SELECTION MODEL FOR COVID-19 RECOVERY AND INFORMATIVE DROPOUT GIVEN WEB SURVEY DATA MNAR

Presenter: Serenity Budd Advisor: Dr. Yongyun Shin

COVID-19 Survey

- Nationwide longitudinal web-based survey
- 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*
- Five surveys
  - *Baseline, 14 days, 1 month, 3 months, 6 months*

GOAL

Predict recovery of sense of smell at six-month survey

PROBLEM

High dropout rates at the six-month survey

HYPOTHESIS

Dropout depends on sense of smell at baseline and at six-month survey

Outcome Variables

- $Y_{i2} =$ sense of smell for person *I* at *2nd* timepoint
  - *Binary outcome*
    - $0 =$ Abnormal sense of smell
    - $1 =$ Normal sense of smell
- $D_i =$ dropout indicator
  - *Binary outcome*
    - $0 =$ Participant did not drop out $\rightarrow Y_2$ observed
    - $1 =$ Participant dropped out $\rightarrow Y_2$ missing

$$f(y_2, d|y_1, x) = f(d|y_2, y_1, x)f(y_2|y_1, x)$$

SELECTION MODEL

Distributional Assumptions

$$Y_2 \sim Bernoulli(p)$$

$$logit(p) = \beta_0 + \beta_1 y_1 + \boldsymbol{\beta_2^T} \mathbf{x}$$

$$D \sim Bernoulli(q)$$

$$logit(q) = \gamma_0 + \gamma_1 y_1 + \gamma_2 y_2 + \boldsymbol{\gamma_3^T} \mathbf{x}$$

Missing Data Mechanisms

$$logit(q) = \gamma_0 + \gamma_1 y_1 + \gamma_2 y_2 + \boldsymbol{\gamma_3^T \mathbf{x}}$$

- Missing Completely At Random (MCAR)
  - $\gamma_1 = \gamma_2 = 0$
- Missing At Random (MAR)
  - $\gamma_1 \neq 0,\ \gamma_2 = 0$
- Missing Not At Random (MNAR)
  - $\gamma_1 \neq 0,\ \gamma_2 \neq 0$

(Diggle and Kenward, 1994)

Likelihood

$$\ell = \sum_{i=1}^{n} \ln[f(d_i|y_{i2})f(y_{i2})]$$

$$= \sum_{\{i:\ d_i=0\}} \ln(1-q_i) + y_{i2}\ln(p_i) + (1-y_{i2})\ln(1-p_i)$$

$$+ \sum_{\{i:\ d_i=1\}} \ln(q_{i0}(1-p_i) + q_{i1}p_i)$$

$$\theta_{n+1} = \theta_n - \ell''(\theta_n)^{-1}\ell'(\theta_n)$$

NEWTON-RAPHSON ALGORITHM

Initial Values

- Fit $Y_2 \sim y_1 + \mathbf{x}$ in glm() using observed values
  - *Method 1: Predict probabilities for missing* $y_2 \rightarrow \tilde{y}_2$

    OR
  - *Method 2: Impute* y_2 *by predictive mean matching* $\rightarrow \tilde{y}_2$
- Fit in $D \sim y_1 + \tilde{y}_2 + \mathbf{x}$ in glm()
- Fitted coefficients are initial values

Simulation Parameters

- 1,000 iterations
- N = 1231
- k = 2, 4, or 6 predictors
- Dropout rates: 20%, 40%, 60%
- Normal smell rate at baseline: 20%
- Normal smell rate at six-month survey: 65%

Simulate Data

- For each dropout rate
  - *Manipulate coefficients to set dropout rate*
  - *Simulate 1,000 datasets:*
    1. Six x predictors
    2. $Y_1 \sim x_1$
    3. $Y_2 \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + y_1$
    4. $D \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + y_1 + y_2$
    5. If $D = 1$, then set Y_2 missing

Newton-Raphson Algorithm

- For each of the 1,000 datasets
  - *Run Newton-Raphson algorithm six times*
    - 2, 4, or 6 predictors
    - Two choices of initial values

SIMULATION RESULTS

Convergence Failure Rate

Missing Percent	Number of X Predictors	Failure Percent Method 1	Failure Percent Method 2
20 %	2	32.9 %	33.4 %
	4	25.2 %	24.7 %
	6	14.2 %	17.8 %
40 %	2	25.6 %	26.3 %
	4	12.5 %	10.9 %
	6	11.4 %	11.1 %
60 %	2	74.5 %	75.8 %
	4	48.7 %	52.2 %
	6	22.4 %	25.3 %

