Overview:

The goal of this project is to outperform existing techniques in the literature related to nonmonotone missing data.

Initial Simulations:

- **Implemented simulation of monotone MAR data**: This is correspondingly easier than the subsequent nonmonotone MAR simulation. 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 $R_2 = 1$.

  At the end of the algorithm, we have determined the values of binary variables R_1 and R_2 for each i and if either of them are equal to 1, the corresponding level of Y_k. As is common in this literature, the values of R_1 and R_2 determine if the corresponding variable Y_1 or Y_2 is missing or observed with $R = 1$ indicating Y being observed.

- **Implemented simulation of nonmonotone MAR data**: Following the approach of [2], I construct a nonmonotone MAR simulation with two response variables Y_1 and Y_2 and one covariate X. 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 $R_2 = 1$.

  Like the monotone MAR simulation this algorithm produces similar final results with the determination of binary variables R_1 and R_2 and variables X, Y_1, and Y_2. Unlike the monotone MAR case, the nonmonotone MAR includes observations with Y_2 observed and Y_1 missing.

- **Simulation 1 with Monotone MAR**: Following the algorithm described in the monotone MAR simulation bullet, we first generate data from the following distributions:

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

  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:

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

  We estimate θ using the following procedures:

  - Oracle: This computes $\bar{Y}$ 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: This is the monotone semiparametric efficient estimator from Slide 11 (Equation 2) of Dr. Kim's Nonmonotone Missingness presentation.

  We run this simulation with different values of θ, sample size of 2000, and 2000 Monte Carlo replications. Each algorithm for each replication generates $\hat{\theta}$. In the subsequent tables, we compute the bias, standard deviation (sd), t-statistic (where we test for a significant difference between the Monte Carlo mean $\hat{\theta}$ and the true θ) and the p-value of the t-statistic.

Table 1: 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 2: 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 3: 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

Overall, these results are mostly what I would have expected. All the algorithms estimate the true value of θ correctly in each case, with the oracle estimate having the smallest variance followed by the semiparametric algorithm. If there is anything surprising it is that the IPW estimator has better performance with the estimated weights compared to the true weights. However, I think that this is a known phenomenon.

- **Simulation 1 with Nonmonotone MAR**:

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

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

  Then,

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

  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 $p_{21} = \text{logistic}(y_2)$. The goal remains to estimate θ. We continue to use the Oracle algorithm and the IPW-Oracle algorithm. Since we have nonmonotone MAR data, we use the "Proposed" algorithm that is described on Slide 25 (Equation 12) of Dr. Kim's presentation. The outcome models were estimated using logistic regression and OLS and correctly specified. The response model used the oracle estimates of the probabilities. This yields the following results:

Table 4: 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 5: 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 6: 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 with Nonmonotone MAR**: We also want to simulate data that is correlated. For this simulation, we focus on $\text{Cov}(\varepsilon_1, \varepsilon_2)$. The data generating process now has $\rho_{yy} \neq 0$. 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 7: 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 8: 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 9: 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

- **Simulation 3 with Nonmonotone MAR**: This simulation aims to see if the proposed algorithm is doubly robust. First, we check with a misspecified outcome model. In this case the data generating procedure is the following:

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

Then, the true outcome model is,

$$y_{1i} = x_i + x_i^2 \varepsilon_{1i} \text{ and } y_{2i} = \theta - x_i + x_i^3 + \varepsilon_{2i}.$$

This procedure causes X to influence both Y_1 and Y_2 and we still have correlation in the error terms of Y_1 and Y_2. However, since neither Y_1 nor Y_2 are linear in X, the model will be misspecified. The response mechanisms are first generated MCAR with a probability of either Y_1 or Y_2 being the first variable observed to be 0.4. (There is a 0.2 probability neither is observed.) Then the probability of the other variable being observed is proportional to logistic(y_k) where y_k is the y that has been observed. To ensure that the proposed method has the correct propensity score we use the oracle probabilities instead of estimating them. This yields the following:

Table 10: True Value is 0. Cor(Y1, Y2) = 0

algorithm	bias	sd	tstat	pval
oracle	0.000	0.075	0.014	0.494
ipworacle	0.002	0.107	0.876	0.191
proposed	-0.002	0.084	-1.063	0.144

Table 11: True Value is 0. Cor(Y1, Y2) = 0.1

algorithm	bias	sd	tstat	pval
oracle	-0.002	0.074	-1.479	0.070
ipworacle	0.000	0.106	-0.196	0.422
proposed	-0.003	0.083	-1.464	0.072

Table 12: True Value is 0. Cor(Y1, Y2) = 0.5

algorithm	bias	sd	tstat	pval
oracle	-0.003	0.074	-1.567	0.059
ipworacle	-0.002	0.108	-0.818	0.207
proposed	-0.003	0.083	-1.633	0.051

Thus, the proposed method is unbiased with a misspecified outcome model. We now show a simulation where the outcome model is correctly specified, but the response model is not.

