Data Appendix for
"Firm Age, Investment Opportunities, and Job Creation" *

Manuel Adelino
Fuqua School of Business
Duke University

Song Ma
Fuqua School of Business
Duke University

David T. Robinson
Fuqua School of Business
Duke University and NBER

September 21, 2015

1 The LEHD Quarterly Workforce Indicators Data

1.1 Overview

We use the publicly-available Quarterly Workforce Indicators (QWI) published by the Longitudinal Employer-Household Dynamics (LEHD) program of the U.S. Census Bureau to compute total employment by firm age.[1] This dataset provides total employment[2] in the private sector tabulated for 5 firm age categories—start-ups (0-1 year-olds), 2-3 year-olds, 4-5 year-olds, 6-10 year-olds, and firms over 11 years old. The totals are provided by county, quarter and industry, where industry is defined at the 2-digit National American Industry Classification System (NAICS) level. For the purposes of our analysis, and as we point out in the paper, we aggregate the county-level observations in each age category to the Commuting-zone level,[3] and we transform this quarterly data into annual data by using the data in the last quarter of each year.

The analysis focuses on firms in the non-tradable sector, namely Retail Trade (2-digit NAICS 44-45), and Accommodation and Food Services (2-digit NAICS 72). This definition matches the definition of non-tradable industries in Mian and Sufi (2012) as closely as is possible given that the LEHD data is not broken down by 4-digit NAICS industries.[4]

[*] Please address correspondence to Manuel Adelino, Fuqua School of Business, 100 Fuqua Drive, Durham, NC 27708.

[1] We use 2014Q4 release.

[2] We mainly use `empend`, which is the number of jobs on the last day of each quarter.

[3] We use the 2000 version of the commuting-zone definition, in which the United States is covered by 709 CZs.

[4] As a robustness check, we also perform our analysis on the age-sorted commuting zone-level employment for all

1.2 Missing Values and Coverage

There are missing values in the publicly-available QWI dataset for confidential protection. Specifically, the missing pattern is usually one or more missing employment values of firms in certain county-year-industry-age. To the best of our knowledge, we think there are two ways of handling the missing data. In the first method, we drop any county-year-industry as long as there are missing values in one or more age categories. In the second method, we impute the missing employment in a county-year-industry-age if we are given the total employment of certain county-year-industry and the employment in all other ages (by assuming this, we are implicitly assuming that the missing is from data construction errors rather than from confidential-protection procedures). The second method (imputation) will save around 15% of the data points. In the paper, we implement the first method (no imputation), as we consider this to be the most conservative way of construction. We also conduct the same set of analysis using the imputed data, and get both qualitatively and quantitatively similar results.

The QWI data coverage increases through time. The data set covers 18 states in 1995, 42 states in 2000 (the first year in our analysis), and 50 states (including the District of Columbia) in 2007 (the last year we consider). Massachusetts is not covered by the LEHD data. We use yearly observations of the data as of the fourth quarter of each year. When we have the most coverage, our dataset includes 555 commuting zones. We do not have full coverage of each CZ in the data. On average, the counties in the LEHD cover about 62% of the population of the CZs included in the dataset. Once a CZ enters the regression sample, we do not add counties to the CZ, even if they become available in the LEHD in later years. This avoids inflating the net employment growth from one year to the next by including new counties in a CZ that were not present before.

1.3 Constructing Job Creation from Employment Data

Net job creation data is constructed from the raw LEHD dataset by exploiting the mechanical transition of firms across firm age categories. As elaborated in Figure 1, the firms in the "start-up" category (0-1 year-olds) in year $t-2$ are the same firms in the 2-3 year-old category at t, conditional on having survived that far. The difference in the total number of jobs in these categories at $t-2$

industries and for construction-only (2-digit NAICS 23) in the LEHD. We get qualitatively similar and quantitatively close results.

and t represents the net job creation by these firms over the two years. There are, of course, firms that die, and this measure picks up this source of job destruction as well. Firms in the "2-3 year-old" category at $t-2$ move into the "4-5 year-old" category at t. Finally, firms in the "4-5 year-old", "6-10 year-old" and "11+ year-old" categories combined at $t-2$ will be the firms in the "6-10 year-old" and "11+ year-old" categories at t. The category "0-1 year olds" at time t includes firms that did not exist as of $t-2$ and this is our measure of job creation by newly formed firms over the 2 year period.

