May 2022

README FILE FOR:

Language Training and Refugees' Integration

Mette Foged, Linea Hasager, Giovanni Peri, Jacob Nielsen Arendt and Iben Bolvig

Contents

1 Data 2

1.1 Data access 2

1.2 Data sources 2

1.3 Variable definitions 2

1.4 Software and operating system 3

2 Programs and instructions 4

2.1 Definition of refugee population 4

2.2 Definition of outcome and control variables, and sample selection 4

2.3 Balancing tests 5

2.4 Language treatment, estimated effects on outcomes and robustness tests 6

Contact person for replication files: Linea Hasager, University College London, l.hasager@ucl.ac.uk

1 Data

1.1 Data access

Our paper is based on proprietary administrative data with restricted access through Statistics Denmark. To gain access to Statistics Denmarks data researchers must be affiliated with a Danish authorized research environment.
Please, find more information about the rules as well as the step by step procedures for obtaining access here: https://www.dst.dk/en/TilSalg/Forskningsservice .

1.2 Data sources

We use administrative data from multiple registers to construct our population, background variables and outcomes. Below we list the registers we use. The name in parenthesis is the official name of the register at Statistics Denmark.

- Integrated Database for Labor Market Research (IDA).
- Income Register (*IND*).
- Classification of employment (*AKM*)
- Databases on completion of schooling (*UDDA*).
- Databases on charges and convictions for criminal activities (*KRSI* and *KRAF*).
- Population Register (*BEF*).
- Admission Register (*OPHG*).
- Language Training Database (*Danskuddannelsesdatabasen*) includes information on the mother tongue and the initial assignment to language training track 1, 2 or 3.
- Adult Education (*VEUV*) includes the initial assignment to language training track 1, 2 or 3.
- Registers on internal mobility (*BEFPOP* and *BEFADR*).
- Migration Register (*VNDS*).
- Labor Market Policy Measures (*AMFORA*) includes the services received in the municipality of settlement.

VNDS is only used for sample selection and AMFORA is only used to look at impact on hours of employment support in the Online Appendix. The variables used from the remaining registers are described in the next section.

A complete list of registers available through Statistics Denmark can be found here: https://www.dst.dk/extranet/forskningvariabellister/Oversigt\%20over\%20registre.html.

1.3 Variable definitions

Below we list the variables used in the paper, their definition as well as the name and source/register of the variable at Statistics Denmark. Variable and register names are in *italic*.

- Earnings (*loenmv_13_y*) is the annual gross earnings converted to the price level in 2000 using the Danish consumer price index and converted to USD using an exchange rate of 6.6 DKK/USD. It is based on *loenmv_13* from *IND*.
- Employment (*employment_y*) is the first difference of labor market experience. It is based on *erhver* from *IDAP*.
- Occupations in ISCO groups 1 to 5 are defined as complex (*complex_job_y*). This leaves ISCO groups 6 to 9 as non-complex (*non_complex_job_y*). Both variables are based on *DISCO_ALLE_INDK_13* and *DISCO08_ALLE_INDK_13* from *AKM*.
- Education obtained in Denmark (*udda_dk*) is based on *hfaudd* and *hf_kilde* from *UDDA*.
- Criminal convictions (*convicted*) is based on *afg_ger7* from *KRAF*.
- Criminal charges (*charged*) is based on *sig_ger7* from *KRSI*.
- Disposable income (*dispon_13*) is the personal income net of tax and including welfare benefits. It is converted to the price level in 2000 using the Danish consumer price index and converted to USD using an exchange rate of 6.6 DKK/USD. It is based on *dispon_13* from *IND*.
- Left initial municipality of residence (*moved_new*) is based on *kom* from *BEF, BEFPOP* and *BEFADR*.
- Date of admission normalized at January 1, 1999 (*day_nr*) is based on *tilladelsesdato* from *OPHG*.
- Asylum processing time (*waittime*) is based on *ansøgningsdato* og *tilladelsesdato* from *OPHG*.
- Age (*age*) is based on *foed_dag* from *BEF*.
- Married (*married*) is based on *civst* from *BEF*.
- Number of children aged 0-2 years (*kids02*) and number of children aged 3-18 years (*kids018*) are based on *familie_id* and *foed_dag* from *BEF*.
- Origin country dummies (*afghanistan*, *iraq* and *other_country*) are based on *opr_land* from *BEF*.
- Mother tongue dummies (*arabic* and *dari*) are based on *sprogkode* from *Danskuddannelsesdatabasen*.
- Initial placement in Danish language track 1, 2 or 3 (*danishf1*, *danishf2* and *danishf3*) is based on *udd* from *VEUV*.
- Detailed admission categories / refugee "types" (*quota*, *convention* and *other_permit*) are based on *forklar* from *OPHG*.
- Family-reunified to refugee already in Denmark (*familyref*) is based on *familie_id* from *BEF* and *forklar* from *OPHG*.

