Online Appendix for The Origins of Top Firms

Rafael Guntin *University of Rochester*

Federico Kochen *CEMFI & Banco de España*

A Data Appendix

A.1 The Orbis Database

This appendix provides definitions and measurement details about the Orbis database. We primarily focus on Spain, the country we use for our baseline results.

A.1.1 Variable Definitions and Measurement

Firm Age. We measure firm i's age at time t as: $age_{it} = t - \tau_{i0}$, were τ_{i0} is firm i year of incorporation. To account for incomplete reporting spells at entry, if firms' first year has a time spell of less than 12 months we define $age_{it} = t - (\tau_{i0} + 1)$, and drop the first negative age observation.[45] For some companies, the incorporation year in Orbis and the actual foundation year may differ because of changes in the company's legal name due to restructurings, or mergers and acquisitions (M&A). In addition, as we focus on unconsolidated accounts, there could be cases of subsidiaries of older parent companies that are not truly new firms. As explained in Appendix A.2 below, we address this issue by searching for the true foundation year of all the top 1.5 percent firms at age 20 in our sample using a large language model (LLM)-assisted search of publicly available information.

Wage bill. Regarding firms' income statement variables, we measure firms' wage bill as

$$wl_{it} = \texttt{staf}_{it}$$

where $\texttt{staf}_{it}$ are total labor costs of firm i at year t, using Orbis acronyms. For what follows, we follow BvD acronyms to refer to the variables in Orbis Historical.

Output. We measure firms' output as value-added using a comprehensive measure of costs. In detail, we measure the output of firm i at time i as the sum of EBITDA (earnings before interest, taxes, depreciation, and amortization), acronym `ebta`, and labor costs

$$\begin{aligned} y_{it} &= \texttt{ebta}_{it} + \texttt{staf}_{it} \\ &= (\texttt{oppl}_{it} + \texttt{depr}_{it}) + \texttt{staf}_{it} \\ &= (\texttt{opre}_{it} - \texttt{cost}_{it} - \texttt{oope}_{it}) + \texttt{depr}_{it} + \texttt{staf}_{it} \end{aligned}$$

[45] For most companies, data is recorded annually in December. Hence, e.g., if a firm is founded in August 1991, we drop the first observation in December 1991, corresponding to only 4 months of operation, and define $age_{it} = t - (1991 + 1)$, for $t \geq 1992$.

$$= \underbrace{\texttt{opre}_{it}}_{\text{Operating revenue}} - \underbrace{[(\texttt{cost}_{it} + \texttt{oope}_{it}) - \texttt{staf}_{it} - \texttt{depr}_{it}]}_{\text{Costs net of labor and capital}} \tag{11}$$

where, as the last expression in (11) shows, the definitions of the variables in Orbis imply that our measure of output is equal to the sum of operating revenue (`opre`) minus a comprehensive measure of costs excluding those related to labor and capital: costs of goods sold (`cost`) net of labor costs (`staf`) and capital depreciation (`depr`) plus other operating expenses (`oope`).[46] We set to missing the few observations with negative output values.

Capital, Equity, and Debt. It is important to note that capital and the rest of the balance sheet variables are reported at the end of each year. Hence, as in Hsieh and Klenow (2009), we measure the stock of the variable at t as the average between the end-of-year values of $t-1$ and t. For our baseline results, we follow Kochen (2025) and measure capital as equity plus net financial debt

$$k_{it} = a_{it} + b_{it}.$$

Using Orbis acronyms, equity is equal to

$$a_{it} = (0.5 + 0.5L) \times (\texttt{toas}_{it} - \texttt{culi}_{it} - \texttt{ncli}_{it})$$

where `toas` denotes total assets, `culi` is current liabilities, `ncli` is non-current liabilities. The term $(0.5 + 0.5L)$ indicates the average between current and previous year balance sheet variables, where L is the lag operator.

We measure firms' net financial debt as

$$b_{it} = (0.5 + 0.5L) \times (\texttt{loan}_{it} + \texttt{ltdb}_{it} - \texttt{cash}_{it})$$

wher `loan` is short-term financial debt (payable within a year), `ltdb` is long-term financial debt, and `cash` denotes the firm's cash and cash equivalents.

In addition to our baseline definition of capital, we also present results for capital measured by tangible fixed assets, $(0.5 + 0.5L) \times \texttt{tfas}_{it}$, and alternatively by the sum of tangible and intangible fixed assets, $(0.5 + 0.5L) \times (\texttt{tfas}_{it} + \texttt{ifas}_{it})$. As Appendix A.2 of Kochen (2025) shows, our baseline measure of capital is broadly equal to the sum of tangible and intangible assets plus inventories (`stok`).

Profits. Our baseline analysis focuses on a model-consistent definition of economic profits. Noting that the firm's capital is the sum of equity plus debt, $k_{it} = a_{it} + b_{it}$, we can rewrite economic profits, defined in (4), as

$$\begin{aligned} \pi_{it} &= y_{it} - wl_{it} - Rk_{it} \\ &= y_{it} - wl_{it} - \delta k_{it} - rb_{it} - ra_{it} \end{aligned} \tag{12}$$

[46] This definition of output coincides to that used in Boar, Gorea, and Midrigan (2023), with the exception that we do not subtract for taxes.

where the rental cost of capital is $R = r + \delta$. We take the definition of economic profits in (12) to the data by setting $\delta = 0.05$ and $r = 0.02$, as in the model, and then directly measuring output y_{it}, labor costs wl_{it}, capital k_{it}, and equity a_{it} as defined above. Lastly, we measure the cost of debt rb_{it} using net financial expenses: $-\texttt{fipl}_{it} = \texttt{fiex}_{it} - \texttt{fire}_{it}$, where `fiex` are financial expenses and `fire` financial revenue.

In addition to the analysis using economic profits, we present results for profits measured by net income of the year, available in the data and equal to

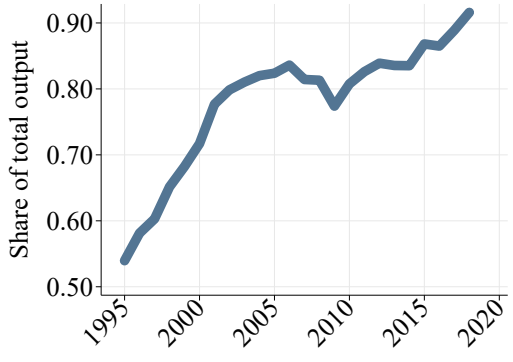
$$\pi_{it} = \texttt{pl}_{it} = \texttt{plat}_{it} + \texttt{extr}_{it}$$

where `plat` is profits after taxation and `extr` is extraordinary and other profit.

A.1.2 Coverage and Representativeness for Spain

Coverage. To analyze the coverage of Orbis Historical, we follow Kalemli-Özcan et al. (2024) and compute the share of total revenue in our data relative to the *OECD STAN Database*. As we study the non-financial private sector, we focus on matched two-digit ISIC and year pairs in both Orbis and OECD STAN. Figure A.1 reports the share of total revenue (turnover) in our data relative to the official statistics in OECD STAN. The figure shows that coverage increases by year. The firms in our sample represent 54% of aggregate revenue in 1995 and 92% by 2018. On average, across all years, Orbis Historical covers 78% of total revenue in Spain, considering the matched two-digit sectors.

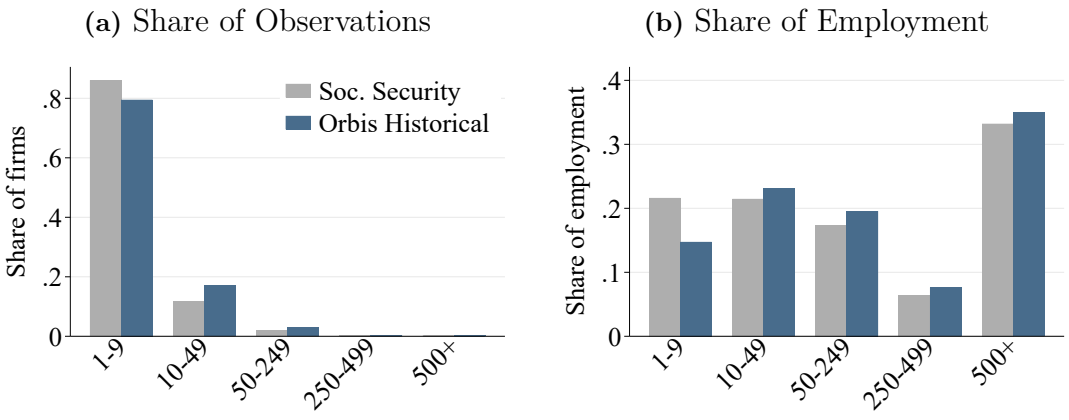
Figure A.1: Orbis Historical Share of Spain Total Revenue



Notes: Share of total revenue relative to OECD STAN data for matched 2-digit sector-year pairs.

Size Distribution. To test the representativeness of the data in terms of the size distribution, we compare the share of firms and share of total employment by firm size in Orbis to official statistics from the Spanish Social Security (*Estadística de Empresas Inscritas en la Seguridad Social*). As before, we focus on employer firms. Figure A.2a reports the share of observations by different employment bins in these two datasets for 2019. Figure A.2b presents the share of total employment. These two panels show that while Orbis slightly underrepresents the smallest firms (1-9 employees), it captures well the full employment distribution.

Figure A.2: Employment Distribution in Orbis Historical and Social Security Data



Notes: Shares for Spain in the year 2019 considering employer firms. Soc. Security data comes from Estadística de Empresas Inscritas en la Seguridad Social published by the Spanish Labor Ministry.

A.1.3 Firm-Owner Linkages

As Section 3.5 describes, we present a taxonomy for top and bottom firms given their owners' characteristics. We use firm-owner linkages in Orbis Historical, available from 2007 onward, which include annual records with owners' names and equity shares. The data also specifies whether the owner is an individual, another company, a financial institution, or a government. Given firms' sometimes complex ownership structures, we identify firms' ultimate owners by sequentially matching the ownership files. The main objective of this iterative procedure is to account for the cases where other companies own firms by assigning the respective ownership shares of the parent company.[47] As we show below, solely studying direct owner linkages would underestimate the importance of foreign and multiple firm owners.

A.2 Foundation-Year Data

The year of incorporation of the current legal entity reported in Orbis may differ from the firm's true foundation year due to corporate restructurings, mergers and acquisitions (M&A), or changes in legal form due to regulation. In addition, because we focus on unconsolidated accounts, subsidiaries of older parent companies may appear as young firms even though they are part of the parent firm's core operations. These discrepancies can be more common among the largest firms in the economy. Since our life-cycle analysis requires an accurate measure of firm ages, we supplement the incorporation year in Orbis with newly collected foundation-year data for all top 1.5 percent firms in our sample at age 20. For Spain, we searched for a total of 2,615 companies, including an initial sample of 1,740 from the top 1 percent. Whenever the newly collected foundation year differs from the incorporation year, we use the true foundation year to compute firm age.

[47] Hence, for example, if 100% of company A is owned by company B, the procedure assigns the ownership information of B to A. If x% of company A is owned by company B, only x% of the ownership of B will be transferred to company A.

Examples. To illustrate the importance of properly measuring companies' foundation years, we present examples we identified in our analysis. For this Appendix, we focus on Spain, but similar considerations apply to the other countries in our sample.

