Data and Code Material for "Entry Barriers, Idiosyncratic Distortions, and the Firm Size Distribution"

Roberto N. Fattal Jaef

December 4, 2020

1 Data Availability

The data for this project is composed of the various firm-level databases, the Penn World Tables for aggregate data, the Doing Business Indicators for the regulation based barriers to entry, and the Small Business Administration for inferring the US's distribution of firm shares across 2 digit industries. Software requirements for replication are specified in section 5

1.1 Penn World Tables

The paper uses data from the Penn World Tables version 9.0 (Feenstra et al. 2015). Data can be downloaded from Zeileis [2019]

The subset of data from Penn World Tables 9.0 utilized in the project is recorded in the following files

1. GDP_percapita_sector_analysis.csv for the real GDP per capita in the year 2014 for the 21 countries in the sample
2. ctfp_pwt.dat for the PPP based measure of TFP in the year 2014 for the 21 countries in the sample
3. pwt90_gdp_kstock.csv allows to compute TFP for the countries for which the Penn World Table's own computation of PPP based TFP is missing. As stated in the manuscript, in countries with missing data the TFP was computed as $\frac{Y}{K^{1/3}L^{2/3}}$, using $rgdpna$ as the real GDP, $rkna$ as the real capital stock, and emp as the labor input.
4. pwt_labshare_2014.csv allows to transform the World Bank's Doing Business Indicator's cost of starting a firm in model equivalent units (see below)

The first first file was created straight from the PWT's excel sheet, defining GDP per capita as $\frac{rgdpna}{pop}$, and filtering countries to keep only the 21 countries under study in the paper. Similarly, the third and forth files are also filtered from the original PWT databse. The second file, in turn, was created from the following R code

- explore_ctfp_pwt.R

1.2 World Bank's Doing Business Indicators

The World Bank's Doing Business Indicators' (World-Bank 2019) cost of starting a firm are reported in the following files:

1. Entry_Costs_sector_analysis.csv reports the costs of starting a firm and the cost of accessing electricity in the year 2014, as proportion of GDP per capita, for the 21 countries in the sample
2. db_indicator_laborunits.txt reports these entry costs, translated into model-equivalent entry costs in units of labor, as explained in the manuscript.

The first file was constructed simply by filtering the full database by country in the year 2014, and selecting the 21 countries under study in the paper. Then, across all doing business metrics, we select the ones for the construction of the costs of starting a formal firm, given by:

- cost of starting a business, as % of income per capita, labeled start_gdp in the replication file
- cost of getting electricity as % of income per capita, labeled electr_gdp in the replication file.
- totalec_gdp in the replication file is simply the addition of the above to elements

The second file was constructed with the following *R*code:

- explore_db_laborunits.R

1.3 Small Business Administration, NAICS-ISIC rev3. Crosswalk, and ISIC rev.4 to ISIC rev3 Crosswalk

The file with 2007's data from the Small Business Administration , used for the computation of firm-shares in the U.S., which act as ingredient for the aggregation of average firm size across countries controlling for industrial composition, is provided in the above-mentioned repository. The data can be downloaded from: https://www.sba.gov/node/12162., under U.S. static data, detailed industry data NAICS2002.

The R-file producing such computation of firm-shares is

- construct_small_business_administration.R

which uses the following table to convert from NAICS to ISIC revision 3 industry classification codes:

- 2002 NAICS_to_ISIC_rev3.csv

The NAICS 2002 to ISIC revision 3 conversion table can be downloaded from the U.S.Census Bureau's website, https://www.census.gov/eos/www/naics/concordances/concordances.html.

Running the *R* file creates the following file:

- Manufacturing_Data_2digit_ISIC31.csv

which is then loaded by subsequent files for the computation of average firm sizes (see section 3).

In some countries, the firm level data classifies industries according to ISIC revision 4. To convert the US's average firm size across 2 digit industries from ISIC revision 3 to ISIC revision 4, I provide the following file

- manufacturing_2digit_analysis.R

which uses the following conversion table at the 2-digit level

- ISIC31_to_ISIC4_2digits.csv

also downloadable from the United Nation's Statistic Division, https://unstats.un.org/unsd/classifications/Econ/ISIC.csht

1.4 Firm-Level Databases

The firm level databases are composed of commercial, confidential, and publicly accessible sources. The commercial database is AMADEUS, which is developed by Bureau Van Dijk (van Dijk, 2018). These can be purchased from the developer or accessed through an affiliation with an institution with an active membership to the data. As explained in the data appendix of the article, the countries selected from AMADEUS for the analysis are: Bulgaria, Belgium, Finland, Portugal, Spain, Latvia, France, Hungary, Romania, and Italy.

