Multiple Imputation of an Expensive Covariate in Outcome Dependent Sampling Designs for Longitudinal Data

Chiara Di Gravio, Ran Tao, Jonathan S Schildcrout

Vanderbilt University
Virtual ENAR, 2020

March 25, 2020

Outline

Motivation

- In longitudinal studies when exposure ascertainment costs limit sample size, it is desirable to target a sample of informative subjects
- Different methods have been developed to efficiently select patients for exposure ascertainment
- Analysis is usually done using only subjects in whom exposure was collected, or combining partial data on those not sampled and complete data on those sampled using full likelihood approaches or multiple imputation

Motivation

- For today, we focus on multiple imputation (MI)
- MI could be more **efficient** than conditional likelihood analysis, and often **easier to implement** than full likelihood approaches

Lung Health Study

- Lung Health Study (LHS) data, a multi-center RCT of smokers with mild chronic obstructive pulmonary disease
- We focus on a single SNP found to be a modifier of lung function decline. This is the expensive exposure.
- We are interested in the association between SNP and $FEV\%$, and how the association changes over time
- We consider a scenario in which phenotype and covariate data are available on all subjects but resource constraints only permit SNP to be collected in 20% of the subjects.

Lung Health Study

- The mixed effects model used for our analyses is:

$$Y_{ij} = \beta_0 + \beta_s SNP_i + \beta_t t_{ij} + \beta_{st} SNP_i t_{ij} + \beta_c c_i + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}$$

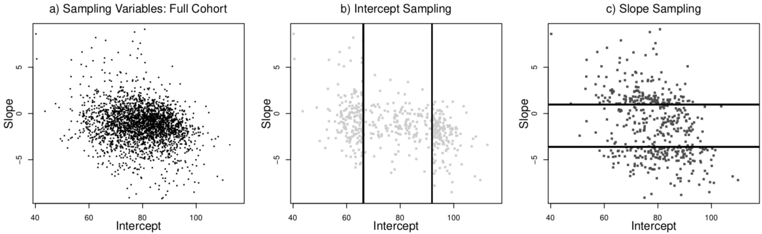
- Y_{ij} is $FEV\%$ for subject i at visit j
- snp_i is an indicator for the presence of at least one copy of the allele at rs177852
- t_{ij} is the time variable
- (b_{0i}, b_{1i}) are the random intercept and slope for subject i
- c_i is a continuous baseline covariate
- ϵ_{ij} is assumed to be normally distributed and independent of the random effects

Outcome Dependent Sampling

- Longitudinal outcome data and basic covariate data ($\boldsymbol{Y}_i$, $\boldsymbol{T}_i$, $\boldsymbol{C}_i$) are available, but resource constraints allow us to collect SNP on 20% of the subjects
- We want to select the most informative individuals using outcome dependent sampling (ODS)
- Sampling is based on strata defined by low-dimensional summaries of $\boldsymbol{Y}_i$:
  - $E(Y_{ij}) = q_{0i} + q_{1i} t_{ij}$
  - q_{0i} is the subject-specific mean of $FEV\%$ at baseline and q_{1i} is the subject-specific rate of change
  - Sort values of q_{0i} and/or q_{1i} and introduce cut-points that define sampling strata from which we sample with different probabilities

Different Sampling Scheme

- **Random sampling**
- **ODS: intercept sampling and slope sampling**



MI Background

- ODS allows us to select the most informative individuals for whom SNP_i will be collected, and to increase the estimates efficiency
- In many circumstances, we can improve estimates efficiency by using all available data and imputing SNP_i in those who were not sampled $(S_i = 0)$
- Because sampling only depends on $\boldsymbol{X}_{oi}$ and $\boldsymbol{Y}_i$:

$$pr(SNP_i|\boldsymbol{X}_{oi}, \boldsymbol{Y}_i, S_i = 0) = pr(SNP_i|\boldsymbol{X}_{oi}, \boldsymbol{Y}_i) = pr(SNP_i|\boldsymbol{X}_{oi}, \boldsymbol{Y}_i, S_i = 1)$$

- Multiple imputation should provide unbiased and valid estimates

Imputation Model

- We construct the imputation model in a straightforward way using available data on subjects. By Bayes' theorem:

$$\frac{pr(SNP_i{=}1|\boldsymbol{X}_{oi},\boldsymbol{Y}_i,S_i{=}0)}{pr(SNP_i{=}0|\boldsymbol{X}_{oi},\boldsymbol{Y}_i,S_i{=}0)} = \frac{f(\boldsymbol{Y}_i|SNP_i{=}1,\boldsymbol{X}_{oi},S_i{=}1)}{f(\boldsymbol{Y}_i|SNP_i{=}0,\boldsymbol{X}_{oi},S_i{=}1)} \frac{pr(SNP_i{=}1|\boldsymbol{X}_{oi},S_i{=}1)}{pr(SNP_i{=}0|\boldsymbol{X}_{oi},S_i{=}1)}$$

