Replication material for "A tale of one tail: business size and the barriers to development and equity in Latin America" – Marcela Eslava, Marcela Meléndez, Laura Tenjo, and Nicolás Urdaneta

This folder contains all Stata codes and public domain datasets necessary to fully replicate the results of this paper. The empirical analysis in this paper relies primarily on data publicly available or available upon request from national household surveys. We provide information on how to access the surveys in case of being publicly available or how to request access if not already available. Additionally, we provide two datasets: one including the raw household surveys for the Latin American countries for which data is public (all but Costa Rica and Dominican Republic) only with the relevant variables to replicate the paper. Second, we provide a dataset with the share of employment in each category for each income decile for all the countries in our sample.

Data Availability and Provenance Statement

The labor force survey for each country is primarily accessible through the statistical agency's website of each country. In certain instances, access to these surveys may require prior registration. In the following sections, we will outline the steps to access each survey and provide a detailed description of the registration process, including any necessary application (and approval) for access.

Software Requirements and Runtime

The programs require Stata. The code was run on Stata MP version 17.0. The processing of all results, including processing the datasets, takes approximately 30 minutes.

Description of Data Availability and Access

1. Household Surveys

Argentina - Encuesta Permanente de Hogares 2019. Microdata for the full survey is available here. Data comes by trimesters and in different modules.

Bolivia - Encuesta de Hogares 2019. Microdata for the full survey is available here. Data files come for the whole year but in different modules.

Brazil - Pesquisa Nacional por Amostra de Domicílios Contínua 2019. Microdata for the full survey is available here. The data we use comes in a single file.

Chile - Encuesta de Caracterización Socioeconómica Nacional 2017. Microdata for the full survey is available here. Data comes in a single file.

Colombia - Gran Encuesta Integrada de Hogares 2019. Microdata for the full survey is available here. This labor force survey is run monthly and includes multiple modules that have to be appended and merged.

Costa Rica - Encuesta Nacional de Hogares 2019. Microdata for the full survey is available here but access requires registering first. Data comes in a single file.

Dominican Republic - Encuesta Nacional Continua de la Fuerza de Trabajo 2019. Microdata for the full survey is available under request to the country's central bank (Banco Central República Dominicana). Access can be requested here. We received an excel file for the survey in 2018 and 2019. There is a possibility that the data received must be pre-processed as some variables come in string format instead of numeric. There is also a possibility that the secondary income variable may have some implausible values. While our dataset did not present any of these two issues, we include a line of code that censors such secondary income observations. We note that doing this has almost no effect on the results displayed in the paper.

Mexico - Encuesta Nacional de Ingresos y Gastos de los Hogares 2018. Microdata for the full survey is available here. Data comes in a single year file with 12 separate modules.

Paraguay - Encuesta Permanente de Hogares Continua 2019. Microdata for the full survey is available here. We use the yearly version that contains two data files (registries) that are merged.

Peru - Encuesta Nacional de Hogares 2019. Microdata for the full survey is available here. Data comes in separate files for each module.

Uruguay - Encuesta Continua de Hogares 2019. Microdata for the full survey is available here. Data comes in a single year file, there are three separate modules that we merge together.

For all Latin American Countries (except Costa Rica and Dominican Republic) we include the raw datasets with the variables required to replicate results. Researchers can access the complete datasets with all variables from each country's statistical department website. For each country, the dataset we provide has appended all survey waves of the year if there are multiple waves and we have merged survey modules if they come separate in the raw downloaded files.

European countries: EU-Statistics on Income and Living Conditions (SILC).

Eurostat compiles household surveys from all European Union countries designed to be comparable and harmonized. Access to the EU-SILC is restricted, application to access the data is required, and is only permitted for individuals with scientific research purposes. The application process depends on each case and may take approximately 12 weeks. Public use microdata files for certain countries and older years are available here. For additional information about the EU-SILC, please consult the following resource here.

We use information from the following available countries: Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Luxembourg, Sweden, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Greece, Latvia, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, and Spain. We exclude Netherlands in our analysis as

requested by Eurostat and we also exclude the UK and Ireland because the 2019 survey was not available at the time.

United States

- Consumer Population Survey (CPS) 2018

We accessed the CPS-ASEC modules microdata available through IPUMS. It is necessary to register, accept the terms of service, and create a data extract to download here. IPUMS registration, and data request take some minutes/hours to be approved. We rely on the ASEC modules since these include information on firm size and self-employment status for workers.

