SUG Conference 2006

A Stata procedure for the de-duplication of individual records: the INPS archive case

Orietta Dessy

(Università Bocconi, fRDB)

Abstract

A quite common problem in collecting individual data is the duplication of individuals. Often the attribution of a code that identifies an individual is very sensitive to variables such as name, surname, date and place of birth, address. This information is imputed when the person is registered the first time in the archive, and then it is checked and updated each time that the same person is contacted. The problem is that any mistake in reporting individuals' identifying information gives rise to a new identifying code, therefore creating wrongly a new person in the archive. Therefore the need for de-duplicating observations for the same individual arises. The construction of an appropriate program for solving this problem can be developed at different stages. First of all, some general checks of coherence have to be implemented, using all the available information in the archive (name, surname, fiscal or social security codes, sex, date and place of birth, variables of assessment of quality of the collected information, etc.). Then, at a second stage, some general criteria of phonetic assonance can be used for de-duplicating observations in an appropriate probabilistic environment. Our program does not correct imputed wrong data with the right ones, but simply generates individual identifying codes that can sensibly reduce the cases of duplication of individual records, making their identity anonymous at the same time. This is sufficient and useful for carrying out any kind of statistical, econometric and, in particular, panel-data analysis on data subject to privacy restrictions. Further research should be devoted to the possibility of correcting for the right information, possibly using pre-constructed universal vocabularies, so that the program can be extended to the cases where individuals' details are needed for the purposes of the analysis. As an example of application of our routine, we use the Italian administrative archive of the National Institute of Social Security (INPS).