State of the Art of Missing Data in Meta-analysis

4/29/2020

Aims to answer 3 questions:

1. What are people talking about theory of missing data in meta-regression?
2. What are they doing in practice? Why?
3. What should they be doing?

## Introduction

## 1. Methods for Handling missing data in research synthesis

This section will attempt to answer question 1. But will be mostly focused on missing predictor variables.

- 1.1 "Ad-hoc" methods:  
 1.1.1 Listwise deletion or Complete Case Analysis

In a listwise deletion (Complete Case Analysis), all cases with one or more missing values on the analysis variables are eliminated. Thus, only those cases with all variables fully observed are used.

Default procedure for handling incomplete data in many statistical packages, such as SPSS, SAS and Stata.

This procedure is one of the most common methods used when missing data on predictors occurs. In meta-analysis, researchers typically use only those studies reporting on a given predictor when estimating a model of effect-size variation. However, selecting only complete cases can limit the number of observations available for the analysis.

If the data are MCAR, listwise deletion produces unbiased estimates of means, variances and regression weights. If the data are not MCAR, listwise deletion can bias estimates (Little and Rubin 2002, pp.41-44)

1.1.2 Pairwise deletion or Available-case Analysis

A pairwise deletion, also named available-case analysis, calculates the means and variances on all observed data.

Under MCAR produces consistent estimates of mean, correlation and covariances (Little and Rubin, 2002, p.55). However, if the data are MAR it will produce biased estimates.

Further, the covariance and/or correlation matrix may not be positive definite, which is requirement for mot multivariate procedures.

1.1.3 Single-Value Imputation Methods  
   
 3.1 Complete Case Mean

Replace the missing value with the mean of the observed cases in each variable.

In meta-analysis, researchers may fill in the complete case mean for a missing predictor value.

Fast and simple fix for the missing data.

Underestimate the variance, disturb the relations between variables, bias almost any estimate other than the mean and bias the estimate of the mean when data are not MCAR.

3.2 With Conditional means (regression model)

Regression imputation also known as conditional mean imputation, includes information of other variables with the idea of constructing improved imputations. Under this procedure, each missing value is imputed with the predicted value from a regression model using the observed variables as predictors and the missing variable as the outcome.

Under MCAR regression imputation generates unbiased estimates of the means and the regression weights. Additionally, the regression weights are unbiased under MAR if the factors that influence the missingness are part of the regression model.

However, regression imputation artificially strengthens the relations in the data. Correlations are biased upwards. Variability is underestimated.

3.3 LOC and BOCF

Last observation carried forward (LOCF) and baseline observation carried forward (BOCF) use the previous observed value as a replacement for the missing data and are usually implemented with longitudinal data.

LOCF can produce biased estimates even under MCAR (Molenberghs and Kenward, 2007, p. 47-50).

“Any single-value imputation strategy that does not adjust standard errors for the uncertainty caused by missing data will produce biased estimates no matter what mechanism leads to the missing observations”(Pigott, 2019)

- 1.2 "Model Based" Methods:  
 1.2.1 Maximun Likelihood Methods Using EM Algorithm  
 1.2.2 Multiple Imputation for Multivariate Normal Data

**NOTE FROM JAKE:** You might mention something like hot-deck imputation, last-observation-carried-forward, and random imputation (replacing a cell from another row chosen at random, which can happen with hot-deck) in your “ad-hoc” methods.

**NOTE FROM JAKE:** There are some great statistical papers from the 1990s by Joe Ibrahim (see our Zotero repo) on EM for missing predictors in GLM, including methods that incorporate MNAR. It would be *tough*, but a worthwhile contribution to derive and implement those for meta-analysis.

**NOTE FROM JAKE: It’s important to point out “missing data in meta-analysis” is inherently ambiguous. You could be missing data from individual participants within studies. You could be missing entire studies (that are in the filedrawer). You could be missing effects from studies that you don’t even know are missing (i.e., some effects from a study are in the file drawer). You could also be missing information about effects or covariates, which is what I think you’re getting at. Not a huge thing to fix, but seems like it would help situate what you’re talking about.**

**NOTE FROM JAKE:** It strikes me there could be a neat paper to write about missingness within studies versus missingness between studies. Like, if there is attrition within studies, how does that affect a meta-analysis? Conversely, if there is little attrition within studies, but we’re missing effects (e.g., from publication selection), how much can that screw up a meta-analysis?

## 2. Current Practices in meta-analysis

This section will try to summarized current practices for handling missing data in meta-anylisis.(What methods are prefered and why?)

- 2.1 Commonly used methods:   
 "45% of studies report missing data problems. When missing data were found, the majority (86%) of synteses reported using ad hoc approach to handle missingess. Only 3% reported using multiple imputation." (Tipton et al., 2019).  
 - 2.2 Fundamental problems of ad hoc approaches  
 - 2.3 Advantages of using Model based methods   
 - 2.4 Remaining work needed   
 "Remains a need for translational work on missing data methods, including compelling cases studies, examples of software implementation, and simulation studies on the comparative performance of current practies versus methods such as multiple imputation."(Tipton et al., 2019)

**NOTE FROM JAKE:** Per point 2.1, are they talking about missingness within studies (i.e., we’re missing individual participant data in a given study) or missingness from the meta-analytic dataset itself? That would be an important thing to know.

**NOTE FROM JAKE:** A key thing to keep in mind is that missing data corrections are all about assumptions. There’s the standard assumptions about why data are missing, which we refer to as MNAR/MAR/MCAR. But, there’s also assumptions about what we *would have* observed. That is, model-based methods make assumptions about *why* data are missing and *what we would likely have seen* if there was no data missing. Neither of those are particularly testable in an explicit sense. However, we can think about ways to examine if they are feasible (e.g., from a Bayesian perspective we could look at posterior predictive distributions).

## References