Analytical Approaches to Targeted Sub-sampling Designs with Longitudinal Continuous Response Data

Jonathan S. Schildcrout
Department of Biostatistics
Vanderbilt University

June 3, 2014

Motivation

- Availability of administrative databases, cohort study data, electronic medical records data is on the rise.
- These resource could be used to address novel study hypotheses.
- Often, we need to collect an exposure or confounder and ascertainment costs limit sample size.
  - Analysis of blood samples required for biomarker research
  - Manual chart reviews required for EMR research
  - Recontacting patients and additional clinic visits may be required for cohort studies

Motivation (cont.)

- Genetic determinants of statin effectiveness.
  - Longitudinal lipids (LDL) data on 1000s of patients on statins for years.
  - Most effective drug / dose required to lower LDL to normal range is unknown
  - It is common to study genetic determinants but this is still expensive.
    - Q1: If we can analyze blood samples on a subset of individuals, who should we choose?
    - Q2: Once we pick the (biased) sample, how do we analyze the data so that we can generalize results to the entire population.
- Designs discussed today require retrospective exposure ascertainment
- Interest is in a continuous, longitudinal outcome that is also used to develop the sampling scheme
- Similar in spirit to other epidemiological designs (e.g., case-control, case-cohort) because who we observe depends on response values.

Childhood Asthma Management Program (CAMP)

- Examined long term effects of anti-inflammatory meds on lung growth in children with mild to moderate asthma.
- 1041 children in eight cities were randomized to one of two anti-inflammatory medications or to placebo.
- A primary aim was to compare lung function at the end of the study period.
- For the primary endpoint, there was no observed treatment effect

CAMP (cont.)

- Genetic ancillary substudies
  - Sought to examine genetic factors for asthma severity and lung function
  - To conduct such analyses, genotype data were ascertained retrospectively ⇒ additional costs.
  - Obtained genetic data for inflammatory cytokines in nearly all kids.
    - ★ Only 555 kids data are available for loci of the IL-10 cytokine.
- In other studies, retrospective ascertainment of a key exposure would limit sample size.
  - Outcome dependent sampling (ODS) designs are known to be highly efficient relative to random sampling designs.

CAMP Analysis

- Analytical goal
  - To examine *FVC* trajectories over the course of 4 years of followup for those with and without at least one variant allele on a locus of the IL 10 cytokine.
- The target, population model
  - Linear mixed effects model that includes
    - Fixed effects: time since randomization, presence/absence of the variant allele, their interaction, and potential confounders.
    - Random Effects: intercept and slope for time
- We use the CAMP data to evaluate study designs / estimation procedures
  - *FVC* and covariate data are available for 555 kids
  - Genotype data are expensive to ascertain ⇒ sample size limited to $\sim$ 250 children

The approach...

- Subsample individuals from a cohort based on features of the response vectors.
    - Calculate summary statistics from the response vectors
    - Use summary statistics to define sampling strata
    - Conduct a stratified sampling approach
- Conduct statistical analyses that acknowledge the biased sampling design
    - Ascertainment corrected (conditional) maximum likelihood
    - MI extensions that use unsampled subject
- Design combined with analysis procedures can be highly efficient compared to standard designs.

Outline

- Population model
- A class of ODS designs for continuous, longitudinal data
- Analysis
  - Ascertainment corrected maximum likelihood
  - Extension from ACML to MI
  - Direct MI
- Relative efficiency of designs/estimation procedures via simulations
- CAMP data
- Summary

The population model

- N subjects in the original cohort (representative of the target population)
- The random intercept and slope linear mixed effect model of Laird and Ware (1982).

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \tag{1}$$

- $\mathbf{X}_i$: $n_i \times p$ design matrix for the fixed effects,
- $\boldsymbol{\beta}$: $p-$vector of fixed-effect coefficients
- $\mathbf{Z}_i$: $n_i \times q$ design matrix for the random effects.
- $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$
- $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma})$: we assume $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_{n_i}$

The population model for CAMP

- Fixed effects: $\mathbf{X}_i = [\mathbf{1},\ \mathbf{T}_i,\ \mathbf{X}_{ei},\ \mathbf{T}_i\mathbf{X}_{ei},\ \mathbf{X}_{oi}]$
  - $\mathbf{T}_i = \{T_{ij}\}_{j \in 1,2,\ldots n_i}$: vector of times subject i was observed
  - $\mathbf{X}_{ei}$: expensive, time-invariant target variable
  - $\mathbf{X}_{oi}$: matrix of pre-existing / inexpensive confounders
- $\mathbf{Z}_i = [\mathbf{1},\ \mathbf{T}_i]$
- $\mathbf{b}_i = (b_{0i}, b_{1i}) \sim N_2(\mathbf{0}, \mathbf{D})$
  - $\mathbf{D}_i$ is the 2×2 covariance matrix that contains the variance components (σ_0^2, σ_1^2) along the diagonal, and the covariance $\rho \cdot \sigma_0 \cdot \sigma_1$ in the off diagonal.

