simstudy: Illuminating research methods through data generation

Keith Goldfeld\textsuperscript{1} and Jacob Wujciak-Jens\textsuperscript{2}

\textsuperscript{1} NYU Grossman School of Medicine. \textsuperscript{2} Independent Researcher

Summary

The \texttt{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 \texttt{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 \texttt{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 \texttt{simstudy} gamma-distributed data are generated based on the specification of a mean $\mu$ (or $\log(\mu)$) and a dispersion $d$, rather than shape $\alpha$ and rate $\beta$ parameters that more typically characterize the gamma distribution. When we estimate the parameters, we are modeling $\mu$ (or some function of $\mu$), so we should explicitly recover the \texttt{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.

\texttt{simstudy} is available on CRAN and can be installed with:

\begin{verbatim}
install.packages("simstudy")
\end{verbatim}

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

\begin{verbatim}
# install.packages("devtools")
devtools::install_github("kgoldfeld/simstudy")
\end{verbatim}
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

Goldfeld et al., (2020). simstudy: Illuminating research methods through data generation. Journal of Open Source Software, 5(54), 2763. https://doi.org/10.21105/joss.02763
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

Goldfeld et al., (2020). simstudy: Illuminating research methods through data generation. *Journal of Open Source Software*, 5(54), 2763. doi:https://doi.org/10.21105/joss.02763