There is some room for debate about what constitutes a "new firm" or a "start-up". First, the data classifies subsidiaries of existing firms as start-ups, as long as they are separate legal entities. Second, a new McDonald's franchisee opening her first McDonald's location is also classified as a startup. Ideally, we would like to run our tests excluding these firms, but we are constrained by what the Census makes available. We are not aware of any dataset that would allow researchers to fully capture these sources of potential misclassification in the setting of young private firms (including the confidential version of the LEHD, or even confidential IRS data). By exploiting these transitions of firms across age bins, we can calculate the net job creation over each 2-year period in each CZ in the non-tradable sector for 4 firm age categories (age measured at observing)—start-ups (0-1 year-olds), 2-3 year-olds, 4-5 year-olds and 6+ year-olds.

In order to make the job creation data comparable across different areas, we need to divide the raw job creation data by a CZ size measure. We want this denominator to satisfy two conditions: first, this measure is a proper measure of the size of the non-tradable sectoral employment in the CZ thus the output ratio is immune to the effect of different CZ non-tradable business size; second, we want the size measure be comparable with each other thus not influenced by the fact that CZs enter the QWI database at different times. The denominator we choose is the total employment in non-tradable sector in the CZ as of 2000, calculated from the County Business Pattern (CBP) database.

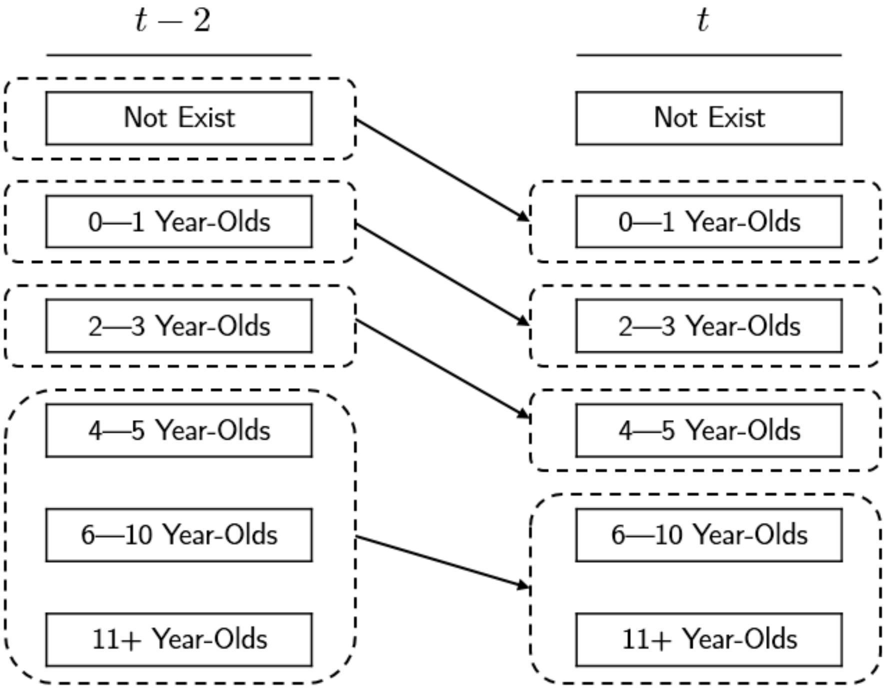

Figure 1: Mechanical Transition of Firms Across Firm Age in QWI Data

2 Investment Opportunities Measure and Its Instruments

2.1 Measuring Income Growth

We use the 2-year growth rate of total income in a commuting zone as our measure of changes in investment opportunities for the local non-tradable sector. Given that firms in this sector depend primarily on local demand, higher local income should create more opportunities for those businesses. We use a 2-year growth rate to be consistent with the 2-year net job creation data.