There are a number of firm-level censuses whose access is confidential, due to agreements between the corresponding statistical agencies and the World Bank. The firm-level Censuses in this category are: El Salvador (2005) , Kenya (2010) , Ethiopia (2010) , Ghana (2003) , Peru (2008) , Pakistan (2005) , Bangladesh (2012) , and Malaysia (2015) .

The remaining firm-level datasets are freely accessible from the countries' statistical agency websites. These are:

- Colombia's Annual Manufacturing Survey for 2016 (Departamento Administrativo Nacional de EstadÃstica, 2016),http://microdatos.dane.gov.co/index.php/catalog/MICRODATOS/about_collection/6/2
- Chile's Manufacturing Survey for 2013 (Instituto Nacional de Estadistica, 2013), accessible at https://datosabiertos.ine.cl/developers/
- India's Annual Survey of Industry for 2005 (Central Statistics Office , Industrial Statistics Wing), accessible at http://microdata.gov.in/nada43/index.php/catalog/ASI. While access is unrestricted, users are required to create and register an account with Central Statistics Office.

In addition to the raw firm level data, some of which are accessible and some are not, the replication repository offers all the aggregate statistics computed from the firm-level data that are necessary for the inference of entry barriers and computation of counterfactuals. Hence, readers should be able to replicate the results, given these inputs, from the computational codes described later. More information on data sources can be found in the Data Appendix of the manuscript.

2 Quantitative Analysis

There are two main files for the computation of the numerical exercises. The first one identifies the model-based entry barrier and computes allocations under various combinations of distortions (that is, undistorted, distorted with both distortions, and distorted with one distortion and a time). The second one solves the same allocations but adopting the World Bank's Doing Business Indicator's measure of entry barriers, translated into model-equivalent labor units. A third file is provided for the simulation of the life-cycle dynamics of firms

2.1 Baseline Quantitative results

The file performing the identification of entry barriers is the following:

- master_avs10_ind_weighted_reg_tr5.f90

The file is compiled using gfortran, the GNU Fortran compiler for Linux distributions

As inputs to the computation, the files loads the average firm size across countries and the WLS regression coefficient between $log\,(TFPR)$ and $log\,(TFPQ)$. The files containing this information are

- avsize10_usa_weight.txt
- regcoeff_weighted_reg_tr5.txt

As said, the average firm size is computed conditional on the sample of 10+ worker firms in each country. Furthermore 2 digit average firm sizes are aggregated, within each country, according to the US distribution of firms across 2-digit industries, as explain in the manuscript.

The slope of the idiosyncratic distortion profile is given by the WLS estimate of the regression between $log\,(TFPR)$ and $log\,(TFPQ)$, where $TFPR$ and $TFPQ$ are demeaned by their respective 4-digit averages. Furthermore, prior to the regression, the tails of the distribution of $TFPR$ and $TFPQ$ are trimmed at the 5%

2.2 Quantitative Results under World Bank's Doing Business Indicators' Entry Barrier

A similar code solves the equilibrium allocations under the various combination of distortions imputing the entry barrier from the World Bank's Doing Business Indicators. The file is the following:

- master_avs10_ind_weighted_reg_tr5_dbindicator.f90

The file is compiled using gfortran, the GNU Fortran compiler for Linux distributions

To get at a model equivalent entry cost from the Doing Business Indicator (DBI), we proceed as follows. We start by adding the total cost of starting a firm and acquiring electricity as a proportion of income per capita from the Doing Business Database, $DB = \frac{(Start+Electricity)}{(Y/L)}$. Multiplying the DBI by the inverse of the labor share, which we take from the Penn World Table version 9.0, we get the level of the cost of entry in units of labor: $DB^L = \left(\frac{start+electricity}{Y/L}\right) * \frac{Y}{w*L} = \left(\frac{start+electricity}{W}\right)$. Expressed in this fashion, the Doing Business's cost of entry is comparable to $f_e\tau^e$ in the model. Thus, to isolate the Doing Business counterpart of τ^E, we divide DB^L by the calibrated value of the technological component of the cost of entry, f_e (i.e. $\tau^E_{DB} = \frac{DB^L}{f_e}$).