1.4 Software and operating system

Software: StataMP 16.1, SAS 9.4, and SAS Enterprise Guide 7.1. Operating system: Windows Server 2016 Standard.

2 Programs and instructions

2.1 Definition of refugee population

The raw register data are stored in SAS. Hence, programs drawing directly on the register data are SAS programs, and the remaining program files are in STATA. Find programs and output data described in this section at L:\Workdata\707455\Papers\ABFHP_1999_reform\Code\ReStat.

File: 1. Admission.sas	
Description	Loads admission data and defines first admission category.
Input	ophg1997-2019.sas7bdat
Output	Admission_latest_record.dta

File: 2. Migration.sas	
Description	Loads migration data and defines first registration in Denmark.
Input	vnds.sas7bdat
Output	Immigrations.dta

File: 3. First_immigration_and_admission.do	
Description	Combines individual admission and immigration data.
Input	Admission_latest_record.dta, Immigrations.dta
Output	First_immigration_and_admission_latest_record.dta

File: 4. Refugee population.sas	
Description	Combines admission/migration data with population register (CPR) records on address and movements within Denmark.
Input	First_immigration_and_admission_latest_record.dta, bef1997-2019.sas7bdat, befpop2017.sas7bdat , befadr2017.sas7bdat
Output	Refugees1997_.dta, municipalitykey.dta

2.2 Definition of outcome and control variables, and sample selection

File: 5. Clean DKDB data.sas	
Description	Defines hours of assigned language instruction, class id and individuals' first languages.
Input	sik_silo_indv_uv__99_ku_t.sas7bdat, sik_silo_indv_uv__99_kuho_t.sas7bdat, sik_silo_indv_uv__99_ho_t.sas7bdat, sik_silo_indv_uv__00_ku_t.sas7bdat, sik_silo_indv_uv__00_kuho_t.sas7bdat, sik_silo_indv_uv__00_ho_t.sas7bdat, kurs2001total.sas7bdat-kurs2003total.sas7bdat, kurshold2001total.sas7bdat-kurshold2003total.sas7bdat, hold2001total.sas7bdat-hold2003total.sas7bdat
Output	_dkdb1999_2003.dta

File: 6. Clean first language spoken.do

Description	Defines individual mother tongues for refugees and test whether treated refugees are in the same language training classes as untreated refugees.
Input	_dkdb1999_2003.dta, Refugees1997_.dta
Output	dkdb1999_2003.dta

File: 7. IDAN_AKM isco crosswalk.sas

Description	Creates crosswalk between ISCO08 codes to ISCO88 codes and implements it.
Input	idan1997-2018.sas7bdat, akm1997-2018.sas7bdat (annual registers)
Output	idanakm.sas7bdat