We obtain data from the Internal Revenue Service (IRS) County Income Data published by the Statistics of Income (SoI) Division. The data is processed by the SoI using the income tax returns from the IRS's Individual Master File, and is accessible as county-year observations.[5] In the database, we can observe both adjusted gross income (AGI) and wages and salaries (in total and per-capita format), and we deflate all those income level measure to 2007 US Dollars using the Consumer Price Index (CPI). We use real income growth as the measure of investment opportunities, and we

[5] For a detailed introduction of the data, please refer to http://trac.syr.edu/data/irs/countyIncome.html.

use “total wages and salaries” as the main income measure.[6]

2.2 Constructing the Instrument

We instrument for changes in local (CZ) income following Bartik (1991) and Blanchard and Katz (1992). Our main instrument is constructed by interacting changes in nationwide employment in the manufacturing sector with the preexisting sectoral composition of manufacturing in a local region. Our second instrument used for robustness check, is constructed in the same spirit as the main IV except exploiting the changes in import penetration as the exogenous shock, which is then interacted by the preexisting sectoral composition in the CZ. The exclusion restriction rests on the fact that the composition of the manufacturing sector is predetermined at the time of the income shock, and that nationwide shocks (or the import penetration) are exogenous to each individual CZ or county.

2.2.1 Bartik IV—Version Using National Manufacturing Shock

In order to construct the instrument we use County Business Pattern (CBP) data published by the Census that includes total employment for each 4-digit NAICS industry at the county and year level. We start by calculating the nationwide employment growth in each 4-digit NAICS industry in the manufacturing sector (codes between 3100 and 3399). We then create a “predicted” local growth in manufacturing employment by multiplying the national growth rate in each 4-digit NAICS industry by the local manufacturing sector composition at the beginning of the period.

Formally, the instrument is given by:

$$\widehat{\Delta_\tau e_{it}^m} = \sum_j \omega_{ij(t-\tau)} \times \Delta_\tau e_{jt} \tag{1}$$

where $\widehat{\Delta_\tau e_{it}^m}$ is the predicted growth rate in total manufacturing employment in CZ i between $t-\tau$ and t. This is calculated as the percent change in the number of jobs nationwide in the four-digit NAICS manufacturing sector j between $t-\tau$ and t, denoted Δe_{jt}, weighted by region i's ratio of jobs in that manufacturing sub-sector to overall employment as of time $t-\tau$, $\omega_{ij(t-\tau)}$. The degree of

[6] As a robustness check, we use the growth rate of per-capita wages and salaries as the income growth measure, the results are both qualitatively and quantitatively similar.

CZ income growth explained by the CZ-specific sector composition of manufacturing employment is our instrument for the income growth in a CZ.

2.2.2 Bartik IV—Version Using Import Penetration

This "Import Bartik" is constructed in the same vein as our main instrument (shown in Equation (1)), except that we replace the change in nationwide employment by industry $\Delta_\tau e_{jt}$ with the change in import penetration in each industry j (which we denote as $\Delta_\tau imp_{jt}$). The import penetration measure is constructed as the net imports (total imports minus total exports) over the total US shipments for each 4-digit NAICS manufacturing sub-sector in each year. This is then used as the shock to local manufacturing employment. The instrument is formally defined as

$$\widehat{\Delta_\tau e_{it}^m} = \sum_j \omega_{ij(t-\tau)} \times (-\Delta_\tau imp_{jt}) \tag{2}$$

because we implicitly assumes that a increasing import penetration will impede the employment in the domestic sector. We obtain import, export, and total shipments data at the 4-digit NAICS level from Peter Schott's webpage.[7]

3 FDIC Summary of Deposits Data

To study how access to finance interacts with firms' ability to pursue investment opportunities, we use the share of local banks in a CZ as a measure of local access to finance. We use the Summary of Deposits (SoD) data from the Federal Deposit Insurance Corporation (FDIC)[8] to compute the share of a bank's deposits that are located in a CZ. The SoD data is an annual survey of branch office deposits for all FDIC-insured institutions, and we can observe both geographical and business information for each branch. For the purpose of this paper, we are mostly relying on the information on the branch location (county) and the deposits in each branch.

Our main measure of the access to finance is the share of local banks in the CZ. A "local" bank is defined as one that has 75% or more deposits concentrated in one CZ (following Cortes (2013)). We then construct the local bank share of a CZ, defined as the share of all deposits in a CZ that are

[7] `http://faculty.som.yale.edu/peterschott/sub_international.htm`.

