VIRGINIA COMMONWEALTH UNIVERSITY

Selection Model for COVID-19 Recovery and Informative Dropout Given Web Survey Data Missing Not At Random

Serenity Budd, M.S., Yongyun Shin, Ph.D.

Department of Biostatistics, Virginia Commonwealth University

Selection Model for MNAR Data

Variables

- y_{ij} = Binary outcome for person i at j^{th} timepoint, $j \in \{1, 2\}$
  - y_{i1} is fully observed, y_{i2} has missing values
  - $y_{i2} \sim Bernoulli(p_i),\ logit(p_i) = \beta_0 + \beta_1 y_{i1} + \boldsymbol{\beta_2^T x_i}$
- d_i = Binary dropout indicator for person i
  - 0 = Participant did not drop out → y_{i2} observed
  - 1 = Participant dropped out → y_{i2} missing
  - $d_i \sim Bernoulli(q_i),\ logit(q_i) = \gamma_0 + \gamma_1 y_{i1} + \gamma_2 y_{i2} + \boldsymbol{\gamma_3^T x_i}$
- $\boldsymbol{x_i}$ = Vector of covariates for person i

Missing Not at Random (MNAR)

- Dropout depends on the previously observed responses (y_1) and the unobserved responses (y_2)
- **Hypothesis**: $\gamma_1 \neq 0,\ \gamma_2 \neq 0$

Selection Model

$$f(y_{i2}, d_i | y_{i1}, \boldsymbol{x_i}) = f(d_i | y_{i2}, y_{i1}, \boldsymbol{x_i})\ f(y_{i2} | y_{i1}, \boldsymbol{x_i})$$

- Decomposes the joint distribution into **dropout conditional on recovery** and **marginal recovery**, controlling for baseline covariates

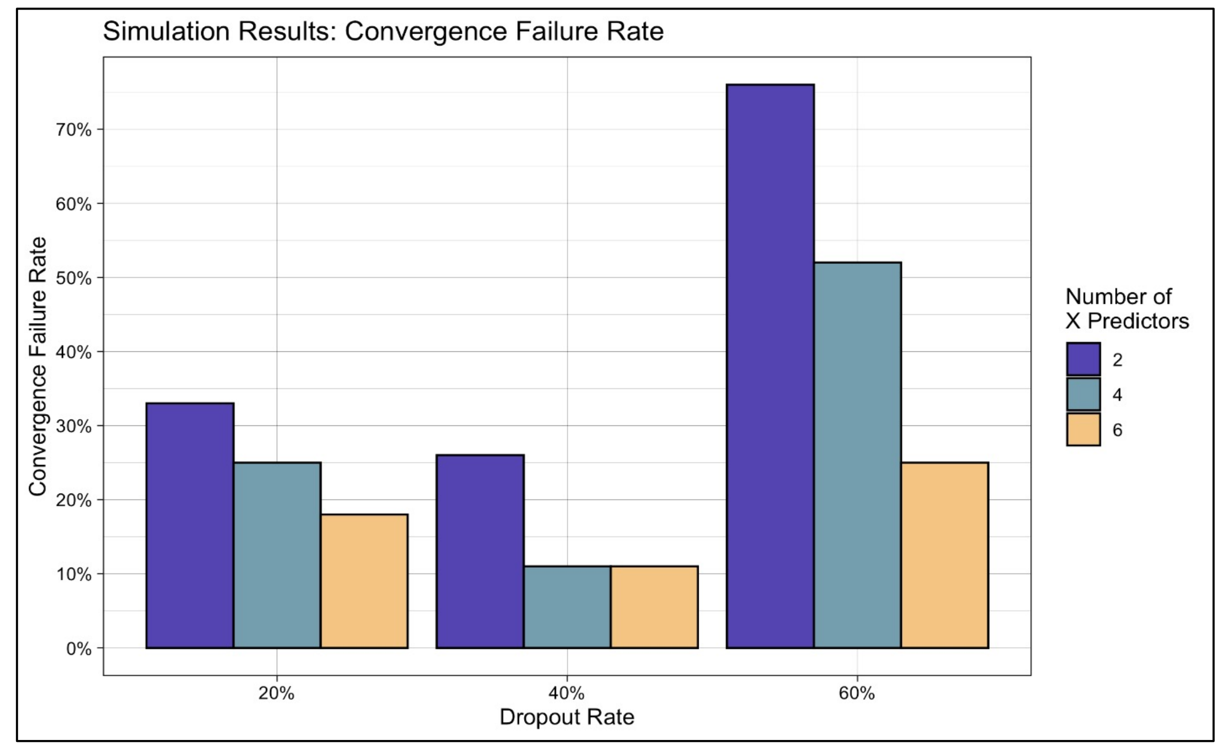
Method

- Derive likelihood and estimate via Newton-Raphson algorithm
- Initial values were chosen through predictive mean matching
- Run simulation to determine efficacy of the method

