Social Science & Medicine xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

The treatment of incomplete data: Reporting, analysis, reproducibility, and replicability

Yulia Sidi, Ofer Harel*

Department of Statistics, University of Connecticut, 215 Glenbrook Road, Unit 4120, Storrs, CT 06269-4120, United States

ARTICLE INFO

Keywords:
Missing data
Incomplete data
Multiple imputation
Reproducibility
Replicability

ABSTRACT

Proper analysis and reporting of incomplete data continues to be a challenging task for practitioners from various research areas. Recently Nguyen, Strazdins, Nicholson and Cooklin (NSNC; 2018) evaluated the impact of complete case analysis and multiple imputation in studies of parental employment and health. Their work joins interdisciplinary efforts to educate and motivate scientists across the research community to use principled statistical methods when analyzing incomplete data. Although we fully support and encourage work in parallel to NSNC's, we also think that further actions should be taken by the research community to improve current practices. In this commentary, we discuss some aspects and misconceptions related to analysis of incomplete data, in particular multiple imputation. In our view, the missing data problem is part of a larger problem of research reproducibility and replicability today. Thus, we believe that improving analysis and reporting of incomplete data will make reproducibility and replicability efforts easier. We also provide a brief checklist of recommendations which could be used by members of the scientific community, including practitioners, journal editors, and reviewers to set higher publication standards.

1. Introduction

A recent study by Nguyen, Strazdins, Nicholson and Cooklin (Nguyen et al., 2018; hereafter referred as NSNC) discusses the importance of properly handling missing data in studies of parental employment and health. The authors provide an excellent overview of the impact of missing data in this research area. Moreover, they compare, using a case study, two commonly used missing data techniques: complete case analysis (CCA) and multiple imputation (MI). Given the fact that CCA is the most common approach in such studies, their example signifies the implications associated with poorly handling incomplete data.

Although the missing data problem is not new, it continues to be overlooked in many research settings (Bell et al., 2014; Eekhout et al., 2012; Harel et al., 2012; Harel and Boyko, 2013; Karahalios et al., 2012; Little et al., 2012; Perkins et al., 2018; Peugh and Enders, 2004; Powney et al., 2014; Sullivan et al., 2017; Wood et al., 2004). For example, Harel et al. (2012) demonstrated that out of 57 HIV-prevention randomized trials with biological outcomes published between 2005 and 2010 in refereed journals, none mentioned missing data assumptions in their analyses, 74% performed a CCA; most seriously, only 12% are expected to report unbiased results. Eekhout et al. (2012) performed a systematic review of 262 studies published in 2010 in the three leading epidemiology journals that used questionnaires: 85% of these articles had no mention of the missingness assumption at all, and 81% used CCA. Masconi et al. (2015) considered 48 prevalent diabetes risk studies published between 1997 and 2014, where they found that 62.5% of the reported articles offered no information in regard to missing data. Nicholson et al. (2017) evaluated 541 papers related to attrition in developmental psychology, published between 2009 and 2012, to assess whether the *Publication Manual* of the American Psychological Association (2010; APA), which recommended reporting, assessment, and appropriate handling of missing data, had any effect in practice. They found that the Manual did not alter improvement in this area; only 18.3% of the articles they reviewed discussed missing data mechanisms. The common goal of the studies above, as in NSNC's, was to underline the importance of appropriate handling and reporting of incomplete data. Moreover, the variety of mentioned research fields shows that overlooking the missing data problem is not specific for a particular research area. We believe this also has a great impact on reproducibility and replicability in research.

Reproducibility, which is the ability to compute the same result, could be compromised when a published manuscript doesn't share analyzed data or programming code. In contrast, *replicability* refers to a chance of obtaining consistent results by independent studies having similar design and research questions. In general, the latter is a more

* Corresponding author.
E-mail address: ofer.harel@uconn.edu (O. Harel).

https://doi.org/10.1016/j.socscimed.2018.05.037
Received 8 March 2018; Received in revised form 11 May 2018; Accepted 19 May 2018
0277-9536/ © 2018 Elsevier Ltd. All rights reserved.

severe problem, as it puts the research community's credibility in question (Leek and Peng, 2015). The chances of achieving high replicability can be lowered by the lack of scientific reasoning for the analysis assumptions used in a previous research.

Although there are ongoing efforts in the statistical community to increase reproducibility and replicability (Leek and Peng, 2015), it needs to be extended to all areas of science. We believe that this goal could be achieved if scientists, journal editors, and reviewers set higher publication standards. Improving current practices along with more rigorous publication requirements would make reproducibility straight forward and would more importantly increase chances of replicability.

We consider the recent paper by NSNC a great addition to the interdisciplinary effort to emphasize the importance of considering the complications that arise from incomplete data in social sciences and medical research. In this commentary, we expand on some aspects of reporting and handling of incomplete data, and share our thoughts about the currently available practices mentioned by NSNC, specifically when MI is used in the statistical analysis.

