Previous Work on Non-montone Missingness

Caleb Leedy

August 15, 2023

Outline

- Notation
- Simulation Studies

Notation

Case 1: Monotone missingness

	X	Y_1	Y_2	R_1	R_2
A_{11}	✓	✓	✓	1	1
A_{10}	✓	✓		1	0
A_{00}	✓			0	0

Case 2: Non-monotone missingness

	X	Y_1	Y_2	R_1	R_2
A_{11}	✓	✓	✓	1	1
A_{10}	✓	✓		1	0
A_{01}	✓		✓	0	1
A_{00}	✓			0	0

The goal is to estimate $\theta = E[g(X, Y_1, Y_2)]$.

Proposed Estimator: Monotone

For, $b(X_i, Y_{1i}) = E[g_i \mid X_i, Y_{1i}]$ and $i = \{1, \dots, n\}$,

$$
\begin{aligned}
\hat{\theta}_{\text{eff}} = n^{-1} \sum_{i=1}^{n} E[g_i \mid X_i] \\
+ n^{-1} \sum_{i=1}^{n} \frac{R_{1i}}{\pi_{1+}(X_i)} (b(X_i, Y_{1i}) - E[g_i \mid X_i]) \\
+ n^{-1} \sum_{i=1}^{n} \frac{R_{1i} R_{2i}}{\pi_{11}(X_i)} (g_i - b_i)
\end{aligned}
$$

Proposed Estimator: Non-monotone

Define,

$$
\begin{gathered}
b(X_i, Y_{1i}) = E[g_i \mid X_i, Y_{1i}], \\
a(X_i, Y_{2i}) = E[g_i \mid X_i, Y_{2i}], \text{ and} \\
i = \{1, \dots, n\},
\end{gathered}
$$

Proposed Estimator: Non-monotone

$$
\begin{aligned}
\hat{\theta}_{\text{eff}} &= n^{-1} \sum_{i=1}^{n} E[g_i \mid X_i] \\
&+ n^{-1} \sum_{i=1}^{n} \frac{R_{1i}}{\pi_{1+}(X_i)} (b(X_i, Y_{1i}) - E[g_i \mid X_i]) \\
&+ n^{-1} \sum_{i=1}^{n} \frac{R_{2i}}{\pi_{2+}(X_i)} (a(X_i, Y_{2i}) - E[g_i \mid X_i]) \\
&+ n^{-1} \sum_{i=1}^{n} \frac{R_{1i} R_{2i}}{\pi_{11}(X_i)} (g_i - b_i - a_i + E[g_i \mid X_i])
\end{aligned}
$$

Simulation Outline: Monotone MAR

For this simulation we use the following approach:

1. Generate X, Y_1, and Y_2 for elements $i = 1, \ldots, n$.
2. Using the covariate X, determine the probability p_1 of Y_1 being observed for each element i.
3. Based on p_1, determine if $R_1 = 1$.
4. If $R_1 = 0$, then $R_2 = 0$. Otherwise, using variables X and Y_1, determine the probability p_{12}.
5. Based on p_{12} determine if $R_2 = 1$.

Simulation Outline: Non-monotone MAR

Following the approach of [2], the algorithm to generate the data is the following:

1. Generate X, Y_1, and Y_2 for elements $i = 1, \ldots, n$.
2. Using the covariate X_i, generate probabilities for each element i p_0, p_1, and p_2 such that $p_0 + p_1 + p_2 = 1$.
3. Select one option based on the three probabilities for each element i. If 0 is selected: $R_1 = 0$ and $R_2 = 0$; if 1 is selected $R_1 = 1$; if 2 is selected, $R_2 = 1$.
4. We take the next step in multiple cases. If 0 was selected, we are done. If 1 was selected, we generate probabilities p_{12} based on X and Y_1. Then based on this probability, we determine if $R_2 = 1$. In the same manner, if 2 was selected in the previous step, we generate probabilities p_{21} based on X and Y_2. Then based on this probability, we determine if $R_2 = 1$.

Simulation 1: Monotone MAR

We generate data from the following distributions:

$$X_i \stackrel{iid}{\sim} N(0,1)$$
$$Y_{1i} \stackrel{iid}{\sim} N(0,1)$$
$$Y_{2i} \stackrel{iid}{\sim} N(\theta,1)$$

Then, we create the probabilities $p_1 = \text{logistic}(x_i)$ and $p_{12} = \text{logistic}(y_{1i})$. Since, both x_i and y_1 are standard normal distributions, each of these probabilities is approximately 0.5 in expectation.

