Data Documentation for the Countries in GRID

May 17, 2022

Phase 1 Countries:
Argentina
Brazil
Canada
Denmark
France
Germany
Italy
Mexico
Norway
Spain
Sweden
UK
US

Argentina

Data Sources

The primary data source for Argentina consists of employer-employee matched panel data based on administrative records from Argentina's social security system, called *Sistema Integrado Previsional Argentino* (SIPA). Records come from sworn statements that employers must present by law each month to Argentina's tax authority, *Administración Federal de Ingresos Públicos* (AFIP). These records contain information about payroll for which employers pay social security contributions (i.e., formal workers). Statistics come from a 3% random, anonymized subsample of employees in the private sector spanning the 1996–2015 period. The random sample is called the *Registered Employment Longitudinal Sample* (RELS), and it is compiled by the Ministry of Labor, Employment and Social Security of Argentina at a monthly frequency. The microdata and documentation are publicly available at http://www.trabajo.gob.ar/estadisticas/oede/mler.asp.

The dataset is representative of the formally employed population at private firms in all sectors and regions and covers all types of contracts (e.g., full-time workers, internships, temporary workers). It contains data from about 130,000 workers in 1996 to 230,000 in 2015. With formal private employment accounting for roughly 30% to 40% of total employment over the period (including independent and self-employed workers), the sample amounts to about 1% of the employed population in any given year.

The income measure used in the database is gross labor earnings, inclusive of all forms of compensation subject to taxation and social security contributions (i.e., base salary, overtime compensation, performance and seasonal bonuses, paid vacations, paid sick leaves, and severance payments). Earnings information for individuals above the 98th percentile of the within-industry earnings distribution is anonymized to protect the privacy of employees in the data.

Sample Selection

To enhance harmonization and allow meaningful comparisons across countries in the project, the following sample selection criteria are adopted. First, the focus is on workers between 25 and 55 years old. Second, only earnings observations above a threshold are retained (the threshold equals what a worker would earn if they were to work part-time for one quarter at the national minimum wage).

When computing longitudinal statistics, two additional restrictions are applied. First, only a subsample of workers for which 1-year and 5-year earnings changes can be computed is retained. Second, only workers who have been in the sample for the previous three consecutive years are retained (to allow computation of a permanent income measure, see below).

Brazil

Data Sources

The main data source for Brazil is RAIS, which contains administrative records from Brazil's Labor Statistics Dissemination Program (*Programa de Disseminação das Estatísticas do Trabalho*) within the Brazilian Ministry of the Economy (*Ministério da Economia*), formerly the Ministry of Labor (*Ministério do Trabalho*). RAIS covers nearly the entire universe of workers in tax-registered firms. It excludes informal workers and firms, firm owners and shareholders (unless they are self-employed), and certain less-populated regions in the years before 1994. The RAIS data are based on legally required annual reports made by firms that transmit information to the Brazilian government on all employees who were on the payroll in the previous year.

Each observation in RAIS is a worker-establishment match, or job, in a given year. The income variable includes wage, salary, and (holiday and performance) bonus payments before taxes from all (formal) employment.

Each worker has a unique identification number in RAIS, so that the full formal work history of all individuals in the database can be recovered. Data is from 1985 to 2018. RAIS is very large, with an average of around 40 million observations per year, which sums to approximately 1.2 billion job records for the 1985–2018 period. For the year 1996, this corresponds to around 42.5 percent of the labor force, 46.0 percent of employment, and close to all formal employment.

Sample Selection

Some standard filters are applied to the administrative data to enhance harmonization and allow meaningful comparisons across countries in the project. First, all workers without valid identification numbers or with zero earnings are dropped. The data is then restricted to workers in the 25–55 age range. Earnings data in RAIS are censored (i.e., reported as missing) above 120 times the national minimum wage. To focus on workers with a meaningful attachment to the labor market, those with total annual nominal earnings below earnings associated with part-time employment at the minimum wage for three months are dropped. Any observation of an individual i in year t with total annual nominal earnings y_{it} if $y_{it} < \frac{1}{2} \times 40 \frac{\text{hours}}{\text{week}} \times 40 \frac{\text{weeks}}{\text{month}} \times 3 \text{ months} \times MW_t$, where MW_t is the mean prevailing minimum wage over the individual i's period of employment in year t is also dropped.

