

# Replication Package for Decomposing Duration Dependence in a Stopping Time Model

Fernando Alvarez                      Katarína Borovičková  
University of Chicago              Federal Reserve Bank of Richmond

Robert Shimer  
University of Chicago

September 22, 2023

## 1 Content of the Replication Package

The replication package includes

1. Stata .do files which transform raw data into format used by the main estimation routine
2. Matlab .m files which use prepared data files to estimate the distribution of worker types
3. Processed data files which are called in .m files

Detailed description of each of these items follows.

## 2 Data and Software

We use panel data set from the Austrian social security registry (Zweimuller, Winter-Ebmer, Lalive, Kuhn, Wuellrich, Ruf and Buchi, 2009), the Austrian Social Security Database (ASSD). The ASSD covers the universe of workers in the private sector from 1972 to 2007. In our analysis, we use data 1986 – 2007 because ASSD has spells of registered unemployment

only starting from 1986. This data set is confidential, we cannot share it and we do not know how researchers can gain access to it.

There exists another data set coming from the Austrian social security records called Arbeitsmarktdatenbank (AMDB, Labor Market Database). AMDB starts in 1986 and covers workers until nowadays. This data set is confidential but an interested researcher can apply to gain access to it. The application and approval process is handled by Public Employment Service Austria (Arbeitsmarktservices Österreich, AMS), website:

<https://arbeitsmarktdatenbank.at/>.

Both ASSD and AMDB are based on social security records, but they are not identical. For example, we counted workers employed on a randomly chosen day, May 25, in year  $y$  for  $y = 1986, \dots 2007$  by age and gender, and find a mean absolute difference of 1% on average for men and 2% for women. These differences tend to be bigger for younger workers. We also counted the number of workers registered as unemployed collecting benefits on May 25 in year  $y$ , for  $y = 1986, \dots 2007$ , by age and gender, and find identical numbers. We conclude that these data set are very close but not identical, and so estimating our model on AMDB will not result in exactly the same estimates. In the data description that follows we list the name of the data source and relevant variable not only in ASSD, but also in AMDB.

We use ASSD rather than AMDB in this project because it contains detailed information on education for workers who ever registered as unemployed, information that is not provided in AMDB. Hence, it is not possible to replicate the top left panel of Figure 11, where we use education to conduct the hazard rate decomposition with observable characteristics.

### 3 Description of Data Files

There are four original ASSD data files used in the sample construction, described in Table 1. These files are NOT INCLUDED in the replication package because we are not allowed to share them.

Table 1 contains the name of the data sets as it is used in the provided Stata code, as well as the name of the corresponding AMDB file. For each data set, we list names of variables used in the provided Stata code, the description of the variable, and the name of the corresponding variable in AMDB.

Variable `lfstatus` is created from variable `ins_type` and it categorizes insurance episodes into 5 labor market categories: 1) employment, 2) registered unemployment, 3) maternity, 4) pension, 5) other. Variable `educ_cat` is created from variable `educ` and categorizes education into five categories: 1) compulsory schooling, 2) middle school, 3) technical secondary school, 4) academic secondary school, 5) college. Education can vary by year. Variables `educ_max`

| Name in Stata files                                     | Name in AMDB | Description                    |
|---|--------------|--------------------------------|
| Dataset epi-all.dta (AMDB name HV-EPI-UNI-ROH)          |              |                                |
| pid   | PENR         | worker identifier              |
| fid   | BENR         | firm identifier                |
| ins_type  | AMP          | insurance type                 |
| lfstatus  | —            | labor force status             |
| bdate   | A_DAT        | starting date of the episode   |
| edate   | E_DAT        | ending date of the episode     |
| Dataset pid-sex-birthyear.dta (AMDB name HV-PN)         |              |                                |
| pid   | PENR         | personal identification number |
| sex   | GESL         | sex at birth                   |
| birthyear   | GEBJ         | year of birth                  |
| Dataset fid-industry-region.dta (AMDB name HV-DG-KONTO) |              |                                |
| fid   | BENR         | firm identifier                |
| industry  | NACE         | industry identifier            |
| fid_nuts  | NUTS         | location code                  |
| Dataset educ-pid.dta                                    |              |                                |
| pid   | —            | worker identifier              |
| year  | —            | year                           |
| educ  | —            | education                      |
| educ_cat  | —            | education category             |
| educ_max  | —            | highest achieved education     |

Table 1: List of main data sets and variables used in attached Stata files.

is the highest education a worker has achieved.

As the `epi-all.dta` is large, two smaller files were created, both of which are called in the provided Stata code.

- `unemployed_days_VA.dta` contains only registered unemployment spells from `epi-all.dta`, that is, those with `lfstatus=2`
- `epi_134_nooverlap_woagerestrict.dta` contains only spells from `epi-all.dta` corresponding to labor force status of employment, maternity and pension (`lfstatus=1,3,4`). Overlapping spells corresponding to the same labor force status were merged into one spell.

All other files are created within the provided code and their names start with “RP”.