- *Corporate Restructuring and Subsidiaries.* One reason for corporate restructuring is that, as companies grow, they divide their operations into separate legal entities. For example, "Telefónica de España S.A.U.", the largest telecommunications company in Spain, separated its domestic and international operations thirty years ago. Hence, in the data, we observe an incorporation year in the late 1990s. However, according to the company website, this company was founded in 1924. Another example in the data is a large energy producer founded in the 1900s that has two relatively recent legal entities, one specializing in energy generation and the other in distribution. In both of these examples, we assigned the true foundation year to be that of the older parent company.
- *M&A.* Another common case is Spanish-founded companies that change legal entities, and sometimes brands, when foreign companies acquire them. In this case, we kept the foundation year of the original Spanish company that was later acquired. For mergers that result in a new legal entity, we set the foundation year to the oldest foundation year of the two pre-merge companies.
- *Regulation.* Regulatory changes can also drive companies to change their legal entities. In our sample, we identified eight well-known Spanish football clubs, most founded at the beginning of the 20th century, that had to convert to Sociedad Anónima Deportiva (SAD) in the early 1990s because of Spain's 1990 Sports Law (Ley del Deporte 10/1990).

LLM-Assisted Search. Given the scale of this task, which involves searching for information on more than 2,600 firms, we employed Anthropic's *Claude Opus 4.6* large language model (LLM) as the primary research assistant. The LLM was given access to web search capabilities and instructed to query company websites, Wikipedia, LinkedIn company pages, Spanish business registries (Informa D&B, Axesor, eInforma), the official gazette of the commercial registry (BORME), and press archives. Each company was researched individually, and the LLM was required to document the source for every verified founding year. We conducted this process under continuous supervision, working on batches of 100 companies at a time. In addition to the foundation year, we asked the LLM to report whether the firm originated in Spain, if it is foreign-owned, and whether a parent company owns it. In those cases, it also had to report the parent's name and foundation year, and whether the company is independent of the parent's core operations, in accordance with the rules described below. Once the LLM retrieved all the information, we "manually" verified all companies with discrepancies between their incorporation and foundation years. We also contrast this information with that available in Orbis for firms' `overview` and `history`, which, in some cases, include the true foundation year, to validate the results of our LLM-assisted search. Finally, for cases in

which the LLM foundation year was the same as the incorporation year in Orbis, we manually double-checked that the information was correct for a sample of randomly selected firms.

Classification Rules. When reporting the previous variables, we instructed the LLM to apply the following classification rules that distinguish among the different types of firms in the data. For standalone Spanish domestic companies, the verified foundation year records the true operational founding year of the business, even when this predates the current legal entity by decades. For foreign multinational companies that established a Spanish subsidiary from scratch, we measure firm age since entering into the Spanish market; the foundation year therefore records the year the Spanish entity was incorporated, while the parent company's founding year is recorded separately. For Spanish-founded businesses later acquired by foreign companies, the original Spanish founding year is preserved as the foundation year to reflect the true age of the business activity. When a parent company was itself formed by merging older entities, the founding year of the oldest constituent is used.

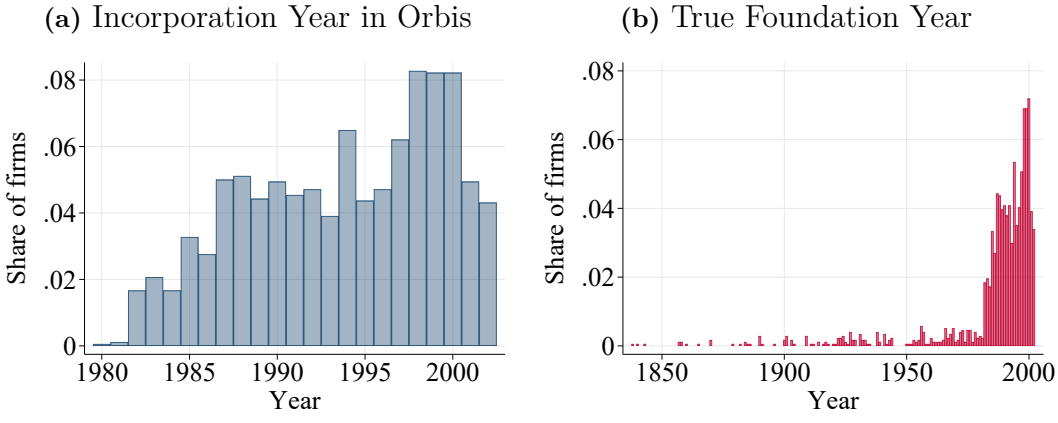
Business Groups and Subsidiaries. Equipped with the verified foundation years of the top companies at age 20, along with additional variables obtained from the LLM-assisted search, we need to determine the foundation years of companies within business groups (i.e., that are owned by a parent company). For Spanish business groups, we treat the companies that are independent of the parent company's core activities as different firms, with their own foundation years. When they are not independent and are merely an organizational unit within the group, we set the subsidiary's foundation year to match that of the parent.

We asked the LLM to determine whether a company is independent using the following four criteria. 1) Distinct market-facing activity: if the entity's name combines the parent name with functional suffixes (e.g., Finance, Leasing, Capital, Services, etc.), it is not considered an independent company. 2) Brand identity: if the company name corresponds to a recognized separate brand, it is considered an independent firm. 3) Economic purpose: if the entity serves primarily as a legal or financial vehicle for the group, it is not considered independent. 4) Standalone viability: if the parent could sell the company as a standalone business unit, it is considered an independent firm. A firm is then classified as independent when the preponderance of these criteria points toward a distinct, self-standing firm. While we classify as separate, independent firms those that are part of a business group satisfying the previous criteria in our baseline analysis, the bottom panel of Figure A.18 shows that our main results are robust to excluding all firms owned by multiple firm owners.

Foreign Multinationals. For foreign multinationals with origins outside Spain, we use the Spanish subsidiary's date of incorporation as the foundation year, as our interest is in measuring firm age since entering the domestic market. Yet, the top panel of Figure A.18 shows that our key findings are robust to excluding all foreign-owned firms.

Results. Following these steps, from the initial sample of 1,740 firms in the top 1 percent at age 20, we corrected for the foundation year and excluded 291 (16.7%) from this group. For those 291 cases, the difference in foundation years was considerable, averaging 46 years. Figure A.3a presents the cohort distribution of the initial sample using the incorporation year in the data. Given our sample selection, focusing on firms we observe at age 20 and can track back for at least 10 years, these are firms with cohorts between 1981 and 2002. Figure A.3b shows the distribution of corrected foundation years. It shows that several firms date back to the 1800s and the first half of the 1900s, underscoring the importance of accurately measuring the foundation years of top firms. Once we excluded companies with an incorrect measure of firm age, we recomputed the firm size distribution conditional on age and defined the final sample as the top 1 percent of firms at age 20.

Figure A.3: Cohort Distribution of Initial Sample of Top 1 at Age 20



Notes: Distribution of incorporation and foundation years for the initial sample of 1,740 top 1 percent firms at age 20 in Spain. Incorporation year is the one reported in Orbis. Foundation year is our newly collected data obtained from the LLM-assisted search.

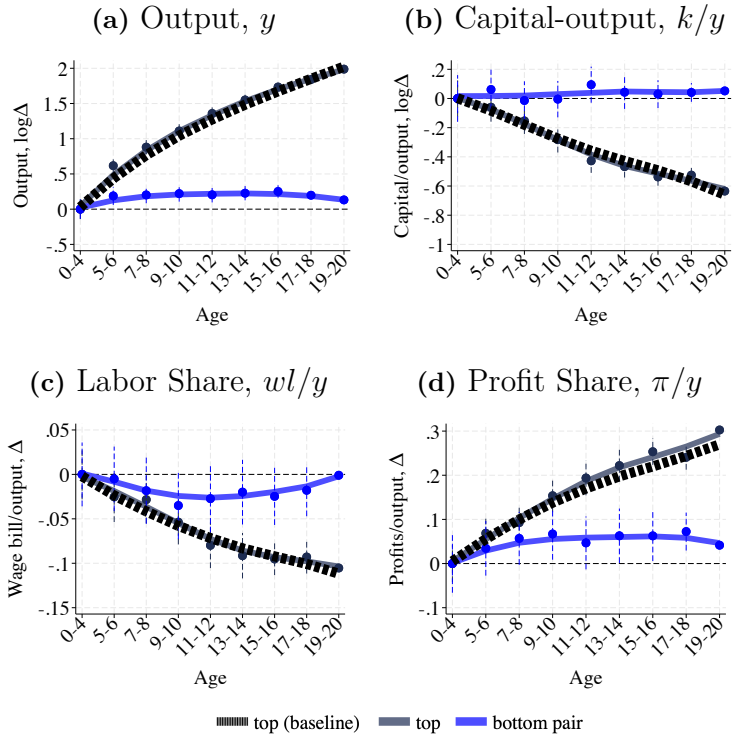
A.3 Top Firms' Pairs and Predictors

In this appendix, we study firms that are similar to top firms in their early years (i.e., bottom pairs) but end up in the bottom 20 years later. Furthermore, we examine which observable characteristics predict whether firms are top at age 20.

Pairs Analysis. Using a propensity score matching approach, we match each top firm with the closest bottom firm in its initial years (0-4 years). The variables used for matching are output level (y), input usage (k/y, wl/y, and k/wl), profitability (π/y), and financing (b/k, $\mathcal{I}_{\{d<0\}}d/k$, b/y, $\mathcal{I}_{\{d<0\}}d/y$), after removing sector and year fixed effects. We condition on bottom firms that survive to age 20 to perform the life cycle analysis. To obtain the propensity scores, we run a logit regression using these variables and then match each observation to its nearest neighbor, ensuring that propensity scores differ by no more than 0.05. This procedure matches a relatively large sample of more than 350 top firms to bottom counterparts. Note that each bottom firm can be matched to more than one top firm (this occurs for roughly

15% of paired bottom firms).[48]

Figure A.4: Life Cycle of Top Firms and Their Bottom Pairs



Notes:Life cycle trajectories of inputs and profit shares for the top 1% and their pairs in the bottom 99% of firms at age 20, estimated using (1), with the omitted age group being [0,4]. The regression for the output and capital-output ratio is in logs, while the labor and profit shares are in levels. The solid lines represent smoothed scatterplots generated through locally weighted regressions. The thick black dashed line is the baseline profile for top firms. The dashed vertical lines indicate 95% confidence intervals considering firm-level clustered standard errors.

We estimate an empirical model based on (1) for the top and bottom pairs. Figure A.4 presents the results for output, input usage, and profits. Panel (a) shows that, as expected, the output growth of bottom firms that were similar to top firms at the start remains flat. Panels (b) and (c) show that their capital-output ratio and labor share are roughly flat as well. Consequently, in panel (c), the profit share is much flatter. Furthermore, the matched top firms behave very similarly to our baseline estimates. This pattern, like our baseline result in Section 3, is consistent with large initial capital-output ratios and labor shares reflecting input-specific fixed costs that become relatively smaller as firms grow. Consistent with this interpretation, initially similar but unsuccessful firms (the bottom pairs) exhibit much flatter trajectories in input usage and profit backloading than the successful ones (top firms).

