README for: “Firm Heterogeneity in Skill Returns”

Data availability and provenance

The main data used in our paper *“Firm Heterogeneity in Skill Returns”* is from confidential register records provided by Statistics Sweden (SCB). The sources of specific variables are explained in our paper and summarized further below.
There is a standardized procedure to apply for data access with SCB. An application includes a detailed research plan, a list of variables from the different registers, and an approval of the ethical review board. The official instructions can be found at https://www.scb.se/en/services/guidance-for-researchers-and-universities/. The agency’s microdata unit can be contacted at mikrodata@scb.se.

SCB performs a judicial clearance of the order and quotes a price. Once the data has been processed, it can be accessed via Statistics Sweden’s access servers (the system is called Microdata Online Access – MONA). MONA is currently only accessible from EU countries or countries meeting the GDPR requirements of the EU.

Source data list

LISA (SCB administrative data): Statistics Sweden’s “Longitudinal Integration Database for Health Insurance and Labour Market Studies” – LISA. We use data from 1990 to 2017. Information about the data and variables can be found at http://www.scb.se/en_/Services/Guidance-for-researchers-and-universities/SCB-Data/Longitudinal-integration-database-for-health-insurance-and-labour-market-studies-LISA-by-Swedish-acronym/.

Krigsarkivet and Rekryteringsmyndigheten (SCB administrative data): Data from Military Archives and Swedish Defence Recruitment Agency. We use this for skill measures of cohorts enlisted from 1969 to 1983 and 1983 to 2010, respectively. Linked to the other data within MONA via the individual’s unique person number. Documentation of this data can be obtained directly from the respective institutions. The data is also described in detail in Lindqvist and Vestman (2011).

Analysis datasets construction

estimSmpl2020.dta and **estimSmpl1999_2008.dta** as base samples

- All males aged 20-60 in LISA (i.e., all residents of Sweden from age 16 onward) who work dependently employed in the non-primary private sector and earn above the basic taxation value (prisbasbelopp) in the respective calendar year. They also work in firms that exist for at least 5 years and have at least 10 employees on average, and can be matched with cognitive and noncognitive skills from the military data.
- Output: estimSmpl2020.dta as annual panel data during 1990-2017, estimSmpl1999_2008.dta as 1999-2008 subsample.

estim_norm_prd2.dta for grouped analyses

- finalsample.do (line25ff): Use estimSmpl2020.dta as input and estimate the 100 group-level returns.
- Output: Annual panel data during 1999–2008 in Stata's .dta format.

/full/XX.txt for firm-level analyses

1. finalsample.do (line134ff): Use estimSmpl1999_2008.dta and split into exclusive and exhaustive skill types S={0,1,2}. Export as network-full-sS.txt to create leave-out-connected subgraph in python.
2. networkconstruction.py: find the leave-out-connected subgraph in each skill type and save as result-full-sS.txt.
3. finalsample.do (line162ff): keep only the intersection of result-full-sS.txt for all S={0,1,2}.

- Iterate over steps 1.-3. until we have dual-connected leave-one-out connected set (i.e., {0,1,2} firm sets coincide).
- finalsample.do (line213ff): collapse the final connected data to the match level for main analyses.
- Output: txt files with /full/XX = worker_ID, firm_ID, time_ID, wage, worker_C, worker_CID, worker_N, worker_NID, age_ID.

Final figures and tables

Table 1: Standard deviations of firm parameters: estimates from clustered sample and from firm-level sample with quadratic-form correction.

- groupedanalysis.do (line77ff): compute grouped estimates for columns (1) and (3). Uses as data inputs /out/estim_norm_prd2.dta.
- firmestimation.py: compute firm-level estimates for columns (2) and (4). Uses as data inputs /full/XX.txt.
- Output: hand-formatted table in latex.

Table 2: Projection of average skills onto grouped returns.

- groupedanalysis.do (line107ff): compute projections of skills onto grouped returns. Uses as data inputs /out/estim_norm_prd2.dta.
- Output: hand-formatted table in latex.

Table 3: Contributions of firm heterogeneity to dispersion and levels of earnings.

- groupedanalysis.do (line77ff): compute grouped estimates for columns (1) and (3). Uses as data inputs estim_norm_prd2.dta.
- firmestimation.py: compute firm-level estimates for columns (2) and (4). Uses as data inputs /full/XX.txt.
- Output: hand-formatted table in latex.

Table 4: Gains from sorting across returns ${{\lambda}}_{j}^{\textup{c}}$ for different cognitive skill levels.

- groupedanalysis.do (line226ff): compute allocation gains/losses separately by skill level in grouped returns. Uses as data inputs /out/estim_norm_prd2.dta.
- Output: hand-formatted table in latex.

Table 5: Moments due to skill returns under random versus actual sorting.

- groupedanalysis.do (line269ff): compute skills' wage distribution contributions under actual and random allocation. Uses as data inputs /out/estim_norm_prd2.dta.
- Output: hand-formatted table in latex.

Figure 1: Firm effects heterogeneity: cognitive and noncognitive skills.

- groupedanalysis.do (line20ff): build the grouped version of the plot and tests. Uses as data inputs /out/estim_norm_prd2.dta.
- Output (for plotting in Matlab): groupedtest2004_2007-cognitive.csv, groupedtest2004_2007-noncognitive.csv.

Figure 2: Distributions of firm returns for different sets of worker skills.

- groupedanalysis.do (line155ff): compute CDF over returns by skills in grouped estimates (Panels A and B). Uses as data inputs /out/estim_norm_prd2.dta.
- groupedanalysis.do (line107ff): compute average returns against average skills in grouped estimates (Panels C and D). Uses as data inputs /out/estim_norm_prd2..dta.
- Output (for plotting in Matlab): cdfskill_lamC_midN.csv, cdfskill_lamN_midC.csv, lamC_avgC.csv, lamN_avgN.csv.

Figure 3: Gains from sorting for workers with different cognitive skill ranks.

- Computed by hand from numbers in Table 4 and plotted in Matlab.

Computational details

The code for the analysis of the administrative data was run on Statistics Sweden's secure server MONA, which is a virtual server in VMware with 40 threads, 500 GB of RAM, 500 GB of fast storage.

The software Stata 17MP and Python (Spyder IDE) was used.

References

Lindqvist, Erik, and Roine Vestman. 2011. "*The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment.*" American Economic Journal: Applied Economics, 3 (1): 101-28.

Statistics Sweden (SCB). 2022. "*Longitudinal Integration Database for Health Insurance and Labour Market Studies (LISA), 1990-2017 [database].*" accessed 2020.

Statistics Sweden (SCB). 2022. "*Military Archives (Krigsarkivet), 1969-1983 [database].*" accessed 2020.

Statistics Sweden (SCB). 2022. "*Recruitment Agency (Rekryteringsmyndigheten), 1983-2010 [database].*" accessed 2020.