}_{\theta^{(m-1)}}(\boldsymbol{Y_j}, \boldsymbol{X_l}, \boldsymbol{Z_j})}$$

where $h^{\hat{\pi}}_{\theta^{(m-1)}} \frac{f(Y_j|\boldsymbol{X_l},\boldsymbol{Z_j};\boldsymbol{\theta})}{P(R=1|\boldsymbol{X_l},\boldsymbol{Z_j})}$

3) Obtain a new estimate $\boldsymbol{\theta}^{(m)}$ by fitting a parametric regression model. Assign weight 1 to subjects sampled in phase two and $\omega_{ij}\left(\boldsymbol{\theta}^{(m-1)}\right)$ to those not sampled in phase two

4) Repeat 2) and 3) until convergence

Secondary Analysis

- Researchers might want to re-use data from a two phase ODS design to study the association between a secondary outcome and the expensive exposure
- To perform valid inference, analysis of a secondary outcome needs to account for the biased nature of the sample

Estimating Equation

- Solving the estimating equation

$$\sum_{i \in V} \frac{1}{\hat{\pi}_i} \left(\frac{\partial \boldsymbol{X}}{\partial \boldsymbol{\theta}} \right)^T \hat{\boldsymbol{Q}}^{-1} (\boldsymbol{Y_i} - \boldsymbol{X_i \theta})$$

- The equation can be extended to include non sampled subjects

Multivariate Outcome

- Analysis of secondary outcome can be done using the methods discussed with a bivariate outcome $\boldsymbol{(Y_1, Y_2)}$ where $\boldsymbol{Y_1}$ is the outcome used for two phase ODS and $\boldsymbol{Y_2}$ is the secondary outcome

Data Augmentation Results

N = 2,000 subjects in phase one, and n_V = 400 subjects in phase two

$$Y_{1ij} = \beta_{10} + \beta_{11} snp_i + \beta_{12} c_i + \beta_{13} t_{ij} + \beta_{14} snp_i t_{ij} + \beta_{15} c_i t_{ij} + a_{1i} + b_{1i} t_{ij} + \epsilon_{1ij}$$

$$Y_{2ij} = \beta_{20} + \beta_{21} snp_i + \beta_{22} c_i + \beta_{23} t_{ij} + \beta_{24} snp_i t_{ij} + \beta_{25} c_i t_{ij} + a_{2i} + b_{2i} t_{ij} + \epsilon_{2ij}$$

Data Augmentation Results

N = 2,000 subjects in phase one, and n_V = 400 subjects in phase two

$$Y_{1ij} = \beta_{10} + \beta_{11} snp_i + \beta_{12} c_i + \beta_{13} t_{ij} + \beta_{14} snp_i t_{ij} + \beta_{15} c_i t_{ij} + a_{1i} + b_{1i} t_{ij} + \epsilon_{1ij}$$

$$Y_{2ij} = \beta_{20} + \beta_{21} snp_i + \beta_{22} c_i + \beta_{23} t_{ij} + \beta_{24} snp_i t_{ij} + \beta_{25} c_i t_{ij} + a_{2i} + b_{2i} t_{ij} + \epsilon_{2ij}$$

$c_i \sim N(-0.15 - 0.05 snp_i, 1)$ is a continuous variable measured at baseline and $P(snp_i = 1) = 0.3$