[48]In the first step, we estimate a logit model for the probability that firm i becomes a top firm by age 20: $\Pr(\text{top}_i = 1 \mid \mathbf{x}_{i0}) = \mathcal{L}(\beta' \mathbf{x}_{i0})$, where top_i indicates whether firm i is a top firm at age 20 and $\mathbf{x}_{i0}$ is a vector of firm characteristics measured at the initial age group. Using the estimated probabilities, $\hat{p}_i = \mathcal{L}\left(\hat{\beta}' \mathbf{x}_{i0}\right)$, we first identify, for each top firm i, the nearest bottom firm in propensity-score space: $j^*(i) = \arg\min_{\{j \,:\, \text{top}_j = 0\}} |\hat{p}_i - \hat{p}_j|$. We keep this match only if the propensity-score distance is below the caliper, $\left|\hat{p}_i - \hat{p}_{j^*(i)}\right| < 0.05$. Top firms that do not satisfy this condition are left unmatched.

Table A.1: Predictors of Top Firms

	$\beta: \ \mathcal{P}(\beta \mathbf{x}_1) = \text{pr}[\text{top} = 1 \mid \mathbf{x}_1]$	
	(a) Baseline	(b) Including Exit
$\ln y$	0.738***	0.567***
	(0.000)	(0.000)
$\ln k/y$	0.072***	0.114***
	(0.002)	(0.000)
wl/y	0.175	0.112
	(0.143)	(0.213)
π/y	-0.275***	-0.149**
	(0.000)	(0.016)
b/k	0.011	-0.086***
	(0.737)	(0.001)
$\mathbb{I}_{\{d<0\}}\lvert d\rvert/k$	0.914***	0.625***
	(0.000)	(0.000)
Observations	93,818	257,251
Pseudo R-2	0.40	0.33

Notes: The table shows the coefficients β and their p-values in parentheses for the joint probit model, where the outcome variable indicates whether the firm is top at age 20. The predictors correspond to observable characteristics between ages 0 and 4, residualized on year and sector fixed effects. Column (a) conditions on firms that survive to age 20 (baseline); column (b) includes firms that exit before age 20.

Predictors. Next, we examine what initial observable characteristics predict whether a firm becomes a top firm at age 20. Based on our main empirical findings, we consider six predictors measured between ages 0 and 4: log output, the capital-output ratio, the labor share, profitability, leverage, and equity injections.[49] Each variable is residualized by projecting it onto year and sector fixed effects. We estimate a probit model in two samples: first conditional on firms surviving to age 20, and then including firms that exit before reaching that age. Table A.1 reports the results. In the joint probit estimation, we find that larger initial size, a higher capital-output ratio, lower profitability, and greater equity injections are associated with a higher probability of becoming top, while the labor share is not significant. These patterns hold whether we condition on firms surviving to age 20 or include those that exit before, though leverage enters negatively only in the latter sample. The large R^2 is mostly explained by initial size ($\ln y$), which is naturally a strong predictor of future top status and is consistent with substantial ex-ante heterogeneity across firms.

These findings are broadly consistent with our baseline results in Section 3, showing that top firms, relative to other firms, start larger, have higher capital-output ratios, similar labor

[49]Note that we use debt and equity injections relative to capital rather than output. Since they are the counterpart of investment, debt and equity injections are highly correlated with capital. Scaling them by output would mechanically reduce the predictive power of the capital-output ratio.

shares, lower profitability, and rely more on external financing when young.

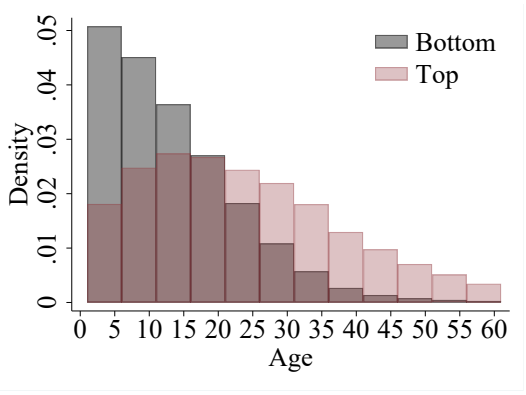
A.4 Additional Figures and Tables

Table A.2: Descriptive Statistics

	Bottom 99		Top 1	
	At age 20	Overall	At age 20	Overall
Panel (a): Averages				
Age	15.5	12.7	15.8	26.9
Output (USD millions)	0.7	0.5	30.9	49.9
Sales (USD millions)	2.6	2.1	118.3	196.3
Employment	14.0	11.8	518.8	629.3
Capital-output	2.65	2.85	2.69	3.38
Labor share	0.67	0.68	0.58	0.52
Profit share	0.15	0.13	0.24	0.26
Debt-output	0.61	0.78	0.93	1.23
Equity injections-output	0.02	0.02	0.02	0.01
Leverage	0.24	0.28	0.35	0.37
Panel (b): Share of Firms				
Manufacturing	0.33	0.32	0.21	0.27
Retail and Transport	0.35	0.31	0.28	0.28
Services	0.28	0.33	0.46	0.40
Other	0.04	0.04	0.06	0.05
Listed Firms	0.00	0.00	0.01	0.03
Public Firms	0.15	0.13	0.50	0.54
Private Firms	0.85	0.87	0.49	0.43

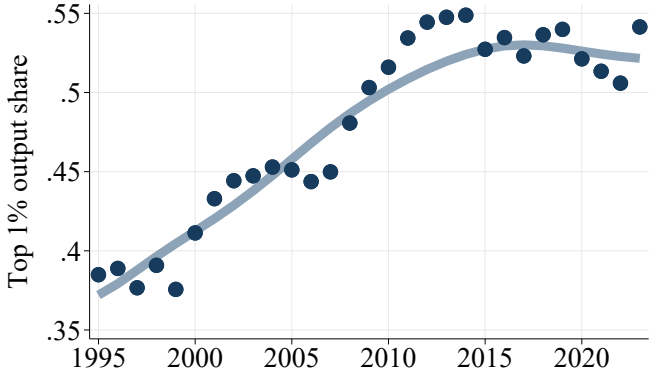
Notes: Descriptive statistics for two samples of top 1 and bottom 99 percent firms: "At age 20" denotes the firms used in our longitudinal sample and classified according to their size at age 20; "Overall" considers the full sample and classifies firms, year by year, using the full size distribution. Variables in dollars are in 2015 USD, using constant prices and exchange rates. The average capital-output ratio, the labor share (labor costs over output), profit share (profits over output), debt-output, and equity injections over output are means weighted by output. Sectors are based on NACE 2-digit classifications: Manufacturing is `nace` $\in [10, 33]$ and includes Construction `nace` $\in [41, 43]$, Retail and Transport is `nace` $\in [45, 53]$, and Services excluding Retail and Transport is `nace` $\in [55, 82]$, `nace` $\in [85, 88]$, and `nace` $\in [90, 96]$. Other corresponds to those sectors not listed before. Listed Firms are those publicly listed throughout our sample. Public Firms are public limited companies (*Sociedades Anónimas*), analogous to U.S. C-corporations. Private Firms are partnerships and private limited companies (*Sociedad de Responsabilidad Limitada*). See Appendix A.1 for definitions and measurement of these variables.

Figure A.5: Age distribution: all and top firms



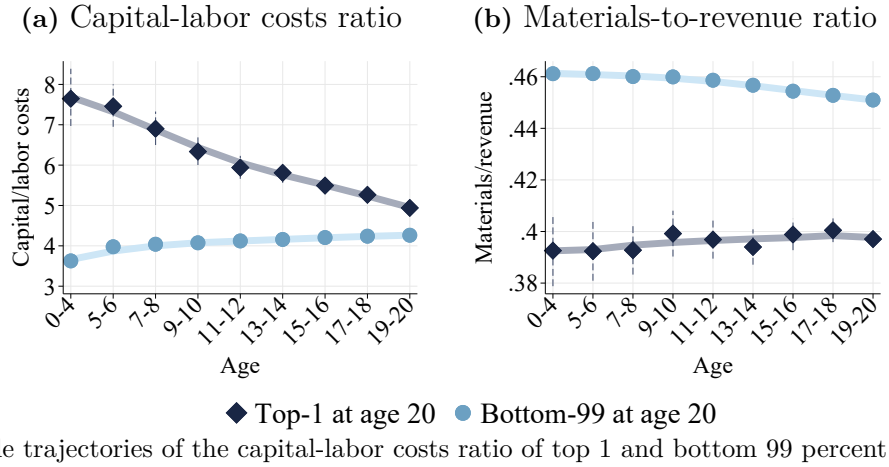
Notes: The figure shows the age distribution for all the firms and conditional on top firms.

Figure A.6: Output Share of Top 1 Percent Firms Over Time



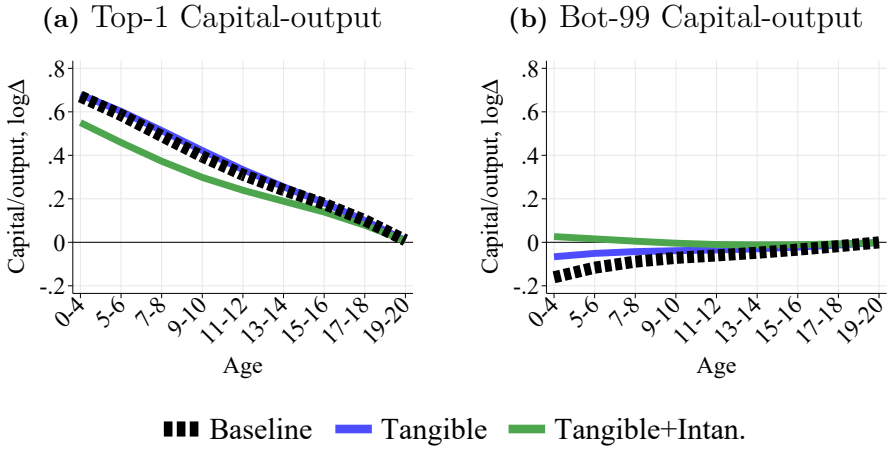
Notes: Top 1 percent separately defined for each year controlling for sector fixed effects. Results for Spain using Orbis Historical. Firm-level output is measured as value added, defined as revenue minus a comprehensive measure of non-capital and non-labor costs.

Figure A.7: Additional Input Variables of Top and Bottom Firms



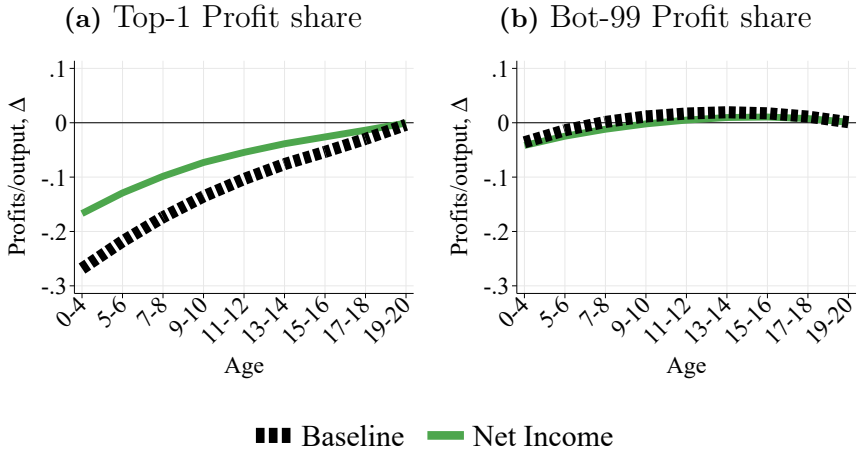
Notes: Life cycle trajectories of the capital-labor costs ratio of top 1 and bottom 99 percent firms at age 20 are estimated using (1). Results are scaled by adding the average of the omitted group (age 19-20) for the top and bottom firms. The solid lines are smoothed scatterplots using locally weighted regressions.