- **Simulation 4 with Nonmonotone MAR**: Continuing to test if the proposed algorithm is doubly robust, this simulation checks a misspecified response model. Instead of using oracle weights as in Simulation 3, we estimate the weights for the proposed method. However, unlike the true probabilities of being proportional to logistic(y_k), this simulation has the true probabilities being proportional to logistic(x_i). Thus, the true response model is the following:

  1. (Stage 1) Choose a variable observe. We choose Y_1 with probability 0.4, Y_2 with probability 0.4 and neither with probability 0.2. If neither, $R_1 = 0$ and $R_2 = 0$. Otherwise, continue to Step 2.
  2. (Stage 2) With probability $p_i \propto$ logistic(x_i), choose to observe the other Y variable.

  This sequence generates the missingness indicators R_1 and R_2. Since, the Stage 1 probabilities are fixed and known and the Stage 2 probabilities only depend on x_i, the missingness is MAR and only a function of x_i. The algorithms to which we compare still use the oracle weights.

Table 13: True Value is 0. Cor(Y1, Y2) = 0

algorithm	bias	sd	tstat	pval
oracle	0.000	0.032	-0.318	0.375
ipworacle	0.002	0.066	1.179	0.119
proposed	0.000	0.036	0.012	0.495

Table 14: True Value is 0. Cor(Y1, Y2) = 0.1

algorithm	bias	sd	tstat	pval
oracle	0	0.031	0.394	0.347
ipworacle	0	0.065	-0.230	0.409
proposed	0	0.035	-0.082	0.467

Table 15: True Value is 0. Cor(Y1, Y2) = 0.5

algorithm	bias	sd	tstat	pval
oracle	0	0.031	0.318	0.375
ipworacle	0	0.065	-0.094	0.462
proposed	0	0.036	-0.056	0.478

The previous version of this simulation used the Stage 2 probability $p_i \propto \text{logistic}(y_i)$ where y_i was the observed Y value in Stage 1. However, under this setup, our method is biased. This is because

$$E\left[E\left[\frac{R_1R_2}{\pi_{11}(X,Y_1)}Y_1 \mid X\right]\right] \neq E\left[\frac{E[Y_1 \mid X]}{\pi_{11}(X,Y_1)}E[R_1R_2 \mid X]\right]$$

but

$$E\left[E\left[\frac{R_1R_2}{\pi_{11}(X)}Y_1 \mid X\right]\right] = E\left[\frac{E[Y_1 \mid X]}{\pi_{11}(X)}E[R_1R_2 \mid X]\right].$$

Thus, there is strong evidence that the proposed method is doubly robust because it is robust to both misspecification in the outcome and response model.

Missingness Mechanism

- First I am going to reproduce the proof of double robustness that we talked about during our last meeting. I think it is insightful for future comments:

$$
\begin{aligned}
E[\hat{\theta}_{eff} - \theta_n] &= E\left[n^{-1}\sum_{i=1}^{n} E[g_i \mid X_i] - g_i\right] \\
&\quad + E\left[n^{-1}\sum_{i=1}^{n} \frac{R_{1i}}{\pi_{1+}(X_i)}(b_2(X_i, Y_{1i}) - E[g_i \mid X_i])\right] \\
&\quad + E\left[n^{-1}\sum_{i=1}^{n} \frac{R_{2i}}{\pi_{2+}(X_i)}(a_2(X_i, Y_{2i}) - E[g_i \mid X_i])\right] \\
&\quad + E\left[n^{-1}\sum_{i=1}^{n} \frac{R_{1i}R_{2i}}{\pi_{11}(X_i)}(g_i - a_2(X_i, Y_{2i}) - b_2(X_i, Y_{1i}) + E[g_i \mid X_i])\right] \\
&= n^{-1}\sum_{i=1}^{n}(E[E[g_i \mid X_i]] - E[g_i]) \\
&\quad + n^{-1}\sum_{i=1}^{n} E\left[E\left[\frac{R_{1i}}{\pi_{1+}(X_i)}(b_2(X_i, Y_{1i}) - E[g_i \mid X_i]) \mid X_i\right]\right] \\
&\quad + n^{-1}\sum_{i=1}^{n} E\left[E\left[\frac{R_{2i}}{\pi_{2+}(X_i)}(a_2(X_i, Y_{2i}) - E[g_i \mid X_i]) \mid X_i\right]\right] \\
&\quad + n^{-1}\sum_{i=1}^{n} E\left[E\left[\frac{R_{1i}R_{2i}}{\pi_{11}(X_i)}(g_i - a_2(X_i, Y_{2i}) - b_2(X_i, Y_{1i}) + E[g_i \mid X_i]) \mid X_i\right]\right]
\end{aligned}
$$

Since $R_{1i} \perp Y_{1i} \mid X_i$, $R_{2i} \perp Y_{2i} \mid X_i$, $(R_{1i}, R_{2i}) \perp (Y_{1i}, Y_{2i}) \mid X_i$ and π_{1+}, π_{2+} and π_{11} are all free of Y_{1i} and Y_{2i}.

