The Role of Own-Group Density and Local Social Norms for Ethnic Marital Sorting: Evidence from the UK

REPLICATION FILES – Readme

List of Included Files

Stata .do files
1. masterfile.do
2. reduced_forms.do
3. predictedfuturesorting.do
4. Homogamy_index.do
5. Spec_tests.do

Stata .dta files
1. eth01q1.dta
2. qual15key.dta
3. qualkey04.dta
4. qualkey05plus.dta
5. qualkey08plus.dta
6. qualkey11plus.dta
7. qualkey15plus.dta
8. qualkey93_96.dta
9. qualkey96plus.dta

Matlab .m files
1. choices.m
2. complementarities.m
3. FIG4.m
4. FIG6.m
5. FIG7.m
6. FIG8.m
7. internalpredict.m
8. loglik_spec1.m
9. loglik_spec2.m
10. loglik_spec3.m
11. setup.m
12. spec1.m
13. spec2.m
14. spec3.m
15. TAB4.m

SECTION 1 – Data Cleaning (Stata)

All LFS data used in this project is freely available from the UK data service repository at the following link:
https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000026#!/access-data

Storing the data
We use data for survey years 1996-2015. All survey waves need to be downloaded and saved in the users 'indata' directory, this directory is where the raw data files will be stored. Within the 'indata' directory we recommend creating additional sub-directories for each survey wave according to the following syntax: yr`z'Q`j'
where `z' corresponds to the survey year i.e. survey year 1999 -> z = 99. And `j' corresponds to the survey quarter i.g. quarter 1 -> j = 1.
Within each sub-directory we recommend saving the corresponding survey wave according to the following syntax: qlfsjm`z'.dta, qlfsaj`z'.dta, qlfsjs`z'.dta, and qlfsod`z'.dta.
Here 'jm' corresponds to January-March i.e. Q1. etc …

IMPORTANT – "masterfile.do" must be executed before executing any other .do file

Masterfile.do
This .do file is our main script for cleaning the data and creating marriage matrices that will be exported to Matlab for the model estimation.
We recommend that the user creates additional directories that correspond to the Global variables defined in lines 30-37 of the .do file "Masterfile.do" before proceeding.

Section 0 of the .do file is for programme definitions. These are programmes that will be subsequently executed on each survey wave in order to clean the data.

Section 1 applies the pre-defined programmes to each survey wave, keeping only the relevant variables and then saves each cleaned wave as a temporary dataset. In order for this to execute it is required to have the qualification and ethnicity merge files stored in the correct directory. Specifically, all qualifications files (provided) need to be stored in a folder 'qualpath' and the ethnicity merge file 'eth01q1.dta' needs to be stored in the 'tempdata' folder. The merge files are user created according to LFS mapping documents; we require the use of the merge .dta files in order to create a consistent definition of qualifications across all survey waves.

Section 2 loops over survey waves creating a dataset of men that is subsequently merged with a created dataset of potential female partners. It also creates a dataset of women that is subsequently merged with a dataset of potential male partners. Section 2 concludes by appending the male and female datasets together into a single 'matched' dataset for each survey wave.

Section 3 appends the matched datasets for each survey wave together into a single dataset.

Sections 4 and 5 inspect and clean the data respectively.

Section 6 uses the cleaned data to create matrices, by region, that are exported as .txt files to be used in matlab for the model estimation. Specifically, the .do file creates matrices for:

- Population distribution over ethnicity and education, by gender – Qualprereg`r'.txt
- Population counts by ethnicity education and gender – cnt_prereg`r'.txt
- Male marriage rates by ethnicity and education – M_pre`r'.txt
- Female marriage rates by ethnicity and education – F_pre`r'.txt
- Male and female singles rates – s_m`r'.txt, s_f`r'.txt
- Male and female marriage rates to partners born outside of the UK – o_m`r'.txt, o_f`r'.txt

Table 1: The population counts shown in Table 1 are created in 1882-1910 of Masterfile.do.

Figure 1: The data used in Figure 1 is generated at this stage by summarising the ethnicity dummies by region. (eth and govtof respectively in the cleaned data).

