Modeling for Nonresponse and Measurement Error with Survey-Weighted Data When Some Margins Are Known From External Data Sources

Jerry Reiter
Department of Statistical Science
Duke University

Collaborators: Jiurui Tang, Olanrewaju Akande, Gabriel Madson, and Sunshine Hillygus

Research supported by NSF SES-1733835

Motivation

- Many surveys have seen a steep decline in response rates.
- Yet less resources available for nonresponse follow-up.
- Agencies forced to account for missing values using methods that rely on strong assumptions. Examples:
  - Missing at random (MAR)
  - Not missing at random (NMAR) according to a selection model with a restrictive specification
- Such assumptions can be unrealistic.

What About "Big Data"?

- Digital revolution has seen a proliferation of information on individuals and populations.
    - Censuses
    - Administrative databases
    - Private sector data aggregators
- How can agencies use such information when accounting for missing data in their particular surveys?

How Might Such Data Help? A Simple Illustration

- We have simple random sample where question on sex suffers from item nonresponse.
    - 70% of respondents report “female”
- We know from auxiliary information that the target population includes around 50% “female” and 50% “male.”
- We likely should impute more “male” than “female” to get the empirical margin in the completed data closer to 50-50.
- Without using margin, we easily could generate unreliable imputations
    - e.g., MAR model likely to impute more “female” than ”male”

Generalizing the Illustration

- Agency has accurate estimates of population percentages or counts for some variables in the survey.
- Agency seeks to take advantage of this auxiliary information in its methods for handling missing values, which could be due to both item and unit nonresponse.
- Agencies routinely find themselves in this scenario.
    - Population counts used as the basis for post-stratification adjustments for unit nonresponse
    - Usually not used in imputations for item nonresponse.

Research Goal

- Develop a framework that leverages population-based marginal information to:
  - Allow for more flexible modeling and potentially nonignorable nonresponse mechanisms.
  - Handle both item and unit nonresponse simultaneously.
  - Allow distinct specifications of missingness mechanisms for different blocks of variables.
- Use Bayesian modeling or multiple imputation to propagate uncertainty.

Notation for Survey Variables

- $\mathcal{D}$ comprises data from survey of $i = 1, \ldots, n$ individuals.
- $\mathcal{A}$ comprises information from auxiliary database.
- $X = (X_1, \ldots, X_p)$ represents the p variables in $\mathcal{A}$ and $\mathcal{D}$.
- $Y = (Y_1, \ldots, Y_q)$ represents the q variables in $\mathcal{D}$ but not $\mathcal{A}$.

$\mathcal{A}$ contains sets of marginal probabilities for some X_k.

Notation for Response Indicators

- $U_i = 1$ if individual i does not respond to the survey, and $U_i = 0$ otherwise.
- $R^x_{ik} = 1$ if individual i would not respond to question on X_k in the survey, and $R^x_{ik} = 0$ otherwise.
- $R^y_{ik} = 1$ if individual i would not respond to question on Y_k in the survey, and $R^y_{ik} = 0$ otherwise.

The MD-AM Framework

- MD-AM = Missing Data with Auxiliary Margins
- Characterize joint distribution of survey variables and indicators for nonresponse using a factorization of sequential conditional models.
- Allow $\mathcal{A}$ to guide the specification of the conditional distributions, using models that encode potentially nonignorable nonresponse mechanisms.
- Require the models to be identifiable as described in Sadinle and Reiter (2019).
- Two-step approach for specifying joint distribution.

The MD-AM Framework

Step 1: Specify model for the observed data

- Specify a model for the survey variables and nonresponse indicators that is identifiable from the observed data alone.
- Generally, use default choices for handling nonresponse absent auxiliary data.
- I show results for selection model specifications, but also could use pattern mixture models.

The MD-AM Framework

Step 2: Incorporate auxiliary margins

- Add sets of parameters to the conditional models in Step 1, ensuring that the model as a whole still can be identified with the auxiliary information.
- Typically, multiple identifiable models, determined by the nature of $\mathcal{A}$.
- Choose among these models according to interpretability and plausibility for data at hand.

