MULTIVARIATE BEHAVIORAL RESEARCH
2019, VOL. 54, NO. 1, 146
https://doi.org/10.1080/00273171.2018.1555747

Routledge
Taylor & Francis Group

Check for updates

Abstract: An Evaluation of Planned Missing Data Designs in Large Surveys

Dan Su

University of Wisconsin-Madison

Planned missing data designs in large surveys can efficiently reduce respondents' burden and lower the cost associated with data collection, without cutting down on the questionnaire items. If the missing data are not appropriately planned, descriptive and potential causal parameter estimates will be biased. This paper implemented two studies which use simulated data based on the PISA 2006 data. The two simulation studies investigate how the bias in marginal means, correlations and regression coefficients depend on the chosen planned missing data designs and related characteristics. The first study shows that, for a fixed sample size, the extent of bias depends on the three major properties of design and data: overlap percentages (i.e., the portion of cases where two variables are jointly observed), missing percentages, and distributions of variables. The second study applies the properties of design to illustrate that an optimal incomplete block design that ensures overlap can be a better choice than a multiple-form design given the same amount of missingness. In large surveys, the available planned missing data designs have not yet been systematically investigated with respect to the missing percentage, overlap percentage, distribution of variables, and sample sizes. The findings will guide researchers in choosing a design with optimally arranged items in forms. With appropriately chosen planned missing data designs and multiple imputations, the number of survey questions can be increased without harming the validity of parameter estimates.

The results of the first simulation show that the estimates of means are unbiased for large survey data (i.e., sample sizes exceeding 1000 cases) even when overlap is zero and the missing percentage is high (e.g., 80%). However, the recovery of correlations and regression coefficients requires sufficient overlap. The bias in correlations is negligibly small when there is 20% or more overlap for continuous data. Regarding regression coefficients, the bias is negligibly small when overlap exceeds 20% for multivariate normal data. For skewed and categorical data, the coefficients are estimated with larger bias and less reliability. The second simulation study compares a two-form design, a three-form design, and two optimal block designs with 50 and 33% missingness. The results show that all designs recover the means of variables without bias. The biases in correlations are negligibly small for all designs except for the two-form design which has no overlap across forms. With the same amount of missingness, the optimal block design with 50% missingness largely reduces the bias in regression coefficients due to its 20% overlap compared to the two-form design. Furthermore, the three-form design and the optimal block design with 30% missingness produce negligibly small bias and more reliable estimates compared to the other two designs.

Article information

Conflict of interest disclosures: The author signed a form for disclosure of potential conflicts of interest. The author did not report any financial or other conflicts of interest in relation to the work described.

Ethical principles: The author affirms having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant #DRL-1749275 from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The author would like to thank David Kaplan for comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the author alone, and endorsement by the author's institution or the American Educational Research Association is not intended and should not be inferred.

CONTACT Dan Su ✉ dsu4@wisc.edu Graduate School, University of Wisconsin-Madison, Madison, WI 53706.

© 2019 Taylor & Francis Group, LLC

gwimberly@aera.net - June 24, 2024 - Read articles at www.DeepDyve.com