Instrumental variable quantile regression under random right censoring: replication package

Jad Beyhum, Lorenzo Tedesco, Ingrid Van Keilegom
ORSTAT, KU Leuven

Overview

The code in this replication package is written in R and can be used to obtain the tables and figures of the paper.

Data Availability and Provenance Statements

Statement about Rights

- ☑ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- ☑ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

Summary of Availability

- ☑ All data **are** publicly available.

Details on each Data Source

Data.Name	Data.Files	Location	Provided	Citation
"The National Job Training Partnership Act Study: Dataset"	jtpa/	empirical/	TRUE	U.S. Department of Labor (1993)
"The National JTPA Study Data"	data.csv	empirical/	TRUE	

Data on The National JTPA Study were provided by the W.E. Upjohn Institute for Employment Research. The raw data were down-

loaded from https://www.upjohn.org/data-tools/employment-research-data-center/national-jtpa-study on the 7th March 2023, and it is included in the replication package. The data are in the public domain. Code for data cleaning and analysis is provided as part of the replication package. The resulting cleaned version of the data corresponds to `empirical/data.csv`.

Computational requirements

Software Requirements

- R 4.0.2
  - `Survival` (3.2.11)
  - `foreach` (1.5.1)
  - `doParallel` (1.0.16)
  - `quantreg` (5.86)
  - `dplyr` (1.0.7)
  - `foreign` (0.8.81)

Controlled Randomness

- Random seed is set at line 184 of program cqr.R
- Random seed is set at line 188 of program script.R
- Random seed is set at line 140 of program bootstrap.R
- Random seed is set at line 142 of program estimation.R

Memory and Runtime Requirements

Summary Approximate time needed to reproduce the analyses on a standard 2022 desktop machine:

☑ Not feasible to run on a desktop machine, as described below.

Details Portions of the code were last run on a **18 cores - Xeon Gold 6140 CPUs@2.3 GHz with 192 GB RAM, 200 GB SSD local disk**. The total amount of time computation is around 2000 hours.

Description of programs/code

The programs require to set the variable `path` at Line 1 equal to the containing folder of the file program. Each program contains two distinct parameter settings: a "test setting" for code verification and a "paper setting" to reproduce the findings of the research paper. The purpose of the “test setting” is simply to demonstrate that the programs can run successfully. It does so by lowering the number of replications and value sof u at which the results are computed. Instead, the “paper setting”, now selected by default, is designed to reproduce the actual results of the paper. The "test setting" can be completed in under 60 minutes on a typical desktop computer from 2022, while running sucessfully the code

in “paper setting” requires a supercomputer. Line 3 of each program initializes a variable named `test`. When `test` is set to TRUE, the "test setting" runs, and when `test` is set to FALSE, the "paper setting" runs. By default, `test` is FALSE.

- Programs in `simulation/script.R` will replicate the simulation results of the proposed estimator of Table 2. The program obtains a final output called `simulations.csv`.
- Programs in `simulation/cqr.R` will replicate the simulation results of the alternative estimator of Wang and Wang (2009) in Table 2. The program obtains a final output called `cqr_results.csv`
- The file `empirical/cqr_estimation.R` produces estimation points which are later utilized in `empirical/estimation.R`. The resulting output, named `cqr.csv`, is generated by the program.
- Programs in `empirical/estimation.R` will replicate the empirical estimation results of the proposed estimator of Table 1 of the supplementary material. The algorithm uses the estimation included in `empirical/cqr.csv`. The program obtains an intermediate output called `raw_simulations.csv`, which consists of all the estimations, and a final output called `estimation.csv`, corresponding to the aimed result.
- Programs in `empirical/bootstrap.R` will replicate the bootstrap procedure for the estimation of confidence intervals for the proposed estimator of Table 1 of the supplementary material. The bootstrap algorithm procedure uses the estimation results included in `empirical/estimation.csv`. The program obtains a final output called `boostrap_results.csv` and corresponds to Table 1 of the supplementary material.
- Programs in `empirical/clean_data.R` will generate the dataset used for the empirical application (`empirical/data.csv`) from The National JTPA Study raw data.

Description of Dataset

The dataset `empirical/data.csv` contains the following variables.

- *age*: integer variable indicating the age of the subject at the start of the experiment;
- *treatment*: boolean variable indicating the treatment status;
- *jtpa*: boolean variable indicating the participation in JTPA programs;
- *delta*: boolean variable indicating the censoring of the unemployment duration;
- *days*: integer variable indicating the unemployment duration.

License for Code

Instructions to Replicators

For `simulation/script.R`, `simulation/cqr.R`, `empirical/cqr_estimation.R` `empirical/estimation.R`, and `empirical/bootstrap.R`, the same following procedure applies.

- Set the value of the variable `test` on line 3 equal to `TRUE` or `FALSE`, depending on the intent of the replicator.
- Run the program.

In order to recreate the dataset `empirical/data.csv`, the following procedure applies.

- Set the path of the subfolder `/empirical/` on line 1 of `empirical/clean_data.R`
- Run the program.

References

U.S. Department of Labor, 1993. *The National Job Training Partnership Act Study: Dataset.* W.E. Upjohn Institute for Employment Research. https://www.upjohn.org/data-tools/employment-research-data-center/national-jtpa-study