File: 8. Controls and outcomes.sas

Description	Matches the population data to AMFO, BEF, IDAP, IDAN, AKM, UDDA, IND, KRAF, KRSI, DKDB and VEUV to create variables measuring predetermined covariates, annual outcomes and average cumulative outcomes the first 1-18 years after admission (1999 to 2016). Annual employment rates are calculated as first-differenced labor market experience. Annual earnings are gross earnings deflated to 2000-level by the consumer price index (CPI) and converted to USD using an exchange rate of 6.6 DKK/USD.
Input	Refugees1997_.dta, dkdb1999_2003.dta, idanakm.sas7bdat, udda2002-2016.sas7bdat, bef1996-2016.sas7bdat, ind1996-2016.sas7bdat, amfo1997-2006.sas7bdat, idap1995-2016.sas7bdat, kraf1997-2016.sas7bdat, krsi1997-2016.sas7bdat (annual registers), veuv.sas7bdat, opholdstilladelse_asyl_1997_2018.sas7bdat, opholdstilladelse_asyl_2019.sas7bdat, AARD.csv (AARD from statistikbanken.dk), CPI.csv (PRIS112 from statistikbanken.dk)
Output	Population_jan_feb2001.dta, Population_full.dta, Population.dta

File: 9. Sample selection.do

Description	Implements sample selection criteria: Individuals receiving permits 1997-2000. Persons from Somalia are excluded, due to irregularities in asylum processing of this group. Age 18-49 at admission.
Input	Population.dta, municipalitykey.dta
Output	Analysis_sample.dta

2.3 Balancing tests

File: 10. Monthly admissions.do

Description	Creates a histogram of monthly admissions.
Input	Population_jan_feb2001.dta
Output	
Table/graph	Online Appendix Figure 1a

File: 11. Demeaned monthly admissions.do

Description	Creates a plot of the demeaned annual share of admissions and RD estimates of monthly admissions.
Input	Population_full.dta
Output	
Table/graph	Online Appendix Figure 1b, Table 1a

File: 12. Balancing tests.do	
Description	Creates binned scatter plots and RD estimates of predetermined covariates for the population of refugees (balancing tests). Tests of changing placement of refugees occurring later than timing of reform.
Input	Analysis_sample.dta
Output	
Table/graph	Online Appendix Figure 1c, Online Appendix Figure 2, Online Appendix Figure 3, Table 1b, Table 1c, Online Appendix Table 3

2.4 Language treatment, estimated effects on outcomes and robustness tests

File: 13. Language treatment.do	
Description	Estimated impacts on individual participation and assigned hours in language training and employment support.
Input	Analysis_sample.dta
Output	
Table/graph	Online Appendix Table 1

File: 14. Main results.do	
Description	Estimates impact on earnings and employment outcomes using preferred specification, simple difference in means, controlling for a second order polynomial in the running variable and conditioning on individual controls. We also show results on long-run criminal convictions, and we reproduce the main results conditional on municipality fixed effects.
Input	Analysis_sample.dta
Output	
Table/graph	Table 2, Online Appendix Table 4, Online Appendix Table 5

File: 15. RD plots and dynamics.do	
Description	Creates binned scatter RD plots and RD estimates of outcome variables for the population of refugees showing the dynamics of treatment effects.
Input	Analysis_sample.dta
Output	
Table/graph	Online Appendix Figure 4, Figure 1, Table 4

File: 16. Heterogeneity analysis.do

Description	Performs heterogeneity analysis for main results by alphabet of mother tongue, initial placement and gender. The bandwidths for estimation by subgroups are fixed at the corresponding bandwidths for the full sample.
Input	Analysis_sample.dta
Output	
Table/graph	Table 3

File: 17. Robustness bandwidth.do	
Description	Robustness tests of main results using bandwidths between 30-360 days for estimation.
Input	Analysis_sample.dta
Output	
Table/graph	Online Appendix Figure 5

File: 18. Robustness placebo.do	
Description	Robustness tests of main results using fictive (placebo) cutoff points. For placebo cutoffs before January 1st, 1999 we only use data for admissions in 1998-1999. For placebo cutoffs after January 1st, 1999 we only use data for admissions from 1999-2000
Input	Analysis_sample.dta
Output	
Table/graph	Online Appendix Figure 6

File: 19. Attrition.do	
Description	Tests of whether attrition rates differ by treatment status and whether individual characteristics of attriters differ across treatment status.
Input	Analysis_sample.dta
Output	
Table/graph	Online Appendix Table 2