Dissecting Multiple Imputation from a Multi-phase Inference Perspective: What Happens When There are Three Uncongenial Models Involved?

Xianchao Xie and Xiao-Li Meng

Harvard University

Abstract: Real-life data are almost never really *real.* By the time the data arrive at an investigator's desk or disk, the raw data, however defined, have most likely gone through at least one "cleaning" process, such as standardization, re-calibration, imputation or de-sensitization. Dealing with such a reality scientifically requires a more holistic multi-phase perspective than is permitted by the usual framework of "God's model versus my model." This article provides an in-depth look, from this broader perspective, into multiple-imputation (MI) inference (Rubin, 1987) under uncongeniality (Meng, 1994). We present a general estimating-equation decomposition theorem, resulting in an analytic description (asymptotically) of MI inference as an integration of the knowledge of the imputer and the analyst, and establish a characterization of self-efficiency (Meng, 1994) for regulating estimation procedures. These results help to reveal *how* the quality of and relationship between the imputer's model and analyst's procedure affect MI inference, including how a seemingly perfect procedure under the "God-versus-me" paradigm is actually inadmissible when there are three uncongenial models involved: God's, imputer's, and analyst's models. Our theoretical investigation also leads to procedures that are as trivially implementable as Rubin's combining rules, yet with confidence coverage guaranteed to be minimally the nominal level, under any degree of uncongeniality. We reveal that the relationship is very complex between the validity of approaches taken for individual phases and the validity of the final multi-phase inference. These results and many open problems are presented to raise the general awareness that the multi-phase inference paradigm is an uncongenial forest populated by thorns, as well as some fruits, many of which are still low-hanging.

Key words and phrases: Confidence validity, data cleaning, estimating equation decomposition, incomplete data, multiphase inference, pre-processing, self-efficiency, strong efficiency, uncongeniality