Abstract

“Analysis of multiple source/multiple informant data in *Stata*” di Nick Horton

We describe regression-based methods for analyzing multiple-source data arising from complex sample survey designs in Stata. We use the term multiple-source data to encompass all cases where data are simultaneously obtained from multiple informants, or raters (e.g., self-reports, family members, health care providers, administrators) or via different/parallel instruments, indicators or methods (e.g., symptom rating scales, standardized diagnostic interviews, or clinical diagnoses). This is an important problem in many social science and medical research areas. We review regression models for analyzing multiple source risk factors and multiple source outcomes and show that they can be considered special cases of generalized linear models, albeit with correlated outcomes. We show how these principled data combination methods can be extended to handle the common survey features of stratification, clustering, and sampling weights as well as missing reports, and how they can be fit within Stata. The methods are illustrated using data from health services research.