Simple Illustrative Example

Suppose we have data comprising two binary variables, with no unit nonrespondents and X_1 subject to item nonresponse. We know the auxiliary marginal distribution for X_1 but not for Y_1.

	X_1	Y_1	R_1^x
Observed Data	✓	✓	0
Observed Data	?	✓	1
Auxiliary margin →	✓	?	?

"✓" represents observed, and "?" represents missing.

Applying the MD-AM Framework

Step 1: Specify model for the observed data

$$(X_1, Y_1) \sim f(X_1, Y_1|\Theta) \tag{1}$$

$$\Pr(R_1^x = 1|X_1, Y_1) = h_1(\eta_0 + \eta_1 Y_1). \tag{2}$$

Step 2: Incorporate auxiliary margins

Auxiliary margin $\Pr(X_1 = 1)$ provides one additional piece of information. We can add one term involving X_1 to (2),

$$\Pr(R_1^x = 1|X_1, Y_1) = h_1(\eta_0 + \eta_1 Y_1 + \eta_2 X_1). \tag{3}$$

Additive Nonignorable Model

- This is the additive nonignorable (AN, Hirano et al., 2001) model developed for attrition in longitudinal studies with refreshment samples
- Interaction between X_1 and Y_1 disallowed to enable identification.
- Special cases of the AN model are informative.
  1. $(\eta_1 = 0, \eta_2 = 0)$ results in MCAR.
  2. $(\eta_1 \neq 0, \eta_2 = 0)$ results in MAR.
  3. $\eta_2 \neq 0$ results in NMAR

MD-AM Framework in Practice

- AN model weakens reliance on assumptions compared to MAR and NMAR models that are special cases.
- For each variable with a univariate auxiliary margin, can add one main effect parameter belonging to that variable
- If we instead know $Pr(X_1, Y_1)$, we have two additional pieces of information about the joint distribution and can add $\eta_2 X_1$ and $\eta_3 X_1 Y_1$.

Now With Unit Nonresponse

Two binary variables with X_1 and X_2 subject to item nonresponse. Data have unit nonresponse as well. Univariate margins for X_1 and X_2 known.

<table>
<tr><th></th><th>X_1</th><th>X_2</th><th>R_1^x</th><th>R_2^x</th><th>U</th></tr>
<tr><td rowspan="5">Observed Data</td><td>✓</td><td>✓</td><td>0</td><td rowspan="2">0</td><td rowspan="4">0</td></tr>
<tr><td>?</td><td>✓</td><td>1</td></tr>
<tr><td>✓</td><td>?</td><td>0</td><td rowspan="2">1</td></tr>
<tr><td>?</td><td>?</td><td>1</td></tr>
<tr><td>?</td><td>?</td><td>?</td><td>?</td><td>1</td></tr>
<tr><td>Auxiliary margin →</td><td>✓</td><td>?</td><td>?</td><td>?</td><td>?</td></tr>
<tr><td>Auxiliary margin →</td><td>?</td><td>✓</td><td>?</td><td>?</td><td>?</td></tr>
</table>

Joint Distribution

- Joint distribution can be fully parameterized using 31 parameters (and sum to one constraint)
  - $\theta_{xr_2r_1u} = \Pr(X_2 = 1|X_1 = x, R_2^x = r_2, R_1^x = r_1, U = u)$,
  - $\pi_{r_2r_1u} = \Pr(X_1 = 1|R_2^x = r_2, R_1^x = r_1, U = u)$,
  - $q_{r_1u} = \Pr(R_2^x = 1|R_1^x = r_1, U = u)$,
  - $s_u = \Pr(R_1^x = 1|U = u)$, and $p = \Pr(U = 1)$.
- We can uniquely estimate eight parameters from the observed data alone: $p, s_0, q_{00}, q_{10}, \pi_{000}, \pi_{100}, \theta_{0000}$, and θ_{1000}.
- Auxiliary margins $\Pr(X_1 = 1)$ and $\Pr(X_2 = 1)$ add two constraints, allowing us to add two parameters.