- We assume the Gaussian linear mixed model

- After log-transforming both sides of the equations and doing some algebra, the imputation model is:

$$\boldsymbol{Y_i^T V_i^{-1}(\mu_{1,i} - \mu_{0,i}) - \frac{1}{2}(\mu_{1,i}^T V_i^{-1}\mu_{1,i} - \mu_{0,i}^T V_i^{-1}\mu_{0,i})} + log\left(\frac{P(SNP_i{=}1|\boldsymbol{X}_{oi})}{P(SNP_i{=}0|\boldsymbol{X}_{oi})}\right)$$

where $\boldsymbol{\mu_{x,i}} = E[\boldsymbol{Y}_i|SNP_i = x, \boldsymbol{X_{oi}}]$ and $\boldsymbol{V}_i = Cov(\boldsymbol{Y}_i|SNP_i, \boldsymbol{X}_{oi})$

Simulation Settings

- We simulate data based on the Lung Health Study
- We consider a cohort of 2,000 subjects from which we sampled 400

$$Y_{ij} = \beta_0 + \beta_s SNP_i + \beta_t t_{ij} + \beta_{st} SNP_i t_{ij} + \beta_c c_i + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}$$

  - $(\beta_0, \beta_s, \beta_t, \beta_{st}, \beta_c) = (75, -0.5, -1, -0.5, -2)$
  - $(b_{0i}, b_{1i}) \sim N(\mathbf{0}, \boldsymbol{D})$
  - $\sigma^2_{b0} = 81, \sigma^2_{b1} = 1.56,\ \sigma_{b0,b1} = 0$
  - $\epsilon_{ij} \sim N(0, \sigma^2_e = 12.25)$
- We consider 3 different sampling designs
- We consider balanced and complete data, balanced and incomplete data, unbalanced data

Case 1: Balanced and Complete Data

- We impute *SNP* using the collected information on Y_{ij} at all time points
- Imputation model:

$$log\left(\frac{SNP_i=1|c_i,\boldsymbol{Y}_i}{SNP_i=0|c_i,\boldsymbol{Y}_i}\right) = \gamma_0 + \gamma_1 c_i + \gamma_2 y_{i1} + \gamma_3 y_{i2} + \gamma_4 y_{i3} + \gamma_5 y_{i4} + \gamma_6 y_{i5}$$

- If we use M = 50 imputations, coefficients' estimates and standard errors:

Sampling	β_0	β_t	β_s	β_{st}	β_c
Random	75.00 (0.38)	-1.00 (0.05)	-0.49 (1.00)	-0.49 (0.14)	-2.00 (0.22)
Intercept	75.00 (0.29)	-1.00 (0.05)	-0.51 (0.66)	-0.49 (0.14)	-2.01 (0.21)
Slope	75.00 (0.38)	-1.00 (0.04)	-0.49 (1.02)	-0.50 (0.09)	-2.00 (0.23)
Truth	75.00	-1.00	-0.50	-0.50	-2.00

Case 2: Balanced and Incomplete Data

- Since not every subject is measured at all visits we cannot use the same model as in the complete data case.
- What about using the mean of Y_{ij}?

$$log\left(\frac{snp_i = 1|c_i, \boldsymbol{y}_i}{snp_i = 0|c_i, \boldsymbol{y}_i}\right) = \gamma_0 + \gamma_1 c_i + \gamma_2 \bar{y}_i$$

- If we use M = 50 imputations, coefficients' estimate and standard errors:

Sampling	β_0	β_t	β_s	β_{st}	β_c
Random	75.25 (0.37)	-1.10 (0.04)	-1.32 (0.98)	-0.18 (0.09)	-1.93 (0.22)
Intercept	75.07 (0.36)	-1.10 (0.04)	-0.49 (0.87)	-0.16 (0.09)	-2.00 (0.21)
Slope	75.53 (0.33)	-1.05 (0.04)	-0.75 (0.84)	-0.33 (0.10)	-1.85 (0.22)
Truth	75.00	-1.00	-0.50	-0.50	-2.00

Case 2: Balanced and Incomplete Data (cont'd)

- Imputation Model

$$\boldsymbol{Y}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_{0,i}) - \frac{1}{2}(\boldsymbol{\mu}_{1,i}^T \boldsymbol{V}_i^{-1} \boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_{0,i}^T \boldsymbol{V}_i^{-1} \boldsymbol{\mu}_{0,i}) + log\left(\frac{P(SNP_i = 1|X_{oi})}{P(SNP_i = 0|X_{oi})}\right)$$