## 4 Description of Codes

We use two softwares in this project: Stata 17 and Matlab R2022a. We use Stata to construct joint distribution of non-employment spells from ASSD. We use Matlab to run the main estimation routine.

### 4.1 Description of .do Files

The replication package contains two .do files:

- `RP_sample_construction.do`
- `RP_sample_observable_characteristics.do`

The file `RP_sample_construction.do` constructs a file containing joint distribution of two non-employment spells in relevant population for  $\bar{T} = 104$ , `IG_raw_0_105_RP.txt`. This file contains 3 columns: column 1 is the duration of the first spell  $t_1$ , column 2 is the duration of the second spell  $t_2$ , and column 3 is the number of workers with such durations. Durations  $t_1$  and  $t_2$  between 0 and 104 weeks correspond completed spells. Duration of  $t_2 = 105$  indicates right-censored spell, that is, a spell which is at least 105 weeks long. Finally there are three special entries in this file which provide information on incomplete spells. The entry with  $t_2 = 300$  corresponds to the number of people with long first non-employment spell and short uncensored employment spell (entry B.1 in the Table 2 in the paper);  $t_2 = 400$  corresponds to long first non-employment spell censored employment spell and very long first non-employment spell (entry B.3a in Table 2 in the paper);  $t_2 = 500$

| name                                   | description   |
|--|---|
| <code>pid_new</code>                   | worker identifier   |
| <code>duration1,duration2</code>       | duration of first and second spell                                  |
| <code>age1,age2</code>                 | age at the beginning of first and second spell                      |
| <code>recall1,recall2</code>           | dummy for returning to the same employer                            |
| <code>NACE101,NACE102</code>           | 1-digit industry code for the first and second spell                |
| <code>region1,region2</code>           | region for the first and second spell                               |
| <code>precall_OLS1,precall_OLS2</code> | recall prob. for the first and second spell, estimated using OLS    |
| <code>precall_LOG1,precall_LOG2</code> | recall prob. for the first and second spell, estimated using LOGIT  |
| <code>precall_PRO1,precall_PRO2</code> | recall prob. for the first and second spell, estimated using PROBIT |
| <code>A1,B1,B3a,B4</code>              | equals 1 if worker belongs to A.1, B.1, B.3a, B.4 in Table 2        |
| <code>educ</code>                      | education code (1–5)  |
| <code>male</code>                      | dummy variable for being a male                                     |

Table 2: List of variables in the file with observable characteristics.

corresponds to long first non-employment spell and no employment spell (entry B.4 in Table 2 in the paper).

This code also constructs file `IG_raw_0_261_RP.txt` for  $\bar{T} = 260$ , analogous to `IG_raw_0_105_RP.txt`. Durations between 0 and 260 weeks correspond to completed spells, duration of 261 weeks represents spells lasting at least 261 weeks. The codes  $t_2 = 300$ ,  $t_2 = 400$ ,  $t_2 = 500$  have the same meaning as in `IG_raw_0_105_RP.txt`.

The code `RP_sample_observable_characteristics.do` creates a file with observable characteristics, `IG_raw_0_105_RP_with_observables.csv`. The structure of this file is described in Table 2. We use this file to conduct hazard rate decomposition with observable characteristics.

## 4.2 Description of .m Files

The replication package contains the main file called `main_file.m` and several subroutines which are called in the main file. The code `main_file.m` produces Tables 1 and 3, and all Figures 1–14 in the paper.

The structure of `main_file.m` is as follows. The beginning of the code sets parameters for the minimum-distance (MD) and expectation-maximization (EM) algorithms. After that, the code does the following:

- Main estimation on the sample with  $\bar{T} = 104$  weeks. This takes around 10 minutes.
- Bootstrap. We draw  $N_s = 500$  random samples with replacement and estimate the model for each sample. This part takes 26 hours on our computer.

- Estimation on the sample with  $\bar{T} = 260$  weeks. This takes 3 hours on our computer.
- Estimation on artificial sample to check accuracy of the estimation method. This takes less than 10 minutes.
- Producing tables and figures. This takes less than 10 minutes.

## 5 How To Use These Files

The replication package does not contain original ASSD data and hence it is not possible to run the provided Stata codes, unless one is able to gain access to these data.

The replication package contains processed data, namely the distribution of two non-employment spells for  $\bar{T} = 104$  and  $\bar{T} = 260$ , as well as the file of workers with observable characteristics. In particular, these data files are:

- `IG_raw_0_105_RP.txt`
- `IG_raw_0_261_RP.txt`
- `IG_raw_0_105_RP_with_observables.csv`

To estimate the model, one should run `main_file.m` in Matlab. This code produces Tables 1 and 3, and Figures 1–14 in the paper. More detailed description of `main_file.m`, together with estimated times to finish it, is provided in Section 4.2.

## References

**Zweimuller, Josef, Rudolf Winter-Ebmer, Rafael Lalive, Andreas Kuhn, Jean-Philippe Wuellrich, Oliver Ruf, and Simon Buchi**, “Austrian Social Security Database,” April 2009. Mimeo.