Alternative indicators for the risk of non-response bias

Federal Committee on Statistical Methodology
2018 Research and Policy Conference

Raphael Nishimura, Abt Associates
James Wagner and Michael Elliott, University of Michigan

Acknowledgements

This research was sponsored by The Eunice Kennedy Shriver National Institute of Child Health and Human Development

Grant number R03HD070012-02

Published paper:

Nishimura, R., Wagner, J., & Elliott, M. (2016). Alternative indicators for the risk of non-response bias: a simulation study. *International Statistical Review*, 84(1), 43-62.

Introduction

- Non-response: threat against quality of survey data
- Non-response bias = response rate x differences between respondents and non-respondents
- Declining response rates in surveys
- In the absence of other guidance → Response rates as indicator of the risk of non-response bias
- Poor indicator of non-response bias (Groves and Peytcheva, 2008)
- Response rate as a tool for monitoring data collection or post-survey adjustments: inefficient, biasing or both

Introduction (cont.)

- Alternative indicators proposed in the survey literature to evaluate the risk of non-response bias (e.g., Schouten et al., 2009; Wagner, 2010)
- Limited research regarding
  - The utility of these alternative measures
  - The conditions/missing mechanisms under which these indicators may prove to be helpful or misleading
- Goal: to assess the ability of various measures to indicate the risk of non-response bias in a variety of missing mechanisms
  - What are the properties of these indicators under different survey conditions?
  - Can a single or a set of these measures reliably indicate whether there is or not a risk of non-response bias?

Indicators for non-response bias

- Response rate
- Subgroups response rates
- Coefficient of variation of subgroups response rates
- Variance of non-response weights
- R-Indicator
- Area Under the Curve (AUC) of the logistic regression predicting response propensity
- Fraction of Missing Information (FMI)
- Correlation between non-response weights and survey variable

Methods: overview

- Two simulation studies using each k = 1,000 SRS's of size n = 1,000 to estimate the population mean of a survey variable Y with two explanatory variables (observed X and unobserved Z) varying:
  - Missing mechanism
  - Response rates
  - Correlation between explanatory and survey variables
  - Correlation between response propensities and explanatory variables
- Simulation and analysis performed in `R 2.13.2` (R Core Team, 2013) with `survey` (Lumley, 2004, 2012) and `mice` (van Buuren & Groothuis-Oudshoorn, 2011) and `rms` (Harell, 2014) packages

Methods: simulation studies

- Simulation study I:
  - k = 1,083 simulations
  - 3 missing mechanisms: MCAR, MAR, MNAR (Z only)
  - 19 response rates varying from 5% to 95% (increments by 5%)
  - 19 correlations between auxiliary variable (X or Z) and survey variable varying from 0.05 to 0.95 (increments by 0.05)

- Simulation study II:
  - k = 243 simulations
  - Missing mechanism: MNAR (Z and X)
  - 3 response rates: 20%, 40% and 70%
  - 3 correlations (X,Y): low, medium and high (0.05, 0.2, 0.7)
  - 3 correlations (Z,Y): low, medium and high
  - 3 correlations (X, ρ): low, medium and high
  - 3 correlations (Z, ρ): low, medium and high

Methods: data generation

- Variables (*Y*, *X*, *Z*) generated independently by

$$\begin{pmatrix} Y_i \\ X_i \\ Z_i \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 100 \\ 10 \\ 10 \end{pmatrix}, \begin{pmatrix} 25 & \sigma_{yx} & \sigma_{yz} \\ \sigma_{yx} & 4 & 0 \\ \sigma_{zy} & 0 & 4 \end{pmatrix} \right)$$