Reduced_forms.do
This .do file creates the data inputs for Table 2: The effect of the own ethnicity share on the probability of being intra-ethnically Married. The file requires the data set "completesample_regs.dta" to execute. This dataset is created by executing the first .do file "Masterfile.do".

Section 2 – Estimation (Matlab)

IMPORTANT – "setup.m" must be executed before executing any other script file

setup.m
Before estimating the model, the user is required to run the matlab script 'setup.m'. This is a script that takes the .txt files created in section 1 and imports them into Matlab. The resulting data will be stored in a structure called "dat" and will be saved on the user's path in a file named "data.mat".

choices.m
once the script "setup.m" has been executed and the file "data.mat" has been created, the user can now generate the data used to create Figures 2 and 3. This can be achieved by running the script file "choices.m". The script uses the data counts and marriage rates to derive aggregate partner choices and aggregate marital status.

FIG4.m
The user can execute the script file "FIG4.m" to create figure 4. The figure shows the proportion of intra-ethnic marriages by region and gender.

SPECIFICATION TESTS – FIGURE 5 & TABLE C.1 (Stata)
Spec_tests.do

Executing this .do file allows the user to replicate the results of the specification tests shown in Table C.1.

Homogamy_index.do
Executing this .do file allows the user to replicate the results of the specification tests shown in Figure 5.

ESTIMATION (Matlab)

spec1.m (loglik_spec1.m)
The script "spec1.m" is used to estimate specification 1 shown in table A.1 (the model without region effects parameters and without peer effect parameters). The script loads in the data "data.mat" and estimates the model by using the unconstrained minimization solver on the likelihood function "loglik_spec1.m". It is therefore crucial that the function script is saved in the same path as the script "spec1.m". All parameter estimates and standard errors are stored in a structure "Spec1", the predicted marriage rates, singles rates etc .. are stored in a structure "pred1".

spec2.m (loglik_spec2.m)
The script "spec2.m" is used to estimate specification 2 shown in table A.1 (the model with region effects parameters and without peer effect parameters). The script loads in the data "data.mat" and estimates the model by using the unconstrained minimization solver on the likelihood function "loglik_spec2.m. All parameter estimates and standard errors are stored in a structure "Spec2", the predicted marriage rates, singles rates etc .. are stored in a structure "pred2".
The script also performs a likelihood ratio test to highlight how the model fit is significantly improved with the inclusion of region effects parameters.

spec3.m (loglik_spec3.m)
The script "spec3.m" is used to estimate our preferred specification 3 shown in table A.1 (the model with region effects parameters and with peer effect parameters). The script loads in the data "data.mat" and estimates the model by using the unconstrained minimization solver on the likelihood function "loglik_spec3.m. All parameter estimates and standard errors are stored in a structure "Spec3", the predicted marriage rates, singles rates etc .. are stored in a structure "pred3".
The script also performs a likelihood ratio test to highlight how the model fit is significantly improved with the inclusion of region effects parameters and peer effects parameters.

PARAMETER ESTIMATES
All parameter estimates and standard errors included in tables A.2 and A.3 correspond to the values stored in the structure "Spec3".

MODEL FIT

FIG6.m
We demonstrate the model fit by plotting the observed and predicted proportions of intra-ethnic marriages against own ethnic shares across regions by ethnicity and gender. The user can execute the script "FIG6.m" to generate these figures. Specifically, the code will only

generate however one sub-figure i.e. Asian women. The user can easily generate the other sub-figures by changing the input in line 121 of the script.

TABLES AND FIGURES

complementarities.m
The script "complementarities.m" is used to generate the estimated complementarities and the corresponding standard errors shown in Table 3. The script uses the data files "data.mat" and "Spec3.mat" as inputs to estimate complementarities with respect to ethnicity and with respect to education.

FIG7.m
The script "FIG7.m" is used to generate figure 7. It shows the principal marriage utility and the distribution of social norms by marriage Type-Profile.

