The Journal of Open Source Software

simstudy: Illuminating research methods through data generation

Keith Goldfeld[1] **and Jacob Wujciak-Jens**[2]

1 NYU Grossman School of Medicine. **2** Independent Researcher

DOI: 10.21105/joss.02763

Software

- Review
- Repository
- Archive

Editor: Mikkel Meyer Andersen

Reviewers:

- @gagolews
- @brunaw

Submitted: 18 October 2020
Published: 27 October 2020

Summary

The `simstudy` package is a collection of functions for R (R Core Team, 2020) that allow users to generate simulated data sets in order to explore modeling techniques or better understand data generating processes. The user defines the distributions of individual variables, specifies relationships between covariates and outcomes, and generates data based on these specifications. The final data sets can represent randomized control trials, repeated measure designs, cluster-randomized trials, or naturally observed data processes. Many other complexities can be added, including survival data, correlated data, factorial study designs, step wedge designs, and missing data processes.

Simulation using `simstudy` has two fundamental steps. The user (1) **defines** the data elements of a data set and (2) **generates** the data based on these definitions. Additional functionality exists to simulate observed or randomized **treatment assignment/exposures**, to create **longitudinal/panel** data, to create **multi-level/hierarchical** data, to create datasets with **correlated variables** based on a specified covariance structure, to **merge** datasets, to create data sets with **missing** data, and to create non-linear relationships with underlying **spline** curves.

The overarching philosophy of `simstudy` is to create data generating processes that mimic the typical models used to fit those types of data. So, the parameterization of some of the data generating processes may not follow the standard parameterizations for the specific distributions. For example, in `simstudy` *gamma*-distributed data are generated based on the specification of a mean μ (or $\log(\mu)$) and a dispersion d, rather than shape α and rate β parameters that more typically characterize the *gamma* distribution. When we estimate the parameters, we are modeling μ (or some function of (μ)), so we should explicitly recover the `simstudy` parameters used to generate the model - illuminating the relationship between the underlying data generating processes and the models. For more details on the package, use cases, examples, and function reference see the documentation page.

`simstudy` is available on CRAN and can be installed with:

```r
install.packages("simstudy")
```

Alternatively, the newest development version can be installed from GitHub with:

```r
# install.packages("devtools")
devtools::install_github("kgoldfeld/simstudy")
```

Statement of need

Empiricism and statistical analysis are cornerstones of scientific research but they can lead us astray if used incorrectly. Choosing the right methodology for the hypothesis and expected data is crucial for useful, valid results. Data simulated with `simstudy` under the assumptions derived from a hypothesis enables researchers to test and refine their analysis methodologies without the need for time-intensive, expensive pre-tests or collection of actual data. Additionally data generated with `simstudy` can be used in generalized, theoretical simulation studies to further the field of methodology.

There are several `R`-packages that allow for data generation under different assumptions. Most of these packages have a narrower scope that focuses on a specific class of data, like `ICCbin` (Hossain & Chakraborty, 2017), `BinNonNor` (Inan, Demirtas, & Gao, 2020) and `genSurv` (Meira-Machado & Faria, 2014). Some do not seem to be actively maintained (Alfons, Templ, & Filzmoser, 2010; Bien, 2016; Chan, 2014; Hofert & Mächler, 2016), which can cause compatibility issues. Some target specific fields of study and their needs, like the psychology-focused `psych` package (Revelle, 2020) or the `conjurer` package (Macherla, 2020) that provides methods to generate synthetic customer data for industry use. `simstudy` is unique with its philosophy of data generating processes that mimic the models used in analysis and allowing for the possibility of generating a wide range of complex data through these processes. The `SimDesign` Package (Chalmers & Adkins, 2020) and the related `MonteCarlo` Package (Leschinski, 2019) follow a similar line of thought but focus on easy replication of the analyses and providing summaries of simulated data.

`simstudy` has been used in a variety of fields for theoretical exploration of research methodology (Anderson, Wennberg, & McMullen, 2019; El Alili et al., 2020; Kirasich, Smith, & Sadler, 2018; Krzykalla, Benner, & Kopp-Schneider, 2020; Liu, Chrysanthopoulou, Chang, Hunnicutt, & Lapane, 2019; Nickodem, 2020; Thoya, Waititu, Magheto, & Ngunyi, 2018; Wang & Ma, 2020), power calculation for trials (Wei et al., 2019) and other simulation tasks supporting researchers (Chukwu, 2019; Forthun et al., 2020; Horry, Fitzgerald, & Mansour, 2020; Renson, Schubert, & Bjurlin, 2017).

Acknowledgements

We acknowledge contributions from James Balamuta, Michael Bradley, Gertjan Verhoeven. For the generation of multivariate binary data the algorithm by Emrich & Piedmonte (1991) is used.