The data request should include the samples for IPUMS-CPS, ASEC 2018. We downloaded the data in Stata format in cross-sectional rectangular structure. When making the data request only the following variables should be included: YEAR, SERIAL, MONTH, HWTFINL, CPSID, ASECFLAG, ASECWTH, PERNUM, WTIFNL, CPSIDV, CPSIDP, ASECWT, AGE, SEX, EMPSTAT, LABFORCE, OCC, IND, CLASSWKR, EDUC, OCCLY, INDLY, CLASSWLY, WKSWORK1, UHRSWORKLY, FULLPART, PENSION, FIRMSIZE, FTOTVAL, INCTOT, INCWAGE, INCBUS, INCFARM, INCSS, INCWELFR, INCRETIR, INCSSI, INCINT, INCUNEMP, INCWKCOM, INCVET, INCSURV, INCDISAB, INCDIVID, INCRENT, INCEDUC, INCCHILD, INCOTHER.

- Business Dynamics Statistics (BDS) 2018

We used publicly available tables from the BDS that categorize employment by firm size categories. The data tables are available here. We provide the file we downloaded and the data we used in the "Public datasets" subfolder in the file "BDS firm size.xlsx".

Australia – Household, Income and Labour Dynamics (HILDA) Survey – Release 20

This survey is conducted by the Department of Social Services (DSS). To access any of the DSS Longitudinal Studies datasets it is necessary to sign a confidentiality agreement. For more information, please refer to the Australian Data Archive Dataverse website here. For our analysis, we used wave 'S' of the HILDA survey, which corresponds to the year 2019.

The data package for each wave includes four types of data: Household, Enumerated person, Responding person, and Combined files. In our case, we specifically selected the Combined file, which combines the information from the Household, Enumerated person, and Responding person files.

Japan - Labor Force Survey 2019

For the Japanese LFS we do not have access to microdata. We downloaded tabulation data table containing employed persons by firm size. We provide the tables download in

the "Public data" folder. The two datasets we download are Table 2-1-1 (available here) and Table II-5 (available here).

South Korea – Korean Labor & Income Panel Study (KLIPS) Survey 2019

Data from the KLIPS survey is publicly available but researchers must register in the Korean Labor Institute's website (here) in order to download the data files. We downloaded files in Stata's file format and used wave 22 which corresponds to the year 2019. We focus on the individual-level data which can be identified by the letter "p". Nevertheless, we also used the Work History file, identified by the letter "w", to merge main-job and side-job data.

India – Periodic Labour Force Survey 2018-2019

Data for the PLFS is publicly available, but researchers must register in the National Data Archive website of the Ministry of Statistics & Programme Implementation. The PLFS 2018-2019 microdata and registration are available here.

Pakistan - Labour Force Survey 2018-2019

The microdata for the Pakistani LFS is publicly available here. We include the data in the replication folder in the subfolder "Raw microdata".

Nepal - Labour Force Survey 2017-2018

Data for the Nepali LFS is publicly available, but researchers must register in the Central Data Catalog website of the National Statistics Office. The PLFS 2018-2019 microdata and registration are available here.

2. Country-level indicators

We also rely on country level statistics: GDP per capita, GINI, and PPP conversion factor that were obtained from the World Development Indicators (WDI, 2023) in July of 2023. We include this publicly available data in the subfolder "Public Datasets" in the file "WDI data.xlsx".

Description of Code and Data Processing

This folder contains the codes to produce the processed household surveys we make available from Latin America, the supplementary dataset for employment shares in each productive unit size by income deciles, and the tables and figures for the paper (and appendix). For replication, follow these steps:

1. Unzip this replication material folder, let's call it "main folder". This zip folder contains the subfolder structure necessary to replicate the results of the paper.
2. Register and request access to the datasets that are not publicly available. Download all these data files and place them in the subfolder "Raw microdata". You will find

in this subfolder the microdata that is publicly available for 9 Latin American countries. Make sure that the datasets downloaded from each country have the following name and format (with the exception of EU countries which are described further below):

- Costa Rica: *ENAHO 2019.sav*
- Dominican Republic: *Base ENCFT 20181 - 20194.xlsx*
- United States: *cps_18.dta*
- Australia: *Combined_s200c.dta*
- South Korea: *eklips22w.dta, eklips22p.dta*
- India: *PerV1.txt*
- Nepal (since there are multiple files, place them under the folder "Raw microdata/Nepal"): *S00_rc.dta, S02_rc.dta, S03_rc.dta, S04_rc.dta, S05_rc.dta, S06_rc.dta, weights.dta*

Since the process to access the EU-SILC data is different from the rest and the data delivered is already organized in a certain structure it is only necessary to place the files received in the "Raw microdata" folder. By doing this, there must be a subfolder inside "Raw microdata" for each EU country with the two-letter abbreviation of the country, followed by a folder named "2019" that contains three files. As an example, for Italy (**IT**), the folder and file structure should be as follows:

- Raw microdata/**IT**/2019/UDB_c**IT**19P.csv
- Raw microdata/**IT**/2019/UDB_c**IT**19H.csv
- Raw microdata/**IT**/2019/UDB_c**IT**19D.csv

