Behaviour Research and Therapy 98 (2017) 4–18

Contents lists available at ScienceDirect

Behaviour Research and Therapy

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

Multiple imputation as a flexible tool for missing data handling in clinical research

Craig K. Enders

UCLA, Department of Psychology, 3586 Franz Hall, 502 Portola Plaza, Los Angeles, CA 90095, United States

ARTICLE INFO

Article history:
Received 9 September 2016
Received in revised form
16 November 2016
Accepted 17 November 2016
Available online 18 November 2016

Keywords:
Missing data
Multiple imputation
Attrition
Maximum likelihood estimation

ABSTRACT

The last 20 years has seen an uptick in research on missing data problems, and most software applications now implement one or more sophisticated missing data handling routines (e.g., multiple imputation or maximum likelihood estimation). Despite their superior statistical properties (e.g., less stringent assumptions, greater accuracy and power), the adoption of these modern analytic approaches is not uniform in psychology and related disciplines. Thus, the primary goal of this manuscript is to describe and illustrate the application of multiple imputation. Although maximum likelihood estimation is perhaps the easiest method to use in practice, psychological data sets often feature complexities that are currently difficult to handle appropriately in the likelihood framework (e.g., mixtures of categorical and continuous variables), but relatively simple to treat with imputation. The paper describes a number of practical issues that clinical researchers are likely to encounter when applying multiple imputation, including mixtures of categorical and continuous variables, item-level missing data in questionnaires, significance testing, interaction effects, and multilevel missing data. Analysis examples illustrate imputation with software packages that are freely available on the internet.

The methodological literature on missing data handling spans many decades, but the modern era of this work arguably began when Rubin (1976) established a theoretical framework for missing data problems. Since then, there has been a substantial increase in missing data research, and most software applications now implement one or more sophisticated missing data handling routines. Despite the uptick in methodological research and the concurrent publication of several missing data texts (Allison, 2002; Carpenter & Kenward, 2013; Enders, 2010; Graham, 2012; Little & Rubin, 2002; van Buuren, 2012), the migration to better analytic practices has understandably been slow. Going back to the 2000s, literature reviews revealed that researchers relied primarily on deletion methods that remove cases with missing data (Jelicic, Phelps, & Lerner, 2009; Peugh & Enders, 2004; Wood, White, & Thompson, 2004), despite warnings that these "are among the worst methods available for practical applications" (Wilkinson & Taskforce on Statistical Significance, 1999, p. 598). Although reporting practices have definitely improved in recent years, the application of modern missing data handling techniques is far from uniform in psychology and related disciplines. Consequently, the primary goal of this manuscript is to promote the awareness and application of analytic methods that enjoy strong support in the methodological literature.

Broadly speaking, the recent missing data literature supports the use of maximum likelihood estimation and multiple imputation (Schafer & Graham, 2002).[1] Maximum likelihood estimation (also known as full information maximum likelihood, or FIML) employs an iterative optimization algorithm that identifies parameter estimates that maximize fit to the observed data. For example, in a regression analysis, the maximum likelihood estimates are coefficients that minimize the sum of the squared standardized distances between the observed data and the regression line. Some methodologists have characterized maximum likelihood estimation as "implicit imputation" because it does not produce a filled-in data set (Widaman, 2006). Rather, the procedure uses all of the available data to estimate a specific set of model parameters and

[1] Bayesian estimation is a third option that I do not consider here in the interest of space. A Bayesian analysis mimics maximum likelihood estimation in the sense that it generates estimates and standard errors for a specific analysis model. However, the missing data handling aspect of Bayesian estimation resembles multiple imputation because each cycle of the iterative algorithm generates a filled-in data set.

E-mail address: cenders@psych.ucla.edu.

http://dx.doi.org/10.1016/j.brat.2016.11.008
0005-7967/

their standard errors. For example, to apply maximum likelihood to an ANOVA-type analysis, a researcher need only use a capable software package to estimate a regression model from the incomplete data. Structural equation modeling software packages are particularly useful for implementing maximum likelihood because they can accommodate a range of missing data patterns (e.g., missing values on explanatory and outcome variables).

In contrast, multiple imputation creates several versions of a data set, each of which contains different estimates of the missing values. As explained later, most incarnations of multiple imputation use a regression model to fill in the data, treating incomplete variables as outcomes and complete variables as predictors. To avoid imputations based on a single set of regression parameters, an iterative algorithm uses Bayesian estimation to update the regression model parameters, and it uses new estimates to generate each set of imputations. Having generated a set of filled-in data sets, the researcher then performs one or more statistical analyses on each complete data set to obtain imputation-specific estimates and standard errors. The final step pools the estimates and standard errors into a single set of results.