$$Y_{ij} = \beta_0 + \beta_t t_{ij} + \beta_e x_{ei} + \beta_{te} t_{ij} x_{ei} + \mathbf{x}_{oij}\boldsymbol{\beta}^o + b_{0i} + b_{1i} t_{ij} + e_{ij}$$

The population model for CAMP

- The multivariate density for this model can be written,

$$f(\mathbf{Y}_i \mid \mathbf{X}_i; \boldsymbol{\theta}) = (2\pi)^{-n_i/2} \left|\mathbf{V}_i\right|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu}_i)^t \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)\right\}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_0, \sigma_1, \rho)$, $\boldsymbol{\mu}_i = \mathbf{X}_i \boldsymbol{\beta}$, $\mathbf{V}_i = \mathbf{Z}_i \mathbf{D}_i \mathbf{Z}_i^t + \sigma^2 \mathbf{I}$.

- With a random / representative sample of N_s subjects, inferences could be made by maximizing the log-likelihood

$$l(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{N_s} l_i(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i) = \sum_{i=1}^{N_s} log \; f(\mathbf{Y}_i \mid \mathbf{X}_i; \boldsymbol{\theta}).$$

A class of ODS designs

- X_{ei} is expensive: can only collect it on a subset of subjects.
- Subsample individuals based on features or a summary of their available data: Q_i.
- We will discuss, $Q_i = \mathbf{W}_i \mathbf{Y}_i$ (linear in the response).
  - $\mathbf{W}_i = \frac{1}{n_i}\mathbf{1}$, Q_i is an average.
  - $\mathbf{X}_{Ti} = (\mathbf{1}, \mathbf{T}_i)$ and $\mathbf{W}_i = (\mathbf{X}_{Ti}^t \mathbf{X}_{Ti})^{-1} \mathbf{X}_{Ti}^t$,
    - $\mathbf{Q}_i$: intercept and slope of subject i's regression of $\mathbf{Y}_i$ on $\mathbf{T}_i$
    - $\mathbf{Q}_i[1]$: intercept
    - $\mathbf{Q}_i[2]$: slope

A class of ODS designs (cont.)

- Choice of Q_i determines the parameters that are estimated efficiently.
- Split the distribution of Q_i into regions
- Conduct a stratified sampling procedure s.t.

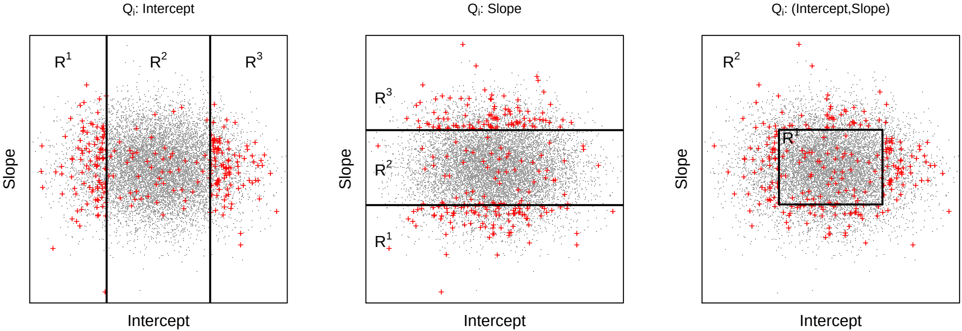
$$pr(S_i = 1 \mid \mathbf{Y}_i, \mathbf{X}_i) = pr(S_i = 1 \mid q_i \in R^k) = \pi(q_i \in R^k)$$

- In the univariate Q_i case,

$$\pi(q_i) = \begin{cases} \pi(q_i \in R^1), & q_i \leq k_1 \\ \pi(q_i \in R^2), & k_1 < q_i \leq k_2 \\ \pi(q_i \in R^3), & q_i > k_2. \end{cases}$$

  Oversampling towards the extremes of $Q_i \rightarrow$ efficiency improvements.
- This also applies to bivariate $\mathbf{Q}_i$