Figure A.8: Alternative Measures for Firms' Capital



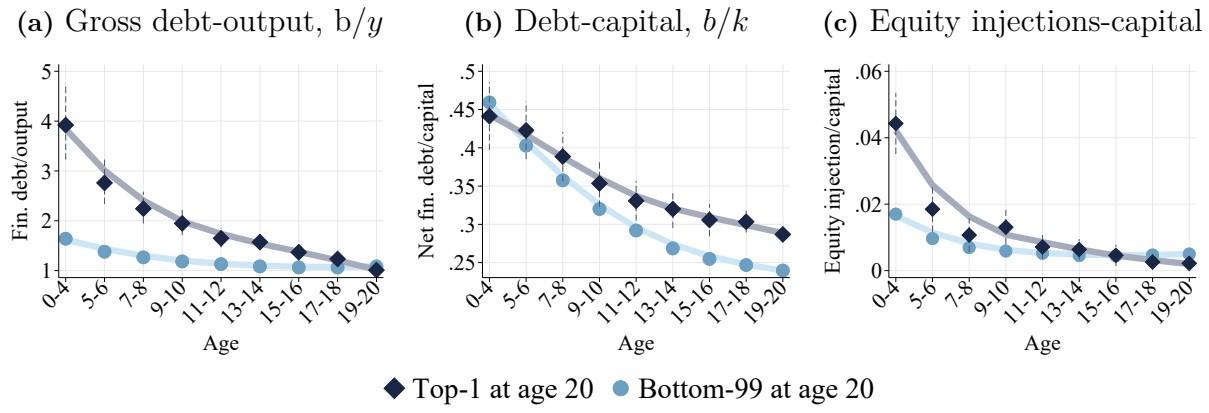
Notes: Life cycle changes for the capital-output ratio of top 1% and bottom 99% firms at age 20 considering two alternative capital measures: tangible capital (`tfas`); and sum of tangible and intangible capital (`tfas`+`ifas`). Changes are relative to the omitted group (age 19-20). The dashed dark line denotes the baseline estimation. All lines are smoothed scatterplots generated through locally weighted regressions.

Figure A.9: Alternative Measure for Firms' Profits



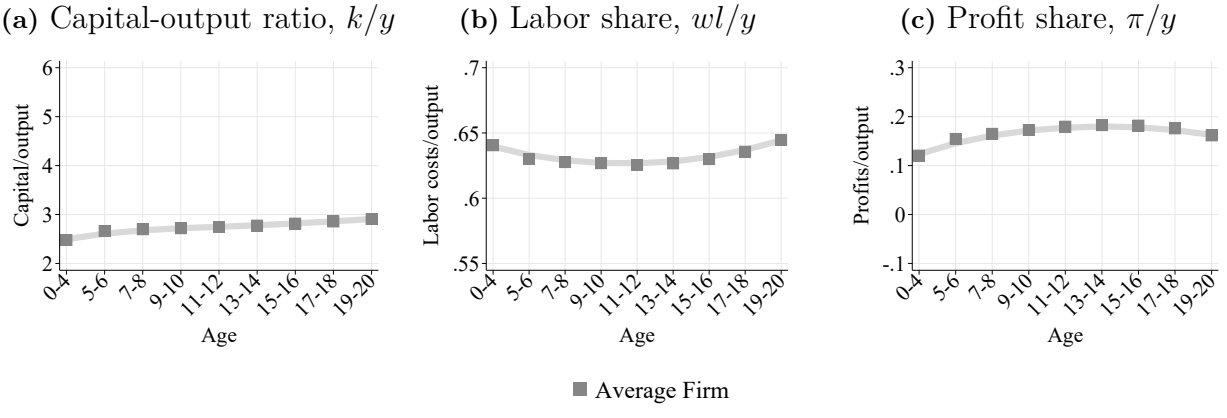
Notes: Life cycle changes for the profit share of top 1% and bottom 99% firms at age 20 considering firms' net income (variable `pl` in Orbis) as alternative measure for profits. Changes are relative to the omitted group (age 19-20). The dashed dark line denotes the baseline estimation. All the lines are smoothed scatterplots generated through locally weighted regressions on the estimated parameters.

Figure A.10: Additional Financing Variables of Top and Bottom Firms



Notes: Life-cycle trajectories of gross debt (financial debt) over output, net debt (financial debt minus cash) over capital, and equity injections over capital for top 1 percent and bottom 99 percent firms at age 20, estimated using (1). Results are scaled by adding the average of the omitted group (age 19-20) for the top and bottom firms. The solid lines represent smoothed scatterplots generated through locally weighted regressions. The dashed vertical lines indicate 95% confidence intervals considering firm-level clustered standard errors.

Figure A.11: Inputs and Profit Share Over the Life Cycle of the Average Firm



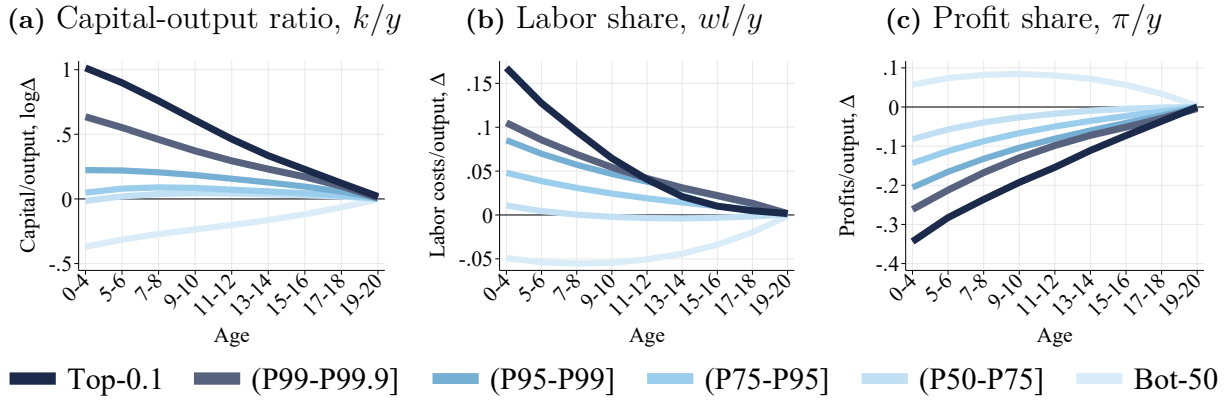
Notes: Life cycle trajectories for the inputs and profit shares of the average firm estimated using a version of (1) putting together all the firms that survive to age 20. The regression for the capital-output ratio is in logs, while the labor and profit shares are in levels. Results are scaled by adding the average of the omitted group (age 19-20). The solid lines represent smoothed scatterplots generated through locally weighted regressions. The dashed vertical lines indicate 95% confidence intervals considering firm-level clustered standard errors.

Table A.3: Top Status Persistence in Data and Model

		Panel (a): Data			**Panel (b): Model**		
			$t+10$			$t+10$	
			Top-1	Bot-99		Top-1	Bot-99
Age=5 at t	t	Top-1	0.736	0.264	t	Top-1	0.734	0.266
		Bot-99	0.006	0.994		Bot-99	0.007	0.993
			$t+10$			$t+10$	
			Top-1	Bot-99		Top-1	Bot-99
Age=10 at t	t	Top-1	0.747	0.253	t	Top-1	0.739	0.261
		Bot-99	0.007	0.993		Bot-99	0.005	0.995
			$t+10$			$t+10$	
			Top-1	Bot-99		Top-1	Bot-99
Age=20 at t	t	Top-1	0.772	0.228	t	Top-1	0.746	0.254
		Bot-99	0.011	0.989		Bot-99	0.004	0.996
			$t+10$			$t+10$	
			Top-1	Bot-99		Top-1	Bot-99
All ages	t	Top-1	0.768	0.232	t	Top-1	0.726	0.274
		Bot-99	0.009	0.991		Bot-99	0.006	0.994

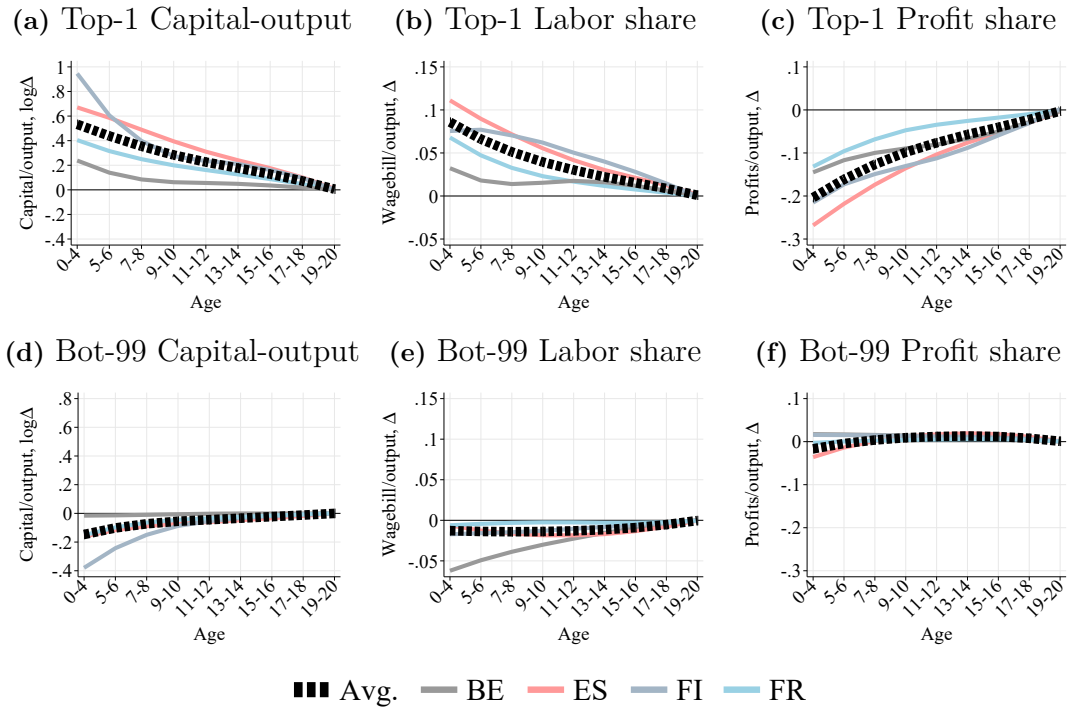
Notes: Transition probabilities matrices for top firm status between years t and $t+10$, conditional to different ages in year t and unconditionally (All ages). In all cases, we define top firms as top if they are in the top 1 percent of the overall firm size distribution.