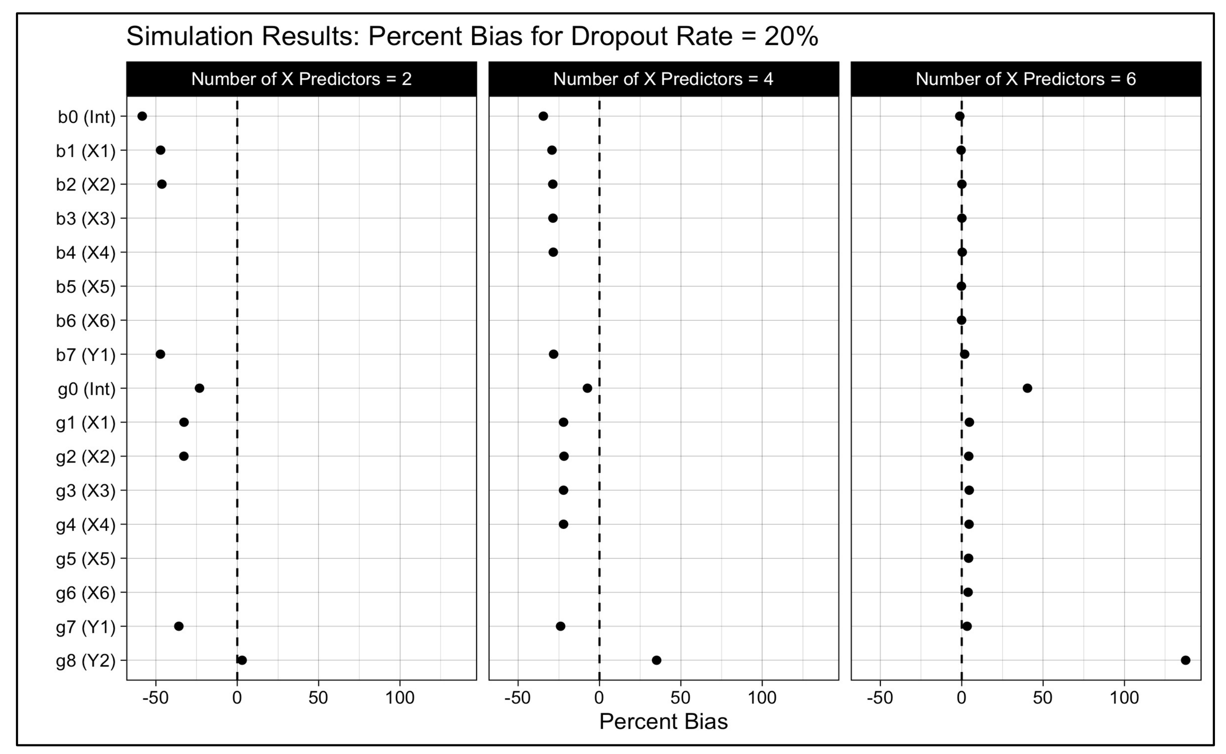
Simulation

- Simulation Parameters
  - N = 1,231
  - Success rate at timepoint 1: 20%
  - Success rate at timepoint 2: 65%
  - Number of predictors: 2, 4, or 6
  - Dropout rates: 20%, 40%, 60%
  - 1,000 samples were simulated for each setup
- For each dropout rate
  - Manipulate coefficients to set dropout rate
  - Run the algorithm using 2, 4, or 6 predictors