3. Run the master dofile, "Master.do", in the programs folder. Change the directory in line 28 to your computer's directory. This code runs all the codes to replicate the paper. They are structured as follows:
   - Section 1 runs the codes that process each country's (or sets of countries) household survey.
   - Section 1.1 takes the LATAM dataset processed before and produces the analysis for the Latin American countries.
   - Section 2 runs the codes that aggregate results for countries in some regions for which we have separate programs:
     - 2. Aggregate Asia.do
       Here we aggregate results for Australia, South Korea, and Japan for Figure 1 - Panel A and Figure 3.
     - 2. Aggregate South Asia.do
       Here we aggregate results for India, Pakistan, and Nepal for Figure 1 – Panel A and Figure A2

- Section 3 runs the code that takes the results from previous sections to construct the figures and tables displayed in the paper.
- Section 4 runs the code that organizes the dataset we make available with employment shares in each productive unit size by income deciles for each country.

There are a few additional results mentioned in the text of the paper that can be found in the following programs (the rest of numeric results mentioned can be found in their corresponding tables):

1 Introduction

- In the first paragraph of page 3 we describe that "Much of the left-out activity is informal in LDCs (3/4 of the segment in our LATAM data)". This result comes from obtaining the fraction of informal employment in the first two categories of Figure A1.

3.2 The joint distribution of individual earnings and business size

- Before Table 1 we make the following statement: "The additional regressor with the highest explanatory power is years of schooling (…) and reduces the earnings gap between elf-employed and the 50+ class to a factor close to 1.8, rather than 2 in the unconditional regression." This result is produced in "Programs/1.1 LATAM Table 1.do" and the result can be found in "Output/Tables.xlsx" in sheet "Table 1B".

References

Banco Central de la República Dominicana. 2019. "Encuesta Nacional Continua de la Fuerza de Trabajo, ENCFT". Banco Central de la República Dominicana.

Central Bureau of Statistics (2017-2018): "Nepal Labour Force Survey," National Planning Commision, Central Bureau of Statistics, Government of Nepal.

DANE. 2019. "Gran Encuesta Integrada de Hogares, GEIH." Departamento Administrativo Nacional de Estadística, Colombia.

Department of Social Services. 2019. "Household, Income and Labour Dynamics in Australia Survey, HILDA." Department of Social Services, Government of Australia.

EUROSTAT. 2019. "European Union Statistics on Income and Living Conditions, EU-SILC." EUROSTAT.

Flood, S., King M., Rodgers R., Ruggles S., Warren J.R., and Westberry M. (2021) Integrated Public Use Microdata Series, Current Population Survey: Version 9.0 [dataset]. Minneapolis, MN: IPUMS.

IBGE. 2019. "Pesquisa Nacional por Amostra de Domicílios Contínua." Instituto Brasileiro de Geografía e Estatística, Brazil.

INDEC. 2019. "Encuesta Permanente de Hogares, EPH." Instituto Nacional de Estadística y Censos, República Argentina.

INE Bolivia. 2019. "Encuesta de Hogares." Instituto Nacional de Estadística, Bolivia.

INE Paraguay. 2019. “Encuesta Permanente de Hogares Continua, EPHC.” Instituto Nacional de Estadística, Paraguay.

INE Uruguay. 2019. “Encuesta Continua de Hogares.” Instituto Nacional de Estadística, Uruguay.

INEC. 2019. “Encuesta Nacional de Hogares, ENAHO.” Instituto Nacional de Estadística y Censos, Costa Rica.

INEGI. 2018, “Encuesta Nacional de Ingresos y Gastos de los Hogares.” Instituto Nacional de Estadística y Geografía, México.

INEI. 2019. “Encuesta Nacional de Hogares, ENAHO.” Instituto Nacional de Estadística e Informática, Perú.

Korea Labor Institute. 2019. “Korean Labor & Income Panel Study, KLIPS.” Korea Labor Institute.

Ministerio de Desarrollo Social y Familia. 2017. “Encuesta de Caracterización Socioeconómica Nacional, CASEN.” Ministerio de Desarrollo Social y Familia, Chile.

National Statistical Office (2018-2019): “Periodic Labour Force Survey,” Ministry of Statistics and Programme Implementation, National Statistical Office, Government of India.

Official Statistics of Japan. 2019. “Labour Force Survey.” Portal Site of Official Statistics of Japan. Ministry of Internal Affairs and Communications, Japan.

Pakistan Bureau of Statistics (2018-2019): “Labour Force Survey,” Pakistan Bureau of Statistics, Government of Pakistan.

U.S. Census Bureau. 2023. “Business Dynamics Statistics.” United States Census Bureau.

WDI World Bank, “World Development Indicators”. (2023). World Bank.