Introduction

This repository includes the replication package for the empirical and numerical results of the paper "Distribution Regression with Sample Selection and UK Wage Decomposition," by Victor Chernozhukov, Ivan Fernandez-Val and Siyi Luo.

Code Files

All the computations were carried out in R (version 3.3.2).

The repository includes 13 files:

1. **DRS_Source.R:** file containing all the auxiliary functions called by other programs.
2. **DRS_Emp_Summary.R:** code that produces plots of descriptive statistics for the covariates.
    - Input: drs_fesdata.csv.
    - Output: plots with descriptive statistics including Inequality.pdf, Summary_Stats_Trend.pdf, and LeavingSchool.pdf.
3. **DRS_Emp_Est.R**: code to generate estimates of coefficients and counterfactual functionals.
    - Input: drs_fesdata.csv.
    - Output: Est_xxx.RData, where xxx refers to the specification (see below).
4. **DRS_Emp_Inf.R:** code to construct confidence bands for the coefficients and counterfactual functions using multiplier bootstrap.
    - Input: Est_xxx.RData file.
    - Output: Results_xxx.RData.
5. **DRS_Emp_Plot.R:** wrapper that combines the graphs of estimates and counterfactuals produced by DRS_Emp_Plot_para.R, DRS_Emp_Plot_DFQF.R, and DRS_Emp_Plot_DFQF.R
    - Input: Results_xxx.RData
    - Output: plots in pdf files: Estimates_xxx.pdf, DFQF_xxx.pdf, CondDFQF_xxx.pdf.
6. **DRS_Emp_Plot_para.R:** generates the graph of the estimates and confidence bands for the correlation coefficient rho and marginal distribution coefficients beta.
7. **DRS_Emp_Plot_DFQF.R:** generates the graph of the estimates and confidence bands for the latent distribution and quantile functions and their decompositions.
8. **DRS_Emp_Plot_Cond.R:** generates the graph of the estimates and confidence bands for the observed distribution and quantile functions and their decompositions.
9. **DRS_Sim_Null.R**: calibrates parameters of the simulation design to the empirical application.
    - Input: drs_fesdata.csv

- Output: Est_female_xxx.RData, which includes the estimates of the coefficients that will be used to generate the data in the simulations.

10. **DRS_Sim_f.R:** generates datasets using the calibrated coefficients and obtains estimates and confidence bands.
    - Input: Est_female_xxx.RData.
    - Output: Sim_f_xxx_n.rds, where n refers to the simulation count.
11. **DRS_Sim_f_Merger.R:** merges simulation results to produce table and figures.
    - Input: Est_female_xxx.RData, Sim_f_xxx_n.rds.
    - Output: Sim_f_xxx.RData, Sim_f_xxx_coef.pdf.
12. **rbatch_est.sh:** program to submit job in batch to computer cluster to produce empirical point estimates
13. **rbatch_inf.sh:** program to submit job in batch to computer cluster to produce confidence bands by multiplier bootstrap

Specification XXX

The specification XXX has 3 components:

1. str.sample: determines the sample used with possible options:
    - For analysis comparing male and female using specific interval of sample period, the specification is _{end year}{start year}, where {end year} and {start year} are the year counts from 1900 (the main results are based on the specification _11378, which means the entire sample period from 1978 to 2013).
    - For analysis comparing first and last decade of sample period within gender, the specification options are _m for male and _f for female (results not included in the paper).
    - For analysis comparing first and second half of sample period within gender, the specification options are _mhalf for male and _fhalf for female.

2. str.rho: determines the specification of the covariates included in the sorting function with 22 possible options:
    - 1 for no covariates.
    - 2 for education indicators (results not included in the paper).
    - 3 for marital status indicators.
    - 4 for year indicators (results not included in the paper).
    - 5 for linear expression of all covariates (results not included in the paper).
    - 6 for age (results not included in the paper).
    - 7 for number of kids in 6 different categories by kid age (results not included in the paper).
    - 8 for years of schooling (results not included in the paper).
    - 9 for 4 education indicators (results not included in the paper).
    - 10 for 4 education indicators interacted with marital status indicators (results not included in the paper).

- 11 for age interacted with marital status indicators (results not included in the paper).
- 12 for 4 age indicators (results not included in the paper).
- 13 for number of kids interacted with marital status indicators (results not included in the paper).
- 14 for 4 age indicators interacted with marital status indicators (results not included in the paper).
- 15 for linear trend.
- 16 for quadratic trend (results not included in the paper).
- 17 for years of schooling interacted with marital status indicators (results not included in the paper).
- 18 for linear trend interacted with marital status indicators (the main results are based on this specification).
- 19 for 6-year indicators (results not included in the paper).
- 20 for propensity score of employment (results not included in the paper).
- 21 for propensity score of employment interacted with marital status indicators (results not included in the paper).
- 22 for instruments in specification str.robust = empty.

3. str.robust: determines the instruments or exclusion restrictions used in the analysis with possible options:

    - Empty for potential out-of-work income interacted with marital status indicators, where the out-of-work income corresponds to the tax unit benefit income when the individual is out of work or the simulated tax unit benefit if the individual were out of work when the individual is working (this specification is used in all the results reported in the paper).
    - _blundell for individual benefit income, simulated by TAXBEN (results not included in the paper). This specification is similar to Blundell et al. (2003).
    - _blundell2 for tax unit benefit income if both individuals were out of work, constructed as the variable in _blundell summed by tax unit (results not included in the paper). This specification is similar to Blundell et al. (2007).
    - _blundell3 for same as_blundell2, except that _blundell2 has one extra step dropping the observations according to whether the default instrument is negative or not (results not included in the paper).

Data

The data used in the empirical analysis come from the U.K. Family Expenditure Survey (FES) for the years 1978 to 2001, Expenditure and Food Survey (EFS) for the years 2002 to 2007, and Living Costs and Food Survey (LCFS) for the years 2008 to 2013, supplemented with variables constructed with the tax and welfare-benefit model (TAXBEN) by the Institute for Fiscal Studies (IFS).

A data replication package is available at 10.5255/UKDA-SN-9355-1 and is accessible upon registration in the UK Data Service. This package includes the file drs_fesdata.csv, which contains the final dataset in Comma Separated Values format used for the empirical analysis and calibration of the numerical simulations.

Replication

To replicate empirical study results, there are 2 options:

1) Run the files DRS_Emp_Summary.R, DRS_Emp_Est.R, DRS_Emp_Inf.R, and DRS_Emp_Plot.R in this order.
2) Submit the files rbatch_est.sh and rbatch_inf.sh to a computer cluster.

To replicate simulation results, run the files DRS_Sim_Null.R, DRS_Sim_f.R and DRS_Sim_f_Merger.R in this order.

Computation time

The replication of the empirical results takes about 10-20 hours in a computer with 16 processors and 16G memory per core for each specification.

References

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): “Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds,” *Econometrica*, 75(2), 323-363.

Blundell, R., H. Reed, and T. M. Stoker (2003): “Interpreting Aggregate Wage Growth: The Role of Labor Market Participation,” *American Economic Review*, 93(4), 1114-1131.