[8] FDIC SoD Data can be viewed and downloaded from http://www2.fdic.gov/sod/dynaDownload.asp?barItem=6.

held by banks local to that CZ. The identifying assumption is that, as shown by Petersen and Rajan (1994, 2002), small (local) banks are more likely to be able to lend to small firms, and especially so to more opaque firms. Lending to old (established) firms is likely to require less screening and monitoring than lending to new firms in an area, so young firms in CZs with a higher proportion of local banks are likely to have better access to financing. In order to mitigate the effect of labor market dynamics on the evolution of the local banking sector, we use a time-invariant CZ-level measure by calculating the time-series median of the deposit concentration in local banks for each CZ.

4 The Census Business Dynamics Statistics (BDS) Data

Though LEHD-QWI data described in Section 1 provides rich data on employment at very detailed level (county-sector-age), one downside of the database is that we do not observe the size of the firms. The Business Dynamics Statistics (BDS) data provided by the Census LBD program, however, provides employment data double sorted by age and size of the employer. This BDS-only feature enables us to study the joint age/size distribution.

The Business Dynamics Statistics (BDS) data can be downloaded from the Census website[9], and we are mainly exploiting the "*Firm Age by Firm Size by MSA*" version of the data. This data is at the MSA-year-size-age level, where both size and age are categorized at a very fine level.[10] In order to be consistent with the QWI database, we reorganize the BDS data into age categories of start-ups (0-1 year-olds), 2-3 year-olds, 4-5 year-olds, 6-10 year-olds, and firms over 11 years old. Net job creation in each age category is then calculated in the same way as in Section 1.3, in a *fixed* size bin. After repeating the algorithm in every size bin, we obtain the net job creation from firms at certain age-size pair, i.e., the observation is at MSA-year-age-size level. If we aggregate this age-size-sorted data by size, then we obtain an age-sorted BDS data covering the whole United States (better coverage than LEHD-QWI), and the observation is at MSA-year-age level.[11]

[9] http://www.census.gov/ces/dataproducts/bds/data_firm.html

[10] BDS provides age data at age 0, 1, 2, 3, 4, 5, 6-10, 11-15, 16-20, 21-25, 26+. The size is categorized by employee numbers 1-4, 5-9, 10-19, 20-49, 50-99,100-249, 250-499, 500-999, 1,000-2,499, 2,500-4,999, 5,000-9,999, and 10,000+.

[11] We calibrate the BDS data with the LEHD-QWI data in the paper. We construct MSA-level income growth measure and instruments, and perform the analysis on LEHD-QWI data and on BDS data separately. The results are quantitatively very similar.

5 Some Other Data

5.1 House Price Index and Saiz Elasticity

The housing prices used in robustness tests come from the Federal Housing Finance Agency (FHFA) House Price Index (HPI) data at the Metropolitan Statistical Area (MSA) level. The FHFA HPI is a weighted, repeat-sales index, and it measures average price changes in repeat sales or refinancings on the same properties. We use data on the MSA-level index between 1999 and 2007. As an alternative to using the change in housing prices during the period, we also use the housing supply elasticity measure developed by Saiz (2010). This measure varies at the MSA level and it is constructed using geographical and local regulatory constraints to new construction. This measure is available for 269 MSAs that we match to 776 counties using the correspondence between MSAs and counties for the year 1999 as provided by the Census Bureau, and then aggregate up to the CZ level.

References

Bartik, Timothy J, 1991, Who benefits from state and local economic development policies?, *Books from Upjohn Press* .

Blanchard, Olivier Jean, and Lawrence F Katz, 1992, Regional evolutions, *Brookings papers on economic activity* 1992, 1–75.

Cortes, Kristle Romero, 2013, New firms, job creation and access to local finance, *Federal Reserve Bank of Cleveland Working Paper* .

Mian, Atif R, and Amir Sufi, 2012, What explains high unemployment?—the aggregate demand channel, *National Bureau of Economic Research Working Paper* .

Petersen, Mitchell A, and Raghuram G Rajan, 1994, The benefits of lending relationships: Evidence from small business data, *The Journal of Finance* 49, 3–37.

Petersen, Mitchell A, and Raghuram G Rajan, 2002, Does distance still matter? the information revolution in small business lending, *The Journal of Finance* 57, 2533–2570.