Missing Data and Applied Solutions (May 23-24, 2012)

Instructor: Fred Boehmke, Political Science, University of Iowa
Where: B&E, 148, 105
When: Wed-Thurs, May 23 & 24, 9 am to 4:30 pm, with a break for lunch (12 pm to 12:45).
Who is eligible? Graduate students and faculty at the University of Kentucky can enroll for free.
You must register for both the morning and afternoon sessions. Seating for the computer lab is limited to 50 seats. The lecture hall will seat 100+

Description: (Materials will be added in the next few weeks)

The two-day workshop will discuss different forms of missing data, attendant problems and various solutions. The presentation will include lectures in the morning session (in B&E 148) and computer lab sessions in the afternoon (in B&E 105). The basic principles of missing data will be covered, as well as how to address the problem theoretically and practically using various software solutions. Emphasis will be on acquiring a practical understanding for applied researchers.

Missing data is ubiquitous in applied social science research. Survey questions often go unanswered, businesses may be exempted from reporting requirements, and countries do not share their national statistics. Missing data is a serious problem and traditional “solutions” such dropping all observations with incomplete data raise questions of bias, validity and sample selection.

This workshop will provide an overview of the primary issues associated with missing data as well as various proposed solutions. Missing data can be categorized in three ways: (1) missing at random (MAR), (2) missing completely at Random (MCAR), and (3) non-ignorable (NI). Each has different underlying mechanisms that produce the observed pattern of missing data and implies different concerns for empirical analysis as well as different solutions. MCAR, for example, poses few problems other than larger standard errors; MAR generally produces biased estimates of sample means but unbiased estimates of regression coefficients; NI leads to both biased sample statistics and usually leads to biased regression coefficients.

Various ways to address to these problems will be discussed, including listwise deletion, interpolation, extrapolation, multiple imputation through various means, and maximum likelihood, including Heckman's solution for non-random sample selection for continuous and binary outcome variables. Software to implement and compare these solutions will be discussed, in particular various Stata commands for interpolation, extrapolation, and multiple imputation through chained equations. As necessary, additional software such as Amelia II will be covered for alternate applications, such as time-series-cross-sectional data sets . Examples will be discussed for both real-world data sets and hypothetical data sets in which we know the true parameter values and patterns of missingness.

Participants should have a basic knowledge of data manipulation and analysis in Stata (or similar software), including generating variables and running common regression models for continuous and binary data.

Supported by QIPSR, Statistics, Agricultural Economics, and the Center for Poverty Research, as well as Sociology, Political Science and the College of Arts & Sciences.