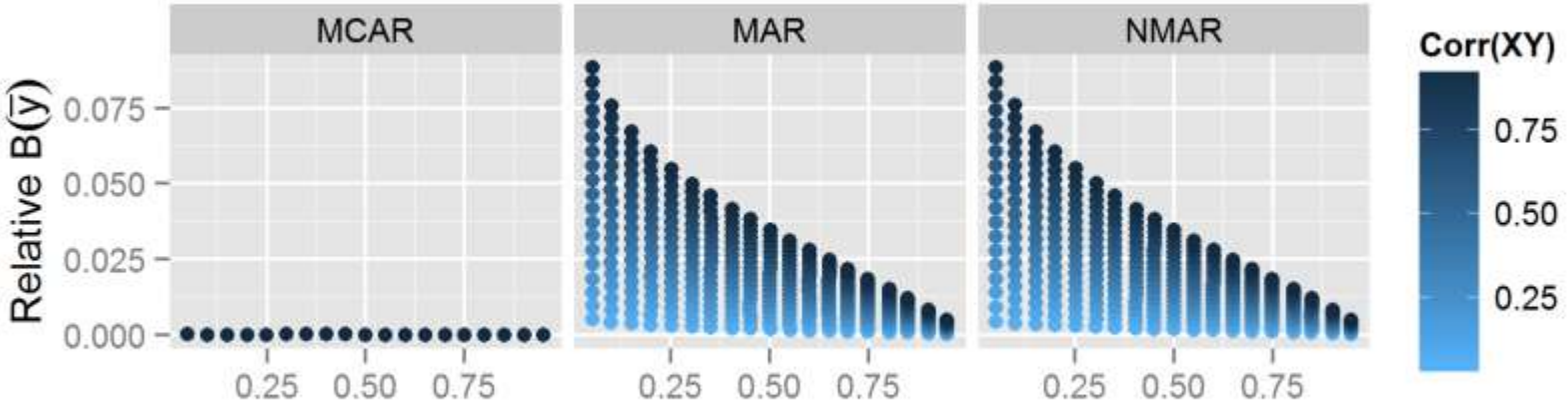
- Missing mechanism using response probabilities given by

$$\text{logit}\left(\rho_i\right) = \beta_0 + \beta_1 x_i + \beta_2 z_i$$

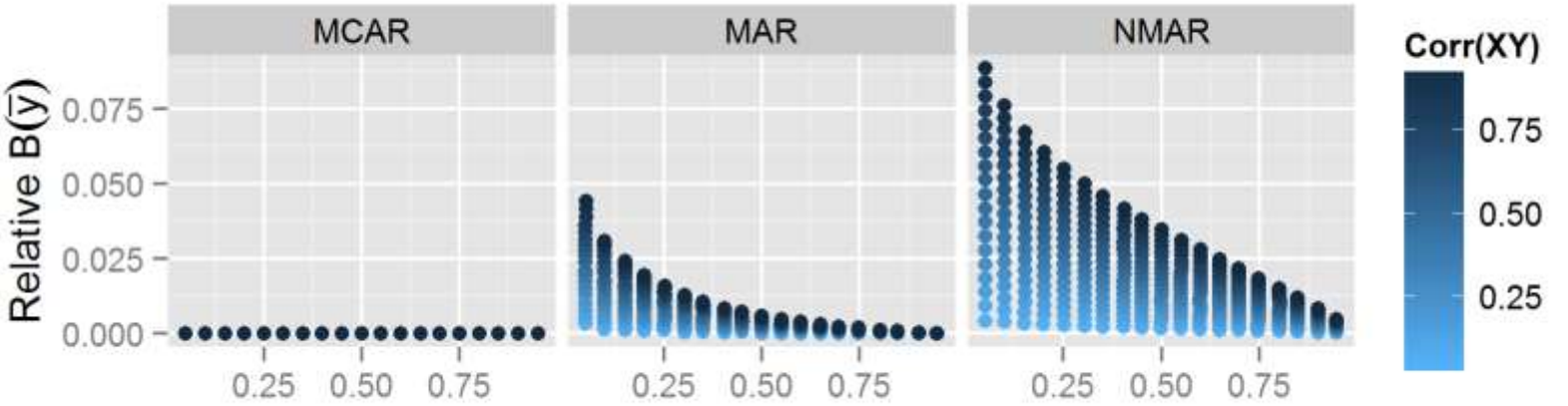
- Imputation model: $Y \sim X$
- Multivariate Imputation by Chained Equation ($M = 10$)