$$
\begin{aligned}
&= n^{-1}\sum_{i=1}^{n} E\left[\frac{R_{1i}}{\pi_{1+}(X_i)} E\left[(E[g_i \mid X_i, Y_{1i}] - E[g_i \mid X_i]) \mid X_i\right]\right] \\
&\quad + n^{-1}\sum_{i=1}^{n} E\left[\frac{R_{2i}}{\pi_{2+}(X_i)} E\left[E[g_i \mid X_i, Y_{2i}] - E[g_i \mid X_i] \mid X_i\right]\right] \\
&\quad + n^{-1}\sum_{i=1}^{n} E\left[\frac{R_{1i}R_{2i}}{\pi_{11}(X_i)} E\left[(g_i - E[g_i \mid X_i, Y_{2i}] - E[g_i \mid X_i, Y_{1i}] + E[g_i \mid X_i]) \mid X_i\right]\right]
\end{aligned}
$$

Since $E[E[g_i \mid X_i, Y_{ki}] \mid X_i] = E[g_i \mid X_i] = 0$,

$$= 0.$$

Thus, if the outcome models are correctly specified $\hat{\theta}_{eff}$ is unbiased. If the response models are correctly specified it is easy to see that $\hat{\theta}_{eff}$ is also unbiased. This means that $\hat{\theta}_{eff}$ is doubly robust.

- However, one of the key steps is that *all* the response models are free of Y. In a previous iteration of Simulation 4, we had adopted the framework of [2] where we first to observe the first variable and see if we observe the second variable. In this case, the second step can depend on the result of the first step and this is what we did. However, this makes it the case that $\pi_1 1$ is a function of X_i and Y_1 and Y_2. In this case $\hat{\theta}_{eff}$ is not unbiased if the response model is misspecified (or even just estimated).

- If we modify Simulation 4, such that the second step of observed the second variable is proportional to logistic(y_k) then we get the same result as before:

Table 16: True Value is 0. Cor(Y1, Y2) = 0

algorithm	bias	sd	tstat	pval
oracle	0.000	0.032	-0.318	0.375
ipworacle	-0.001	0.079	-0.475	0.317
proposed	0.002	0.037	2.851	0.002

Table 17: True Value is 0. Cor(Y1, Y2) = 0.1

algorithm	bias	sd	tstat	pval
oracle	0.000	0.031	0.394	0.347
ipworacle	0.001	0.082	0.560	0.288
proposed	0.008	0.037	9.204	0.000

Table 18: True Value is 0. Cor(Y1, Y2) = 0.5

algorithm	bias	sd	tstat	pval
oracle	0.000	0.031	0.318	0.375
ipworacle	0.001	0.093	0.683	0.247
proposed	0.017	0.039	19.062	0.000

Minimizing the Variance

- The goal of this section is to find optimal values of $b_2(X, Y_1)$ and $a_2(X, Y_2)$ such that the variance of $\hat{\theta}_{eff}$ is minimized.
- Recall:

$$\begin{aligned}\hat{\theta}_{eff} - \hat{\theta}_n &= n^{-1}\sum_{i=1}^{n} E[g_i \mid X_i]\left(1 - \frac{R_{1i}}{\pi_{1+}} - \frac{R_{2i}}{\pi_{2+}} + \frac{R_{1i}R_{2i}}{\pi_{11}}\right) \\ &+ n^{-1}\sum_{i=1}^{n} b_2(X_i, Y_{1i})\left(\frac{R_{1i}}{\pi_{1+}} - \frac{R_{1i}R_{2i}}{\pi_{11}}\right) \\ &+ n^{-1}\sum_{i=1}^{n} a_2(X_i, Y_{2i})\left(\frac{R_{2i}}{\pi_{2+}} - \frac{R_{1i}R_{2i}}{\pi_{11}}\right) \\ &+ n^{-1}\sum_{i=1}^{n} g_i\left(\frac{R_{1i}R_{2i}}{\pi_{11}} - 1\right) \\ &\equiv A + B + C + D.\end{aligned}$$

- Notice that we will suppress the fact that response models are functions of X (i.e. we write π_{11} instead of $\pi_{11}(X)$).
- To compute the variance, we first solve for each covariance combination. Basically, all these computations rely on the following ideas. First, we assume that the response model is correctly specified. Consequently, $E[A] = E[B] = E[C] = E[D] = 0$ and things work out better. This helps when we take the covariance conditional on X because the inner expectations are zero. The second key insight it to notice that $E[R_j^k] = E[R_j]$ for $j \in \{1, 2\}$ and $k \in \mathbb{N}$. This is because R is a binary variable. Third, since we assume that the response models are correctly specified, we have $E[R_1 \mid X] = \pi_{1+}$, $E[R_2 \mid X] = \pi_{2+}$, and $E[R_1, R_2 \mid X] = \pi_{11}$.

