9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Introducing `nrobust`: A multiversal library with model selection, averaging, and out-of-sample analysis.

Keywords: Multiverse, Specification-Curves, Robustness, Reproducibility, Selection and Averaging

Extended Abstract

There has recently been an increasing amount of attention paid to the levels of uncertainty around estimates produced in the academic social sciences, with the issue of researcher-induced uncertainty addressed in several high-profile papers [1, 2]. A key issue in researcher-induced uncertainty is the sensitivity of the estimates to different model specifications. These model specifications usually include all possible combinations of 'control' variables, along with other analytical choices. In order to address the sensitivity of these estimates, all possible specifications are typically computed and presented in a curve, showcasing the variability of the output space (both in terms of their absolute value and the variation in their statistical significance). This has come to be known as 'multiverse analysis' when examining multiple types of 'researcher degrees of freedom' or 'specification-curve analysis' when exclusively considering specification choices [5, 6, 7]. In order to reliably estimate and collect the possible specifications present in a single analysis, researchers usually write ad-hoc routines in the programming language of their choice, or use limited existing libraries (predominantly in `R`). The varying quality and stability of such routines usually affects replication efforts, undermining their robustness; ironically one the primary goals of conducting multiverse/specification-curve analysis in the first place. Here we present `nrobust`, an accessible Python library that allows researchers to conduct advanced multiverse analysis with a multitude of new, exciting and much needed features. It aims to provide standard, stable and highly reproducible methods with a user-friendly interface. Given a standard multiple regression model:

$$Y_i = \alpha_0 + \beta_1 X_{i,1} + \sum_{k=2}^{K} \beta_k X_{i,k} + \sum_{j=1}^{J} \lambda_j c_{i,j} + \varepsilon_i$$

For each observation $i = 1, \ldots, N$ with $\{x_{i,1}, \ldots x_{i,K}\} \in X$ being a fixed set of predictors, our approach allows us to:

- Isolate the variation in a 'feasible' output space pertaining to a specific coefficient of interest ($\hat{\beta}_1$).
- Include a set of variables $\{x_{i,2} \ldots x_{i,K})$ variables which should always be included in the model space.
- Estimate the model with all possible combinations of subsets of control variables $\{c_1, \ldots, c_J\}$ of any length up to J, assuming $\{c_1, \ldots, c_J\} \in C$, is a reasonably identified set of auxiliary variables which might – subjectively – be included.
- Compute the model with the arithmetic mean of all possible combinations of $\{y_1, \ldots, y_M\}$ as the dependent variable.

- Select a posteriori models or weight combinations of well-specified models for reliable inference.

- Conduct out-of-sample prediction to analyse the relative and absolute fit of models.

Unlike similar multiverse/specification-curve libraries [7, 4], `nrobust` computes confidence interval using bootstrap re-sampling instead of assumed distributions. While this considerably increases the computational burden, it also allows for more robust results as well as in-depth inspection of the variability of bootstrap estimates. Additionally, `nrobust` computes the Akaike, Hannan–Quinn and Bayesian information criteria for each specification and conducts the Joint-Inference test proposed by Simmonsohn, et al. [5]. It also conducts out-of-sample analysis and compares predictive models of $\hat{y}_i$ against two baselines (null and fully specified models) using the Inter-Model Vigorish [3]. This all allows us to consider posterior model performance to make more reliable inferences based on the possible space of $\hat{\beta}$, an essential advancement given that existent multiversal libraries imply a naively uniform probability distribution across all 'feasible' spaces. We demonstrate the use and interpretation of the outputs of our library by applying it to hypotheses tested in three well-published studies. These relate to labour unionisation, the provision of social care, and children's use of electronic devices, where each application showcases a different facet of our library. We show that `nrobust` can be used to either replicate and expand in new directions the results of studies that have already used multiverse/specification curve analysis in the past, or to expand the understanding of studies that have only used single model estimation for their inferences to date. This not only allows us to consider whether our existent understandings of social phenomenom are statistically reliable, but also to gain a deeper understanding into both their uncertainty and as to how different models fit the data.

