Austrian Journal of Statistics
MMMMMM YYYY, Volume VV, 1–xx.
http://www.ajs.or.at/

Simulation Tools for Small Area Estimation: Introducing the R-package saeSim

Sebastian Warnholz
Freie Universität Berlin

Timo Schmid
Freie Universität Berlin

Abstract

The abstract of the article in English

Keywords: package, small area estimation, reproducible research, simulation, R.

1. Introduction

A brief overview of what the reader can expect of this article. What is the aim of this article and where can it be located in science. Discussion of the reproducible research movement and where the package may be beneficial. Packages like Sweave, knitr, R-markdown. How can simulation studies (the R-code) be published? As script? As stand-alone-package? How can ideas for specific data generation tools be shared amongst the small area community and thus be more transparent: saeSim (far-fetched but anyway).

2. Simulation studies in small area estimation

The role of simulation studies in the field. Separation into model-based and design-based simulation studies. How does saeSim address these 2 perspectives? Identifying steps/phases/components of simulation studies (data generation, sampling, computation).

3. saeSim: A simulation framework

Introducing a simulation study as a data manipulation process. From data generation to estimating models. How can this process be mapped into the R-language. Present the simulation phases as flow-diagram and map the function names to the problem domain. Defining interfaces between phases: data.frame in, data.frame out. What is the difference between a design-based (fixed-population) and model-based (random-population) approach.

Why is this package helping in terms of literate programming (code is written for humans not machines) and reproducible research? Package naming conventions. Use the %>% operator to compose simulation set-ups. Reuse defined scenarios to compose new scenarios: What is a contamination scenario? A standard scenario plus contamination:

```
> contaminatedSetup <- standardSetup %>% sim_gen_cont()
```

3.1. Data generation

This is not a data generation tool. However, it supports some useful functions to add random variates to the data. Definition of outliers, see Rao (?) – how is it addressed in the package. More than linear models, define any response with:

```
> setup %>% sim_resp_eq(y = g(2 * x) + e)
```

3.2. Drawing samples

Is present in unit-level simulation studies. Draw with simple random sampling from the whole population or within cluster/domains. Specify the sample size as integer or fraction and add weights if necessary. Wrappers around dplyr::sample_n and dplyr::sample_frac.

3.3. Adding customized computations

Adding user specified functions to the simulation process is what separates this package from a mere data generation tool. The interface is simple: Add functions with one argument which is the data at that moment and which return the modified data. I am not sure if this and the sampling section is necessary – maybe it should just be summarized in the overview section...

3.4. Separation of designing and running simulations

First develop a scenario outside any looping structures, then run it R-times:

```
> simulationResults <- simSetup %>% sim(R = 500)
```

There is an easy back-end to run simulations in parallel (in Windows with special care):

```
> simulationResults <- simSetup %>% sim(R = 500, parallel = TRUE, mc.cores = 8)
```

4. Outlook

Use this package to share and publish simulation studies alongside papers. Contribute to the package to make your ideas available. Contribute to the package and make your whole simulation study available.

Affiliation:

Sebastian Warnholz
Department of Economics
Freie Universität Berlin
D-14195 Berlin, Germany
E-mail: Sebastian.Warnholz@fu-berlin.de
URL: http://www.wiwiss.fu-berlin.de/fachbereich/vwl/Schmid/Team/Warnholz.html

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume VV *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd