

Robust specification testing for rank-based linear regression: README for the replication package

Overview

This replication package is intended to reproduce the main numerical results in “Robust specification testing for rank-based linear regression”.

The replication package contains three folders: code, data, and output.

- The code folder includes the scripts:
0-setup.R,
1-runtables.R,
2-empirical.R,
simulation-table1.R,
simulation-table2.R,
specificationtest_lm.R,
specificationtest_robust.R.
- The data folder contains the empirical dataset Salary Data.csv.
- The output folder is empty and is intended to store the reproduced result files and figures.

Two main files (1-runtables.R and 2-empirical.R) reproduce the main numerical outputs reported in the paper.

Data Availability and Provenance Statements

This paper uses both author-generated simulation data and one public-use empirical dataset.

- The simulation results are generated by code.
- The empirical application uses a salary dataset hosted on Kaggle (Sukumar et al., 2023).
Data can be downloaded from
<https://www.kaggle.com/datasets/mohithsairamreddy/salary-data>, keeps four visibly aberrant observations, and randomly subsamples to $n = 100$ observations for the analysis.
In this replication package, the empirical data file is stored as data/Salary Data.csv.

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.
- ☐ Some data **cannot be made** publicly available.
- ☐ **No data can be made** publicly available.

Dataset list

Data file	Source	Notes	Provided
data/Salary_Data.csv	https://www.kaggle.com/datasets/mohithsairam/reddy/salary-data	Empirical source data for Table 3 and Figure 1	Yes

Computational requirements

Software Requirements

- R 4.4.1
 - extraDistr (1.10.0)
 - parallel (4.3.1)
 - MASS (7.3-60)
 - Rfit (0.27.0)
 - stats (4.3.1)
 - methods (4.3.1)
 - quantreg (5.97)
 - Rstudioapi (0.18.0)
 - the file “0-setup.R” will install all dependencies (latest version), and should be run once prior to running other programs.

Controlled Randomness

- To ensure consistency in data generation, the random seed was set to 888.
- For the parallel simulation, the number of cores was fixed at 9, with 1234 as the seed for the parallel random-number generator.

Memory and Runtime Requirements

Summary

Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine:

- ☐ <10 minutes
- ☐ 10-60 minutes
- ☐ 1-2 hours
- ☒ 2-8 hours
- ☐ 8-24 hours
- ☐ 1-3 days
- ☐ 3-14 days
- ☐ > 14 days
- ☐ Not feasible to run on a desktop machine, as described below.

Details

The hardware environment for the experiments included a system with an Intel(R) Core(TM) i5-1340P processor (1.9 GHz) and 16-GB RAM.

The system operated on a 64-bit Windows 11 environment.

Description of programs/code

The replication archive has the following structure.

- `code/0-setup.R` installs all dependencies (latest version).
- `code/1-runtables.R` reproduces the simulation results for Tables 1 and 2.
- `code/2-empirical.R` reproduces the empirical application results, including Table 3 and Figure 1.
- `code/simulation-table1.R` contains the simulation code for Table 1.
- `code/simulation-table2.R` contains the simulation code for Table 2.
- `code/specificationtest_lm.R` contains supporting routines for the classical linear-model specification test.
- `code/specificationtest_robust.R` contains supporting routines for the proposed robust rank-based specification test.

Instructions to Replicators

- Run `code/0-setup.R` to install all packages required by the scripts in the code folder.
- Run `code/1-runtables.R` to reproduce the simulation outputs for Tables 1 and 2.

- Run code/2-empirical.R to reproduce the empirical outputs.

List of tables and figures

Figure/Table #	Program file	Output file(s)
Table 1	1-runtables.R	spec_results_table1_KS_n100.csv; spec_results_table1_CvM_n100.csv; spec_results_table1_KS_n200.csv; spec_results_table1_CvM_n200.csv
Table 2	1-runtables.R	spec_results_table2_KS_n100.csv; spec_results_table2_CvM_n100.csv; spec_results_table2_KS_n200.csv; spec_results_table2_CvM_n200.csv
Table 3	2-empirical.R	p-values.csv
Figure 1	2-empirical.R	linear.eps; quadratic.eps; interaction.eps

References

Sukumar, J., M. S. R. Reddy, N. Sambangi, S. Abhishek, et al. (2023). Enhancing salary projections: a supervised machine learning approach with flask deployment. In 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 693–700. IEEE.