Since the period from 1985 to 1994 was characterized by high inflation and multiple currency switches, focus is first placed on monthly earnings reported in multiples of the prevailing minimum wage. This multiple is then multiplied by the mean nominal minimum wage in current Brazilian Reais during the months of a given job spell. As the hiring and separation dates for each job spell are known, total annual nominal earnings is constructed for each individual by summing their mean monthly earnings over all months of employment and all jobs recorded in RAIS in a given calendar year. Finally, total annual real earnings is obtained by deflating total annual nominal earnings by the mean IPCA, normalized in December 2018, during the months of employment. The same method is applied each year throughout the whole period 1985–2018.

Canada

Data Sources

The data set for Canada is the Canadian Employer–Employee Dynamics Database (CEEDD). The CEEDD is a linkable environment developed by Statistics Canada that consists of several administrative and tax files for individuals and firms. The individual-level data are drawn from the T1 Personal Master File (T1PMF), which contains annual personal income tax records for all Canadian tax filers who filed their tax returns before a specified cut-off date, usually in December one year after the tax reference year. Only about 3.5-4.8% of all tax filers do not file tax returns before this date; these late filers are not included in T1PME. Over the sample period, 1983–2016, the T1PMF includes records each year for 89–93% of all 25- to 55-year-old Canadians.

The income measure is annual (pre-tax) individual earnings, given by the sum of employment income (wages, salaries, bonuses, overtime pay, paid vacation, and commissions) reported on T4 slips from all jobs and other taxable receipts from employment (e.g., tips, gratuities, director's fees) that are not reported on T4 slips. Self-employment income is not considered. All earnings from 1983 to 2016 are denominated in 2018 Canadian dollars using the Canadian Consumer Price Index.

Sample Selection

To ensure cross-country comparability, the analysis uses earnings observations only for individuals ages 25–55. Very low earnings observations are dropped. In particular, earnings have to be above a minimum earnings threshold in year $t, \underline{y}_t$, defined as the amount a worker could earn by working 20 hours per week for a quarter of the year (13 weeks) at the real minimum wage for that year.

Denmark

Data Sources

The data source for Denmark is based on merged administrative data covering the entire Danish population for the period 1987 to 2016. The various administrative registers are collected by Statistics Denmark from relevant public authorities and made available to researchers. The core data set is compiled by the Danish Tax Agency, which collects information about earnings for all employees directly from all employers in Denmark. Earnings include the value of fringe-benefits, severance payments, and the value of stock options, but they do not include contributions to employer pension accounts, since these are tax deductible and are subtracted at the payroll level. As taxes are calculated based on gross income, including transfer income, earnings are measured before taxes. None of the income measures are top-coded. Earnings are deflated using the consumer price index with 2018 as the base year.

Sample Selection

Only individuals who are aged 25-55 and who have positive earnings are included. Next, only individuals who have earnings amounting to at least 28,500DKK annually (2018 prices) are included. In terms of earnings, this roughly compares to the level for a part-time job held for one quarter.

France

Data Sources

The data source for France is DADS (*Déclaration Annuelle des Données Sociales*), a French Linked Employer-Employee dataset for the period 1991-2016. These administrative data are based on mandatory employer reports of the earnings of each employee subject to French payroll taxes. They include all employers and their (declared) employees. Because of legal constraints the full panel version does not include all workers. The panel combines a random sample (individuals born in October of an even year) from the DADS with data on central government public employees, similarly selected. The sample size increased by two in 2002 by including individuals born in October of an odd year.

The data are aggregated at the job spell level (in an establishment in a given year for a given individual). Hence, the data on earnings use this employment (job) spell level. For each individual, total earnings in year *t* is defined as the sum of earnings across all employment spells in that year. Earnings are measured at gross level (i.e. net labor earnings plus workers' mandatory social contributions). This measure includes the sum of wages, over-time hours, paid leaves, bonuses, in-kind benefits, and several kinds of compensations (sickness, short-time work, severance payments, etc.). It does not include stock options nor employer-paid payroll taxes. Earnings are expressed in 2018 euros deflated using the CPI computed by the French Statistical Office (INSEE).