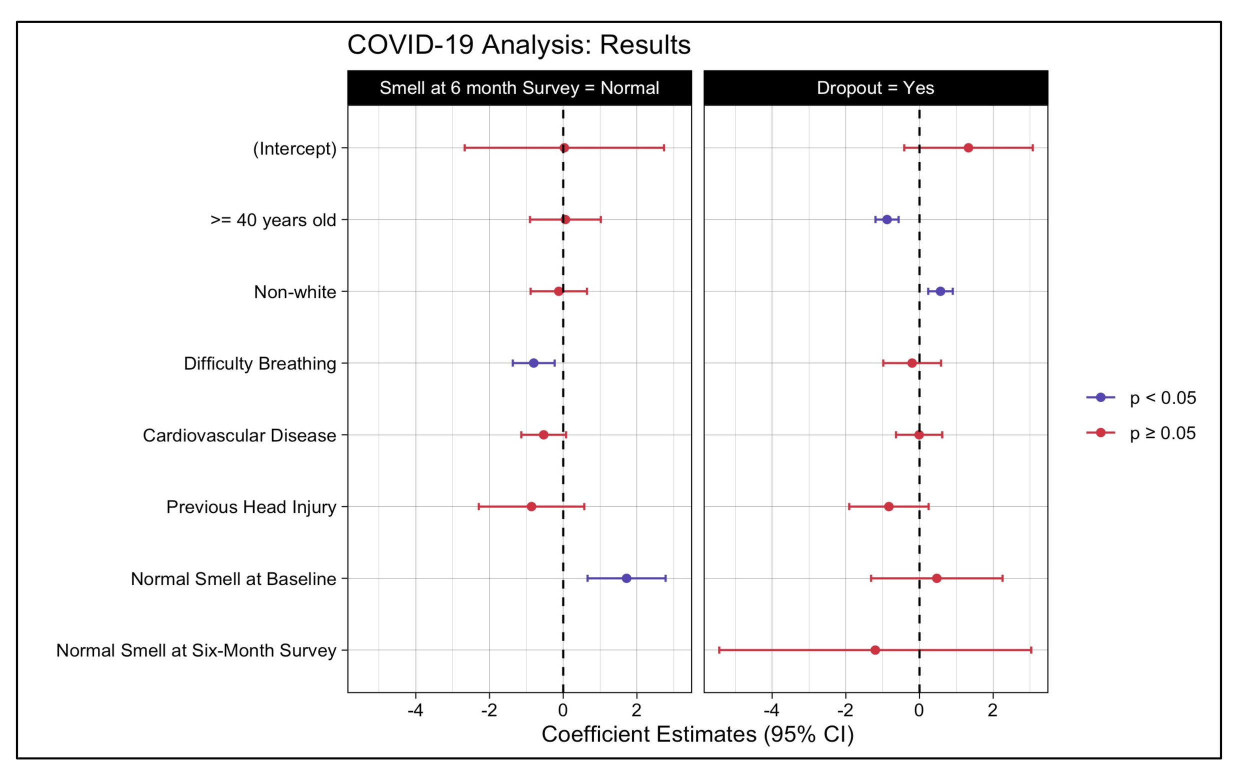
Results



- Higher number of predictors leads to more successful convergence



- Higher number of predictors leads to decreased bias in coefficients



- Difficulty breathing during COVID-19 symptoms → less likely to be recovered
- Normal sense of smell at baseline → more likely to be recovered
- Greater than or equal to 40 years old → less likely to drop out
- Non-white participants → more likely to drop out

COVID-19 Survey Analysis

COVID-19 Survey

- Nationwide longitudinal web-based survey
  - Conducted by Virginia Commonwealth University
  - Two timepoints: baseline, 6 months after baseline
  - N = 1,231
  - Dropout rate: 62%
- Participant population
  - Normal sense of smell before 01/2020
  - COVID-19 diagnosis between 01/2020 & baseline survey
  - Abnormal sense of smell during COVID-19 symptoms
- **Goal**: predict recovery of sense of smell at six-months
- **Assumption**: dropout depends on sense of smell at both timepoints
- **Method**: estimate recovery of sense of smell while accounting for dropout using the selection model

Discussion

Selection Model

- More predictors help recover information missing in $\boldsymbol{y_2}$
- **Limitation**
  - High uncertainty in $\boldsymbol{y_2}$ coefficient due to missing values
- **Future Research**
  - Apply multiple imputation
  - Conduct sensitivity analysis

COVID-19 Analysis

- **Limitation**
  - High standard errors in estimate of normal smell at six-month
  - Not enough evidence to reject the null hypothesis

References

1. Diggle, P., & Kenward, M. (1994). Informative Drop-Out in Longitudinal Data Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 43*(1), 49-93.
2. Little, R. (2008). Selection and Pattern Mixture Models. In Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.), *Longitudinal Data Analysis* (1st ed.). Chapman and Hall/CRC.