References

Alfons, A., Templ, M., & Filzmoser, P. (2010). An object-oriented framework for statistical simulation: The R package simFrame. *Journal of Statistical Software*, *37*(3), 1–36. doi:10.18637/jss.v037.i03

Anderson, B. S., Wennberg, K., & McMullen, J. S. (2019). Enhancing quantitative theory-testing entrepreneurship research. *Journal of Business Venturing*, *34*(5), 105928. doi:10.1016/j.jbusvent.2019.02.001

Bien, J. (2016). The simulator: An engine to streamline simulations. *Submitted*. Retrieved from http://faculty.bscb.cornell.edu/~bien/simulator.pdf

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. doi:10.20982/tqmp.16.4.p248

Chan, T. J. (2014). *Ezsim: Provide an easy to use framework to conduct simulation*. manual. Retrieved from https://CRAN.R-project.org/package=ezsim

Chukwu, V. (2019). Safety constraint optimization of combination drug therapy in hypertension clinical trials. Retrieved from https://digitalcommons.georgiasouthern.edu/etd/2007/

El Alili, M., van Dongen, J. M., Goldfeld, K. S., Heymans, M. W., van Tulder, M. W., & Bosmans, J. E. (2020). Taking the analysis of trial-based economic evaluations to the next level: The importance of accounting for clustering. *PharmacoEconomics*, 1–15. doi:10.1007/s40273-020-00946-y

Emrich, L. J., & Piedmonte, M. R. (1991). A Method for Generating High-Dimensional Multivariate Binary Variates. *The American Statistician*, *45*(4), 302–304. doi:10.1080/00031305.1991.10475828

Forthun, I., Lie, R. T., Strandberg-Larsen, K., Solheim, M. H., Moster, D., Wilcox, A. J., Mortensen, L. H., et al. (2020). Parental education and the risk of cerebral palsy for children: An evaluation of causality. *Developmental Medicine & Child Neurology*. doi:10.1111/dmcn.14552

Hofert, M., & Mächler, M. (2016). Parallel and other simulations in R made easy: An end-to-end study. *Journal of Statistical Software*, *69*(4), 1–44. doi:10.18637/jss.v069.i04

Horry, R., Fitzgerald, R. J., & Mansour, J. K. (2020). "Only your first yes will count": The impact of pre-lineup instructions on sequential lineup decisions. *Journal of Experimental Psychology: Applied*. doi:10.31234/osf.io/59uaq

Hossain, A., & Chakraborty, H. (2017). *ICCbin: Facilitates clustered binary data generation, and estimation of intracluster correlation coefficient (ICC) for binary data*. manual. Retrieved from https://CRAN.R-project.org/package=ICCbin

Inan, G., Demirtas, H., & Gao, R. (2020). *BinNonNor: Data generation with binary and continuous non-normal components*. manual. Retrieved from https://CRAN.R-project.org/package=BinNonNor

Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), 9. Retrieved from https://scholar.smu.edu/datasciencereview/vol1/iss3/9/

Krzykalla, J., Benner, A., & Kopp-Schneider, A. (2020). Exploratory identification of predictive biomarkers in randomized trials with normal endpoints. *Statistics in Medicine*, *39*(7), 923–939. doi:10.1002/sim.8452

Leschinski, C. H. (2019). *MonteCarlo: Automatic parallelized monte carlo simulations*. manual. Retrieved from https://CRAN.R-project.org/package=MonteCarlo

Liu, S.-H., Chrysanthopoulou, S. A., Chang, Q., Hunnicutt, J. N., & Lapane, K. L. (2019). Missing data in marginal structural models: A plasmode simulation study comparing multiple imputation and inverse probability weighting. *Medical care*, *57*(3), 237. doi:10.1097/MLR.0000000000001063

Macherla, S. (2020). *Conjurer: A parametric method for generating synthetic data*. manual. Retrieved from https://CRAN.R-project.org/package=conjurer

Meira-Machado, L., & Faria, S. (2014). A Simulation Study Comparing Modeling Approaches in an Illness-Death Multi-State Model. *Communications in Statistics - Simulation and Computation*, *43*(5), 929–946. doi:10.1080/03610918.2012.718841

Nickodem, K. (2020). Use of aggregated covariates in propensity score analysis of clustered data. Retrieved from https://conservancy.umn.edu/handle/11299/216126

R Core Team. (2020). *R: A language and environment for statistical computing*. manual, Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Renson, A., Schubert, F. D., & Bjurlin, M. A. (2017). Lack of insurance is associated with lower probability of diagnostic imaging use among US trauma patients: An instrumental variable analysis and simulation. *bioRxiv*, 215889. doi:10.1101/215889

Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research*. manual, Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Thoya, D., Waititu, A., Magheto, T., & Ngunyi, A. (2018). Evaluating methods of assessing optimism in regression models. *Am. J. Appl. Math. Stat*, *6*, 126–134.

Wang, L., & Ma, W. (2020). Improved empirical likelihood inference and variable selection for generalized linear models with longitudinal nonignorable dropouts. *Annals of the Institute of Statistical Mathematics*, 1–25. doi:10.1007/s10463-020-00761-4

Wei, X., Hicks, J. P., Pasang, P., Zhang, Z., Haldane, V., Liu, X., Yin, T., et al. (2019). Protocol for a randomised controlled trial to evaluate the effectiveness of improving tuberculosis patients' treatment adherence via electronic monitors and an app versus usual care in Tibet. *Trials*, *20*(1), 273. doi:10.1186/s13063-019-3364-x