Figure A.12: Life Cycle Changes in Inputs and Profit Share by Size at Age 20



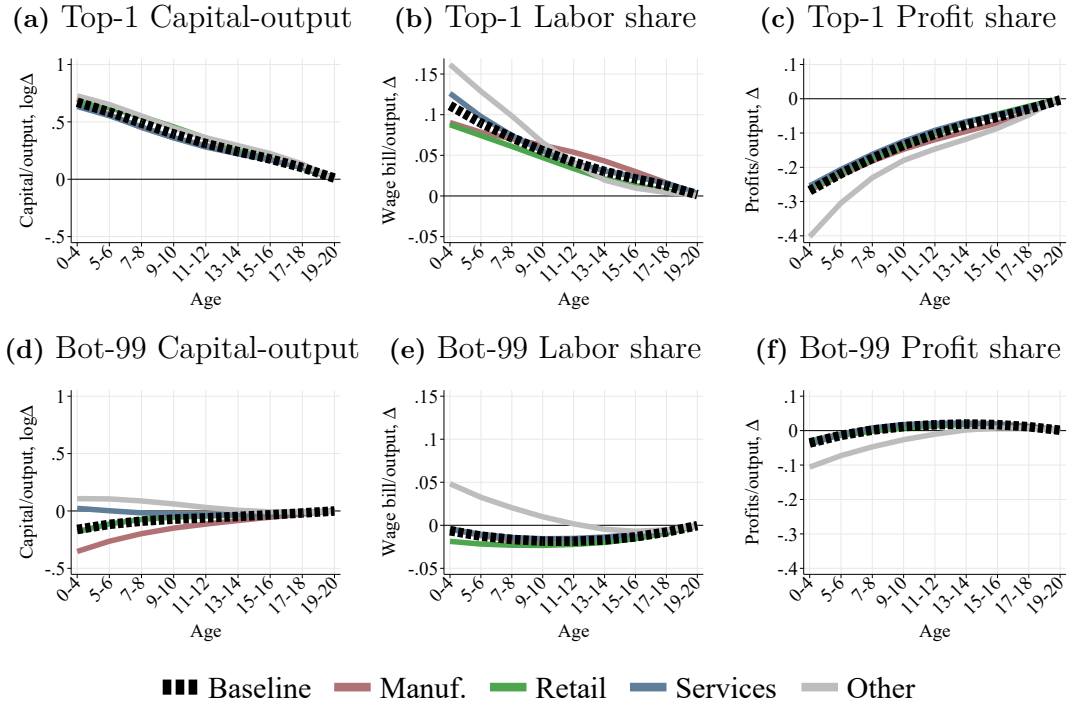
Notes: Life cycle changes for the inputs and profit shares of the top 0.1, percentile 99 to 99.9, percentile 95 to 99, percentile 75 to 95, percentile 50 to 75, and the bottom 50 percent firms at age 20. Life cycle changes are estimated using a version of regression (1) with these six groups of firms. Changes are relative to the omitted group (age 19-20). The lines are smoothed scatterplots generated through locally weighted regressions.

Figure A.13: Life Cycle Changes of Top and Bottom Firms in Other Countries



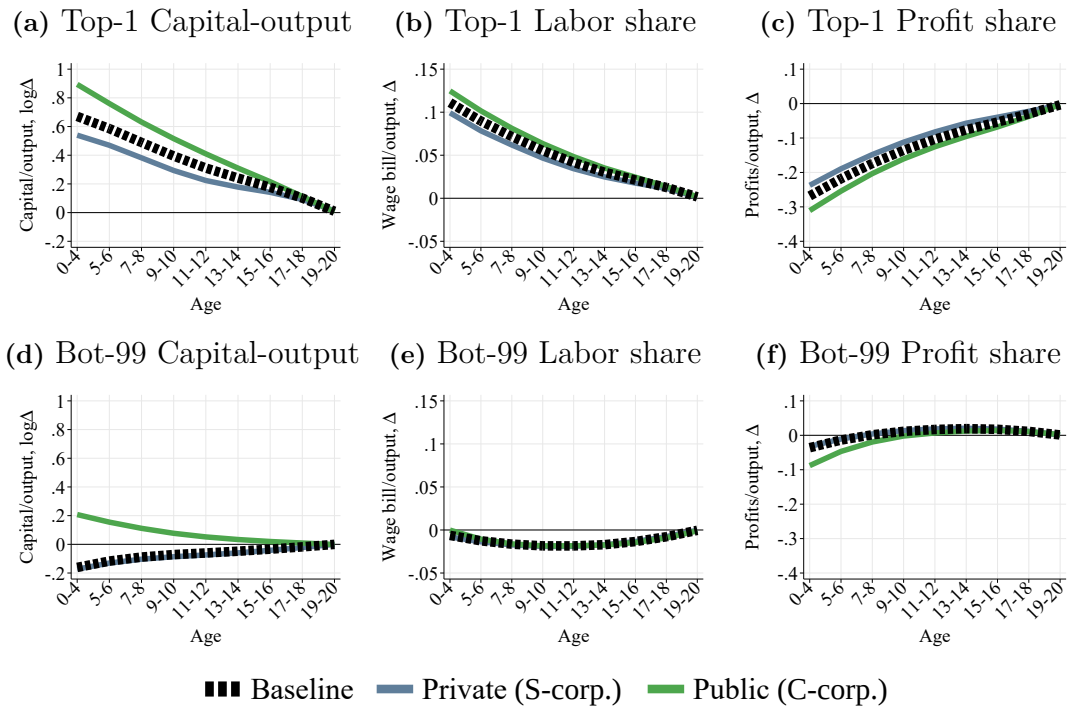
Notes: Life cycle changes for the inputs and profit shares of the top 1 and bottom 99 percent firms at age 20 estimated using (1), separately for each country. Changes are relative to the omitted group (age 19-20). The selected countries are Belgium (BE), Spain (ES), Finland (FI), and France (FR). The dashed dark line is the average value across countries for the life cycle estimates. The lines are smoothed scatterplots generated through locally weighted regressions on the estimated parameters.

Figure A.14: Life Cycle Changes of Top and Bottom Firms by Sector



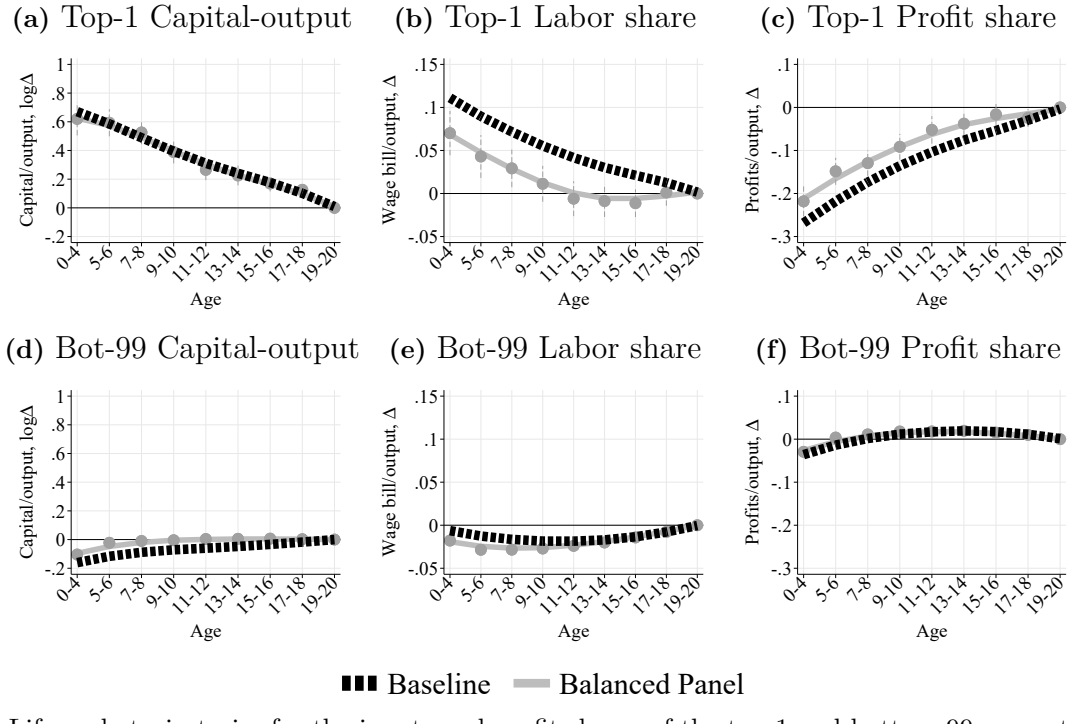
Notes: Life cycle changes of the top 1 and bottom 99 percent firms at age 20 estimated using a modified version of (1) with interactions by sector. Changes are relative to the omitted group (age 19-20). See Table A.2 for the definition of the sectors. The dashed dark line denotes the baseline estimation. All the lines are smoothed scatterplots generated through locally weighted regressions on the estimated parameters.

Figure A.15: Life Cycle Changes of Top and Bottom Firms by Legal Status



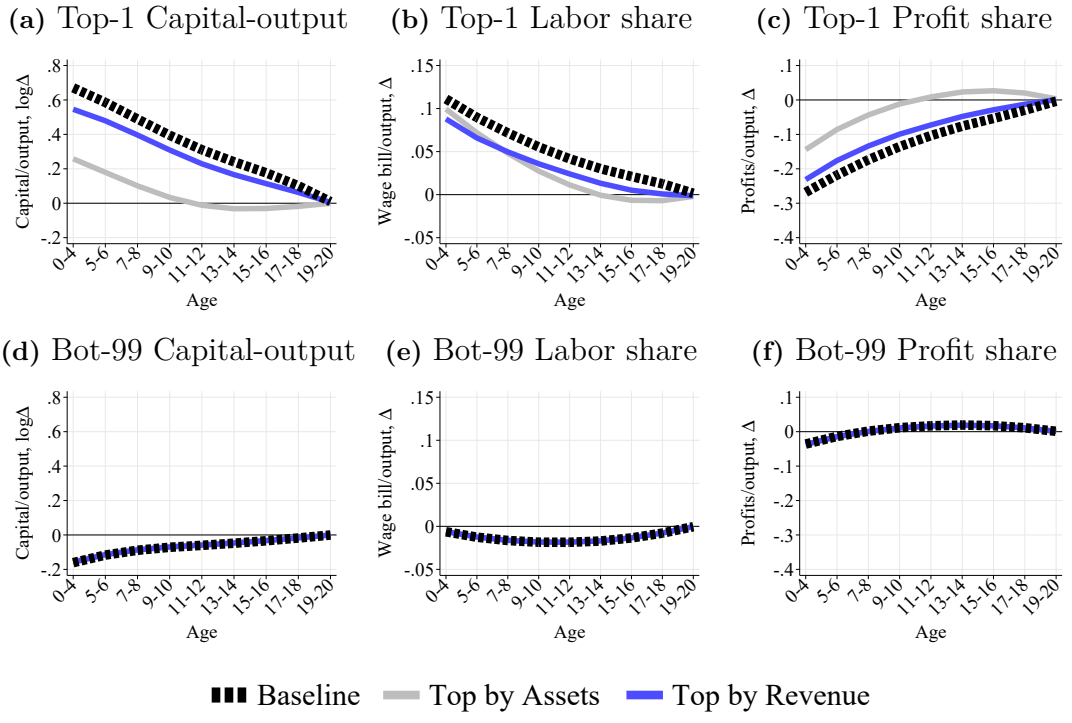
Notes: Life cycle changes of the top 1 and bottom 99 percent firms at age 20 estimated using a modified version of (1) with a public firm dummy. Changes are relative to the omitted group (age 19-20). Public Firms are analogous to U.S. C-corporations, and Private Firms are partnerships and private limited companies, analogous to U.S. S-corporations. The dashed dark line denotes the baseline estimation. All the lines are smoothed scatterplots generated through locally weighted regressions on the estimated parameters.