Results: Study I, Non-response bias by RR

Respondent Mean

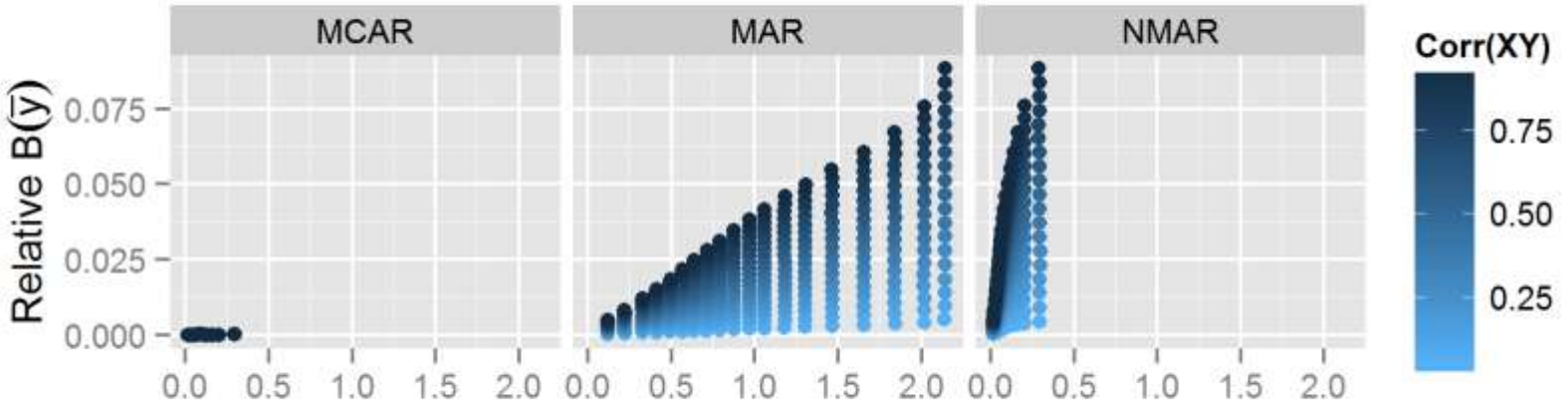


Weighted Mean

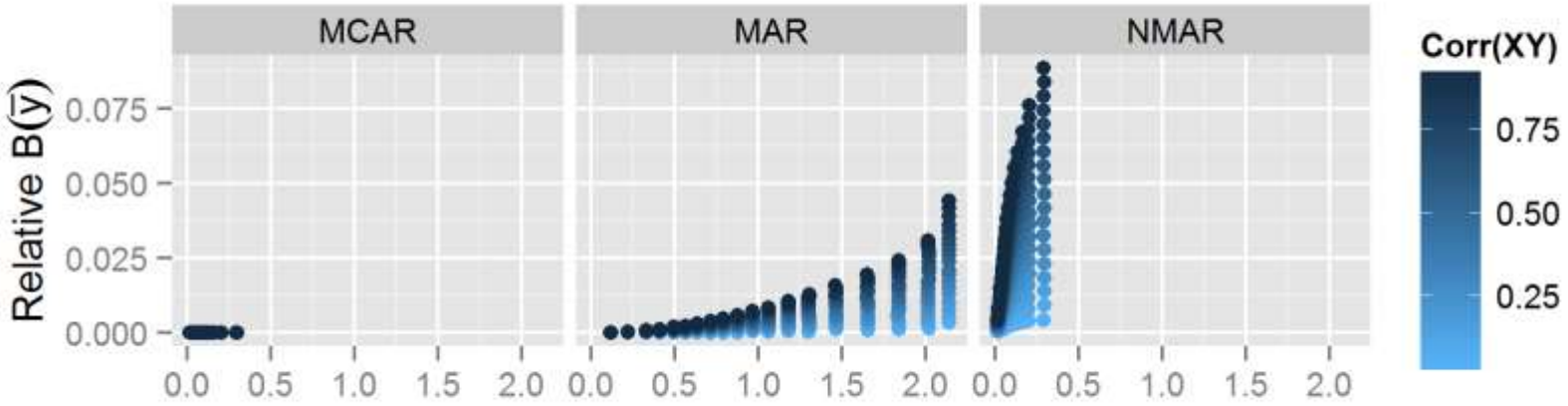


Results: Study I, Non-response bias by CV(RR_{sub})