The goal of this simulation is to estimate θ. Alternatively, we can express this as solving the estimating equation:

$$g(\theta) \equiv Y_2 - \theta = 0.$$

Simulation 1: Monotone MAR

We estimate θ using the following procedures:

- Oracle: This computes $\bar{Y}_2$ using *both* the observed and missing data.
- IPW-Oracle: This is an IPW estimator using only the observed values of Y_2. The weights (inverse probabilities) use the actual probabilities.
- IPW-Est: This is an IPW estimator using the probabilities that have been estimated by a logistic model.
- Semi: The monotone efficient estimator.

- Sample size (n): 2000
- Monte Carlo replications: 2000

Table: True Value is -5

algorithm	bias	sd	tstat	pval
oracle	0.001	0.033	0.680	0.248
ipworacle	-0.012	0.392	-0.973	0.165
ipwest	0.007	0.186	1.178	0.120
semi	0.001	0.074	0.538	0.295

Table: True Value is 0

algorithm	bias	sd	tstat	pval
oracle	-0.001	0.031	-1.091	0.138
ipworacle	-0.001	0.085	-0.201	0.420
ipwest	0.000	0.085	-0.029	0.488
semi	0.000	0.079	0.112	0.455

Table: True Value is 5

algorithm	bias	sd	tstat	pval
oracle	0.000	0.033	-0.468	0.320
ipworacle	0.010	0.383	0.857	0.196
ipwest	-0.006	0.176	-1.020	0.154
semi	0.000	0.077	-0.049	0.481

Simulation 1: Non-monotone MAR

We generate variables (X, Y_1, Y_2) using the following setup:

$$\begin{bmatrix} X_i \\ \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ \theta \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \sigma_{yy} \\ 0 & \sigma_{yy} & 1 \end{bmatrix} \right).$$

Then,

$$y_{1i} = x_i + \varepsilon_{1i} \text{ and } y_{2i} = x_i + \varepsilon_{2i}.$$

Since we have nonmonotone data, our “Stage 1” probabilities are different. We compute the true Stage 1 probabilities being proportional to the following values:

$$\begin{aligned} p_0 &= 0.2 \\ p_1 &= 0.4 \\ p_2 &= 0.4 \end{aligned}$$

However, we keep the same structure for the Stage 2 probabilities with: $p_{12} = \text{logistic}(y_1)$ and $p_{21} = \text{logistic}(y_2)$.

Table: True Value is -5. Cor(Y1, Y2) = 0

algorithm	bias	sd	tstat	pval
oracle	0.000	0.032	0.285	0.388
ipworacle	-0.003	0.381	-0.318	0.375
proposed	0.000	0.038	0.492	0.311

Table: True Value is 0. Cor(Y1, Y2) = 0

algorithm	bias	sd	tstat	pval
oracle	0.000	0.032	0.285	0.388
ipworacle	0.000	0.076	-0.237	0.406
proposed	0.001	0.038	0.894	0.186

Table: True Value is 5. Cor(Y1, Y2) = 0

algorithm	bias	sd	tstat	pval
oracle	0.000	0.032	0.285	0.388
ipworacle	-0.001	0.098	-0.479	0.316
proposed	0.000	0.037	0.505	0.307

Simulation 2: Non-monotone MAR

For this simulation, we focus on Cov(Y_1, Y_2). The data generating process now has $\sigma_{yy} \neq 0$. We are still interested in $\bar{Y}_2$ and we still run 2000 simulation with 2000 observations. In all the next simulations the true value of $\theta = 0$. The results are the following:

Table: True Value is 0. Cor(Y1, Y2) = 0.1

algorithm	bias	sd	tstat	pval
oracle	0.001	0.031	1.623	0.052
ipworacle	0.001	0.077	0.762	0.223
proposed	0.001	0.037	1.366	0.086

Table: True Value is 0. Cor(Y1, Y2) = 0.5

algorithm	bias	sd	tstat	pval
oracle	0.001	0.032	1.486	0.069
ipworacle	0.004	0.086	1.890	0.029
proposed	0.000	0.041	0.172	0.432