With normally distributed data, a common set of input variables, and a sufficiently large sample size, there is no theoretical reason to expect differences between maximum likelihood estimation and multiple imputation (Gelman et al., 2014; Meng, 1994; Schafer, 2003), and empirical studies suggest that the two methods usually yield similar estimates and standard errors (Collins, Schafer, & Kam, 2001).[2] All things being equal, maximum likelihood estimation is probably preferable for many situations on the basis of simplicity alone – as noted previously, a researcher need only translate the desired analysis to a capable software package. However, psychological data sets often feature complexities that are currently difficult to handle appropriately in the likelihood framework. A regression analysis with mixtures of categorical and continuous variables is a very simple, yet common, scenario where maximum likelihood estimation is not optimal. For example, consider a model with a nominal covariate (e.g., race, diagnostic category, gender) and a continuous outcome. A complete-data regression analysis uses a set of dummy codes to represent the nominal covariate, and it does so without imposing distributional assumptions on predictors. In contrast, maximum likelihood missing data handling requires distributional assumptions for the incomplete variables, and software packages would typically force the user to treat a set of incomplete dummy codes as though they were multivariate normal (and some software programs will simply exclude cases with missing predictor scores). An analysis that features scale scores computed from a set of questionnaire items is another common situation where maximum likelihood missing data handling is surprisingly difficult. Because it does not fill in the data, maximum likelihood effectively encourages the user to treat the scale as missing when one or more of its component items is missing. Specifying an analysis that leverages the typically-strong correlations among the items can be difficult, and ignoring this source of information can decimate power (Gottschall, West, & Enders, 2012; Mazza, Enders, & Ruehlman, 2015).

In my experience, multiple imputation is often a better tool for behavioral science data because it gives researchers the flexibility to tailor the missing data handling procedure to match a particular set of analysis goals. For example, mixtures of categorical and continuous variables (e.g., a regression analysis with an incomplete nominal covariate) pose no problem for multiple imputation, and composite scores with incomplete item responses are similarly benign. Because a number of accessible descriptions of maximum likelihood estimation appear in the literature (Enders, 2010, 2013; Graham, 2012; Schafer & Graham, 2002), I limit the scope of this manuscript to multiple imputation, focusing on practical issues that clinical researchers are likely to encounter in their work. Throughout the paper, I use a series of data analysis examples to illustrate the application of multiple imputation to problems that are not necessarily easy to handle with maximum likelihood estimation. Although multiple imputation is widely available in most general-use software packages, I use the Blimp application (Enders, Keller, & Levy, 2016; Keller & Enders, 2014) because it is flexible enough to accommodate a variety of scale types (nominal, ordinal, and continuous) with single-level and multilevel data, and it can be used in conjunction with any analysis program. Blimp is available for the Mac and Windows operating systems and is available for free download at www.appliedmissingdata.com/multilevel-imputation.html.

1. Motivating example

The analysis example comes from a study of an online chronic pain management program (Ruehlman, Karoly, & Enders, 2012), where individuals were randomly assigned to an intervention condition ($n = 167$) or a wait-listed control group ($n = 133$). The primary focus of this example is a 6-item depression measure, which researchers administered at pretest, 7-week follow-up, and 14-week follow-up. The data set also includes a number of background variables (e.g., gender, age, education) and baseline measures of pain severity and pain interference with daily life activities. Table 1 gives the percentage of observed values for a subset of variables that I use throughout the paper. So that interested readers can work through the data analysis examples, I used the means and correlations from real data to create an artificial data set that mimics the original. The data set and analysis scripts are available

Table 1
Percentage of observed data for analysis variables.

Variable	Name	% Complete	Range
Intervention code	TXGRP	100.0	0–1
Gender	Male	100.0	0–1
Age	Age	100.0	18–78
Education	Educ	95.0	1–7
Exercise frequency	Exercise	93.3	1–8
Pain interference	Interf	100.0	6–42
Pain severity rating	Severity	93.7	1–7
Wave 1 depression item 1	T1DEP1	100.0	1–5
Wave 1 depression item 2	T1DEP2	94.7	1–5
Wave 1 depression item 3	T1DEP3	94.7	1–5
Wave 1 depression item 4	T1DEP4	100.0	1–5
Wave 1 depression item 5	T1DEP5	100.0	1–5
Wave 1 depression item 6	T1DEP6	100.0	1–5
Wave 2 depression item 1	T2DEP1	84.7	1–5
Wave 2 depression item 2	T2DEP2	86.7	1–5
Wave 2 depression item 3	T2DEP3	91.3	1–5
Wave 2 depression item 4	T2DEP4	91.3	1–5
Wave 2 depression item 5	T2DEP5	91.3	1–5
Wave 2 depression item 6	T2DEP6	86.7	1–5
Wave 3 depression item 1	T3DEP1	76.3	1–5
Wave 3 depression item 2	T3DEP2	76.3	1–5
Wave 3 depression item 3	T3DEP3	76.3	1–5
Wave 3 depression item 4	T3DEP4	72.7	1–5
Wave 3 depression item 5	T3DEP5	72.7	1–5
Wave 3 depression item 6	T3DEP6	74.0	1–5
Wave 1 depression scale	DEP1	89.7	6–30
Wave 2 depression scale	DEP2	77.7	6–30
Wave 3 depression scale	DEP3	68.7	6–30

[2] It is difficult to identify a rule of thumb for a "sufficiently large" sample, but my experience suggests that multiple imputation and maximum likelihood can yield equivalent estimates in samples that are typical in psychological research (e.g., $N = 200$).