Respondent Mean

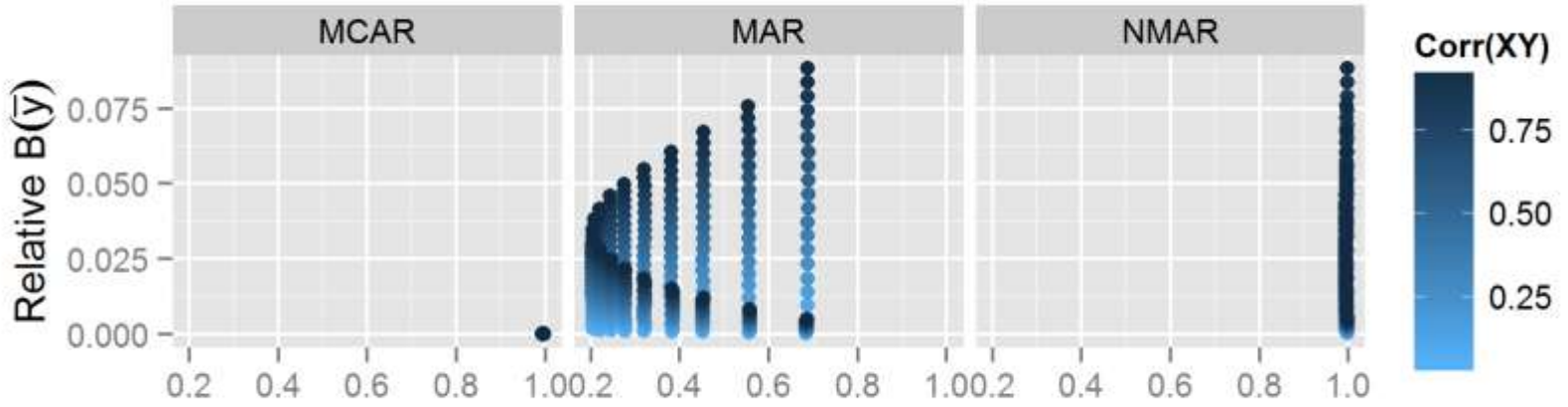


Weighted Mean

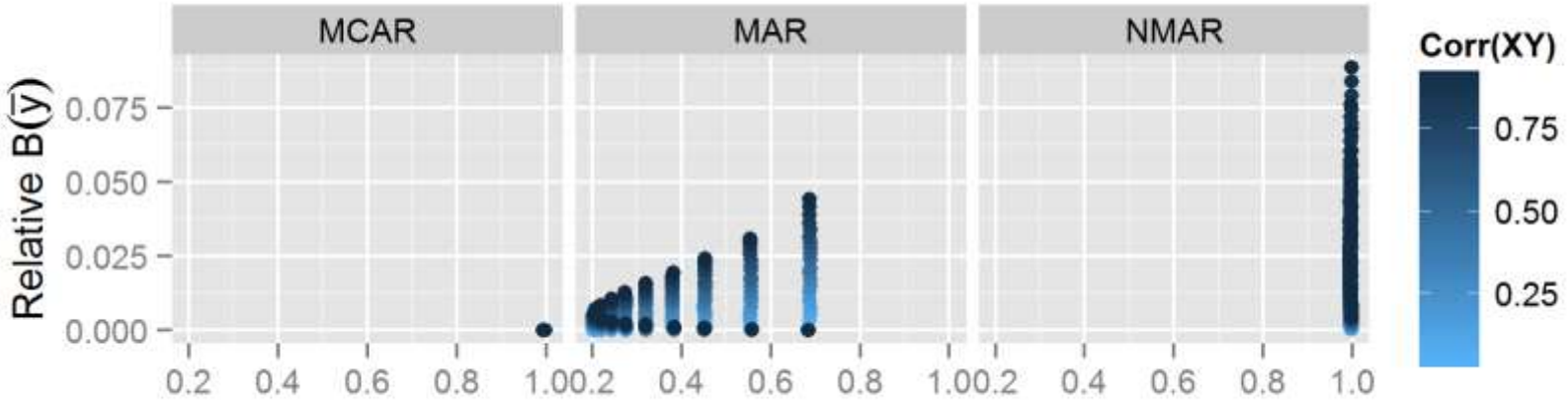


Results: Study I, Non-response bias by R-Indicator