Sample Selection

In line with other countries requirements for this project, there is an age requirement (25-55) to be in the sample and a minimum level of annual earnings for an individual to be included in the data. More precisely, an observation must have earnings above the equivalent of 260 hours paid at the French minimum wage, as it corresponds approximately to a part-time job for one quarter. Every year, between 6 and 7% of the observations of the sample are excluded.

Germany

Data Description

The database for Germany combines two high quality administrative data sources: social security data (IAB) and personal income tax records (TPP).

The first source of data (IAB), is the Integrated Employment Biographies (IEB, version 13.01) supplied by the Institute for Employment Research (*"Institut fur Arbeitsmarktund Berufsforschung", IAB*). A 10% random sample of the IEB for the years 1993-2018 is available, which gives 87,012,649 observations. Focus is placed on the period 2001-2016 (65,900,481 observations on 6,250,877 individuals).

The second source of data is the German Taxpayer Panel (TPP), which is an administrative dataset based on the universe of personal income tax returns. The dataset covers all tax units filing tax returns in Germany in the period 2001-2016. The panel has a total of 58,808,899 unique records for which information is available for at least two years. A 25% random sample of these records is used.

Sample Selection

For comparability with other countries covered in the GRID project, focus of the analysis is on individuals who are between 25 and 55 years old. Following the GRID guidelines, the focus is on labor earnings excluding self-employment. The definition of gross annual labor earnings is the same for both IAB and TPP data: annual earnings is broadly defined and include, among others, overtime pay, bonuses, 13th month pay, paid sick leave, severance pay, and vacation allowance. Workers with weak attachment to the labor force are dropped by trimming annual earnings below a threshold $\underline{y_t}$, which corresponds to working part-time for one quarter at the national minimum wage (2,300 Euro in 2018). This cuts approximately 7% of all workers. All incomes are deflated using the CPI and Euro figures in the text, tables and figures refer to 2018 Euro.

Italy

Data Sources

The data source for Italy is INPS (*Istituto Nazionale di Previdenza Sociale*), the equivalent of the US Social Security Administration. The data cover the period 1985-2016. Statistics come from a 6.6% sample of the INPS universe based on workers born on 24 randomly selected birth dates. Public sector jobs, as well as self-employment, are not in the INPS archives. These account for 16% and 20% of total employment respectively.

The basic observation is a job relationship within a year, based on mandatory employer reports. The main measure of earnings is the sum of all regular and irregular income received by the employer that is subject to social security contributions. This includes base pay, COL adjustments, overtime work, paid vacation and sick leave, bonuses and profit-sharing payments, and the monetary value of in-kind payments, across all jobs of a worker within a given year.

Sample Selection

For comparability with other countries in the project, the sample is restricted to workers aged 25 to 55 who have positive earnings and worked a minimum of 4 weeks over the year. There is no additional minimum earnings threshold imposed. Despite that, the first percentile of earnings in the sample never falls below 800 euros. The sample includes 2.3 million unique workers (1.4 million men and 0.9 million women) and a total of 22.4 million worker-year observations – approximately 700,000 observations per year.

Mexico

Data Sources

The data source for Mexico is based on social security records from the Instituto Mexicano del Seguro Social (IMSS), one of the main Mexican social security institutions. All formal private sector workers who receive a salary are required, by law, to register with IMSS. The set of workers affiliated with IMSS represents approximately 80% of formal sector workers with access to social security, according to estimates from the *Secretaría de Trabajo y Previsión Social* (the Mexican Ministry of Labor) but does not include government workers or workers employed in the informal sector. Since informality is prevalent in Mexico, a large portion of the labor force is not included in the social security data. Self-employed workers —individuals that work on their own and without employees— can register with IMSS to obtain access to some parts of the social security system and hence may appear in the social security records. By default, they are recorded having a wage equal to the minimum wage. For any given month, the share of enrolled workers that are "self-employed" is roughly 0.1% of the total observations.