Table: True Value is 0. Cor(Y1, Y2) = 0.9

algorithm	bias	sd	tstat	pval
oracle	0.001	0.032	0.706	0.240
ipworacle	0.003	0.098	1.395	0.082
proposed	-0.002	0.062	-1.339	0.090

Comparing with a Calibration Estimator

The efficient monotone estimator should be very similar to the following calibration estimator, for $\sum_{i=1}^{n} w_i y_{2i}$,

$$\text{argmin}_w \sum_{i=1}^{n} w_i^2 \text{ such that}$$

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} R_{1i} w_{1i} x_i$$

$$\sum_{i=1}^{n} w_{1i}(x_i, y_{1i}) = \sum_{i=1}^{n} R_{1i} R_{2i} w_{2i}(x_i, y_{1i})$$

The reason that these should be the same is because they are similar in relationship to a calibration and regression estimator which are equivalent.

Calibration Comparison: Monotone

To test the idea that the monotone regression estimator is similar to the calibration estimator we run several simulation studies. In the monotone case data is generating in the following steps:

1. The variables X, Y_1, and Y_2 are simulated from the following distributions:

$$X_i \overset{iid}{\sim} N(0,1)$$
$$Y_{1i} \overset{iid}{\sim} N(0,1)$$
$$Y_{2i} \overset{iid}{\sim} N(\theta,1).$$

2. After the variables have been simulated, we see which variables are observed. We always observe X_i. We observed Y_1 with probability $p_{1i} \propto$ logistic(x_i). If Y_{1i} is observed, then we observe Y_{2i} with probability $p_{2i} \propto$ logistic(y_{1i}). If Y_{1i} is not observed, we do not observe Y_{2i}.

Additional Estimators

- HT estimator of $\theta = E(Y_2)$:

$$\hat{\theta}_{\mathrm{HT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_{1i}R_{2i}}{\pi_{11}(X_i)} y_{2i}$$

- The three-phase regression estimator of θ:

$$\begin{aligned}
\hat{\theta}_{\mathrm{reg}} &= \frac{1}{n}\sum_{i\in A_2}\frac{1}{\pi_{2i}}\left\{y_i - \hat{E}(Y \mid x_i, z_i)\right\} \\
&+ \frac{1}{n}\sum_{i\in A_1}\frac{1}{\pi_{1i}}\left\{\hat{E}(Y \mid x_i, z_i) - \hat{E}(Y \mid x_i)\right\} + \frac{1}{n}\sum_{i\in U}\hat{E}(Y \mid x_i) \\
&= \bar{x}_0'\hat{\beta} + \left(\bar{x}_1'\hat{\gamma}_x + \bar{z}_1'\hat{\gamma}_z - \bar{x}_1'\hat{\beta}\right) + \{\bar{y}_2 - (\bar{x}_2'\hat{\gamma}_x + \bar{z}_2'\hat{\gamma}_z)\} \\
&= \bar{y}_2 + \{\bar{x}_1'\hat{\gamma}_x + \bar{z}_1'\hat{\gamma}_z - (\bar{x}_2'\hat{\gamma}_x + \bar{z}_2'\hat{\gamma}_z)\} + \left(\bar{x}_0'\hat{\beta} - \bar{x}_1'\hat{\beta}\right)
\end{aligned}$$

- We can view the above three-phase regression estimator as a projection estimator of [1].

Table: True Value is -5

algorithm	bias	sd	tstat	pval
oracle	0.001	0.032	0.849	0.198
ipworacle	0.009	0.410	0.678	0.249
ipwest	0.012	0.191	1.974	0.024
semi	0.002	0.076	0.907	0.182
reg2p	0.002	0.072	0.857	0.196
reg3p	0.004	0.127	0.992	0.161
calib	0.003	0.075	1.339	0.091

Table: True Value is 0

algorithm	bias	sd	tstat	pval
oracle	0.001	0.031	1.122	0.131
ipworacle	-0.003	0.085	-1.002	0.158
ipwest	-0.003	0.088	-1.131	0.129
semi	-0.001	0.077	-0.298	0.383
reg2p	-0.001	0.072	-0.440	0.330
reg3p	-0.004	0.118	-1.078	0.141
calib	0.000	0.076	0.080	0.468