Respondent Mean

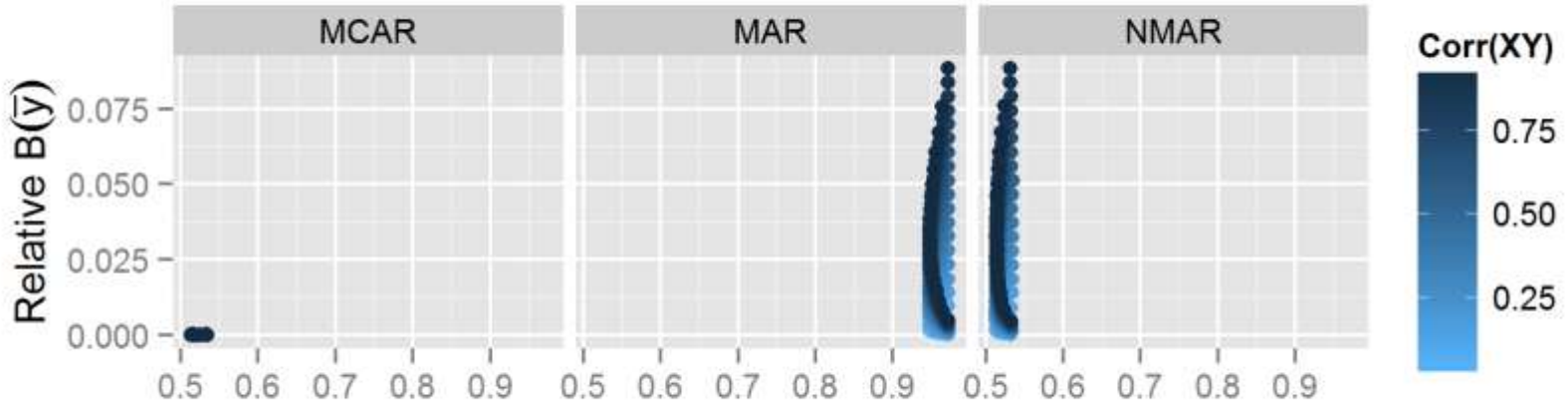


Weighted Mean

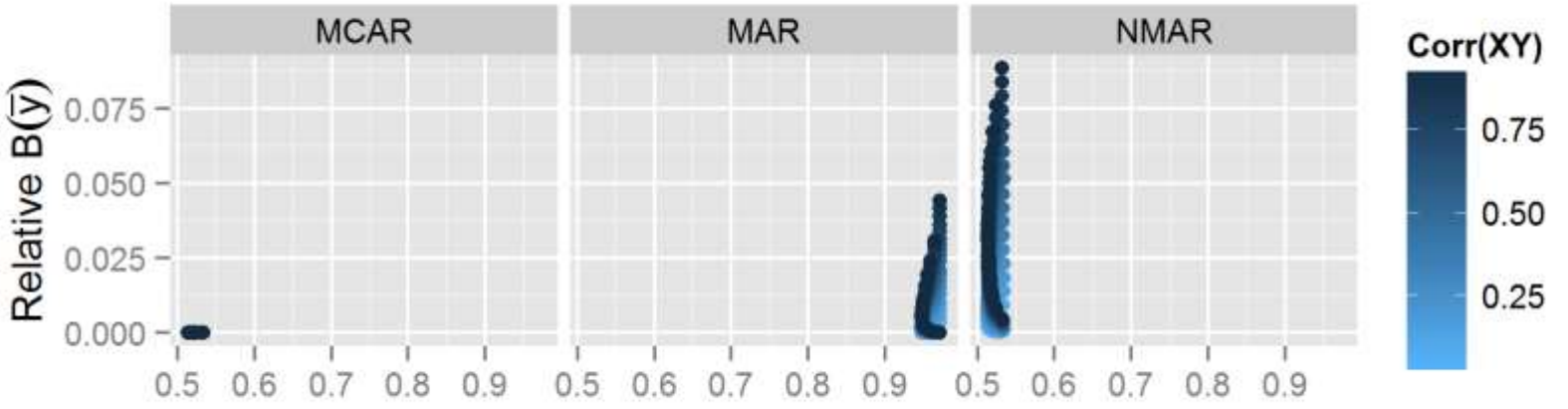


Results: Study I, Non-response bias by AUC