References

[1] Nate Breznau et al. "Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty". In: *Proceedings of the National Academy of Sciences* 119.44 (2022), e2203150119.

[2] Colin F Camerer et al. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015". In: *Nature human behaviour* 2.9 (2018), pp. 637–644.

[3] Benjamin W Domingue et al. "InterModel Vigorish (IMV): A novel approach for quantifying predictive accuracy when outcomes are binary". In: ().

[4] Philipp K. Masur and Michael Scharkow. *specr: Conducting and Visualizing Specification Curve Analyses (Version 1.0.0)*. 2020. URL: `https://CRAN.R-project.org/package=specr`.

[5] Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. "Specification curve analysis". In: *Nature Human Behaviour* 4.11 (2020), pp. 1208–1214.

[6] Sara Steegen et al. "Increasing transparency through a multiverse analysis". In: *Perspectives on Psychological Science* 11.5 (2016), pp. 702–712.

[7] Cristobal Young and Katherine Holsteen. "Model uncertainty and robustness: A computational framework for multimodel analysis". In: *Sociological Methods & Research* 46.1 (2017), pp. 3–40.

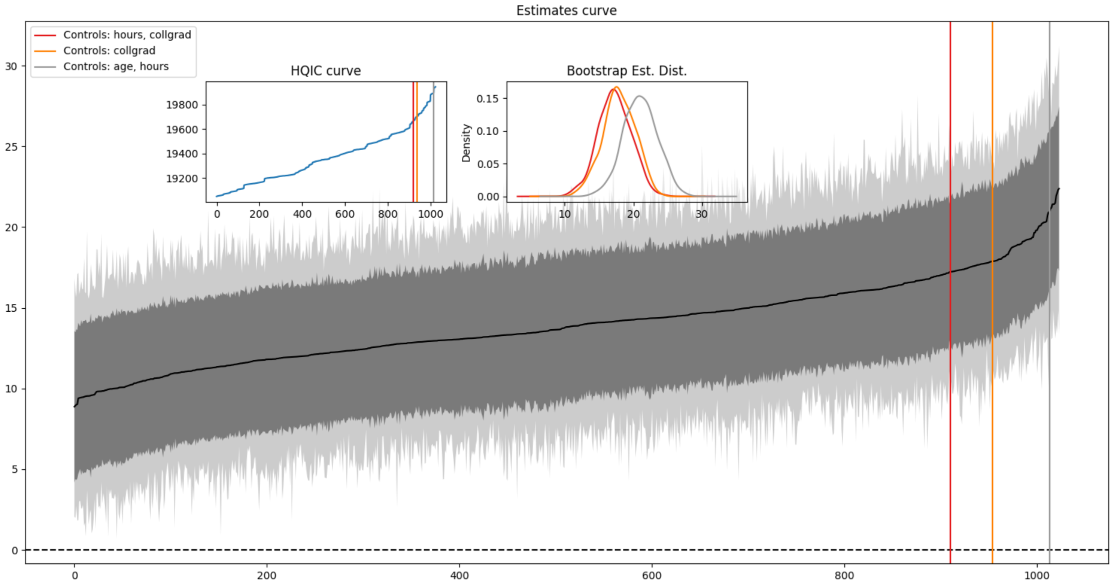

Figure 1: Preliminary output based on a canonical labour union example. Estimates across specifications are shown in solid black, while bootstrapped 95% C.I. are shown in dark-grey. Maximum and minimum bootstrapped estimates are shown in light-grey. The vertical lines represents the position of singular highlighted specifications of interest with covariates = $\{hours, age\}, \{collgrad, hours\}, \{collgrad\}$. The figure also includes embedded plots of the BIC curve across specifications, with the position of the selected specifications highlighted with vertical lines and the distribution of the bootstrap estimates of said specifications.