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:
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 \texttt{R 2.13.2} (R Core Team, 2013) with \texttt{survey} (Lumley, 2004, 2012) and \texttt{mice} (van Buuren & Groothuis-Oudshoorn, 2011) and \texttt{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, \( \rho \)): low, medium and high
  – 3 correlations (Z, \( \rho \)): low, medium and high
Methods: data generation

• Variables \((Y, X, Z)\) generated independently by

\[
\begin{align*}
Y_i & \sim N_3(100, \begin{pmatrix} 25 & \sigma_{yx} & \sigma_{yz} \\ \sigma_{yx} & 4 & 0 \\ \sigma_{yz} & 0 & 4 \end{pmatrix}) \\
X_i & \sim N_3(10, \begin{pmatrix} 10 & 0 & 0 \\ 0 & \sigma_{yx} & 0 \\ 0 & 0 & \sigma_{zy} \end{pmatrix}) \\
Z_i & \sim N_3(10, \begin{pmatrix} 10 & 0 & 0 \\ 0 & \sigma_{yx} & 0 \\ 0 & 0 & \sigma_{zy} \end{pmatrix})
\end{align*}
\]

• Missing mechanism using response probabilities given by

\[
\text{logit}(\rho_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i
\]

• 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(\(RR_{\text{sub}}\))
Results: Study I, Non-response bias by R-Indicator
Results: Study I, Non-response bias by AUC

![Graph showing Non-response bias by AUC for different scenarios: MCAR, MAR, NMAR. The graphs display the relationship between Respondent Mean and Weighted Mean, with Correlation (XY) values ranging from 0.25 to 0.75.](image)
Results: Study I, Non-response bias by FMI
Results: Study I, Non-response bias by Corr($W_{nr}, Y$)
Results: Study I, Maximal absolute bias

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

\[ B_m(\rho) = \frac{[1 - R(\rho)]S(y)}{2\bar{\rho}} \]
### Results: Study II, Bias of the FMI under MNAR

<table>
<thead>
<tr>
<th></th>
<th>Corr(Y,X)</th>
<th></th>
<th>Corr(Y,Z)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Corr(R,X)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>-0.99%</td>
<td>0.24%</td>
<td>35.79%</td>
<td>0.55%</td>
<td>1.16%</td>
</tr>
<tr>
<td>Medium</td>
<td>-3.74%</td>
<td>-2.51%</td>
<td>29.58%</td>
<td>-3.08%</td>
<td>-0.69%</td>
</tr>
<tr>
<td>High</td>
<td>-29.20%</td>
<td>-29.41%</td>
<td>-18.83%</td>
<td>-30.42%</td>
<td>-30.04%</td>
</tr>
<tr>
<td>Medium</td>
<td>0.57%</td>
<td>1.40%</td>
<td>32.65%</td>
<td>0.84%</td>
<td>-0.13%</td>
</tr>
<tr>
<td>Medium</td>
<td>-4.22%</td>
<td>-2.34%</td>
<td>27.13%</td>
<td>-3.40%</td>
<td>-1.56%</td>
</tr>
<tr>
<td>High</td>
<td>-27.63%</td>
<td>-27.96%</td>
<td>-19.54%</td>
<td>-29.26%</td>
<td>-27.81%</td>
</tr>
<tr>
<td>High</td>
<td>-29.20%</td>
<td>-29.41%</td>
<td>-18.83%</td>
<td>-30.42%</td>
<td>-30.04%</td>
</tr>
<tr>
<td>Medium</td>
<td>-1.90%</td>
<td>-1.84%</td>
<td>10.21%</td>
<td>-1.53%</td>
<td>-0.83%</td>
</tr>
<tr>
<td>High</td>
<td>-15.34%</td>
<td>-15.08%</td>
<td>7.55%</td>
<td>-14.94%</td>
<td>-14.78%</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>-0.10%</td>
<td>6.33%</td>
<td>1250.96%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-6.78%</td>
<td>-1.04%</td>
<td>1111.53%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-40.53%</td>
<td>-40.64%</td>
<td>200.48%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

• Most of the indicators, as expected, are survey variable/statistic-independent
• FMI and $\text{corr}(W_{\text{NR}}, Y)$ are the only indicators that are sensitive to $\text{corr}(Y, X)$
• 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

Thank you

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