Figure A.16: Life Cycle Changes of Top and Bottom Firms in Balanced Panel



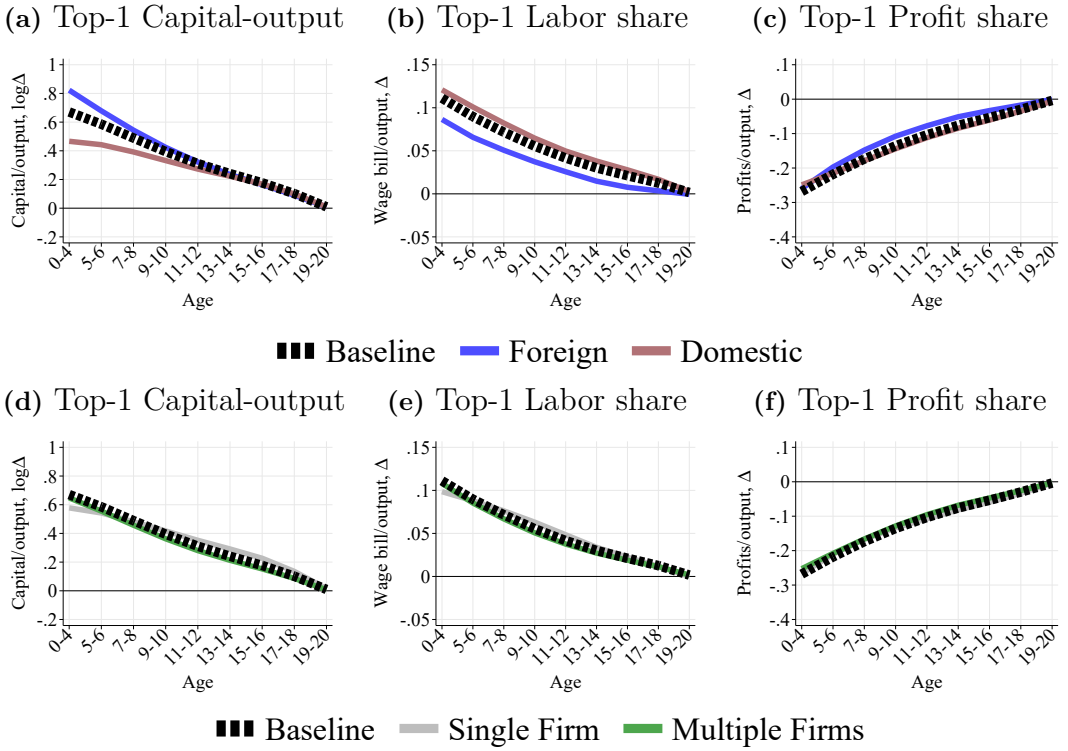
Notes: Life cycle trajectories for the inputs and profit shares of the top 1 and bottom 99 percent firms at age 20 estimated using (1) for a balanced panel of firms. The solid lines represent smoothed scatterplots generated through locally weighted regressions. The dashed vertical lines indicate 95% confidence intervals considering firm-level clustered standard errors. The dashed dark line denotes the baseline estimation.

Figure A.17: Life Cycle Changes of Top and Bottom Firms by Assets and Revenue



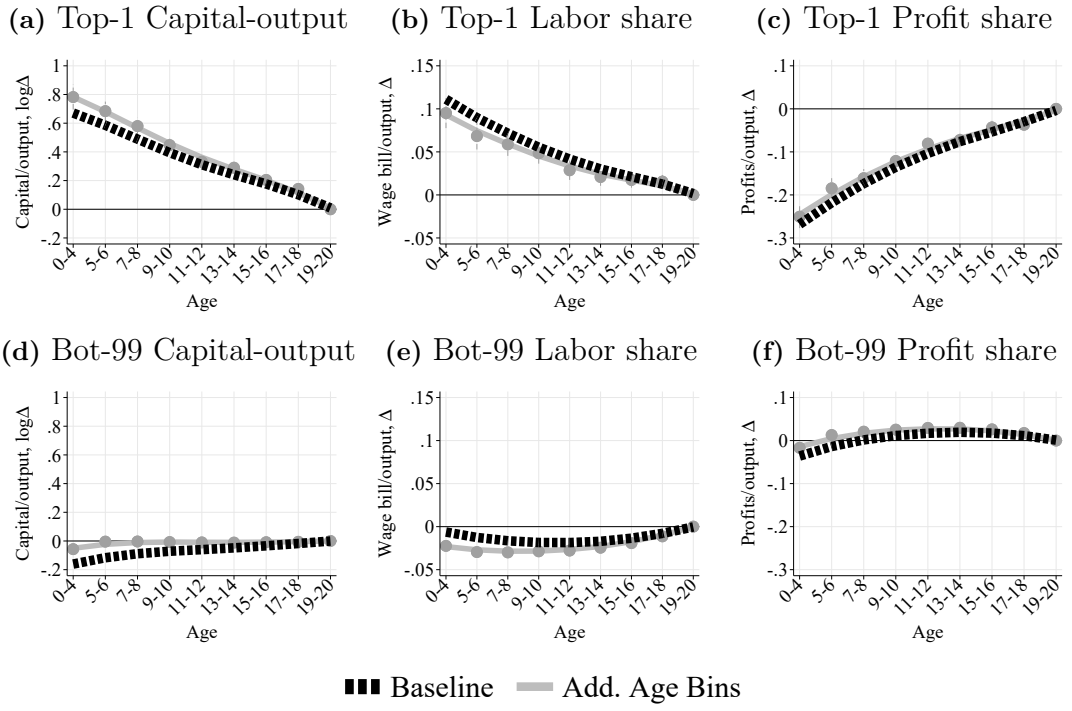
Notes: Life cycle changes of the top 1 and bottom 99 percent firms at age 20 estimated using (1), classified using alternative measures of firm size. The solid lines represent smoothed scatterplots generated through locally weighted regressions. The dashed dark line denotes the baseline estimation.

Figure A.18: Life Cycle Changes of Top Firms by Owners' Characteristics



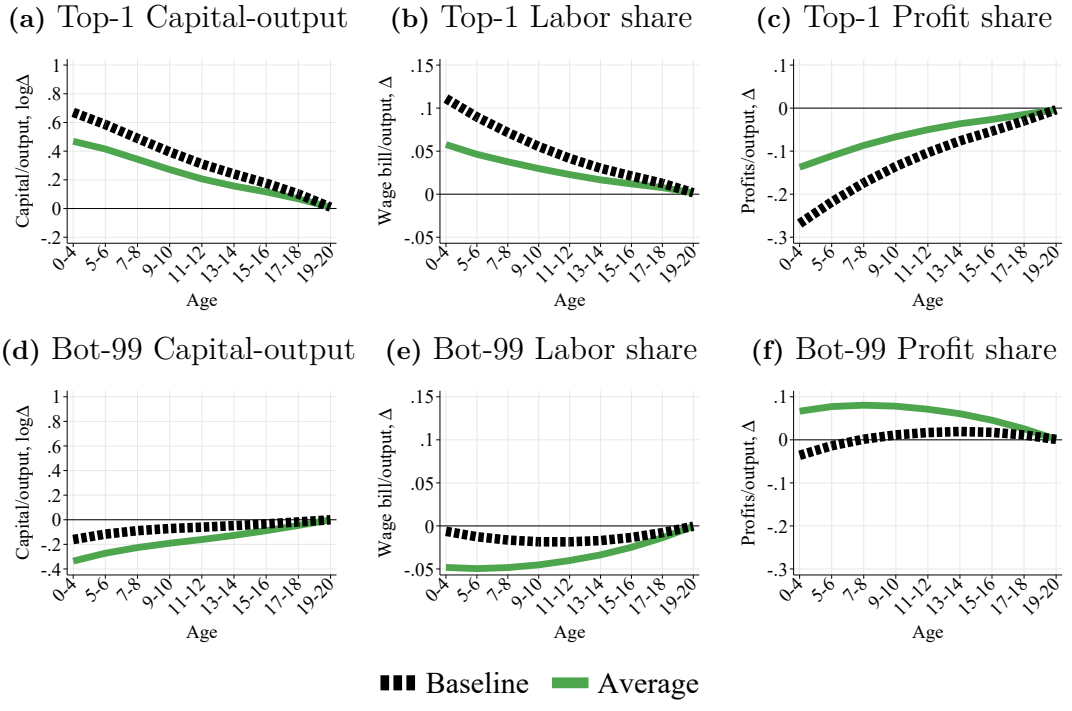
Notes: Life cycle changes of top firms at age 20 estimated using a modified version of (1). The solid lines represent smoothed scatterplots generated through locally weighted regressions. The dashed dark line denotes the baseline estimation. Ownership classifications are done using ultimate ownership linkages.

Figure A.19: Life Cycle Changes of Top and Bottom Firms With Additional Age Bins



Notes: Life cycle changes of the top 1 and bottom 99 percent firms at age 20 estimated using a modified version of (1) with additional age bins: 21-22, 23-24, 25-26, 27-28, 29-30, and 31+, interacted with firm group dummies. Changes are relative to the omitted group (age 19-20). The solid lines represent smoothed scatterplots generated through locally weighted regressions. The dashed vertical lines indicate 95% confidence intervals considering firm-level clustered standard errors. The dashed dark line denotes the baseline estimation.

Figure A.20: Average Firm-Level Change Relative to Age 19-20 of Top and Bottom Firms



Notes: Simple average of firm-level changes relative to the age 19-20 value. We use residualized variables, after controlling for industry and time fixed effects. For each firm, we first compute the average at age 19-20, $\overline{x}_{i,19-20}$; we then compute firm-level changes $x_{it} - \overline{x}_{i,19-20}$; finally, we compute the average across firms for each age bin. The lines represent smoothed scatterplots generated through locally weighted regressions. The dashed dark line denotes the baseline estimation.

Table A.4: Firms' Taxonomy by Owners' Characteristics

	By Size at Age 20		
	Top-1	Bot-99	Rest
(a) *Direct Ownership Linkages*			
Domestic owner	0.73	0.80	0.63
Foreign owner	0.23	0.02	0.03
Multiple firms	0.55	0.19	0.17
Multiple firms + Domestic	0.49	0.18	0.16
Multiple firms + Foreign	0.06	0.00	0.01
(b) *Ultimate Ownership Linkages*			
Domestic owner	0.55	0.77	0.60
Foreign owner	0.39	0.03	0.04
Multiple firms	0.78	0.33	0.26
Multiple firms + Domestic	0.46	0.27	0.20
Multiple firms + Foreign	0.29	0.02	0.02

Notes: Share of firms by owners' characteristics. Domestic (Foreign) owner equals one if the owner is registered in (outside) Spain. Both variables are zero if the owner's country is unknown. Multiple firms is equal to one if the owner owns at least one other firm inside Spain in the same year. Shares are averages at the firm-level, after taking each firm's equity-weighted average over different owners and years. For the baseline sample of firms, we focus on the years before age 20 or the closest year with available ownership data. The ownership data is available for 97% of top 1 percent firms at age 20, 64% of the bottom 99 percent, and 46% of the remaining firms. Additional details are provided in Appendix A.1.

B Model Appendix

B.1 Profit Maximization Problem

For ease of notation, we remove the time t and individual firm i subscripts.