  The overall approach to each of these computations is the following: (1) take conditional expectations with respect to X (the $\text{Cov}(E[\cdot])$ term is zero), (2) expand the covariance to $E[XY] - E[X]E[Y]$ (the second term is also zero), (3) by the MAR assumption g, a_2, b_2 are independent of R_1 and R_2 and we can take the latter out of the expectation, (4) evaluate and simplify expressions involving $E[R]$.

$$\begin{aligned}\text{Cov}(A, B) &= n^{-2}\sum_{i=1}^{n} E\left[\text{Cov}\left(E[g_i \mid X]\left(1 - \frac{R_{1i}}{\pi_{1+}} - \frac{R_{2i}}{\pi_{2+}} + \frac{R_{1i}R_{2i}}{\pi_{11}}\right) \mid X_i,\right.\right. \\ &\qquad \left.\left. b_2(X_i, Y_{1i})\left(\frac{R_{1i}}{\pi_{1+}} - \frac{R_{1i}R_{2i}}{\pi_{11}}\right) \mid X_i\right)\right] \\ &= n^{-1}E\left[E\left[E[g \mid X]\left(1 - \frac{R_{1i}}{\pi_{1+}} - \frac{R_{2i}}{\pi_{2+}} + \frac{R_{1i}R_{2i}}{\pi_{11}}\right) b_2(X_i, Y_{1i})\left(\frac{R_{1i}}{\pi_{1+}} - \frac{R_{1i}R_{2i}}{\pi_{11}}\right) \mid X\right]\right] \\ &= n^{-1}E\left[E[g \mid X]E[b_2(X, Y_1) \mid X]\left(\frac{1}{\pi_{1+}} + \frac{1}{\pi_{2+}} - \frac{1}{\pi_{11}} - \frac{\pi_{11}}{\pi_{1+}\pi_{2+}}\right)\right].\end{aligned}$$

By symmetry,

$$\mathrm{Cov}(A,C) = n^{-1}E\left[E[g \mid X]E[a_2(X,Y_2) \mid X]\left(\frac{1}{\pi_{1+}} + \frac{1}{\pi_{2+}} - \frac{1}{\pi_{11}} - \frac{\pi_{11}}{\pi_{1+}\pi_{2+}}\right)\right].$$

$$\begin{aligned}\mathrm{Cov}(A,D) &= n^{-1}E\left[E\left[E[g \mid X]\left(1 - \frac{R_{1i}}{\pi_{1+}} - \frac{R_{2i}}{\pi_{2+}} + \frac{R_{1i}R_{2i}}{\pi_{11}}\right)g\left(\frac{R_1R_2}{\pi_{11}} - 1\right) \mid X\right]\right] \\ &= n^{-1}E\left[E[g \mid X]^2\left(\frac{-1}{\pi_{1+}} - \frac{1}{\pi_{2+}} + 2\right)\right].\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(B,C) &= n^{-1}E\left[E[b_2(X,Y_1) \mid X]E[a_2(X,Y_2) \mid X]E\left[\left(\frac{R_1}{\pi_{1+}} - \frac{R_1R_2}{\pi_{11}}\right)\left(\frac{R_2}{\pi_{2+}} - \frac{R_1R_2}{\pi_{11}}\right) \mid X\right]\right] \\ &= n^{-1}E\left[E[b_2(X,Y_1) \mid X]E[a_2(X,Y_2) \mid X]\left(\frac{\pi_{11}}{\pi_{1+}\pi_{2+}} - \frac{1}{\pi_{1+}} - \frac{1}{\pi_{2+}} + \frac{1}{\pi_{11}}\right)\right].\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(B,D) &= n^{-1}E\left[E[b_2(X,Y_1) \mid X]E[g \mid X]E\left[\left(\frac{R_1}{\pi_{1+}} - \frac{R_1R_2}{\pi_{11}}\right)\left(\frac{R_1R_2}{\pi_{11}} - 1\right) \mid X\right]\right] \\ &= n^{-1}E\left[E[b_2(X,Y_1) \mid X]E[g \mid X]\left(\frac{1}{\pi_{1+}} - \frac{1}{\pi_{11}}\right)\right].\end{aligned}$$

By symmetry,

$$\mathrm{Cov}(C,D) = n^{-1}E\left[E[a_2(X,Y_2) \mid X]E[g \mid X]\left(\frac{1}{\pi_{2+}} - \frac{1}{\pi_{11}}\right)\right].$$