2. Important aspects regarding the reporting and handling of incomplete data

2.1. Missingness mechanism - stating the assumptions

Before diving into methodological details, let us first present an artificial example that will be referred to throughout this commentary. Suppose we are interested in determining whether mental health status is associated with physical activity (high/low) and the mental health status is assessed by a questionnaire with possible outcome scores of 0–20, with lower scores representing a better mental health. We further assume that the data are collected in one wave and while the physical activity is recorded for all the subjects in the study, mental health scores are missing for 25% of the participants. In order to analyze such a dataset with incomplete data, a researcher needs to determine plausible statistical assumptions before performing any statistical analysis. These assumptions, which are embedded in any statistical method, must be explicitly noted and justified, in particular when related to missing data.

Following a general notation, let Y_{com} denote a complete dataset we aim to collect, that is, physical activity and mental health status scores. Y_{com} could be conceptually partitioned into observed Y_{obs} and missing Y_{mis} parts, while in practice we see only values for Y_{obs}. Also, let's define θ as parameter of interest (e.g., mean, regression coefficient, or odds ratio) for which we use Y_{obs} to estimate it. Further suppose that R is a matrix of the same dimension as Y_{com}, which consists of 1s where the data values are missing, and 0 otherwise. In our dataset, we would have a missing values indicator for *Mental Status* score as it is the only variable with missing data. In context of the above example, the data structure appears in Table 1.

In general, we would like to infer about θ from $P(Y_{com}|\theta)$; however, because the database is incomplete, a joint model of the data (Y_{com}) and missing data mechanism (R) needs to be considered. Consider the following joint model:

$$P(Y_{com}, R; \theta, \varphi) = P(Y_{com}|\theta)P(R|Y_{com}, \varphi)P(\theta, \varphi), \quad (1)$$

where φ is a nuisance parameter which characterizes the distribution of R. Due to the missing values, we are unable to summarize information in $P(Y_{com}; \theta,)$ (e.g., regression or ANOVA of the complete data) and need to evaluate θ from Equation (1) instead. Of course, it is a much more complicated situation.

The missing data mechanisms could be specified as: missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR) (Rubin, 1976; Little and Rubin. 2014). MCAR indicates that missing data mechanism, R, neither depends on the data Y_{com} we tried to collect, nor on any other information outside the study. In the context of our example, if some records of the mental health scores were deleted by mistake due to a technical problem it will imply that MCAR could be considered as the missing data mechanism. MAR indicates that R depends only on the observed information (Y_{obs}). In our example, this assumption suggests that only exercise status is responsible for the missing information in the mental health scores. Finally, MNAR indicates that R may depend on information that is not available to us, which either is missing due to incomplete variable(s) we are collecting or is missing due to other factors outside the study. MNAR in our example could imply that people with worse mental health status refused to answer this questionnaire. As can be seen, different reasons imply different missing data mechanisms, which consequently lead to different assumptions being made in the statistical analysis.

Certainly in practice it is hard to know the underline reasons (mechanism) that cause missing data, and while some of the missing data mechanism assumptions are testable (MCAR) (Little, 1988), others are not (MAR, MNAR) (Molenberghs et al., 2008). In particular, it is impossible to distinguish between MAR and MNAR structure, with the observed data alone. Yet, this problem makes the assumptions choice argument even more important. Thus, practitioners are encouraged to clearly specify the assumptions they use in the analysis, as well as to justify them in the context of the specific problem they study.

Table 1
Dataset structure for mental health vs. physical activity study.

Y_{obs}- Observed values in the data		R-missing values indicator	
Physical activity	Mental status score	Physical activity	Mental status score
Yes	7	0	0
Yes	?	0	1
No	15	0	0
…	…	…	…
No	13	0	0
No	?	0	1

Note. The character "?" implies a missing datum.

2.2. Ignorability — commonly confused with missing at random

Many researchers confuse ignorability with MAR, mostly because MAR is a needed component (and mostly argued) for ignorability. Yet, the two concepts, while related, differ and should be understood by those dealing with incomplete data. While missing data assumptions are necessary for proper analysis of incomplete data, these are not sufficient for ignorability. Ignorability plays a central role in the analysis of incomplete data, and is usually incorporated as default option for multiple imputation procedures in many statistical software programs (e.g., PROC MI in SAS (SAS Institute Inc, 2011), norm package in R (Novo and Schafer, 2013), and the suite of MI commands in STATA (StataCorp, 2013). As Little and Rubin (2014) described, *ignorability* consists of two assumptions: (a) MAR and (b) distinctness (or *a priori* independence in Bayesian framework) in parameters of the data model (θ) and missing data mechanism (φ). *Distinctness* can be thought of as meaning that a change in one parameter will not influence the other. For example, the mean difference in the mental scores between people with high and low physical activity (θ) is not related to the proportion of study participants who couldn't complete the mental health questionnaire due to time constraints (φ).

Consequently, non-ignorability could be attributed to either MNAR or non-distinctness (or both). Although, non-distinctness is more likely to appear in longitudinal settings, where observations are collected for the same individuals repeatedly over time (or clustered data in general), it can still lead to inefficiency in other types of studies (Little and Rubin, 2014). As was recently evaluated through a thorough simulation study conducted by Yucel (2017), non-distinctness can cause serious reductions in the coverage rates when evaluated in relation to MI. Thus,

Download English Version:

https://daneshyari.com/en/article/7327387

Download Persian Version:

https://daneshyari.com/article/7327387

Daneshyari.com