SIMULATION RESULTS

Number of Iterations Until Convergence

Missing Percent	Number of X Predictors	Num. of Iterations Method 1	Num. of Iterations Method 2
20 %	2	11.5	11.5
	4	10.1	10.2
	6	7.9	8.5
40 %	2	11.6	11.8
	4	12.1	12.3
	6	6.5	7.2
60 %	2	18.8	18.5
	4	12.2	12.4
	6	7.1	7.2

% Bias (Median SE) for Dropout Rate = 20 %

	k = 2		k = 4		k = 6	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
b0 (Int)	-59.3 (0.2)	-58.4 (0.20)	-34.6 (0.15)	-34.5 (0.15)	-1.7 (0.13)	-1.2 (0.13)
b1 (X1)	-47.4 (0.11)	-47.1 (0.11)	-29.2 (0.10)	-29.2 (0.10)	-0.8 (0.11)	-0.4 (0.11)
b2 (X2)	-46.8 (0.11)	-46.3 (0.11)	-28.6 (0.10)	-28.7 (0.10)	-0.3 (0.11)	0.1 (0.11)
b3 (X3)			-28.6 (0.10)	-28.6 (0.10)	-0.3 (0.11)	0.1 (0.11)
b4 (X4)			-28.5 (0.10)	-28.4 (0.10)	-0.1 (0.11)	0.3 (0.11)
b5 (X5)					-0.4 (0.11)	-0.2 (0.11)
b6 (X6)					-0.2 (0.11)	-0.1 (0.11)
b7 (Y1)	-47.6 (0.23)	-47.2 (0.23)	-27.9 (0.24)	-28.2 (0.24)	1.3 (0.27)	1.8 (0.27)

	k = 2		k = 4		k = 6	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
g0 (Int)	-31.4 (1.02)	-23.2 (1.06)	-8.4 (1.20)	-7.4 (1.21)	35.6 (1.35)	40.4 (1.39)
g1 (X1)	-32.0 (0.15)	-32.7 (0.15)	-22.1 (0.13)	-22.1 (0.13)	5.0 (0.14)	4.7 (0.14)
g2 (X2)	-32.0 (0.14)	-32.8 (0.14)	-21.8 (0.12)	-21.8 (0.13)	4.8 (0.13)	4.3 (0.13)
g3 (X3)			-22.1 (0.13)	-22.1 (0.13)	5.1 (0.13)	4.6 (0.13)
g4 (X4)			-22.0 (0.13)	-22.1 (0.13)	4.9 (0.13)	4.5 (0.13)
g5 (X5)					4.7 (0.13)	4.2 (0.13)
g6 (X6)					4.5 (0.13)	3.9 (0.13)
g7 (Y1)	-35.2 (0.21)	-35.9 (0.21)	-23.8 (0.22)	-23.9 (0.22)	4.0 (0.25)	3.3 (0.25)
g8 (Y2)	-30.5 (1.56)	3 (1.59)	31.7 (1.50)	35.1 (1.50)	118.6 (1.56)	137.6 (1.62)

COVID-19 Analysis: Predictors

Age Group	Race	Difficulty Breathing	Cardiovascular Disease	Previous Head Injury
• *< 40 years old* • *≥ 40 years old*	• *White* • *Non-white*	• *Without* • *With*	• *Without* • *With*	• *Without* • *With*

COVID-19 Analysis: Results

		Estimate (SE)	p-value
Y_2 Model	(Intercept)	0.03 (1.38)	0.98
	>= 40 years old	0.06 (0.49)	0.90
	Non-white	-0.12 (0.39)	0.77
	Difficulty Breathing	**-0.80 (0.29)**	**0.01**
	Cardiovascular Disease	-0.53 (0.31)	0.09
	Previous Head Injury	-0.86 (0.73)	0.24
	Normal Smell at Baseline	**1.72 (0.54)**	**<.01**
D Model	(Intercept)	1.33 (0.89)	0.13
	>= 40 years old	**-0.88 (0.16)**	**<.01**
	Non-white	**0.57 (0.17)**	**<.01**
	Difficulty Breathing	-0.20 (0.40)	0.61
	Cardiovascular Disease	-0.01 (0.32)	0.98
	Previous Head Injury	-0.83 (0.55)	0.13
	Normal Smell at Baseline	0.47 (0.91)	0.61
	Normal Smell at Six-Month Survey	-1.20 (2.16)	0.58

Log likelihood: -1053.0

Number of iterations: 8

Discussion

- More predictors help recover information in missing Y_2
  - *Reduced failed convergence rate*
  - *Reduced bias*
- Limitation
  - *High uncertainty in Y_2 coefficient due to missing values*

Future Research

- Multiple imputation in the algorithm
  - Y_2 *model estimated with accuracy*
  - *Impute* Y_2
  - *Given imputed* Y_2*, fit* D *model*
- Sensitivity analysis

References

- 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. doi:10.2307/2986113
- 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. https://doi.org/10.1201/9781420011579

THANK YOU!
QUESTIONS?