We also compute the variance terms,

$$\begin{aligned}\mathrm{Cov}(A,A) &= n^{-1}E\left[E\left[E[g \mid X]^2\left(1 - \frac{R_{1i}}{\pi_{1+}} - \frac{R_{2i}}{\pi_{2+}} + \frac{R_{1i}R_{2i}}{\pi_{11}}\right)^2 \mid X\right]\right] \\ &= n^{-1}E\left[E[g \mid X]^2\left(-1 + \frac{2\pi_{11}}{\pi_{1+}\pi_{2+}} - \frac{1}{\pi_{1+}} - \frac{1}{\pi_{2+}} + \frac{1}{\pi_{11}}\right)\right].\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(B,B) &= n^{-1}E\left[E\left[b(X,Y_1)^2\left(\frac{R_1}{\pi_{1+}} - \frac{R_1R_2}{\pi_{11}}\right)^2 \mid X\right]\right] \\ &= n^{-1}E\left[E[b_2(X,Y_1)^2 \mid X]\left(\frac{-1}{\pi_{1+}} + \frac{1}{\pi_{11}}\right)\right].\end{aligned}$$

$$\mathrm{Cov}(C,C) = n^{-1}E\left[E[a_2(X,Y_2)^2 \mid X]\left(\frac{-1}{\pi_{2+}} + \frac{1}{\pi_{11}}\right)\right].$$

$$\begin{aligned}\mathrm{Cov}(D,D) &= n^{-1}E\left[E\left[g_i^2\left(\frac{R_1R_2}{\pi_{11}}-1\right)^2 \mid X\right]\right]\\ &= n^{-1}E\left[E[g^2 \mid X]\left(\frac{1}{\pi_{11}}-1\right)\right].\end{aligned}$$

This means that

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_{eff}-\hat{\theta}_n) &= \mathrm{Cov}(A,A)+2\mathrm{Cov}(A,B)+2\mathrm{Cov}(A,C)+2\mathrm{Cov}(A,D)+\mathrm{Cov}(B,B)\\ &+2\mathrm{Cov}(B,C)+2\mathrm{Cov}(B,D)+\mathrm{Cov}(C,C)+2\mathrm{Cov}(C,D)+\mathrm{Cov}(D,D)\\ &= n^{-1}E\left[E[g \mid X]^2\left(-1+\frac{2\pi_{11}}{\pi_{1+}\pi_{2+}}-\frac{1}{\pi_{1+}}-\frac{1}{\pi_{2+}}+\frac{1}{\pi_{11}}\right)\right]\\ &+2n^{-1}E\left[E[g \mid X]E[b_2(X,Y_1) \mid X]\left(\frac{1}{\pi_{1+}}+\frac{1}{\pi_{2+}}-\frac{1}{\pi_{11}}-\frac{\pi_{11}}{\pi_{1+}\pi_{2+}}\right)\right]\\ &+2n^{-1}E\left[E[g \mid X]E[a_2(X,Y_2) \mid X]\left(\frac{1}{\pi_{1+}}+\frac{1}{\pi_{2+}}-\frac{1}{\pi_{11}}-\frac{\pi_{11}}{\pi_{1+}\pi_{2+}}\right)\right]\\ &+2n^{-1}E\left[E[g \mid X]^2\left(\frac{-1}{\pi_{1+}}-\frac{1}{\pi_{2+}}+2\right)\right]\\ &+n^{-1}E\left[E[b_2(X,Y_1)^2 \mid X]\left(\frac{-1}{\pi_{1+}}+\frac{1}{\pi_{11}}\right)\right]\\ &+2n^{-1}E\left[E[b_2(X,Y_1) \mid X]E[a_2(X,Y_2) \mid X]\left(\frac{\pi_{11}}{\pi_{1+}\pi_{2+}}-\frac{1}{\pi_{1+}}-\frac{1}{\pi_{2+}}+\frac{1}{\pi_{11}}\right)\right]\\ &+2n^{-1}E\left[E[b_2(X,Y_1) \mid X]E[g \mid X]\left(\frac{1}{\pi_{1+}}-\frac{1}{\pi_{11}}\right)\right]\\ &+n^{-1}E\left[E[a_2(X,Y_2)^2 \mid X]\left(\frac{-1}{\pi_{2+}}+\frac{1}{\pi_{11}}\right)\right]\\ &+2n^{-1}E\left[E[a_2(X,Y_2) \mid X]E[g \mid X]\left(\frac{1}{\pi_{2+}}-\frac{1}{\pi_{11}}\right)\right]\\ &+n^{-1}E\left[E[g^2 \mid X]\left(\frac{1}{\pi_{11}}-1\right)\right].\end{aligned}$$

Differentiating yields:

