Appendix: Replicable Data Example

Supplement to 'Valid Simultaneous Inference in High-Dimensional Settings (with the HDM Package for R)' (Bach et al., 2018)

Philipp Bach

2020-12-09, Updated version: December 2020

Introduction

This data appendix is provided in order to make the real-data example in the paper "*Valid Simultaneous Inference in High-Dimensional Settings (with the HDM Package for R)*" fully replicable.

The R-scripts required for replicability are available at the website https://www.bwl.uni-hamburg.de/en/statistik/forschung/software-und-daten.html .

The R package *hdm* can be downloaded from CRAN (https://CRAN.R-project.org/package=hdm) and R-Forge (http://r-forge.r-project.org/projects/hdm/). The multiple testing adjustment methods presented in the paper are available in the *hdm* package starting with Version `0.3.0`.

Data Extract from IPUMS-USA

###1. Register at IPUMS-USA

The data set in the example is taken from the 2016 American Community Survey (ACS). The ACS is collected and managed by IPUMS-USA. To download data from ACS, it is necessary to create an account and declare the interest in the data. After approval through IPUMS, the data can be accessed online (click on "SELECT DATA").

2. Select Samples

Click on "Select Samples" and select the 2016 ACS. Click on "SUBMIT SAMPLE SELECTIONS".

3. Select Variables

To replicate the data, it is necessary to select the following variables. Variables can be selected either by first clicking on the correct variable group, e.g. Person > RACE, ETHNICITY, AND NATIVITY VARIABLES > Race (click on the plus symbol) and proceed this way or by searching variables and adding them to the data cart.

List of variables needed for replication (data preprocessing and analysis):

- age
- degfield
- educd
- empstat

- hispanic
- incwage
- ind1990
- marst
- metro
- nchlt5
- occ2010
- race
- region
- schltype
- school
- sex
- speakeng
- uhrswork
- vetstat
- wkswork2
- yngch

4. Create and download IPUMS-USA Data Extract

The data selection is saved in the data cart. Proceed with a click on "CREATE DATA EXTRACT" and on "SUBMIT EXTRACT" in the next step. It takes some minutes until the data extract is available.

There are different ways to download the data. For instance, it is possible to download the data using the R package "ipumsr" (available at CRAN). An alternative is to take a detour by downloading the data extract (including the codebook and the STATA do-file) and extracting the data set with STATA and then importing the data (in format .dta) into R using the R package "foreign". Since at the time of the data extraction the R package "ipumsr" has not been available we extracted data according to the second alternative.

5. Data Prepocessing

The data can be loaded with R (in format .dta) using the R package "foreign". For data preprocessing, the following R scripts are used.

- "preproc_clean_siminf.R": R script for basic data cleaning (requires functions defined in "helpers_siminf.R")
- "helpers_siminf.R": Contains helper functions needed for data preprocessing when running "preproc_clean_siminf.R".
- "subsample_siminf.R": The data example is performed for a subset of the full data sample in oder to offer an example that is replicable in a reasonable amount of time (and with reasonable compuational power). The selection of the subsample is performed in the R script "subsample_siminf.R".

All files and the data excerpt from IPUMS-USA should be saved in one directory. First run the file "preproc_clean_siminf.R" (specify the working directory in `path` and adjust the name for the data excerpt). After that, run the file "subsample_siminf.R" (specify the working directory in `path`) to save the file "ACS2016_gender.rda").