MD-AM Framework Step 1

- A default is the specification:

$$(X_1, X_2) \sim f(X_1, X_2|\Theta) \quad (4)$$

$$\Pr(U = 1|X_1, X_2) = g(\eta_0) \quad (5)$$

$$\Pr(R_1^x = 1|X_1, X_2, U) = h_1(\zeta_0 + \zeta_1 X_2) \quad (6)$$

$$\Pr(R_2^x = 1|X_1, X_2, U, R_1^x) = h_2(\gamma_0 + \gamma_1 X_1). \quad (7)$$

- At most eight parameters, which is the maximum identifiable from $\mathcal{D}^{obs}$ alone.
- Special case of the itemwise conditionally independent (ICIN) mechanism (Sadinle and Reiter, 2017).

MD-AM Framework Step 2

We have several options for adding the two parameters.

1. Add to unit nonresponse model:

$$\Pr(U = 1 | X_1, X_2) = g(\eta_0 + \eta_1 X_1 + \eta_2 X_2). \tag{8}$$

2. Add to item nonresponse models:

$$\Pr(R_1^x = 1 | X_1, X_2, U) = h_1(\zeta_0 + \zeta_1 X_2 + \zeta_2 X_1) \tag{9}$$

$$\Pr(R_2^x = 1 | X_1, X_2, U, R_1^x) = h_2(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2). \tag{10}$$

3. Add one parameter to unit nonresponse model and one to item nonresponse model (not shown here)

Tailoring Missingness Mechanisms via MD-AM

- Two choices for each variable with auxiliary margin
  - Encode a relationship between the variable and its item nonresponse indicator
  - Or, encode a relationship between the variable and its unit nonresponse indicator
- Tailor nonresponse models for unit and item nonresponse to the data at hand
  - e.g., more flexibility to model for U when unit nonresponse greater concern than item nonresponse, or vice versa
- No need to pick only one: MD-AM framework amenable to sensitivity analysis

Application to Voter Turnout

- Current Population Survey allows one to estimate voter turnout by state and demographic sub-groups.

	Unit	Item		
		Vote	Sex	Age
FL	.28	.18	.00	.07
GA	.21	.16	.00	.05
NC	.24	.11	.00	.03
SC	.25	.10	.00	.03

Table 1. Person-level unit and item nonresponse rates by state in the CPS data. Only 7 total cases (six in FL and 1 in SC) are missing sex.

Details about this application in Akande et al. (2021).

Our Alternative: Use MD-AM Framework!

- Voter turnout for complete cases in CPS (unweighted):
  FL = 75%, GA = 73%, NC = 77% and SC = 73%.
- Auxiliary margins for turnout by state:
  FL = 62.8%, GA = 59%, NC = 64.8 % and SC = 56.3%.
- Auxiliary margins for age by state from the 2010 census (adjusted to remove population ineligible to vote).

Notation

- S = state (FL, GA, NC, SC)
- G = sex (0 = male; 1 = female)
- A = age (1 = 18 - 29; 2 = 30 - 49; 3 = 50 - 69; 4 = 70+)
- V = vote (0 = did not vote; 1 = voted)
- U = unit nonresponse indicator
- R^G = item nonresponse indicator for sex
- R^A = item nonresponse indicator for age
- R^V = item nonresponse indicator for vote

Model for (G, A, V)

- Model for G
  - Logistic regression on S
- Model for A
  - Ordered logistic regression on main effects of S and G
- Model for V
  - Logistic regression on main effects for (S, G, A), and interactions of (S, A)

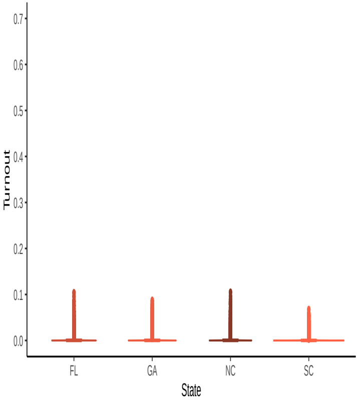
Model for (U, R^G, R^A, R^V)