$$\begin{aligned}
\frac{\partial}{\partial a_2}\mathrm{Var}(\hat{\theta}_{eff} - \hat{\theta}_n) &= E\left[E[g \mid X]\left(\frac{1}{\pi_{1+}} + \frac{2}{\pi_{2+}} - \frac{2}{\pi_{11}} - \frac{\pi_{11}}{\pi_{1+}\pi_{2+}}\right)\right] \\
&\quad + E\left[E[b_2(X, Y_1) \mid X]\left(\frac{\pi_{11}}{\pi_{1+}\pi_{2+}} - \frac{1}{\pi_{1+}} - \frac{1}{\pi_{2+}} + \frac{1}{\pi_{11}}\right)\right] \\
&\quad + E\left[E[a_2(X, Y_2) \mid X]\left(\frac{-1}{\pi_{2+}} + \frac{1}{\pi_{11}}\right)\right] \\
&\equiv 0, \text{ and} \\
\frac{\partial}{\partial b_2}\mathrm{Var}(\hat{\theta}_{eff} - \hat{\theta}_n) &= E\left[E[g \mid X]\left(\frac{2}{\pi_{1+}} + \frac{1}{\pi_{2+}} - \frac{2}{\pi_{11}} - \frac{\pi_{11}}{\pi_{1+}\pi_{2+}}\right)\right] \\
&\quad + E\left[E[a_2(X, Y_2) \mid X]\left(\frac{\pi_{11}}{\pi_{1+}\pi_{2+}} - \frac{1}{\pi_{1+}} - \frac{1}{\pi_{2+}} + \frac{1}{\pi_{11}}\right)\right] \\
&\quad + E\left[E[b_2(X, Y_2) \mid X]\left(\frac{-1}{\pi_{1+}} + \frac{1}{\pi_{11}}\right)\right] \\
&\equiv 0.
\end{aligned}$$

Substitution shows that these constraints are equivalent to:

$$\begin{aligned}
&E\left[E[g \mid X]\left(\frac{-1}{\pi_{1+}} + \frac{1}{\pi_{2+}}\right)\right] + E\left[E[b_2(X, Y_1) \mid X]\left(\frac{\pi_{11}}{\pi_{1+}\pi_{2+}} - \frac{1}{\pi_{2+}}\right)\right] \\
&\quad - E\left[E[a_2(X, Y_2) \mid X]\left(\frac{\pi_{11}}{\pi_{1+}\pi_{2+}} - \frac{1}{\pi_{1+}}\right)\right] \equiv 0
\end{aligned}$$

where is the same as,

$$\begin{aligned}
&E\left[E[b_2(X, Y_1) \mid X]\left(\frac{\pi_{11}}{\pi_{1+}\pi_{2+}} - \frac{1}{\pi_{2+}}\right)\right] + E\left[E[g \mid X]\left(\frac{1}{\pi_{2+}}\right)\right] \\
&\quad = E\left[E[a_2(X, Y_2) \mid X]\left(\frac{\pi_{11}}{\pi_{1+}\pi_{2+}} - \frac{1}{\pi_{1+}}\right)\right] + E\left[E[g \mid X]\left(\frac{1}{\pi_{1+}}\right)\right].
\end{aligned}$$

- These constraints can be satisfied (this is sufficient but maybe not necessary) if

$$\begin{aligned}
&E\left[\left(E[b_2(X, Y_1) \mid X] - E[a_2(X, Y_2) \mid X]\right)\left(\frac{\pi_{11}}{\pi_{1+}\pi_{2+}}\right)\right] = 0 \\
&E\left[\left(\frac{1}{\pi_{1+}} - \frac{1}{\pi_{2+}}\right)\left(E[a_2(X, Y_2) \mid X] + E[b_2(X, Y_1) \mid X] - 2E[g \mid X]\right)\right] = 0.
\end{aligned}$$

Comparison with Calibration Estimator

Monotone Case

- In the monotone case the efficient estimator is

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

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

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

- The three-phase regression estimator of θ:

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

- We can view the above three-phase regression estimator as a projection (= mass imputation) estimator of Kim and Rao (2012, Biometrika).

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

$$\begin{aligned}&\text{argmin}_w \sum_{i=1}^{n} w_i^2 \text{ such that} \\ &\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} R_{1i} w_{1i} x_i \\ &\sum_{i=1}^{n} w_{1i}(x_i, y_{1i}) = \sum_{i=1}^{n} R_{1i}R_{2i} w_{2i}(x_i, y_{1i})\end{aligned}$$

- The reason that these should be the same is because they are similar in relationship to a calibration and regression estimator which are exactly the same.
- 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 $p_{1i} \propto \text{logistic}(x_i)$. If Y_{1i} is observed, then we observe Y_{2i} with probability $p_{2i} \propto \text{logistic}(y_{1i})$. If Y_{1i} is not observed, we do not observe Y_{2i}.

- The goal of this simulation study is the estimate $\theta = E[Y_2]$. We use the previous monotone data generating process with different true values of θ and compute the bias, standard deviation, T-statistic and p-value. (The T-statistic and p-value test if the estimated value of $\hat{\theta}$ is significantly different from the true value of θ.)

Table 19: 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 20: 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 21: 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

Nonmonotone Case

- Similar to the monotone case, we have an idea of the efficient estimator. Now we want to show that it is similar to a calibration equation. Unlike the monotone case where $R_{1i} = 0$ implies $R_{2i} = 0$, the nonmonotone case does not have this relationship. Instead we believe that we have the following calibration equations:

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

- We still have the same goal of the simulation study: estimate $\theta = E[Y_2]$. We use the previous nonmonotone data generating process with different true values of θ to estimate the bias, standard deviation, T-statistic, and p-value. For clarity here is a reminder of the simulation setup.

