Replication package for BLM

This folder contains all the code to replicate the results of **Bonhomme Lamadon and Manresa "A distributional Framework for matched employer-employee data"**, forthcoming at **Econometrica**. The working-paper version is available here. This package is also available online at github:blm-replicate. Virtually all code is based on the R platform.

If you are looking for the R package to use the method of the paper, you should use the rblm package. It includes most of the estimators available here, and we keep updating it.

The present replication package is built as an R package that can be easily installed on any system. All package dependencies can be handled using packrat. This option guarantees that results can be reproduced using the exact versions of all the libraries that were used at the time the paper was written. We also provide a Docker container to ensure full portability. This provides a full linux stack with RStudio and all libraries installed and configured.

Importantly, reproducing the results on Swedish data **requires access to the administrative data from Sweden**. Researchers need to apply to get access to such data. We recommend contacting the IFAU. The institute is hosting this replication package that can be accessed and ran on the data on their servers. The reference name for our project is `"IFAU-2015-65 ("dnr65/2015")`. See at the end of this page for more info.

If you have any question or comment, please contact us or use directly the issue page on the github repository.

How do I run this?

In R, run the following commands:

```
# installing the package locally in your R env.
# make sure you are running this from within the package folder

install.packages("pakcrat") # make sure that packrat is available
install.packages("devtools") # make sure that devtools is available

source("packrat/init.R") # initialize the packrat environment
packrat::restore()       # make sure all is up to date

devtools::install(".")   # build the replication package

source("inst/main.R")    # fire up the replication
```

Overview of the replication package

The main entry point is inst/main.r. It will **automatically** run all the necessary steps in the other files in order to reproduce all the results of the paper. Note however that this would take a very long time as it will start some bootstrap procedures. The code will generate all figures and tables and put them into a folder called `tmp` .

We invite resesearchers to read through inst/main.r which has explicit calls for each subsets of the paper.

Organization of the code

- All the heavy lifting such as the estimators and simulation codes are in the R/*.r folder. This is the usual way to store functions in an R package.
- inst/server/estimation-static.r contains the code that runs the estimations for the **static** version of the model
- inst/server/estimation-dynamic.r contains code that runs the different estimations for the **dynamic** version of the model.
- inst/server/fig-blm.R contains functions that generate all of the **figures and tables** in the paper.

Replicating the results on Swedish data

Data availability requirements – requests for replication

From the IFAU:

> Due to strict regulations regarding access to and processing of personal data, the Swedish microdata cannot be uploaded to journal servers. However the IFAU ensures data availability in accordance with requirements by allowing access to researchers who wish to replicate the analyses.
>
> Researchers wishing to perform replication analyses can apply for access to the data. The researcher will be granted remote (or site) access to the data to the extent necessary to perform replication, provided he/she signs a reservation of secrecy. The reservation states the terms of access, most importantly that the data can only be used for the stated purposes (replication), only be accessed from within the EU/EEA, and not transferred to any third party. The authors will be available for consultation.
>
> Apart from allowing access for replication purposes, any researcher can apply to Statistics Sweden to obtain the same data for research projects, subject to their conditions.

Researchers can directly apply for access to `data-static.dat` and `data-dynamic.dat` by contacting us and the IFAU. These two files are the inputs to the replication code and a copy is stored as part of the replication package on the servers at the IFAU. Our two data sets (data-static.dta and data-dynamic.dta) will be stored on a server at IFAU, as part of the project`"IFAU-2015-65 ("dnr65/2015")`. The files will be in a separate folder that can be accessed by anyone who gets clearance from IFAU.

Researchers could also re-construct these data sets from the original files, which are available on a server at IFAU, as part of the project `dnr167/2009` that was put together by Benjamin Friedrich, Lisa Laun, Costas Meghir, and Luigi Pistaferri. This project and ours are linked. The main data source should be the following list of files: `selectedf0educ1.dta`, `selectedf0educ2.dta`, `selectedf0educ3.dta`, `selectedf1educ1.dta`, `selectedf1educ2.dta`, `selectedf1educ3.dta`, `selectedfirms9708.dta`.

The following two scripts use these data sources to construct the two data files `data-static.dat` and `data-dynamic.dat`:

- inst/server/data-section-static.r contains the code that **processes the data inputs** to prepare the data for the static estimation.
- inst/server/data-section-dynamic.r contains the code that **processes the data inputs** to prepare the data for the dynamic estimation.

Using your own data source

This is similar to using the Swedish data. You only need to provide two data sources in the form of a `data.frame`. One should be called `sdata` and contain information on all workers, and one should be called `jdata` and contain information only about the movers. The sdata and jdata frames should be saved into `data-tmp/data-static.dat` and `data-tmp/data-dynamic.dat` for the static and the dynamic estimation.

We recommend to have a look at the function `generate_simulated_data` in inst/server/server-utils.R. It creates synthetic data simulated from our main specifications and saves files to the same format as the actual data. This is your best source to match the structure exactly.