Respondent Mean

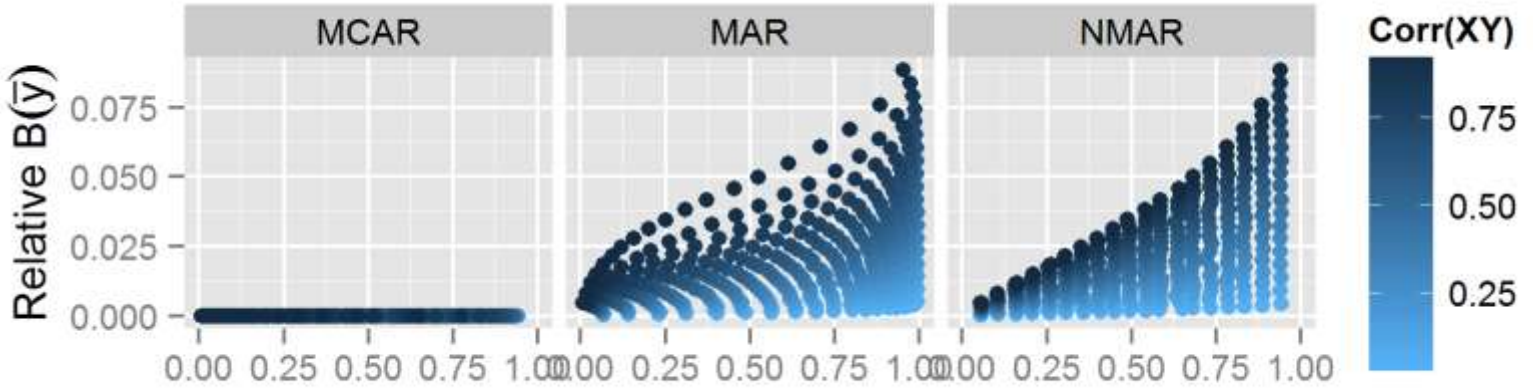


Weighted Mean

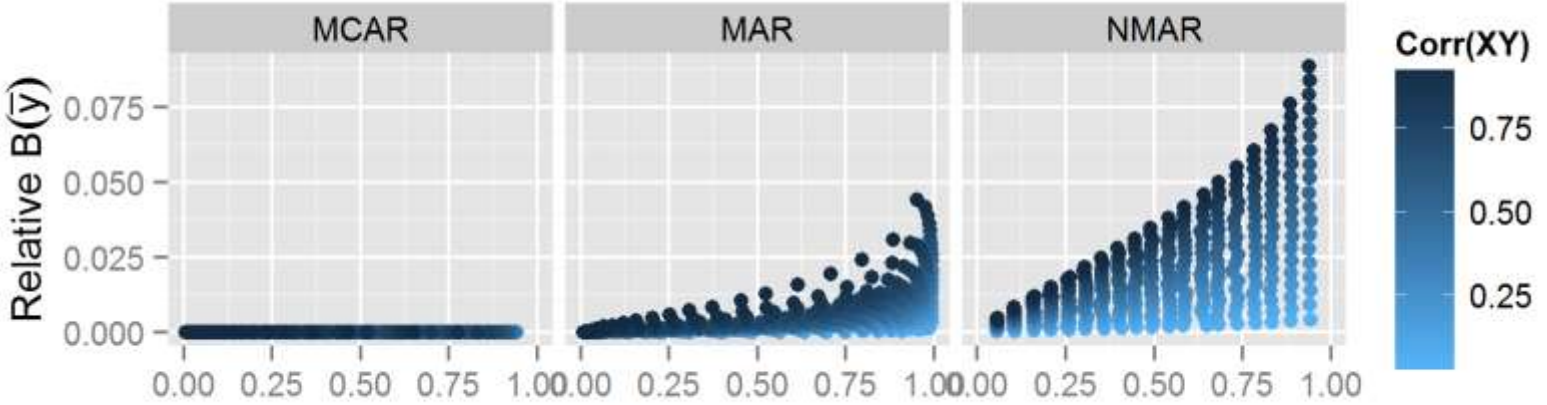


Results: Study I, Non-response bias by FMI