ODS designs based on subject-specific linear regressions



- Oversample relatively 'informative' subjects for the estimation targets.
- Choice of Q_i is a reflection of who you think is informative

Analyses that acknowledge the ODS designs

- We observe a biased sample.
- How to analyze the data so that inferences generalize?
  - Ascertainment corrected likelihoood
  - MI extensions

An ascertainment corrected likelihood

- If $f(\mathbf{Y}_i \mid \mathbf{X}_i; \boldsymbol{\theta})$ is the MV density for subject i under random sampling from a population, a density for those who are included in the ODS is given by

$$\begin{aligned} f(\mathbf{Y}_i \mid \mathbf{X}_i, S_i = 1; \boldsymbol{\theta}) &= \frac{\pi(q_i) f(\mathbf{Y}_i \mid \mathbf{X}_i; \boldsymbol{\theta})}{pr(S_i = 1 \mid \mathbf{X}_i; \boldsymbol{\theta})} \\ &= \frac{\pi(q_i) f(\mathbf{Y}_i \mid \mathbf{X}_i; \boldsymbol{\theta})}{\sum_{k=1}^{K} \pi(q_i \in R^k) pr(q_i \in R^k \mid \mathbf{X}_i; \boldsymbol{\theta})} \end{aligned}$$

where q_i is subject i's observed value of the sampling variable Q_i.

An ascertainment corrected likelihood

- If a total of N_s subjects are selected into the ODS, the ascertainment corrected log-likelihood is given by,

$$l^C(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{N_s} \left[l_i(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i) - log \left\{ \underbrace{\sum_{k=1}^{K} \pi(q_i \in R^k) \int_{R^k} f(q_i \mid \mathbf{X}_i; \boldsymbol{\theta}) dq_i}_{AC_i} \right\} \right],$$

  where AC_i is an ascertainment correction.

- Carroll et al (1995) and Lawless et al (1999) refer to this conditional likelihood as the complete data (CD) likelihood
- We are not exploiting the incomplete data from subjects in whom X_{ei} was not observed.

An ascertainment correct likelihood

- Since $Q_i = \mathbf{W}_i \mathbf{Y}_i$ is a linear function of the response profile.

$$\mathbf{Y}_i \mid \mathbf{X}_i \sim N(\boldsymbol{\mu}_i, \mathbf{V}_i) \Rightarrow Q_i \mid \mathbf{X}_i \sim N(\boldsymbol{\mu}_{qi}, \Sigma_{qi})$$

  where $\boldsymbol{\mu}_{qi} = \mathbf{W}_i \boldsymbol{\mu}_i$ and $\Sigma_{qi} = \mathbf{W}_i \mathbf{V}_i \mathbf{W}_i^t$.

- If Q_i is univariate, $\boldsymbol{\mu}_{qi} = \mu_{qi}$ and $\Sigma_{qi} = \sigma_{qi}^2$, and we can write

$$AC_i = \sum_{k=1}^{K} \pi(q_i \in R^k) \left\{ F_{Q_i|\mathbf{X}_i}(k_k) - F_{Q_i|\mathbf{X}_i}(k_{k-1}) \right\}$$

  where $F_{Q_i|\mathbf{X}_i}(c)$ is the cumulative distribution function.

Ascertainment Corrected Maximum Likelihood Estimation

- Score equation for the ascertainment corrected likelihood ,

$$\frac{\partial l_i^c}{\partial \boldsymbol{\theta}} = \frac{\partial l_i}{\partial \boldsymbol{\theta}} - \frac{\partial AC_i}{\partial \boldsymbol{\theta}} \cdot [AC_i]^{-1}$$

  where $\frac{\partial l_i}{\partial \boldsymbol{\theta}}$ is the derivative of the standard log-likelihood.

- For univariate Q_i, score function is given by the equations

$$\frac{\partial l_i^c}{\partial \boldsymbol{\beta}} = \frac{\partial l_i}{\partial \boldsymbol{\beta}} + \left[\frac{1}{\sigma_{qi}} \frac{\partial \boldsymbol{\mu}_{qi}}{\partial \boldsymbol{\beta}} \sum_{k=1}^{K-1} \left\{ \pi(R^{k+1}) - \pi(R^k) \right\} \cdot \phi\left(\frac{k_k - \mu_{qi}}{\sigma_{qi}} \right) \right] \cdot [AC_i]^{-1}$$

  for fixed effect parameters $\boldsymbol{\beta}$ and

$$\frac{\partial l_i^c}{\partial \alpha_m} = \frac{\partial l_i}{\partial \alpha_m} + \left[\frac{1}{\sigma_{qi}^2} \frac{\partial \sigma_{qi}}{\partial \alpha_m} \sum_{k=1}^{K-1} \left\{ \pi(R^{k+1}) - \pi(R^k) \right\} (k_k - \mu_{qi}) \cdot \phi\left(\frac{k_k - \mu_{qi}}{\sigma_{qi}} \right) \right] \cdot [AC]^{-1}$$

  for variance component parameters in $\boldsymbol{\alpha} = (\sigma_0, \sigma_1, \rho)$