Section 3 – Predicted Future Ethnic Homogamy (Stata & Matlab)

predictedfuturesorting.do (Stata)
For section 3 the user is first required to run the Stata .do file "predictedfuturesorting.do". This file begins by loading in the sample used for the estimation. As many of the individuals included in the prediction cohorts had not completed their education by the time they were observed, we impute the proportion with a high qualification by gender, region, and ethnicity in these recent cohorts. To do this the .do file estimates a regression model for holding a high qualification (A-level+) using the estimation cohort sample observed in the QLFS 1996-2015, born in the UK and aged 25 or higher when observed. The results from the regression in turn generate the imputed distribution of ethnicity and qualifications in the estimating and cohorts shown in Table F.1 and presented in Figure F.1.

The .do file finally creates vectors by region for:
- The type distribution by gender, ethnicity and education – "Qualprereg`r'.txt"
- The population counts by gender, ethnicity, and education – "cnt_prereg`r'.txt"

The .txt files directly correspond to the .txt files created for the estimation cohorts in section 1. We recommend the user ensures that the .txt files are saved in the directory that contains the Matlab script "counterfactuals.m".

counterfactuals.m (Matlab)
This Matlab script is used to perform the exercise of predicting future ethnic homogamy. The script begins by loading in the .txt files for the prediction cohorts. It then creates predictions for future marriage patterns when social norms terms are contemporaneous, and when they are fixed at the level in the estimation cohorts. The predictions for each case are saved, respectively, in structures "predcount1" and "predcount2". The script file will then save, for each ethnicity-gender combination, a .mat file which can be used to produce figure 8. For example, the file "wmcount.mat" is a data file that includes: the proportion of white individuals by region and the proportion of intra-ethnic marriages for white males, by region, in each of the two prediction scenarios.

We recommend that the user save the created .mat files in the directory that includes the matlab scripts "FIG6.m", "FIG8.m".

TAB4.m (Matlab)

This script is used to create the data for table 4. As table 4 includes both the prediction and the estimation cohorts it is advised to complete all previous steps before running this script file. The script is set up for the estimation cohorts. To replicate for the prediction cohorts the user has two options. The first is to save the corresponding data that is created when running the script "counterfactuals.m" and load the data into this script file. The second option is to execute the code in lines 404-454 of "counterfactuals.m" as this creates the values in table 4 for the estimation cohorts.

FIG8.m (Matlab)

This script creates the sub-figures included in figure 8. Specifically, the code will generate one sub-figure i.e. Asian women. The user can easily generate the other sub-figures by changing the inputs in lines 131, 140, and 141 of the script.

Section 4 – Robustness Checks and Additional Tables and Figures (Stata & Matlab)

internalpredict.m - Table A.4 (Matlab)

If the user would like to replicate the results for the robustness check in table A.4 - Predicted Intra-Ethnic Marriage Rates for the West Midlands Region when Included and Excluded from the Estimation, it can be achieved in a few simple steps. The first is to run the script "spec3.m" again, but excluding the West Midlands from the estimation (the west midlands is region 7 in the data). Doing so will generate a new set of parameters which the user should save e.g. "Spec3woWM.mat". These are the parameter estimates shown in table E.3. Then the user can open and run the script "internalpredict.m" which uses the newly estimated set of parameters for the model that excludes the West Midlands. The script uses the estimated parameters to predict marriage rates for the West Midlands, it then compares these with the predicted marriage rates for the West Midlands that were obtained using the full model. The marriage rates for both of these cases are then used to create the values seen in Table A.4.

Figure E.1 (Matlab)

To replicate both sub-figures of figure E.1 the user can use the script "choices.m" but instead change the input from the marriage rates that are observed in the data to the predicted marriage rates i.e. "Spec3.mat".

Table E.1 and E.2 (Matlab & Stata)

The user can generate the parameter estimates in tables E.1 and E.2 by running the stata .do file "Masterfile.do" but instead change the threshold for singlehood (in line 1871 of the .do file) to "age<20". Executing the .do file will create new .txt files for the modified sample that the user can load into matlab by using "setup.m". The user can then recover the

parameters shown in tables E.1 and E.2 by estimating the model for the alternative sample by executing the script "spec3.m" on the new sample.