Table: True Value is 5

algorithm	bias	sd	tstat	pval
oracle	-0.001	0.031	-1.015	0.155
ipworacle	-0.003	0.399	-0.213	0.416
ipwest	-0.011	0.189	-1.914	0.028
semi	-0.002	0.077	-0.775	0.219
reg2p	-0.004	0.075	-1.494	0.068
reg3p	0.000	0.122	0.033	0.487
calib	-0.001	0.075	-0.518	0.302

Non-monotone Calibration

For the non-monotone case, we believe that we have the following calibration equations:

$$
\begin{aligned}
\sum_{i=1}^n E[g_i \mid X_i] &= \sum_{i=1}^n R_{1i} w_{1i} E[g_i \mid X_i] \\
\sum_{i=1}^n E[g_i \mid X_i] &= \sum_{i=1}^n R_{2i} w_{2i} E[g_i \mid X_i] \\
\sum_{i=1}^n R_{1i} w_{1i} E[g_i \mid X_i, Y_{1i}] &= \sum_{i=1}^n R_{1i} R_{2i} w_{ci} E[g_i \mid X_i, Y_{1i}] \\
\sum_{i=1}^n R_{2i} w_{2i} E[g_i \mid X_i, Y_{1i}] &= \sum_{i=1}^n R_{1i} R_{2i} w_{ci} E[g_i \mid X_i, Y_{2i}] \\
\sum_{i=1}^n E[g_i \mid X_i] &= \sum_{i=1}^n R_{1i} R_{2i} w_{ci} E[g_i \mid X_i].
\end{aligned}
$$

Non-monotone Calibration

We still have the same goal of the simulation study: estimate $\theta = E[Y_2]$.

1. Generate X_i, ε_{1i}, and ε_{2i} from the following distributions:

$$
\begin{aligned}
x_i &\stackrel{iid}{\sim} N(0,1) \\
\varepsilon_{1i} &\stackrel{iid}{\sim} N(0,1) \\
\varepsilon_{2i} &\stackrel{iid}{\sim} N(\theta,1)
\end{aligned}
$$

Then we have

$$y_{1i} = x_i + \varepsilon_{1i} \text{ and } y_{2i} = x_i + \varepsilon_{2i}.$$

2. Then we have to select the variables to observe. We always observe X_i. Then we choose to either observe Y_1 with probability 0.4, Y_2 with probability 0.4 or neither with probability 0.2.
3. If neither then $R_{1i} = 0$ and $R_{2i} = 0$. If we observe Y_1 then $R_1 = 1$ and if we observe Y_2 then $R_2 = 1$.
4. If we observe either Y_1 or Y_2 then with probability $p \propto$ logistic(Y_k) where Y_k is the observed Y variable we choose to observe the other Y variable.
5. If the other Y variable is observed then the corresponding $R_k = 1$. Otherwise, $R_k = 0$.

Table: True Value is -5. Cor(Y1, Y2) = 0

algorithm	bias	sd	tstat	pval
oracle	-0.001	0.045	-0.953	0.170
ipworacle	0.011	0.552	0.627	0.266
proposed	-0.002	0.055	-0.873	0.191
reg2p	-0.008	0.099	-2.512	0.006
reg3p	-0.007	0.099	-2.127	0.017
calib	-0.002	0.054	-1.312	0.095

Table: True Value is 0. Cor(Y1, Y2) = 0

algorithm	bias	sd	tstat	pval
oracle	-0.001	0.044	-0.945	0.173
ipworacle	0.001	0.112	0.178	0.429
proposed	-0.001	0.053	-0.363	0.358
reg2p	0.004	0.069	1.809	0.035
reg3p	0.005	0.069	2.372	0.009
calib	-0.001	0.052	-0.508	0.306

Table: True Value is 5. Cor(Y1, Y2) = 0

algorithm	bias	sd	tstat	pval
oracle	-0.002	0.045	-1.358	0.087
ipworacle	-0.002	0.141	-0.409	0.341
proposed	-0.002	0.051	-1.531	0.063
reg2p	-0.003	0.052	-1.589	0.056
reg3p	-0.003	0.052	-1.565	0.059
calib	-0.002	0.051	-1.401	0.081

References I

[1] Jae Kwang Kim and Jon NK Rao. "Combining data from two independent surveys: a model-assisted approach". In: *Biometrika* 99.1 (2012), pp. 85–100.

[2] James M Robins and Richard D Gill. "Non-response models for the analysis of non-monotone ignorable missing data". In: *Statistics in medicine* 16.1 (1997), pp. 39–56.