Ascertainment Corrected Maximum Likelihood Estimation

- Since all parameters in $\boldsymbol{\alpha}$ are subject to constraints e.g., variance components and variances must be positive and ρ must fall within $[-1, 1]$, for our analyses, we transform $\boldsymbol{\alpha}$ and estimate the following parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_\rho) = (\mathit{log}(\sigma_0), \mathit{log}(\sigma_1), \mathit{log}\{(1-\rho)/(1+\rho)\})$.

Multiple Imputation approaches

- The ACML approach does not exploit any of the available data from those in whom X_{ei} was not sampled.
- MI approaches may be able to recover some of that information
- Two MI approaches
  - an extension of the CD analysis
  - a direct MI approach
- To conduct imputation, we need the model for $[x_{ei} \mid \mathbf{x}_{oi}, \mathbf{y}_i, S_i = 0]$ but from our design, we know

$$\begin{aligned} pr(x_{ei} \mid \mathbf{x}_{oi}, \mathbf{y}_i, S_i = 0) &= pr(x_{ei} \mid \mathbf{x}_{oi}, \mathbf{y}_i, S_i = 1) \\ &= pr(x_{ei} \mid \mathbf{x}_{oi}, \mathbf{y}_i). \end{aligned}$$

so it's not that bad.

Multiple Imputation approaches (cont): Binary X_{ei}

- Complete Data Analysis then MI (CD+MI)
  - Conduct the CD analysis using ACML
  - Use results to build the model $pr(x_{ei} \mid \mathbf{x}_{oi}, \mathbf{y}_i, S_i = 0)$
  - General location family of imputation models
- Direct MI (D-MI)
  - Directly impute X_{ei} by building the model from the sampled subjects
  - Imputation by chained equations
- Why not always do D-MI?
  - Decision should be made on what you believe you can do well.
  - If using MI, the MI model needs to be correct.
    - With D-MI, we are attempting to impute a time-invariant exposure with longitudinal data.
    - Can be non-trivial if distributional assumptions not correct
  - If assumptions are correct, and with balanced and complete data, D-MI imputation model has the same form as linear or quadratic discriminant analysis.

Multiple Imputation approaches (cont): Binary X_{ei}

- CD+MI approach:

$$\frac{pr(X_{ei} = 1 \mid \mathbf{x}_{oi}, \mathbf{y}_i, S_i = 0)}{pr(X_{ei} = 0 \mid \mathbf{x}_{oi}, \mathbf{y}_i, S_i = 0)} = \frac{f(\mathbf{y}_i \mid X_{ei} = 1, \mathbf{x}_{oi}, S_i = 1)}{f(\mathbf{y}_i \mid X_{ei} = 0, \mathbf{x}_{oi}, S_i = 1)} \cdot \frac{pr(X_{ei} = 1 \mid \mathbf{x}_{oi}, S_i = 1)}{pr(X_{ei} = 0 \mid \mathbf{x}_{oi}, S_i = 1)}.$$

- First term on rhs comes directly from ACML analysis
- Second term is not totally simple because we've done biased sampling which can induce unintuitive relationships

$$\frac{pr(X_{ei} = 1 \mid \mathbf{x}_{oi}, S_i = 1)}{pr(X_{ei} = 0 \mid \mathbf{x}_{oi}, S_i = 1)} = \frac{pr(S_i = 1 \mid X_{ei} = 1, \mathbf{x}_{oi})}{pr(S_i = 1 \mid X_{ei} = 0, \mathbf{x}_{oi})} \cdot \frac{pr(X_{ei} = 1 \mid \mathbf{x}_{oi})}{pr(X_{ei} = 0 \mid \mathbf{x}_{oi})}.$$

- First term on rhs comes from the ACML analysis
- It can be used as an offset in an offsetted logistic regression analysis

Recap...