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

$$x_i \overset{iid}{\sim} N(0,1)$$
$$\varepsilon_{1i} \overset{iid}{\sim} N(0,1)$$
$$\varepsilon_{2i} \overset{iid}{\sim} N(\theta,1)$$

     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 $p \propto \text{logistic}(Y_k)$ 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, $R_k = 0$.

- For this simulation setup, we estimate $\theta = E[Y_2]$. Like the previous nonmonotone simulations, we compare this calibration estimator to the oracle estimator which uses the average value of Y_2 if $R_2 = 0$ or $R_2 = 1$, an IPW estimator with the correct weights, and the proposed regression estimator. These are currently run with a sample size of $n = 1000$ with the number of Monte Carlo simulations of $B = 1000$.

Table 22: 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 23: 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 24: 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

Efficiency of Proposed Estimator

One of our main goals is to show that we have an efficient estimator. We want to ensure that our estimator is more efficient (lower MSE or really zero bias and a smaller variance) than other competing estimators. We have already demonstrated superior performance compared to the IPW estimator with known weights. Now, we want to compare our estimator with two-phase and three-phase regression estimators.

Two and Three Phase Regression Estimators

- The simulation setup that we have been running generates monotone and nonmonotone missing patterns because we always observe X, yet we observe Y_1 and Y_2 with missingness. This creates the need to extend the traditional two-phase estimator to become a three-phase estimator. The two-phase regression estimator is really just a regression estimator with the "finite population" being Phase 1 of the sample. Thus, a valid two-phase estimator for $\theta = E[Y_2]$ is

$$\hat{\theta} = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2)\hat{\beta}$$

where $\bar{y}_2 = n_2^{-1} \sum_{i \in U} I(i \in A_2) y_2$ and $\bar{x}_k = n_k^{-1} \sum_{i \in U} I(i \in A_k) x_i$ where n_k is the number of elements in A_k where A_k is the Phase k sample. In this case, $\hat{\beta}$ solves the following equation for β_0 and β_1,

$$\sum_{i \in A_2} (y_{2i} - \beta_0 - \beta_1 x_i)^2 = 0.$$

- Notice that the previous construction of the two-phase estimator ignored the variable Y_1. We can incorporate this into the model using a three-phase estimator. From [1], the three-phase estimator is

$$\bar{y}_{2,2p} = \bar{y}_2 + (\bar{x}_0 - \bar{x}_2)\hat{\beta}_1 + (\bar{y}_{1,reg} - \bar{y}_{1,2p})\hat{\beta}_2$$

where

$$\bar{x}_0 = n^{-1} \sum_{i \in U} x_i$$

$$\bar{x}_1 = \left(\sum_{i \in A_1} \pi_{1i}^{-1} \right)^{-1} \sum_{i \in A_1} \pi_{1i}^{-1} x_i$$

$$\bar{x}_2 = \left(\sum_{i \in A_2} \pi_{2i}^{-1} \right)^{-1} \sum_{i \in A_2} \pi_{2i}^{-1} x_i$$

$$\bar{y}_{1,reg} = \bar{y}_{1,1p} + (\bar{x}_0 - \bar{x}_1)\hat{\beta}_{1p}$$

$$\bar{y}_{1,1p} = \left(\sum_{i \in A_1} \pi_{1i}^{-1} \right)^{-1} \sum_{i \in A_1} \pi_{1i}^{-1} y_{1i}$$

$$\bar{y}_{1,2p} = \left(\sum_{i \in A_2} \pi_{2i}^{-1} \right)^{-1} \sum_{i \in A_2} \pi_{2i}^{-1} y_{1i}$$

$$\hat{\beta}_{1p} = \left(\sum_{i \in A_1} (x_i - \bar{x}_1)^2 \pi_{1i}^{-1} \right)^{-1} \sum_{i \in A_1} (x_i - \bar{x}_1) \pi_{1i}^{-1} (y_{1i} - \bar{y}_{1,1p})$$

$$\hat{\beta}_1 = \left(\sum_{i \in A_2} (x_i - \bar{x}_2, y_{1i} - \bar{y}_{1,2p})' \pi_{2i}^{-1} (x_i - \bar{x}_2, y_{1i} - \bar{y}_{1,2p}) \right)^{-1} \sum_{i \in A_2} (x_i - \bar{x}_2) \pi_{2i}^{-1} (y_{2i} - \bar{y}_2)$$

$$\hat{\beta}_2 = \left(\sum_{i \in A_2} (x_i - \bar{x}_2, y_{1i} - \bar{y}_{1,2p})' \pi_{2i}^{-1} (x_i - \bar{x}_2, y_{1i} - \bar{y}_{1,2p}) \right)^{-1} \sum_{i \in A_2} (y_{1i} - \bar{y}_{1,2p}) \pi_{2i}^{-1} (y_{2i} - \bar{y}_2)$$

- Notice that the three-phase estimator implicitly assumes a monotone missingness model. It ignores the values of y_2 that have an unobserved y_1.