Capital and Labor. The FOC of problem (4) are

$$\theta^h (1-\alpha)^{\frac{1}{\sigma}} \frac{y}{(x^h)^{\frac{\sigma-1}{\sigma}}} \left(l - \kappa_l^h\right)^{-\frac{1}{\sigma}} = w$$
$$\theta^h \alpha^{\frac{1}{\sigma}} \frac{y}{(x^h)^{\frac{\sigma-1}{\sigma}}} \left(k - \kappa_k^h\right)^{-\frac{1}{\sigma}} = R + \mu,$$

where μ is the lagrange multiplier of the capital collateral constraint. Combining the FOC, we get that the optimal labor choice as a function of capita is

$$l = \left(\frac{1-\alpha}{\alpha}\right) \left(\frac{R+\mu}{w}\right)^{\sigma} \left(k - \kappa_k^h\right) + \kappa_l.$$

Using this, the input aggregator x^h can be written as

$$x^h = \underbrace{\left[\alpha^{\frac{1}{\sigma}} + (1-\alpha)^{\frac{1}{\sigma}}\left[\left(\frac{1-\alpha}{\alpha}\right)\left(\frac{R+\mu}{w}\right)^{\sigma}\right]^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}}_{A_k}\left(k-\kappa_k^h\right)$$
$$= A_k\left(k-\kappa_k^h\right).$$

Finally, using the FOC of capital we get the optimal capital choice as

$$k = \left[\frac{\theta^h \alpha^{\frac{1}{\sigma}} \exp(u^h+z)}{A_k^{\frac{\sigma-1}{\sigma}-\theta^h}(R+\mu)}\right]^{\frac{1}{1-\theta^h}} + \kappa_k^h.$$

To find the *unconstrained* solution we set $\mu = 0$, then the optimal labor and capital choices are

$$l^* = \left(\frac{1-\alpha}{\alpha}\right)\left(\frac{R}{w}\right)^{\sigma}(k^*-\kappa_k) + \kappa_l^h \tag{13}$$

$$k^* = \left[\frac{\theta^h \alpha^{\frac{1}{\sigma}} \exp(u^h+z)}{(A_k^*)^{\frac{\sigma-1}{\sigma}-\theta^h} R}\right]^{\frac{1}{1-\theta^h}} + \kappa_k^h, \tag{14}$$

where $A_k^* = \left[\alpha^{\frac{1}{\sigma}} + (1-\alpha)^{\frac{1}{\sigma}}\left[\left(\frac{1-\alpha}{\alpha}\right)\left(\frac{R}{w}\right)^{\sigma}\right]^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}$. To solve for the *constrained* problem, first, we need to check if the solution is constrained and feasible. Thus, if $k^* < \bar{k}^h\left(a, s^h\right)$, then $k = k^*$; otherwise, $k = \bar{k}^h\left(a, s^h\right)$. Moreover, if $\bar{k}^h\left(a, s^h\right) > \kappa_k^h$, the solution is feasible; otherwise, we set the value of the firm to $-\infty$, forcing it to exit.[50] Under the constrained solution, we know that $\mu > 0$, which we find implicitly from

$$\bar{k}^h\left(a, \mathbf{s}^h\right) = \left[\frac{\theta^h \alpha^{\frac{1}{\sigma}} \exp(u^h+z)}{A_k^{\frac{\sigma-1}{\sigma}-\theta^h}(R+\mu)}\right]^{\frac{1}{1-\theta^h}} + \kappa_k^h.$$

Once we find μ, we can compute the optimal solution as

$$l = \left(\frac{1-\alpha}{\alpha}\right)\left(\frac{R+\mu}{w}\right)^{\sigma}\left(\bar{k}^h\left(a, \mathbf{s}^h\right) - \kappa_k\right) + \kappa_l^h \tag{15}$$

$$k = \bar{k}^h\left(a, \mathbf{s}^h\right). \tag{16}$$

B.2 Input Non-Homotheticity Microfoundations

In this appendix, we provide a microfoundation for the production technology with input non-homotheticities using homothetic technologies. Without loss of generality, we simplify the non-homothetic technology to be $y = z\left(x - \kappa_x\right)^{\theta}$, where x input rental price is p_x, and

[50] An alternative assumption is that if the firm is unable to satisfy the capital non-homotheticity requirement, it remains dormant for that period and may produce later if it manages to obtain sufficient funds.

there are no financing frictions. For this technology, we can solve analytically for the x/y ratio as:

$$\frac{x}{y} = \frac{\theta}{p_x}\left[1 + \frac{\kappa_x}{\left[\frac{\theta z}{p_x}\right]^{\frac{1}{1-\theta}}}\right]. \tag{17}$$

Input Indivisibilities. First, we can microfound this technology by assuming that firms produce using of two types of inputs x: a standard divisible input, $\tilde{x}$, and an indivisible input with a renting cost of $p_x\kappa_x$. The production technology is $y = z\tilde{x}^\theta \mathbf{1}_x$ such that $\mathbf{1}_x = 1$ if the indivisible input was purchased. Thus, the firm profits are now

$$\pi = \max\left\{\underbrace{\max_{\tilde{x}} z\tilde{x}^\theta - p_x\left(\tilde{x} + \kappa_x\right)}_{\text{rent indivisible input}}, \underbrace{0}_{\text{not produce}}\right\}.$$

Thus, if the firm decides to produce, the optimal choice of $\tilde{x} = \left[\frac{\theta z}{p_x}\right]^{\frac{1}{1-\theta}}$ and input-to-output ratio is simply $\frac{\tilde{x}}{y} = \frac{\theta}{p_x}$. However, if we redefine the input as $x = \tilde{x} + \kappa_x$, then doing some algebra we get

$$\frac{x}{y} = \frac{\theta}{p_x} + \frac{\kappa_x}{y} = \frac{\theta}{p_x}\left[1 + \frac{\kappa_x}{\left[\frac{\theta z}{p_x}\right]^{\frac{1}{1-\theta}}}\right],$$

which is equivalent to (17).

Input-TFP Complementarities. Next, we show how a technology featuring *learn-by-doing*, which decreases as firms grow, can serve as a microfoundation for the non-homothetic technology. Now we assume that the firm's TFP is $z\left(\mathbf{s}, x\right)$ and the production technology is $y = z\left(\mathbf{s}, x\right)x^\theta$, then the profits are

$$\pi = \max_x z\left(\mathbf{s}, x\right)x^\theta - p_x x$$

the FOC is

$$\frac{\partial z\left(\mathbf{s}, x\right)}{\partial x}x^\theta + \theta z\left(\mathbf{s}, x\right)x^{\theta-1} = p_x.$$

Note that the term $\frac{\partial z(\mathbf{s},x)}{\partial x}x^\theta$ captures additional marginal benefit if there are complementarities. Also, the input-to-output ratio is now

$$\frac{x}{y} = \frac{\theta}{p_x} + \frac{\epsilon_{z,x}}{p_x},$$

where $\epsilon_{z,x} = \frac{\partial z(\mathbf{s},x)/z(\mathbf{s},x)}{\partial x/x}$ is the elasiticity of the productivity to the input choice. We can interpret this elasticity as capturing potential *learn-by-doing*. Finally, if the elasticity at

the optimum is $\epsilon_{z,x} = \theta \frac{\kappa_x}{[\theta z/p_x]^{\frac{1}{1-\theta}}}$, so that complementarities decline as productivity rises, equilibrium input use is equivalent to that in (17).

B.3 Top Firms' Taxation with Externalities

We provide additional details on the firm problem under size-dependent taxation.

Firm problem. Denote the profit function π_{it}, given state variables $(\mathbf{s}_{it}^h, a_{it})$, such that

$$\pi_{it}(y) \equiv \max_{l,k \leq \bar{k}^h\left(a_{it},\mathbf{s}_{it}^h\right)} \exp\left(u_{it}^h + z_{it}\right) f^h(k,l) - wl - Rk - c_F$$

subject to $\exp\left(u_{it}^h + z_{it}\right) f^h(k,l) = y$, where $\{k_{it}(y), l_{it}(y)\}$ solve the standard cost minimization problem. Denote $y^* \equiv \arg\max_y \pi_{it}(y)$. For $\tau > 0$, the firms decide between maximizing output and potentially paying taxes (if $y^* > \bar{y}$) or producing at the threshold and not paying corporate taxes (i.e., bunching):

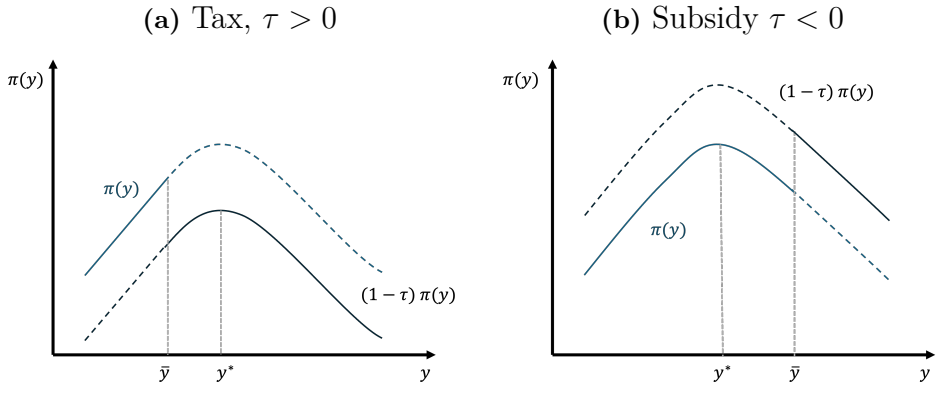
$$\max\left\{\pi_{it}(\bar{y}), \mathbb{I}_{\{y^*>\bar{y},\pi_{it}\geq 0\}}(1-\tau)\pi_{it}(y^*) + \left(1 - \mathbb{I}_{\{y^*>\bar{y},\pi_{it}\geq 0\}}\right)\pi_{it}(y^*)\right\},$$

where $\mathbb{I}_{\{y^*>\bar{y},\,\pi_{it}\geq 0\}} = 1$ if output is above the threshold and profits are nonnegative, and 0 otherwise. Notice that taxes are applied only to positive profits. Analogously, if $\tau < 0$ (subsidy) the choice is

$$\max\left\{\pi_{it}(\bar{y})(1-\tau), \mathbb{I}_{\{y^*>\bar{y},\pi_{it}\geq 0\}}(1-\tau)\pi_{it}(y^*) + \left(1 - \mathbb{I}_{\{y^*>\bar{y},\pi_{it}\geq 0\}}\right)\pi_{it}(y^*)\right\}.$$

As a result, the tax may induce some firms to scale down to avoid paying it, while a subsidy may encourage others to scale up to receive it. Figure B.1 illustrates this distortion for the case of tax, in panel (a), and a subsidy, in panel (b).

Figure B.1: Top Firms' Static Distortions from Taxation



Beyond the static output distortion, taxes also affect firm value, influencing $\bar{k}^h\left(a_{it}, \mathbf{s}_{it}^h\right)$ disproportionately more for firms that are currently at the top or more likely to reach it. Additionally, by distorting firm returns, taxes impact equity financing and savings.

These financing distortions can, in turn, lead to greater capital distortions for such firms. Furthermore, it is straightforward to see that distortions in firm value would also impact entry and exit decisions. The rest of the firm problem remains unchanged, as tax revenues are not rebated to firms.