- Model for U: logistic regression on S and A
- Model for R^G: Logistic regression on S
- Model for R^A: Logistic regression on S, G, V
- Model for R^V: Logistic regression S, G, A, and V
- Rationale for model choices
  - Simple model for R^G since only 7 missing values
  - For A margin, add term to model for U since unit nonresponse more prevalent than item nonresponse for age.
  - For V margin, add term to model for R^V in case people who do not vote are more likely not to answer (MD-R)
  - For V margin, instead add term to model for U (MD-U)

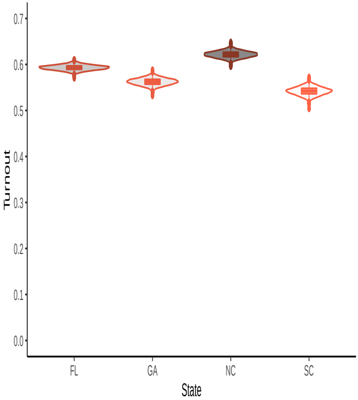
Estimation

- Non-informative priors for all parameters.
- Base inferences on MCMC sampler using 5,000 post burn-in posterior samples.
- Incorporate marginal information by augmenting the observed data with data generated to match the marginals (Schifeling and Reiter, 2016).

Turnout Among Nonrespondents: MD-R

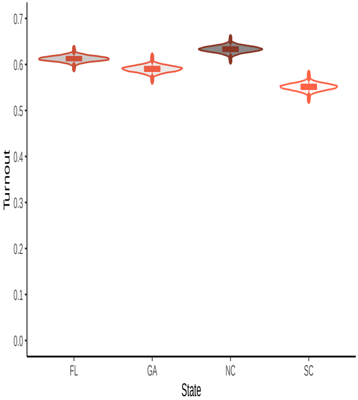


(a) Predicted turnout among item nonrespondents for MD-R model.

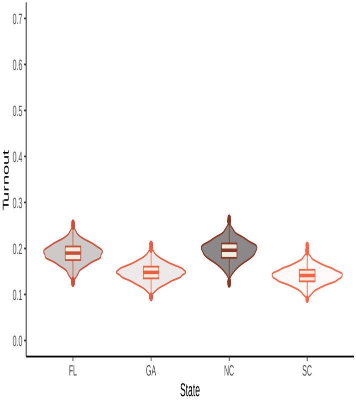


(b) Predicted turnout among unit nonrespondents for MD-R model.

Turnout Among Nonrespondents: MD-U



(c) Predicted turnout among item nonrespondents for MD-U model.



(d) Predicted turnout among unit nonrespondents for MD-U model.

Sub-population Estimates

Table 2. Population-level margin is 64.8% in NC.

	MD-R	MD-U	CC
Full	.65 (.64, .66)	.64 (.63, .65)	.77
M	.63 (.61, .64)	.61 (.59, .62)	.76
F	.67 (.66, .68)	.67 (.66, .69)	.79
<30	.50 (.45, .56)	.45 (.40, .51)	.64
30-49	.65 (.61, .68)	.62 (.58, .65)	.76
50-69	.72 (.68, .75)	.75 (.71, .79)	.82
70+	.76 (.71, .81)	.79 (.74, .85)	.84
<30(F)	.53 (.47, .58)	.48 (.43, .54)	.68
30-49(F)	.67 (.63, .70)	.64 (.61, .68)	.81
50-69(F)	.73 (.70, .77)	.77 (.73, .81)	.82
70+(F)	.78 (.73, .82)	.81 (.76, .86)	.82

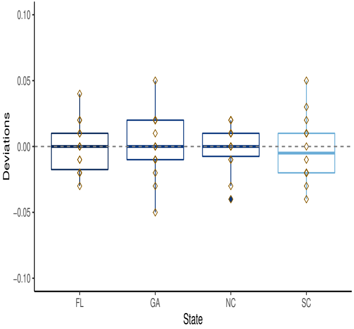
Model Checking

- Construct 95% posterior predictive intervals for all 64 four-way observable joint probabilities in the contingency table for $\{(S_i, G_i, A_i, V_i) : U_i = 0, R^y_{iG} = 0, R^x_{iA} = 0, R^x_{iV} = 0\}$.
- Percentage of intervals containing corresponding observed data point estimates:
    - MD-R: 94%
    - MD-U: 92%
- No really bad misses. Both models fit observed data well.