The file collecting the DBI's cost of starting a firm in units of labor,which is loaded by the Fortran file, is given by

- db_indicator_laborunits.txt

2.3 Simulation of Life-Cycle Dynamics of Employment

Figure 7 in the manuscript reports the life-cycle dynamics of employment for a subset of countries in the sample. These life-cycles are first simulated using the following Fortran file:

- life_cycles_weighted_reg_tr5.f90

This program loads the productivity, employment, and innovation profiles across productivity levels across countries, under their given pair of distortions, generated in the master Fortran Program. These files are provided to the reader in the following files:

- Innovation probability for each productivity level: innovdist_weighted_reg_tr5.txt
- Total labor demand for each productivity level: emptotdist_weighted_reg_tr5.txt
- Mass of firms in each productivity level: massdist_weighted_reg_tr5.txt
- Production labor demand for each productivity level: lpdist_weighted_reg_tr5.txt

The outcome from the Fortran files that are loaded by the R-file are:

- LdAWeightedRegTr5.pc
- LdAFlessAvs10ind.pc

These files are in turn loaded by replication_figures.R

3 Analysis Data Files

The replication repository provides a series of files that are loaded in the process of constructing the figures in the manuscript.

3.1 Average Firm Size Weighted by US's Shares of Firms across 2 Digit Industries

The average firm size across countries, constructed as dictated by equation 17 in the article, are reported in the following file:

- Av_Size_bySector_All_short_WeightedReg_tr5.csv

Because the firm-level data that allows for the computation of average firm size is to a large extent confidential, I provide the *R* code that computes the average firm size for one of the countries where the firm-level data is publicly available, Chile.

The firm-level analysis in Chile is performed in the following file:

- firm_analysis_chile.R

For running the firm analysis in Chile, the reader must

1. Download the firm-level data from the official sources, as explained above.
2. Run construct_small_business_administration.R to compute average firm sizes in the US across 2 digit manufacturing industries classified according to ISIC rev3
3. Run manufacturing_2digit_analysis.R to convert the average size distribution from the ISIC rev3 classficiation to ISIC rev 4
4. Run firm_analysis_chile.R

3.2 Regression Coefficient $log\,(TFPR) - log\,(TFPQ)$

The employment-weighted least squares regression coefficients of $log\,(TFPR)$ against $log\,(TFPQ)$ are stored in the following file:

- RegCoeffs_by_Country_WeightedReg_tr5.csv

Again, using Chile as example, the R program just described for the computation of the average firm sizes also performs the derivation of the idiosyncratic distortions, following Hsieh and Klenow (2009)'s methodology , and estimates the WLS regression.

3.3 Employment Histograms of Firm Size Distribution

As validating evidence for the mechanisms in the model, figure 3 in the article shows the share of firms with 250 workers or more across countries in the sample. Continuing with Chile as example of program that allows to replicate the construction of the cumulative distribution function from which the shares are computed, the repository provides the following file for replication:

- employment_histograms_chile.R

This R-file constructs histograms of the firm size distribution with employment bins that are comparable to the bins reported in the Small Business Administration and the Business Dynamics Statistics databases. After constructing the cumulative distribution functions corresponding to the size bins, the file then computes the share of firms with 250 workers or more. The output of this computation, applied to the Census-based and Amadeus-based databases, which are in turn taken as ingredients in the replication_figures.R file for the construction of figure 3, are the following:

- Mshare250censuses.txt
- Mshare250amadeus.txt

While the raw firm-level data is not provided in the replication repository, these text files are, and hence the reader should be able to replicate the corresponding figure.

3.4 Ingredients for Fortran Programs

A number of files provided in the repository act as inputs to be loaded by the Fortran programs that compute the various counterfactuals in the model.

These are:

1. avsize10_usa_weight.txt, the average firm size across countries
2. regcoeff_weighted_reg_tr5.txt, the $log(TFPR) - log\,(TFPQ)$ regression coefficients
3. db_indicator_laborunits.txt, the World Bank's cost of starting a firm translated into model equivalent units, as explained in section 1.2

The first two files contain information that was reported section 3.1, but are presented in a separate txt file of ease of access by the corresponding Fortran programs. Section 2.1 provides more details.