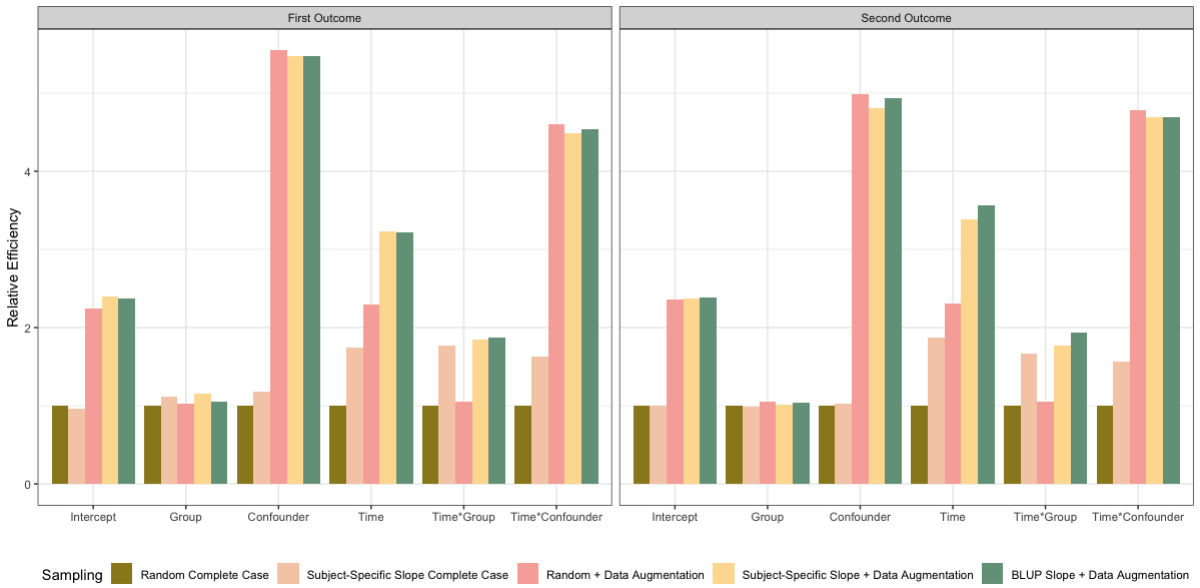
$(\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}) = (80, 0.5, -2.5, -1.5, -0.25, -0.10)$;

$(\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25}) = (65, -0.6, -2, -1, -0.15, -0.15)$.

The random effects $(b_{01i}, b_{11i}, b_{02i}, b_{12i})$ were generated from a multivariate normal distribution with mean 0 and unique elements of the variance and covariance matrix $\boldsymbol{\sigma}$ = (20.25, 0.25, 7.50, 0.125, 1, 0.75, 0.250, 9, 0.375, 0.25).

The error components were normally distributed with mean 0 and variance $\boldsymbol{\Sigma}_i = (\sigma_1^2 \boldsymbol{I}, \sigma_2^2 \boldsymbol{I})$ with $\sigma_1^2 = 2.25$ and $\sigma_2^2 = 1$.

Relative efficiency of multiple sampling designs compared to simple random sampling



EM Algorithm

Using sampled subjects fit the linear mixed effects model of interest and estimate the coefficients $\boldsymbol{\theta}^{(0)}$ and the variance components $\boldsymbol{\alpha}^{(0)}$

At the m^{th} iteration:

a) use $\left(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\alpha}^{(m-1)}\right)$ to calculate the offset for the conditional exposure log-odds model $\texttt{offset}^{(m)}$

b) On sampled subjects fit $logit(pr(x_i = 1|\boldsymbol{y_i}, \boldsymbol{z_i})) = \gamma \boldsymbol{Z} + \texttt{offset}^{(m)}$ to estimate γ. Using this values calculate $pr^{(m)}(x_i = 1|\boldsymbol{y_i}, \boldsymbol{z_i}))$ and $pr^{(m)}(x_i = 0|\boldsymbol{y_i}, \boldsymbol{z_i}))$

c) Calculate the estimated/expected log-likelihood:

$$
\begin{aligned}
l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = & \sum_{i \in V} l_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\
& + \sum_{i \in \bar{V}} l_i(\boldsymbol{\beta}, \boldsymbol{\alpha}; g_i = 1) pr^{(m)}(x_i = 1 | \boldsymbol{y_i}, \boldsymbol{z_i})) \\
& + \sum_{i \in \bar{V}} l_i(\boldsymbol{\beta}, \boldsymbol{\alpha}; g_i = 0) pr^{(m)}(x_i = 0 | \boldsymbol{y_i}, \boldsymbol{z_i}))
\end{aligned}
$$

d) Maximise the estimated log-likelihood

e) Repeat a) to f) until convergence