The social security data cover, approximately, between 13 million workers at the start of the sample and 20 million workers toward the end. The key variable contained in the social security data is the information on wages, reported as a worker's daily taxable income ("*salario base de cotización*"). This means that the data on daily wages can include various forms of compensation received by the worker other than wages (usually paid vacation and end of the year bonus) but may exclude others (in general any additional benefit or compensation that is not subject to labor income taxation), hence not necessarily reflecting the total labor income a worker receives.

Sample Selection

Since the information on wages is available as daily wage ("*salario base de cotización*"), monthly wages are obtained by multiplying the daily wage by 30; monthly wages are then added up to obtain the annual labor income for each worker in a given year. For the period 2005–2019, this results in over 315 million worker-year pairs with observations per year ranging between 17 and 26 million for workers aged 14–75 years old. This constitutes the universe of potential observations. Two "admissibility" conditions are imposed: (i) individuals must be between 25 and 55 years of age (i.e. the prime-age labor force), and (ii) individuals must display *meaningful attachment to the labor force*, in the sense that their earnings must be above a minimum earnings threshold $Y_{\min,t}$. Since in Mexico the minimum wage is defined as a daily wage, rather than as an hourly wage as it is common in other countries, $Y_{\min,t}$ is set equal to 45 days of minimum wage, which corresponds to half a quarter of full-time minimum wage employment. Within the subset of the sample that satisfies the first admissibility condition, the fraction of observations that are above the minimum earnings threshold varies between 97.5 and 98.5% throughout the sample period.

Norway

Data Sources

The data source for Norway consists of several high-quality administrative registers covering the entire Norwegian population. The measure of income used is labor earnings, available between 1993 and 2017. It is a comprehensive measure of labor income from all jobs (except for self-employment income). More precisely, *labor* earnings include: (i) salaries and hourly wages; (ii) fees received by board members, bonuses, commissions; (iii) overtime, piecework, performance, caregiver, severance, and holiday payments; (iv) fixed wage and irregular supplements (linked and not linked to working hours). Data come from annual tax records and are third-party reported by employers. The earnings reported on tax forms also include certain work-related transfers such as sickness and parental leave benefits. These benefits are deducted from the income measure so as to reflect only labor earnings.

All nominal incomes are deflated to their 2018 real values using the Consumer Price Index in Norway.

Sample Selection

The baseline sample includes all residents in Norway between ages 25 and 55 who have a personal identification number. The number of individual-year observations between 1993 and 2017 is about 51.3 million in total. Observations below a certain time-varying annual minimum earnings threshold (Y_t^{min}) are trimmed to focus on workers with a meaningful labor force attachment. Norway does not have a national minimum wage defined by law. Thus, Y_t^{min} is defined as the annual earnings of an individual who works 40 hours a week for a full quarter at half the US minimum wage, which roughly equals NOK 12,000 in 2017. 16% of the earnings observations between 1993 and 2017 are below this threshold.

Spain

Data Sources

The data source for Spain comes from the Continuous Work History Sample (*Muestra Continua de Vidas Laborales*, MCVL, in Spanish), which is a 4% non-stratified random sample from the Spanish population registered with the social security administration in the reference year. Since 2005, individuals who are present in a wave and subsequently remain registered with the social security administration stay as sample members. In addition, the sample is refreshed with new sample members so it remains representative of the population in each wave. For the years prior to 2005, income records are top and bottom coded, so focus is on the period 2005-2018 where uncensored annual earnings from tax information are observed.

Since 2005, the MCVL is matched to data from the tax authority, which provides uncensored individual pre-tax income from paid employment accumulated in a calendar year, as reported by employers to the tax authority, as well as unemployment benefits and subsidies. The tax information comes from "model 190", the "Annual summary of retentions and payments for the personal income tax on earnings, economic activities, awards and income imputations." This form is required of all entities that pay wages, pensions or unemployment benefits. It covers all beneficiaries, including those whose wages fall below the legal minimum of exemption for the obligation to declare personal income taxes. Reported earnings include all taxable payments of the employer to the employee including overtime pay, bonuses, paid vacation, and sick leave benefits. All earnings measures are deflated to 2018 euros using the Spanish consumer price index.