- Described a class of very simple ODS study designs for longitudinal continuous response data
  - Summarize response vector or profile based on key features.
  - Split Q_i into regions (coarsening) and sample with equal probability within each region
- Described a relatively simple ascertainment corrected ML approach to estimation and two MI extensions
- Expectation is that sampling towards the extremes of the distribution should lead to efficiency gains

Data generating model

- N=750 subjects with $n_i = n = 10$ observations.
- Population model (at time j):

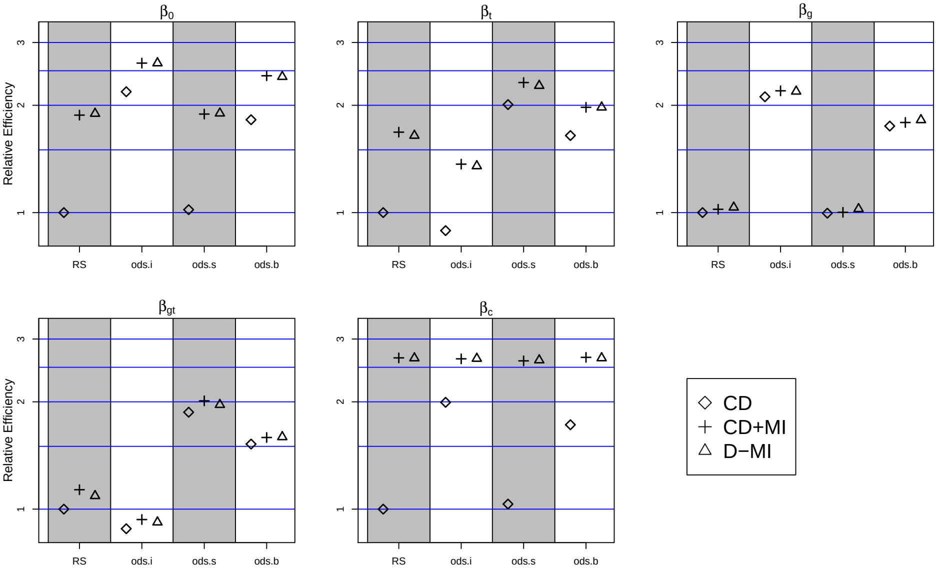
$$Y_{ij} = \beta_0 + \beta_t t_{ij} + \beta_g g_i + \beta_{gt} g_i t_{ij} + \beta_c c_i + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}. \qquad (2)$$

- $\mathbf{t}_i = \{t_{i1}, \ldots t_{in_i}\}$: equally spaced times ranging from -2 to 2.
- C_i : binary with $pr(C_i = 1) = 0.5$.
- G_i : binary with $pr(G_i = 1 | C_i = 1) = 0.4 + 0.15c_i$. X_{ei} from before.
- $(\beta_0, \beta_g, \beta_t, \beta_{gt}, \beta_c) = (5, -2.5, 1.0, 0.75, 1)$
- $\mathbf{b}_i = (b_{0i}, b_{1i})^t \sim N(\mathbf{0}, \mathbf{D})$ with variance components $(\sigma_0^2 = 5, \sigma_1^2 = 1)$ and with correlation parameter $\rho = 0$.
- $\epsilon_{ij} \sim N(0, \sigma^2)$ with σ set to 5.