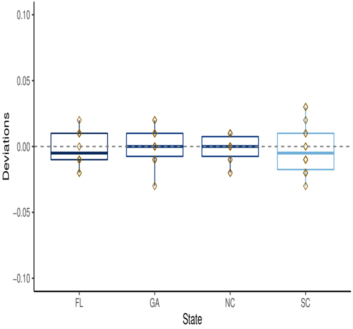
A (Modest) Evaluation

- Compare demographic characteristics of *voters* to those in voter file maintained by Catalist.

- Plot differences in point estimates across many sub-groups.



(e) Results for MD-R.



(f) Results for MD-U.

Extension: Measurement Error

- Other studies estimate over-reporting in turnout, with rates dependent on education level.
- Handle reporting error and missing data simultaneously via a hierarchical, measurement error model specification.
- $Z_i = 1$ for reported voter and $Z_i = 0$ for reported nonvoter.
- $C_i \in \{1, 2, 3\}$ is three-level educational attainment.
- $Pr(Z_i = 1 \mid V_i = 0, C_i = c) = \theta_c$, where $c = 1, 2, 3$.
  - Beta distribution prior on θ_c reflecting evidence in literature.
- Couple with models from slides 23 and 24.
- After allowing for measurement error, estimates are similar with a few increasing or decreasing by a point or two.

Extension: Survey-weighted Data

- General strategy: Require design-based estimates based on completed data to be plausible
  - Draw $\hat{t}_x^* \sim N(t_x, V)$, where V is analyst-determined
  - Impute missing x_i so that weighted estimate with completed data $\approx \hat{t}_x^*$
  - Use design weights for unit respondents and create equal “pseudo-weights” for unit nonrespondents
  - Repeat for all variables with margins, using imputation models with regression coefficients (other than intercepts) estimated from unit respondents’ data
  - Impute remaining variables (without margins) using models estimated with unit respondents’ data

Extension: Survey-weighted Data

- Implemented for stratified (Akande and Reiter, 2022) and Poisson sampling (Tang et al., 2024)
- Recent work: Replace item nonresponse modeling with MICE and "remaining variables" parts of unit nonresponse imputation by hot deck (Yang and Reiter, 2024)
- Ongoing work: Methods for use when design weights are available for unit respondents and nonrespondents

References I

Akande, O., G. Madsen, D. S. Hillygus, and J. P. Reiter (2021). "Leveraging auxiliary information on marginal distributions in nonignorable models for item and unit nonresponse." *Journal of the Royal Statistical Society, Series A* 184, pp. 643–662.

Akande, O. and J. P. Reiter (2022). "Multiple imputation for nonignorable nonresponse in complex surveys using auxiliary margins." *Statistics in the Public Interest-–In Memory of Stephen E. Fienberg*. Ed. by A. Carriquiry, W. Eddy, and J. Tanur. New York: Springer, pp. 289–306.

Hirano, K., G. Imbens, G. Ridder, and D. Rubin (2001). "Combining panel data sets with attrition and refreshment samples." *Econometrica* 69, pp. 1645–1659.

Sadinle, M. and J. P. Reiter (2017). "Itemwise conditionally independent nonresponse modelling for incomplete multivariate data." *Biometrika* 104 (1), pp. 207–220.

References II

Sadinle, M. and J. P. Reiter (2019). "Sequentially additive nonignorable missing data modeling using auxiliary marginal information." *Biometrika* 106, pp. 889 –911.

Schifeling, T. S. and J. P. Reiter (2016). "Incorporating marginal prior information into latent class models." *Bayesian Analysis* 11, pp. 499–518.

Tang, J., D. S. Hillygus, and J. P. Reiter (2024). "Using auxiliary marginal distributions in imputations for nonresponse while accounting for survey weights, with application to estimating voter turnout." *Journal of Survey Statistics and Methodology* 12, pp. 155–182.

Yang, Y. and J. P. Reiter (2024). *Imputation of nonignorable missing data in surveys using auxiliary margins via hot deck and sequential imputation*. arXiv: 2406.04599.