Sample Selection

The analysis focuses on workers who are between 25 and 55 years old, are not self-employed, and do not live in the Basque Country or Navarra (for which the tax data does not provide coverage). Whether the individual has declared herself as self-employed but not how much is earned in self-employment income is observed. Annual earnings below a threshold $\underline{y}_t$ are trimmed, which corresponds to working part-time for one quarter at the national minimum wage.

Sweden

Data Sources

The data source for Sweden is the administrative register LOUISE, provided by Statistics Sweden. The main measure of earnings is annual individual labor earnings including positive self-employment income, which is uncensored. The earnings measure includes all payments from the employer to the employee during the calendar year, including overtime pay, bonuses, paid vacation, and sick leave pay, among others. The coverage is 1985–2016.

Earnings are deflated using the CPI with base year 2018.

Sample Definitions

The sample is harmonized following guidelines by the GRID Project to facilitate cross-country comparison and focus on the core labor force, aged 25–55, over 1985–2016. Employment is defined as having total annual earnings of at least 1.5 times the monthly earnings at the retail minimum wage, and include all employed workers each year.

United Kingdom

Data Sources

The data source for the United Kingdom is the Annual Survey of Hours and Earnings (ASHE), formerly known as the New Earnings Survey (NES). This is the premier source of earnings information in the UK and forms the basis for many official wage statistics. It is a 1% sample of all employees, with a panel structure which makes it possible to follow workers over time.

The panel dataset is available back to 1975 and is administered by the Office for National Statistics (ONS). Workers enter the sample frame by having a particular pair of digits at the end of their National Insurance Number (NIN), the UK equivalent of a Social Security Number, randomly assigned to all workers upon labor market entry. Surveyors then identify the employer(s) of these individuals by Her Majesty's Revenue and Customs (HMRC) Pay As You Earn (PAYE) system, the UK government's income tax withholding system. This takes place each January.

Earnings include overtime pay and any bonuses related to the surveyed work week. Earnings are deflated by CPI 2018 base. To retain anonymity, where percentiles or percentile-based statistics are reported, data are binned into groups of 10 individuals and each individual is then assigned the group mean.

Sample Selection

A number of workers at the bottom of each year's earnings distribution is dropped, in order to select only those with reasonable labor market attachment. For years 1999 onwards, those with weekly earnings below 7x the minimum wage are dropped, corresponding to one day's work for a low-wage worker. For the years before then, the 1999 minimum wage is scaled by median real earnings growth to calculate a pseudo minimum wage.

United States

Data Sources

The data source for the United States is the Longitudinal Employer-Household Dynamics (LEHD) infrastructure files, developed, and maintained by the U.S. Census Bureau. Reporting covers private employers and state and local government. There are no self-employment earnings unless the proprietor draws a salary, which is indistinguishable from other employees in this case. UI earnings are reported by the firm every quarter. Although there is some variability across states, UI earnings generally include regular hourly earnings or salary, overtime, bonuses, reported tips, and sick/vacation pay. Federal workers participate in the unemployment insurance system, but their earnings are not included in the data.

The LEHD program is based on a voluntary federal-state partnership. When a state becomes a member of the partnership, current as well as all available historical data for that state are ingested into the LEHD internal database. By 2004, LEHD data represent the complete universe of statutory jobs covered by the UI system in the United States. To construct person-year level annual real (deflated by the Personal Consumption Expenditures Index (PCE), base 2018) earnings files covering the period 1998-2019 all jobs are used.

Sample Selection

The data set includes workers earning above an earnings floor $m_t = 260 \times \text{federal hourly minimum wage}(t)$ (about $1,900 in 2018) and there is also a ceiling imposed by winsorizing earnings at the 99.999999th quantile. The sample contains approximately 2 billion person-year earnings records; it includes all years and available states from 1998 to 2019; it includes workers each year that are age 25-55., while the reference year for sample 2 is 2010.

To meet Census Bureau disclosure avoidance standards for the common code cross-country earnings comparisons, the reported earnings values are replaced with the earnings from at least 10 adjacent persons. The various earnings and change in earnings variables are first calculated on the not-binned data at the person level prior to binning. Each earnings variable is then binned separately.