

ReGenesees: an software for computing estimates and sampling errors



Diego Zardetto
Technology and Methodological Support

What is ReGenesees?

- ReGenesees acronym:
 - R Evolved Generalised Software for Estimates and Errors in Surveys
- Scope:
 - Design-Based and Model-Assisted analysis of complex sampling surveys
- Programming Language:
 - Entirely developed in interpreted R code
- System Architecture:
 - 2 integrated R packages:
 - ✓ Package **ReGenesees**: implements the application layer of the system
 - ✓ Package **ReGenesees . GUI**: implements the presentation layer of the system

Minimum Objectives of the ReGenesees Project

- Build an R-based software system able to:
 - cover the main functionalities of GENESEES/SAS (a former - non open - Istat system), i.e. calibration and calculation of estimates and standard errors for Estimators of Totals
 - extend GENESEES/SAS by adding sw modules implementing new statistical methods:
 - ✓ Ratios
 - ✓ Quantiles
 - ✓ Automated Linearization of Complex Analytic Estimators
 - ✓ Variance Estimation of Non-Analytic Estimators (e.g. Poverty) by means of Replication Methods (e.g. DAGJK)
 - re-engineer the old system in order to enhance its overall quality (mainly in terms of robustness, usability and maintainability)

Main Statistical Functions of the System (1/2)

- ✓ Complex Sampling Designs
 - Multistage, stratified, cluster sampling designs
 - Unequal inclusion probabilities, with or without replacement
 - Mixed sampling designs (with SR and NSR strata)
- ✓ Calibration
 - Global and partitioned calibration
 - Unit-level and cluster-level calibration
- ✓ Estimators
 - Horvitz-Thompson (and functions of them)
 - Calibration Estimators (and functions of them)
- ✓ Sampling Variance Estimation
 - Multistage (Bellhouse recursive descent algorithm)
 - Ultimate Cluster Approximation
 - “GENESEES-like” for mixed designs
 - Taylor-series linearization for nonlinear Estimators

Main Statistical Functions of the System (2/2)

- ✓ Estimates and Sampling Errors (Standard Errors, Coefficients of Variation, Variances, Confidence Intervals, Design Effects) for:
 - Totals
 - Means
 - Marginal Distributions (absolute and relative frequency)
 - Joint Distributions (absolute and relative frequency)
 - Ratios between Totals
 - Quantiles (Variance estimation by Woodruff method)
- ✓ Estimates and Sampling Errors for Complex Estimators, interactively defined by the user:
 - Analytic Functions of (HT or Calibrated) Estimators of Totals and Means
 - Variance estimation by Automated Linearization
- ✓ Estimates and Sampling Errors by Domains (subpopulations)

System Architecture

- The “2 package” architecture of the ReGenesees system aims at favouring versatility and flexibility of usage

✓ ReGenesees can be used under different setups:

Configuration	Interaction	User Skills			
		Stat ▼ R ▼	Stat ▲ R ▼	Stat ▼ R ▲	Stat ▲ R ▲
ReGenesees . GUI + ReGenesees	GUI	×	×	×	×
	GUI + CLI	–	×	×	×
ReGenesees	CLI	–	–	–	×

- Advantages under the IT perspective
 - ✓ Parallel and (almost) independent development / testing / evolution of the 2 packages
 - ✓ Support to users and maintenance can be managed by distinct human resources

ReGenesees package structure

- Package **ReGenesees** contains ~ 7.000 lines of R code
- Package skeleton is made up of ~ 170 named functions
 - ~ 20 of them are public (i.e. fully “user visible”) and are intended to be invoked directly by the user
 - ✓ the GUI of the system communicates with the application layer only through such ~ 20 functions
 - remaining ~ 150 functions are private (i.e. “hidden”, so to speak) and are intended to be called by other functions
- Package **ReGenesees** provides a Namespace management mechanism:
 - ✓ avoids undesired, accidental masking of internal functions
 - ✓ ensures implicit, automatic loading of needed libraries

ReGenesees package design principles (1/2)

- Formulas and computation techniques in the realms of Design-Based and Model-Assisted Survey Sampling involve **both survey data** and a lot of **meta-data**
- Typical **meta-data** are e.g. information describing the adopted sampling design (stages, cluster, strata, inclusion probabilities, fpcs...) or the desired calibration procedure (assisting model, auxiliary variables, population totals, calibration metric...)
- Thus, **binding survey data to meta-data** in an effective and persistent way has to be considered a major task when designing a good, general purpose survey software
- All statistical analysis functions operating on survey data should **automatically** find and use the meta-data appropriate to **that survey** and **that analysis**

