**DATA AND PROGRAMS FOR "TRADE, INEQUALITY, AND THE ENDOGENOUS SORTING OF HETEROGENEOUS WORKERS"**

Eunhee Lee

March 2020

## Data file and variable description

For the detailed construction of the data, see Appendix Section B of the paper. The data file final_data.mat includes all variables that are necessary to derive the quantitative results of the paper. If a variable was calculated using any information either from the IPUMS-International database or from the UN Comtrade database, the variable only provides the information of the dimension size in this file, because the re-distribution of these databases in any form is not allowed legally. The Appendix Section B explains how to construct those variables from the IPUMS-International database or from the UN Comtrade database.

Most variables have the same size dimensions $(N \times T \times J \times O)$, where $N$ is the number of countries, $T$ is the number of worker types, $J$ is the number of sectors, and $O$ is the number of occupations, in order to facilitate matrix operations. If a certain variable has less size dimensions, then that variable repeated along the irrelevant dimensions. For example, $L_{i,\tau}$ is of $(N \times T)$ in the model, and the variable in the data file for $L_{i,\tau}$ is repeated $(J \times O)$ times. Note that these expansions of variables are done using the repmat function of Matlab.

I explain each variable in the data file below. All data variables are for the base year 2000, otherwise mentioned.

- **B_data:** Together with the variable C_data that will be explained below, this constructs $\pi_{i,\tau}^{j,o}$. In order to make the program compatible for the case with the nested logit specification, I define $\pi_{i,\tau}^{j,o}$ in a more general way.

$$\begin{aligned}\pi_{i,\tau}^{j,o} &= \frac{B_{i,\tau}^{j,o}[\sum_{j'} B_{i,\tau}^{j',o}]^{\delta_\tau^o - 1}}{\sum_{o'}[\sum_{j'} B_{i,\tau}^{j',o'}]^{\delta_\tau^{o'}}} \\ &= \frac{B_{i,\tau}^{j,o}}{\sum_{j'} B_{i,\tau}^{j',o}} \times \frac{[\sum_{j'} B_{i,\tau}^{j',o}]^{\delta_\tau^o}}{\sum_{o'}[\sum_{j'} B_{i,\tau}^{j',o'}]^{\delta_\tau^{o'}}} \equiv C_{i,\tau}^{j,o} \times D_{i,\tau}^{o},\end{aligned}$$

where $B_{i,\tau}^{j,o} \equiv T_{i,\tau}^{j,o}(p_i^{j,o})^{\frac{\theta_{i,\tau}}{\delta_\tau^o}}$, $C_{i,\tau}^{j,o} \equiv \frac{B_{i,\tau}^{j,o}}{\sum_{j'} B_{i,\tau}^{j',o}}$ and $D_{i,\tau}^{o} \equiv \frac{[\sum_{j'} B_{i,\tau}^{j',o}]^{\delta_\tau^o}}{\sum_{o'}[\sum_{j'} B_{i,\tau}^{j',o'}]^{\delta_\tau^{o'}}}$. If $\delta_\tau^o = 1$ for all $\tau$ and $o$, this corresponds to $\pi_{i,\tau}^{j,o}$ defined in equation (3). This variable in the data file is equal to $C_{i,\tau}^{j,o}$,

which is the probability of type $\tau$ workers in country $i$ having occupation $o$ conditional on working in sector $j$.

- **C_data:** This variable is equal to $D^o_{i,\tau}$ which is the share of workers with occupation $o$ within type $\tau$ in country $i$.
- **dissim_est:** The dissimilarity parameter estimated from the nested logit extension of the model using the U.S. data. The estimates are as reported in the Online Appendix. In the baseline specification, this parameter is set to be equal to 1.
- **eta1_baseline:** $\eta_1 = 0.75$
- **gamma_baseline:** $\gamma = 0.9$
- **GDP_data:** Dollar values of each country's gross output, as the baseline model abstracts from intermediate inputs.
- **GDP_data_after:** for year 2007
- **J:** the total number of sectors
- **L_data:** the number of workers within each worker type. The sum for each country is normalized to match the labor income matches GDP.
- **lambda_change_data:** $\hat{\lambda}^j_{in}$ (size dimension: $(N \times N \times T \times J \times O)$ by using repmat of a data of size $(N \times N \times J)$)
- **lambda_data:** $\lambda^j_{in}$ (size dimension: $(N \times N \times T \times J \times O)$ by using repmat of a data of size $(N \times N \times J)$)
- **lambda_ind_data:** $\lambda^j_i$
- **N:** the total number of countries
- **pi_data:** $\pi^{j,o}_{i,\tau}$, this is consistent with the one that is constructed by B_data and C_data as described above.
- **Sigma:** the total number of occupations
- **T:** the total number of worker types
- **theta_trade_baseline:** $\nu^j$
- **w_data:** $w_{i,\tau}$

- **X_data:** $X_i^j$
- **Xbilat_data:** $X_{in}^j$ for 2000 (size dimension: $(N \times N \times T \times J \times O)$ by using repmat of a data of size $(N \times N \times J)$)
- **Xbilat_data_after:** for 2007 (size dimension: $(N \times N \times T \times J \times O)$ by using repmat of a data of size $(N \times N \times J)$)
- **xsi_data:** $\xi_i^{j,o}$

## Matlab programs

All programs for the benchmark model are in the folder named "benchmark", and all programs for the two limit cases of the model are in the folder named "limitcases". Below I explain each file in the benchmark folder. The files in the limitcases folder have the same name as their counterparts in the benchmark folder.

- **main:** The main program that runs all sub-programs for the benchmark model.
- **theta_estimation_MLE:** Using the data file resid_l_wage_data.mat, which is the residual log hourly wage as explained in this paper, this file estimates $\theta_{i,\tau}$ for five countries and five worker types. Since the residual wage variable is from the estimation using variables in the IPUMS-International, it is not included in this package.
- **wageL:** This file calculates empirical likelihood for the ML estimation done in theta_estimation_ML
- **dhat_calibration:** This file calibrates $\hat{d}_{in}^j$.
- **dhat_summary_stats:** Using the calibrated $\hat{d}_{in}^j$ from dhat_calibration.m, this file computes summary statistics for Table 4 and Table A1 of the paper.
- **paramCF:** Based on the model specification specified by variables like THETAcase, CORRcase, CFcase, caseE, caseG, and caseN in main.m, this file sets all parameter values and counterfactual shocks. This file in the limitcases folder also specifies each of two limit cases using the variable LIMITcase.
- **phat_solver:** This file solves for $\hat{p}_i^{j,o}$.
- **LMC:** This file calculates excess demands of each occupation in each sector and each country, conditional on a guess of $\hat{p}_i^{j,o}$.
- **CFoutcomes:** For the solution of $\hat{p}_i^{j,o}$ from phat_solver.m, this file computes counterfactual changes in all equilibrium variables of interest.