Directions: To replicate the analysis, execute the original R build scripts **(any file other than ecimergeclean.R)** first; these prepare all intermediate datasets. Once they complete, run **ecimergeclean.R**, which consolidates the outputs and generates finaldata.dta. Load that file in Stata and run the supplied do-file to reproduce the results.

Explanations of intermediary datasets

1. Establishment data (entr.R)

Loads the needed R packages, sets the working directory, reads the eight yearly entYYYY_sub.sas7bdat files (2008-2015), adds a year column, fixes the column name NBPNA_SEXECSR → NBPNA_SEXECS, stacks the datasets into one clean_entr table, converts all variable names to lower-case, and saves the result as clean_entr.rds.

2. Entreprise Data (firmyear.R, for calculating firmage)

Reads the 2010-2015 entreYYYY_sub.sas7bdat files, stacks them, lower-cases column names, for each firm (siren) keeps the first row with a non-missing dat_crea, converts that date to a numeric creation year (creation), drops dat_crea and the original year column, and saves the result as firmage.rds.

3. Panel Tous Salarie Data (for calculating longtenure)

Loads required packages, reads ptsclean.sas7bdat, keeps observations with ANCSIR > 2, (which gives the people with more than 2 years of tenure in that firm) aggregates by firm (SIR) to count (i) all workers with >2-year tenure and (ii) those with ≥4-year tenure, computes their ratio as longtenure, and exports the resulting firm-level table to longtenure.csv.

Final Dataset and .do file

1. Eci data and merging it with other datasets

ecimergeclean.R is the streamlined, fully-commented build script that cleans and merges the ECI survey, BTS-Entreprises, and ancillary files to produce finaldata.dta for the probit analysis. It has been pruned of unused chunks but reproduces every variable listed in the manuscript. Please run this version first. **If you encounter any unexpected errors** (e.g., missing objects created by the old imputation blocks), **simply fall back to ecimerge.R**, the original untrimmed script; it contains the same logic plus the redundant steps and will still generate an identical finaldata.dta.

2. Probit Regressions (probit.do file)

Opens the final analysis file (finaldata.dta) and converts the two-digit industry code from string to numeric. Runs five probit specifications on the binary outcome direc (hiring difficulty) with standard errors clustered at the firm level (siren). After each model, outreg2 appends the results to results.xls. Excessturnover variable is dropped due to multicolinearity.

Short definitions of the variables that are used:

- **direc** – 1 if the establishment reported recruitment difficulty that year, 0 otherwise.
- **i.year** – dummy indicators for each calendar year (capture business-cycle shocks).
- **pexec / pinter / pworker** – shares of executives, intermediate staff, and blue-collar workers in total employment in the establishment, respectively.
- **loglabor** – log of (1 + yearly average employment in an establishment).
- **wageprem** – firm's log wage premium relative to the mean wage in its industry-LMA-year cell.

- **longtenure** – fraction of employees with ≥4 years' tenure among those who have >2 years.
- **direc1 / direc2** – lagged values of *direc* from one and two years earlier (state dependence).
- **netgrowth** – LMA net job growth:
- **loglt** – log of average employment stock in the local industry-LMA (market size).
- **i.industry** – industry fixed-effect dummies (two-digit NAF codes).
- **firmage** – years since the firm's legal creation date.
- **eff** – average headcount at the firm in the year, firm level control for firm size.
- **cl(siren)** – clustered standard errors at the firm (SIREN) level.

On the topic of word count

The main text of our report contains approximately 1,950 words, excluding tables, their footnotes, and references, in line with the word limit guidelines. Word counts were verified by omitting non-text elements from the final PDF.

Programme.sas

We have also used a SAS program in order to select the variables, thus reduce the dimensions of the dataset so that we can do everything faster. It looks like this for example BTS Etab. Dataset ;

```
libname raw "C:/Users/Public/Documents/bekircan gonzalo/entreprisesraw";
libname subset "C:/Users/Public/Documents/bekircan gonzalo/entreprises";

%macro subset_by_year;
        %do year = 2008 %to 2008;
                data subset.ent&year._sub;
                        set raw.ent&year (keep=APET EFF_MOY_ET NBPNA_SEXECSR_13
NBPNA_SEXECSR_14 NBPNA_SEXECSR_16 NBPNA_SEXECSR_23 NBPNA_SEXECSR_24
NBPNA_SEXECSR_26 S_BRUT SIRET ZEMPT);
                run;
        %end;
%mend
%subset_by_year

%macro subset_by_year2;
        %do year = 2013 %to 2015;
                data subset.ent&year._sub;
                        set raw.ent&year (keep=APET EFF_MOY_ET NBPNA_SEXECS_13
NBPNA_SEXECS_14 NBPNA_SEXECS_16 NBPNA_SEXECS_23 NBPNA_SEXECS_24
NBPNA_SEXECS_26 S_BRUT SIRET ZEMPT);
                run;
        %end;
%mend
%subset_by_year2
```