Respondent Mean

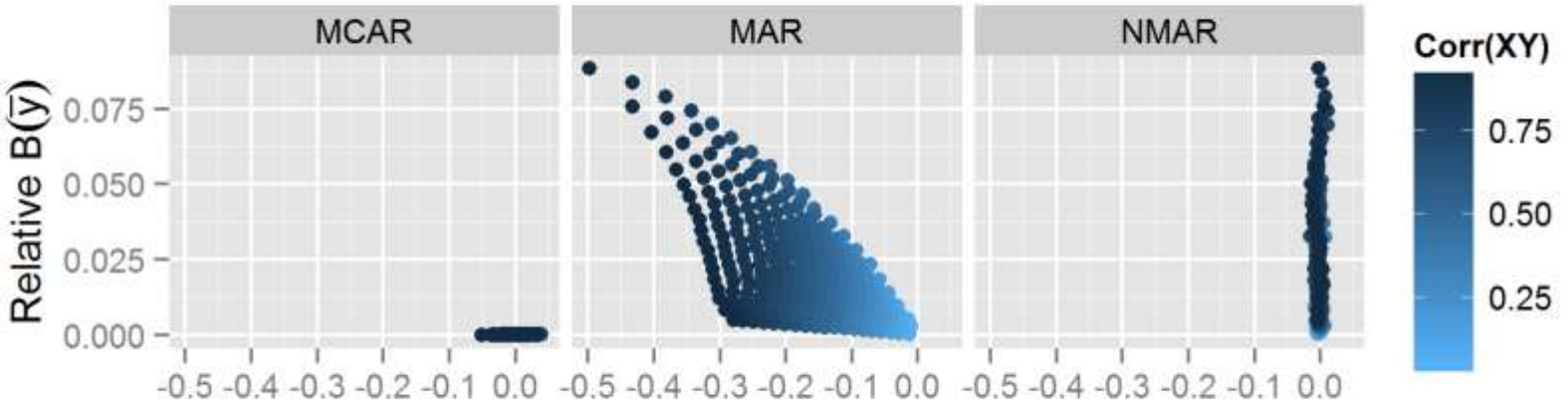


Weighted Mean

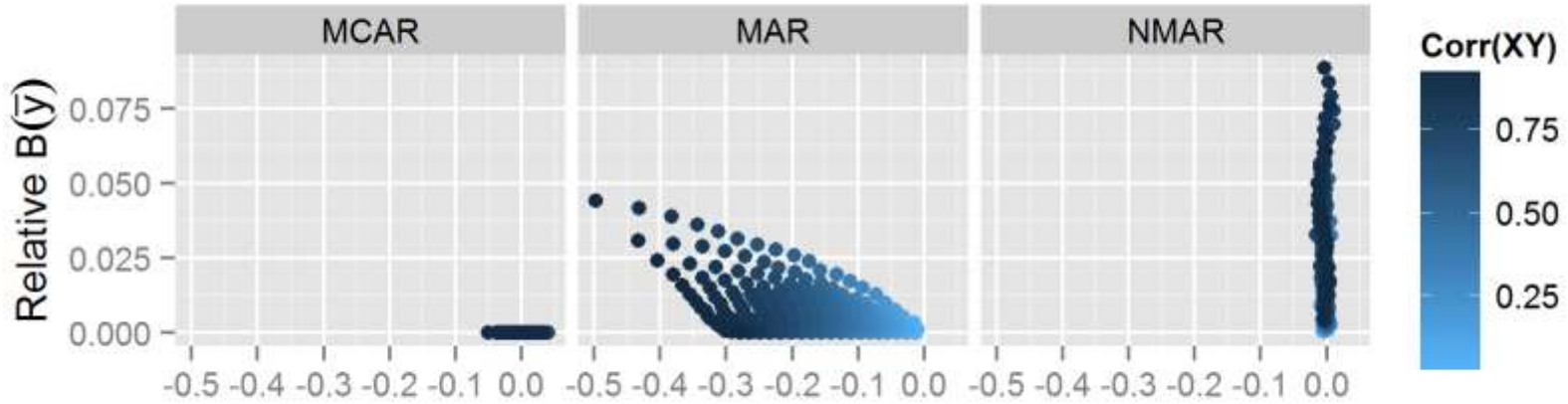


Results: Study I, Non-response bias by Corr(W_{nr},Y)

