Management and Analysis of Large Survey Data Sets Using the `memisc` Package

Martin Elff
Universität Mannheim

March 25, 2008

Abstract

One of the aims of the `memisc` package is to make life easier for useRs who have to work with (large) survey data sets. It provides an infrastructure for the management of survey data including value labels, definable missing values, recoding of variables, production of code books, and import of (subsets of) SPSS and Stata files. Further, it provides functionality to produce tables and data frames of arbitrary descriptive statistics and (almost) publication-ready tables of regression model estimates. Also some convenience tools for programming and simulation are provided, as well as some miscellaneous probability distributions, statistical models, and graphics.

Based on an example analysis of the cumulated ALLBUS 1980-2004 data set (ZA-No. 4243), it is demonstrated how even large data sets can be handled without much pain using the `memisc` package. The cumulated ALLBUS comprises data of 44,526 respondents and 1,141 (!) variables. The proposed presentation shows the workflow of analysis of such a large data set: First, variables that are relevant for the analysis are loaded selectively into the workspace, thus minimizing the overall memory footprint. Second, attributes of variables in such a data set, like variable labels, value labels and user-defined missing values are retained and used for data management conducive for typical social science data analysis. Third, tables of descriptive statistics are produced for preliminary or exploratory analyses using the `genTable` function of the package. Fourth, estimates of statistical models are formatted in a way suitable for publication in social science journals using the `mtable` function.