ReGenesees package design principles (2/2)

- From this viewpoint the **ReGenesees** package is a best-practice example:
 - the user is **first** asked to simultaneously specify data and meta-data so that both can be bound in a complex **object**
 - any subsequent analysis has **then** to be operated on that complex object by means of a dedicated statistical **method**
 - the request of a statistical analysis (e.g. compute the standard error of an estimator) on a given object is **automatically dispatched** to the **suitable program**, according to the **class** of the object
- From an Object Oriented perspective, **ReGenesees** programs naturally fall into the following clean-cut decomposition:
 - Object (**class instance**) builders
 - Statistical analysis functions (**“main” methods**) operating on objects
 - Utility tools (**“auxiliary” methods**)

ReGenesees: Object-Oriented overview

- ✓ Object (**class instance**) builders
 - `e.svydesign`
 - `e.calibrate`
 - ...
- ✓ Statistical Analysis Functions (package **“main” methods**)
 - `svystatTM`
 - `svystatR`
 - `svystatQ`
 - `svystatL`
 - `aux.estimates`
 - ...
- ✓ Utility Tools (package **“auxiliary” methods**)
 - `des.addvars`, `collapse.strata`
 - `cv`, `SE`, `VAR`, `deff`, `confint`, `coef`
 - `write.svystat`
 - `pop.template`, `population.check`, `fill.template`
 - `bounds.hint`
 - `g.range`
 - ...

ReGenesees vs. former survey analysis systems (1/3)

Calibration Process

- ✓ user-system interaction at an higher level of abstraction
 - the user specifies the calibration model simbolicly, via model formulae
→ the system automatically deduces auxiliary variables and calibration domains
- ✓ no need of pre-processing survey data files
 - by using meta-data (calibration model, auxiliary variables types and interactions, calibration domains, ...) the system automatically transforms the survey data → the user doesn't need to manage: disjunctive representations of categorical variables, cartesian products between modalities, elimination of linearly dependent variables, ...
- ✓ support in building / checking / filling known totals datasets
 - driven by meta-data, the system generates a “template” dataset to store the known totals → the user doesn't need to understand the “standard format” required: he has only to fill-in the template with the actual totals he has at hand
 - whenever a sampling frame is available → the system handles also the operation of computing the known totals and correctly filling the template

ReGenesees vs. former survey analysis systems (2/3)

Calibration Process

✓ cluster-level calibration

- the system allows to constrain the calibrated weights to be equal inside clusters selected at a given stage → the user doesn't need to “aggregate” survey data by cluster before calibrating, nor to “re-expand” the data after calibration, when computing estimates

✓ better computational efficiency

- the calibration algorithm performs also “partial” computations (i.e. QR decomposition of the model-matrix) which are not directly addressed to determine final weights → such informations are stored in order to be subsequently used in the estimation phase

ReGenesees vs. former survey analysis systems (3/3)

Estimation Process

- ✓ fully integrated with calibration
 - the meta-data describing the calibration estimator are encapsulated inside the calibrated object: they need not to be asked again to the user → the system prevents the user from applying variance estimation algorithms in a methodologically inconsistent way
- ✓ computational efficiency gain in estimating the variance of calibration estimators
 - the most intensive computation performed when calculating regression coefficients is factorised: $(X^t D X)^{-1}$ → matrix inversion is performed only once, i.e. not repeated when changing interest variable
- ✓ user-friendly management of non-linear estimators
 - the linearized variable associated to the complex estimator (the so called “Woodruff transform”) is computed transparently by the system → the user doesn’t need to perform calculations “on paper”, nor to subsequently implement them by writing code

ReGenesees: an R software for computing estimates and sampling errors

- Further enrichments of the system:
 - ✓ Integrating inside **ReGenesees** the **EVER** package → sampling variance estimation for non-analytic estimators (e.g. poverty) by the DAGJK method
 - ✓ Synthetic presentation of sampling errors → can exploit R powerful facilities for fitting and plotting regression models
- Current and planned actions:
 - ✓ ReGenesees has been widely beta-tested and has already been used in production for 3 structural business surveys
 - ✓ ReGenesees beta-version is currently available on Istat intranet → testing and validation on a larger scale
 - ✓ Official release of ReGenesees on Istat website for General Availability → December 2011