Monotone Results

- First, we test the two-phase and three-phase regression estimators with a monotone simulation. Like the previous monotone simulations, we first generate data from the

following distributions:

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

Then, we create the probabilities $p_1 = \text{logistic}(x_i)$ and $p_{12} = \text{logistic}(y_{1i})$. Including the calibration estimator from the previous section (note, these are the same as the previous monotone tables) this yields,

Table 25: 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 26: 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

- The proposed semiparametric estimator and the calibration estimator seem to both outperform the regression estimators. With smaller bias and smaller variance our estimators do better.
- It may be puzzling to see that the three-phase estimator does worse than the two-phase; however, I think that there is a good reason for this. In the simulation, $y_2 = x + \varepsilon$. This is the regression from the two-phase estimator. However, the three-phase estimator measures $y_2 \sim x + y_1 + \varepsilon$, which just adds noise. This is why I think the standard deviation of the three-phase regression estimator is larger.

Table 27: 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

Nonmonotone Results

- Similar to the monotone results, we also use the same simulation as the nonmonotone calibration estimators. Repeating the simulation outline, we are trying to estimate $\theta = E[y_2]$ and we have

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

  $$x_i \overset{iid}{\sim} N(0,1)$$
  $$\varepsilon_{1i} \overset{iid}{\sim} N(0,1)$$
  $$\varepsilon_{2i} \overset{iid}{\sim} N(\theta,1)$$

  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 $R_{2i} = 0$. Otherwise, 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 $p \propto \text{logistic}(Y_k)$ 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, $R_k = 0$.

- Overall, it seems that the proposed estimator and calibration estimator outperform the two regression estimators. The regression estimators display slight bias (perhaps because of the nonmonotonicity).

Table 28: 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 29: 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 30: 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

Multiphase Estimators

- We notice that there are two multiphase estimators that we want to compare. The first is the following:

- The three-phase regression estimator of θ:

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

- We can view the above three-phase regression estimator as a projection (= mass imputation) estimator of Kim and Rao (2012, Biometrika).

- The second multiphase estimator is the previous three-phase regression estimator:

$$\bar{y}_{2,2p} = \bar{y}_2 + (\bar{x}_0 - \bar{x}_2)\hat{\beta}_1 + (\bar{y}_{1,reg} - \bar{y}_{1,2p})\hat{\beta}_2$$

where

$$\bar{x}_0 = n^{-1}\sum_{i\in U} x_i$$

$$\bar{x}_1 = \left(\sum_{i\in A_1}\pi_{1i}^{-1}\right)^{-1}\sum_{i\in A_1}\pi_{1i}^{-1}x_i$$

$$\bar{x}_2 = \left(\sum_{i\in A_2}\pi_{2i}^{-1}\right)^{-1}\sum_{i\in A_2}\pi_{2i}^{-1}x_i$$

$$\bar{y}_{1,reg} = \bar{y}_{1,1p} + (\bar{x}_0 - \bar{x}_1)\hat{\beta}_{1p}$$

$$\bar{y}_{1,1p} = \left(\sum_{i\in A_1}\pi_{1i}^{-1}\right)^{-1}\sum_{i\in A_1}\pi_{1i}^{-1}y_{1i}$$

$$\bar{y}_{1,2p} = \left(\sum_{i\in A_2}\pi_{2i}^{-1}\right)^{-1}\sum_{i\in A_2}\pi_{2i}^{-1}y_{1i}$$

$$\hat{\beta}_{1p} = \left(\sum_{i\in A_1}(x_i - \bar{x}_1)^2\pi_{1i}^{-1}\right)^{-1}\sum_{i\in A_1}(x_i - \bar{x}_1)\pi_{1i}^{-1}(y_{1i} - \bar{y}_{1,1p})$$

$$\hat{\beta}_1 = \left(\sum_{i\in A_2}(x_i - \bar{x}_2, y_{1i} - \bar{y}_{1,2p})'\pi_{2i}^{-1}(x_i - \bar{x}_2, y_{1i} - \bar{y}_{1,2p})\right)^{-1}\sum_{i\in A_2}(x_i - \bar{x}_2)\pi_{2i}^{-1}(y_{2i} - \bar{y}_2)$$

$$\hat{\beta}_2 = \left(\sum_{i\in A_2}(x_i - \bar{x}_2, y_{1i} - \bar{y}_{1,2p})'\pi_{2i}^{-1}(x_i - \bar{x}_2, y_{1i} - \bar{y}_{1,2p})\right)^{-1}\sum_{i\in A_2}(y_{1i} - \bar{y}_{1,2p})\pi_{2i}^{-1}(y_{2i} - \bar{y}_2)$$

- Question: Are these estimators the same?

References

[1] Wayne A Fuller. *Sampling statistics*. John Wiley & Sons, 2009.

[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.