Tax revenues. Corporate tax revenues, $\text{TR} = \int_{\{y_i > \bar{y}, \pi_i \geq 0\}} \tau \pi_i \, \mathrm{d}i$, are not rebated to firms; instead, it is as if they were rebated to households. Since we assume that the labor supply decision is not affected by income effects (e.g., GHH preferences), it remains undistorted by changes in output and tax revenues.[51] Thus, from the perspective of production, this formulation is equivalent to assuming that the tax entails a deadweight loss.

Externalities. There are two sources of externalities. Assume firm level productivity now is Pp_i where $p_i = \exp(u_i^h + z_i)$ and P is an aggregate common component, that is

$$P = \bar{P} \underbrace{\left(\int_{i \in h_{\text{top}}} p_i \mathrm{d}i \right)^{\Gamma_1}}_{\text{top tech spillovers}} \underbrace{\left(\frac{\int_{i \in \text{top } 1\%} y_i \mathrm{d}i}{Y} \right)^{-\Gamma_2}}_{\text{concentration}},$$

where $\bar{P}$ is a scaling parameter calibrated such, given externality parameters $\{\Gamma_1, \Gamma_2\}$, $P = 1$ when $\tau = 0$.

First, the top-technology spillover component of P depends on the aggregate level of top technologies, $\int_{i \in h_{\text{top}}} p_i \, \mathrm{d}i = \Omega_{\text{top}} \mathbb{E}\left[p_i \mid i \in \text{top} \right]$, which reflects both the measure of top-technology firms, Ω_{top}, and their average productivity. This component can be interpreted as the potential benefit of having more advanced technologies in the economy, which may spill over to the rest in various ways. It is an externality because top firms do not internalize that their entry could affect the productivity of other firms. The importance of this externality depends on Γ_1, which is set to $\Gamma_1 = 0$ in the baseline.

The concentration externality component of P depends on the share of top 1% firms output, $\int_{i \in \text{top } 1\%} \frac{y_i}{Y} \mathrm{d}i$. The underlying idea is that higher concentration may reduce aggregate productivity by weakening firms' incentives to innovate, consistent with the literature on market power and innovation, including the escape-competition mechanism in Aghion et al. (2005). The largest firms do not internalize that, by capturing a greater share of output, they reduce the productivity of other firms. The importance of this externality depends on Γ_2, which is set to $\Gamma_2 = 0$ in the baseline.

Counterfactuals. In our baseline exercises, we set $\bar{y}$ equal to the 99th percentile of output in the no-tax steady state and vary $\tau \in [-0.4, 0.4]$ across different externality levels. Because the externalities make P an equilibrium object, we solve for it at each new value of τ. For simplicity, however, we hold $\bar{y}$ fixed across counterfactuals. Allowing it to vary would make

[51] More generally, we can assume that $L^s \equiv \bar{L} \frac{w^{\gamma_L}}{(Y + \text{TR})^{\gamma_Y}}$, where γ_L and γ_Y capture substitution and wealth effects, respectively. In the baseline, we assume $\gamma_L = 2$ and $\gamma_Y = 0$.

the tax scheme an additional equilibrium object.

B.4 Additional Figures and Tables

Table B.1: Input Usage Over Firms' Life Cycle in Data, Model, and Alternative Theories

		Life Cycle Δ		
	σ	k/y	wl/y	k/wl
Data				
Top 1 percent firms		$(-)$	$(-)$	$(-)$
Bottom 99 percent firms		$(+)$	$(\approx)$	$(+)$
Baseline Model				
High growth + input-specific fixed costs*	any value	$(-)$	$(-)$	$(-)$
Decreasing K shadow cost	$\sigma \approx 1$	$(+)$	$(\approx)$	$(+)$
Alternative Theories for Top Firms				
Increasing markups	any value	$(-)$	$(-)$	$(=)$
Increasing L shadow cost	$\sigma < 1$	$(-)$	$(-)$	$(+)$
K-biased growth	$[\sigma < 1, \sigma > 1]$	$[(-),(+)]$	$[(+),(-)]$	$[(-),(+)]$
High growth + non-homothetic CES†	$\sigma < 1$	$(-)$	$(+)$	$(-)$

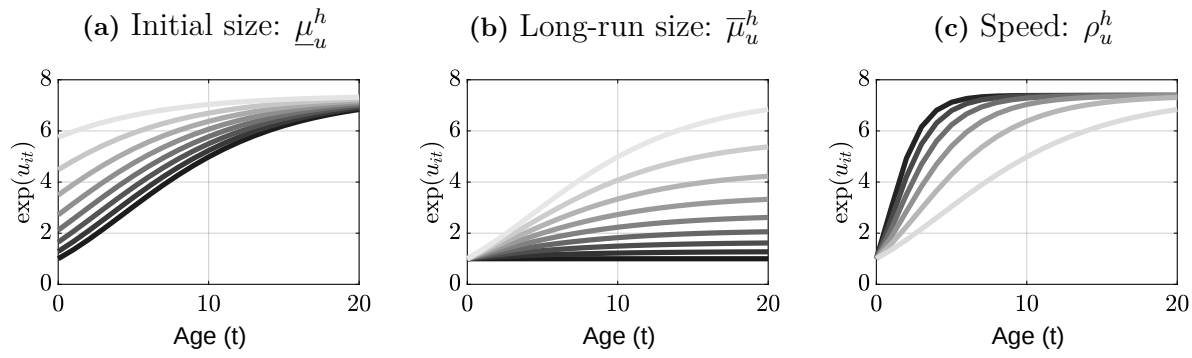
Notes: $(+)$ implies that the variable increases over firms' life cycle, $(-)$ decreases, and $(=)$ denotes no change. *Data* refers to the evidence documented in Section 3, *Model* to the baseline model implications, and *Alternative Theories for Top Firms* to the alternative hypotheses discussed in the text for the top firms input composition. *For k/wl to decline, κ_k must be larger than κ_l. †We assume that the income elasticities are such that $\epsilon_k < \epsilon_l$.

Table B.2: Interest Rate Changes and Aggregate Outcomes

	Real interest rate: r				
	0.01	0.02	0.03	0.04	0.05
Relative to $r = 0.05$ (=1)					
Share Top-h firms	5.69	2.86	1.54	1.06	1.00
Output per worker, Y/L	1.31	1.19	1.10	1.04	1.00
Labor share, wL/Y	0.92	0.94	0.98	1.00	1.00
Profit share, Π/Y	1.44	1.30	1.14	1.04	1.00
Output share top 1	2.12	1.72	1.31	1.03	1.00

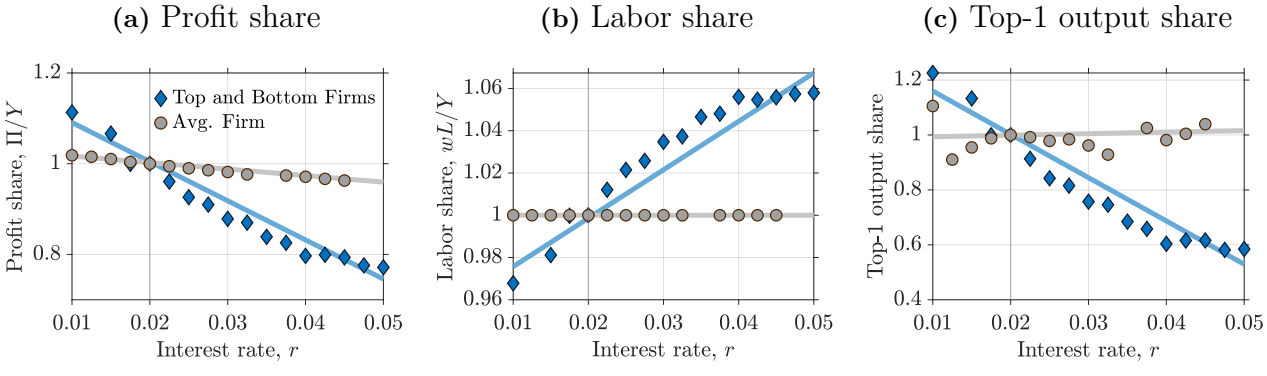
Notes: Results are reported relative to the counterfactual with an interest rate of $r = 0.05$. The "Share Top-h firms" is calculated as the measure of top-h firms divided by the total measure of firms in the economy. The baseline calibration has $r = 0.02$. We assume that the change in the interest rate, Δr, also induces a change in the implicit discount rate, Δr_β, where $r_\beta = \beta^{-1} - 1$, so that the spread $r - r_\beta$ remains unchanged.

Figure B.2: Ex-ante TFP Component



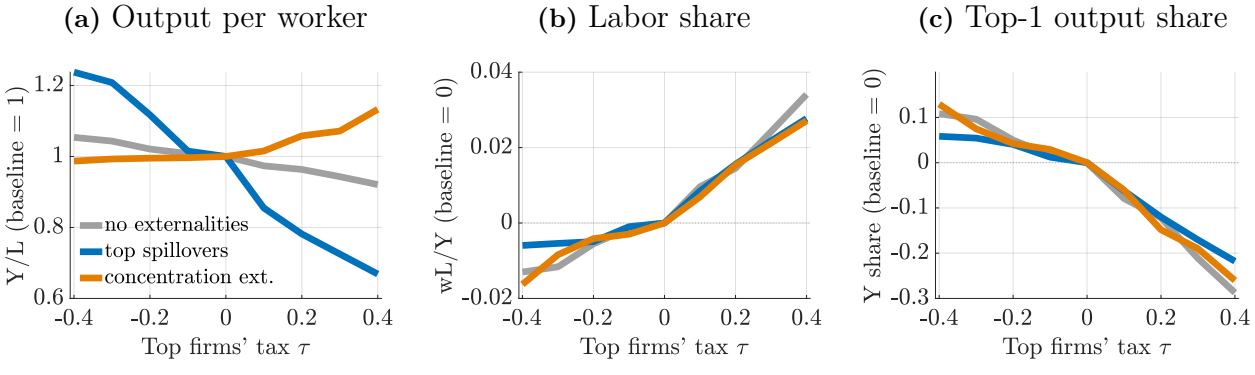
Notes: The figure shows different paths of the deterministic ex-ante TFP component $\exp(u_{it})$ for different values of $\underline{\mu}_u^h$ (panel a), different values of $\overline{\mu}_u^h$ (panel b), and different values of $\overline{\mu}_u^h$ (panel c). Lighter lines indicate higher values. Numerical example.

Figure B.3: Interest Rate Changes in Benchmark and Avg. Firm Model



Notes: 'Top and Bottom Firms" refers our baseline model that targeting the life cycle of top and bottom firms. "Avg. Firm" corresponds to the model where we target the average firm life cycle. For each calibration, all variables are normalized relative to the baseline quantification. We assume that the change in the interest rate, Δr, also induces a change in the implicit discount rate, Δr_β, where $r_\beta = \beta^{-1} - 1$, so that the spread $r - r_\beta$ remains unchanged in both models.

Figure B.4: Top Firms' Taxes with Externalities



Notes: Steady state comparisons in GE solving for new equilibrium wage w. Aggregate output in Panel (a) are normalized relative to the no-tax steady state, which is the same across counterfactuals. $\Gamma_1 = 0.2$ is the case with positive technological spillovers. $\Gamma_2 = 0.3$ is the economy with negative concentration externalities.