ReGenesees: symbolic computation for calibration and variance estimation

Diego Zardetto

1. Istat – The Italian National Institute of Statistics
*Contact author: zardetto@istat.it

Keywords: Complex estimators, variance estimation, automated linearization, symbolic computation

ReGenesees (R Evolved Generalized Software for Sampling Estimates and Errors in Surveys) is a full-fledged R software for design-based and model-assisted analysis of complex sample surveys. It is the outcome of a long-term research and development project, aimed at defining a new standard for calibration, estimation and sampling errors assessment to be adopted in all Istat large-scale sample surveys. The system is distributed as Open Source Software, under the EUPL license. It can be freely downloaded from Istat website1, as well as from JOINUP2. ReGenesees is rather different from existing estimation platforms developed by NSIs (mostly based on SAS) from both the application logic and the user experience standpoints. In a nutshell:

(1) User interaction with the new system takes place at a very high level of abstraction. ReGenesees users, indeed, no longer need to preprocess the survey data relying on ad-hoc programs; instead, they only have to feed the software with (i) the data as they are, plus (ii) symbolic metadata that describe the adopted sampling design and calibration model. At that point, it is up to the system itself to transform, in an automatic and transparent way, the survey data into the complex data structures required to solve the calibration problem and to compute estimates and errors.

(2) Besides Totals, and Absolute Frequency Distributions (linear estimators that are covered by all traditional platforms), ReGenesees allows to compute estimates and sampling errors with respect to Means, Ratios, Multiple Regression Coefficients, Quantiles, and, more generally, with respect to any Complex Estimator, provided it can be expressed as a differentiable function of Horvitz-Thompson or Calibration Estimators. It is worth stressing that such Complex Estimators can be defined in a completely free fashion: the user only needs to provide the system with the symbolic expression of the estimator as a mathematical function. ReGenesees, indeed, is able to automatically linearize such Complex Estimators, so that the estimation of their variance comes at no cost at all to the user.

Traditional estimation softwares did not give any support to the users in preparing auxiliary variables and population totals for calibration, nor in deriving the Taylor expansion of non-linear estimators and in computing the corresponding linearized variable for variance estimation purposes. As a consequence, ad-hoc (often very complex) programs for data preparation, transformation and validity check were developed and maintained outside the scope of the estimation system: a time-consuming and error-prone practice. ReGenesees frees its users from such needs, with an evident gain in terms of workload reduction, better usability and increased robustness against possible errors. Interestingly, both the innovative ReGenesees features sketched above leverage a peculiar strong point of the R programming language, that is its ability to process symbolic information.

1 http://www.istat.it/it/strumenti/metodi-e-software/software/regenesees
2 https://joinup.ec.europa.eu/software/regenesees/description