3.5 Output from Fortran Programs

The following list of files collect the results from the Fortran programs that are used for the creation of the figures in the paper (more on this below). The output files are:

1. ResMasterAvs10indWeightedRegTr5.dat
2. dist_weighted_reg_tr5_DB.txt
3. res_db_weighted_reg_tr5.txt
4. resgains_db_weighted_reg_tr5.txt
5. LdAWeightedRegTr5.pc
6. LdAFlessAvs10ind.pc

File number 1 is the main output file from the Fortran programs that compute the undistorted stationary equilibrium, the distorted equilibrium with both model-based entry barriers and idiosyncratic distortions, the distorted equilibrium with idiosyncratic distortions only, and the distorted equilibrium with model-based entry barriers only.

File number 2 is the output file from the Fortran program that solves the distorted equilibrium with World Bank's Doing Business Indicators cost of starting a firm and idiosyncratic distortions.

File number 3 presents the TFP gains from removing both the World Bank's entry barriers and the idiosyncratic distortion, from removing the entry barrier only, and from removing the idiosyncratic distortion only.

File number 4 presents the same information as file number 3, as well as the gains from removing each distortion assuming there other distortion is set to zero.

Files 5 and 6 are the output from the Fortran program that computes the life-cycle employment dynamics of firms under every country's estimates of distortions.

4 Figures

Lastly, the replication materials provide a file for the creation of all the figures reported in the paper. The file name is the following:

- replication_figures.R

The file is was created in $R-Studio$ Version 1.0.153 for Linux. Required libraries are declared at the beginning of the file.

All the input files that are loaded for the construction of the figures are provided as part of the replication materials.

5 Software Requirements

- R version 3.6.3 (2020-02-29) -- "Holding the Windsock" Copyright (C) 2020 The R Foundation for Statistical Computing Platform: x86_64-pc-linux-gnu (64-bit).

- $R - Studio$ Version 1.0.153 – © 2009-2017 RStudio, Inc. Required libraries are declared at the beginning of the file, and can be installed through the R-studio interface
- Fortran compiler: GNU Fortran (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0 Copyright (C) 2019 Free Software Foundation, Inc.
- R libraries can be downloaded running the global-libraries.R file, provided in the repository, or through the $R - studio$ interface.

References

U.S. Small Business Administration. Firm Size Data. 2015. URL https://www.sba.gov/node/12162.

Ethiopia's Central Statistical Agency. Large and Medium Manufacturing and Electricity Industries Survey. 2010.

Government of India Central Statistics Office (Industrial Statistics Wing) MOSPI. *Annual Survey of Industries*. 2005.

Instituto Nacional de Estadistica e Informatica de Peru. Censo Nacional Economico. 2008.

Direccion General de Estadisticas y Censos de El Salvador. *Censo Economico*. 2005. URL http://www.censos.gob.sv/cecon/#.

Gobierno Nacional Colombia Departamento Administrativo Nacional de EstadÃstica. *Encuesta Anual Manufacturera*. DANE, 2016.

Robert C. Feenstra, Robert Inklaar, and Marcel P. Timmer. The next generation of the penn world table. *American Economic Review*, 105(10):3150–3182, 2015. URL http://www.ggdc.net/pwt/.

Chang-Tai Hsieh and Peter J. Klenow. Misallocation and manufacturing tfp in china and india. *The Quarterly Journal of Economics*, 124(4):1403–1448, November 2009. URL http://ideas.repec.org/a/tpr/qjecon/v124y2009i4p1403-1448.html.

Chile Instituto Nacional de Estadistica. *Encuesta Nacional Industrial Anual*. INE, 2013. URL https://datosabiertos.ine.cl/developers/.

Bangladesh Bureau of Statistics. Survey of Manufacturing Industries. 2012.

Kenya's National Bureau of Statistics. Census of Industrial Production. 2010.

Pakistan's Federal Bureau of Statistics. Census of Manufacturing Industries. 2005.

Department of Statistics Malaysia. Census of Manufcturing Sector. 2015.

Ghanaian Statistical Service. National Industrial Census. 2003.

Bureau van Dijk. *AMADEUS*. 2018. URL https://www.bvdinfo.com/en-us/our-products/data/international/amadeus.

World-Bank. *Doing Business 2019*. Number 30438 in World Bank Publications. The World Bank, June 2019. URL https://www.doingbusiness.org/.

Achim Zeileis. *pwt9: Penn World Table (Version 9.x)*, 2019. URL https://CRAN.R-project.org/package=pwt9. R package version 9.1-0.