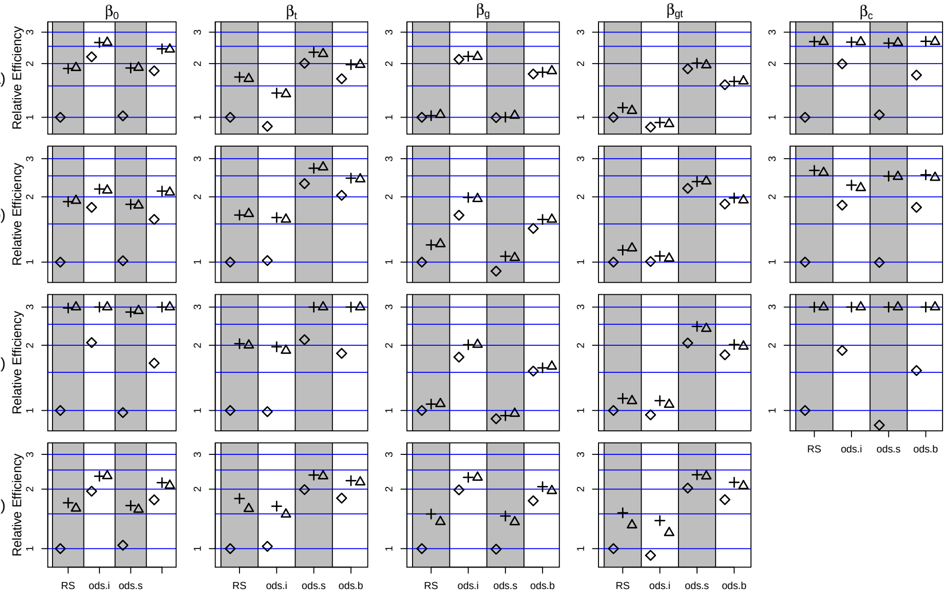
Study Designs and Sampling

- Sample approximately 250 individuals from the original cohort into the proposed substudy.
- Consider 4 total designs
  - Random sampling (RS, standard ML analysis) and three ODS designs based on Q_i :
    1. Intercept: *ods.i*
    2. Slope: *ods.s*
    3. Intercept and slope: *ods.b*
  - Each design is analyzed with either a CD analysis or two MI analyses
  - 12 design by analysis procedure combinations
  - Sample $\sim$ 70 from the central region and $\sim$ 180 from outlying regions.
  - Regions are defined so that $pr(S_i = 1) = 1$ if in the outlying region.

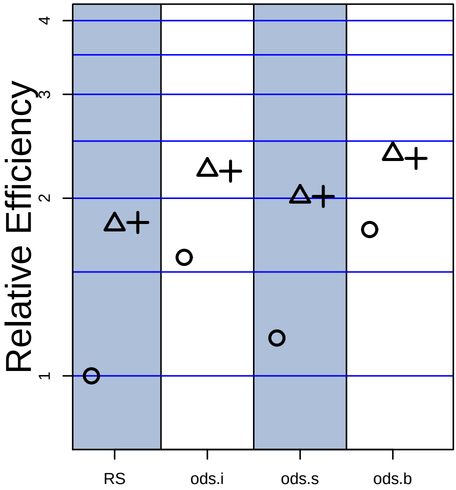
Parameter Estimation Efficiency



Parameter Estimation Efficiency



End of Study Predicted Value Efficiency



CAMP

Table: Demographic Characteristics of the CAMP Study Cohort. Continous variables are summarized with the $10^{th}, 50^{th}, 90^{th}$ percentiles, and categorical variables with proportions.

Variable	Summary
Cohort size (N)	555
Age at randomization (years)	6.23, 8.81, 11.71
Male gender	0.65
Black race	0.10
Other (non-caucasian) race	0.26
Randomized treatment	
Placebo	0.50
Budesonide	0.32
Nedocromil	0.17
IL-10 Variant Allele	0.50
Observations per subject	9, 10, 10
Follow-up time (years)	3.85, 3.99, 4.1
Post BD Percent Predicted	92, 105, 116

CAMP: Summary from 100 replicates

Variable	RS		ods.s		ods.i	
	CD	CD+MI	CD	CD+MI	CD	CD+MI
Intercept	104.95	105.15	105.90	105.19	104.45	105.28
	(2.07)	(1.59)	(2.00)	(1.57)	(1.60)	(1.41)
Time (per year)	0.12	0.09	0.15	0.10	-0.05	0.10
	(0.23)	(0.19)	(0.17)	(0.16)	(0.22)	(0.18)
IL10 SNP	-1.50	-1.68	-2.07	-2.00	-1.72	-1.96
	(1.71)	(1.71)	(1.63)	(1.62)	(1.30)	(1.30)
Time by IL10	-0.34	-0.29	-0.33	-0.31	-0.36	-0.31
	(0.33)	(0.31)	(0.24)	(0.24)	(0.30)	(0.29)
...						
Male (vs female)	-0.98	-1.19	-1.46	-1.12	-0.63	-1.16
	(1.08)	(0.73)	(1.07)	(0.72)	(0.85)	(0.72)

Summary

- Increasingly we are in situations where we may have some but not all of the data needed to address a question... we need to collect the missing pieces of data... this is expensive!
- With limited study resources, efficient designs are crucial.
- Our designs sample based on features of the response vector Q_i.
- Circumstances for which we gain efficiency makes sense
- Efficiency improvements can be very large.
- Limitations/future work
  - Sensitivity to different flavors of misspecification (MI model, likelihood)
  - Extensions to
    - Different response distributions
    - Multivariate longitudinal data
    - Sampling on an auxiliary variable dynamically
    - Sampling adaptively (i.e., altering the study design after doing interim analyses)
    - etc.