- Let ν_{ijk} the $(j,k)^{th}$ element of $\boldsymbol{V}_i^{-1}$. We need:
  - the weighted sum of Y_{ij}: $\sum_{j=1}^{n_i} \sum_{i=1}^{n_i} \nu_{ijk} Y_{ij}$
  - the weighted sum of $Y_{ij} t_{ik}$: $\sum_{j=1}^{n_i} \sum_{i=1}^{n_i} \nu_{ijk} Y_{ij} t_{ik}$
  - the weighted sum of t_{ij}: $\sum_{j=1}^{n_i} \sum_{i=1}^{n_i} \nu_{ijk} t_{ij}$
  - the weighted sum of $t_{ij} t_{ik}$: $\sum_{j=1}^{n_i} \sum_{i=1}^{n_i} \nu_{ijk} t_{ij} t_{ik}$
  - the confounder: c_i
  - the interaction between the weighted sum of t_{ij} and the confounder
  - the weighted sum of the confounder: $\sum_{j=1}^{n_i} \sum_{i=1}^{n_i} \nu_{ijk} c_i$
  - the sum of all ν_{ijk}

Case 2: Balanced and Incomplete Data (cont'd)

- Imputation Model

$$\boldsymbol{Y_i^T V_i^{-1}(\mu_{1,i} - \mu_{0,i})} - \tfrac{1}{2}\boldsymbol{(\mu_{1,i}^T V_i^{-1} \mu_{1,i} - \mu_{0,i}^T V_i^{-1} \mu_{0,i})} + log\left(\frac{P(SNP_i=1|X_{oi})}{P(SNP_i=0|X_{oi})}\right)$$

- Let ν_{ijk} the $(j,k)^{th}$ element of $\boldsymbol{V}_i^{-1}$. We need:
  - the weighted sum of Y_{ij}: $\sum_{j=1}^{n_i}\sum_{i=1}^{n_i} \nu_{ijk} Y_{ij}$
  - the weighted sum of $Y_{ij} t_{ik}$: $\sum_{j=1}^{n_i}\sum_{i=1}^{n_i} \nu_{ijk} Y_{ij} t_{ik}$
  - the weighted sum of t_{ij}: $\sum_{j=1}^{n_i}\sum_{i=1}^{n_i} \nu_{ijk} t_{ij}$
  - the weighted sum of $t_{ij} t_{ik}$: $\sum_{j=1}^{n_i}\sum_{i=1}^{n_i} \nu_{ijk} t_{ij} t_{ik}$
  - the confounder: c_i
  - the interaction between the weighted sum of t_{ij} and the confounder
  - the weighted sum of the confounder: $\sum_{j=1}^{n_i}\sum_{i=1}^{n_i} \nu_{ijk} c_i$
  - the sum of all ν_{ijk}

Case 2: Balanced and Incomplete Data (cont'd)

- We estimate ν_{ijk} iteratively
- We estimate the initial set of weights ν_{ijk} from a model that does not include *SNP* and $SNP \times time$
- We use MI to impute *SNP*
- We fit the linear mixed effect model of interest and estimate the new set of ν_{ijk}
- In a scenario where subjects are observed 4, 5 or 6 times, coefficient's estimate and standard error:

Sampling	β_0	β_t	β_s	β_{st}	β_c
Random	75.00 (0.54)	-1.00 (0.06)	-0.50 (1.01)	-0.49 (0.14)	-2.00 (0.22)
Intercept	75.00 (0.42)	-1.00 (0.06)	-0.50 (0.66)	-0.49 (0.15)	-2.01 (0.21)
Slope	75.00 (0.53)	-1.00 (0.04)	-0.48 (1.02)	-0.49 (0.10)	-2.01 (0.21)
Truth	75.00	-1.00	-0.50	-0.50	-2.00

Case 3: Unbalanced Data

- The same algorithm can be used for cases where subjects are observed at random time points
- For each subject i we generate n_i observations from $t_i \sim U(0, 10)$

Sampling	β_0	β_t	β_s	β_{st}	β_c
Random	74.99 (0.39)	-1.00 (0.06)	-0.49 (1.04)	-0.48 (0.14)	-2.01 (0.23)
Intercept	74.99 (0.30)	-1.00 (0.05)	-0.48 (0.71)	-0.50 (0.14)	-2.01 (0.22)
Slope	75.01 (0.40)	-1.00 (0.04)	-0.52 (1.07)	-0.50 (0.10)	-2.00 (0.23)
Truth	75.00	-1.00	-0.50	-0.50	-2.00

Summary

- When exposure ascertaiment cost limit the sample size, it is desirable to target a sample of informative subjects
- Analysis is usually done using complete data, or combining partial data and complete data using full likelihood approaches or multiple imputation
- We introduced an MI approach that is easy to implement and applicable to different sampling schemes
- We presented a series of simulation studies to look at the performance of the MI approach

Reference

- Schildcrout J, Rathouz P, Zelnick L, Garbett S, and Heagerty P. Biased Sampling Design To Improve Research Eefficiency: Factors Influencing Pulmonary Function Over Time In Children With Asthma. The Annals of Applied Statistics, 9(2):731753, 2015.
- Schildcrout J, Haneuse S, Tao R, Zelnick L, Schisterman E, Garbett S, Mercaldo N, Rathouz P, and Heagerty P. Two-Phase, Generalized Case-Control Designs for the Study of Quantitative Longitudinal Outcomes. American Journal of Epidemiology. 2019.

Thank you!

- chiara.di.gravio@vanderbilt.edu