Respondent Mean



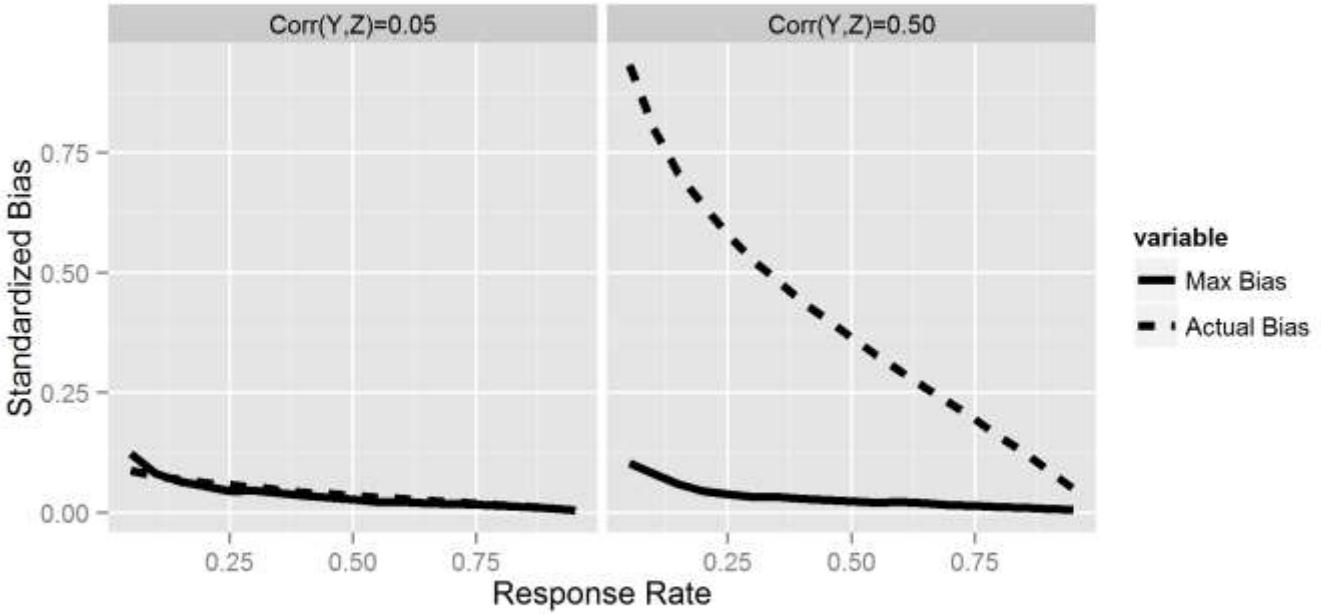
Weighted Mean



Results: Study I, Maximal absolute bias

Minimize the “maximal absolute bias”
(Schouten, et al., 2009; Buellens and Loosveldt, 2012):

$$B_m(\rho) = \frac{\left[1 - R(\rho)\right] S(y)}{2\bar{\rho}}$$



Results: Study II, Bias of the FMI under MNAR

		Corr(Y,X)								
		Low			Medium			High		
		Corr(Y,Z)								
Corr(R,X)	*Corr(R,Z)*	Low	Medium	High	Low	Medium	High	Low	Medium	High
Low	Low	-0.99%	0.24%	35.79%	0.55%	1.16%	38.06%	-0.10%	6.33%	1250.96%
	Medium	-3.74%	-2.51%	29.58%	-3.08%	-0.69%	34.36%	-6.78%	-1.04%	1111.53%
	High	-29.20%	-29.41%	-18.83%	-30.42%	-30.04%	-20.13%	-40.53%	-40.64%	200.48%
Medium	Low	0.57%	1.40%	32.65%	0.84%	-0.13%	40.57%	-0.36%	5.09%	1208.11%
	Medium	-4.22%	-2.34%	27.13%	-3.40%	-1.56%	32.43%	-5.11%	1.18%	1105.40%
	High	-27.63%	-27.96%	-19.54%	-29.26%	-27.81%	-17.23%	-39.46%	-37.86%	212.31%
High	Low	-29.20%	-29.41%	-18.83%	-30.42%	-30.04%	-20.13%	-40.53%	-40.64%	200.48%
	Medium	-1.90%	-1.84%	10.21%	-1.53%	-0.83%	11.77%	-2.55%	2.45%	522.15%
	High	-15.34%	-15.08%	***-7.55%***	-14.94%	-14.78%	***-5.36%***	-21.53%	-20.47%	353.43%

Conclusions

- Most of the indicators, as expected, are survey variable/statistic-independent
- FMI and corr(W_{NR}, Y) are the only indicators that are sensitive to corr(Y, X)
- In general, we observe that none of the indicators or a set of them can clearly pick up situations where there is a risk of non-response bias either because:
  - There is no association with the indicators and the non-response bias or
  - We cannot distinguish the missing mechanisms (especially between MCAR and MNAR)

Conclusions

- Indicators such as the *maximum bias* are sensitive to model assumptions and should be used with care
- Other indicators, such as the *FMI*, might be biased, but somehow useful to detect the possibility of non-response bias
- The general pattern of the indicators don’t change whether it is about the non-response bias in the respondent unweighted mean or the non-response weighted mean

References

- Beullens, K. and Loosveldt, G. (2012) Should high response rates really be a primary objective? *Survey Practice*, 5(3).
- Groves, R. M. and Peytcheva, E. (2008) The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, 72(2): 167-189.
- Harrell Jr., F. E. (2014). rms: Regression Modeling Strategies. R package version 4.2-0.http://CRAN.R-project.org/package=rms
- Lumley T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*. 9 (1): 1-19
- Lumley, T. (2012). survey: analysis of complex survey samples. R package version 3.28-2.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
- Schouten, B., F. Cobben and J. G. Bethlehem (2009) Indicators for the representativeness of sur-vey response. *Survey Methodology*, 35(1): 101-113.
- Wagner, J. (2010) The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data. *Public Opinion Quarterly*, 74(2): 223-243.

Thank you

- Raphael Nishimura: raphael_nishimura@abtassoc.com
- James Wagner: jameswag@umich.edu
- Michael Elliott: mrelliot@umich.edu